

Configuration Manual

MSc Research Project
Data Analytics

Princy Dcunha
Student ID: x18135889

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing



Student Name:	Princy Dcunha
Student ID:	x18135889
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	12/12/2019
Project Title:	Configuration Manual
Word Count:	903
Page Count:	9

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission , to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Princy Dcunha
x18135889

Introduction

This Configuration Manual aims at illustrating all the steps done to get to the end of this project. Right from stating the hardware 1 used to the various softwares 2 and tools used, is mentioned in the next pages. The outputs and R Scripts that could not be put into the technical report are attached below as well.

1 Hardware Configuration

1.1 MacBook Pro, 2018

The Figure 1 shows the Mac OS configuration used for all processes of this project. It is updated to the latest version 10.14.6 (18G1012) MacOS Mojave and has a 2.2 GHz Intel Core i7 processor.



Figure 1: MAC OS

2 Software Configuration

2.1 RStudio

Figure 2 shows the RStudio version used for running all the R scripts from cleaning and preparing the data to implementing and evaluating the models. The version used was RStudio Desktop 1.2.5019. Figure 3 Shows the successful installation of RStudio Software.

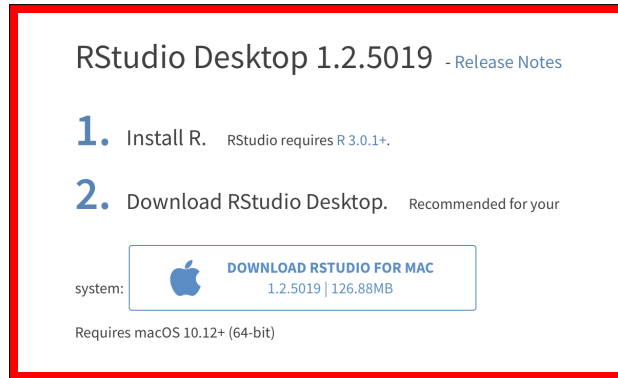


Figure 2: RStudio

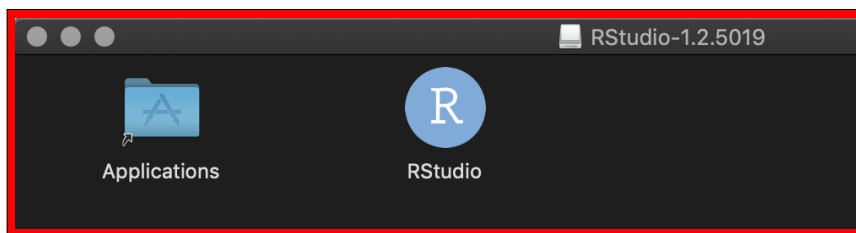


Figure 3: RStudio Installed

2.2 Overleaf

This Figure 4 online documentation tool was used for all the documentations related to this project. It has an inbuilt library and is like an html code with tags and labels. Everything done on overleaf is automatically saved on the cloud, eliminating the risk of losing a drafted document.

