

MSDS 6371 Project

Table of Contents

Team Members	3
Introduction	3
Data Description.....	3
Analysis Question 1.....	3
Restatement of the Problem.....	3
Build and Fit the Model.....	4
Checking Assumptions	4
Residual Plots	4
Influential point analysis (Cook's D and Leverage).....	5
Assumptions.....	5
Comparing Competing Models	6
Adj R^2	6
Parameters	6
Estimates and Interpretation	6
Confidence Intervals	6
Conclusion.....	7
R Shiny: Price v. Living Area Chart.....	7
Analysis Question 2:.....	8
Step 1: Understand the Data.....	8
Step 2: Use Variable Selection Methods.....	8
Step 3: Custom Model Selection.....	9
Kaggle Score	9
References.....	9
Appendix	10
Analysis 1	10
Build and Fit Model.....	10
Checking Assumptions	11
Residual Plots	11

Influential point analysis (Cook's D and Leverage).....	12
Assumptions.....	15
Comparing Competing Models	16
R Shiny: Price v. Living Area Chart	17
Analysis Question 2:	17
Step 1: Understand the Data.....	17
Backward Results	19
Forward Results.....	19
Stepwise Results.....	19
Custom Results.....	20

Team Members

Lani Lewis, Jake Rastberger and Gwonchan Jason Yoon

Introduction

In the realm of real estate valuation and market analysis, the accurate prediction of property sale prices plays a pivotal role in decision-making processes for both buyers and sellers. As the housing market continuously evolves, leveraging advanced data science methodologies becomes imperative to unravel the intricate relationships between a multitude of factors that contribute to property valuation. This paper embarks on a comprehensive exploration of housing dynamics within Ames, Iowa, employing a rich dataset encompassing 79 distinct explanatory variables. Our primary objective is to harness the power of linear regression techniques to construct a robust predictive model capable of estimating house sale prices with a high degree of precision.

Data Description

At the heart of our study lies the Ames Housing dataset, meticulously compiled by Dean De Cock to enhance data science education. This dataset offers a comprehensive glimpse into the real estate landscape of Ames, Iowa, boasting an intricate web of 79 distinct features. These features meticulously encapsulate various facets of housing attributes and contextual variables that collectively sway property valuations.

With an extensive ensemble of 1460 observations in the training dataset alone, the dataset provides a holistic view of residential properties. Its depth and breadth, combined with its accessibility on Kaggle's platform, render it an invaluable resource for delving into the intricate interplay of variables that dictate house sale prices in Ames. To explore the dataset further, please refer to the link:

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>. In the ensuing sections, we harness the potency of linear regression techniques to distill profound insights from this repository of data, with the ultimate aim of constructing a predictive model that unravels the multifaceted dynamics underlying property valuations in Ames, Iowa.

We used many different variables in our analysis if you need further clarification on what the variables mean please reference this [document](#).

Analysis Question 1

Restatement of the Problem

Entrusted by Century 21 Ames, a prominent real estate company located in Ames, Iowa, we are tasked with addressing a pivotal inquiry that holds immense significance for their operations. This inquiry revolves around unraveling the intricate connections between house sale prices and specific attributes of

residential properties. In particular, Century 21 Ames exclusively deals with properties nestled within the NAmes, Edwards, and BrkSide neighborhoods. Their interest lies in comprehending the impact of living area square footage (GrLivArea) on the sale price of a house, while also discerning if this relationship is contingent on the neighborhood in which the property is situated.

In response to this crucial question, our objective is to construct and calibrate a predictive model that succinctly captures the interplay between living area square footage and sale prices. This model will be tailored to accommodate the preferences of real estate professionals, who commonly prefer discussing living area in increments of 100 square feet. The crux of our analysis will not only yield estimates of the relationship between sale prices and living area for the overall dataset but will also shed light on potential variations in this relationship across the NAmes, Edwards, and BrkSide neighborhoods. Through this meticulous analysis, we aim to equip Century 21 Ames with actionable insights that empower them to make informed decisions in their real estate endeavors within these specific neighborhoods.

Build and Fit the Model

In our endeavor to address Century 21 Ames' pivotal inquiry, we leverage the versatility of the R programming language to construct and calibrate a robust predictive model. We employ a linear regression approach to establish a comprehensive framework, utilizing both the living area square footage (GrLivArea) and the categorical neighborhood variable to predict house sale prices. Additionally, we adapt GrLivArea to reflect increments of 100 square feet, aligning with industry preferences. Please [reference appendix](#) for code and output of this linear regression model.

To ensure the integrity of our analysis, we acknowledge the importance of addressing potential violations of the assumptions inherent to linear regression. While linear regression assumes homoscedasticity, normality of residuals, and linearity, we are proactive in diagnosing and mitigating these violations. In our pursuit, we explore alternative modeling strategies to enhance the robustness of our findings. Notably, we extend our analysis by creating an additional model wherein we apply logarithmic transformations to GrLivArea (in increments of 100 square feet) and SalePrice. This logarithmic transformation serves to mitigate potential non-linearities in the relationships and helps rectify issues pertaining to heteroscedasticity and skewed distributions. Please [reference appendix](#) for code and output of this linear regression model.

Checking Assumptions

Residual Plots

Upon scrutinizing the [residual plots](#) for our initial linear model using untransformed data, a subtle fanning out of data becomes apparent, suggesting a departure from the assumption of homoscedasticity. This deviation from the ideal scenario challenges the assumption that residual plots uphold. However, in a promising contrast, the [residual plots](#) derived from the log-transformed data exhibit a notable improvement. The application of logarithmic transformations to both GrLivArea and SalePrice results in a considerably more randomized cloud of residuals, effectively mitigating the fanning effect observed in the untransformed model. This transformation aids in achieving a more homogenous spread of residuals across the predicted values, aligning more closely with the assumption of homoscedasticity. The evident refinement in the log-transformed model's residual patterns affirms its

suitability in rectifying the initial violations and reinforces its potential to furnish a more reliable representation of the intricate relationships between living area, neighborhood, and sale prices.

