Dear SPRocket Central Team,

My name is Thanak Rattanopastkul, a data analyst intern from KPMG Virtual Internship. First of all, thank you for providing the raw customer data and transaction data. Explored the data, I’m sure there are really good insights in the dataset waiting to be discovered. However please allow me to summarize the data just to make a reference. From the dataset I received, there is only one xls file called **“KPMG\_VI\_New\_raw\_data\_update\_finala.xlsx”**. In the spreadsheet there are 4 sheets called:

* Transactions - 20000 observations
* CustomerDemographic - 4000 observations
* CustomerAddress - 3999 observations
* NewCustomerList - 1000 observations

**Define the Goal and Deliveries**

The main task for KPMG interns is to identify prospects from NewCustomerList using the other 3 dataset as features.

The delivery will consist of:

* Powerpoint presentation consisted of 3 phases
  + EDA
  + Model Development
  + Insight Interpretation
* All of the phases need to be described thoroughly about the strategy and activities included in a given phase

**Some issue about the raw data**

While the dataset is vast and expansive. There are still some problems with data integrity. I would like to put the problems found in the data set below and also provide some ideas to mitigate the issue in the future.

* Accuracy issue - [fix input mistakes]
  + Impossible Date of birth
* Completeness issue - [Trim the missing data]
  + Many missing data in the dataset
* Consistency issue - [Validate the data]
  + Some are not properly format such as gender
* Currency issue -[Update the dataset if able]
  + The dataset is from 2017, a bit too old.
* Relevancy issue - [Trim out the unnecessary data]
  + Default column with no context data
* Validity issue - [Verify the source, format, and correctness of the data]
  + Product\_first\_sold\_date
  + Some dates aren’t in datetime format
* Uniqueness issue - [Check data duplication]
  + Country column are all “Australia”

All the issues mentioned above are all fixed during the data cleaning process. Our team is currently working on exploratory data analysis. The presentation will be ready soon.

Best Regard,

Thanak Rattanopastkul

Data Analyst Intern