EDA-AMES HOUSING PRICING

This project involves performing EDA on the housing dataset.

The Ames Housing dataset contains information about home sales in Ames, Iowa between 2006 and 2010.

Documentation

The data is available on Kaggle. Sign up and get access to all the material by clicking the link below.

- (https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

Question 1. Load the Dataset with Pandas

Import pandas with the standard alias `pd` and load the data into a data frame with the standard name `df`.

Data columns (total 80 columns):

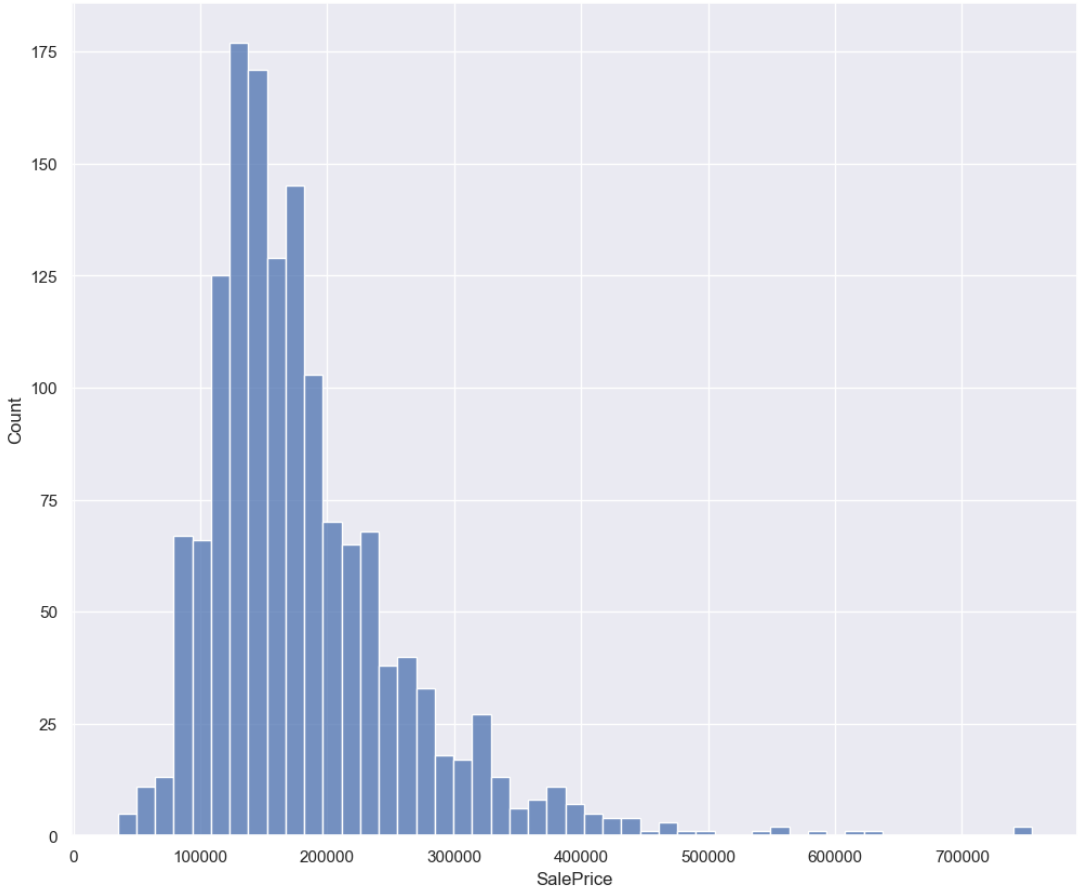
The Ames housing Dataset had 1460 records(observations) and 80 variables.

Although the dataset has some missing values as shown below, the main variables used in this analysis had no missing values:

Variable	No. of Missing values
PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	690

Question 2. Explore Data Distributions

Produce summary statistics, visualizations, and interpretive text describing the distributions of `SalePrice`, `TotRmsAbvGrd`, and `OverallCond`.



