---Lefei Liu

Report

1. Introduction:

Provide an individual report based on project file data and complete the 4 steps to complete project requirements using the Ames Housing Dataset:

NAME: AmesHousing.txt

TYPE: Population

SIZE: 2930 observations, 82 variables

ARTICLE TITLE: Ames Iowa: Alternative to the Boston Housing Data Set

DESCRIPTIVE ABSTRACT: Data set contains information from the Ames

Assessor's Office used in computing assessed values for individual

residential properties sold in Ames, IA from 2006 to 2010.

State the objective: Predicting house prices using multiple regression methods.

the significance of this analysis in real estate valuation: The significance of this analysis in real estate valuation lies in its ability to provide accurate and reliable estimates of property values. Accurately predicting house prices is crucial for various stakeholders in the real estate market, including buyers, sellers, investors, and financial institutions. It helps in making informed decisions about buying, selling, investing, and lending.

2. Analysis:

In the preprocessing phase, missing values in the Ames Housing dataset were handled, primarily through median imputation, to ensure data integrity. This step was crucial to prepare the dataset for accurate analysis. Then, **20 continuous variables** were carefully selected based on their relevance and potential impact on house pricing,

handling missing data: we found high correlated vars and marked these.

Moreover, The exploratory data analysis involved generating **boxplots** and **correlation plots**. Boxplots were used to understand the distribution and identify outliers in these variables, while correlation plots helped in examining the relationships and dependencies among them.

3. **Results:** the main outcomes from the regression models

Stepwise Regression (AIC and BIC):

The models selected based on AIC and BIC have included variables that the stepwise algorithm determined to be statistically significant contributors to predicting the SalePrice.

Model Summaries:

The model in summary we can get the coefficients for the variables included in the final models after the stepwise selection. Also provide statistics the standard error, t-values, and p-values for each coefficient, indicating the reliability and significance of each predictor. The residual standard error and R-squared values give an indication of the model's fit, with the R-squared value representing the proportion of variance in the SalePrice that is explained by the model.

Lasso and Elastic Net Plots:

The plots show the paths of the coefficients as the penalty parameter is varied in Lasso and Elastic Net models. The x-axis represents the L1 norm of the coefficients or the penalty applied, while the y-axis shows the coefficient values.

These plots help in understanding how each predictor's importance changes as the model complexity is adjusted. and we find that trying to plot the lasso model we get **"1 or less nonzero coefficients; glmnet plot is not meaningful",** I guess This occurs when the penalty imposed during the regularization process is so large that it drives all or all but one of the coefficients to zero

Calculate mean prediction error for each model: The output shown shows the average prediction error for each of the four regression models applied to the Ames Housing dataset: AIC, BIC, Lasso, and Elastic Net. Average prediction error is a measure of the average difference between actual and predicted values; quantifies the average magnitude of the model's prediction error.

AIC Model: The mean prediction error for the model selected by the Akaike Information Criterion is approximately 1.15 billion.

BIC Model: The mean prediction error for the model selected by the Bayesian Information Criterion is also about 1.15 billion, which suggests it may be very similar or the same as the AIC model in this case.

Lasso Model: The Lasso regression model, which applies an L1 penalty to achieve sparsity and potentially better generalization, has a mean prediction error of approximately 1.04 billion.

Elastic Net Model: The Elastic Net regression model, which is a combination of L1 and L2 penalties, has a mean prediction error of approximately 1.27 billion.

4. Conclusion:

Based on these results, the Lasso model had the lowest average prediction error, which may indicate that it has the best prediction performance among the models evaluated

Future work:

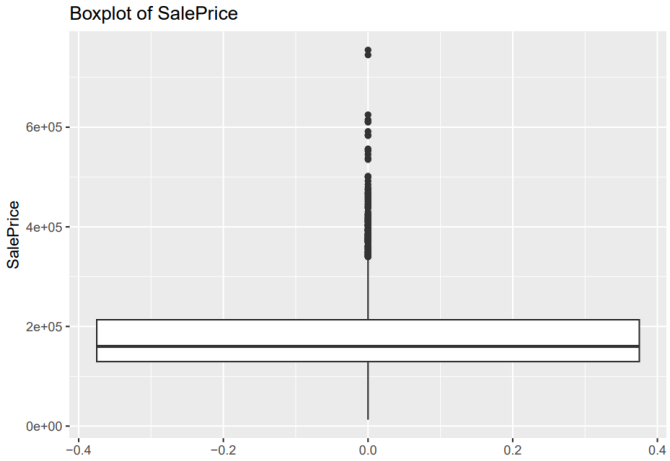
Incorporating Additional Data: Expanding the dataset with more recent sales,

