DALHOUSIE UNIVERSITY

CSCI 5408 DATA MANAGEMENT WAREHOUSING AND ANALYTICS

PREDICTIVE ANALYSIS ON 911 CALLS OF MONTOGEMERY COUNTY

Submission Date: 05-08-2018

Submitted by

Manojha Panneerselvam	B00783628
Lakshmi Maligireddy	B00773792

GitHub https://github.com/Manojha18/DW_project.git

Table of Contents

List of figures

List of Tables

1. SUMMARY

We worked on 911 calls of Montgomery dataset and did exploratory and geographical analysis also decided to do predictive analysis based on latitudes and longitudes to identify which part of the township requires more attention. We worked with Clustering algorithm based on unsupervised learning and predicted the population density for each cluster of the Montgomery township. Based on the predicted population and the extracted station number from the time stamp we were able to identify the pattern and conclude that the cluster with more population density requires more emergency and fire services when compared to the clusters with less population density.

2. PROBLEM STATEMENT

Montgomery County is one of the populous Counties in The United States, so it is more prone to accidents, crime, and disaster. To help Montgomery police department serve their people better and keep their cities safe we decided to analyze the Montgomery 911 calls and found patterns and trends on which portion of the county requires more attention.

3.TOOLS AND RESOURCES

We worked with data analytical tools and techniques which helped us with the flow of the project.

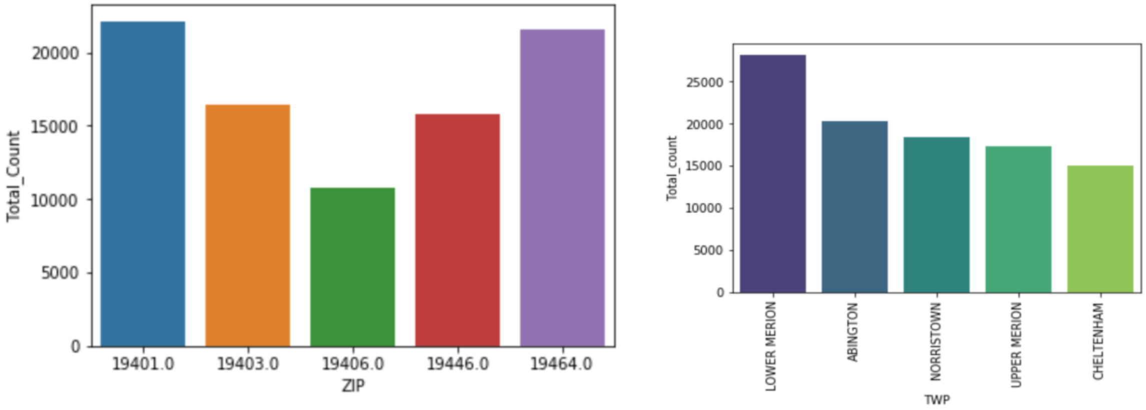
- Scripting Language: Python
- Libraries used:
    - Pandas for database manipulation
    - NumPy for mathematical manipulation
    - Matplotlib and seaborn for plotting
    - Sklearn for machine learning algorithms
- IDE: PyCharm, Jupyter Notebooks
- Visualization: Tableau
- Project Management Tool: GitHub

4.DATA ACQUISITION AND CLEANING

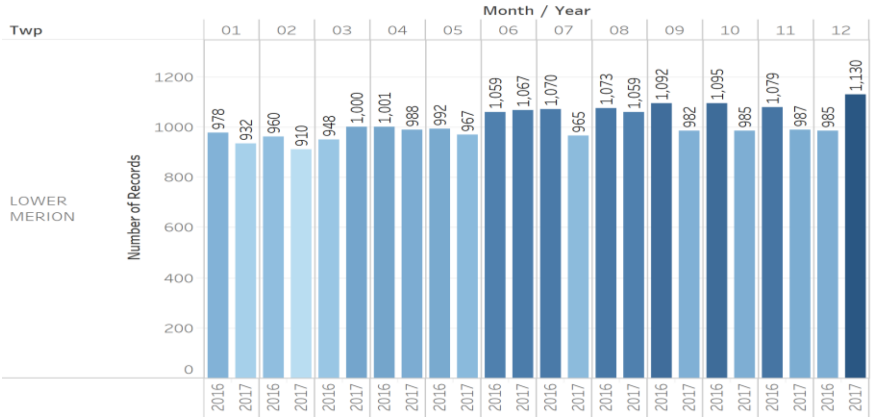
We got 911 calls of Montgomery dataset from Kaggle. It is platform for predictive modelling and analytics. The range index of the data set is around 3,26,425. The dataset comprises of 9 data columns latitude as lat, longitude as lng, address with time, date and station ID as desc, zip code as zip, call description as title, time and date of the call as timestamp, township as twp, street address as addr, e. Most of them are string values whereas very few are numerical. The data set is from 2015 to 2018. We removed the incomplete data, dirty data, missing values, noisy data, incorrect format, too much data, inconsistent data and non-integrated data. We also extracted certain useful features from the attributes like station Id and Timestamp. Through station ID we got the station number so that it was useful to find the location of the station in the township. With Timestamp we extracted time, hours, years, month and days so that we were able to identify the volume and the range of the data points in the dataset. We understood that we have a complete 911 calls data only for the year 2016 and 2017 whereas, 2015 and 2018 have very few data points.

