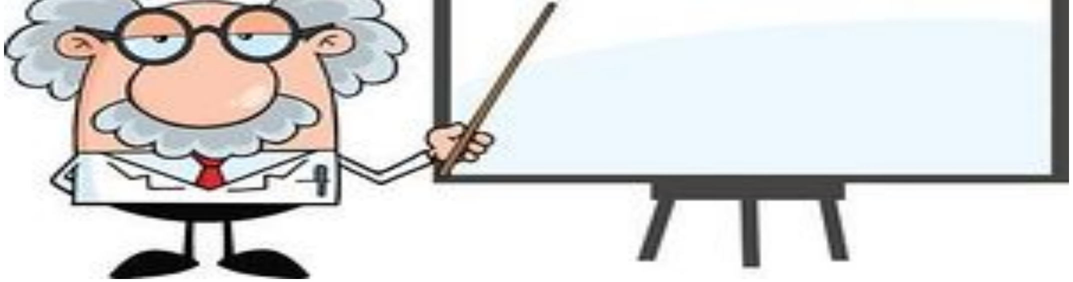
Crime Trends in Seattle

Anirudh Rao
Karan Murthy
Paromita Banerjee
Shreya Agarwal

Introduction

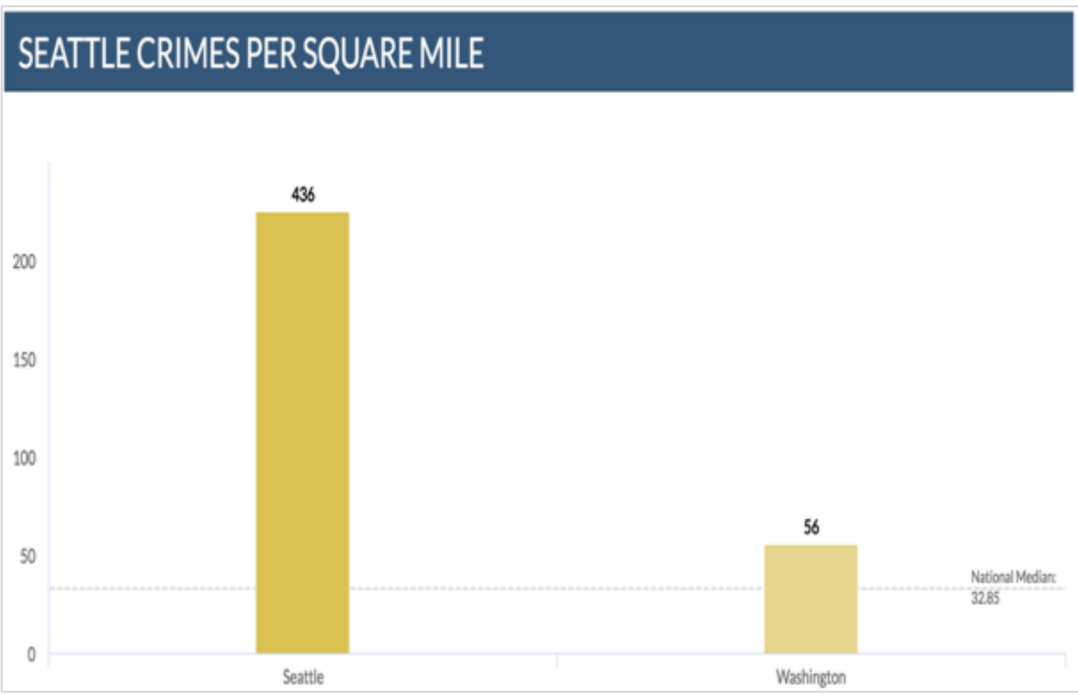


amazon

STARBUCKS COFFEE

Microsoft

BOEING



CRIME INDEX

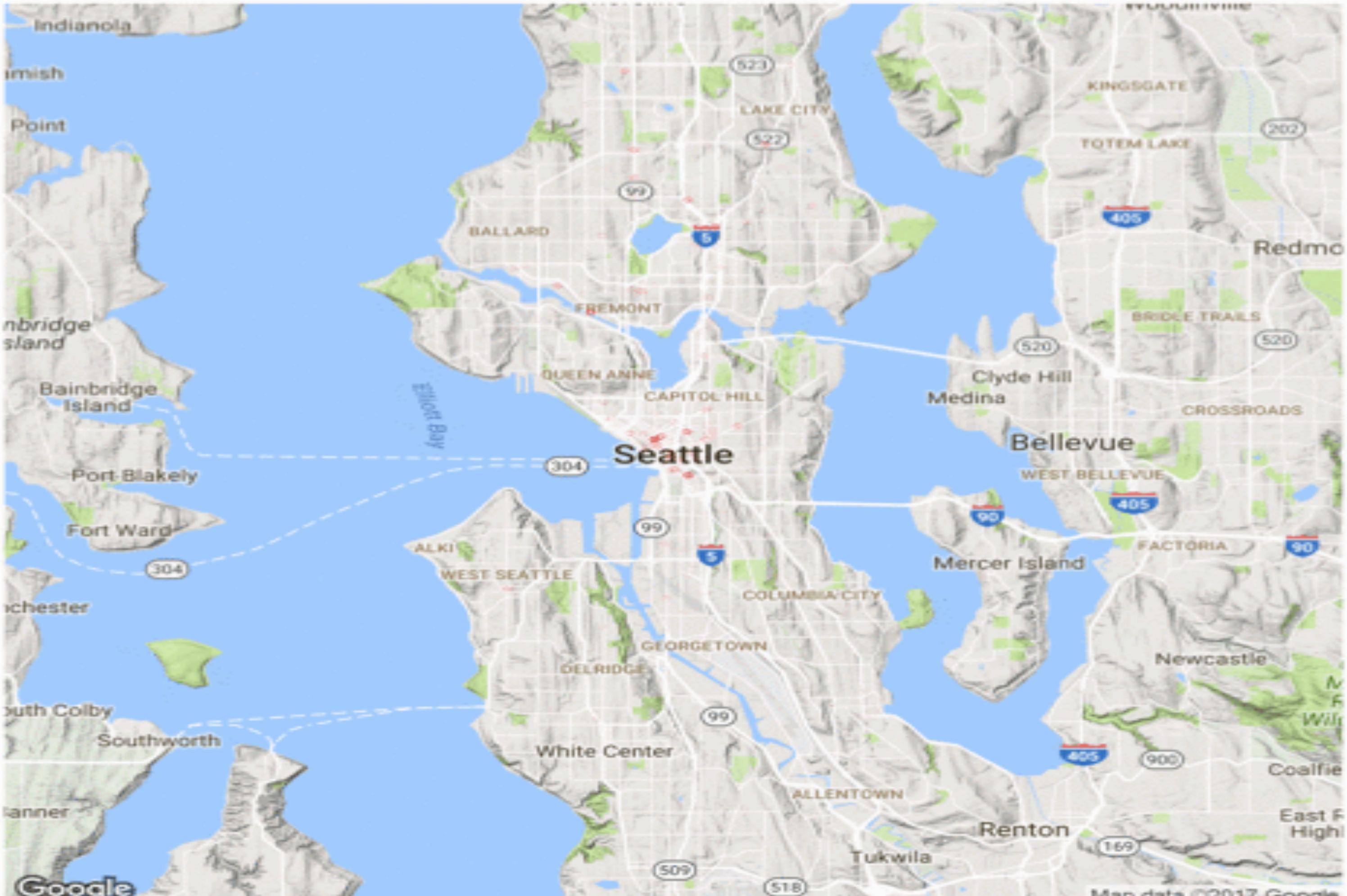
3

(100 is safest)

Safer than 3% of U.S. Cities

Month wise variation in crime:

Crime in Seattle for the month of January 2016



About the data

seattle.gov

Zillow

- For crime and school data
- The crime data has ~81k rows with 19 columns
- School data has 116 rows with 11 columns

- For housing property data
- 99 rows with 73 columns

Data Cleaning

- Used shape file to extract the neighborhoods for Latitude and Longitude.

- Multiple datasets joined by Neighborhood, hence common homogenizing the nomenclature.

- Rows with NA values for the variables of interest have been removed.

