Daily Screen Time

Jefferson, Violet, Leah and Isaac

11/25/2020

Data Manipulation

Import Screen Time Data

We chose to collect our data through a Google form that was sent out to friends and family across various social media platforms. The survey simply asked students to type in their total screen time and their social media screen time in minutes for an average day. We gathered the responses on a Google Sheet and converted it to a csv file.

```
my_data <- read.csv(file.choose())
```

Check Screen Time Data

The `head()` function displays 6 data columns by default.

```
head(my_data)
```

```
##            Timestamp Total_Screen_Time Social_Media
## 1 11/23/2020 13:13:10               180           88
## 2 11/23/2020 13:16:42               644          427
## 3 11/23/2020 13:21:37               210           70
## 4 11/23/2020 13:26:05               375          264
## 5 11/23/2020 13:42:25               300          180
## 6 11/23/2020 13:43:24               400          300
```

Check Screen Time Data (with parameter)

```
head(my_data,10)
```

```
##             Timestamp Total_Screen_Time Social_Media
## 1  11/23/2020 13:13:10               180           88
## 2  11/23/2020 13:16:42               644          427
## 3  11/23/2020 13:21:37               210           70
## 4  11/23/2020 13:26:05               375          264
## 5  11/23/2020 13:42:25               300          180
## 6  11/23/2020 13:43:24               400          300
## 7  11/23/2020 14:10:16               150           30
## 8  11/23/2020 14:11:35               240          120
## 9  11/23/2020 14:11:45               240          150
## 10 11/23/2020 14:12:27              1080          240
```

Descriptive Statistics

Measures of Central Tendency

The means of both Total_Screen_Time and Social_Media are greater than the medians of either variable. This implies that the distribution is right skewed.

Mean

```
mean(my_data$Total_Screen_Time)
```

```
## [1] 434.1193
```

```
mean(my_data$Social_Media)
```

```
## [1] 215.1193
```

Median

```
median(my_data$Total_Screen_Time)
```

```
## [1] 360
```

```
median(my_data$Social_Media)
```

```
## [1] 122
```

Mode

```
require(modeest)
```

```
## Loading required package: modeest
```

```
## Warning: package 'modeest' was built under R version 4.0.3
```

```
mfv(my_data$Total_Screen_Time)
```

```
## [1] 300 400
```

```
mfv(my_data$Social_Media)
```

```
## [1] 120
```

Measures of Variablity

Range

The ranges below provide the maximum and minimum values of the Total_Screen_Time and the Social_Media. Social media screen time has a smaller and lower range than total screen time, illustrating that it is a subset of total screen time.

```
range(my_data$Total_Screen_Time)
```

```
## [1]   45 1500
```

```
range(my_data$Social_Media)
```

```
## [1]    0 1300
```

Interquartile Range

The Interquartile Range gives us the middle 50% of our data for both Total_Screen_Time and Social_Media. A smaller interquartile range indicates a less variable data set. This is evident as the Social_Media IQR is smaller than the IQR of Total_Screen_Time and the variability of Social_Media is also much smaller than Total_Screen_Time.

```
# Compute the quantiles/percentiles
IQR(my_data$Total_Screen_Time)
```

```
## [1] 260
```

```
IQR(my_data$Social_Media)
```

```
## [1] 223
```

Variance

```
#Compute the Variance
var(my_data$Total_Screen_Time)
```

```
## [1] 88556.11
```

```
var(my_data$Social_Media)
```

```
## [1] 45813.51
```

Standard Deviation

```
#Compute the Standard Deviation
sd(my_data$Total_Screen_Time)
```

```
## [1] 297.5838
```

```
sd(my_data$Social_Media)
```

```
## [1] 214.0409
```

Summary Data

```
#Compute the Sepal Width Summary
summary(my_data)
```

```
##   Timestamp         Total_Screen_Time  Social_Media
##  Length:109         Min.   :  45.0    Min.   :   0.0
##  Class :character   1st Qu.: 240.0    1st Qu.:  77.0
##  Mode  :character   Median : 360.0    Median : 122.0
##                     Mean   : 434.1    Mean   : 215.1
##                     3rd Qu.: 500.0    3rd Qu.: 300.0
##                     Max.   :1500.0    Max.   :1300.0
```

Plots

Box Plots

Looking at the Screen Time outliers there are some that would make sense to eliminate. For example, the data point of 1500 minutes is impossible since there are 1440 minutes in a day. Therefore, it would make sense to eliminate this data point. Both boxplots also indicate that the data is right skewed.

```
#Outlier Rows
#Outlier rows for Average Total Screen Time
message("Screen Time Outliers:")
```

```
## Screen Time Outliers:
```

```
boxplot.stats(my_data$Total_Screen_Time)$out
```

```
##  [1] 1080  960 1500 1000 1400 1039  900 1080 1220  900  900 1000 1000
```

