NIKOLA ILIC

END-TO-END ANALYTICS WITH MICROSOFT POWER BI

CRASH COURSE ON BUILDING POWERFUL ANALYTIC SOLUTIONS

DATA MOZART
Make Music from your Data!

TABLE OF CONTENTS

FOREWORD

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.

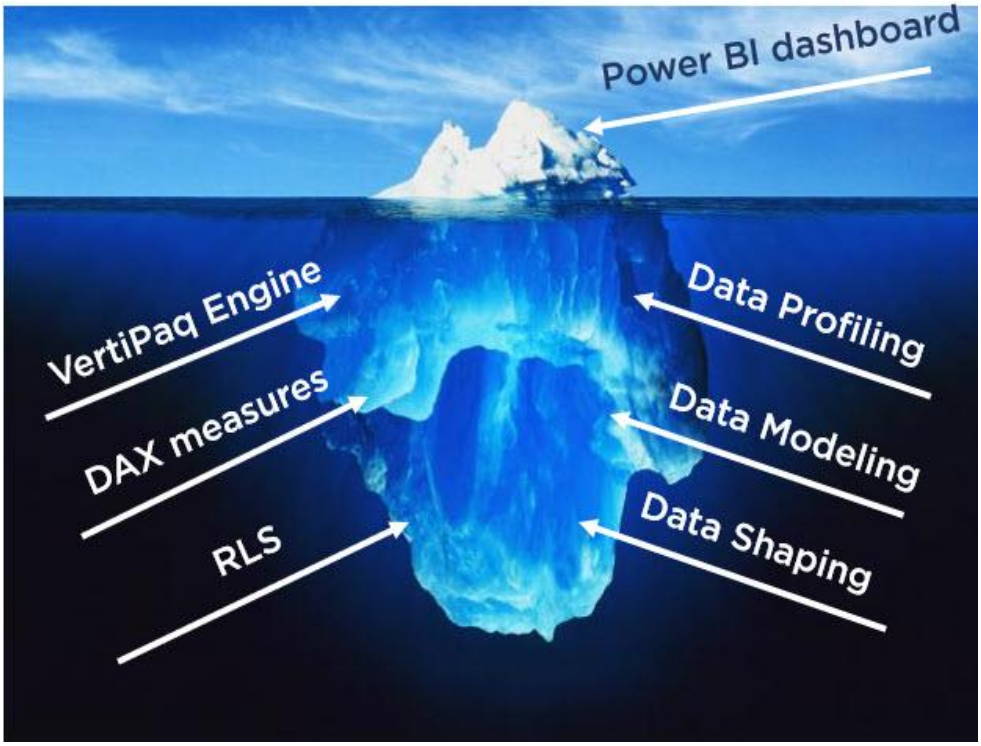
In this brochure, I'll show you ***how Power BI can be used to create a fully-fledged analytic solution****. Starting from the raw data, which doesn't provide any useful information, to building, not just nice-looking visualizations, but extracting insights that can be used to define proper actions –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

INTRODUCTION

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…

INTRODUCTION



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.

INTRODUCTION

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 here. 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

INTRODUCTION

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

Data Preparation

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

INTRODUCTION

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 Star-schema 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

INTRODUCTION

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.

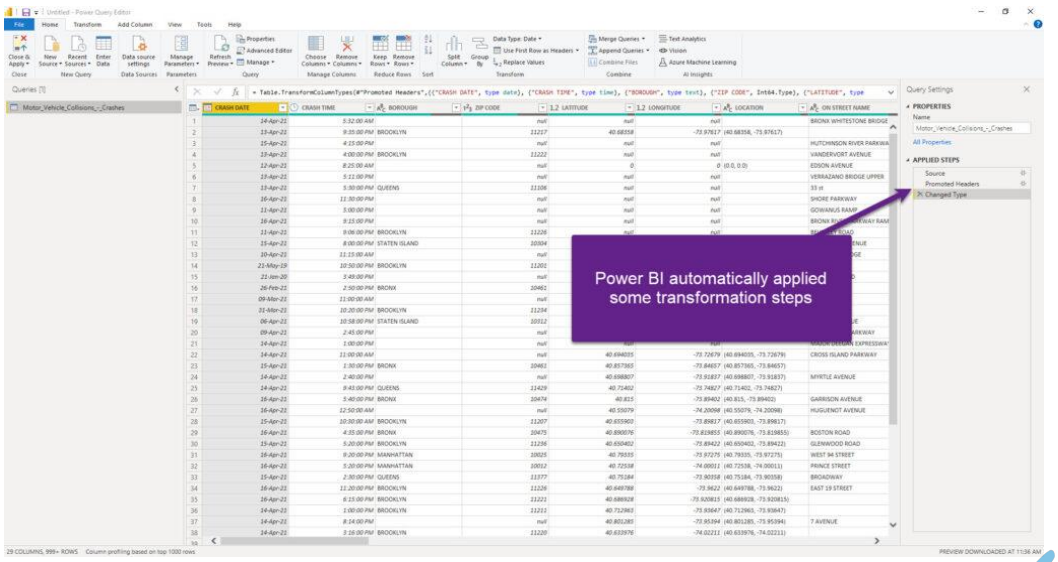
DATA PREPARATION

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 here.

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:



DATA PREPARATION

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 shaping our data. 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.

DATA PREPARATION

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!

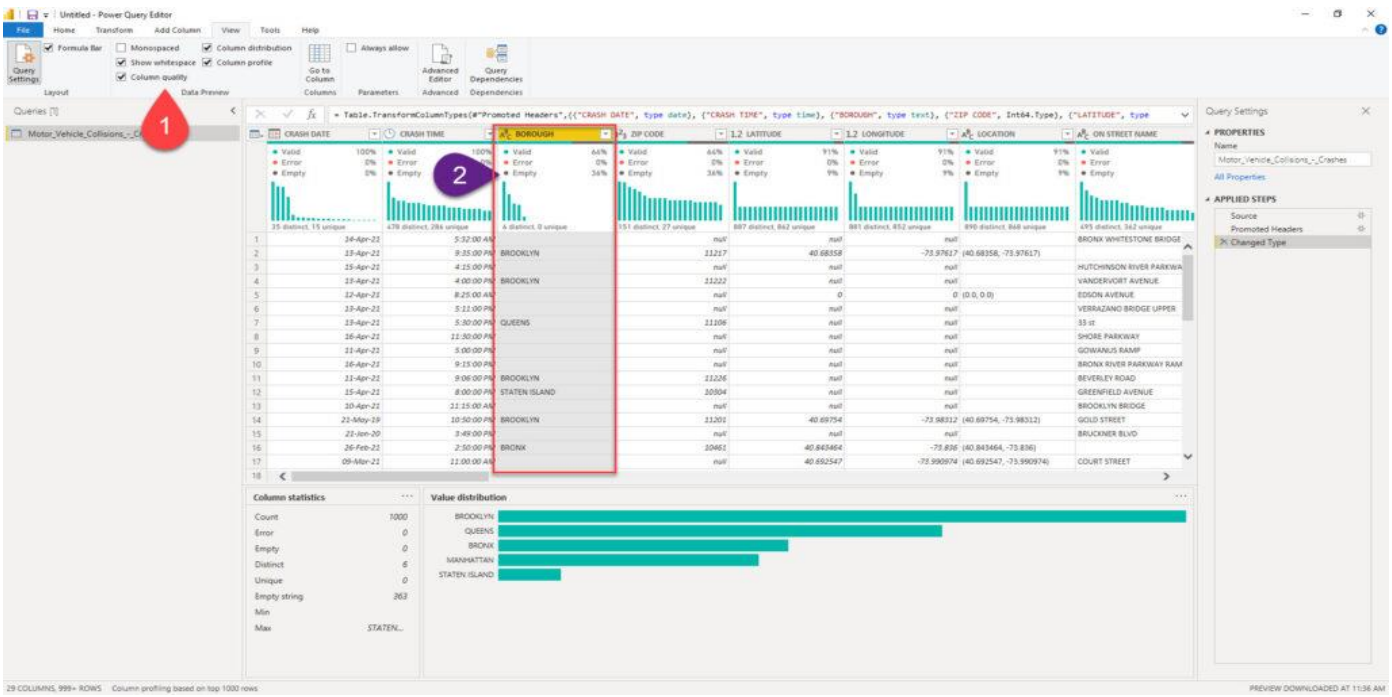
Here, as a best practice, I'll quote Matthew Roche's famous "maxim":

> ***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 data profiling. 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.

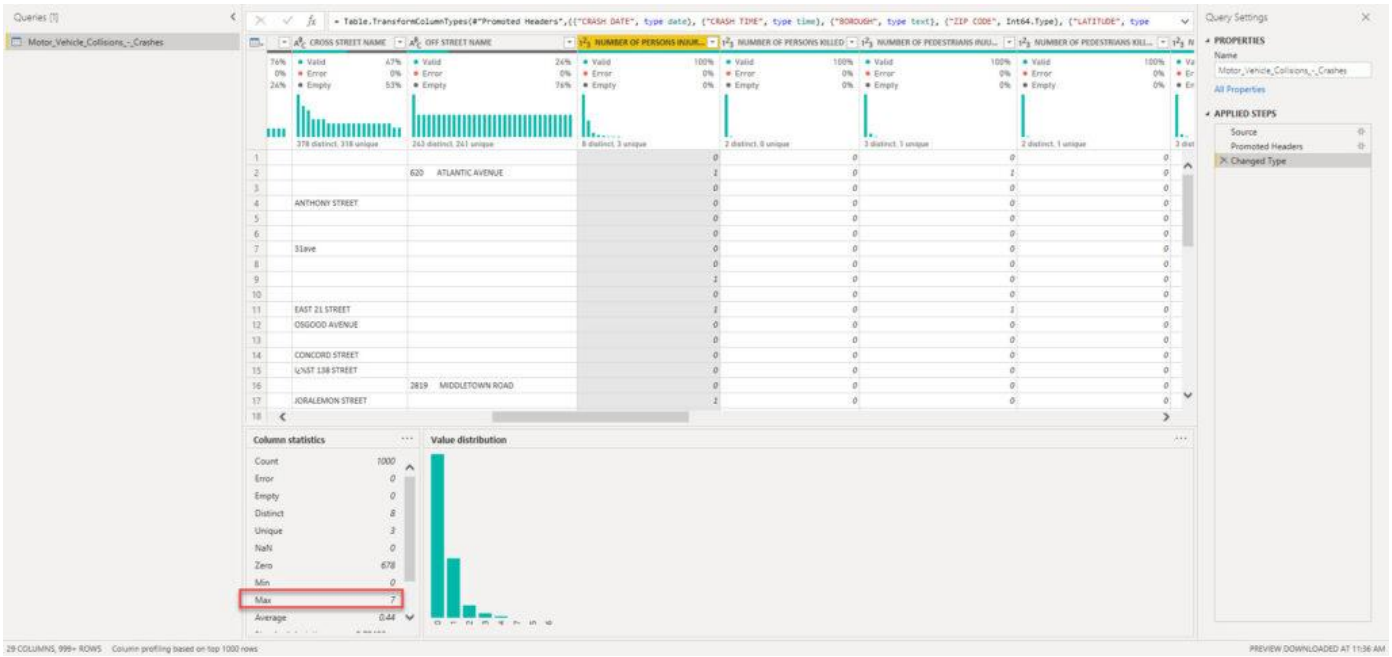
DATA PREPARATION



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:

DATA PREPARATION

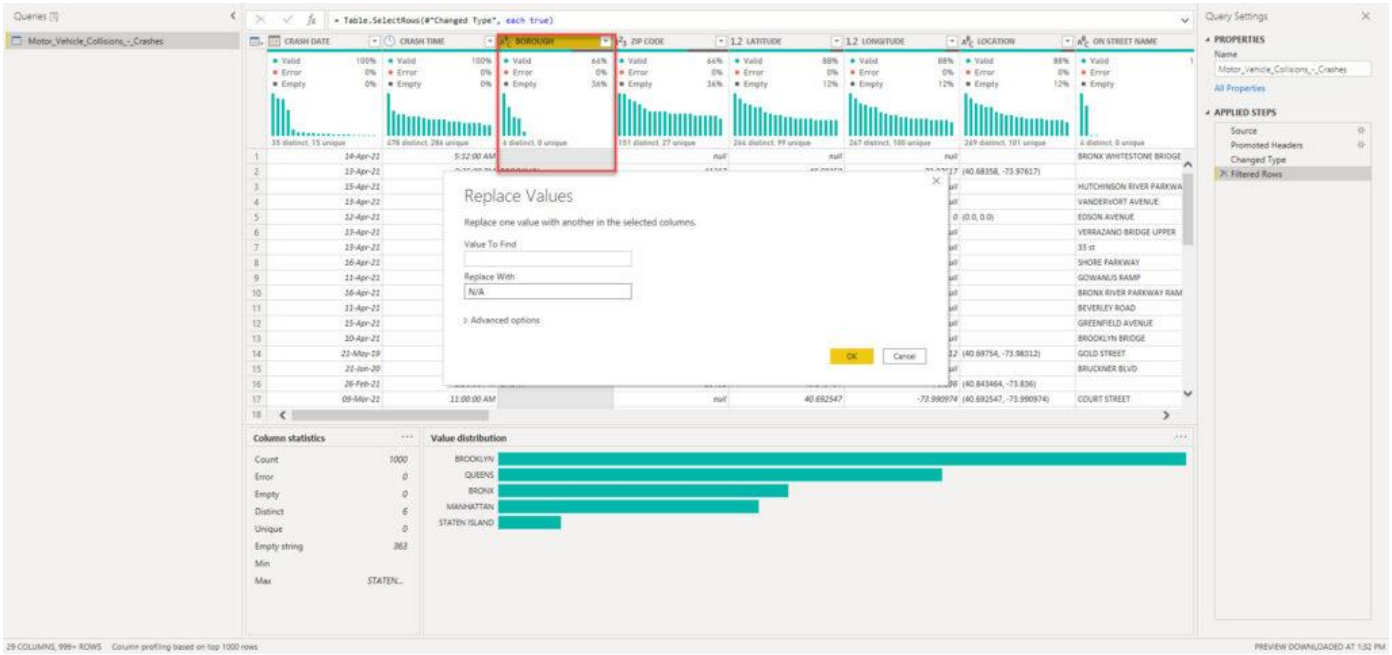


If there were some data anomalies (i.e. instead of 7 for the Max number of injured persons, let's say 7000), we would be able to spot that right away and react accordingly!

Data Shaping

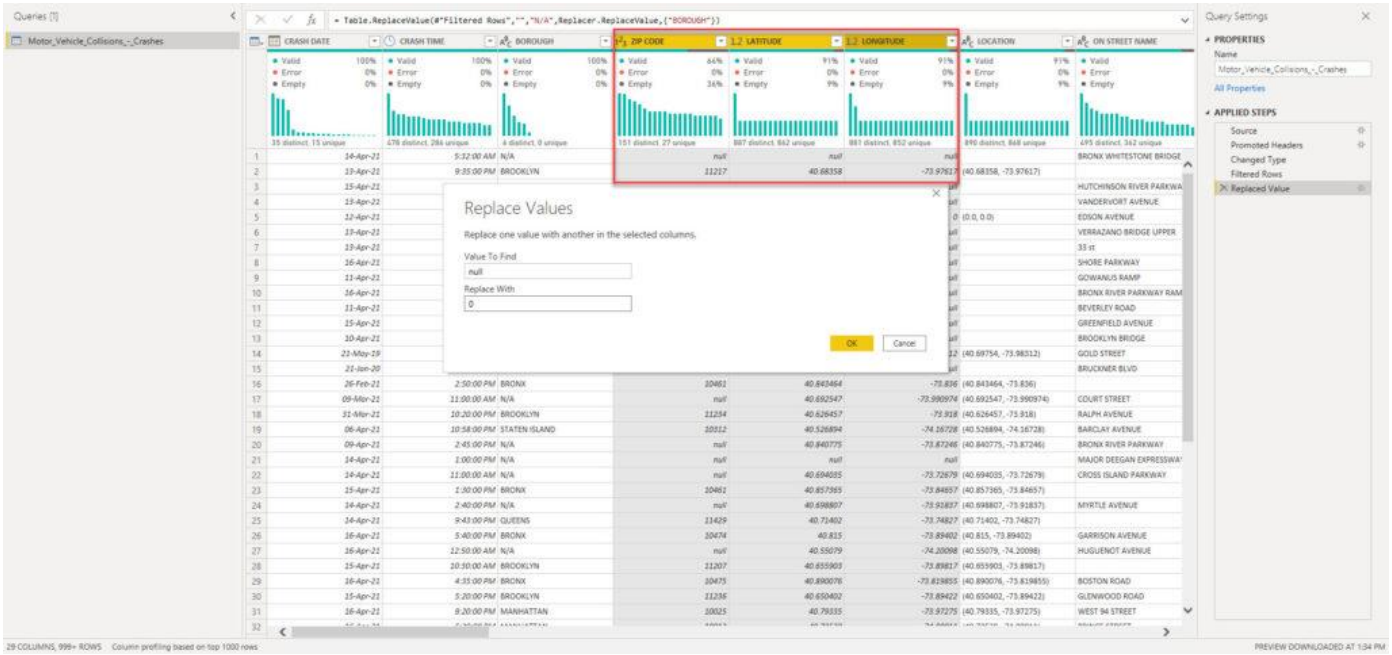
It's time to enhance our dataset and invest some additional effort to improve the data quality. Let's start with replacing blank values with N/A in the *Borough* column:

