**DATA REPORT**

**Introduction**

The task is a research problem where as a data scientist working for MTN Cote d’Ivoire, a leading telecommunication company in the country, I am required to study the datasets given in order to correctly advise the company on the best strategy to go about the upgrade of its infrastructure in the cities. Below is a documented process that I followed while performing analysis. I used the CRISP-DM methodology to optimize the results during the data mining process. This is a link to the whole documented process on [GitHub](https://github.com/Jenn-mawia/MTN-Data-Analysis).

1. **BUSINESS UNDERSTANDING**

This is where I would seek to get answers to gather information as to why MTN seeks to upgrade its infrastructure. It would also be the best time to analyse the current situation and what improves after an upgrade to the infrastructure. The main task here is to try and gain much insight into the business goals for the data mining process. Some of the questions that I would ask are:

* Why does MTN seek to improve its infrastructure? Is it to keep up with competitors in the market?To improve customer satisfaction?To improve on profits?
* What is the current state of MTN’s infrastructure?Is it up-to-date with the current technology?Does the current infrastructure in place satisfy customers?
* What is the current influx of customers subscribed to MTN’s services?How often do they use the services provided byMTN? What are the locations of these customers?To determine whether there are more or few customers at which areas and at what frequency do these customers use the services from MTN
* How does the upgrade determine profitability in the company? To clarify on the long-term cost-benefit analysis.
* Which kind of upgrade is needed in the infrastructure? It could be on the voice, sms, internet bundles, cabling and fiber optics infrastructure.

All these and many more questions can be asked to get a clear concept of the business goal. Some of these questions will be answered during the data preparation and analysis process.

Therefore, the ***main goal for the project*** is to get insights from the relevant data provided in the datasets to identify the best way in which MTN can improve its infrastructure for its mobile users within the given cities.

1. **DATA UNDERSTANDING**

Data Collection

Data required for the data mining process was readily provided. There were no challenges on data acquisition as I did not conduct any data collection activities.

Data Exploration & Description

I worked with six datasets.

* [cells\_geo](https://drive.google.com/file/d/11OW07ZRMsksnsmbZ32oaEN2MkFfRmNRe/view?usp=sharing) - this dataset contains 3974 records and with 11 attributes. The attributes on this dataset are as described in the “cells\_geo\_description” dataset
* [cells\_geo \_description](https://drive.google.com/a/moringaschool.com/file/d/1-rIM5ihDu79RaH7rAs-d-7SQSAQhrY9N/view?usp=sharing) - this comprises 11 records, with three attributes - Column name, description and format. This dataset describes the data in the “cells\_geo” dataset. The values in the Column name is a list of all attribute names used in the “cells\_geo” dataset. Values in the description column give a brief description of what the column names mean and the kind of data they hold in the “cells\_geo” dataset. The format column contains the datatype of listed column names as string or float.
* [CDR\_description](https://drive.google.com/open?id=1cVoNXl25IO5-_yQk97ThdeqhE6yw8YTD) - this comprises 11 records and with 3 columns. This data gives a brief description of what is contained in the following three datasets. The values in the column “Column name” are viewed as rows in the datasets below. Values in the description column give a brief description of what the column names mean and the kind of data they hold in the 3 datasets containing mobile usage log. The format attribute contains the datatype of listed column names as string or integer.
* [CDR 20120507](http://bit.ly/Telcom_dataset1) - this data comprises 5001 records and a total of 10 columns. The column names of this dataset are as described in the “CDR\_description” dataset. It is a log of MTN’s mobile usage by customers on the date 2012/05/06 and 2012/05/07.
* [CDR 20120508](http://bit.ly/Telcom_dataset2) - this data comprises 5001 records and a total of 10 columns. The column names of this dataset are as described in the “CDR\_description” dataset. It is a log of MTN’s mobile usage by customers on the date 2012/05/07 and 2012/05/08.
* [CDR 20120509](http://bit.ly/Telcom_dataset3) - this data comprises 5001 records and a total of 10 columns. The column names of this dataset are as described in the “CDR\_description” dataset. It is a log of MTN’s mobile usage by customers on the date 2012/05/08 and 2012/05/09.

