

3421 Spring 2023

Final Project: National housing price determinants in the U.S



Question: When evaluating housing values, school quality, crime statistics, neighborhood house values, and a variety of other factors are typically used. Can house values be predicted based independently of the common factors? Could these factors influence gentrification affecting some neighborhoods and not others? A better understanding of non-traditional factors may help address this problem.

Algorithms:

- Linear Regression
- Support Vector Regression
- Decision tree
- Random Forest

Data:

- Zillow API
- U.S. airport dataset
- DollarTree locations
- restaurant location data from Yelp
- U.S. census data for median household income In [3]:

Objective #1: Determine correlation between local infrastructure, institutions and retail outlets to housing prices.

Objective #2: Determine methods to collect data containing location information from a variety of data sources and combine into a common data repository for analysis.

Example hypotheses:

- Is housing price correlated to proximity to an airport?
- Is housing price correlated to proximity to a dollar store?
- Is housing price correlated to census tract household income?
- Is housing price correlated to relative proximity to an inexpensive restaurant vs expensive restaurant?

To make this determination, I first selected criteria so the houses were of a common size and age. The criteria I selected (arbitrarily) by searching on the Zillow.com website and filtering on:

- 2250-2750sf single family home
- 3+ bedrooms
- 3+ baths
- 1/4-1/2 acre lot
- built 1998-2010
- located in the contiguous U.S.
- currently for sale

I used a chrome plugin from Bardeen (<https://www.bardeen.ai/>) to allow me to scrape some data from properties that were returned from an interactive Zillow search. This plugin populates a Google sheet with core data. Of particular note is the URL in Column F. The last segment of the URL is the zpid for each property. This is the primary key to the Zillow database. The Zillow website limits the number of properties returned in each search so successive searches were performed by year on construction to accumulate a dataset of sufficient size.

A1	fx Location								
	A	B	C	D	E	F	G	H	I
1	Location	Price	Sq. ft.	Ba	Br	Link			
2	195 Shoreline Pkwy, Tega Cay, SC 29708	\$650,000	2,699 sqft	3 ba	4 bds	https://www.zillow.com/homedetails/195-Shoreline-Pkwy-Tega-Ca			
3	6024 Courtland Dr, Erie, PA 16509	\$399,000	2,637 sqft	4 ba	4 bds	https://www.zillow.com/homedetails/6024-Courtland-Dr-Erie-PA-1			
4	3936 Long Branch Ln, Apopka, FL 32712	\$499,900	2,586 sqft	3 ba	4 bds	https://www.zillow.com/homedetails/3936-Long-Branch-Ln-Apopk			
5	3933 Fox Meadow Ln, Hope Mills, NC 28348	\$300,000	2,636 sqft	3 ba	4 bds	https://www.zillow.com/homedetails/3933-Fox-Meadow-Ln-Hope-			
6	788 Locks Way, Martinez, GA 30907	\$375,000	2,706 sqft	3 ba	4 bds	https://www.zillow.com/homedetails/788-Locks-Way-Martinez-GA			
7	2254 Angler Ln, Chesapeake, VA 23323	\$549,000	2,626 sqft	3 ba	5 bds	https://www.zillow.com/homedetails/2254-Angler-Ln-Chesapeake			

This spreadsheet with about 1500 rows was then exported to a CSV stored locally as 'propertyAddresses.csv'.

To access the full data record for each Zillow zpid, an API must be used. There are no longer any free methods to acquire this data so paid access is necessary. A third-party API from RapidAPI was selected.

The API key is obtained via the RapidAPI.com signup process.

In []:

Interim results were saved as both a csv and a pickle file as pulling this data successive times may lead to exceeding the monthly quota allowed under the license and therefore incur additional charges. Also of note, various licensing levels have differing data requests allowed per second. My license allows two requests per second. Exceeding the licensed threshold causes an error so the code must invoke a wait state between requests.

In [42]:

Out[42]:

	listingProvider	zpid	buildingPermits	propertyTaxRate	contact_recipients	zipcode	longitude	zestimateLowPercent	cityId	timeOnZillow	..	listingProvider.disclaimerText	listingProvider.sourceText	listing
0	None	71262940	None	0.63	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	29708	-81.007170	5	29454	8 days	..	NaN	NaN	NaN
0	None	67940085	None	1.90	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	16509	-80.063385	8	11335	8 days	..	NaN	NaN	NaN
0	None	71061935	None	1.09	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	32712	-81.534570	5	50741	8 days	..	NaN	NaN	NaN
0	None	70931874	None	1.12	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	28348	-78.931980	5	56956	8 days	..	NaN	NaN	NaN
0	None	83348915	None	1.01	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	30907	-82.073360	5	19223	8 days	..	NaN	NaN	NaN
..
0	None	68098746	None	1.03	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	33993	NaN	None	30742	305 days	..	NaN	NaN	NaN
0	None	227223070	None	0.75	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	39120	-91.371920	None	12909	305 days	..	NaN	NaN	NaN
0	None	68335546	None	1.10	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	43446	-82.818520	None	12703	322 days	..	NaN	NaN	NaN
0	None	80265690	None	1.03	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	34420	-82.076470	12	3614	347 days	..	NaN	NaN	NaN
0	None	37631190	None	1.05	[{'agent_reason': 1, 'zpro': None, 'recent_sal...	21821	-75.947586	None	24334	585 days	..	NaN	NaN	NaN