Mean: 180921.19589041095

Median: 163000.0

Standard Deviation: 79442.50288288662

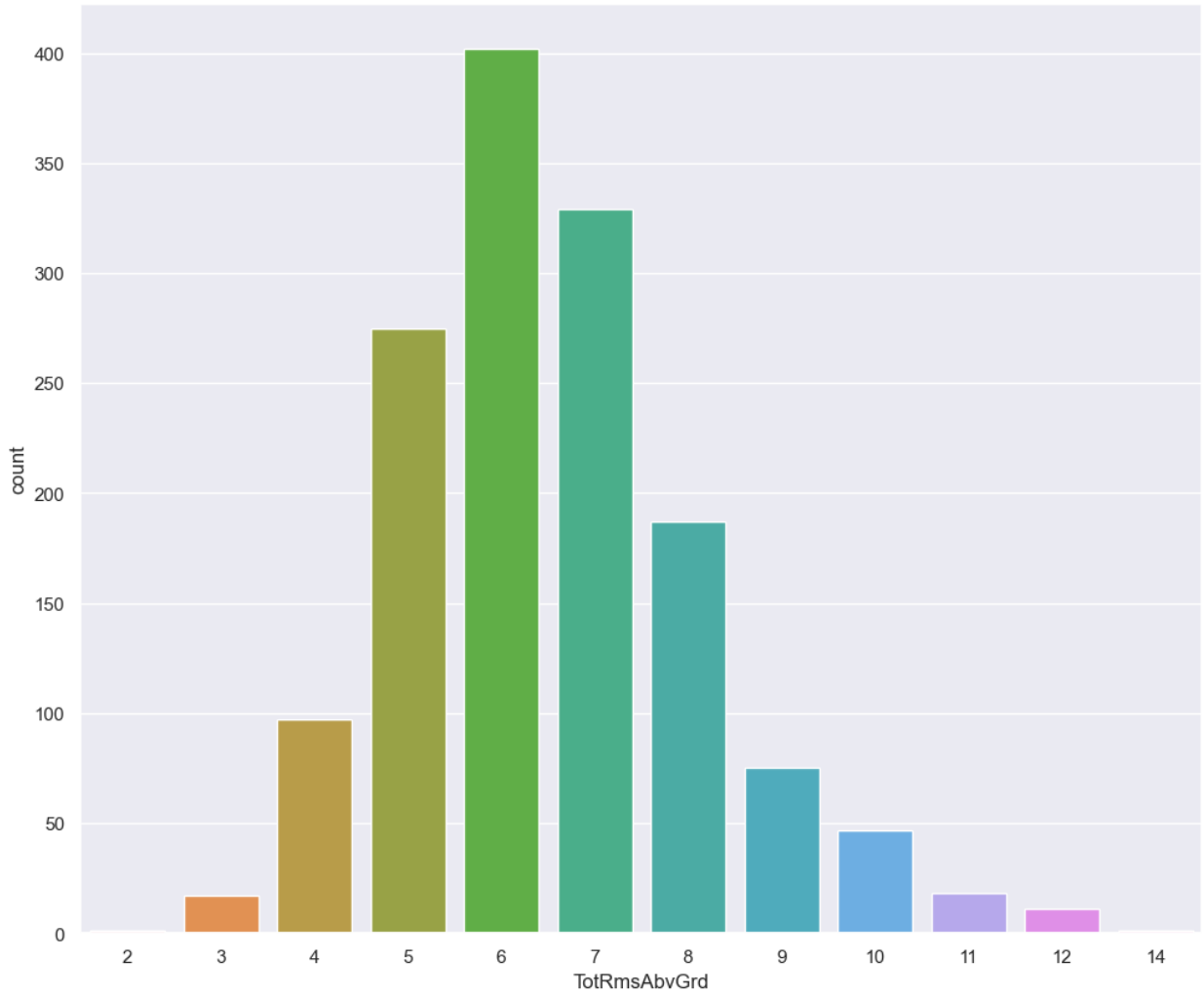
Interpretation of the statistics:

The data distribution shows a symmetrical pattern resembling a bell curve, indicative of a normal distribution.

Most of the houses in the sample are clustered around the median value of $163,000.

Nevertheless, the mean value is higher than the median at over $180,000 due to the influence of high-priced properties.

