KPMG Virtual Internship Notes Document

Overview

Task 1

- Perform a data quality assessment and write an email to the client outlining any issues as well as mitigation strategies to improve the quality of insights that can be gained in later data processes.

Task 2

- Create a presentation that will help educate the client on why they should sign off on the 3-week project aimed to dive deeper into determining which *new* customers are likely to drive more value and thus, the client should target in marketing efforts.

Task 3

- Create an interactive dashboard that highlights key findings
  - *I also plan to create a report to expand on these findings in a succinct manner.*

Project Context

Sprocket Central Pty Ltd is a long-standing KPMG client who specializes in high-quality bikes and accessible cycling accessories to riders. Their marketing team is looking to boost business by analyzing their existing customer dataset to determine customer trends and behavior.

Task 1 - Data Assessment

Email Draft

Hi Sprocket Team,

Attached is the comprehensive Data Quality Assessment. Below, is an overview of key issues, assumptions and mitigation strategies:

Table	***Key Issues***	***Mitigation Actions***
CustomerDemo	Last Name	Update high-value fields to be

	● 125 blank records Gender ● Non-streamlined data (ex: male could be ‘M’ or “Male’) DOB ● Incorrect data type ● One record with 1843 as birth year ● 87 blank records Tenure ● 87 blank records	required to have a value prior to form/checkout submission. Also, streamline data inputs for fields like Gender and DOB with dropdowns.
Transactions	Incorrect Data Types ● Product First Sold Date Key Fields with Blanks ● Online Order ● Brand, Product Line, Product Class, Product Size, Standard Cost & Product First Sold Date ○ *All of these rows have Order Status = ‘Approved’*	It is unclear which field should be used to determine total revenue (ex: list_price). Recommended to ensure fields are setup to accept specific data types only. ● Also, it may be valuable to understand why there are no values for fields in the Brand, Standard Cost,... grouping and what type of service was performed to generate revenue.
NewCustomerL	No straightforward primary key to join this table with others. Data Types Converted to Number: ● Postcode ● Property Valuation ● Past 3 Years Bike Related Purchases	Data in this table is valuable for understanding your customers better and improving future customer acquisition - identify a primary key to merge this data with other tables. Ensure data types are proper and set certain fields to ‘required.’ Set a Universal Rounding Rule for numeric fields (ex: Value).
CustomerAddr	There are 3 customer_id records in this table that are not referenced in other tables;	1. Understand how some data is added to this table and not others. Determine if it is an

	○ 4001 ○ 4002 ○ 4003	issue. 2. If it is an issue, update ETL processes for organizing collected data.

Key Assumptions

Overall

- Performing this analysis on January 1, 2018.
- GDPR/AU's data policies comply with regulations.

CustomerDemographics

- Tenure represents how long they have been a customer of Sprocket for.
- "Past 3 Years Bike Related Purchases" - this field represents the number of purchases the customer made in the past 3 years, related to bikes.

Transactions

- Use the List_Price field to gain insights into total revenue.

NewCustomerList

- Create a compound primary key using Address, Postcode, State to join with/use the CustomerAddress table as a junction table to the CustomerDemographic & Transactions tables

CustomerAddress

- The primary key, cusotmer_id, should line up seemingly seamlessly with the CustomerDemogrpahic table.
- Duplicate addresses, containing unique customer_ids, represent different people in the same household

Please review, confirm field definitions and our assumptions, and reach out with further questions. As next steps, we plan to clean the data based on your insight and provide you with a presentation outlining the approach we will take to identify the 1000 customers you should target.

Best,
Kevin Gregersen

KPMG Data Analyst Team

Data Quality Assessment

CustomerDemographic Table

4000 total rows

DOB

- Customer_id has DOB in 1843
    - Likely a typo for 1943
- 87 blank values
- Converted the data type to date format

Last Name

- This field is blank for 125 records
- **Mitigation Action**
    - Last names are typically valuable to capture. To ensure they are captured update the field to be required to have a value before the user is able to submit form/purchase.

Gender

- Non-streamlined values. Ex: Female can have the input of 'F', 'Femal', 'Female'
    - 'F' only appears as a value for cusotmer_id 1
    - 'Femal' value only appears for customer_id 54
- Similar case with Male values being specified as 'M' or 'Male'
    - 'M' only appears for customer_id 57
- There seems to be an unspecified value 'U' that appears 88 times in the dataset. (determine if Sprocket offered this field as an option for a user to select or they have a way to define it)
- **Mitigation Action**
    - To gather data for the gender field, it is recommended to this field be a dropdown with predefined values of either M/F or Male/Female

Job Title

- Note: 506 blank values. If this field is expected to provide value via something like segmenting your customer base, try making it a required field or provide options for a user to select in a dropdown.

Job Industry Category

- Note: 656 rows with 'N/A' values

Wealth Segment

- Side question: how is this determined?

Default

- Messy data in this column with special characters in some rows, dates in others and file paths in others. Determine what should be in it or remove it.

Tenure

- What type of tenure does this refer to? Maybe how long they have been coming to Sprocket for?
- There are 87 blanks.

Transactions Table

20,000 total rows covering transaction data between 1/1/2017 - 12/30/2017

Online Order
- 360 blanks

Order Status
- Note: 179 records were 'Cancelled'

Brand, Product Line, Product Class, Product Size, Standard Cost & Product First Sold Date
- 197 row are blank. All of these rows have Order Status = 'Approved'
    - Maybe these were for service only charges?
    - They all have a List Price value

Product First Sold Date
- Incorrect data type. Should be updated to Date

NewCustomerList Table

1000 rows

Issue: There is no straightforward primary key to join this table with others. Possibility: create a compound primary key using Address, Postcode, State to join with/use the CustomerAddress table as a junction table to the CustomerDemographic & Transactions tables

Last Name
- 29 blanks

Past 3 Years Bike Related Purchases
- Converted data type from text to integer

DOB
- 17 blanks

Postcode
- Converted data type from text to integer

Property Valuation
- Converted data type from text to integer

Value
- Recommended to set a universal rounding rule

CustomerAddress Table

4000 rows. Data looks pretty clean.

- One thing to note is there are 3 customer_id records in this table that are not referenced in other tables;
    - 4001
    - 4002
    - 4003
- ***State***
    - This observation was made while completing Task 2:
        - NSW = New South Wales
        - VIC = Victoria
    - Mitigation: clean the data by replacing New South Wales with NSW & Victoria with VIC.
- ***Postcode***
    - This observation was made while completing Task 3:
        - Postcodes are preceded by a 0 which, when trying to display on a density map, makes these postcodes exist in the US instead of AU.
        - Data cleaning task - remove the preceding 0s.
    - Scratch that, the issues was Tableau imposed as it thought the postcodes, by themselves, were located in the US. Mitigation: include country as a Detail Mark.

Task 2 - Targeting High Value New Customers

Link to Presentation (in progress)

Objective

Reveal useful insights, in presentation format, aimed at helping the client optimize resource allocation for targeted marketing by focusing on identifying potential new high value customers. Of the list of 1000 new customers, determine which customers are likely to drive more value for Sprocket and thus, target marketing towards.

The presentation is what will be used to receive approval from the client to proceed with the 3 week scope.

Present an outline of the approach for completing the analysis activities:

- Understand the data distributions (ex: age) - Data Exploration
- Feature engineering - Data Exploration
- Data transformations - Data Exploration
- Modeling - Model Development
- Results interpretation - Interpretation
- Reporting - Interpretation

Data Exploration

Completing data exploration in Excel:

1. Create an Age column
    a. First removed the rows where there was a blank in the DOB column (87)
    b. =DATEDIF([@DOB], TODAY(), "Y")
2. Defined Age Ranges based on marketing standards
3. Created a histogram of the client's customers age distribution
4. Created an Age Range column to classify the customers
5. Pivoted the data and generated insights for…
    a. Customer Wealth
    b. Customer Wealth based on Age Range
    c. Average Tenure
    d. Gender breakdown based on Age Range
        i. Auto transformed the data to be 1 of 3 options (M, F, U)

Installed Parallels to Run a Virtual Windows Machine on my Mac and Work with Power Pivot.

1. Loaded the tables, excluding NewCustomerList, into Power Pivot

2. Defined relationships using Customer ID
3. Created measures for total profit, unique customers that made a purchase, average profit for the customers that made a purchase

Went back to the workbook with the tables.

1. Created a Pivot Cache based on the Power Pivot data source.
    a. This allowed me to pivot across tables and generate deeper insights about both the client's customers and their product preferences.
    b. See sheet 'PowerPivotTable' to view the staging ground of the visuals creation.

Model Development

RFM Analysis

A way of identifying high-value customers based on purchase recency, frequency of purchases and overall monetary value of each customer.

Correlation Matrix

Use to help support the identification of valuable customer attributes.

Linear Regression

Examining the relationship between Profit and Bike Related Purchases in the Past 3 Years:

What models are more valuable for this type of data?

Interpretation

Link to Presentation (in progress)

How is the data to be understood and measured?

Taking Task 2 Further

Customer Insights

- *Data Exploration*
    - Find the most valuable zip codes in each AU state.
    - Determine distance between customers' home and work offices to gain insights into understanding who is more likely to bike.
- *Model Development*

        - Correlation Matrix
        - Multiple Regression between profit and key variables to determine statistical significance

Product Insights

- *Data Exploration*
    - Explore the data from the perspective of which products you should prioritize marketing & selling
        - Product Recommendations
        - Most profitable product lines & brands
- *Model Development*

What else I would like to do…

- Create a more comprehensive profile of current & prospective customers.
    - Research/merge population-based data (from the Australian Bureau of Statistics). Things to understand about the people and areas of value:
        - Age, Gender, Property Value, Count in each location
    - Get more granular through using “postcode” rather than state
- Perform analysis from the perspective of revealing products to sell:
    - What sells?
        - Cross-sell/up-sell strategies & opportunities
        - Develop product recommendations
    - Products with the best margins
        - Relate it to customer purchasing behavior
- Identify what a data science team could help with. Ideas:
    - Statistical significance of attributes
    - Enhanced modeling (using python for statistical approaches)
        - Regression
        - Clustering / KNN
        - Classification (likelihood of customer-lifetime value)

Task 3 - Dashboarding

The client is happy with the analysis plan and would like us to proceed. After building the model we need to present our results back to the client. Visualisations such as interactive dashboards often help us highlight key findings and convey our ideas in a more succinct manner. A list of customers or algorithm won’t cut it with the client, we need to support our results with the use of visualisations.