Research questions

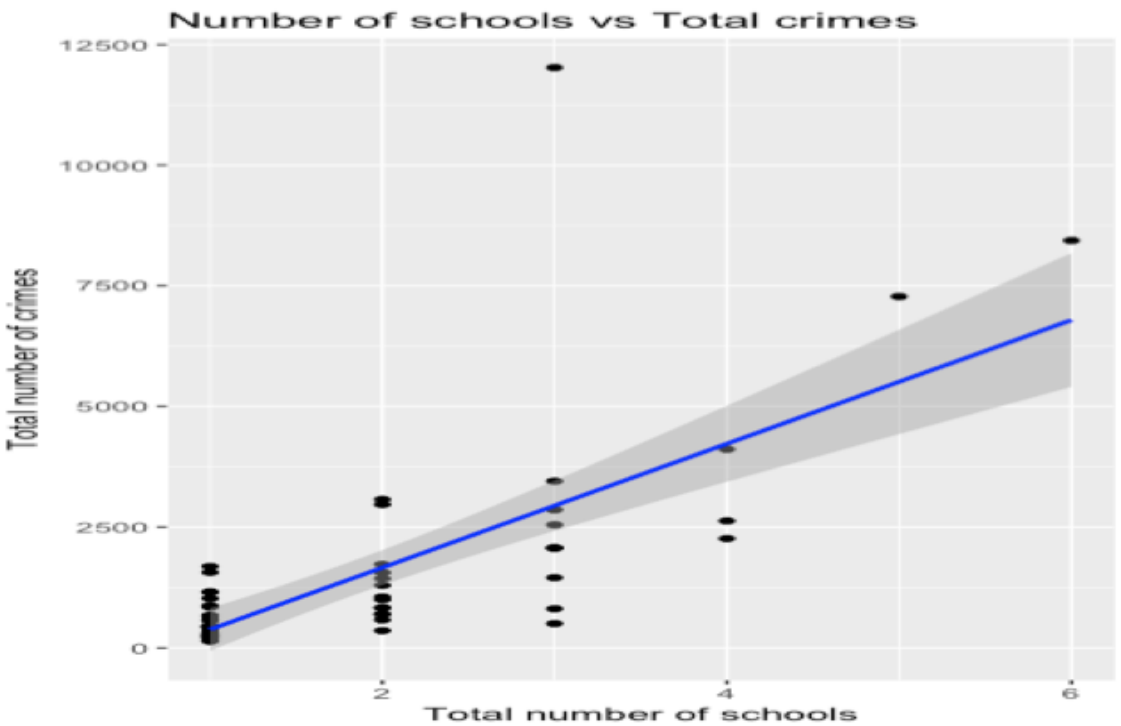
1. **Does the increase in the number of schools in an area increase the occurrence of crimes in that area?**
2. **Does the time of the day correlate with the occurrence of assaults?**
3. **Does crime in a neighborhood correlate with its property value?**
4. **Are there any seasonal variations in the distribution of Crime?**

Does the increase in the number of schools in an area increase the occurrence of crimes in that area?

H_0: The increase in the number of schools in an area does not increase the incidence of crimes in that area.

H_A: The increase in the number of schools in an area increases the incidence of crimes in that area.

Linear Regression Model




```
Call:
lm(formula = area_wise_school_crime$count_crime ~ area_wise_school_crime$Total_Schools)

Residuals:
    Min      1Q  Median      3Q     Max
-2439.8  -449.2   -96.5   189.0  9083.2

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -900.8      341.0  -2.642   0.0104 *
area_wise_school_crime$Total_Schools   1281.6      160.2   8.001 3.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1430 on 62 degrees of freedom
Multiple R-squared:  0.508,	Adjusted R-squared:  0.5001
F-statistic: 64.02 on 1 and 62 DF,  p-value: 3.942e-11
```

Multiple Regression Model

```
Call:
lm(formula = ml_join$count_crime ~ ml_join$Current + ml_join$count_school)

Residuals:
    Min      1Q  Median      3Q     Max 
-2464.9  -416.3     9.0   171.6  8959.5 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -7.334e+02  6.635e+02  -1.105    0.274    
ml_join$Current      -4.403e-04  8.691e-04  -0.507    0.614    
ml_join$count_school  1.350e+03  1.659e+02   8.137 3.21e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1428 on 59 degrees of freedom
Multiple R-squared:  0.5316,	Adjusted R-squared:  0.5158 
F-statistic: 33.49 on 2 and 59 DF,  p-value: 1.914e-10
```

Does the time of the day correlate with the occurrence of assaults?

H_0 *: Time of the day does not correlate with the occurrence of assaults*
H_A *: Time of the day correlate with the occurrence of assaults*

Assaults vs. Time of the day

▷ **Logistic Regression:**

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Response: Assault
Predictor: Time of day

```
crime_zillow_merged_df$Crime.Response <- rep(0, nrow(crime_zillow_merged_df))
crime_zillow_merged_df$Crime.Response[crime_zillow_merged_df$Summary.Offense.Code %in% c(1300)] <- 1
sub <- sample(nrow(crime_zillow_merged_df), (nrow(crime_zillow_merged_df) * 0.75))
training <- crime_zillow_merged_df[sub, ]
testing <- crime_zillow_merged_df[-sub, ]
```

```
> nrow(training)
[1] 44605
> nrow(testing)
[1] 14869
```

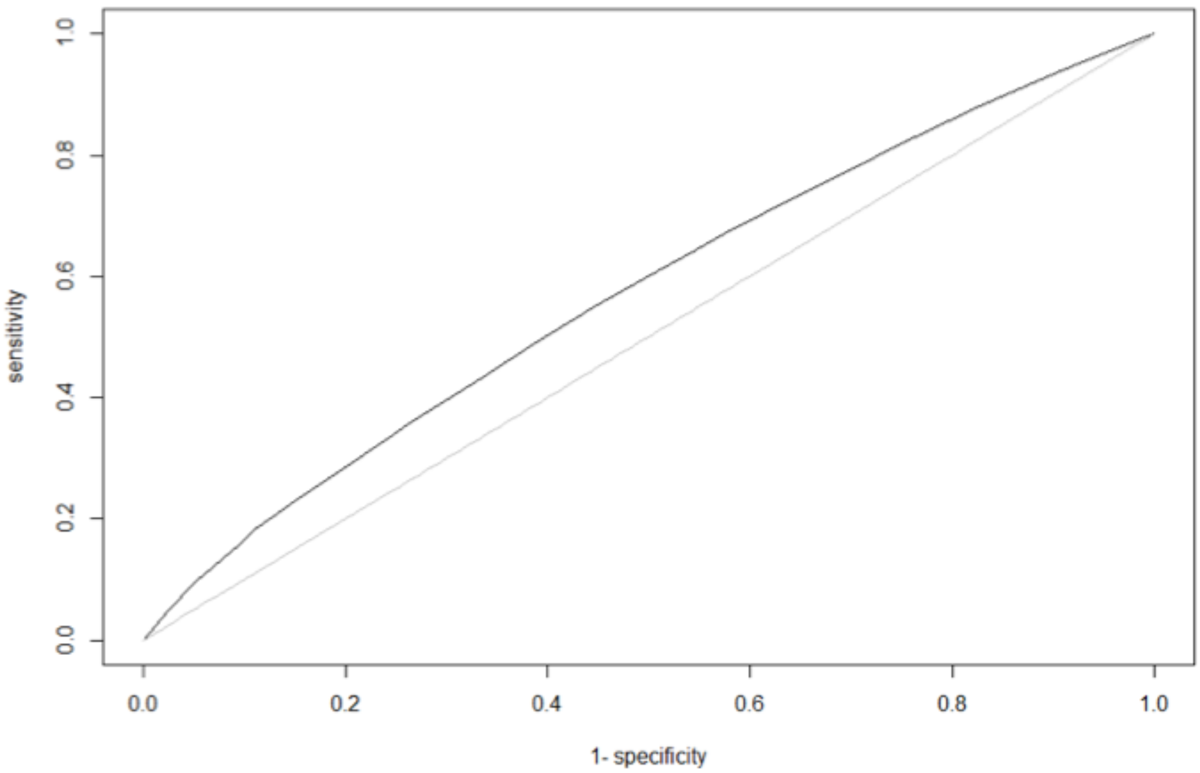
Model:

```
logmod<-glm(formula = training$Crime.Response ~ training$time_slots,
            data=training, family="binomial")
```

