HR_SALARY_PREDICTION REPORT

Final Report

Contents

List of Figures:

List of Tables:

Problem Statement

Predicting and analyzing the price of houses based on several independent and dependent features, that is: demand & supply factors that could influence US home prices.

1. Introduction: Brief introduction about the problem statement and the need to solve it.

Defining problem statement:

- ***The goal here is to build a model to give predicted US home prices and use it to explain how these factors impacted home prices over the last ~20 years.***
- ***The prediction will be based on a number of demand & supply factors related to the Unemployment rate, Housing unit experience, Working Age Population, Rental Vacancy and other factors.***

Need of the study/project:

- ***The main reason to build such a model is to understand the factors that have influenced housing prices over the years.***

2. EDA and Business Implication:

Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables.

Data collection:

- ***Collected the data from different sources for the past 20 years, over the period of 2002-01-01 to 2022-01-01. Frequency: Quarterly, Aggregation method:(Sum)***
- ***Sources: Retrieved from FRED, U.S. Census Bureau, UC Irvine Machine Learning Repository, Census Survey Explorer.***
- ***Collected from the US's past record of home sale transactions.***

Inspection of data (rows, columns, descriptive details).

- ***For this project the data we have is:***

- ***The number of rows (observations) is 81.***
- ***The number of columns (variables) is 17.***

Table_1: Data Dictionary:

Price	**US home prices, Quarterly(sum)**
Unemployment Rate	**The number of unemployed as a percentage of the labor force.**
Employment Rate	**The employment-to-population ratio, Aged 15-64**
GDP	**Gross domestic product, is the market value of the goods and services produced by labor and property located in the United States**
Working Age Population	**The population in which people are typically engaged in either paid or unpaid work, Aged 15-64**
Dwellings and Residential Buildings	**Residential building includes all buildings intended for private occupancy whether on a permanent basis or not. Dwellings are divided into the following types: single-family, mobile, cottage, semi-detached, row house, and apartment building**
Essential Utilities Cost Average	**Consumer Prices for Housing, Water, Electricity, Gas, and Other Fuels over the years**
Housing starts	**Housing Starts are a measure of new residential construction**
No. Housing units	**Housing unit is a house, an apartment, a group of rooms, or a single room occupied or intended for occupancy as separate living quarters.**
Mortgage Rate(30-Year)	**A mortgage rate is the amount of interest determined by a lender to be charged on a mortgage. 30-year fixed rate is a mortgage that will be completely paid off in 30 years if all the payments are made as scheduled.**
Homes for sale	**Quarterly Supply of New Houses in the United States for sales**
Building Permits	**housing units authorized by a building or zoning permit.**
Rental Vacancy	**No. of vacant units offered for rent and those offered both for rent and sale.**
Homeownership Rate	**The homeownership rate is the proportion of households that is owner-occupied.**
Homeownership Rate by Race and Ethnicity: All Other Races	**Households that is owner-occupied by non-natives such as Asian, Native Hawaiian or Other Pacific Islander, or American Indian or Alaska Native regardless of whether they reported any other race, as well as all other combinations of two or more races.**
Homeowner Vacancy Rate	**Homeowner vacancy rate is the proportion of the homeowner inventory that is vacant for sale.**

Data information:

RangeIndex: 81 entries, 0 to 80

Data columns (total 17 columns):

#	Column	Non-Null Count	Dtype
0	DATE	81 non-null	object
1	Price($)	81 non-null	float64
2	Unemployment Rate	81 non-null	float64
3	Employment Rate	81 non-null	float64
4	GDP	81 non-null	float64
5	Working Age Population	81 non-null	float64
6	Dwellings and Residential Buildings	81 non-null	float64
7	Essential utilities Cost average	81 non-null	float64
8	Housing starts	81 non-null	int64
9	No. Housing units	81 non-null	int64
10	Mortgage Rate(30-Year)	81 non-null	float64
11	Homes for sale	81 non-null	float64
12	Building Permits	81 non-null	int64
13	Rental Vacancy	81 non-null	int64
14	Homeownership Rate	81 non-null	float64
15	Homeowner non-natives	81 non-null	float64
16	Homeowner Vacancy Rate	81 non-null	float64

dtypes: float64(12), int64(4), object(1)

memory usage: 10.9+ KB

Data Size:

- ***The total number of elements in the dataset housing_price_US is 1377.***

Table_2: Describe the data.

	count	mean	std	min	25.00%	50.00%	75.00%	max
Price$	81	519.9	106.7	353.7	438.9	507.6	562.1	871.6
Unemployment_Rate	81	6.2	2	3.6	4.7	5.7	7.4	13.2
Employment_Rate	81	69.5	2.1	62.4	67.8	70.1	71.3	72.2
GDP	81	4114652.4	432406.2	3348727.5	3820381.3	4044992	4453140	4951572.5
Working_Age_Population	81	199075199.8	6939986.3	182810138	195104005.9	201116681.2	205405713.6	207104199.1
Dwellings_and_Residential_Buildings	81	150434.3	37370.6	92679.5	120424.5	149354.3	174738.3	223861.3
Essential_utilities_Cost_average	81	93.2	15	65	82.5	92.9	104	123.7
Housing_starts	81	3786.8	1390.6	1577	2802	3561	5037	6361
No._Housing_units	81	132234.3	6264.9	119061	128439	132619	136818	142939
Mortgage_Rate30_Year	81	4.7	1.1	2.8	3.8	4.4	5.8	7
Homes_for_sale	81	17.8	5.6	10.2	13.7	16.4	19.9	34.2
Building_Permits	81	3959	1447.1	1616	2965	3842	5143	6685
Rental_Vacancy	81	3600.6	455.8	2491	3243	3673	3865	4625
Homeownership_Rate	81	66.3	1.9	63.1	64.7	66	68	69.4
Homeowner_non_natives	81	56.4	2.1	51.2	55	56	58.1	60.6
Homeowner_Vacancy_Rate	81	1.9	0.5	0.8	1.7	1.9	2.4	2.9

Exploratory Data Analysis.

EDA - Univariate Analysis: Numerical

- ***Our objective is to derive the data and define and analyze the pattern present in each variable separately.***

Dist and boxplot of all variables: Price($)

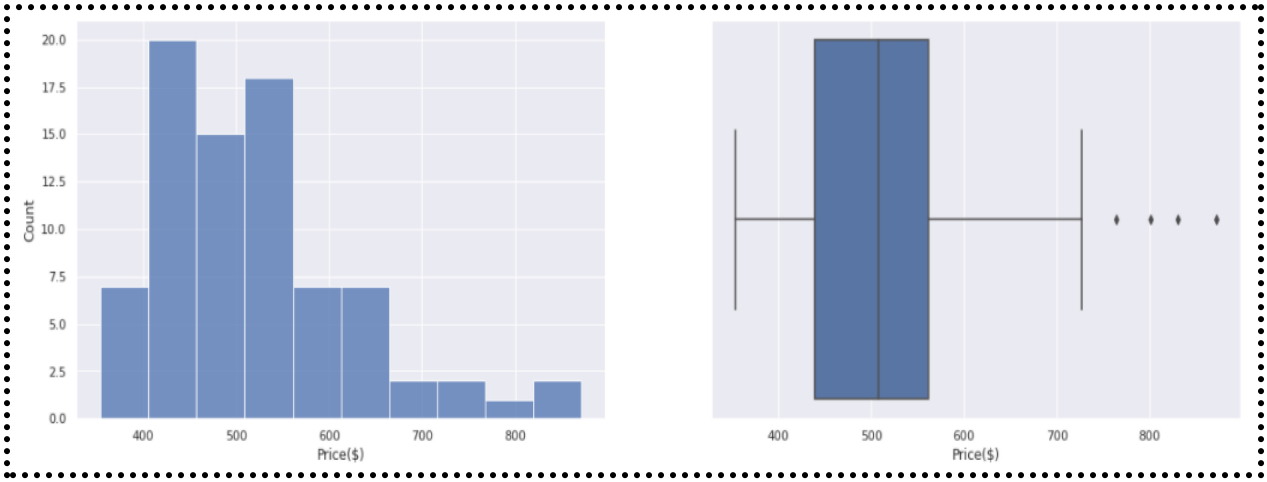


Fig_1: Dist and boxplot: Price($)

- ***The Boxplot tells us there are a few outliers in the Price($) distribution.***
- ***The distplot distribution can be said to indicate towards range format for this variable. The distribution ranges from a sum of 400 to 800(Quarterly).***

Unemployment Rate

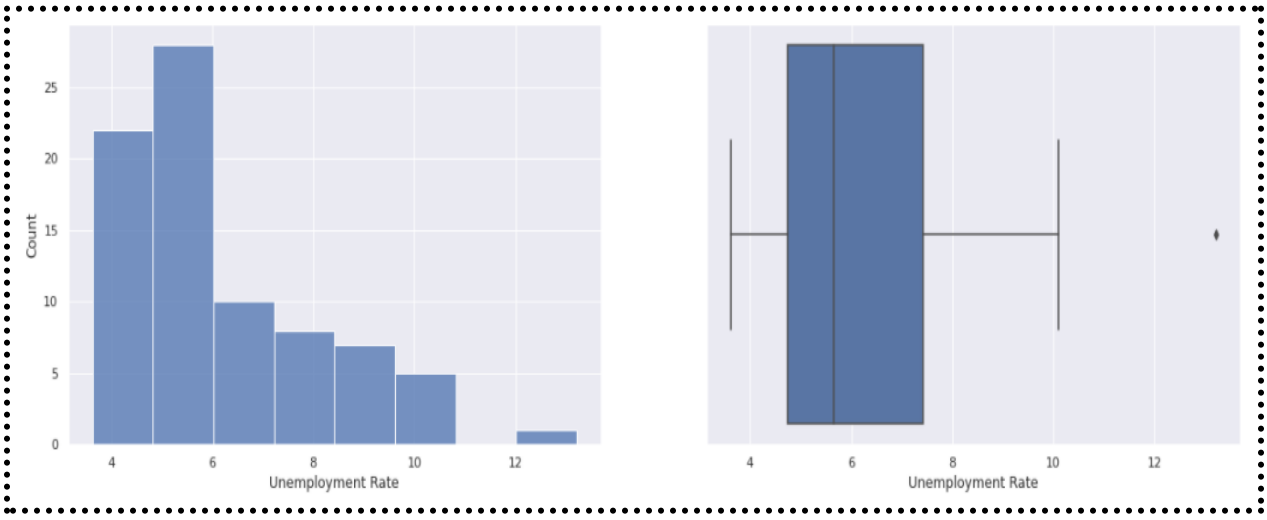


Fig_2: Dist and boxplot: Unemployment Rate

- ***The Boxplot tells us there is an outlier Unemployment Rate distribution.***
- ***The distplot distribution can be said to be highly right-skewed. The distribution ranges from 4 to 12(%).***

