INSAID

INTERNATIONAL SCHOOL OF AI AND DATA SCIENCE

thecapstoneproject

CONSULTING ASSIGNEMENT REPORT
DATA ANALYSIS

PREPARED BY:
TEAM1003

TABLE OF CONTENTS

INTRODUCTION

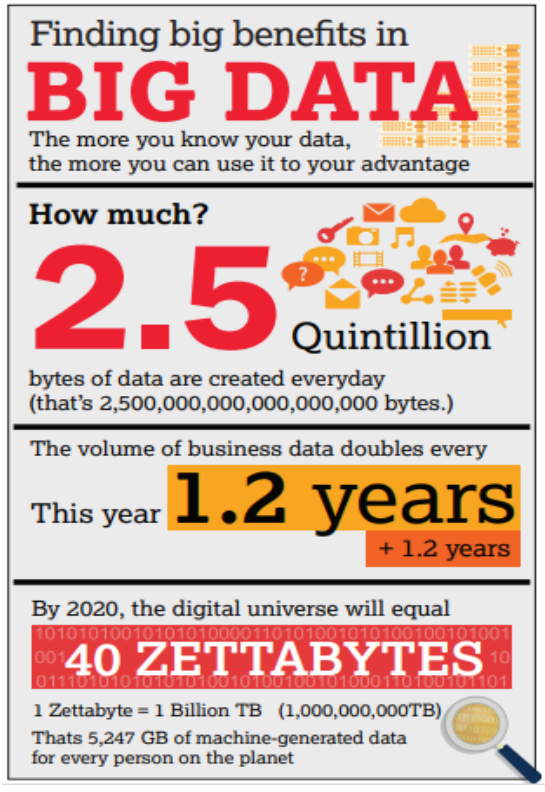
The telecommunications sector has become one of the main industries in developing countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.

Three main strategies have been proposed to generate more revenues:

- Acquire new customers
- Upsell the existing customers
- Increase the retention period of customers.

However, comparing these strategies taking the value of return on investment of each into account has shown that the third strategy is the most profitable strategy. Retaining an existing customer costs much lower than acquiring a new one.

With the increasing adoption of smart phones and growth in mobile internet, Telecoms today have access to exceptional amounts of data sources including –customer profiles, device data, network data, customer usage patterns, location data, apps downloaded, etc. All this data combined together becomes the Big Data. Big data has the potential to place the telecoms in a position to win the battle to earn more customers and create new revenue streams. It provides them with a wealth of information about their customer behaviours, preferences and movements. Yet, many telecoms still struggle to fully derive the greatest value of big data.



PROJECT DESCRIPTION

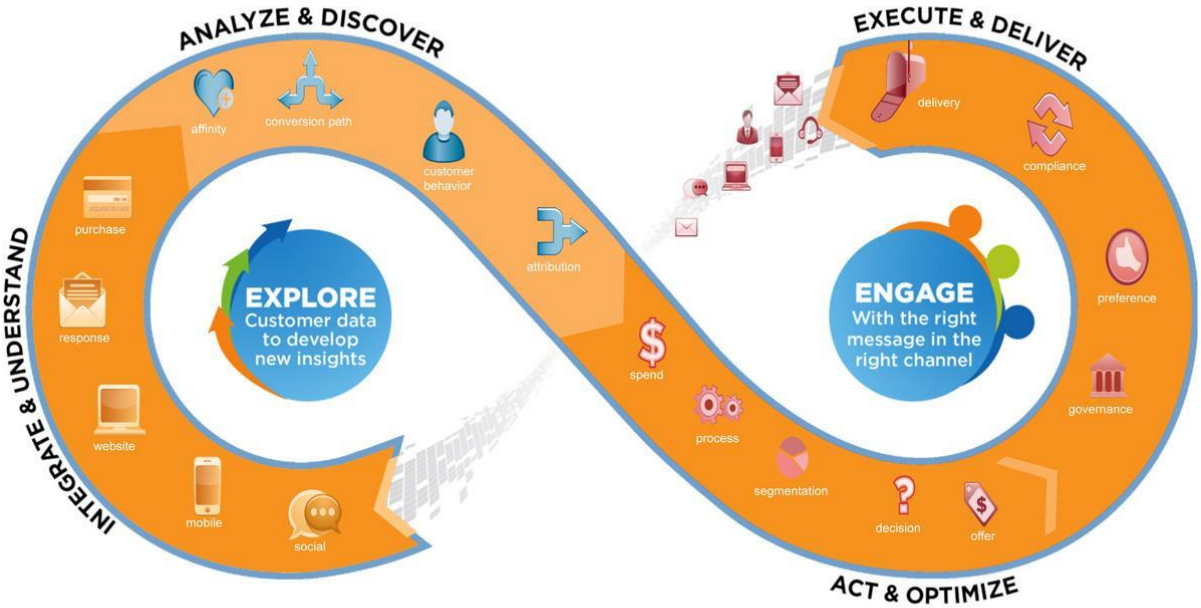
Insaid Telecom, one of the leading telecom players, understands that customizing offerings is very important for its business to stay competitive.

In this Project, Insaid Telecom is seeking to leverage behavioural data from more than 60% of the 50 million mobile devices, active daily in India, to help better understand and interact with their audience.

This **consulting assignment** aims to use the data to build a dashboard to understand users' demographic characteristics based on their age, gender, geolocation, mobile usage and mobile device properties.

Doing so can potentially help millions of developers and brand advertisers around the world to pursue data-driven marketing.

This project aims to help Insaid Telecom understand the right way forward with actionable insights for Marketing and Product teams to pursue data-driven marketing.



PROBLEM STATEMENT

The purpose of this report is to analyse the demographic data of telecom users in 6 states in India viz, **Tamil Nadu, Manipur, Chandigarh, Tripura, Uttar Pradesh, Arunachal Pradesh** , and find out their usage pattern in terms of the phones that they use and in particular which model they are using and what is the frequency of their usage by state, age and gender. By analysing this, we wish to find out meaningful business outcomes that can be utilised by Insaid Telecom to further develop products that are appropriate for market usage and would build upon the current pattern of consumer usage.

PROBLEM ANALYSIS

To Analyse the problem, the team followed few strategic steps:

1. The data has to be looked at closely. See what information can be obtained from each data set. Understand the data. Make a connection with the data.
2. Find the challenges with the data.
    a. Find missing data and find ways to repopulate the missing data.
    b. Find incorrect data and correct them.
    c. Find ways to join the data sets
3. Less is more. Reduce, consolidate, net out the data. Drop the irrelevant data.
4. Question the dataset. Brainstorm and find insights.
5. Identify relationships between variables that are particularly interesting or unexpected.
6. Using effective visualizations, communicate results and findings

These are some of the questions asked about the data:

For the States: <u>Tamil Nadu, Manipur, Chandigarh, Tripura, Uttar Pradesh, Arunachal Pradesh</u>

1. Where we have the most frequency of events /users?

2. What time of the week and day are the consumers most active on their phones?
3. What is the demographic (age and gender) distribution of these events, mobile brands and mobile phones?
4. Which brand and models of phone is being utilised the most based on age and gender of users?
5. Analysis of geographical density of users

SOURCES OF DATA

In this assignment, INSAID has provided demographic data of users (gender and age) for 6 states (**Tamil Nadu, Manipur, Chandigarh, Tripura, Uttar Pradesh, Arunachal Pradesh).**

The following three data sets were provided to perform exploratory data analysis and to ascertain patterns if any:

- Event data set: Comprising of consumers location coordinates as well as timestamp of usage
- Gender data set: Comprising of gender and age of the consumers
- Phone Brand-Model: Comprising of phone brands and models used by consumers

S. No	Event data set	Description
1	Event ID:	Represents an event whenever consumer accesses telecom network
2	Device ID	Represents the mobile device
3	Timestamp	Represents the date and time.
4	Longitude	Represents geographical coordinate
5	Latitude	Represents geographical coordinate
6	City	Represents the city where consumer is located
7	State	Represents the state where consumer is located.

S. No	Brand-Model data set	Description
1	Device ID	Represents the mobile device
2	Phone Brand	Represents the mobile brand that a consumer uses
3	Model	Represents the model of the mobile a consumer uses

S. No	Gender Age data set	Description
1	Device ID	Represents the mobile device
2	Gender	Represents whether the consumer is a male or a female
3	Age	Represents the age of the consumer
4	Age Group	Represents the age group of the consumer

SUMMARY OF DATA MINING

Data Processing:

❑ Events dataset

- The event dataset contained **3,252,950 records** and **7 columns**
- Some values for Latitude and longitude were incorrect
- There were incorrect values for City-State pairs
- Following values were missing :
    - State: 377
    - Longitude : 423
    - Latitude : 423
    - Device ID : 453

To resolve the above issues following steps were taken:

1. Filling in Missing Values:
    a. State: For state, we used the city data and found the same city elsewhere in the dataset. The corresponding values for State were then used to fill in the missing values.
    b. Latitude and Longitude: Missing values for latitude and longitude were filled by finding the same Device ID in the dataset. Corresponding latitude and longitude values from the rows with the same Device ID were used to fill the missing values using forward fill.
    c. Device ID: The missing values for Device IDs were filled using the latitude and longitude values and finding the same latitude and longitude values elsewhere in the dataset. Corresponding Device IDs were then used to fill the missing Device ID values using forward fill.

2. Since we had incorrect City-State pairs as well as incorrect latitude and longitude values, we then asked if we should use coordinates to find the correct city/state or use city/state to find correct

coordinates? So, we assumed that the coordinates are correct and calculated the correct city and state using reverse Geocoder. We then filtered out the 6 states and created a new dataframe.

