Home Assignment #2 Documentation

Course: Data Visualization 1
Author: Márton Nagy

1. Data source and description

During my second home assignment, I decided to work with the **NYC Traffic Accidents** dataset from Maven Analytics Data Playground. The dataset **presents the motor vehicle collisions** reported **by the New York City Police Department** from **January 2021 to April 2023**. Each record represents an individual collision, including the date, time, and location of the accident (borough, zip code, street name, latitude/longitude), vehicles and victims involved, and contributing factors[1]. The table has **18 fields and 238,421 records** in total.

2. Analytical goals

With my dashboard, I wanted to present some **key features of deadly traffic collisions in NYC**. That is, I focused only on accidents for which there was at least one death reported. I also separated **deaths into two parts: motorists'** (that is people in/on a motorized vehicle) **and non-motorists'** (that is pedestrians and cyclists) death. The dashboard aims to provide insights into the following questions:

1. What are the **trends in traffic collision deaths in the dataset for motorists and non-motorists**?
2. What is the **geographical distribution of traffic collision deaths** (for motorists and non-motorists) **amongst New York's five boroughs**?
3. What is the likelihood of an accident being deadly for motorists and non-motorists during each hour of the day?

3. Dashboard layout

The dashboard is **structured according to the best practices** learnt during the course. The key elements are the following:

- **Title and description** in the upper left corner.
- **Date slicer** in the upper right corner.
- **KPIs** below the title, centered.
- The **three charts related to** the above-mentioned **analytical goals.**

You can find below some more detailed information on the elements (other than the title and description). You can find the **picture of the dashboard at the end of this section on a separate page**.

[1] Description taken from Maven Analytics' website.

3.1. Slicers

I included a **between date slicer** in my dashboard with which the user can filter to any date interval that they deem relevant. The slicer **affects all the charts and KPI cards** on the dashboard.

I also included a **dropdown-list slicer for vehicle category** (a field grouping vehicle types to five categories). As there are **17 distinct vehicle types** in the dataset, the resulting **graphs when the user would select only one vehicle type, could have been not quite informative**. This issue **was resolved by grouping vehicle types** according to some common sense logic by a calculated column (VehicleCategory) to have fewer options. The syntax for the column, showing the exact matching between vehicle type and vehicle category, is the following:

```
VehicleCategory =
     SWITCH(
          TRUE(),
          'NYC_Collisions'[Vehicle Type] IN {"Bicycle", "Scooter",
               "Motorcycle"}, "Light Personal",
          'NYC_Collisions'[Vehicle Type] IN {"Limousine", "Taxi", "Passenger
               Vehicle"}, "Passenger",
          'NYC_Collisions'[Vehicle Type] IN {"Delivery Vehicle", "Utility
               Vehicle", "Construction", "Transport"}, "Commercial & Utility",
          'NYC_Collisions'[Vehicle Type] IN {"Emergency Services", "Fire
               Services", "Bus"}, "Emergency & Public Service",
          'NYC_Collisions'[Vehicle Type] IN {"Other", "Other (Open
               Passenger)", "Not Reported", "Unknown"}, "Other",
          "Uncategorized"
)
```

Also, note that **cross visual filtering is also available** in the dashboard.

3.2. Non-visible filters

I filtered the dashboard to **only include records until 2024-03-31**, as records for April were incomplete. I also **filtered out records where no borough was reported**.

3.3. KPIs

My dashboard includes four KPIs:

1. **No. of Collisions**: quite straightforward, I only had to add a count of the collision IDs.
2. **No. of Deadly Collisions**: same as above, but **I filtered the card** to only include records where **at least one person was killed**.
3. **Avg. of People Killed By 1K Collisions**: a simpler way would have been to add only the **average number of people killed** – however, this is a **very small number** (luckily), so I created a **measure that simply multiplies this by 1,000** to get the average by 1K collisions. The measure's syntax is the following:

   ```
   AvgDeathby1000 = AVERAGE(NYC_Collisions[Persons Killed])*1000
   ```

4. **Chance of a Collision Being Deadly**: for this, I added a **new column called Deadly**.
   The column's syntax is the following:
   `Deadly = If(NYC_Collisions[Persons Killed]>0,1,0)`
   This column takes **1 if at least one person died in the accident, and 0 otherwise**. Taking an **average of this** gives us the proportion of deadly accidents (which also **can be interpreted as the chance of a collision being deadly** in the sample). I achieved this with a measure:
   `ChanceDeadly = AVERAGE(NYC_Collisions[Deadly])`

3.4. Charts

Below you can find a detailed description on the three charts included in my dashboard. The **coloring of the visuals follows the same pattern**: **motorists** are colored in **taxicab yellow color**, while **non-motorists** are presented with a **dark grey color** resembling New York's pavement.

