CAPSTONE PROJECT

CONSULTING REPORT

SUBMITED BY: - Veena Srivatsa For CAPSTONE - Group 1020

**Table of Contents**

[**STUDY OF INDIAN MOBILE USER DEMOGRAPHICS USING DATA ANALYTICS** 2](#_Toc35297994)

[**FOR INSAID TELECOM** 2](#_Toc35297995)

[**1.** **INTRODUCTION** 2](#_Toc35297996)

[**1.1** **PROJECT DESCRIPTION** 2](#_Toc35297997)

[**1.2** **OBJECTIVE/PROBLEM STATEMENT** 2](#_Toc35297998)

[**1.3** **SOURCE OF DATASET:** 2](#_Toc35297999)

[**1.4** **SOURCIG THE DATASET** 2](#_Toc35298000)

[**2** **DATASET CHALLENGES** 6](#_Toc35298001)

[**2.1** **DATA CLEANSING**: 6](#_Toc35298002)

[**2.2** **JOINING DATAFRAMES** 7](#_Toc35298003)

[**2.3** **UPDATE MISSING PARAMETERS** 8](#_Toc35298004)

[**2.4** **DROP DUPLICATE ‘device\_id’** 8](#_Toc35298005)

[**2.5** **SELECT DATA CORRESPONDING TO STATES TO BE ANALYSED** 9](#_Toc35298006)

[**3** **DATA ANALYSIS W.R.T SPECIFIED STATES** 9](#_Toc35298007)

[**3.1** **CUSTOMER DISTRIBUTION WITH GENDER ACROSS THE 7 STATES:** 9](#_Toc35298008)

[**3.2** **ORDER OF CUSTOMER DENSITY ACROSS THE 7 INDIAN STATES** 10](#_Toc35298009)

[**3.3** **CUSTOMER GENDER WITH AGE\_GROUP DISTRIBUTION ACROSS 7 STATES** 10](#_Toc35298010)

[**3.4** **DISTRIBUTION OF USERS ACROSS PHONE- BRANDS** 10](#_Toc35298011)

[**3.5** **PHONE BRANDS AND THEIR MODELS** 11](#_Toc35298012)

[**3.6** **PHONE USER GENDER DISTRIBUTION** 11](#_Toc35298013)

[**3.7** **HOURLY DISTRIBUTION OF PHONE EVENTS** 11](#_Toc35298014)

[**4** **STATEWISE DATA VISUALIZATION** 12](#_Toc35298015)

[**4.1** **STATEWISE DISTRIBUTION OF PHONE-BRANDS Vs AGE-GROUP:** 12](#_Toc35298016)

[4.1.1 UTTAR PRADESH: 12](#_Toc35298017)

[4.1.2 RAJASTHAN: 13](#_Toc35298018)

[4.1.3 HARYANA: 13](#_Toc35298019)

[4.1.4 JHARKHAND: 14](#_Toc35298020)

[4.1.5 ORISSA: 14](#_Toc35298021)

[4.1.6 ASSAM: 15](#_Toc35298022)

[4.1.7 TELANGANA: 15](#_Toc35298023)

[**5** **CONCLUSION** 16](#_Toc35298024)

[**6** **RECOMMENDATION** 16](#_Toc35298025)

# **STUDY OF INDIAN MOBILE USER DEMOGRAPHICS USING DATA ANALYTICS**

# **FOR INSAID TELECOM**

# **INTRODUCTION**

## **PROJECT DESCRIPTION**

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 behavioral 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

## **objective/PROBLEM STATEMENT**

* To build a dashboard to understand user's demographic characteristics based on their mobile usage, geolocation, and mobile device properties.
* To help millions of developers and brand advertisers around the world to pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.
* States to be considered for analysis are Rajasthan, Uttar Pradesh, Uttarakhand, Haryana, Orissa, Assam and Telangana

## **Source of Dataset:**

The Data is collected from mobile apps that use Insaid Telecom 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, details on how the gender and age data was obtained

has not been provided.

