Predicting US Vacancy Rate by Zip Code

Capstone 2

Joseph Frasca

the problem See problem id worksheet

What is the current vacancy rate in a certain zip code?

Lack of specific current vacancy rate data hampers real estate investors ability to definitively know a key factor in their decision for where to invest. Currently many investors rely on local knowledge of an area - ie. local real estate agents, property managers etc. for this knowledge - so the info is hard to get and hard to tell if accurate. There is vacancy rate data for the US as a country in up to the current year on the Federal Reserve of Economic Development website (FRED). There is also vacancy rate data by zip code available up to 2018 through the American Housing Community, but currently there does not seem to be vacancy rate data by zip code for the current year.

the approach,

The goal was to predict current vacancy rate by zip code for 2019-2020 using housing market indicators and other econometrics.

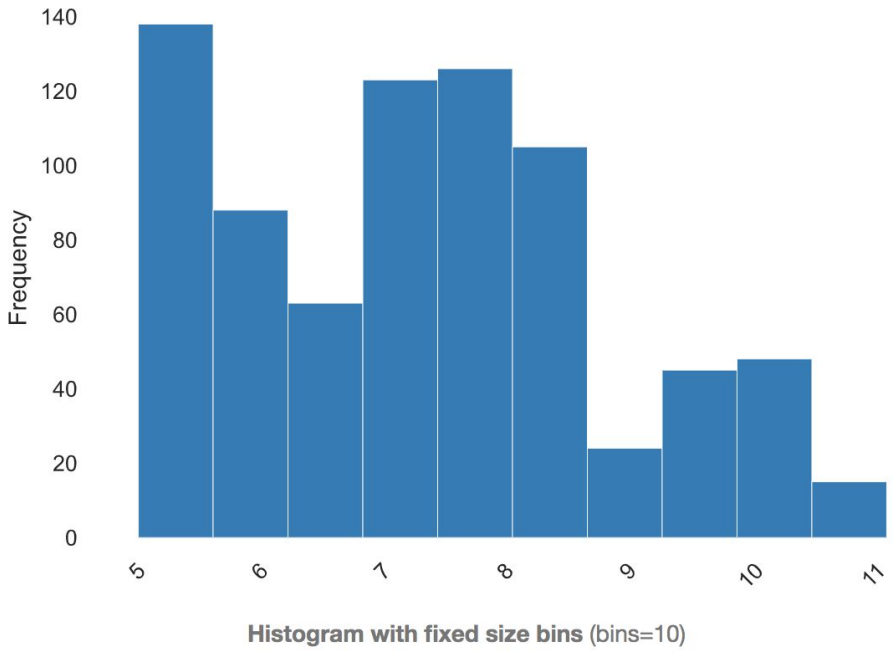
US Rental Vacancy Rate 2011-2020 (Quarterly)

Overview of Jupyter Notebooks

Juptyer Notebooks 1.0-2: National US Data

First I wrangled US data from the Federal Reserve Bank of St. Louis (FRED) for several key econometrics related to the health of the US housing market and the target variable rental vacancy rate. This includes population growth, median household income, unemployment rate, interest rate, rent price CPI, home price index CPI, new housing starts, residential construction price index PPI, and residential construction spending. Then I proceeded to clean data types, column names, etc. I saw that different variables had different timelines, see reference section for Data Definition and hence some variables with limited time spans had a lot of missing data. I merged the different datas into a single dataframe that was broken down by month. After exploring the data I dropped all data before 1956, as 1956 was when the target variable, vacancy rate, started and most of the data would be preserved.

Frequency of Rental Vacancy Rates from 1956 - 2018

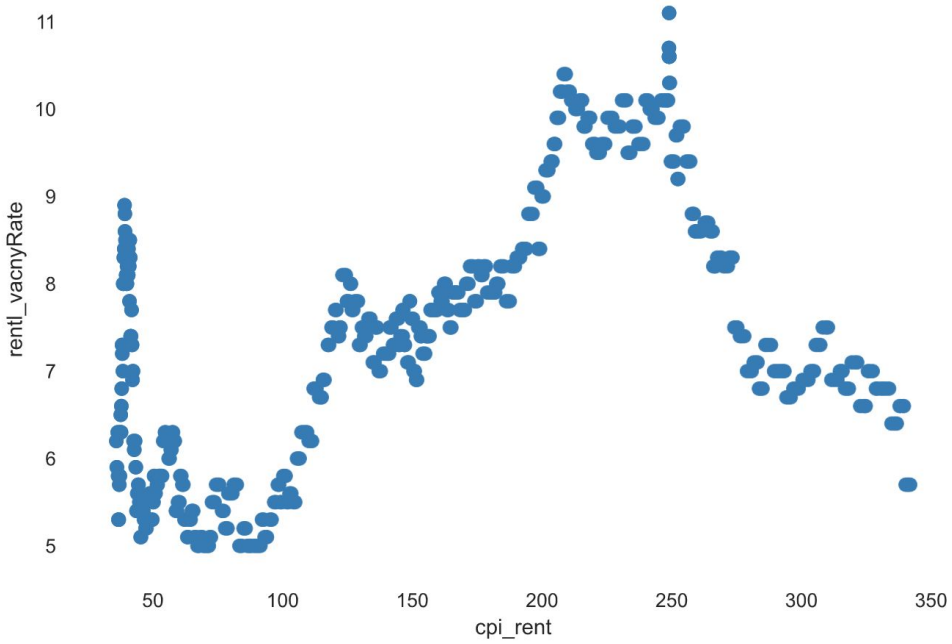


Rental Vacancy Rate (%)

I explored the data using pandas profiling, see [html version in reports folder]. The **variables that had stronger correlations with rental vacancy rate were: median household income (positive correlation), interest rates (negative), and cpi rent prices (positive).** I also noted that 'cpi_rent' was strongly positively correlated with 'homePrice_index' and 'ppi_resConstruct'. **Overall rental vacancy rates seemed to have more of a "sideways" quadratic correlation with cpi_rent and home prices vs. a pure linear correlation**. From all this I formed the following hypothesis:

- **Null hypotheses:** rental vacancy rates ARE NOT correlated with median household income (positive), interest rates (negative), or cpi rent (positive). And can therefore not be used to model future rental vacancy rates.
- **Alternative hypothesis:** rental vacancy rates ARE correlated with median household income (positive), interest rates (negative), or cpi rent (positive). And can therefore be used to model future rental vacancy rates

Rent vs. Vacancy Rate 1956-2018

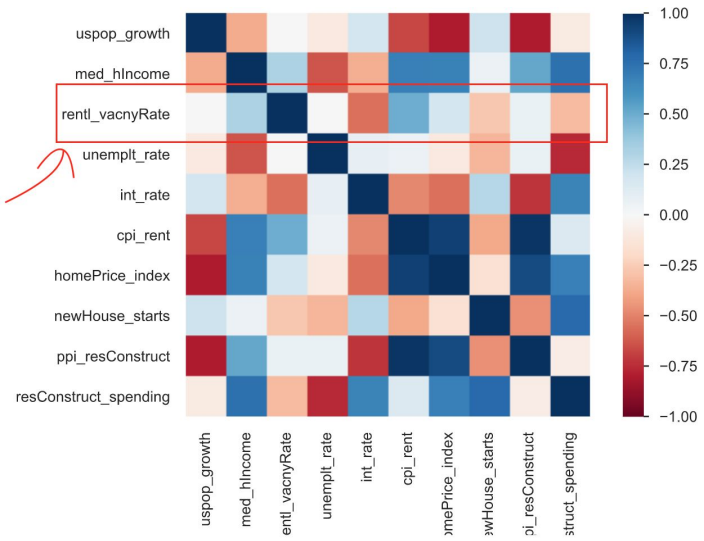


Due to different variables in the time series having data beginning and ending at different times, I created 4 separate DataFrames - some with longer timeframes and less variables and some with more variables and shorter timeframes, as this was the tradeoff that had to be made. I then proceed to see which of the 4 data frames might produce the best predictive model for vacancy rate. I dealt with remaining NaNs in data (due to some variables starting later than others) by simply trimming the data sets to the time periods that had all data for those variables. I then compared the correlations of the different data frames and **saw different trends/correlations between variables depending on the different time periods. This was interesting to see that during different economic periods in US history the variables had very different relationships**. I noted these trends as comments in the notebook.

