GENERAL ASSEMBLY

AUSTIN RESTAURANTS HEALTH INSPECTION

Project Final Presentation

by Nikos Vergos

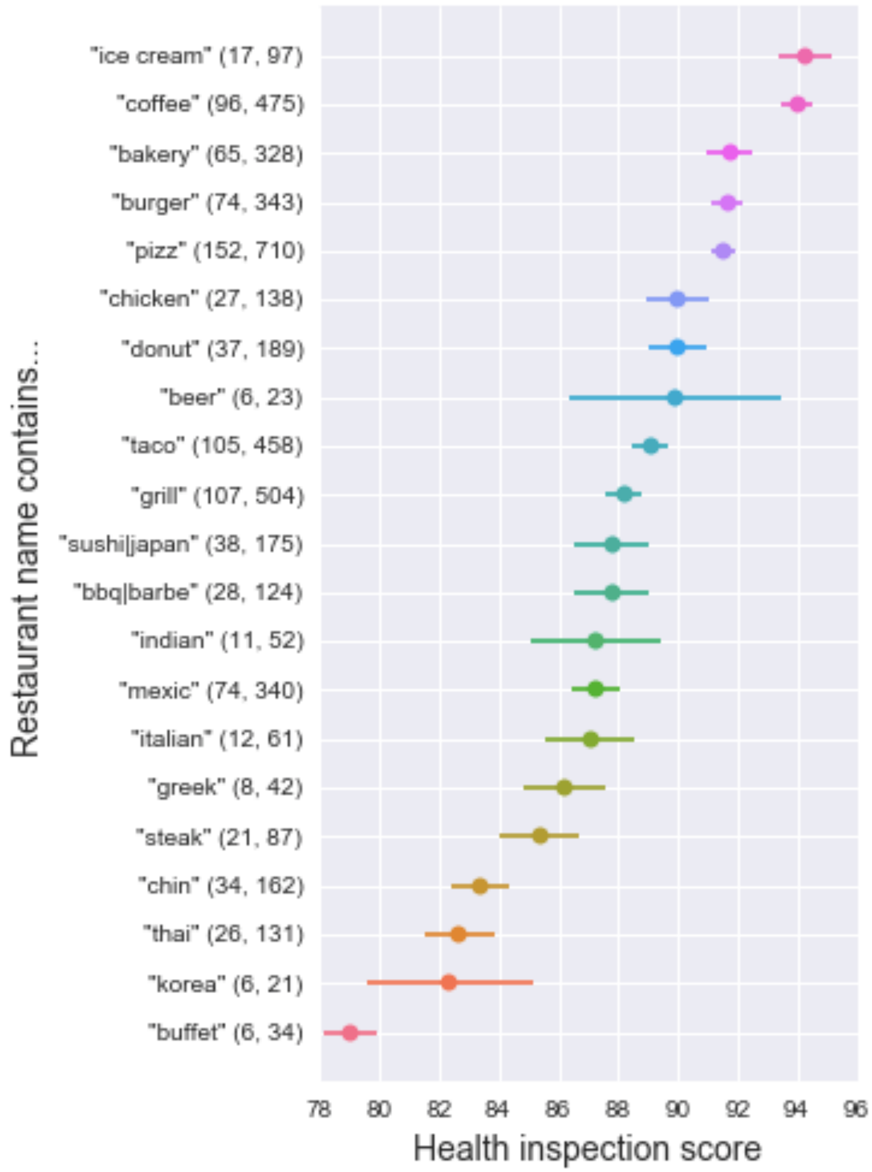
PREVIOUSLY…

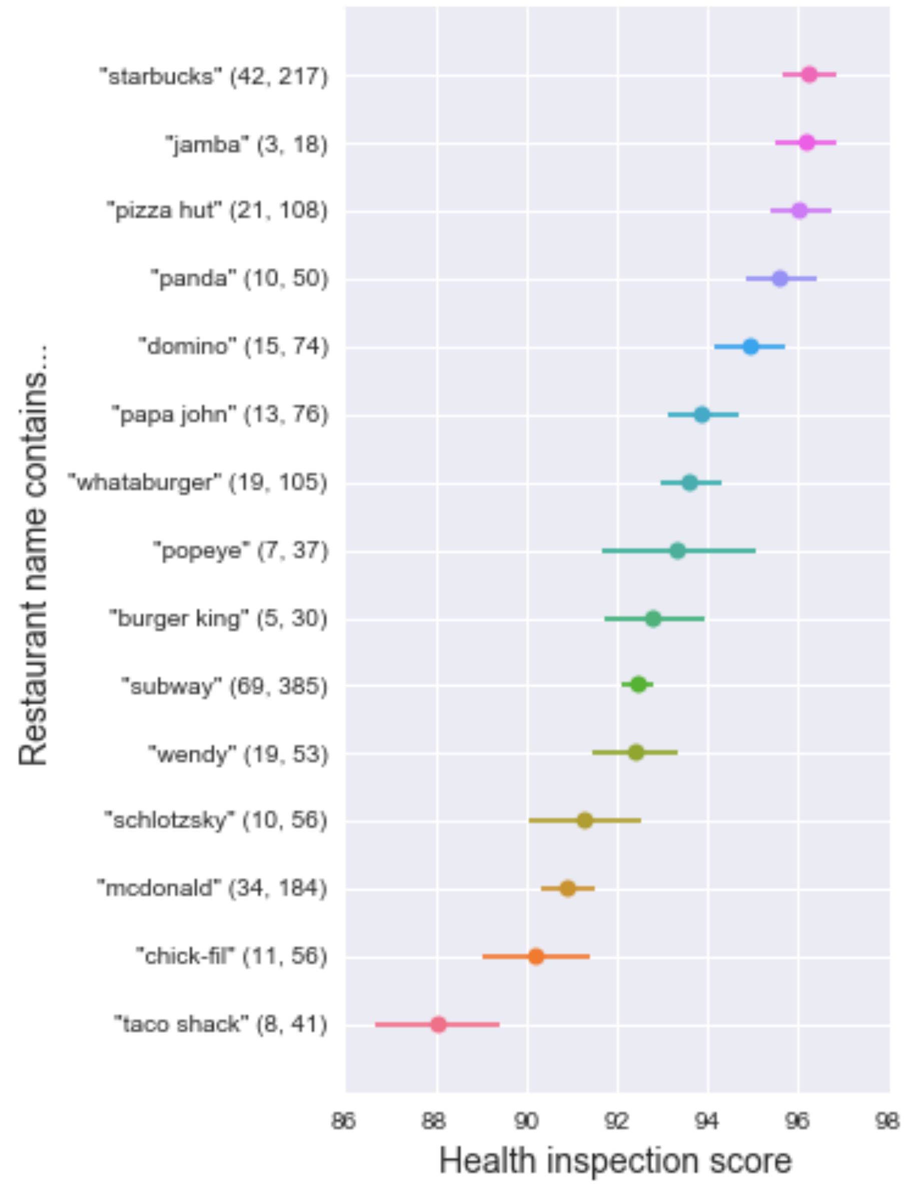
GENERAL ASSEMBLY

- Using open data from data.austintexas.gov
- Data set: City of Austin/Travis County Restaurant Inspections
  - 22,875 rows (inspections)
  - 4,774 establishments inspected (roughly 2x a year)

PREVIOUSLY...

