Yelp Business Insights & Hybrid Restaurant Recommendation Engine

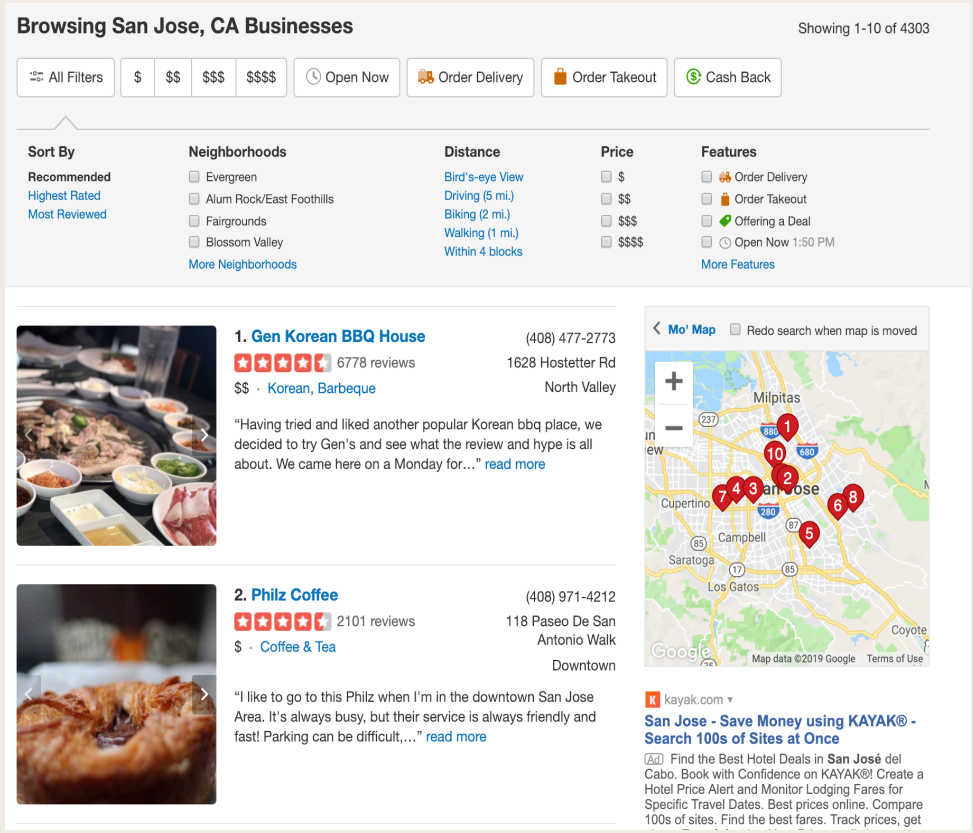
Jing Zhao
Capstone Project
Springboard Data Science
Jan, 2019

Recommender Systems

- **Recommender systems are everywhere**
  - Online shopping: Amazon, Walmart, Target, etc.
  - Entertainment: YouTube, Netflix, Pandora, Spotify, etc.
  - Online listings: Yelp, LinkedIn, Airbnb, Zillow, etc.

- **Yelp relies heavily on its recommender system**
  - Keyword filtering on rich collections of business attributions
  - Crowd wisdom: average star ratings, number of reviews
  - Generic recommendations

- **Desires for improvement**
  - personalized recommendations
  - improved ranking metrics



The Approach

- **Data acquisition & wrangling:**

  json raw data files → csv files → Pandas dataframe

- **EDA:**

  Restaurant & user patterns: location, cuisine, style, price range, etc
  Review & tip & checkin trends

- **Interactive data visualizations:**

  Interactive visualization hosted by Bokeh server based on EDA findings

- **Hybrid recommendation engine:**

  Module 1: non-personalized keyword-search recommender:
  Module 2: personalized collaborative recommender
  Module 3: personalized restaurant content-based recommender
  Integration: interactive user interface to wrap and navigate the three submodules

The Client

- **The hybrid recommendation engine is beneficial to both Yelp and Yelp users**

  Yelp:

  provide all levels of interactions to its user: both generic and personalized
  improved rankings metrics lead to higher rank effectiveness
  make use of its rich user data and restaurant attributes

  Yelp Users:

  more relevant recommendations from personalization
  flexible options and better user experience from various options

- **The methodology is transferrable and adaptive to a wide range of applications**

  Business with rich user data and product attributes, e.g.: Amazon, Target, Netflix, etc.

The Dataset

- **Source:** Yelp Dataset Challenge
- **Link:** https://www.yelp.com/dataset
- **Raw data:**
  - **five individual JSON files:**
    'business.json', 'user.json', 'review.json', 'tip.json', 'checkin.json'
  - **a total record of 5,996,996 reviews, 1,518,169 users, 188,593 businesses, 1,185,348 tips, and over 1.4 million business attributes for each of the 188,593 businesses**
  - **the total size of raw datasets is > 7 GB**

Data Wrangling

- **Raw JSON to CSV and import as Pandas dataframe**
  - Five raw JSON files are converted to five CSV files by running a python script https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/json_to_csv.py
  - Nested JSON dictionaries are also extracted during file conversion
  - Five CSV files are imported as Pandas dataframes

