# CUSTOMER SATISFACTION – ISSUES AND ANALYSIS: A COMPARATIVE STUDY IN TELECOMMUNICATION INDUSTRY

## DATA COLLECTION PROCESS:

In this project, we have done comparative analysis of two telecommunication companies, namely, AT&T and Verizon. As per Statista, the top two wireless carrier/operator subscriber share in the U.S. 2020 were AT&T and Verizon, hence we considered these two companies for our analysis.

We have collected data from Consumer Affairs (https://www.consumeraffairs.com/). Consumer Affairs provides verified reviews of customers of AT&T and Verizon. We collected a total of 13,749 reviews through web scraping. The total of AT&T reviews is 6838 and Verizon reviews is 6911.

Our dataset consists of 4 columns namely, Date, Reviewer, Content, and Rating. Date mentions the date on which the review was written. Reviewer consists of the name of the individual who has written the review. Content is the detailed review mentioned by the reviewer about the respective telecom companies. Rating is the rating given to the telecommunication provider on the scale of 1 to 5, 1 being the least rating score and 5 being the highest rating score.

## PREPROCESSING PROCESS:

Through data collection process we collected raw data. Data Preprocessing is a data mining technique through which raw data is transformed into an understandable format to do effective and efficient data analysis.

The dataset collected through web scraping was incomplete, inconsistent, and contained few errors. Through data preprocessing we have tried to resolve these issues.

The AT&T and Verizon dataset consisted of 6838 reviews and 6911 reviews, respectively. The datasets consisted of 4 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6838 entries, 0 to 6837
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      6838 non-null   object
 1   Reviewer  6838 non-null   object
 2   Content   6838 non-null   object
 3   Rating    6101 non-null   float64
dtypes: float64(1), object(3)
memory usage: 213.8+ KB
```

**AT&T**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6911 entries, 0 to 6910
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      6911 non-null   object
 1   Reviewer  6911 non-null   object
 2   Content   6911 non-null   object
 3   Rating    5601 non-null   float64
dtypes: float64(1), object(3)
memory usage: 216.1+ KB
```

**Verizon**

**Convert the Date "object" datatype to "datetime64" datatype:**

As we can see from the above table, the Date, Reviewer, Content is in "object" datatype and Rating is in "float64" datatype. Hence, for further analysis we shall convert Date from "object" datatype to "datetime" datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6838 entries, 0 to 6837
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      6838 non-null   datetime64[ns]
 1   Reviewer  6838 non-null   object
 2   Content   6838 non-null   object
 3   Rating    6838 non-null   float64
dtypes: datetime64[ns](1), float64(1), object(2)
memory usage: 213.8+ KB
```

**AT&T**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6911 entries, 0 to 6910
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      6867 non-null   datetime64[ns]
 1   Reviewer  6911 non-null   object
 2   Content   6911 non-null   object
 3   Rating    5601 non-null   float64
dtypes: datetime64[ns](1), float64(1), object(2)
memory usage: 216.1+ KB
```

**Verizon**

**Handling missing values:**

```
Date          63
Reviewer       0
Content        0
Rating       737
dtype: int64
```

**AT&T**

```
Date          44
Reviewer       0
Content        0
Rating      1310
dtype: int64
```

**Verizon**

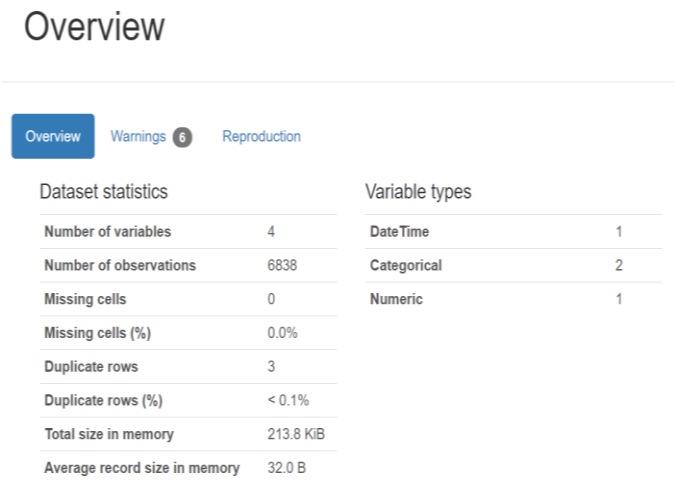
As we can see from the above table, in AT&T dataset there are 63 null values in Date column and 737 null values in Rating column, in Verizon dataset there are 44 null values in Date column and 1310 null values in Rating column.

We shall fill these null values so that there are no inconsistencies in the dataset and analysis can be done properly. For Date column we shall forward fill the null values. The reason for using forward fill method is because it is a timeseries dataset and forward fill method will propagate last valid observation forward in the Date column.

