
SCHOOL OF ENGINEERING AND TECHNOLOGY

FINAL ASSESSMENT FOR THE BSC (HONS) INFORMATION SYSTEMS (BUSINESS ANALYTICS); YEAR 3

ACADEMIC SESSION AUGUST 2021; SEMESTER 7, 8, 9

IST2334: WEB AND NETWORK ANALYTICS

DEADLINE: 3rd DECEMBER 2021 4:00PM

GROUP: _____ 17 _____

INSTRUCTIONS TO CANDIDATES

- This project will contribute 50% to your final grade.
- This is a group project. Each group consists of 4-5 members.

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Students' declaration:

(Name) (ID) (Signature) We

- 1) Lim Wei Zheng (19033067) LWZ
- 2) Lee Jia Yin (19027879) LJY
- 3) Ngoi Yi Wen (19028992) NYW
- 4) Thejal A/P L.Ramesh (18004036) TR
- 5) Yashinnie A/P R Ganeswaran (19022649) YG

received the assignment and read the comments.

Academic Honesty Acknowledgement

“We (names stated above) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties (*refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme*) for any kind of copying or collaboration on any assignment.”

1) Lim Wei Zheng (19033067)

2) Lee Jia Yin (19027879)

3) Ngoi Yi Wen (19028992)

4) Thejal A/P L.Ramesh (18004036)

5) Yashinnie A/P R Gnaneswaran (19022649)

.....ACGTY.../ 12/2/2021..... (Student's signature / Date)

1.0 Introduction	4
2.0 Elaboration of the data sets	5
Summary Statistics & Graph	7
3.0 Presentation of Analyses	10
(i) Sentiment Analysis	10
(ii) Social Network Analysis	12
(iii) Wordcloud	19
4.0 Coding In R	22
5.0 Lesson & Conclusion	28
6.0 Reflection	29
19027879 Lee Jia Yin	29
19028992 Ngoi Yi Wen	29
19033067 Lim Wei Zheng	30
19022649 Yashinnie A/P R Gnaneswaran	30
18004036 Thejal A/P L.Ramesh	31


1.0 Introduction

For this assignment, we are going to analyze a dataset about the reviews of various wine brands by users. Wine is the second most popular alcoholic drink in the world behind beer. As taste and scent are so intricately linked to an individual's own preference, wine tasting will always have a subjective and personal quality. A majority of people would love to drink wine to unwind after a long tiring day, some would have it during a celebration or even for fun. Thus, wine shall be appreciated by truly tasting the texture, scent and quality. Therefore, this analysis is to explore the opinions of different users after tasting different brands of wine and other alcohol such as beer. Sentiment analysis is a submachine learning task of mining reviews, opinions, and emotions from the text or data. It classifies the text of positive, neutral and negative emotion. We will be working on detecting subjectively that is to determine the positive and negative sentiment of the dataset, social network analysis and further determine the polarity score also called sentiment prediction. The dataset used to explore, understand and analyze for this assignment is the wine reviews dataset from kaggle, <https://www.kaggle.com/krrai77/wine-reviews>.

In addition to that, sentiment analysis research has been increasing tremendously due to the wide range of business and social applications. Techniques employed by sentiment analysis models can be broadly categorized into machine learning and semantic orientation approaches. The motivation behind the wine review dataset is to transform the wine reviews from data to useful information and understand the data frames. Besides that, to identify the opinion polarities (positive or negative) expressed on various wine brands and the opinion polarities of reviews or sentences. Furthermore, to find the best wine brand based on reviews and ratings. Hence, we were curious to explore the word cloud for the wine dataset to look at the frequency of words and study the most dominant wine brand among a vast number of users, based on their positive and negative reviews. This will also be beneficial to us, to be familiar with various wine brands out in the market and increase the popularity of the brand by analyzing the top most dominant wine brand.

2.0 Elaboration of the data sets

For the wine review dataset, it consists of 2551 reviews especially for wines. The variables for this data set are SI.No., Brand, Name, Reviews Data Added, Reviews do Recommend, Reviews Num Helpful, Reviews Rating, Reviews Text and Weight.

# Reviews Rating	Reviews Text	Reviews Title
 1 5	2551 unique values	Five Stars 2% [null] 2% Other (2780) 96%
5	This a fantastic white wine for any occasion!	My Favorite White Wine
5	Tart, not sweet...very refreshing and delicious!	Yum!!
5	I was given this wine so it was a delightful surprise to find that it has a flavorful and delicious ...	A New Favorite!
5	This is a phenomenal wine and my new favorite red.	Bold, Flavorful, Aromatic, Delicious
5	4 750ml bottles for the price of two With way less packaging YES PLEASE! I was nervous it was too go...	Yum! Plus, Environmentally Friendly!

From the dataset, we can see that the dataset consists of reviews text and reviews title , this links us to our motivation and interest which is to do sentiment analysis. The reviews text is filled with bags of words that describe the flavour of the wine. We can utilise all these datas to determine whether the text is positive or negative. We also use reviews for social network analysis to see the connection between specific words.

▲ Brand		▲ Name	
Carmex	36%	Carmex Lip Balm ...	36%
Master of Mixes	5%	Bittermens Xocolatl ...	4%
Other (1707)	59%	Other (1717)	59%
Gallo		Ecco Domani174 Pinot Grigio - 750ml Bottle	
Fresh Craft Co.		Fresh Craft174 Mango Citrus - 4pk / 250ml Bottle	
1000 Stories		1000 Stories174 Zinfandel - 750ml Bottle	
1000 Stories		1000 Stories174 Zinfandel - 750ml Bottle	

From here we can see the variable brand, this also brings us to our next motivation to use this dataset which is to produce word clouds. This variable allows us to visualise frequent words in a text where the size of words can represent their frequency using Rstudio.