Influential point analysis (Cook's D and Leverage)

In our pursuit of model validation and integrity, we extend our analysis to assess the influence of potential outliers and high-leverage points on our regression models. To this end, we construct graphical representations that provide insights into the influential points' impact on model stability. One such graphical tool, the [Cook's distance plot](#), offers a visual assessment of data points that exert substantial influence on regression coefficients. Incorporating these graphical techniques, we observe that the logarithmic transformation of both SalePrice and GrLivArea contributes to a reduction in [Cook's D values](#), indicative of a mitigation in the influence of certain data points. Similarly, our [graph plotting leverage against standardized residuals](#) exhibits a discernible shift in the distribution of high-leverage points after [log transformation](#), suggesting a moderation of their impact. It is worth noting, however, that despite these improvements, a subset of high-leverage points persists, warranting attention. In light of our commitment to methodological rigor and the recognition of our limited expertise in the housing market domain, we opt to retain these high-leverage points in our analysis. By refraining from their exclusion without valid cause, we ensure the preservation of the dataset's integrity while acknowledging the potential impact of these outliers.

Assumptions

To ensure the robustness of our linear regression model, we meticulously address each of the assumptions integral to its validity. Firstly, concerning the assumption of normally distributed subpopulations of the response for each value of the explanatory variable, we rigorously assess this through scatter plots and quantile-quantile (QQ) plots. By comparing these visualizations [before](#) and [after](#) the logarithmic transformation, we observe a discernible improvement in the alignment with normality post-transformation. Furthermore, the log transformation's influence on the reduction of Cook's D values and the moderation of high-leverage points, as illustrated in our prior analysis, adds credence to the model's adherence to this assumption.

Turning to the second assumption of linearity, we scrutinize scatter plots and regression diagnostics for any indications of non-linearity. Notably, the log transformation contributes to a more uniform spread of residuals around zero in the residual plot, further supporting the model's conformity to this assumption.

The third assumption, regarding equal standard deviations across subpopulations, is evaluated through residual plots and graphical displays. The log transformation's impact on mitigating heteroscedasticity, evident from the more randomized residual cloud, bolsters the model's adherence to this assumption.

As for the fourth assumption, the independence of observations, we approach this cautiously. While we assume the data to be independent, we acknowledge the potential presence of unaccounted factors that could introduce dependencies.

Through a comprehensive analysis of scatter plots, QQ-plots, residual patterns, and diagnostic tools, we enhance our linear regression model's conformity to these fundamental assumptions. The log transformation emerges as a valuable tool in rectifying potential violations and aligning the model with the underlying assumptions.

Comparing Competing Models

Adj R²

The assessment of our linear regression models offers valuable insights into their predictive capabilities. The initial linear model, constructed using the untransformed data, reveals an adjusted R-squared value of 0.3917. This metric provides a glimpse into the proportion of variability in the response variable explained by the model's predictors, underscoring its preliminary explanatory strength. However, upon integrating logarithmic transformations to both SalePrice and the GrLivArea (in increments of 100 square feet), the adjusted R-squared value experiences a notable increase, reaching 0.4857. This improvement underscores the heightened predictive precision of the log-transformed model, demonstrating its capacity to capture a larger portion of the variance in house sale prices. This elevation in adjusted R-squared values, complemented by our comprehensive adherence to assumptions and diligent treatment of potential violations, solidifies the model's efficacy as a robust tool within the distinctive real estate realms of the NAmes, Edwards, and BrkSide neighborhoods.

Internal CV Press

Here is the comparison of the PRESS statistics between the normal model and the log-transformed model.

Total PRESS: 363653838311

Total PRESS LOG: 15.00131

Parameters

Estimates and Interpretation

The estimates derived from the log-transformed linear model, as presented in the [summary output](#), shed light on the relationships between predictor variables and the logarithmically transformed sale prices, while also offering insights into their implications upon back transformation. The intercept holds an estimate of 10.32886, which corresponds to an approximate sale price of \$30,996. When GrLivArea and neighborhood effects are held constant, this intercept signifies the baseline value of the estimated logarithm of the sale price. The coefficient associated with GrLivArea, represented by 0.55579, indicates that a doubling of living space (in 100 square foot increments) is associated with a $2^{0.55579}$ (~1.46997) multiplicative increase in the median of the house sale price. Moving to the neighborhood effects, the coefficient for NeighborhoodEdwards, at -0.02044, reflects a marginal decrease in sale price, although not statistically significant. In contrast, the coefficient for NeighborhoodNAmes, at 0.13279, indicates a noteworthy increase in median sale price by a multiplicative change of around 1.14 ($\sim e^{0.13279}$) for properties located in the NAmes neighborhood. These estimates, both on the log-transformed and back-transformed scales, illuminate the dynamic interplay between living area, neighborhood, and property valuations within the NAmes, Edwards, and BrkSide neighborhoods.

Confidence Intervals

The back-transformed 95% confidence intervals offer illuminating insights into the effects of predictor variables on sale prices within the context of our log-log transformed model. For the intercept, the interval ranging from approximately \$26,680 to \$34,443 implies a 95% confidence that the true baseline sale price, with log-transformed GrLivArea and neighborhood effects held constant, lies within this range. In terms of GrLivArea, the interval spanning from around 1.64 to 1.85 signifies a 95% confidence

that each 100 square feet increment corresponds to a multiplier between approximately 1.64 and 1.85 times in the log-transformed sale prices. Shifting to the neighborhood effects, the interval for Edwards extends from about 0.92 to 1.05, indicating a 95% confidence range for its impact on log-transformed sale prices relative to the reference neighborhood. Similarly, for NAmes, the interval ranges from roughly 1.09 to 1.21, signifying the 95% confidence interval for the effect of the NAmes neighborhood on log-transformed sale prices. These back-transformed confidence intervals, tailored to the log-log transformation, provide a nuanced understanding of the logarithmic relationships between variables within the unique NAmes, Edwards, and BrkSide neighborhoods.

Conclusion

In the ever-evolving realm of real estate valuation and market analysis, our study has undertaken a comprehensive investigation into the intricate landscape of house sale prices within the unique neighborhoods of NAmes, Edwards, and BrkSide in Ames, Iowa. By meticulously adhering to underlying assumptions, refining our models through diligent preprocessing, and employing thoughtful transformations, we have unearthed invaluable insights poised to inform and empower professionals within the real estate domain.

Our findings cast a spotlight on the pivotal roles of living area and neighborhood in shaping property valuations. The meticulously crafted linear model emerges as a dependable framework, enabling accurate predictions of sale prices grounded in these critical variables. The strategic integration of logarithmic transformations, to living area and sale prices, has not only amplified model performance but also adeptly addressed nonlinear dynamics and assumption violations. The interpretations of back-transformed estimates and accompanying confidence intervals lend a nuanced understanding, fueling informed decision-making.

