STA 199
Final Project Presentation

Ivy Shi, Katie Wilbur, Miles Turpin

1.

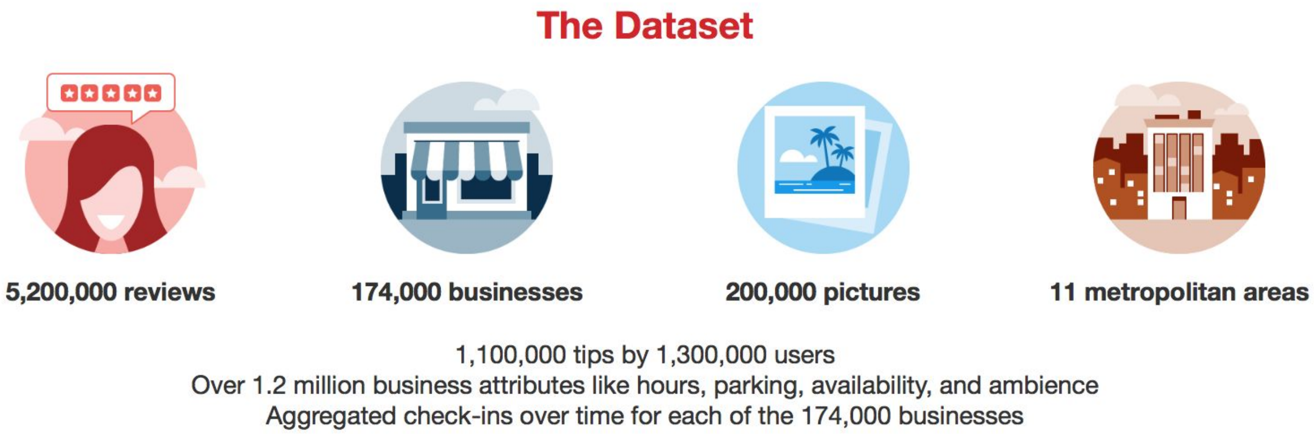
Dataset and Background Information

What is Yelp?

Yelp is a web and mobile platform that publishes crowd-sourced reviews about local businesses, as well as online reservation service through Yelp reservations.

Dataset

The data is released by the Yelp Dataset Challenge to encourage student to conduct research and analysis. It contains a subset of Yelps’ businesses, reviews, and user data.



The Dataset

5,200,000 reviews

174,000 businesses

200,000 pictures

11 metropolitan areas

1,100,000 tips by 1,300,000 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 174,000 businesses

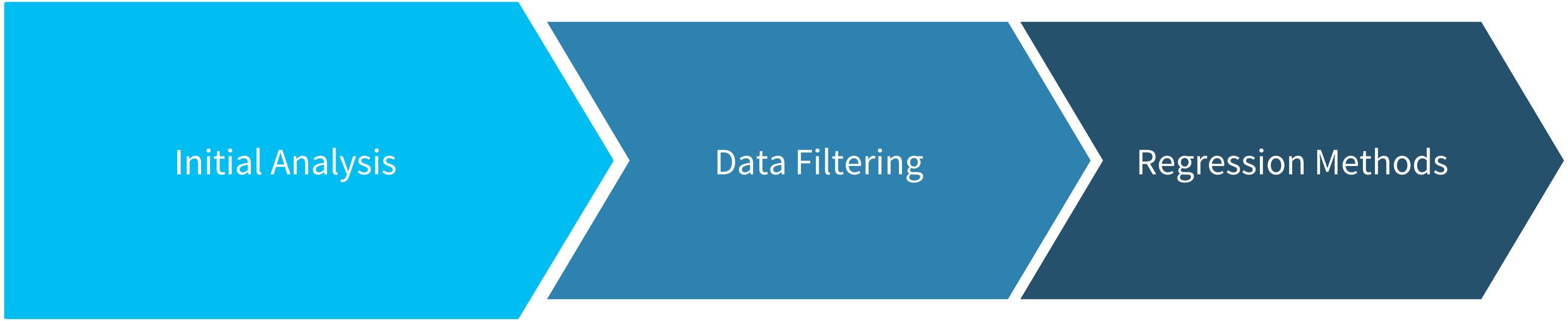
Big Dataset → Narrow down to restaurant data within the business subset

2.