- **Data cleaning**
  - **Dataframe 'business':**

    Data quality check

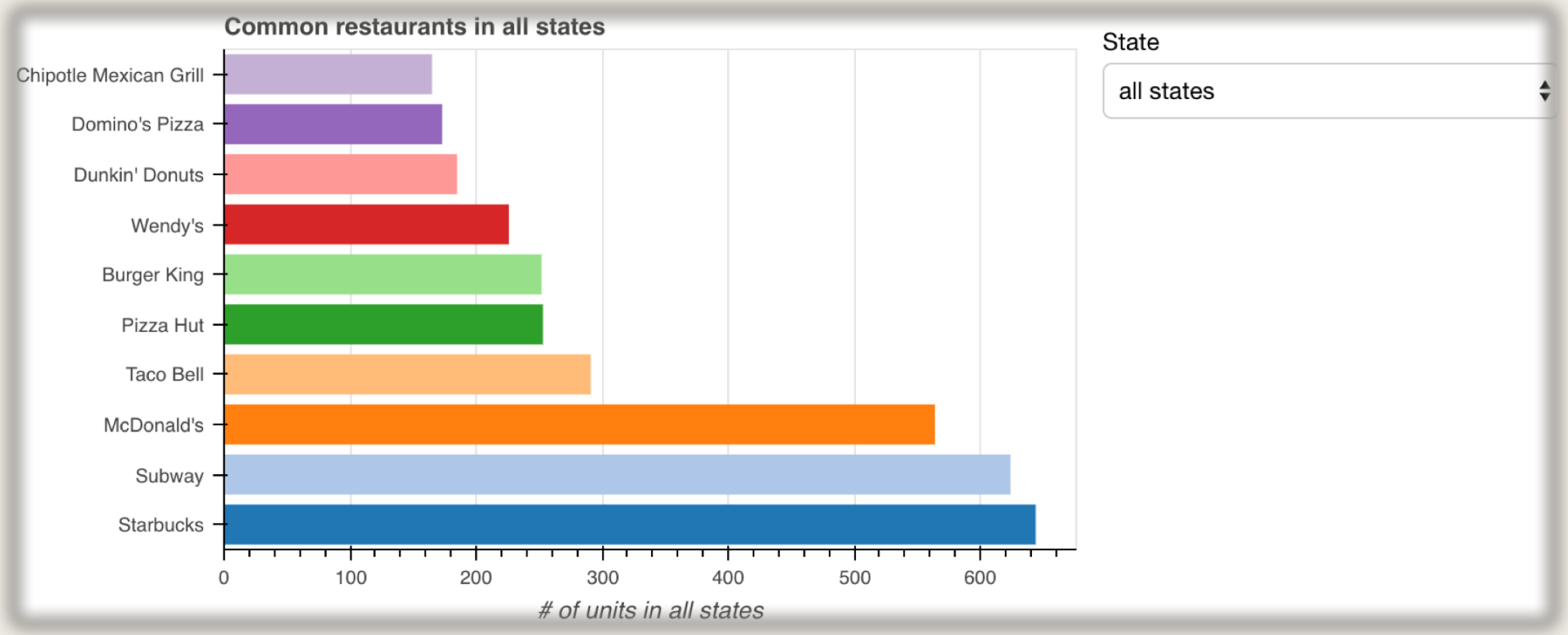
    Filter to restaurant businesses in the United States only (188,593 businesses → 47,554 US restaurants)

    Extract new columns 'cuisine' and 'style' from the 'categories' column

  - **Dataframes 'user', 'review', 'tip' & 'checkin':**

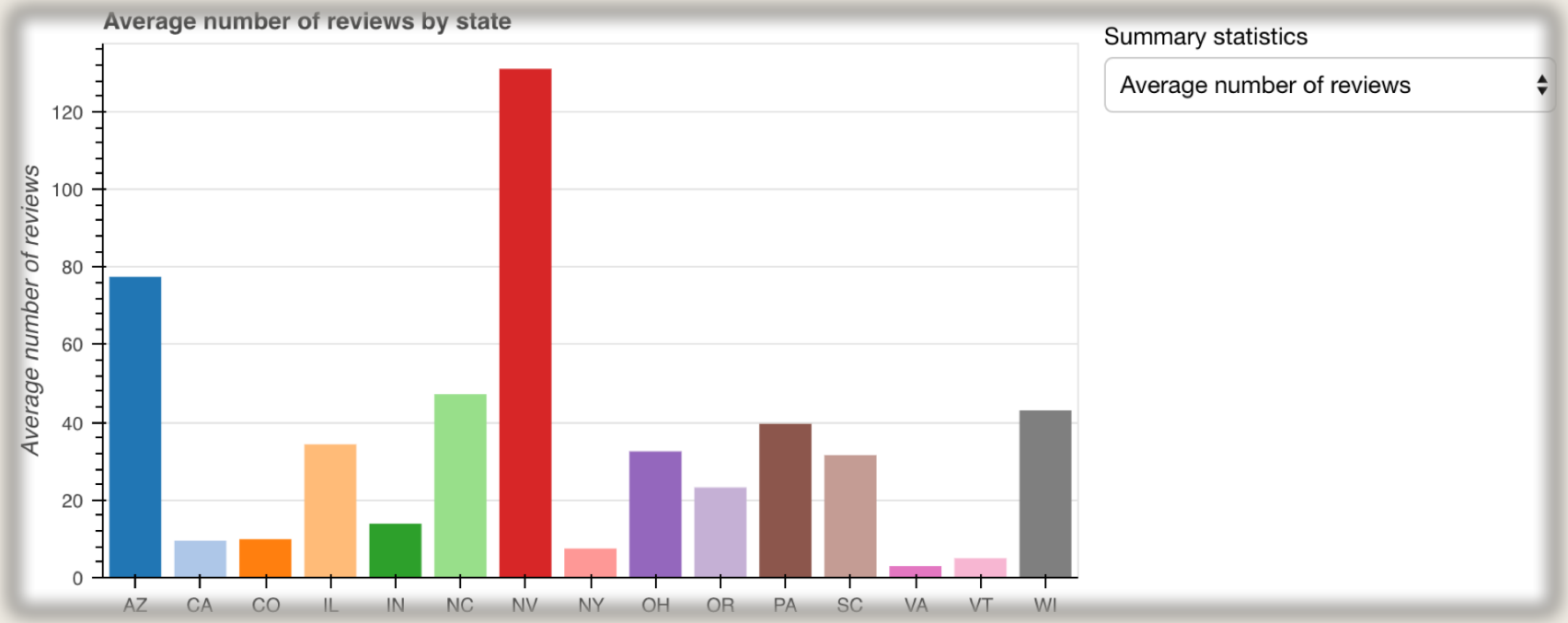
    Data quality check and fix missing values and outliers

