Problem Set1

Himani Anil Deshpande

9/7/2021

Loading the first four columns from the Dataset

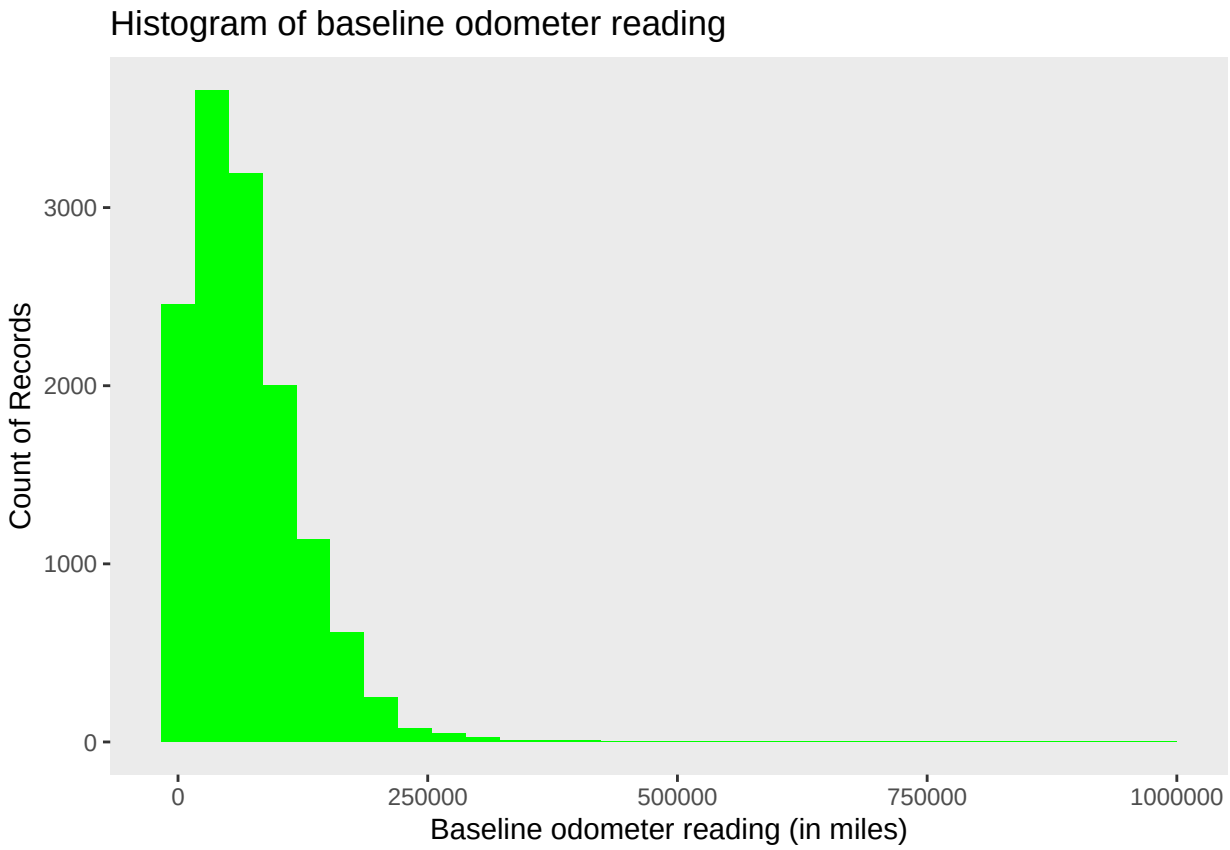
```
data <- read_excel("Assignment 1/DrivingdataAll.xls", range = cell_cols("A:D"))


#View(data)
```

