Assignment 3

Nimish Gaurish Khandeparker — IMT2021077
Jashwanth Kadaru — IMT2021095

Instructor: Prof. Jaya Sreevalsan Nair *Date:* December 14, 2023

Contents

1 Overview

The project deals with performing visual analytics on data from Montgomery County's Crash Report Dataset. The dataset has a total of 37 columns and each row corresponds to an accident/road incident. The dataset contains rich features and has plentiful information to be extracted from it. For moving on with the project, We have set an objective of finding crash-prone areas and clustering the incident data corresponding to that area to show the demographics of the accidents that happened in the area, providing useful insights that would help reduce the crashes or take precautionary measures to prevent fatality and property damage. Our visual analytics workflow is based on Keim's visual analytics model. We have used supervised and unsupervised ML models, Visualisations, and followed feedback loops to come up with a dashboard that summarises our findings and helps other professionals to gain valuable insights and improve road safety in certain areas of the county.

2 Introduction

The problem statement for the following visual analytics can be broken down into the following 3 smaller questions:

- The data when plotted on the map as such is very difficult to comprehend and difficult to attack. To approach the problem of improving traffic road safety, we need data that distributed across finite locations and each location should be provide scope for further assessment to reveal the demography of accidents, in turn revealing what could be the cause or contributor to vehicle accidents at a location. So, we pose ourselves the question, Can the accidents be grouped together to form clusters, such that every accident in a cluster is closely related to each other in terms of location and the environment/surrounding conditions?
- Is/Are there traffic metric(s), like speed limit, weather, surface conditions, etc. , that influence the fatality of the accident? If so, we can use this information to make changes in infrastructure, at certain places, to improve the road safety. We can also adjust the rules such that the chances of fatality are reduced.
- Is there correlation between traffic and time of the year? We have found during A1 that number of accidents is directly proportional to the traffic and is more or less independent of other factors. So, if we know the behaviour of traffic, the periodic bursts and falls in traffic with time, we can better adapt to the traffic requirements according to time.

Before diving into the document here is the link to A1 report. The rest of the document summarises how we approached the above objectives and used Keim's visual analytics flow in doing so.

3 Data Overview

The dataset use for this assignment (also for A1) can be downloaded from the following website cat.data.gov. For more information on the dataset, visit: data.montgomerycountymd.gov.

Briefly put, This dataset provides information on vehicle (drivers) involved in traffic collisions occurring on county and local roadways. The dataset reports details of all traffic collisions occurring on county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police

4 Preprocessing

The dataset had NaN and Null values in several rows and columns. Some of the columns with high null percentages ($\geq 15\%$) were removed from the dataset. We have dealt with remaining null values in the dataset by using various metrics from statistics, to appropriately impute values in place of nulls. This is the basic preprocessing done for all the tasks that were preformed. For training the models, we had to do some additional preprocessing which involved encoding categorical columns to numerical columns and dropping/extracting only useful features from the extended feature set based on the task.

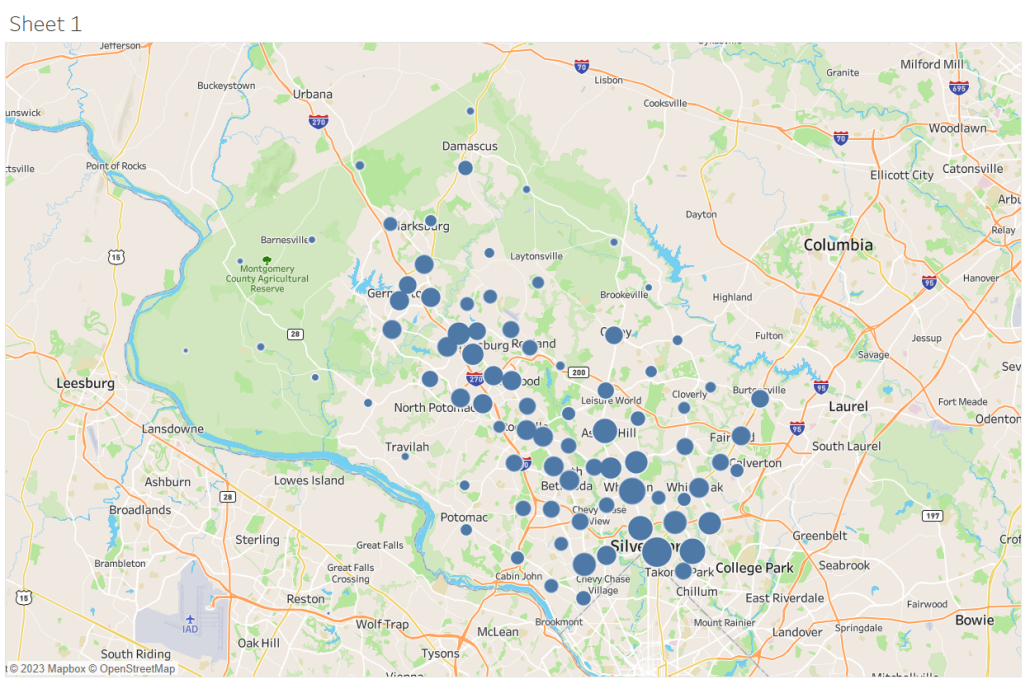


Figure 1: K-means visualisation

5 EDA

We plotted pie charts and bargraphs to analyse the distribution of data across various categorical columns. We noted some common observations about how accidents are distributed across categories in each case. For all of the categorical columns, Most of the accidents fall under 1 or 2 of the available categories making the data uneven. Training models on this type of dataset is challenging due to the biased nature of the underlying data. This was noted and necessary care was taken while training models to counter this effect.

