

Data Science Project

Deep Shah, Gaurung Vasani, and Dean Wise

ENGR 241: Probability and Statistics with Data Science Applications

26 April 2024

“We pledge our honor that we have abided by the Stevens Honor System.” ~ DS, GV, DW

Table of Contents

Introduction to the data (source, description of variables)	4
Exploratory Data Analysis	5
Graphs	5
Histogram of Number of Houses Based on the House Prices	5
Scatter Plot of Price vs. Area with SLR Line	6
Bar Plot of Number of Houses With or Without Hot Water Heating	6
Pie Chart of Houses Based on Furniture	7
Housing Price Statistics for 1-Variable from Code Output	7
Discussion	7
Missing or incorrect values and how you would fix them in such a scenario	7
Histogram	8
Descriptive statistics for at least one variable	8
Scatter plot	9
Bar Graph	9
Pie Chart	10
Linear Regression Modeling	10
Select a significance level and fit a Simple/Multilinear Regression Model	10
Simple Linear Regression Model of Price vs Area	10
Equation for Simple Regression Model of Price vs Area	11
ANOVA Table for Simple Linear Regression	11

	3
Multilinear Regression Model of Price vs Area and Parking	12
Equation for MultiLinear Regression Model of Price vs Area and Parking	12
ANOVA Table for MultiLinear Regression Model	13
Discussion for Simple Regression Model and MultiLinear Regression Model:	13
Interpretation and Conclusions	14
Lessons Learned	15
Appendix	16
Link to Google Colab: Python Code for Data Science Project:	16

Introduction to the data (source, description of variables)

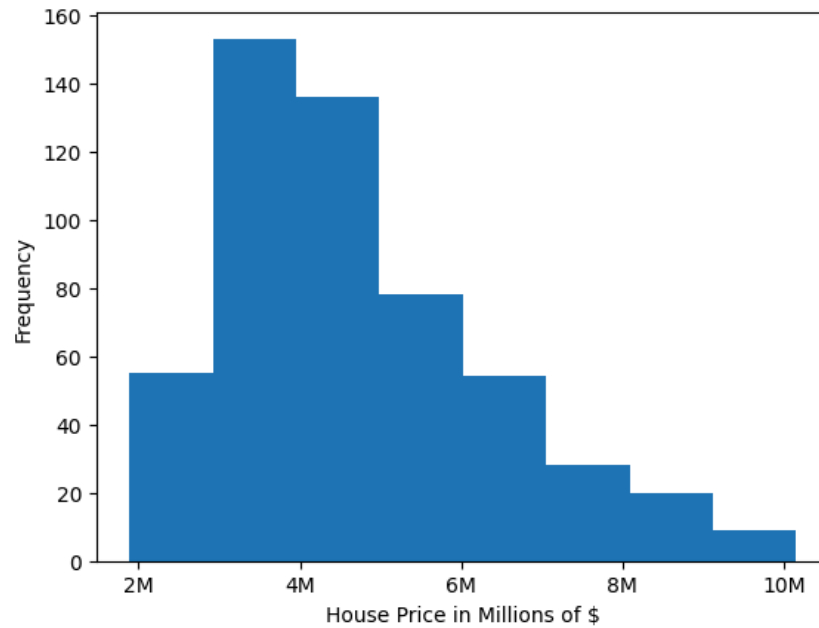
The purpose of this project is to utilize concepts in probability and statistics to make meaningful analyses of given data sets and identify any possible relations between variables. For this project, we have chosen the housing data set, taking into consideration several variables such as hot water heating, air conditioning, as well as parking to better understand the given data set as a whole.

Python and its Pandas library will be utilized to organize this large dataset and produce several graphs to better understand and visualize the given data. For example, a histogram will be used to determine the amount of houses that fall within specific ranges of prices. A scatterplot is utilized in this case to examine the relationship between the size of a house and its price, along with any possible linear relationships these variables may have. Finally, linear regression models are leveraged to conduct significance testing and to find out whether there is any statistical significance between several variables. In this analysis, housing price and parking data will be analyzed with the dependent variable being the housing area. Following the creation of regression models, an ANOVA table will be constructed to display several statistical parameters, including the p-value, the F-statistic, and the sum of mean squares, all of which will be used to determine the statistical significance (or lack thereof) of the data.

