**KPMG VIRTUAL INTERNSHIP TASK 1**

As we all know that poor quality data can be a hinderance to one’s business growth. What is in demand , Whether online mode is preferred more or offline mode, which brand is in trend, and many more. The answer to all these queries can be solved only the observing the data sharply and then find the ways to tackle the quality issues. So, it is very important to analyse the data quality and fix the issues with the help of the techniques available. This step plays a vital role in development and growth of a company.

Now, we all have one questions that what kind of issues data can have?

So, there can be many issues present. I have tried to list some of the problems below:

* Huge amount of data,
* Incompleteness of data,
* Inaccurate data,
* Duplicate Values,
* Presence of null values,
* Improper organisation of data,
* Presence of Irrelevant and extra data, and many more.

So, to address these issues present in the dataset of Sprocket Central Pty Ltd , I deployed python libraries such as pandas, matplotlib, seaborn and extracted information regarding the issues present. I gathered the below mentioned issues from the respective sheets of the dataset given.

**#ISSUES IN DIFFERENT SHEETS:**

**1. From Customer Demographic sheet:**

* Out of 4000 names, 3139 first names are unique and rest are same and 3725 last names are unique
* Last name count is 3875 , it means 125 people don’t have their last names.
* DOB of 87 people are not given.
* Job title of 506 people are not mentioned and job industry category of 656 are not mentioned
* Tenure of 87 people are not mentioned
* There are 6 types of gender mentioned
* ‘U’ and ‘Femal’ gender types don't make any sense at all. Futher , We need to mention whether ‘F& female’ or ‘M &male’ are same else computer won't consider it as equal entity
* One DOB is 21-12-1843, 167 year aged customer,which is quite impossible as of today's scenario.
* Deceased\_indicator and value column is of no use.
* After describing df1, we get Nan as unique value for tenure but tenure can’t be Nan.

**2.From New customer list :**

Number of null values present in different columns

* #last\_name- 29
* #DOB- 17
* #job\_title - 106
* #job\_industry\_category- 165
* deceased\_indicator, Unnamed: 16,Unnamed: 17, Unnamed: 18, Unnamed: 19,Unnamed: 20 has no significance .
* Country is same for all, so there is no need to have a column for country.
* ['Male', 'Female', 'U'], U gender type creates ambiguity here.
* On what basis you are characterising customer with rank and value columns.

**3.From Transactions’ sheet:**

* 179 orders were cancelled out of 1000.
* Sale of different product line is not balanced. Standard has the highest count wheras Mountain has the least. This needs to be balanced.
* Norco Bicycles is the least.

After doing the assessment of all the above data present, I have come up with some of the mitigation strategies that are mentioned below:

**Strategies:**

1. The name columns has no relevancy with further analysis so we can drop it out.
2. The gender columns need to be managed as it contains 6 unique values, which doesn’t seem correct at all. It may take 3 values at max.
3. The idea is to compare between old sales and recent sales and analyse them but it is unnecessary to take too much old transactions and sales into consideration.
4. Remove the DOB which gives ambiguous ages.The DOB column needs to be converted into age column so that we can easily get an idea which age group we have to target in order to maximize sales.
5. Job\_title column has too many categories in it so it will be difficult to analyse sales according to it so in place of that we will only consider the job\_industry column, making our analysis easier and effective.
6. Deceased indicator, default indicator and tenure has no relevancy with our further analysis.
7. Unnamed: 16,Unnamed: 17, Unnamed: 18, Unnamed: 19,Unnamed: 20 has no significance in data analysis.
8. In new customer list also above points are applicable.
9. In new customer list, Country is same for whole data so it can be dropped ,also rank and value has no relevancy.
10. State column can be used to analyse sales in different states of country.
11. In transaction , transaction date ,order status, Brand, product lines, Product class, product\_size and cost price are most relevant data for analysis .
12. Or Fill the missing data with mean of total of the respective column provided the column has integer or float values.
13. Drop the values which are not of int or float datatype.