CUNY Baruch College

Traffic Signals and Vehicle Collisions by Borough

By: Karissa Maharaj

**My Datasets:**

Main (311 Complaints): DOT Street Lights and Traffic Signals

([https://data.cityofnewyork.us/Transportation/DOT-Street-Lights-and-Traffic-Signals-311-Servic e-/jwvp-gyiq](https://data.cityofnewyork.us/Transportation/DOT-Street-Lights-and-Traffic-Signals-311-Service-/jwvp-gyiq))

Secondary: Motor Vehicle Collisions, specifically Crashes

(<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data>)

**Description:**

With the growing number of traffic accidents in New York City, I want to see whether or not traffic and street lights are one of the few factors that are responsible for the recent growth. We all know how chaotic driving in the city is, especially with plenty of traffic and congestion on the roads. However, driving is even more chaotic and dangerous with broken street lights. One could be stuck at an intersection, not knowing whether they have the right of way, whose turn it is to cross the street, or if a driver is responding to a green light that should have been red. If broken street lights and traffic signals correlate with such an increase in a traffic accident, I hope my findings can help those in charge to take the appropriate course of action to prevent such accidents.

My primary dataset is from the 311 open data dataset called the ‘Department of

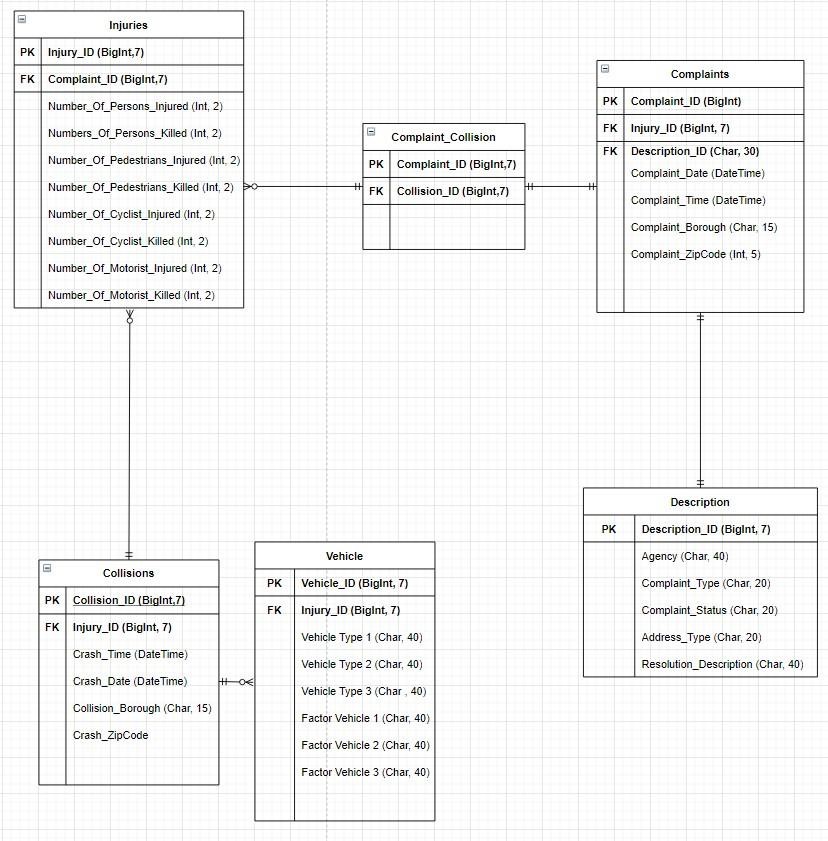
Transportation street lights and traffic signals’ report. My second dataset will be the ‘Motor Vehicle Collisions’ information, specifically for crashes, to integrate my understanding of how the primary dataset relates to the secondary in question. I will be looking at how the two relate to each other (specifically by borough).

**KPI’s of Interest:**

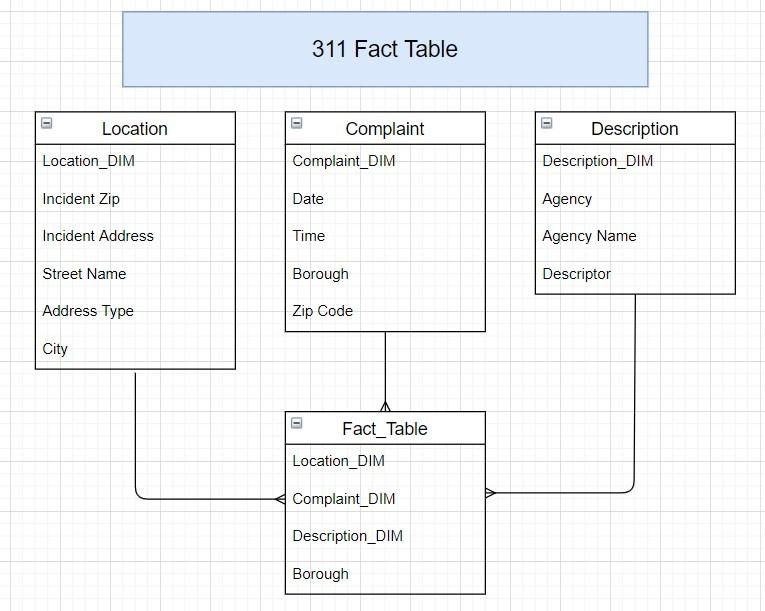
|  |
| --- |
| Percentage of Traffic Accidents within each borough (Motor Vehicle Crashes database). |
| Percentage of complaints about street/traffic lights being out within each borough (311 database). |
| Rate of injury/fatality (Motor Vehicle Crashes database). |

