Driving Revenue Strategy Through "Market-Based Approach"

Research Project

13th November 2019

Susheel Aakulu

Acknowledgement

This research project was supported by The Royal Melbourne Institute of Technology and the Australasian Institute of Mining and Metallurgy. I thank our course co-ordinator Dr. Yang Wang and course instructor Mr. Denwick Munjeri who allowed us to gain industry experience as a part of our curriculum. We also extend our vote of thanks to AusIMM for believing in us and allowing us to provide insights and expertise that greatly served the research and addressing the business problem.

I would also like to show gratitude to Natalie, Margaret, Jim, Dave and Sam for sharing their pearls of wisdom with us during the course of this research. It wouldn't have been possible to complete the project within the designated time frame without the support of our industry supervisors. AusIMM and RMIT held us along all the way through the research project in mastering the hurdles in our way to successfully reach the given obstacles.

Susheel Aakulu

Table of Contents

Table of Figures

1. Executive Summary

The Australasian Institute of Mining and Metallurgy (**AusIMM**) provides services to professionals engaged in all facets of the global minerals sector. AusIMM required a customer segmentation model to understand the likelihood of members, the sensibility and responsiveness of members, participation of members in their conferences and ascertaining the cost to serve a conference. The scope of this project is to use an unsupervised machine learning algorithm to segment the members into different groups based on their areas of interest, to determine the participation and activeness of members towards the company. We worked on natural language processing techniques i.e., using textual analysis to understand the speech emotions of responses received through emails sent by members. Also, extracted insights from google analytics for understanding the digital engagement of members.

AusIMM provided us a semi-structured data, which had redundancies and inconsistencies. We worked on data governance policies to handle inconsistencies from the database. This historical data was used to run machine learning algorithms that were later able to segment the members/customers into different personas. After several algorithms, unsupervised K-means clustering model was found to be the most efficient way to segment the customers/members. After segmentation, the segments were used to ascertain the activeness/participation of members in events, cost to serve analysis for an event, events operational support and analysis. We next worked on NLP text mining to understand the emotions of members towards the company, based on different levels in the organization. We also forecasted the cost and member attendance for any given event, so that AusIMM would be ready for the future. We hope that our clustering model, sentiment analysis and inferences from google analytics for digital engagement can help AusIMM in better decision making while expanding roots in the future.

2. Introduction

i. Background and Business Problem

The growth of data analytics has been transforming the viewpoints of the business by helping people in making informed decisions, understand business direction and objectives, explain and analyze the cause of certain events based on data findings, present technical insights which in turn helps in making business decisions.

Our client, The Australasian Institute of Mining and Metallurgy in the background has a privilege of 125-year history or a royal charter, acts as a leader for resource professionals in the mining and metallurgy field. AusIMM provides services to professionals engaged in all facets of the global minerals sector. A community of 13000+ professionals is passionately contributing to the global resource industry, of which just over 1000 are students. Their online presence has been redeveloped to render a modern digital age where interaction is done with technology away from home and on the go. Although the Institute and its members are primarily Australian based, the international reach of the organization has expanded over the years. The AusIMM is uniquely positioned to serve its members in the future and is the reason for initializing this project.

AusIMM intends to foresee and understand their customer segmentation, the cost to serve an event and a member, and their participation and involvement towards the events of the company. Essentially, data analytics plays a pivotal role in understanding customer/member segmentation, understanding customer behavior and market behavior through opinion mining. So, along with the segmentation, AusIMM also plans the migration from data warehousing to the cloud. Thus, we need to clean the data provided by the organization for migration purposes and help them understand the member's behavior, member's activeness in the events conducted by the organization and digital engagement of the members.

The AusIMM membership is basically divided into 5 categories: -

- **Student** – Helps in equipping future leaders of the sector with the skills needed to succeed in the workplace.
- **Associate** – helps in expanding professional network.
- **Member** – Gain a globally recognised credential with the primary professional grade membership.
- **Fellow** – Be recognised for your leadership in resources sector.
- **Chartered Professional** – Accredited leaders across the resource sector.

ii. Objective of this study

The objectives of this research are as follows: -

- Data governance to eliminate redundancies and inconsistencies from the database, to turn the unstructured data to structured data, to present business insights from the cleaned data, and assist in transforming data from warehouse to cloud [Microsoft Azure – Dynamics 365]
- To create persona segments using unsupervised machine learning techniques based on targeted factors like age, education level, areas of interest, professional experience and gender.
- To understand the emotions of the members by analyzing the in-bound mails to the company, i.e., the study of texts using "Emotion detection Sentiment Analysis".
- To ascertain the cost to serve an event and cost to serve a member by analyzing the behavior of events attended by the members considering the demographics and considering the target features from persona segments.
- To create an interactive dashboard to understand the digital engagement, behavior, interests and web activeness of members/users with the help of google analytics, by monitoring their clicks and movements from one web page to another.
- Analysis and operational support of events/conferences. To forecast the future revenue collected from events/conferences, to predict the number of attendants for future events.

iii. Unsupervised Machine Learning Algorithms

Unsupervised learning is a machine learning technique, where you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabelled data. It is a modelling technique in which the algorithm learns about the underlying structure of the data in order to make informed decisions. (referred from the literature review)

Reasons for using Unsupervised Learning:

- Unsupervised machine learning seeks all kinds of unknown patterns in data.
- Unsupervised methods help in feature selection which is useful for segmentation.
- Unsupervised learning algorithm is useful in modelling unlabelled data.

Two of the main methods used in unsupervised learning are principal component analysis and cluster analysis. (Clustering — Unsupervised Learning, n.d.)

Step by step process of unsupervised learning

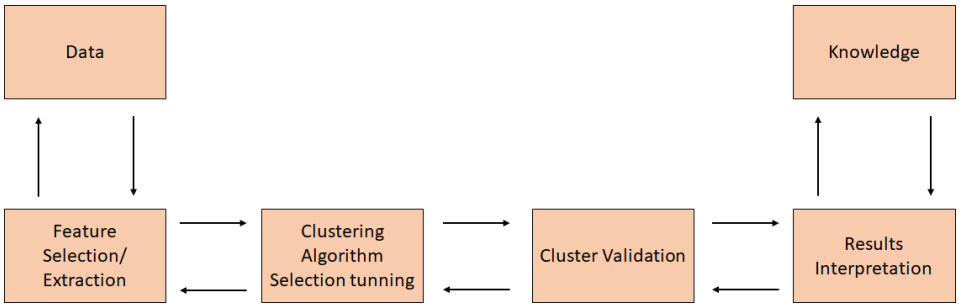

Figure 1 - Process of Unsupervised Machine Learning

iv. Sentiment Analysis

Sentiment analysis is contextual mining of text which recognizes and extracts subjective information for getting insights from social media comments and product reviews, and making data-driven decisions

It is the automated process that uses 'Artificial Intelligence' to identify positive and negative responses/opinions from text. Sentiment analysis is widely used in social media and customer centric industries.

There are many types and flavours of sentiment analysis and tools ranging from polarity to systems that detect feelings and emotions or identify intentions like interested or not depending on the text.

- Fine-grained Sentiment Analysis
- Emotion detection
- Aspect-based Sentiment Analysis
- Intent analysis
- Multilingual sentiment analysis

v. Google Analytics

The Google Analytics is a free tracking tool offered by Google, and it shows us how visitors use a certain website. By using the Google Analytics, we can observe the number of visitors who have visited the website, which device they are using to visit the website and much more. There are various advantages of using Google Analytics: -

- Helps in measuring website's performance.
- Helps in dividing users into different segments (like age, gender, country, device used)
- It optimizes website pages to boost conversions.
- Helps in understanding if the marketing tactics which will drive most traffic to a website.

vi. Forecasting

Forecasting is defined as a technique which uses historical data as inputs to make informed estimates which help in predicting the direction of future trends. Basically, it is used to predict the future events as accurately as possible with all the information (i.e. historical data and/or knowledge of any future events) that may impact the forecasts.

Forecast follows the following process: -

- First a data point or problem is chosen.
- An ideal dataset and theoretical variables are chosen. It helps us to understand and identify what variables must be considered for the forecasting and further helps in deciding how to collect the data.
- We also must make explicit assumptions to simply the forecasting process by cutting down on time and data which is needed to do the forecast.
- After understanding the data, a model is chosen such that the model chosen to fit the dataset, selected variables and assumptions.
- By using the model, the data is analysed and the forecasted.

vii. Analytical approach

For the analysis of this project, the data was obtained from the customer database (MySQL) in which the schema was observed to be inconsistent and does not satisfy the integrity constraints. So, to analyze this data, we will be using RStudio, Microsoft Excel and Power Query Editor for the analysis. By using the unsupervised machine learning techniques (using the K-Means method), the membership database will be segmented based on the number of clusters ascertained with the elbow method. The elbow method is one way to validate the number of clusters with the idea to run K-means clustering on the dataset for a range of values of K. The optimal K number of segments created will be analyzed individually to understand their action, activeness and their benefaction to the company. Further, we will be also using the emotions involved in the customer emails to the company i.e. text mining using Natural Language Processing [NLP] techniques. The NLP or Natural Language Processing is a branch of artificial intelligence that deals with the interaction between the humans and the computers with the main objective to read, decipher, understand and make sense of the human languages in a valuable manner. Therefore, by using NLP we will be analyzing the customer emails received by different departments of the company. The use of Google Analytics will help us in analyzing the behavior of members based on most hit webpages, their interests, the devices used and how frequently a member or non-members visit the website of AusIMM. Therefore, comprehending ways to convert those non-members to members. Further, to understand the operations, [future revenue and conference attendance by members] we will be using the forecasting techniques. The forecasting techniques will be based on the segments and the demographics to understand projected income and the participation of members observed in the future conferences organized by the company.

