Contents

I. Problem 1

```r
data <- read.csv("train.csv")
my_data <- data[ , c("SalePrice", "LotFrontage", "LotArea", "BsmtFinSF1",
                     "BsmtFinSF2", "BsmtUnfSF", "X1stFlrSF", "X2ndFlrSF",
                     "GrLivArea", "GarageArea", "WoodDeckSF")]
my_data[is.na(my_data)] <- 0
test_data<-read.csv("test.csv")
```

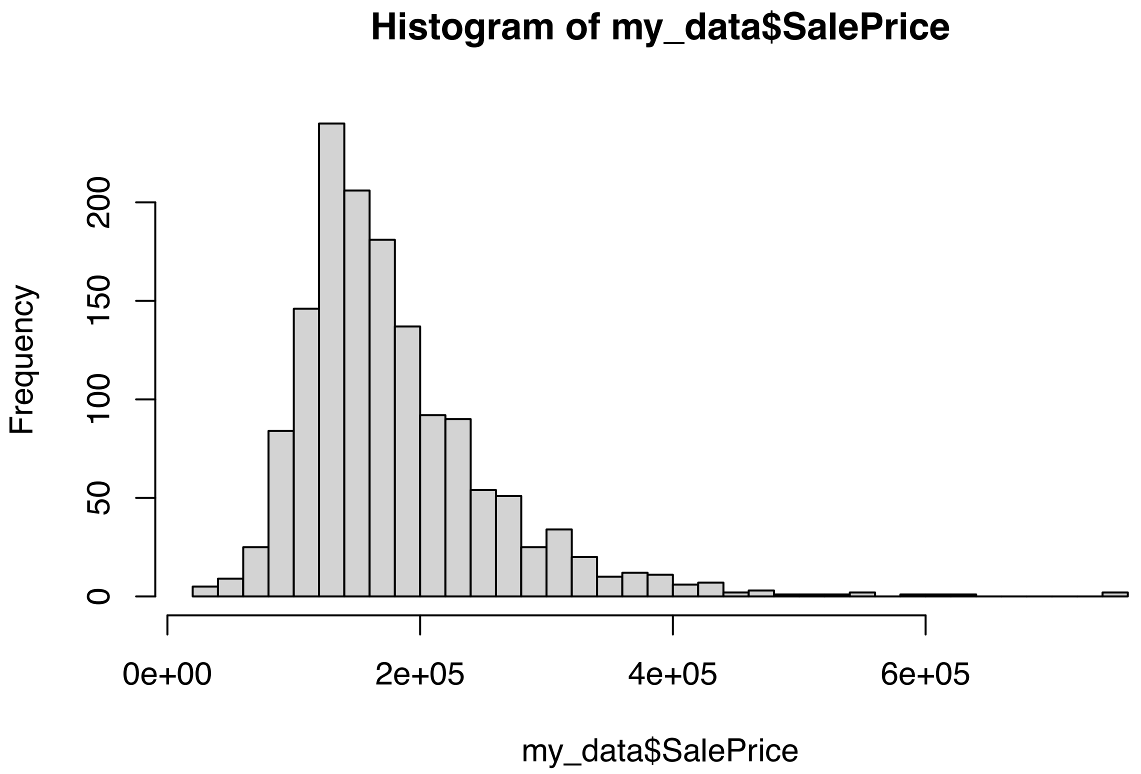
REMARK: The ten variables I am selecting to explain changes in the variable SalePrice are as follows: "LotFrontage", "LotArea", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "X1stFlrSF", "X2ndFlrSF", "GrLivArea", "GarageArea", and "WoodDeckSF". I selected these values as I wanted to isolate the effect only physical characteristics of the home and property. I thought that although this might not yield the best model, it would allow to me isolate the important physical variables, and then I could run another model of ten variables that looks at more intangible characteristics like neighborhood type and general zoning classification which impact home prices but are not directly related to the physical size or shape of the home.

