CS 732: Data Visualization Assignment 3 Report

M Srinivasan
IMT2021058
IIIT Bangalore
m.srinivasan@iiitb.ac.in

Siddharth Kothari
IMT2021019
IIIT Bangalore
siddharth.kothari@iiitb.ac.in

Sankalp Kothari
IMT2021028
IIIT Bangalore
sankalp.kothari@iiitb.ac.in

***Abstract*—Assignment 3 report made by group "Are you able to visualize?" for the course CS-732 Data Visualization. This assignment intends to use Visual Analytics to provide us with information as to which factors can cause a change in the number of accidents observed per month. We first observe the trends of number of accident per month, and then supplement our data using a number of datasets to explore contributing factors to these accidents. We mainly explore weather based, demographic and geographical factors.**

***Index Terms*—Visual Analytics, Visualization, Machine Learning and Clustering, Statistics**

I. Dataset

A. Motor Vehicle Collision Dataset-NYC

The given dataset is about all the motor vehicle accidents in NYC from the year 2012-23 [1]. In the previous assignments, we considered only the data from 2020-23. However, for this assignment, we consider all the years. Some of the important columns in the dataset are listed below:

- Date of collision
- Time of collision
- Borough of New York City in which the accident had taken place.
- The latitude and longitude of the location where the accident took place.
- Number of people injured and killed (pedestrians / motorists / cyclists)
- Major contributing reason to the accident.
- Street where the accident occurred in the borough.
- Vehicle types involved in the accident.

A grouping similar to the one used for assignment 1 for the accident categories has been used for this assignment as well.

Apart from this dataset, to support our inferences, and to draw new conclusions, we have supplemented the dataset with many additional data sources.

B. US Weather Dataset

This repository contains a comprehensive collection of weather events data across 49 states in the United States. The dataset comprises of about 8.6 million events, ranging from regular occurrences like rain and snow to extreme weather phenomena such as storms and freezing conditions. The data spans from January 2016 to December 2022 and is sourced from 2,071 airport-based weather stations nationwide. [2]

We have filtered out and used the data only for New York City. We have used this dataset to supplement the existing Motorcycle accidents dataset to further investigate upon the weather conditions and their co-relation with the accidents happening in the city.

C. Monthly Precipitation and Temperature Data in NYC

The rainfall dataset represents the rainfall in inches in each of the months ranging from January 2012 to December 2019.

The data contains 3 columns - Date, which is the month pertaining to the data, Value, which is the precipitation value, and Anomaly. We only use the first 2 columns for our analysis.

The temperature dataset represents the average temperature in fahrenheit in each of the months ranging from January 2012 to December 2019.

The data contains 3 columns - Date, which is the month pertaining to the data, Value, which is the temperature value, and Anomaly. We only use the first 2 columns for our analysis.

We had 3 regions in New York City (JFK, La Guardia and Central Park) from which we could use the data, and we picked all three of them and averaged them out for our analysis. [3]

D. Daylight NYC Dataset

Another weather based feature which we considered for our analysis is the Daylight Info dataset [4], which contains the number of daylight hours per day from 2012-2019. We included this as we feel that this can well explain some of the trends observed in accident types.

E. NYC Motor Vehicle Collisions - Person Dataset

In order to more accurately observe the effects of the demographics of the population on the accidents, we decided to pick a dataset pretty similar to our original dataset.

This dataset is based on the people involved in the accidents in New York City, rather than the crashes [5]. This allows us to fully investigate the various demographic factors which can lead to the number of accidents observed.

II. Tasks

We have a broad objective for the analysis - to observe monthly trends in the accident numbers, and observe a variety of factors which contribute to these trends. The factors considered are as follows -

1) Weather info. The main factors considered are - Precipitation, Temperature, Daylight Hours, Snow, and Storms.
2) Demographics info. The main factors considered here are sex and age of the drivers.

3) Geographic factors. For this, we form clusters of the data, and study the clusters to be able to explain why certain areas may have greater accident numbers.

III. Libraries

We have used Matplotlib, Folium and Geopandas libraries of Python for all our visualizations.

IV. Visualization

A. Initial Step

We first plot the number of accidents per month for each of the years from 2013-2023. The results for the same are observed in Figure 1.

Two of the plots for the accident types have been shown here - for Distractions (Figure 2) and Medical/Fatigue (Figure 3). The others have been added to the images folder.

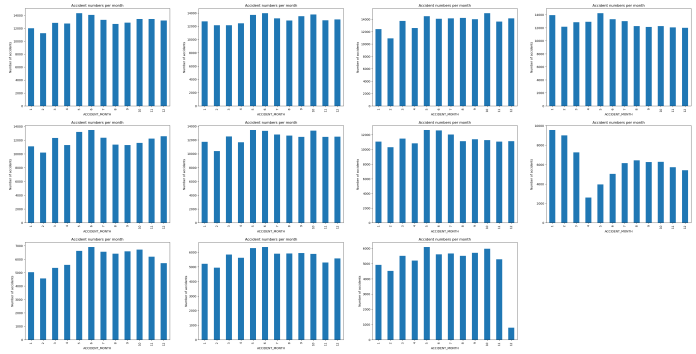

Fig. 1. Monthly trends of accidents

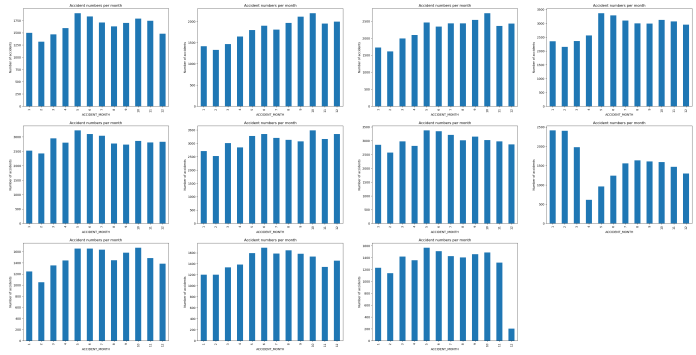

Fig. 2. Monthly trends of accidents caused by Distractions

We immediately observe the following through the plots -

1) There is a clear trend observed across each of the years, regardless of the actual numbers. The trend has some distinct features. (Figure 1)
2) 2020 is a clear outlier to any trend due to the onset of the Covid 19 pandemic and the subsequent lockdown. This is clearly visible by the sudden dip in the number of accidents in April 2020 when compared to March 2020. This is visible in all the figures (1, 2, 3).
3) In almost all the years, there is a clear decrease in the number of accidents in February, followed by a very sharp increase in May and June. May, June and October are almost always the months with the highest number of accidents.
4) The various accident types also follow a similar trend. (Figure 2) The only exceptions to this trend are the accidents due to medical/fatigue reasons. (Figure 3)
5) This seasonal distribution of the accidents suggest an influence of the weather conditions. Hence we began to analyse various weather conditions which can lead to these trends.

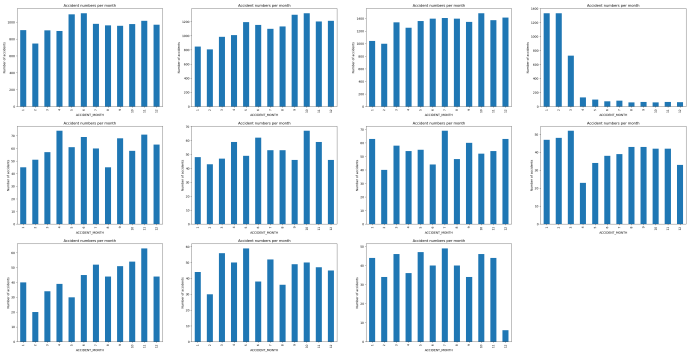

Fig. 3. Monthly rends of accidents caused by Medical/Fatigue factors

B. First Step

From the initial step, we now further analyse the effect of weather conditions of the city during the accident dates. We first filtered out the US Weather Data [2] and only kept the data for New York city. We then performed the merge of this new dataset [2] with the existing dataset [1]. We performed the merging of two datasets based on the ‘accident date‘ attribute, following which we did a grouping on the basis of ‘Accident Month’ (retrieved from Crash Date), and the kind of weather present during that particular day.