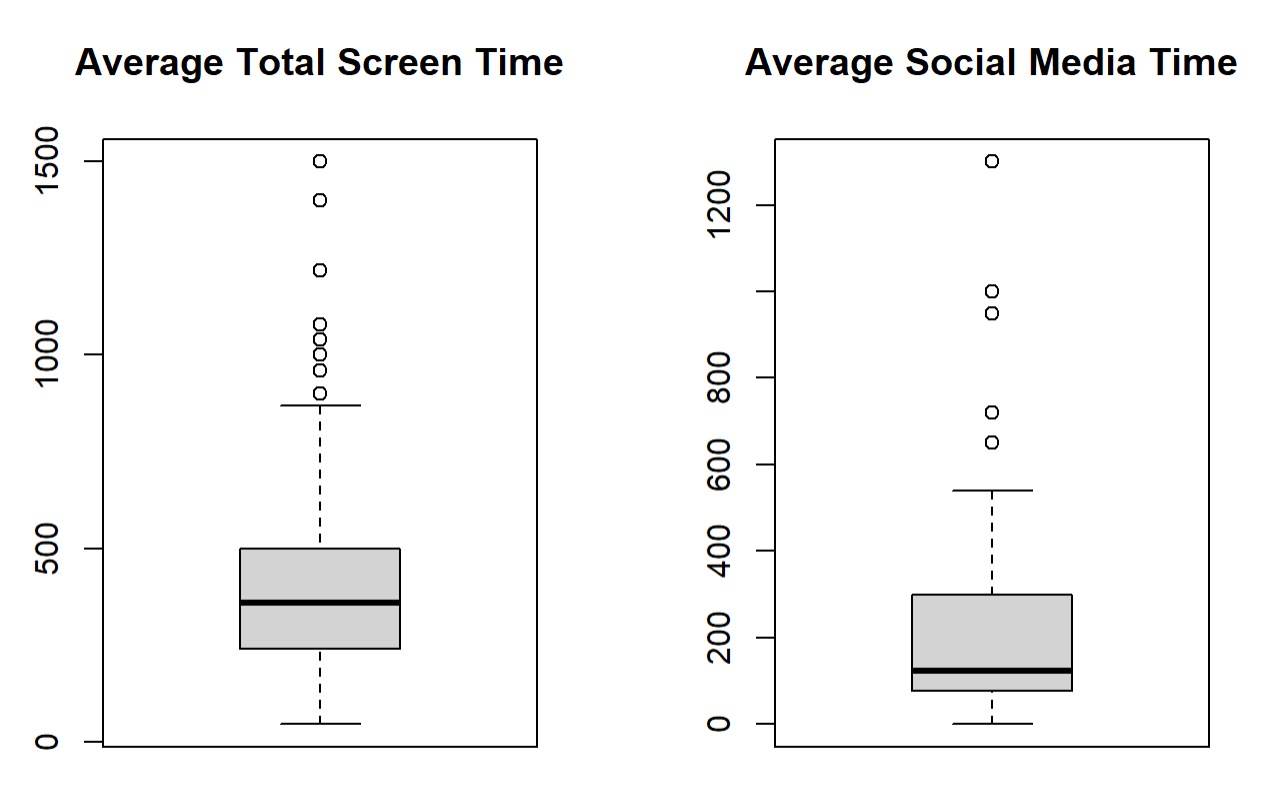
```
message("Social Media Time Outliers:")
```

```
## Social Media Time Outliers:
```

```
#Outlier rows for Average Social Media Time
boxplot.stats(my_data$Social_Media)$out
```

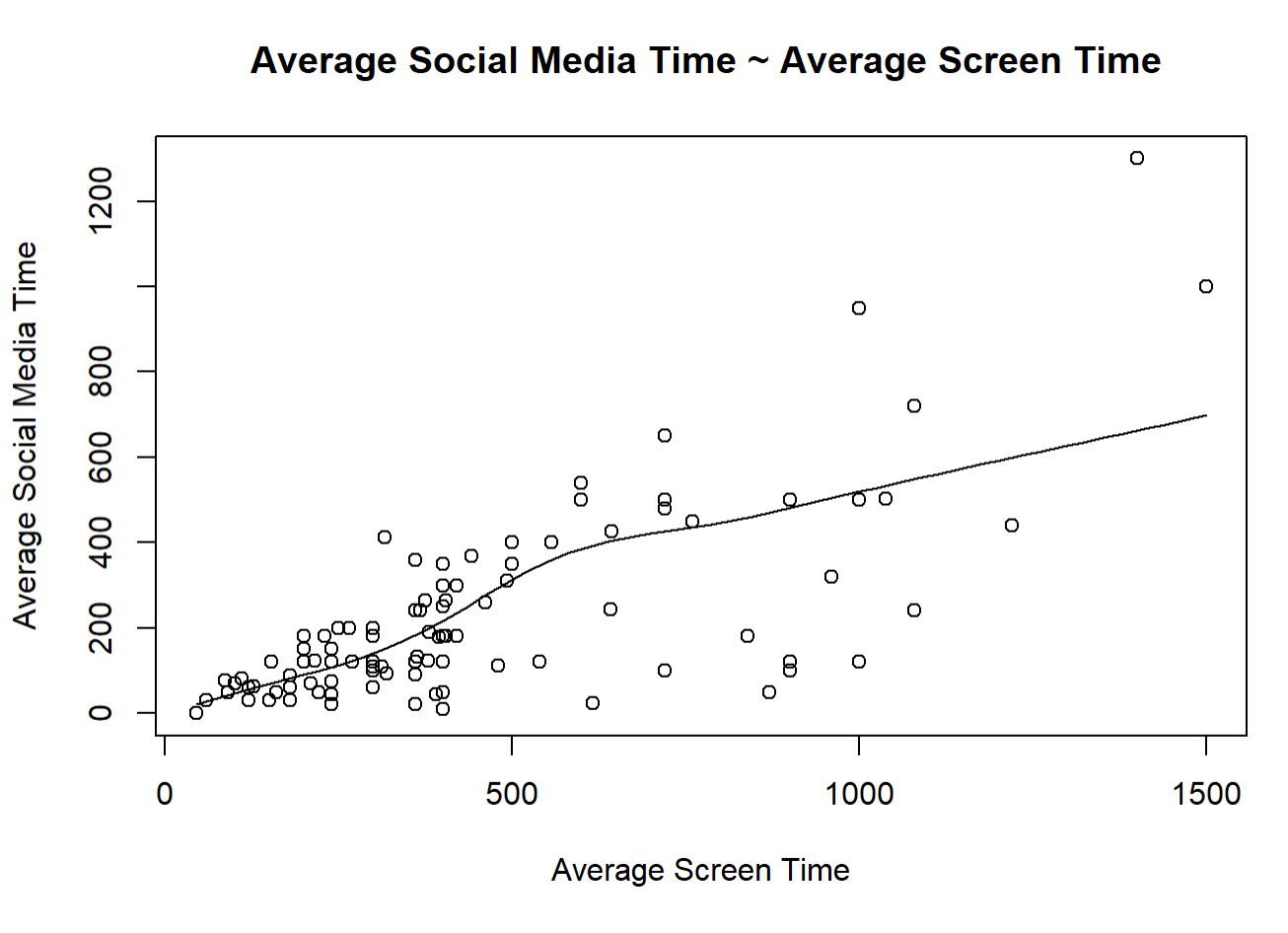
```
## [1]  650 1000 1300  720  950
```

```
#box plots
par(mfrow=c(1, 2))  # divide graph area in 2 columns
boxplot(my_data$Total_Screen_Time, main="Average Total Screen Time")
boxplot(my_data$Social_Media, main="Average Social Media Time")
```



