Week 10 Project: Twitter

Libraries

```
library(tidyverse)
library(twitteR)
```

The tidyverse package is used for the data import and cleaning. The twitteR package is used to pull and process the tweets.

Data Import and Cleaning

```
# Authentication
apikey <- "PyeImX7Gt4rpjmjcb03mPap5z"
apisecret <- "saPpNxxnDRKP27GVpfrbi9ataqLSJNjf0nsFLnKvsRhECEmtdG"
token <- "859134854-5UgDzirz1QTUS9d4h4rZDMCJ3a90JgvNylk9z9PF"
secrettoken <- "9mV8UMlesNgRnjUhImvW1kLBkfrsE956G8eljg80gpKyg"

setup_twitter_oauth(apikey, apisecret, token, secrettoken)
```

```
## [1] "Using direct authentication"
```

```
# Pull the tweets from twitter
tweets_raw <- searchTwitteR("#rstats", n = 1000)
# remove retweets and select variables of interest
tweets_tbl <- strip_retweets(tweets_raw) %>%
                twListToDF() %>%
                select(screenName, text, favoriteCount, retweetCount) %>%
                as_tibble()
```

The data cleaning process included removing all retweets, converting the tweets into a data frame, selecting the four variables of interest (screen name, text of the tweet, the number of favorites, and the number of retweets), and finally converting the data frame to a tibble.

Analysis

```
# Is there a correlation between length of tweet and number of retweets?
length_retweet <- cor.test(str_length(tweets_tbl$text), tweets_tbl$retweetCount)
length_retweet
```

```
##
##  Pearson's product-moment correlation
##
## data:  str_length(tweets_tbl$text) and tweets_tbl$retweetCount
## t = 2.0759, df = 106, p-value = 0.04033
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.00901092 0.37270170
## sample estimates:
##       cor
## 0.1976483
```

```
# Is there a correlation between length of tweet and number of favorites?
length_favorite <- cor.test(str_length(tweets_tbl$text), tweets_tbl$favoriteCount)
length_favorite
```

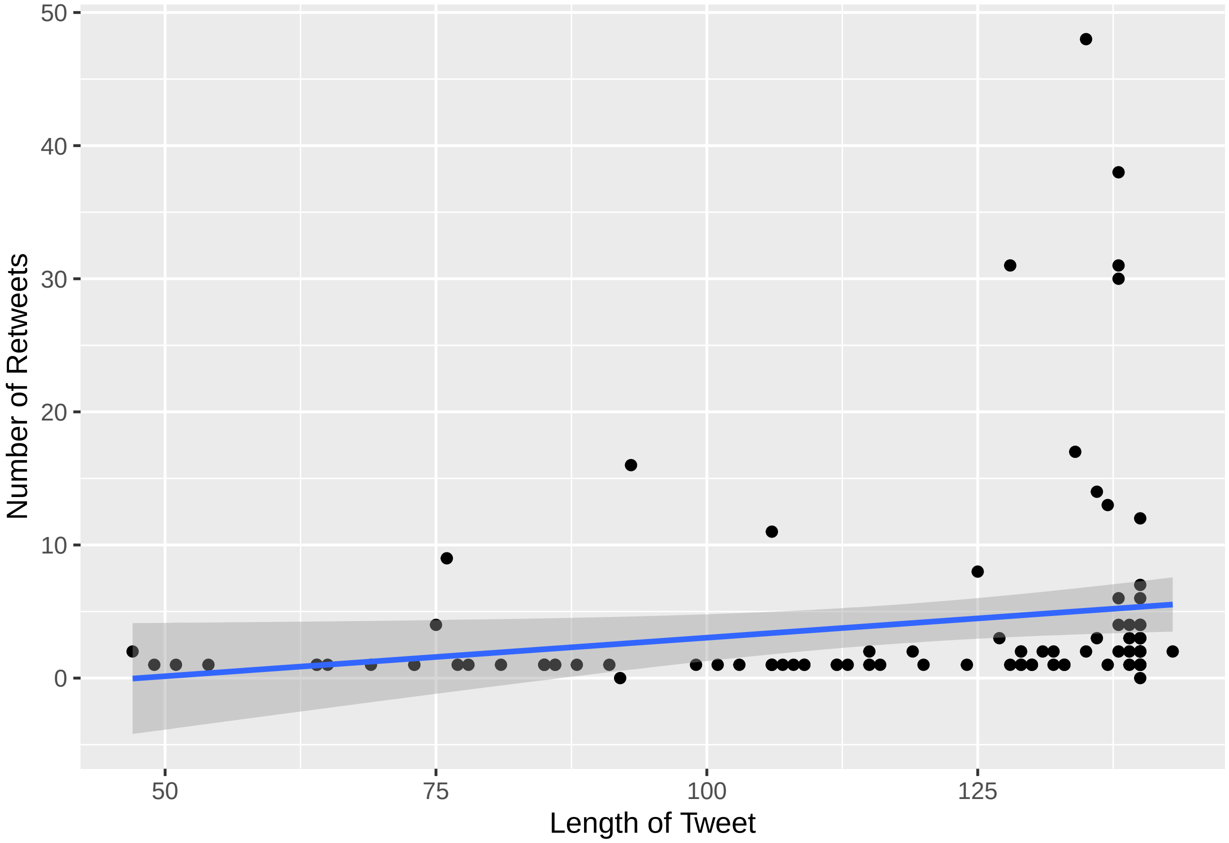
```
##
##  Pearson's product-moment correlation
##
## data:  str_length(tweets_tbl$text) and tweets_tbl$favoriteCount
## t = 2.466, df = 106, p-value = 0.01527
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04597868 0.40411460
## sample estimates:
##       cor
## 0.2329288
```

