NEW COFFEE SHOP LOCATION

Steven Asten

CONTENTS

PROBLEM

- Find an available storefront property in DC area which would be suitable for a new coffee shop.

HYPOTHESES

1. If I can find where people are coming from when they visit coffee shops, then I can look for a geographical area with a large presence of these features, and this location would be suitable for a new coffee shop.
2. If I can predict characteristics of a coffee shop-goer (including age, income, means of transportation), then I can find a geographical area with the most of this type of person, and this would be a suitable location for the new coffee shop.

DATA DICTIONARY - YELP API

Data ID	Data Type	Data Description
Business_id	String	unique id
Name	String	Business' name
Neighborhood	String	Neighborhood's name
Address	String	Full address of business
City	String	City
State	String	2-char state abbreviation
Postal code	String	5-char postal code
Latitude	Float	Latitude
Longitude	Float	Longitude
Categories	Array	List of categories

DATA DICTIONARY - FOURSQUARE

Data ID	Data Type	Data Description
user_id	Integer	Unique ID
utc_time	Datetime	Full datetime timestamp
latitude	Float	Latitude of checkin
longitude	Float	Longitude of checkin
venue_category	String	Category description
country_code	String	Abbreviated country

DATA DICTIONARY – 2016 ACS DATASET

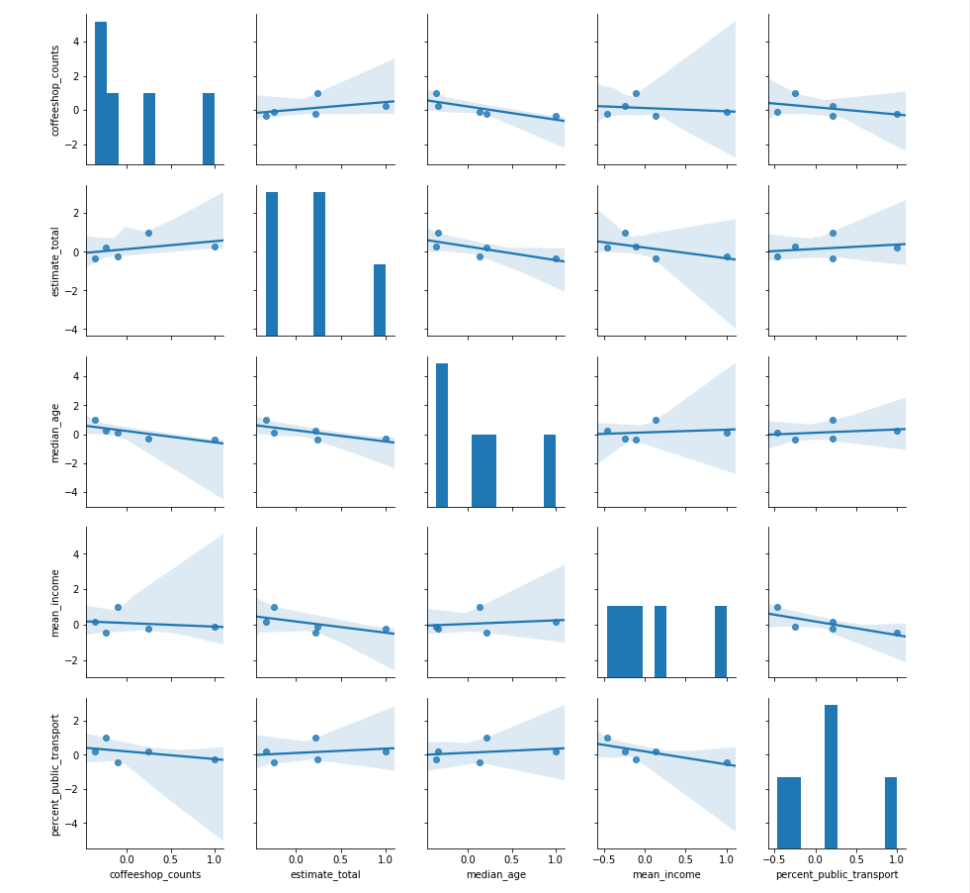
Data ID	Data Type	Data Description
zip_code	String	Zip code
population	Integer	Total Population of zip code
median_age	Integer	Median age of zip code
mean_income	Float	Mean income of zip code
percent_public_transport	Float	Percentage of people who use public transportation as a primary means of transport

EXPLORATORY DATA ANALYSIS

- Foursquare dataset contained Latitude and Longitude, no other location data. Reverse geocoding 3-million records unrealistic.
  - Python package "reverse_geocoder" returns county
  - Zillow Neighborhood Shapefile returns name of neighborhood
  - Convert neighborhood names to zip code
- Yelp dataset missing data
  - Yelp API 50 results per query, 5000 queries per day

