# Abstract

The project is entitled "Segmentation of Prepaid Mobile Subscribers" and was implemented within the Tunisian company Tunisie Telecom, a services company specialized in telecommunication.

This work forms part of the graduation project presented to obtain the Engineer's degree under the specialty data science at the Private Higher School of Engineering and Technology.

The aim of this project is to offer personalized plans adapted to customers' needs through understanding their behaviour. It was realized with the data science and BI teams at Tunisie Telecom Technopole Ghazala, it covers essentially components of data science: machine learning and data visualisations.

**Keywords:** telecommunication services, insight, market segmentation, profiling segments, customer behaviour, machine learning, k-mean clustering, relationship marketing, data analysis, data analytics, data visualisations

# Table of Contents

# Table of Figures

# Table of Tables

# List of Abbreviations and Acronyms

| | |
|---|---|
| AI: Artificial Intelligence | 10 |
| API: Application Programming Interface | 28 |
| BI: Business Intelligence | VII |
| BTS: Base Transceiver Station | 11 |
| CEO: Chief Executive Officer | 3 |
| CRISP-DM: Cross-Industry Standard Process for Data Mining | 5 |
| CRM: Customer Relationship Management | 9 |
| CSP: Communications Service Provider | 1 |
| CSV: Comma-Separated Values | 39 |
| DGD: Deputy General Director | 3 |
| EDA: Exploratory Data Analysis | 15 |
| ICT: Information and Communications Technology | 2 |
| SQL: Structured Query Language | 28 |

# General Introduction

Today, due to the expansion of smartphones, mobile devices, and telecommunications services over the internet, CSPs need to handle massive data. They must process, store and quickly obtain insight from high volumes of data. This data generally travels through wide extensive networks.

Telecommunications companies collect large amounts of data. This data includes mobile phone usage, records, network equipment, server logs, billing, and social media. This data provides a lot of information about their customers and their network.

The content and strategy for each part of our customer demographic in the market is not be the same. So, telecommunications companies generally segment their market and then target the content according to each group.

Segmentation and targeting help in identifying the needs, preferences, and customer's reactions to the telecommunications services and products.

Once the main customer motivators are identified, products can be branded appropriately. The purpose of market segmentation is to reach the target and to improve the brand image of the company. Promoting the product with a well-adjusted brand strategy allows for putting the company ahead of its competitors.

# Chapter 1: General Context and Presentation

## Introduction

In the first chapter, we will present the host organization where we carried out our internship and define the object and the requirements, discuss existing solutions as well as the choice of the methodology for our work processes.

## 1. Presentation of the Company

### 1.1. History of the Company

Tunisie Telecom is the trade name of the national communications office created on April 17, 1995. Since its creation, Tunisie Telecom has been working to consolidate the telecommunication infrastructure in Tunisia, to improve the coverage rate and boost its competitiveness. It also actively contributes to the promotion of the use of ICT and the development of innovative companies in the telecommunications sector.

Figure 1: Tunisie Telecom Logo

A pioneer of the telecommunication sector in Tunisia, Tunisie Telecom has established a defining value that place the customer at the centre of his priorities. The adoption of these values is reflected in particular by a continuous improvement of the standards of the company and the quality of services. (Tunisie Telecom, 2022)

## 1.2. Services of the Company

Tunisie Telecom's mission mainly is to ensure telecommunications services for its clients. In particular, the mission is divided as following:

The installation, development, maintenance and operation of the networks of telephone and data transmission.

Promotion of new telecommunications services.

The offer of all public or private telecommunications services which responds to various social and economic needs.

Participation in the national efforts to improve higher education at the level of the telecommunications sector.

Promotion of cooperation in all areas of telecommunications.

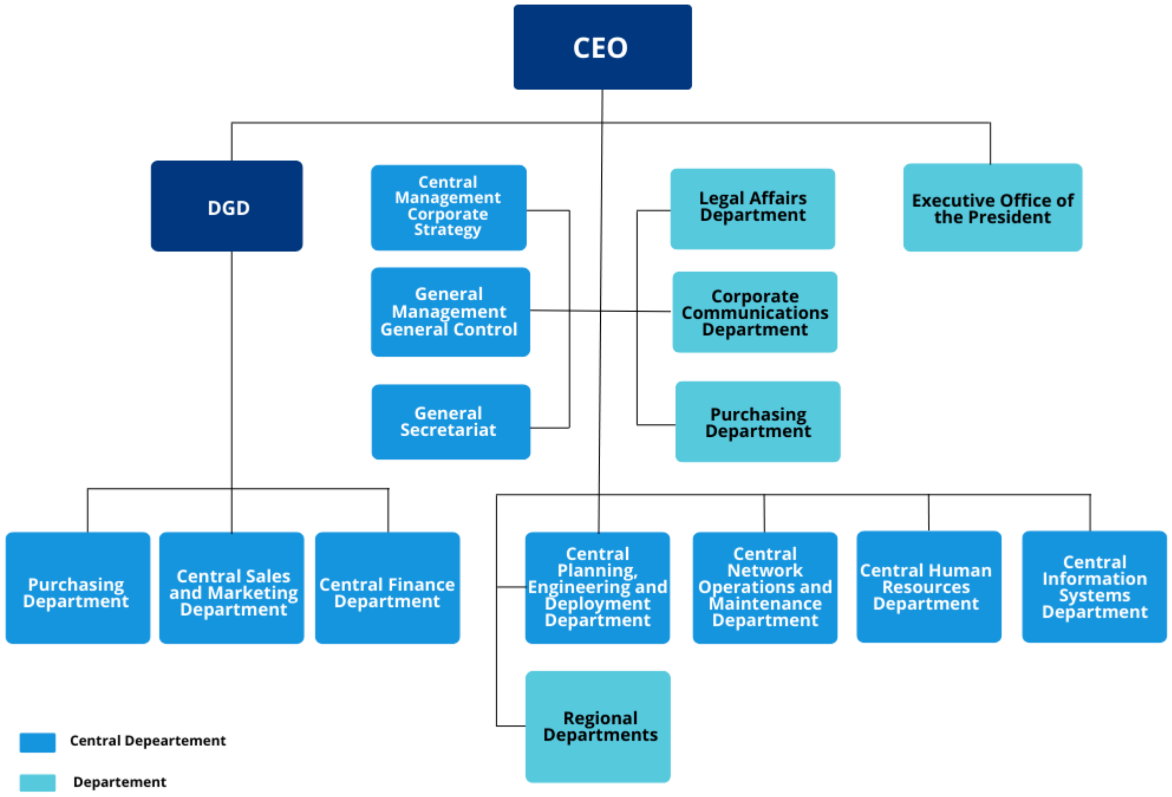

CEO
DGD
Central Management Corporate Strategy
General Management General Control
General Secretariat
Legal Affairs Department
Corporate Communications Department
Purchasing Department
Executive Office of the President
Purchasing Department
Central Sales and Marketing Department
Central Finance Department
Central Planning, Engineering and Deployment Department
Central Network Operations and Maintenance Department
Central Human Resources Department
Central Information Systems Department
Regional Departments
Central Depeartement
Departement

Figure 2: Tunisie Telecom Organizational Chart

# 2. Context of the Project

## 2.1. Problem Statement

We need to launch customized offers that match customer needs. To know customers and their behaviour is the best way to deliver an excellent service.

We have a big number of customers however we cannot rely on typical generalisations. Customers are anonymous so we know very little about them but we have data that documents their behaviour.

## 2.2. Object and Requirements

Our object is to segment customers into homogeneous groups or segments according to their behavioural data.

We are required to provide:

An unsupervised segmentation model: to detect stereotypical customer behaviour.

Supervised classification model: to classify new customers in the classes that correspond to them.

Dashboards and a dashboard application: to visualise and study patterns within the data.

## 2.3. Proposed Solution

Consumers segmentation is a marketing strategy it consists of dividing a broad consumer or business market into sub-groups of consumers based on some type of shared attributes rather than seeing them as a homogenic population with similar profiles and needs.

This strategy allows us to provide different market segments with different marketing programs. If provided with big and high-quality data it can identify purchase motivations, however, the data for a demographic or psychographic segmentation is difficult to obtain. In our project, we are applying a behavioural segmentation.

### 2.4. Methodology and Process

In this section, we will present the methodology and the process used in our project which will determine the next modules of our work.

Generally, data science projects go hand-in-hand with traditional software development. Data science is much more experimental in nature than software development. This may seem wasteful in software development, but it is essential for a successful data science project. It's often advantageous and recommended to delay or completely avoid, if possible, committing to a single data science technique. Instead, data science related components should be built around experimentation and designed for flexibility if a better performing alternative becomes available.

#### 2.4.1. Agile

Agile is an iterative software development methodology that aims to reduce time to market (the time it takes for a product being conceived until its being available for sale).

Scrum is one of the many frameworks that can be used to implement rapid development. In Scrum Agile, development takes place in sprint courses, and at the end of each sprint a minimum viable product is deployed. A sprint usually ranges anywhere from 1 to 4 weeks.

#### 2.4.2. CRISP-DM

The cross-industry standard process for data mining or CRISP-DM is an open standard process framework model for data mining project planning. It is the most widely-used analytics model.

CRISP-DM conceptualizes data science as a cyclical process. And, while the cycle focuses on iterative progress, the process doesn't always flow in a single direction. In fact, each step may cause the process to revert to any previous step, and often, the steps can run in parallel.

CRISP-DM divides the data mining process into six major phases but insure also different iterations (going back and forth in the different steps):

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

The CRISP-DM approach helps ensure that business objectives remain at the heart of the project. Also, it provides an iterative approach, including frequent opportunities to evaluate the progress of the project against its original objectives. (Wirth, 2000)

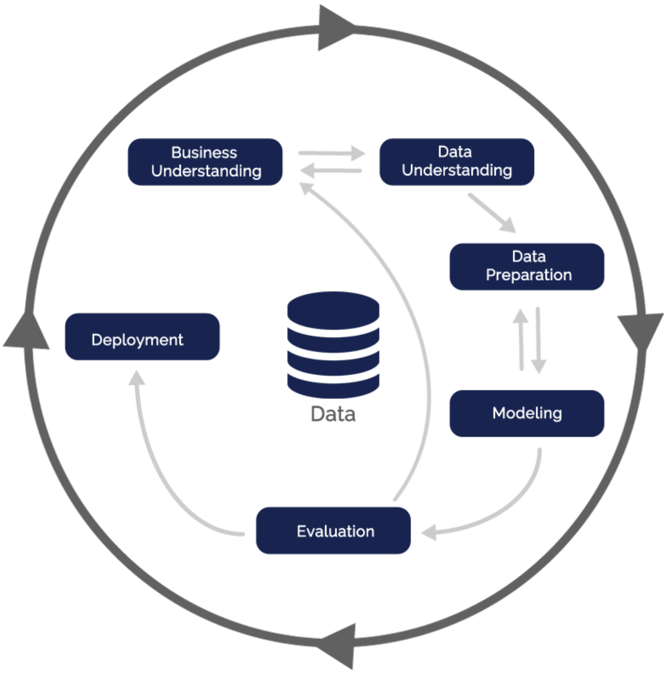


Figure 3: Process Diagram of the CRISP-DM Phases

#### 2.4.2.1. Business Understanding

Focuses on understanding the project objectives and requirements from a business perspective. The analyst formulates this knowledge as a data mining problem and develops preliminary plan.

#### 2.4.2.2. Data Understanding

Starting with initial data collection, the analyst proceeds with activities to get familiar with the data, identify data quality problems and discover first insights into the data. In this phase, the analyst might also detect interesting subsets to form hypotheses for hidden information.

#### 2.4.2.3. Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. The data preparation task is the final selection of the data, acquiring it and doing all the possible cleaning, formatting and integration. it also should be extended to some transformation and enrichment (for wider possibilities of analysis). It could sometimes be the longest part of your data mining project.

#### 2.4.2.4. Modelling

The analyst evaluates, selects and applies the appropriate modelling techniques. Since some techniques like neural networks have specific requirements regarding the form of the data. There can be a loop back here to data preparation.

#### 2.4.2.5. Evaluation

The analyst builds and chooses models that appear to have high quality based on loss functions that were selected. The analyst them tests them to ensure that they can generalise the models against unseen data. Subsequently, the analyst also validates that the models sufficiently cover all key business issues. The end result is the selection of the champion model(s).

#### 2.4.2.6. Deployment

Generally, this will mean deploying a code representation of the model into an operating system. This also includes mechanisms to score or categorise new unseen data as it arises. The mechanism should use the new information in the solution of the original business problem. Importantly, the code representation must also include all the data preparation steps leading up to modelling. This ensures that the model will treat new raw data in the same manner as during model development.

## Conclusion

During this introductory chapter, we presented the company in which we carried out our end-of-study internship, we stated the problems faced, the objectives to be achieved as well as the work methodology adopted. In the next chapter, we will we study the business environment.

# Chapter 2: Business Understanding

## Introduction

The first step of the CRISP-DM process is business understanding. This step includes the basic groundwork for the rest of the project, such as determining goals and objectives.

## 1. Project Background

Businesses now face severe competition in a highly dynamic and unstable marketing environment. In order to be successful and hold a leading place on the market, they have to provide quality service and respond to changes in their customers' needs, wishes, characteristics and behaviours. So, instead of viewing customers as homogenous and engaging all customers in the same marketing campaigns or incentives, companies should approach customers differently, based on their needs, characteristics, and behaviours. (Bose, 2010)

Given the fierce competition and the highly dynamic environment, companies can no longer content themselves with attracting new customers. They should place equal emphasis on strategies that focus on retaining customers and returning former subscribers, rather than increasing their market share. Most of the time, building customer loyalty and increasing profitability rates are just as important and easier to achieve than attracting new customers from the competition. (Tsiptsis, 2009)

To develop a proper relationship with the subscribers, telecom operators need to implement and apply the principles of customer relationship management. Introducing CRM enables operators to customize products, services and communication, to create and offer higher customer value, increase customer loyalty, improve business stability with higher customer retention rates and create value, boost sales and reduce the number of dissatisfied clients. (Diller, 2000)

Relationship marketing focuses on identifying the most suitable and profitable customers to develop mutually beneficial long-term relationships. (Niarn, Agnes, & Bottomley, 2003)

Segmentation is an extremely important marketing concept because it is needed, together with better understanding of customers' needs, in order to improve the relationship with existing customers. (Storbacka & Kaj, 1997)

## 2. Business Objective

Segmentation and data analytics give us AI insight into customer behaviour, avoiding false generalizations, and making better business decisions for the purpose of maintaining and improving the service.

The business goal is to launch personalized offers adapted to the customer's needs. One of the best ways to know the customer's needs is to know his behaviour. As a consequence, telecommunications companies often build their segmentation around actual customer behaviour as observed via their networks and systems.

The concept of segmentation relies grouping individuals into segments based on common demands and behaviours to have a similar response to marketing strategies. Market segmentation is a fundamental component in the companies' strategic marketing planning in industrialized countries because goods and services can no longer be produced and retailed without taking into consideration the customers' needs and wishes and the fact that they differ. (Wedel, Michel, Kamacura, & Wagner, 2000)

## 3. Data Science Objectives

To offer personalized plans adapted to the customer's needs we need a model to classify the customers according to their behaviour and provide them the best suitable services.

Mobile operators use basically the following segmentation types: subscriber value-based segmentation, subscriber behaviour-based segmentation, subscriber lifecycle-based segmentation and subscriber (possible) migration-based segmentation. They are used for different situations and focus on different aspects. (Bayer & Judy, 2010)

In this context, we will be segmenting customers according to their behavioural data and combine that with a classifier model that can be used on new data. This segmentation is called a consumption or usage-based segmentation which is a subcategory of behavioural segmentation.

## 4. Technical Environment

A prepaid customer at a telecommunications company is often anonymous. Despite attempts to add descriptive properties to them (for example, by encouraging self-registration online through mobile applications), the majority of the customers usually stay anonymous or provide very little identifying information. Mainly, data is automatically collected in BTS using software.

The number of plans or segments depends on business factors and strategies; however, we can determine the ideal number of segments according to the data through algorithms.

The technical tools we are going to use consist of Python libraries to explore our data, a clustering algorithm, a classification algorithm to generate a model and data visualisations to study the segments of the demographic.

## Conclusion

In this chapter, we have studied the business background of the project, we presented the business objectives as well as data science goals and the technical environment we are using in the process. We then begin, in the next chapter, working on the data.