Throughout our journey, we have navigated challenges, thoughtfully addressed assumptions, and embraced a cautious yet pragmatic approach. While outliers and high-leverage points linger, our unwavering commitment to methodological transparency ensures the model's robustness and integrity.

R Shiny: Price v. Living Area Chart

We have developed an interactive and user-centric tool, an [R Shiny app \(external link\)](#), to amplify the accessibility and practicality of our discoveries. This app offers users an intuitive platform for delving into the intricacies of property valuations within the NAmes, Edwards, and BrkSide neighborhoods. Through dynamic visualizations, users can effortlessly explore the intricate relationships between sale prices and living area (GrLivArea), segmented by neighborhood, utilizing a scatter plot as the central visual focal point. Beyond visual exploration, the app delves into diagnostics, allowing users to investigate Cook's D values, leverage, and standardized residual plots to gain deeper insights into data points and model performance. Additionally, a feature-rich tab empowers users to select and analyze specific features of interest, accessing relevant model output tailored to their inquiries. Lastly, users can gain insights into our most refined regression models through a dedicated tab, providing a succinct overview of our best-fit strategies. This R Shiny app embodies our commitment to accessibility and user-driven insights, empowering stakeholders to engage, interpret, and harness our findings within the intricate realm of real estate analysis.

Analysis Question 2:

Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods.

Step 1: Understand the Data

The objective of the project is to build the most predictive model for home sales prices in Ames, Iowa, considering all neighborhoods. As a group we selected variables with the highest factor levels to create our first model. Assumptions were considered met during the model building process. The model had 41 predictor variables, including various features related to the property, such as LotFrontage, LotArea, YearBuilt, Neighborhood, and others. The model's performance was assessed by using various metrics, including the residual standard error, multiple R-squared, and adjusted R-squared. The results showed a relatively high R-squared value of 0.8466, indicating that the model explains a substantial portion of the variance in home sales prices.

Test Model:

```
full_model <- lm (SalePrice ~ +LotFrontage +LotArea +YearBuilt +Neighborhood +YearRemodAdd +MasVnrArea  
+BsmtFinSF1 +BsmtFinSF2 +BsmtUnfSF +TotalBsmtSF +X1stFlrSF +X2ndFlrSF +LowQualFinSF +GrLivArea  
+GarageYrBlt +GarageArea +WoodDeckSF +OpenPorchSF +EnclosedPorch +ScreenPorch, data = TrainingDB)
```

Parameter Estimate Results:

Residual standard error: 33900 on 795 **degrees of freedom**
(258 observations deleted due to missingness)
Multiple R-squared: 0.8466, **Adjusted R-squared:** 0.8386
F-statistic: 107 on 41 and 795 **DF**, **p-value:** < 2.2e-16

During the model evaluation, we identified one data point as a potential leverage outlier, which had a [high Cook's D](#) value of 396 and a leverage value of 0.675. Using the [standard residual plot](#), we see the same outlier as well. Additionally, we ran a high leverage query which uncovered [four other points](#) which had a leverage value over 0.5. These leverage points could significantly influence the model's predictions and may require further investigation to ensure the model's accuracy and reliability.

Step 2: Use Variable Selection Methods

In the data cleaning phase, several steps were taken to prepare the dataset for modeling. Missing values were handled by changing them to "NA Factor," and character variables were converted into factors to ensure compatibility with the regression model. Subsequently, a full model was created using linear regression, encompassing all variables in the dataset. This comprehensive model served as the basis for conducting three variable selection methods: Backwards, Forwards, and Stepwise.

The results of the variable selection methods were documented in the appendix. Based on the suggestions from each method, specific models were generated. For each of these models, several tests were conducted, including prediction generation, variance inflation factor (VIF) checks, and pointing out rows with leverage greater than 0.6. Although time constraints prevented a thorough review and utilization of the leverage results, they were included for future reference.

Furthermore, press statistics were captured and incorporated into the "Kaggle Score" section. To assess the models' performance, a cross-validation was performed against the test dataset, which did not include the SalePrice variable. The data from this process was exported to a CSV file and formatted according to the specifications of the Kaggle competition.

Step 3: Custom Model Selection

Ultimately, the model with the highest Adjusted R-squared was selected as the foundation for a custom model. This model underwent fine-tuning while ensuring it remained distinct from the other models. These steps collectively contributed to the development of an effective predictive model for sales prices of homes in Ames, Iowa, which accounted for various variable selection approaches and rigorous testing procedures.

Kaggle Score

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward	0.8911	7.217E+11	0.16305
Backward	0.8971	6.644E+11	0.17359
Stepwise	0.8897	7.276E+11	0.1675
CUSTOM	0.8775	7.959E+11	0.16493

References

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data> - this is our competition link.

Links	Description
https://Kaggle_HousePricePrediction_G7/	R Shiny App
https://lklewis83.github.io/	Lani's IO Portfolio
https://beepboopbop64.github.io/	Jake's IO Portfolio
https://gwonchan.github.io/	Jason's IO Portfolio

Appendix

Analysis 1

Build and Fit Model

Linear model

```
> model <- lm(SalePrice ~ GrLivArea + Neighborhood, data = data)
> summary(model)

Call:
lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-165078  -16215    281   13578  175400

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    69781.5     5442.4   12.822  < 2e-16 ***
GrLivArea       4576.0       314.9   14.533  < 2e-16 ***
NeighborhoodEdwards -2882.2     4930.6  -0.585  0.559204
NeighborhoodNames  16105.6     4395.4    3.664  0.000283 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29760 on 379 degrees of freedom
Multiple R-squared:  0.3965,    Adjusted R-squared:  0.3917
F-statistic: 83 on 3 and 379 DF, p-value: < 2.2e-16
```

