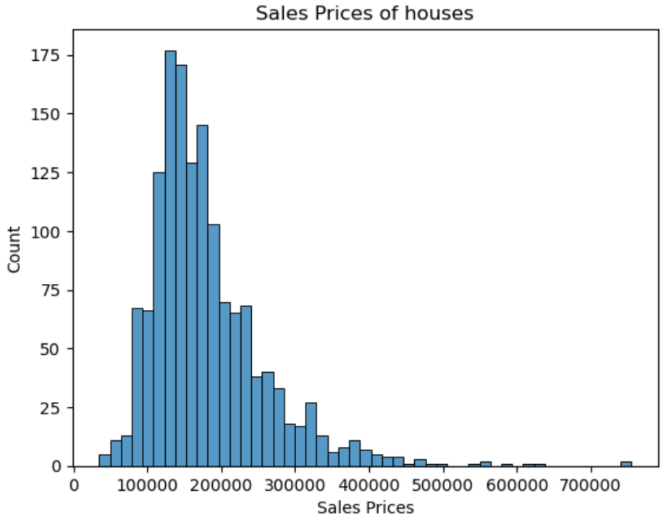
Phan Hoang An Nguyen
PDD Data Analytics
Langara College

Assignment 3 - part 3

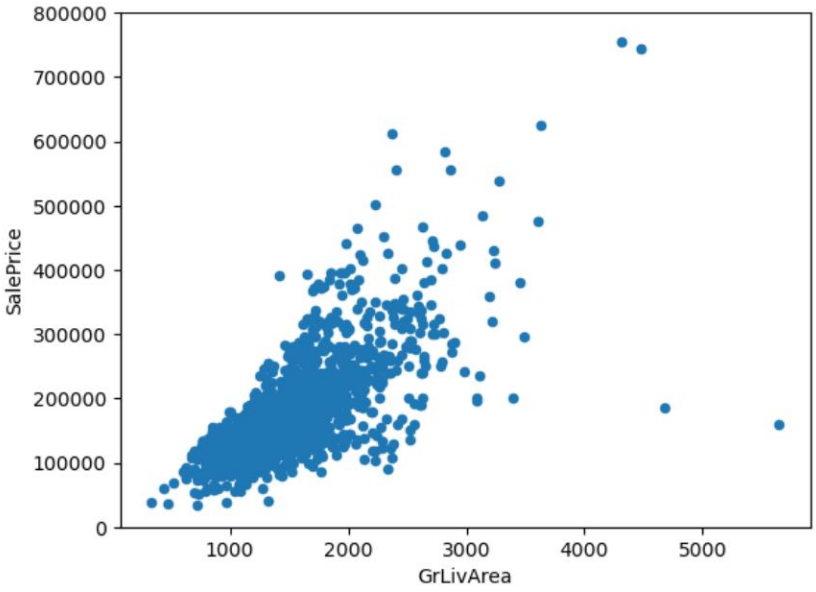
Name of dataset: House Price Data Set Source of dataset: Kaggle

Summary of dataset: total 81 columns x 1460 rows

The main variable of this dataset is the sales price (SalePrice) so below is the correlation between this variable with 3 numerical variables including Ground Living Area (GrLivArea), Total area in square feet of basement (TotalBsmtSF) and Overall quality (OverallQual).



1. **Determine if the sales price is associated to the Ground Living Area.**



⇨ Strong positive, linear relationship between Ground Living Area and Sales price. The larger ground living area will result in higher sales price.

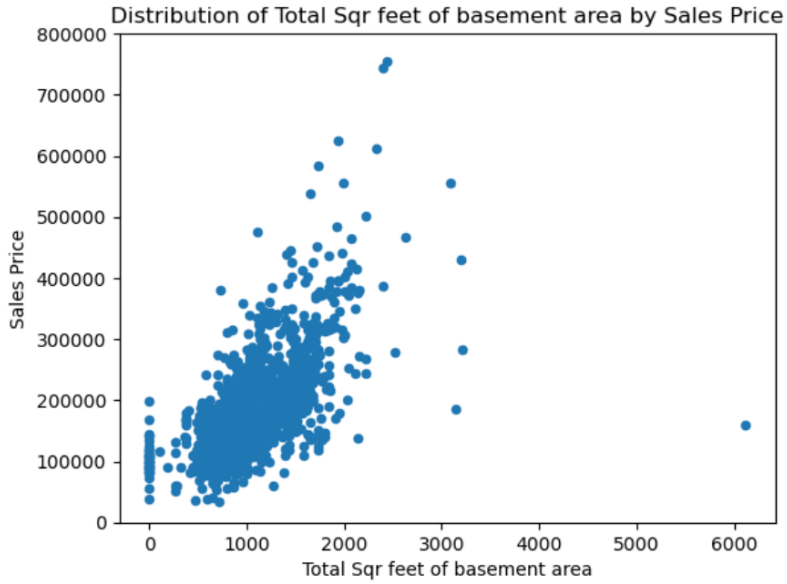
Phan Hoang An Nguyen
PDD Data Analytics
Langara College

