Jose Montoro - Data Visualization - Capstone Project

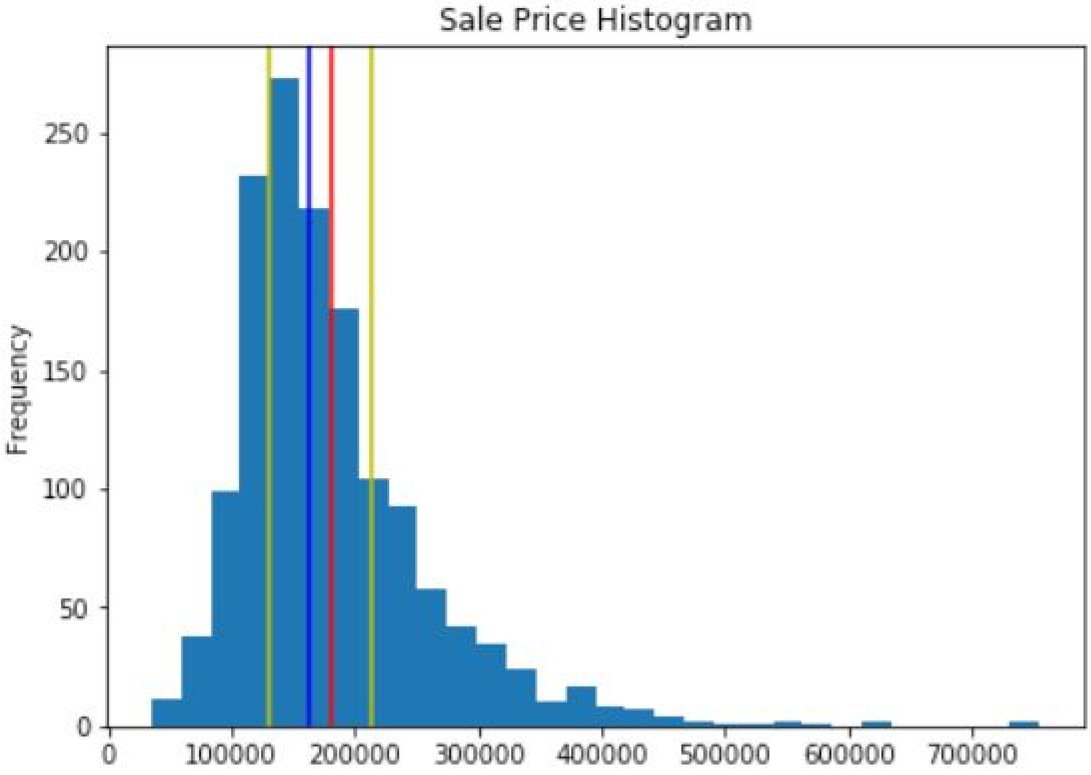
After the data is clean, it's time to visualize different variables to see how they are related.

Predicted variable

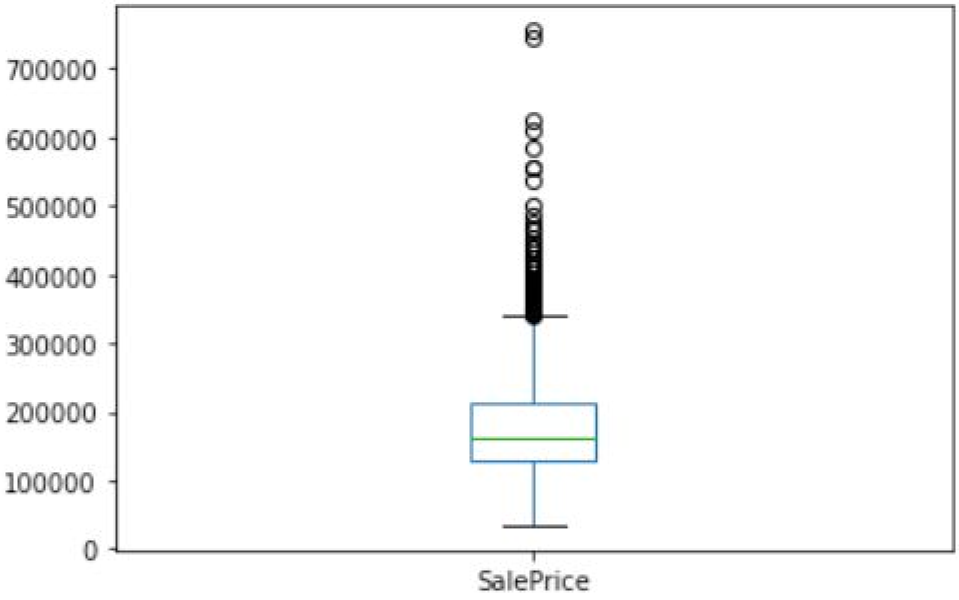
As part of the training data, I was curious to know more about the SalePrice variable. We have 1460 observations in the training data. Here's the distribution of the SalePrice.

