**IP WEEK3 – *link to GitHub*** *(https://github.com/therealellam/My-Moringa.git)*

**BUSINESS UNDERSTANDING**

**Business Understanding Overview**

MTN Côte d’Ivoire (MTNCI), a subsidiary of South Africa-based MTN Group, is an Ivorian telecommunications company that provides post-paid and prepaid GSM services including fixed line, mobile, and internet services.

Usually, an enterprise mobile solution is expansive; meaning that it can be able to provide competitiveness, productivity, and profitability advantages for enterprises that have the capability to successfully deploy a well-conceived and well-executed mobile infrastructure strategy. Growth in the mobility hardware, software, and service markets speaks to how organizations are prioritizing mobility. Strategically aligning technologies and services to leverage as a foundation upon which to build a mobile strategy requires insight and guidance

The main objective of this project is to give recommendations on how MTN Cote d'Ivoire can go about the upgrade of their technology infrastructure strategy within the given cities for its mobile users.

**Desired outputs of the project.**

The primary objective is to recommend MTN Cote d’Ivoire on how they upgrade their technology infrastructure strategy. Some of the related business questions include Which ones were the most used city for the three days? Which cities were the most used during business and home hours? Most used city for the three days?

The project plan included but not limited to

1. Initial assessment of tools and techniques where in this case python was used on google Collaboratory to carry out Data wrangling and analysis

2. Determining Data mining goals and business success criteria. The intended output of the project was to advise MTN Cote d’Ivoire to invest more in cities that use their product

3. Data understanding stage. Data used in this project was provided through the following URLs.

- CDR 20120507 [http://bit.ly/TelecomDataset1]

- CDR 20120508 [http://bit.ly/TelecomDataset2]

- CDR 20120509 [http://bit.ly/TelecomDataset3]

The dataset comprised three CSV files. This stage also included loading, exploratory data analysis and verification of the quality of data provided in terms of completeness, accuracy and if there are missing values.

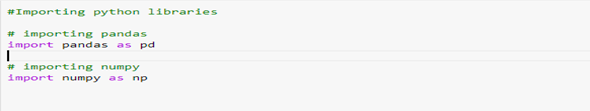
4. Data preparation. As mentioned in the previous stage, the data to be used had been provided which were four data sets. Here, among the activities carried out included data cleaning, integrating and merging datasets.

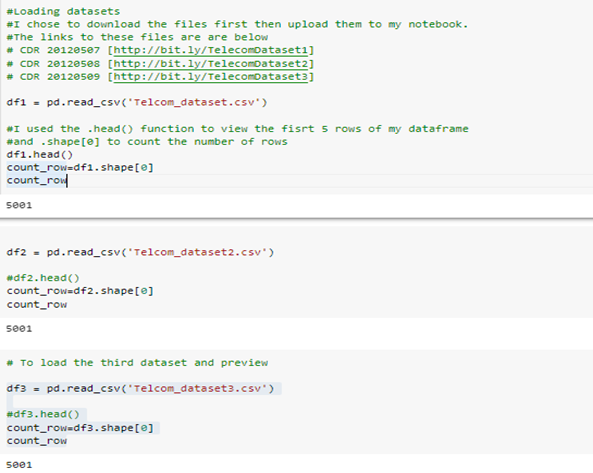
5. Data analysis. This stage included analysing the data to generate insights and stating the assumptions made if any. An assessment of technique according to the data mining success criteria and desired test design was also done

6. Lastly was the evaluation of results and giving recommendations. A thorough review of the data mining engagement was carried out in order to determine if there was any important factor or task that has somehow been overlooked.

**DATA UNDERSTANDING.**

The datasets were downloaded from the provided URL and then uploaded in the Notebook which is easier than loading data directly using URLs. For the CDR datasets (Telcom\_Dataset, Telcom\_Dataset2 & Telcom\_Dataset3) the three datasets had similar columns of data type string and integer only. The code used to read the data is as follows,

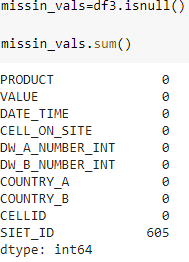
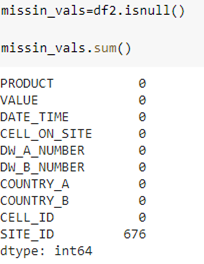
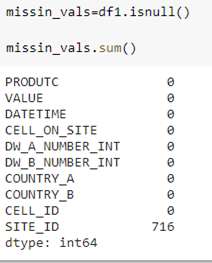




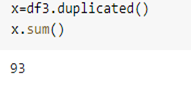
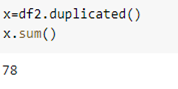
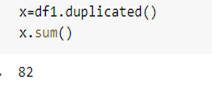
The three CDR datasets had equal number of entries, 5001. The columns were ten in number and their description is as below

| **Column name** | **Description** | **Format** |
| --- | --- | --- |
| PRODUCT | Voice or SMS | String |
| VALUE | Billing price | Integer |
| DATE\_TIME | Time in format yyyy-MM-dd hh:mm: ss.0 | String |
| CELL\_ON\_SITE | Which cell in the site was used (not needed here) | Integer |
| DW\_A\_NUMBER\_INT | Anonymized phone number of the person for which the CELL\_ID and SITE\_ID are given | String |
| DW\_B\_NUMBER\_INT | Anonymized phone number of the counterparty | String |
| COUNTRY\_A | Country of party A (useless here) | String |
| COUNTRY\_B | Country of party B (useless here) | String |
| CELL\_ID | ID of the cell | String |
| SITE\_ID | ID of the SITE | String |

The next step was to check the dataframes for missing values



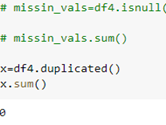
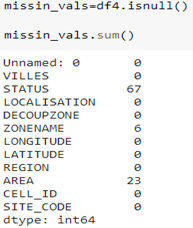
All the 3 dataframes did not have missing values except for the SITE ID column where there was a total of 1,997 missing values.



