Exploratory Data Analysis by Kira Lapides

Preparation

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
ncbirths <- openintro::ncbirths
```

Introduction

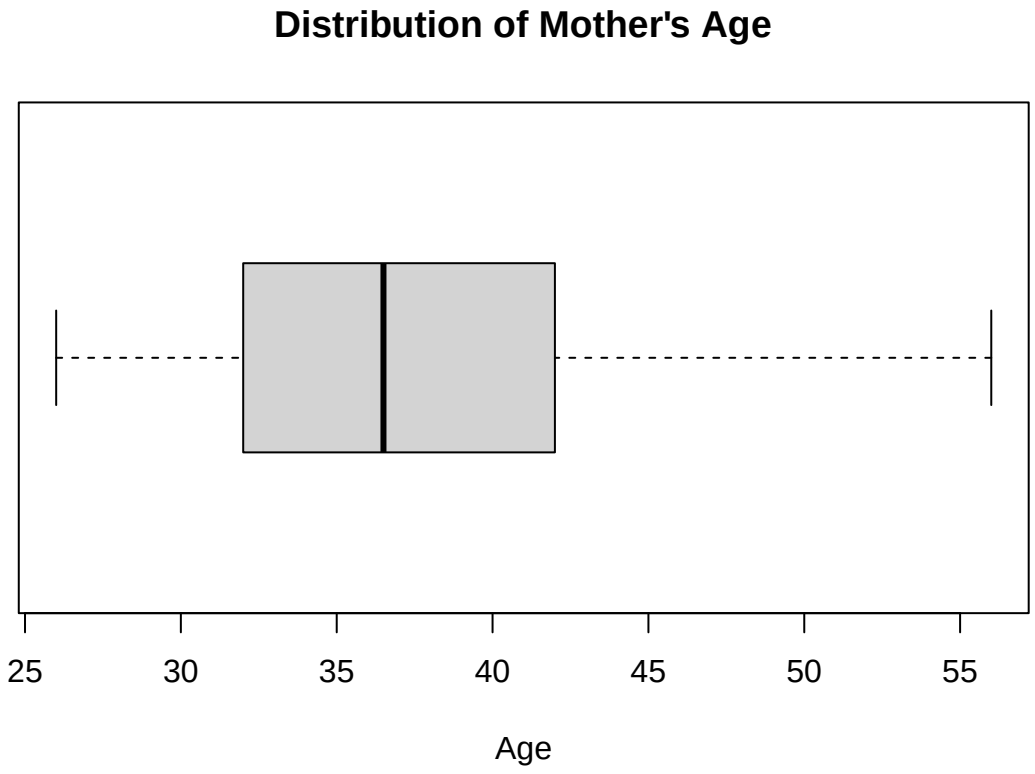
The data set I will be analyzing is the Lung Function data set, containing 150 observations and 32 variables within the Los Angeles area. This data set is exploring the effects of smog on lung function. I will be exploring the variables weight and height of females and their effect on the forced vital capacity. The forced vital capacity (FVC) is the total amount of air exhaled when measuring the patient's volume of air exhaled during a forced breath. I have chosen these variables because I would like to determine if there is a correlation or patterns between weight and location and if it has an effect on their total lung capacity and the strength of their lungs. My key question is are there prominent differences between lung capacity of different females based on their weight and height? Is there are correlation between these variables and their forced vital capacities?

