Mobile Usage Analysis

Consulting Report

Table of Content

1. Introduction

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

Descriptive analytics is the interpretation of historical data to better understand changes that have occurred in a business. Descriptive analytics describes the use of a range of historic data to draw comparisons.

Most commonly reported financial metrics are a product of descriptive analytics—for example, year-over-year pricing changes, month-over-month sales growth, the number of users, or the total revenue per subscriber. These measures all describe what has occurred in a business during a set period.

Analytics often favours data visualization to communicate insight. Dashboards are a data visualization tool that allow all users to understand the analytics that matter to their business, department or project. Even for non-technical users, dashboards allow them to participate and understand the analytics process by compiling data and visualizing trends and occurrences. Data dashboards provide an objective view of performance metrics and serve as an effective foundation for further dialogue. A dashboard is a business intelligence tool used to display data visualizations in a way that is immediately understood.

Behind the scenes, a dashboard connects to your files, attachments, services and API's, but on the surface displays all this data in the form of tables, line charts, bar charts and gauges. A data dashboard is the most efficient way to track multiple data sources because it provides a central location for businesses to monitor and analyse performance. Real-time monitoring reduces the hours of analysing and long line of communication that previously challenged businesses.

2. Project Description

It is always wonderful to see services customized to your needs. Businesses try to understand your behaviour and adjust their offerings so as to ensure you feel attached to their services.

InsaidTelecom, one of the leading telecom players, understands that customizing offering is very important for its business to stay competitive. Currently, InsaidTelecom is seeking to leverage behavioural data from more than 60% of the 50 million mobile devices active daily in India to help its clients better understand and interact with their audiences.

3. Problem Statement

Objective of this consulting assignment is to build a dashboard to understand user's demographic characteristics based on their mobile usage, geolocation, and mobile device properties. Doing so will help millions of developers and brand advertisers around the world

pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

4. Problem Analysis

Over the past couple of years, the Indian telecom industry has been going through a paradigm shift from a voice-centric market to a data-centric market. Therefore, telcos have started shifting beyond traditional telecom business to wider digital consumer space such as content, mobile banking solutions, etc. Due to the competition intensity in telecom industry, the fittest will survive.

Some of the industry opportunities include mobile penetration, increase in internet users, untrapped rural market, exploring adjacent businesses in an evolving environment (digital consumer space like content and mobile banking solutions) and customized offerings.

InsaidTelecom has understood the importance of customized offerings and wants to analyse the behavioural data. This will aid in not only customizing offerings but also lead to innovations in developing new or additional services.

Descriptive analytics will be applied on the user data to find out underlying insights and facts. Dashboards are used to visualize the data. This allows anyone to understand the analysis.

5. Sources of Data

We are going to study the demographics of a user (gender and age) based on their app download and usage behaviours.

The Data is collected from mobile apps that use InsaidTelecom services. Full recognition and consent from individual user of those apps have been obtained and appropriate anonymization have been performed to protect privacy. Due to confidentiality, InsaidTelecom won't provide details on how the gender and age data was obtained.

InsaidTelecom has given the permit to access their 'MySQL' server in order to extract the data needed for the analysis. Here are the server and database details:

host	'cpanel.insaid.co'
user	'student'
passwd	'student'
database	'Capstone1'

The data schema can be represented in the following table:

Table 1, **gender_age_train** contains the customers age and gender information.

Feature Name	Description
device_id	Handset device id
gender	User's gender

age	User's age
group	User's age group

Tab.1: gender_age_train features and description

Table 2, **phone_brand_device_model** holds phone brand and models details.

Feature Name	Description
device_id	Handset device id
phone_brand	Phone brand
device_model	Handset model

Tab.2: phone_brand_device_model features and description

Table 3, **events_data** contains event id, location geo codes and location data.

When a user uses mobile on INSAID Telecom network, the event gets logged in this data. Each event has an event id.

Feature Name	Description
event_id	Event id
device_id	Handset device id
timestamp	Time of the user activity
longitude	Longitude of user location
latitude	Latitude of user location
city	User's city
state	User's state

Tab.3: events_data features and description

6. Summary of Data Mining

Data mining is the exploration and analysis of large data to discover meaningful patterns and rules. It's considered a discipline under the data science field of study and differs from predictive analytics because it describes historical data, while data mining aims to predict future outcomes. Additionally, data mining techniques are used to build machine learning (ML) models that power modern artificial intelligence (AI) applications such as search engine algorithms and recommendation systems.

Data preparation is the process of data cleansing, and missing data is included to ensure it is ready to be mined. Data processing can take enormous amounts of time depending on the amount of data analysed and the number of data sources. Therefore, distributed systems are used in modern database management systems (DBMS) to improve the speed of the data mining process rather than burden a single system. They're also more secure than having all an organization's data in a single data warehouse. It's important to include failsafe measures in the data manipulation stage so data is not permanently lost.

1. There are null/zeros in "device_id" of events_data which should not be the case.
2. There are null/zeros in "latitude" and "longitude".
3. Some of the "latitudes" and "longitudes" are wrong.

4. "state" column of events_data has null values in the DataBase. Retrieve those rows and fill them appropriately.
5. Phone brand and device models got some non-English names which need translation.

Missing value is availability of incomplete observations in the dataset. This is found because of reasons like, incomplete submission, wrong input, manual error etc. These missing values affect the accuracy of model. So, it becomes important to check missing values in our given data.

Missing Value Analysis in Given Data:
In the given dataset it is found that there are lot of values which are missing. It is found in the following types:

- Blank space: Which are converted to NA and NaN in R and Python respectively for further operations
- Zero Values: This is also converted to NA and Nan in R and python respectively prior further operations

Following the standards of percentage of missing values, we now have to decide to accept a variable or drop it for further operations. Industry standards ask to follow following standards:
1. Missing value percentage < 30%: Accept the variable
2. Missing value percentage > 30%: Drop the variable

Null values present in the following features:
"device_id", "latitude", "longitude" and “state”

Impute the missing value:
After the identification of the missing values the next step is to impute the missing values. Imputation is normally done by following methods.

- Central Tendencies: by the help of Mean, Median or Mode
- Distance based or Data mining method like KNN imputation
- Prediction Based: It is based on Predictive Machine Learning Algorithm

In order to choose the best of the above methods, it is necessary to know which method predicts, values close to the original data. This done by taking a subset of data, noting down its original value and the replacing that value with NA. Now apply all methods and note down every output. For the given dataset, data mining method was applied for the missing value imputation.

Wrong/Outlier Analysis:
Outlier is an abnormal observation that stands or deviates away from other observations. It may occur due to exceptional case (correct value but exceptional data) or manual updating error or poor quality of data. This may result in incorrect predictions. So, it is recommended that outlier check is important. The outlier values can either be removed or replaced.

Outliers are found in “longitude” and “latitude” features. Best way to look for wrong values is to plot for easy recognition. “longitude” and “latitude” are plotted on the map and scatter plot in Fig.1. This helped in finding and eliminating the wrong values. Fig.2, shows the geo code distribution after the wrong values elimination.

