Analysis of Motor Collisions in NYC

Parv Khatri
pk2321@nyu.edu
New York University
Brooklyn, NY

Purva Kondaji
pk2312@nyu.edu
New York University
Brooklyn, NY

Tensaiye Zelealem
taz231@nyu.edu
New York University
Brooklyn, NY

ABSTRACT

The goal of this project is to analyse and present motor vehicle collisions data for the city of New York. We plan to analyse collisions frequency and the factors contributing towards such accidents. The initial phase involves cleaning the data using techniques like finding missing values, mismatched values, etc. and correcting data discrepancies. Data cleaning has been done using Spark and Python. The final phase involves analysing the cleaned data to gather useful insights and understand the factors that result in vehicle collisions as well as any relationships we can find that leads to an increase/decrease in vehicle collisions.

1 INTRODUCTION

The world has seen a massive growth of data in the recent years. New York City, being a densely populated city in the United States, generates a lot of data. The data collected from various sectors like the Transportation, Crime data, Banking, Education etc. is published in data lakes and made available to public. In this project, we plan to extract and analyze data for motor vehicle collisions in NYC available in one such data lakes called NYC OpenData. We will analyse the frequency of collisions based on the areas around NYC and their impact on people involved in the crash.

This data has been provided by the New York Police Department and contains the details of Vehicle crashes and their impact.The entire data contains around 6 million rows and hence, will be analysed using Spark framework for Big Data. Initially, we perform data wrangling and cleaning where we extract data from various sources and check for data discrepancies and missing and mismatched values in it. We also perform data profiling to examine the data available so that it can be transformed into a format that can be easily analysed to gather actionable and useful insights.

Once we clean the data, we will perform analysis on the various columns available in the data sets to understand the vehicle collision causes and effects on people and assets.

2 GITHUB REPOSITORY

The code for data cleaning and analysis can be found at the **Github Repository**.

3 DATA SETS USED

For this project we decided to work with the following data sets:

1. Motor Vehicle Collisions – Crashes: This data set contains information that has been reported by the Police Department of New York for motor vehicle collisions in New York City. All collisions that involve an injured or killed person are published in this data set. Also, if the collision damage is worth at least $1000, it is reported in this data set. The data dates back to the year 2016.
The total number of rows available are: 1,766,159
The total number of columns in the data set are: 29
Source of data: NYC OpenData
The dataset for Motor vehicle collisions - Crashes can be found at **Collisions - Crashes**.

2. Motor Vehicle Collisions – Person: This data set contains the information of people involved in the crash. Each row here represents a person (driver, occupant, pedestrian, bicyclist,..) involved in the crash. The data dates back to the year 2016.
The total number of rows available are: 4,236,940
The total number of columns in the data set are: 21
Source of data: NYC OpenData
The dataset for Motor vehicle collisions - Person can be found at **Collisions - Person**

3. Motor Vehicle Collisions – Vehicle: This data set contains details on each vehicles involved in the crash. The data goes as far back as April 2016.
The total number of rows available are: 3,559,738
The total number of columns in the data set are: 25
Source of data: NYC OpenData
The dataset for Motor vehicle collisions - Vehicles can be found at **Collisions - Vehicles**

4. Latitude and Longitude data: This data set contains city, zip code, longitude, latitude and other columns which will be used to map location of crashes.
The Latitude Longitude dataset can be found at **Locations**

4 DATA CLEANING AND INTEGRATION

For each column in the data sets, we first checked their data type to know if they were valid or invalid. For those that were invalid we found a more appropriate data type that better displayed the values and converted the original data type to the appropriate one. In our first data set which contains data for crashes, the following are the list of columns that had were changed to more appropriate data types:

CRASH DATE: from string to date
CRASH TIME: from string to timestamp
NUMBER OF PERSONS INJURED: from string to integer
NUMBER OF CYCLIST KILLED: from string to integer
NUMBER OF MOTORIST INJURED: from string to integer

For our second data set which contains person data, the following are the list of columns that had were changed to more appropriate data types:

CRASH DATE: from string to date
CRASH TIME: from string to timestamp

For our third data set which contains vehicle data, the following are the list of columns that had were changed to more appropriate data types:

CRASH DATE: from string to date
CRASH TIME: from string to timestamp

After that we proceeded to check the values of each column if they had any missing values or null values. Some columns did not require any cleaning and did not have any missing or null values. For those that did, we changed the null values to appropriate values instead of removing them from the data set. For example, when dealing with column 'VEHICLE TYPE 1' instead of removing the rows in which there are null values we instead converted all null values to 'Unknown' so that those rows can still be used for analysis later. If we removed a row for each null value, we saw that most of our data would be removed which is not ideal for analysing data and producing meaningful insights. We surmised that although null value does not tell us much about the current column, the other information on the row could be important, thus, we just converted the null values to more appropriate names.

Data profiling was done to make sure values aren't incorrect and don't create obstacles while analysing data. One such change was made for the column 'PERSON AGE' in the person data set where age was a negative value or more than 100 years. We modified the negative values and higher incorrect values and replaced them by -1 if the values are not between the range 0 and 100.

We also used scikit-learn in order to detect outliers in each column. Since different representation of the same values can be come in data sets, we needed to detect these to prepare our data appropriately. The methodology we used to identify different representations are clustering and violations of functional dependencies.

5 DATA ANALYSIS

In this section we will discuss how we analyzed our data sets in order to find solution to our main questions. Each data set has been carefully analyzed separately to get a better understanding of the relationship of each column to one another. Below we have provided some charts and graphs that will help us better portray and analyze the data.

5.1 Vehicle

The below chart shows the types of vehicles with the highest number of collisions. From this chart we are able to easily tell Sedans have the highest number of crashes followed by passenger vehicles.

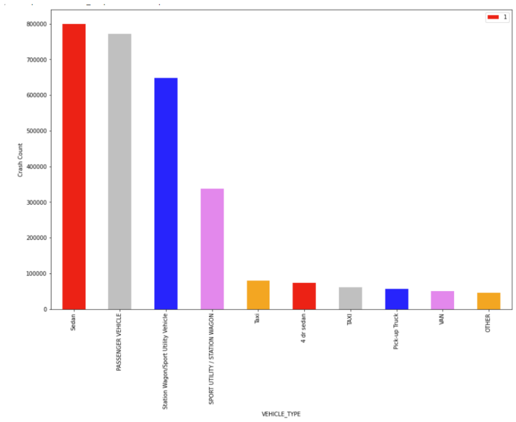

Figure 1. Collisions based on Vehicle Type

The graph below shows us the type of accidents that cause damage to public property in New York city. Going straight ahead, backing and making a left turn are the top three reason people.
In order to reduce the property damages government can take initiatives by adding more training modules to avoid these kind of accidents and can subsidise sensors cost so more vehicle owners can buy them and would be cautious before accidents happen. There are wide range of senors to assist drivers to not to sleep while driving, if they do then they start alarming the driver.Similarly for parking and backing camera module and proximity sensors are there to avoid accidents in such cases.
By taking such initiatives government can reduce losses from public property damages and save the citizens.

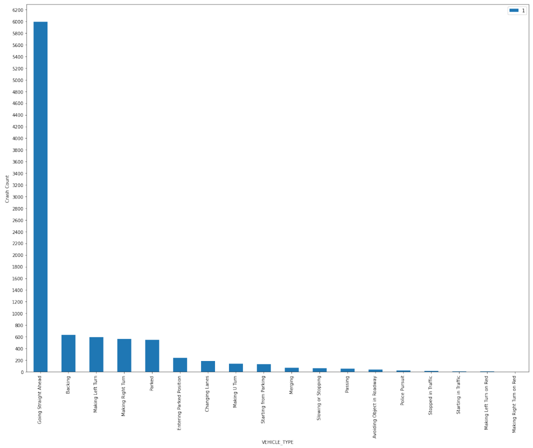

Figure 2. Accident type for public property damage

5.2 Collisions

As we can see from the graph below, The number of crashes differ greatly based on the boroughs.Brooklyn takes the first place for highest amount of crashes just above 350,000. Queens is at second place at above 300,000 followed by Manhattan which is slightly below 300,000.