We have then plotted the accidents month wise for each of the year from 2016-2019, the colours in the stacked bar chart represent the kind of weather during that accident and the y-axis represents the number of accidents.

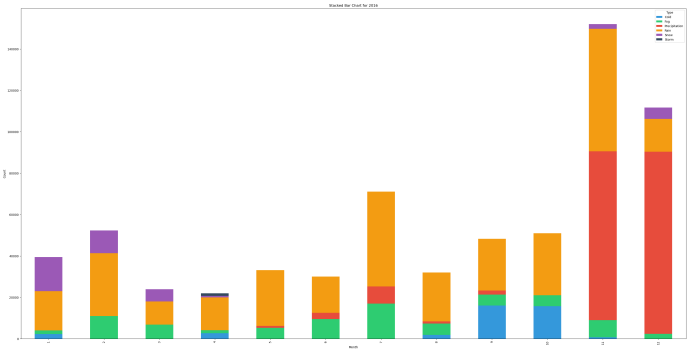

Fig. 4. Monthwise Accidents for 2016 along with the weather conditions

Figure 4 shows the accidents happened in 2016 monthwise and along with that it shows the weather on that particular

day(colour of the stacked bar chart).

A major portion of accidents in this year shows that rain has been one of the contributing factors across all the months and snow which is a common weather condition during the cold months does contribute in smaller quantities to the accidents taking place in the colder months(November to March).

Fog is also one factor that has been present all throughout the year in varied amounts across the months.

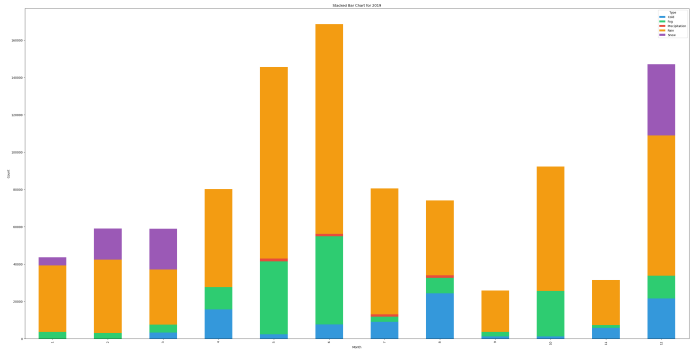

Fig. 5. Monthwise Accidents for 2019 along with the weather conditions

Figure 5 shows the accidents happened in 2019 (the same representation as Figure 4 but for 2019).

A major portion of accidents in this year also shows that rain has been one of the contributing factors across all the months and snow has been one of the major contributors second only to rain during the cold months of New York as expected. The spring months of the city also has more accidents occurring during the presence of fog. The other years show similar trends and their images have been added to the images folder.

Based on these images, we are led to conclude that precipitation plays a constant role in accidents, while the other factors have very fluctuating correlations with the number of accidents, and are as such not good indicators of the accident trends.

However, very often, correlation does not imply causation. Hence in the second step, we further analyse the trends of rain, and see whether the levels have a correlation with the number of accidents.

Along with this, we also begin to analyse other possible factors which can explain these trends, such as temperature and daylight hours.

C. Second Step

In this step, we now begin to analyse the levels of precipitation, temperature and the daylight hours, and try to find a correlation between the accident numbers and these indicators.

For this, we first found datasets pertaining to these factors. The precipitation and temperature datasets contain monthly data of precipitation levels (in inches) and the average temperature (in Fahrenheit) for 2012-2019. This data is obtained from 3 regions in New York City (JFK, La Guardia and Central Park) and have been averaged out for the analysis. [3]

We also found a dataset pertaining to the daylight hours per day, starting from 2013 to 2019 [4]. To carry out a monthly analysis, we have averaged out the daytime hours per month.

Equipped with this data, we begin analysing per year, whether we can find a steady correlation between these factors and the accident numbers. We attach the plots for the years 2018 and 2019 here, while the others have been added to the images folder.

