# Motor Vehicle Accidents in New York- End-to-end Analytical Solution Project using Power BI

Contents

[Motor Vehicle Accidents in New York- End-to-end Analytical Solution Project using Power BI 1](#_Toc105277616)

[I. Purpose of the project 1](#_Toc105277617)

[II. Dataset 1](#_Toc105277618)

[III. Analysis process 1](#_Toc105277619)

[1. Identifying Right Business Questions 1](#_Toc105277620)

[2. Data Preparation 1](#_Toc105277621)

[3. Data Modelling 4](#_Toc105277622)

[a. Scope of project 4](#_Toc105277623)

[a. Data Visualization 12](#_Toc105277624)

[b. Insight 14](#_Toc105277625)

[c. Extent Recommendation 14](#_Toc105277626)

[IV. Summary 14](#_Toc105277627)

# Purpose of the project

Building an end-to-end analytic solution with Power BI. Visualization and Recommendation are included.

# Dataset

From NYC national website contain 2M lines and not yet cleaned

# Analysis process

## Identifying Right Business Questions

## Data Preparation

1. Data Profiling

*View* tab-> Turn on *Column quality, Volume distribution, and Column profile.*This step helps to figure out the potential issues

* **Identify missing data:**

Then, we can see that in the Borough column, there are 37% in the Empty spot. It means there is 37% of missing data. This is quite a big percentage, so I decide to *replace* all blank or null data with *N/A:*

Click *Replace Value*-> Put nothing on the box *Replace:* -> Put *N/A* to the box *With:*

**- Identify Outliner/ Abnormality:**

Click NUMBER OF PERSONS INJURED, in Column Statistics, Max contains 7. It means the max number of persons injured is just 7. It is a quite small number. However, if this Max number is 70 0000 or kind of, the further step to solve this abnormality should be taken.

Graphical user interface, application, table, Excel

Description automatically generated

1. Data Shaping

* **Clean the numeric column:**

Data types of:

+ Zip Code: whole number

+ Latitude, Longitude: decimal number  
There are a lot of Null and 0 values. So, we replace all Nulls with 0

Click Replace Values, put null in the box Replace: and put 0 in the box With:

* **Find the most important column and make sure the highest data quality of this column**

That column is On Streat Name. The reason is this column would help to answer the core question ‘Which locations are at the highest risk of collision in New York?

Click the arrow down to check the Street Name there and check if any abnormal. We can see an issue with Capital and Non-Capital Letter. ‘BELT PARKWAY’ and ‘Belt parkway’ are listed in different lines. There are 2 uppercase values for BELT PARKWAY because of the CSV dataset. In that kind of data, some characters might be hidden including new line, space, tabulator, etc. The solution is to use Power Querry in Power BI to deal with that.

Graphical user interface, application

Description automatically generated

To solve it, use TRIM.

Right-click to the arrow down on the column On Streat Name (also Off Street Name and Cross Street Name-> Choose Transform-> Choose Trim. The trim transformation will remove lead and trail blanks. We can double-check if these transformation helps to remove all duplication cases there.

In this situation, an Uppercase Transformation will be made:

Right-click to the arrow down on the column On Streat Name (also OffStreet name and Cross Street Name-> Choose Transform-> Choose UPPERCASE. The result looks like this:

Then, click Close &Apply.

* **Unnecessary data: Remove**

Unnecessary data is defined as data that does not contribute to finding answers to the core business questions.

We can see that column including Contributing Factor Vehicle 3-> 5, and Vehicle Type Code 3-> 5 have an Empty rate is more than 90%, meaning that they are totally empty columns. So, we make the Remove Column action.

Shift+ Click to these first 3 column (and then the last 3 column)-> Right click to the arrow-> Choose Remove Column.

## Data Modelling

## Scope of project

As an out-of-scope aspect, we would not talk about data warehouse design including OLTP or OLAP model database design. OLAP systems are designed for use to support business intelligence (BI), data mining, and other decision support applications. OLTP is optimized for processing a massive number of transactions. This broad topic will be discussed deeply in the next project, I hope.

1. Models

There are 2 models in my consideration:

* Using **Dimensional Modelling** for this analytic solution project. It means our tables need to be defined as Fact Table (contain events, and observations) or Dimension Table(contain description).
* Using **Star Schema.** It means using 1 single flat fat table containing all data. It helps gain accurate and efficient results from data.

However, using 1 single flat table can produce incorrect results ( from my experience). Therefore, we choose the Star Schema model.

We can see some potential Dimension Tables:

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

We will find the Fact Table latter.

* Empty data: remove (done at the previous step)

Another data cleaning action carried out is Hour Transformation in the Crash Time column. The idea is to analyze data on the hour level, instead of the minute level.

Right Click to Crash Time volume-> Choose Transform -> Choose Hour-> Choose Start of Hour.

