Consulting Report for INSAID Telecom

- An EDA project by Capstone_1014 Team

Report Date: 4th September 2022

Table of Contents

1. Introduction 3
2. Project Description 3
3. Problem Statement 3
4. Problem Analysis 3
5. Sources of Data 4
6. Summary of Data Mining 5
7. Proposed Solution for Customers 6
8. Tools 14
9. Conclusion 14

1. Introduction

INSAID Telecom is major Telecom service provider based out of Gurugram. It needs some data driven insights for its product design and marketing team. These insights must act as an enabler for its decision making processes as it looks forward to expand its customer base in the states of West Bengal, Karnataka, Kerala, Gujarat, Bihar & Punjab.

2. Project Description

INSAID has been able to gather three data sets in form of CSV files to draw inferences from. The “event.csv” is the master file while the “gender.csv” & “mobile.csv” contain demographic and mobile specific data respectively. The three data sets are linked together by customer specific “device_ids”, which identify individual customers.

3. Problem Statement

In the past few years, the Telecom space has seen intense competition. The customer churn is observed along with a rise in customer expectation and the demand for better service at a lesser tariff.

- Most value added services are now being extended to the customers through a mobile
- Customers buying a new mobile phone and telecom service package looks for a choice worth the money
- Customers are increasingly demanding more capabilities at an affordable rate from their mobiles
- There is a need for a method to predict the customer behaviour that determines the mobile usage pattern and to effectively determine the need for the type of service extended
- Even though INSAID Telecom uses internal methodologies based upon managerial instinct to predict the user demand and product service rollout, **the design of the service and their prediction method may not be the best**
- Besides the product offering, the marketing effort may not be contributing to the revenue generation effort
- It is important to know the impact of these demographic influences to tailor a suitable product
- Team 1014 is undertaking this task with respect to six states – West Bengal, Karnataka, Kerala, Gujarat, Punjab, Bihar

4. Problem Analysis

The Capstone_1014 team has decided to follow the classic strategy followed by INSAID consisting of critically examining the available data and subjecting it to data wrangling processes to glean the customer data using a statistical approach

5. Sources of Data

The data received by the Capstone_1014 team was in the form of three "csv" files named

- "events_data.csv"
- "gender_age_train.csv"
- "phone_brand_device_model.csv"

The "events_data.csv" data set has the following features:

Column Name	Description
event_id	Unique Instance Identifier This identifies the each event for the data to be generated in the data set.
device_id	Unique to each customer An anonymised number that identifies each unique device linked to a customer
timestamp	The instance at which a device_id is used for making a call
longitude	Geographical Variable to determine location of customer
latitude	Geographical Variable to determine location of customer
city	Location of the recorded event_id identified by City
state	Location of the recorded event_id identified by State

The "gender_age_train.csv" data set has the following features:

Column Name	Description
device_id	Unique to each customer An anonymised number that identifies each unique device linked to a customer
gender	Related to the record of device_id mapped while applying for a connection
age	Age of the customer identified by the device_id
group	A segmented data linking gender and age features

The "phone_brand_device_model.csv" data set has the following features:

Column Name	Description
device_id	Unique to each customer An anonymised number that identifies each unique device linked to a customer
phone_brand	The Brand of the Phone Manufacturer
device_model	The particular model of the Brand

6. Summary of Data Mining

Challenge 1

On examining the three data sets it was seen that these data sets had the "device_id" column that provided the common link between the data sets.

However, when the three data sets were merged using inner join to identify the customers common to all the data sets, there were only 406 such customers were found to be present. Further, when filtration was undertaken for the target states, this figure was dropping to 120.

Hence it was seen that the data was complete in respect of only 406 customers. As per the problem statement, insights into the customer behaviour using demographic data was the goal.

Solution

There were two courses of action available. The first one was to use 406 values (of which only 120 were available in the data set pertaining to the target states) and the second one of imputing entire column filling missing values.

In order to minimise the error and considering "60865 to 406" value gap of 60459 as unacceptable, it was decided to impute 60459 values via a random selection from gender and phone data sets in sequence as separate columns. This was expected to provide balance to the data while ensuring adequate spread thus reducing the overall error and increase the accuracy of insights. Thus, the missing values in combined data set after dropping duplicates have been imputed from the mobile and gender data sets respectively.

Challenge 2

It was evident that no useful insights could be drawn from the segmentation evident in the group column in the "gender_age_train.csv" file.

Solution

A new column (new_group) was generated based on the Age column.

Challenge 3

The "phone_brand" column in the "phone_brand_device_model.csv" had characters in Mandarin language.

Solution

These were translated into English. (The customer had indicated that he needed data only for top 10 most popular phone brands.)

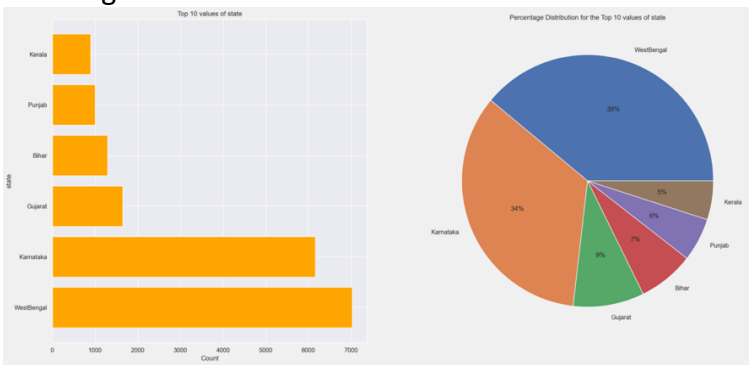
7. Proposed Solution for Customers

Analysis

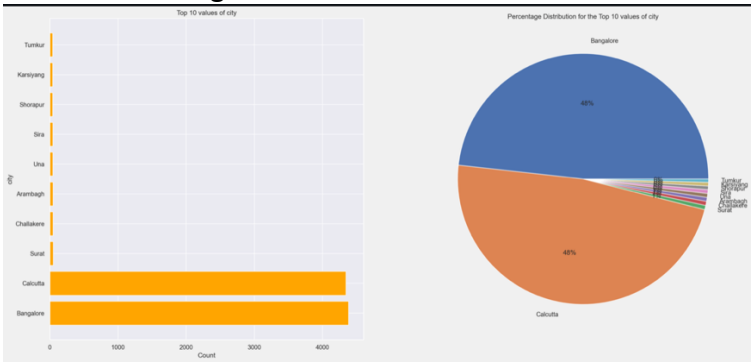
(a) At the outset the dataset comprises of primarily categorical variables of demographic nature

	Feature Name	Observations
0	timestamp	The highest freq recorded is 2016-05-01 10:02:04
1	city	The highest freq recorded is Bangalore
2	state	The highest freq recorded is WestBengal
3	gender	The highest freq recorded is M
4	new_group	The highest freq recorded is (20, 30]
5	phone_brand	The highest freq recorded is Xiaomi
6	device_model	The highest freq recorded is 红米note

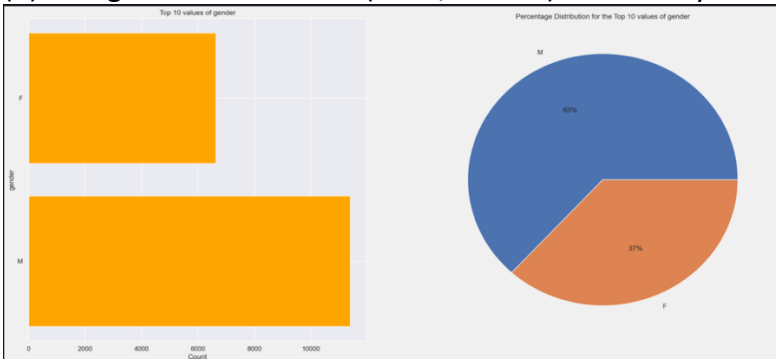
(b) Over 73% of the customers in the relevant data set belong to two states of West Bengal & Karnataka