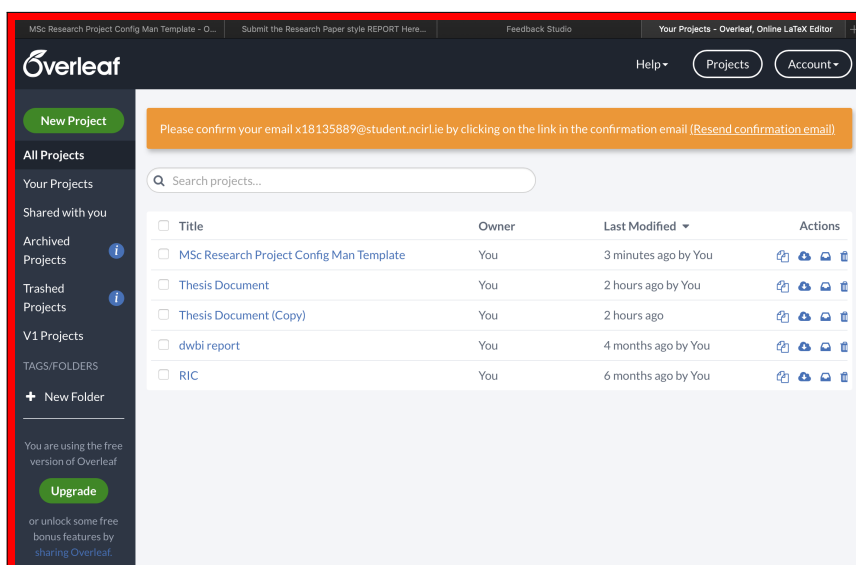


Figure 4: Overleaf

2.3 Microsoft Excel

Figure 5 displays the version of Microsoft Excel for Mac used, which is Version 16.30. Figure 6 shows an excel sheet containing the project dataset. Microsoft Excel was used to do minor tweaks and adjustments and VLOOKUP() was used to combine two sheets having common columns.

