NIKOLA ILIC

END-TO-END ANALYTICS WITH DICROSOFT POVER B

CRASH COURSE ON BUILDING POWERFUL ANALYTIC SOLUTIONS





Foreword

Introduction

Understanding Business Problem

Data preparation

Data modeling

Data visualization

Data analysis

13(28) (50) (60)

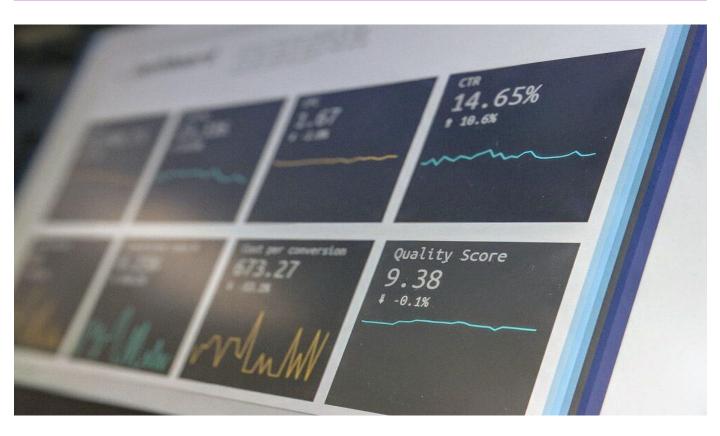
2022 All Rights Reserved



According to all relevant researches, Microsoft Power BI became is a leading tool when it comes to providing insights from the data. However, when I talk to people who are not deep into the Power BI world, I often get the impression that they think of Power BI as a visualization tool exclusively.

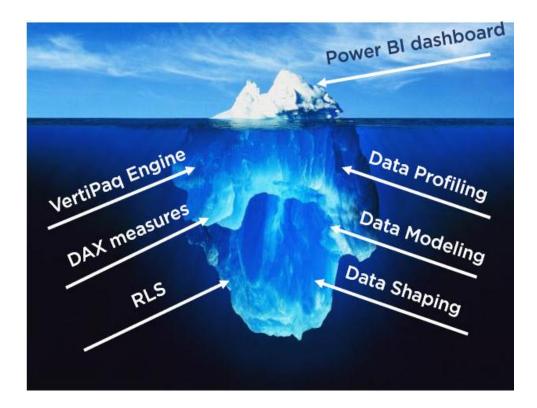
However, there is a lot more to it, as the most powerful features are expanding beyond nice visualizations.

In this brochure, I'll show you **how Power BI can be used to create a fully-fledged analytic solution**. Starting from the raw data, which doesn't provide any useful information, to building, not just nice-looking visualizations, but extracting insights that can be used to define proper actions – something that we call informed decision making.



When I talk to people who are not deep into the Power BI world, I often get the impression that they think of Power BI as a visualization tool exclusively. While that is true to a certain extent, it seems to me that they are not seeing the bigger picture – or maybe it's better to say – they see just a tip of an iceberg! This tip of an iceberg is those shiny dashboards, KPI arrows, fancy AI stuff, and so on.

However, there is a lot more to it, as the real thing is under the surface...



This underneath portion, which consists of multiple individual, but cohesive parts, enables the above-the-surface piece to shine!

In this brochure, I'll show you **how Power BI can be used to create a fully-fledged analytic solution**. Starting from the raw data, which doesn't provide any useful information, to building, not just nice-looking visualizations, but extracting insights that can be used to define proper actions – something that we call informed decision making.

Setting the stage

In this brochure, I'll use an open dataset that contains data about car collisions in New York City, and can be found <u>here</u>. This dataset contains ~1.8 million rows. Each row represents one accident that happened in New York City, where at least one person was injured/killed, or the overall damage was at least 1000\$. Data comes in the CSV file, containing 29 columns.

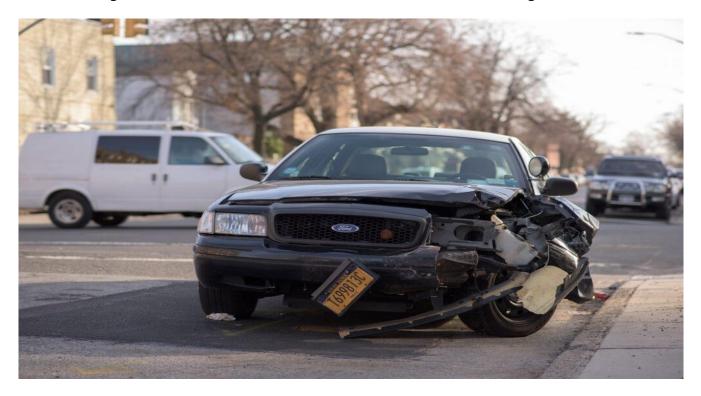


Photo by Michael Jin at Unsplash

Now, before we start building our solution, we need to define the workflow and identify specific stages in the process. So, the first and most important task is to set the steps necessary to create the final outcome. Here is my list:

Understanding Business Problem This is the starting point, as without understanding the business problem, our solution won't be able to address business needs. Do I want to increase the sales? Is customer retention my main goal? What will happen if I discard some services in the next quarter? These are some typical examples of the business questions that need to be answered using data insights. In this example, our "business" problem is to identify critical locations for collisions, and try to prevent accidents in the future

In this stage, we need to perform some steps to make our data ready for further digest. Starting with <u>data profiling</u>, so we can identify possible outliers and anomalies, then applying various <u>data shaping</u> techniques to prepare the data BEFORE it becomes part of our data model

Data Preparation

Data Modeling

As we are building an analytic solution, data model must satisfy (or at least SHOULD satisfy) some general postulates related to data modeling. For most analytical systems, including Power BI, dimensional modeling is the way to go – so, we need to decompose our original wide fact table and leverage Starschema concept to establish the proper data model

This is the stage that folks from the beginning of the brochure will like most:)...It's time to please our eyes with numbers and display them using convenient Power BI visuals

Data Visualization

Data Analysis

Having a nice visual is fine, but it needs to provide some insight to a person looking at it. Therefore, the main purpose of this phase is to provide the insight – for example, what are the peak hours for car accidents in NYC? What are the most risky locations? How many pedestrians were injured in Queens? And, so on...

This is an optional phase and could've been excluded from this solution and left completely to business stakeholders. But, hey, let's play our Data Analyst role till the end and give some recommendations based on the insights we obtained in the previous phase!

Informed Business Decisions

UNDERSTANDING BUSINESS PROBLEM

The first and most important step for building your (successful) analytic solution, in order to serve its purpose and **be adopted by the users**, is to give answers to key business questions. No one needs pretty dashboards and cool visuals if they don't provide insight and help decision-makers understand what is happening and why.

How can I increase my sales? Why did so many customers leave us in the previous quarter? What can I do to improve the delivery process? When is the best period to target the market with promotions?

These are just a few most frequent questions asked by business stakeholders. Not just that – maybe an insight into the underlying data can help users identify completely new patterns and ask a question: *are we solving the right problem?*

No one needs pretty dashboards and cool visuals if they don't provide insight and help decision-makers understand what is happening and why.

UNDERSTANDING BUSINESS PROBLEM

Therefore, it is extremely important to identify the key questions at the very beginning, so we can shape and model our data to answer those questions in the most effective way. For our dataset, we don't have to deal with "classic" business questions- as there are no sales, products, promotions... However, it doesn't make it less "worth", let alone allowing us to skip some of the steps defined above. Some of our "business" questions could be:

- ✓ What are the riskiest locations in the city?
- ✓ Which time of the day is the most critical?
- ✓ What is the percentage of pedestrians among all injured persons?
- ✓ Which city boroughs have the highest rate of accidents?
- ✓ What car types are most frequently involved in the accidents?

The final goal in finding the answers to these questions would be to identify the key indicators that cause collisions (Data Analysis stage), and somehow try to act and prevent future accidents, or at least reduce their number (making informed decisions).

UNDERSTANDING BUSINESS PROBLEM

Summary

Power BI is much more than a visualization tool! Keep repeating this sentence, and don't forget the illustration of the iceberg from the beginning.

In this chapter, we laid the theoretical background and explained the concepts which are the key pillars of every successful analytic solution. In the next chapter, we will start exploring our dataset, try to identify the possible anomalies, check if some parts of the dataset need to be enhanced or restructured, and finally shape the data in the form that will enable us to build an efficient data model for the subsequent phases in the process.

Introduction

In the previous chapter, we laid some theoretical background behind the process of building an end-to-end analytic solution and explained why it is of key importance to understand the business problems *BEFORE* building a solution. Now, it's time to pull our sleeves up and start real work with our dataset. As a reminder, we will use an open dataset about motor vehicle collisions in NYC, which can be found <u>here</u>.

First look into the dataset

Data is stored in the CSV format, and we have one flat table containing ~1.8 million rows and 29 columns. Let's take a quick look at the data once it's imported into Power BI:

	Manape Parameters • Parameters	Refresh Preview * Manage * Query	Choose Remove Keep Remove 5	11 Dise First Row as Headers *	X Append Queries • 43	Test Analytics P Yilian A gaure Machine Learning Al Insights			
Queries [1]	< ×	V fr - Table.Tra	sformColumnTypes(#"Promoted Headers",{{"CRA	SH DATE", type date), ("CRASH TIME", ty	of time), ("BOROUGH",	type text), ("ZIP CODE", Int64.Type)	, ("LATITUDE", type	Query Settings	32
Motor_Vehicle_CollisionsCrashes	.	CRASH DATE	CRASH TIME - AS BOROUGH	- 12 LATITUDE	- 1.2 LONG		· AP- ON STREET NAME	+ PROPERTIES	
C. monetering compared community		14-Apr-22	5.32.00 AM		44	nul	BRONK WHITESTONE BRIDGE	- Name	
	2	13-Apr-21	9-35-00 AM BROOKLYN	nut 11217	40.68558	-75.97617 (40.68358, -75.97617)	BRUNK WHITESTONE BRUNE	Motor_Venice_CollsionsCras	she
	4	15-Apr-22	#35.00 PM BROOKTN # 75.00 PM	11217	40.58558	-73.97617 (Hd.58356) -73.97617)	HUTCHINSON RIVER PARKING	All Properties	
	1	13-Apr-21	4 00:00 PM BROCKLYN	11222	100	145	VANDERVORT AVENUE		
	-	12 Apr 22	825 00 AM	ngi	0	0 10.0.0.01	EDSON AVENUE	APPLIED STEPS	
	-2	17-40-21	5 11 00 PM	-	and its	000,000	VERRAZANO BROGE UPPER	Source	6
	7	13-Apr-21	5.30-00 PM QUEENS	11106	100	nal	33 rt	Promoted Headers	- 0
	8	16-Apr-21	11:30:00 PM	Sine Sector	nat	[01]	SHORE PARKWAY	X Changed Type	
	8	11-40-21	5 00:00 PM	nur nur	and a	100	GOWANE/S FAMP		
	10.	16 Apr-22	3-15-00 PM	nat	nut	eut	BRONK ROAD OF WAY RAM		
	11	11-Apr-22	B-05-00 PM BROOKLYN	11226		647	TO THE ROAD		
	12	15-Apr-21	8-00-00 PM STATEN ISLAND	10304	1997		INUE		
	13	10-Apr-21	11:15:00 AM	null .			100		
	14	21-Mar-19	10-50-00 PM BROOKLYN	11201					
	14	21-Jan-20	3 49:00 PM	1100	D		all and the second		
		21.000-22	2 50 00 PM	20462	Power BI	automatically app	ollea		
	10	09-Mer-21	11-00-00 AM	10001	como tr	ansformation ste	ne		
		21-Mer-21	10-20-00 PM BROCKLYN		some u	ansionnation ste	ps		
	18		10-SR-00 PM STATEN ISLAND	11254					
		06-Apr-21		10312			· · · · · · · · · · · · · · · · · · ·		
	20	69-Apr-22	2.45.00 PM	nut			ARMAY		
	21	34-Apr-22	1.00.00 PM	nuV		101			
	22	14-Apr-21	11:00:00 AM	mali	40.694035	-73.72679 (40.894035, -73.72679)	CROSS ISLAND PARKWAY		
	23	15-Apr-21	1:30:00 PM BRONX	20452	40.857365	-73.84657 (40.857365, -73.84657)			
	24	14 Apr-31	2.40.00 PM	naV	40.598307	-73.91837 (40.698807, -75.91837)	MYRTLE AVENUE		
	-25	14-Apr-21	\$143.00 PM QUEENS	11429	40,75402	-75 74827 (40.71402, -75.74827)			
	26	16-Apr-33	5.40.00 PM BRONX	30474	40.825	-75.89402 (40.815, -75.89402)	GARRISON AVENUE		
	27	16-Apr-21	12:50:00 AM	nut	40.55079	-74,20098 (40,55079, -74,20098)	HUGUENOT AVENUE		
	28	15-Apr-21	10:30:90 AM BROOKLYN	11207	40.655900	-73.89837 (40.655903, -73.89817)			
	- 29	16 Apr 21	4.35.00 PM BRONX	20475	40.890076	-73.819855 (40.890076, -73.819855)	BOSTON ROAD		
	30	15-Apr-21	5.20:00 PM BROOKLYN	11296	40.650402	-75.89422 (40.650402, -73.89422)	GLENWOOD ROAD		
	31	16-Apr-21	8-20-00 PM MANHATTAN	20025	40 79535	-73 97275 (40.79335, -73.97275)	WEST 94 STREET		
	32	16-Apr-21	5-20-00 PM MANHATTAN	10012	40.72538	-74.00011 (40.72538, -74.00011)	PRINCE STREET		
	33	15-Apr-21	2.30.00 PM QUEENS	11377	40.75284	-73.90358 (40.75184, -73.90358)	BROADWAY		
	34	16-Apr-21	12.20:00 PM BROCKLYN	11226	40.649788	-73.9622 (40.649788, -73.9622)	EAST 19 STREET		
	35	16-Apr-21	6 25 00 PM BROOKLYN	23222	40.586928	-73.920815 (40.688928, -73.920815)			
	36	14-Apr-22	1.00-00 PM BROOKLYN	11211	40.722963	-73.93647 (40.712965, -73.93647)			
	37	14-Apr-21	# 14:00 PM	nut	40.801285	-73.95394 (40.801285, -75.95394)	7 AVENUE		
	38	14-Apr-21	3-25-00 PM BROOKLYN	11220	40.633976	-74.02211 (40.63397674.02211)			

Before we go deeper into specific challenges related to data modeling, let me briefly stop here and state a few important things:

Power BI (or to be more specific, Power Query Editor), automatically applied some transformation steps and start <u>shaping our data</u>. As you can see, Promoted Headers transformation took first row values and set them as column names, while Power Query also changed the type of various columns

Data Preparation

Here starts our journey! This is the first station to debunk the myth about Power BI as a visualization tool only. Let me quickly explain why: you could just hit that *Close & Apply* button in the top left corner of the Power Query Editor and start building your visualizations right away!

But, the fact that you **CAN** do something, doesn't mean that you **SHOULD**...For some quick ad-hoc analysis, you may sneak through without applying additional steps to shape and prepare your data, but if you plan to build a robust and flexible analytics solution, that would be able to answer a whole range of different business questions, you would be better spending some time to face-lift your data and establish a proper data model.

Since we are dealing with CSV file in our example, Power Query Editor is the obvious place to apply all of our data preparation work. If we were to use, for example, SQL database as a data source, we could've also performed data shaping on the source side – within the database itself!

Here, as a best practice, I'll quote Matthew Roche's famous "maxim":

Data should be transformed as far upstream as possible, and as far downstream as necessary...

