Aneeka Latif
IST 718
Lab 6

Introduction

Forecasting is a useful business tool for production planning, inventory, and pricing. One application of forecasting is for business investments, if you can predict whether a product's value will increase or decrease over time, you can make better informed decisions about your investments. In this project, we'll be using real estate data to identify the best zip codes to purchase investment properties in.

About the data

The base data file is from Zillow (Zip_Zhvi_SingleFamilyResidence) and is a set of 30,435 zip codes and the average median housing value per month from 1996 to 2019. Each zip code also includes locality and metropolitan information. In addition to this base set, I included monthly employment information per zip code from the U.S. Bureau of Labor Statistics. The metrics provided included the total labor force (in numbers of people), number of people employed, number of people unemployed, and the unemployment rate. As all these attributes are highly correlated, going forward, I focused solely on the unemployment rate.

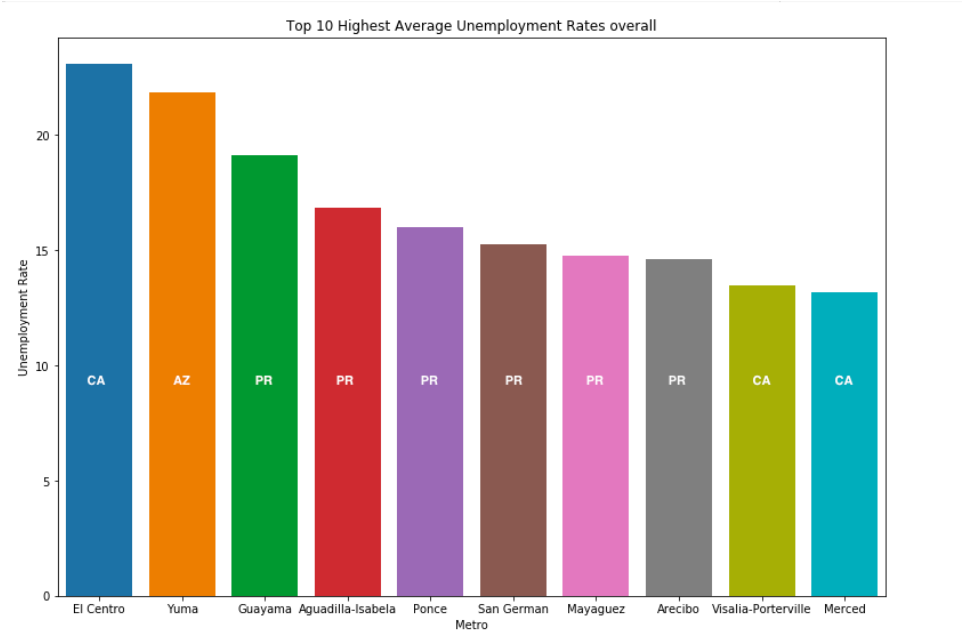
For the base datafile, because we're working with data from 1997 – 2019, I dropped the columns earlier than 1997. I then excluded rows with missing data/non-continuous data. I transformed the file using Pandas Melt which changed the structure of the file from being indexed by zip code and having a column for each monthly datapoint to rows of monthly datapoints for each zip code. I joined the BLS unemployment data to the base file and ended up with a final dataset of 2,169,912 observations in 7295 unique zip codes.

