Hello,

I am Dishant Naik a Data Analyst at KPMG. First, we all at KPMG would like to thank you for choosing us and sharing your valuable data with us. As an initial stage of any project we look at the quality of the data and if we find anything that does not look good, we try to fix it after consulting it with the client. So as per process I have analyzed the dataset and found some issues and some ways to overcome them in it which I have mentioned below.

In Transactions dataset the main issue is unavailable or null values. Following are the columns with null values online\_order, Brand, product\_line, product\_class, product\_size, Standard\_cost and Product\_first\_sold\_date. Apart from this data in Product\_first\_sold\_date dose not make any sense as it contains random number then dates. One important finding is that exactly 197 values are missing in Brand, product\_line, product\_class, product\_size and Standard\_cost colums. And my assumption is that your data is missing information about one brand. To overcome these issues, we can either get rid of all the missing value which may lead to missing data about a unknow brand or we can feed in some assumed values like average of the column.

In NewCustomerData dataset also major issue is missing value but less missing values compare to Transactions dataset. Apart from this there is three important findings which I would like to discuss with you. First In some rows there is no correlation between job\_title and job\_industry\_category for example Nurse -> Property and Librarian -> Entertainment. Second in welth\_segment where two category Affluent customer and High net worth serves same purpose so we should replace both with similar name. And third issue is in Deceased\_indicator column where all values were N which may indicate that this data is not up to date and to avoid error in future finding we should get rid of this column or we should update this column. Lastly same as Transactions data there was one column named Value which is missing context or purpose of the column.

In CustomerDemographic dataset Last\_name and job\_title columns have considerable amount of missing values. Apart from it in Gender column there is redundant categories such as M and Male, F and Female, here we should replace all Male with M and all Female with F or vice versa. default is one more column whose purpose is unclear so I think we should get rid of it. And lastly one minor issue in this dataset is in DOB column which have one instance of the year 1843 which looks like mistake to me as no other data available of 1800 decade, here we can replace 1843 with 1943 as it looks logical solution.

Lastly in CustomerAddress data there is just two minor issues. First in state column abbreviation is used for state’s name except for Victoria which have both Vic and Victoria. In this case we should replace all Victoria instances with Vic. Finally, property\_valuation is column whose measure is unclear, or we can say it is unclear how it was calculated and how it will be helpful in future findings?

I have also attached a doc file which describes issues and solutions in depth for your reference. Also, you can contact as for any queries you may have. Again, thank you for choosing KPMG and I or someone from our team will be in touch with you for further process.

Thank You

Dishant Naik