Dear Sir or Madam,

Thank you for providing us with three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please let us know if you have any queries surrounding the issues presented.

We used Python Pandas package to identify the data quality issues, and making modifications.

Summary Table

	Accuracy	Completeness	Consistency	Relevancy	Validity
Transactions	1. missing profit	1. online order: null values 2. Brand: null values			1. Product first sold date: format
NewCustomerList			1. gender: inconsistency	1. unnamed value columns: delete	
CustomerDemographic	1. missing age	with null values 1. last_name 2. DOB 3. job_title / category	1. gender: inconsistency	1. default: delete	1. deceased_indicator : drop "Y"
CustomerAddress			1. state: inconsistency		

Below are more in-depth descriptions of data quality issues and methods of mitigation used. Recommendations and explanations have also been included to avoid further data quality.

Accuracy Issues:

- In the Transactions table, we need a new column to record the sales profit.
- In the Customer Demographic table, we need a column for customer age.

Accuracy Mitigations:

- Added profit column by subtracting standard cost from list price

```
#Add profit
#df_1.dtypesB
df_1['profit'] = df_1['list_price'] - df_1['standard_cost']
```

```
df_1.head()
```

_line	product_class	product_size	list_price	standard_cost	product_first_sold_date	profit
ıdard	medium	medium	71.49	53.62	1970-01-01	17.87
ıdard	medium	large	2091.47	388.92	1970-01-01	1702.55
ıdard	low	medium	1793.43	248.82	1970-01-01	1544.61
ıdard	medium	medium	1198.46	381.10	1970-01-01	817.36
ıdard	medium	large	1765.30	709.48	1970-01-01	1055.82

- Add age column by subtracting DOB from current date

```
#attempt to calculate age
import datetime as DT
now = pd.Timestamp('now')
df_3['age'] = (now - df_3['DOB']).astype('<m8[Y]')
```

```
df_3.head()
```

job_title	job_industry_category	wealth_segment	deceased_indicator	owns_car	tenure	age
Executive Secretary	Health	Mass Customer	N	Yes	11.0	67.0
nistrative Officer	Financial Services	Mass Customer	N	Yes	16.0	39.0
ecruiting Manager	Property	Mass Customer	N	Yes	15.0	66.0
NaN	IT	Mass Customer	N	No	7.0	59.0
ior Editor	NaN	Affluent Customer	N	Yes	8.0	43.0

Completeness Issues

- In the Transactions table, null values exist in online order status, and brand / product relevant columns.

```
df_1.isnull().sum()
```

```
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
```

- In the customer demographic table, null values exist in last name, job title, category and tenure

```
df_3.isnull().sum()
```

```
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
owns_car                               0
tenure                                87
age                                   87
dtype: int64
```

Completeness Mitigations:

- For transactions table, since we are difficult to find unified replacement value, we can choose to drop null values, or we need the update of data table to provide all the data information.
- For customer demographic table, we can either drop null values or update the table to ensure all information to be filled without null values.

Consistency Issues:

- Inconsistency in gender for NewCustomerList and customer demographic tables respectively. Details as shown in below screenshots.

```
1 df_2['gender'].value_counts()
```

```
Female    513
Male      470
U          17
Name: gender, dtype: int64
```

```
1 df_3['gender'].value_counts()
```

```
Female    2036
Male      1871
U           88
F            1
Femal        1
M            1
Name: gender, dtype: int64
```

Consistency mitigations:

- Replace “U” with “Unspecified” for gender in New customer list. Also replace other inconsistent values in the customer demographic table.

```
1 df_2['gender'] = df_2['gender'].replace('U', 'Unspecified')
```

```
1 df_2['gender'].value_counts()
```

```
Female         513
Male           470
Unspecified     17
Name: gender, dtype: int64
```

```
1 .replace('F','Female').replace('M','Male').replace('Femal','Female').r
```

```
1 df_3['gender'].value_counts()
```

```
Female         2038
Male           1872
Unspecified      88
Name: gender, dtype: int64
```

Relevancy Issues:

- New customer list has 5 unnamed columns
- Customer demographic table has default column.

Relevancy mitigations:

- Delete unnamed columns and the default column.

Validity Issues:

- In the transactions table, the product first sold date values has float datatype which needs to be converted to datetime.
- In the customer demographic table, the deceased indicator has 2 “Y” values.

Validity mitigations:

- Convert first sold date from float to datetime.

```
st_sold_date']=pd.to_datetime(df_1['product_first_sold_date']).dt.date
```

- Drop rows with “Y” in deceased indicators.

```
df_3['deceased_indicator'].value_counts()
```

```
N    3998
Y       2
Name: deceased_indicator, dtype: int64
```

```
df_3 = df_3[df_3['deceased_indicator'] == 'N']
```

```
df_3['deceased_indicator'].value_counts()
```

```
N    3998
Name: deceased_indicator, dtype: int64
```

That summarises all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis.

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

Kind regards

Hang Liu