

**ISSS616 Applied Statistical Analysis with R – G3**

**Final Report:**

**Glassdoor Rating and Reviews**

**Submitted by**

|  |  |
| --- | --- |
| Joey Chua | joey.chua.2020@mitb.smu.edu.sg |
| Fang Dong | fang.dong.2020@mitb.smu.edu.sg |
| Hongyi Li | hongyili.2020@mitb.smu.edu.sg |
| Di Lu | dilu.2020@mitb.smu.edu.sg |
| Yuchen Yao | yuchen.yao.2020@mitb.smu.edu.sg |
| Weimin Zhang | wmzhang.2020@mitb.smu.edu.sg |

**15 November 2020**

**Content**

[1. Abstract 3](#_Toc56353340)

[2. Overall Concept 3](#_Toc56353341)

[3. Data 4](#_Toc56353342)

[3.1 Kaggle 4](#_Toc56353343)

[3.1.1 Data Cleaning 5](#_Toc56353344)

[3.1.2 Data Preview: 5](#_Toc56353345)

[3.2 Web Scraping 5](#_Toc56353346)

[3.2.1 Data Cleaning 6](#_Toc56353347)

[3.2.2 Data Preview: 7](#_Toc56353348)

[4. Packages 7](#_Toc56353349)

[5. Exploratory Data Analysis 7](#_Toc56353350)

[5.1 Glassdoor Rankings 7](#_Toc56353351)

[5.2 Ratings Trend 8](#_Toc56353352)

[5.3 Distribution 9](#_Toc56353353)

[6. Ratings 10](#_Toc56353354)

[6.1 Correlation 10](#_Toc56353355)

[6.2 Multiple Linear Regression 10](#_Toc56353356)

[6.2.1 Confirmatory Data Analysis 13](#_Toc56353357)

[6.3 Predictions 13](#_Toc56353358)

[7. Review 15](#_Toc56353359)

[7.1 Word Cloud 15](#_Toc56353360)

[7.2 AFINN Lexicon 16](#_Toc56353361)

[7.3 Summary of Review Findings 16](#_Toc56353362)

[Word Cloud and Sentiment Reviews from *Annex D* are summarized: 16](#_Toc56353363)

[7.4 Linear Regression 17](#_Toc56353364)

[8. Conclusion 19](#_Toc56353365)

# 1. Abstract

Glassdoor is one of the world’s largest job and recruiting sites. As of 2020, the number of companies on Glassdoor reached 1 million, with their ratings and reviews totalling 55 million[[1]](#footnote-1). These ratings and reviews provide valuable insights such as jobs, salaries, and insights into a company’s culture. Additionally, with 12 million job listings and 64 million monthly users, employers can recruit and hire candidates which are of good fit. A summary of how Glassdoor works is shown in *Annex A: Figure 1*. The insights on Glassdoor shed light on an organisation’s culture, which is helpful in attracting talents and driving employees’ satisfaction. This in turn can be used to predict a range of corporate outcomes, including profitability, innovation, as well as financial fraud. In this report, we will explore the trends of FANMAG companies’ ratings over the year and investigate the factors that contribute to these ratings. Additionally, we will conduct sentiment analysis to understand the emotional intent of words based on Glassdoor reviews, and observe how they affect overall rating.

# 2. Overall Concept

**Problems 1:** Glassdoor final aggregated ratings for each company is based on Glassdoor’s proprietary awards algorithm, and each employer’s rating is evaluated based on the quantity, quality, and consistency of reviews. Additionally, heavier weightage is placed on more recent ratings. However, the aggregated rating does not provide insights into company reviews such as work-life balance and career opportunities.

**Problem 2:** Glassdoor’s API is not publicly available, and they do not provide reviews that are of valuable insights into a company’s culture. For this project, we requested for the API to access the data but have not received approval as it is based on a case-by-case evaluation.

**Literature Review:** Currently there are far and few between comprehensive research on Glassdoor ratings and reviews. SoftwareAdvice’s article on “How Job Seekers Use Glassdoor Reviews”[[2]](#footnote-2) sheds light on how 50% of respondents use Glassdoor when Job Hunting. They also identified the most positive and negative predictors that job hunters look for when applying to jobs. However, it is unclear which industries these respondents are in, and there is a lack of statistical analysis. Another article by MITSloan[[3]](#footnote-3) used machine learning and human expertise to analyse culture on Glassdoor. While they identified that Culture 500 industries have IT services making up majority of the companies *(Annex A: Figure 2)*, with an attempt to understand corporate culture, most part of the article explained the spikes in Glassdoor rating and not the predictors of Glassdoor ratings.

**Purpose and Target Audience:** This report aims to solve the above problems and limitations, and to provide a comprehensive insight on trends, factors affecting ratings, and comprehend nuanced emotions of reviews. With this information, **individuals** can position themselves to be an informed candidate in their job search. Similarly, **companies** can better reflect based on ratings and reviews on Glassdoor. Following, concerted effort to improve the culture of companies can increase the productivity of employees, and to attract talents.

**FANMAG:** Inparticular, we will explore **Facebook, Apple, Netflix, Microsoft, Amazon, Google**. According to Research Affiliates, Over US$4.0T in combined market cap places these six companies above all but 2 of 61 countries in the Morningstar Global Markets Index. With their increasing popularity as tech companies, and rising stock market performance (*Annex A: Figure 3*), we will focus on FANMAG companies.

**Year 2015 – 2018:** The timeline for this report is from year 2015 to 2018. While the source provides data backdated to 2008, given the rapidly changing environment, we will look at the most recent 4 years.

**Overview:** The following table shows an overview of the approach in this report. Data is explained in Section 3, EDA conducted in Section 4, and CDA in section 5.

|  |  |  |
| --- | --- | --- |
| Overview | | |
| Data | Kaggle | Web scraping |
| Trend | Top companies to work for based on glassdoor;  Ranking over the years for FANMAG;  Ratings over the years for FANMAG | |
| Exploratory  Data Analysis | Overall Ratings  Work Balance Rating  Culture  Company Benefit  Management  Opportunities | Word Cloud  Sentiment Analysis (AFINN, Bing)  Within Industry Comparison  Between Industries Comparison |
| Confirmatory Data Analysis | 3x3 Chi Square:  Predictors of Overall Rating  Predictors of Level of Ratings  Levels of Ratings | Multinomial Regression:  Reviews affect Overall Rating |
| Conclusion | | |

# 3. Data

## 3.1 Kaggle

Dataset can be downloaded from Kaggle. The dataset contains 67k reviews and ratings for FANMAG from the year 2008 to 2018. Variables are shown in the following table:

### 3.1.1 Data Cleaning

* Files are imported using read.csv() function
* Variables are selected (overall.ratings, work.balance.stars, culture.values.stars, career.opporunitiy. stats, comp.benefit.stars, senior.management.stars)
* Data types for variables are changed to numeric
* Date is changed to mdy format
* Date from 2015 to 2018 are obtained using filter() function.

