Travel Tide - Executive Summary

Introduction

Elena Tarrant has given the task of using demographic information to find those who would most prefer the following perks:

- Free Checked Bag
- Exclusive Discounts
- No Cancellation Fees
- Free Hotel Meal
- 1 Night Free Hotel With Flight

The overarching goal is to increase return customers using the above segments for personalized advertising of services.

Objectives

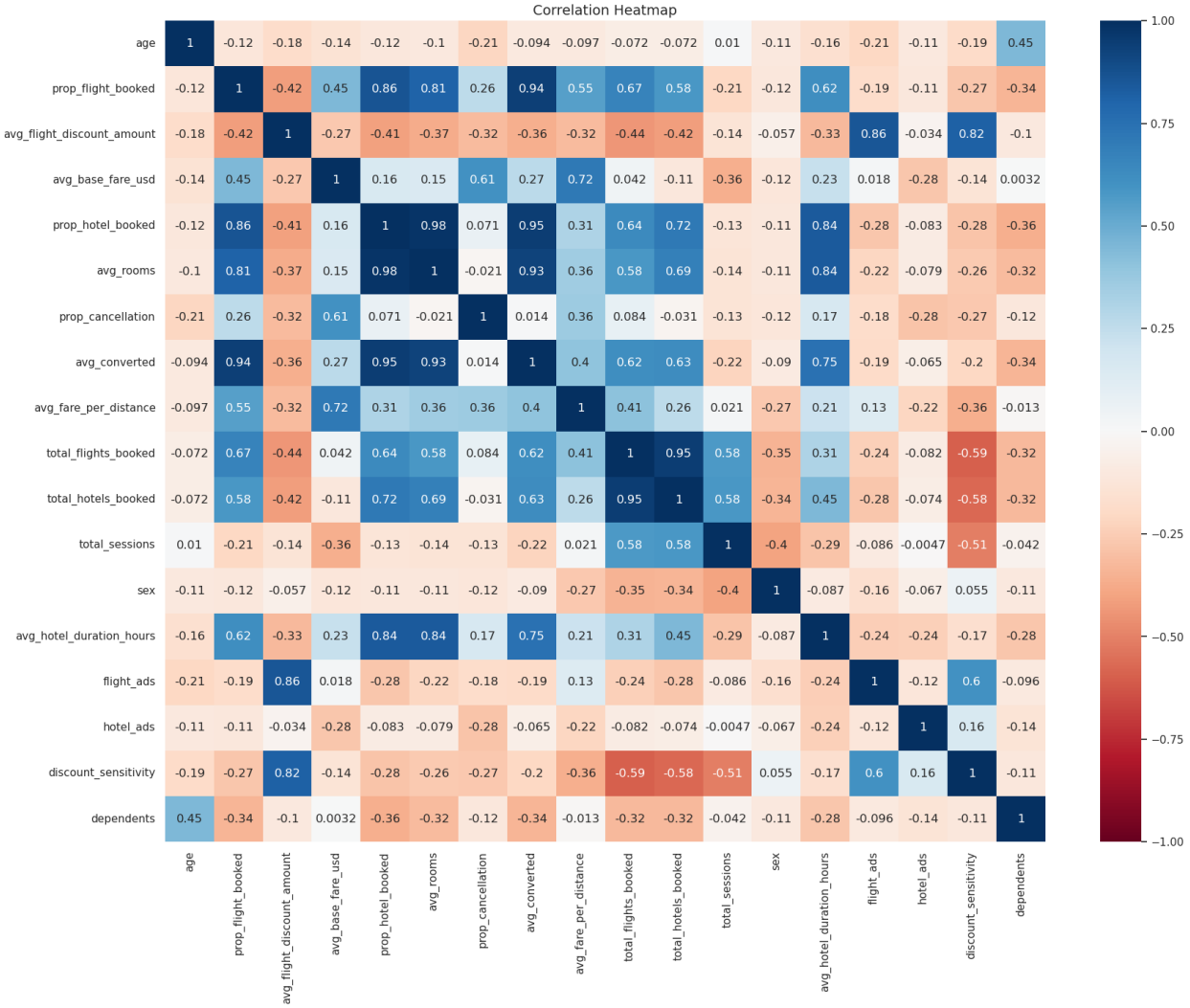
- Find metrics that are best suited to determining the segment a user belongs to.
- Attribute those metrics to each user.
- Share the results and identify any problems or add any recommendations on changes of segments.

