NASDAQ Scraper & Analysis

Shahzad-Ali & Hassaan Ahmed

12/3/2019

Introduction

The project aims at downloading the NASDAQ website which includes data for approximately 6000+ companies world wide listed on the stock market. The scraped data incldues the company names, sector, market capital and some sentiments regarding the decision to buy, sell or hold each company's shares. After the data scraping, we have done some sentiment analysis as well to decide the top 10 winners in each sector and a final decision with score to hold, buy or sell stocks for all the companies.

NASDAQ Scraper Code

The code below was used to scrape the NASDAQ data

```
library(rvest)
library(jsonlite)
library(data.table)
library(tidyverse)
library(pbapply)
library(dplyr)
library(purrr)

all_urls <- list()

for (i in 1:304) {
    one_url <- paste0("https://www.nasdaq.com/api/v1/screener?page=", i, "&pageSize=20")
    all_urls <- rbind(all_urls, one_url)
}


one_page_data <- function(x) {

    jdata <- fromJSON(x, flatten = TRUE)

    jdata$data$priceChartSevenDay <- NULL
    jdata$data$articles <- NULL
    jdata$count <- NULL

    Ticker <- jdata$data$ticker
    Company <- jdata$data$company
    Market_capital <- jdata$data$marketCap
    Market_share <- jdata$data$marketCapGroup
    Sector <- jdata$data$sectorName
    Analyst_advice <- jdata$data$analystConsensusLabel
    News_Sentiment <- jdata$data$newsSentimentData.signal
    News_Score <- jdata$data$newsSentimentData.score
    Media_Buzz <- jdata$data$mediaBuzzData.signal
    Media_Score <- jdata$data$mediaBuzzData.score
    HedgeFund_Sentiment <- jdata$data$hedgeFundSentimentData.signal
    HedgeFund_Score <- jdata$data$hedgeFundSentimentData.score
```

```
    Investor_Sentiment <- jdata$data$investorSentimentData.signal
    Investor_Score <- jdata$data$investorSentimentData.score

    one_nasdaq <- data.frame(Ticker, Company, Market_capital, Market_share, Sector, Analyst_advice,
                             News_Sentiment, News_Score, Media_Buzz, Media_Score, HedgeFund_Sentiment,
                             HedgeFund_Score, Investor_Sentiment, Investor_Score)

    return(one_nasdaq)
}


nasdaq_data <- rbindlist(pblapply(all_urls, one_page_data))
```

Here's how the downlaoded data looks like:

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
nasdaq_data <- read_csv("nasdaq.csv")
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Ticker = col_character(),
##   Company = col_character(),
##   Market_capital = col_double(),
##   Market_share = col_character(),
##   Sector = col_character(),
##   Analyst_advice = col_character(),
##   News_Sentiment = col_character(),
##   News_Score = col_double(),
##   Media_Buzz = col_character(),
##   Media_Score = col_double(),
##   HedgeFund_Sentiment = col_character(),
##   HedgeFund_Score = col_double(),
##   Investor_Sentiment = col_character(),
##   Investor_Score = col_double()
## )
```

```
nasdaq_data
```

```
## # A tibble: 6,080 x 15
##       X1 Ticker Company Market_capital Market_share Sector Analyst_advice
##    <dbl> <chr>  <chr>            <dbl> <chr>        <chr>  <chr>
##  1     1 AAPL   Apple    1187463907500 Mega         Consu~ Moderate Buy
##  2     2 MSFT   Micros~  1154849257800 Mega         Techn~ Strong Buy
```

```
##  3     3 AMZN   Amazon    892831237600 Mega         Servi~ Strong Buy
##  4     4 GOOGL  Alphab~   838763302100 Mega         Techn~ Strong Buy
##  5     5 FB     Facebo~   575172050800 Mega         Techn~ Strong Buy
##  6     6 BRK.A  Berksh~   538799567900 Mega         Finan~ Moderate Buy
##  7     7 BRK.B  Berksh~   538726026000 Mega         Finan~ Moderate Buy
##  8     8 BABA   Alibaba   523574000000 Mega         Servi~ Strong Buy
##  9     9 JPM    JPMorg~   413263922400 Mega         Finan~ Moderate Buy
## 10    10 TCEHY  Tencen~   399412135500 Mega         Techn~ Strong Buy
## # ... with 6,070 more rows, and 8 more variables: News_Sentiment <chr>,
## #   News_Score <dbl>, Media_Buzz <chr>, Media_Score <dbl>,
## #   HedgeFund_Sentiment <chr>, HedgeFund_Score <dbl>,
## #   Investor_Sentiment <chr>, Investor_Score <dbl>
```

Analysis

Companies per Sector:

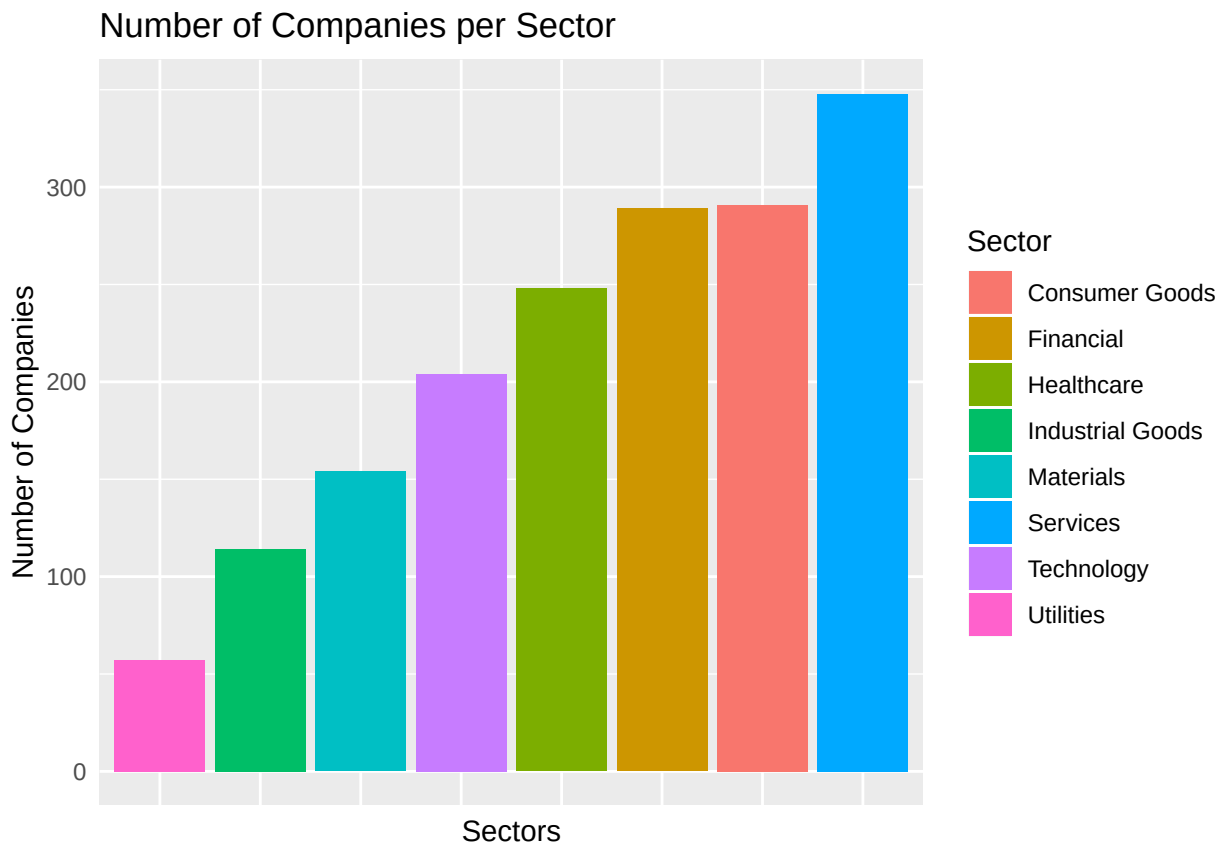
To start off, we first see the number of companies in each Sector