Analysing the different Odometer readings using Histograms

Baseline Odometer Reading

```
ggplot(data, aes(x = `Odom Reading 1 (Previous)`)) +
  geom_histogram(fill = 'green') +
  xlab("Baseline odometer reading (in miles)") +
  ylab("Count of Records") +
 ggtitle("Histogram of baseline odometer reading") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

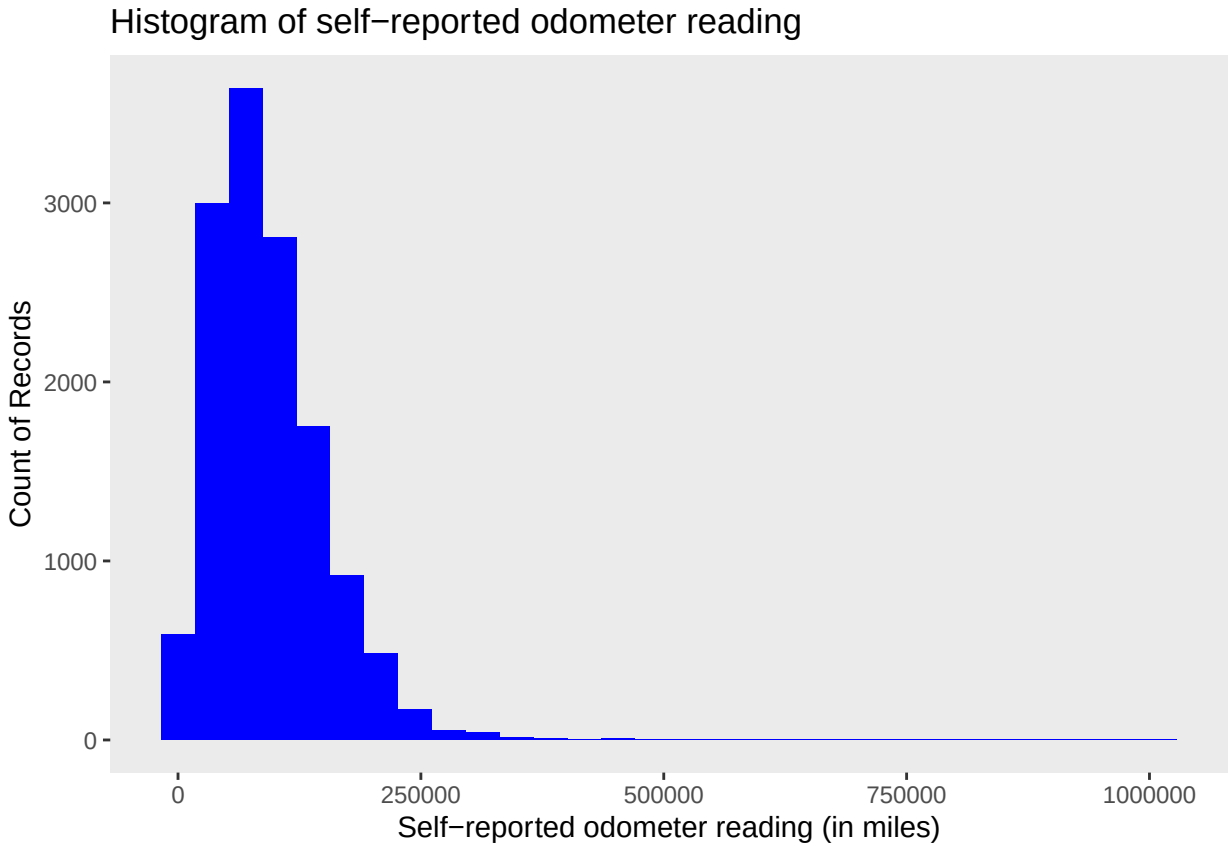
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Self-reported Odometer Reading

```
ggplot(data, aes(x = `Odom Reading 1 (Update)`)) +
  geom_histogram(fill = 'blue') +
  xlab("Self-reported odometer reading (in miles)") +
  ylab("Count of Records") +
  ggtitle("Histogram of self-reported odometer reading") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Both the histograms are right skewed. So, the mean is greater than median.

