CS521 C1 Information Structures with Python (Summer 1 2022)
GitHub Username: final-project-FinnG13
Names: Finn Granham, Khudeja Begum, Jiehua Liu
Final Project: Data Analysis with Python (Predicting Housing Prices (Linear Regression))
Due Date: August 12, 2022

Introduction

For our Final Project, we decided to go with the Predicting Housing Prices (Linear Regression). We have concluded a dataset of 2919 rows by 31 columns. We used the House Prediction Data to get a better understanding of the prediction analysis over the time period. We were able to scope out the relevant data to create scatter plots and bar charts of the sale prices. Our primary packages for this final project are pandas for data processing and sklearn to gather the entail. We will use the house price dataset that contains numerous features and detailed statistics about the house and its sale price.

```
import pandas as pd
import sklearn
```

```
# Loaading our dataset
dataset = pd.read_csv("House Prediction Data.csv")

# loading first 5 lines to verify data is formatted correctly csv was read properly
dataset.head()
```

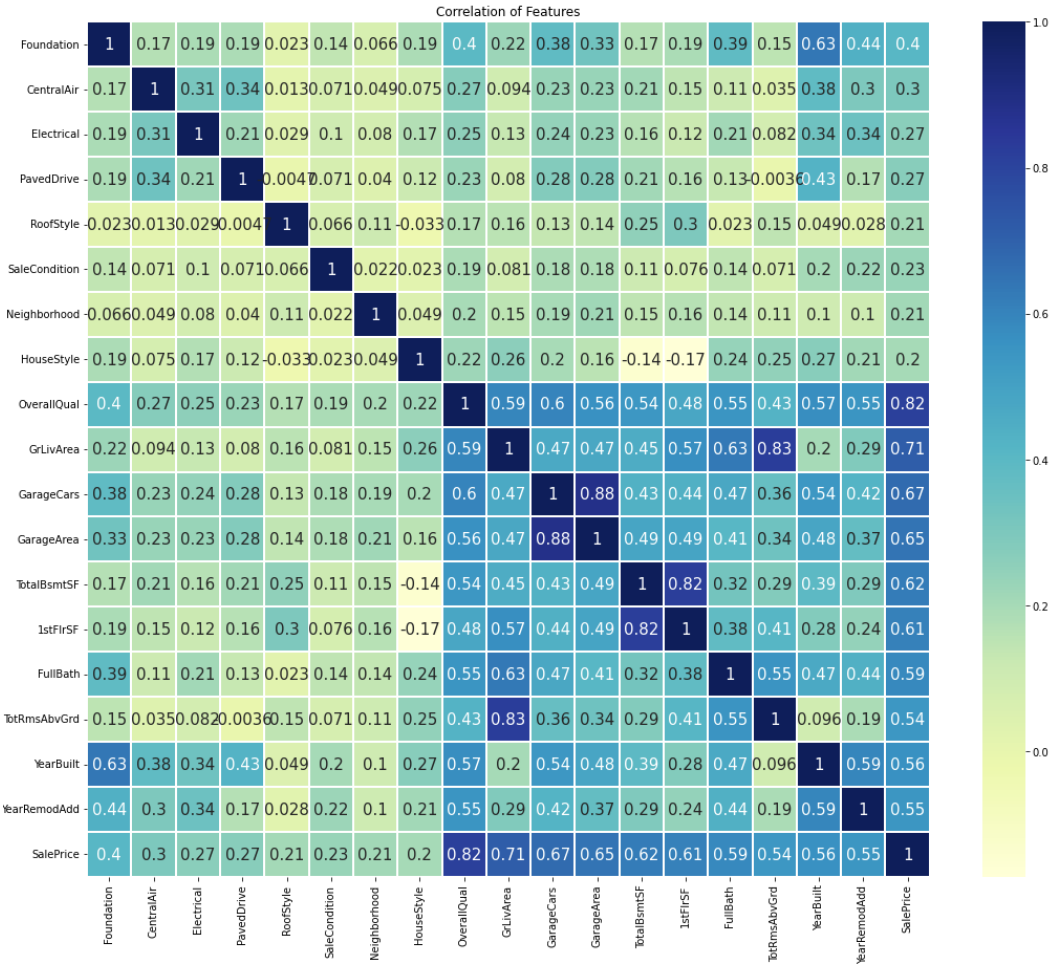
Data Processing