Employment Rate

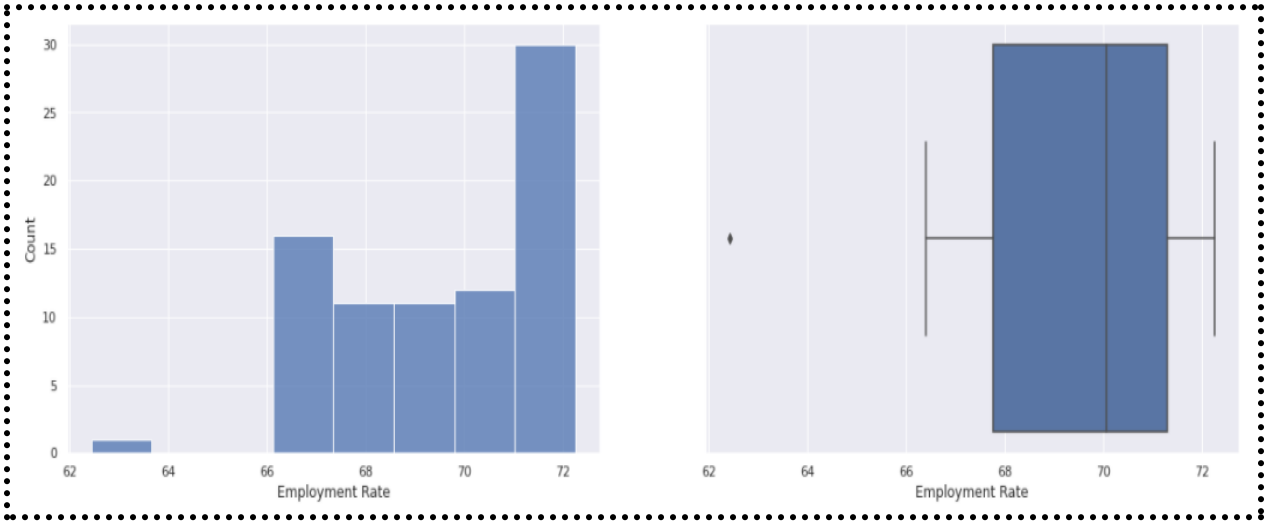


Fig_3: Dist and boxplot: Employment Rate

- ***The Boxplot tells us there is an outlier Unemployment Rate distribution.***
- ***The distplot distribution can be said to be highly left-skewed. The distribution ranges from 62 to 72(%).***

GDP

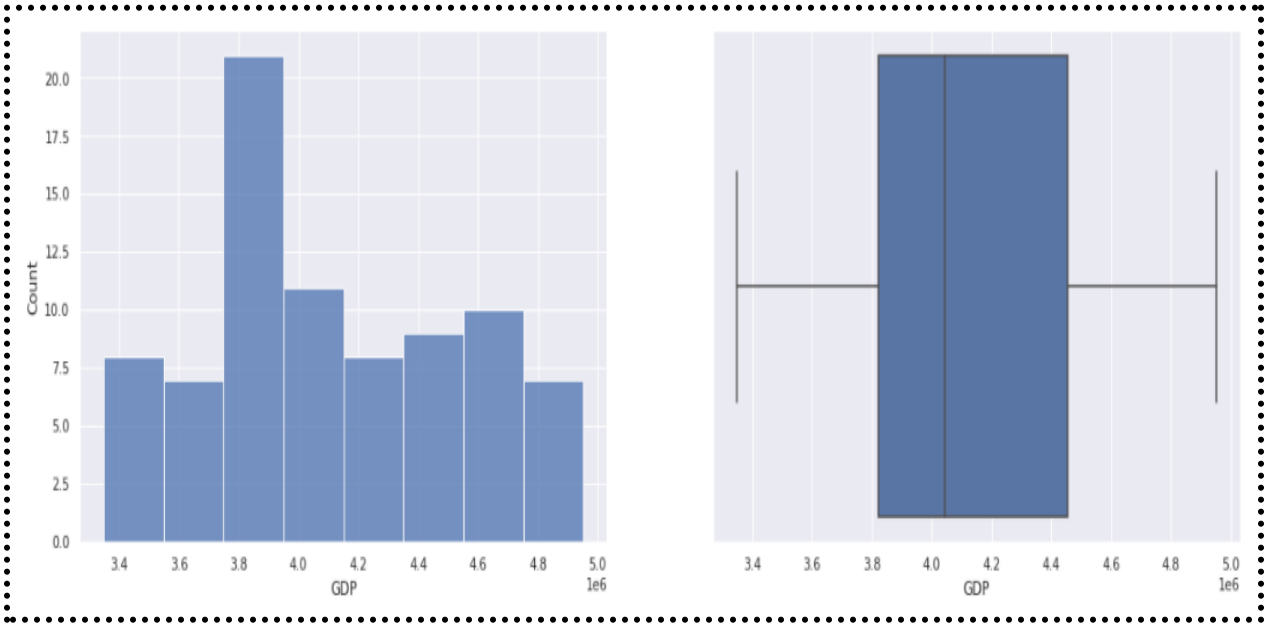


Fig_4: Dist and boxplot: GDP

- ***The Boxplot tells us there are no outliers for GDP distribution.***
- ***The distplot distribution can be said to be slightly right-skewed. The distribution ranges from 3.4e+06 to 5.0e+06.***

Working Age Population

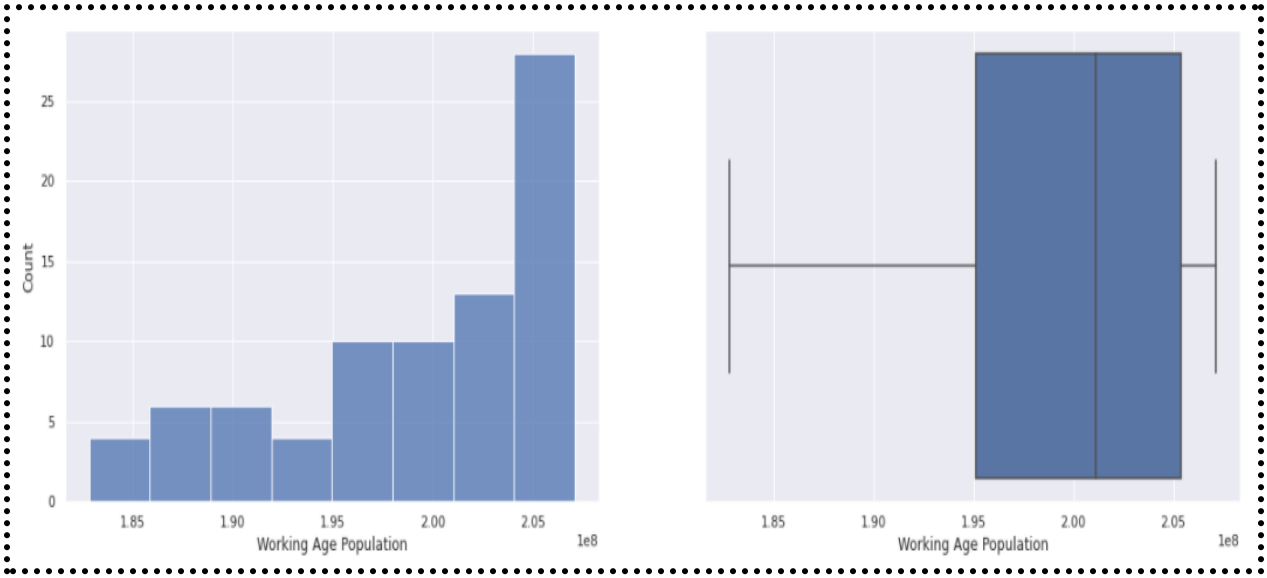


Fig_5: Dist and boxplot: Working Age Population

- ***The Boxplot tells us there are no outliers for Working Age Population distribution.***
- ***The distribution can be said to be highly left-skewed distributed. The distribution ranges from 1.85e+08 to 2.05e+08.***

Dwellings and Residential Buildings

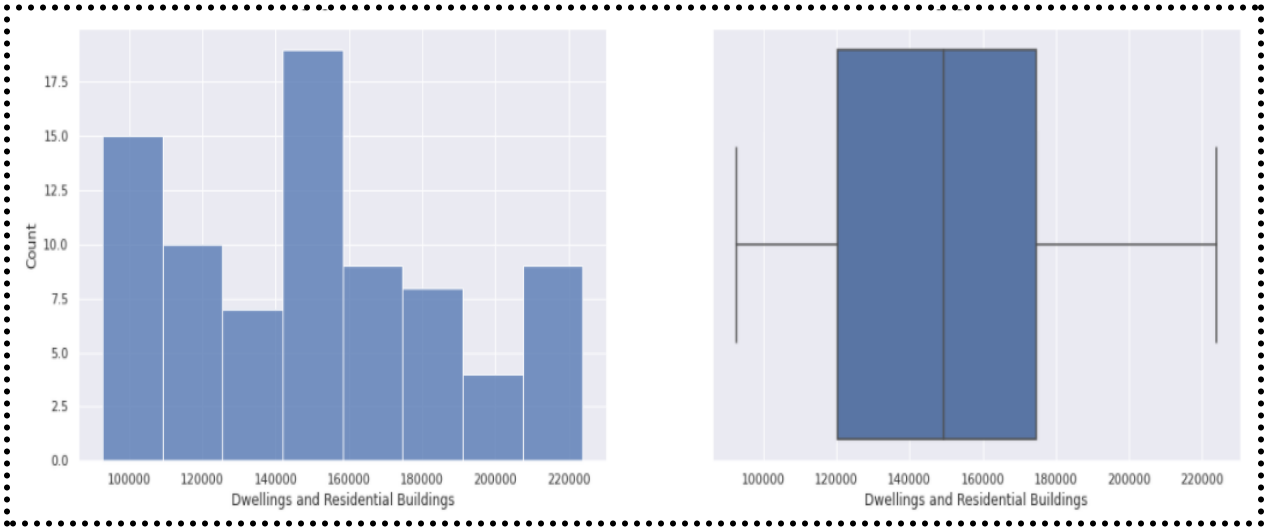


Fig_6: Dist and boxplot: Dwellings and Residential Buildings

- ***The Boxplot tells us there are few outliers for the Dwellings and Residential Buildings distribution.***
- ***The distribution can be said to be slightly right-skewed. The distribution ranges from 100000 to 220000 units.***

