Subject: Data Review and Suggestions for Improvement

Hello,

I hope this email finds you well. I wanted to share some observations and recommendations after going through the datasets from Sprocket Central Pty Ltd. Here's a summary of the data and a few suggestions for improvement:

Data Quality Issues:

1. **Additional Customer IDs:**
    - **Issue:** Some customer IDs in 'Transactions' and 'Customer Address' are not in 'Customer Demographic.'
    - **Mitigation:** Ensure all tables are from the same period. Only customers in the 'Customer Demographic' will be used for our model.
    - **Recommendation:** Please refer to the 'data_outliers.xlsx' file for the list of outliers between tables.
2. **Empty Values in Certain Records:**
    - **Issue:** Some records have empty values, especially in key columns like brand or job title.
    - **Mitigation:** Remove records with a small number of empty values from the training set. For core fields, consider imputing based on the distribution in the training dataset.
    - **Recommendation:** Addressing missing fields in key datasets, such as transactions, where less than 1% of transactions have missing fields.
3. **Inconsistent Values for the Same Attribute:**
    - **Issue:** Inconsistent representations of the same attribute (e.g., "Victoria" represented as "V," "Vic," and "Victoria").
    - **Mitigation:** Used regular expressions to standardize values.
    - **Recommendation:** Enforce drop-down lists to ensure consistency. Cleaned gender records where 'U' based on the distribution from the training dataset.
4. **Inconsistent Data Types for the Same Attribute:**
    - **Issue:** Different data types for the same attribute (e.g., numeric values and strings).
    - **Mitigation:** Converted selected records to numeric. Removed non-numeric characters from strings.
    - **Recommendation:** Ensure fact tables have constraints on data types for better interpretation.

Next Steps:

Our team will continue with the data cleaning, standardization, and transformation process for model analysis. We may have questions and assumptions along the way. After completion, it would be valuable to sync up with your data SME to ensure alignment.

Looking forward to making progress together.

Best regards,

Kshitij C