Statistical Data Analysis Report

Overview

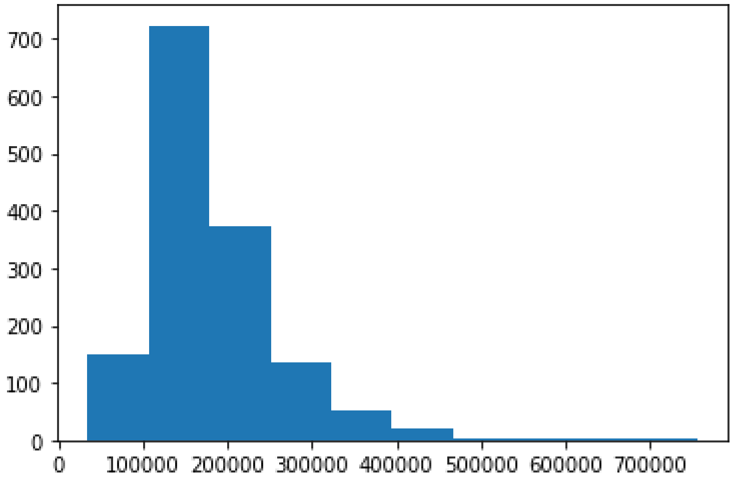
This section describes investigating the data , finding trends and relationships between features, stating and testing a hypothesis applied to the house-prices-advanced-regression-techniques” data.

Summary Files

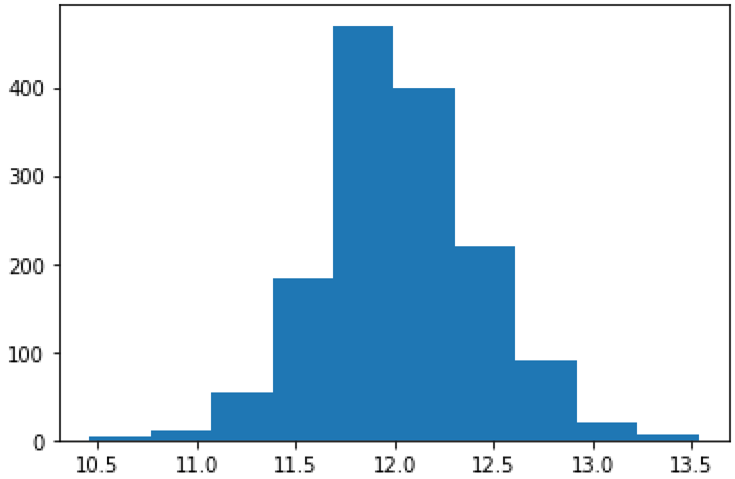
This dataset is split into two separate files:train and test sets. I took a train .csv file since I have a SalePrice target variable in there to investigate dependencies and this file is basically used to train a ML model.
Original version of the dataset
https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Exploratory data analysis

For EDA I started to explore how my SalePrice distribution looks visually and testing it for normality.



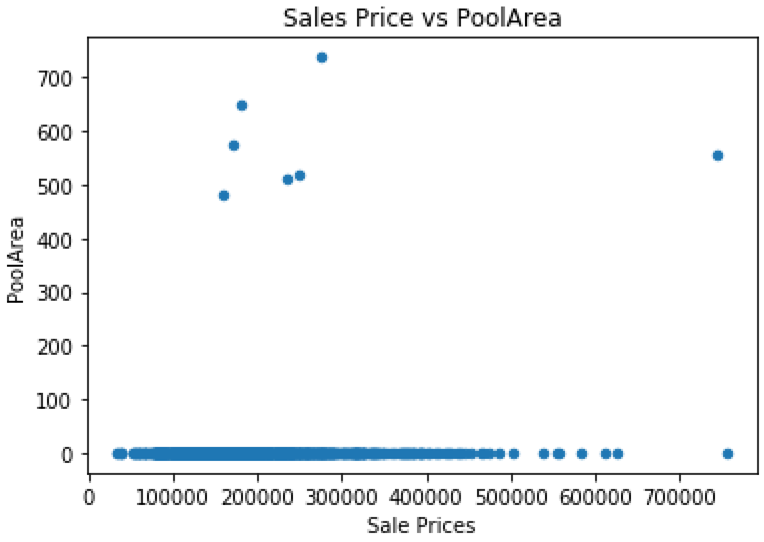
Also, I changed this distribution using a log-transformation to make it as close to normal as possible. This step is reserved for training a ML model for later.



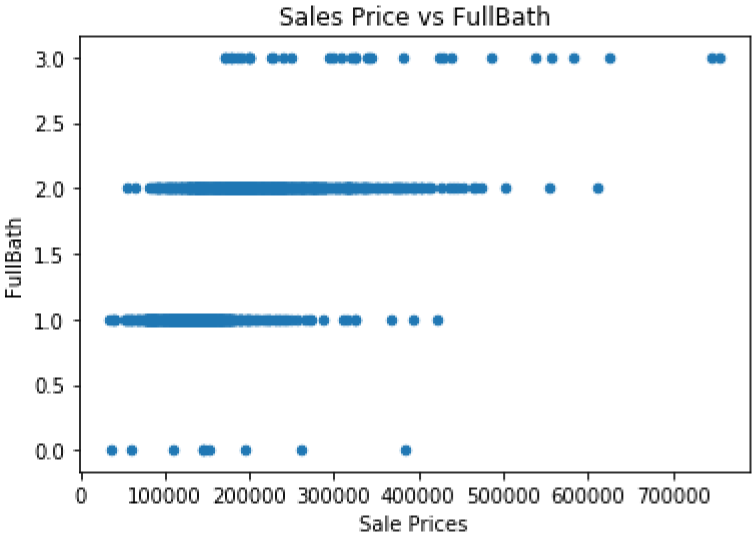
Original plot tells us trends of SalePrices over the number of houses. So data tend to radically positively grow in numbers and to decay as prices go higher. To fit our model , not to overfit it, would be good to normalize our data. For this purpose we do a log-transformation. Now our data looks almost a normal distribution and it gives us a good distribution of our data to fit a trained model.

Next, I compared two related quantities or more. I visualized it to have an idea of how correlated the data and what else interesting or facts to notice. I made several scatterplots and it is a really helpful tool that tells a lot about the data.

Only a few houses have Pool Area but almost all of them do not have so I should drop and remove this column to avoid a noise creation in my dataset.



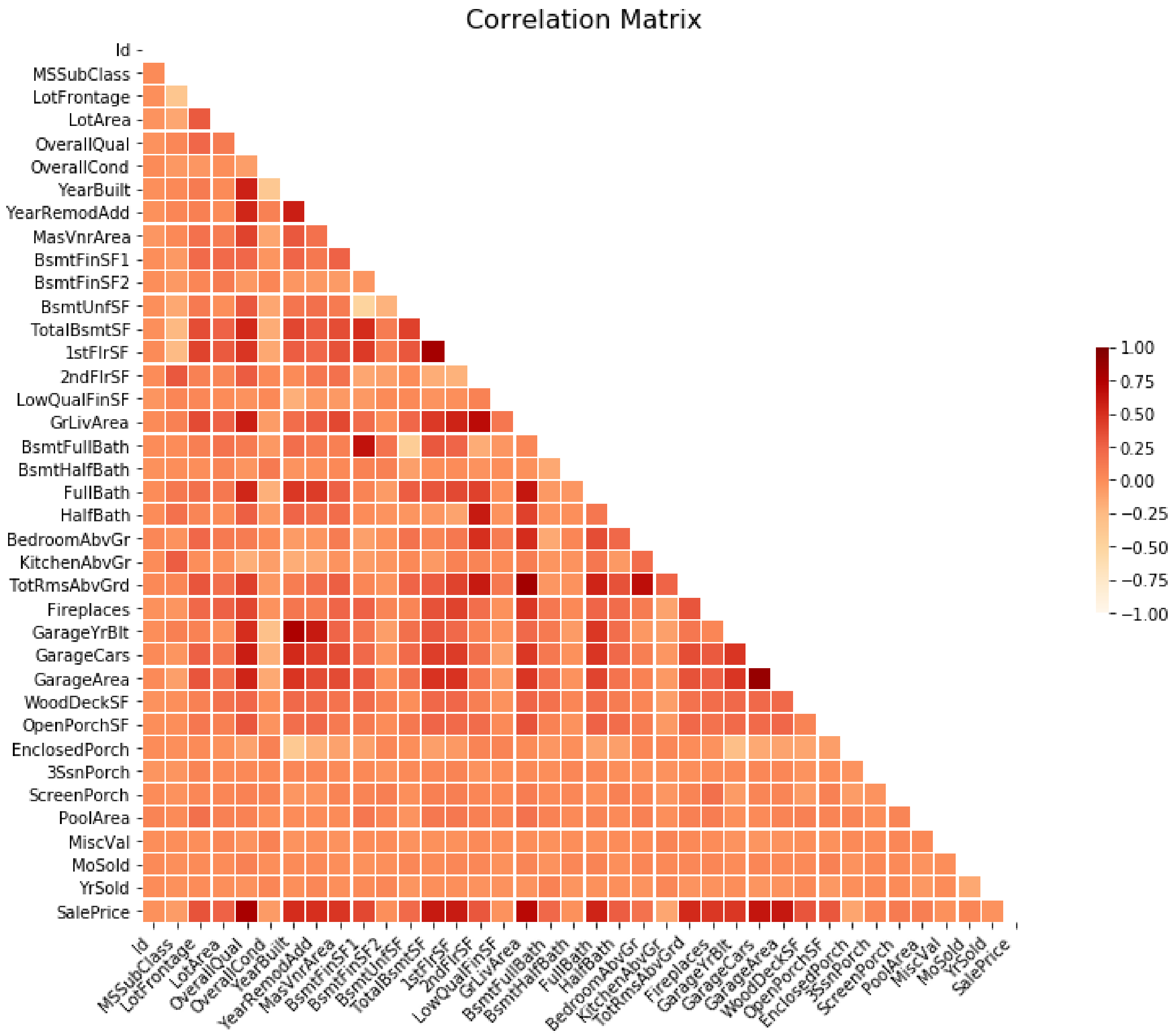
Obviously, Houses that have 3 full baths are more expensive.



But that is not all , there is a lot more to research.

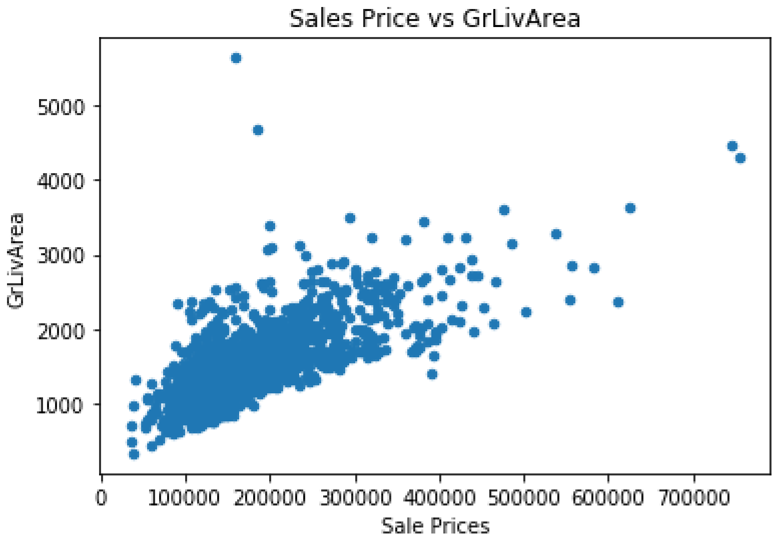
Moreover, in addition to scatterplots, I made a correlation matrix, a powerful exploratory tool, that tells me so much about correlation of features in the dataset. I conveyed what SalePrice is correlated to visually using a heatmap.

Let's look at the correlation matrix and what it will convey.

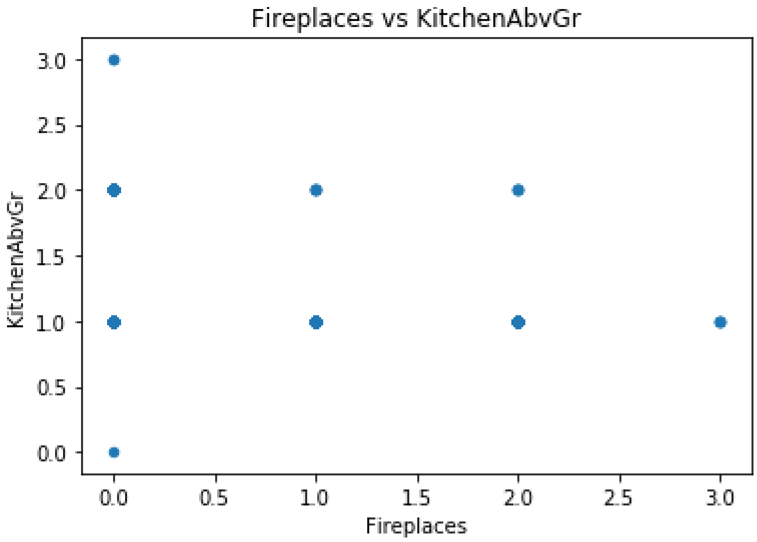


SalePrice (target) is strongly correlated to OverallQual, YearBuilt,YearRemodAdd,MasVnrArea,BsmntFinSF1,TotalBSmntSF, 1stFlrSF, GrLivArea, FullBath,GarageCars,GarageArea.

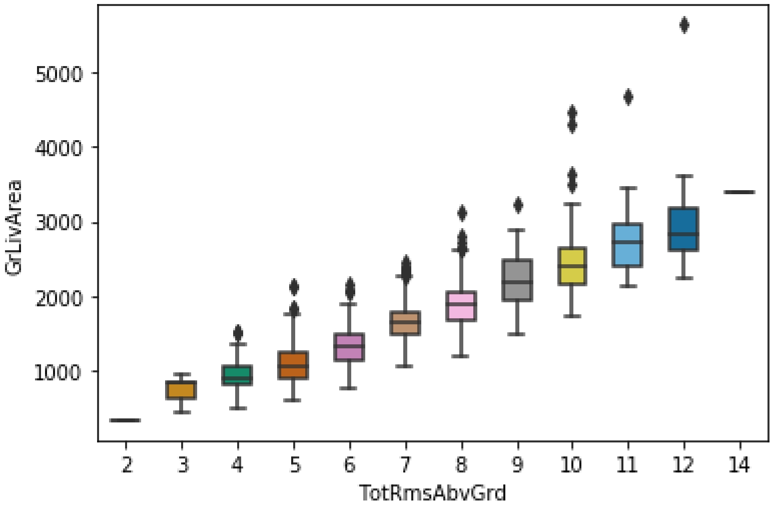
This is how plots look if there is a strong correlation.



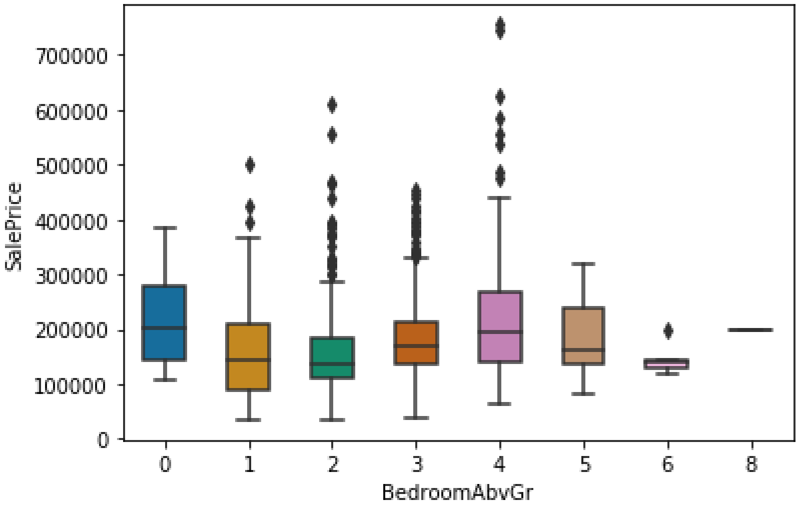
According to the Correlation Matrix I have data that are not correlated and if to visualize them plots will tell that.



According to the Correlation Matrix I have data that are strongly correlated between ech other and if to visualize them plots will tell that.



But I want to investigate further if SalePrice means 2 and 3 BedroomAbvGr have a statistical significance. In other words, I want to investigate if there is no difference in their SalePrices.



Testing a hypothesis

Finally, I have a hypothesis that houses and their SalePrice means for those that have 2 and 3 bedrooms above grade only do not have a difference.

To test my hypothesis I need to perform a ttest of 2 groups. To perform the ttest I have to compute standard deviations of tested groups.

I computed a standard deviation and mean for houses that have 2 Bedrooms only.

I computed a standard deviation and mean for houses that have 3 Bedrooms only.

So knowing that information I performed a t-test, trying to find evidence of a significant difference between population means (2-sample t). Put another way, T is simply the calculated difference represented in units of standard error. The greater the magnitude of T, the greater the evidence against the null hypothesis. This means there is greater evidence that there is a significant difference. The closer T is to 0, the more likely there isn't a significant difference.

My ttest computed that 1.5299459699837254e-06 is smaller than .05, so we're going to reject the null hypothesis which asserts there is no difference between our sample mean and the population mean. For this two-tail test, we reject the Null and we conclude that there is a statistically significant difference .