Housing Dataset

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Deep Shah

Gaurung Vasan

Dean Wise

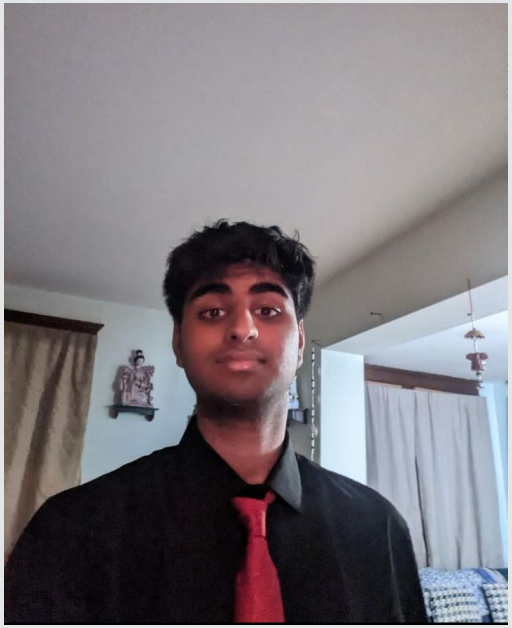
“We pledge our honor that we have abided by the Stevens Honor System.” ~ DS, GV, DW

Group Members

Deep Shah

2/4 Computer Engineer

ENGR 241 - WS



Gaurung Vasan

1/4 Software Engineer

ENGR 241 - B

Dean Wise

2/4 Chemical Engineer

ENGR 241 - WS

Introduction

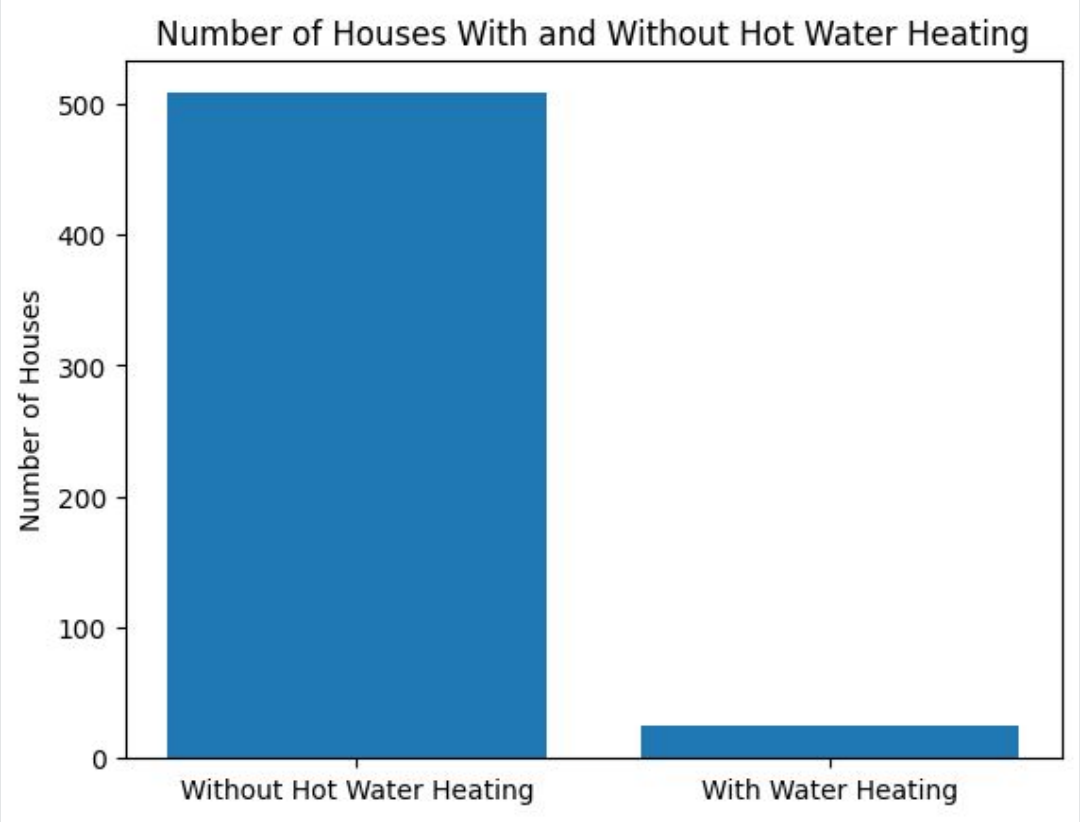
- Make meaningful analysis of housing data set
- Variables: hot water heating, area size, furnished, price
- Use of Python and Panda Library
- Variety of Models:
  - Bar Plot
  - Pie Chart
  - Histogram
  - Scatter Plot
  - Simple Linear Regression
  - Multilinear Regression
  - ANOVA Table -> p-value, F-stat, sum of mean squares