❑ Age_Gender Dataset:

1. This was a relatively clean dataset. All DeviceIDs were unique and there were no missing values.
2. The data was slightly right skewed and there were some outliers as observed using a box plot.
3. In Age Groups, male age groups and female age groups were grouped together, hence there was a need to make these uniform by creating new bins. We decided to solve this during the EDA (please see EDA point no. 5C)

❑ Phone Brand Dataset:

1. Both Brand and model had Chinese names – These were translated in English before merging the dataset.

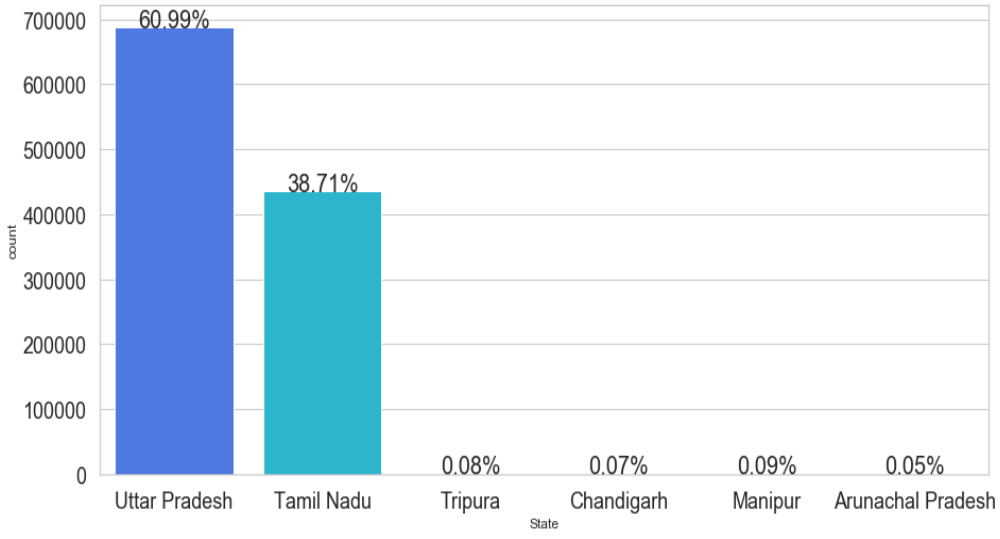
After having processed the three datasets, the datasets were joined using 'INNER JOIN'. Device ID was used as common (primary-foreign) key for merging three data sets. After merging, total data set had 1,127,005 rows (Events) on which we used for our EDA.

EDA

NOTE: The data set is imbalanced with 99% of users belonging to only two states – Uttar Pradesh and Tamil Nadu. The data for the rest of the 4 states, Manipur, Tripura, Chandigarh and Arunachal Pradesh has data only for ~30 users each. This imbalance in data distribution can make the conclusions drawn in this report biased.

1. **Distribution of Events:**

```
State
Uttar Pradesh        687387
Tamil Nadu           436290
Manipur                 989
Tripura                 892
Chandigarh              839
Arunachal Pradesh       608
Name: EventID, dtype: int64
```

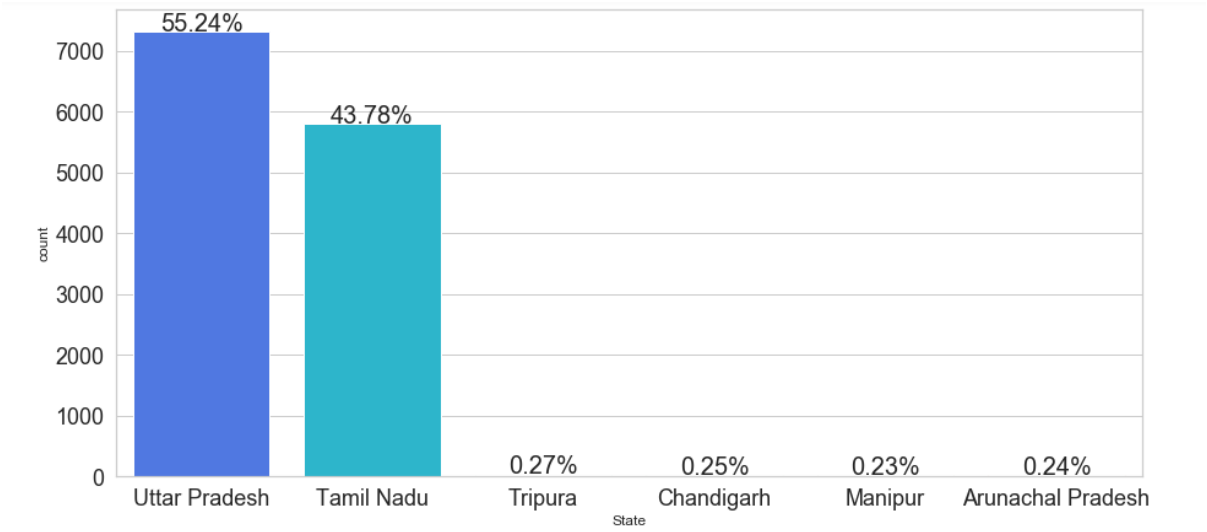


Conclusion: The dataset is imbalanced since 99.7% of data is from only two states – Uttar Pradesh and Tamil Nadu. Data for rest of the four states is less than 1%.

2. **Distribution of Users:**

Since there are multiple events recorded from same Device ID, we narrowed down the data to unique device IDs to get unique User data for all states. The unique dataset had 13,256 events

```
EventID          13256 non-null int64
DeviceID         13256 non-null int64
Timestamp        13256 non-null datetime64[ns]
Longitude        13256 non-null float64
Latitude         13256 non-null float64
```



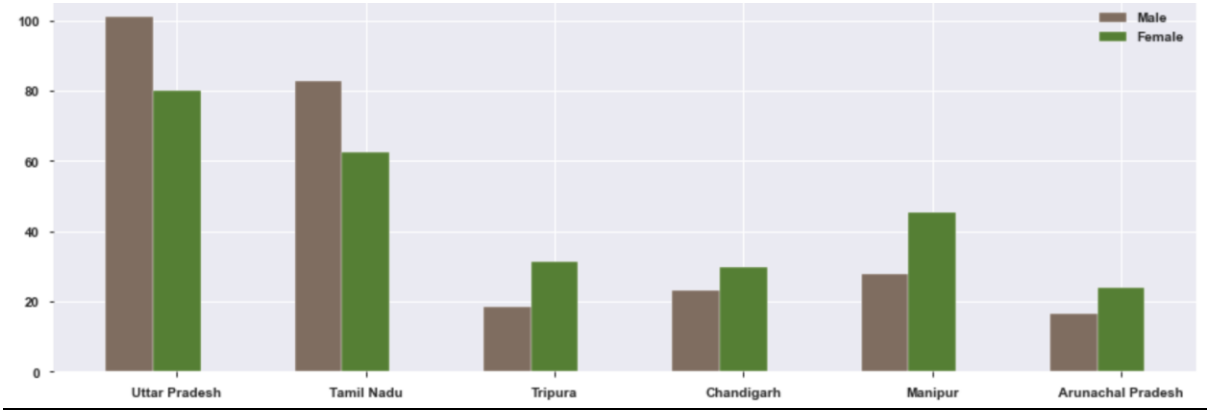

```
State
Uttar Pradesh        7322
Tamil Nadu           5803
Tripura                36
Chandigarh             33
Arunachal Pradesh      32
Manipur                30
Name: DeviceID, dtype: int64
```

Conclusion:

Similar to the distribution of events, distribution of users is also imbalanced with 99% of users belonging to only two states – Uttar Pradesh and Tamil Nadu.

3. Analysis of Events/User per state:

55% of Uttar Pradesh users are responsible for 60% of the events generated. So it is evident that UP users are much more active compared to other state users.

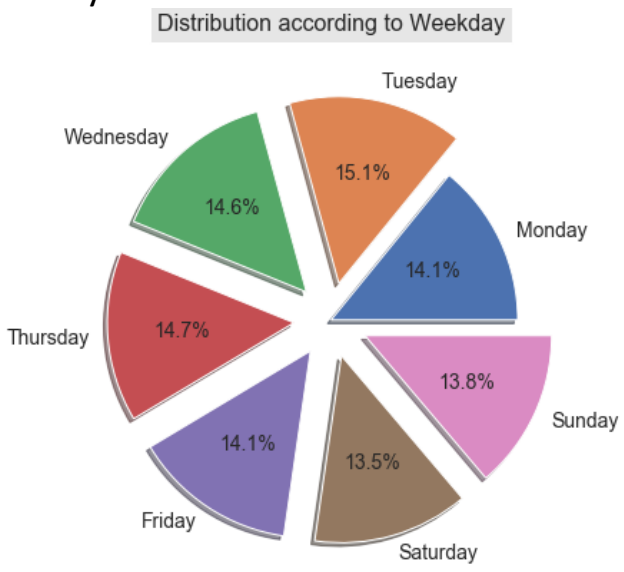


Conclusion: By analysing the data, it looks like both male and female UP users use the mobile the most followed by TN and Manipur. In Uttar Pradesh and Tamil Nadu, males use the mobile more than female. But in Tripura, Chandigarh, Manipur and Arunachal Pradesh it is females who are more active.

4. Analysis of activity during the weekdays:

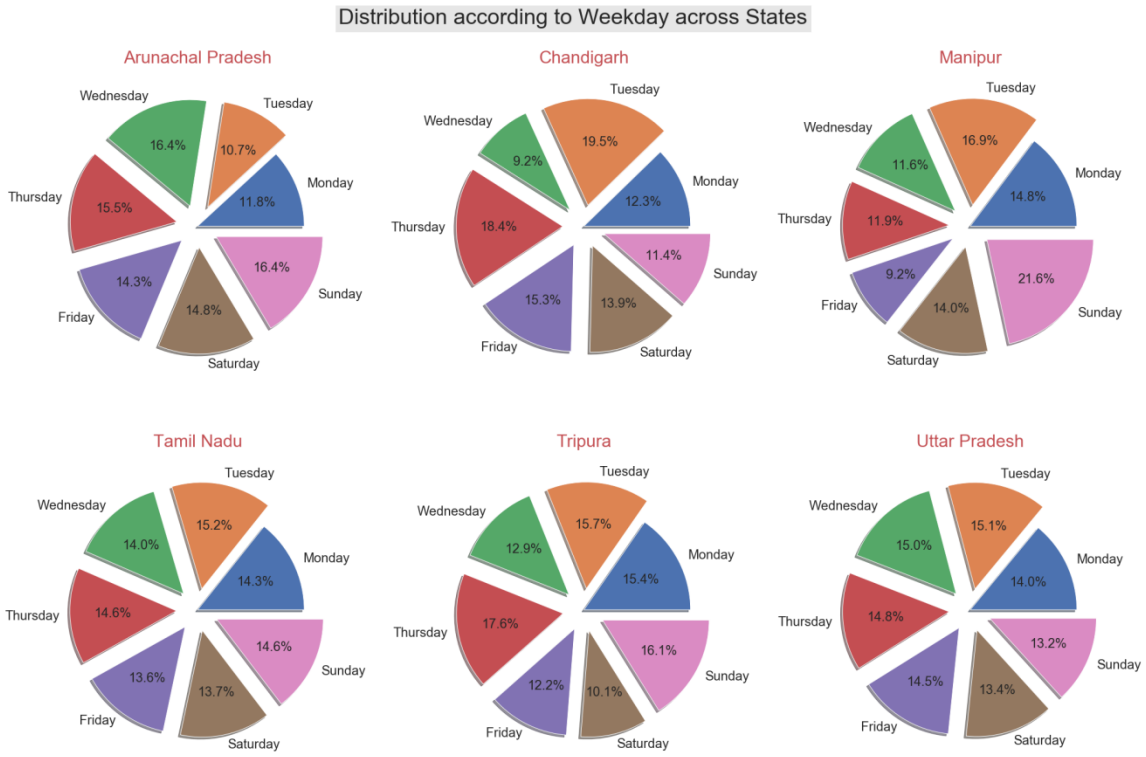
A. Cumulative Analysis of activity during the week for all 6 states:

To analyse the activity during the week, we have used the timestamp data and extracted the activity based on Weekday as well as time of the Day.