Now, we duplicate the original flat table’ Collison’ 4 times for making 4 more tables with names equivalents to the above-listed dimensions (Time dimension, Location dimension (Borough + ZIP Code), and Contributing Factor dimension, Vehicle Type dimension, except for Date dimension.). These tables before changing the equivalent names looks like this:

Graphical user interface, table, Excel

Description automatically generated

We will keep only relevant columns in each of our dimensions and remove all the others. Particularly:

* Table Collision (2)-> Change name into table ‘ Location’. Then keep columns: ‘Borough’ and ‘ Zip Code’ and then remove the rest of the columns here.
* Table Collision (3)-> Change name into table ‘ Contributing Factors’. Then keep columns: ‘Borough’ and ‘ Zip Code’ and then remove the rest of the columns here.
* Table Collision (4)-> Change name into table ‘ Vehicle Type’. Then keep columns: ‘Borough’ and ‘ Zip Code’ and then remove the rest of column here. The result looks like this:

Graphical user interface, application, Word

Description automatically generated

* Table Collision (2)-> Change name into table ‘Time’. Then keep columns: ‘Borough’ and ‘ Zip Code’ and then remove the rest of the columns here

**- Unique Values in all dimensions:**

To do so, we can create proper ***1-M relationships*** *between* ***dimensions and the fact table  
->*** Right-click and choose ‘Remove Duplicate’ in all dimensions of each table.

Therefore, there is no ‘classic’ or ‘normal’ key column on the original flat table. So, we assume building relationships on text columns

* Create a ***surrogate key***. It means a simple integer/ bigint value that increases sequentially and uniquely identifies the row in the table. There is no business meanings contained in the surrogate key

Click table ‘Contributing Factor’-> Click ‘Add Column’-> Choose ‘Index Column’-> Choose ‘From 1’.

Normally, using ***Index column transformation*** will break ***query folding.*** (Querry folding is the ability of Power Query’s Mashup engine to create a single SQL statement combining all M statements behind your transformations)

<>In this situation, Index column transformation is ok to use here. The reason is we use a CSV file, and no query folding is supported by this file type.

Add this integer column to the fact table and use it as a foreign key to our dimension table, instead of the text value. To do so, we merge the Location dimension with my Collisions fact table:

Click table Location and Table Collison-> Click Home-> Click Merge Queries-> Choose Merge Queries. A box ‘Merge’ appear-> On ‘Collisions’, click ‘BOROUGH’ and then‘ZIP CODE’ ->u see there is a number 1 next to ‘BOROUGH’ and number 2 next to ‘ZIP CODE’-> Look down and you can see a white box with an arrow-> click to this arrow and choose ‘Location’-> Click ‘‘BOROUGH’ and ‘ZIP CODE’ and you can see number 1 and 2 taking a turn to show on ‘BOROUGH’, ‘ZIP CODE’-> On’ Join Kind’, choose Inner (only) matching rows-> click the box ‘Use fuzzy matching from the merge’

Graphical user interface, application

Description automatically generated

We need to make sure we have successfully merged like this:

Graphical user interface, application

Description automatically generated

I got an error on that stage so I try to make sure I clean well data of these 2 column (replace value null-> N/A or not, etc). It was not suceffull. So, I check the DAX and make sure it is right: Table.NestedJoin(#"Calculated Start of Hour", {"BOROUGH", "ZIP CODE"}, Location, {"BOROUGH", "ZIP CODE"}, "Location.1", JoinKind.Inner)

Then, I done with the merge.

A screenshot of a computer

Description automatically generated with medium confidence

Prepare Index on column ‘Location’ on the merged Collision:

Click table ‘Location’-> Click ‘Add Column’-> Choose ‘Index Column’-> Choose ‘From 1’.

Now, we can expand the ***merged Location table*** and take the ***Index*** column from there

Click table ‘Collisions’-> Click the Expand icon of ‘Location 1’ ( its origin is the arrow icon)-> Unclick ‘Borough’ and ‘Zip code’, Click ‘Index’-> Click Ok

Graphical user interface, application, table

Description automatically generated

So, column Location 1 is my integer column. It is a foreign key to my Location dimension table.

Remove attribute columns: BOROUGH and ZIP CODE

-> It means we replace 2 text columns with 1 integer column

-> Save memory space

-> make tables look clearer

So the Location Key column now is considered a Foreign Key.

Then, do the same logic with Tables: Contributing Factor and …… It means we make index columns as foreign keys and then remove original text attributes.

* **Enhancing the data model with Date dimension and relationships**

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.

To create a Date dimension by using DAX and a calculated table, it is important to set up some rules:

* The table contains 1 DateTime type in the column.
* All the dates need to be shown
* Each date is located on its own row.
* There are no holes/gaps even if a date is not referenced by an event.

There are some functions to use in creating a date table: CALENDAR, CALENDARAUTO:

- Similarities: return: a table with a column ‘Date’ and values of dates

- Differences:  
+ CALENDAR: needs boundaries of some dates

+ CALENDARAUTO: figures out the first and the last year’s references after seeking among dates within the model.

