Subject: Data Quality Assessment and Recommendations for Sprocket Central Pty Ltd

Dear Sprocket Central,

I hope this email finds you well. I would like to express my gratitude for providing us with the datasets from Sprocket Central Pty Ltd. As requested, our team has conducted a thorough assessment of the data quality and identified several key issues that may impact our upcoming analysis. In this email, I will outline these data quality issues and provide strategies to mitigate them, ensuring that we have a solid foundation for our analysis in phase two.

Customer Demographics Dataset

1. **Data Completeness:**
    - We can see that the table consists of some missing values in certain columns : 'last_name', 'DOB', 'job_title', 'job_industry_category', 'default', 'tenure'.
    - **Mitigation:**
        - As the percentage of missing values in the datasets is low as compared to the whole dataset we can proceed by removing them.
        - Remove the redundant data as it may skew the distribution of the dataset.
2. **Data Accuracy:**
    - In order to ensure more relevancy and readability it's better to use ‘Male’, ‘Female’ and ‘Unidentified’ for the ‘gender’ column.
    - There are some first and last names that contain non-alphabetical characters that are not typical in a person's name. Examples of unexpected characters might include numbers, special symbols, or punctuation marks.
    - It includes job titles that do not conform to common industry roles or standard job titles. Examples include "Budget/Accounting Analyst I," and many others. It also includes job titles that contain special characters, numbers, or symbols not typical in job titles. Examples include "Media Manager I," .
    - The job category for agriculture in the Customer Demographics dataset has been misspelled and there are customers with a missing job category.
    - It is recommended to drop unnecessary columns like ‘default’ columns in the Customer Demographic dataset.
3. **Data Redundancy:**
    - There seems to be redundant information in the datasets which may skew our analysis results and should be deduplicated.
4. **Data Integrity:**
    - Missing values can affect the integrity of your analysis results. For example:
        - If you plan to calculate the average age of customers and a significant portion of DOB values are missing, your analysis might not accurately reflect the true average age.

        - If you intend to segment customers by job title and a substantial number of job titles are missing, the segmentation may be incomplete or less informative.
    - **Mitigation**
        - Verify the consistency of customer IDs across datasets. Any records with unmatched IDs should be investigated and resolved.
5. **Data Currency**
    - There are some customers with an extremely distant past DOB. I have used a reasonable range of the last 100 years for example customer_id 34 with DOB of 1843-12-21.
    - Assuming Sprocket has been in existence for 20 years, there are customer records with tenure values above 20. These records represent customers who have been associated with your company for a longer period than expected based on your business's history.
    - The presence of deceased customers in the dataset has implications for data currency. There are 2 customers marked as deceased, suggesting that there are individuals in the dataset who are no longer alive, according to the recorded information

Customer Addresses Dataset

1. **Data Accuracy:**
    - There are some addresses on the "address" column in your Customer Addresses dataset that don't follow expected patterns or formats such as 123 Main Street
2. **Data Consistency:**
    - Some states in the Customer Addresses dataset are invalid.
3. **Data Redundancy:**
    - There seems to be redundant information in the datasets.
4. **Data Relevancy**
    - How relevant is the Property_Valuation column in the Customer Addresses dataset?

Transactions Dataset

1. **Data Completeness**
    - There are gaps in the Transaction data for certain time periods, potentially leading to incomplete insights into recent customer behavior.
    - "online_order", "brand," "product_line," "product_class," "product_size," "standard_cost," and "product_first_sold_date" have missing values which may affect analyses related to online purchasing behavior. For example, if you want to analyze online vs. offline purchase trends, missing values may skew the results.

2. **Data Accuracy:**
    - The presence of outliers in the "standard_cost" column in the Transactions dataset indicates that some data points have values significantly different from the majority of the data.
        - If outliers are due to data errors, consider correcting or removing those data points.
        - If outliers represent valid but extreme values, decide whether to keep or transform them for analysis or modeling purposes
3. **Data Consistency:**
    - There are 197 dates missing from the “product_first_sold_date” in the Transactions dataset.
    - There should be a separate dataset that holds product information with product_id as the foreign key in Transactions dataset.
    - In the Transactions dataset, brand, product_line, product_size and product_class have a category called nan which should be invalid.
4. **Data Integrity:**
    - We found several records in the Transaction dataset that do not have matching customer IDs in the Customer Demographic dataset.
5. **Data validity**
    - The "online_order" column contains only values, 0, 1 and nan. It should only contain 0 and 1 which represents binary online order status.
    - The “brand” column contains unique categories 'Solex', 'Trek Bicycles', 'OHM Cycles', 'Norco Bicycles', 'Giant Bicycles', 'WeareA2B' and nan.
    - The ‘product_first_sold_date’ has been assigned to the wrong datatype of float instead of datetime.

In addition to these strategies, our team will also perform data profiling and exploratory data analysis to uncover any other data quality issues that may arise during the analysis process.

Please feel free to reach out if you have any questions or require further clarification on these recommendations. We look forward to collaborating with you on this project and delivering actionable insights that drive company growth.

Thank you for entrusting us with this opportunity.

Best Regards,

Lynsey Bwisa
Data Scientist