6 Methodology

We have used Keim's visual analytics workflow model to do our project. The rest of the section describes various building blocks used in the visual analytic work flow, like ML models & data visualisations used to analyze and gain insights into data and the inferences drawn from each and descriotion of the feedback loop.

6.1 ML Models

6.1.1 Unsupervised Models:

K-Means: K-means algorithm for clustering data points was used to cluster the accidents by their geographic location/proximity. The algorithm takes latitude and longitude column as inputs, and clusters the data points on the 2d plane. After clustering, the k-cluster centers (longitude, latitude pairs) along with their accident count (aggregate of number of accidents belonging to the cluster) are returned (K=20 & K=100). These clusters were useful in analysing where the accidents cluster around (i.e. where does the density peak) on the map. It is seen by underlaying the street map with the cluster points that most of the accidents happen in cities and larger municipalities and p-arks where the population (vehicle/traffic) density is high. The fig 1. shows the visualisation of clusters (for k=100).

6.1.2 Supervised Models:

Model 1: We use RandomForest bagging to predict Injury severity based on the feature set: . Upon training and testing the algorithm, we get an accuracy of 0.82 and f1 score of 0.74. train_test_split() from scikit-learn was

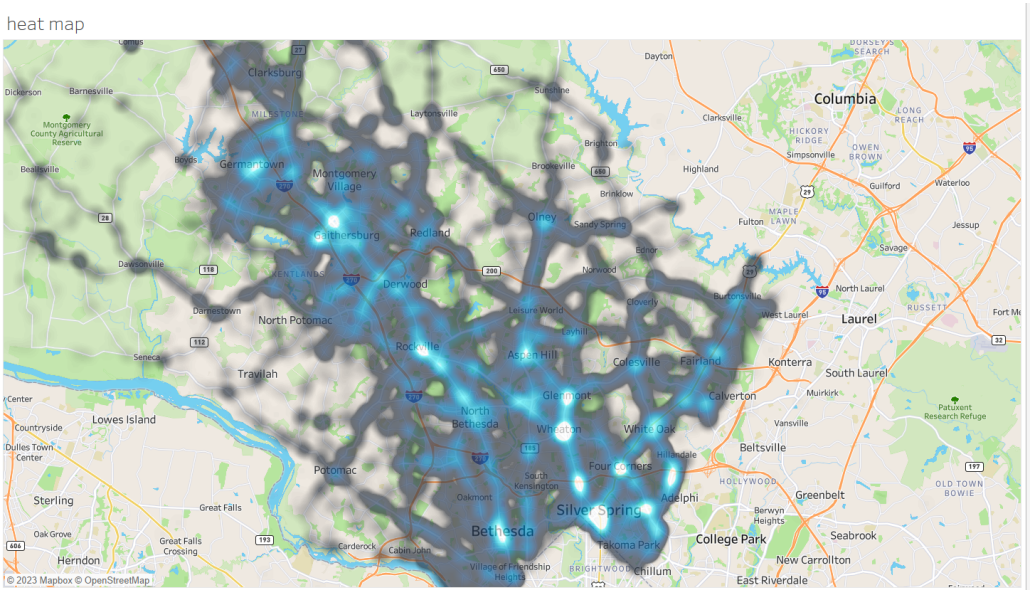


Figure 2: Accident Heat Map

used to do the train-test split. After the training, we also extracted the feature importance of all the processed features (involves dummy columns generated to encode categorical values). Going by the list in order, we have speed limit as the most important feature for predicting injury severiry in accidents, followed by traffic light (which has very weak effect). Hence, we concluded that speed-limit is an important factor to keep in mind while analysing road-safety for various intersections of roads on the map.

Model 1: We use RandomForest bagging to predict Damage extent based on the feature set: . Upon training and testing the algorithm, we get an accuracy of 0.46 and f1 score of 0.44. train_test_split() from scikit-learn was used to do the train-test split. After the training, we also extracted the feature importance of all the processed features (involves dummy columns generated to encode categorical values). After looking at the prediction and making several failed attempts to improve accuracy, we have concluded that the feature set was not enough to predict the Damage extent of a vehicle. This implies that traffic conditions & vehicle speed do not play much role in vehicle damage/safety, as much as they do in preventing accidents and driver injury.

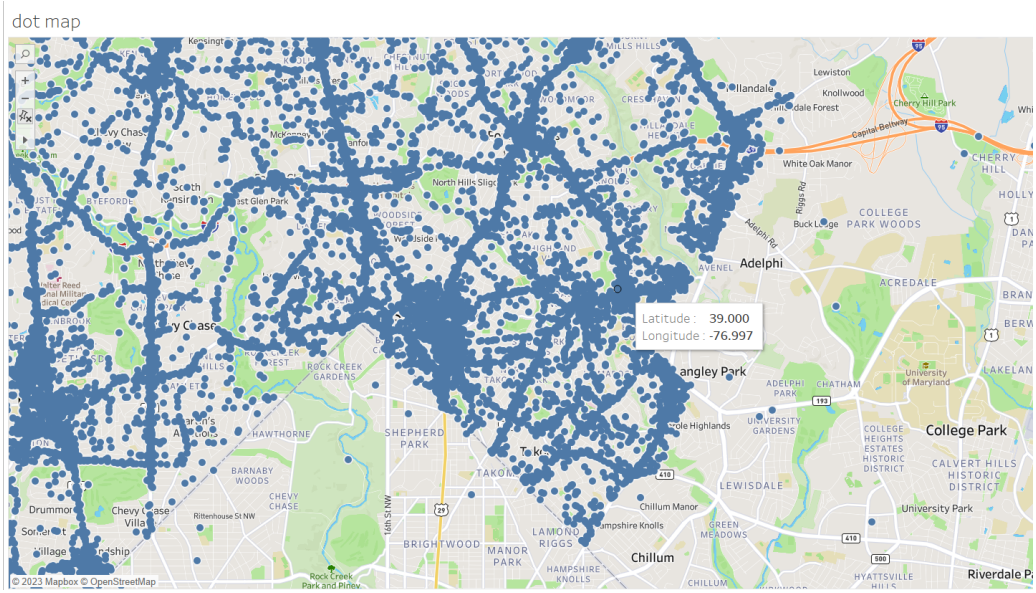
6.2 Visualizations:

6.3 heatmap of accidents based on location:

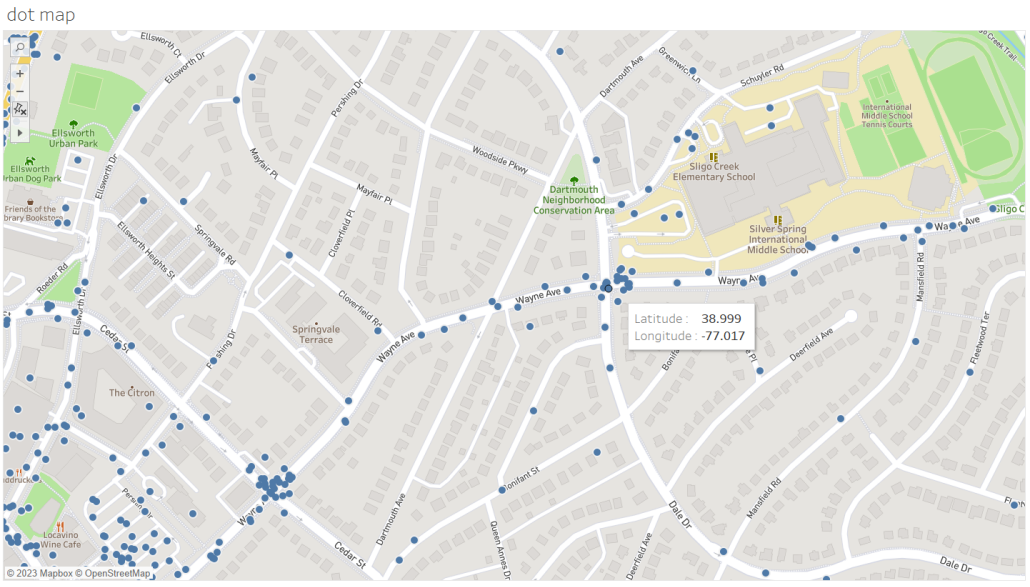
Fig 2. shows the heat map obtained by plotting accidents on map (by location). The high luminous areas show higher density of accidents in the area and dull areas show lower density of accidents. Upon zoom, this indicates that accidents occur mainly at city centers, which are packed with roads and vehicles. If we zoom more closely we will also see that accidents happen around intersections of 2 or more roads or at traffic junctions.

6.4 map of accidents based on location:

Fig 3. shows the map obtained by plotting accidents on map (by location). The areas with more number of circles packed together shows high density of accidents in the area and areas with sparse circle population show lower density of accidents. Upon zoom, we can clearly see that these bigger clusters are made up of smaller clusters which are centered around various intersections and cross streets. This indicates that accidents occur mainly at junctions & intersections. Since, road networks are more branched and densely connected at the city centers, these small clusters make up the bigger clusters that we see on zoom out. The results arrived at in above 2 sub-sections, help us arrive at conclusion with respect to our second question or objective stated in the introduction.



(a)



(b)

Figure 3: (a) Individual Accident location Map, (b) Individual Accident location Map. (Zoomed)

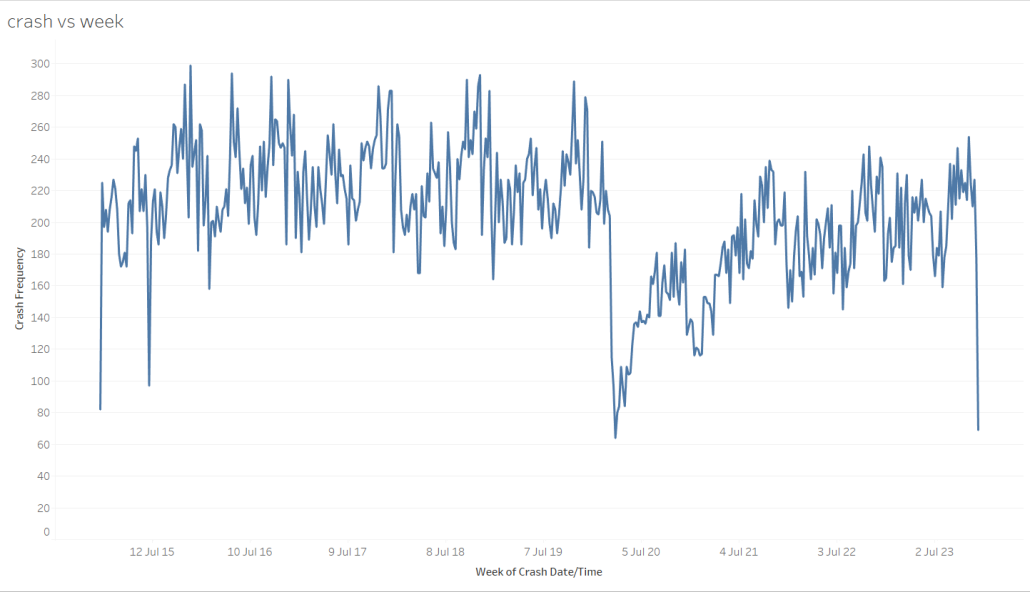


Figure 4: Crashes vs Week (Time series)