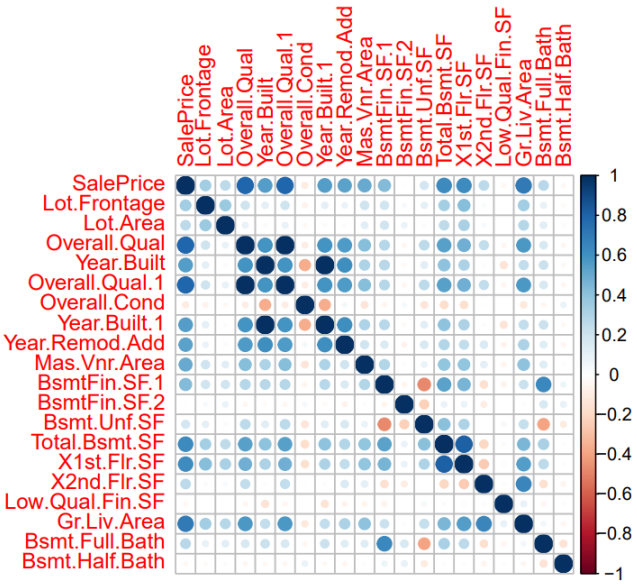
additional geographic locations, or more granular data such as neighborhood characteristics could improve model robustness and predictive power.

Feature Engineering: Developing new features from existing data, such as creating interaction terms or polynomial features, could uncover complex relationships between variables and the target outcome.

5. Appendix(not include in pages):

- **Boxplot and corrplot**





---Lefei Liu

- **Highly Correlated vars code:**

```
# Split Data into Training and Testing Sets
set.seed(2023)
test_indices <- sample(nrow(housing), round(nrow(housing)/4))
train_set <- housing[-test_indices, ]
test_set <- housing[test_indices, ]

# Check for Highly Correlated Predictors
corr_matrix <- cor(train_set[, selected_vars], use = "complete.obs")
high_corr <- findCorrelation(corr_matrix, cutoff = 0.75)
high_corr_vars <- names(train_set[, selected_vars])[high_corr]
high_corr_vars
```

```
## [1] "SalePrice"     "Overall.Qual"  "Total.Bsmt.SF" "Year.Built"
```

```
selected_high_corr_vars <- setdiff(selected_vars, c("SalePrice", "Overall.Qual", "Year.Built", "Total.B
```

- **Summarize AIC and BIC models:**

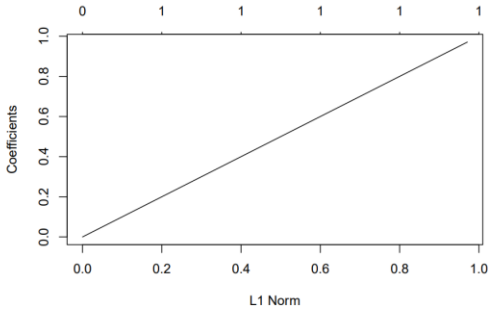
```
# Summarize AIC and BIC models
summary(aic_model)
```

```
##
## Call:
## lm(formula = SalePrice ~ Lot.Frontage + Lot.Area + Overall.Qual +
##     Year.Built + Overall.Cond + Year.Remod.Add + Mas.Vnr.Area +
##     BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF + X1st.Flr.SF +
##     X2nd.Flr.SF + Low.Qual.Fin.SF + Bsmt.Full.Bath, data = train_set[,
##     selected_vars])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -577788  -17002   -2041   13882  256052
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.281e+06  9.002e+04 -14.224  < 2e-16 ***
## Lot.Frontage     1.084e+02  3.957e+01   2.740  0.00619 **
## Lot.Area         6.146e-01  9.519e-02   6.456 1.32e-10 ***
## Overall.Qual     2.102e+04  8.670e+02  24.249  < 2e-16 ***
## Year.Built       3.841e+02  4.039e+01   9.509  < 2e-16 ***
## Overall.Cond     4.610e+03  7.932e+02   5.811 7.11e-09 ***
## Year.Remod.Add   2.140e+02  5.169e+01   4.140 3.61e-05 ***
## Mas.Vnr.Area     2.854e+01  4.748e+00   6.012 2.15e-09 ***
## BsmtFin.SF.1     2.318e+01  3.564e+00   6.505 9.58e-11 ***
## BsmtFin.SF.2     1.489e+01  5.438e+00   2.738  0.00623 **
## Bsmt.Unf.SF      1.059e+01  3.217e+00   3.292  0.00101 **
## X1st.Flr.SF      6.098e+01  3.616e+00  16.861  < 2e-16 ***
## X2nd.Flr.SF      5.256e+01  2.136e+00  24.609  < 2e-16 ***
## Low.Qual.Fin.SF  2.734e+01  1.640e+01   1.667  0.09575 .
## Bsmt.Full.Bath   8.192e+03  1.965e+03   4.169 3.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34810 on 2181 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.8048
```

