GSBO city collision data analysis

Report 3

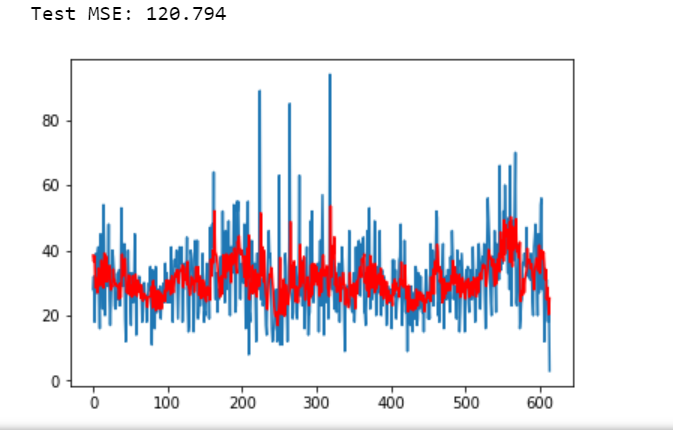
Sahithi Vanteru

1. **Task** - Machine learning -

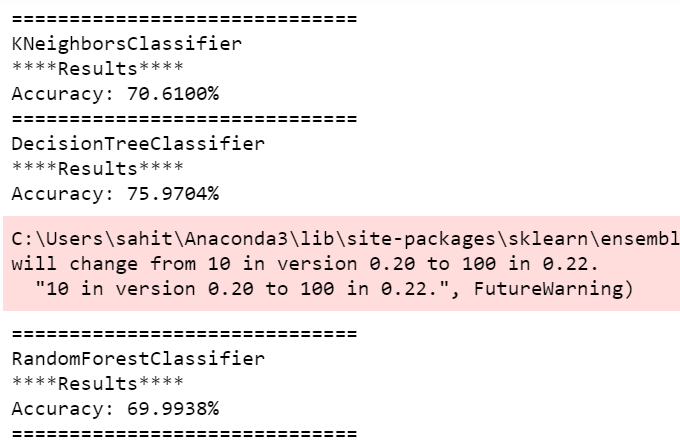
a. Prediction of accidents based on time series data.

b. Random forest classification based on weather attributes. - Vanteru Sahithi.

a.Prediction of accidents based on time series data.

* As we have accidents data from year 2014, using ARIMA(Autoregressive Integrated Moving Average Model), based on the trends we have of past years, I predicted future trends of accidents of years 2019 and 2020.
* I considered the accident\_time column and number of accidents on that particular day columns to use for ARIMA model.
* Attributes - ARIMA(5,1,0) , training - 66% data,testing - 34% data.
* Results - Mean square error = 120.794
* The below figure shows the time series trends of actual values (Blue color) against predicted values ( Red color) of number of values.
* 

