Capstone 2: Yelp

Dan Duda
Springboard Career Track

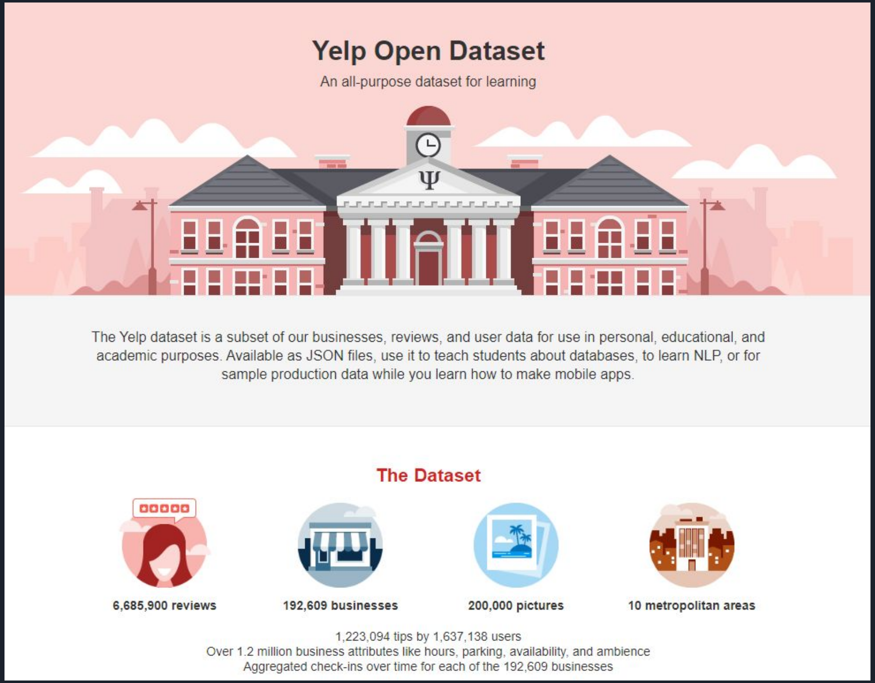
Problem Statement

I will build a restaurant recommendation system for users utilizing a combination of two methods: content-based filtering and collaborative filtering.

UPDATED: Provide insight as to whether a user would like (give a rating greater than 3) as well as if a user would give a rating greater than the business' average. This will to also identify which attributes would impact that result as well.

The Data

- Yelp Public Dataset for Yelp Challenge
  - Round 13
- 12 Metropolitan Areas
- 156,000 Local Businesses
- Recommendation System is the Goal



