Pattern Study of Signal Strength and Data Speed in India

Final Report (CS F415 – Data Mining) Group 39

Nakul Gupta
2017A7PS0211H

Sundeep Ammisetti
2017A7PS1218H

Pranav Sista
2017A7PS1225H

Satish Kumpata
2017AAPS0284H

***Abstract*—The objective of this project is to analyze data provided to us and to seek out meaningful results. The data sets our group chose to work on were related to mobile data speeds in various parts of India. We used the data sets available on "The Open Government Data (OGD) Platform India" and chose data sets from different years and different months. A number of processing techniques and data analysis techniques were used on the data to make the data sets more coherent and to extract any useful information. The goal of this analysis is to get results which will be useful to customers for deciding what network providers provide the best service wherever they live and for network providers to identify where their services are lacking so that they can work on improving it. In total, three pre-processing techniques were used and two data analysis techniques were used.**

Keywords—Outliers, Analysis, Data Speed, Signal Strength

I. Introduction

The Open Government Data Platform India (data.gov.in) was a platform started by the government under the Open Data initiative. Under this initiative, various data sets were uploaded to the web and made available for the public to use. The files were uploaded as CSV files and one such set of CSV files were related to mobile data speeds. Using these data sets, we have begun to draw out whatever information we can from the data sets and put it to good use.

To analyse the data, the data had to go through different phases – cleaning, pre-processing, clustering and outlier analysis. During pre-processing, the data was cleaned, reduced and transformed. The purpose of this processing was to –

- Determine how fast different network providers were and to get an idea as to where different network providers were better and where they weren't.
- Observe the differences in performance between 3G and 4G and during the download or upload of data.
- Observe how data speeds fluctuated at different values of signal strength and whether there were a considerable number of outliers.
- Observe the performance of different network providers and states across different years and different months.

To also help with the above-mentioned tasks, various visualization plots, primarily bar graphs, will also be used to show these differences.

II. Nature of the Data

The data sets had six features - "Network Provider", "3G/4G", "Upload/Download", "Data Speed", "Signal Strength" and "state". We used data sets from the years 2018 and 2019 and compared data between the two across three months - March, May, and July. We chose these three months because they were the only common months between the two years which had data. Each month had at least two lakh entries. All six datasets got sampled and cleaned and were merged into one final set, which had two extra features - month and year. It was on this data set that we began our data analysis and drawing different visualization plots.

The features present in the data sets were both categorical and continuous. There were six network providers - Airtel, Vodafone, Idea, Jio, Cellone, Uninor, Aircel and Dolphin. The signal strengths and speeds were recorded for two separate scenarios - during download and upload. Furthermore, the network being provided was either 3G or 4G. The data was collected across 23 states/regions in India.

The attributes which were continuous were "Signal Strength" and "Data Speed". Data speed was recorded in kilobytes per second (kbps) and signal strength in decibels(dB). On observing these two attributes at face value, the magnitude of the data speed decreased as the signal strength decreased (that is,. It was also observed that the signal strength values lied between –50dB to -120dB. This is a noteworthy observation since any values greater than -120dB are considered dead signals.

Other than the features, there were other aspects of the data which were worth noting. Most of the data sets had missing values. In cases where the state was missing, the records were discarded since we wouldn't be able to use it in different comparisons. There were cases where the signal strength value was missing. There was also an uneven distribution in records with respect to the network providers. Most of the records came under JIO while barely any were recorded for more obscure providers such as Cellone, Uninor, Aircel and Dolphin. This gave an idea as to which network provider was used more as compared to the others.

III. Pre-Processing Techniques Used

A total of four pre-processing techniques were used, including cleaning. The data was read and stored in data frames with the help of Pandas. The techniques used were -

A. *Data Reduction*

All the data sets were sampled using stratified data sampling. This sampling technique was used since we had many sub-groups to work with. We took every combination of the attribute values and selected 5% of the data from each sub-group. There were cases where the values obtained from sampling were far below those compared to other sub-groups. For example, the obscure network providers mentioned previously had only a total of 3-10 entries total in the final data set. This was in contrast with providers like Jio which had close to 72,000 final entries in the data set, making them mostly useless in future comparisons. However, it did give an idea as to how rarely these providers are used.