****

### 3.1.2 Data Preview:

|  |  |  |
| --- | --- | --- |
| No. | Variable | Description |
| 1 | Company | Name of company that adheres to Glassdoor’s guidelines for employers’ profiles. They typically do not include business extensions (i.e. Inc, LLC, Pvt, Ltd, etc.). |
| 2 | Date | MM DD YYYY format |
| 3 | Overall Rating | 1 - 5 |
| 4 | Work/Life Balance Rating | 1 - 5 |
| 5 | Culture and Values Rating | 1 - 5 |
| 6 | Career Opportunity Rating | 1 - 5 |
| 7 | Comp and Benefits Rating | 1 - 5 |
| 8 | Senior Management Rating | 1 - 5 |

## 3.2 Web Scraping

Date and Reviews are obtained via web scrapping and saved as CSV files.



**Landing Page** is <https://www.glassdoor.sg/Reviews/Google-Reviews-E9079.htm>;

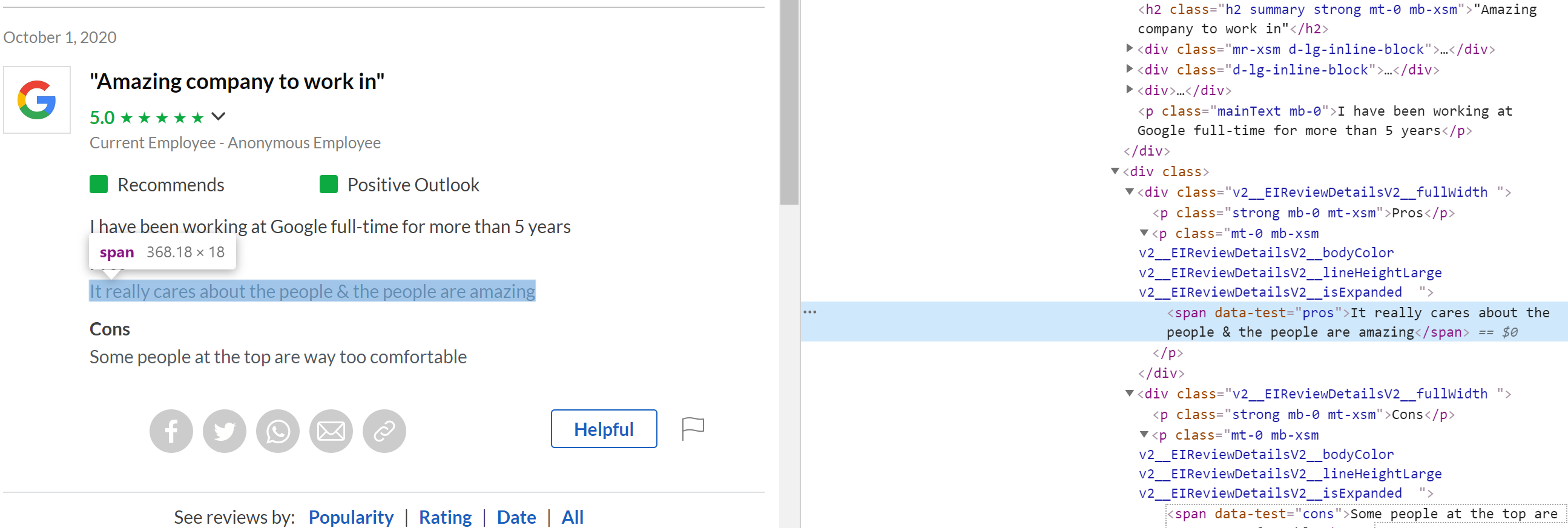
**subsequent pages** are <https://www.glassdoor.sg/Reviews/Google-Reviews-E9079_P2.htm>; “\_P2” is added towards the endpoint.

Inspection of the elements through Hyper Text Markup Language (HTML) and Cascading Style Sheets (CSS), class selectors (Figure X) are as follow:

**Date**: <div class="d-flex align-items-center"><time class="date subtle small" datetime="Thu Oct 01 2020 07:00:23 GMT-0700 (Pacific Daylight Time)">October 1, 2020</time></div>

