Final_Assignment

Nicole Conrad, Nurjahan Hera and Manna Toth

5/26/2019

`THE RELATIONSHIP BETWEEN BENEFITS AND JOB SATISFACTION/EMPLOYEE WELL-BEING`
`Analyzing Benefit Comments of Glassdoor`

Introduction

During good economic times, companies are competing to get the best possible workforce. They may want long-term, highly skilled, innovative or a combination of many attributes for their company's own success. The United States labor market is relatively deregulated compared to other areas of the world (i.e. in Europe). Therefore, workers are not guaranteed certain 'benefits' in the US– for example, health insurance, life insurance, paid maternity leave, and paid vacation days are all optional for a company to provide for it's employees. During a stable economy, the benefits provided company to company are governed by a competitive marketplace to attract and retain the 'best' talent. We want to use these optional company benefits to explain employee satisfaction in order to see how this modern deregulated labor market is providing benefits that might be compulsory in other markets. Are US companies responding to the job market by offering benefits and then are they seeing higher employee satisfaction and attracting their target workforce? Is a deregulated labor market able to still provide high employee satisfaction without mandated benefit requirements?

Therefore, we need to research how important benefits are to the average worker's satisfaction in their employer. If employees have high satisfaction in their employer, how much can we attribute to the benefits offered, and which benefits are most important? Benefits range from health insurance coverage to free coffee and snacks, and in today's marketplace, a comprehensive benefit plan can make or break a company's reputation for prospective employees.

We believe that analysing the benefit comments would allow us to examine a different, more subjective dimension of what the offered benefits mean for the individuals (employees). Text analysis may allow us to discover patterns "over" the quantitative relationships - the association between the benefit scores and the general company ratings.

Research Questions

1. How important are benefit comments to employee's rating in a certain company?
2. Which text analysis technique predicts the best: sentiment analysis, topic modeling or readability score?

Data

Data source and quality

To find out how benefits play into employee well-being, we decided to use Glassdoor, a California-based company rating website that offers information from users all over the world. There were 424,244 companies and 6,057,729 ratings as of January 23rd, 2019, but the data is easily changeable over time. It is written in JSON format.

We limited our scope to companies operating within the United States because we know labor regulation and health benefits are more similar across this nation that they would be worldwide. The website depends on user rankings and comments for each company and allows the user to independently rate and comment on different aspects of a company's working environment—for example, interview questions, salaries, and overall ratings of many different categories. Users either specify themselves as current or former employees, but importantly, there is no verification process to prove they are/were associated with the company. Out of the total information on the website, we focused on the rating information generated by users, general company data provided by Glassdoor, and the company benefits pages that have information from both employers and employees. The benefit comments are also listed here, employees can write as much as they want in a free format.

The data generation process is a combination of voluntary reviews, information that Glassdoor itself has collected, and information that the employer itself has verified (the company attests to whether or not it offers particular benefit categories). We believe that the benefit categories that are verified by a representative of the company is reliable, so we decided to use the verified benefit information for our research. We also used Glassdoor-provided data on the company like information on the companies' revenue, industry, and location. Finally, we used the more subjective data of the individual users that have specified themselves as either employees, potential employees, or former employees of the company in the form of comments and ratings. We believe this data is subjective because it is possible for anyone to participate in the community whether or not they actually worked for or interacted with the company. Also, the company could potentially create fake reviews or ratings in order to increase their company-wide scores, so part of the information may be fictitious from either the employee or employer side.

Our assumptions going into this project are that the general reviews of a workplace indicate employee satisfaction and well-being. We also assume that users are submitting real data, write their own opinion, and they actually work or used for work for the companies indicated in order for us to gain any useful results.

Datasets

We used our own data base which was scrapped from the Glasdoor website by using the rvest package. We collected data on a company level in one dataset and in another, individual benefit comments that indicate which company they correspond to.

1. Company-level dataset (name of dataset attached is company_data.rds): we just kept companies with more than 1000 reviews. After cleaning the data we had 429 companies.In order to best describe our full project, we will explain all of the variables. It has 63 variables all together:

- General rating per company, which is ranked between 1-5.
- 54 employer verified benefit categories, these are binary variables show which benefits are offered in that companies;
- Sum of the verified benefit categories per companies;
- Benefit rating per company which is ranked between 1-5 (it was ranked independently from the general rating);
- US or international headquarter (binary);
- Size of the company, which is ranked from 0-7: “1 to 50 employees”,“51 to 200 employees”, “201 to 500 employees”, “501 to 1000 employees”,“1001 to 5000 employees”, “5001 to 10000 employees”, and “10000+ employees”;
- Revenue of the company which is ranked from 0-10: There were 10 different classify cations of revenue. The top were also followed: “Less than $1 million (USD) per year”, “$5 to $10 million (USD) per year”," $25 to $50 million (USD) per year“, “$50 to $100 million (USD) per year”,” $100 to $500 million (USD) per year“,” $500 million to $1 billion (USD) per year“, “$1 to $2 billion (USD) per year”, “$2 to $5 billion (USD) per year”,” $5 to $10 billion (USD) per year“, and“$10+ billion (USD) per year”;
- Region code of the company's headquarter;
- Industry (there are 78 different industries)

2. Individual benefit comments dataset (name of the dataset attached: all_comments.rds) The following section shows how we cleaned, transformed and discovered the data.

```r
knitr::opts_chunk$set(echo = TRUE, error = TRUE)
```

Opening, cleaning and getting to know the database

```r
Sys.setlocale(locale = "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8"
```

```
library(foreign)
library(stringr)
library(quanteda)
library(readtext)
library(topicmodels)
library(ggplot2)
library(data.table)
library(Hmisc)
library(dplyr)

#setwd("~/Documents/CEU/2018_2019_spring/Text/final")

comments <- readRDS("all_comments copy.rds")
str(comments)
```

```
## Classes 'data.table' and 'data.frame':   145463 obs. of  2 variables:
##  $ companynames: chr  "3M" "3M" "3M" "3M" ...
##  $ comments    : chr  " 2 weeks of vacation, 1 week of \"personal time,\" 4 days of personal holidays
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr  "companynames" "comments"
```

```
names(comments) <- c("name", "text")
corpus <- corpus(comments)
corpus <- corpus_subset(corpus, texts(corpus) != "")
ndoc(corpus)
```

```
## [1] 145453
```

```
head(docvars(corpus, "name"))
```

```
## [1] "3M" "3M" "3M" "3M" "3M" "3M"
```

```
#unique(docvars(corpus, "name"))

table <-table(docvars(corpus, "name"))
#data.frame(table)
describe(table)
```

```
## table
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      429        0      290        1    339.1    306.9     80.0    106.8
##      .25      .50      .75      .90      .95
##    150.0    219.0    363.0    653.6   1128.0
##
## lowest :   10   30   33   38   44, highest: 1979 2021 2119 2210 2555
```

```
summary(table)
```

```
## Number of cases in table: 145453
## Number of factors: 1
```

The database had 145 463 individual comments, but some comments were just emptry strings - we have discovered it later during the analysis - so we have remained with 145 453 comments.

Since these are benefit comments, all of them covers a very specific topics, opinion, evaluation of description of the different benefits offered in that company. Comments are usually not so long, they range from a few words to a maximum 10 sentences. All comments were written in english, using rather informal and simple language.

The company with the lowest number of comments has 10 comments, while the company with the highest number of comments has 2555 comments. The mean number of comments per company is 339.1.

We have decided to create two dfm files, as we will use a stemmed one for the topic modeling and for the complexity analysis, while use a non-stemmed version for the sentiment analysis. The stemmed file has 20 436 features, while the non-stemmed one has 28 668. The top 50-50 features can be seen above of both the dfm files.

```
corpus.dfm_cloud <- dfm(corpus, stem = FALSE, remove=stopwords("english"), remove_punct=TRUE, ngrams = 
corpus.dfm <- dfm(corpus, stem = TRUE, remove=stopwords("english"), remove_punct=TRUE, ngrams = 1, tolo

nfeat(corpus.dfm) #20 436
```

```
## [1] 20436
```

```
nfeat(corpus.dfm_cloud) #28 668
```

```
## [1] 28668
```

```
#topfeatures(corpus.dfm)
topfeatures(corpus.dfm_cloud, 10)
```

```
##  benefits    health      401k insurance      time      good     great 
##     65181     41155     39293     37697     32534     30176     29948 
##  vacation   company   package 
##     21990     20351     18849
```

```
topfeatures(corpus.dfm, 10)
```

```
## benefit  health    401k   insur    time    good   great   match employe 
##   73276   41174   39312   38225   33495   30245   30076   25757   25756 
## compani 
##   25199
```

Since we have a great number of comments and we would like to concentrate on words that majority of the employees used to describe benefits in order to discover some general patters, we kept those tokens that appear in at least 20 documents. We have remained with 3035 features for the stemmed and 4087 features for the non-stemmend document.

```
library(quanteda)
corpus.dfm = dfm_trim (corpus.dfm, min_docfreq = 20)
corpus.dfm_cloud = dfm_trim (corpus.dfm_cloud, min_docfreq = 20)

nfeat(corpus.dfm) #3 035
```

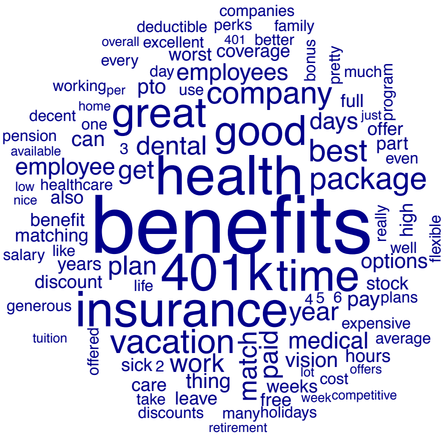
```
## [1] 3035
```

```
nfeat(corpus.dfm_cloud) #4 087
```

```
## [1] 4087
```

The following graph shows the wordcloud of the 100 most frequent words in the non-stemmed dfm. As we saw in the top features, we can see that benefit comments has two most important dimensions: 1) the different types of health and pension contributions and 2) time, holiday and vacation are among the most frequently occured words.

```
textplot_wordcloud (corpus.dfm_cloud, max_words = 100)
```



Methodology

Text Analyis on the Comments Dataset

We are specifically going to use text analysis to analyze the companies' comments on their benefits to see if they tell us anything about how the company is generally scored. We would like to use three different techniques to see if they help us predict a company's rating. Each of these techniques of text analysis should give us variables to add to an overall linear analysis of our data, which we can then see which is the most useful.

First, we will conducted complexity analysis, which will tell us how sophisticated or how easily readable the average comments are per company. For determining the complexity of comments across companies, we will apply the Flesch reading-ease score method which is measured in the scale of 1-100. 1 denoting text with more complexity and 100 denoting easier readability. The manual formula for getting the readability score is the following: 206.835 - 1.015 x (words/sentences) - 84.6 x (syllables/words). We can use the textstat_readability() code of quanteda in R to determine the readiability scores for all comments across companies. In the next step aggregated the readability score at company level. We expect to see certain companies with either very pleased employees or very disgruntled employees with higher complexity comments. Concerns: as we mentioned before, we are analysing comments which are generally shorter and use simplier language. However, there still can be a different patterns accross certain groups.

Then, we will do a sentiment analysis. We will use two different dictionaries - the Neal Caren and the Lexicoder - in order to have more robust results. We are going to check the word polarity of the comments: words in the article are assigned to a polarity score based on the lexicons and then the scores will be added to determine the sentiment of the comment. It allows us to see generally what the overall sentiment of each companies' comments will be and if that corresponds to the overall score or the benefit scores. As mentioned before we are interested in the polarity (positive or negative opinion) in the text and sentiment analysis is one of the best approach for it. In case of comments it is a useful method, since it allows to study and understand deeper the affective states and the opinion behind the quantitative scores. We expect that companies with higher avarege sentiments will have higher general ratings on average. Concerns: words can be used in different context or in a sarcastic way - this is relevant in case of comments - what can change the polarity and this method is not able to treat this problem.