Assaults vs. Time of the Day

```
Call:
glm(formula = training$Crime.Response ~ training$time_slots,
    family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5859  -0.4451  -0.3999  -0.3841   2.4151

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -1.81040    0.07909 -22.889  < 2e-16 ***
training$time_slots1 PM to 2 PM       -0.14935    0.10437  -1.431 0.152451
training$time_slots10 AM to 11 AM     -0.69692    0.12210  -5.708 1.14e-08 ***
training$time_slots10 PM to 11 PM     -0.73300    0.11293  -6.491 8.54e-11 ***
training$time_slots11 AM to 12 PM     -0.42711    0.11249  -3.797 0.000146 ***
training$time_slots11 PM to midnight  -0.40097    0.10930  -3.669 0.000244 ***
training$time_slots12 AM to 1 AM      -0.87732    0.11214  -7.823 5.14e-15 ***
training$time_slots12 PM to 1 PM      -0.75959    0.10813  -7.025 2.15e-12 ***
training$time_slots2 AM to 3 AM        0.13505    0.11304   1.195 0.232212
training$time_slots2 PM to 3 PM       -0.47564    0.10911  -4.359 1.31e-05 ***
training$time_slots3 AM to 4 AM       -0.43712    0.13935  -3.137 0.001708 **
training$time_slots3 PM to 4 PM       -0.61009    0.11108  -5.492 3.97e-08 ***
training$time_slots4 AM to 5 AM       -1.05037    0.18079  -5.810 6.25e-09 ***
training$time_slots4 PM to 5 PM       -0.42940    0.10712  -4.009 6.11e-05 ***
training$time_slots5 AM to 6 AM       -0.16273    0.13856  -1.174 0.240220
training$time_slots5 PM to 6 PM       -0.71449    0.10900  -6.555 5.57e-11 ***
training$time_slots6 AM to 7 AM       -0.45185    0.13818  -3.270 0.001076 **
training$time_slots6 PM to 7 PM       -0.54479    0.10556  -5.161 2.46e-07 ***
training$time_slots7 AM to 8 AM       -0.67535    0.13157  -5.133 2.85e-07 ***
training$time_slots7 PM to 8 PM       -0.59261    0.10885  -5.444 5.20e-08 ***
training$time_slots8 AM to 9 AM       -0.88729    0.13138  -6.754 1.44e-11 ***
training$time_slots8 PM to 9 PM       -0.94417    0.11473  -8.230  < 2e-16 ***
training$time_slots9 AM to 10 AM      -0.74542    0.12141  -6.140 8.28e-10 ***
training$time_slots9 PM to 10 PM      -0.47552    0.10452  -4.550 5.37e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26207  on 44604  degrees of freedom
Residual deviance: 25953  on 44581  degrees of freedom
AIC: 26001

Number of Fisher Scoring iterations: 5
```

Assaults vs. Time of the Day




```
> auc(rr_assault)
[1] 0.5725203
```

Assaults vs. Time of the day

Multiple Logistic Regression:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Response: Assault
Predictors: Time of day, Neighborhood

Model:

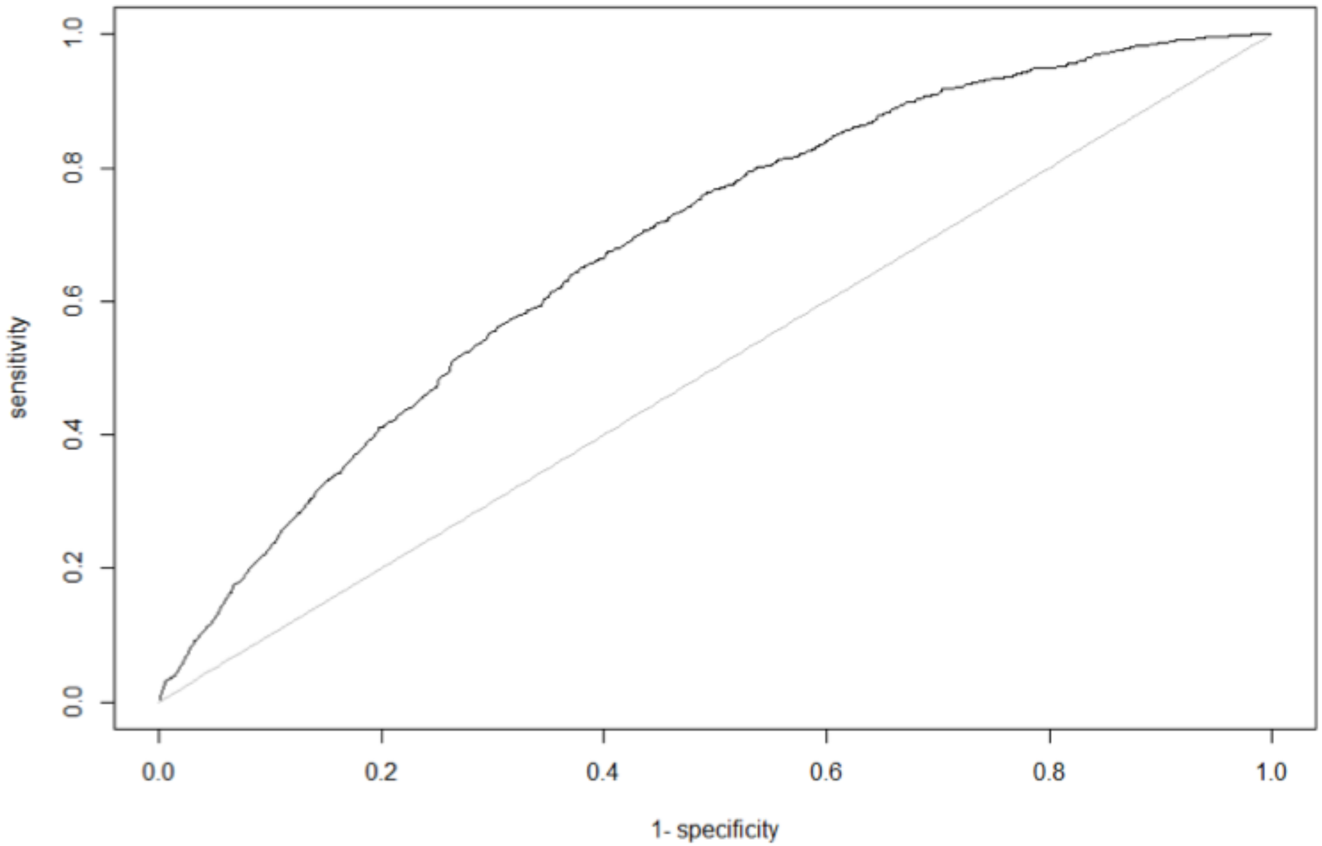
```
logmod_new<-glm(formula = training$Crime.Response ~ training$time_slots + training$Neighborhood,
                data=training, family="binomial")
```