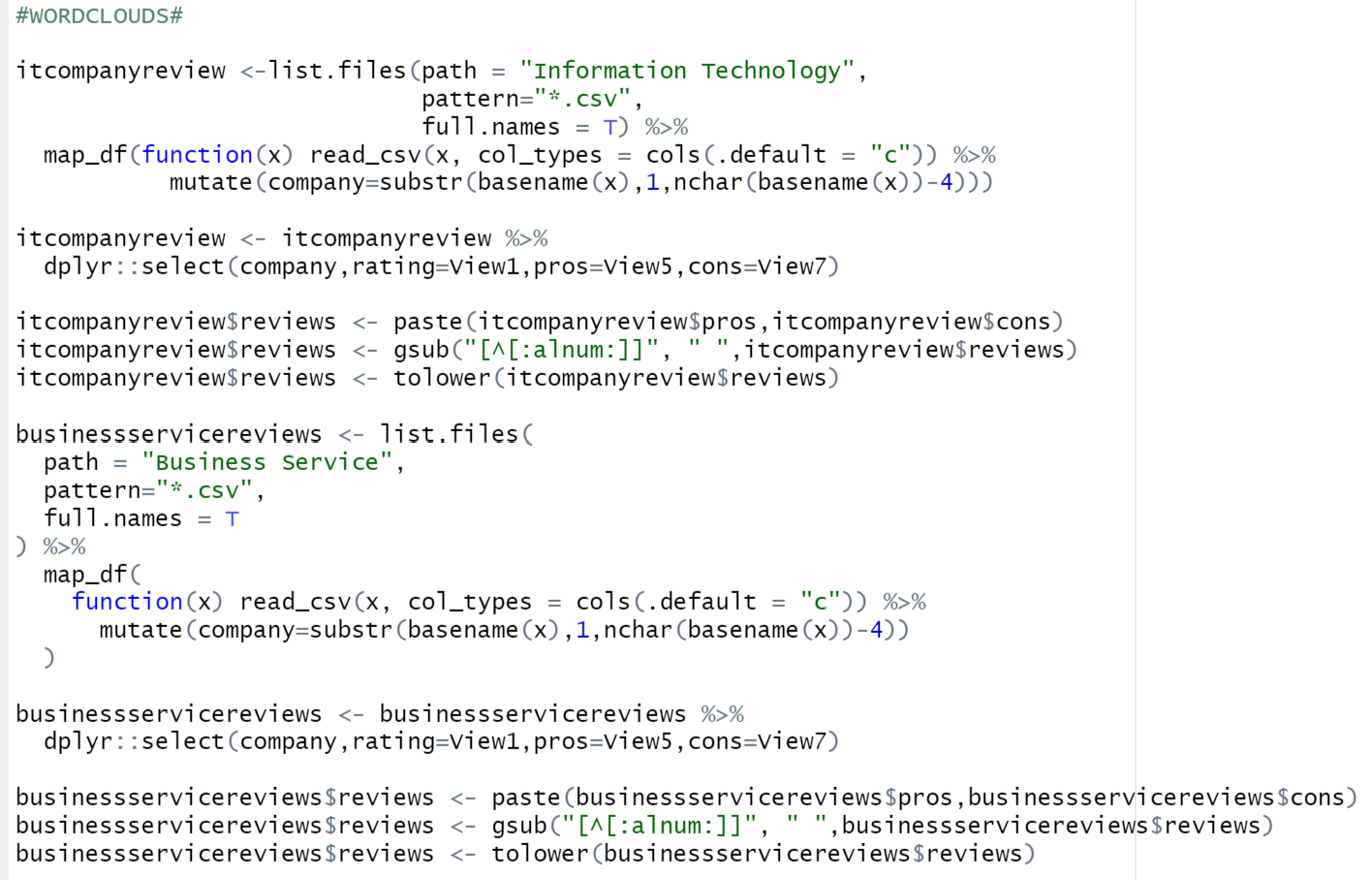
**Pros**: <span data-test="pros">It really cares about the people &amp; the people are amazing</span>

**Cons**: <span data-test="cons">Some people at the top are way too comfortable</span>

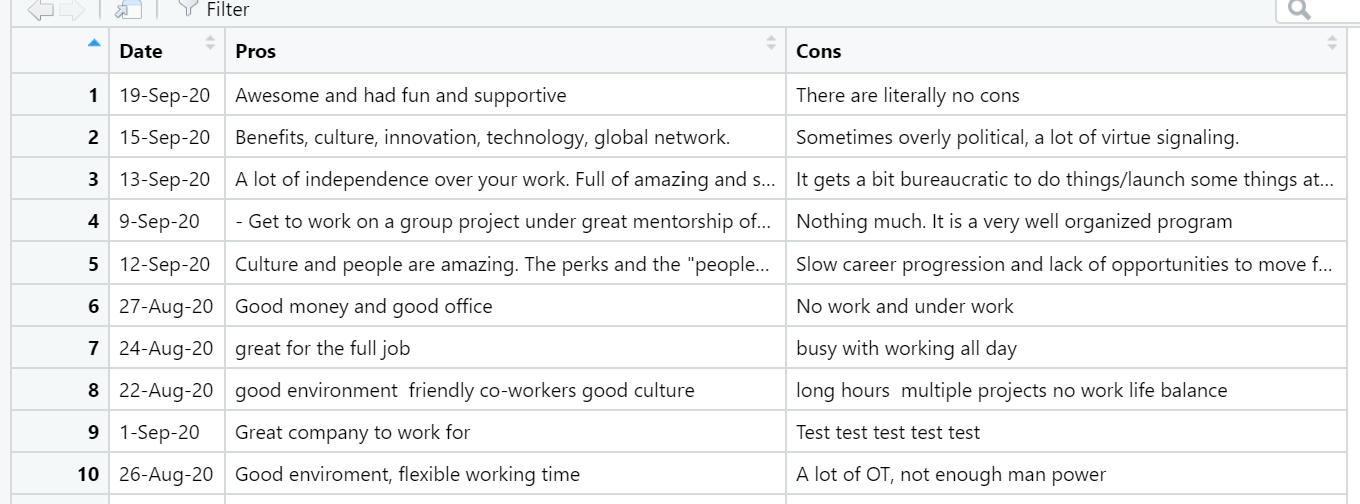


### 3.2.1 Data Cleaning

* Function is created to import multiple CSV files
* New column “company” is created using file’s name
* Columns of interest are selected using select() function
* Anything but alphabets and numerical numbers are removed using REGEX in reviews

****

### 3.2.2 Data Preview:



# 4. Packages

Packages are listed in Annex B. R package, Shiny is used to build interactive web pages.

# 5. Exploratory Data Analysis

## 5.1 Glassdoor Rankings

Glassdoor only provides historical rankings of the current top 100 companies. As only Apple, Facebook and Google are in the current top 100 companies, the following contains a brief analysis of these 3 companies. From Figure 1, while Apple’s ranking has fallen greatly during this period, Facebook and Google hold steady between the top 20 companies ranked by Glassdoor’s proprietary algorithm.

