Greetings from KPMG,

Dear Sir/Ma’am,

I am Hardik Prajapati former intern at KPMG, from the team of Data analytics. I was assigned to analyse your data and evaluate the quality of your dataset.

After reading your T&C and the requirements mentioned in the documents, I have made some analysis on your data to find behaviour and nature of the data. I have commented on various quality standards that are missing in your dataset. I have also suggested some proven steps/tips to improve the quality of the data. In addition to that I have also requested you to provide us some more information about your business in order to understand some labels(variables).

I request you to kindly go through my report once. I hope you understand this report very well and provide me necessary information for the analysis. This will help me and my team to understand your requirements in depth. Also, if you have questions, please feel free to contact to me and my team.

Regards,

KPMG.

**Transactions**



As we can see that the data contains too many missing values, hence this data is **incomplete**. Although there is no duplicated entries in the data, hence it holds the quality of **uniqueness**.

Also there is one important thing about this dataset regarding validity of the data

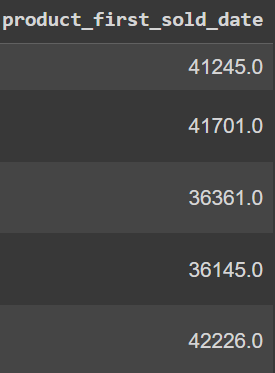
**Column name**: product\_first\_sold\_data

**Problem**: The values present in this column are not appropriate and it seems incorrect.

**Solution:** Can be stored in DD/MM/YYYY or MM/DD/YYYY format.

**Null Values:** 197

**Snap Shot:**

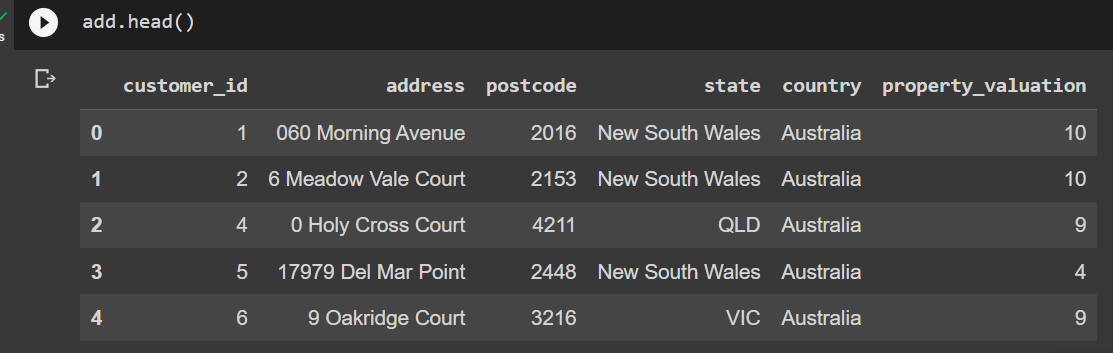


**Findings:**

Since the number 197 is common in most of the column means this is something observable in the data. I would suggest you to investigate about this. This unusual behaviour in the data may reveal something important and it can play important role in your business.

**Customer Addresses**

This is most cleaned and managed data set it does not any issues in my honest opinion.

****

**Suggestions:** The only suggestions I would give you is if you can Fill the column named as state with full forms of the name of the corresponding state.

**Quality:** The data holds **Uniqueness, Completeness, Validity**. I cannot comment on the accuracy of the data. I would request you to provide more information on this data set for checking the accuracy. I my honest opinion the data is fine.

**Customer Demographic**



**Description:** This data of 4000 entries contains too many missing values for some columns i.e job\_title and job\_industry\_category.

**Problems:**

1. **Column name:** gender

**Problem:** it contains both short forms and full-forms for representing the same gender. Also

there are some entries where even full-forms are not complete. Hence, this column can cause

the data inconsistency in your system.

**Possible solutions:** Filtering out each entry and store either only short forms or full-forms

For example, if the entry is named as “Femal” or “Mal” we can restore them as “F” or “M”

respectively.

1. **Column Name:** job\_title

**Problem:** The column contains too many missing values about the job\_title of the

help customer. It is very important to know about the customer’s background for any business

**Why:** This can help you to decide which customers to target. As your business is all about bike products. If someone has good job and wealth, he would require more luxuries bike products (provided that he/she is someone who loves to ride on bike). You can also design special products just for these kinds of targets. Also, if someone has low paid job and the basic survival is difficult for him/her. Your company can launch some low budget products for this people. I suggest you to investigate and find out their job roles.

1. **Column Name:** job\_industry\_category

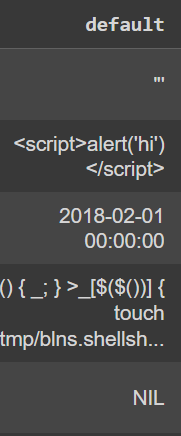
**Problem:** This column contains the greatest number of null cells in whole dataset. i.e., 656 out of 4000 entries. It is important to know about these values

**Solution:** Since compare to “job\_title” it has more null values. there is one solution that could work here. We can remove those entries which have null value in both columns from the data set (if and only if common entries lie in count of 50). Then we can select for those entries which have the null value only in “job\_industry\_category” but have something in “job\_title”, through this we can try to guess the category of the job.

1. **Column Name:** Default

**Suggestion:** I am not aware of this column what this is, and why it is there in your data set. If it is redundant and some garbage variable. And not useful to you. I would suggest you to remove it from the dataset.

**Request:** If the variable is useful to you then I would request you to provide me some information about this.



Please Find the link to the python notebook for the analysis: <https://colab.research.google.com/drive/1PxYrhqK5-qQ6HJRMLII6XiVwGkQ201JZ#scrollTo=_o1pN8HMlkJ9>