GENERAL ASSEMBLY





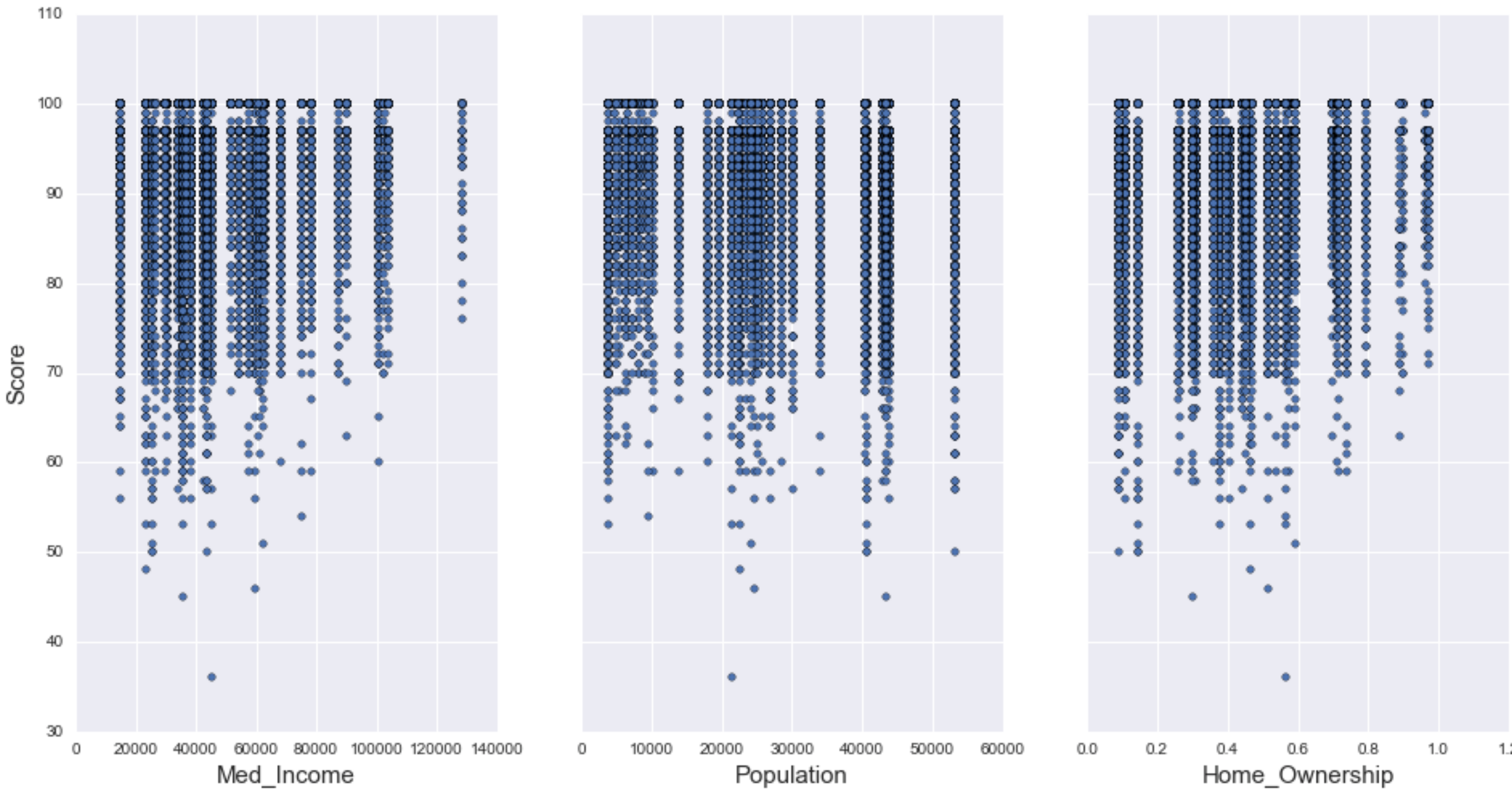
PROJECT OUTLINE

GENERAL ASSEMBLY

- ~~First and foremost: Clean and "enrich" data set~~
- ~~Part I: Exploratory Data Analysis~~
  - ~~Scores distribution~~
  - ~~Correlation among inspections~~
  - ~~Confidence intervals for scores by type of restaurant~~
- Part II: Machine Learning
  - A) [*Supervised*] Classification Algorithms: can we **predict** a restaurant's performance based on past scores?
  - B) [*Unsupervised*] PCA & k-means clustering: are there any **patterns** in names/addresses? Do they affect the score?
- Part III: Fancy stuff (time permitting)
  - Geocoding - visualizations: Heat maps by zip code / area
  - Linear regression: web scraping for yelp scores, add zip codes' median income - can I predict health inspection score?

ADDING NUMERICS

GENERAL ASSEMBLY



LINEAR REGRESSION

GENERAL ASSEMBLY

- *Attempted to predict score given numerics - terrible idea (2% of variance explained)*
- *Switched to "averaged" data - 36 rows, one per zip code of interest:*
  - *Score vs. Median Income: 40%*