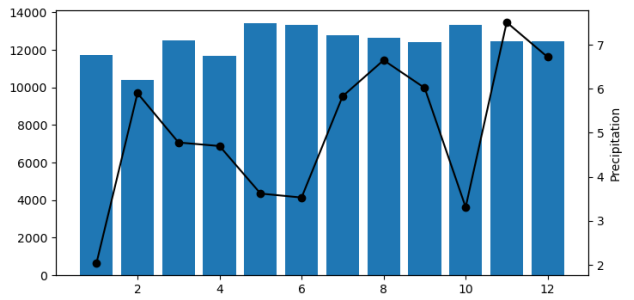

Fig. 6. Monthwise Accidents for 2018 along with the Precipitation levels

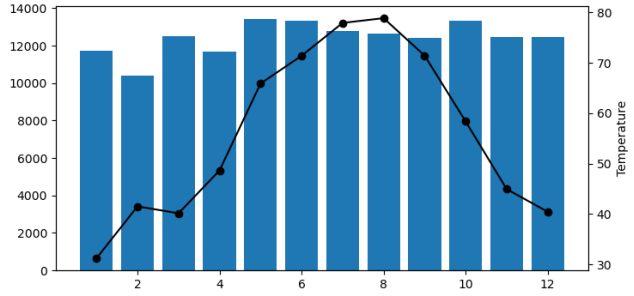

Fig. 7. Monthwise Accidents for 2018 along with the Temperature levels

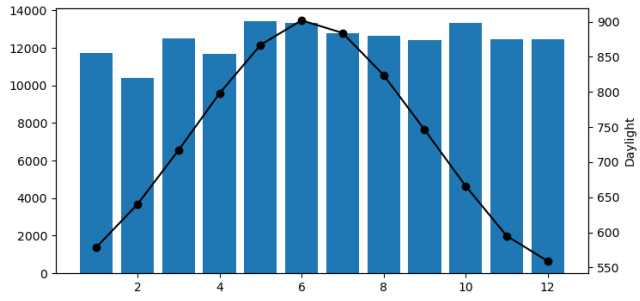

Fig. 8. Monthwise Accidents for 2018 along with the Daylight hours

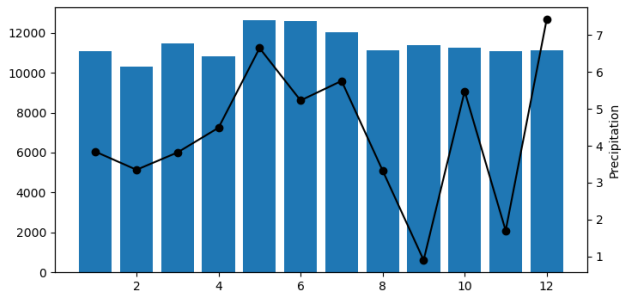

Fig. 9. Monthwise Accidents for 2019 along with the Precipitation levels

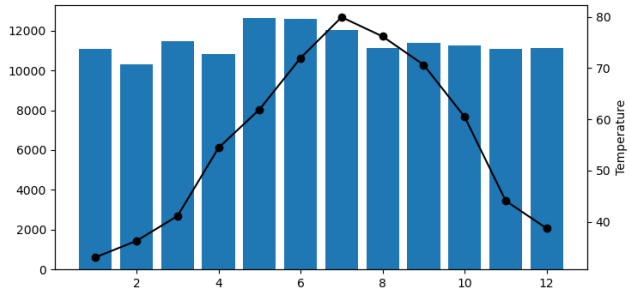

Fig. 10. Monthwise Accidents for 2019 along with the Temperature levels

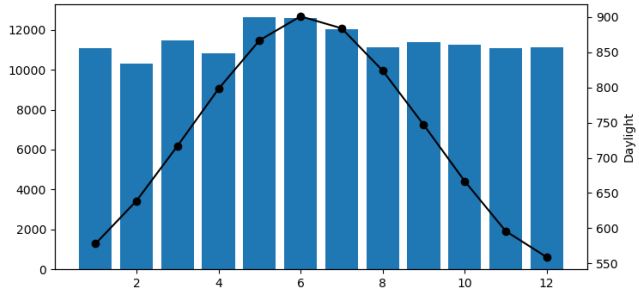

Fig. 11. Monthwise Accidents for 2019 along with the Daylight conditions

We have tabulated the results of the correlation values for each year and factor as shown in Table I

Through the table I and through the Figures 6, 7, 8, 9, 10, 11, we observe the following -

1) Precipitation has a very fluctuating relationship with the accident numbers. For certain years, (2013,2017,2019), it shows a high positive correlation with the accident numbers, for some years (2015, 2016), there is almost no correlation between the rain levels and the number of accidents, while for other years (2014,2018), a negative correlation is observed. (Table I)
2) These values lead us to conclude that precipitation is, in general, not a good factor in New York City to be able to determine the accident numbers.

TABLE I
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR

Year	Factor	Correlation Value
2013	Precipitation	0.4546577084604513
2013	Temperature	0.5339582587066065
2013	Daylight	0.4781428213638841
2014	Precipitation	−0.3204701747545821
2014	Temperature	0.686560119382722
2014	Daylight	0.4007662217819287
2015	Precipitation	0.13037009246379044
2015	Temperature	0.7316800892213952
2015	Daylight	0.34544744354149076
2016	Precipitation	0.10051308578791764
2016	Temperature	0.012776674056908794
2016	Daylight	0.4117139973729996
2017	Precipitation	0.4515321790367224
2017	Temperature	0.2163162435932823
2017	Daylight	0.4187806308493553
2018	Precipitation	−0.19186309268451623
2018	Temperature	0.5628329723171775
2018	Daylight	0.45704807426151833
2019	Precipitation	0.4175562666823352
2019	Temperature	0.5722216033424102
2019	Daylight	0.6967194630225572

3) Temperature and Daylight, on the other hand, show promising signs of being strong factors in determining the number of accidents occurring in a month, as can be seen by the strong positive values of correlation for each of the years. (Table I)
4) The only exceptions to this analysis are the years 2016 and 2017, where no proper relation is observed between the temperature and the number of accidents. However, as we shall observe later, some months of 2016 are anomalies in other analysis as well. This can indicate the influence of certain factors only in that year, which we may be unaware of (similar to the effect the Covid 19 lockdown had on the accident numbers). (Table I)

This analysis leads us to conclude that Temperature and Daylight are strong factors in influencing the number of accidents. To further confirm our findings, we perform this analysis again on the data, but the data is now subdivided into 5 time intervals -

1) Midnight (12 am to 6 am)
2) Morning (6 am to 12 pm)
3) Afternoon (12 pm to 4 pm)
4) Evening (4 pm to 8 pm)
5) Night (8 pm to 12 am)

D. Third Step

In this step of the workflow, we now analyse the effect that temperature and daylight have on the accident numbers in a day. We also include precipitation once again in this analysis to be able to compare and contrast our results.

The methodology used for generating the visualisations is the same as the previous, except that we now filter the accidents before observing correlations.

We have attached plots for the temperature and daylight hours for the year 2019, for each of the time intervals. Images

for Precipitation have not been included due to the main focus being the Daylight and Temperature. However, we do make tables for each of the time intervals where precipitation correlation values have been included.

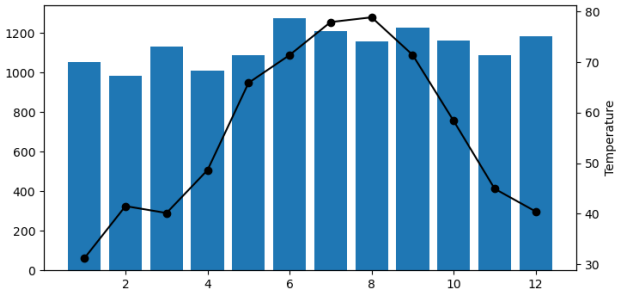

Fig. 12. Monthwise Midnight Accidents for 2018 along with the Temperature levels

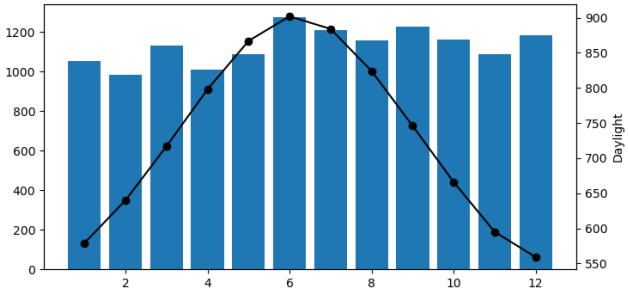

Fig. 13. Monthwise Midnight Accidents for 2018 along with the Daylight hours

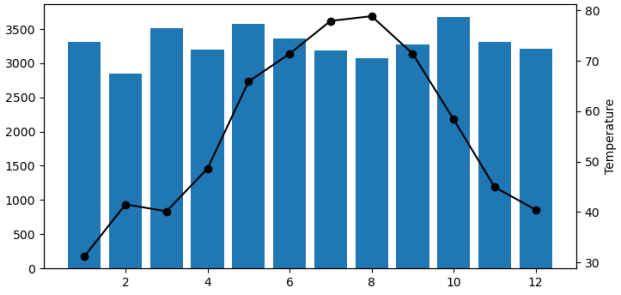

Fig. 14. Monthwise Morning Accidents for 2018 along with the Temperature levels

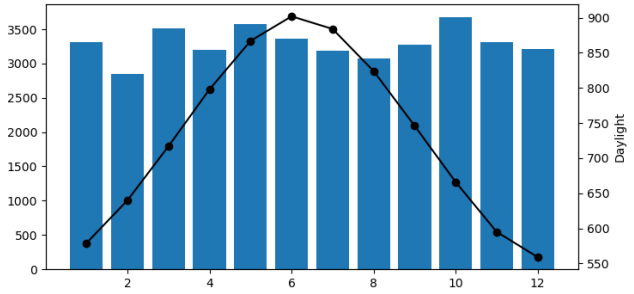

Fig. 15. Monthwise Morning Accidents for 2018 along with the Daylight hours

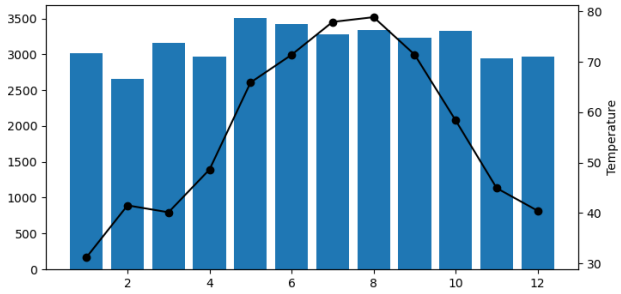

Fig. 16. Monthwise Afternoon Accidents for 2018 along with the Temperature levels

