## Introduction

As data volumes growing, it requires more applications to scale out to large clusters. In both commercial and scientific fields, new data sources and devices (e.g. RFID, gene sequencers and the web) are producing rapidly growing amounts of information. Telecoms industry like Reliance Jio carried with 16,000 terabytes (TB) per day of traffic, social media like Facebook users send an average of 31.25 million messages and view 2.77 million videos every minute. Google alone processes 40,000 search results per second. Unluckily, the processing and Input output power of single machines have not kept up with this growth. As a consequence, more companies and organizations have to scale out their computations over the clusters. The cluster environment comes with various challenges for programmability. At first, Google's MapReduce presented a simple and general model for batch processing that automatically handles faults. This MapReduce model is used in very famous distributed framework know as Apache Hadoop. There are also other distributed frameworks too, like Apache Spark, Apache Flume etc.

Big data analytics is the use of advanced analytic techniques against very diverse, large datasets that comprise different types such as unstructured/structured and batch/streaming and with different sizes from terabyte to zettabyte. Big data is a term applied to dataset whose type or size is beyond the ability of traditional relational databases to catch, manage, and process the data with low latency. And it has one or more of the following characteristics high volume, high velocity, or high variety. Big data come from various sources, like sensors, devices, video or audio, networks, log files, transactional applications, the web, and social media much of it generated in real time and in a very big scale. Analysing big data allows researchers, analysts and business users to take better and quicker decisions using data that was previously unreachable or unusable. With the help of advanced analytics techniques such as machine learning, prescriptive analytics, data mining, statistics, and natural language processing, businesses can analyse earlier untapped data sources independently or together with their existing enterprise data to obtain new insights resulting in significantly reliable and quicker decisions.

The cumulating amount of data has pressured researchers and practitioners to devise new techniques and data processing models to tap into the invaluable source of Big Data. One such usage in extracting knowledge from the large amount of data is in Prescriptive Analytics which is used in optimization and simulation algorithms to advice on possible outcomes and answer: "What should we do?. Within the context of Data Mining, Prescriptive Analytics pairs with statistical analysis to provide a very interesting combination of techniques for knowledge discovery.

## Overview and Problem Statement

The study concentrates on finding the consumption of different telecom services (SMS, Call, and Internet) by analyzing the very large unstructured data generated over the city of Milano by the computation of Call Detail Record (CDR) and based on the Outcomes obtained, suggest the best suited prescription.

## Motivation

Businesses and research era take great interest in furthering the use of the Prescriptive analytics in enriching the Business Intelligence by forecast ability across a wide range of applications. As data is growing so fast every day, analysis of big data has become a problem for traditional analysis technique. Data generated from various resources is tremendous in volume and highly unstructured in nature, it is thus important to structure the data and leverage its actual potential. This requires a need for new technologies and frameworks to aid humans in automatically and intelligently analyzing large datasets to acquire useful information. The telecoms service providers are required in estimating the various trends in order to plan future upgrades and deployments driven by real data.

As data volumes growing, it requires more applications to scale out to large clusters. In both commercial and scientific fields, new data sources and devices (e.g. RFID, gene sequencers and the web) are producing rapidly growing amounts of information. Telecoms industry like Reliance Jio carried with 16,000 terabytes (TB) per day of traffic, social media like Facebook users send an average of 31.25 million messages and view 2.77 million videos every minute. Google alone processes 40,000 search results per second. Unluckily, the processing and Input output power of single machines have not kept up with this growth. As a consequence, more companies and organizations have to scale out their computations over the clusters.
This immense amount of Telecom data is not being used due to unavailability of resources for computing such a large and unstructured data. So, although having the data, it was not been used for any of the goodwill and profit driven decisions of the industry. Hence, the big question in front of every telecom provider was "How we can use this data for the betterment of the organization?"

The Prescriptive analytics carried out in this study will help Telecom Industry to get the insights of whole data collected over the city of Milano, specifically will get the Region wise analysis - which part of city having maximum consumption of services and which one is having minimum. Besides, The Time-Series analysis will help the industry to get the insights for Day wise analysis-which day of month is having good consumption and similarly, the Hour wise analysis-what time of the day has most and least consumption of different services. Based on all the outcomes obtained, the best suited prescription has been suggested so as to overall improve the profit of the organization.

## Contributions of this work

As already been discussed, Businesses and research era take great interest in furthering the use of the Prescriptive analytics in enriching the Business Intelligence by forecast ability across a wide range of applications. As data is growing so fast every day, analysis of big data has become a problem for traditional analysis technique. Data generated from various resources is tremendous in volume and highly unstructured in nature, it is thus important to structure the data and leverage its actual potential. This requires a need for new technologies and frameworks to aid humans in automatically and intelligently analyzing large datasets to acquire useful information. The telecoms service providers are required in estimating the various trends in order to plan future upgrades and deployments driven by real data.

All the existing data analysis techniques. Due to some or other constraint, are not able to process this huge amount of data and hence not able to generate the desired outcomes. The constraints may wary from processing capacity of the system to the amount of time required for the processing. As, data is very huge and its volumes is growing rapidly, it requires more applications to scale out to large clusters. In both commercial and scientific fields, new data sources and devices (e.g. RFID, gene sequencers and the web) are producing rapidly growing amounts of information. Telecoms industry like Reliance Jio carried with 16,000 terabytes (TB) per day of traffic, social media like Facebook users send an average of 31.25 million messages and view 2.77 million videos every minute. Google alone processes 40,000 search results per second.

To suffice the need of this huge data, we need a system that should have the capacity to process very huge amount of data. Therefore, in spite of using our traditional data mining techniques using typical Databases, we have to look for the distributed framework that can simultaneously process this large amount of data with very much ease. So, as per our need now we will going to process this huge amount of data using Distributed Framework specifically Apache Hadoop Distributed framework where we are going to divide this data into smaller blocks and then will perform the same set of processing on each of these blocks.

The Prescriptive analytics carried out in this study will help Telecom Industry to get the insights of whole data collected over the city of Milano, specifically will get the Region wise analysis- which part of city having maximum consumption of services and which one is having minimum. Besides, The Time-Series analysis will help the industry to get the insights for Day wise analysis-which day of month is having good consumption and similarly, the Hour wise analysis-what time of the day has most and least consumption of different services. Based on all the outcomes obtained, the best suited prescription has been suggested so as to overall improve the profit of the organization.

# 2 Background and Related Work

## 2.1 2.1 Big Data Analytics in Telecommunication

Hundreds of decisions have made every day by business industries. Most of them are routine decisions, but what if a decision needs to be made is beyond the day to day operations that may shift the plan of a business or even an industry and reshape the world. It has been noted that decision-making ability lacks the sophistication and speed required to make sure competitive advantage. The technology makes potential to pile up the art of instinct and skill with the science of knowledge and analytics. So that decision makers are ready to model opportunities based on particular product and market characteristics like whether or not to grow or shrink a business or to collaborate with competitors to obviously determine opportunities and associated risks. In a business where opportunities are a lot of advanced and complex and a lot of accelerated, technology can still deliver with analytical power for those decisions that basically count by bending the art of leadership in judgment with the science of analytics excellence can build the best and quicker decisions.

## 2.2 3.1.1Need of Big Data Analytics in the Telecoms Industry

Telecoms providers have an interest in estimating varied trends so as to set up future upgrades and deployments driven by real information. As associate degree example, a typical provider would wish to grasp the instruments, like a cell tower, that originates the majority of the calls. Another valuable information point is determining the base stations that function the busiest switch hubs throughout varied times of the day, particularly the daily traffic patterns and also the kinds of calls often created to varied locations. This study correlates base station and cell towers these 2 huge information sources that are generally within the order of many of gigabytes on a daily basis for a moderate-sized cellular network, this contains records for each cellular transaction being made on any cellular network. Basically, the service supplier landscape is ever-changing. There are many additional devices and subscribers than the past as well as much more data on a network. That increases have generated an enormous data revolution that's having a sway on telecoms. The data comes in varied forms: application and device-generated data and consumer data from services offered through service providers, industrial IOT, machine-to-machine data, even the network and service data from devices like routers and switches.

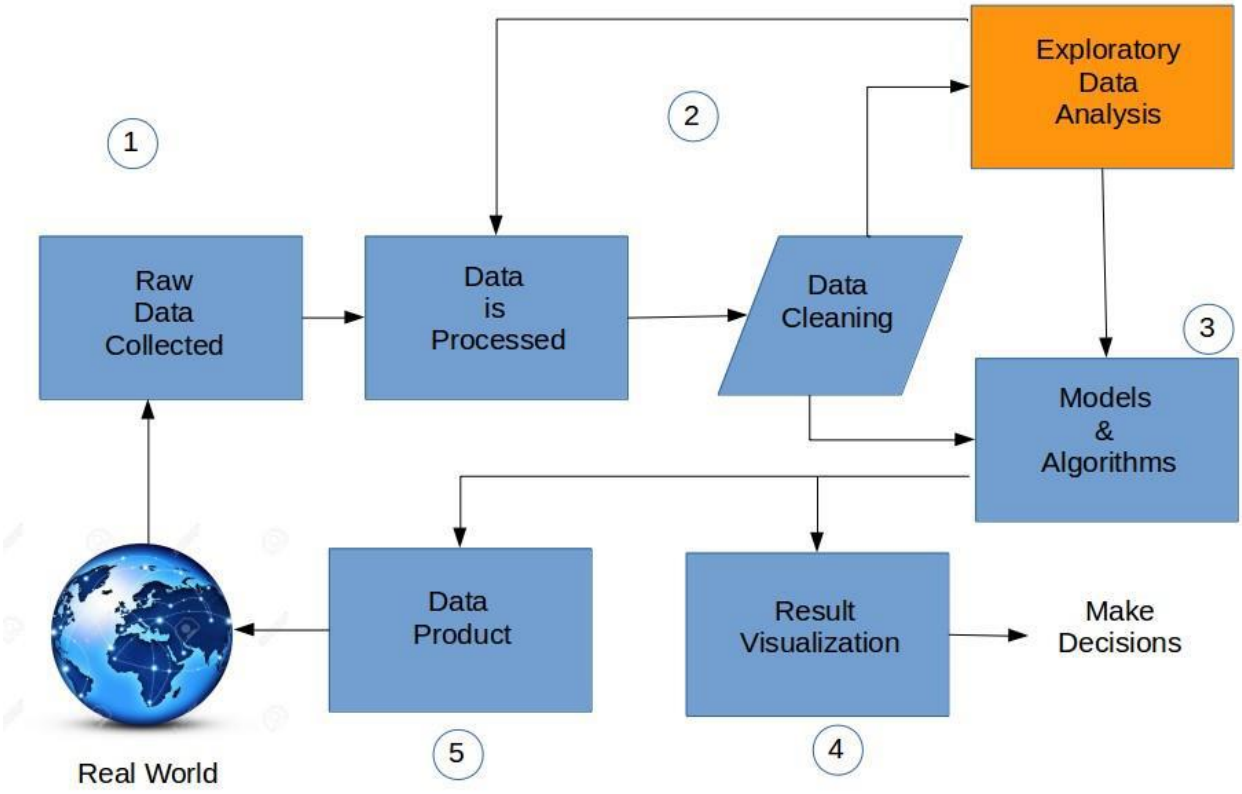



Figure 1: Data Analysis Steps

In this work the pattern detection and trends in telecoms usage of the customers and prediction of the usage for making business decisions is carried out.

### 2.3 Methodology

The following are the steps in data analysis process to achieve the objective of detecting pattern and trends in telecoms usage of the customers and predicting the usage for making business decisions. Figure 3.1 shows the workflow of this research work. It is divided into 5 levels from a work point of view.
The following steps have been applied for every algorithm:
(a) Step-I is collection of data.

(b) Step-II is exploring and preparing the data.

(c) Step-III is training and applying a model on the data.

(d) Step-IV is visualization of result.

(e) Step-V is decisions and result to real world.

# 3 Proposed Method/Algorithm

## 3.1 Problem Definition

Finding the consumption of different telecom services (SMS, Call, and Internet) by analyzing the very large unstructured data generated over the city of Milano by the computation of Call Detail Record (CDR) and based on the Outcomes obtained, suggest the best suited prescription.

## 3.2 Proposed Idea/System

### 3.2.1 Big Data Analytics in Telecommunication

Hundreds of decisions have made every day by business industries. Most of them are routine decisions, but what if a decision needs to be made is beyond the day to day operations that may shift the plan of a business or even an industry and reshape the world. It has been noted that decision-making ability lacks the sophistication and speed required to make sure competitive advantage. The technology makes potential to pile up the art of instinct and skill with the science of knowledge and analytics. So that decision makers are ready to model opportunities based on particular product and market characteristics like whether or not to grow or shrink a business or to collaborate with competitors to obviously determine opportunities and associated risks.

### 3.2.2 Need of Big Data Analytics in the Telecoms Industry

Telecoms providers have an interest in estimating varied trends so as to set up future upgrades and deployments driven by real information. a typical provider would wish to grasp the instruments, like a cell tower, that originates the majority of the calls. Another valuable information point is determining the base stations that function the busiest switch hubs throughout varied times of the day, particularly the daily traffic patterns and also the kinds of calls often created to varied locations. This study correlates base station and cell towers these 2 huge information sources that are generally within the order of many of gigabytes on a daily basis for a moderate-sized cellular network, this contains records for each cellular transaction being made on any cellular network.

Need of Big Data Analytics in Telecom Industry...

- To Understand the potential of new product offerings
- To Improve customer experiences
- To Reduce service truck rolls while improving customer service
- To Forecast network capacity and demand faster and more accurately
- To Implement value-based network capacity planning
- To reduce customer churn.

### 3.2.3 Methodology

The following are the steps in data analysis process to achieve the objective of detecting pattern and trends in telecoms usage of the customers and predicting the usage for making business decisions.

(a) Step-I is collection of data.
(b) Step-II is exploring and preparing the data.
(c) Step-III is training and applying a model on the data.
(d) Step-IV is visualization of result.

(e) Step-V is decisions and result to real world.

### 3.3 System Architecture

#### 3.3.1 Apache Hadoop

Apache Hadoop is a software framework that supports data-intensive distributed applications. It has been used by many big technology companies, such as Facebook, Amazon, Yahoo, and IBM. Apache Hadoop is best known for a computational framework (MapReduce) and its distributed file system (HDFS). The main objective of Apache Hadoop is to focuses on tasks that require all the available data for examination. Apache Hadoop is a scalable framework for processing data and storing on a cluster of commodity hardware system. To scale up from a single node to the thousands of nodes cluster Apache Hadoop is designed. HDFS uses the commodity server nodes and storage drives to store the data. These same set of server nodes are used for computation. This enables scalable and efficient ways of storing and processing data. By adding more servers, capacity of storage, computation along with I/O bandwidth can also be scaled. The high-level components of a Apache Hadoop cluster is showed in Below Fig.

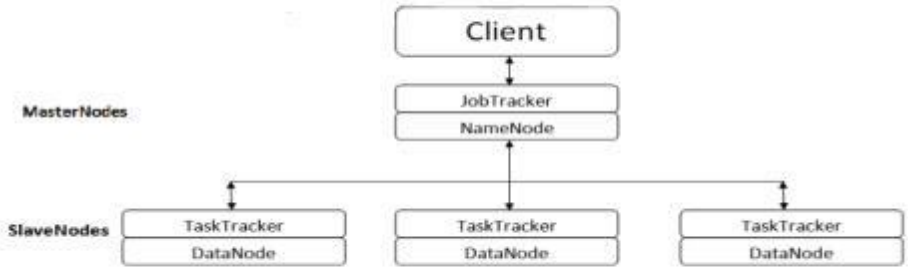



Figure 2: Hadoop Architecture

Above Figure gives simple illustration of the Apache Hadoop cluster architecture. NameNode is a single master server for storing the file system metadata in HDFS. The file stored on HDFS is split into blocks. These blocks are stored on slave nodes referred as DataNodes. For sake of data reliability, multiple replicas of blocks are stored on Data Nodes. At the Name-Node client does an operation such as producing, altering, and erasing files at the file system. The Name-Node monitors the Data-Nodes actively, if in case a replica of a block is lost due to any failure, new replicas will be created. Computation framework supports parallel, distributed programming models over which huge amount of data can be easily processed in a reliable and fault-tolerant way. Typically, the data is processed in parallel by using multiple tasks where each task processes a subset of the data.

#### 3.3.2 Apache Pig

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

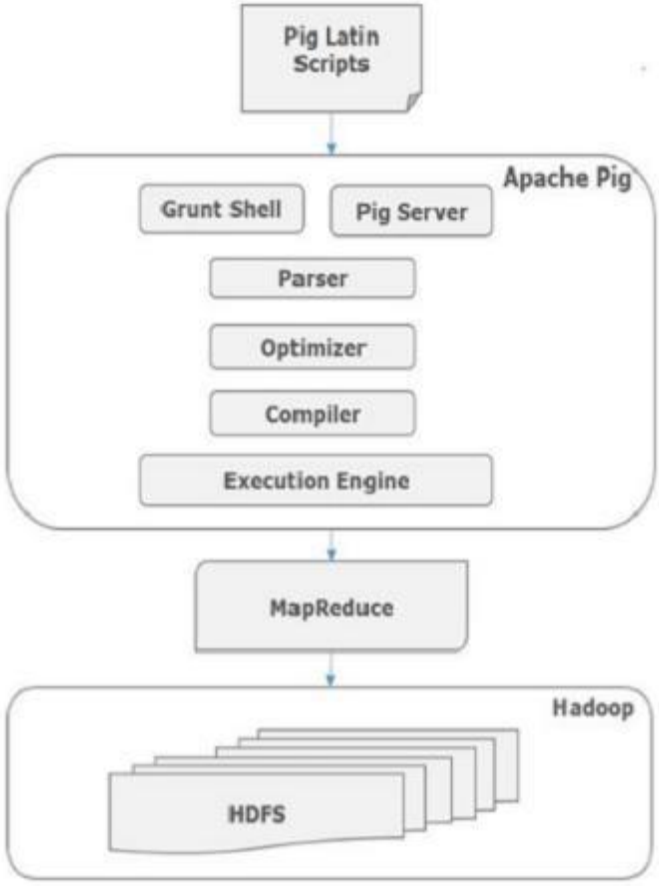
Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist. Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

 **Ease of programming**: It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

 **Optimization opportunities:** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

 **Extensibility**: Users can create their own functions to do special-purpose processing.

It is a high-level data processing language which provides a rich set of data types and operators to perform various operations on the data.
To perform a particular task Programmers using Pig, programmers need to write a Pig script using the Pig Latin language and execute them using any of the execution mechanisms (Grunt Shell). After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.



**Algorithms and Pseudo Code**

**K-Means algorithm**.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

# 4 Performance Study

## 4.1 Implementation/Simulation Environment

### 4.1.1 Data Collection and Description

In this thesis work the open big data is used that is published at the location https://dandelion.eu/datamine/open-big-data/ under Open Data Commons Open Database License (ODbL) license. This dataset provides information regarding the telecommunication activities over the town of Milano. The dataset is that the results of a computation over the Call Detail Records (CDRs) generated by the Telecoms Italian Republic cellular network over the city of Milan (Milano). CDRs log the user activity for charge purposes and network management. There are many varieties of CDRs. For the generation of this dataset, the subsequent activities are considered:

(a) SMS Received: a CDR is generated whenever a user receives an SMS.
(b) SMS Sent: a CDR is generated whenever a user sends an SMS.
(c) Incoming CALL: a CDR is generated whenever a user receives a call.
(d) Outgoing CALL: CDR is generated whenever a user issues a call.
(e) Internet: a CDR is generated every time.
 A user starts an internet connection.
 A user ends an internet connection.
 During the same connection one of the following limits is reached:

15 minutes from the last generated CDR.
5 MB from the last generated CDR.

By aggregating the same records, it had been created this dataset that provides SMSs, calls and Internet usage traffic activity. It measures the level of interaction of the users with the mobile phone network; for example, the higher is the number of SMS sent by the users, the higher is the activity of the sent SMS. Measurements of call and SMS activity have the same scale (therefore are comparable); those referring to Internet traffic do not.

(f) Spatial aggregation: different activity measurements are provided for each square of the Milano GRID.
(g) Temporal aggregation: activity measurements are obtained by temporally aggregating CDRs in timeslots of ten minutes.
(h) Milano GRID: The Milano city is divided into grid of 100 by 100 small squares like shown in figure 4.1.

Where d=235 meter

Squares are numbered with ids. The square id numbering starts from the bottom left corner of the grid and grows till its right top corner. Figure 4.2 shows map of Milano city and figure 4.3 shows the grid that overlaid on the city map.

