Part C: Regression

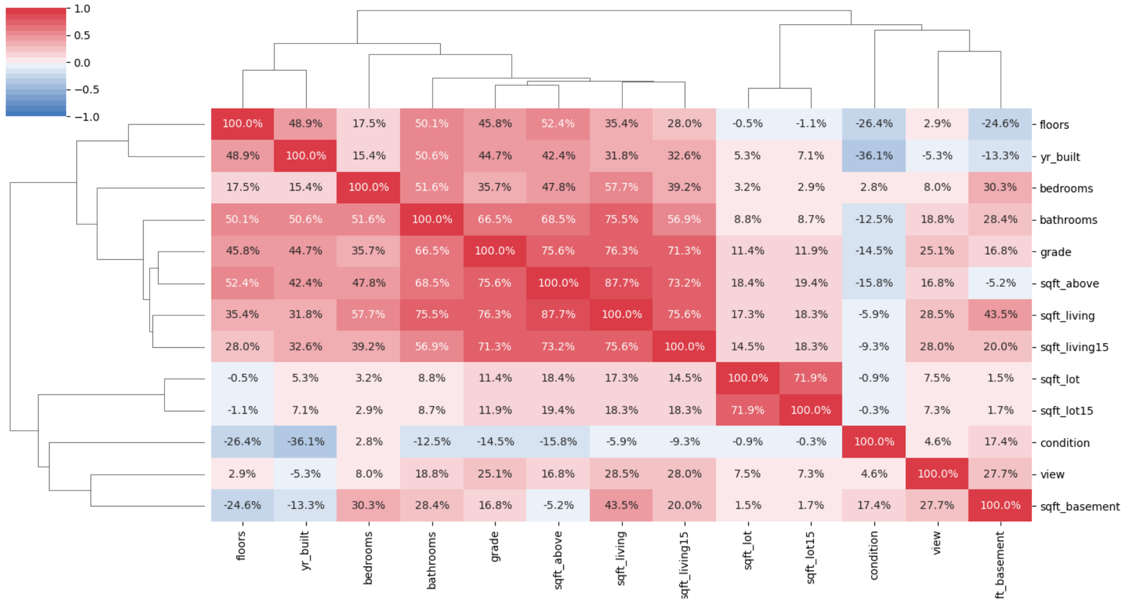
Our aim in this part of the project was to find a suitable dataset and apply various Machine Learning Algorithms for Regression. The dataset that we chose was the “House Sales in King County, USA” and it contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. (House Sales in King County, USA, 2020).

After searching for some time to find a good dataset where the class we are trying to predict has continuous values and its features have meaningful data we found the aforementioned dataset. It didn’t have any missing values also which made it a very good choice for this part of the project.

The dataset has 21 columns from which one represents the price of the houses and the rest the various features of the houses like the number of bedrooms, bathrooms, the square feet of the various rooms, the zip code and more.

Our goal is to use various Regression algorithms to predict the Price (dependent variable) based on the rest 20 features of the dataset (independent variables).

After importing the dataset in a pandas DataFrame we created a heatmap to see the correlation between its features:



	floors	yr_built	bedrooms	bathrooms	grade	sqft_above	sqft_living	sqft_living15	sqft_lot	sqft_lot15	condition	view	ft_basement
floors	100.0%	48.9%	17.5%	50.1%	45.8%	52.4%	35.4%	28.0%	-0.5%	-1.1%	-26.4%	2.9%	-24.6%
yr_built	48.9%	100.0%	15.4%	50.6%	44.7%	42.4%	31.8%	32.6%	5.3%	7.1%	-36.1%	-5.3%	-13.3%
bedrooms	17.5%	15.4%	100.0%	51.6%	35.7%	47.8%	57.7%	39.2%	3.2%	2.9%	2.8%	8.0%	30.3%
bathrooms	50.1%	50.6%	51.6%	100.0%	66.5%	68.5%	75.5%	56.9%	8.8%	8.7%	-12.5%	18.8%	28.4%
grade	45.8%	44.7%	35.7%	66.5%	100.0%	75.6%	76.3%	71.3%	11.4%	11.9%	-14.5%	25.1%	16.8%
sqft_above	52.4%	42.4%	47.8%	68.5%	75.6%	100.0%	87.7%	73.2%	18.4%	19.4%	-15.8%	16.8%	-5.2%
sqft_living	35.4%	31.8%	57.7%	75.5%	76.3%	87.7%	100.0%	75.6%	17.3%	18.3%	-5.9%	28.5%	43.5%
sqft_living15	28.0%	32.6%	39.2%	56.9%	71.3%	73.2%	75.6%	100.0%	14.5%	18.3%	-9.3%	28.0%	20.0%
sqft_lot	-0.5%	5.3%	3.2%	8.8%	11.4%	18.4%	17.3%	14.5%	100.0%	71.9%	-0.9%	7.5%	1.5%
sqft_lot15	-1.1%	7.1%	2.9%	8.7%	11.9%	19.4%	18.3%	18.3%	71.9%	100.0%	-0.3%	7.3%	1.7%
condition	-26.4%	-36.1%	2.8%	-12.5%	-14.5%	-15.8%	-5.9%	-9.3%	-0.9%	-0.3%	100.0%	4.6%	17.4%
view	2.9%	-5.3%	8.0%	18.8%	25.1%	16.8%	28.5%	28.0%	7.5%	7.3%	4.6%	100.0%	27.7%
sqft_basement	-24.6%	-13.3%	30.3%	28.4%	16.8%	-5.2%	43.5%	20.0%	1.5%	1.7%	17.4%	27.7%	100.0%

Figure 1 of Part 3: Heatmap showing correlation between features

We then removed the outliers from our dataset and made a matrix with scatter plots to get a better understanding of how price is related to the rest of the features:

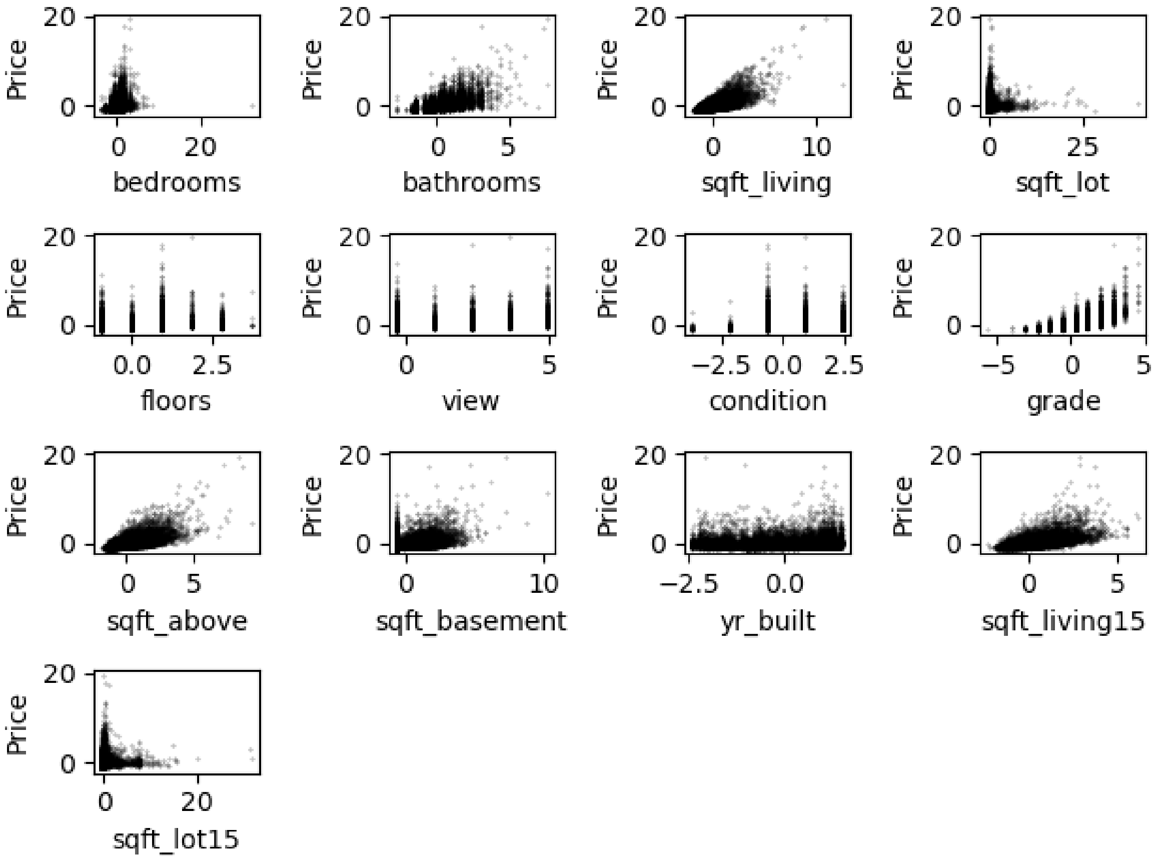


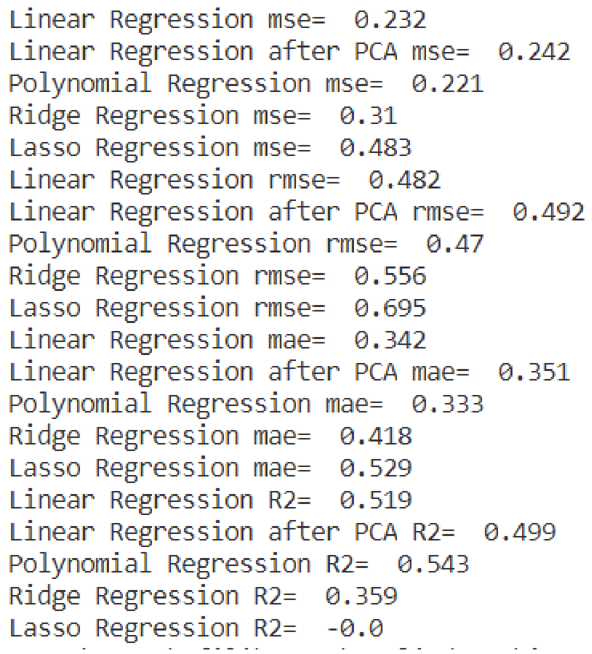
Figure 2 of part 3: Scatter plot of Price vs features

After that we performed feature selection using Principal component analysis and we decided to use 7 components which led us to an Explained Variance of 0.895 .

We then split the dataset with all the features and the dataset after performing PCA with a 50-50 train test split and we run 4 Machine Learning Regression algorithms:

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression

We then measured the Mean Square Error(MSE), Root Mean Square Error(RMSE), Mean Absolute Error(MAE) and the R2 Score.



```
Linear Regression mse=  0.232
Linear Regression after PCA mse=  0.242
Polynomial Regression mse=  0.221
Ridge Regression mse=  0.31
Lasso Regression mse=  0.483
Linear Regression rmse=  0.482
Linear Regression after PCA rmse=  0.492
Polynomial Regression rmse=  0.47
Ridge Regression rmse=  0.556
Lasso Regression rmse=  0.695
Linear Regression mae=  0.342
Linear Regression after PCA mae=  0.351
Polynomial Regression mae=  0.333
Ridge Regression mae=  0.418
Lasso Regression mae=  0.529
Linear Regression R2=  0.519
Linear Regression after PCA R2=  0.499
Polynomial Regression R2=  0.543
Ridge Regression R2=  0.359
Lasso Regression R2=  -0.0
```

Figure 3: MSE, RMSE, MAE and R2 Score

From the above measures we can clearly see that the Polynomial Regression had the best prediction for the value of the Price

We then created two plots to visualize the residuals and the coefficients:

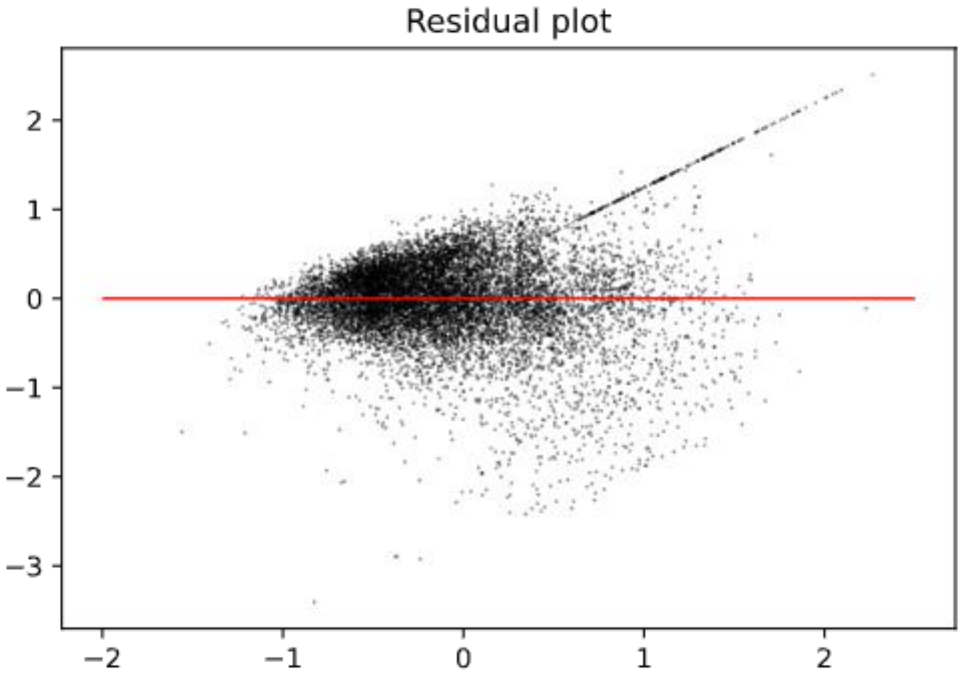


Figure 4 of part 3: Residual plot

In the residual plot of Figure 4 we have the predicted values of the price as the x-axis and the predicted values minus the actual values as the y-axis. The values on the red line, that is the 0 for the y-axis, are the ones that the predicted value is the same as the actual value. The values above the red line, positive y- axis values, show that the predicted values were lower than the actual values where the values that are below the red line, negative y-axis values, show that the predicted values were higher than the actual values.

In our case, Figure 4 exhibits “heteroscedasticity”, meaning that as the prediction value increases so does the residuals. That indicates that we might have to transform a variable or that a variable from our dataset is missing.

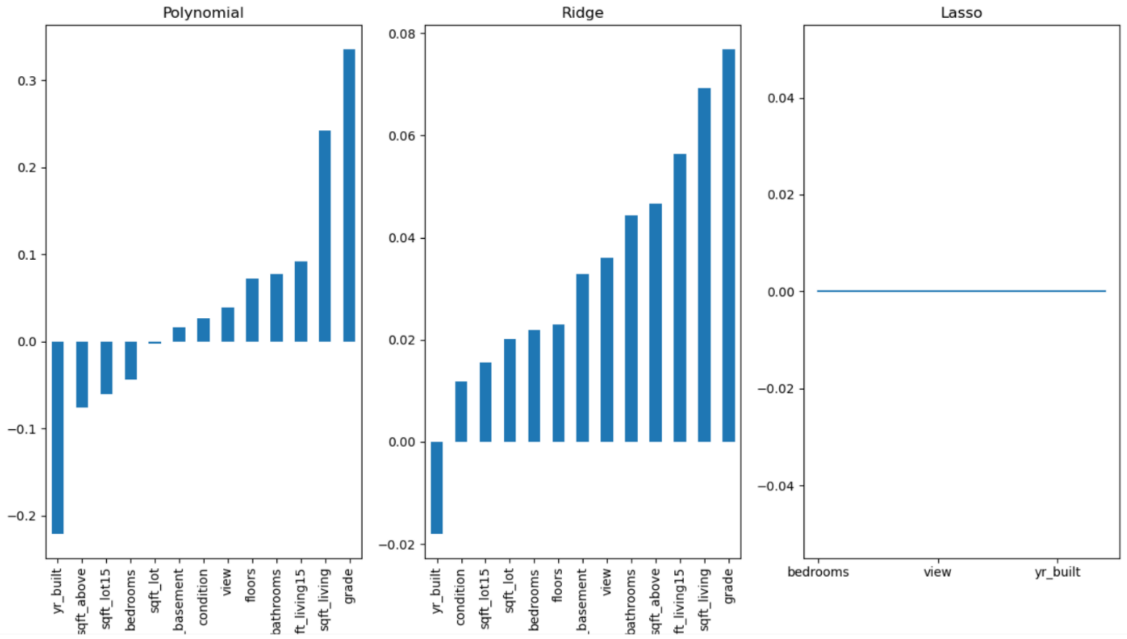


Figure 5 of part 3: Coefficients plot

In Figure 5 we can see the different values of the coefficients based on the type of regression. We can see that for the same coefficient ridge regression is penalizing the coefficients that take large values but almost leaves the same the coefficients with smaller values compared to Polynomial Regression. That is because Ridge Regression is performing a form of regularization and favors small value parameters. On the other hand, we can see that Lasso's Regression coefficient values are zero because Lasso works very similar to Ridge Regression but with a big difference. They both have a penalizing factor and favor small value parameters but, in contrast to Ridge Regression, Lasso Regression can turn many parameter values to zero. By doing that it excludes all the useless parameters and keeps the most useful ones.