# Chapter 3: Data Understanding and Preparation

## Introduction

This chapter is dedicated to the second step in the CRISP-DM process, in which we will go through dissecting the data involved in our project through inspecting and exploring it to fully comprehend the significance of each feature.

Data collection is an intricate process that is essential to provide good service to mobile customers these days. Mobile operators' activity consists of gathering and managing a large amount of information and data. Thus, millions of people, in millions of places can perform tens or hundreds of transactions in a short period resulting in billions of events to be recorded. In order to handle such an enormous quantity of data, special analyses methods need to be involved. These have appeared and grown at the same pace with the information technology. (Bryan & Merlin, 2001)

## 1. Data Collection

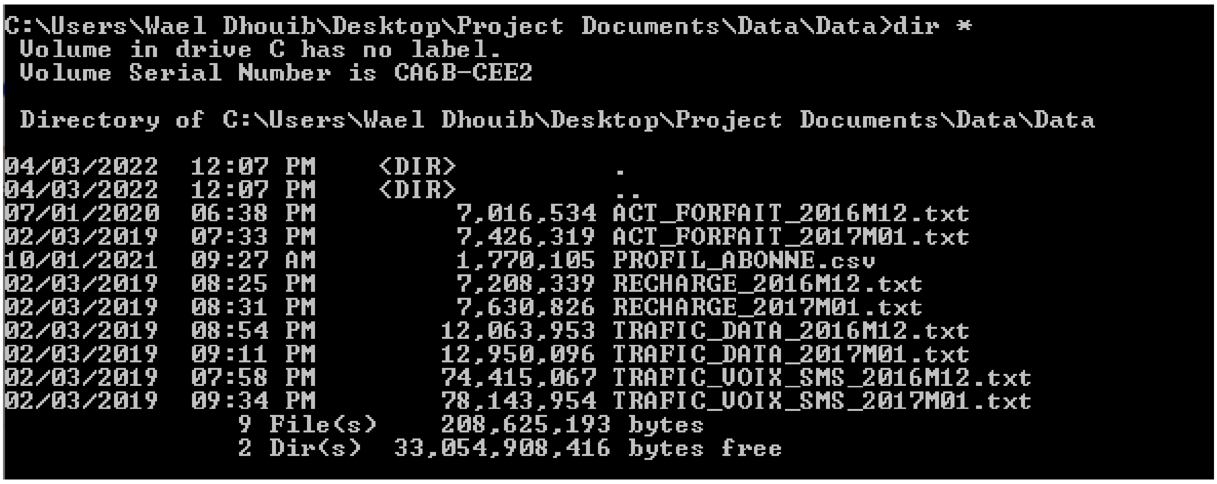
Telecommunication companies, like other large businesses, may have millions of customers. This necessarily means maintaining a database of information on their customers. This data includes the descriptions of the calls that traverse the telecommunication networks, network data and customer data.

Telecommunication companies maintain a great deal of data about their customers. In addition to the general customer data that most businesses collect, telecommunication companies also store call detail records, which precisely describe the calling behaviour of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and/or forecasting purposes. (Weiss, 2022)

For the data we are working with on this project, we have the profile information which is documented at the moment of buying the mobile subscription but the rest of data like plan activation, line recharge or any traffic is collected automatically through software in BTS.

## 2. Data Description

The data provided to us is 198 MB in size divided into 9 datasets. One dataset describes customer profiles and the rest is a record of customer activity over the two months of December 2016 and January 2017.



```
C:\Users\Wael Dhouib\Desktop\Project Documents\Data\Data>dir *
 Volume in drive C has no label.
 Volume Serial Number is CA6B-CEE2

 Directory of C:\Users\Wael Dhouib\Desktop\Project Documents\Data\Data

04/03/2022  12:07 PM    <DIR>          .
04/03/2022  12:07 PM    <DIR>          ..
07/01/2020  06:38 PM         7,016,534 ACT_FORFAIT_2016M12.txt
02/03/2019  07:33 PM         7,426,319 ACT_FORFAIT_2017M01.txt
10/01/2021  09:27 AM         1,770,105 PROFIL_ABONNE.csv
02/03/2019  08:25 PM         7,208,339 RECHARGE_2016M12.txt
02/03/2019  08:31 PM         7,630,826 RECHARGE_2017M01.txt
02/03/2019  08:54 PM        12,063,953 TRAFIC_DATA_2016M12.txt
02/03/2019  09:11 PM        12,950,096 TRAFIC_DATA_2017M01.txt
02/03/2019  07:58 PM        74,415,067 TRAFIC_VOIX_SMS_2016M12.txt
02/03/2019  09:34 PM        78,143,954 TRAFIC_VOIX_SMS_2017M01.txt
               9 File(s)    208,625,193 bytes
               2 Dir(s)  33,054,908,416 bytes free
```

Figure 4: Data Files in Command-line

The following tables contain the description of each feature of every dataset we have received:

| Feature | Description |
|---|---|
| CODE_CONTRAT | Subscription contract code |
| DATE_OPERATION | Plan activation date |
| FORFAIT | Mobile plan title |
| TYPE_FORFAIT | Mobile plan type |
| NBR_OPERATION | Number of times the plan has been activated or renewed |
| MNT_OPERATION | Total price |

Table 1: Plan Activation Dataset Fields

| Feature | Description |
|---|---|
| CODE_CONTRAT | Subscription contract code |
| DATE_RECHARGE | Recharge date |
| TYPE_REFILL | Refill type |
| NBR_RECHARGE | Number of line recharge operations |
| MONTANT_RECHARGE | Recharged amount |
| MONTANT_BONUS_RECHARGE | Bonus amount from line recharge |

Table 2: Line Recharge Activity Datasets Fields

| Feature | Description |
|---|---|
| CODE_CONTRAT | Subscription contract code |
| DATE_APPEL | Call date |
| TYPE_TAXATION | Taxation type |
| RESEAU_APPEL | Called network |
| NBR_APPEL | Number of calls |
| VOLUME_DATA | Data volume |
| COUT_TTC | Price all tax included |

Table 3: Data Traffic Activity Datasets Fields

| Feature | Description |
|---|---|
| CODE_CONTRAT | Subscription contract code |
| DATE_APPEL | Call date |
| TYPE_TAXATION | Taxation type |
| RESEAU_APPEL | Called network |
| TYPE_TRAFIC | Traffic type |
| DESTINATION_TRAFIC | Traffic destination |
| NBR_APPEL | Number of calls |

| | |
|---|---|
| DUREE_APPEL | Call duration |
| COUT_TTC | Price all tax included |

Table 4: Voice and SMS Traffic Activity Datasets Fields

| Feature | Description |
|---|---|
| CODE_CONTRAT | Subscription contract code |
| DATE_ACTIVATION | Contract activation date |
| DESC_SEGMENT_CLIENT | Subscriber segment description |
| DESC_OFFRE | Giveaway description |

Table 5: Subscriber Profile Dataset Fields

## 3. Exploratory Data Analysis

The datasets were imported into the Python kernel where we explored the features using conventional EDA practices. The data is structured, clean with little to no missing values due to the fact that it was collected and documented automatically by software in BTS.

In Figure 5, we have a time series plot of plan activation during the month of December 2016, we observe and linear increase towards the end of the month, and minor peeks at the beginning of each week.

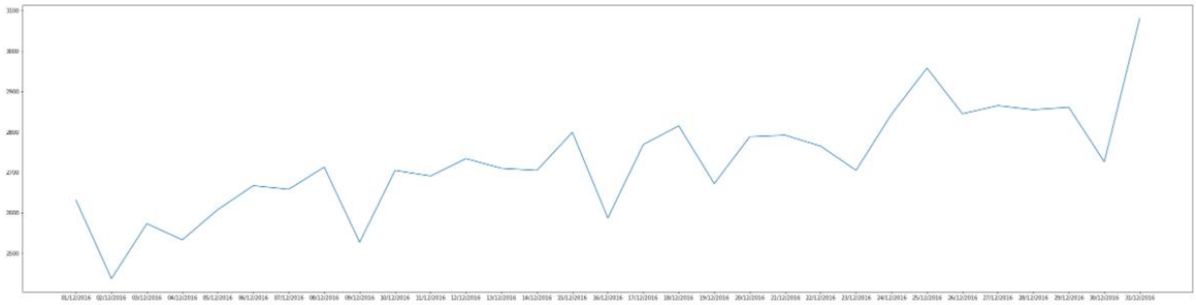


Figure 5: Plan Activation Time Series of December 2016

Form the bar plot in Figure 6, we observe that data is the main type of plan used by costumers as this period of time smartphones had a big surge in sales. Also, we can observe the abandonment of SMS. Voice communication remains used but not as much as data.

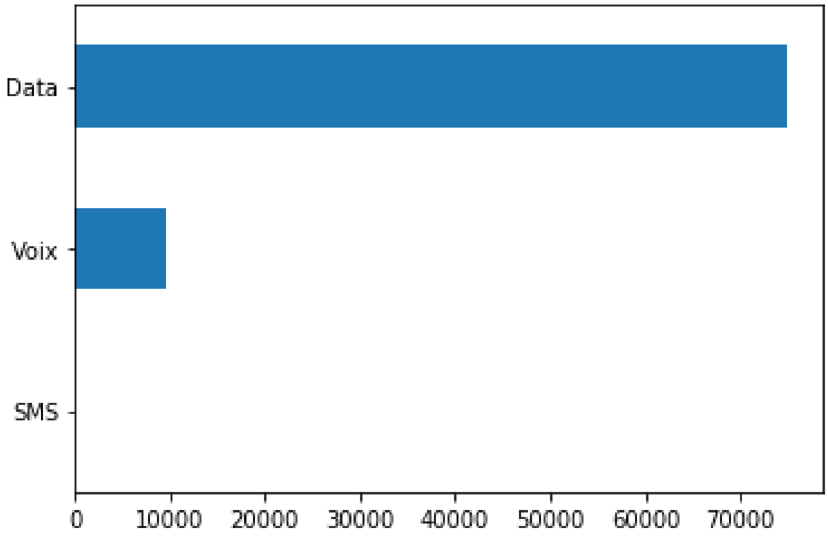
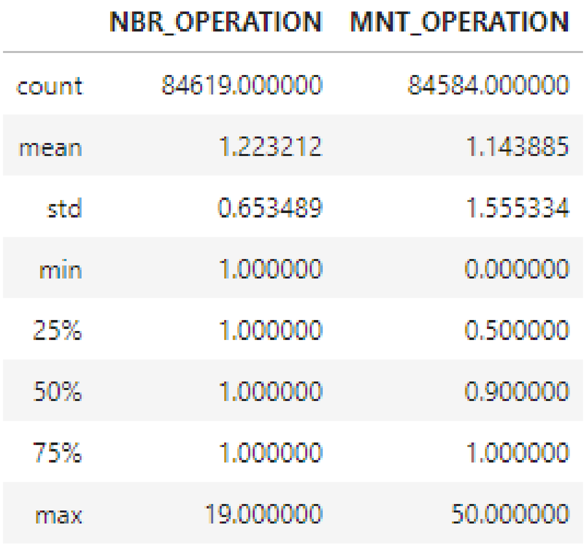


Figure 6: Plan Type Bar Plot

The average cost for a plan activation is 1.144 TND. Also, the average number of operations for a plan activation is 1.223.



| | NBR_OPERATION | MNT_OPERATION |
|---|---|---|
| count | 84619.000000 | 84584.000000 |
| mean | 1.223212 | 1.143885 |
| std | 0.653489 | 1.555334 |
| min | 1.000000 | 0.000000 |
| 25% | 1.000000 | 0.500000 |
| 50% | 1.000000 | 0.900000 |
| 75% | 1.000000 | 1.000000 |
| max | 19.000000 | 50.000000 |

Figure 7: Plan Activation Dataframe Description

In the dataframe description, as well as in Figures 8 and 9 below, we observe that the variables of the number of operations and the price of operations have asymmetrical distributions.

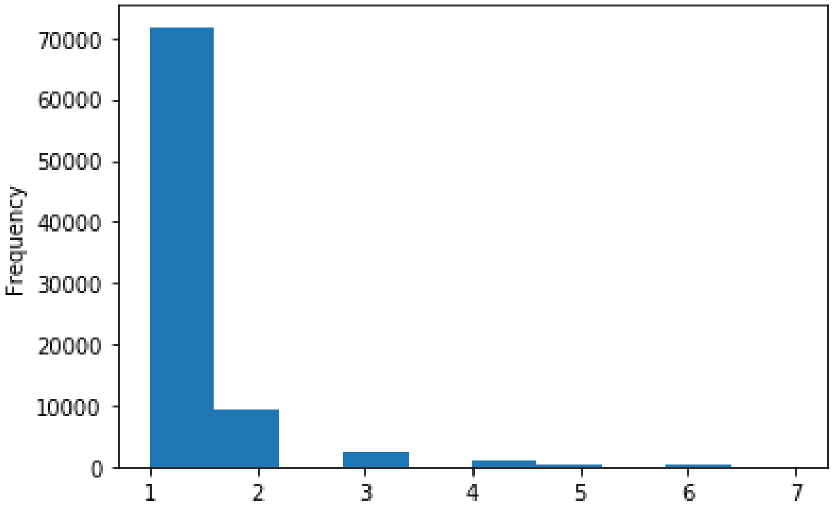


Figure 8: Number of Operations per Plan Activation Distribution

The bar plot shows that the two variables have positive skewed distributions, the mode equals 1 meaning that the most activated plan has the price of 1 TND.

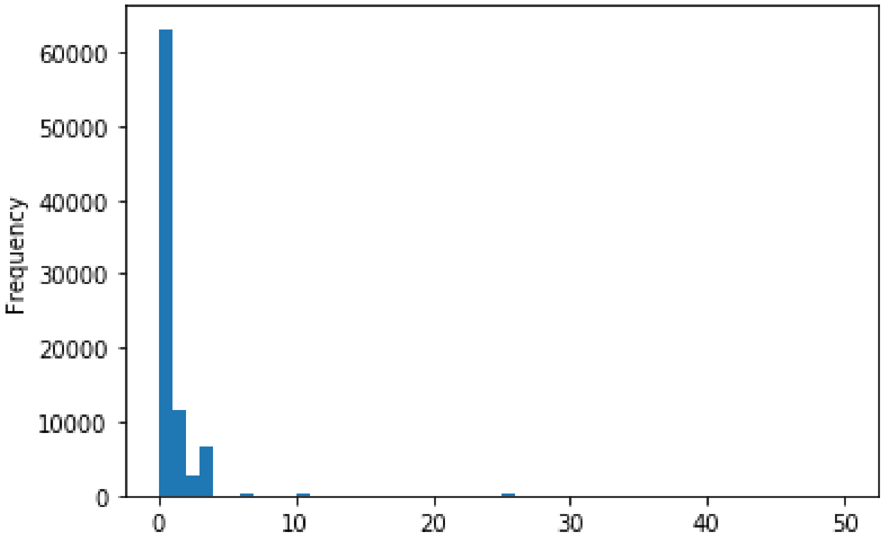


Figure 9: Plan Activation Price Distribution

From the bar plot in Figure 10, we notice that customers in late 2016 and early 2017 still predominantly use scratch cards. While electronic line recharge is used but significantly less than scratch cards, data electronic recharge is very little used that is not visible on the chart.

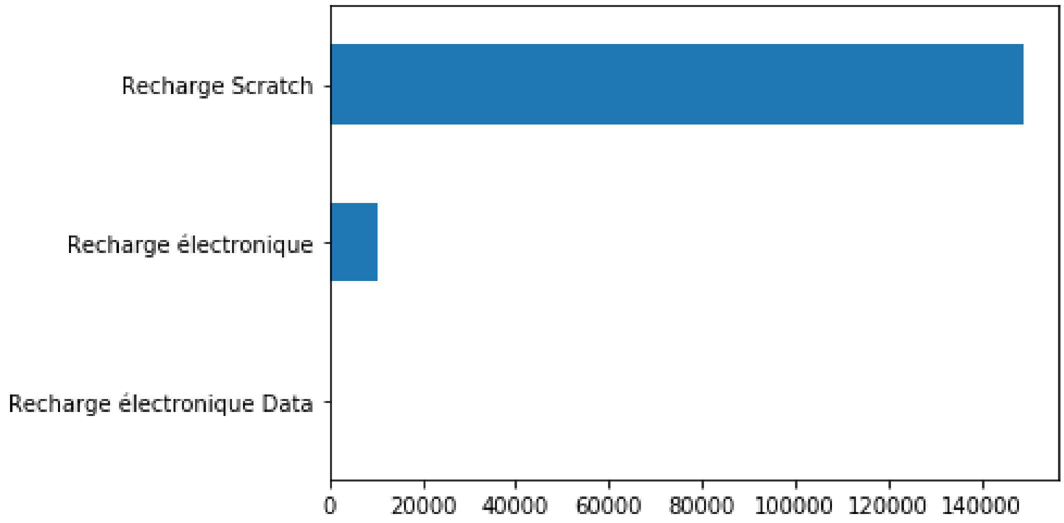


Figure 10: Line Recharge Type Bar Plot

According to the bar graph in Figure 11, we report that approximately half of data traffic is free for both December 2016 and January 2017. During both months, the total of traffic is around 250000 operations.