Research Question & Methods

Q: What attributes contribute to high restaurant star ratings? -

A: Build a regression model to predict star ratings based on restaurant characteristics

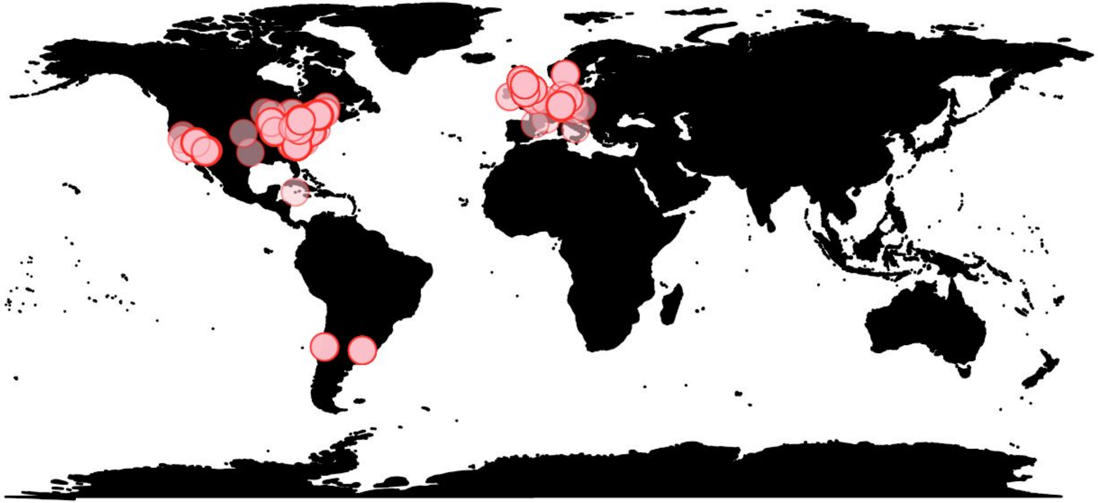


Initial Analysis

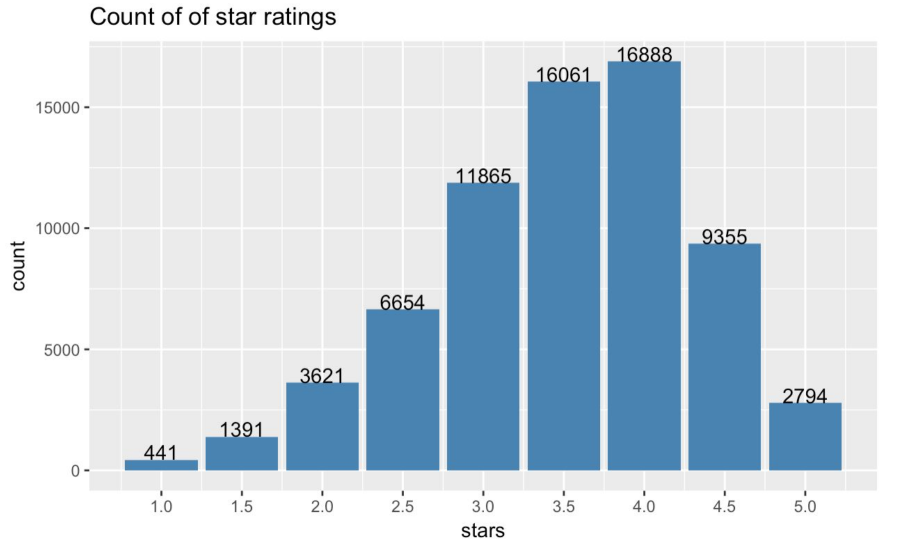
Data Filtering

Regression Methods

Initial Analysis of Data