Differing Correlations Dependent on Time Period

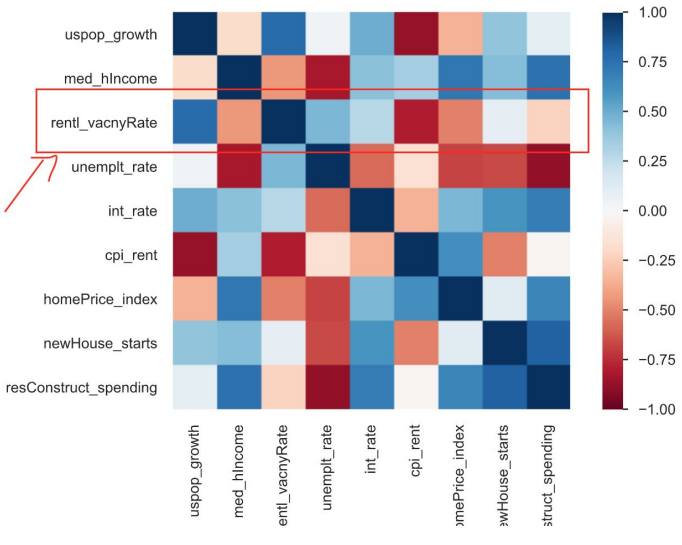
Correlation Heatmap, 1965-2018

(Vacancy Rate Highlighted)



Correlation Heatmap, 2002-2018

(Vacancy Rate Highlighted)



I then split all 4 data frames into testing and training datasets, dropping ppi_res construct because of it's high correlation with cpi_rent. Next I established Baseline Measurement Comparisons using a Dummy Regressor to see what the R2, MSE, and MAE would be if the mean of the DataFrames were used. Then I standardized the magnitude of numeric features using a scaler and trained a linear regressor on the four data frames. **I found the best performing dataset was the 8 variable dataset from 2002-2018** (labeled 9 variables, see note at top of notebook), **on the test set this model had an R2 of 92%, a mean absolute error of .29, and an mean squared error of .14**

Juptyer Notebooks 2.0.0-2.0.3: Wrangling by Zipcode Data

While cleaning the data I **calculated two new variables; vacancy rate (using the calculation # vacant/#total**), and also the margin of error (MOE) of vacancy rate using an error propagation formula. While checking for NaNs, I saw Vacancy_Rate NaNS due to Estimate!!Total == 0 AND the Margin of Error of Vacancy_Rate NaN due to Estimate!!Total!!Vacant == 0. I then set these NaNs to 0 because either the corresponding zip code had 0 population or 0 homes vacant. I noted there were some MOE Vacancy Rates larger than Vacancy Rates, which doesn't make practical sense, because then your range of potential vacancy rates would include negative numbers. I dealt with these occurrences by setting MOE-VacancyRate = Vacancy_Rate, which then would have the bottom end of the range be 0, which makes practical sense.

Throughout the cleaning/wrangling process my goal was to see if I could get a model that would be able to predict vacancy rate, so I often opted to drop rows with NaNs because quickness in getting to the modeling stage was what I was looking for. I knew I could always come back and do a more thorough job of imputing missing data to preserve variables later on after being more certain the model would be robust/effective.

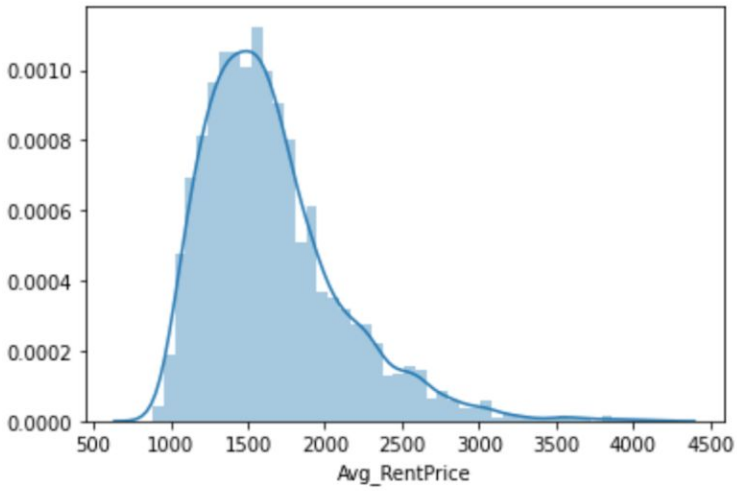
In the process of wrangling the rent price by zip code data I realized the **Zillow Public Data only had rent 2014-2020 price data for ~3,000 of the ~33,000 zip codes**. So I found 2011-2018 rent price data by zip code from the ACS. However I had to do some work to calculate the average rent price by zipcode from the ACS data because this data had columns with different income ranges (ie. $1000-1490, $1500-$1999) and then the values were the number of people in that row's zip code who paid that column's listed rent . So I created a column with average rent price per zip code by summing median rent of each column's given cash rent range, ie. if the range was 100 to 149 I used average rent price = 125. Then multiplied this by number of people (value of the column), ie.125*p. Finally I divided this by the total *n* number of people in that zipcode to get the average rent price in that zip code. Note: for rents above $3500 (the ACS range listed this as $3500+), I used $5401.47 as this was 2 std higher than the mean from zillow rental data. Some zip codes had average rent prices as NaN, I couldn't figure out why at first, I assume because the *n* number of people residing in that zipcode may have been 0. I did not pursue this further as I had decided not to use this data for the model for the reasons explained below.

After cleaning the Zillow rental data I compared Zillow to the ACS data. I then shifted the ACS data based on the difference in medians of both datasets, (ie. added 218.85 to every ACS rent price to match the 50% of the zillow data). I did this because I trusted the zillow data more as this was live market data for those areas. After the shift the datasets had similar 50% quantile rent prices and 25% quantile rent prices but the zillow dataset had a lot more variability especially on the higher ends of rent which contributed to its overall mean being around $150 higher than the ACS datasets. I then created a function for getting the shift for each yearly ACS dataset, as the relationship between the ACS and Zillow data may have been different depending on the year. This turned out to be true, and the shifts ranged from -$30 to +$218 (the average across 2014-2018 was a $156.74 difference in medians). So the 2014-2018 ACS data was shifted based on the difference between corresponding zillow year's median so the datasets then had the same median. For 2011-2013 ACS data there wasn't a corresponding zillow dataset to compare it to so I used the average difference in medians between the zillow and ACS data ($156.74) to shift these datasets.