Essential utilities Cost average

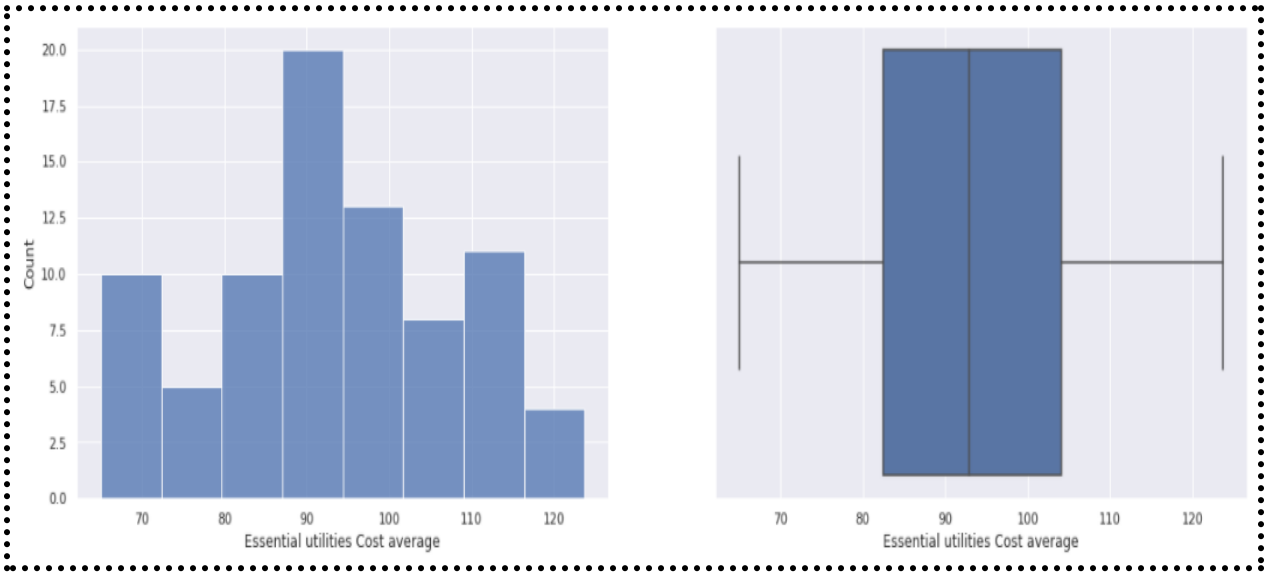


Fig_7: Dist and boxplot: **Essential utilities Cost average**

- ***The Boxplot tells us there are no outliers for Essential utilities Cost average distribution.***
- ***The distribution can be said to be slightly left-skewed. The distribution ranges between 70 to 120 (price avg $)***

Mortgage Rate(30-Year)

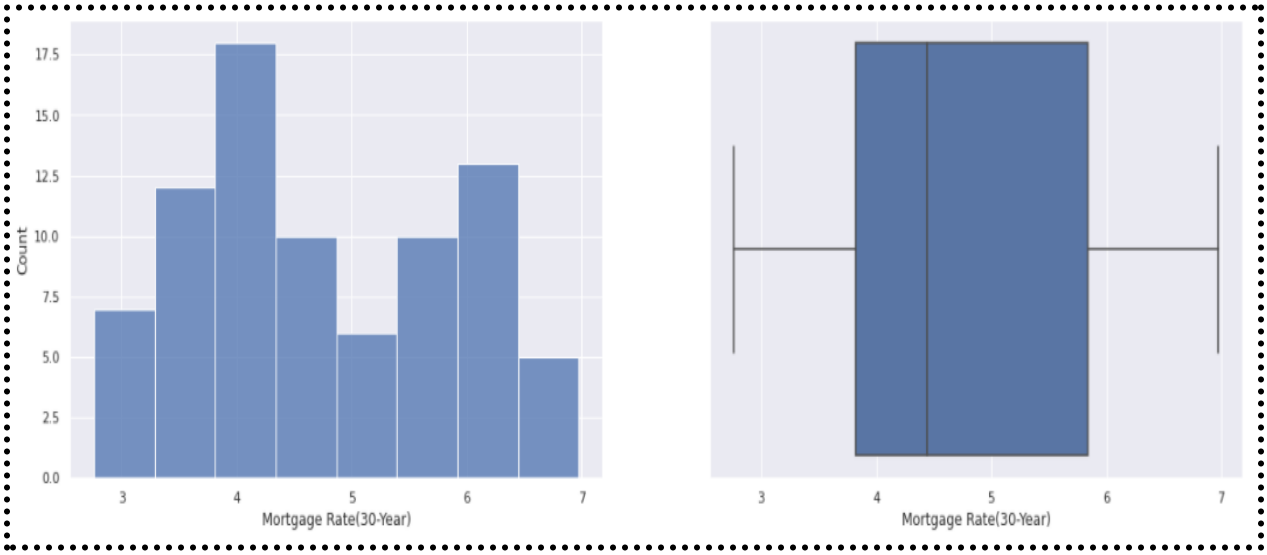
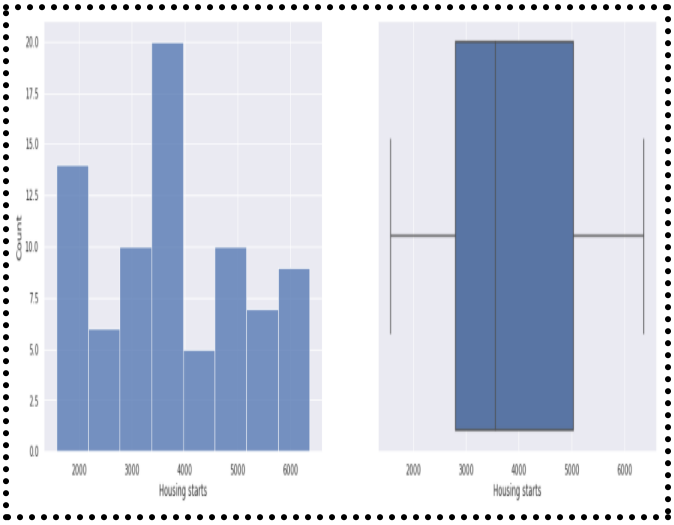


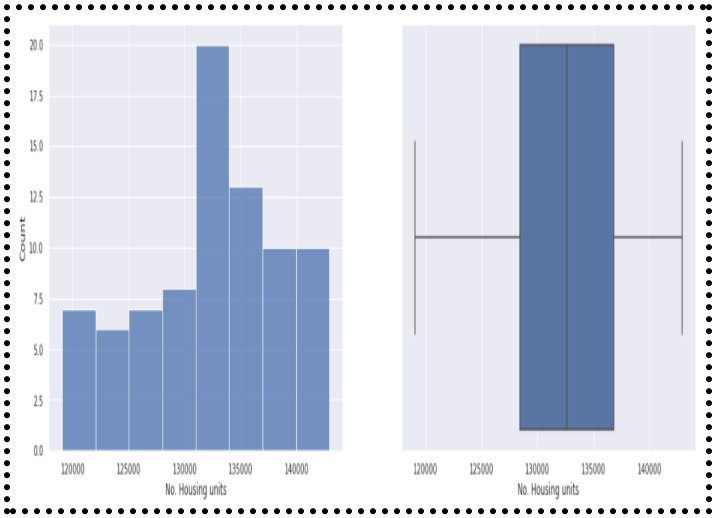
Fig_8: Dist and boxplot: **Mortgage Rate(30-Year)**

- ***The Boxplot tells us there are no outliers for Essential utilities Cost average distribution.***
- ***The distribution can be said to be slightly left-skewed. The distribution ranges between 70 to 120 (price avg $)***

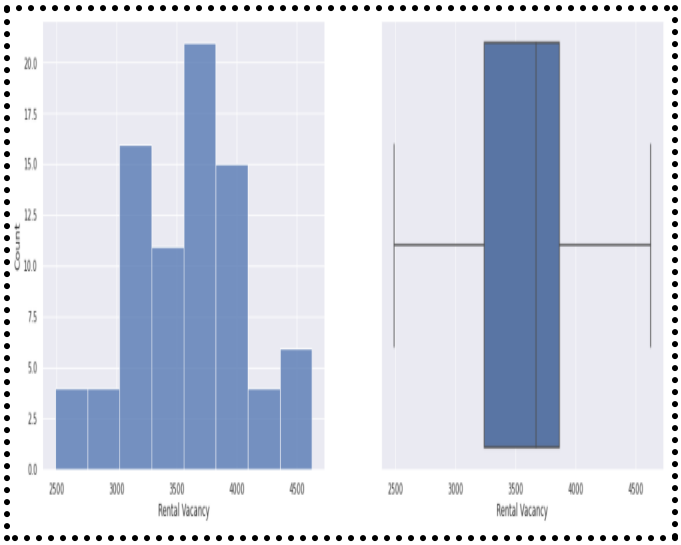
Housing starts, No. Housing units, Rental Vacancy, and Building Permits.



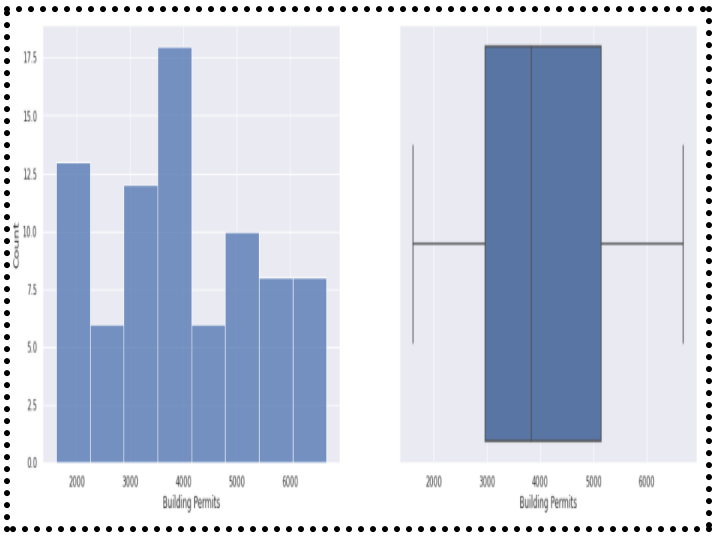
- ***The Boxplot tells us there are no outliers for Housing starts distribution.***
- ***The distribution can be said to be slightly right-skewed. The distribution ranges between 2000 to 6000(units)***



- ***The Boxplot tells us there are no outliers for No. Housing units distribution.***
- ***The distribution can be said to be slightly left-skewed. The distribution ranges between 120000 to 140000 (units)***



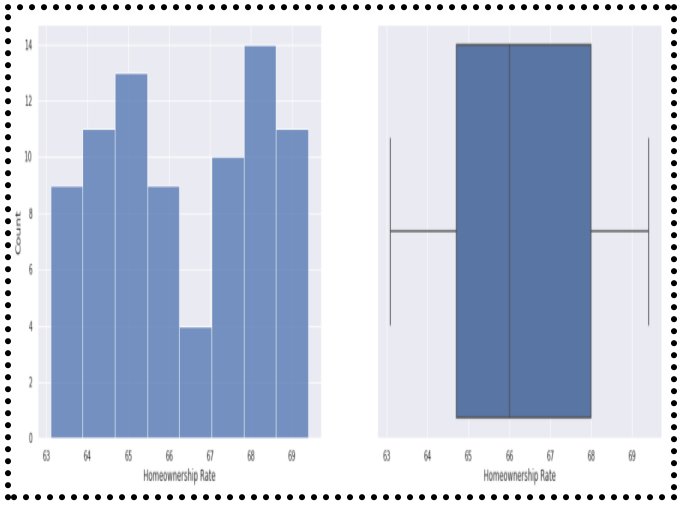
- ***The Boxplot tells us there are no outliers for Rental Vacancy distribution.***
- ***The distribution can be said to be normally distributed. The distribution ranges between 2500 to 5000(units)***



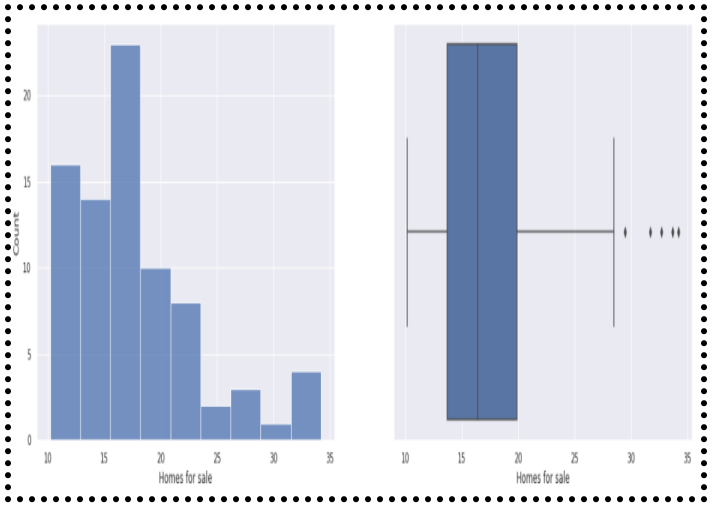
- ***The Boxplot tells us there are no outliers for Building Permits distribution.***
- ***The distribution can be said to be slightly right-skewed. The distribution ranges between 2000 to 6000(issued)***

Fig_9: Dist and boxplot: **Housing starts, No. Housing units, Rental Vacancy, and Building Permits.**

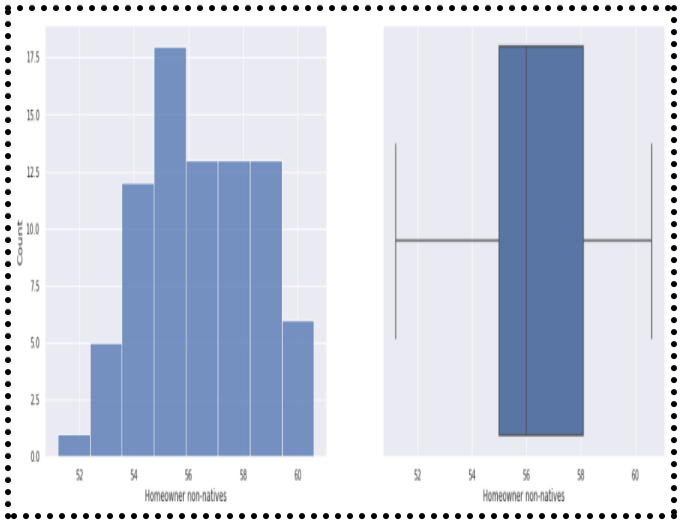
Homeownership Rate, Homes for sale, Homeowner non-natives, and Homeowner Vacancy Rate.



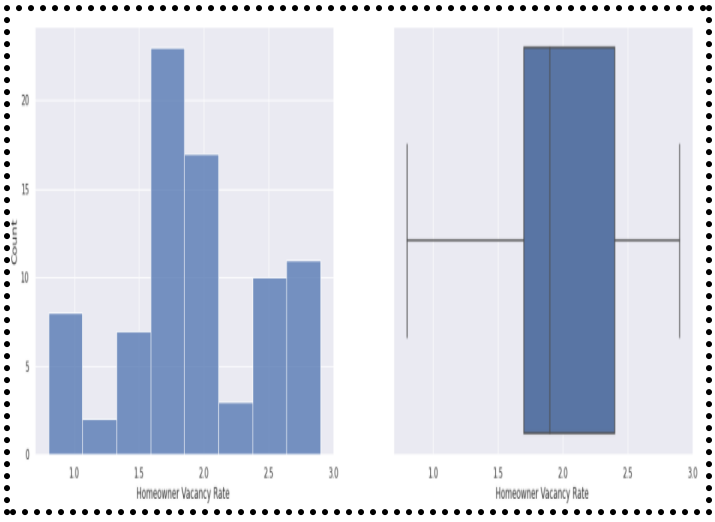
- ***The Boxplot tells us there are no outliers for Homeownership Rate distribution.***
- ***The distribution can be said to be slightly right-skewed. The distribution ranges between 63 to 69(%)***



- ***The Boxplot tells us there are no outliers for Homes for sale for sale distribution.***
- ***The distribution can be said to be highly right-skewed. The distribution ranges between 10 to 35(%)***



- ***The Boxplot tells us there are no outliers for Homeowner non-natives distribution.***
- ***The distribution can be said to be normally distributed. The distribution ranges between 52 to 60 (%)***



- ***The Boxplot tells us there are no outliers for Homeowner Vacancy Rate distribution.***
- ***The distribution can be said to be slightly right-skewed. The distribution ranges between 1.0 to 3.0 (%)***

Fig_10: Dist and boxplot: **Homeownership Rate, Homes for sale, Homeowner non-natives, and Homeowner Vacancy Rate.**

Bivariate and Multivariate Analysis:

US home prices over years

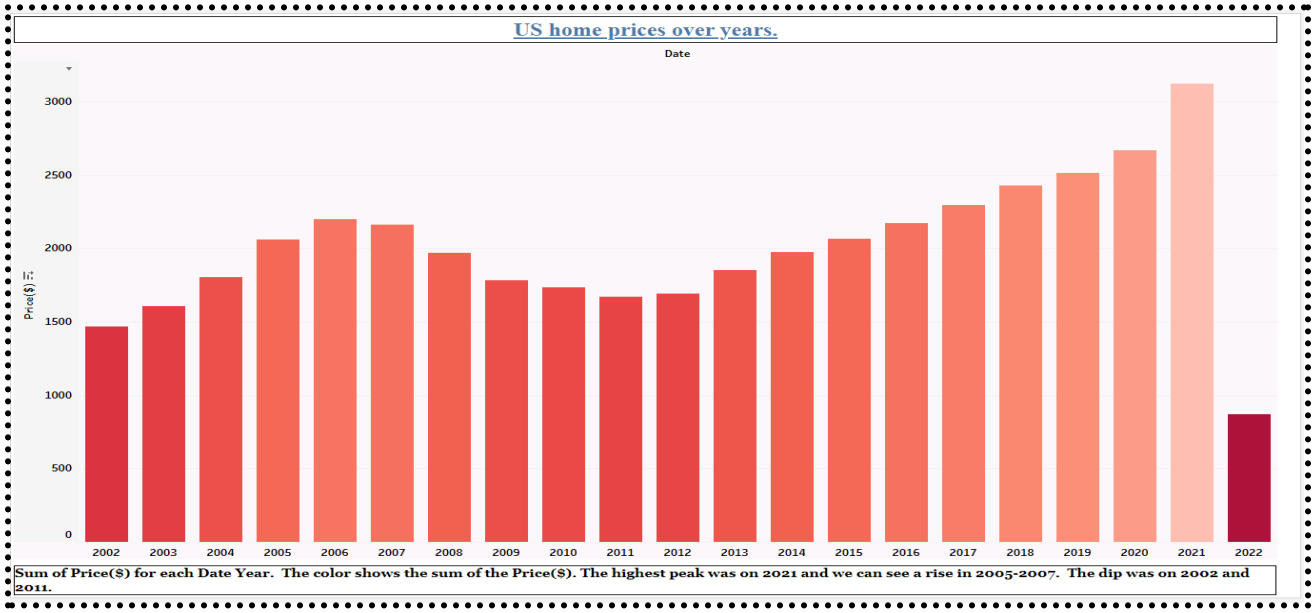


Fig_11: US home prices over years.

Unemployment Rate & Mortgage Rate(30-Year) on price.