Summary Statistics & Graph

```
> dim(wine_reviews)
[1] 2890 10
> str(wine_reviews)
'data.frame': 2890 obs. of 10 variables:
 $ Sl.No.      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Brand       : chr   "Gallo" "Fresh Craft Co." "1000 Stories" "1000 Stories" ...
 $ Name        : chr   "Ecco Domani174 Pinot Grigio - 750ml Bottle" "Fresh Craft174 Mango Citrus - 4pk / 250ml Bottle"
 $ Reviews.Date.Added : chr   "2018-01-09T13:24:04Z" "2018-01-09T17:31:52Z" "2018-01-09T17:31:51Z" "2017-10-04T18:03:12Z" ...
 $ Reviews.do.Recommend: logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ Reviews.Num.Helpful : int  1 NA NA NA 1 NA 1 1 0 NA ...
 $ Reviews.Rating  : int  5 5 5 5 5 5 3 2 5 5 ...
 $ Reviews.Text    : chr   "This a fantastic white wine for any occasion!" "Tart, not sweet...very refreshing and delicious
and delicious taste! A new favorite!!!!" "This is a phenomenal wine and my new favorite red." ...
 $ Reviews.Title    : chr   "My Favorite white wine" "Yum!!" "A New Favorite!" "Bold, Flavorful, Aromatic, Delicious" ...
 $ Weight          : chr   "1.0 lbs" "2.45 lbs" "3.09 lbs" "3.09 lbs" ...
```

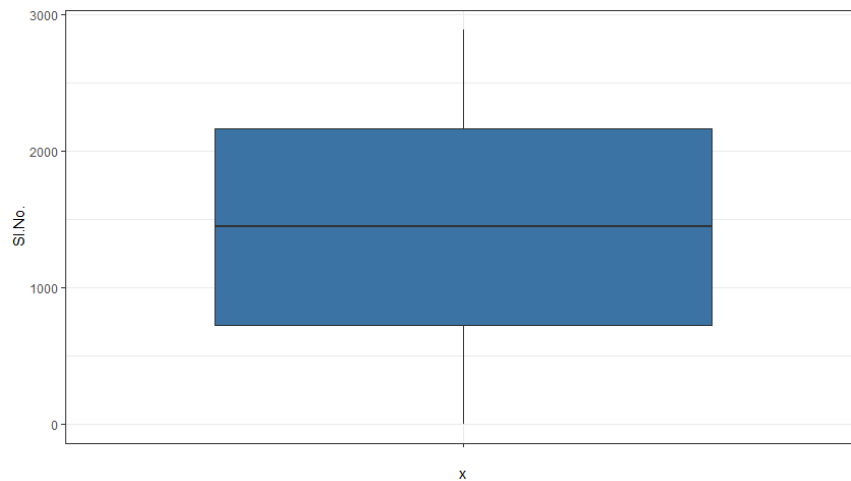
The dataset contains a total of 2890 observations and 10 variables. It includes common variables such as serial number, brand, name, reviews date added, reviews do recommend, reviews number helpful, review rating, reviews text and weight.

```
> summary(wine_reviews)
  Sl.No.      Brand      Name      Reviews.Date.Added  Reviews.do.Recommend  Reviews.Num.Helpful
Min.   : 1.0    Length:2890  Length:2890      Length:2890      Mode :logical      Min.   : 0.000
1st Qu.: 723.2   Class :character  Class :character  Class :character FALSE:90      1st Qu.: 0.000
Median :1445.5   Mode :character   Mode :character   Mode :character  TRUE :1821      Median : 0.000
Mean   :1445.5                                     NA's :979          Mean   : 1.283
3rd Qu.:2167.8                                     3rd Qu.: 1.000
Max.   :2890.0                                     Max.   :29.000
                                     NA's   :2264

Reviews.Rating  Reviews.Text      Reviews.Title      weight
Min.   :1.000    Length:2890      Length:2890      Length:2890
1st Qu.:5.000    Class :character  Class :character  Class :character
Median :5.000    Mode :character   Mode :character   Mode :character
Mean   :4.691
3rd Qu.:5.000
Max.   :5.000
NA's   :445
```

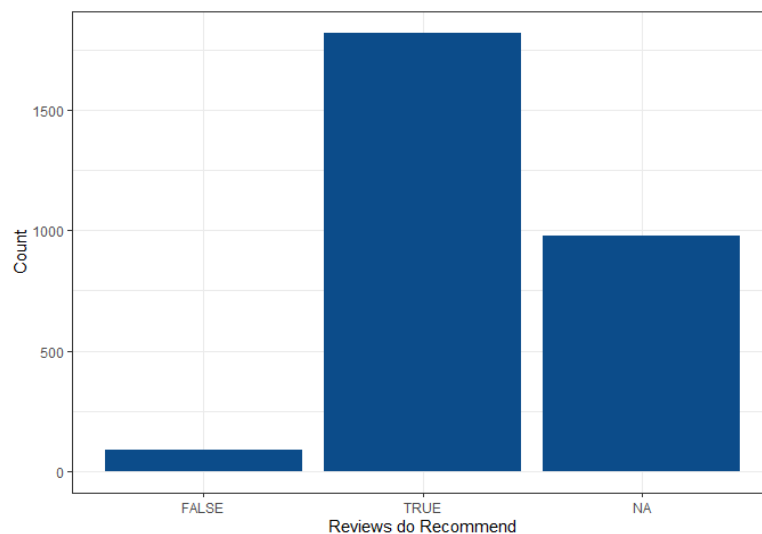
The summary statistics above summarize and provide information for each variable in the wine_reviews dataset.

- *SI.No.*



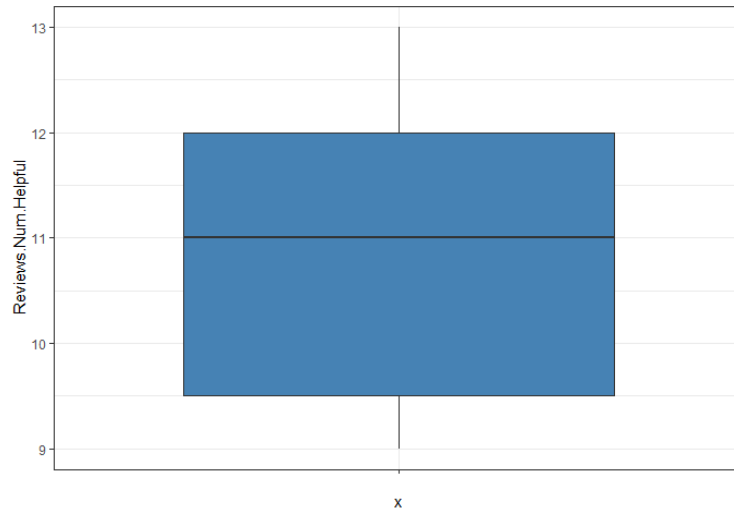
The figure above shows the box plot of the series number for wines. It has an upper quartile of 2167.8 and lower quartile of 723.2. The median is 1445.5. It is normally distributed where 75% of the number of series for wine are between 723.2 and 2167.8. There are no outliers.