1458 rows × 332 columns

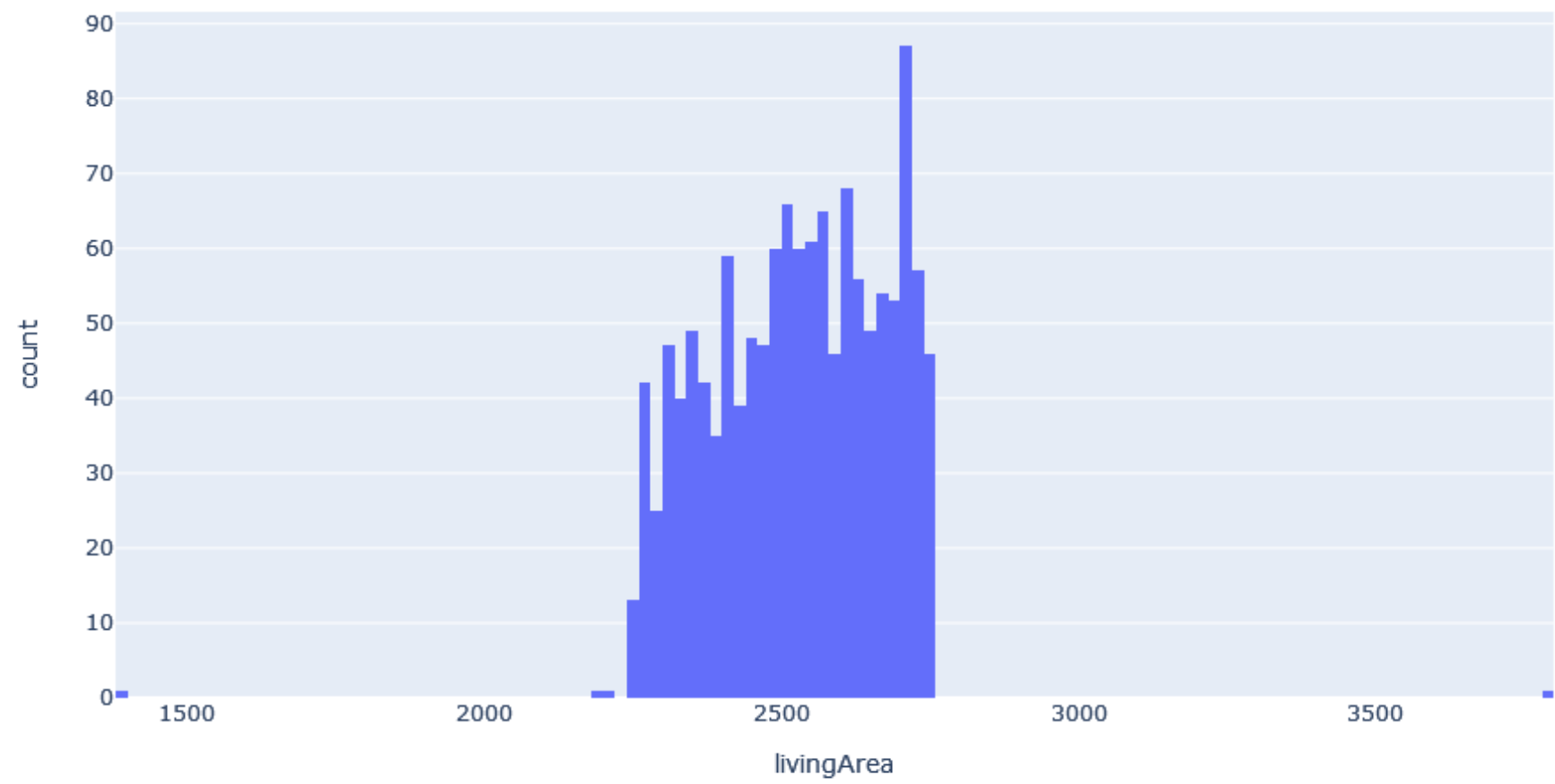
Of note is that latitude and longitude are contained in the zillow data. This will later be used for distance comparisons.

Clean the zillow data

There were some duplicate rows due to selection of overlapping regions while manually searching Zillow via a browser. The duplicate rows were removed.

Standard House Living Area Distribution

II

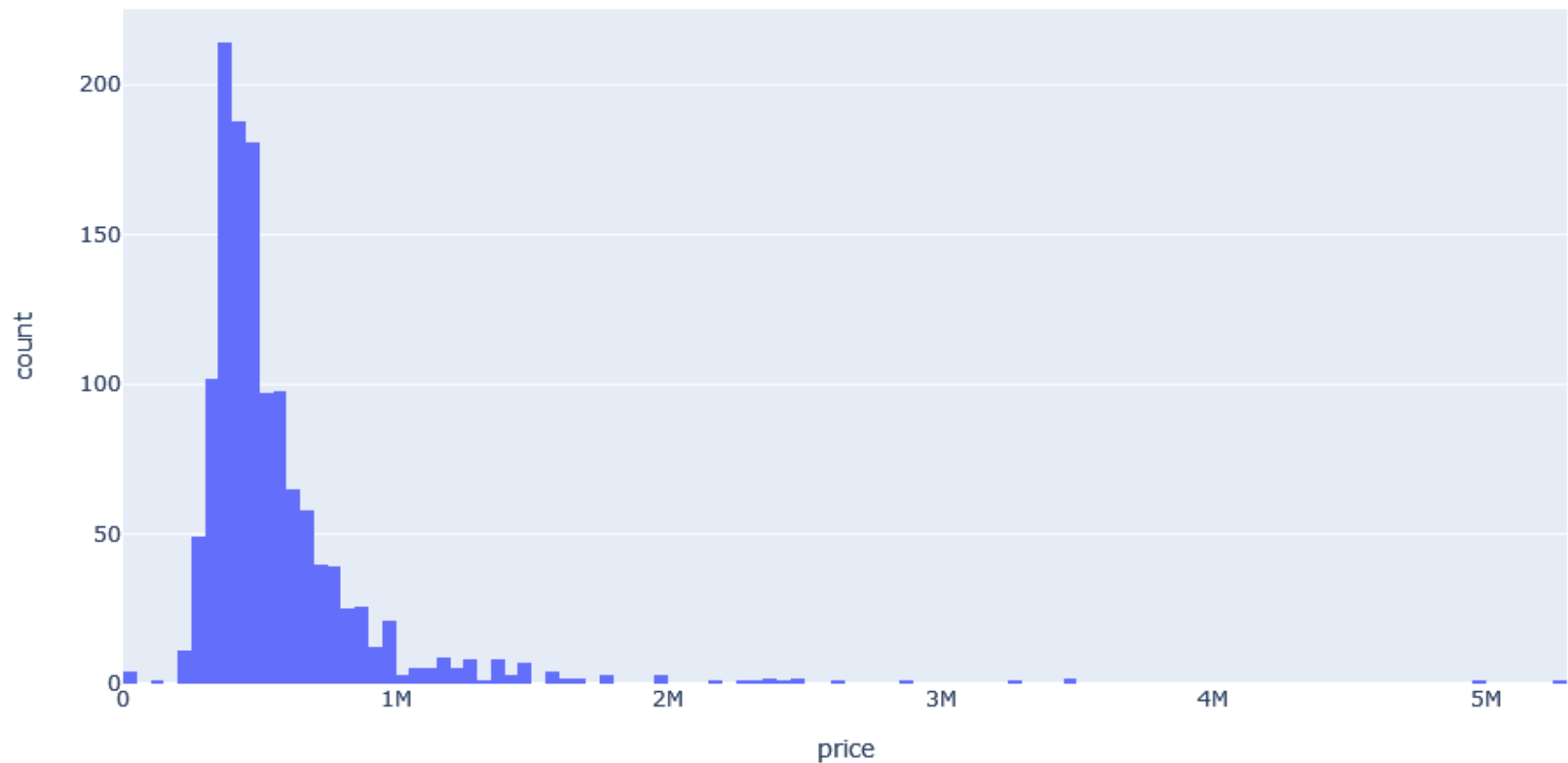


In spite of the original query being restricted to very specific square footage, there were a few outliers. These were removed.