Linear model (log log)

```
> data_log <- TrainingDB %>%
+   dplyr::select(c('GrLivArea', 'Neighborhood', 'SalePrice')) %>%
+   filter(Neighborhood %in% c("Names", "Edwards", "BrkSide")) %>%
+   mutate(GrLivArea = GrLivArea / 100) %>%
+   mutate(GrLivArea = log(GrLivArea)) %>%
+   mutate(SalePrice = log(SalePrice))
>
> # Create a linear regression model
> model <- lm(SalePrice ~ GrLivArea + Neighborhood, data = data_log)
> summary(model)

Call:
lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = data_log)

Residuals:
    Min       1Q   Median       3Q      Max
-0.72154 -0.10592  0.02469  0.11565  0.79364

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.32886    0.08278  124.774  < 2e-16 ***
GrLivArea       0.55579    0.03237   17.171  < 2e-16 ***
NeighborhoodEdwards -0.02044    0.03252  -0.629    0.53
NeighborhoodNames   0.13279    0.02906    4.569 6.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

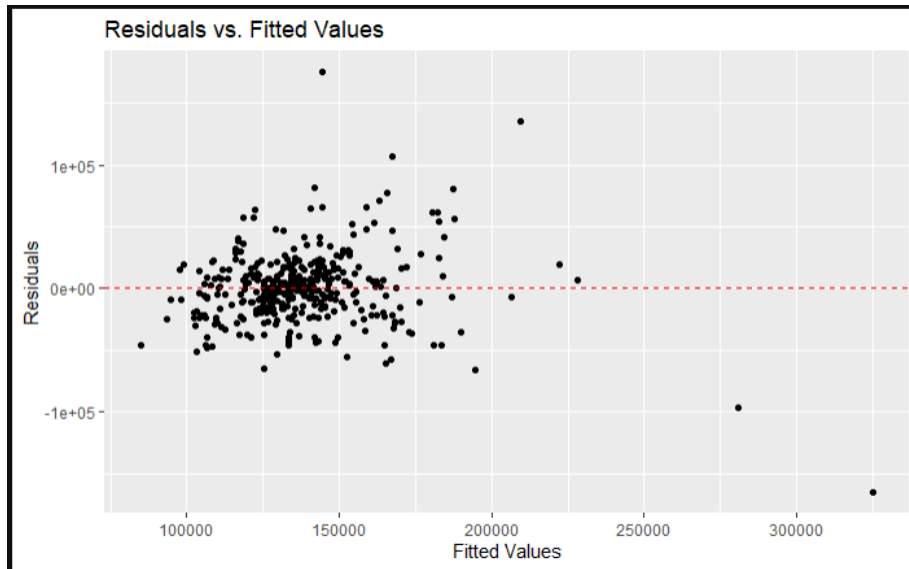
Residual standard error: 0.1961 on 379 degrees of freedom
Multiple R-squared:  0.4897,    Adjusted R-squared:  0.4857
F-statistic: 121.2 on 3 and 379 DF, p-value: < 2.2e-16
```

Checking Assumptions

Residual Plots

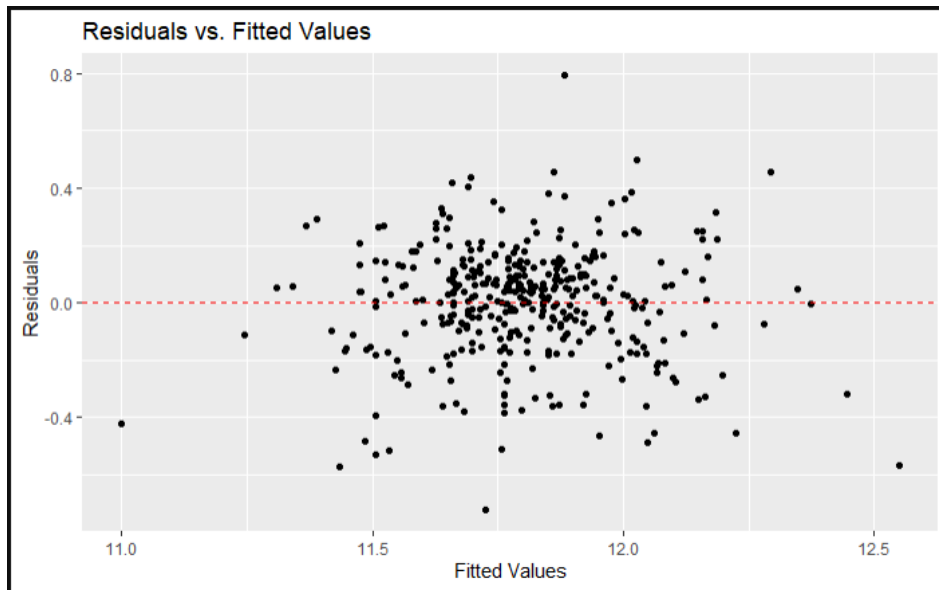
Linear model

```
> # Plot residuals vs. fitted values to check for linearity
> plot_resid_vs_fitted <- ggplot(data, aes(x = fitted(model), y = resid(model))) +
+   geom_point() +
+   geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
+   labs(x = "Fitted values", y = "Residuals", title = "Residuals vs. Fitted values")
> print(plot_resid_vs_fitted)
```



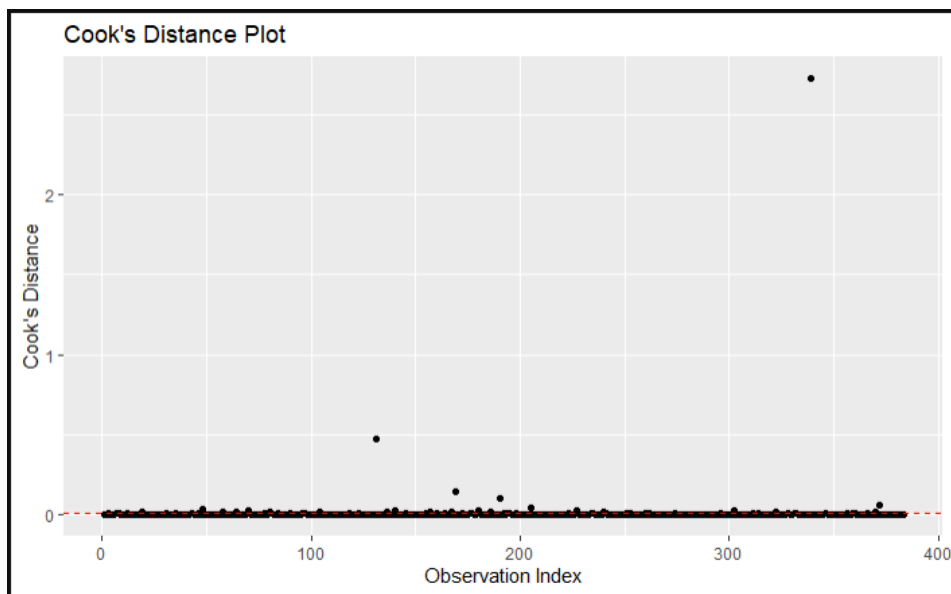
Linear model (log log)