Assaults vs. Time of the Day

```
Call:
glm(formula = training$Crime.Response ~ training$time_slots +
    training$Neighborhood, family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8445  -0.4926  -0.3804  -0.2630   3.3180

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.69377    0.14619 -11.586  < 2e-16 ***
training$time_slots1 PM to 2 PM    -0.15968    0.10680  -1.495 0.134887
training$time_slots10 AM to 11 AM  -0.77360    0.12421  -6.228 4.71e-10 ***
training$time_slots10 PM to 11 PM  -0.64103    0.11499  -5.575 2.48e-08 ***
training$time_slots11 AM to 12 PM  -0.48182    0.11483  -4.196 2.72e-05 ***
training$time_slots11 PM to midnight -0.36475  0.11141  -3.274 0.001060 **
training$time_slots12 AM to 1 AM   -0.83064    0.11398  -7.287 3.16e-13 ***
training$time_slots12 PM to 1 PM   -0.75215    0.11004  -6.835 8.20e-12 ***
training$time_slots2 AM to 3 AM     0.13573    0.11553   1.175 0.240032
training$time_slots2 PM to 3 PM    -0.48914    0.11147  -4.388 1.14e-05 ***
training$time_slots3 AM to 4 AM    -0.32399    0.14209  -2.280 0.022599 *
training$time_slots3 PM to 4 PM    -0.59311    0.11345  -5.228 1.71e-07 ***
training$time_slots4 AM to 5 AM    -0.98471    0.18325  -5.374 7.72e-08 ***
training$time_slots4 PM to 5 PM    -0.38962    0.10937  -3.562 0.000368 ***
training$time_slots5 AM to 6 AM    -0.19000    0.14142  -1.343 0.179125
training$time_slots5 PM to 6 PM    -0.68601    0.11118  -6.170 6.81e-10 ***
training$time_slots6 AM to 7 AM    -0.39658    0.14119  -2.809 0.004972 **
training$time_slots6 PM to 7 PM    -0.51788    0.10781  -4.804 1.56e-06 ***
training$time_slots7 AM to 8 AM    -0.69367    0.13395  -5.179 2.24e-07 ***
training$time_slots7 PM to 8 PM    -0.51421    0.11111  -4.628 3.70e-06 ***
training$time_slots8 AM to 9 AM    -0.87695    0.13339  -6.575 4.88e-11 ***
training$time_slots8 PM to 9 PM    -0.86026    0.11681  -7.364 1.78e-13 ***
training$time_slots9 AM to 10 AM   -0.75962    0.12352  -6.150 7.76e-10 ***
training$time_slots9 PM to 10 PM   -0.45237    0.10674  -4.238 2.26e-05 ***
training$NeighborhoodAlki          -0.50568    0.29291  -1.726 0.084279 .
training$NeighborhoodArbor Heights -1.78740    0.72487  -2.466 0.013670 *
training$NeighborhoodBeacon Hill   -1.13533    0.27534  -4.123 3.73e-05 ***
training$NeighborhoodBelltown       0.34743    0.13683   2.539 0.011113 *
```

```
training$NeighborhoodOlympic Hills        -0.05521    0.20778  -0.266 0.790453
training$NeighborhoodPhinney Ridge        -1.07295    0.30779  -3.486 0.000490 ***
training$NeighborhoodPinehurst            -0.37805    0.18088  -2.090 0.036611 *
training$NeighborhoodPortage Bay          -0.76272    0.47588  -1.603 0.108990
training$NeighborhoodRainier Beach        -1.19296    0.34373  -3.471 0.000519 ***
training$NeighborhoodRainier View         -2.47470    0.72091  -3.433 0.000598 ***
training$NeighborhoodRavenna              -1.07504    0.24487  -4.390 1.13e-05 ***
training$NeighborhoodRiverview             0.39888    0.22925   1.740 0.081874 .
training$NeighborhoodRoosevelt            -0.85898    0.24981  -3.439 0.000585 ***
training$NeighborhoodRoxhill              -0.46429    0.21282  -2.182 0.029138 *
training$NeighborhoodSeaview              -2.41103    0.59284  -4.067 4.76e-05 ***
training$NeighborhoodSeward Park          -3.21347    1.00983  -3.182 0.001462 **
training$NeighborhoodSouth Beacon Hill    -1.26840    0.43240  -2.933 0.003353 **
training$NeighborhoodSouth Delridge       -0.69407    0.23203  -2.991 0.002779 **
training$NeighborhoodSouth Park           -0.62834    0.24713  -2.543 0.011004 *
training$NeighborhoodSunset Hill          -1.29463    0.46994  -2.755 0.005871 **
training$NeighborhoodUniversity District  -0.18549    0.14427  -1.286 0.198549
training$NeighborhoodVictory Heights      -0.85701    0.29930  -2.863 0.004192 **
training$NeighborhoodView Ridge           -0.81022    0.40536  -1.999 0.045632 *
training$NeighborhoodWallingford          -0.75868    0.18815  -4.032 5.53e-05 ***
training$NeighborhoodWedgwood             -1.44687    0.43092  -3.358 0.000786 ***
training$NeighborhoodWest Queen Anne      -0.30337    0.24872  -1.220 0.222564
training$NeighborhoodWest Woodland        -0.44668    0.21567  -2.071 0.038348 *
training$NeighborhoodWestlake             -1.19989    0.52275  -2.295 0.021713 *
training$NeighborhoodWhittier Heights     -0.96989    0.34517  -2.810 0.004956 **
training$NeighborhoodWindermere          -14.34547  231.09733  -0.062 0.950503
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26207  on 44604  degrees of freedom
Residual deviance: 24641  on 44508  degrees of freedom
AIC: 24835

Number of Fisher Scoring iterations: 15
```

Assaults vs. Time of the Day




```
> auc(rr_new)
[1] 0.6839471
```

```
> auc(rr_assault)
[1] 0.5725203
```

Does crime in a neighborhood correlate with its property value?

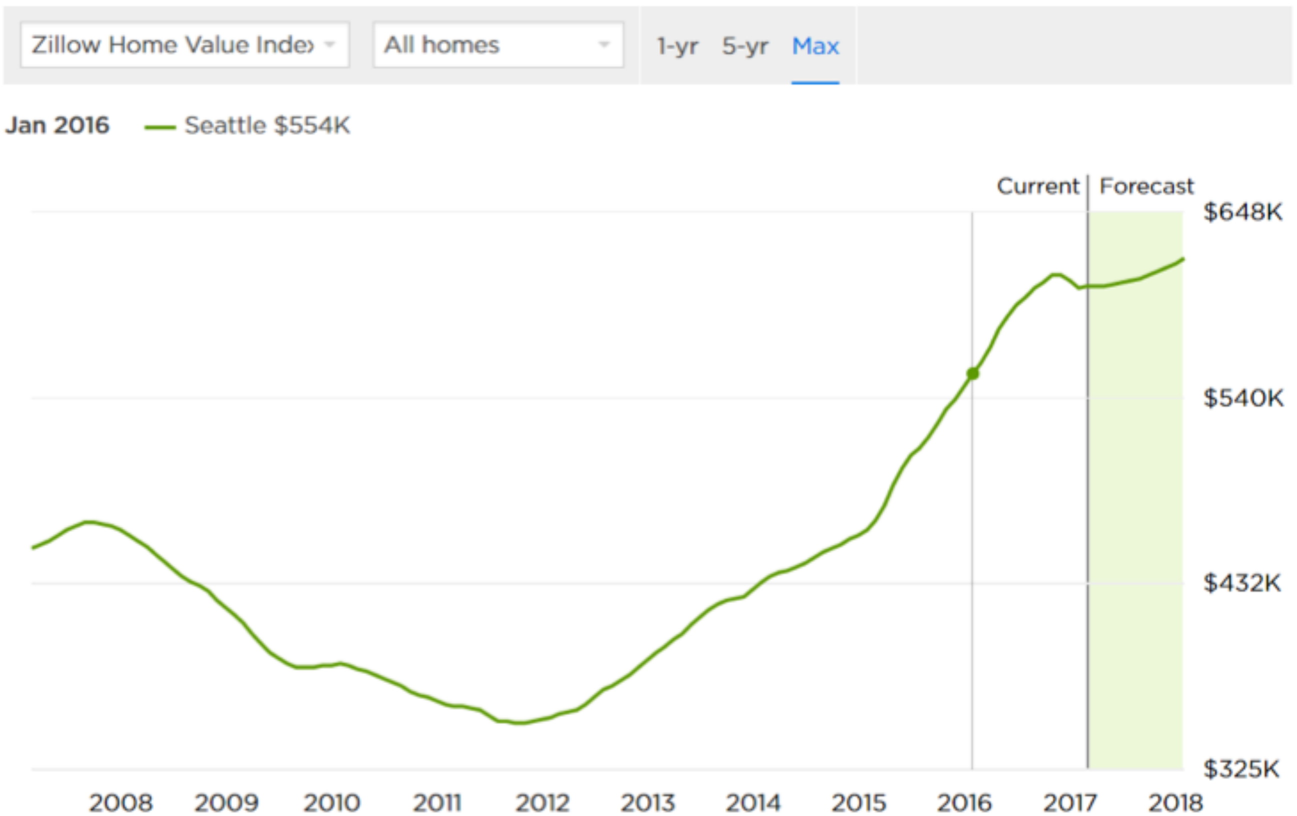
H_0: Crime in a neighborhood does not correlate with its property value
H_A: *Crime in a neighborhood correlates with its property value*