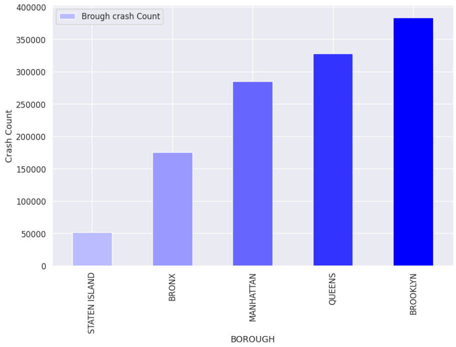

FIGURE 3. Crashes based on boroughs

5.3 Person

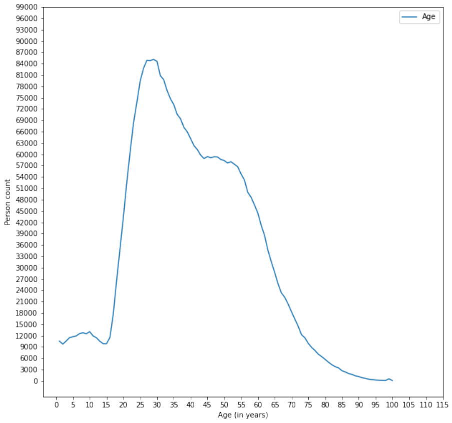

FIGURE 4. Age of person being affected by the crash

The above chart shows the age of people who are involved in car collisions. Those that are ages from 25-30 are involved in car crashes at a much higher frequency that the other age groups.

5.4 Does the amount of cars on the road affect collisions?

Figure 5 shows the number of people killed during a crash versus the number of people injured overs span of 8 years. The number of people killed during a crash has stayed consistent under 400 throughout 8 years. But the number of people injured had an increase from 2012 until about the beginning of 2020. In 2020, you can see a sudden decrease in collisions reaching its lowest point in 8 years. Through research and analysis we found the reason for this is due to the current Corona virus pandemic. During the beginning of 2020, the corona virus outbreak had occurred and there was a mandatory lockdown initiated in New York City. This decreased the number of vehicles being driven daily which in return decreased the number of collisions.

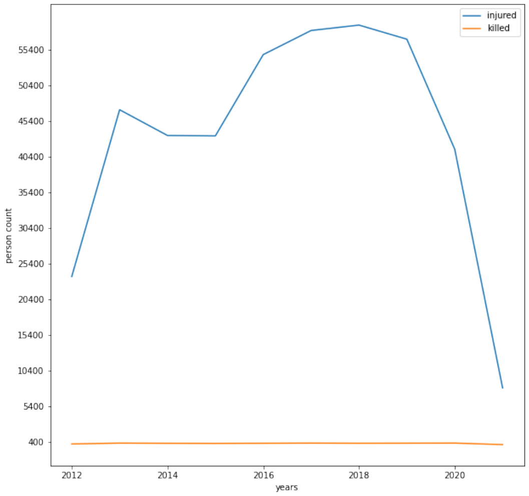

FIGURE 5. Number of injured and killed over time

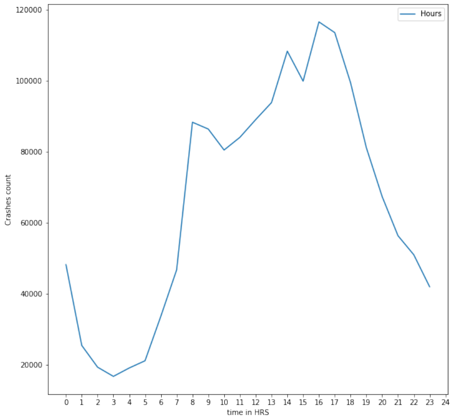

FIGURE 6. Amount of collisions at different times of the day

Figure 6 shows us that during certain hours of the day car crashes tend to peak. As you can see in the graph around 8:00AM -9:00AM crashes start to increase afterwards the decrease for a little bit before they spike even higher at around 2:00PM then eventually at 5:00pm they reach their peak. This is because those are times when people usually go to work or leave work.

In conclusion, we saw that during times where their were an increase in amount of cars their were also an increase in car collisions.

5.5 What are the major factors contributing towards the collisions?

In the graph below, we can see that the major contributing factors for collision is Driver not being attentive at the time of the accident or being distracted by some factors. There is a possibility that the driver might get distracted due to other people in the vehicle or due to some event happening in the surroundings. The second major reason for collisions is the failure to yield right of way, followed by the driver following another vehicle too closely or being followed by some other vehicle with very short distance between the two vehicles. The other reasons for collisions could be improper turning of vehicle, driver being fatigued or drowsy, overtaking another vehicle too closely, changing the lane in an unsafe manner, etc.