Scatter Plots

```
#observed positive correlation
scatter.smooth(x=my_data$Total_Screen_Time, y=my_data$Social_Media, main=" Average Social Media Time ~ Average Sc
reen Time",xlab="Average Screen Time",ylab="Average Social Media Time")
```



Density Plots

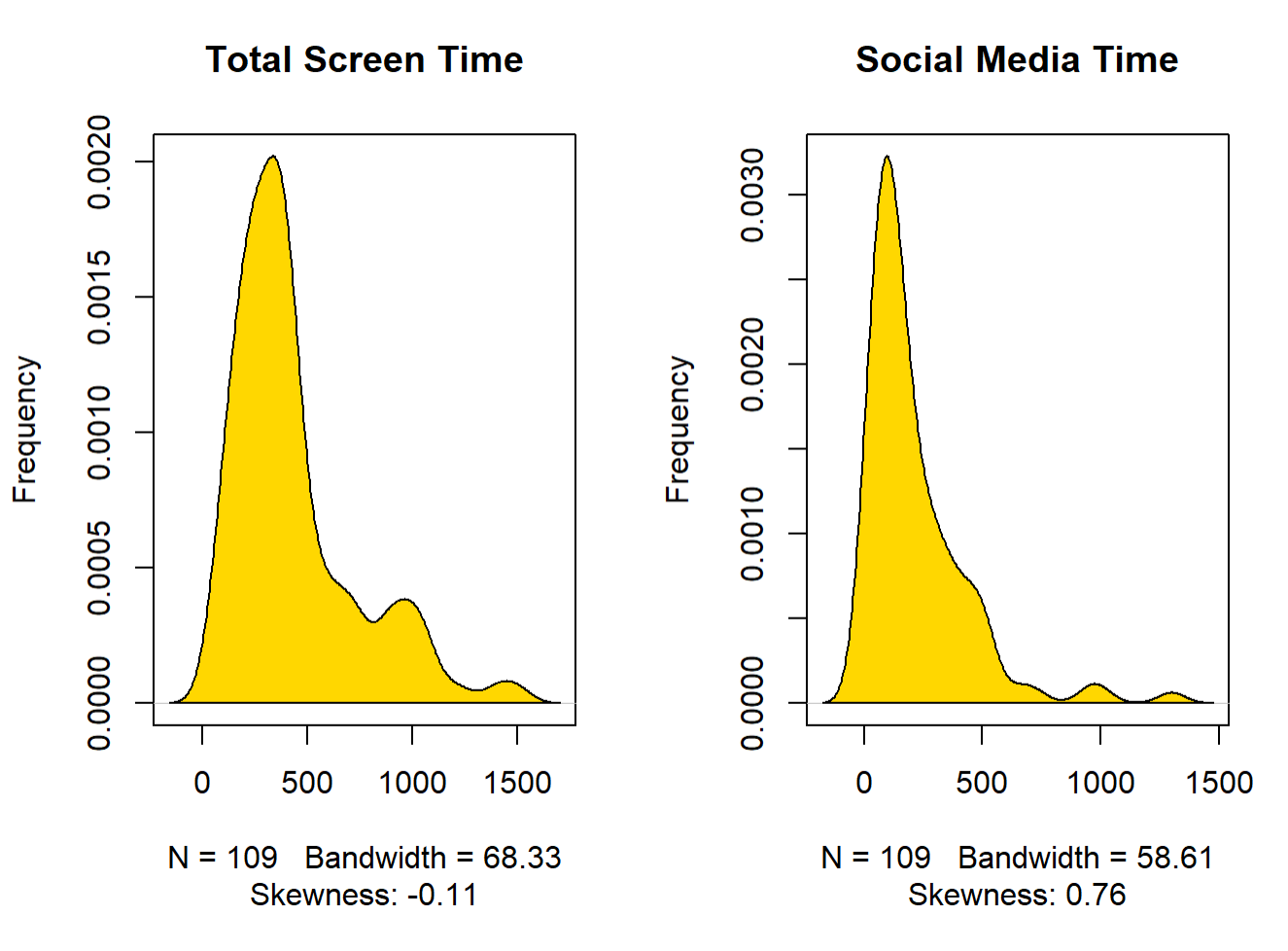
```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.0.3
```

```
##
## Attaching package: 'e1071'
```

```
## The following object is masked from 'package:modeest':
##
##     skewness
```

```
par(mfrow=c(1, 2))  # divide graph area in 2 columns
plot(density(my_data$Total_Screen_Time), main="Total Screen Time", ylab="Frequency", sub=paste("Skewness:", round
(e1071::skewness(cars$speed), 2)))
polygon(density(my_data$Total_Screen_Time), col="gold")
plot(density(my_data$Social_Media), main="Social Media Time", ylab="Frequency", sub=paste("Skewness:", round(e107
1::skewness(cars$dist), 2)))
polygon(density(my_data$Social_Media), col="gold")
```



