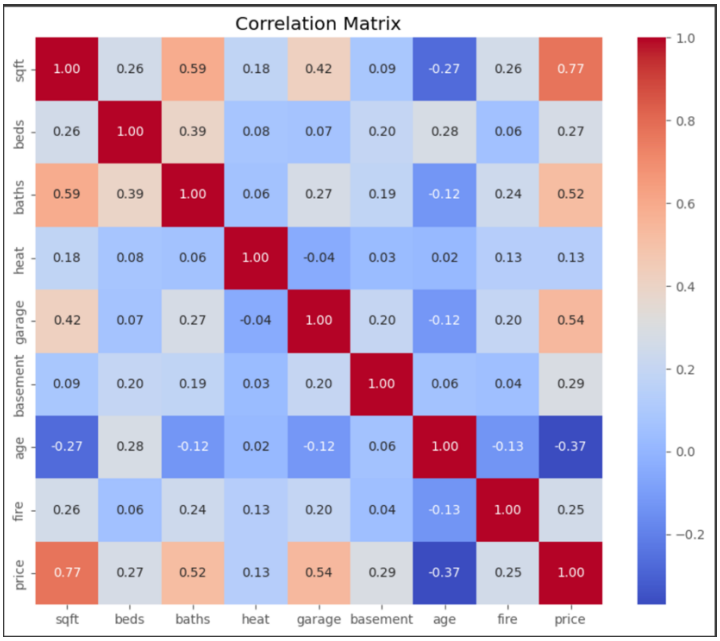
ISOM 352 Group 18: Pranit Aggarwal, Ivy Chen, Caroline Ma, Louisa Ma, and Joshua Qian

ISOM 352 Lab 3 Report

Step 2: Descriptive Analytics

We explored how the numerical variables relate to one another using a correlation heatmap.



We started by separating quantitative and qualitative variables. Then we computed a correlation matrix on the numeric features and visualized it using a heatmap.

From the correlation matrix, we can see that **Sqft and Price** have the strongest relationship (r=0.77), with the positive and large correlation value suggesting that larger homes tend to be more expensive. There is also a moderate positive correlation between Sqft and Bath (r = 0.59) and Baths and Price (r=0.52). **Age and Price** show a negative correlation, showing that older homes tend to be cheaper.

To get a better understanding of the distribution of our quantitative variable, we calculated the **descriptive statistics**:

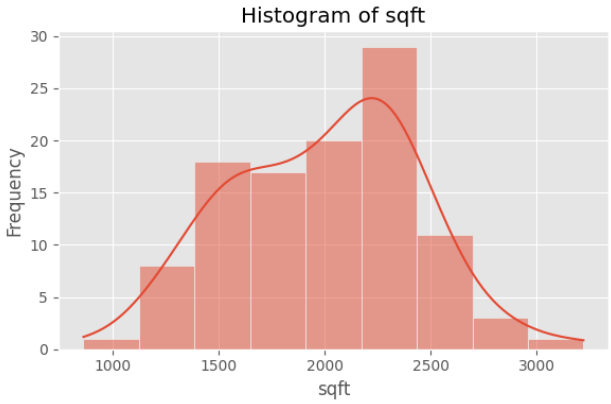
ISOM 352 Group 18: Pranit Aggarwal, Ivy Chen, Caroline Ma, Louisa Ma, and Joshua Qian

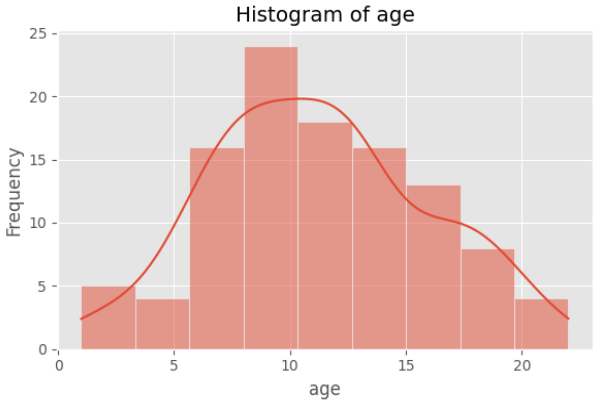
Descriptive Stats for Quantitative Variables:

	count	mean	std	min	25%	50%	75%	max
sqft	108.0	1994.481481	443.710975	861.0	1649.25	2034.5	2304.75	3222.0
beds	108.0	3.601852	0.682763	3.0	3.00	3.5	4.00	6.0
baths	108.0	2.648148	0.552223	2.0	2.00	3.0	3.00	4.0
heat	108.0	0.138889	0.347443	0.0	0.00	0.0	0.00	1.0
garage	108.0	2.037037	0.332814	1.0	2.00	2.0	2.00	3.0
basement	108.0	0.842593	0.365882	0.0	1.00	1.0	1.00	1.0
age	108.0	11.231481	4.655630	1.0	8.00	11.0	14.00	22.0
fire	108.0	0.805556	0.397618	0.0	1.00	1.0	1.00	1.0
price	108.0	341394.629630	82092.910396	197300.0	285017.50	331715.0	377335.00	675030.0