Initial Analysis of Data



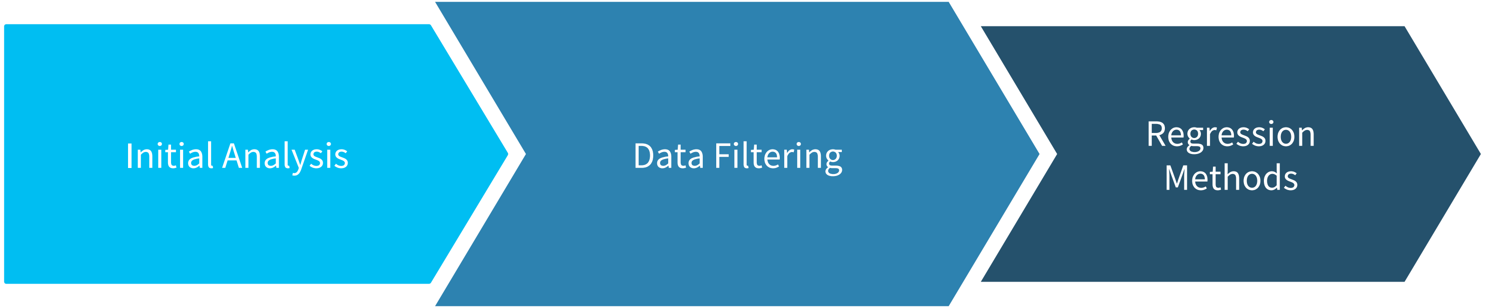
Initial Restaurant Data

60970 observations

88 variable

Mean:3.5 stars

Median :3.5 stars

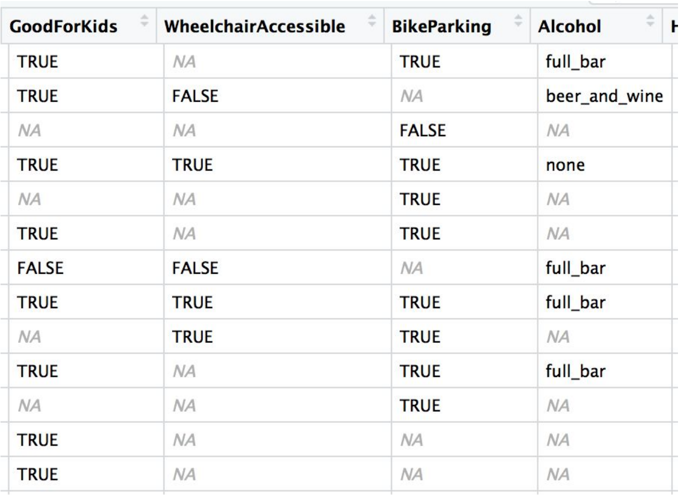


Initial Analysis

Data Filtering

Regression Methods

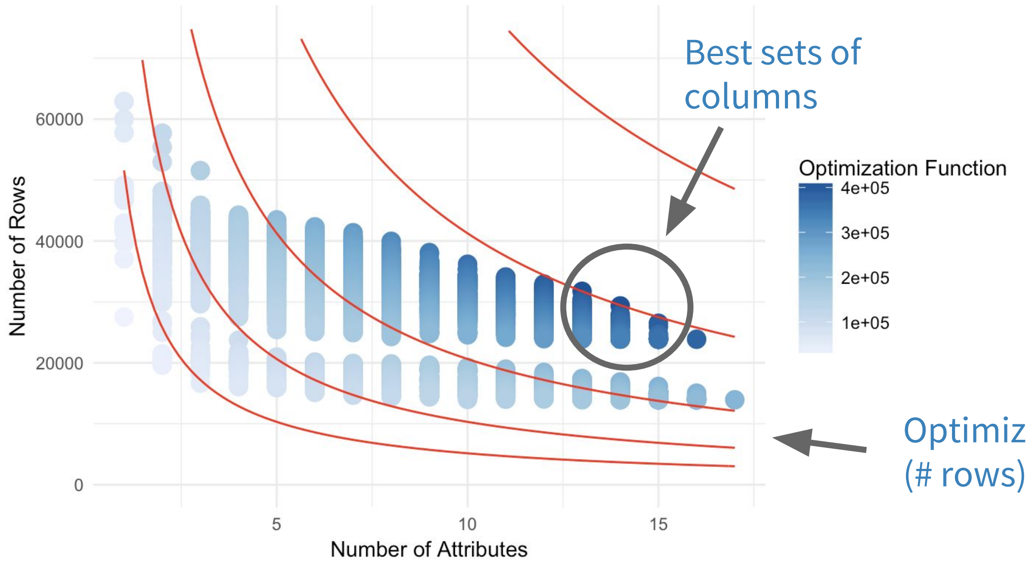
Problem – Too Many NAs, Too Many Columns