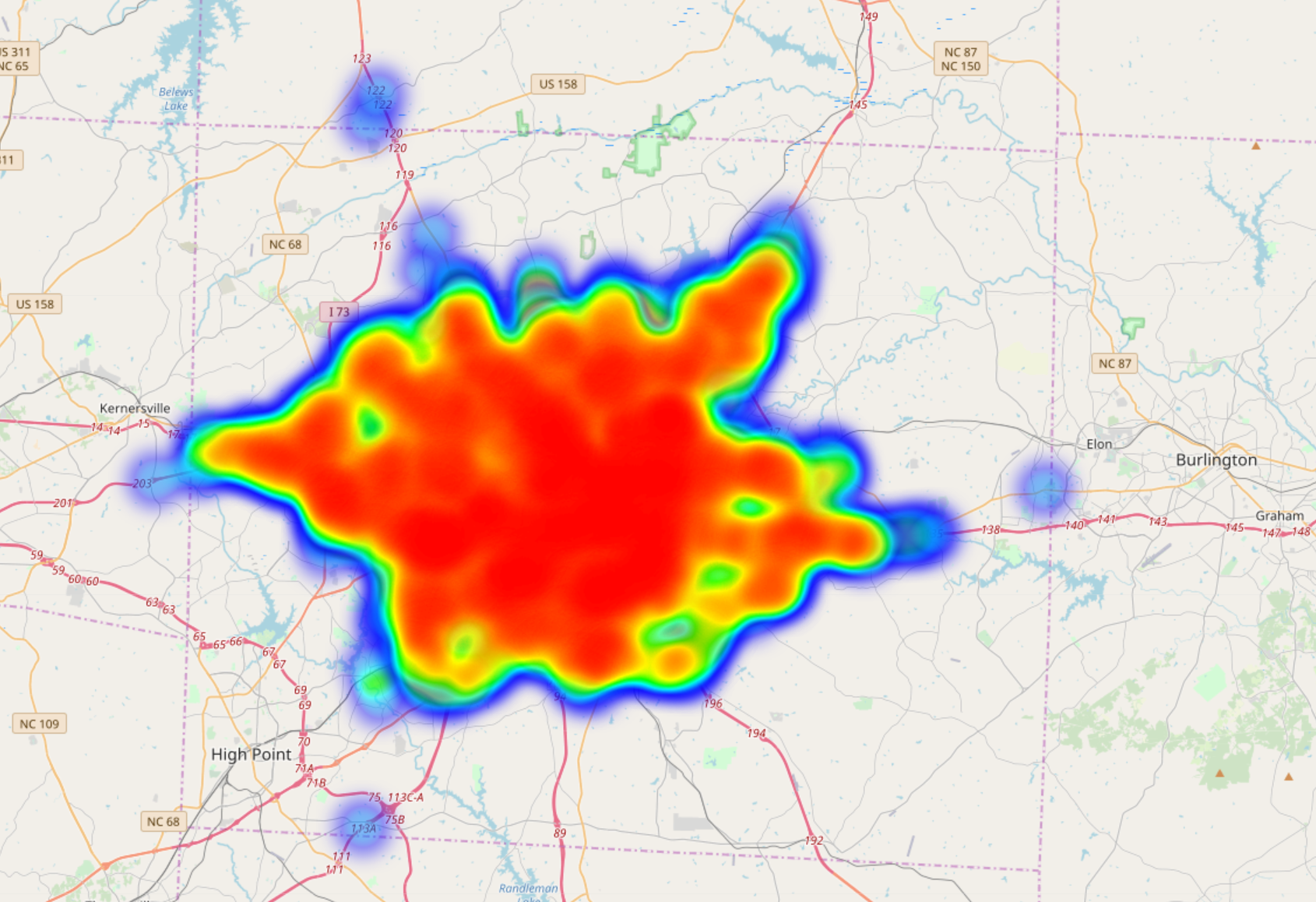
b. Random forest classification based on weather attributes.

* The weather has 3 attributes - cloudy,clear,rain.
* Predicted the type of weather that caused accidents.
* Trained data- 75%, test data - 25%.
* Compared with KNearest classifier and decision tree classifier. The accuracies are shown below.
* 

Daniel Frye

Task: Visualization of data - dashboard

Developing an interactive heat map using folium, currently trying to display the data of all collisions and fatalities by Month-Day-Year with an autoplay slider. Current state of the heat map is a layered heat map of all collisions in 2014 till 2018. This map will be imbedded into the dashboard once completed.