```
fev <- read.delim("Lung_081217.txt", header=TRUE,sep="\t")
dim(fev)
```

```
## [1] 150  32
```

Univariate Analysis

```
boxplot(fev$MAGE, horizontal = TRUE, main="Distribution of Mother's Age", xlab="Age")
```

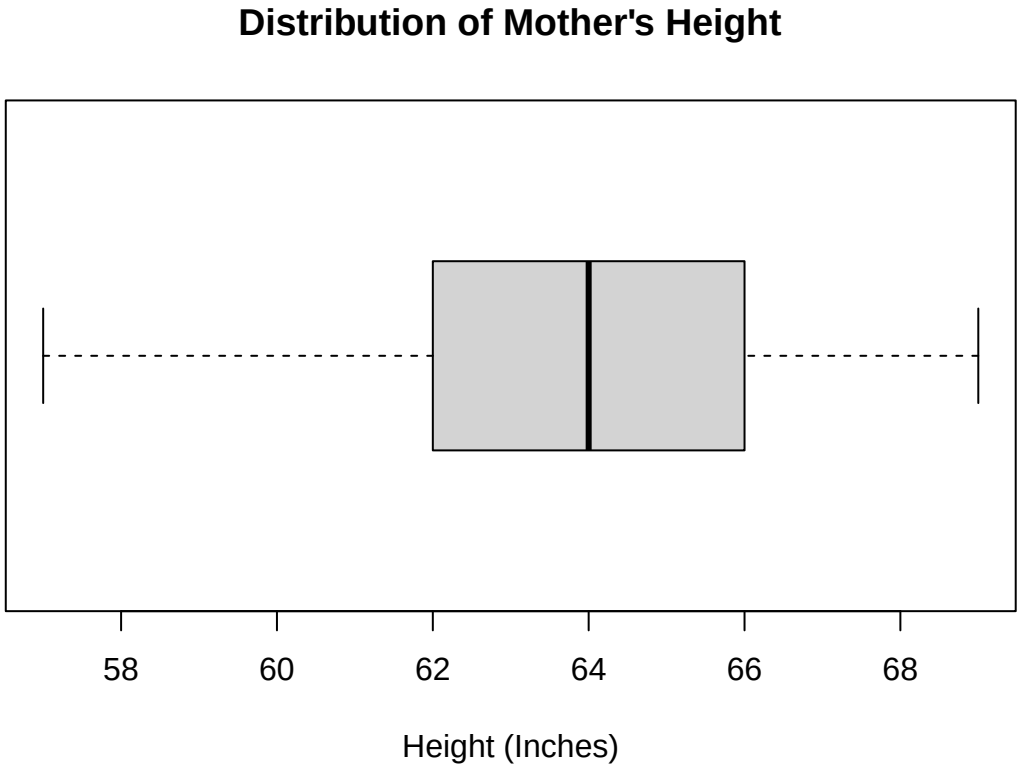



```
summary(fev$MAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   32.00   36.50   37.56   42.00   56.00
```

Within the data of mother's, there is a range of 26-56 years with a 36 year median. There is an average of 37 years.

```
boxplot(fev$MHEIGHT, horizontal = TRUE, main="Distribution of Mother's Height", xlab="Height (Inches)")
```

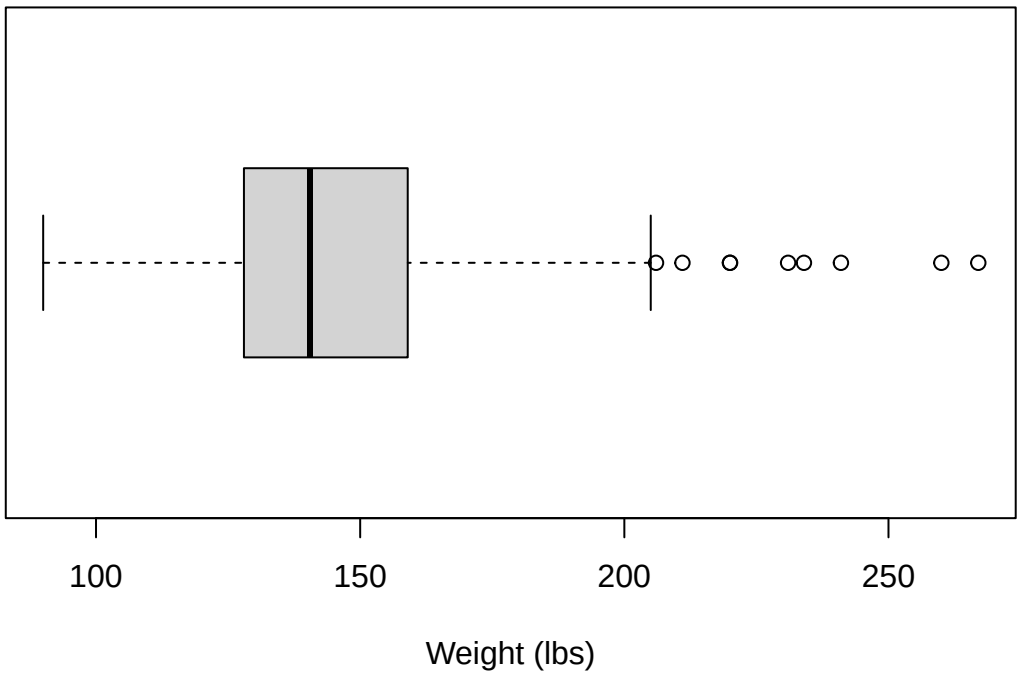



```
summary(fev$MHEIGHT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   62.00   64.00   64.09   66.00   69.00
```

Within the data of mother's, there is a range of 57-69 inches (4'9ft-5'9ft) with a 64 inch (5'4 ft) median. There is an average of 64 inches(5'4ft).