Exploratory Questions

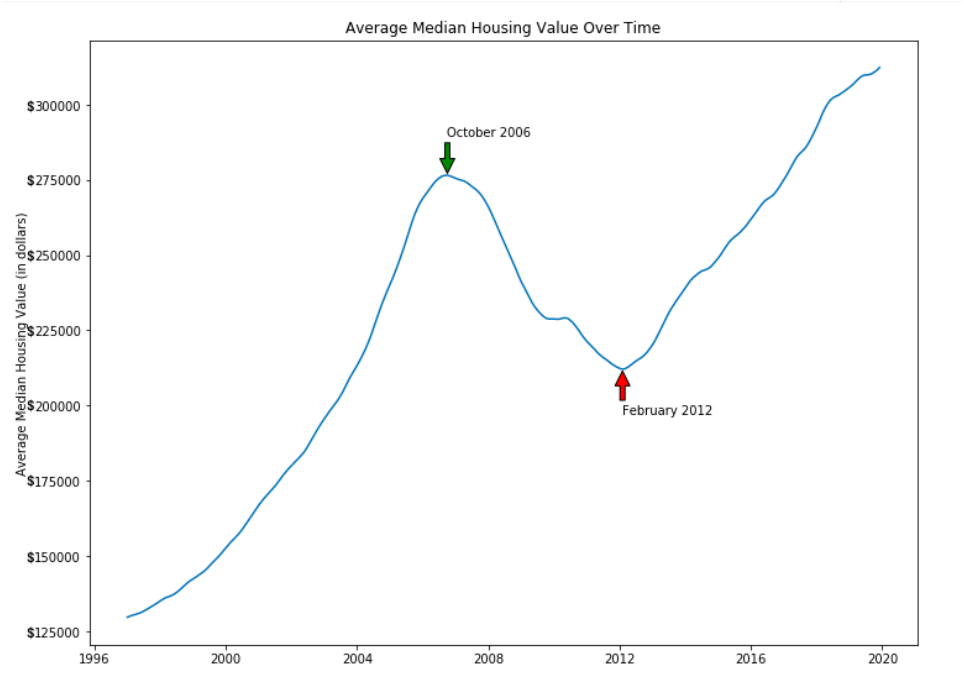
Which Metropolitan areas have the highest overall average unemployment rate?

The counties that have the highest overall average unemployment rate are mainly in California and Puerto Rico. This is unfortunately expected, as according to the U.S. Census, Puerto Rico's poverty rate is 44.9% (U.S. Department of Commerce). California's poverty rate is 18.2%, the highest in the contiguous 48 states. Counties with high poverty rates are correlated with high unemployment rates and would likely not be good investment opportunities.

Aneeka Latif
IST 718
Lab 6



How has the average median housing value changed overall?

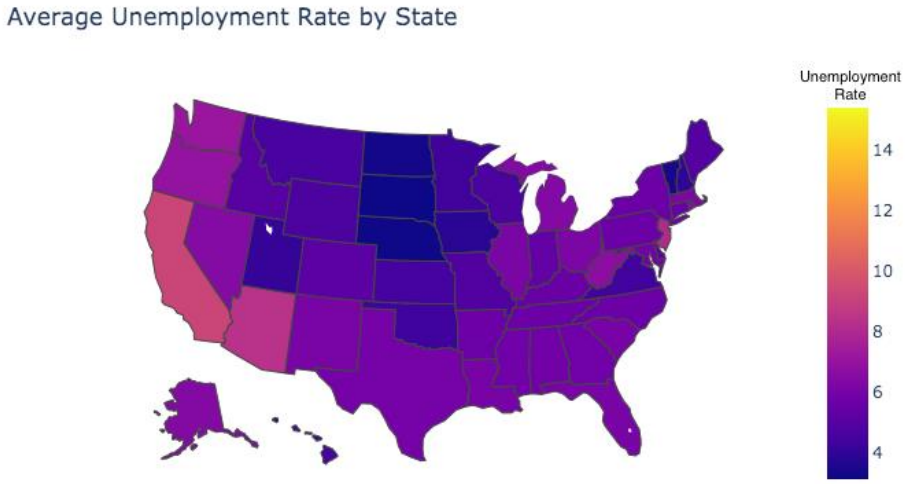


From the graph we can see that overall, the average median value of houses increased until October 2006, at which point the overall value in the United States began to fall. This coincides with the U.S. housing bubble – houses increased in value until 2006, started declining between 2006-2007, and hit its lowest values in 2012. In our dataset, the lowest overall dip was in February 2012. After which point, the housing market started to get better and housing value began to increase again.