```
companies_sector <- na.omit(nasdaq_data) %>%
    group_by(Sector) %>%
    summarise('No_of_Companies' = n())

companies_sector
```

```
## # A tibble: 8 x 2
##   Sector           No_of_Companies
##   <chr>                      <int>
## 1 Consumer Goods               291
## 2 Financial                    289
## 3 Healthcare                   248
## 4 Industrial Goods             114
## 5 Materials                    154
## 6 Services                     348
## 7 Technology                   204
## 8 Utilities                     57
```

```
# To view this data as a nice plot
ggplot(data = na.omit(nasdaq_data)) +
    geom_bar(mapping = aes(x = Sector, fill = Sector)) +
    scale_x_discrete(limits=c("Utilities", "Industrial Goods", "Materials", "Technology",
                              "Healthcare", "Financial", "Consumer Goods", "Services")) +
    theme(axis.text.x = element_blank(), axis.ticks = element_blank()) +
    labs(x = "Sectors",
         y = "Number of Companies",
         title = "Number of Companies per Sector")
```

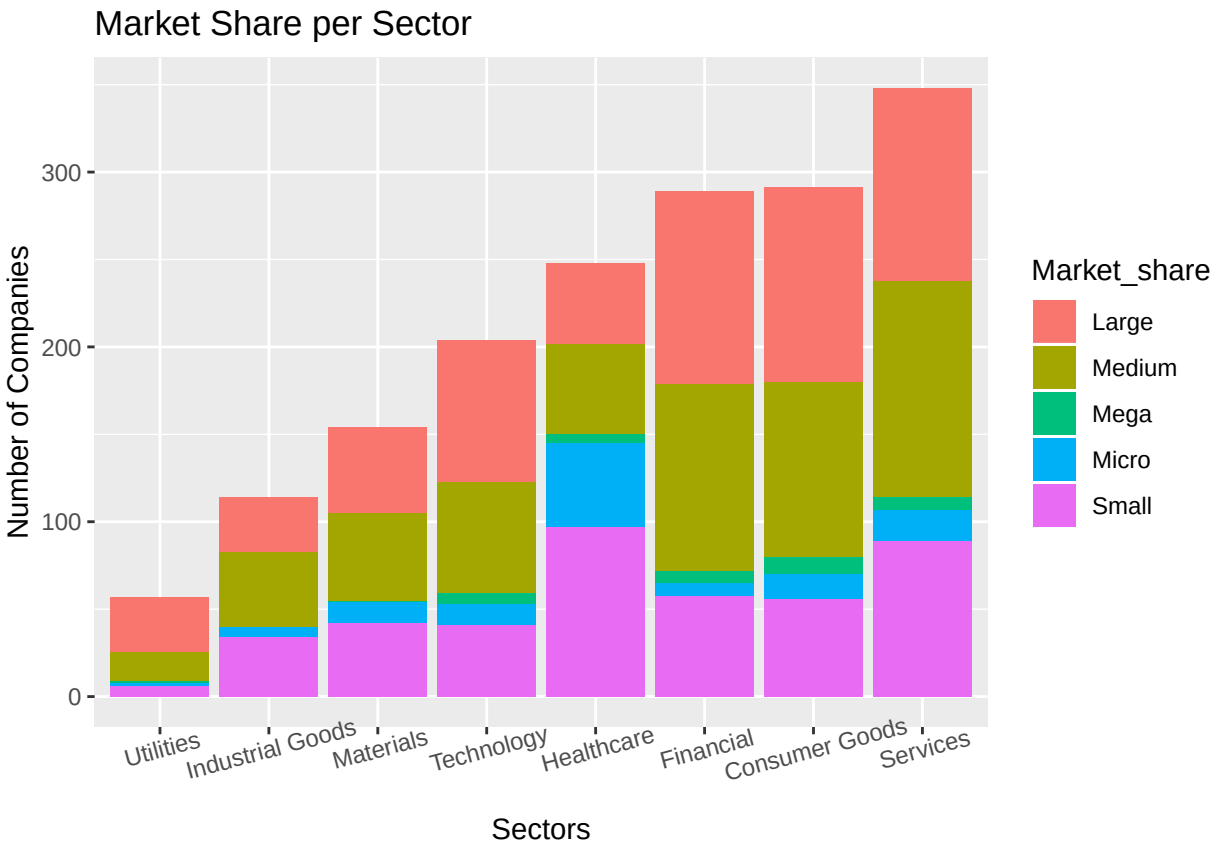