Top 3 crime categories in Seattle in 2016

Theft – 21254 incidences

Burglary – 10374 incidences

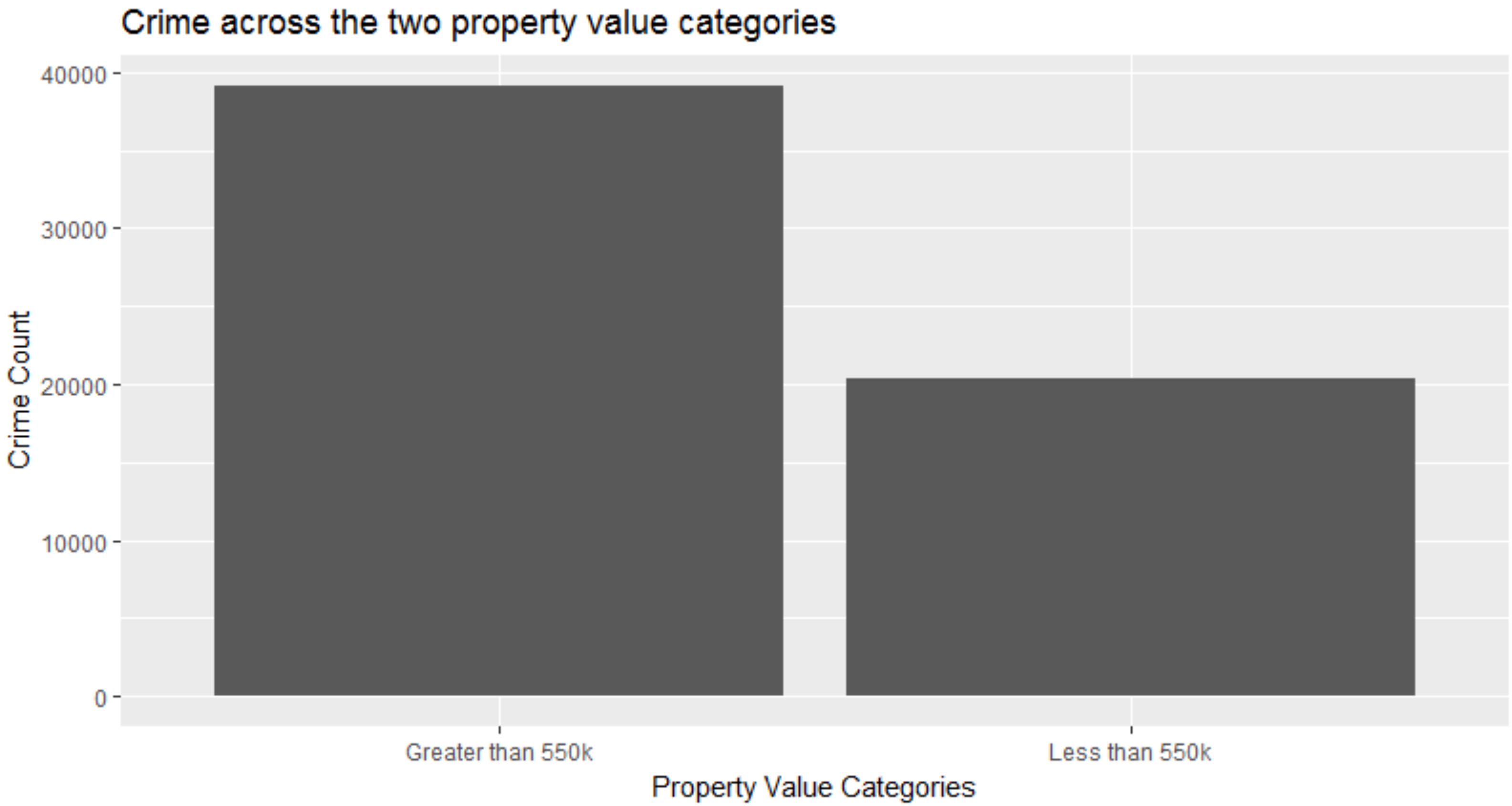
Vehicle Theft – 7058 incidences incidences



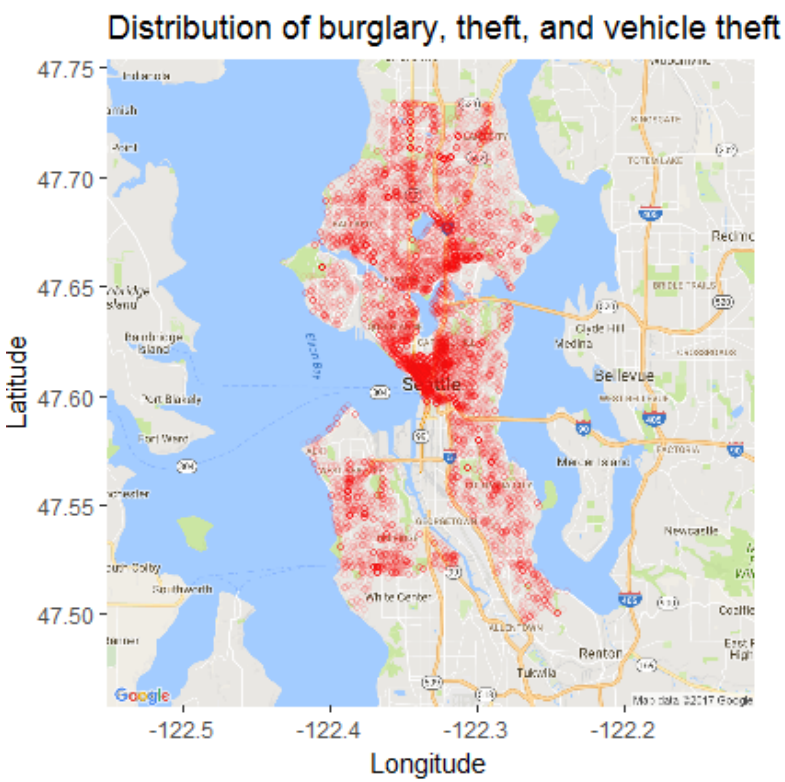
Zillow Data: Property values across 99 neighborhoods in Seattle

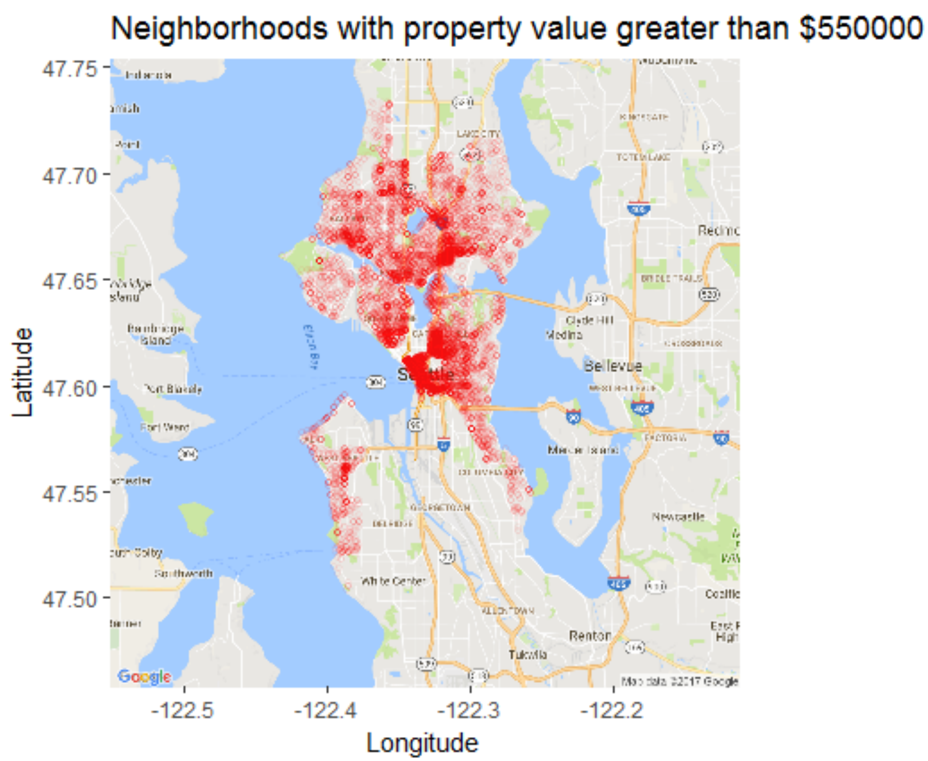
Property value categories

- 1. Neighborhoods with property value greater than or equal to $5,50,000
- 2. Neighborhoods with property value less than $5,50,000



Property vs Crime Comparison





Logistic Regression:

- Response : Top 3 crimes (burglary, theft, and vehicle theft)
- Predictor: Property values

```
glm(formula = Indicator ~ Current.Value, family = "binomial",
    data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8280  -1.3336   0.9475   1.0250   1.1450

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.826e-01  4.389e-02  -8.717   <2e-16 ***
Current.Value  1.330e-06  7.123e-08  18.665   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59906  on 44604  degrees of freedom
Residual deviance: 59543  on 44603  degrees of freedom
AIC: 59547

Number of Fisher Scoring iterations: 4
```

Model Accuracy?

- ▹Training data = 75% of total observations
- ▹Testing data = 25% of total observations
- ▹Cut-off = 60%

```
train <- sample(1:nrow(crime_zillow_merged_df), 3/4 * nrow(crime_zillow_merged_df))
test <- -train
training_data <- crime_zillow_merged_df[train,]
testing_data <- crime_zillow_merged_df[test,]
log.mod <- glm(formula = Indicator ~ Current.Value, family = "binomial", data = training_data)
summary(log.mod)
predicted_Prob <- predict(log.mod, testing_data, type = "response")
predictedIndicators = rep(0, nrow(testing_data))
predictedIndicators[predicted_Prob > 0.60] <- 1
table(predicted = predictedIndicators, actual = testing_data$Indicator)
mean(predictedIndicators==testing_data$Indicator)
```

```
         actual
predicted    0    1
        0 3077 4284
        1 2727 4781
```

- ▹Accuracy = 52.84%
- ▹Is there any better model?