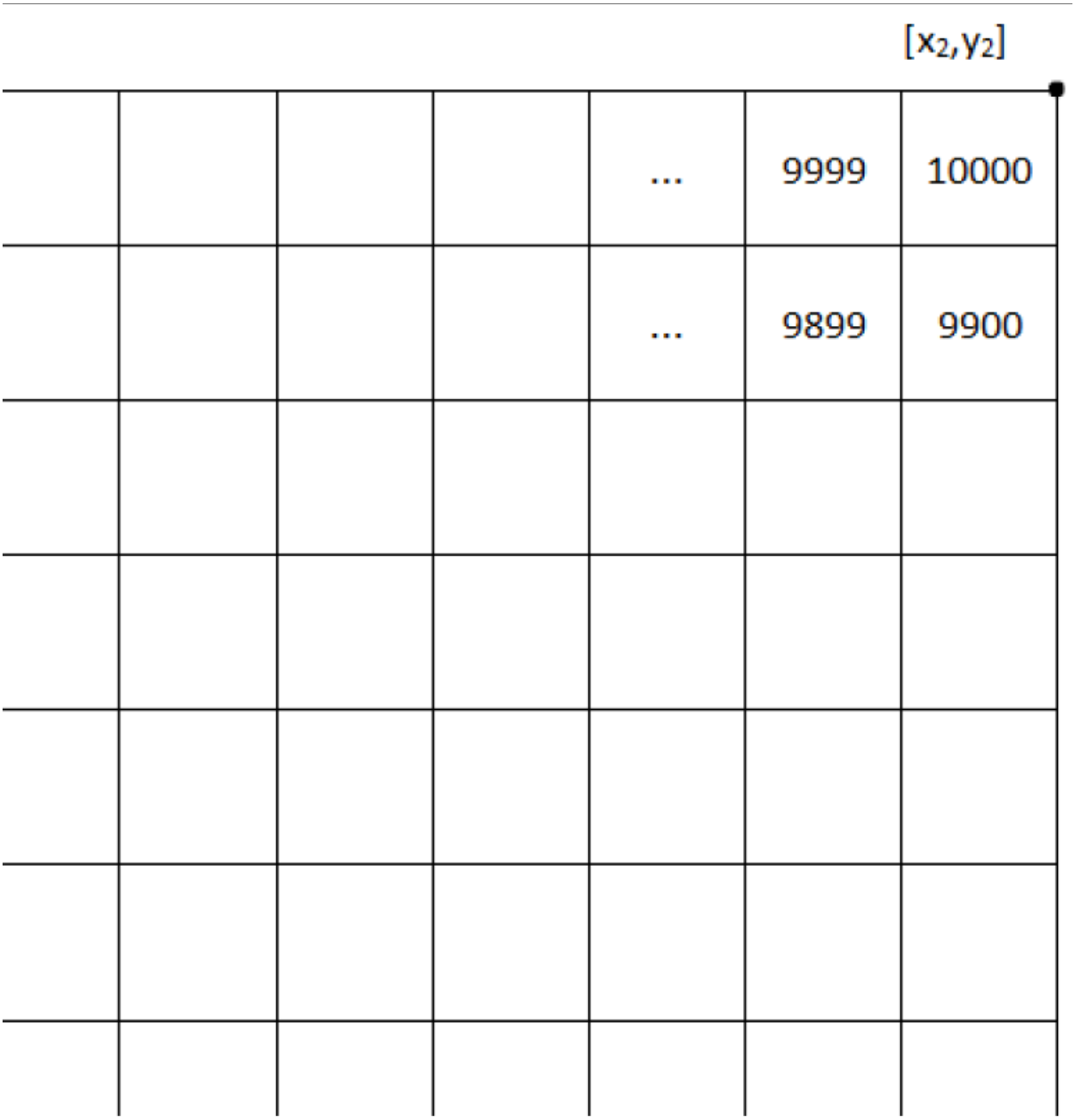


Figure 9: Milano city Structure

The data contain 8 features, those are listed below:

(i) **Square id**: the id of the Sq. that's a part of the Milano city GRID; TYPE: numeric

(ii) **Time interval**: the start of the time interval expressed as the number of milliseconds passed on from the Unix Epoch on January first, 1970 at UTC. The end of the time interval are often obtained by adding 600000 milliseconds (10 minutes) to this value. TYPE: numeric

(iii) **Country code**: the phone country code of a nation. Depending on the measured activity this value assumes completely different meanings that are explained later.
TYPE: numeric

(iv) **SMS-IN** activity: the activity in terms of received SMS-IN side the Sq. id, during the time interval and sent from the nation known by the Country code.
TYPE: numeric

(v) **SMS-OUT** activity: the activity in terms of sent SMS-IN side the Sq. id, during the Time interval and received by the nation known by the Country code.
TYPE: numeric
(vi) **CALL-IN** activity: the activity in terms of received calls inside the Sq. id, during the Time interval and issued from the nation known by the Country code.
TYPE: numeric
(vii) **CALL-OUT** activity: the activity in terms of issued calls within the Sq. id, during the Time interval and received by the nation known by the Country code.
TYPE: numeric
(viii) **Internet traffic** activity: the activity in terms of performed INTERNET traffic within the Sq. id, throughout the Time interval and by the nation of the users performing the connection known by the Country code.
TYPE: numeric

**Exploring and preparing the data**

Data in raw form (e.g., from a warehouse) are not always the best for analysis, and especially not for predictive data mining. The data must be preprocessed or prepared and transformed to get the best mineable form. Data preparation is very important because different predictive data mining Techni ques behave differently depending on the preprocessing and transformational methods. There are many techniques for data preparation that can be used to achieve different data mining goals.

### 4.1.3 Missing value imputation

Data may contain some missing or empty values, missing value imputation is also depending upon how data is collected. Commonly used missing value methods are
(a) Complete Case Analysis
(b) Available Case Analysis
(c) Single value Imputation
(d) Model Based Methods for multivariate Normal missing data
(e) Multivariate Imputation by Chained Equation
There are more methods also. In thesis work single value imputation is used. Most commonly single value imputation technique involves replacing any missing value with the mean of that variable for all other cases, which has the benefit of not changing the sample mean for that variable.

After Pre-processing the data, it can be used for further Analysis. We can use this data for
- Region-wise analysis
- Day-wise analysis
- Hourly analysis

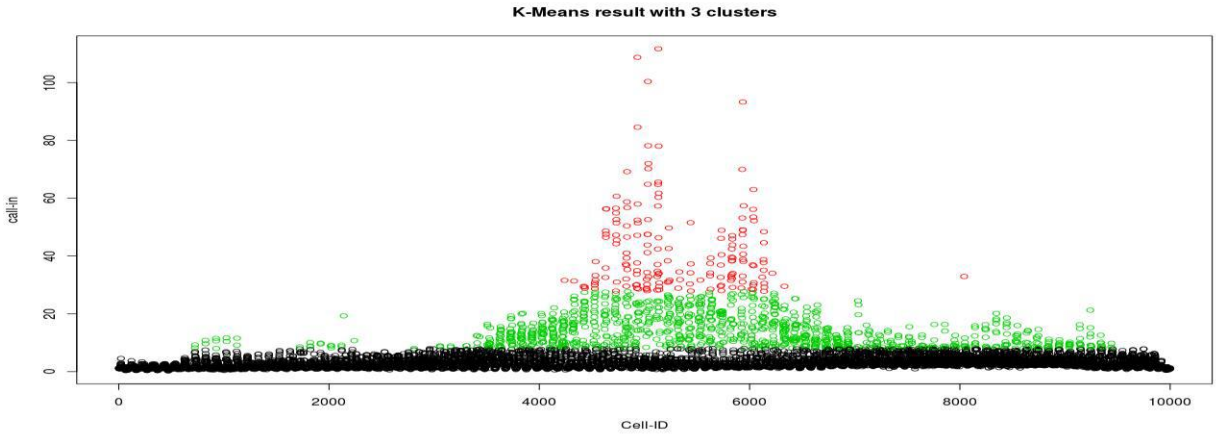
## 4.2 Results and Analysis

### 4.2.1 Region Wise Analysis

**K-means on daily usage of month**

The data for month is already prepared using apache pig. The arrangement of data is for each country Code for whole month on daily basis. For example, for COUNTRY CODE=39 and CELL-

ID=1 will have 1 record. The total records are 10000. After k-means with k=3 the result is shown in figure 4.9.




