Final Project

Tasheena Narraidoo, Jeffrey Lancaster, Luke Haggerty

December 22, 2017

Load all packages

We have used `tidyverse` and `lubridate` for data cleaning & manipulation and for basic EDA, `ggmap` for additional EDA and `nnet` for our model. To compute the multi-class log loss for our cross validation, we used the `ModelMetrics` package.

Load data and perform data cleaning

We are going to create three more variables (`Year` , `Month` , `Hour`) from the given `Dates` variables for EDA purposes.

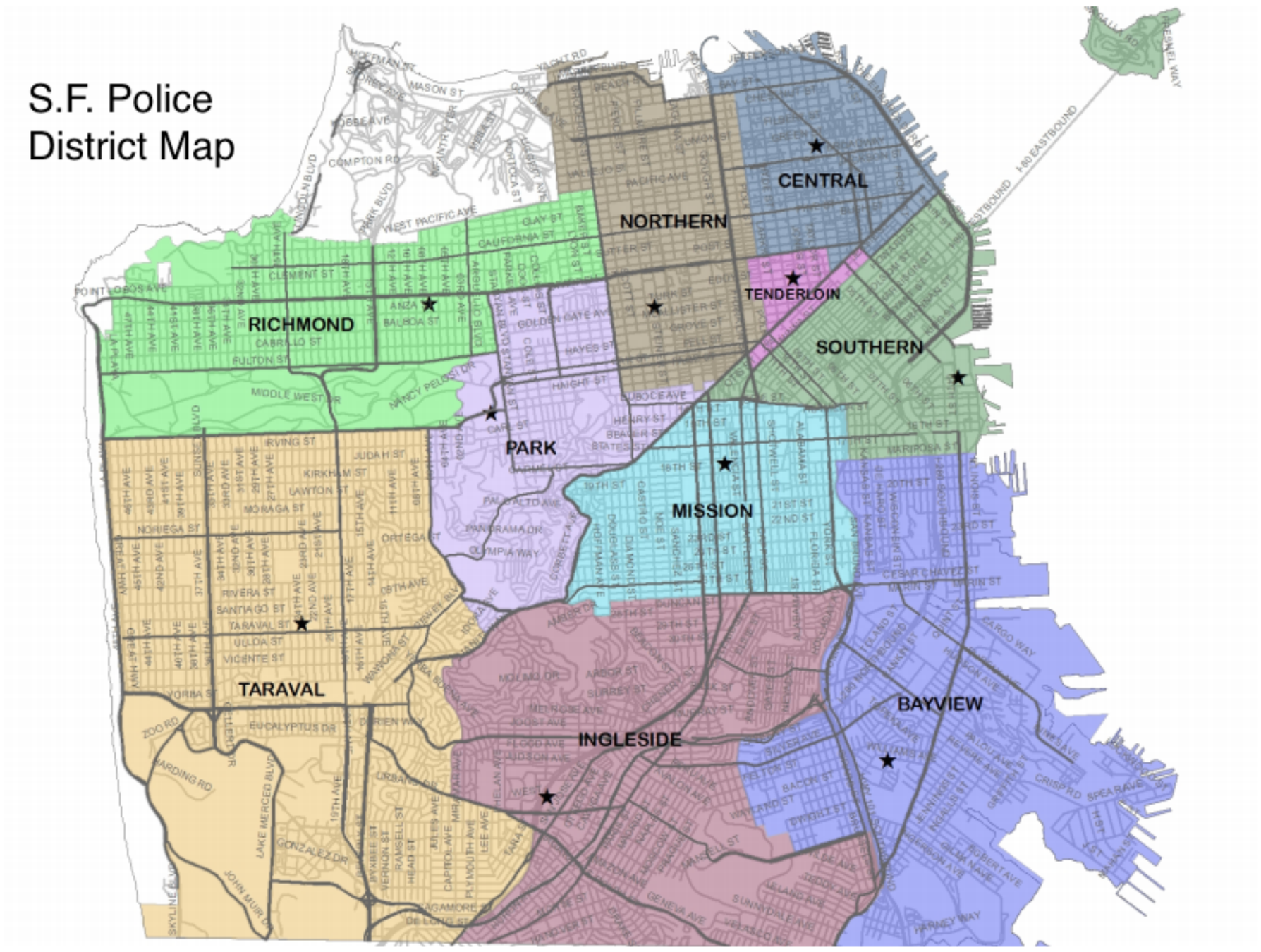
EDA visualizations and tables

Our original test data has nine variables as seen in the table below. Our test set has the same variablees except for `Category` (our response variable), `Descrip` and `Resolution` as the two latter variables would