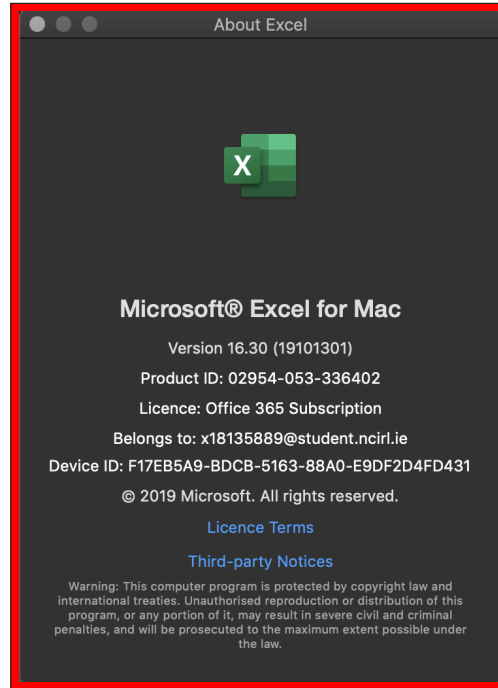


Figure 5: Microsoft Excel Version

Hotel Address	Hotel Name	Sentiment Type	Average Score	Total Number of Reviews	Reviewer's Location
Fulham Road Stamford Bridge Hammersmith and Fulham	Millennium Copthorne Hotels at Chelsea Football Clu	0	8.2	Cleanliness	1842
14 rue Armand 11th arr 75011 Paris France	Les Jardins Du Marais	0	8.1	Our room fa	638
Corso Concordia 1 Milan City Center 20129 Milan Italy	Chateau Monfort Hotels Chateau	1	9	Bed isn't com	1003
Via Panfilio Castaldi 7 Central Station 20124 Milan Italy	Mokinba Hotels Baviera	0	7.7	No Negative	1240
Ylefer Graben 34 20 01 Innere Stadt 1010 Vienna Austria	Hotel Des Tigra	1	8.9	There is bull	1817
20 Newem Square Kensington and Chelsea London SW5	Twenty Newem Square Hotel	0	8.3	Friendly stal	1718
Gran Via de les Corts Catalanes 668 Eixample 08010 Barcelona	Hotel Palace GL	1	9.4	The concierg	1266
Via Fabbricatorelli 21 Milan City Center 20121 Milan Italy	Hotel Canova	1	8.5	Shower plus	2463
69 Vincent Square Westminster Borough London SW1P	Grange Rochester Hotel	0	8.2	The room fu	1046
11 Shortlands Hammersmith and Fulham London W6	Novotel London West	0	8.3	Really g	2443
202 220 Cromwell Road Kensington and Chelsea London NW	Novotel Kensington	0	8.2	No Negative	879
Via Fabbricatorelli 21 Milan City Center 20121 Milan Italy	Hotel Canova	1	8.5	Nothing	2463
350 Oxford Street Westminster Borough London W1C	Radisson Blu Edwardian Berkshire	0	8.1	No Negative	1600
43 51 Wembley Hill Road Brent London HA9	BAU Unite St George's Hotel Wembley	1	8.9	WIFI didn't v	2274
5 More London Place Tooley Street Southwark London E11	London Tower Bridge	0	8.7	Cost a bit m	1705
53 59 Kilburn High Road Maida Vale London Camden	LoBEST WESTERN Maltree Hotel Maida Vale	0	7.1	WINDOWLE	1877
97 Cromwell Road Kensington and Chelsea London SW	Holiday Inn London Kensington Forum	0	7.8	The hotel in	3867
40 Trinity Square City of London London EC3N	40s Unite citizenM Tower of London	1	9.1	Very comfy	4872
George Street Westminster Borough London W1S	581 (Dorlands Hotel	0	8.1	We were ple	1811
Avenida Catedral 7 Ciutat Vella 08002 Barcelona	Spain Col n Hotel Barcelona	1	8.9	We had a co	1300
Half Moon Street Westminster Borough London W1J	71 Hilton London Green Park	0	7.3	No Fridge in	1139
Viale Certosa 104 100 Certosa 20156 Milan Italy	Best Western Hotel Mirage	0	8.5	The staff we	785
18 Albert Embankment Lambeth London SE1	771 Unite Park Plaza London Riverbank	0	8.3	Large queue	4684
16 avenue de Tourville 7th arr 75007 Paris France	Le Tourville Eiffel	1	8.8	No Negative	545
41 54 Ruckingham Gate Westminster Borough London	St James Court A Taj Hotel London	0	8.7	No Negative	2394
257 Rue de Vaugirard 15th arr 75015 Paris France	Novotel Paris Vaugirard Montparnasse	0	7.6	The toiletter	1313
Via Spadari 11 Milan City Center 20123 Milan Italy	Hotel Spadari Al Duomo	1	9.3	Perfect loca	755
50 Lancaster Gate Westminster Borough London W2	3F Commodore Hotel	0	6.7	Location	2400
Upper Woburn Place Camden London WC1H	98H Ur Ambassadors Bloomsbury	0	7.9	No Negative	1231
Via Spadari 11 Milan City Center 20123 Milan Italy	Hotel Spadari Al Duomo	1	9.3	Perfect loca	755
Scarsdale Place Kensington Kensington and Chelsea	Lor Copthorne Tara Hotel London Kensington	0	8.1	The staff we	7105
Spaansaat 288 292 Amsterdam City Center 1012	9K An Nh City Centre Amsterdam	0	8.2	Great locati	3417
Westminster Bridge Road Lambeth London SE1	7UT Ur Park Plaza Westminster Bridge London	0	8.7	Having stay	12158
625 Chiswick High St Chiswick London W4	58Y United K Clayton Hotel Chiswick	0	8.5	No Negative	1944
Empire Way Wembley Brent London HA9	805 United K Holiday Inn London Wembley	0	8.3	Swimming i	3469
7 Gracchurch Street City of London London EC3N	006 U Club Quarters Hotel Gracchurch	0	8.2	Just the off	2986
La Rambla 128 Ciutat Vella 08002 Barcelona Spain	Hotel Sams Rivoli Rambla	0	8.1	No Negative	1957

Figure 6: Microsoft Excel Sheet

2.4 Web

The web plays an important role in supporting the completion of any project. Numerous amount of data and knowledge was gained form the web for stating and implementing this project.

2.4.1 Chrome

The Google Chrome Figure 7 was used for finding datasets and to explore various possibilities to complete and complement the project.

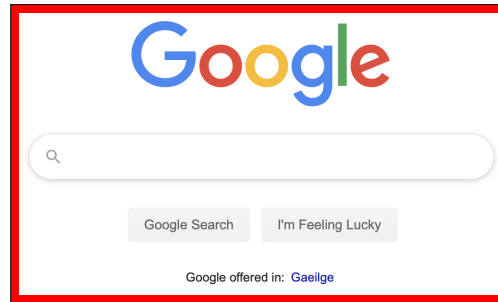


Figure 7: Google

2.4.2 Safari

The Safari Figure 8 is a web browser made for MacOS and comes in very handy when in need. It was the default browser used for most of the job.



Figure 8: Safari

2.5 Outputs and Visualizations

Figure 9 shows the distribution of reviewer scores in the dataset. This was a part of exploratory data analysis.

