Executive Summary

A dataset of houses will be explored and analyzed displaying visualizations of relationships by datamining. There will be four machine learning regression algorithms used to predict the price of houses based on their features. The question to be answered is as follows:

Research Question: Can the price of a house be predicted by regression and which model is best?

House prediction is important because a person might want to estimate how much money will be spent based on features of the house. Other reasons include how much to sell a house for, or home buyers might want to estimate a price range so they can plan their finances. House prediction is also beneficial for property investors to know the trend of housing prices.

In this analysis, we will use one outcome variable which is the price of the house. The rest of the variables will be used to predict the price of the house. The conclusion of this analysis will be based on predictor variables to predict the outcome variable. The regression models to be used are: *Random Forest, Decision Tree, K-Nearest Neighbors, and Ridge Regression.*

Data Source: Kaggle – https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Dataset:

- *Train dataset:* 1460 observations, 81 variables
- *Test dataset:* 1460 observations, 81 variables
  - Both datasets were combined into one dataset and then split into a training and testing set.

- *Combined dataset:* 2920 observations, 81 variables
  - To make this analysis simpler, used numeric variables that made sense to use.
- *Final dataset:* 2920 observations, 12 variables
- *Variable Description:* 12 variables
  - *LotArea* – Lot size in square feet.
  - *OverallQual* – Rates the overall material and finish of the house.
  - *OverallCond* – Rates the overall condition of the house.
  - *TotalBsmtSF* – Total square feet of basement area.
  - *1stFlrSF* – First floor square feet.
  - *FullBath* – Full bathrooms above grade.
  - *BedroomAbvGr* – Bedrooms above grade (does not include basement bedrooms).
  - *TotRmsAbvGrd* – Total rooms above grade (does not include bathrooms).
  - *Fireplaces* – Number of fireplaces.
  - *GarageCars* – Size of garage in car capacity.
  - *GarageArea* – Size of garage in square feet.
  - *SalePrice* – Output variable (Sale price of the house).

Exploratory Data Analysis

The ‘housing’ dataset has 12 variables, which all of them are integers.

	LotArea	OverallQual	OverallCond	TotalBsmtSF	1stFlrSF	FullBath	BedroomAbvGr	TotRmsAbvGrd	Fireplaces	GarageCars	GarageArea	SalePrice
0	8450	7	5	856	856	2	3	8	0	2	548	208500
1	9600	6	8	1262	1262	2	3	6	1	2	460	181500
2	11250	7	5	920	920	2	3	6	1	2	608	223500
3	9550	7	5	756	961	1	3	7	1	3	642	140000
4	14260	8	5	1145	1145	2	4	9	1	3	836	250000

Let’s plot a histogram of the target variable, ‘SalePrice’ and look at a heat map of the correlation between the target variable and the input variables.

