Milestone 1

July 9, 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 cities and states, to see where properties are the highest valued over time and how time has affected the sale prices.

[1]:
```
import pandas as pd
```

[4]:
```
# Read data
data = pd.read_csv('Sale_Prices_City.csv')
```

Here we can see a sample of the data.

[33]:
```
data.head()
```

[33]:

	Unnamed: 0	RegionID	RegionName	StateName	SizeRank	2008-03	\
0	0	6181	New York	New York	1	NaN	
1	1	12447	Los Angeles	California	2	507600.0	
2	2	39051	Houston	Texas	3	138400.0	
3	3	17426	Chicago	Illinois	4	325100.0	
4	4	6915	San Antonio	Texas	5	130900.0	

	2008-04	2008-05	2008-06	2008-07	…	2019-06	2019-07	2019-08	\
0	NaN	NaN	NaN	NaN	…	563200.0	570500.0	572800.0	
1	489600.0	463000.0	453100.0	438100.0	…	706800.0	711800.0	717300.0	
2	135500.0	132200.0	131000.0	133400.0	…	209700.0	207400.0	207600.0	
3	314800.0	286900.0	274600.0	268500.0	…	271500.0	266500.0	264900.0	
4	131300.0	131200.0	131500.0	131600.0	…	197100.0	198700.0	200200.0	

	2019-09	2019-10	2019-11	2019-12	2020-01	2020-02	2020-03
0	569900.0	560800.0	571500.0	575100.0	571700.0	568300.0	573600.0
1	714100.0	711900.0	718400.0	727100.0	738200.0	760200.0	NaN
2	207000.0	211400.0	211500.0	217700.0	219200.0	223800.0	NaN
3	265000.0	264100.0	264300.0	270000.0	281400.0	302900.0	309200.0
4	200800.0	203400.0	203800.0	205400.0	205400.0	208300.0	NaN

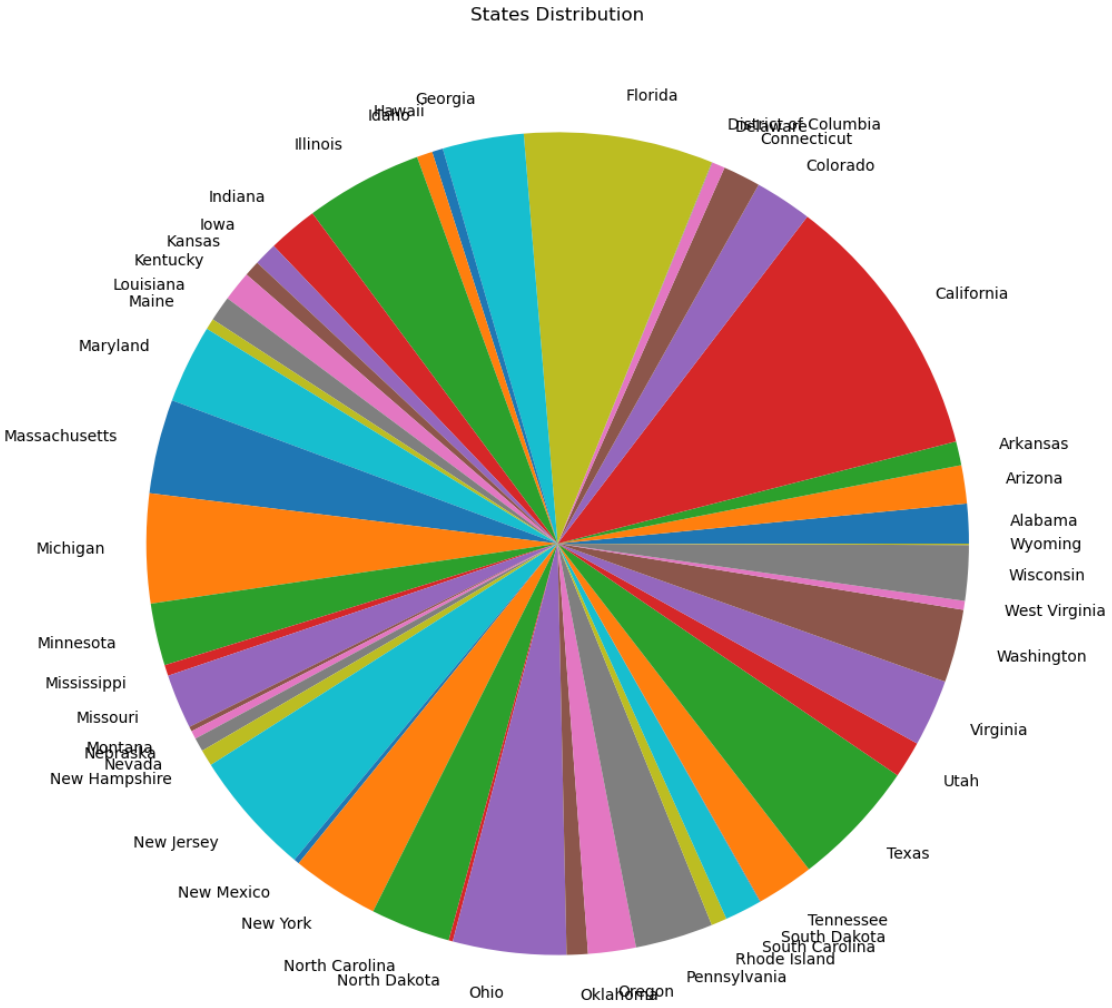
[5 rows x 150 columns]