-Matthew Roche-

Data Profiling

For the starter, Power Query Editor offers you a very handy set of features to perform <u>data</u> <u>profiling</u>. I'll go to the View tab and turn on *Column quality, Column distribution and Column profile* features to help me better understand the data and identify potential issues that need to be resolved.

ny ngn Layout Data Preve	lumin profile Go tr Go tr Colum	kh-	Advanced Q Editor Deper	very sdencies sdencies							
eries (1)			Construction of the second second second	Provide the second s		a desident lines that is in the star		BOROUGH", type text), ("2			Query Settings
Notor_Vehicle_CollisionsCh	CRASH DAT	(Interference	CIUASH TIME	AC BOROUGH		and the second se				ARE ON STREET NAME	Name
-	• Valid	100%		 Valid 						Valid	Motor_Vehide_CollsionsCrashes
	Error Error		Error Empty 2	Error Empty						Error Error	
	1		2	1		11.	1	1	1	1	All Properties
			him			Innouse		1		Iture	→ APPLIED STEPS
				111.						IIIIIIIIIIIIIIIIIIIIIIIIIIIII	Source
	35 distinct, 15 u	rique i	179 distinct, 286 unique	A distinct, 8 unique		151 distinct, 27 unique	887 distinct, 862 unique	BB1 distinct, #52 unique	890 distinct, 868 wrigae	495 distinct, 362 unique	Promoted Headers
	1	34-Apr-22	5:32:00 AA			nut				BRONK WHITESTONE BRIDGE	X Changed Type
	2	13-Apr-21		BROOKLYN		11217			(40.68358, -73.97617)		
	3	15-Apr-21	4 15 00 PM			nul				HUTCHINSON RIVER PARKWA	
	4	12-Apr-22		BROOKLYN		11222				VANDERVORT AVENUE	
	5	12-Apr-21	8.25.00 AN			nul			(0.0, 0.0)	EDSON AVENUE	
	6	13-Apr-21	5-11-00 PM			nul				VERRAZANO BRIDGE UPPER	
	7	13-Apr-21	5:30:00 PA			11106				33 st	
		16-Apr-21	11:30:00 PA	-		nuf				SHORE PARKWAY	
	9 10	11-Apr-22 16-Apr-21	5.06.00 Pt 9-15:00 Pt	1 /		null null				GOWANUS RAMP BRONX RIVER PARKWAY RAM	
	51	10-Apr-21		BROOKLYN		11226				BEVERLEY ROAD	
	12	15-Apr-21		STATEN ISLAND		10304				GREENFIELD AVENUE	
	13	10-Apr-21	21 15 00 AM			nul				BROOKLYN BRIDGE	
	54	22-May-29		BROOKLYN		11201			(40.69754, -73.98312)	GOLD STREET	
	15	21-Jon-20	3 49:00 PM			nuk				BRUCKNER BLVD	
	16	36-Feb-22	2:50:00 PM			10451			(40,843464, -73,836)		
	17	09-Mpr-21	11.00.00 AM	1		nul	40.69254		(40.692547, -73.990974)	COURT STREET	
	10 4				_	1 (AC)				>	
			W.								
	Column statistics		···· Value distribut	lion							
	Count		1000 BROOKLYN								
	Greor		Ø QUEENS								
	Empty		0 BROND								
	Distinct		6 MANHATTAN								
	Unique		0 STATEN ISLAND								
	Empty string		363								
	Min		100								
	Max	STA78	N.								

This will enable me to immediately spot that, for example, there are 36% of missing values for the Borough column. Based on the findings, I can decide to leave it like that, or apply some additional transformations to fix the missing or incomplete data. For example, I can decide to replace all blank or null values with N/A or something similar.

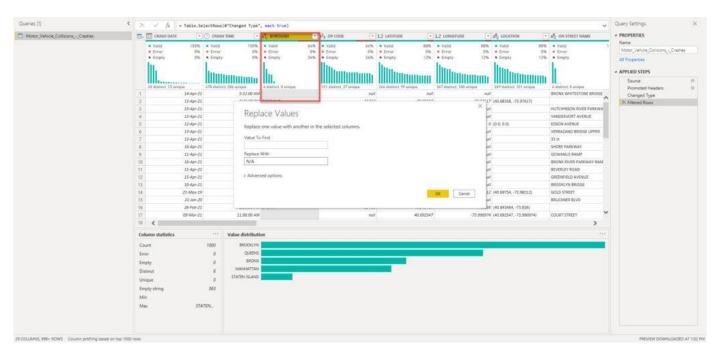
I could also quickly identify outliers or anomalies (if any). Let's imagine that we profile *Number* of *Persons Injured* column:

Motor_Vehicle_CollisionsCrashes		 A⁰_C CROSS STREET NAME 	A ^B C OFF STREET NAME	121 NUMBER OF PERSONS ININ	H - 123 NUMBER OF PERS	ONS KILLED • 123 NUMBER OF PEDEST	RIANS INJU 💌 1 ² 3 NUMBER OF PEDESTRIA	NS XIII = 1 ² 5 N	
				Valid	100% Valid	100% • Va6d	102% • Yald	100% • Va	
				Error Empty	0% Error 0% Errorty	0% Error 0% Error	0% Error 0% Error	0% * E= 0% * E=	
		L L L L L L L L L L L L L L L L L L L	I complete	a singer	I I	and a strategy	and a singly		All Properties
		li .		di ser si					 APPLIED STEPS
		n IIIIIIIIIIIIIIIII		1.					Source
		378 distinct, 318 unique	263 distinct 261 unique	8 distinct, 3 unique	2 distinct, 0 unique	3 distinct, 3 sestase	2 distinct, 1 unique	2 dist	Promoted Headers
	1				0	0	0	° ^	X Changed Type
	2		620 ATLANTIC AVENUE		2	0	2	0	
	3	ANTHONY STREET			0	0	0 0	0	
	4	HIGHTURET SCHEET,			0	0	0	0	
	6				0	0	0	0	
	7	Slave			0	0	0	0	
	8				Ø	0	0	0	
	9				1	0	0	0	
	10				0	0	0	0	
	11	EAST 21 STREET			1	0	3)	0	
	12	OSGOOD AVENUE			0	0	0	0	
	13				0	0	0	0	
	14	CONCORD STREET			0	0	0	0	
	15	WAST 138 STREET			0	0	0	0	
	16	JORALEMON STREET	2819 MIDDLETOWN ROAD		0	0	0	° ~	
	37 18				1	0	0	2	
	1.19.11	•	and the second						
	Colum	un statistics	··· Value distribution					A37.	
	Count	1000							
	Error	0	<u>^</u>						
	Empty	0							
	Distin								
	Uniqu								
	NaN								
	Zero	678							
	Min	0.48							
	provide statements	0							
	Max	7							
	Avera		Y OF NATE NO.						

If there were some data anomalies (i.e. instead of 7 for the Max number of injured persons, let's say 7000), we would be able to spot that right away and react accordingly!

Data Shaping

It's time to enhance our dataset and invest some additional effort to improve the data quality. Let's start with replacing blank values with N/A in the *Borough* column:



The next step will be to clean the numeric columns. ZIP Code is the whole number column, while Latitude and Longitude are represented as decimal values. That being said, we will replace nulls with 0 value in each of these columns:

Motor_Vehicle_CollisionsCrashes	🛄 🥅 CRASH DAT	E CRASH TH	at	· 121 26 COOL	* 12 JATTUOK	* 12.10NG	TUDE .	A LOCATION	· AC ON STREET NAME	PROPERTIES
	. Valid	100% • Valid		10% • Valid	sim • Valid	91% • Vator	91%		· Valid	Name
	• Error	0% • Error		0% · Error	0% Error	0% · Error	0%		Mi · Error	Motor_Vehicle_Collisions, - Crasher
	• Empty	0% Empty	0% Empty	0% • Empty	36% Empty	9% Empty		• Empty 3	% Empty	All Properties
				illin					li.	APPLIED STEPS
			HILLING IN		HINT MANAGEMENT	annan Ham			n IIIIIIIIIIIIIII	
	35 distinct, 15 u		A unique à distinct 0 unique	151 distinct 27 unique	US7 distinct 64	2 uninum HB2 diation	1, 852 unique	890 distinct, 848 unique	495 distinct, 362 unique	Source Promoted Headers
	1	14-Apr-71	5:32:00 AM N/A	Contraction of the second	naf	Ban	nul		BRONX WHITESTONE BRIDGE	Changed Type
	2	19-Apr-21	9-35-00 PM BROOKLYN		11217	40.68358	-73.97617	(40.68158, -73.97617)	^	Filtered Rows
	3	15-Apr-21		-			10		HUTCHINSON RIVER PARKWA	> Replaced Value
	4	13-Apr-22	Dealers Malves				×		VANDERVORT AVENUE	
	5	12-Apr-21	Replace Values				0	(0.0, 0.0)	EDSON AVENUE	
	6	13-Apr-21	Replace one value with another	in the selected columns			Arti-		VERRAZANO BRIDGE UPPER	
	7	19-Apr-21					LOT.		33 st	
	8	16-Apr-21	Value To Find				Law.		SHORE PARKWAY	
	9	11-Apr-22	nul				2.4Y		GOWANUS RAMP	
	10	16-Apr-21	Replace With				part .		BRONX RIVER PARKWAY RAM	
	11	11-Apr-21	0				A.F.		BEVERLEY ROAD	
	12	15-Apr-25					207		GREENFIELD AVENUE	
	13	10-Apr-21				OK	Cancel 47		BROOKLYN BRIDGE	
	14	22-May-29					42	(40.69754, -73.98312)	GOLD STREET	
	15	21-Jon-20					ist		BRUCKNER BLVD	
	16	26-Feb-22	2:50:00 PM BRONK		20462	40.843454	-72.836	(40,843464,-73,836)		
	将	09-Mor-22	11:00:00 AM N/A		nuť	40.692547	-73.990974	(40.692547, -73.990974)	COURT STREET	
	18	31-Mor-21	10:20:00 PM BROOKLYN		21254	40.626457	-73.918	(40.626457, -75.918)	RALPH AVENUE	
	19	06-Apr-22	10-58-00 PM STATEN ISLAND		20312	40.526894		(40.526894, -74.16728)	BARCLAY AVENUE	
	20	09-Apr-21	2:45:00 PM N/A		nuli	40.840775	-73.87245	(40.840775, -73.87246)	BRONX RIVER PARKWAY	
	21	14-Apr-21	1:00:00 PM N/A		nut	nut	nai		MAJOR DEEGAN EXPRESSWA'	
	22	34-Apr-25	11.00.00 AM N/A		nuť	40.694035	-73.72579	(40.694035, -73.72679)	CROSS ISLAND PARKWAY	
	23	25-Apr-21	2:30:00 PM BRONK		20462	40.857355	-73.84657	(40.857365, -73.84657)		
	24	34-Apr-23	2.40.00 PM N/A		nuš'	40.698807	-73.92837	[40.698807, -73.92837]	MYRTLE AVENUE	
	25	34-Apr-23	9-43-00 PM_QUEENS		11429	40,72402	-73.74827	(40.71402, -73.74827)		
	26	16-Apr-21	5:40:00 PM BRONK		20474	49.815	-78.89402	(40.815, -73.89402)	GARRISON AVENUE	
	27	16-Apr-21	22 50:00 AM N/A		nuť	40,55079		(40.55079, -74.20098)	HUGUENOT AVENUE	
	28	15-Apr-21	20-50-00 AM BROOKLYN		21207	40.655903		(40.655903, -73.89817)		
	29	16-Apr-21	4.35-00 PM BRONX		20475	40.890076		(40.890076, -75.819855)	BOSTON ROAD	
	30	15-Apr-21	5-20:00 PM BROOKLYN		11235	40.650402		(40.650402, -73.89422)	GLENWOOD ROAD	
	31	16-Apr-21	9.20:00 PM MANHATTAN		10025	40.79335		(40.79335, -73.97275)	WEST 94 STREET	
	12	47.8.11.84	P. S. S. Sec. Les Accession Press		addeb.	10 201 00	The delates of	Last WARTER The Milder of	ADDALOW ADDALOW	

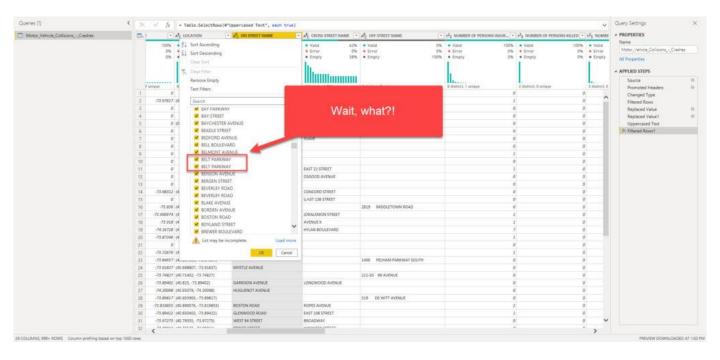
That was quick and easy, right? Now, let's move on and try to profile other columns and check if some more sophisticated transformations are needed. Column *On Street Name* is extremely important, because it's needed to answer one of the crucial business questions: what are the riskiest locations in the city? Therefore, we need to ensure that this column has the highest level of data quality.

Motor_Vehicle_CollisionsCrashes	. OE - 12 LONGITUDE	- A ^B C LOCATION -	CON STREET NAME	NAME - A ^R C OFF STREET NAME	T 123 NUMBER OF PERSO	INS POUR \$23 NUMBER	 PROPERTIES
	88% • Vaid 0% • Error 12% • Error	89년 21 Sort Ascending 이체 12년 3년 Sort Descending Clear Sort	• Valis • Error • Erroty	1% Valid 0% Ecror 90% Employ	0% Valid 0% Empr 100% Empty	100% • Valid 0% • Error 0% • Empty	Name Motor_Vencie_Collsons,*_Crashes All Properties • APPLIED STEPS
	1. P9 unique 247 distinct 109 u	nique 2 nique 2 nique 2	12 district, 11 seld		10 shatruct, 3 ornigoat	2 diatout, 0 ur 0	Source Promoted Headers Changed Type
	2 muli	mu0 Search	150 street			0	X Filtered Rows
	3 null	Aud BAY STREET	Van wyck expresse	nay		0	
	4 null	AUF				0	
	5 null	null 🐱 Beadle Street	Rockaway bouleva	rd .		0	
	6 null 7 null	AND BEDFORD AVEN				2	
	7 mwl	Aud Sell BOULEVARD				.0	
	¢	BERGEN STBEET BEVERLEY ROAD BEVERLEY ROAD BEVERLEY ROAD BOADEN AVENU BORDEN AVENU BORDEN AVENU	🗸 🛛 In rea	ality, this is one s	same street!	>	
	Column statistics	Value di Perint and strette List may be incor				4.7	
	Count 1000	AL Cat may be inco					
	Error Ø		OK Cancel				
	Empty 0	ex-spectrum					
	Distinct 4	Belt Parkway	1				
	Unique 0						
	Empty string 0						
	Min BELT PA						
	Max Belt park.						

Wait, what?! Belt Parkway is the same as Belt parkway, right? Well, in reality – YES! But, in Power Query M language, case sensitivity will make these two as completely different entities! So, we need to conform the values to be able to get correct results in our reports:

Motor_Vehicle_CollisionsCrashes	al cocation		Copy	A . OFF STREET NAME	123 NUMBER OF PERSONS INJUR	PERSONS KILLED + 123 NUMBE	 PROPERTIES
	0% * Error (Valid Valid Error Error Error Error Error Error Error Error Error	Remove Remove Other Columns Duplicate Column Add Column From Examples Remove Duplicates	tor DN	Valid 102% Valid Free ON Free Copy One Free One One Free One Free	100% • Yalid 0% = Error 5% = Erroty	Name Motor Vence Colscon, Crathes Al Properties • APPLED STEPS Sound Promoted Headers
	1 0 2 -72.57627 (40.64334, -72.97617) 4 0 5 0 (40.05) 6 0 (40.05) 7 0 (40.05) 8 0 (40.05) 9 0 (40.05) 11 0 (40.05) 12 0 (40.05) 13 0 (40.6754, -73.96312) 15 0 (40.6754, -73.96312)	BRONK WHITESTONE BROG HUTCHINSON BIVER PARKW VANDERVORT AVENUE EDSON AVENUE VERACIAND BRIDGE UPPER 33 st	Remove Energ Change Syse Tonofform Replace Volues Replace form Split Column Group By Fill Unpixet Columns Unpixet Columns Unpixet Columns Unpixet Columns Remunta Move	ATLANTE AVENUE Inverse UPPERCASE Copitalie Each mark Tom Coopitalie Each mark JSON JANE		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Changel Type Changel Type Filtered Rows Replaced Value > Replaced Value1
	16 -73.836 (40.843464, -73.836) 17 -73.990974 (40.692547, 73.990974) 18 C	COURT STREET	Drill Down Add as New Query	MIDDLETOWN ROAD	0 2	° ~ ~	
	Column statistics ···· Count 1000 Error 0 Emply 0 Distinct 495 Unique 382 Emply string 243 Min peromyte	Value distribution BELT Net ROTO LONG GLAND ENRISES BROOM ATACHT, AN MAJOR DESAN ENRISE BROOM BOULT BROOK INGER BOULT	RIVE			Î	

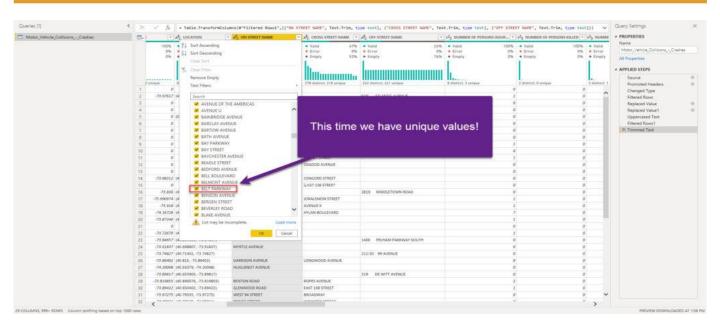
As you can see, I will apply Uppercase transformation to all the columns containing street names, and now we should be good to go:



Why do we have two exactly the same uppercased values for BELT PARKWAY? Well, the original CSV file sometimes can contain hidden characters, such as tabulator, new line, or space. Don't worry, I have good news for you: Power Query enables you to solve this specific issue with one click!