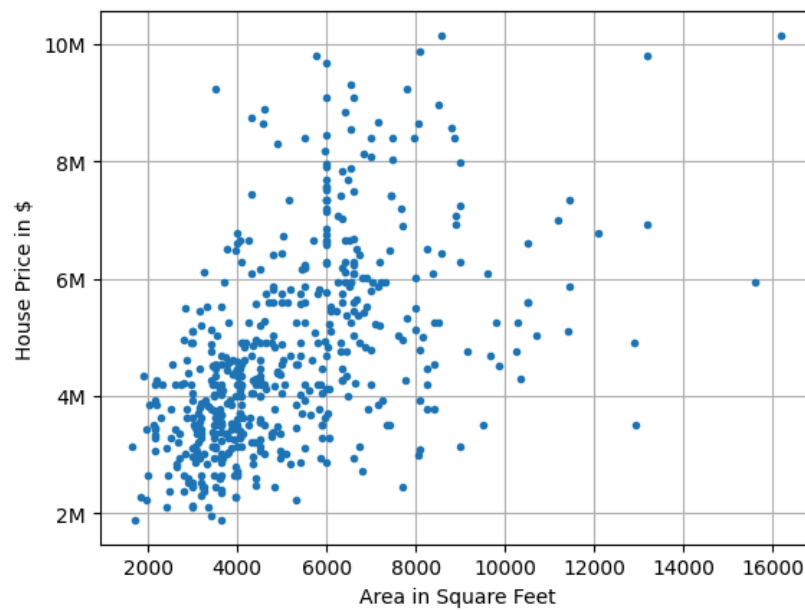
Exploratory Data Analysis

Graphs

Histogram of Number of Houses Based on the House Prices



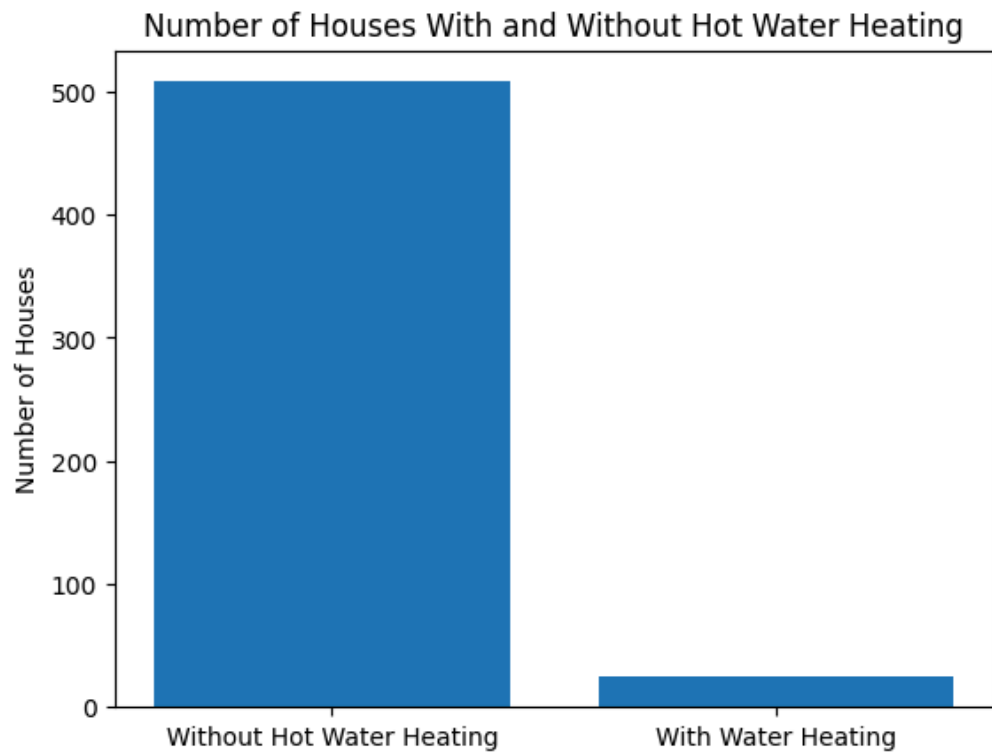
Scatter Plot of Price vs. Area



Scatter Plot of Price vs. Area with SLR Line

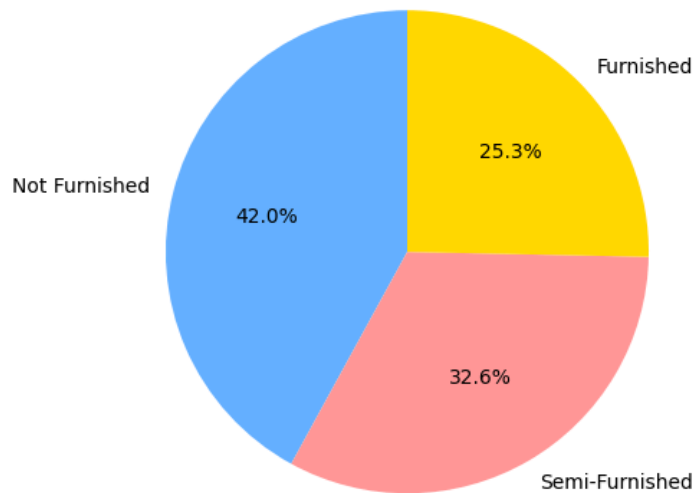


Bar Plot of Number of Houses With or Without Hot Water Heating



Pie Chart of Houses Based on Furniture

Percentage of Houses That are Fully Furnished, Semi-Furnished, and Not Furnished in Housing Dataset



Housing Price Statistics for 1-Variable from Code Output

Discussion

Missing or incorrect values and how you would fix them in such a scenario

In the event that a dataset has missing or incorrect values, there are several methods that can help mitigate issues that would arise from this improper data. It is important to remember that such errors would not be as impactful on the values of statistical parameters if a set of data is larger. According to the Law of Large Numbers, the average of a set of data converges to its true value as the amount of points becomes infinitely large. Thus, one solution could be taking more data points and masking the impact of an incorrect value. Another way to actually transform such values would be to identify which values are incorrect and replace them with the mean. This is a technique known as imputation, where an incorrect value is replaced by a value that better reflects a majority of the set. When performing imputation, it is also possible to replace the value with the median or mode based on what would make the most sense for that specific dataset.

Histogram

The histogram above describes the frequencies of the ranges of house prices in the horizontal axis. Through the creation of this, we can interpret the data accordingly as to which price is more preferred or common to purchase. As seen in the histogram, the distribution of frequencies for houses purchased is skewed right, which indicates a preference for cheaper houses. The peak frequency for a given range is from 3 to 4 million dollars, which is towards the left of the x-axis. Through this graph, it can be inferred that cheaper houses are bought more often. This makes sense in the context of what is going on because cheaper housing is more accessible to people and will naturally be bought more frequently than more expensive houses.

