Dear Manager,

Thank you for providing us with the three datasets from Sprocket Central Ltd. The following table reflects the data we received, please reach out if there is a misunderstanding.

Dataset	No. of Rows	No. of Columns	Distinct Customer IDs
Customer Demographic	4000	13	4000
Customer Addresses	3999	6	3999
Transaction Data	20000	13	3494

I have evaluated the datasets according to the data quality dimensions framework as follows:

- Accuracy – correct values(Inaccuracy: not reasonable values)
- Completeness – data fields with values(Incompleteness: missing values)
- Consistency – values free from contradiction
- Timeliness – values up to date
- Relevancy – data item with value meta-data(Not relevancy: Might not be relevant to the topic)
- Uniqueness – records that are not duplicated
- Validity – data containing allowable values

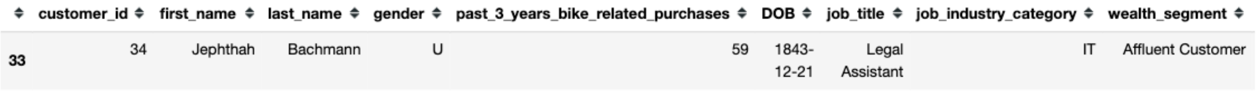
The summary table below highlights key data quality issues we have discovered in the data cleaning process. Please let us know if you have any queries concerning the issues presented.

Dataset	Customer Demographic	Customer Addresses	Transaction
Accuracy	DOB		
Completeness	last_name, DOB, job_title, job_industry_category, Tenure		transaction_date, online_order, Brand, product_line, product_class, product_size, standard_cost, Product_first_sold_date
Consistency	Gender	State	standard costs
Timeliness			
Relevancy			Cancelled orders
Uniqueness			
Validity	DOB, Default		Product_first_sold_date

In the following, I set out a more in-depth description of the data quality issues we have discovered and the strategies used to mitigate any data inconsistencies, along with recommendations and explanations as to how to improve the accuracy of the data sources to avoid data quality issues in the future. This will improve the accuracy of the data available to inform any future business decisions.

Accuracy issues

In Customer Demographic, the DOB column has an inaccurate value. DOB is 1843, which means this customer is over 120 years old.



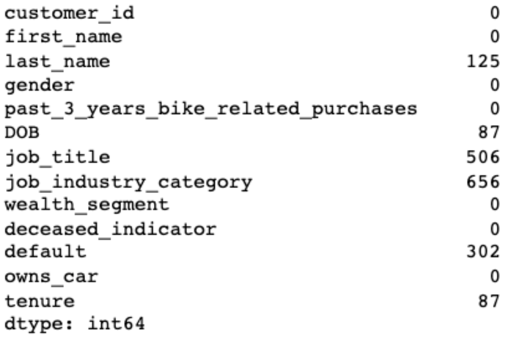
	customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category	wealth_segment
33	34	Jephthah	Bachmann	U	59	1843-12-21	Legal Assistant	IT	Affluent Customer

Mitigation: I treated this DOB data as a missing value. Then, I fill it by forward/backward filling methods.

Recommendation: Create an age column from DOB for easier comprehensible data.

Completeness issues

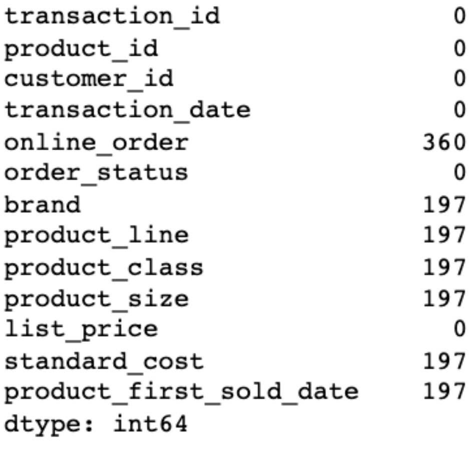
In Customer Demographic dataset, there are some missing values in each column below. Ex:last_name has 125 missing values



```
customer_id                              0
first_name                               0
last_name                              125
gender                                   0
past_3_years_bike_related_purchases      0
DOB                                     87
job_title                              506
job_industry_category                  656
wealth_segment                           0
deceased_indicator                       0
default                                302
owns_car                                 0
tenure                                  87
dtype: int64
```

Mitigation: I filled in the missing values using forward/backward filling methods for these columns but tenure column, which is used the median of tenure to fill the missing values.