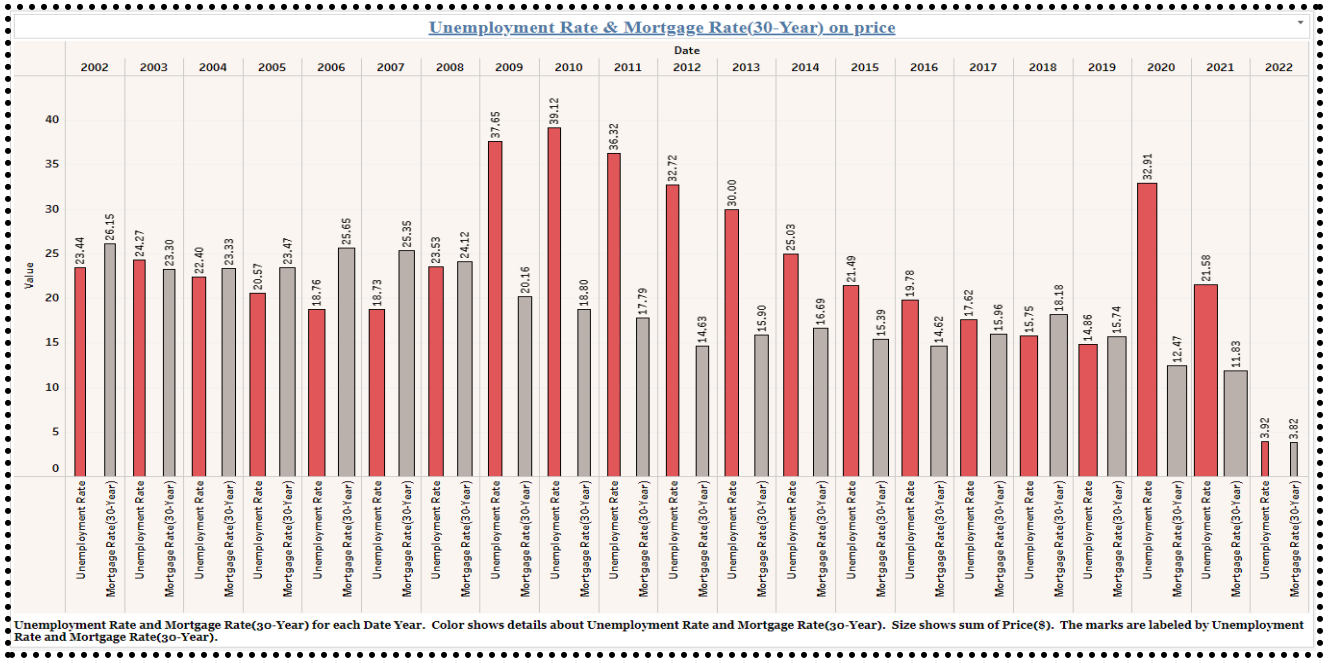


Fig_12: Unemployment Rate & Mortgage Rate(30-Year) on price.

- ***As we saw early, the dip in price during 2002 had a high Mortgage and Unemployment Rate. The same can be said for 2011, Though the Mortgage, Rate was low, the Unemployment rate was at its peak.***

Homeownership Rate & GDP on price

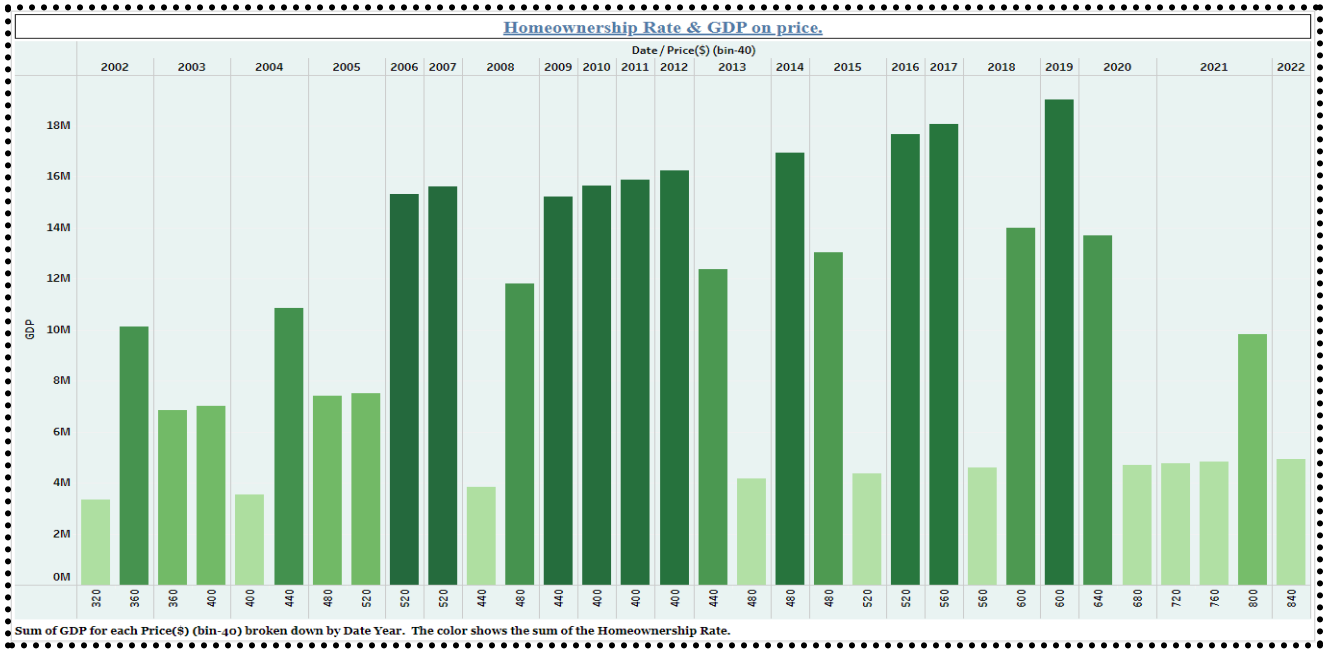


Fig_13: Homeownership Rate & GDP on price.

➢ ***We can see here that higher GDP, higher homeownership rate, in 2019 was the highest GDP rise recorded.***

Homeowner non-natives effect on price

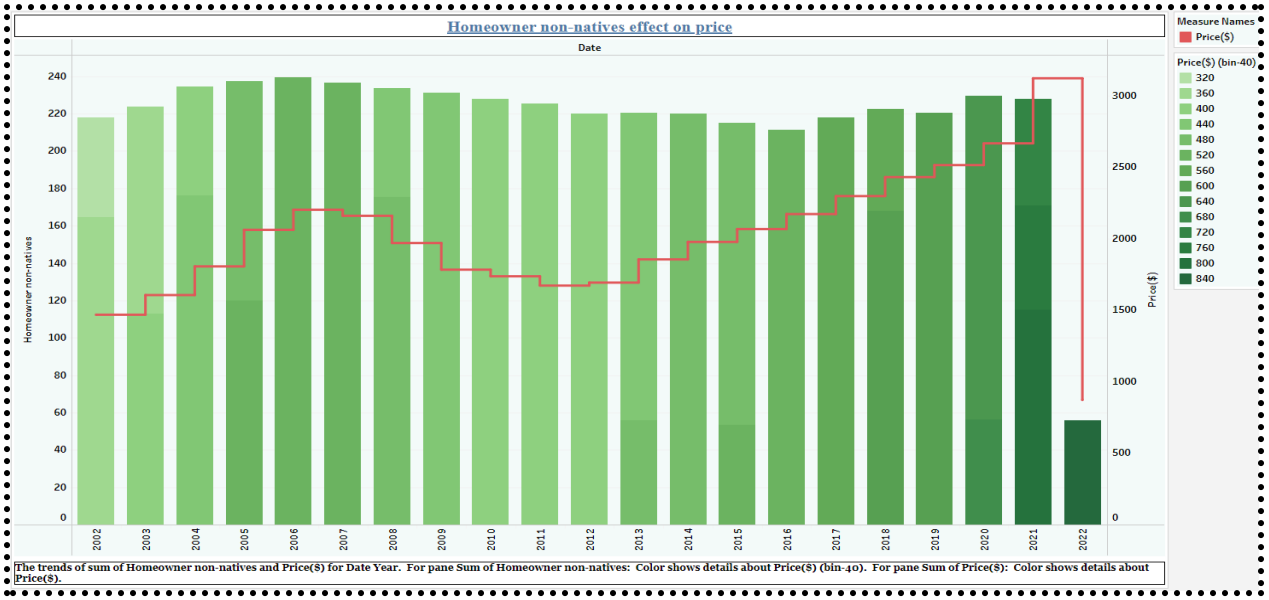


Fig_14: Homeowner non-natives effect on price.

➢ ***We can see that as the rate of Homeowner non-natives the home price also increases.***

Rental Vacancy effect on price

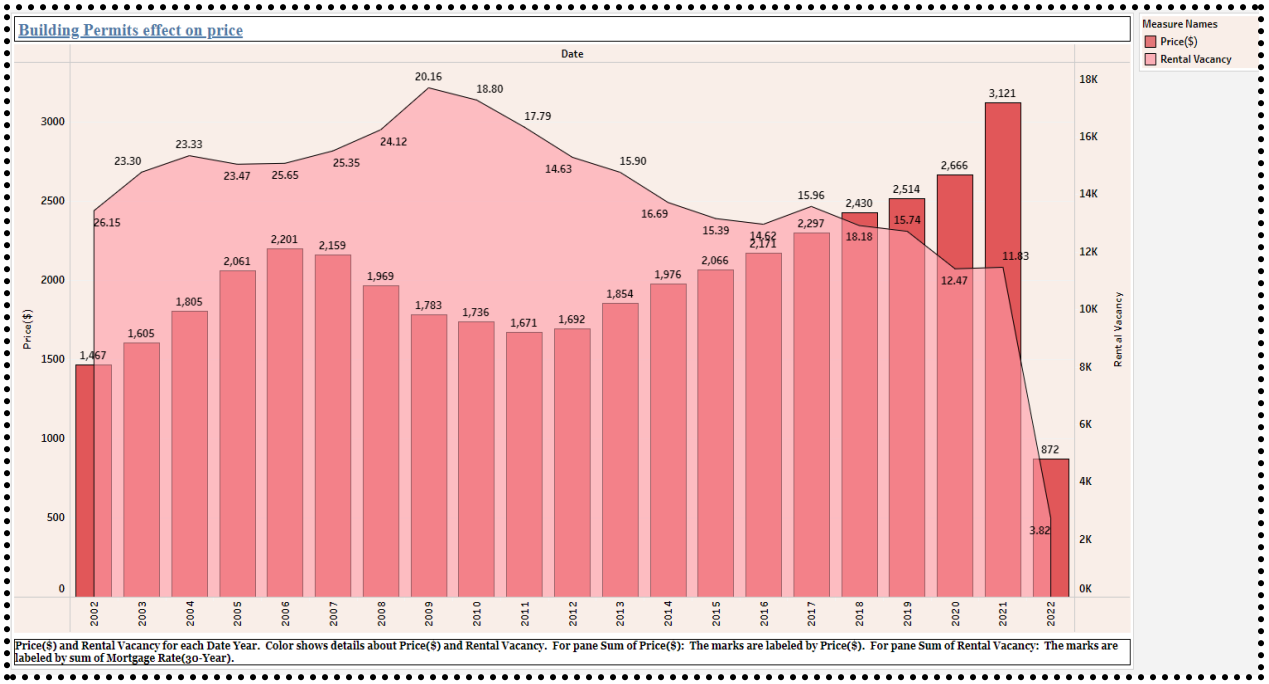


Fig_15: Rental Vacancy effect on price

➢ ***We can see that as the rate of Rental Vacancy the home price also increases.***

