The KPMG forage virtual internship

Task 1.

Data Quality assessment

Business background:

Sprocket Central Ptl Ltd, is a fiction company name for the purpose of the virtual internship. It is a medium size bikes and cycling organization. The company needs help with its customer and transaction data. The organization has a large datasets relation to its customers, but their team is unsure how to effectively analyze it to help optimize its marking strategies.

**Datasets:**

Customer demographic

Customers addresses

Transactions data for the past 3 months

Data exploration: identifying ways to increase the quality of spocket plt data.

Highlights of the data cleaning process:

Transaction data table

1. Accuracy – correct values: the column online orders, has text values formatted as general. It contains a lot of blanks
2. Uniqueness- no duplicate was found
3. Completeness - data field with values: the column brand, product\_line, product\_size, product\_class, online\_order has empty values
4. Consistency – values free from contradictions: The values in the columns ‘prodcut\_class and product\_size’ are similar and can be confusing.
5. Currency – values up to date: All the information provided were up to date
6. Relevancy – data items with values, mate-data:
7. Validity – data containing allowable values

New customers data table:

The columns past 3 year bike related purchase, postcode, property\_valuation has numbers formatted as text.

The datasets contains large amount of blank entries, the column name DOB is not complete.

The dataset contains abbreviations such as DOB, QLD, NSW, U and VIC which makes it difficult to infer the meaning of the words.

**Conclusion:**

I will recommend that the company provide access to some of the abbreviated words in the datasets, also allow me to rearrange the date to suite the analysis I will perform afterwards

After viewing the model solution, I noticed that the mode of my presentation was poor, as I could not properly highlight the areas that needed to be revaluated.

Points to note:

* All the data sets contains empty values(blanks), and because they are much, we can't just remove them.
* The inconsistency in the datasets "text formatted as general, and numbers formatted as text" should be addressed while recoding the entries.
* If there is need for abbreviation, a descriptive note should be included so as to understand most of the abbreviated words.
* The entries that requires known answers like, Female, Male or name of cities can be incorporated into a dropdown list for easy selection.

The last part is, data cleaning is paramount to effective data analysis and decision making.

