| **Mubashir Bello**  Data Analyst  **ㅡ**  (+234) 816 858 9208  bellomubashir4@gmail.com | horizontal line **Sprocket Central Pty Ltd**  ***KPMG***  Dear Sir,  Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.   | Table name | No. of records | | --- | --- | | Customer Demographic | 3139 | | Customer Address | 3996 | | Transaction Data | 20000 |   I wanted to reach out to you regarding some data quality issues that we have identified in your company's dataset, the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.  During our recent analysis, we found that the following issues are affecting the accuracy and completeness of your data:   1. Inconsistent Data Entry: We noticed that there is a lack of standardization in the data entry process, leading to variations in how data is recorded across the dataset. The following are some of the lists of inconsistent data entries in the dataset    1. The gender data field (Customer Demography table) has a value that has no full meaning in the context of the field and the value is “**u**”. Also, gender has fields with similar meanings but typed in the system with different formats, we have ‘female’, ‘femal’, and ‘f’ which are all similar, and also ‘male’ and ‘m’ are similar.   *Mitigation: Using regular expressions to replace extended values with abbreviations to ensure consistency across addresses.*  *Recommendation: Enforce a dropdown box for the user to select their gender instead of a free box where users can input anything they want.*   * 1. DOB (Customer Demography table): Jephthah's date of birth is 1843-12-21 which is way beyond the average age and it seems to be a mistake because his age will be 180 as of December 2023.   *Mitigation: Since it's just a single individual, it may be due to a typo or an error.*  *Recommendation: Create a constraint where a user can’t be older than 100 or 90 years of age, so it can create a range of ages that the user can select.*   * 1. Job\_industry\_categories (Customer Demography table) This field has a value ‘na’ which is allocated to some job whose category is already defined, for example ‘web developer’ which should be in the category of ‘IT’, 520 users filled na for their job.   *Mitigation: Checking for each job with its relevant categories, using an “IF” statement*  *Recommendation: Enforce a dropdown box for the user to select their job category based on their Job description instead of a free box where users can input anything they want.*   * 1. The default field (Customer Demography table) has no significant meaning in the dataset.   *Mitigation: This field is meaningless or was changed due to some constraints which is not visible to us in the analytics team*  *Recommendation: Meta Data should be provided explaining every field in a dataset.*   * 1. The customer's address (customer address), has some inconsistent data, New souths wales and NSW is similar, also VIC and Victoria are similar.   *Mitigation: Using regular expressions to replace extended values with abbreviations to ensure consistency across addresses.*  *Recommendation: Enforce a dropdown box for the user to select their Location instead of a free box where users can input anything they want.*   * 1. The transaction customer\_id and customer\_address customer\_id have a value that doesn’t exist in the customer\_demography table.   *Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.*   1. Missing Data: The following are fields with their tables with missing values.    1. The customer demography table has lots of missing data some filed with missing values are:       1. The last name field has 125 missing values which may occur because the field is not compulsory       2. The DOB field has 87 missing values.    2. From the transaction table       1. The brand, product line, product class, product size, standard cos, and product first sold date fields all have 197 empty fields.   *Mitigation: If the field that are empty are few (1% of the dataset), then filter out the empty field, if its a field that is need for the analysis then we add (Impute) a value based on some distribution (mean, median or mode).*  Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.  Please let us know if you have any questions or concerns, and we will be happy to discuss this matter with you in more detail. Best regards Mubashir Bello |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |