1. INTRODUCTION

The KPMG AU Data Analytics virtual internship offers an immersive learning experience designed to equip participants with practical skills in data analysis, visualization, and problem-solving. Through hands-on projects and real-world datasets, participants gain insights into the field of data analytics and its applications across various industries.

1.1. PROJECT BACKGROUND

Sprocket Central Pty Ltd[1], a medium size bikes & cycling accessories organization, has approached Tony Smith (Partner) in KPMG's Lighthouse & Innovation Team.

Relying on KPMG's expertise in its Analytics, Information & Modelling team, the Sprocket Central Pty Ltd needs help with its customer and transactions data. The organization has a large dataset relating to its customers, but their team is unsure how to effectively analyze it to help optimize its marketing strategy.

In order to support the analysis, the following stages are designed:

Phase 1 – Data Quality Assessment

The task requires to review the data quality to ensure that it is ready for our analysis in phase two. It is mandatory to issue the list of notes with any assumptions or issues that needed to go back to the client on. As well as recommendations going forward to mitigate current data quality concerns. The Data Quality Framework indicate a list of Data Quality Dimensions for evaluating the dataset.

Phase 2 – Exploratory Data Analysis

The stage aims to provide a comprehensive understanding of the dataset, offering valuable insights that can guide strategic decisions and drive business growth. The primary focus of this phase is to perform a detailed exploration of the dataset to identify correlations, distributions, and anomalies in the data.

Phase 3 – Data Insights and Visualization

The core analysis revolves around an RFM (Recency, Frequency, Monetary) analysis, a customer segmentation technique. This involves segmenting customers based on transaction behaviors and assigning scores to distinct categories.

Phase 4 – High-value customers prediction

By leveraging customer data, historical transaction data, and demographic information, the predictive models can be developed to identify and target high-value customers.

1.2. APPROACH AND TOOLS

Throughout this analysis, Python's pandas, seaborn, and matplotlib libraries play a pivotal role in manipulating, visualizing, and gaining insights from the dataset.

[1] Sprocket Central Pty Ltd is a fictional company named for the purposes of this virtual internship

2. DATA QUALITY ASSESSMENT

As part of the KPMG AU Data Analytics virtual internship, one of the tasks assigned was focused on data cleaning and preparation. The objective was to perform a comprehensive data quality assessment and cleansing process to ensure the dataset was accurate, complete, and ready for subsequent analysis.

2.1. DATASET AND DATA QUALITY DIMENSIONS

Sprocket Central Pty Ltd has a large dataset relating to its customers, but their team is unsure how to effectively analyze it to help optimize its marketing strategy.

The client provided KPMG with 4 datasets:

- Customer Demographic
- Customer Addresses
- Transaction's data
- NewCustomerList

The following list of the Data Quality dimensions has been used to evaluate dataset: Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, Uniqueness.

The following sub-chapters describe outputs of the preliminary data exploration and identify ways to improve the quality of Sprocket Central Pty Ltd.'s data.

2.2. ASSESSMENT RESULTS

The main characteristics of given datasets are shown it the table below:

Table 1 – Data Profiling

Dataset	# of records	# of unique records	# of columns	percentage of missing values
Transactions	20,000	20,000	13	< 1%
NewCustomerList	1,000	1,000	23	1.4%
CustomerDemographic	4,000	4,000	13	3.4%
CustomerAddress	3,999	3,999	6	0%

Transactions table contained 20,000 records providing information on transactions made by 3,494 distinct customers for 101 distinct products and 6 brands from the year 2017. No duplicate transactions were found. The following data quality issues has been defined during assessment:

- 358 records (2% of transactions) had unspecified *online_status*, which can be filled as 'unspecified' to keep those transactions for further analysis.
- 197 records (1% of transactions) represent missing product attributes (brand, size, class, standard costs) and could be removed from the dataset.
- more information needed on what the column *product_first_sold_date* refers to.
- it should be mentioned that there is no column for quantity sold was observed.