Finally, we will use LDA topic modeling to see if we can distiguish topics in the comments. We will also examine if there is a relationship between topic prevalence and company or benefit ratings. We choose this method becasue it can help discover patterns in the text we can not be notice otherwise. Seince it differentiates topics across the corpus by using unsupervised, statistical methods. We will try more models with different numbers of topics (2,4,5,6,7,8 and 9), and decide which is the best model based on the perplexity score and our based on own individual decision whether the offered topics make sense. We expect that our model will be able to distinguish meaningful topics across the comments. Concerns: The results are not robust since they are probabilistic based on this certain dataset. Furthermore the results are not always explicable, words may be assigned to one topic without any semantic relevance (see more: Sam Tazzyman:https://moj-analytical-services.github.io/NLP-guidance/LDA.html).

Predictive Analysis on the Company Dataset

After we are done with the Text Analyis, we are going to aggregate the results on a company level and merge them with the company dataset.We are going to use the agregated results of the text analyis as predictors for overal company scores. We are going to run a multiple linear regression model to examine the relationship betweeen the general score of the company and the compexity and the sentiment analyis - since these variables are numberic. We need to run an ANOVA to check the relationship between the topic models (the most prevalent topic among the comments of the company) and the ratings - since the topics are categorical variables. We expect our analyis to answer our questions in the following way:

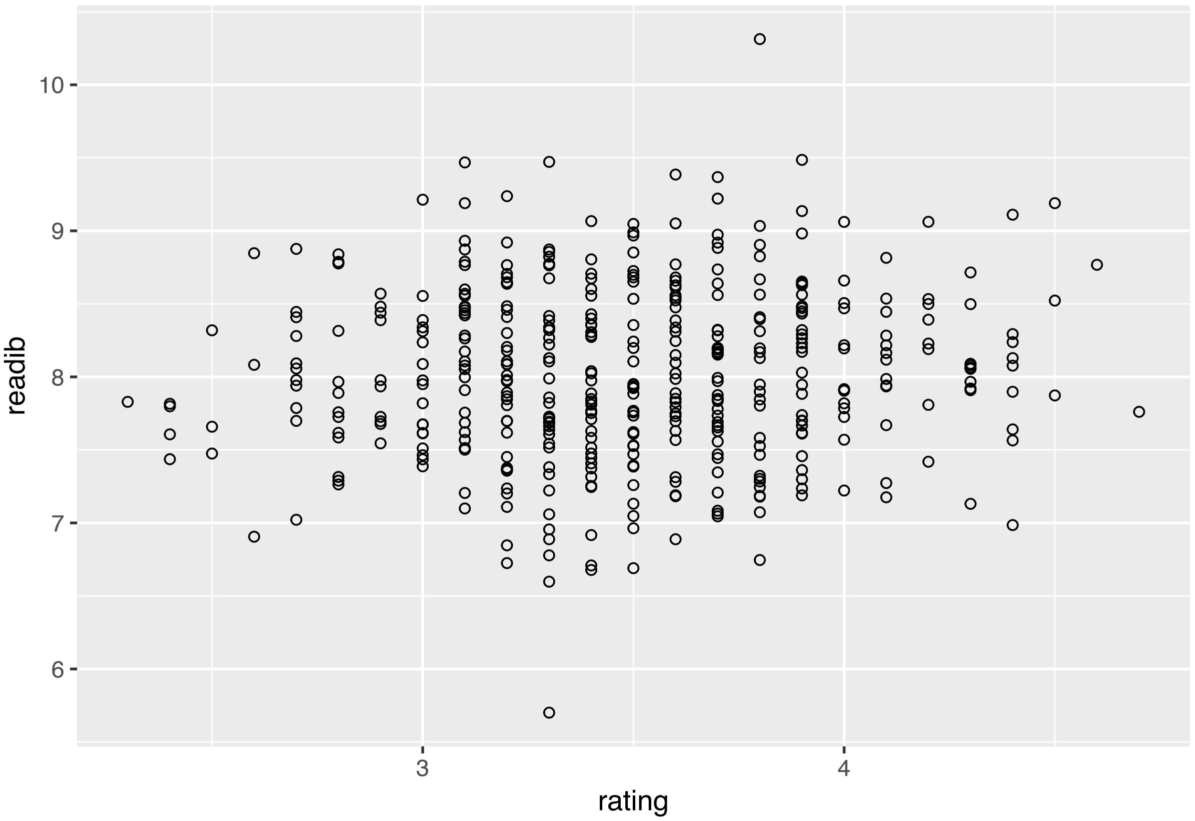
1. How important are benefit comments to employee's rating in a certain company? If any of the variables is significant - the p value is lower than 0.05 - then we can say that the comments are important. We can use them as predictors of general scores.

2. Which text analysis technique predicts the best: sentiment analysis, topic modeling or readability score? In case of the complexity and the sentiment analysis we define that the higher the r2 the better the certain predictor is. In case of the topic models we could calculate chi2, but chi2 and r2 can be compared only in relative terms. The p value may also be a proxy of the strenght of the relationship: the smaller the p value, the better the predictor is (this is relevant in all the three predictors)

```r
# Opening the company database
company_data <- readRDS("~/Documents/CEU/2018_2019_spring/Text/final/company_data.rds")
```

Complexity Analysis

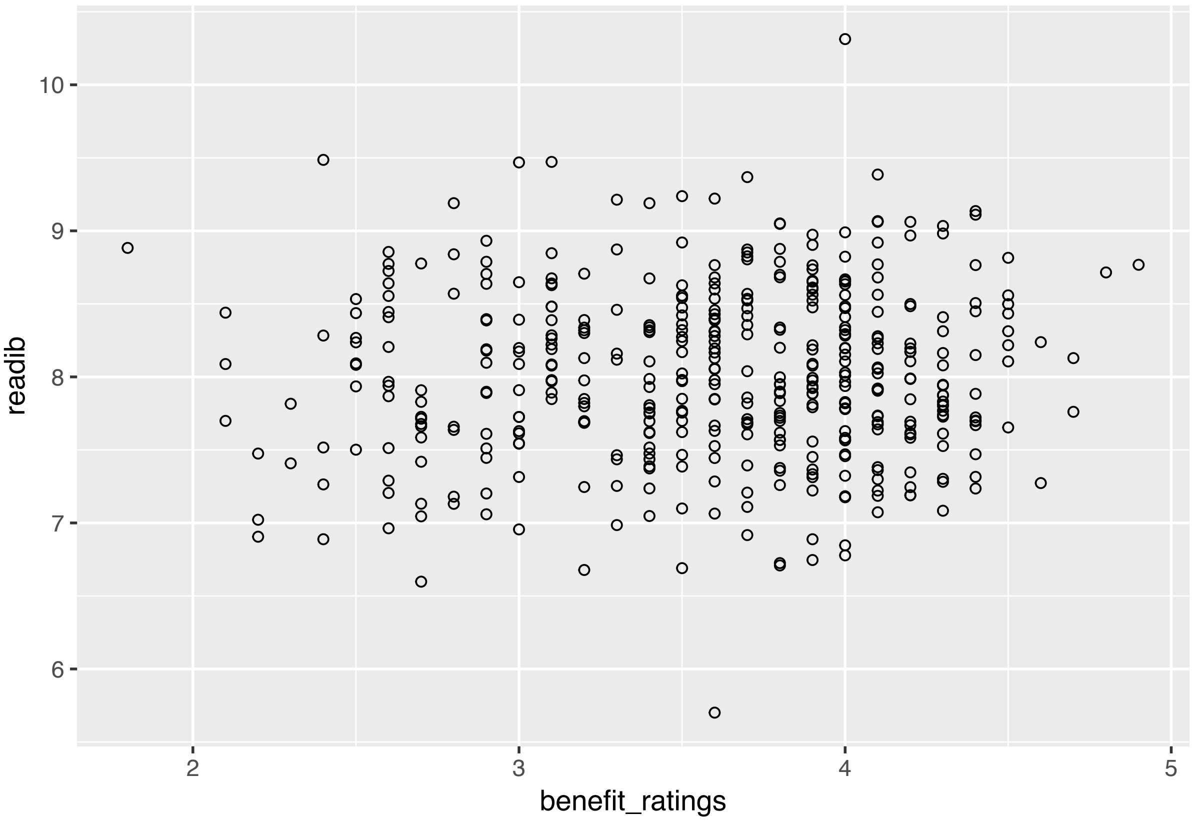
```r
#we have realized the issue with the empty strings here, that was the reason why the flesch kincaid fun
#rows with empty strings needed to be deleted from the original as well
comments_2 <- comments[!apply(comments, 1, function(x) any(x=="")),]
comments <- comments_2

docvars(corpus, "flesch.kincaid") <- textstat_readability(corpus, "Flesch.Kincaid")[,2]
net.read <- docvars(corpus, "flesch.kincaid")
readibility <- data.frame("name" = comments$name, net.read)

#aggregating results on the company level and adding them to company data
readib_2 <- aggregate(readibility$net.read, by=list(readibility$name),FUN = mean)
names(readib_2) <- c("name", "readib")
company_data$readib <- readib_2$readib

#Plots distribution of readibility by general company rating
read.plot <- ggplot(company_data, aes(x= rating, y = readib)) + geom_point(shape = 1)
print(read.plot)
```



The data from readability scores shows that the highest score is 10.31 and the lowest one is 5.7 which is close to 1 in the scale of 1-100. This means that comments in Glassdoor by the employees are complicated according to Flesch reading-ease score. This outcome can be due to use of complicated industrial terminologies, complicated words or longer syllables. Additionally, it is also possible that people who write comments to Glassdoor are the ones who use more complicated language - and people who use less complicated language would less likely to register or write comments

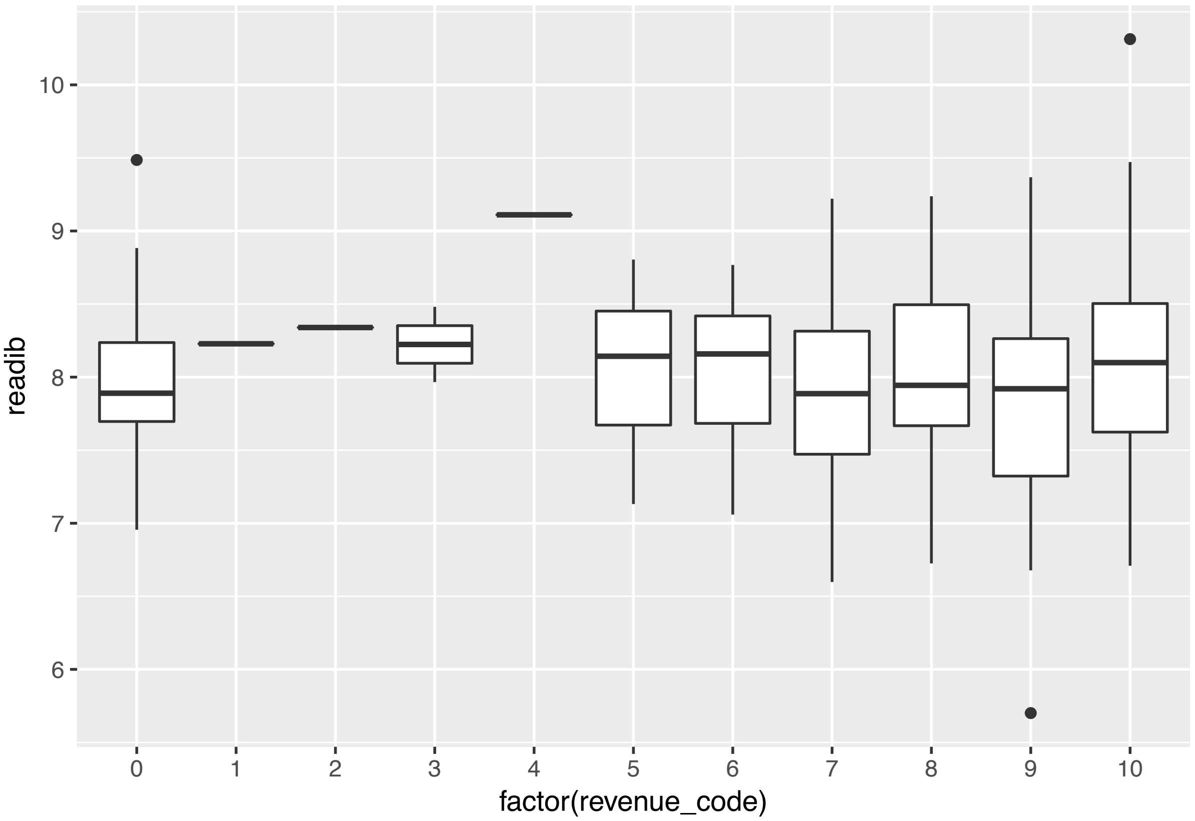
This scatter plot depicts the relation between readability scores at company level with the general rating assigned to the companies. The plot shows that companies with ratings between 3 and 4 (scale 1-5) have readability scores of 7 to 9 (1-100) for comments aggregated at company level. We can also assume that there is no strong relationship between readibility score and overal rating of the company from this graph.

```
#distribution of readibility by benefit ratings
read.plot2 <- ggplot(company_data, aes(x= benefit_ratings, y = readib)) + geom_point(shape = 1)
print(read.plot2)
```



