

Final Project Part 4:

Final Report

Peter Brown, Danielle Dane, Trang Tran

College of Professional Studies, Northeastern University Roux Institute

ALY6015

Professor Steven Morin

May 15, 2023

Real estate is an ever-changing, fast-paced industry. As new real estate investors in the Ames, Iowa community in 2011, we have limited market experience in this region and need help familiarizing ourselves with the local market. Our objective is to utilize the Ames Housing data set of recent home purchases in Ames from 2006-2010 to give us guidance for which properties to purchase or sell (De Cock 2011). This analysis will look at a smaller aspect each of three major themes in real estate: 1- Features; 2- Location; and 3- Size.

- 1- **Features-** we will look at if the varying types of basements impact the sale price of a property using the Kruskal-Wallis test.
- 2- **Location-** we will look at if there is a meaningful difference in the average sale price of a property in 3 target neighborhoods using the Kruskal-Wallis test.
- 3- **Size-** we will look at how accurately we can predict the sale price of a property with only predictor variables of size (like square footage and room counts). We will use LASSO and Ridge regressions to compare the best fitting model.

Exploratory Data Analysis & Data Cleaning

The Ames Housing data set contains 82 variables and 2,930 observations. Each observation is a residential property, and each variable is a feature or measurement of interest for a property. In the initial data cleaning, using *skim()* we identified the type of variables, descriptive stats on each variable, and removed variables with >20% missing values as well as the identifier column. We also dropped all character variables not used in the analysis, and finally for the remaining numeric variables that had missing data, we imputed their mean.

Neighborhood and *Bsmt.Exposure* were changed to factors. This left us with a clean data set shown in Figure 1 of 44 variables.

Figure 1: Skim
summary of
cleaned data set

-- Data Summary -----	
Name	values
Number of rows	df
Number of columns	2930
	43
Column type frequency:	
character	5
factor	2
numeric	36
Group variables	
	None

Features: Basements & Sale Prices

This part of our analysis will look into if the features of a basement significantly impact the price of a property. The method used for this question is a Kruskal-Wallis test.

The Dependent variable is 'SalePrice'. The Independent variable is 'Bsmt.Exposure' (Ordinal). Basement Exposure refers to walkout or garden level walls in a basement. The classes are as follows, and the counts of each class are displayed in figure 2.

- Gd = Good Exposure
- Av = Average Exposure (split levels or foyers typically score average or above)
- Mn = Minimum Exposure
- No = No Exposure
- NA = No Basement

Figure 2: Skim
summary of
cleaned data set
showing each class

```
> df %>% group_by(Bsmt.Exposure) %>% summarise(count = n())
# A tibble: 6 x 2
  Bsmt.Exposure count
  <chr>         <int>
1 ""              4
2 "Av"           418
3 "Gd"           284
4 "Mn"           239
5 "No"          1906
6 "NA"            79
```

There are total 79 'NA' values in this variable, so we replace 'NA' with 'No.Bsmt' and we dropped the 4 observations with null values for the analysis. After that, we filter the dataframe with only two variables: 'Bsmt.Exposure' and 'SalePrice'.

First, we chose to use the one-way ANOVA test for this analysis. The ANOVA test assumes that the data is normally distributed, the variances are equal across groups, and the observations are independent. However, the assumption of normality is not met, specifically when the "SalePrice" variable is not normally distributed per each group, the Kruskal-Wallis test may be a more appropriate choice. The figure 3 below is one of the results of the Shapiro-Wilk normality test:

Figure 3: Skim
summary of
cleaned data set
showing each class

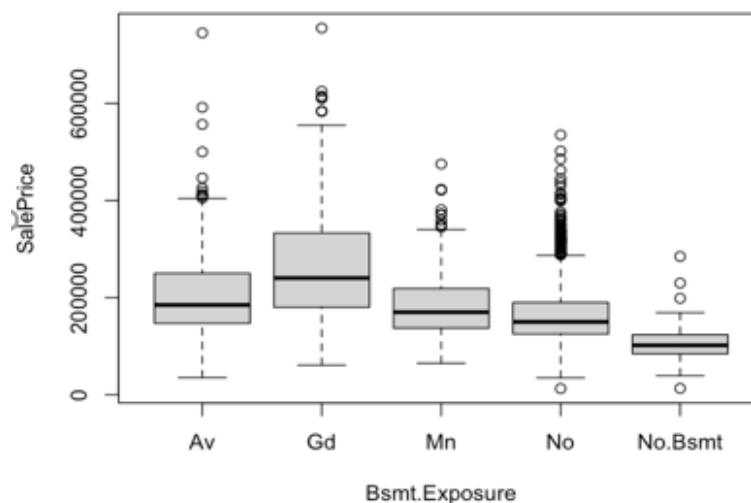
```
[1] "Gd"

      Shapiro-Wilk normality test

data:  df_bsmt[df_bsmt$Bsmt.Exposure == Bsmt.Exposure, "SalePrice"]
W = 0.92492, p-value = 0.00000000009051
```

The p-value in the "Good Exposure Basement" group is much lower than the significant level (0.05) that is also similar to all other groups.

Figure 5: Boxplot of Basement exposure and Sale price



Location: Neighborhoods & Sale Prices

The second part of our analysis explores the location of a home in Ames, compared to the sale price. As investors, we want to know which neighborhoods in Ames consistently sell for the highest amounts, so that we can focus on purchasing properties in the best neighborhoods where resale values will be the highest and take advantage of good deals when we see them. To start we calculated the mean sale price for each neighborhood to get a sense of which neighborhoods had the highest average selling prices. This analysis showed three neighborhoods that were reliably averaging above \$300K: Northridge, Northridge Heights, and Stone Brook. We wanted to know whether any difference in mean sales prices between these neighborhoods was meaningfully different. We chose to use ANOVA to investigate this question, because ANOVA tests the difference of two or more means in the relationship between a categorical and a numerical variable. In our case, we would perform a one-way ANOVA tests to compare the mean sale prices (numeric variable) of these three highest-selling neighborhoods (categorical variable).

After creating a dataframe for the three \$300K neighborhoods (Northridge, Northridge Heights, and Stone Brook), we checked to see if the dataset met the assumptions necessary for ANOVA:

- The samples must be simple random samples, one from each population. ☒
- The samples must be independent of one another. ☒
- The variances of the populations must be equal.
 - The ratio between the largest (119273.02) and smallest (95932.35) standard deviations was calculated. The quotient (1.243) fell between 0.5 and 2.0, so the assumption of homogeneity of variance was met. ☒
- The populations from which the samples were obtained must be normally or approximately normally distributed.
 - Shapiro-Wilks tests were performed on the salesprice for each neighborhood. (Figure 6) None of the neighborhoods in the \$300K dataframe had p-values above 0.05, which are necessary to establish normal distribution. ☒

Figure 6: Shapiro-Wilkes Test Results

```
*****
Check one-way annova assumptions:
*****
Test for normality - Shapiro-wilk normality test (H0: normal)

[1] "Northridge"

      Shapiro-wilk normality test

data:  nghbhd_300k[nghbhd_300k$Neighborhood == nh, "salePrice"]
W = 0.72739, p-value = 3.573e-10

[1] "Northridge_Hts"

      Shapiro-wilk normality test

data:  nghbhd_300k[nghbhd_300k$Neighborhood == nh, "salePrice"]
W = 0.9626, p-value = 0.0001926

[1] "Stone_Brook"

      Shapiro-wilk normality test

data:  nghbhd_300k[nghbhd_300k$Neighborhood == nh, "salePrice"]
W = 0.9492, p-value = 0.02918
```

Unable to perform the one-way ANOVA due to a failure to met the assumption of normal distribution, we ran a Kruskal-Wallis test, which is a non-parametric alternative to the one-way ANOVA test. Results from the Kruskal-Wallis test can be found in Figure 7.

Figure 7:
Kruskal-Wallis
test results

```
> p_value <- results$p.value
> print(p_value)
[1] 0.9563666
>
> writeLines('\n*****')
*****
> # compare the p-value to alpha and make decision
> if (p_value > alpha) {
+   decision = 'fail to reject H0'
+ } else {
+   decision = 'reject H0'
+ }
> writeLines(paste('decision: ', decision))
decision: fail to reject H0

alpha = 0.05

# hypotheses
# H0: mu1 == mu2 == mu2
# H1: one or more mu different
# claim is H1

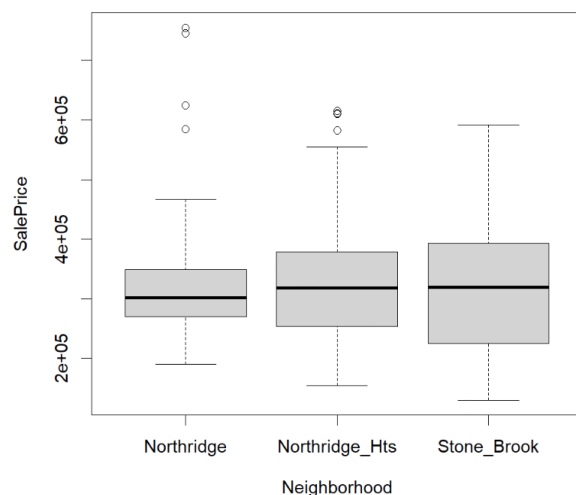
# kruskal.test
results <- kruskal.test(list('a' = df_300K$NoRidge,
                             'b' = df_300K$NridgHt,
                             'c' = df_300K$StoneBr))

test_statistic <- results$statistic
print(test_statistic)

kruskal-wallis chi-squared
0.08922787
```

The Kruskal-Wallis test calculated a p-value of 0.956. Because this p-value is greater than alpha (0.05), we failed to reject the null hypothesis that the mean sale price of the neighborhoods is the same. There is not enough evidence to support the claim of the alternative hypothesis. Our analysis showed no statistically significant difference in the mean sale price of properties in Northridge, Northridge Heights and Stone Brook. While Figure 8 shows there is a difference in range of prices in each neighborhood, the means are not significantly different enough from each other to say one neighborhood is higher-priced than another.

Figure 8: Boxplot of Sale prices for
the \$300k Neighborhoods



Size: Predicting Sale Price by Square Footage & Room Counts

The third part of our analysis looks at how accurately we can predict the sale price of a house solely on the variables that indicate the size of a property, such as square footage and number of rooms. This can be a valuable analysis because most assessor offices do not publicly publish detailed quality-ranking variables of an off-market home, only measureable square footage or room counts. We want to build a model that uses these simple square footage and room count variables that indicate the size of a home, and relate that to a predicted sale price.

We will evaluate two model methods, LASSO and Ridge, and will decide which provides the best fitting model to predict sale price using only the square footage and room count predictor variables in the data set. Figure 9 displays the variables being used. We will evaluate each value of lambda (λ) from each method: λ_{\min} and λ_{1se} to see which option provides the best fitting model for each method and will use 10 folds for testing each model.

Figure 9: 27 variables that refer to the size of a house are used in the LASSO and Ridge regressions. SalePrice is the dependent variable.

Columns: 27	\$ Full.Bath
\$ Lot.Frontage	\$ Half.Bath
\$ Lot.Area	\$ Bedroom.AbvGr
\$ Mas.Vnr.Area	\$ Kitchen.AbvGr
\$ BsmtFin.SF.1	\$ TotRms.AbvGrd
\$ BsmtFin.SF.2	\$ Garage.Cars
\$ Bsmt.Unf.SF	\$ Garage.Area
\$ Total.Bsmt.SF	\$ Wood.Deck.SF
\$ X1st.Flr.SF	\$ Open.Porch.SF
\$ X2nd.Flr.SF	\$ Enclosed.Porch
\$ Low.Qual.Fin.SF	\$ X3Ssn.Porch
\$ Gr.Liv.Area	\$ Screen.Porch
\$ Bsmt.Full.Bath	\$ Pool.Area
\$ Bsmt.Half.Bath	\$ log_SalePrice

Using the selected variables, there are 523 observations with NA values. Because none of these variables had 20% or more missing and were numeric, we imputed the mean for each NA value, as done in the initial data cleaning stage.

Next, we looked at the distribution of *SalePrice*, our dependent variable. The data looks slightly normally distributed, but it has a long right tail, and does not pass the Shapiro-Wilkes

test of normality. The null hypothesis of the Shapiro-Wilks is that the data is normally distributed. We rejected the null hypothesis because the p-value was less than 0.05. Therefore, we decided to try transforming the data to get to normality. We tested log, square root and cube root transformations (Zach, 2020). The log transformation (with a base of 10) performed the best and had the highest W statistic at 0.986, however it still did not pass the Shapiro-Wilks test for normality, but visually it looks more normally distributed than the original. Figure 9 compares the Shapiro-Wilks tests of *SalePrice* vs. *log_SalePrice*. While P is still <0.05 , the W statistic in the *log_SalePrice* improved to 0.98. The closer to 1, the closer the distribution is to normal (King, 2019). Figure 11 visually compares the *SalePrice* variable before and after the log transformation; it is a decent improvement. We replaced *SalePrice* with *log_SalePrice* with a base of 10, as our transformed dependent variable into our data frame.

Figure 10: Shapiro-Wilk tests of *SalePrice* before and after log transformation.

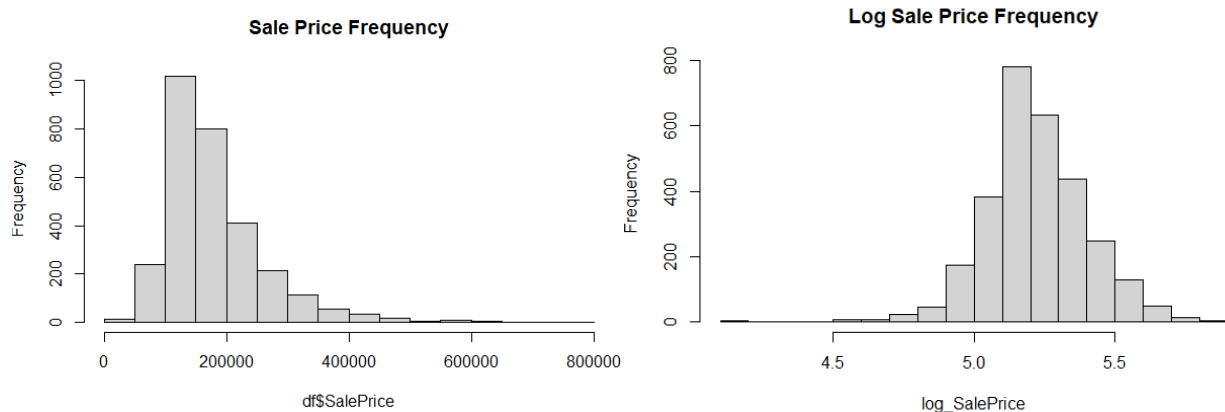
Before transformation:

```
Shapiro-wilk normality test
data:  df$SalePrice
W = 0.87626, p-value < 0.00000000000000022
```

After transformation:

```
Shapiro-wilk normality test
data:  log_SalePrice
W = 0.98579, p-value < 0.00000000000000022
```

Figure 11: *SalePrice* frequency distribution before and after log transformation.



Finally, we split the data randomly, 70% train set and 30% test set and put aside the test set.

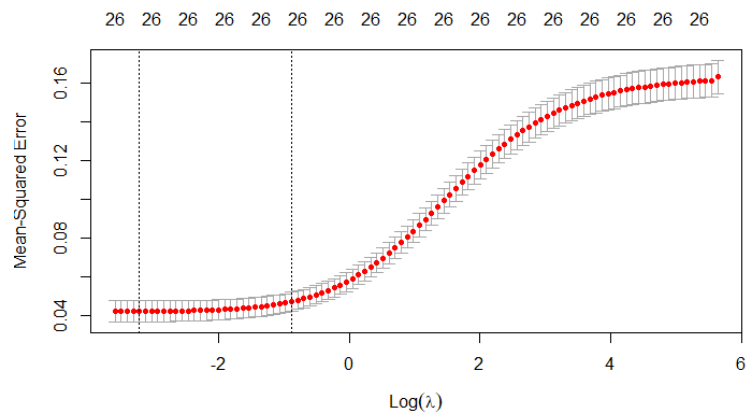
Ridge Regression:

The Ridge regression is not a variable selection tool, so we used all 27 independent variables in the Ridge attempts. Using the train set, the λ_{1se} and λ_{min} values, their log values, and RMSE values are displayed in Figure 12. Figure 13 shows the CV plot.

Figure 12: Test set Ridge model λ_{1se} and λ_{min} values and their corresponding log & RMSE values

Ridge Train Set	λ	$\text{Log}(\lambda)$	RMSE of $\log_SalePrice$
λ_{1se}	0.415	-0.88	0.21415
λ_{min}	0.045	-3.20	0.19824

Figure 13: CV plot for Ridge train set.



Next we ran the Ridge model against the test set. Figure 14 displays the RMSE values for the test set. The RMSE's between the train and test sets are close, indicating there is no over fitting. Now we will compare these results later to the LASSO model.