Conclusion: Although users used the network almost the same on all days of the week, but Tuesday seems to have the most activity (0.5 to 1 % higher than other days).

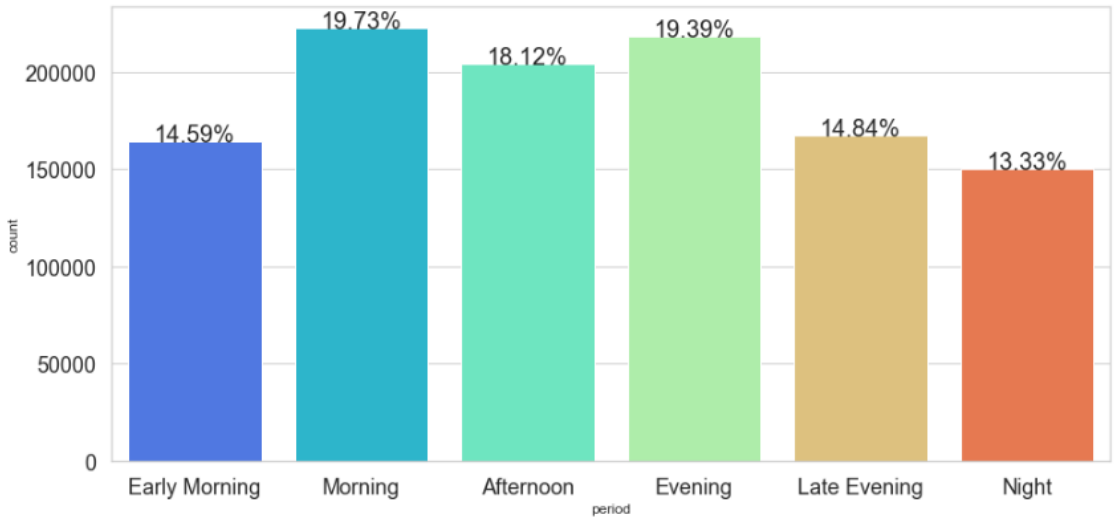
B. **Analysis of activity during weekdays across each State:**



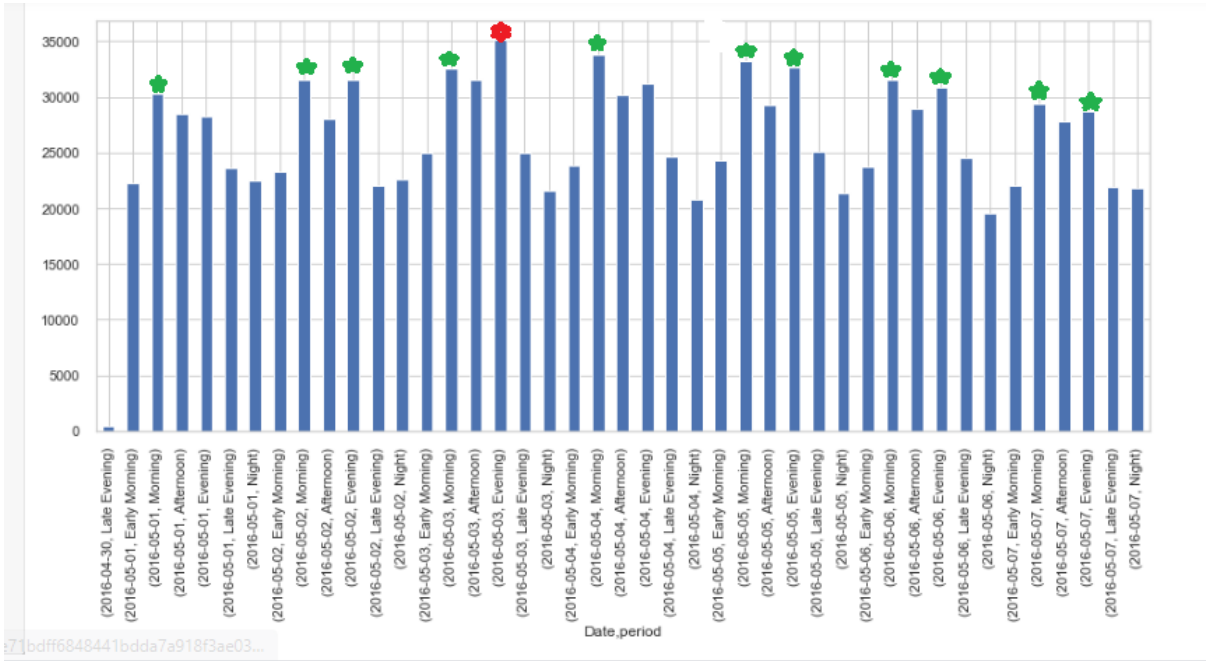
Conclusion: For Tamil Nadu, Tuesday seems to have a 1% higher activity than other weekdays, while Tuesday and Wednesday had the highest activity at 15%, Thursday and Friday were only 0.5% less.

5. Analysis of activity during the day:

A. Cumulative analysis for all states:



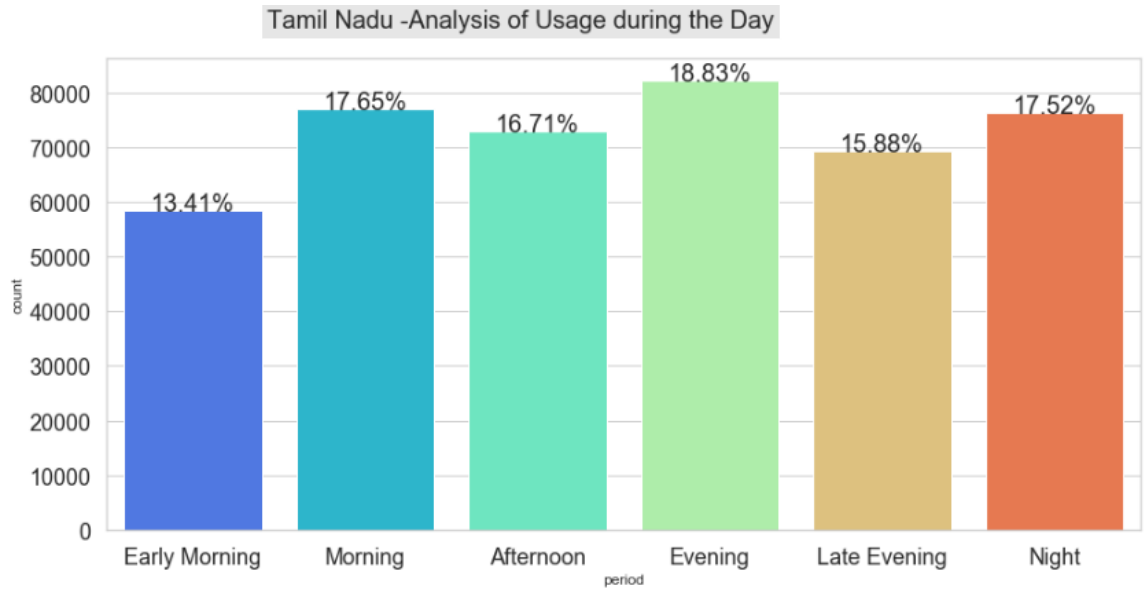
Conclusion: Mornings (4AM-8AM) and Evenings (4PM-8PM) are the peak hours for network, closely followed by afternoons.

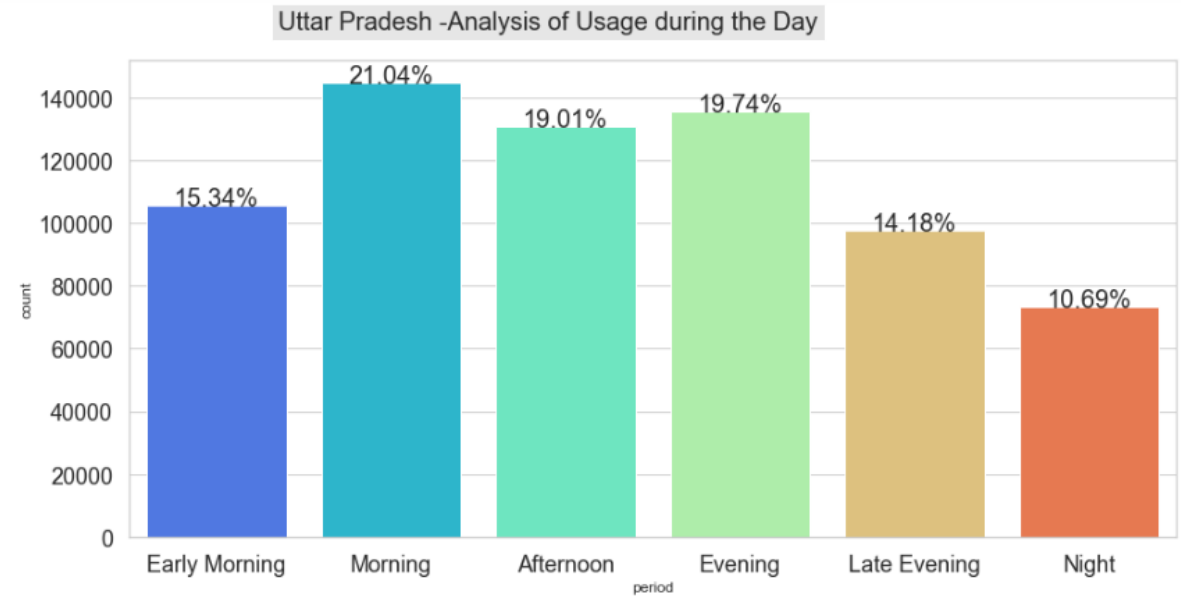


Conclusion: For the dataset, which spans across 30th April to 7th May 2016, the activity on the network peaked during the evening (4PM to 8PM) of 3rd of May 2016, which was a Tuesday (**Red star**). For other

days in the week, the activity peaked either during mornings (4-8AM) or evenings (4-8PM) or both (**Green stars**).

