KPMG’s Sprocket Central Pty Ltd.: Customer Analytics for Strategic Targeting Project Report

Data Analysis Life Cycle to be followed:

1. Understanding Business Problem
2. Data Collection and Preparation (Data Cleaning and Standardization Using MS Excel)
3. Exploratory Data Analysis (Deep Dive Using SQL)
4. Generating Insights (Finalizing Potential Customers)
5. Communicating Findings (Conclusion)

Problem Statement (Understanding Business Problem):

- Help Sprocket Central Pty Ltd identify which 1000 new customers they should target based on historical and new customer data.
- Find the top 1000 *potential customers* who are most likely to be profitable, engaged, or aligned with Sprocket Central’s ideal customer profile—based on patterns from existing high-value customers.

Data Collection:

The datasets used in this project were sourced from Kaggle, originally provided as part of KPMG’s virtual internship program hosted on the Forage platform. Although the internship is no longer accepting submissions, I undertook the project independently to gain practical, hands-on experience with real-world datasets and to better understand the typical challenges a data analyst encounters during the course of an end-to-end project.

Data Preparation:

This section outlines the comprehensive data preparation and cleaning procedures undertaken for the KPMG’s Sprocket Central Pty Ltd. datasets. Initial preprocessing and transformation were meticulously performed in Microsoft Excel, providing detailed control over each step before importing the cleaned data into MySQL Workbench for analysis.

1. Transaction Dataset (transaction dataset kpmg)

The following key data preparation steps were executed to enhance the quality and usability of the transaction data:

- **Missing Value Management:**

- ✓ Identified a total of 1,542 missing values across the dataset.
- ✓ Specific columns with notable missing values included online_order (360 missing), brand, product_line, product_class, product_size, updated_standard_cost, and updated_product_first_sold_date (197 missing each). These were either imputed or flagged for consideration in subsequent analysis.

- **Date Format Standardization:** Converted the original transaction dates into a consistent YYYY-MM-DD format, creating the updated_transaction_date column for improved analytical compatibility.
- **Binary Flag Validation:** Verified the online_order column, confirming 19,641 valid entries as either "TRUE" or "FALSE" to ensure accurate binary representation.
- **Order Status Quality Check:** Confirmed all 20,001 order_status entries were consistently categorized as either "Approved" or "Cancelled," ensuring data integrity for transaction outcomes.
- **Numeric Field Standardization:** Transformed list_price and standard_cost into new columns, updated_list_price and updated_standard_cost, respectively, ensuring all values were correctly formatted as numerical data.
- **Product Sold Date Formatting:** Standardized the product_first_sold_date into a YYYY-MM-DD format, creating updated_product_first_sold_date.

2. Customer Address Dataset (customer_address kpmg)

The customer address data underwent the following preparation to ensure accuracy and consistency:

- **Missing Value Verification:** Conducted a thorough check for missing values across all columns and confirmed that the dataset contained no blanks, indicating high data completeness.
- **State Name Standardization:** Introduced an updated_state column to convert abbreviated state names (e.g., "NSW", "QLD", "VIC") into their full, descriptive names ("New South Wales", "Queensland", "Victoria") for enhanced readability and uniformity.

3. Customer Demographics Dataset (customer_demographics kpmg)

Extensive cleaning and transformation were applied to the customer demographics data to prepare it for detailed analysis:

- **Missing Value Management:** Identified and addressed 805 missing values across various demographic fields, requiring careful consideration for subsequent analysis.
- **Name Consolidation:** Merged first_name and last_name into a single customer_name column, applying proper capitalization for consistent presentation.
- **Gender Normalization:** Created an updated_gender column to unify inconsistent gender labels, standardizing entries like "F", "Femal", and "Female" to "Female"; "M" and "Male" to "Male"; and mapping "U" to "Not Specified."
- **Age Calculation from Date of Birth (DOB):** Calculated customer_age using the DOB field. Robust error handling was implemented to manage outliers and blank

entries, categorizing them as "Date Not Mentioned" or "Date Too Old" where applicable.

- **Job Industry Category Cleaning:** Standardized the job_industry_category column by replacing "n/a" entries with "Other Industry" to ensure all customers were assigned to a recognized category.
- **Deceased Indicator Mapping:** Transformed the binary deceased_indicator values ("Y" and "N") into more descriptive labels, "Deceased" and "Not Deceased," respectively, for clarity.

