Correlation and Prediction Factors for US Housing Prices

Abhishek Bada
Abhijeet Chawhan
Matt Litz

Agenda

- Project Scope
- The Data
- Exploratory Data Analysis
- Correlation Analysis and Development
- Machine Learning and Regression
- Results

Project Scope

During the COVID pandemic we have all witnessed the housing market grow wildly, houses are selling sight unseen and $20,000 to $50,000 over asking price. This market piqued our interest and drove us to see what relationships whether economically, socially, and even geographically affected the housing market.

Objectives:

- Analyze correlations between our predictor variables and "Price" to establish which ones most heavily influence house prices.
- To convert House Index Data into useful information and subsequently into knowledge to find correlation and prediction factors.
- Explore the influence from COVID and other factors on House price index in the granularity of geography.

The Data

Data Set Predictors

- Unemployment Rate
  - % unemployment per county from the years 2000 to 2020
  - 67599 data values with no missing data
  - Quantitative continuous data with a range from 0.0% to 29.4% and mean of 6.23%
  - Data Source: Geo FRED

- Crime Rate
  - Crime rate per 100,000 in each state from the years 2000 to 2020
  - 1050 data values with no missing data - 50 crime rates by state for 21 years
  - Quantitative continuous data with a range from 78.2 to 891.7 and mean of 385.7
  - Data Source: Crime Data Explorer

- Mortgage Rate
  - % Mortgage Rate by month for 30-Year Fixed Mortgage from 2000 to 2020
  - 252 data values with no missing data
  - Quantitative continuous data with a range from 3.11 to 8.05 and mean of 5.09
  - Data Source: Freddiemac

Data Set Predictors (Continued)

- Bedroom
  - Number of data values by bedroom size
  - 8 million rows of data
  - Quantitative discrete data with a range from 1 to 5
  - Data Source: Kaggle
- S&P 500
  - Market Index average for each month from 2000 to 2020
  - 252 data values with no missing data
  - Quantitative continuous data
  - Data Source: Web Scrape from Yahoo Finance
- House Price Index **(Response variable)**
  - 8 million rows of data
  - Quantitative discrete data with a range from 1 to 5
  - Data Source: Kaggle

Additional Financial Metrics (Continued)

- S&P 500
- M2 Money Stock
- S&P/Case-Shiller U.S. National Home Price Index
- Consumer Price Index for All Urban Consumers: All Items in U.S. City Average
- Effective Federal Funds Rate
- 10-Year Treasury Constant Maturity Rate
- US Unemployment Rate
- Crime Rates
- 30-Year Fixed Rate Mortgage Average in the United States
- Personal Saving Rate of US Households
- Crude Oil Prices: West Texas Intermediate (WTI)
- Commercial and Industrial Loans, All Commercial Banks
- Moody's Seasoned Aaa Corporate Bond Yield

Data Cleaning and Transformation

- Kaggle data set for Housing Prices and Bedrooms had many empty rows, NaN values, and house prices that were $0.
  - Used dropna() function to remove empty rows and filtered out any rows where the "Price" column had values of $0.
- Many data sets acquired from government sites were clean and had no missing data however the format of the data sets was not conducive to Time-Series Analysis.
  - Used pandas melt function to transpose many date columns into 1 column.
  - Used pandas merge function to combine different data sets based on either the date, county, or state

