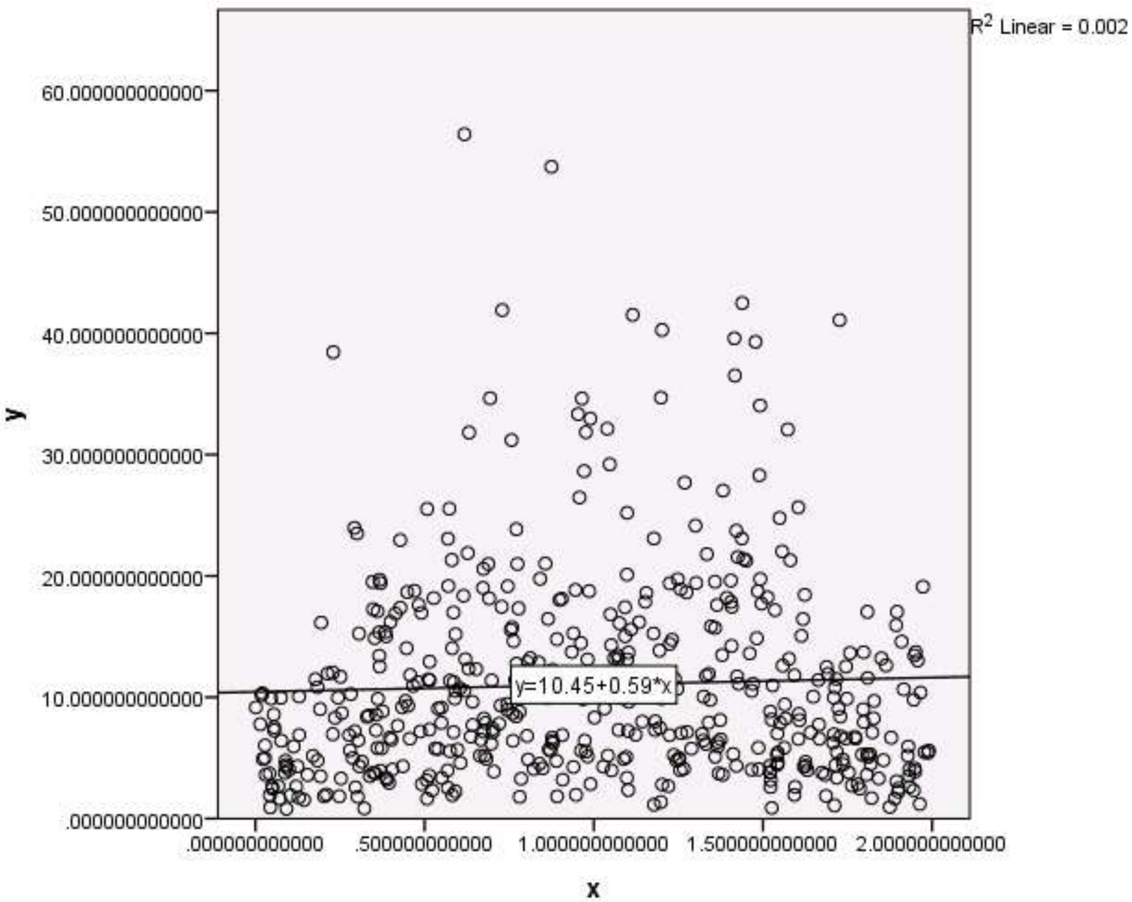
HOMEWORK-4

1. **I first created a scatter plot to display the relationship between independent variable x and dependent variable y. After fitting the trend line into the scatterplot, almost no correlation was found between x and y. This implies that variance of x did not sufficiently explain the variance of y.**