  - *Score vs. Population: 17.4%, negative slope*
  - *Score vs. Home Ownership Percentage: 13.7%*
- *Multiple Linear Regression Model: 49.4% - Almost as good as a coin toss!*

- *No predictions allowed - more of a toy problem*

LOGISTIC REGRESSION

GENERAL ASSEMBLY

- *Features: City Areas (5)*
- *Target: Pass/Fail*
  - *Also: assigned letter grades (like in New York City)*
  - *Class Imbalance: only 1% of 18610 entries have failing scores*
    - *linear_model.LogisticRegression(class_weight='balanced') - **55%***
    - *Oversampling the minority class - **56%***
- *Changing the classification threshold:*

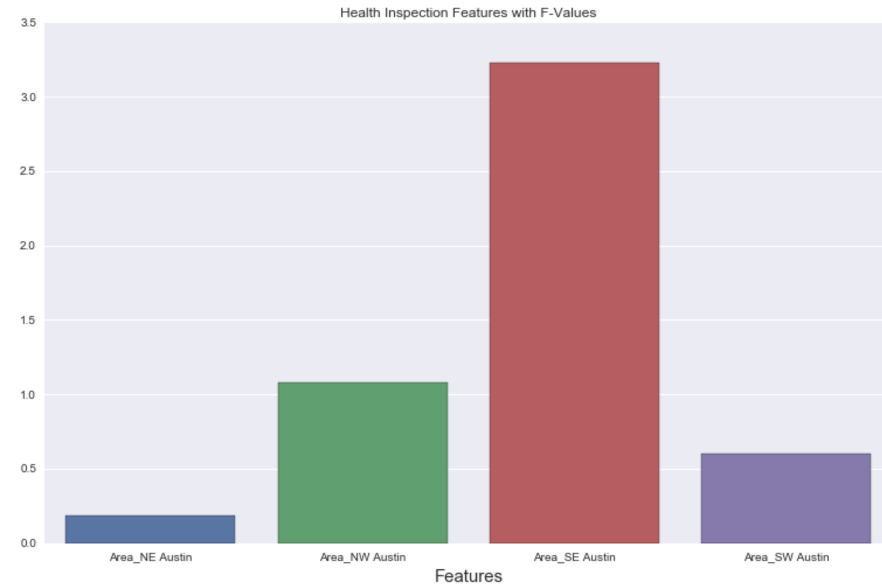
```
The number of restaurants predicted to pass inspection with logistic threshold = 0.5 is 13384
The number of restaurants predicted to pass inspection with logistic threshold = 0.6 is 11999
The number of restaurants predicted to pass inspection with logistic threshold = 0.7 is 5423
The number of restaurants predicted to pass inspection with logistic threshold = 0.8 is 0
The number of restaurants predicted to pass inspection with logistic threshold = 0.9 is 0
```