Chi-square test

- **Null hypothesis:** Ground Living Area is independent with Sales Price. In other words, there is no association between Ground Living Area and Sales Price
- **Alternative hypothesis:** Ground Living Area is not independent with Sales Price. In other words, there is an association between Ground Living Area and Sales Price

⇨ P-value = 0 < α (=0.05) → Reject H0 → There is an association between Ground Living Area and Sales Price

2. Determine if the Total square feet of basement area is associated to sales price.



⇨ Strong positive, linear relationship between Total square feet of basement area and Sales price. The larger Total square feet of basement area will result in higher sales price.

Chi-square test

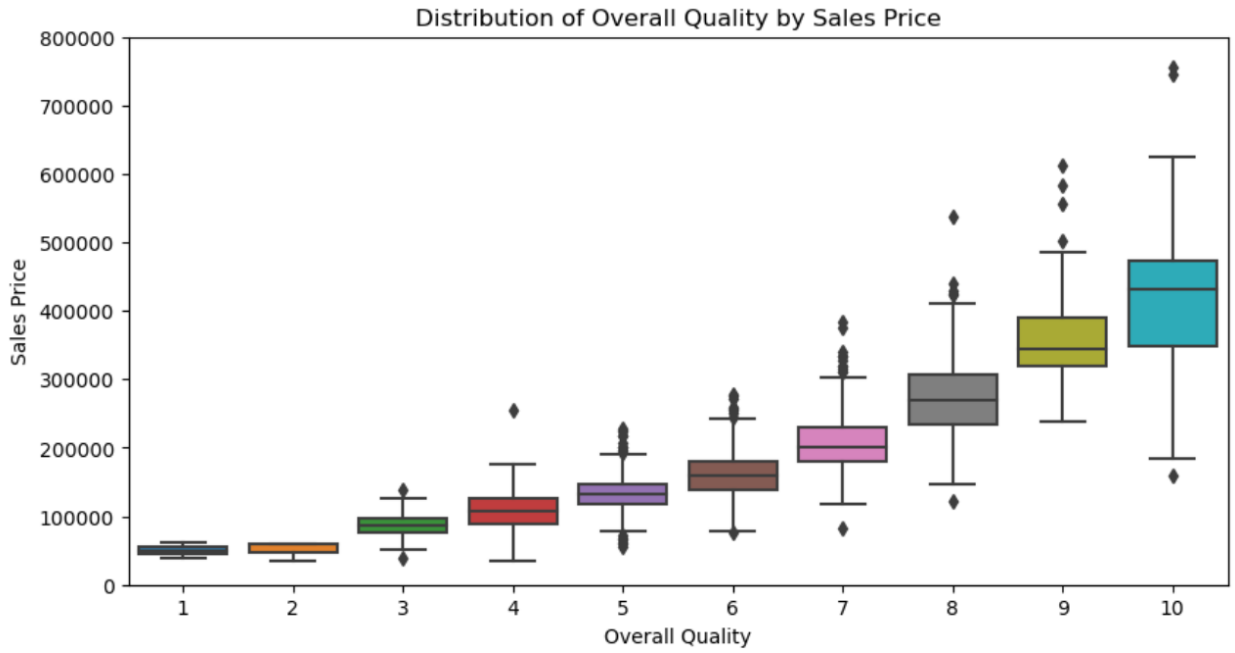
⇨ **Null hypothesis:** Total square feet of basement area is independent of Sales Price. In other words, there is no association between Total square feet of basement area and Sales Price

⇨ **Alternative hypothesis:** Total square feet of basement area is not independent with Sales Price. In other words, there is an association between Total square feet of basement area and Sales Price

⇨ P-value = 0 < α (=0.05) → Reject H0 → There is an association between Total square feet of basement area and Sales Price

3. Determine the Sales price is associated to the Overall Quality.

Phan Hoang An Nguyen
PDD Data Analytics
Langara College



⇨ Higher Overall Quality might result in high sales price with plenty outliers.

Chi-square test

- **Null hypothesis:** Overall Quality is independent of Sales Price. In other words, there is no association between Overall Quality and Sales Price
- **Alternative hypothesis:** Overall Quality is not independent of Sales Price. In other words, there is an association between Overall Quality and Sales Price

⇨ P-value = 1.0 > α (=0.05) → Accept H0 → There is no association between Overall Quality and Sales Price