This project of Moringa Data science Programs Week 3

In this project we are required to analyse data with CRISP-DM Methodology. We will also use python and Pandas to get insight on the data.

CRISP-DM has the following steps:

1. Business understanding.
2. Data understanding.
3. Data Preparation.
4. Modelling.
5. Evaluation.
6. Deployment.

**Business Understanding**

MTN Cote d'Ivoire would like to upgrade its technology infrastructure for its mobile users in Ivory Coast within the given cities.

It is important for their business and I am expected to be their KPIs.

Considering the upgrading of technology for a service provider, we have to identify the factors that affect the customers indirectly and directly. This project have to discover what is the best strategy to upgrade the technology by answering questions like:

* Determine how many users will be in service when upgrading the technology?
* Following the trend of voice and sms usage,we can determine the favourable time to upgrade the technology without inconveniencing the customers.
* Determine the favourable time for voice calls
* Determine the number of voice call and sms
* Determine the total number of users using sms and voice calls.

**Data Understanding**

We have two data:

* Cell geo data: including city,Geographical zone,Longitude & Latitude,Region,area Id of the cell and site
* CDR data: Products category ie sms or voice,billing price,time format,ID of the cell,Id of the site etc.

Now after identifying our data we load it in our Collab Notebooks to analyze it.

Cell Geo Data

I have discovered that our data has 10 columns.

A notable characteristic of the data is that rows with empty content in the Area columns are recorded as NaN.

This brings a question to understand how many Regions in this dataset are Associated with Area.

We have 18 Areas associated with regions

CDR Data:

The CDR data is in three datasets to ease the load when loading for analysis due to the huge number of the records in the dataset and memory usage.

These datasets have 9 columns,their most frequent data type is Strings.

Their characteristics are:

* Product column contains of voice and sms data categories
* Value consists of billing price
* Date time column consists of date and time.
* Cell\_ID consists of a unique character to identify a specific cell
* Site\_ ID also consists of a unique ID
* The other column consists of irrelevant data to this project.

After understanding the columns in these dataset we Load them in Collab for cleaning and analysis.

**Data Preparation**

Since we cannot use data with null value columns we check and fix those columns.

From cell geo dataset the columns with null values are:

* STATUS with 67 Null values and a percentage of 1.69%
* ZONENAME has 6 null Values 0.15 %
* AREA has 23 null Values 0.58 %

After identifying the above there are columns with data in our dataset that are irrelevant to our project and they need to be removed to make work easier..

The CDR Datasets shall also go through the same process as above

That is cleaning and replacing the missing values with a Zero.

Telecom dataset 1 has a column with null value:

* SITE\_ID has 716 null Values 18.02%

Telecom dataset 2:

* SITE\_ID has 676 null Values 17.01 %

Telecom dataset 3:

* SITE\_ID has 605 null Values 15.22 %