Histogram of ACS rent price data and Zillow rent price data for overlapping zip codes

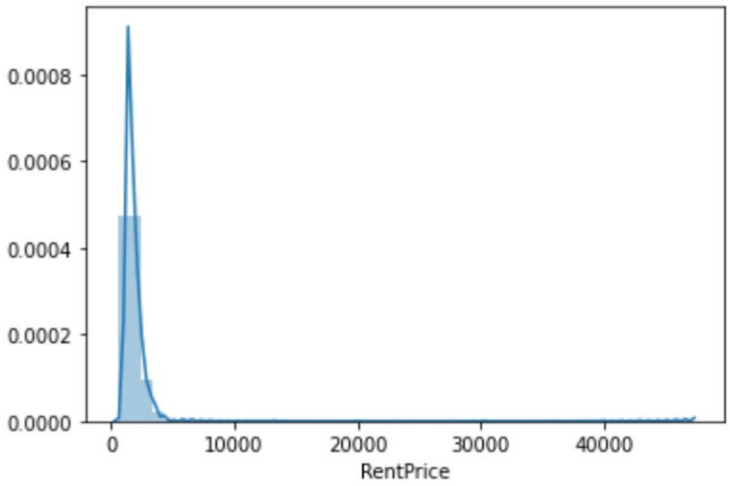
```
#look at histogram of ACS rent prices (of subset that match zillow data with same zipcode)
sns.distplot(check_diff2.Avg_RentPrice)
```

```
<AxesSubplot:xlabel='Avg_RentPrice'>
```




```
#look at histogram of zillow rent prices
sns.distplot(check_diff2.RentPrice)
```

```
<AxesSubplot:xlabel='RentPrice'>
```

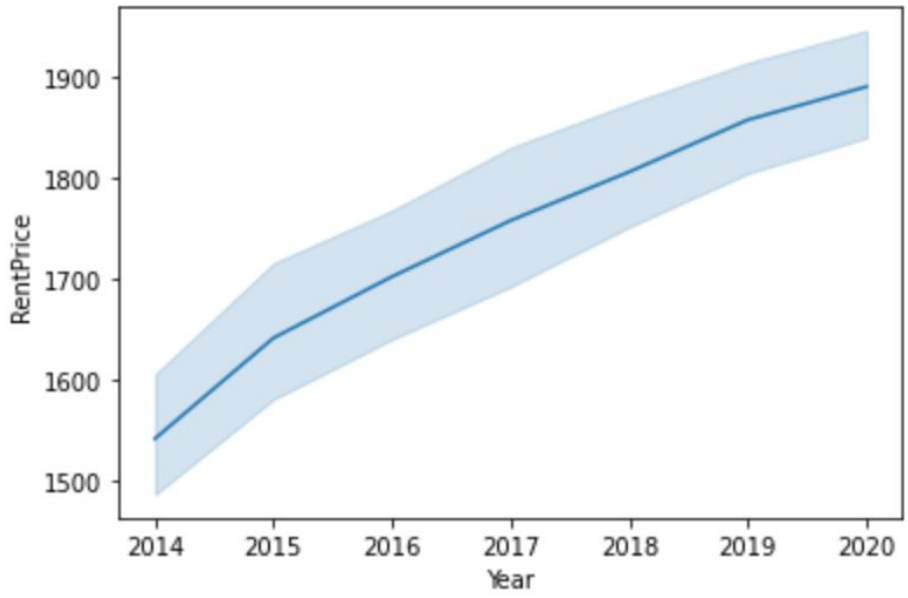


After the shifts I checked the Zillow and ACS data sets for a zip code I knew to see how they compared. They were different by around $300 with Zillow being the higher of the two and seemed more accurate for the rent prices I knew of in the area. Using this I decided to keep all the Zillow data for the zipcodes we had (only ~3000 zip codes) and then have the ACS data for the rest (~27,000 of the ~30,000 zip codes), again **because I trusted the Zillow data more than the ACS especially after seeing the truer variability observed in the zillow data.** I also plotted the ACS + zillow joined data, **the ACS data prices are vastly different from 2013 and 2014 it seems because the ACS possibly used different data collection metrics between those years. This being another reason I trusted the ACS data less.**

Line Plot showing Zillow Rent Price vs. Year

```
#plot zillow rental data from 2014-2020
sns.lineplot(x='Year', y='RentPrice', data=df_rents_2014_2020)
```

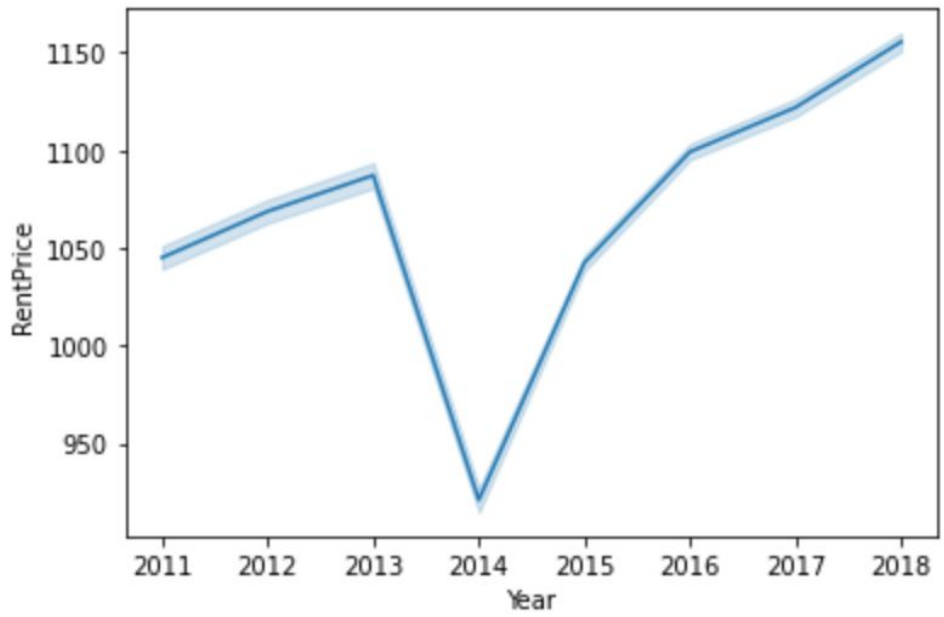
```
<AxesSubplot:xlabel='Year', ylabel='RentPrice'>
```



Line Plot showing ACS Rent Price vs. Year

```
#plot ACS rental data from 2011-2018
sns.lineplot(x='Year', y='RentPrice', data=df_ACSrents2011_2018)
```

```
<AxesSubplot:xlabel='Year', ylabel='RentPrice'>
```



During this work I decided to abandon the idea of getting rent prices for all the zip codes as I realized I did not have 2019-2020 data for all the zip codes, only the ~3,000 zip codes from Zilllow. Furthermore the ACS data seemed less trustworthy than the Zillow data (see reasons stated above). I would first need to use a model to predict rent prices by zip code (which seems do-able because home prices are so correlated with rent prices, and Zillow has current home price data on around > 30,000 zip codes. So I instead created one large zillow dataframe with rental data from 2014-2020 for later predictive modeling.

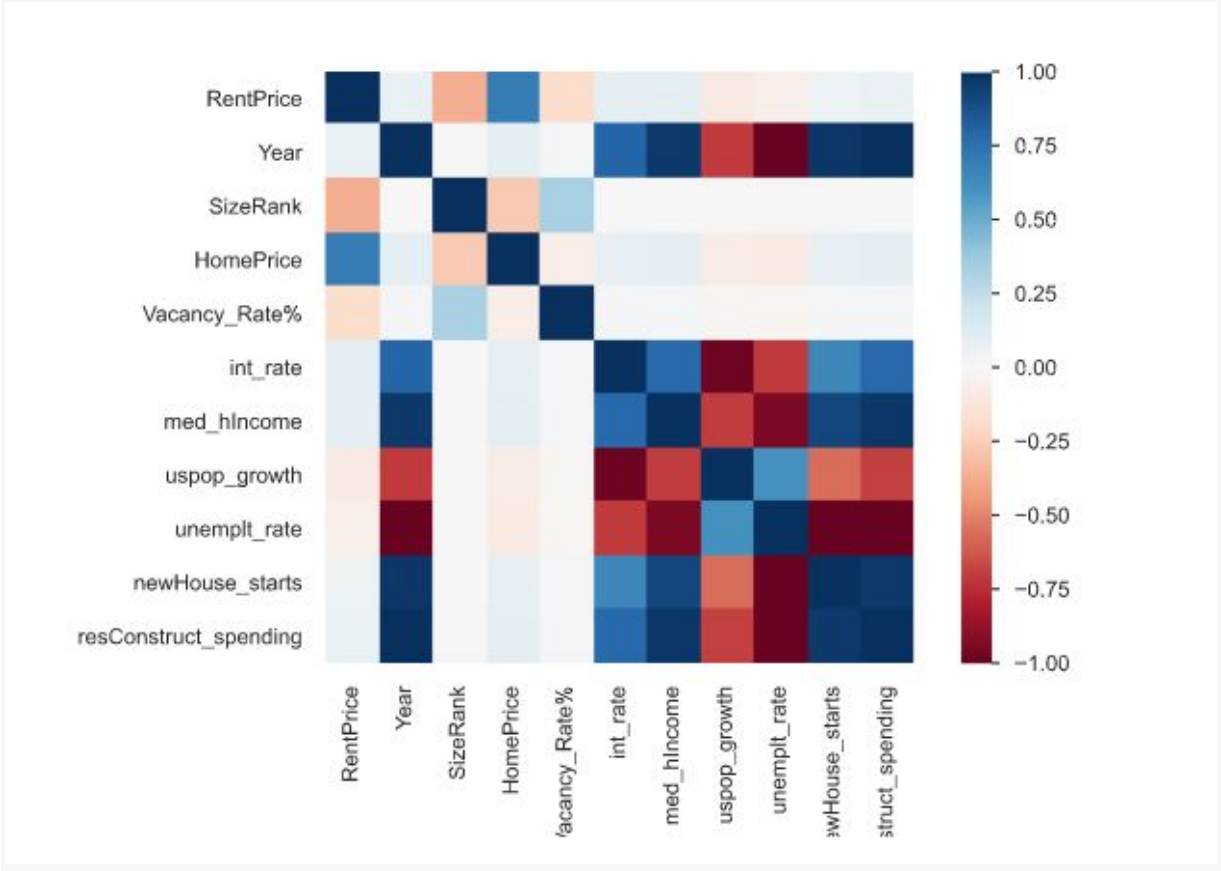
I also inspected and cleaned FRED national annual economic data and Zillow home price data to help predict vacancy rate by zip code.

Juptyer Notebooks 3.0-3.2: Attempt to Combine National Econometrics & Data by Zipcode

Note, during wrangling the primary goal was the merge the separate dataframes into one dataframe from 2011-2018 (matching the dates from the ACS vacancy rate by zip code data) that contained vacancy rate, home prices, rent prices by zip code and the national economic data from FRED filled for all zip codes in the corresponding year. During EDA this seemed to be less fruitful than anticipated. So as a second pass I created and wrangled another data frame from 2014-2020 with home prices and rent prices from zillow and vacancy rates from ACS data.

During EDA, I created a pandas profiling report for vacancy rate, home and rent prices, and econometrics (zipcode master notebook). **I noted that all the national FRED econometric data was highly correlated, this correlation was most likely the result of a single variable from national data being used for every zip code in that year (around ~33k zip codes each year). I concluded this would most likely not be a good data set to build a predictive model.** Then I created line plots of vacancy rate data, rent price and home price data and noted vacancy rate and home prices had a steady upwards trend. Home prices dipped from 2011-2012 most likely from the housing market collapse. There was a significant dip in rent prices from 2013 to 2014 may be due to a different data collection method being used by the ACS those years. I ran the pandas profiling report from 2014-2018 to take out the large dip in rent prices from 2013-2014 and see if the data has any different relationships, the strong correlations between national data remained so this didn't seem especially helpful.

Correlation Heatmap showing high correlation of national econometrics



During modeling, I attempted to fit the training data on a linear regression model... Note: this notebook was not completed because preliminary results from EDA suggested it may not be helpful and also because the .fit() function took far too long and did not yield helpful results even as the data was fit in batches.

Juptyer Notebooks 4.1-4.4: Using Zillow Data to Create Final Model

Performed EDA of the merged vacancy rate and zillow home/rent prices dataframe. I split into two dataframes for future modeling 2014-2018 and predicting vacancy rates in 2019-2020. After checking NaNs as a % of the column, I saw each column had less than .2%, and decided to just drop them for the initial model.

Pandas profiling EDA showed vacancy rate had a positive correlation with rent prices, and a smaller correlation with SizeRank (of the zip code). There didn't seem to be much of a correlation with vacancy rate and home price. Vacancy rate was highly correlated with Margin of Error (MOE) Vacancy rate. Also note that the categorical variables of City, County, Metro Area, State etc were not included in the correlation heatmap and this may affect future models.

I then examined line plots of average vacancy rate over time, average rent price over time, and average home price over time. The rent price line plot from Zillow data seemed much smoother than the ACS data seen in earlier notebooks and aligned with the national FRED data trends, giving me more confidence in this data.

During preprocessing, I dropped the margin of error (MOE) of vacancy rate variable, as we do not have this data for 2019-2020 and therefore would not be helpful with a predictive model, also it was highly correlated with vacancy rate. I performed a Time Series train/test split on the data. I also decided not to use scaled data as it led to negative values for R2 scores. Negative R2 may be because I tried to scale data after creating dummy variables. Then I used a linear regression with an R2 score of 74.4% and mean absolute error (MAE) of 2.44 on the test set, this compares to a 5.19 MAE using a Dummy regressor.