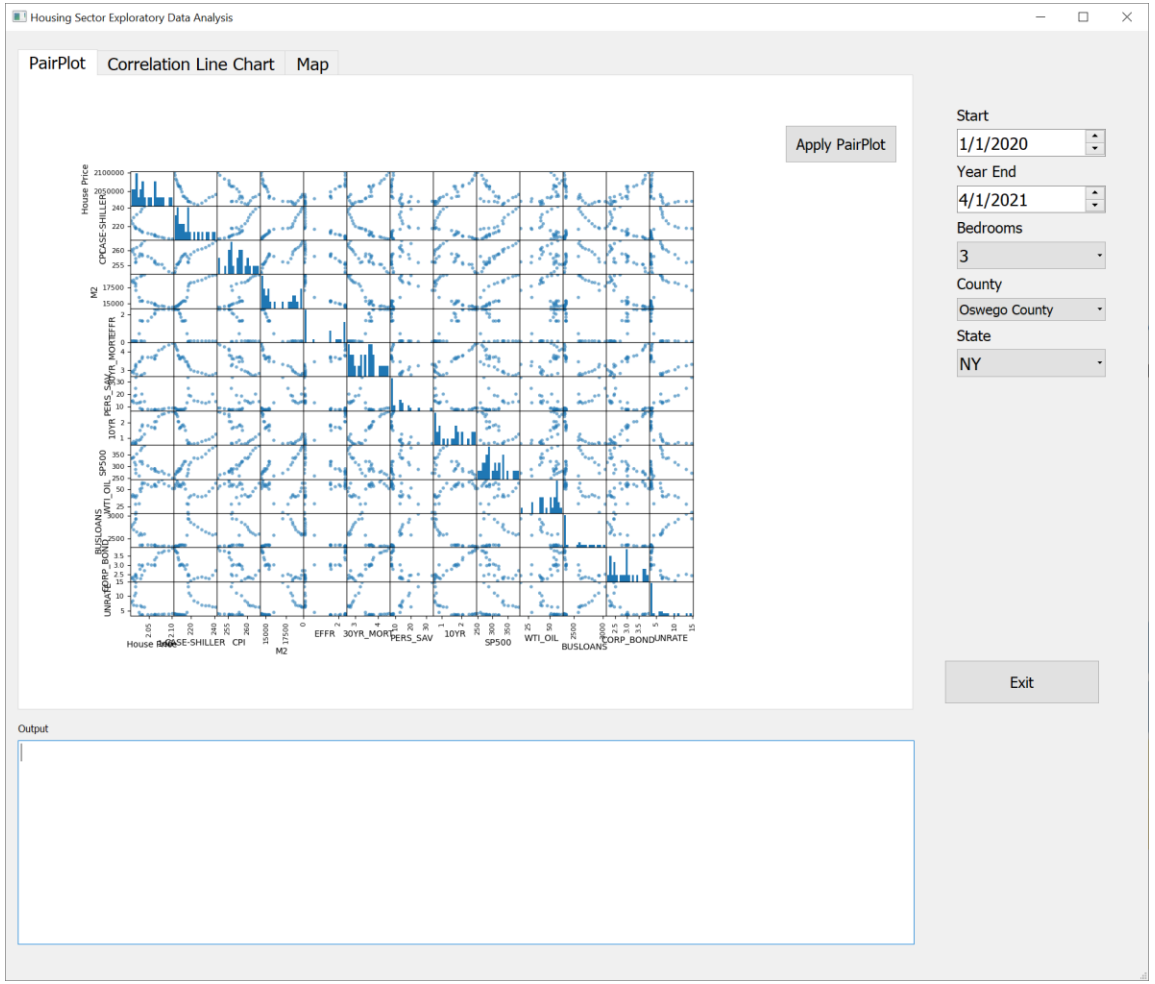
untyName	State	City	Bedrooms	RegionName	Price	Year	Month	Unnamed: 0.1	Region Name	Unemployment Rate	StateName	Crime Rate	Rate	S&P 500
beville County	SC	Abbeville	2	29620	59608.0	2010.0	1.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.03	1073.87
beville County	SC	Abbeville	3	29620	98736.0	2010.0	1.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.03	1073.87
beville County	SC	Abbeville	4	29620	147164.0	2010.0	1.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.03	1073.87
beville County	SC	Abbeville	5	29620	232582.0	2010.0	1.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.03	1073.87
beville County	SC	Abbeville	2	29620	59460.0	2010.0	2.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.99	1104.49
beville County	SC	Abbeville	3	29620	97550.0	2010.0	2.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.99	1104.49
beville County	SC	Abbeville	4	29620	146869.0	2010.0	2.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.99	1104.49
beville County	SC	Abbeville	5	29620	230781.0	2010.0	2.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.99	1104.49
beville County	SC	Abbeville	2	29620	58965.0	2010.0	3.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.97	1169.43
beville County	SC	Abbeville	3	29620	96787.0	2010.0	3.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.97	1169.43
beville County	SC	Abbeville	4	29620	146119.0	2010.0	3.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.97	1169.43
beville County	SC	Abbeville	5	29620	229557.0	2010.0	3.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	4.97	1169.43
beville County	SC	Abbeville	2	29620	58816.0	2010.0	4.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.1	1186.69
beville County	SC	Abbeville	3	29620	96633.0	2010.0	4.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.1	1186.69
beville County	SC	Abbeville	4	29620	146473.0	2010.0	4.0	32190.0	Abbeville County, SC	13.8	South Carolina	602.2	5.1	1186.69

Data Structures

- Used pandas data frame to hold and manipulate our data set
- Python lists, dictionaries, tuples.