The dataframes had 82,78,93 duplicates respectively.

For the Cell\_Geo.csv dataset, the dataset has 11 columns and 3974 rows. The description and format of each column is as below

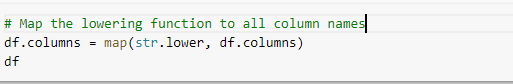
| **Column name** | **Description** | **Format** |
| --- | --- | --- |
| VILLES | City | String |
| STATUS | In Service or not | String |
| LOCALISATION | in ABIDJAN or not | String |
| DECOUPZONE | Geographical Zone | String |
| ZONENAME | Name of Zone | String |
| LONGITUDE | Longitude | Float |
| LATITUDE | Latitude | Float |
| REGION | Region | String |
| AREA | Area | String |
| CELL\_ID | ID of the cell | String |
| SITE\_CODE | Site (there are several cells per site, several sites per City, several Cities per Zone, etc.) | String |



There were 23 missing values in the Area column and 67 on the status column for this dataframe. Also, to note is that there were Zero duplicates in this Dataframe.

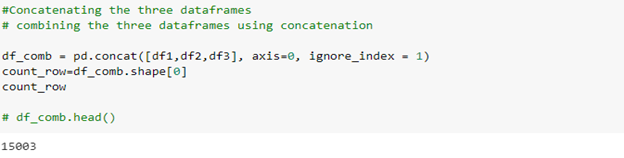
**DATA PREPARATION**

The first step in this stage was to change all the column names for the CDR (Telcom\_Dataset, Telcom\_Dataset2 & Telcom\_Dataset3) Dataframes into lowercase by mapping the lowering function to all column names in respective dataframes.

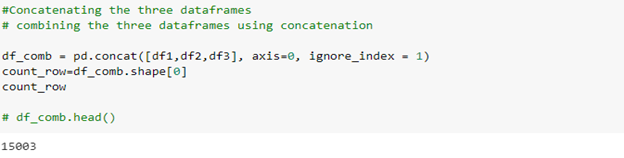


This step is also done on the dataframe created from the Cell\_geo.csv dataset.

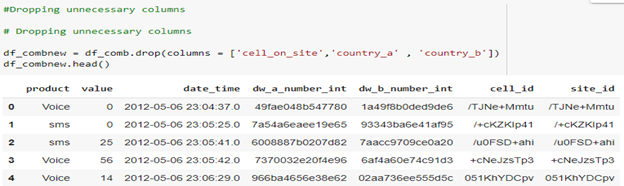
The next procedure carried out was concatenating the three CDR datasets into one. This was done using the code below.



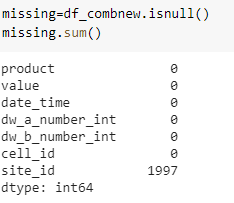
The dataframe *df\_comb* had 10 columns and 15003 rows. Duplicates were then dropped from the Combined dataframe.



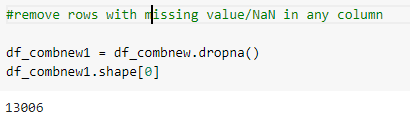
Unnecessary columns that were not needed in this project were removed using the code below.



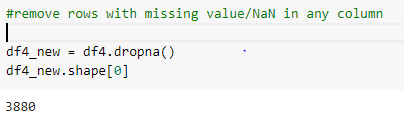
As part of data preparation to ensure the data quality is raised to the level required for analysis, missing values were removed from the combined dataframe as follows,



The column site\_id was the only one with NaN values totaling to 1997 out of the 15003.

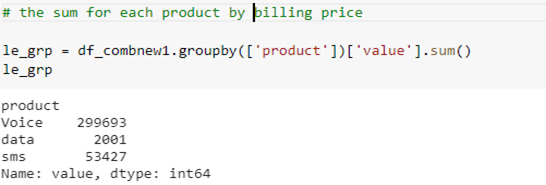


Missing values were also removed from the df4 dataframe which had the *cells\_geo* data.

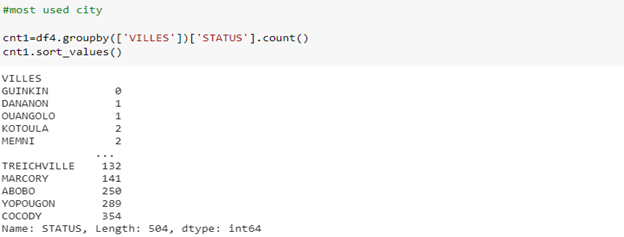


**ANALYSIS**

To understand the product offering by MTN Cote d’Ivoire an analysis of the Product column was done grouping each product by total Billing price. The results showed that the total billing price for Voice, data and SMS were 299,693, 2001 and 53,427 respectively.



All the cities were grouped by cells whose status was ‘In service’ to find out Which city were most used for the three days and the outcome of the analysis showed that COCODY was the most used city.



**RECOMMENDATIONS.**

MTN Cote d’Ivoire should focus their investments in Cocody as it was the most used city for the three days. Most of their infrastructure should go to their voice product since voice has the highest billing income.

**EVALUATION**

The analysis should be re-evaluated to find cities which were most used during business and home hours as merging the two dataframes resulted in missing values. In addition , knowing the total billing price per city would help in pinpointing which cities to focus on based on the income they bring in.