ANLY 699: Assignment 1

Code

Subhash Pemmaraju

05/31/2020

Key Variables

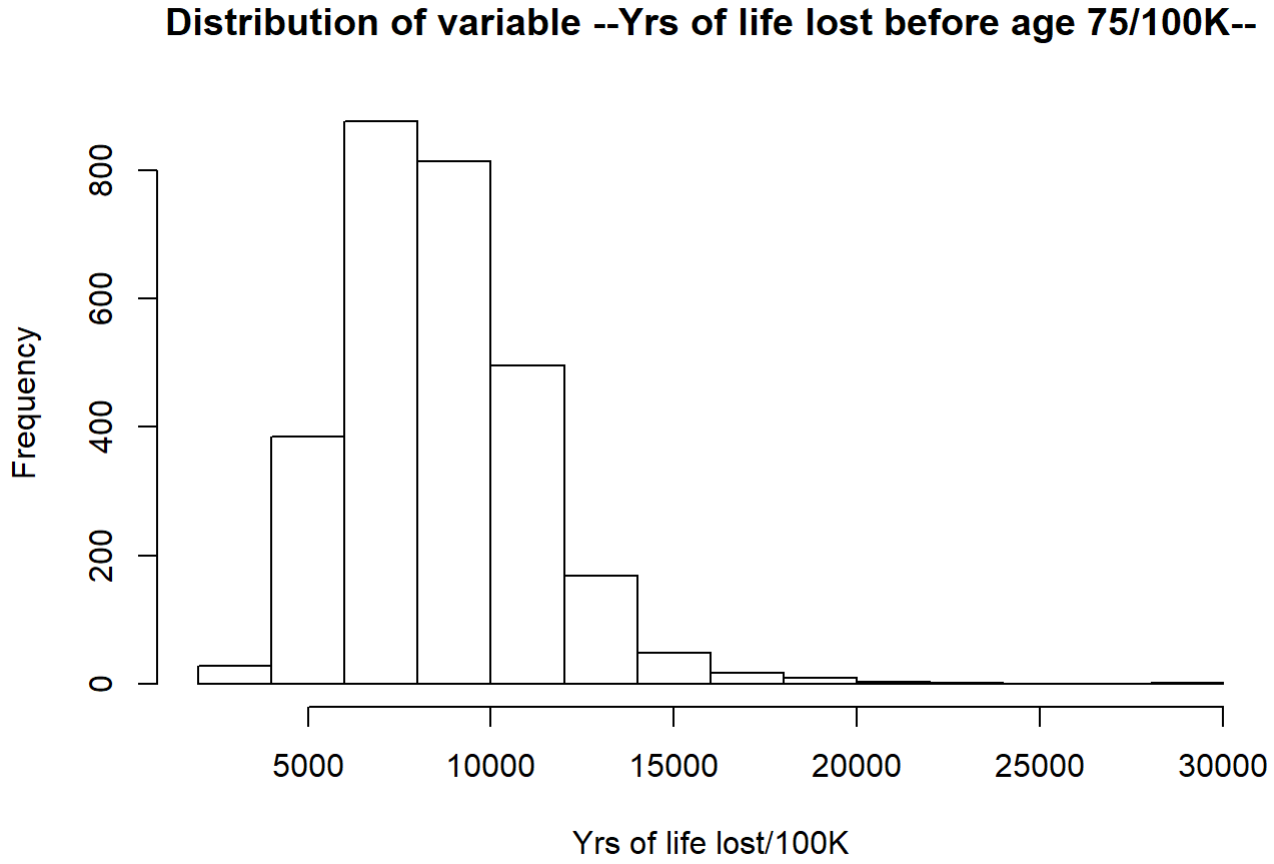
The key output variable being considered for this analysis is: “Years of life lost before the age of 75 per 100,000 population”

The key input variables are

- % of the population living within half a mile of a park
- % of population that are smokers

Hide

```
hist(health_ind_cty$yrs_plr, main="Distribution of variable --Yrs of life lost before age 75/100K--", xlab = "Yrs of life lost/100K")
```



Hide

```
summary(health_ind_cty$yrs_plr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2731    6793    8310    8583   10077   29138     293
```

Summary

From the histogram, we can see that the data is not normally distributed. It is positively skewed with a long tail. We can confirm this fact by looking at the skewness and kurtosis.