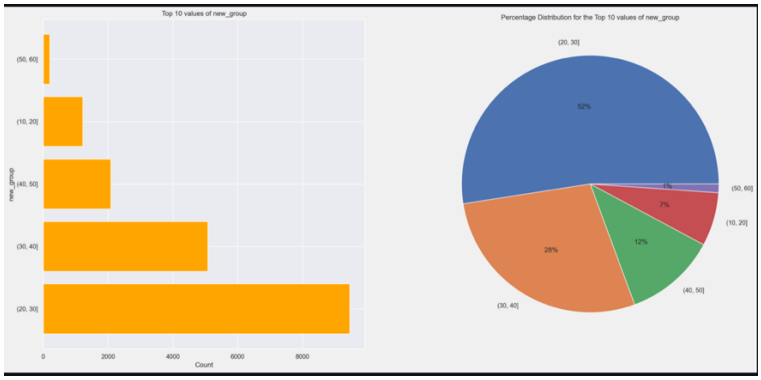
(c) The city wise distribution indicates that over 96% of customers belong to the two cities of Bengaluru & Kolkata



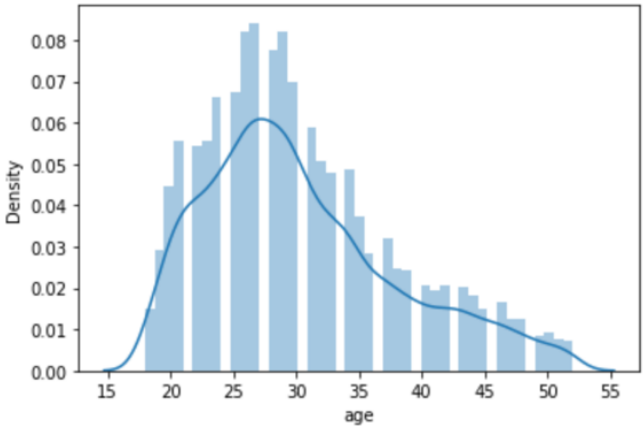
(d) The gender distribution (Male/Female) is distinctly in favour of Males by 26%



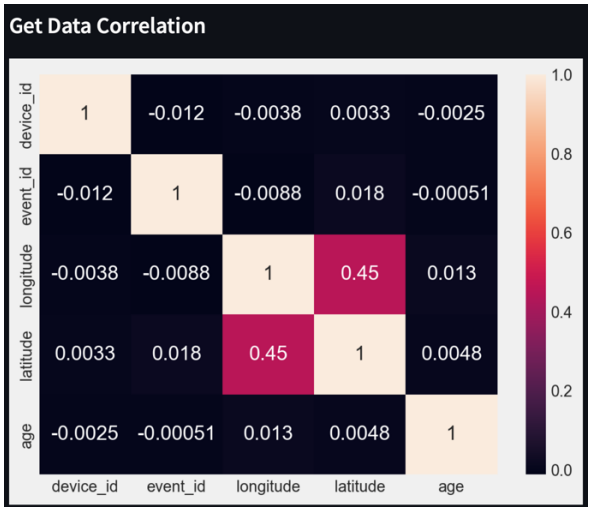
(e) Over 80% of customers belong to age group 20-40 years with over 52% belonging to 20-30 age group. This number increases to 87% for 18-40 year group



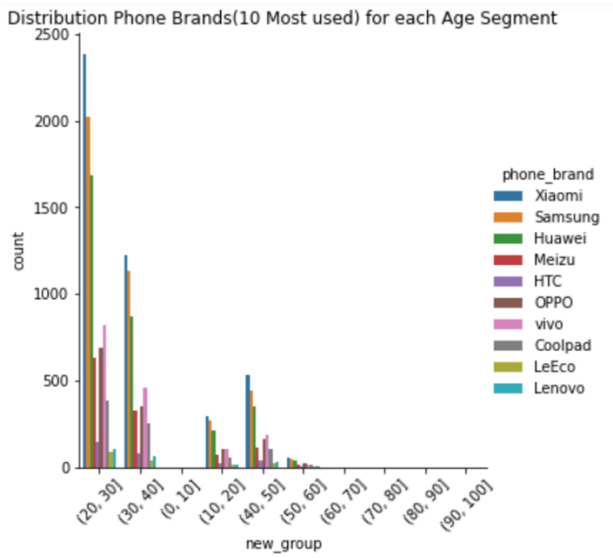
(f) The age data despite having been corrected for outliers, is seen to be right skewed