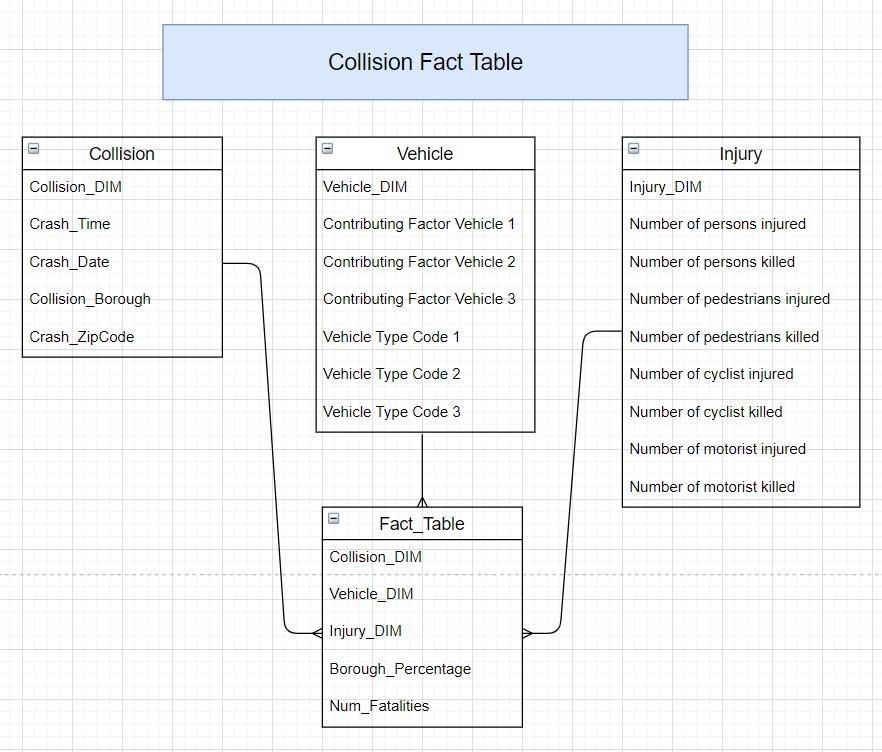
\*I will be looking into cases by borough for both datasets.

**Dimensional Model Draft (1st Attempt, I made a mistake and combined both tables\*\*)**

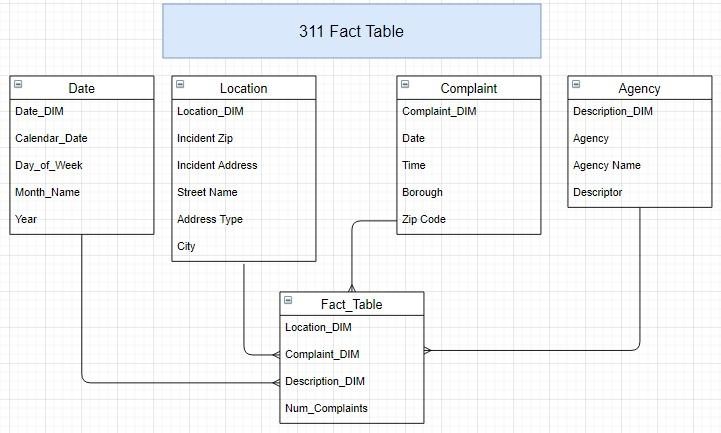


**Fact Tables (Attempt #1)**

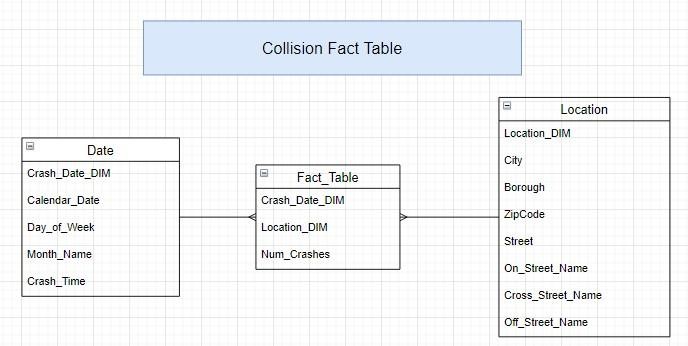




**Revised Fact Tables:**



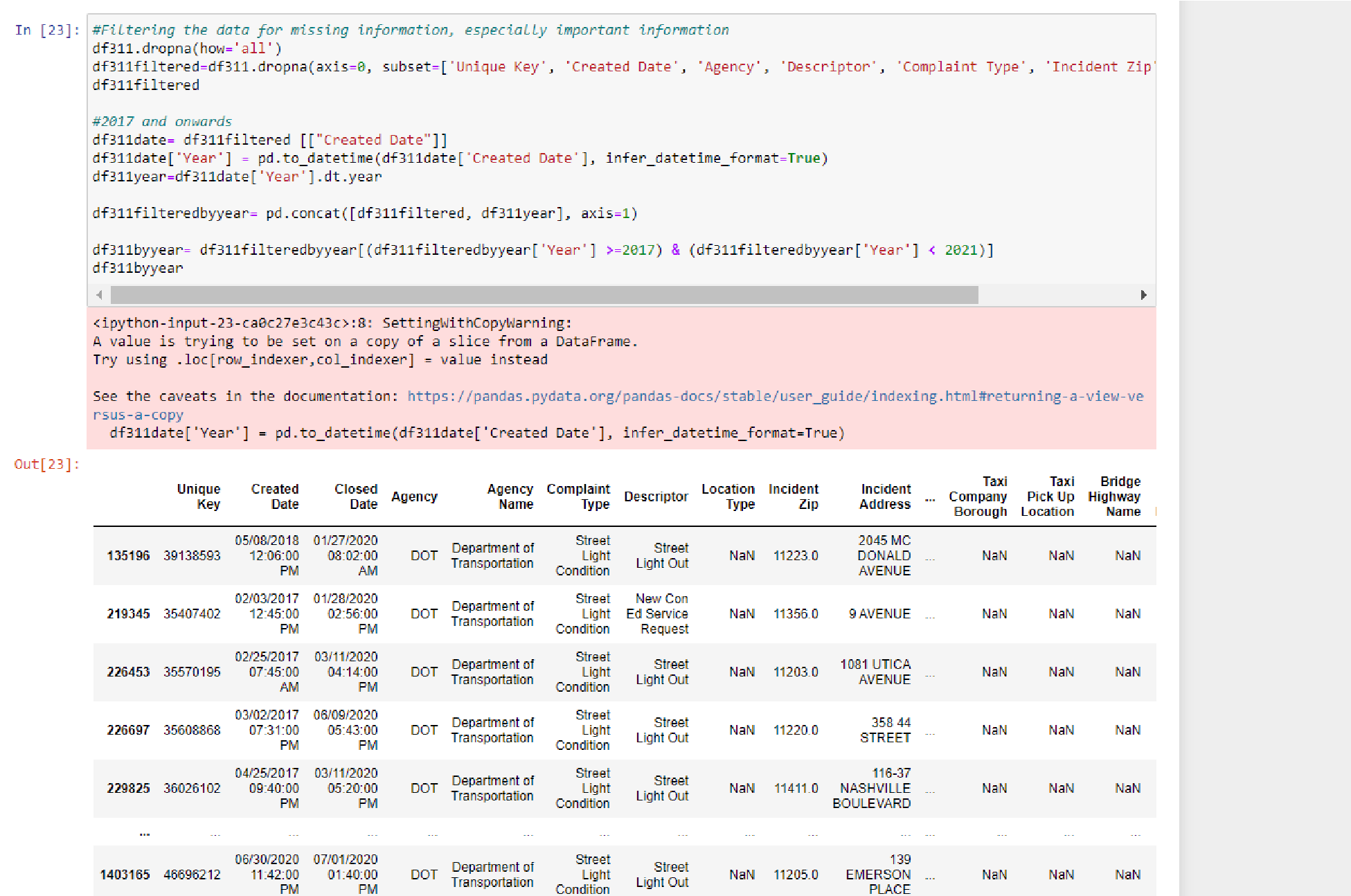
**Collision Data - Periodic Snapshot Grain**