3.4.1. Trends in the No. of People Killed in Traffic Collisions

This chart shows the **number of motorists and non-motorists killed in traffic collisions for each month**. I opted for a **stacked column chart** to visualize this, as this way I think the **users can more easily compare deaths of the two groups, as well as see the overall number of deaths by month**. (A line chart may have been another option, but that would have made these comparisons much harder.)

I included **direct data labels for both categories** so that users do not have to work hard to figure out, for example, the number of motorist deaths for a certain month. However, I **did not include total labels** to avoid the chart being **overcrowded**. Instead, the Y-axis is visible with gridlines, so that the totals can be read off that way.

While making this visual, I encountered the **problem**, that when I included the whole Date field, the **months showed up on the X-axis with their full names**, which took up a lot of space from the dashboard. I found a **workaround for this by creating a new column with only the number of the month**, and I included that instead of the date hierarchy's month. The syntax for this column is the following: `Month = MONTH(NYC_Collisions[Date])`. I did the same for years as well: `Year = YEAR(NYC_Collisions[Date])`. This way, months are shown only by their number, and not their full name, **saving some space**.

Also, I had to include a **calculated column** called NonMotorists_killed, which **sums up cyclists and pedestrians killed for each accident**. I also use this column in the second visual. The syntax for this column is: `NonMotorists_killed = NYC_Collisions[Cyclists Killed] + NYC_Collisions[Pedestrians Killed]`

3.4.2. No. of People Killed in Traffic Collisions by Boroughs

This chart presents the **number of motorists and non-motorists killed in traffic collisions in the five boroughs of New York**. The visual type chosen to represent this is a **clustered bar chart**, as the **main message I want to convey** with the visual is the **differences between motorist and non-motorist deaths between boroughs**.

As I added **direct labels** to the bars, I **removed the X-axis completely**, only keeping the axis title.

This chart also makes use of the previously mentioned NonMotorists_killed column (see the syntax above).

The chart **is ordered by total number of people killed**. As I did not want this field to be shown directly on the chart, I added it to the tooltips so that I can use it to order the Y-axis. I did this by adding a new column called Total_killed[2]:

```
Total_killed = NYC_Collisions[NonMotorists_killed] + NYC_Collisions[Motorists Killed]
```

3.4.3. Chance of a Traffic Collision Being Deadly during a Typical Day

The chart presents **how the likelihood of a traffic accident being deadly changes throughout the day for motorists and non-motorists**. Preparing this chart required adding some calculated columns and measures to my data model.

MotoristDeadly takes the **value 1 if the accident had at least 1 motorist fatality, 0 otherwise**. Similarly, **NonMotoristDeadly** takes the **value 1 if the accident had at least 1 non-motorist fatality, 0 otherwise**. The syntaxes for these are the following:

```
MotoristDeadly = IF(NYC_Collisions[Motorists Killed]>0,1,0) and
NonMotoristDeadly = If(NYC_Collisions[NonMotorists_killed]>0,1,0)
```

Taking the **average of these columns** (with the ChanceMotoristDeadly and the ChanceNonMotoristDeadly measures) by hour **gives us the proportion of deadly accidents in each category** (which then can be **interpreted as a sample probability**). The syntaxes of these measures are the following:

```
ChanceMotoristDeadly = AVERAGE(NYC_Collisions[MotoristDeadly]) and
ChanceNonMotoristDeadly = AVERAGE(NYC_Collisions[NonMotoristDeadly])
```

For the **X-axis**, I created a column with **only the hour component of the time** of the accident:

```
Hour = HOUR(NYC_Collisions[Time])
```