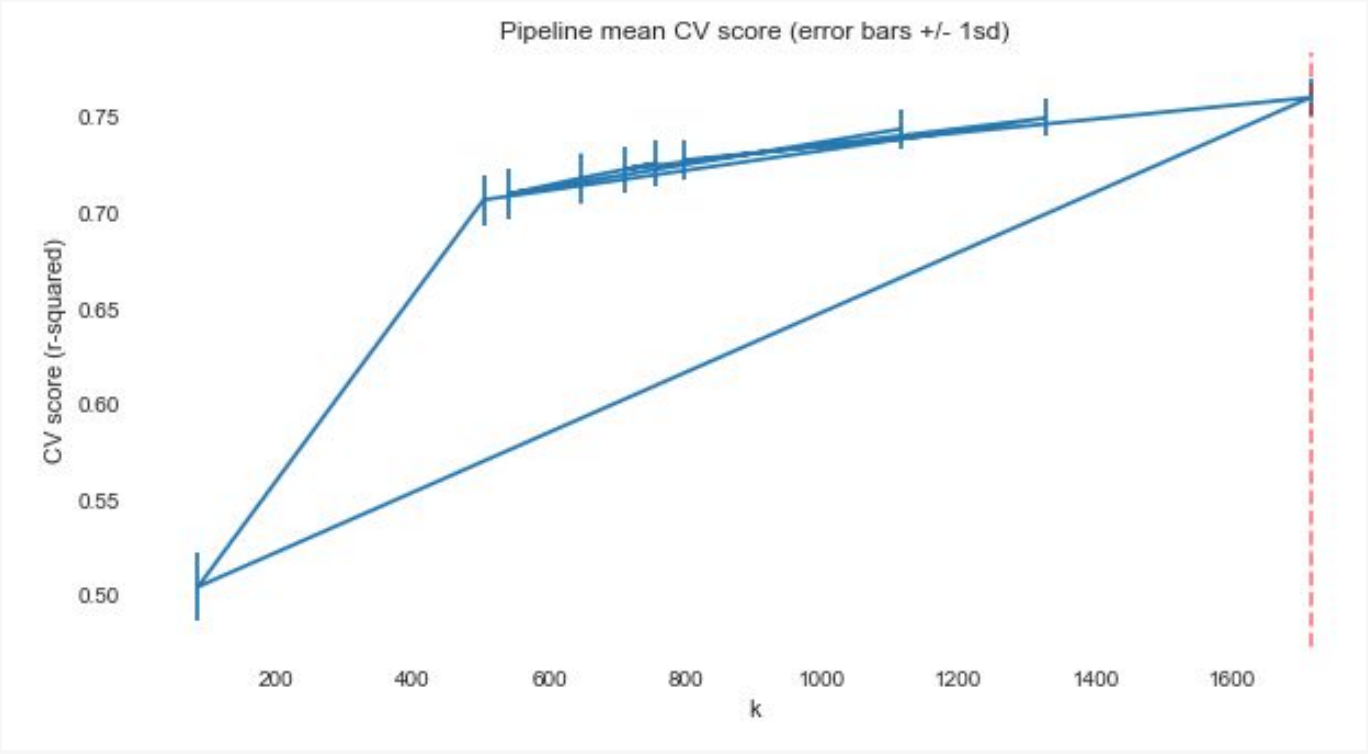
During modeling, I decided not to include national data for this model because I wanted to see what kinds of preliminary results we got and since it consisted of only 15 rows it may not have made a major impact. After loading the data, I defined

X, y variables and did TimeSeriesSplit using 5 = n_splits for the 5 years from 2014-2018. As before I decided not to use scaled data because variables were in the similar format and scaled data provided -4 or so results for r^2 scores (possibly because dummy variables had been created before the scaling).

LINEAR REGRESSION MODEL

I re-examined the linear regression model tested in the preprocessing notebook using adjusted R2 for training/testing sets (because data set is rather large) and got adjusted r2 for train set: 73.6%, while only an adjusted r2 for test set: 43.9%. **It seems the linear regression model is not performing that well when looking at adjusted r2 test performance vs. r2 test performance, possibly because the data set is larger. I** tried tuning some hyperparameters using SelectKBest where k=10, but the model performed worse with around 23% with 5 fold cross validation. Then I used RandomizedSearchCV to find the best k, plotted k vs. cv scores and it **seemed like the best CV score approached 75% CV scores with k=1715** (for reference there are 1754 total variables). See graph here.

Selecting best 'k' for Linear Regression Model



Then I inspected the feature importance of the linear model with: `coefs` as the values and `features` as the index. Results suggested that Cities (in Hawaii), States (in the northeast - PA, NY, NJ, CT), and Metro Boston, Cambridge, Newton are the features with greatest importance in the model. Further exploration of these variables is needed.

RANDOM FOREST MODEL

I used 5 fold cross validation on training data with the out-of-the-box random forest regressor and found a mean cross validation score of 93.66% and std of 0.02%, a marked improvement from the linear regression model. The Mean absolute error was 1.46, also an improvement from the linear regression model. The **model had significantly improved performance on random forest test data R2: .92. However Adjusted R2 on test data showed a score of .78 which seems to suggest some of performance is skewed by the large number of variables, and that some of these variables may be unnecessary to the model.** I worked on hyperparameter tuning using RandomizedSearchCV to search for 'randomforestregressor__n_estimators' and found the best n_estimators was 615. I ran the model with the optimized n_estimators and found not much difference in performance. Since this took so much longer to load vs. the out of the box model, I decided to use the out of the box random forest model and did not do further testing with these hyperparameters. I tried to see feature importance on the graph, but seemed to have difficulty visualizing, and used the Lasso Model instead (see below).

XGBoost Model

The XGBoost Model took several days of running and still did not load, so it tried data in smaller chunks. The model fitting for XGBoost, on 100 rows took 80.19 seconds and got R2: -0.7574101368649369. Model fitting for XGB classifier, on 200 rows took 304.025593996048 seconds and got an R2: -0.96. I stopped running this as it was not yielding better results in a realistic amount of time.

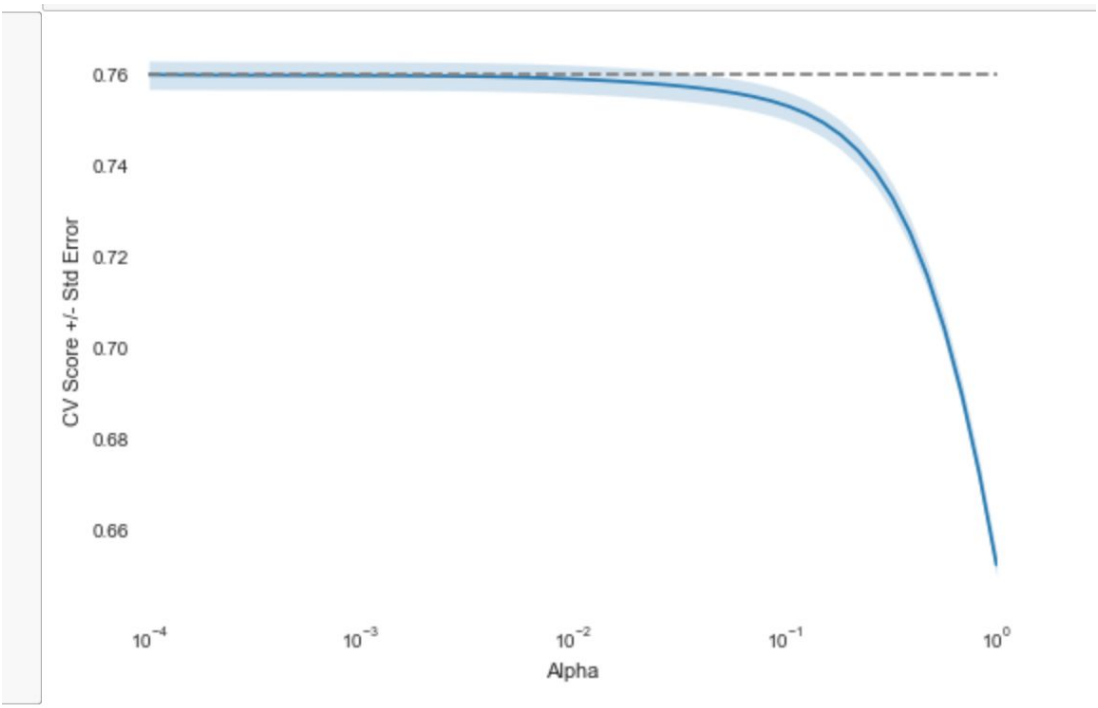
LASSO MODEL (to create a feature selector)

It seems from this that the **top 10 features in predicting vacancy rate from the Lasso Model comprise: zipcode, rent price, year, size rank, home price, State(s).** Note that none of these features had a coefficient much greater than 0.

