Dear Manager,

Thank you for providing us the datasets. We have discovered some issues within the datasets during our analysis with Python. Please take a look at our data quality description below.

“**Transactions” data set**

**Accuracy issues**

* “Profit” column can be added instead of “list\_price” and “standart\_cost”.

**Completeness**

* Some columns have NA values.

|  |  |
| --- | --- |
| Column Name | NA Values |
| online\_order | 360 |
| brand | 197 |
| product\_line | 197 |
| product\_class | 197 |
| product\_size | 197 |
| standard\_cost | 197 |
| product\_first\_sold\_date | 197 |

**Relevancy**

* If we just want to analyze the valid orders, then we must not be analyzing the data where “order\_status” is “Cancelled”. If we drop the data with “order\_status” is “Cancelled”, the column will only have “Approved” values. This brings up an another problem because the column will only have 1 unique value and it won’t carry any information. Can be dropped.

**Validity**

* There seems to be a format problem with “list\_price” and “product\_first\_sold\_date”.
* “list\_price” needs a dollar sign before every value or every dollar sign must be removed before the values in “standart\_cost”.
* “product\_first\_sold\_date” needs proper date formatting like “00/00/00”.

**“CustomerDemographic” data set**

**Accuracy issues**

* “DOB” has an inaccurate value.
* “Age” column can be added instead of “DOB”.

**Completeness**

* There are some NA values.

|  |  |
| --- | --- |
| Column Name | NA Values |
| last\_name | 125 |
| DOB | 87 |
| job\_title | 506 |
| job\_industry\_category | 656 |
| default | 302 |

* Besides empty NA values, there are strings like “n/a” as well in the ”job\_industry\_category”.

**Consistency**

* There are 6 unique values, I guess the main values are “Male” and “Female” but there are values like “Femal”, “F” and “M”. This values might be refer to same values and there is an unidentified value like “U” which can be meaningless.

**Currency**

* If a data has “deceased\_indicator” == “Y”, this can mean that this data might not be up to date. Should be checked.

**Relevancy**

* “default” column has meaningless values and doesn’t carry any information. Can be dropped.

**“CustomerAddress” data set**

**Completeness**

* “customer\_id” has values that go up to 4003, but the data set has 4000 rows, so there must be some “customer\_id” missing. The column is incomplete. Should be filled with right information.

**Consistency**

* There are 5 unique values in “state”. These values are “NSW”, “VIC”, “QLD”, “New South Wales”, “Victoria”. Assuming “NSW” and “New South Wales”, “VIC” and “Victoria” are the same, “New South Wales” and “Victoria” values should be shortened.
* In the “address” column, we have detected 3 repeated values.

They are:

|  |  |
| --- | --- |
| Address Value | Repetition |
| 3 Mariners Cove Terrace | 2 |
| 3 Talisman Place | 2 |
| 64 Macpherson Junction | 2 |

This is an anomaly, comparing the rest of the data set.

**Relevancy**

* “country” column has just 1 unique variable. It doesn’t carry any info. Can be dropped.

This description includes all data quality issues discovered through the first stage of the data quality analysis. Suggested strategies are simple and effective ways of improving data quality for future analysis.

Please let me know if you have questions regarding any data quality issues identified.

Best of luck,

Virtual Intern Lokman Efe

KPMG Data Analytics