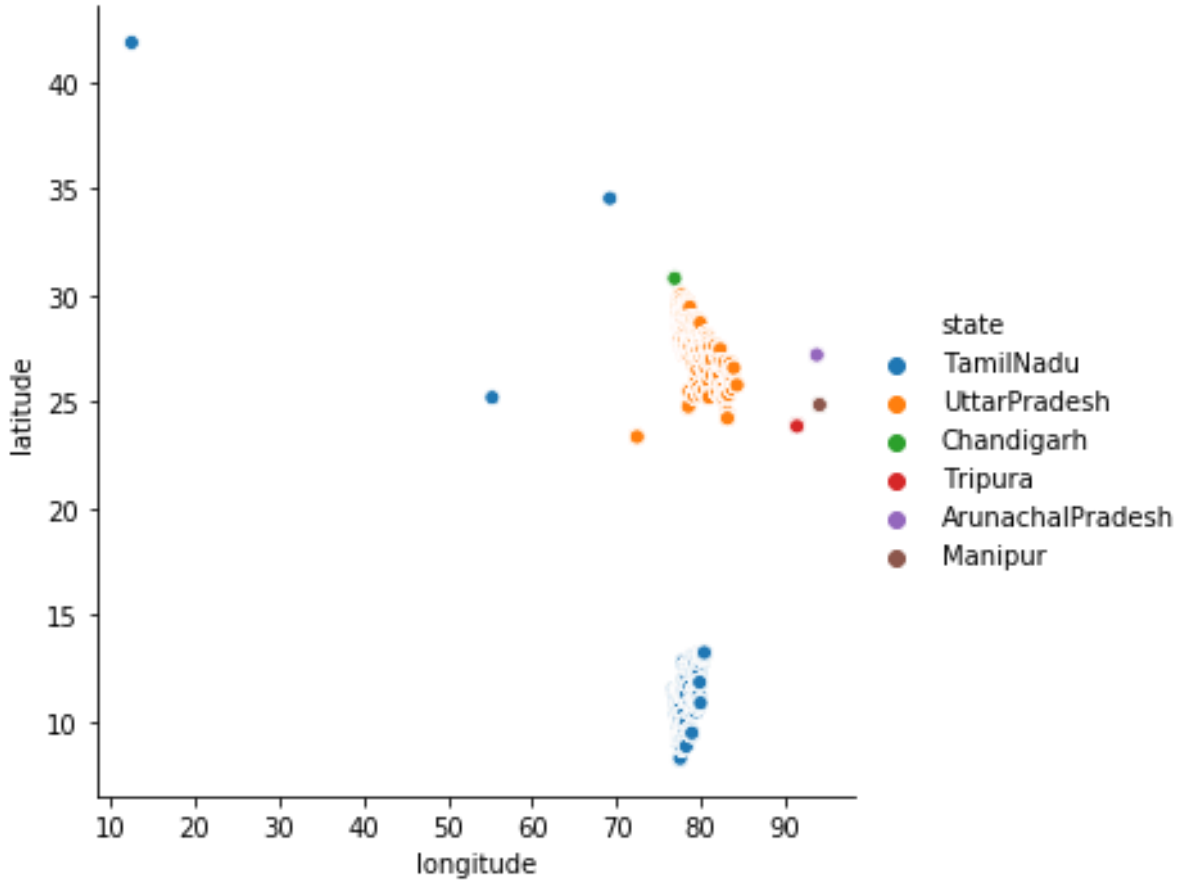

Fig 1: lattitude vs longitude

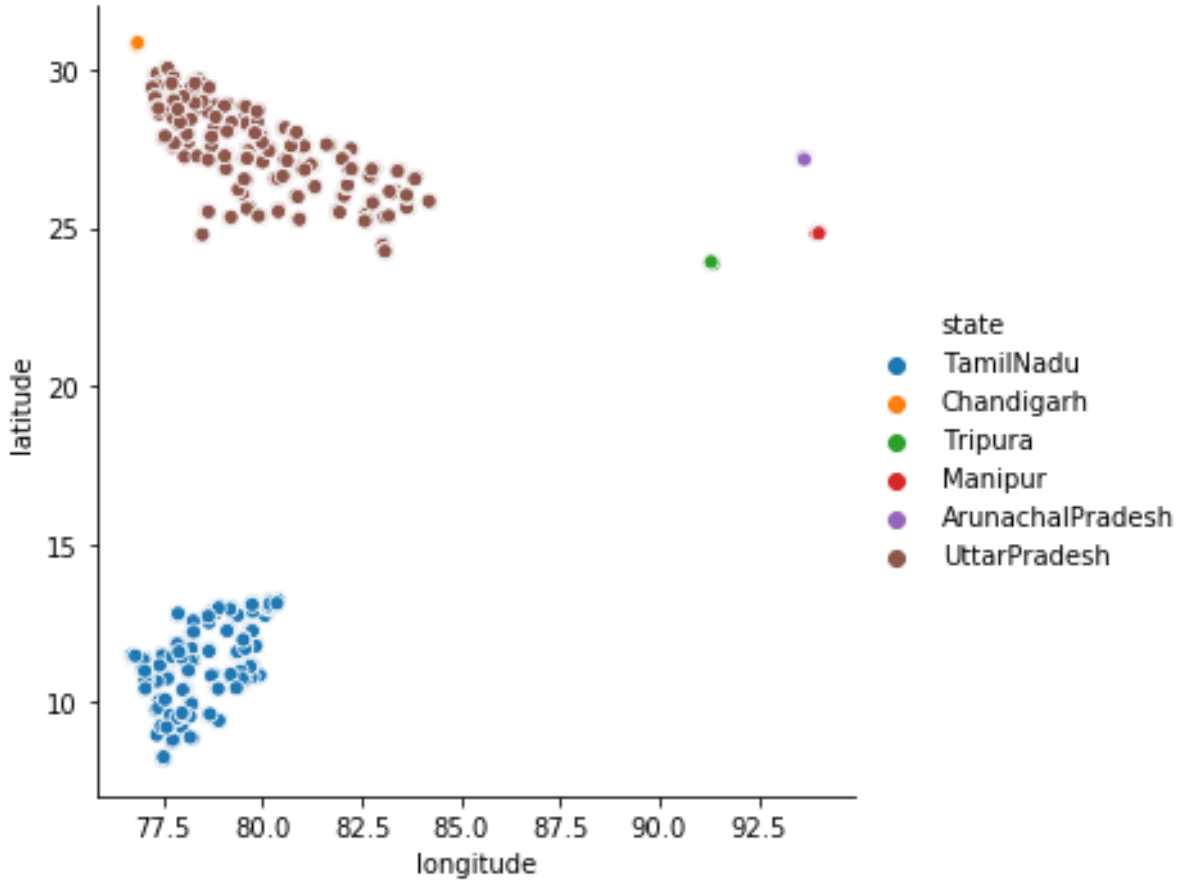

Fig 2: lattitude vs longitude after wrong geo code elimination

Brand data frame includes, non-English phone brand and device models. These need English translation for readability. Google translator package was used and failed due to API time restriction. Using the predefined translations provided by INSAIDTelcom in Tab.4.

Brand Name	Brand English Mapping
'华为'	'Huawei'
'小米'	'Xiaomi'
'三星'	'Samsung'
'vivo'	'vivo'
'OPPO'	'OPPO'
'魅族'	'Meizu'
'酷派'	'Coolpad'
'乐视'	'LeEco'
'联想 '	'Lenovo'
'HTC'	'HTC'

Tab. 4: Chinese-English translation

7. Proposed Solution for Customers

The analysis is limited to certain factors:

- Only 6 states have been chosen for the analysis namely TamilNadu, Manipur, Chandigarh, Tripura, UttarPradesh and ArunachalPradesh.
- The comparison of data is only for one year i.e., 2016.