DATA PREPARATION



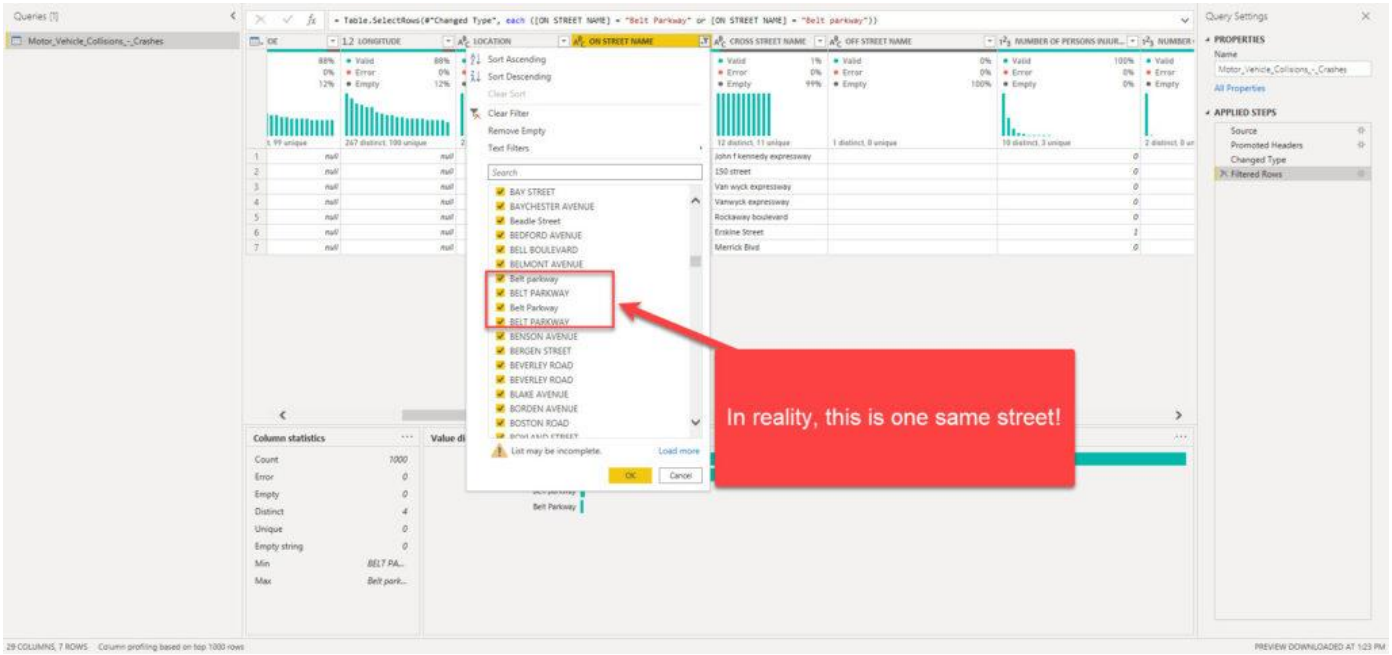
The next step will be to clean the numeric columns. ZIP Code is the whole number column, while Latitude and Longitude are represented as decimal values. That being said, we will replace nulls with 0 value in each of these columns:

DATA PREPARATION



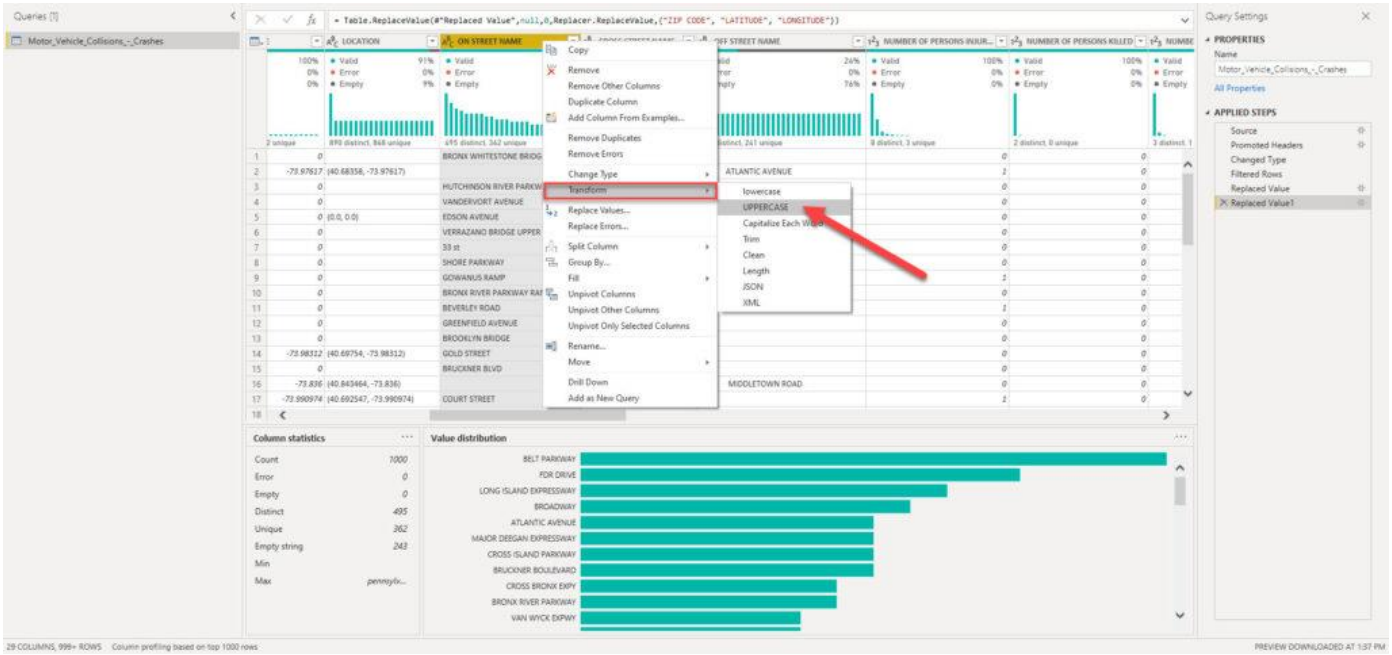
That was quick and easy, right? Now, let's move on and try to profile other columns and check if some more sophisticated transformations are needed. Column *On Street Name* is extremely important, because it's needed to answer one of the crucial business questions: what are the riskiest locations in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

DATA PREPARATION



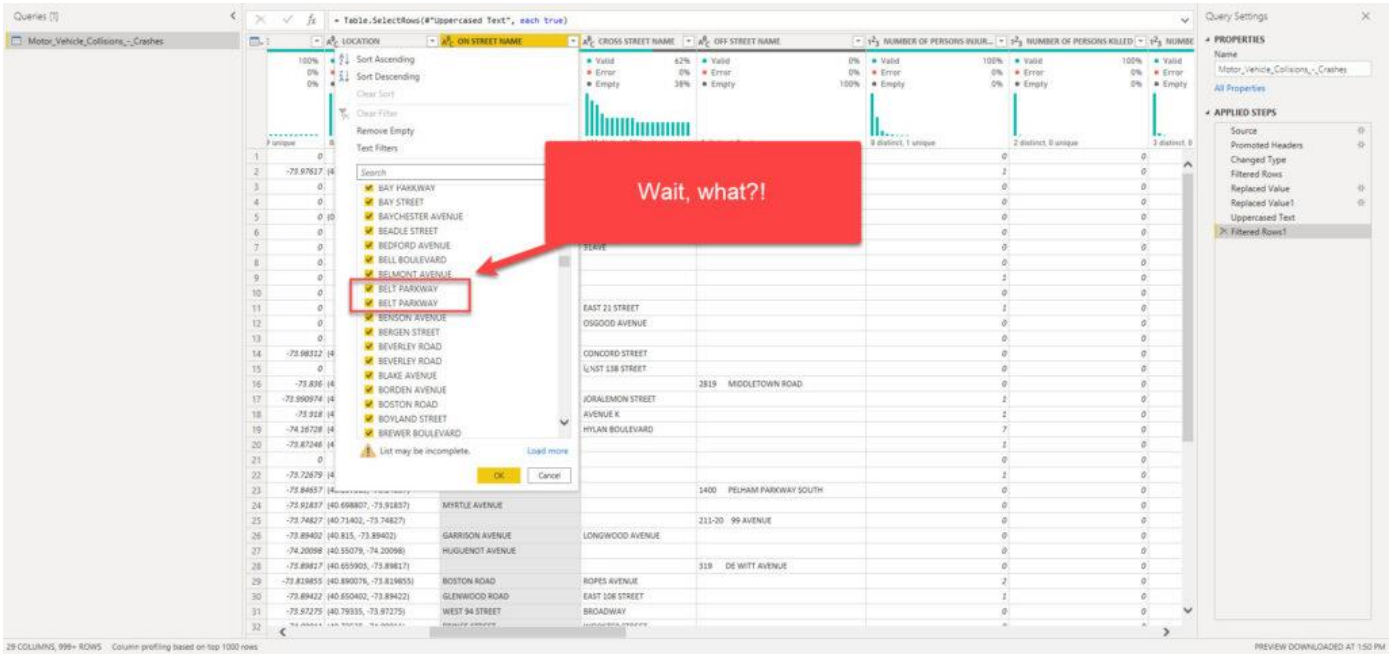
Wait, what?! Belt Parkway is the same as Belt parkway, right? Well, in reality – YES! But, in Power Query M language, case sensitivity will make these two as completely different entities! So, we need to conform the values to be able to get correct results in our reports:

DATA PREPARATION



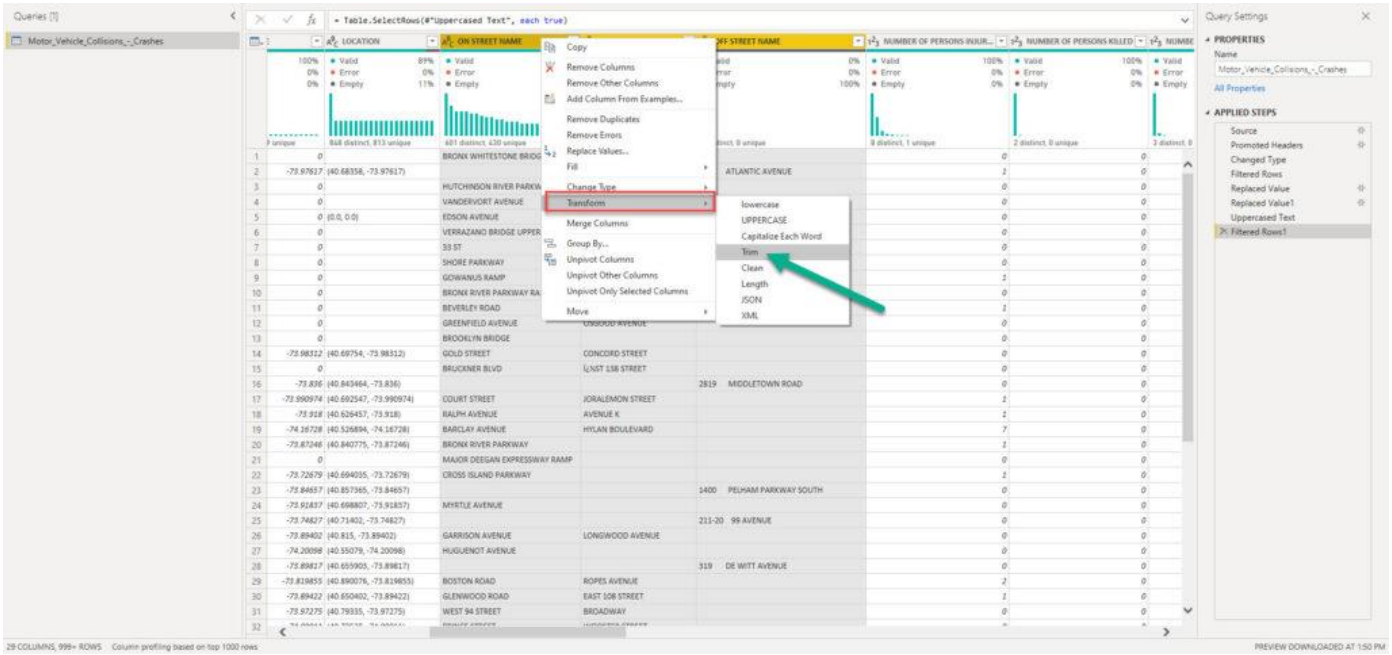
As you can see, I will apply Uppercase transformation to all the columns containing street names, and now we should be good to go:

DATA PREPARATION



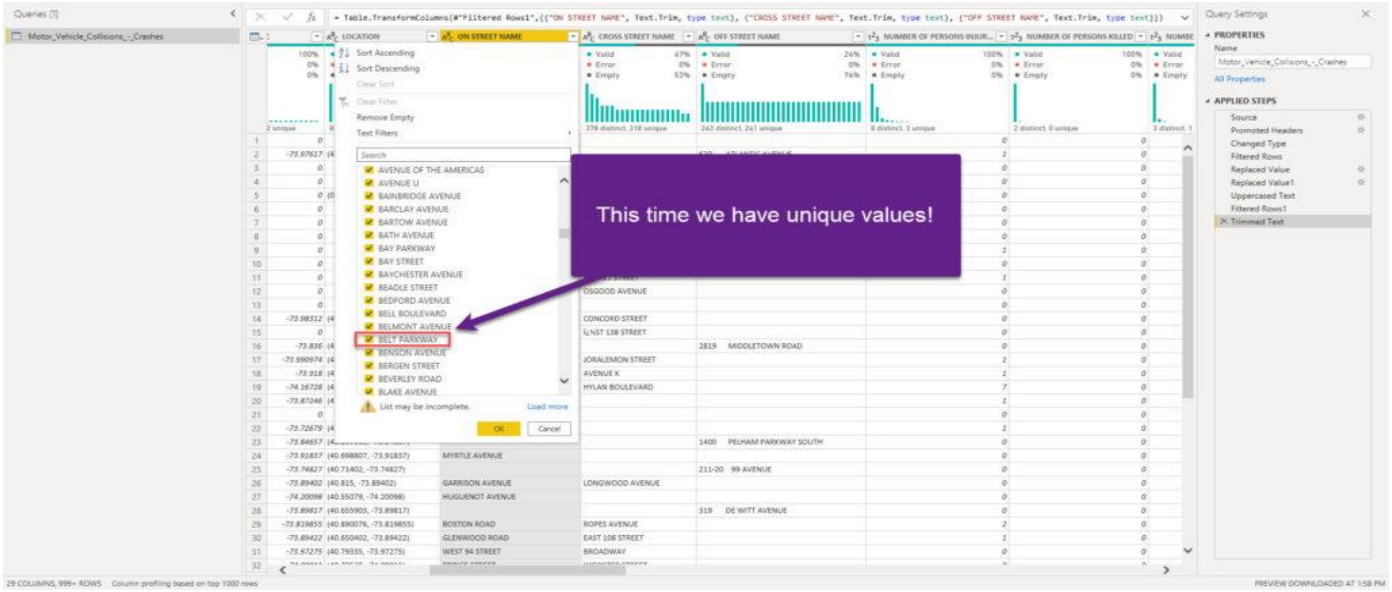
Why do we have two exactly the same uppercased values for BELT PARKWAY? Well, the original CSV file sometimes can contain hidden characters, such as tabulator, new line, or space. Don't worry, I have good news for you: Power Query enables you to solve this specific issue with one click!

DATA PREPARATION



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:

DATA PREPARATION



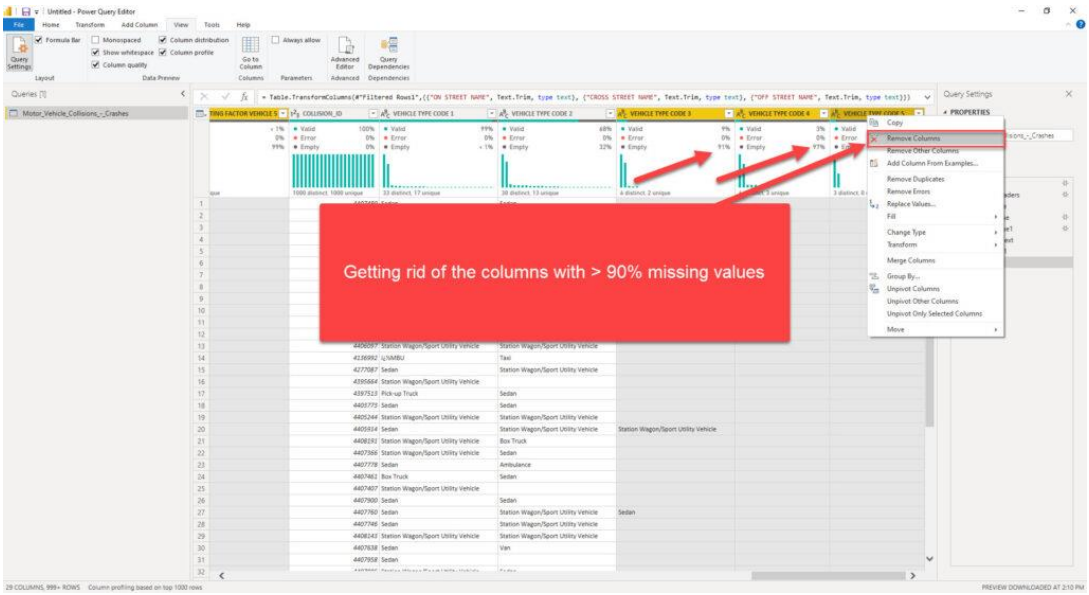
Finally, our column looks as expected: we have unique values!

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 best practices regarding data model optimization, 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?!

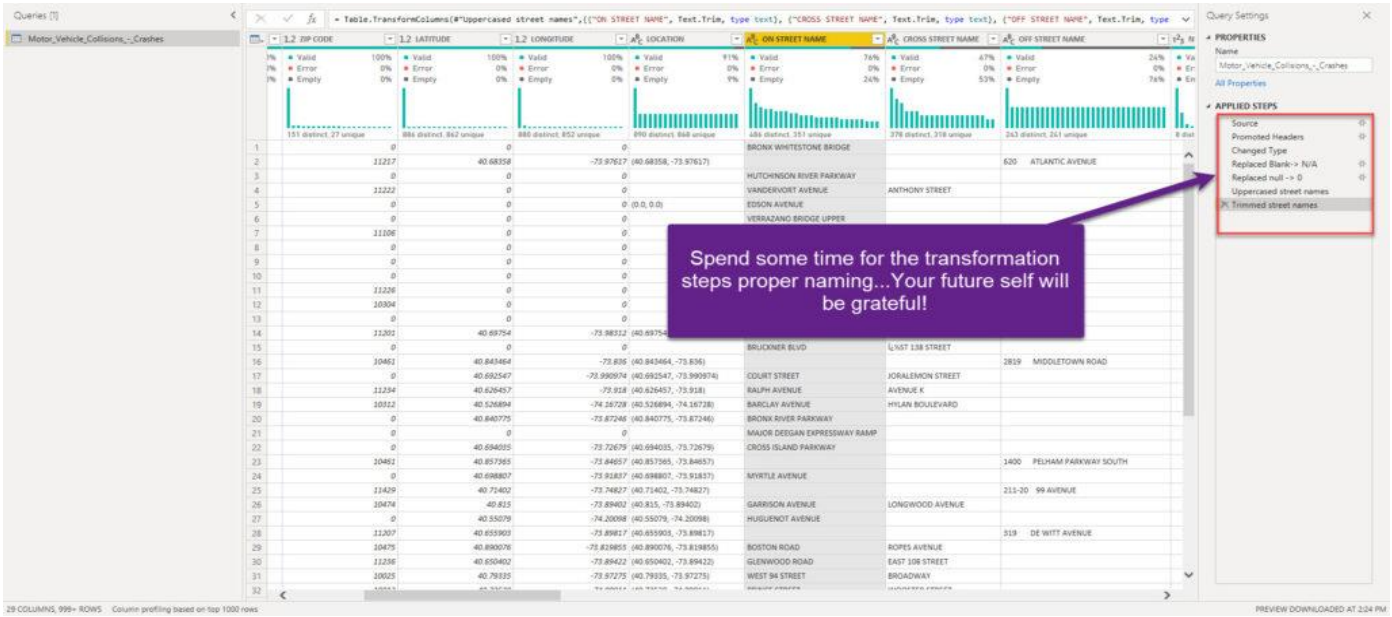
DATA PREPARATION



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 best practices when working with Power Query Editor.

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:

DATA PREPARATION



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!

DATA PREPARATION

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.

DATA MODELING

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 here.

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…

DATA MODELING

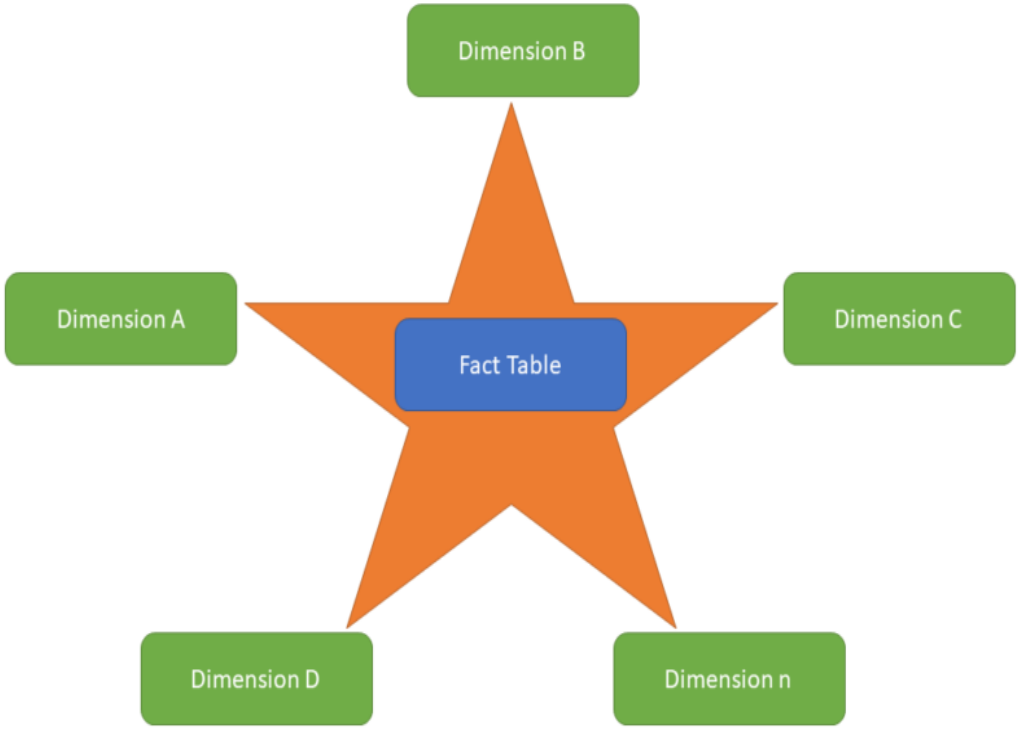
It's important to keep in mind that this concept is not exclusively related to Power BI – it's a general concept that's being used for decades in various data solutions!

If you're serious about working in the data field (not necessarily Power BI), I strongly recommend reading the book: **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling** 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.

DATA MODELING

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

DATA MODELING

In the illustration above, you can see a typical example of dimensional modeling, called Star-schema. 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 here. 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.

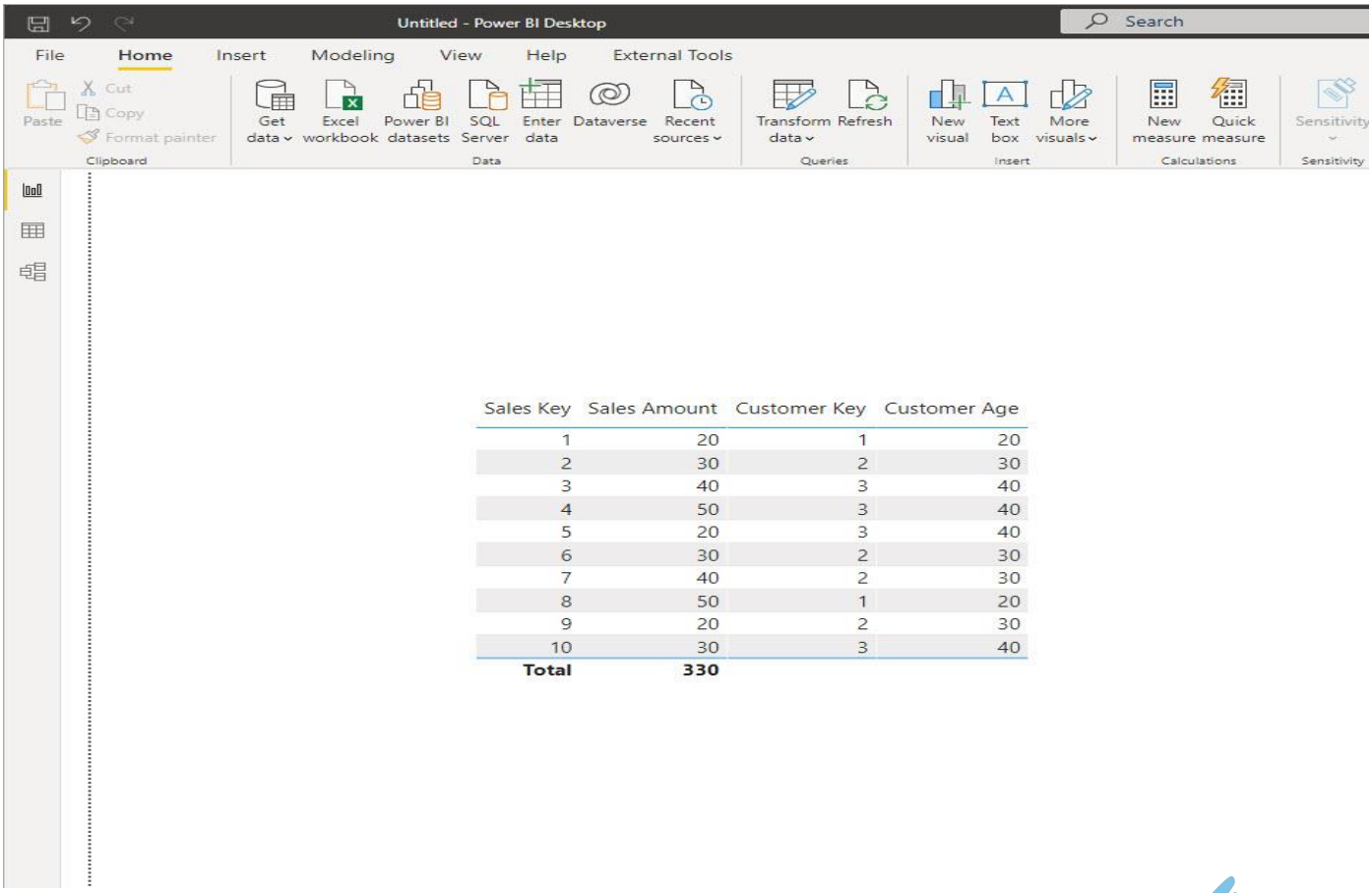
And, then, *Amir Netz*, CTO of Microsoft Analytics and one of the people responsible for building a VertiPaq engine, cleared all the uncertainties on Twitter:

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 this video from Patrick (Guy in a Cube), or this one from Alberto Ferrari (SQL BI).

DATA MODELING

And, it's not just about efficiency – it's also about getting accurate results in your reports! In this article, 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!

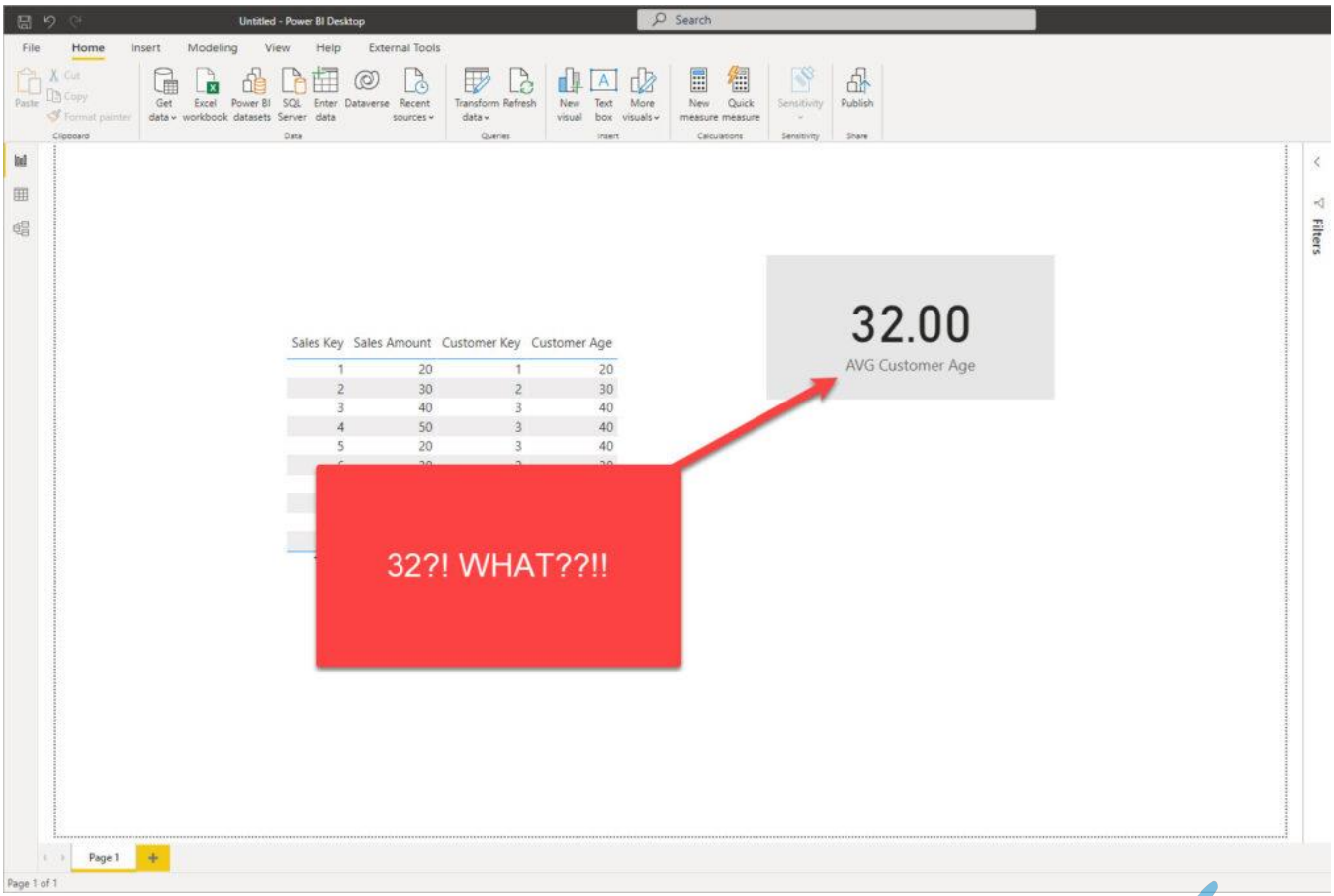


Sales Key	Sales Amount	Customer Key	Customer Age
1	20	1	20
2	30	2	30
3	40	3	40
4	50	3	40
5	20	3	40
6	30	2	30
7	40	2	30
8	50	1	20
9	20	2	30
10	30	3	40
Total	**330**		

DATA MODELING

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…

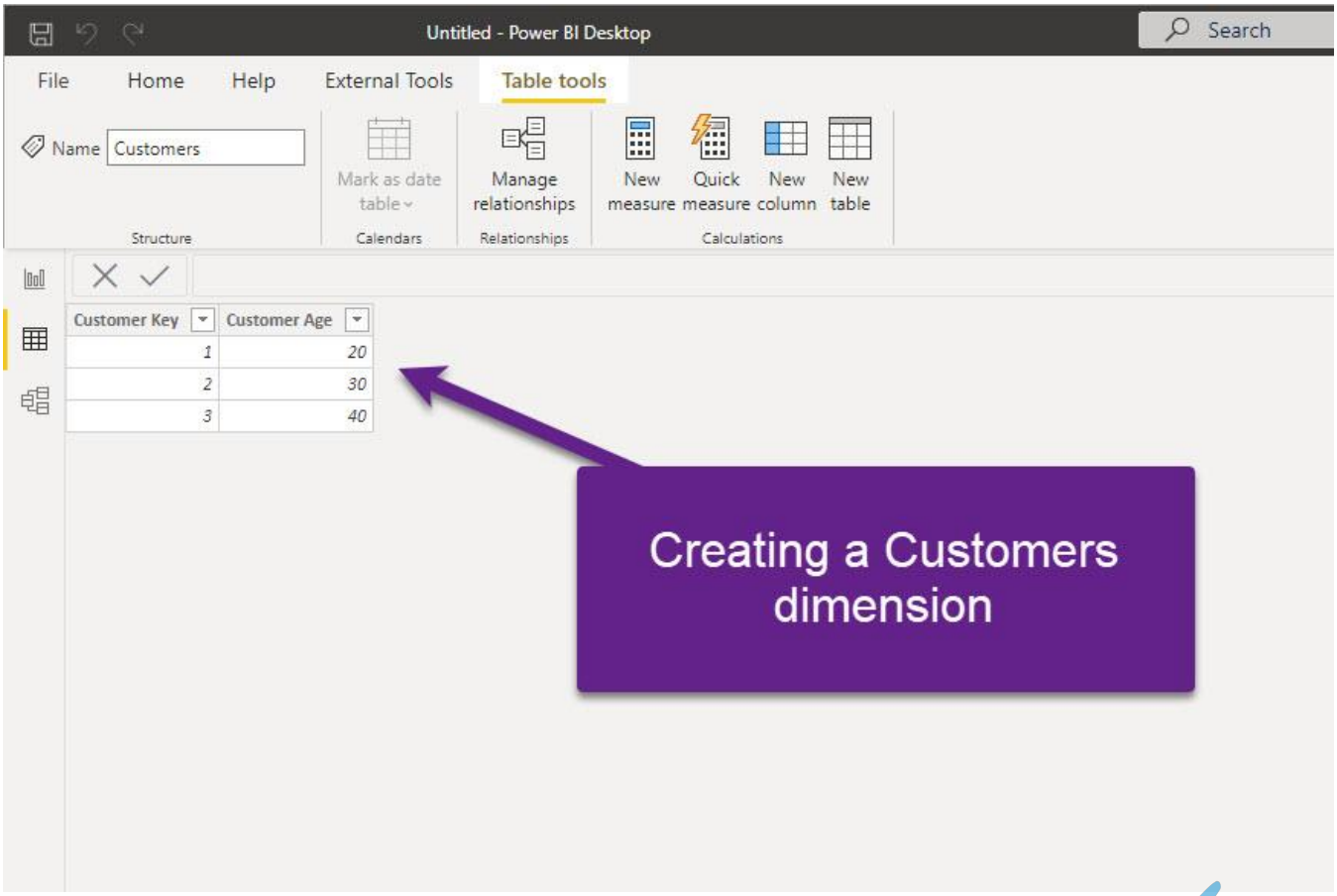
AVG Customer Age = AVERAGE(Table1[Customer Age])



DATA MODELING

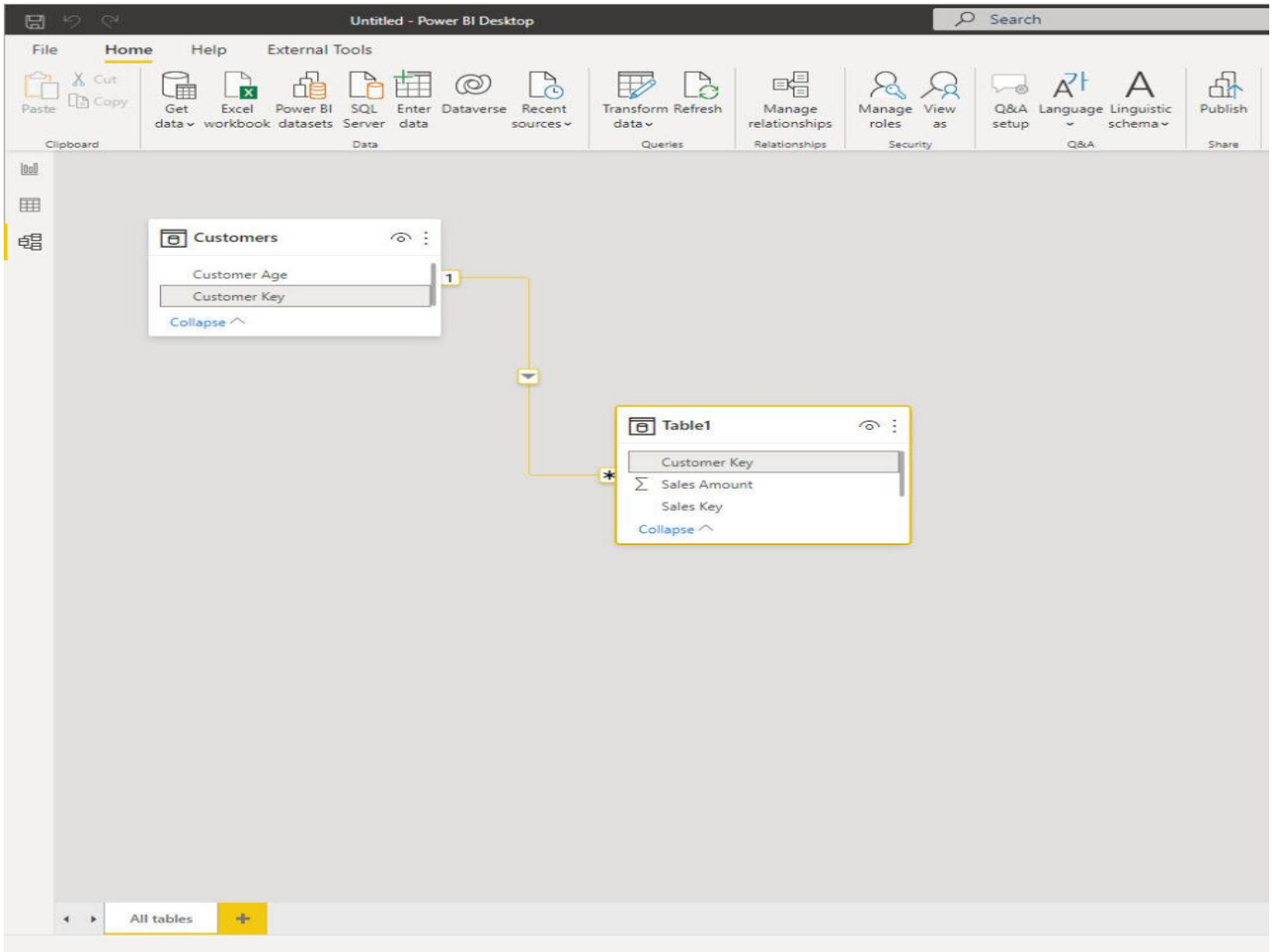
How the hell is this possible?! 32, really?! Let's see how we got this unexpected (incorrect) number...If we sum all the Customer Age values, we will get 320...320 divided by 10 (that's the number of sales), and voila! There you go, that's your 32 average customers' age!