[29]:
```
# Import libraries
import matplotlib.pyplot as plt
import matplotlib
```

The first graph shown will be a pie chart. This chart can give an overview of the distribution of States to see which have the most density (sales).

[73]:
```
# Create data only containing StateName
pie_data = data.groupby('StateName').size()

# Make the pie plot with pandas
pie_data.plot(kind='pie', subplots=True, figsize=(12,12),  labeldistance=1.1)
plt.title("States Distribution")
plt.ylabel("")
plt.show()
```

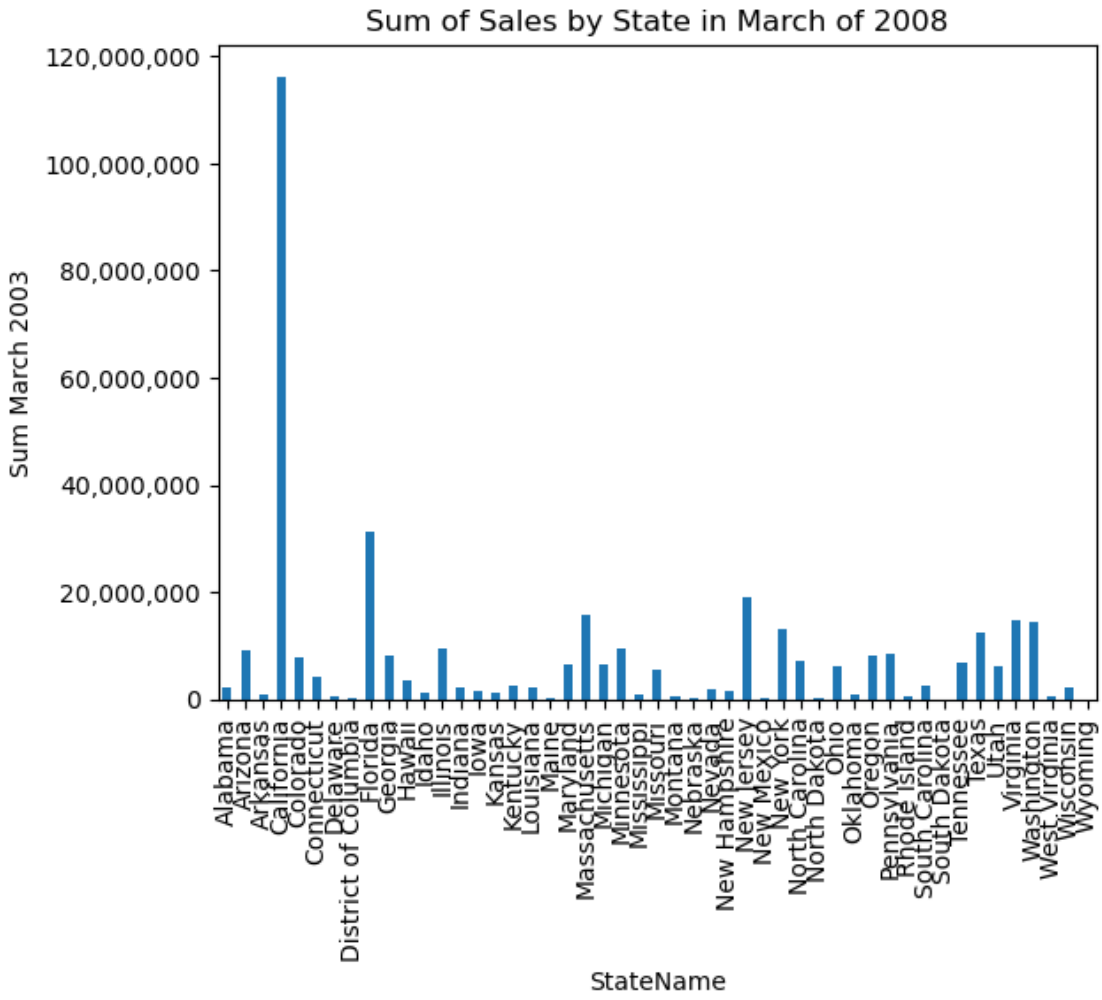


When analyzing this chart, we can see that there are a few states that stick out in terms of size. These are California, Florida, and Massachusetts.

The next graph that will be created, will be the sales of the first data point, March of 2008. This can be a good starting point to being to understand the changes of price over time.

[74]:
```
# Groups by state and sums the prices of March 2008
df_grouped = data.groupby('StateName').sum()['2008-03']
```

[32]:
```
# Creates plot using grouped data
ax = df_grouped.plot(kind='bar')
# Formats numbers to regular notation instead of scientific
ax.get_yaxis().set_major_formatter(
    matplotlib.ticker.FuncFormatter(lambda x, p: format(int(x), ',')))
# Creates Labels
plt.xlabel('StateName')
plt.ylabel('Sum March 2003')
plt.title('Sum of Sales by State in March of 2008')
# Rotates X labels 90 degrees
plt.xticks(rotation=90)
# Displays plot
plt.show()
```



