Dear Client,

I’m a member of KPMG’s AIM team. Thanks for providing us with three datasets from Sprocket Central Pty Ltd. There are several issues of data quality I want to talk with you later that may impact the next data analysis.

1. The dataset “CustomerAddress” includes 3 ***Customer\_Ids*** that are not shown in other two datasets. The dataset “Transactions” also includes an extra ***Customer\_Id*** that isn’t shown in other datasets. It will generate many missing values when we join the tables. Since the number of extra Customer Ids is very small, we could delete corresponding record before analyzing.
2. There are some duplicate value of attribute ***Gender*** in “CusotmerDemographic”, ***State*** in “CustomerAddress”, and ***Property*** in “NewCutomerList”. These duplicate values would disrupt the result of analysis. We could replace similar value with same one.
3. The ***DOB*** in “CustomerDemographic” include some unreasonable date such as “1843-12-21” that means this customer is 175 years old. We should check the accuracy of these birth date and delete the wrong records.
4. Most values of ***Default*** in “CustomerDemographic are unreadable and useless. It will be better if we delete this attribute.
5. Some attributes include few missing values. We could delete relevant records or fill with mean value
6. The ***product\_first\_sold\_date*** in “Transactions” does not use data format. It will be difficult to analyze the effect of dates.

The above is the database quality issue I found and the corresponding solutions. Our team will continue with the data preparation, data mining process. I hope this will enable us to better analyze the datasets and provide you with professional advice.

If you have any advice or idea, please email me at any time.

Have a nice day!

Sincerely,

Mia