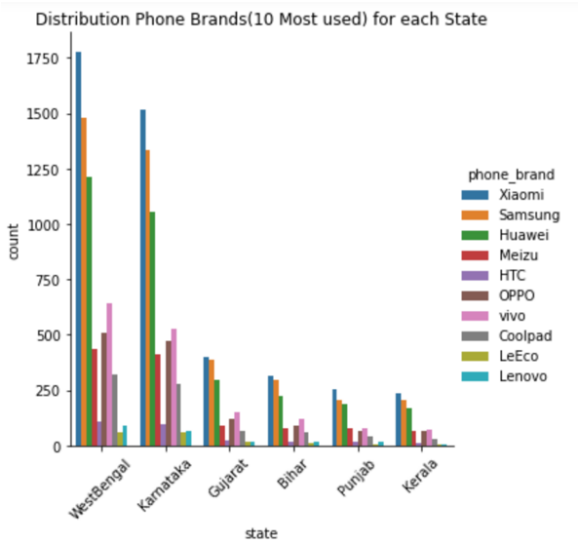
(g) A heat map generated expected results. Only latitude and longitude data is correlated



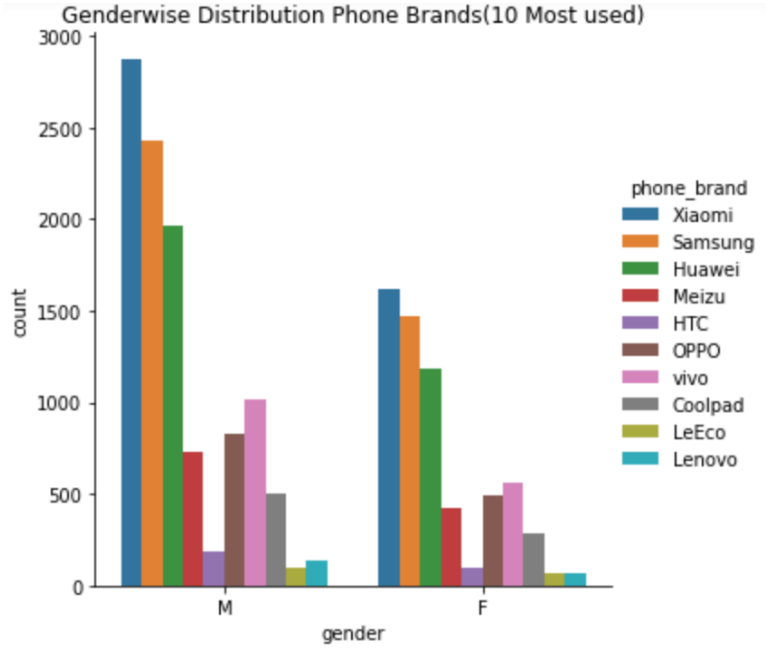
(h) Customers showed distinct preference for Xiaomi followed by Samsung across age groups. The customer preferences did not change across the groups



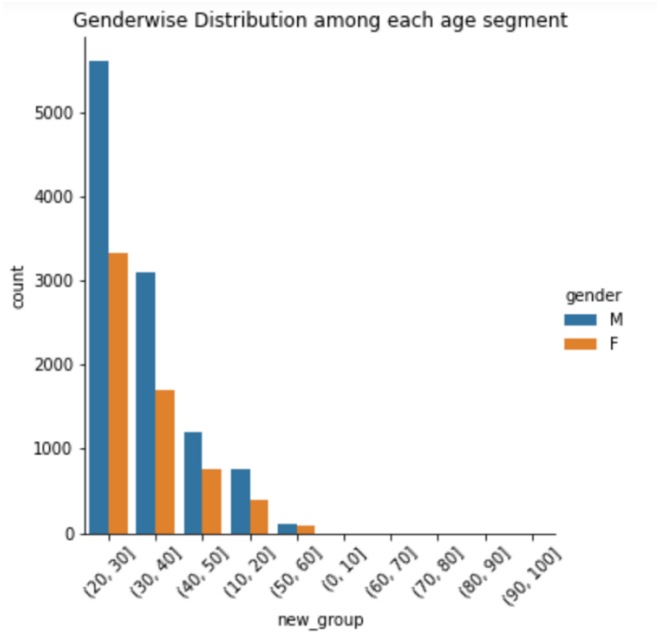
(i) The phone_brand preferences across the states remain the same