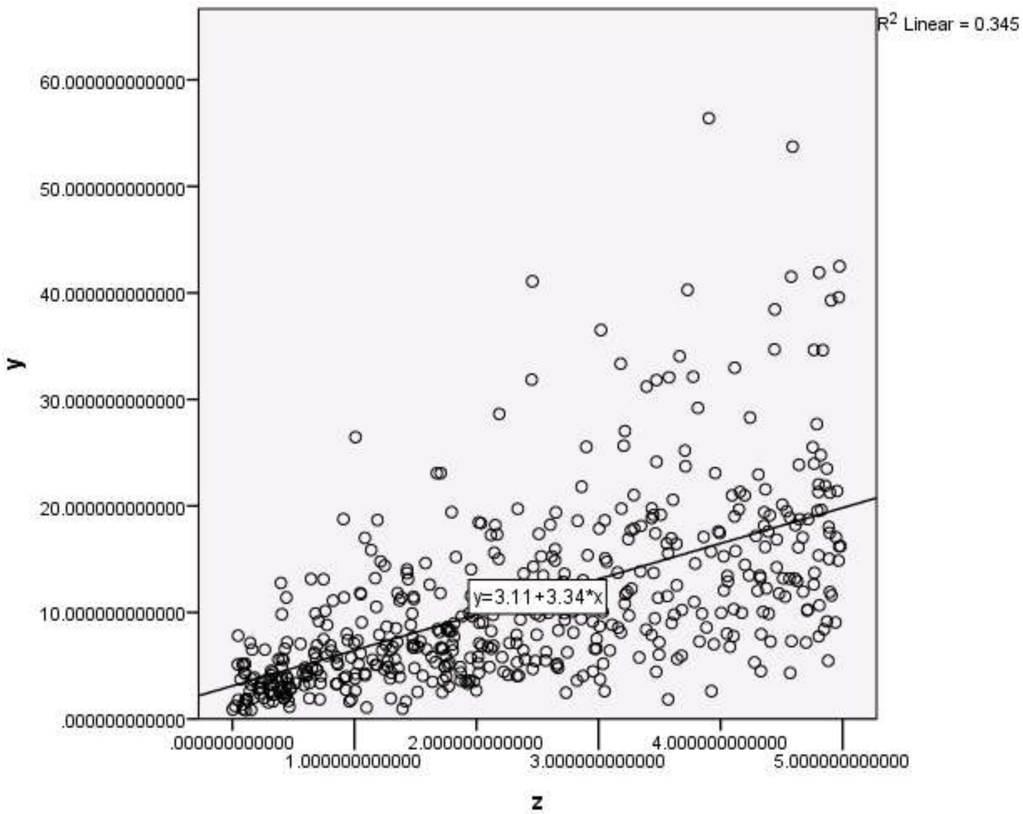
```
NEW FILE.
DATASET NAME DataSet1 WINDOW=FRONT.

GET DATA
  /TYPE=XLS
  /FILE='C:\Users\rrane1\Desktop\sample.xls'
  /SHEET=name 'sample'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME DataSet3 WINDOW=FRONT.
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=x y MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: x=col(source(s), name("x"))
  DATA: y=col(source(s), name("y"))
  GUIDE: axis(dim(1), label("x"))
  GUIDE: axis(dim(2), label("y"))
  ELEMENT: point(position(x*y))
```



END GPL.

2. **I again created a scatter plot to display the relationship between independent variable z and dependent variable y. A strong positive correlation was found was found between x and y. After fitting the trend line into the scatter plot, we can say that 34.5% of the variance in y can be explained by variance in z.**

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=z y MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: z=col(source(s), name("z"))
  DATA: y=col(source(s), name("y"))
  GUIDE: axis(dim(1), label("z"))
  GUIDE: axis(dim(2), label("y"))
  ELEMENT: point(position(z*y))
END GPL.
```



3. **According to the Correlations table, we can figure out that there is a significant correlation between z and y (0.587), whereas there is a very less correlation between x and y (0.039).**

```
DATASET ACTIVATE DataSet3.
DATASET CLOSE DataSet1.
CORRELATIONS
  /VARIABLES=x z y
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

		x	z	y
x	Pearson Correlation	1	.063	.039
	Sig. (2-tailed)		.160	.383
	N	500	500	500
z	Pearson Correlation	.063	1	.587**
	Sig. (2-tailed)	.160		.000

	N	500	500	500
y	Pearson Correlation	.039	.587**	1
	Sig. (2-tailed)	.383	.000	
	N	500	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

4. **Next, I carried out the Multiple Linear Regression with x and z as the independent variables(predictors) and y as the dependent variable (response variable). Together, x and z accounted for 34.5% of the variance of y. However, taken individually z also accounted for 34.5 % variance of y.**

The Multiple Linear Regression Equation for this model is as follows:

y = 3.076 + 0.032x + 3.342 z + E

This implies that for each unit increase in x, the estimated average amount of y is increased by 0.032 units, holding z constant.
This implies that for each unit increase in x\z, the estimated average amount of y is increased by 3.342 units, holding x constant.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x z.
```

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	z, x[b]	.	Enter

a. Dependent Variable: y

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.587[a]	.345	.342	6.903845902157768

a. Predictors: (Constant), z, x

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12473.455	2	6236.728	130.850	.000[b]
	Residual	23688.555	497	47.663		
	Total	36162.010	499			

a. Dependent Variable: y
b. Predictors: (Constant), z, x

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.076	.776		3.962	.000
	x	.032	.549	.002	.058	.954
	z	3.342	.207	.587	16.141	.000

a. Dependent Variable: y

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x z
```