include the category of the crime reported. Each observation in our data set represents a crime reported between 1/1/2003 and 5/13/2015. Our train set has `878,049` observations while our test set has `884,262` observations. The weeks were numbered and crimes reported in odd-numbered weeks formed the train set and the rest formed our test set.

Original Variables with Description

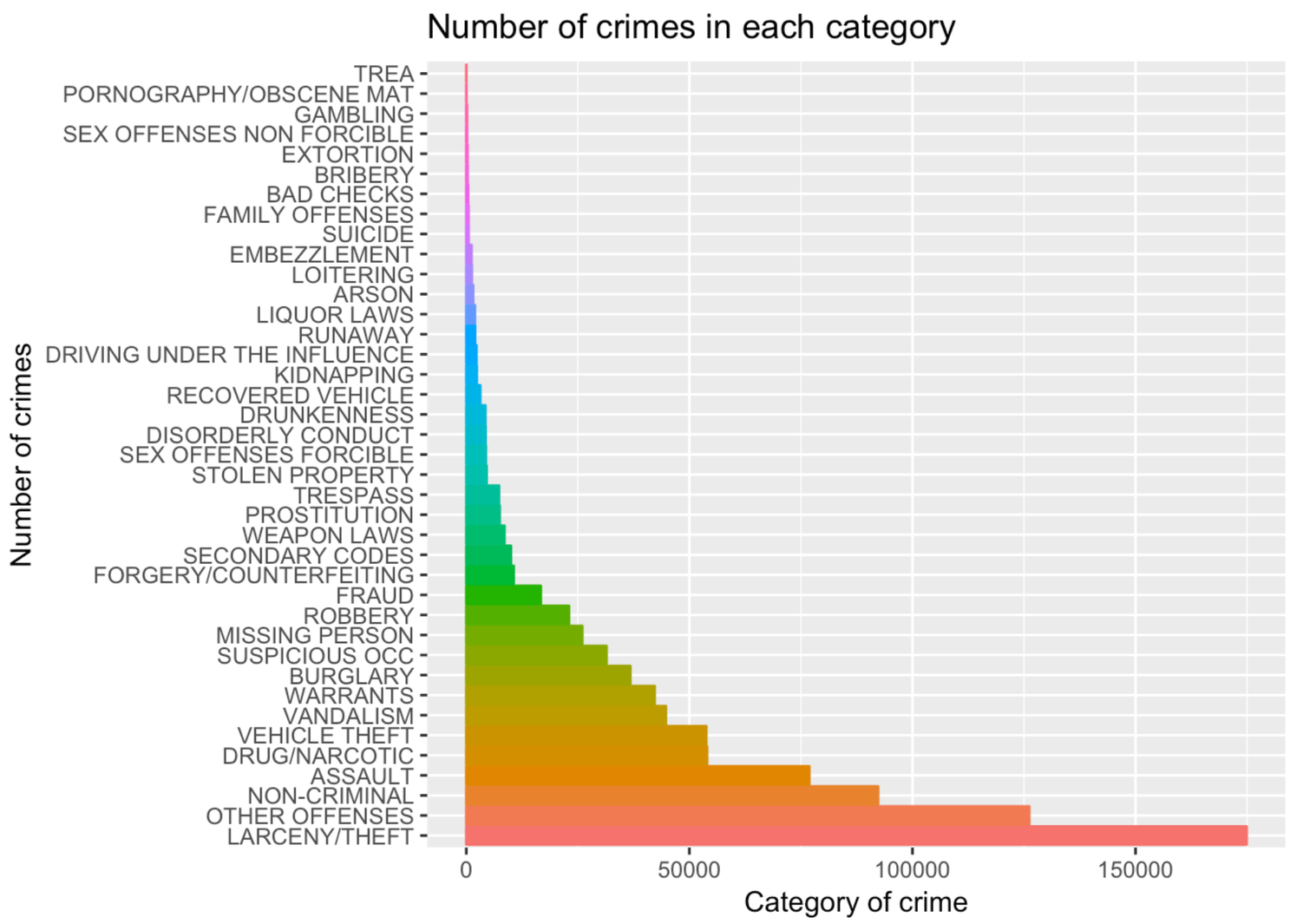
Name	Description
Dates	timestamp of the crime incident
Category	category of the crime incident (our reponse variable)
Descript	detailed description of the crime incident
DayOfWeek	the day of the week
PdDistrict	name of the Police Department District
Resolution	how the crime incident was resolved
Address	the approximate street address of the crime incident
X	Longitude
Y	Latitude