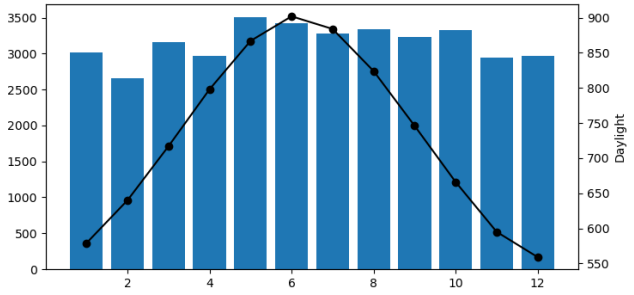

Fig. 17. Monthwise Afternoon Accidents for 2018 along with the Daylight hours

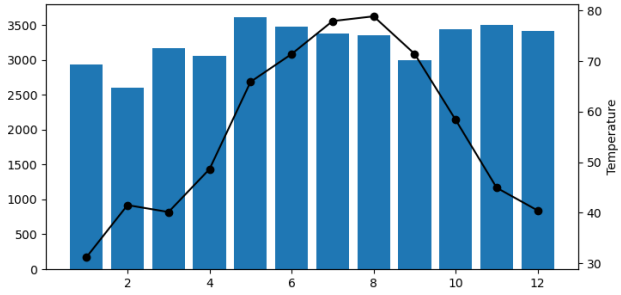

Fig. 18. Monthwise Evening Accidents for 2018 along with the Temperature levels

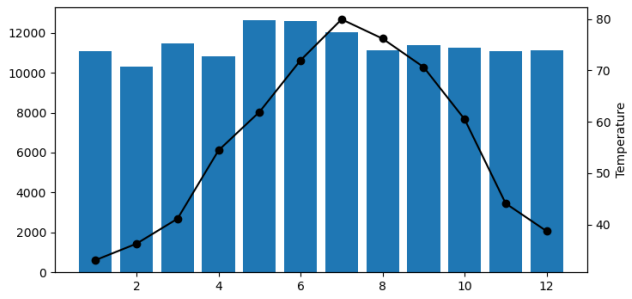

Fig. 19. Monthwise Evening Accidents for 2018 along with the Daylight hours

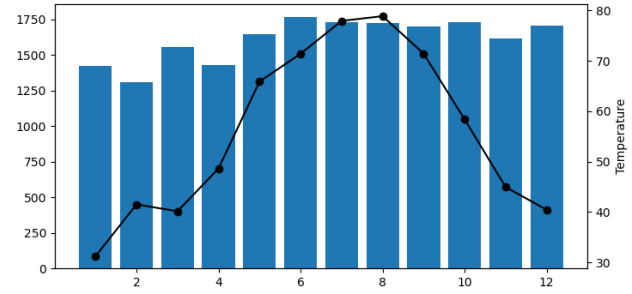

Fig. 20. Monthwise Night Accidents for 2018 along with the Temperature levels

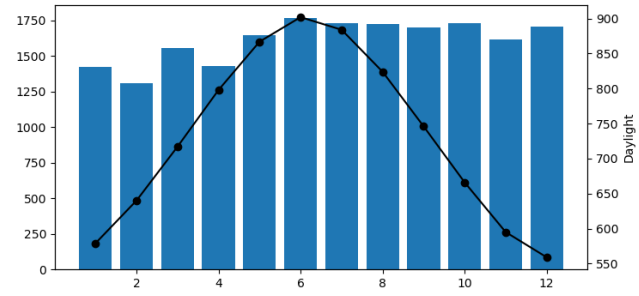

Fig. 21. Monthwise Night Accidents for 2018 along with the Daylight hours

Based on the tables and the figures, we conclude the following -

1) The tables II, IV, V, VI show that the accidents in the Midnight, Afternoon, Evening and Night follow a trend similar to the overall accident trends. They too have a high positive correlation with the temperature and daylight hours, for each year, while precipitation continues to be fluctuating in its relationship with the number of accidents.
2) In all these 4 tables, we observe that 2016,2017 continue to be anomalies to the trend.
3) We gain one insightful observation from this step - the accidents in the morning time deviate highly from the otherwise established trends. Many a times, a negative or almost zero correlation is observed between the daylight hours, temperature levels and the accident numbers. (Table III)
4) This goes to show that these trends apply to most, but not all of the data, and generalisations must be made carefully.
5) In the later workflows, we shift our focus from weather based factors to demographic and geographic factors, and try to observe correlations using those features.

TABLE II
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR FOR MIDNIGHT ACCIDENTS

Year	Factor	Correlation Value
2013	Precipitation	0.7064813250169413
2013	Temperature	0.5431136894225498
2013	Daylight	0.4868214706385231
2014	Precipitation	−0.007525847919453359
2014	Temperature	0.49944828472516306
2014	Daylight	0.2094590852251228
2015	Precipitation	−0.10101389153860882
2015	Temperature	0.8053690813895327
2015	Daylight	0.34790282979653836
2016	Precipitation	0.49522113240177057
2016	Temperature	0.5273903906877158
2016	Daylight	0.36028570346897837
2017	Precipitation	0.36720082482666516
2017	Temperature	0.25009910653041906
2017	Daylight	0.4099547830818516
2018	Precipitation	0.0756401299076366
2018	Temperature	0.623237621496069
2018	Daylight	0.38103941020277915
2019	Precipitation	0.14794735298332604
2019	Temperature	0.5431715417382577
2019	Daylight	0.4858279255047328

TABLE III
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR FOR MORNING ACCIDENTS

Year	Factor	Correlation Value
2013	Precipitation	0.08672277738455672
2013	Temperature	−0.24524112850765842
2013	Daylight	−0.16281486814308538
2014	Precipitation	−0.4167105593880939
2014	Temperature	−0.3029747610985446
2014	Daylight	−0.3482514600767481
2015	Precipitation	0.35082275324175055
2015	Temperature	−0.023790967651436425
2015	Daylight	−0.25442231615006505
2016	Precipitation	−0.06796009156899073
2016	Temperature	−0.49019868060944133
2016	Daylight	−0.08642549616978625
2017	Precipitation	0.22880289786183147
2017	Temperature	−0.18336515938262749
2017	Daylight	−0.04205164248032423
2018	Precipitation	−0.5352412841102921
2018	Temperature	0.04506822257153796
2018	Daylight	0.10790048487131691
2019	Precipitation	0.14703699409145954
2019	Temperature	−0.004437477964179068
2019	Daylight	0.2198252958825902

TABLE IV
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR FOR AFTERNOON ACCIDENTS

Year	Factor	Correlation Value
2013	Precipitation	0.3530296018290465
2013	Temperature	0.5596041034939421
2013	Daylight	0.5327534068173387
2014	Precipitation	−0.29618263490089364
2014	Temperature	0.7074352019862964
2014	Daylight	0.6076961595043188
2015	Precipitation	0.21732924616356308
2015	Temperature	0.6853516043709585
2015	Daylight	0.42033597221976376
2016	Precipitation	0.10631304873212832
2016	Temperature	0.07445307445828034
2016	Daylight	0.47341741953067523
2017	Precipitation	0.5310399420337727
2017	Temperature	0.20446920120193604
2017	Daylight	0.44312124218387866
2018	Precipitation	−0.3411604489713847
2018	Temperature	0.7390734646439807
2018	Daylight	0.7030384896923172
2019	Precipitation	0.2065421782239833
2019	Temperature	0.6719528796788914
2019	Daylight	0.8461211599525158

TABLE V
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR FOR EVENING ACCIDENTS