Performing Log transformation to further anlayze the distribution

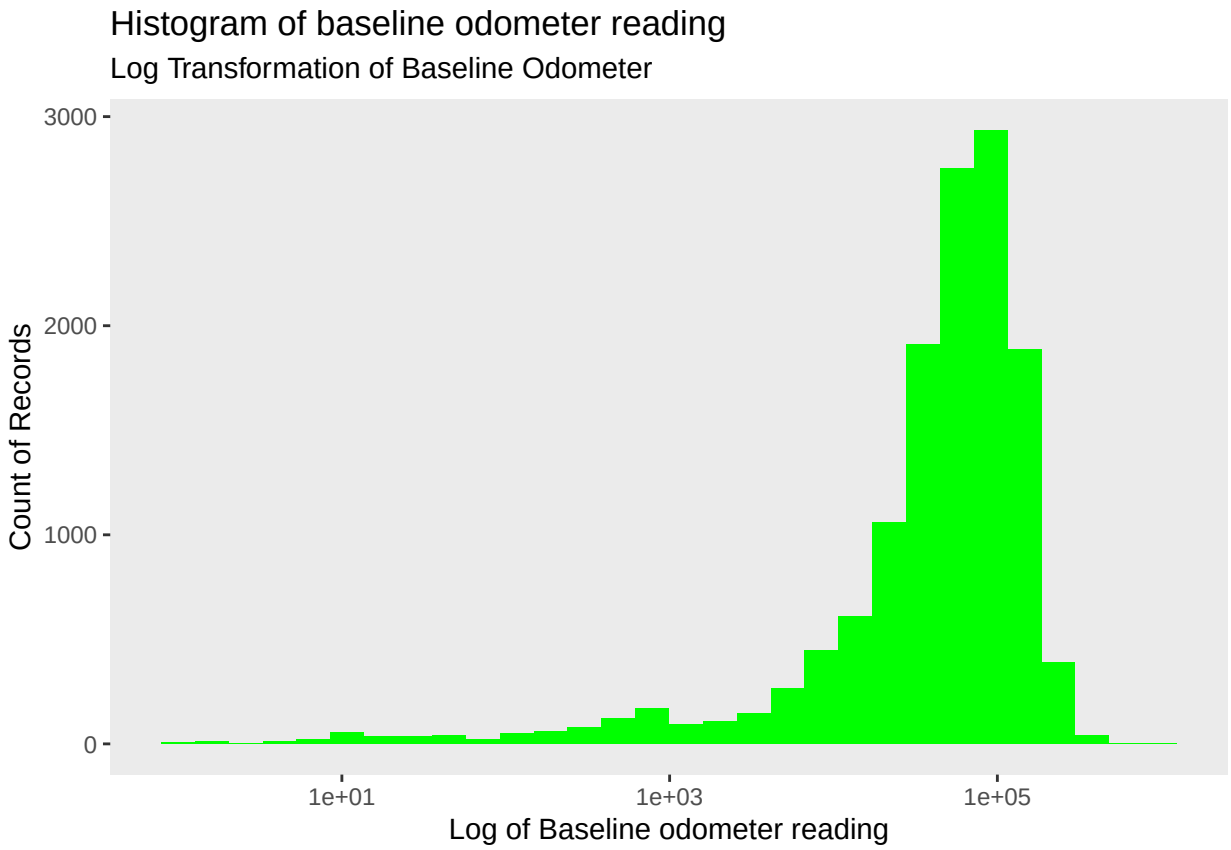
Baseline Odometer Reading

```
ggplot(data, aes(x = `Odom Reading 1 (Previous)`)) +
  geom_histogram(fill = 'green') +
  scale_x_log10() +
  xlab("Log of Baseline odometer reading") +
  ylab("Count of Records") +
 ggtitle("Histogram of baseline odometer reading") +
  labs(subtitle = "Log Transformation of Baseline Odometer") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

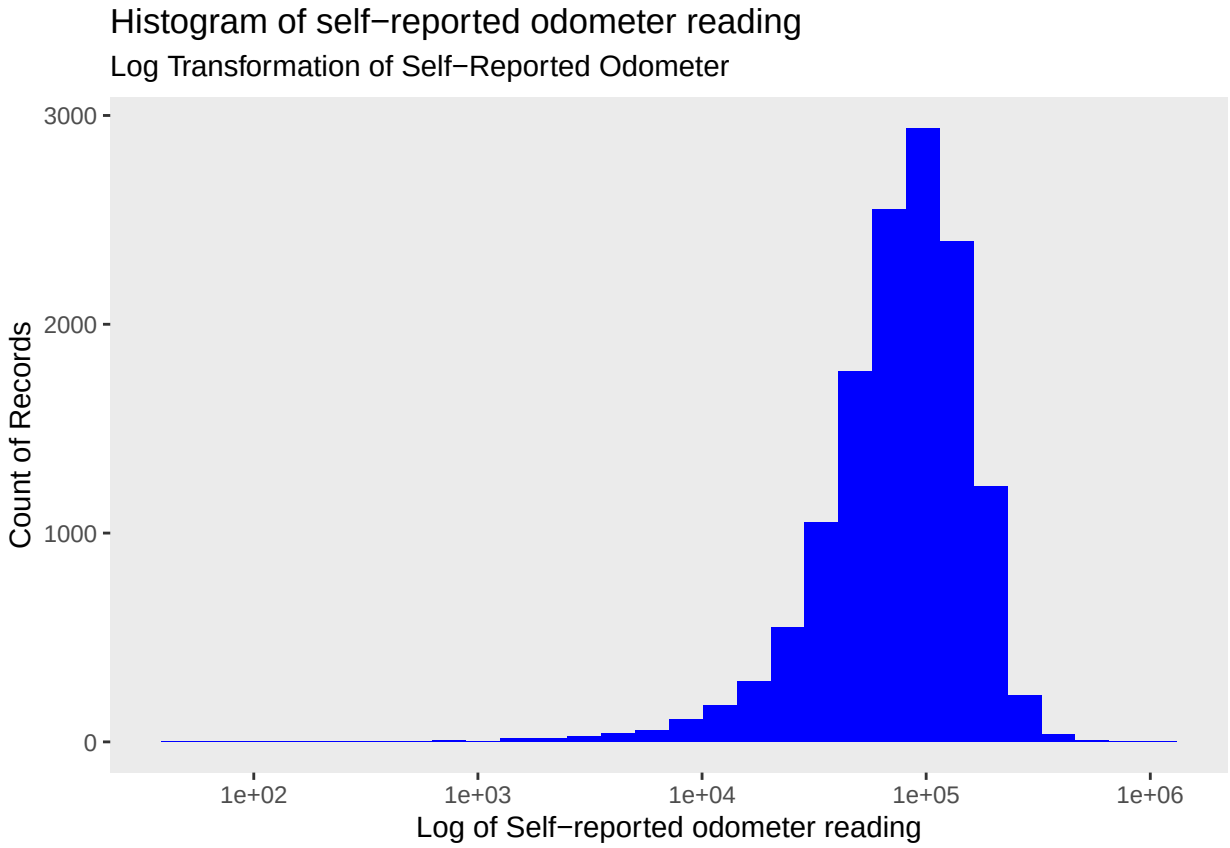
```
## Warning: Removed 117 rows containing non-finite values (stat_bin).
```



Self-reported Odometer Reading