→ Step by step analytical process design

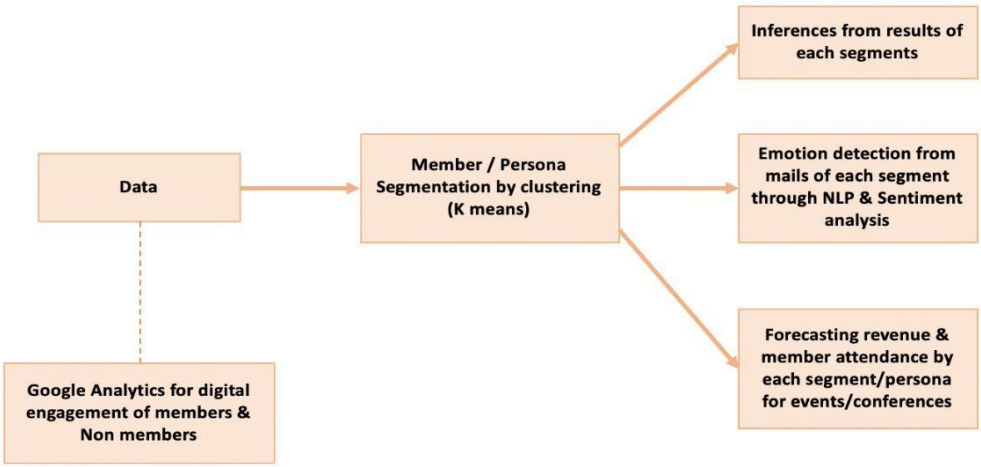

Figure 2 - Analytical approach to the business problem

3. Methodology

i. Data description

The data was provided by the AusIMM organization. The data provided contains three different databases with 1912+ tables in each of the database with over 72000+ records approximately. The data provided is mostly unstructured and contains several nested tables. Hence making we would be applying data pre-processing techniques to clean the data. As AusIMM has a **data privacy policy,** we would not be revealing any information, visualizations and dashboards worked on site. We have used alike dummy data and picturized the report.

Data governance was one of the most important business problems as the schemas in the database were redundant, inconsistent and does not satisfy integrity constraints. Hence data cleaning and segmentation were done by using tools R, Power query editor on Excel. This process was done to pull out raw insights from the structured data. The behavior of members, their profile based on age groups, demographic frequencies, the revenue gap between memberships and events, finance at a glance were done. Also, as the company was moving from data warehouse system to cloud, we provided them a clean and structured data for digital transformation to Microsoft Azure, Dynamics 365 platform.

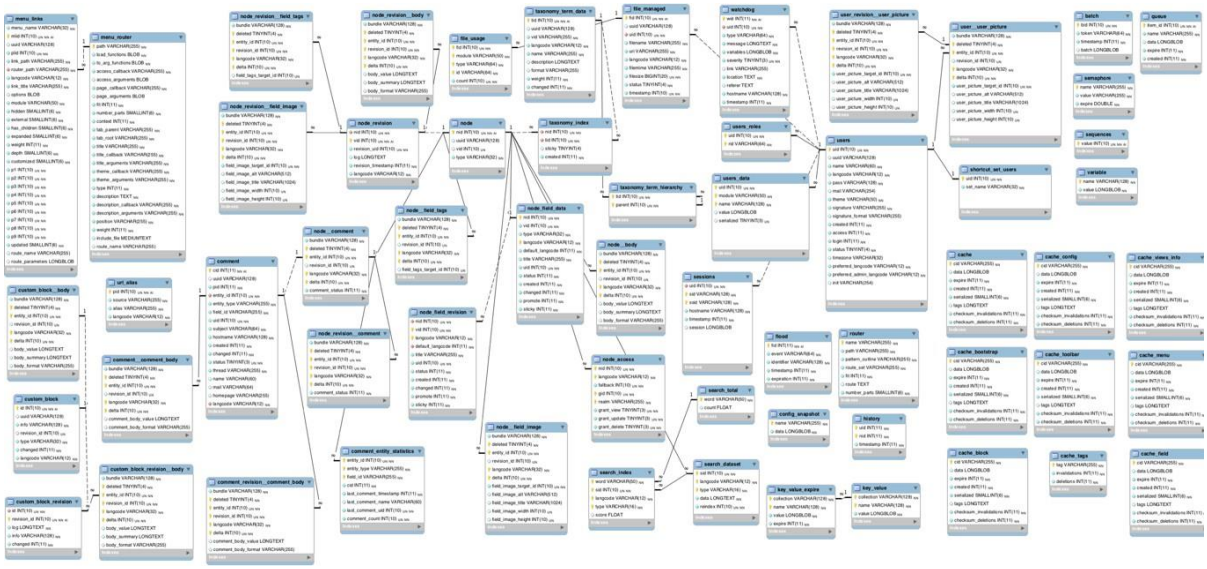

Figure 3 - Integrity constrained nested SQL schema

We worked on similar schema [Figure 3] at AusIMM which had integrity constraints.

We have used ODBC package on R for an efficient and easy setup of connection to MySQL database with ODBC drivers in R to extract tables and convert them to Data frame. (Databases using R from RStudio, n.d.)

```
odbc <- dbConnect(odbc::odbc(), dsn = "PostgreSQL")
```

Figure 4 - Connection to SQL server in R with ODBC

Reading tables – "*dbReadTable()"* will read full tables into an R data frame.

```
data <- dbReadTable(con, [redacted])
```

Figure 5 - Reading SQL tables in R

We have performed unsupervised machine learning techniques, i.e., clustering k-means method to segment and structure the redundant, inconsistent and semi structured database.

ii. Analysis and Machine Learning techniques for 'Persona Segmentation'

We have used unsupervised machine learning techniques, i.e., clustering and segmentation to the data that describes member purchase behavior, membership types, member gender, member occupation, member areas of interests … We used the clustering method as this will give us insights into the underlying patterns of different groups. Also, it helps in reducing the dimensionalities of the data. We worked on both Hierarchical Clustering and K-mean Clustering for the analysis and found the results of k-means to be better with meaningful segments for the available data.

Literature Review:

Customer Segmentation: -

Companies that capture customer and purchase information use such information to analyze and market to their customer base and analysis of this information has become the foundation of database marketing practice. A deeper understanding of customers has validated the value of focusing on them. It is now generally accepted that it costs about five times more to gain a new customer than to keep an existing one, and ten times more to get a dissatisfied customer back **(Massnick, 1997)**. Studies across numerous industries have also shown that a five-point increase in customer retention can increase profits by more than 25 percent **(Reichheld, 1996)**. For every business, the Customer Value Matrix provides an affordable, easy to implement segmentation methodology that delivers substantial value relative to the amount of effort involved. **(Claudio Marcus, 1998)**

- ***K-mean Clustering***:

K-means clustering is the most commonly used unsupervised machine learning algorithm for dividing a given dataset into "K**"** clusters. Here, k represents the number of clusters and must be provided by the user. It starts with K as the input, which is how many clusters, we want to find from the data. We find the value of K using elbow method. For each value of K, we calculate the sum of squared errors (SSE). Once we get the optimal count of clusters, i.e., the value of K, we place K centroids in random locations in our space. Using the "Euclidean distance" between data points and centroids, assign each data point to the cluster which is close to it. We recalculate the cluster centres as a mean of data points assigned to it. The algorithm repeats the same process until there are no further changes occurred.

The standard algorithm that defines the total within-cluster variation as the sum of squared distances, Euclidean distances between items and the corresponding centroid:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- x_i design a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k

Here each observation in x_i is assigned to a given cluster such that the sum of squares distance of the observation to their assigned cluster centres μ_k is minimized.

The total within sum of square or the total within-cluster variation is defined as:

$\sum kk=1W(Ck)=\sum kk=1\sum xi\in Ck(xi-\mu k)2$

The above formula is the summation of all clusters over the sum of squared Euclidean distances between items and their corresponding centroids. (K-Means Clustering in R: Algorithm and Practical Examples - Datanovia, n.d.)

- ***Practical Approach***

We have used R software for clustering analysis with k-means

Step 1: Data Preparation

```
# Load  and prepare the data
data([redacted])

my_data <- [redacted] %>%
  na.omit() %>%          # Remove missing values (NA)
  scale()                # Scale variables

# View the firt 3 rows
head(my_data, n = 3)
```

Figure 6 - Step 1 - Data Preparation for clustering

In step 1 we remove the missing data and scale the variables to make them comparable. (K-Means Clustering in R Tutorial, n.d.)