We chose linear regression because it is an algorithm used to predict values that are continuous in nature. Predicting sale prices of houses is an ongoing topic so we would need to use something that is beneficial to predict continuous variables. After processing the data, we have encountered nulls. There are 6 features in this data that contain more than 1000 null values, one of which is the target information, only delete the null value row operation, it cannot be deleted. Some of the features that contained more than 1000 nulls were 'Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature', 'SalePrice'. When some of the features were deleted others would delete them as well. For example, when MiscFeature was deleted, MiscVal also vanished.

Each graph represents the value of the houses over several categories. For example, the scatter plots of ov's numerical features and houses price shown below are representing the values over time. The larger their values the higher the house prices are. As you can see the sale price prediction is based OverallQual, depending on the number of garage cars, layout platform, amount of bathrooms/architecture of the blueprint, and the year it was built. We have constructed multiple datasets but here are three examples listed below. Those three are heatmap, scatter plot and distribution plot.

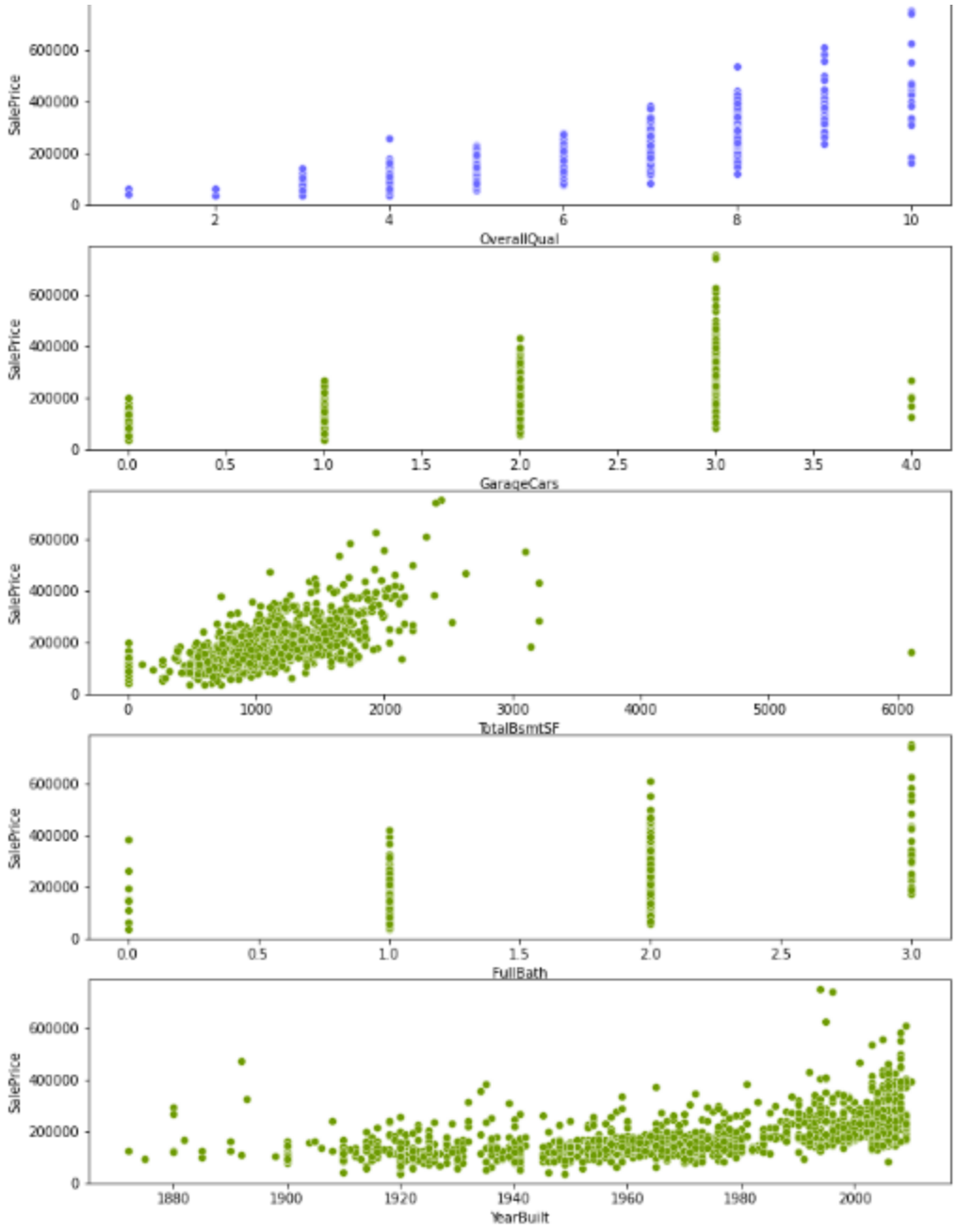
Heatmap:

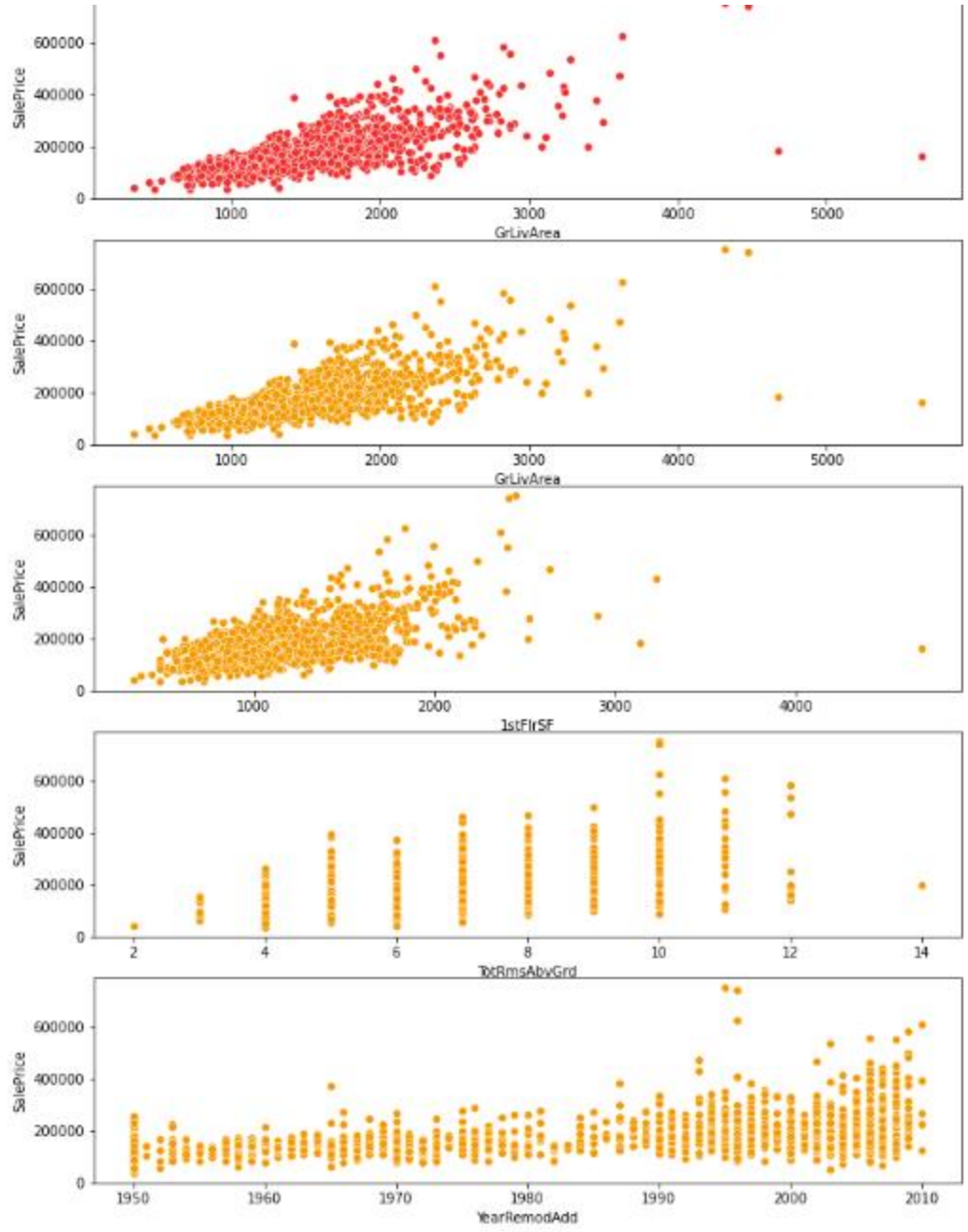
Heatmaps are very useful to find relations between two variables in a dataset.



