Dear John Smith,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The summary table below highlights the key quality issues that was discovered in the three datasets. Please let me know if you have any questions about the issues.

Summary Table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer Demographic | **DOB**: inaccurate | **Job title**: blanks  **Customer id**: incomplete  **DOB**: Blanks | **Gender**:  inconsistency | **Deceased customers**: filtered out | **Default column**: deleted |  |
| Customer Address |  | **Customer id**: incomplete | **States**: inconsistency |  |  |  |
| Transactions | **Profit**: missing | **Customer id**: incomplete  **Online order**: blanks  **Brand**: blanks |  |  | **Order status**: Removed cancelled orders | **List Price**: format  **Product first sold date**: format |

Below are more in-depth descriptions of data quality issues discovered and methods used to resolve them. Recommendations have also been included to avoid further data quality issues in the future. Following recommendations will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

**Accuracy Issues:**

* DOB was inaccurate for Customer Demographic dataset and was missing an age column.
* Transactions dataset was missing a profit column.

Solutions:

* Filtered out outlier DOB.

Recommendations:

* Create an age column, allowing for more comprehensible data and easier to check for errors.
* Create a profit column to check accuracy of sales.
* Create dropdown options to easier verify there are no accuracy issues.

**Completeness:**

* Customer ids were incomplete for all 3 datasets.
* Job title, DOB, Online order and brands all had blanks.

Solutions:

* Filtered out all blanks from datasets.
* Filter customer ids from 1 to 3500

Recommendations:

* Make sure tables are up to date.
* Create dropdown options to easier verify that no blanks are in datasets.
* To prevent future occurrences of incompleteness it is encouraged to cross check datasets and sync the data. Incomplete data can cause the data to be skewed.

**Consistency:**

* Inconsistent in gender for customer demographic dataset and states for customer address dataset.

Solutions:

* Filter all ‘M’ entries under ‘Male’, Filter all ‘F’ and ‘Femal’ entries under ‘Female’.
* Filter all ‘New South Wales’ entries under ‘NSW’, and all ‘Victoria’ entries under ‘VIC’.

Recommendations:

* Create dropdown options to easier verify that no mistakes or errors have been made.

**Currency:**

* In Customer Demographic dataset, the entries with ‘Y’ in deceased indicator are no longer current customers.

Solutions:

* Filter out all ‘Y’ entries in deceased indicator.

Recommendations:

* Deceased customers are no longer customers, removing them from data will increase the currency of data and will result in more accurate estimates in future analysis.

**Relevancy:**

* Default column in Customer Demographic and cancelled transactions in order status provided no relevant information.

Solutions:

* Deleted the Default column from the Customer Demographic dataset.
* Filtered out ‘Cancelled’ entries in order status for Transaction dataset.

Recommendations:

* Check for incomprehensible metadata and delete or format to make comprehensible.
* Cancelled order status is irrelevant information for future analysis. Remove so it doesn’t skew the data.

**Validity:**

* Format for list price and product first sold date in Transaction dataset.

Solutions:

* Format list price to currency.
* Format product first sold data to short date.

Recommendations:

* Setup columns so formats like price and date are already in place when entering new data. Makes the data easier to interpret.

That summarizes all the data quality issues discovered through the first stage of data quality analysis. The solution methods are simple and effective ways to improve data quality for future analysis. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let us know if you have any questions regarding methods used or any data quality issues identified.

Best Regards,

Alex Rodriguez