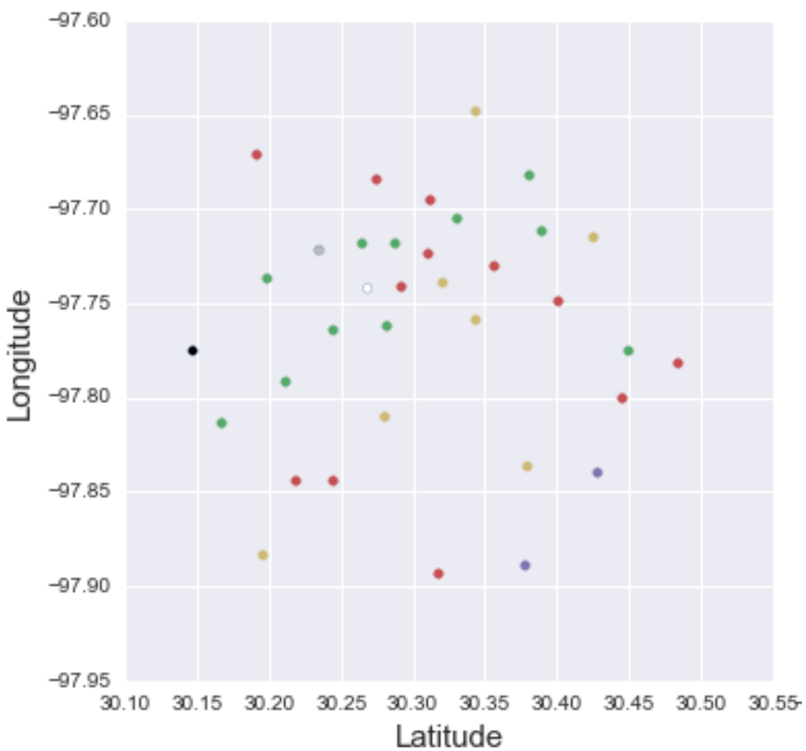
- *Switching to a different prediction target: Letter Grade A or (not A)*

