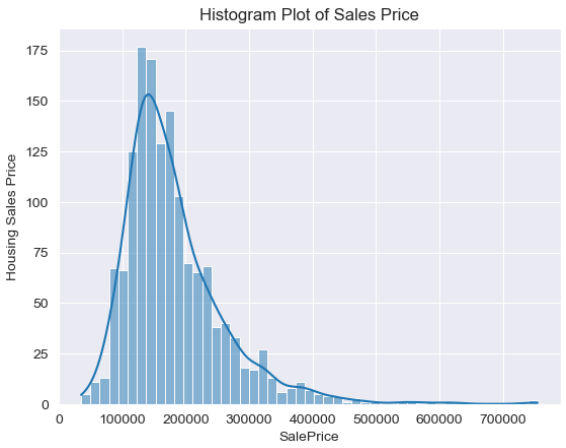
CPSC 4800 – Assignment 3 – Part 3

EDA Report

1. Analysis on the distribution of Sales Price.

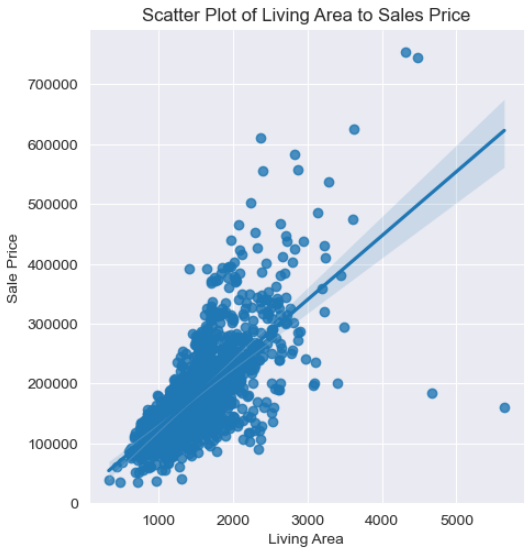


We can see from the histogram on the left that the distribution of housing sale prices looks to be rightly skewed. This basically means that much of the data is located on the left side of the graph i.e., there are more houses with lower prices.

	SalePrice
count	1460.000000
mean	180921.195890
std	79442.502883
min	34900.000000
25%	129975.000000
50%	163000.000000
75%	214000.000000
max	755000.000000

We can see from the statistics above on the right that indeed the mean (180921.19) is greater than the median (163000), so the sale prices are skewed to the right.

2. Determining whether living area is related to the sales price.

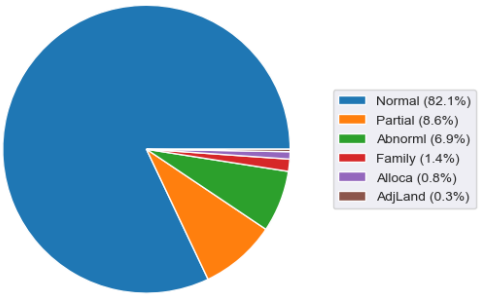


From the scatterplot on the left we can observe that the living area and sales price have a strong positive linear relationship with a few outliers.

```
PearsonRResult(statistic=0.7086244776126513, pvalue=4.518033646784566e-223)
```

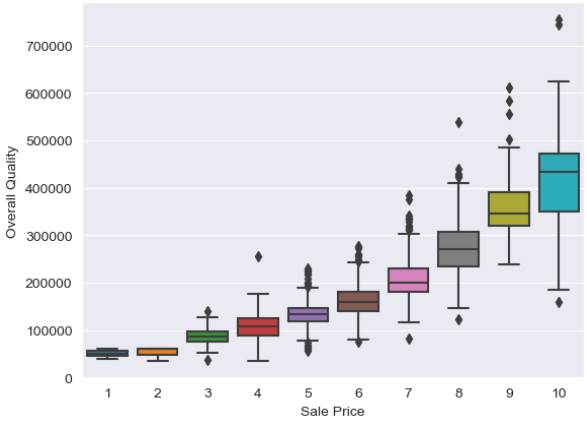
By checking the correlation between the living area and sales price we can verify that there is indeed a strong positive linear relationship as the regression statistic is 0.71 which is close to +1. Therefore, an increase in living area leads to an increase in the sale price.

3. Analyzing the distribution of sale condition of the houses.



As we can see from the pie chart on the left, most of the houses sold are in Normal condition whereas Adjacent Land houses are the least common.

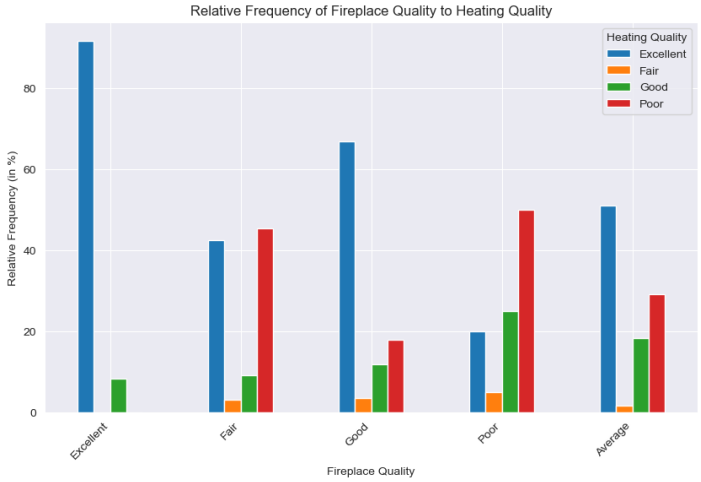
4. Determining if the overall quality of houses is associated with their sales price.



From the side-by-side box plot on the left we can observe that as the quality of the houses increases their median price also increases. This also seems to lead to an increase in the variance of the price as we can see from the difference between the interquartile ranges.

We can conclude that as the house quality increases it leads to an increase in the sales price.

5. Determining if fireplace quality in the houses is associated with their heating quality.



A side-by-side bar graph has been created to determine whether the fireplace quality in the houses is associated with the heating quality. Here, as the group of bars do not look alike, they are not independent from each other.

```
Chi-squared value: 56.30095777816627
Degree of freedom: 12
```

After this a chi-squared analysis has been run to verify our results. Here, we look at the chi-square statistic and compare it to the decision point value corresponding to a degree of freedom of 12 which is 21.03. Since the test statistic (56.30) is more than the DP, we have enough evidence to conclude that the variables (fireplace quality and heating quality) are not significantly independent to each other among all subjects in the population. Based on this analysis we can conclude that fireplace quality of houses is associated with the heating quality.