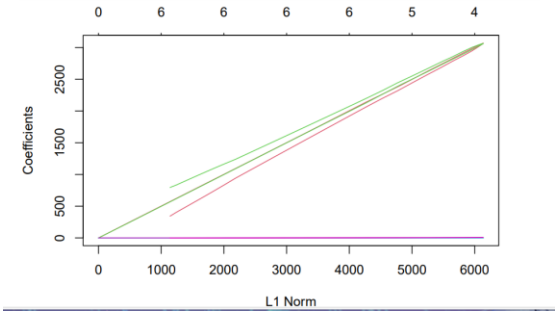
```
## F-statistic: 647.3 on 14 and 2181 DF,  p-value: < 2.2e-16
summary(bic_model)

##
## Call:
## lm(formula = SalePrice ~ Lot.Frontage + Lot.Area + Overall.Qual +
##     Year.Built + Overall.Cond + Year.Remod.Add + Mas.Vnr.Area +
##     BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF + X1st.Flr.SF +
##     X2nd.Flr.SF + Low.Qual.Fin.SF + Bsmt.Full.Bath, data = train_set[,
##     selected_vars])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -577788  -17002   -2041   13882  256052
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.281e+06  9.002e+04 -14.224  < 2e-16 ***
## Lot.Frontage     1.084e+02  3.957e+01   2.740  0.00619 **
## Lot.Area         6.146e-01  9.519e-02   6.456 1.32e-10 ***
## Overall.Qual     2.102e+04  8.670e+02  24.249  < 2e-16 ***
## Year.Built       3.841e+02  4.039e+01   9.509  < 2e-16 ***
## Overall.Cond     4.610e+03  7.932e+02   5.811 7.11e-09 ***
## Year.Remod.Add   2.140e+02  5.169e+01   4.140 3.61e-05 ***
## Mas.Vnr.Area     2.854e+01  4.748e+00   6.012 2.15e-09 ***
## BsmtFin.SF.1     2.318e+01  3.564e+00   6.505 9.58e-11 ***
## BsmtFin.SF.2     1.489e+01  5.438e+00   2.738  0.00623 **
## Bsmt.Unf.SF      1.059e+01  3.217e+00   3.292  0.00101 **
## X1st.Flr.SF      6.098e+01  3.616e+00  16.861  < 2e-16 ***
## X2nd.Flr.SF      5.256e+01  2.136e+00  24.609  < 2e-16 ***
## Low.Qual.Fin.SF  2.734e+01  1.640e+01   1.667  0.09575 .
## Bsmt.Full.Bath   8.192e+03  1.965e+03   4.169 3.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34810 on 2181 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.8048
## F-statistic: 647.3 on 14 and 2181 DF,  p-value: < 2.2e-16
```

- **Plot Lasso and Elastic Net models:**




```
plot(elastic_net_model)
```



- **Calculate mean prediction error for each mode:**

---Lefei Liu

```r
# Calculate Mean Prediction Error
predict_error <- function(model, x, y) {
  predictions <- predict(model, newx = x)
  mean((predictions - y)^2)
}

# Calculate mean prediction error for each model
mean_error_aic <- predict_error(aic_model, x_test, y_test)
mean_error_bic <- predict_error(bic_model, x_test, y_test)
mean_error_lasso <- predict_error(lasso_model, x_test, y_test)
mean_error_elastic_net <- predict_error(elastic_net_model, x_test, y_test)

# Output the errors
c(AIC = mean_error_aic, BIC = mean_error_bic, Lasso = mean_error_lasso, ElasticNet = mean_error_elastic
```

```
##         AIC         BIC       Lasso  ElasticNet
## 11536668354 11536668354  1040106431  1272404579
```

AmesHousing.txt

AmesDataDocumentation.txt

project2.rmd

project 2 with output.pdf