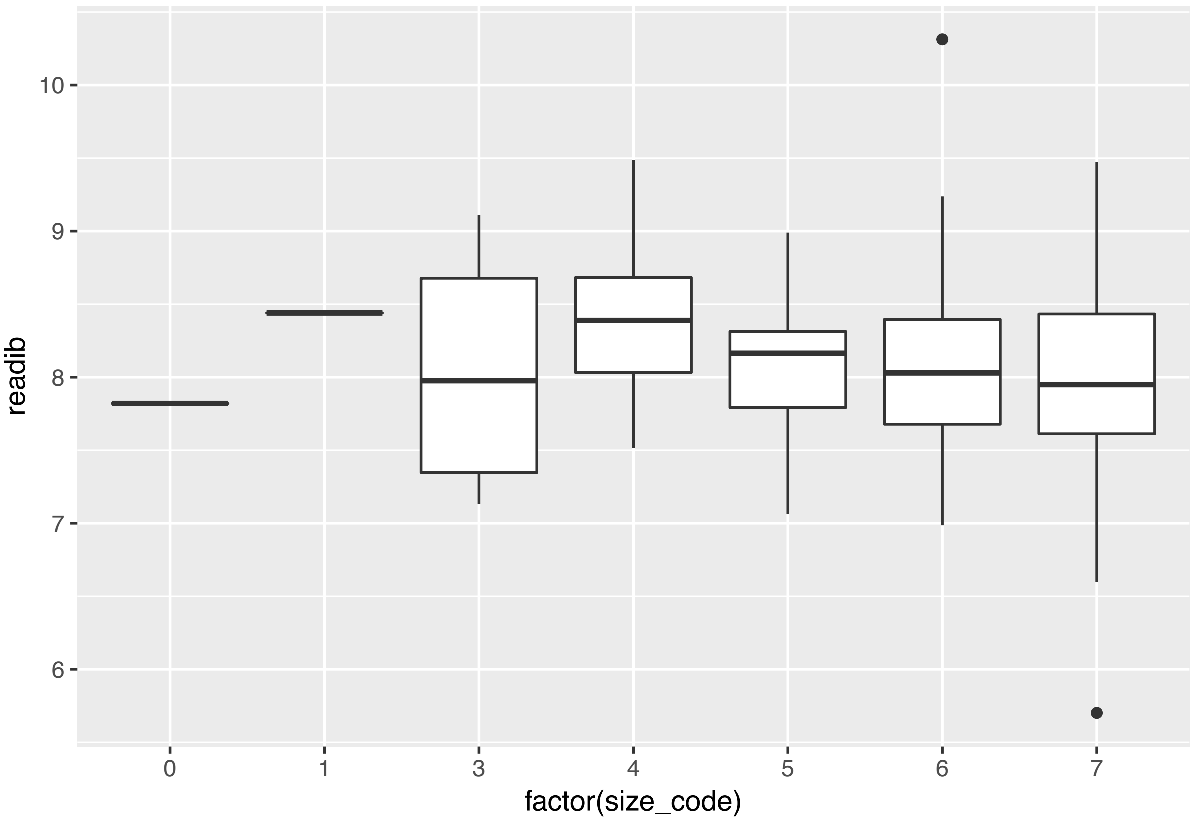
The scatter plot above portrays the spread of readability scores per company across the benefit rating per company. From this graph we can see that, companies with benefit rating between 3-4.5 have readability scores between 7-9. The scatter plot also gives some outliers for which we can say that there is no correlation between readability scores of the comments per company and benefit rating assigned to each company.

```
#distribution of readibility by revenue
read.plot3 <- ggplot(company_data, aes (x = factor(revenue_code), y = readib)) + geom_boxplot()
print(read.plot3)
```



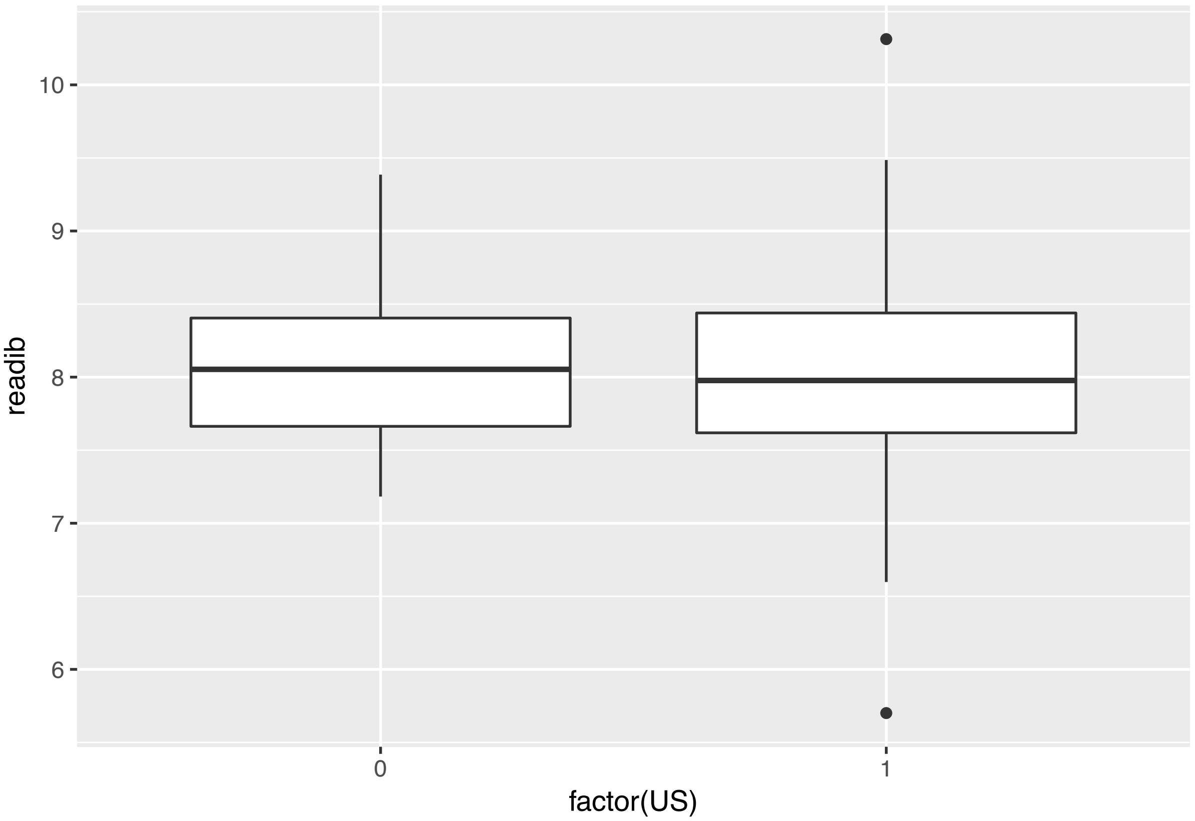
The above box plot provides distribution of readability scores across revenue code per company. It is evident from the box plots that there is no strong relationship between revenues and readibility scores.

```
#distribution of readibility by size
read.plot4 <- ggplot(company_data, aes (x = factor(size_code), y = readib)) + geom_boxplot()
print(read.plot4)
```



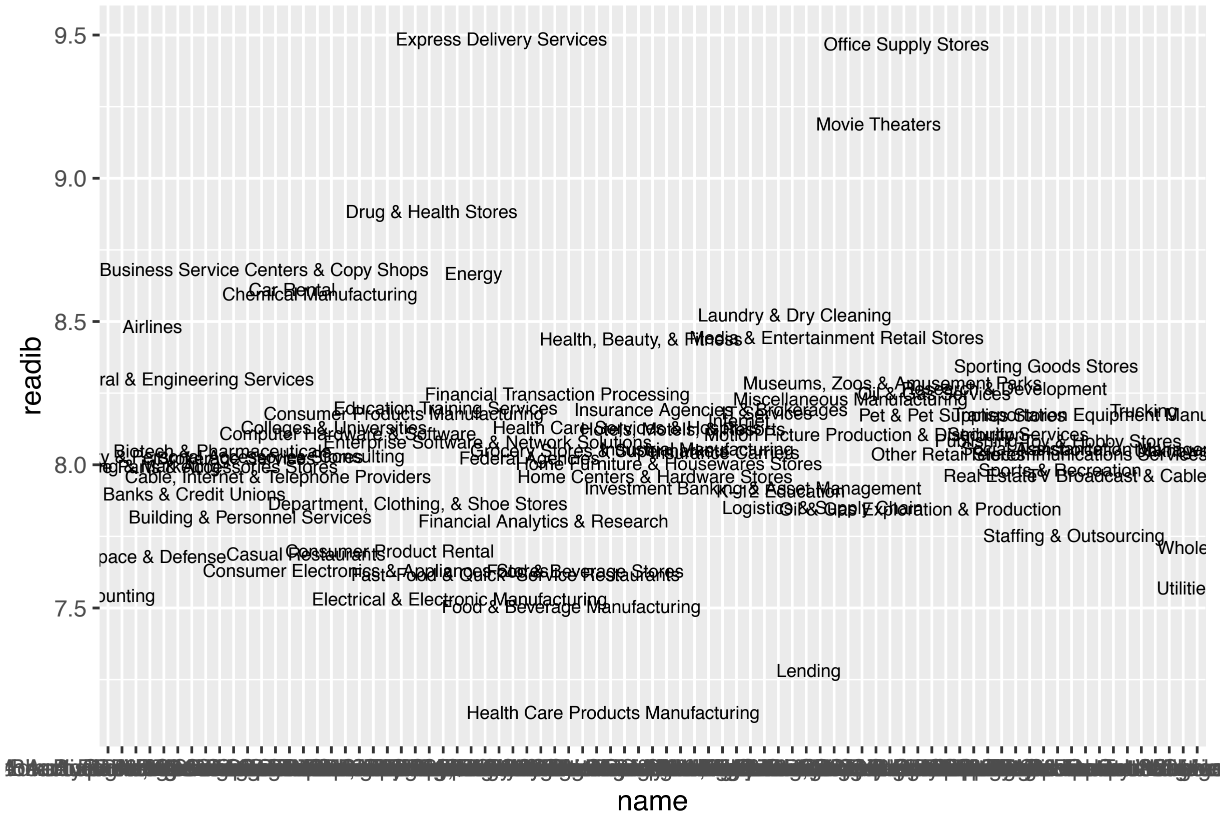
The above box plot provides distribution of readability scores across size code per company. It seems that there are no significant relationship between these two, however after company size reaches 500 employee it seems that and increase in size is assiciated with lower readibility score.

```
#distribution of readibility by US
read.plot5 <- ggplot(company_data, aes (x = factor(US), y = readib)) + geom_boxplot()
print(read.plot5)
```



The above box plot provides distribution of readability scores across companies with US association. It is evident from the box plots that irrespective company being US based, the median readability score is close to 8.

```r
#distribution of readibility by Industry
#calculationg industry average
industry_a <- aggregate(company_data$readib, by=list(company_data$industry),FUN = mean)
names(industry_a) <- c("name", "readib")
#plot
read.plot6 <- ggplot(industry_a, aes(x = name, y = readib)) + geom_text(aes(label=name), size = 2.3)
print(read.plot6)
```



We can see that Express Delivery, Office Supply Stores and Movie Theaters have the highest industry avarage readibiliy while Lending and Health Care Product Manufacturing have the lowest. None of the highest scored industries are assiciated with writing or intellectual work except the Movie Theaters.