Step 2: Determining the number of clusters "k" using Elbow method

```
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
```

Figure 7 - Determining the number of clusters using Elbow method

Step 3: Compute and visualise k-means clustering

```
set.seed(123)
km.res <- kmeans(my_data, 3, nstart = 25)
# Visualize
library("factoextra")
fviz_cluster(km.res, data = my_data,
             ellipse.type = "convex",
             palette = "jco",
             ggtheme = theme_minimal())
```

Figure 8 - Computing and visualising K-means clustering

By the results of elbow method, we found the ideal number of clusters to be 8. Hence, we segment the membership database into 8 segments and continue with the analysis. (5 Amazing Types of Clustering Methods You Should Know - Datanovia, n.d.)

---

iii. Sentiment Analysis

Sentiment analysis is important because emotions and attitudes can be used as an actionable piece of information useful in marketing plan and research.

- **Emotion Detection:**

Emotion is expressed in many ways such as facial expression, speech and text documents. Contextual mining for emotion Detection is essentially a content–based classification problem. We have used natural language processing on emails to understand the sentiment the emails deliver to its members. Emotion recognition based on textual data and the techniques used in emotion detection are discussed.

- **Practical Approach**

We have used R software for clustering analysis with k-means

Step 1: Data Preparation

Combining various emails from different sources to form a data frame and extracting the text from the emails into single words.

```
# Combine emails in a data frame

Emails <- do.call(rbind, Map(data.frame, year=seq(2015, 2019),
                  text=c(Email_Marketing, Email_CEO, Email_CustomerSupport)))

Emails$text <- as.character(Emails$text)
```

Figure 9 -Data Preparation for Emotion Detection

Step 2: Descriptive Statistics

Using the code below we plot a boxplot of various emotion and their usage.

```
### Box plot to indicate distribution of emotions
cols  <- colorRampPalette(brewer.pal(7, "Set3"), alpha=TRUE)(8)
boxplot2(emo_box[,c(2:9)], col=cols, lty=1, shrink=0.8, textcolor="red",
         xlab="Emotion Terms", ylab="Emotion words count (%)",
         main="Distribution of emotion words count in inbound emails"
```

Figure 10 - Descriptive Statistics for Emotion Detection

Step 3 Use of emotions over time

```
##  Line chart to depict emotions over time
ggplot(emotions, aes(x=year, y=percent, color=sentiment, group=sentiment)) +
geom_line(size=1) +
geom_point(size=0.5) +
xlab("Year") +
  ylab("Emotion words count (%)") +
  ggtitle("Emotion words expressed over time")
```

Figure 11 - Visualizing Emotions over time

Step 4 Emotions for emails per department.

```
ggplot(emotions_diff, aes(x=year, y=difference, colour=difference>0)) +
geom_segment(aes(x=year, xend=year, y=0, yend=difference),
size=1.1, alpha=0.8) +
geom_point(size=1.0) +
xlab("Emotion Terms") +
     ylab("Net emotion words count (%)") +
     ggtitle("Emotion words expressed from office of CEo emails") +
theme(legend.position="none") +
facet_wrap(~sentiment, ncol=4)
```

Figure 12 - Use of various Emotion per department

iv. Forecasting techniques for operational support [Events/Conferences]

For the analysis, we have followed the following points for building a forecast model and for the forecasting purpose: -

1. The class of the data has to be converted to a time series object thus with the help of ts() function we will change the class of the data.
2. To analyse the existence of trend, presence of seasonality, any evidence of changing variance or any sign of intervention, we will have to plot a Time Series plot.
3. To further understand the pattern in the data, ACF and PACF plot will be used such that to identify the trend in the data.
4. To deal with the serial correlation and heteroskedasticity in the time series, we will be using the ADF (Augmented Dickey-Fuller) test and PP (Phillips-Perron) test.
5. To investigate existing components and effects occurred due to historical data, we will be using the decomposition technique i.e. classical, X12 and STL decompositions.
6. After the decomposition process, we will be building Distributed Lag Models namely, Distributed Lag Model, Polynomial, Koyck and Autoregressive Distributed Lag Model.
7. The Exponential smoothing methods helps in modelling time series data by fitting different trend and seasonality patterns. The state space models include the additive and multiplicate errors to possible exponential smoothing models. Thus, for the exponential smoothing we will be building the few models i.e. Simple Exponential Smoothing, Holt's Linear Method, Holt's Damped trend Model, Holt's Exponential Trend Model, Holt's Winter method and lastly the State Space Model.
8. The residual analysis of the best model selected will be conducted so that to understand the behaviour of the errors of the model.
9. We will be using the selected model to forecast the revenue generated by the company and the attendance of members in the events for the next 2 years.

• Practical Approach

We have used R software for forecasting Revenue and member attendance.

→ Checking stationarity of the series

```
# Unit Root Test
adf.test(rad_ts,k=ar(rad_ts)$order)
```

```
## Warning in adf.test(rad_ts, k = ar(rad_ts)$order): p-value smaller than
## printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  rad_ts
## Dickey-Fuller = -4.557, Lag order = 25, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(Ppt_ts,k=ar(Ppt_ts)$order)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  Ppt_ts
## Dickey-Fuller = -3.2594, Lag order = 28, p-value = 0.07769
## alternative hypothesis: stationary
```

Figure 13 - R chunk for checking stationarity of the series

→ Modelling approach

```
model_dlm  = dlm(x=as.vector(Data1_ts[,2]) ,y=as.vector(Data1_ts[,1]),q = 4 )
summary(model_dlm)
```

```
model_koyckdlm = koyckDlm(x=as.vector(Data1_ts[,2]) ,y=as.vector(Data1_ts[,1]) )
summary(model_koyckdlm$model, diagnostics = TRUE)
```

```
model_polydlm = polyDlm(x=as.vector(Data1_ts[,2]) ,y=as.vector(Data1_ts[,1]) , q = 12 , k = 2, show.beta = TRUE )
```

```
#State-space models for Holt's method
Statespace_1 <- ets(rad_ts, model="ANN")
summary(Statespace_1)

Statespace_2 <- ets(rad_ts, model="MNN")
summary(Statespace_2)

Statespace_3 <- ets(rad_ts, model="AAA")
summary(Statespace_3)

Statespace_4 <- ets(rad_ts, model="MAA")
summary(Statespace_4)

Statespace_5 <- ets(rad_ts, model="MMM" , damped = TRUE)
summary(Statespace_5)
```

Figure 14 - DLM and SES models

```
#Simple Exponential Smoothing
model_Ex <- ses(rad_ts, alpha=0.8, initial="simple", h=2)
summary(model_Ex)

#Holt's Linear Method
model_holt <- holt(rad_ts, alpha=0.8, beta=0.1, initial="simple", h=2)
summary(model_holt)

#Holt's Exponential Trend Method
model_Ex_trend <- holt(rad_ts, alpha=0.1, beta=0.8, initial="simple", exponential=TRUE, h=2)
summary(model_Ex_trend)

#Holt's Damped Trend Method
model_damped <- holt(rad_ts, alpha=0.8, beta=0.3, damped=TRUE, initial="simple", h=2)
summary(model_damped)

#Holt Winter Additive
fit_1 <- hw(rad_ts,seasonal="additive", h=2*frequency(rad_ts))
summary(fit_1)

#Holt Winter Additive Damped
fit_2 <- hw(rad_ts,seasonal="additive",damped = TRUE, h=2*frequency(rad_ts))
summary(fit_2)

#Holt Winter Multiplicative
fit_3 <- hw(rad_ts,seasonal="multiplicative", h=2*frequency(rad_ts))
summary(fit_3)
```

Figure 15 - State space modelling with Holt's Method

v. Google Analytics for digital engagement

The Google Analytics is a tool which offers information about the user behaviour that can be critical for the business. The user must add a tracking code of their website first. The tracking code (made up of a programming language i.e. JavaScript) is a code through which Google Analytics tracks the website visitors and any action performed by the visitors. The activity of the users is collected by the cookie which Google Analytics drop on the user's browser. By using these cookies, Google Analytics will know how a user behaves on the website and then collects this information to show different reports. The Google Analytics gives a quick overview of how the website is performing. It shows the following things: -

- Users: - Number of visitors who visited the website in past 7 days.
- Sessions: - Number of interactions a visitor makes with the website in a time frame i.e. viewing a page, clicking a link or purchasing a product.
- Bounce Rate: -The number of users who hit the back button or closed the website without performing a single interaction.
- Session Duration: - The average time user spends on the website.
- Active Users: - Number of users who are currently active on the website.

The Google analytics has five reporting options: -

- Audience: - It helps in breaking down the web traffic according to the age of the visitors or the device they are using.

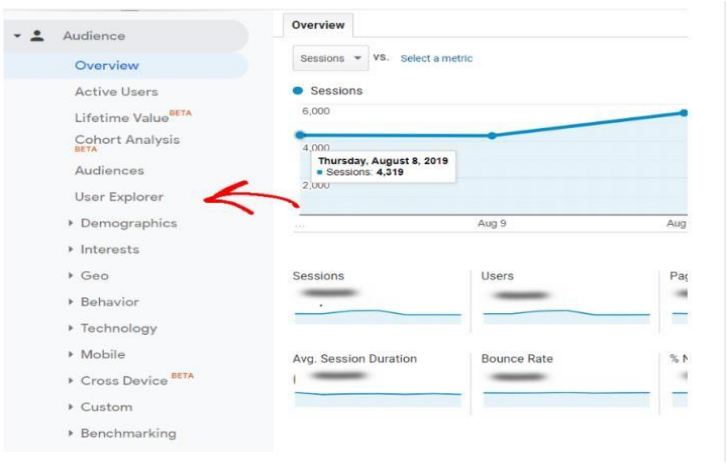

Figure 16 - Screen of Audience

- Acquisition: - The acquisition reports tells how the traffic reaches the website. The traffic is divided into four categories namely, Organic Search (traffic comes from search engines), Direct(traffic comes when someone types the website's URL or when Google cannot recognise the traffic source), Referral (traffic which comes from any source other than search engines) and Social (traffic which comes from social media websites like Facebook or Twitter).