**Notes:** I plan on using Python to import the CSV files through the 311 sites, then creating two separate dataframes through the use of Pandas to store each of the separate data tables in. My process to clean up the data will include an ‘if’ statement that will parse through the columns defined in the dataframes. If any rows are missing any important columns that I designate, such as Zip Code or Date, I will remove such rows. Other less important columns will be filled with N/A to ensure data integrity. After my data is cleaned up, I will then export the data to Google Bigquery in order to explore my data using SQL.

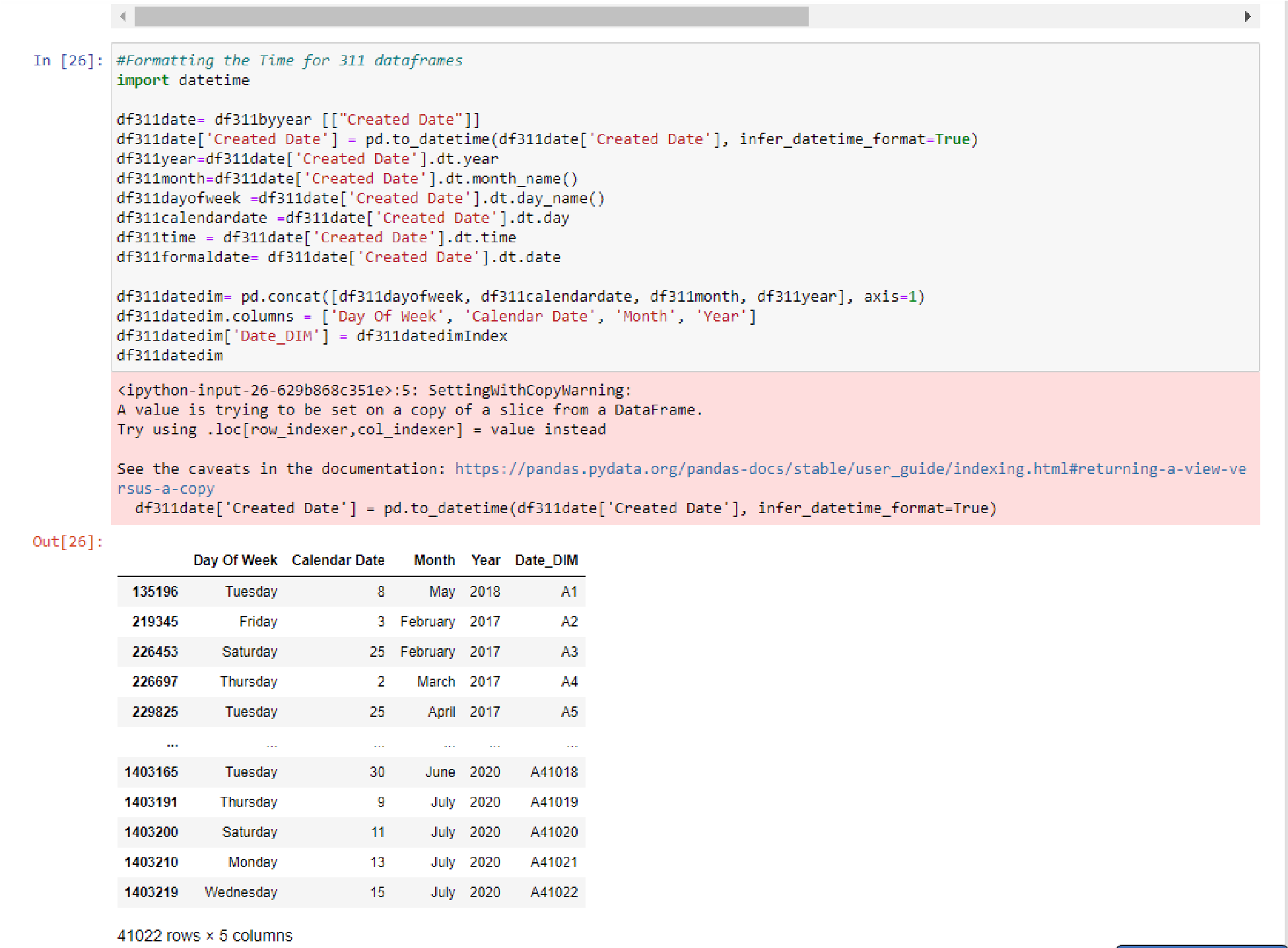
# ETL Process

For the ETL process, I used Python to create a main dataframe for each of the two datasets. After getting the two datasets into Python, I was able to filter out the data using the ‘dropna’ function to parse through each column looking for important data that is missing and restricting the data to the years of 2017 to 2020.



After I was able to filter out the data, the number of rows in the dataset went from almost 3,000,000 combined rows between both datasets to just under 500,000 rows in total.

