☑️

Step1: Data quality assessment and cleaning

Tools: Excel(Google Sheet, copied excel file) [google sheets 运行太慢了]

Tasks:

Sprocket Central Pty Ltd , a medium size bikes & cycling accessories organisation, has approached Tony Smith (Partner) in KPMG’s Lighthouse & Innovation Team. Sprocket Central Pty Ltd is keen to learn more about KPMG’s expertise in its Analytics, Information & Modelling team.

Smith discusses KPMG’s expertise in this space (you can read more [**here**](https://home.kpmg/au/en/home/services/advisory/management-consulting/digital/data-analytics-modelling.html)). In particular, he speaks about how the team can effectively analyse the datasets to help Sprocket Central Pty Ltd grow its business.

Primarily, 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.

However, in order to support the analysis, you speak to the Associate Director for some ideas and she advised that “*the importance of optimising the quality of customer datasets cannot be underestimated. The better the quality of the dataset, the better chance you will be able to use it drive company growth*.”

The client provided KPMG with 3 datasets:

* Customer Demographic
* Customer Addresses
* Transactions data in the past 3 months

You decide to start the preliminary data exploration and identify ways to improve the quality of Sprocket Central Pty Ltd’s data.

[Voicemail transcript below]

“*Hi there – Welcome again to the team! The client has asked our team to assess the quality of their data; as well as make recommendations on ways to clean the underlying data and mitigate these issues. Can you please take a look at the datasets we’ve received and draft an email to them identifying the data quality issues and how this may impact our analysis going forward?*

*I will send through an example of a typical data quality framework that can be used as a guide. Remember to consider the join keys between the tables too. Thanks again for your help*.”

[Read email below]

**Draft an email** to the client identifying the data quality issues and strategies to mitigate these issues. Refer to ‘Data Quality Framework Table’ and resources below for criteria and dimensions which you should consider.

Instructions:

1. review the excel and delete unuseful information
   1. training datasets: Transactions, CustomerDemographic(customer\_id) and CustomerAddress
   2. NewCustomerList: find the high-value customers (Problem)
   3. delete the guide sheet
   4. delete the first row for Transactions, CustomerDemographic and CustomerAddress
2. Transactions
   1. check completeness: delete rows containing null values(in python, delete standard\_cost.isnull())
   2. delete customer\_id not in CustomerDemographic (in python, merge three datasets, the base isCustomerDemographic customer\_id )
   3. delete order\_status = cancelled
   4. after [delete order\_status = cancelled], 496 customer in CustomerDemographic, but not in transactions
   5. list\_price（标价） and standard\_cost（成本）: into number
   6. product\_first\_sold\_date: into date
   7. add two columns: recency and profit
      1. recency: max(transaction\_date)-transaction\_date
      2. profit: list\_price - standard\_cost
   8. unuseful columns: order\_online, product\_first\_sold\_date,[brand, product\_line, product\_class, product\_size]
3. CustomerDemographic
   1. check completeness: delete rows containing null values
   2. gender: changed into Female, Male, U(Unknown/Unspecified)
   3. add age columns
      1. change DOB into date format
      2. now-DOB
      3. delete wrong rows
   4. delete default
   5. delete values: deceased\_indicator = Yes (people who were dead)
   6. tenure: the months of customers subcribed?
   7. unuseful columns: job\_title, DOB, first\_name, last\_name
4. CustomerAddress
   1. check completeness: no null values
   2. there are 3 customer in CustomerDemographic, but not in CustomerAddress; 3 in CustomerAddress, but not in CustomerDemographic
   3. state: NSW, VIC, QLD (changed the expression to be consistent)
   4. postcode: add the sa3\_name (in Python)
   5. unuseful columns: address, country

Python(see customer.py)

Step2: Data Insights (exploratory data analysis)

Tools: Python, Jupyer Notebook

Instructions: see Jupyer Notebook

Task and tips:

Sprocket Central Pty Ltd has given us a new list of 1000 potential customers with their demographics and attributes. However, these customers do not have prior transaction history with the organisation.

The marketing team at Sprocket Central Pty Ltd is sure that, if correctly analysed, the data would reveal useful customer insights which could help optimise resource allocation for targeted marketing. Hence, improve performance by focusing on high value customers.

For context, Sprocket Central Pty Ltd is a long-standing KPMG client whom specialises in high-quality bikes and accessible cycling accessories to riders. Their marketing team is looking to boost business by analysing their existing customer dataset to determine customer trends and behaviour.

Using the existing 3 datasets (Customer demographic, customer address and transactions) as a labelled dataset, please recommend which of these 1000 new customers should be targeted to drive the most value for the organisation.

In building this recommendation, we need to start with a PowerPoint presentation which outlines the approach which we will be taking. The client has agreed on a 3 week scope with the following 3 phases as follows - Data Exploration; Model Development and Interpretation.

Prepare a detailed approach for completing the analysis including activities – i.e. understanding the data distributions, feature engineering, data transformations, modelling, results interpretation and reporting. This detailed plan needs to be presented to the client to get a sign-off. Please advise what steps you would take.

Please ensure your PowerPoint presentation includes a detailed approach for our strategy behind each of the 3 phases including activities involved in each - i.e. understanding the data distributions, feature engineering, data transformations, modelling, results interpretation and reporting. This detailed plan needs to be presented to the client to get a sign-off.

--

Tips: Raw data fields may be transformed into other calculated fields for modelling purposes (i.e. converting D.O.B to age or age groups).  Tips: You may source external data from the ABS / Census to add additional variables that may help support your model. [like postcode with area]

Step3: Data Modelling (RMF and clustering)

Tools: Python, Jupyer Notebook

Instructions: see Jupyer Notebook

Task:

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.

Please develop a dashboard that we can present to the client at our next meeting. Display your data summary and results of the analysis in a dashboard (see tools/references for assistance). Maximum of 3 dashboard views/tabs, creativity in layout and presentation is welcome.

As this is not a KPMG branded deliverable, please find attached a client logo for incorporation.

It is important to keep in mind the business context when presenting your findings:

* What are the trends in the underlying data?
* Which customer segment has the highest customer value?
* What do you propose should be Sprocket Central Pty Ltd ’s marketing and growth strategy?
* What additional external datasets may be useful to obtain greater insights into customer preferences and propensity to purchase the products?

Specifically, your presentation should specify who Sprocket Central Pty Ltd’s marketing team should be targeting out of the new 1000 customer list as well as the broader market segment to reach out to.

Step4: Presentation

Tools: Python, Jupyer Notebook or Tableau Dashboard