Year	Factor	Correlation Value
2013	Precipitation	0.4702152052120923
2013	Temperature	0.6879479039837523
2013	Daylight	0.5751465359582213
2014	Precipitation	−0.21419839025251194
2014	Temperature	0.8239124196701807
2014	Daylight	0.5650101228730179
2015	Precipitation	0.010419043963706787
2015	Temperature	0.8276207829506816
2015	Daylight	0.5026288521979808
2016	Precipitation	0.05403597991288825
2016	Temperature	0.0326030468031104
2016	Daylight	0.48054280829551815
2017	Precipitation	0.5120073522684374
2017	Temperature	0.2792013646686929
2017	Daylight	0.505006430512172
2018	Precipitation	0.0377753233455735
2018	Temperature	0.4320747671890667
2018	Daylight	0.3351592129196695
2019	Precipitation	0.5569225182532576
2019	Temperature	0.5529433886518881
2019	Daylight	0.6762539497351767

TABLE VI
CORRELATION VALUES BETWEEN THE FACTOR AND THE YEAR FOR NIGHT ACCIDENTS

Year	Factor	Correlation Value
2013	Precipitation	0.39522336747532266
2013	Temperature	0.6617831330890132
2013	Daylight	0.5333171043893498
2014	Precipitation	0.08025515950590534
2014	Temperature	0.7685050679239125
2014	Daylight	0.435368086711856
2015	Precipitation	0.04476501498780864
2015	Temperature	0.8387412462173723
2015	Daylight	0.3794032626788023
2016	Precipitation	−0.06608570074630299
2016	Temperature	0.3753515870461533
2016	Daylight	0.6895664258069499
2017	Precipitation	0.30047122031274254
2017	Temperature	0.6095263220109464
2017	Daylight	0.6881324945293475
2018	Precipitation	0.1568778312633951
2018	Temperature	0.7000193489159856
2018	Daylight	0.39748209863365924
2019	Precipitation	0.598056628930444
2019	Temperature	0.6328844312639671
2019	Daylight	0.5549553292630716

E. Fourth Step

In this step, we will try and see whether the demographic data of the people involved in the crash, such as age, gender etc, can tell us something or point out some interesting observations that we may not be able to infer from the crash data itself. For this we use the NYC Motor Vehicle Collisions - Persons dataset [5]. The data used is limited to only 2016-19, which isn't as broad as the other datasets used in this assignment, but provides a good enough understanding of the data none the less. This is done as the dataset does not well document the accidents in the years before 2016 and after 2019.

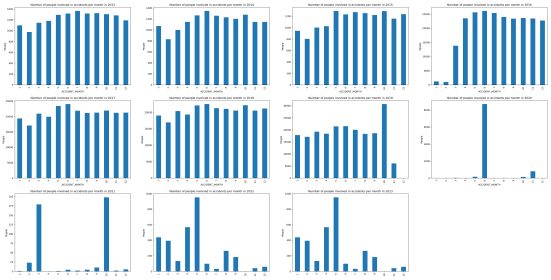

Fig. 22. All people involved in accidents

We will first look at the gender of the people. Figure 22 shows us the numbers of all people involved in accidents in the 2016-19 period, whereas the figures 23, 24 show the number of men and women involved in accidents during the same period. As we can see, the numbers of men and women involved in the accidents is different, but the distribution is more or less the same. The difference in numbers can be attributed to the actual number of male vs female drivers. But the overall distribution remains roughly the same. We can thus conclude that gender is no real factor affecting the number of accidents. Here, we have also debunked the popular sexist myth that women are bad drivers!

We now look at how age affects the number and distributions of accidents. Figure 25 shows people involved in accidents of various age groups as a stacked bar chart, while figures 26, 27 show 2 different age groups → 25-35 and 55-65 respectively. Again, as was the case with gender, although the

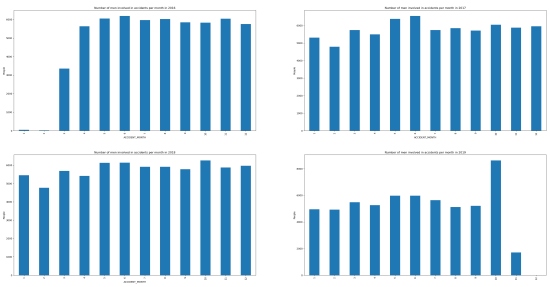

Fig. 23. Men involved in crashes from 2016-19

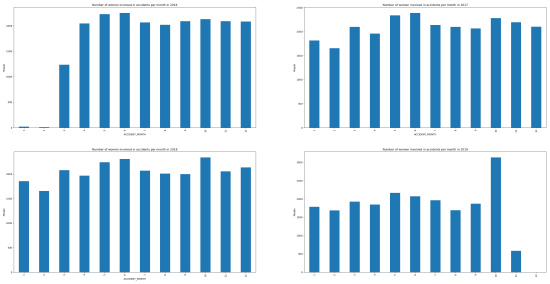

Fig. 24. Women involved in crashes from 2016-19

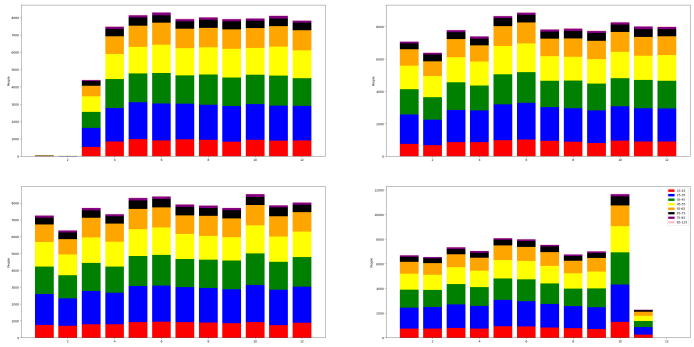

Fig. 25. People of different ages involved in accidents

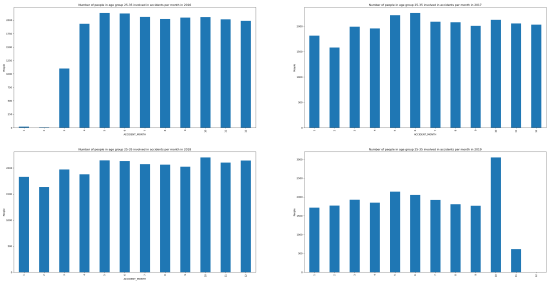

Fig. 26. People aged 25-35 involved in accidents

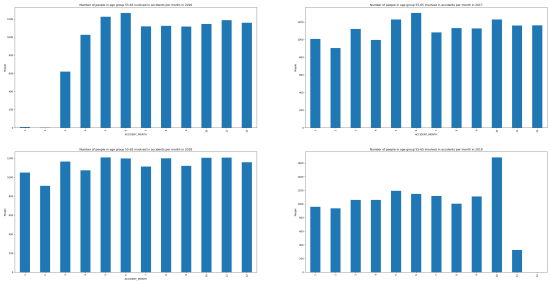

Fig. 27. People aged 55-65 involved in accidents

number of people involved in the accidents is different (due to the actual populations of those age groups and how often they drive), but the distributions remain largely the same, indicating that the age of a person, same as gender, does not really affect their chances of being involved in accidents.

Following this analysis, we now move on to analyse the final set of features, and see how the geography of New York City can influence the accidents happening in certain months.

F. Fifth Step

In this final step, we begin with the analysis of the clusters formed using the longitude and latitude data of the crashes dataset. Using those clusters formed, we will try and infer the causes of why certain areas have larger number of accidents.

The clusters represent hotspot maps that can represent the number of accidents happening in an area and its vicinity.

