Term Project Write-up

Felipe Rodriguez

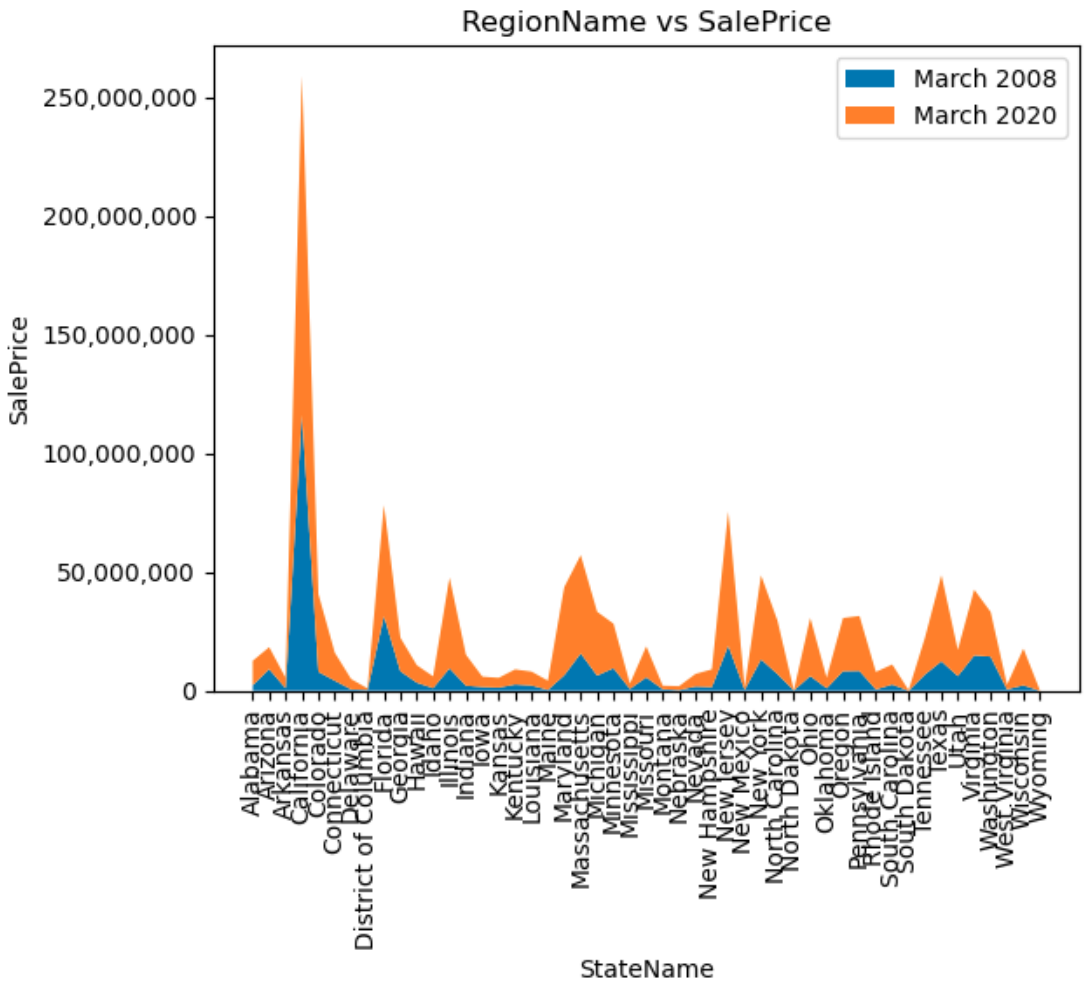
DSC 550 Data Mining

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.

```
In [13]:
1 data2.head()
```

Out[13]:

	Unnamed: 0	RegionID	RegionName	StateName	SizeRank	2008-03	2008-04	2008-05	2008-06	2008-07	...	2019-06	2019-07	2
0	0	6181	New York	New York	1	NaN	NaN	NaN	NaN	NaN	...	563200.000	570500.000	572
1	1	12447	Los Angeles	California	2	507600.000	489600.000	463000.000	453100.000	438100.000	...	706800.000	711800.000	717
2	2	39051	Houston	Texas	3	138400.000	135500.000	132200.000	131000.000	133400.000	...	209700.000	207400.000	207
3	3	17426	Chicago	Illinois	4	325100.000	314800.000	286900.000	274600.000	268500.000	...	271500.000	266500.000	264
4	4	6915	San Antonio	Texas	5	130900.000	131300.000	131200.000	131500.000	131600.000	...	197100.000	198700.000	202

5 rows × 150 columns

The initial dataset has three columns that provide information that will not be needed for this study. These are the columns "Unnamed:", "RegionId" and "SizeRank".

```
In [14]:
1 data2 = data2.drop(columns='Unnamed: 0')
```

```
In [15]:
1 data2 = data2.drop(columns='RegionID')
```

```
In [16]:
1 data2 = data2.drop(columns='SizeRank')
```

```
In [17]:
1 data2.head()
```