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:



DATA MODELING

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

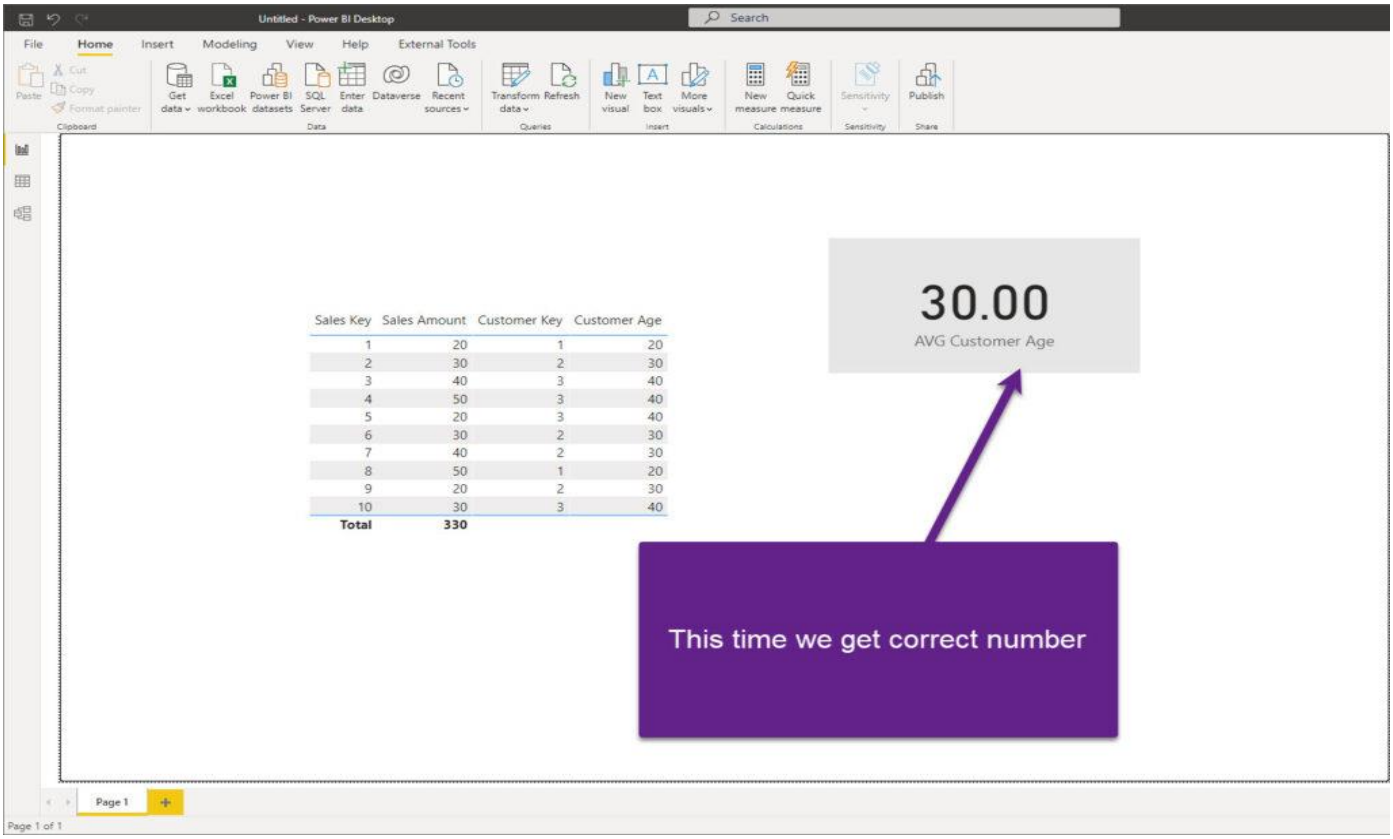


Finally, I just need to rewrite my measure to refer to a newly created dimension table:

DATA MODELING

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:



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.

DATA MODELING

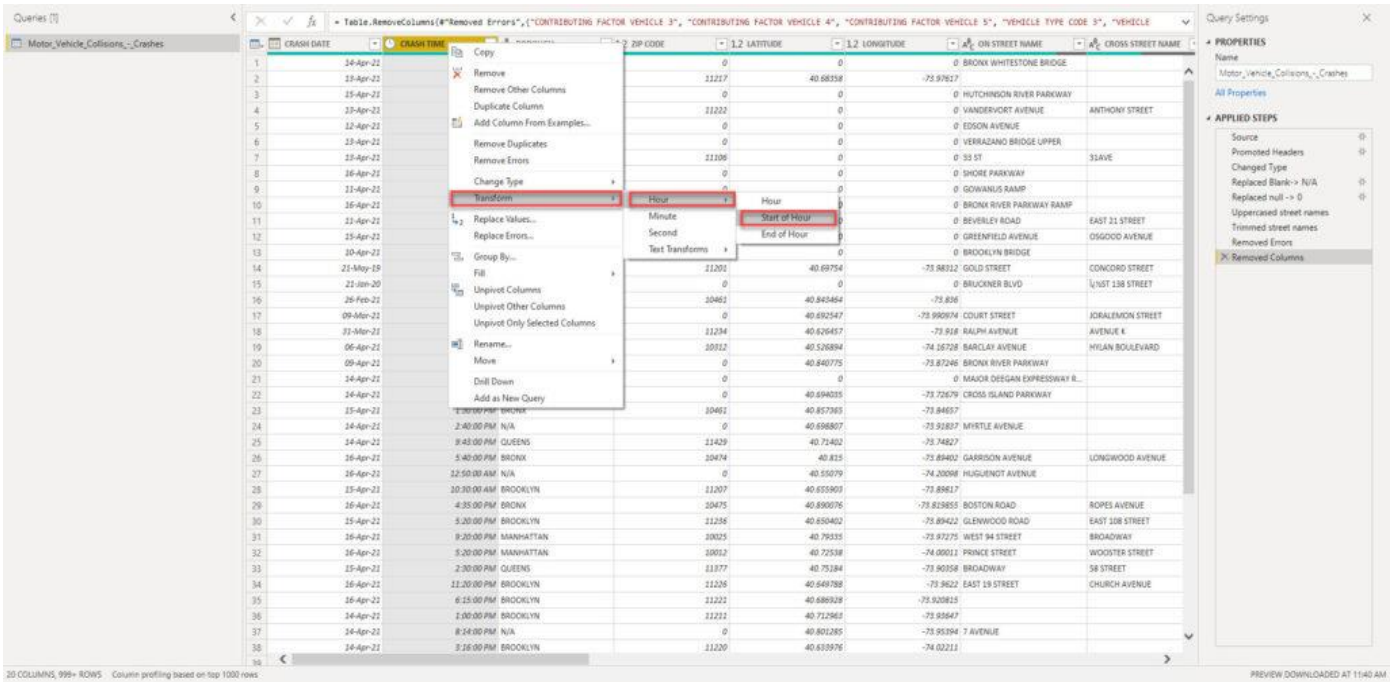
As we concluded that the Star schema is the way to go, let's start building the optimal data model for our dataset. The first step is to get rid of the columns with >90% missing values, as we can't extract any insight from them. I've removed 9 columns and now I have 20 remaining.

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

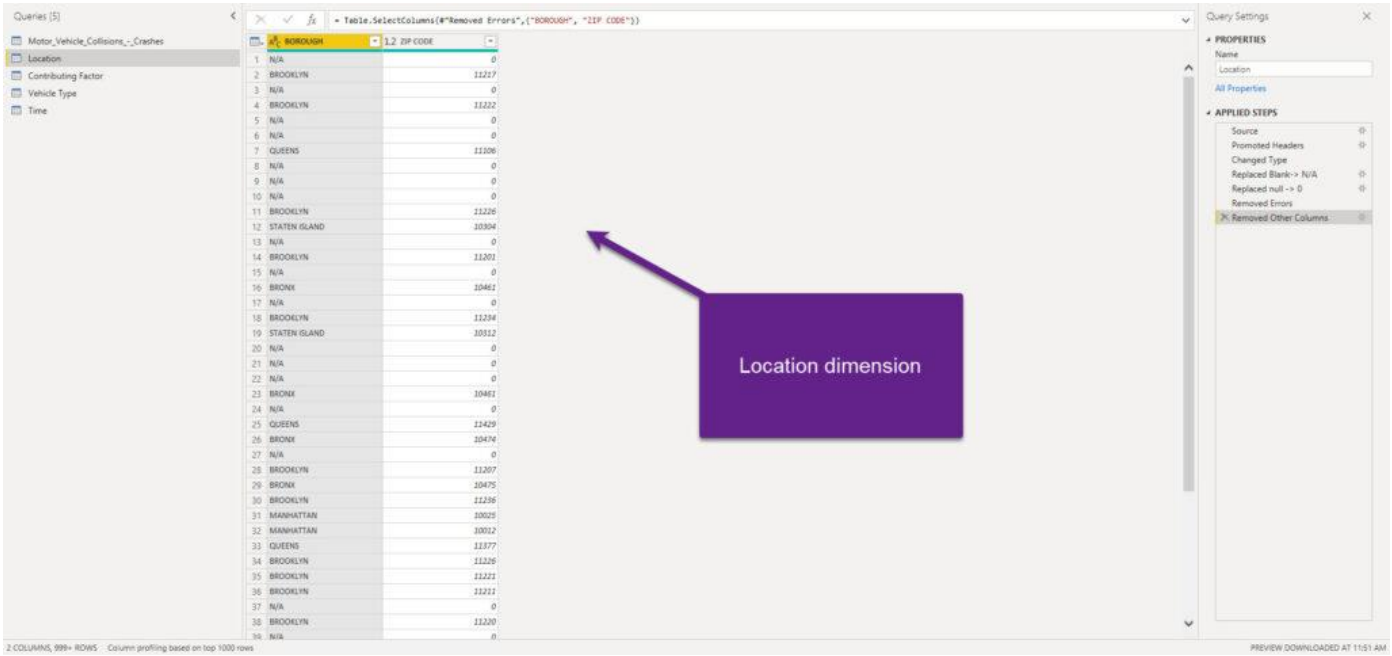
But, before we proceed to create them, I want to apply one additional transformation to my Crash Time column. As we don't need to analyze data on a minute level (hour level of granularity is the requirement), I'll round the values to a starting hour:

DATA MODELING



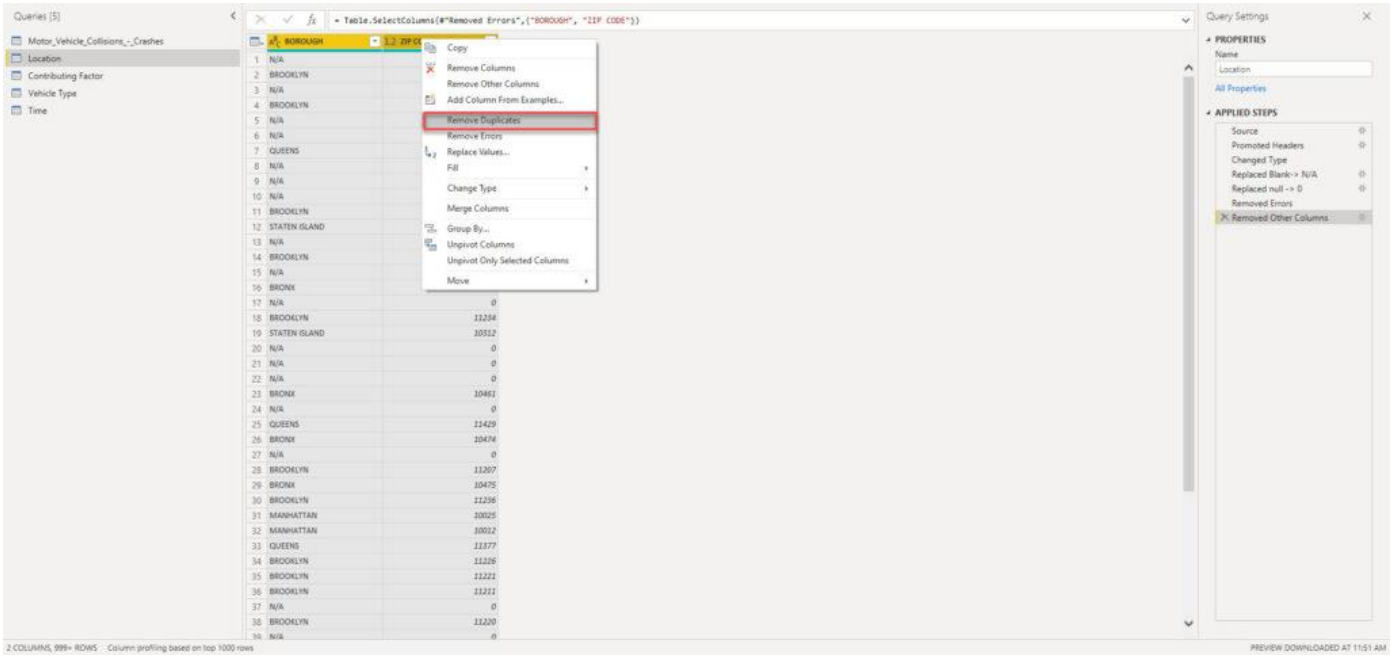
I'll now duplicate my original flat table 4 times (for each of the dimensions needed, except for the Date dimension, as I want to use a more sophisticated set of attributes, such as day of the week for example). Don't worry, as we will keep only relevant columns in each of our dimensions and simply remove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

DATA MODELING



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:

DATA MODELING

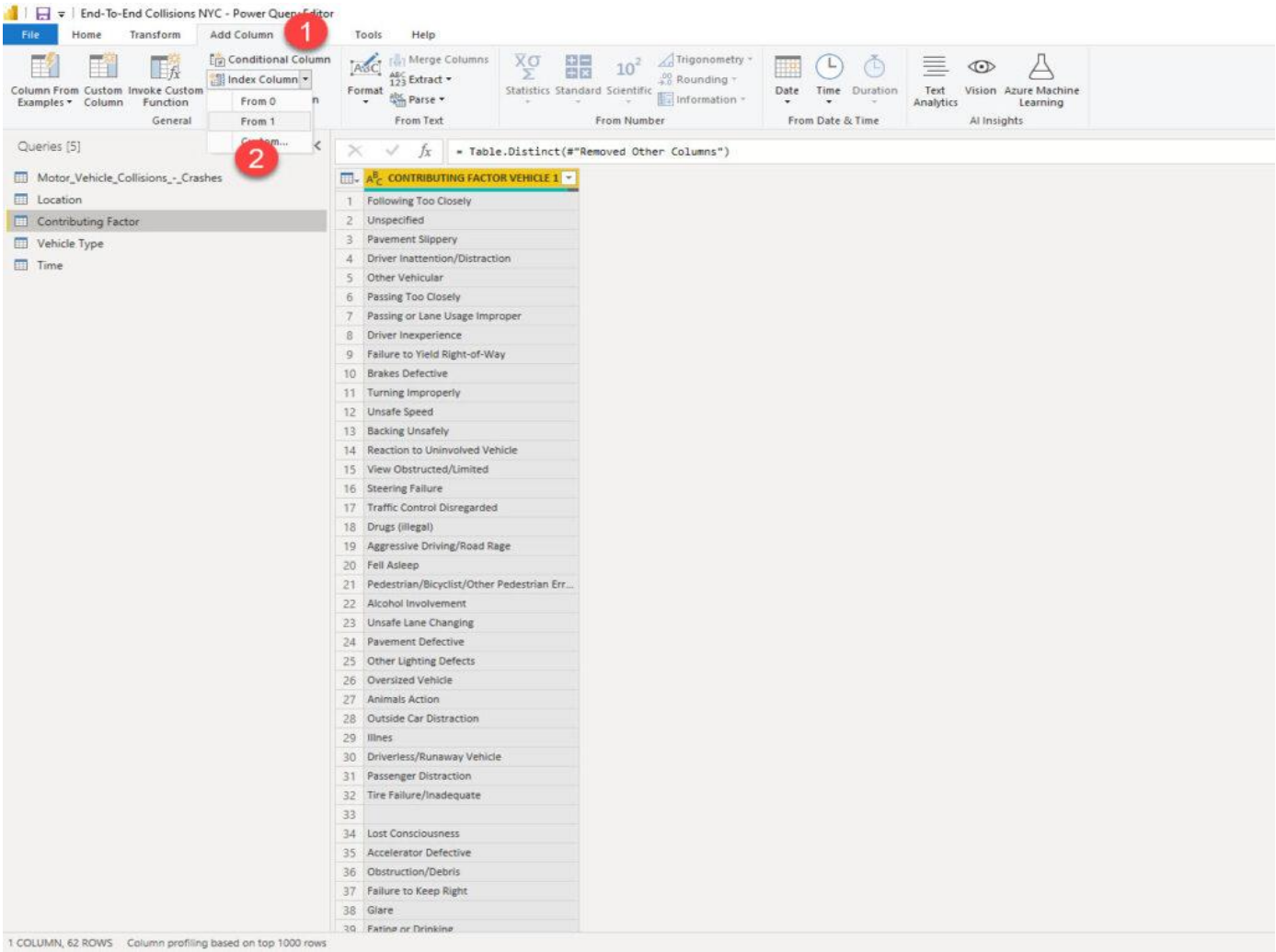


We need to do this for every single dimension in our data model! From here, as we don't have "classic" key columns in our original table (like, for example, in the previous case when we were calculating average customers' age and we had Customer Key column in the original flat table), there are two possible 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.