Homeowner vacancy effect on price

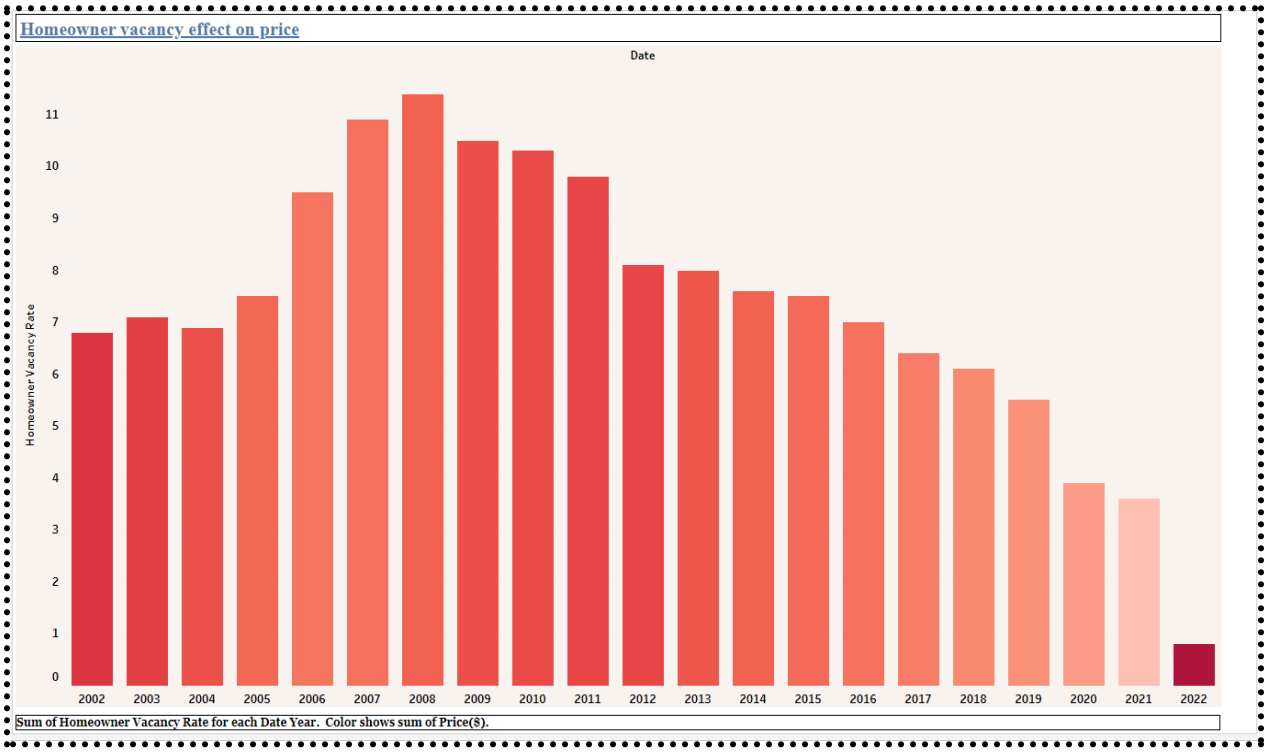


Fig_16: Homeowner vacancy effect on price.

Correlation Heatmap:

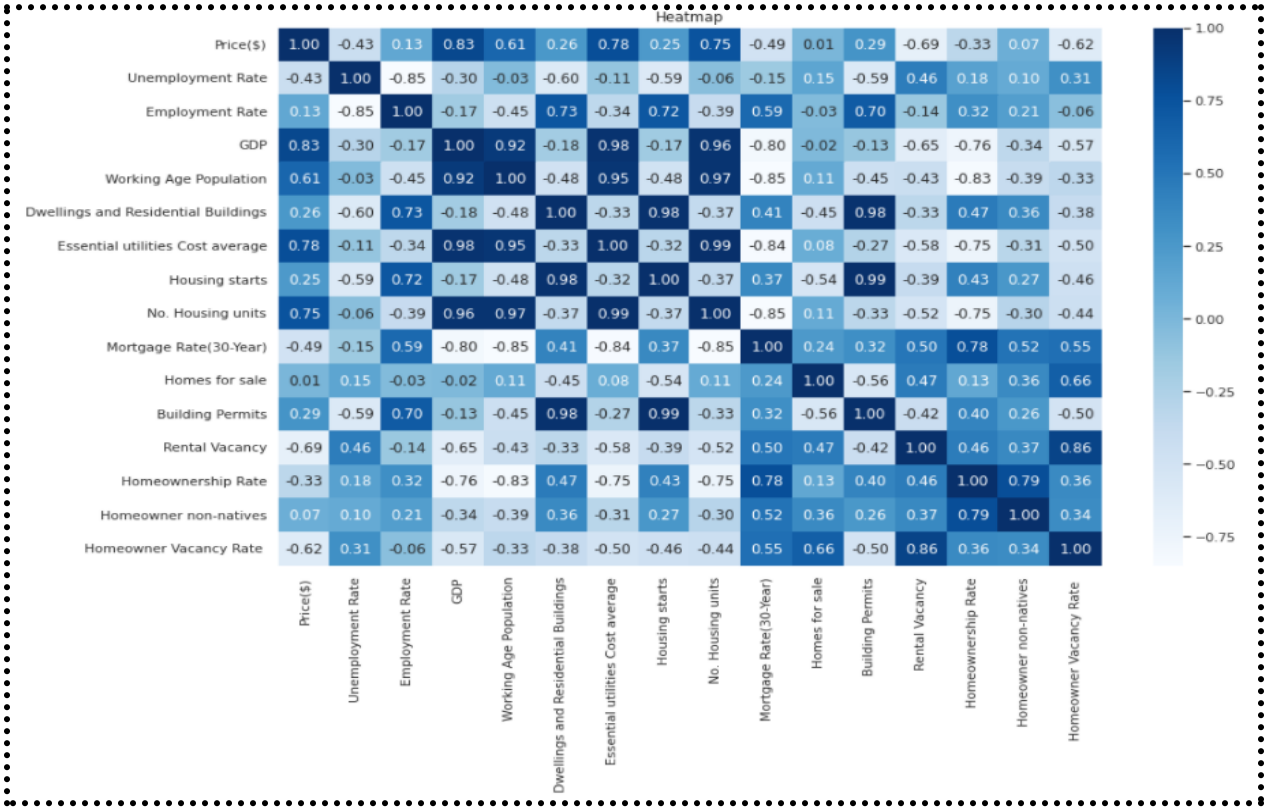


Fig_17: Heatmap

- ***The relation between pairs of numeric variables is given by the heatmap.***
- ***The correlation between the following variables is highly positive: (variables are directly proportional)***
  - ***Essential utilities Cost average and GDP***
  - ***Housing starts and Building Permits***
  - ***GDP and Price($)***
  - ***Building Permits and Dwellings and Residential Buildings***
  - ***Homeowner Vacancy Rate and Rental Vacancy***
- ***The correlation between the following variables are negative: (variable are inversely proportional)***
  - ***Homeownership Rate and No. Housing units***
  - ***Rental Vacancy and Price($)***
  - ***Homeowner Vacancy Rate and Price($)***
  - ***Working Age Population and Mortgage Rate(30-Year)***
  - ***GDP and Mortgage Rate(30-Year)***

Interpretation: Both visual and non-visual understanding of the data:

- ***The data has 81 rows and 17 columns with 12 Floats, 4 Ints and 1 Date as the data types.***
- ***The total number of elements in the dataset is 1377.***
- ***Our target variable is Price($), which is continuous and non-null.***
- ***The total number of missing values in the dataset is 0.***
- ***We see from the data record the max of 7% Mortgage Rate(30-Year).***
- ***After the year 2008 there is a dip in homeowner vacancy and a rise in home prices.***
- ***We will be dropping the ID columns, Date to move forward.***

3. Data Cleaning and Pre-processing:

- ***We will start by dropping the ping DATE, as this is an identification variable we will not need them for model building.***

Outlier treatment (if required):

Homes for sale, Unemployment Rate, and Employment Rate with Outliers:

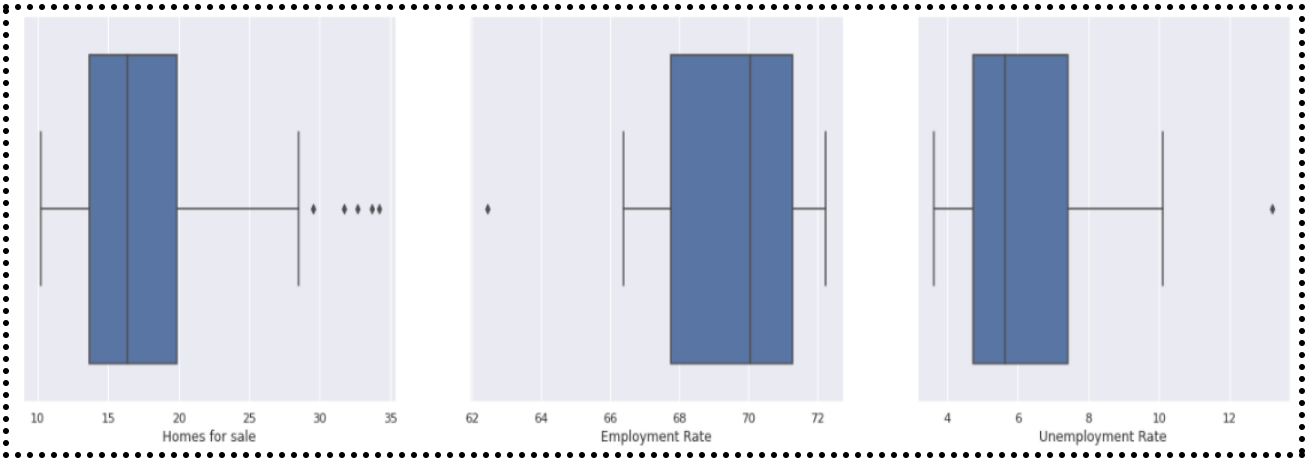


Fig_18: Homes for sale, Unemployment Rate, and Employment Rate with Outliers

- ***These 3 variables along with Homes for sale, Unemployment Rate, and Employment Rate show outliers, we will not be treating the outliers, as they are few and could impact the output.***

Inference:

Our target Price($) variable is a continuous variable, it is positively right-skewed with, Skewness at: 1.08.