- Disadvantages:   
+ CALENDAR: Due to the above differences, this function needs MIN and MAX of dates of all transactions ( sales, etc.) and then, transfers them to the given first and the last day of the given years -> Ok to use this function  
  
+ CALENDARAUTO: Due to the above differences, this function might consider Customer Birth Dates are also the values that need to be found-> contains Irrelevant years-> should not use this function  
  
=> Solution: Use CALENDARAUTO to:  
+Seek for dates  
+ Then, remove from all found dates that do not include/ refer to the period of interest:  
Click to ‘New Table’-> a box opens-> write these codes:

DateTable =

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

Passing above codes into another data model because using 2 variables here MinYeat, Maxyear, the rest of the codes do not rely on this data model.   
  
After writing these codes on the box, a new table is created. It is ‘DateTable’

Table

Description automatically generated

* Then, click ‘Mark as a table’-> choose ‘Dates’
* Then, create a Schema Model on Power Bi Model view, not Transformation view
* ***Problem:*** Cannot make the relationship among tables  
  -> Try to create new relationship on ‘Manage Relationship-> Error: ‘Many-many’ relationships, not ‘one-many’ or one-one’
* Create a key (with name: original table name+ key) by using Add Column-> Choose Index-> Choose Start with 1.  
  So we have these key column  
  - On the Location column, we create the Location Key

- On Contributing Factor column, we create Contributing Factor Key

- On Vehicle Type column, we create the Vehicle Type key  
- On Time column, we create Time key

-Then do the task ‘Expand querries’ as mentioned above and task ‘Expanded Location.1’. It would take so much time to run these tasks, especially the second task.

-Next, we click’ Make Relationship’ to create a Star Scheme Model in Model view. I make this manually because somehow, I cannot make autodetect function to create this model. Next time, I will try to do this kind of model automatically

Graphical user interface, application

Description automatically generated

* **Reflection from this part:**

I made some better improvements than applying the Power BI tool only. It is quite a nice feeling. I still go closer to the goal I create an efficient and user-friendly end-to-end analytic solution using Power BI only.

- What I done:

+ Necessary data cleaning action and data shaping (building hierarchical relationships between entities in a query)

-> even make Star Model

=> Benefits: increase usability (less complex)

+ Writing simple DAX code for different calculations.

- What I have not yet done:

+ Cannot merge the 'Borough' column and ' Zip Code' column from 2 tables 'Collisions' and 'Location. The running time of this task is so long (still running after 1 hour and so many clicks to refresh.  
-> Next time, will try making ‘Merge’ by using DAX, not M language

+ Improvement is still slowly

## Data Visualization

I am not a visualization wizard, a good aesthetician. My visualization improvement is thanks to books such as the storytelling data of Cole, and some influencers from the BI community.

* Making Card Visualize:  
  + Total Collisions

+ Deaths: data takes from NUMBER OF PERSONS KILLED

+ Injured: Click Modelling-> a box appears-> Write:

Deaths = SUM(ALL('Collisions'[NUMBER OF MOTORIST KILLED], [NUMBER OF CYCLIST KILLED], [NUMBER OF PEDESTRIANS KILLED]))

* On Model view, at Collisions field, click ‘Modeling’-> choose add column-> write:

COUNT CONTRIBUTION = COUNT('Collisions'[CONTRIBUTING FACTOR VEHICLE 1])

-RANK CONTRIBUTION FACTOR = RANKX(ALL(‘Collisons'[CONTRIBUTING FACTOR VEHICLE 1])),[COUNT CONTRIBUTION],,0)

- VALUE = IF([RANK CONTRIBUTION FACTOR]<=[SELECTEDTOPNVALUE],[COUNT CONTRIBUTION])

* This visualization made from Power BI

A picture containing timeline

Description automatically generated

On the main page of this dashboard (the only page now), besides the Line chart and Column Bar chart, the Multi-row card shows clearly who is the most endangered in the traffic. The key category is the time of the day. This key is used to move within different metrics on the same visual

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.

-> Improvement: Next time, I will create a Detail page including information about accidents with slicers for Borough and ZIP Code.

Some visuals will be shown based on 2 categories – Person Type and Borough. These factors are also relevant to the mentioned Main page.

* **Conclusion part a:**

Though I have just done the first main page (the Power BI dashboard link is not active, and the detail page is not yet done), we would come closer to the end goal. The key data points are taken full use of the visualization and hope it helps for better communication to stakeholders/ customers.

This visualization plays a good role to get meaningful insights and create recommendations to answer the initial questions in the beginning.

## Insight

* From Card ‘Death % by person type’, it shows the pedestrians are the most endangered traffic participants – 8x times more pedestrians died, compared to cyclists
* All other causes combined are lower than the cause ‘Driver Inattention’

## Extent Recommendation

* Highest penalties for offenders
* Additional training for drivers
* Additional Street lights/ Traffic Officers

# Summary

I have almost completed the mission: Building an end-to-end analytic solution with Power BI. The improvement mentioned above will be accomplished on next sessions