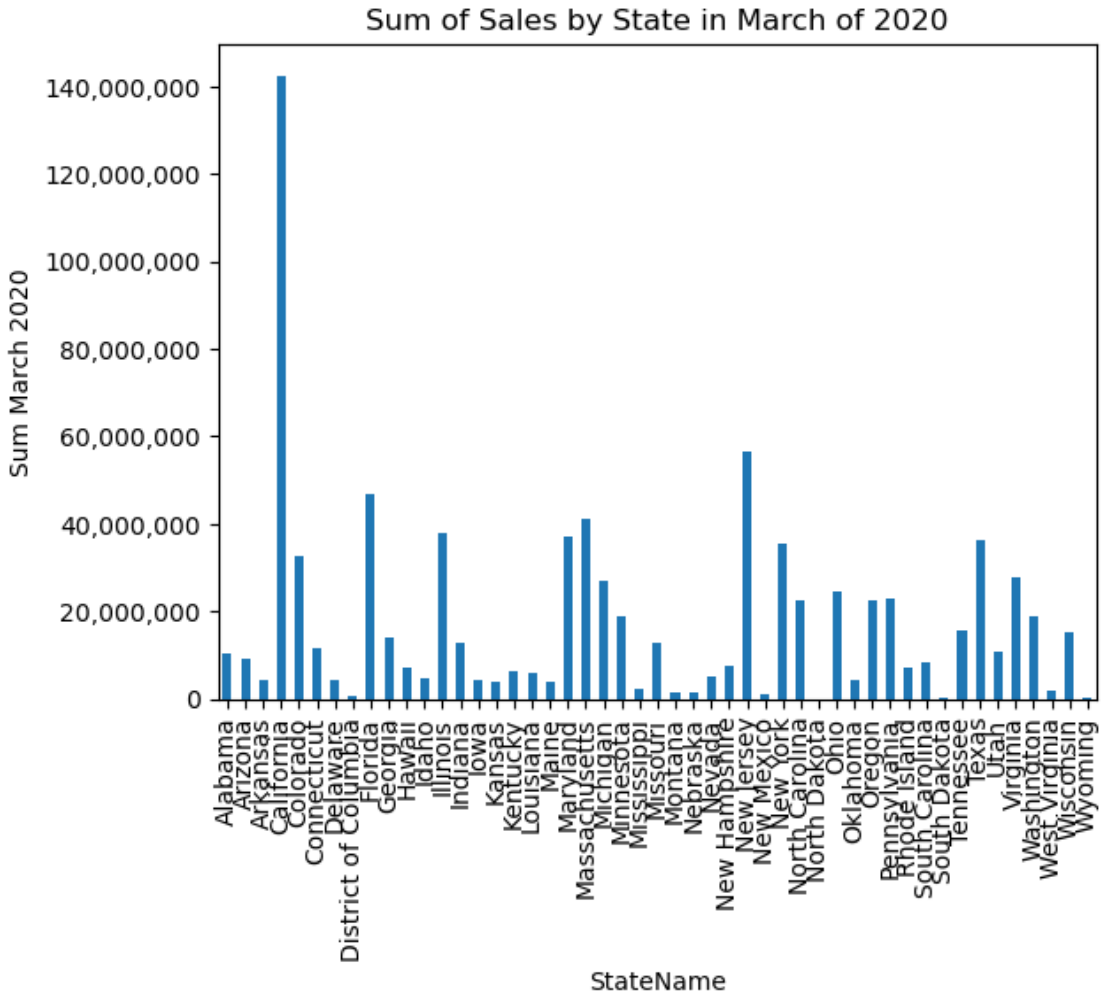
This graph confirms some of the assumptions seen in the previous plot. This shows that in the month of March in 2008, the sum of total sales are the highest in California, Florida, New Jersey, and Massachusetts.

Next, a graph using more recents sales, March of 2020.

[36]:
```
# Grouping by State and sum of March of 2020
df_grouped2 = data.groupby('StateName').sum()['2020-03']
```

[75]:
```
# Creates plot using the grouped data
ax = df_grouped2.plot(kind='bar')
# Formats numbers to regular notation instead of scientific
ax.get_yaxis().set_major_formatter(
    matplotlib.ticker.FuncFormatter(lambda x, p: format(int(x), ',')))
# Creates Labels
plt.xlabel('StateName')
```