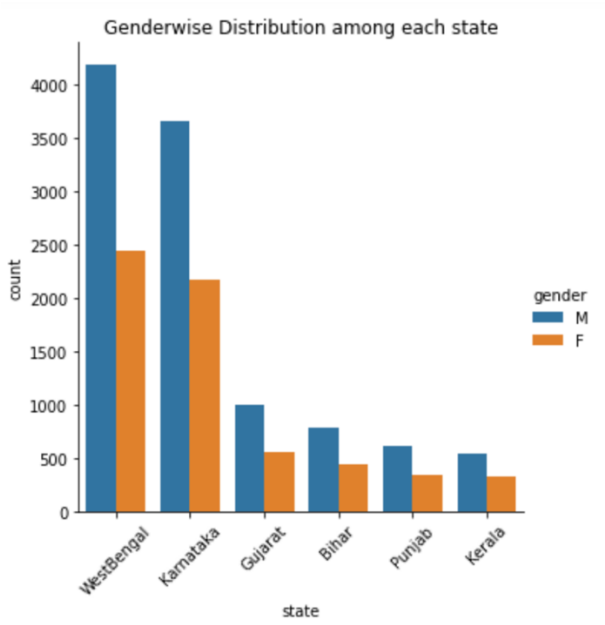
(j) There was no change noticed in the phone brand preference across the genders. Xiaomi followed by Samsung and the other eight considered



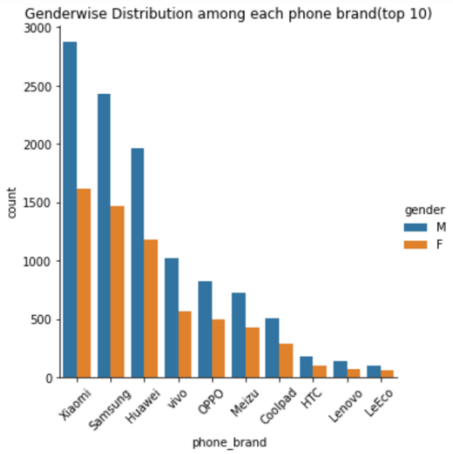
(k) Across the age segments the gender gap remains but reduces imperceptibly as age increases



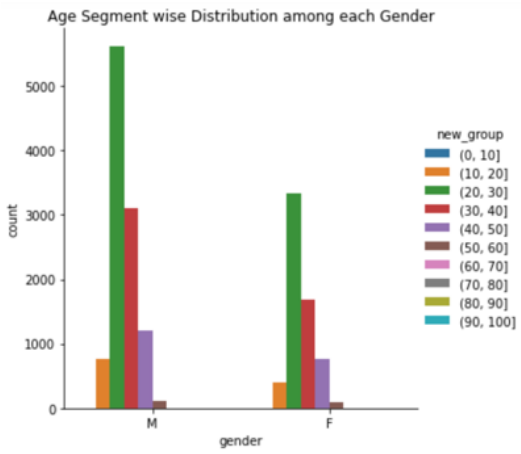
(l) The gender gap continues to be seen and male subscribers distinctly outnumber the female subscribers