- ***GDP is directly proportional to Price($), it increases with GDP. While a decrease in Rental vacancy indicates a price rise for homes.***

- ***Consumer Prices for Housing, Water, Electricity, Gas, and Other Fuels over the years had a positive effect on US home prices.***
- ***As for demand factors, we can look into variables like:***
  - ***Unemployment Rate***
  - ***Mortgage Rate(30-Year)***
  - ***GDP***
  - ***Working Age Population***
- ***As for supply factors, we can look into variables like:***
  - ***Building Permits***
  - ***Homes for sale***
  - ***Rental Vacancy***
  - ***Homeowner Vacancy Rate***

4. Model building and interpretation.

Our target Price($) variable is a continuous variable, the models we are going to build are regression models listed below.

- ***Linear Regression.***
- ***Decision Tree Regression.***
- ***Random Forest Regression.***
- ***ANN.***
- ***Ensemble model.***

Linear Regression

- ***Variables:*** `'Unemployment Rate', 'Employment Rate', 'GDP', 'Working Age Population', 'Dwellings and Residential Buildings', 'Essential utilities Cost average', 'Housing starts', 'No. Housing units', 'Mortgage Rate(30-Year)', 'Homes for sale', 'Building Permits', 'Rental Vacancy', 'Homeownership Rate', 'Homeowner non-natives', 'Homeowner Vacancy Rate '`
- ***Splitting data into Train and Test at the default radio of 30:70% with the random state as 123.***
- ***After splitting the data into test and train, distribution is:***
  - `X_train (56, 15)`
  - `X_test (25, 15)`

- ***y_train (56,)***
- ***y_test (25,)***

Model Iteration:1

- ***Linear Regression Model score for R Square:***
  - ***The model score on traing set: 0.981***
  - ***The model score on test set: 0.977***
- ***97.7% of the variation in the Price is explained by the predictors in the model for the test set.***
- ***Linear Regression Model score for Adj R Square:***
  - ***Combining X and Y to get one dataset for Adj R Square, implemented using stats model.***

Regression model summary:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.981
Model:	OLS	Adj. R-squared:	0.974
Method:	Least Squares	F-statistic:	140.3
Date:	Fri, 19 Aug 2022	Prob (F-statistic):	1.06e-29
Time:	13:09:07	Log-Likelihood:	-229.31
No. Observations:	56	AIC:	490.6
Df Residuals:	40	BIC:	523.0
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	250.1359	120.975	2.068	0.045	5.637	494.635
Unemployment_Rate	-111.3935	112.490	-0.990	0.328	-338.743	115.956
Employment_Rate	-122.5917	103.076	-1.189	0.241	-330.917	85.733
GDP	-103.7944	208.716	-0.497	0.622	-525.625	318.037
Working_Age_Population	-385.3882	98.258	-3.922	0.000	-583.975	-186.801
Dwellings_and_Residential_Buildings	39.8125	103.488	0.385	0.702	-169.344	248.969
Essential_utilities_Cost_average	340.7700	209.924	1.623	0.112	-83.502	765.042
Housing_starts	59.5194	101.947	0.584	0.563	-146.523	265.562
No_Housing_units	610.9284	348.446	1.753	0.087	-93.306	1315.163
Mortgage_Rate30_Year	44.0723	39.554	1.114	0.272	-35.869	124.014
Homes_for_sale	47.9450	35.859	1.337	0.189	-24.529	120.419
Building_Permits	154.5452	129.865	1.190	0.241	-107.922	417.013
Rental_Vacancy	-21.4312	26.471	-0.810	0.423	-74.932	32.070
Homeownership_Rate	4.2978	61.063	0.070	0.944	-119.114	127.710
Homeowner_non_natives	-1.4292	43.202	-0.033	0.974	-88.744	85.885
Homeowner_Vacancy_Rate	100.4833	45.415	2.213	0.033	8.697	192.270

Omnibus:	6.718	Durbin-Watson:	1.767
Prob(Omnibus):	0.035	Jarque-Bera (JB):	8.725
Skew:	0.343	Prob(JB):	0.0127
Kurtosis:	4.808	Cond. No.	432.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig_19: OLS 1st Iteration

Interpretation:

- ***According to our result above, the p-value is greater than alpha, hence null hypothesis is accepted, we will have to further tuning our formula to get a better result.***
- ***For this model, we are removing the variables with high p values than alpha.***

`Iteration:2`

- ***For this model, we are removing the variables with higher p values than alpha 0.05.***
- ***Variables we are taking will be:***
- ***Working_Age_Population+Essential_utilities_Cost_average+No_Housing_units+Homes_for_sale+Building_Permits+Homeowner_Vacancy_Rate***

Regression model summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.974
Method:                 Least Squares   F-statistic:                     350.7
Date:                Fri, 19 Aug 2022   Prob (F-statistic):           1.82e-38
Time:                        13:24:21   Log-Likelihood:                -234.88
No. Observations:                  56   AIC:                             483.8
Df Residuals:                      49   BIC:                             497.9
Df Model:                           6
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
Intercept                               125.5292     23.965      5.238      0.000      77.371     173.688
Working_Age_Population                 -310.2474     54.219     -5.722      0.000    -419.204    -201.290
Essential_utilities_Cost_average        298.1787    140.065      2.129      0.038      16.708     579.649
No_Housing_units                        477.9991    132.668      3.603      0.001     211.393     744.605
Homes_for_sale                           63.2452     26.577      2.380      0.021       9.836     116.654
Building_Permits                        262.3277     15.409     17.024      0.000     231.362     293.294
Homeowner_Vacancy_Rate                  102.5559     32.697      3.137      0.003      36.850     168.262
==============================================================================
Omnibus:                        4.659   Durbin-Watson:                   1.728
Prob(Omnibus):                  0.097   Jarque-Bera (JB):                5.660
Skew:                           0.067   Prob(JB):                       0.0590
Kurtosis:                       4.552   Cond. No.                         135.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig_20: OLS 2nd Iteration

Interpretation:

- ***The overall P-value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression coefficient is not 0. Here all regression coefficients are not 0.***
  - ***The model RMSE on training set: 16.045***
  - ***The model RMSE on the test set: 11.78***

- *price= (125.53) * Intercept + (-310.25) * Working_Age_Population + (298.18) * Essential_utilities_Cost_average + (478.0) * No_Housing_units + (63.25) * Homes_for_sale + (262.33) * Building_Permits + (102.56) * Homeowner_Vacancy_Rate*

Decision Tree Classifier

- *Creating a model with default feature value with criterion as Gini (It is calculated by subtracting the sum of squared probabilities of each class from one.)*
- *The parameters are: { criterion="squared_error", splitter="best", max_depth=3, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_features=None, random_state=123, max_leaf_nodes=None, min_impurity_decrease=0, ccp_alpha=0 }*
- *Decision Tree Regressor Model score for R Square:*
  - *The model R-squared score on train set: 0.928*
  - *The model R-squared score on test set: 0.921*
- *Mean_squared_error*
  - *The model MSE on train set: 446.6*
  - *The model MSE on test set: 446.6*
- *Root_mean_squared_error*
  - *The model RMSE on train set: 21.133*
  - *The model RMSE on test set: 38.813*

Interpretation:

- *As we see, the accuracy of training and test results are very close to each other. Training data was 92.9%, followed by test data at 92.1%.*

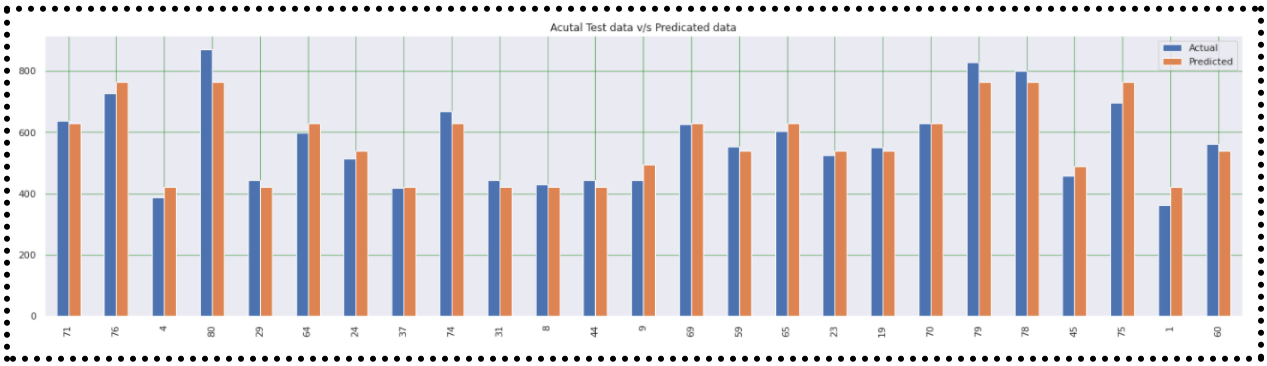


Fig_21: DT_Acutal v/s data

Random Forest Regressor

- *Creating a model with default feature value with criterion as squared_error(It is for the mean squared error, which is equal to variance reduction as feature selection criterion.)*

- ***The parameters are: { n_estimators=100, *, criterion="squared_error", max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_features="auto", max_leaf_nodes=None, min_impurity_decrease=0, bootstrap=True, oob_score=False, n_jobs=None, random_state=123 , verbose=0, warm_start=False, ccp_alpha=0, max_samples=None }***
- ***Random Forest Regressor Model score for R Square:***
  - `The model R-squared score on train set:  0.973`
  - `The model R-squared score on test set:  0.821`
- ***Mean_squared_error***
  - `The model MSE on train set:  166.302`
  - `The model MSE on test set:  166.302`
- ***Root_mean_squared_error***
  - `The model RMSE on train set:  12.896`
  - `The model RMSE on test set:  58.357`