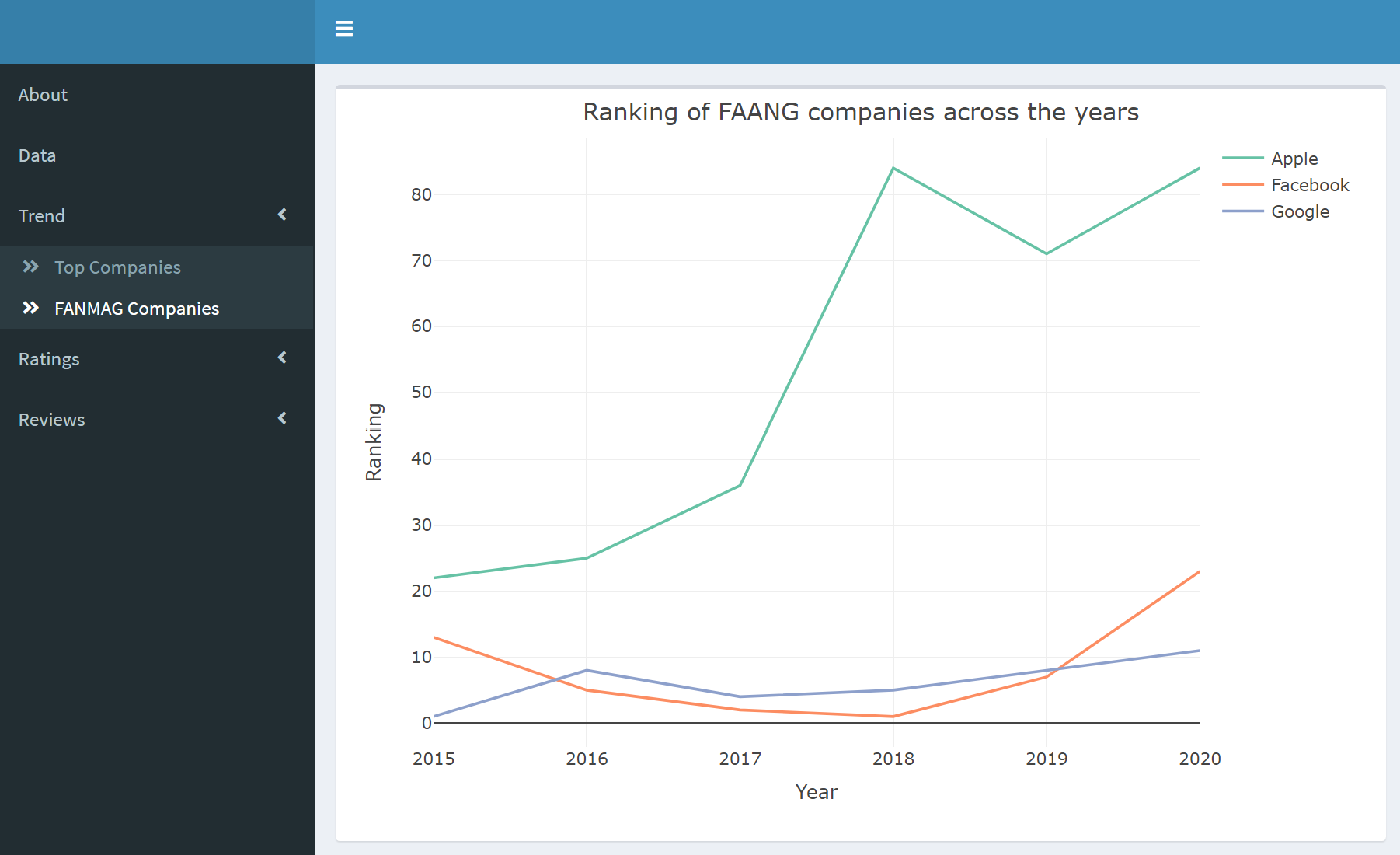


Figure 1

## 5.2 Ratings Trend

From Figure 2, the mean rating for **Amazon** showed an uptrend between 2015 and 2018. There is a sharp increase from 2015 to 2017 followed by a small increase from 2017 to 2018. A decreasing trend is observed for **Facebook**, with a sharp decrease from 2017 to 2018.

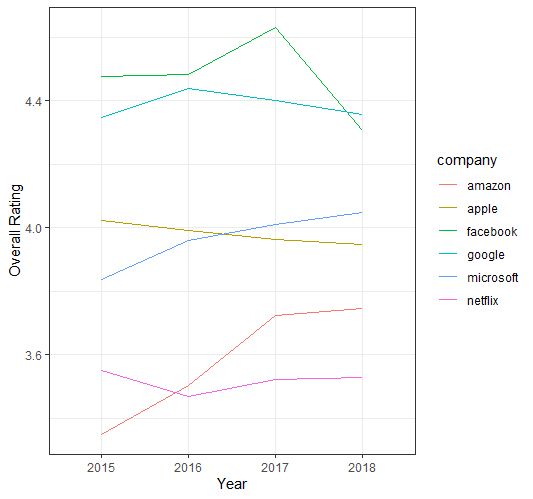


Figure 2

From Figure 3, a general trend can be observed with the ratings increasing around the start and end of year, while in the middle hovers low. Netflix displays the largest fluctuation in ratings.

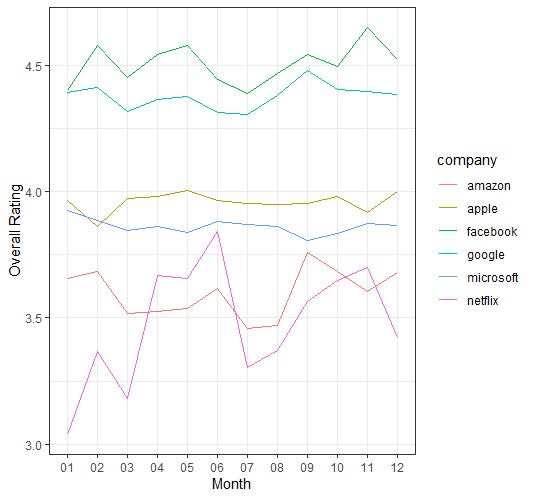


Figure 3

## 5.3 Distribution

In FANMAG industries, the distributions are as follows. **Overall ratings** and **Company Benefits** have a higher proportion of 4s and 5s**. Work Balance and Senior Management**  have higher quantity of 3s, 4s, and 5s. **Culture Values** and **Career Opportunities** scored higher in 5 and subsequently decrease down the ratings. However, in **Culture Values**, there is an **exception** of 1s being slightly higher than 2s.

|  |  |  |
| --- | --- | --- |
| Overall Rating | Work balance | Culture Values |
|  |  |  |
| Senior Management | Company Benefits | Career Opportunities |
|  |  |  |

# 6. Ratings

## 6.1 Correlation

For the preparation of the model, we created and ran a correlational matrix, to see how our variables are related. The correlation coefficient *r* measures the strength and direction of a linear relationship between two variables, with 0 indicating no correlation, 1 indicating a perfect positive correlation, and –1 indicating a perfect negative correlation.

|  |
| --- |
| Hypothesis |
| H0: There is no correlation between two variables  H1: There is correlation between variables |
| Test |
|  |
| Results |
|  |
| Interpretation |
| From the correlation plot, there is **strong correlation** between **culture** and **overall rating**. This is followed by career opportunities, senior management, company benefit then work balance. |

## 6.2 Multiple Linear Regression

First, we define the linear model for ANOVA, specified with the lm function. A two-way anova can investigate the *main effects* of each of two independent factor variables, as well as the effect of the *interaction* of these variables. Using glance, we can see that the p value is 0, reject H0 and conclude that there are differences in the means between variables.