Scatterplots:

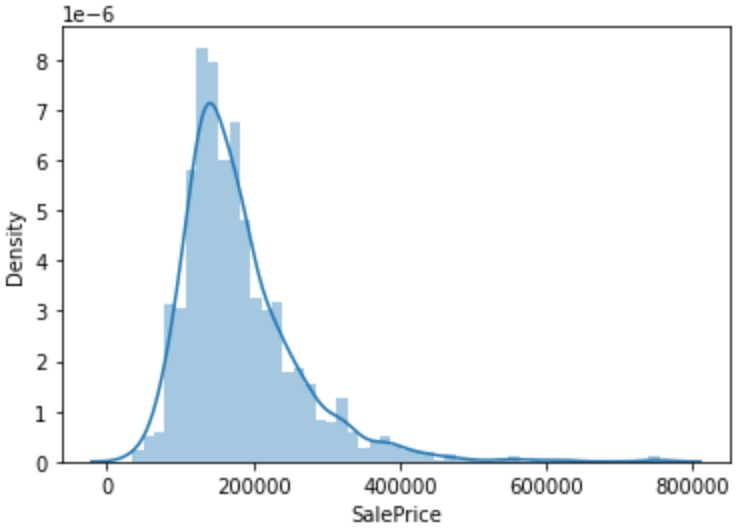
A scatter plot is also used to observe linear relations between two variables in a dataset. In a scatter plot, the dependent variable is marked on the x-axis and the independent variable is marked on the y-axis. In our case, the ‘SalePrice’ attribute is the dependent variable, and every other is the independent variable.


SalePrice
OverallQual
SalePrice
GarageCars
SalePrice
TotalBsmtSF
SalePrice
FullBath
SalePrice
YearBuilt

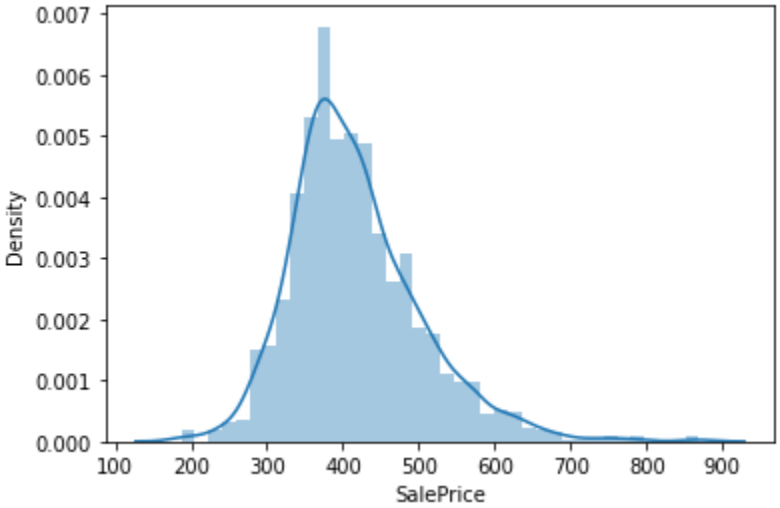


Distribution Plot:

Distribution plots are used to check how well a variable is distributed in the dataset. Listed below is a production distribution plot using the *'density'* function to check the distribution of the *'SalePrice'* variable in the dataset.



It can be seen from the above figure that the target information, that is, the housing price is basically a normal distribution.



By taking the square root of the house price data, a more uniform normal distribution is obtained.


Linear Regression
actual values
actual values
1000
900
800
700
600
500
400
300
200
0
100
200
300
400