```
ggplot(data, aes(x = `Odom Reading 1 (Update)`)) +
  geom_histogram(fill = 'blue') + scale_x_log10() +
  xlab("Log of Self-reported odometer reading") +
  ylab("Count of Records") +
  ggtitle("Histogram of self-reported odometer reading") +
  labs(subtitle = "Log Transformation of Self-Reported Odometer") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

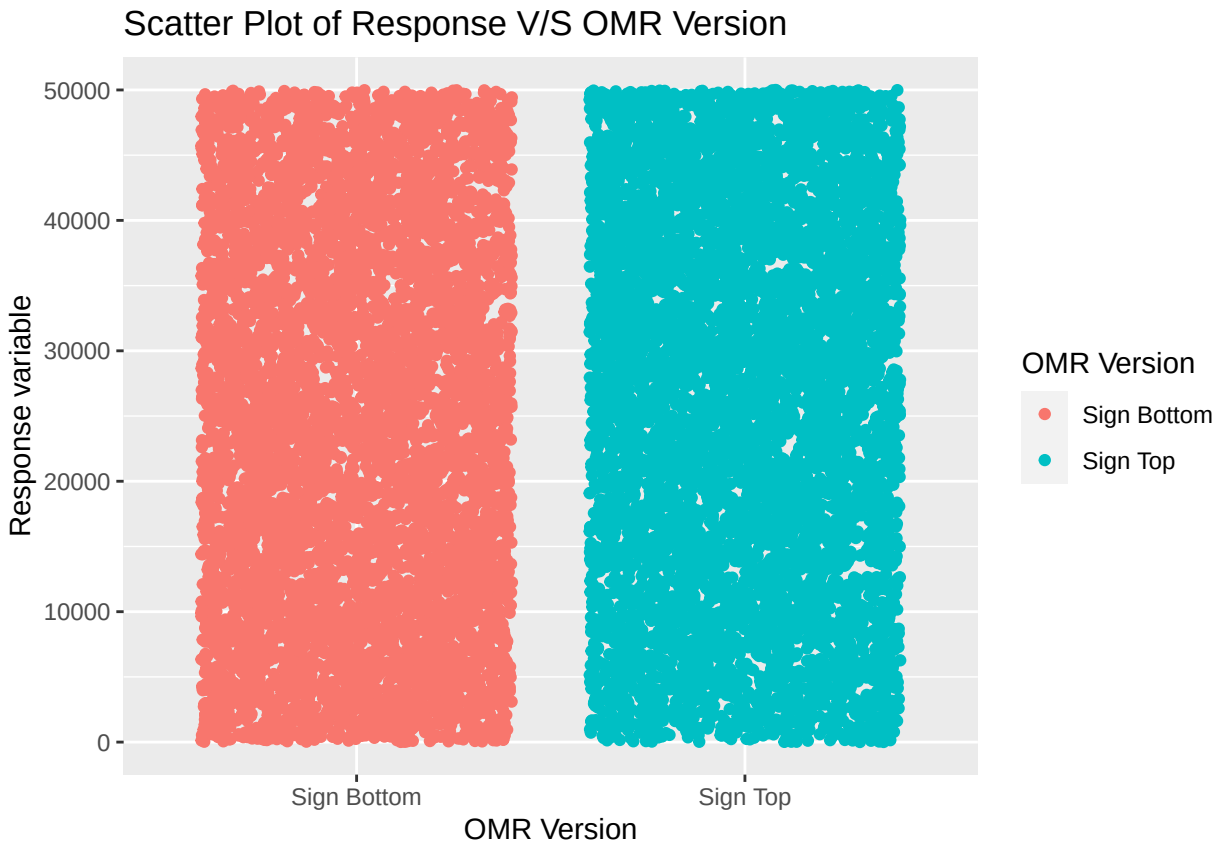


The log transformation makes the data left skewed. So we cannot use mean for analysis.