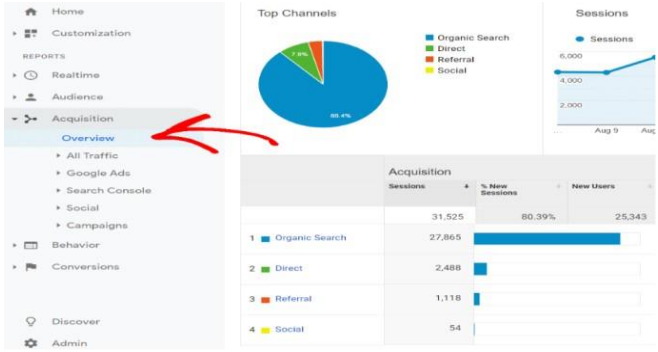

Figure 17 - Image explaining Acquisition

- Behaviour: - The behaviour report helps in understanding what the visitors are doing on the website. The behaviour of the visitors can be understood by looking at Pageviews (total number of pages viewed by the visitor), Unique pageviews (when an individual user views a certain page at least once on the website), Average Time on Page (the average amount of time spend by user on viewing a webpage), Bounce Rate (the percentage of users that viewed only a single page and left after interacting with it) and Percentage Exit (how often visitors exit the websites' page.

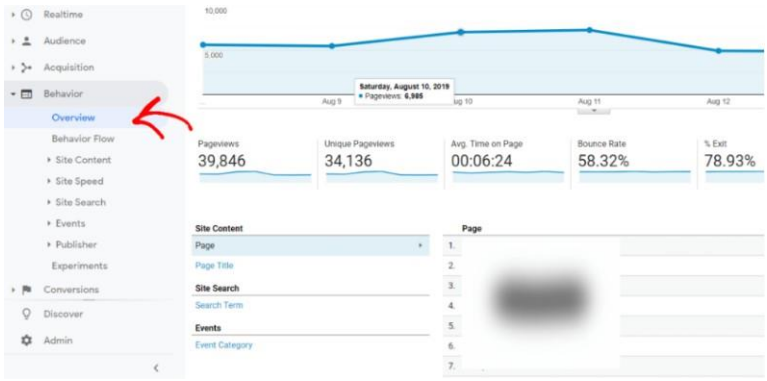

Figure 18 - screen of Behaviour in Google analytics

- Conversions: - It tells how the conversion rate of the website is performing. The conversion rate is simply any activity completed by a visitor. It can include anything like downloading a video, buying a product or subscribing to a newsletter.

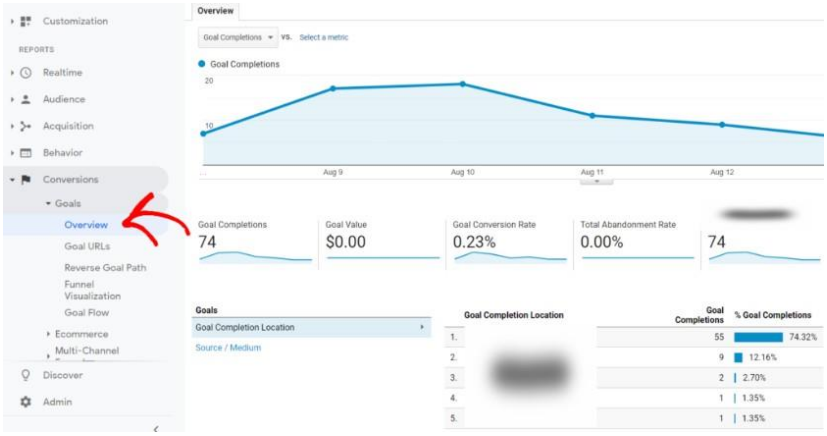

Figure 19 - Conversions heat on google analytics

4 Results

→ **K-mean Clustering**

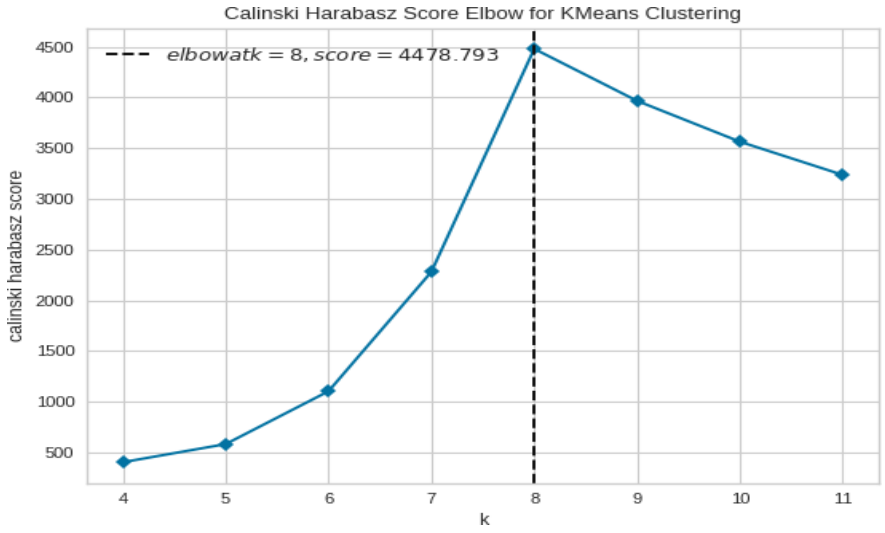

Figure 20 - Elbow method

From the results of clustering analysis, we found that the optimal number of clusters to be 8. We decided to segment the data into 8 parts and understand customer segments clearly. (Determining The Optimal Number Of Clusters: 3 Must Know Methods - Datanovia, n.d.) (Elbow Method, n.d.)

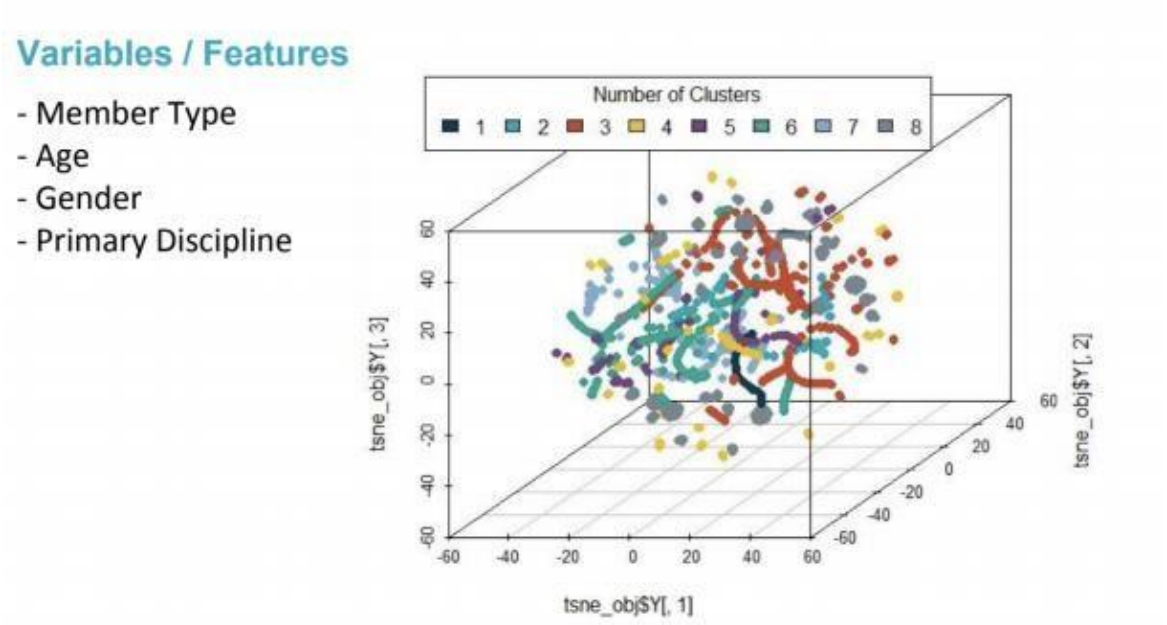

Figure 21 – Clustered 8 segments in 3D pane

We conceived various features for improving the clustered data to better describe the segments. The new features included “Branch Location” (we assigned everyone into Regional, Urban or International based on their postcode); “Professional Level” (we parsed the Job Title discipline and assorted everyone into one of 14 widespread levels).

→ Segments at AusIMM before Modelling approach.