B. *Data Cleaning*

The process of cleaning the data had two parts - dropping all records which had an N/A value for “state” and replacing N/A values for signal strengths in certain records. The records without a state were dropped since we had no way of figuring out the source of the data. To resolve the N/A values in the “Signal Strength” column, we decided to replace it with the mean value of the signal strength using a combination of unique values of the features. A decent number of records were removed for this reason but not many as compared to the total number of data entries.

This part of pre-processing the data took a very long time to run. In some of the bigger data sets (twenty lakh entries), there were instances where it took more than an hour to run the data and having to run it on five other data sets was more time consuming.

C. *Data Transformation*

Once all the data was sampled and cleaned, the entire data set was normalized. For this, both min-max normalization and z-score normalization were used separately. Z-score normalization was used to help aid in finding the outliers since the algorithm used for outlier analysis required the data to be normalized through this method. Otherwise, min-max normalization was used in for the different clustering methods. For min-max normalization, the values were normalized into the range 0 – 100 for both features.

Aggregation was performed as well. We chose to merge the all records which had Idea as a network provider into Vodafone. This was done since Idea was merged into Vodafone in 2018 and we thought that it would be better to consider them as one company. One interesting observation made was that despite merging two network providers with a moderate number of records, Vodafone still had a fewer number of records as compared to Jio.

Two additional features were included into the data set – “Month” and “Year”. This was when all the data sets were being merged with one another and so that future comparisons could be made easier and kept track off. These were all the techniques used in pre-processing the data.

IV. Conclusions Drawn From Pre-Processing the Data and Visualization

Visualization techniques were used to draw comparisons between how the data speeds and signal strengths changed across different months and years. The data was visualized using Seaborn. We used a violin plot to also get an idea of what range most of the values were present in and univariate distributions between signal strength, data speed and year, observing how the former two varied across the years. Listed below are some notable results.

A. *Variations Across Different Months*

The following graphs were plotted to compare how the continuous variable varied depending on what kind of technology was used and whether the data was being uploaded or downloaded. Notable differences and similarities were observed.

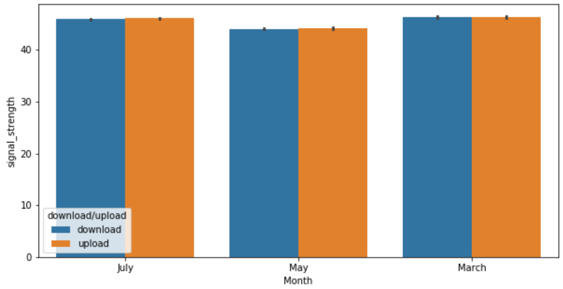


Fig.1. Month vs Signal Strength (Hue = download/upload)

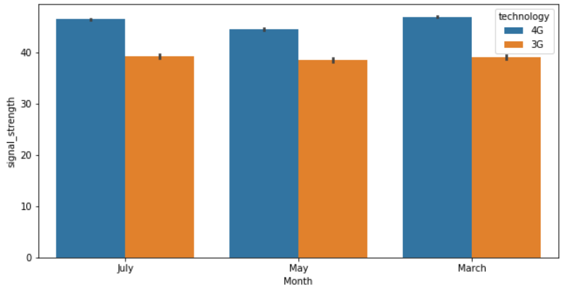


Fig.2. Month vs Signal Strength (Hue = 4G or 3G)

It was observed across all months (both years combined) that the signal strength values were similar for a given technology or whether it was during the upload or download of data. During the upload/download of data, the signal strength is more or less the same (Fig. 1) while there is a slight difference for different technologies (Fig. 2).