4. New Customer List Dataset (new_customer_list kpmg)

Similar to the historical datasets, the new customer list underwent critical preparation steps:

- **Missing Value Management:** Identified and addressed 152 missing values across this dataset, ensuring completeness for profiling.
- **Name Construction:** Created new_customer_name by concatenating first_name and last_name, applying proper case formatting.
- **Gender Standardization:** Unified "U" entries in the gender column as "Not Specified" for consistency with the historical demographics.
- **Age Calculation:** Derived customer_age from the DOB and implemented handling for empty DOB cells, marking them as "Date Not Mentioned."
- **Job Industry Category Update:** Applied the same standardization logic as the demographic dataset, replacing "n/a" values in job_industry_category with "Other Industry."
- **Deceased Indicator Mapping:** Converted the binary "Y"/"N" deceased_indicator values into "Deceased"/"Not Deceased."
- **State Name Expansion:** Replaced abbreviated state names with their full names to maintain consistency across all address-related data.

Exploratory Data Analysis (Using SQL):

1. Introduction and Project Context

This report details the Exploratory Data Analysis (EDA) performed on KPMG's Sprocket Central customer datasets. The primary objective of this phase is to gain a deep understanding of both historical and potential new customers to inform targeted marketing and customer acquisition strategies. This analysis is the first stage in a four-part project, laying the foundation for profiling ideal customers, analyzing new customer segments, and ultimately, defining final customer targeting recommendations.

The analysis leveraged SQL queries within MySQL Workbench, building upon initial data cleaning and preparation.

2. Data Overview

The analysis utilized the following key datasets:

- transaction dataset kpmg: Contains historical transaction records.

- customer_demographics kpmg: Provides demographic and personal information for historical customers.
- new_customer_list kpmg: Contains demographic and a pre-calculated "Value" score for potential new customers.
- customer_address kpmg: Contains address details for customers.

3. Historical Customer Dataset Analysis

3.1. Duplicate Value Analysis

A crucial first step was to ensure data integrity by checking for duplicate entries across all primary keys in the datasets. Queries were executed for transaction_id in transaction dataset kpmg, customer_id in customer_demographics kpmg and customer_address kpmg, and new_customer_name in new_customer_list kpmg.

Finding: No duplicate values were found in any of the primary key columns across the datasets, confirming the successful data cleaning and ensuring the reliability of subsequent analysis.

3.2. Customer Demographics & Segmentation

Analysis of the customer_demographics kpmg dataset provided foundational insights into the existing customer base:

- **Wealth Segment Distribution:** The "Mass Customer" segment holds the largest share of customers (1848), followed closely by "High Net Worth" (933) and "Affluent Customer" (905). This indicates a broad customer base across various wealth tiers, with a notable concentration in the mass market.
- **Job Industry Category Distribution:** "Manufacturing" (743 customers) and "Financial Services" (729 customers) are the most dominant job industries, suggesting these sectors are significant sources of customers. "Health" (567) and "Other Industry" (614) also contribute substantially.
- **Age Group Distribution:** The majority of customers fall into the "Above 35" age group (2919 customers), followed by "26-35" (657 customers), and "18-25" (110 customers). This highlights that the customer base skews older and more mature.
- **Car Ownership:** (Data not provided in current report, but was part of analysis scope.)
- **Average Customer Tenure:**

  - ➢ **Age-wise:** A strong positive correlation was observed between age and tenure. "Above 35" customers have the longest average tenure (12 years), significantly higher than "25-35" (5 years) and "18-25" (1 year).
  - ➢ **Gender-wise:** Female and Male customers show a highly consistent average tenure of 11 years.
  - ➢ **Job Industry Category-wise:** Most industries maintain an average tenure of 11 years, with slight variations (e.g., IT and Telecommunications at 12 years, Other Industry and Retail at 10 years).
  - ➢ **Wealth Segment-wise:** All wealth segments ("Mass Customer," "Affluent Customer," "High Net Worth") exhibit a consistent average tenure of 11 years.

3.3. High-Value Customer Identification & Profiling

Historical customers were segmented into three high-value categories to understand different facets of customer worth. The demographic findings across these three categories showed remarkable consistency, allowing for the synthesis of a robust "Ideal Historical Customer Profile."

Key Findings Across High-Value Historical Segments:

Highest Total Spending Customers (Top "Whales"):

- **Age:** Predominantly "Above 35" (797 customers), followed by "26-35" (176).
- **Wealth Segment:** "Mass Customers" are the largest group (500), with "High Net Worth" (251) and "Affluent Customer" (249) following closely.
- **Job Industry:** "Financial Services" (206) and "Manufacturing" (199) are leading.
- **Gender:** Females (503) slightly outnumber Males (496).