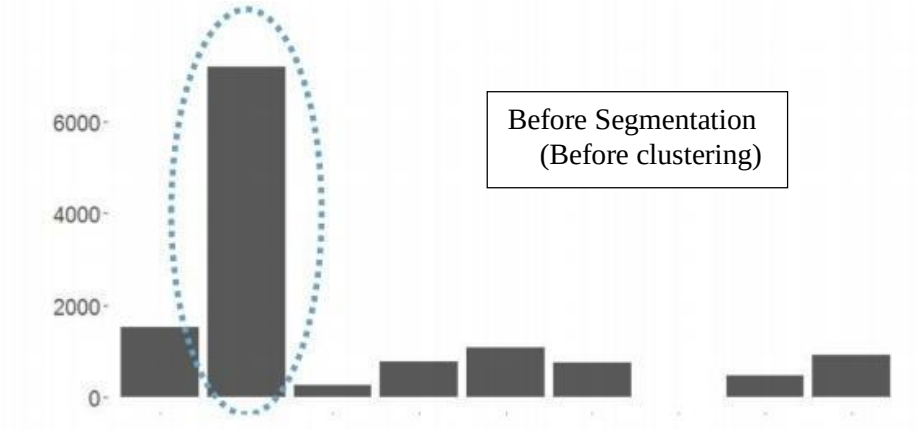

Figure 22 - Membership status before segmentation

→ Customer/Member Segmentation after clustering

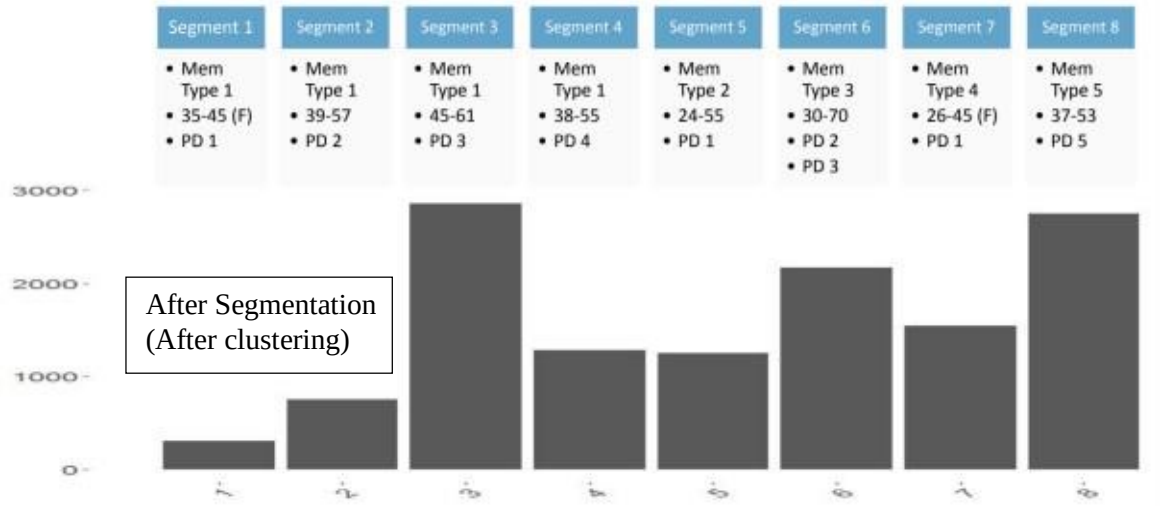

Figure 23 - Results of Clustering

From Figure 22 we understood that one of the segment had a huge spike which had around 6000+ members. To break that segment and understand segment in better manner we used k-means. From the results of k-means [figure 23], i.e. 8 segments it is clear to understand the demographics and areas of interests of people in each segment.

Our investigation didn't end with the clustering analysis. We extended proving descriptive titles for each cluster so that the company could understand them thoroughly, and we laid out clear company rules to classify members into the clusters so that they wouldn't need to rely on the clustering algorithm to classify new people.

→ Sentiment Analysis

Sentiment analysis steps of textual data encapsulates word tokenization, pre-processing of tokens to exclude stop words, followed by aggregation and presentation of results. Word tokenization is the process of separating text into single words or unigrams.

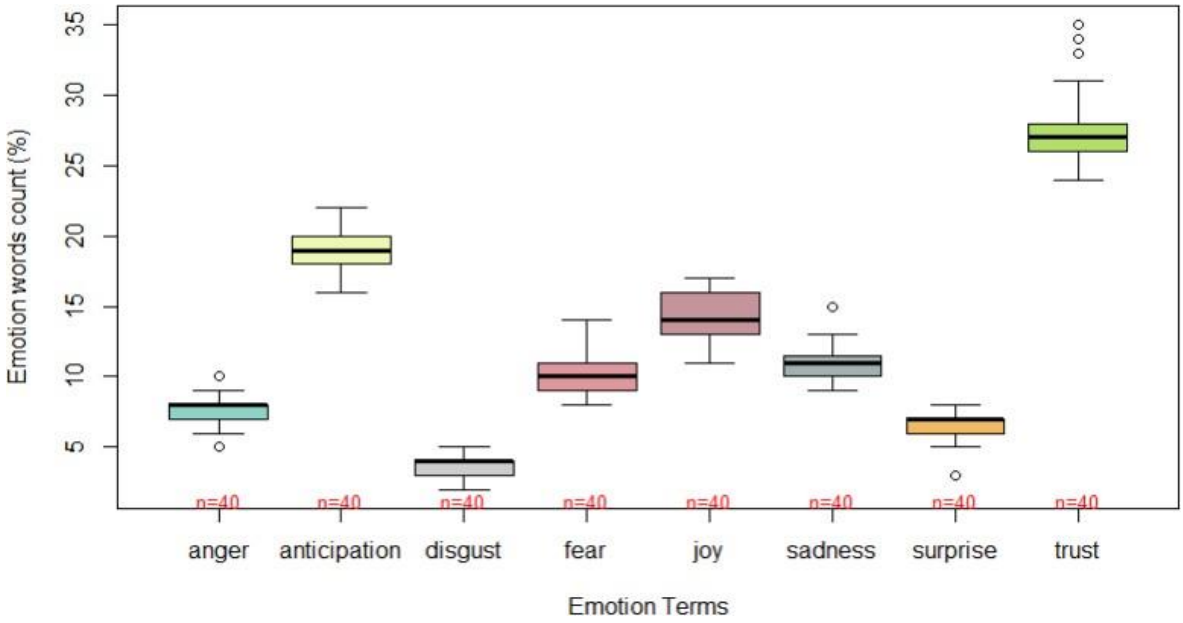

Figure 24 - Boxplot of emotions by count

Terms for all eight emotions types were expressed with the help of boxplot. It indicates that count for positive emotions are quite evidently higher. Anger, sadness, surprise and trust had few outliers, whereas Joy was skewed to the right.

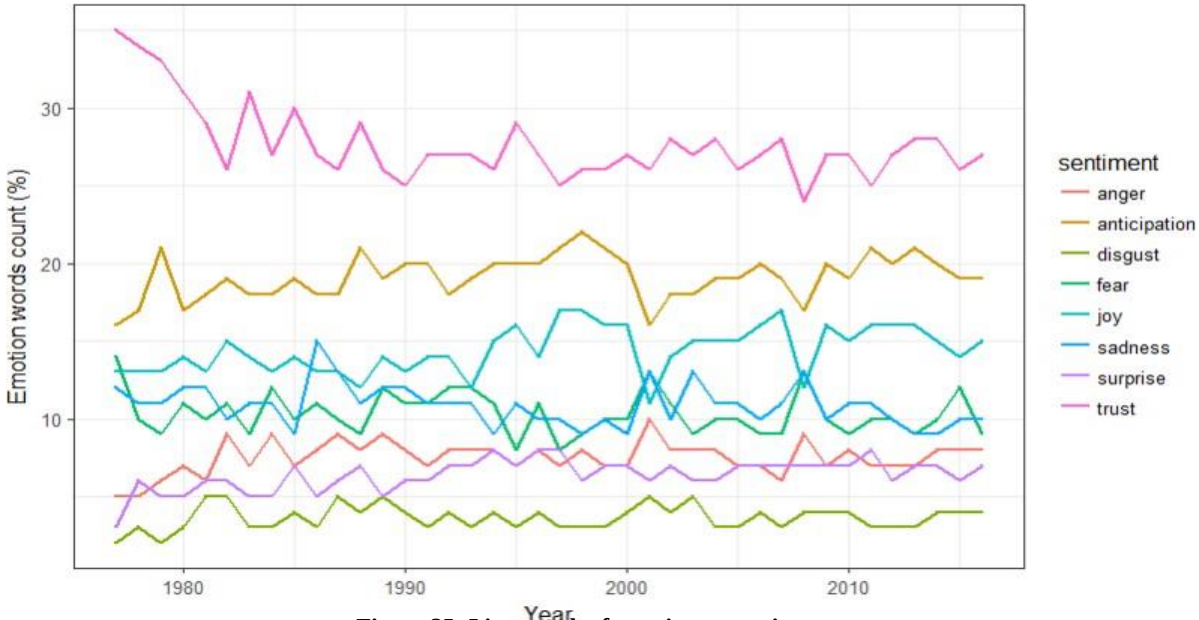

Figure 25 -Line graph of emotions over time

From the line graph it's quite clear that emotion terms referring to trust and anticipation were expressed consistently higher than the other emotion terms in all the emails sent by the AusIMM organization. Furthermore, emotions referring to disgust, anger and surprise were expressed consistently lower.

