Milestone Report
By Kavin Sampath

Introduction

The goal of this project is just to display that you've gotten used to working with the data and that you are on track to create your prediction algorithm. Please submit a report on R Pubs (http://rpubs.com/) that explains your exploratory analysis and your goals for the eventual app and algorithm.

1. Demonstrate that you've downloaded the data and have successfully loaded it in.
2. Create a basic report of summary statistics about the data sets.
3. Report any interesting findings that you amassed so far.
4. Get feedback on your plans for creating a prediction algorithm and Shiny app.

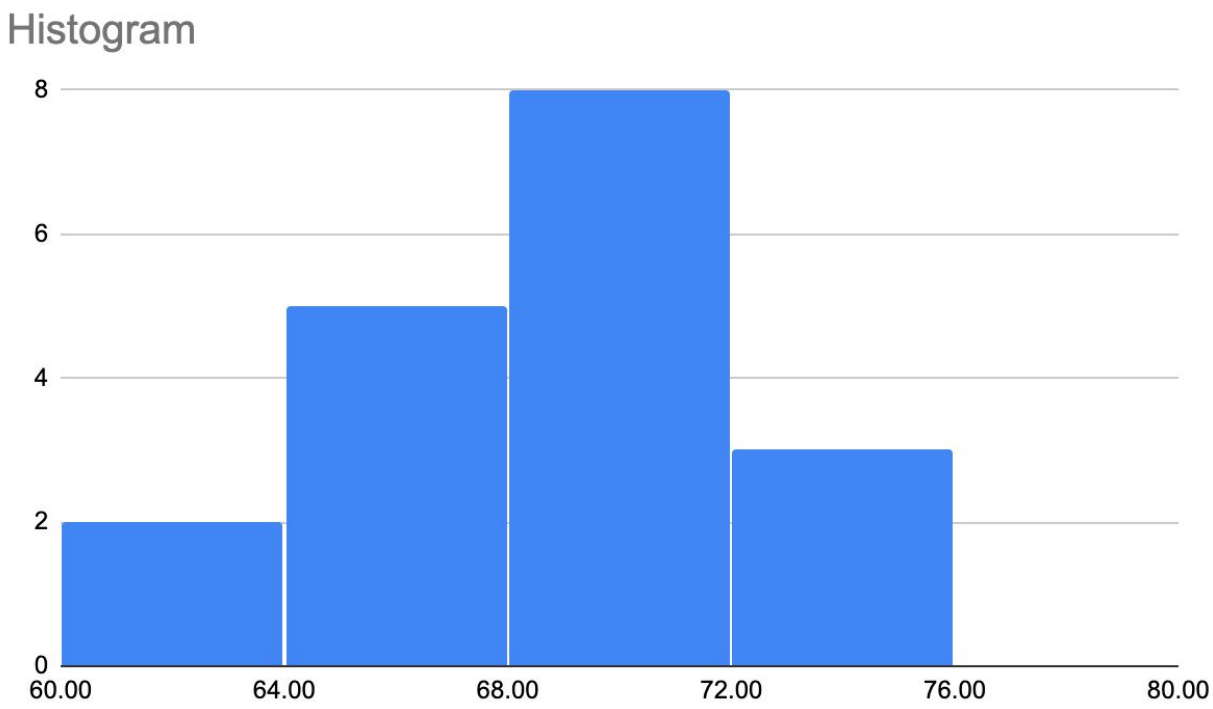
Creating a summary report

```
file.list = c("final/en_US/en_US.blogs.txt", "final/en_US/en_US.news.txt",
"final/en_US/en_US.twitter.txt")

text <- list(blogs = "", news = "", twitter = "")


data.summary <- matrix(0, nrow = 3, ncol = 3, dimnames = list(c("blogs", "news",
"twitter"),c("file size, Mb", "lines", "words")))

for (i in 1:3) {

  con <- file(file.list[i], "rb")

  text[[i]] <- readLines(con, encoding = "UTF-8",skipNul = TRUE)

  close(con)

  data.summary[i,1] <- round(file.info(file.list[i])$size / 1024^2, 2)

  data.summary[i,2] <- length(text[[i]])

  data.summary[i,3] <- sum(stri_count_words(text[[i]]))

}

library(knitr)
```

kable(data.summary)

Data filing and data cleaning

```
NotKnown <- grep("NotKnown", iconv(data_sample, "latin1", "ASCII", sub="NotKnown"))
# doing some simple cleaning
data_sample <- gsub("&amp", "", data_sample)
data_sample <- gsub("RT :|@[a-z,A-Z]*: ", "", data_sample) # remove tweets
data_sample <- gsub("@\\w+", "", data_sample)
```



Conclusion

I also find it interesting to know that everyone has different height words in the data. How many words do I have to know to cover half of the text? How many people do I need to have a good sample size? I can combine the data in one set and analyse how many unique words are used and how frequently a certain height occurs.