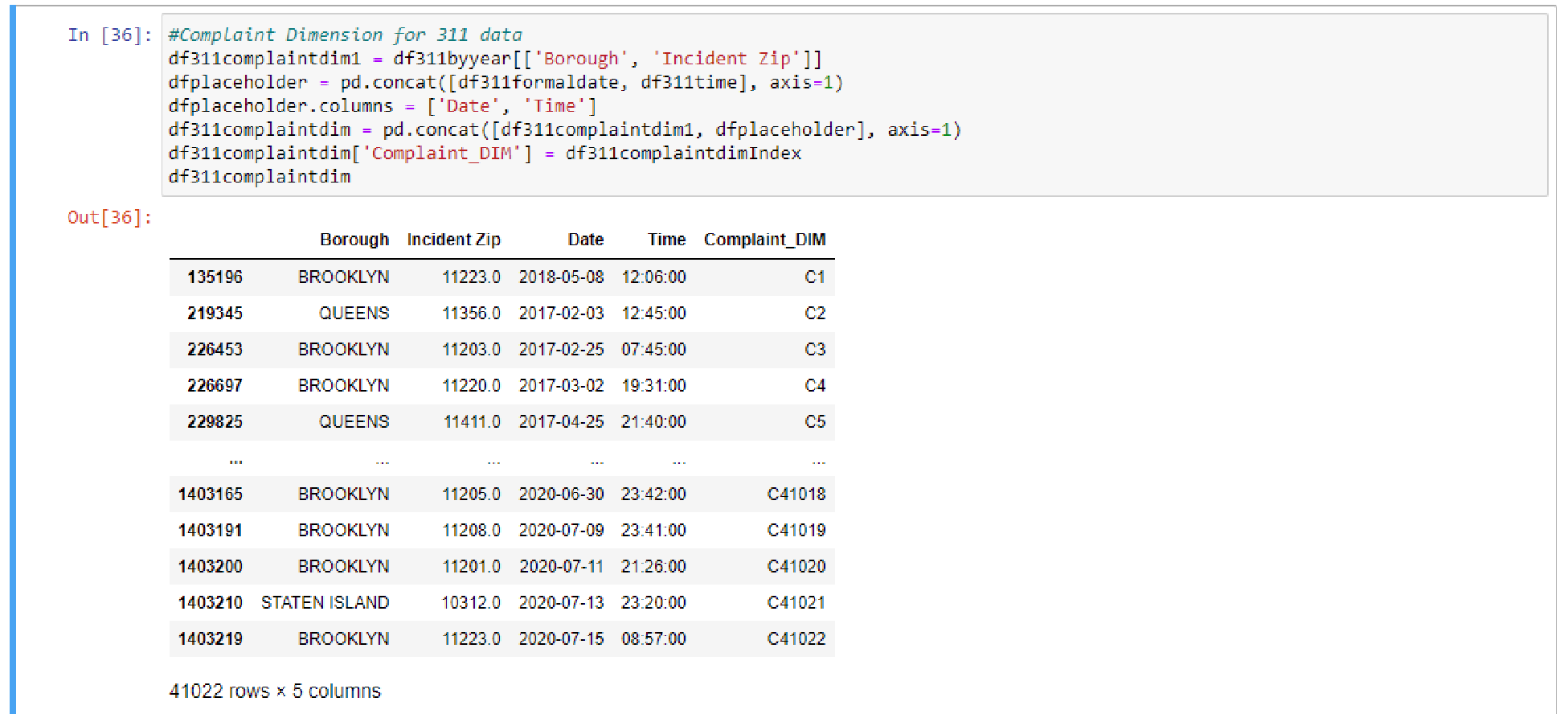
To create each of the six dimensional models, I created six unique data frames that would take the data from their respective dataframe. I also incorporated the date\_dim identifier, which is a unique identifier that would help us separate each dataset from one another. The screenshot below shows my data dimensional model for the 311 dataset.



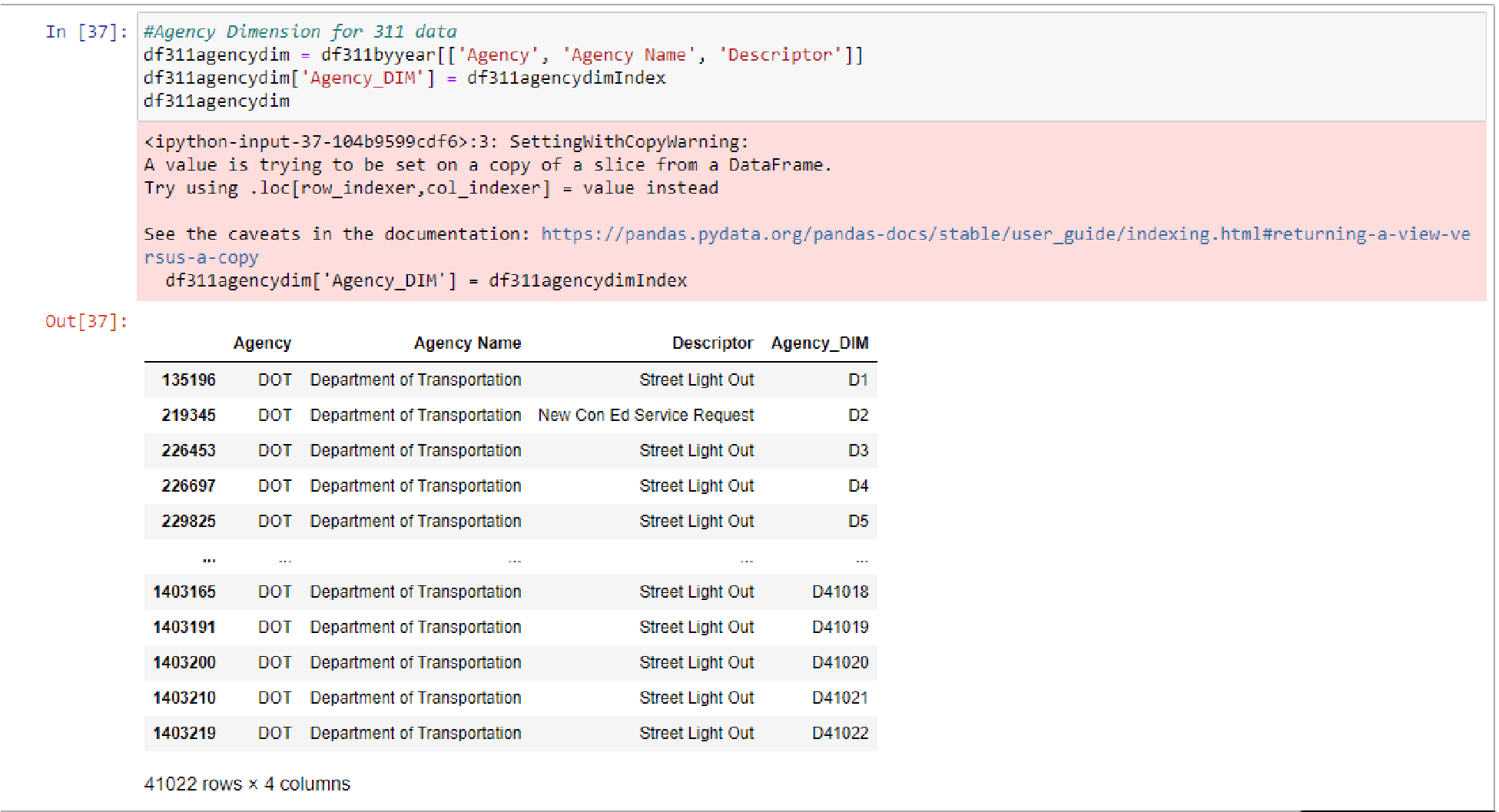
This screenshot showcases the location dimensional model for the 311 dataset with its own Location\_dim identifier.