We used the K-Means Clustering Algorithm on the latitude and longitude data to get nearly 8-9 clusters for each month's data for the years between 2013-19. We first simply outlined the various boroughs with a color coding as follows -

- Brooklyn - blue
- Bronx - red
- Manhattan - green
- Queens - purple
- Staten Island - yellow

For the outline coordinates, we referred to - [6], and the data for the same has been integrated using a python library called geopandas [11].

Then we plotted the clusters based on the centres (means) returned by the k-means algorithm, with a circle around each, the radius of which represents the size of the cluster (number of data points in the cluster).

We have attached a few of the figures (Fig 28-34) here which we will refer to in order to show a most of our conclusions, the others are present in the images folder of the assignment submission. Inferences from this data are written later in the inferences section.

This marks the end of our analysis. We now shift our focus on the various insights gained into the data, and the monthly trends, and how we used visual analytics for the same.

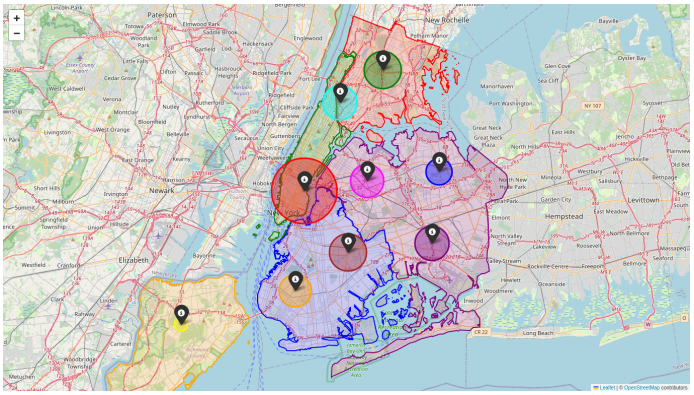

Fig. 28. Hotspot map for April 2016

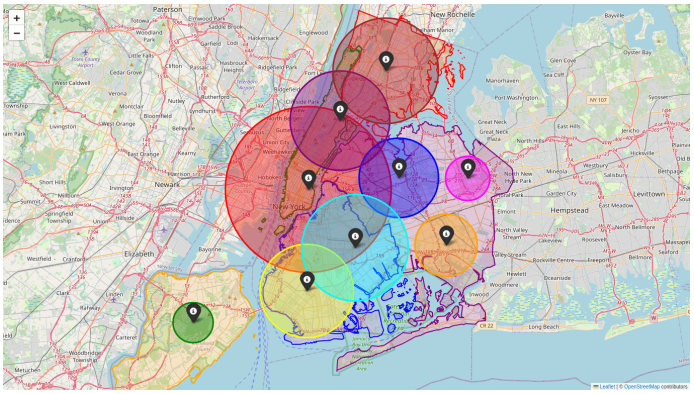

Fig. 29. Hotspot map for May 2017

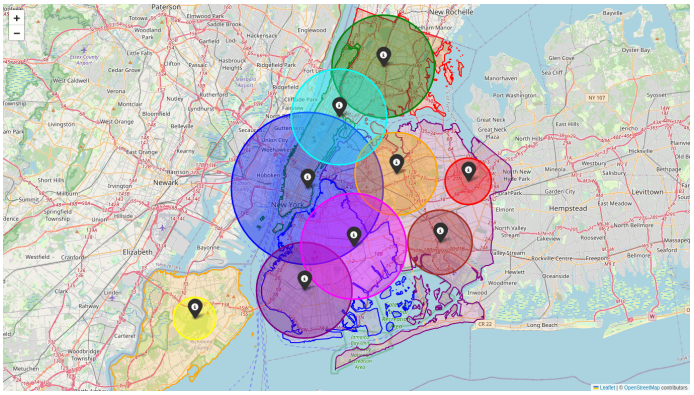

Fig. 30. Hotspot map for May 2018

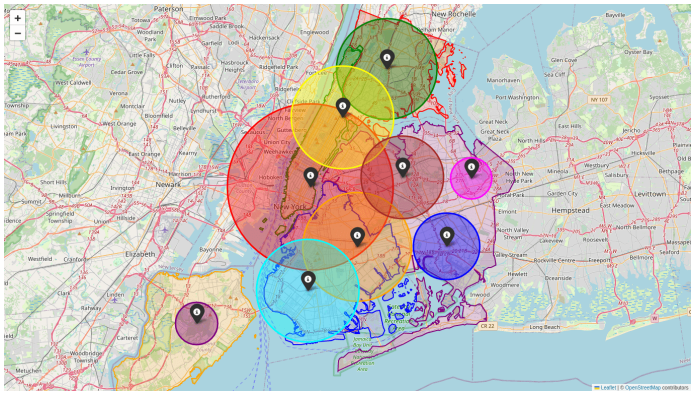

Fig. 31. Hotspot map for June 2017

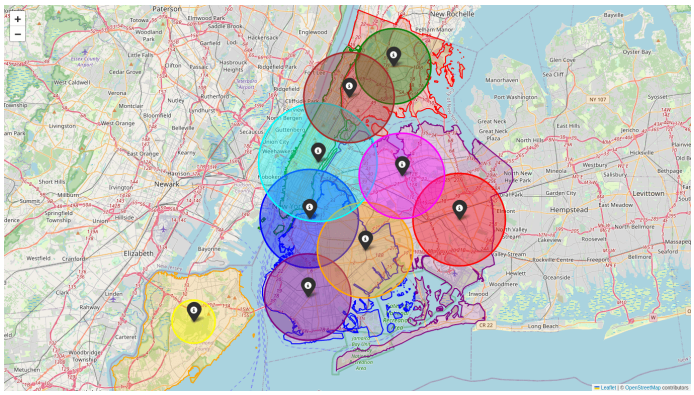

Fig. 32. Hotspot map for June 2018

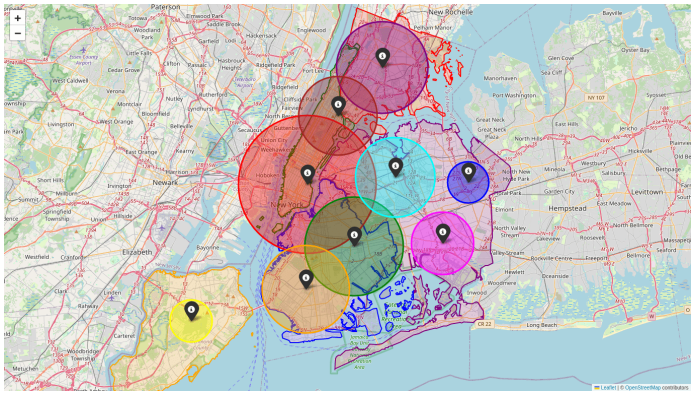

Fig. 33. Hotspot map for December 2018

V. Inferences and Workflows Used

1) The visual analytics workflow described by Keim et al. [12] is shown in Figure 35.
2) From our initial step IV-A, we see that there is a pattern in the number of accidents happening in NYC. The months of May,June and October have in general a higher number of accidents compared to other months. After months of bitter cold and harsh weather, the milder temperatures of May and June bring more New Yorkers out of their houses. Whether traveling by car, riding their bikes, or walking, more people are out on or near the roads. The varying trends across the months of the year, lead us to look through the weather aspects of New York City. So the **Feedback loop** lead us to think in the direction of supplementing our dataset with details regarding weather. (Figure 36)
3) From the first step IV-B we see that there is a possibility of weather events relating to the accidents taking place in