5.DATA EXPLORATION

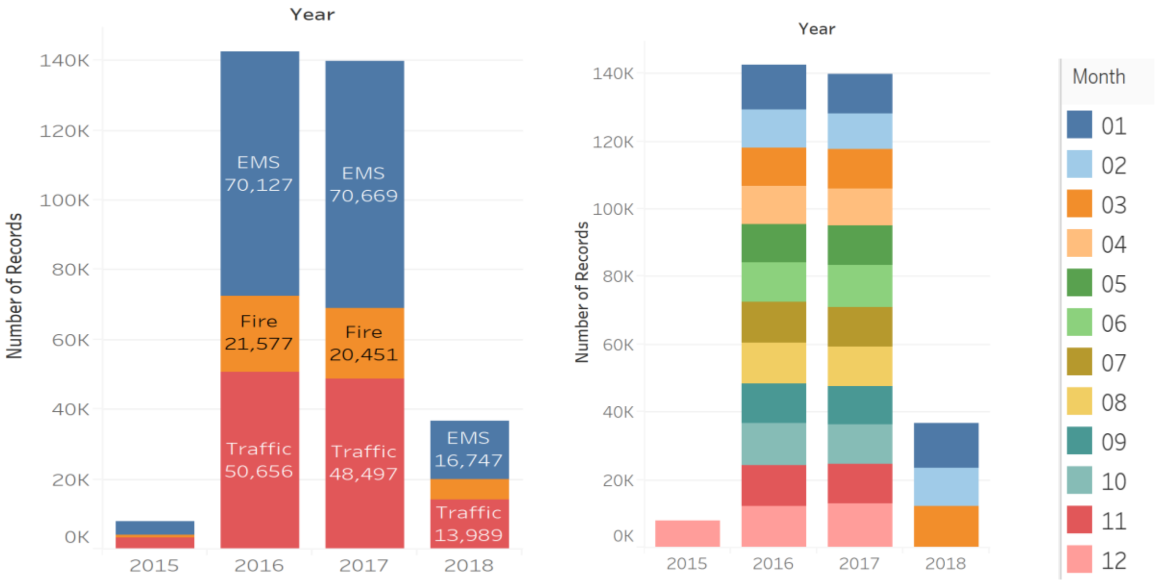
We explored the data using various visualizations and feature extraction. Initially, we found the top five zip code using value_count. Similarly, we found the top five cities for 911 emergency calls. Through the plot, we came to an understanding that Lower Merion is the top city for a frequent number of 911 emergency calls. So, we decided to take a deeper look into this city and created a separate plot which is plotted against the years 2016, 2017 against the number of records. We can clearly see that there is a consistent number of 911 calls throughout both the years. There is no dramatic increase or sudden breakdown. The first half of the year receives the same number of emergency calls as that of the second half of the year. We also explored the 911calls in Montgomery based on the time stamp and the extracted year. We plotted a graph with a number of years against a number of records, we could identify that 2016 and 2017 have data points for all the months in the year whereas 2015 and 2018 have very few data points. We took a deeper look into months of 2015 and 2018 and came to know that only December month data points are present for 2015 and for 2018 months from January to March is considered.