Bar chart for TotRmsAbvGrd



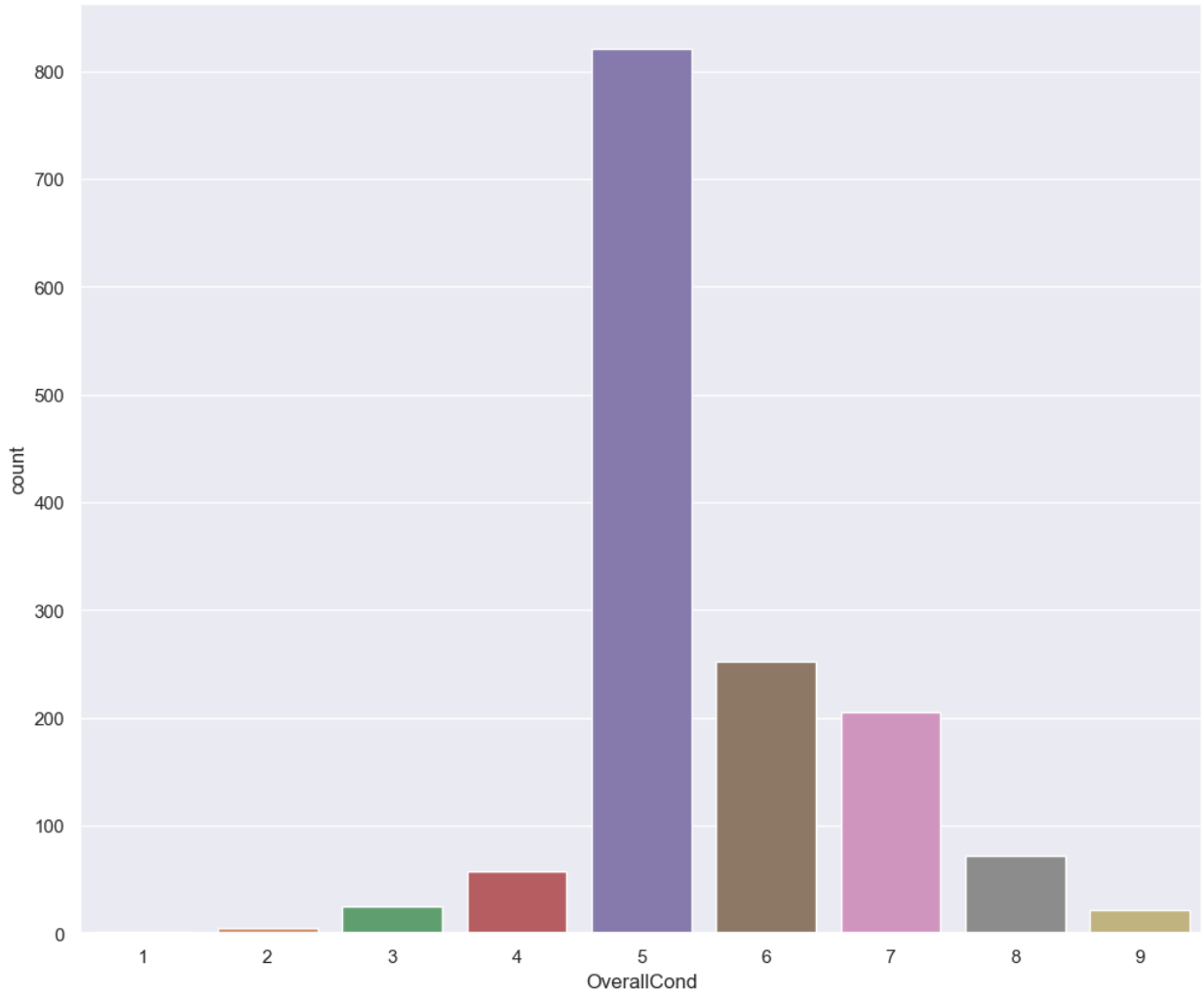
Examining the spread of total rooms above grade

Mean:	6.517808219178082
Median:	6.0
Standard Deviation:	1.6253932905840505

Interpret the above information.

The distribution of the number of rooms in houses is quite similar to a normal distribution, with the average and median both around 6 rooms. Although there are some (12) houses with twice as many rooms as the average, the overall distribution is less skewed than the distribution of sale prices.

