Hi,

This is Zefan Feng who is from KPMG Analytics Team. We have reviewed the data set provided and found some places where the data quality can be improved during our data quality assessment.

The data quality of the dataset is key for the analysis, poor quality may skew the analysis results. Please refer to the following table and IPYNB file ‘KPMG Module 1’ for the issues and mitigations to improve the data quality.

**Transactions**

|  |  |  |
| --- | --- | --- |
| **Issue** | **Mitigation** | **note** |
| Data type of column ‘product\_first\_sold\_date’ | Change from int to date | we should refer to the author on how the data is collected since the transactions all happened in the same day at different times which is not accurate. |
| Null values in some columns | Drop or fill with value | We should drop blank online\_orders which may be incomplete or fake orders. |
|  |  |  |

**Customer\_demographic**

|  |  |  |
| --- | --- | --- |
| **Issue** | **Mitigation** | **note** |
| Invalid values in ‘default’ | Drop columns | validity |
| Missing values in ’last\_name’ | Assign a uniform last name |  |
| Missing values in ‘DOB’ | Assign a mode year value |  |
| Missing values in ‘tenue’ | Assign the mean value | We can assign the mean value of the valid values to blank to assure consistency. |
| Missing values in ‘job\_title’ and ‘job\_category’ | Drop or assign values |  |
| Inaccurate data in ‘DOB’ | Drop or assign a mode year | The oldest customer is 187 years old which is incorrect. |
| Inconsistent values in ‘gender’ | Replace values into consistent ones | ‘F’ and ‘Female’ refer to the same gender |

**Customer\_address**

|  |  |  |
| --- | --- | --- |
| **Issue** | **Mitigation** | **note** |
| Inconsistent values in ‘state’ | Replace values into consistent ones | State have both full name and abbr refer to the same ones |
| Contains customer\_id not in table ‘customer\_demographic’ | Lest Join table and drop outliers to ensure all data are collected in the same period. | This indicates data may not collected in the same period and skew the results. |
|  |  |  |

**New\_customer\_lists**

|  |  |  |
| --- | --- | --- |
| **Issue** | **Mitigation** | **note** |
| Unnamed columns | Drop columns | We should drop unnamed columns |
| Missing values in ‘last\_name’ | Assign a uniform last name |  |
| Missing values | Drop or assign values |  |
| Lack of customer\_id | Assign consistent id from 4001 after customer\_id in ‘customer\_demographic’ | Assign id to better analyze |

Best,

Zefan Feng