**Data quality issues of KPMG report on the dataset for Sprocket Central Pty Ltd**

**Duplicate**: The first thing I did was to check if there is any element of duplicate, a duplicate data can easily affect the desired result of our analysis. Since the dataset is a very large one, one or more duplicate is possible and if found it should be removed instantly. Fortunately, no duplicate was found.

**Spacing**: Irrelevant spacing in vital column or rows of our dataset can affect the outcome of our Analysis, since it’s a large data set I calculated the number rows that contain empty cells so because they are very small compare to the large dataset so I decided to remove entire row that contains empty cell.

Irrelevant columns were removed e.g. Column K was removed because is irrelevant to the data.

**Variables**: I filter the Gender variables and I found 7 different variables gender variable should not be more than 2 and at most 3(male, female or undisclosed).

**D.O.B**: I converted the date of birth to age so as to calculate the average age of people who buy each product. I noticed Bjephthah Strettle was 180years old from the current date, this is practically impossible so I thought of predicting that, it could have been a mistake of a digit 1943 instead of 1843 but to avoid irregularities and doubt I removed the entire row and I work with the rest of the Dataset

**Conclusion and recommendation**

* **According to the result of our analysis, it’s been proven that people from nsw especially have greater influence on the company’s product hence more reason why the company should focus on them followed by vic then qld.**
* **I do not recommend the company to focus on victoria and new south wales because the population is already against them.**