**Subject**: Data Quality Assessment of Customer Demographic, Customer Address, and Transaction Data for Sprocket Central Private Limited

Manager,

I have received your request to perform a data quality assessment on the datasets comprising Customer Demographic, Customer Address, and Transaction data for the last three months at Sprocket Central Private Limited.

Here are the details after quality assessment of each dataset , and strategies to mitigate these issues:

***CUSTOMER DEMOGRAPHIC DATA***

**Shape** :

* 4000, 13

**Columns** :

* customer\_id
* first\_name
* last\_name
* gender
* past\_3\_years\_bike\_related\_purchases
* DOB
* job\_title
* job\_industry\_category
* wealth\_segment
* deceased\_indicator
* default
* owns\_car
* tenure

**Correctness** :

first\_name ,gender ,past\_3\_years\_bike\_related\_purchases ,wealth\_segment ,owns\_car

these columns provide correct values , upto date values , free from contradiction ,

Hence provides Uniqueness , Accuracy , Completeness and valid data

**Data Quality Assessment of other different columns :**

* **last\_name** has **125** missing values
* **DOB** has 87 missing values and a person with 180 years of age , with mean age of 36 and median age of 46 , seems good
* **gender** has 88 Gender not Specified Values specified with U ,where Male are specified with Male , M and Female are specified with Female , F , Femal.
* **job\_title** has 506 missing values .
* **deceased\_indicator** has 2 person who are not alive, but have significant purchases does not provide significant details .
* **job\_industry\_category** has **605** missing values .
* **default** column has irrelevant , most random values that seems to provide no information at all
* **tenure** column again has **87** missing values and matches with customers who have not specified gender , dob , job\_title , seems these customers don’t want to provide anything

**Conclusion** :

* people who have not specified gender has not specified DOB and most of them belong to IT industry and tenure is also not available for them.

Hence can be removed from the dataset since not providing any significant information or can be kept separately to get their data or analyzed later.

* deceased\_indicator does not provide significant information , most of customers are alive .
* default column doesn’t specify anything so can be remove
* job\_industry\_category has 605 missing values , can be replaced with the mod value

but not included in segment who have not specified other values or not specified

so most of this customer can be imputed with mean , median or mode values

* Hence this seems to be customers who don't want to share their personal details ,So can be removed from our customer demographic dataset or can be kept separately in different table and analysed later

***CUSTOMER ADDRESS***

**Shape :**

3999, 6

**Columns:**

* customer\_id
* address
* postcode
* state
* country
* property\_valuation

**Correctness :**

Everything seems good

* No missing values
* valid address
* valid postcode
* no requirement for country column , all belongs to Australia
* 101 products mentioned
* No duplicate values

***LAST 3 MONTHS TRANSACTIONS***

**Shape :**

(20000, 13)

**Columns:**

* transaction\_id
* product\_id
* customer\_id
* transaction\_date
* online\_order
* order\_status
* brand
* product\_line
* product\_class
* product\_size
* list\_price
* standard\_cost
* product\_first\_sold\_date

**Data Quality Assessment of other different columns :**

* **online\_order** has 360 missing values
* **brand , product\_line , product\_class , product\_price , standard\_cost , product\_first\_sold\_date and product\_size** all have missing value , all this segment belong to product\_id '0' .Removing this product id will be better as provide no significant information
* No duplicate values
* **product\_first\_sold\_date** datatype is in **floating** , converting it to datetime will be better
* **online\_order** can replaced with mode value
* **customer\_id** has id 5034 , but customer demographic dataset has only 4000 customers with last 4000 max value , hence invalid , removing it would be a better option

Best regards,

Shourya Nagpal

Data Analyst KPMG