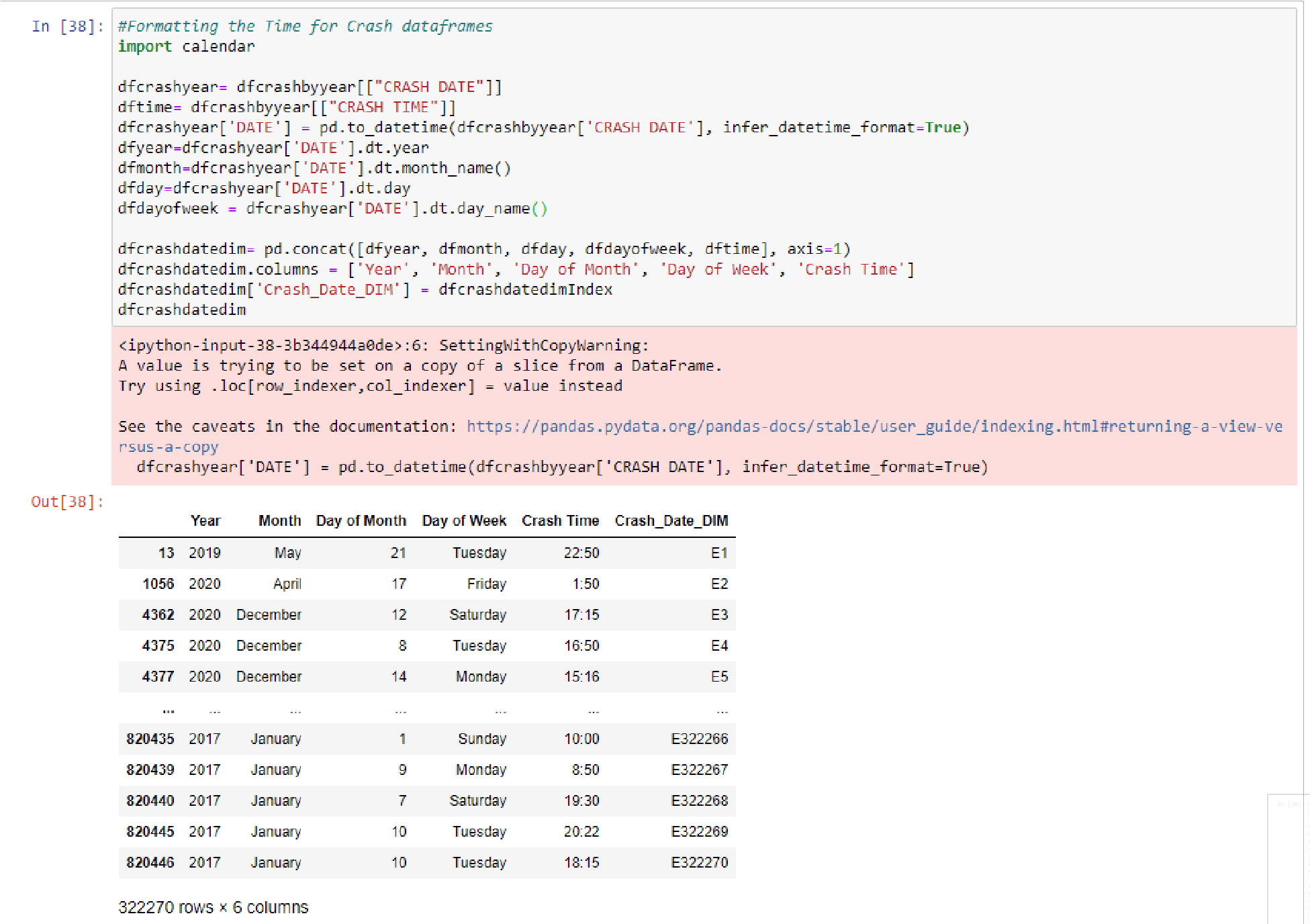
My third dataframe, which encompasses the compliant dimensional model for the 311 dataset, along with the Compliant\_DM identifier.



This is my last dimensional model for the 311 dataset. This showcases the agency dimension. Throughout all four of my dimensional models, the rows stay consistent. It is also important to keep in mind that when trying to code these data frames I tried to decide on how I were going to identify each dataset from another. I decided to use the alphabet followed by an incremental increase of 1 for each row. So it would start from D1 -> D2 -> D3 and so on.