It is evident from the graph that the Services Sector has the most number of companies listed on the stock market followed by Consumer Goods and Financial Sector.

Market Share per Sector with count of Companies:

Next we see the market share for each of the sectors in the NASDAQ data

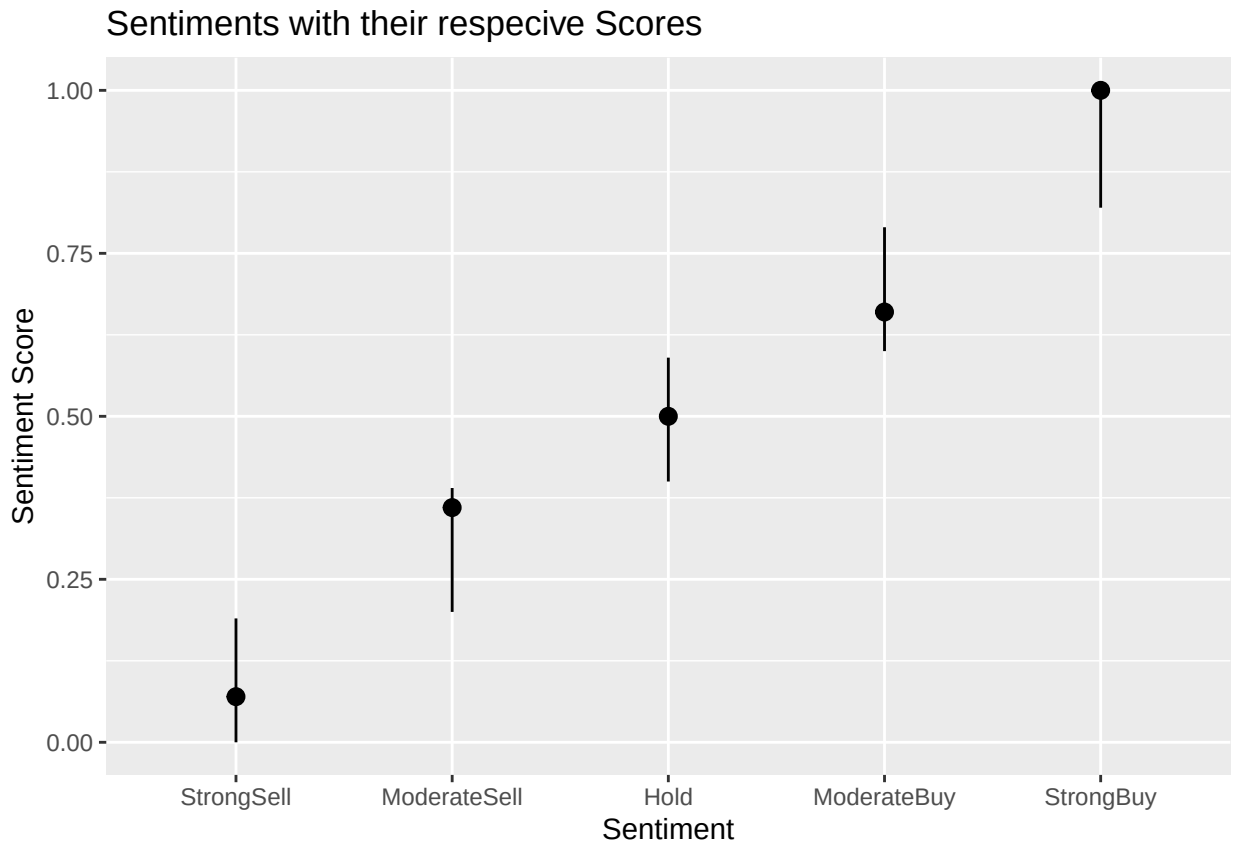
```
ggplot(data = na.omit(nasdaq_data)) +
    geom_bar(mapping = aes(x = Sector, fill = Market_share)) +
    scale_x_discrete(limits=c("Utilities", "Industrial Goods", "Materials", "Technology",
                              "Healthcare", "Financial", "Consumer Goods", "Services")) +
    theme(axis.text.x = element_text(angle = 15)) +
    labs(x = "Sectors",
         y = "Number of Companies",
         title = "Market Share per Sector")
```



Sentiment Statistics:

As discussed earlier, each company has sentiments attached to it from various sources such as News, Media, Analysts and Hedgefunds. All these sources give a sentiment for each company in terms of Strong sell, Moderate Sell, Hold, Moderate Buy and Strong Buy and along with it a score for the sentiment given. Below is a plot to see the range of scores attached to each sentiment given to get an overall picture of how the scoring criteria goes