6.5 Accident frequency vs week, vs year:

Fig 4. shows accident frequency vs weeks (from 2015 to 2023). It is hard to see any patterns in this graph, but we notice that the graph is has many peaks and lows and fluctuates rapidly. We would like to look for some general patterns across years, which could be useful for predicting traffic behaviours with respect to time of the year. Fig 5. shows the Accident frequency vs year graph, which is way to simple to reveal any patterns. This is obvious because we have only data worth just 8 years with us.

6.6 Accident frequency vs month, for all years:

Fig 6. shows Accident frequency vs month graphs plotted for each year. This visualization shows us that there is a pattern in traffic behaviour and it is repated every year. The traffic peaks every year in the months of May-June and October-November and dips in January-February and August-September. This is very useful for implementing varying road safety precautions depending on the time of the year, to cope up with traffic surges and to save expenditure by traffic control department of county by reducing traffic crew & budget for the months which ehibit a dip in traffic.

6.7 Plot of junctions on the map:

Fig 7. is a plot of circles (of varying size) on the map. These circles correspond to road intersections present in the dataset. These intersections (places where 2 road/streets cross each other) are arrived at by groupig the 'Cross-street Name' and 'Road Name' columns of the dataset. Look at the dataset description at data.montgomerycountymd.gov for more information regarding the interpretation and meaning of each of the cilumns in the dataset. The circle size indicates the accident count. Thus, bigger circles correspond to junctions with higher accidents/incidents. This helps us achieve a acceptable solution to our first objective.

7 Inferences:

The inferences were all mentiond in the section before, under each subsection corresponding to the model or visualisation, from where the inference came from. However, we list them here again briefly:

1. Most of the Accidents happen near the city centers or in larger municipalities in the county. The accidents, plotted as circles on map in fig 1, are clustered around sepcific geography locations and around specific places or municipalities which are large in size and have high vehicle traffic. Rest of the map is, relatively,

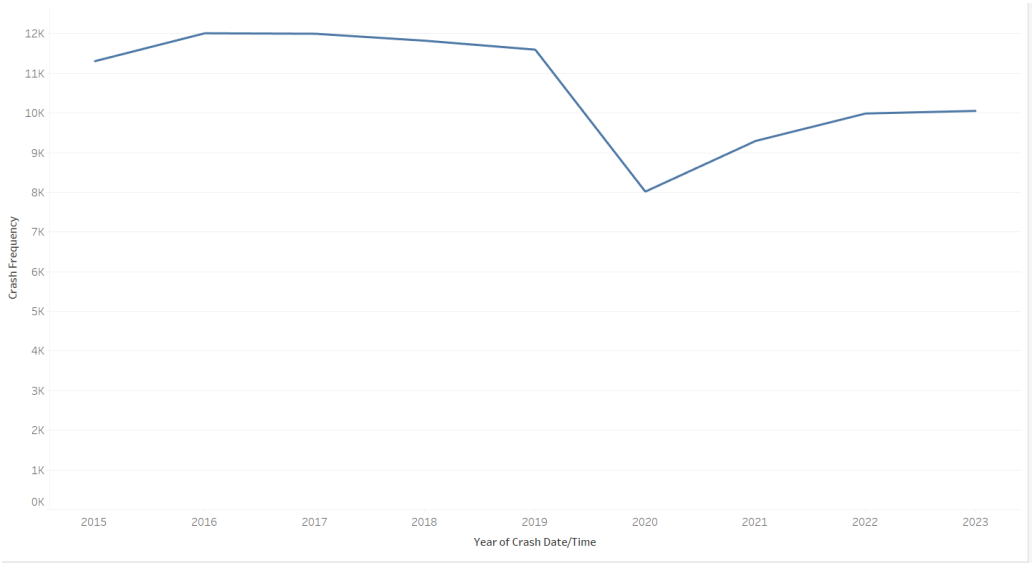


Figure 5: Crashes vs Year (Time series)