Random Forests

- Training data = 75% of total observations
- Testing data = 25% of total observations
- Cut-off = 60%

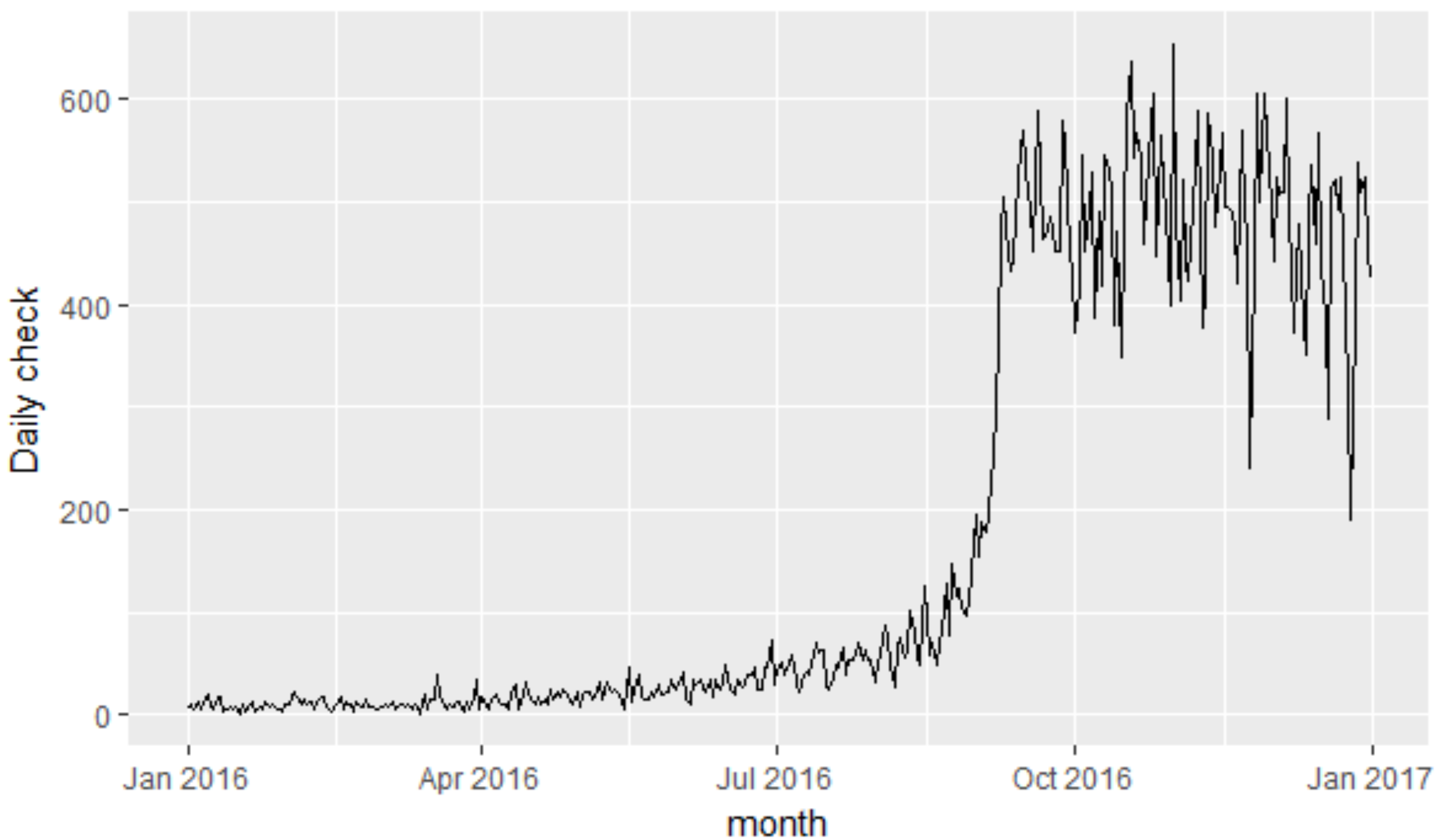
```
rf_model <- randomForest(Indicator~Current.Value, ntree=20, training_data)
predicted_Prob <- predict(rf_model, testing_data, type = "response")
predictedIndicators = rep(0, nrow(testing_data))
predictedIndicators[predicted_Prob > 0.60] <- 1
confusion_matrix <- table(predictions = predictedIndicators, actual = testing_data$Indicator)
mean(predictedIndicators==testing_data$Indicator)
```

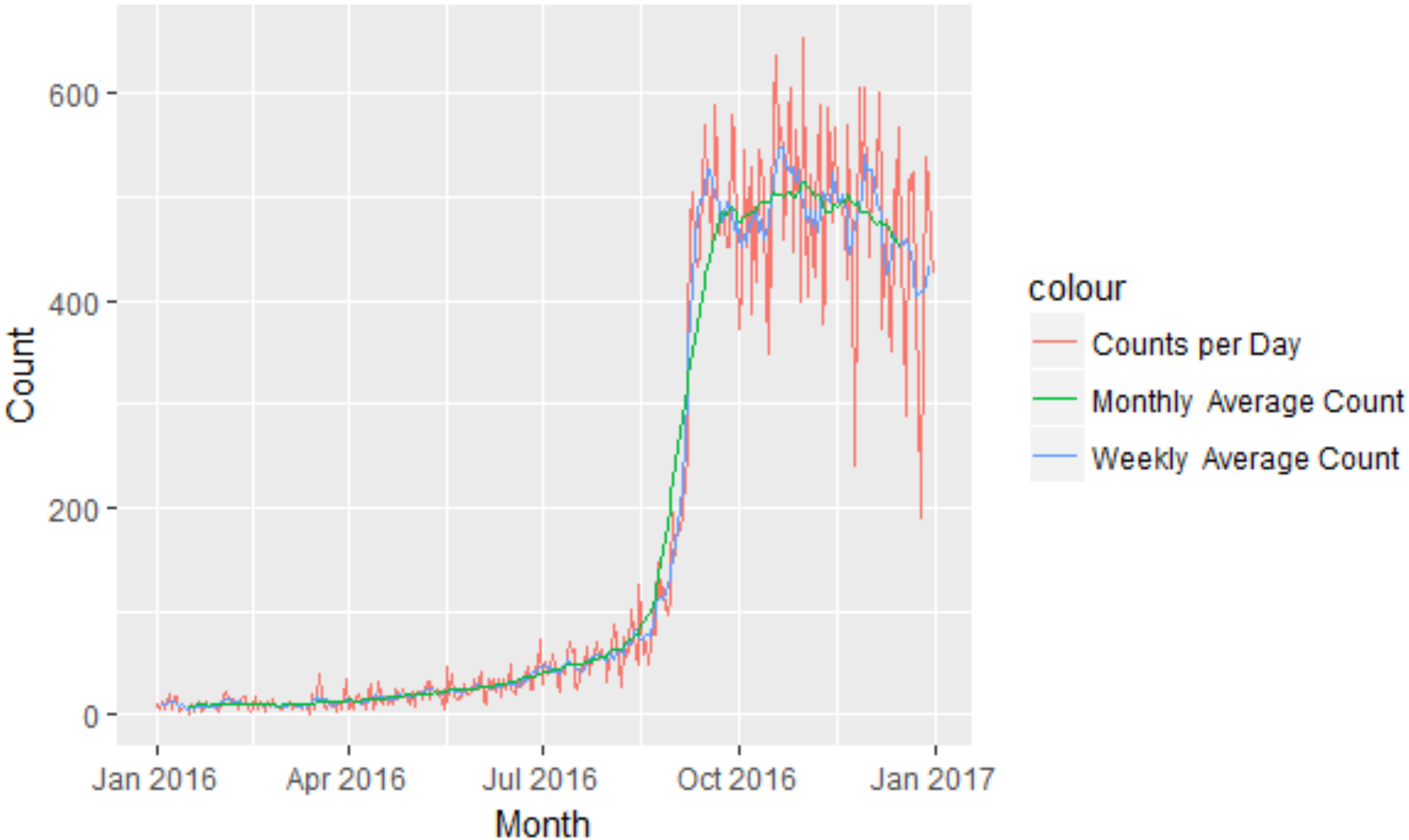
```
           actual
predictions    0    1
          0 3849 4057
          1 1955 5008
```

- Accuracy: 62.56%

Are there any seasonal variations in the distribution of Crime?