Data From Housing Data

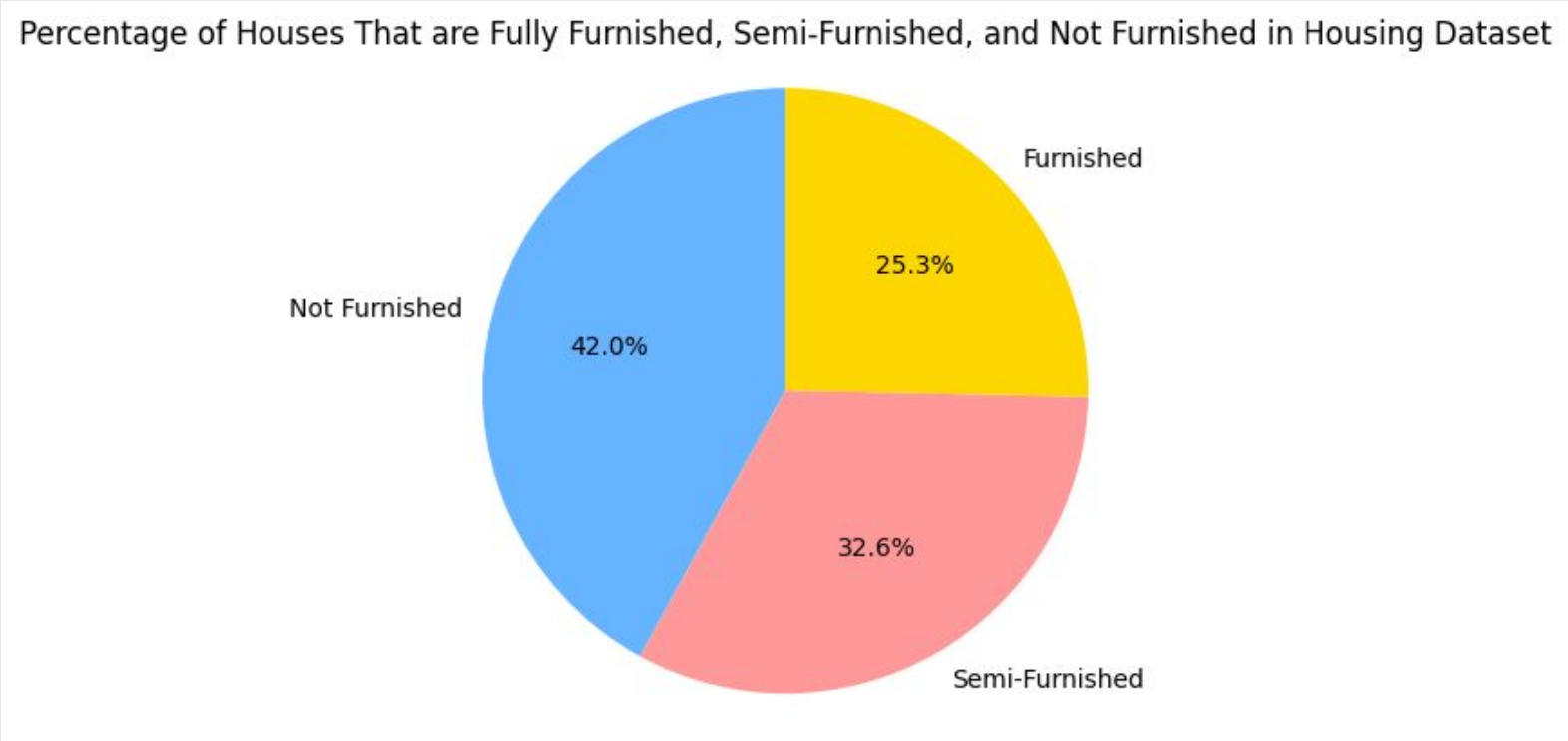
```
Housing.head(15)
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
6	10150000	8580	4	3	4	yes	no	no	no	yes	2	yes	semi-furnished
7	10150000	16200	5	3	2	yes	no	no	no	no	0	no	unfurnished
8	9870000	8100	4	1	2	yes	yes	yes	no	yes	2	yes	furnished
9	9800000	5750	3	2	4	yes	yes	no	no	yes	1	yes	unfurnished
10	9800000	13200	3	1	2	yes	no	yes	no	yes	2	yes	furnished
11	9681000	6000	4	3	2	yes	yes	yes	yes	no	2	no	semi-furnished
12	9310000	6550	4	2	2	yes	no	no	no	yes	1	yes	semi-furnished
13	9240000	3500	4	2	2	yes	no	no	yes	no	2	no	furnished
14	9240000	7800	3	2	2	yes	no	no	no	no	0	yes	semi-furnished
15	9100000	6000	4	1	2	yes	no	yes	no	no	2	no	semi-furnished
16	9100000	6600	4	2	2	yes	yes	yes	no	yes	1	yes	unfurnished
17	8960000	8500	3	2	4	yes	no	no	no	yes	2	no	furnished
18	8890000	4600	3	2	2	yes	yes	no	no	yes	2	no	furnished