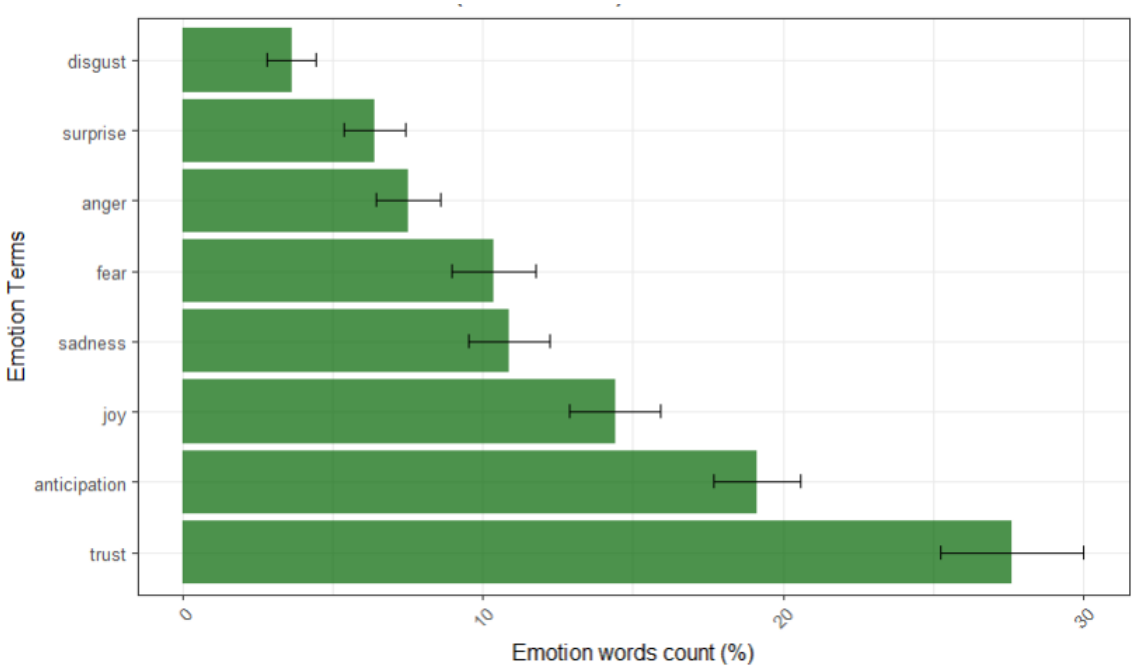

Figure 26 - Bar graph of emotions use and error

The above plot indicates the average of emotion words expressed using bar charts with error bars. Emotions referring to trust, anticipation and joy accounted for approximately 60% of all emotion words in all the emails.

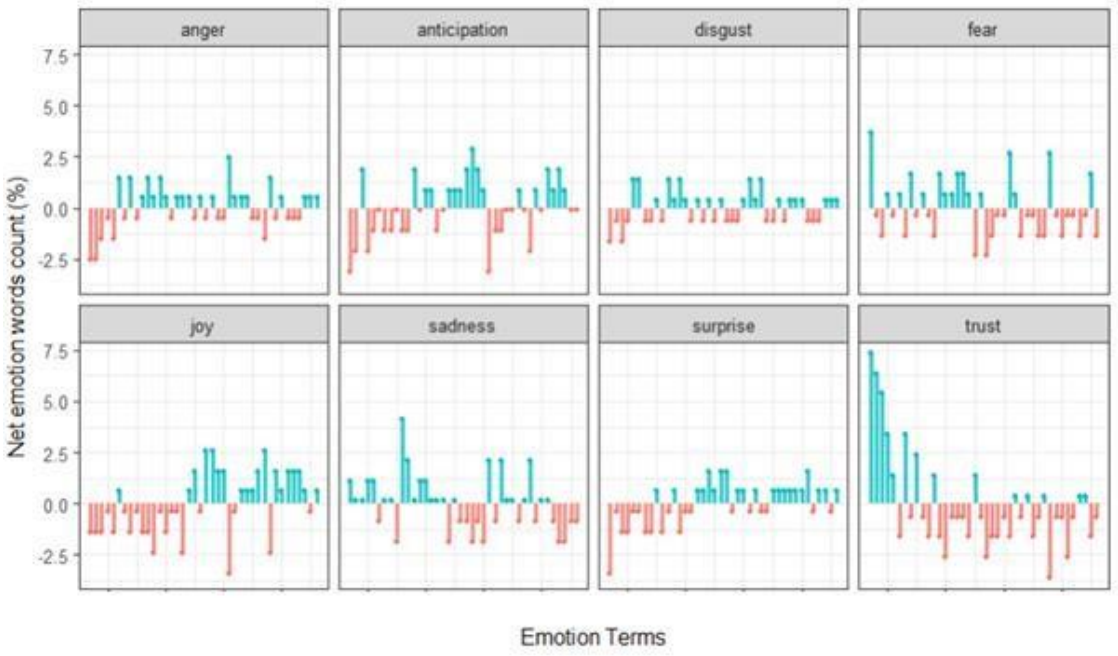

Figure 27 - Emotion term in emails per department

In conclusion, approximately 1 in 4 words in the emails represented emotion terms. Clearly emotion terms referring to positive sentiments like trust, anticipation and joy where higher than negative emotions like sadness, fear and anger. Moreover, there were very limited emotions of fear approximately 1 in 10 emotion terms.

→ Forecasting and Time Series Analysis

Time Series analysis and decomposition for event cost and number of people attending the event was performed to understand the underlying structure of the cost to host a networking conferences and events for the Metallurgy and mining industry. (Srivastava, 2016)

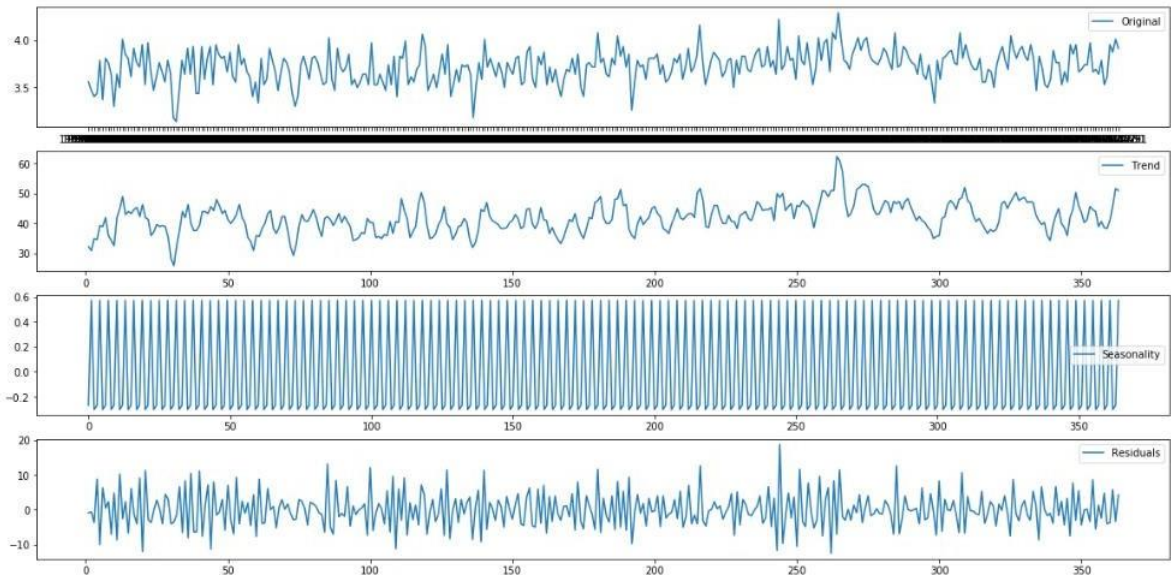

Figure 28 - Event cost time-series decomposition

Both Distributed lag models and state space model were used to assess the cost to serve an event. Error Rate for all the forecasting techniques are recorded in the form of Mean Absolute Scaled error (MASE) value. MASE is a scale free error rate providing error measure with no bias.

	n <dbl>	MASE <dbl>
model.dlm	88	4.6472800
model.koyck	88	0.8818111
model.ar	86	0.8633163
model.poly	81	3.8838937

Figure 29 - Model Summary statistics

State Space model for Additive error and Additive trend with no seasonality was the selected model with the least MASE values.