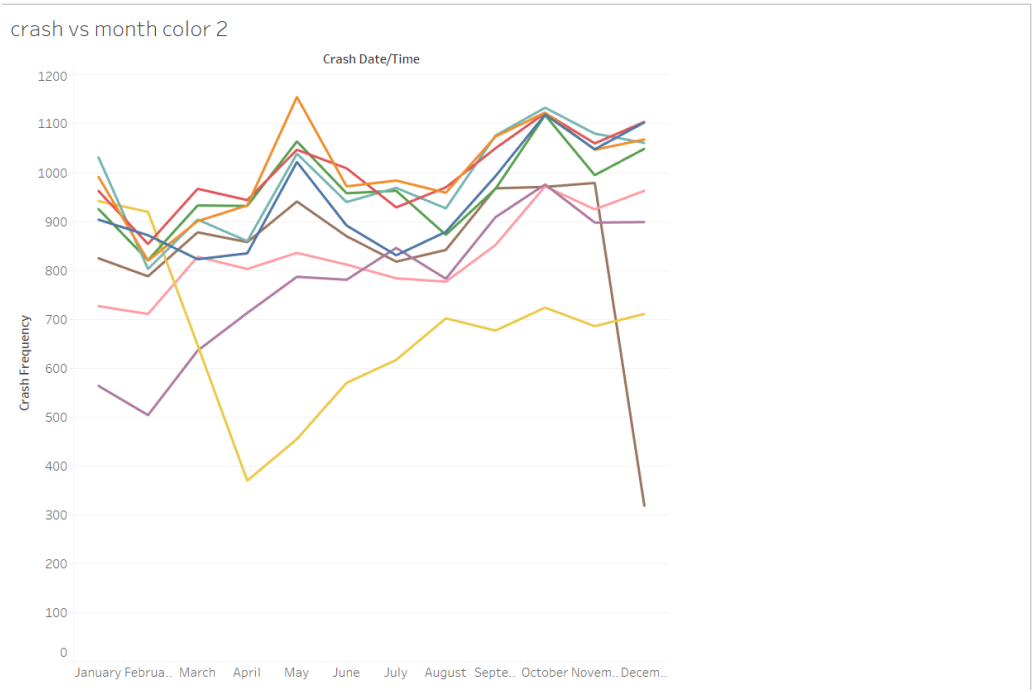


Figure 6: Crashes vs Month (Time series)

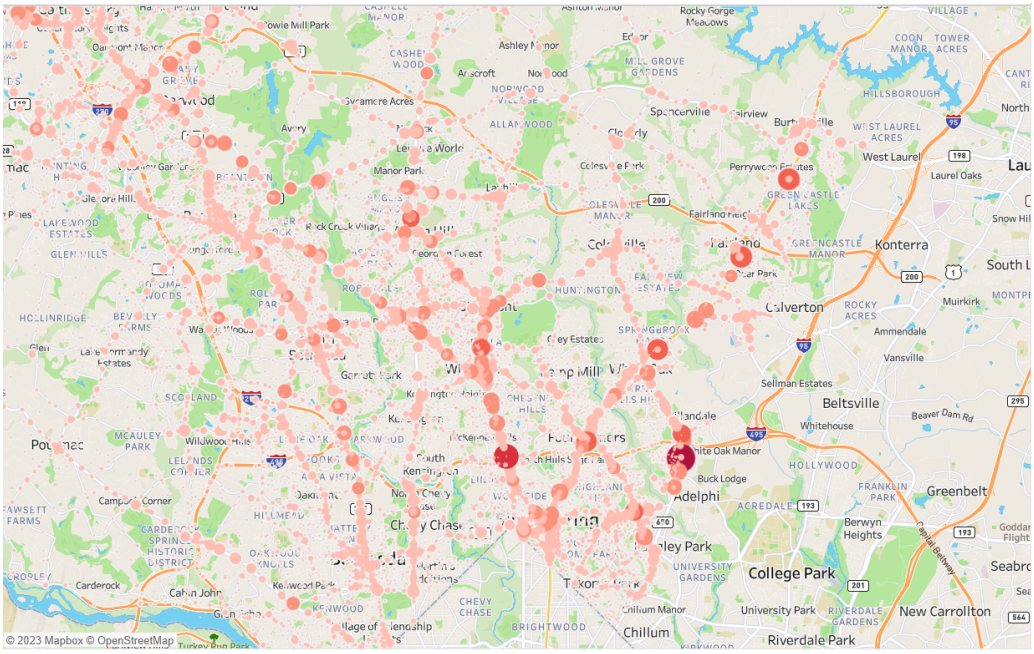


Figure 7: Plot of junctions with varying size over accident count.

very sparsely populated. Focus on meeting traffic needs (& improving road safety) in cities and large municipalities would benefit the police in bringing down the accident rate and fatality rate.

2. Fatality depends majorly on speed of vehicle than any other traffic parameter. Traffic controls at the place also play a significant role in deciding fatality of accidents. Rest do not seem to have much effect on fatalities. This is a useful insight from the point of view of ploice who are interested in improving road safety and bringing down fatality rate on roads.

3. Most of the accidents happen around or at the intersection of 2 roads or at traffic junctions (which is nothing but intersection of more than 2 roads). This is a useful insight. Now, we can confidently proceed on to focus and improve traffic safety measures at the major road intersections which are indicated by size of circle on map. (in Fig 3.)

4. Traffic bursts in the months of May and June, and falls in January-February and October-November, every year. This insight into temporal behaviour of traffic is very useful for traffic police for effectively planning, budgeting, and allocating the resources across the year.

8 Visual Analytics Dashboard:

We now present the final visual analytics dashboard that wa finalised after going through 2 feed back loops (described in the next section). The fig 7 - 12 shows the pictures of dashboard.

8.1 features and visualization mantras used:

Features:

1. User can see accident clusters based on road intersections and junctions. Helps user easily group data and reduce complexity and focus on important areas(roads) of map and find specific info regarding them by interaction.

2. Line chart to describe temporal data about accidents and area chart to describe how accidents are distributed across speed limits.

3. suitable legends and eye-pleasing colors.

