**KPMG Task 1**

***Draft an email*** *to the client identifying the data quality issues and strategies to mitigate these issues. Refer to ‘Data Quality Framework Table’ and resources below for criteria and dimensions which you should consider.*

Subject: Quality issues and Analysis for Sprocket Central

Greetings from KPMG,

We have analyzed the data which was sent to us and below are some of the issues present in it and our recommendations on how to deal with them. Let’s go through the database table by table

1. Transactions table
2. The product\_first\_sold\_date is not stored as date time values and therefore it has to be converted into date time before processing the data
3. product\_id does not uniquely identify each product, that is product\_id can’t be used as a primary key. The database can be made much cleaner if we create a new table called ‘inventory’ which has all the product details and each product is identified with product \_id.
4. There are missing values in the following columns:

[online\_order,brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date].

1. NewCustomer List
2. Unlabeled column numbers 16,17,18,19,20. It is not clear what is being achieved here.
3. The table does not contain a primary key and hence we cannot uniquely identify records
4. This table can further be decomposed into new\_customer\_address and new\_customer\_demographics
5. The exact purpose of value and rank columns is unclear.
6. As all of the customers present in the database are from Australia. It is redundant to have a country column.
7. This table contains data of customers who have been associated with Sprockets for up to 22 years and hence cannot be called NewCustomer list
8. Missing values in the following columns: [last\_name, DOB, job\_title, job\_industry\_category]
9. Unknown values of Gender column has been labelled as ‘U’ which should be replaced by NaN
10. CustomerDemographic
11. Default column contains data which does not provide any value and hence can be removed.
12. There are missing values in the following columns: [last\_name, DOB, job\_title, job\_industry\_category, tenure].
13. Gender column contains [F, M, Female, Male, Femal ,U] as the unique values i.e multiple representations for the same gender. Therefore, these errors must be fixed and unique values should only be [male,female,NaN].
14. There are multiple representations for Boolean values: ex: Y/N, Yes/No. It is better to have a standard Boolean value representation across the database for the ease of processing.
15. There are discrepancies in the DOB column where one of the observation states DOB of a customer who is not deceased as 21-12-1843 which cannot be true.
16. Customer address
17. There are multiple representations for the same state, for example, NSW – New South Wales, VIC – Victoria which needs to be rectified.
18. There are 3996 customers whose details are present in both the customer demographics and customer address tables. But there are 7 customers with incomplete data i.e, 4 customers do not have their addresses recorded and 3 do not have their demographic data recorded

Thank you,

Regards