Clean the zillow data (cont'd)

There were a few low outliers (near zero) in the asking price parameter:

Standard House Price Distribution

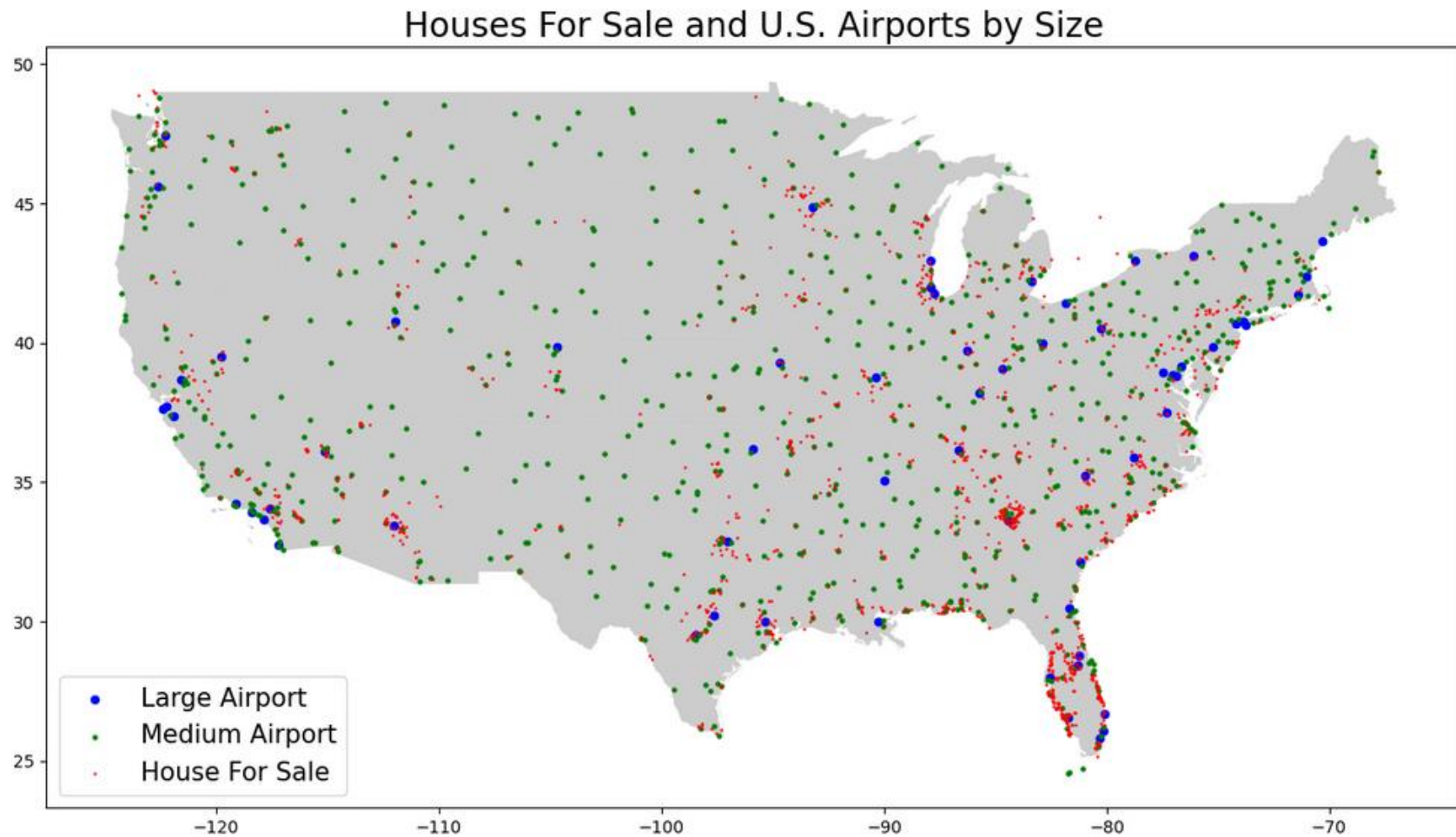


These were removed.