***

Correlation

There is sufficient evidence to conclude that there is a significant linear relationship between Total_Screen_Time and Social_Media because the correlation coefficient is significantly different from zero. We determined this by calculating the p value which is less than the significance level. Therefore, the null hypothesis is rejected. The correlation co-efficient was calculated using the `cor()` function.

```
message("Correlation Co-efficient: ",cor(my_data$Total_Screen_Time, my_data$Social_Media))
```

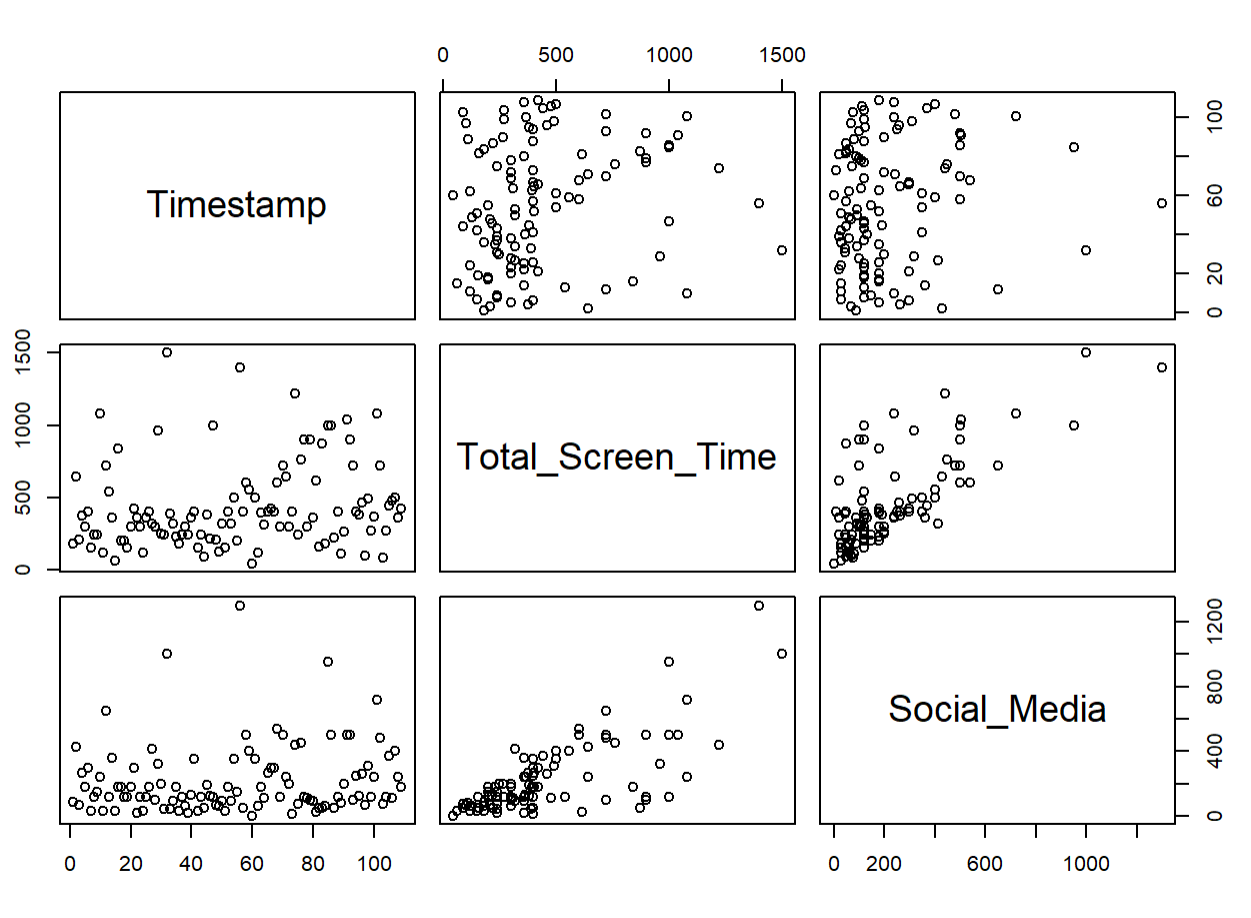
```
## Correlation Co-efficient: 0.731913398104009
```

```
cor.test(my_data$Total_Screen_Time, my_data$Social_Media)
```