FittedModels <chr>	AIC <dbl>	AICc <dbl>	BIC <dbl>	HQIC <dbl>	MASE <dbl>
ANN	430.8430	431.1254	438.3089	430.8492	0.2317327
MNN	420.2628	420.5452	427.7287	420.2690	0.2317153
AAN	423.8938	424.6167	436.3370	425.9062	0.2202870
AAA	450.2939	458.9137	492.6008	464.3435	0.2285203
MAN	415.2485	415.9714	427.6917	417.2609	0.2203648
MMM	462.1425	471.9139	506.9379	477.1952	0.2364370

Figure 30 - Error rates of state space models

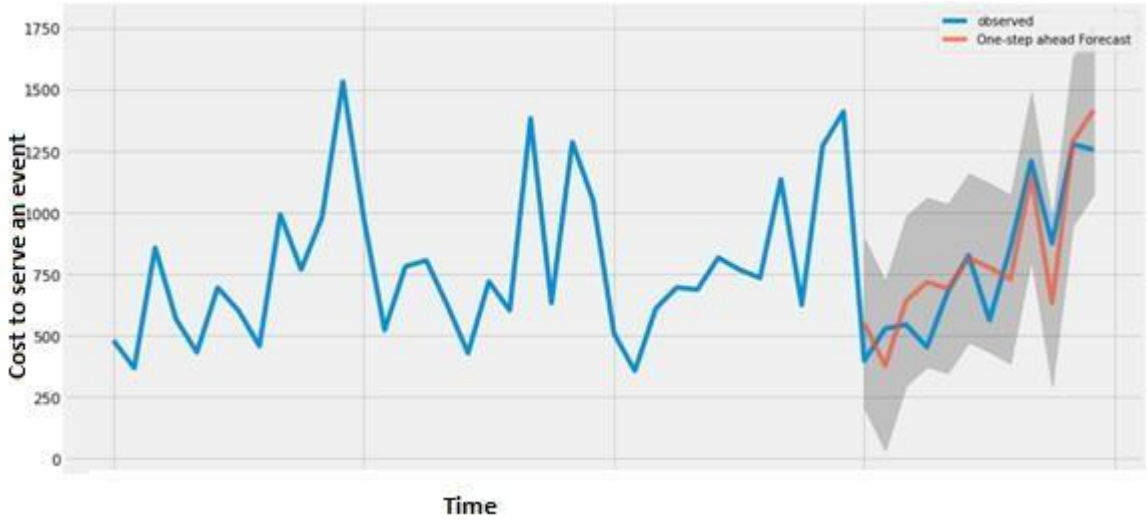

Figure 31 - Forecasting for cost to serve an event

Forecasting the number of attendees provides a good indicator in providing the cost to serve that big or small-scale events and conferences. Events that caters for the iron ore industry professionals, students, industry leaders and others had a significant number of attendees both member and non-member. (Oehm, 2018)

Various forecasting methods were modelled to find the best suitable forecasting technique with the least error rate within a 95% confidence interval.

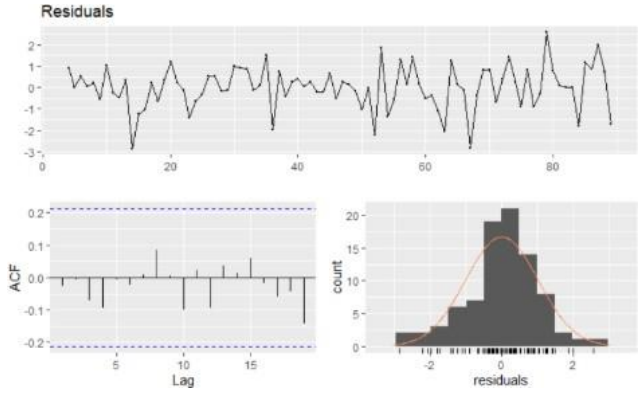

Figure 32 - Residual Analysis of the best model

The residual analysis for the selected model using state space modelling shows there is no autocorrelation left within the series. Also, we can conclude that the residuals are randomly distributed with a histogram following an approximate normal bell cure lightly skewed to the right.

→ Google Analytics

Google analytics dashboard provided us with peculiar insights like the difference in demographics, area of interest and behaviour with respect to member information. Age information from google analytics along with gender have different segments.

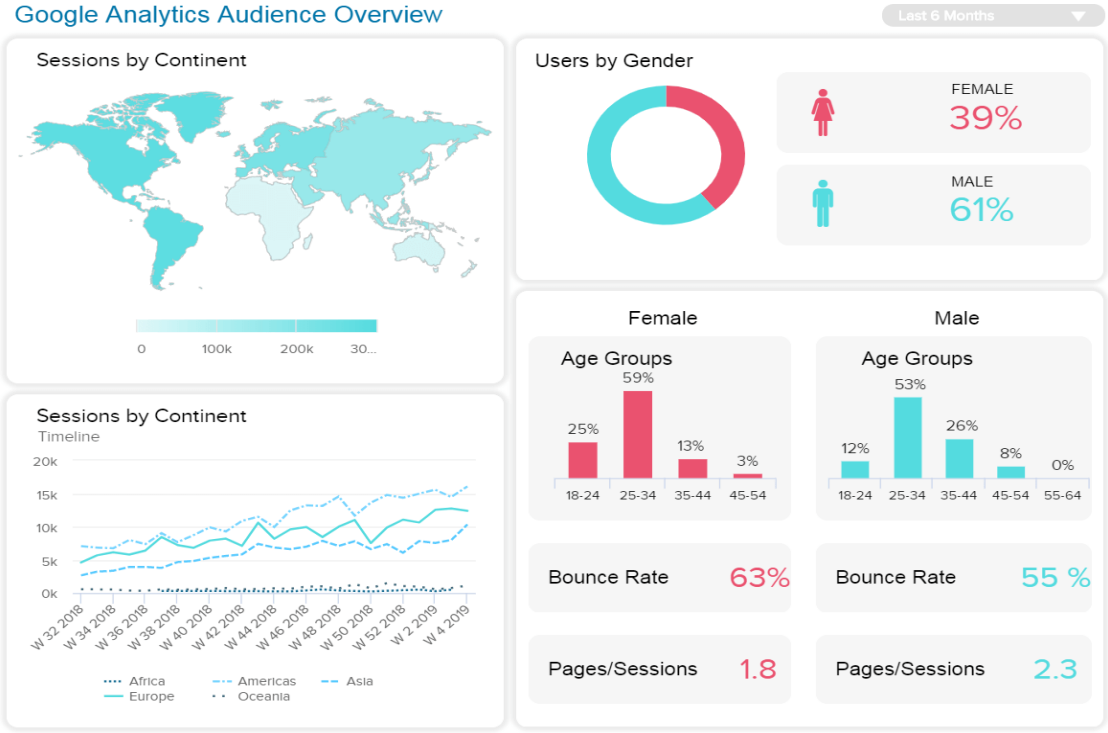

Figure 33 -Dashboard for demographic overview

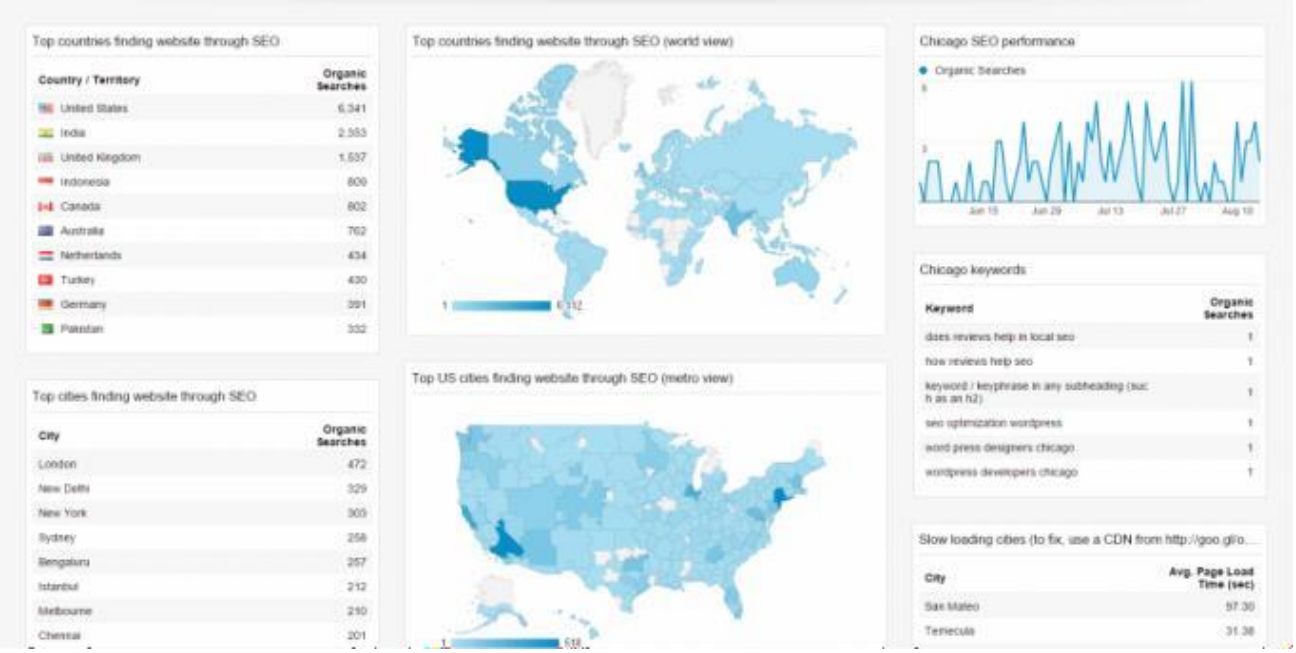

Figure 34 - Geography of Customer base

Google analytics geography dashboard also shows the various language preferences, country which can be targeted as prospective members and cities with slow load time. Furthermore, it also indicates what pages people enter the most, what pages people exit from the most, average pages viewed per visit, bounce rates and pages with the highest bounce rate to understand the what news and events interests' people the most visit.