```
boxplot(fev$MWEIGHT, horizontal = TRUE, main="Distribution of Mother's Weight", xlab="Weight (lbs)")
```

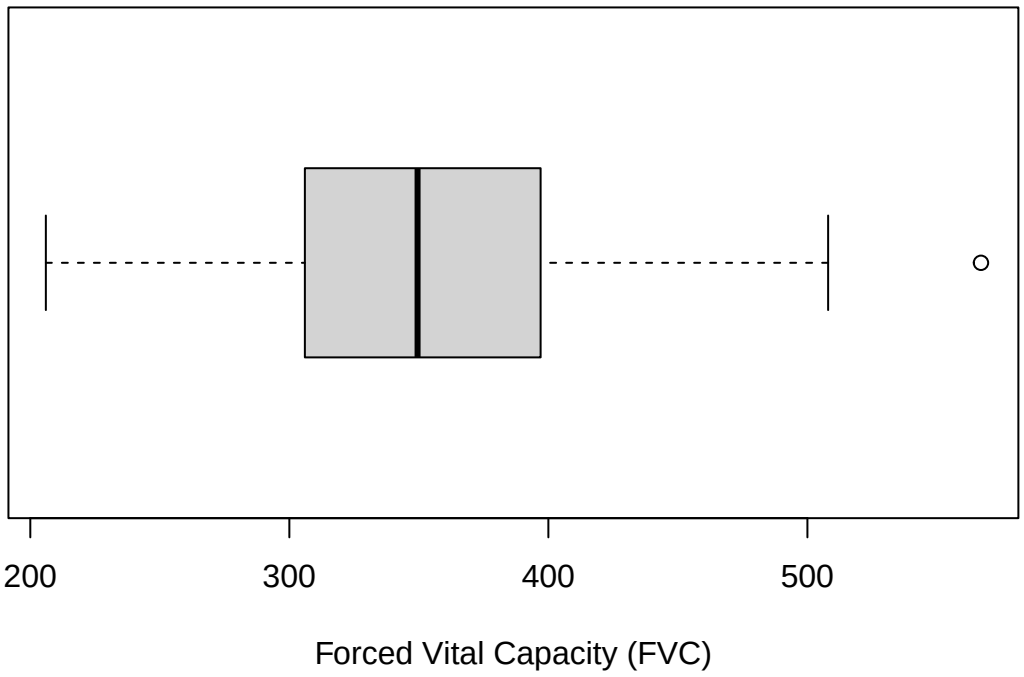



```
summary(fev$MWEIGHT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    90.0   128.0   140.5   147.0   159.0   267.0
```

Within the data of mother's, there is a range of 90-267lbs with a 140.5lbs median. There is an average of 147 lbs.

```
boxplot(fev$MFVC, horizontal = TRUE, main="Distribution of Mother's Forced Vital Capacity (FVC)", xlab=
```




```
summary(fev$MFVC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   206.0   306.0   349.5   350.2   396.2   567.0
```

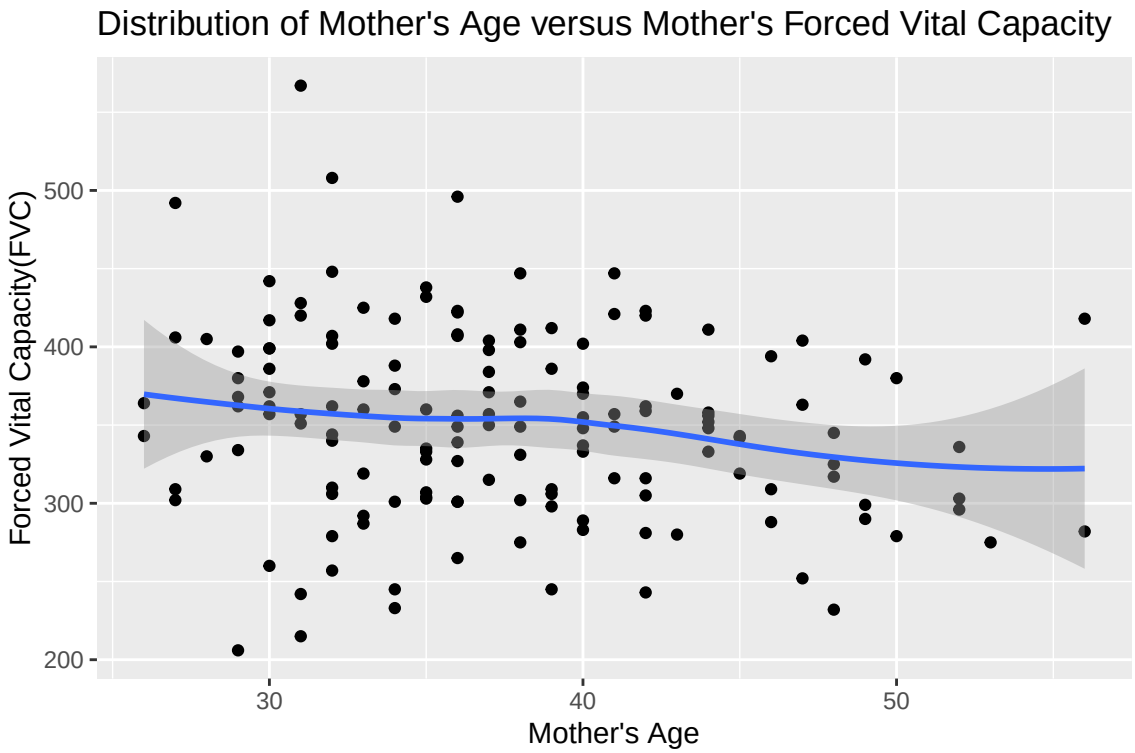
On average the FVC was 350.2. It ranged from 206-567.

Bivariate Analysis

Age vs FVC