7.1. Number of records

Post the data pre-processing, the number of records analysed are **411599.**

7.2. Distribution of Customers State-wise

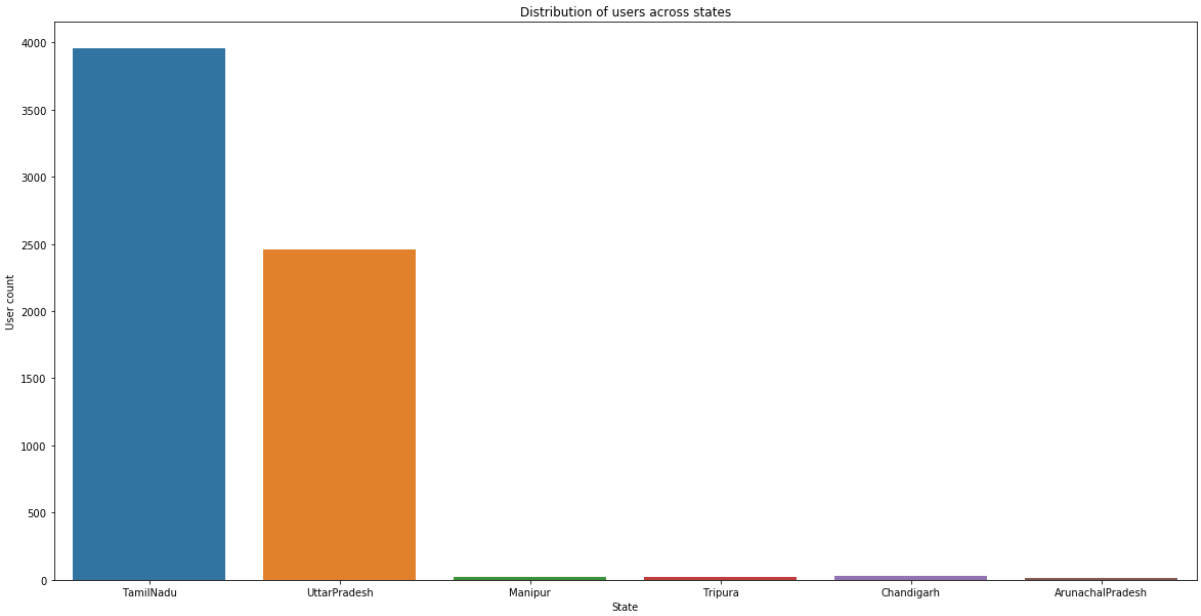

Fig.3: Distribution of Customers State-wise

7.3. Customer's distribution across India

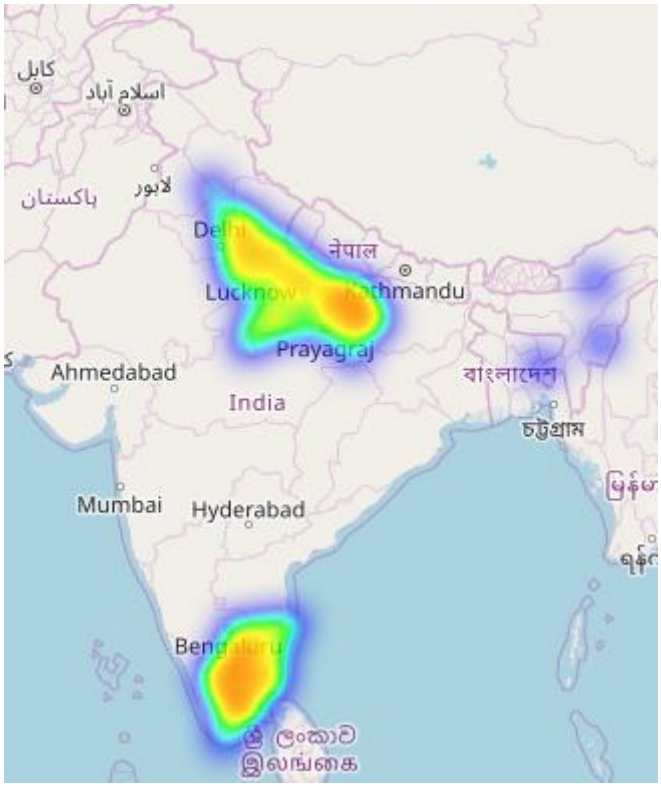

Fig.4: Customer's distribution across India

7.4. Frequency of Events with respect to Time

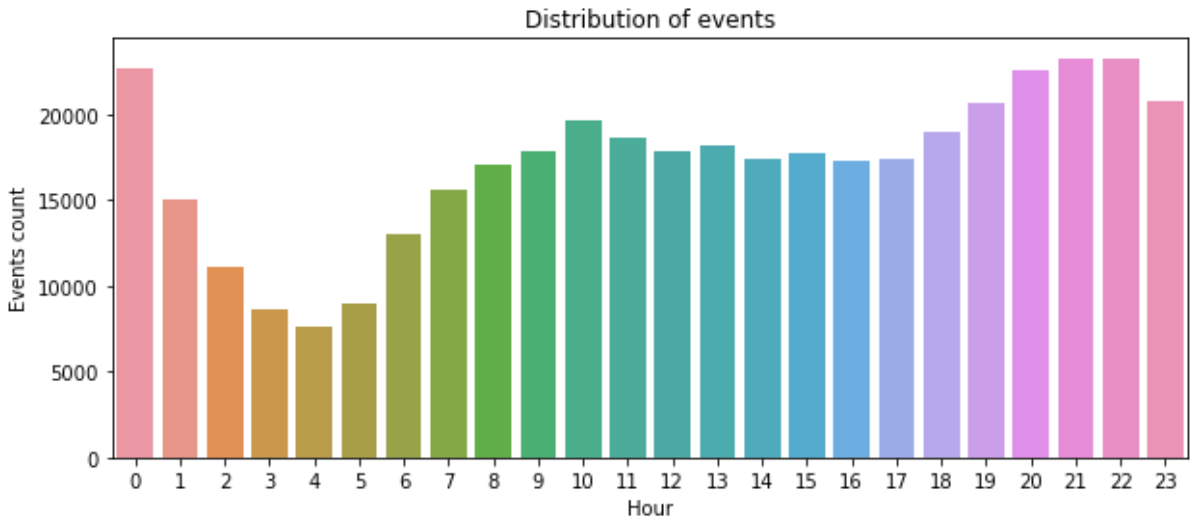

Fig.5: Distribution of events across the day

7.5. Distribution of Mobile Phone Brands among Customers

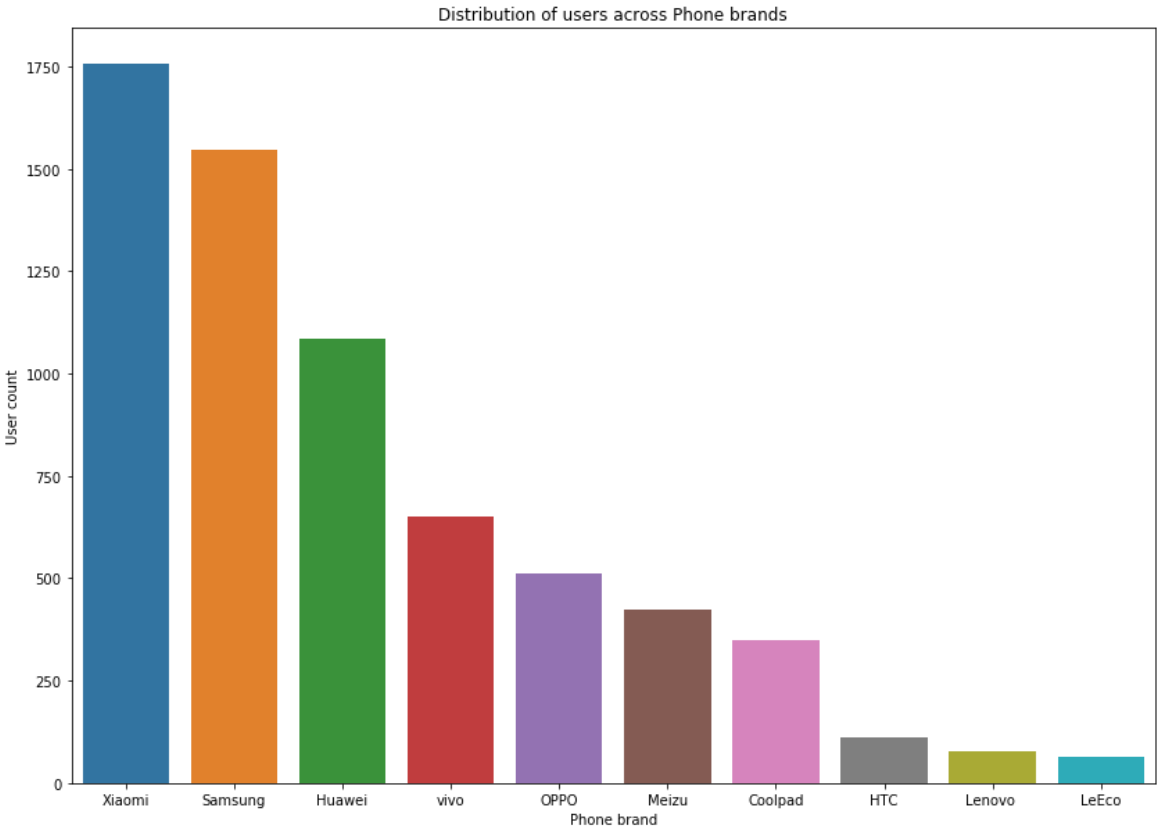

Fig.6: Distribution of mobile phone brands among customers

7.6. Mobile Phone users based on Gender

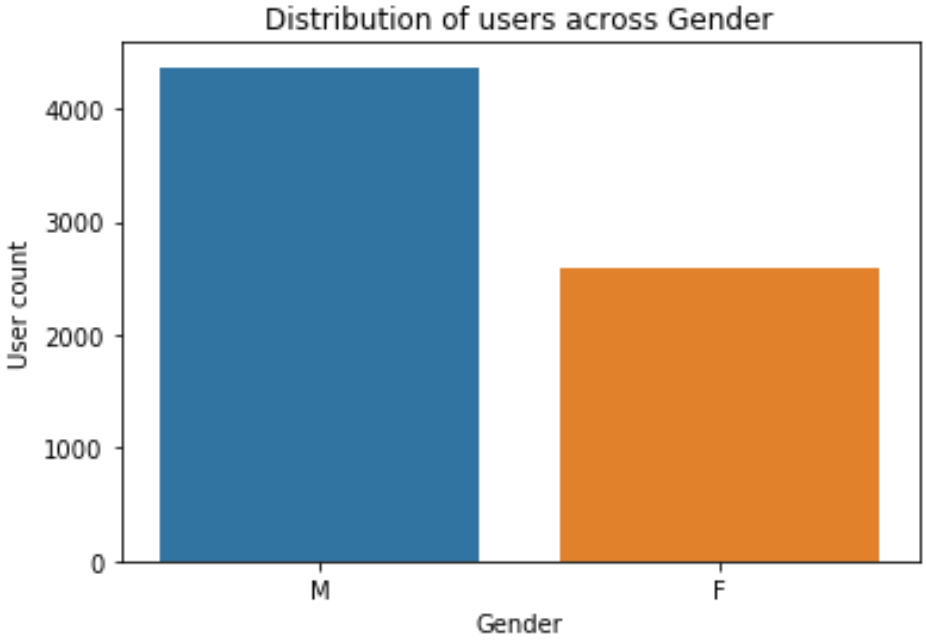

Fig.7: Mobile phone users based on gender