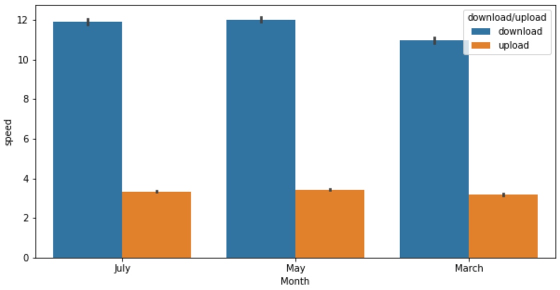


Fig.3. Month vs Speed (hue = sownload/upload)

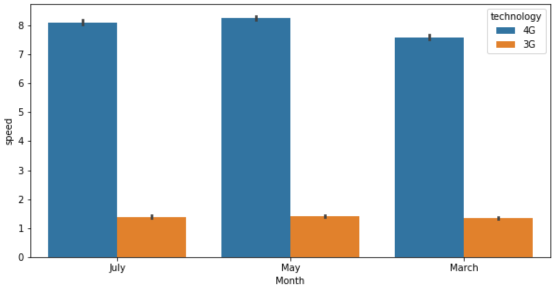


Fig.4. Month vs Speed (hue = 4G or 3G)

These results were also reflected when comparing the data speeds across the months. The values were similar across all months for a given hue but between hues there was a very high difference. Download speeds were far greater than upload speeds (Fig. 3) and data speeds were far greater for the 4G technology (Fig. 4).

B. Variations Across Different States

We compared how the signal strengths and data speeds varied across different states of India. We compared the values across different months and merged all the months within a year to get the values. The graphs are given below. These graphs were plotted using values normalized with min-max normalization. Keep in mind that if the signal strength lowers then the mobile data speed will increase, that is the lower the signal strength the better.

We found that in a little more than half of the states the signal strengths had become better. There was a notable decrease in signal strength for Delhi. Bihar had by far the worst change in signal strength among all the states, having the largest positive increase. Among other states, the signal strengths improved in five states, remained mostly stagnant in six states and the quality reduced in the remaining 12 states. (Fig. 5)