Above Figure shows that 1.5 % of city region comes under high traffic, 9.5 % comes under medium traffic and 89% comes under low traffic. The scale for figure 4.10 is on X-axis 1 unit= 235 meter and on Y-axis 1 unit= 235 meter. The better visualize in city grid plot. Figure 4.10 shows very good result, the insights are as follow,

(a) Middle of city has high CALL-IN traffic, because crowd is more in middle of city, and generally all commercial business offices, government offices etc. are in middle of city.

(b) Around the Middle of city CALL-IN traffic is low because there is mostly domestic area, houses etc.

(c) Outside of city gets low CALL-IN traffic, because outside have low crowd.

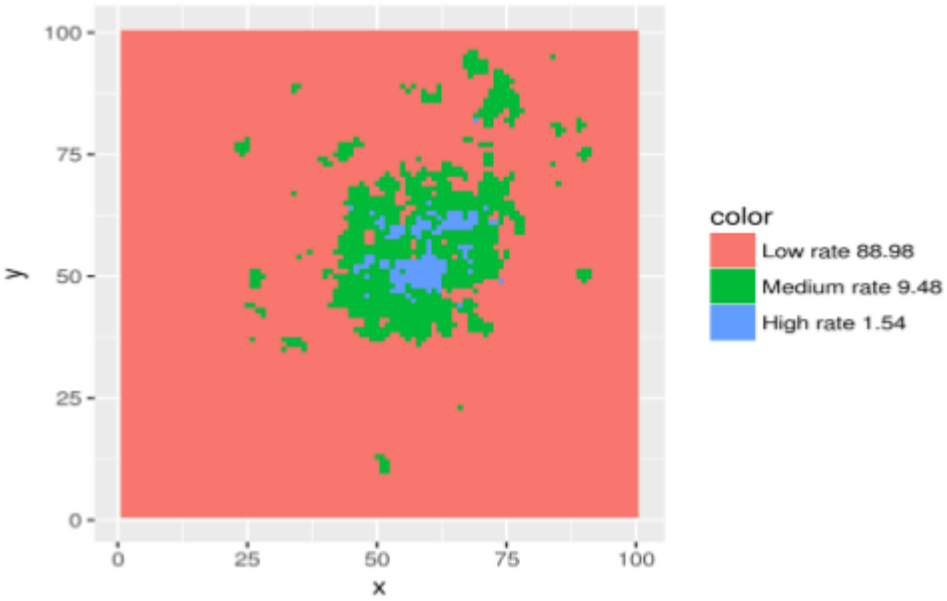
This is the prediction after k-means model. The prescription from that result to respective company will likely,

(a) Telecom company can take appropriate action to reduce traffic in middle of city.

(b) It can alter their service rate charges based on use.

(c) They can alter their offers in respective region to attract new customer etc.

The scale for above plots is, on X-axis 1 unit= 235 meter and on Y-axis 1 unit= 235 meter. From above cluster plot for city, Telecoms service provider can the appropriate actions in particular region. In this work, cluster plot for all Country Code has plotted.




**Day wise Analysis**

To analyze, the data is again pre-processed. Already pre-processed data is present in day-wise form from 1 Nov 2013 to 31 DEC 2013. For daily analysis of activity for whole Milano city, means value of activity has taken for all cells and country codes.

