Hello,

This is Muhammad Owais Akram from KPMG Data Analytics (Virtual Internship) team. I am sharing my report after reviewing the dataset provided by your company.

Data set had several quality issues and we will talk in terms of standard data quality dimensions

1- Correct Values:
Most of the columns has correct values where as there are columns whose data types were not correct.

2- Completeness:
Data set “Transactions” and “Customer Demographic” were not completed several columns has null values.

```
# to check for null values
cusdemo.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
default                            302
owns_car                             0
tenure                              87
dtype: int64
```

```
tra.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
```

3- Consistency:
Some columns of Data set “Transactions” and “Customer address” were not consistent and same values were being written with different styles.

```
#checking for gender column
cusdemo["gender"].value_counts()
```

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

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

4- Relevancy:
There was a column that was not relevant to the study and contained garbage value.

default
''
<script>alert('hi') </script>
2018-02-01 00:00:00
() { _; } >_[(())] { touch 'tmp/blns.shellsh...
NIL

5- Uniqueness:
All three data frames did not have any duplicated values.

6- Interpretability
Some columns were not interpretable, like property valuation does not have any unit. Similarly, product class column does not clarify what it refers too.

property_valuation
10
10
9
4
9
product_class
medium
medium
low
medium
medium

Furthermore, it is important to note that in transaction data set no of customer_id is not equal to no of customer_id in customer demographic data set, which means when combined we will lose some data.

Our recommendation is that data validation method to be used while collecting and inserting data. Limits should be clear and datatypes should be defined prior to data collection in order to avoid the above mistakes.

Currently some null values can be filled with median, mean values but some categorical values can not be filled as filling them, in order to preserve data quantity will impact on the quality of data and can skew the result of our analysis.

Best Regards

Engr. Muhammad Owais Akram