Airport Data

Gather U.S. airport data from <https://data.humdata.org/dataset/ourairports-usa>

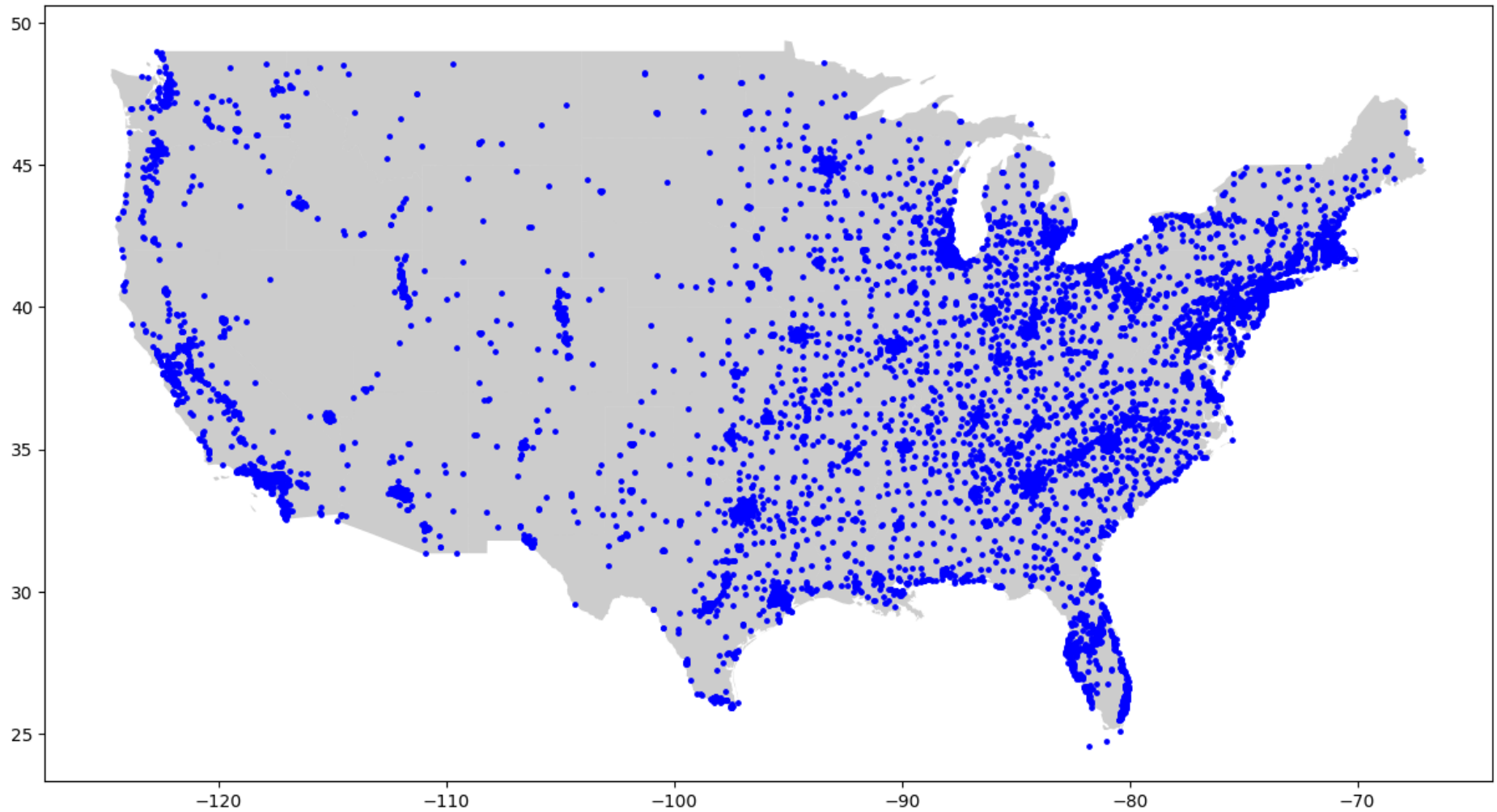
Visualization of US Airports by size w/house locations from the dataset:



Discount store (Dollar Tree) locations:

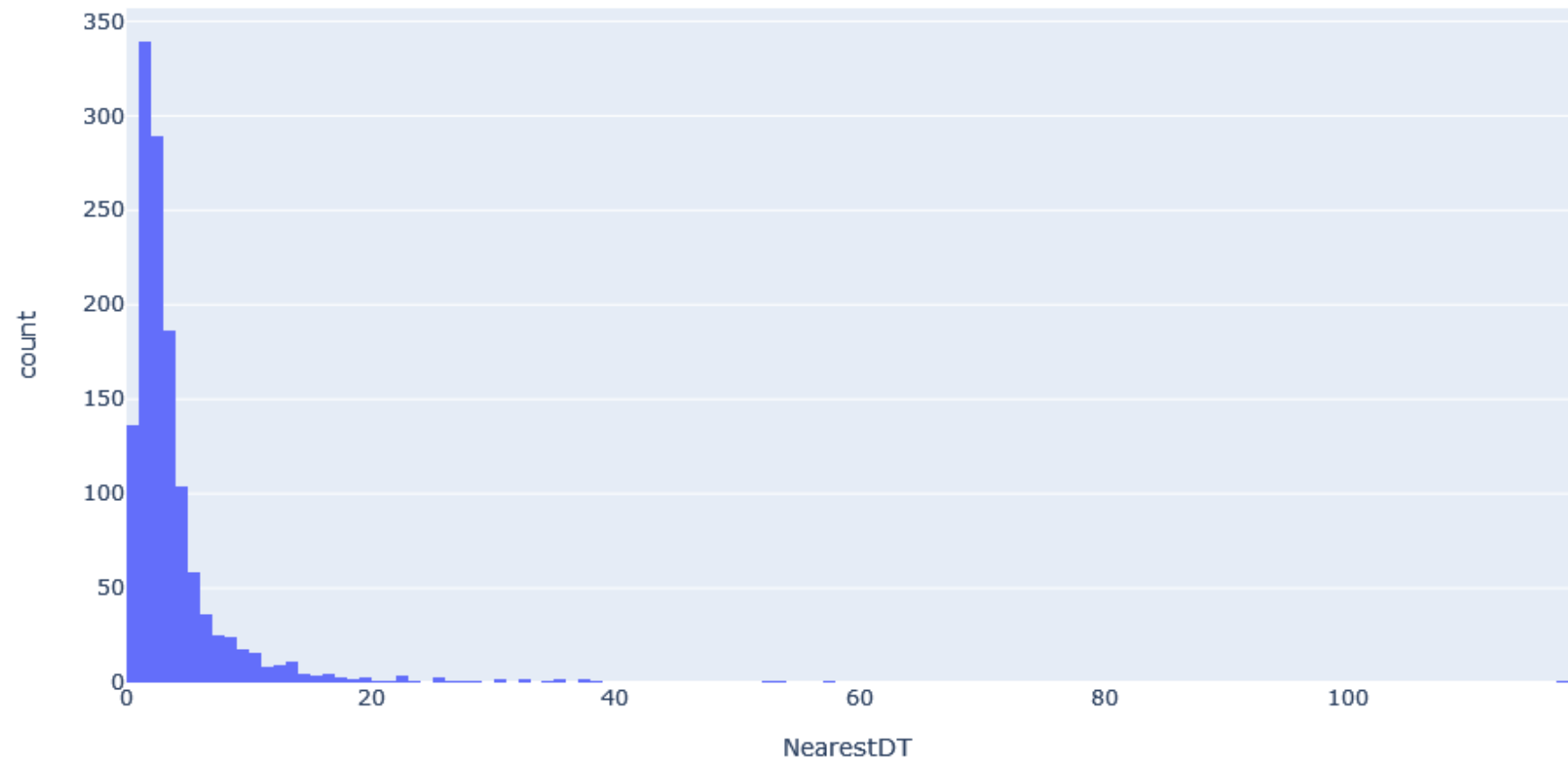
Location data was captured using python library BeautifulSoup and leveraging the Google Geocoding API to convert street addresses to latitude/longitude coordinates.