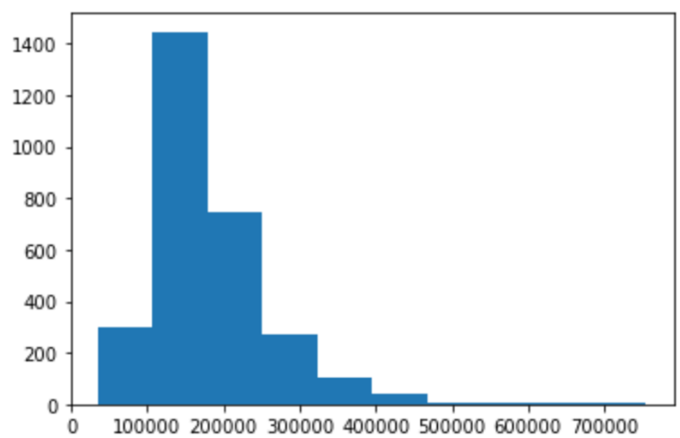

Figure 1

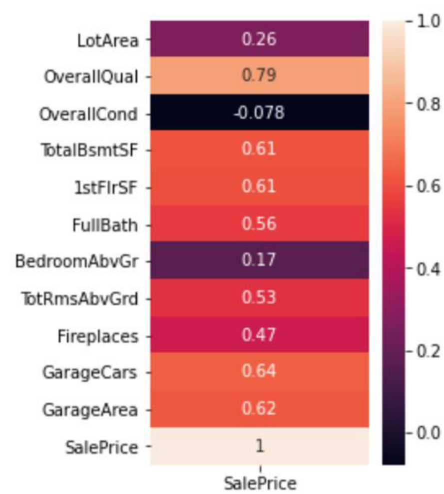

Figure 2

The target variable is right skewed as shown above in figure 1 and figure 2 displays that the best correlated variable to the target variable is ‘OverallQual’.

Let’s examine these relationships by using scatterplots.