- *Reviews.do.Recommend*



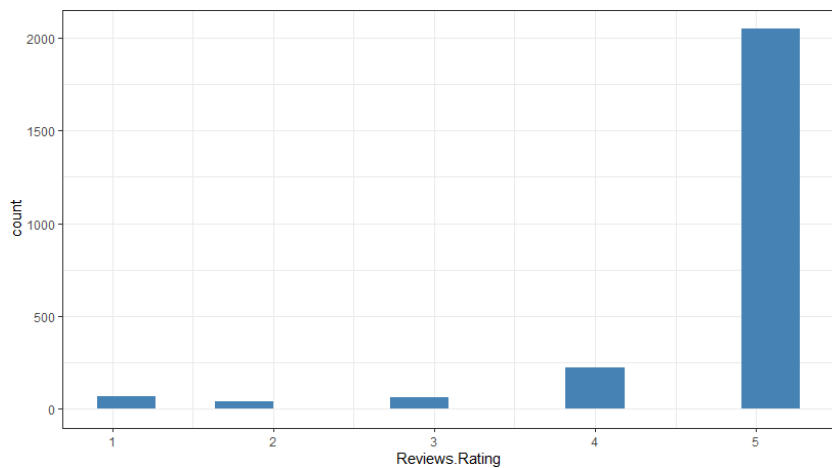
The figure above shows a bar chart for the number of recommended reviews for the wine. From the bar chart above, we can see that there are a total of 1821 bottles of wines with recommended reviews and 90 bottles of wines without reviews. There are 979 missing records for the data.

- *Reviews Num Helpful*



The figure above shows the box plot of the number of helpful reviews for each wine. It has an upper quartile of 3 and lower quartile of 0. The median is 1.283. It is a negative skewed where 75% of the number of series for wine are between 0 and 29. There are 2264 missing records for the data.

- *Reviews Rating*

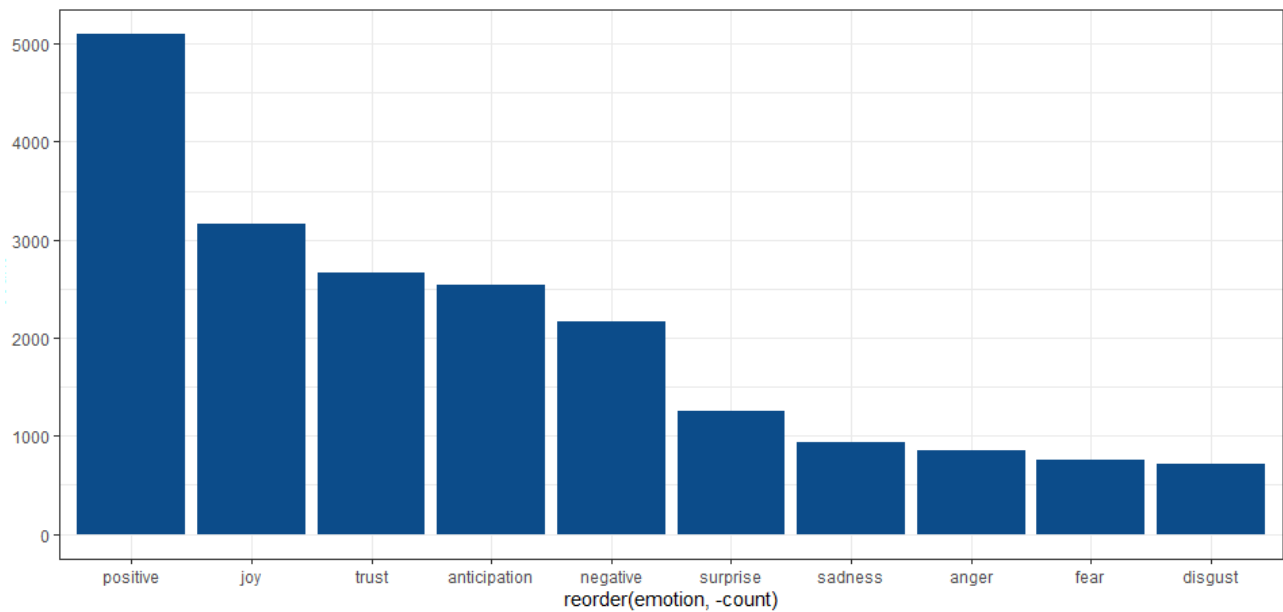


The figure above shows a bar chart for the number of reviews rating for the wine. It has an upper quartile of 5 and lower quartile of 1 for the rating. The median is 4.691. The highest majority of the rating for wine is 5 which means most of the wines are superior and exceptional. There are 445

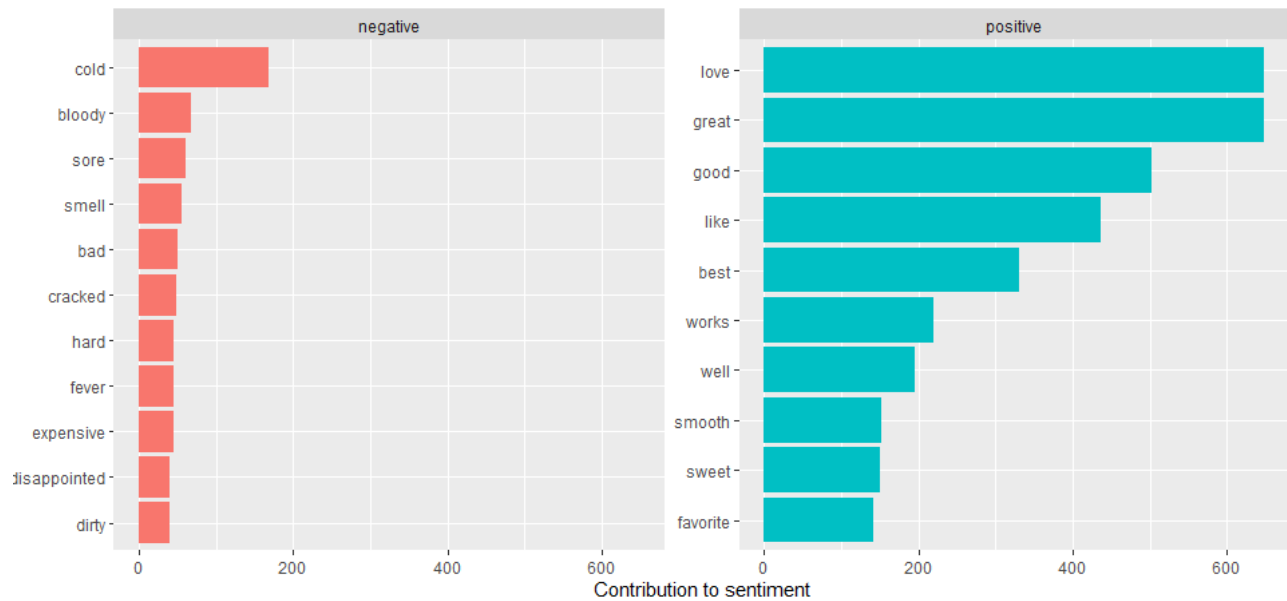
missing records for the data.