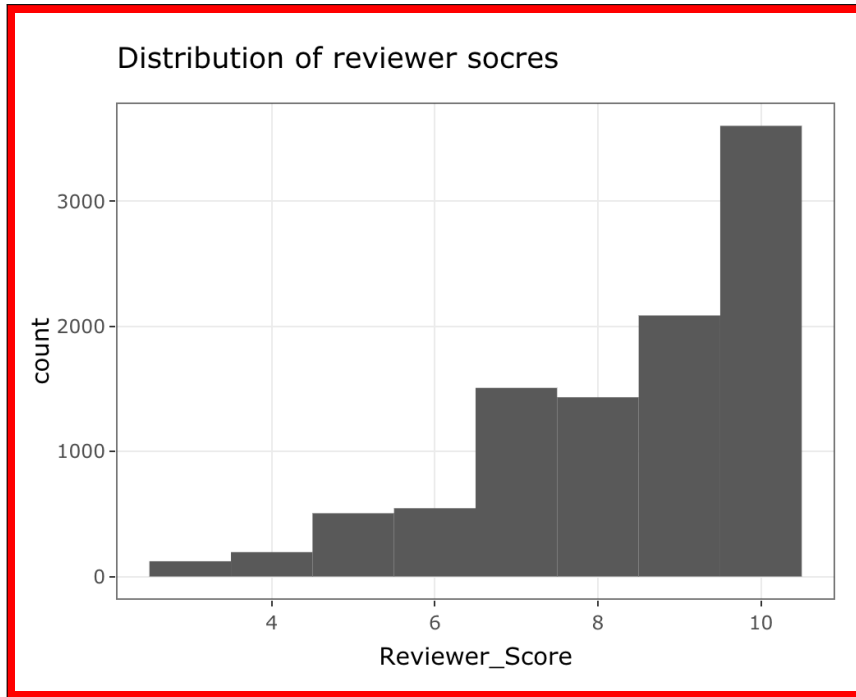


Figure 9: Reviewer_Score Plot

Figure 10 plots the Average Reviewer Score from the dataset. This also was a part of Exploratory Data Analysis (EDA).

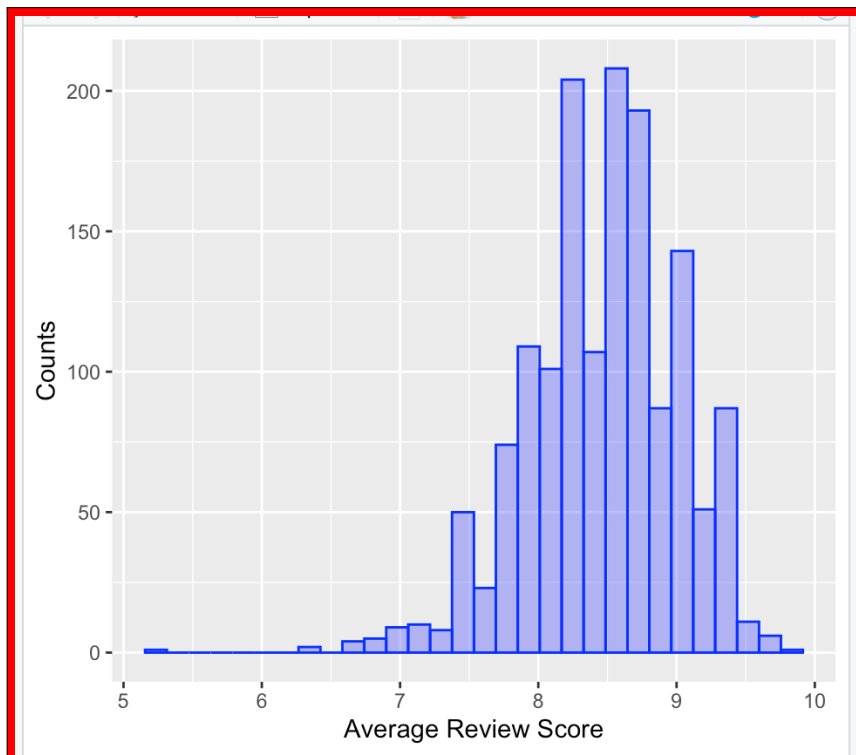


Figure 10: Average Review Score

Figure 11 shows the Top most score of the ratings and Figure 12 shows the bottom most

score of the user ratings.

```
> data.frame(percentile=perc,score=score)->d
> print("Top rating scores are:")
[1] "Top rating scores are:"
> d%>%arrange(desc(score))%>%head(5)
percentile score
1      98%    9.4
2      99%    9.4
3      96%    9.3
4      97%    9.3
5      93%    9.2
```

