Boston Housing Analysis

Jamese Brown

03 August, 2024

```
summary(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

```
Boston %>% group_by(age == 100) %>% summarise(count = n())
```

```
## # A tibble: 2 × 2
##   `age == 100` count
##   <lgl>        <int>
## 1 FALSE          463
## 2 TRUE            43
```

```
Boston %>% group_by(age >= 90) %>% summarise(count = n())
```

```
## # A tibble: 2 × 2
##   `age >= 90` count
##   <lgl>       <int>
## 1 FALSE         336
## 2 TRUE          170
```

```
Boston %>% group_by(age <= 10) %>% summarise(count = n())
```

```
## # A tibble: 2 × 2
##   `age <= 10` count
##   <lgl>       <int>
## 1 FALSE         492
## 2 TRUE           14
```

```
Boston %>% group_by(crim <= 5) %>% summarise(count = n())
```

```
## # A tibble: 2 × 2
##   `crim <= 5` count
##   <lgl>       <int>
## 1 FALSE         106
## 2 TRUE          400
```

```
cor(Boston$medv,Boston$crim)
```

```
## [1] -0.3883046
```

Introduction

This data set contains 14 variables recorded from 506 Boston suburbs. This dataset was initially created to predict housing values in Boston Suburbs, but here it will be used to see what factors are the best predictors of crime. The remaining 13 variables will be analyzed to determine their ability to predict crime rates.

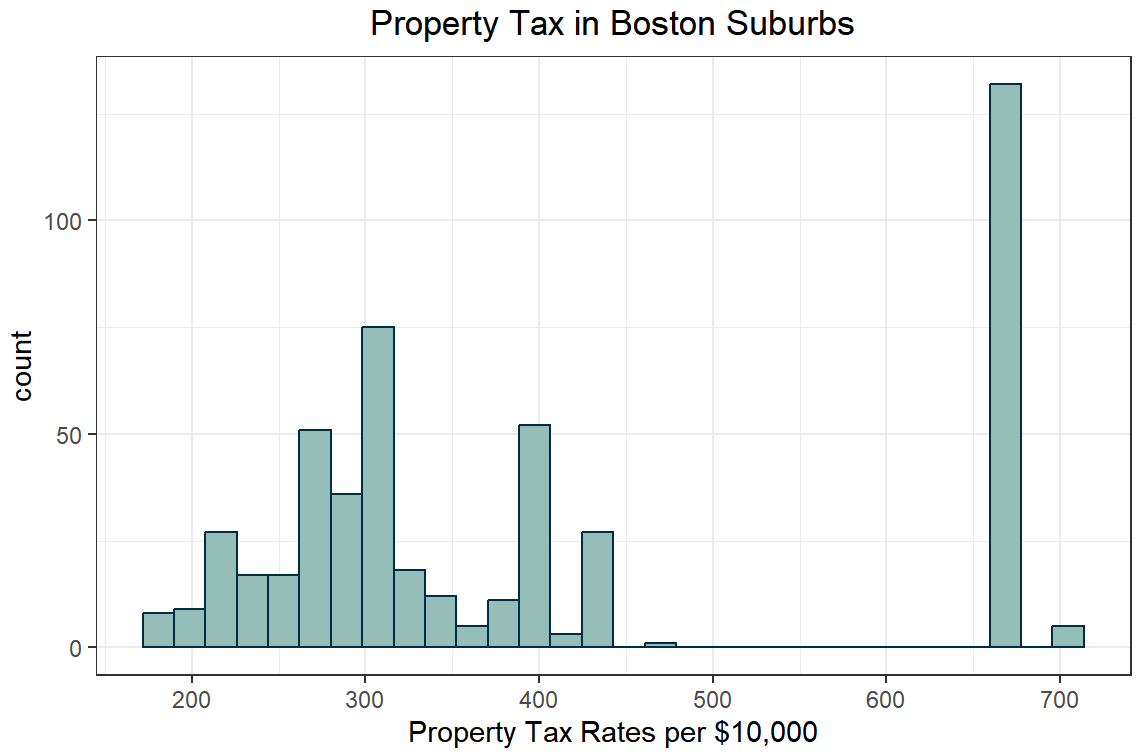
Data Highlights

- The highest per capita crime rate in a town is 88.97 while the lowest crime rate per capita is .006. The median crime rate is .25 while the mean is 3.61.
    - This difference between the mean and median is indicative of a right skew meaning that while the majority of Boston suburbs have low crime rates, there are a few with very high crime rates.