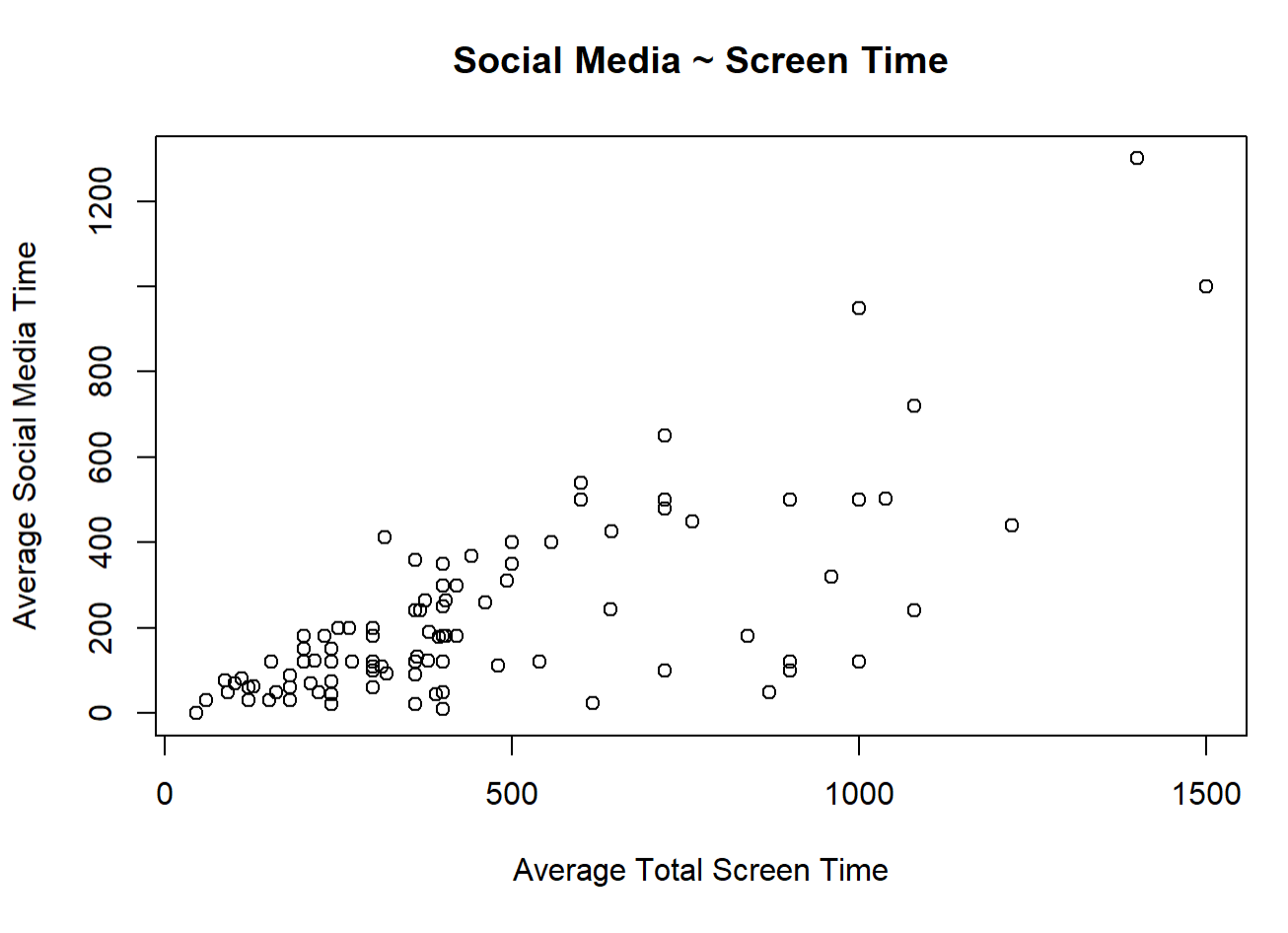
```
##
##  Pearson's product-moment correlation
##
## data:  my_data$Total_Screen_Time and my_data$Social_Media
## t = 11.111, df = 107, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6306336 0.8086807
## sample estimates:
##       cor
## 0.7319134
```