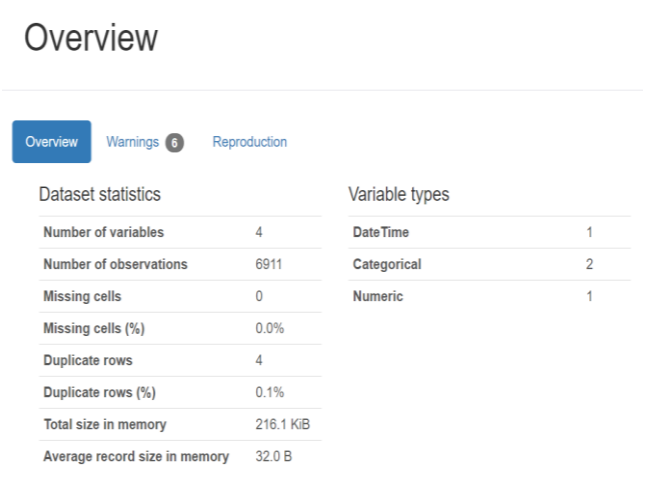
For Rating column, we shall fill the missing values in it by replacing the missing values with a constant variable known as "0". The "0" denotes that the rating is not given by the reviewer.

## Pandas Profiling:

Pandas profiling an open source python module is used for exploratory data analysis. It helps us in understanding the distribution of data.

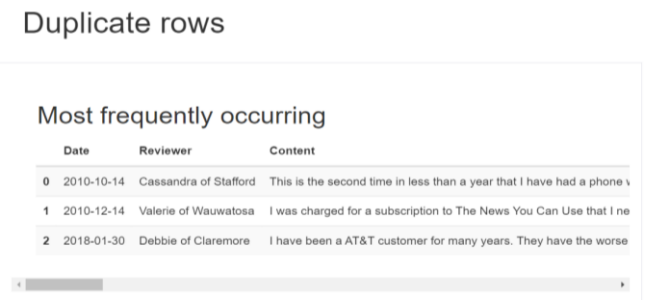



**AT&T**

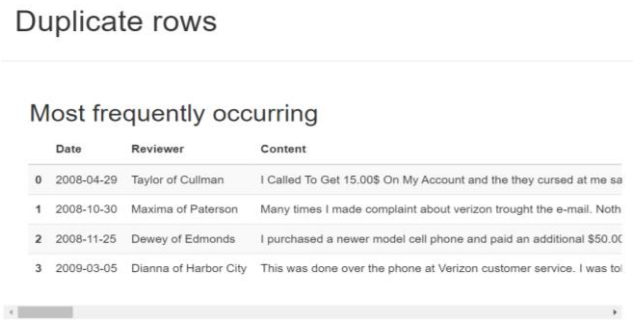



**Verizon**

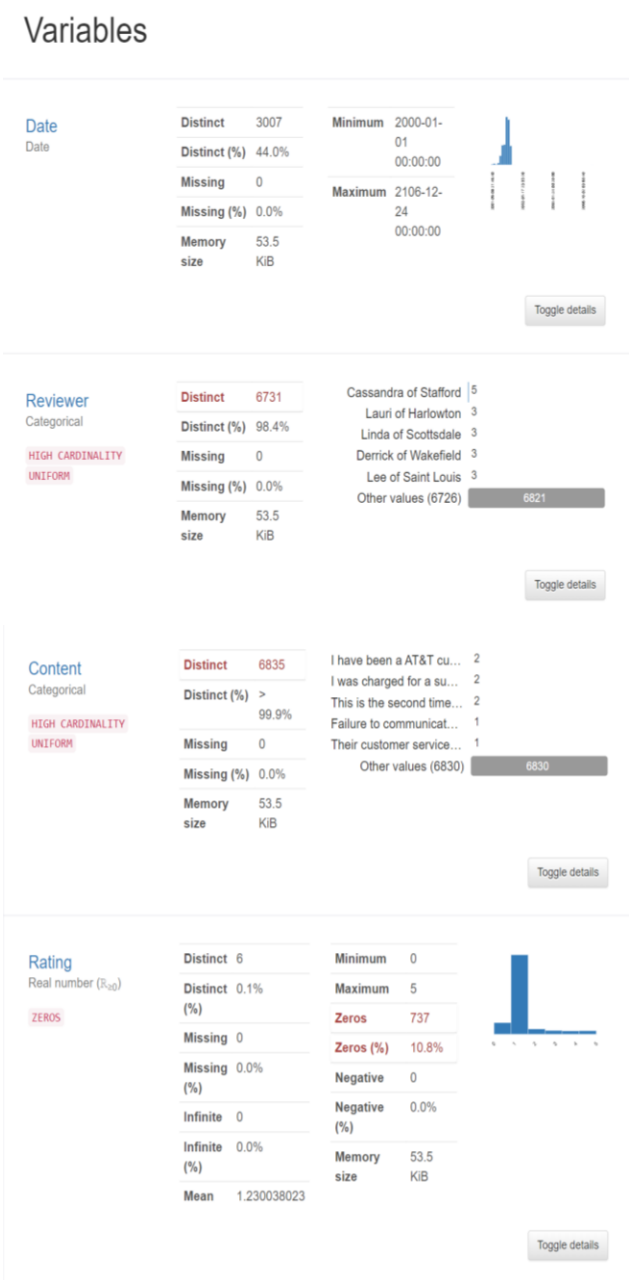
From the above table, we get to see an overview about the dataset. Also, we can see that there are duplicate rows present in the dataset. The AT&T dataset consists of 3 duplicate rows and Verizon dataset consists of 4 duplicate rows. There is 1 DateTime variable, 2 categorical variable, and 1 numeric variable in both the datasets.