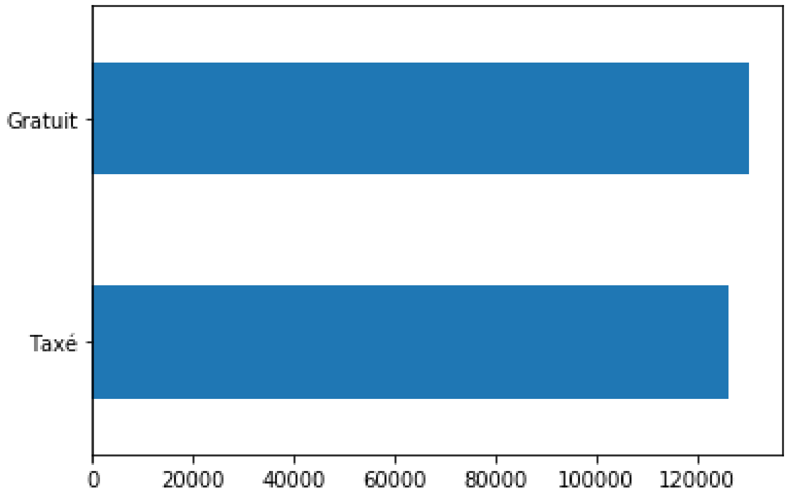


Figure 11: Taxation Type Bar Plot

From the bar graph in Figure 12, we note that roaming data traffic is totally insignificant compared to local data traffic.

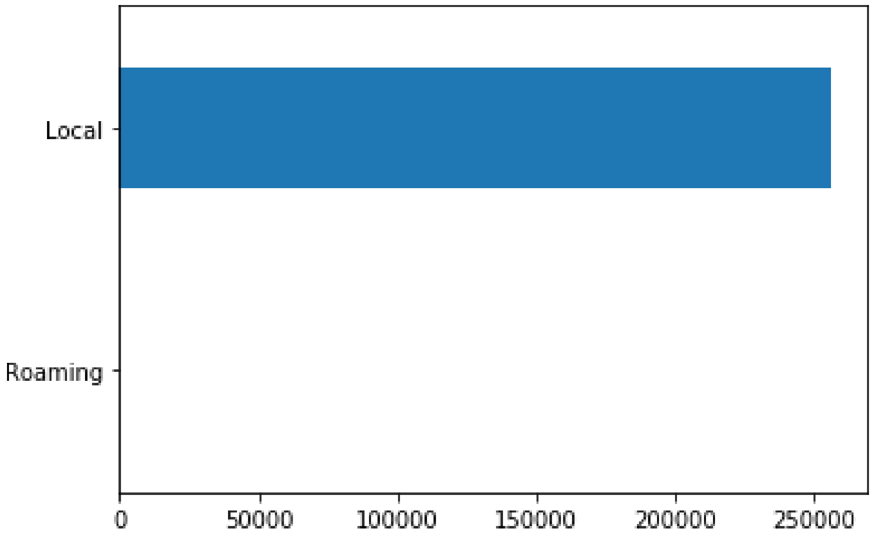


Figure 12: Called Network Bar Plot

In Figures 13 and 14 below, we observe the distributions of the number of calls and the volume of data from the data traffic dataset. The distributions are nearly identical which indicates that the variables are in high correlation.

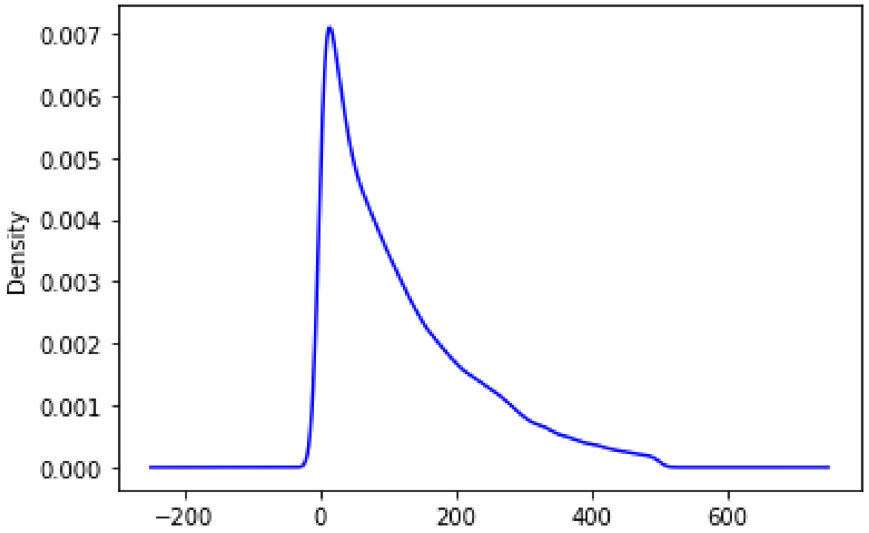


Figure 13: Data Traffic Calls Distribution

Both plots present positively skewed leptokurtic unimodal distributions. The average of data traffic calls is 146 call and for data traffic volume it is an average of 75917 kbit/s for a data traffic transaction.

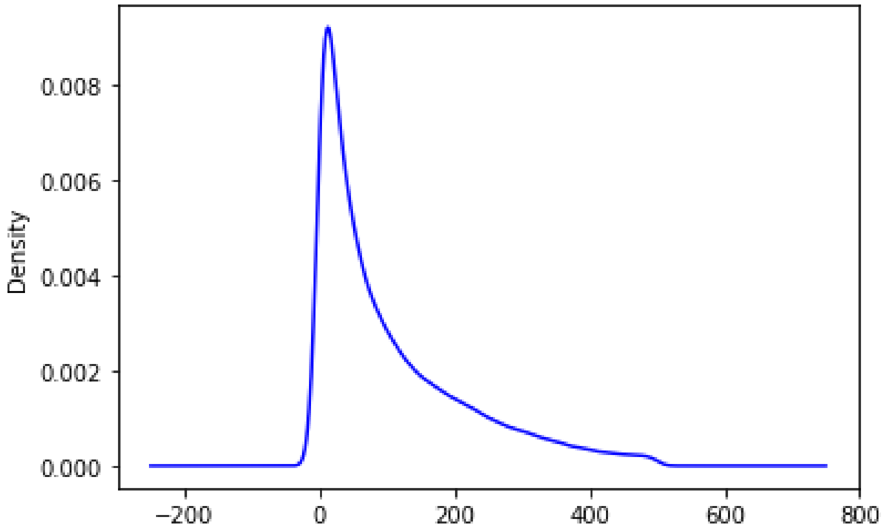


Figure 14: Data Traffic Volume Distribution

As seen on in Figure 15 the value axis for the heatmap shows that the values of the correlation matrix range from 0.7 to 1 which indicates that there are very strong correlations between the three quantitative variables of the data traffic dataframe.

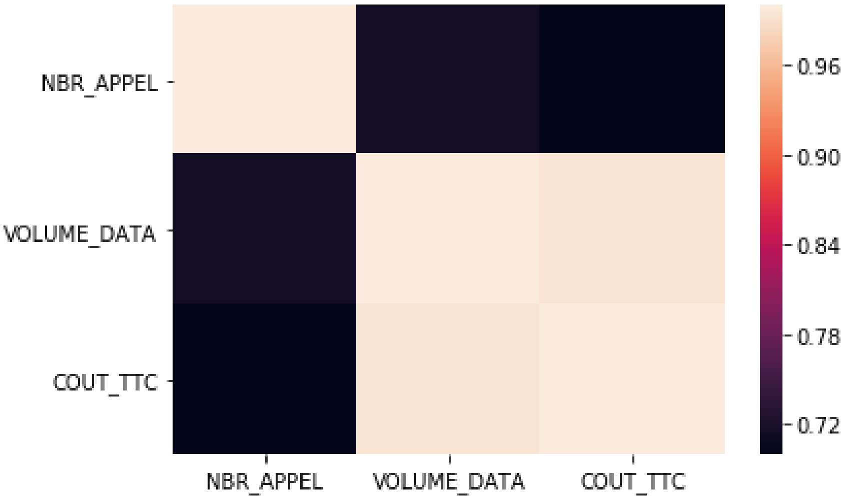


Figure 15: Data Traffic Dataframe Correlation Heatmap

The correlation between the traffic volume and cost is 0.99 while between the traffic cost and the number of calls it is 0.7 and between the number of calls and the traffic volume it is 0.71.

# 4. Data Preparation

This chapter is for the third step in the CRISP-DM process. We combined the December and January datasets for each category of datasets (plan activation, recharge line, data traffic and voice traffic and SMS) to obtain broader insights for each customer during segmentation, because the data provided is over a longer scope of time.

For every new dataframe we extracted new features that provide more insights about the costumer behaviour. The process includes a hot-encoding for categorical data variables so the result dataframe contains only numeral values and is ready for the modelling algorithm.

## 2.5. Feature Engineering

Through the date feature, we were able to generate weekly activity count variables for all dataframes. These variables can be useful for analytics or models that consider customers' recent activity.

The description of the engineered features in the plan activation activity dataframe:

| Feature | Description |
|---|---|
| ACTIVE_DAYS | The number of days the costumer is active |
| FORFAIT_1 to FORFAIT_10 | The count of activation for the top 10 used plans |
| DATA_FORFAIT | The count of activation of data plans |
| VOICE_FORFAIT | The count of activations of voice plans |
| SMS_FORFAIT | The count of activations of SMS plans |
| NBR_OPERATION_AVG | The average number of operations per active day |
| NBR_OPERATION_SUM | The total number of operations per active day |
| MNT_OPERATION_AVG | The average price of operations per active day |
| MNT_OPERATION_SUM | The total price of operation per active day |
| FORFAIT_AVG | The average activation of notable plans |
| TYPE_FORFAIT_AVG | The average plans activation per type |

| | |
|---|---|
| FORFAIT_WEEK_AVG | The weekly average of plans activation |

Table 6: Engineered Features in the Plan Activation Dataframe

| Feature | Description |
|---|---|
| RECHARGE_DAYS | The number of days the costumer recharged his line |
| REFILL_SCRATCH | The count of scratch recharge |
| REFILL_ELECTRONIC | The count of electronic recharge |
| REFILL_ELECTRONIC_DATA | The count of electronic data recharge |
| NBR_RECHARGE_SUM | The total of line recharge |
| NBR_RECHARGE_AVG | The average of line recharge |
| MONTANT_RECHARGE_AVG | The average price of recharges |
| MONTANT_RECHARGE_SUM | The total price of recharges |
| MONTANT_BONUS_RECHARGE_AVG | The average bonus amount per recharge |
| MONTANT_BONUS_RECHARGE_SUM | The total bonus amount per recharge |
| FORFAIT_AVG | The average of recharge count per type refill |
| RECHARGE_WEEK_AVG | The weekly recharged average |

Table 7: Engineered Features in the Line Registration Activity Dataframe

| Feature | Description |
|---|---|
| DATA_APPEL_DAYS | The number of days the costumer used data |
| DATA_TYPE_TAXATION_FREE | The count of free data transmissions |
| DATA_TYPE_TAXATION_TAXED | The count of taxed data transmissions |
| DATA_RESEAU_APPEL_LOCAL | The count of data transmissions on local network |
| DATA_RESEAU_APPEL_ROAMING | The count of roaming data transmissions |
| DATA_NBR_APPEL_AVG | The average number of calls |
| DATA_NBR_APPEL_SUM | The total number of calls |
| VOLUME_DATA_AVG | The average data volume |

| | |
|---|---|
| VOLUME_DATA_SUM | The total data volume |
| DATA_COUT_TTC_AVG | The average data price all tax included |
| DATA_COUT_TTC_SUM | The total data price all tax included |
| DATA_WEEK  1 to 10 | The count of data transmissions per week |
| DATA_TYPE_TAXATION_AVG | The average of taxed data transmissions |
| DATA_WEEK_AVG | The weekly average of data transmissions |

Table 8: Engineered Features in the Data Traffic Dataframe

| Feature | Description |
|---|---|
| APPEL_DAYS | The number of days the costumer used voice call or SMS |
| TYPE_TAXATION_FREE | The count of free call transmissions |
| TYPE_TAXATION_TAXED | The count of taxed call transmissions |
| RESEAU_APPEL_LOCAL | The count of call transmissions on local network |
| RESEAU_APPEL_ROAMING | The count of roaming call transmissions |
| TYPE_TRAFIC_VOICE | The count of voice traffic |
| TYPE_TRAFIC_SMS | The count of SMS traffic |
| TYPE_TRAFIC_MMS | The count of MMS traffic |
| DESTINATION_TRAFIC 1 to 12 | Traffic destination variables |
| NBR_APPEL_AVG | The average number of calls |
| NBR_APPEL_SUM | The total number of calls |
| DUREE_APPEL_AVG | The average duration of calls |
| DUREE_APPEL_SUM | The total duration of calls |
| COUT_TTC_AVG | The average price of calls all tax included |
| COUT_TTC_SUM | The total price of calls all tax included |
| APPEL_WEEK 1 to 10 | The count of call transmissions per week |
| TYPE_TAXATION_AVG | The average of taxed call transmissions |
| RESEAU_APPEL_AVG | The average called network score |

| | |
|---|---|
| DESTINATION_TRAFIC_AVG | The average destination traffic score |
| APPEL_WEEK_AVG | The weekly average of call transmissions |
| ONNET_TRAFIC | The count of on network traffic |
| OFFNET_TRAFIC | The count of off network traffic |
| OTHER_TRAFIC | The count of other network traffic |

Table 9: Engineered Features in the SMS and Voice Traffic Dataframe

## 2.6. Feature Selection

Many of the features in our data have high correlation hence they have a double effect in the generation of our model. When two features have high correlation, we drop one of the two features to decrease bias.

