Dear team,

Hope you’re doing well!

Thanks, first of all for sharing that data with us. I ran through that data and observed some anomalies related to the quality of data.

Below are the data quality issues identified and strategies to mitigate the same:

Table name : NewCustomerList data quality issues

1. **DOB** is not standardized.
2. **past\_3\_years\_bike\_related\_purchases** somewhat follows Gaussian Distribution but has integers in text format
3. It has 5 unnamed columns from Q to U containing values originating from random function.
4. Null/NA Values:
   1. **DOB** = 17
   2. **job\_title** = 106
   3. **job\_industry\_category** = 165

Recommendations to fix data quality issues:

1. Unnamed columns in sheet can be removed from dataset as they have the least correlation with output feature.
2. While reading the data from csv we can standardize and rectify the data type of features: **DOB**, **past\_3\_years\_bike\_related\_purchases, postcode**, and **property\_valuation**.
3. Missing Job titles can be populated with the mode of **job\_title** that falls under the **job\_industry\_category** of missing job title. Ex – For a missing **job\_title**, we will pick its **job\_industry\_category**, let’s say Manufacturing, and find out the mode of **job\_title** across dataset. This mode will be put in missing cell.
4. If the **job\_industry\_category** is also missing, in that case we will take mode of both **job\_title** and **job\_industry\_category** in union.
5. Missing **job\_industry\_category** can be populated with the mode of **job\_industry\_category**.

Table: Transactions data quality issues

1. Null/NA values:
   1. **online\_order** = 360
   2. **brand** = 197
   3. **product\_line** = 197
   4. **product\_class** = 197
   5. **product\_size** = 197
   6. **standard\_cost** = 197
   7. **product\_first\_sold\_date** = 197
2. **product\_first\_sold\_date** is in integer format.

Recommendations to fix data quality issues:

1. **product\_first\_sold\_date** feature will be converted into DateTime.
2. Rows having null **brand**, **product\_line**, **product\_class**, **product\_size, standard\_cost, online\_order,** & **product\_first\_sold\_date** can be deleted. In total we will delete 557 rows out 20,000.

Table: CustomerDemographic data quality issues

1. Null/NA values:
   1. Last name: 125
   2. DOB: 87
   3. Job title: 506
   4. Job industry category: 656
   5. Default: 302
   6. Tenure: 87
2. Gender has multiple values for same gender i.e. F, Femal, Female represents Female and M, Male represent Male.
3. Gender also has U.
4. At index 2346, Last name contains semicolon.

Recommendations to fix data quality issues:

1. Missing Job titles can be populated with the mode of **job\_title** that falls under the **job\_industry\_category** of missing job title. Ex – For a missing **job\_title**, we will pick its **job\_industry\_category**, let’s say Manufacturing, and find out the mode of **job\_title** across dataset. This mode will be put in missing cell.
2. If the **job\_industry\_category** is also missing, in that case we will take mode of both **job\_title** and **job\_industry\_category** in union.
3. Missing **job\_industry\_category** can be populated with the mode of **job\_industry\_category**.
4. We will take mode of **tenure** to fill any missing values in **tenure.**
5. Encode all values for **gender** into either Female or Male or U.

Table: CustomerAddress data quality issues

1. **state** is not standardized. Ex: VIC is as same as Victoria but they both have separate labels.

Recommendations to fix data quality issues:

1. State names will need to be replaced by their abbreviations. Ex: **New South Wales** will be converted into **NSW**.