RIDGE REGRESSION MODEL

I created a ridge regression model and computed scores over range of alphas, performing 5-fold CV. It seems like CV scores maxed out around .76 with an Alpha of 10^-4. I would like to run again with smaller alphas to see if there is even better performance.

Ridge Regression: CV vs. Alpha



Concluded that it seemed the best model was the out of the box random forest regression, as this had the highest performance and took shorter to run compared to the models where hyperparameters were optimized. I would be interested to run this model again but with only the most important features. Performed Data quantity assessment to check if we needed to undertake further data collection, it did not seem that more data would be useful. I saved the random forest regression model to make the final predictions for vacancy rate 2019-2020.

2019-2020 VACANCY RATE PREDICTIONS

***NOTE THE FOLLOWING DATA IS NOT REPRESENTATIVE OF ALL US ZIP CODES* b/c Zillow only had ~3000 zipcodes with rent prices vs. ~33k with home prices. Therefore the data is most likely representative of more dense areas due to those areas having Zillow rental price data (ie. cities etc.)**

I loaded the model and refit model with all available data (2014-2018). Then performed 5 fold cross validation on all available data, this yielded a mean test cv score of 0.95 (std: 0.03) and a mean absolute error: 1.15 (std: 0.25). After calculating expected Vacancy Rate for 2019-2020 from the model noted that unique US States in 2019-2020 Vacancy Rate Data was 43 and that the missing states were AK, ME, MT, ND, SD, VT, WV, and WY. **For the reasons noted above further research would be needed to ensure the model was representative of the entire USA. The goal of this project was simply to see if a model that could meaningfully predict vacancy rate by US zip code could be created, which from the outcomes I believe can.**

EXPLORING 2020 VACANCY RATE DATA:

Exploring the data further I looked at **5 highest/lowest 2020 vacancy rate** by:

	Highest	Lowest
zip code	61-67.3% vacancy. Includes tourist areas/destinations in Long Island NY, around Disney World/ Sarasota beaches FL, and Las Vegas NV	1.3-2.3% vacancy. Includes suburban areas outside major urban areas Mahawah, NJ (outside NYC); Austin, TX; San Francisco Bay Area; Philadelphia, PA; Colorado Springs, CO
State	12.1-17.2% vacancy. Includes LA, FL, RI, NV, NY	4.6-5.2% vacancy. Includes UT, MN, CO, NH, ID
County	24-33.5% vacancy. Includes many tourist areas/destinations. Includes counties of **Osceola County (**central Florida near Disney World); **El Dorado County (**in the Sierra Nevada wilderness in CA); **Sarasota County,** FL; **Lee County** (southwest coastal Florida); **Martin County** (east coast of Florida	3.2-3.5% vacancy. Seems to include populous counties and/or relatively wealthy areas outside major cities. **Loudoun County** (wealthy area outside Washington, DC); **Dakota County Dakota County (**outside Minneapolis, MN); **Broomfield County, CO; Manassas Park City** (outside Washington, DC**); Weld County** (outside Denver, CO)
Metro Area	18.1-24.8% vacancy. Includes many areas in FL, may near coastal areas near beaches: **Cape Coral-Fort Myers FL; North Port-Sarasota-Bradenton FL; Lakeland-Winter Haven, FL; New Orleans-Metairie, LA**	3.2-4.4% vacancy. Includes many areas in the west: **Colorado Springs, CO; Provo-Orem, UT; Fort Collins, CO, Stockton-Lodi, CA, Greeley, CO**
City	58.9-67.4% vacancy. Includes many beach areas, especially in Long Island, NY and FL: **Westhampton Beach, NY; Siesta Key, FL, Water Mill, NY, Longboat Key, FL; Town of Shelter Island, NY**	1.8-2.5% vacancy, seems to include (some rapidly growing) suburbs outside major cities: **Chantilly, VA; West Jordan, UT; Burke, VA; Mahwah Township, NJ, Wynnewood, PA**

Results seem to suggest tourist areas near beaches have the highest vacancy rates while wealthier, faster growing suburbs outside major cities have the least vacancy. This seems to make sense generally and also with trends seen during the COVID-19 Pandemic.

I then calculated the rent to price ratio for each zip code. This is a metric that real estate investors use to determine how profitable an area or property may be. To calculate I took each zip code's average gross yearly rent and divided it by the zip code's average home price. Then I calculated rent to price ratios adjusted for vacancy for each zip code, where the gross rent is multiplied by the percent non-vacant of each zip code (1 - our vacancy rate variable) before being divided by average home price. This gives us a more accurate picture of a location or property's potential profitability. I looked at **5 highest/lowest price to rent ratios vacancy adjusted (higher number is generally better)** by:

	Highest	Lowest
zip code	20.5-23.6% rent/price ratio vacancy adjusted. Includes zip codes with homes at $40k or below and high vacancy rates in: Detroit, MI (3 zip codes); Toledo, OH; Jennings, MO	0.88-1.29% rent/price ratio vacancy adjusted. Includes zip codes with homes above $1.5 million and low vacancy rates mostly in CA: Los Angeles, San Mateo, Palo Alto, Newport Beach, CA; Paradise Valley, AZ;
State	8.1-9.7% rent/price ratio vacancy adjusted. Includes: MS, MI, KY, OH, NE	3.4-4.3% rent/price ratio vacancy adjusted. Includes: HI, DC, CA, WA, OR
County	9.7-17.4% rent/price ratio vacancy adjusted. Includes: **Lucas County, OH; Luzerne County, PA; Wayne County, MI; Baltimore City, MD; DeSoto County, MS**	2.1-3.1% rent/price ratio vacancy adjusted. Includes counties ALL in CA: **El Dorado County, CA; Santa Clara County, CA; Marin County, CA; San Francisco County, CA; San Mateo County, CA**
Metro Area	9.1-17.4% rent/price ratio vacancy adjusted. Includes: **Toledo, OH; Scranton--Wilkes-Barre--Hazleton, PA; Dayton, OH; Memphis, TN; Detroit-Warren-Dearborn, MI**	2.9-3.7% rent/price ratio vacancy adjusted. Includes: **Boulder, CO; Los Angeles-Long Beach-Anaheim, CA; Urban Honolulu, HI; San Francisco-Oakland-Hayward, CA; San Jose-Sunnyvale-Santa Clara, CA**
City	17.8-20.5% rent/price ratio vacancy adjusted. Includes: **Jennings, MO; Detroit, MI; Hampton Bays, NY; Northwoods, MO; Park Forest, IL**	0.97-1.6% rent/price ratio vacancy adjusted. Includes **Laguna Beach, CA; Beverly Hills, CA; Los Gatos, CA; Burlingame, CA; Paradise Valley, AZ**

Results seem to show that places that may be good to invest are in cities with low home prices, relatively higher vacancy rates where more suburban places with higher home prices (especially in CA), are not as good potential areas to invest in real estate. *Note: Rent/Price ratios are only one factor when looking for potential areas to invest, one would also want to consider a variety of other factors (ie. unemployment, crime rate, job/population growth, etc)*

Finally I calculated the difference from rent/price and rent/price vacancy adjusted ratios for each zip code. This gave me a variable that showed places in **2020 where Rent/Price Ratios were most impacted by vacancy. Here too I looked at 5 highest/lowest differences by:**