    - 400 of the recorded suburbs had a crime rate per capita of 5% or less.
    - There is negative correlation between crime rate and median house value (r = -.39). The crime rate outliers occur in suburbs with lower median house values and that are closer to an employment center.
- Many of the suburbs in Boston are older homes rather than new builds.
    - On average about 77% of the owner-occupied units in each suburb were built before 1940.
    - 43 of the 506 suburbs had 100% of the owner-occupied units built before 1940, and 107 suburbs had at least 90% of the homes built before 1940.
    - Only 14 of the recorded suburbs had 10% or fewer owner-occupied units built before 1940.

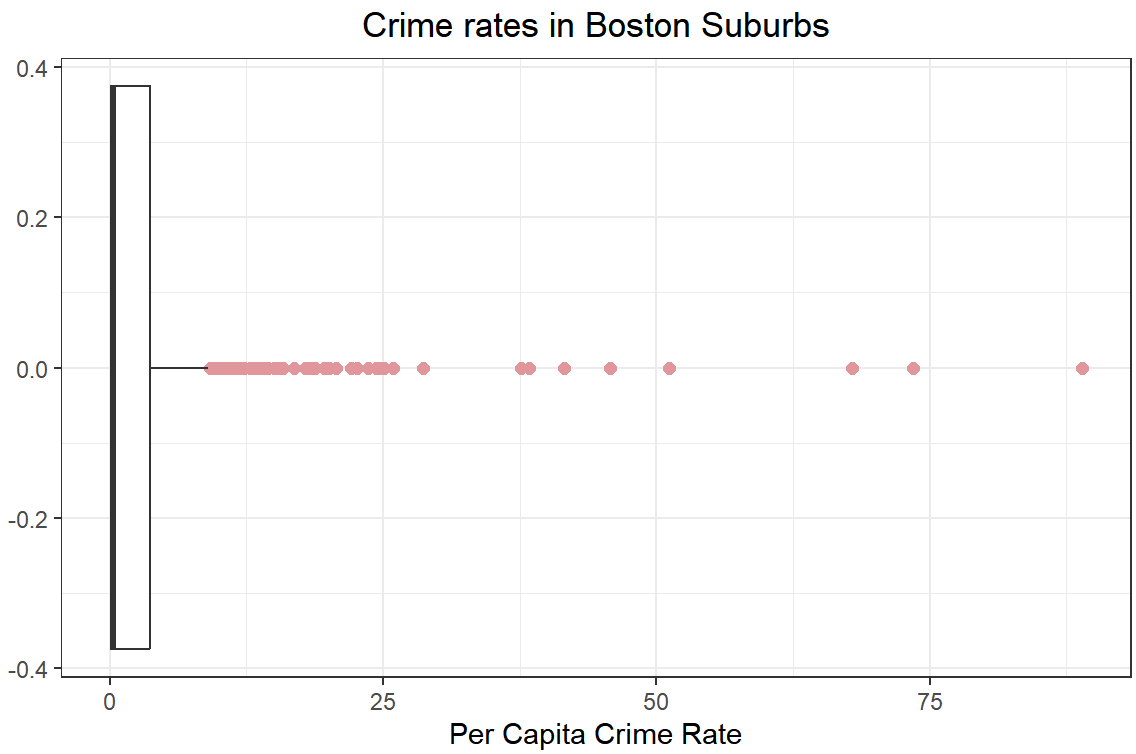
- Left skew on the black population in Boston suburbs which might be indicative of segregated neighborhoods as a few suburbs had a much lower proportion of black residents.
- Right skew on property tax rate meaning that higher property taxes were rarer for the suburbs in this dataset. While most of the suburbs had a property tax rate between $187/$10,000 and $450/$10,000, just over 125 neighborhoods had a property tax rate between $650/$10,000 and $711/$10,000.
- There is a positive linear correlation (r=.76) between nitrogen oxide concentration(nox) and the proportion of non-retail businesses in a neighborhood(indus). So, if a variable is correlated with nox, we will likely see a correlation between the variable and indus and vice versa.

Graphs

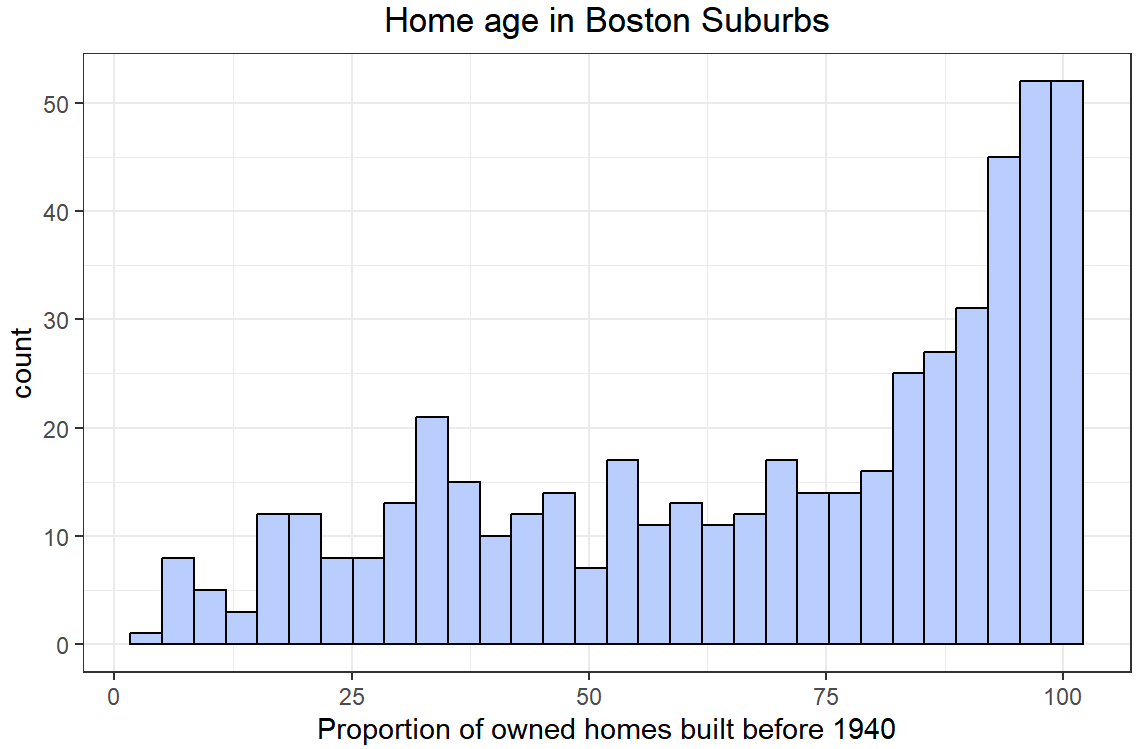
```
ggplot(Boston, aes(x = tax)) + geom_histogram(fill = '#99c1b9', col = I('#003049')) + labs(x = 'Property Tax Rates per $10,000', title = 'Property Tax in Boston Suburbs') + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