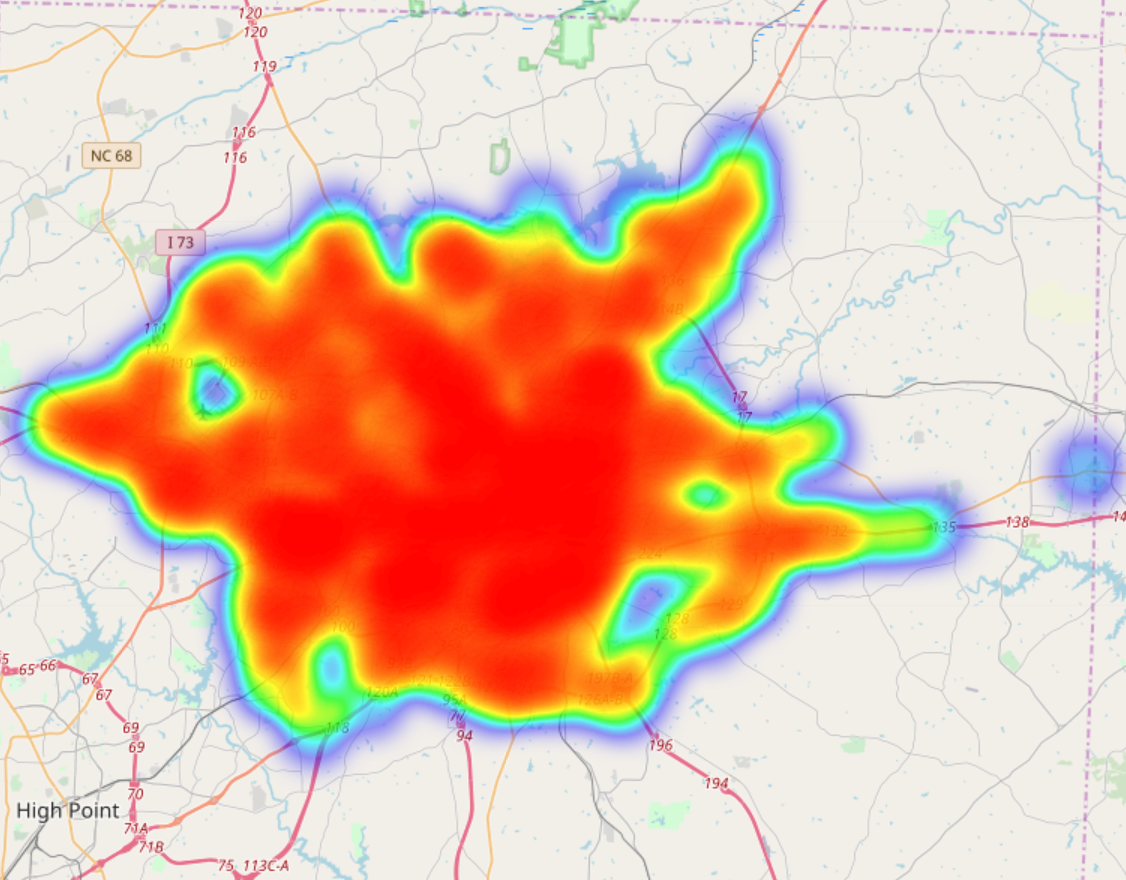
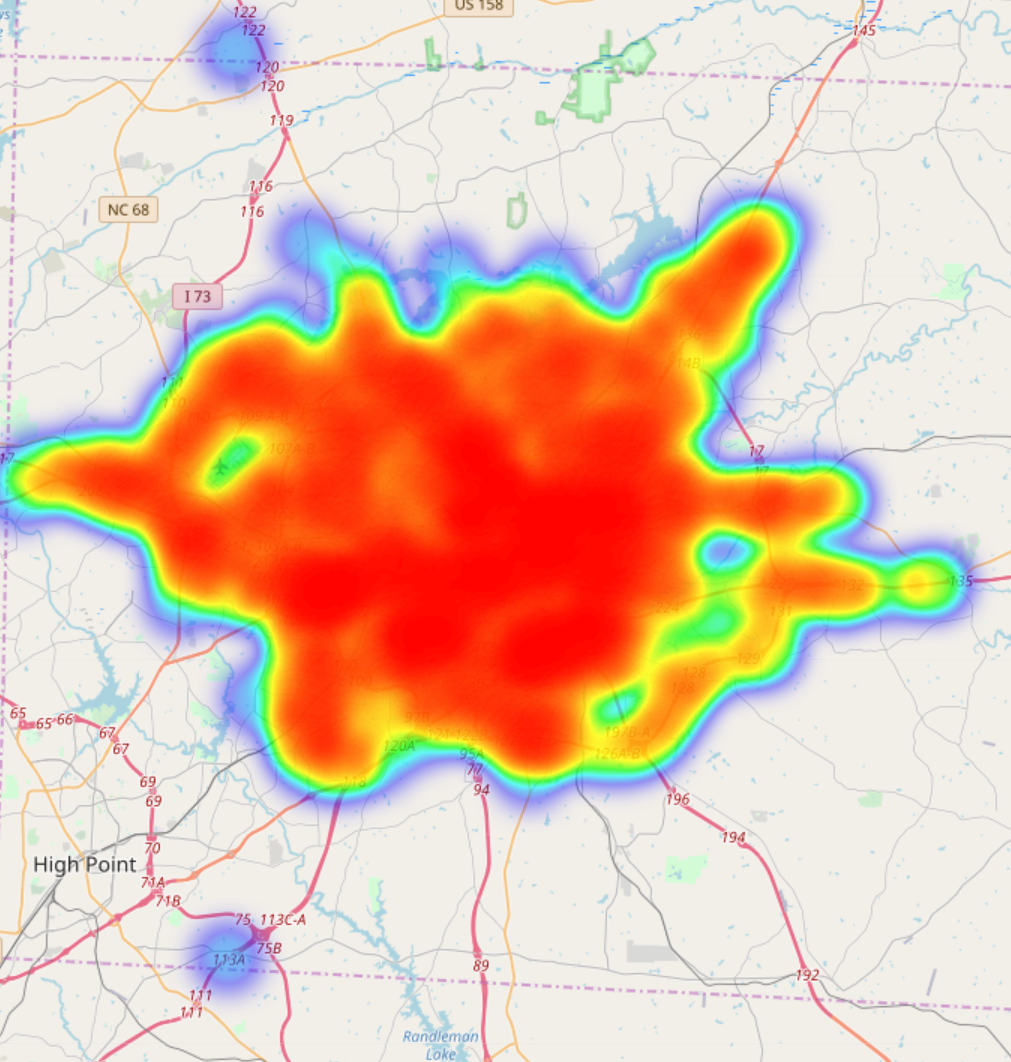