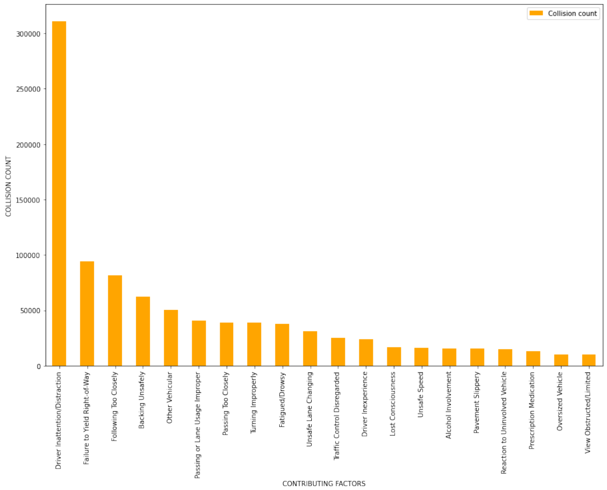

Figure 7. Major contributing factors of car collisions

5.6 Impact on public property and reasons for the public property damage?

From the crash records, we found the top 5 reasons for public property damages by crashes. From Figure 8: Top Contributing reasons for public property we can see that the major reason for crash are drivers inattentiveness/distraction, Unsafe Speed, Alcohol Involvement, Backing Unsafely, Inexperienced driver. The type of property that majorly got damaged are Fences,Light Poles, Utility Poles, Fire Hydrants, trees.

Government can focus on reducing the public property damage cost by taking care of public property positioning and add additional protection to highly prone public properties. Also, there is need to mitigate crashes by considering reasons like Unsafe speed, Alcohol Involvement etc.

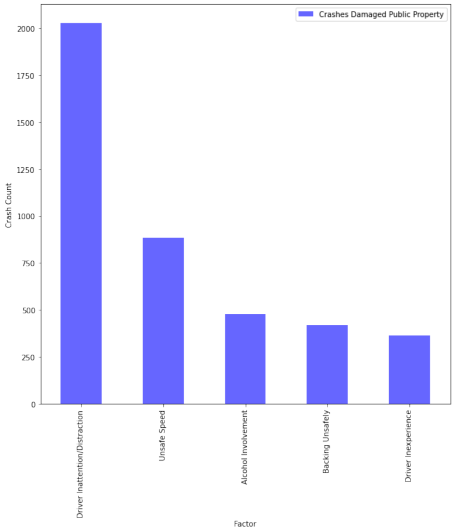

Figure 8. Top contributing reasons for public property damage

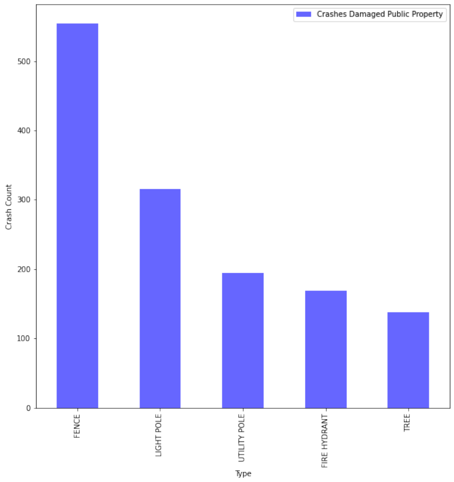

Figure 9. Type of public property damaged

5.7 Does injury to the head account for the main reason for death?

In this sub-section we want to find the main type of injury in car crashes that cause death.

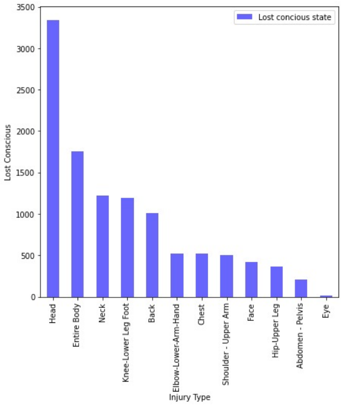

Figure 10. Top reasons for loss of consciousness

We were able to build the graph above to see the main reason for loss of consciousness in car accidents is head injuries. This still does not help us find if head injuries actually lead to death. Since lost consciousness does not imply death we need to look at the relationship between car accident deaths and the injuries incurred.