The cheapest house costs $34,900 and the most expensive one costs $755,000. The mean value is $180921 and the median $163000, relatively close. Here's a histogram with those values plotted (mean in red, median in blue, 25% and 75% quantiles in yellow).

You can see that most houses' price is between $100,000 and $250,000.



Here's a boxplot, too. We can see that there are two very clear outliers. We'll probably discard those so they don't interfere with our linear regression model.



Feature selection

This dataset has 81 variables. They are too many to keep track of. For that reason, one goal of the visualization part of the project will be to reduce the number of variables to the most significative ones.

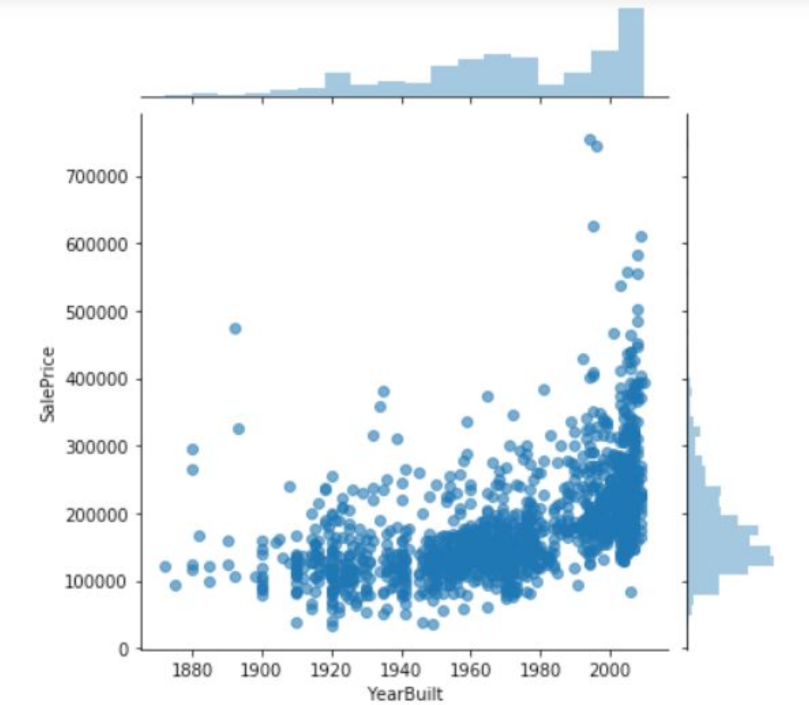
Since we're trying to predict Sale price, the ones with high correlation are good candidates.

