Subject: Data Quality Assessment Report and Identified Issues

Dear Manager,

I hope this email finds you well. As part of the data quality assessment process, I've thoroughly reviewed the provided datasets and identified several key issues that required attention.

Please find attached Data Quality Assessment Report, which outlines the issues encountered and the steps taken to address them. This will provide a quick overview of the main concerns that were addressed during the assessment.

I believe that addressing these data quality issues will significantly improve the reliability and accuracy of our analyses moving forward. Your guidance and input would be invaluable in ensuring the data is optimized for the upcoming tasks.

Thank you for your time and consideration. I'm looking forward to discussing these findings and moving towards a more robust and accurate dataset.

Best regards,

Nikita Timoshenko

**DATA QUALITY ASSESSMENT REPORT**

The main characteristics of given datasets are shown it the table below:

**Table 1 – Data Profiling**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | # of records | # of unique records | # of columns | percentage of  missing values |
| Transactions | 20,000 | 20,000 | 13 | < 1% |
| NewCustomerList | 1,000 | 1,000 | 23 | 1.4% |
| CustomerDemographic | 4,000 | 4,000 | 13 | 3.4% |
| CustomerAddress | 3,999 | 3,999 | 6 | 0% |

**Transactions** table contained 20,000 records providing information on transactions made by 3,494 distinct customers for 101 distinct products and 6 brands from the year 2017. No duplicate transactions were found. The following data quality issues has been defined during assessment:

* 358 records (2% of transactions) had unspecified *online\_status*, which can be filled as 'unspecified' to keep those transactions for further analysis.
* 197 records (1% of transactions) represent missing product attributes (brand, size, class, standard costs) and could be removed from the dataset.
* more information needed on what the column *product\_first\_sold\_date* refers to.
* it should be mentioned that there is no column for quantity sold was observed.

**CustomerDemographic** table represents data related to 4,000 customers indicating names, genders, birthdates, job titles and other information. The following data quality issues has been defined during assessment:

* the *gender* column contains not allowable values (misspelling and different format) which can be replaced with F/M/U
* the following attributes contain missing values: *job\_title* (497 customers), *job\_industry\_category* (656 customers), *last\_name* (125 customers). Depending on analysis purposes it is possible to keep those customers replacing blank values with ‘unspecified’ category.
* DOB doesn’t match the range constraint for 1 customer with 1843-12-21 indicated. Also, DOB is missing for 87 customers, for whom data in *default* and *tenure* columns are also missed.
* the column default is not interpretable thus cannot be used for further analysis.

**CustomerAddress** table provides data related to addresses, postcodes, states and countries for 3,999 customers referring to *customer\_id* as foreign key. The following data quality issues has been defined during assessment:

* the *state* column doesn’t meet the validity requirements due to different approaches in state naming. It`s possible to replace values and bring the data into one standardized format (e.g., NSW, QLD etc.)
* the following IDs are missing when *customer\_id* used as foreign key and data has been merged with **CustomerDemographic** table: 3, 10, 22, 23, 4001, 4002, 4003.
* more information needed on what the column *property\_valuation* refers to.

**NewCustomers** table expanded **CustomerDemographic** data with 1,000 new customers. The following data quality issues has been defined during assessment:

* there is no ID column for customers which could be used as primary key for further analysis.
* DOB data is missed for 17 customers.
* there are 5 unnamed columns contained numeric data with a lack of context.

The Data Quality Assessment results are shown in the table below:

**Table 2 – Data Assessment Matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity | Uniqueness |
| Transactions | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| NewCustomerList | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |
| CustomerDemographic | Close with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |
| CustomerAddress | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |

|  |
| --- |
| Apart from mentioned inaccuracies and data quality issues the following recommendations should be used to improve data quality and reliability for further analysis:   * determine issue root-causes for further resolution; * define Data Quality rules, thresholds to perform data quality assessments; * review data policies and procedures; * identify typical data quality issues and keep issue log updated; * implement validation control for specified fields, set fields with missing data as mandatory. |