Two-Step Transformation to Normality using R.

Samantha Deokinanan
CUNY MSDS DATA 621 Blog #1
September 13th, 2020

Introduction

The distribution that is inarguably the most important and the most frequently used distribution in both theory and application of statistics is the Normal (Gaussian) Distribution. The assumption of normality needs to be checked for many statistical procedures, namely parametric tests, because their validity depends on it. When these assumptions do not hold, it is impossible to draw accurate and reliable conclusions about reality.

According to the central limit theorem, if the sample data are approximately normal then the sampling distribution too will be normal; in large samples (> 40), the sampling distribution tends to be normal, regardless of the shape of the data; and means of random samples from any distribution will have a normal distribution. However, research studies in Health, Education, and Social Sciences often deal with small sample sizes and non-normal data. Blanca *et al.* (2013) analyzed the shape of 693 distributions from real psychological data by examining the values of the third and fourth central moments as a measurement of skewness and kurtosis in small samples. They found that most distributions were non-normal. Moreover, considering skewness and kurtosis jointly, the results indicated that only 5.5% of the distributions were close to expected values under normality. Overall, 74.4% of distributions presented either slight or moderate deviation, while 20% showed more extreme deviation.

While there are many accepted diagnostics tests (e.g., Kolmogorov-Smirnov, P-P plot, skewness, kurtosis) that are used to access normality of a parameter, if a test fails, data transformations (e.g., power and logarithm) are usually pursued. However, it may not always produce acceptable results, and this trial-and-error approach can become frustrating. Therefore, this blog aims to demonstrate a simple yet powerful approach referred to as the Two-Step as derived by Templeton, G.F. (2011). Because this method is not widely known, this blog goes further and presents the R way of carrying out this method.

Methodology

Two-Step may be used to transform many non-normally distributed continuous variables toward statistical normality (i.e., satisfies the preponderance of appropriate diagnostics tests for normality). The proposed transformation can achieve statistically acceptable kurtosis, skewness, and an overall normality test in many situations and improve normality in many others. As the name suggests, the following are the two steps:

Step 1 - transform the original variable toward statistical uniformity by calculating the percentile rank of each score.

Step 2 - any variable found to conform to statistical uniformity is Normal-Feasible. Uniform probabilities may be transformed to normal using the inverse normal distribution function:

$$p = \mu\sqrt{2}\sigma(erf^{-1})(-1 + 2Pr)$$

where
p is the z-score,
μ is the mean of p,

σ is the standard deviation of p,
erf^{-1} is the inverse error function, and
Pr is the percentile rank from *Step 1*.

R code

In `R`, the following function can be used to transform a variable into a normal distribution using Two-Step:

```
two.step = function(variable, mean, sd){
  # Step 1
  variable.prob = 1 - (rank(variable, na.last = TRUE, ties.method = "average")/length(variable))
  variable.prob[variable.prob == 0] = 0.0001
  variable.prob[variable.prob == 1] = 0.9999

  # Step 2
  t.variable = qnorm(variable.prob, mean, sd, lower.tail = FALSE)
  return(t.variable)
}
```

Example

The Dataset

The Ames Housing dataset describes the sale of individual residential property in Ames, Iowa from 2006 to 2010, was compiled by Dean De Cock for use in data science education. The dataset provides 79 explanatory variables and a sample size of n = 1460. Some of the variables of interest are listed below.

Variables	Descriptions
LotFrontage	Linear feet of street connected to property
LotArea	Lot size, ft^2
WoodDeckSF	Wood deck area, ft^2
OpenPorchSF	Open porch area, ft^2
EnclosedPorch	Enclosed porch, ft^2
ScreenPorch	Screen porch area, ft^2
PoolArea	Pool area, ft^2
GarageArea	Size of garage, ft^2
OverallCond	Rates the overall condition of the house
SalePrice	Selling Price of the House

```
library(tidyverse)
library(psych)
library(knitr)

df = read.csv("train.csv", header = TRUE, sep = ",")
house = data.frame(df$LotFrontage,df$LotArea,df$WoodDeckSF,df$OpenPorchSF,df$EnclosedPorch,df$Sc
reenPorch,df$PoolArea,df$GarageArea,df$OverallCond,df$SalePrice)
names(house) = substring(names(house), 4)
```

Exploratory Analysis

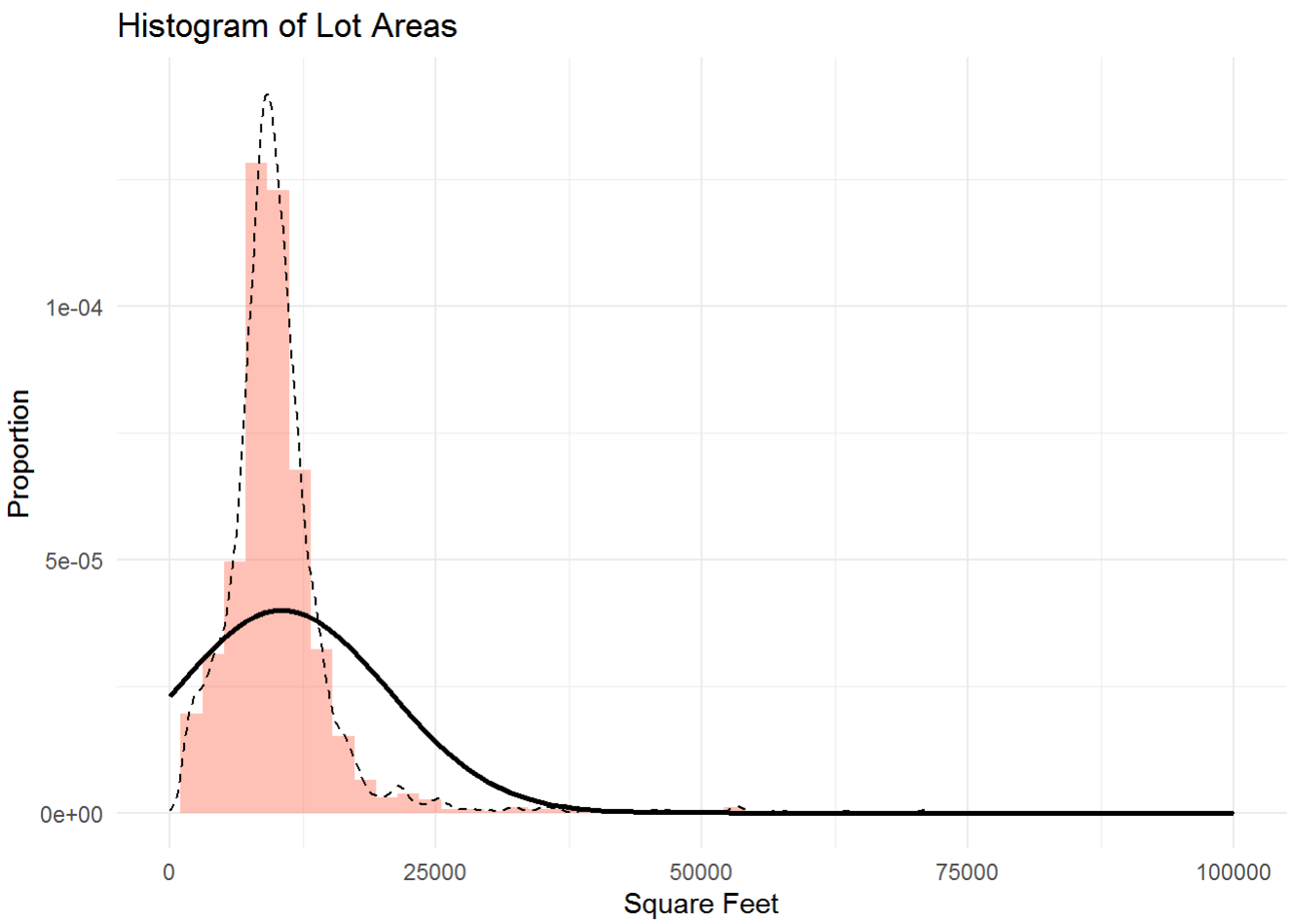
From the descriptive statistics, there are a few problematic variables that are highly skewed, and therefore would not be normally distributed.

```
describe(house, na.rm = TRUE)[,-c(1,6:10)] %>% kable(digits = 2L)
```

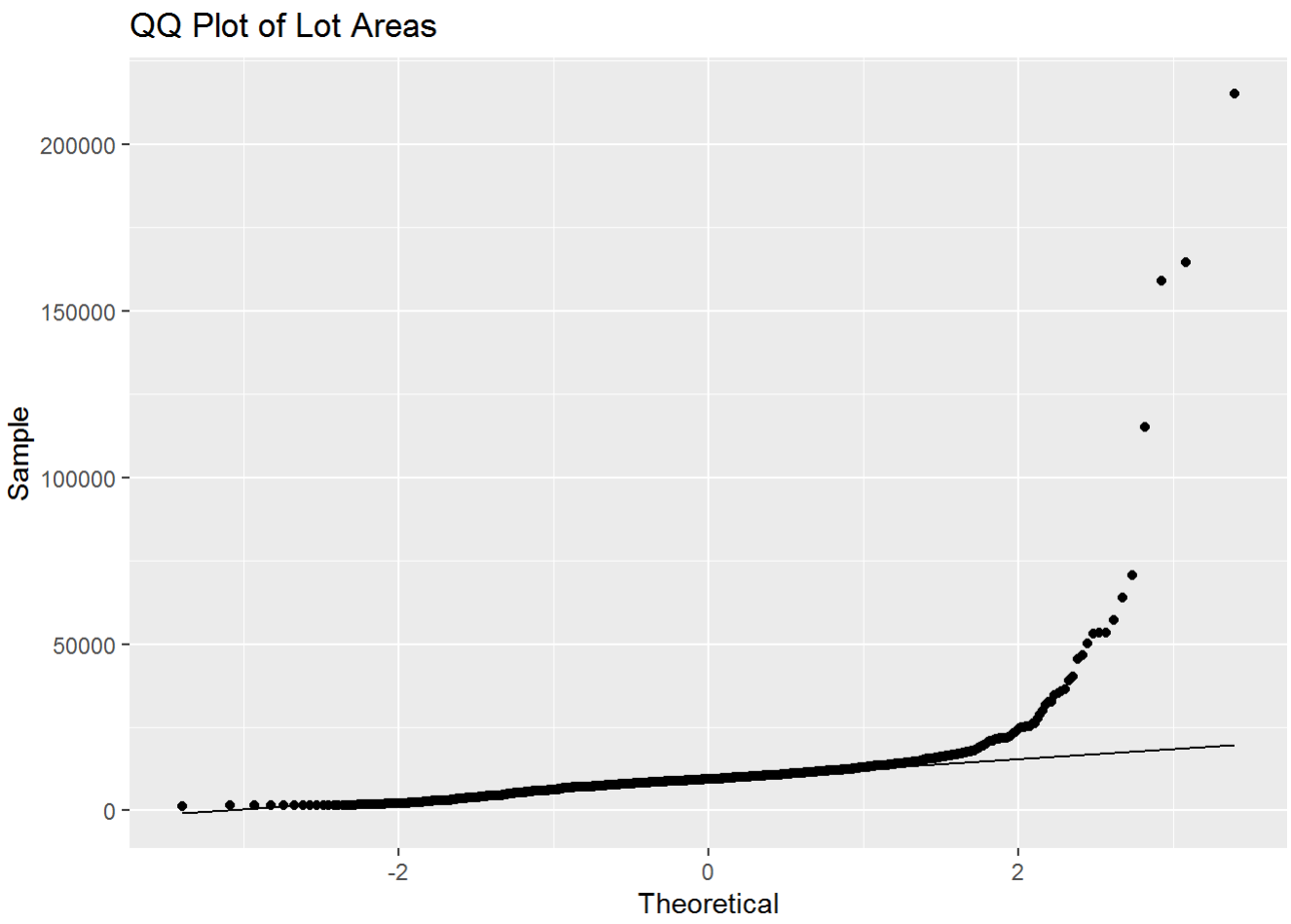
	n	mean	sd	median	skew	kurtosis	se
LotFrontage	1201	70.05	24.28	69.0	2.16	17.34	0.70
LotArea	1460	10516.78	9981.23	9478.5	12.18	202.27	261.22
WoodDeckSF	1460	94.24	125.34	0.0	1.54	2.97	3.28
OpenPorchSF	1460	46.66	66.26	25.0	2.36	8.44	1.73
EnclosedPorch	1460	21.95	61.12	0.0	3.08	10.37	1.60
ScreenPorch	1460	15.06	55.76	0.0	4.11	18.34	1.46
PoolArea	1460	2.76	40.18	0.0	14.80	222.19	1.05
GarageArea	1460	472.98	213.80	480.0	0.18	0.90	5.60
OverallCond	1460	5.58	1.11	5.0	0.69	1.09	0.03
SalePrice	1460	180921.20	79442.50	163000.0	1.88	6.50	2079.11

One of the variables of interest is positively skewed, namely `LotArea`, with skewness of 12.18. From the histogram below, the dashed line shows the current density of the raw data while the solid line is an imposed normal distribution with the same mean and standard deviation. The difference between the two highlights lack of a normal distribution. The non-normality and positive skew are further emphasized by the QQ plot. And lastly, the Shapiro-Wilk test of normality for the variable resulted in the p-value < 0.05. These are just three of several methods that can be used to assess the normality of a variable. Altogether, these methods confirm that the distribution of the data is significantly different from a normal distribution. In other words, normality cannot be assumed.

```
ggplot(house, aes(x = LotArea)) + geom_histogram(aes(y = ..density..), bins = 50, alpha = 0.4, f
ill = "tomato") + theme_minimal() + theme(legend.title = element_blank()) + labs(title = "Histog
ram of Lot Areas", x = "Square Feet", y = "Proportion") + xlim(0, 10^5) + geom_density(alpha =
0.4, linetype = 2, size = 0.5) + stat_function(fun = dnorm, args = list(mean = mean(house$LotAre
a), sd = sd(house$LotArea)), size = 1)
```




```
ggplot(house, aes(sample = LotArea)) + stat_qq() + stat_qq_line() + labs(title = "QQ Plot of Lot Areas", x = "Theoretical", y = "Sample")
```




```
shapiro.test(house$LotArea)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  house$LotArea
## W = 0.35105, p-value < 2.2e-16
```

Now, the plots suggest that outliers are present and are greatly influencing the distribution. Thus, outliers are removed for `LotArea` and the normality assessment is conducted once again.

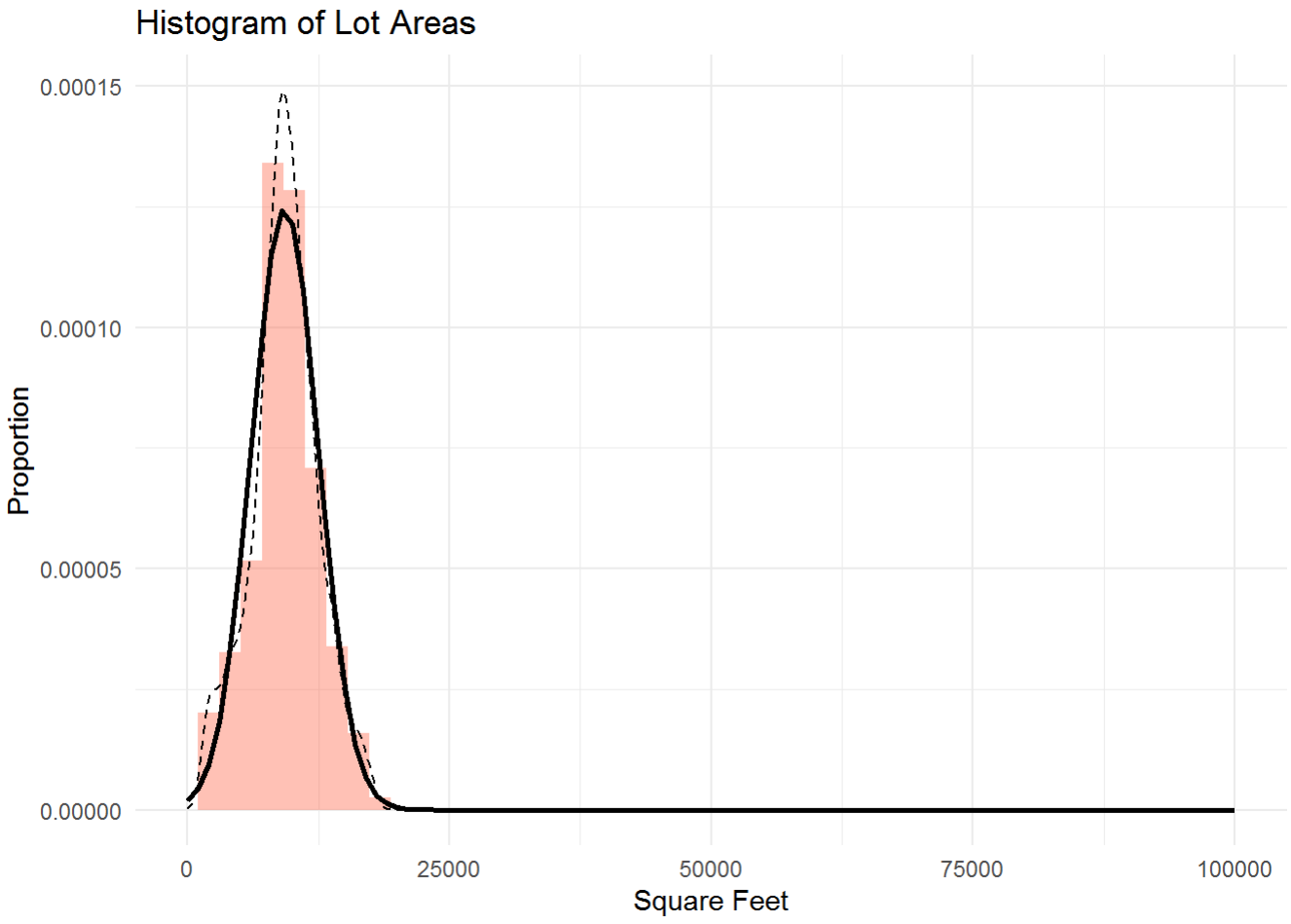
```
# Outlier removed
house.outlier.removed = house[-which(house$LotArea %in% boxplot(house$LotArea, plot = FALSE)$ou
t),]

describe(house.outlier.removed)[,-c(1,6:10)] %>% kable(digits = 2L)
```

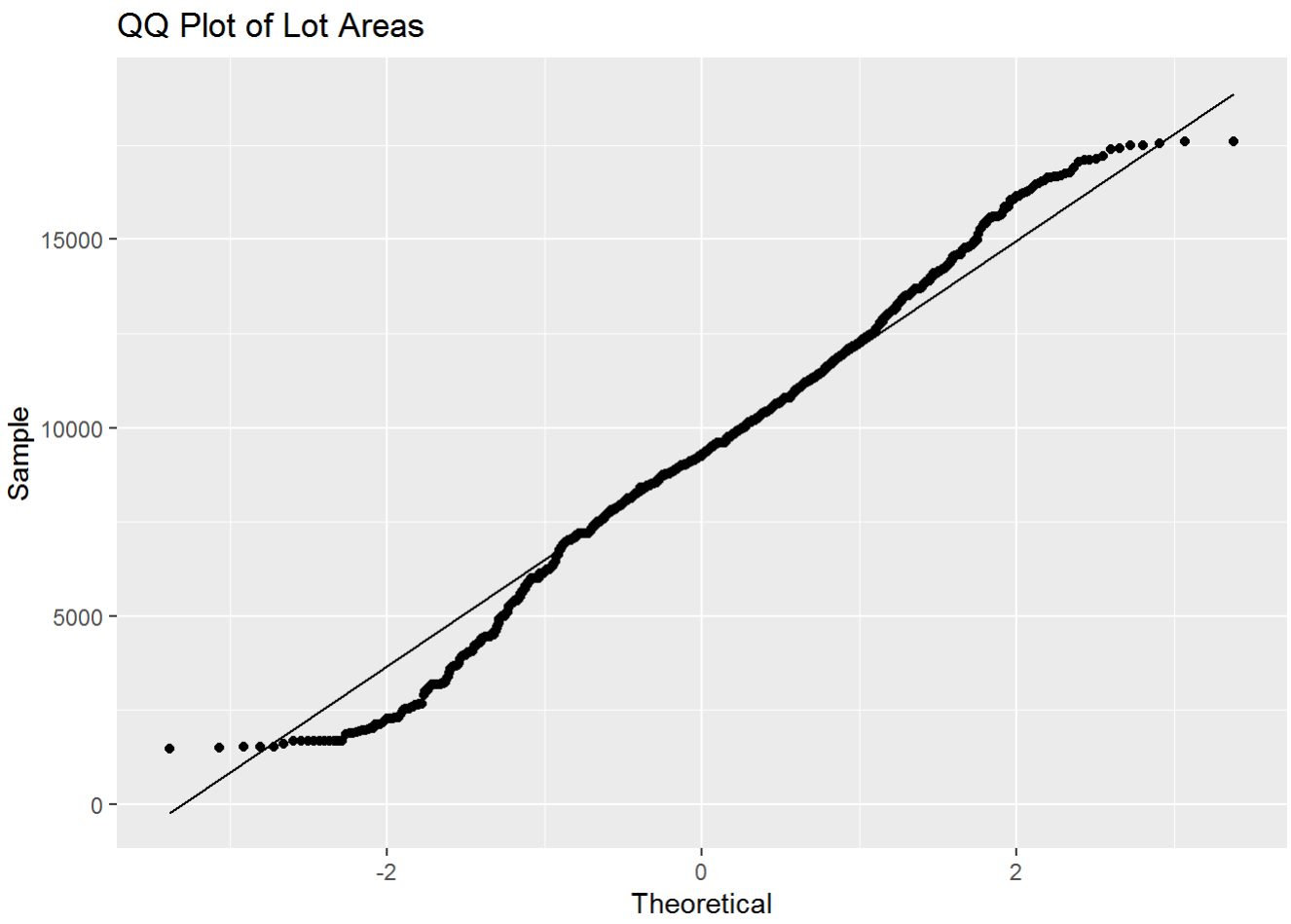
	n	mean	sd	median	skew	kurtosis	se
LotFrontage	1159	68.79	20.92	69	0.49	2.04	0.61
LotArea	1392	9266.89	3202.27	9274	-0.08	0.06	85.83
WoodDeckSF	1392	90.55	119.71	0	1.44	2.36	3.21

	n	mean	sd	median	skew	kurtosis	se
OpenPorchSF	1392	45.14	63.49	24	2.35	8.96	1.70
EnclosedPorch	1392	21.64	59.40	0	2.85	7.49	1.59
ScreenPorch	1392	14.43	54.18	0	4.17	18.95	1.45
PoolArea	1392	1.65	30.87	0	18.75	352.14	0.83
GarageArea	1392	466.72	210.17	474	0.06	0.54	5.63
OverallCond	1392	5.58	1.11	5	0.67	1.11	0.03
SalePrice	1392	178065.79	75811.56	160000	1.69	5.20	2031.96

```
ggplot(house.outlier.removed, aes(x = LotArea)) + geom_histogram(aes(y = ..density..), bins = 50
, alpha = 0.4, fill = "tomato") + theme_minimal() + theme(legend.title = element_blank()) + labs
(title = "Histogram of Lot Areas", x = "Square Feet", y = "Proportion") + xlim(0, 10^5) + geom_d
ensity(alpha = 0.4, linetype = 2, size = 0.5) + stat_function(fun = dnorm, args = list(mean = me
an(house.outlier.removed$LotArea), sd = sd(house.outlier.removed$LotArea)), size = 1)
```




```
ggplot(house.outlier.removed, aes(sample = LotArea)) + stat_qq() + stat_qq_line() + labs(title =
"QQ Plot of Lot Areas", x = "Theoretical", y = "Sample")
```




```
shapiro.test(house.outlier.removed$LotArea)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  house.outlier.removed$LotArea
## W = 0.99034, p-value = 6.119e-08
```

With the removal of the outliers in `LotArea`, there is a noticeable change in the skewness, i.e. skewness is now -0.08. From the histogram, the dashed and the solid line are more similar than the previous graph. Unfortunately, the skewness is still present, as seen in the QQ plot. Additionally, the Shapiro-Wilk test resulted in the p-value < 0.05. Overall, the removal of outliers greatly improved the normality of the data, however, it is still significantly different from a normal distribution.

Now, applying the Two-Step transformation on the variable:

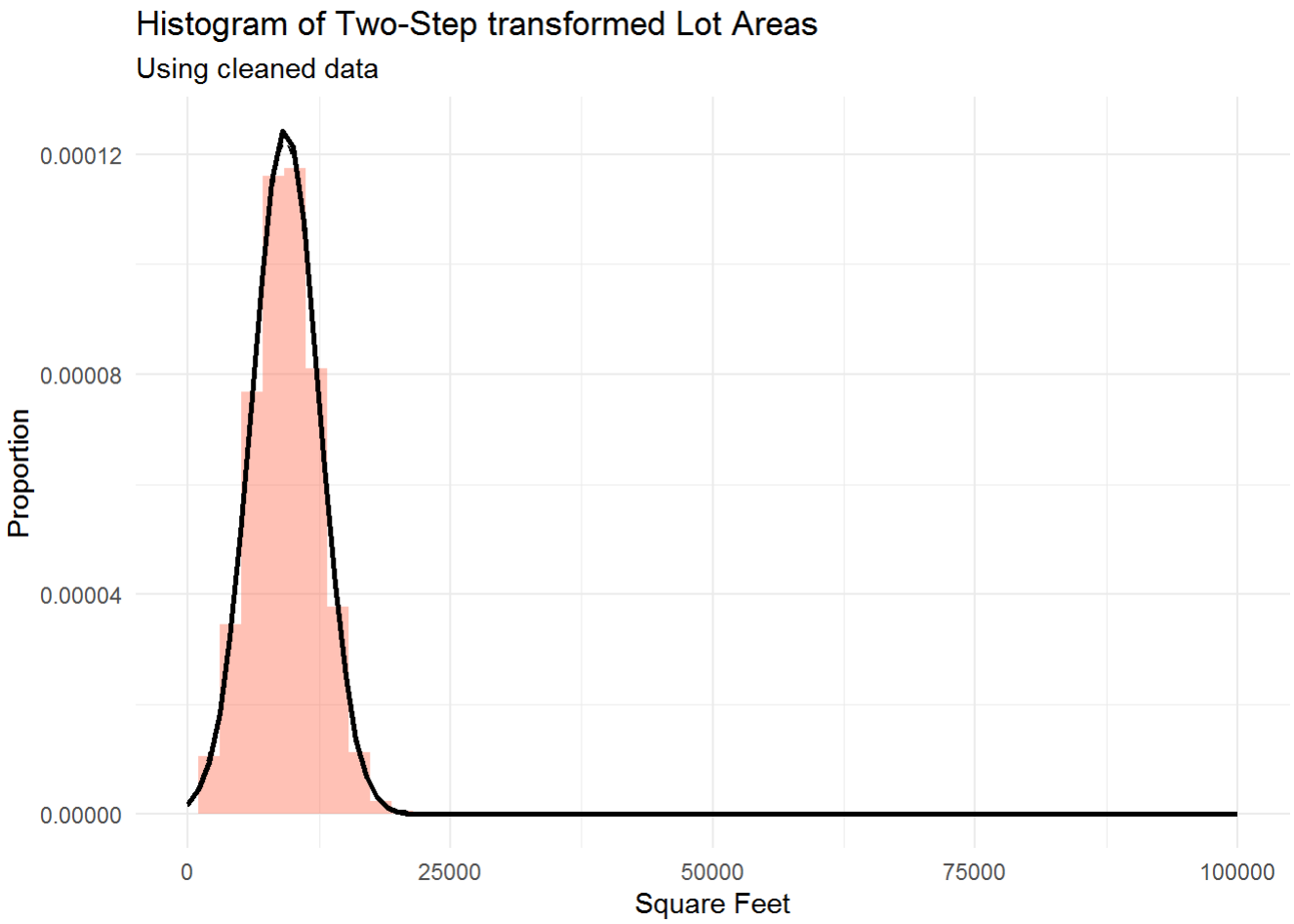
```
# With cleaned data
LotArea.mean.clean = mean(house.outlier.removed$LotArea)
LotArea.sd.clean = sd(house.outlier.removed$LotArea)
house.outlier.removed$LotArea.t = two.step(house.outlier.removed$LotArea, LotArea.mean.clean, Lo
tArea.sd.clean)
```

By comparing the statistics among the raw, cleaned and Two-Step transformed variables, it is evident that the transformed data from both the raw and cleaned data is not significantly different from a normal distribution. Visually, from the histogram, the dash and solid lines are in line, and a majority of the points fall on the line in the QQ plot. The Shapiro-Wilk test of normality for the variable resulted in the p-value > 0.05. Overall, these methods confirm that the distribution of the Two-Step transformed data is not significantly different from a normal distribution.

```
results = cbind(LotArea = c('Original', 'Cleaned', 'Two-Step Cleaned'),
                rbind(describe(house)[2,-c(1,6:10)], describe(house.outlier.removed)[c(2,11),-c(
1,6:10)]))
rownames(results) = NULL
results %>% kable(digits = 2L)
```

LotArea	n	mean	sd	median	skew	kurtosis	se
Original	1460	10516.78	9981.23	9478.50	12.18	202.27	261.22
Cleaned	1392	9266.89	3202.27	9274.00	-0.08	0.06	85.83
Two-Step Cleaned	1392	9275.39	3201.65	9269.77	0.03	-0.03	85.81

```
ggplot(house.outlier.removed, aes(x = LotArea.t)) + geom_histogram(aes(y = ..density..), bins =
50, alpha = 0.4, fill = "tomato") + theme_minimal() + theme(legend.title = element_blank()) + la
bs(title = "Histogram of Two-Step transformed Lot Areas", subtitle = "Using cleaned data", x =
"Square Feet", y = "Proportion") + xlim(0, 10^5) + geom_density(alpha = 0.4, linetype = 2, size
 = 0.5) + stat_function(fun = dnorm, args = list(mean = LotArea.mean.clean, sd = LotArea.sd.clea
n), size = 1)
```




```
ggplot(house.outlier.removed, aes(sample = LotArea.t)) + stat_qq() + stat_qq_line() + labs(title
= "QQ Plot of Two-Step transformed Lot Areas", subtitle = "Using cleaned data", x = "Theoretica
l", y = "Sample")
```


```
shapiro.test(house.outlier.removed$LotArea.t)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  house.outlier.removed$LotArea.t
## W = 0.99971, p-value = 0.9998
```

It is noteworthy that if a Two-Step transformation was performed on the raw data, it will still result in a closely, normally distributed data.

```
##
##  Shapiro-Wilk normality test
##
## data:  house$LotArea.t
## W = 0.99979, p-value = 1
```

LotArea	n	mean	sd	median	skew	kurtosis	se
Original	1460	10516.78	9981.23	9478.50	12.18	202.27	261.22
Two-Step Original	1460	10542.89	9983.24	10525.35	0.03	-0.01	261.27
Cleaned	1392	9266.89	3202.27	9274.00	-0.08	0.06	85.83

LotArea	n	mean	sd	median	skew	kurtosis	se
Two-Step Cleaned	1392	9275.39	3201.65	9269.77	0.03	-0.03	85.81

Limitations

Step 1 is a critical step since the achievement of statistical uniformity is required before *Step 2* will result in statistical normality. Some situations will not allow for the achievement of statistical uniformity, which is a very high standard according to the norms of non-STEM research.

Ratio data is most amenable to successful transformations into normal distributions using this approach. Ordinal and interval data types with greater numbers of levels will also be more successful to be transformed, and categorical data types cannot logically be transformed.

When using the technique, be cautious of the frequency and influence of mode values. If modes are found to impair results, consider replacing mode values with missing values, and retry the transformation. In count variables, influential modes are typically represented by values of zero.

Works Cited

- Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). *Skewness and kurtosis in real data samples*. Methodology 9, 78–84. doi:10.1027/1614-2241/a000057 (doi:10.1027/1614-2241/a000057)
- Cock, D. D. (2011) *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*. Journal of Statistics Education 19:3. doi:10.1080/10691898.2011.11889627 (doi:10.1080/10691898.2011.11889627)
- Templeton, G. F. (2011). *A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research*, Communications of the AIS, 28:4.