	Highest	Lowest
zip code	11-28.1% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: All 5 zip codes in Long Island, NY	0.07-0.09% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: Zip Codes in San Jose, Oakland, & Burlingame, CA; Colorado Springs, CO; Wynnewood, PA

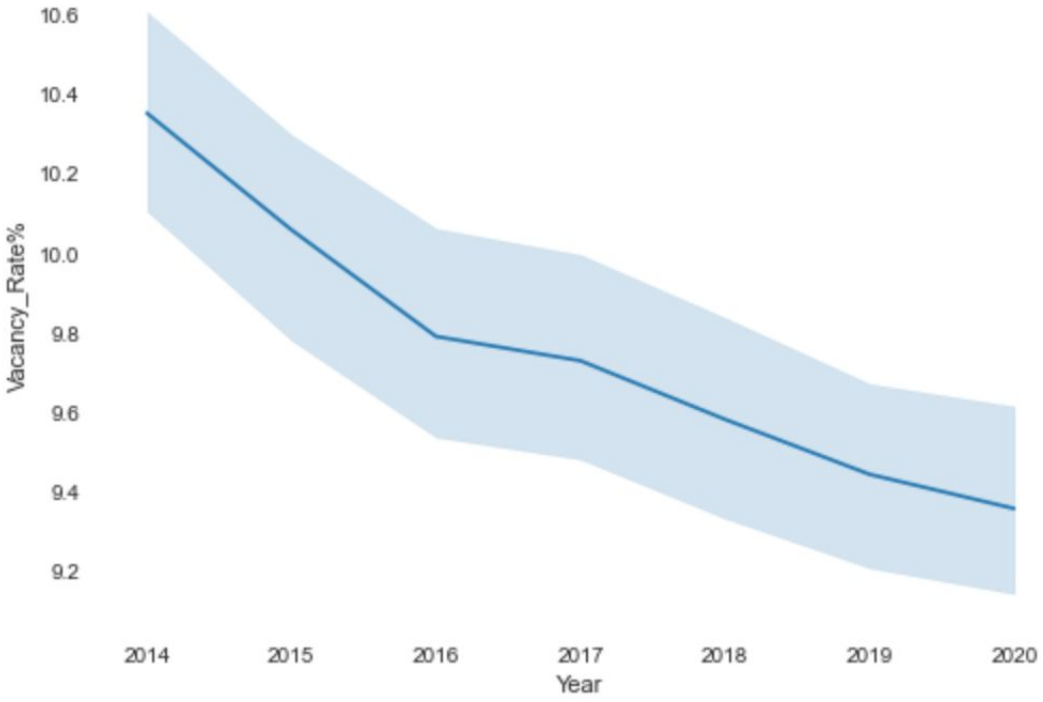
State	1.1-1.3% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: FL, NY, KY, DE, OH	0.22-0.25% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: ID, WA, OR, CO, UT
County	13.0-17.4% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Lucas County, OH; Suffolk County, NY; Baltimore City, MD; Osceola County, FL; Luzerne County, PA;**	2.1-4.8% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Loudoun County, VA; Contra Costa County, CA Broomfield County, CO; Santa Clara County, CA; San Mateo County, CA**
Metro Area	1.76-3.68% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Toledo, OH; Scranton--Wilkes-Barre--Hazleton, PA; Cape Coral-Fort Myers, FL; Dayton, OH; North Port-Sarasota-Bradenton, FL**	0.14-0.19% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Provo-Orem, UT; Fort Collins, CO; Boulder, CO; Greeley, CO; San Jose-Sunnyvale-Santa Clara, CA**
City	11-28.1% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Westhampton Beach, NY; Southampton, NY; East Hampton, NY; Town of Shelter Island, NY**	0.07 -0.1% difference in rent/price ratio and rent/price ratio vacancy adjusted. Includes: **Cupertino, CA; Los Alamitos, CA; Redwood City, CA; Burlingame, CA, Wynnewood, PA**

Seems like zip codes most affected by vacancy may be vacation areas.

EXPLORING 2014-2020 DATA

I then combined the 2014-2018 & 2019-2020 datasets for exploration and plotted average vacancy rate over time and noted the line plot of 2019-2020 predictions seems to roughly follow trend of 2019-2020 national data from FRED - which to me was a good, albeit rough, sign regarding the accuracy/usability of the model results. I then examined average vacancy rate, rent price, and home price vs. years 2014-2020 and plotted the mean and standard deviation of those 3 variables. **Over 2014-2020 the mean of home prices and rent prices increased while vacancy rate decreased. Also the standard deviation of rent prices and vacancy rate decreased while home prices increased. This suggests a trend of less variation from 2014-2020 in rent prices and vacancy rates and more variation in home prices. This seems to make sense in today's current market where in some areas home prices have increased rapidly during covid-19 pandemic and others areas have rapidly decreased.**

Vacancy Rate % by Year



I then examined Rent to Price Ratios over Time (2014-2020) and plotted rent/price ratios by year, plotted rent/price ratios adjusted for vacancy by year, and plotted difference from rent/price and rent/price vacancy adjusted ratios by year. From these plots it **seems that the US as a whole has gotten a harder place to find good investments, even as rent prices have increased and vacancy rates have decreased. Potentially this is due to home prices having increased faster than rent prices. It also seems that vacancy rate has less of an impact on rent/price ratios over 2014-2020 across the US.**

Juptyer Notebooks 5.0-5.2: Attempt to Create Vacancy Rate ARMA Time Series Model

During the EDA I made a line plot of vacancy rate vs. year and noted a large drop in vacancy rate from 2019-2020, this seems to be due that only the national data is there for 2019-2020. A potential better way to show this data could be to use the median vacancy rate as the data for this variable is highly skewed with a lot of outliers. I created a function to visualize the time series and test for stationarity and ran it for vacancy rate, the results of Dickey-Fuller Test suggest that we can reject the null hypothesis that vacancy rate is non-stationary.

During preprocessing, I made a 'Year' column as a datetime index to prepare for time series analysis and performed TimeSeries train test split. Then I established baseline comparisons using dummy regressors. I trained a linear regression model on the training data and this yielded an R2 of .48 on test set, MAE of 8.33 on test set, and a MSE of 153.36 on test set. This was markedly better performance than when using Dummy variable/mean but the model underperforms the other multiple regression models, ie. linear regression, ridge regression, random forest models in notebook 4.3.

For modeling, I computed pct_change and added a constant to the DataFrame for the regression intercept. Then ran an OLS regression with vacancy rate pct_change but the coef and std error were both NaN. I decided to try only vacancy_rate as when I ran the dickey-fuller test before it seemed to already be a stationary object - implying that calculating percent change may not be necessary. So when using Vacancy Rate to compute autocorrelation: 0.2758303730509692 and plotted autocorrelation function and ran the dickey fuller test and got 0.0. I estimated an AR model and estimated parameters from data.

- coef const: 17.8, coef ar.L1.Vacancy_Rate: .28
- std. error const: 0.043, std. error ar.L1.Vacancy_Rate: 0.002
- AIC 2197829.721
- BIC 2197861.161

I decided not to use a time series model, because with an annual time series model for 2011-2018 there are only 7 data points for each zip code. Instead opting to use the multiple linear regression model.

Juptyer Notebooks 6:0 Further research to improve on original model

Note: This notebook is not currently being utilized in the model as is, it was an exploration for further research to gather more variables to make the model more robust and able to predict more zipcodes going forward.

I started wrangling median household income by zip code from ACS data. I compared ACS data median income to the estimate from US national data. Because I was less trustworthy of ACS data and decided to shift the median of ACS data to match that of the US national data from FRED.