19	8855000	6420	3	2	2	yes	no	no	no	yes	1	yes	semi-furnished
20	8750000	4320	3	1	2	yes	no	yes	yes	no	2	no	semi-furnished

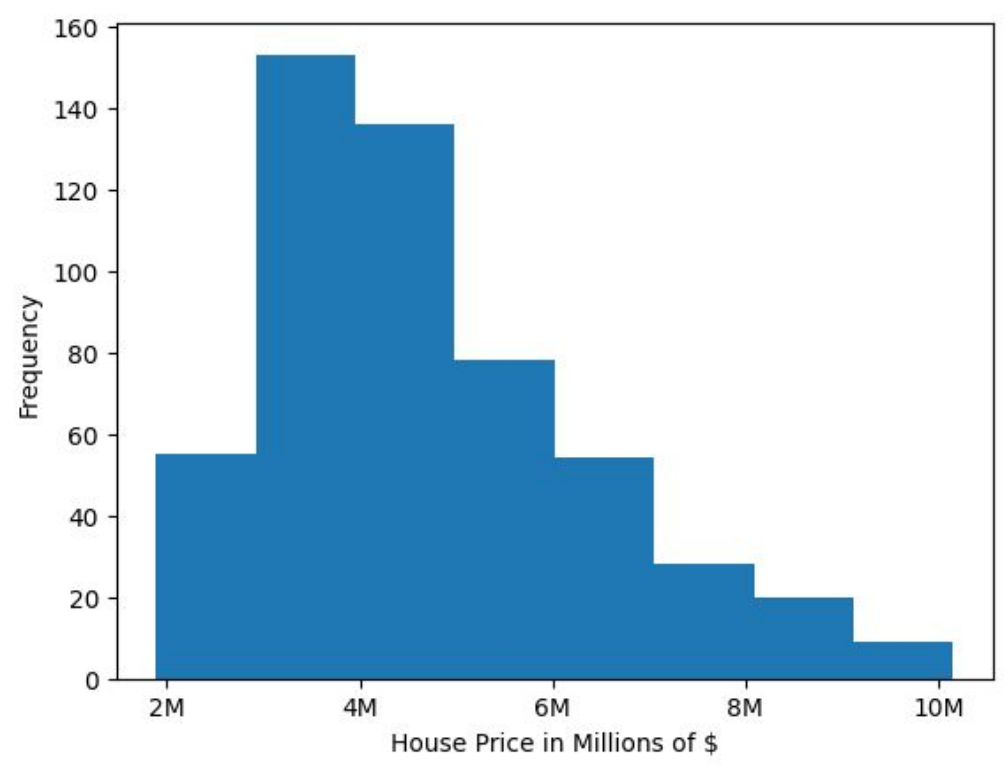
Bar Plot of Number of Houses With or Without Hot Water Heating



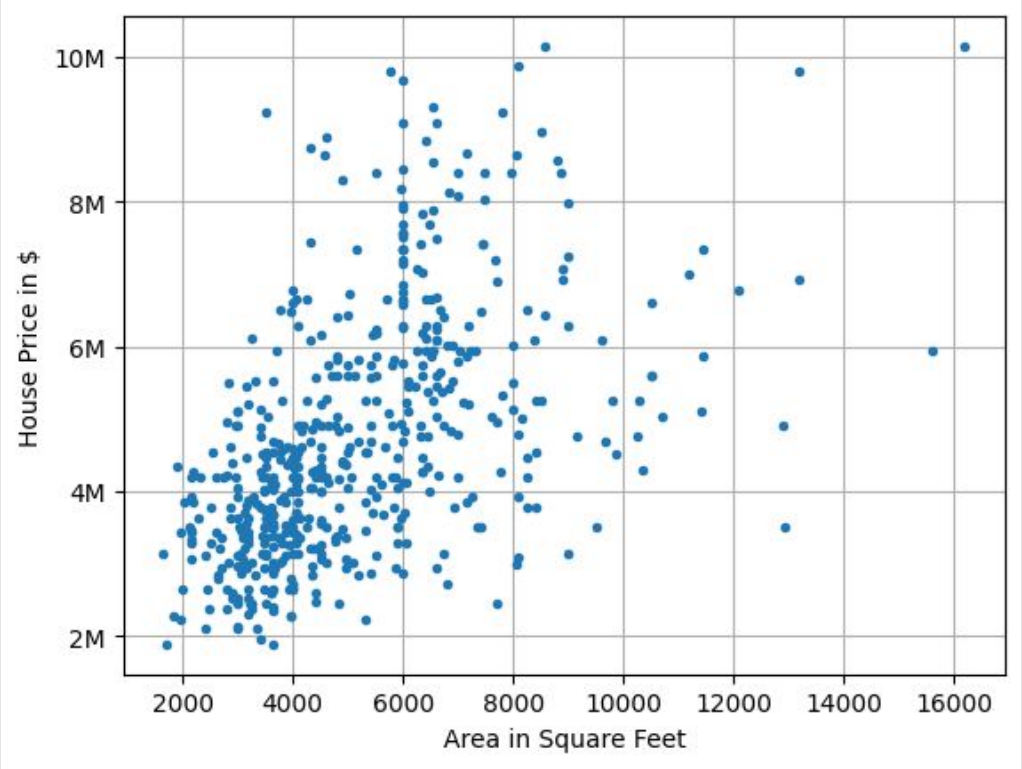
Pie Chart of Houses Based on Furniture



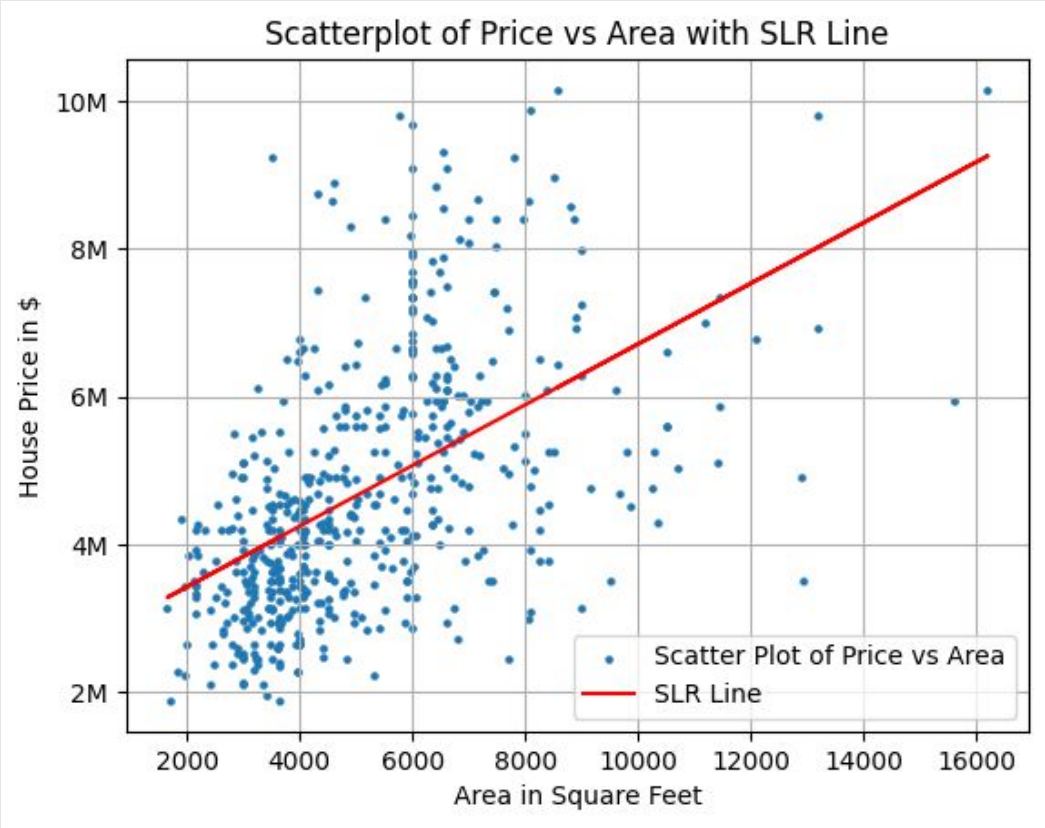
Histogram of Number of Houses Based on the House Prices



Scatter Plot of Price vs. Area



Scatter Plot of Price vs. Area with SLR Line



Simple Linear Regression Model of Price vs Area

Equation for Simple Regression Model of Price vs Area:
y = 0.0007x1+1999.3588

```
[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.

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
```

Multilinear Regression Model of Price vs Area and Parking

Equation for Multilinear Regression Model of Price vs Area and Parking

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

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

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
```

Conclusion

- Concluded that:
  - People gravitate towards cheaper prices
  - Price and parking are not linearly related but are statistically significant and directly proportional
  - Majority of houses in the area do not have hot water heating
- Next steps can include making improvements for the housing community, including implementation of heating and price adjustments

Lesson Learned

This project succeeded in teaching students the following:

- Application of statistical knowledge
- Critical thinking and understanding data
- Professional skills used in industry

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK YOU

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030