EDA on Odom data

2023-08-30

```
odom_data<-read_excel("C:/Users/beebe/OneDrive/Documents/MS DS IUB/EDA/Assignment 1 data.xls")
```

Calculating difference between self reported and baseline miles

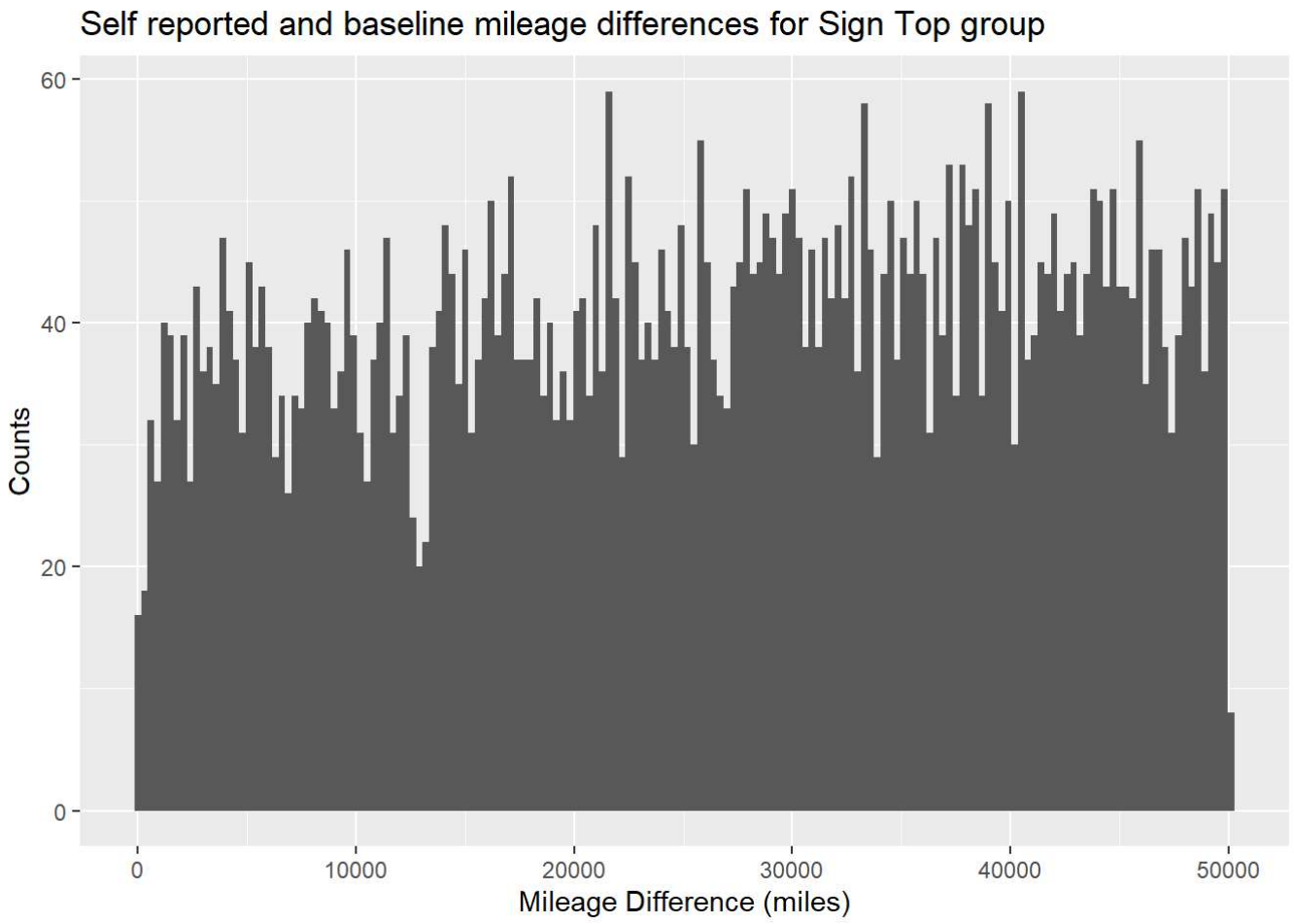
```
group <- odom_data$`OMR Version`
self_reported_miles<- odom_data$`Odom Reading 1 (Update)`
baseline_miles<- odom_data$`Odom Reading 1 (Previous)`
difference<- self_reported_miles- baseline_miles
```

Splitting into two groups based on where they signed(top/bottom).