We next looked at a map showing San Francisco broken down by Police Districts. This map gives us a better idea on the scope of the project and helps us to visually identify the Police District names found in our data set. One interesting observation is that the richest district in the city, Presidio, is the only part of the map below without a police district assigned to it. Due to this, we have decided to use Police District as our variable for crime location instead of `Address` or `longitude` and `latitude` .

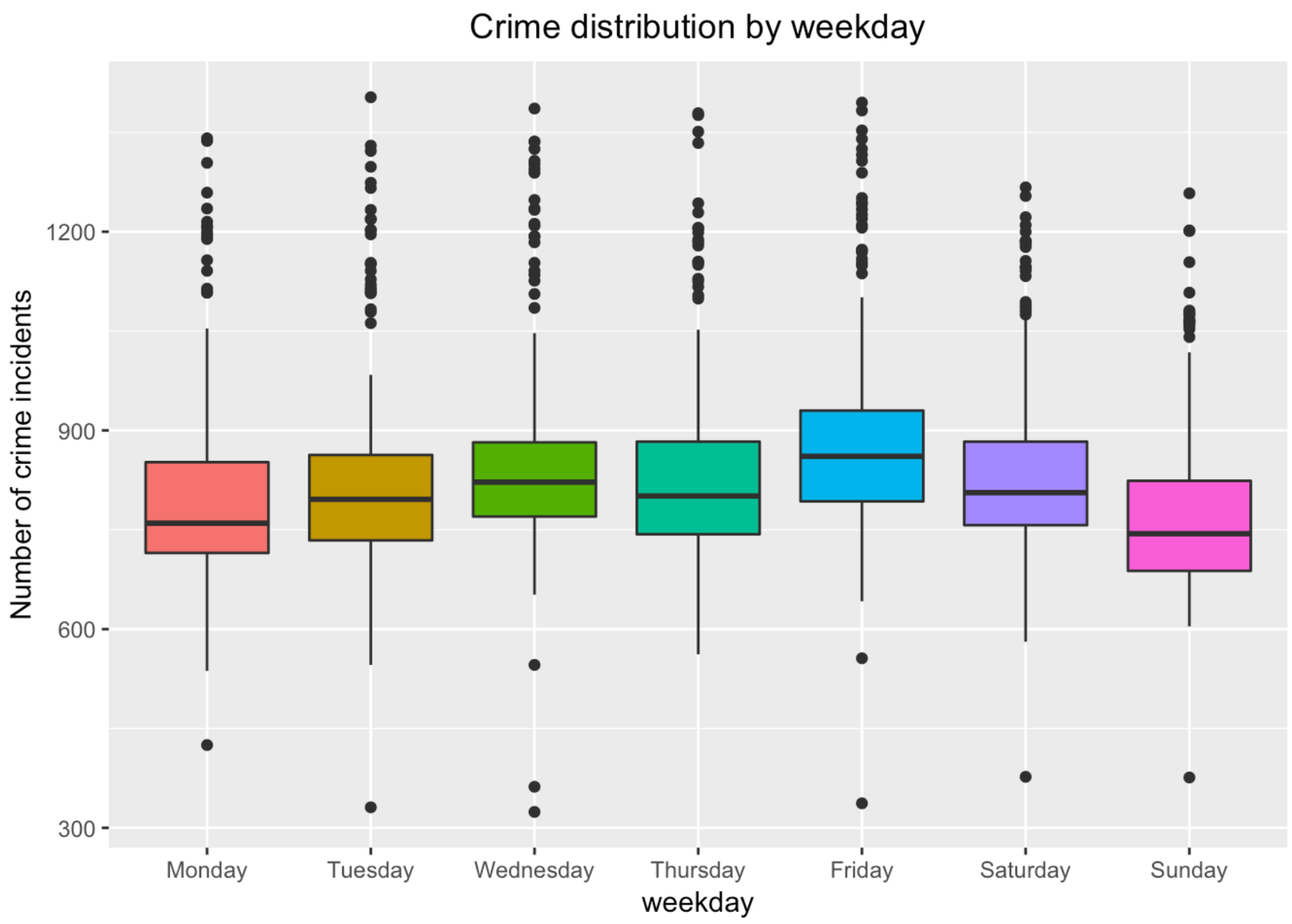


San Francisco Police District Map.

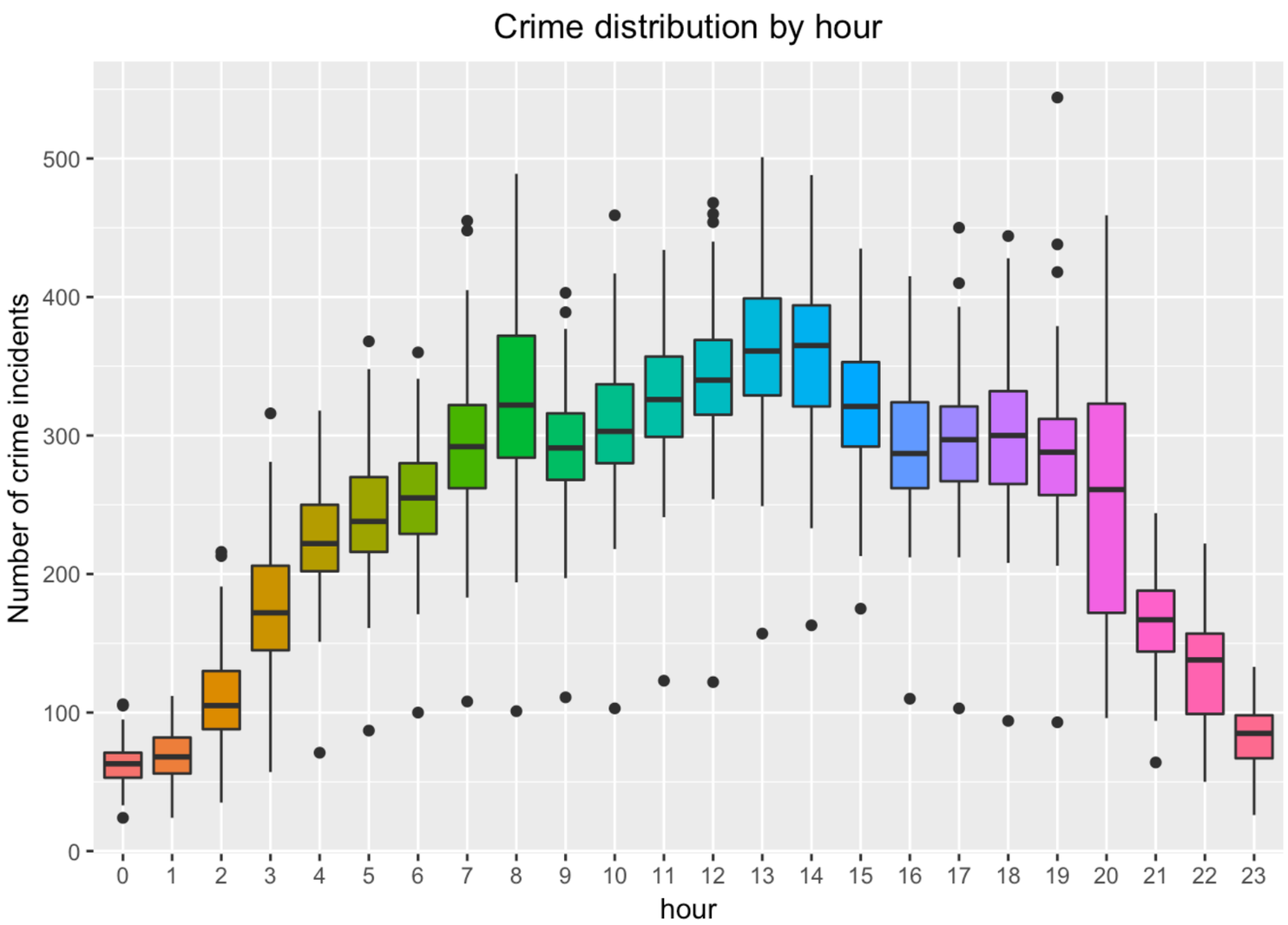
Our next step was to look at the breakdown of crime by `Category` . This allows us to identify the most and least common type of crimes reported. We are also able to understand how our observations are divided.



