EMS Data - Predicting Ambulance Response Times

Anya Tralshawala
Rensselaer Polytechnic Institute, Information Technology & Web Science (ITWS), Troy, NY, United States

Abstract and Problem Area

As we know, quicker ambulance response times lead to greater survival rates. If we decrease response time by one minute, survival rates of patients increase by 20-30%. We can use predictive analytics to estimate ambulance response times as a function of time, location, operational and environmental factors. By predicting response times, we can determine where to assign resources.Current models don't account for socio-demographic or political factors and by analyzing these we can improve models and increase the chance of survival.
It would also be interesting to see how political alignment and congressional district may impact response times. We hypothesize that response times would be slower in districts that lean red or more republican since democrats believe more in government intervention and invest in services that are controlled by the government such as EMS. During pre-processing some data were merged and joined on the congressional district column to allow us to add congressional district to the EMS data.

The Data

- The data were obtained by NYC Open data.
  - Collected from EMS Computer Aided Dispatch System in NYC from 2008-2016.
  - This data has 23.3 million rows with 31 columns

Field Name	Field Description
INITIAL_CALL_TYPE *	The call type assigned at the time of incident creation.
INITIAL_SEVERITY_LEVEL_CODE	The segment(priority) assigned at the time of incident creation.
FINAL_CALL_TYPE *	The call type at the time the incident closes.
FINAL_SEVERITY_LEVEL_CODE	The segment(priority) assigned at the time the incident closes.
INCIDENT_RESPONSE_SECONDS_QY	The time elapsed in seconds between the incident_datetime and the first_on_scene_datetime.
INCIDENT_TRAVEL_TM_SECONDS_QY	The time elapsed in seconds between the first_assignment_datetime and the first_on_scene_datetime.
BOROUGH	The borough of the incident location.
INCIDENT_DISPATCH_AREA	The dispatch area of the incident.
ZIPCODE	The zip code of the incident.
POLICEPRECINCT	The police precinct of the incident.
CONGRESSIONALDISTRICT	The congressional district.

- New York State Elected Officials and Congressional District Data obtained by election.ny.gov
  - These data were collected in Nov. 2022 following the midterm elections in early 2022