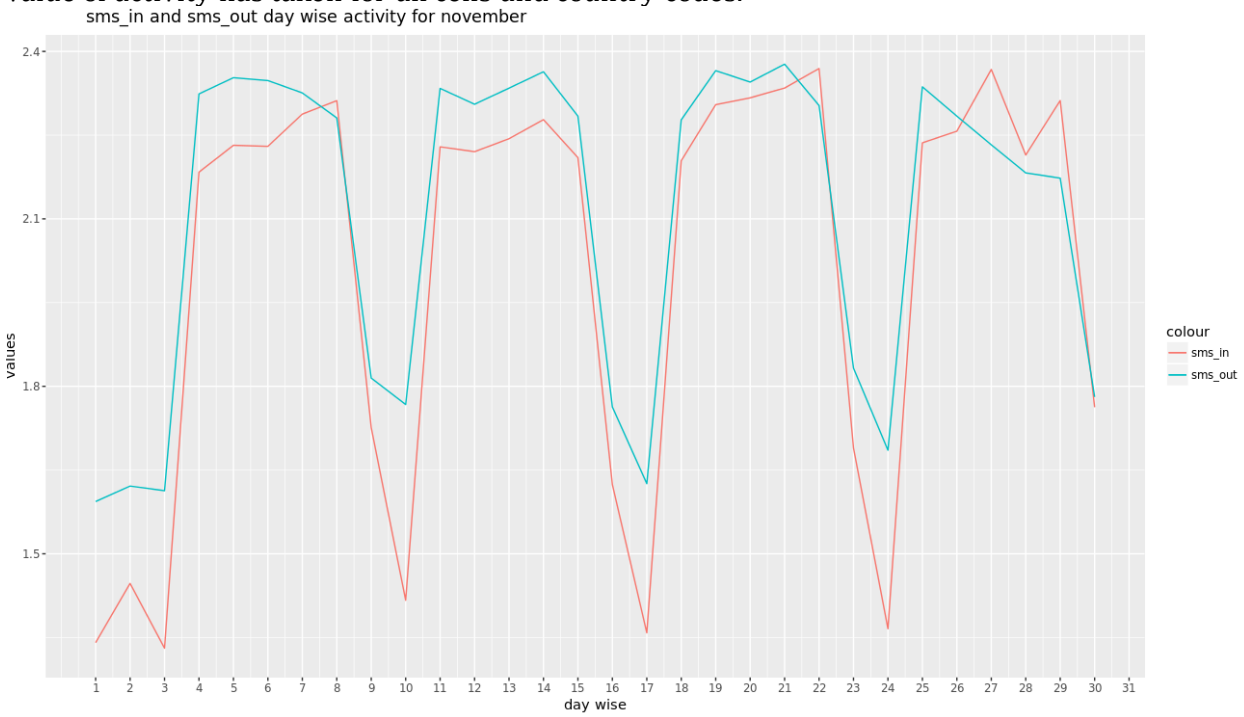



Figure 13: SMS IN & SMS OUT Day Wise Activity for November

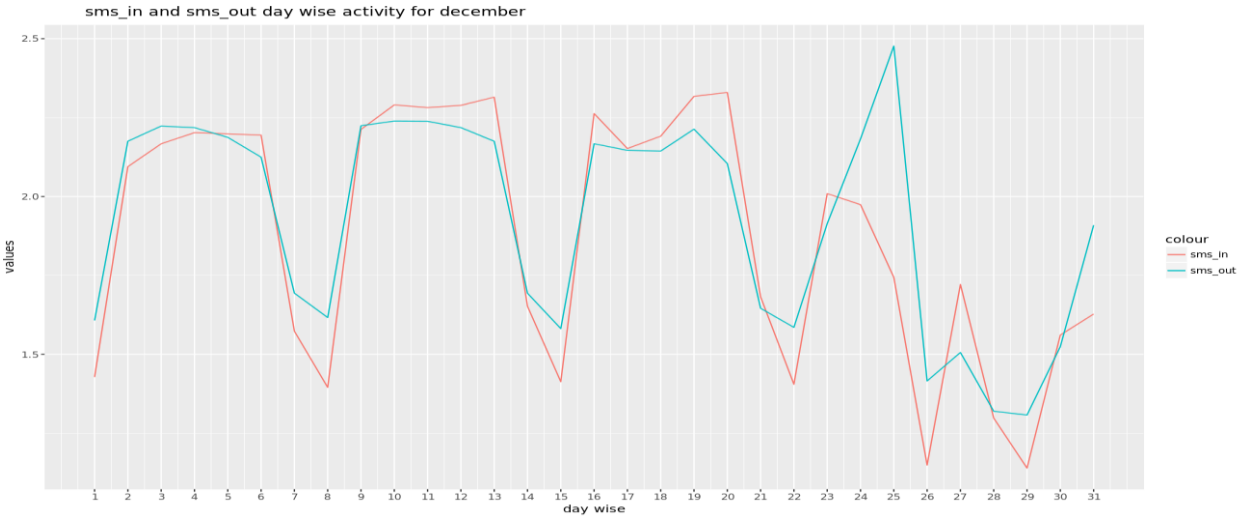



Figure 14: SMS IN & SMS OUT Day Wise Activity for December

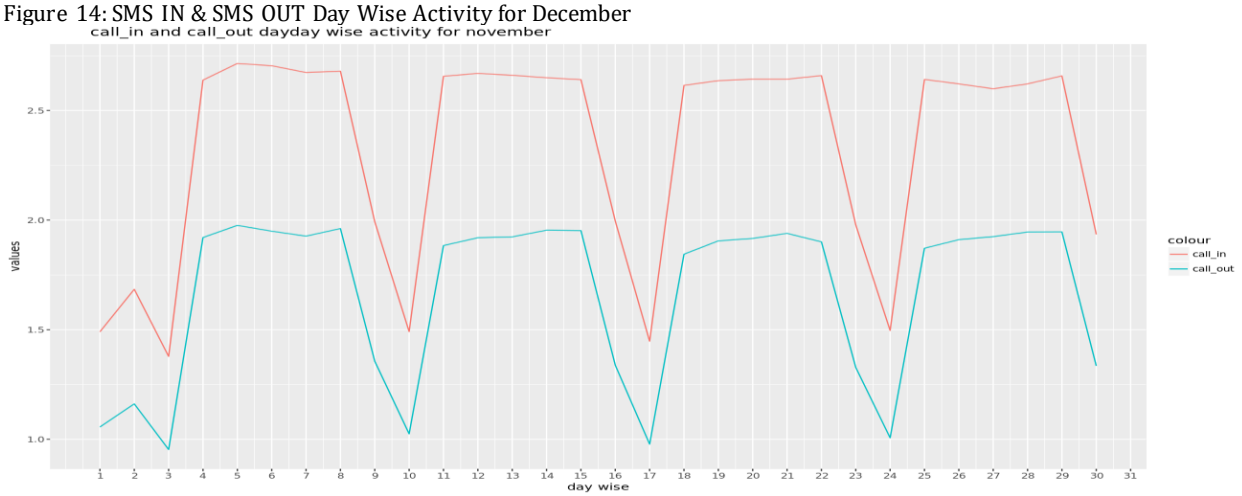



Figure 15: Call In & Call Out Day Wise Activity for November

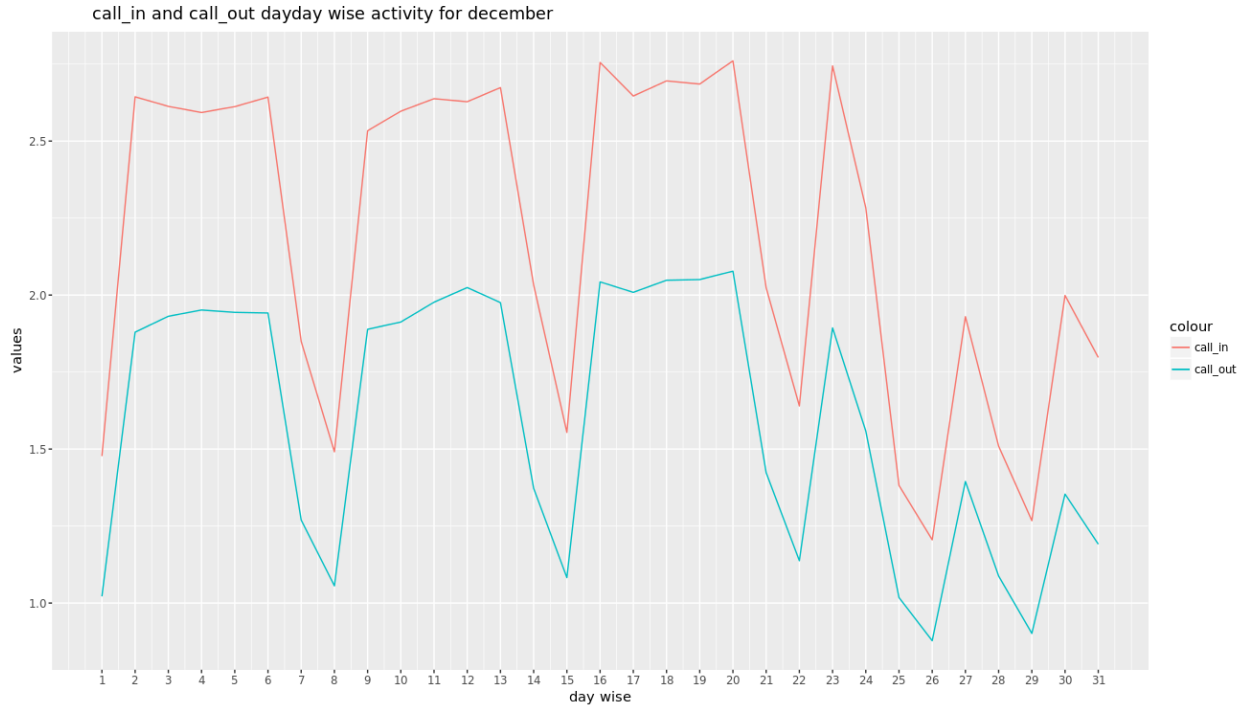



Figure 16: Call In & Call Out Day Wise Activity for December

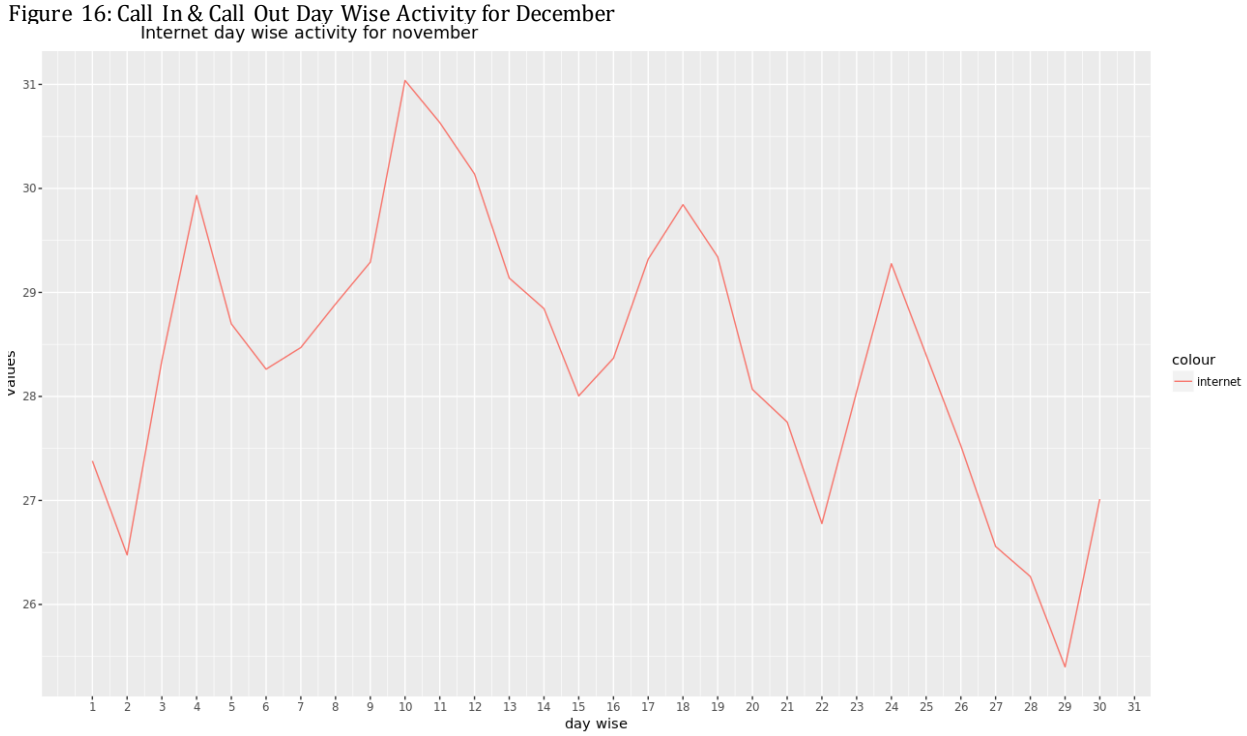



Figure 17: Internet Day Wise Activity for November

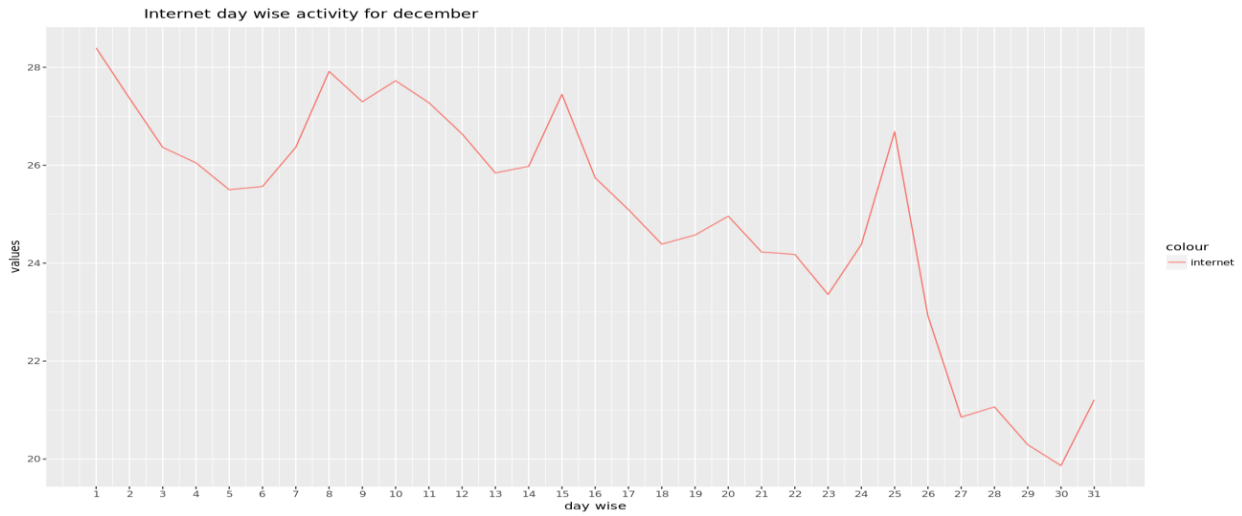



Figure 18: Internet Day Wise Activity for December

## Hourly analysis

Hourly analysis for each cell has been consider separately, it is analyzed that how particular cell region is behaving for activities like SMS, CALL, INTERNET DATA in hours of days. To do this, the data is again required to preprocess according to cell-Id wise for hourly. For this analysis, data of DEC 2013 day is taken, as each day has separate input file. Total cells are 10000, for this analysis only chose 100 and 6000 cells. Prepared the data for CELL-ID 100 and 6000, CELL-ID are randomly selected.