Figure 11: Top Reviews

```
> print("Bottom rating scores are:")
[1] "Bottom rating scores are:"
> d%>%arrange(score)%>%head(5)
percentile score
1         1%    6.9
2         2%    7.1
3         3%    7.1
4         4%    7.3
5         5%    7.4
>
```

Figure 12: Bottom Reviews

Figure 13 displays the output of topic modelling. It is not clear because it does not have unique topics, rather it has unique full sentences, which makes it difficult to plot as per the topic. This was not a very successful approach and would be recommended as future work.

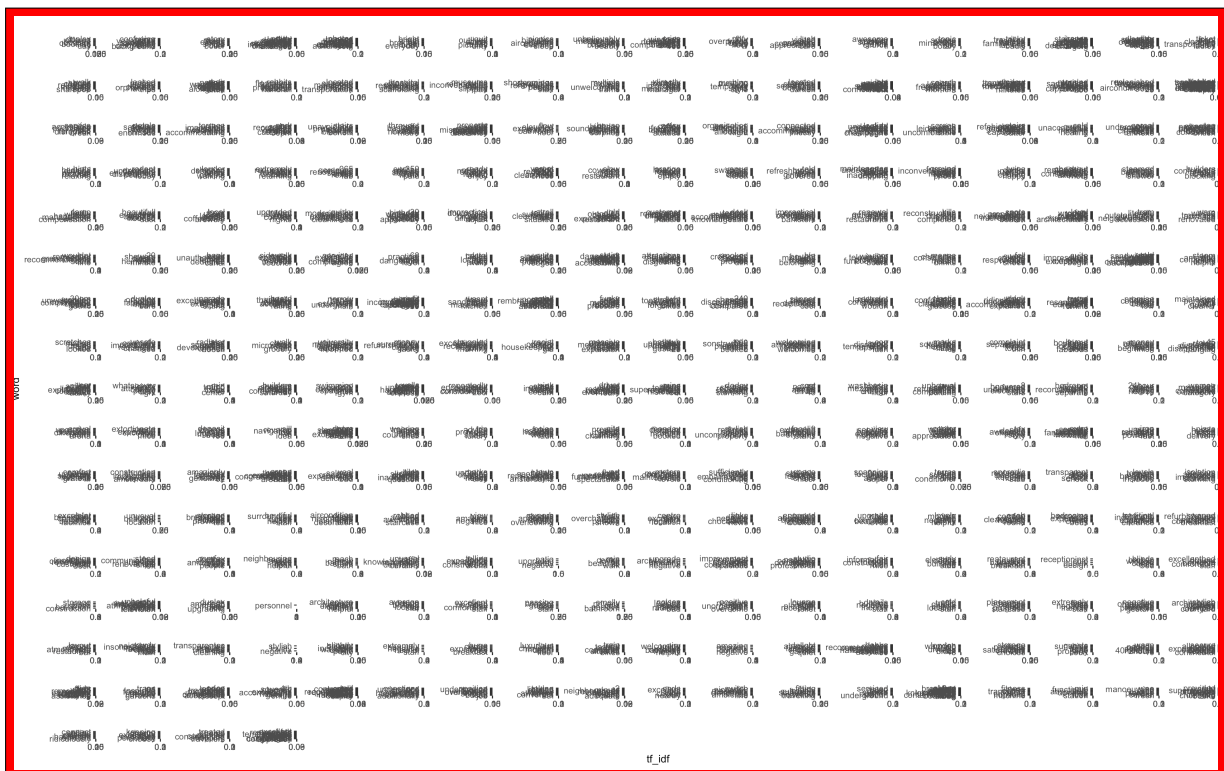


Figure 13: Topic Modelling

Figure 14 shows the extraction of countries from the hotel address.


```

> countries=paste(unique(df$country.x),collapse=",")
> message=paste("The countries mentioned in the dataset are:", countries)
> print(message)
[1] "The countries mentioned in the dataset are: Netherlands,United Kingdom,France,Spain,Italy,Austria"
>