As mentioned in the paper, The response variable was the difference between self-reported miles and baseline miles(Column D minus Column C.) So I am going to analyze the response variable and its histogram

Analyzing the reponse Variable

```
ggplot(data,  aes(y =`Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)` ,
                  x = `OMR Version`, col = `OMR Version`)) +
  ylab("Response variable") +
  ggtitle("Scatter Plot of Response V/S OMR Version") +
             geom_jitter(position = position_jitter(height = 0, width=0.4))
```

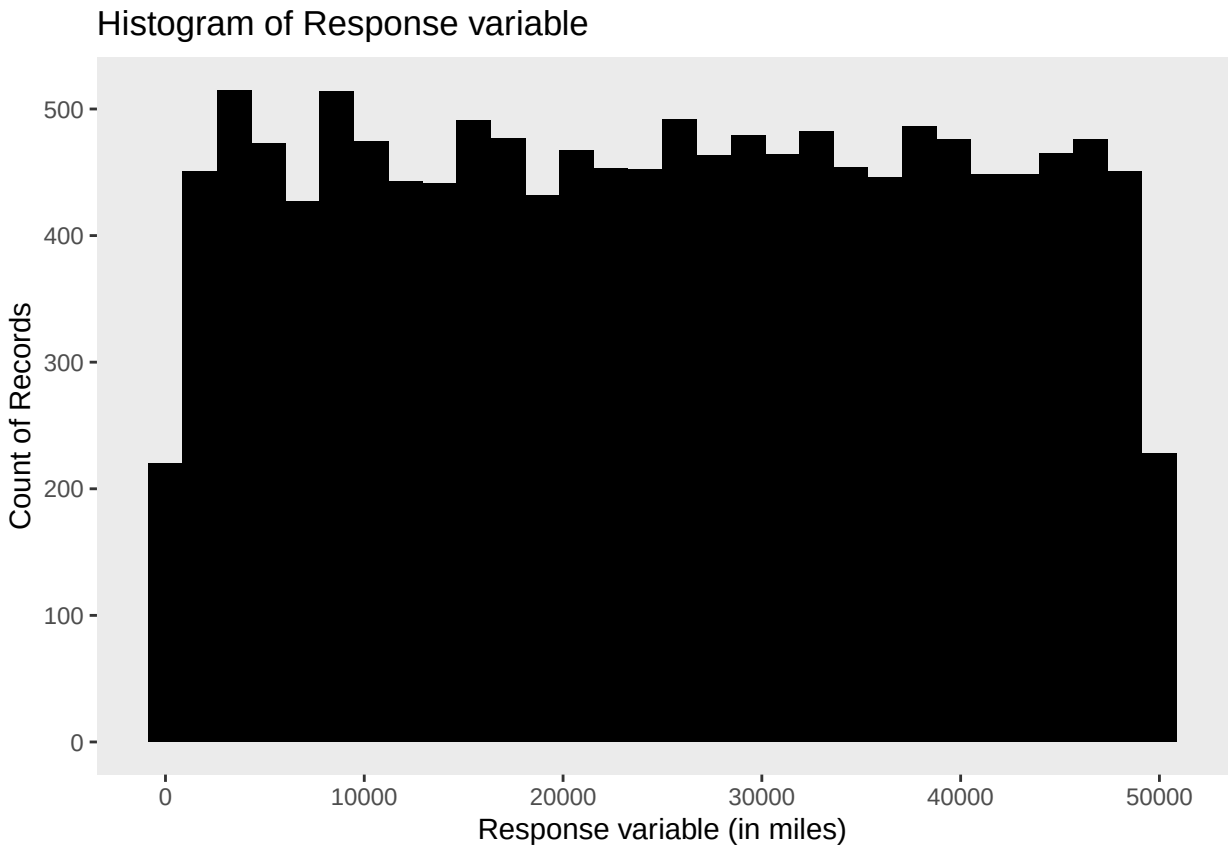


The batches share the same range So I will check the spread using the histogram

Histogram of Response Variable