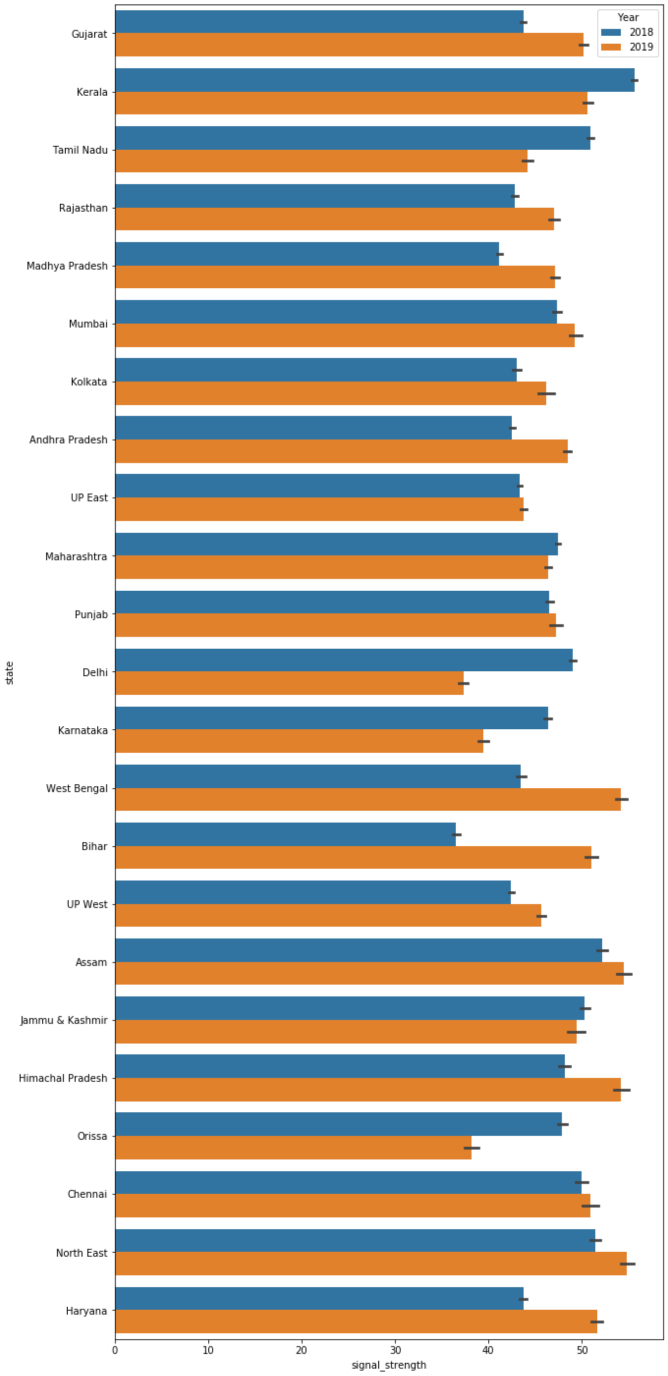


Fig.5. Signal Strength vs State (hue = Year)

The results of how the data speeds varied across different states were reflective of how the signal strengths changed. Wherever the signal strength dropped, the data speed increased and vice versa. As a result, Delhi saw the greatest increase in data speed among all states and Bihar saw the greatest decrease in data speed across all states (Fig. 6).

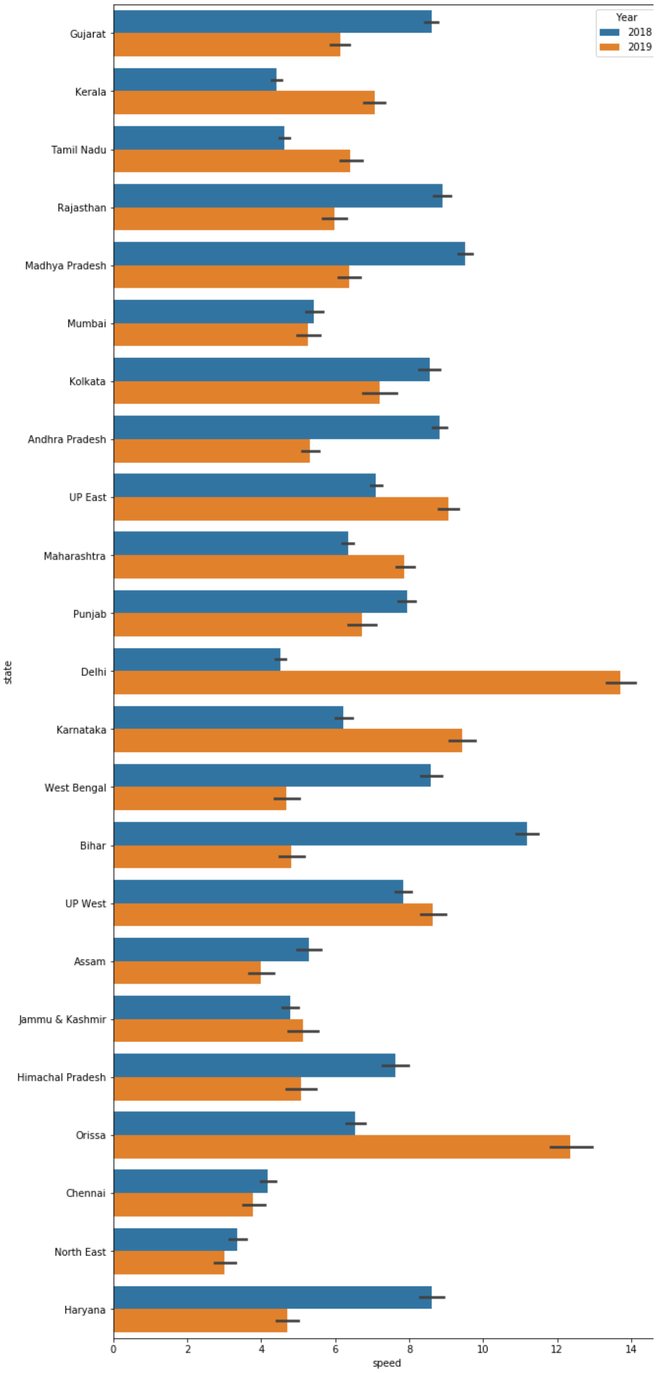


Fig.6. Speed vs State (hue = year)