(a)

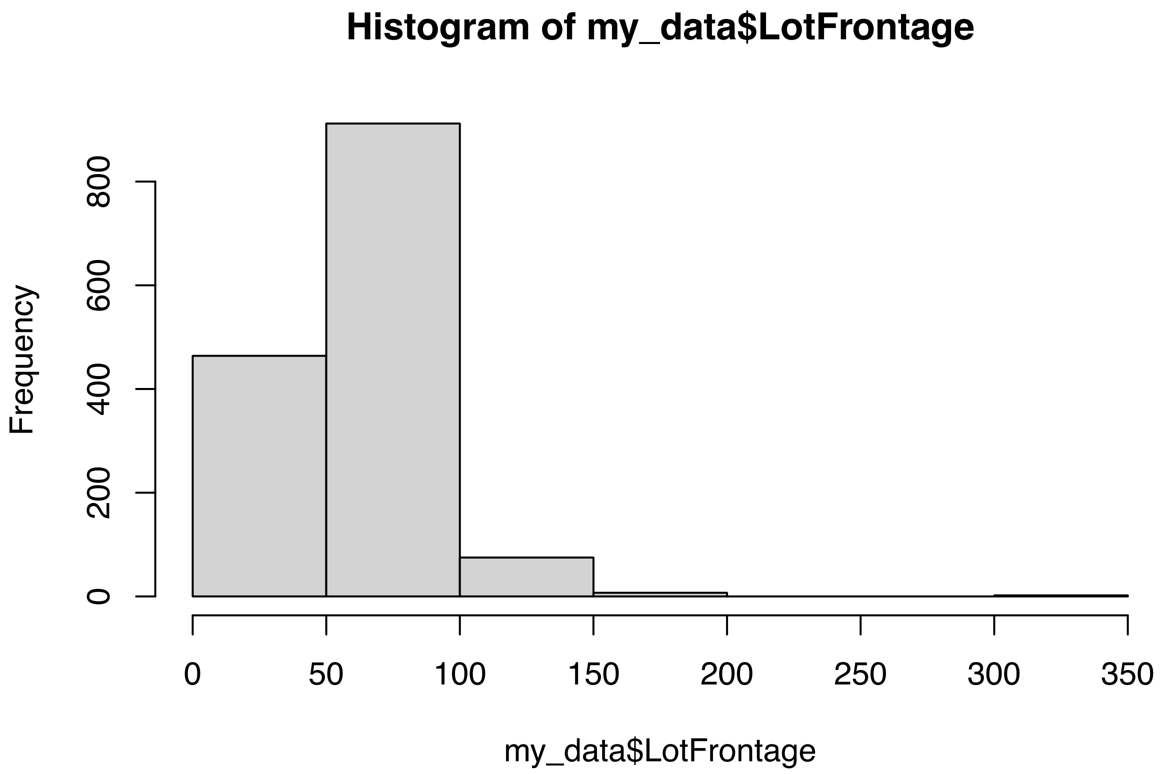
```
#5 number summary for all variables as a part of descriptive statistics
summary(my_data)
```

```
##    SalePrice       LotFrontage        LotArea          BsmtFinSF1    
##  Min.   : 34900   Min.   :  0.00   Min.   :  1300   Min.   :   0.0  
##  1st Qu.:129975   1st Qu.: 42.00   1st Qu.:  7554   1st Qu.:   0.0  
##  Median :163000   Median : 63.00   Median :  9478   Median : 383.5  
##  Mean   :180921   Mean   : 57.62   Mean   : 10517   Mean   : 443.6  
##  3rd Qu.:214000   3rd Qu.: 79.00   3rd Qu.: 11602   3rd Qu.: 712.2  
##  Max.   :755000   Max.   :313.00   Max.   :215245   Max.   :5644.0  
##    BsmtFinSF2        BsmtUnfSF        X1stFlrSF      X2ndFlrSF   
##  Min.   :   0.00   Min.   :   0.0   Min.   : 334   Min.   :   0  
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 882   1st Qu.:   0  
