The first part of this project involves merging two data sets and analyzing the resulting data set.We are testing the null hypothesis that the slope is zero in order to discover if there exists a relationship between the IV and DV. There are two separate data sets; one with an ID number and the corresponding independent variable, and the other with an ID and the corresponding dependent variable.

In order to manipulate the data sets, the programming language R was used with RStudio as the chosen developer environment. After accessing the two data sets, the merge function was used to combine the data sets by the ID number. That is, the independent variables and dependent variables were linked if they had the same ID number. Then, using the MICE package, the ID numbers missing both IV and DV were removed from the data set, as no information could be obtained from them. As seen in Table 1, 35 data points were missing IV only, 25 were missing DV only, and 26 were missing both. Those 26 were removed leaving 536 data points . Then, the remaining missing data was imputed by linear regression using bootstrap. The resulting dataset was complete and was used to get the OLS estimators.

The completed data set was used to analyze the data points. The summary of the data was calculated and shown in Table 2. In addition, the ANOVA table was calculated and shown in Table 3. The resulting confidence intervals for 95% and 97.5% are shown in Table 4.

The experiment aimed to test the null hypothesis that the slope was zero. Given the low p-value of 2.2e^-16, the null hypothesis was rejected. Thus, we can conclude that the slope of the data set was not zero. This is shown in the plot of the estimated regression line in Table 5, as the line was not horizontal and seemed to have a positive slope. Thus, this indicates there exists a relationship between the IV and the DV. The correlation coefficient is 0.72842295405.

PART A APPENDIX

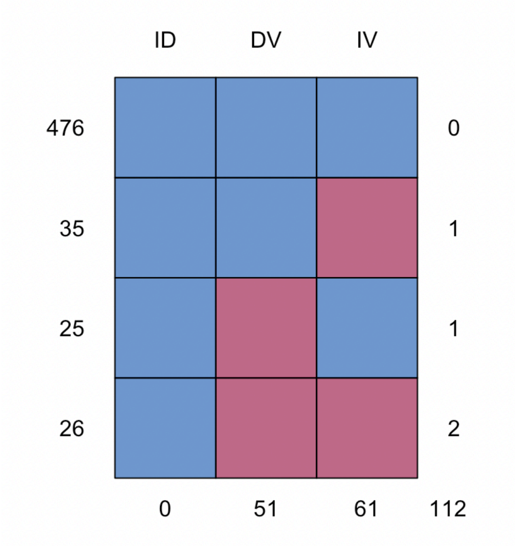


Table 1: Missing Data

```
Call:
lm(formula = DV ~ IV, data = PartA_complete)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.977  -6.934   0.001   6.565  33.754 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  47.5878     1.1675   40.76   <2e-16 ***
IV            5.2439     0.2135   24.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.36 on 534 degrees of freedom
Multiple R-squared:  0.5306,	Adjusted R-squared:  0.5297 
F-statistic: 603.5 on 1 and 534 DF,  p-value: < 2.2e-16
```

Table 2: Simple Regression Summary

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	64795.14	64795.1424	603.5044	0
Residuals	534	57332.81	107.3648	NA	NA

Table 3: ANOVA Table

```
> confint(M, level = 0.95)
               2.5 %    97.5 %
(Intercept) 45.373382 49.989153
IV           4.795851  5.639934
> confint(M, level = 0.99)
               0.5 %    99.5 %
(Intercept) 44.644213 50.718322
IV           4.662508  5.773277
```

Table 4: CI

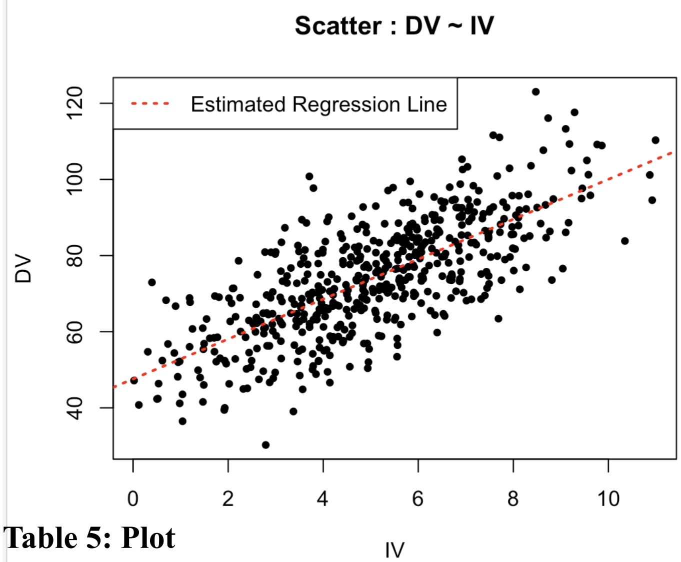


Table 5: Plot

The second part of the project involved the transformation of a data set and the analysis of the subsequent data set. The transformation of the set allowed for a better fit, which we then performed a Lack of Fit test in order to find the best regression model. We are then testing the null hypothesis that the slope is zero to see if there exists a relationship between the two variables.

Firstly, the data set needed to be transformed. The prominent issue was to find the correct transformation. Common transformations of the IV and DV were tested on the 475 observations of the data set. Each transformation was graded on the p-value and R squared of the transformed data set. Each transformation and its results are listed in Table 1. The best transformation seemed to be x^(1/3), as it led to a p-value of 0.6200322 and a R squared value of 0.5905. The IV data points seemed to cluster around intervals of 0.025, so the binning range was 0.025. The cut command was used to create groups with intervals of 0.025. The resulting groups and the frequency of each group are listed in Table 2. These new values are plotted in Table 3, which show a relatively linear relationship. Table 3 shows the before and after of the binning.

The high p-value of 0.6200322 indicates that the null hypothesis of good fit is not rejected. In addition, the R squared value of 0.5905 is relatively higher than the other transformations, meaning that it is a better fit than other options. The low p-value on the F-statistic indicates that the null hypothesis that the slope is zero should be rejected. Thus, this indicates that there exists a relationship between the two variables. The coefficient of correlation between the variables is 0.76896033707.

Part B Appendix

Table 1: Transformation Testing

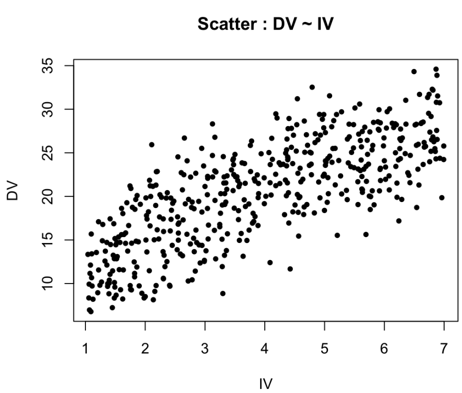
TRANSFORMATION	ASSOCIATED P-VALUE	R2 adjusted
x	0.003616626	0.5734
x^2	0.0003476335	0.5177
x^3	0.0001689719	0.4548
x^(-1)	INF	0.5352
x^(-2)	0.002876507	0.4362
x^(-3)	NaN	0.3384
x^(-½)	NaN	0.5704
x^(-⅓)	NaN	0.5785
x^(½)	0.2991456	0.5886
x^(⅓)	0.899648	0.5905
y	0.003616626	0.5734
y^2	0.06016825	0.5587
y^3	0.05528897	0.5162
y^(-1)	2.798467e-09	0.4747
y^(-2)	1.093478e-10	0.3784
y^(-3)	2.409707e-10	0.2832
y^(-½)	7.653519e-08	0.515
y^(-⅓)	2.737987e-07	0.5264
y^(½)	0.0001776507	0.5655
y^(⅓)	5.288008e-05	0.5603

y^(-⅔)	2.298715e-08	

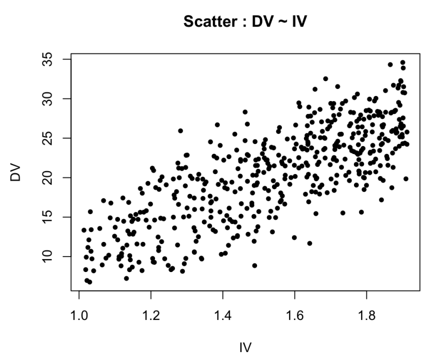
Table 2: Binning Groups

```
groups
(-Inf,1.04] (1.04,1.06] (1.06,1.09] (1.09,1.11] (1.11,1.14] (1.14,1.16]
         11           2           7           6          10          12
(1.16,1.19] (1.19,1.21] (1.21,1.24] (1.24,1.26] (1.26,1.29] (1.29,1.31]
         10          10          10           6          12          11
(1.31,1.34] (1.34,1.36] (1.36,1.39] (1.39,1.41] (1.41,1.44] (1.44,1.46]
         11          10          14          12          11          13
(1.46,1.49] (1.49,1.51] (1.51,1.54] (1.54,1.56] (1.56,1.59] (1.59,1.61]
         14          14          17          14          10          11
(1.61,1.64] (1.64,1.66] (1.66,1.69] (1.69,1.71] (1.71,1.74] (1.74,1.76]
         18          20          14          19          15          13
(1.76,1.79] (1.79,1.81] (1.81,1.84] (1.84,1.86] (1.86, Inf]
         23          19          18          16          42
```

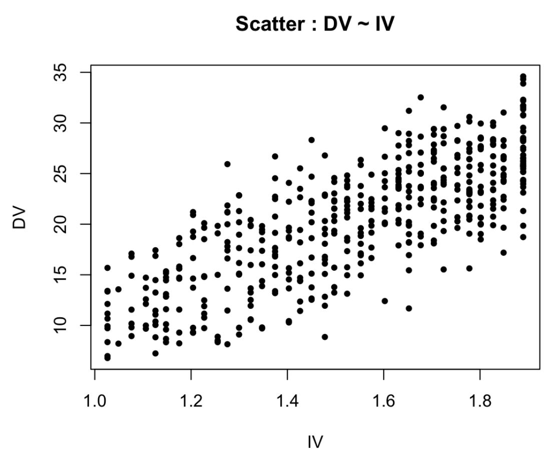
Table 3: Data to Transformed to Binned



Data



Transformed



Binned

Table 4: ANOVA

```
Lack of Fit F Test
--------------
Response :   y
Predictor:   x

                         Analysis of Variance Table
----------------------------------------------------------------------------
                  DF      Sum Sq      Mean Sq      F Value        Pr(>F)
----------------------------------------------------------------------------
x                  1    10097.82     10097.82     675.4672   3.677131e-93
Residual         473    7024.867     14.85173
 Lack of fit      33    447.1355     13.54956    0.9063621      0.6200322
 Pure Error      440    6577.732     14.94939
----------------------------------------------------------------------------
```

Table 5: Simple Regression Summary

```
Call:
lm(formula = ytrans ~ xtrans, data = data_trans)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6437  -2.8444   0.1236   2.6537  10.2251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.8892     1.0966  -7.194 2.48e-12 ***
xtrans       18.4001     0.7033  26.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.846 on 473 degrees of freedom
Multiple R-squared:  0.5913,    Adjusted R-squared:  0.5905
F-statistic: 684.4 on 1 and 473 DF,  p-value: < 2.2e-16
```