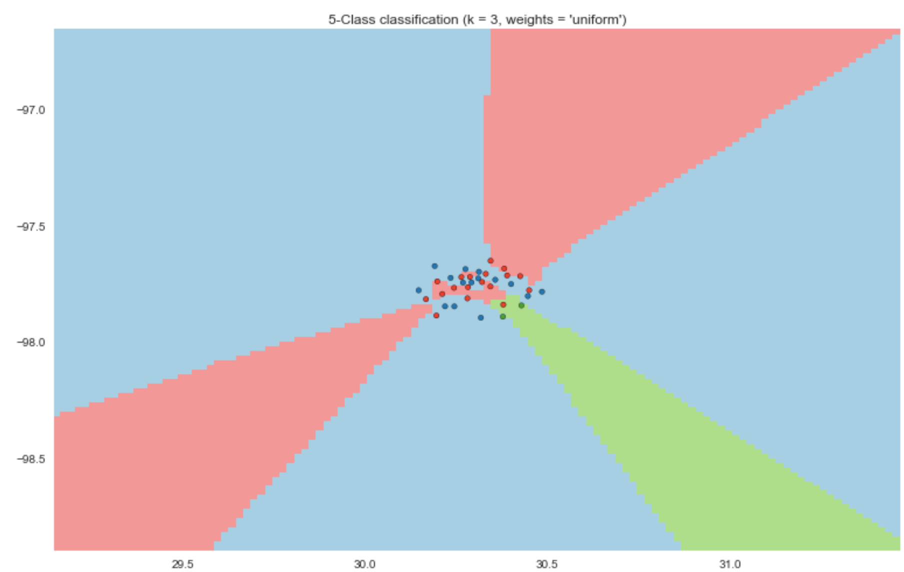
KNN ON AVERAGE SCORES

GENERAL ASSEMBLY

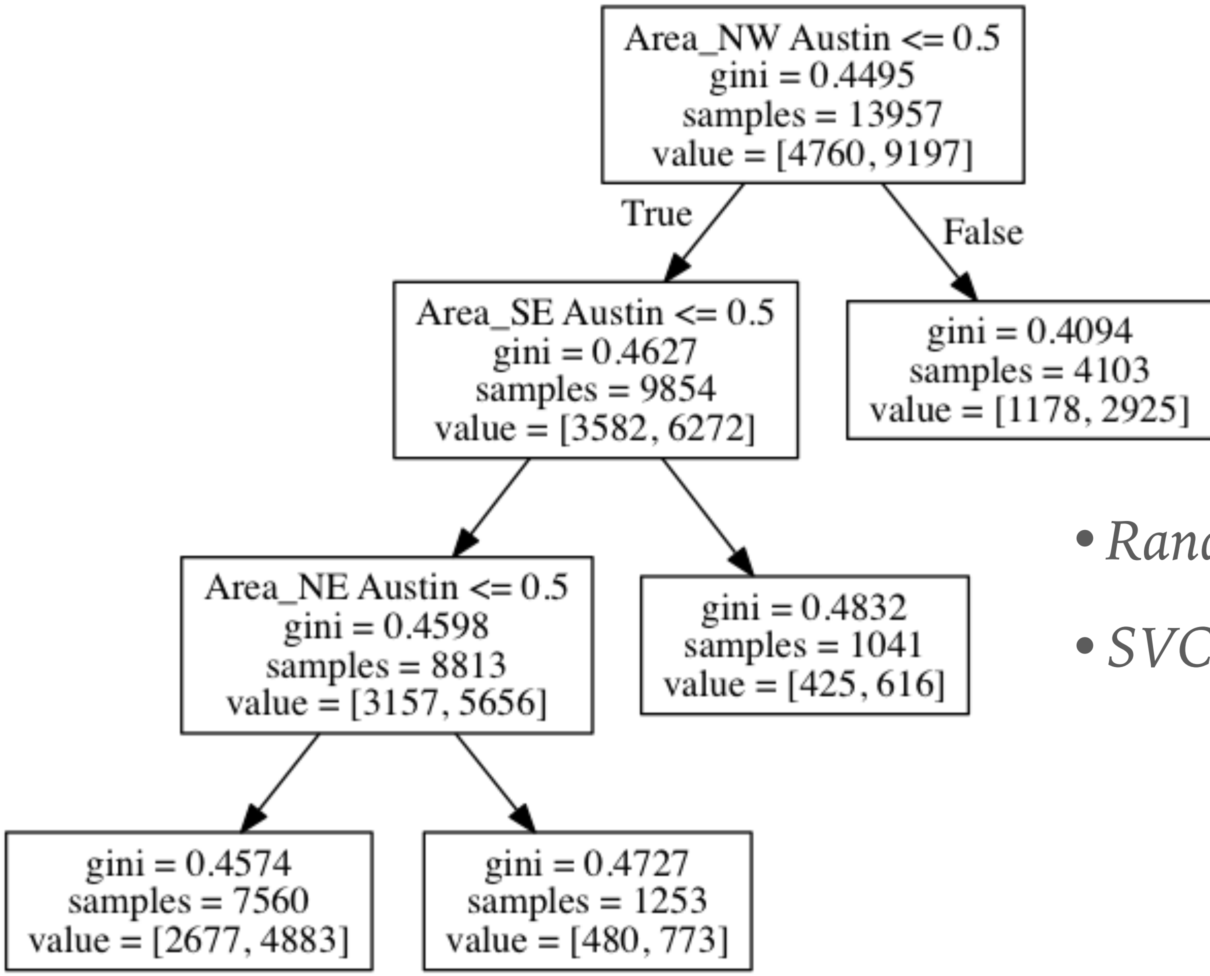
- *Score on holdout set: **85%!***
- *A recurring theme:*
- *SE/NW Austin seem to be "important"*