Exploratory Data Analysis

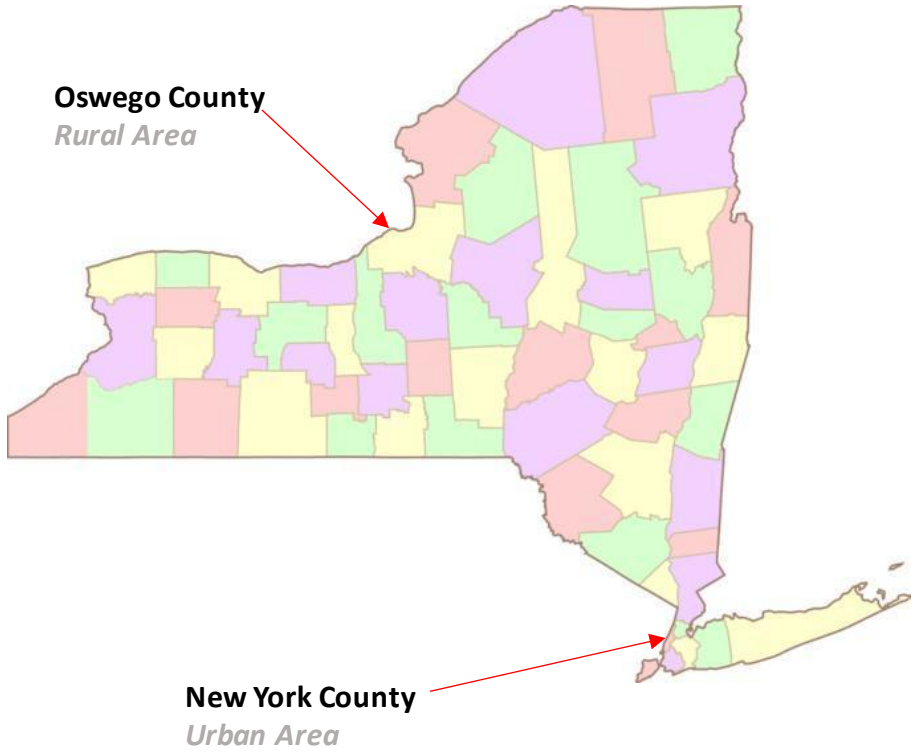
User Interface (UI) for EDA



- A User Interface was developed to enable users to analyze the relationships between the financial metrics pulled from FRED
- **PyQt** was used to build the UI
- PairPlots and Line plots from **matplotlib** were embedded in the application
- The **pandas** library is used for all data manipulation

Focus: New York

- During the start of the COVID pandemic, there were media reports that people were leaving urban areas
- Urban flight suggests a decrease in house prices
- We decided to explore if these reports were true by analyzing New York County housing prices from the start of pandemic to the present