Yelp Open Dataset

An all-purpose dataset for learning

The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

The Dataset

6,685,900 reviews

192,609 businesses

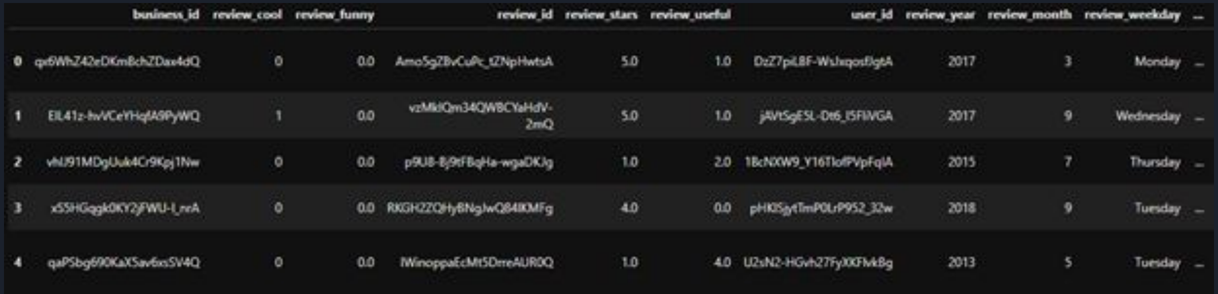
200,000 pictures

10 metropolitan areas

1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses

Cleaning The Data

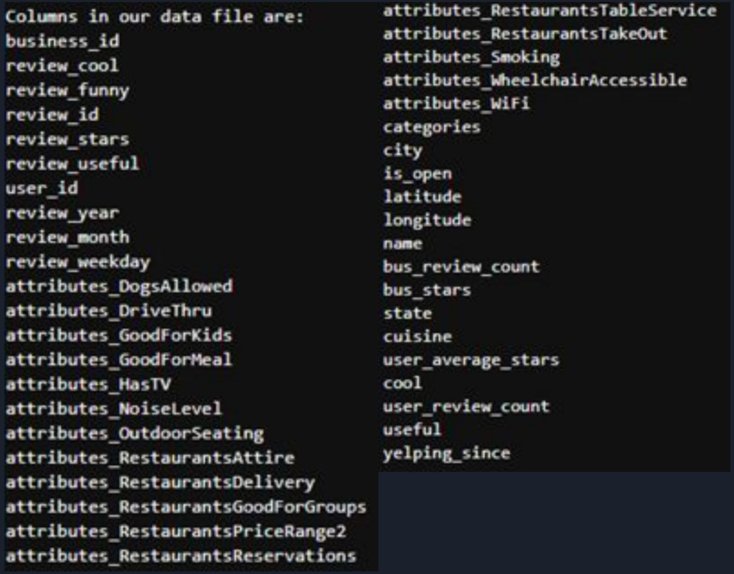
- Read in necessary JSON files using Pandas
    - review.json
    - business.json
    - User.json
- Converted to .csv format
- Removed empty user_id data
- Only kept business data within the ‘Food’ or ‘Restaurant’ categories
- Issues with computation power and timeliness of running code on such large files
    - Reduced dataset to just Arizona businesses and other datasets related to Arizona.



	business_id	review_cool	review_funny	review_id	review_stars	review_useful	user_id	review_year	review_month	review_weekday	...
0	qx6WhZ42eDKmBchZDax4dQ	0	0.0	Amo5gZBvCuPc_tZNpHwtsA	5.0	1.0	DzZ7piLBF-WsJxqosfJgtA	2017	3	Monday	...
1	EIL41z-hvVCeYHqtA9PyWQ	1	0.0	vzMkIQm34QWBCYaHdV-2mQ	5.0	1.0	jAVtSgESL-Dt6_I5FIiVGA	2017	9	Wednesday	...
2	vhIJ91MDgUuk4Cr9Kpj1Nw	0	0.0	p9U8-Bj9tFBqHa-wgaDKJg	1.0	2.0	1BcNXW9_Y16TlofPVpFqIA	2015	7	Thursday	...
3	x5SHGqgk0KY2jFWU-I_nrA	0	0.0	RKGH2ZQHyBNgJwQ84lKMFg	4.0	0.0	pHKISgytTmP0LrP952_32w	2018	9	Tuesday	...
4	qaPSbg690KaXSav6xsSV4Q	0	0.0	lWinoppaEcMt5DmeAUR0Q	1.0	4.0	U2sN2-HGvh27FyXKFlvkBg	2013	5	Tuesday	...

Columns in Merged and cleaned Dataset

- Columns were a merging of:
  - User data,
  - Business data
    - Many of which were attributes of the businesses,
  - Review data