DollarTree Locations in the U.S.



Dollar tree proximity:

Proximity (miles) to Nearest Dollar Tree Distribution



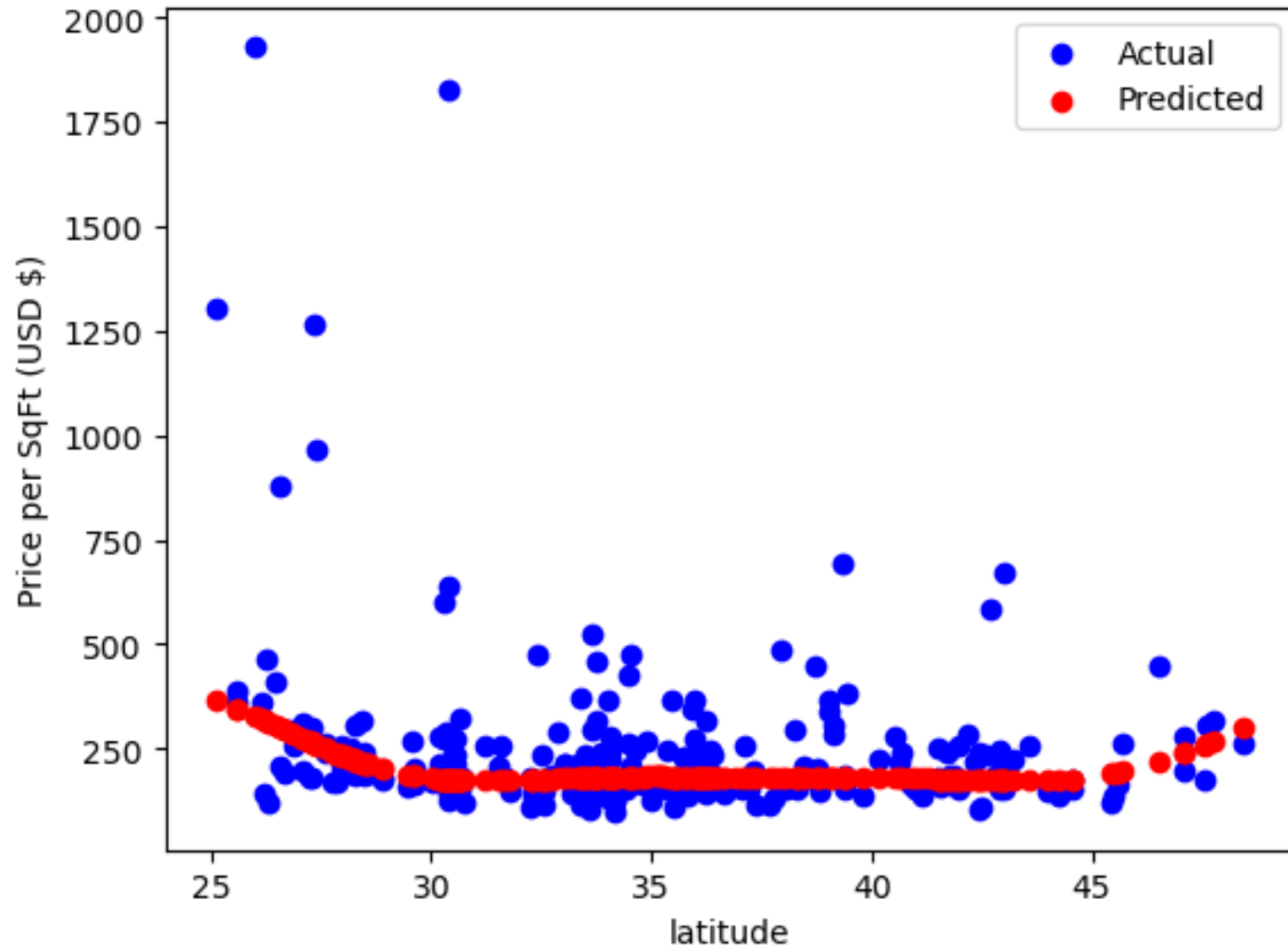
Correlation between house price and airport proximity:

Correlation Method	Airport Type	Correlation Value
Pearson	Large	-0.0966
Spearman	Large	-0.1935
Kendall	Large	-0.1310
Pearson	Medium	0.0343
Spearman	Medium	0.0026
Kendall	Medium	0.0014

There were no correlations of significance found between house price and proximity to airports.