I stopped working at this point, the next step would be to join the separate dataframes into one, also I could also get other economic factors by zip code (ie. unemployment)**. The issue is we don't have these economic factors by zip code in the current year, only up to 2018, so these variables would not be helpful in a multiple regression predictive model.**

the findings.

- I found vacancy rates had different trends/correlations between econometric variables depending on the different time periods. This was interesting to see that during different economic periods in US history the variables had very different relationships.
- The best performing model was the Random Forest regressor, with R2 on the test set of .92. However Adjusted R2 on test data showed a score of .78 which seems to suggest some of performance is skewed by the large number of variables, and that some of these variables may be unnecessary to the model. Not only did this model have the highest performance, it also took shorter to run compared to the models where hyperparameters were optimized.
- From the Lasso Regression model it seems that the top 10 features in predicting vacancy rate comprise: zip code, rent price, year, size rank, home price, State(s).
- When examining projected 2020 vacancy rate data, results seem to suggest tourist areas near beaches have the highest vacancy rates while wealthier, faster growing suburbs outside major cities have the least vacancy. This seems to make sense generally and also with trends witnessed during the COVID-19 Pandemic.
- When examining vacancy adjusted rent/price ratios, a metric used for rental real estate investment, results seem to show that places that may be good to invest are in cities with low home prices, relatively higher vacancy rates. More suburban areas with higher home prices (especially in CA), may be less desirable in terms of potential areas to invest in real estate. But real estate investors should also be sure to consider other variables when choosing an area to invest (ie. crime rate, unemployment, population and job growth etc.)

Include ideas for further research

- The current model is only predicting vacancy rates for ~10% of US zipcodes and 43/51 US states, this is because I do not have 2019-2020 rent price data for all the zip codes, only the ~3,000 zip codes from Zilllow. In the future if I wanted to be able to predict vacancy rate for all US zip codes I would either need:
    - to create a model to predict rent prices by zip code (which seems do-able because home prices are so correlated with rent prices, and Zillow has current home price data on around > 30,000 zip codes). This would then allow me to update my current model and predict vacancy rates for the rest of the zip codes.
    - A simple way to preserve more zip codes would be to instead of simply dropping all NaNs could be to group by year and zip code, and then linear fill the data for home/rent prices. This change would add only a few zip codes, not a meaningful amount.
    - Another thought would be to drop rent prices from the model altogether and see how much the model's predictive ability is impacted. If it is not by much then it would be very easy to have have the model predict ~30k zip codes because we have that many variables from zillow's home price data

- Another idea would be to measure the current model's performance's predictions as the real vacancy rate data comes out from ACS, ie. as the ACS releases their 2019 results I could test and see how accurate the 2019 predictions were. As needed I could update the model to improve performance.
- Because it seems that geographical factors are important in the ability to predict a specific zip codes vacancy rate, I could consider creating different models for different regions, ie. northeast, southwest, etc. or whatever divisions would yield the most predictive models.
- I could try running the model with less variables to see if the difference between R2 & adjusted R2 scores gets smaller.
- Also I could go back and look for more data/variables to improve the model. Current economic and housing market data for the US nationally is plentifully available, and I could possibly incorporate that or try to impute data by zip code for some variables (ie. ACS has zip code level data for median income, unemployment from 20111-2018). Possibly adding in some of these variables that we used in the US national model may yield better predictive results.
- It would be interesting to see which areas have had the most vacancy rate change from 2014-2018 (by zip code, county, metro area, and city)
- I could also use the current vacancy rate data I have created in a larger real estate investor dataframe where I have other information by zip code that real estate investors utilize when choosing areas to invest (ie. unemployment, price to rent ratios, crime rates, job growth and diversity, population growth etc.). This could be a useful tool to help explore/identify areas that seem good places for potential investment.

Recommendations on how your client can use your findings:

Background on using vacancy rate data to determine areas for potential real estate investing with a rent/price ratio adjusted for vacancy:

Many times real estate investors will use a rent to price ratio to determine if an area or property may be a good place to invest. They would use the gross annual rent divided by the home price to get this ratio. A higher ratio is better as investors look for higher rents and lower home prices, which generally yield greater rental income. Generally ratios above 12% are good places to do more research and look into.

However if real estate investors only use rent/price ratios they are missing an incredibly important factor: how vacancy eats into their annual rental income. So a more nuanced statistic is rent/price ratio adjusted for vacancy, calculated as such:

((avg. monthly rent * 12) * (1 - vacancy rate))/ avg. home price

This is one of the indicators I would use when selecting zip codes for potential real estate investments. Here is an example of how this statistic can be important:

Two homes A&B:

- Both homes rent for $1000/month
- Home A sale price at $125k (9.6% rent/price ratio)
- Home B sale price at $100k (12% rent/price ratio)

At this point if I stopped here I would go for Home B as it should give higher returns on my investment, but looking at average vacancy rates in the area:

- Home A vacancy rate - 5% (now a 9.12% ratio adjusted for vacancy)
- Home B vacancy rate - 35% (now a 7.8% rent/price ratio adjusted for vacancy)

Now I would choose Home B and look deeper into why the vacancy rate in Home B's neighborhood is so high.
Some times higher vacancy may mean higher crime rate, higher turnover in renters, more potential for vandalism (ie. higher home insurance premiums), and more truly vacant (abandoned or otherwise) properties in the area.

1. Places to potentially invest

a. States: Mississippi, Michigan, Kentucky, Ohio, Nebraska

b. Counties: Lucas County, OH; Luzerne County, PA; Wayne County, MI; Baltimore City, MD; DeSoto County, MS
c. Cities: Jennings, MO; Detroit, MI; Hampton Bays, NY; Northwoods, MO; Park Forest, IL

2. Places to you may to avoid when investing

a. States: Hawaii; Washington, DC; California; Washington; Oregon
b. Counties: El Dorado County, CA; Santa Clara County, CA; Marin County, CA; San Francisco County, CA; San Mateo County, CA
c. Cities: Cupertino, CA; Los Alamitos, CA; Redwood City, CA; Burlingame, CA, Wynnewood, PA

As noted above, data seems to show that places that may be good to invest are in cities with low home prices, relatively higher vacancy rates. Conversely, areas that are more suburban with higher home prices (especially in CA), seem to be not as good potential areas to invest in real estate.

Note: Rent/Price ratios are only one factor when looking for potential areas to invest, one would also want to consider a variety of other factors (ie. unemployment, crime rate, job/population growth, etc)

3. Use the current predicted vacancy rate data in a larger real estate investor dataframe with other information by zip code to create an "investability" metric for each zip code.

Rent prices, home prices, and vacancy rates would be a few of the data points you would want to collect when deciding to invest in an area. Other variables could include:

- crime rates
- unemployment rates
- variation in job types (higher variation shows job security in a market)
- avg. yearly rent/median income
- Population growth, etc.

You could combine and weight these features to develop some type of single "investability" metric for each zip code that aggregates all your individual variables. For example you could scale all the variables and then add them together (while giving the more important metrics higher weights). So if you believe rent/price ratio adjusted for vacancy and crime rate are more important you may choose to weigh these higher than unemployment. This would then give you a quick and easy to access database that would give you the best potential areas to invest but also the background statistics that make up that larger "investability" metric. You could also have it automatically update as this data came in from zillow etc. This could save considerable time and money, and generate considerable potential profits, while allowing investors to expand their real estate portfolio across the USA.