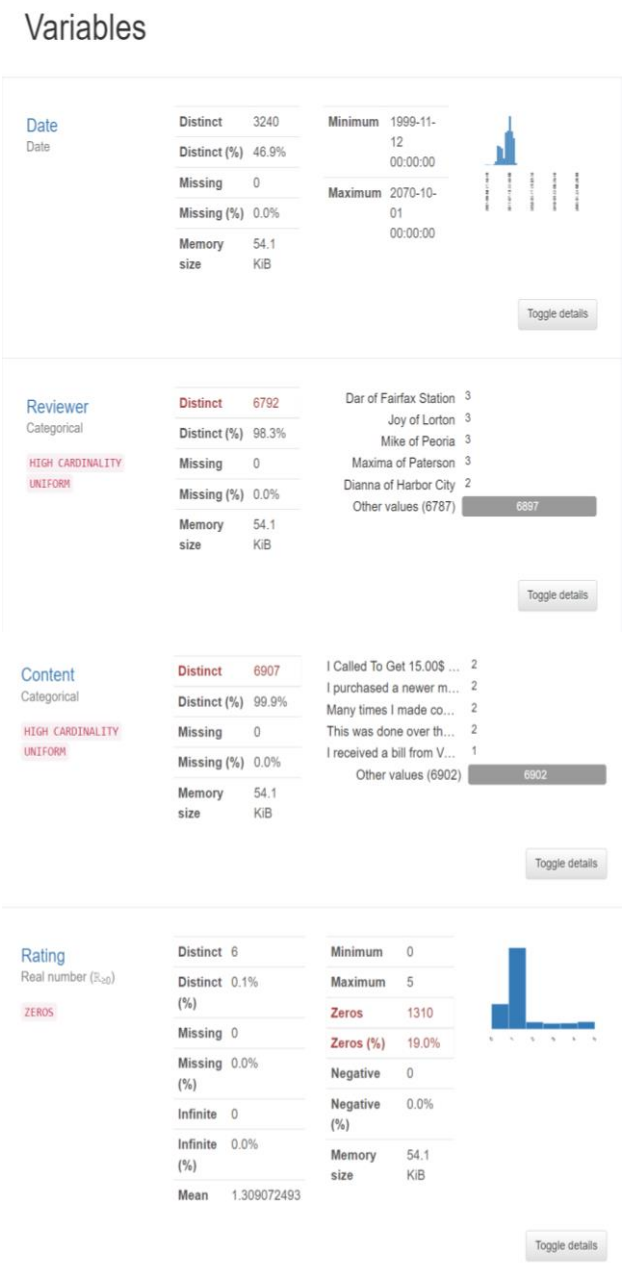



**AT&T**




**Verizon**




**AT&T** **Verizon**

As we can see from the above tables, detailed description of each variable is provided through pandas profiling. Also, we get to know there are invalid date present in the datasets. The description of date variable shows that the maximum date present in AT&T dataset is “2106-12-24” and in Verizon dataset is “2070-10-01”. These dates consist of discrepancy because the years mentioned in both the dates are incorrect.

**Handling Duplicated Data:**

The dataset consisted of a total of 7 duplicate entries, 3 in AT&T dataset and 4 in Verizon dataset. The 7 duplicated rows are being deleted. Therefore, AT&T consist of 6835 rows and Verizon consist of 6907 rows after dropping duplicate rows.

**Handling Invalid Values:**

| Date | |
|---|---|
| 2000 | 15 |
| 2001 | 7 |
| 2002 | 11 |
| 2003 | 36 |
| 2004 | 9 |
| 2005 | 3 |
| 2006 | 17 |
| 2007 | 28 |
| 2008 | 77 |
| 2009 | 109 |
| 2010 | 339 |
| 2011 | 365 |
| 2012 | 366 |
| 2013 | 190 |
| 2014 | 636 |
| 2015 | 1070 |
| 2016 | 794 |
| 2017 | 989 |
| 2018 | 787 |
| 2019 | 637 |
| 2020 | 310 |
| 2021 | 38 |
| 2024 | 1 |
| 2106 | 1 |

Name: Date, dtype: int64

**AT&T**

| Date | |
|---|---|
| 1999 | 1 |
| 2000 | 9 |
| 2001 | 6 |
| 2002 | 13 |
| 2003 | 12 |
| 2004 | 3 |
| 2006 | 35 |
| 2007 | 41 |
| 2008 | 303 |
| 2009 | 367 |
| 2010 | 412 |
| 2011 | 306 |
| 2012 | 254 |
| 2013 | 133 |
| 2014 | 463 |
| 2015 | 900 |
| 2016 | 621 |
| 2017 | 1119 |
| 2018 | 717 |
| 2019 | 537 |
| 2020 | 551 |
| 2021 | 103 |
| 2070 | 1 |

Name: Date, dtype: int64

**Verizon**

As we can see from the above table, AT&T consist of 2 invalid information in the date columns and Verizon consists of 1. These entries have been deleted from the dataset. Now the AT&T dataset consist of 6833 rows and 4 columns and Verizon dataset consist of 6906 rows and 4 columns.