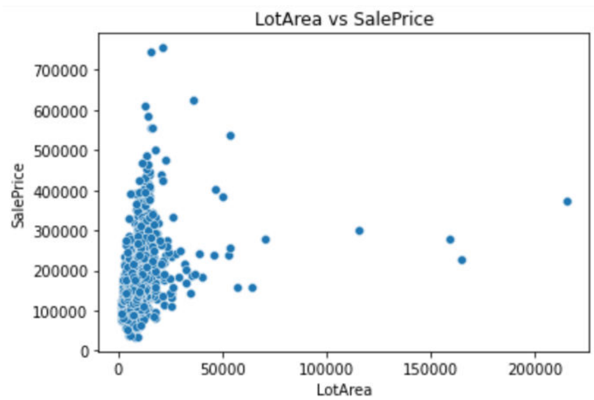

Figure 3

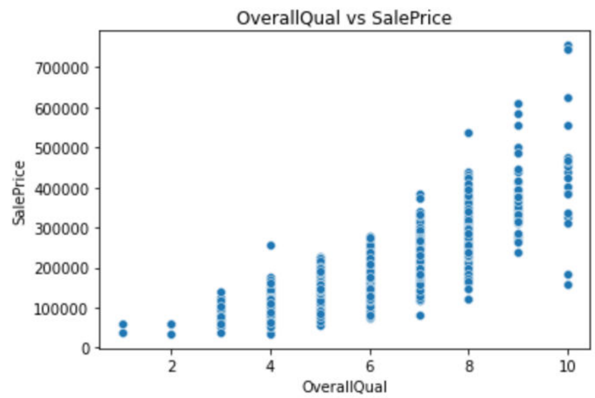

Figure 4

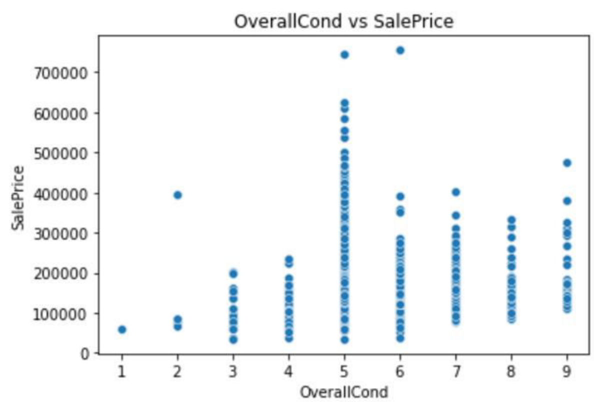

Figure 5

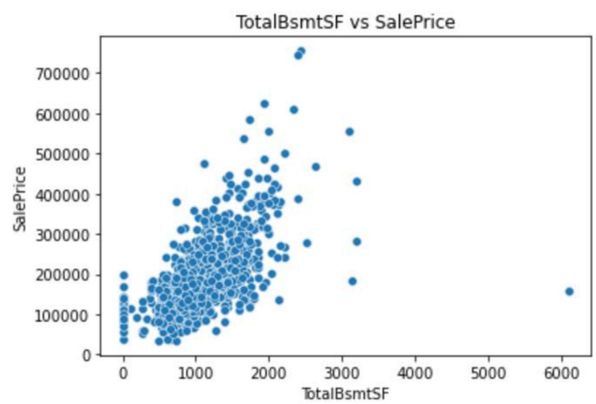

Figure 6

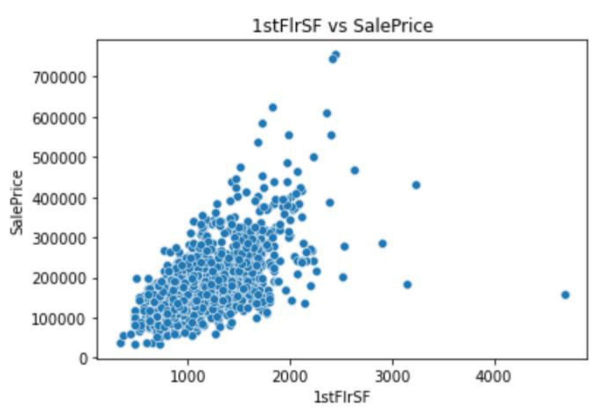

Figure 7

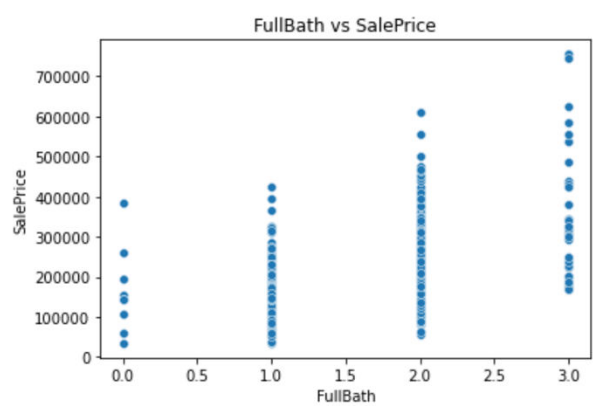

Figure 8

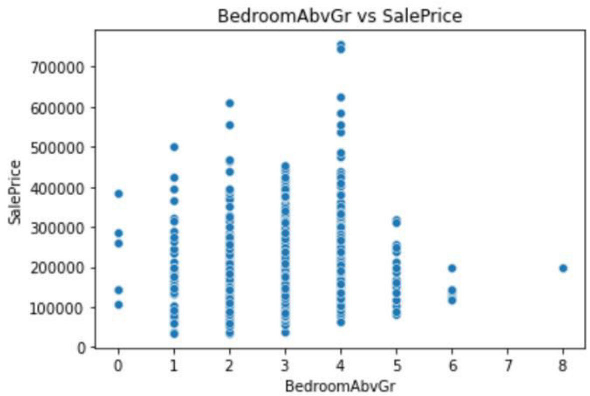

Figure 9

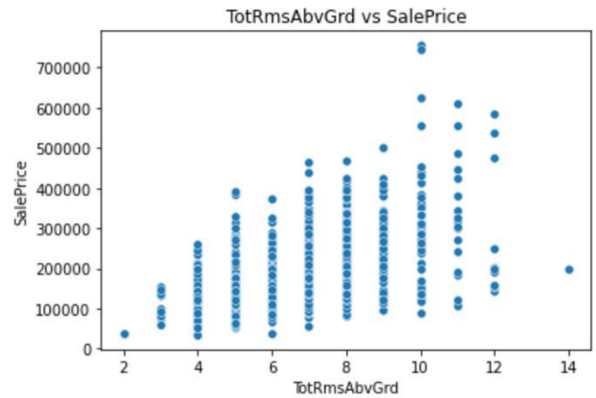

Figure 10

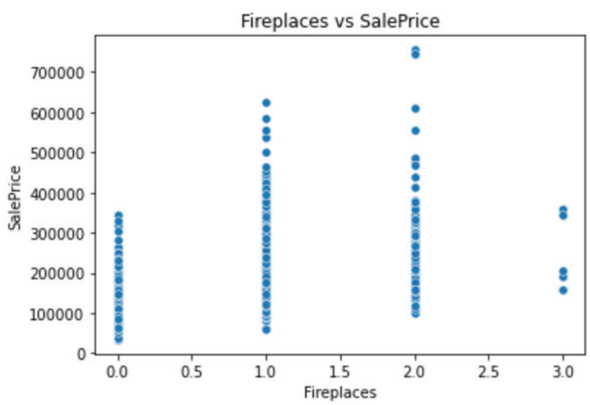

Figure 11

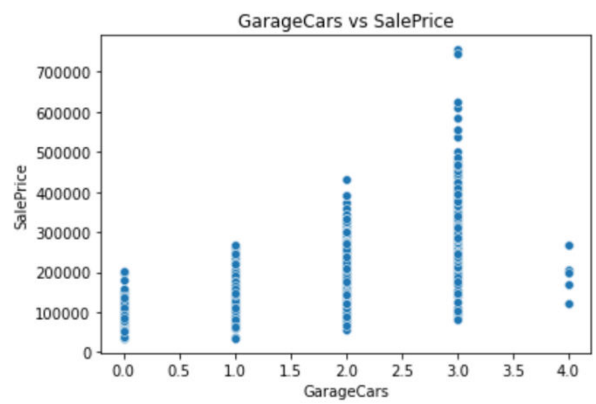

Figure 12

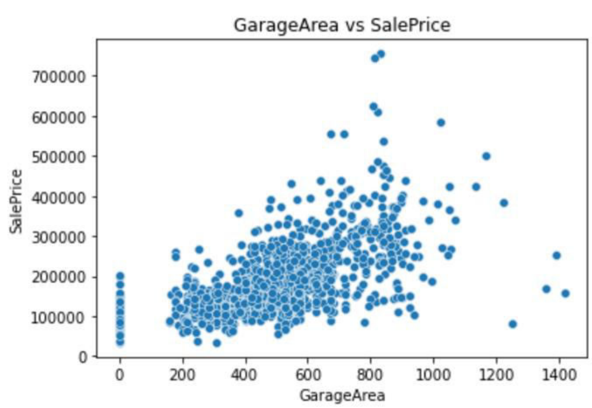

Figure 13

From the scatterplots above, Figure 4 has the best linear relationship which is ‘OverallQual’.

Data Processing

Check for null values

```
data.isna().sum().sum()
```

```
0
```

Scaling the Data

```
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)
```

After dropping the variables that we are not going to use, there are 0 null values, so we scale the data using the StandardScaler in SKLearn. Once the data is scaled, we can fit the models.

Data Modeling

We will get the X and y values of the data and then split them into training and testing sets.

```
X = scaled_data[:,:-1]
y = scaled_data[:,-1]

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3)
```

Random Forest Model

```
regressor_rf = RandomForestRegressor().fit(X_train, y_train)
pred_rf = regressor_rf.predict(X_test)
```

Decision Tree Model

```
regressor_tree = DecisionTreeRegressor().fit(X_train, y_train)
pred_tree = regressor_tree.predict(X_test)
```

K-Nearest Neighbors Model

```
regressor_knn = KNeighborsRegressor(n_neighbors=5).fit(X_train, y_train)
pred_knn = regressor_knn.predict(X_test)
```

Ridge Regression Model

```
ridge = Ridge().fit(X_train, y_train)
pred_ridge = ridge.predict(X_test)
```