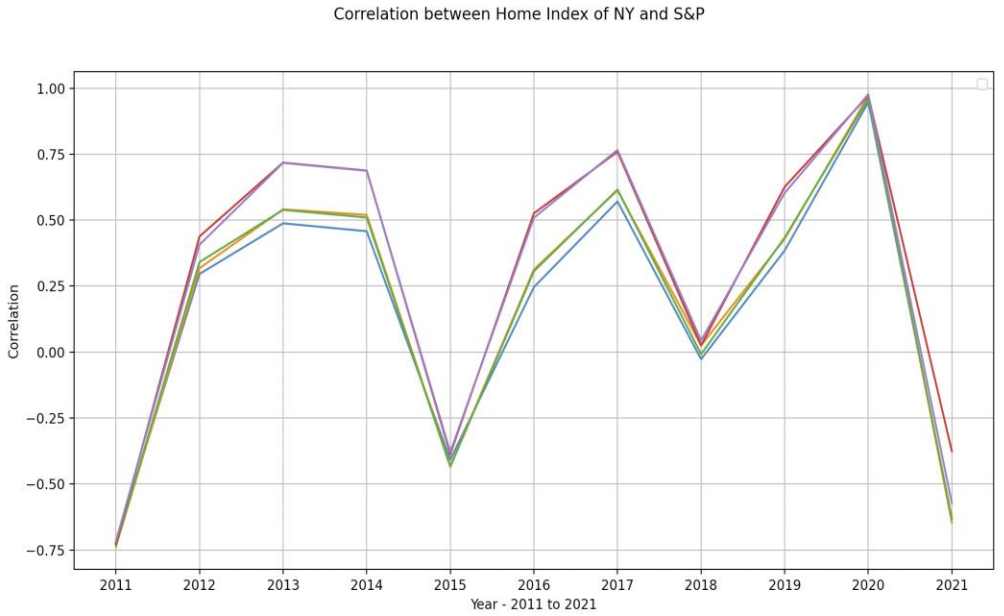
Correlation Analysis and Development

- Utilized Object oriented programming throughout.
- Data from the captured dataset drawn into Pandas data frame.
- Aggregated to but not limited to State, County and City columns.
- Implemented Time series string and date time conversion for generating period wise data.
- Calculation based on the User Input for State, year, City
- Correlation calculated between columns from multiple dataframes created out of separate multiple datasets.
- Calculates the correlation between the house price index and some of the factors viz. S&P500, crime rate, Unemployment rate, etc using the **corr** function.
- Seaborn and Matplotlib for Plots

Correlation Analysis

Let's take an example of NY state --

Correlation between S&P500 and House Price Index for the state of NY shows a positive correlation for the year 2020 indicating that the house prices went up when the S&P500 index went up and vice versa.

