**Sub:** Data quality issues and observations

Hi,

After receiving all three datasets from your side, I have performed a data quality check based on our predefined quality parameters. In this email, I have focused entirely on data quality, observations, and identified issues with the data. I have also outlined the strategies we will be implementing to address these issues.

Please take a moment to review the seven points mentioned below and provide your views and opinions on them:

* The "**NewCustomerList**" data requires a "**customer\_id**" for each new customer. This is necessary as the "**customer\_id**" serves as a unique reference for each customer, and without it, we may not be able to retrieve their transaction details from the database.
* There are incorrect data entries in the "**CustomerDemographic**" data. For example, the DOB for "**customer\_id**" 33 is recorded as 1843-12-21. It is highly unlikely for a person to be around 179 years old and still marked as "**deceased\_indicator**": N. We will assume that the DOB values need to be corrected for further analysis.
* The basis for classifying customers according to their wealth is unclear. We require a detailed understanding of this feature to properly interpret and analyse the data.
  + Taking "**customer\_id**" 108 as an example, the customer is marked as "High Net Worth" under the "**wealth\_segment**" column. However, they fall under a valuation bin of 8, while the highest class value for "**property\_valuation**" in the "**CustomerAddress**" data sheet is 12. Additionally, this customer does not own a car, as indicated in the "owns\_car" data in the "**CustomerDemographic**" datasheet.

| **customer\_id** | **first\_name** | **last\_name** | **gender** | **past\_3\_years\_bike\_related\_purchases** | **DOB** | **job\_title** | **job\_industry\_category** | **wealth\_segment** | **deceased\_indicator** | **owns\_car** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 108 | Kayle | Mingaud | Female | 4 | 1994-03-14 |  | n/a | High Net Worth | N | No |

* The "**gender**" column in the "**CustomerDemographic**" data needs to be modified. We will replace the M and F classes with Male and Female, respectively, to align with the gender format in the "**NewCustomerList**" datasheet.
* The "**defaul**t" column in the "**CustomerDemographic**" data contains irrelevant and unintelligible encoded characters. These characters do not provide any useful information for solving the problem at hand.
* The "**deceased\_indicator**" column in both the "CustomerDemographic" and "**NewCustomerList**" data has only one class, N. Since there is no variation in the column's values, we will not consider it for our analysis.
* Lastly, some of the data points are missing. When the amount of missing data is small, it can be handled using replacement methods. We have observed a negligible amount of missing data points compared to the total number of data points. The table below shows the missing points:

| **Sheet Name** | **Column Name** | **Blank Cells** |
| --- | --- | --- |
| Transactions | online\_order | 360 |
| brand | 197 |
| product\_line | 197 |
| product\_class | 197 |
| product\_size | 197 |
| standard\_cost | 197 |
| product\_first\_sold\_date | 197 |
| NewCustomerList | last\_name | 29 |
| DOB | 17 |
| job\_title | 106 |
| job\_industry\_category | 165 |
| CustomerDemographic | last\_name | 125 |
| DOB | 87 |
| job\_title | 506 |
| job\_industry\_category | 656 |
| default | 302 |
| owns\_car | 0 |
| tenure | 87 |
| deceased\_indicator | 0 |

Please review the information provided and share your feedback on these points.

Thank You

Regards

Sachin Rana

Virtual Intern @KPMG