V. Data Analaysis Techniques

Once the data was pre-processed, the next step was to analyze the data. In this section, the various data mining techniques used will be described and the results they yielded will also be discussed. In total, two unique data analysis techniques were applied – clustering and outlier analysis. Within clustering, we performed two different clustering techniques – DBSCAN clustering and fuzzy C-means clustering. As for outlier analysis, we only used one algorithm which we will go in depth into later.

A. *Clustering*

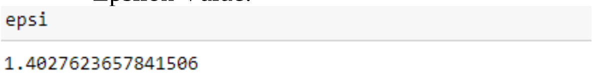
Clustering was done with the intention of classifying the data into two separate clusters. This would be done by comparing the speed and signal strength of a record while keeping the other attributes fixed. The goal of this exercise was to get an idea of how many entries clustered into having either 4G or 3G technologies. An important point to note is that when we ran the clustering algorithm, we did not remove the outliers. This is because the outliers could be shown as outliers in the output graph and because they didn't affect the clusters much. Initially, we ran the DBSCAN algorithm for figuring this out.

1) DBSCAN Clustering:- Prior to performing this, we first checked the clustering tendency of our data by checking the Hopkin's statistic. The Hopkin's statistic is a measure which returns a value between 0 to 1, indicating whether the given data points can be clustered or not. If the value lies close to one, it indicates that the data can be highly clustered, it it's close to 0.5 it indicates that the data is random and if it is close to 0, it will indicate that the data is uniformlt distributed. Using this statistic, we decided to choose a sub group of data accordingly. In case the reader want to experiment, they can change the given values for the different attributes in the code accordingly.

Once the subset of data was chosen, the code for clustering was run. Attached below are some of the results we got.

Example: The following attribute values and variables were considered and computed –

- State – Delhi
- Month – March
- Year – 2018
- Download/Upload – Upload
- Service Provider – Airtel
- Hopkin's Statistic: 0.7531
- Epsilon Value:



```
epsi
```

```
1.4027623657841506
```