B. Analysis for activity during the day for Tamil Nadu and Uttar Pradesh states

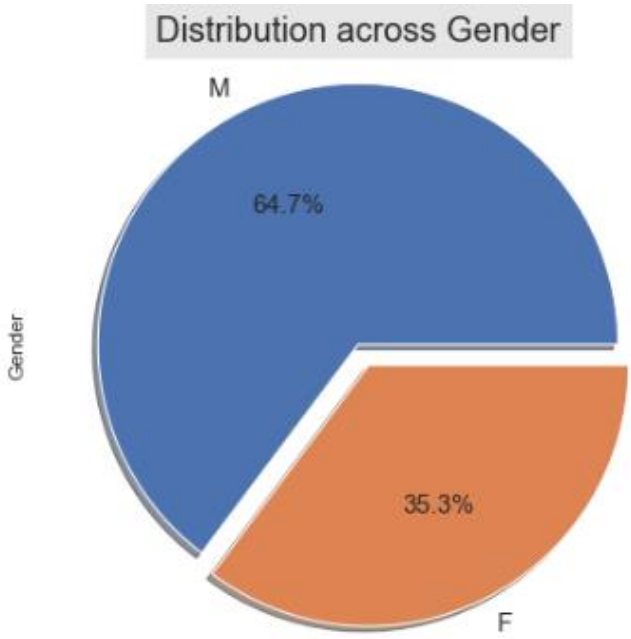




Conclusion: Mornings (4AM-8AM) and Evenings (4PM-8PM) are the peak hours for activity for both Tamil Nadu and Uttar Pradesh, followed by afternoons. Data for other states is too small to draw any conclusions.

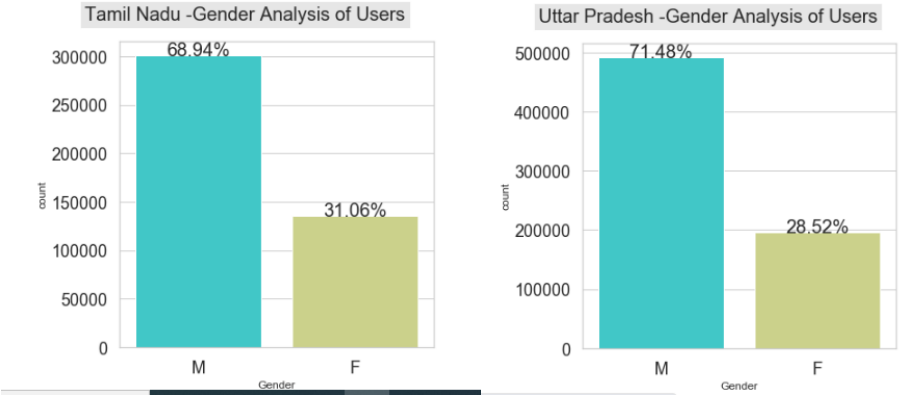
6. Demographic Analysis of Users across States

A. Cumulative Analysis of Gender



Conclusion: Across the states, there are almost double Male users than Female users in the dataset. While there are 64.7% Male users, only 35.3% are Female users.

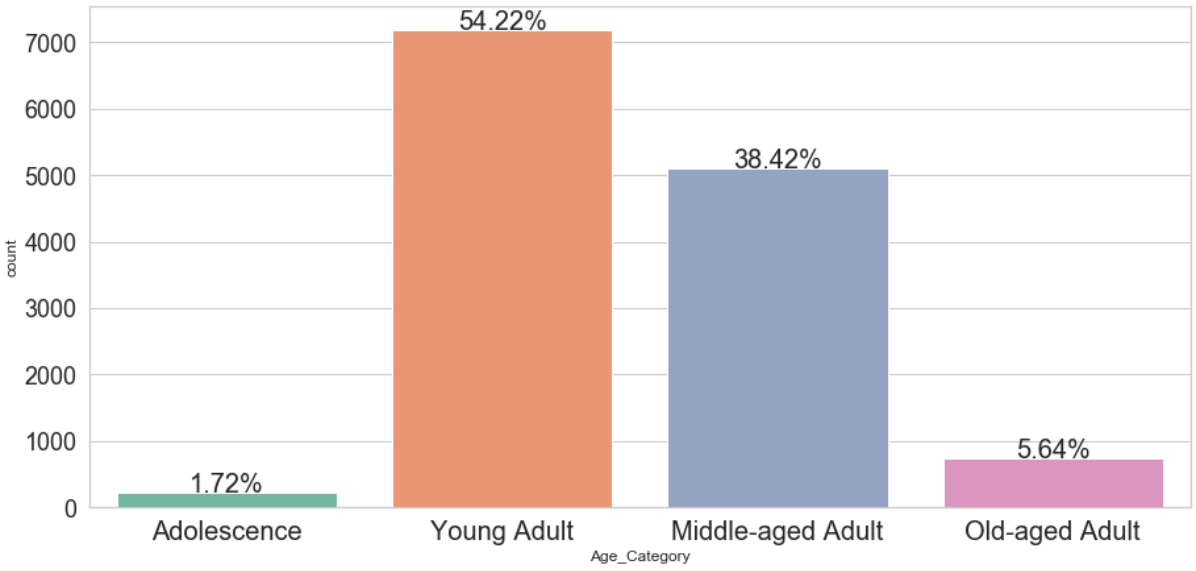
B. Gender analysis for Users in each state



Conclusion: In Tamil Nadu and Uttar Pradesh, Male users are the dominant users. Data for other states is too small to draw any conclusions.

C. **<u>Cumulative analysis of Age Group for Users</u> :** New Groups were created to analyze age distribution of users. The groups created are as below:

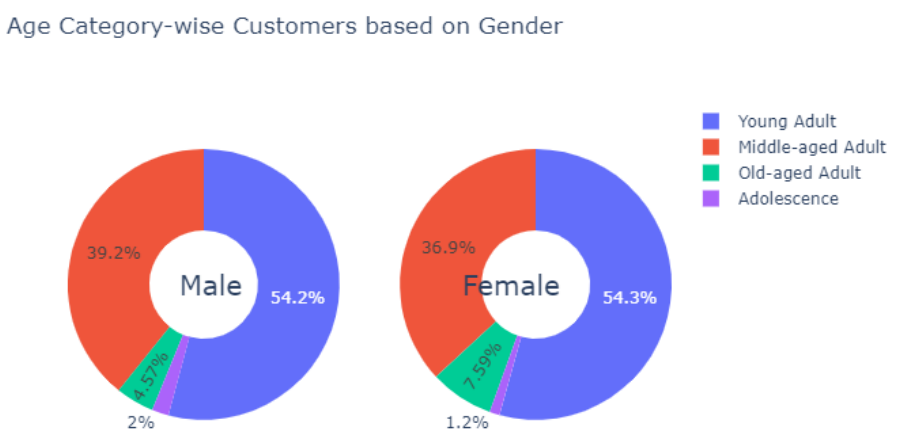
Age	Age Category
6 - 18	Adolescence
19 - 30	Yound Adult
31 - 50	Middle-aged Adult
51 - 95	Old-aged Adult



Conclusions:

Young & Middle-aged Adults constitute most of the customers.

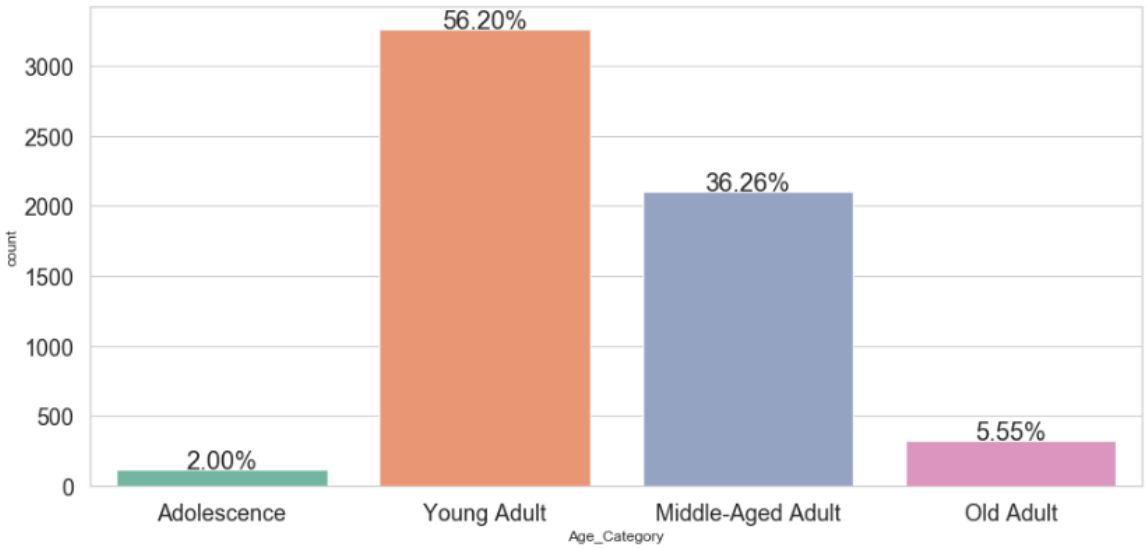
D. Cumulative analysis of Age Group for Male and Female Users:



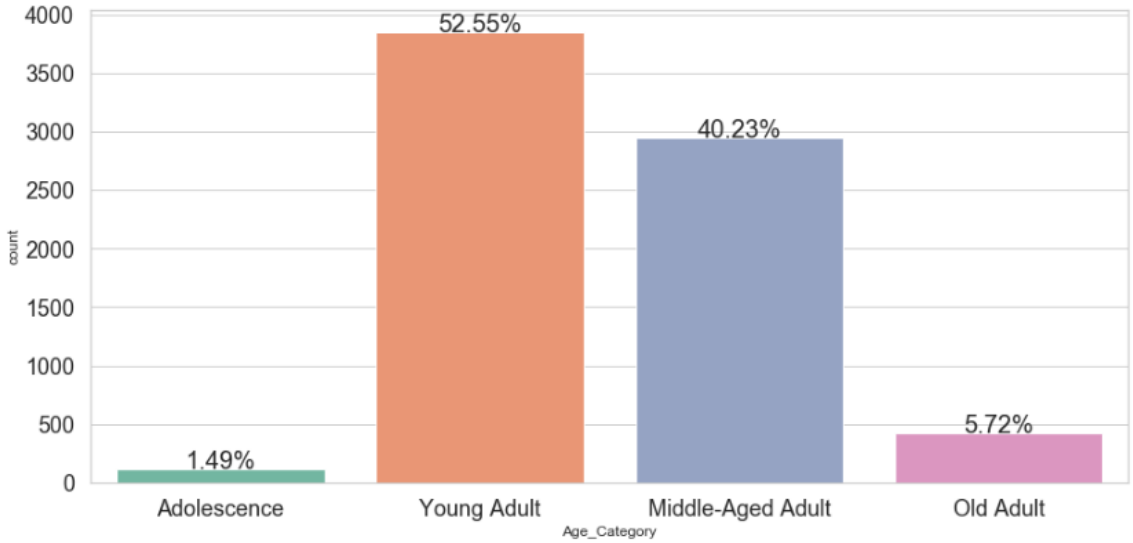
Conclusions: Young & Middle-aged Adults constitute most of the customers for both the genders.

E. Age Distribution of Users across Tamil Nadu and Uttar Pradesh:

Tamil Nadu – Age Distribution of Users



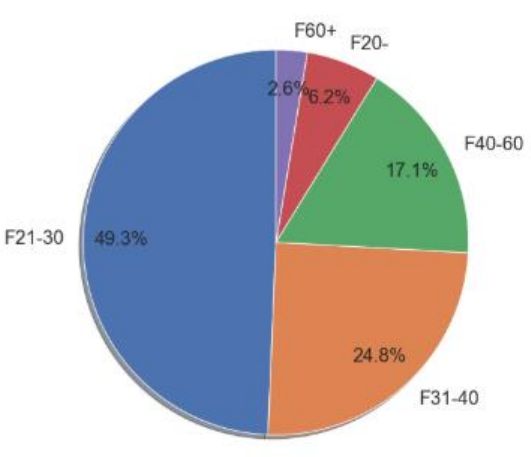
Uttar Pradesh – Age Distribution of Users



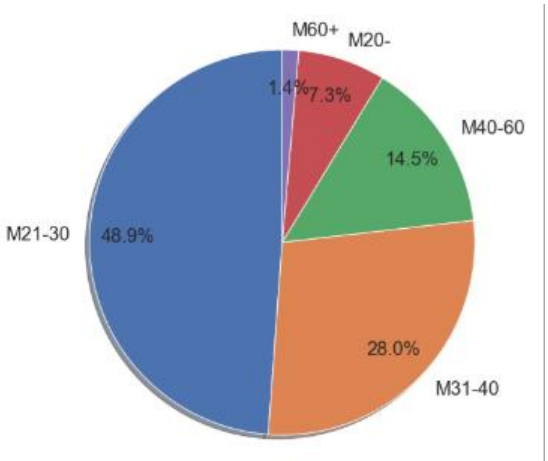
Conclusions: Young & Middle-aged Adult categories constitute most of the customers in Tamil Nadu as well as Uttar Pradesh. Data for other states is too small to draw any meaningful conclusions.

F. Age Distribution of Users across Gender:

New Groups were created for both Males and Females : 20-, 21-30, 31-40, 40-60,60+ for age group plus Gender analysis of users in this data set.



Proportion of each age group type(Female)

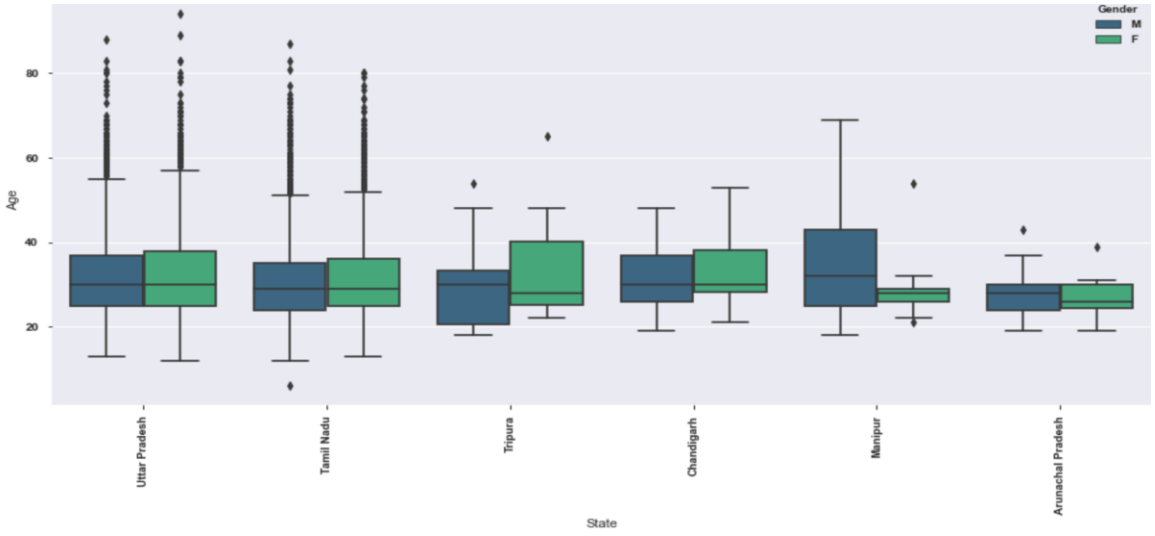


Proportion of each age group type(Male)

Conclusions:

- It is very evident that almost 50% of the users, both in Male and Female category are between the age group 21-30 years.
- Next biggest category is between (1/4th of the users in this dataset) are from 31-40 years.
- Users aged above 40 years and under-20 years constitute the remaining customers.
- The telecom company needs to focus on age groups, 31-40 and 40-60 to expand their business.

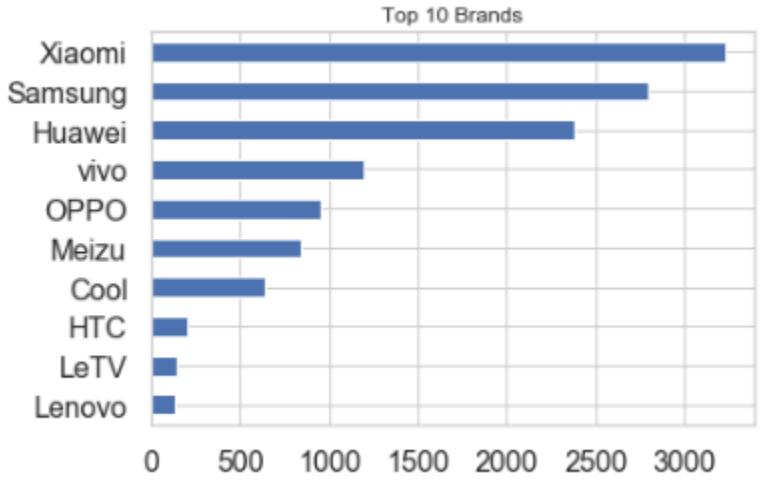
G. Age and Gender distribution across the states:



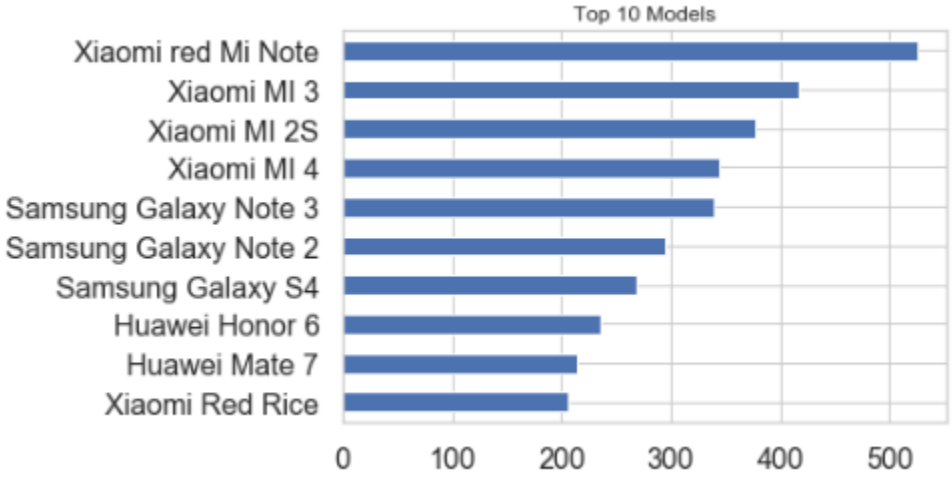
Conclusion: Most of the users are between 20 and 55. Uttar Pradesh, Tamil Nadu and Manipur have users older than 60 years. When Comparing Gender vs Age, Uttar Pradesh and Tamil Nadu have almost similar distribution of Male/Female users. Whereas Female User's Age of Manipur and Arunachal Pradesh are lesser than the Male User's Age. Also, Boys start using the phone at a much younger age in the states of Tripura, Chandigarh and Manipur.