/SAVE PRED.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	z, x[b]	.	Enter

a. Dependent Variable: y

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.587[a]	.345	.342	6.903845902157768

a. Predictors: (Constant), z, x

b. Dependent Variable: y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12473.455	2	6236.728	130.850	.000[b]
	Residual	23688.555	497	47.663		
	Total	36162.010	499			

a. Dependent Variable: y

b. Predictors: (Constant), z, x

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.076	.776		3.962	.000
	x	.032	.549	.002	.058	.954
	z	3.342	.207	.587	16.141	.000

a. Dependent Variable: y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.13382625579834	19.74876213073731	11.03555140993509	4.999690379536319	500
Residual	-14.041371345520020	40.258914947509766	-.000000000000005	6.889996648688157	500
Std. Predicted Value	-1.580	1.743	.000	1.000	500
Std. Residual	-2.034	5.831	.000	.998	500

a. Dependent Variable: y

5. **Next, I calculated the Unstandardized Predicted Value based on the unstandardized coefficients (for both x and z) found in the previous section. I then found the correlation between the unstandardized predicted value (of both x and z) and the dependent variable y. I again found that the correlation between the dependent and the independent variables was 0.587.**

```
CORRELATIONS
 /VARIABLES=y PRE_1
 /PRINT=TWOTAIL NOSIG
 /MISSING=PAIRWISE.
```

Correlations

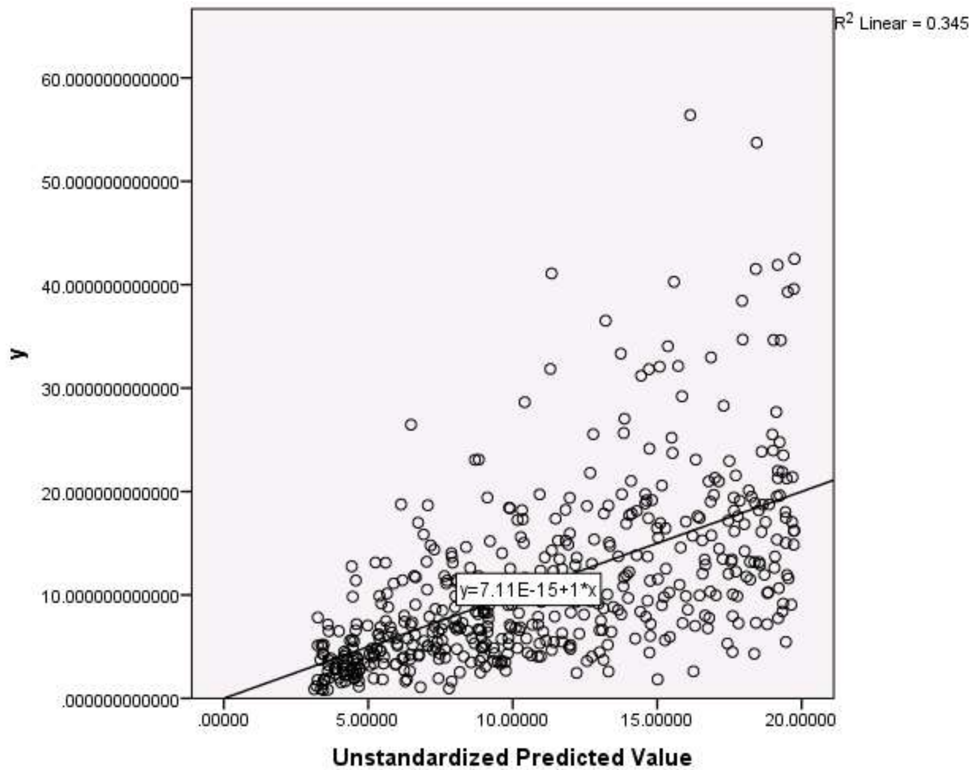
	y	Unstandardized Predicted Value

y	Pearson Correlation	1	.587**
	Sig. (2-tailed)		.000
	N	500	500
Unstandardized Predicted Value	Pearson Correlation	.587**	1
	Sig. (2-tailed)	.000	
	N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

6. **I then derived the scatter plot between the unstandardized predicted values and dependent variable. This time again I found that 34.5 % of the variance of y was explained by the variance of unstandardized predicted values.**