Out[17]:

	RegionName	StateName	2008-03	2008-04	2008-05	2008-06	2008-07	2008-08	2008-09	2008-10	...	2019-06	2019-07
0	New York	New York	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	563200.000	570500.000
1	Los Angeles	California	507600.000	489600.000	463000.000	453100.000	438100.000	423200.000	407800.000	396300.000	...	706800.000	711800.000
2	Houston	Texas	138400.000	135500.000	132200.000	131000.000	133400.000	135400.000	138000.000	136400.000	...	209700.000	207400.000
3	Chicago	Illinois	325100.000	314800.000	286900.000	274600.000	268500.000	264400.000	267100.000	268400.000	...	271500.000	266500.000
4	San Antonio	Texas	130900.000	131300.000	131200.000	131500.000	131600.000	132300.000	131600.000	131800.000	...	197100.000	198700.000

5 rows × 147 columns

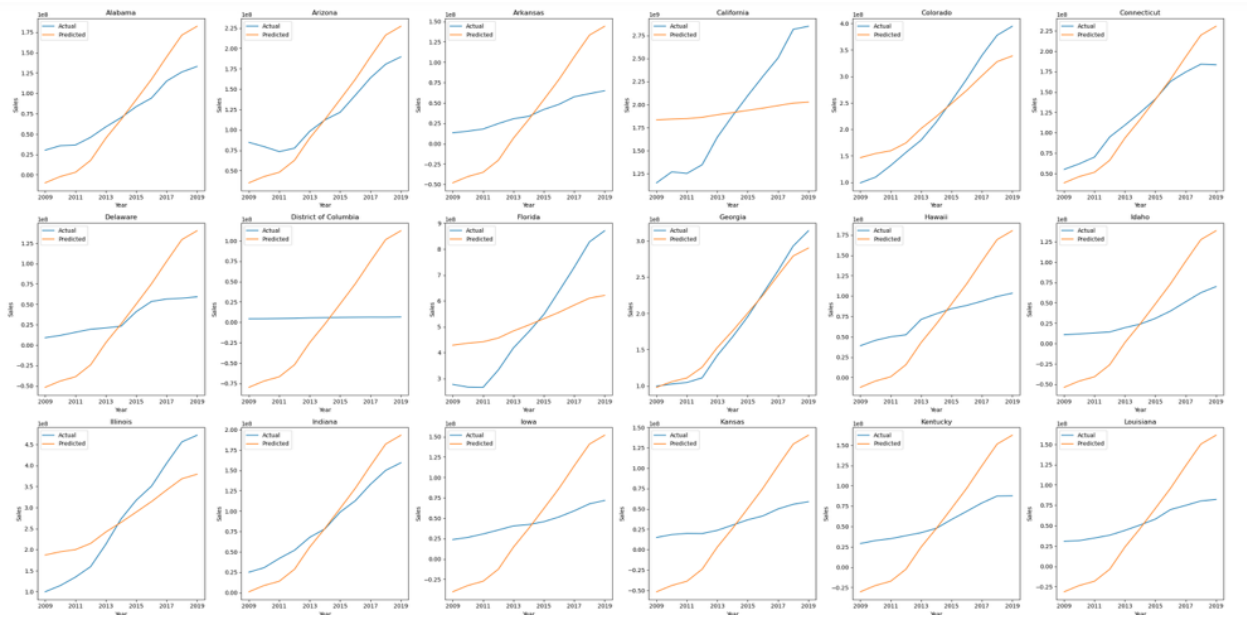
The data also needed to be grouped and aggregated by state so that predictions can be made over time. Grouping the data by year, allows for a more condensed view as a consistent variable for modeling.

```
In [31]:
1 # Groups by State and Sums values
2 StateDf = result.groupby(['StateName'], as_index=False).sum()
3 StateDf.head()
```

Out[31]:

	StateName	2008	2009	2010	2011	2012	2013	2014	2015	
0	Alabama	23540500.000	30190100.000	35690200.000	36529000.000	46068100.000	58877400.000	70141500.000	83774700.000	9413710
1	Arizona	82255000.000	84684400.000	79430400.000	73322700.000	77522600.000	98266200.000	112274100.000	121594400.000	14247600
2	Arkansas	10166100.000	13352700.000	15423600.000	18001500.000	24754700.000	30623200.000	33655900.000	42085100.000	4838820
3	California	1062767100.000	1151730100.000	1270672000.000	1254149500.000	1347562200.000	1646128100.000	1877307500.000	2095765300.000	230513530
4	Colorado	78392800.000	98927500.000	109680500.000	131729600.000	156346900.000	179651600.000	213806800.000	253655300.000	29510360

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.



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.