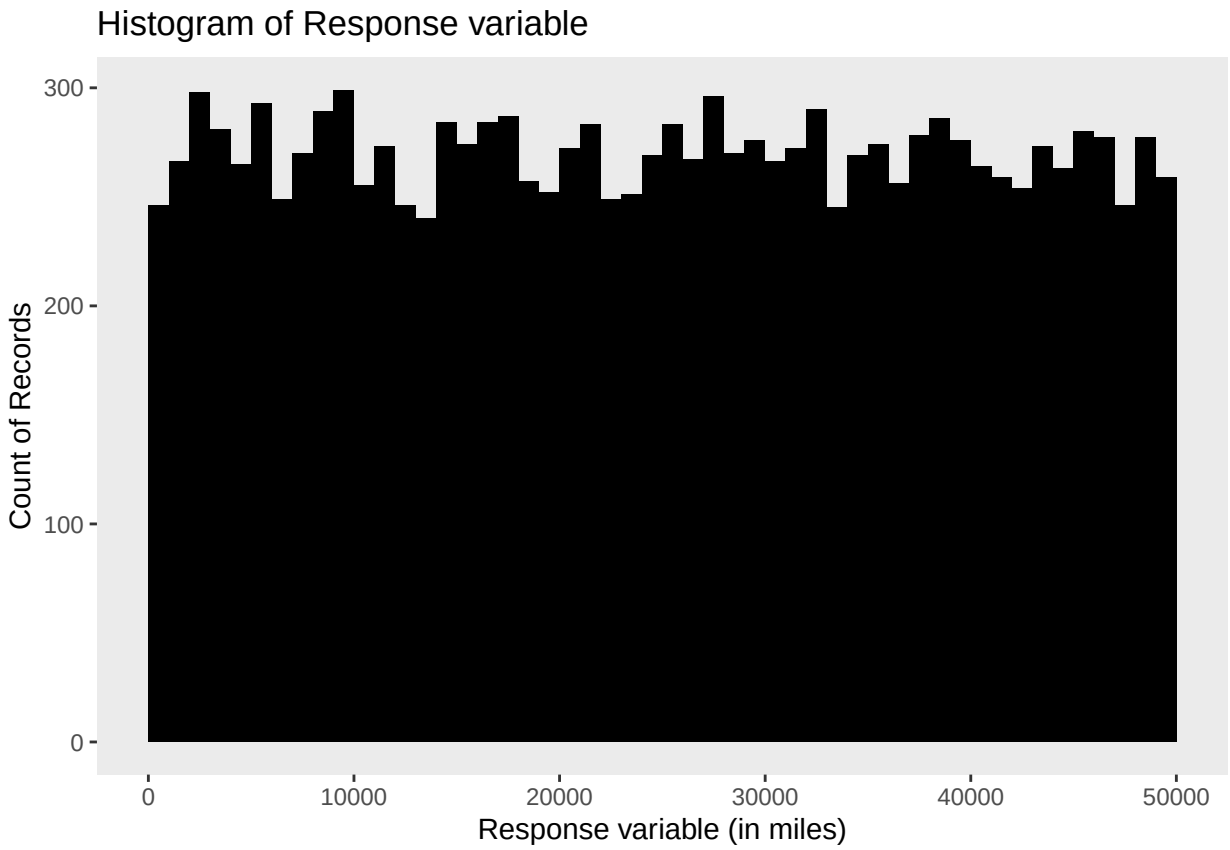
```
ggplot(data, aes(x =`Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)`)) +
  geom_histogram(fill = 'black') +
  xlab("Response variable (in miles)") +
  ylab("Count of Records") +
 ggtitle("Histogram of Response variable ") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Trying Histogram with different Bin sizes

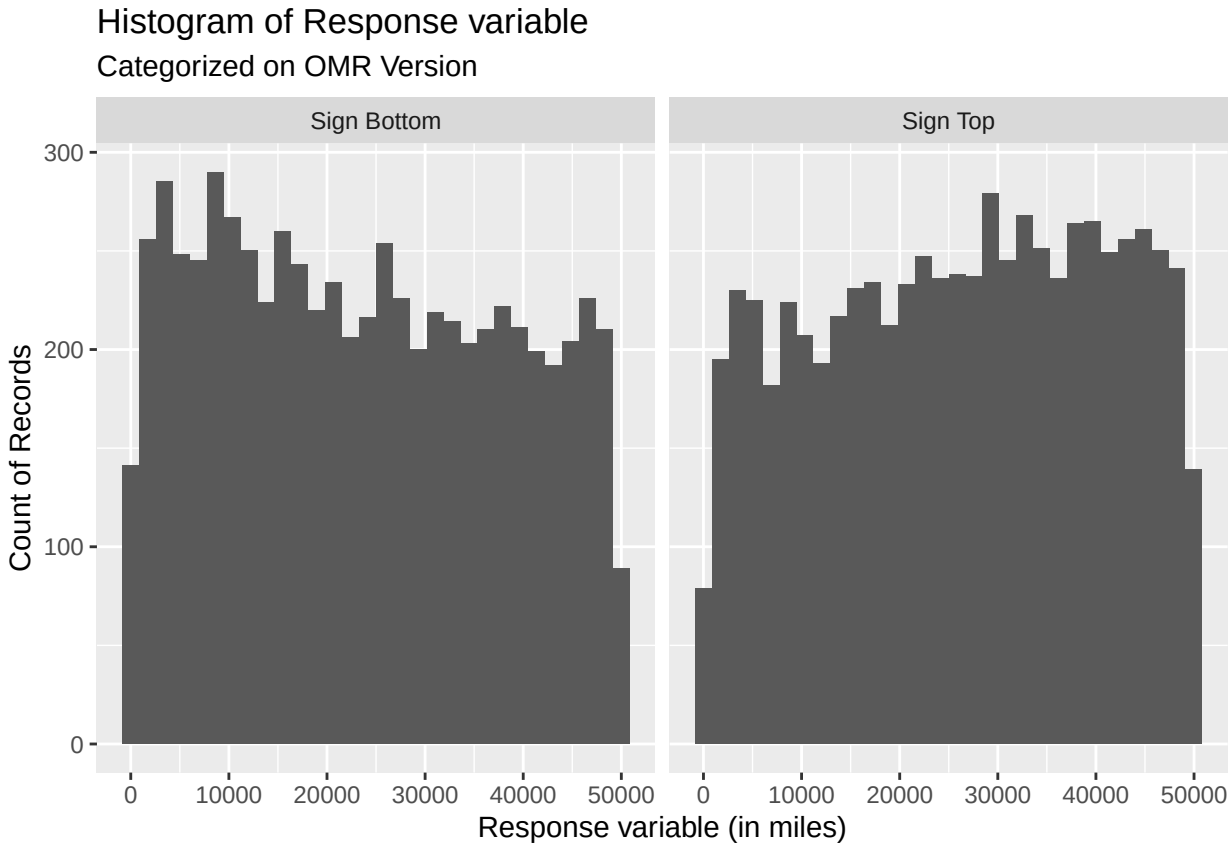
```
ggplot(data, aes(x =`Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)`)) +
  geom_histogram(fill = 'black', breaks = seq(-1, 50500, 1000)) +
  xlab("Response variable (in miles)") +
  ylab("Count of Records") +
 ggtitle("Histogram of Response variable ") +
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```