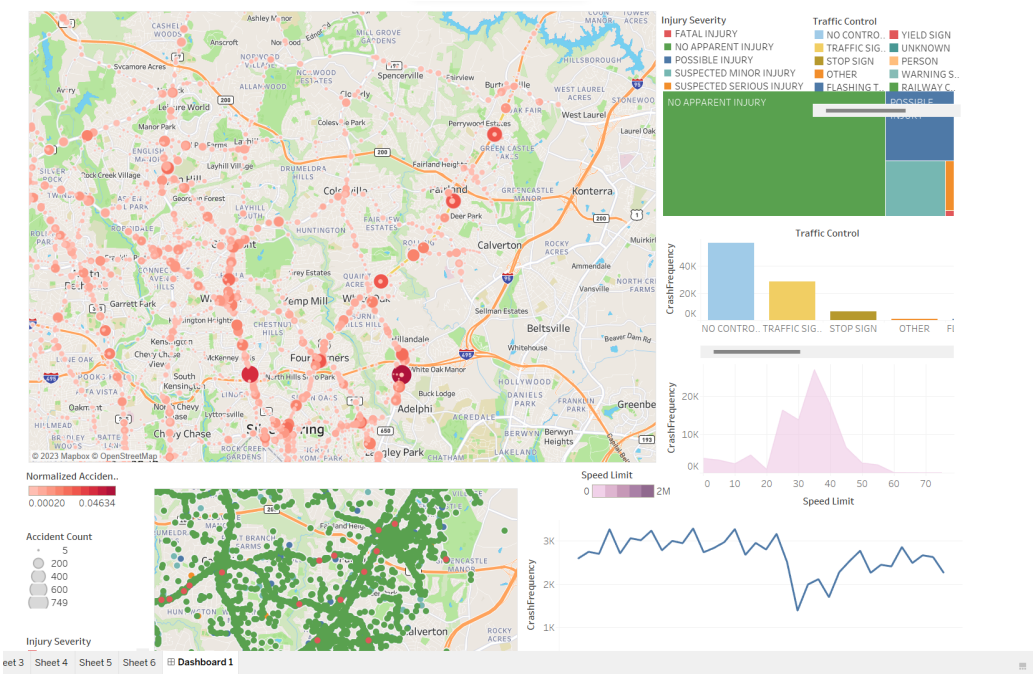


Figure 8: Visual Analytics Dashboard (i)

Visual mantras:

1. Overview First, Zoom and Filter, Then Details-on-Demand:
2. Use of Effective Visualization Types for Specific Data Types.
   (a) Map with Clusters for Geospatial Data
   (b) Tree Map for Hierarchical Data
   (c) Bar Graphs for Comparative Analysis
   (d) Line Charts for Temporal Data
3. Dynamic Filtering. The ability of the dashboard to update other charts based on user interaction with the cluster map is a form of responsive interaction.
4. Facilitating Comparison: By updating the other visualizations based on the selected cluster, the dashboard facilitates direct comparison of data within that cluster versus the general demographics of all accidents.
5. Data-Ink Ratio and Clutter Avoidance

8.2 user interactions:

1. User can filter data being outputted on all visualizations, by clicking or selecting a particular junction or category in one of the visualisations.
2. User can see a tool tip on hovering over the marks in visualisations. This helps us to convey extra information about the mark to user upon demand.

8.3 visualizations:

1. **Junction Map:** Shows each of the junctions wothc cirlces of varying size. The size of the mark corresponds to the accident count.
2. **Tree-Map of Driver Injury Severity:** shows distribution of accidents across injury severity levels.

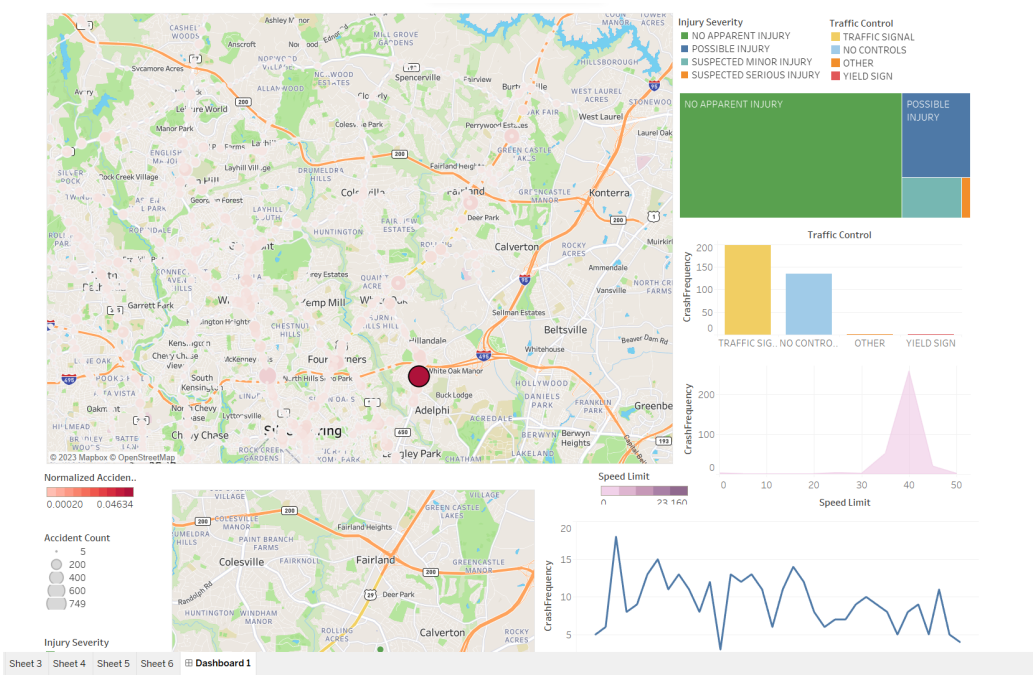


Figure 9: Visual Analytics Dashboard (ii)