|  |
| --- |
| Hypothesis |
| H0: Coefficients associated with variables = 0  H1: Coefficients are not equal to zero  i.e. there exists a relationship between independent variables and dependent variable. |
| Linear Model 1 regressed **without** **interaction effects.** |
| Results |
|  |
| Linear Model 2 |
| Linear model 2 regressed **with** **interaction effects** |
| Results |
|  |
| Comparison |
| Comparison of two models by an analysis of variance (ANOVA). Model 2, with interaction terms, is checked to determine whether it enhances the explanatory power of the model. |
| Interpretation |
| Since **p value is less than significance level (< 0.05)**, we **reject the null hypothesis** that the co-efficient β of the predictor is zero. It can be concluded that the **model is statistically significant**.  The results of analysis also suggest that interaction terms significantly increases the R-squared of Model 2 compared to Model 1. |
| Model 2 |
| With the chosen model, a summary of the model with estimates of the coefficients and  *p*-value for the model is obtained. |

### 6.2.1 Confirmatory Data Analysis

In practice, the model should conform to the assumptions of linear regression.

|  |
| --- |
| Linearity assumptions: |
|  |
| Results |
|  |
| Interpretation |
| Linearity assumption not satisfied. |

**Diagnostic plots** to check linear regression assumption are conducted and shown in *Annex C*.

## 6.3 Predictions

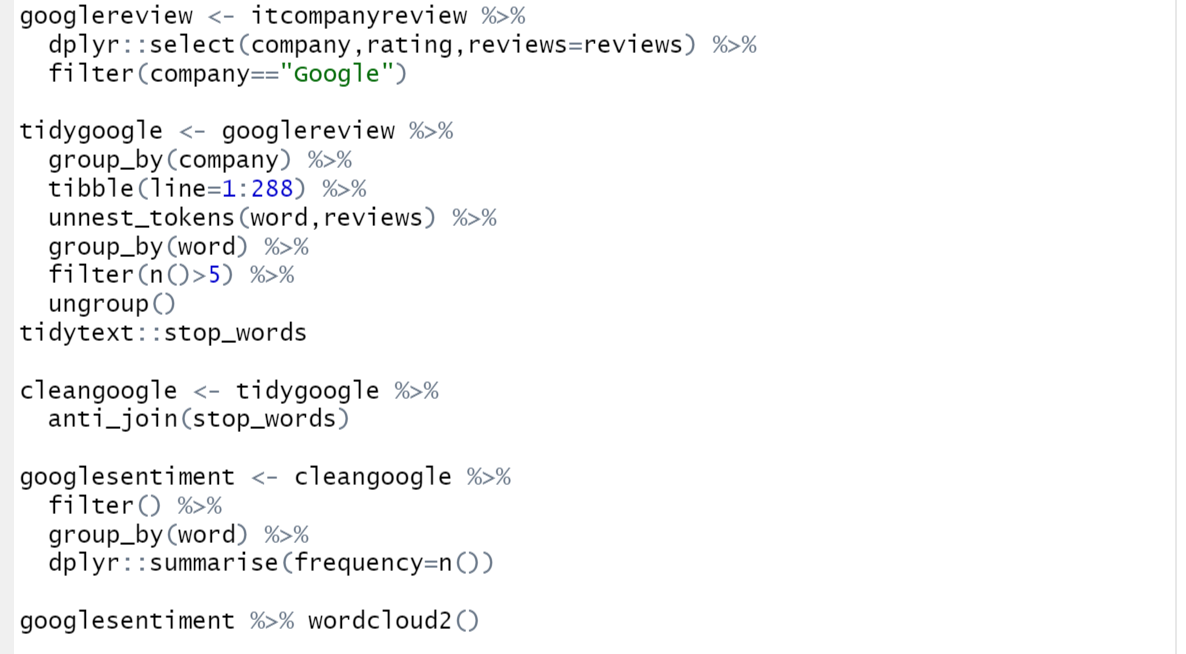
Correlation between the actual and **predicted values** can be used as a **measurement of accuracy**. A higher correlation accuracy implies that the actual and predicted values have similar directional movement, i.e. when the actual values increase the predicted values also experience an increase and vice-versa.

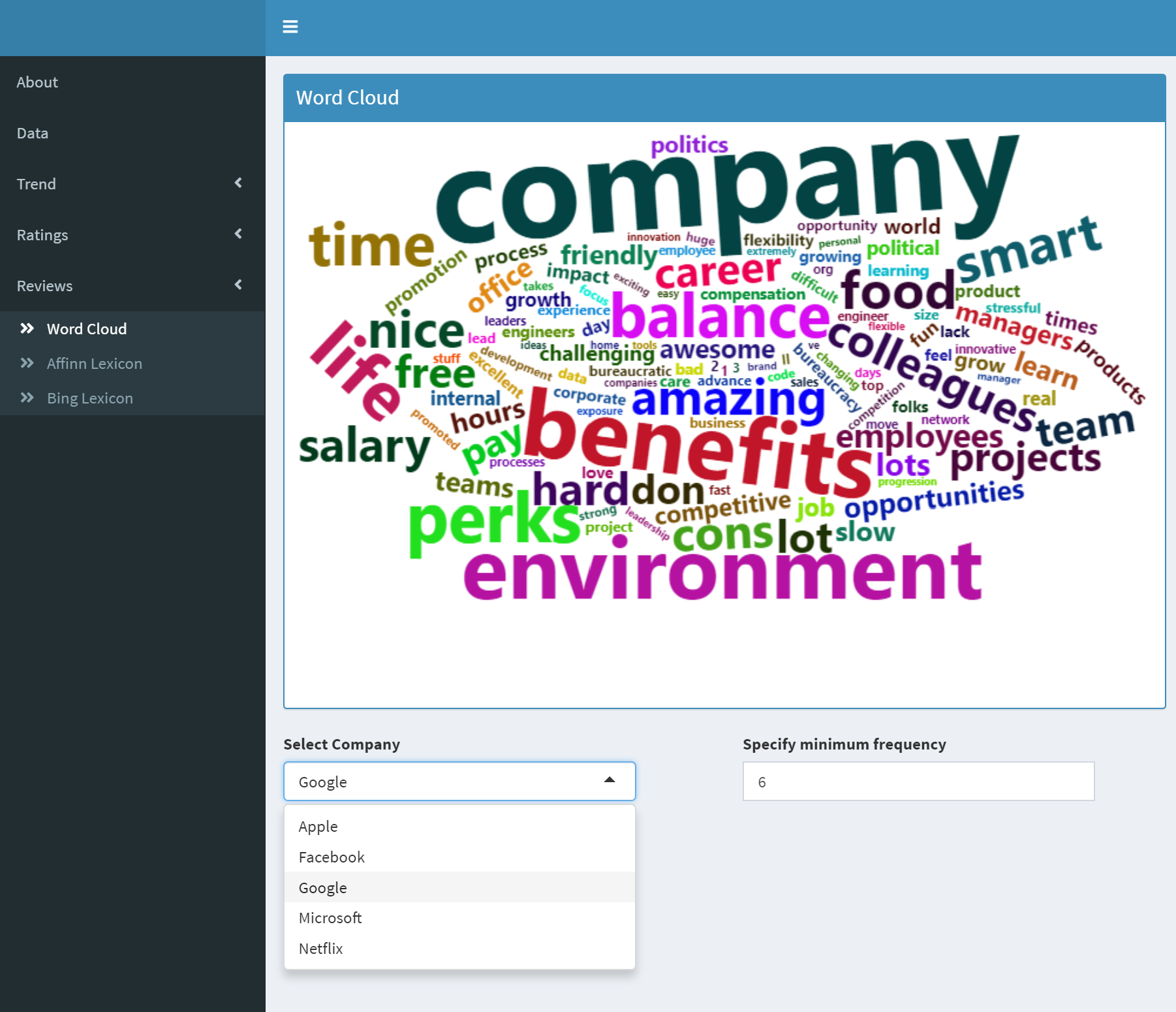
|  |
| --- |
| **Predicted Values** |
| A new column “pred” is added with the function add\_predictions() which shows the predicted value based on the correlation. |
| With the existing dataset, a grouping by 2 variables “Factors” and “Ratings” were made to obtain the aggregated mean of the predicted ratings. With this new dataset, we plot a linegraph with “Actual Ratings” against “Predicted Ratings”. The higher the gradient, the more the factor contributes to the actual ratings. From this, we can see that “Culture” and “Career Opportunities” contribute more greatly to the company’s ratings as compared to the other 3 factors. |