Overall Condition



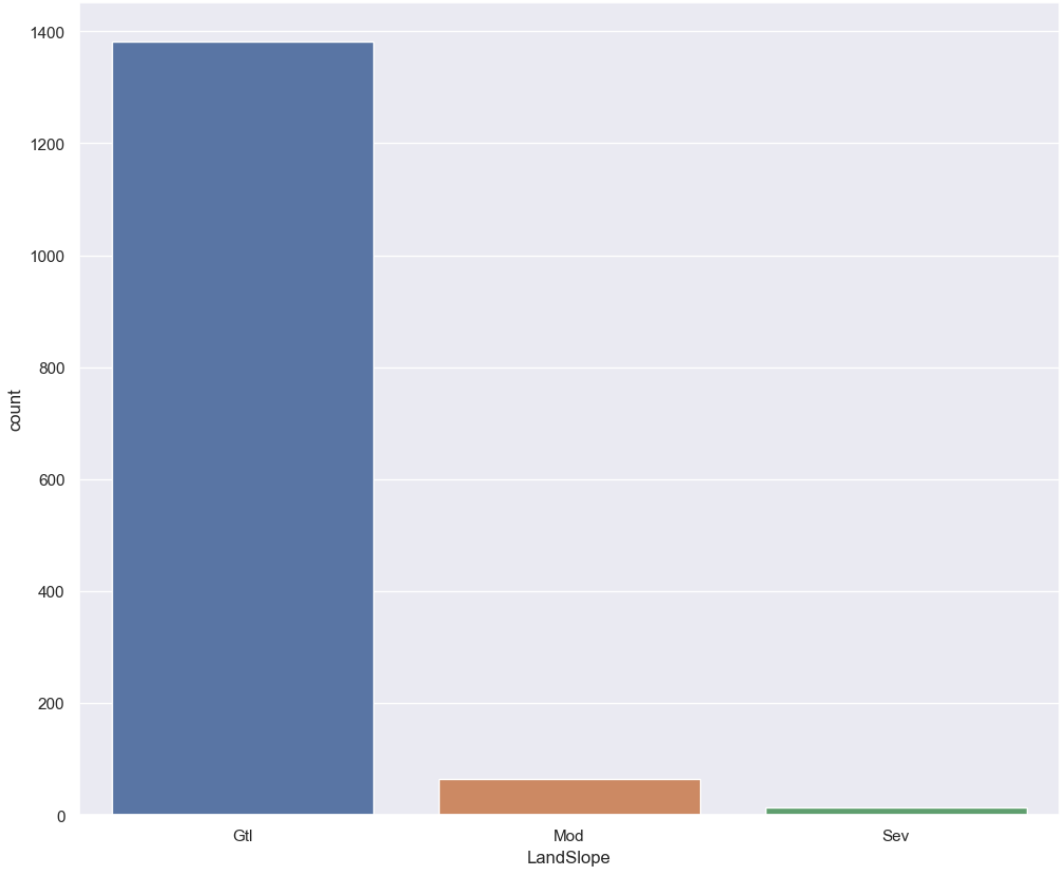
In the produce a histogram above for `OverallCond`.

Mean:	5.575342465753424
Median:	5.0
Standard Deviation:	1.1127993367127316

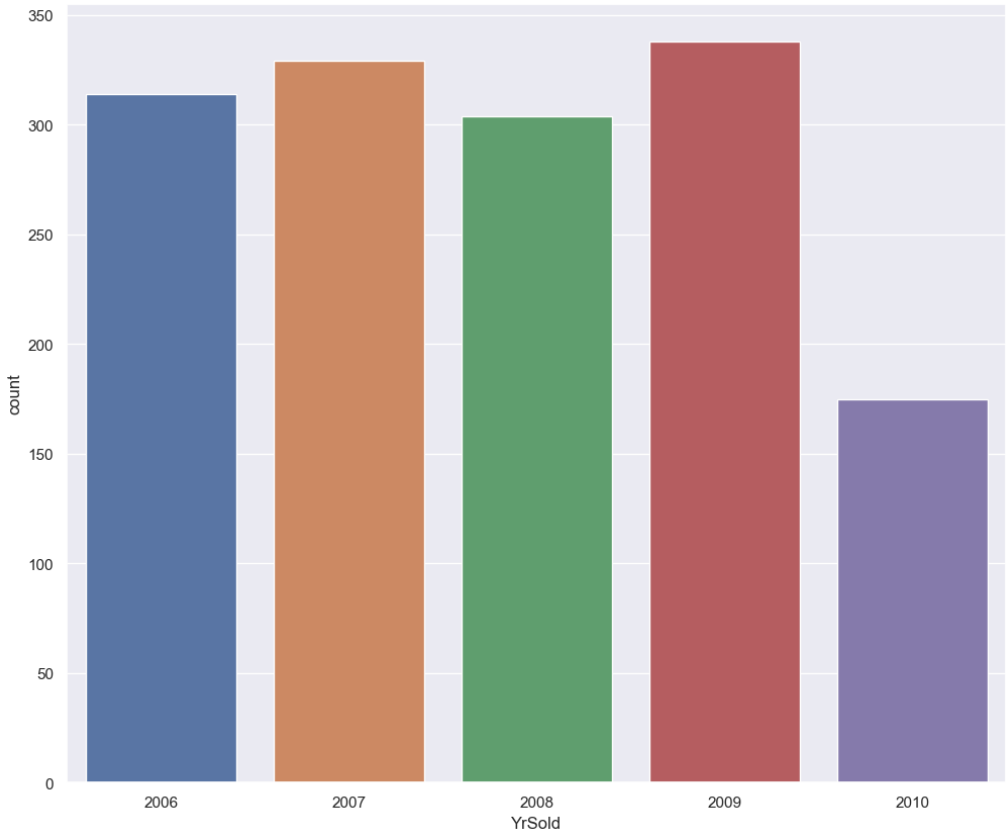
Interpretation of the above information.