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=PRE_1 y MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: PRE_1=col(source(s), name("PRE_1"))
  DATA: y=col(source(s), name("y"))
  GUIDE: axis(dim(1), label("Unstandardized Predicted Value"))
  GUIDE: axis(dim(2), label("y"))
  ELEMENT: point(position(PRE_1*y))
END GPL.
```



Descriptives

			Statistic	Std. Error
y	Mean		11.03555140993508	.380707125518191
	95% Confidence Interval for Mean	Lower Bound	10.28756493107124	
		Upper Bound	11.78353788879892	
	5% Trimmed Mean		10.13148278142971	
	Median		8.75753733676500	
	Variance		72.469	
	Std. Deviation		8.512870121772210	
	Minimum		.803793532026	

	Maximum	56.401666828441	
	Range	55.597873296415	
	Interquartile Range	10.081229470144	
	Skewness	1.746	.109
	Kurtosis	4.161	.218

```
EXAMINE VARIABLES=y
  /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

7. **Next, I assessed the normality of the distribution of the dependent variable y.**

Skewness = 1.746 , which is greater than 1. This implies that the data for y is not normally distributed.

Significance level for both the Kolmogorov and Shapiro Wilk Test is less than 0.05 (0.000). This means that the data is not normally distributed.

The data also did not fit around the normal Q-Q plot line, indicating that the data was not normalized.

The Histogram also clearly depicts that the data is skewed.