Descriptive statistics for at least one variable

One variable to consider in this data set is house price, the mean and standard deviation of which is displayed above. Price is an important variable for the consumers as a deciding factor for purchasing a certain house. According to the scatter plot, as well as the histogram, people tend to purchase cheaper houses more frequently. In the histogram, about 160 people chose houses in the range between 3-4 million dollars, which is the highest frequency for a given range in this set of data. Of the possible houses for sale in the area, these houses are on the cheaper end. The mean is around 4.7 million dollars, and the standard deviation is 1.7 million dollars. This means that while the mean price is relatively cheap, the prices of houses generally vary by a large range. In this set of data, the price distribution is skewed right, which implies that people are more compelled to purchase cheaper properties, and that the frequency with respect to price does not follow a normal distribution.

Scatter plot

The scatter plot above indicates the housing price in dollars for various area sizes of houses. Although there are some outliers, the trend in the plot is positively skewed indicating a direct relationship between the two axes (the more area size of the house, the higher its price). Within the Area Size of Houses ranging from 2000-6000 with the housing price ranging from approximately \$2,000,000.00 to approximately \$6,000,000.00, a variety of data is plotted, suggesting this is the typical range of values people prefer to purchase houses within. As stated before, as the area size of the house increases, so does the price. As such, this can be a factor for which people have certain preferences of housing, especially for outlier data such as the point near (16000, 0.6). The way data is plotted in the scatter plot makes sense as the distribution of frequency of house prices is more for lower prices between $0.2-0.6 \times 10^7$ in the histogram. Finally, apart from the histogram, a scatter plot allows for data to be separated and viewed more accurately, but having both, especially for different variables on each axis, is useful for proper interpretation of the data.