3.0 Presentation of Analyses

(i) Sentiment Analysis



The figure above shows the analyzed sentiments for the wine dataset, using the syuzhet package based on the NRC sentiment dictionary. There are eight different emotions which are joy, trust, anticipation, surprise, sadness, anger, fear, disgust and positive/negative ratings. The positive emotion has the highest number of counts which is 5000, while the disgust emotion has the lowest number of counts which is 700. Hence, there are more positive emotions like joy, trust, anticipation and surprise towards the review of the wine compared to the negative emotions like sadness, anger, fear and disgust.



The figure above shows sentiment analysis with the tidytext package using the "bing" lexicon. The contribution to sentiment shows the top 10 words by sentiment, distributed in positive and negative. The top 10 words for negative sentiment are cold, bloody, sore, smell, bad, cracked, hard, fever, expensive, disappointed and dirty. Moreover, the top 10 words for positive sentiment are love, great, good, like, best, works, well, smooth, sweet, favorite. The positive reviews for the wine outweighs the negative reviews. As such, the top contribution to positive sentiment is love and great with a count of 630. While, the least contribution to negative sentiment is disappointment and dirty with a count of approximately 40. Thus, there are a majority of positive reviews towards the wine compared to negative reviews.

(ii) Social Network Analysis

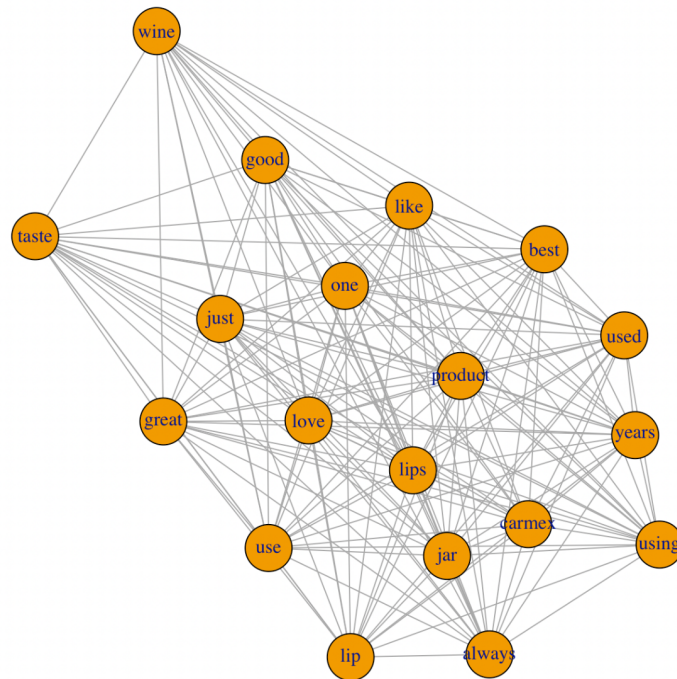
Term document matrix

```
> termM[1:10,1:10]
```

Terms		Terms								
Terms	wine	taste	good	like	great	just	love	one	best	always
wine	206	38	50	45	48	43	40	26	21	8
taste	38	267	68	54	84	43	46	38	31	12
good	50	68	438	78	47	82	47	63	33	17
like	45	54	78	344	63	80	71	57	32	19
great	48	84	47	63	555	70	93	55	45	40
just	43	43	82	80	70	335	72	50	46	30
love	40	46	47	71	93	72	538	71	55	48
one	26	38	63	57	55	50	71	291	60	41
best	21	31	33	32	45	46	55	60	311	36
always	8	12	17	19	40	30	48	41	36	253

As seen from the above term document matrix, the term matrix used is [1:10, 1:10] (rows and columns) this shows that the term extracted is the first ten variables from the website and the first 10 observations following the variables. Some of the most recorded observations regarding wine are good(50) and great(48). Taste appeared the most (84) times in great.

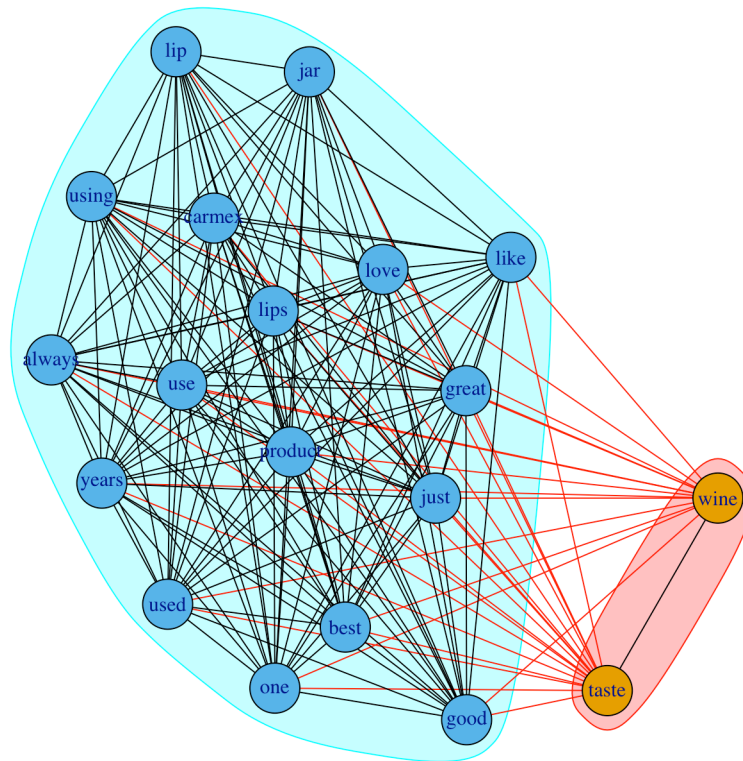
Network Diagram



A network diagram as the one seen above is used to detect groups that are densely connected with nodes. The strength of the connections is represented by varying the display of the line where width, colour, and arrow heads communicate aspects of the relationships. The above network diagram shows the most frequent terms users use to describe wine. The ones in the centre are the most used terms and as seen above, love, lips and products have been vastly used.