**Training and Testing Dataset:**

The training and testing dataset has been created for our analysis. The training dataset if from the period 2015 to 2020 and testing dataset is of 2021. The training dataset of AT&T contains 4587 rows and 4 columns, and the testing dataset contains 38 rows and 4 columns. The training dataset of Verizon consists of 4445 rows and 4 columns and testing dataset consists of 103 rows and 4 columns.

**EXPLORATORY DATA ANALYSIS:**

Exploratory data analysis helps in investigating and summarizing datasets. It also helps in determining the appropriateness of the statistical analysis method being considered for data analysis.

## Description of the variable "Rating":

The mean of rating for AT&T is 1.435 and for Verizon is 1.712. The standard deviation for AT&T is 1.060 and for Verizon is 1.318. The minimum rating for both AT&T and Verizon is 0 that is few reviewers have not given any rating and the maximum rating received is 5. The percentile distribution of rating can also be seen in the below table.

| | Rating |
|---|---|
| count | 4587.000000 |
| mean | 1.435579 |
| std | 1.060913 |
| min | 0.000000 |
| 25% | 1.000000 |
| 50% | 1.000000 |
| 75% | 1.000000 |
| max | 5.000000 |

**AT&T**

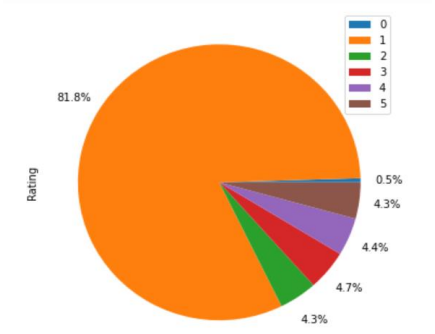
| | Rating |
|---|---|
| count | 4445.000000 |
| mean | 1.712036 |
| std | 1.318651 |
| min | 0.000000 |
| 25% | 1.000000 |
| 50% | 1.000000 |
| 75% | 2.000000 |
| max | 5.000000 |

**Verizon**

## Distribution of Rating:

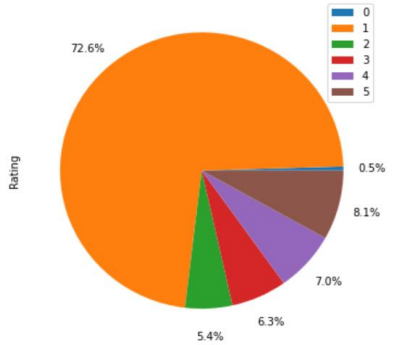
**AT&T**

```
1.0    3752
3.0     216
4.0     203
2.0     199
5.0     195
0.0      22
Name: Rating, dtype: int64
```


**Verizon**

```
1.0    3229
5.0     361
4.0     313
3.0     282
2.0     239
0.0      21
Name: Rating, dtype: int64
```

From the above charts, we get to know about the distribution of rating over the span of 6 years from 2015 to 2020. 3752 reviewers (81.8%) have rated AT&T as 1, 199 (4.3%) rated it as 2, 216 (4.7%) rated it as 3, 203 (4.4%) rated it as 4, and 195 (4.3%) rated it as 5 and 22 (0.5%) reviewers just wrote the review and did not gave any rating. 3229 reviewers (72.6%) have rated Verizon as 1, 239 (5.4%) rated it as 2, 282 (6.3%) rated it as 3, 313 (7%) rated it as 4, and 361 (8.1%) rated it as 5 and 21 (0.5%) reviewers just wrote the review and did not gave any rating. From 2015 to 2020 Verizon has received better ratings when compared to AT&T.

**Year wise Rating Distribution:**

| Date | Rating | Rating |
|---|---|---|
| 2015 | 1.0 | 996 |
| | 2.0 | 45 |
| | 5.0 | 16 |
| | 0.0 | 5 |
| | 3.0 | 4 |
| | 4.0 | 4 |
| 2016 | 1.0 | 729 |
| | 5.0 | 18 |
| | 4.0 | 15 |
| | 2.0 | 13 |
| | 3.0 | 13 |
| | 0.0 | 6 |
| 2017 | 1.0 | 611 |
| | 3.0 | 115 |
| | 4.0 | 104 |
| | 5.0 | 83 |
| | 2.0 | 71 |
| | 0.0 | 5 |
| 2018 | 1.0 | 614 |
| | 4.0 | 53 |
| | 3.0 | 50 |
| | 2.0 | 35 |
| | 5.0 | 32 |
| | 0.0 | 3 |
| 2019 | 1.0 | 593 |
| | 2.0 | 19 |
| | 5.0 | 11 |
| | 3.0 | 9 |
| | 4.0 | 4 |
| | 0.0 | 1 |
| 2020 | 1.0 | 209 |
| | 5.0 | 35 |
| | 3.0 | 25 |
| | 4.0 | 23 |
| | 2.0 | 16 |
| | 0.0 | 2 |