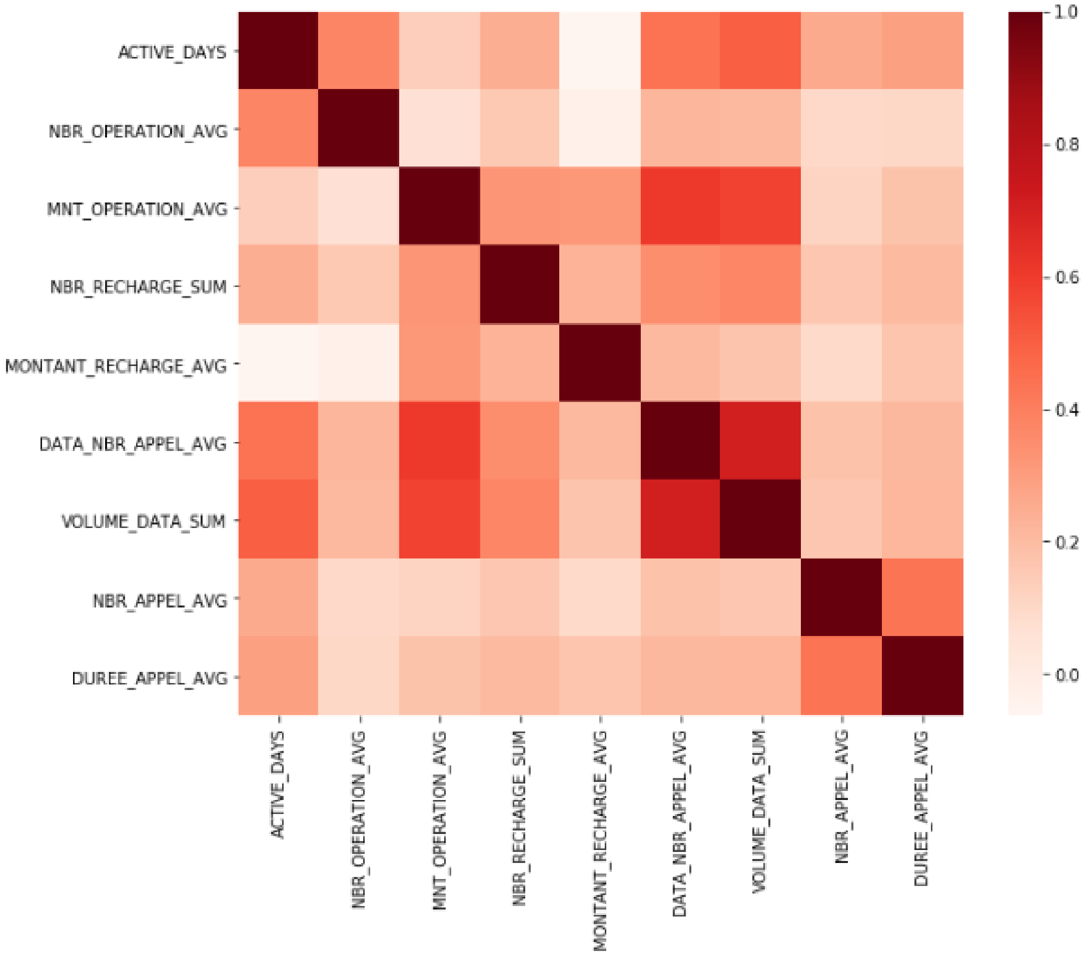


Figure 16: Correlation Heatmap of the Modelling Data

In Figure 21, we observe the correlation heatmap of the selected features from our data. The highest correlation on the modelling dataset is 0.411397. Variables strongly correlated to the number of active days, such as the number of line recharge days and the number of data usage days were eliminated.

## Conclusion

In this chapter, we described each feature of our data, we also performed an exhaustive analysis of the data, then we generated several features and made a selection of key features to prepare the data to run on our algorithm.

# Chapter 4: Modelling

## Introduction

In this chapter we present the fourth and fifth step in the CRISP-DM process, which is the pinnacle of the project, where we run the algorithms and generate the models and assess them.

## 1. Unsupervised Learning: Cluster Analysis

### 1.1. Algorithm Description

The k-means algorithm is an iterative algorithm that attempts to divide datasets into *k* pre-defined non-overlapping sets of clusters. In this case, each data point belongs to one group. It minimizes intra-cluster variance while maximizing inter-cluster variance at the same time, keeping the clusters as different as possible. It assigns data points to a cluster, so that the sum of the squared distance between the data points and the cluster's centroid is minimum. (Alibuhtto & Mahat, 2020)

### 1.2. Algorithm Selection

The clustering algorithm of choice is k-means for the generation of our model. The k-means algorithm is used to partition the input data set into *k* partitions. Also, k-means is based on calculating distance between data points which is the appropriate choice in this case because we have numerical data.

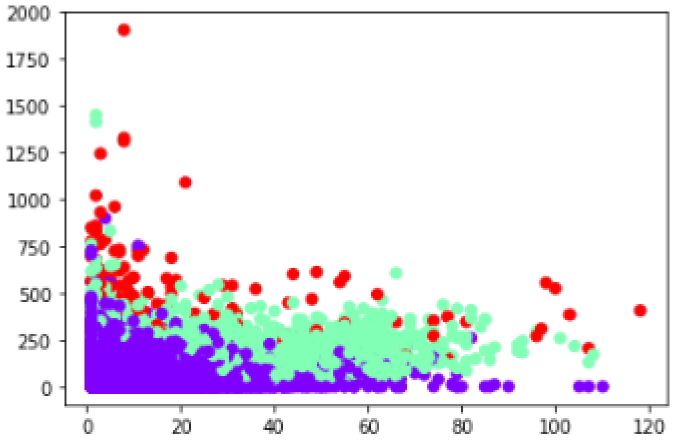


Figure 17: Scatter Plot of Customer Clusters Across Two Dimensions

A study at Faculty of Economics and Business Administration in Cluj-Napoca, Romania pursued the applicability of the k-means cluster method to provide business intelligence. They concluded that the method employed proved its efficiency in processing large data volumes leading to obtaining consumer segments featuring high internal homogeneity of the segments and high heterogeneity among segments. (Bacila, Adrian, & Ioan, 2012)

## 2. Optimal Number of Clusters

### 2.1. Elbow Method

The elbow method is used to determine the optimal number of clusters in k-means clustering. The optimal number of clusters is the point represented by the bend of the curve. As seen in Figure 18, the elbow can be represented by *k*=3. Generally, the elbow point is that of the number of clusters from which the inter-cluster variance is no longer significantly reduced. In Figure 17 above, we visualise a scatter plot of the clusters across two dimensions.

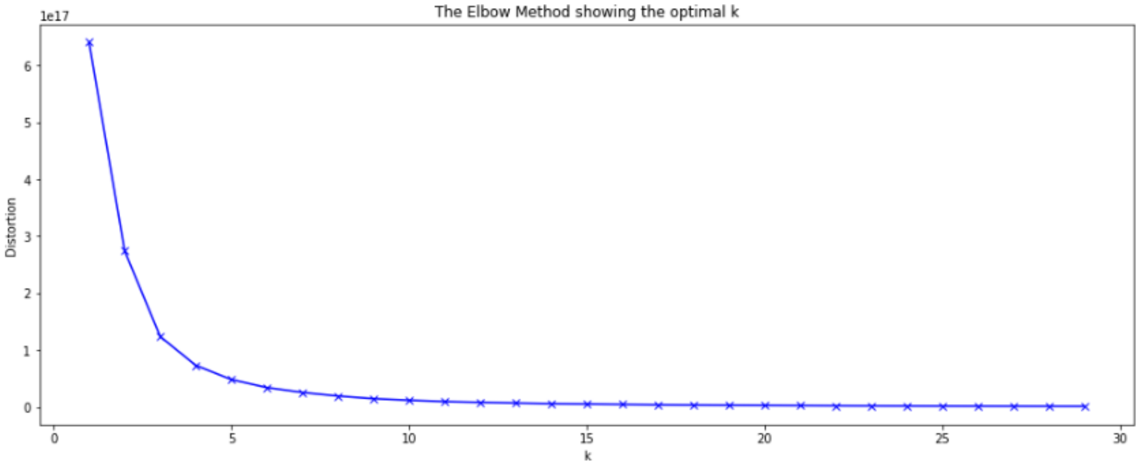


Figure 18: The Elbow Method Using Distortion

The elbow method plots the value of the cost function produced by different values of *k*. As you know, if *k* increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as *k* increases. The value of *k* at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters. (Dangeti, 2017)

# 3. Supervised Learning: Classifier Model

## 3.1. Apache Spark

Apache Spark is a unified analytics engine for large-scale data processing. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, pandas API on Spark for pandas workloads, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for incremental computation and stream processing. (Spark 3.3.0 Documentation, 2022)

Apache Spark is an environment designed for large-scale data processing on programming clusters with implicit data parallelism and fault tolerance, which makes an ideal choice for our project as well as similar big data projects.

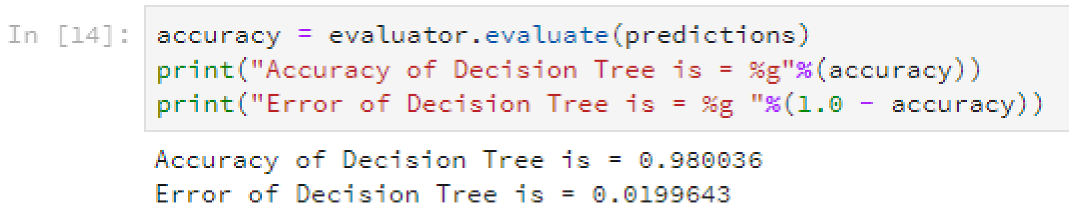
## 3.2. Modelling Algorithm Description

Decision tree is a supervised learning algorithm that can be used for both classification and regression problems. Decision trees are widely used since they are easy to interpret, handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions. Tree ensemble algorithms such as random forests and gradient-boosted trees are among the top performers for classification and regression tasks. (Spark 3.3.0 Documentation, 2022)

Decision tree classifier is the chosen algorithm for generating our classifier model as it is one of the tree ensemble algorithms. For the selection of the test and validation samples we used a random splitter with 80% of the data for test and 20% for validation.

# 4. Model Evaluation

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.



```
accuracy = evaluator.evaluate(predictions)
print("Accuracy of Decision Tree is = %g"%(accuracy))
print("Error of Decision Tree is = %g "%(1.0 - accuracy))
```

```
Accuracy of Decision Tree is = 0.980036
Error of Decision Tree is = 0.0199643
```

Figure 19: Accuracy of the Generated Classifier Model

We used a multiclass classification evaluator to calculate the accuracy of our model which value is 0.98 so the loss is 0.2 as shown Figure 19 above.

## Conclusion

After preparing the data, we run our segmentation algorithm, we also generated a classifier model for future use to assign the subscriber to an appropriate segment.

# Chapter 5: Deployment

## Introduction

In the final chapter, we present the last step of the CRISP-DM process, in which we will go through comparing the clusters and examining the charts presented in our dashboards.

## 1. Comparative Cluster Analysis

In this section, we examine the clusters to identify the criteria by which the algorithm unbundled subscribers through data analytics and visualisations.

In Figure 20, we observe a packet bubble chart. The size of the bubble is proportional to the number of subscribers in the cluster. It is clear that cluster number 1 is the largest with 25,449 subscribers followed by cluster number 2 with 1,964 subscribers and finally cluster number 3 with 153 subscribers.

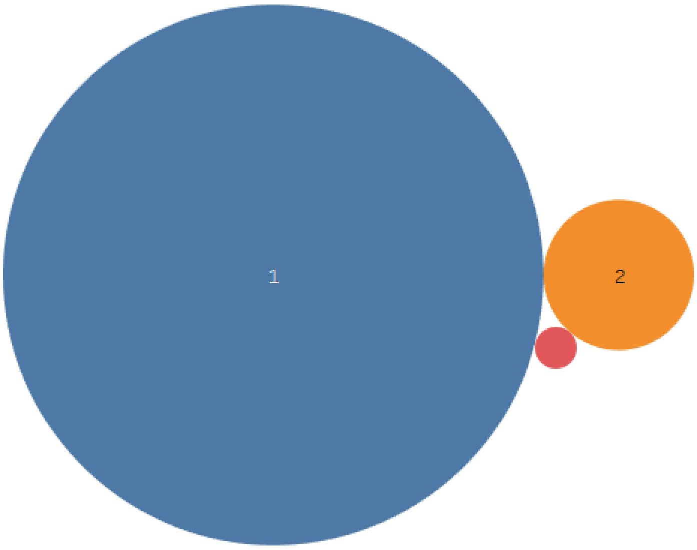


Figure 20: Clusters Packet Bubbles Chart

From the first line chart in Figure 21, we observe that cluster number 3 have the highest average data costs with 22.497 TND followed by the cluster number 2 with an average of 4.930 TND and then the cluster number 1 with 0.209 TND.

From the following observation, we come to conclude the that the subscribers of cluster number 3 are high spenders in terms of data costs followed by the subscribers of cluster 2 and the cluster 1.

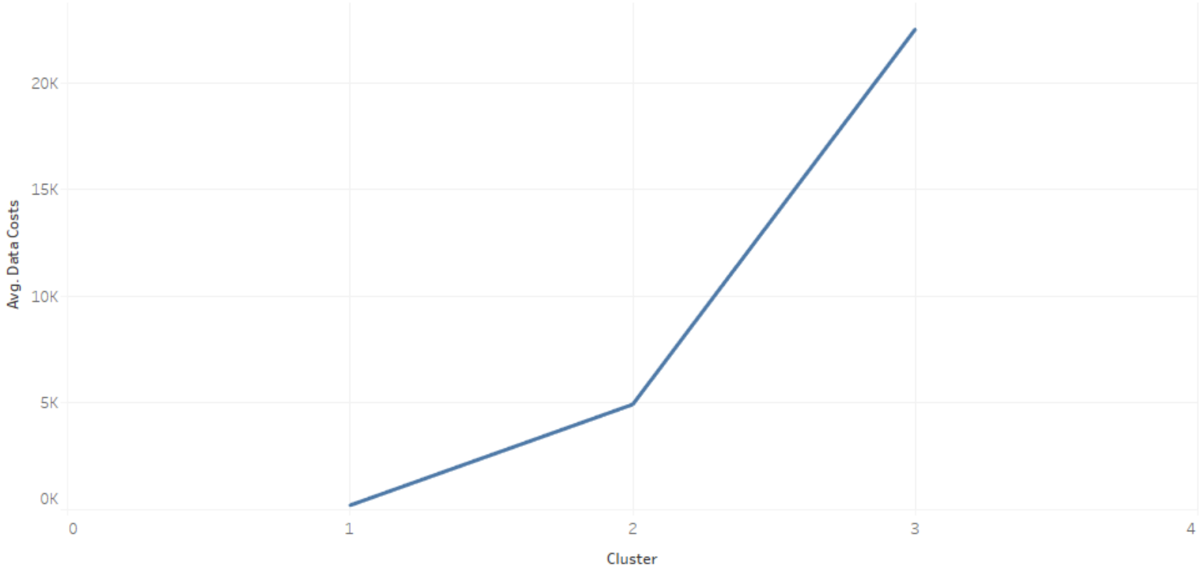


Figure 21: Average Data Costs Line Chart

In the second line chart in Figure 22, we notice that cluster number 2 have the highest average plan activation days with 34.81 followed by the cluster number 3 with an average of 21.95 and then the cluster number 1 with 4.74.

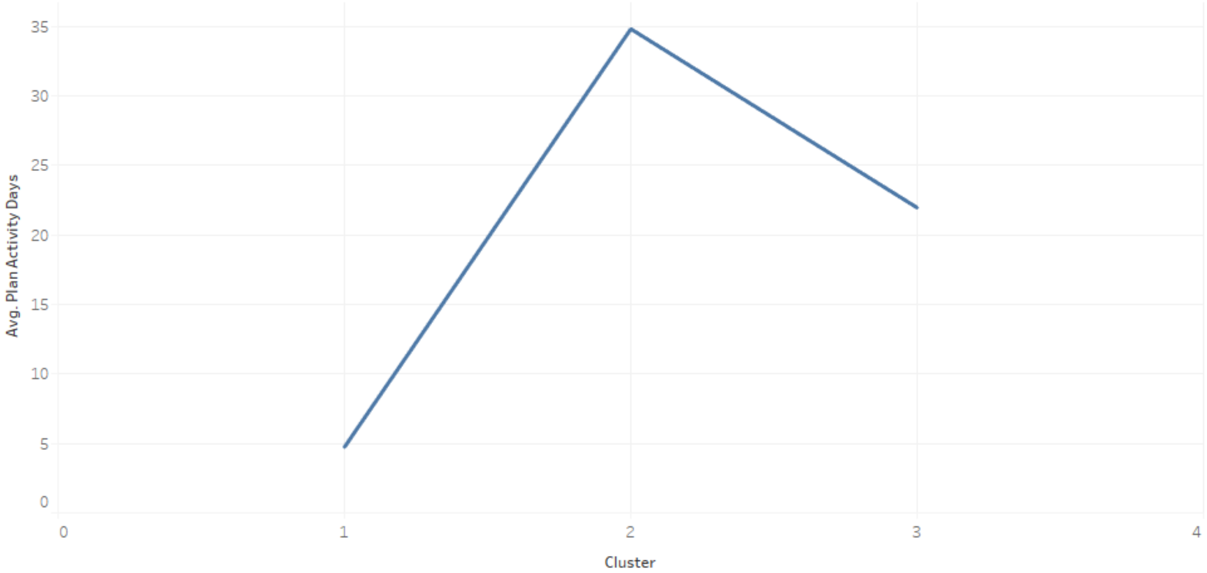


Figure 22: Average Plan Activity Line Chart

We explain this anomaly chart, as most chart have cluster 3 as top (see dashboard in Figure 44 in appendix), with the fact that subscribers of cluster 3 activate less plans as the plans they activate are high value plans so it provides them with services for longer time.