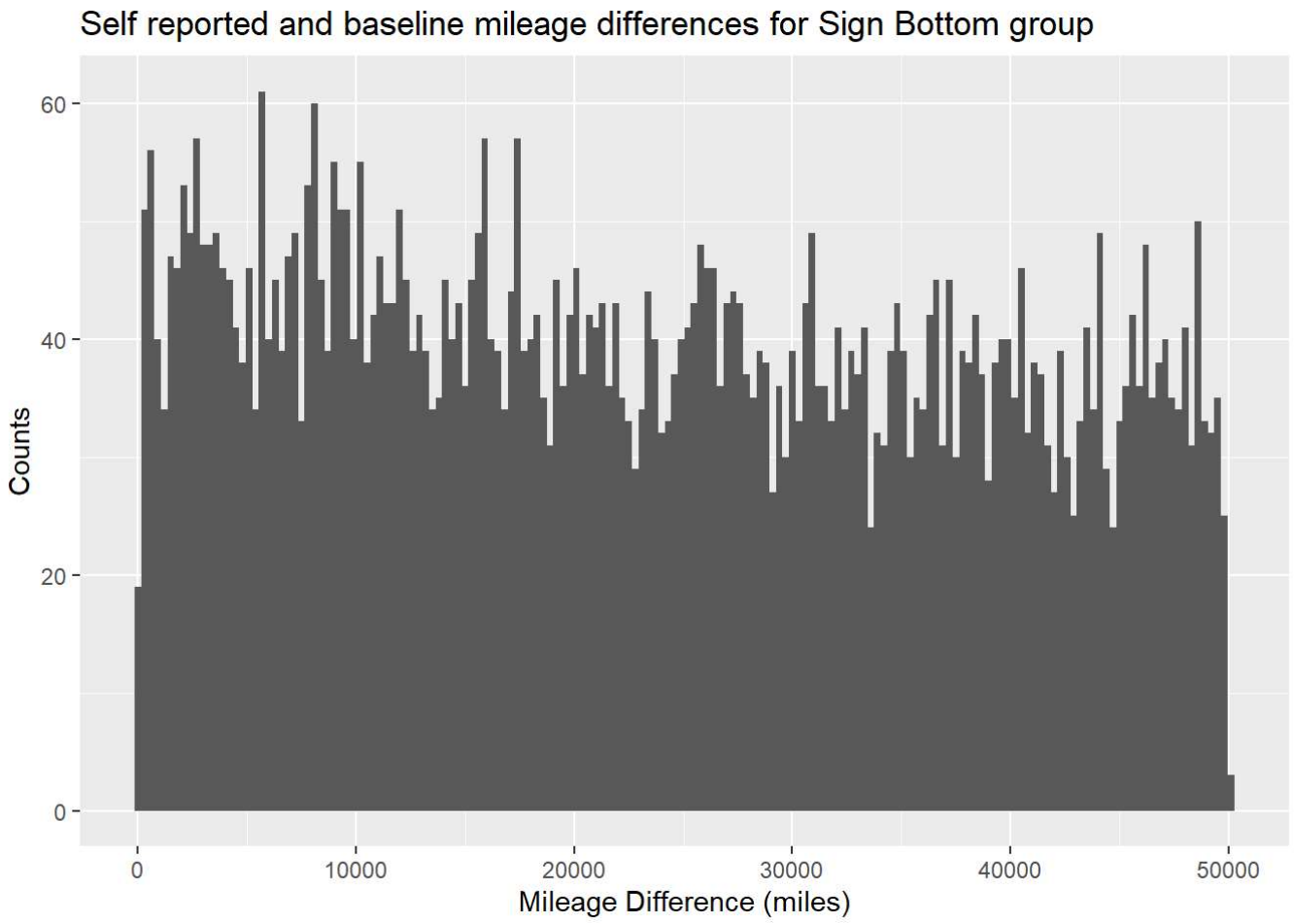
```
sign_top<- difference[group=="Sign Top"]
sign_bottom<- difference[group=="Sign Bottom"]
```

#Histograms for mileage difference

```
ggplot(data = data.frame(difference = sign_top), aes(x = difference)) +
  geom_histogram(binwidth = 300)+
  labs(title = "Self reported and baseline mileage differences for Sign Top group",
       x = "Mileage Difference (miles)",
       y = "Counts")
```




```
ggplot(data = data.frame(difference = sign_bottom), aes(x = difference)) +
  geom_histogram(binwidth = 300)+
  labs(title = "Self reported and baseline mileage differences for Sign Bottom group",
       x = "Mileage Difference (miles)",
       y = "Counts")
```



