Dear John Doe,

This is Qianwen Qian from KPMG Data Analytics (Virtual Internship) team. Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

SUMMARY TABLE:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer Demographic | Age, DOB | Job Title |  | Decreased Customers | Default column |  |
| Customer Address |  | Customer id | States |  |  |  |
| Transactions | Profit | Customer id, Online order, brand |  |  |  | List price |

Below are more in-depth descriptions of data quality issues discovered and methods of mitigation used. Following recommendations will improve accuracy of data used to influence business decisions of your company in the future.

**Accuracy**

* **DOB was inaccurate for ‘Customer Demographic’ and missing an age\_column; missing a profit column for ‘Transactions’**

*Mitigation:* Filter out outlier in DOB

*Recommendation:* Creating additional columns for age and profit will allow for easier identification of errors. The profit\_column will assist in future monetary analysis

**Completeness**

* **Additional customer\_ids were inconsistent among three datasets**

*Mitigation:* Filter all customer\_ids from 1-3500, filter all the blanks job title, online order, brand as ‘Unknow’

*Recommendation:* merge all the datasets in to one spreadsheet to check the incomplete data. Because incomplete data analysis results may be skewed. Kept other blank data, because some customers leave the job title blanks for private purpose which is not a missing value.

**Consistency**

* **Gender and states are inconsistency for ‘Customer Demographic’ and ‘Customer Address’ respectively**

*Mitigation:* Filter all ‘M’ as ‘Male’, ‘F’ as ‘Female’ and ‘U’ as ‘Unknow’. Filter all ‘New South Wales’ as ‘NSW’ and ‘Victoria’ as ‘VIC’ for states.

*Recommendation:* Minimizes manual entry and human errors.

**Currency**

* **There are two ‘Y’ in deceased\_indicator means they are not current customers.**

*Mitigation:* Filter out there two customers

*Recommendation:* Removing them from datasets can focus on accurate customers analysis

**Relevancy**

* **There are some wired messages in the ‘Customer Demographic’**

*Mitigation:* Delete the default column

**Validity**

* **Format of list\_price for ‘transactions**

*Mitigation:* format list\_price to currency

That summaries all data quality issues discovered through the first stages of the data quality analysis. Please let us know if you have any question regarding mitigation or data quality issues.

Best Regards,

Qianwen Qian