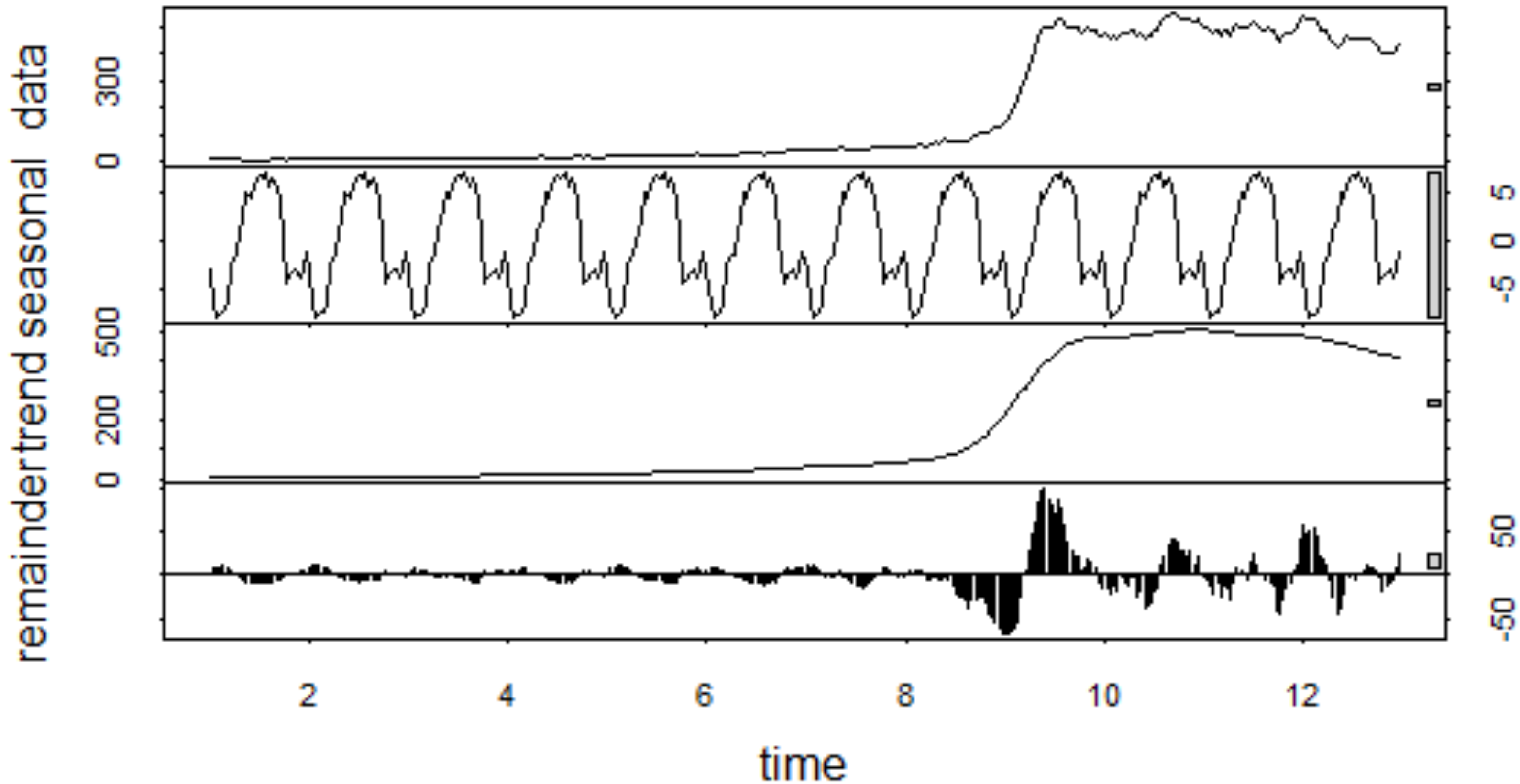
Time Series Analysis of Seattle Crime data of 2016



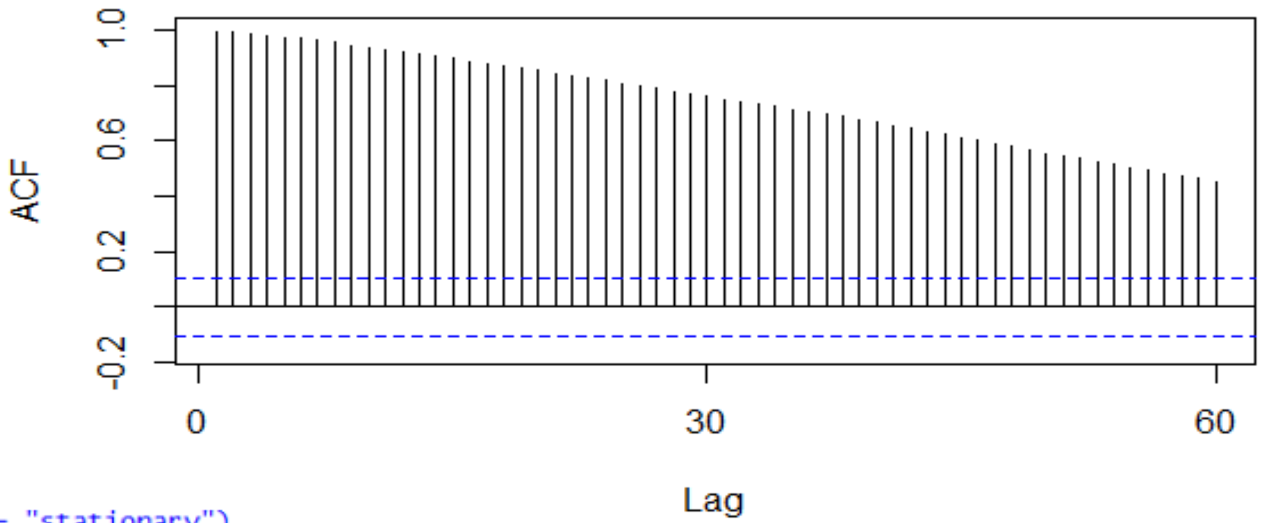


600
400
200
0
Count
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
Month
colour
Counts per Day
Monthly Average Count
Weekly Average Count