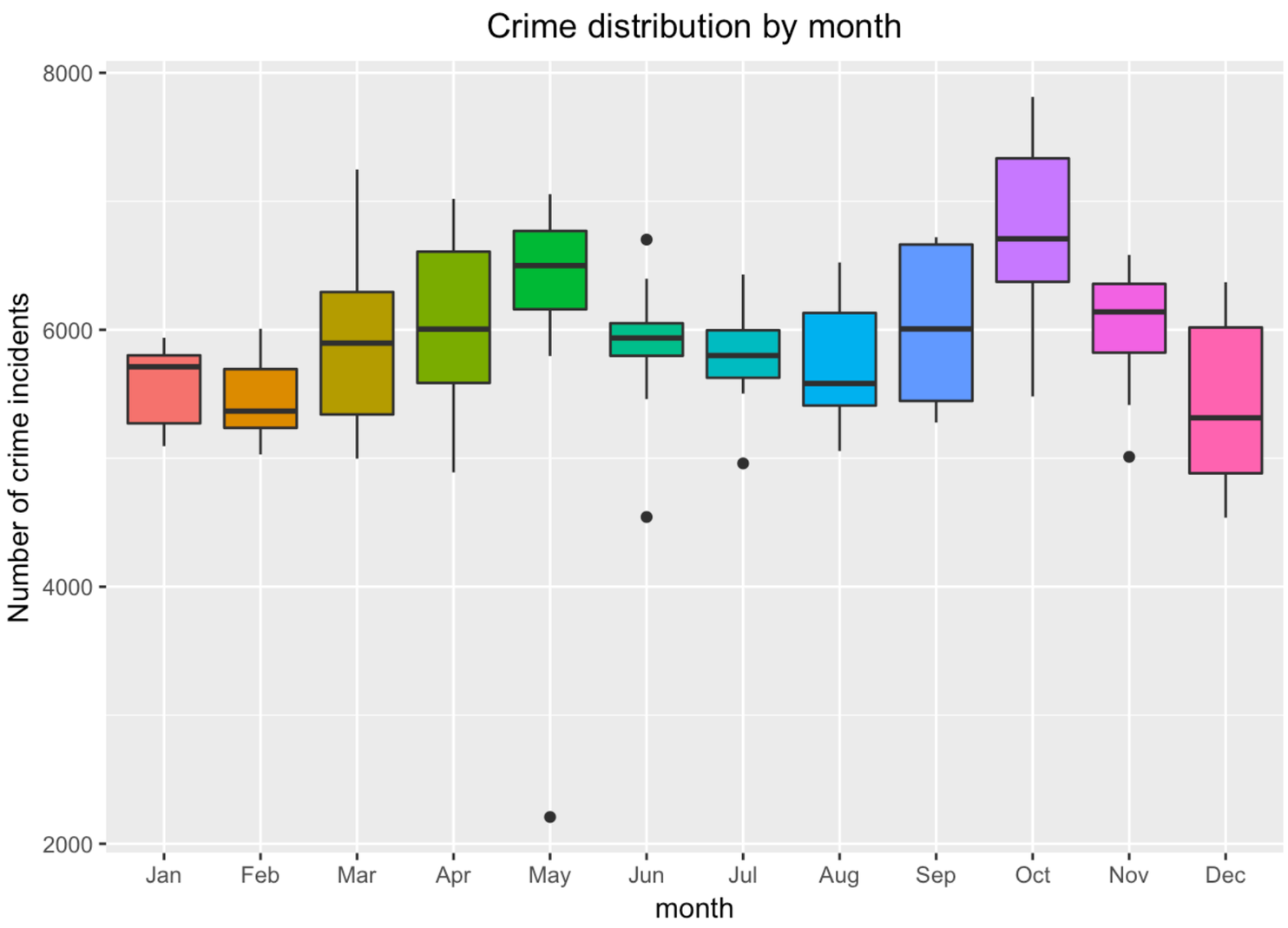
From the graph above we see that `LARCENY/THEFT` is the Number 1 most reported crime category. The bottom 10 categories seem to have low reported number of incidents. Indeed, the top 10 crime categories accounted for over 83 % of total number of crimes reported and the top 20 crime categories accounted for over 96%. We have therefore decided to concentrate on predicting top crimes when making our model to increase accuracy. Theft seems to be the number 1 crime reported as `LARCENY/THEFT`, `BURGLARY`, `VEHICLE THEFT`, `ROBBERY` are all thefts and they are in our Top 15 crimes reported. We will use this broad categorization to analyze type of crime against the different predictor variables such as the hour crimes were committed, day of week crimes occurred, and area in which crimes were reported using Police Dept District.



