|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

|  |
| --- |
| **KPMG DATA ANALYTICS VIRTUAL INTERNSHIP PROGRAM ON FORAGE** |
| **Final Draft Report** |
| Task 1 – Data Quality Assessment  Task 2 – Data Insights  Task 3 – Data Visualization  Issued by: Nikita Tymoshenko  Document date: 07.08.2023 |

**Contents**

[1. Introduction 2](#_Toc142263383)

[1.1. Project Background 2](#_Toc142263384)

[1.2. Approach and Tools 2](#_Toc142263385)

[2. DATA QUALITY ASSESSMENT 2](#_Toc142263386)

[2.1. Task overview 2](#_Toc142263387)

[2.2. Assessment results 2](#_Toc142263388)

[3. DATA INSIGHTS 2](#_Toc142263389)

[4. DATA VISUALIZATION 2](#_Toc142263390)

# 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.

## Project Background

Sprocket Central Pty Ltd[[1]](#footnote-1), a medium size bikes & cycling accessories organisation, 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 organisation has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise 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 – Data Insights

Some text about task

Phase 3 – Data Visualization

Some text about task

## Approach and Tools

Some text about python colab github RFV analysis and what else.

# 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.

## Task overview

Sprocket Central Pty Ltd has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise 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.

## 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 | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| NewCustomerList | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |
| CustomerDemographic | Close with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |
| CustomerAddress | Checkmark with solid fill | Checkmark with solid fill | Close with solid fill | Checkmark with solid fill | Close with solid fill | Close with solid fill | Checkmark with solid fill |

|  |
| --- |
| Apart from mentioned inaccuracies and data quality issues the following recommendations should be used to improve data quality and reliability for further analysis:   * determine issue root-causes for further resolution; * define Data Quality rules, thresholds to perform data quality assessments; * review data policies and procedures; * identify typical data quality issues and keep issue log updated; * implement validation control for specified fields, set fields with missing data as mandatory. |

# DATA INSIGHTS

# DATA VISUALIZATION

Subject: Data Quality Assessment and Identified Issues

Dear Manager,

I hope this email finds you well.

I wanted to provide you with an update on the recent data quality assessment I conducted on the provided datasets. I have outlined the main issues I identified along with the steps I took to address them in the attached report.

2. New Customer Table:\*\*

1. The table includes 1,000 customers without assigned IDs.
2. 17 customers lack specified dates of birth (DOB).
3. Columns without names containing calculated values were removed.
4. Customers Demographic Table:

\* Initially, 4,000 distinct customer IDs were present.

\* No duplicated records were found.

\* Gender values were replaced with F/M/U.

\* Missing data in job-related columns (job\_title, job\_industry\_category, last\_name) was filled with 'unspecified'.

\* DOB values before 1931 and inconsistent values were removed.

\* The unclear 'default' column was dropped.

\*\*Customer Address Table:\*\*

\* Addresses are available for 3,999 customers.

\* No duplicate records were identified.

\* No missing values were present.

\* State names were standardized (New South Wales to NSW, Victoria to VIC).

\* Differences in customer IDs between tables were noted (e.g., missing IDs: 3, 10, 22, 23, 4001, 4002, 4003).

\* The context of the property\_valuation column is unclear.

\* Your valuable feedback and suggestions for further improvement are welcome.

1. Sprocket Central Pty Ltd is a fictional company named for the purposes of this virtual internship [↑](#footnote-ref-1)