7. Brand and Model Analysis across all Users

A. Top 10 Brands across all users

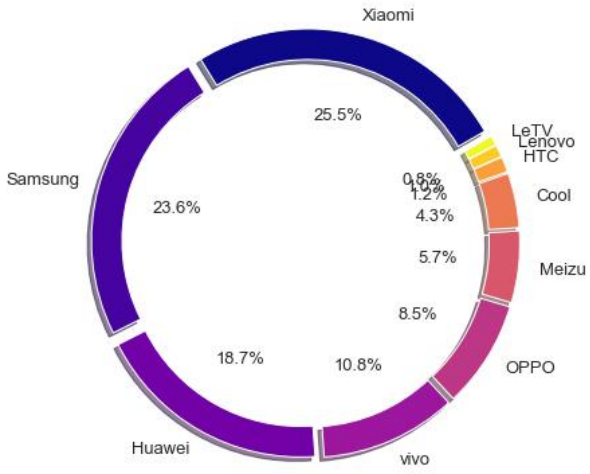


B. Top 10 Models across all users

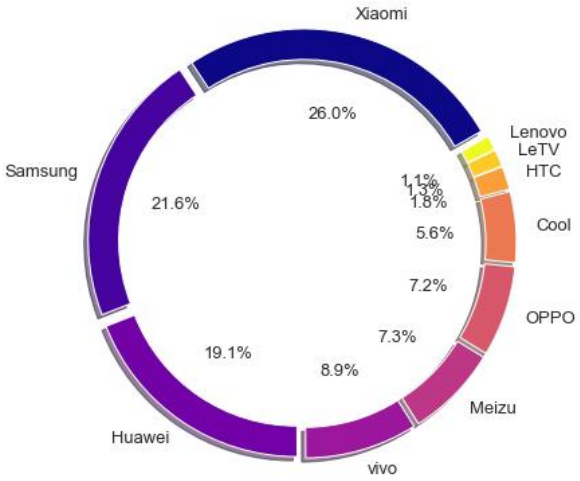


Conclusion: Xiaomi, Samsung and Huawei are the top 3 brands of mobiles used by the users in this dataset. Top 10 models used also belong to these 3 Brands.