House Price/SqFt by Latitude

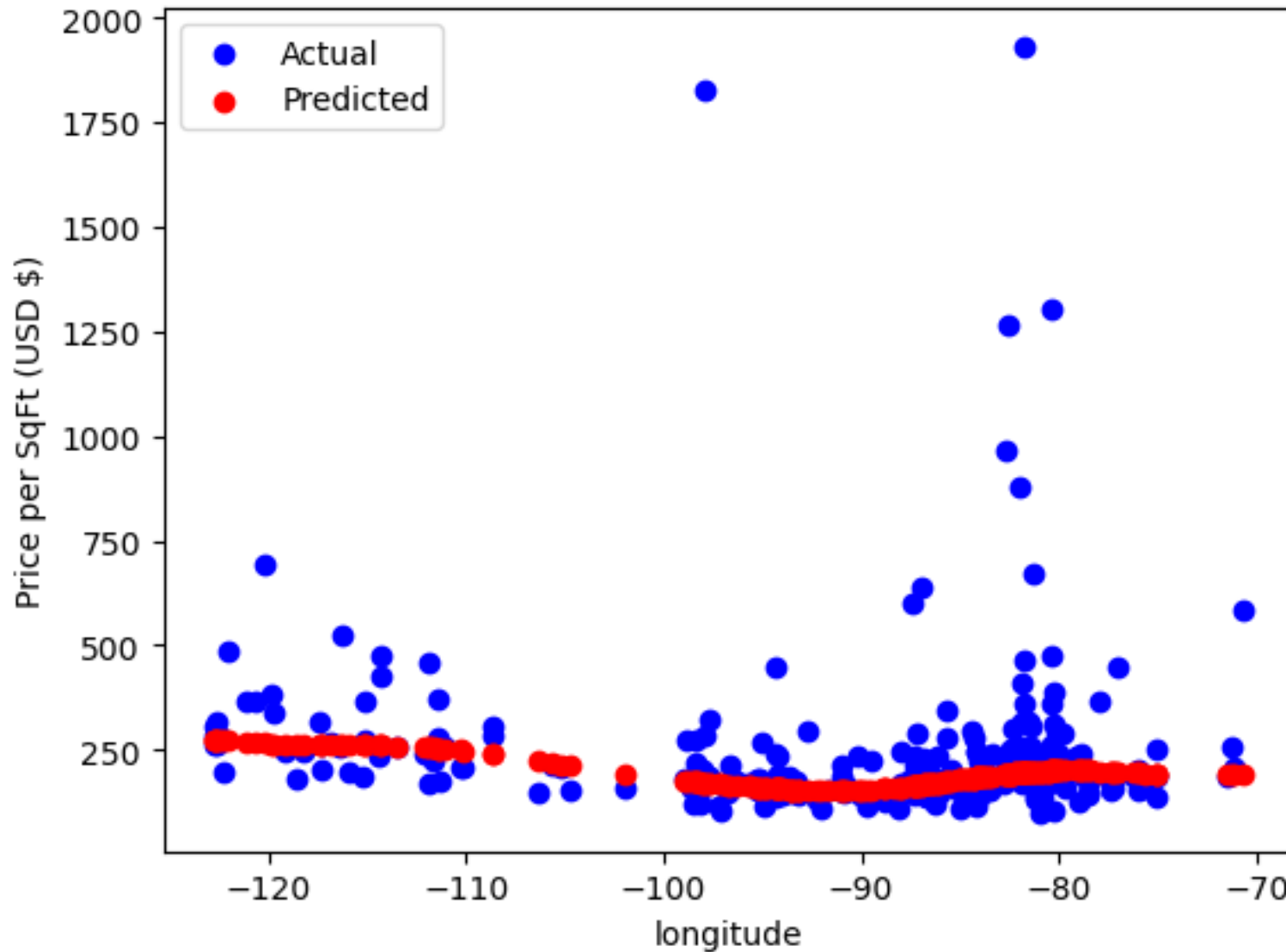
Support Vector Regression (SVR)



House prices in the far south and far north are elevated in comparison to the middle latitudes.

House Price/SqFt by Longitude

Support Vector Regression (SVR)



House prices west of the Rocky Mountains and in the approximate longitude of Miami are slightly higher than the middle of the country.

chatGPT4 prompt: I'd like to explore the possible correlation of single-family house prices based on proximity to certain types of retail outlets like Whole Foods.

To explore the possible correlation of single-family house prices based on proximity to certain types of retail outlets like Whole Foods, you can follow these steps:

Define your research question: You need to clearly state your research question, which in this case is, "Is there a correlation between the prices of single-family houses and their proximity to Whole Foods stores?"

Data collection: Collect data on single-family house prices and their locations. You may use real estate websites, APIs, or public records to gather this information. Additionally, collect data on the locations of Whole Foods stores.

Geocoding: Convert addresses of single-family houses and Whole Foods stores to geographic coordinates (latitude and longitude). There are various online geocoding services available, such as Google Maps API, OpenStreetMap's Nominatim, or Mapbox.

Calculate distances: Compute the distance from each single-family house to the nearest Whole Foods store. You can use various distance metrics such as Euclidean distance, Manhattan distance, or haversine distance for geographic coordinates.

Data analysis: Analyze the data to find any correlation between single-family house prices and their proximity to Whole Foods stores. You can use statistical methods like Pearson's correlation coefficient, Spearman's rank correlation coefficient, or Kendall's tau coefficient. You may also want to control for other variables that could influence house prices, such as the size, age, and condition of the property, local crime rates, and school quality.

Visualization: Create visual representations of the data to better understand any trends or patterns. You can use scatter plots, heatmaps, or choropleth maps to display the relationship between house prices and proximity to Whole Foods stores.

Interpretation: Evaluate the results of your analysis and draw conclusions based on the data. If you find a significant correlation between house prices and proximity to Whole Foods stores, you can hypothesize possible reasons behind this relationship. For example, Whole Foods stores may be located in more affluent neighborhoods, or their presence could signal a demand for healthier food options and drive up property values.

Report findings: Summarize your findings and share them with others. You can create a report or presentation outlining your methodology, data analysis, visualizations, and conclusions.

Remember that correlation does not imply causation, so even if you find a correlation between house prices and proximity to Whole Foods stores, it may not necessarily mean that the presence of a Whole Foods store directly causes an increase in property values. Further research would be required to determine any causal relationships.

chatGPT4 prompt: provide some examples of retail companies that may have a correlation to house prices

Various retail companies may have a correlation to house prices, either due to their presence in more affluent neighborhoods or by providing amenities that positively influence the desirability of an area. Some examples include:

Whole Foods Market: As an upscale grocery store, Whole Foods often locates its stores in affluent areas, and its presence may indicate a demand for healthier food options, which could contribute to higher property values.

Trader Joe's: This grocery store chain is known for its unique products and affordable prices. Like Whole Foods, its presence in a neighborhood might be a positive signal to potential homebuyers, leading to higher house prices.

Starbucks: As a popular coffee chain, the presence of a Starbucks may indicate a thriving, walkable community with a demand for leisure and social spaces. This can contribute to higher property values in the area.

Apple Store: Apple stores are often found in high-end shopping centers and affluent areas. Their presence may signify a wealthier demographic and contribute to higher house prices.

Target: This retail chain is known for offering a wide variety of products at affordable prices. The presence of a Target store can be a desirable amenity for potential homebuyers, potentially leading to higher house prices in the surrounding area.

