Final Project(Group Activity)

Kent Hervey D. Gener BSIT2-B

2023-12-22

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

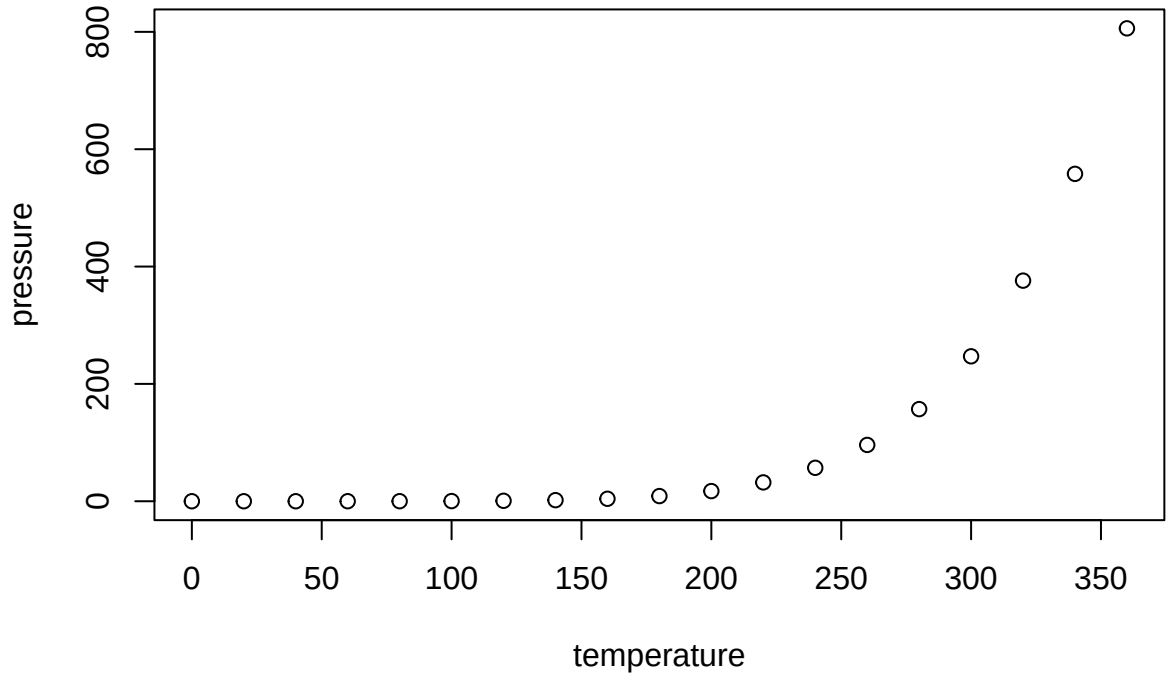
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Select a website that you want to scrape 300 reviews. You can select 1 product from Amazon Or you can select 1 movie Or you can select reviews from SkyTrax https://www.airlinequality.com/review-pages/a-z-airline-reviews/ ; the reviews are for the different airlines, but you can only select 1 airline

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
scrape <- function(url) {
  page <- read_html(url)

  user <- page %>% html_nodes(".display-name-link") %>% html_text()
  date <- page %>% html_nodes(".review-date") %>% html_text()
  rating <- page %>% html_nodes(".rating-other-user-rating") %>% html_text()
  title <- page %>% html_nodes(".title") %>% html_text()
  comment <- page %>% html_nodes(".text.show-more__control") %>% html_text()

  reviews <- data.frame(User = user, Date = date, Rating = rating, Title = title, Comment = comment)
  return(reviews)
}

scrape_pages <- function(base_url) {
  all_reviews <- data.frame()
  reviews_per_page <- 10
  total_reviews_target <- 300
  num_pages <- ceiling(total_reviews_target / reviews_per_page)

  for (page_num in 1:num_pages) {
    url <- paste0(base_url, "&start=", (page_num - 1) * reviews_per_page)
    reviews <- scrape(url)
    all_reviews <- bind_rows(all_reviews, reviews)

    if (nrow(all_reviews) >= total_reviews_target) {
      break
    }
  }

  return(all_reviews)
}

base_url <- "https://www.imdb.com/title/tt6166392/reviews/?ref_=tt_ql_2"
reviews_data <- scrape_pages(base_url)

reviews_data <- head(reviews_data, 300)

write.csv(reviews_data, "imdb_reviews.csv", row.names = FALSE)
```

After scraping 300 reviews, have a basic sentiment analysis with your own analysis and data visualization. each visualization shall have its own description.