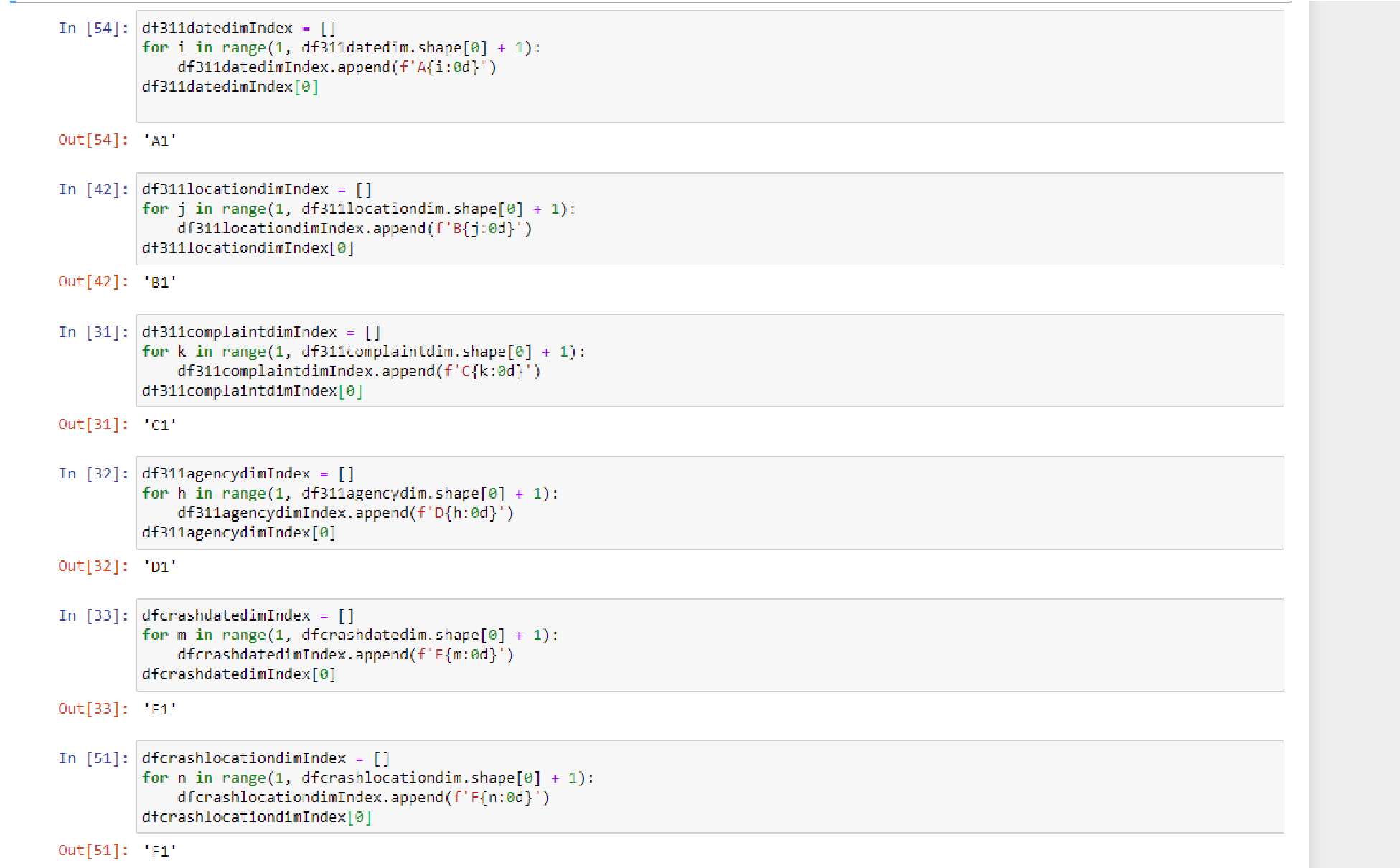
Moving onto the crash dataset where I only have two dimensional models I plan on creating. This first model shows the crash date dimension from the crash dataset. One thing to note is that this dataset, even after it was filtered, contains over 300k rows.



Last of the six dimensional models. This last model is my crash location dimension of the crash dataset.



What I used to create the identifiers for each of the dimensions.



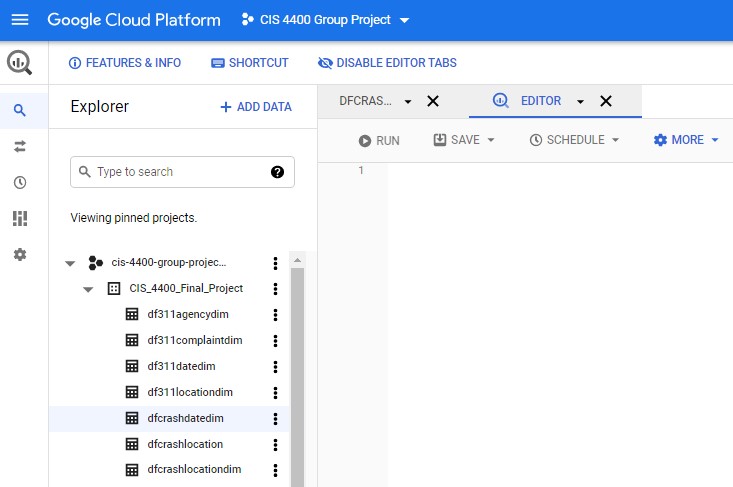
# Dashboard Programming



For this part of the project, I imported the necessary Python packages to import my data over to Google Console.