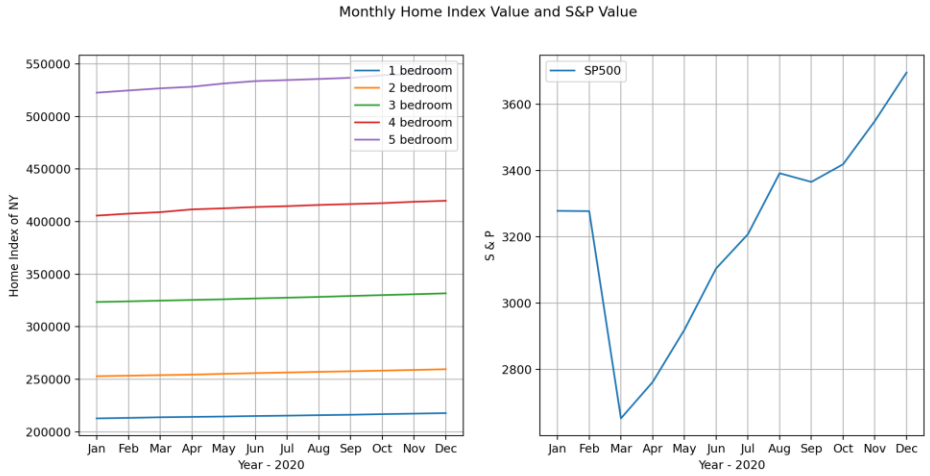


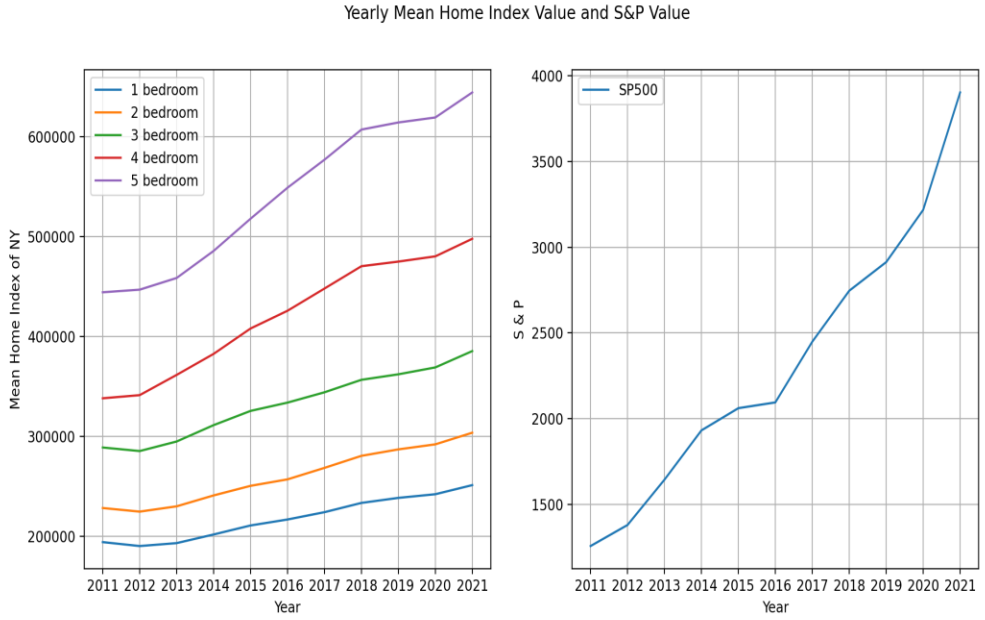
Correlation between 1 bed single family house with SP500 for the year 2021 is	-0.632
Correlation between 2 bed single family house with SP500 for the year 2021 is	-0.647
Correlation between 3 bed single family house with SP500 for the year 2021 is	-0.631
Correlation between 4 bed single family house with SP500 for the year 2021 is	-0.376
Correlation between 5 bed single family house with SP500 for the year 2021 is	-0.574
Correlation between 1 bed single family house with SP500 for the year 2020 is	0.947
Correlation between 2 bed single family house with SP500 for the year 2020 is	0.971
Correlation between 3 bed single family house with SP500 for the year 2020 is	0.96
Correlation between 4 bed single family house with SP500 for the year 2020 is	0.972
Correlation between 5 bed single family house with SP500 for the year 2020 is	0.977
Correlation between 1 bed single family house with SP500 for the year 2019 is	0.385
Correlation between 2 bed single family house with SP500 for the year 2019 is	0.429
Correlation between 3 bed single family house with SP500 for the year 2019 is	0.434
Correlation between 4 bed single family house with SP500 for the year 2019 is	0.626
Correlation between 5 bed single family house with SP500 for the year 2019 is	0.603