The third and last line chart in Figure 23, we see that cluster number 3 have the highest average call costs with 503.2 TND followed by the cluster number 2 with an average of 388.2 TND and then the cluster number 1 with 131.6 TND. The rest of the variables are correlated to average call costs and have similar chart.

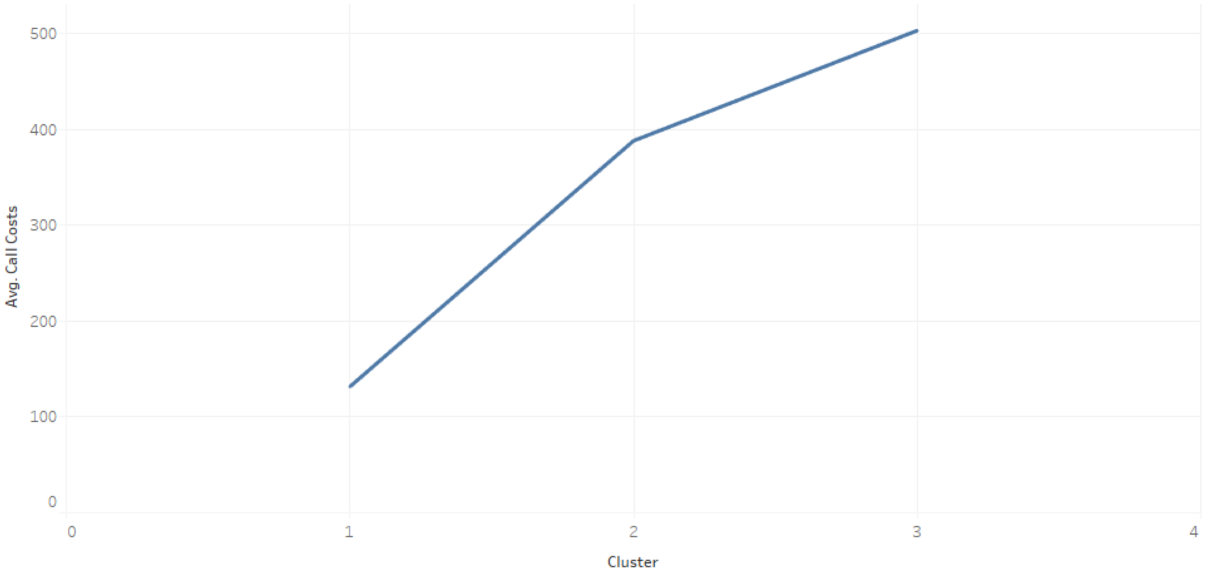


Figure 23: Average Call Costs Line Chart

In conclusion, cluster number 1 contains the average spending subscribers, cluster number 2 contains the above average spending subscribers, and cluster number 3 contains the high spending subscribers.

# 2. Analytics Dashboards

## 2.1. Plan Activity Dashboard

In the plan activity dashboard (the dashboard in Figure 40 in the appendix), we include plan activity income line chart during the two months of December 2016 and January 2017 shown in Figure 24.

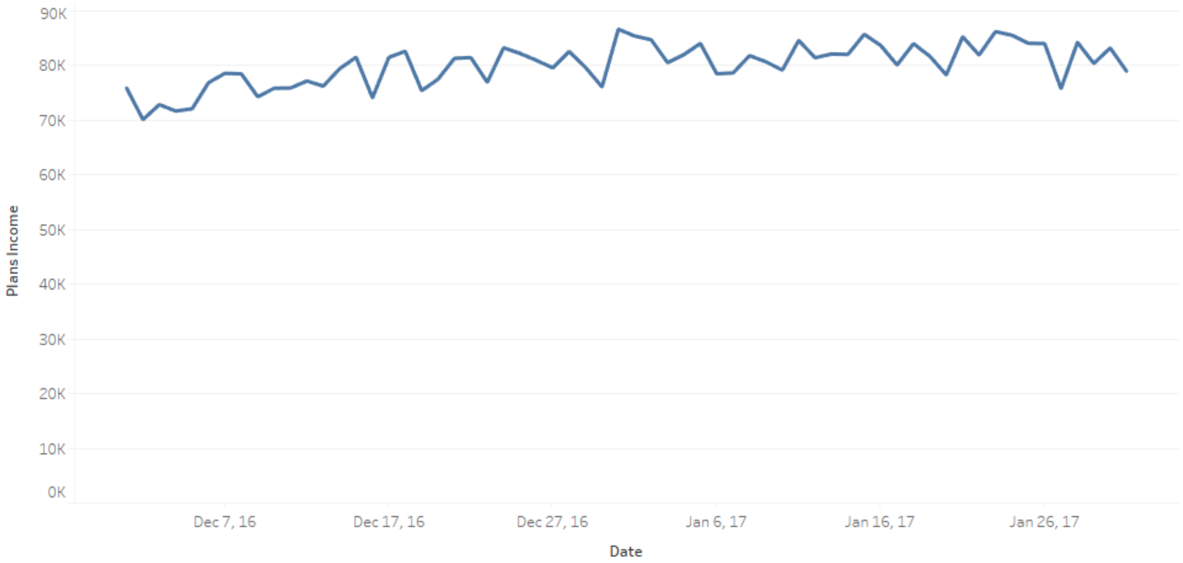


Figure 24: Plan Income Line Chart

In Figure 25, we observe the treemap of the plan type during the two months of December 2016 and January 2017 from the dashboard. Data are the dominant plan type followed by Voice and then SMS which is very minimal.

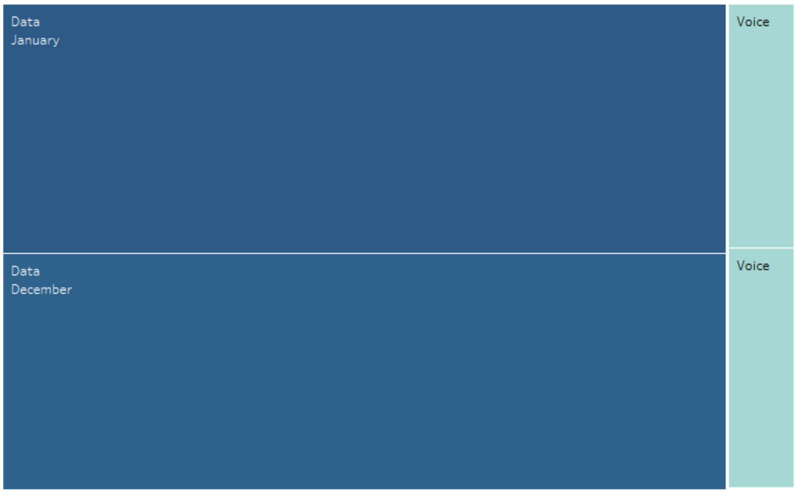


Figure 25: Treemap of Plan Income per Plan Type

### 2.2. Line Recharge Dashboard

In the line recharge dashboard (the dashboard in Figure 41 in the appendix), we include line recharge income line chart during the two months of December 2016 and January 2017 shown in Figure 26.

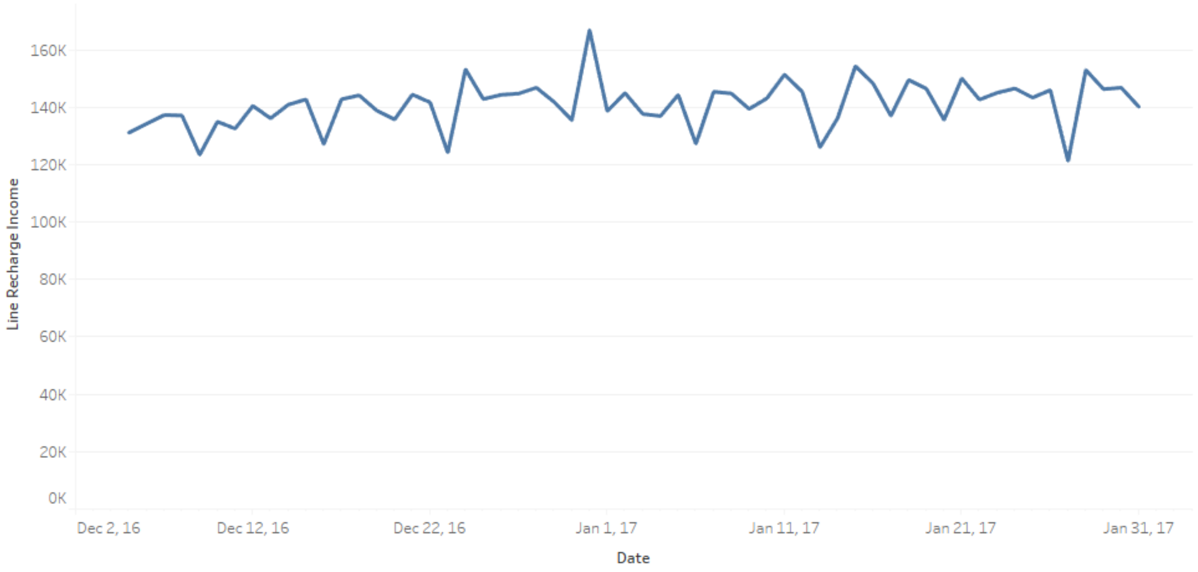


Figure 26: Line Recharge Income Line Chart

In Figure 27, we observe the treemap of the refill type during the two months of December 2016 and January 2017 from the dashboard. Scratch Recharge are the dominant line refill type followed by Electronic Recharge and then Data Electronic Recharge which is very minimal.

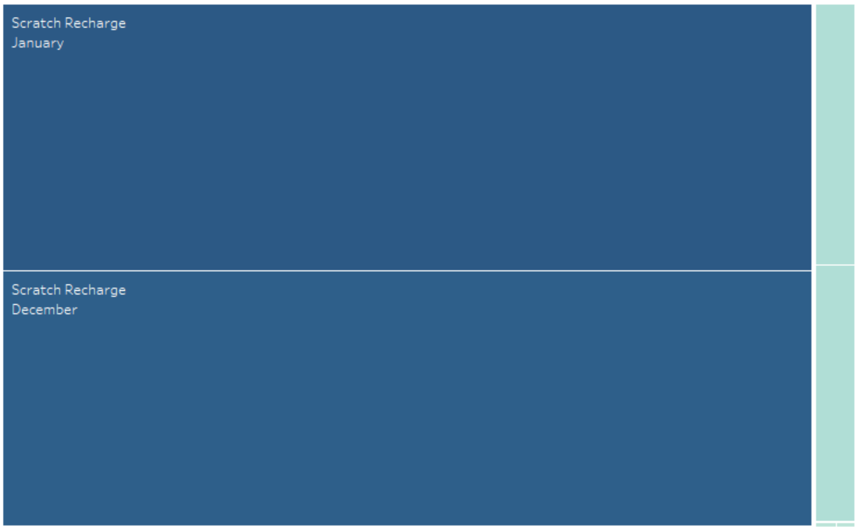


Figure 27: Treemap of Line Recharge Count per Refill Type

### 2.3. Data Traffic Dashboard

In the data traffic dashboard (the dashboard in Figure 42 in the appendix), we include multiple line chart first we have data transmission income line chart during the two months of December 2016 and January 2017 shown in Figure 28.

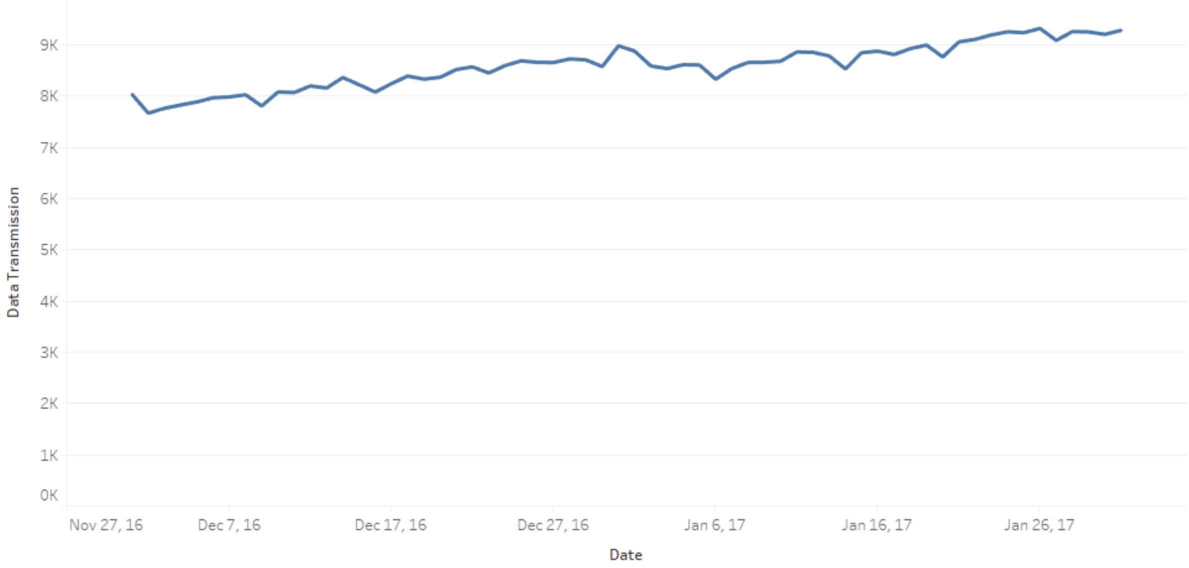


Figure 28: Data Transmission Line Chart

Also, we have data volume line chart during the two months of December 2016 and January 2017 shown in Figure 29.

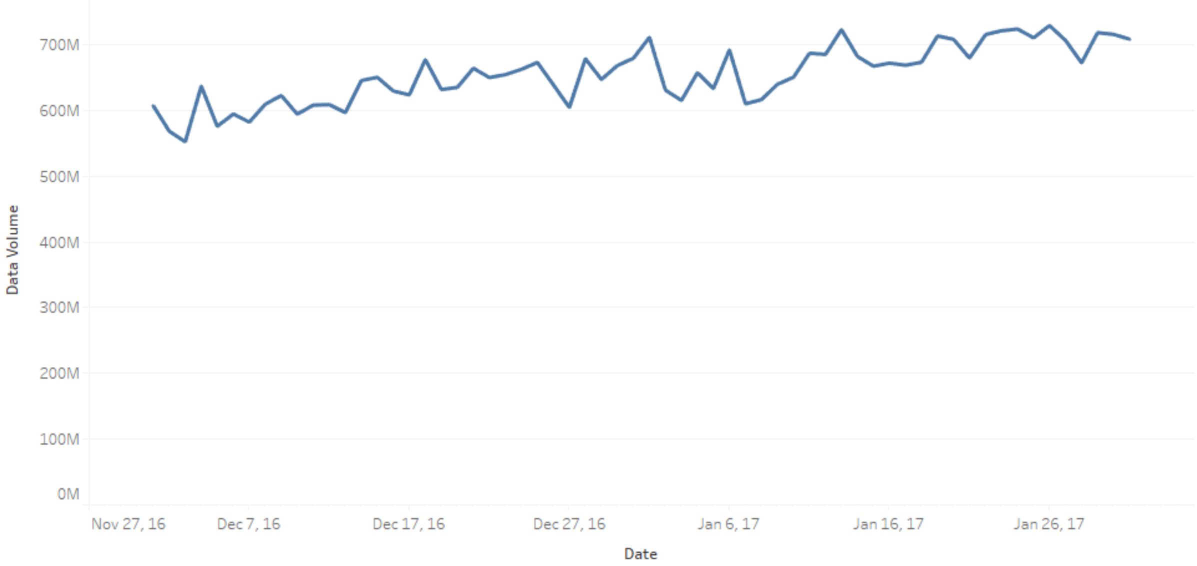


Figure 29: Data Volume Line Chart

Finally, we include data calls line chart during the two months of December 2016 and January 2017 shown in Figure 30.