```
ggplot(data = na.omit(nasdaq_data)) +
    stat_summary(
    mapping = aes(x = reorder(Investor_Sentiment, Investor_Score), y = Investor_Score),
    fun.ymin = min,
    fun.ymax = max,
    fun.y = median
  ) + labs(x = "Sentiment",
           y = "Sentiment Score",
           title = "Sentiments with their respecive Scores")
```



So we see the scores range from 0 to 1 indicating Strong Sell to Strong Buy respectively.

sentiment Analysis for Companies:

Our aim here is to come up with a unified sentiment score for each company and accordingly assign a final decision to either Buy, Hold or Sell the company's stocks for each company

How we calculate the final score for each company is by taking the mean of all the differnent scores from the various sources listed in the dataset. From the graph above we see that the scores for each sentiment increase very linearly therefore the cut off scores for each decision range linearly from 0-1 with 5 steps of 0.2 points each.

```
Company_sentiments <- nasdaq_data %>%
    mutate(mean_Score= rowMeans(data.frame(News_Score, Media_Score, HedgeFund_Score, Investor_Score)))

Company_sentiments$Decision <- cut(as.numeric(Company_sentiments$mean_Score), 5,
                        labels = c("Strong_sell", "Moderate_sell", "Hold", "Moderate_buy", "strong_buy"

Companies_Decision <- Company_sentiments %>%
    group_by(Sector, Company, Decision) %>%
    select(Company, mean_Score, Decision)
```