The correlation between length of tweet and the number of retweets was was 0.198 (p=0.04), which is statistically significant.

The correlation between length of tweet and the number of favorites was was 0.233 (p=0.015), which is statistically significant.

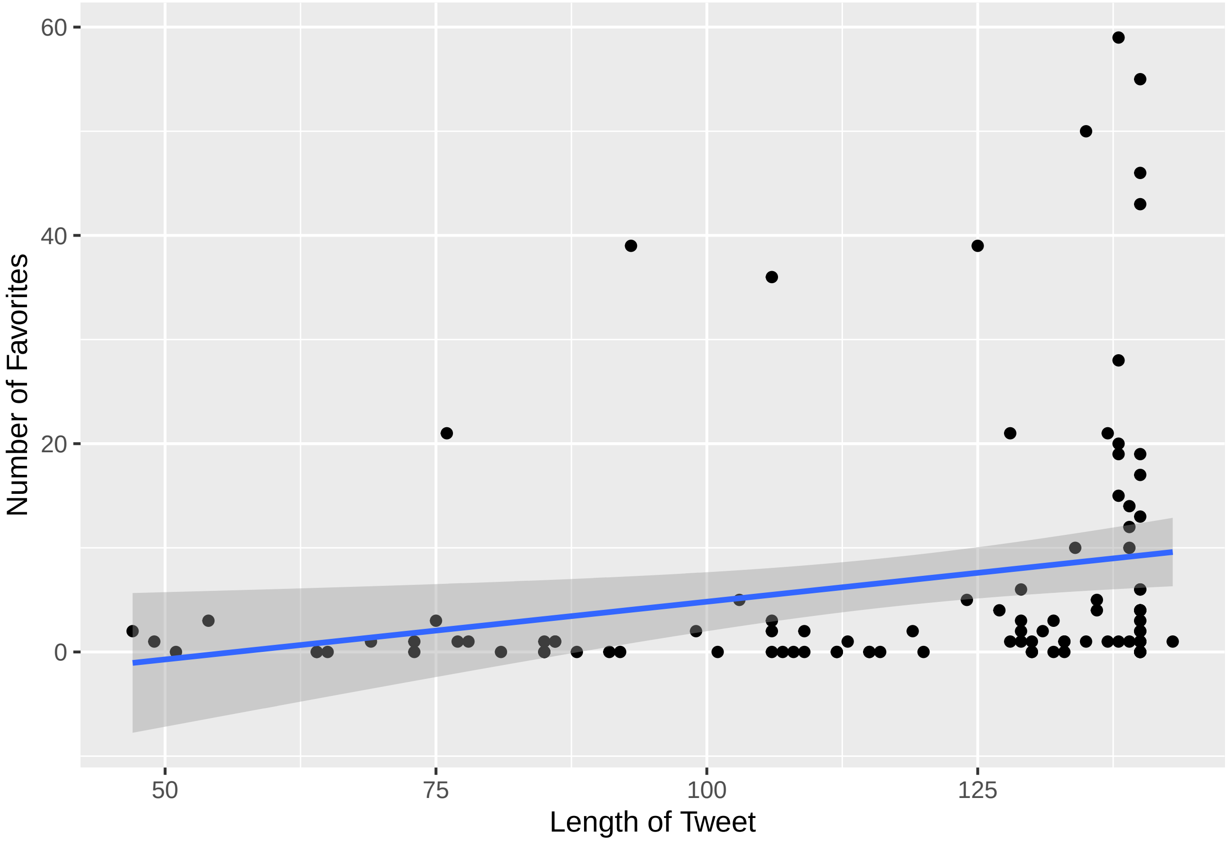
Visualization

```
# plot of length of tweet and number of retweets
ggplot(tweets_tbl, aes(x = str_length(tweets_tbl$text), y = tweets_tbl$retweetCount)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Length of Tweet", y = "Number of Retweets")
```



The plot shows the relationship between the length of the tweet and the number of retweets. Each point represents an individual tweet, and the line is the OLS regression line with a 95% confidence band.

```r
# plot of length of tweet and number of favorites
ggplot(tweets_tbl, aes(x = str_length(tweets_tbl$text), y = tweets_tbl$favoriteCount)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Length of Tweet", y = "Number of Favorites")
```



The plot shows the relationship between the length of the tweet and the number of favorites. Each point represents an individual tweet, and the line is the OLS regression line with a 95% confidence band.