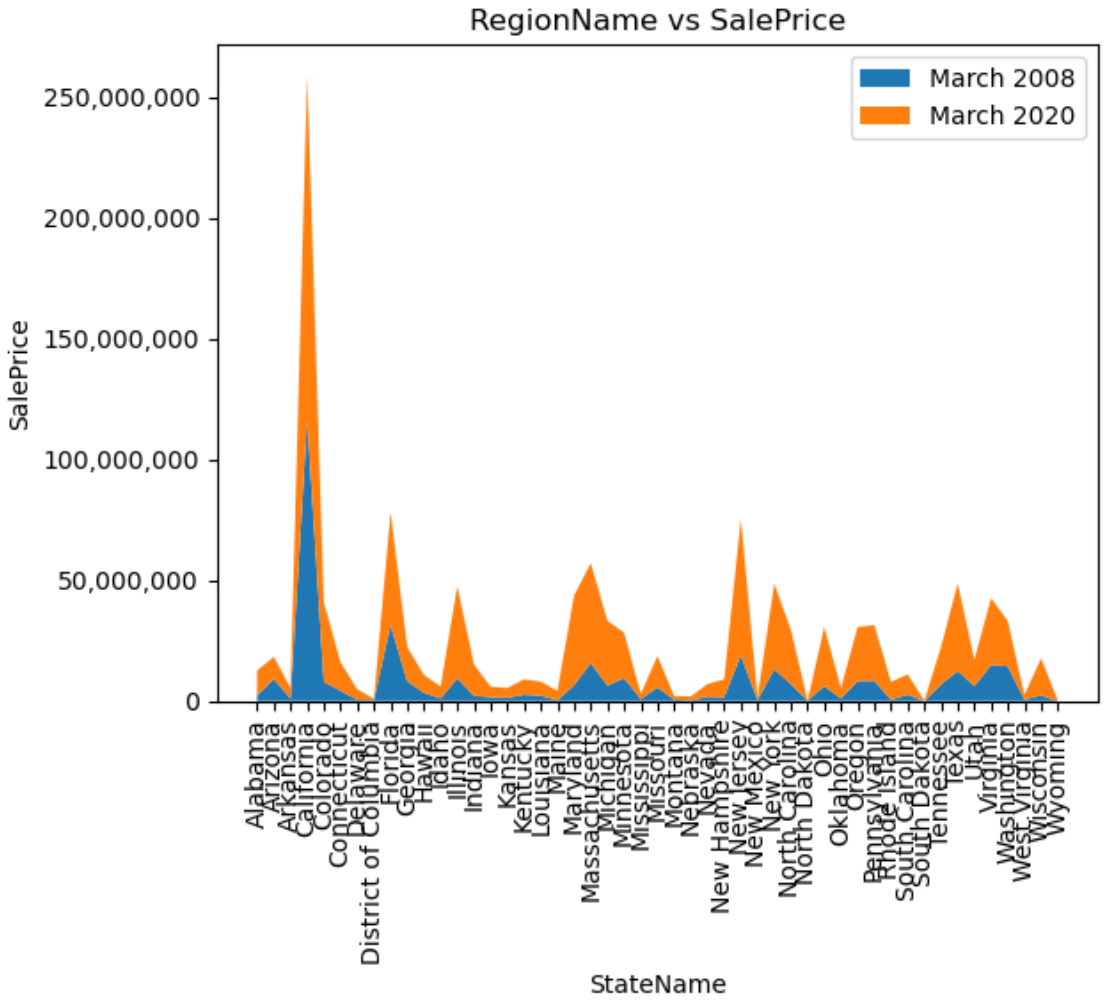
```
plt.ylabel('Sum March 2020')
plt.title('Sum of Sales by State in March of 2020')
# Rotates X-Labels 90 degrees
plt.xticks(rotation=90)
# Displays Plot
plt.show()
```



As expected, we see a rise in almost all states that are reporting data. Now, there are more states that share a similar amount in sales. Next, a stack plot can be created to see the two data sets overlayed.

[49]:
```
# Create figure and axes
fig, ax = plt.subplots()
# Plot data
ax.stackplot(df_grouped.index, df_grouped.values, df_grouped2.values,␣
 ↪labels=['March 2008', 'March 2020'])
```

```
# Formats numbers to regular notation instead of scientific
ax.get_yaxis().set_major_formatter(
    matplotlib.ticker.FuncFormatter(lambda x, p: format(int(x), ',')))
# Set labels
ax.set_xlabel('StateName')
ax.set_ylabel('SalePrice')
ax.set_title('RegionName vs SalePrice')
# Rotates X-Labels 90 Degrees
plt.xticks(rotation=90)
# Creates Legend
ax.legend()
# Displays plot
plt.show()
```



The graph above shows the changes in sales side by side from March 2008 to March 2020.

In more recent times, it can be seen that sales are higher than they were in 2008. The graphs created have demostrated those changes and show that certain states increased more than others. With this information, continuing analysis can be done to see which states have increased the most over time. These graphs begin to give some insight but more research will be conducted to see where the highest valued areas are located and where the increase is the highest.

Data Source: https://www.kaggle.com/datasets/paultimothymooney/zillow-house-price-data/discussion