C. Top 10 Brands amongst Female and Male Users

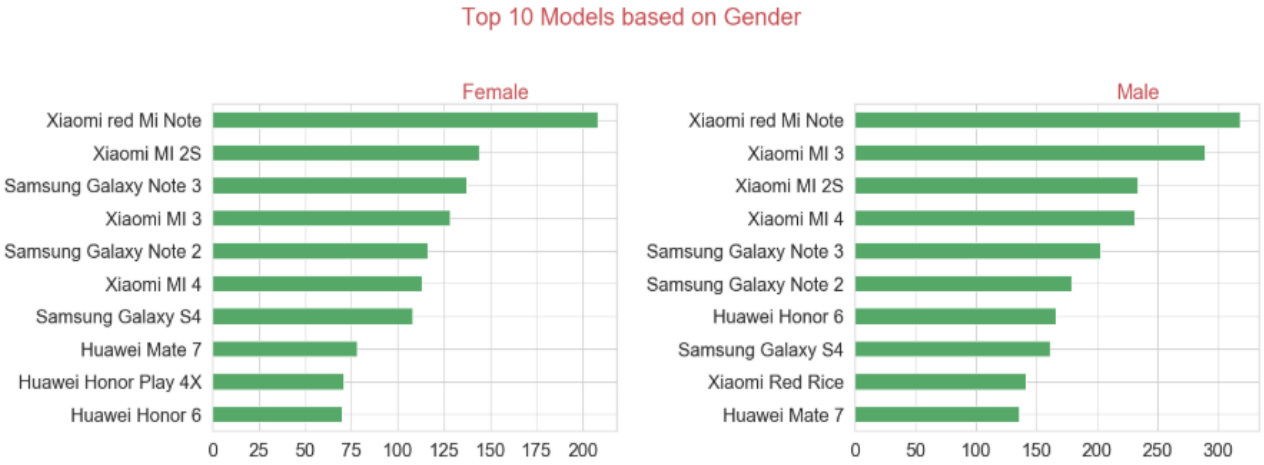


Top 10 Brands used by Female

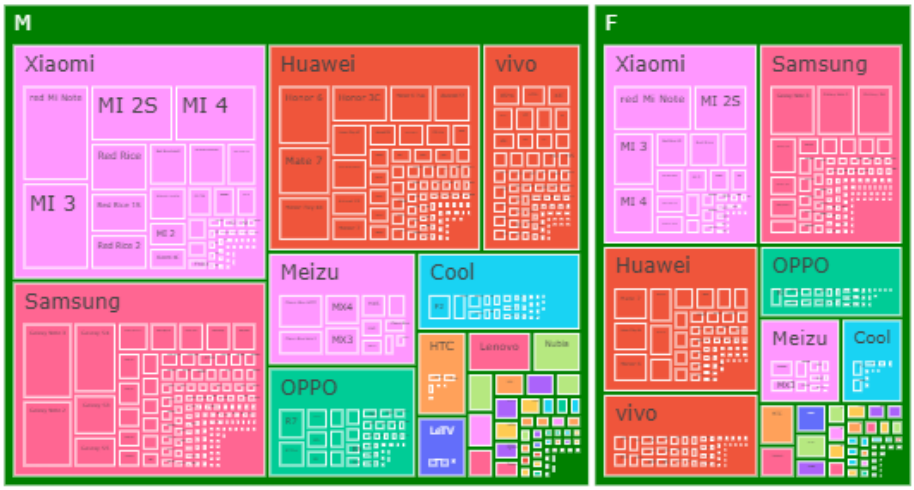


Top 10 Brands used by Male

D. Top Mobile Models amongst Female and Male Users

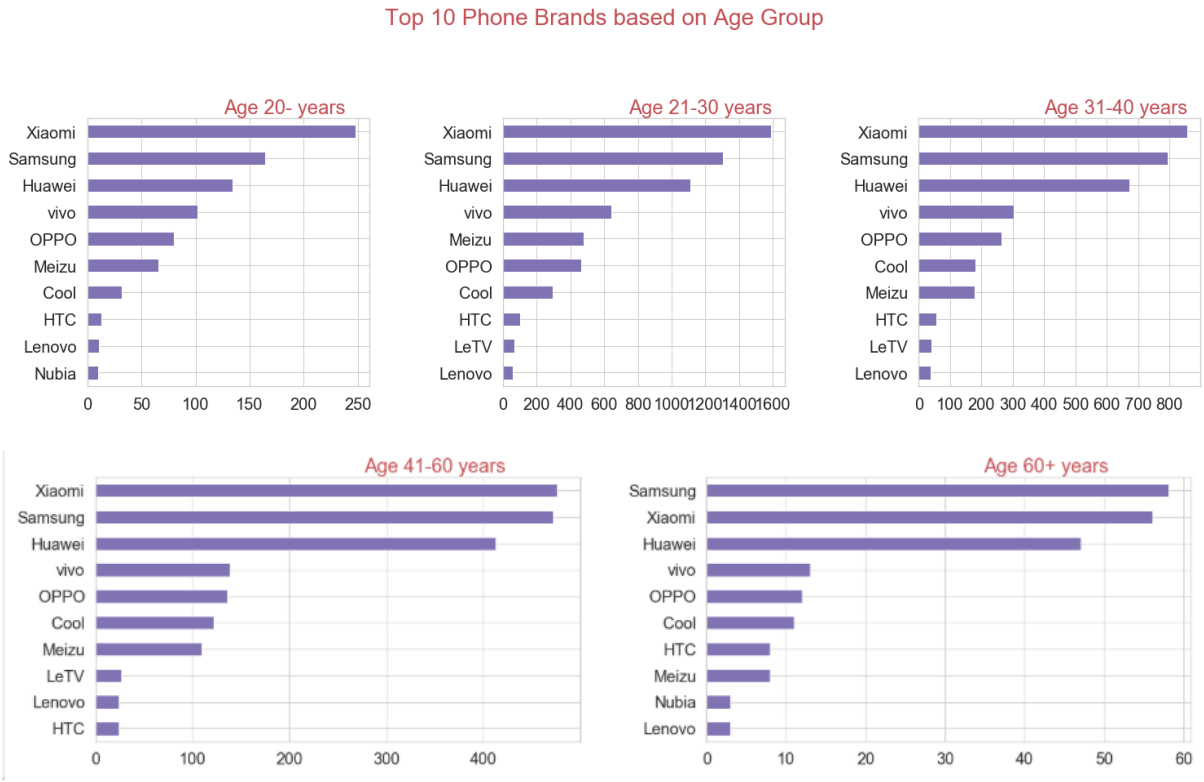


TreeMap showing customers with all brands & their models based on gender



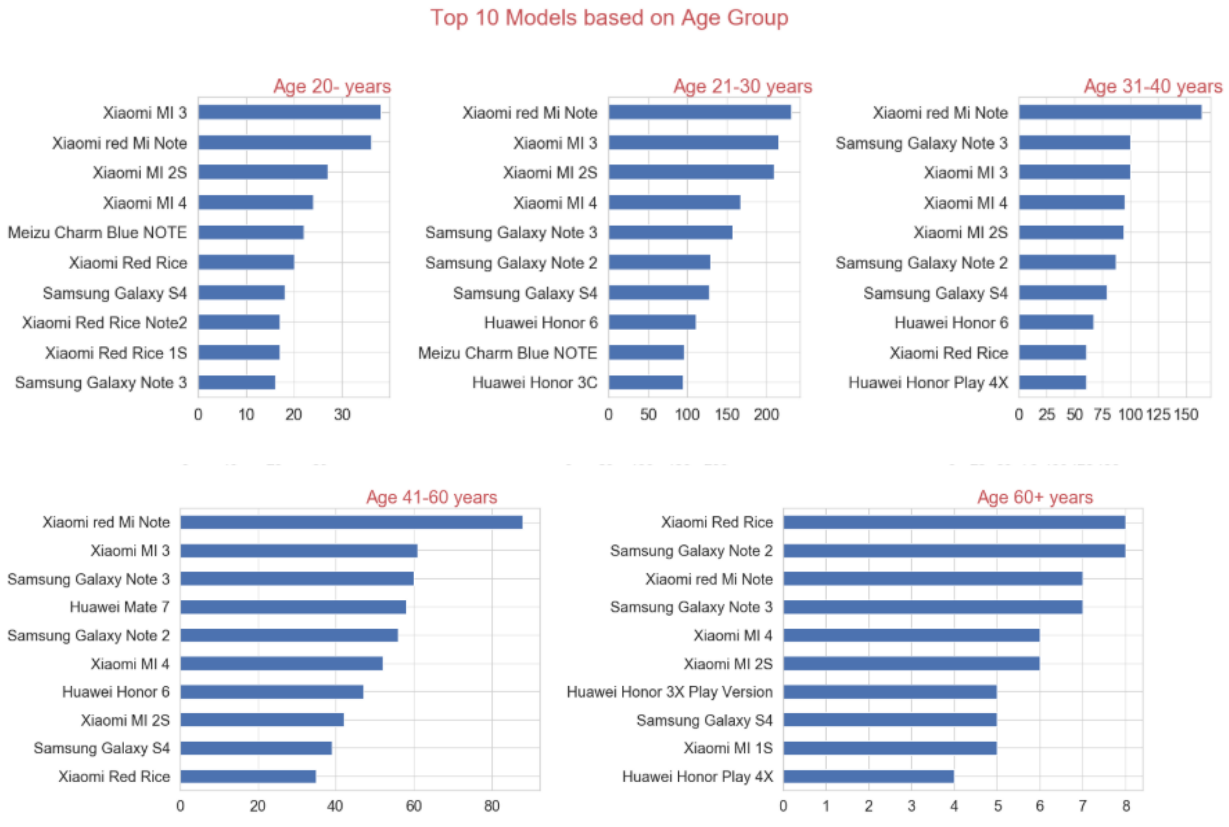
Conclusion: Top 10 brands of mobile used by users in this dataset are same for Male and Female users in this dataset. Except for Huawei Honor Play 4X and Xiaomi Red Rice, Top 10 models are the same for Male and Female users in this dataset.

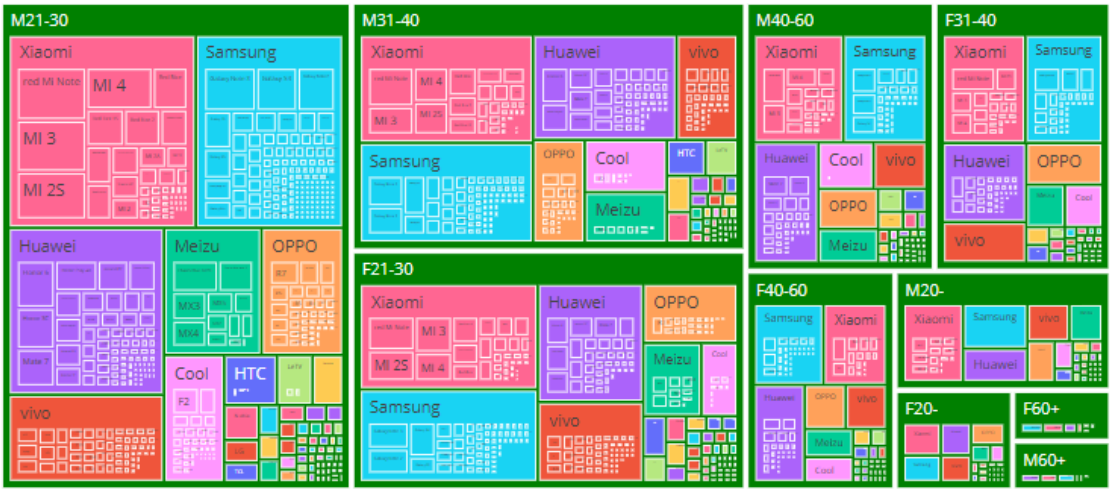
E. Top 10 Brands across Users of different Age Groups:



Conclusion: Except for Nubia, Top 10 brands remain the same amongst the users for various age groups. Nubia appears in Top 10 amongst the youngest and the oldest users, but that could be because of small number of users in those two groups.

F. Top 10 Models across Users of different Age Groups:

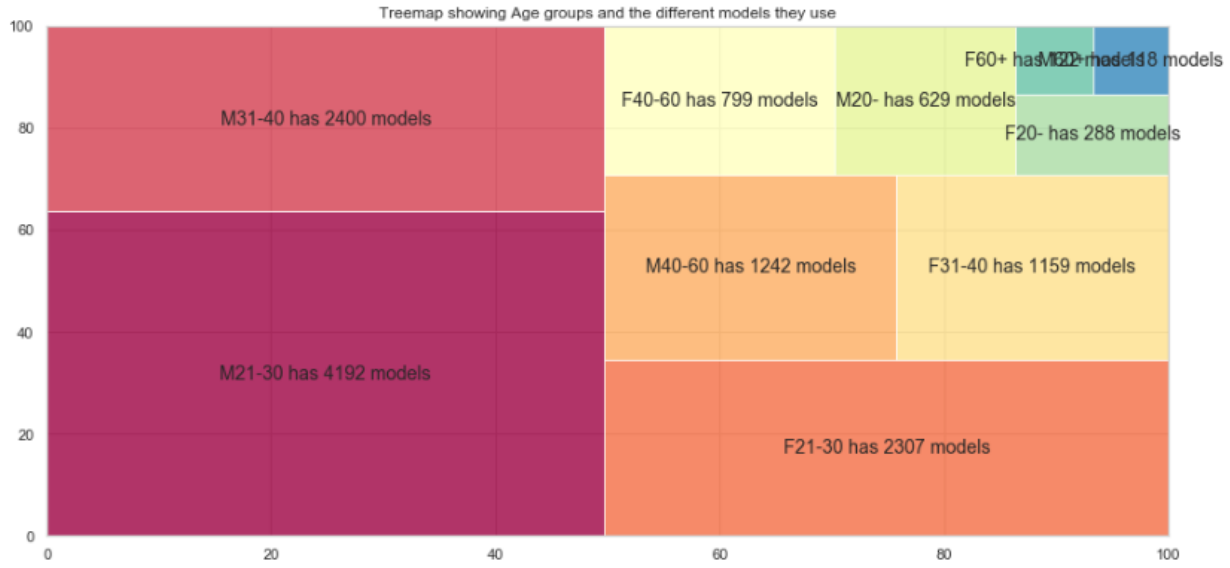




Conclusion: Xiaomi Red Rice seems to be more popular amongst the oldest age group (60+ years) but this could be because of small number of users in this group. Top 10 models remain mostly the same amongst the users for various age groups.

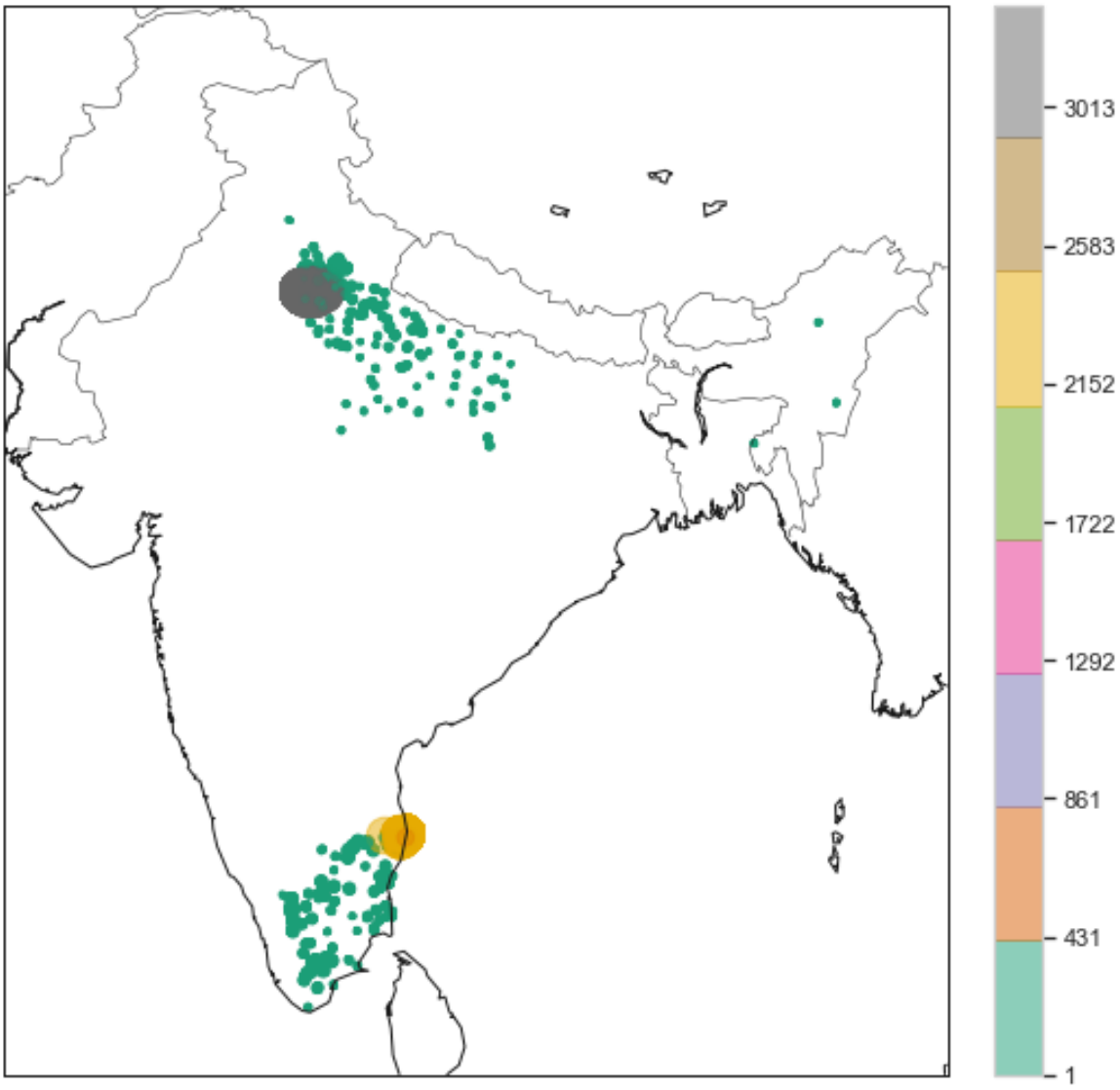
G. Diversity of Models Used by different Age Groups