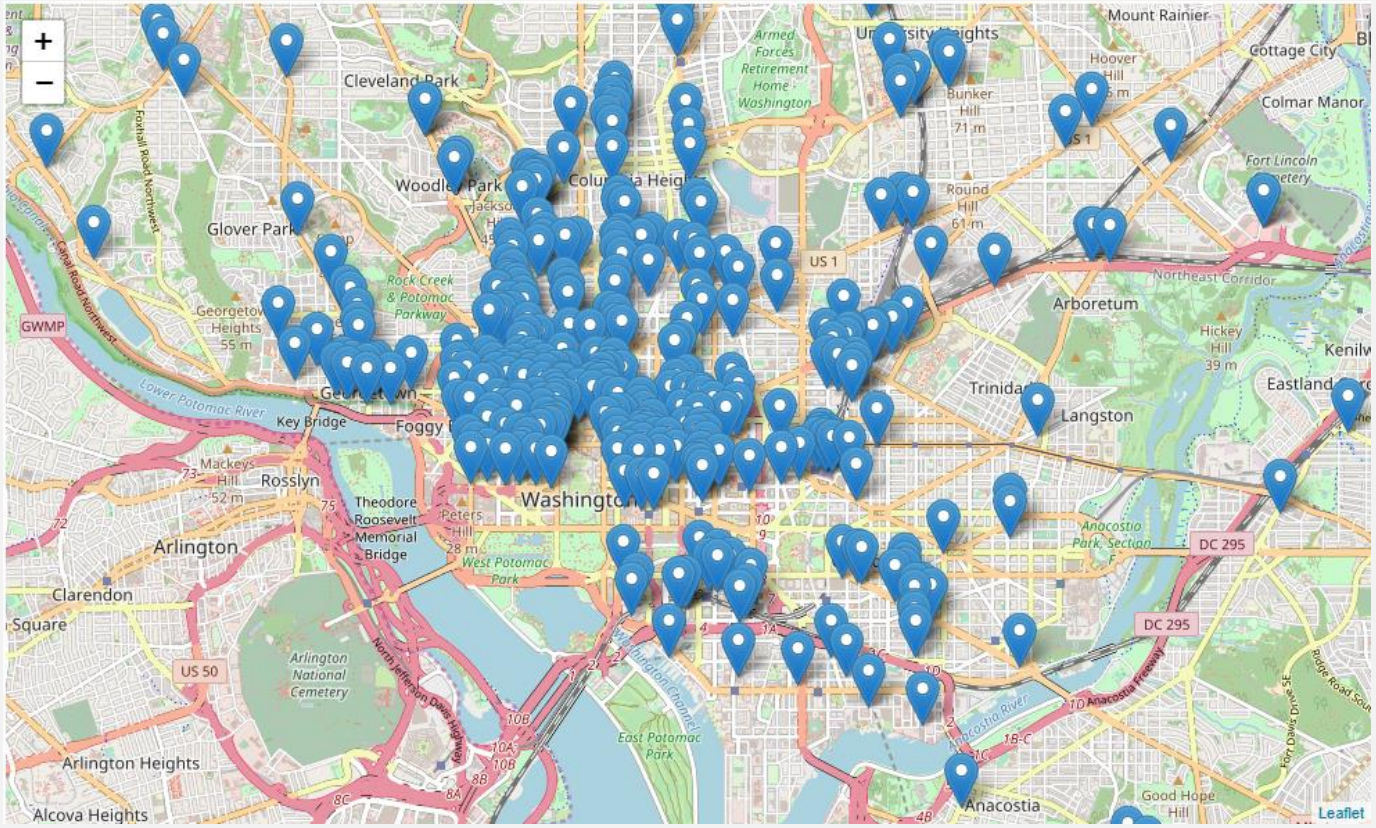
VISUALIZATIONS I

Correlations between count of coffee shop visits and zip code demographic features