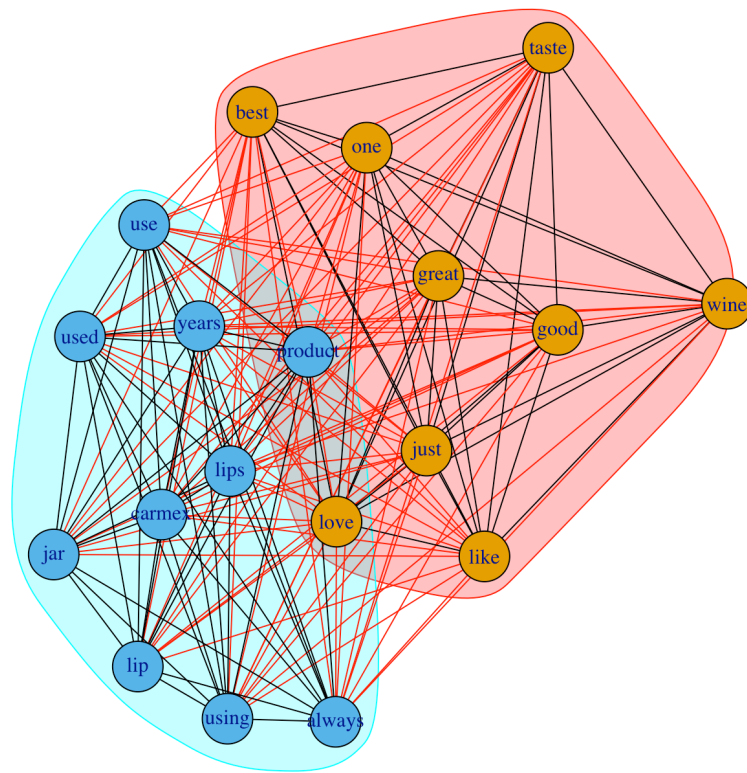
Community detection

- *Using cluster_label_prop*



Community detection is used to reveal the hidden relations among the nodes in the network. In this situation, community detection techniques are useful for finding out what are the most used terms for how people feel towards wine. Community detection is specially tailored for network analysis which depends on a single attribute type called edges. The top highest terms that have derived from the dataset is wine and taste.

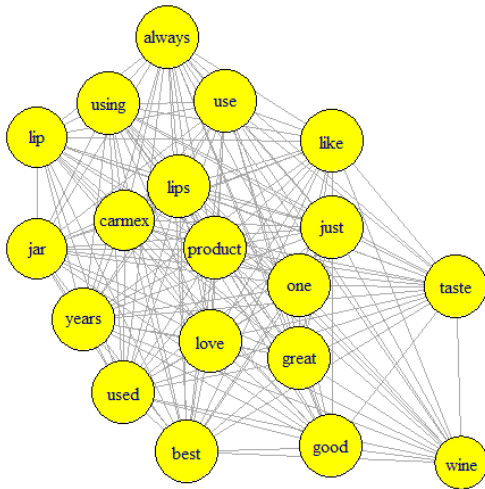
- *Using cluster_fast_greedy*



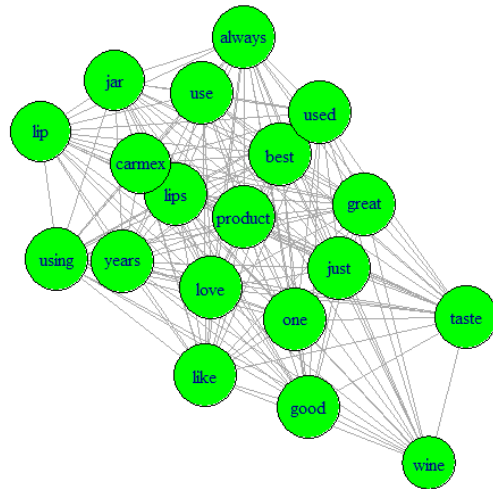
The greedy method of modularity is used to return a group of communities that forms a cluster. The function used to show the graph is used to find subgraphs of the communities via directly optimizing a modularity score. In the network analysis above, it can be seen that the terms highlighted in blue are grouped together as opposed to the terms grouped together in red. The distinguishing factor is represented by the different colour of grouping as seen above.

Hub and authorities

Hubs



Authorities



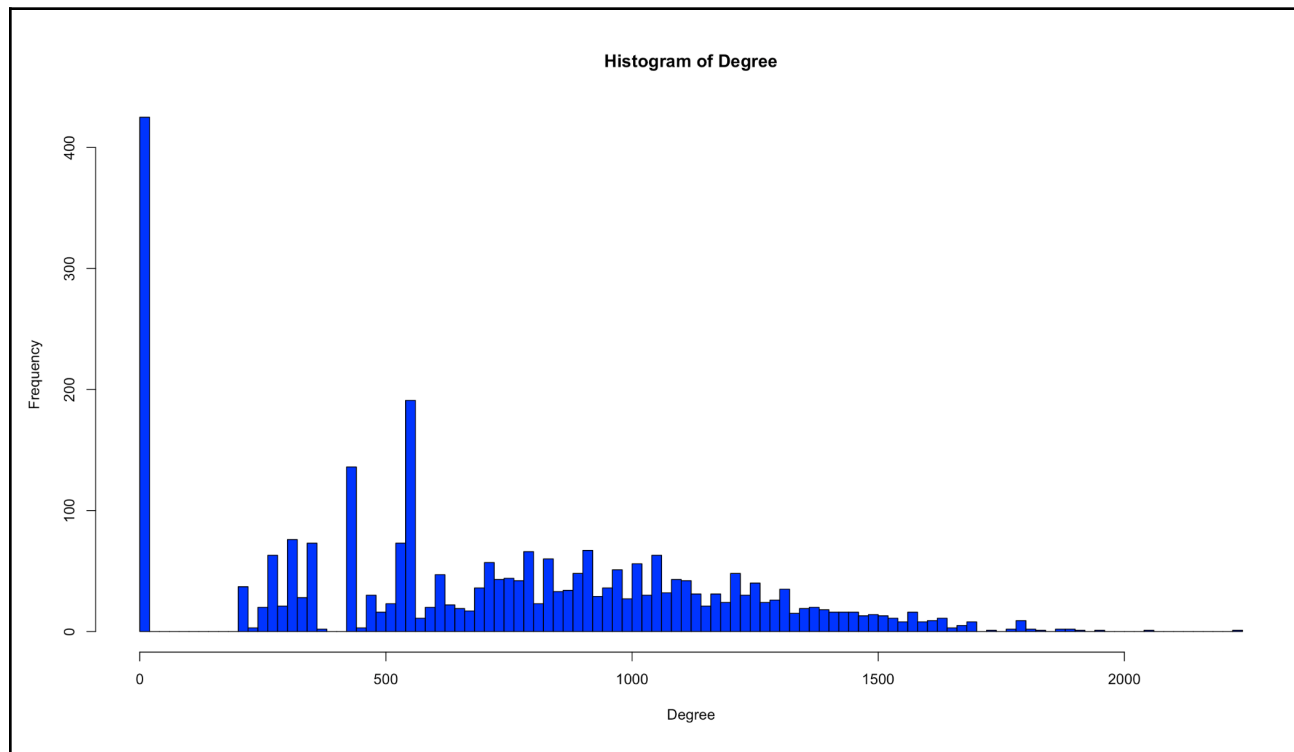
Hub and authorities as seen from the above networks are used to define in terms of one another in mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to.

