Coronavirus Impacts on Housing Market:
- Predicting Median Ppsf

•••

September, 2020

Overview of Assignment

- Dataset of interest.
  - Redfin Data Center - Public Housing Data from the Real Estate consumer website, Redfin.
    - Import Coronavirus Case History for data calculations and projections.
  - Predict Median Price per square feet
- Hypothesis: There is no significant relationship between Median Price per Square Feet and COVID total cases data
- Explore the data
- Model your outcome of interest.

Capstone objective:

Predict Median Price per square feet - There is heightened talk regarding Real Estate and the effects that COVID is taking on the market. New homeowners are facing obstacles getting a mortgage, selling their home, and worrying about all the unknowns during this time. My objective is to provide a resource to Buyers and Sellers showing what the future holds regarding the Median Ppsf.

Main Features:

Housing Data

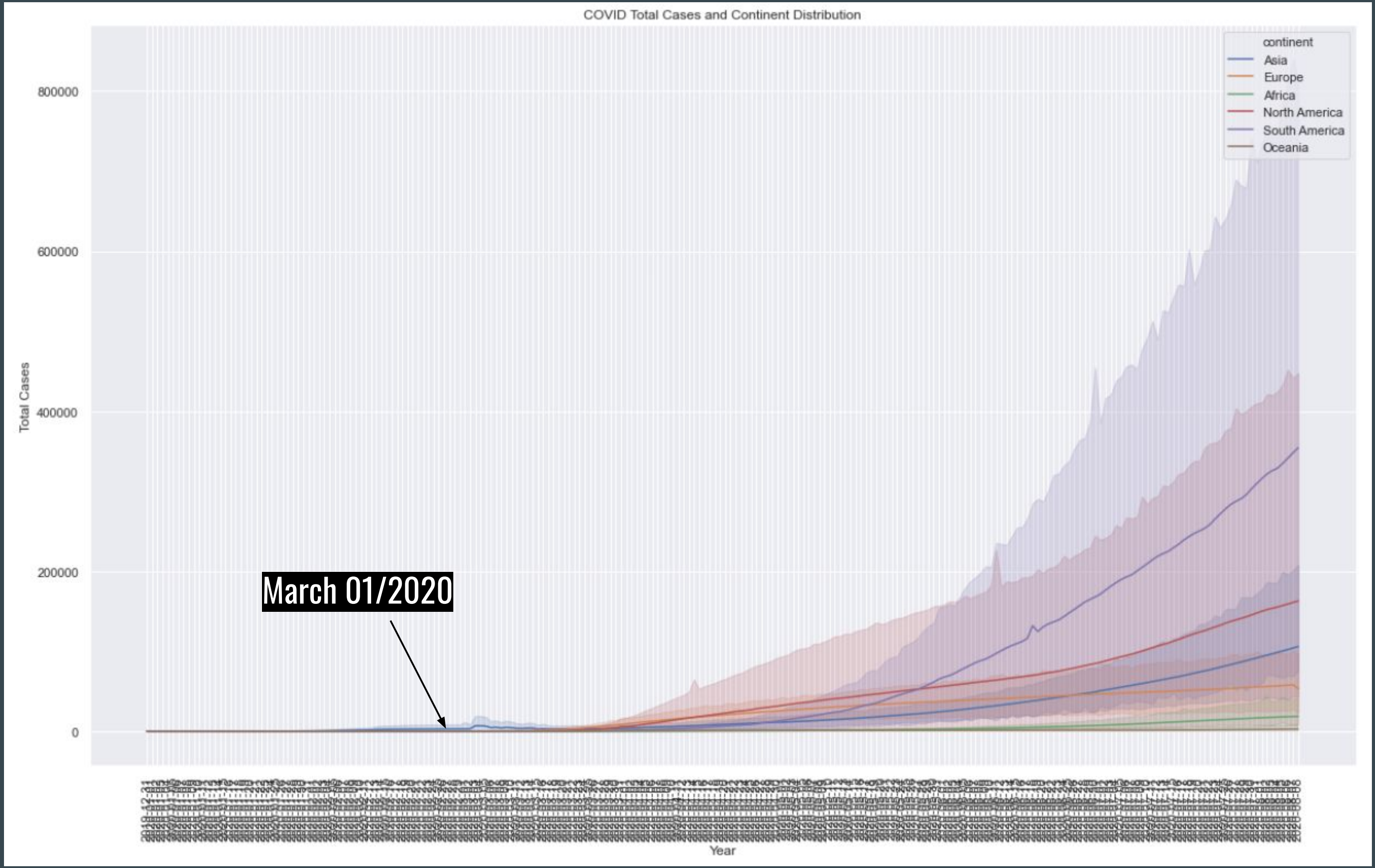
- Median Ppsf
- Median Sale Price
- Home Sales
- New Listings
- Inventory
- Active Listings
- Months Supply
- Days on Market
- Price Drops
- Sale-to-List ratio

COVID Data

- Location
- Total Cases
- Total Deaths
- New Tests
- Positive Rate
- Median Age

Exploration:

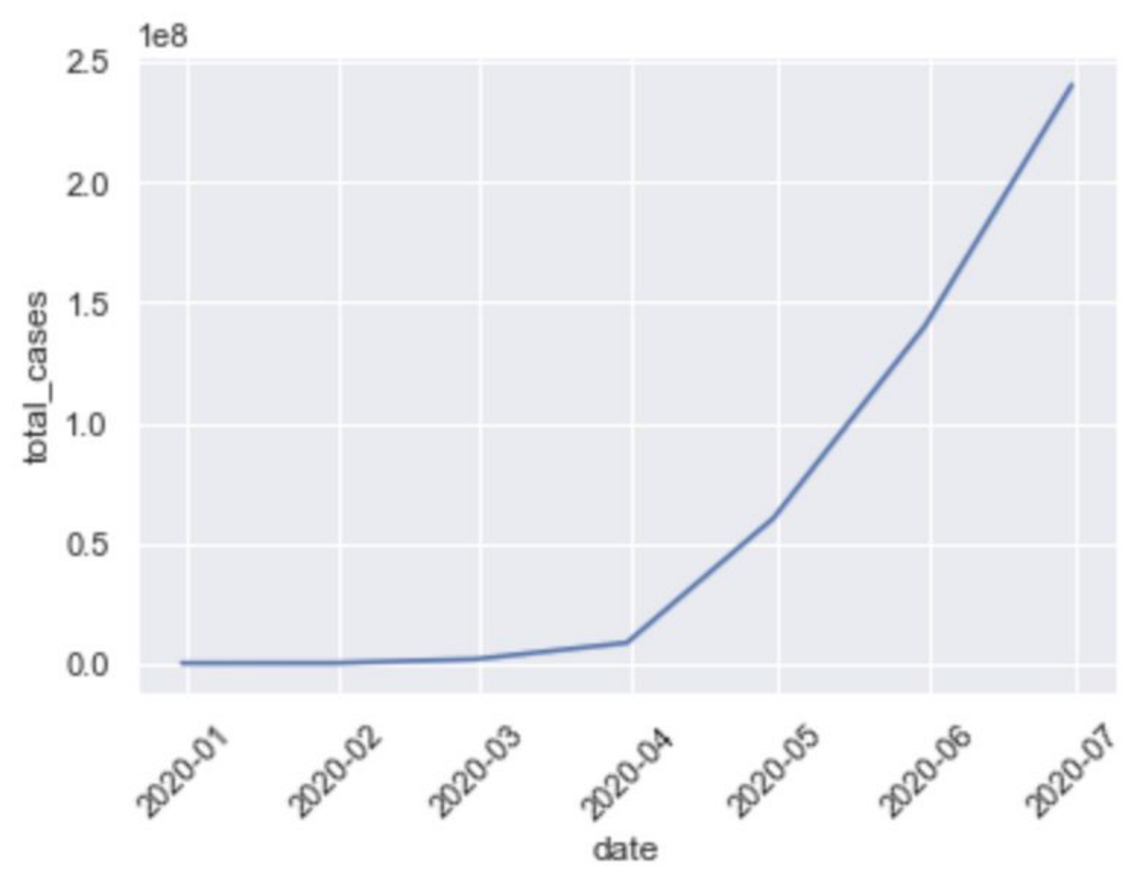
COVID Total Cases and Continent Distribution



March 01/2020

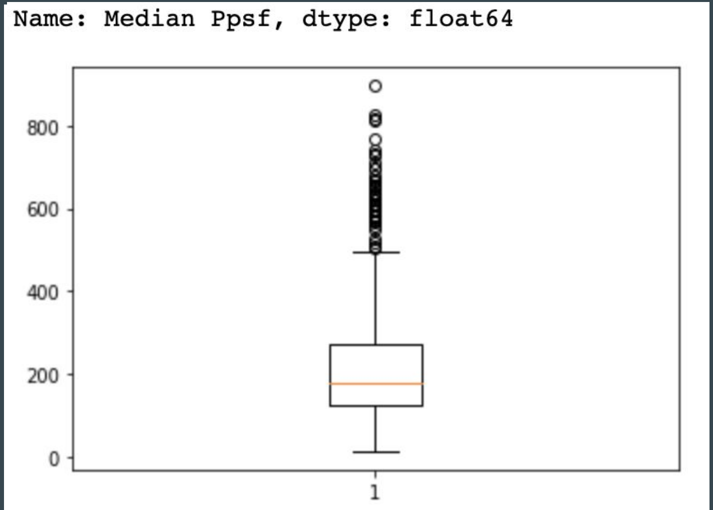
Line Plot of Total Cases of COVID in the World

- There are no signs of the disease slowing down when looking at total cases and date.

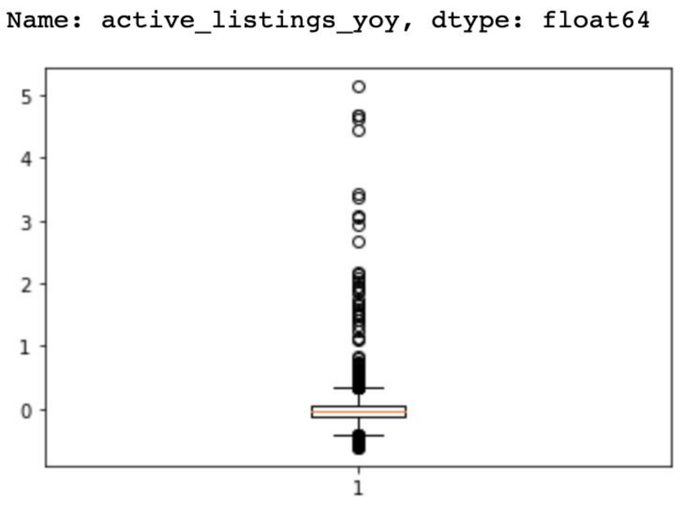


Check Variables for Outliers via Boxplot

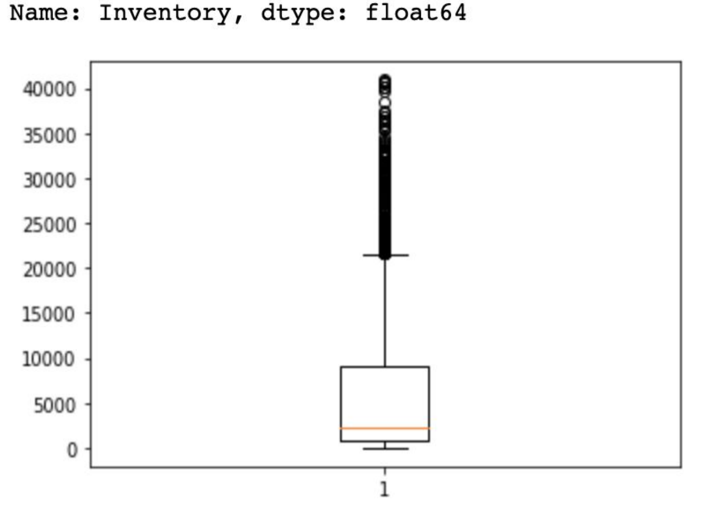
- Outliers represent a real phenomenon within the data set.
- Particular major markets contains this data and although it looks like an outlier when all the data is combined, it's not an outlier within its own markets.



```
Name: Median Ppsf, dtype: float64
```



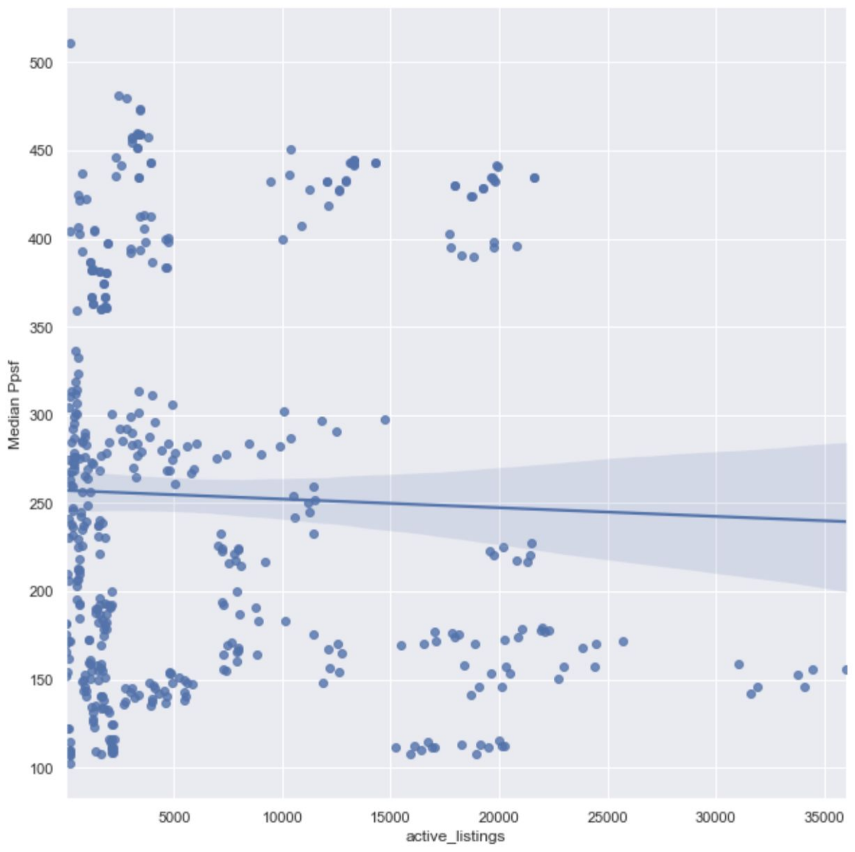
```
Name: active_listings_yoy, dtype: float64
```



```
Name: Inventory, dtype: float64
```

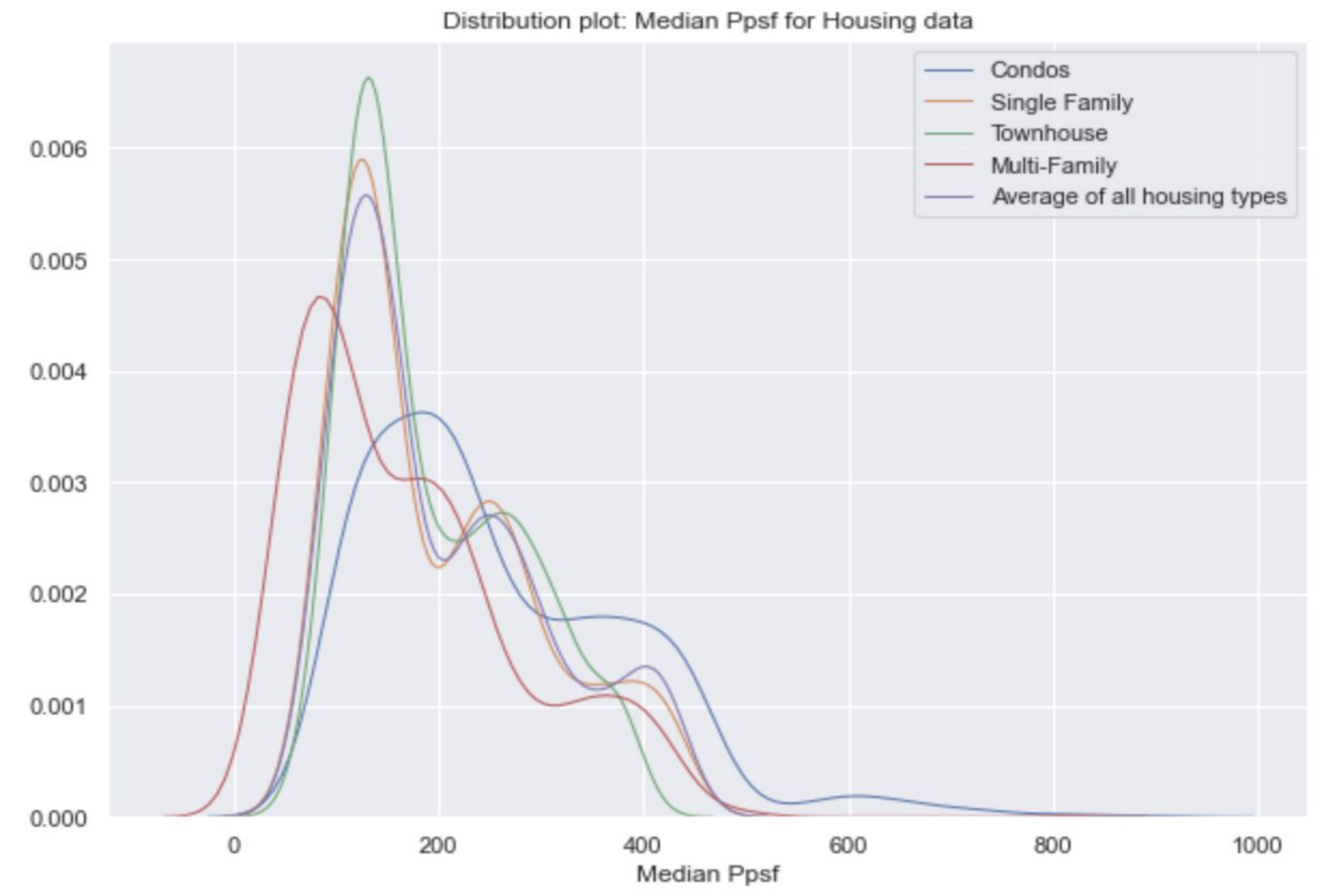
Look at relationships between Median Ppsf and Active Listings

- With lower Active Listings, the Median Ppsf increases. Supply and Demand

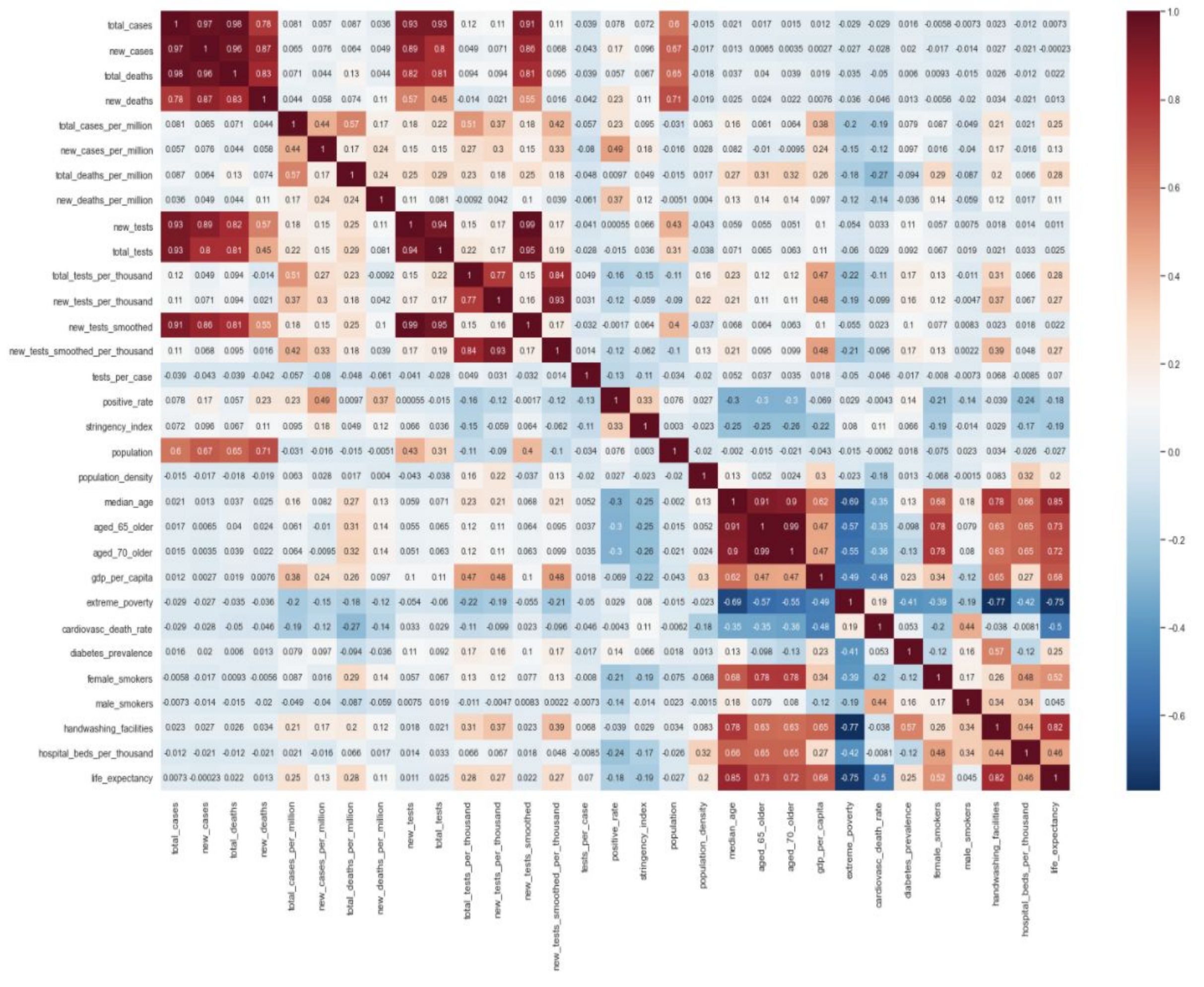


Median Ppsf
Dist. Plot

- Housing Types are fairly similar
- Condos have highest Median Ppsf.

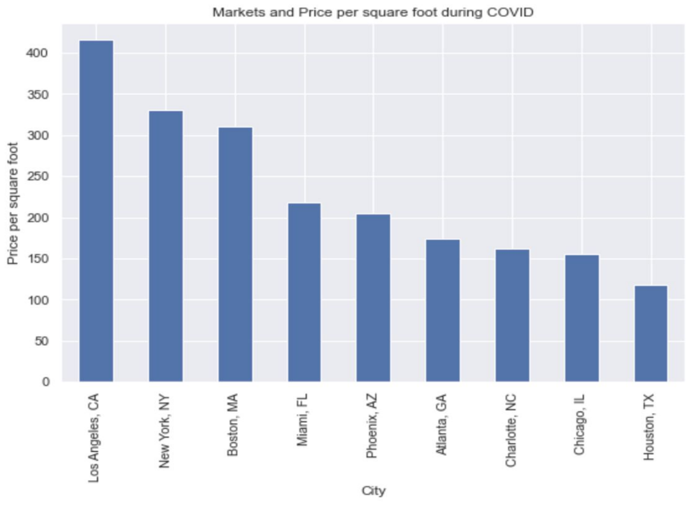


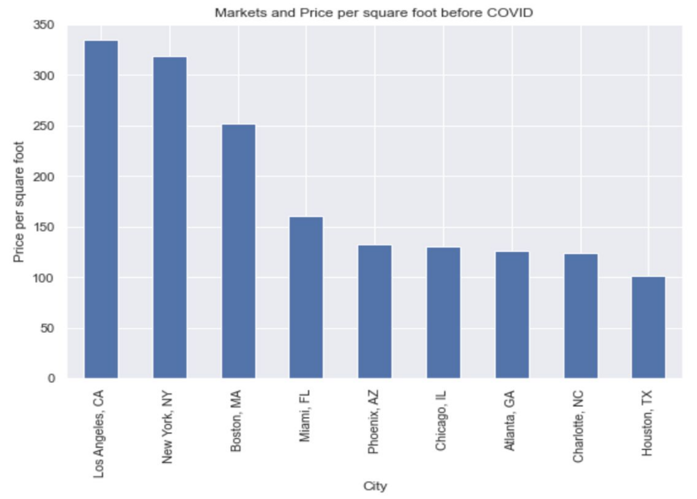
Strong Correlation between multiple features



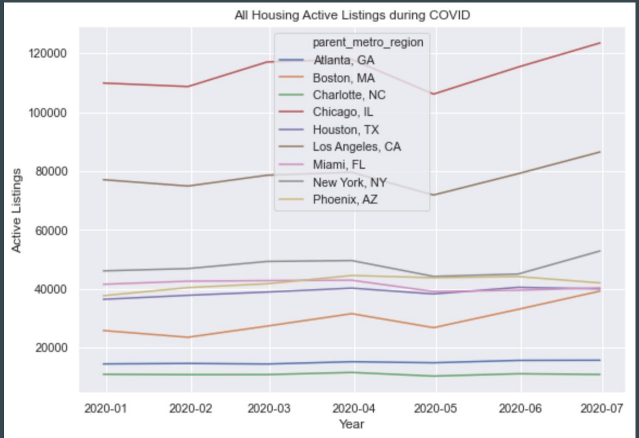
Major Markets and Median Ppsf

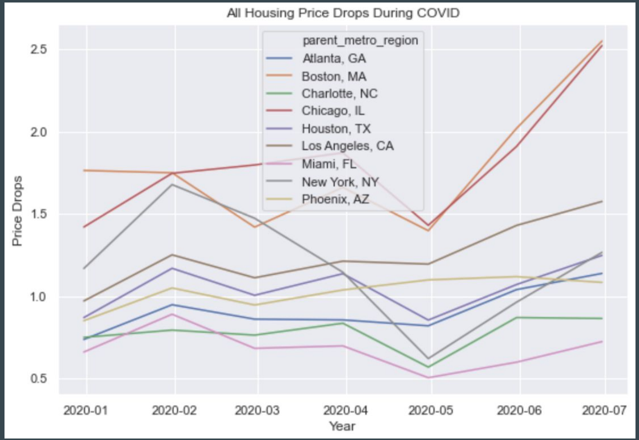
- In all major market cities I gathered, except New York, NY, the Median Ppsf increased during COVID.
- A reason for this is that New York, NY is known for being heavily Condos based, There is a possibility that there is less of a demand for housing now that Stay at Home order is in affect and people can work from home. The majority of Buyers are looking for space.

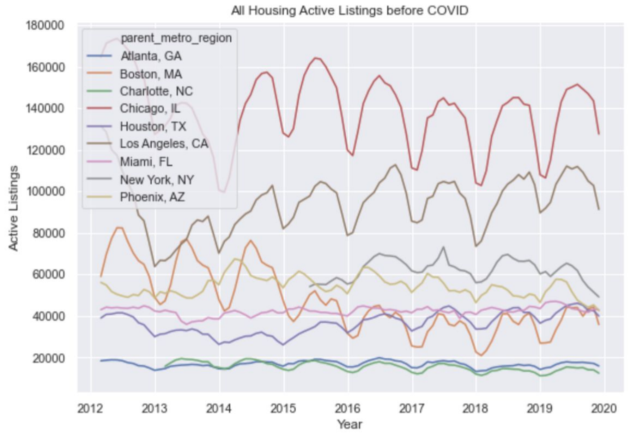


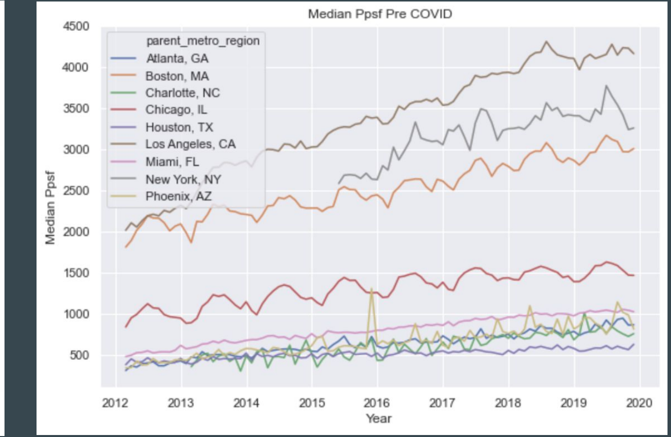


Line Plot of Major Markets and Features



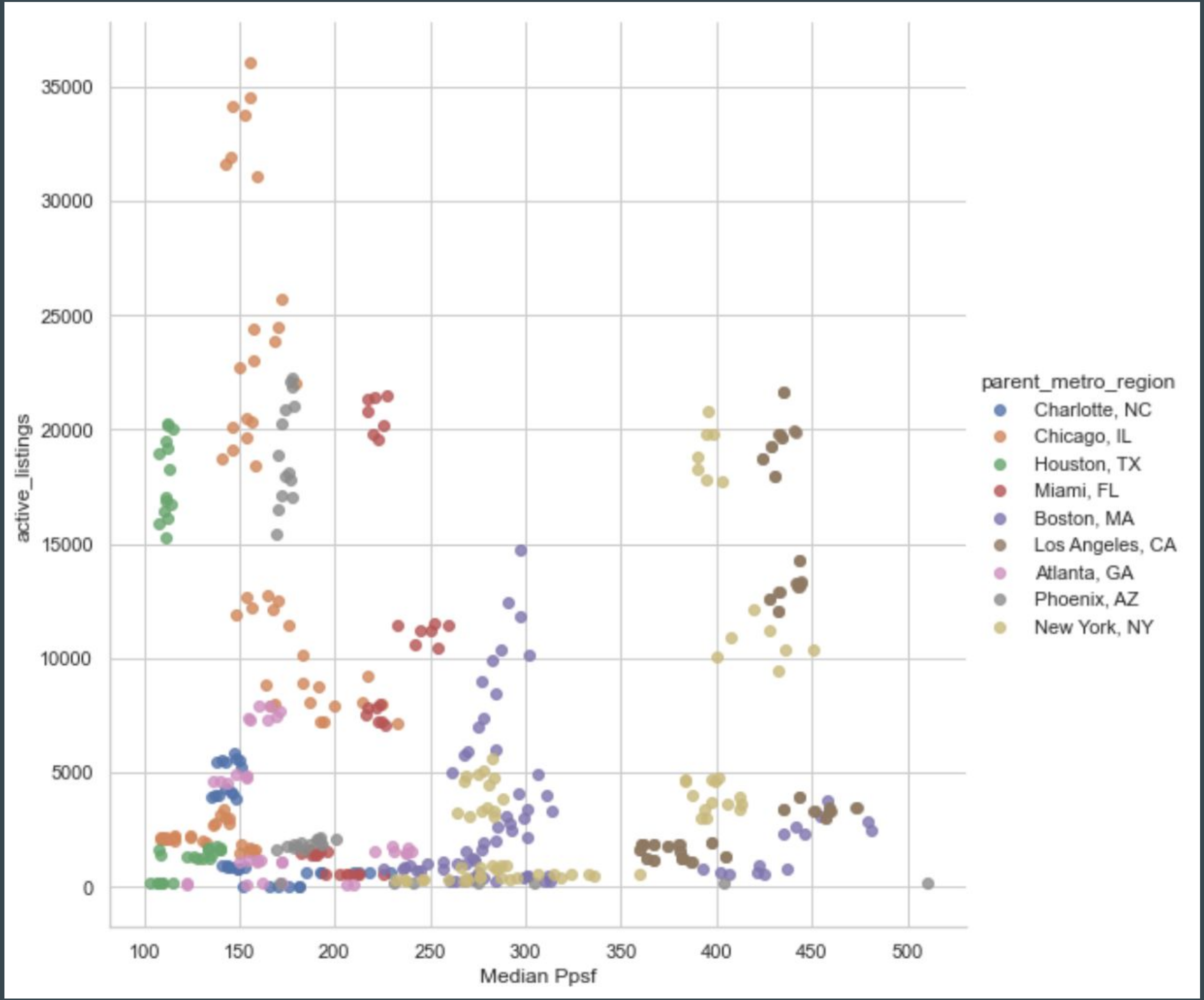




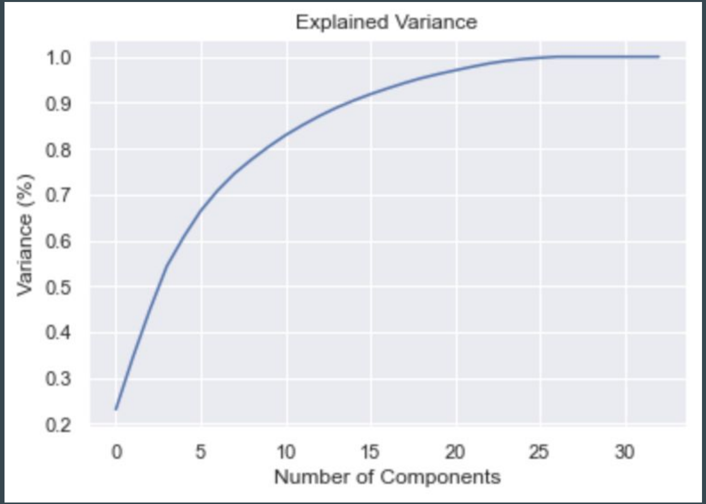


Major Market and Median Ppsf during COVID

- Many of the major markets haven’t fluctuated much in Median Ppsf during COVID.
- The largest difference in Median Ppsf is Boston jumping from 225 ppsf to 475 ppsf

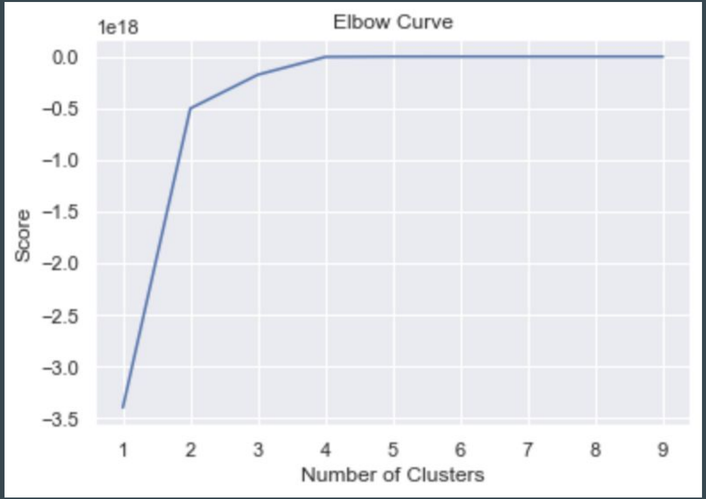


Machine Learning Models:



Model Preparation

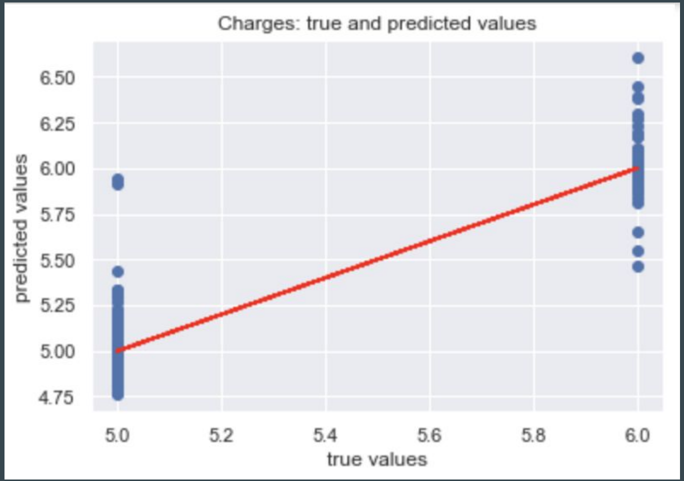
- 20 components explain 95% of the variance
- Elbow plot shows between 2 clusters will yield best results



Linear Regression

- Looking at the graph, the model doesn’t predict values very well visually.

- Looking at the F-statistic and R-squared, the features add information to the model

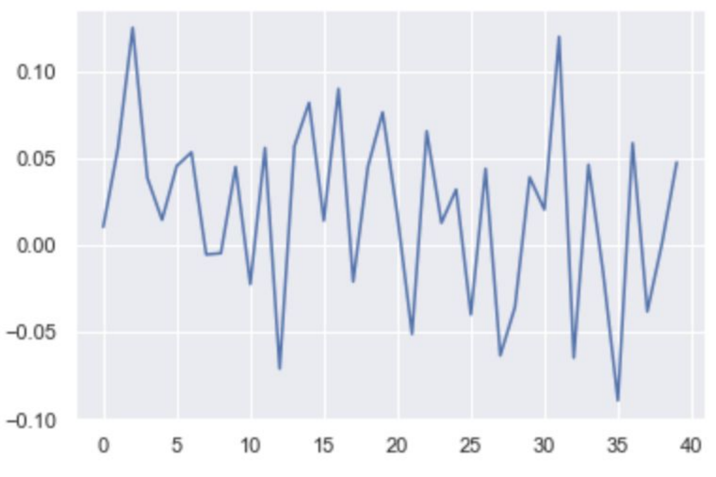


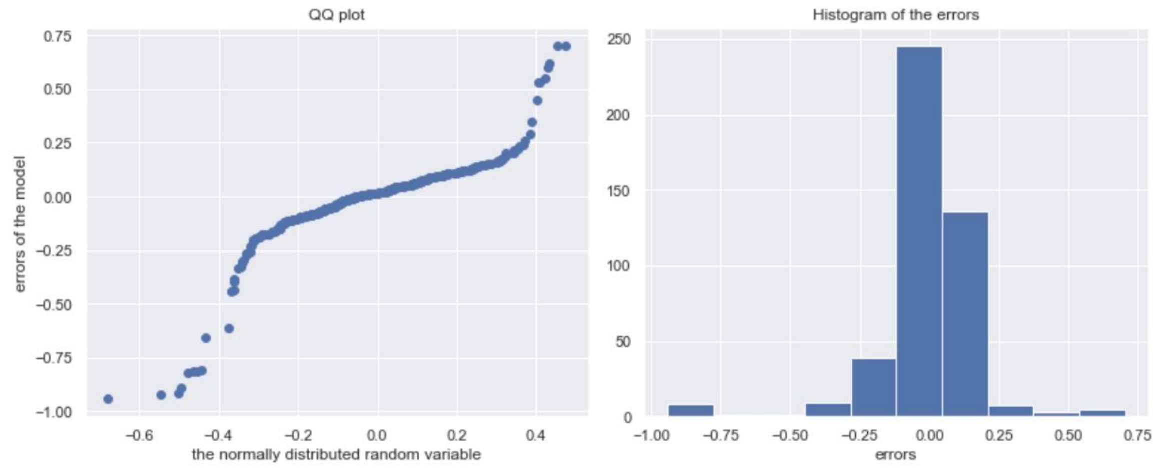
R-Squared Training Set	0.889
F-statistic	65.12
P-Value	3.78e-122
R-Squared Test Set	0.809

Linear Regression

- Autocorrelation between errors of Median Ppsf ranges between -0.10 and 0.10

- Errors are not normally distributed

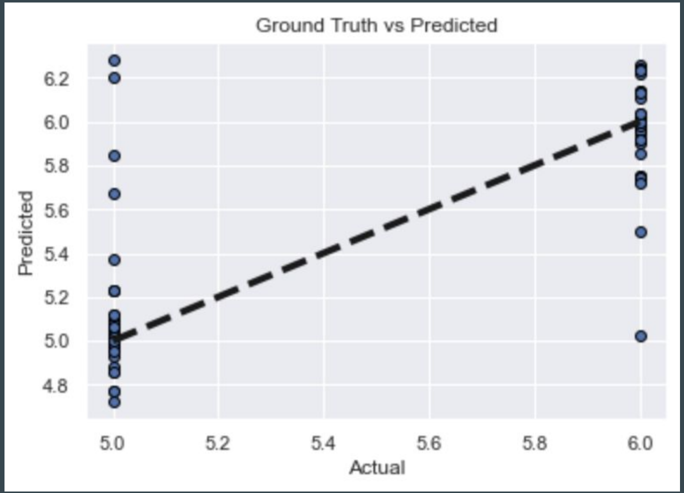




Gradient Boost with Parameter Tuning

- Looking at the graph, the model doesn't predict values very well visually.

- Looking at the R-squared, Gradient Boost did a promising job predicting the model.



R-Squared Training Set	0.999
R-Squared Test Set	0.823
MAE	0.100

Models Summary

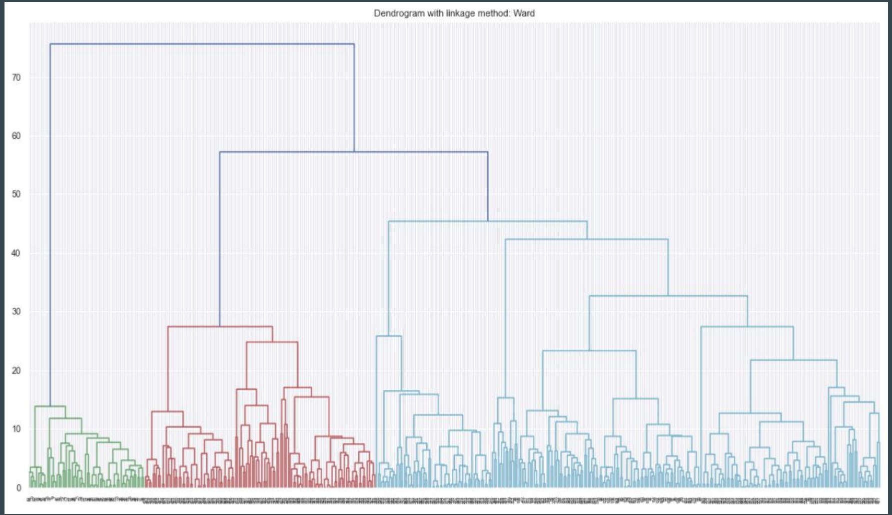
Model	R-squared Training set	R-squared Test set
Decision Tree Regressor	1.000	0.786
Random Forest Regressor	0.971	0.853
Random Forest w. Normalization	0.993	0.870
Gradient Boost	0.999	0.823
Linear Regression	0.889	0.810

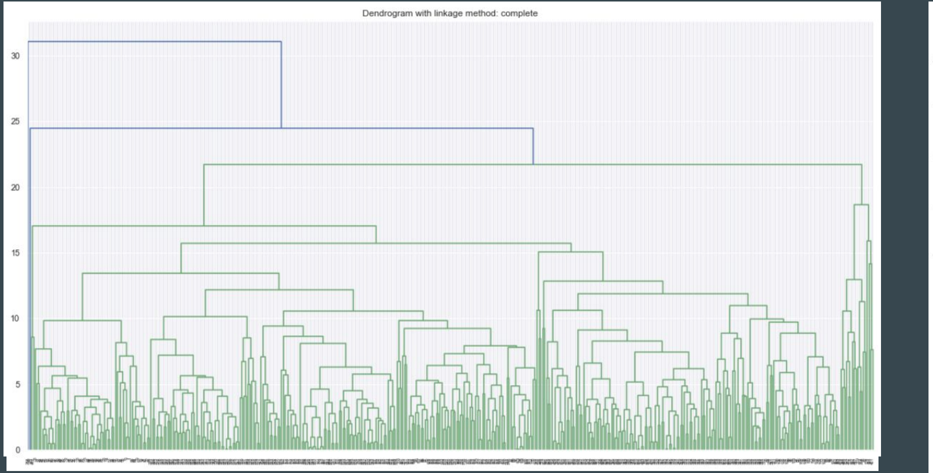
Unsupervised Learning

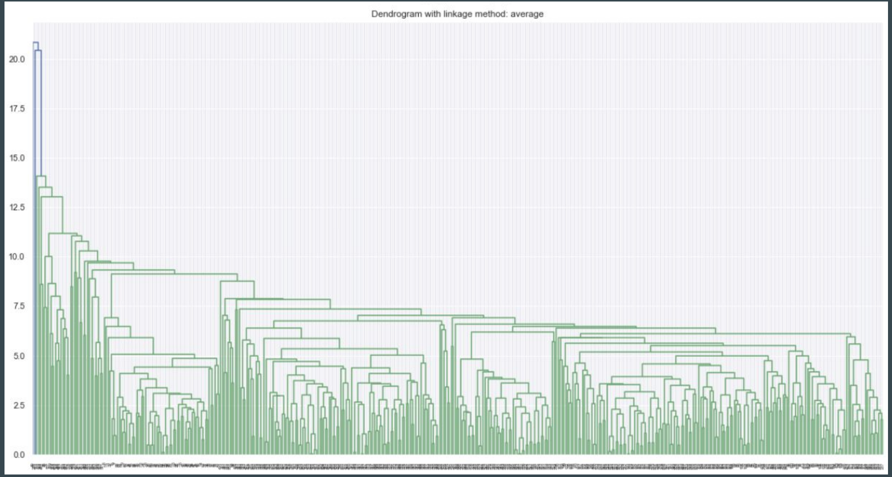
Dendrogram with various linkage methods

- hierarchical relationship between objects

- Ward is the most uniform and nodes are best distinguishable

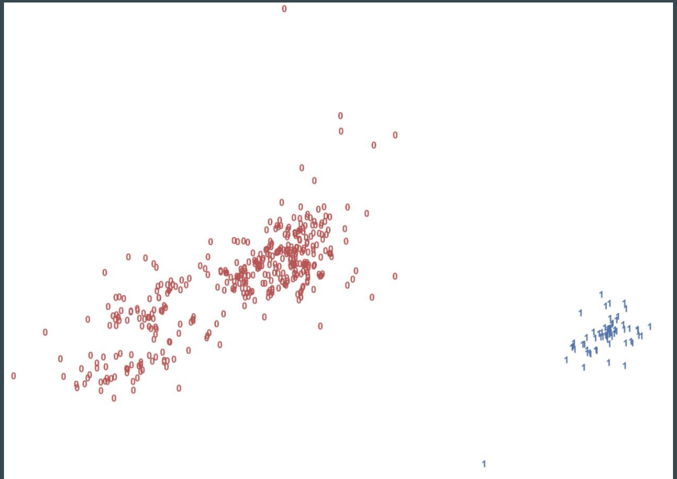






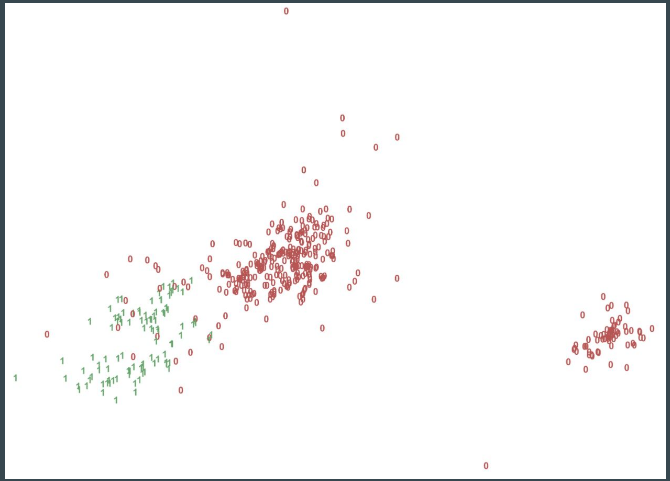
KMeans

- Silhouette Score of 0.31. Closer to similar data points in same cluster than other clusters.



Gaussian Mixture

- Silhouette Score of 0.11. KMeans does a better job clustering than Gaussian Mixture.



Select K Best

- The 3 most important features (using Select K Best) in predicting Median Ppsf::
  - Avg Sale To List
  - Price Drops
  - Price Drops Yoy

Results

- Random Forest, Decision Tree and Gradient boost performed the best predicting Median Ppsf amongst the other models used on the baseline parameters..
- The models know the training set very well, which could be a sign of slight overfitting.
- Using the models with PCA did not have a significant difference in accuracy scores

Real Estate Facts:

- There are 44 million renter-occupied homes in the US to 75 million owner-occupied homes.
- In 2007, around two-thirds of investors were primarily focused on the stock market. That number has fallen to 50%, with many Millennials choosing to invest in real estate instead.
- The short term rental market was valued at $167.9 billion in revenue in 2019.
- A Zillow survey of real estate experts and economists, half are expecting a recession in 2020.
- By 2025, Millennials are expected to form more than 20 million new households.

Conclusion:

- Opportunity to explore data trends, that are happening now.
- People are moving out of cities (Condos) due to job losses, relocation, low mortgage rate opportunities and for a change in lifestyle.
- Clustering data was not very informative.
- Results of a significant relationship between Median Ppsf and COVID were promising, so the alternative hypothesis is accepted.
- Analyzing trends and modeling the distribution greatly benefits new homeowners, sellers, investors and lenders alike.

Future Application and Research

- Gather additional Housing data such as:
  - Bedroom count
  - Bathroom count
  - Condition of home
  - Sqft of home
- Gather data from Mortgage rates
- Further monitor COVID trends..
- Use Deep Learning on a large dataset.
- Update every few weeks with the latest datasets on Redfin
- Pick specific city and run models for more specific Median Ppsf.

Questions?