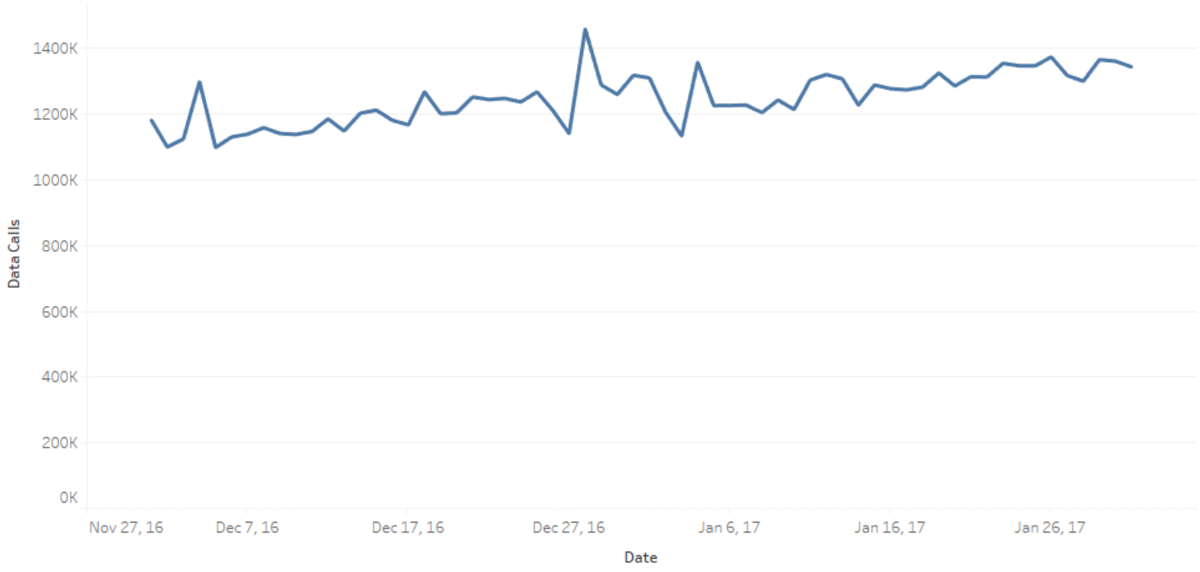


Figure 30: Data Calls Line Chart

## 2.4. Voice and SMS Traffic Dashboard

In the voice and SMS traffic dashboard (the dashboard in Figure 43 in the appendix), we include multiple line chart first we have data transmission income line chart during the two months of December 2016 and January 2017 shown in Figure 31.

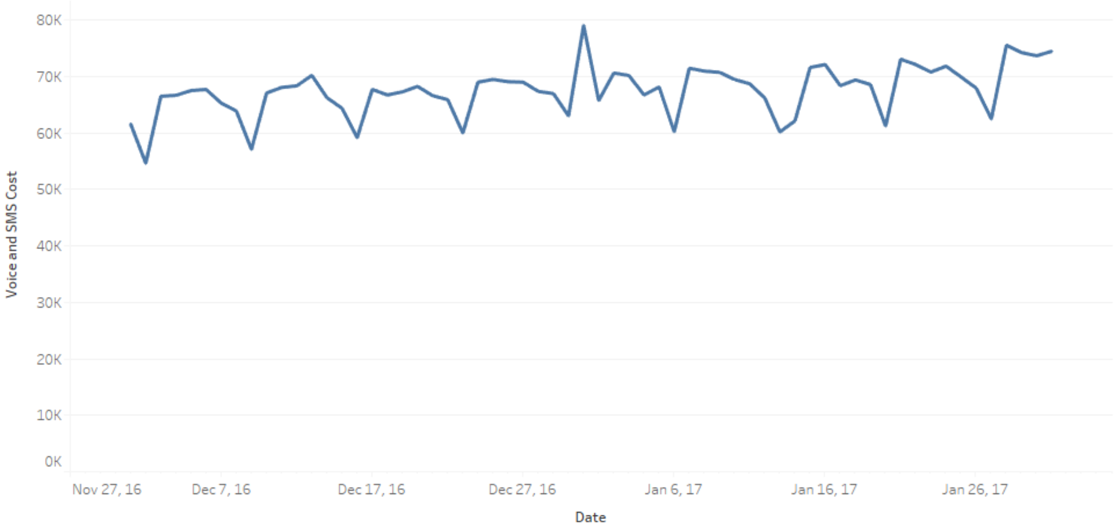


Figure 31: Voice and SMS Cost Line Chart

Also, we have call duration line chart during the two months of December 2016 and January 2017 shown in Figure 32.

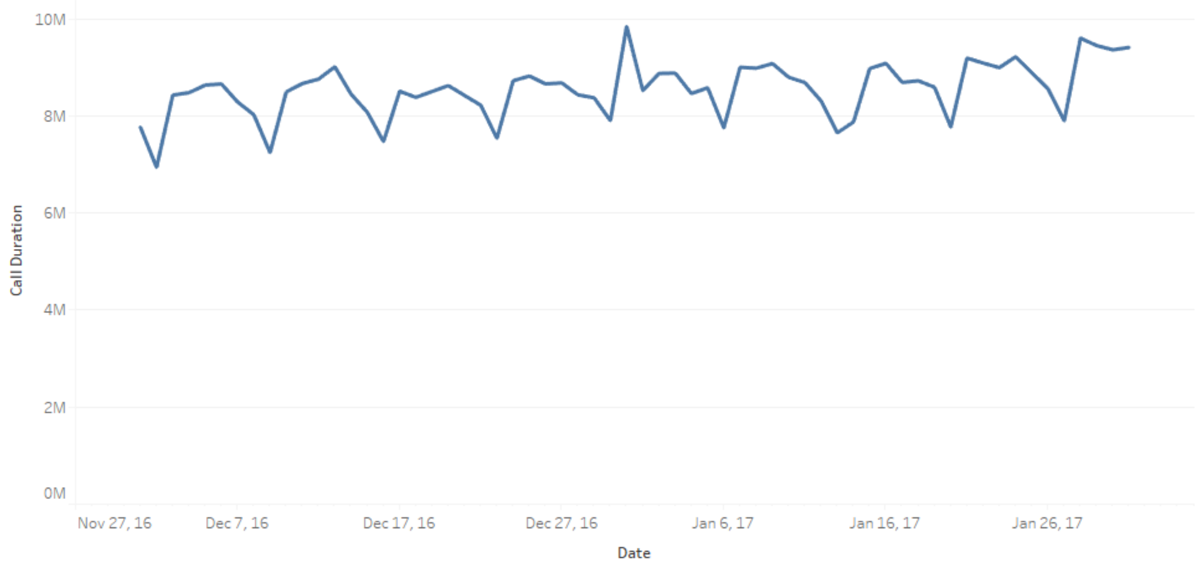


Figure 32: Call Duration Line Chart

Finally, we include number of calls chart during the two months of December 2016 and January 2017 shown in Figure 33.

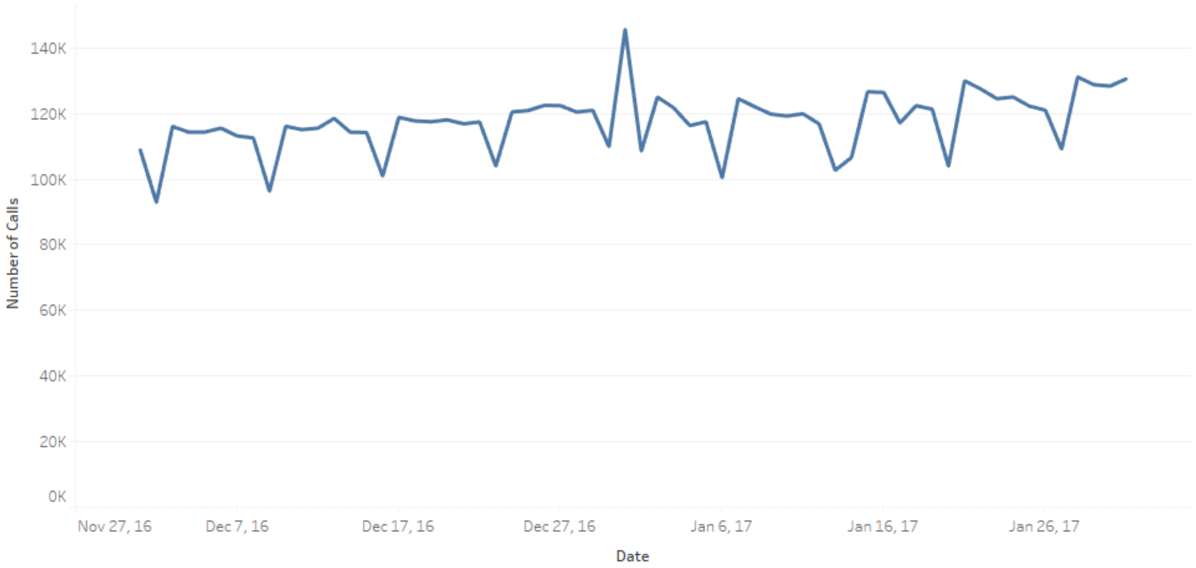


Figure 33: Number of Calls

## 2.5. Clusters Dashboard

In the clusters dashboard (the dashboard in Figure 45 in the appendix), we include a pie chart and a bar chart of the clusters. The line chart of the subscription from 2011 until 2019 as seen in Figure 34.

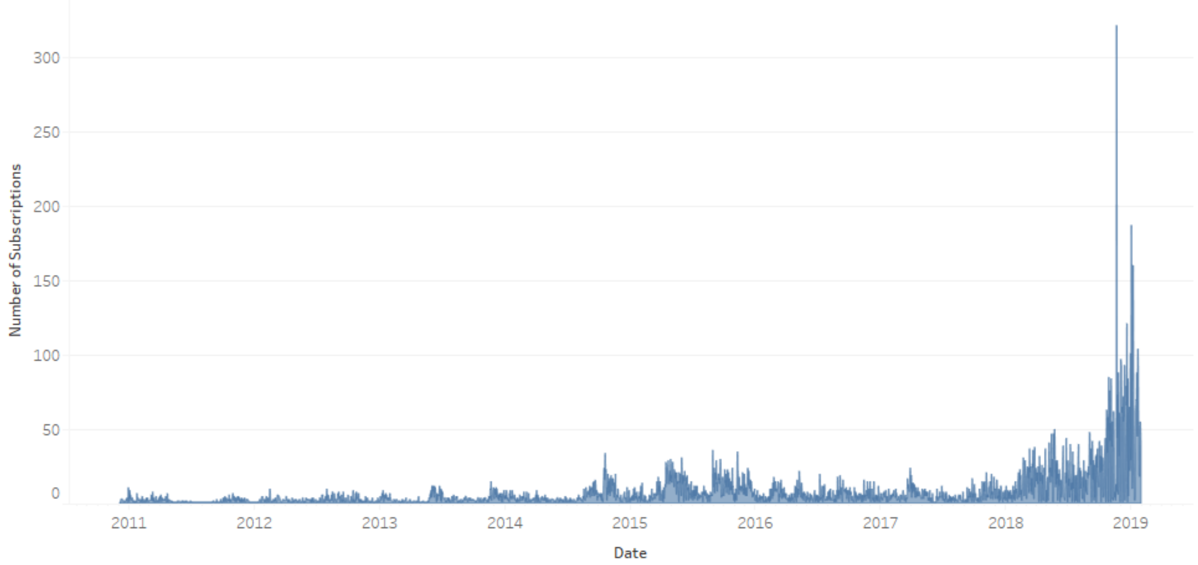


Figure 34: Subscription Line Chart

In Figure 35, we observe the scatter plot from the dashboard.

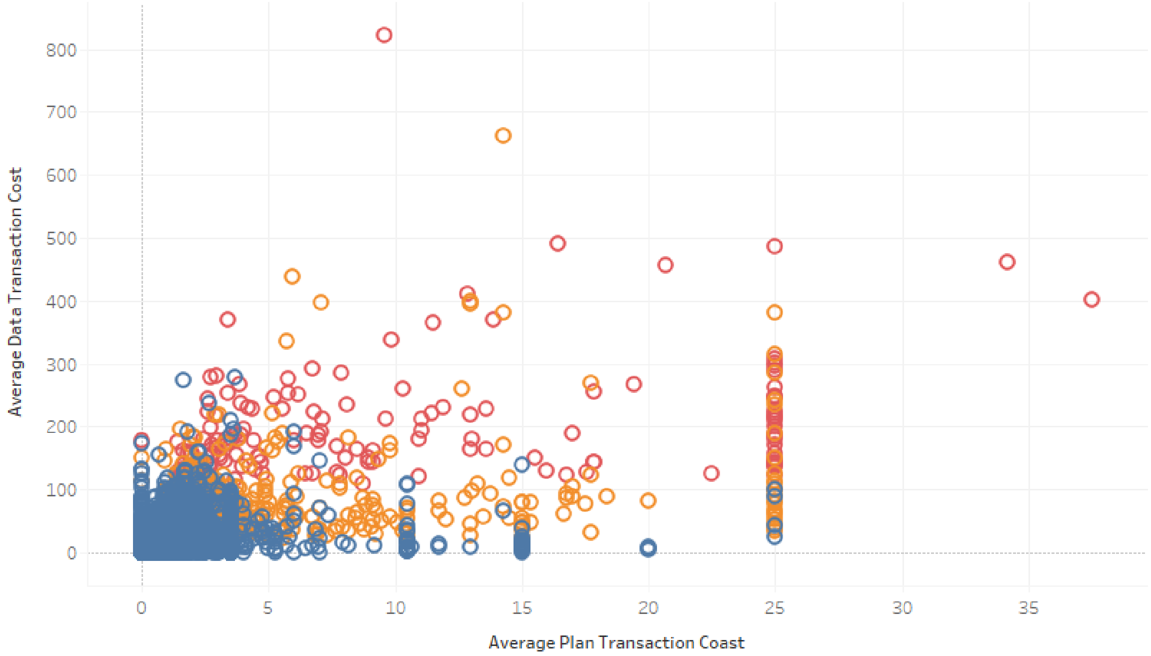


Figure 35: Subscriber Clusters Scatter Plot

## 3. Dashboard Application

We have developed a Django application that allow the administrator to import a CSV dataset and visualise a dashboard.

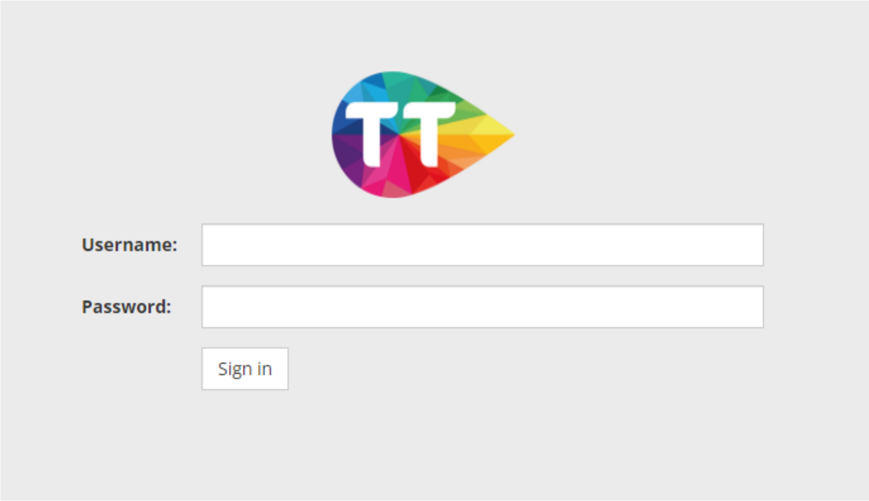
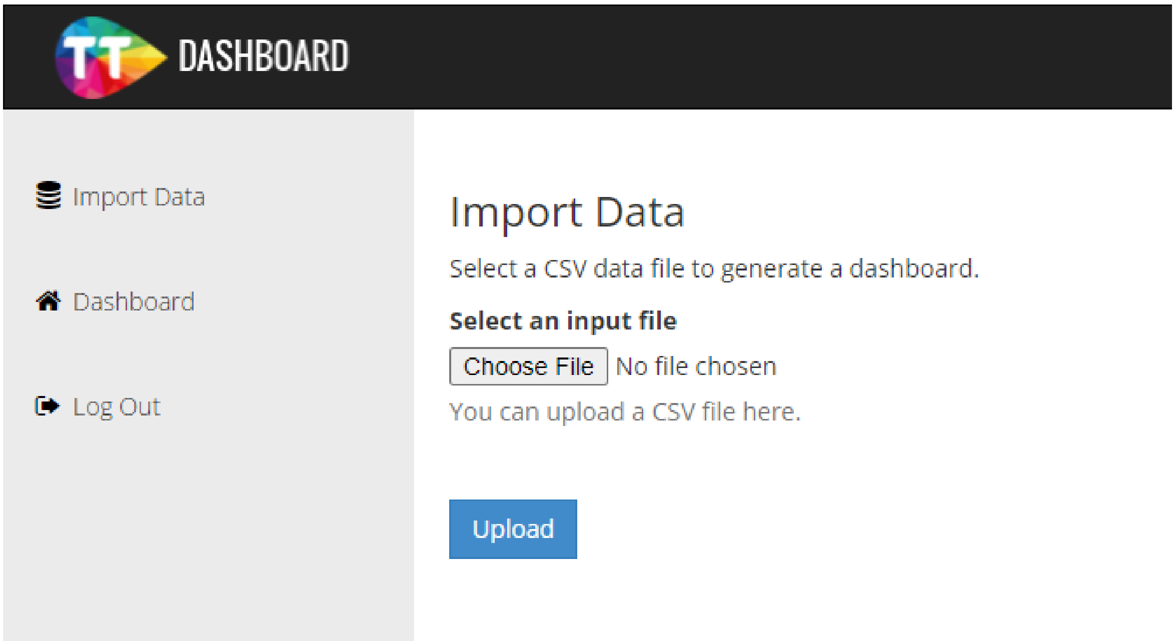


Figure 36: Application Login Page

In the data import menu, we have a select file button. The application accepts only CSV file format. It shows an alert if the file isn't a CSV file and goes to the dashboard menu and presents the visualisations if the file is coherent.