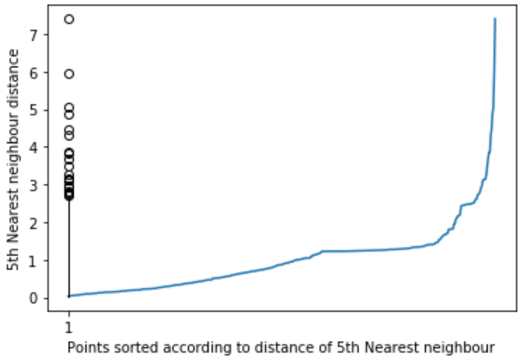


Fig.7. Graph to find the value of EPS

Graphs Obtained: -

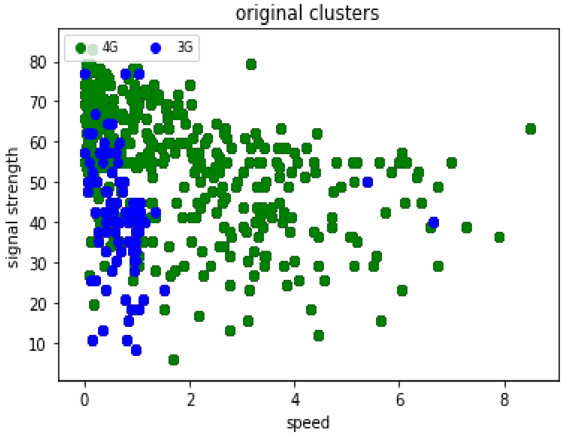


Fig.8. Plot of all the points of the sub-group before clustering

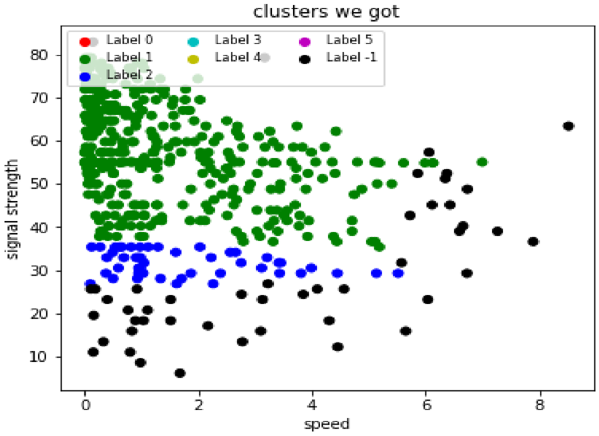


Fig.9. Plot of all the points of the sub-group after running clustering

In the above diagram, the black points are outliers (label -1), the green points are points which belong to the 4G cluster (label-1) and the blue points are the points which belong to the 3G cluster (label-2) in Fig. 9. Unfortunately, we were unable to implement cluster validation for the generated graphs so the only comparisons one can make is by comparing the graphs and observing how the accurate the clusters were. In the above case, we can see that most of the 3G entries got classified as entries having 4G technology and similar results were observed in other graphs. (Fig. 8)

2) Fuzzy C-means Clustering: Aside from DBSCAN clustering, we used fuzzy C-means clustering. This was done due to the observations in DBSCAN clustering, with many 3G points being classified as 4G points. This algorithm was run with the hope that it would give better results than the DBSCAN algorithm. This algorithm was run taking the number of clusters as 2 (the two clusters representing 3G and 4G points). However, upon running the algorithm for different sub groups of data, all the data entries ended up being shared by both of the clusters equally, that is, having a membership value of 0.5 As a result, we couldn’t make much progress with this technique. The membership matrix was the final result of the algorithm.

B. *Outlier Analysis*

This was the second data analysis technique we used to identify what outliers were present in the data. We considered this technique necessary since there were many random spikes in the data and for a single strength value, there was a range of data speed values for different records (the values ranged from small values to high values).

The outlier analysis algorithm uses the final combined data set from preprocessing before normalization. We apply z score normalization to its speed and signal strength values so we can have a better understanding of how they vary across all the states, service providers, months, etc. After applying z score normalization, we can find outliers by checking if any values are higher than the threshold of z score equal to 3 or less than z score equal to -3. These thresholds are the generally accepted ones to find outliers. After checking for outliers, we found that the only outliers which existed were the ones having the z score values of speed greater than 3. From this we can say that there are no outliers in the signal strength values which means they are properly distributed across all states and service providers. By running the algorithm, we were able to obtain two separate CSV files – one comprising of only outliers and input data set without outliers. Given below is a visual representation of how the outliers were distributed.