Improvements that are needed:

-slider instead of buttons

-continue to implement month and day in the map

Example of years side by side:



2014 2015

Jorge Salas

Task - Model of Fatalities Locations, weather, date, and time (unsuccessful)

What combination of factors are most likely to cause a crash/fatality was my question? The method should be K-means clustering using scikit-learning module. However,

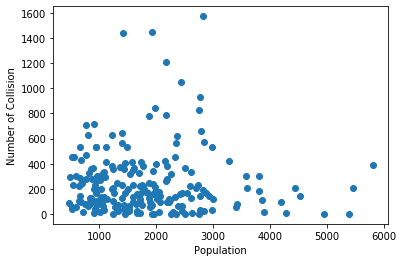
I currently do not have any findings at this time. Perhaps in the final presentation I would have that ready to show.

Jin Kang

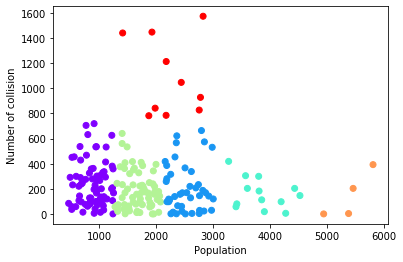
Task - Investigating the number of collisions and block populations of Guilford county where blocks are designed by census

My question was which areas are more likely to have collisions per capita.

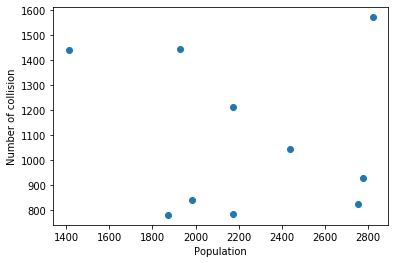
After ignoring blocks that have population but no collision, I could get a graph like below.



To capture points that have enough population and high number of collisions, I used clustering method. And when Kmeans are 6, it was good enough to fit the data.