Exploratory Data Analysis

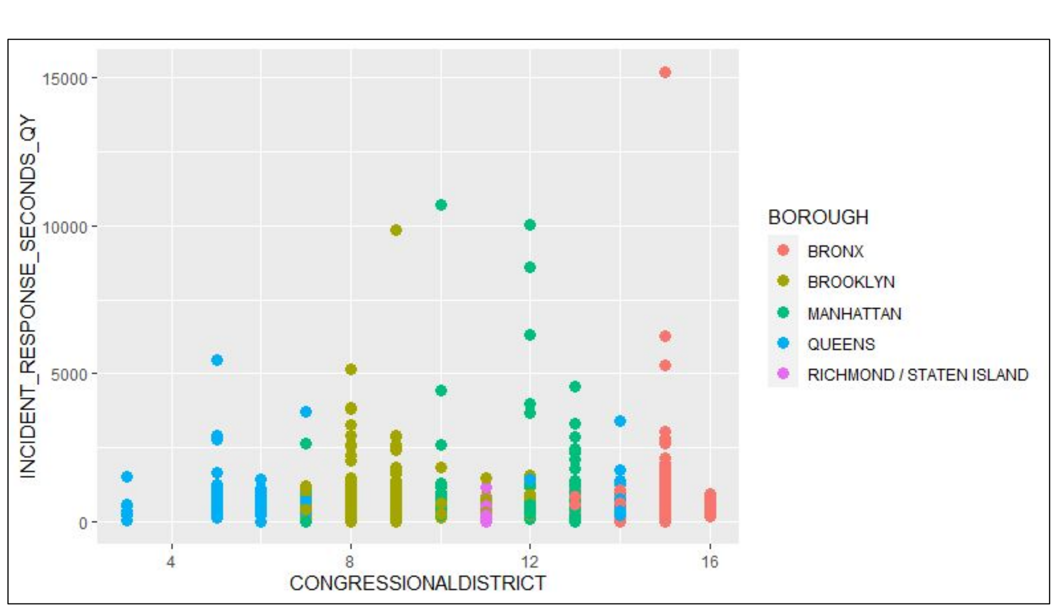

Figure 1 - Congressional District Vs Response Time classified by Borough

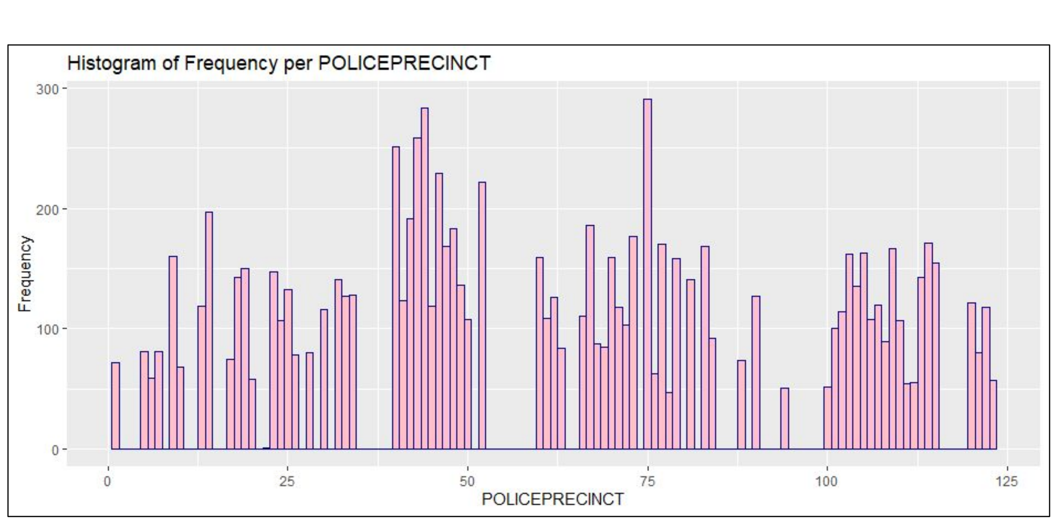

Figure 2 - Histogram of Frequency per Police Precinct

Models

1. Multivariate Regression

where INCIDENT_RESPONSE_SECONDS_QY is the **dependent/ response variable** for the regression model and INITIAL_SEVERITY_LEVEL_CODE, FINAL_SEVERITY_LEVEL_CODE, INCIDENT_TRAVEL_TM_SECONDS_QY are **independent/ predictors variables**

Residual standard error: *913.5 on 9557 degrees of freedom*
Adjusted R-Squared Value: 0.2529
p-value: 2.2e-16

```
> summary(model1)

Call:
lm(formula = INCIDENT_RESPONSE_SECONDS_QY ~ INITIAL_SEVERITY_LEVEL_CODE +
    FINAL_SEVERITY_LEVEL_CODE + INCIDENT_TRAVEL_TM_SECONDS_QY,
    data = sample)

Residuals:
    Min      1Q  Median      3Q     Max
-2287.8  -259.4  -139.9   -15.8 21669.0

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -181.31499   26.89010  -6.743 1.64e-11
INITIAL_SEVERITY_LEVEL_CODE    81.35121   14.44120   5.633 1.82e-08
FINAL_SEVERITY_LEVEL_CODE      -1.64091   14.39666  -0.114    0.909
INCIDENT_TRAVEL_TM_SECONDS_QY   1.15830    0.02231  51.930  < 2e-16

(Intercept)                   ***
INITIAL_SEVERITY_LEVEL_CODE   ***
FINAL_SEVERITY_LEVEL_CODE
INCIDENT_TRAVEL_TM_SECONDS_QY ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 913.5 on 9557 degrees of freedom
  (439 observations deleted due to missingness)
Multiple R-squared:  0.2531,    Adjusted R-squared:  0.2529
F-statistic:  1079 on 3 and 9557 DF,  p-value: < 2.2e-16
```

Figure 3 - Multivariate Regression for Model 1

2. K-Means Clustering Model

on a subset of EMS data, including the following fields: INCIDENT_RESPONSE_SECONDS_QY, INITIAL_SEVERITY_LEVEL_CODE, FINAL_SEVERITY_LEVEL_CODE, INCIDENT_TRAVEL_TM_SECONDS_QY, POLICEPRECINCT, ZIPCODE, CONGRESSIONALDISTRICT, & BOROUGH