Data Plots

```
plot(my_data)
```

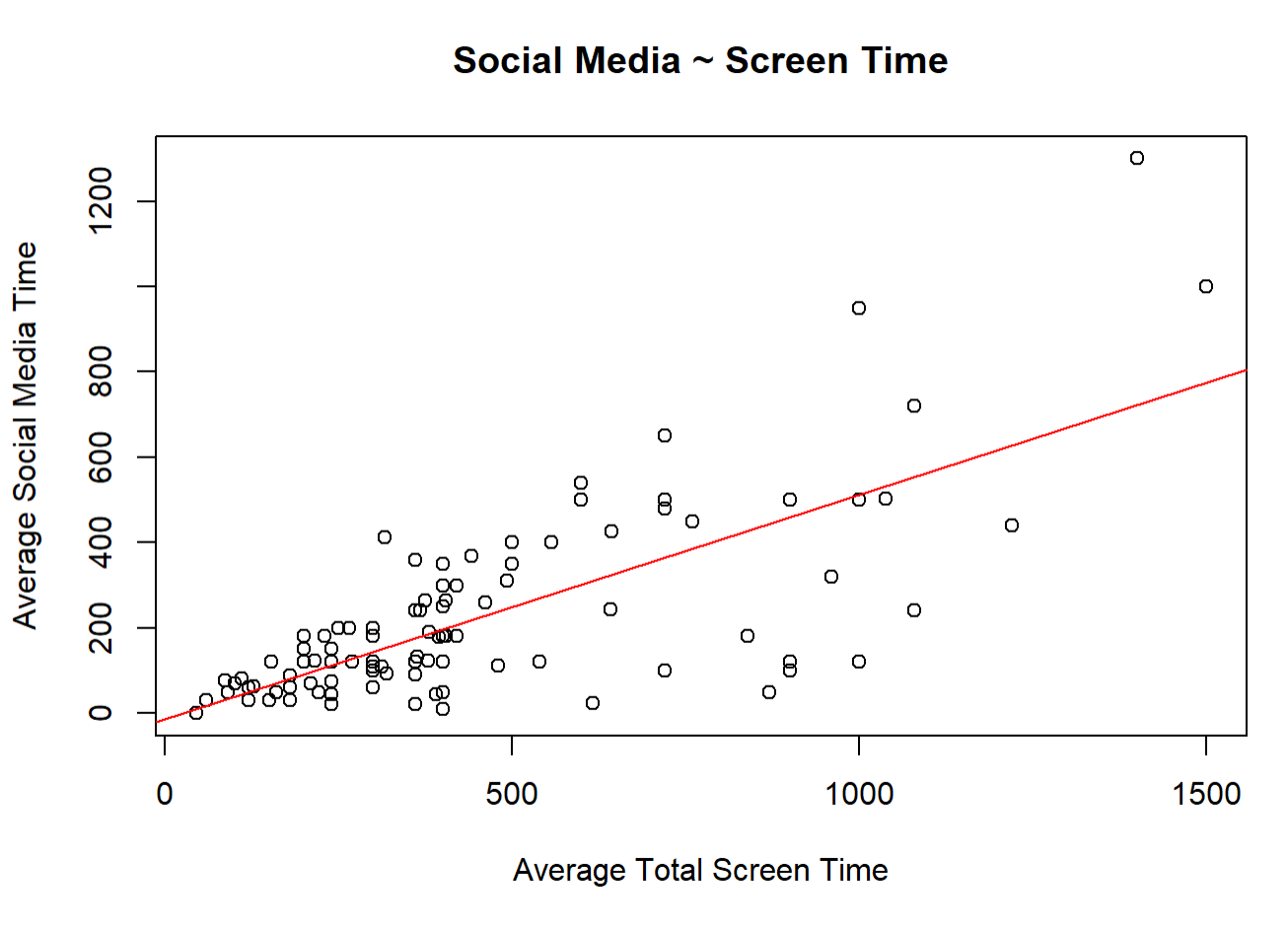



```
plot(x=my_data$Total_Screen_Time, y = my_data$Social_Media, xlab = "Average Total Screen Time", ylab = "Average S
ocial Media Time", main = "Social Media ~ Screen Time")
```



Linear Model

```
plot(x=my_data$Total_Screen_Time, y = my_data$Social_Media, xlab = "Average Total Screen Time", ylab = "Average S
ocial Media Time", main = "Social Media ~ Screen Time")
linearmod = lm(my_data$Social_Media~my_data$Total_Screen_Time)
abline(linearmod, col ="red")
```




```
message("Least Square Line:")
```

```
## Least Square Line:
```

```
linearmod
```

```
##
## Call:
## lm(formula = my_data$Social_Media ~ my_data$Total_Screen_Time)
##
## Coefficients:
##               (Intercept)  my_data$Total_Screen_Time
##                  -13.4176                     0.5264
```

The linear regression line does fit the data. The line has a similar number of points above and below it at any given interval. The variables begin strongly correlated when screen time is low and the correlation weakens as screen time increases, once more supporting that social media time could effectively be a subset of screen time since the spread of social media time is dependent on screen time.

```
summary(linearmod)
```

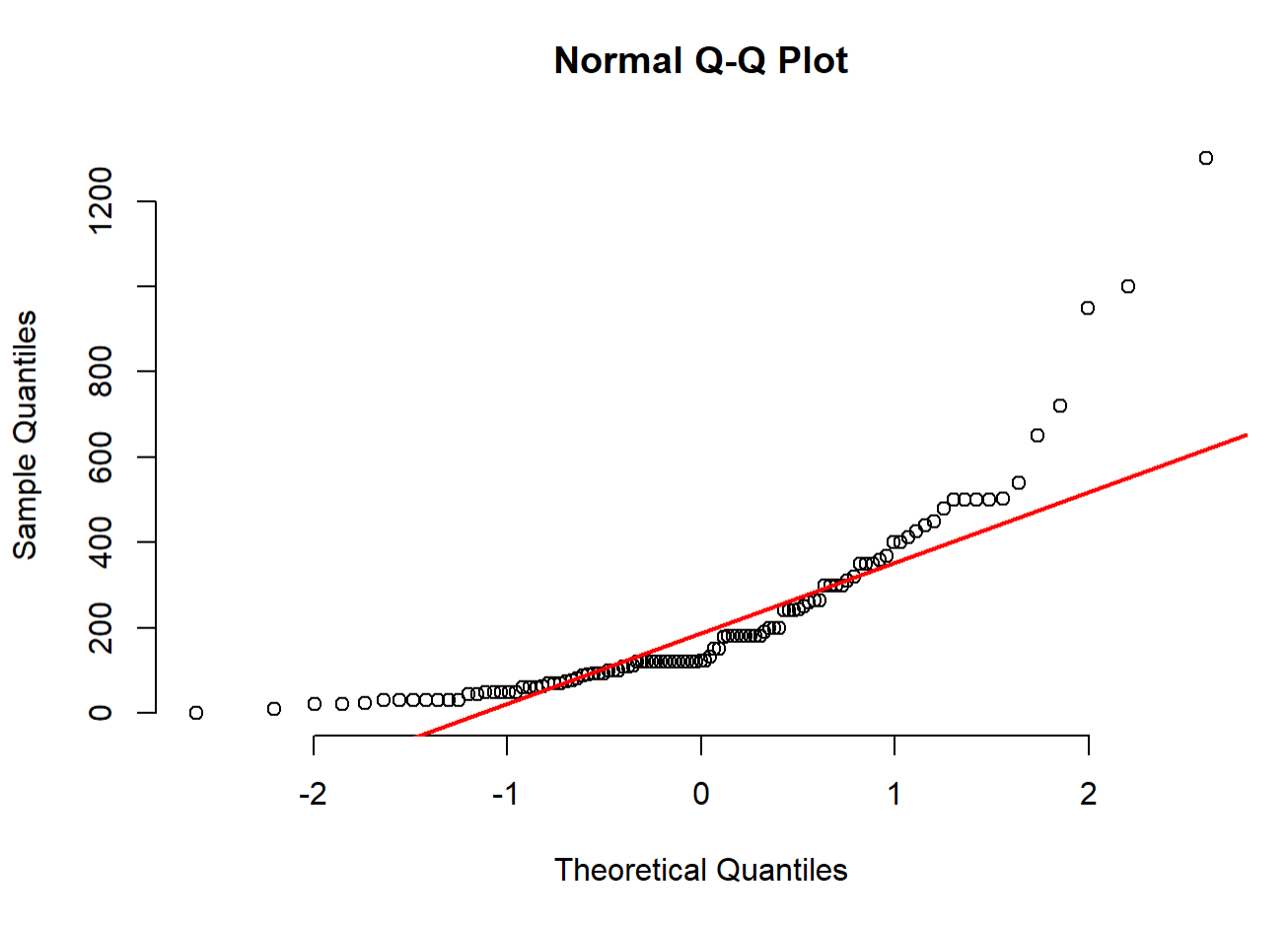
```
##
## Call:
## lm(formula = my_data$Social_Media ~ my_data$Total_Screen_Time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -394.58  -54.93    3.37   64.74  576.40
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -13.41762   24.90065  -0.539    0.591
## my_data$Total_Screen_Time   0.52644    0.04738  11.111   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.5 on 107 degrees of freedom
## Multiple R-squared:  0.5357, Adjusted R-squared:  0.5314
## F-statistic: 123.5 on 1 and 107 DF,  p-value: < 2.2e-16
```