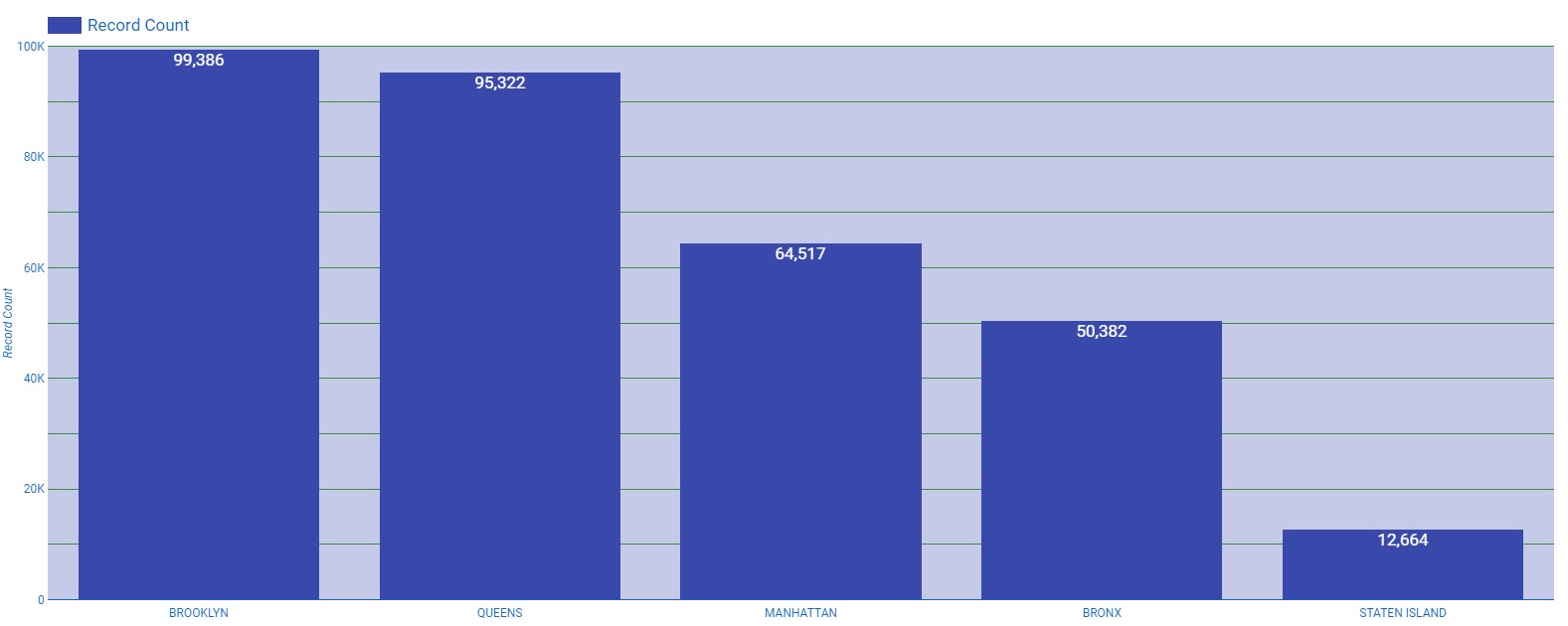
I then uploaded the dataframes one by one to my Google Console project.



**Google Studio Visualizations**

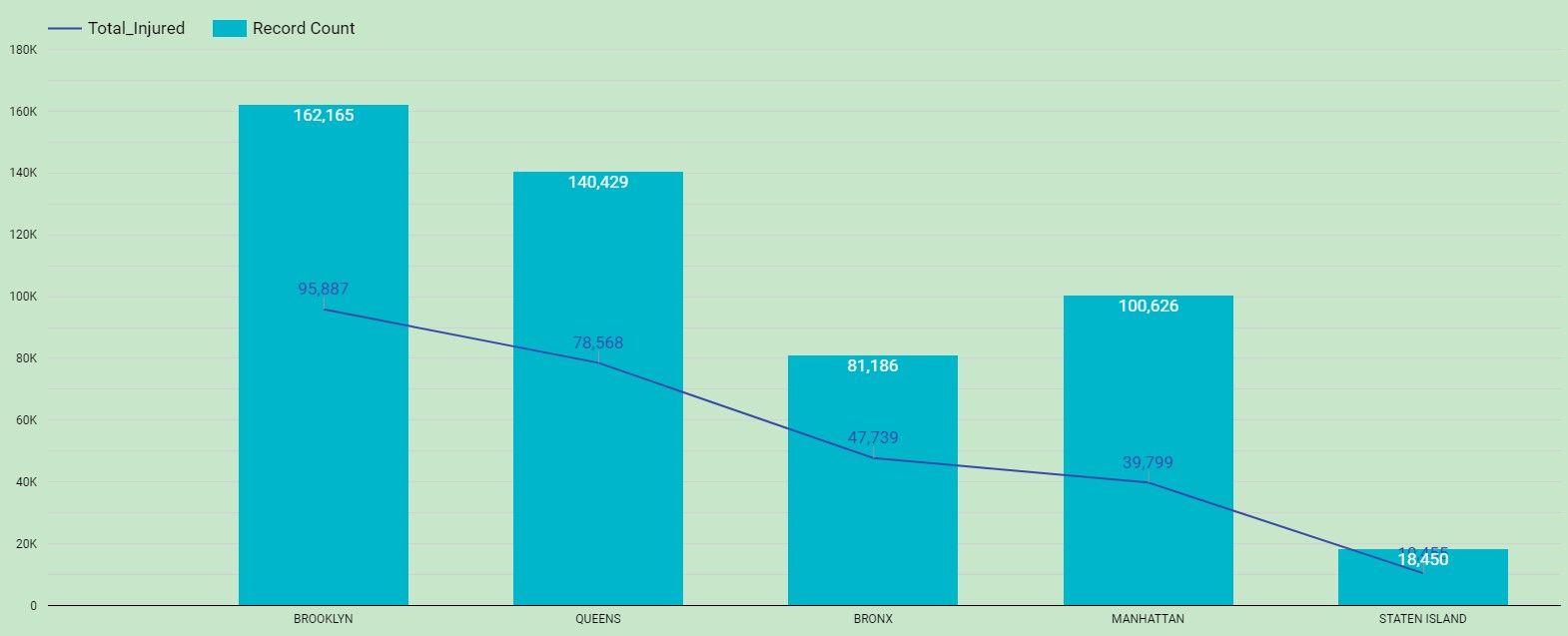
I used Google Data Studio to create visualizations of my KPIs.

Number of Traffic Accidents Within Each Borough:



**Caption:** This figure shows my KPI of the number for traffic accidents within each borough. As you can see, each borough has its own count for how many traffic accidents occurred per borough. In Brooklyn, between 2017 and 2020, there are 99,386 records of accidents. In Queens, there are 95,322 records of accidents, with 64,517 in Manhattan, 50,382 in the Bronx, and 12,664 in Staten Island.