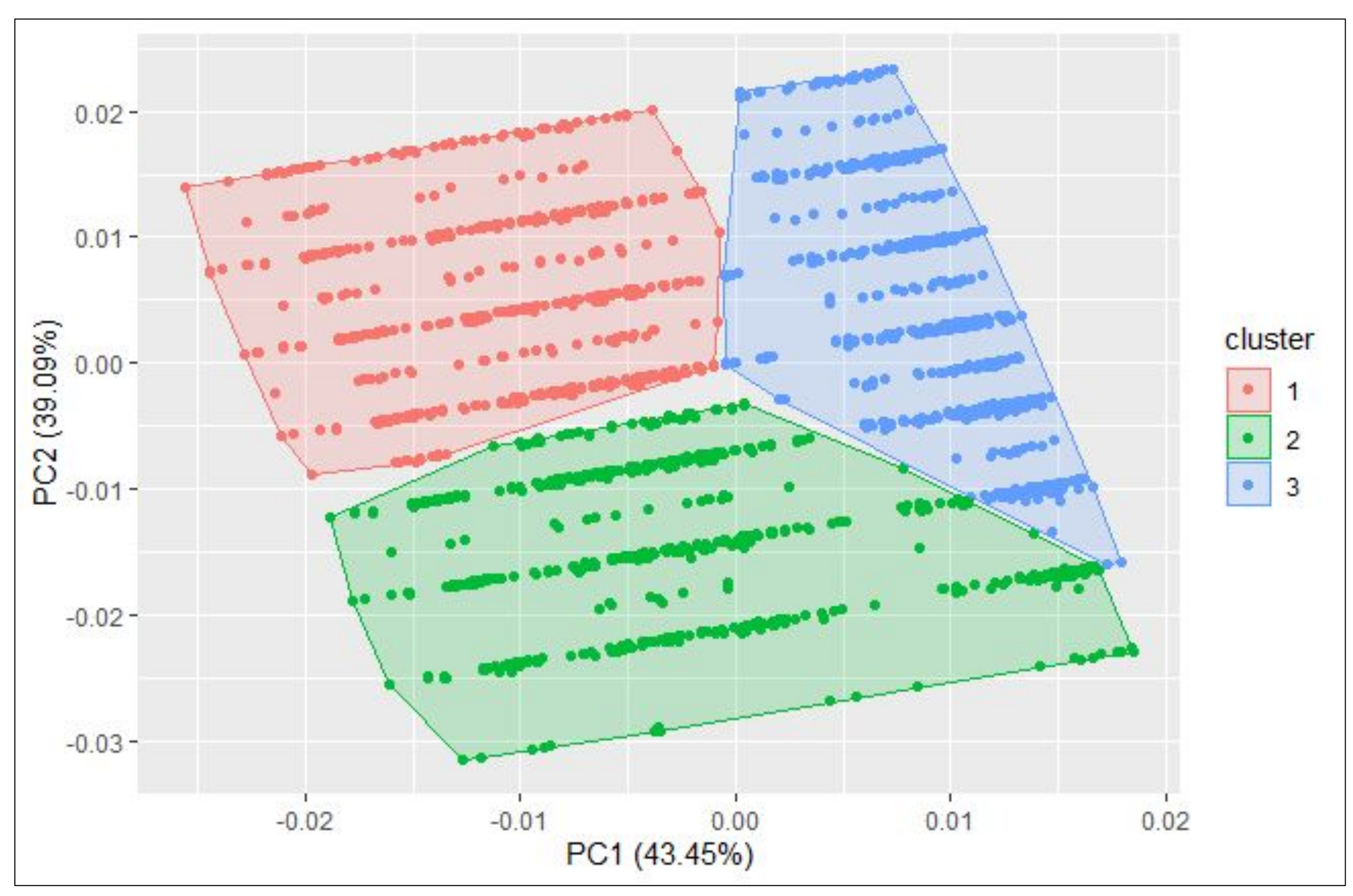

Figure 4 - K-Means Cluster Plot where K =3

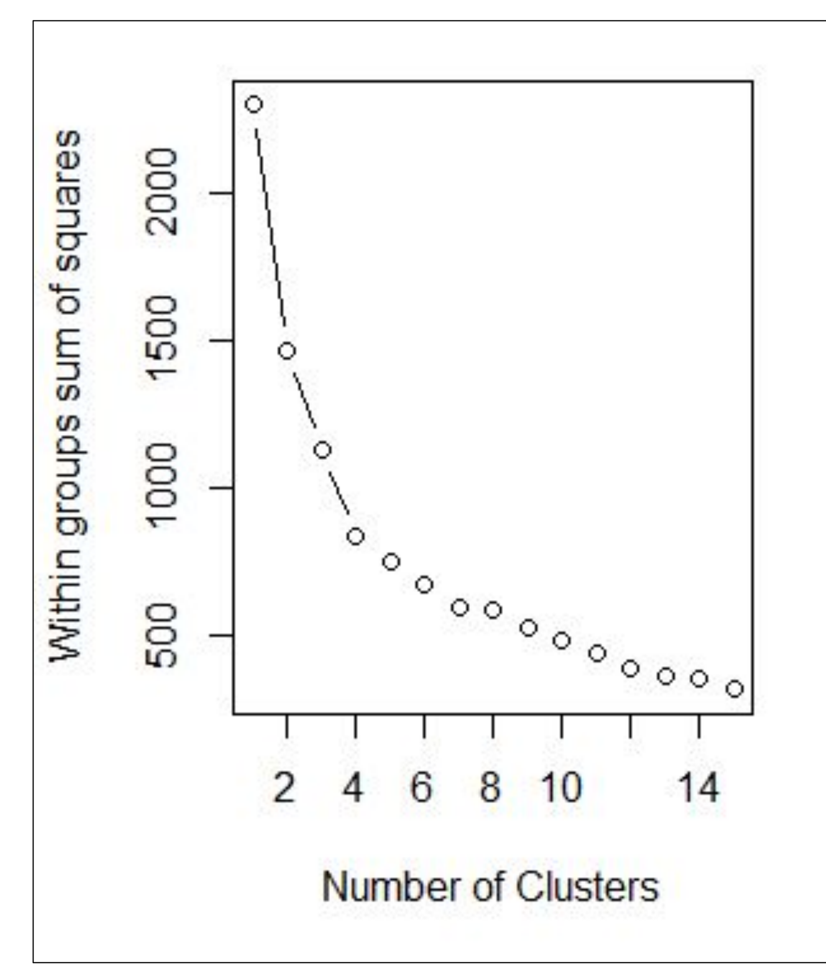

Figure 5 - Number of Clusters vs WSS Plot

3. KNN Classification

by response time speed classification:

fast, mid, or slow
fast → ***0 - 5 mins***
mid → ***5-15 mins***
slow → ***> 15 mins***

Based on the model here the misclassification error was **0.14** therefore accuracy = **86%**

```
KNNpred
fast  mid slow
 537 1902  387
```

Figure 6 - KNNpred response time

```
        KNNtestlabel
KNNpred fast  mid slow
   fast  426  107    4
   mid   139 1661  102
   slow    3   38  346
```

