Coffee offers: Advanced customer segmentation with Python

MADIHA DANISH

Customer Segmentation

Customer Segmentation can be a powerful means to identify unsatisfied customer needs.

The most common ways in which businesses segment their customer base are:

- **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
- **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence.
- **Psychographics**, such as social class, lifestyle, and personality traits.
- **Behavioral data**, such as spending and consumption habits, product/service usage, and desired benefits.

The Challenge

You are owing a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. You want to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

K Means Clustering Algorithm

- Specify number of clusters *K*.
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

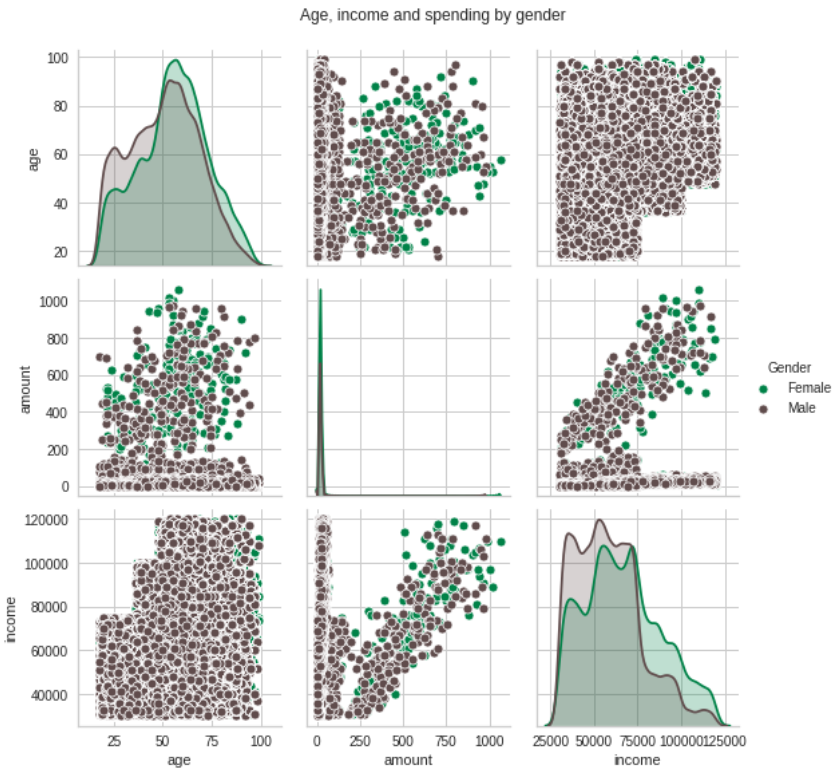
Data wrangling

Before I could visualize and model the data, I've had to do some preprocessing both outside, and in Python.

- **Removed empty lines** in transcript.json using search \n\n & replace \n in Visual Studio Code
- **Imputed** empty income values with the mean ($65,404), and added a separate feature that tracks missing income values with 1s and 0s.
- **Engineered** a new feature for the year when the user became a member
- **One-hot-encoded** channels using the *MultiLabelBinarizer*
- **One-hot-encoded** offer types, genders, years joined and event types using *get_dummies*
- **Dropped age outliers** (a number of outlier customers had their age set to 118, and were missing data for several of the other fields)
- **Engineered** first receipt, first view and first completion time features (a customer can receive and interact with the same offer multiple times)
- **Dropped misattributions** (completion without view, completion before view, or view before receipt)
- **Calculated RFM** *Recency* and *Frequency* scores— a common method used for analyzing customer value in retail and e-commerce
- **Engineered** view and conversion rate features for each offer type
- **Merged all data** into one dataframe grouped by customers, including means and sums for all available data, as well as additional columns for the average number of exposures per offer-type

Data exploration

- The mean age across all customer groups, is 53 years. Male customers in the dataset tend to be younger than this average.
- Incomes range from $30,000 to $120,000, with a mean of $61,800. Female customers tend to have higher incomes than male customers, likely correlated with their higher average age.
- The average transaction value is just $14, but there are long tail transactions of up to $1062.



Conclusion

- I’ve identified four segments showcasing distinct purchasing habits and reactions to marketing offers.
- Most importantly, we’ve identified an entire segment of subpar targets that we can exclude in our paid marketing campaigns to optimize our Customer Acquisition Cost.
- Many challenges in working with this dataset resulted from repeat exposures to the same offer over different channels, and imperfect conversion attribution. In future experiments, it would be desirable to generate more accurate data on the source of each conversion, and confirm completion through coupon codes