KPMG Internship Task 1 Data Quality Assessment

Subject: Review of Data Quality and strategies to mitigate Data Quality Issues

Dear Client,

I hope this mail finds you well. Out team have thoroughly gone through the quality assessment of data provided. Here are some data Quality Issues we find,

1. **Accuracy** : We observed that certains fields contain incorrect data. For example

- Some Addresses provided in data might not exist
- In transactions sheet, same product id corresponds to different brand, product line etc
- There is extra column name **default** which is of no use.

2. **Completeness** : There is some important data is missing. For example,

- Job title and Industry is not mentioned for some customers.
- In some cases, It’s not confirmed whether it was an online order or no.’
- Some orders doesn’t contain important information like brand name, product line and product class.
- For some customers DOB was not mentioned.

3. **Consistency** : There is some data which gives conflicting information.

- In transactions sheet, (product id = 0) corresponds to different brands which is not possible.
- There were some cases where Order Status was **“Approved”** but there was no further information mentioned about product and payment status.

4. **Currency** : There are some values which are not up to date. For example,

- In CustomerDemographic Sheet, one customer have DOB “1843-12-21” which means the person is 180 years old. The customer’s deceased status is “N” that means he is alive which is almost impossible.

5. **Relevancy** : There are some columns which lacks relevance to our analysis objectives. For example,

- There is column named “default” in CustomerDemographic which doesn’t seem relevant to our analysis.
- There is column property valuation in Customer Address. We also find that irrelevant to out analysis objectives.

6. **Validity** : There was some data that did not contain allowable values. For example,

- Zip Code, Date was not formatted as number which can cause problems.
- Column name product_first_sold_date was not formatted as Date.

7. **Uniqueness** : There were some duplicate values in the data. For example,

- There were multiple names mentioned for same state e.g. NSW and New South Wales are same.
- Same Productid was assigned to different products.

Strategies to mitigate these Issues

1. **Data Profiling and Cleansing** : Identify and correct inaccurate values. e.g. Formatting Date/Time, converting “New South Wales to NSW” in state etc.

2. **Setup Validation Rules :** The data which have limited options to be entered can be restricted to certain values e.g. Gender column can have only two options “Male” or “Female”, Name of states can also be selected from given options.
3. **Remove Duplicate Values :** Remove customers which have same customer id, products which have same transaction id etc.
4. **Update Data Regularly** : Maintain customer’s living status up-to-date.