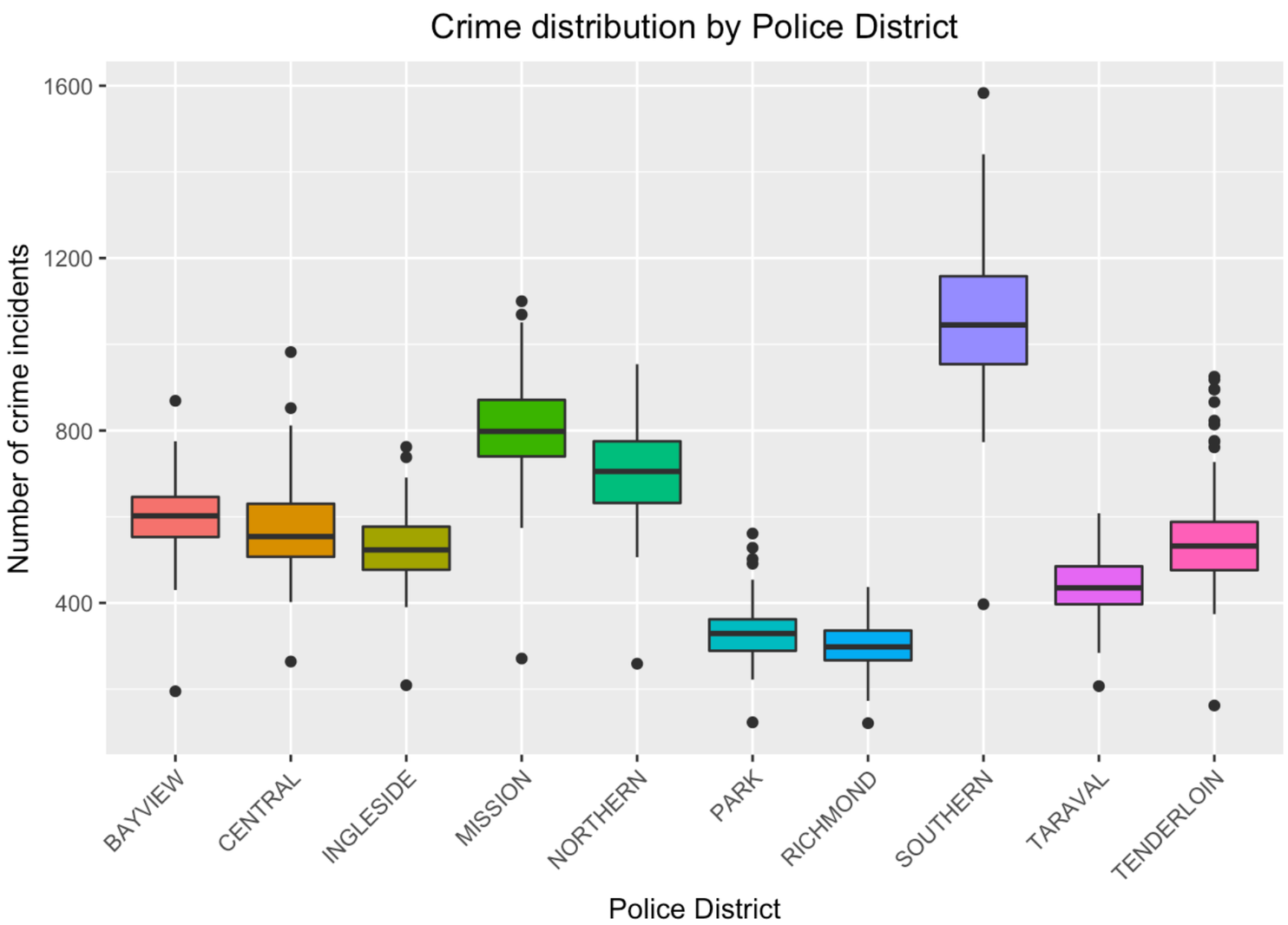
From the weekday graph it would seem that Sunday has seen the lowest median of crimes committed. This is probably because most people are at home on Sunday. Friday has the highest median, and this is probably because people are most likely to go out on Fridays. So `weekday` would be a good predictor of theft, and crimes by extension.