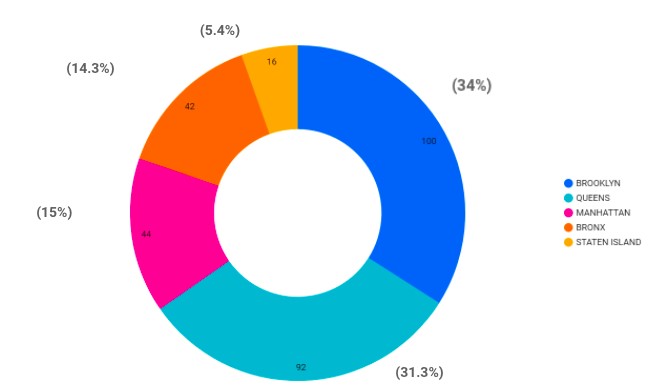
Rate of Injuries in Accidents within Each Borough:



**Caption:** This figure is illustrating the rate for how many people who get injured per accident, filtered by borough. The light blue bars indicate the amount of total accidents by borough, and the dark blue line indicates the amount of injured people from those accidents. The number of accidents is in white text, while the amount of injured people is in blue text. For instance, out of the 162,165 vehicle accidents in Brooklyn, 95,887 people got injured. [Disclaimer: For Staten

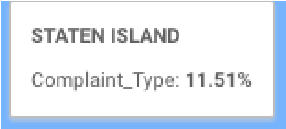
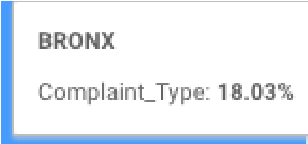
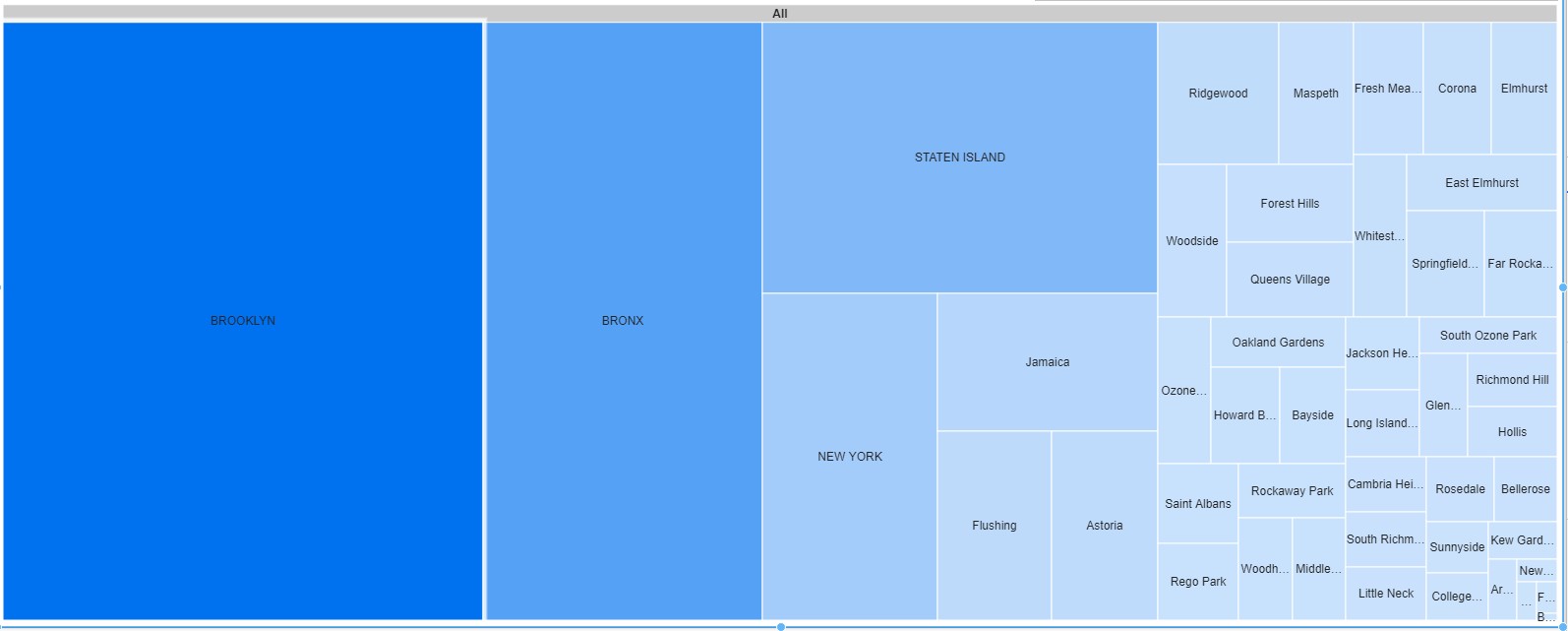
Island, the amount of injured people was 10,455]

Percentage of Fatalities Across the Boroughs



**Caption:** This figure illustrates the total number and percentage of fatalities in vehicle accidents, broken down by borough. Brooklyn has the highest rate of fatalities, with 34% of the total fatalities, with 100 deaths. Queens has the second highest rate with 31.3% of the total fatalities from accidents, with 92 deaths. Staten Island accounts for 5.4% of the total fatalities with an overall 16 deaths.

Percentage of complaints about street/traffic lights being out within each borough:



**Caption:** This figure is a tree map that demonstrates the percentage of complaints by borough. The darkest shades of blue are the boroughs with the higher percentages of complaints, and as the shades of blue become lighter in color, the percentage of complaints by borough drops. Here, we can see that Brooklyn has the highest percentage of complaints, with 30.83% of the total. Improper Street Light conditions ended up being the most representative complaint type across each borough. [Disclaimer: The visualization recognized all cities in Queens and Manhattan to be individual entities, thus leading to many different boxes to show up on the graph. Due to this, the true percentage of Queens is not shown on the treemap.]

**Conclusions**

One conclusion I can draw from my analysis is that Brooklyn needs a lot more attention when it comes to 311 traffic signal complaints. Brooklyn has the most accidents, highest death rates, and highest injury rates of all of the boroughs. Brooklyn also contained most of the complaints about traffic signal conditions.

My data provides a way to measure how many accidents there are per area. This is important to note, as this can help give some direction as to what areas need the most attention with street light conditions.

**References:**

Crash Data:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data 311 Data:

https://data.cityofnewyork.us/Transportation/DOT-Street-Lights-and-Traffic-Signals-311-Service -/jwvp-gyiq

Uploading my data from Python to Bigquery:

https://stackoverflow.com/questions/44838239/upload-to-bigquery-from-python/44851534

Tools used:

* Lucidchart to create the fact tables
* Python - Cleaned all my data and prepared it for uploading into the Google Big Query environment.
* Google Bigquery - I used Bigquery as a means to upload and manipulate the data.
* Google Data Studio - Created all of my data visualizations.