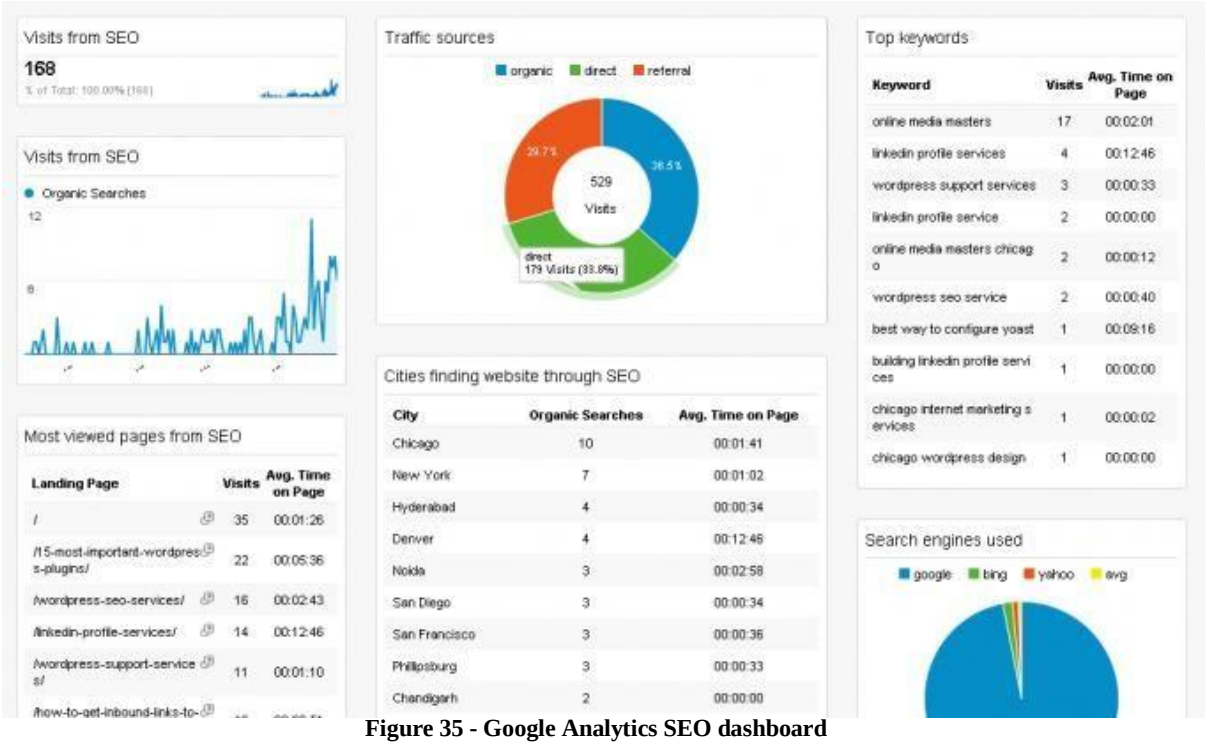

Figure 35 - Google Analytics SEO dashboard

The Google Analytics SEO dashboard indicates unique visitors, Top keyword, Search engines used and Cities finding for website. Various traffic source along with connection provider details along with countries.

5 Conclusion and Recommendation

→ I suggested some key data validation that they should introduce in their new database so that their data requires less cleaning in future.

→ I discussed with them the idea of adding a Data Warehouse layer to their new Azure infrastructure plans, to help them combine their different data sources and allow their business lines to easily query it using Power BI.

→ They are trying to decide between a few different options for storing their member data in the new system, including storing it in multiple places and syncing between them, or storing them only in one place and reducing access to the data from other sources. We helped them work through the pros and cons of these options from a data integrity and reporting perspective.

→ Segment Analysis provided us with 8 distinct segments of the member/customer base for the Australian Institute of Metallurgy and Mining. These segments have specific and special characteristics that will stand out and help the organization in devising a marketing plan. This would heighten the member base and act as a revenue-generating factor.

→ Sentiment Analysis of the emails for each department unearths the emotional effect the emails carry for its reader. The emotion detection method used in the analysis showcases that emotions referring to trust, anticipation, and joy are more frequent and recurring than emotions fear, anger, and disgust.

→ Events and conferences hosted by the organization for networking and collaborations are modelled using forecasting models to help us estimate a cost to service an event as well as the average number of members attending an event. It also provides information for prospective members attending the events.

→ Google Analytics helps in interpreting the behavior and understanding the demographics for both members and non-members. The dashboard additionally symbolizes the diverse behavior, social engagement, targeted keywords, and web analytics for the Australian Institute of Metallurgy and Mining site.

→ The Forecasting summary provided to the company will help in budgeting their expenses, bringing in more investors for events considering the number of attendants in the events/conferences, assists in strategic planning, boosts decision making for production scheduling and to learn customer behavior. Overall forecasting results speak the eventuality and guide the company to be ready for the doom

→ Lifetime value can be formulated if the data governance model is incorporated.

As Data analytics is a potent, most advanced and arising notion for the mining industry, we couldn't find any tangible literature review associated with the realm. We are the first to correlate the mining business with customer segmentation, natural language processing, and machine learning techniques. The research factor from our judgment is that the mining industry can implement a "Market-based" approach [with booming machine learning and Data Analytics] to drive revenue and understand its customer's in more precise terms.

6. Reflection of the work completed - Academic vs Operational environment

→ We successfully used a range of analytic tools and models in the actual real-world environment. (Text Mining, Clustering, Segment Profiling, etc.)

→ Validation of the fact that 70% of a data analyst's time is spent on preparing the data.

→ Soft skills and the ability to articulate queries were indeed helpful in discussion with various departments to understand their expectations from analytics.

As AusIMM has a data privacy policy, we are not deemed to accord any data or codes used for the research done on-site, so there is no appendix attached with R-codes for this report. I have bestowed all the inferences acquired from the Work Integrated AusIMM Project in this report.

7 References

(n.d.). Retrieved from https://www.drupal.org/files/Drupal8_UPsitesWeb_Schema_10-19-2013.png

(n.d.). Retrieved from https://www.drupal.org/files/Drupal8_UPsitesWeb_Schema_10-19-2013.png

5 Amazing Types of Clustering Methods You Should Know - Datanovia. (n.d.). Retrieved from Datanovia: https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/

Akhtar, A. (2019, Sep 17). *How does google analytics work.* Retrieved from Monster Insights: https://www.monsterinsights.com/how-does-google-analytics-work-beginners-guide/

Beattie, A. (2019, June 25). *Business Forecasting: Understanding the Basics*. Retrieved from Investopedia: https://www.investopedia.com/articles/financial-theory/11/basics-business-forcasting.asp

Clustering — Unsupervised Learning. (n.d.). Retrieved from Towards Data Science: https://towardsdatascience.com/clustering-unsupervised-learning-788b215b074b

Databases using R from RStudio. (n.d.). Retrieved from odbc: https://db.rstudio.com/

Determining The Optimal Number Of Clusters: 3 Must Know Methods - Datanovia. (n.d.). Retrieved from Datanovia: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Elbow Method. (n.d.). Retrieved from YellowBrick: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

Garbade, D. J. (2016, Oct 16). *A Simple Introduction to Natural Language Processing*. Retrieved from Medium: https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Gove, R. (2017, Dec 26). *Using the elbow method to determine the optimal number of clusters for k-means clustering*. Retrieved from Robert Gove's Block: https://bl.ocks.org/rpgove/0060ff3b656618e9136b

guru99. (n.d.). Retrieved from Unsupervised Machine Learning: What is, Algorithms, Example: https://www.guru99.com/unsupervised-machine-learning.html

Guru99. (n.d.). Retrieved from https://www.guru99.com/unsupervised-machine-learning.html

Hines, K. (2015, june 24). *The Absolute Beginner's Guide to Google Analytics*. Retrieved from MOZ: https://moz.com/blog/absolute-beginners-guide-to-google-analytics

K-Means Clustering in R Tutorial. (n.d.). Retrieved from Datacamp: https://www.datacamp.com/community/tutorials/k-means-clustering-r

K-Means Clustering in R: Algorithm and Practical Examples - Datanovia. (n.d.). Retrieved from Datanovia: https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/#computing-k-means-clustering-in-r

Oehm, D. (2018, may 8). *State space models for timeseries analysis and dlm package*. Retrieved from Gradient descending: http://gradientdescending.com/state-space-models-for-time-series-analysis-and-the-dlm-package/

Roman, V. (2012, March 7). *Un supervised machine learning - Clustering analysis*. Retrieved from Towards Data science : https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e

Sarkar, D. (. (n.d.). *Emotion and Sentiment Analysis: A Practitioner's Guide to NLP - KDnuggets.* Retrieved from kdnuggets: https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html

Sentiment Analysis: . (n.d.). Retrieved from MonkeyLearn: https://monkeylearn.com/sentiment-analysis/

Sentiment Analysis: Concept, Analysis and Applications. (n.d.). Retrieved from towardsdatascience: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Srivastava, T. (2016, December 16). *A complete tutorial on time series modeling in r*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

Tuovila, A. (2019, Aug 8). *Forecasting*. Retrieved from Investopedia: https://www.investopedia.com/terms/f/forecasting.asp