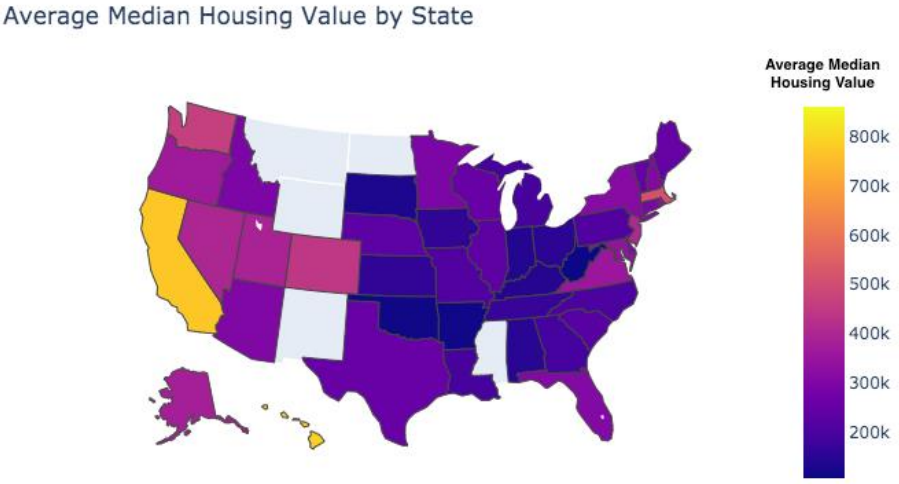
Aneeka Latif
IST 718
Lab 6

What's the overall unemployment rate by state?

We already know that California and Puerto Rico are going to be among the states with the highest unemployment rates. How do the other states compare?



Puerto Rico is not pictured but we can see California and the west coast in general seem to have very high Unemployment rates. On the east coast, all states seem to average around 5-6% unemployment with New Jersey being the highest averaging at an unemployment rate of 8.06%.

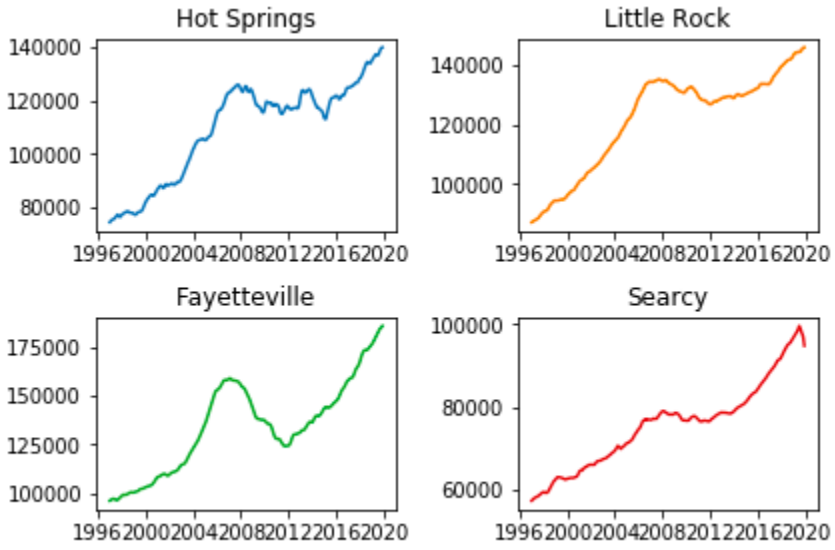


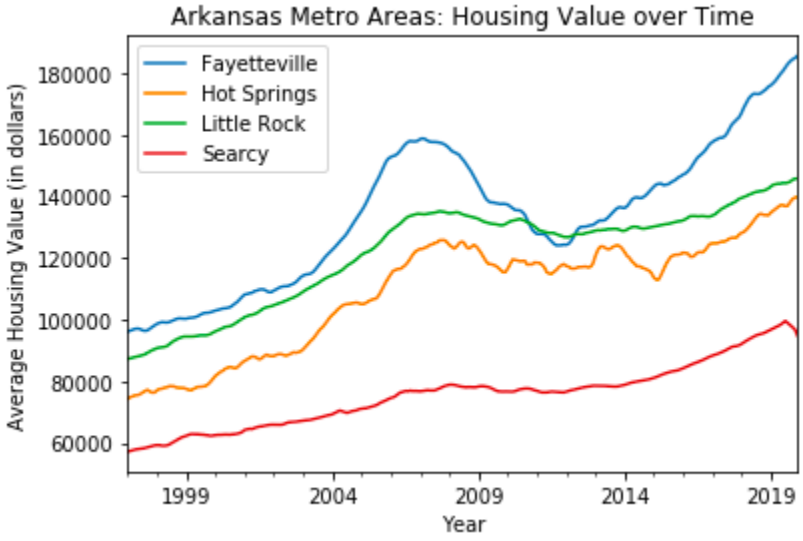
Aneeka Latif
IST 718
Lab 6

Plotting the average median housing value by state, we see many of the same hotspots for high unemployment rates also being locations for high average median housing values – namely California, Massachusetts and New Jersey.

