Term Project Write-up

Felipe Rodriguez

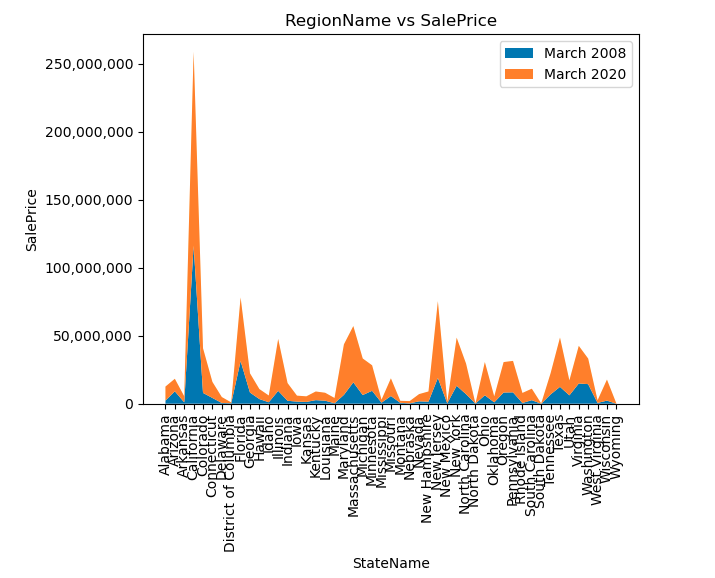
DSC 550 Data Mining

Dr. Brett Werner

August 12, 2023

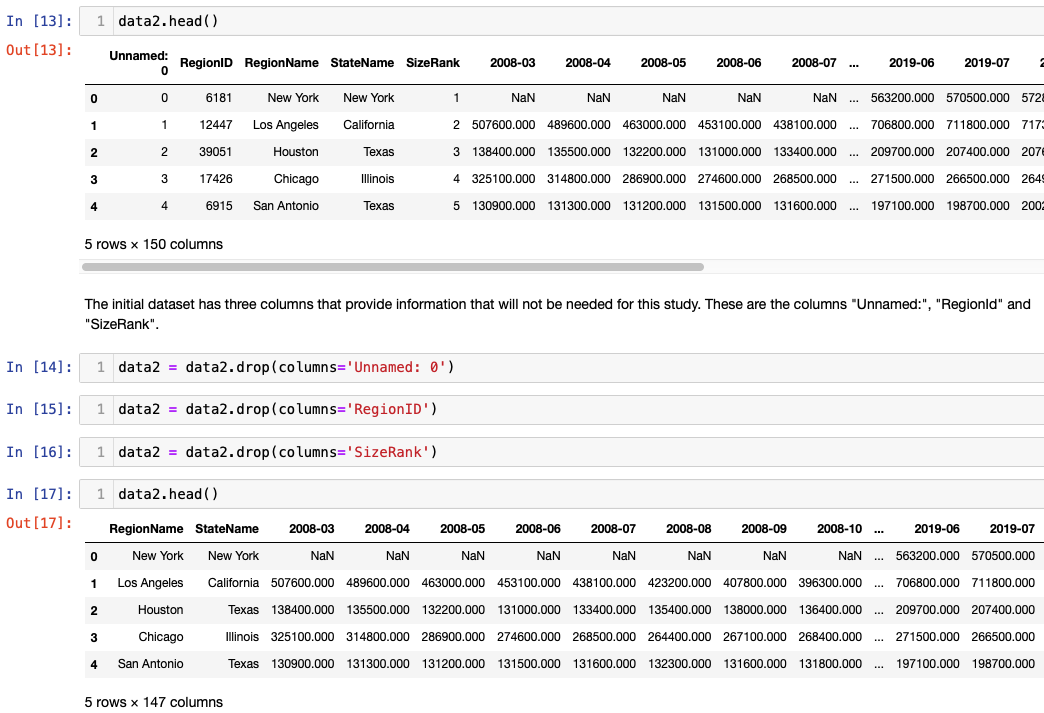
The housing market has always been in a constant state of change and hard to predict. Using housing data from Zillow, an analysis on sales over the last 10 years can be conducted to see if there can be any identification of states that are increasing in price. More importantly, the analysis can include the trends over time. The problem this analysis will solve is to research the trends of sales throughout time in different states, to see where properties are the highest valued over time and how time has affected the sale prices. By analyzing these, investment opportunities can be determined to generate the most profit. The data was obtained from Kaggle which included data from Zillow from 2008 to 2020.

In Milestone 1, an overall distributional analysis was conducted on the data. This involved creating plots that displayed a distribution of the different states and their sales. The data was grouped by State and the Sales were aggregated. An example of a plot created is the stack plot of Sales and State.

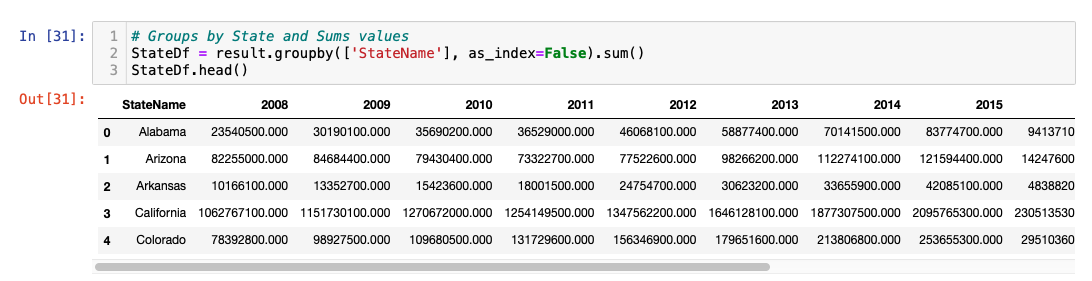


This milestone showed that the data indicted an increase over time, justifying the continuation of the project as well as the need for additional data cleansing and modeling.

Milestone 2 was a data cleansing stage and preparation for a model. Most of the work in this Milestone was transformation and creating new features. A few important pieces were removing unwanted elements from the data set that would not be needed in modeling.



The data also needed to be grouped and aggregated by state so that predictions can be made over time. Grouping the data by year, allows for a more condensed view as a consistent variable for modeling.



Milestone 3 created the model. The model selected was a liner regression model. This method was being used because it can allow us to uncover patterns and relationships of the data. By exploring the different states, we can see how the predictions trend over time and what assumptions can be made of future data. The model was built by using Year and State as feature and Sales as the target. This data was not split. The reason the data was not split is because when analyzing the predictions over actual values, there was the possibility of null values since the splitting involves taking random values from the sets. The predictions from the model and the actual values were graphed by state to see where the predictions were closest to the actual data. A group of graphs showing different types of data

Description automatically generated with medium confidence

The model provided 90% accuracy. This gives confidence in the model and allows for further exploration to continue. It is clearly seen that most states increased in sales as time progressed. The next steps are to identify which states the predictions fit the closest to, for example Georgia and Minnesota. By identifying these, more research can be conducted on the individual states and areas can be identified where the model fits the best. Although Linear Regression was used for our modeling, other methods can be explored as well. The recommendations are to continue to observe the states in that match the predictions versus actual values. Based on the predictions, it can also be recommended to exploring investing in real estate in these states since the model accuracy was so high.