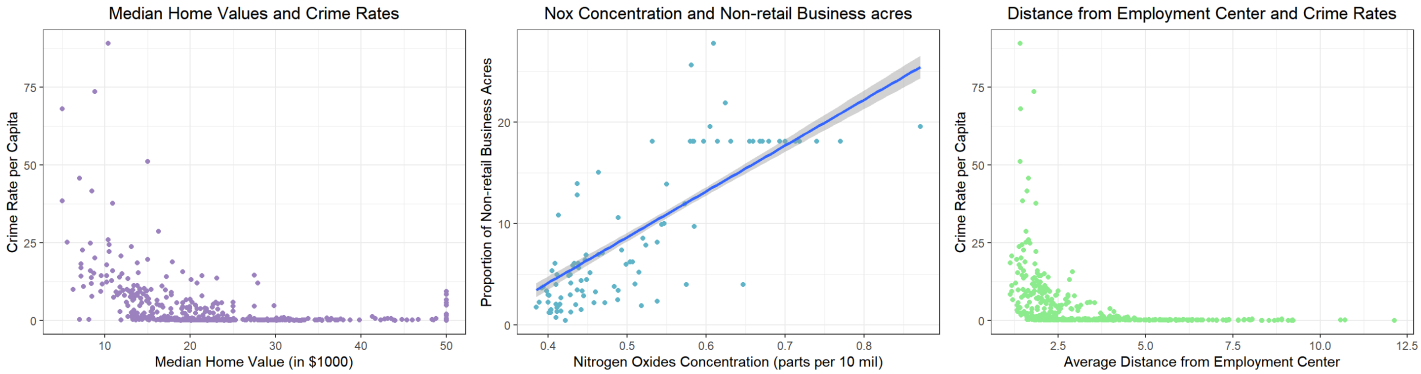

```
ggplot(Boston, aes(x = crim)) + geom_boxplot(outlier.color = '#E5989B', outlier.size = 2) + labs(x = 'Per Capita Crime Rate', title = 'Crime rates in Boston Suburbs') + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(Boston, aes(x = age)) + geom_histogram(fill = '#bbd0ff', col = I('black')) + labs(x = 'Pr
oportion of owned homes built before 1940', title = 'Home age in Boston Suburbs') + theme_bw() +
theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