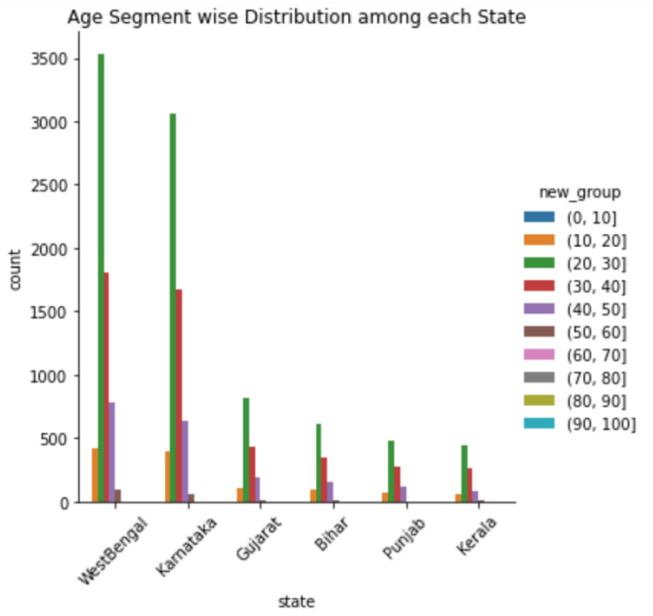
(m) Phone_brand preferences remain same irrespective of the gender. So Gender has no impact on phone_brand preferences



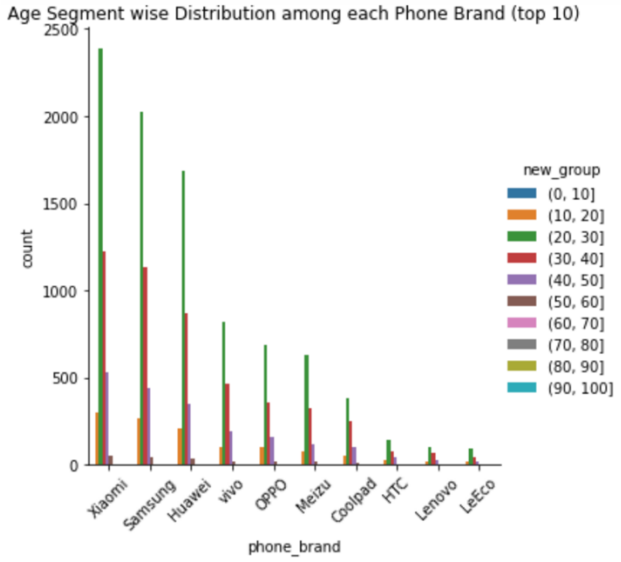
(n) Age segment wise distribution when viewed from gender perspective indicates maximum number of users in range 20-30 years followed by 30-40 age group



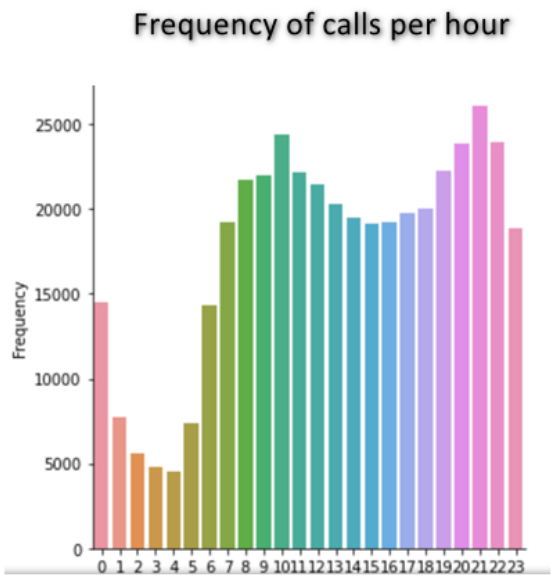
(o) No significant change in distribution observed across the states