Interpretation:

- ***As we see, the accuracy of training and test results are very close to each other. Training data is 97.3%, followed by test data at 82.1%.***
- ***Records show us that the Prediction closely follows our test sample value, which is great for our model.***

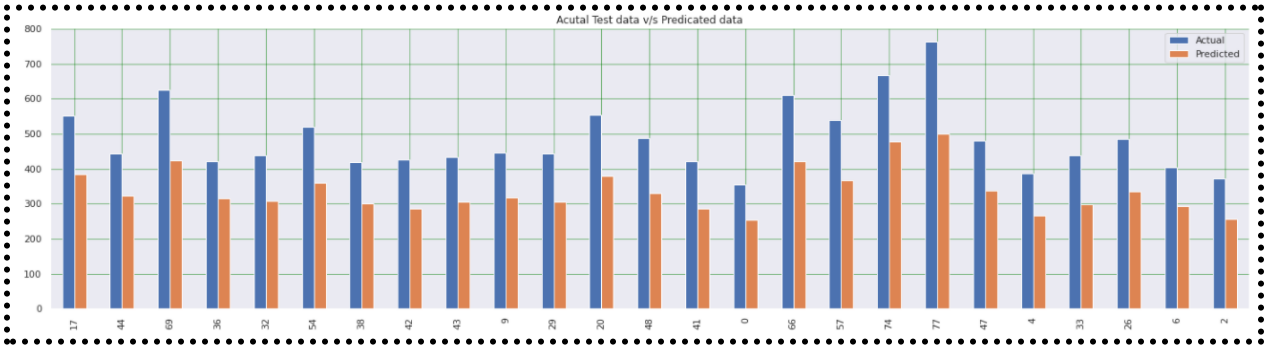


Fig_22: RF_Acutal v/s data

Artificial Neural Network Regressor

- ***We will start with scaling our X independent variables.***
- ***Creating a model with default feature value with activation="relu"(It is the rectified linear unit function, returns f(x) = max(0, x))***

- ***The parameters are: { hidden_layer_sizes=(100,), activation="relu", *, solver="adam", alpha=0.0001, batch_size="auto", learning_rate="constant", learning_rate_init=0.1, power_t=0.5, max_iter=200, shuffle=True, random_state=123, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-8, n_iter_no_change=10, max_fun=15000 }***
- ***Random Forest Regressor Model score for R Square:***
  - ***`The model R-squared score on train set:  0.958`***
  - ***`The model R-squared score on test set:  0.961`***
- ***Mean_squared_error***
  - ***`The model MSE on train set:  480.314`***
  - ***`The model MSE on test set:  480.314`***
- ***Root_mean_squared_error***
  - ***`The model RMSE on train set:  21.916`***
  - ***`The model RMSE on test set:  18.994`***

Interpretation:

- ***As we see, the accuracy of training and test results are very close to each other. Training data is 95.8%, followed by test data at 96.1%. Let's improve it.***
- ***Records show us that the Prediction closely follows our test sample value, which is great for our model.***

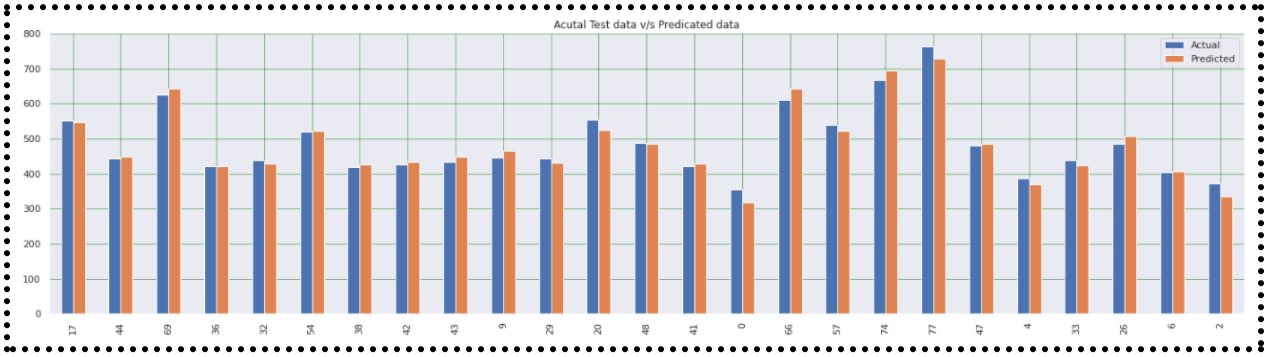


Fig_23: ANN_Acutal v/s data

Ensemble modeling:

- ***We have created an Ensemble model for our dataset, using VotingRegressor.***
- ***With three regression models:***
  - ***Linear Regression***
  - ***Decision Tree Regressor***

➢ *Artificial Neural Network Regressor*

- ***The ensemble model is created to get better output with simple base versions of the above models.***
- ***Our model has given a score of 98%. Which is higher when compared to other models we have built so far.***
- ***We have used Average on prediction to get our output. (prediction_model1+prediction_model2+prediction_model3+prediction_model4)/4***
- ***The final ensemble model predictions on test set value is as follows: [548.78436601 461.17846531 623.61814652 442.67701968 437.06142252 521.78049382 429.11436487 420.06537604 442.44421161 457.96624981 434.85209225 537.28076917 479.8272219 415.66255124 347.66241365 621.02309131 529.90405792 694.6778449 729.19026347 486.50291849 375.74050218 425.94018125 488.59080748 413.9343083 356.44851683]***

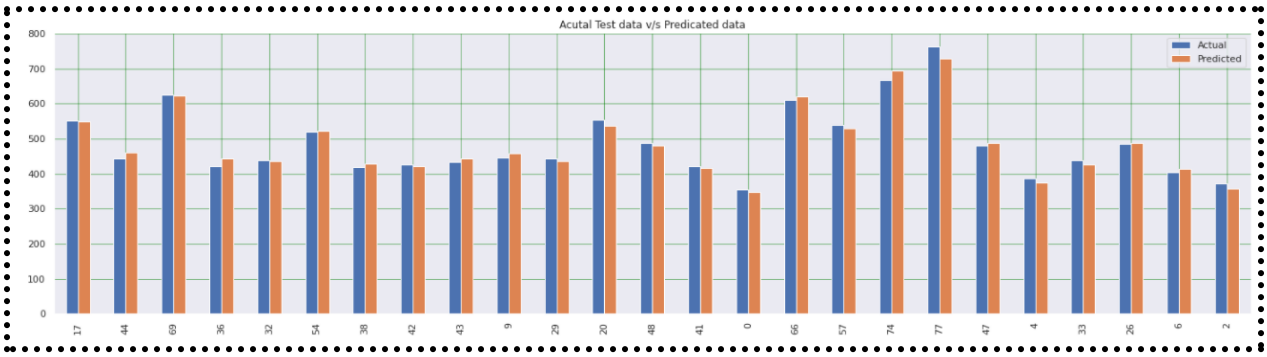


Fig_24: Ensemble_Acutal v/s data

5. Model validation

Table_4: Final comparison for all the models.

	LR Train	LR Test	Decision Tree Train	Decision Tree Test	Random Forest Train	Random Forest Test	ANN Train	ANN Test	Emsemble Train	Emsemble Test
MAE	10.94	8.63	15.11	30.38	5.4	30.01	15.2	15.32	7.9	10.72
MSE	211.01	138.76	446.6	1506.48	166.3	3405.49	480.31	360.76	123.23	174.33
RMSE	14.53	11.78	21.13	38.81	12.9	58.36	21.92	18.99	11.1	13.2
Score	0.98	0.98	0.93	0.92	0.97	0.82	0.96	0.96	0.99	0.98

- ***Mean absolute error(MAE): it measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables.***

- ***Mean Squared Error (MSE): The Mean Squared Error (MSE) is perhaps the simplest and most common loss function, often taught in introductory Machine Learning courses. It takes the difference between your model's predictions and the ground truth, square it, and average it out across the whole dataset.***
- ***Root Mean Square Error(RMSE): Is the square root of the mean of the square of all of the errors. The use is very common, and it is considered an excellent general purpose error metric for numerical predictions.***
- ***The final Linear Regression equation is at 98% score:***
  - ***Expected_CTC = (125.53) * Intercept + (-310.25) * Working_Age_Population + (298.18) * Essential_utilities_Cost_average + (478.0) * No_Housing_units + (63.25) * Homes_for_sale + (262.33) * Building_Permits + (102.56) * Homeowner_Vacancy_Rate***
- ***R-sqaure score: Is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).***
- ***We have the best model with the best score and the least error is our Linear Regression with a score of around 98%, would recommend going ahead with it.***
- ***This is our measure of model quality. In this example, we can say that our model predictions are off by about $138.76. The lower the error, the better for us.***
- ***Random Forest gives us better results than Decision Tree.***

6. Final interpretation/recommendation.

Inference :

- ***We have the best model with the best score and the least error is our linear Regression can go ahead with it.***
- ***The variables that affect the US home Prices, most over the past 20 years are:***
  - ***Working_Age_Population***
  - ***Essential_utilities_Cost_average***
  - ***No_Housing_units***
  - ***Homes_for_sale***
  - ***Building_Permits***
  - ***Homeowner_Vacancy_Rate***
- ***When Working_Age_Population increases by 1 unit, Price decreases by 310.25 units, keeping all other predictors constant.***
- ***Similarly, when Essential_utilities_Cost_average increases by 1 unit, Price increases by 298.18 units, keeping all other predictors constant, and so on.***

- ***The feature that has the most effect on the US home Price is No_Housing_units that are occupied or intended for occupancy as separate living quarters.***
- ***This is our measure of model quality. In this example, we can say that our model predictions are off by about $138.76. The lower the error, the better for us. More data can give us more improved prediction model.***
- ***Can be used to effectively understand and work on the US homing market.***
- ***Mortgage_Rate30_Year is indirectly proportional to Price, this further proves the point we draw from EDA.***
- ***By analyzing the past we can make better decision and give advice which may result in increased and faster transaction and deal in the housing market, making a positive impact on revenue.***
- ***Also decrease the buyer cancellation or fall-throughs in property deals.***

END !