Highlighting degrees



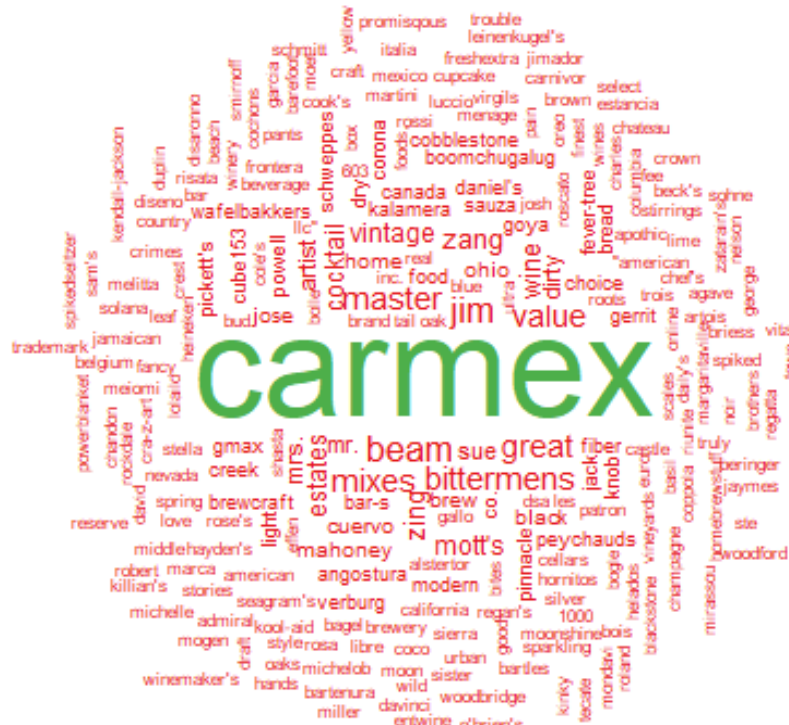
The highlighting degrees network show the focus of the visualization on a particular node or a group of nodes. The term with the most connection/degree seems to be lips, products, love. Carmex appeared in the network however it was an error in the review as it has nothing to do with the wine that we are examining.

Histogram for Network of Wine_Text



The histogram above shows that the frequency at its highest peak is at 400 when the degree is 0.

(iii) Wordcloud

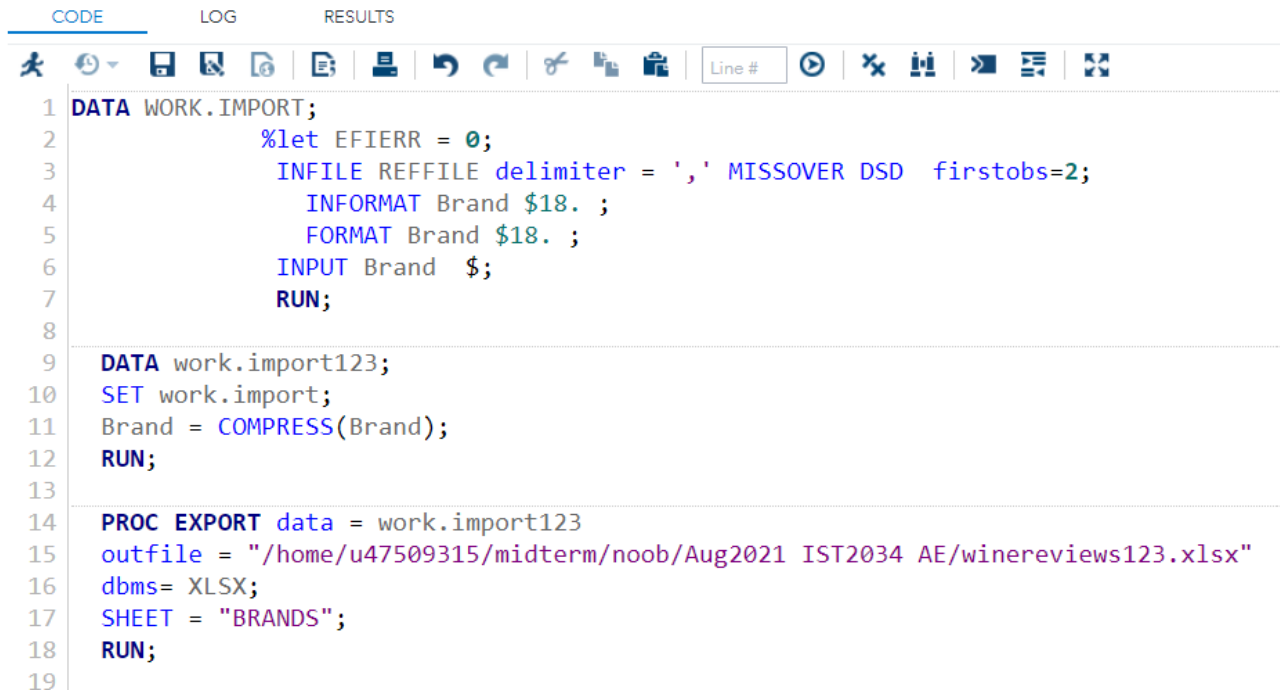


Before modification

For the third presentation analysis, we have chosen to do word cloud analysis. Word Cloud refers to a graphic depiction of the frequency of words. The larger the keyword appears in the graphic generated, the more frequently it appears in the dataset being analysed. However, we have spotted a few issues in our first attempt doing the word cloud. The diagram above shows our first attempt at word cloud, we have found out that Carmex is actually not a brand of wine. After several research, we realised that word cloud analysis separates reading every value in Brand. For example, Great Value is supposed to be one brand, but in the result above, they are separated. This will lead to inaccurate analysis. Therefore, we have figured out some ways to fix this issue.