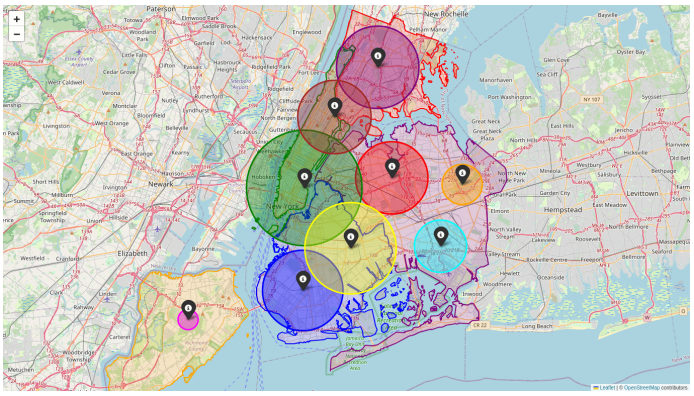

Fig. 34. Hotspot map for December 2019

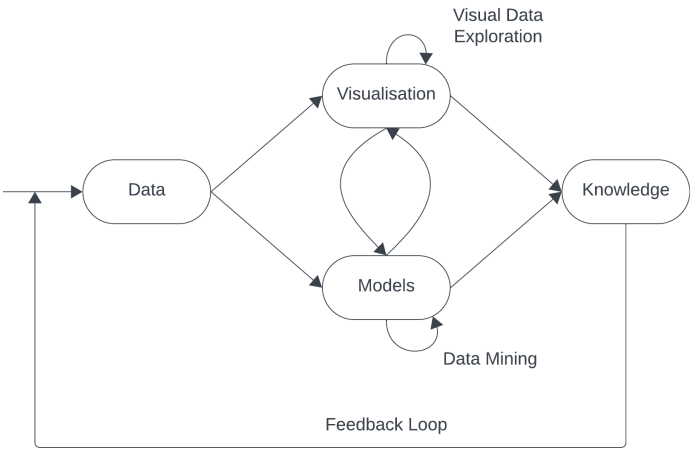

Fig. 35. Visual Analytics Workflow

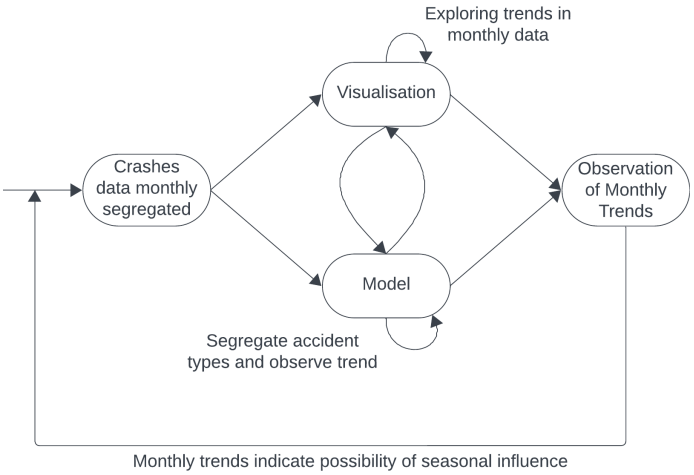

Fig. 36. Initial step of workflow

the city. We have used US weather dataset [2] and used it to get the weather conditions of the city during the time of accident and then used it to get the weather condition that was prevailing during the day of accident. (There might be a slight error, because weather is dynamic and it might change from time to time and the weather the dataset says might not be the actual weather that was present during the time of accident, but that's the slight trade-off we have done to get our visualizations more meaningful).

4) The visualization from the first step IV-B, from the Figure 4 and Figure 5 seemed to indicate that rain and precipitation (of any kind) are highly attributed to the accidents happening in the city. This is clearly seen when there are huge yellow bars in the stacked bar charts of Figure 4 and Figure 5. The other conditions that were of importance was 'snow' and 'fog' which were present in little amounts during the day of accidents.
5) But the problem with this was that there were huge number number of fluctuations for other factors when you see the plots for other years. So we decided that weather conditions such as fog, snow and rain aren't good indicators and decided to move to newer indicators such as temperature and daylight duration.(**Feedback loop**) (Figure 37)

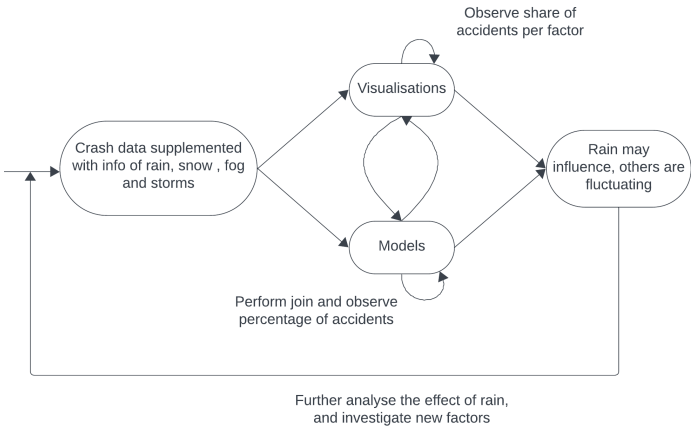

Fig. 37. First step of workflow

6) For this, we supplemented our dataset with precipitation, temperature and daylight info using methods previously described [3], [4] in subsection IV-C. (**Data Mining and Supplementing the data**) We found that the data is largely positively correlated with the temperature and daylight hours, indicating that these factors may directly influence the number of accidents. The precipitation continued to show a fluctuating trend similar to previous, indicating its inadequacy in explaining the monthly trends. 2016 and 2017 were outliers to this trend, where no clear correlation was observed even with temperature and daylight hours (Table I). Equipped with this knowledge, we now move back to perform this

analysis on the data segregated by the time intervals. (**Feedback Loop**)

7) Following this analysis, we decided to try and extend this model to see whether this analysis is true for every kind of accident, or only for a few. For this, five time intervals as mentioned earlier were created and the analysis were done on these intervals separately. The inference from this were that the accidents not in the morning followed the same general trend (Tables II, IV, V, VI). However the accidents in the morning show a very large deviation from this trend (Table III). This goes to show that some trends may be true for general data, but may not generalise well to specific portions of the data, and one must be careful about the same before making generalisations.
8) Having successfully concluded the detailed analysis of the weather conditions and its influence on the accidents, we moved on to investigate the influence of other factors such as demographic info and geographic factors such as tourism or social factors such as national holidays etc. (Figure 38)

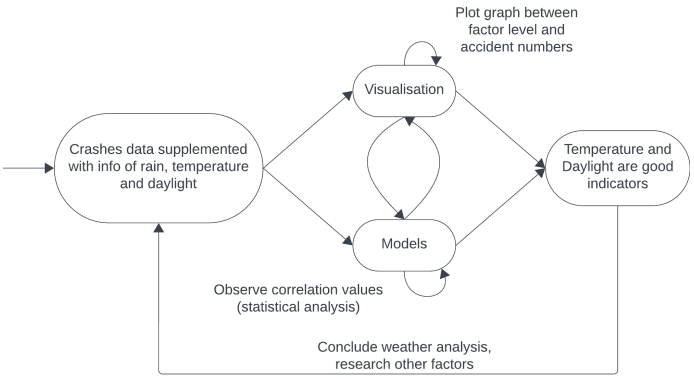

Fig. 38. Second and Third steps of workflow

9) We now begin with the analysis of demographic factors for which we used the dataset [5] (**Data Mining and Supplementing the Data**). On comparing the crashes data with the demographic data, we see a few interesting observations. As discussed before, the figures 22-27 show, demographic data such as age and gender do not really affect the distribution of accidents throughout the year. The number of accidents across these categories only differs because of their actual population and driving frequency. (Figure **??**)
10) Finally to conclude our analysis, we analysed the effect that the geography of New York City may have on the accident numbers in certain regions.
11) We used K-Means Clustering Algorithm for these next few inferences (**Data Mining**). From the Fig. 28, we can very clearly see that the number of accidents in April 2016 is way fewer than any other month, which is kind of an anomaly when looking at the data for New York City, which is consistent with the trend we have seen so far that the data from 2016 is not really consistent with the other years, and is kind of an outlier.