Highest Average Spending Customers (Top "Big Ticket" Purchasers):

- **Age:** Similar to total spending, "Above 35" dominates (782), followed by "26-35" (181).
- **Wealth Segment:** "Mass Customers" again lead (504), with "Affluent Customer" (253) and "High Net Worth" (243).
- **Job Industry:** "Manufacturing" (201) and "Financial Services" (197) remain top.
- **Gender:** Females (515) slightly lead Males (484).

Top Most Frequent Purchasers (Top "Loyalty" Customers):

- **Age:** Consistently "Above 35" (793), followed by "26-35" (184).
- **Wealth Segment:** "Mass Customers" are the largest (507), then "High Net Worth" (255) and "Affluent Customer" (238).
- **Job Industry:** "Manufacturing" (200) and "Financial Services" (192) are prominent.
- **Gender:** Females (514) slightly lead Males (485).

3.4. Ideal Historical Customer Profile:

Based on the consistent trends across the high-value historical customer segments, an "Ideal Historical Customer Profile" can be defined:

- **Age:** Predominantly **Above 35 years old**. This age group consistently represents the largest and most valuable demographic.
- **Wealth Segment:** Primarily **"Mass Customers"**, with significant contributions from "High Net Worth" and "Affluent Customers."
- **Job Industry Category:** Strong concentration in **"Financial Services"** and **"Manufacturing"**, with "Health" and "Other Industry" also being key contributors.
- **Gender:** Slightly more **Female** customers are identified as high-value, but **Males** are a very close second, indicating that both genders contribute significantly.

- **(Pending additional analysis like Car Ownership, Tenure, Past 3 Years Bike Purchases, and States/Regions for a more complete picture).**

4. New Customer Dataset Analysis

4.1. Duplicate Value Analysis

As noted in Section 3.1, no duplicates were found in the new_customer_name column, indicating data cleanliness.

4.2. Analysis of Top Potential New Customers (using "Value" Column)

Since no transactional history is available for new customers, the provided "Value" column, a pre-calculated score representing potential, was used as the primary metric to identify the top prospective customers. A demographic analysis was then performed on this top segment (approximately 1000 customers with the highest "Value" score).

Key Findings (Top Potential New Customers):

- **Age-wise:** Overwhelmingly "Above 35" (787 customers), followed by "26-35" (163), and "18-25" (34).
- **Wealth Segment:** "Mass Customer" is dominant (500), followed by "High Net Worth" (249) and "Affluent Customer" (235).
- **Job Industry Category:** "Financial Services" (202) and "Manufacturing" (199) are the leading categories, with "Other Industry" (165) and "Health" (152) also significant.
- **Gender:** Females (513) lead Males (471) in this segment.

5. Comparative Analysis & Key Insights

The most critical insight from this EDA is the **striking alignment between the "Ideal Historical Customer Profile" and the demographic profile of the "Top Potential New Customers."**

- **Age Consistency:** Both groups are predominantly "Above 35."
- **Wealth Segment Similarity:** "Mass Customers" lead in both, with higher wealth segments showing similar proportional representation.
- **Job Industry Overlap:** "Financial Services" and "Manufacturing" are consistently strong in both historical high-value customers and new potential customers.
- **Gender Balance:** A slight female lead, with strong male representation, is observed in both analyses.

This strong alignment provides significant validation:

1. **Validation of 'Value' Column:** The "Value" score for new customers appears to be an effective predictor, as it identifies individuals whose demographic traits closely mirror those of historically valuable customers.
2. **Confidence in Targeting:** Sprocket Central Pty Ltd. can have high confidence in targeting these identified "Top Potential New Customers" because their profiles

suggest they are likely to exhibit similar purchasing behaviors and become valuable, long-term customers.

This Exploratory Data Analysis has successfully identified the key characteristics of KPMG's Sprocket Central Pty Ltd.'s most valuable historical customers and, crucially, found a strong demographic resemblance in the top potential new customers. This alignment provides a robust, data-driven foundation for highly effective customer acquisition and targeting strategies, aiming to maximize marketing ROI by focusing on the most promising leads.

Generating Insights (Finalizing Top Potential New Customers):

Following the comprehensive analysis of the historical customer dataset and the subsequent establishment of an "Ideal Historical Customer Profile," the focus shifted to identifying and profiling the top potential customers from the new customer dataset.

Methodology for Identifying Top Potential New Customers

Given the absence of transactional history for new customers, the pre-calculated **"Value"** column within the new_customer_list kpmg dataset was utilized as the definitive metric for assessing customer potential. This column provides a proprietary score indicating a customer's likely worth or desirability to Sprocket Central. The top approximately 1000 new customers were selected based on their highest "Value" scores.