DATA MODELING

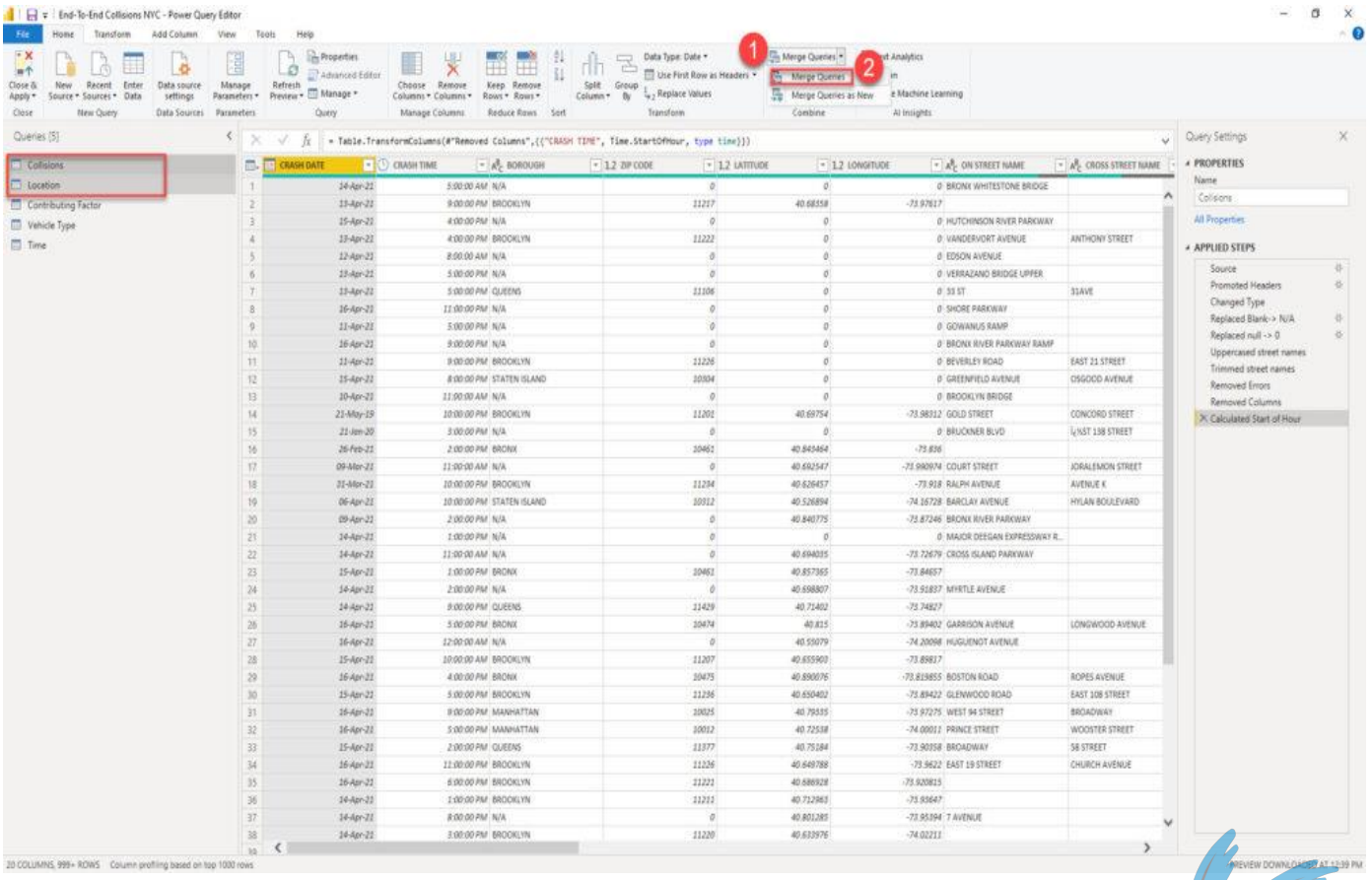
Creating a surrogate key in Power Query is quite straightforward using Index column transformation.



DATA MODELING

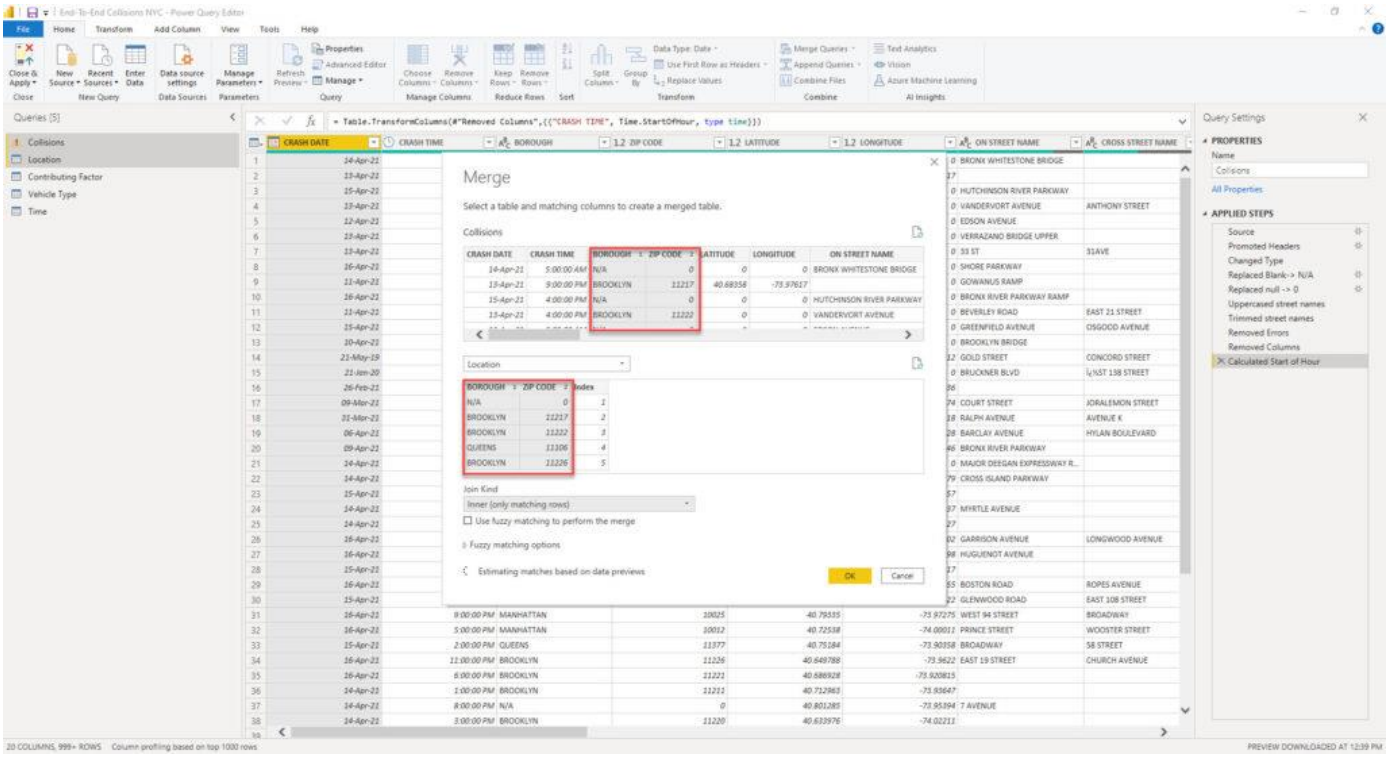
Just one remark here: by default, using an Index column transformation will break query folding. However, as we are dealing with CSV file, which doesn't support query folding at all, we can safely apply Index column transformation.

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:



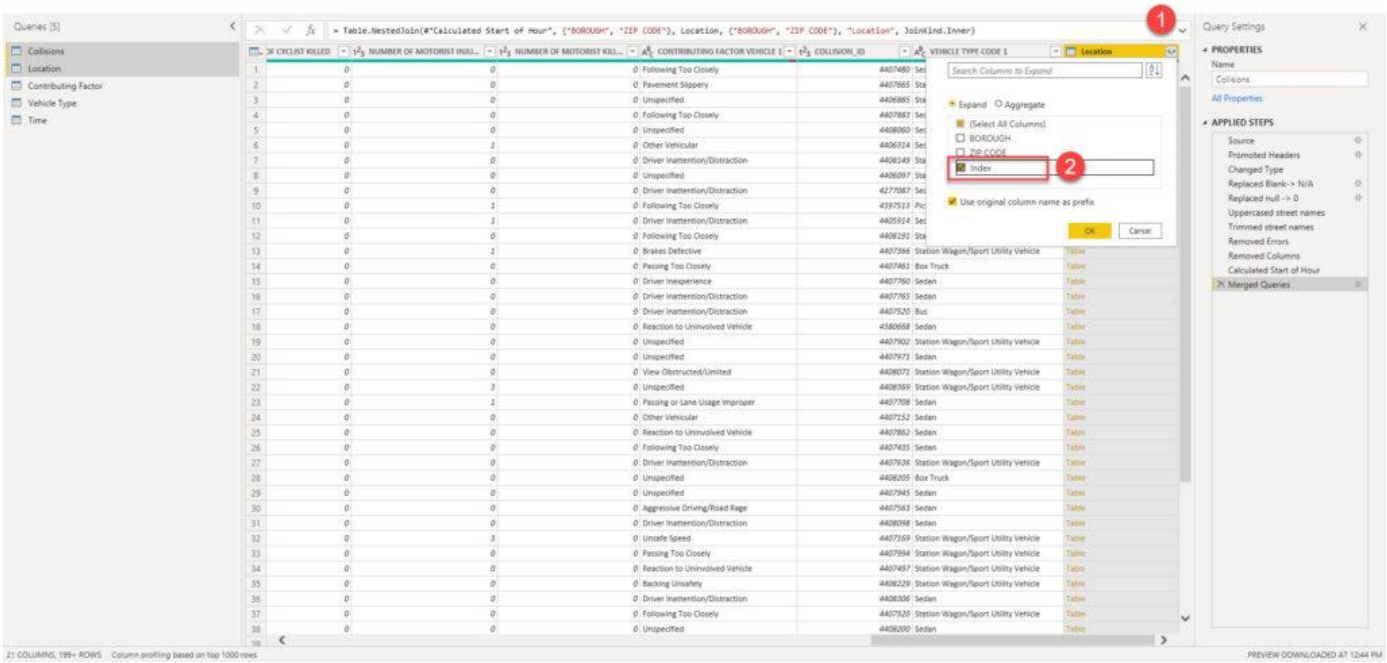
DATA MODELING

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):



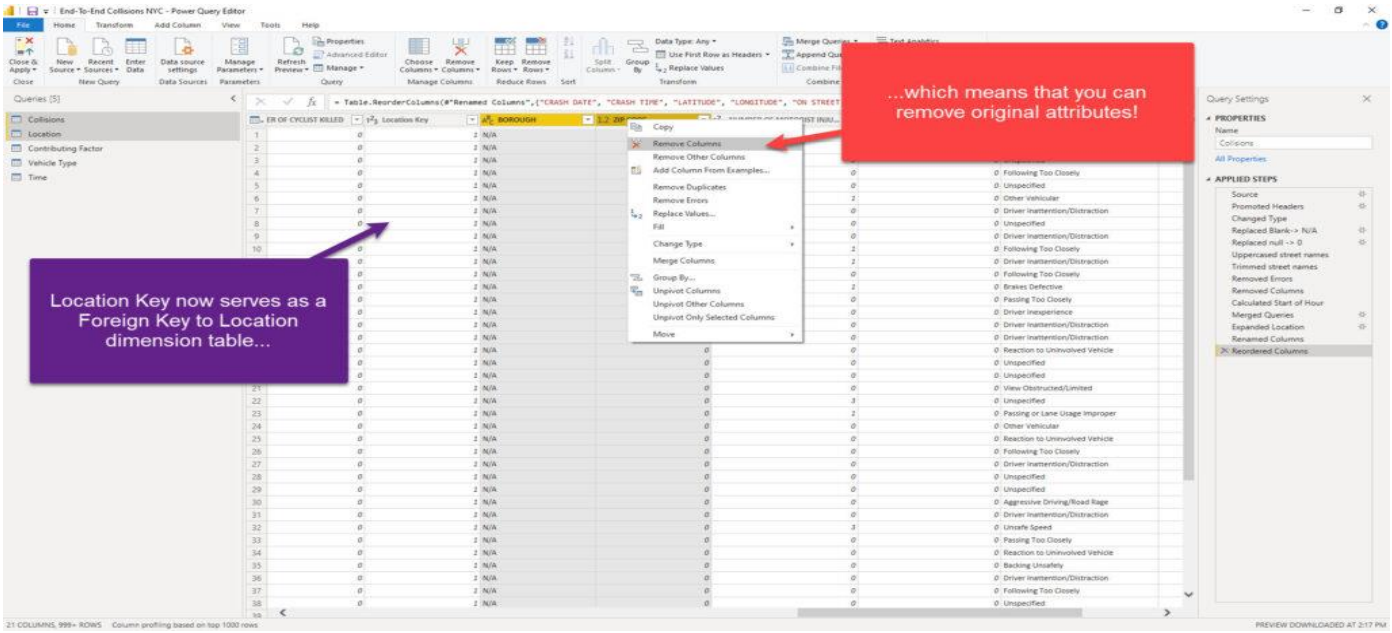
And after Power Query applies this transformation, I will be able to expand the merged Location table and take the Index column from there:

DATA MODELING



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!

DATA MODELING



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 set a proper Date/Calendar dimension, in order to enable DAX Time Intelligence functions to work in a proper way.

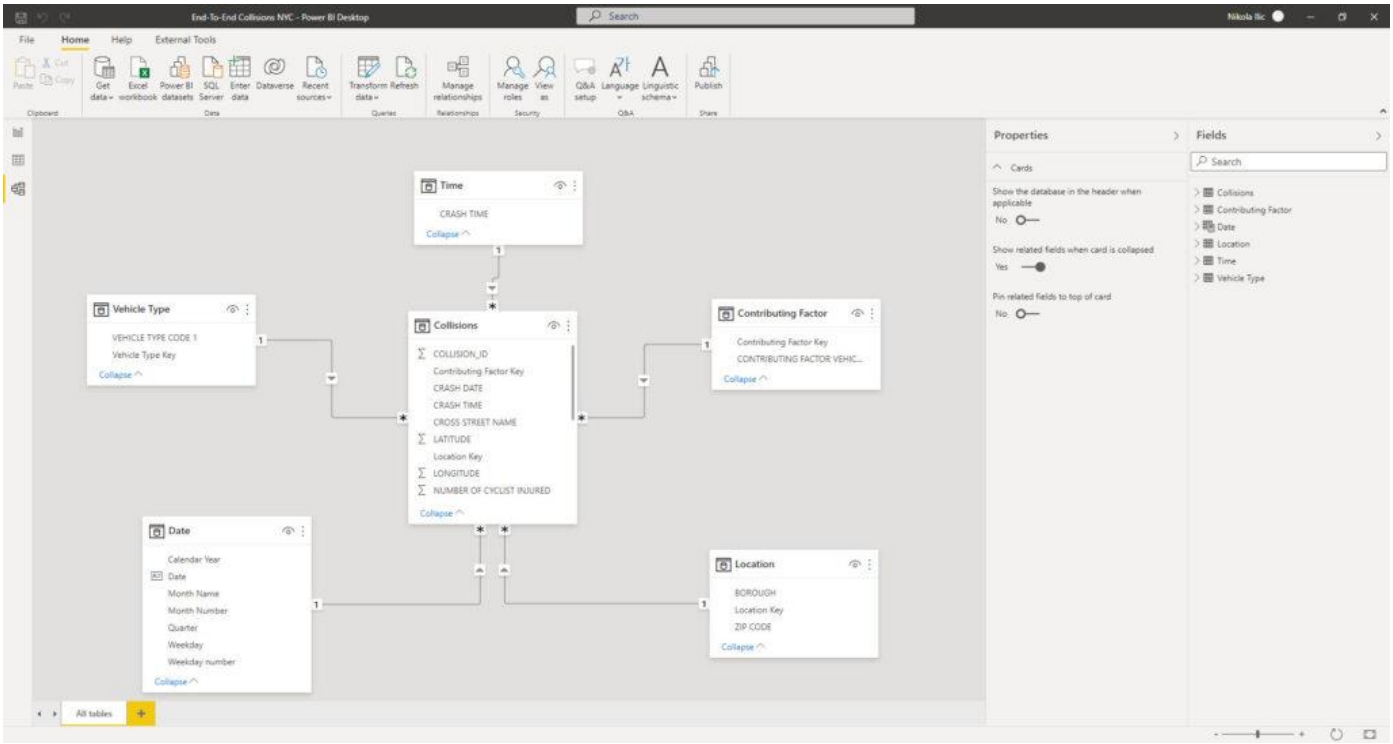
DATA MODELING

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

```
Date =
        VAR MinYear = YEAR ( MIN ( Collisions[CRASH DATE] ) )
        VAR MaxYear = YEAR ( MAX ( Collisions[CRASH DATE] ) )
RETURN
ADDCOLUMNS (
        FILTER (
                CALENDARAUTO( ),
                AND ( YEAR ( [Date] ) >= MinYear, YEAR ( [Date] ) <= MaxYear )
        ),
        "Calendar Year", "CY " & YEAR ( [Date] ),
        "Month Name", FORMAT ( [Date], "mmmm" ),
        "Month Number", MONTH ( [Date] ),
        "Weekday", FORMAT ( [Date], "dddd" ),
        "Weekday number", WEEKDAY( [Date] ),
        "Quarter", "Q" & TRUNC ( ( MONTH ( [Date] ) - 1 ) / 3 ) + 1
)
```

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

DATA MODELING



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:)

DATA MODELING

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.

DATA VISUALIZATION

Introduction

Good news, folks – slowly, but steadily, we are nearing our goal – to build an efficient end-to-end 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 (this one, for example), blogs, and inspired by some amazing community members, such as Armand Van Amersfoort, Daniel Marsh-Patrick, Kerry Kolosko, Ried Havens, Andrej Lapajne (Zebra BI), or folks from powerbi.tips.

DATA VISUALIZATION

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

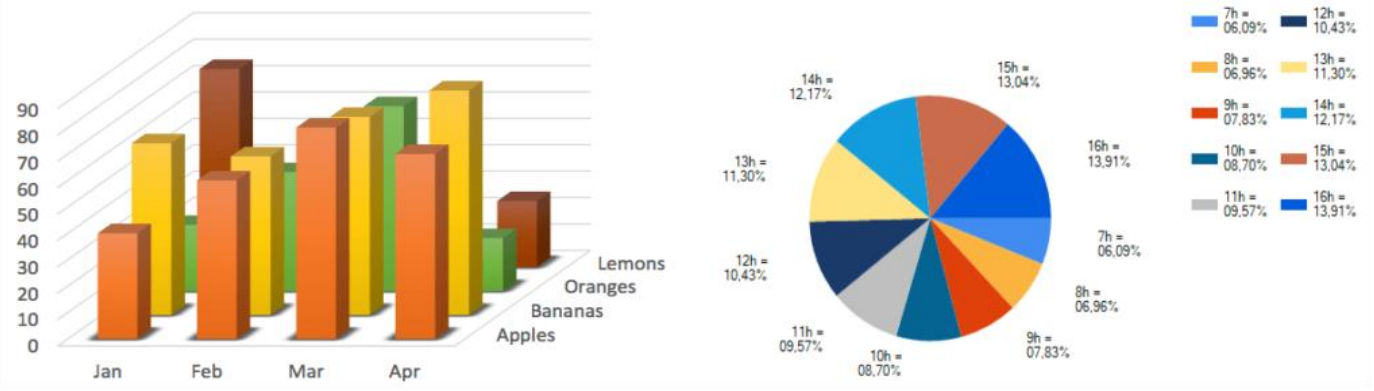
This one applies not to data visualization exclusively, it's more of a general rule. There is no single solution to satisfy each and every business request! Period! As a first step, you should determine the purpose of the dashboard – ***operational dashboards*** provide time-critical data to consumers. 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.

DATA VISUALIZATION

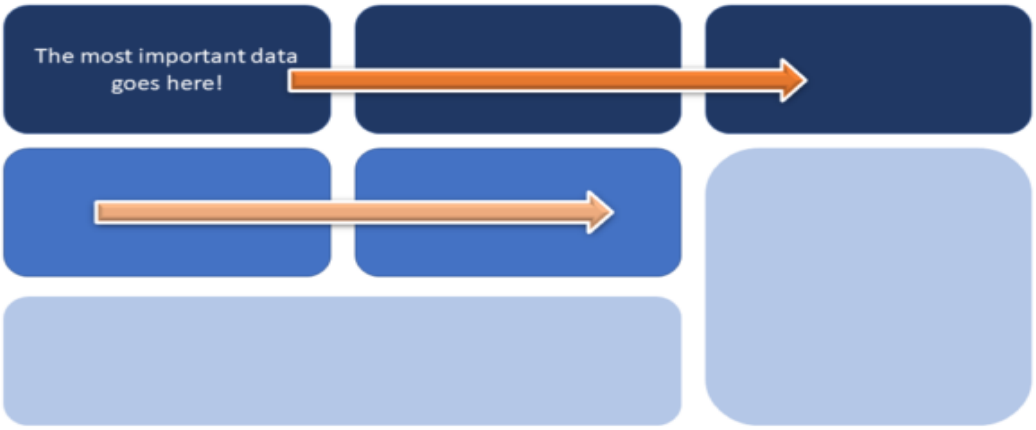


I've intentionally used the word SHOULD, as no one can prohibit you from using, let's say, gauges, pie-charts, 3D-charts in your dashboards, even though a lot of recognized experts advised against that – just be careful and mindful in which scenario to use what visual type.

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.

DATA VISUALIZATION



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.

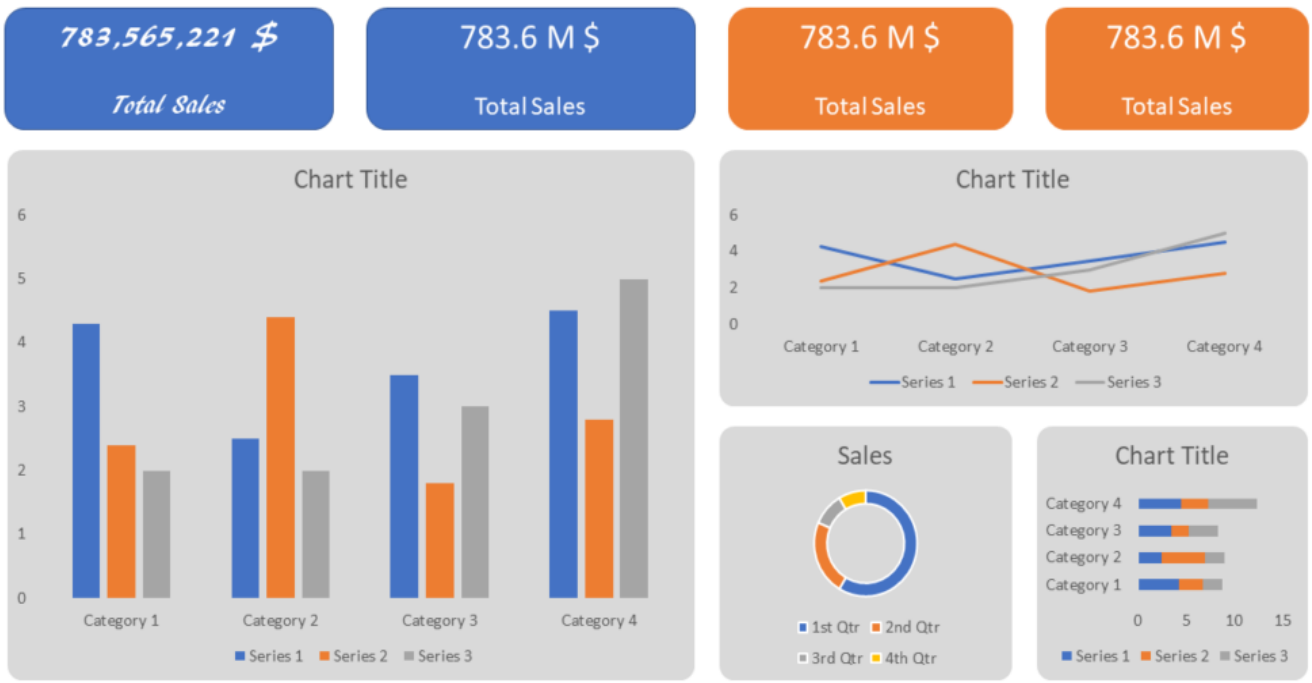
5. Remove the distraction

I've already written about one specific case of removing distractions from the Power BI report. 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.

DATA VISUALIZATION

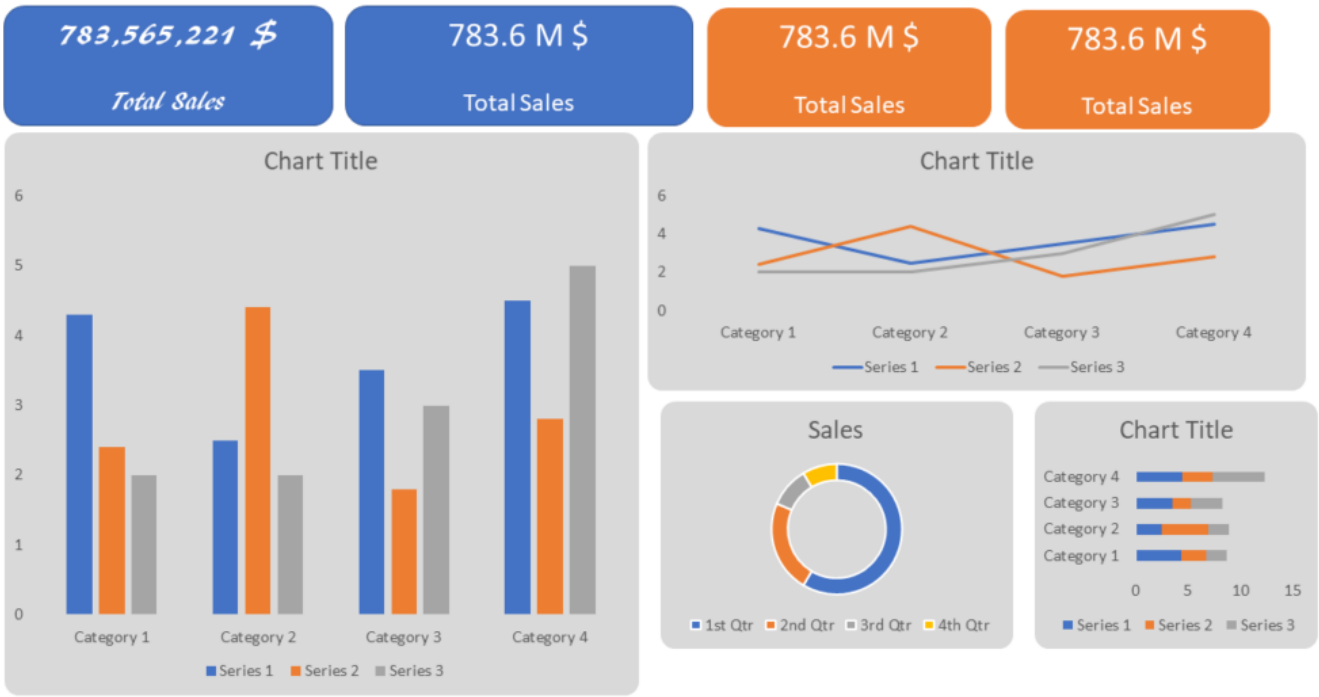
In addition, take care of the proper alignment and give your visuals some space in between:



Proper alignment and space between visuals will improve the clarity

DATA VISUALIZATION

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



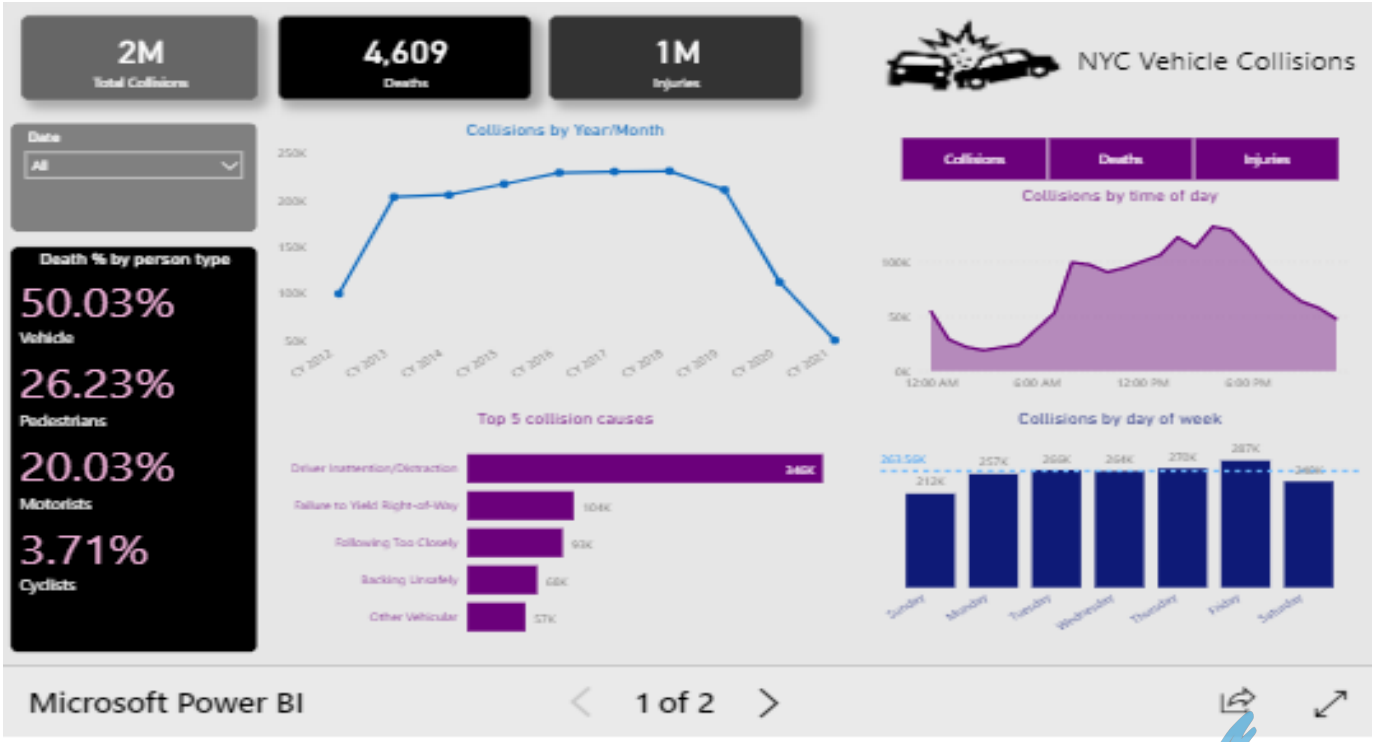
There are many more best practices, tips, and recommendations when it comes to data visualization. As I've already stressed, I don't consider myself a "data viz" wizard, but I'm still trying to stick with some of the general rules mentioned in the previous chapter.

DATA VISUALIZATION

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.



DATA VISUALIZATION

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 here 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.

DATA VISUALIZATION

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.

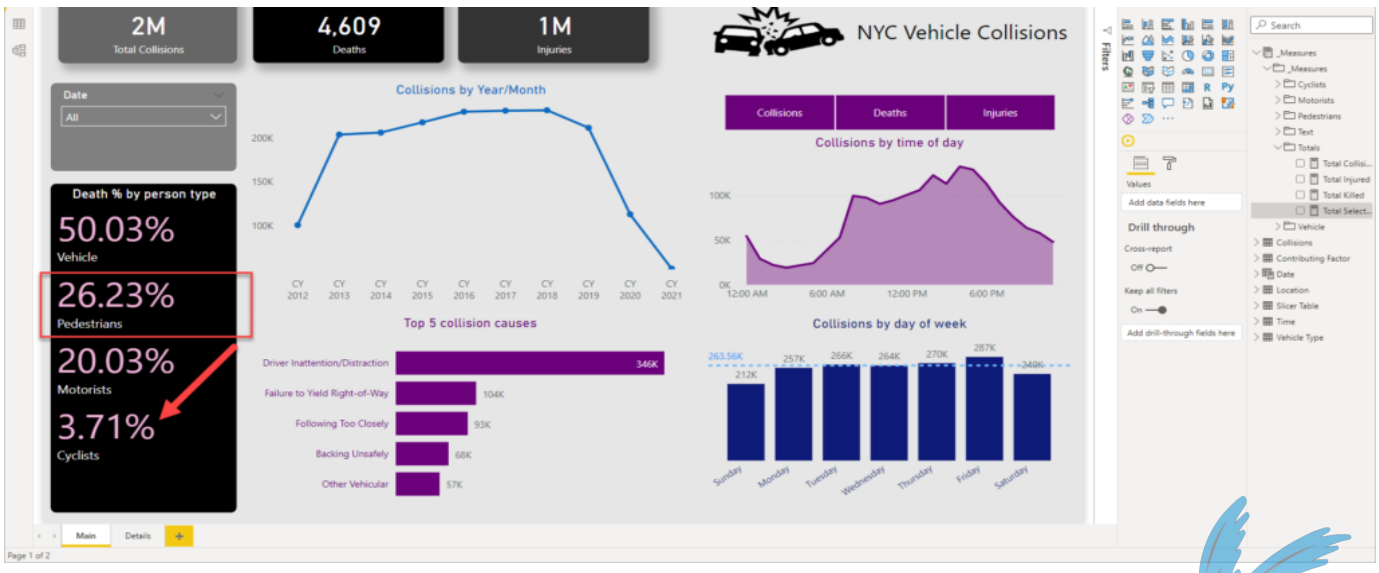
DATA ANALYSIS

Introduction

Here we are – after taking raw data in the form of a CSV file, 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!