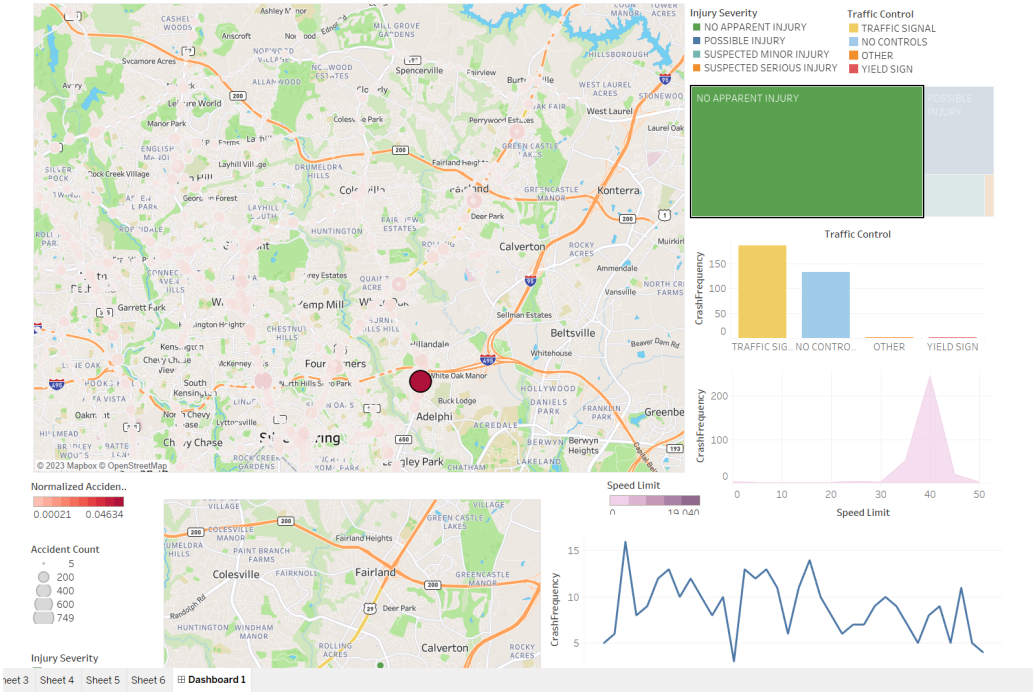


Figure 10: Visual Analytics Dashboard (iii)

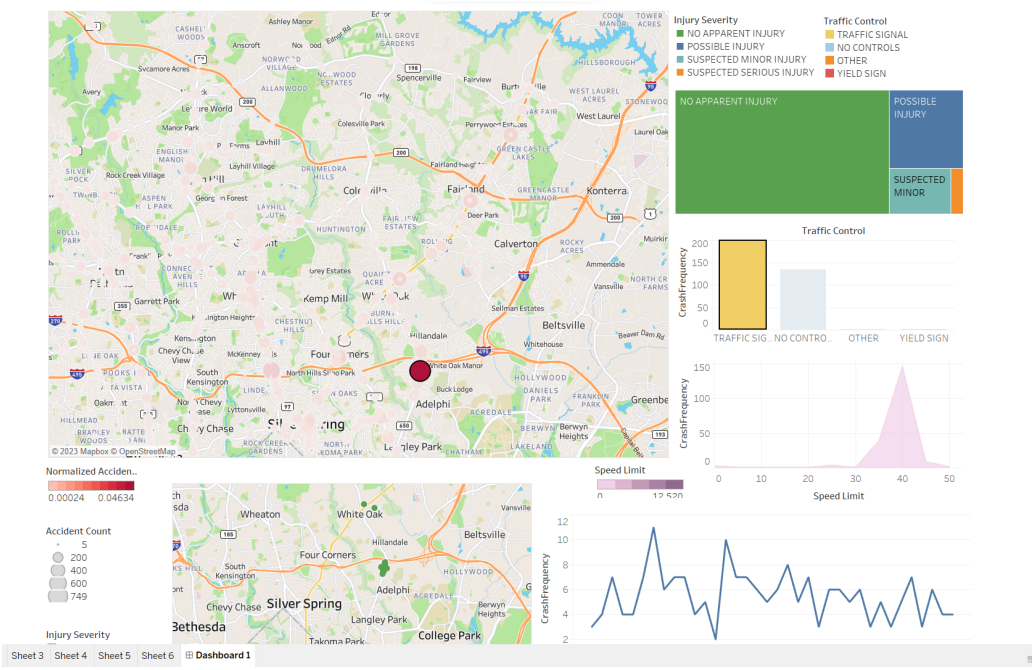


Figure 11: Visual Analytics Dashboard (iv)

3. **Line chart of accidents vs time:** shows how accident rate varied over months.
4. **Bar chart for Traffic controls:** shows distribution of accidents across different traffic control categories.
5. **Area chart:** shows how accidents are distributed over speed limits imposed at the location of accident.
6. **Individual Accidents map:** shows individual accidents of a selected junction on the map. Displays all accidents if none is selected.