Most homes have a condition of 5 and above with majority being of average condition.

Slope-Majority of the houses were built on a gentle slope. A few houses were built on land with moderate slope and very few houses were constructed on land with sever slope.



Year sold- Most houses were sold in 2009,2007, 2006, 2008 and 2010 respectively.

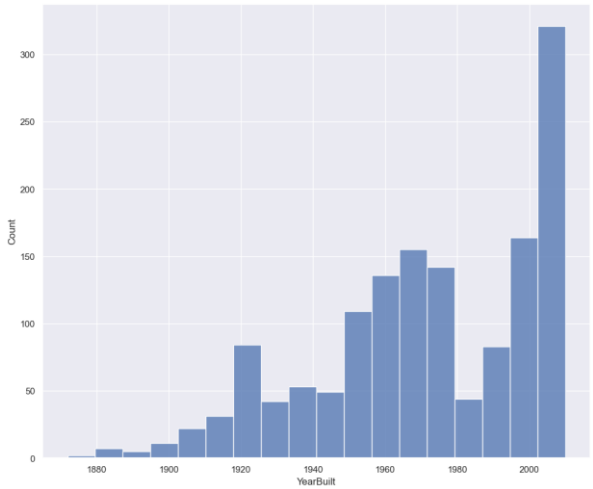


Mean: 2007.8157534246575

Median: 2008.0

Standard Deviation: 1.328095120552104

Year Built- There was a steady increase of new buildings as years progressed. Most buildings were built post year 2000.



Question 3. Explore Differences between Subsets

The overall condition of the house should be a categorical variable.

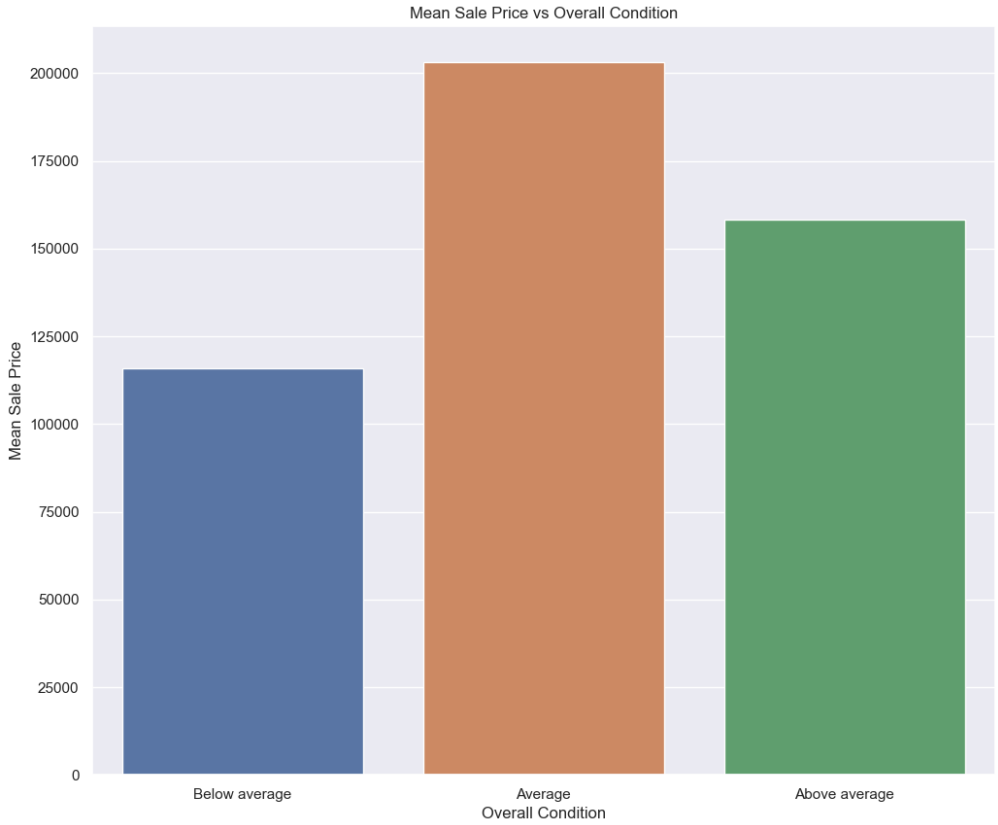
Creation of categories of the full dataset based on that categorical variable, then plotting their distributions based on `SalePrice` as the variable.