Sentiment analysis

```
#Lexicoder dictionary
#str(data_dictionary_LSD2015)
sentiment.dfm <- dfm(corpus.dfm_cloud, dictionary = data_dictionary_LSD2015)
dim(sentiment.dfm)
```

```
## [1] 145453      4
```

```
head(sentiment.dfm)
```

```
## Document-feature matrix of: 6 documents, 4 features (62.5% sparse).
## 6 x 4 sparse Matrix of class "dfm"
##        features
## docs    negative positive neg_positive neg_negative
##   text1        0        1            0            0
##   text2        0        4            0            0
##   text3        0        3            0            0
##   text4        1        2            0            0
##   text5        1        6            0            0
##   text6        2        1            0            0
```

```
docvars (corpus.dfm_cloud, "prop.neg.words") <- as.numeric(sentiment.dfm [,1])/ntoken(corpus.dfm_cloud)
docvars (corpus.dfm_cloud, "prop.pos.words") <- as.numeric(sentiment.dfm [,2])/ntoken(corpus.dfm_cloud)

docvars (corpus.dfm_cloud, "net.sentiment") <- docvars (corpus.dfm_cloud, "prop.pos.words") - docvars(co

net.sentiment <- docvars (corpus.dfm_cloud, "prop.pos.words") - docvars(corpus.dfm_cloud, "prop.neg.word

# cleaning, aggregating results on the company level and adding them to company data
net.sentiment.df <- data.frame("name"=comments$name, net.sentiment)
x <- na.omit(net.sentiment.df)
sent_company <- aggregate(x$net.sentiment, by=list(x$name), FUN=mean)
names(sent_company) <- c("name", "sentiment")
company_data$sentiment1 <- sent_company$sentiment


###############################################################################################
#Neal Caren dictionary


lexicon <- read.csv(file = "lexicon.csv",
                     header = TRUE,
                     stringsAsFactors = FALSE,
                     sep = ",")
#library(readr)
#lexicon <- read_csv("lexicon.csv")

#cols(
 # word = col_character(),
  #polarity = col_character()
names(lexicon) <- c("word", "sentiment")
lexicon <- as.dictionary(lexicon)

sentiment_2.dfm <- dfm(corpus.dfm_cloud, dictionary = lexicon)
dim(sentiment_2.dfm)
```

```
## [1] 145453      2
```

```
head(sentiment_2.dfm)
```

```
## Document-feature matrix of: 6 documents, 2 features (33.3% sparse).
## 6 x 2 sparse Matrix of class "dfm"
##        features
## docs    negative positive
##   text1        0        0
##   text2        0        1
##   text3        0        3
##   text4        1        2
##   text5        1        7
##   text6        1        1
```

```
docvars (corpus.dfm_cloud, "prop.neg.words_2") <- as.numeric(sentiment_2.dfm [,1])/ntoken(corpus.dfm_cl

docvars (corpus.dfm_cloud, "prop.pos.words_2") <- as.numeric(sentiment_2.dfm [,2])/ntoken(corpus.dfm_cl

docvars (corpus.dfm_cloud, "net.sentiment_2") <- docvars (corpus.dfm_cloud, "prop.pos.words_2") - docva
```

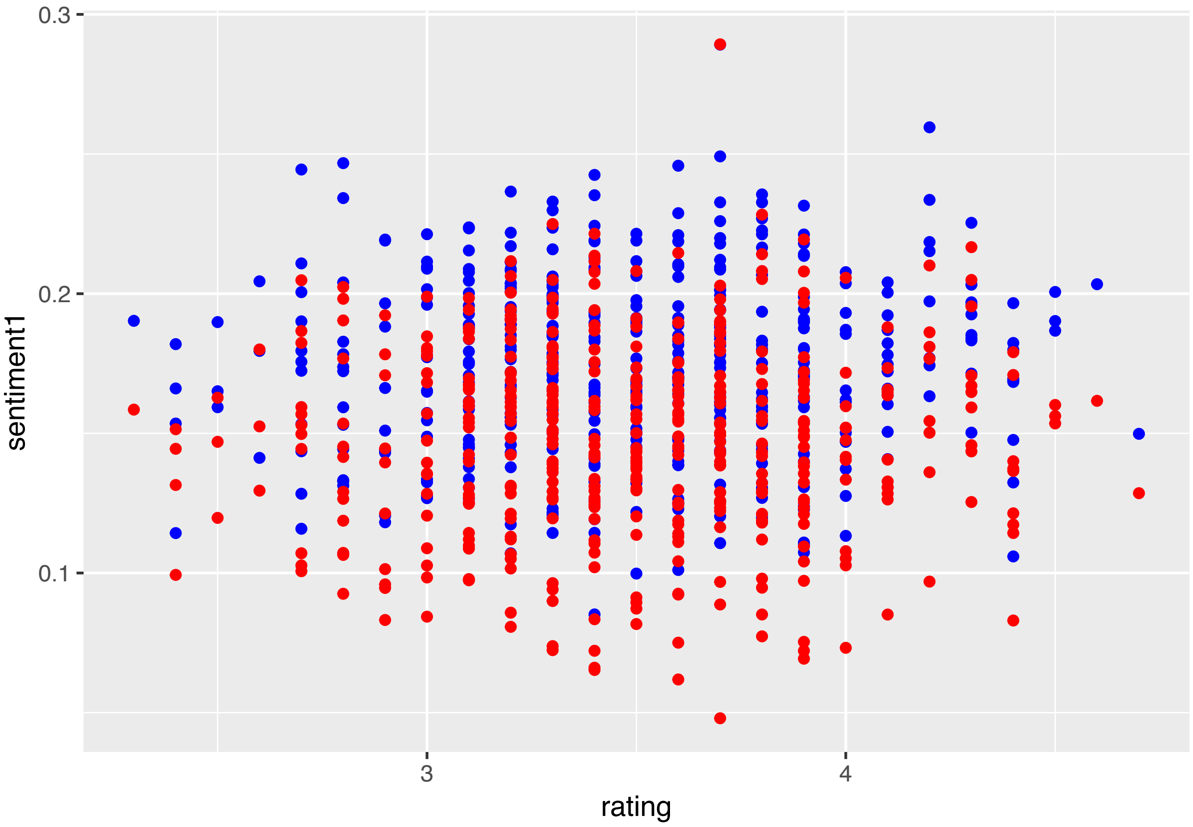
```
net.sentiment_2 <- docvars (corpus.dfm_cloud, "prop.pos.words_2") - docvars(corpus.dfm_cloud, "prop.neg

# cleaning, aggregating results on the company level and adding them to company data
net.sentiment_2.df <- data.frame("name"=comments$name, net.sentiment_2)
x <- na.omit(net.sentiment_2.df)
sent_company_2 <- aggregate(x$net.sentiment_2, by=list(x$name), FUN=mean)
company_data$sentiment2 <- sent_company_2$sentiment
```

```
## Warning in `[<-.data.table`(x, j = name, value = value): Adding new column
## 'sentiment2' then assigning NULL (deleting it).
```

```
names(sent_company_2) <- c("name", "sentiment")
company_data$sentiment2 <- sent_company_2$sentiment

#comparing results of the two dictionary
company_data$sent_dif <- company_data$sentiment1 - company_data$sentiment2
describe(company_data$sent_dif) #on average sentiments counted by the Lexicoder dictionary are 0.026 hi
```

```
## company_data$sent_dif
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      429        0      429        1  0.02758  0.01538 0.001668 0.010839
##      .25      .50      .75      .90      .95
## 0.020615 0.027658 0.036672 0.044068 0.049538
##
## lowest : -0.032459209 -0.015700318 -0.009919693 -0.009047057 -0.008625898
## highest:  0.060996938  0.062695885  0.063088135  0.069630212  0.077420206
```

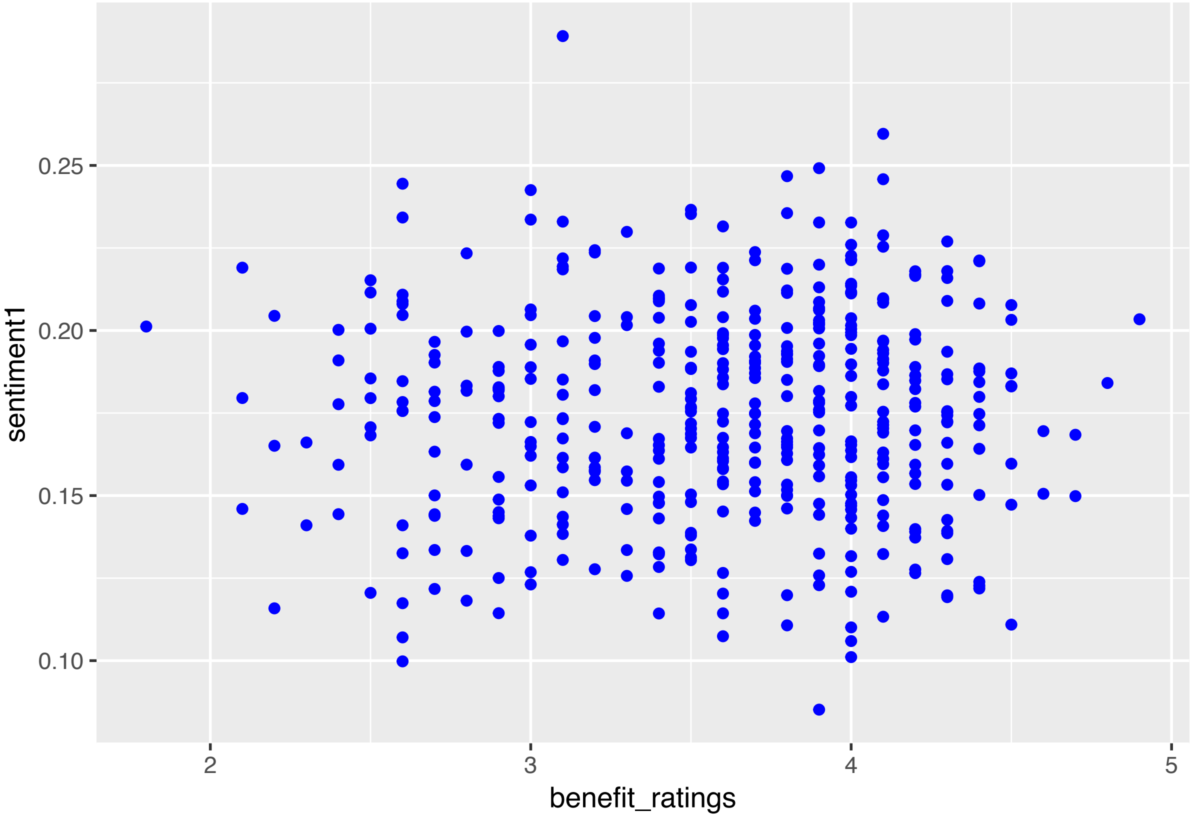
```
#Plot
sent.plot <- ggplot(company_data) + geom_point(aes(x = rating, y = sentiment1), color = "blue") + geom_
print(sent.plot)
```



First we checked the difference between sentiment results of the two dictionaries. Our results show that results counted by the Lexicoder dictionary are higher, by 0.026 on average. This graph shows the scatterplot of the average rating of companies for both sentiments: sentiments counted by the Lexicoder dictionary are blue, and sentiments counted by the Neal Caren dictionary are red. We can see that that the Lexicoder sentiments are higher on average, there are more blue dots in the upper part of the graph, whlie Neal Caren sentiments are lower on average, there are more red dots in the bottom part of the graph. From the graph it seems that there is no relationship between sentiments and general ratings of the company.