OTHER CLASSIFIERS



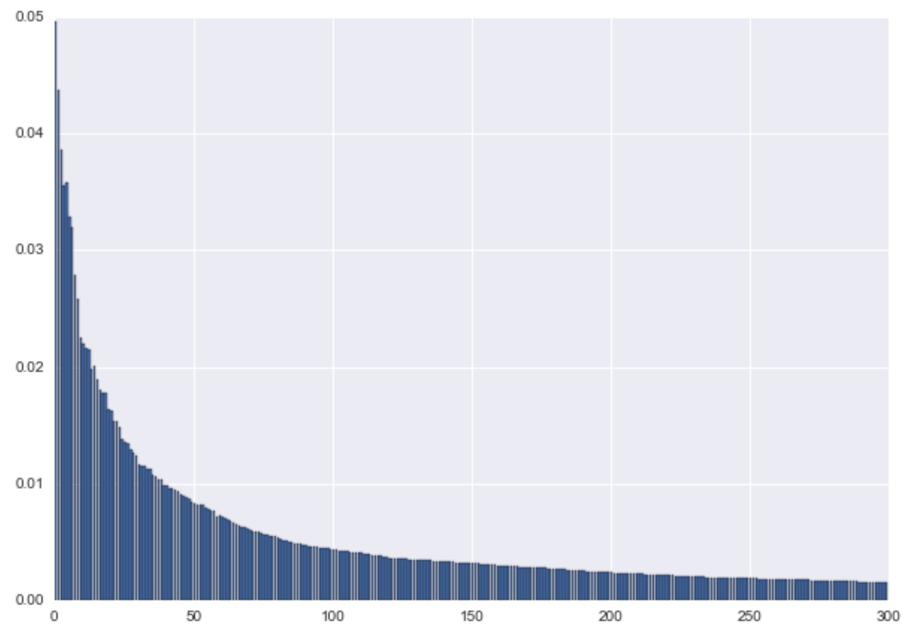
- *Random Forest with 250 trees: 65.8%*
- *SVC: 65.8%*

TEXT MINING FOR CLASSIFICATION

GENERAL ASSEMBLY

- *Bag-of-Words model for:*
  - *Restaurant Name (~3500 features): 66.8%*
  - *Street (~540 features): 61.9%*
  - *Using: CountVectorizer, Multinomial Naive Bayes*

- *Dimensionality Reduction: Truncated SVD*
  - *Similar score with 10% of the features!*



IMPORTANT FEATURES

GENERAL ASSEMBLY

Class: A

austin: -4.6089, the: -4.7837, elementary: -4.8022, food: -4.8064, mart: -4.9958, cafe: -4.9958, **pizza:** -5.0059, coffee: -5.1214, market: -5.1953, center: -5.2554, bar: -5.3193, school: -5.4331, food mart: -5.4449, **subway:** -5.5099, club: -5.5795, **bakery:** -5.6690, grill: -5.6739, **house:** -5.6990, catering: -5.7195, and: -5.7247, deli: -5.7247, **starbucks:** -5.7299, restaurant: -5.7892, foods: -5.8060, inn: -5.8289, of: -5.8347, wl: -5.8464, taco: -5.8523, child: -5.8823, la: -5.8823

Class: B

restaurant: -4.4523, cafe: -4.5059, **austin:** -4.6419, the: -4.8050, **bar:** -4.8518, market: -4.8884, pizza: -4.9658, la: -4.9658, food: -4.9725, grill: -5.1414, el: -5.1991, mart: -5.2876, **taco:** -5.3448, house: -5.3746, **subway:** -5.4924, and: -5.4924, mexican: -5.5271, in: -5.5271, bakery: -5.5877, food mart: -5.6930, coffee: -5.8108, deli: -5.8265, taqueria: -5.8265, grocery: -5.8588, inn: -5.8921, kitchen: -5.9092, club: -5.9092, stop: -5.9443, star: -5.9807, thai: -5.9993

Class: C

restaurant: -3.9294, cafe: -4.3385, **grill:** -4.5645, the: -4.5930, bar: -4.5930, la: -4.5930, market: -4.8381, taqueria: -4.8568, el: -4.8758, house: -4.9559, mexican: -5.0204, **austin:** -5.0429, thai: -5.2435, and: -5.3006, taco: -5.4257, food: -5.4596, mexican restaurant: -5.5688, in: -5.6488, bakery: -5.6488, cuisine: -5.6488, pizza: -5.6913, kitchen: -5.7358, pho: -5.7358, los: -5.7358, **sushi:** -5.7823, bar grill: -5.7823, meat: -5.8310, meat market: -5.8310, shop: -5.9363, china: -5.9935

Class: F

la: -3.7515, restaurant: -3.8002, market: -4.3107, **taqueria:** -4.3107, el: -4.3107, **austin:** -4.3977, grill: -4.5981, thai: -4.7158, meat: -4.8491, bar: -4.8491, meat market: -4.8491, bakery: -5.1850, los: -5.1850, **la michoacana:** -5.1850, kitchen: -5.1850, **michoacana:** -5.1850, cafe: -5.4077, cuisine: -5.4077, house: -5.4077, lion: -5.6945, taco: -5.6945, mexican: -5.6945, **taqueria los:** -5.6945, del: -5.6945, el meson: -5.6945, and: -5.6945, shop: -5.6945, buffet: -5.6945, vietnamese: -5.6945, la catracha: -5.6945

HOT OFF THE PRESS!