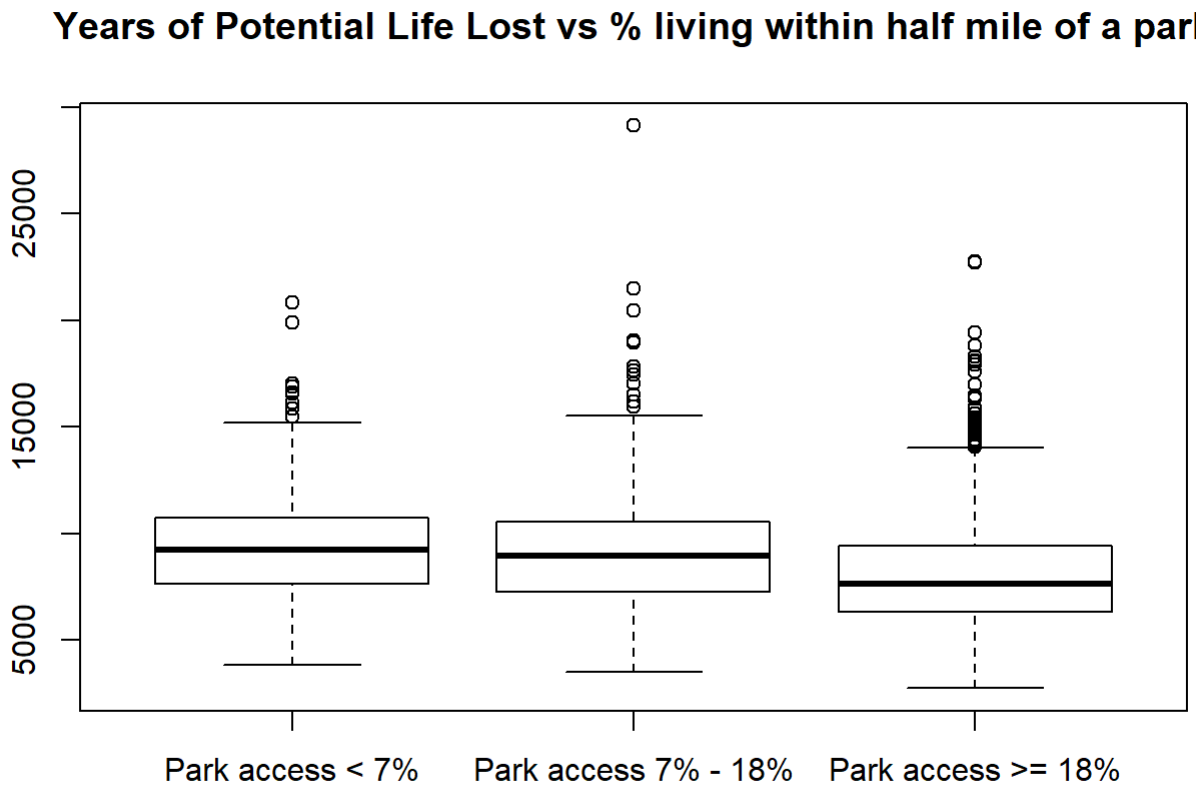
The skewness of the data is 1.1083303

The kurtosis of the data is 6.9618585

The high kurtosis of the data (>3) points to fat tails and the skew indicates a positive skew.

Hide

```
boxplot(merged_data$yrs_plr~merged_data$park_access_seg, main="Years of Potential Life Lost vs % living within half mile of a park")
```