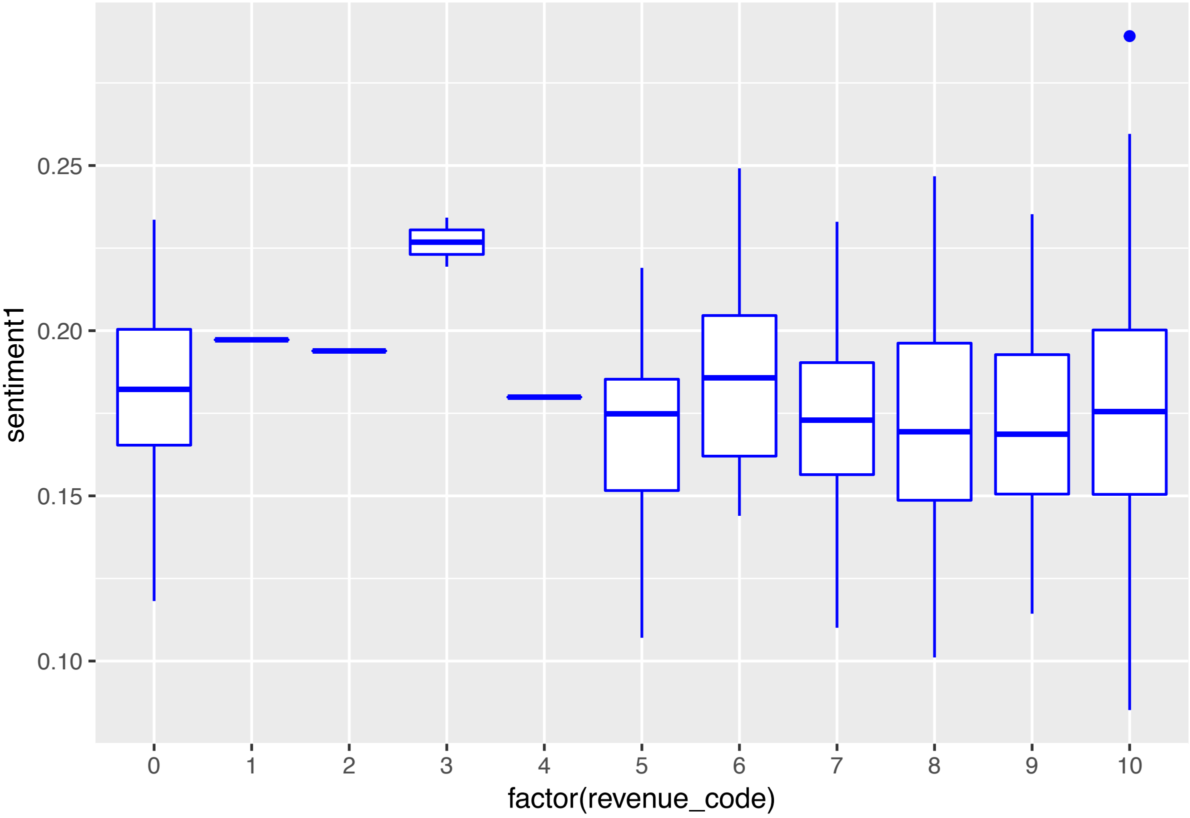
As there are no significant difference in the pattern of the sentiments counted by different dictionaries, in the following graphs we are going to plot the Lexicoder sentiments only.

```
#distribution of sentiments by benefit ratings
sent.plot2 <- ggplot(company_data, aes(x= benefit_ratings, y = sentiment1)) + geom_point(color = "blue"
print(sent.plot2)
```



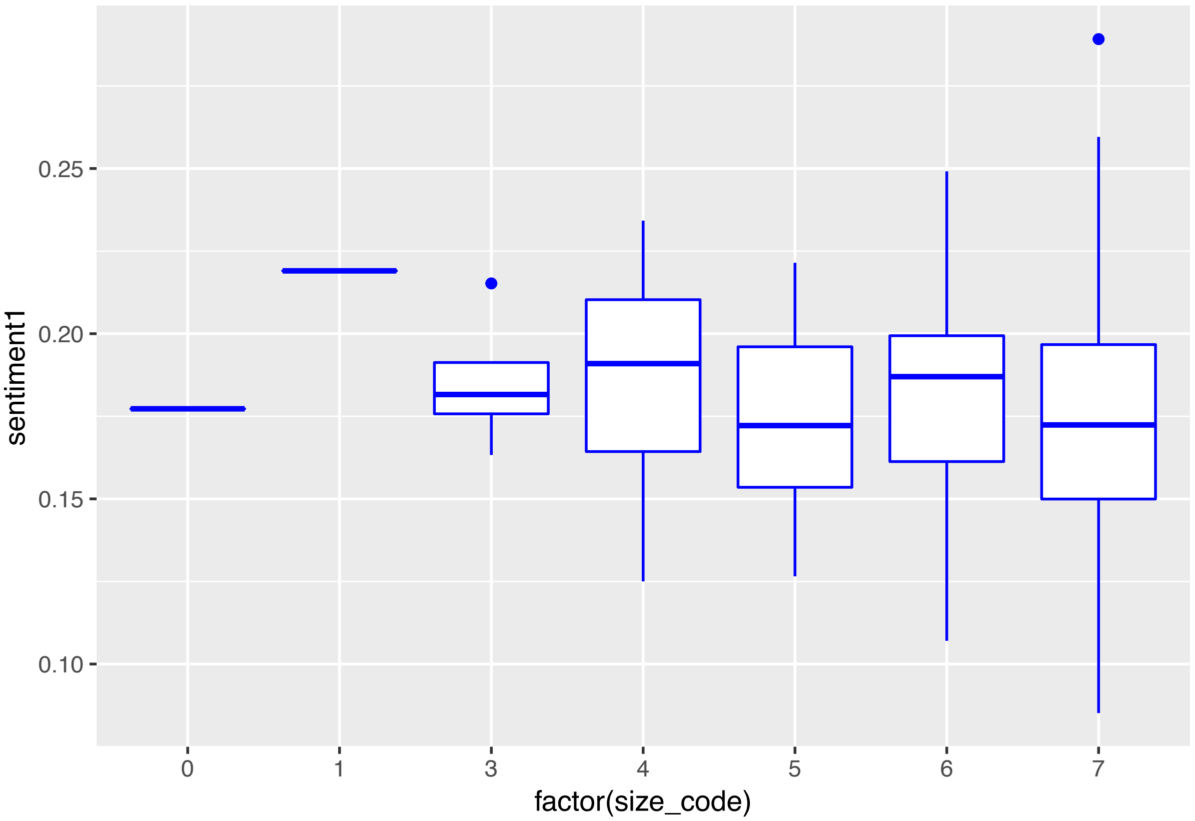
We expected that higher benefit ratings will correspond with more positive comments. However our results do not support this theory, it seems that there is no relationship between the two.

```
#distribution of sentiments by revenue
sent.plot3 <- ggplot(company_data, aes (x = factor(revenue_code), y = sentiment1)) + geom_boxplot(color
print(sent.plot3)
```



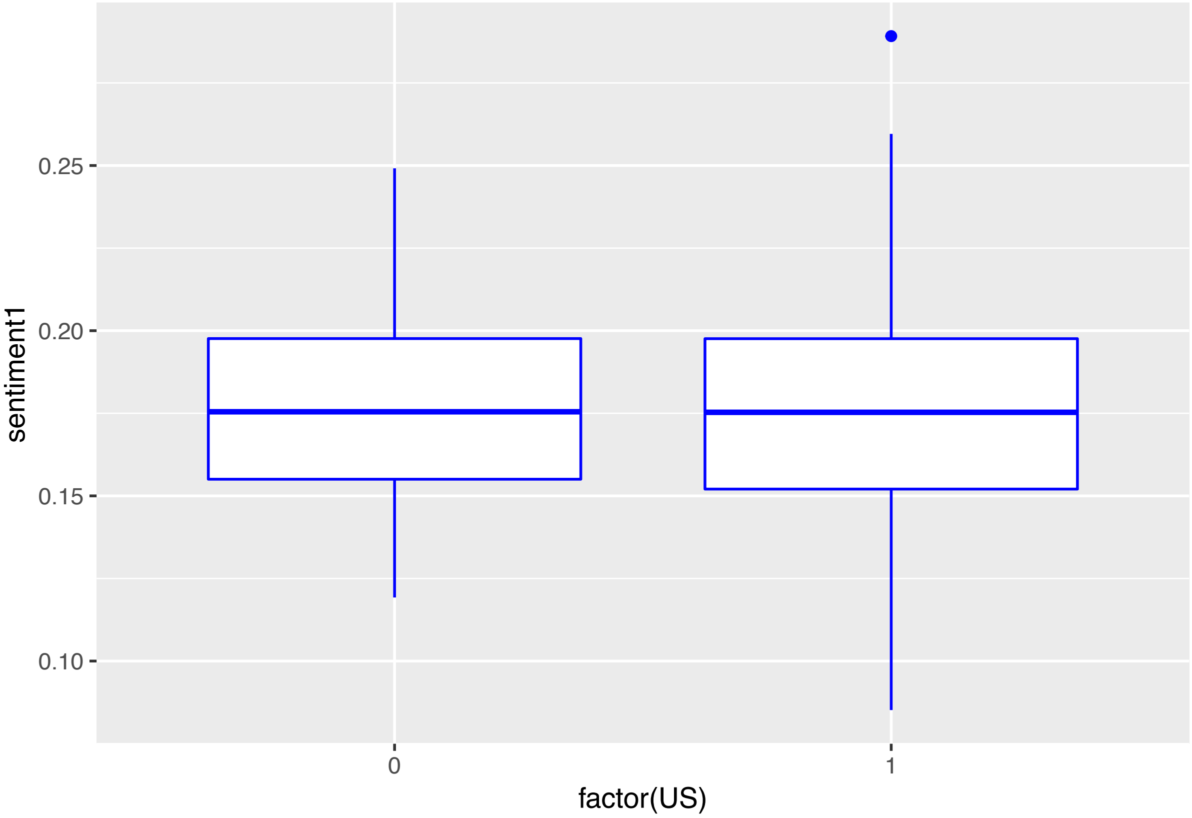
There is no relationship between revenue and sentiments, the median sentiment per revenues does not change significantly (except for some categories where the number of companies is only one or very few compared to the other groups). The median sentiment seem to be higher in companies with25-50 million AND 500 hundred to 1 billion AND 10billion + USD per a year.

```
#distribution of sentiments by size
sent.plot4 <- ggplot(company_data, aes (x = factor(size_code), y = sentiment1)) + geom_boxplot(color =
print(sent.plot4)
```