```

Figure 14: Countries mentioned in dataset

Figure 15 show the extraction of cities from the address and country names.

```

> cities=paste(unique(df$city.x),collapse=",")
> message=paste("The cities mentioned in the dataset are:", cities)
> print(message)
[1] "The cities mentioned in the dataset are: Amsterdam,London,Paris,Barcelona,Milan,Vienna"
>

```

Figure 15: Cities in dataset

Figure 16 gives the bar plot of aspects extracted from topic modelling, showing its count and relevance in that specific review/sentence. Its the Term Frequency-Inverse Document Frequency (TF-IDF) in which we can see how important a word was to that sentence.

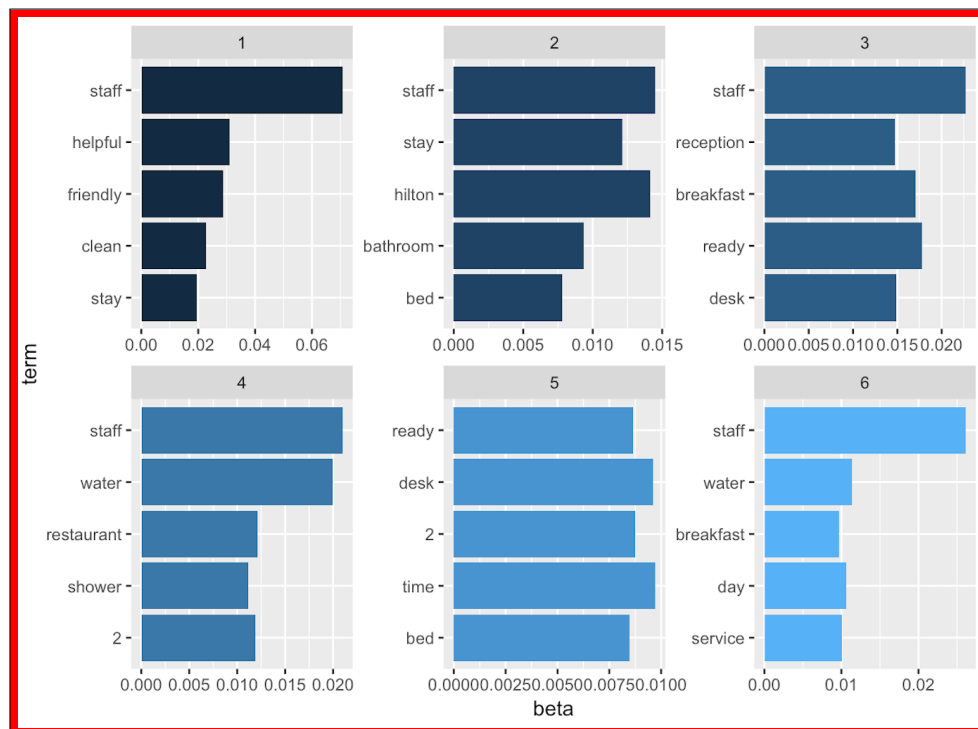


Figure 16: TF-IDF

Figure 17 shows how much an aspect has an effect on the given sentence.