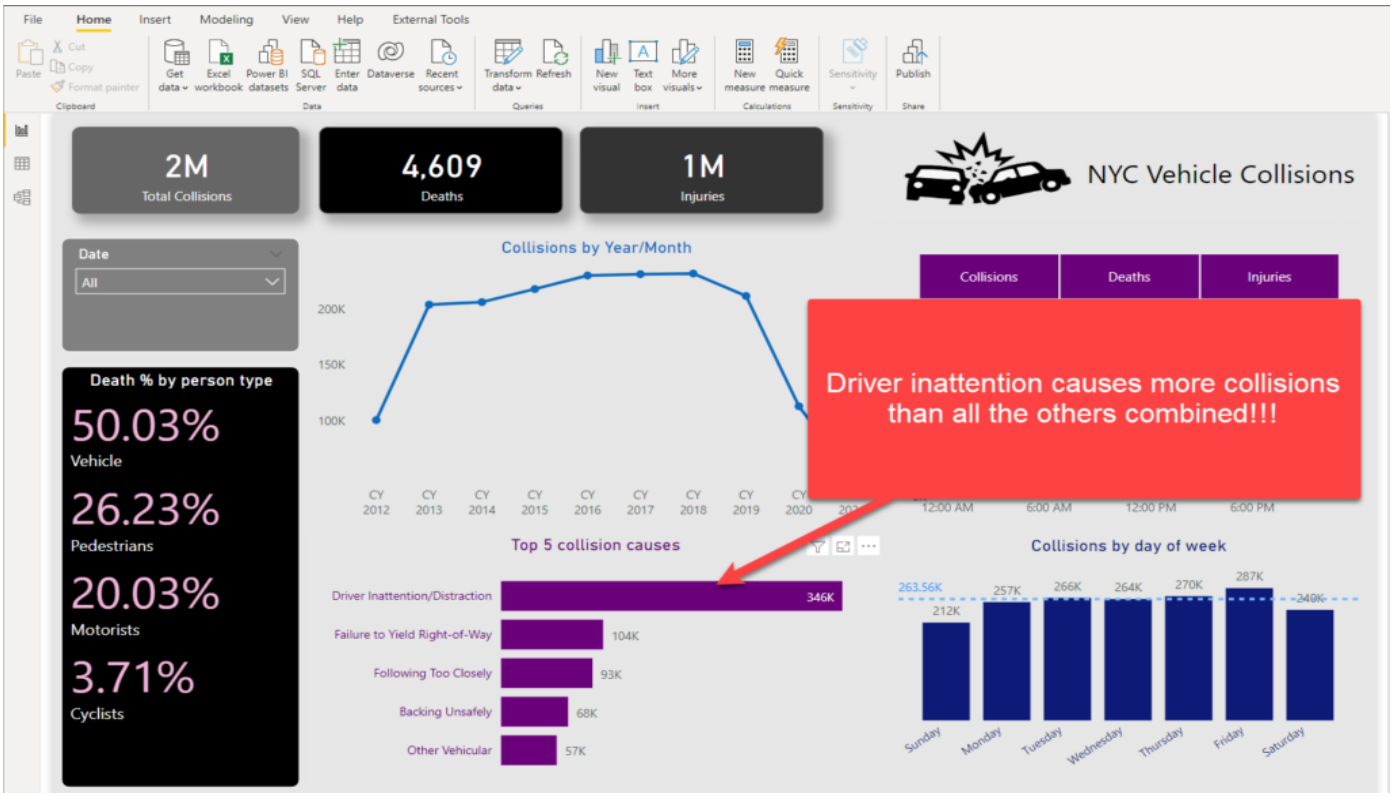
Extracting the insights

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!



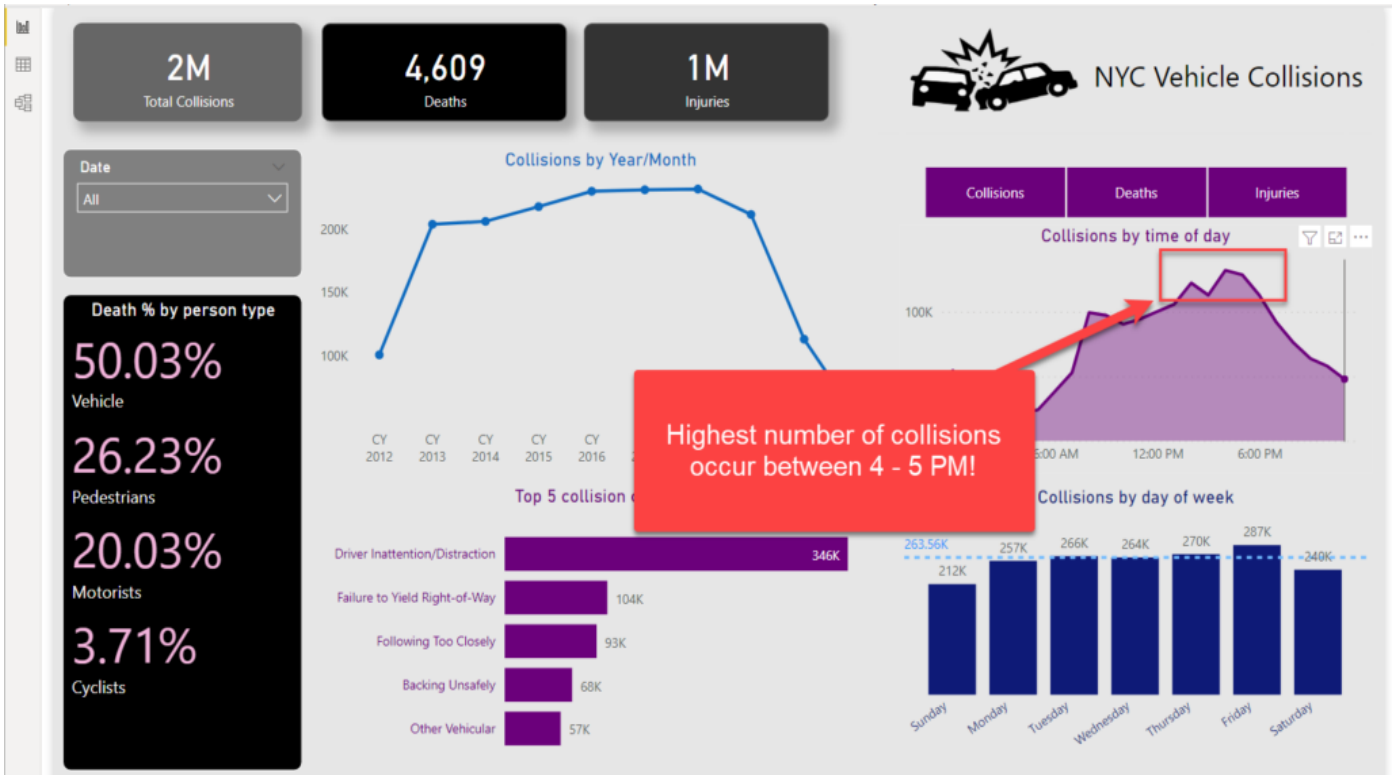
DATA ANALYSIS

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:

DATA ANALYSIS

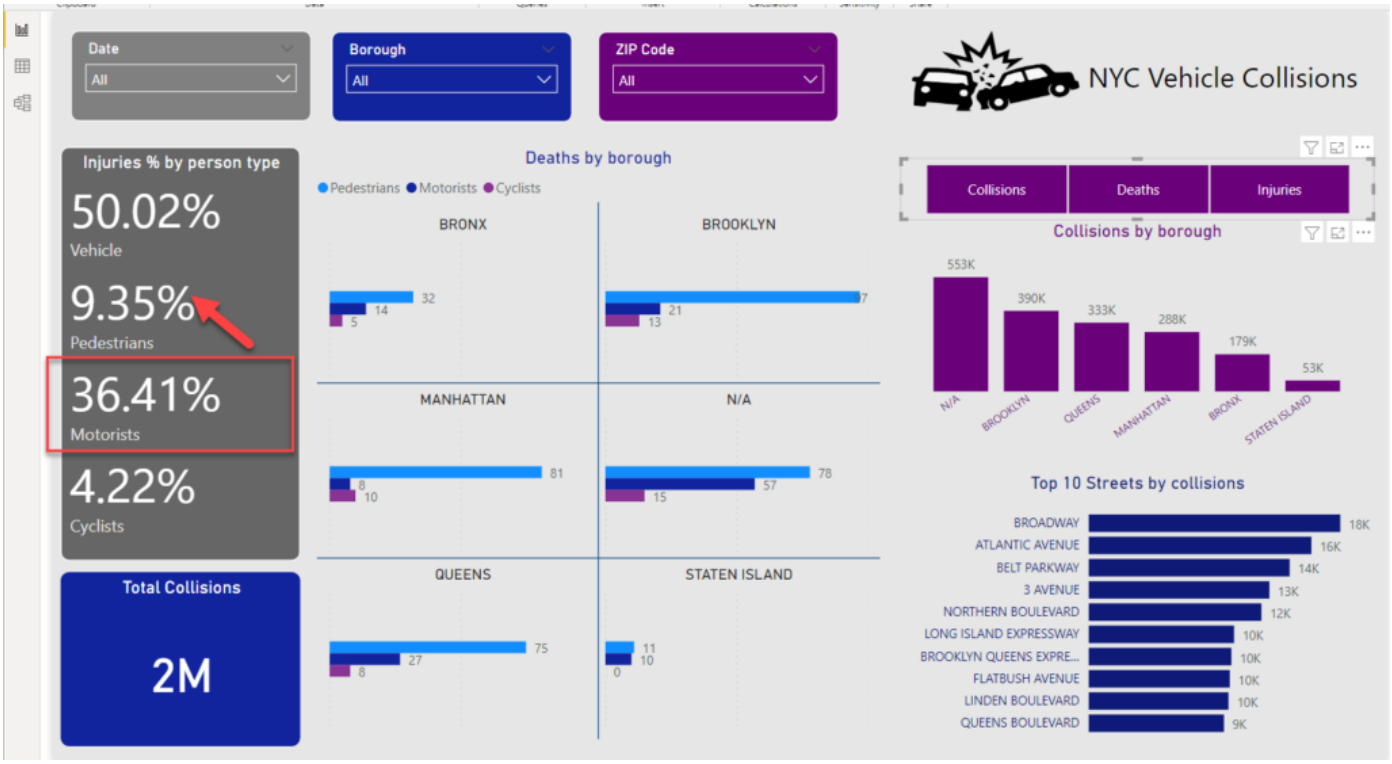


It's approximately 30% higher than in the morning (8-9 AM) when traffic participants are probably not tired and distracted after a hard-working day.

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.

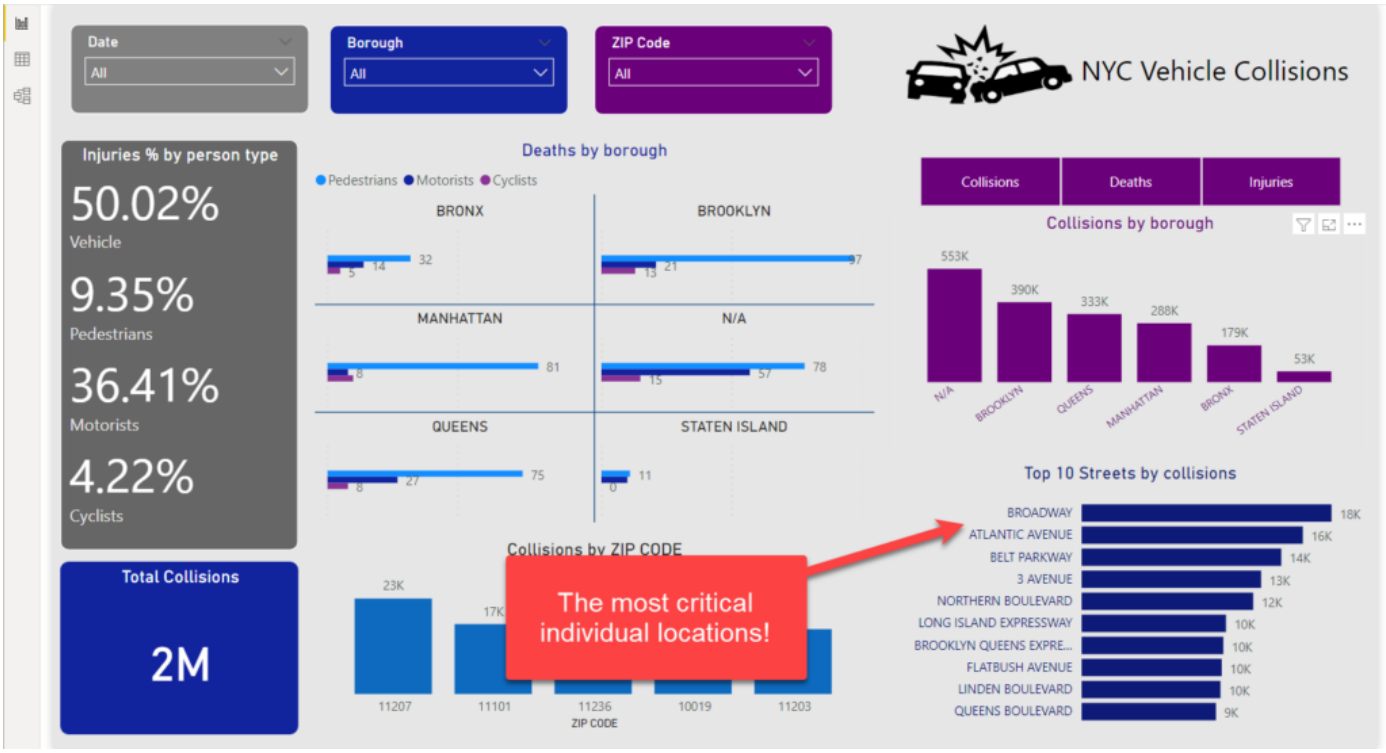
DATA ANALYSIS

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!

DATA ANALYSIS

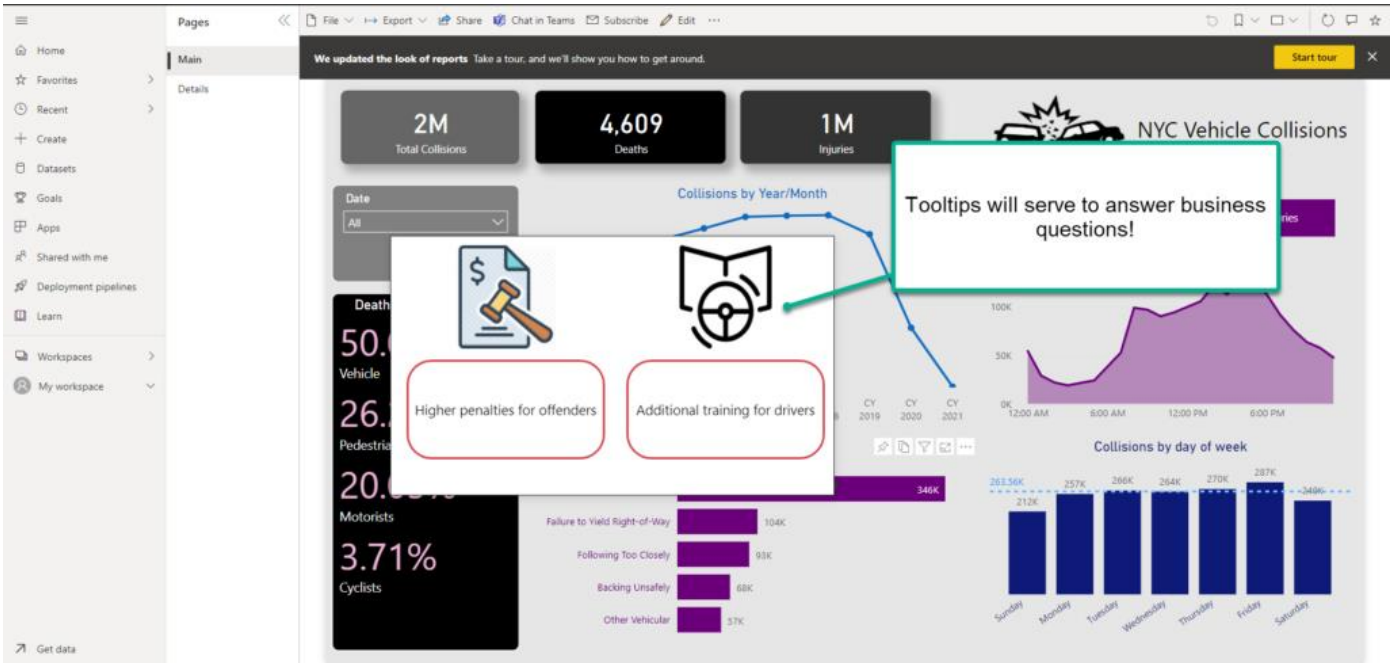


Action, please!

