**Data quality issues** in KPMG\_VI\_New\_raw\_data\_update\_final.xlsx

Hello,

Thank you. I received with 3 datasets attached from Sprocket Central Pty. Ltd.

* Customer Demographic
* Customer address
* Transaction from past three months

The following is the data quality from each sheet:

Transaction sheet-

Total data available = 20000

Transaction data available from past 3 months = 5080 (25.4% of total data)

* **Missing data - should be changed**

Columns\_name No. Of empty data(out of 5080)

|  |  |
| --- | --- |
| Transaction\_id | 0 |
| Product\_id | 0 |
| Customer\_id | 0 |
| Transaction\_date | 0 |
| Online\_order | 94 (1.85%) |
| Order\_status | 0 |
| Brand | 48(0.95%) |
| Product\_line | 48(0.95%) |
| Product\_class | 48(0.95%) |
| Product\_size | 48(0.95%) |
| List\_price | 0 |
| Standard\_cost | 48(0.95%) |
| Product\_first\_sold\_date | 48(0.95%) |

Same rows missing multiple data = 48

Column names No. of empty cells

|  |  |
| --- | --- |
| Brand | 48 |
| Product\_line | 48 |
| Product\_class | 48 |
| Product\_size | 48 |
| Standard\_cost | 48 |
| Product\_first\_sold\_date | 48 |

* **Inaccurate data - format has to be changed.**

Column\_name Issue

Product\_first\_sold\_date Not in date format

* **Non - standard data - check on decimal places**

|  |  |
| --- | --- |
| Column\_names | Column\_names |
| Transaction\_id | standard\_cost |
| 8708 | More than 2 decimal places |

Customer demographic sheet:

* Total data available - 4000
* **Incomplete data** -

customer\_id 0

first\_name 0

last\_name 125(3%)

gender 0

past\_3\_years\_bike\_related\_purchases 0

DOB 87(2%)

job\_title 506(12.65%)

job\_industry\_category 656(16.4%)

wealth\_segment 0

deceased\_indicator 0

default 302(7.5%)

owns\_car 0

tenure 87(2%)

* **Junk values** -

|  |  |
| --- | --- |
| Column\_name | Junk values |
| Default | 3364(84.1%) |

* Same gender entry : F, Femal, Female, etc.
* **Inaccurate data** -

1. Customer\_id-34 : invalid date of birth

|  |  |  |  |
| --- | --- | --- | --- |
| First\_name | Last\_name | Gender | DOB |
| Jephthah | Bachmann | U | 1843-12-21 |

Customer address sheet:

Total data available: 3999

* Same place with more than one name. Like VC, Victoria, etc.

For the analysis part, there will be issues, some of which mentioned above. It will be interesting to see the MSE of the analysis. Hoping to see assumptions aligning with your Company’s understandings.

Kind regards,

Subhodeep Sinha