# 7. Review

## 7.1 Word Cloud

The data is put into a data frame as a dependency to using the tidy text dataset. It is convert to  **one-token-per-document-per-row**. Tokenization[[4]](#footnote-4) and removal of stop words[[5]](#footnote-5) are applied to the dataset.For tokenization, tidytext’s [unnest\_tokens()](https://rdrr.io/pkg/tidytext/man/unnest_tokens.html) function is applied which transforms the data to a one-word-per-row format. The removal of stop words is now applied using the anti\_join() function. Following which, comparisons of frequency of the texts are made.

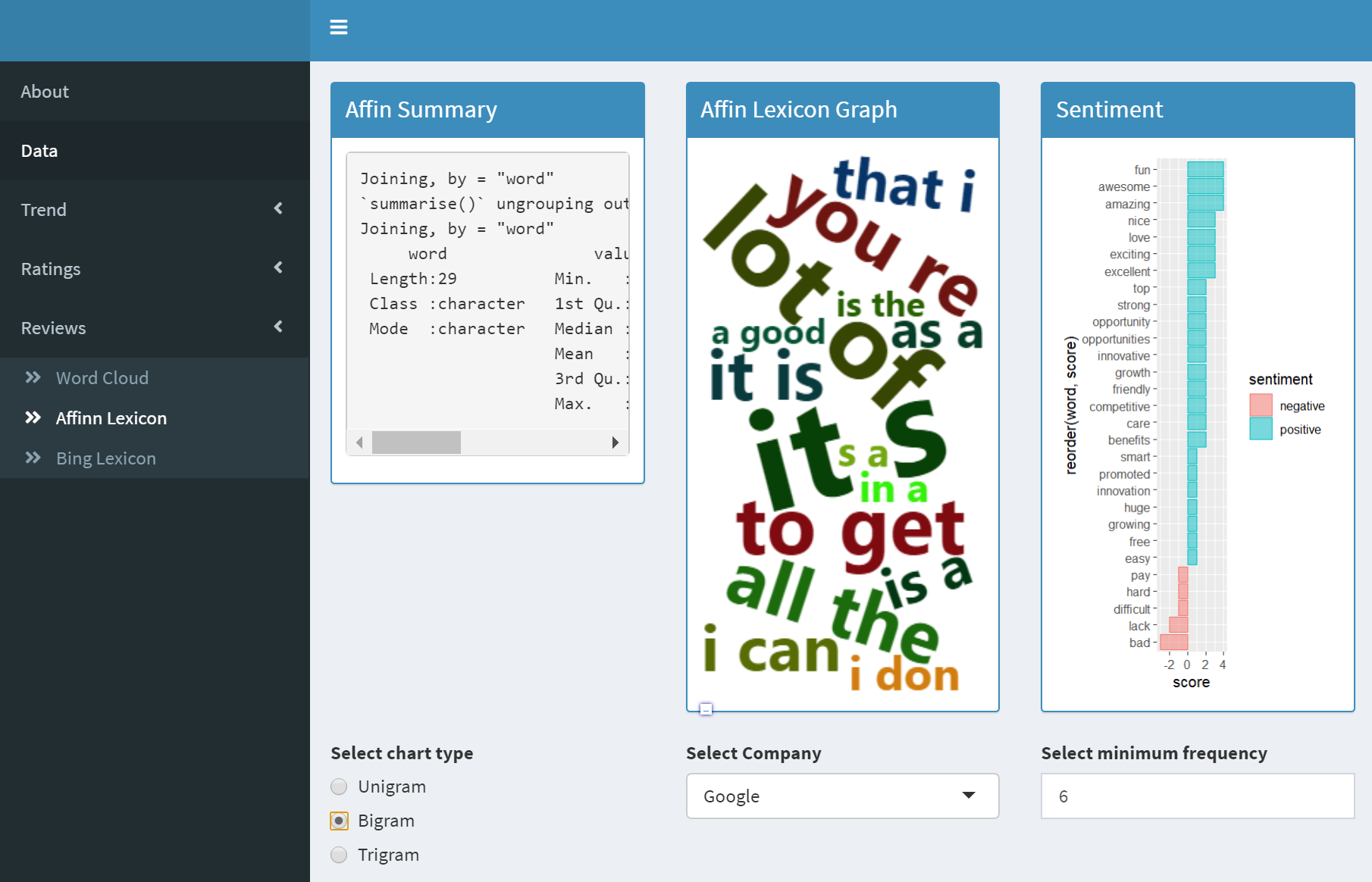
****



Detailed description of **each company’s word cloud** can be found in *Annex D*.