1.1) Top 5 zipcodes and top 5 cities



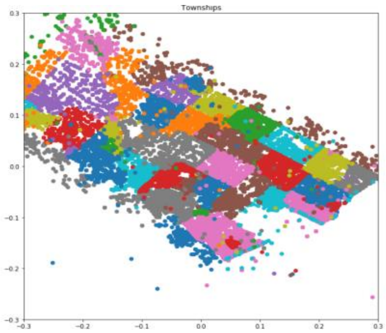
1.2) Lower Merion 911 calls for 2016 and 2017



1.3) visualization based on year and months

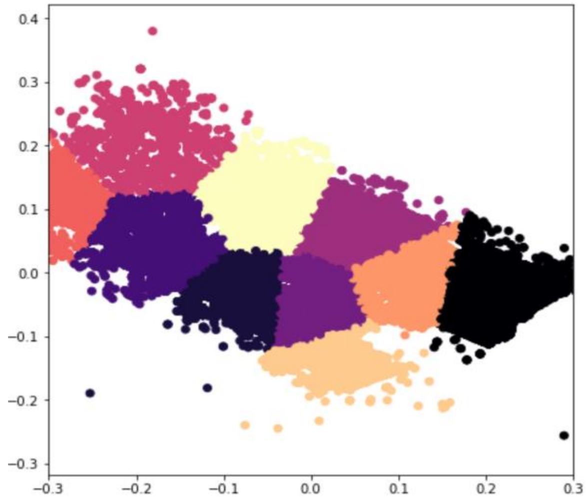
6.ALGORITHM

Grouping the cities based on towns seems to be unrealistic and inefficient with fewer data points so we decided to follow a different grouping mechanism.



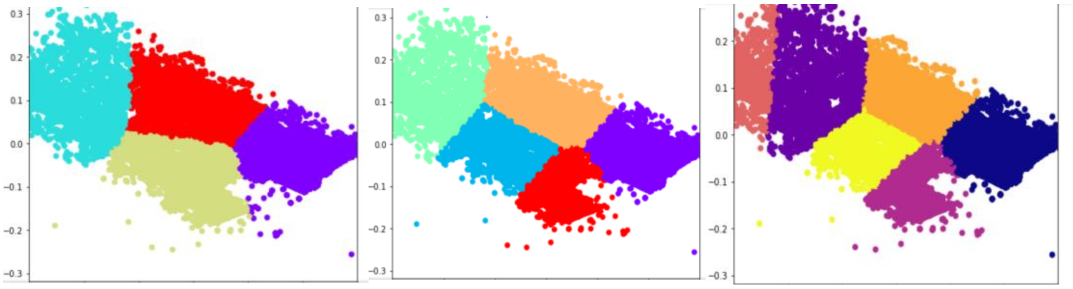
1.4) Datapoints based on towns

We decided to do predictive analysis based on unsupervised learning and we used Clustering algorithm. Clustering is nothing but grouping of similar features or characteristics. As the number of clusters increases the members within each cluster become more like each other moreover, the nearby clusters become more similar with the neighbourhood clusters. This do not provide us with proper insight, so we should carefully decide on choosing the number of clusters. We initially chose the cluster of size 10. From the figure 1.1 we can see ten different clusters for Montgomery county with 10 distinct colors.



1.5)Number of cluster = 10

Later we began to test with different number of clusters ranging from 4 to 8. Finally, we decide to choose the number of clusters to be 6 because the southern part or the lower part of the county seems to have lesser number of data points whereas the middle portion seems to have higher density. So, it is better to have a greater number of divisions in the middle than towards the end. All these observations seem to be satisfactory only when we choose n=6. So, we decided to proceed our findings with six number of clusters.



n=4 n=5 n = 6

1.6)clusters based on the range of 4 to 6

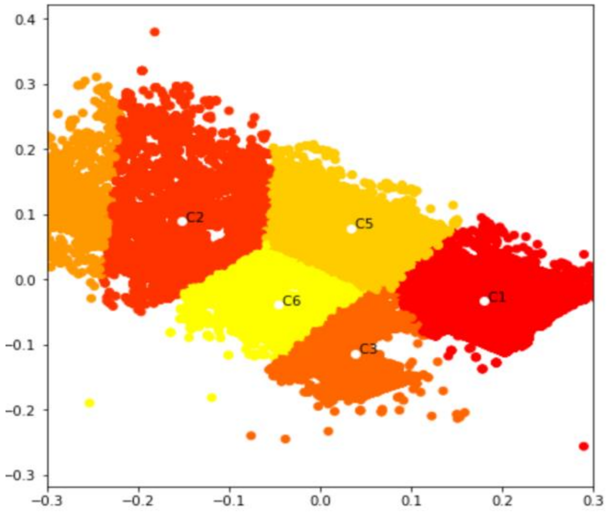
To predict the population density of each cluster. We decided to calculate