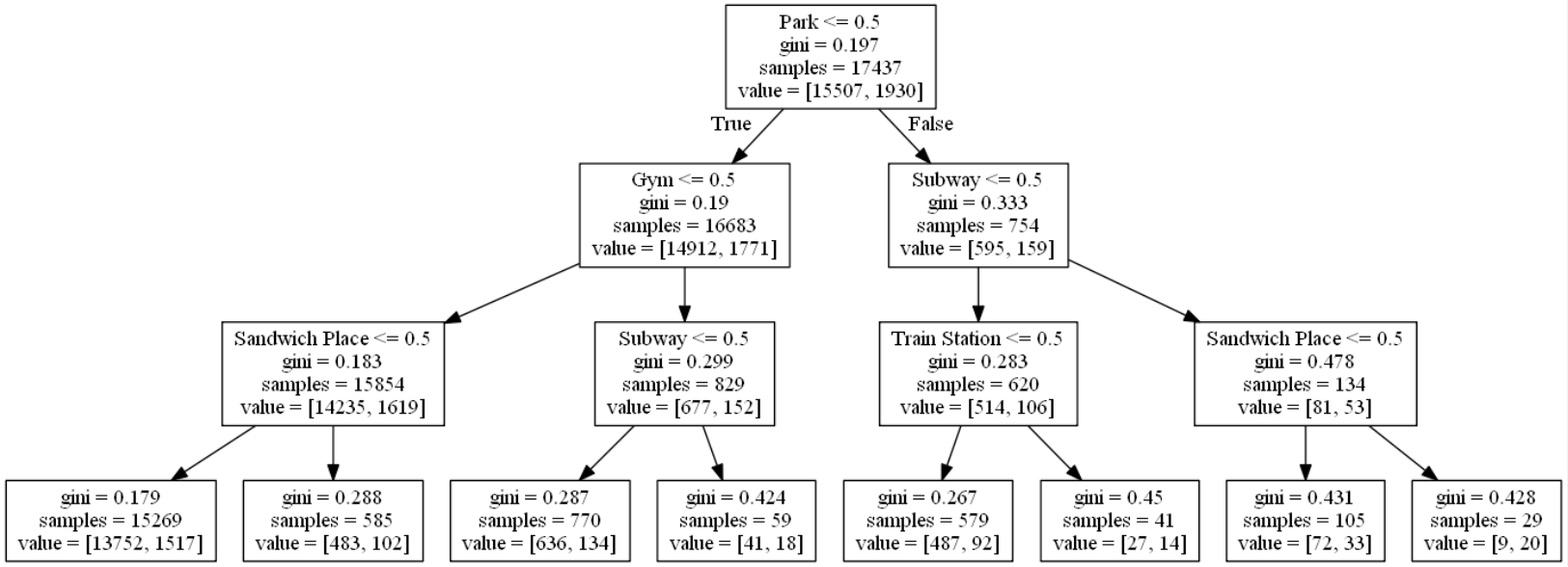
VISUALIZATIONS 2

All coffee shops in DC plotted on interactive Folium map



MACHINE LEARNING 1

1. Model 1: Decision Tree. Transposed top 23 categories to columns with binary values and used a Decision Tree to predict where people are coming from when they go to coffee shops.



	feature	importance
11	Park	0.290257
1	Subway	0.238393
13	Sandwich Place	0.233781
7	Gym	0.190744
6	Train Station	0.046825
20	Road	0.000000
19	Baseball Stadium	0.000000
18	Neighborhood	0.000000
17	Pub	0.000000
16	Pizza Place	0.000000

```
Accuracy:  0.889908256881
```

MACHINE LEARNING 2

- Model 2: Random Forest

	feature	importance
1	Subway	0.072851
0	Office	0.065243
2	American Restaurant	0.058225
8	Grocery Store	0.057380
9	Home (private)	0.054446
11	Park	0.053394
4	Government Building	0.053389
6	Train Station	0.051777
12	Mexican Restaurant	0.050790
10	Residential Building (Apartment / Condo)	0.050631
13	Sandwich Place	0.048432
5	Hotel	0.045965
7	Gym	0.044408
18	Neighborhood	0.043520

```
accuracy: 0.889220183486
```

- Model 3: Linear Regression. I found that Median Age and Percentage of people who use Public Transport were most correlated with coffee shop check-ins, so I used these two features in my linear regression model.
- As median age increases by 1 year, number of coffee shop check-ins decreases by 21.

MACHINE LEARNING PERFORMANCE

Models run on Foursquare Dataset	Model	Accuracy
	Decision Tree	Baseline ≈ 88% Modelled ≈ 88%
	Random Forest	Baseline ≈ 88% Modelled ≈ 89%
	Gradient Boosting	Baseline ≈ 88% Modelled ≈ 89%