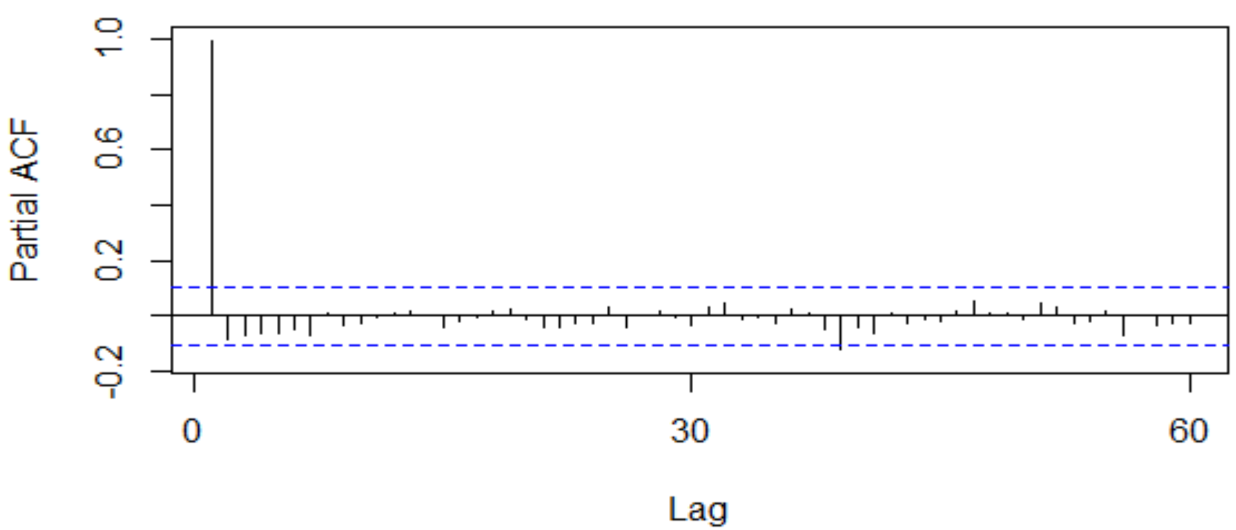
Decomposition of the Time Series



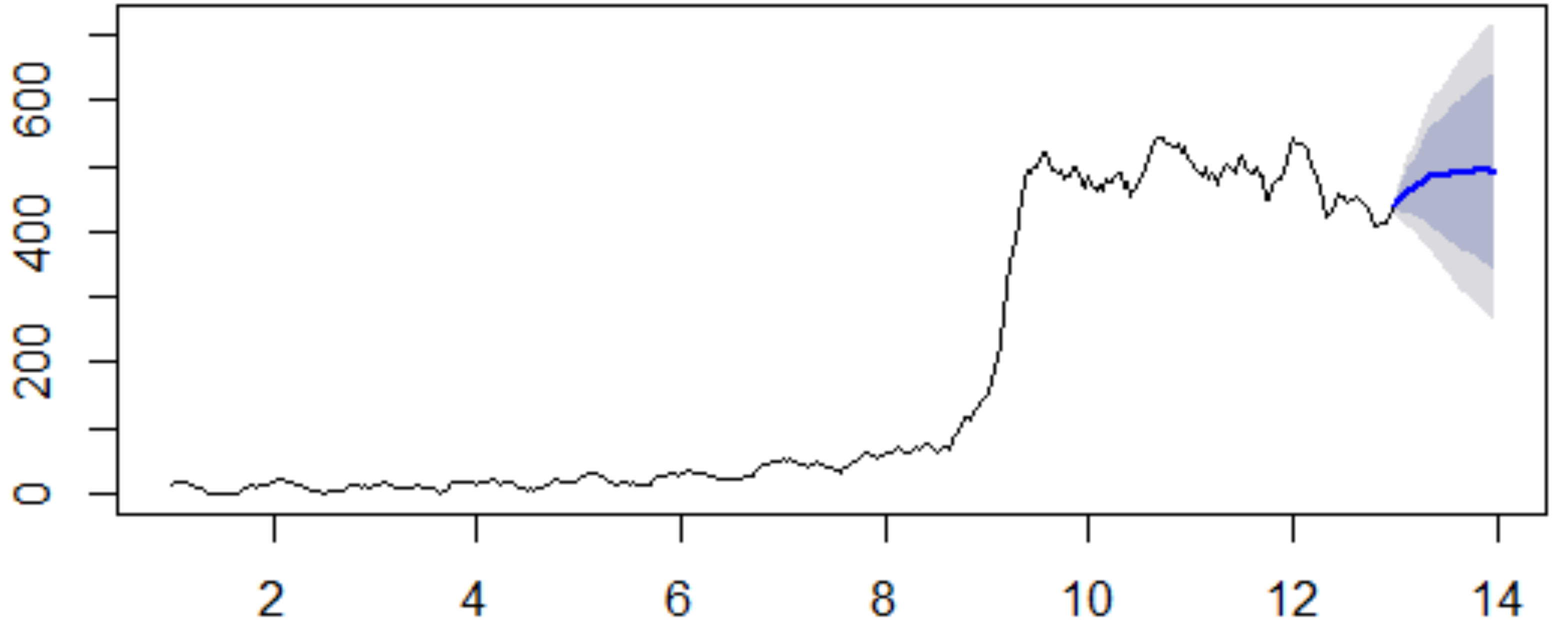
Determining the Stationarity of the Series



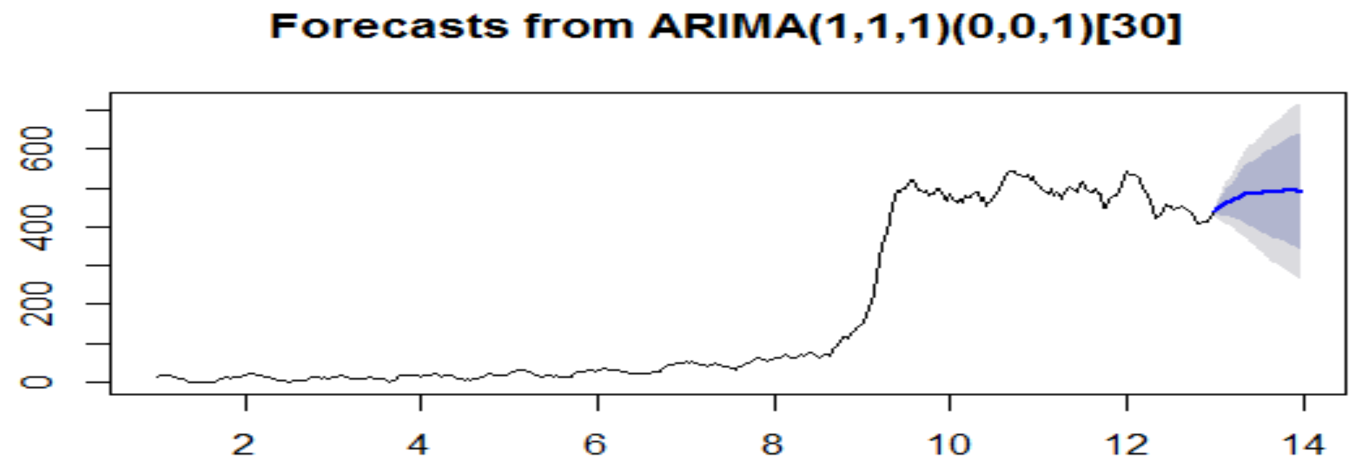

```
> adf.test(count_ma, alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  count_ma
Dickey-Fuller = -1.6664, Lag order = 7, p-value = 0.7177
alternative hypothesis: stationary
```

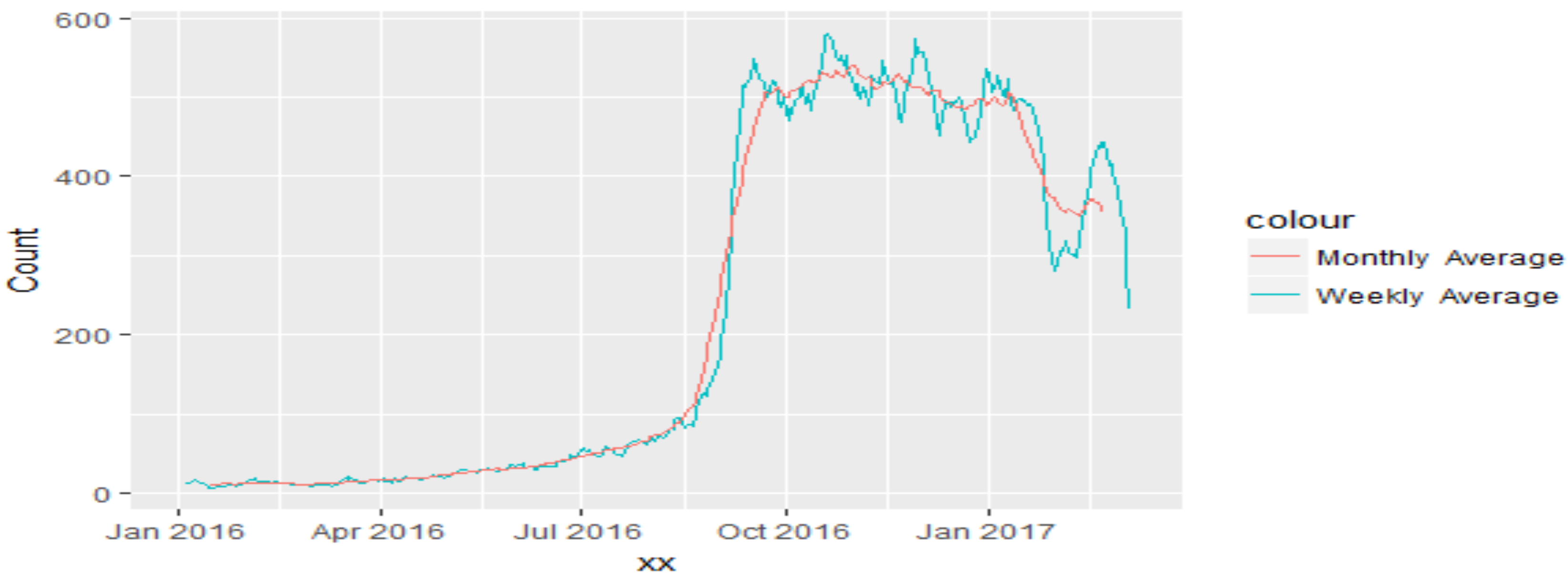


Forecasts from ARIMA(1,1,1)(0,0,1)[30]




```
> fit_w_seasonality
Series: deseasonal_cnt
ARIMA(1,1,1)(0,0,1)[30]

Coefficients:
         ar1      ma1     sma1
      0.8400  -0.4890  -0.1182
s.e.  0.0454   0.0712   0.0575

sigma^2 estimated as 56.1:  log likelihood=-1231.2
AIC=2470.41   AICc=2470.52   BIC=2485.94
> fit_w_seasonality
```


Forecasts from ARIMA(1,1,1)(0,0,1)[30]
0
200
400
600
2
4
6
8
10
12
14


600
400
200
0
Count
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
xx
colour
Monthly Average
Weekly Average

Conclusion

Significant findings:

- Although the presence of schools in an area is statistically significant, it might not be a strong predictor of crimes in that particular area

- Property value is statistically significant and is a significant predictor of crimes in that particular area

- Time series analysis shows that there is a trend of the crime occurring over the months.

Limitations

- Data Bias - We find significant amount of skewness in the data
- Unavailability of data in standard format
- Forecasting is based on one year data

Future Scope

Include other predictors that could help understand the crime scenario in Seattle. Such as :

- Demographic Data
- Census Data
- Weather Data

Collaborate with SPD

Thank you !