From the hour of the day breakdown we notice that most crimes occur during the day (from 8 am to around 6 pm). This is the time window when people commute to work and are at work. So `hour` would be a good predictor of theft, and crimes by extension.



From the `Crime distribution by month` we notice that October has the highest median. That seemed odd considering there are no major holidays in October. We thought this might be due to the way the training and test set were split. Odd numbered weeks were assigned to the train set and the rest to the test set. So major holiday weeks might have gone in the test set. So `Month` would not be a good predictor.



The Southern District seems to be leading in terms of number of crimes being reported. From the distribution of reported crime by district, it would seem that Police District would be a good preditor of crime as we clearly seem different patterns in the overall distribution.

From the above plots, `PdDistrict`, `DayOfWeek` and `Hour` seem to indicate when thefts are highly to occur. Since our broad definition of thefts account for most crimes reported, we believe that these variables would be good indicators of the type of crime in general, likely to have been committed.

Ultimate model

For our ultimate model, we have decided to use `multinomial logistic regression` because our reponse variable is a categorical outcome variable. We have decided to use `PdDistrict`, `DayOfWeek` and `Hour` variables from our EDA as our predictor variables. Also, they were all significant predictor variables when we ran our `Anova` test. We used a sample of size 200,000, which is roughly a quarter of our train set. This decision was made due to computation time and we also used `maxit` to get more accurate results. Multinomial logistic regression is used to model categorical outcome variables, where the log odds of the outcomes are modeled as a linear combination of the predictor variables. We used the `multinom()` function from the `nnet` library to estimate a multinomial logistic regression model. We ran our multinorm model over 200 iterations to get a good model.

```
m1 <- multinom(Category ~ PdDistrict + DayOfWeek +  Hour,
               data = train_samp, maxit = 200)
```

Our Anove results:

```
## Analysis of Deviance Table (Type II tests)   ##    ## Response: Category   ##           LR Chisq  Df Pr(>Chisq)       ## PdDistrict  3154
```

Crossvalidation of ultimate model

For our cross-validation, we have decided to use 5-fold coss-validation. We divided our data into five equal folds each containing 40,000 unique observations. The sample were generated from the training sample of 200,000 observations we used to run our ultimate model on. Our cross-validation results was 2.8, in terms of multi-class log loss error.

```
## [1] "cvscore: "
```

```
## [1] 2.8
```