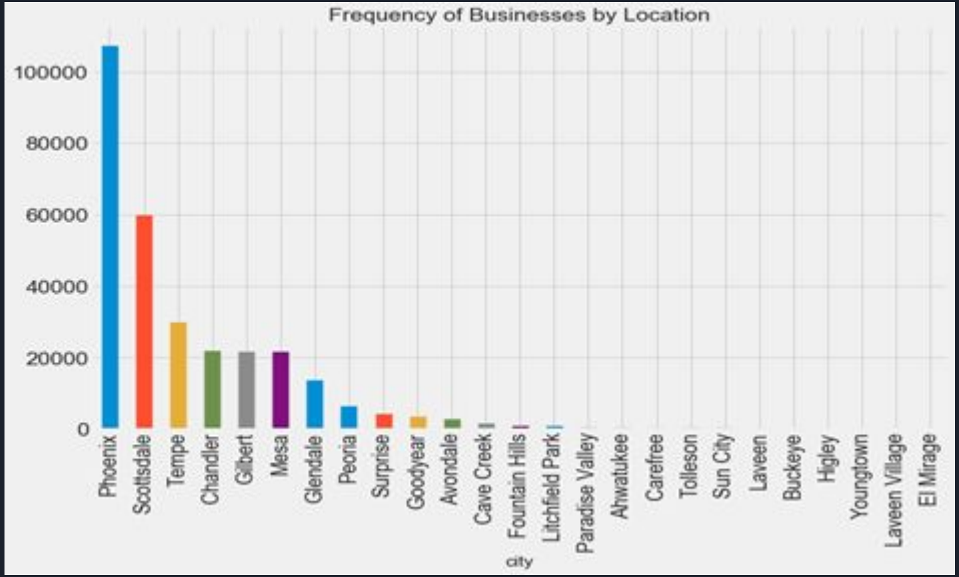


```
Columns in our data file are:
business_id
review_cool
review_funny
review_id
review_stars
review_useful
user_id
review_year
review_month
review_weekday
attributes_DogsAllowed
attributes_DriveThru
attributes_GoodForKids
attributes_GoodForMeal
attributes_HasTV
attributes_NoiseLevel
attributes_OutdoorSeating
attributes_RestaurantsAttire
attributes_RestaurantsDelivery
attributes_RestaurantsGoodForGroups
attributes_RestaurantsPriceRange2
attributes_RestaurantsReservations
```

```
attributes_RestaurantsTableService
attributes_RestaurantsTakeOut
attributes_Smoking
attributes_WheelchairAccessible
attributes_WiFi
categories
city
is_open
latitude
longitude
name
bus_review_count
bus_stars
state
cuisine
user_average_stars
cool
user_review_count
useful
yelping_since
```

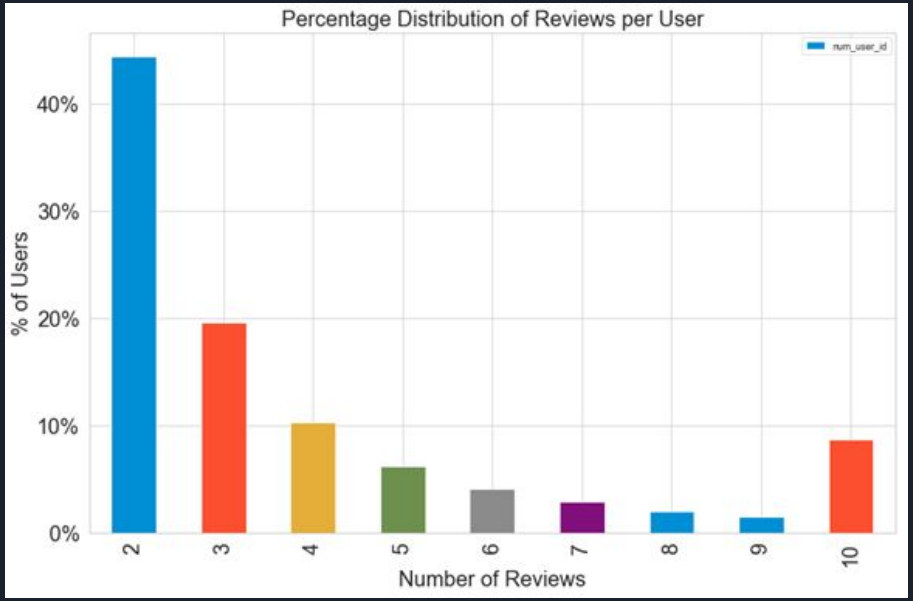
Data Exploration

- Frequency of restaurants based on local locations.