Data Quality Verification

There were missing data in the various datasets. These values were coded as a non-response($NAME?).

There also existed some coding inconsistencies, especially on the cells\_geo dataset. This data contained information that was not splitted into columns and with delimiters “;”. These delimiters were however not consistent throughout the data. There were also some inconsistencies in the naming conventions of the columns on the three datasets CDR 202120507, CDR 202120508, CDR 202120509. The columns were not named using a consistent format.

Another inconsistency that I observed in the cells\_geo dataset was that the values in each column did not correspond to what it was supposed to have. They were all jumbled up together in the various columns.

1. **DATA PREPARATION**

The activities that I undertook at this stage were mainly aimed at consolidating the data into a meaningful format that can then be used for analysis. I used Python and Microsoft Excel.

Selecting data & Data Cleaning

The first step I took here was to rename the column names in each of the three datasets to be all in a consistent format. On CDR 202120507, I renamed the the attribute “PRODUTC” to “PRODUCT” and “DATETIME” to “DATE\_TIME”. On CDR 202120508, I renamed “DATETIME” to “DATE\_TIME”, “DW\_A\_NUMBER” to “DW\_A\_NUMBER\_INT” and “DW\_B\_NUMBER” to “DW\_B\_NUMBER\_INT”. On CDR 202120509, renamed “SIET\_ID” to “SITE\_ID” and “CELLID” to “CELL\_ID”. This step would make it easier to consolidate all the datasets into one dataframe. I then went ahead to combine the three datasets CDR 202120507, CDR 202120508, CDR 202120509 into one dataframe . This I did using the append function in Python.

On the cells\_geo dataset, I removed the first column which did not have any meaning to the whole dataset. I went on to rename the field “SITE\_CODE” TO “SITE\_ID”. Renaming this would help in combining the new dataframe from the three datasets CDR 202120507, CDR 202120508, CDR 202120509 and the cells\_geo data frame into one new dataframe.

Having the data in one dataframe would make it easier to analyse the data as a whole as opposed to analysing just one at a time.

I then went on to clean the data by deleting unnecessary columns in the data. These were “CELL\_ON\_SITE”, “COUNTRY\_A","COUNTRY\_B", "CELL\_ID\_x", "CELL\_ID\_y"

These fields were particularly not useful and excluding them did not affect future analysis in any way.

1. **ANALYSIS**

I did further analysis on the data and this is what I gathered:

* The most used product in the three days is the sms, followed by the voice.
* The most used cities in the three days are: YOPOUGON, ABOBO, COCODY, ADJAME, KOUMASSI
* The most used city in the three days is : ZUENOULA
* The least used cities in the three days were: NGUINOU, BLANFLA, ROBANGO, KOTOULA, DANANON
* The maximum billed amount was 6750

1. **RECOMMENDATION**

* Since the most used product is the sms, it can be assumed that the infrastructure at place does not handle good connection over voice. It could probably be a 2G infrastructure which limits users from using the voice product. Recommendation here is to upgrade infrastructure to a high-value network that ensures good connections and is much faster such as the 4G/LTE technology.
* The most used cities were : : YOPOUGON, ABOBO, COCODY, ADJAME, KOUMASSI while the least used cities were: NGUINOU, BLANFLA, ROBANGO, KOTOULA, DANANON. It is highly likely that in the least used cities, the infrastructure is not well built as in the most used cities. The best recommendation here would be to upgrade infrastructure in these areas to enable more mobile user connection.

1. **EVALUATION**

This data mining project is considered a success because the business objectives were met. The recommendations provided will be the best strategy that MTN can use to improve their infrastructure as it focuses on the areas that need most improvement. This strategy deems to be satisfactory in meeting the company’s current needs and should therefore be implemented by first upgrading infrastructure in the least used cities then onto the most used cities.