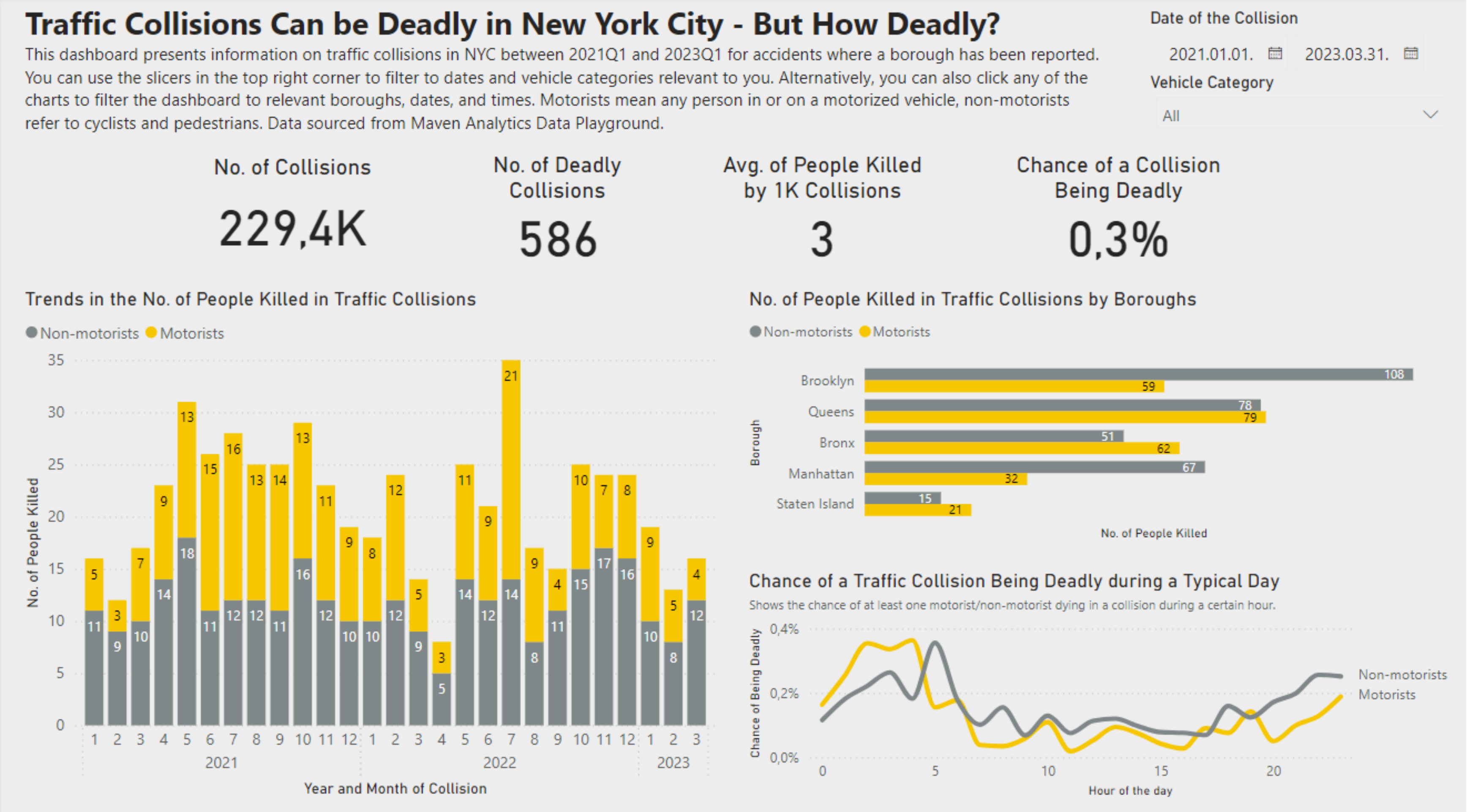
As for the visual type, I opted for a **line chart,** as I wanted **to emphasize which parts of the day are more deadly for each group**. I included series labeling so that the legend can be turned off.

[Please find the picture of the whole dashboard on the next page.]

[2] This was necessary as somehow the original Persons killed column does not always equal the total of motorist, pedestrian and cyclist deaths.

3.5. Whole dashboard

You can find below the picture of the dashboard with no slicers or cross-filtering applied[3].



Traffic Collisions Can be Deadly in New York City - But How Deadly?

This dashboard presents information on traffic collisions in NYC between 2021Q1 and 2023Q1 for accidents where a borough has been reported. You can use the slicers in the top right corner to filter to dates and vehicle categories relevant to you. Alternatively, you can also click any of the charts to filter the dashboard to relevant boroughs, dates, and times. Motorists mean any person in or on a motorized vehicle, non-motorists refer to cyclists and pedestrians. Data sourced from Maven Analytics Data Playground.

Date of the Collision: 2021.01.01. – 2023.03.31.

Vehicle Category: All

No. of Collisions	No. of Deadly Collisions	Avg. of People Killed by 1K Collisions	Chance of a Collision Being Deadly
229,4K	586	3	0,3%

Trends in the No. of People Killed in Traffic Collisions

No. of People Killed in Traffic Collisions by Boroughs

Chance of a Traffic Collision Being Deadly during a Typical Day

Shows the chance of at least one motorist/non-motorist dying in a collision during a certain hour.

[3] Note, that the decimal separator is a comma as my computer's default localization is Hungarian.