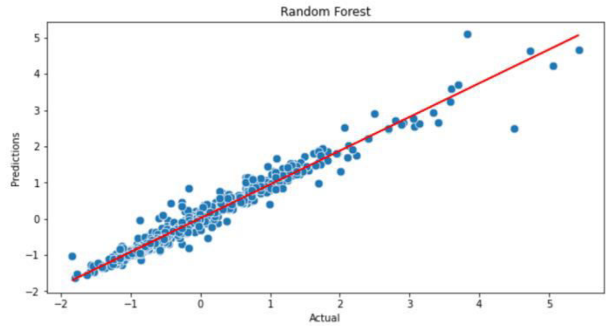

Figure 14

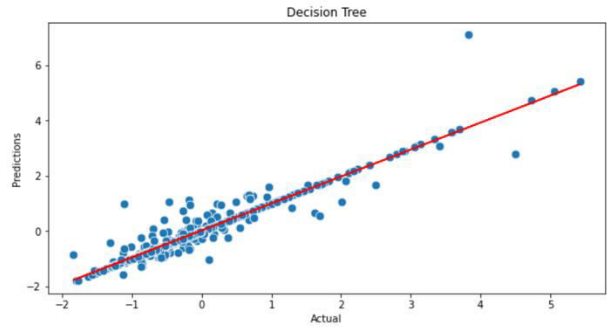

Figure 15

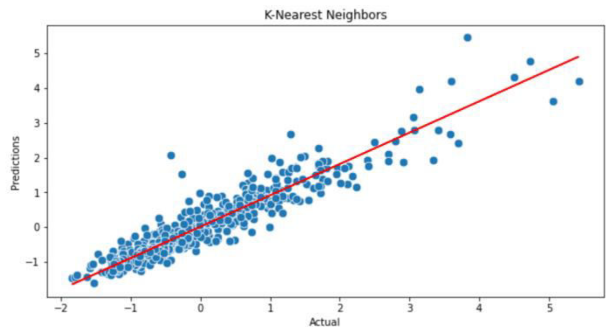

Figure 16

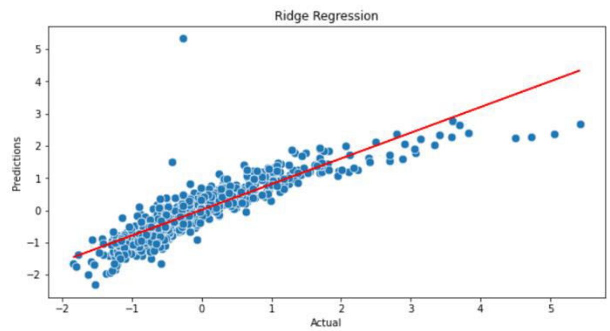

Figure 17

All four regression models represent a good fit for the data.