M21-30	4192
M31-40	2400
F21-30	2307
M40-60	1242
F31-40	1159
F40-60	799
M20-	629
F20-	288
F60+	122
M60+	118



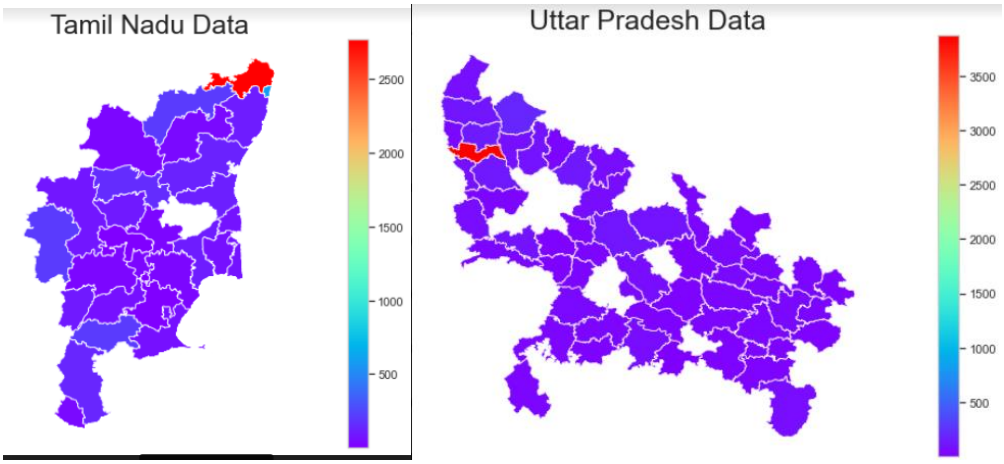
Conclusion: Males in the age Group 21-30 years had the most diversity in the models used by them. Amongst Females, 21-30 year olds used the most number of models.

8. Geographical Analysis of Users:



Conclusion: Most of the users are from Tamil Nadu and Uttar Pradesh.

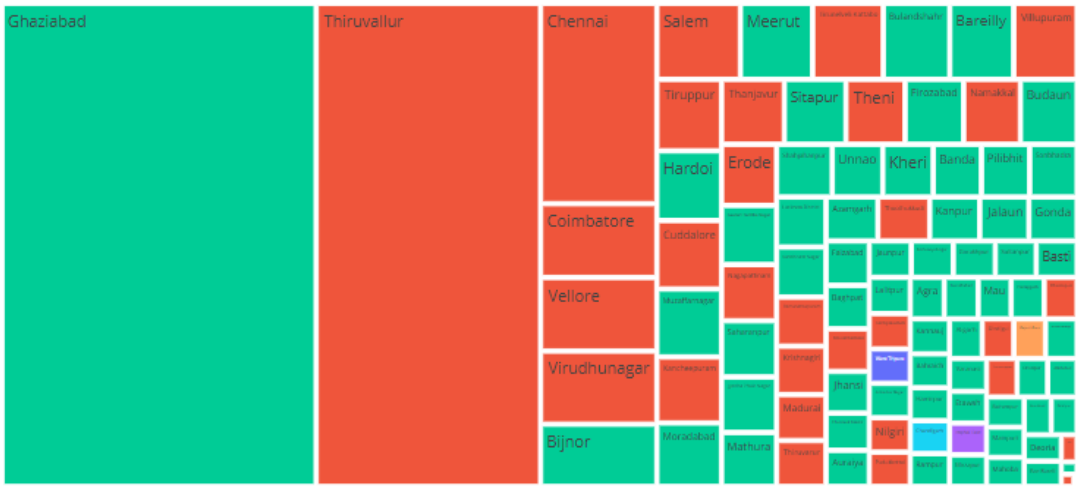
A. Density of Users across Tamil Nadu and Uttar Pradesh



Conclusion: Most of the users in Tamil Nadu in this data set are from near Chennai and most of the users in Uttar Pradesh in this data set are from near Delhi-NCR.

B. Analysis of Top 5 Cities for users in Tamil Nadu and Uttar Pradesh

TreeMap showing customer count in each city based on City



Conclusion: Top 5 cities in Tamil Nadu based on the highest number of Users in this data set are Thiruvallur, Chennai, Coimbatore, Vellore and VirudhuNagar.

Top 5 cities in Uttar Pradesh based on highest number of Users in this data set are Ghaziabad, Bijnor, Meerut, Bijnor, Hardoi and Muzzafarnagar.

PROPOSED SOLUTION FOR CUSTOMER

This report aims to increase the level of awareness regarding the customer data for Insaid Telecom. In this section we will summarize some of the key conclusions from the analysis and provide actionable insights for Insaid Telecom. The insights are aimed to help Insaid Telecom's Product and Marketing teams to follow data-driven and targeted marketing, which would in turn help in expanding the customer base and increasing the revenue for Insaid Telecom.

The analysis of the user data helped us reach the following conclusions:

- Amongst the 6 states, Uttar Pradesh & Tamil Nadu holds the highest number of customers.
    - Within Uttar Pradesh, Ghaziabad was the top city in terms of highest number of users.
    - Within Tamil Nadu, highest number of users came from Thiruvallur
- From the given dataset, we found that the highest number of users for Insaid Telecom were Males in the Age group of 20-40 years.
- Xiaomi, Samsung & Huawei were the top 3 most preferred brands amongst both men and women, across all age –groups.
- Mobile models used did not change much with Gender or Age.
- Network Usage peaked during mornings (4AM to 8AM) and Evenings (4PM to 8PM) during the time period of this dataset.

Actionable Insights:

- Insaid Telecom needs to make sincere efforts to increase their presence in Chandigarh, Arunachal Pradesh, Tripura and Manipur – since these states have only a handful of users for their network. This is a huge opportunity for Insaid Telecom to increase their market share and revenues.
- The network needs to improve usage amongst Female customers, since currently the network has only 33% of Total users who are Female. Hence, Insaid Telecom needs targeted marketing campaigns to reach out to Female customers.
- In terms of the Age Group, Insaid Telecom needs to focus on people older than 50 years, since Insaid Telecom does not have many users in this age group. This is another opportunity for Insaid Telecom, as the adoption of smart phone is now slowly increasing in this age group.
- Insaid Telecom has a third opportunity to expand its user base in Uttar Pradesh and Tamil Nadu by targeting smaller cities in both these states, where there are few users for their network.
- Insaid Telecom should also consider partnering with Xiaomi, Samsung and/or Huawei (for selected models) to provide value added service to its customers for retention, since these are the top 3 brands that Insaid's customers are using.

TOOLS

❖ DS TOOLS

Data science tools can be of two types. One for those who have programming knowledge and another for the business users. Tools which are for business users, automate the analysis.

Programming language tool used for our project was Python and the IDE used was Jupyter Notebook. Some of the other tools used in this project are mentioned below along with a basic description about them.

Data Processing and Modelling

NumPy

NumPy (Numerical Python) is a perfect tool for scientific computing and performing basic and advanced array operations.
The library offers many handy features performing operations on n-arrays and matrices in Python. It helps to process arrays that store values of the same data type and makes performing math operations on arrays (and their vectorization) easier. In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

Highlights:

- Easy to use and interactive
- Open-source contribution and ample community support
- Simplifies the process of complex mathematical implementations

Pandas

Pandas is a library created to help developers work with "labelled" and "relational" data intuitively. It's based on two main data structures:

"Series" (one-dimensional, like a list of items) and "Data Frames" (two-dimensional, like a table with multiple columns). Pandas allows converting data structures to DataFrame objects, handling missing data, and adding/deleting columns from DataFrame, imputing missing files, and plotting data with histogram or plot box. It's a must-have for data wrangling, manipulation, and visualization.

Highlights:

- Ability to perform custom types of operations
- Ensures that the entire process of data manipulation is easier
- Offers high flexibility and functionality when used with other Python libraries and tools
- Outstanding speed indicators
- Selects the best-suited output for the apply method
- Supports aggregations, concatenations, iteration, re-indexing, and visualizations operations

Data Visualization

Matplotlib

This is a standard data science library that helps to generate data visualizations such as two-dimensional diagrams and graphs (histograms, scatterplots, non-Cartesian coordinates graphs). Matplotlib is one of those plotting libraries that are really useful in data science projects — it provides an object-oriented API for embedding plots into applications.

It's thanks to this library that Python can compete with scientific tools like MatLab or Mathematica. However, developers need to write more code than usual while using this library for generating advanced visualizations. Note that popular plotting libraries work seamlessly with Matplotlib.

Highlights:

- Full control of axes properties, font properties, line styles, etc. via an object-oriented interface
- Legend for scatter
- Provides a MATLAB-like interface for simple plotting
- Secondary x/y axis support
- Works great with several graphics backends and operating systems

Seaborn

Seaborn is based on Matplotlib and serves as a useful Python machine learning tool for visualizing statistical models – heatmaps and other types of visualizations that summarize data and depict the overall distributions. When using this library, you get to benefit from an extensive gallery of visualizations (including complex ones like time series, joint plots, and violin diagrams).

Highlights:

- Automatic estimation as well as the plotting of linear regression models
- Comfortable views of the overall structure of complex datasets
- Eases building complex visualizations using high-level abstractions for structuring multi-plot grids
- Options for visualizing bivariate or univariate distributions
- Specialized support for using categorical variables

Plotly

This web-based tool for data visualization that offers many useful out-of-box graphics – you can find them on the Plot.ly website. The library works very well in interactive web applications. Its creators are busy expanding the library with new graphics and features for supporting multiple linked views, animation, and crosstalk integration.

Highlights:

- Extreme level of customizability
- Extreme level of interactivity
- Plotly has a clean interface that is explicitly designed to allow you to build your own APIs

CONCLUSION

After a thorough user base analysis, this report highlights the market opportunities that can be exploited by Insaid Telecom to increase their market share.

Insaid Telecom needs to:

- Use targeted Marketing for those demographic segments where they currently do not have a large user base
- Increase presence in smaller cities in bigger states like Tamil Nadu and Uttar Pradesh
- Increase presence in smaller states like Arunachal Pradesh, Tripura and Manipur
- Provide Value-added services in partnership with Phone Brands like Xiaomi and Samsung, for retention of customers as well as to gain new customers