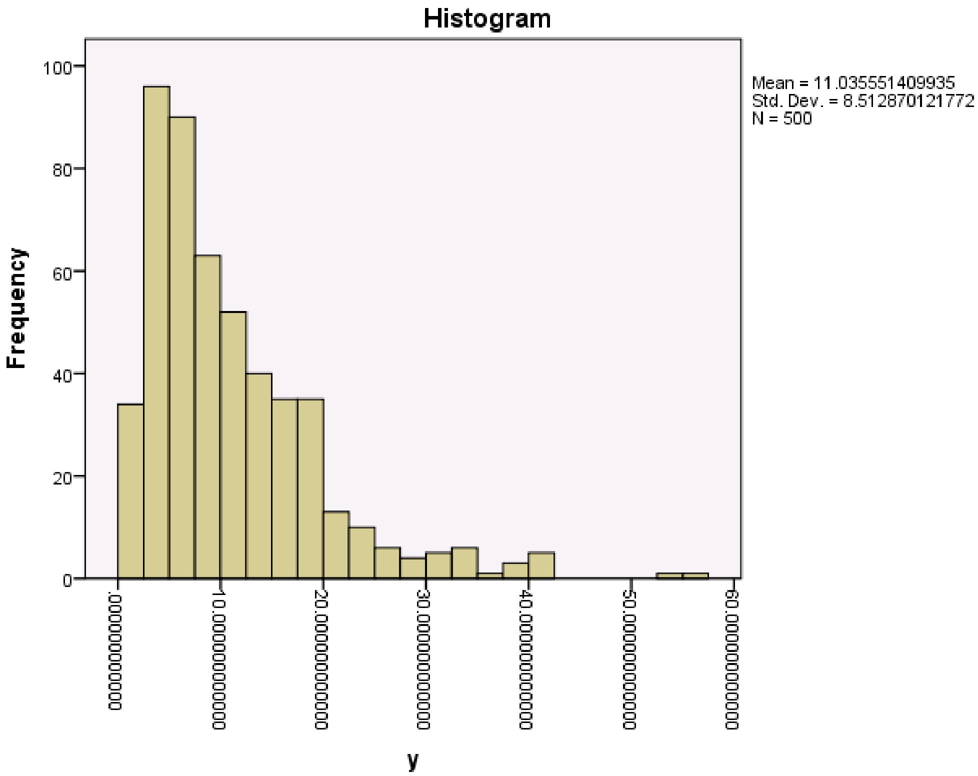
Case Processing Summary

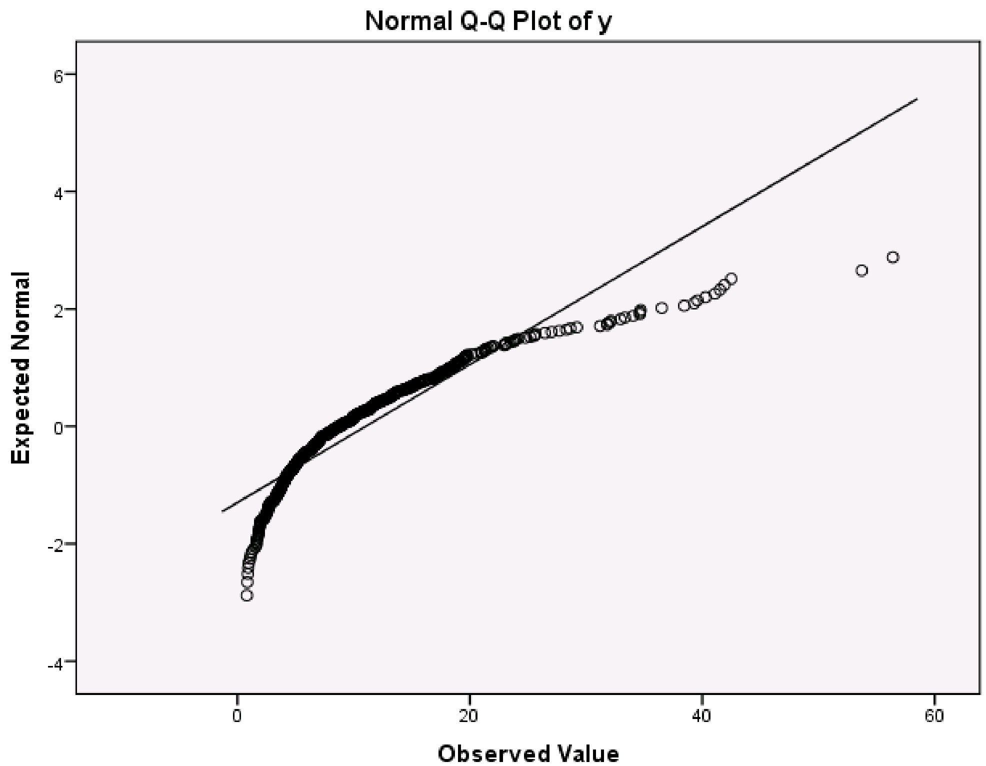
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
y	500	100.0%	0	0.0%	500	100.0%

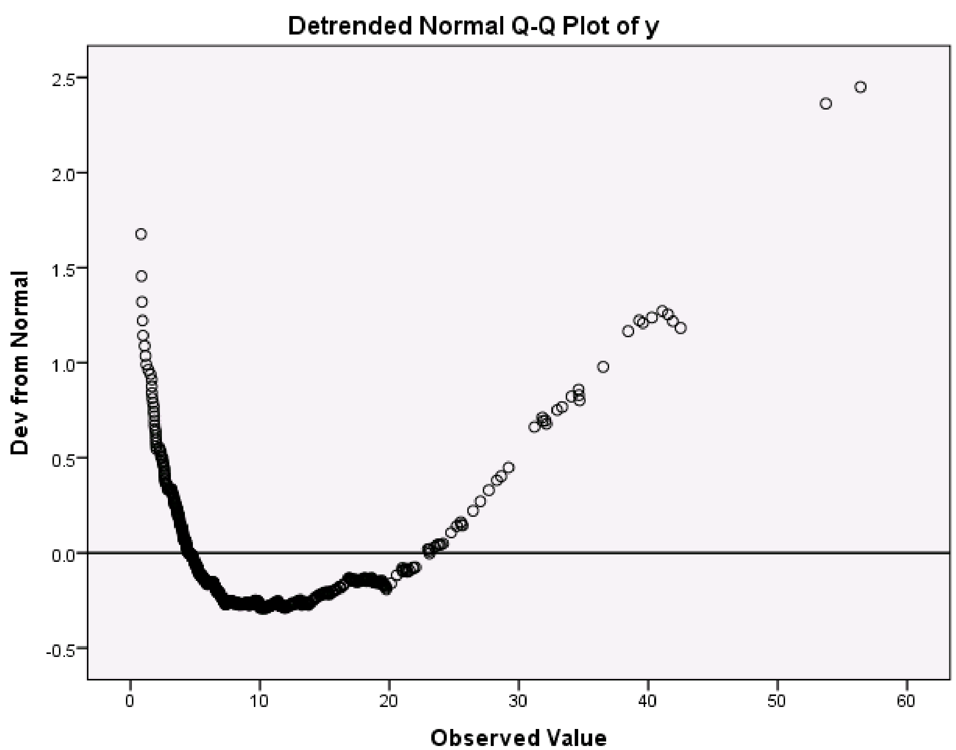
Tests of Normality

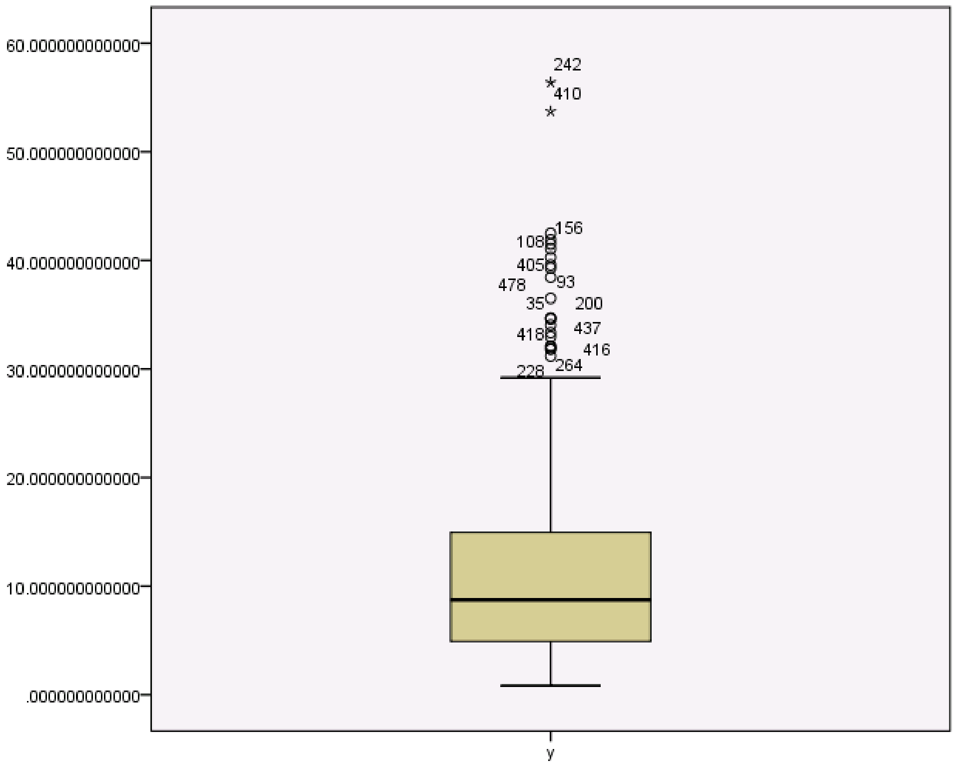
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
y	.117	500	.000	.852	500	.000

a. Lilliefors Significance Correction


Histogram
Mean = 11.035551409935
Std. Dev. = 8.512870121772
N = 500
Frequency
100
80
60
40
20
0
.000000000000
10.000000000000
20.000000000000
30.000000000000
40.000000000000
50.000000000000
60.000000000000
y


Normal Q-Q Plot of y
Expected Normal
Observed Value
6
4
2
0
-2
-4
0
20
40
60


Detrended Normal Q-Q Plot of y
Dev from Normal
Observed Value
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
10
20
30
40
50
60



8. **So, I transformed the dependent variable y by using log y. This enabled me to get normalized data for y.**

Skewness was -0.3 (which lies between the range of -0.1 and 0.1). This implies that the data followed a normal distribution.
The Histogram also shows that most of the data is centered around the mean and is normally distributed.
The Q-Q plot indicates a good fit between the data points and the line, indicating that the data is now transformed to follow a normal distribution.
Transformation was carried out so that the data could more closely meet the assumptions of a statistical inference procedure and to improve the interpretability of graphs.

```
GET DATA
  /TYPE=XLS
  /FILE='C:\Users\rrane1\Desktop\sample.xls'
  /SHEET=name 'sample'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
```

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
y_log	500	100.0%	0	0.0%	500	100.0%

Descriptives

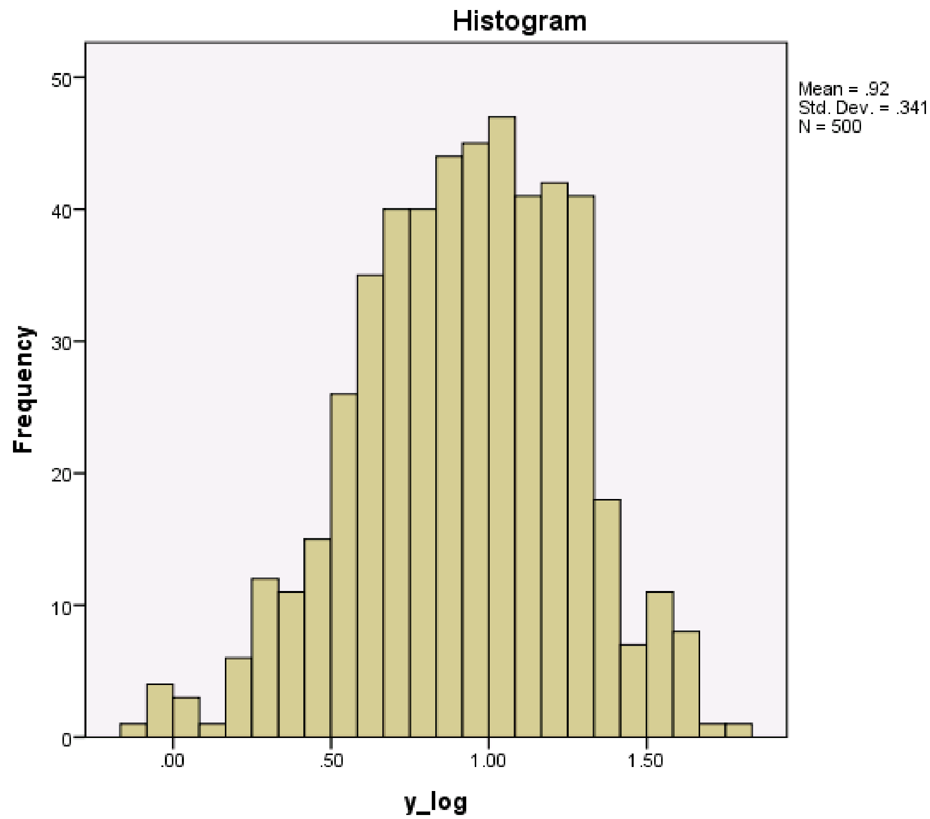
			Statistic	Std. Error
y_log	Mean		.9206	.01523
	95% Confidence Interval for Mean	Lower Bound	.8906	
		Upper Bound	.9505	
	5% Trimmed Mean		.9272	
	Median		.9424	
	Variance		.116	
	Std. Deviation		.34057	
	Minimum		-.09	
	Maximum		1.75	
	Range		1.85	
	Interquartile Range		.48	
	Skewness		-.311	.109
	Kurtosis		-.091	.218

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
y_log	.041	500	.045	.991	500	.005