Motor_Vehicle_CollisionsCrashes	at at tocation	ATC ON STREET NAME	apy .	OF STREET NAME	123 NUMBER OF PERSONS INJUR	OF PERSONS KILLED - 123 NUMBE	
	0% # Error 0%	• Valid • Error • Erroty Erroty Re	move Columns move Other Columns Id Column From Examples move Duplicates	ever 0%	6 Valid 102% Valid 6 Error 0% Error 6 Empty 0% Errory	100% • Valid O% • Error S% • Erropy	Name Motor_Venicle_Collisions,Crashes All Properties • APPLIED STEPS
		Real Providence in the International Providence in the Interna	move Errors		ll.		Source
	P unique B48 distinct, 813 unique	BRONX WHITESTONE BRIDG	place Values	brict, 8 unique	8 distinct, 1 unique 2 distinct, 8 un	tique 3 distinct, 0	Promoted Headers
	2 -78.97617 (40.68358, -73.97617)	Fil	E	* ATLANTIC AVENUE		· ·	Changed Type
	3 0	HUTCHINSON RIVER PARKA	unge Type			0	Filtered Rows Replaced Value
	4 0	- 1045-55-56-66-55-56-56-56-56-56-56-56-56-56	iniform	lowercase			Replaced Value1
	5 0 (0.0, 0.0)	Encode and		UPPERCASE	0	0	Uppercased Text
	6 0	VERRAZANO BRIDGE UPPER	erge Columns	8831400 States	0	0	> Filtered Rows1
	7 0		eup By	Capitalize Each Word	0	8	
	8 0		ipivot Columns	Trim	0	0	
	0 0		pivot Other Columns	Clean	1		
	10 0		spivot Only Selected Columns	Length	0	0	
	11 0		ove	JSON		0	
	12 0	GREENFIELD AVENUE	UNSUUD AVENUE	XML	0	0	
	13 0	BROOKLYN BRIDGE			0	0	
	1472.98322 (40.69754; -73.98312)	GOLD STREET	CONCORD STREET		0	0	
	15 0	BRUCKNER BLVD	LENST 138 STREET		0	0	
	16 -73.836 (40.843464, -73.836)		Transfer and	2819 MIDDLETOWN ROAD	0	0	
	17 -73.990974 (40.692547, -73.990974)	COURT STREET	JORALEMON STREET		1	0	
	18 -75 918 (40.626457, -75.918)	ILAUPH AVENUE	AVENUE K		1	0	
	19 -74.26728 (40.526894, -74.16728)	BARCLAY AVENUE	HYLAN BOULEVARD		7	0	
	20 -73.87248 (40.840775, -73.87246)	BRONK RIVER PARKWAY			1	0	
	21 0	MAJOR DEEGAN EXPRESSWAY RAMP			0	0	
	22 -73.72679 (40.694035, -73.72679)	CROSS ISLAND PARKWAY			1	Ø.	
	23 -75.84657 (40.857365, -73.84657)			1400 PELHAM PARKWAY SOUTH	0	0	
	24 -75.92837 (40.698807, -75.92837)	MHETLE AVENUE			0	0	
	25 -73.74827 (40.71402, -73.74827)			211-20 99 AVENUE	0	ø	
	26 -73.89402 (40.815, -73.89402)	GARRISON AVENUE	LONGWOOD AVENUE		0	0	
	27 -74.20098 (40.55079, -74.20098)	HUGUENOT AVENUE			0	0	
	28 -75.89827 (40.655903, -73.89817)			319 DE WITT AVENUE	0	0	
	29 -73.819855 (40.890076, -73.819855)	BOSTON ROAD	ROPES AVENUE		2	0	
	30 -73.69422 (40.650402, -73.89422)	GLENWOOD ROAD	EAST LOB STREET		1	0	
	31 -75.87275 (40.79335, -73.97275)	WEST 94 STREET	BRCADWAY		0	o 🗸	
	12	frankr.emera.	HITCH THE OWNER			· · ·	

This time we used Trim transformation to remove or leading and trailing blank characters. And, let's check again if that resolved our issue with duplicate values:

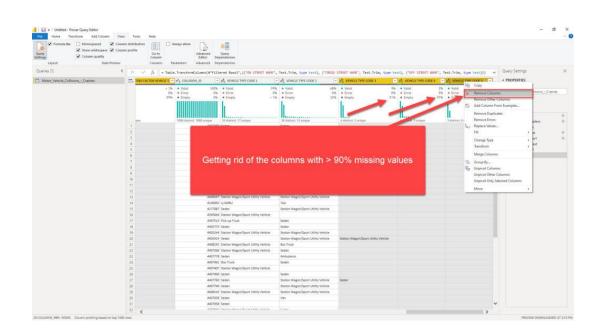


Finally, our column looks as expected: we have unique values!

Thinking forward

Now, you can be tempted again to hit that *Close & Apply* button and start building nice visualizations in Power BI. But, please be patient, as we need to do put some additional effort before closing Power Query Editor.

First consideration – do we need all 29 columns for our analytic solution? I'll put my money that we don't. So, let's follow <u>best practices regarding data model optimization</u>, and get rid of the unnecessary data. There are 6 columns with more than 90% empty values (thanks again Power Query Editor for enabling me to spot this in literally a few seconds) – so, why on Earth should we bloat our data model with these columns when they can't provide any useful insight?!



Now it looks much better! Before we proceed to the next stage of our process and start building an efficient data model, there is one more thing that should be done, to stay aligned with the <u>best practices when working with Power Query Editor</u>.

I will rename each transformation step, so that if someone (or even I) opens this file in a few months, I know exactly which step performs which transformation! I mean, it's easy when you have just a few transformation steps (even though you should follow the recommendation to rename them in that case too), but once you find yourself within tens of transformation steps, things quickly become more cumbersome...Instead of walking through each of the steps trying to understand what each of them does, you will be able to easily catch the logic:

Motor_Vehicle_CollisionsCrashes	- m. (+)	1.2 70P CODE -	L2 IATITUDE	L2 LONGITUDE	OCATION	- A- ON STREET NAME	AR CROSS STREET NAME	A ⁸ r OFF STREET NAME	- 122 M	* PROPERTIES
	100	• Valid 100%	Valid 100%	• Valid 100% • Va	14	6 • Valid 76%	Valid 47%	• Valid 249	Va	Name
		# Error D%	· Error 0%	• Error 0% • Er	rtor D	6 Error D%	• Error 0%		N . Er	Motor_Vehicle_CollisionsCrasher
	2%	 Empty 0% 	 Empty 0% 	 Emply Emply Emply Emply 	mpty Pt	6 Empty 24%	 Empty 53% 	• Erripity 749	h • En	All Properties
										✓ APPLIED STEPS
						Hilling	III III III III III III III III III II		E la la	Provide and an and a second se
			Bhi district. 862 unique		fatinet bill unique	Adia distinct. 357 unique	376 distinct, 318 unique	263 distinct 261 unique	R dist	Source
	4	Distantine and and a	D D	C.	report and model	BRONX WHITESTONE BRIDGE	are merred, are writed	they managed and musicing	1 1 1 1 1	Promoted Headers
	2	11217	40.68358	-75 97617 (40.68	8358 -73.976171			620 ATLANTIC AVENUE	^	Changed Type Replaced Blank-> N/A
	3	0	0	0		HUTCHINSON RIVER FARKWAY			100	Replaced null -> 0
	4	11222	p	0		VANDERVORT AVENUE	ANTHONY STREET		-	Uppercased street names
	5	p	0	0.00.0	0.0x	EDSON AVENUE				X Immed street names
	6	0	0	0		VERRAZANO ERIOGE UPPER				
	7	11106	Ø	0	0	and the second			_	L
	8	0	0	0	10000			and the second		
	9	0	0	0	Sper	id some time for	the transfor	mation		
	10	0	0	0		proper naming				
	11	11236		0	steps	proper naming.	. I our iuture	Sen win		
	1.11	11008								
	11	10304	0	0		be gra	teful!			
			0	0 0		be gra	teful!			
	12		0 0 40 59754		8754	be gra	teful!			
	12	10304 0	0 0 40.65754 0	0	8754	be gra	teful!			
	12 13 14 15 16	10304 0	0 40.843464	0 -73.98312 (40.65 0	8754 43464, -73,836)		LANST 138 STREET	2819 MIDDLETOWN ROAD		
	12 13 14 15	10304 0 11301 0	0	0 -73.98312 (40.65 0	43464, -73,836)			1819 MIDDLETOWN ROAD		
	12 13 14 15 16 17 18	10304 0 11203 0 10461 0 11234	0 40,843464 40,692547 40,625457	0 73.98512 (40.60 72.836 (40.94 72.936)74 (40.95 77.918 (40.65	45464, -73,856) 91547, -73,990974) 26457, -75,918)	BRUCONER BLVD COURT STREET BALTH AVENUE	LEXIST EINE STREET JORIALEMON STREET AVENUE K	3819 MIDOLETOWN BOAD		
	12 13 14 15 16 17 18 19	9 1000 11201 0 1001 9	0 40,843464 40,692547 40,626457 40,526494	0 -73.9833.3 (40.69 0 -72.935 (40.94 -72.930974 (40.69	45464, -73,856) 91547, -73,990974) 26457, -75,918)	BRUCKNER BLVD COLURT STREET	LUNST EINE STREET	2819 MIDDLETOWN ROAD		
	12 13 14 15 36 17 18 19 20	10304 0 11203 0 10461 0 11234	0 40.843464 40.692547 40.825457	0 -73.88532 (40.83 -73.855 (40.94 -73.856 (40.94 -73.996074 (40.95 -73.8724 (40.95 -74.55728 (40.94	43464, -73,836) 91547, -73,990974) 26457, -73,9181 26694, -74,16728)	BRUDONER BLVD ODURT STREET RAUTH AVERUE BARCLAV AVERUE BRONK BV/ER PARKWAY	LEXIST EINE STREET JORIALEMON STREET AVENUE K	2819 MIDDLETOWN ROAD		
	12 13 14 15 16 17 18 19 20 21	10304 0 11203 0 10461 0 11234	0 40.843464 40.562547 40.826457 40.526894 40.566775 0	0 -73,8852 (40,89 -72,885 (40,94 -72,8950 (40,94 -73,982 (40,94 -74,8729 (40,93 -75,8724 (40,94 -75,8724 (40,94	43464, -73, 836) 83547, -73, 993974) 26457, -73, 9381 26694, -74, 16728) 40775, -73, 87246)	BRUCKNER BLVD COURT STREET BARLYF AVENUE BARCKAF AVENUE BRONK BYJR FARKKAY MAUOR DEEGAN DOPRESSWAY RAMP	LEXIST EINE STREET JORIALEMON STREET AVENUE K	2859 MIDDLETOWN ROAD		
	12 13 14 15 16 17 18 19 20 21 22	10804 0 17305 0 1005 0 12254 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 40,843454 40,642547 40,55657 40,55654 40,55654 60,84075 0 40,64035	0 75 (882) (40, 89 75 (882) (40, 98 75 (896) (40, 98 74 (897)) (40, 99 74 (897)) (40, 99 75 (872) (40, 99 75 (726)) (40, 99	43464, -73,836) 82547, -73,999974) 26457, -73,918) 26894, -74,16728) 40775, -73,87246) 94035, -73,72679)	BRUDONER BLVD ODURT STREET RAUTH AVERUE BARCLAV AVERUE BRONK BV/ER PARKWAY	LEXIST EINE STREET JORIALEMON STREET AVENUE K			
	12 13 14 15 16 17 18 19 20 21 21 22 23	20004 0 32005 0 9 31205 20052 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 40,845454 40,8925477 40,825457 40,825457 40,840775 0 40,840775 40,840775 40,840775 40,85555	0 -73,8852 (40,89 -77,886 (40,84 -72,89674 (40,64 -72,87,84 (40,64 -73,87,84 (40,64 -73,87,246 (40,64 -73,72,65 (40,64 -73,72,65 (40,64) -73,72,65 (40,64) -73,72,65 (40,64)	43464, -73,836) 92547, -73,9399974j 26457, -75,938) 26894, -74,15728) 40775, -73,87246) 94035, -75,72679j 57365, -73,84657)	BRUCKIER BLVD COLAIT STREET BARCUM AVENUE BARCUM AVENUE BARCUM SVIES MAXIMAT MAUGE DEGIAN ERMESSIWAT BARE CROSS SILAND PARKINAT	LEXIST EINE STREET JORIALEMON STREET AVENUE K	2859 MIDDLETOWN ROAD		
	12 13 14 15 56 17 18 19 20 20 21 22 23 23 24	20004 0 21201 0 21204 21204 0 21204 0 0 20022 0 0 0 20422 0 0 0 0 0 0 0	0 40.845464 40.825457 40.826457 40.826457 40.826457 0 40.846055 40.85055 40.658057	73 8832 (40 85 73 8832 (40 85 77 885 (40 84 77 885 (40 84 78 899 (40 84 78 87 85 78 87 85 78 72 87 78 72 87 74 865 75 72 87 74 865 75 87 87 75 88 9 75 88 9 76 78 9 76	45464, -73, 836) 52547, -73, 8393974) 26457, -75, 9381 16994, -74, 16728) 40775, -73, 87246) 94035, -73, 872679 57565, -73, 918657) 84807, -73, 91837)	BRUCKNER BLVD COURT STREET BARLYF AVENUE BARCKAF AVENUE BRONK BYJE FARKKAY MAUCH DEEGAN EDPRESSWAY RAMP	LEXIST EINE STREET JORIALEMON STREET AVENUE K	1400 PELHAM PARKWAY SOUTH		
	12 13 14 15 15 16 17 18 19 20 21 22 23 24 23 24 25	20084 0 31200 0 300451 300451 300451 0 0 0 0 0 200451 0 0 0 0 0 0 0 0 0 10409 0 10409 0 10409	0 40,84544 40,845457 40,8526457 40,825689 40,85457 9 40,854035 40,857455 40,857455 40,875455 40,0584035	0 -73,8952 (40,89 -73,895 (40,84 -74,895 (40,84 -74,895 (40,84 -74,895 (40,85 -74,847 (40,85 -73,8487 (40,85 -75,8487 (40,85 -73,8487 (40,85) -73,8487 (40,85)	45464, -73, 836) 32547, -73, 939374) 26457, -73, 9381 8094, -74, 16728) 40775, -73, 87246) 94035, -75, 75679) 37365, -73, 8657) 84007, -75, 93837) 1402, -73, 74827)	BRUCKNER SLUD COLINT STREET BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA DER SARANGAN CROSS SLUNG FARKWARF ANTRELE AVENUE	UNST EIN STREET JORALEMON STREET ANDRUK K HYLAN BOULEVARD			
	12 13 14 15 16 17 18 19 20 21 22 23 24 24 25 25	20084 0 21383 0 0 21385 12384 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 40 84346 40 83547 40 636547 40 53695 40 586775 0 40 686775 40 686775 40 68675 40 68675 40 68675 40 68675 40 68655 40 68655 40 68655 40 68655 40 6855	0 173 586722 (482,587 173 5857 (683,697 173 5857 (683,697 173 5872 (683,597 173 5872 (683,597 173 5872 (683,597 173 5872 (683,597 173 58,5877 (483,897 173 58,5877 (483,897 173 58,5877 (483,897 173 58,5877 (483,897 173 58,5877 (483,897 173 58,5877 (483,897 173 58,5877 (483,897) 173 58,5877 (483,897) 173 58,5877 (483,897) 173 58,597 (483,997) 173 57,597 (483,997) 17	45464, -73, 856) 92547, -73, 890974) 94697, -75, 9181 26094, -74, 16728) 94035, -75, 726739 94035, -75, 726739 94035, -75, 73, 84837) 94807, -73, 94827) 15, -75, 89402)	BRUCHER BLVD COURT STREET BAUM ANDNE BROCK AVENUE BROCK AVENUE BROCK STUB ARXIVAT ANNOP DEBAU PRESSINAT AMAP CROSS SLAND PARVIARY AVENTLE AVENUE GARRICO AVENUE	LEXIST EINE STREET JORIALEMON STREET AVENUE K	1400 PELHAM PARKWAY SOUTH		
	12 13 14 15 15 16 17 18 19 20 21 22 23 23 24 23 23 24 25 25 25 25 25 25 25 25 25 25 25 25 25	20085 0 31205 0 30651 20051 20051 20051 0 0 20051000 20051 200500 200510	0 0 40,84344 40,85547 40,55587 40,55587 40,55587 40,85585 40,85575 40,98807 40,75507 40,75507 40,7555 40,98807 40,7555 40,7557 40,7557	0 73,8823, (40,80 72,960)74, (40,60 72,960)74, (40,60 73,9738), (40,70 74,9738), (40,70 74,9738), (40,70 74,9738), (40,70 73,9738), (40,70 73,9738), (40,70 73,9738), (40,70 73,9738), (40,70 73,9738), (40,70 73,9748), (40,70 74,9748), (40,70), (4	4546473.836) 82547.73.83080740 82547.73.83080740 26894.74.57.281.81 26894.74.57280 4007573.872460 94005.73.726799 57365.73.846377 140273.748277 140273.748279 5079.74220980	BRUCKNER SLUD COLINT STREET BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA VINTING BANCHA DER SARANGAN CROSS SLUNG FARKWARF ANTRELE AVENUE	UNST EIN STREET JORALEMON STREET ANDRUK K HYLAN BOULEVARD	3400 PEJHAM PARKWAY SOUTH 213-20 99 AVENUE		
	12 13 14 15 15 16 17 18 19 20 21 22 23 24 25 25 26 27 27 28	2008 0 27285 0 0 2005 2005 0 0 2005 0 2005 0 21459 0 21459 0 21459 0 21459 0 21459 0 21459 0 21459 0 21451 0 21451 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 46 (26 44 64 40 65 05 47 40 55 95 7 40 55 96 7	9 73,9452,240,241 73,955 (60,000,400,00 77,9532 (60,00 74,9572 (60,00 74,97242 (74,952) 74,9724 (74,950) 74,9724 (74,950)	45464, -73,856) 45547, -73,8550 4557, -75,8181 26894, -74,87281 40775, -73,87246) 40025, -78,726795 5766, -72,846575 1402, -73,74827 1402, -73,74827 1402, -73,74827 1402, -73,74827 1402, -75,74827 1402, -75,74827 1	BRUDNER BUD COURT STREET RAILHY AVDNE BRICHY AVDNE BRICHY AVDNE BRICHY RAVENE BRICHY RAVENE ROOS SUAND PARYNEY CROSS SUAND PARYNEY AVIRTLE AVENUE SUARDSON AVENUE HUBUENOT AVENUE	LENGT LIK STREET JORALEMON STREET ANENUE K HITAR BOULEVARD	1400 PELHAM PARKWAY SOUTH		
	12 13 14 15 15 16 17 18 19 20 21 22 23 24 25 25 25 25 25 25 25 25 25 25 25 25 25	2008 0 31205 0 3063 2022 2022 2022 0 0 0 2025 0 0 2045 0 0 21459 0 21459 0 21459 0 21459 0 21459 0 21459 0 21205 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 40,843444 40,82547 40,52547 40,5254894 40,852589 40,555895 40,5559540,55595 40,55595 40,5559540,55595 40,55595 40,5559540,55595 40,55595 40,5559540,55595 40,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,55595 40,5559540,55595 40,5559540,55595 40,55595 40,5559540,55595 40,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540,55595 40,5559540 40,55595 40,55595400 40,55	0 73,8532,648,85 74,854,648,64 74,854,648,74 74,854,648,74 74,87246,633 74,87246,634 73,87246,634 73,8427,663 73,8427,663 73,8427,663 73,8402,663 73,8402,663 73,8402,663 73,8402,663 73,8402,663 73,8402,663 73,8402,663 73,8402,663 74,9402,663 74,9	4546473.836) 4254773.809074) 4254773.809074) 4264773.9381 4669474.87780 4603574.72679 5756573.846377 4860773.9184577 1460275.9184577 1460275.9184577 1460275.9184577 95078,-72.8184577 96076,-73.81984551	BRUDDRER SLUD ODURT STREET AAUTH AVIDINE BARCLAY AVIDINE BARCHAY AVIDINE BARCHAY AVIDINE BARCHAY AVIDINE BARCHAY AVIDINE CROSS SULAND FARCHARY AVITEL AVIDUE GARROOK AVIDINE BARTON BOAD	LINGT 238 STREET JORALDONG STREET ANDYOR K HILAN BOULEVARD LONGWOOD AVENUE ROPES AVENUE	3400 PEJHAM PARKWAR SOUTH 213-20 99 AVENUE		
	12 13 14 15 15 16 17 18 19 20 21 22 23 24 25 25 26 27 27 28	2008 0 27285 0 0 2005 2005 0 0 2005 0 2005 0 21459 0 21459 0 21459 0 21459 0 21459 0 21459 0 21459 0 21451 0 21451 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 40,843464 40,653547 40,553547 40,553659 40,553659 40,557365 40,557365 40,557365 40,6557365 40,657365 40,557365 40,555901	9 73,9452,240,241 73,955 (60,000,400,00 77,9532 (60,00 74,9572 (60,00 74,97242 (74,952) 74,9724 (74,950) 74,9724 (74,950)	4546473.836) 4545473.8399374) 2645773.939374) 2645773.939311 2669474.40728) 4007573.87246) 9402573.87246) 9402573.848577 9800773.913577 1402,-73.948271 507974.20098) 5950573.848477 5950574.848479 507974.20098) 5950573.848477 507074.20098) 5950574.848479 507074.20098) 5950574.848479 507074.200980 5950574.848479 507074.200980 50707	BRUDNER BUD COURT STREET RAILHY AVDNE BRICHY AVDNE BRICHY AVDNE BRICHY RAVENE BRICHY RAVENE ROOS SUAND PARYNEY CROSS SUAND PARYNEY AVIRTLE AVENUE SUARDSON AVENUE HUBUENOT AVENUE	LENGT LIK STREET JORALEMON STREET ANENUE K HITAR BOULEVARD	3400 PEJHAM PARKWAR SOUTH 213-20 99 AVENUE		