9 Feedback Loops:

Following Keim's Visual Analytic workflow link, we have gathered the insights and feedbacks from our visualizations and improved change our models & visualisations. Listed below are the 2 feedback loops:

1. **Feedback Loop 1:** The first inference came from using K-means model and Map visualisations described. This made it clear that visualising data as junctions/clusters would be very useful than doing it individually on the map. Also, from vizzes in EDA, we arrived at some subset of features that could be responsible for predicting Injury Severity and Damage extent.
2. **Feedback Loop 2:** The next 3 three inferences came from the 2 supervised models and there results (described above), the junctions map (arrived at by grouping rows over 'Cross-street name' and 'Road Name'.), Crash frequency vs Month (across years) visualisation.

These feedbacks led to major changes resulting in the final dashboard of improved visualisations.

10 Conclusion:

Visualising accident data as clusters (junctions) is more useful than as individual marks on the map. Fatality depends mainly on speed limit followed by traffic controls in place at the location. This is useful in improving road safety. Traffic shows periodic behaviour, the pattern occurs every year, with peaks in May and June and Lows in Jan- Feb and Oct - Nov months. Overalol, the visual analytics is a very powerful process to gain insight and make useful interactive visualizations for users to explore and interact with.

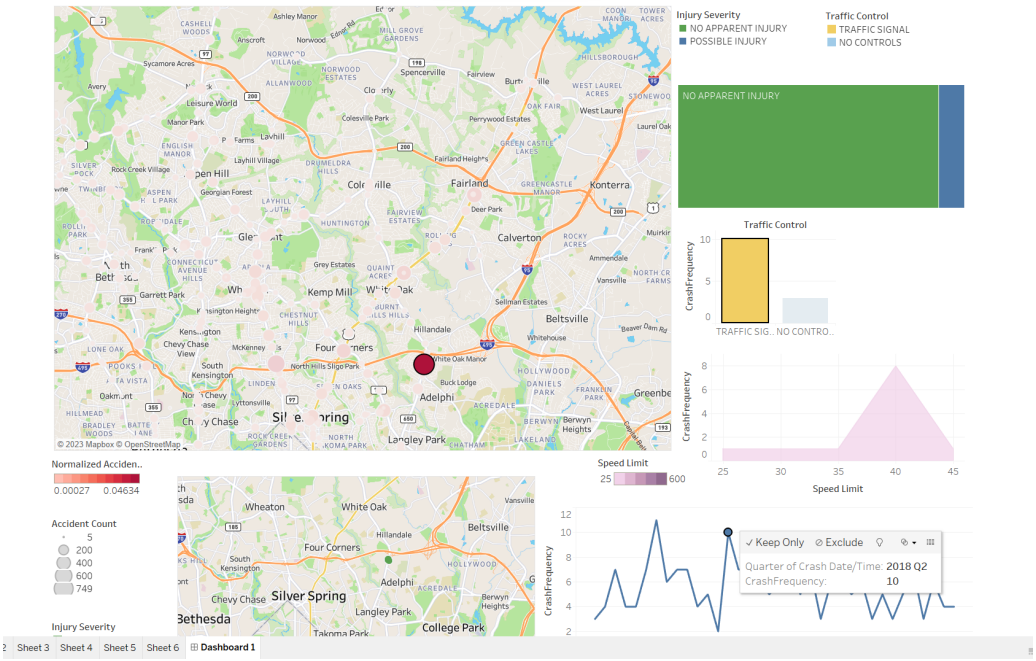


Figure 12: Visual Analytics Dashboard (v)

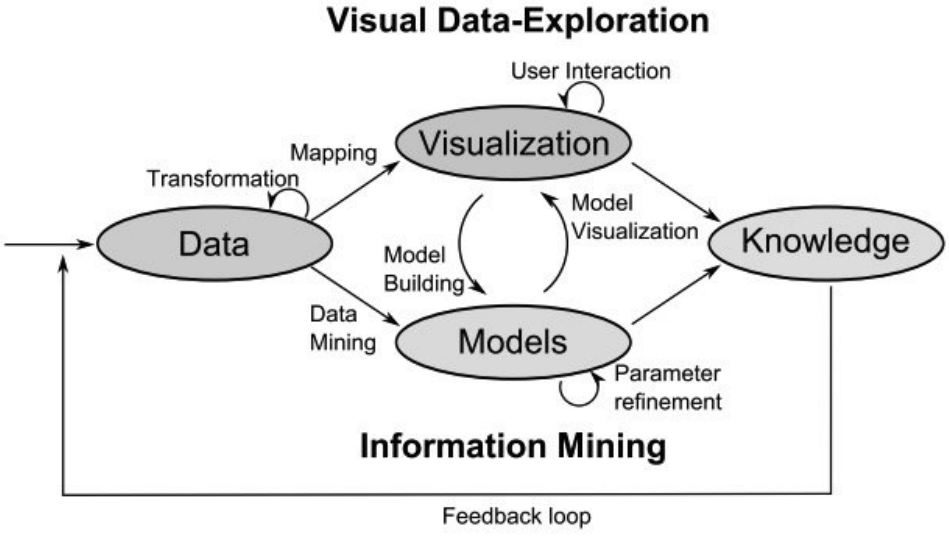


Figure 13: Work flow diagram

11 References:

1. For Dataset source: cat.data.gov.
2. For more info on dataset: data.montgomerycountymd.gov.
3. A1 report link, DVTeam117795: link to A1 report
4. Keim's Visual Analytic workflow, Research Gate. https://www.researchgate.net/

12 Appendices:

link to project folder along with video demo: click this link.