a. Lilliefors Significance Correction

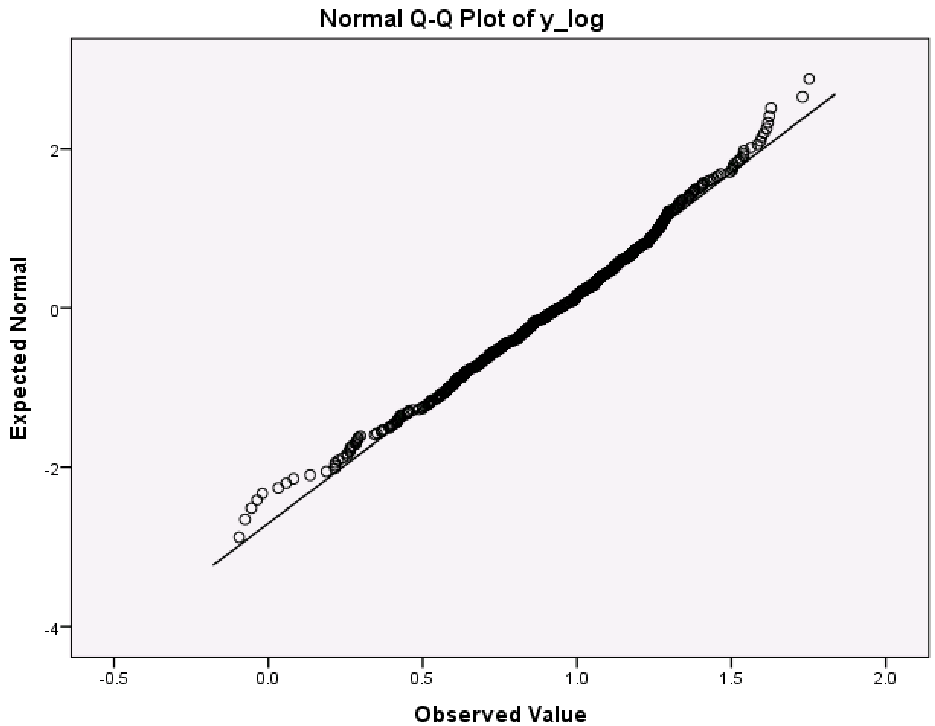
y_log

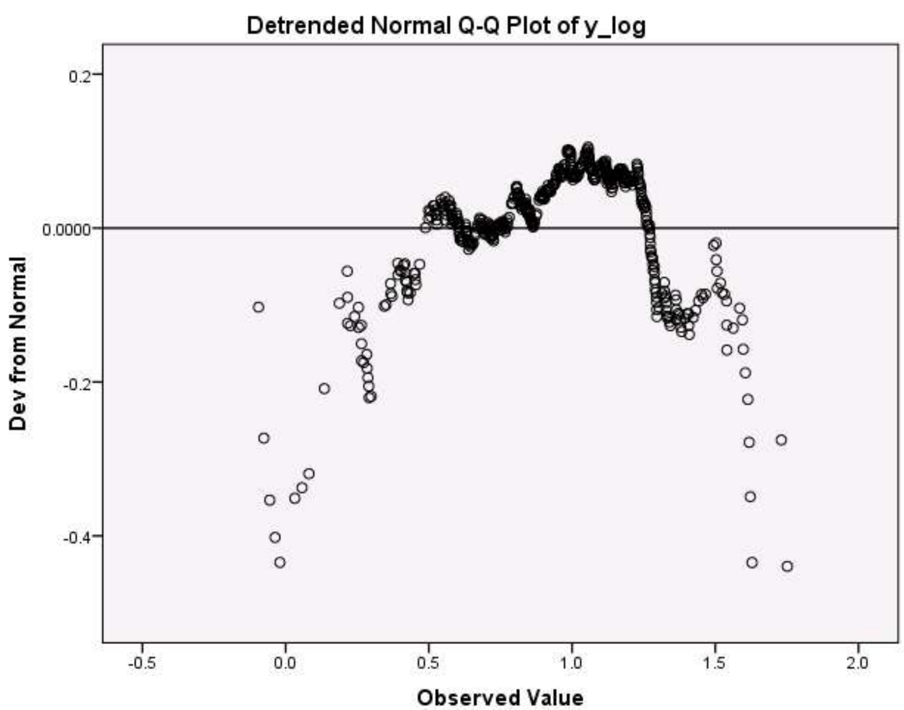


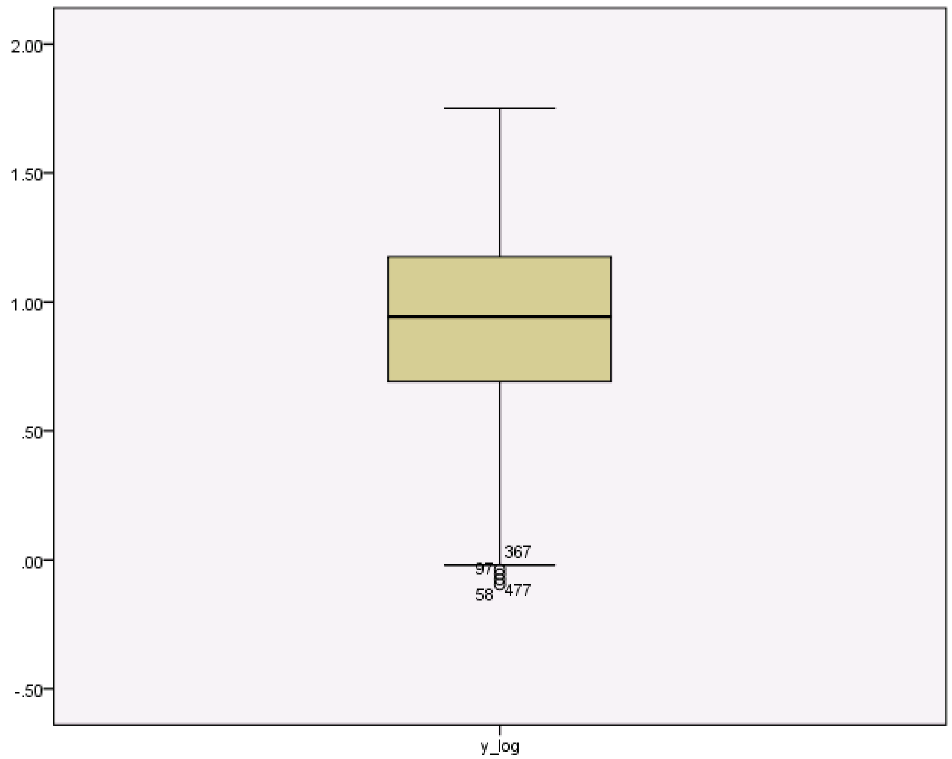
In order to make the variable better fit the assumptions underlying regression, we need to transform it.

Analyzing the histogram, we can conclude that the data is now normally distributed after the log transformation of y.

After transformation, the distribution is significantly closer to the normal probability distribution. It's a bit skewed to one side, but it's a big improvement.


Normal Q-Q Plot of y_log
Expected Normal
Observed Value
2
0
-2
-4
-0.5
0.0
0.5
1.0
1.5
2.0


Detrended Normal Q-Q Plot of y_log
Dev from Normal
Observed Value
0.2
0.0000
-0.2
-0.4
-0.5
0.0
0.5
1.0
1.5
2.0


2.00
1.50
1.00
.50
.00
-.50
367
97
58
477
y_log