Figure 7 - KNNpred Confusion Matrix

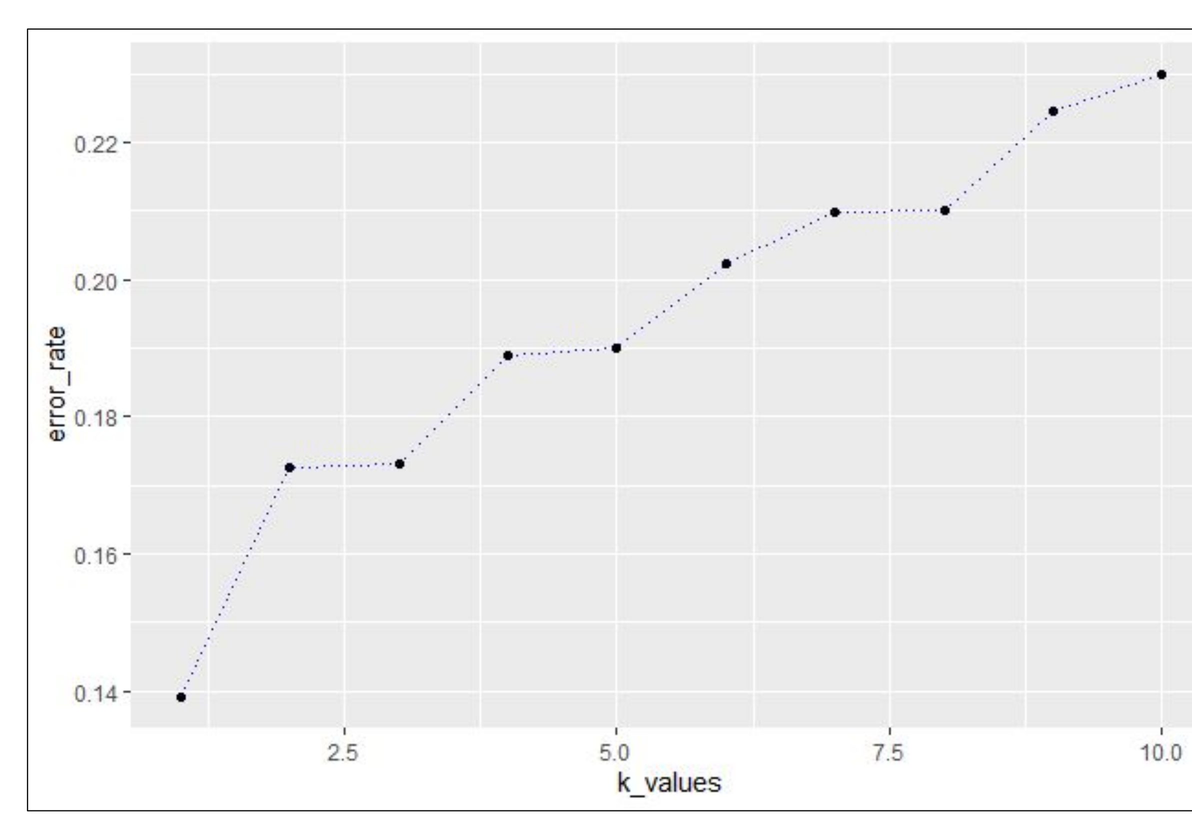

Figure 8 - K-Values vs Error_rate

Conclusion

Prior to analysis, it was hypothesized that political factors and which way the district leaned; red or blue would make an impact on EMS response times along with data like severity level, travel time to the incident, which borough the incident was in, and police precinct to name a few. After conducting multivariate regression and feeding in a model using initial, final severity level, and incident travel time it was found that there is not a very strong linear correlation. The adjusted R-Squared value as seen in the modeling above was 0.2529. This means that there is some correlation, just not a strong one. The next model generated was K-means clustering. PC1 had a variability of 43.45 and PC2 had a variability of 39.09. Currently, the K-means clustering model was generated with K = 3. The clustering could be modeled using various K values as per the elbow plot to see if variability increased. Lastly, according to the models you can see that KNN classification is pretty accurate when you select the appropriate k-value and parameters to train the model. In the case of the model generated above the accuracy was 86%. More factors could have been added to this model to see if perhaps more parameters increase accuracy. Multiple KNN classification models could be run simultaneously with different factors to see which factors are influential. In the future, if we were to merge congressional district data with EMS data and re-run these models we would make sure the data are accurate to the 2008-2016 time frame and not 2022. Some districts may have swung to the opposite political party and this might have impacted the model and resulted in no clear correlation being present when the incidents actually occurred. This was clearly an oversight when generating the datasets and models. In conclusion, by generating models we can determine where to assign resources to decrease response times and increase rates of survival. We can refine these models and increase accuracy to clearly see how factors like political affiliation, sociodemographics and incident severity impact response times.

Glossary:
R – A program to process data and perform statistical analysis
ggplot library - an open-source package in R used for displaying statistical data

Resources:
EMS Data: https://mediasite.mms.rpi.edu/mediasite/Play/77c72d7a29d44838a872d3a19e3086f31d
NYC Open Data: https://data.cityofnewyork.us/Public-Safety/EMS-Incident-Dispatch-Data/76xm-jjuj
Congressional District Data: https://www.elections.ny.gov/district-map.html
ggplot library: https://ggplot2.tidyverse.org/
GitHub: https://github.com/anya-tralshawala/DataAnalytics_TermProject_Anya_Tralshawala