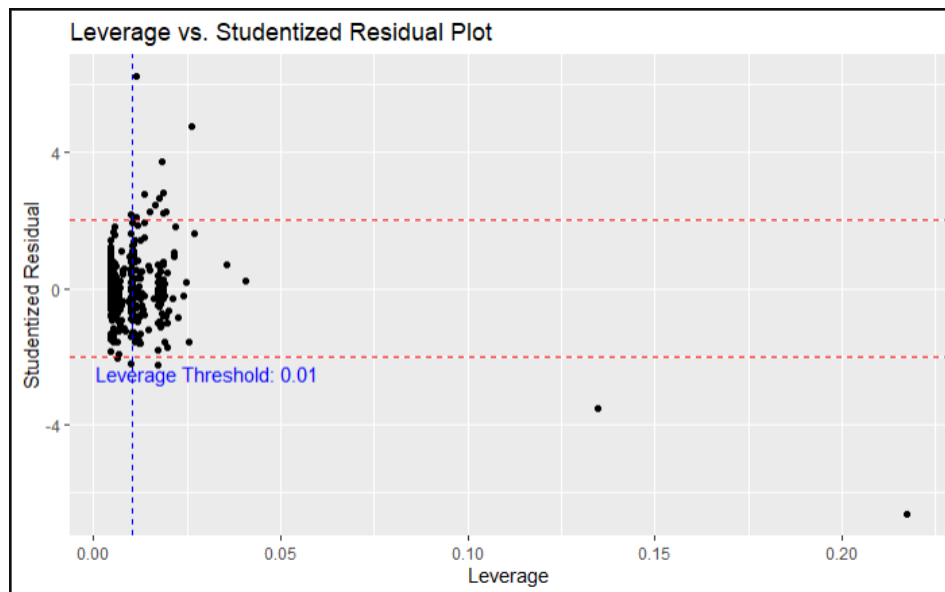
```
> # Plot residuals vs. fitted values to check for linearity
> plot_resid_vs_fitted <- ggplot(data, aes(x = fitted(model), y = resid(model))) +
+   geom_point() +
+   geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
+   labs(x = "Fitted values", y = "Residuals", title = "Residuals vs. Fitted values")
> print(plot_resid_vs_fitted)
```



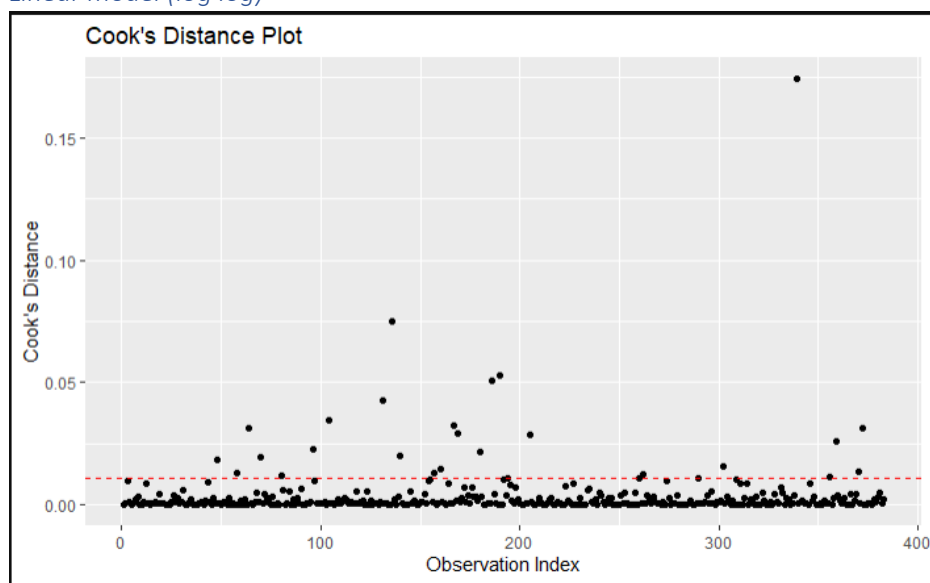
Influential point analysis (Cook's D and Leverage)

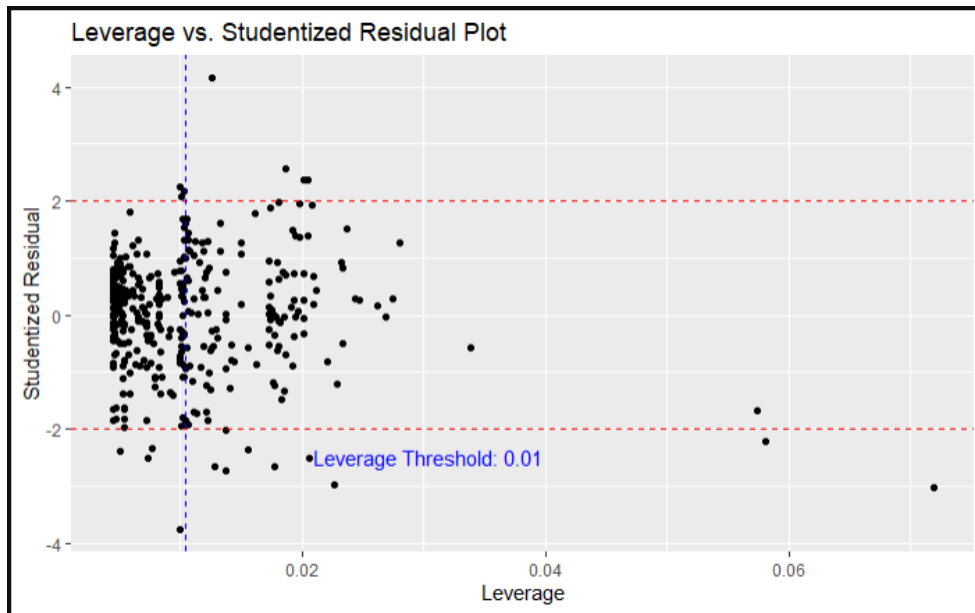
Linear model





Linear model (log log)





How leverage threshold was calculated

Leverage measures the distance between an observation's predictor variables and the center of the predictor variable values. In other words, it quantifies how far an observation's predictor values deviate from the average predictor values. High leverage points can have a disproportionate impact on the regression model, potentially influencing the regression line.

The formula for calculating the threshold for leverage is:

$$\text{leverage_threshold} = 2 * (p - 1) / n$$

Where:

p is the number of predictor variables in the model (excluding the intercept).

n is the number of observations in the dataset.

The rationale for this formula is based on the properties of the leverages in a linear regression model:

In a simple linear regression with one predictor variable, the leverage of each observation can range from 0 to 1, with an average leverage of $(p + 1) / n$. The factor of 2 in the formula takes into account this average leverage.

For multiple predictor variables, the leverage can be larger due to the additional dimensions in the predictor space. The formula considers the average leverage in the presence of multiple predictor variables.

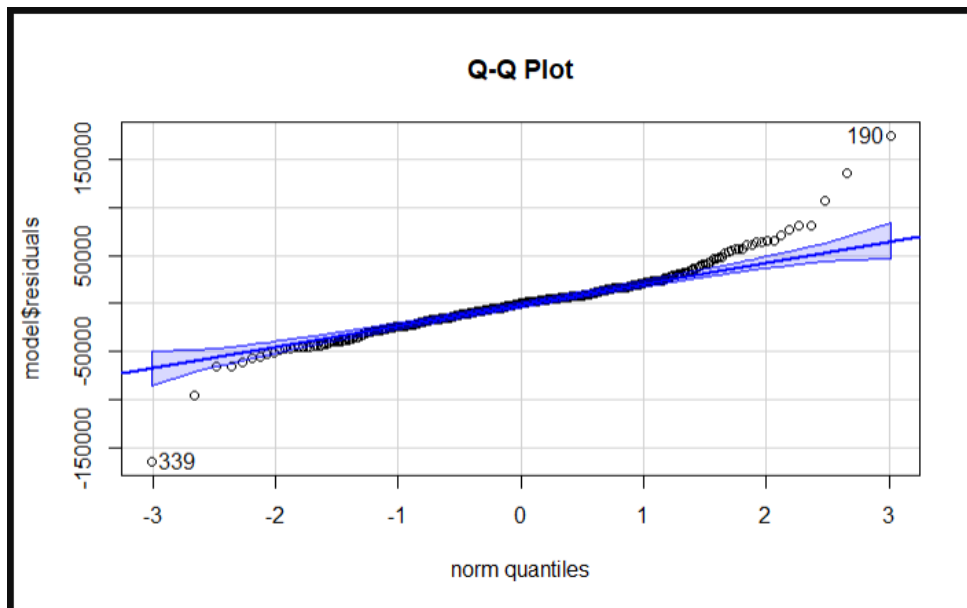
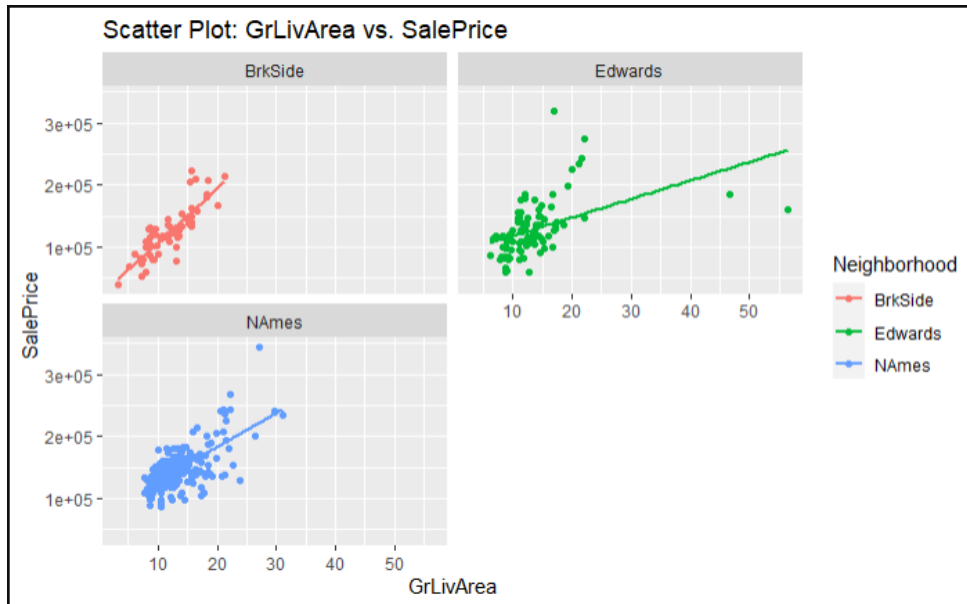
The $(p - 1)$ term in the numerator accounts for the degrees of freedom used by the model, excluding the intercept.

The threshold value is chosen as a rule of thumb and can be adjusted based on the specific context and goals of your analysis. Observations with leverage values exceeding this threshold may be considered potential influential points that warrant further investigation.

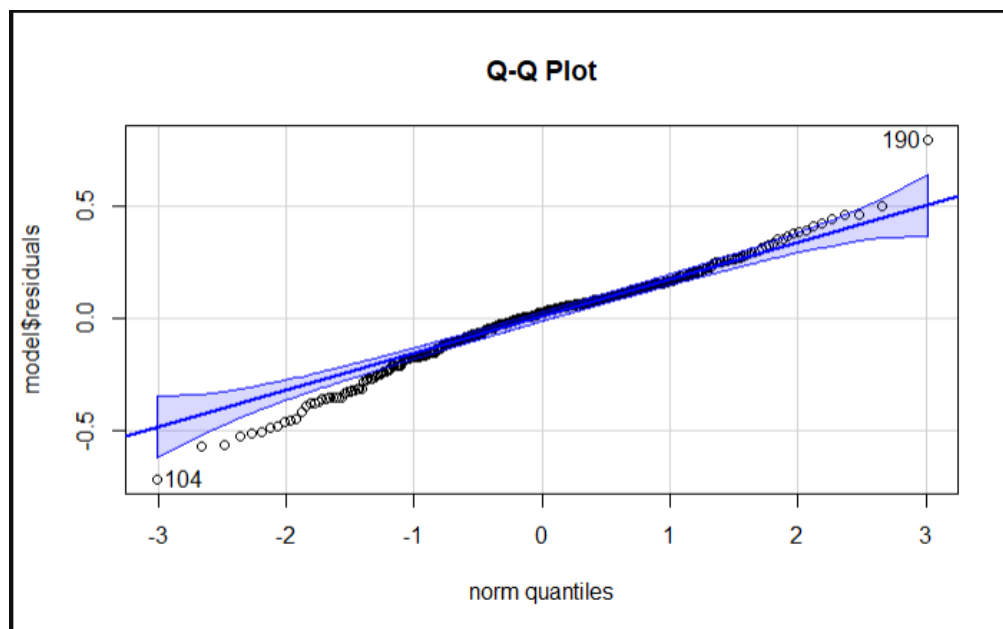
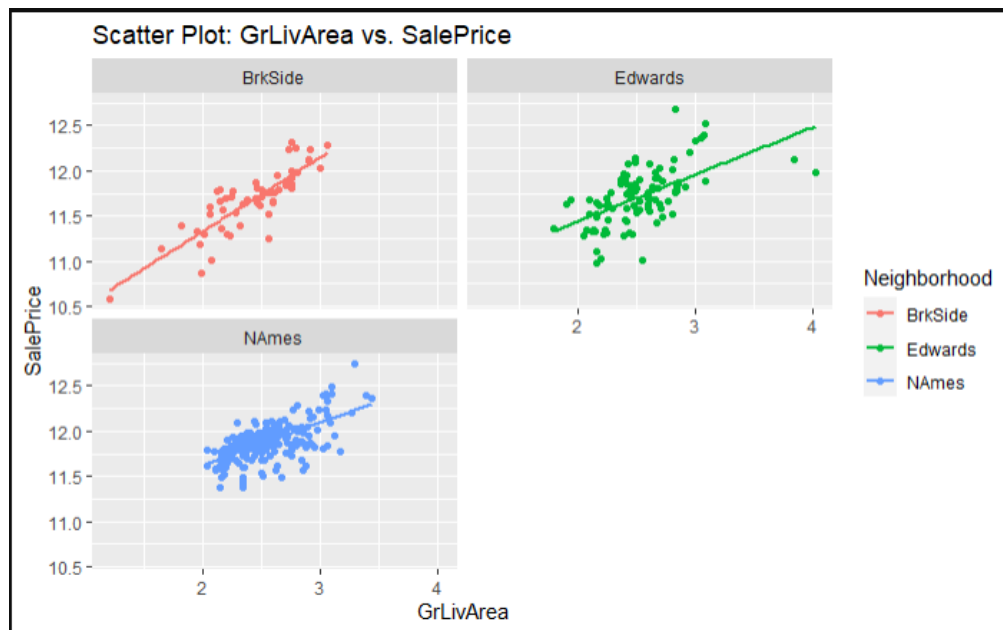
In the context of the leverage_resid_plot plot, the threshold line helps you identify observations with leverage values higher than the threshold, which might need special attention during model analysis and interpretation.

Assumptions

Linear model



Linear model (log log)