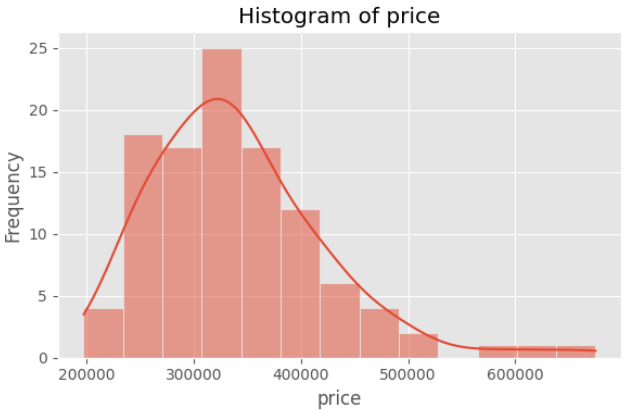
From the summary, we can see that the average house price is around $341,395, ranging from $197,300 to $675,030. Homes also vary in size, with an average of about 1994 sqft and a max of over 3200 sqft. Most homes have 2 to 3 bathrooms and 3 to 4 bedrooms. The average home age is around 11 years.

To better understand how each numerical variable is distributed, we plotted **histograms** for all quantitative features.





ISOM 352 Group 18: Pranit Aggarwal, Ivy Chen, Caroline Ma, Louisa Ma, and Joshua Qian



The histogram for **sqft** shows a slightly right-skewed distribution, meaning most homes are around 2000 sqft, with a small portion of the population are very large homes making the average higher than the median. The distribution of age appears nearly normally distributed, with peak around 10–12 years and fewer homes at the youngest and oldest ends. The price shows a right skewed distribution with fewer higher prices homes. This skew is important to keep in mind when analyzing relationships between price and other features.

Lastly, we printed **frequency tables** for each categorical variable and visualized them with bar plots.

```
Frequency Tables and Descriptive Stats for Qualitative Variables:


--- style ---
Frequency Table:
style
Ranch        44
Cape Cod     39
Two-story    25
Name: count, dtype: int64

Descriptive Stats:
count       108
unique        3
top       Ranch
freq         44
Name: style, dtype: object

--- school ---
Frequency Table:
school
Apple Valley    65
Plum Ridge      43
Name: count, dtype: int64

Descriptive Stats:
count             108
unique              2
top       Apple Valley
freq                65
```

The style column shows that **Ranch** is the most common home style, followed by Cape Cod and then Two-story. This balance helps us later analyze how style affects price. In the school column, we see that 65 homes are in **Apple Valley** and 43 are in Plum Ridge. This matters because homes in Apple Valley were priced higher on average. The frequency table supports that Apple Valley is the more dominant school zone in the dataset, which may partially explain its stronger influence on price.

ISOM 352 Group 18: Pranit Aggarwal, Ivy Chen, Caroline Ma, Louisa Ma, and Joshua Qian

Step 3: Multiple Regression Model

To better understand how different features impact housing prices, we built a **multiple linear regression model**. We converted categorical variables like style and school into dummy variables and defined price as the target variable (y-variable in the regression model).

```
# Drop identifier column if it's not useful for modeling
data_1 = data.drop(columns=['house'])

# Create dummy variables for categorical columns
data_dummies = pd.get_dummies(data_1, drop_first=True)

# Display the transformed dataset
print(data_dummies.head())
```

```
   sqft  beds  baths  heat  garage  basement  age  fire   price  style_Ranch  \
0  1610     3      2     0       1         1   12     1  234280        False
1  2151     3      2     1       2         0   13     1  246360        False
2  1718     4      2     0       2         1   17     0  265650        False
3  1534     3      2     0       2         1   11     1  237420        False
4  1527     3      2     0       1         0    6     1  259170        False

   style_Two-story  school_Plum Ridge
0            False              False
1            False               True
2            False               True
3            False               True
4            False               True
```

```
# Convert all boolean columns to int
for col in data_dummies.columns:
    if data_dummies[col].dtype == bool:
        data_dummies[col] = data_dummies[col].astype(int)
```

```
y = data_dummies['price']
x = data_dummies.drop(columns=['price'])
X = sm.add_constant(x)
model = sm.OLS(y, X).fit()
print(model.summary())
```

ISOM 352 Group 18: Pranit Aggarwal, Ivy Chen, Caroline Ma, Louisa Ma, and Joshua Qian

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.787
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     32.30
Date:                Fri, 28 Mar 2025   Prob (F-statistic):           1.70e-27
Time:                        03:33:53   Log-Likelihood:                -1291.2
No. Observations:                 108   AIC:                             2606.
Df Residuals:                      96   BIC:                             2639.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -1.174e+04   3.36e+04     -0.349      0.728   -7.85e+04     5.5e+04
sqft             114.9015     14.082      8.159      0.000      86.948     142.855
beds            1.388e+04   7207.382      1.925      0.057    -431.516    2.82e+04
baths          -3683.5601   1.03e+04     -0.357      0.722   -2.42e+04    1.68e+04
heat            4909.1095   1.16e+04      0.424      0.672   -1.81e+04    2.79e+04
garage          4.938e+04   1.35e+04      3.645      0.000    2.25e+04    7.63e+04
basement        1.978e+04    1.2e+04      1.647      0.103   -4062.792    4.36e+04
age            -4082.8770    951.725     -4.290      0.000   -5972.037   -2193.717
fire            5198.5376   1.03e+04      0.505      0.615   -1.52e+04    2.56e+04
style_Ranch     1.649e+04   1.14e+04      1.446      0.152   -6151.090    3.91e+04
style_Two-story 4.288e+04   1.27e+04      3.374      0.001    1.77e+04    6.81e+04
school_Plum Ridge -2.301e+04   8709.357     -2.642      0.010   -4.03e+04   -5720.960
==============================================================================
Omnibus:                        4.413   Durbin-Watson:                   1.755
Prob(Omnibus):                  0.110   Jarque-Bera (JB):                3.773
Skew:                           0.392   Prob(JB):                        0.152
Kurtosis:                       3.472   Cond. No.                     1.84e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.84e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Regression Results:

- sqft: each additional sqft adds $114.9 to the home price. The p value is 0.00 which shows that this variable is **significant** (p-value < 0.05).
- baths: each bathroom decreases the home price by $3683.56. However, the p value is 0.722 which means that it is **insignificant** (> 0.05). It is likely correlated with beds or sqft which is causing the unpredictable behavior.
- style: ranch style homes increase the home price by $16,490 over Cape Cod. The p value is 0.152 which means it is not significant. Two-story homes increase the home price by $42,880 over Cape Cod. The p value is 0.001 which shows it is significant.
- basement: having a basement increases the home price by $19,780. The p value is 0.103 meaning it is not quite significant. Again, its correlation with other variables such as sqft may be causing this.
- school: the Plum Ridge school district decreases the home price by $23,010 compared to the Apple Valley school district. The p value is 0.01 which shows it is significant.

