Subject: Completion of Data Preprocessing for the Project

Dear Sir/Mam,

I hope this email finds you well. I wanted to provide you with an update on the progress of our project. I'm pleased to inform you that the preprocessing phase for KPMG\_VI\_New\_raw\_data\_update\_final datasets has been successfully completed. In this email, I will outline the key steps we undertook during the preprocessing phase to ensure the quality and readiness of the data for further analysis.

1. Data Integration:

- We consolidated the four datasets, namely [Transactions], [NewCustomerList], [CustomerDemographic], and [CustomerAddress], into a unified structure. This integration process involved merging relevant fields, ensuring consistent data formats, and creating a single, comprehensive dataset for analysis.

2. Data Cleaning:

- We performed extensive data cleaning to address various data quality issues. This included handling missing values, outliers, and inconsistent data formats. By employing appropriate techniques such as imputation, removal of duplicates, and outlier treatment, we have enhanced the reliability and accuracy of the dataset.

3. Data Transformation:

- We carried out necessary transformations to make the data more suitable for analysis. This involved standardizing units of measurement, normalizing variables, and encoding categorical variables. These transformations ensure the compatibility of data across different sources and enable meaningful comparisons and insights.

4. Feature Engineering:

- To enrich the dataset and capture important information, we engineered new features based on domain knowledge and data characteristics. By extracting relevant information and creating derived variables, we have enhanced the dataset's predictive power and the potential for meaningful analysis.

5. Data Validation:

- Throughout the preprocessing phase, we implemented rigorous data validation checks to ensure the integrity and validity of the dataset. This included validating data ranges, identifying inconsistencies, and cross-referencing with external sources to verify accuracy.

By undertaking these preprocessing steps, we have laid a strong foundation for the subsequent stages of the project. The cleaned, transformed, and validated dataset is now primed for in-depth analysis, model development, and extracting valuable insights.

Moving forward, our focus will shift to the exploratory data analysis and modeling phases, where we will leverage this refined dataset to uncover patterns, correlations, and key findings. We are confident that the efforts invested in data preprocessing will contribute significantly to the success of the project and the reliability of our results.

The summary table below highlights key data quality issues we have discovered in the data cleaning process. Please let us know if you have any queries concerning the issues presented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Timeliness** | **Relevancy** | **Uniqueness** | **Validity** |
| Customer Demographics | Erroneous: DOB | Missing: last name, job title, job industry, tenure | Format: gender | Out of date: deceased customer | Corrupted:  default column deleted |  | Unreliable:  age |
| Customer Address |  |  | Format: state |  |  |  |  |
| Transactions | Missing:  profits (cost - price) | Missing:  customer ID (3501-4999), online order, brand, product line/class/size, standard cost, first sold date | Format:  list price, standard cost |  | Filtered: cancelled orders |  | Format:  product sold date |
| New Customer |  | Missing: last name, DOB, job title, job industry |  |  | Irrelevant:  unnamed columns | Duplications:  customers |  |

I have evaluated the datasets according to the data quality dimensions framework as follows:

* Accuracy – correct values
* Completeness – data fields with values
* Consistency – values free from contradiction
* Timeliness – values up to date
* Relevancy – data item with value meta-data
* Uniqueness – records that are not duplicated
* Validity – data containing allowable values

### Accuracy issues:

In Customer Demographics, the DOB is inaccurate, e.g. a data entry indicated a customer was born in 1843. An age column could be derived to improve comprehensibility.

Mitigation: I filtered out the outliers in DOB.

Recommendation: Create an age column for more comprehensible data and to allow checking for errors

### Completeness issues:

In Transactions, the customer IDs are inconsistent among all datasets, with some customer IDs being null.

Mitigation: I included only customer ID from 1-3500.

Recommendation: Ensure tables are up to date. Only complete data will be used. If the data is not in-sync across all spreadsheets, data analysis with incomplete data may skew the results. To prevent future occurrences, it is advised to cross-check spreadsheets for completeness.

In Customer Demographics and New Customer, some records of last name, job title, job industry (as well as tenure and DOB, respectively) are missing.

In Transactions, some records of the online order, brand, product line, product class, product size, standard cost and first sold date are missing.

Mitigation: I filled in the missing null values using forward/backward filling methods or an average of the cluster group where appropriate.

Recommendation: Provide drop-down options for job title, online order and brand column. Convert the first sold date into a standard format. The introduction of filled data may skew the result of data analysis. The use of pre-defined options will allow for more complete data.

### Consistency issues:

For Customer Demographics, the gender was in inconsistent formats.

For Customer Address, the state was in inconsistent formats.

For Transactions, the list price and standard costs were in inconsistent formats.

Mitigation: I found and replaced variations of Men under the category of ‘Male’, and all variations of Women under ‘Female’. Similarly I replaced the names of states to abbreviations, e.g. ‘Victoria’ to ‘VIC’ to ensure consistency across the datasets. The prices were restricted to 2 decimal places for the currency $.

Recommendation: To avoid different representations of the same value, the data type should be categorical rather than a variable text field. Dropdown options minimise inconsistencies and human error in manual entries by different personnel and improves the data interpretability and readability. As gender is a protected characteristic, those identified as others may fall under the category of ‘U’.

### Timeliness issues:

In Customer Demographics, a few customers were reported as deceased so they are not current customers.

Mitigation: I filtered out the customers marked as deceased.

Recommendation: It may be difficult to verify this information, but where available this information should be updated as soon as possible.

### Relevancy issues:

For Customer Demographics, there was a default column with incomprehensible or corrupted data.

For Transactions, the order status showed cancelled orders.

For New Customer, there were several hidden columns.

Mitigation: I dropped the corrupted data columns, filtered out cancelled order status and dropped the hidden columns showing intermediate calculations.

Recommendation: Remove or reformat any incomprehensible meta-data to make it comprehensible. Cancelled order status may be ignored if it is not relevant to the analysis.

### Uniqueness issues:

For New Customer, some of the customers may be duplicated

Mitigation: Some records were filled in in a way which suggests that they relate to the same customer, e.g. same customer ID with the same first name but missing last name.

Recommendation: Ensure each customer record includes first name and last name, with a uniquely identifiable ID that is consistent.

### Validity issues:

For Customer Demographics, there is currently no age column.

For Transactions, the product sale date is an integer which may cause confusion.

Mitigation: I standardised the product sale date and converted list price to currency format.

Recommendation: Ensure that all datasets are from the same time period, otherwise any duplicate or missing data records may skew the data analysis. The datasets have been merged to a single master dataset in a consistent format

### Other data quality issues:

There were many missing datapoints across various features/columns.

Some of the data was also out of sync, i.e. there was some mismatch between the datasets.

Inconsistent data types were used for the same attributes, e.g. integer for some fields and float for others which can introduce unintended bugs due to discrepancy in precision.

Mitigation: If the number of null-value is small, I have filled the records using appropriate statistical methods. Otherwise, if the number of null-value is significant, the records have been dropped from the master datasets. The only exception I made was if the sample size is small and the datapoints are critical. This achieved a standardisation of all fields to achieve constraints on the permitted data types.

I have attached the pre-processed datasets for your review. If you have any questions, suggestions, or specific requirements regarding the preprocessing steps or the upcoming analysis phase, please don't hesitate to reach out. We highly value your input and look forward to further collaboration.

Thank you for your continued support and trust in our team. We remain committed to delivering high-quality results that meet your objectives.

Best regards,

Tamilarasu P

Data Analyst

9789929179