##  Median :   0.00   Median : 477.5   Median :1087   Median :   0  
##  Mean   :  46.55   Mean   : 567.2   Mean   :1163   Mean   : 347  
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1391   3rd Qu.: 728  
##  Max.   :1474.00   Max.   :2336.0   Max.   :4692   Max.   :2065  
##    GrLivArea      GarageArea       WoodDeckSF    
##  Min.   : 334   Min.   :   0.0   Min.   :  0.00  
##  1st Qu.:1130   1st Qu.: 334.5   1st Qu.:  0.00  
##  Median :1464   Median : 480.0   Median :  0.00  
##  Mean   :1515   Mean   : 473.0   Mean   : 94.24  
##  3rd Qu.:1777   3rd Qu.: 576.0   3rd Qu.:168.00  
##  Max.   :5642   Max.   :1418.0   Max.   :857.00
```

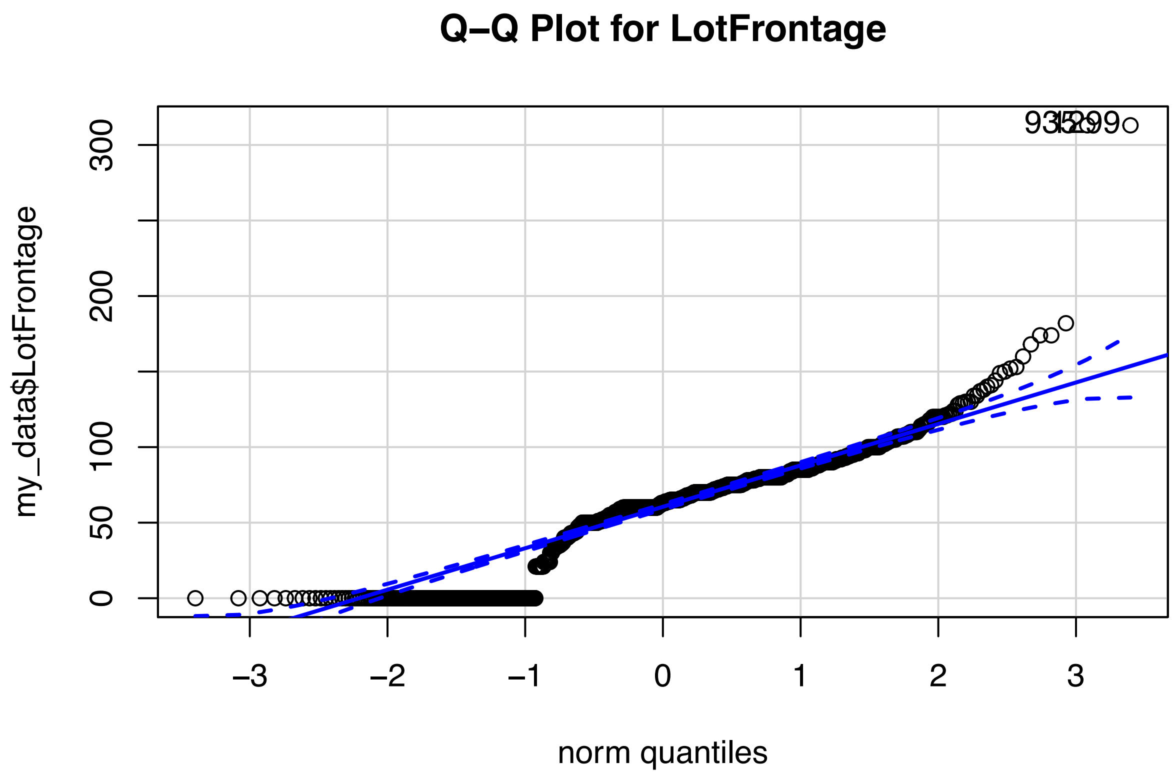
```
#setting optimal bin width
n=1460
k=1+log2(n)

hist(my_data$SalePrice, breaks="FD")
```



```
#For our SalePrice variable we see a fairly right skewed distribution with most
#values occurring around 100,000-150,000.
#We will consider a log transformation.