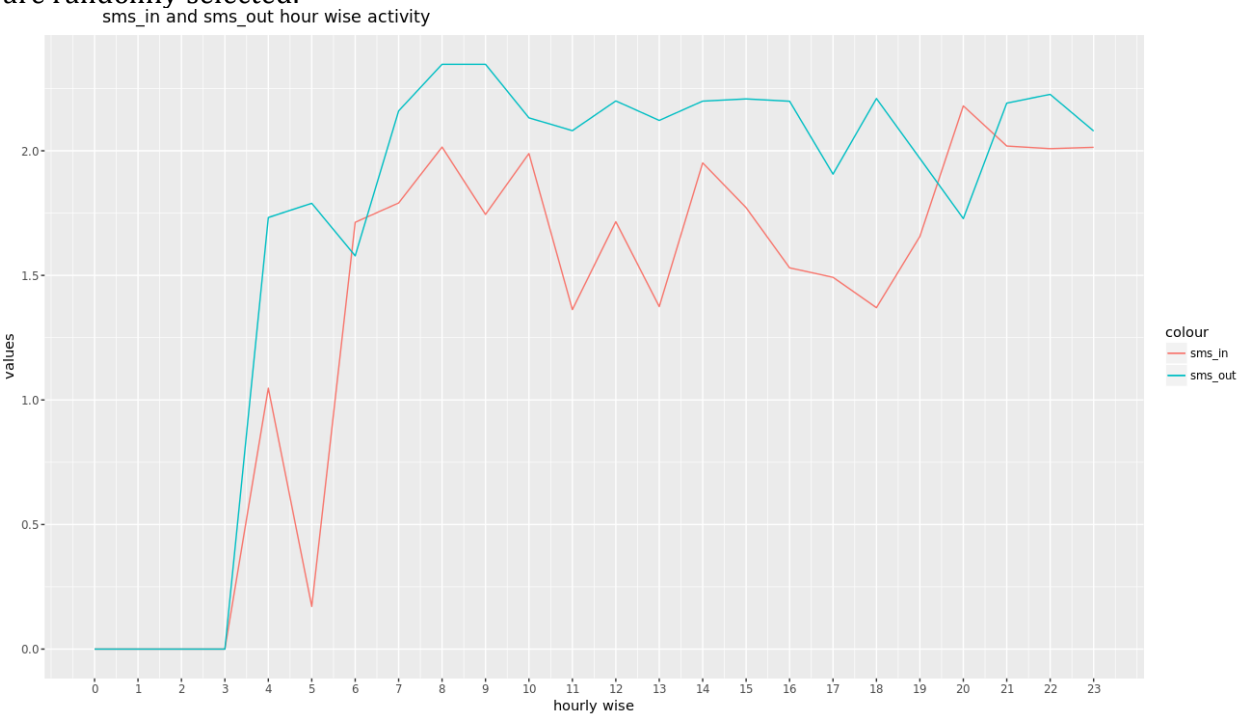



Figure 19: SMS IN & SMS OUT Hour wise Analysis for December

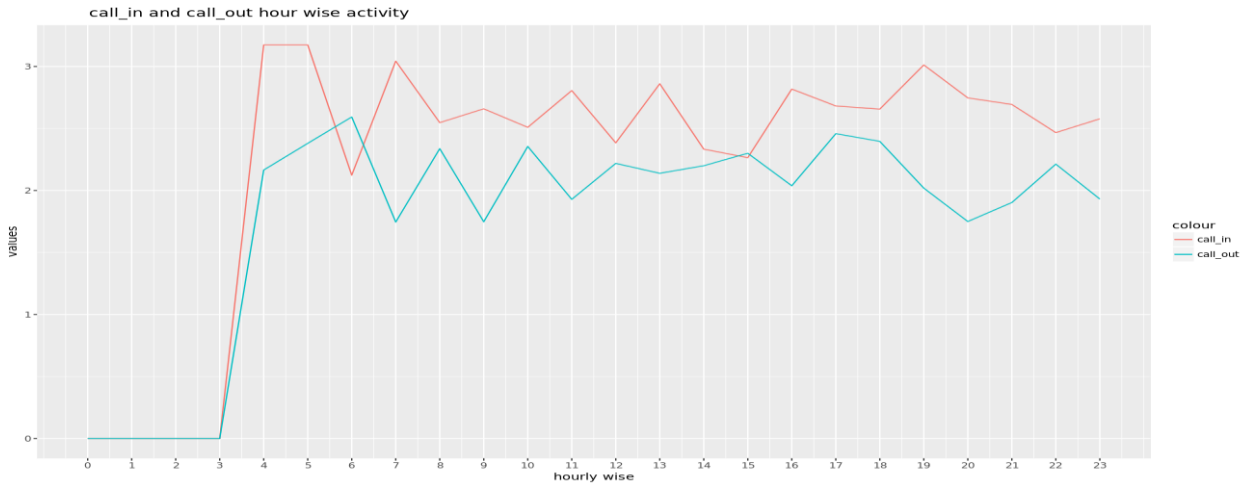



Figure 20: Call IN & Call OUT Hour wise Analysis for December

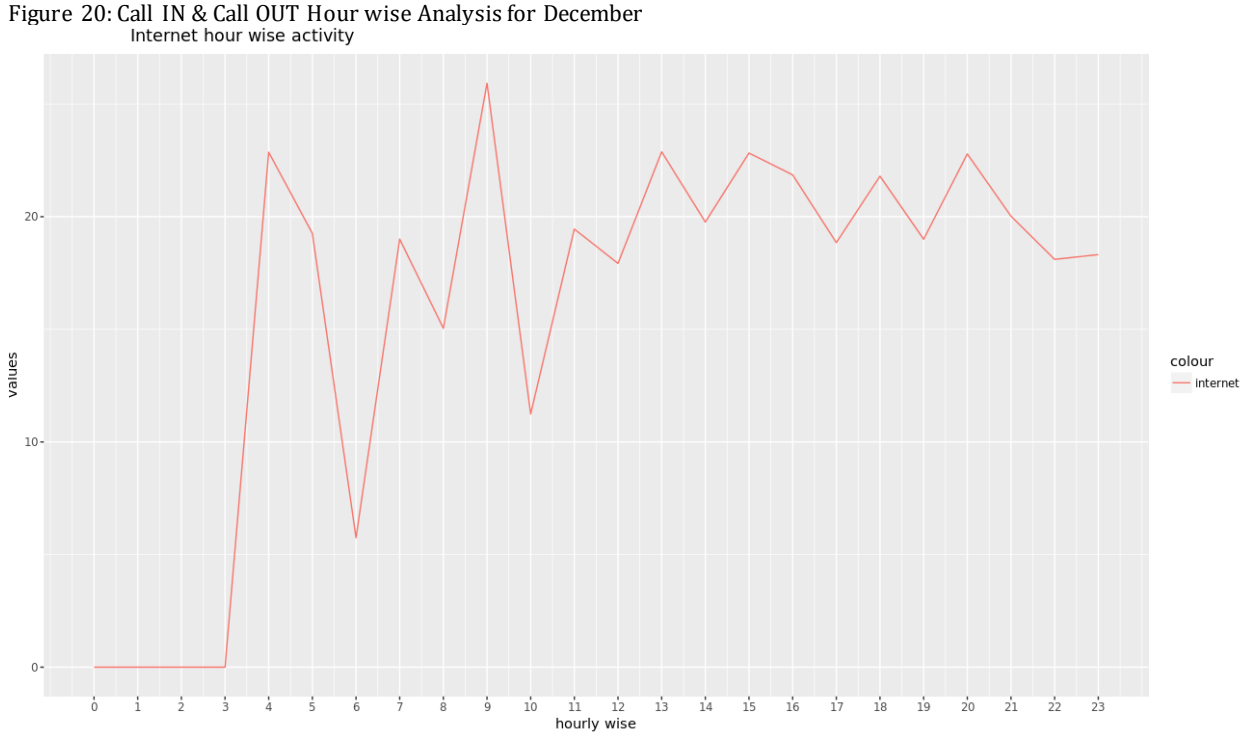



Figure 21: Internet Activity Hour wise Analysis for December

### 4.3 Summary of performance study

The prescriptive analytics has been performed on the data for the following purposes:

1. Region-wise Analysis.
2. Monthly/Day-wise analysis.
3. Hourly analysis.

**Region-wise Analysis**

It has been observed that,

- Middle of city has highest consumption of services.
- Around the city has medium consumption of services
- Outskirts have least consumption of services

**Monthly/Day-wise analysis**
It is been observed that,
 On Sundays (3, 10, 17, 24 November) & (8, 15, 22, 29 December) the consumption of services is very less.
 24 December onwards till 28, the consumption of all services decreases except for SMS-Out and Internet.
 Extreme peak for SMS-Out and slight increase of Internet Usage from 24 to 26 December.

**Hourly analysis**
It has been observed that,
 Negligible usage between 0 to 4.
 **Call_in:**
- High activity between 4 to 5.
- Normal activity between 7 to 19.

 **Internet:**
- Negligible usage between 0 to 4 AM.
- Good usage between 9 to 10.
- Normal usage between 4 to 11.

# 5 Conclusions and Future Work

## 5.1 Conclusion

As we know, Telecom data is growing tremendously every day, but the traditional analytic techniques are not capable of analyzing this huge data due to which the power of big data is not been utilized to its full extent. In this study, the Apache Hadoop distributed framework has been used for processing this huge amount of data and performing the analysis. Furthermore, Apache Pig has been used to improve the efficiency that is to reduce the amount of time required for processing.
The prescriptive analytics has been performed on the data to recommend the best suited prescriptions to the Telecom Industry based on the outcomes obtained after processing. The K-means algorithm has been used for Clustering to get the Region-based prescription on the data. Similarly, the Time-series analysis has been performed on the data to get the Daily and Hourly usage of the different services that is, how particular service is used with respect to time. From all the analysis performed on the data, the Telecom provider can take the decisions to have the maximum utilization of resources and ultimately to increase the profit.

## 5.2 Future work

As Data pre-processing is crucial and one of the important steps in the data analytics, it consumes more than 60% of total work. So, to improve the time efficiency that is the amount of time required for pre-processing the raw data, another framework Apache Spark can be used (Because Apache Spark works better than Apache Hadoop in many cases).
Besides, in future study, we can consider the many years data as a Training set and can use the different Machine Learning algorithms so as to predict what could be the Future pattern for the particular service.