## **sourcing the datasets**

The datasets ‘**events\_data’ , ‘gender\_age\_train’** & ‘**phone\_brand\_device\_model’** had to be downloaded onto Python by connecting to the MySQL instance provided.

MySQL is an open-source database and one of the best type of relational database (RDBMS) with Structured Query Language - a set of commands used to store and retrieve information using a relational database. This is called relational database because all the huge volumes of data is stored into different tables and relations are established using primary keys or other keys known as Foreign Keys.

A Relational Database Management System (RDBMS) is a software that −

* Enables you to implement a database with tables, columns and indexes.
* Guarantees the Referential Integrity between rows of various tables.
* Updates the indexes automatically.
* Interprets an SQL query and combines information from various tables.

**mysql\_connect()** function is used to open a database connection

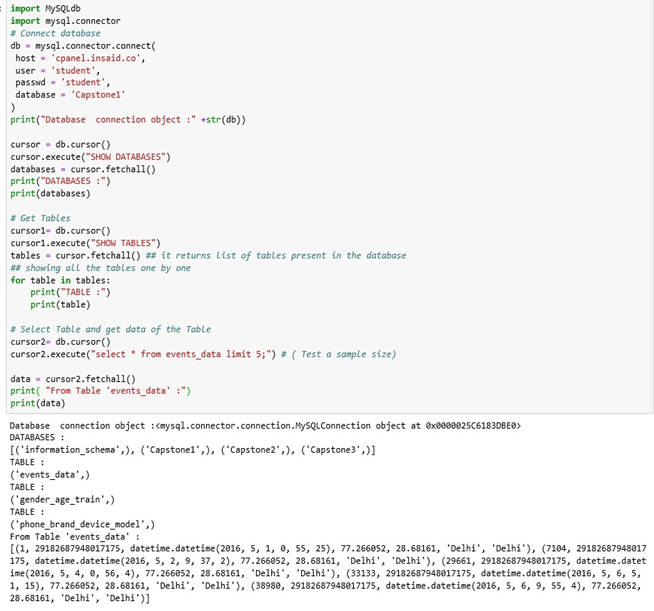


Figure 1: SQLserver Connection, databases, tables

The screenshot( Figure1) above shows the commands used to connect to the SQL server, the

available Databases and the Tables with a brief sample of data from Table “events\_data’.

Further to this, upon querying, the number of rows in each of the required Tables was found as

below:

Number of Rows in Table events\_data : 3252950

Number of Rows in Table gender\_age\_train: 74645

Number of Rows in Table phone\_brand\_device\_model: 87726

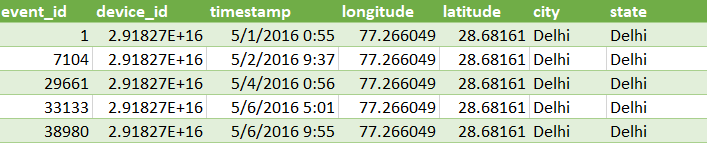


Figure 2 Records & their sizes

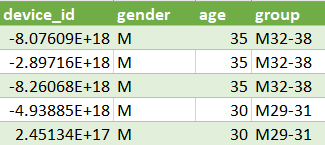
It can be observed that the size of the records is high, and this resulted in server time out errors when downloading of complete file was tried. So, approach of downloading of the big sized file in smaller chunks, followed by joining of files was tried successfully. Sample code for the same is shown in figure(3).

**Representation of data schema is as follows:**

**events\_data** - when a user uses mobile on INSAID Telecom network, the event gets logged in this data. Each event has an event id, location (lat/long), and the event corresponds to frequency of mobile usage timestamp, when the user is using the mobile.