The categories created were:

* below_average_condition: home sales where the overall condition was less than 5

* average_condition: home sales where the overall condition was exactly 5

* above_average_condition: home sales where the overall condition was greater than 5

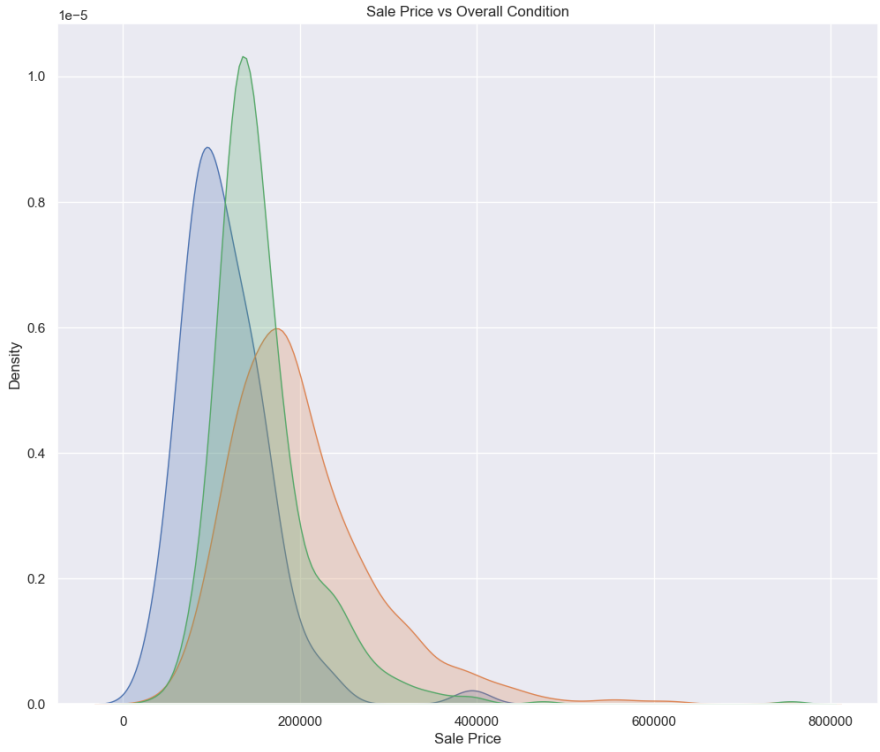


Firstly, it's worth noting that most of the houses in the dataset are considered to be in average condition, with around one-third of them being in above-average condition and less than 20% being in below-average condition.

This suggests that the average condition group covers a wider range of sale prices compared to the other two groups.

As expected, houses in below-average condition have lower sale prices compared to the other two groups.

However, it's surprising that houses in above-average condition don't have significantly higher average sale prices than those in the average condition group.



Interestingly, above-average condition houses appear to be concentrated in the $100,000 to $200,000 price range, while houses in the average condition group tend to sell more frequently for prices above $200,000.

Further investigation is necessary to better understand the characteristics of above-average condition houses, as this finding challenges the common belief that better condition always leads to higher prices."

Generally, the plot clearly shows that most houses are below average and average in terms of their overall condition with 1 or 2 kitchens above grade.

Houses that have two kitchens were sold for less than $200,000, whereas some homes with only one kitchen were sold for considerably more

Sale Price increases with an increase in ranking on the basis of the overall condition of the house.

Examining the spread of total rooms above grade

The distribution of the number of rooms in houses is quite similar to a normal distribution, with the average and median both around 6 rooms.

Although there are some houses with twice as many rooms as the average, the overall distribution is less skewed than the distribution of sale prices.

Question 4. Explore Correlations

To understand more about the variables of these homes leading to higher sale prices, correlation knowledge was key.