Figure 14: Ridge test set RMSE values

Ridge Test Set	RMSE of $\log_SalePrice$
λ_{1se}	0.21102
λ_{min}	0.19634

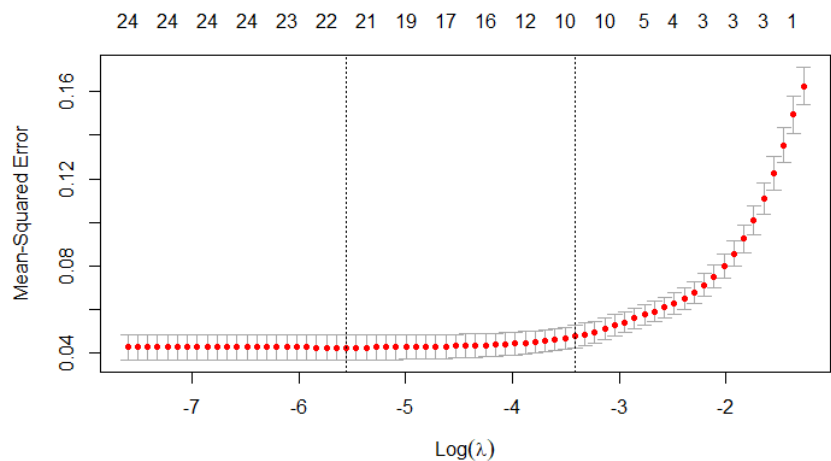
LASSO Regression:

The Lasso regression is a variable selection tool, so it narrows down the set of predictor variables to utilize in the model. For LASSO, using the train set, the λ_{1se} and λ_{min} , their log values and RMSE values are displayed in Figure 15 and the cv plot in Figure 16. As compared to the Ridge models, the LASSO models are less complex because they have less variables, and are more flexible because they have lower $\log \lambda$ values. However, both Ridge and LASSO models are known as low flexibility model types because they are linear.

Figure 15: Train set LASSO λ_{1se} and λ_{min} values and their corresponding log & RMSE values

LASSO Train Set	λ	Log	RMSE of $\log_SalePrice$	Non-zero coefficients
λ_{1se}	0.033	-3.4	0.21443	10
λ_{min}	0.004	-5.6	0.19823	22

Figure 16: CV plot for LASSO train set.



Next, we ran the same model against the test set. Figure 17 displays the RMSE values for the test set. Again, like Ridge there is no significant difference between the train and test RMSEs, so the models are not over fit.

Figure 17: LASSO test set
RMSE values

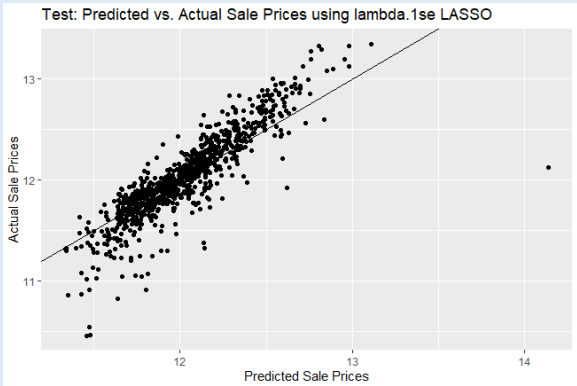
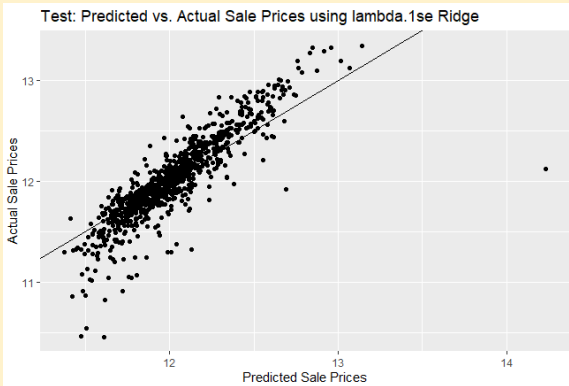
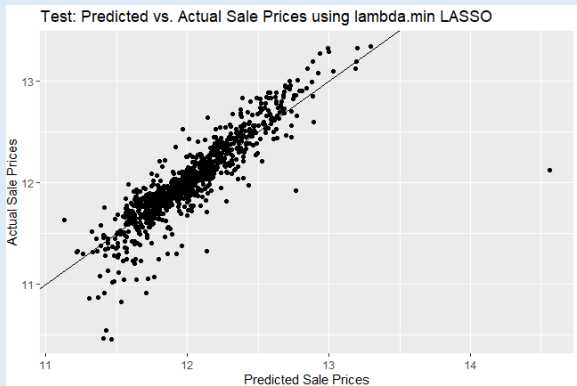
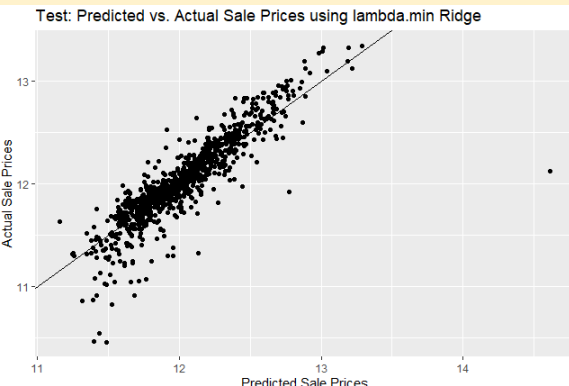
LASSO Test Set	RMSE of <i>log_SalePrice</i>
λ_{1se}	0.20859
λ_{min}	0.19533

Summarized in figure 18, next page, is a side-by-side comparison of the models predicting sale price using square footage/area and room count predictor variables. It also shows visual plots on how the regression line fits among the predicted vs. actual values of LASSO and Ridge models' test set.

All four models visually have a fairly good fit along the regression line and one does not appear visually worse than another. The RMSEs between the LASSO and Ridge models are not significantly different. Technically LASSO's λ_{min} has the lowest test RMSE of all 4 models, but because LASSO λ_{1se} RMSE is not much higher and has only 10 coefficients instead of 22, LASSO λ_{1se} is the best fit model due to parsimony.

The LASSO λ_{1se} model essentially proves that we can predict the *log_SalePrice* of a home using the 10 selected variables and the predicted log of sale price could have up to a 0.21 difference from the actual sale price. Considering the average log sale price of a home in the Ames data set is 12.0, the error of +/- 0.21 is only 1.8% of the average log sale price. So the accuracy is pretty good.

Figure 18: Comparison table of models. Actual vs. predicted plots of test set.

Model	LASSO ($\alpha = 1$)				Ridge ($\alpha = 0$)			
	non-zero coefficients	train RMSE	test RMSE	RMSE difference (irreducible error)	non-zero coefficients	train RMSE	test RMSE	RMSE difference (irreducible error)
λ_{1se}	10	0.21443	0.20859	0.00584	26	0.21415	0.21102	0.00313
								
λ_{min}	22	0.19823	0.19533	0.0029	26	0.19824	0.19634	0.0019
								

Conclusion

As real estate investors in Ames, Iowa, we have been able to validate answers to some valuable questions as we begin our investment journey:

- **Features:** As for basement features, we were able to detect which features had a meaningful difference in impacting sale price, so we can target properties with those features when investing (such as good basement exposure).
- **Location:** As for our location analysis, we uncovered that, out of all 3 of the “best” neighborhoods in Ames averaging over a \$300k sale price, there is no meaningful difference between those three neighborhoods’ sale prices, so they are all equal and there wouldn’t be one neighborhood we should focus on solely by average sale price.
- **Size:** As for our size analysis, we validated that we can quite accurately predict the log value sale price of a home using only size-indicating variables. We have not translated these into actual sales prices. This model allows for us to not have to rely on quality-indicating variables to predict price, because they are less accessible in general property assessments.

References

De Cock D. 2011. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*; 19(3).

<https://jse.amstat.org/v19n3/decock.pdf>

King, A.P., Eckersley, R.J. (2019) Inferential Statistics IV: Choosing a Hypothesis Test.

Statistics for Biomedical Engineers and Scientists. 7.(3.4).

<https://www.sciencedirect.com/topics/mathematics/wilk-test#:~:text=The%20Shapiro%E2%80%9393%20Wilk%20test%20statistic,1%20being%20a%20perfect%20match.>

Zach. 2020, October 13. How to transform data in R (log, square root, cube root). *Statology*.

<https://www.statology.org/transform-data-in-r/>