EDA – Common restaurant names



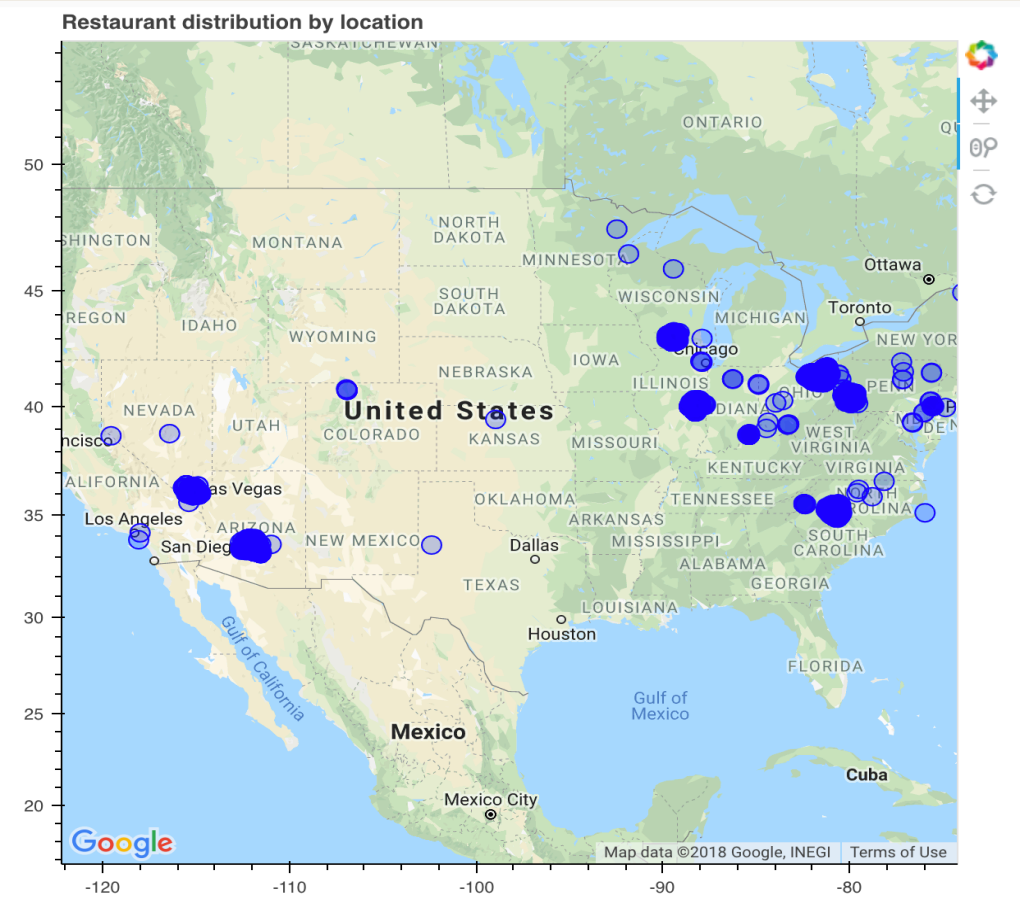
- Top 10 common restaurant names are popular chain or franchised restaurants, fast food or coffee shops
- Regional restaurant chains show up in the top list only in certain states (e,g. Filibertos ranked # 8 in AZ)

EDA - Restaurant statistics by state



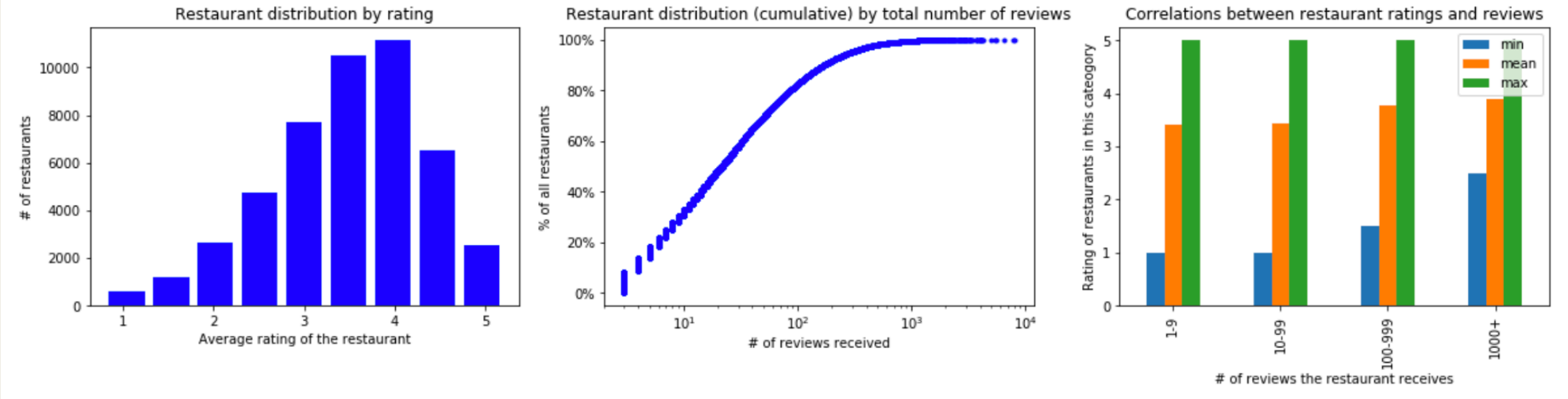
- This dataset only contains a subset of restaurants on Yelp from 15 states
- Only Arizona, Nevada, Ohio, North Carolina & Pennsylvania have a rich catalog of > 5000 restaurants in this dataset
- Nevada has a much higher restaurant review counts on average than other states, as a result of the popularity of Las Vegas as a resort town. The average restaurant rating is very similar among five states, close to 3.5.