```
ggplot(fev, aes(x=MAGE, y=MFVC)) + geom_point() + geom_smooth() + ggtitle("Distribution of Mother's Age
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

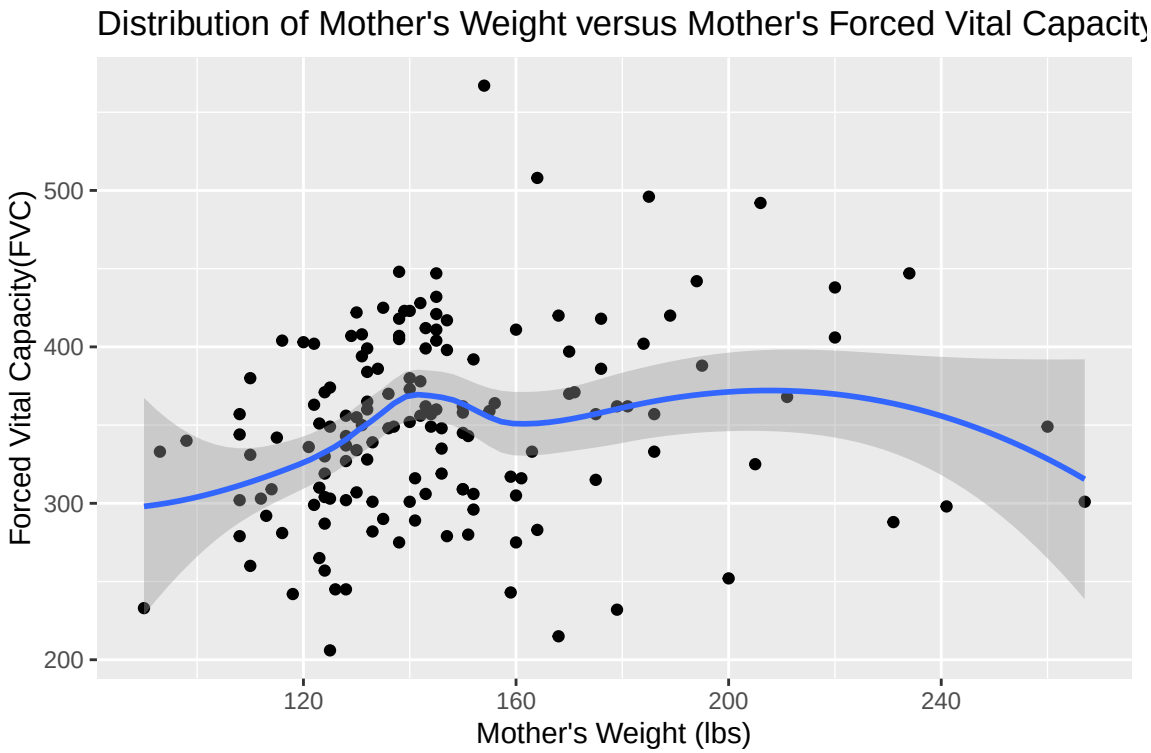


In this scatterplot, I compared the mother's age with their FVC measurements. I had hoped to see a correlation between these two variables. I had hypothesized that the more mature the women were (the older) the more their bodies would optimize the oxygen intake. However there are also many other confounding variables within the environment that would affect this.