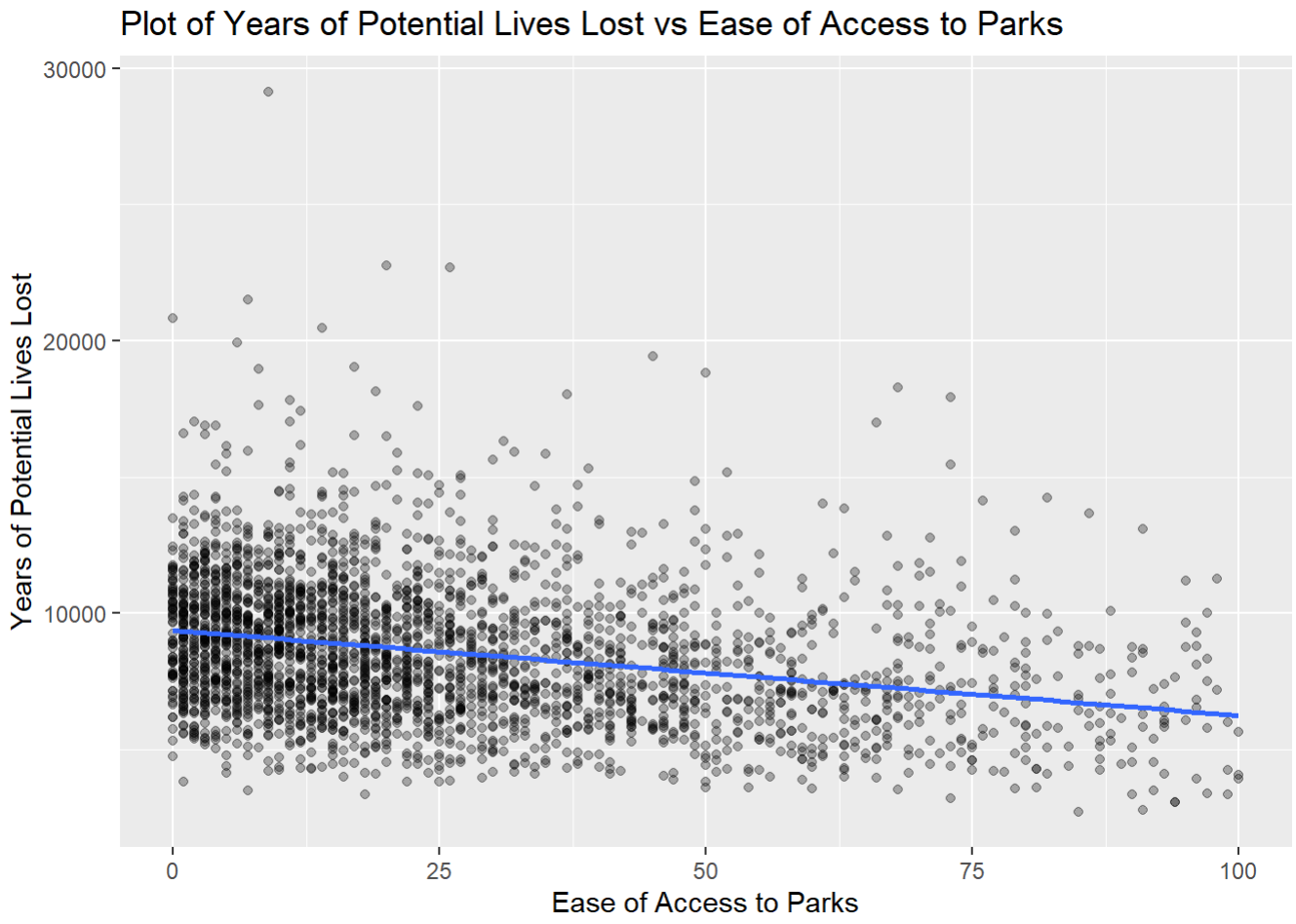


Summary

From the boxplot, we can see that there are a few outliers in the data. Furthermore, we can see that the median moves lower as % of people with access to parks goes up. This is a good sign that there is a potential directional relationship between these two variables indicating that more access to parks leads to fewer years of potential lives lost.

Hide

```
ggplot(data=merged_data, aes(x=Value,y=yrs_plr))+geom_point(alpha=0.3)+
  ggtitle("Plot of Years of Potential Lives Lost vs Ease of Access to Parks")+xlab("Ease of Acce
ss to Parks")+ylab("Years of Potential Lives Lost")+
  geom_smooth(method = 'lm', se=F)
```

