DATA SCIENCE FINAL PROJECT

Max Comer

The project is based on the yelp academic challenge dataset

- **My goal is to create customer segments for restaurant goers based on their preferences**
- This is an unsupervised classification problem, I use Kmeans
- The raw data I need is in json format in two datasets:

User/Review Level Data


```
{
    "votes": {
        "funny": 0,
        "useful": 0,
        "cool": 0
    },
    "user_id": "4CgusCZkipvUhvBZrRD46w",
    "review_id": "aT-ogKlfaUb42QDn42pn_w",
    "stars": 5,
    "date": "2014-01-26",
    "text": "We have been using Casey for about two yea
    Amazing job!  Thanks Casey.  \nThe Gelbmans",
    "type": "review",
    "business_id": "gXkQQ6-XpATxAk7kOp7CjA"
}
```


Business Level Data


```
{
    "business_id": "stH6XAn2Drzol1H5oGvL2A",
    "full_address": "23623 N Scottsdale Rd\nScottsdale, AZ 85255",
    "open": false,
    "categories": [
        "Chinese",
        "Restaurants"
    ],
    "city": "Scottsdale",
    "review_count": 69,
    "name": "Jade Palace",
    "neighborhoods": [],
    "longitude": -111.92528,
    "state": "AZ",
    "stars": 4,
    "latitude": 33.7019607,
    "attributes": {
        "Take-out": true,
        "Wi-Fi": "no",
        "Good For": {
            "dessert": false,
            "latenight": false,
            "lunch": true,
            "dinner": true,
            "brunch": false,
            "breakfast": false
        },
```