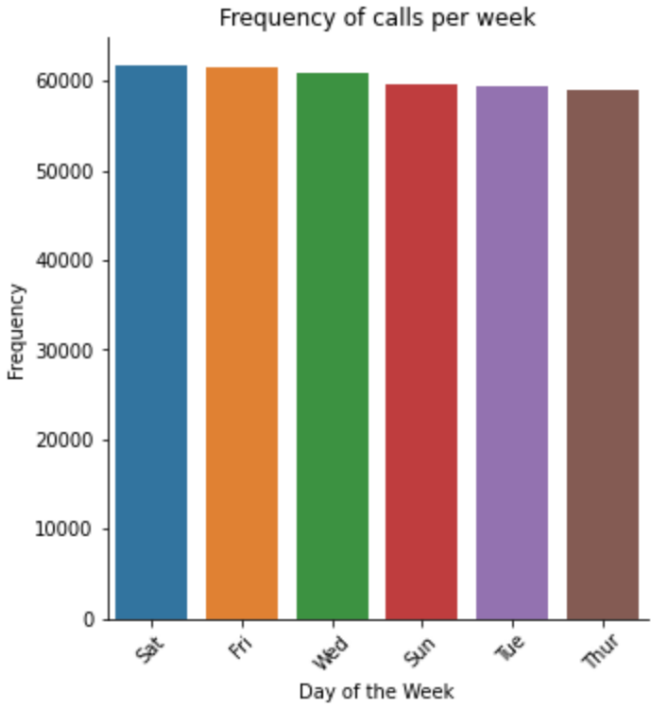
(p) The top ten brand preferences remain same for the differing age segments across the sample set



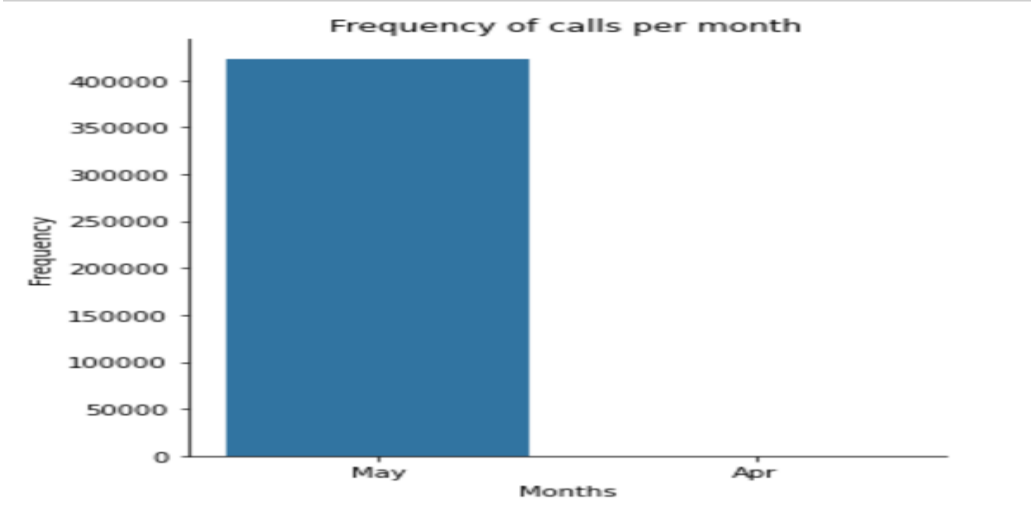
(q) The frequency of calls were maximum between 0800 to 2200 hrs peaking at 1000 and 2100 hrs. Significant number of calls have been observed around midnight 0000 hrs