Correlation Analysis

- Correlation calculations for previous years

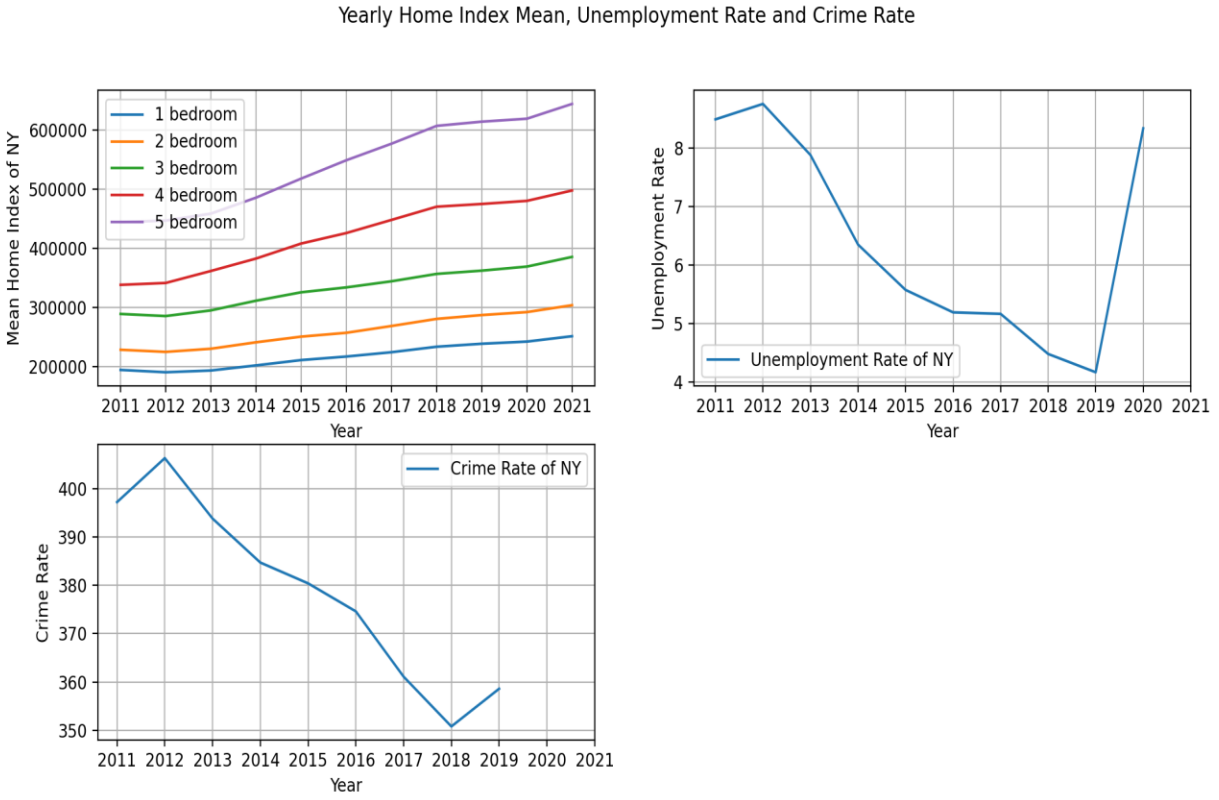
```
Correlation between 2 bed single family house with SP500 for the year 2013 is 0.541
Correlation between 3 bed single family house with SP500 for the year 2013 is 0.539
Correlation between 4 bed single family house with SP500 for the year 2013 is 0.718
Correlation between 5 bed single family house with SP500 for the year 2013 is 0.717
Correlation between 1 bed single family house with SP500 for the year 2012 is 0.296
Correlation between 2 bed single family house with SP500 for the year 2012 is 0.318
Correlation between 3 bed single family house with SP500 for the year 2012 is 0.341
Correlation between 4 bed single family house with SP500 for the year 2012 is 0.438
Correlation between 5 bed single family house with SP500 for the year 2012 is 0.407
Correlation between 1 bed single family house with SP500 for the year 2011 is -0.725
Correlation between 2 bed single family house with SP500 for the year 2011 is -0.738
Correlation between 3 bed single family house with SP500 for the year 2011 is -0.734
Correlation between 4 bed single family house with SP500 for the year 2011 is -0.727
Correlation between 5 bed single family house with SP500 for the year 2011 is -0.717
```





More Correlations Factors

- Correlation between house price index and some more factors like crime rate and the Unemployment rate




```
/Users/abhijeet/Desktop/CS5010/venv/bin/python /Users/abhijeet/Desktop/CS5010/Homework/Semester_project/RealEstateCor/RealEstateCorr
Correlation between 1 bed single family house of NY with Unemployment Rate from the year 2011 to 2020 is -0.575
Correlation between 1 bed single family house of NY with Crime Rate from the year 2011 to 2020 is -0.969
Correlation between 2 bed single family house of NY with Unemployment Rate from the year 2011 to 2020 is -0.569
Correlation between 2 bed single family house of NY with Crime Rate from the year 2011 to 2020 is -0.973
Correlation between 3 bed single family house of NY with Unemployment Rate from the year 2011 to 2020 is -0.62
Correlation between 3 bed single family house of NY with Crime Rate from the year 2011 to 2020 is -0.977
Correlation between 4 bed single family house of NY with Unemployment Rate from the year 2011 to 2020 is -0.652
Correlation between 4 bed single family house of NY with Crime Rate from the year 2011 to 2020 is -0.978
Correlation between 5 bed single family house of NY with Unemployment Rate from the year 2011 to 2020 is -0.633
Correlation between 5 bed single family house of NY with Crime Rate from the year 2011 to 2020 is -0.975
```