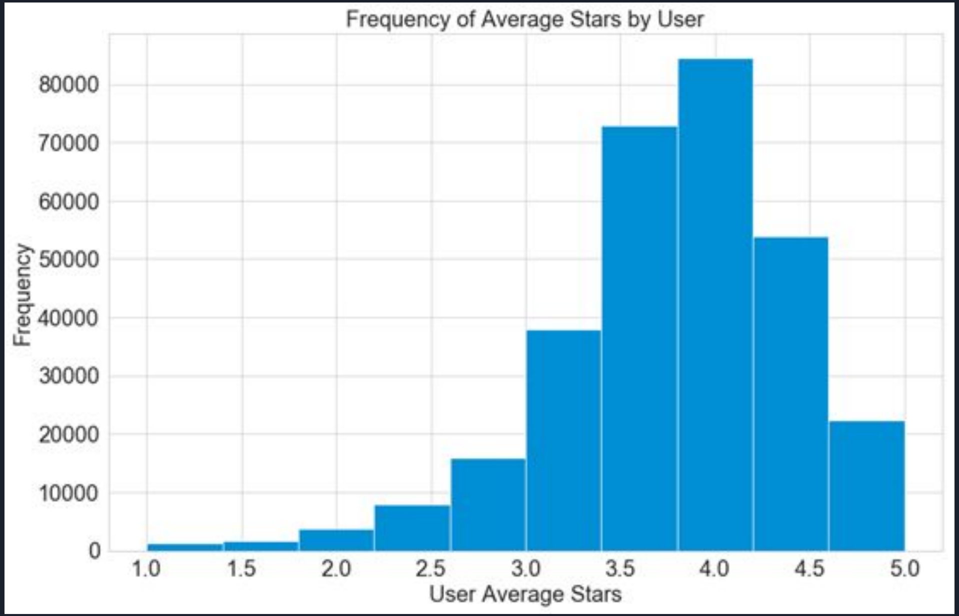
Data Exploration (cont.)

- This is a histogram plot showing the percentage distribution of reviews per user.
  - We can see the majority of users actually have 3 or less reviews. There is a noticeable spike for 10+ reviews.
- It seems most Yelper's are either very active or not very active, with little room between.



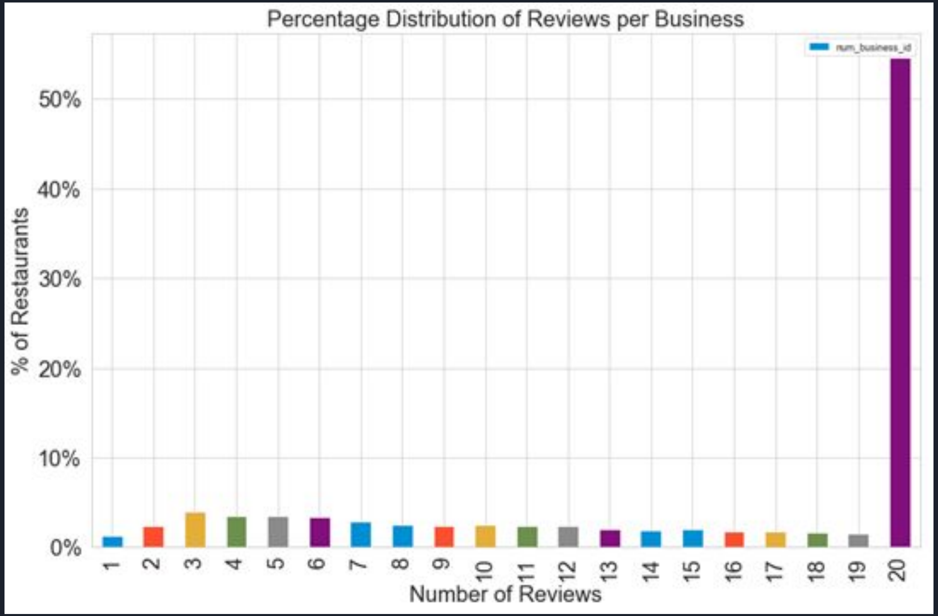
Data Exploration (cont.)

- A distribution of average stars by user
- Potential bias to keep in mind as majority of users give fairly high ratings
- Average Stars Given by Users: 3.781



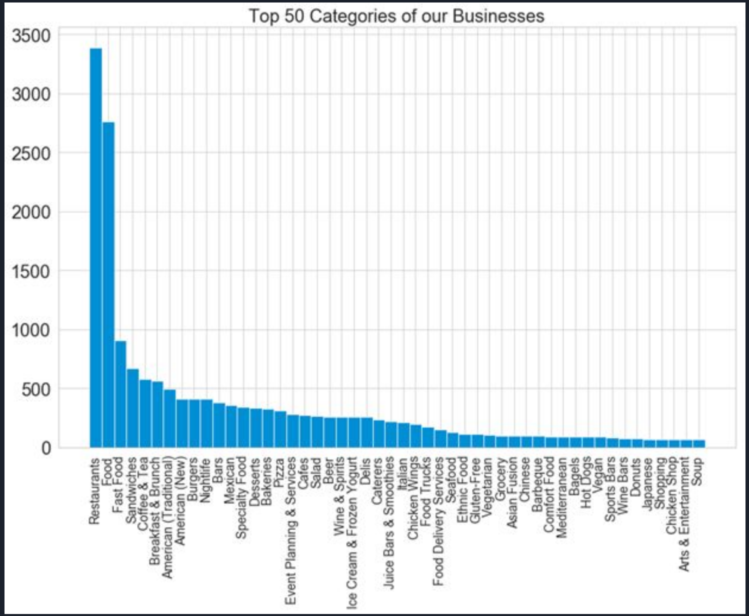
Data Exploration (cont.)