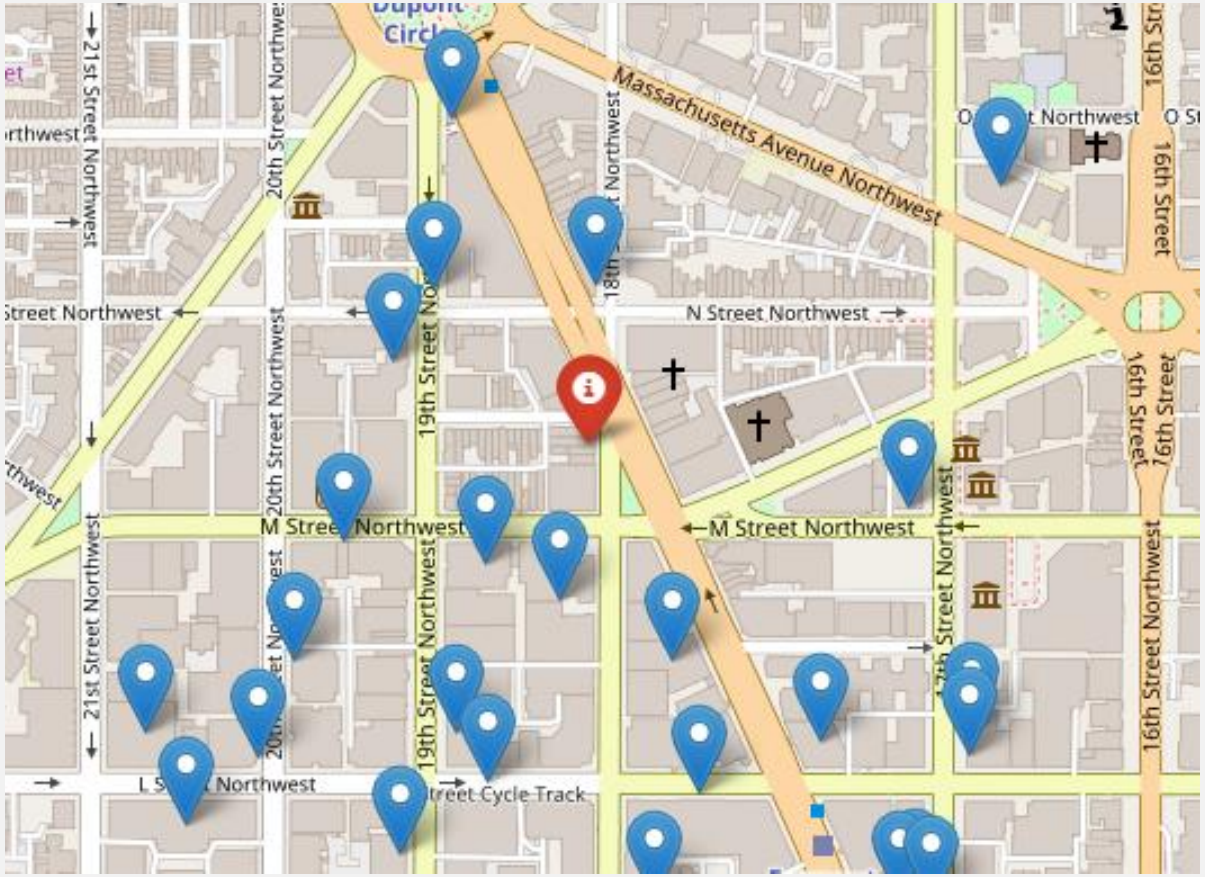
Models run on ACS dataset	Model	RMSE
	Linear Regression	Baseline: 243.6 Modelled: 117.9

CONCLUSION 1

- Hypothesis 1: People who visit coffee shops also visit parks and gyms, and they use public transportation. They do not like sandwich shops.
- Hypothesis 2: Lower median age results in a higher number of coffee shop visits.

	zip_code	sandwich	park	coffee	median_age	percent_public_transport
0	20001	54	8	43	30.6	0.353349
4	20005	36	2	33	33.8	0.264241
5	20006	34	5	31	20.6	0.287554
8	20009	30	4	41	32.3	0.433303
1	20002	28	5	33	34.0	0.361972
3	20004	23	1	20	41.0	0.312452
2	20003	22	9	20	34.8	0.401540
18	20036	21	1	27	32.7	0.363870
16	20024	21	7	15	38.2	0.441809
6	20007	18	6	20	33.2	0.233151

CONCLUSION 2



APPENDIX

- GitHub Repo: https://git.generalassemb.ly/astensteven/project-final
- Yelp API Docs: https://www.yelp.com/developers/documentation/v3
- Foursquare Data: https://sites.google.com/site/yangdingqi/home/foursquare-dataset
- ACS Data: https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml
- Zillow Neighborhood Shapefiles: https://www.zillow.com/howto/api/neighborhood-boundaries.htm
- Reverse Geocoder: https://github.com/thampiman/reverse-geocoder
- Property Listing: http://www.cityfeet.com/cont/listing/retail-space-for-lease/1214-18th-st-nw-washington-dc-20036/CS4135317