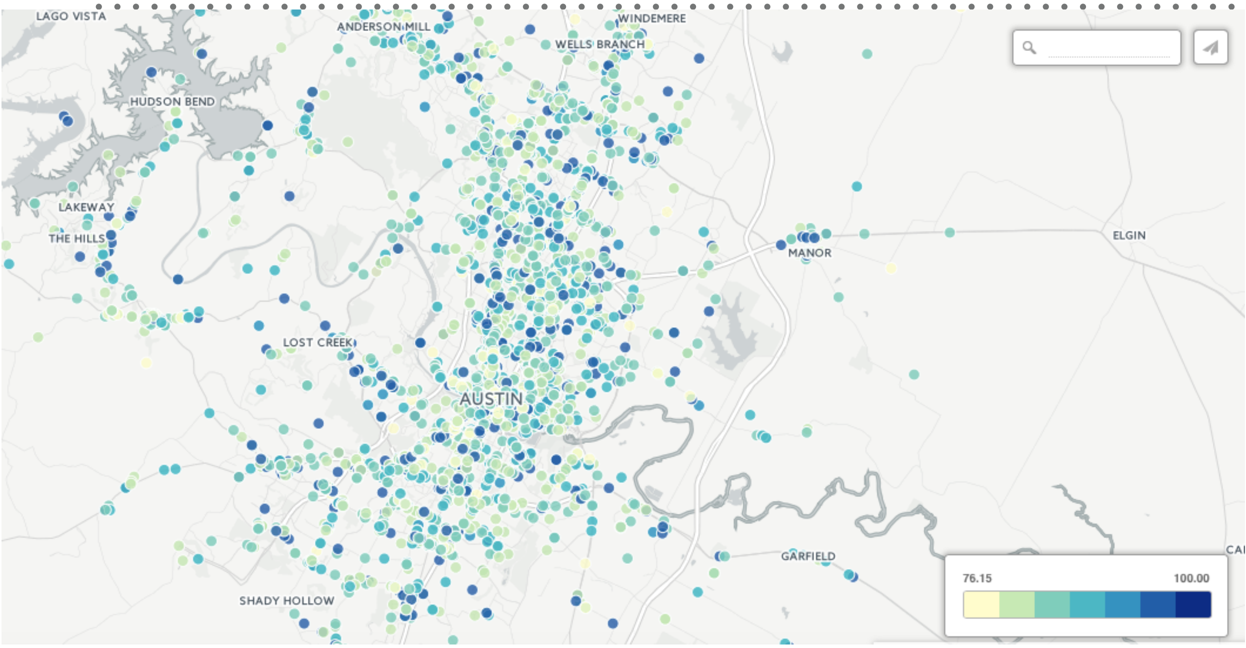
GENERAL ASSEMBLY

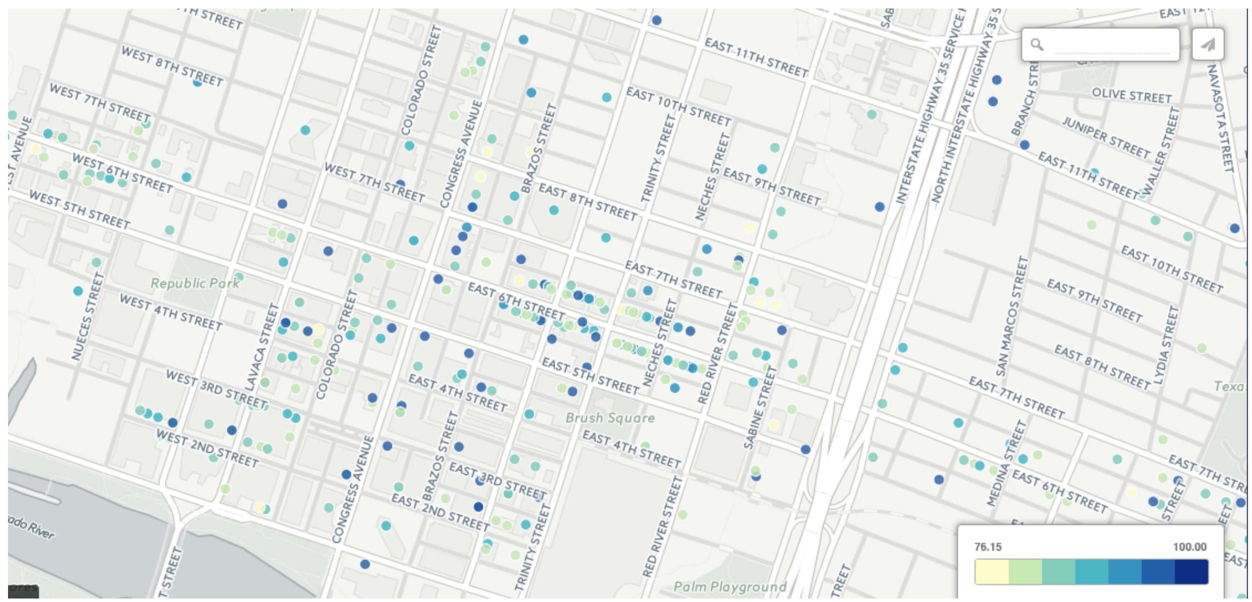
- *Filter out words with len < 3:*

A: catering grill **club subway** market **center coffee** pizza cafe austin
B: **subway** mexican **taco** house grill **pizza** market austin cafe restaurant
C: taco **thai** house austin mexican **taqueria** market grill cafe restaurant
F: kitchen bakery **michoacana** meat **thai** market austin grill **taqueria** restaurant