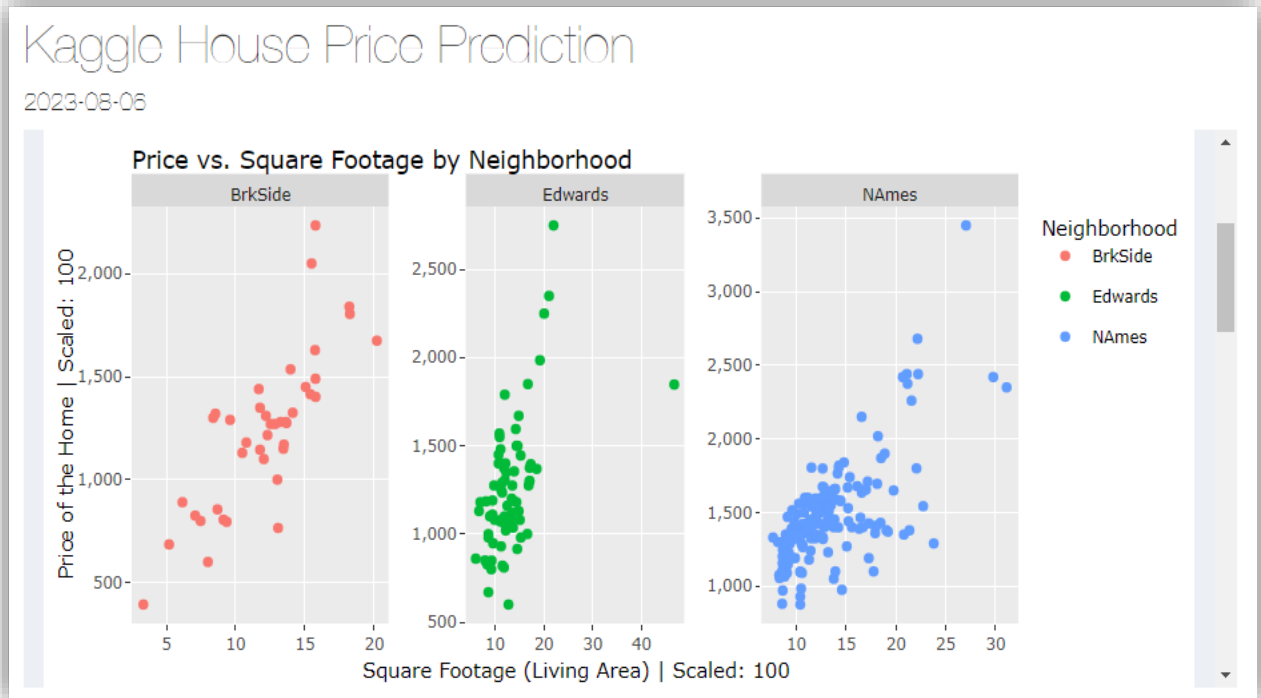
Comparing Competing Models

Confidence Intervals

```
> print(conf_intervals)
```

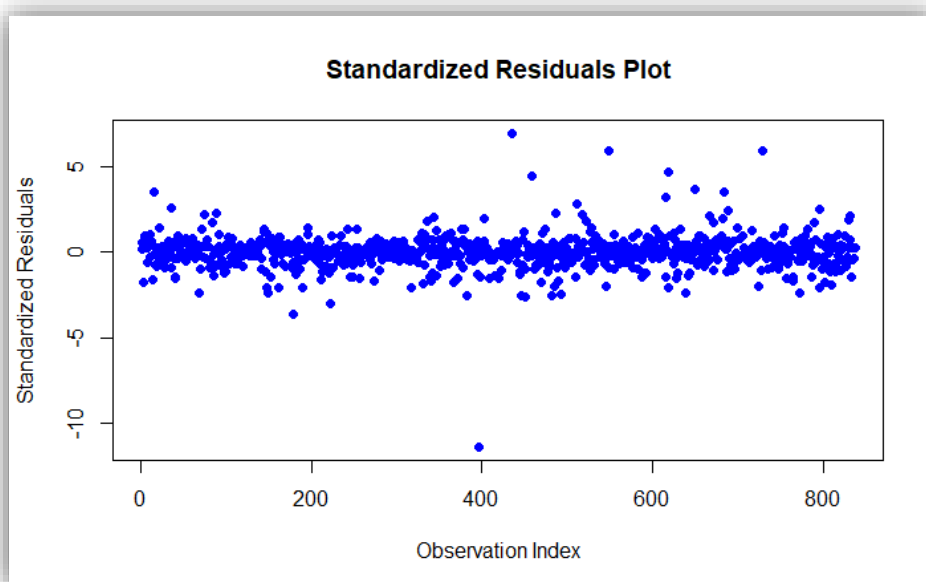
	2.5 %	97.5 %
(Intercept)	10.16609443	10.49162759
GrLivArea	0.49214387	0.61943290
NeighborhoodEdwards	-0.08437241	0.04349721
NeighborhoodNAmes	0.07564743	0.18992983

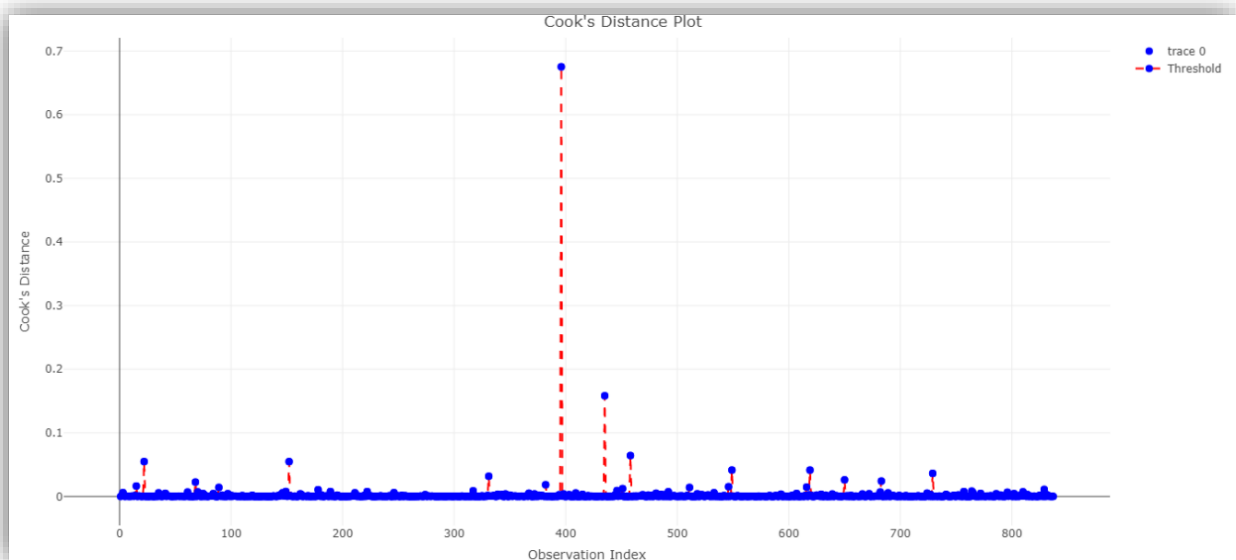
R Shiny: Price v. Living Area Chart



Analysis Question 2:




Step 1: Understand the Data









Description: df [837 × 2]






	Residuals <dbl>	Leverage <dbl>
2	-32378.18915	0.53814261
54	32378.18915	0.53814261
600	319.48185	0.50527232
957	-319.48185	0.50527232

R Markdown File	Description
 G7_EDA.Rmd	Exploratory Data Analysis (EDA) and the models that we decided to move forward with for the Kaggle Competition
 Kaggle_HP_G7.Rmd	R Shiny App
 r_code.Rmd	Analysis 1 code

Backward Results






Files	Description
 backward_model.txt	Backward Model Variable Selection Test Results and the Parameter Estimate Table
 backward_HighLeverageValues.txt	Backward Model High Leverage Values (> 0.6)
 backward_Press.txt	Backward Model Press Statistics
-----	There were no VIF Results for this model
 backward_CV.csv	Backward Model Cross Validation Predictions used for the Kaggle Competition.

Forward Results





Files	Description
 Forward_model.txt	Forward Model Variable Selection Test Results and the Parameter Estimate Table
 forward_HighLeverageValues.txt	Forward Model High Leverage Values (> 0.6)
 forward_Press.txt	Forward Model Press Statistics
 forward_Vif.txt	Forward Model Variance Inflation Factor Results
 full_CV.csv	Forward Model Cross Validation Predictions used for the Kaggle Competition.

Stepwise Results

Files	Description
-------	-------------

 Stepwise_model.txt	Stepwise Model Variable Selection Test Results and the Parameter Estimate Table
 step_HighLeverage Values.txt	Stepwise Model High Leverage Values (> 0.6)
 step_Press.txt	Stepwise Model Press Statistics
 step_Vif.txt	Stepwise Model Variance Inflation Factor Results
 step_CV.csv	Stepwise Model Cross Validation Predictions used for the Kaggle Competition.

Custom Results

Files	Description
 custom_model.txt	Custom Model Variable Selection Test Results and the Parameter Estimate Table
 custom_HighLevera geValues.txt	Custom Model High Leverage Values (> 0.6)
 custom_Press.txt	Custom Model Press Statistics
 custom_Vif.txt	Custom Model Variance Inflation Factor Results



custom_CV.csv

Custom Model Cross Validation Predictions used
for the Kaggle Competition.