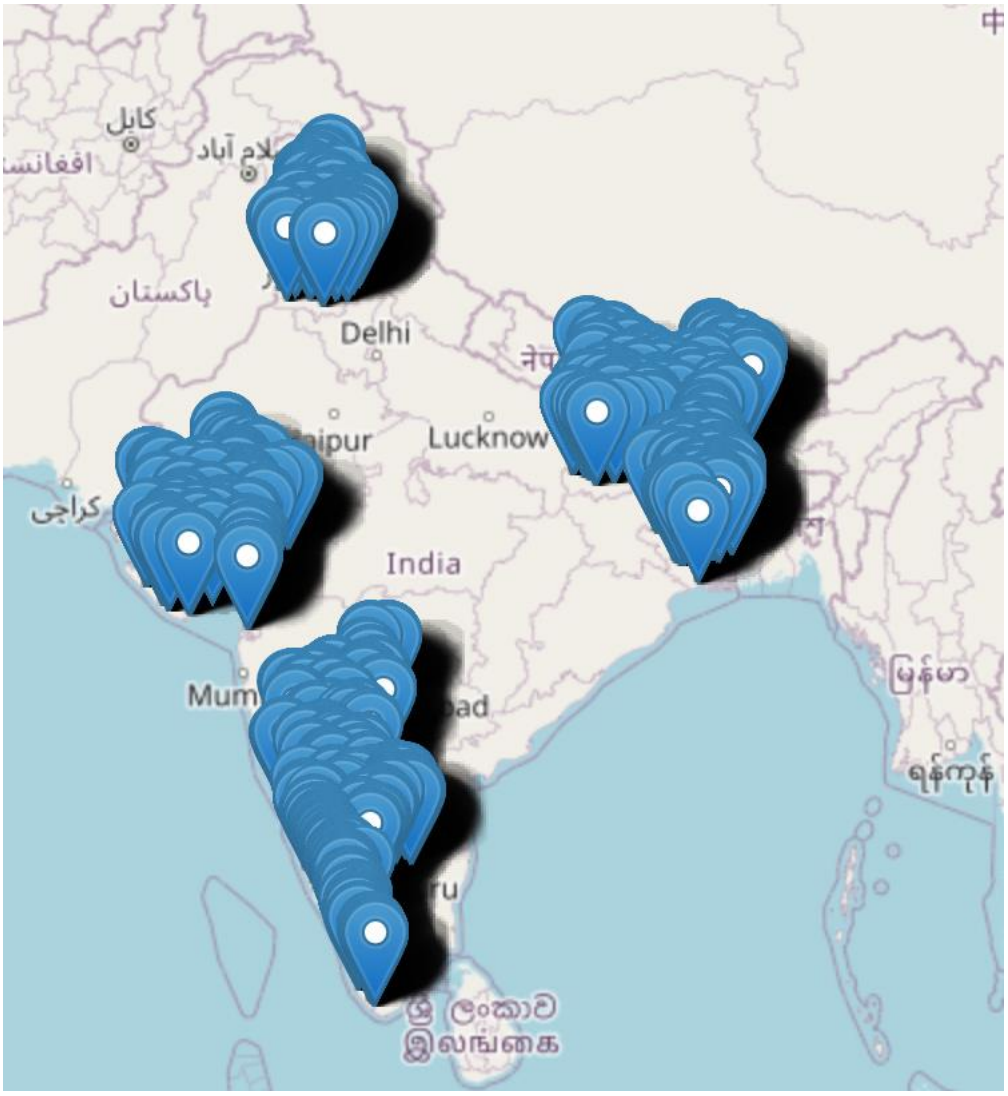
(r) The week wise summary appears to be erroneous as no calls seem to have been made on Monday with almost 60,000 plus calls having been made every day of the week except on Monday



(s) The month wise summary indicates that all the calls in the data set were made in the months of April and May with most calls having been made in May



(t) Distribution of Customers State-wise



Proposed Solution

- The telecom company must focus on expanding its presence beyond the cities of Bengaluru and Kolkata and focus on Tier II and Tier III cities across the target states

- The highest number of subscribers present in Kolkata and Bengaluru cities could also be an indication of the poor tower coverage in other cities of the target states. This needs to be further evaluated

- The gender disparity needs to be addressed through suitable product design accompanied by a focussed marketing effort targeting females

- As a part of co-branding effort, Xiaomi and Samsung may be approached as these are preferred brands across genders and age groups as well as across states

- The targeting of age groups above 40 years must be a focussed area of expansion as the customer base in these regions is relatively low

- The low utilisation of available capacity between 0100 hrs to 0800 hrs may be increased using appropriated marketing tools such as offers of free internet during this period

8. Tools

DS Tools

- Anaconda Environment
- Jupyter Notebook
- Pycharm
- Git
- Streamlit
- Heroku

Web UI Tools

- PHP
- JavaScript

9. Conclusion

- This project was undertaken by the Capstone_1014 team to provide actionable insights for INSAID Telecom as per the data provided
- The same was evaluated using EDA and statistical approach and some proposed solutions have been suggested
- It is also highlighted that the data collection practices of INSAID telecom needs to be improved, to avoid data imbalance and to gain insights for better analysis and identifying business improvement opportunities