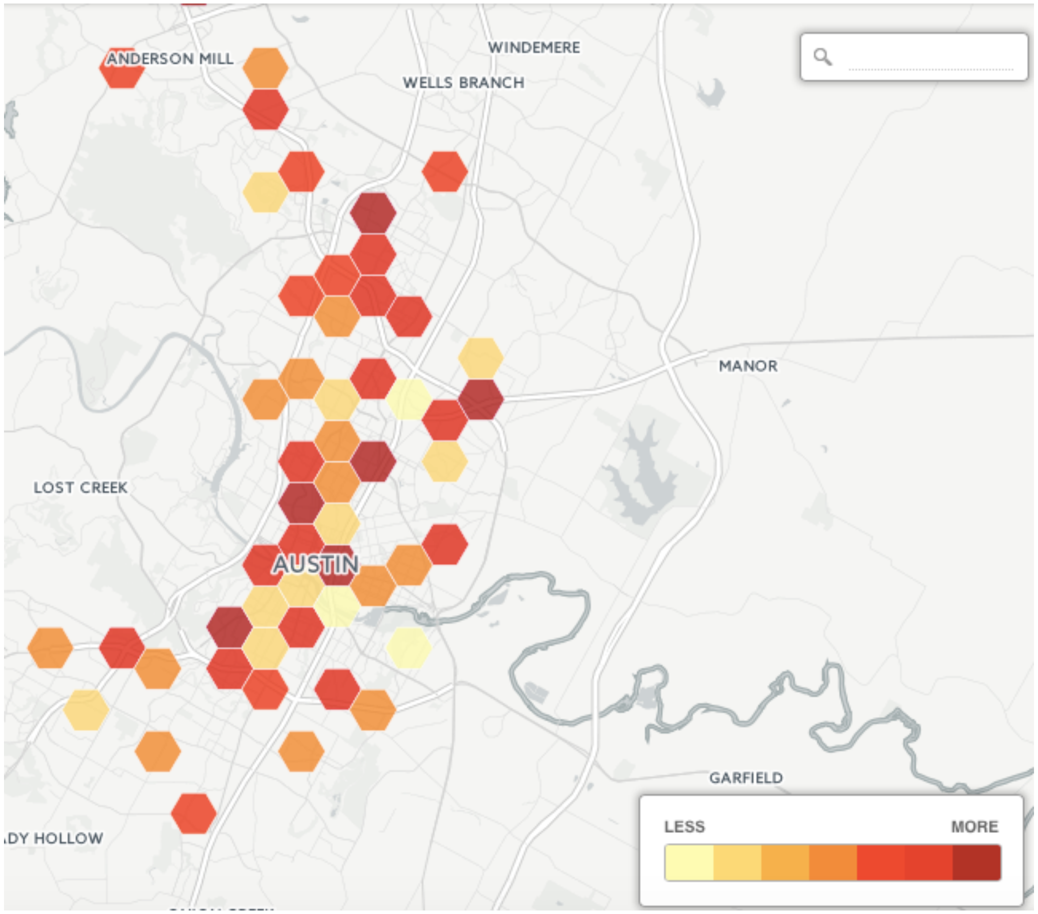
A: **slaughter anderson mopac caves capital** congress william parmer lamar blvd
B: oltorf riverside springs **capital anderson** parmer congress william blvd lamar
C: slaughter parmer stassney oltorf riverside springs william congress blvd lamar
F: **cesar** parmer springs congress martin **anderson** riverside blvd oltorf lamar

GEOCODING & MAPPING





GEOCODING & MAPPING



CONCLUSION – NEXT STEPS

GENERAL ASSEMBLY

- *Austin is very diverse in its neighborhoods*
- *Managed to predict the letter grade of an ungraded restaurant with 60 - 67% accuracy by only knowing the restaurant's area of town and name.*
- *Next: Improve NB Classifier; Try unsupervised learning; clustering of restaurants by name/street*
- *Fork out the $ for good geocoding of all 19,000 data points (less than $20)*
  - *Add "neighborhood" data as a better(?) feature than "Area" - "Street"*

- *Publish: https://github.com/nvergos, CartoDB (nvergos.cartodb.com)*
- *Do more **Data Science!** (Note to self: next data set should have some non categorical variables)*