Trust me, your future self will be extremely grateful after a few months:)

Before we conclude the Data Preparation phase, I've intentionally left the best thing for the end:

All of the transformation steps you defined will be saved by Power Query Editor, and every time you refresh your dataset, these steps will be applied to shape your data and will always bring it to the desired form!

Summary

After we emphasized the importance of understanding the business problems that need to be solved by the analytic solution, in this part we got our hands dirty and started to shape our data in order to prepare it to answer various business questions.

During the data preparation process, we performed data profiling and identified different issues that could potentially harm our final solution, such as missing or duplicate values. Using an extremely powerful built-in transformation tool – *Power Query Editor* – we were able to quickly resolve data inconsistencies and set the stage for the next phase – data modeling! Don't forget that Power Query Editor, which is an integral part of Power BI, enables you not just to apply complex transformations using a simple UI, without any coding skills, but also offers you a possibility to enhance your data model significantly by using very powerful M language if needed.

Therefore, when someone tells you that the Power BI is a "visualization tool only", ask her/him to think again about it.

In the next chapter, we'll continue our journey on building an end-to-end analytic solution using Power BI, by focusing on the data modeling phase.

Introduction

After we laid some theoretical background behind the process of building an end-to-end analytic solution and explained why it is of key importance to understand the business problems *BEFORE* building a solution, and applied some basic data profiling and data transformation, it's the right moment to level up our game and spend some time elaborating about the best data model for our analytic solution. As a reminder, we use an open dataset about motor vehicle collisions in NYC, which can be found <u>here</u>.

Data Modeling in a nutshell

When you're building an analytic solution, one of the key prerequisites to create an **EFFICIENT** solution is to have a proper data model in place. I will not go deep into explaining how to build an enterprise data warehouse, the difference between OLTP and OLAP model design, talking about normalization, and so on, as these are extremely broad and important topics that you need to grasp, nevertheless if you are using Power BI or some other tool for development.

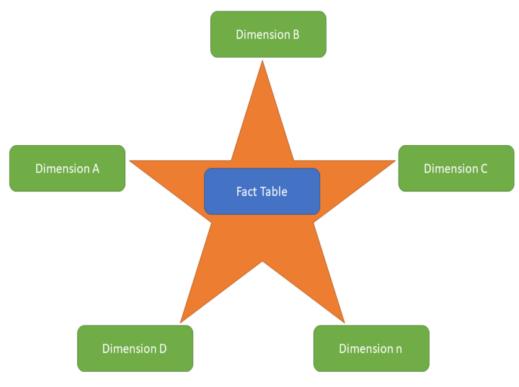
The most common approach for data modeling in analytic solutions is *Dimensional Modeling*. Essentially, this concept assumes that all your tables should be defined as either fact tables or dimension tables. Fact tables store events or observations, such as sales transactions, exchange rates, temperatures, etc. On the other hand, dimension tables are descriptive – they contain data about entities – products, customers, locations, dates...

It's important to keep in mind that this concept is not exclusively related to Power BI – it's a general concept that's being used for decades in various data solutions!

If you're serious about working in the data field (not necessarily Power BI), I strongly recommend reading the book: <u>The Data Warehouse Toolkit: The Definitive Guide to</u> <u>Dimensional Modeling</u> by Ralph Kimball and Margy Ross. This is the so-called "Bible" of dimensional modeling and thoroughly explains the whole process and benefits of using dimensional modeling in building analytic solutions.

Star schema and Power BI – match made in heaven!

Now, things become more and more interesting! There is an ongoing discussion between two confronted sides – is it better to use one single flat table that contains all the data (like we have at the moment in our NYC collisions dataset), or does it make more sense to normalize this "fat" table and create a dimensional model, known as Star schema?



Building an End-To-End Analytic Solution with Power BI

In the illustration above, you can see a typical example of dimensional modeling, called Starschema. I guess I don't need to explain to you why it is called like that:) You can read more about Star schema relevance in Power BI <u>here</u>. There was an interesting discussion whether the Star schema is a more efficient solution than having one single table in your data model – the main argument of the Star schema opponents was the performance – in their opinion, Power BI should work faster if there were no joins, relationships, etc.

And, then, *Amir Netz*, CTO of Microsoft Analytics and one of the people responsible for building a <u>VertiPaq engine</u>, cleared all the uncertainties on Twitter:



If you don't believe a man who perfectly knows how things work under the hood, there are also some additional fantastic explanations by proven experts why star schema should be your preferred way of modeling data in Power BI, such as <u>this video from Patrick (Guy in a Cube</u>), or this one from <u>Alberto Ferrari (SQL BI)</u>.

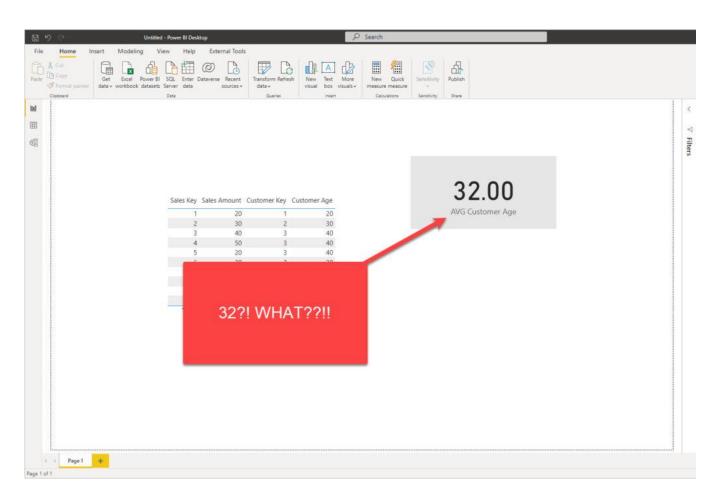
And, it's not just about efficiency – it's also about getting accurate results in your reports! In this <u>article</u>, Alberto shows how writing DAX calculations over one single flat table can lead to unexpected (or it's maybe better to say inaccurate) results.

Without going any deeper into explaining why you should use Star schema, let me just show you how using one single flat table can produce incorrect figures, even for some trivial calculations!

IJ	すべ		Untitled - Power BI Desk	top		<u>م</u> آ	Search	
File	Home	Insert Modelin	g View Help	External Tools				
Paste	X Cut Copy Format paints Clipboard	Get Excel data ~ workbook		Oataverse Recent sources ~	Transform Refre data ~ Queries		New Quick measure measure Calculations	Sensitivity
000								
6777								
III								
릡								
			Sales Key 1 2 3 4 5 6 7 8	Sales Amount 20 30 40 50 20 30 40 50	Customer Key 1 2 3 3 3 2 2 2 1	Customer Age 20 30 40 40 40 30 30 20		
			9	20	2	30		
			10	30	3	40		
			Total	330				

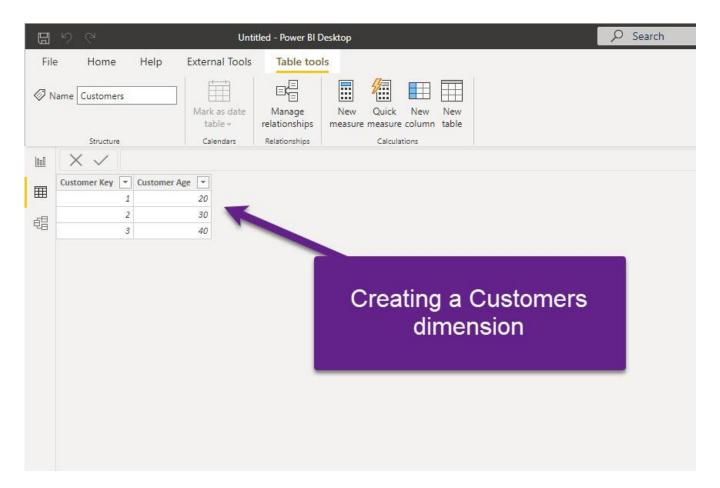
This is my flat table that contains some dummy data about Sales. And, let's say that the business request is to find out the average age of the customers. What would you say if someone asks you what is the average customers' age? 30, right? We have a customer 20, 30, and 40 years old – so 30 is the average, right? Let's see what Power BI says...

AVG Customer Age = AVERAGE(Table1[Customer Age])



How the hell is this possible?! 32, really?! Let's see how we got this unexpected (incorrect) number...If we sum all the Customer Age values, we will get 320...320 divided by 10 (that's the number of sales), and voila! There you go, that's your 32 average customers' age!

