- # Summary of my method

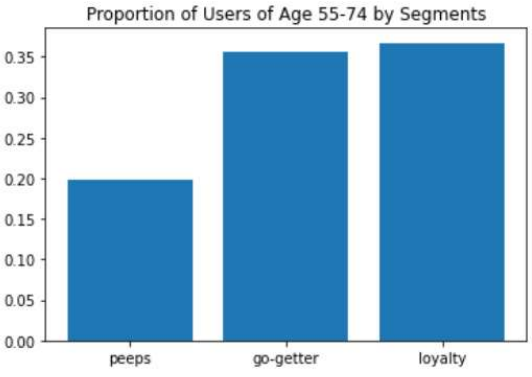
1. **Data cleaning & preprocess**:
   - initial scan on the dataset and handled **missing values** by removing the rows that have <10% null values considering the nature of survey data (not sure if it's missing because of respondents refuse to answer or the answer is just null or respondents did not complete the survey)
   - converted **categorical features that are ordinal** by nature to ordinal values
   - experimented with the two **set features** (primary_language_motivation_followup and other resources) with **PCA** to extract principal components
   - some visualization on the numerical values to gain insights on their distribution
   - applied **scaler** to numerical values
2. **Clustering Analysis**:
   - based on the nature of the dataset, which is a mix of numerical and categorical values, I've chosen the **KPrototypes** method in sklearn.
   - choose best # of clusters using the **elbow method**
   - assigning labels to each data point after the analysis
3. **Result Interpretation:**
   - look at the **centroids** of the three clusters
   - look at **summary statistics and distribution** of each feature to discover pattern

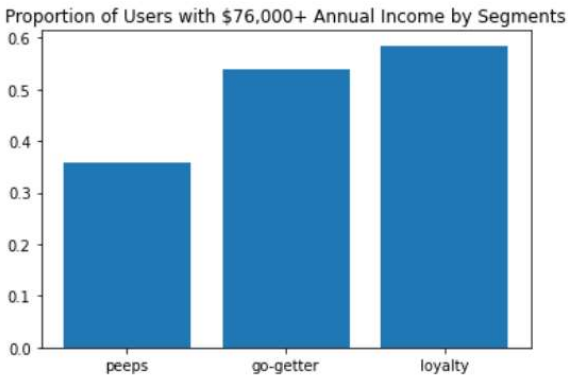
- # Proposed segments

The following table shows our current user breakdown and main characteristics of each segments with regards to three aspects.

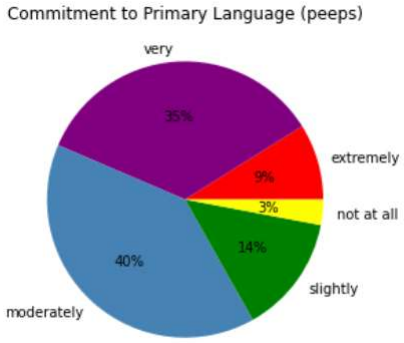
	***peeps* (64.2%)**	***go-getter* (30%)**	***loyalty* (6.8%)**
Demographics	younger, lower income, some students	middle age, middle to high income	older age, middle to high income
Commitment	mostly moderate, mostly never pay, low goal	very to extremely, mostly pay, high goal	moderate to very, mostly pay, medium goal
Progress	few days on platform, low number of streak, low course progress	some days on platform, some number of streak, some course progress	many days on platform, high number of streak, high course progress

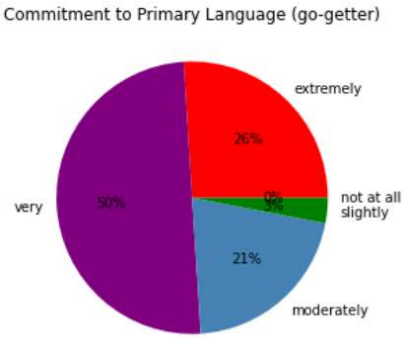
Demographics

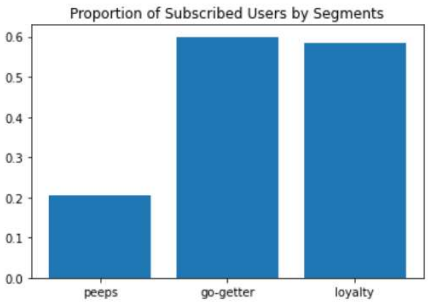


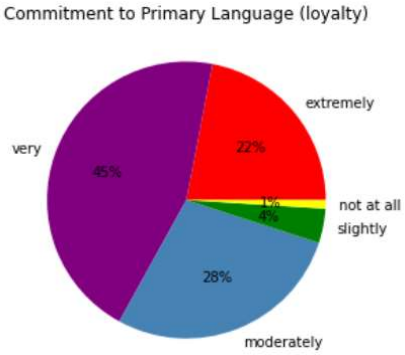


Comitment

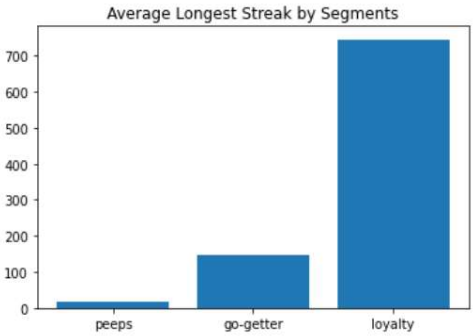


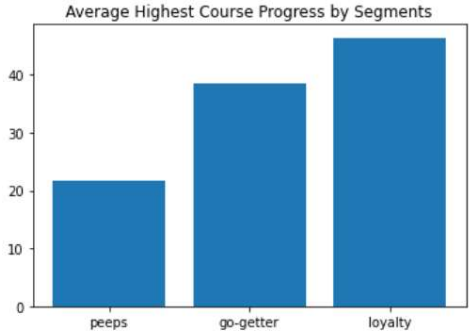






Progress





- # Product recommendation

Based on the characteristics of each segment and the breakdown of our current users, I give the following recommendation:

- For the *peeps*: this is the largest group in our users.
  - get more peeps to start paying: this is to generate profits, some common techniques are free trials, cheaper basic plans, etc.
  - increase retention and activity rate: after all this makes up the majority of our user base. Could try introducing fun mechanisms (which Duolingo is already doing a good job at) such as games, ranking, etc.
- For the *go-getter*: this is the second largest group and most committed
  - improve their learning experience with Plus: these users are highly committed to learning and tend to pay for Plus, we should help them learn better and try to hold onto them (eventually transform to the *loyalty*).
- For the *loyalty*: this is the smallest group and have high retention
  - reward them for being loyal: these are probably the earliest ones to start using Duolingo and are still using. Should reward them with free Plus experience or exclusive perks.