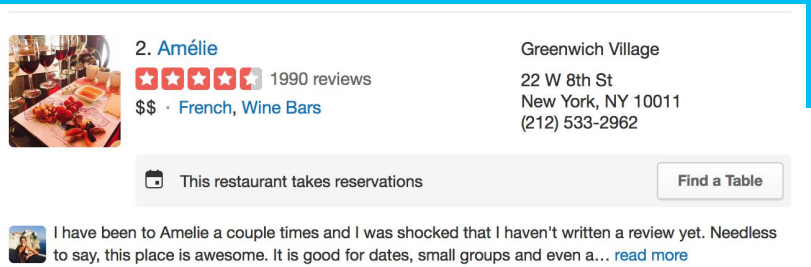
STA 199
Final Project Presentation

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

business_id	BusinessAcceptsCreditCards	RestaurantsPriceRange2	GoodForKids	WheelchairAccessible	BikeParking	Alcohol
PfOCPjBrlQAnz__NXj9h_w	TRUE	2	TRUE	NA	TRUE	full_bar
o9eMRCWt5PkpLDE0gOPtcQ	TRUE	3	TRUE	FALSE	NA	beer_and_wine
EsMcGiZaQuG1OOvL9iUFug	TRUE	1	NA	NA	FALSE	NA
XOSRcvtaKc_Q5H1SAzN20A	TRUE	1	TRUE	TRUE	TRUE	none
xcgFnd-MwkZeO5G2HQ0gAQ	FALSE	1	NA	NA	TRUE	NA
fNMVV_ZX7CJSDWQGdOM8Nw	TRUE	1	TRUE	NA	TRUE	NA
l09JfMeQ6ynYs5MCJtrcmQ	TRUE	3	FALSE	FALSE	NA	full_bar
Gu-xs3NIQTj3Mj2xYoN2aw	TRUE	2	TRUE	TRUE	TRUE	full_bar
lHYiCS-y8AFjUitv6MGpxg	TRUE	2	NA	TRUE	TRUE	NA
1K4qrnfyzKzGgJPBEcJaNQ	NA	2	TRUE	NA	TRUE	full_bar
AtdXq_gu9NTE5rx4ct_dGg	TRUE	2	NA	NA	TRUE	NA
Dj0S-Oe4ytRJzMGUPgYUkw	TRUE	NA	TRUE	NA	NA	NA
gAy4LYpsScrj8POnCW6btQ	TRUE	NA	TRUE	NA	NA	NA
nbhBRhZtdaZmMMeb2i02pg	NA	1	TRUE	TRUE	TRUE	none
1_3nOM7s9WqnJWTNu2-i8Q	TRUE	2	TRUE	NA	NA	full_bar
FXHfcFVEf1vVnqW2qVOpw	TRUE	2	NA	FALSE	TRUE	NA

FALSE

Initial restaurant data overview - 60970 observations, 88 variables

2.

Research Question & Methods

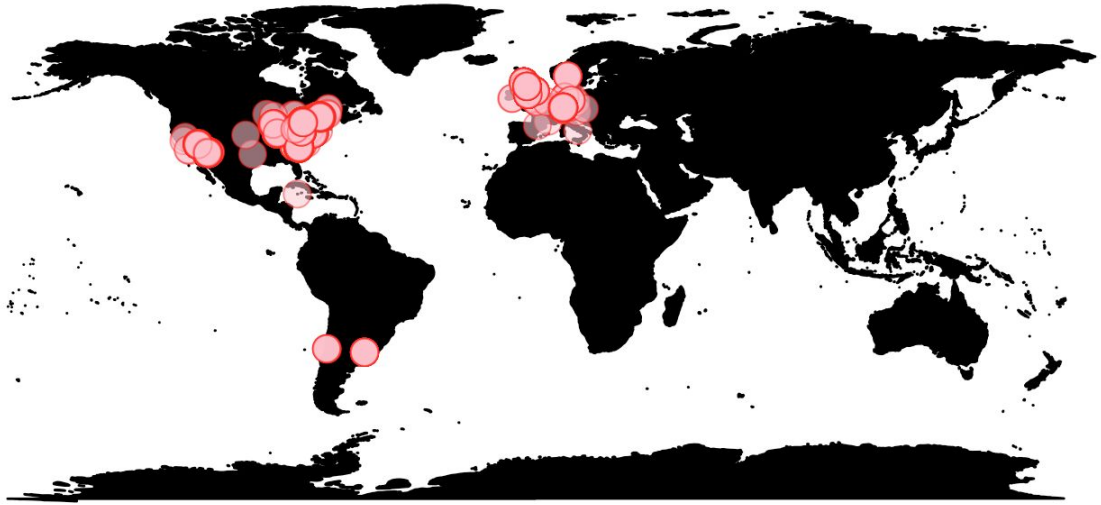
“

Q: What attributes contribute to high restaurant star ratings?

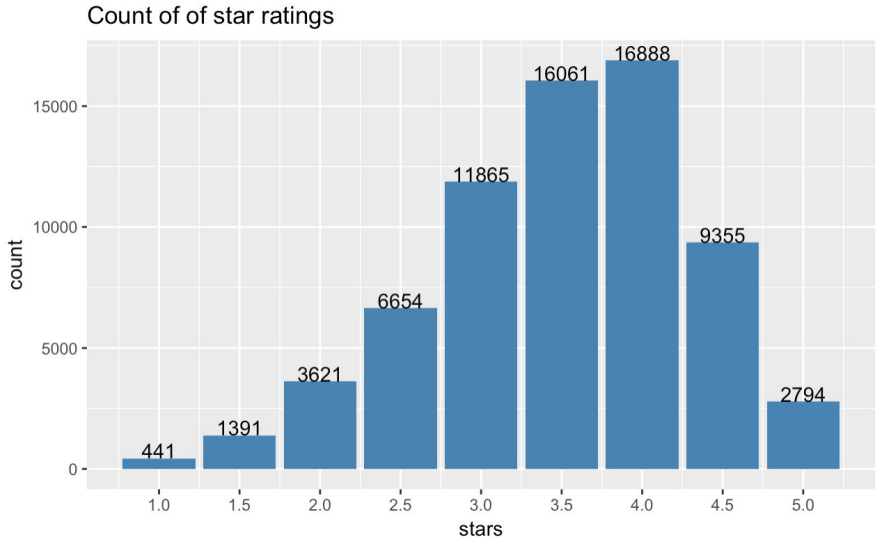
A: Build a regression model to predict star ratings based on restaurant characteristics