```
## Adding missing grouping variables: `Sector`
```

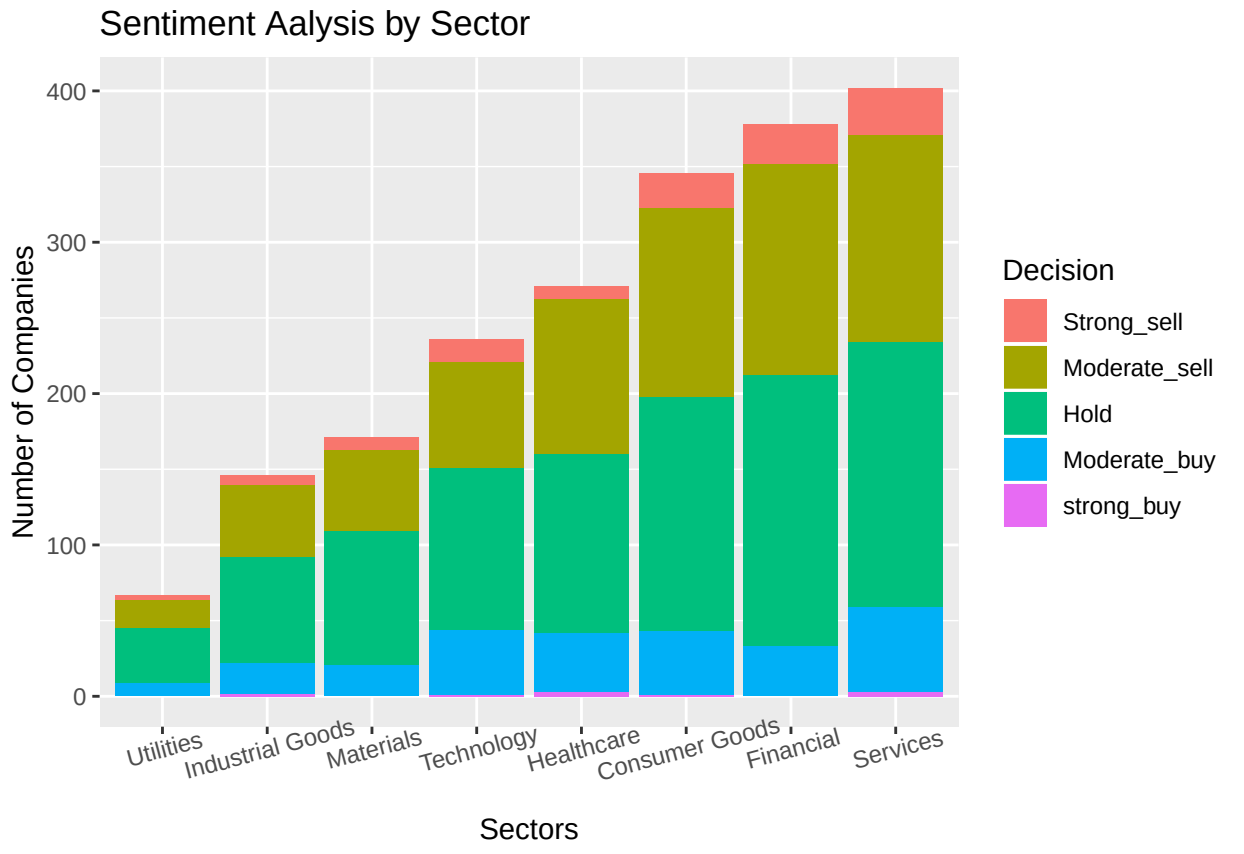
```
Companies_Decision
```

```
## # A tibble: 6,080 x 4
## # Groups:   Sector, Company, Decision [6,022]
##    Sector         Company              mean_Score Decision
```

```
##    <chr>          <chr>                  <dbl> <fct>
##  1 Consumer Goods Apple                  0.605 Hold
##  2 Technology     Microsoft              0.561 Hold
##  3 Services       Amazon                 0.797 strong_buy
##  4 Technology     Alphabet               0.565 Hold
##  5 Technology     Facebook               0.628 Moderate_buy
##  6 Financial      Berkshire Hathaway A   0.419 Moderate_sell
##  7 Financial      Berkshire Hathaway B   0.677 Moderate_buy
##  8 Services       Alibaba                0.798 strong_buy
##  9 Financial      JPMorgan Chase & Co.   0.485 Hold
## 10 Technology     Tencent Holdings       0.698 Moderate_buy
## # ... with 6,070 more rows
```

For each listed company, we now have a mean score and a final decision for stock trading. A mean score of zero indicates that now is the perfect time to sell the shares and a score of 1 favors the buying of more stocks of a company. To finally summarise our findings in a single graph, we see the different decisions per sector for each company to see which Sectors have the best opportunity to buy, hold or sell stocks

```
ggplot(data = na.omit(Companies_Decision)) +
    geom_bar(mapping = aes(x = Sector, fill = Decision)) +
    scale_x_discrete(limits=c("Utilities", "Industrial Goods", "Materials", "Technology",
                              "Healthcare", "Consumer Goods", "Financial", "Services")) +
    theme(axis.text.x = element_text(angle = 15)) +
    labs(x = "Sectors",
         y = "Number of Companies",
         title = "Sentiment Aalysis by Sector")
```



We can conclude that the best selling opportunites are for the Services, Financial and the Consumer Goods Sector. The dataset indicates that very few companies currently have good buying opportunities and overall holding the stocks for almost all sectors is most advisable.

The final submission contians a Companies_Decision.csv file for the complete list of decisions for each company.