Part 1: Arkansas

I subset the data for the Arkansas metro areas in Hot Springs, Little Rock, Fayetteville, and Searcy. I then aggregated each data into an average data point per metro area per monthly data point. Using this data, I plotted the four metro areas as time series graphs.





Aneeka Latif
IST 718
Lab 6

Between these four metropolitan areas, Searcy has the lowest value, but seems to be the most stable over time. Fayetteville has the highest value, and the quickest rate of increase, but it also is the most unstable with very extreme highs and lows -- choosing to invest in Fayetteville would be very risky but has the potential for the highest reward. Investing in Searcy would result in the lowest potential reward but is the least risky. The shape os Little Rock and Hot Springs are very similar, with Little Rock being smoother and more stable over time. If I were to make my decision on a guaranteed profit, I would choose Searcy. If I were to make my decision on what would give me the best return for my investment, despite the risk, I would choose Fayetteville.

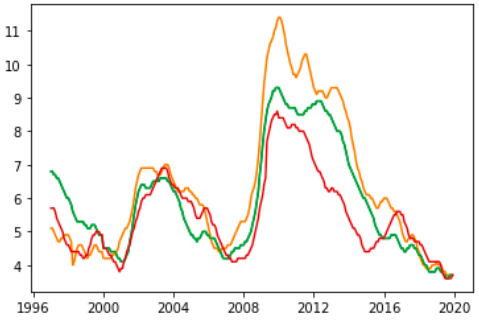
Part 2: United States

Methods and Analysis:

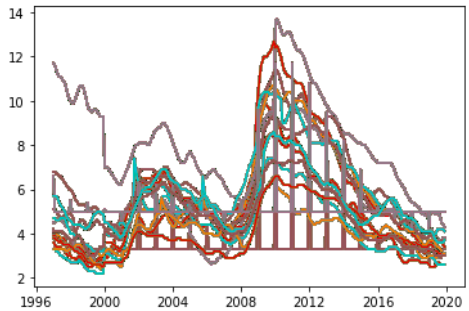
The model we're using to forecast is Facebook's Prophet which is an additive model ideal for nonlinear trend data. For the base model, I organized the observations into sets of data by zip code. I then generated some future dates for Prophet to plot against and generated a model for each zip code. Prophet's built-in performance metrics generate an error value using cross validation; only a portion of the data is used for model creation, and the forecast is validated against an unused portion of the dataset. Instead, I decided to calculate my own error values by finding the Mean Absolute Percentage Error – the mean absolute difference between the forecasted y value and the actual y values. I determined that the most accurate models would have the lowest MAPE values.

To better the base model, I added unemployment rate as a regressor. The caveat to predicting using Facebook's Prophet for multivariate time series is that you need to predict future values of your regressor variables before fitting your Prophet model to the data. This was such a large dataset, it was impossible to use standard ARIMA / ARMA models to predict these values because we are not able to plot and evaluate each model to evaluate its optimal parameters.

Originally, I mistakenly thought that if I plotted the time series for each zip code, I would be able to approximate order values.



Attempt: plotting four time series to identify lag patterns



Attempt: Plotting all time series data to identify trends

As you can see, this was not the case, there's no clear pattern of lags, and no clear similarity between the zip codes. This indicated that I needed to find another forecasting method that did not require lag identification.

Instead, I used the pmdarima package and auto_arima method to generate random fit models for each zip code. This would automatically find the best order parameters and create the best model by iterating through a set of constraints. Knowing that the unemployment data required transformation, I created two variations of random_fit models.

Model 1	Model 2
Logarithmic	Logarithmic Difference
Log of Unemployment Rates	Shifted Delta Log of Unemployment Rates

In model 1, I first normalized the data by taking the log of the unemployment rates. Once the data was transformed, it was prepared for the auto_arima function. Once the models were created for each zip code, I forecasted the unemployment values for the next 13 time periods – effectively bringing the forecast into 2020.

In model 2, I again took the log of the unemployment rates. To further induce stationarity, I additionally took the difference in log values by shifting the dataset and recording the difference between an observation and its preceding value.

I had one final step in Prophet preparation in joining the calculated future dates and predicted corresponding future unemployment rates. I joined the two sets of data by zip code, renamed the columns as appropriate for Prophet, and predicted the median average housing value for each zip code. I calculated MAPE scores for each model as an accuracy metric and plotted the top 10 models; the best 10 models are those with the calculated lowest MAPE score.