**gender\_age\_train** - Devices and their respective user gender, age and age\_group



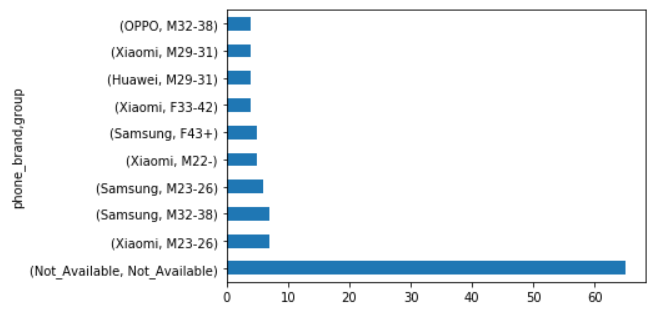
**phone\_brand\_device\_model** - device ids, brand, and models with phone\_brand: note that few

brands are in Chinese



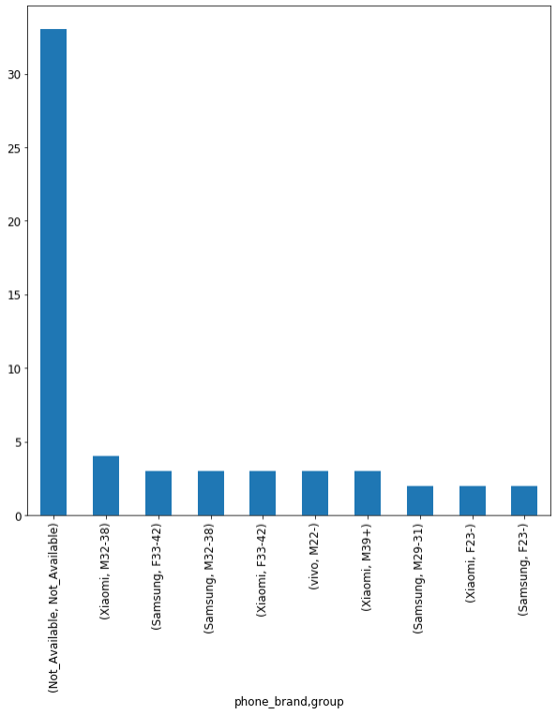
|  |  |
| --- | --- |
|  |  |
|  |  |
| Figure 3 Sample code to download dataset in chunks **DATASET CHALLENGES**   It is observed that parameter ‘device\_id’ is present in each of the dataset available and can be used as common index to combine all the three. This requires that parameter to be in the same format across all the three datasets available and hence converted to ‘float’ and later to int64 for uploading. **Data Cleansing**:   **Problem:**  The dataframe containing event\_data is found to have nulls at device\_id, state, longitudes and  latitudes. However, there is no null at ‘city’.  **Observation 1:**  Where longitudes and latitudes are NaN, but city and state for the same are appropriate:    Where all parameters including longitudes and latitudes are appropriate for the same city and state:    **Strategy1:** To update the nulls at state, longitude and latitude with the corresponding non-null values from the same matching city  **Observation 2:**  The device\_id is like a ‘User\_ID’ and is the same for a pair of latitudes and longitudes. |  |
| **Strategy2:** To update the nulls at ‘device\_id’ with non-nulls of ‘device\_id’, based on matching longitude and latitudes. **JOINING DATAFRAMES** Join all the three dataframes keeping ‘device\_id’ , the only common parameter across all the three dataframes, as the index. |  |
|  |  |
| **UPDATE MISSING PARAMETERS** Check for Nulls in the integrated dataframe and fill them appropriately as there is no common device\_id available to match the other categorical missing data.  There is the option of updating the missing categorical data by imputing the same using :  from sklearn\_pandas import CategoricalImputer using strategy = 'most\_frequent' for 'device\_model' 'phone\_brand','group' and 'gender'  However, this appears incorrect as this would lead to leveraging a particular popular brand and hence these parameters are updated as ‘Not\_Available’   **DROP DUPLICATE ‘device\_id’** There are multiple entries of the same ‘device\_id’ corresponding to multiple transactions/events with the phone-App handling.  As the parameter ‘device\_id’ is like the ‘User-ID’, keeping multiple entries of the same will be leading to improper analysis w.r.t phone user density.  Number of Records before dropping the duplicates of ‘device\_id’:    Number of Records after dropping the duplicates of ‘device\_id’ and filtering required states:   **SELECT DATA CORRESPONDING TO STATES TO BE ANALYSED**  **DATA ANALYSIS W.R.T SPECIFIED STATES****CUSTOMER DISTRIBUTION WITH GENDER ACROSS THE 7 STATES:** |  |
| **ORDER OF CUSTOMER DENSITY ACROSS THE 7 INDIAN STATES**   Telangana has the highest number of phone using customers, even though Hyderabad is the only city considered in the entire state **Customer gender with AGE\_GROUP distribution across 7 states**   Highest gender and age group belongs to male in the age range of 23 -26 years, followed by older men of 39+ years, while female age group is in the range of 33 to 42 years, followed by young women of age 23 years.. **Distribution of Users across Phone Brands** ConsiderING only 10 Most used Phone brands:  The above indicates that Xiaomi tops the list among phone users, followed by Samsung at number 2 and Huawei at 3. **PHONE BRANDS AND THEIR MODELS**   Xiaomi has the highest number of models in the field followed by Samsung and Huawei **PHONE USER GENDER DISTRIBUTION**   The graph above indicates higher usage of mobile phones by males. Also, there is a high population of users whose gender details are not available. **HOURLY DISTRIBUTION OF PHONE EVENTS** |  |
|  |  |
| The hourly distribution of phone events shows a behavioral pattern of growing activity during the day and sloping downwards towards nightfall. **STATEWISE DATA VISUALIZATION****Statewise Distribution of Phone Brands Vs AGE-GROUP:****Uttar Pradesh:**   The above hart indicates that Samsung and Xiaomi are the popular brands among young customers mostly male . Huawei has captured 22-year-old boys. Data is unavailable for the phone brand as well as the age group of users that falls under the highest users’ category  in Uttar Pradesh |  |