hist(my_data$LotFrontage, breaks=k)
```



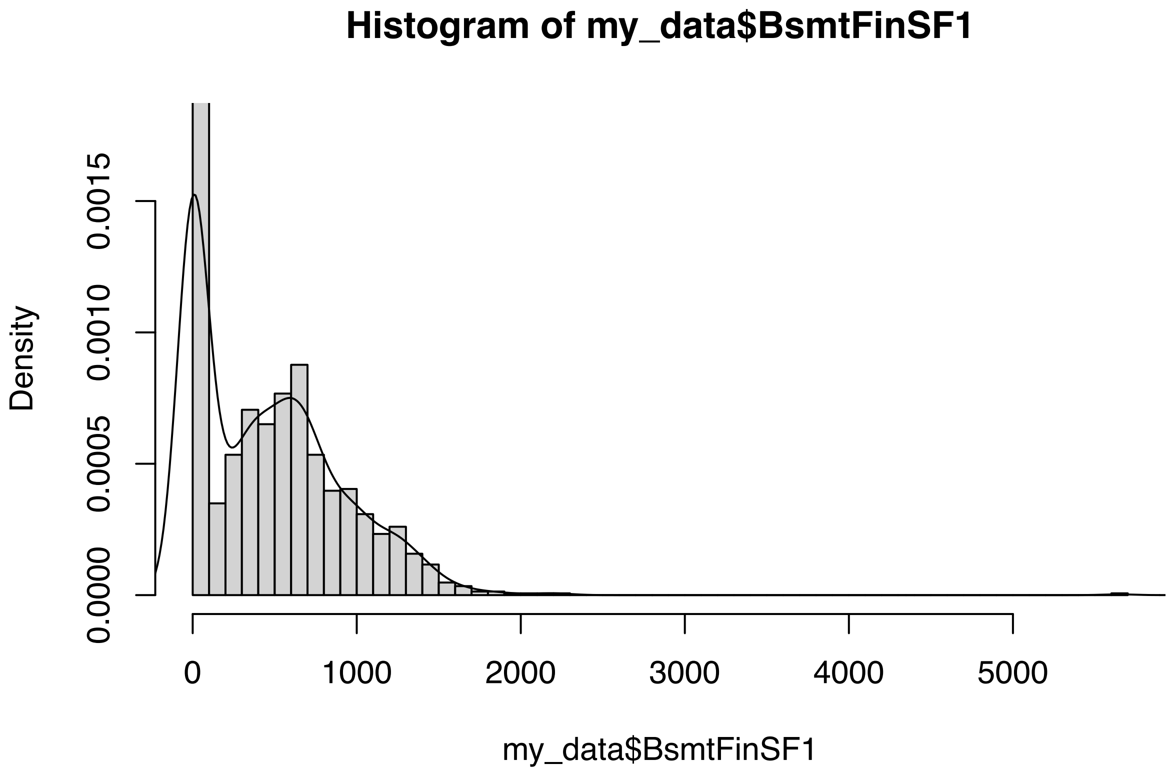
```r
#using k breaks showed the distributions trend more clearly
qqPlot(my_data$LotFrontage, main="Q-Q Plot for LotFrontage")
```



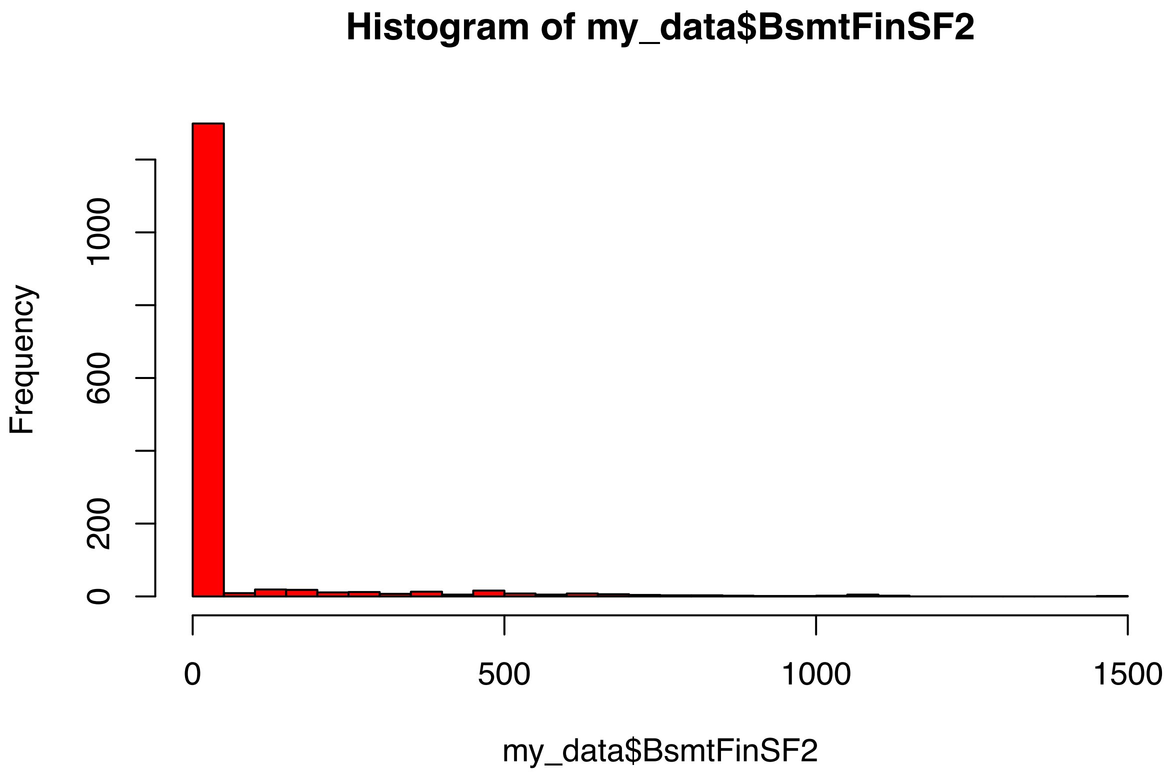
```
## [1]  935 1299
```

```
#This appears to have disagreement in the tails. The upper tail
#specifically is giving us more issues because all of the observations beyond
#1 standard deviation occur outside of normal distribution. By looking at the
#histogram we can see why this is occurring because of the fairly skewed shape
#of our distribution. We will consider a transformation