## 7.2 AFINN Lexicon

AFINN Lexicon is used, and the words with positive and negative sentiments for each company are explored. To add on, bigram and trigram are observed.



## 7.3 Summary of Review Findings

## Word Cloud and Sentiment Reviews from *Annex D* are summarized:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Facebook | Apple | Netflix | Microsoft | Amazon | Google | Mean |
| Word Cloud Summary | Benefits,  Amazing,  Employees,  Diversity,  Balance,  Impact,  Growth | Balance,  Hard,  Nice,  Hours,  Employees,  Flexible | Benefits, Freedom, Amazing,  Employees,  Salary,  Hours,  Smart | Benefits,  Growth,  Comfortable,  Encouraged | Culture,  Balance,  Fast,  Hard,  Growth | Culture, Benefits,  Amazing, Environment,  Smart,  Salary | - |
| Impact Comparison | Benefits (+)  Boring (-) | Benefits (+)  Pay (-) | Benefits (+)  Pay (-) | Opportunities (+)  Worse; Bullying (-) | Benefits (+)  Pay (-) | Amazing, Benefits (+)  Bad, Hard (-) | - |
| Score | 3.8 | 2.0 | 2.6 | 3.3 | 1.7 | 2.3 | 2.6 |
| Deviation | +1.2 | -0.6 | +0.01 | +0.7 | -0.9 | -0.3 | - |

## 7.4 Linear Regression

|  |
| --- |
| Linear Regression |
|  |
| Boxplot (How sentiment scores affect overall rating) |
|  |
| Linear Plot (Predictions) |
|  |
|  |
| Interpretation |
| It can be observed lower review scores lead to lower ratings and vice versa. With the mean of the sentiment scores closer to +2, ratings are more likely to be closer to 5. Conversely, with lower mean sentiment scores closer to 0, ratings are more likely to be to be closer to 1. |

# 8. Conclusion

In conclusion, “Culture” and “Career Opportunities” factors have the greatest impact on overall rating. Further analysis of sentiment analysis revealed that FANMAG does not deviate much from the mean sentiment scores, with Facebook attaining greater positive impact on company overall ratings.

For **individuals**, they can decide what they value the most and not rely on the overall rating as it is impacted most heavily by “Culture” and “Career Opportunities” factors. For example, if they value opportunities, they should look beyond solely the overall ratings and focus on the word cloud and the rating for greater insight into their company choice. For instance, “opportunities” occurred frequently with a high impact score for Microsoft and this might pose as a good fit for this individual.

For the 3 **companies** not in the top 100, they can look to focus on “Culture” and “Career Opportunities” factors to increase their overall rating and hence ranking in the near term to gain greater visibility and potentially attract better talents. For the 3 companies currently in the top 100, they can instead look beyond their overalls ratings and focus on the aspects which are lacking which contribute less to their overall ratings. For example, Microsoft has a huge proportion of negative sentiments of “bullying” and they can look to improve their ratings for “senior management”.