Calculating mean of the mileage differences of two groups.

```
mean_top<- mean(sign_top)
print(mean_top)
```

```
## [1] 26204.83
```

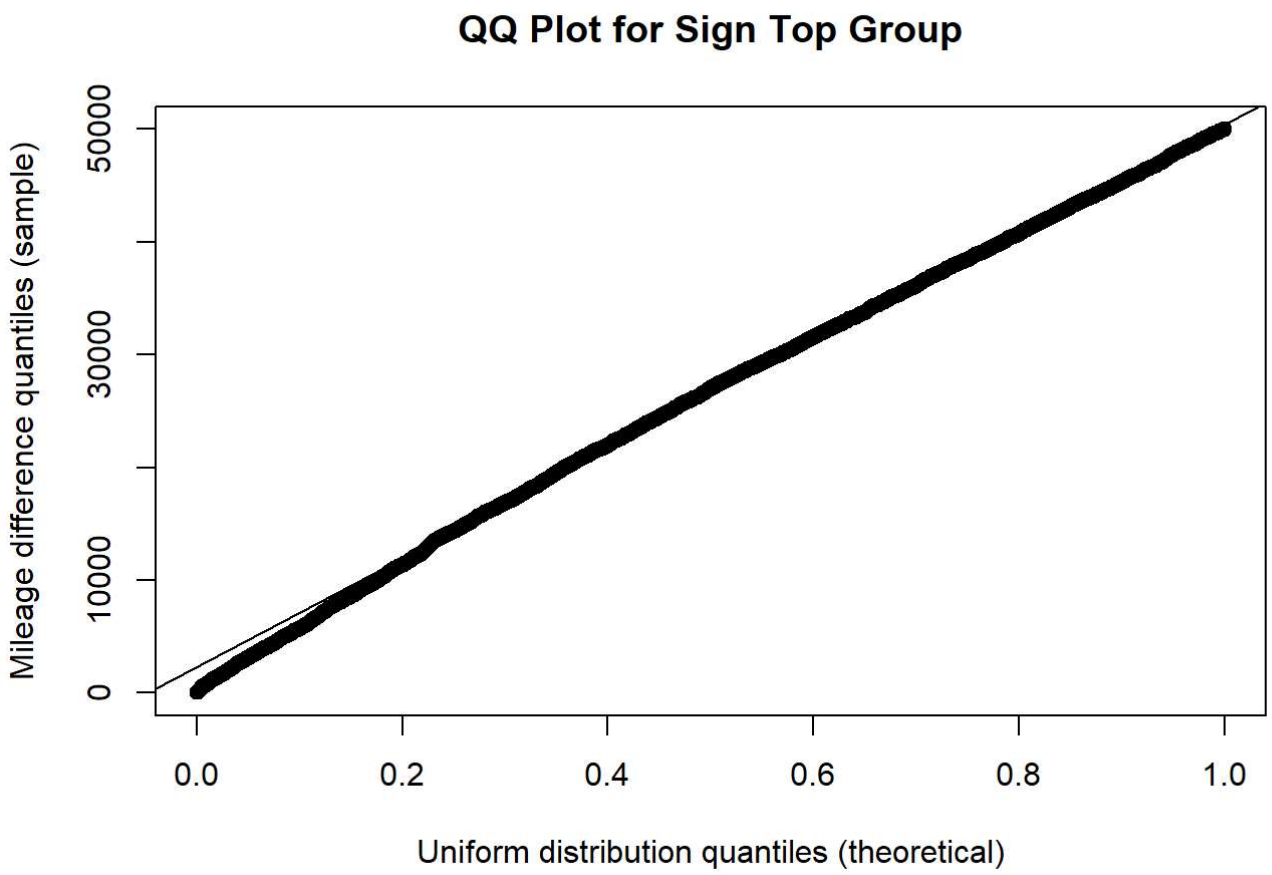
```
mean_bottom<- mean(sign_bottom)
print(mean_bottom)
```

```
## [1] 23622.55
```