Numerical Variables Scatterplots

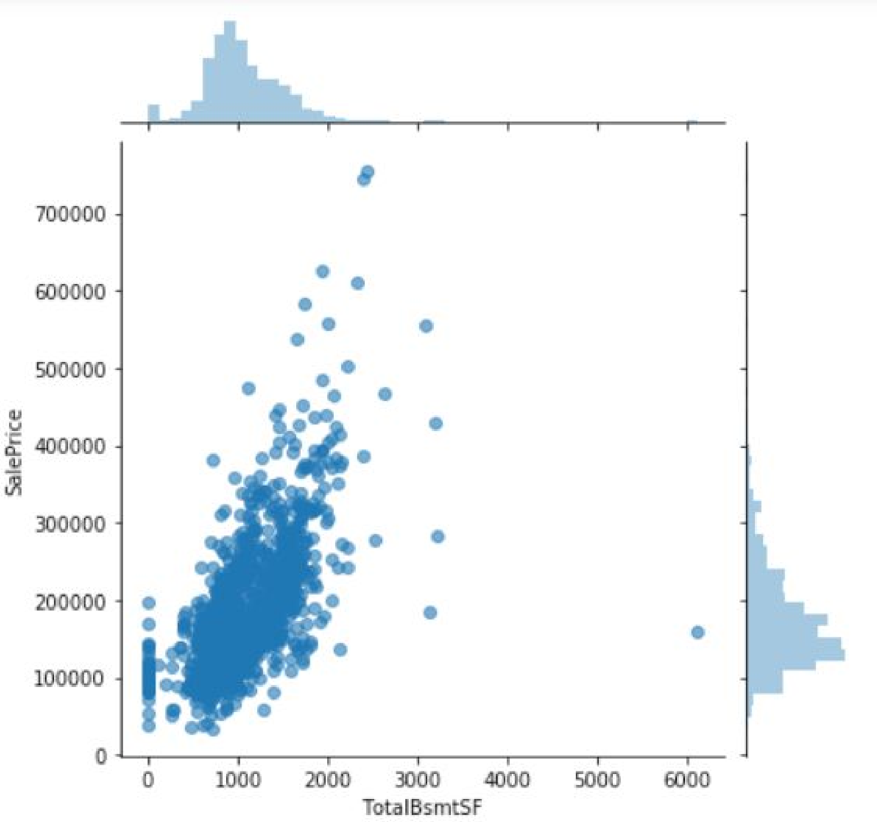
I plotted scatterplots of the numerical variables vs the SalePrice variable. Here are some examples.

These features present a strong correlation with SalePrice, therefore are good candidates to keep.

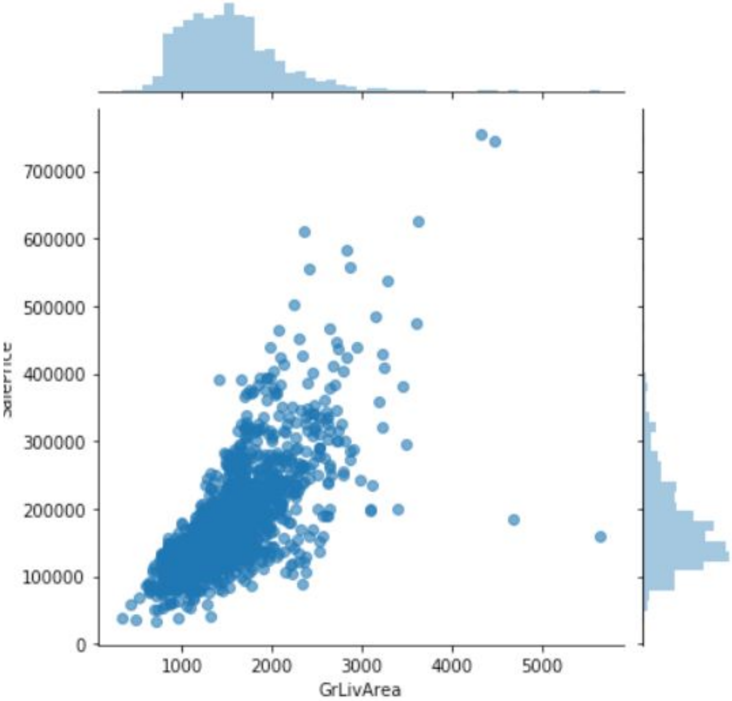
- Year Built. We can also see in the histogram that a lot of the houses are new, built in the 2000s.



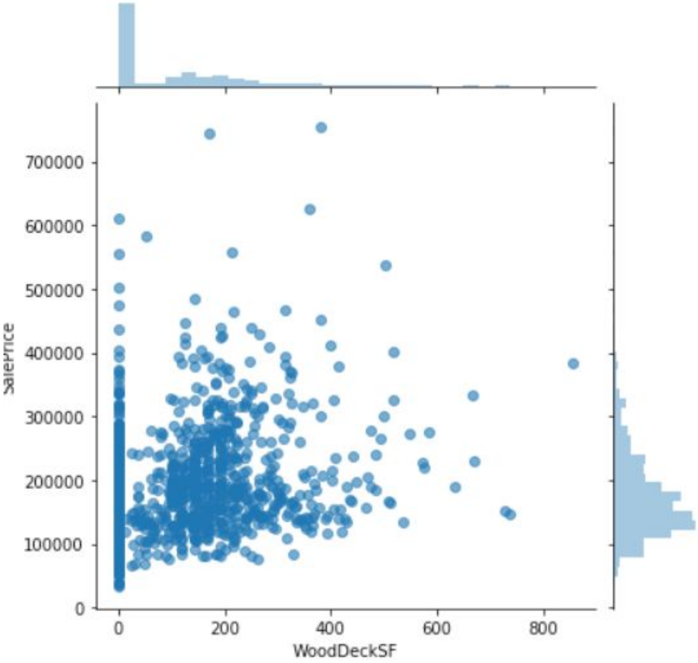
- Total Basement Square Footage: How big the basement is. Again, we have a huge outlier here with +6K sqf. We'll have to investigate.



- General Living Area: sqf of the house itself. This seems like a great predictor. Some outliers there, again.



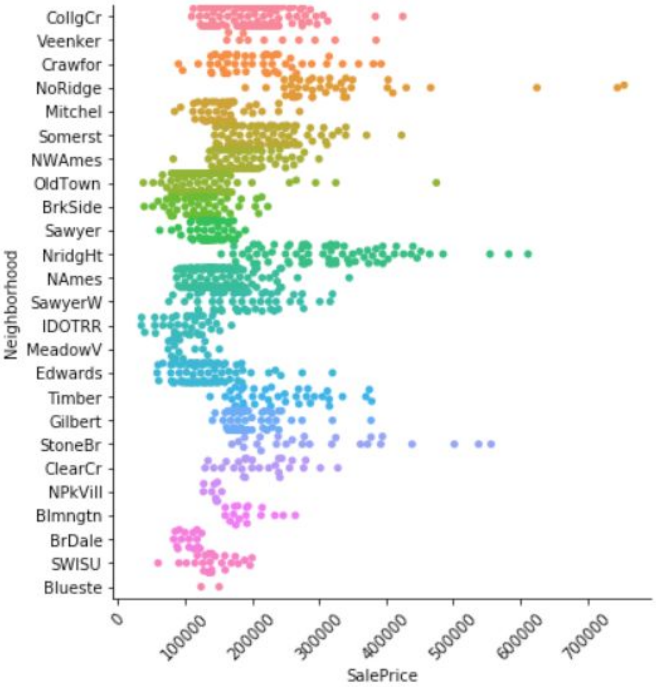
- Wood Deck sqf: Not all houses have a wood deck. A lot of values here are 0. But if they do, it seems related to the house price.



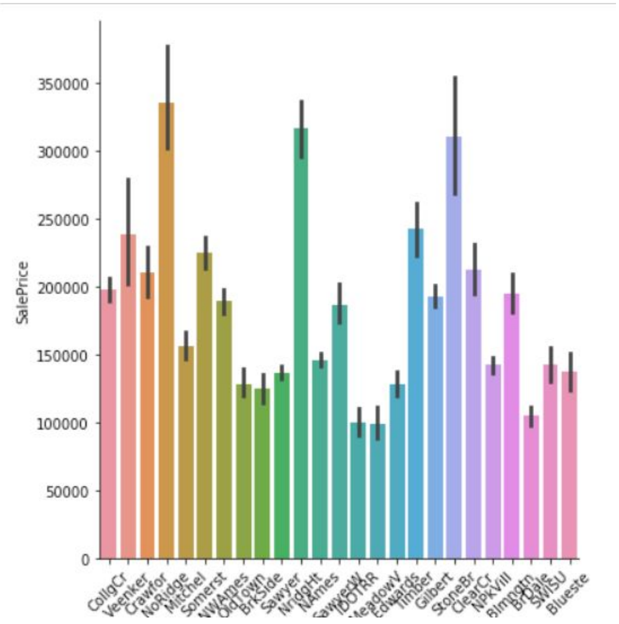
Categorical Variables

These are plots that show the correlation between SalePrice and a categorical variable. They also show, to some extent, how many datapoints are in each variable.

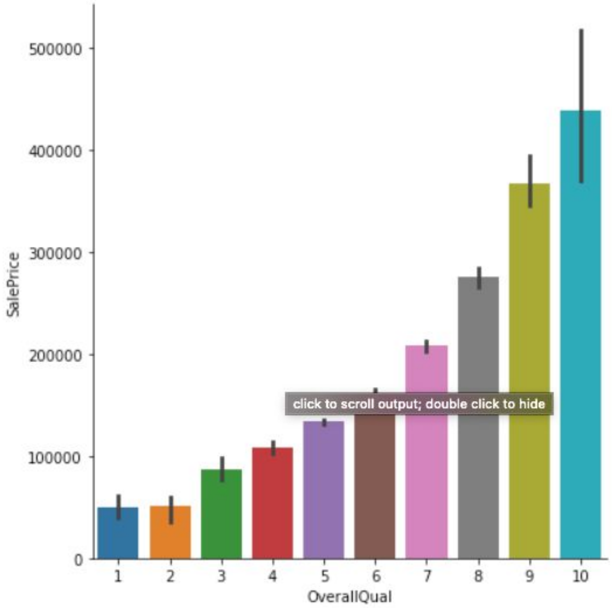
- Neighborhood: Seems to be a good predictor, some of them have higher house prices than the others. Also, some have many more houses, too.



Here’s a barplot of the same data, showing the means and the CIs.



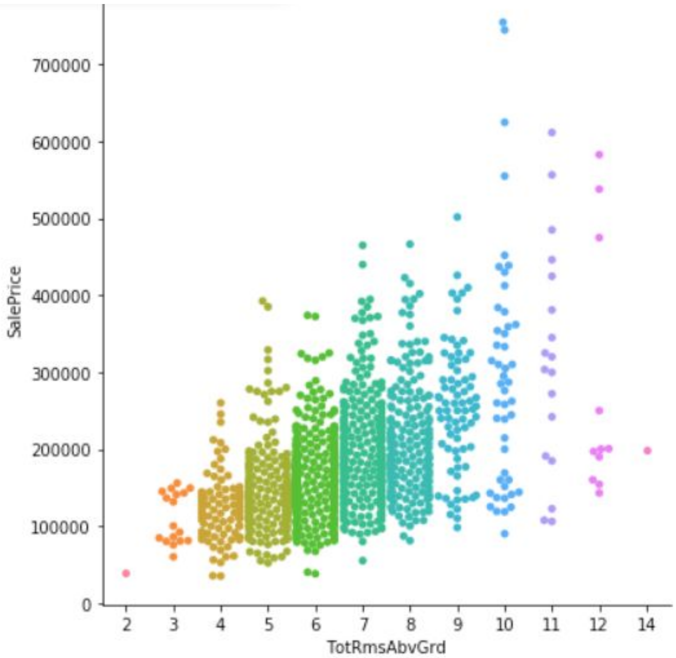
- Overall quality: Seems to be highly related to the SalePrice. Here it is:



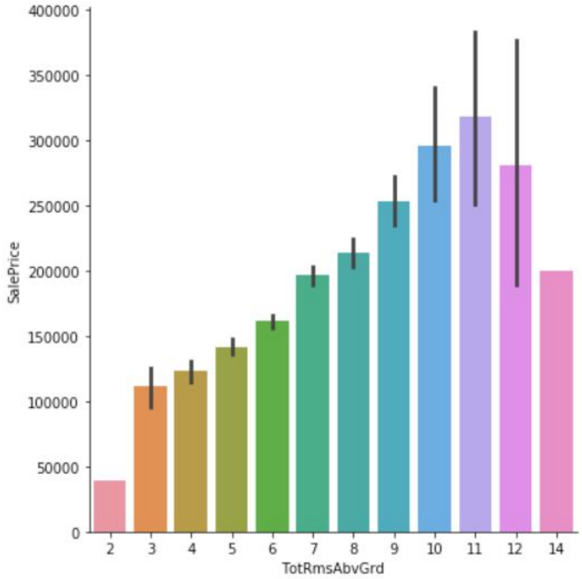
However, Overall Condition doesn't seem to be a very good prescriptor, therefore I'm not showing it here.

- Number of rooms: More rooms mean a higher house price, as expected.

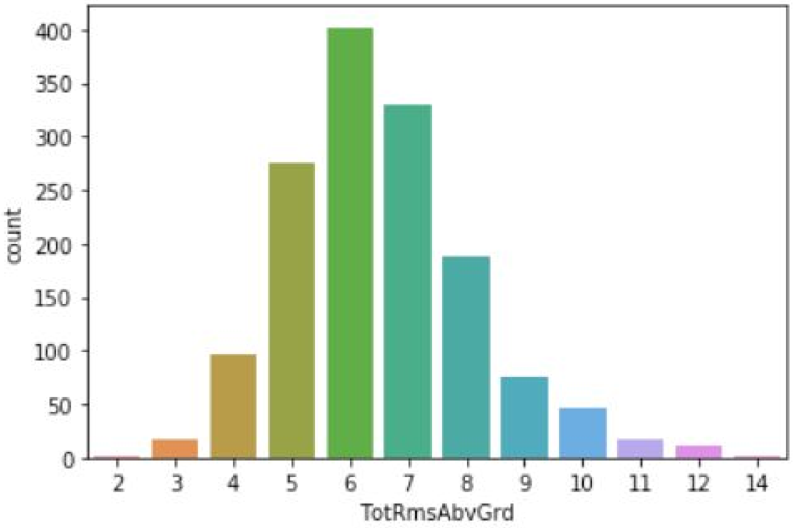
Here's a swarm plot.



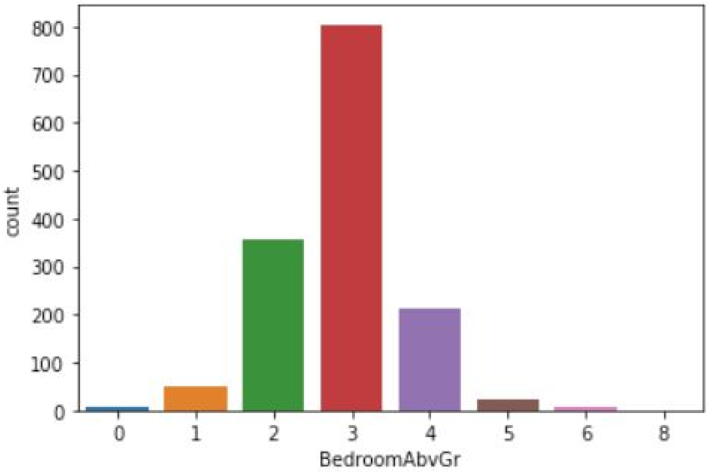
And below is a barplot with the mean of each group.



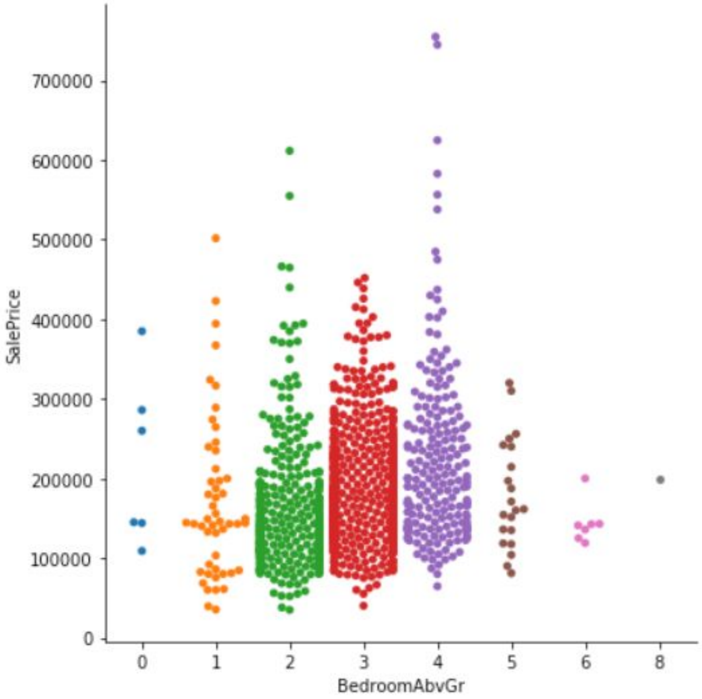
I’m curious to see the distribution of the variable itself. Below is a count plot of TotRmsAbvGrd. It looks like a normal distribution, with most values in the “middle ground”: between 5 and 8 rooms.



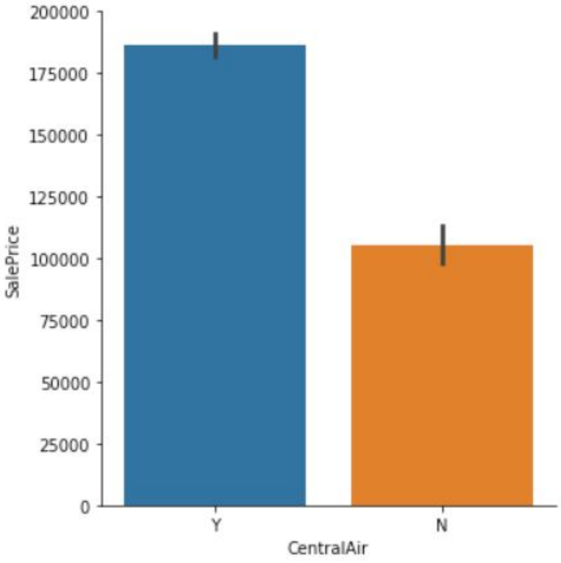
Note that this variable refers to all rooms, including kitchen, bathrooms, etc (not just bedrooms). If we only include bedrooms, this is the distribution:



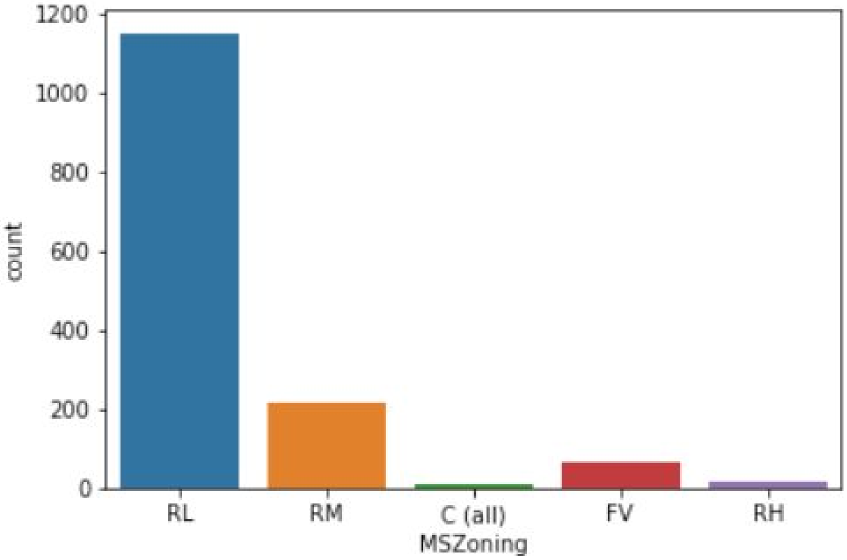
And the correlation with price. Surprisingly, doesn't seem as strong as all rooms included.



- Central air: having A/C seems to increase the value of the property. Here's a bar chart with the mean price for each group.



- Zoning: the last feature I wanted to explore is the Zoning. Looks like most houses are residential (RL, RM, RH) and only some are Commercial (C) or boat houses (FV), floating vehicles. It's worth considering if we should drop those 2 groups and keep only the residential properties.



Conclusion

After the data exploration and visualization, I have a better understanding of the distribution of the variable I'm trying to predict. I'm also more familiar with the different predicting variables, and I can decide which ones to keep for my first baseline model that appear to be the most informative.

There were some surprises, like LotArea or number of bedrooms not having a strong correlation with price. Other variables were related as expected, like total number of rooms, or livable square feet.