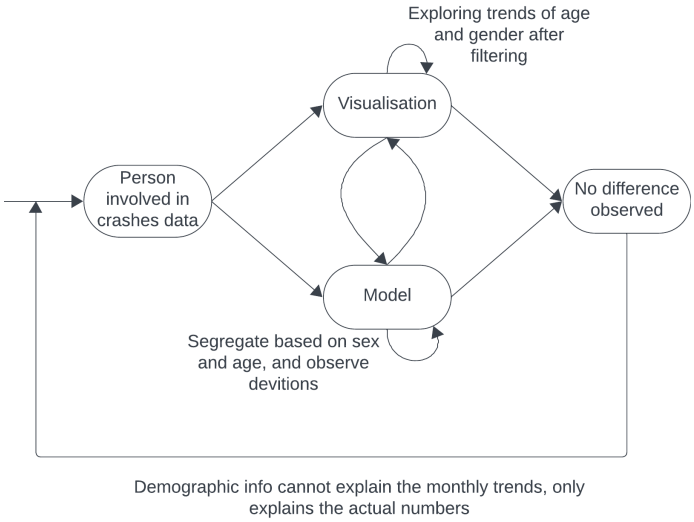

Fig. 39. Fourth step of workflow

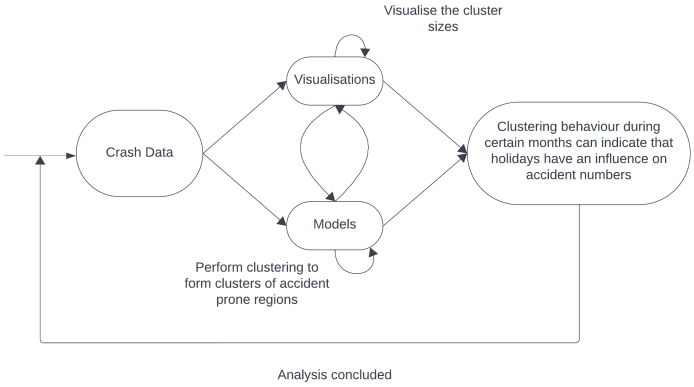

Fig. 40. Fifth step of workflow

12) From the figures 29-34, we see that the clusters are more concentrated around Manhattan, and the reason for that can be attributed to tourism. In the months of May-June and late December, many people visit NYC during school / university breaks and want to visit the tourist spots in NYC, particularly the Times Square which is located in midtown Manhattan and the shopping centres in downtown Manhattan. This tourism spike in these months means there are more vehicles on the streets, inevitably leading to a larger number of accidents and crashes.
13) One more inference we gather from the clustering of this data, consistent with our inferences from A1, is that if we look at Staten Island, we see always a singular cluster with a small radius indicating that the number of accidents is way less when compared to the other clusters in the ”busier” boroughs of NYC. This can be attributed to Staten Island being a suburb.

14) This concludes our analysis. (Figure 40)

VI. Work Distribution

- The initial workflow to observe the monthly trends was done by all three of us collectively.
- Siddharth analysed the factors - precipitation, temperature and daylight hours, while Srinivasan worked on the effects of fog, snow and storms.
- Siddharth and Sankalp worked on the temporal (time of day based) analysis of the trends.
- Srinivasan worked on the demographic based analysis.
- Sankalp worked on the clustering of the data for the geographical analysis, and the generation of videos for the same.

Appendix

This is a follow up to the Assignment 1 report by our team [13]. The important inferences made there have been listed out here -

1) By looking into the pie charts plotted for the number of accidents in NYC, we observe that the majority of accidents are taking place in the night (evening included). This is usually the time when people get back from work or head outside for personal work. We also observe that the accidents that take place during the midnight are relatively more fatal compared to the ones that take place in the morning/afternoon. The plots shown here are for the entire city. Similarly we plotted them for each of the boroughs of NYC, and they have been added to the images directory in our submission. The inferences for the boroughs are similar.
2) Another inference worth mentioning here is that there is a very less percentage of accidents that take place in the mornings, given that is peak traffic time and number of vehicles on the road would be relatively higher. The accidents taking place in the mornings have neither high fatalities nor high injuries suggesting that most of them are minor accidents.
3) On further view on the borough wise geographical scatter plots (put in the images directory), we observe a lot of red points on each map compared to other colours, indicating that most accidents that take place are due to traffic rule violations. Other causes such as substance abuse etc. are relatively less in number.
4) By plotting the co-relation between deaths and causes of accident, we observe that the number of deaths corresponding to traffic violations are high and so is the number of injuries which was expected. But to our surprise we observe that the number of injuries are high for the distraction case, indicating that distraction accidents are mostly non-fatal or minor accidents.
5) Plotting the correlation between time of the accident and cause of the accident yields some interesting information. Usual accidents are somewhat evenly distributed across all times of day. However, accidents due to Medical Reasons/Fatigue and substance abuse occur very frequently in the midnight hours, which is expected. People who drive will usually be tired/sleepy during the late night hours which will make them loose control over driving. Most of the drug abuse, drink and drive issues happen only during the odd hours.
6) Plotting the number of accidents and injuries for each of the boroughs, we observe that the number of deaths, injuries and accidents are higher for Brooklyn and Queens. Staten Island has the least number of deaths as well as accidents among all the boroughs, which is expected given that it is a suburb and is away from the bustling city of NYC.
7) On analysis of the number of cyclists, pedestrians and motorists killed/injured, we realise that cyclists are relatively much safer in New York City than pedestrians and motorists.
8) There is a very low ratio of the number of deaths vs injuries. However, this may be due to the fact that all people injured, regardless of the severity of the injury, have been put into the same category.

References

[1] US data.gov catalog - City of New York - Motor Vehicle Collisions-Crashes [Online] Available: https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

[2] US Weather Dataset: https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

[3] NCEI Time Series Data for Precipitation and Temperature: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00014732/tavg/all/1/2012-2019?base_prd=true&begbaseyear=1991&endbaseyear=2020

[4] Daylight Info New York City: https://aa.usno.navy.mil/data/Dur_OneYear

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

[6] NYC Open Data - https://opendata.cityofnewyork.us/data/ https://data.cityofnewyork.us/api/geospatial/tqmj-j8zm?method=export&format=GeoJSON

[7] Matplotlib Documentation https://matplotlib.org/stable/index.html

[8] Folium Documentation https://python-visualization.github.io/folium/latest/

[9] Scikit Learn Documentation - https://scikit-learn.org/stable/modules/clustering.html#clustering

[10] Pandas Documentation https://pandas.pydata.org/docs/

[11] Geopandas Documentation - https://geopandas.org/en/stable/docs.html

[12] Keim, D., Andrienko, G., Fekete, JD., Görg, C., Kohlhammer, J., Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In: Kerren, A., Stasko, J.T., Fekete, JD., North, C. (eds) Information Visualization. Lecture Notes in Computer Science, vol 4950. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

[13] Assignment 1 Report- https://iiitbac-my.sharepoint.com/:b:/g/personal/siddharth_kothari_iiitb_ac_in/EZJ7FCiQsFJHncHFaNwANlEBOEyQLIyAEKAJGG31HUyALg?e=1Gh7pc