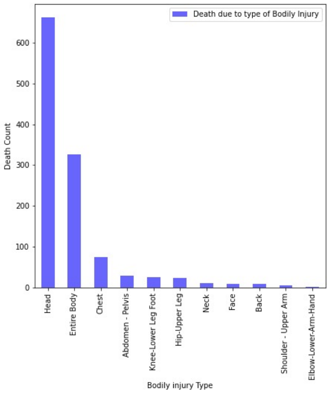

Figure 11. Top injuries during accidents that cause death

The graph above effectively shows the main injury that causes death during car collisions. We can conclude by saying that injury to the head does account for the main injury reason for death in car collisions.

5.8 Spatial Analysis

The map below depicts New York City. This map shows the areas car collisions occurred. The area with dense clusters represent a higher number of car crashes while the area with less dense clusters represent a relatively lower number of car crashes.

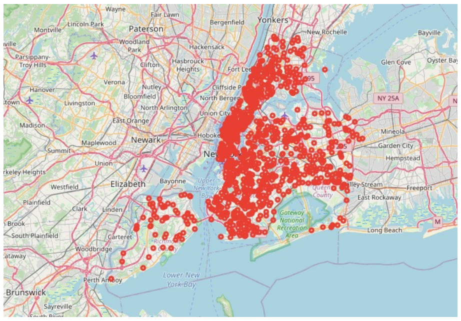

Figure 12. Collisions in New York City

The map below shows the number of car collisions in Manhattan. The legend ranges from a dark red which represents a relatively lower number of collisions to a dark blue which represents a relatively high number of collisions. This map was built using longitudes and latitudes provided during each crash.

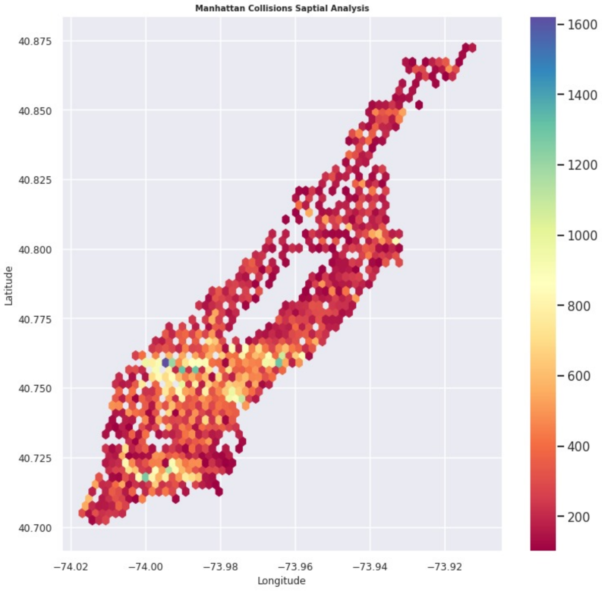

Figure 13. Collisions in Manhattan

5.9 Trend of collisions according to borough

The following are trends we saw in each borough in New York City.The Figure in this sub-section represent the relationship between time and collisions in the Bronx, Brooklyn,Staten Island, Queens and Manhattan. The following are some of the common thing we can see in all the boroughs: They all have show that 4:00PM - 5:00PM is when the collisions are at their peak everyday. They also show that collisions start to increase after 6:00AM. Lastly, we can see a steady decrease in collisions after 7:00PM.