Now, I'll start building a dimensional model and take customers' data into a separate dimension table, removing duplicates and keeping unique values in the Customers dimension:



I've also removed Customer Age from the original Sales table and established the relationship between these two on the Customer Key column:

8 90	Untitled - Power BI Desktop		D Search
File Hom	e Help External Tools		
Paste Cut Clipboard	Get Excel Power BI SQL Enter Dataverse Recent data - workbook datasets Server data		A CR Q&A Language Linguistic setup - schema - Security Q&A Share
1 <u>000</u>			
=			
電	E Customers		
	Customer Age Customer Key Collapse ^	Table1 Customer Key ∑ Sales Amount Sales Key Collapse ∧	
< • A	I tables		

Finally, I just need to rewrite my measure to refer to a newly created dimension table:

AVG Customer Age = AVERAGE(Customers[Customer Age])

And, now, if I take another look at my numbers, this time I can confirm that I'm returning the correct result:

89	r çi	Untitled - Power BI Desktop			Ą	Search			
File	Home Insert Mod	deling View Help Exte	nal Tools						
Passe \$	Get Exce	el Power BI SQL Enter Dataverse pook datasets Server data	Recent Trans		ew Text More sual box visuals v insert	New Quick measure measure Calculations	Sensitivity Sensitivity Sensitivity Sensitivity Share		
Ini									
=									
43									
148									
							20.00		
		Sales Key Sales A	mount Curto	mar Kau Custo	mer Ace		30.00	J	
		Jales Key Jales P	20	1	20		AVG Customer A		
		2	30	2	30			2-	
		3	40	3	40		4		
		4	50	3	40				
		5	20 30	3	40				
		7	40	2	30				
		8	50	1	20				
		9	20	2	30				
		10	30	3	40				
		Total	330						
					The second second second			A 44 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
					This	time we	get correct num	nber	
	L								
i c	> Page 1 +								
Page 1 of 1									

Of course, there is a way to write more complex DAX and retrieve the correct result even with a single flat table. But, why doing it in the first place? I believe we can agree that the most intuitive way would be to write a measure like I did, and return a proper figure with a simple DAX statement.

So, it's not only about efficiency, it's also about accuracy! Therefore, the key takeaway here is: *model your data into a Star schema whenever possible!*



Building Star schema for NYC collisions dataset

Of course, there is a way to write more complex DAX and retrieve the correct result even with a single flat table. But, why doing it in the first place? I believe we can agree that the most intuitive way would be to write a measure like I did, and return a proper figure with a simple DAX statement.

As we concluded that the Star schema is the way to go, let's start building the optimal data model for our dataset. The first step is to get rid of the columns with >90% missing values, as we can't extract any insight from them. I've removed 9 columns and now I have 20 remaining.

At first glance, I have 5 potential dimension tables to create:

- ✓ Date dimension
- ✓ Time dimension
- ✓ Location dimension (Borough + ZIP Code)
- ✓ Contributing Factor dimension
- ✓ Vehicle Type dimension

But, before we proceed to create them, I want to apply one additional transformation to my Crash Time column. As we don't need to analyze data on a minute level (hour level of granularity is the requirement), I'll round the values to a starting hour:

Motor Vehicle Collisions - Crashes	CRASH D		TIME	A second		2IP CODE - 1.	2 LATITUDE	- 12 LOWGITUDE	- AR ON STREET NAME	- AP CROSS STREET NAM	· · PROPER	THES
	1	14-Apr-21	fa.	Cepy	1	0		0	Ø BRONX WHITESTONE BRIDGE		Name	
	2	13-Apr-22	×	Remove		11217	-40	66358	-73.97617		∧ Motor,	Vehicle, Collisions, -, Crash
	3	15-Apr-21		Remove Other Columns	1	0		0	0 HUTCHINSON RIVER PARKWAY		All Prop	erties
	4	13-Apr-21		Duplicate Column		11222		0	Ø VANDERVORT AVENUE	ANTHONY STREET		
	5	12-Apr-21	53	Add Column From Examples		6		0	0' EDSON AVENUE		 APPLIE 	D STEPS
	6	13-Apr-21		Remove Duplicates		0		0	Ø VERRAZANO BRIDGE UPPER		Sou	rce
	7	13-Apr-21		Remove Errors		11106		0	0 33.5T	31AVE	Prov	moted Headers
	8	16-Apr-21						0	D SHORE PARKWAY			inged Type
	9	11-Apr-21		Change Type	1			0	0 GOWANUS RAMP		1000	laced Blank-> N/A
	10	16-Apr-21	- C	Transform		Hour F	Hour	þ	Ø BRONK RIVER PARKWAY RAMP		1000	faced null -> 0
	11	11-Apr-21	£.,	Replace Values		Minute	Start of Hour		Ø BEVERLEY ADAD	EAST 21 STREET	Uppercased street ru	
	12	15-Apr-22		Replace Errors		Second	End of Hour p		0 GREENFIELD AVENUE	OSGOOD AVENUE	Trimmed street nav Removed Errors	
	13	10-Apr-21				Text Transforms +		0	0 BROOKLYN BRIDGE			
	14	21-May-19	- 10	Group By Fall	ा	11201	-40	69754	-73 S8312 GOLD STREET	CONCORD STREET	2.840	noved Columns
	15	21-301-20			- 11	0		0	0 BRUCKNER BLVD	WINST 138 STREET		
	16	25-Fep-21	50	Unpivot Columns Unpivot Other Columns	- 1	10461	40.5	43464	-73.836			
	37	09-Mer-21			- 1	0	40.6	92547	-73 980924 COURT STREET	JORALEMON STREET		
	18	31-Mor-21		Unpivot Only Selected Columns	1	21234	40.4	26457	-73 PIB RALPH AVENUE	AVENUE 6		
	19	06-Apr-21	a)	Rename		10512		26894	-74 16728 BARCLAY AVENUE	HVLAN BOLLEVARD		
	20	09-Apr-21		Move	- 61	0	40.8	40775	-73.87246 BRONK RIVER PARKWAY			
	21	14-Apr-21		Dnill Down		0		0	Ø MAJOR DEEGAN EXPRESSWAT			
	22	14-Apr-21		Add as New Query		0	40.5	94035	-73 72679 CROSS ISLAND PARKWAY			
	23	15-Apr-21	-	ADD AS NEW COLEY	_	10461	40.8	57345	-73.84657			
	24	14-Apr-21	25	40:00 PM N/A		0	40.6	96807	73 92827 MYRTLE AVENUE			
	25	14 Apr 21		43:00 PM QUEENS		11429		71402	-73.74827			
	26	16-Apr-21	5	AD-00 PM BRONX		20474		40.825	-75 89402 GARRISON AVENUE	LONGWOOD AVENUE		
	27	16-Apr-21	12.5	0.00 AM N/A		0	40	\$5079	-74 20098 HUGUENOT AVENUE			
	28	15-Apr-21	10	REDERAM BROCKLYN		11207		55903	-73.89617			
	29	16-Apr-21	4	RS OD PM BRONX		20475		90076	-75.829855 BOSTON ROAD	ROPES AVENUE		
	30	15-Apr-22	5.	20:00 PM BROOKLYN		31236		50402	-73.89422 GLENWOOD ROAD	EAST 108 STREET		
	31	16-Apr-21	10.	20:00 PM MANHATTAN		10025		79335	-73 97275 WEST 94 STREET	BROADWAY		
	32	16-Apr-21	5.	20:00 PM MANHATTAN		10012	40	7253#	-74.00011 PRINCE STREET	WOOSTER STREET		
	33	15-Apr-21	2	NO ROPANT QUEENS		21277	40	75184	-73 90358 BROADWAY	58 STREET		
	34	16-Apr-21	11	AD RO PM BROOKLYN		21226	40.6	49788	-73 9622 EAST 19 STREET	CHURCH AVENUE		
	35	16-Apr-21	6.	15:00 PM BROOKLYN		31223	40.4	86928	-73.920825			
	36	14-Apr-21	11	AD-OD PM BROCKLYN		11211	40.7	12963	-73 93647			
	37	14-Apr-21	8:	SEOD PM N/A		0	40.8	01285	-73.95394 7 AVENUE		~	
	38	14-Apr-21	3	LE OD PM BROOKLYN		11220	40.6	33976	-74.02211		- A	

I'll now duplicate my original flat table 4 times (for each of the dimensions needed, except for the Date dimension, as I want to use a more sophisticated set of attributes, such as day of the week for example). Don't worry, as we will keep only relevant columns in each of our dimensions and simply remove all the others. So, here is an example how the Location dimension looks like:

-		olumns(#"Removed Errors",("BOROUGH", "21P CODE		+ PROPERTIES
Motor_Vehicle_CollisionsCrashes Location				- PROPERTIES
	1 N/A	Ø		 Location
Contributing Factor	2 BROOKLYN	11217		All Properties
Vehicle Type	3 16/6	0		All Properties
Time	4 BROOKLYN	11222		 APPLIED STEPS
	5 N/A 6 N/A	0		Source
		11106		Promoted Headers
	7 GUEENS 8 N/A			Changed Type
		0		Replaced Blank-> N/A
	9 14/6	0		Replaced null -> 0
	10 N/A	0		Removed Errors
	11 BROOKLYN 12 STATEN GLAND	11226		X Removed Other Columns
	12 STATEN OLAND 13 N/A	0		
	14 BROOKLYN	11201		
	15 N/A	0		
	16 BRONK	10461		
	17 N/A	0		
	18 BROOKIN	11234		
	10 STATEN ISLAND	10312		
	20 N/A	0		
	21 N/A	0	Location dimension	
	22 N/A	0	Location dimension	
	23 BRONK	10451		
	24 N/A	0		
	25 QUEENS	11429		
	26 BRONK	20474		
	27 N/A	0		
	25 BROOKLYN	11207		
	29 BRONX	10475		
	30 BROOKLYN	11236		
	31 MANHATTAN	10025		
	32 MANNATTAN	30022		
	33 QUEENS	11377		
	34 BROOKLYN	11226		
	35 BROOKLYN	11221		
	35 BROOKLYN	11222		
	37 N/A	0		
	38 BROOKLYN	11220		~
	10 N/A	0		

The next important step is to make sure that we have unique values in each dimension, so we can establish proper 1-M relationships between dimension and fact table. I will now select all my dimension columns and remove duplicates:

Queries (5)	< 🔀 🗹 fz - Table	SelectColumns(#"Removed Errors",("BOROUSH", "21P (CODE"))	 Query Settings
Motor_Vehicle_CollisionsCrashes	T. A. BOROUSH	12 0PM (12 PM	+ PROPERTIES
Location	1 N/A		Name
Contributing Factor	2 BROOKLYN	Kempue Columns	▲ Location
Vehicle Type	3 N/A	Remove Other Columns	All Properties
	4 BROOKLYN	🔁 Add Column From Examples	
III Time	5 N/A	Remove Duplicates	✓ APPLIED STEPS
	6 N/A	Remove Errors	Source
	7 QUEENS	Ly Replace Values	Promoted Headers
	S N/A	El Contractor	Changed Type
	9 N/A		Replaced Blank -> N/A
	10 N/A	Change Type >	Replaced null -> 0
	11 BROOKLYN	Merge Columns	Removed Errors
	12 STATEN GLAND	S. Group By	Removed Other Columns
	13 N/A	Unpivot Columns	
	14 BROOKLYN	Unpivot Only Selected Columns	
	15 N/A		
	16 BRONK	Move +	
	37 N/A	0	
	18 BROOKLYN	11254	
	10 STATEN ISLAND	10312	
	20 N/A	0	
	21 N/A	0	
	22 N/A	0	
	23 BRONK	10461	
	24 N/A	0	
	25 QUEENS	11429	
	26 BRONK	20474	
	27 N/A	0	
	25 BROOKLYN	11207	
	29 BRONX	10475	
	30 BROOKLYN	11236	
	31 MANHATTAN	10025	
	32 MANNATTAN	20022	
	33 QUEENS	11377	
	34 BROOKLYN	11226	
	35 BROOKLYN	11221	
	35 BROOKLYN	11211	
	37 N/A	0	
	38 BROOKLYN	11220	ý l

We need to do this for every single dimension in our data model! From here, as we don't have "classic" key columns in our original table (like, for example, in the previous case when we were calculating average customers' age and we had Customer Key column in the original flat table), there are two possible ways to proceed: the simpler path assumes establishing relationships on text columns – it's nothing wrong with that "per-se", but it can have implications on the data model size in large models.

Therefore, we will go another way and create a surrogate key column for each of our dimensions. As per definition in dimensional modeling, the surrogate key doesn't hold any business meaning – it's just a simple integer (or bigint) value that increases sequentially and uniquely identifies the row in the table.

Creating a surrogate key in Power Query is quite straightforward using Index column transformation.

Column From Custom Invoke Custom Examples Column Function General From 1		
Queries [5]	< X / fx * Table.Distinct(#"Removed Other Columns")	
(duenes ())		
Motor_Vehicle_CollisionsCrashes	III. A ^B _C CONTRIBUTING FACTOR VEHICLE 1	
Location	1 Following Too Closely	
Contributing Factor	2 Unspecified	
Vehicle Type	3 Pavement Slippery	
Time	4 Driver Inattention/Distraction	
	5 Other Vehicular	
	6 Pessing Too Closely	
	7 Passing or Lane Usage Improper	
	8 Driver Inexperience	
	9 Feilure to Yield Right-of-Way	
	10 Brakes Defective	
	11 Turning Improperly	
	12 Unsafe Speed	
	13 Backing Unsafely	
	14 Reaction to Uninvolved Vehicle	
	15 View Obstructed/Limited	
	16 Steering Failure	
	17 Traffic Control Disregarded	
	18 Drugs (illegal)	
	19 Aggressive Driving/Road Rage	
	20 Feil Asleep	
	21 Pedestrian/Bicyclist/Other Pedestrian Err	
	22 Alcohol Involvement.	
	23 Unsafe Lane Changing	
	24 Pavement Defective	
	25 Other Lighting Defects	
	26 Oversized Vehicle	
	27 Animals Action	
	28 Outside Car Distraction	
	29 Illnes	
	30 Driverless/Runaway Vehicle	
	31 Passenger Distraction	
	32 Tire Failure/Inadequate	
	33	
	34 Lost Consciousness	
	35 Accelerator Defective	
	36 Obstruction/Debris	
	37 Failure to Keep Right	
	38 Giare	

Just one remark here: by default, using an Index column transformation will <u>break query</u> <u>folding</u>. However, as we are dealing with CSV file, which doesn't support query folding at all, we can safely apply Index column transformation.