Results:

Base Model – the average MAPE score for the top 10 base models is 24.61%

Model 1 – the average MAPE score for the top 10 logarithmic models is 31.34%

Model 2: the average MAPE score for the logarithmic difference model is 24.73%

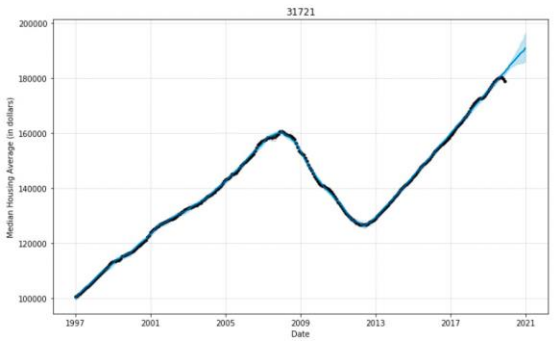
Model 2 resulted in the best top three models, all three incidentally also ranked in the top 10 for the base model.

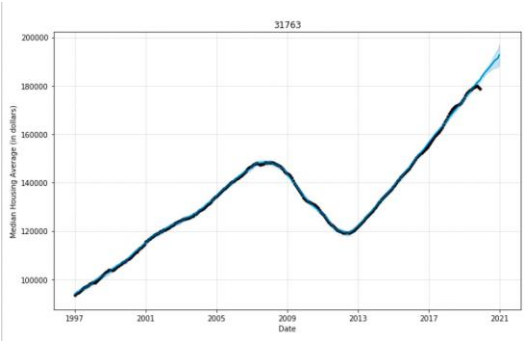
Base Model (no Regressor)		Model 1 (Logarithmic)		Model 2 (Logarithmic Difference)	
Region (Zip code)	MAPE	Region (Zip code)	MAPE	Region (Zip code)	MAPE
31721	22.67%	73013	25.73%	31721	22.04%
72543	23.98%	27603	28.46%	31763	22.72%
72364	24.20%	63376	28.96%	38340	22.76%
50112	24.37%	27587	29.37%	43113	24.61%

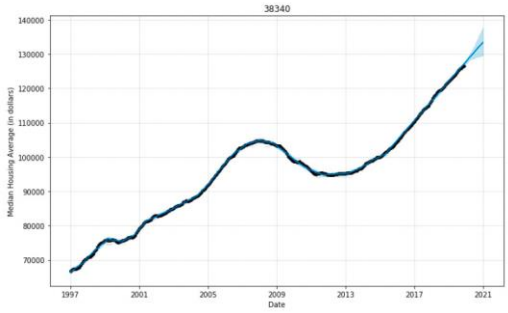
Aneeka Latif
IST 718
Lab 6

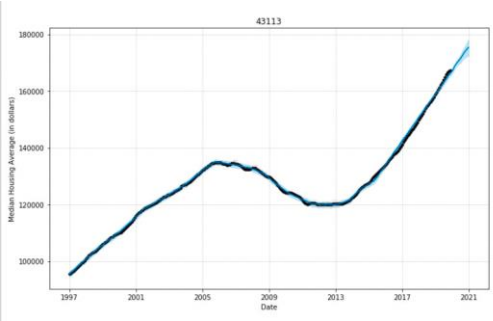
38340	24.76%	74012	30.08%	73170	25.49%
74403	25.12%	40601	31.87%	70047	25.70%
31763	25.18%	77083	32.36%	73013	25.87%
31794	25.22%	74133	34.42%	73025	25.97%
70047	25.28%	45601	35.72%	73072	26.16%
72802	25.34%	29681	36.45%	80421	26.21%

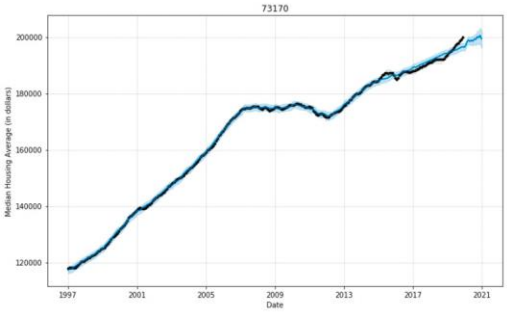
Top Five Forecasted Zip Codes using Logarithmic Difference:

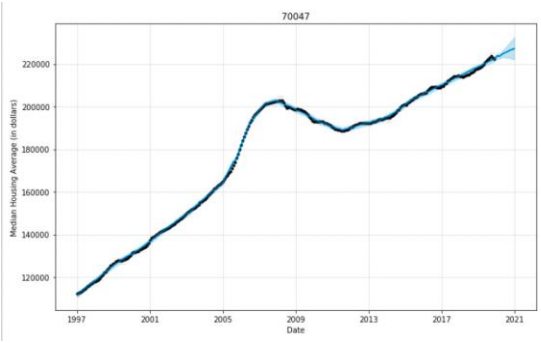




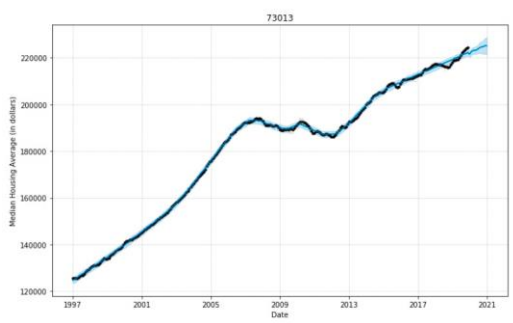


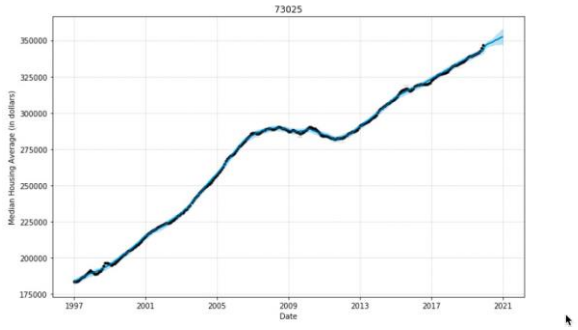


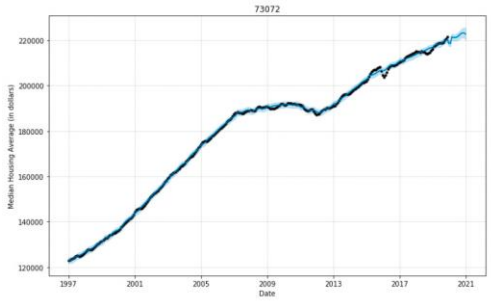


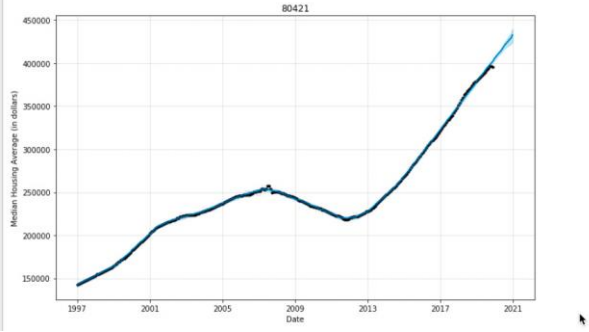


Aneeka Latif
IST 718
Lab 6









The graphs that depict the best investment opportunities have tighter forecasting bounds, are generally more stable (smooth lines), have sharper slopes, and offer the opportunity for a high investment return. From looking at these graphs, 31721, 31763, and 80421 appear to indicate the highest forecasted increase in value over time. 31721 and 31763 have the lowest mape scores and indicate high value increase over time. Though 80421 shows high value increase, it has the lowest mape score of the top 10 – however, the forecasted value is the highest amongst all the models and would result in the highest investment return. 38340 and 43113, though similar in shape, and with low mape scores, have much lower maximum housing values, though the risk is very low, the potential reward is also very low.

Conclusion

In terms of accuracy, choosing the zip code with the most accurate model is the least risky choice and is less likely to result in financial loss. In terms of investment, it's better to choose the zipcode with a model that is not only safe, but suggests a higher investment return in the quickest timeframe. The zip codes that are safe and are expected to return a profit quickly are 31721, 31763, and 80421.

Aneeka Latif
IST 718
Lab 6

Citations

"U.S. Census Bureau QuickFacts: Puerto Rico." Census Bureau QuickFacts, www.census.gov/quickfacts/PR.

"ZHVI Single-Family Homes - Dataset by Zillow-Data." Data.world, 26 Oct. 2017, data.world/zillow-data/zhvi-single-family-homes#__sid=js0.