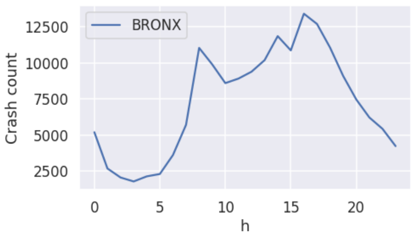

FIGURE 14. Bronx: Collisions vs Time

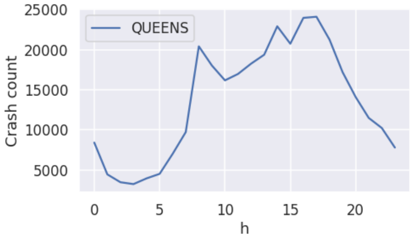

FIGURE 15. Queens: Collisions vs Time

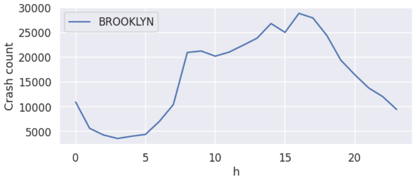

FIGURE 16. Brooklyn: Collisions vs Time

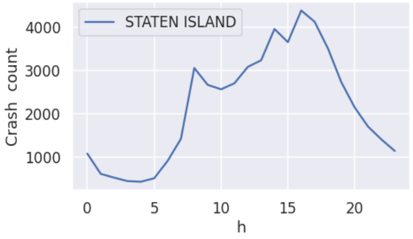

FIGURE 17. Staten Island: Collisions vs Time

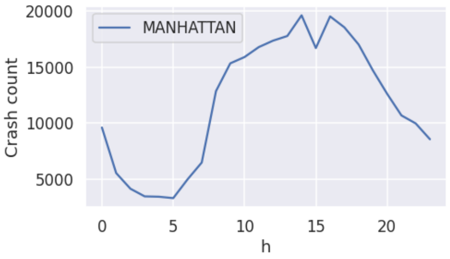

FIGURE 18. Manhattan: Collisions vs Time

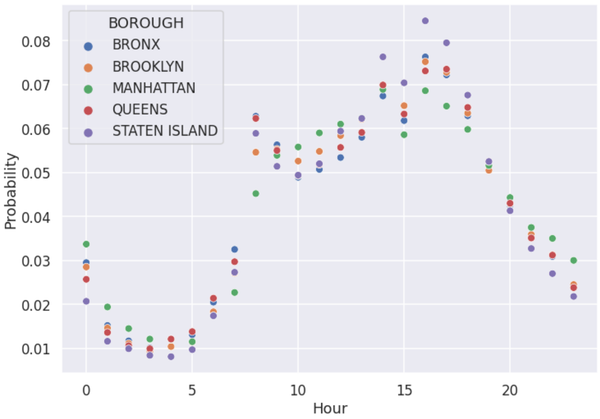

FIGURE 19. Trend of collisions in all boroughs

6 CHALLENGES FACED

1. Finding Geo location values for which some attributes were missing like Zip code, Borough, Latitude, Longitude etc. without which it is difficult to find the location on a map.
2. Some values for columns were not making any sense like 1, 80, -,etc in Vehicle collision contributing factors and other columns.
3. Age of people affected by the crashes was a negative value or very large value which can never be true.
4. Plotting a Map using Plotly was challenging due to rendering engine issues we faced. Then we switched to basemap library and we were unable to install the dependencies. Finally, we used Folium for maps and markers.
5. Debugging the errors at the time of creating graphs from Spark SQL Dataframes were hard, we converted dataframes to pandas dataframe and then we plotted the information .
6. Converting columns with date and time and formatting it according to the types of the graphs sometimes returned the values like Hours, Year and whole date object when it was not needed .

7 CONCLUSION

In conclusion, we answered the questions we had regarding collisions and we provided detailed analysis to back up our hypothesis. We have surmised the peak times at which car collision occurs is during rush hour when people usually leave their office or when they enter their office in the morning. We can inform people about this so they will be more careful at that time to avoid crashes. We have also found the leading contributing factor of collisions in New York City is drivers inattention and drivers distraction. This information is very important. By making people more aware of the leading factor of crashes we might be able to decrease the number of crashes caused by a drivers inattentiveness. We also found that car collisions that result in head injuries have a higher chance of causing death than any other type of injuries. Thus, we should find ways in protecting the heads of the passengers during collisions.

REFERENCES

[1] Data Cleaning,https://github.com/VIDA-NYU/openclean

[2] Geo-Pandas,https://ncar.github.io/PySpark4Climate/tutorials/pyspark-geo-analysis/geopandas-and-spark/

[3] Collision: Crash, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

[4] Collision: Vehicle, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

[5] Collision: Person, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu

[6] Pyspark: Conversion,https://sparkbyexamples.com/spark/spark-sql-how-to-convert-date-to-string-format/

[7] Pyspark,https://spark.apache.org/docs/latest/api/python/index.html

[8] Seaborn,https://seaborn.pydata.org/

[9] Mathlib,https://matplotlib.org/