Linear Regression Analysis

whole US

- Bedrooms
  - Y = 4536 + 68705X
  - R2 = 0.05928
- S&P 500
  - Y = 162248 + 38X
  - R2 = 0.00562
- 30 Year Mortgage Rate
  - Y = 241703 - 467X
  - R2 = 0.00000042607
- Crime Rate
  - Y = 315113 - 12238X
  - R2 = 0.01126
- Unemployment Rate
  - Y = 187650 + 146X
  - R^2 = 0.00222

```
[68705.15306074]
4536.417773565103
The linear regression model for Bedrooms is: y=68705x+4536
[38.74975087]
162248.16820024326
The linear regression model for S&P 500 is: y=38x+162248
[-467.99580772]
241703.75280663522
The linear regression model for Rate is: y=-467x+241703
[-12238.65628658]
315113.2959416428
The linear regression model for Unemployment Rate is: y=-12238x+315113
[146.39235984]
187650.56209040535
The linear regression model for Crime Rate is: y=146x+187650
```

```
def linearRegression():
    df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
    reg = linear_model.LinearRegression()
    predictors = ['Bedrooms',"S&P 500", "Rate", "Unemployment Rate", "Crime Rate"]
    for predictor in predictors:
        reg.fit(df[[predictor]], df['Price'])

        print(reg.coef_)
        print(reg.intercept_)

        print('The linear regression model for ' + predictor + ' is: y=' + str(int(reg.coef_)) + "x+" + str(int(reg.intercept_)))
```

New York County

Inputs

State: NY	Bedrooms: 4
30 YR Rate:2.98 %	S&P 500: 4140.71
Crime Rate: 7089	Unemployment Rate: 5.1%

Multilinear Regression Model Output

- Price : $ 6366422
- Zillow: $6 – 8 million

Multilinear Regression & Machine Learning

- Multilinear Regression Equation
- Price= -3768341*Bedrooms* -2.4*SP500* +14830*Rate* -14126*Unemployement* +209*Crime*
- Machine Learning Test/Train Coeffecient
  - USA – 0.0394
  - New York - 0.0461
  - New York County- 0.3680

```
def machinelearningmodel():
    df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
    df = df[df['State'] == 'VA']
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(df[['Bedrooms',"S&P 500", "Rate","Unemployment Rate","Crime Rate"]],df['Price'],test_size = 0.2)
    from sklearn.linear_model import LinearRegression
    clf = LinearRegression()
    clf.fit(X_train,y_train)
    print(clf.score(X_test,y_test))
```

Unit testing

```python
import unittest

class gradesTestCases(unittest.TestCase):

    def test_columns_have_no_null_values(self):
            df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
            predictors = ['Bedrooms',"S&P 500", "Rate","Unemployment Rate","Crime Rate"]
            for pred in predictors:
                for boolean in df[pred].isna():
                    self.assertTrue(boolean == False)

    def test_dataset_has_no_duplicates(self):
            df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
            for boolean in df.duplicated():
                self.assertTrue(boolean == False)

    def test_price_column_has_no_zeros(self):
        df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
        for boolean in df['Price'].isin([0]):
                self.assertTrue(boolean == False)

    def test_that_unemployment_has_no_zeros(self):
        df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
        for boolean in df['Unemployment Rate'].isin([0]):
            self.assertTrue(boolean == False)

    def test_that_SP500_has_no_zeros(self):
         df = pd.read_csv('/Users/abhishekbada/Desktop/CS Project/Main1.csv')
         for boolean in df['S&P 500'].isin([0]):
             self.assertTrue(boolean == False)


if __name__ == '__main__':
    unittest.main()
```

```
In [22]: runfile('/Users/abhishekbada/Desktop/MSDS/Spring 2020/CS 5010/AggBedroom.py', wdir='/Users/
abhishekbada/Desktop/MSDS/Spring 2020/CS 5010')
.....
----------------------------------------------------------------------
Ran 5 tests in 105.580s

OK
```

Results

Conclusions

- S&P 500 and House Price Index correlations are not affected by quantity of bedrooms.
- Machine learning analysis demonstrated that regions affect house prices.
- The analysis can be used by the secondary mortage and real estate marketing companies to advertise efficiently.
- Federal Reserve Monetary Policy affect Housing Prices
  - Affected predictors: M2, Federal Funds Rate, 10-Year Treasury, 30-Year Mortgage Rate
- We have determined it is possible to build predictive models for US House Prices through our analysis of house index price data by converting into information(data frame manipulation) and subsequently into knowledge (regression model).

Future Improvements

- Increase granularity of data to county level for all predictors
- Embed Multilinear-regression and machine learning in UI
- Use One-Hot encoding for the states categorical predictor
- Statistical analysis of linear regression models
- Increase the number of predictors
  - Social: Schools, Hospitals and Parks
  - Economic: NASDAQ, Average Household Income
  - Geographic: Close to busy roads

Questions