Brandon Marcure
2023 - 08 - 30

Methodology

The process used for completing the task was using jupyter-notebook which is a python based notebook editor, mostly used in the case of analytics and machine learning. The main modules used were pandas and seaborn. A few modules were used throughout the completion of the project and, even though not all were used in the final notebook, they were used throughout the testing phases. SQL queries were placed in the queries.py file as they were also used in the notebook to pull the data into a pandas DataFrame. The notebook begins with all the imports followed by settings, then basic setup, metrics, segmentation, and lastly, a k-means algorithm.

Correlations were mostly found using the following seaborn heatmap:



Correlations of 1 are simply comparing the same metrics, which is the cause of ones going down diagonally from left to right.
Blue is + correlation and Red is - correlation.

Certain Metrics were more useless than one would normally consider them to be in the context of travel i.e sex, dollars saved metrics, and marriage.
This heatmap was used when considering the best suited metrics for each perk.

- Ranking of values was a weighted using max-min averaging from 1 to 5 as there are 5 perks, including decimals, that would be used to determine which segment was ranked highest for each user.

- After the segmentation, a csv is created containing the user_id, segment values, their highest ranking value and then the name of the category that maximum value correlates to.

- Exclusive Discounts was associated with discount sensitivity as that was what was discussed with Elena Tarrant.

- No Cancellation Fees was associated with the proportion of cancellation as those are the people who would most likely care about cancellation fees.

- Free Checked Bag was associated with the proportion of flights booked as that would be the focus of those who travel a lot but not necessarily booking hotels a whole lot.

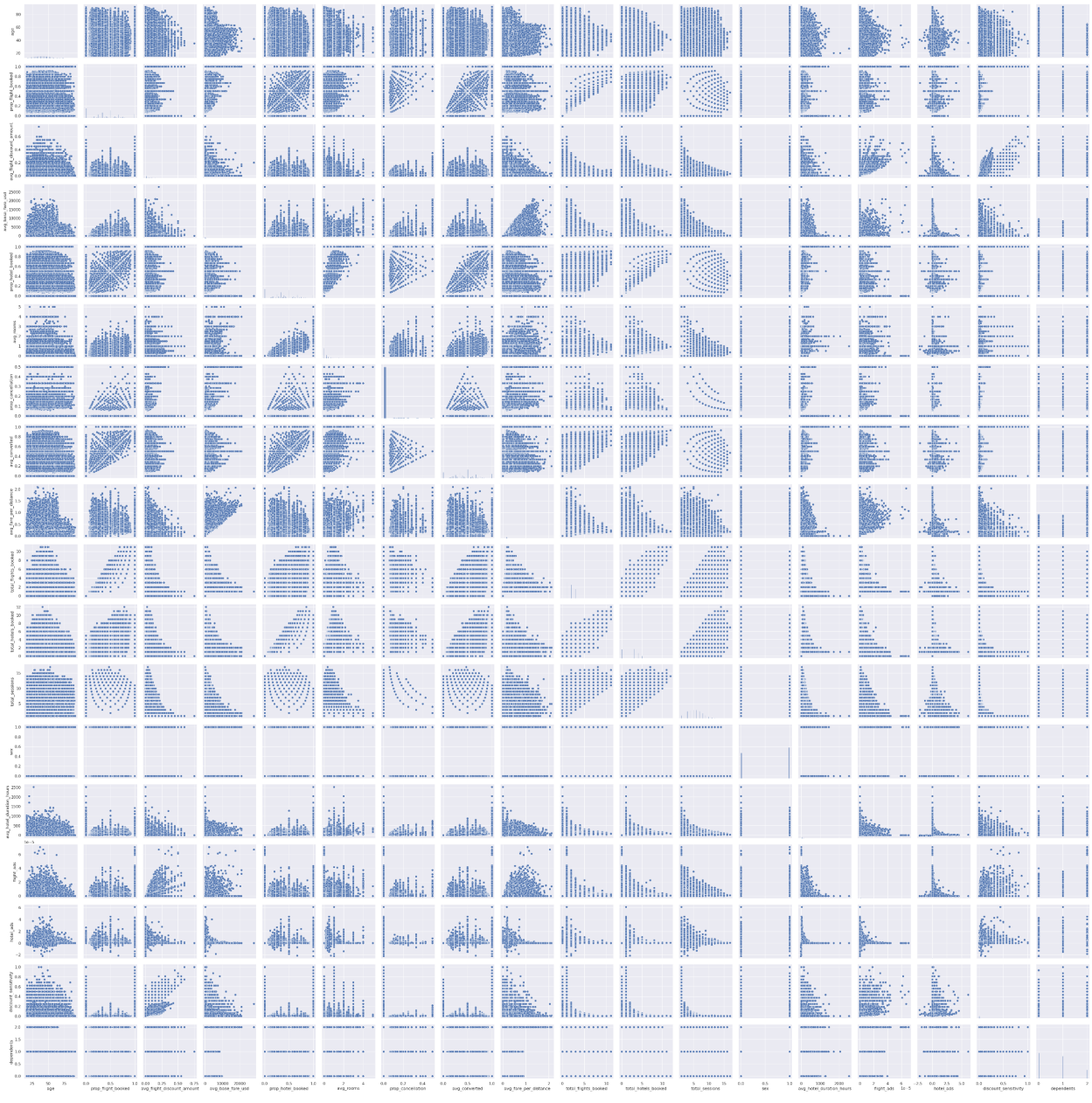
- Free Hotel Meal was associated with the average hotel duration in hours as those who would most benefit from meals would be those who booked the most hotels