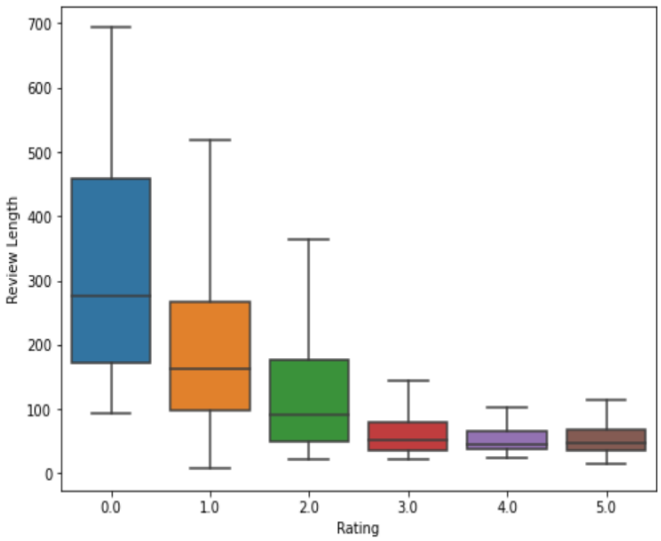
**AT&T**

| Date | Rating | Rating |
|---|---|---|
| 2015 | 1.0 | 824 |
| | 2.0 | 50 |
| | 5.0 | 15 |
| | 0.0 | 6 |
| | 3.0 | 3 |
| | 4.0 | 2 |
| 2016 | 1.0 | 552 |
| | 2.0 | 20 |
| | 4.0 | 15 |
| | 3.0 | 14 |
| | 5.0 | 12 |
| | 0.0 | 8 |
| 2017 | 1.0 | 558 |
| | 4.0 | 184 |
| | 3.0 | 156 |
| | 5.0 | 152 |
| | 2.0 | 68 |
| | 0.0 | 1 |
| 2018 | 1.0 | 456 |
| | 3.0 | 77 |
| | 5.0 | 70 |
| | 4.0 | 66 |
| | 2.0 | 47 |
| | 0.0 | 1 |
| 2019 | 1.0 | 464 |
| | 5.0 | 34 |
| | 2.0 | 25 |
| | 3.0 | 6 |
| | 4.0 | 5 |
| | 0.0 | 3 |
| 2020 | 1.0 | 375 |
| | 5.0 | 78 |
| | 4.0 | 41 |
| | 2.0 | 29 |
| | 3.0 | 26 |
| | 0.0 | 2 |

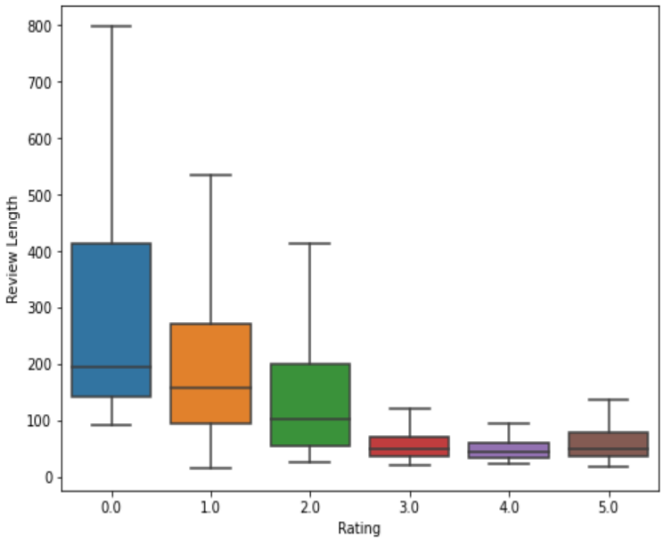
**Verizon**

From the above table, AT&T received highest number of 1 rating in the year 2015 and least in 2020. It received highest number of 5 rating in the year 2017 and least in 2019. Verizon received highest number of 1 rating in the year 2015 and least in 2020. It received highest number of 5 rating in the year 2017 and least in 2020. If we shall compare these tables, we can conclude that the Verizon has performed better than AT&T in terms of the rating received.

## Length of Reviews based on Ratings:




**AT&T**




**Verizon**

As we can see from the above graph, for AT&T the largest review is written by reviewers who did not give any ratings consisting of length from approximately 170 to 450 word count, then the reviewers who rated AT&T as 1, have written the second largest reviews length wise and the review length ranges from 100 to approximately 270 word counts. The reviewers who have given rating 4 has written the shortest reviews. For Verizon, the largest review is written by reviewers who did not give any ratings consisting of length from approximately 150 to 410 word count, then the reviewers who rated Verizon as 1, have written the second largest reviews length wise and the review length ranges from 100 to approximately 270 word counts. The reviewers who have given rating 4 has written the shortest reviews.