Weight vs FVC

```
ggplot(fev, aes(x=MWEIGHT, y=MFVC)) + geom_point() + geom_smooth() + ggtitle("Distribution of Mother's
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

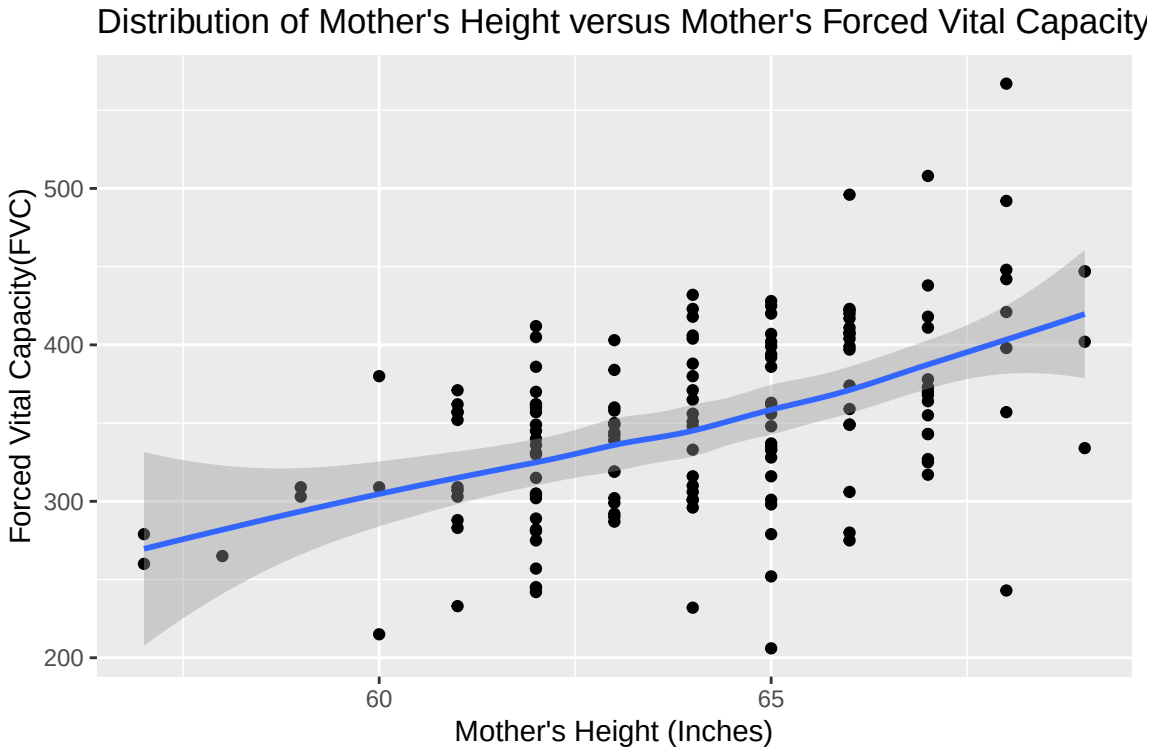


In this scatterplot, I compared Mother's weight with their FVC measuremnts. I had hypothesized that at higher weights the body would need more oxygen within their body's chemistry translating to higher FVC measurements. Based on this scatterplot there is little correlation. Within the weight range of 120-160lbs there is varying FVC measurements, there is no sure pattern.

Height vs FVC

```
ggplot(fev, aes(x=MHEIGHT, y=MFVC)) + geom_point() + geom_smooth() + ggtitle("Distribution of Mother's
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In this scatterplot, Mother's height was compared to their FVC Measurements. I based my hypothesis on similar reasoning as the Weight vs. FVC; Larger heights require more oxygen to support the body systems. Being physically bigger can also translated to having bigger lungs. This was the only scatterplot that seemed to have any correlation, so it supported my hypotheis.

Conclusion

At the beginning of my data analysis of the lung function data set, I was curious to determine whether their were correlations with the female's lung function and their ages, height and weight. I had hypothesized that these would have positive correlation because it made sense to me that at older ages, larger weights and taller people would have more mature lungs or simply require more oxygen to support their body's systems. Overall, there was only correlation found within the Height vs FVC scatterplot. This made sense that at taller heights/being physically bigger would call for physically bigger lungs that could intake more air. I had hoped to find correlation to better understand my own lung function, however unfortunately given the data it was inapplicable to myself. With the variables of age and weight there are a lot of other factors that influence lung function, so data was quite scrabbled for theese scatterplots.