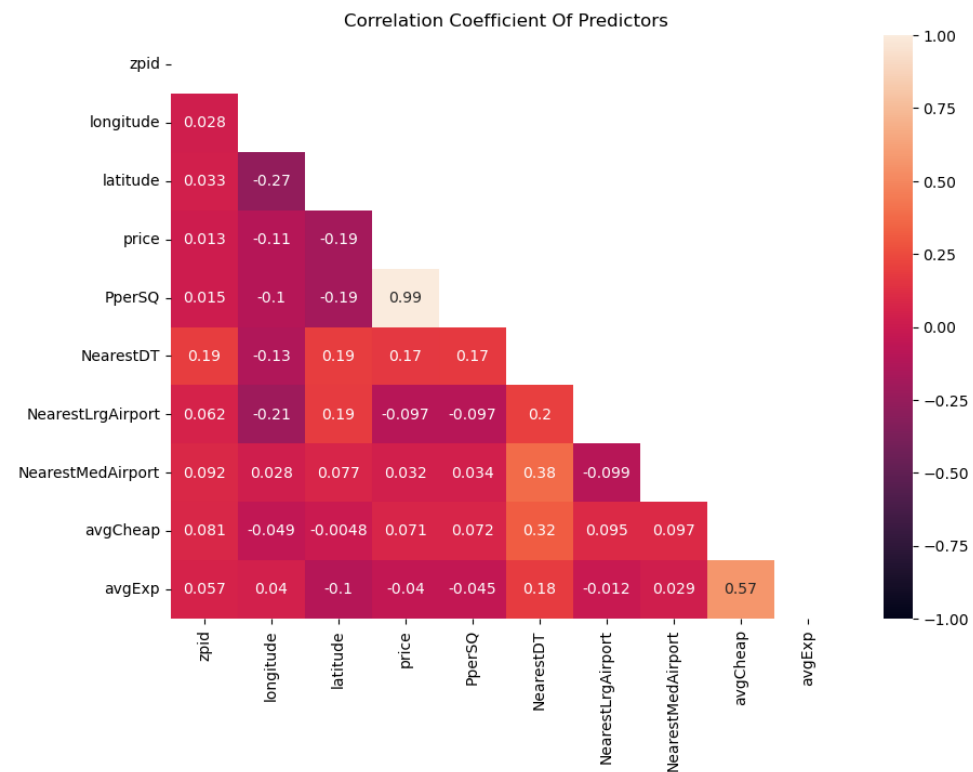
High-end fitness centers (e.g., Equinox, SoulCycle): These fitness centers target a more affluent clientele and may locate their facilities in wealthier areas. Their presence might signal an active, health-conscious community, which could contribute to higher property values.

Luxury department stores (e.g., Neiman Marcus, Saks Fifth Avenue): These upscale retailers typically locate their stores in affluent areas or high-end shopping centers. Their presence can be a status symbol, indicating a wealthy community and potentially higher house prices.

Please note that the presence of these retail companies does not guarantee a direct correlation or causation with house prices. Other factors, such as local economic conditions, demographics, and neighborhood amenities, can also influence property values.

Acquire restaurant location data

Leveraged the yelp API to acquire the three closest expensive restaurants to each house location and the three closest inexpensive restaurants.



For absolute values of r, 0-0.19 is generally regarded as very weak, 0.2-0.39 as weak, 0.40-0.59 as moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation, but these are arbitrary limits

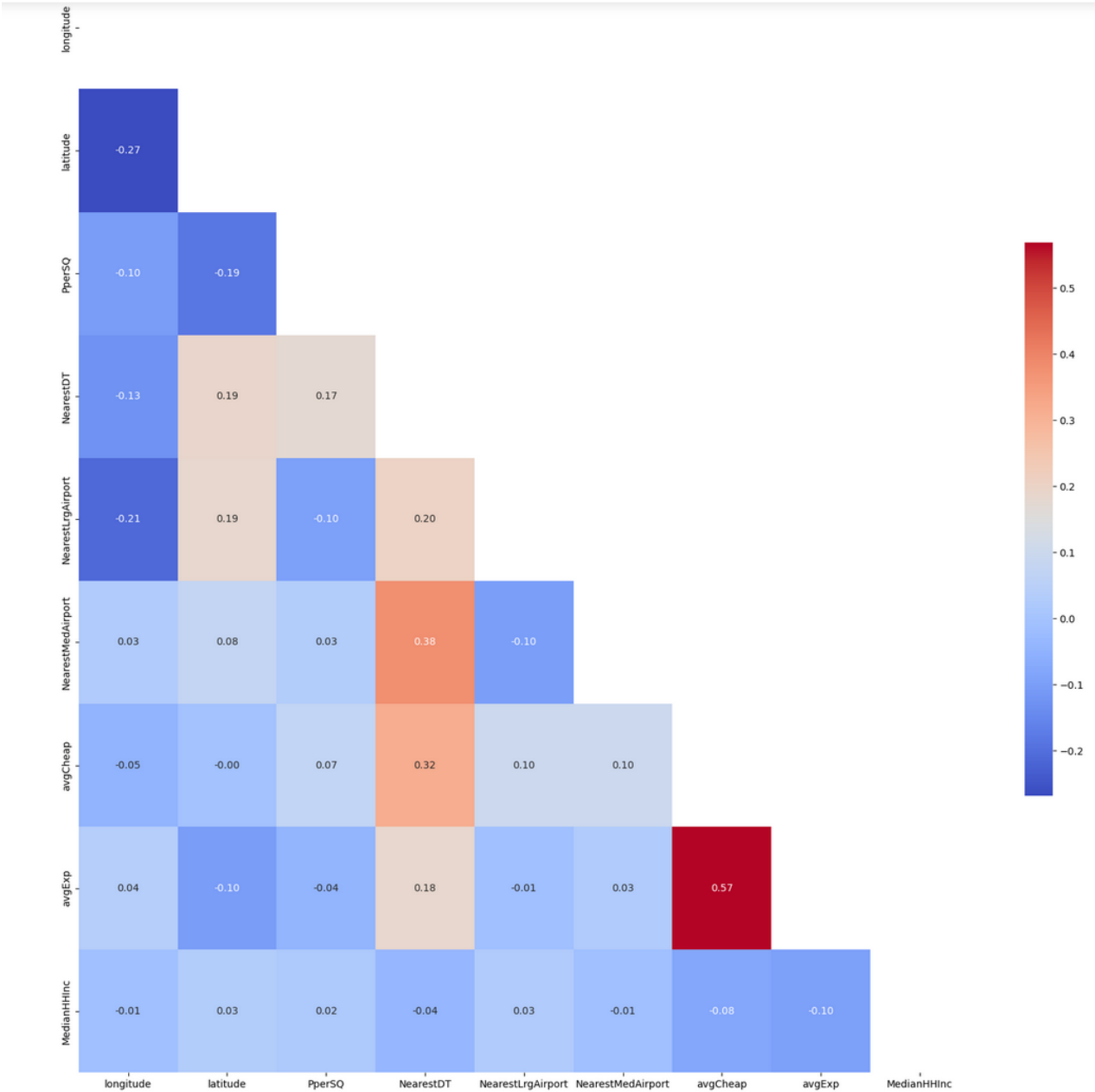
Reviewing this correlation matrix and ignoring the almost perfect correlation between price and price per square foot, the only notable correlation is a moderate correlation between the distance to a cheap restaurant and the distance to an expensive restaurant. Therefore, pursuing this as a significant differentiator is unlikely to lead to a significant finding.

