

A DATA MINING PROJECT

YELP US! WE'RE DROWNING

TEAM MEMBERS

- ▶ Devin Arnold: Banh Mi
- ▶ Matthew Coker: White Widow
- ▶ Taylor Gunter: McDougle inside Spicy McChicken, extra pickles
- ▶ Keaton Whitehead: Eggs and Ham with a side of Ketchup and Mustard in a tiny cup

PROJECT DESCRIPTION

- ▶ The restaurant business is notoriously difficult. Margins are slim, risk is high, and people are picky; but over time many restaurants do find success. We are setting out to determine what makes a restaurant successful. Are there objectively better ways to run a restaurant to increase the chance of success, or is the secret just in the sauce?

OBJECTIVE

- ▶ Prove successful restaurants have more in common than just great food
- ▶ What do successful restaurants have in common?
- ▶ What failed restaurants have in common?
- ▶ How does mean and median household income of the area a restaurant is located in affect restaurant success?

PRIOR WORK

- ▶ Predicting New Restaurant Success- <https://web.stanford.edu/class/cs221/2017/restricted/p-final/wizeng/final.pdf>
- ▶ Predicting New Restaurant Failure- <https://towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6>
- ▶ Review rating prediction- <https://cs.uwaterloo.ca/~nasghar/886.pdf>
- ▶ Yelp Visualization- <https://blog.exploratory.io/working-with-json-data-in-very-simple-way-ad7ebcc0bb89>

DATASETS

- ▶ Yelp Dataset
- ▶ Mean and median household income, and population by zip code

DATASET SOURCE

- ▶ <https://www.kaggle.com/yelp-dataset/yelp-dataset>
- ▶ <https://www.yelp.com/dataset>
- ▶ <https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/>

IS THE DATA DOWNLOADED?

- ▶ The data is downloaded.
- ▶ CSV
- ▶ SQL
- ▶ JSON

PROPOSED WORK- DATA CLEANING

- ▶ Discard sparse tuples
- ▶ Discard non-restaurant reviews
- ▶ Discard outliers.
 - ▶ Restaurants with fewer than 30 reviews

PROPOSED WORK- DATAPREPROCESSING: GARBAGE IN; GARBAGE OUT

- ▶ Create consistency rules
 - ▶ Cash Only: Yes and Accepts Credit Cards: Yes = False
- ▶ Fill in missing values with content of other attributes
- ▶ Normalize tables to make joins make sense
- ▶ Add additional attributes
 - ▶ Price > \$50 and Atmosphere: Quiet = Family Friendly: False
- ▶ Transform strings to ints or string to bools

PROPOSED WORK- DATA INTEGRATION

- ▶ Integrate with household income data
- ▶ Join tables

METHODS

- ▶ Null Hypothesis testing based on chosen attributes
- ▶ Descriptive Modeling
- ▶ Pattern Mining
- ▶ Correlation Analysis
- ▶ Visualization- JSON heat map

TOOLS

- ▶ Python
- ▶ Numpy
- ▶ Scipy
- ▶ Pandas
- ▶ mySQL
- ▶ Tableau

EVALUATION

- ▶ Was restaurant failure or success confirmed by our descriptive analyses?
- ▶ Model restaurant parameters and map with success or failure.
- ▶ What restaurant attributes really matter?