GoodForKids	WheelchairAccessible	BikeParking	Alcohol	H
TRUE	NA	TRUE	full_bar	
TRUE	FALSE	NA	beer_and_wine	
NA	NA	FALSE	NA	
TRUE	TRUE	TRUE	none	
NA	NA	TRUE	NA	
TRUE	NA	TRUE	NA	
FALSE	FALSE	NA	full_bar	
TRUE	TRUE	TRUE	full_bar	
NA	TRUE	TRUE	NA	
TRUE	NA	TRUE	full_bar	
NA	NA	TRUE	NA	
TRUE	NA	NA	NA	
TRUE	NA	NA	NA	

- Models need all columns to be non-null
- Very sparse dataset; unclear what is important

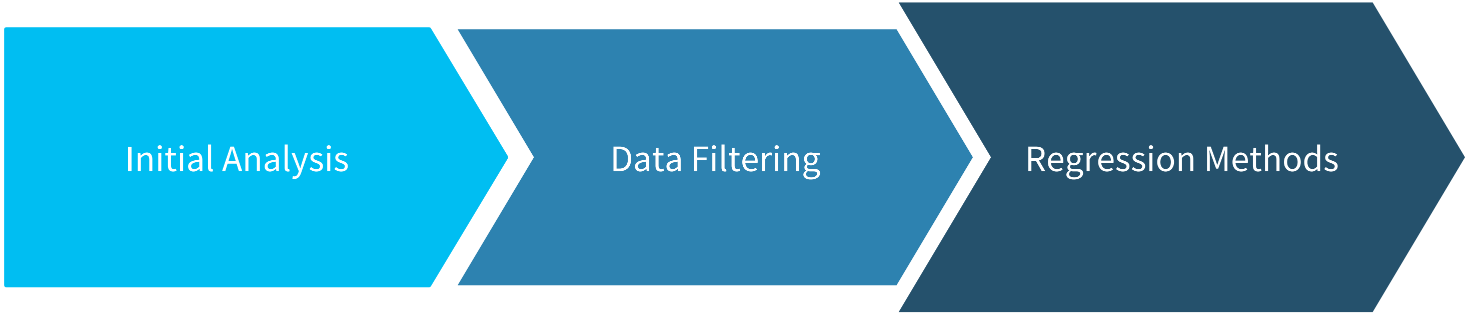
Solution - Optimizing Columns



Best sets of columns

- Each point represents combination of columns
- Find subset of columns that maximizes data

Optimization Function = (# rows) x (# columns)