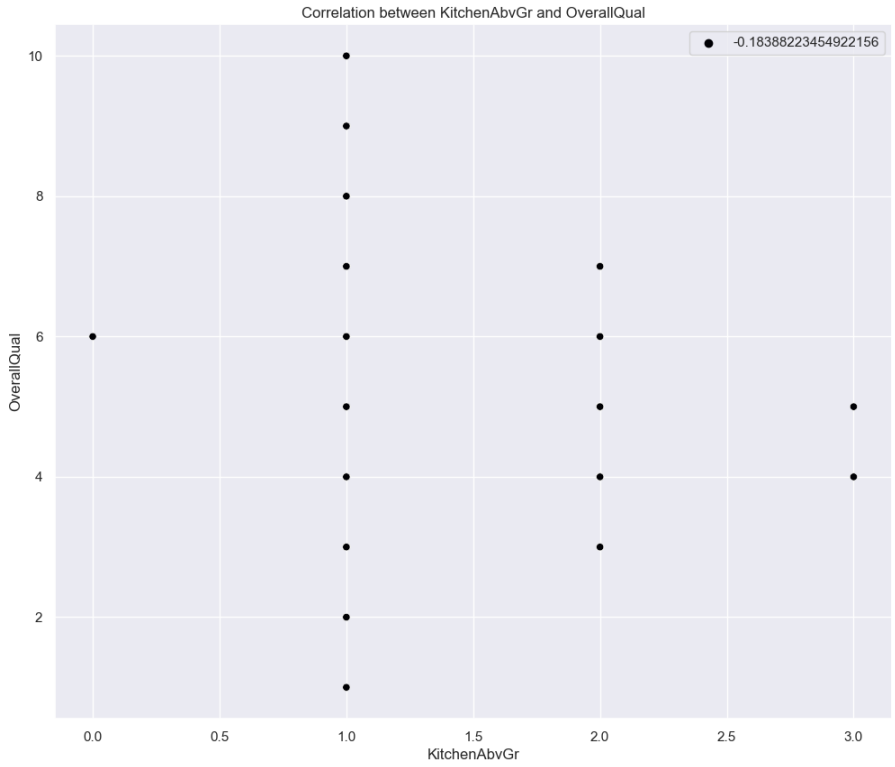
Check the correlations with numeric data types.

Most positively correlated column: OverallQual

Maximum correlation value: 0.7909816005838047

Most negatively correlated column: KitchenAbvGr

Minimum correlation value: -0.1359073708421411



<u>Interpretation of the results is as below with reference from the data dictionary.</u>

The column with the highest correlation is overall quality.

According to the data description/data dictionary:

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

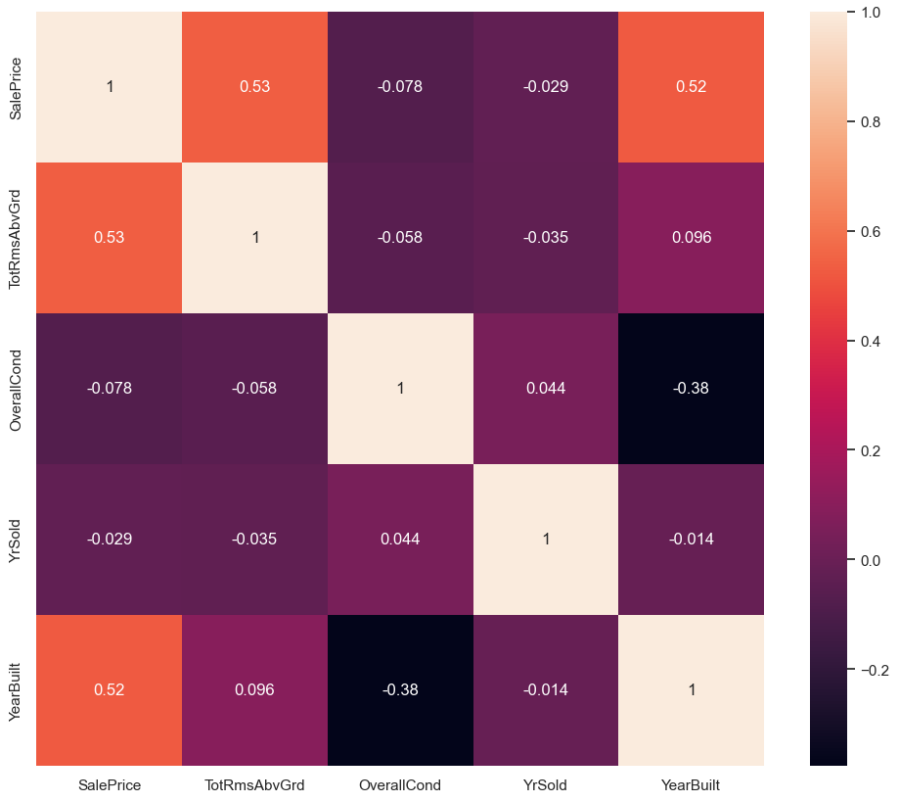
4 Below Average

3 Fair

2 Poor

1 Very Poor

	SalePrice	TotRmsAbvGrd	OverallCond	YrSold	YearBuilt
SalePrice	1	0.533723	-0.07786	-0.02892	0.522897
TotRmsAbvGrd	0.533723	1	-0.05758	-0.03452	0.095589
OverallCond	-0.07786	-0.05758	1	0.04395	-0.37598
YrSold	-0.02892	-0.03452	0.04395	1	-0.01362
YearBuilt	0.522897	0.095589	-0.37598	-0.01362	1



Interpretation

Sale Price has a correlation coefficient of 0.53 with total rooms above grade indicates a moderate positive correlation between two variables. This means that as one variable increases, the other variable tends to increase as well, but the relationship is not particularly strong.