Initial Analysis
Data Filtering
Regression Methods

Initial Analysis of Data



Initial Analysis of Data



Initial Restaurant Data

60970 observations

88 variables

Mean:3.5 stars

Median :3.5 stars

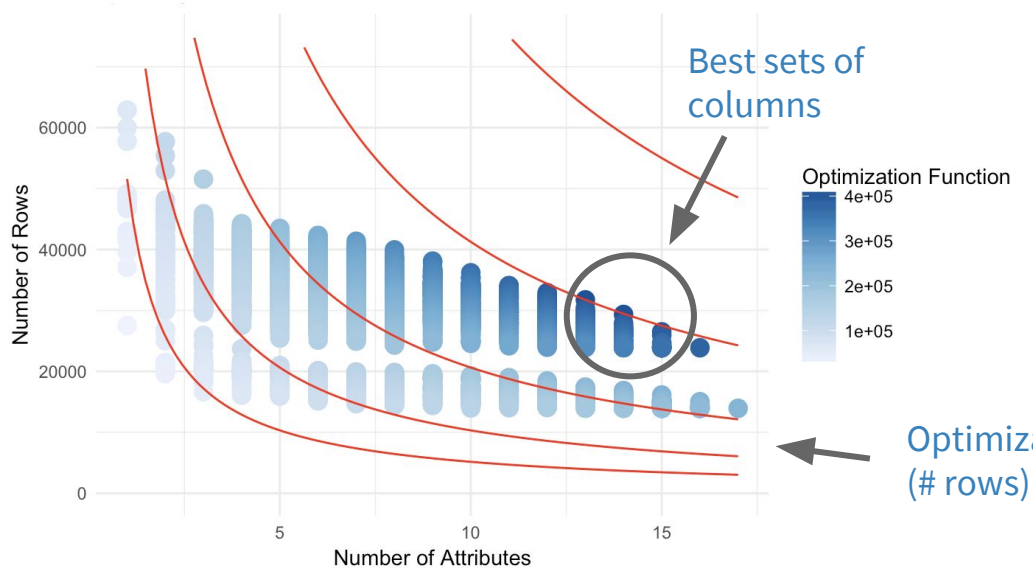

Initial Analysis
Data Filtering
Regression Methods

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



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

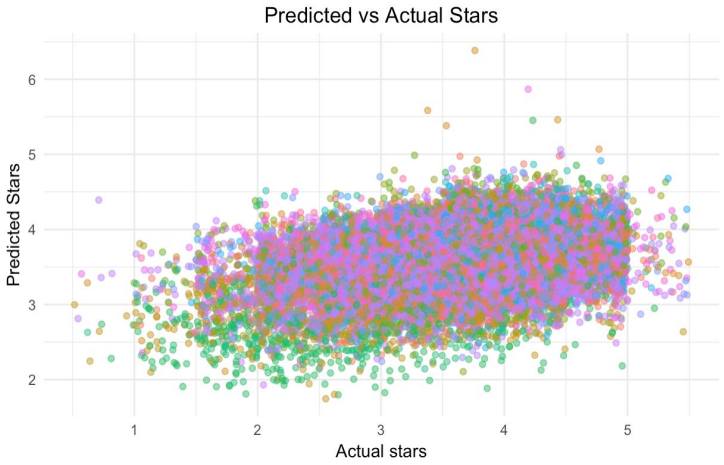

Initial Analysis
Data Filtering
Regression Methods

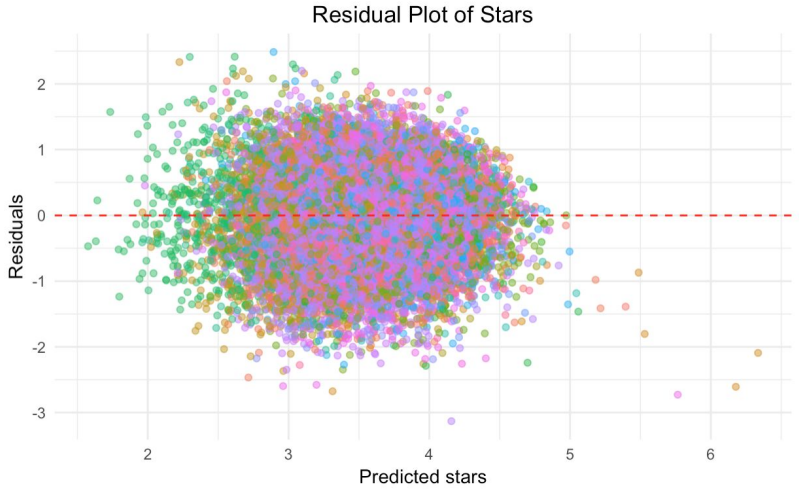
Multiple Linear Regression Model


```
[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)

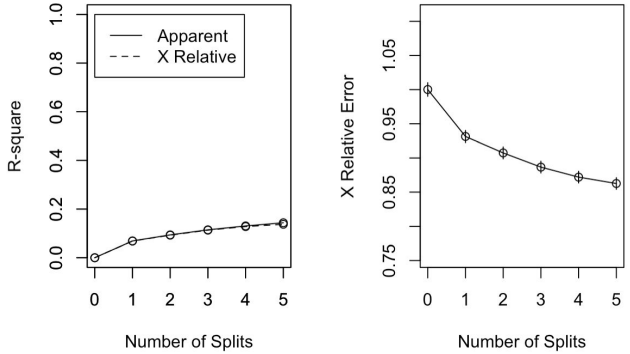


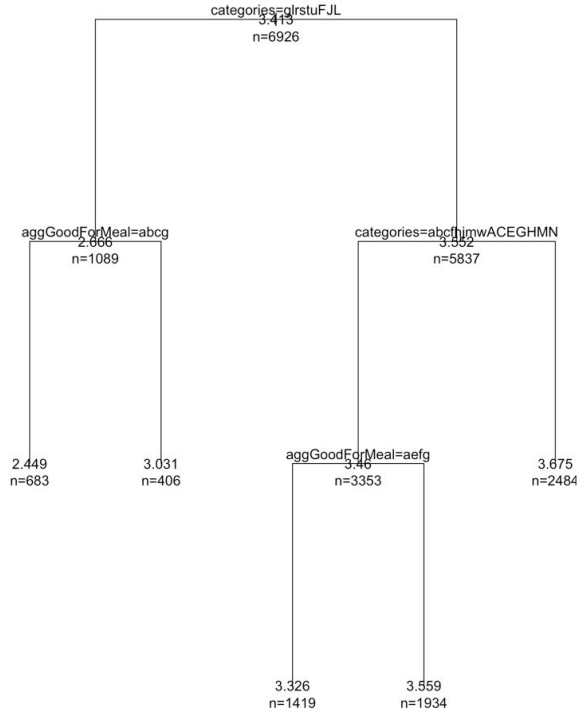


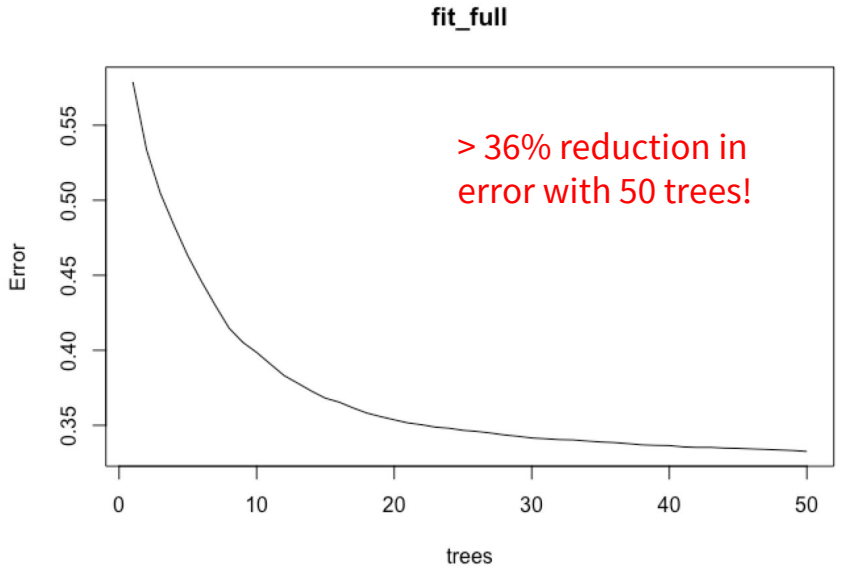
Regression Decision Tree Algorithm

Concept

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







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)

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