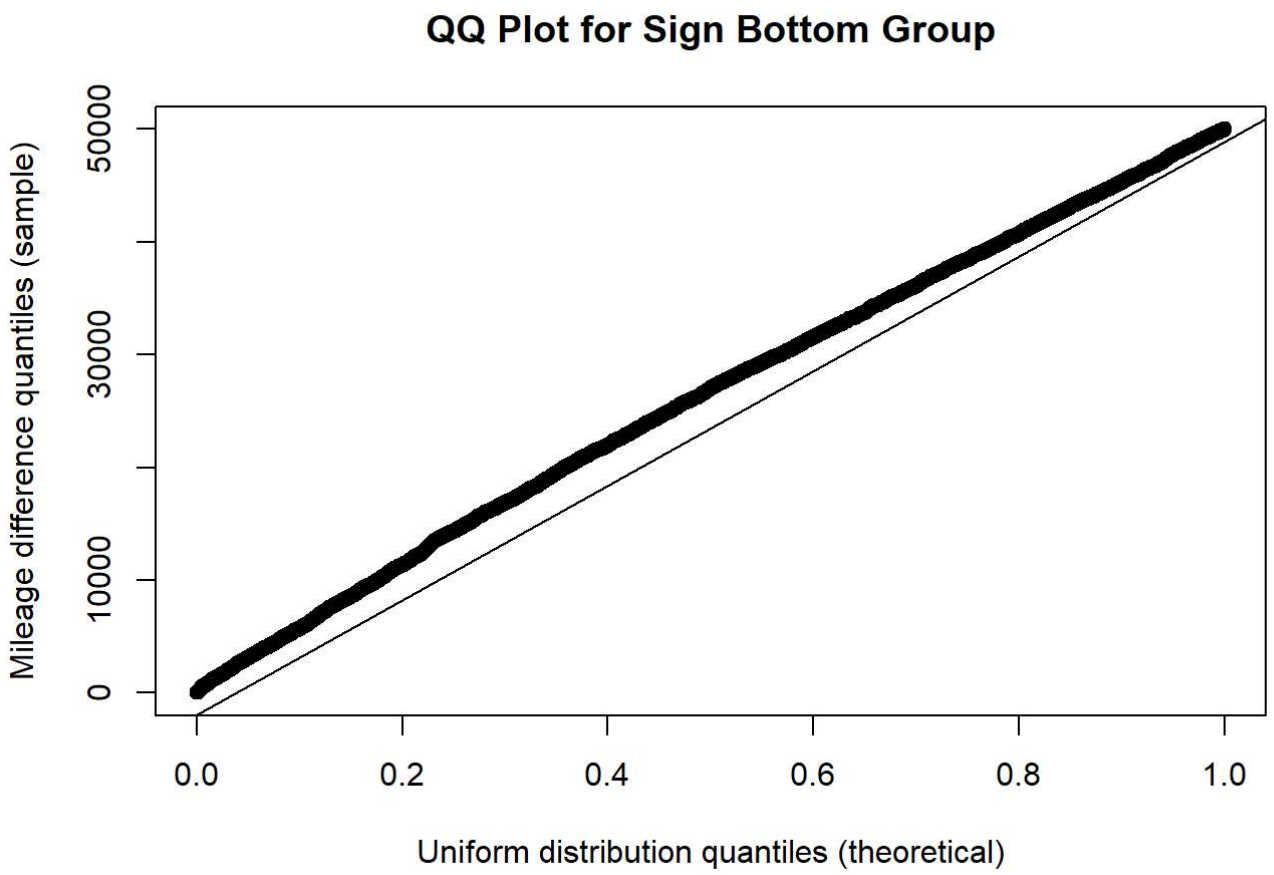
The mean mileage difference is greater for Sign_Top group.

#QQ plot

```
qqplot(qunif(ppoints(length(sign_top))), sign_top, ylab = "Mileage difference quantiles (sampl
e)",xlab="Uniform distribution quantiles (theoretical)", main="QQ Plot for Sign Top Group")
qqline(sign_top, distribution = qunif)
```




```
qqplot(qunif(ppoints(length(sign_bottom))), sign_top, ylab = "Mileage difference quantiles (samp
le)",xlab="Uniform distribution quantiles (theoretical)", main="QQ Plot for Sign Bottom Group")
qqline(sign_bottom, distribution = qunif)
```



#CONCLUSION The qq plot shows that the quantile-quantile intersection for the data set and uniform distribution aligned more to the straight line than any other distributions suggesting that the mileage difference of the Sign Top group and Sign Bottom group follows a uniform distribution. In this scenario we could think that the data could be fraudulent as there are high chances that the data could be fabricated to fit the uniform distribution. Because the chances of these many people reporting miles with similar mileage difference is very rare. Since this data is randomly collected the differences should have fit a different distribution. Another possibility is that the insurance company could have shown bias in selecting the participants to get their expected outcome or they could have omitted some of the data points(reports of people) that would disagree with the expected results(uniform distribution). So there are high chances that the data is fraudulent.