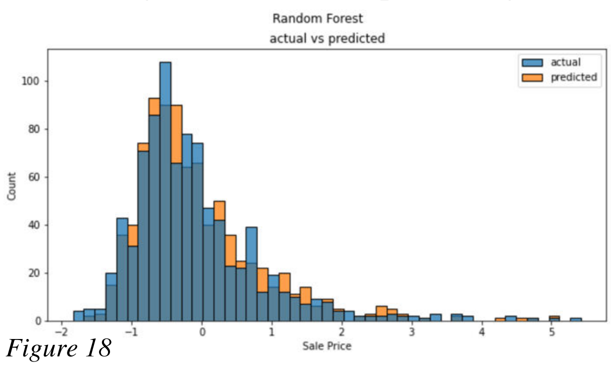

Figure 18

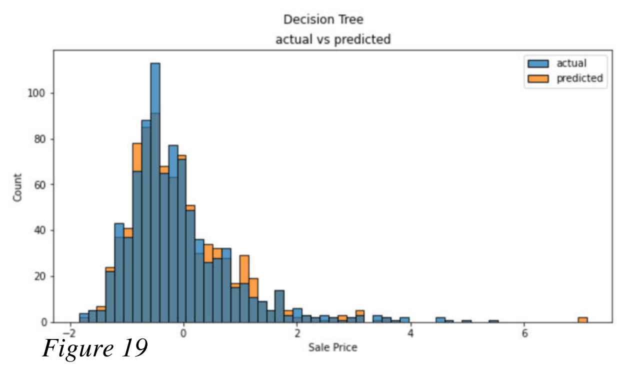

Figure 19

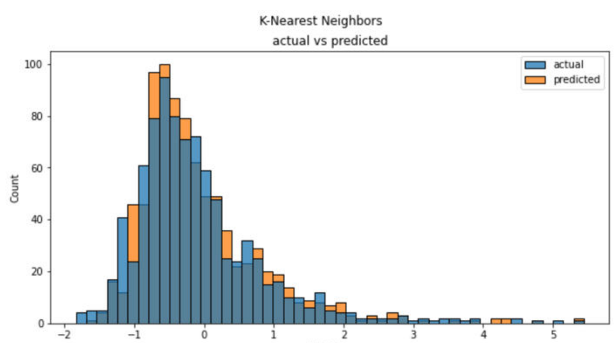

Figure 20

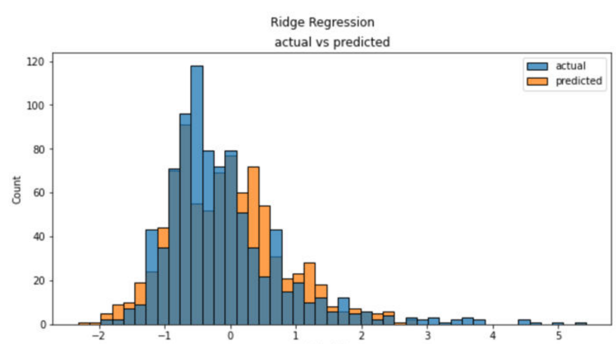

Figure 21

All four regression models make good predictions.

The variable importance plot below in figure 22 shows that 'OverallQual' is the most important variable as expected.

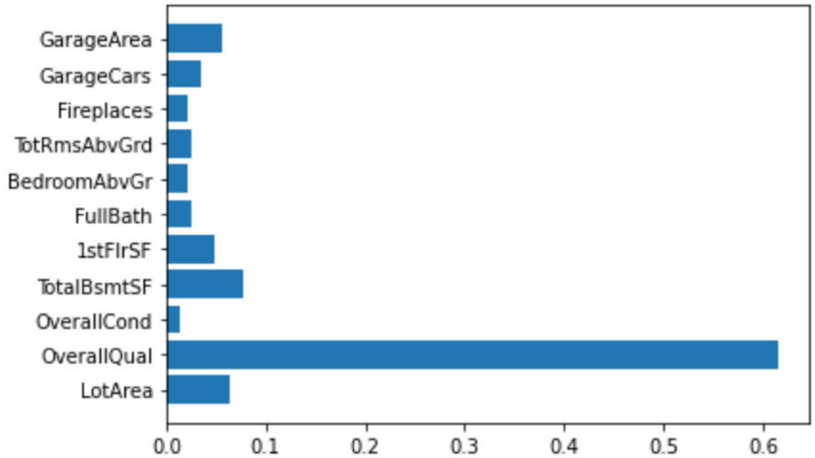

Figure 22

The metric scores for all four models are shown in the table below.

	Metric	Random Forest	Decision Tree	K-Nearest Neighbors	Ridge Regression
0	explained_variance	0.942	0.88937	0.877	0.77061
1	max_error	2.011	3.27393	2.497	5.59862
2	mean_abs_error	0.140	0.11612	0.227	0.30621
3	mean_sq_error	0.054	0.10134	0.112	0.20936
4	med_abs_error	0.081	0.00000	0.158	0.24778
5	R2	0.941	0.88847	0.877	0.76958

Conclusion

All four models represent a linear model. All models predict a good outcome. From the metric table, the Random Forest model performs the best with an explained variance of 0.942 and an R2 of 0.941. The conclusion to our question is yes, the price of a house can be predicted by regression.