### **RAJASTHAN:**



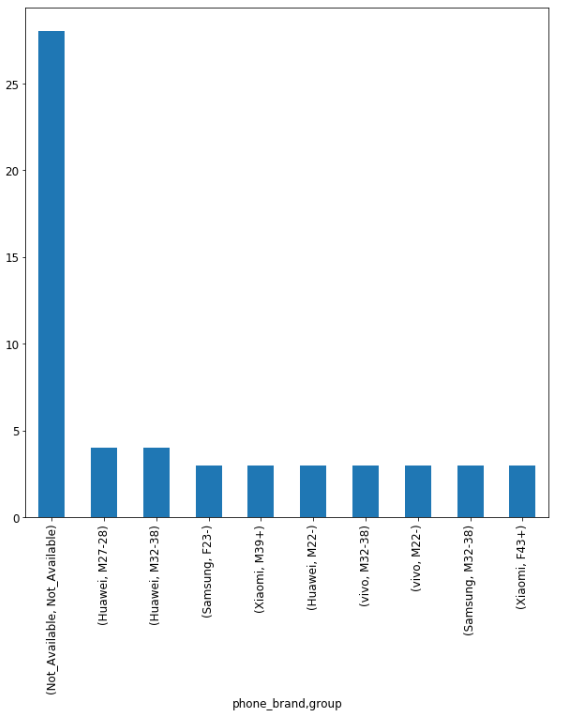
The above hart indicates that Samsung and Xiaomi are the popular brands among young customers mostly male . Huawei is used by men in the age group of 29-31 years. Data is unavailable for the phone brand as well as the age group of users that falls under the highest users’ category in Rajasthan. Rajasthan also has women Xiaomi phone users in the age group 33-42 and Samsung users aged over 43 years. Oppo also has considerable its presence with young men in Rajasthan