## Normalization, Tokenization, Collocation:

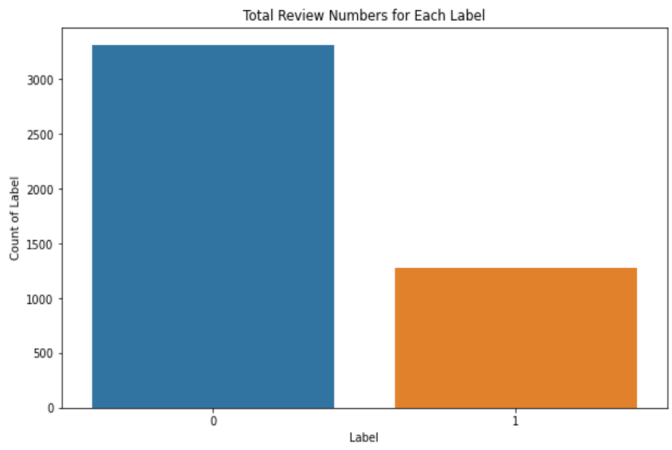
The content of review was normalized using the lemmatization technique. The tokens were created, punctuations and stop words were removed from the list of tokens and then collocations of part of speech tagging as adjective-noun and noun-noun were discovered from the content.

For AT&T the top most collocations discovered were based on several aspects. Customer service aspect included customer service, service department, service call, call center, sale rep, customer care, customer support. The negative collocations discovered were cancel service, termination fee, early termination, late fee, activation fee, worst customer, worst company, cancellation fee, horrible customer, poor service, waste time, poor customer, worst service, worst experience, service issue. There were few positive words discovered too such as good customer, gift card, customer loyalty. Different service offered related collocations were discovered too namely, internet service, data plan, billing cycle, wireless service, family plan, international call, tv internet, directv service, internet speed, installment plan, home internet.

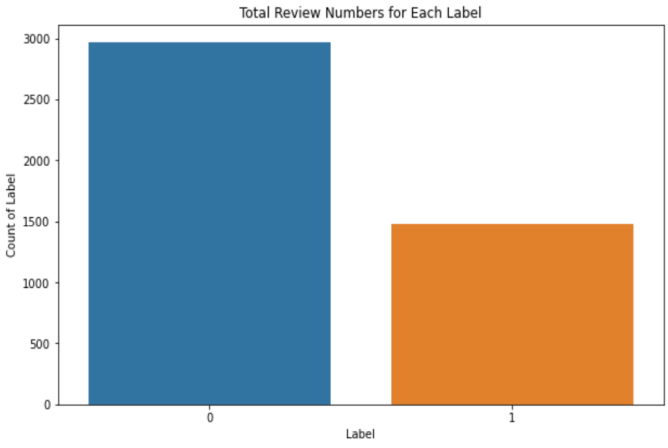
For Verizon, the top negative collocations were bad experience, termination fee, late fee, cancel service, poor customer, worst customer, fraud department, horrible customer, bad experience, poor service. The positive colocations are loyal customer, good service, good customer, great service, best coverage, great customer. There were collocations related to different service features offered by Verizon such as data plan, unlimited data, billing cycle, unlimited plan, activation fee, data usage, wireless service, internet service, family plan, payment plan, prepaid plan, auto pay. The customer support related collocations discovered were service number, service representative, call center, service department, technical support, customer service, tech support, service rep, sale rep, text message, customer support, customer care. There were certain financial aspects too discovered in Verizon's collocation such as credit card, credit report, account number, bank account, business practice, financial service, social security.

The collocations discovered captures multiple aspects of AT&T and Verizon. The negative aspect, positive aspect, customer satisfaction, service features offered can be discovered through the top hundred collocations.

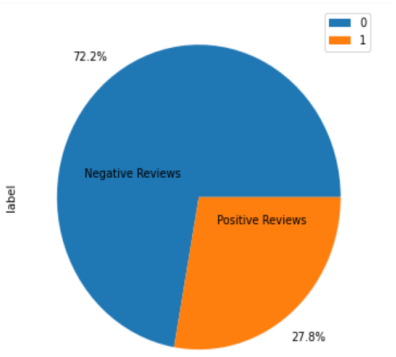
## Naïve Based Sentiment Analysis:



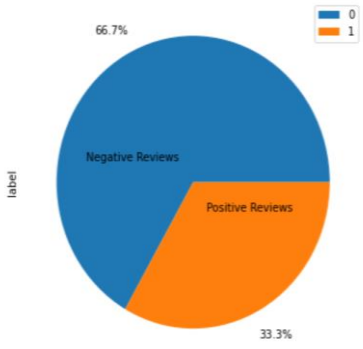

```
0    3312
1    1275
Name: label, dtype: int64
```




```
0    2965
1    1480
Name: label, dtype: int64
```




**AT&T**




**Verizon**

From the above charts, we conclude that AT&T received 3312 negative reviews and 1275 positive reviews while Verizon received 2965 negative reviews and 1480 positive reviews over a period of six years. The percentage of negative and positive reviews received by AT&T is 72.2% and 27.8% respectively while Verizon received 66.7% negative and 33.3% positive reviews, respectively.