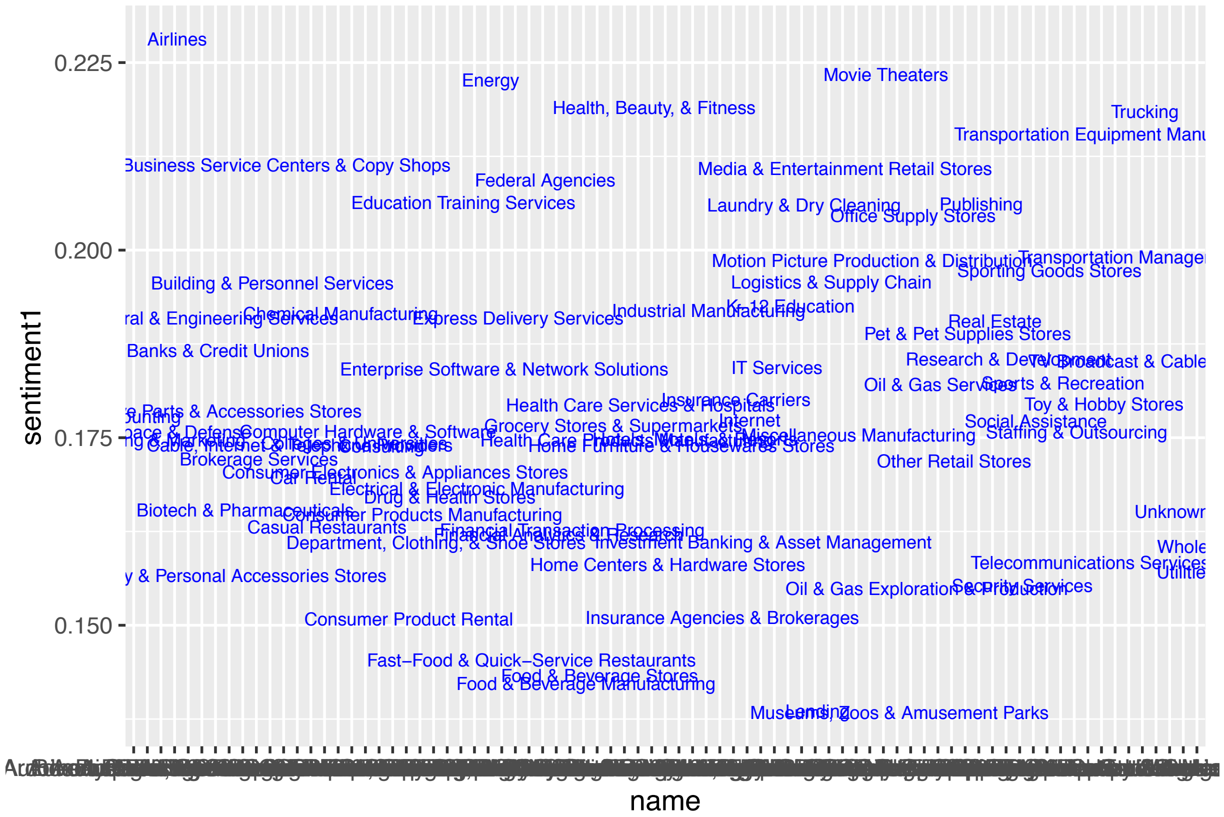
There is no linear pattern in the relationship between size and sentiments. In the first two categories the number of companies is to small to compare with the rest. The median sentiment is higher in companies between 501 and 1000 employews AND between 5001 and 10000 employees; and lower for companies between 1001 and 5000 AND 10000+ employees.

```r
#distribution of sentiments by US
sent.plot5 <- ggplot(company_data, aes (x = factor(US), y = sentiment1)) + geom_boxplot(color = "blue")
print(sent.plot5)
```



There is no difference in the median of sentiments between US and non-US companies, all are across 0.175.

```
#distribution of sentiments by Industry
#calculationg industry average
sentiment_a <- aggregate(company_data$sentiment1, by=list(company_data$industry),FUN = mean)
names(sentiment_a) <- c("name", "sentiment1")
#plot
sent.plot6 <- ggplot(sentiment_a, aes(x = name, y = sentiment1)) + geom_text(aes(label=name), size = 2.3
print(sent.plot6)
```



If we compare industry sentiment avereges we can see that majority of them are located in the lower part of the graph, which shows that they have lower sentiment results - it is important to mentioned that these are simple industry avarages, we did not weighted them by the number of companies. The following industries have the highest industry sentiment averages: Airlines, Energy, Movie Theaters, Health&Beauty and Fitness, Trucking and Transportation Equipment Manufacturing. These are high prestigous, well-paid or self-employed - so quite flexible - jobs. The following industries have the lowest industry sentiment averages: Fast-Food & Quick-Service Restaurant, Food & Bevarege Stores, Food & Bevarege Manufacturing, Museums & Zoos & Amusement Parks and Lending. These are lower-prestigious, low-paid jobs with stricter schedule and control and low work authority and individual decisions.

Topic models

```
#setting seed and creating a train and a test sample
set.seed(1)
docvars(corpus.dfm, "id") <- 1:ndoc(corpus)
id_train <- sample (1:nrow(corpus.dfm), nrow(corpus.dfm)*0.7, replace = FALSE)
train.dfm <- dfm_subset(corpus.dfm, id %in% id_train)
dim(train.dfm)
```

```
## [1] 101817   3035
```

```
test.dfm <- dfm_subset(corpus.dfm, !id %in% id_train)
train.lda.dfm <- convert(train.dfm, to = "topicmodels")
test.lda.dfm <- convert(test.dfm, to = "topicmodels")

#numer of topic models we have run is seven with the following number of topics: 2,3,5,6,7,8,9
#speeches.lda.2 <- LDA(train.lda.dfm,
                      #method = "Gibbs",
                      #k = 2)

#speeches.lda.4 <- LDA(train.lda.dfm,
                      # method = "Gibbs",
                      # k = 4)

#speeches.lda.6 <- LDA(train.lda.dfm,
                      # method = "Gibbs",
                      # k = 6)

#speeches.lda.5 <- LDA(train.lda.dfm,
                     #  method = "Gibbs",
                     # k = 5)

#speeches.lda.7 <- LDA(train.lda.dfm,
                       # method = "Gibbs",
                       # k = 7)

speeches.lda.8 <- LDA(train.lda.dfm,
                       method = "Gibbs",
                       k = 8)

#speeches.lda.9 <- LDA(train.lda.dfm,
                      # method = "Gibbs",
                      # k = 9)

#terms(speeches.lda.2)
#terms(speeches.lda.4)
#terms(speeches.lda.5)
#terms(speeches.lda.6)
#terms(speeches.lda.7)
#terms(speeches.lda.8)
#terms(speeches.lda.9)

#checking the 10 highest loading terms for topics for each models
#terms (speeches.lda.2, 10)
#terms (speeches.lda.4, 10)
#terms (speeches.lda.5, 10)
#terms (speeches.lda.6, 10)
#terms (speeches.lda.7, 10)
terms (speeches.lda.8, 10) # our choice
```

```
##       Topic 1   Topic 2  Topic 3     Topic 4   Topic 5   Topic 6
##  [1,] "benefit" "get"    "401k"      "pay"     "time"    "discount"
##  [2,] "packag"  "can"    "match"     "high"    "work"    "employe"
##  [3,] "compani" "take"   "plan"      "deduct"  "best"    "also"
##  [4,] "offer"   "manag"  "compani"   "cost"    "benefit" "free"
##  [5,] "great"   "use"    "stock"     "expens"  "employe" "perk"
##  [6,] "good"    "need"   "good"      "plan"    "thing"   "lot"
##  [7,] "pretti"  "realli" "bonus"     "coverag" "hour"    "well"
##  [8,] "averag"  "make"   "contribut" "famili"  "worst"   "reimburs"
##  [9,] "better"  "job"    "option"    "cover"   "part"    "servic"
## [10,] "overal"  "just"   "salari"    "low"     "full"    "program"
##       Topic 7   Topic 8 
##  [1,] "year"    "health"
##  [2,] "vacat"   "insur" 
##  [3,] "day"     "great" 
##  [4,] "paid"    "good"  
##  [5,] "week"    "dental"
##  [6,] "pto"     "medic" 
##  [7,] "leav"    "vision"
##  [8,] "sick"    "401k"  
##  [9,] "holiday" "care"  
## [10,] "3"       "option"
```

```
#terms (speeches.lda.9, 10)
```

We have run 7 topic models (with 2,4,5,6,7,8 and 9 topics). We only kept one that we believe is the most suitable for the final paper. We have choosen the model with 8 topics. It has the second lowest perplexity score (379) - however, the difference is very small between the perplexity scores of the different models. We can distinguish the 8 topics in the following way: 1. Overall idea of benefits (first word - benefit) 2. Action words (first word - get) 3. Salary and other income (first word 401k) 4. Types and effectiveness of health insurance coverage (first word pay) 5. Part time and hourly employee (first word time) 6. Perks and discounts (first word discount) 7. Time off (first word year) 8. What insurance is offered (first wolrd health)

(We have listed the first words in case it would change during the knitting)

```
#calculating the perplexity scores of the models
#perplexity(speeches.lda.2, test.lda.dfm) #436
#perplexity(speeches.lda.4, test.lda.dfm) #410
#perplexity(speeches.lda.5, test.lda.dfm) #401
#perplexity(speeches.lda.6, test.lda.dfm) #392
#perplexity(speeches.lda.7, test.lda.dfm) #385
perplexity(speeches.lda.8, test.lda.dfm) #379
```

```
## [1] 379.3558
```

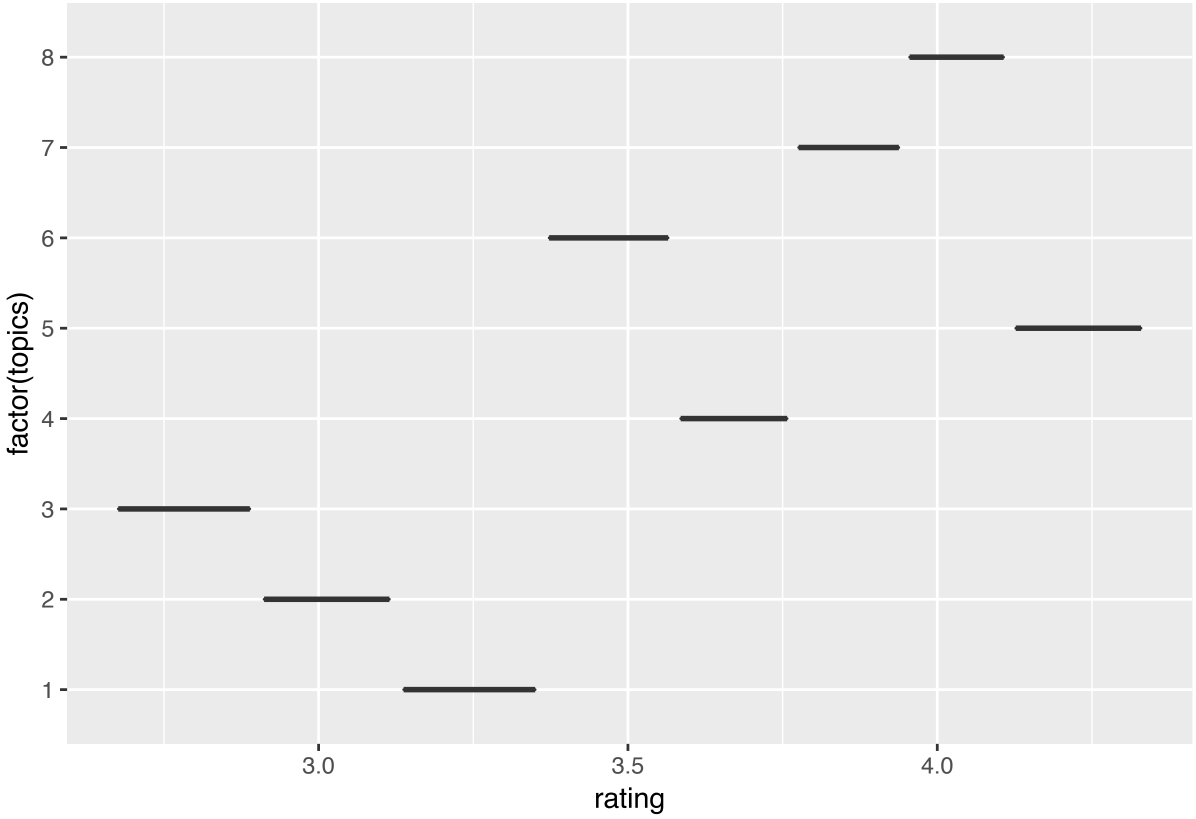
```
#perplexity(speeches.lda.9, test.lda.dfm) #375

#checking which topic load highest in the comments
#topics(speeches.lda.8)
#posterior(speeches.lda.8)$topics
df_top <- data.frame(topics(speeches.lda.8))
#checking the topic prevelance
table(unlist(df_top))
```

```
## 
##     1     2     3     4     5     6     7     8 
## 19388 13114 14313 11543 11800 10455 11042 10153
```

We can see that the first topic (Overall idea of benefits) has the greatest number of topic prevalence, the third topic (Salary and other income) has the second greatest and the second topic (Action words) has the third greatest number of topic prevalence. The remaining 5 topics distribution is relativly equal - the difference between the greatest (19388) and the lowest (10153) number is not as big, we can say that the distribution of the most prevalent topics are quite ballanced.