Profiling of Top Potential New Customers

A detailed demographic analysis was conducted on this identified segment of top potential new customers to understand their inherent characteristics. The key findings are as follows:

- **Gender: Female** customers are the dominant group within the top potential new customers, slightly outnumbering their male counterparts.
- **Age Distribution:** Customers **Above 35 years old** are overwhelmingly dominant in this segment, consistent with the overall customer base. The "26-35" age group follows, with "18-25" being the smallest proportion.
- **Job Industry Category: "Financial Services"** and **"Manufacturing"** remain the most prevalent job industry categories among the top potential new customers, underscoring their significance. Other notable industries include "Other Industry" and "Health."
- **Wealth Segment:** The largest proportion of top potential new customers falls into the **"Mass Customer"** segment, followed closely by "High Net Worth" individuals. "Affluent Customers" constitute a smaller, but still relevant, portion.

Alignment with Ideal Historical Customer Profile

A critical finding from this analysis is the **remarkable consistency and strong alignment** between the demographic characteristics of these top potential new customers and the previously established "Ideal Historical Customer Profile." This congruence is observed across all key parameters:

- **Age:** Both historical high-value customers and **high-potential new customers** are predominantly **"Above 35."**
- **Wealth Segment:** "**Mass Customers**" and "**High Net Worth**" categories are highly represented in both groups, confirming the broad appeal of Sprocket Central's offerings across different economic tiers.
- **Job Industry:** "**Financial Services**" and "**Manufacturing**" consistently emerge as leading industries for both historically valuable and newly identified potential customers.
- **Gender:** A similar slight lead by **female** customers, with strong representation from **males**, is observed in both datasets.

The strong demographic resemblance between the top potential new customers (identified by the "Value" column) and the most valuable historical customers provides significant validation. This indicates that the "Value" score is effectively identifying individuals who share the core characteristics of Sprocket Central's most engaged and profitable customer segments. This clarity empowers highly targeted acquisition strategies, ensuring that marketing efforts are concentrated on individuals most likely to convert into valuable, long-term customers. The finalized list of these top potential new customers, rich with their demographic details, serves as the direct output for targeted outreach.

Communicating Findings (Conclusion):

This comprehensive analysis for KPMG's Sprocket Central Customer Analytics and Targeting project has successfully moved from meticulous data preparation to deriving actionable insights from both historical and new customer datasets.

The initial phase involved rigorous data cleaning and standardization across all key datasets (transaction dataset kpmg, customer_demographics kpmg, customer_address kpmg, and new_customer_list kpmg). This critical step ensured data integrity, addressing missing values, standardizing formats, and normalizing inconsistent entries, thereby providing a robust foundation for reliable analysis.

Our Exploratory Data Analysis (EDA) on historical customer data revealed a clear "Ideal Historical Customer Profile." This profile consistently describes Sprocket Central's most valuable customers (those with highest total spending, highest average spending, and most frequent purchases) as predominantly **Above 35 years old, primarily "Mass Customers" (with significant contributions from "High Net Worth" and "Affluent" segments), working in "Financial Services" or "Manufacturing," and showing a slight female lead in gender distribution.** These insights are crucial for understanding the core loyal customer base.

Crucially, the analysis of the new customer dataset, leveraging the pre-calculated "Value" score to identify top potential leads, demonstrated a **striking and highly significant demographic alignment** with this "Ideal Historical Customer Profile." The top potential new customers exhibit remarkably similar characteristics across age, wealth segment, job industry category, and gender distribution.

This strong congruence leads to a powerful conclusion: The "Value" score for new customers is effectively identifying individuals who mirror the traits of KPMG's most valuable and engaged existing customer base. This validation provides high confidence in the effectiveness of focusing acquisition efforts on these identified individuals.

In summary, our key findings enable clear and confident recommendations:

- **Targeted Acquisition:** KPMG can confidently prioritize its customer acquisition strategies by focusing marketing efforts on the approximately 1000 new potential customers identified by their high "Value" scores.
- **Profile-Driven Messaging:** Campaigns should be specifically tailored to resonate with the shared demographic profile of "Above 35" individuals, primarily "Mass Customers" (including High Net Worth and Affluent), and those in "Financial Services" or "Manufacturing."
- **Maximized ROI:** By concentrating resources on these highly aligned and validated segments, KPMG is positioned to maximize its return on marketing investment, converting new leads into valuable, long-term customers who mirror the success observed in the historical customer base.

This project phase successfully delivers a data-driven understanding of who Sprocket Central's most valuable customers are and, critically, identifies precisely which new prospects share those characteristics, setting a solid foundation for impactful customer targeting.

Author: Himanshu Kumar

Gmail: ds.himanshu.kumar@gmail.com