From naïve based sentiment analysis, we can conclude that Verizon performed better than AT&T from the year 2015 to the year 2020.

**Year wise Bifurcation of Positive and Negative Reviews:**

| Date | label | label |
|---|---|---|
| 2015 | 0 | 818 |
| | 1 | 252 |
| 2016 | 0 | 607 |
| | 1 | 187 |
| 2017 | 0 | 654 |
| | 1 | 335 |
| 2018 | 0 | 555 |
| | 1 | 232 |
| 2019 | 0 | 473 |
| | 1 | 164 |
| 2020 | 0 | 205 |
| | 1 | 105 |

**AT&T**

| Date | label | label |
|---|---|---|
| 2015 | 0 | 656 |
| | 1 | 244 |
| 2016 | 0 | 443 |
| | 1 | 178 |
| 2017 | 0 | 668 |
| | 1 | 451 |
| 2018 | 0 | 465 |
| | 1 | 252 |
| 2019 | 0 | 384 |
| | 1 | 153 |
| 2020 | 0 | 349 |
| | 1 | 202 |

**Verizon**

In 2015, AT&T received highest number of negative reviews and in 2020 it received lowest number of negative reviews. In 2017, AT&T received highest number of positive reviews and in 2020 it received lowest number of positive reviews.

In 2017, Verizon received highest number of negative reviews and in 2020 it received lowest number of negative reviews. In 2017, Verizon received highest number of positive reviews and in 2019 it received lowest number of positive reviews.

**Year wise Comparison with respect to Label and Rating:**

From the below table, we get to see the year wise comparison with respect to label and rating. After comparing the information in the table, we get to know that the reviews which are positive do have lesser rating such as 1 and 2, however in most cases there is a positive correlation among “Label” and “Rating”. This further needs to be examined during the data analysis process.

| Date | label | Rating | Rating |
|---|---|---|---|
| 2015 | 0 | 1.0 | 769 |
| | | 2.0 | 30 |
| | | 5.0 | 11 |
| | | 4.0 | 4 |
| | | 0.0 | 3 |
| | | 3.0 | 1 |
| | 1 | 1.0 | 227 |
| | | 2.0 | 15 |
| | | 5.0 | 5 |
| | | 3.0 | 3 |
| | | 0.0 | 2 |
| 2016 | 0 | 1.0 | 581 |
| | | 3.0 | 7 |
| | | 2.0 | 6 |
| | | 4.0 | 6 |
| | | 5.0 | 4 |
| | | 0.0 | 3 |
| | 1 | 1.0 | 148 |
| | | 5.0 | 14 |
| | | 4.0 | 9 |
| | | 2.0 | 7 |
| | | 3.0 | 6 |
| | | 0.0 | 3 |
| 2017 | 0 | 1.0 | 467 |
| | | 3.0 | 80 |
| | | 2.0 | 49 |
| | | 4.0 | 31 |
| | | 5.0 | 25 |
| | | 0.0 | 2 |
| | 1 | 1.0 | 144 |
| | | 4.0 | 73 |
| | | 5.0 | 58 |
| | | 3.0 | 35 |
| | | 2.0 | 22 |
| | | 0.0 | 3 |
| 2018 | 0 | 1.0 | 468 |
| | | 3.0 | 33 |
| | | 2.0 | 25 |
| | | 4.0 | 15 |
| | | 5.0 | 11 |
| | | 0.0 | 3 |
| | 1 | 1.0 | 146 |
| | | 4.0 | 38 |
| | | 5.0 | 21 |
| | | 3.0 | 17 |
| | | 2.0 | 10 |
| 2019 | 0 | 1.0 | 452 |
| | | 2.0 | 15 |
| | | 3.0 | 3 |
| | | 0.0 | 1 |
| | | 4.0 | 1 |
| | | 5.0 | 1 |
| | 1 | 1.0 | 141 |
| | | 5.0 | 10 |
| | | 3.0 | 6 |
| | | 2.0 | 4 |
| | | 4.0 | 3 |
| 2020 | 0 | 1.0 | 158 |
| | | 3.0 | 17 |
| | | 2.0 | 12 |
| | | 4.0 | 9 |
| | | 5.0 | 9 |
| | 1 | 1.0 | 51 |
| | | 5.0 | 26 |
| | | 4.0 | 14 |
| | | 3.0 | 8 |
| | | 2.0 | 4 |
| | | 0.0 | 2 |

**AT&T**

| Date | label | Rating | Rating |
|---|---|---|---|
| 2015 | 0 | 1.0 | 618 |
| | | 2.0 | 29 |
| | | 0.0 | 5 |
| | | 5.0 | 3 |
| | | 3.0 | 1 |
| | 1 | 1.0 | 206 |
| | | 2.0 | 21 |
| | | 5.0 | 12 |
| | | 3.0 | 2 |
| | | 4.0 | 2 |
| | | 0.0 | 1 |
| 2016 | 0 | 1.0 | 415 |
| | | 2.0 | 12 |
| | | 3.0 | 6 |
| | | 0.0 | 5 |
| | | 5.0 | 4 |
| | | 4.0 | 1 |
| | 1 | 1.0 | 137 |
| | | 4.0 | 14 |
| | | 2.0 | 8 |
| | | 3.0 | 8 |
| | | 5.0 | 8 |
| | | 0.0 | 3 |
| 2017 | 0 | 1.0 | 418 |
| | | 3.0 | 82 |
| | | 4.0 | 73 |
| | | 5.0 | 51 |
| | | 2.0 | 43 |
| | | 0.0 | 1 |
| | 1 | 1.0 | 140 |
| | | 4.0 | 111 |
| | | 5.0 | 101 |
| | | 3.0 | 74 |
| | | 2.0 | 25 |
| 2018 | 0 | 1.0 | 354 |
| | | 3.0 | 42 |
| | | 2.0 | 32 |
| | | 4.0 | 21 |
| | | 5.0 | 15 |
| | | 0.0 | 1 |
| | 1 | 1.0 | 102 |
| | | 5.0 | 55 |
| | | 4.0 | 45 |
| | | 3.0 | 35 |
| | | 2.0 | 15 |
| 2019 | 0 | 1.0 | 351 |
| | | 2.0 | 17 |
| | | 3.0 | 5 |
| | | 5.0 | 5 |
| | | 4.0 | 4 |
| | | 0.0 | 2 |
| | 1 | 1.0 | 113 |
| | | 5.0 | 29 |
| | | 2.0 | 8 |
| | | 0.0 | 1 |
| | | 3.0 | 1 |
| | | 4.0 | 1 |
| 2020 | 0 | 1.0 | 283 |
| | | 2.0 | 23 |
| | | 5.0 | 17 |
| | | 3.0 | 14 |
| | | 4.0 | 11 |
| | | 0.0 | 1 |
| | 1 | 1.0 | 92 |
| | | 5.0 | 61 |
| | | 4.0 | 30 |
| | | 3.0 | 12 |
| | | 2.0 | 6 |
| | | 0.0 | 1 |

**Verizon**

## Feature Engineering:

Feature engineering was done using CountVectorizer technique. The CountVectorizer technique is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire content of the review.


**AT&T**


**Verizon**

From the above WordCloud chart, we get to know the positive features of AT&T and Verizon. The positive features of AT&T are love company, service helpful, good experience, easy use, better prices, excellent service, great deals and the positive features of Verizon are always reliable, extra mile, pleased service, always friendly, helpful call, best one, reliable network, always works, great plan, great experience, signal everywhere.

## PROBLEM STATEMENT:

The use of the internet has increased to a great extent especially during the time of COVID-19 when classes are conducted online and work from home culture has increased tremendously. Hence, companies need to understand what features and services are important for the customers. This will lead to improved service offerings, better customer service, and retention of old customers as well as attracting new customers. As per Statista U.S. 2020 research records, AT&T and Verizon are the top two companies with the majority of the market share. Hence, we decided to analyze the reviews of these customers to understand what benefits are offered by these companies and what are the drawbacks of them. Once we understand this, we shall be able to highlight the competitive advantage these companies have over others. Also, the research shall enable the telecom companies to understand where they are lacking and therefore are unable to grab the top spot. Through our research, we shall provide a set of attributes that makes a company stand out from others and highlight key issues faced by the customers that need to be worked upon.

**Hypothesis:**

After conducting the exploratory data analysis, we came up with the below hypothesis:

H1 (Alternative Hypothesis): Verizon provides better overall service than AT&T.

H0 (Null Hypothesis): AT&T provides better overall service than Verizon.

We came up with the above hypothesis because, after comparing the results of the past six years from 2015 to 2020, Verizon has received higher positive reviews when compared to AT&T. Also, the ratings received by Verizon is higher than AT&T. Assuming that better customer service and product features lead to better reviews and ratings thus providing better overall service.

## METHODOLOGY:

The reviews with rating 1 and 2 are considered as negative reviews and reviews with the rating 4 and 5 are considered as positive reviews. The reviews with the rating 3 are manually labeled.

We shall be conducting supervised sentiment analysis using Support Vector Machine and Naïve Bayes learning models. A classification report shall help us in performance evaluation of the models. A confusion matrix shall be created to see how many reviews are correctly labeled.

## PROJECT PLAN & STEPS OF COMPLETION:

Datasets are collected through web scraping. Then preprocessing and exploratory data analysis has been done. In next stage, machine learning algorithms shall be applied, and performance evaluation of the models shall be done.

| Particulars | Team Member |
|---|---|
| Data Set Collected | Equal contribution by each Team Member |
| Data Preprocessing | |
| Exploratory Data Analysis | |
| Project Report | |

## PLANNING FOR NEXT STAGE:

| Particulars | Team Member |
|---|---|
| Methodology | Each Team Member shall contribute equally |
| Performance Evaluation | |
| Analysis of Experiment Results | |
| Conclusion and Future Work | |
| Term Paper | |
| Presentation | |