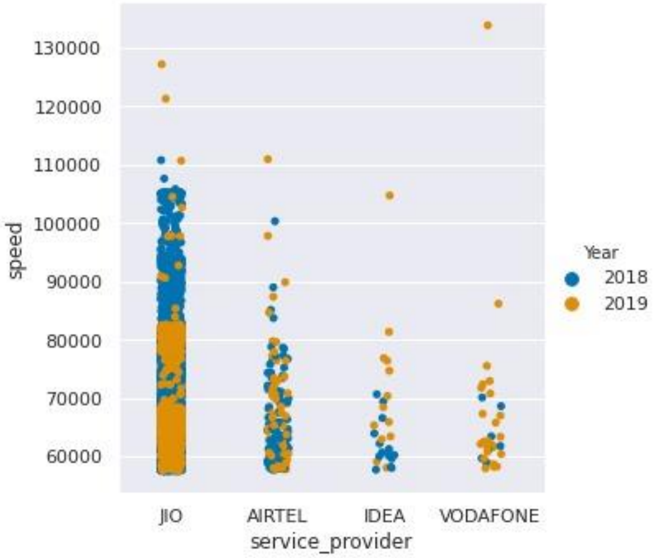


Fig.10. A graph of outliers for each network provider

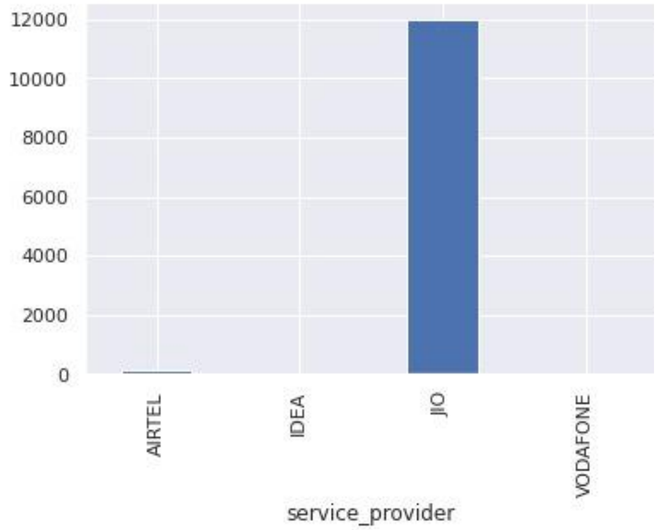


Fig.11. The frequency of outliers for each network provider

A total of 12,189 outliers were obtained. For speed values, the outliers exist only for high extreme values not for low values. We've noticed that approximately 98% of the outliers are from Jio for all the respective months and 2 years that we've sampled from. The remaining 2% are from Airtel, Vodafone, and Idea but they're all negligible in comparison to Jio's outliers. Despite Jio having the most outliers, they only comprise of approximately 6% of its total frequency in the entire dataset. In conclusion, despite the outliers in total being a small fraction of the total dataset, we can say that choosing Jio will at least give the highest chance of possibly getting extremely high speeds in comparison to the other service providers across India.

VI. Conclusions Of Data Analysis

Given the nature of our data, we were unable to apply as many data analysis techniques as we had hoped. We attempted association rule mining but there was no point in doing it since most of the data was already classified. We attempted to use linear regression as well but data ended up being too scattered to construct a linear model. The only techniques we could use were clustering and outlier analysis and the latter gave us useful results.

We couldn't get much out of clustering since we had not successfully implemented cluster validation. However, we were able to conclude that there isn't much of a difference in values between data records in 3G and 4G clusters. This was supported by the fact that membership of the records in most cases was equal in both clusters. Not to mention, many data points which were in the 3G cluster got assigned to the 4G cluster during analysis. It could be that these points were entries which had an unusually high value of data speed for a given signal strength. Despite removing outliers, we were unable to notice any significant changes in the clustering so we left those points as they were.

As for outlier analysis, we received very good results. We observed that most of the points were points under Jio. However, we couldn't automatically discard the points since it isn't natural to have approximately 12,000 outliers from a single network provider. Given Jio's general popularity, it was clear that Jio in general had higher data speeds than other providers. This graph went on to show that Jio provides the best internet speeds all across India.

Overall, from the results we can conclude that without a doubt, Jio is the best network provider in the country.

Acknowledgment

We thank Dr. Manik for giving us the opportunity to learn and apply our knowledge to real world applications. We also thank BITS Pilani Hyderabad Campus for giving us the opportunity to take this course. Lastly, we thank the government of India in providing us the datasets for others to analyse and research on.

References

[1] Open Government Data (OGD) Platform India - https://data.gov.in/

[2] Pandas Documentation - https://pandas.pydata.org/docs/

[3] Seaborn Documentation - https://seaborn.pydata.org/.

[4] Introduction to Data Mining, Pang-Ning, Tang and others, Pearson Education