*Annex A*

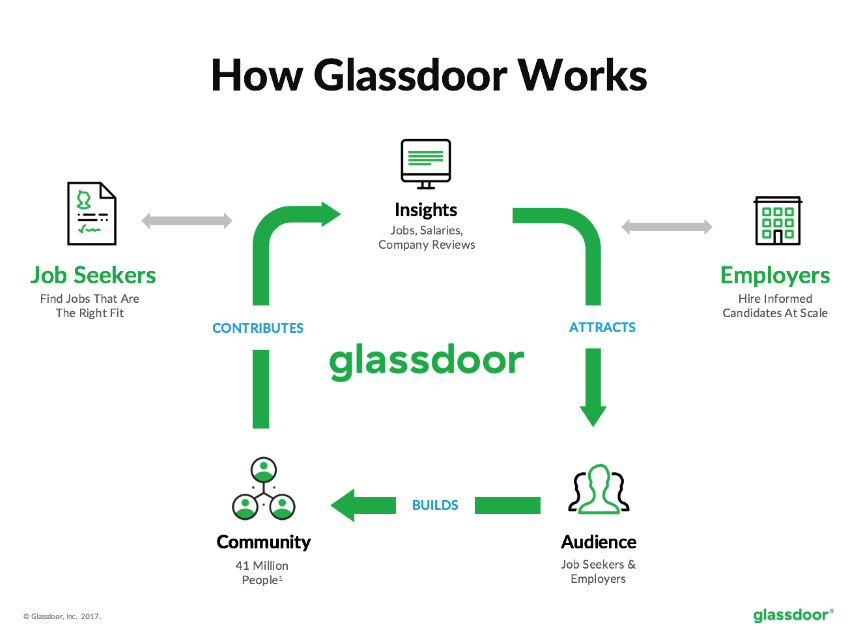


Figure 1

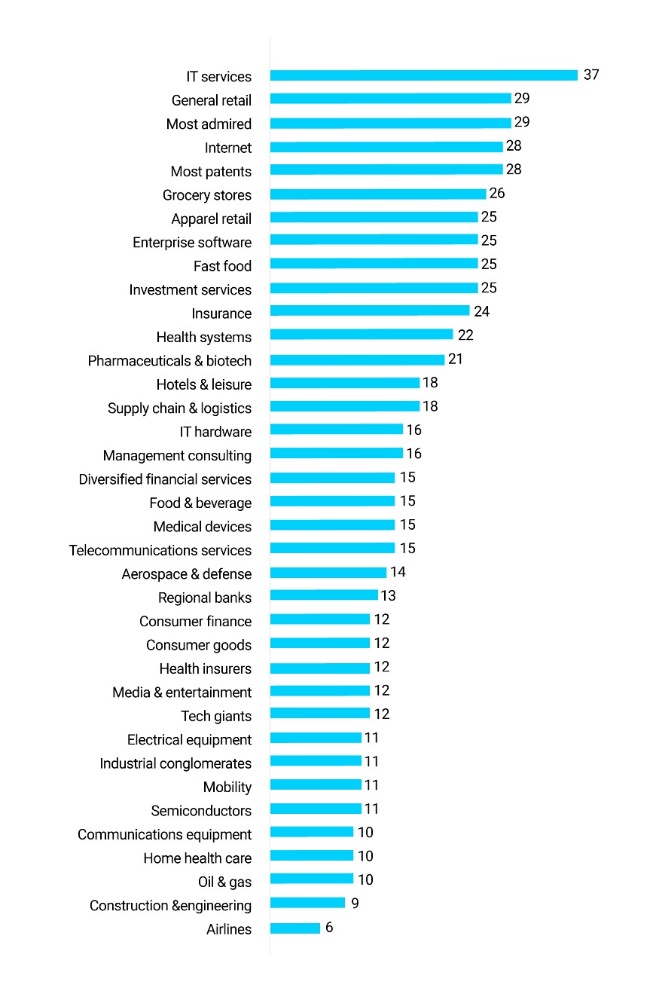


Figure 2

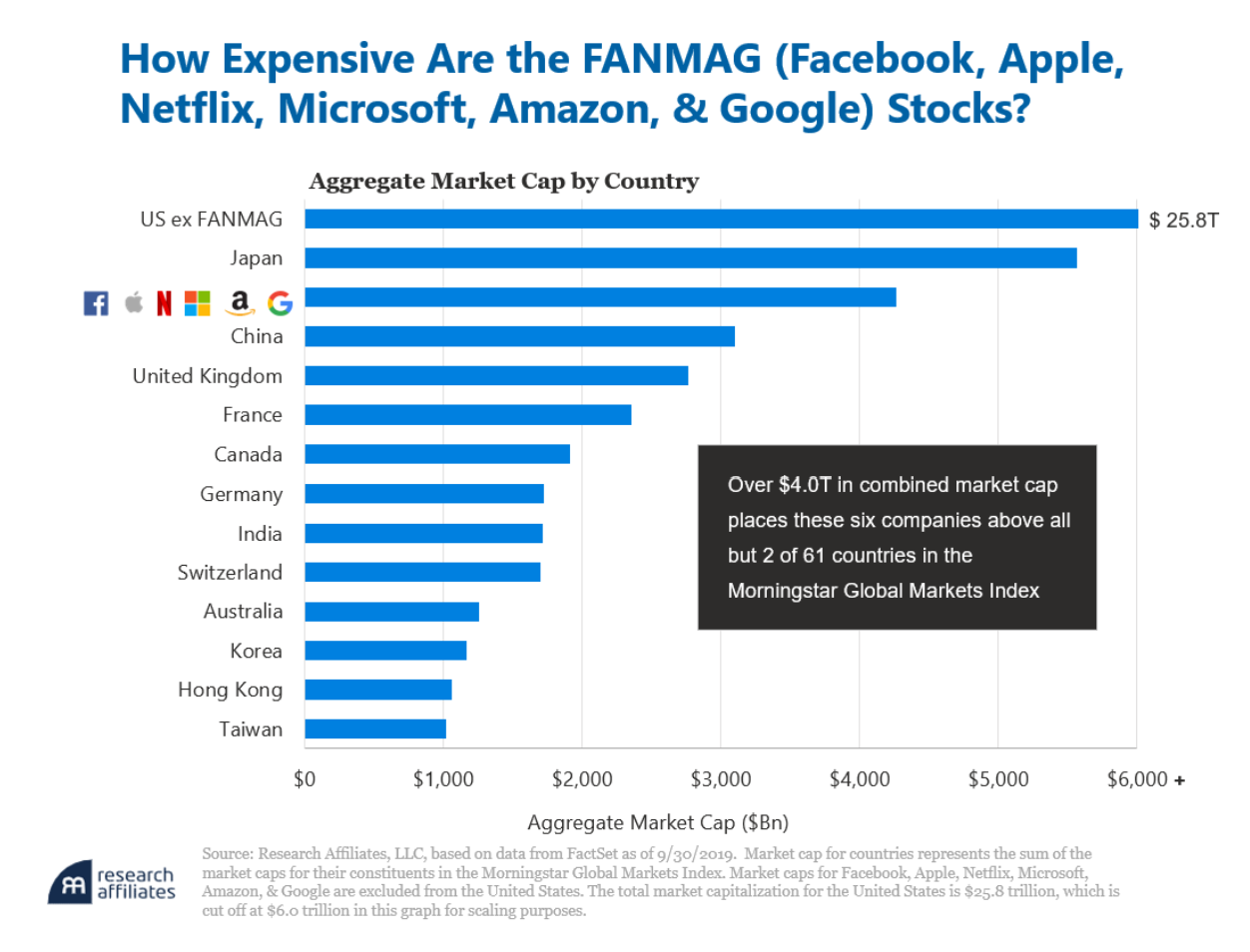


Figure 3

*Annex B*

library(shiny)

library(shinydashboard)

library(readr)

library(dplyr)

library(tidyr)

library(purrr)

library(stringr)

library(lubridate)

library(reshape2)

library(wordcloud)

library(wordcloud2)

library(tidytext)

library(DT)

library(ggplot2)

library(plotly)

library(corrplot)

library(vcd)

library(rcompanion)

library(gvlma)

library(ggplot2)

library(lattice)

library(lava)

library(purrr)

library(caret)

library(nnet)

library(tidyverse)

library(broom)

library(modelr)

library(stringr)

library(Hmisc)

library(tidytext)

library(modelr)

library(ggpubr)

*Annex C*

|  |  |
| --- | --- |
| **Diagnostic Plot** | |
|  | |
| **Residuals vs Fitted** | |
|  | |
| **Normal QQ** | |
|  | |
| **Scale-Location** | |
|  | |
| **Residuals** | **Cook’s Distance** |
|  |  |

*Annex D*

|  |
| --- |
| Facebook |
| Word Cloud |
|  |
| Sentiment |
|  |
| Apple |
| Wordcloud |
|  |
| Sentiment |
|  |
| Netflix |
| Word Cloud |
|  |
| Sentiment |
|  |
| Microsoft |
| Word Cloud |
|  |
| Sentiment |
|  |
| Amazon |
|  |
|  |
| Sentiment |
|  |
| Google |
| Word Cloud |
|  |
| Sentiment |
|  |

1. <https://expandedramblings.com/index.php/numbers-15-interesting-glassdoor-statistics/> [↑](#footnote-ref-1)
2. https://www.softwareadvice.com/resources/job-seekers-use-glassdoor-reviews/ [↑](#footnote-ref-2)
3. https://sloanreview.mit.edu/projects/measuring-culture-in-leading-companies/ [↑](#footnote-ref-3)
4. Process of splitting text into tokens [↑](#footnote-ref-4)
5. Words that are not useful for an analysis, typically extremely common words such as “the”, “of”, “to”, and so forth in English. [↑](#footnote-ref-5)