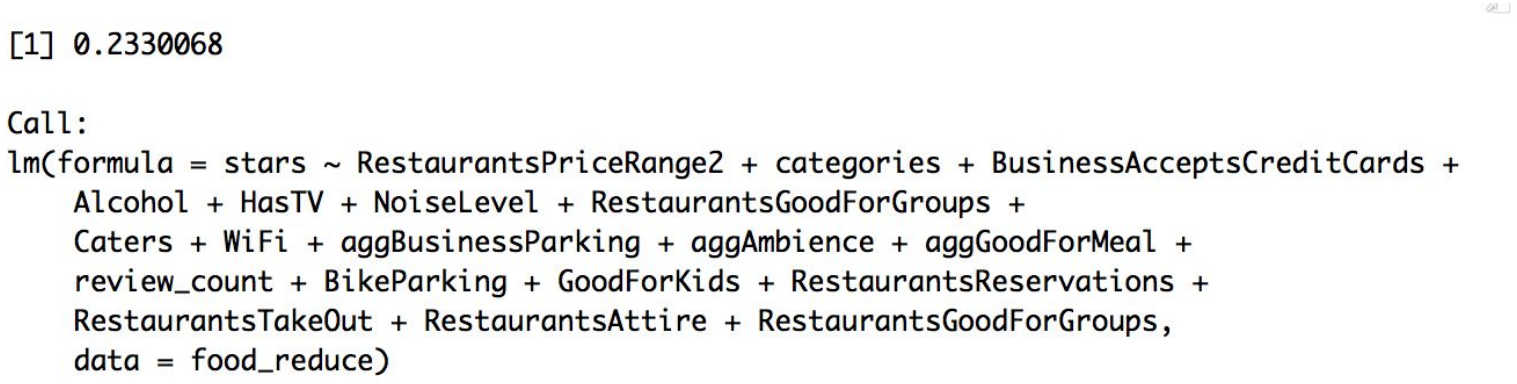


Initial Analysis

Data Filtering

Regression Methods

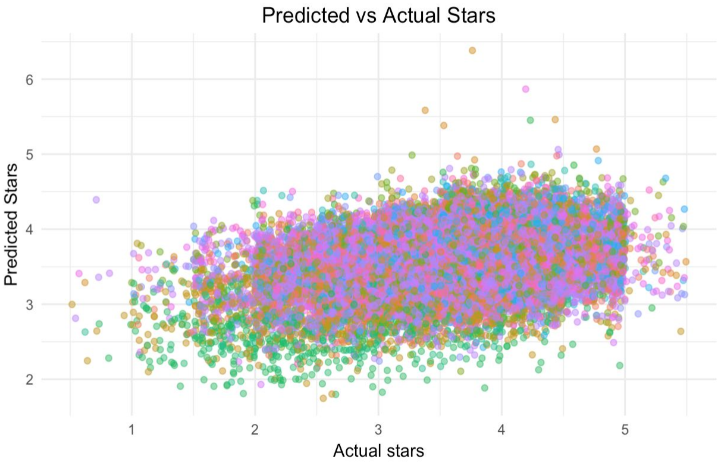
Multiple Linear Regression Model

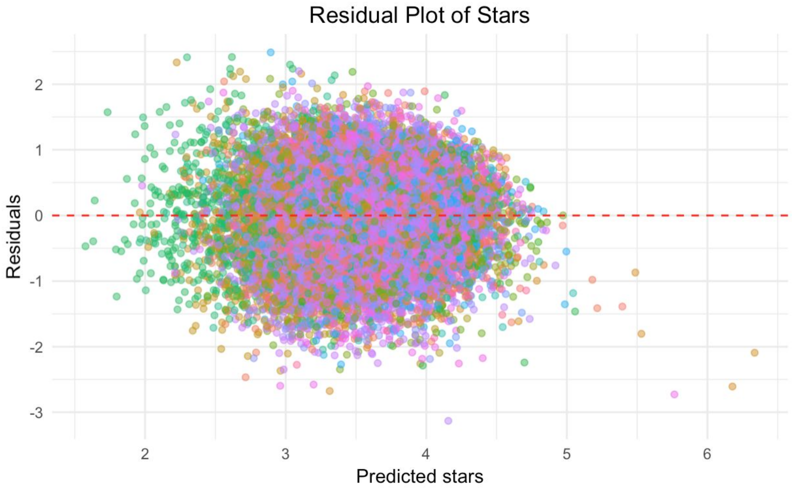


```
[1] 0.2330068

Call:
lm(formula = stars ~ RestaurantsPriceRange2 + categories + BusinessAcceptsCreditCards +
    Alcohol + HasTV + NoiseLevel + RestaurantsGoodForGroups +
    Caters + WiFi + aggBusinessParking + aggAmbience + aggGoodForMeal +
    review_count + BikeParking + GoodForKids + RestaurantsReservations +
    RestaurantsTakeOut + RestaurantsAttire + RestaurantsGoodForGroups,
    data = food_reduce)
```

Residual vs Fitted Plot (colored by categories)