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 enhance your report using tooltip pages, and here we will follow a similar approach:

DATA ANALYSIS

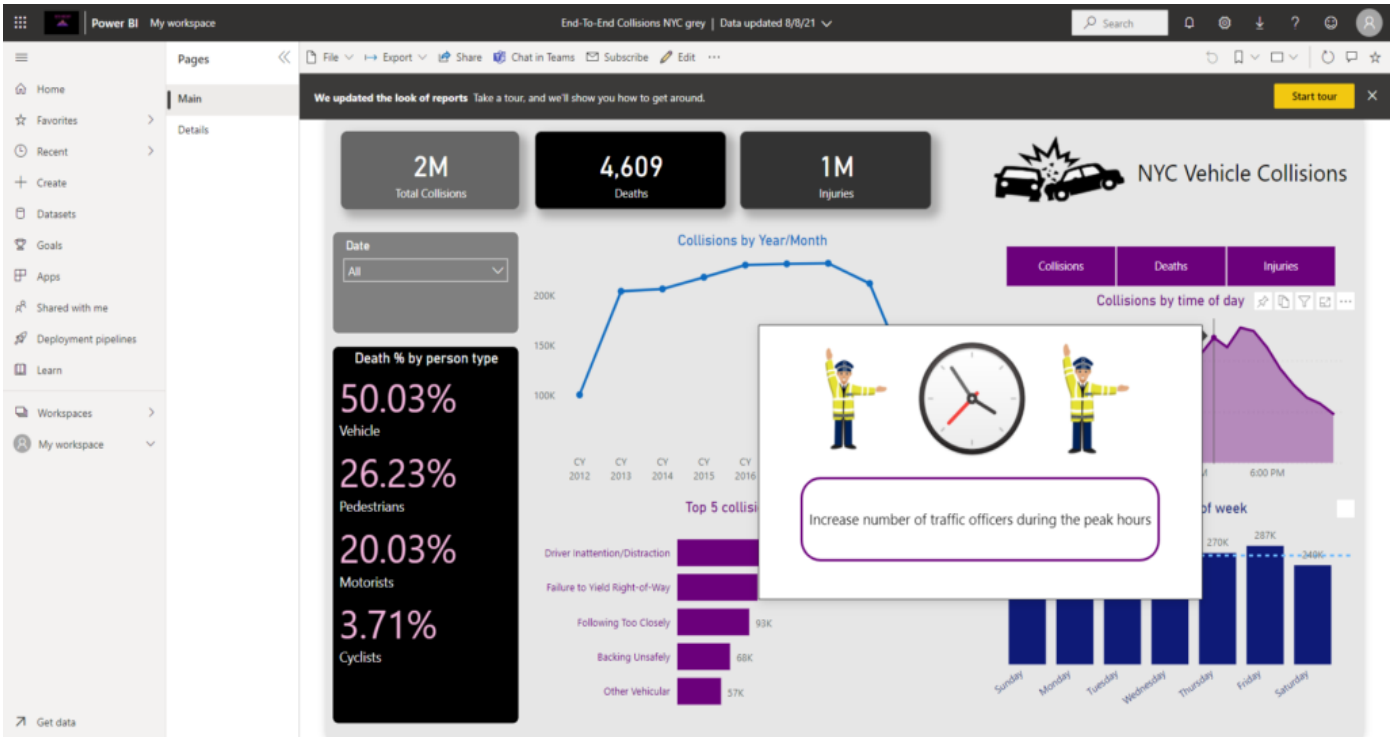


So, once I hover over a visual that shows the top 5 causes for collisions, certain actions will be recommended:

✓ Higher penalties for offenders
✓ Additional training for drivers

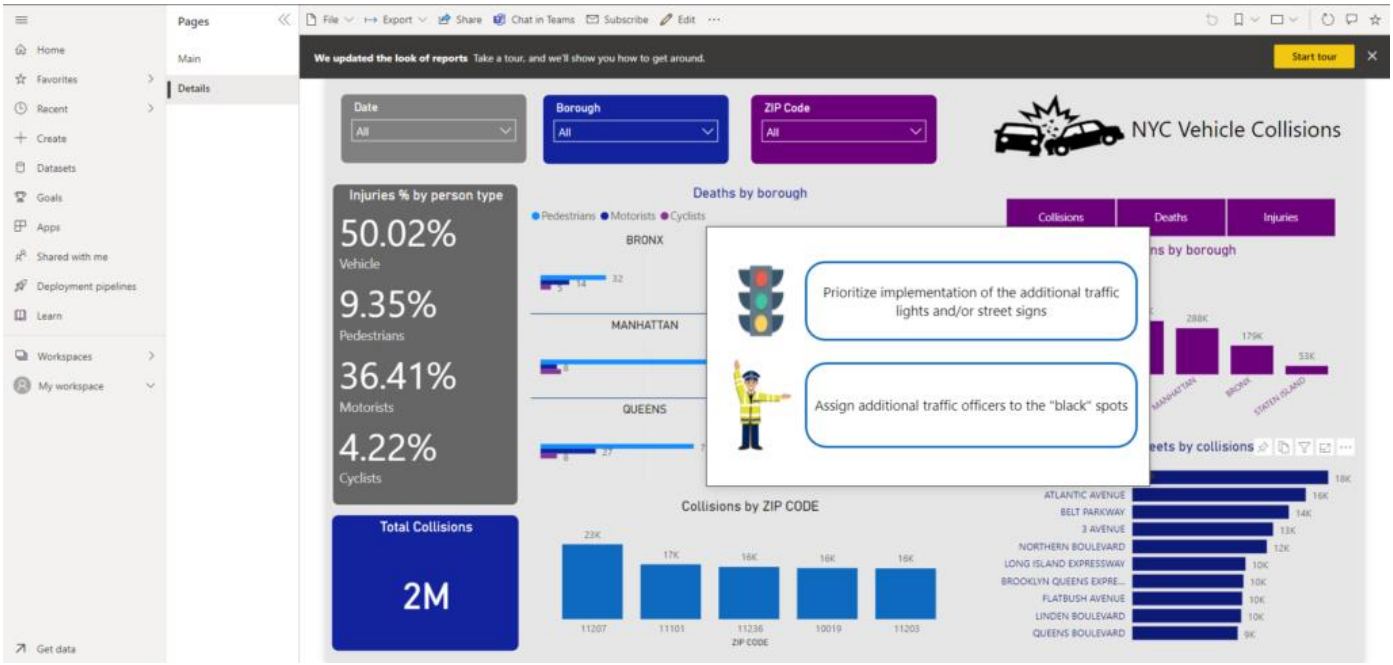
Similarly, if you want to act based on the time when most collisions occur, simply hover over that visual and you will see the suggestion to increase the number of traffic officers during these peak hours:

DATA ANALYSIS



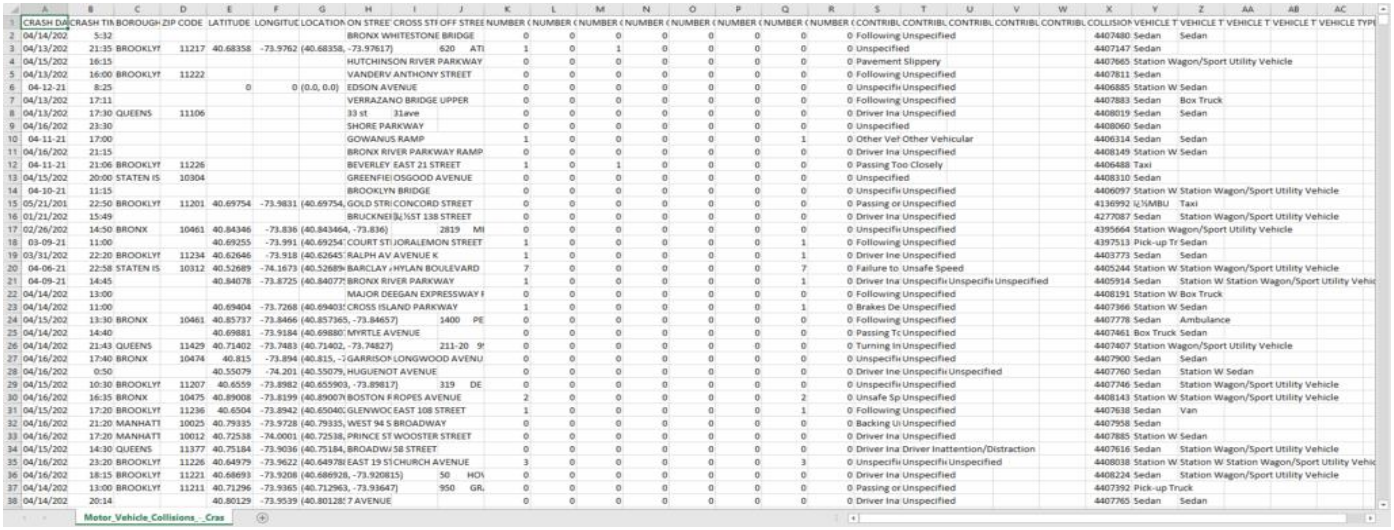
Further down, to be able to reduce the number of collisions and injuries/deaths in specific locations, we emphasized the importance of implementing additional traffic lights and assigning more traffic officers to “black” spots:

DATA ANALYSIS



SUMMARY

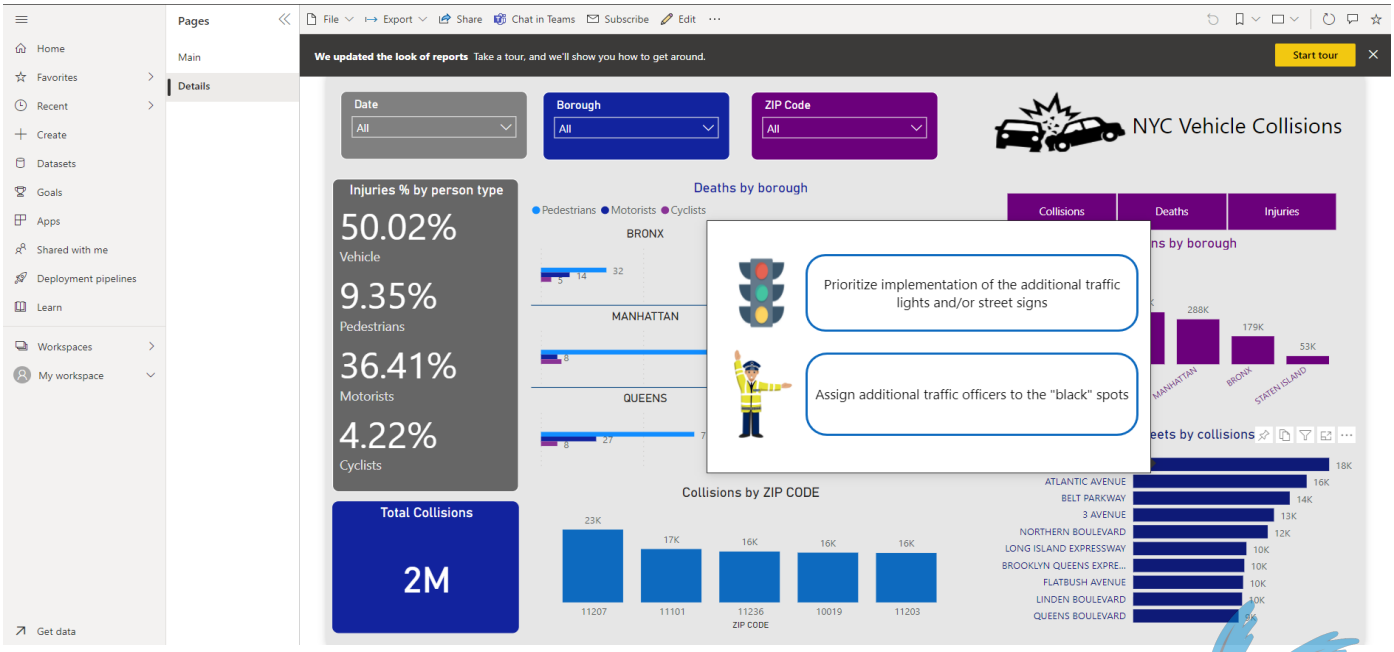
That's a wrap folks! Just to remind you, this is where we've started:



CRASH DA	CRASH TIM	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREE	CROSS ST	OFF STREE	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	NUMBER (	CONTRIBU	CONTRIBU	CONTRIBU	CONTRIBU	CONTRIBU	COLLISION	VEHICLE T	VEHICLE T	VEHICLE T	VEHICLE T	VEHICLE TYPE
04/14/202	5:32						BRONX WHITESTONE BRIDGE			0	0	0	0	0	0	0	0 Following	Unspecified				4407480	Sedan	Sedan			
04/13/202	21:35	BROOKLYN	11217	40.68358	-73.9762	(40.68358, -73.97617)	620	ATI		1	0	1	0	0	0	0	0 Unspecified					4407147	Sedan				
04/15/202	16:15						HUTCHINSON RIVER PARKWAY			0	0	0	0	0	0	0	0 Pavement	Slippery				4407665	Station Wagon/Sport Utility Vehicle				
04/13/202	16:00	BROOKLYN	11222				VANDERV	ANTHONY	STREET	0	0	0	0	0	0	0	0 Following	Unspecified				4407811	Sedan				
04-12-21	8:25			0	0	(0.0, 0.0)	EDSON AVENUE			0	0	0	0	0	0	0	0 Unspecifie	Unspecified				4406885	Station W	Sedan			
04/13/202	17:11						VERRAZANO BRIDGE	UPPER		0	0	0	0	0	0	0	0 Following	Unspecified				4407883	Sedan	Box Truck			
04/13/202	17:30	QUEENS	11106				33 st	31ave		0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4408019	Sedan	Sedan			
04/16/202	23:30						SHORE PARKWAY			0	0	0	0	0	0	0	0 Unspecified					4408060	Sedan				
04-11-21	17:00						GOWANUS RAMP			1	0	0	0	0	0	1	0 Other Vel	Other Vehicular				4406314	Sedan	Sedan			
04/16/202	21:15						BRONX RIVER PARKWAY RAMP			0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4408149	Station W	Sedan			
04-11-21	21:06	BROOKLYN	11226				BEVERLEY	EAST 21	STREET	1	0	1	0	0	0	0	0 Passing Too	Closely				4406488	Taxi				
04/15/202	20:00	STATEN IS	10304				GREENFIE	OSGOOD	AVENUE	0	0	0	0	0	0	0	0 Unspecified					4408310	Sedan				
04-10-21	11:15						BROOKLYN BRIDGE			0	0	0	0	0	0	0	0 Unspecifie	Unspecified				4406097	Station W	Station Wagon/Sport Utility Vehicle			
05/21/201	22:50	BROOKLYN	11201	40.69754	-73.9831	(40.69754,	GOLD STR	CONCORD	STREET	0	0	0	0	0	0	0	0 Passing or	Unspecified				4136992	¼MBU	Taxi			
01/21/202	15:49						BRUCKNEI	¾ST 138	STREET	0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4277087	Sedan	Station Wagon/Sport Utility Vehicle			
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 Unspecifie	Unspecified				4395664	Station Wagon/Sport Utility Vehicle				
03-09-21	11:00			40.69255	-73.991	(40.69254	COURT ST	JORALEMON	STREET	1	0	0	0	0	0	1	0 Following	Unspecified				4397513	Pick-up Tr	Sedan			
03/31/202	22:20	BROOKLYN	11234	40.62646	-73.918	(40.62645	RALPH AV	AVENUE	K	1	0	0	0	0	0	1	0 Driver Ina	Unspecified				4403773	Sedan	Sedan			
04-06-21	22:58	STATEN IS	10312	40.52689	-74.1673	(40.52689	BARCLAY	HYLAN	BOULEVARD	7	0	0	0	0	0	7	0 Failure to	Unsafe Speed				4405244	Station W	Station Wagon/Sport Utility Vehicle			
04-09-21	14:45			40.84078	-73.8725	(40.84077	BRONX RIVER PARKWAY			1	0	0	0	0	0	1	0 Driver Ina	Unspecifie	Unspecified			4405914	Sedan	Station W	Station Wagon/Sport Utility Vehic		
04/14/202	13:00						MAJOR DEEGAN EXPRESSWAY			0	0	0	0	0	0	0	0 Following	Unspecified				4408191	Station W	Box Truck			
04/14/202	11:00			40.69404	-73.7268	(40.69403	CROSS ISLAND PARKWAY			1	0	0	0	0	0	1	0 Brakes De	Unspecified				4407366	Station W	Sedan			
04/15/202	13:30	BRONX	10461	40.85737	-73.8466	(40.857365, -73.84657)		1400	PE	0	0	0	0	0	0	0	0 Following	Unspecified				4407778	Sedan	Ambulance			
04/14/202	14:40			40.69881	-73.9184	(40.69880	MYRTLE AVENUE			0	0	0	0	0	0	0	0 Passing To	Unspecified				4407461	Box Truck	Sedan			
04/14/202	21:43	QUEENS	11429	40.71402	-73.7483	(40.71402, -73.74827)		211-20	9	0	0	0	0	0	0	0	0 Turning In	Unspecified				4407407	Station Wagon/Sport Utility Vehicle				
04/16/202	17:40	BRONX	10474	40.815	-73.894	(40.815, -7	GARRISON	LONGWOOD	AVENU	0	0	0	0	0	0	0	0 Unspecifie	Unspecified				4407900	Sedan	Sedan			
04/16/202	0:50			40.55079	-74.201	(40.55079,	HUGUENOT AVENUE			0	0	0	0	0	0	0	0 Driver Ina	Unspecifie	Unspecified			4407760	Sedan	Station W	Sedan		
04/15/202	10:30	BROOKLYN	11207	40.6559	-73.8982	(40.655903, -73.89817)		319	DE	0	0	0	0	0	0	0	0 Unspecifie	Unspecified				4407746	Sedan	Station Wagon/Sport Utility Vehicle			
04/16/202	16:35	BRONX	10475	40.89008	-73.8199	(40.89007	BOSTON R	ROPES	AVENUE	2	0	0	0	0	0	2	0 Unsafe Sp	Unspecified				4408143	Station W	Station Wagon/Sport Utility Vehicle			
04/15/202	17:20	BROOKLYN	11236	40.6504	-73.8942	(40.65040	GLENWOO	EAST 108	STREET	1	0	0	0	0	0	1	0 Following	Unspecified				4407638	Sedan	Van			
04/16/202	21:20	MANHATT	10025	40.79335	-73.9728	(40.79335,	WEST 94 S	BROADWAY		0	0	0	0	0	0	0	0 Backing Ui	Unspecified				4407958	Sedan				
04/16/202	17:20	MANHATT	10012	40.72538	-74.0001	(40.72538,	PRINCE ST	WOOSTER	STREET	0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4407885	Station W	Sedan			
04/15/202	14:30	QUEENS	11377	40.75184	-73.9036	(40.75184,	BROADWA	58	STREET	0	0	0	0	0	0	0	0 Driver Ina	Driver Inattention/Distraction				4407616	Sedan	Station Wagon/Sport Utility Vehicle			
04/16/202	23:20	BROOKLYN	11226	40.64979	-73.9622	(40.64978	EAST 19	51	CHURCH AVENUE	3	0	0	0	0	0	3	0 Unspecifie	Unspecifie	Unspecified			4408038	Station W	Station W	Station Wagon/Sport Utility Vehic		
04/16/202	18:15	BROOKLYN	11221	40.68693	-73.9208	(40.686928, -73.920815)		50	HOV	0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4408224	Sedan	Station Wagon/Sport Utility Vehicle			
04/14/202	13:00	BROOKLYN	11211	40.71296	-73.9365	(40.712963, -73.93647)		950	GR	0	0	0	0	0	0	0	0 Passing or	Unspecified				4407392	Pick-up Truck				
04/14/202	20:14			40.80129	-73.9539	(40.80128	7 AVENUE			0	0	0	0	0	0	0	0 Driver Ina	Unspecified				4407765	Sedan	Sedan			

Motor_Vehicle_Collisions_-_Cras

And, this is where we finished:



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!