Import Data

Select a CSV data file to generate a dashboard.

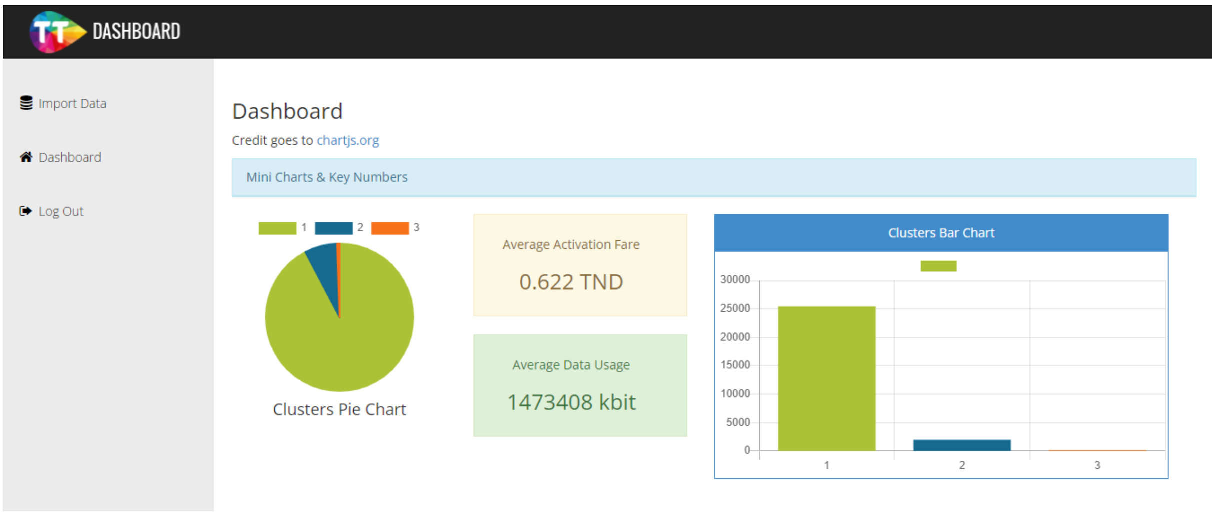
Select an input file

Choose File No file chosen

You can upload a CSV file here.

Figure 37: Data Import Menu

In the Figure 38, we observe clusters pie chart and bar chart as well as the subscriber's average activation fare and average data use.



Dashboard

Credit goes to chartjs.org

Mini Charts & Key Numbers

Average Activation Fare

0.622 TND

Average Data Usage

1473408 kbit

Figure 38: Mini Charts and Key Numbers in the Dashboard Application

In the Figure 39, we have the next section of the dashboard a scatter plot of the clusters in two dimensions, the x-axis represents days of plan activation and the y-axis represents data costs.

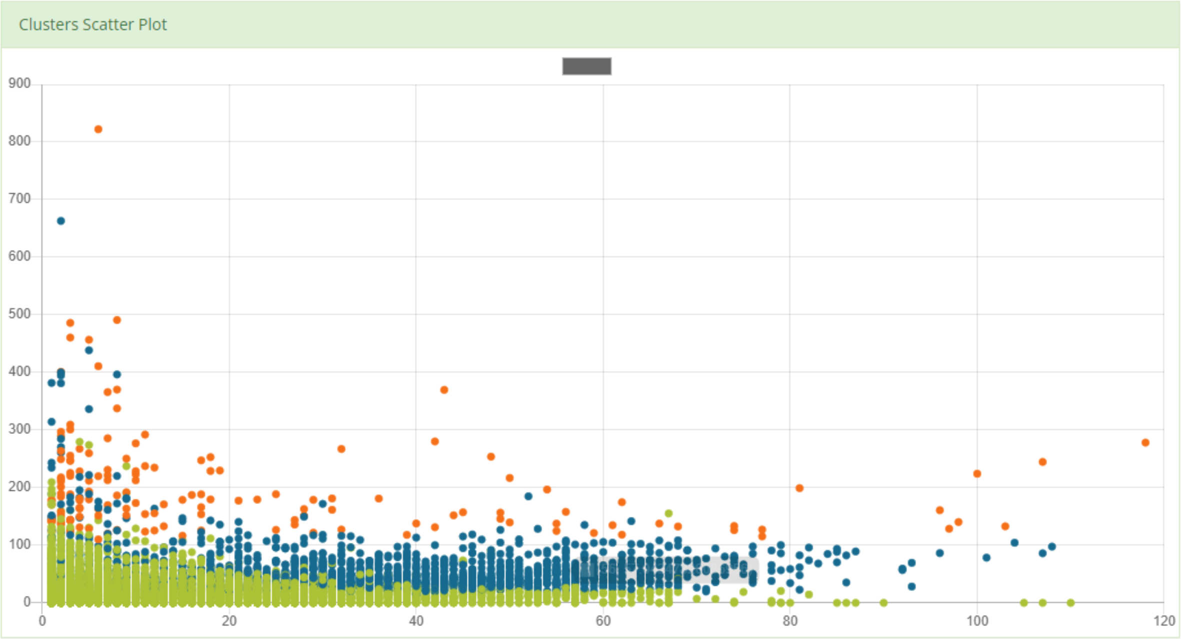


Figure 39: Cluster Scatter Plot in the Dashboard Application

## Conclusion

In this last chapter, we have performed a comparative analysis of the determined segments, we have presented some of the noteworthy charts from our analytics dashboards as well as our visualisation application.

# General Conclusion

Providing better services is a constant preoccupation at Tunisie Telecom. Which is why they collect a large amount of data every day through their networks and systems. This data provides them with detailed information about their customers' behaviour. The business goal is to offer suitable plans for the costumers according to the behavioural activity.

We used a sample of this data the documented the behaviour of 27565 anonymous costumers during the months of December 2016 and January 2017. We applied a consumption-based segmentation of our demographic, we studied the segments and provided multiple visualisations and dashboards as well as a web application that aim to clarify the segmentation criteria and provide a clear understanding of the data, also we generated a classification model that can be used to provide suitable services to the customer according to his class.

In a long-term perspective, geographic and demographic or even psychographic segmentations can be applied in the future if data can be collected, the more data the better. Advanced use of segmentation allows each customer to be part of a micro-segment. This allows precise targeting, with knowledge of what the retention and value drivers are for each customer.