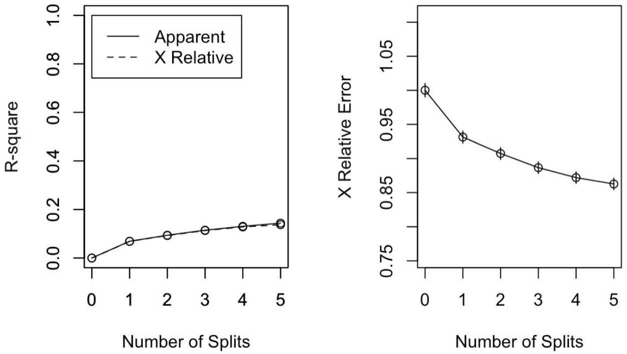


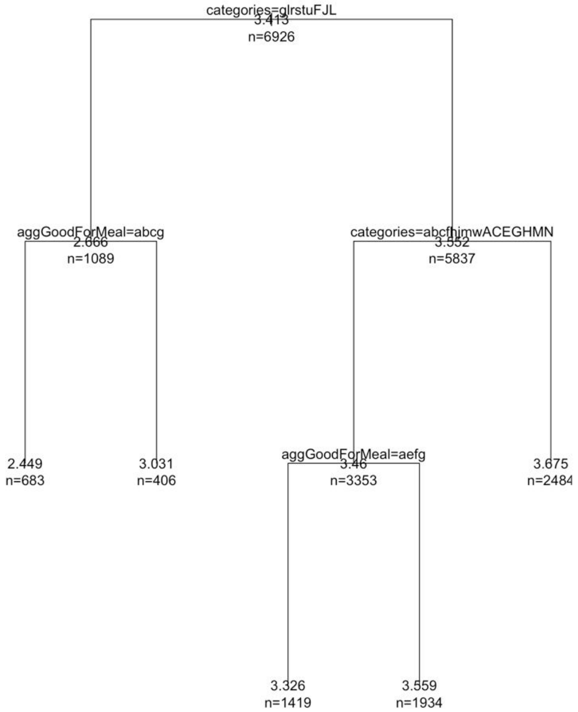


Regression Decision Tree Algorithm

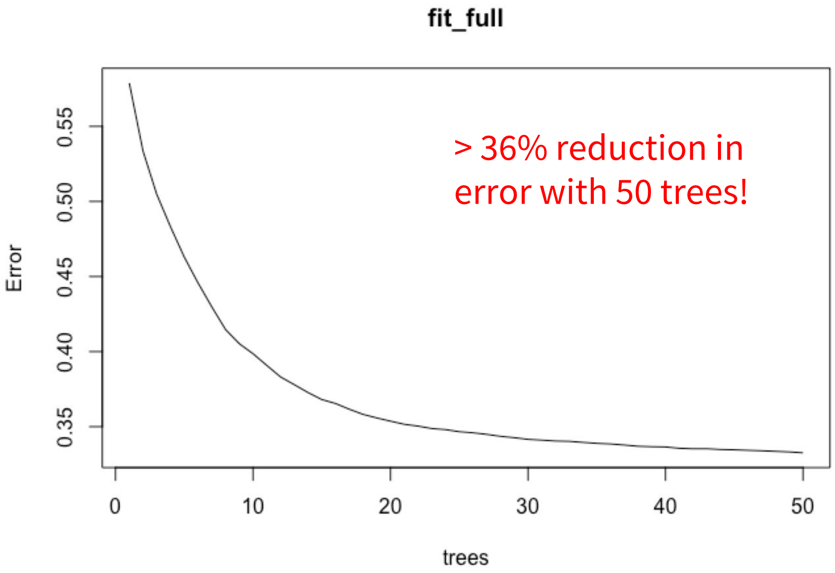
Concept

Non-linear regression method that predicts a value by building a tree





Random Forest Algorithm



> 36% reduction in error with 50 trees!

Concept

Train many decision trees and average their results, cancels out any bias/noise from any single tree

3.

Conclusion

Conclusion for this dataset

- Random forest algorithm most successful (R-squared of 0.277%)
- Prevalence of low R-squared values suggests that data is non-linear

	Random forest	Linear regression	Decision tree
R-squared	0.277	0.230	0.147