Since the p-value is below the 0.05 threshold, the model is statistically significant. We can safely reject the null hypothesis that the coefficient beta of the predictor is zero. We can now go ahead and use it to predict (or estimate) the dependent variable.

QQ Plots

Social Media

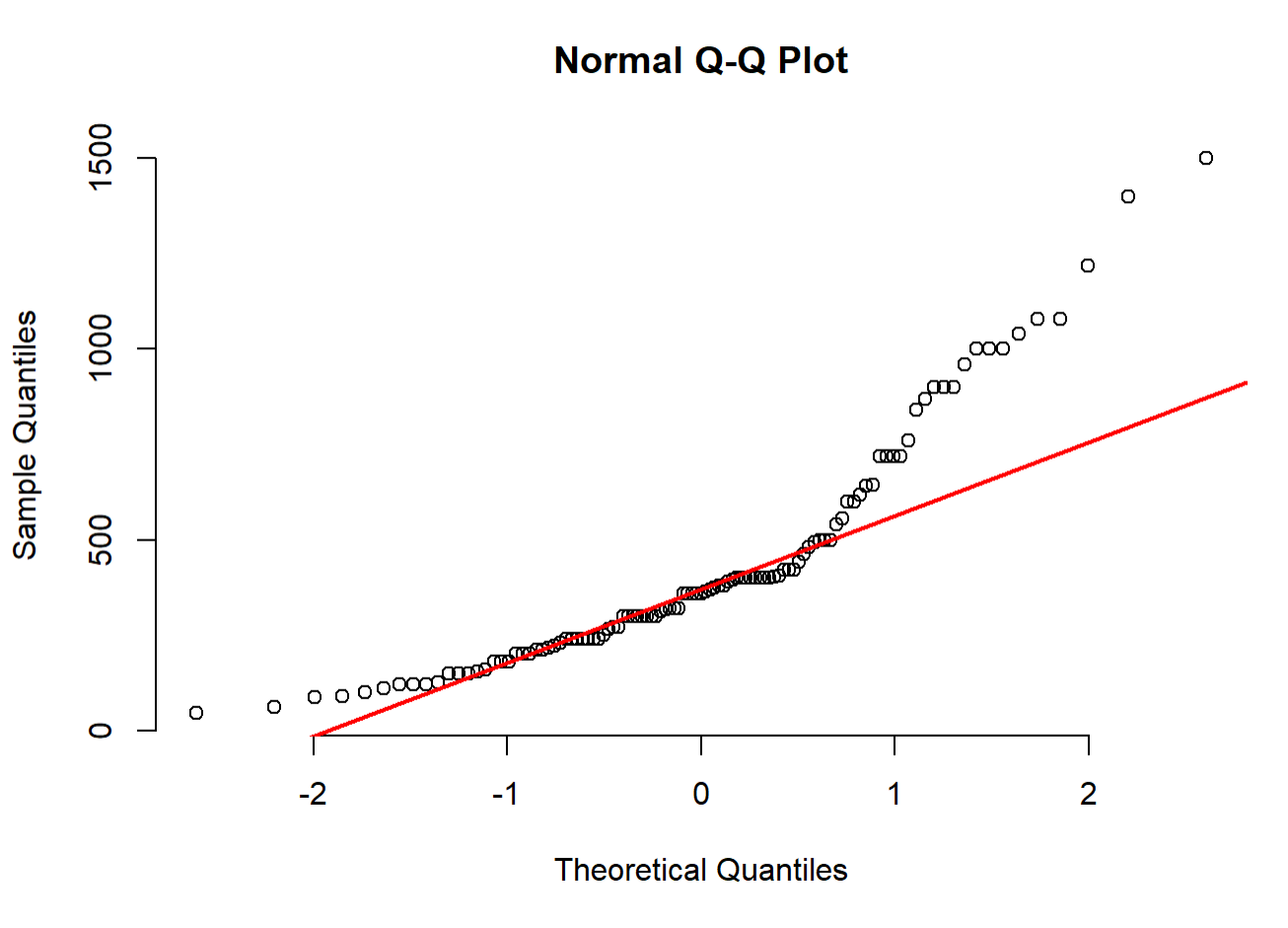
```
qqnorm(my_data$Social_Media, pch = 1, frame = FALSE)
qqline(my_data$Social_Media, col = "red", lwd = 2)
```



Most points do not fall on the line, indicating that Social Media Time fits poorly on the normal distribution.