Histogram of Reponse variable categorized over OMR Version

```
ggplot(data, aes(x = `Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)`)) +
  geom_histogram() +
  xlab("Response variable (in miles)") +
  ylab("Count of Records") +
 ggtitle("Histogram of Response variable ") +
  labs(subtitle = "Categorized on OMR Version") +
  facet_grid(~`OMR Version`)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



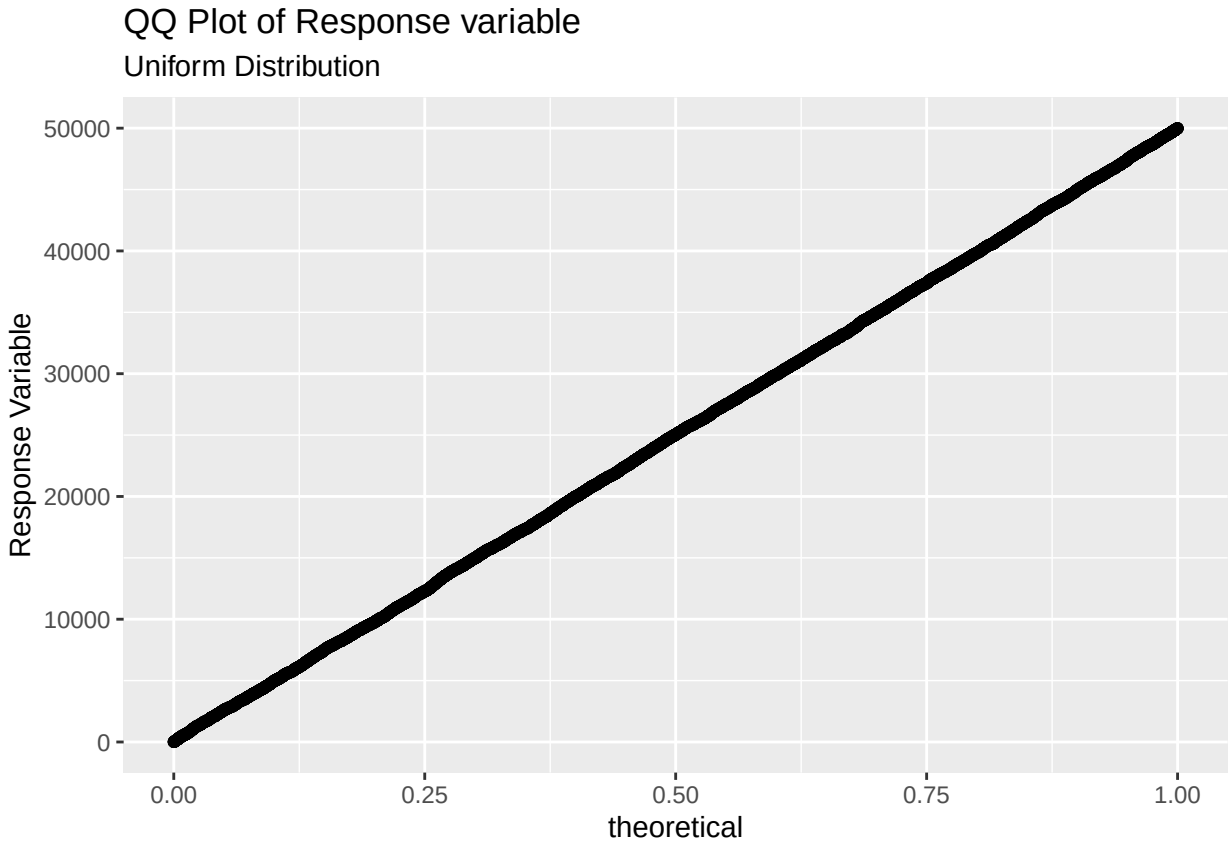
The histogram is not skewed and look symmetric So we dont need to perform any transformations for understanding the distribution of response variable. The histogram of response variable(self reported - baseline odometer readings) clearly shows that it is a uniform distribution.

So we will go ahead and check if the set of data comes from same distribution by plotting QQ plots.

QQ Plots

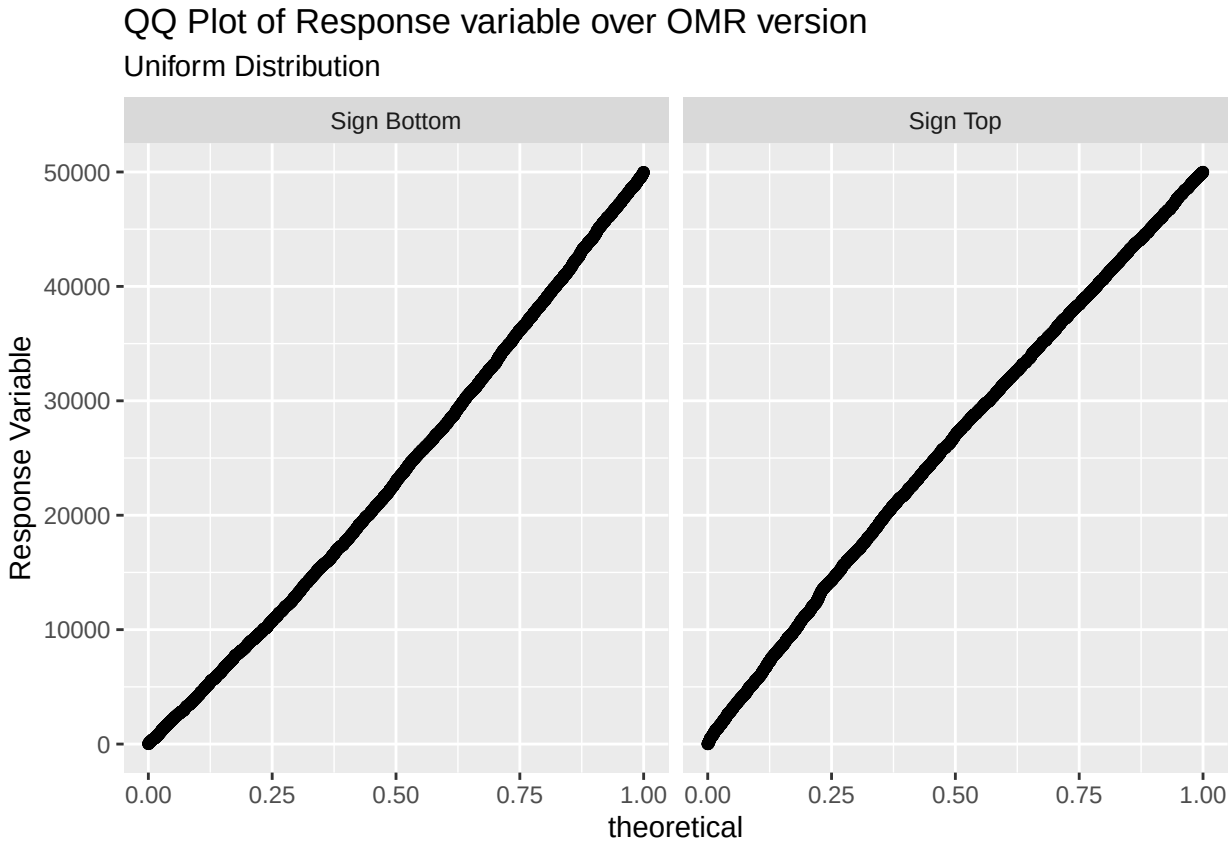
QQ Plot of Response variable

```
ggplot(data, aes(sample = `Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)`))+
  stat_qq(distribution = qunif) +
  ylab("Response Variable") +
  ggtitle("QQ Plot of Response variable") +
 labs(subtitle = "Uniform Distribution")
```



QQ plot of Response variable categorized over OMR Version

```
ggplot(data, aes(sample = `Odom Reading 1 (Update)`- `Odom Reading 1 (Previous)`))+
  stat_qq(distribution = qunif) +
  ylab("Response Variable") +
  ggtitle("QQ Plot of Response variable over OMR version") +
  labs(subtitle = "Uniform Distribution") +
  facet_grid(~`OMR Version`)
```



A straight line suggests a probability model, which implies that both the Self reported and Baseline Odometer reading came from the same distribution. The sample shows an exact straight line for Uniform distribution, which is rare in a real world data. The real world data usually poses some outliers or skewness in data, however the dataset we are working on doesnt have any outliers or any skewness with repect to response variable. All the data points falling on straight line with no deviation or curve implies that the data is fradulent. It seems that the data was created specifically to fit a Uniform distribution. So, any claim that the hypothesis is correct based on this data cannot be deemed trustworthy.