1. Approximate area of the cluster
2. Average population density

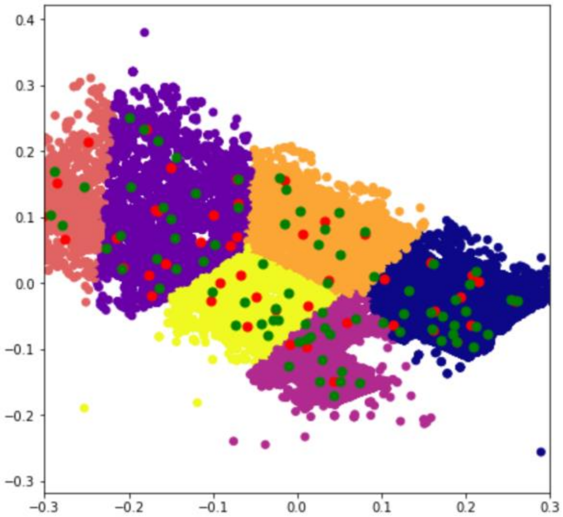
Initially, we calculated the area of the township using the formula

pi.R^2 |sin(lat1)-sin(lat2)| |lon1-lon2|/180

Then we calculated the overall township population based on the USA's average urban population to be 314 people per sq Km. Then we created a model to approximate the population of each cluster. The approximate population density of each cluster is approximated based on kernel density. Kernel density is an inbuilt function in Sklearn, we used to approximate the emergencies with the given data points. Later we calculated the area of each clusters and the centroid of the clusters. Finally, we got the approximate area of the cluster and the predicted population density of each cluster.



1.7) Clusters with centroids



1.8)Clusters with emergency and fire services

After taking a closer look into the clusters we can conclude that clusters with higher population density has lower number of emergency and fire services. Whereas the services with lower population density seems to have more number of services. The green dots represents emergency services while the red one represents the fire services.

7.FUTURE WORK

We decided to follow the same approach with different cities in united cities with more number of data points. It would be helpful to identify more patterns and consider more number of features. We can also future divide the clusters into sub groups and get a deeper view in deploying the necessary emergency services.

8.VALUE PROPOSITION

By predicting the population density - government, police department and intelligent systems of the Montgomery County can serve their people, help the city thrive for better and keep the people safe.

9.SPRINT REVIEW

The sprint review is divided into three phases like exploring, modelling and visualization. We followed the sprint cycle so that we were able to finish the project at the right time.

S.No	Cycle
Sprint 1	We planned to work on data acquisition, cleaning and data exploration
Sprint 2	Predictive analysis based on unsupervised learning, calculated approximate cluster area and predicted the population density of each cluster
Sprint 3	Identified patterns, Visualization

1) Sprint Review

10.LIMITATIONS

- We had lesser number of datapoints, so we were not able to get strong insights
- Since it was unsupervised learning we randomly focused on latitudes and longitudes
- Apart from Clustering algorithm, we did not try any of the other algorithms

11.CRITICAL REVIEW

- We did not perform K-fold cross validation to find the accuracy of the model
- If we had more time, we could have gathered datasets for crime and would have calculated in which portion of the county more crimes are occurring.
- We limited our visualisation to an extend, we could have identified a few more patterns
- We also limited our focus on descriptive analysis

12.ROLE EFFORT

DATA ENGINEER - Lakshmi Maligireddy

- Data Acquisition
- Data Cleaning
- Data Exploration

DATA SCIENTIST - Manojha Panneerselvam

- Statistical analysis
- Model building
- Visualization and identifying the patterns

13.REFERENCE

[1]. *Mathforum.org*, 2018.
[2]"Overview — NumPy v1.15 Manual", *Docs.scipy.org*, 2018
[3]"911 Calls - City Sevices Planning For Emergencies | Kaggle", *Kaggle.com*, 2018.
[4]"Emergency - 911 Calls | Kaggle", *Kaggle.com*, 2018.
[5]"Package overview — pandas 0.23.3+8.g4aa80b6d6 documentation", *Pandas.pydata.org*, 2018.
[6]"2.3. Clustering — scikit-learn 0.19.2 documentation", *Scikit-learn.org*, 2018.