- 1 Night Free Hotel With Flight was associated with “proportion of hotels booked” as those who are most likely to want a free night would be those who are staying in many places and could benefit in more areas as opposed to those who would like Free Hotel Meals due to saving money from longer hotel stays.

Conclusions

- The final conclusion, using a Fuzzy Segmentation Process shows no good reason for why the segments given can’t be acceptable.
- It was noticed that the data did give indications that some categories were more correlated with others than others, however it didn’t seem likely that it was a real cause of concern as the categories to be fitted would entail some overlap but are inherently different and more specific to certain behaviors.

Further Viewing



The Seaborn pair plot above, a plot of plots - given the data passed in are integers, initially used in investigating the data and the pattern of metrics created. It takes a long time to generate (3 mins) so it was placed here for time saving concerns. It gives a good initial investigation for figuring out what should be done.

Other plots can be viewed from the notebook directly as they only take <= 1 second to process.

Notes

- All python based information is stored in the zip file within the python directory.
- Any other data is within the root directory TravelTide.
- Queries were used in the Jupyter Notebook (python) therefore are contained within the queries.py file.
- Distance functions are contained in the self named python file. It contains both a Haversine function as well as a Vincenty function, which was not used as it was not a requirement for this to be completed. It is, however, more accurate as it factors in the differences in width and height of the earth. Haversine is more accurate than the provided SQL function, which is only accurate for short distances.
- Not all functions contained within the .py modules are still used. They were mostly made when organizing the jupyter notebook and were left in case they may be found relevant again. Ex: I have quite a few modules devoted to downloading the SQL database and then transferring them to my own personal database to allow for faster response times. The results could take quite a while to complete so I automated the transferring of the tables into my own psql database - initially.
- If you want to save CSV files or compile the plots then check the “Notebook Specific Settings” in Blue to mark anything you want with a True bool in order to see the graphs display or to save the CSV files.
- It should only take two to three minutes for the whole notebook to finish processing if the charts’ variables are set to False (Depending On Hardware).
- The Python Version Used In The Whole Project Is: 3.11.3
- Video Presentation Link