CustomerDemographic table represents data related to 4,000 customers indicating names, genders, birthdates, job titles and other information. The following data quality issues has been defined during assessment:

- the *gender* column contains not allowable values (misspelling and different format) which can be replaced with F/M/U
- the following attributes contain missing values: *job_title* (497 customers), *job_industry_category* (656 customers), *last_name* (125 customers). Depending on analysis purposes it is possible to keep those customers replacing blank values with 'unspecified' category.
- DOB doesn't match the range constraint for 1 customer with 1843-12-21 indicated. Also, DOB is missing for 87 customers, for whom data in *default* and *tenure* columns are also missed.
- the column *default* is not interpretable thus cannot be used for further analysis.

CustomerAddress table provides data related to addresses, postcodes, states and countries for 3,999 customers referring to *customer_id* as foreign key. The following data quality issues has been defined during assessment:

- the *state* column doesn't meet the validity requirements due to different approaches in state naming. It`s possible to replace values and bring the data into one standardized format (e.g., NSW, QLD etc.)
- the following IDs are missing when *customer_id* used as foreign key and data has been merged with **CustomerDemographic** table: 3, 10, 22, 23, 4001, 4002, 4003.
- more information needed on what the column *property_valuation* refers to.

NewCustomers table expanded **CustomerDemographic** data with 1,000 new customers. The following data quality issues has been defined during assessment:

- there is no ID column for customers which could be used as primary key for further analysis.
- DOB data is missed for 17 customers.
- there are 5 unnamed columns contained numeric data with a lack of context.

The Data Quality Assessment results are shown in the table below:

Table 2 – Data Assessment Matrix

Dataset	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity	Uniqueness
Transactions	✓	✓	✓	✓	✓	✓	✓
NewCustomerList	✓	✗	✓	✓	✗	✗	✓
CustomerDemographic	✓	✓	✓	✓	✓	✓	✓
CustomerAddress	✓	✓	✓	✓	✓	✓	✓

As a conclusion, the following mitigations can be applied to improve the accuracy of the underlying data:

- Empty values within core fields can be entirely filtered out from the resulting set if imputing is irrelevant;
- For categorical fields empty values can be replaced with 'unspecified' category depending on analysis purposes;
- Regular expressions can be used to replaced extended values into abbreviations to ensure consistency;
- Appropriate data transformation can be made to ensure consistent data types for a given field;

2.3. RECOMMENDATIONS FOR IMPROVING DATA QUALITY

The data quality assessment revealed notable issues in the dataset, prompting the implementation of effective strategies to address these inconsistencies. In light of these findings, a set of recommendations has been developed to proactively prevent future data quality issues and enhance the precision of the foundational data crucial for informed business choices.

- Root-cause analysis

Identifying the root causes of data inconsistencies is a critical step in ensuring the accuracy and reliability of the data. The following approaches can be used for determining root causes:

  - mapping the end-to-end data flow,
  - tracking the origin and transformations of data across different systems,
  - examining data sources,
  - thoroughly analyzing the data through data profiling techniques.

By understanding the underlying reasons for data inconsistencies, organization can implement targeted and effective solutions to address them.

- Data Quality Rules

Data quality rules and thresholds play a crucial role in assessing and maintaining the integrity of organizational data with the following key steps:

  - identifying the most critical data attributes for further decision-making process,
  - defining quantifiable quality metrics for each data attribute (data quality dimensions),
  - establishing clear thresholds or ranges for each quality metric,
  - developing validation rules to assess data quality,
  - automating data quality assessment process whenever possible.

In conclusion, establishing data quality rules and thresholds is a foundational step in maintaining trustworthy and valuable data assets.

- Maintain and update Issue log

An issue log serves as a centralized repository to document and manage identified data quality issues. It provides a structured approach to tracking, addressing, and resolving issues, ensuring that data remains accurate and reliable. The following components can be used to properly maintain the issue log:

  - definition of data quality issue,
  - identification of the underlying causes of the issue,
  - assessment of potential impact on business operations,
  - assigned responsibility for addressing and resolving the issue,
  - resolution steps and planned activities,
  - progress of issue resolution status
  - timestamp recorded for issue identification, assignment, resolution and closure.

By proactively addressing data quality challenges, organizations enhance their ability to make well-informed decisions and drive successful outcomes.