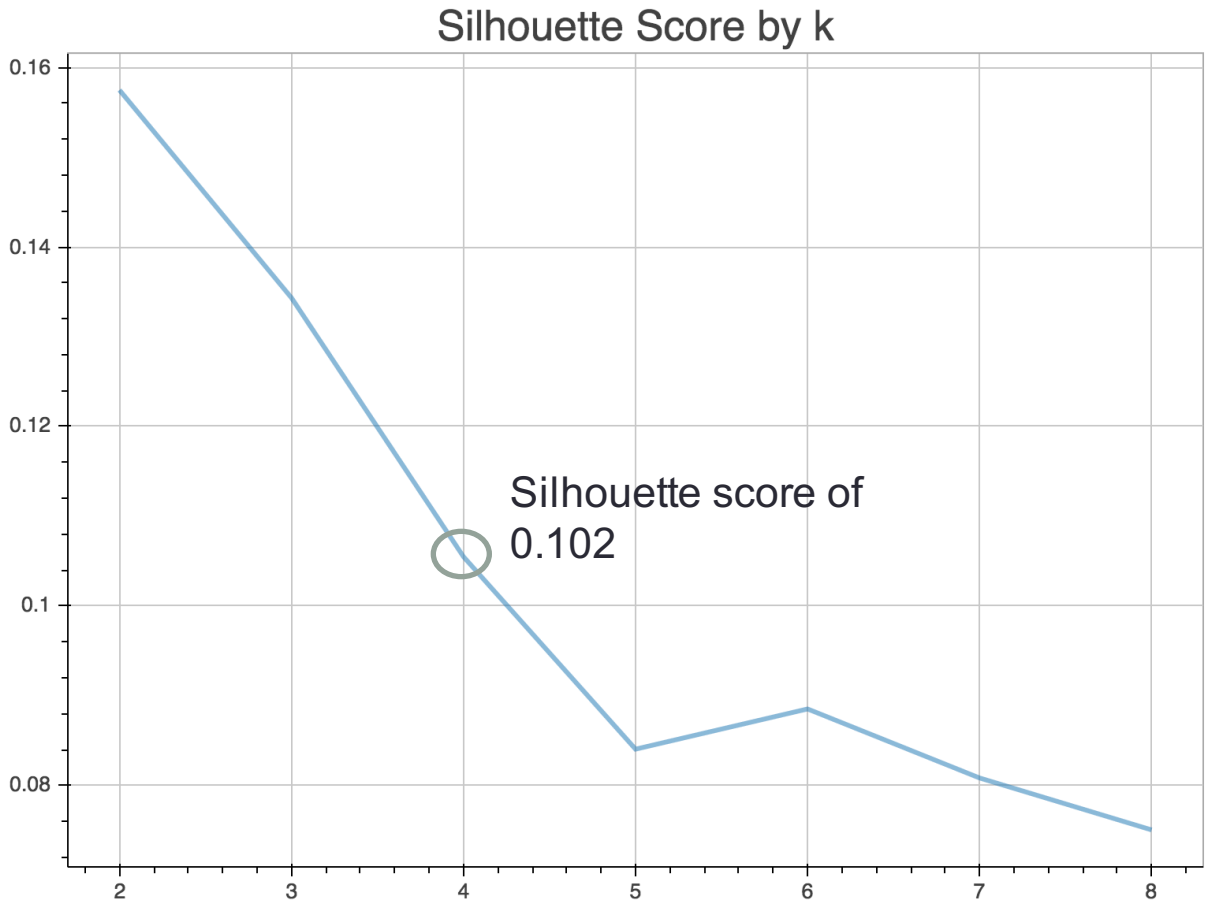

http://www.yelp.com/dataset_challenge

Next step is to join these together, and get a user level dataset with their preferred restaurant characteristics

- I keep only reviews with 4 or 5 stars since I want restaurants users like
- I also limit the dataset to users with at least 10 of these reviews, to make sure they have a reasonable sampling of their interests
- From the business dataset, I pull the business categories, attributes, and ambience (over 300 features)
- I then aggregate at a user level for the reviews the user liked and take the average of the value for each business characteristic (most attributes are 0/1, others are normalized)

	user_id	alcohol	amb_casual	amb_classy	amb_divey	amb_hipster	amb_intimate	amb_romantic	amb_touristy	amb_trend
0	-65q1FpAL_UQtVZ2PTGew	0.265306	0.765306	0.020408	0.091837	0.020408	0.020408	0.010204	0	0.020408
1	VxRvXk3b8FwsSbC2Zpxw	0.400000	0.733333	0.066667	0.066667	0.000000	0.066667	0.000000	0	0.066667
2	-0itF0VWVBe3k2AdfUReGA	0.294872	0.461538	0.038462	0.076923	0.076923	0.076923	0.076923	0	0.153846

On cleaned dataset, I run Kmeans for different k values to choose how many clusters I want to create



- While k=2 has the highest silhouette score, but I need more clusters for this to be an interesting business result, so I choose 4

Kmeans creates these 4 clusters – the next step is to describe their characteristics as segments

- The cluster model output gives me 4 lists of centers containing an average value for each feature for that center
- I compare this with the overall average for each feature in the dataset to see how much each center deviates on each feature
- Since the features are all normalized to a [0,1] interval I can compare the deviations to understand which features in the center deviate the most
- Taking the top 15 features for each center, sorted by the absolute value of their deviations, gives me a the top distinguishing characteristics of each center
- Based on that I can get an idea of what users in that cluster like and don't like, and create a persona that helps to understand the cluster

Segment 1: Refined Dining

Steakhouses
French waiter service
corkage trendy price classy
takes reservations
alcohol

outdoor seating
casual
take-out
has tv good for kids
caters

Segment 2: Fast, Casual, and Kid Friendly

outdoor seating
Mexican casual
good for kids
caters
take-out

American (New)
Nightlife
happy hour trendy
takes reservations
Bars
waiter service

Segment 3: Kid Friendly and Indoors

Japanese

good for kids

casual

wheelchair accessible

outdoor seating

trendy caters Mexican

Bars Nightlife

has tv happy hour

American (New)

dogs allowed

Segment 4: Nightlife and Alcohol

dogs allowed
trendy American (New)
Bars alcohol
happy hour
outdoor seating
has tv Nightlife
takes reservations
waiter service
wheelchair accessible

good for kids
casual
delivery

Next Steps to Investigate

- Outdoor seating seems to be an surprisingly strong – is this related to the different cities in the dataset?
- A simple recommendation engine from this dataset would be relatively easy to build
- Cluster 3 doesn't seem to have that strong a definiting – see how the attributes would look with only 3 clusters