Hello [Client name],

This is Qinghua Liu from KPMG Data Analytics (Virtual Internship) team. We have reviewed the three datasets provided by your company and summarized the statistics in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | No. of records | No. of columns | No. of unique customer IDs |
| Customer Demographic | 4,000 | 13 | 4,000 |
| Transaction Data | 20,000 | 13 | 3,495 |
| Customer Address | 3,999 | 6 | 3,999 |

Meanwhile, we have identified several data quality issues regarding the completeness, consistency, and validity of the records, and are listed as follows, with suggestions of methods to mitigate these issues and optimize the analysis of developing marketing strategy.

* Many fields, such as last name, DOB, job title, online order, are incomplete in Customer Demographic table and Transaction table. We suggest removing the rows with empty values if the amount is considerably small otherwise using data distribution to estimate the blank.
* The ‘customer\_id’ in the Transaction table and in Customer Address table contains customer IDs that are not recorded in Customer Demographic table, which suggests that these tables have inconsistent records, which requires to further check the date when data recorded. For the moment, we are only considering customers included in the Customer Demographic table, if you have updated your customer list, please let us know immediately so we can use the latest data and get more accurate analysis results.
* A value of 1843 is recorded under DOB column in the Customer Demographic table, which is obviously an error. Besides, given some customers have recorded to have been deceased, we reckon they are less influential to our analysis for a new marketing strategy, hence we will remove these rows in the data cleaning process.
* There are duplicate first names in the table. We can assign a default last name for the missing values so that they are clearly recording different individuals.
* There are a few spelling mistakes and inconsistency in columns like ‘gender’ and ‘state’. We have revised them by replacing miss-spellings and abbreviations to the full spelling versions.
* The ‘list price’ and ‘standard cost’ columns have values in inconsistent format. We have format all values under ‘list price’ to number, and all values under ‘standard cost’ to. currency, ensuring the data consistency for analysis.

We will follow the mitigation mentioned above and start data cleaning, data processing and analyzing process. Any assumptions or further operations on the data will be well-documented and emailed you at every stage of our work.

Please feel free to contact us if you have any questions.

Regards,

Qinghua Liu (KPMG Data Analytics Team)