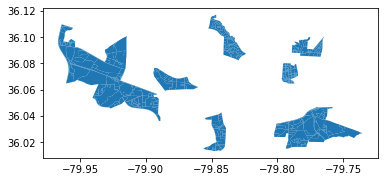


These are the graph and the list of red points.



|  |  |  |
| --- | --- | --- |
| **Display\_Label** | **Population** | **Number of Collision** |
| Block Group 2, Census Tract 103 | 1412 | 1447 |
| Block Group 1, Census Tract 164.05 | 1871 | 782 |
| Block Group 2, Census Tract 108 | 1926 | 1440 |
| Block Group 1, Census Tract 160.11 | 1982 | 842 |
| Block Group 2, Census Tract 126.01 | 2174 | 785 |
| Block Group 1, Census Tract 128.04 | 2176 | 1213 |
| Block Group 1, Census Tract 125.09 | 2440 | 1047 |
| Block Group 1, Census Tract 126.09 | 2755 | 827 |
| Block Group 1, Census Tract 128.05 | 2777 | 928 |
| Block Group 1, Census Tract 165.03 | 2823 | 1573 |

These is a display of captured blocks by using geopandas mapping method.



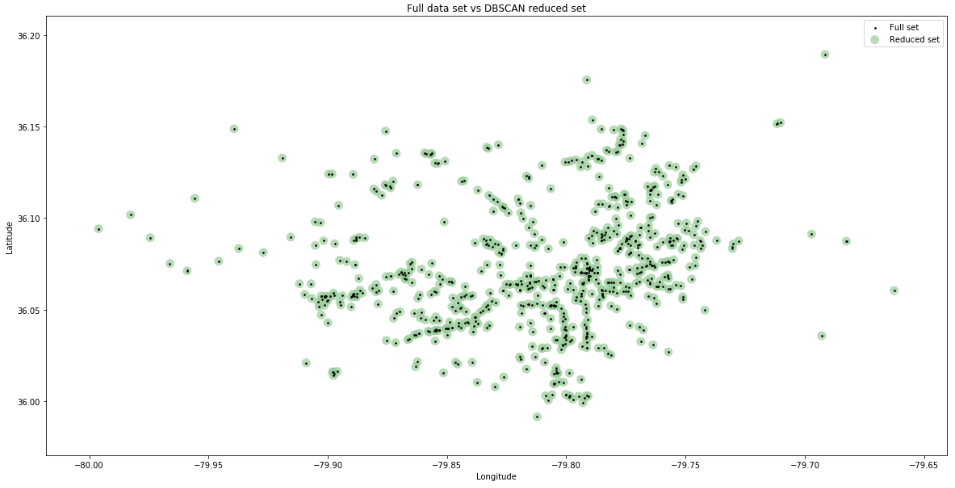
Dimithra Ratnayake

Goal:

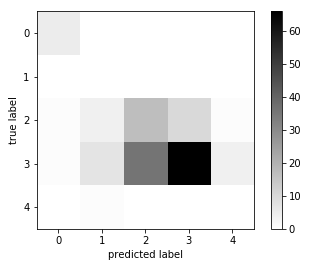
To identify pedestrian collision hotspots and important factors contributing to pedestrian collisions.

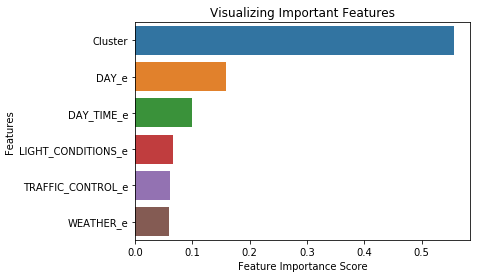
These goals were chosen because the results from the goals would help the Greensboro city to implement steps to mitigate pedestrian collisions in identified locations.

To identify the hotspots, DBSCAN unsupervised learning method was used. Latitude and longitudinal data was used to generate the clusters, where there would be accidents within 25 meters of each other. This resulted in 639 clusters and there were 20 clusters with more than 5 accidents within a cluster. These 20 clusters could be considered as hot spots.



Random forest supervised learning method was used to identify factors which were important in collisions. Two approaches were used. For the first, Severity was used as the dependent variable. It is a categorical variable so the classification model was used. This gave an accuracy level of 0.575 and did not predict all the levels in a consistent manner.





For the second approach, number of collisions in a cluster was used as the dependent variable. By implementing a regression random forest method, the results generated a model with a R-square of 65%. Similar to the previous approach, this also indicated that the most important variable as the ‘Cluster’ with about 80% importance. All the other variables were comparatively less important.