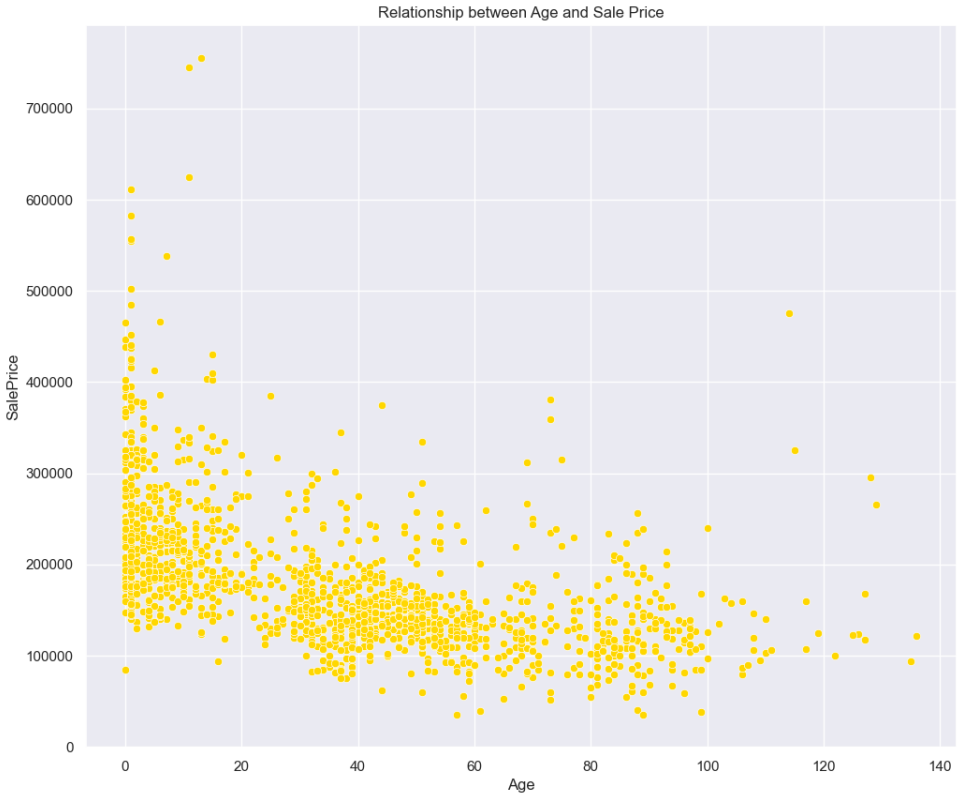
The same applies to correlation between Sale Price and Year the house was built which has moderate positive correlation of strength 0.52.

A correlation coefficient of -0.029(Sale Price vs Year house was sold) indicates a very weak negative correlation between two variables. This means that as one variable increases, the other variable tends to decrease slightly, but the relationship is not significant

Question 5. Engineer and Explore Age

creating a new variable and naming it Age

df["Age"] = df["YrSold"] - df["YearBuilt"]



Interpretation of the plot

In general, newer homes tend to command higher prices, with their values increasing as they age.

However, it's important to note that the variability in sale prices appears to rise after homes reach the age of 100 years.

This is because some houses may have higher-than-average sale prices, but there are relatively few home sales in general.

Moreover, there may have been periods of rapid expansion and decline in the housing market in recent decades.

Evidence of this is apparent in the relatively low number of homes sold that are around 20 years old compared to those that are slightly above 20 years old but less than approximately 25 years old.

More exploration of this pattern may reveal valuable insights which need to be explored.

Acknowledgments & References

- [Documentation](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html)
- [plottinghistograms](https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.hist.html)
- [customizing axes](https://matplotlib.org/stable/api/axes_api.html#axis-labels-title-and-legend)
- [plottingverticallines](https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.axvline.html#matplotlib.axes.Axes.axvline)
- Antony Muiko-Data Science tutor.