Due to limited knowledge on R Studio, unfortunately, we are unable to compress the brand values using R Studio. However, we have learnt a similar function in SAS Studio. So, we have decided to

compress the blanks between brand names in SAS Studio and generate a new Excel file to be used in Word Cloud analysis using R Studio.



```
1 DATA WORK.IMPORT;
2     %let EFIERR = 0;
3     INFILE REFFILE delimiter = ',' MISSOVER DSD firstobs=2;
4     INFORMAT Brand $18. ;
5     FORMAT Brand $18. ;
6     INPUT Brand $;
7     RUN;
8
9 DATA work.import123;
10 SET work.import;
11 Brand = COMPRESS(Brand);
12 RUN;
13
14 PROC EXPORT data = work.import123
15 outfile = "/home/u47509315/midterm/noob/Aug2021 IST2034 AE/winereviews123.xlsx"
16 dbms= XLSX;
17 SHEET = "BRANDS";
18 RUN;
19
```

Compress brand names using SAS Studio

The diagram above shows the code we used for compressing the brand names using SAS Studio. First, we import the original raw excel file into SAS. Then, we used the COMPRESS() function to remove the blank space between each brand name. Lastly, we output the modified excel file that only contains the Brand column for word cloud analysis.

Next, to remove the irrelevant observations, we added a line of code in R Studio while creating the word cloud. The R Code below is to remove Carmex, a lip balm brand which is irrelevant to wine.

```
wines <- tm_map(wines, removewords, c("carmex"))
```


4.0 Coding In R

```
###Code###
#read data
library(ggplot2)
library(esquisse)
library(modeldata)
wine_reviews <- read.csv("C:/Users/Yiwen/Downloads/wine_reviews.csv")
dim(wine_reviews)# check the number of columns and rows
str(wine_reviews)# check the variables
summary(wine_reviews)# print summary statistics

# generate graph for sl.No.
ggplot(wine_reviews) +
  aes(x = "", y = Sl.No.) +
  geom_boxplot(shape = "circle", fill = "#3A73A4") +
  theme_bw()

# generate graph for Review.do.Recommend
ggplot(wine_reviews) +
  aes(x = Reviews.do.Recommend) +
  geom_bar(fill = "#0C4C8A") +
  labs(x = "Reviews do Recommend",
       y = "Count") +
  theme_bw()

# generate graph for Reviews.Num.Helpful
# removed 1040 rows which contain non-finite values
wine_reviews %>%
  filter(Sl.No. >= 580L & Sl.No. <= 1980L) %>%
  filter(Reviews.Num.Helpful >= 8.4 & Reviews.Num.Helpful <=
13.4 | is.na(Reviews.Num.Helpful)) %>%
  ggplot() +
  aes(x = "", y = Reviews.Num.Helpful) +
  geom_boxplot(shape = "circle", fill = "#4682B4") +
  theme_bw()

# generate graph for Reviews.Rating
# removed 445 rows which contain non-finite values
ggplot(wine_reviews) +
  aes(x = Reviews.Rating) +
```

```
geom_histogram(bins = 12L, fill = "#4682B4") +  
theme_bw()
```

Sentiment Analysis

```
# load required packages
library(tidyverse)
library(syuzhet)
library(tidytext)
# import text dataset
wine_reviews <- read.csv("C:/Users/Yiwen/Downloads/wine_reviews.csv")
text.wine_reviews <- tibble(text =
  str_to_lower(wine_reviews$Reviews.Text))
# analyze sentiments using the syuzhet package based on the NRC
sentiment dictionary
emotions <- get_nrc_sentiment(text.wine_reviews$text)
View(emotions)
emo_bar <- colSums(emotions)
emo_bar
emo_sum <- data.frame(count=emo_bar, emotion=names(emo_bar))
emo_sum
# create a graph showing the counts for each of eight different emotions
and positive/negative rating
ggplot(emo_sum, aes(x = reorder (emotion, -count), y = count))+
  geom_bar (stat = "identity", fill = "#0C4C8A")+
  theme_bw()

# sentiment analysis with the tidytext package using the "bing" lexicon
bing_word_counts <- text.wine_reviews %>% unnest_tokens(output = word,
input = text) %>%
  inner_join(get_sentiments("bing"))%>%
  count(word, sentiment, sort = TRUE)

# select the top 10 words by sentiment
bing_top_10_words_by_sentiment <- bing_word_counts %>%
  group_by(sentiment) %>%
  slice_max(order_by = n, n=10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n))
bing_top_10_words_by_sentiment

# create barport
bing_top_10_words_by_sentiment %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment", x = NULL) +
  coord_flip()
```


Social Network Analysis

```
library(tm)
library(NLP)
net <- iconv(wine_reviews$Reviews.Text, to = "utf-8")
net <- Corpus(VectorSource(net))

# clean text
# convert all the character to lowercase
# remove all the punctuation
# remove all the Numeric
# remove common english word
net <- tm_map(net, tolower)
net <- tm_map(net, removePunctuation)
net <- tm_map(net, removeNumbers)
cleandata <- tm_map(net, removeWords, stopwords('english'))

# select row sum more than 270 for term document matrix
tdm <- TermDocumentMatrix(cleandata)
tdm <- as.matrix(tdm)
tdm <- tdm[rowSums(tdm)>270,]
tdm[1:10,1:10]

# network of terms
library(igraph)
tdm[tdm>1] <- 1
termM <- tdm %*% t(tdm)
termM[1:10,1:10]
network_graph <- graph.adjacency(termM, weighted = T, mode =
'undirected')
network_graph
network_graph <- simplify(network_graph)
V(network_graph)$label <- V(network_graph)$name
V(network_graph)$degree <- degree(network_graph)

# network diagram
set.seed(222)
plot(network_graph)

# by cluster_edge_betweenness
clus1 <- cluster_edge_betweenness(network_graph)
plot(clus1, network_graph)

# by cluster_fast_greedy
clus2 <- cluster_fast_greedy(as.undirected(network_graph))
```

```

plot(clus2, as.undirected(network_graph))

# Hub and authorities
hub <- hub_score(network_graph, weights = NA)$vector
authorities <- authority_score(network_graph, weights=NA)$vector
par(mfrow=c(1,2))
plot(network_graph, vertex.size=hub*30, main='Hubs',
      vertex.color="yellow")
plot(network_graph, vertex.size=authorities*30, main='Authorities',
      vertex.color="green")
par(mfrow=c(1,1))

# highlighting degrees
V(network_graph)$label.cex <- 2.2*V(network_graph)$degree /
max(V(network_graph)$degree) + 0.3
V(network_graph)$label.color <- rgb(0, 0, .2, .8)
V(network_graph)$frame.color <- NA
egam <- (log(E(network_graph)$weight)+.4) /
max(log(E(network_graph)$weight) + .4)
E(network_graph)$color <- rgb(.5, .5, 0, egam)
E(network_graph)$width <- egam
plot(network_graph,
      vertex.color='yellow',
      vertex.size = V(network_graph)$degree*.5)

# network of Wine_Text
textn <- t(tdm) %*% tdm
network_graph <- graph.adjacency(textn, weighted = T, mode =
'undirected')
V(network_graph)$degree <- degree(network_graph)
network_graph <- simplify(network_graph)
hist(V(network_graph)$degree,
      breaks = 100,
      col = 'blue',
      main = 'Histogram of Degree',
      ylab = 'Frequency',
      xlab = 'Degree')