Total Screen Time

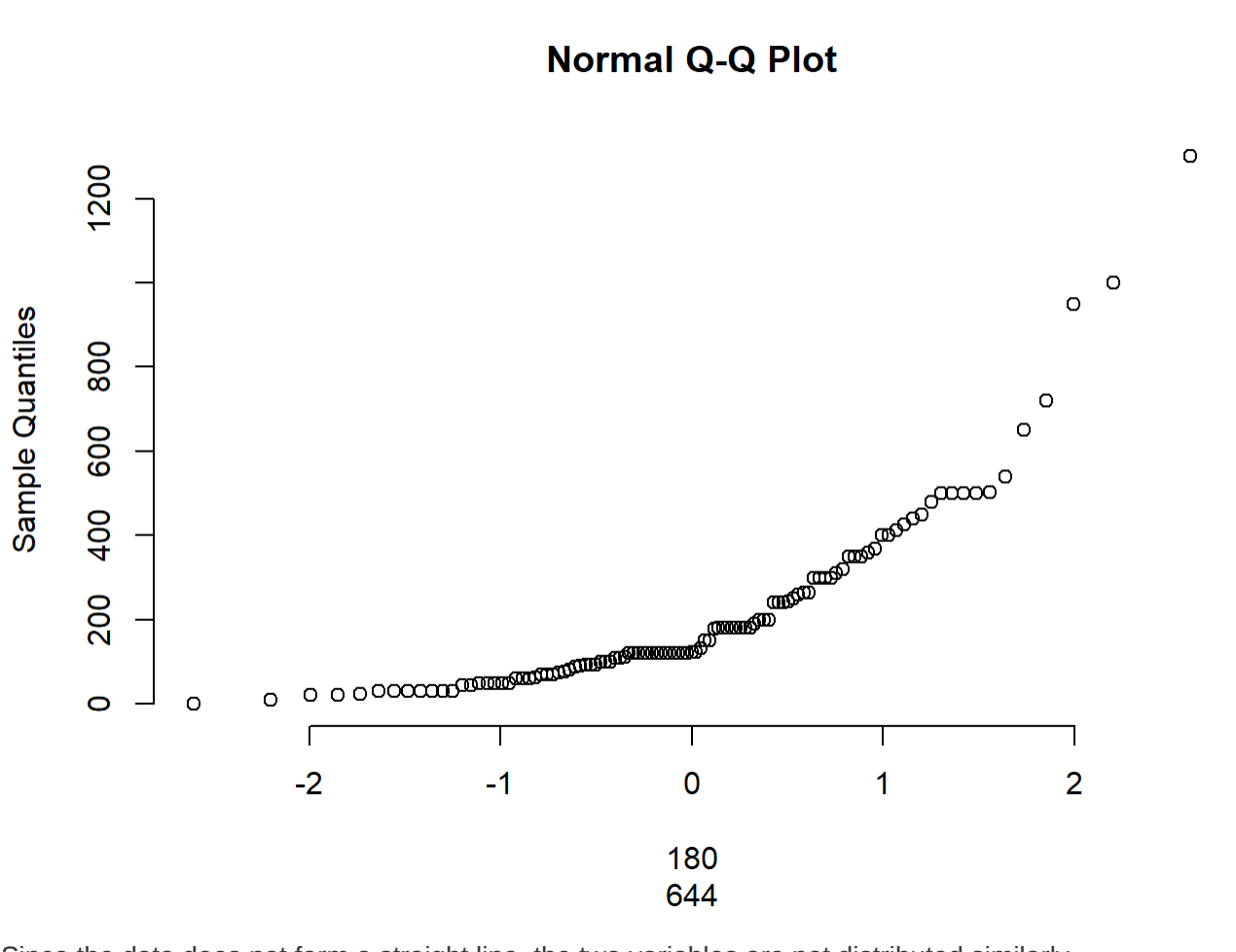
```
qqnorm(my_data$Total_Screen_Time, pch = 1, frame = FALSE)
qqline(my_data$Total_Screen_Time, col = "red", lwd = 2)
```



Most points do not fall on the line, indicating that Total Screen Time fits poorly on the normal distribution. There are more points falling on the line for screen time than for social media time indicating that while social media time does not fit the normal distribution well, it fits the normal distribution better than social media time does.

Total Screen Time vs Social Media Time

```
qqnorm(x = my_data$Total_Screen_Time, y = my_data$Social_Media, pch = 1, frame = FALSE)
```



Since the data does not form a straight line, the two variables are not distributed similarly.

Analysis

The variables are positively correlated with a weaker correlation as screen time increases.This can be attributed to the fact that social media is conventionally accessed via cellphone. Considering this, it is likely that most if not all time spent using social media was also time spent using a phone, effectively making social media time a subset of screen time.

This is a possible explanation for why only a single survey response had higher social media time than screen time. Social media time being a subset of screen time would also explain why there is a greater spread as screen time increases, since there is a higher maximum and the same minimum so the range of social media time values is greater.