Create submission

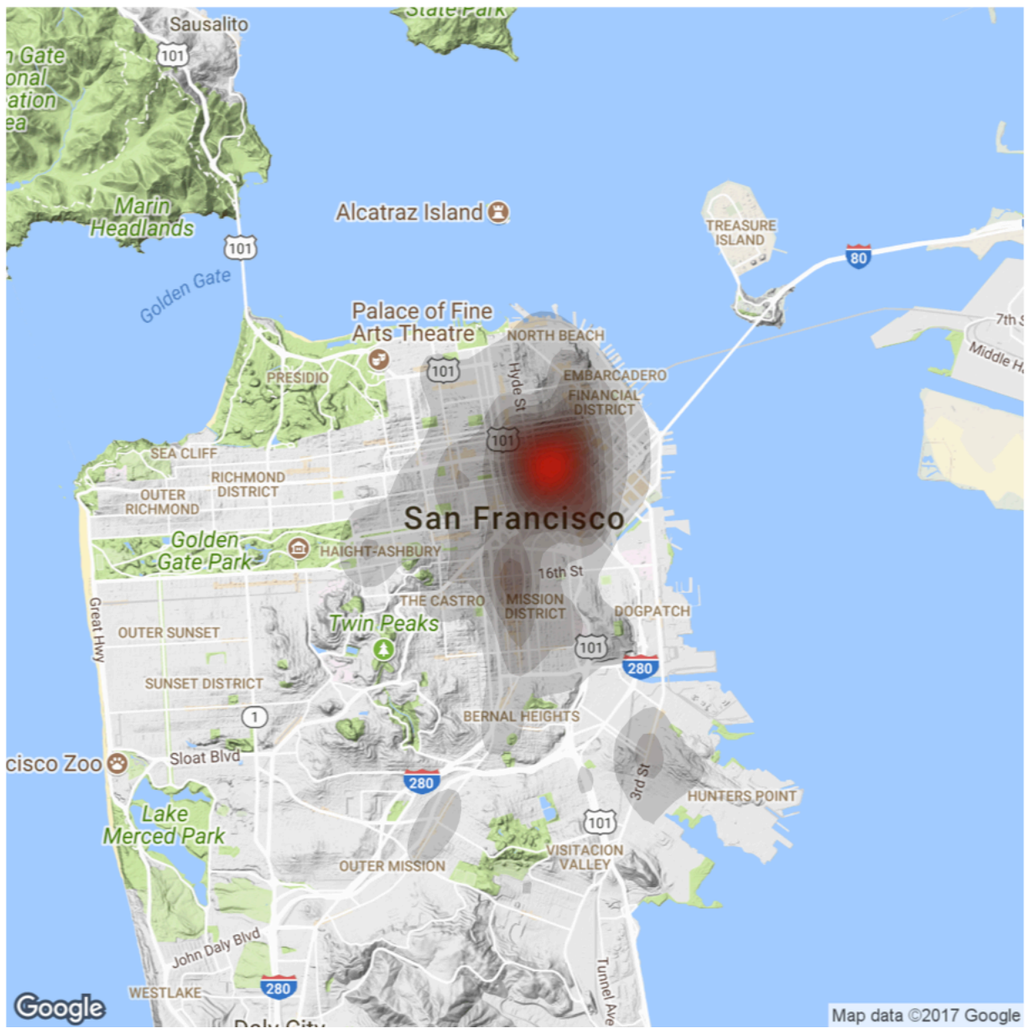
On submitting our best model on Kaggle, we got a score of around 2.6 which is close to our crossvalidation multi-class log loss of 2.8.

Citations and references

- Data sets. Retrieved from https://www.kaggle.com/c/sf-crime (https://www.kaggle.com/c/sf-crime)
- Multinomial Logistic Regression. Retrieved from https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/ (https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/)
- San Francisco Police Dept. District Map. Retrieved from http://hoodline.com/2015/07/citywide-sfpd-redistricting-to-take-effect-sunday (http://hoodline.com/2015/07/citywide-sfpd-redistricting-to-take-effect-sunday)

Supplementary materials

HeatMap of Crime in SanFrancisco



The heatmap plot allows us to visualize crime distribution in SanFrancisco. As we observed in the barplot, most crime occurrs around the districts in the top right of the city, especially the Tenderloin.