EDA – Restaurant distribution on map



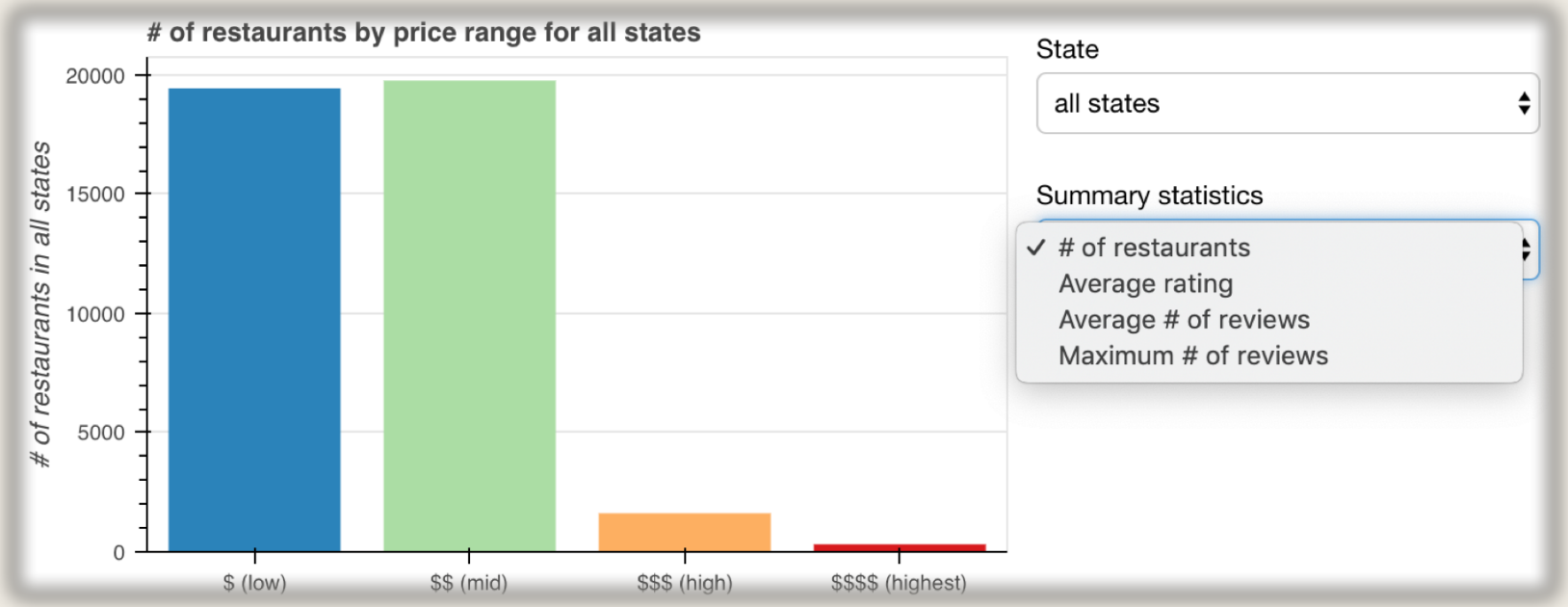
- Google map distribution also confirms that the Yelp open dataset only contains a subset of restaurants on Yelp from a few selected states.
- In particular, restaurants are densely distributed around Phoenix of Arizona, Las Vegas of Nevada, Cleverland of Ohio, Charlotte of North Carolina and Pittsburgh of Pennsylvania.

EDA - Restaurant rating vs. # of reviews



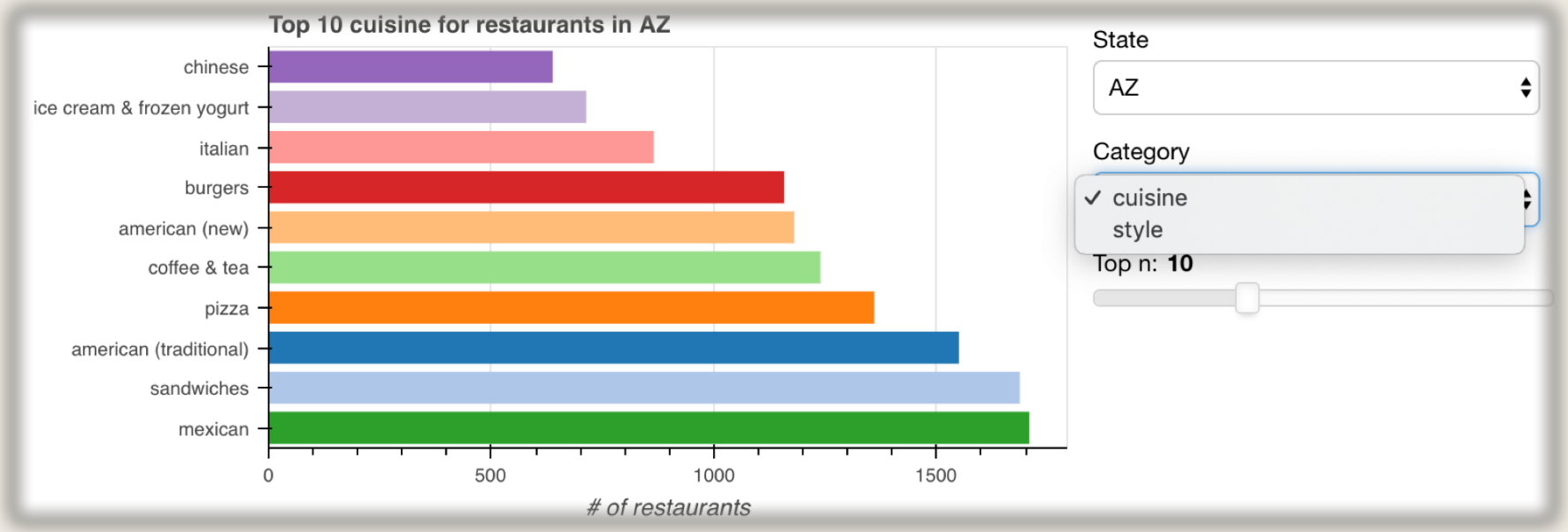
- The majority of the restaurants has ratings between 3.0 and 4.5, with 3.5 and 4.0 being the most common
- Half of the restaurants have less than 30 reviews, although the record number of reviews is as high as 7968
- Restaurant rating is related to # of reviews to some extent, as restaurants with more reviews tend to have higher ratings on average.

EDA – Restaurant by price range



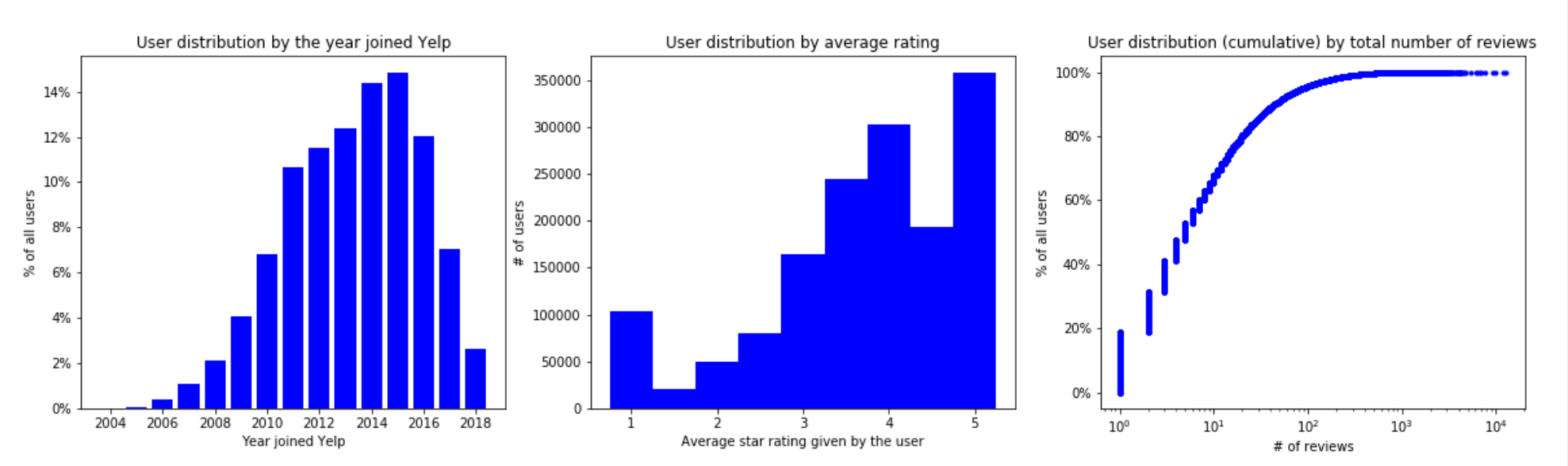
- Most restaurants are in the low (40.9%) and mid (41.6%) price range
- Restaurants in different price ranges have relatively similar average rating around 3.5
- More expensive restaurants tend to receive more reviews on average (average # of reviews)
- State-wise trends are in general in good agreement with the national trends, with only minor variations by state

EDA – Restaurants by category



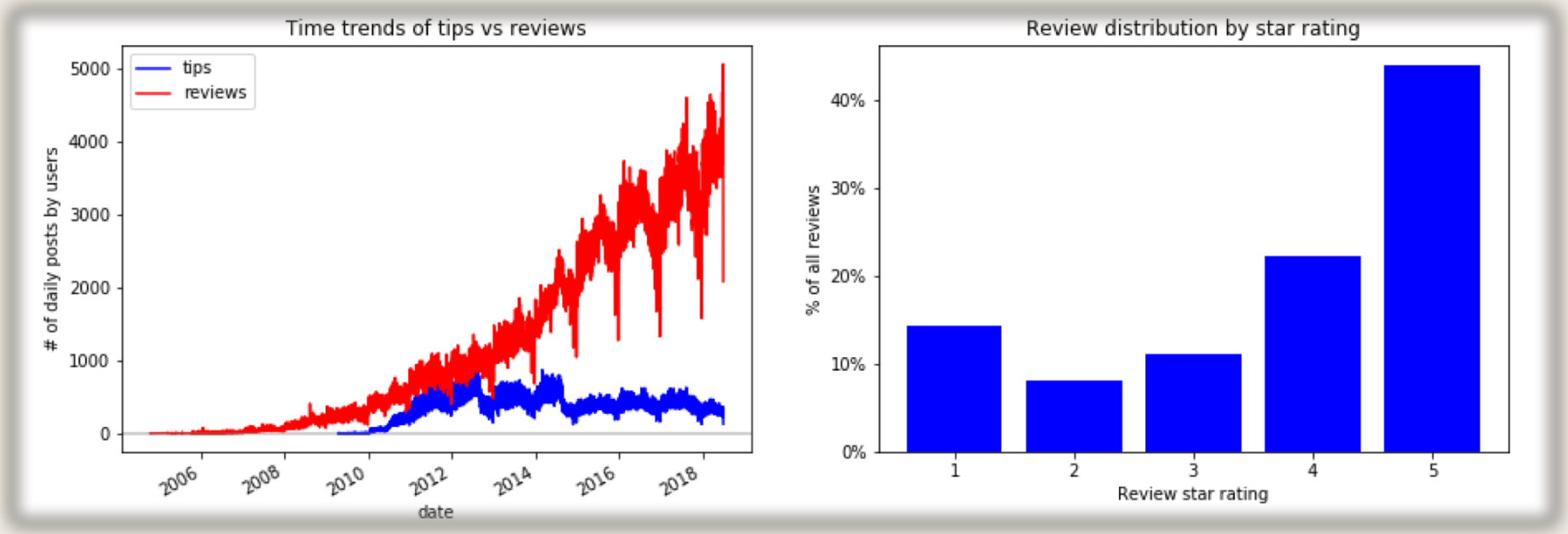
- The most popular cuisines are American (traditional & new), followed by Mexican, Italian and Chinese
- The most popular restaurant styles are the formal restaurant style, followed by nightlife/bar style and fast food
- Restaurant trend by cuisine varies quite a bit by location, suggesting people in different states favor different cuisines. The trend by style remains similar among all states.

EDA – Yelp user patterns



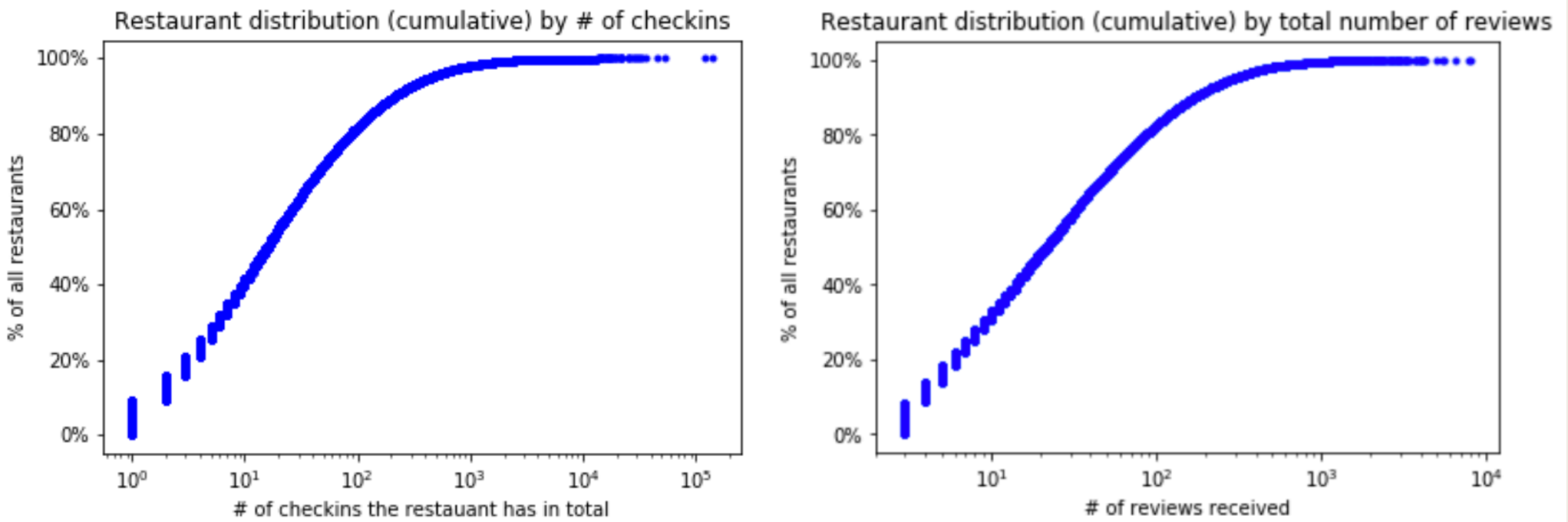
- Yelp witnesses steady increases of new members from the beginning to 2015, followed by significant declines thereafter
- Average rating given by Yelp users is 3.72; 81% of Yelp users are generous with an average rating of >3
- Although the record number of reviews of Yelp users is 12723, 60% of the users have < 10 reviews in total, suggesting that most users post reviews on Yelp only occasionally.

EDA – Yelp review & tip trends



- The popularity of reviews (# of reviews) shows a steady upward trend over time with seasonal fluctuations. Two thirds of the reviews are associated with a positive star rating of 4+.
- The popularity of tips (# of tips) increases in the first four years after introduction in 2009 and slowly dives down afterward. Overall, tip is not as popular as review.

EDA – Yelp checkin distribution



- Half of the restaurants have less than 20 check-ins
- When compared with review, check-in is a less widely used feature on Yelp platform

Recommender – Non-personalized keyword

- Restaurant location-based (zip code, city, state) keyword filtering based on **geodesic distance**
- Restaurant feature-based (cuisine, style, price range) keyword filtering
- The returned recommendations can be customized based on **ranking criteria of user's choice**:
  - Original average restaurant star rating
  - An improved rating metric based on damped mean:
    - average rating of the restaurant (quality)
    - # of ratings by users (popularity)
- Any combination of the above features

Below is a list of the top 3 recommended restaurants for you:

	distance_to_interest	state	city	name
7464	5.399251	AZ	Phoenix	Little Miss BBQ
13261	9.519138	AZ	Scottsdale	Simon's Hot Dogs
16103	1.038229	AZ	Phoenix	Tres Leches Cafe

Below is a list of the top 3 recommended restaurants for you:

	state	city	name	address
25730	AZ	Phoenix	Papa Joe's Fish-N-Que	2019 W Bethany Home Rd
13442	AZ	Surprise	Got Que?	16995 W Greenway Rd, Ste 111
13824	NC	Stallings	Rock Store Bar-B-Q	3116 Old Monroe Rd

	attributes.RestaurantsPriceRange2	cuisine	style	review_count
25730	1.0	barbeque	restaurants	222
13442	1.0	barbeque	restaurants	200
13824	1.0	barbeque	restaurants	103

$$score_i = \frac{\sum_u r_{ui} + k * \mu}{n_i + k}$$

	review_count	stars	adjusted_score
7464	1746	5.0	4.984169
31910	1380	5.0	4.980037
45401	547	5.0	4.950811
7784	520	5.0	4.948360
28162	472	5.0	4.943342

Recommender – Personalized collaborative

- **Methodology:**
  Very sparse User x Business matrix (99.997% empty)
  Matrix factorization is chosen for matrix completion
  Recommendations are generated based on the predicted rating

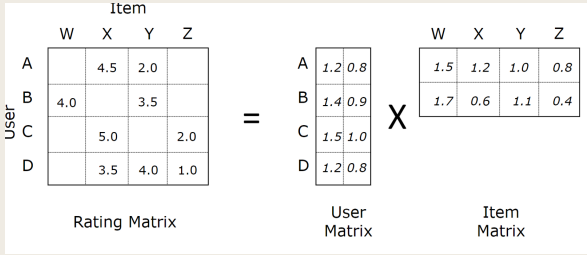
- **Matrix factorization:**
  Several algorithms are experimented
  Scikit-surprise package for handling many implementation details
  SVD with user & business bias terms gives the best performance

- **Evaluation metrics:**
  Accuracy of rating prediction: RMSE
  Recommendation ranking effectiveness: nDCG(normalized Discounted Cumulative Gain)

$$DCG(Rank) = \sum_i u_i * d_i \qquad d_i = \frac{1}{log_2(i+1)} \qquad nDCG(Rank) = \frac{DCG(Rank)}{DCG(PerfectRank)}$$

  RMSEs of the best model are 1.277 and 1.244 for testset with new user/business and testset with no new user/business, respectively
  nDCG of the best model are 0.905 and 0.908 for NDCG@10 and NDCG@5 on testset, respectively

Matrix factorization:



Rating prediction:

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Cost function:

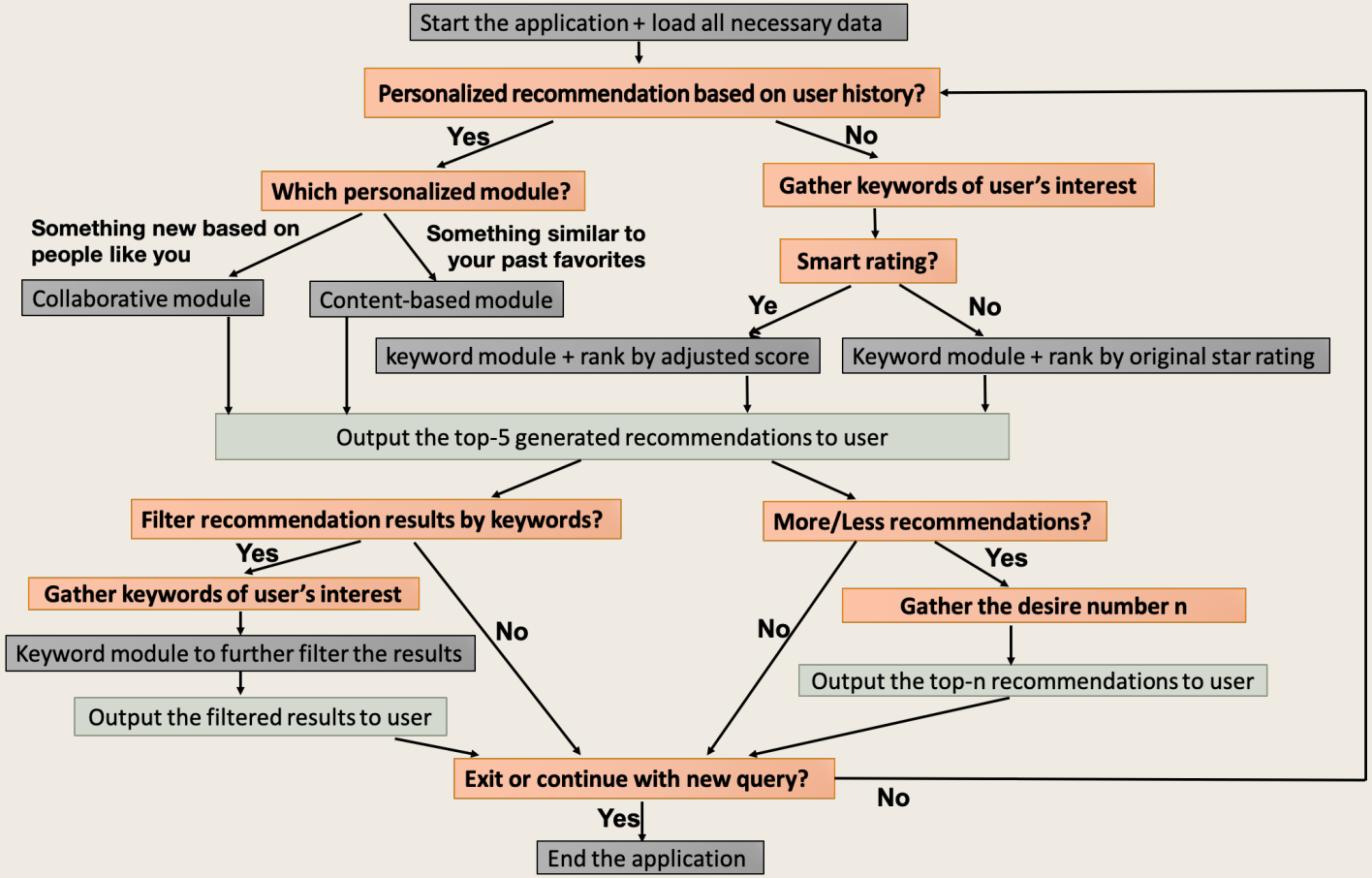
$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2\right)$$

Recommender – Personalized content-based

- **Methodology:**
  Compute restaurant vectors & user vectors from a rich set of Yelp restaurant text reviews & numerical ratings
  Personalized similarity scores are computed based on cosine similarity between user and restaurant vectors
  Recommendations are generated by ranking unrated open restaurants by descending similarity scores
- **Three strategies for computing restaurant & user vectors:**
  - Restaurant vectors space is the top 300 PCA components out of the top 1000 word features (mono & bigrams) from all restaurant reviews using Tfidf vectorizer; user feature vectors presenting user's preference are then computed by aggregating feature vectors of user-rated restaurants weighted by the corresponding user rating
  - Restaurant vector space is the top 300 word features (monogram only) extracted from all available restaurant text-based metadata by count vectorizer; user feature vectors are then computed by aggregating feature vectors of user-rated restaurants weighted by the corresponding user rating
  - The above two cosine similarity scores are used as engineered features to enable personalization; along with all other restaurant numerical metadata, a supervised regression model is built and optimized to predict user's rating of restaurants and generate recommendations
- **Evaluation metrics:**
  Recommendation ranking effectiveness: nDCG
  The best strategy achieves NDCG scores of 0.857 and 0.863 for NDCG@10 and NDCG@5, respectively.

Recommender – Integration & User interface

The user interface is created by creating a recommendation engine as a class object, implementing the three submodules as class methods and incorporating several interactive questions for gathering user interests and navigating users through the application:



Recommender - Demonstration

```
Hey, welcome to the Hybrid Yelp Recommender!
Please wait while we initiate the recommendation engine
 loading...
Yeah, recommendation engine is ready to use!
Want to try a customized recommendation based on your Yelp user history? yes/no
no
That's cool! Let's filter by keywords and generate your recommendations!
What would you like to filter by?
1 location (zipcode, city, state);
2 cuisine;
3 style;
4 price range
Please enter the corresponding numbers. Multiple filtering criteria are supported, please separate the corresponding
numbers by comma.
1,2,3,4
Great! Now let's gather your filtering criteria.
Please follow the instructions to enter your location of interest or use the ENTER/RETURN key to skip.
Please enter the zipcode of interest or use the ENTER/RETURN key to skip

Please enter the city of interest or use the ENTER/RETURN key to skip
las vegas
Please enter the state of interest or use the ENTER/RETURN key to skip

Please enter the max distance allowed between the restaurant and your location of interest or use the ENTER/RETURN ke
y to skip
10
Please select one from the following cuisines as your interest or use the ENTER/RETURN key to skip:
['mexican', 'italian', 'chinese', 'japanese', 'thai', 'indian', 'american (new)', 'american (traditional)', 'french',
'middle eastern', 'korean', 'mediterranean', 'vietnamese', 'cajun', 'greek', 'hawaiian', 'asian fusion', 'vegetaria
n', 'vegan', 'steakhouse', 'barbeque', 'sushi bars', 'tex-mex', 'specialty food', 'gluten-free', 'coffee & tea', 'des
serts', 'seafood', 'ice cream & frozen yogurt', 'bakeries', 'beer', 'wine & spirits', 'soup', 'pizza', 'hot dogs', 'b
urgers', 'donuts', 'cupcakes', 'salad', 'tacos', 'chicken wings', 'sandwiches', 'bubble tea', 'tapas/small plates',
'shaved ice', 'bagels', 'southern', 'local flavor', 'latin american', 'custom cakes', 'ethinic food']
japanese
Please select one from the following styles as your interest or use the ENTER/RETURN key to skip:
['restaurants', 'fast food', 'food stands', 'street vendors', 'nightlife', 'buffets', 'bars', 'food trucks', 'breakfa
st & brunch', 'diners', 'cocktail bars', 'pubs', 'sports bars', 'wine bars', 'beer bars', 'casinos', 'juice bars & sm
oothies', 'caterers', 'delis', 'cafes', 'lounges', 'music venues', 'performing arts', 'food delivery services', 'dive
bars', 'dance clubs', 'breweries']
restaurants

Please indicate your price range of interest:
1 cheap ($);
2 medium ($$);
3 expensive ($$$);
4 most expensive($$$$)
Please enter the corresponding number(s) separated by comma
2,3
```

```
Great! Filtering criteria fetched! Just one more question before generating your recommendations
Wanna rank your recommendations by 'smart' ratings?
'smart' rating adjusts the original restaurnat average star rating by the number of ratings it receives.
Enter no to deactivate smart ratings or any other key to continue

Awesome, all set! Here is your recommendations:
---------
Below is a list of the top 5 recommended restaurants for you:
       distance_to_interest state       city                     name  \
21311              4.897829    NV  Las Vegas  Yonaka Modern Japanese
10345              3.291231    NV  Las Vegas              Sushi Way
33922              1.704563    NV  Las Vegas                Katsuya
43415              0.501641    NV  Las Vegas             Bocho Sushi
28374              9.556272    NV  Las Vegas               Japaneiro

                                  address  \
21311             4983 W Flamingo Rd, Ste A
10345               3900 Paradise Rd, Ste B
33922  SLS Las Vegas, 2535 S Las Vegas Blvd
43415                        124 S 6th St
28374       7315 W Warm Springs Rd, Ste 170

       attributes.RestaurantsPriceRange2  \
21311                                2.0
10345                                2.0
33922                                3.0
43415                                2.0
28374                                2.0

                                            cuisine  \
21311  japanese, tapas/small plates, asian fusion, am...
10345         japanese, asian fusion, sushi bars, salad
33922                             japanese, sushi bars
43415                             japanese, sushi bars
28374           japanese, asian fusion, seafood, french

                            style  review_count  stars  adjusted_score
21311                  restaurants           940    4.5        4.482340
10345                  restaurants           745    4.5        4.477850
33922                  restaurants           427    4.5        4.462163
43415  restaurants, nightlife, bars           405    4.5        4.460214
28374                  restaurants           401    4.5        4.459837
---------
Would you like to display more/less recommendation results? Enter the desire number to continue or any other key to s
kip:


Would you like to further filter your recommendation results by keywords? Enter yes to continue or any other key to s
kip:

Awesome, all done!
Please enter q to quit the recommender engine, or enter c to restart with another recommendation
q
Enjoy your recommendations! See you next time!
```

Potential Improvements

- **Using Spark and/or cloud platform to speed up the computation**
  This Yelp dataset has a significant amount of data that poses a challenge to work with. For practical application and/or deployment, where models need to be refreshed periodically with the updated dataset, it's better to setup the computation using distributed systems.

- **Using unsupervised learning techniques to group restaurants and/or users into clusters**
  This will enable more customized recommender model to be designed and optimized tailoring towards each cluster, leading to better submodels with improved performance.

- **Incorporating data from the 'tip' and 'checkin' dataset**
  The personalization is currently computed based on data from the 'review' dataset. The 'tip' and 'checkin' datasets also contain information that can be potentially used for personalization.

- **Adjusting the restaurant star rating by time relevance**
  Restaurant ownership, food quality, service and surrounding environment can all change over time, newer ratings are considered more relevant than older ratings. Therefore, the rating metric can be further improved by incorporating the age relevance of the rating.

Summary — Yelp Business Insights

The Yelp open dataset of 5,996,996 reviews, 1,518,169 users, 188,593 businesses, 1,185,348 tips and over 1.4 million business attributes is cleaned and analyzed. Interactive visualizations are also created using Bokeh.

The key business findings are:

- Only a subset of Yelp restaurants from a few selected states are available in this dataset. Among them, only AZ, NV, OH, NC and PA have a rich catalog of over 5000 restaurants.
- The most common restaurants are popular franchised businesses, Starbucks, McDonald's and Subway being the top three.
- The average restaurant rating is around 3.5. Half restaurants have less than 30 reviews, but restaurants from Nevada (Las Vegas) have significant more reviews than others. Restaurants with more reviews tend to have higher ratings on average.
- Most restaurants are in the low (40.9%) and mid (41.6%) price ranges. More expensive restaurants tend to receive more reviews on average, but the average rating remains similar.
- The most popular cuisine of restaurants overall is American style (traditional and new), followed by Mexican, Italian and Chinese; the most popular restaurant setting is the formal restaurant style, followed by nightlife/bar and fast food.
- A steady increase of new users has continued since Yelp's debut in 2004 till 2015, followed by a significant decline thereafter. The average rating given by Yelp users is 3.72, and 60% of the users have less than 10 reviews in total.
- Daily # of reviews posted on Yelp shows a steady upward trend with seasonal fluctuations, whereas daily # of tips only increased in the first four years and slowly dived down thereafter.
- Half of the restaurants have less than 20 checkins, indicating that checkin is less popular than review.

Summary — Hybrid Restaurant Recommendation Engine

A hybrid recommendation engine is implemented powered by the Yelp Dataset. User-friendly interface is also created to gather user interests and navigate users through the three submodules in the engine.

Key features of the hybrid recommendation engine include:

- **A non-personalized keyword-search recommender module** supports a combination of restaurant location-based keyword filtering and restaurant feature-based keyword filtering of restaurant catalog.
- **A personalized collaborative recommender module** supports personalized restaurant recommendation given the unique user_id. The personalization is computed based on the user's and all other users' rating history of all Yelp businesses via an optimized matrix factorization model.
  RMSE of rating prediction is 1.2443; NDCG@10 = 0.905 and NDCG@5 = 0.908
- **A personalized restaurant content-based recommender module** supports personalized restaurant recommendation given the unique user_id. The personalization is computed based on the similarity between the user's preference indicated by historical ratings and all restaurants' features extracted from a rich set of Yelp restaurant review texts.
  NDCG@10 = 0.857 and NDCG@5 = 0.863
- **Further filter a recommendation list by keyword** supports further filtering the recommendation results by a combination of restaurant location-based keywords and restaurant feature-based.
- **A user-friendly interface** supports flexible navigation among the three available recommender modules at user's choice

Acknowledgement

- **Acknowledgement**

  Max Sop (mentorship)

  Yelp (open dataset)

  Springboard team (curriculum & administrative support)

  Recommender Systems Specialization from Coursera (curriculum on recommenders systems)

Reference

- **Project final report:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/final_report.ipynb

- **Data wrangling & EDA notebook:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/milestone_report_github.ipynb
- **Non-personalized keyword module notebook:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/recommender_keyword.ipynb
- **Personalized collaborative module notebook:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/recommender_collaborative.ipynb
- **Personalized restaurant content-based module notebook:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/recommender_content.ipynb
- **Recommender module integration and user interface:**
https://github.com/jingzhaomirror/capstone2_hybrid_yelp_recommender/blob/master/hybrid_recommender.ipynb