**MTN Technology Infrastructure Upgrade**

**Business Understanding**

Business understanding requires one to understand the business objectives so as to uncover the important factors that may influence the outcome of the project.

**Project Definition**

In this project as a data scientist i have been tasked with solving how MTN Cote d’Ivoire would go about upgrading of its infrastructure strategy within the given cities

**Business Objectives**

The objective of the project is to analyse the information in the dataset and give the best strategy to MTN Cote d’Ivoire to go about upgrading their infrastructure.

To achieve the main objective i had to look into the following questions:

* Which ones were the most used city for the three days?
* Which hours were the most used during the business or home hours?
* Most used city for the three days?
* Most used time in the cities?

**Assess Current Situation**

For this process to be successful and give the most favourable solution from the research as a Data scientist i require Data,Computer resources and softwares,Evaluate the current constraints and Be dedicated to the process.

Data is provided By MTN Cote d’Ivoire Telecom Company in the form of datasets.

Computer resources and software to be used in this project includes Python programming language Pandas , Numpy and Matplotlib Libraries for data preparation and analysis.

Constraints encountered in this project include datasets being too Large for analysis,computer power to process the data,enough time to analyse the data Thoroughly.

**Data mining Goals**

**Project Plan**

|  |  |  |
| --- | --- | --- |
| **Phases** | **Steps to follow** | **Time Mngt** |
| **Data Sourcing** | The data to be used is provided by the Company MTN Cote d’ivoire |  |
| **Data Preparation** | * Loading the datasets * Selecting the attributes and viewing information of the dataset. * Data cleaning i.e Dropping duplicated records,replacing NA value with Relevant replacement. * Integrating and formatting data |  |
| **Analysis and Evaluation** | Answering the main objective of the research question |  |
| **Conclusion and Recommendation** | Giving the appropriate recommendation to the client according to the findings from the analysis. |  |

**Data Understanding**

In this Phase of the process we explore the dataset and understand the dataset.

The dataset in this project is provided in csv files and the include :

* Cell\_geo.csv
* Telecom\_dataset.csv
* Telecom\_dataset2.csv
* Telecom\_dataset3.csv

The cell\_geo.csv provides the:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Villes | city |
| Status | In service or not |
| Localisation | In Abidjan or not |
| Decoupzone | Geographical Zone |
| Zonename | Name of Zone |
| Longitude | Longitude |
| Latitude | Latitude |
| Region | Region |
| Area | area |
| CELL\_ID | ID of the cell |
| SITE\_CODE | site |

The telecom datasets provides the following information:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Product | Voice and Sms |
| Value | Billing price |
| Date\_time | Time format |
| Cell in site | Which cell in the site was used |
| DW\_A\_number Int | Anonymized Phone Number |
| DW\_B\_number int | Anonymized phone number of the counterparty |
| Country\_A | Country part A |
| Country\_B | Country part B |
| CELL\_ID | ID of the Cell |
| SITE\_ID | ID of the Site |

**Data Verification**

After Keenly looking at the data there were some null values that required to be filled to make the data set complete,the null values in the Cells\_geo.csv datasets in the status column meant that the cell was offline/out of service,the null values in the area column of the same dataset meant that the cells came from anonymized phone number.

By verifying that,it will help to continue to the next phase of the process.

**Data Preparation**

**Data Selection.**

In this research i have four datasets;the first dataset describes the cell geographical location while the other three have similar data.

Cell\_geo.csv dataset has 3974 entries and 11 columns of which four are useful for our research.

Telecom datasets have 5001 entries and 10 columns

**Data Cleaning**

Data cleaning in this project involves checking for null values in the datasets and replacing them with the appropriate value and dropping irrelevant columns.

In the Cell\_geo.csv Dataset I dropped 7 columns that had irrelevant data that was not relevant in this analysis,filled the null values that were in status and Area columns and also changed the name of a column that was wrongly spelt. I then converted the column into lower cases

In the Telecom dataset i dropped 6 columns,rename misspelled column name and checked for the null values

The clean datasets are as follows:

Cell\_geo:

df\_cell.info()

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 cities 3974 non-null object

1 status 3974 non-null object

2 region 3974 non-null object

3 area 3974 non-null object

Telecom Dataset

telco.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 5001 entries, 0 to 5000

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 product 5001 non-null object

1 value 5001 non-null int64

2 datetime 5001 non-null object

3 cell\_on\_site 5001 non-null object

**Analysis**

After cleaning the dataset the data can be used for analysis that will help to draw insight to help guide the company on strategizing on upgrading their infrastructures.

The following are one of most analysis that will help draw an insight:

a). The most used cities in the period of the three days.

df\_cell['cities'].value\_counts()

COCODY 354

YOPOUGON 295

ABOBO 255

MARCORY 141

TREICHVILLE 132

...

SEGUELON 3

MEMNI 2

KOTOULA 2

OUANGOLO 1

DANANON 1

Name: cities, Length: 504, dtype: int64

Cocody City is the most used city from our dataset

b). The common hours of usage are at midnight between 0000hrs and 0100hrs which are

telco['datetime'].value\_counts()

2012-05-07 00:00:10.0 71

2012-05-07 00:00:05.0 61

2012-05-07 00:00:06.0 60

2012-05-07 00:01:41.0 53

2012-05-07 00:00:07.0 52

..

2012-05-06 23:50:29.0 1

2012-05-06 23:37:09.0 1

2012-05-06 23:52:24.0 1

2012-05-06 23:35:11.0 1

2012-05-06 23:45:22.0 1

Name: datetime, Length: 708, dtype: int64

Home hours and this occurred on date 7th of the month.

**Recommendations**

Within those three days of data collection the datasets insight is that the most used city is Cocody meaning that if the company was to upgrade their infrastructure the best city to start with would be the city of Dananon since fewer or close to none usage,which means less inconvenience to users.

The best time also to conduct the upgrade would also be at around 23:30hrs on date 6th of the month at home hours since less users are not using their services at that time.

**Conclusion**

After drawing the insight from the cleaned data it is right to conclude that the data was sufficient to determine the city with most users,time and date at which the usage was at highest,however the data collected was insufficient to draw insight for long-term decision making. The dataset would have covered more relevant details relating to the subject of the collection.

Collab notebook that helped to conduct this processes are at: [github repo Ip3](https://github.com/ArthackA/IPweek3.git)