```r
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.3.2
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

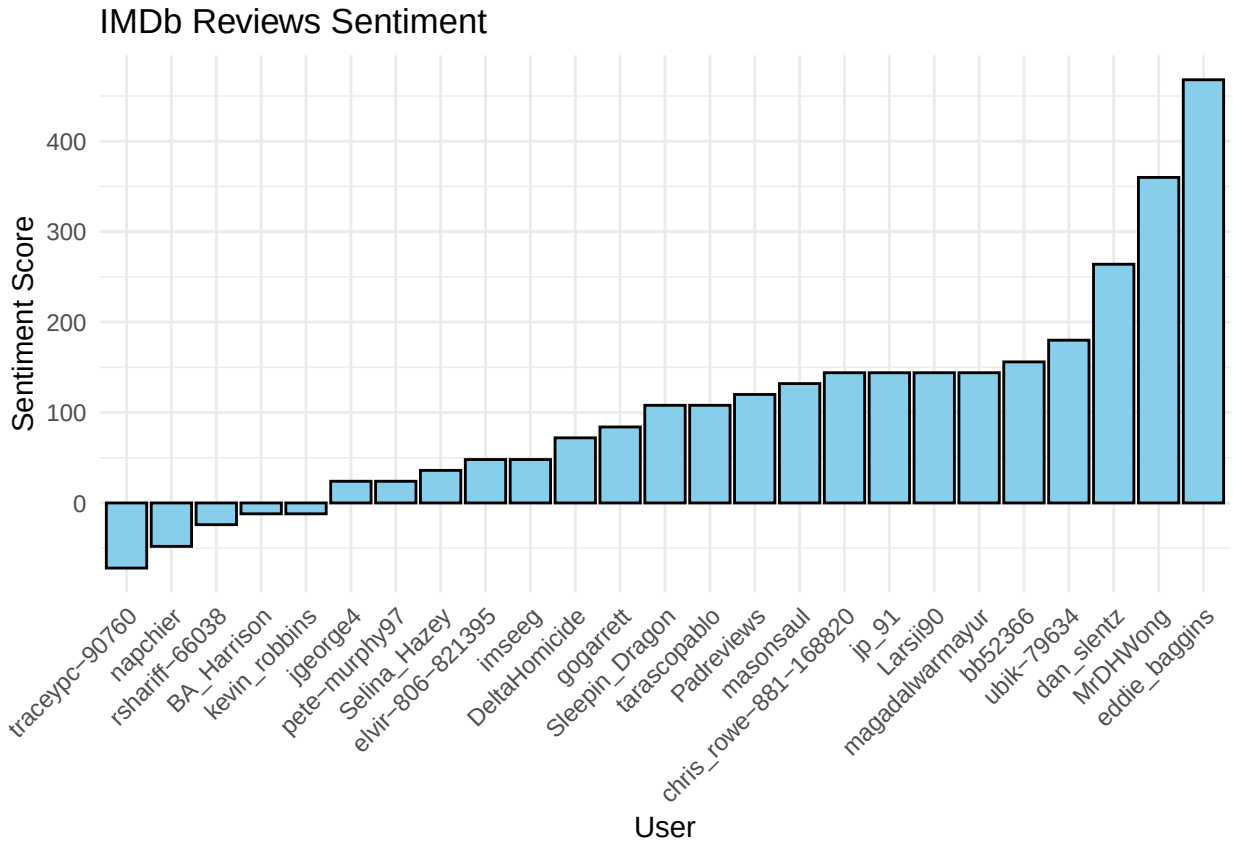
```r
library(dplyr)

reviews_data <- read.csv("imdb_reviews.csv")
reviews_tokens <- reviews_data %>%
  unnest_tokens(word, Comment)

bing <- get_sentiments("bing")

sentiment_data <- reviews_tokens %>%
  inner_join(bing, by = "word") %>%
  group_by(User) %>%
  summarize(SentimentScore = sum(sentiment == "positive") - sum(sentiment == "negative"))

ggplot(sentiment_data, aes(x = reorder(User, SentimentScore), y = SentimentScore)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "IMDb Reviews Sentiment",
       x = "User",
       y = "Sentiment Score") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
ggplot(sentiment_data, aes(x = "", y = SentimentScore, fill = factor(SentimentScore > 0))) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  labs(title = "Sentiment Distribution",
```

```
      x = NULL,
      y = NULL) +
theme_minimal()
```