Bar Graph

We created a bar plot above that represents how many houses in the data set have hot water heating or how many don't. The bar graph has two bars, one that represents houses with no hot water heating and the other with hot water heating. As per the bar plot above, the number of houses with no hot water heating is more than the number of houses with hot water heating. This is important in this data as this could either be a preference of the people or a supply issue of housing due to cheaper prices.

Pie Chart

We created a pie chart above that represents the comparison of the percentage of houses that are fully furnished items, semi-furnished items, and are not furnished. Particularly, the pie chart shows that 42.0% of the houses are not furnished, 32.6% the semi-furnished, and 25.3% of the houses are furnished. The significance of the data set based on the pie chart is that many of the houses they chose either be a preference of the people or the appearance of the house due to cheaper prices.

Linear Regression Modeling

Select a significance level and fit a Simple/Multilinear Regression Model

Throughout this project, a significance level of $\alpha = 0.05$ will be used.

Simple Linear Regression Model of Price vs Area

```
[74] SLR=sm.ols(formula = 'Housing.area ~ Housing.price', data = Housing).fit()
      SLR.summary()
```

```

OLS Regression Results

Dep. Variable:   Housing.area   R-squared:    0.273
Model:          OLS            Adj. R-squared: 0.272
Method:         Least Squares   F-statistic:   199.5
Date:           Thu, 25 Apr 2024 Prob (F-statistic): 1.10e-38
Time:           23:50:58        Log-Likelihood: -4762.5
No. Observations: 533          AIC:               9529.
Df Residuals:    531           BIC:               9538.
Df Model:         1

Covariance Type: nonrobust

   coef    std err   t    P>|t| [0.025   0.975]
Intercept  1999.3588  236.214   8.464   0.000  1535.331  2463.387
Housing.price  0.0007   4.71e-05  14.125   0.000   0.001   0.001

Omnibus:    166.858   Durbin-Watson:   1.982
Prob(Omnibus): 0.000   Jarque-Bera (JB): 495.948
Skew:        1.501     Prob(JB):         2.02e-108
Kurtosis:    6.649     Cond. No.         1.48e+07

```

Notes:

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

[2] The condition number is large, 1.48e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Equation for Simple Regression Model of Price vs Area

$$y = 0.0007x_1 + 1999.3588$$


ANOVA Table for Simple Linear Regression

Anova Table for Simple Linear Regression

```

▶ import statsmodels.formula.api as smf
  from statsmodels.stats.anova import anova_lm
  SLR = smf.ols('Housing.price ~ Housing.area', data = Housing).fit()
  anova_table = anova_lm(SLR)
  print(anova_table)

```



	df	sum_sq	mean_sq	F	PR(>F)
Housing.area	1.0	4.170461e+14	4.170461e+14	199.501504	1.096662e-38
Residual	531.0	1.110024e+15	2.090441e+12	NaN	NaN

According to the data above which is relating the housing area to housing price, the model will not be strongly linear due to its R^2 value. In this simple linear regression model, the data is statistically significant because of a very low p-value and high F statistic for the housing area variable. This comes directly from the ANOVA table for SLR, which lists these variables and corroborates the fact that the data is indeed significant.

Multilinear Regression Model of Price vs Area and Parking

MultiLinear Regression Model of Price vs Area and Parking

```
[77] MLR=sm.ols(formula = 'Housing.price ~ Housing.area + Housing.parking', data = Housing).fit()  
MLR.summary()
```



OLS Regression Results

Dep. Variable:	Housing.price	R-squared:	0.304	
Model:	OLS	Adj. R-squared:	0.301	
Method:	Least Squares	F-statistic:	115.8	
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	1.93e-42	
Time:	23:50:59	Log-Likelihood:	-8303.9	
No. Observations:	533	AIC:	1.661e+04	
Df Residuals:	530	BIC:	1.663e+04	
Df Model:	2			

Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.61e+06	1.59e+05	16.460	0.000	2.3e+06	2.92e+06
Housing.area	360.9828	30.214	11.948	0.000	301.629	420.337
Housing.parking	3.732e+05	7.69e+04	4.854	0.000	2.22e+05	5.24e+05

Omnibus: 32.412 **Durbin-Watson:** 0.586

Prob(Omnibus): 0.000 **Jarque-Bera (JB):** 36.908

Skew: 0.600 **Prob(JB):** 9.67e-09

Kurtosis: 3.470 **Cond. No.** 1.44e+04

Equation for MultiLinear Regression Model of Price vs Area and Parking

$$y = 360.9828x_1 + 3.732 * 10^5 x_2 + 2.61 * 10^6$$

ANOVA Table for MultiLinear Regression Model

ANOVA Table for MultiLinear Regression Model

```
▶ anova_table3 = anova_lm(MLR)
  print(anova_table3)
```

	df	sum_sq	mean_sq	F	PR(>F)
Housing.area	1.0	4.170461e+14	4.170461e+14	207.976884	5.179996e-40
Housing.parking	1.0	4.724046e+13	4.724046e+13	23.558362	1.596474e-06
Residual	530.0	1.062784e+15	2.005252e+12	NaN	NaN

Discussion for Simple Regression Model and MultiLinear Regression Model:

Based on the multilinear regression model between the housing price and parking parameters, an ANOVA chart was derived, which indicates several statistical values that describe the significance of the data set(s). The first values that stick out are the “PR (>F)” values in the last column. These values are extremely low for both categories, which indicates that there is statistical significance between both variables and the area in which the house is located. These p-values are low because according to the F tables for a significance level of 0.05, the F values calculated are much larger than the Fcrit values for the corresponding degrees of freedom. To further indicate statistical significance, the calculated t-stat values in the multilinear model are greater than the t-crit values at their respective degrees of freedom. Although it is true that R^2 is quite low, the regression model also indicates a very high condition number, which points towards a stable set of solutions to a linear equation. With all these variables considered, it can be inferred that there is statistical significance within this model.

Interpretation and Conclusions

As per the given set, we have transformed the data into various models using Python which provide information regarding housing and several of its variables. Overall, as per the interpretation of the models, the histogram, the scatter plot, the Simple Linear Regression Model, the Multilinear Regression Model, as well as the Anova Table, we have made distinct interpretations of each variable relationship and their implications for people's choices (or the dataset). As per the histogram, data is positively skewed when comparing the frequencies to the housing price. The highest frequency is between \$3,000,000 to \$4,000,000, while the lowest frequencies range from prices surrounding \$8,000,000 to \$12,000,000. Further, in the scatter plot, as the Area size of houses increase, the housing price increases. As people mostly stay within the ranges 2000 to 8000, within the price of \$2,000,000 to \$6,500,000, the data is shown to be positive displaying a direct relationship in the increase in both variables. From this data set and its corresponding results, it could be concluded that people gravitate towards cheaper prices, the price and parking are not related in a linear matter but are statistically significant variables, and that a majority of houses do not have hot water heating. Through this new information, the town could take these discoveries and make the right improvements for the housing community to flourish more.

Lessons Learned

This project provided students with the opportunity to apply acquired statistical knowledge and Python programming skills to a real world project involving data analysis of a set. By performing exploratory data analysis and linear regression modeling, students were able to understand and interpret datasets while developing insights about the data through statistical analysis. This project also enabled students to collaborate with one another and prepare this project along with a presentation, which helps strengthen the teamwork and communicative skills that are necessary for engineers in industry later on. Overall, the projects and the tasks associated with it helped a great deal in learning the real-world applications of probability, statistics, and data science, along with professional skills that are necessary in the workplace. It encourages students to utilize their existing knowledge base and critically think in order to understand what is going on in a set of data. Professor Jagupilla did a great job in designing this experience, as it was an extremely helpful tool for understanding the importance of the content learned throughout this semester.

Appendix

Link to Google Colab: Python Code for Data Science Project:

<https://colab.research.google.com/drive/1OnUm6fL002-TF9PnzQNgE2xTPO1tUOsK?usp=sharing>