Acquire median household income data for census tract:

Compare median household income (census tract level) to house price:

Correlation Method	Census Parameter	Correlation Value
Pearson	Median HH Income	0.0198
Spearman	Median HH Income	0.1561
Kendall	Median HH Income	0.1048

Heat map including median household income data for census tract:

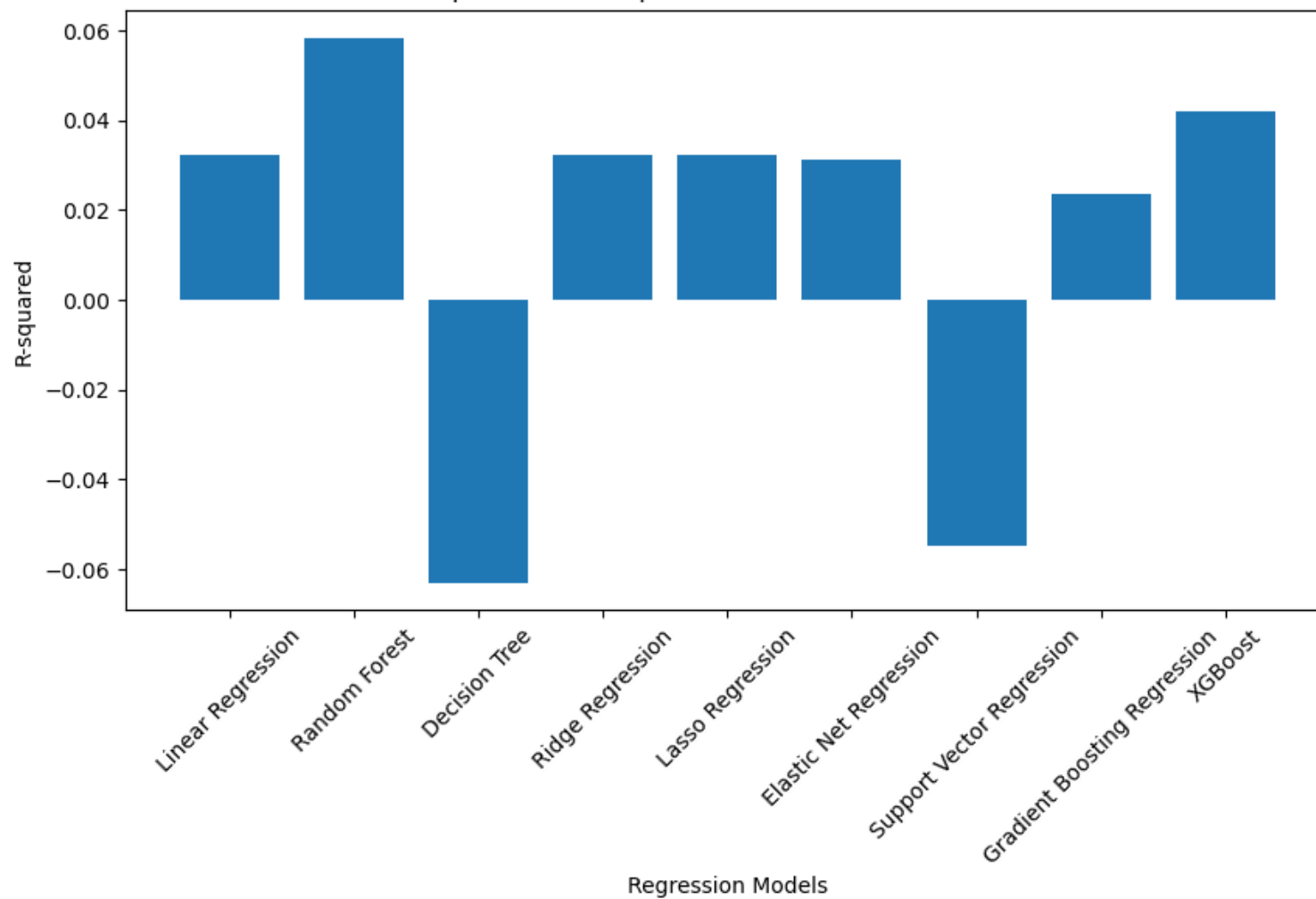


There does not appear to be a correlation amongst these various independent variables and the dependent variable, price. But maybe looking deeper will turn something up...

After scaling the columns, ['NearestDT', 'NearestLrgAirport', 'NearestMedAirport', 'avgCheap', 'avgExp', 'MedianHHInc']
And splitting 80/20 for train/test, the following methods were assessed against price prediction:

Model	Mean Squared Error:	R-Squared
Linear Regression	40798.73	0.03
Random Forest	39701.10	0.06
Decision Tree	44817.90	0.06
Ridge Regression	40799.50	0.03
Lasso Regression	40801.14	0.03
Elastic Net Regression	40849.40	0.03
Support Vector Regression	44468.91	0.05
Gradient Boosting Regression	41165.29	0.02
XGBoost	40390.05	0.04

Comparison of R-squared scores for various models



Conclusion: In the assessment of whether proximity to discount retail, airports, restaurants and neighborhood median household income, was a significant determinant of house price, there was a failure to reject the null hypothesis.

Challenges:

- Scraping data from most high-end retail sites proved to be challenging.
- Attempting to leverage lat/long coordinates to acquire U.S. census data was misguided. Using street address simplified this greatly.
- While reviewing and attempting to recreate online examples of the transfer learning machine learning process, most material older than 2-3 years was no longer usable as libraries and tools have progressed and backward compatibility was limited.

Key takeaways: Scraping data is laborious. Searching for patterns in data that likely don't exist requires a lot of mental fortitude when encountering obstacles so it's best to have a significant passion for the project if not being directly compensated for the effort.