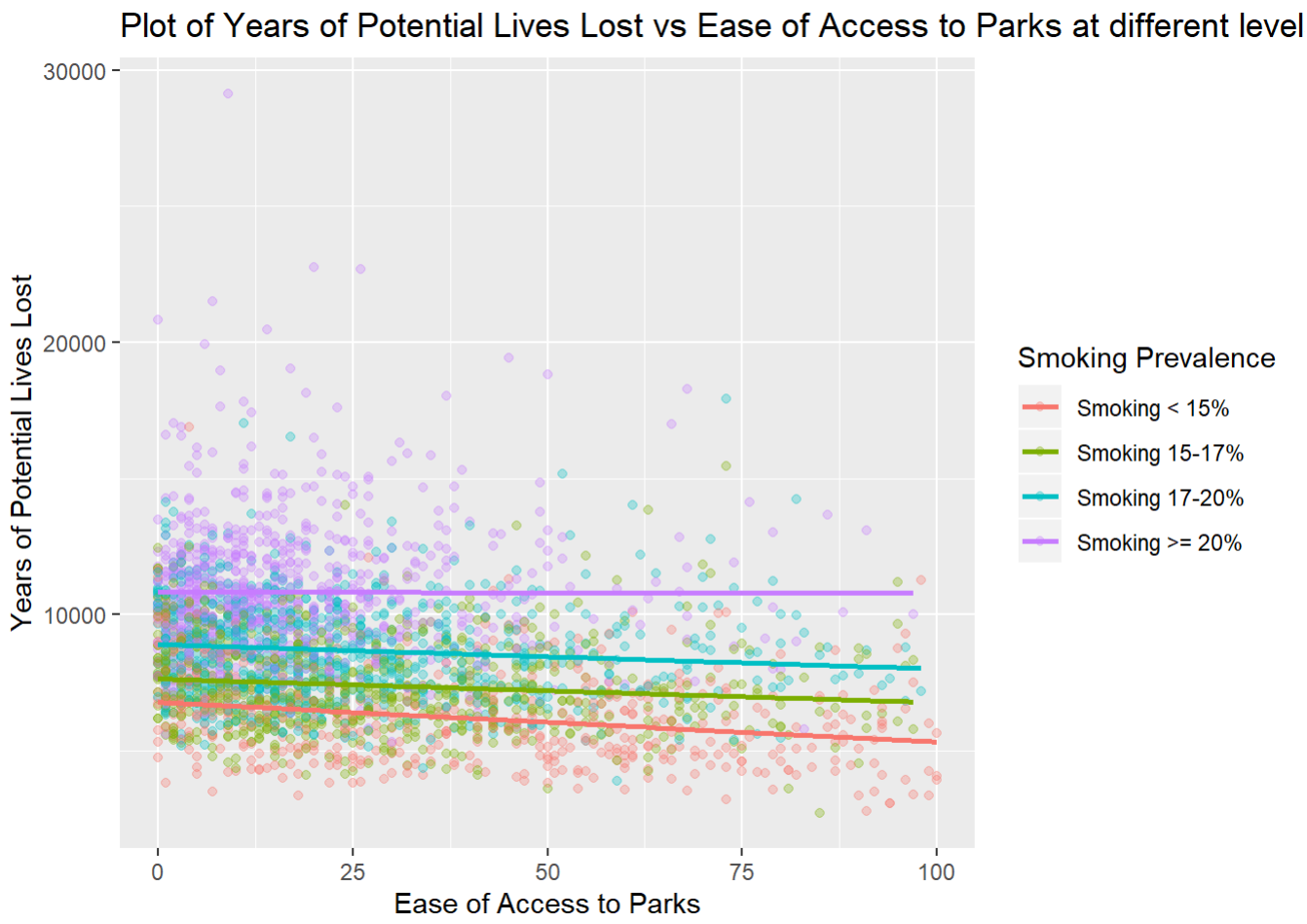


Summary

From the scatterplot and the linearly fitted line, we can see a negative correlation, a confirmation of the initial finding from the boxplot. The scatterplot also shows several outliers.

Hide

```
ggplot(data=merged_data, aes(x=Value,y=yrs_plr,color=smoking_seg))+geom_point(alpha=0.3)+
  ggtitle("Plot of Years of Potential Lives Lost vs Ease of Access to Parks at different levels
 of smoking")+xlab("Ease of Access to Parks")+ylab("Years of Potential Lives Lost")+
  scale_colour_discrete(na.translate = F)+labs(color="Smoking Prevalence")+
  geom_smooth(method = 'lm', se=F)
```



Summary

From the scatterplot and corresponding fitted lines, we see an interesting trend. The negative correlation between ease of access to parks and years of potential lives lost holds at different levels of smoking prevalance but at high levels of smoking prevalence, the slope is flatter. This indicates that the benefits of having access to public parks wears down as unhealthy behaviours such as smoking increase. The levels are also different with more smoking line at a higher level (more years of potential lives lost).