The next step is to add this integer column to the fact table, and use it as a foreign key to our dimension table, instead of the text value. How can we achieve this? I'll simply merge the Location dimension with my Collisions fact table:

Home Transform Add Column 1	/iew Tools	Нер		1	N			
	1	Properties	🔟 UU 🗮 🛤 11 .	Data Type: Date •		Analytics		
k New Recent Enter Data source		efresh		Solt Group	· MerperQuerres	n		
	rameters* Pr	eview * 🛄 Manage *	Columns * Columns * Rows * Rows * Col	lumin + By 42 Replace Values	Je Merge Queries as New	e Machine Learning		
New Query Data Sources P	arametera	Query	Manage Columns Reduce Rows Sort	Transform	Combine	Al insights		
ies (5)	< x	J & - Table, Tran	sformColumns(#"Removed Columns",{{"CRASH TI	NET. Time StartOfNour, type time)1)				Query Settings
Islaw	-	Contraction of the		* 12 ZIP CODE * 12 LATT		A- ON STREET NAME	AP CROSS STREET NAME	A PROPERTIES
2012/00		And in case of the local division in which the local division in which the local division in which the local division in the local din the local division in the local division in the local din the l	And an	Intel Contraction	And a second	Contraction of the local data and the local data an	WE SHIPS STREET HAVE	Name
cation	1	14-Apr-21	5:00:00 AM N/A	0	0	Ø BRONK WHITESTONE BRIDGE	,	Colsons
intributing Factor	2	13-Apr-22	9-00-00 PM BROOKLYN	11217	40.68558	-73.97617		All Properties
hicle Type	3	15-Apr-22	4:00:00 PM N/A	0	0	0' HUTCHINSON RIVER PARKWAY		wit properties.
ne	4	13-Apr-21	4:00:00 PM BROCKLYN	11222	0	Ø VANDERVORT AVENUE	ANTHONY STREET	→ APPLIED STEP5
	2	12-Apr-23	8.00.00 AM N/A	0	0	d EDSON AVENUE		Source
	6	13-Apr-22	5.00:00 PM N/A	0	0	0 VERRAZANO BRIDGE UPPER	11110	Promoted Headers
	7	13-Apr-22	S-00:00 PM QUEENS	31106	0	0 53 ST	STAVE	Changed Type
	8	16-Apr-71	11:00:00 PM N/A	0	0	Ø SHORE PARKWAY		Replaced Blank-> N/A
	9	11-Apr-21	5:00:00 PM N/A	0	0	Ø GOWANUS RAMP		Replaced null -> 0
	10.	16 Apr 21	9:00:00 PM N/A	0	0	0 BRONX RIVER PARKWAY RAMP	COLUMN COLUMN	Uppercased street names
	11	11-Apr-21	P-00-00 PM BROOKLYN	11226	0	Ø BEVERLEY ROAD	EAST 21 STREET	Trimmed street names
	12	15-Apr-22	8:00:00 PM STATEN ISLAND	20304	8	0 GREENFIELD AVENUE	OSGOCO AVENUE	Removed Errors
	13	10-Apr-21	11:00:00 AM N/A 10:00:00 PM BROOKLYN	0	40.69754	Ø BROCKLYN BRIDGE -73.98912 GOLD STREET		Removed Columns
	14	21-May-19 21-Jan-20	3:00:00 PM BROOKLYN 3:00:00 PM N/A	11201	40.59754	23 96312 GOLD STREET	CONCORD STREET	X Calculated Start of Hour
	15		2.00.00 PM BROW	0	0	0 BRUCKNER BLVD -73 #36	12/551 158 518221	
	16	26-feb-21	2.00:00 PM BROWN	20462	40.843454 40.692547	-73.930 -75.980974 COURT STREET	ADRALEMON STREET	
	17	09-Mor-21					AVENUE K	
	18	31-Mor-21 06-Apr-21	10:00:00 PM BROOKLYN 10:00:00 PM STATEN ISLAND	11214	40.826457 40.526894	-78.918 RALPH AVENUE -74.16728 BARCLAY AVENUE	HYLAN BOLLEVARD	
	19		20000 PM SIATEN ISLAND 20000 PM N/A	10312	40.528894 40.840775	-74.15728 BARLLAT AVENUE -73.87246 BRONX RIVER PARKWAY	HYLAN BOULEVARD	
	20	09-Apr-21 14-Apr-21	2.00.00 PM N/A	0	40.840775	0 MAJOR DEEGAN EXPRESSIVAL R		
	23	14-Apr-21	11:00:00 AM N/A	0	40 594015	-73 72679 CROSS ISLAND PARKWAY		
		15-Apr-21	1.00.00 PM BRONX	39462	40.857365	-73.24657		
	23	19-Apr-21	2:00:00 PM N/A	10402	40.598807	-73.91837 MHRTLE AVENUE		
	24	14 Apr 21	2.00.00 PM N/A 9.00.00 PM QUEENS	11429	40.558807	-73 50857 MERILE AVENUE -73 74827		
	25	16-Apr-21	5.00.00 PM BRONX	30474	40,825	-73 89402 GARRISON AVENUE	LONGWOOD AVENUE	
	20	16-Apr-21	12:00:00 AM N/A	0	40.55079	-74 20098 HUGUENOT AVENUE	NAME OF A COMPANY	
	27	15-Apr-21	10:00:00 AM BROOKLYN	11207	40.555900	-78.89817		
	23	16 Apr 21	4.00:00 AM BRONK	50475	40.890075	-73.819855 BOSTON ROAD	ROPES AVENUE	
	30	15-Apr-21	5.00.00 PM BROOKLYN	11236	40.650402	-73.89422 GLENWOOD ROAD	EAST 108 STREET	
	30	15-Apr-21	B 00:00 PM MANHATTAN	10025	40.00402	-73.97275 WEST 94 STREET	BROADWAY	
	33	16-Apr-21	S 30:00 PM MANHATTAN	10012	40.72538	-74.00011 PRINCE STREET	WOOSTER STREET	
	36	15-Apr-21	2:00:00 PM QUEENS	11377	40.75184	-73.90358 BRCADWAY	S8 STREET	
	34	16 Apr 21	12:00:00 PM BROCKLYN	11226	40.549788	-73.9622 EAST 19 STREET	CHURCH AVENUE	
	35	16-Apr-21	5.00.00 PM BROCKLYN	11220	40.585928	-75.920815	SUMPLIT ATCHING	
	35	10-Apr-21	1.00:00 PM BROOKLYN	1122	40.712963	-73 93647		
	30	14-4pr-21	8:00:00 PM N/A	ana a	40.801285	-73.95394 7 AVENUE		
	37	14-Apr-21	3.00:00 PM BROCKLYN	11220	40.633976	-74.02211		e

Once prompted, I'll perform a merge operation on the columns that uniquely identify one row in the dimension table (in this case, composite key of Borough and ZIP Code):

Apply * Source * Source * Data settings Par Close New Query Data Sources Pa		Choose Resoure Keep Restore Columni - Columni - Rows - Rows + Manage Columni - Reduce Rows So	Solit Group Lychepisce Calumn - By Lychepisce It Transform	Row as Headers - Noves	Merge Queries - End Analy Append Queries - Original Combine Files Aduer Mac Combine All Ins	hine Learning			
Queries (5)		ransformColumns(#"Removed Columns",{{"CR	SH TIME", Time.StartOfHour,	type time}})			*		
1. Collisions	CRASH DATE	Сплан тіме 💽 📲 воколен	* 1.2 DP CODE	• 1.2 LATTUDE	E 1.2 LONGITUDE	T AR ON STREET NAME	A ^R C CROSS STREET MAME	 PROPERTIES 	
Location	1 34-Apr-21	£				× 0 BRONK WHITESTONE BRIDGE		Name	
Contributing Factor	2 11-Apr-2	 Merge 				27		an entre	
Vehicle Type	3 15-Apr-20	1				0 HUTCHINSON RIVER PARKWAY		All Properties	
Time	4 33-Apr-2	E Select a table and matching	ig columns to create a merged t	able.		0: VANDERVORT AVENUE	ANTHONY STREET	* APPUED STEPS	
	5 12-Apr-2				0	d EDSON AVENUE			
	6 13-Apr-23	2 Collisions			G	0 VERRAZANO BRIDGE UPPER		Source	
	7 13-Apr-2	CRASH DATE CRASH TIN	I NOROUGH 1 ZIP CODE 2	ATITUDE LONGI	TUDE ON STREET NAME	0 53 ST	31AVE	Promoted Headers	1
	8 16-Apr-71	1 14-Apr-21 5:00:0	DAM N/A O	0	Ø BRONK WHITESTONE BRIDGE	0 SHORE PARKWAY		Changed Type Replaced Blank-> N/A	
	9 11-Apr-21	13-Apr-21 9:00:0	PM BROOKLYN 11217	40.68358 -7	3.37617	Ø GOWANUS RAMP		Replaced null -> 0	
	10. 16 Apr 3	1 15-Apr-21 4:00:0	PM N/A O	0	0 HUTCHINSON RIVER PARKWAY	0 BRONK RIVER PARKWAY RAMP		Uppercased street names	
	11 JJ-Apr-21	1 13-Apr-21 4:00:0	PM BROOKLYN 11222	0	Ø VANDERVORT AVENUE	0 BEVERLEY ROAD	EAST 21 STREET	Trimmed street names	
	12 15-Apr-2		* *** * ***			0 GREENFIELD AVENUE	OSGOOD AVENUE	Removed Errors	
	13 J0-Apr-21				1	Ø BROOKLYN BRIDGE		Removed Columns	
	14 23-May-25	9 Location	*		D	32 GOLD STREET	CONCORD STREET	Calculated Start of Hour	
	15 22 Jan 28	2	10 - C		5.0	Ø BRUCKNER BLVD	IgNST 158 STREET		
	16 26-fep-23	1 BOROUGH = ZIP CODE =	Indea			36			
	17 09-Mor-21	s N/A 0	1			74 COURT STREET	JORALEMON STREET		
	18 JI-Mor-21	BROOKLYN 22227	2			18 RALPH AVENUE	AVENUE K		
	19 06-Apr-2	5 5ROOKLYN 11222	1			28 BARCLAY AVENUE	HYLAN BOULEVARD		
	20 09-Apr-23	2 QUEENS JJ306	4			#6 BRONK RIVER PARKWAY			
	21 34-Apr-23	ERDOKLYN 11226	5			Ø MAJOR DEEGAN EXPRESSWAY R.			
	22 14-Apr-21					79 CROSS ISLAND PARKWAY			
	23 15-Apr-21					\$7			
	24 14-Apr-21					97 MYRTLE AVENUE			
	25 24-Apr-23	Use fuzzy matching to p	erform the merge			27			
	26 26-Apr-33	3 Fuzzy matching options				02 GARRISON AVENUE	LONGWOOD AVENUE		
	27 16-Apr-2	l rolly maching openin				PR HUGUENOT AVENUE			
	28 15-Apr-21	Estimating matches bas	ed on data previews		CK Cancel	27			
	29 16-Apr-21				Lance	55 BOSTON ROAD	ROPES AVENUE		
	30 25-Apr-2	1				22 GLENWOOD ROAD	EAST 108 STREET		
	31 36-Apr-21	B 00:00 PM MANHATTAN		20025	40.79535	3 97275 WEST 94 STREET	BROADWAY		
	32 36-Apr-21	S 00:00 PM MANHATTAN		30012	40.72538 -7	4.00012 PRINCE STREET	WOOSTER STREET		
	33 15-Apr-71	2.00:00 PM QUEENS		11377	40.75284 -7	3.90358 BROADWAY	S8 STREET		
	34 16-April:	I II.00.00 PM BROCKLYN		11226	40.649788	73.9622 EAST 19 STREET	CHURCH AVENUE		
	35 26-Apr-23	S 30:00 PM BROCKLYN		21222	40.586928 -73	920815			
	36 14-Apr-23	1 1-00-00 PM BROOKLYN		11211	40.712963 -7	3.93647			
	37 34-Apr-21	8:00:00 PM N/A		0	40.801285 -7	1.95394 7 AVENUE			
	38 14-Apr-21	3.00.00 PM BROOKLYN		11220	40.633976 -7	11220.4		6	

And after Power Query applies this transformation, I will be able to expand the merged Location table and take the Index column from there:

Queries (5)	* × ×	∫x = Table.Nestedloin	(#"Calculated Start of Hour", ("BOG	00694", "ZIP CODE"), Location, ("8080069", "ZIP CC	Of"), "Location", JoinKi	d.Inner)		~	Query Settings	
Collisions	T. F CHLIST	KILLED - 123 NUMBER OF N	ADTORIST HUU_ + +23 NUMBER OF MOTO	NEST RUL ALC CONTRIBUTING FACTOR VEHICLE 1 - 123 (COLUNION ID - A ^R C V	THICLE TYPE CODE 1	Location	-	+ PROPERTIES	
Location	1	0	0	Ø Following Too Closely	4407480 Set	Search Columns to Espand	24		Name	
Contributing Factor	2	0	ø	0. Pavement Slippery	4407665 Sta		1.44	^	Colisions	
Vehicle Type	3	Ø	0	Ø Unspecified	4406885 Sta	+ Expand O Appregate			All Properties	
Time	4	0	0	0 following Too Closely	4407883 Set	and the second sec			a second second	
targe	5	ø	ø	Ø Unspecified	4408060 Sec	(Select All Columns)			 APPLIED STEPS 	
	6	0	1	Ø Other Vehicular	4406314 Set	D BOROUGH			Source	
	7	0	0	8 Driver Instantion/Distraction	4408149 Sta	and the second se	0		Fromoted Headers	
	8	ø	0	Ø Unspecified	4406097 Sta	12 Index	2		Changed Type	
	9	0	0	Ø Driver inattention/Distraction	4277087 Set		-		Replaced Blank-> N/A	
	10	ø	2	Ø Following Too Closely	4397523 Pic	🖉 Use original column nam	e as prefix		Replaced null -> 0	
	27	0	1	Ø Driver Inattention/Distraction	4405914 Set			0	Uppercased street names Trimmed street names	
	12	ø	0	Ø Following Tao Closely	4406191 Sta		OK Cancel	P.	Removed Errors	
	13	0	1	0 Brakes Defective	4407366 Statio	n Wagon/Sport Utility Vehicle	Table		Removed Columns	
	14	0	0	0 Passing Too Closely	4407463 Box 7	nuck	Table		Calculated Start of Hour	
	15	0	0	Ø Driver Inexperience	4407760 Sedar	6	Tatie		X Merged Queries	
	16	0	0	Ø Driver Inattention/Distraction	4407765 Sedar	0	Table		and the second se	
	17	ø	0	Oriver Instsention/Distraction	4407520 Bus		Table			
	16	0	0	Ø Reaction to Uninvolved Vehicle	4380668 Sedar	Carrier and the second	Table			
	19	0	0	Ø Unspecified	4427902 Static	n Wagon/Sport Utility Vehicle	Table			
	20	Ø	0	0 Unspecified	4407971 Sedar	ē	Table			
	21	0	0	0 View Obstructed/Umited	4409071 Static	n Wagon/Sport Utility Vehicle	Table			
	22	0	3	Ø Unspecified	4408369 Statio	e Wagon/Sport Utility Vehicle	Table			
	23	0	1	0 Passing or Lane Usage Improper	4407708 Sedar	6	Table			
	24	0	0	Ø Other Vehicular	4407152 Sedar	(Take			
	25	0	0	Ø Reaction to Uninvolved Vehicle	4407862 Sedar		Tatte			
	26	Ø	0	Ø Following Too Closely	4407435 Sedar		Table			
	27	0	0	Ø Driver Inattention/Distraction	4407636 Static	e Wagon/Sport Utility Vehicle	Table			
	28	0	0	Ø Unspecified	4408205 Box 1	ruck	Table			
	29	0	0	Ø Unspecified	440/945 Sedar		Table			
	30	ø	0	0 Aggressive Driving/Road Rage	4407563 Sedar	(Table			
	31	ø	0	0 Driver Inattention/Distraction	4408098 Sedar	S.	Table			
	3.2	Q	3	Ø Unsafe Speed	4407159 Static	n Wagon/Sport Utility Vehicle	Table			
	33	0	0	Ø Passing Too Closely	4407994 Static	n Wagon/Sport Utility Vehicle	Table			
	34	Ø	0	Ø Reaction to Uninvolved Vehicle	4407497 Static	n Wagon/Sport Utility Vehicle	Table			
	35	0	0	0 Backing Linsafety	4408229 Static	n Wagon/Sport Utility Vehicle	Tahte			
	36	σ	0	Ø Driver Inattention/Distraction	4408306 Sedar	6	Table			
	37	ø	0	Ø Following Too Closely	4407020 Statio	n Wagon/Sport Utility Vehicle	Table	~		
	38	0	0	Ø Unspecified	4408200 Sedar		Table			

Now, I can use this one integer column as a foreign key to my Location dimension table, and simply remove two attribute columns BOROUGH and ZIP CODE – this way, not that my table is cleaner and less cluttered – it also requires less memory space – instead of having two text columns, we now have one integer column!

