**Sub:** Data quality issues and observations

Hi,

After receiving all the 3 data-sets from your side, I have performed a data quality check as per our set-defined quality parameters.

In this mail, I have focused fully on the **data quality**, **observations**, **identifications** which had been made regarding the data and the **strategies** which we would be taking to overcome the issues.

I would request you to go through all the **7 points** mentioned below and get back with you views and opinions about them.

* The **“NewCustomerList”** data needs **“customer\_id”** for each new customers.

The reason behind this demand is because **“customer\_id”** is an unique reference to each customers.We might not be able to fetch their transaction details from the DB if their ID’s are unknown.

* There are some incorrect data in **“CustomerDemographic”** data. The **DOB** for **“customer\_id”: 34** is **1843-12-21.**

A human being is highly unlikely to be around 174 years and can’t fall under the **“deceased\_indicator”: N** at the same time. We would assume the DOB values to be correct for further analysis.

* The basis of classifying customers according to their wealth is unclear. A detailed understanding on this feature is required.

Taking an example of **“customer\_id”: 108,** the customer is marked as **“High Net Worth”** under the **“wealth\_segment”** column. The customer is seen to fall under a valuation bin of **8** whereas the highest class value for **“property\_valuation”** in the **“CustomerAddress”** data-sheet is **12.** That customer doesn’t own a car as well as per **“owns\_car”** data in **“CustomerDemographic”** sheet.

* The **“default”** column in **“CustomerDemographic”** contains weird encoded characters.

They are irrelevant and are not useful at all to solve the desired problem.

* The **“gender”** column in **“CustomerDemographic**” data has to be modified.

We will replace the **M** and **F** class with **Male** and **Female** respectively in order to follow the **“gender”** format of **“NewCustomerList”** data-sheet.

* The column **“deceased\_indicator”** in **“CustomerDemographic”** and **“NewCustomerList”** datahas only one class **N**.

As there is no variation in the column in terms of values, we will not consider this column.

* Some of the data-points are **missing**. Missing data when in a small amount can be handled by replacement method. As observed, there is a negligible amount of missing data-points as compared to the total number of data-points as a whole.

The columns which has negligible amount of missing points are not included in the table below.

|  |  |
| --- | --- |
| **Column Name** | **Missing values in percentage** |
| DOB | 2.18% |
| job\_title | 16.65% |
| job\_industry\_category | 16.40% |
| tenure | 2.18% |

Thank You

Regards

Adipta Biswas

Virtual Intern @KPMG