7.7. Distribution of different Age Groups

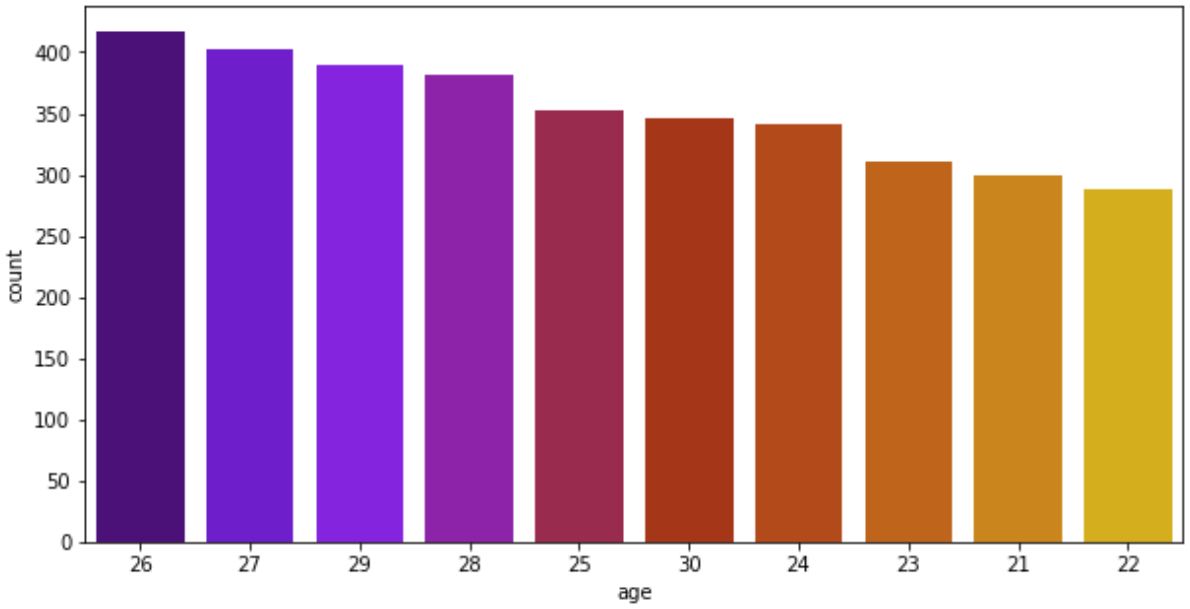

Fig.8: Distribution of customers age groups

7.8. Observations:

- TamilNadu has the highest number of customers followed by UttarPradesh
- ArunachalPradesh is with the least.
- Xiaomi has got most customers followed by Samsung and Huawei
- 26.7% of the market is dominated by Xiamomi
- 62.7% of the users are Male and remaining are female
- Top ten count of users age spanning from 21 through 30 years old.
- Majority of customers are 26 years old followed by 27.
- Highest number of male users fall in M23-26 age group.
- Highest number of female users fall in F33-42 age group.
- The most loved phone brand is Xiaomi regardless of the age, gender and state the user belong to.
- This signifies that Xiaomi phone's features, cost, durability, etc is accepted by wide customers.
- Customized offerings targeting Male users will be profitable; of course, this has to be analysed further with the support of sales data.
- Events count are more during 22nd hour of the day.
- Male users are more compared to the female users throughout the day.
- Mobile usage picks up from 18:00 hours and keeps increasing till 22:00 hour and drops the following hours.
- Minimum events are observed during 04:00 hour.

8. Tools

- **DS tools**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

In this analysis, Jupyter Notebook was used throughout for data cleaning and transformation, numerical simulation and data visualization.

9. Conclusion

Mobile phone has become an integral part of human life. With the increased digital consumer space, it is imperative that the service providers are on top of customer's necessities and move them in the right direction or the right merchant. This not only delight the customer but also service providers by enjoying their benefits. The analysis has revealed a lot of interesting facts about the mobile phone users. 26.7% of the market is dominated by Xiaomi phone brand regardless of user's demography and the region. User activity is high during 18:00 - 22:00. Male customers are the major users across the analysed regions. TamilNadu and UttarPradesh top the list with the largest number of users. INSAIDTelecom may customize their offerings by introducing special packages for evening. It can may be increase/decrease in the call cost, data packs and SMS charges.