```

Word cloud

```
#insert the necessary libraries
```

```
library(NLP)
```

```
library(tm)
```

```
library(RColorBrewer)
```

```
library(wordcloud)
```

```
library(ggplot2)
```

```
#library(dplyr)
```

```
library(data.table)
```

```
library(rJava)
```

```
library(RWeka)
```

```
library(SnowballC)
```

```
#locate the file and create r data frame
```

```
#create textual corpus
```

```
wines <- Corpus(DirSource("E:/wines"))
```

```
#cleaning the brands with text mining(tm) package
```

```
wines <- tm_map(wines, stripWhitespace)
```

```
wines <- tm_map(wines, tolower)
```

```
wines <- tm_map(wines, removeWords, stopwords("english"))
```

```
wines <- tm_map(wines, removePunctuation)
```

```
wines <- tm_map(wines, PlainTextDocument)
```

```
wines <- tm_map(wines, removeWords, c("carmex"))
```

```
#create tdmPromote and termFreqPromote
```

```
tokPromote <- function(x) NGramTokenizer(x, Weka_control(min=2, max=3))
```

```
tdmPromote <- TermDocumentMatrix(wines, control = list(tokenize =  
tokPromote))
```

```
termFreqPromote <- rowSums(as.matrix(tdmPromote))
```

```
termFreqVectorPromote <- as.list(termFreqPromote)
```

```
#calculating the frequency of the respective terms
```

```
wines2 <- data.frame(unlist(termFreqVectorPromote), stringsAsFactors =  
FALSE)
```

```
setDT(wines2 , keep.rownames = TRUE)
```

```
setnames(wines2 , 1, "term")
```

```
setnames(wines2 , 2, "freq")
```

```
wines3 <- head(arrange(brands2, desc(freq)), n = 30 )
```

```
wines3$npstype <- "brands"
```

```
#create wordcloud with the wordcloud package
wordcloud(words = wines2$term, freq = wines2$freq, min.freq = 1,
          max.words=500, random.order=FALSE, rot.per=0.35,

          colors=brewer.pal(3, "Set1"))
```

5.0 Lesson & Conclusion

Through this assignment, our team has learnt that teamwork is extremely significant. From choosing dataset to conducting all sorts of analysis, our team has always been brainstorming together to collect the best ideas and decisions. Our team has chosen 3 different types of analysis which are word cloud, social network analysis and sentiment analysis. Due to time constraints, we have limited knowledge for this subject. Therefore, we are unable to apply some techniques to achieve the particular requirements. For example, while constructing word cloud analysis, we noticed that the brand names were separated leading to inaccurate results. Fortunately, we have learnt a function that allows us to compress the blank space between the data which is the COMPRESS() function in SAS Studio. Thus, we have figured out how to manipulate the dataset with SAS Studio for the word cloud analysis. This had successfully helped us to solve the problem. We truly appreciate the knowledge absorbed from different subjects and through the assignment we learnt more about how to utilise the R studio to get the best analysis from the data provided to us.

6.0 Reflection

19027879 Lee Jia Yin

After completing this assignment, I am amazed by the power of the technology and programming tools. In the beginning, I am quite confused and lost in using programming tools as I don't see how I can apply this knowledge in my future job. However, throughout this assignment, I have learnt the concept of applying all the programming knowledge in digging valuable information from the messy raw data file and turning them into decisions to help a company. I have also seen the importance of knowing different programming tools. By learning a few different programming languages, we will be able to tackle different issues with different tools. Also, I have encountered an unexpected problem which is having to involve another programming language in this assignment. Although we are not sure if this solution will be accepted, we tried our best to utilize all the knowledge we have to solve the issue. I am grateful for this assignment for giving me a whole new level of understanding.

19028992 Ngoi Yi Wen

Through this assignment, I learned how to use text mining and sentiment analysis to obtain valuable insights. I have developed a better understanding of how to use these techniques to provide more complex insights than I have in the past. Now I can interpret free-form text, which includes customer reviews, feedback comments, and survey responses, as well as blog posts, new articles and other written content. Besides, I have explored how people talk about brands within the context of wines. I have utilized the ability to understand customer sentiment about a variety of different brands in wines. This allows me to provide valuable insights into customer feedback on the wine. Next, I referred to the tutorial exercises and Youtube's video to get a better understanding of this coding. All of these resources help me so well to do this report and coding in RStudio. I would like to apply the skills and knowledge I acquired to the professional field. I am very happy to work with my group members as they provide a lot of support and ideas to complete the assignment.

19033067 Lim Wei Zheng

It has been truly a joy to work on this assignment as we have learnt a lot of things such as how to choose the right datasets from kaggle, how to create word clouds, sentiment analysis and social network analysis. Along the way, we were faced with a few problems but with teamwork and google, we were able to solve it and successfully analyze the data in our own unique way. I learnt that I might not be able to find the solution alone immediately, and I need to consult my teammates for opinions and answers. All the stress was worth it as we spent days and nights finishing this assignment while working on other subject's assignments too. I felt that learning R Studio will serve as a foundation for me to learn other programming languages to play with data and learn how to extract data from different ways and perspectives. I am happy that our team did our best and invested our time to learn more about R-studio and data.

19022649 Yashinnie A/P R Gnaneswaran

It has been an amazing experience working on this assignment, although it was a little difficult at the beginning to understand the concept of R programming, eventually, I was able to learn how the raw data was analyzed technically to obtain useful information like the positive and negative emotions for better decision making, social network analysis, and word cloud. Along the way, as we encountered multiple issues, I was able to build strong communication and teamwork between my teammates, by finding a solution together prior to the knowledge we have without hesitance. Although it was a stressful learning process, I am grateful that we managed it well with proper time and team management. As such, I hope I will be able to carry forward this knowledge learned through the R programming language and RStudio in the future for other project execution.

18004036 Thejal A/P L.Ramesh

This assignment has been the highlight of my semester. This is because I have been able to see how data is being visualised in a technical manner, in this assignment's case, R programming language. It is fascinating to me how R programming itself is able to generate real time results of graphs in the form of networks. In addition, I have built a strong foundation of teamwork with my mates and I am able to communicate clearly with them without feeling any sort of hesitation. The teamwork process was smooth as we were able to distribute tasks very well and time management as well was done very smoothly. I hope to be able to apply the skills and knowledge acquired from this assignment in the future when I work with the companies I am hired for.