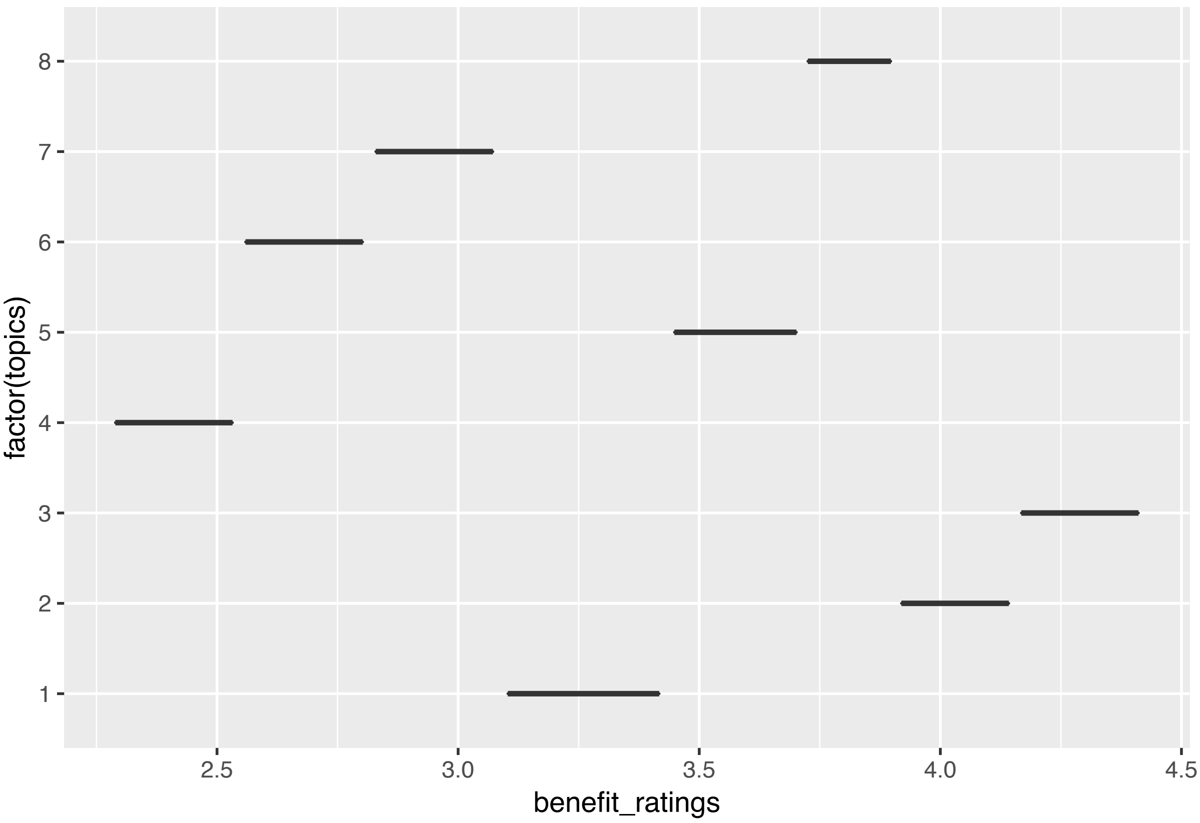
```
# cleaning and aggregating the results and adding to company dataset
#getting and cleaning rownames to get which rows were in the sample and matching with the dataset
zzz <- rownames(df_top)
row_id <- data.frame(gsub("text", "", zzz))
comments$id <- 1:nrow(comments)
```

```r
joo <- comments[match(row_id$gsub..text.......zzz., comments$id),]
jo.df <- data.frame("name"= joo$name, df_top)

#aggregating and calculating mode (the value that appears most often) since there is no bulit in functi
getmode <- function(v) {
            uniqv <- unique(v)
            uniqv[which.max(tabulate(match(v,uniqv)))]}

topics.df <- aggregate(jo.df$topics.speeches.lda.8 ~ jo.df$name, jo.df, getmode)
names(topics.df) <- c("name", "topic")
company_data$topics <- topics.df$topic

saveRDS(company_data, 'company_data_v2.rds')

#most prevalent topics by Industry (it is not as demonstrative in this case, because topics is a catego
#topics_a <- aggregate(company_data$topics ~ company_data$industry, company_data, getmode)
#names(topics_a) <- c("name", "topics")

#Plots
#distribution of most frequent topics across overal company ratings
top.plot <- ggplot(company_data) + geom_boxplot (aes(x = rating, y = factor(topics)))
print(top.plot)
```



Since topics variable is categorical, we can compare it with overal company ratings and benefit scores. Topic 3 (Salary and other incomes) is associated with the lowest overal score and topic 5 (Part time and hourly employees) is associated with the highest overal score. It is possible that those employees commented more who are not satesfied with thier compensation package, and those who are satesfied with their flexible working conditions.

```
#distribution of topics by benefit ratings
top.plot2 <- ggplot(company_data, aes(x= benefit_ratings, y = factor(topics))) + geom_boxplot()
print(top.plot2)
```



Interestingly the occurance of topic 3 (Salary and other incomes) is correlated with the highest benefit ratings while it was associated with the lowest overal company scores.Topic 4 (Types and effectiveness of health insurance coverage) has the highest prevalence among companies with the lowest benefit ratings.