hist(my_data$BsmtFinSF1, breaks="FD", freq=FALSE, ylim=c(0,.0018))
lines(density(my_data$BsmtFinSF1,lwd = 2, col ="red"))
```

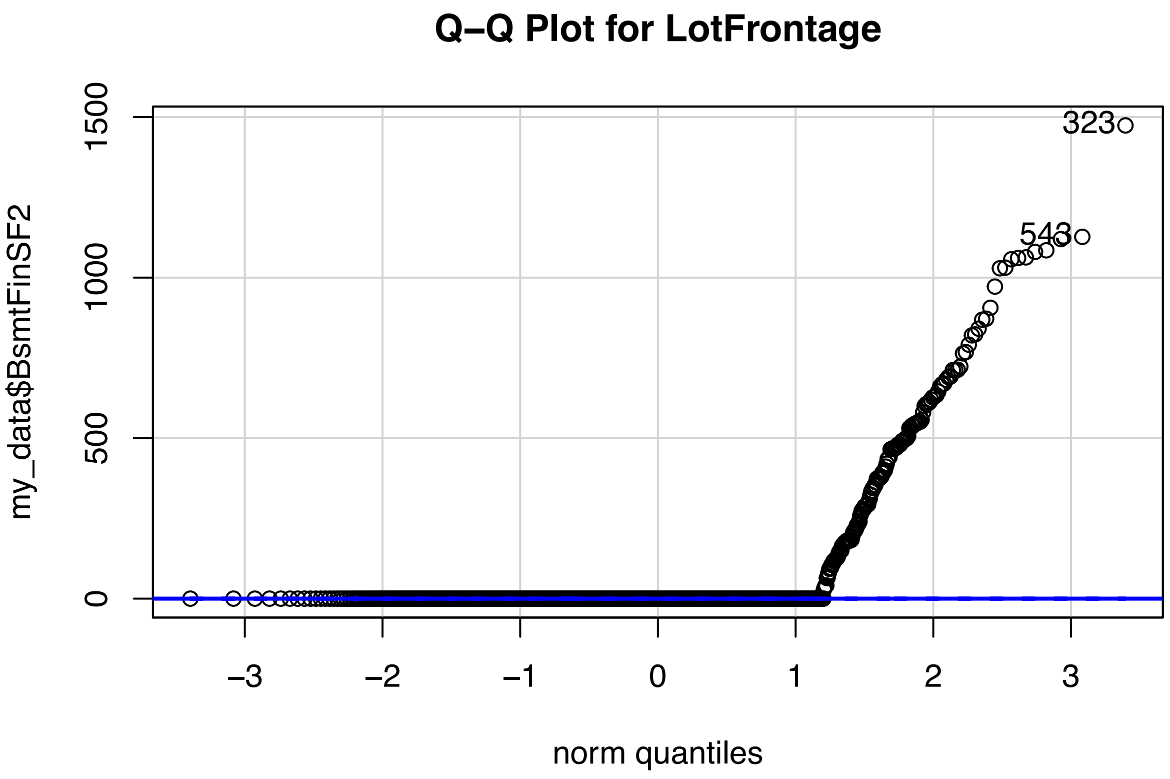
```
## Warning: In density.default(my_data$BsmtFinSF1, lwd = 2, col = "red") :
##  extra arguments 'lwd', 'col' will be disregarded
```



```
#This density plot shows us a right skewed distribution which makes sense with
#a variable like BsmtFinSF1 because we expect the largest number of observations
#to occur close to 0, and then slowly decrease in number as square footage
#increases. We will consider a transformation due to this distribution.