# Appendix

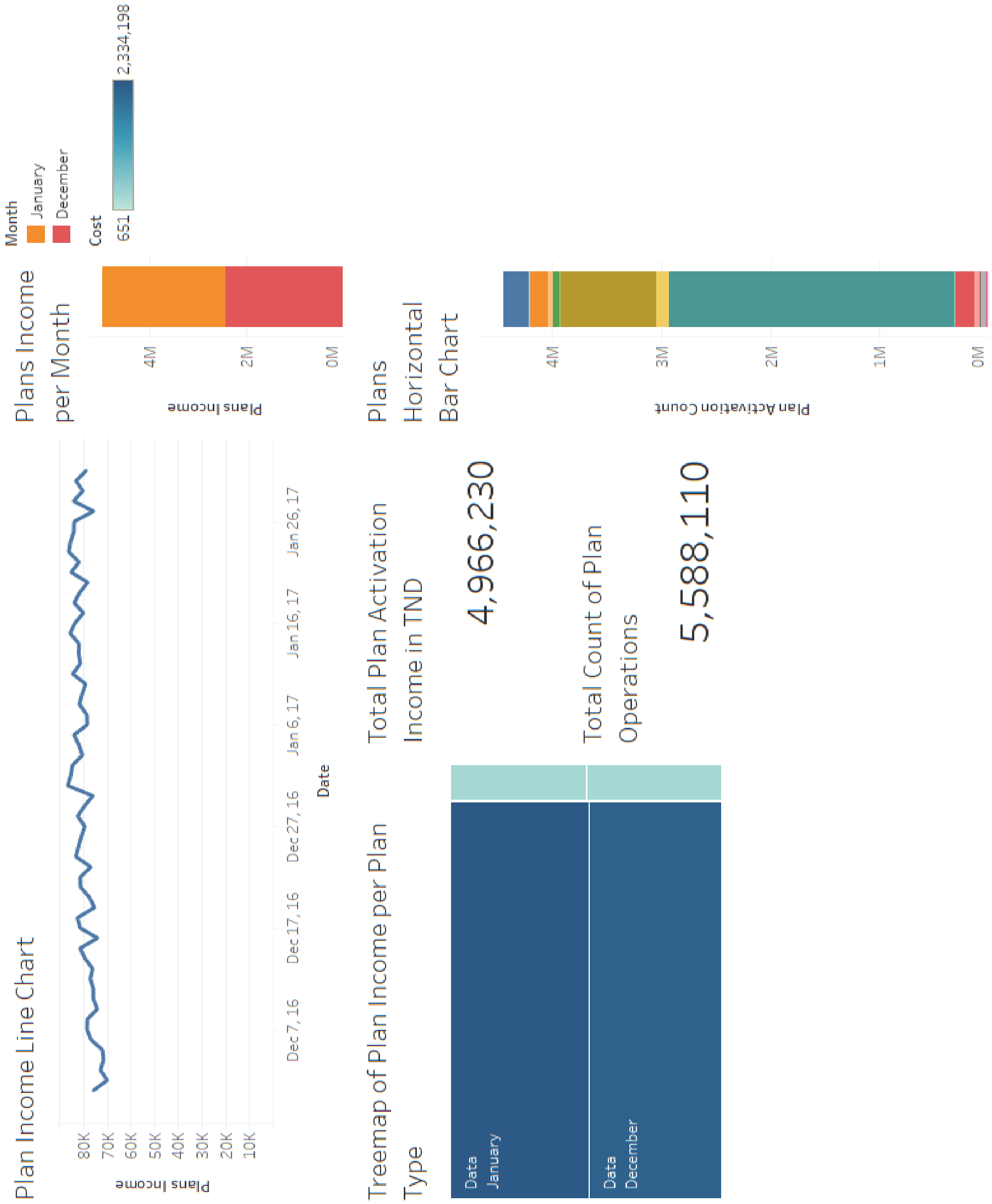


Figure 40: Plan Activity Dashboard

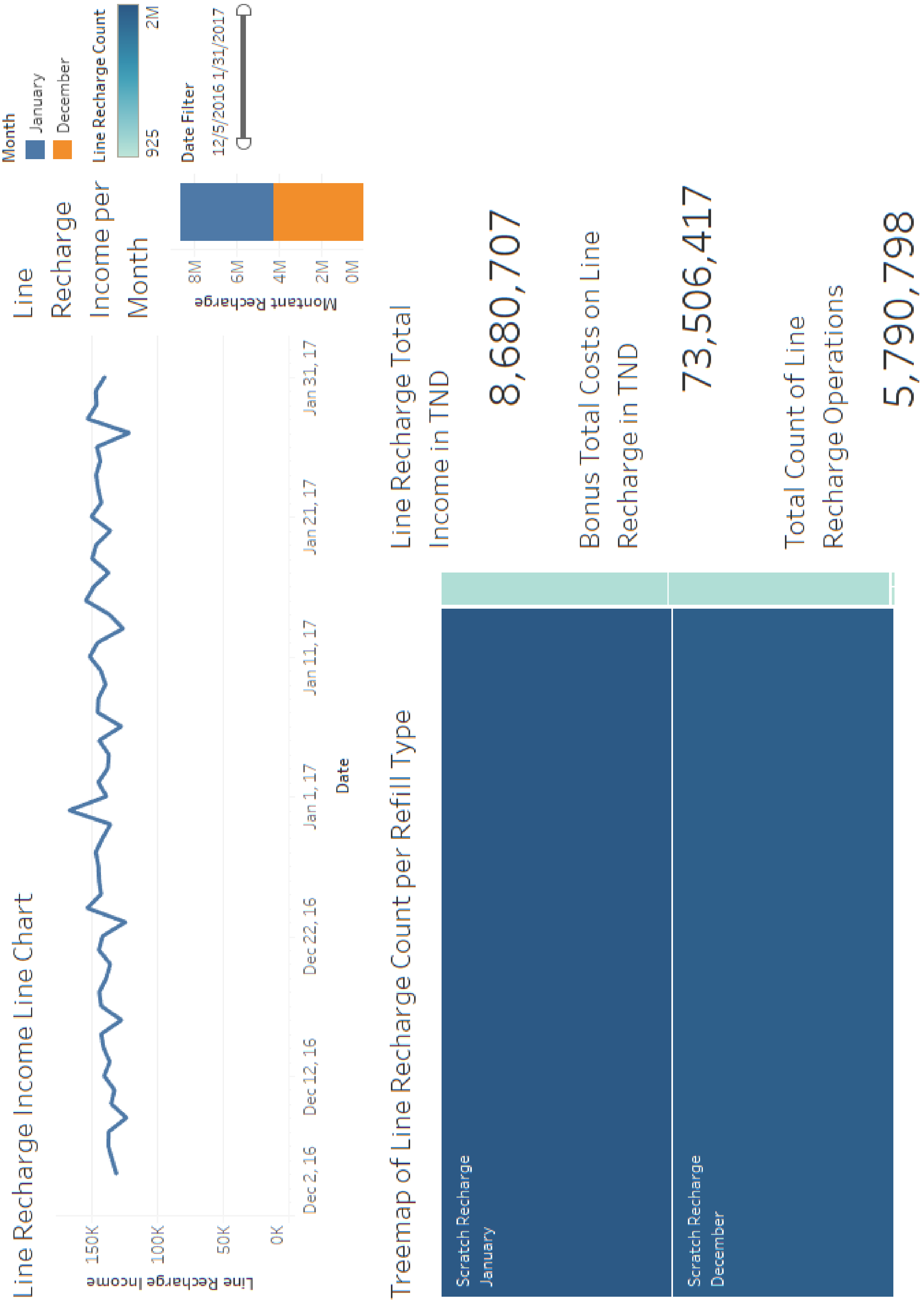


Figure 41: Line Recharge Dashboard

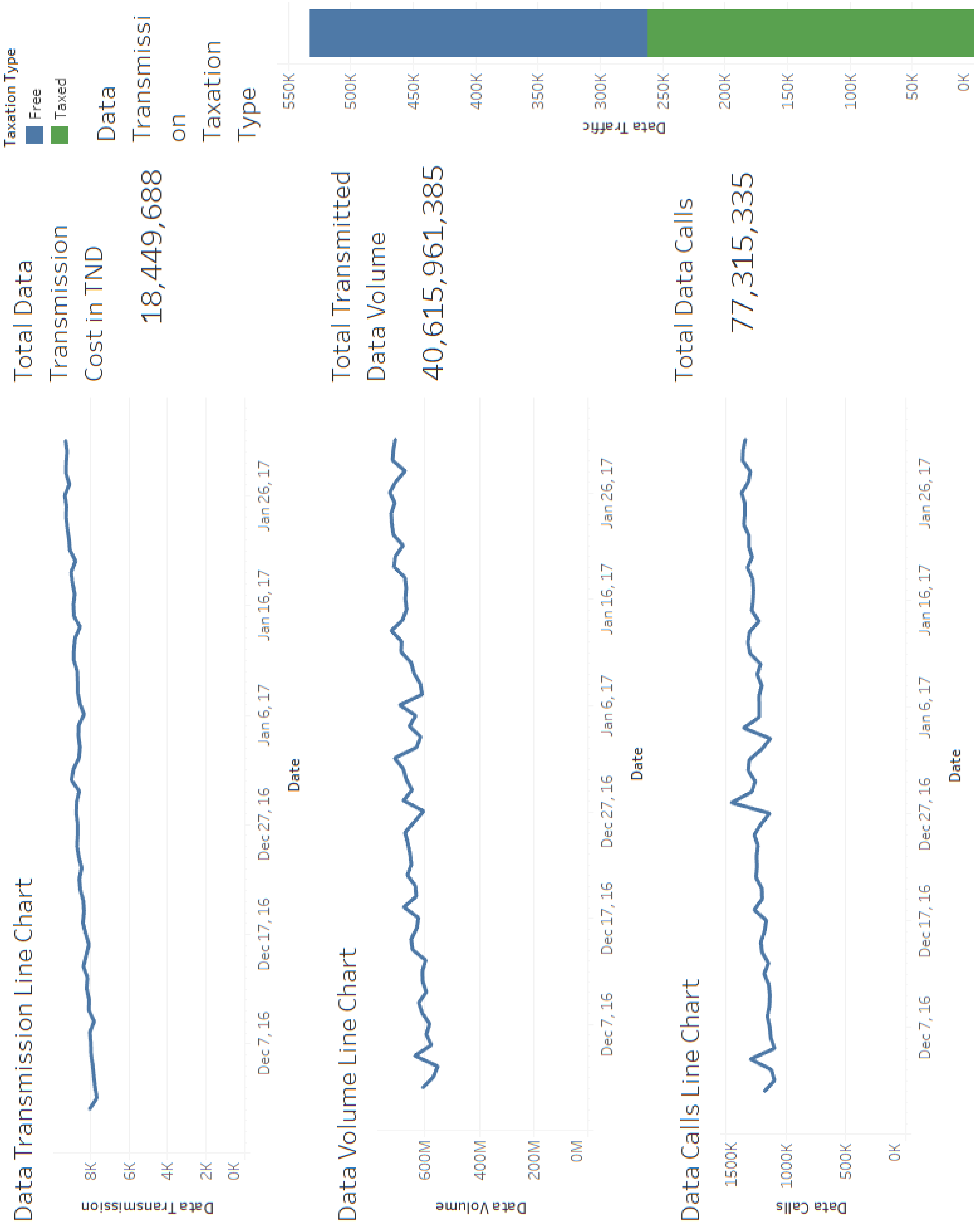


Figure 42: Data Traffic Dashboard

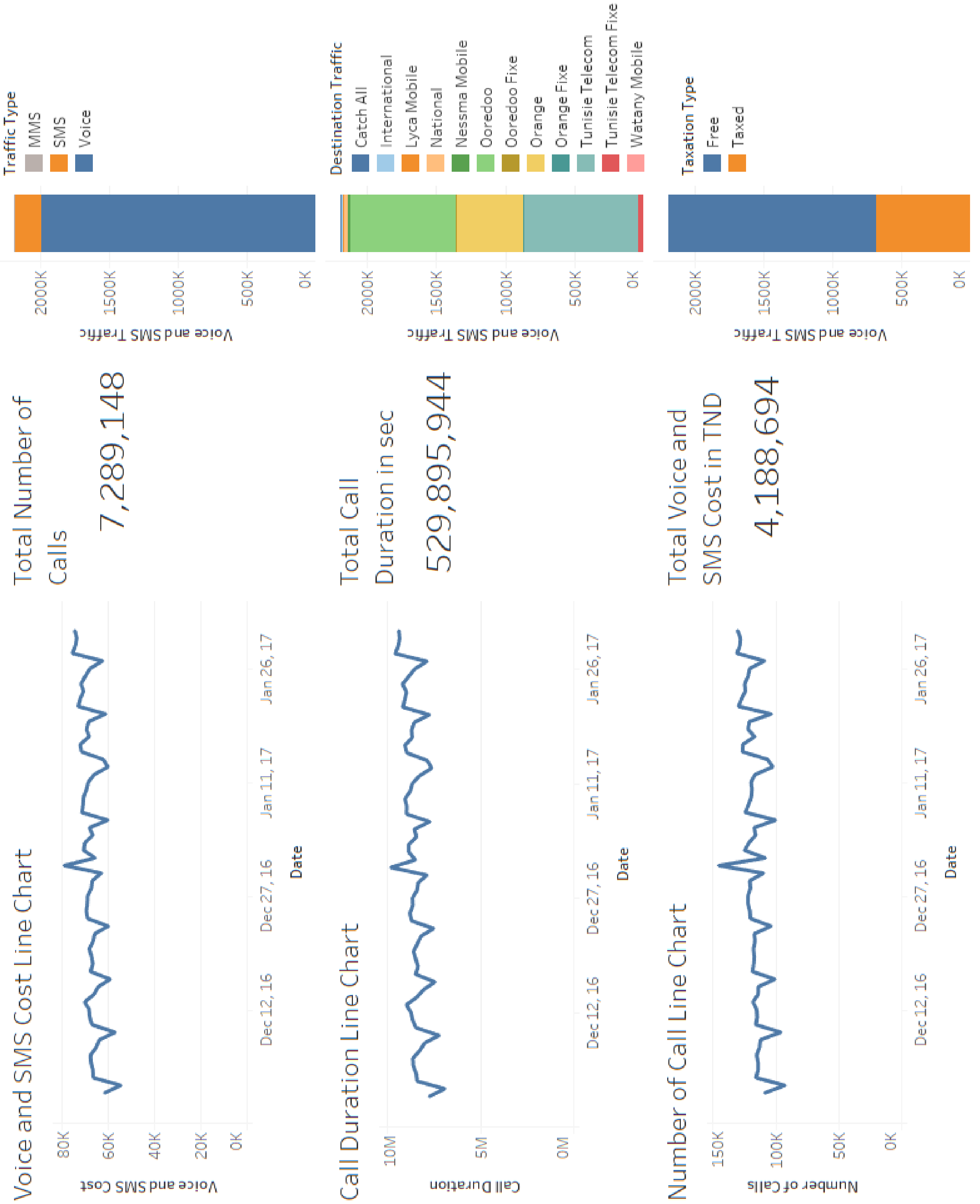


Figure 43: Voice and SMS Traffic Dashboard

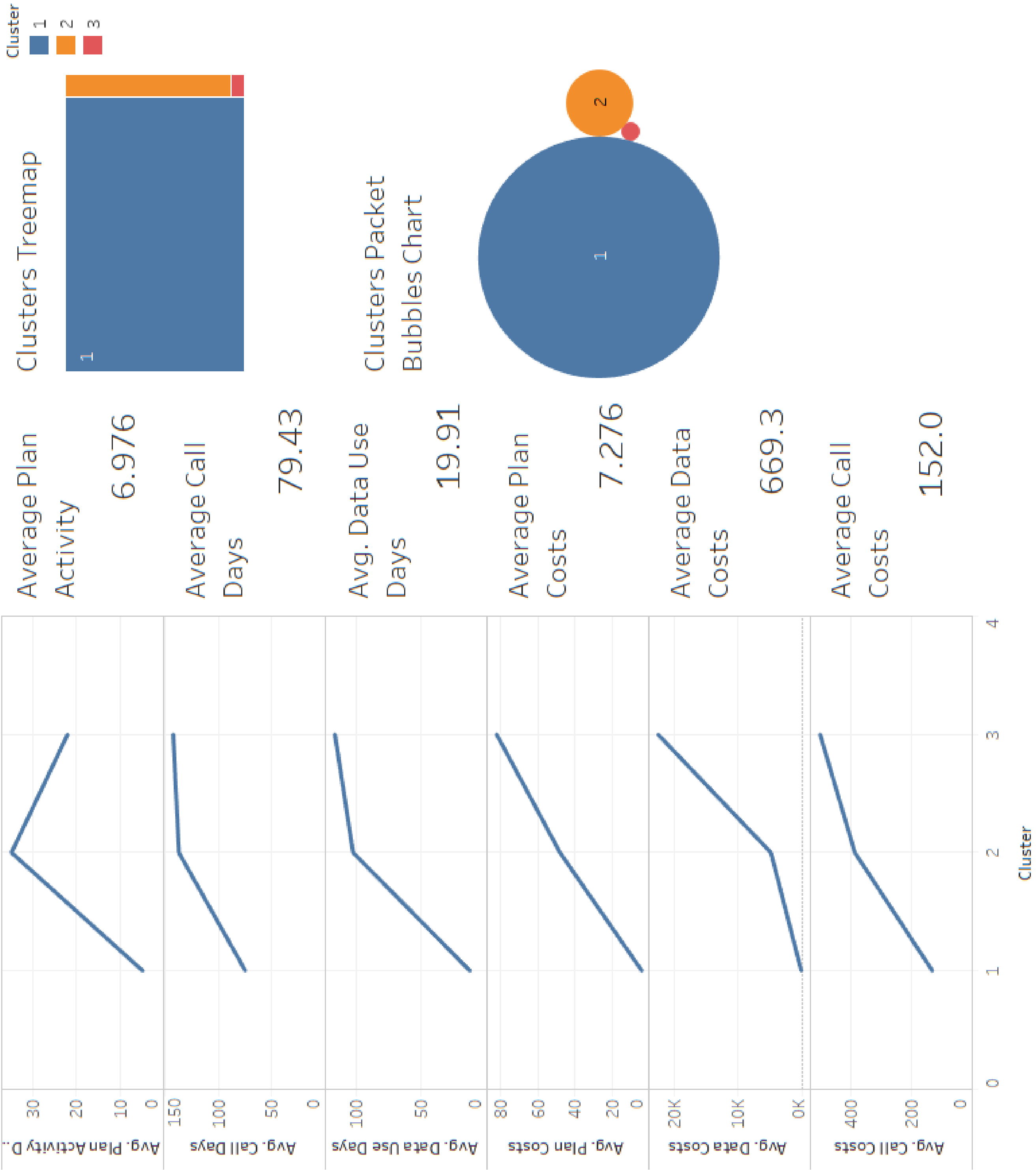


Figure 44: Comparative Cluster Analysis Dashboard

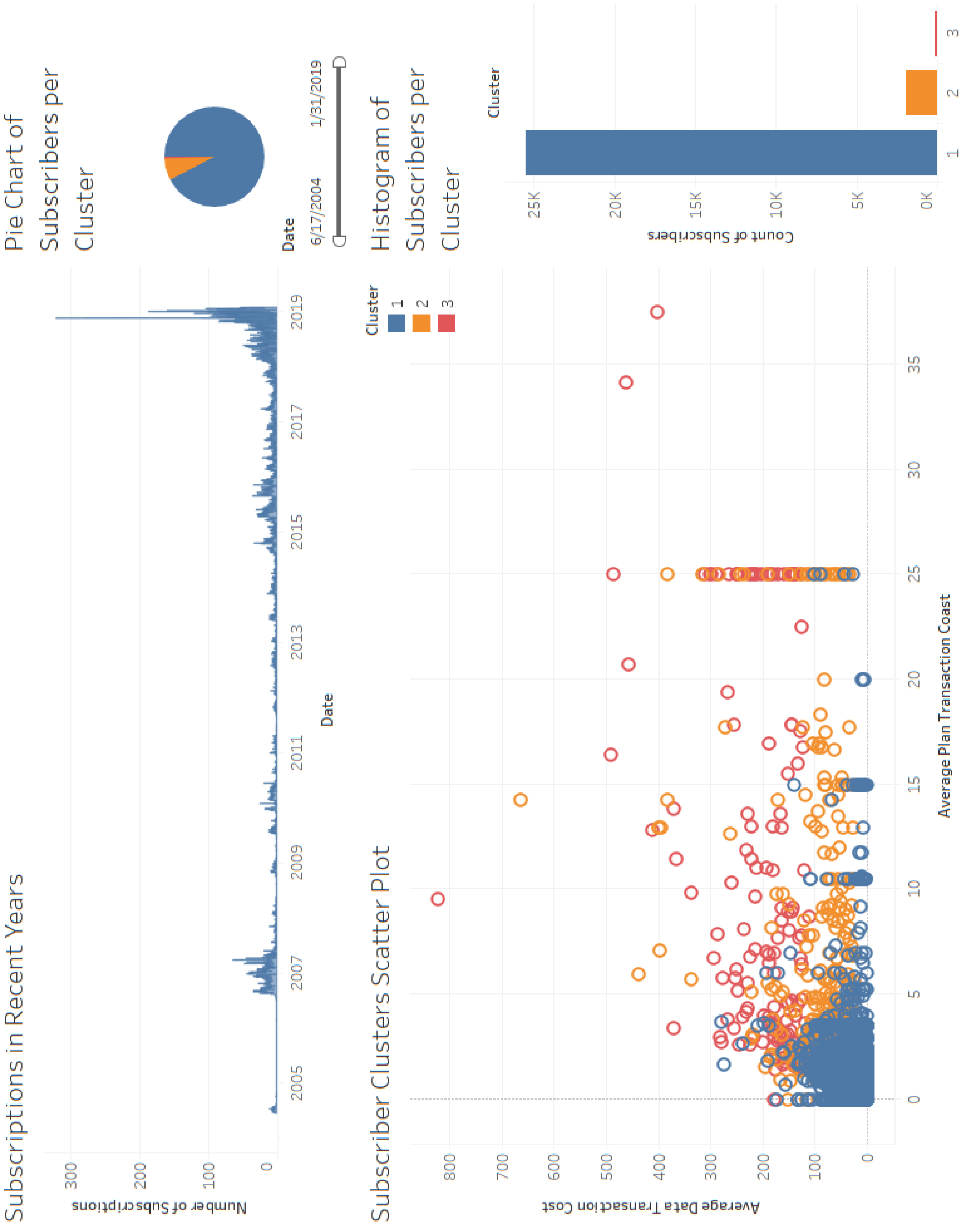


Figure 45: Clusters Dashboard

# References

Alibuhtto, M. C., & Mahat, N. I. (2020). Distance Based k-means Clustering Algorithm for Determining Number of Clusters for High Dimensional Data. *Decision Science Letters*, 52.

Bacila, M. F., Adrian, R., & Ioan, M. L. (2012). *Prepaid Telecom Customer Segmentation Using the K-Mean Algorithm.*

Bayer, & Judy. (2010). Customer Segmentation in the Telecommunication Industry. *Database Marketing & Customer Strategy*, 248.

Bose, I. C. (2010). Exploring Business Opportunities from Mobile Services Data of Customers: An Inter-Cluster Analysis Approach. *Electronic Commerce and Applications*.

Bryan, F., & Merlin, S. (2001). *Successful Customer Relationship Marketing.* London: Kogan Page.

Dangeti, P. (2017). *Statistics for Machine Learning.* Mumbai: Packt Publishing Ltd.

Diller, H. (2000). *Customer Loyalty: Fata Morgana or Realistic Goal? Managing Relationships with Customers In Relationship Marketing: Gaining Competitive Advantage through Customer.* Berlin: Springer-Verlag.

Niarn, Agnes, & Bottomley, P. (2003). Something Approaching Science? Cluster Analysis Procedures in the CRM Era. *Journal of Market Research*, 45.

*Spark 3.3.0 Documentation*. (2022, July). Retrieved from Apache Spark: https://spark.apache.org/

Storbacka, & Kaj. (1997). Segmentation Based on Customer Profitability – Retrospective Analysis of Retail Bank Customer Bases. *Journal of Marketing Management*, 13.

Tsiptsis, K. C. (2009). *Data Mining Tehniques in CRM: Inside Data Mining Tehniques in CRM: Inside*. Chichester: John Wiley & Sons Ltd.

Tunisie Telecom. (2022, July 15). *About Tunisie Telecom*. Retrieved from Tunisie Telecom: https://www.tunisietelecom.tn/

Wedel, Michel, Kamacura, & Wagner, A. (2000). *Market Segmentation: Conceptual and Methodological Foundations.* Boston: Kluver Academic Publishers.

Weiss, G. M. (2022). *Data Mining for Telecommunications.* New York City: Fordham University.

Wirth, H. J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*