* De Da Corre	Manage Refresh Preview *	Advanced Editor Manage *	e Remove * Columns ge Columns Reduce Roms Reduce Roms Series Seri	Split Gr Column	Data Type: Any • Dute First Row as Headers • Outp L ₊₂ Replace Values Transform					
ieries (5)	< × × 5	- Table.ReorderColum	ns(#"Renamed Columns",{"CRASH D	ATE", "CRASH	TIPE", "LATITUDE", "LONGITU	er, "ON STREE		which means that you can	Query Settings	
Colisions		T KILLED T 123 Location Key	A BOROUGH	- 1.2 20	-1.2			move original attributes!	* PROPERTIES	
Location	a the or creation	I write [.] 1-1 coverses and	2 N/A	E STATE	Copy				Name	
		0	2 N/A		Remove Columna	And in case of the local division of the loc			Colisions	
Contributing Factor	2	0	J N/A	-	Remove Other Columns				All Properties	
Vehicle Type	4	0	1 N/A		Add Column From Examples.			Ø Following Too Closely	and properoes.	
Time	-	0	1 1/4				0	0 Unspecified	→ APPLIED STEPS	
		0	1 N/A 1 N/A		Remove Duplicates		e .	0 Other Vishicular	Source	
	0.	0	1 N/A		Remove Errors			0 Driver inattention/Distraction	Promoted Headers	
		a	J N/A J N/A	1	Replace Values		0	0 Driver inattention/Distraction 0 Unspecified	Changed Type	
	8		1 N/A		154				Replaced Blank-> N/A	
	9				Change Type		0	Ø Driver inattention/Distraction	Replaced null -> 0	
	10.		2 N/A		Merge Columns	_		8 Following Too Closely	Uppercased street nan	nes
		0	3 N/A				1	Ø Driver Inattention/Distraction	Trimmed street names	
		D	2 N/A		E. Group By		9	Ø Following Too Closely	Removed Errors	
		0	I N/A	1	Unpivot Columns		1	Ø Brakes Defective	Removed Columns	
Location Key now	serves as	a	2 N/A		Unpivot Other Columns		0	0 Passing Too Closely	Calculated Start of Ho	ur:
Foreign Key to	Location	0	1 N/A		Unpivot Only Selected Column	G	0	8 Driver inexperience	Merged Queries	
		0	I N/A		Move	14	0	0 Driver Inattention/Distraction	Expanded Location	
dimension ta	able	Ø	3 N/A		1940/YE		0	0 Driver inattention/Distraction	Renamed Columns	
		0	I N/A		ø		0	Ø Reaction to Uninvolved Vehicle	X Reordered Columns	
		Ø	1 N/A		o		0	Ø Unspecified		
		0	2 %/A		ø		0	0 Unspecified		
	23	o	J N/A		σ		ø	Ø View Obstructed/Linvited		
	22	Ø	1 N/A		ø		3	0 Unspecified		
	23	ø	3 N/A		o		2	Ø Passing or Lane Usage Improper		
	24	0	1 N/A		0		0	Ø Other Vehicular		
	25	ø	1 N/A		0		0	Ø Reaction to Uninvolved Vahicle		
	26	σ	I N/A		a		0	0 Following Too Closely		
	27	. 0	3 N/A		ø		0	0 Driver insttention/Distraction		
	28	ø	1 N/A		σ		0	0 Unspecified		
	29	0	2 N/A		0		0	0 Unspecified		
	30	0	2 M/A		σ		0	0 Aggressive Driving/Road Rage		
	31	0	z N/A		σ		0	Ø Driver Inattention/Distraction		
	32	ø	2 N/A		ø		3	Ø Uncafe Speed		
	33	0	1 N/A		ø		0	Ø Passing Too Closely		
	34	0	2 N/A		ø		Ø	Ø Reaction to Uninvolved Vehicle		
	35	ø	2 N/A		ø		0	0 Backing Unsafely		
		a	z 50/A		a		ö	0 Driver Inattention/Distraction		
	36									
	36	0	3 N/A		ø		0	Ø Following Too Closely	~	

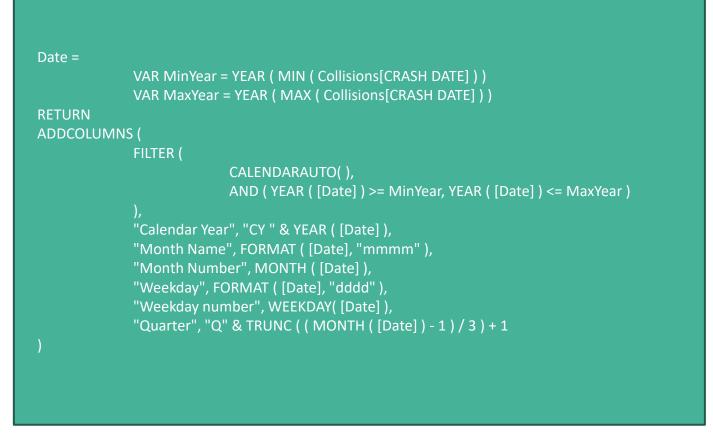
I will apply the same logic to other dimensions (except the Time dimension) – include index columns as foreign keys and remove original text attributes.

Enhancing the data model with Date dimension

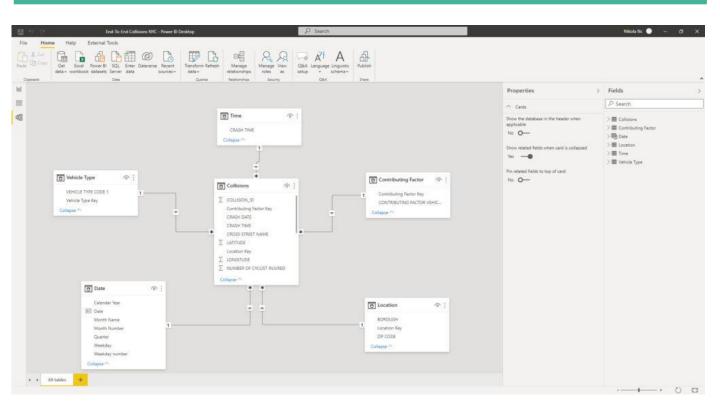
Now, we're done with data modeling in Power Query editor and we're ready to jump into Power BI and enhance our data model by creating a Date dimension using DAX. We could've also done it using M in Power Query, but I've intentionally left it to DAX, just to show you multiple different capabilities for data modeling in Power BI.

It's of key importance to <u>set a proper Date/Calendar dimension</u>, in order to enable DAX Time Intelligence functions to work in a proper way.

To create a Date dimension, I'm using this script provided by SQL BI folks.



After I've marked this <u>table as a date table</u>, it's time to build our Star schema model. I'll switch to a Model view and establish relationships between the tables:



Does that remind you of something? Exactly, looks like the star illustration above. So, we followed the best practices regarding data modeling in Power BI and built a Star schema model. Don't forget that we were able to do this without leaving the Power BI Desktop environment, using Power Query Editor only, and without writing any code! I hear you, I hear you, but DAX code for Date dimension doesn't count:)

Summary

Our analytic solution is slowly improving. After we performed necessary data cleaning and shaping, we reached an even higher level by building a Star schema model which will enable our Power BI analytic solution to perform efficiently and increase the overall usability – both by eliminating unnecessary complexity, and enabling writing simpler DAX code for different calculations.

As you witnessed, once again we've proved that Power BI is much more than a visualization tool only!

In the next chapter, we will finally move to that side of the pitch and start building some cool visuals, leveraging the capabilities of the data model we've created in the background.

Introduction

Good news, folks – slowly, but steadily, we are nearing our goal – to build an efficient end-toend analytic solution using Power BI only! After we emphasized the importance of understanding business problem *BEFORE* creating a solution, performed some simple data cleansing and transformation, in the previous part we learned why Star schema and Power BI are match made in heaven, and why you should always strive to model your data that way.

Now, it's time to build some compelling visualizations that will help us to tell the data story in the most effective way and provide insight to business decision-makers – in the end, based on those insights, they will be able to make informed decisions – decisions based on the data, not on a personal hunch or intuition!

DISCLAIMER: I consider myself as a person not aesthetically talented, so my data visualization solutions are mostly based on the best practices I've read in the books (<u>this one</u>, for example), blogs, and inspired by some amazing community members, such as Armand Van Amersfoort, Daniel Marsh-Patrick, Kerry Kolosko, Ried Havens, Andrej Lapajne (<u>Zebra BI</u>), or folks from <u>powerbi.tips</u>.

Data Visualization – my top list!

Before we pull up our sleeves and start to visualize our data, I would like to point out a few best practices regarding data visualization I picked along the way.

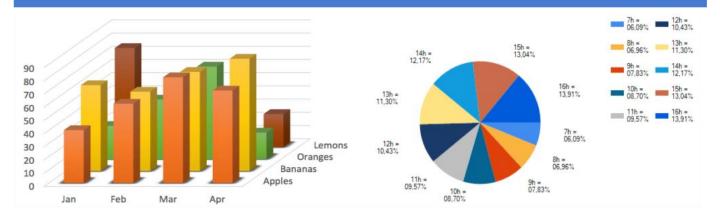
1. One dashboard to rule them all

This one applies not to data visualization exclusively, it's more of a general rule. There is no single solution to satisfy each and every business request! Period! As a first step, you should determine the purpose of the dashboard – **operational dashboards** provide time-critical data to consumers. I like to think about operational dashboards as of cockpit in the car or plane...On the other hand, **analytical dashboards** focus more on identifying trends and patterns from historical data and enable better mid to long-term decision-making.

In our case for this brochure, we are building an analytical dashboard.

2. Picking the right visualization type

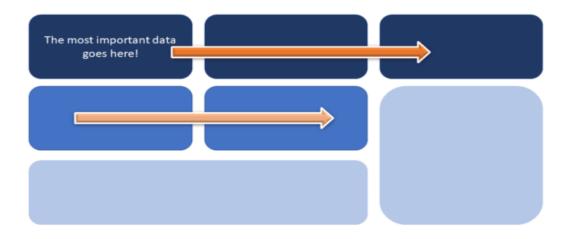
Uh, this one is probably the most complicated to define. There are literally hundreds of blog posts, books, videos from established authors, explaining which visualization type to use for specific data representation. Based on what kind of insight you want to provide – for example, comparison between two data points, distribution of specific value, relationships between different data, changes over time, parts of the whole, and so on – there are certain visual types that *SHOULD* be used.



I've intentionally used the word SHOULD, as no one can prohibit you from using, let's say, gauges, pie-charts, 3D-charts in your dashboards, even though a lot of recognized experts advised against that – just be careful and mindful in which scenario to use what visual type.

3. Define the most important data points

Obviously, some data points have higher importance than others. If your overall revenue is 50% lower than in the previous month, it's definitely way more significant than looking at the chart showing individual numbers per product color. With that in mind, try to place all key data points in the top left corner, as most of the people on our planet read from left to right, and from top to bottom (imagine reading a book, or newspapers), and that position will naturally catch their attention right away.

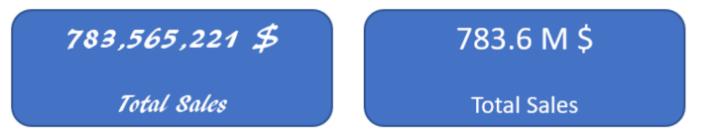


4. Be consistent!

This is one of the key things to keep in mind! What does consistency mean? For example, sticking with a defined layout and design, putting related information close to each other, or using similar visual types for a similar type of information – you don't want to use a pie-chart to display Sales Amount by Region in one dashboard part, and then use a column bar chart to display, let's say, Total Orders by Region.

5. Remove the distraction

I've already written about one specific case of <u>removing distractions from the Power BI report</u>. There are a lot of possible distractors in your dashboards. Let's start with fonts: tend to use standard fonts instead of artistic ones, as they are easier to consume:



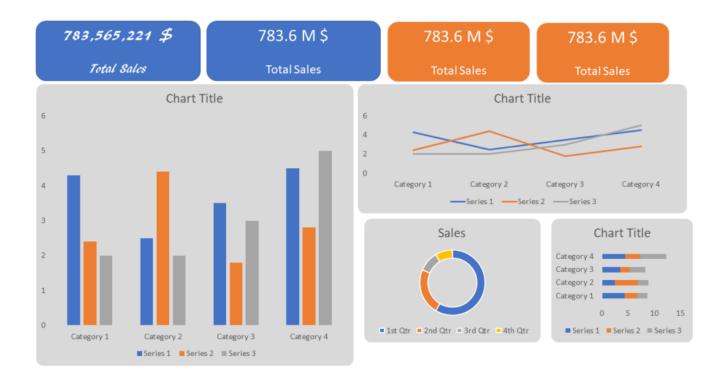
The illustration above demonstrates the easier readability of the card on the right, which uses one of the standard fonts (Calibri). It also demonstrates another point to consider – shortening numbers is also a good way to remove distraction from your dashboard.

In addition, take care of the proper alignment and give your visuals some space in between:



Proper alignment and space between visuals will improve the clarity

I assume we can agree that the dashboard on the illustration above is way more readable than the one below:



There are many more best practices, tips, and recommendations when it comes to data visualization. As I've already stressed, I don't consider myself a "data viz" wizard, but I'm still trying to stick with some of the general rules mentioned in the previous chapter.

Finally, even though that for many dashboard creators the first step is to set the overall dashboard design and then fit data elements in the predefined template, I prefer doing the opposite: first, I create all data elements, and then based on the story I want to "tell" with these elements, I'm building a final solution...

Visualizing vehicle collisions data

Ok, now as we identified some of the general data visualization best practices, it's time to get our hands dirty and use Power BI to tell the story about the vehicle collisions in NYC.



This is how my report looks like. In this part, we won't go deep into the details about each visual, but let me just briefly introduce the overall concept. There are two pages – Main page contains the most important data points, such as the number of collisions, deaths, and injuries. There are also a few "classic" visuals, like the Line chart and Column Bar chart, that will help us to extract the insights looking at the data from different perspectives. Multi-row card visual quickly illustrates who is the most endangered in the traffic.

Time of day is one of our key analytic categories, so report users have full flexibility to switch between different metrics on the same visual (Collisions, Deaths, Injuries) – keep an eye on the dynamic title – this enhances the overall user experience!

Remember, we defined a set of questions <u>here</u> that we'll try to answer using this report. Data can be sliced from a calendar perspective using a Date slicer.

Details page gives a possibility to dive deeper into the details about accidents – introducing additional slicers for Borough and ZIP Code. Small Multiples visual nicely breaks down figures by two categories – person type and borough, while other elements extend the logic from the Main page.

Summary

We've covered a lot in this chapter. Not that we just built our own report to visualize the data from the original dataset, I've also shared with you some of the general best practices when it comes to data visualization, and recommendations from proven experts on this topic.

I'll repeat this: I consider myself far from being talented to be a "designer", and I'm sure that many of you can create a better-looking Power BI report. However, the end goal is to effectively communicate the key data points to report consumers, and enable them to make decisions based on the insights provided by this communication.

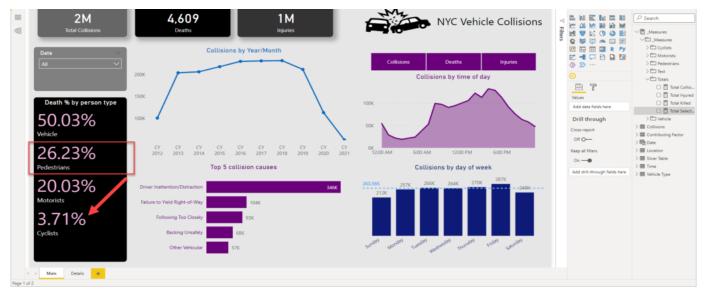
With that in mind, I believe that we built a solid foundation to wrap everything up in the final chapter – we'll try to extract some meaningful information from the report we've just created and recommend certain actions in accordance with the findings.

Introduction

Here we are – after taking <u>raw data in the form of a CSV file</u>, defining a set of business questions that need to be answered using that data, then cleaning and shaping the original dataset and building an efficient data model (Star schema), in the previous part we've created compelling visualizations to provide different insights to business decision-makers. Now, it's time to analyze insights and, based on the information we extract from these insights, recommend some actions!

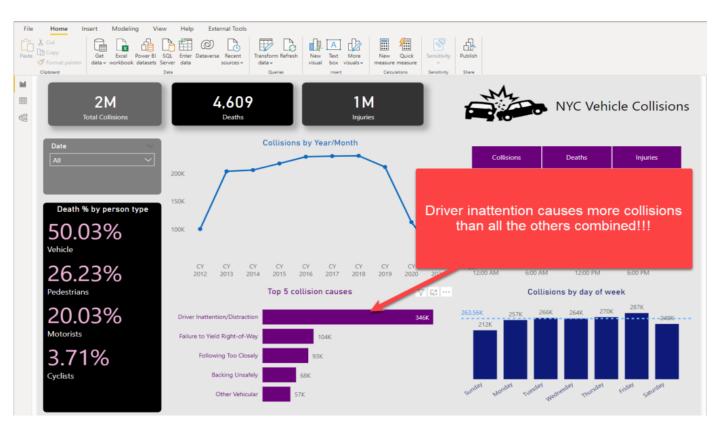
Extracting the insights

Let's start by analyzing deaths caused by collisions. If we exclude persons that were in the vehicles themselves, we can see that the pedestrians are the most endangered traffic participants – almost 8x more pedestrians were killed, compared to cyclists!