### **Haryana**:



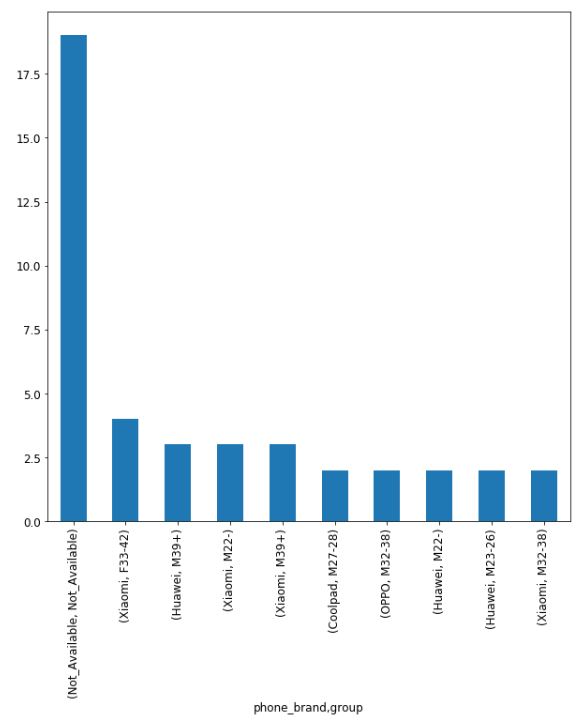
Dominated by Xiaomi and Samsung with Vivo attracting 22-year-old youngsters.

### **Jharkhand**



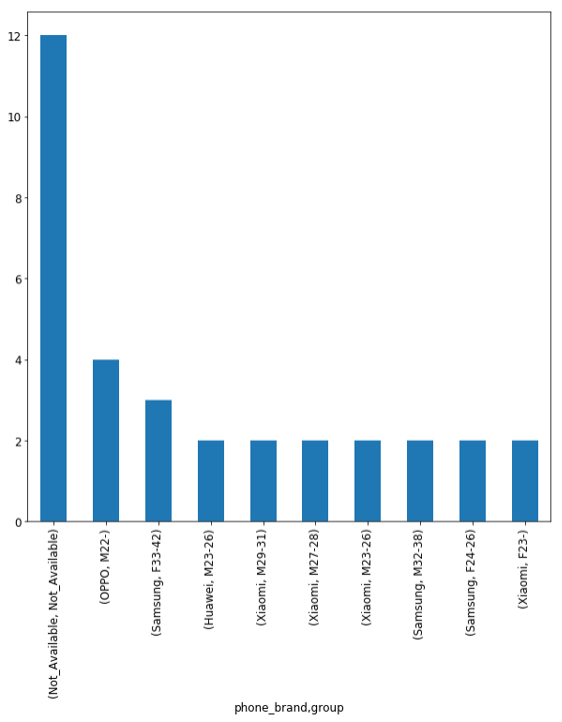
Huawei, Samsung , Xiaomi and Vivo appear to share available customer share almost equally in the youth age group.

### **ORISSA:**



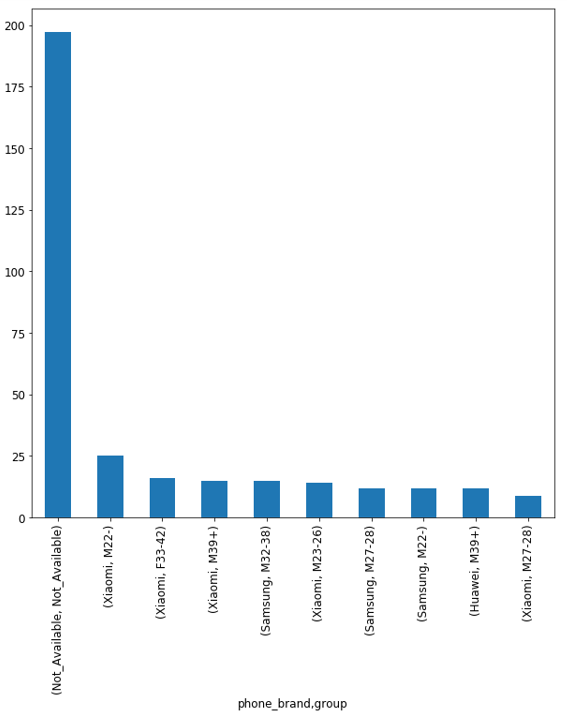
Key players are Xiaomi followed by Huawei, with Coolpad ad Oppo also trying to cross the entry barrier