Code:

```
PartA_IV <- read.csv('849035_IV.csv', header = TRUE)

PartA_DV <- read.csv('849035_DV.csv', header = TRUE)
PartA <- merge(PartA_IV, PartA_DV, by = 'ID')

#View(PartA)

#check if compiled file has same number of values
#str(PartA)
#str(PartA_IV)
#str(PartA_DV)

#562 observations

PartA_incomplete <- PartA

library(mice)

md.pattern(PartA_incomplete)

#35 only IV missing, 25 only DV missing, 26 both missing. 51 total DV missing, 61 total IV
missing

PartA_imp <- PartA[!is.na(PartA$IV)==TRUE|!is.na(PartA$DV)==TRUE,]

#26 observations removed, PartA_imp has 536 observations

imp <- mice(PartA_imp, method = "norm.boot", printFlag = FALSE)

PartA_complete <- complete(imp)

md.pattern(PartA_complete)

#complete with 536 observations

M <- lm(DV ~ IV, data=PartA_complete)
summary(M)
```

```
##
# Call:
#   lm(formula = DV ~ IV, data = PartA_complete)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -31.977  -6.934   0.001   6.565  33.754
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)  47.5878     1.1675   40.76   <2e-16 ***
#   IV           5.2439     0.2135   24.57   <2e-16 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 10.36 on 534 degrees of freedom
# Multiple R-squared:  0.5306,          Adjusted R-squared:  0.5297
# F-statistic: 603.5 on 1 and 534 DF,  p-value: < 2.2e-16
##
```

```
install.packages('knitr')

library(knitr)

kable(anova(M), caption='ANOVA Table')

# Table: ANOVA Table
#
# |          |  Df|   Sum Sq|    Mean Sq|  F value| Pr(>F)|
#   |:---------|---:|--------:|----------:|--------:|------:|
#   |IV        |   1| 64795.14| 64795.1424| 603.5044|      0|
#   |Residuals | 534| 57332.81|   107.3648|       NA|     NA|


plot(PartA_complete$DV ~ PartA_complete$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV',
pch=20)

abline(M, col='red', lty=3, lwd=2)

legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
```

```
confint(M, level = 0.95)

#              2.5 %    97.5 %
#   (Intercept) 45.294264 49.881275
# IV           4.824536  5.663173

confint(M, level = 0.99)

#              0.5 %    99.5 %
#   (Intercept) 44.569639 50.605900
# IV           4.692054  5.795655


############### PART B ####################################


data <- read.csv('849035_PartB.csv', header = TRUE)

plot(data$y ~ data$x, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)

# WHY CHOOSE TRANS?
data_trans <- data.frame(xtrans=data$x^(1/3), ytrans=data$y)

plot(data_trans$y ~ data_trans$x, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)

groups <- cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.025,
max(data_trans$xtrans)-0.025,by=0.025),Inf))

table(groups)

# groups
# (-Inf,1.04] (1.04,1.06] (1.06,1.09] (1.09,1.11] (1.11,1.14] (1.14,1.16]
# 11          2           7           6           10          12
# (1.16,1.19] (1.19,1.21] (1.21,1.24] (1.24,1.26] (1.26,1.29] (1.29,1.31]
# 10          10          10          6           12          11
# (1.31,1.34] (1.34,1.36] (1.36,1.39] (1.39,1.41] (1.41,1.44] (1.44,1.46]
# 11          10          14          12          11          13
# (1.46,1.49] (1.49,1.51] (1.51,1.54] (1.54,1.56] (1.56,1.59] (1.59,1.61]
# 14          14          17          14          10          11
```

```
# (1.61,1.64] (1.64,1.66] (1.66,1.69] (1.69,1.71] (1.71,1.74] (1.74,1.76]
# 18          20          14          19          15          13
# (1.76,1.79] (1.79,1.81] (1.81,1.84] (1.84,1.86] (1.86, Inf]
# 23          19          18          16          42

x <- ave(data_trans$xtrans, groups)

data_bin <- data.frame(x=x, y=data_trans$ytrans)

plot(data_bin$y ~ data_bin$x, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)

#install.packages("olsrr")
library("olsrr")

fit_b <- lm(y ~ x, data = data_bin)

ols_pure_error_anova(fit_b)

# IF P VALUE LOW, then reject good fit

fit_b_final <- lm(ytrans ~ xtrans, data = data_trans)
summary(fit_b_final)
```