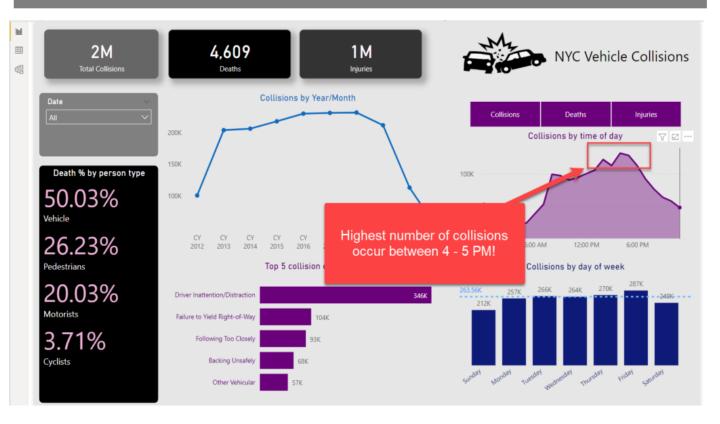


END-TO-END ANALYTICS WITH MICROSOFT POWER BI

The next conclusion we can draw is that the main cause of the collisions is drivers' inattention/distraction! If we take a look at the top 5 collision causes, you will see that all other causes combined are lower than the top one.



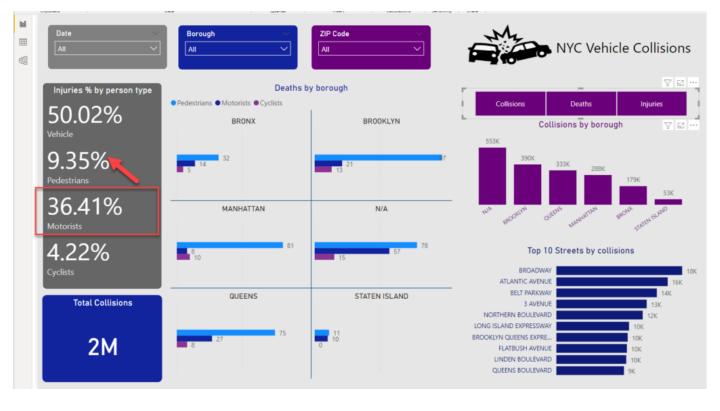
Moving on, and the next pattern we can spot is a significant spike of the collisions in the early afternoon hours, specifically 4 and 5 PM. This makes sense, as a lot of people are driving home in that period, returning from their offices:



It's approximately 30% higher than in the morning (8-9 AM) when traffic participants are probably not tired and distracted after a hard-working day.

Let's move on to a detailed overview of the accidents and try to identify "black" spots in the city. At first glance, most people are being killed in Brooklyn, and more or less, all other boroughs follow the pattern of the "death distribution" between the different traffic participants, except Manhattan, where more cyclists were killed than motorists.

If we analyze the percentage of injuries, the trend is quite different than with fatality rates: now, motorists are the most endangered (again, excluding persons that were in the vehicles) – almost 4x more motorists were injured than pedestrians!



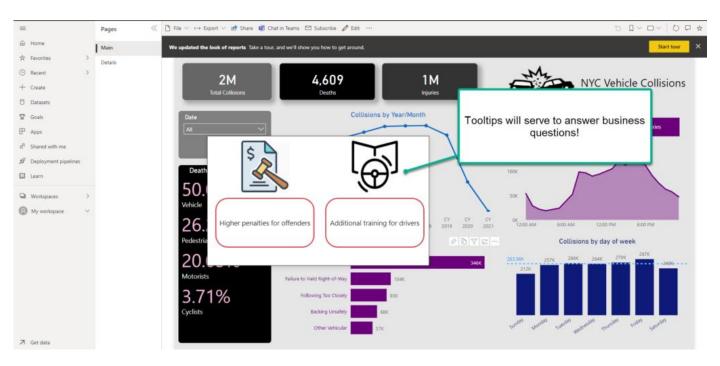
Further down, ZIP Codes with the most frequent collisions are 11207 and 11101 (one in Brooklyn, the other in Queens). If we focus on the specific street, we can see that Broadway (Manhattan) and Atlantic Avenue (Brooklyn, ZIP Code 11207) are the most critical spots in New York City!



Action, please!

Ok, now we have way more information to support our business decisions. And, since we already defined the set of questions that need to be answered, let's focus on providing the proper recommendations for actions!

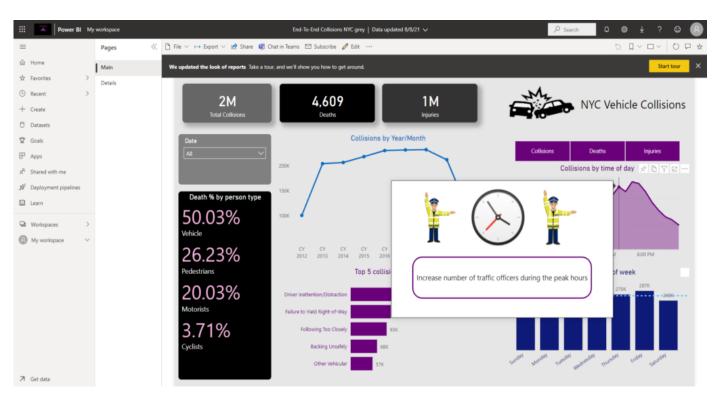
The idea is to show recommendations in the form of tooltips – when someone hovers over a specific visual, the respective action should be displayed! I've already written how to <u>enhance</u> <u>your report using tooltip pages</u>, and here we will follow a similar approach:



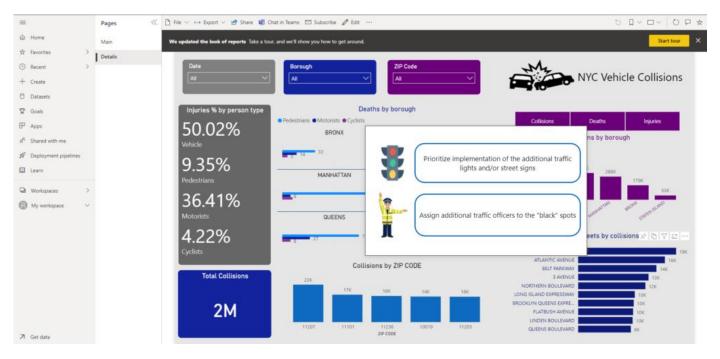
So, once I hover over a visual that shows the top 5 causes for collisions, certain actions will be recommended:

- ✓ Higher penalties for offenders
- ✓ Additional training for drivers

Similarly, if you want to act based on the time when most collisions occur, simply hover over that visual and you will see the suggestion to increase the number of traffic officers during these peak hours:



Further down, to be able to reduce the number of collisions and injuries/deaths in specific locations, we emphasized the importance of implementing additional traffic lights and assigning more traffic officers to "black" spots:

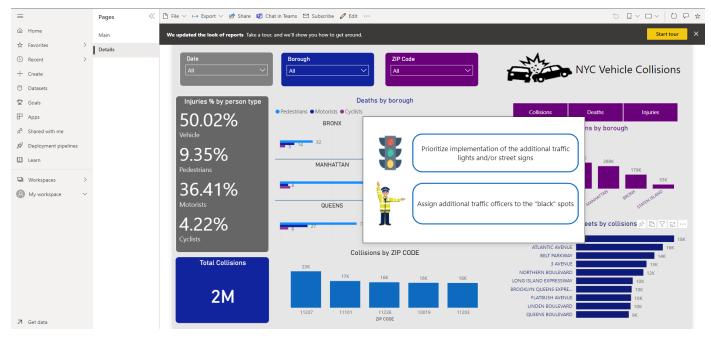


SUMMARY

That's a wrap folks! Just to remind you, this is where we've started:

A	e c	D	6		G			1				M	N	0	P		Q R		T		V	W	X	. ¥		AA.	AB	AC
	RASH TIN BOROUGH	ZIP CODE	LATITUDE	LONGITUD						BER (NUN	BERCN		NUMBER	NUMBER	NUMBE	RINUI	MBER (NUM				CONTRIB	LCONTRIB				VEHICLE T	VEHICLETY	VEHICLE TY
04/14/202	5:32							NE BRIDG		0	0	0	0	2	0	0	0	0 Following		d			4407480 5		Sedan			
04/13/202	21:35 BROOKLY	11217	40.68358	-73.9762				620		1	0	1	0	5 I	0	0	0	0 Unspecifi					4407147 5					
04/15/202	16:15							ER PARKW	AY	0	0	0	0		0	0	0	0 Pavemen							/agon/Spor	t Utility Ve	hicle	
04/13/202	16:00 BROOKLY	11222	2					NY STREET		0	0	0	0	() I	0	0	0	0 Following					4407811 5					
04-12-21	8:25		0	0	(0.0, 0.0)					0	0	0	0		0	0	0	0 Unspecifi					4406885 5					
04/13/202	17:11					VERRAZA	ANO BRID	GE UPPER		0	0	0	0		0	0	0	0 Following					4407883 5	Sedan	Box Truck			
04/13/202	17:30 QUEENS	11106	5			33 st	31ave			0	0	0	0	E 1	0	0	0	0 Driver Ina	Unspecifie	d			4408019 5	Sedan	Sedan			
04/16/202	23:30					SHORE P	ARKWAY			0	0	0	0	÷.	0	0	0	0 Unspecifi	ed				4408060 5	Sedan				
04-11-21	17:00					GOWAN	US RAMP			1	0	0	0		0	0	1	0 Other Vel	Other Veh	icular			4406314 5	iedan	Sedan			
04/15/202	21:15					BRONX R	IVER PAR	KWAY RAS	MP.	0	0	0	0	(I	0	0	0	0 Driver Ina	Unspecifie	d			4408149 5	Itation W	/ Sedan			
04-11-21	21:06 BROOKLY	11226	5			BEVERLE	Y EAST 2	STREET		1	0	1	0	0	0	0	0	0 Passing Te	on Closely				4406488 7	Taxi				
04/15/202	20:00 STATEN IS	10304	i)			GREENFI	EIOSGOO	D AVENUE		0	0	0	0	K I	0	0	0	0 Unspecifi	ed				4408310 5	Sedan				
04-10-21	11:15					BROOKLY	N BRIDG	E		0	0	0	0	()	0	0	0	0 Unspecifi	Unspecifie	d			4406097 5	Itation W	Station W	agon/Spor	t Utility Veh	ticle
05/21/201	22:50 BROOKLY	11201	40.69754	-73.9831	(40.69754,	GOLD STR	RICONCO	AD STREET		0	0	0	0	0	0	0	0	0 Passing or	Unspecifie	d			4136992 %	UBMAKS	Taxi			
01/21/202	15:49					BRUCKNS	EI BUIST	38 STREET		0	0	0	0	6 I	0	0	0	0 Driver Ina	Unspecifie	d			4277087 5	iedan	Station W	agon/Spor	t Utility Veh	licle
02/26/202	14:50 BRONX	10461	40.84346	-73.836	(40.843464	, -73.836)	2819	MI	0	0	0	0		0	0	0	0 Unspecifi	Unspecifie	d			4395664 5	station W	lagon/Spor	t Utility Ve	hicle	
03-09-21	11:00		40.69255	-73.991	(40.69254)	COURT ST	TIJORALE	MON STRE	ET	1	0	0	0	0	0	0	1	0 Following	Unspecifie	d			4397513 4	Pick-up Ti	r Sedan			
03/31/202	22:20 BROOKLY	11234	40.62646	-73.918	(40.62645)	RALPH A	VAVENU	EK		1	0	0	0	8 1	0	0	1	0 Driver Inc	Unspecifie	d			4403773 5	Sedan	Sedan			
04-06-21	22:58 STATEN IS	10312	40.52689	-74.1673	(40.52689)	BARCLAY	HYLAN	BOULEVAR	D	7	0	0	0		0	0	7	0 Failure to	Unsafe Spe	bed			4405244 5	station W	Station W	agon/Spor	t Utility Veh	licle
04-09-21	14:45		40.84078	-73.8725	(40.84077	BRONX R	IVER PAR	KWAY		1	0	0	0		0	0	1	0 Driver Ina	Unspecific	Unspecifi	Unspecifi	ied	4405914 5	iedan .	Station W	Station W	agon/Sport	Utility Ve
04/14/202	13:00					MAJOR	EEGAN E	XPRESSW/	CY F	0	0	0	0	S - 1	0	0	0	0 Following	Unspecifie	d			4408191 5	station W	Box Truck			
04/14/202	11:00		40.69404	-73.7268	(40.69403:	CROSS IS	LAND PA	RKWAY		1	0	0	0		0	0	1	0 Brakes De	Unspecifie	d			4407366 5	station W	/ Sedan			
04/15/202	13:30 BRONX	10461	40.85737	-73.8466	(40.857365	, -73.846	57)	2400	PE	0	0	0	0		0	0	0	0 Following	Unspecifie	d			4407778 5	iedan	Ambuland	ce .		
04/14/202	14:40		40.69881	-73.9184	(40.69880)	MYRTLE /	AVENUE			0	0	0	0	£	0	0	0	0 Passing Tr	Unspecifie	d			4407461 E	Box Truck	Sedan			
04/14/202	21:43 QUEENS	11429	40.71402	-73.7483	(40.71402,	-73,7482	7)	211-20	.91	0	0	0	0		0	0	0	0 Turning In	Unspecifie	d			4407407 5	station W	lagon/Spor	t Utility Ve	hicle	
04/16/202	17:40 BRONX	10474	40.815	-73.894	(40.815, -7	GARRISO	LONGY	OOD AVE	NU.	0	0	0	0		0	0	0	0 Unspecifi	Unspecifie	d			4407900 5	iedan	Sedan			
04/16/202	0:50		40.55079	-74.201	(40.55079)	HUGUEN	OT AVEN	UE		0	0	0	0		0	0	0	0 Driver Inc	Unspecifie	Unspecifi	ed		4407760 5	Sedan	Station W	Sedan		
04/15/202	10:30 BROOKLY	11207	40.6559	-73.8982	(40.655903	73.898	17)	319	DE	0	0	0	0		0	0	0	0 Unspecifi	Unspecifie	d			4407746 5	Sedan	Station W	agon/Spor	t Utility Veh	sicle
04/16/202	16:15 BRONX	10475	40.89008	-73.8199	(40.89007)	BOSTON	FROPES	AVENUE		2	0	0	0		0	0	2	0 Unsafe Sp	Unspecifie	d			4408143 5	station W	Station W	agon/Spor	t Utility Veh	ticle
04/15/202	17:20 BROOKLY	11236	40.6504	-73.8942	(40.650400	GLENWO	CEAST 1	STREET		1	0	0	0	1	0	0	1	0 Following	Unspecifie	d			4407638 5	Sedan	Van			
04/16/202	21:20 MANHAT	10025	40.79335	-73.9728	(40.79335,	WEST 94	SBROAD	WAY		0	0	0	0		0	0	0	0 Backing U	Unspecifie	d			4407958 5	Sedan				
04/16/202	17:20 MANHAT		40,72538		(40.72538.					0	0	0	0		0	0	0	0 Driver Ina					4407885 5	station W	Sedan			
04/15/202	14:30 QUEENS	11377	40,75184	-73.9036	(40.75184,	BROADW	1 58 STRE	ET		0	0	0	0		0	0	0	0 Driver Ina	Driver Inat	tention/E	Nstraction		4407616 5	Sedan	Station W	agon/Spor	t Utility Veh	sicle
04/16/202	23:20 BROOKLY		40.64979		(40.64978)					3	0	0	0		0	0	3	0 Unspecifi									agon/Sport	
04/16/202	18:15 BROOKLY				(40.686928				105	0	0	0	0		0	0	0	0 Driver Ina			1		4408224 5				t Utility Veh	
04/14/202	13:00 BROOKLY		40.71296						GR.	0		0	0		0	0	0	0 Passing of					4407392 1					
04/14/202	20:14				(40.801285					0	0	0	0		0	0	0	0 Driver Ina					4407765 5		Sedan			
	Motor Vehicle C			(+)							~																	

And, this is where we finished:



END-TO-END ANALYTICS WITH MICROSOFT POWER BI

SUMMARY

Along the way, we cleaned and transformed our data, built a proper data model using Star schema, and visualized key data points. And, guess what – **we did ALL of that using one**

SINGLE tool: Power BI! That's the reason why I've called this series of blog posts: Building an end-to-end analytic solution with Power BI – and I believe we can agree that this brochure proved that without exaggeration.

So, next time when you hear about "Power BI as a visualization tool only", just remember what we were able to achieve using this tool exclusively, and draw a conclusion on your own!

THANK YOU!



DATA-MOZART.COM



@DataMozart

2022 All Rights Reserved