### **ASSAM:**



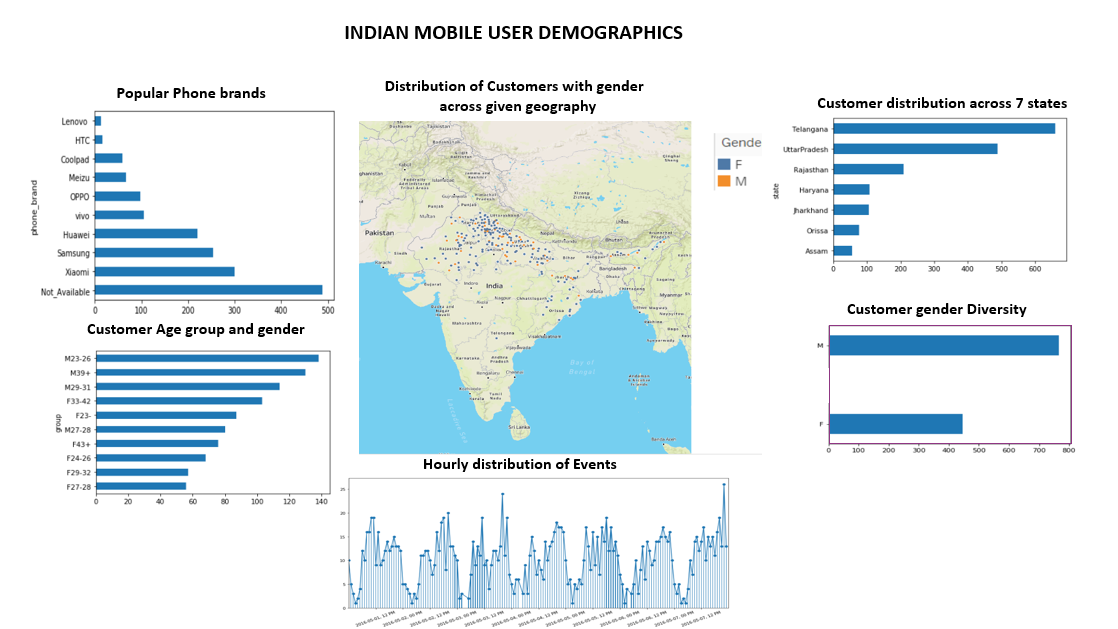
Oppo seems to have managed to get a foothold in Assam having entered the battlefield with strong opponents like Xiaomi, Samsung and Huawei.

### **TELANGANA**

****

Telangana has the highest number of customers coming from a single state, in the dataset, even though data of only one city – Hyderabad is available.

# **CONCLUSION**



# **RECOMMENDATION**

From analyzing the data available for the given 7 states of India, it is evident that phone brands Xiaomi, Samsung and Huawei are ruling the masses and barrier for entry for players like Vivo and Oppo is difficult. However, Oppo has managed its entry into Assam and Orissa, along with Coolpad and Vivo is surfacing in Jharkhand.

Most population belongs to men of age group 22 to 39+ years.

Xiaomi has made a strong presence with its brand being equated to low-cost high-end phone.

It is recommended for the top brands to continue their marketing aiming at young enthusiasts

with their core selling point of providing high configuration and high cost performance.

Companies should also explore expanding their customer base across many more cities, especially in the IT state of Telangana of which only Hyderabad has been considered, to target the young techies.

Data gathering needs improvement to enable better insights by avoiding missing details, unlike in the case here. Pricing details need to be included along with features of models for further analysis.

Companies need to target women with attractive discount, conduct surveys and find out their interests /tastes and promote appropriate Apps and features to them.