As one can see, there are quite a few insignificant variables included in the model above – it can be optimized! From here, we **eliminated non-significant variables using a backwards elimination** process to create the **best version** of our regression model.

```
['style_Ranch', 'style_Two-story']
['school_Plum Ridge']
variable baths is removed
variable heat is removed
variable fire is removed
the dummy group ['style_Ranch', 'style_Two-story'] is kept
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     45.44
Date:                Fri, 28 Mar 2025   Prob (F-statistic):           8.14e-30
Time:                        03:33:55   Log-Likelihood:                -1291.6
No. Observations:                 108   AIC:                             2601.
Df Residuals:                      99   BIC:                             2625.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -1.302e+04   3.24e+04     -0.402      0.689   -7.74e+04    5.13e+04
sqft             115.3223     13.267      8.693      0.000      88.998     141.646
beds            1.302e+04   6607.723      1.971      0.052     -89.016    2.61e+04
garage          4.894e+04   1.31e+04      3.726      0.000    2.29e+04     7.5e+04
basement        1.994e+04   1.19e+04      1.682      0.096   -3586.306    4.35e+04
age            -4074.2865    935.726     -4.354      0.000   -5930.970   -2217.603
style_Ranch     1.496e+04   1.03e+04      1.446      0.151   -5573.227    3.55e+04
style_Two-story 4.259e+04   1.25e+04      3.400      0.001    1.77e+04    6.74e+04
school_Plum Ridge -2.307e+04   8600.484     -2.682      0.009   -4.01e+04   -6003.363
==============================================================================
Omnibus:                        4.297   Durbin-Watson:                   1.748
Prob(Omnibus):                  0.117   Jarque-Bera (JB):                3.650
Skew:                           0.393   Prob(JB):                        0.161
Kurtosis:                       3.438   Cond. No.                     1.80e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.8e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

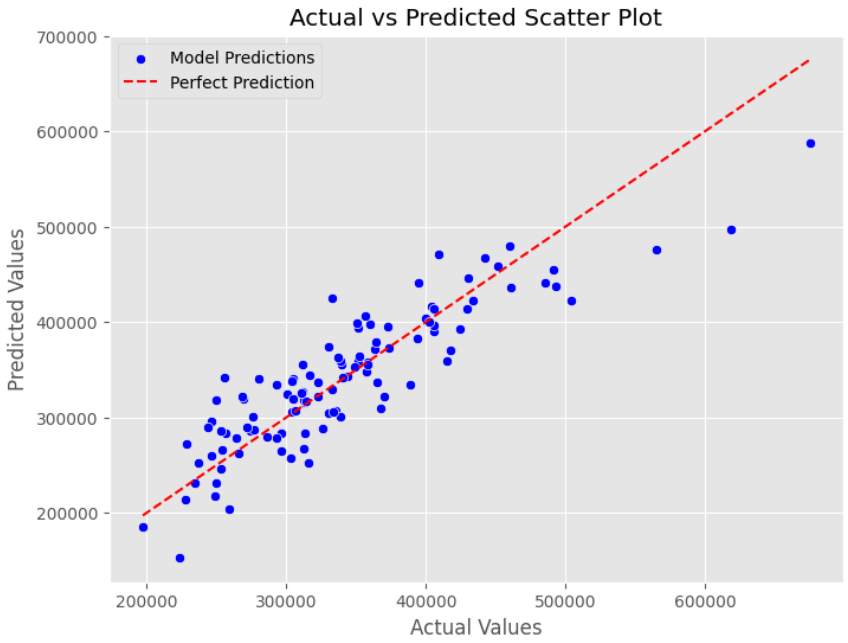
Regression Results:

- After running our backwards elimination process, we can see that variables **baths, heat, and fire have been removed** since their **p-values exceeded** the significance threshold of **0.05**. (Note that the style dummy variables remained because the style dummy group proved to be significant!)
- sqft: In our optimized regression model, each additional sqft now adds $115.32 on average to the home price. The p-value is 0.000, which is less than the significance threshold of 0.05. This makes it **significant.**
- beds: In our optimized regression model, each additional bedroom adds $13,020 on average to the home price. Though its p-value is slightly above the significance threshold of 0.052, it likely remains after backwards elimination due to its **strong multicollinearity** with other variables.

- basement: In our optimized regression model, having a basement increases a home's price by $19,940 on average. Though its p-value of 0.096 exceeds the significance threshold of 0.05, it is kept after backwards elimination likely due to **strong multicollinearity** with other variables.
- age: In our optimized regression model, for every year older a home is, its price decreases by $4074.29. With its p-value of 0.000 being less than the significance threshold of 0.05, it is **significant.**
- style: ranch style homes increase the home price by $14,960 over Cape Cod. The p value is 0.151 which means it is not significant. Two-story homes increase the home price by $42,590 over Cape Cod. The p value is 0.001 which shows it is significant. However, **all the style dummy variables remained** in the optimized model because the style dummy groups altogether proved to be significant (**passed the f-test**)!
- school: In our optimized regression model, the Plum Ridge school district decreases the home price by $23,070 compared to the Apple Valley school district. Since its p-value of 0.009 is less than the significance threshold of 0.05, it is **significant.**

Step 4: Make prediction and comment on the quality of the prediction

We visualized the predicted vs. actual home prices using a scatter plot. Most predictions fall close to the line, suggesting the model captures general pricing trends well, although some variation still exists especially at the higher price range.



ISOM 352 Group 18: Pranit
Aggarwal, Ivy Chen, Caroline Ma,
Louisa Ma, and Joshua Qian

The model explains about **78.6%** of the variation in home prices, based on the **R-squared** value. This indicates a strong model fit, especially given the number of features used.

The Root Mean Squared Error (**RMSE**) is approximately **$39,485**, meaning that, on average, the model's price predictions are off by that amount. While this is a moderate prediction error, it's acceptable considering the range of home prices in the dataset.