- The percentage distribution of reviews per business
- A noticeable jump in the number of businesses that have 20 or more reviews
  - Over 50%



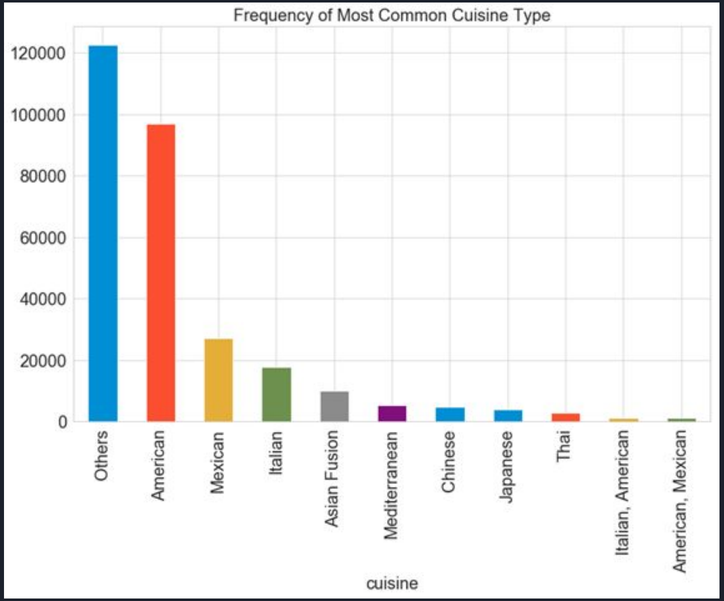
Data Exploration (cont.)

- This is a distribution of the most popular categories and their frequency in which they appear within our businesses
- There are a fair amount of businesses that have just one category tag, such as Restaurants or Food, which is why we see such high frequencies for those categories.



Data Exploration (cont.)

- A breakdown of the most common cuisine types
- This was done by identifying the 10 most common types and then specifying the remainder as “Others”
- You can see there are a lot of “Others” which can be related to the large variety of cuisine types in Arizona.



Important Statistics

- Median value:
  - 4.000000
- Mean value:
  - 3.771028
- Standard deviation value:
  - 1.363917

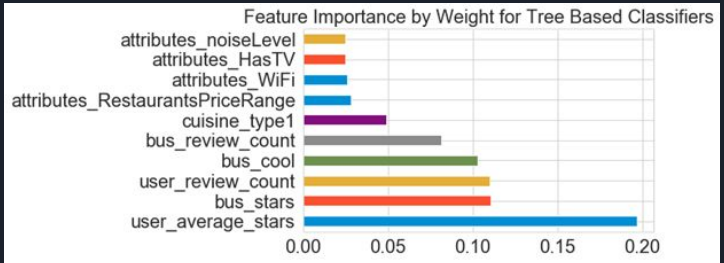
- .10% sparsity of the matrix resulted in the change in project goal.

- Number of users :
  - 62446
- Number of reviews :
  - 301731
- Number of ratings :
  - 301731
- Average reviews/ratings per user:
  - 4.83187

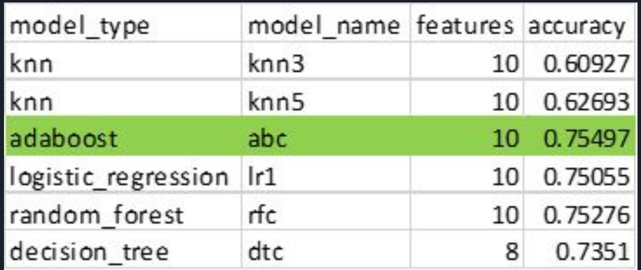
Machine Learning

- Two separate questions being explored:
  - Would a user like a restaurant?
    - Give a rating greater than 3
  - Would a user give a rating greater than the business' average rating?
- These questions were addressed by utilizing classification systems
  - KNN, ADAboost, SVM, Random Forest, Decision Tree
- More insights to be discovered than just the questions?
  - Which restaurant features provide the greatest impact on whether a user would like it?