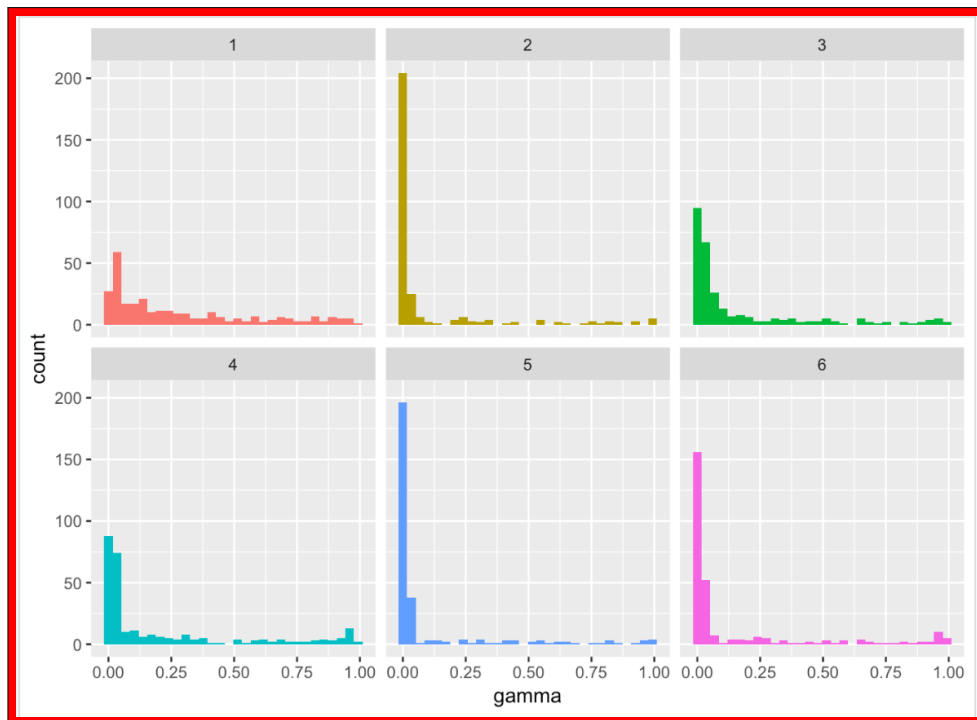


Figure 17: Topic Modelling

Figure 18 shows how the data is split into training set and test set.

```
dataset = read.csv("/Users/princydcunha/Desktop/NewWhole.csv")
#View(dataset)
str(dataset)
dataset = dataset[4:8]
#dataset$Sentiment_Type= factor(dataset$Sentiment_Type, levels = c(0, 1))
dataset$Review = NULL
dataset$Sentiment_Type = as.numeric(dataset$Sentiment_Type)
dataset$Average_Score = as.numeric(dataset$Average_Score )
dataset$Total_Number_of_Reviews = as.numeric(dataset$Total_Number_of_Reviews)
dataset$Reviewer_Score= as.numeric(dataset$Reviewer_Score)

set.seed(123)
split = sample.split(dataset$Sentiment_Type, SplitRatio = 0.75)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
#View(training_set)
#View(test_set)

#Feature scaling
training_set[-1] = scale(training_set[-1])
test_set[-1] = scale(test_set[-1])

training_set$Sentiment_Type = as.factor(training_set$Sentiment_Type)
test_set$Sentiment_Type = as.factor(test_set$Sentiment_Type)
```

Figure 18: Splitting of Train and Test Data

Figure 19 shows the code used to plot the confusion matrix in Random Forest algorithm.

```

#===== Random Forest

model1 <- randomForest(Sentiment_Type ~ ., data = training_set, importance = TRUE)
model1
varImpPlot(model1)

confusionMatrix(predict(model1,test_set), test_set$Sentiment_Type)

```

Figure 19:

Figure 20 shows the ANOVA results using ChiSquare Test.

```

> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Sentiment_Type

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                6999      9284.1
Average_Score          1    3217.1      6998      6067.0 <2e-16 ***
Total_Number_of_Reviews 1     363.4      6997      5703.6 <2e-16 ***
Reviewer_Score          1         1.8      6996      5701.8  0.1744
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 20:

Figure 21 displays the R code used to calculate and predict the accuracy in Logistic Regression.

```

pR2(model)
fitted.results <- predict(model,newdata=subset(test,select=c(2,3,4)),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$Sentiment_Type)
print(paste('Accuracy',1-misClasificError)) # "Accuracy 0.821631878557875"

p <- predict(model, newdata=subset(test,select=c(2,3,4)), type="response")
pr <- prediction(p, test$Sentiment_Type)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]] #auc 0.8875061

```

Figure 21: