SPRINGBOARD - DATA SCIENCE CAREER TRACK

INFERENTIAL STATISTICS

HOUSING PRICES- ADVANCED REGRESSION TECHNIQUES

INTRODUCTION

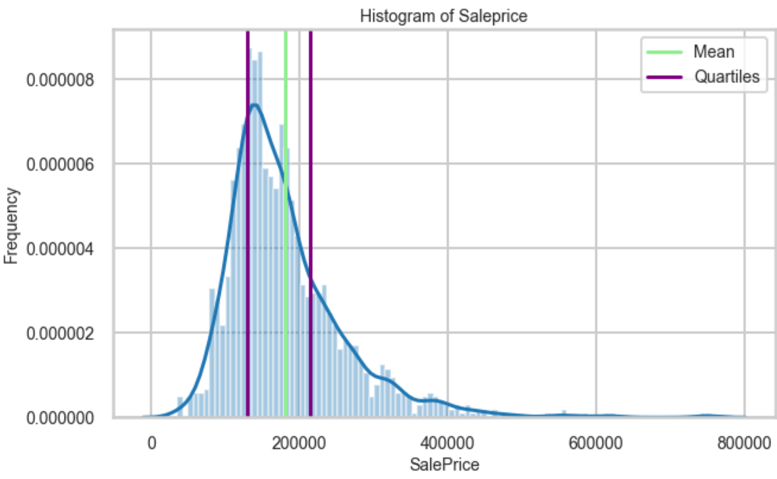
In this project we shall explore the Target variable i.e Sale Price and its relation with some variables. We will analyse a few variables by applying inferential statistics. We will also conduct a Hypothesis Test to validate our assumptions..

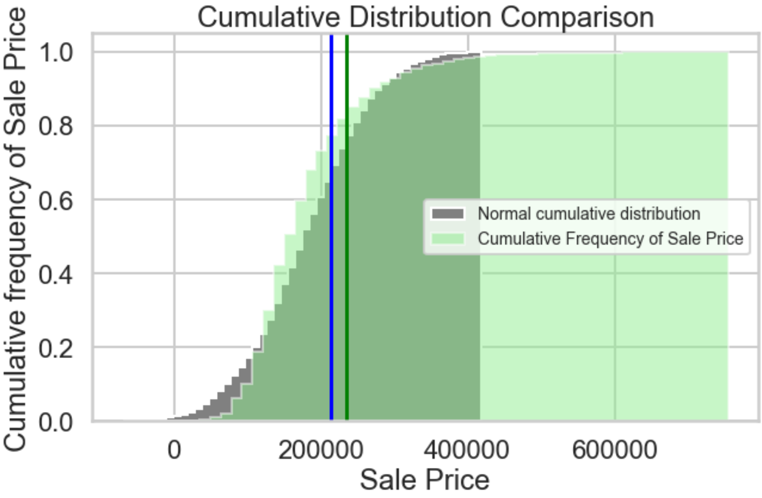
TARGET VARIABLE - Sale Price

We will look at the following:

1. Distribution plot of SalePrice.

2. Develop a cumulative distribution of Sale Price alongside cumulative normal distribution.





Computations:

Mean is: 180921.2 and Standard deviation is: 79442.5

Median of SalePrice is: 163000.0 and Interquartile Range is: -84025.0

Kurtosis of Sale Price is: 6.54 and skewness is: 1.88

Observations:

1. The distribution is not normal. Mean and median, both have lower probabilities of occurrence than the mode.
2. Distribution of SalePrice is leptokurtic. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian
3. The distribution is right skewed. The range of the upper 25% of data is around 60000 which is 3 times more that the Interquartile range.
4. Mean does not seem to be a good representation as there are quite a number of outliers.

95% confidence interval of Mean:

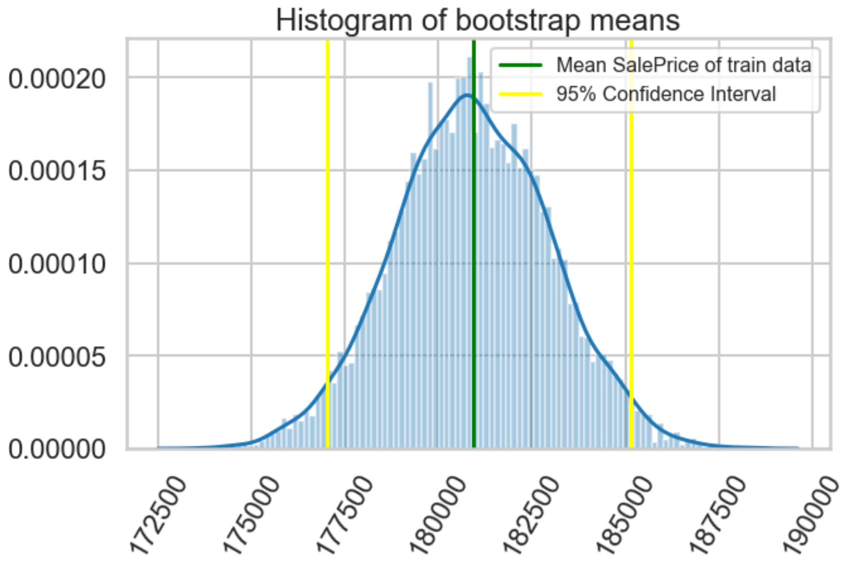
With the above information it's difficult to estimate the population mean. It is more crucial to know if our sample mean is in reasonable range or close to population mean.

However as per Central Limit Theorem, the distribution of the averages from samples of population will be closely approximated by a normal distribution. We can extract the 95% confidence interval of sample means.

If the actual mean is well within this range, we could use this statistic for many base calculations.

For this we will create 10000 bootstrap samples to calculate the sample means.

Then we see the histogram of these bootstrap replicates and calculate the 95% confidence interval.

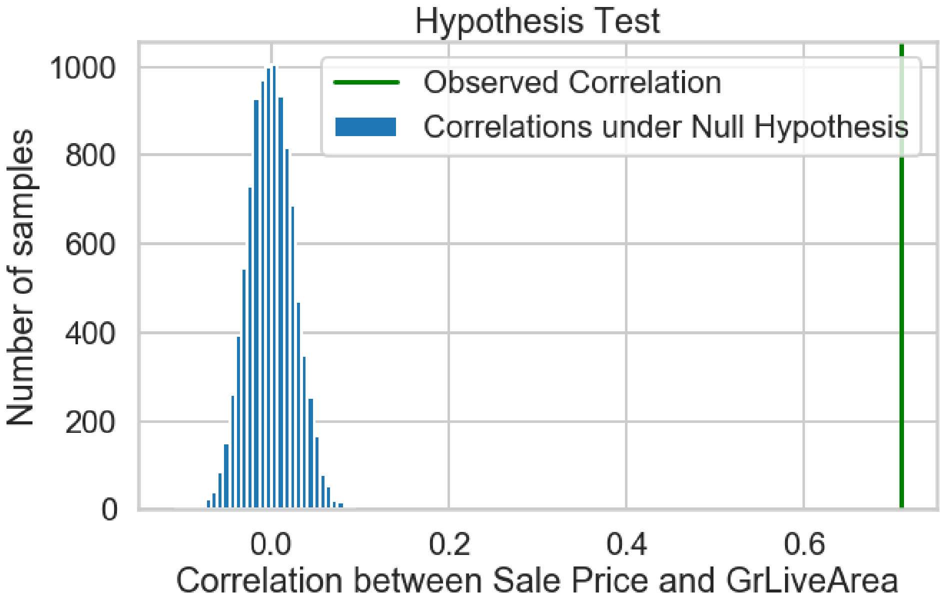


Though as Sale Price has high value outliers in high range, we find that the mean is not just in the 95% confidence interval but is also in the mid high probability range. This gives some confidence of using this mean value for some base calculations.

Hypothesis test on correlation

Since our objective is to predict Sale Price, we looked at the independent variables that have a high correlation with Sale Price and found that GrLivArea has .70 correlation. However, just looking as the correlation is not sufficient proof of a linear relationship as the population data is unknown. For this we need to do a Hypothesis test.

Null Hypothesis H0: There is no correlation between 'GrLiveArea' and 'SalePrice' .
Alternate Hypothesis H1: There is significant correlation between 'GrLiveArea' and 'SalePrice'.
We created permutation samples of GrliveArea and calculated correlations with SalePrice and the p_value. This p value worked out to zero.
P value is zero:
There are no cases like the observed correlation which is close to zero correlation. So we can reject the null hypothesis. This means that there is a significant linear relationship between Sale Price and GrLiveAre and we can go ahead and apply Linear Models. We can also plot it..


Hypothesis Test
Observed Correlation
Correlations under Null Hypothesis
Number of samples
Correlation between Sale Price and GrLiveArea