Question 1: Would a User Like a Restaurant?



- Features to be used for our models



model_type	model_name	features	accuracy
knn	knn3	10	0.60927
knn	knn5	10	0.62693
adaboost	abc	10	0.75497
logistic_regression	lr1	10	0.75055
random_forest	rfc	10	0.75276
decision_tree	dtc	8	0.7351

- ADAboost Classifier was most accurate model
  - ~75% accuracy

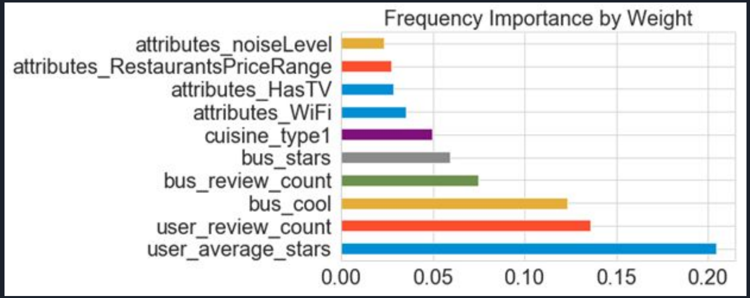
Deeper Dive Into ADAboost Model

Question 1 Model

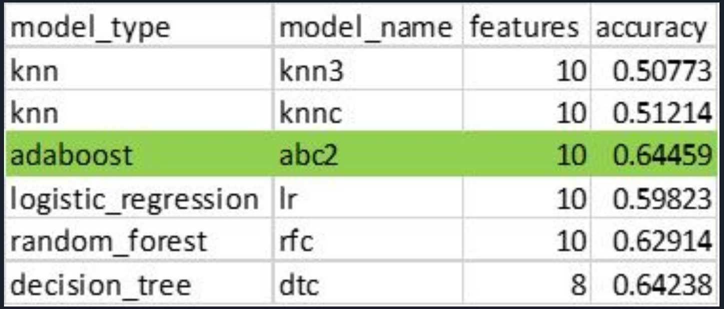
Attribute	Result	Description
Accuracy	.755	How often classifier is correct
Error (misclassification rate)	.245	How often classifier is incorrect
Sensitivity	.8804	When actual value is positive, how often is classifier correct
Specificity	.5066	When actual value is negative, how often is classifier correct

77 True Negatives	75 False Positives
36 False Negatives	265 True Positives

Question 2: Would a User Give a Rating Greater Than the Businesses Average?



- Features to be used for our models.
  - The 10 with the greatest impact



model_type	model_name	features	accuracy
knn	knn3	10	0.50773
knn	knnc	10	0.51214
adaboost	abc2	10	0.64459
logistic_regression	lr	10	0.59823
random_forest	rfc	10	0.62914
decision_tree	dtc	8	0.64238

- ADAboost Classifier was most accurate model
  - ~65% accuracy

Deeper Dive Into ADAboost Model

Question 2 Model

Attribute	Result	Description
Accuracy	.645	How often classifier is correct
Error (misclassification rate)	.355	How often classifier is incorrect
Sensitivity	.696	When actual value is positive, how often is classifier correct
Specificity	.594	When actual value is negative, how often is classifier correct

136 True Negatives	93False Positives
68 False Negatives	156True Positives

Conclusion

- Changed goal due to data and computational issues.
  - Original matrix had .10% Sparsity, too sparse to predict from
- Question 1: Would a user like a restaurant?
  - ADAboost Model was most accurate
    - 10 features
    - ~75% accuracy
- Question 2: Would a user give a rating greater than the businesses average?
  - ADAboost Model was most accurate
    - 10 features
    - ~65% accuracy
- Future Work:
  - Utilize big data entities to work with original data files
  - Test for User Average Star bias
    - Could result in impacting Yelp Challenge on a larger scale