In Transaction dataset, there are some missing values in each column below.



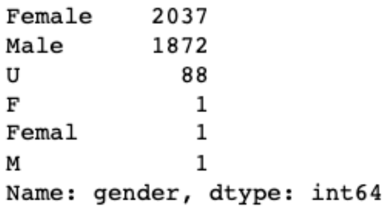
```
transaction_id             0
product_id                 0
customer_id                0
transaction_date           0
online_order             360
order_status               0
brand                    197
product_line             197
product_class            197
product_size             197
list_price                 0
standard_cost            197
product_first_sold_date  197
dtype: int64
```

Mitigation: I deleted all rows in brand, product_line, product_class, product_size, standard_cost, product_first_sold_date which have missing values in this dataset. And use forward/backward filling methods in online_order column.

Recommendation: It would be better to have original complete dataset rather than fill out or delete data for missing values.

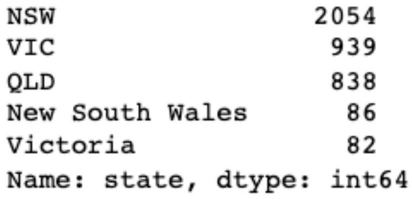
Consistency issues

For Customer Demographic dataset, the gender was in inconsistent format, such as M is Male for short. (Table is shown below.)



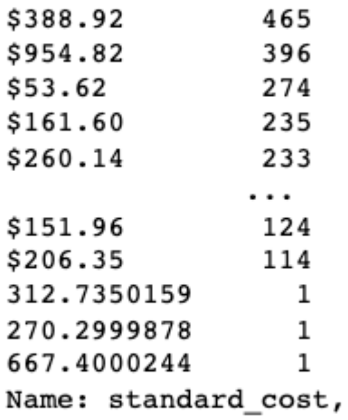
```
Female    2037
Male      1872
U           88
F            1
Femal        1
M            1
Name: gender, dtype: int64
```

For Customer Addresses, the state was in inconsistent format, such as NSW is as same as New South Wales. (Table is shown below.)



```
NSW                2054
VIC                 939
QLD                 838
New South Wales      86
Victoria             82
Name: state, dtype: int64
```

For Transaction, standard_cost was in inconsistent format, such as "$", dollar sign.(Table is shown below.)



```
$388.92        465
$954.82        396
$53.62         274
$161.60        235
$260.14        233
              ...
$151.96        124
$206.35        114
312.7350159      1
270.2999878      1
667.4000244      1
Name: standard_cost,
```

Mitigation: Replace acronym using regular expressions.

Recommendation: Ensure consistency of regular expressions across datasets for categorical fields.

Relevancy issues

For Transaction dataset, the order status showed cancelled orders.

Mitigation: I filtered out cancelled order status .

Recommendation: Cancelled order status may be ignored if it is not relevant to the analysis.

Validity issues

For Customer Demographic dataset, DOB and Default are not valid to analyse.

For Transaction dataset, the product sale date is an integer which may cause confusion.

Mitigation: I changed the DOB format to date format and deleted Default. I also deleted Product_first_sold_date.

Recommendation: Ensure that there is no corrupted data.

The above summarises the key data quality issues discovered through the first, data quality analysis stage.

Could you confirm if there is in-corrupted Default column and Product_first_sold_date column?

Please let us know if you have comments or questions on the above as I would be happy to discuss to ensure that all assumptions applied align with Sprocket Central Ltd.'s understanding.

Kind regards,

Yuan Hao

Data Consultant, KPMG.