```
g1 = ggplot(Boston, aes(x = medv, y = crim)) + geom_point(color = '#9f86c0') + labs(x = 'Median Home Value (in $1000)', y = 'Crime Rate per Capita', title = 'Median Home Values and Crime Rates') + theme_bw() + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
g2 = ggplot(Boston, aes(x = nox, y = indus)) + geom_point(color = '#62b6cb') + geom_smooth(method = lm) + labs(x = 'Nitrogen Oxides Concentration (parts per 10 mil)', y = 'Proportion of Non-retail Business Acres', title = 'Nox Concentration and Non-retail Business acres') + theme_bw() + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
g3 = ggplot(Boston, aes(x = dis, y = crim)) + geom_point(color = 'lightgreen')+ labs(x = 'Average Distance from Employment Center', y = 'Crime Rate per Capita', title = 'Distance from Employment Center and Crime Rates') + theme_bw() + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
ggarrange(g1,g2,g3, ncol = 3)
```



Simple Linear Regressions

```
zoned = lm(crim ~ zn, Boston) #p = 5.5e-06
summary(zoned)
```

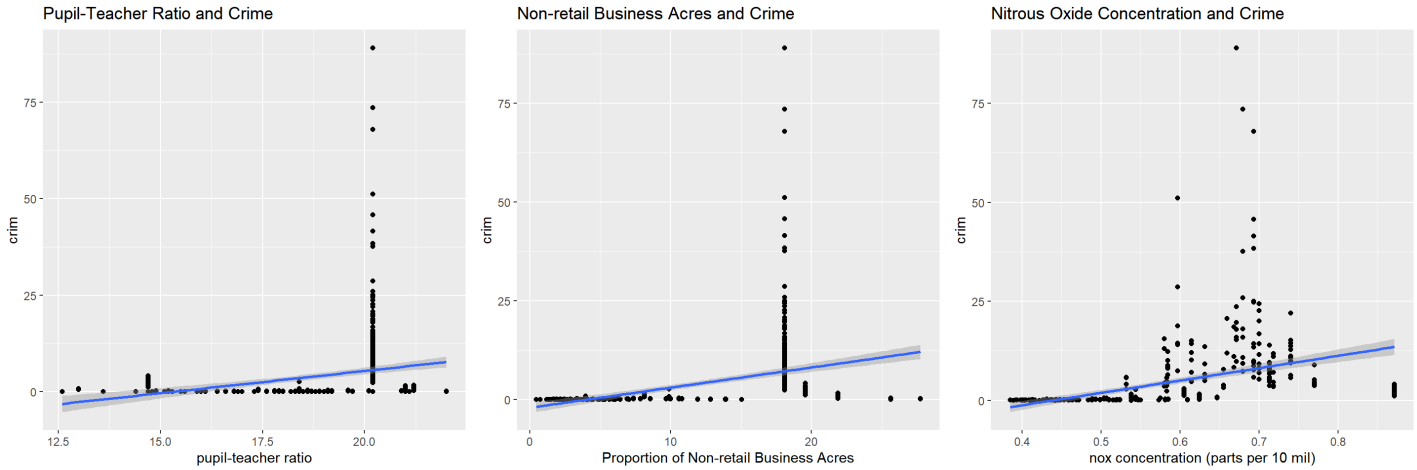
```
##
## Call:
## lm(formula = crim ~ zn, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

```
#repeated for remaining variables
```

There is a statistically significant association between the predictor and response(crime) for every feature except chas, the dummy variable providing a binary response to if the tract bounds the Charles river of not.

Supporting Plots

```
b = ggplot(Boston, aes(y = crim, x = ptratio)) + geom_point() + geom_smooth(method = lm) + labs
(x = 'pupil-teacher ratio')+ ggtitle('Pupil-Teacher Ratio and Crime')
c = ggplot(Boston, aes(y = crim, x = indus)) + geom_point() + geom_smooth(method = lm) + labs(x
= 'Proportion of Non-retail Business Acres')+ ggtitle('Non-retail Business Acres and Crime')
d = ggplot(Boston, aes(y = crim, x = nox)) + geom_point() + geom_smooth(method = lm) + labs(x =
'nox concentration (parts per 10 mil)')+ ggtitle('Nitrous Oxide Concentration and Crime')
ggarrange(b,c,d, ncol = 3)
```



Multiple linear Regression

```
mult_lin = lm(crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio +
                black + lstat + medv, Boston)
summary(mult_lin)
```

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

We would reject the null for zn, nox, dis, rad, black, lstat, and medv. In other words, we have evidence that the proportion of residential land, nox concentration, distance to employment centers, accessibility to radial highways, the proportion of black residents, the percentage of ‘lower status’ individuals, and median home value have significant impacts on crime rates in Boston suburbs.