Predictive Analysis

```
# linear model y: overal company rating x: sentiment1 and readibility, controlling for revenue, size, i
linearmodel1=lm(rating ~ sentiment1 + readib + revenue_code + size_code + industry + location, data=comp
summary(linearmodel1)
```

```
##
## Call:
## lm(formula = rating ~ sentiment1 + readib + revenue_code + size_code +
##     industry + location, data = company_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0925 -0.1480  0.0000  0.1756  0.9089
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                     3.109080   0.548057
## sentiment1                                     -0.503336   0.703666
## readib                                          0.066232   0.039406
## revenue_code                                    0.010950   0.007871
## size_code                                      -0.076589   0.029044
## industry Advertising & Marketing               -0.587682   0.217303
## industry Aerospace & Defense                   -0.004136   0.215162
## industry Airlines                               0.318035   0.253185
## industry Architectural & Engineering Services  -0.620971   0.409798
## industry Automotive Parts & Accessories Stores -0.483330   0.327396
## industry Banks & Credit Unions                 -0.075995   0.213464
## industry Beauty & Personal Accessories Stores  -0.569748   0.273397
## industry Biotech & Pharmaceuticals              0.066120   0.212397
## industry Brokerage Services                    -0.160538   0.363586
## industry Building & Personnel Services         -0.454285   0.326751
## industry Business Service Centers & Copy Shops -0.550668   0.414767
## industry Cable, Internet & Telephone Providers -0.563159   0.232211
## industry Car Rental                            -0.054100   0.447543
## industry Casual Restaurants                    -0.304111   0.193286
## industry Chemical Manufacturing                 0.094012   0.335689
## industry Colleges & Universities                0.407024   0.194743
## industry Computer Hardware & Software          -0.130389   0.172167
## industry Consulting                             0.071354   0.190860
## industry Consumer Electronics & Appliances Stores -0.851321   0.327994
## industry Consumer Product Rental               -0.744580   0.350821
## industry Consumer Products Manufacturing        0.049649   0.252300
## industry Department, Clothing, & Shoe Stores   -0.538303   0.188876
## industry Drug & Health Stores                  -0.731315   0.317307
## industry Education Training Services           -0.339982   0.418017
## industry Electrical & Electronic Manufacturing  0.097439   0.314363
## industry Energy                                 0.414327   0.444061
## industry Enterprise Software & Network Solutions -0.025140   0.188852
## industry Express Delivery Services             -0.119748   0.423533
## industry Fast-Food & Quick-Service Restaurants -0.354564   0.243444
## industry Federal Agencies                       0.050375   0.451711
## industry Financial Analytics & Research         0.095870   0.361114
## industry Financial Transaction Processing      -0.221777   0.250602
## industry Food & Beverage Manufacturing         -0.223445   0.222583
## industry Food & Beverage Stores                -0.022867   0.425109
## industry Grocery Stores & Supermarkets         -0.492115   0.221667
## industry Health Care Products Manufacturing     0.064947   0.287843
## industry Health Care Services & Hospitals      -0.346515   0.178796
## industry Health, Beauty, & Fitness             -1.158667   0.456118
## industry Home Centers & Hardware Stores        -0.391571   0.250708
## industry Home Furniture & Housewares Stores    -0.587907   0.226367
## industry Hotels, Motels, & Resorts              0.284561   0.222695
## industry Industrial Manufacturing               0.064250   0.274991
## industry Insurance Agencies & Brokerages       -0.790263   0.315424
## industry Insurance Carriers                    -0.198994   0.189508
## industry Internet                              -0.207767   0.204413
## industry Investment Banking & Asset Management -0.277545   0.194076
## industry IT Services                           -0.455009   0.188891
## industry K-12 Education                         0.538720   0.412767
## industry Laundry & Dry Cleaning                -0.130449   0.416444
## industry Lending                                    0.023167   0.408599
## industry Logistics & Supply Chain                  -0.461864   0.305224
## industry Media & Entertainment Retail Stores       -0.796810   0.571136
## industry Miscellaneous Manufacturing               -0.105507   0.437176
## industry Motion Picture Production & Distribution -0.737907   0.407367
## industry Movie Theaters                            -0.808051   0.592620
## industry Museums, Zoos & Amusement Parks           -0.514290   0.418464
## industry Office Supply Stores                      -0.726589   0.420762
## industry Oil & Gas Exploration & Production         0.083933   0.276926
## industry Oil & Gas Services                        -0.619576   0.319515
## industry Other Retail Stores                       -0.755623   0.210768
## industry Pet & Pet Supplies Stores                 -0.390740   0.438180
## industry Publishing                                -0.152315   0.317981
## industry Real Estate                               -0.435755   0.252390
## industry Research & Development                    -0.055912   0.442731
## industry Security Services                         -0.928612   0.315976
## industry Social Assistance                         -0.628937   0.268132
## industry Sporting Goods Stores                     -0.206586   0.329533
## industry Sports & Recreation                       -0.774787   0.318201
## industry Staffing & Outsourcing                    -0.272868   0.190341
## industry Telecommunications Services               -0.575348   0.249640
## industry Toy & Hobby Stores                        -0.994931   0.418924
## industry Transportation Equipment Manufacturing   -0.049229   0.324900
## industry Transportation Management                 -0.050299   0.278391
## industry Trucking                                   0.151784   0.513849
## industry TV Broadcast & Cable Networks             -0.403022   0.308480
## industry Unknown                                   -0.215038   0.205792
## industry Utilities                                 -0.078202   0.321008
## industry Wholesale                                 -0.453764   0.250522
## locationAL                                          0.328377   0.581685
## locationAR                                          0.467987   0.530083
## locationAustralia                                   0.825683   0.574478
## locationAZ                                          0.372963   0.452617
## locationCA                                          0.684868   0.425334
## locationCanada                                      0.428254   0.464743
## locationChina                                       0.818069   0.594598
## locationCO                                          0.450943   0.450096
## locationCT                                          0.471105   0.446129
## locationDC                                          0.350612   0.459643
## locationDE                                          0.356626   0.633810
## locationFL                                          0.522700   0.427952
## locationFrance                                      0.241171   0.458540
## locationGA                                          0.652417   0.442710
## locationGermany                                     0.872705   0.471657
## locationIA                                          0.987186   0.584484
## locationID                                          0.232815   0.565030
## locationIL                                          0.492809   0.429756
## locationIN                                          0.480782   0.475905
## locationIndia                                       0.456241   0.565110
## locationIreland                                     0.379951   0.535138
## locationIsrael                                     -0.026702   0.581803
## locationJapan                                       0.517505   0.509976
## locationKS                                          0.751286   0.595282
## locationKY                                          0.733226   0.502635
## locationLA                                        0.347036   0.588266
## locationLuxembourg                                0.188694   0.506327
## locationMA                                        0.620010   0.431578
## locationMD                                        0.557462   0.462111
## locationMI                                        0.535991   0.459884
## locationMN                                        0.866229   0.454434
## locationMO                                        0.486688   0.459196
## locationNC                                        0.425576   0.448213
## locationNE                                        0.633808   0.525005
## locationNetherlands                               0.411799   0.458113
## locationNJ                                        0.766693   0.451582
## locationNY                                        0.841960   0.421507
## locationOH                                        0.489401   0.434840
## locationOK                                        0.980711   0.478086
## locationOR                                        0.698183   0.596120
## locationPA                                        0.618221   0.441933
## locationRI                                        0.066286   0.567422
## locationSC                                        0.196588   0.503401
## locationSpain                                    -0.254489   0.582702
## locationSweden                                    0.947618   0.482174
## locationSwitzerland                               0.442730   0.485233
## locationTN                                        0.533349   0.441028
## locationTX                                        0.632425   0.430044
## locationUnited Kingdom                            0.431497   0.438753
## locationUT                                        0.889547   0.525238
## locationVA                                        0.457879   0.437504
## locationWA                                        0.945214   0.446147
## locationWI                                        0.678121   0.436836
##                                                  t value Pr(>|t|)
## (Intercept)                                        5.673 3.37e-08 ***
## sentiment1                                        -0.715 0.474990
## readib                                             1.681 0.093875 .
## revenue_code                                       1.391 0.165182
## size_code                                         -2.637 0.008811 **
## industry Advertising & Marketing                  -2.704 0.007242 **
## industry Aerospace & Defense                      -0.019 0.984677
## industry Airlines                                  1.256 0.210067
## industry Architectural & Engineering Services     -1.515 0.130772
## industry Automotive Parts & Accessories Stores    -1.476 0.140941
## industry Banks & Credit Unions                    -0.356 0.722089
## industry Beauty & Personal Accessories Stores     -2.084 0.038030 *
## industry Biotech & Pharmaceuticals                 0.311 0.755792
## industry Brokerage Services                       -0.442 0.659149
## industry Building & Personnel Services            -1.390 0.165490
## industry Business Service Centers & Copy Shops    -1.328 0.185325
## industry Cable, Internet & Telephone Providers    -2.425 0.015904 *
## industry Car Rental                               -0.121 0.903866
## industry Casual Restaurants                       -1.573 0.116713
## industry Chemical Manufacturing                    0.280 0.779631
## industry Colleges & Universities                   2.090 0.037475 *
## industry Computer Hardware & Software             -0.757 0.449455
## industry Consulting                                0.374 0.708782
## industry Consumer Electronics & Appliances Stores  -2.596 0.009920 **
## industry Consumer Product Rental                  -2.122 0.034643 *
## industry Consumer Products Manufacturing           0.197 0.844131
## industry Department, Clothing, & Shoe Stores       -2.850 0.004682 **
## industry Drug & Health Stores                      -2.305 0.021879 *
## industry Education Training Services               -0.813 0.416695
## industry Electrical & Electronic Manufacturing      0.310 0.756814
## industry Energy                                     0.933 0.351567
## industry Enterprise Software & Network Solutions   -0.133 0.894188
## industry Express Delivery Services                 -0.283 0.777579
## industry Fast-Food & Quick-Service Restaurants     -1.456 0.146339
## industry Federal Agencies                           0.112 0.911279
## industry Financial Analytics & Research             0.265 0.790822
## industry Financial Transaction Processing          -0.885 0.376895
## industry Food & Beverage Manufacturing             -1.004 0.316269
## industry Food & Beverage Stores                    -0.054 0.957138
## industry Grocery Stores & Supermarkets             -2.220 0.027179 *
## industry Health Care Products Manufacturing         0.226 0.821643
## industry Health Care Services & Hospitals          -1.938 0.053578 .
## industry Health, Beauty, & Fitness                 -2.540 0.011593 *
## industry Home Centers & Hardware Stores            -1.562 0.119400
## industry Home Furniture & Housewares Stores        -2.597 0.009875 **
## industry Hotels, Motels, & Resorts                  1.278 0.202329
## industry Industrial Manufacturing                   0.234 0.815425
## industry Insurance Agencies & Brokerages           -2.505 0.012774 *
## industry Insurance Carriers                        -1.050 0.294558
## industry Internet                                  -1.016 0.310275
## industry Investment Banking & Asset Management     -1.430 0.153758
## industry IT Services                               -2.409 0.016620 *
## industry K-12 Education                             1.305 0.192868
## industry Laundry & Dry Cleaning                    -0.313 0.754317
## industry Lending                                    0.057 0.954824
## industry Logistics & Supply Chain                  -1.513 0.131307
## industry Media & Entertainment Retail Stores       -1.395 0.164032
## industry Miscellaneous Manufacturing               -0.241 0.809463
## industry Motion Picture Production & Distribution  -1.811 0.071102 .
## industry Movie Theaters                            -1.364 0.173765
## industry Museums, Zoos & Amusement Parks           -1.229 0.220061
## industry Office Supply Stores                      -1.727 0.085250 .
## industry Oil & Gas Exploration & Production         0.303 0.762038
## industry Oil & Gas Services                        -1.939 0.053447 .
## industry Other Retail Stores                       -3.585 0.000395 ***
## industry Pet & Pet Supplies Stores                 -0.892 0.373267
## industry Publishing                                -0.479 0.632291
## industry Real Estate                               -1.727 0.085309 .
## industry Research & Development                    -0.126 0.899590
## industry Security Services                         -2.939 0.003556 **
## industry Social Assistance                         -2.346 0.019662 *
## industry Sporting Goods Stores                     -0.627 0.531210
## industry Sports & Recreation                       -2.435 0.015493 *
## industry Staffing & Outsourcing                    -1.434 0.152761
## industry Telecommunications Services               -2.305 0.021882 *
## industry Toy & Hobby Stores                        -2.375 0.018194 *
## industry Transportation Equipment Manufacturing    -0.152 0.879670
## industry Transportation Management                 -0.181 0.856745
## industry Trucking                                   0.295 0.767908
## industry TV Broadcast & Cable Networks            -1.306 0.192415
## industry Unknown                                  -1.045 0.296917
## industry Utilities                                -0.244 0.807700
## industry Wholesale                                -1.811 0.071122 .
## locationAL                                         0.565 0.572827
## locationAR                                         0.883 0.378038
## locationAustralia                                  1.437 0.151707
## locationAZ                                         0.824 0.410601
## locationCA                                         1.610 0.108434
## locationCanada                                     0.921 0.357555
## locationChina                                      1.376 0.169923
## locationCO                                         1.002 0.317227
## locationCT                                         1.056 0.291845
## locationDC                                         0.763 0.446201
## locationDE                                         0.563 0.574089
## locationFL                                         1.221 0.222917
## locationFrance                                     0.526 0.599318
## locationGA                                         1.474 0.141639
## locationGermany                                    1.850 0.065277 .
## locationIA                                         1.689 0.092285 .
## locationID                                         0.412 0.680611
## locationIL                                         1.147 0.252434
## locationIN                                         1.010 0.313210
## locationIndia                                      0.807 0.420120
## locationIreland                                    0.710 0.478266
## locationIsrael                                    -0.046 0.963425
## locationJapan                                      1.015 0.311056
## locationKS                                         1.262 0.207928
## locationKY                                         1.459 0.145702
## locationLA                                         0.590 0.555692
## locationLuxembourg                                 0.373 0.709661
## locationMA                                         1.437 0.151895
## locationMD                                         1.206 0.228661
## locationMI                                         1.165 0.244768
## locationMN                                         1.906 0.057606 .
## locationMO                                         1.060 0.290076
## locationNC                                         0.949 0.343151
## locationNE                                         1.207 0.228313
## locationNetherlands                                0.899 0.369443
## locationNJ                                         1.698 0.090608 .
## locationNY                                         1.997 0.046695 *
## locationOH                                         1.125 0.261309
## locationOK                                         2.051 0.041124 *
## locationOR                                         1.171 0.242465
## locationPA                                         1.399 0.162900
## locationRI                                         0.117 0.907083
## locationSC                                         0.391 0.696436
## locationSpain                                     -0.437 0.662622
## locationSweden                                     1.965 0.050324 .
## locationSwitzerland                                0.912 0.362305
## locationTN                                         1.209 0.227511
## locationTX                                         1.471 0.142471
## locationUnited Kingdom                             0.983 0.326191
## locationUT                                         1.694 0.091403 .
## locationVA                                          1.047 0.296160
## locationWA                                          2.119 0.034964 *
## locationWI                                          1.552 0.121659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3724 on 293 degrees of freedom
## Multiple R-squared:  0.5156, Adjusted R-squared:  0.2924
## F-statistic:  2.31 on 135 and 293 DF,  p-value: 1.548e-09
```

```
#print(linearmodel1)

#2 linear model y: overal company rating x: sentiment2 and readibility, controlling for revenue, size,
linearmodel2=lm(rating ~ sentiment2 + readib + revenue_code + size_code + industry + location, data=comp
#summary(linearmodel2)
#print(linearmodel2)
```

Call: lm(formula = rating ~ sentiment2 + readib + revenue_code + size_code + industry + location, data = company_data)

Residuals: Min 1Q Median 3Q Max -1.0909 -0.1496 0.0000 0.1749 0.9035

Neither sentiment1 and sentiment2 nor readibility were significant in a 0.05 p level. We can conclude from the results of the linear regression that we find that there is no relationship between sentiments and readibility and overal company scores. Readibility is significant on a 0.1 p level. Our significance criteria is 0.05, but if we would accept it, we could say that one unit increase in readibility scrore on average will imply 0.06 increase in the overal company ratings. It means that less complicated comments are associated with higher ratings.

Interestingly, we find no significant relationship between benefit ratings and the sentiments or the readibility score. Benefit ratings express the quantified, general view about the benefits offered by the company what employees can describe in more details in the comments. Thus, we expected a strong relationship between the sentiments and benefit ratings.

```
#3 Just for curiosity, checking the relationship between benefit ratings and the sentiment1 and readibi
linearmodel3=lm(benefit_ratings ~ sentiment1 + readib + revenue_code + size_code + industry + location,
#summary(linearmodel3)
#print(linearmodel3)

### ANOVA for the topic models (categorical variable) and the dependent variables, we need to do an ANO
#factoring
company_data$topics_factor <- factor(company_data$topics)
#table(company_data$rating, company_data$topics_factor)
anova <- aov(rating ~ topics_factor, data = company_data)
summary(anova)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## topics_factor    7   1.10  0.1575   0.801  0.587
## Residuals      421  82.76  0.1966
```

We can examine the relationship between the topics and the ratings by ANOVA analysis, becasue topics is categorical and the ratings are numberic variables. The P value (0.587) is not significant, so we can conclude that there is no statistical, significant relationship between which topic is prevalent and the overal company ratings. There is no relationship between benefit ratings and the topics either.

```
# checking relationship between benefit ratings and topics
anova1 <- aov(benefit_ratings ~ topics_factor, data = company_data)
summary(anova1)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## topics_factor   7   1.93  0.2752   0.806  0.583
## Residuals     421 143.83  0.3416
```

```
#table(company_data$benefit_ratings, company_data$topics_factor)
```

Conclusion

We tried to predict employee satisfaction (overal company ratings) by analyzing benefit comments by three different methodology: 1) readibility scores 2)sentiment analysis 3) topic modeling. Despite the fact that none of them proved to be a good predictor, we had interesting findings: The readibiliy score of glassdoor commenters is quite low (which means they use complex language). The sentiment analysis shows that company commenst are rather negative than positive and there is no relationship between the sentiments and the benefit ratings. Finally, we were able to distinguish 8 different topics by LDA modelling: Overall idea of benefits, Action words, Salary and other income, Types and effectiveness of health insurance coverage, Part time and hourly employee, Perks and discounts, Time off, and What insurance is offered. Discovering the reasons of the lack of significant relationship between the text analyis results and the overal company rating would imply more reasearch and the modification of our model. One limitation of our data that we had only aggragated company data availabe, these patters could be examined better comparing comments and ratings given by the individuals. We have lost a lot of the variance by the aggregation. As regards to the sentiment analysis, we used two pre-prepared dictionary, which were developed by analyzing mostly political text, it is possible that benefits comments are such a special field that a tailored dictionary could better identify its sentiments.