hist(my_data$BsmtFinSF2, breaks="FD", col="red")
```



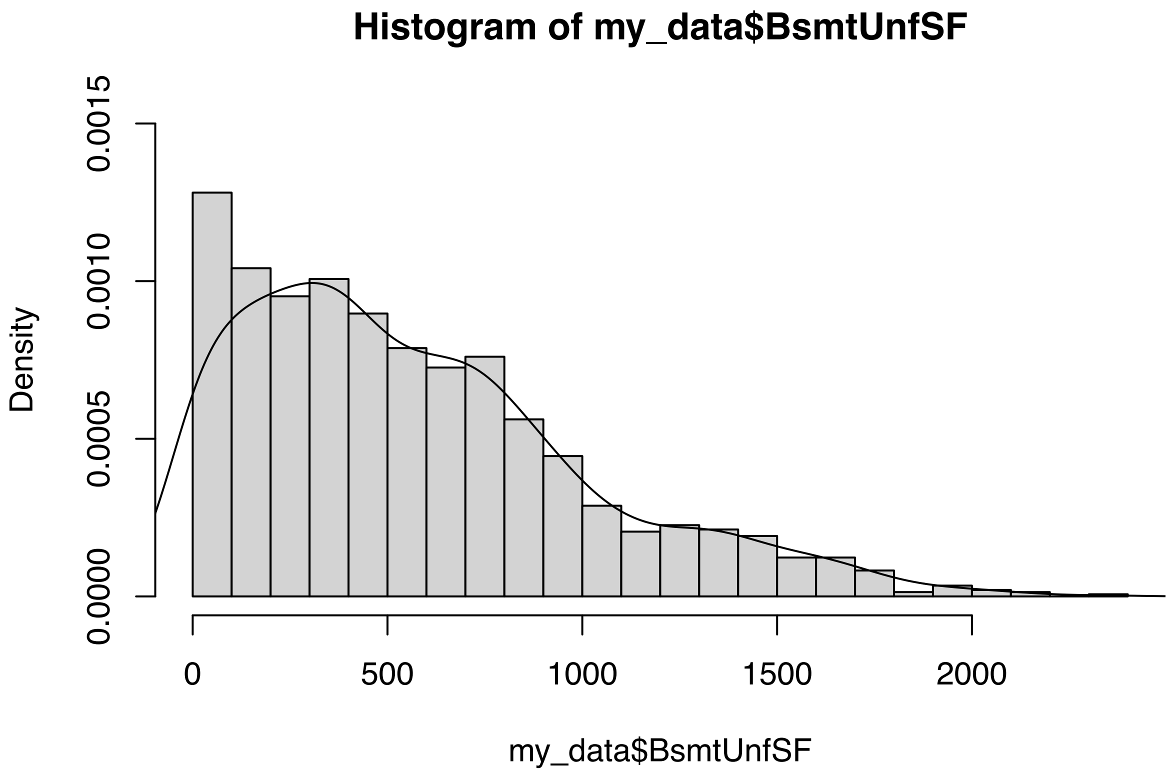
```
#We see that this plot is extremely right skewed in that almost every
#observation occurs within the first bar, and all subsequent observations are
#very limited in terms of numbers so we don't see a gradual "step-down" as
#we move right along this graph. We will consider a transformation.

qqPlot(my_data$BsmtFinSF2, main="Q-Q Plot for LotFrontage")
```



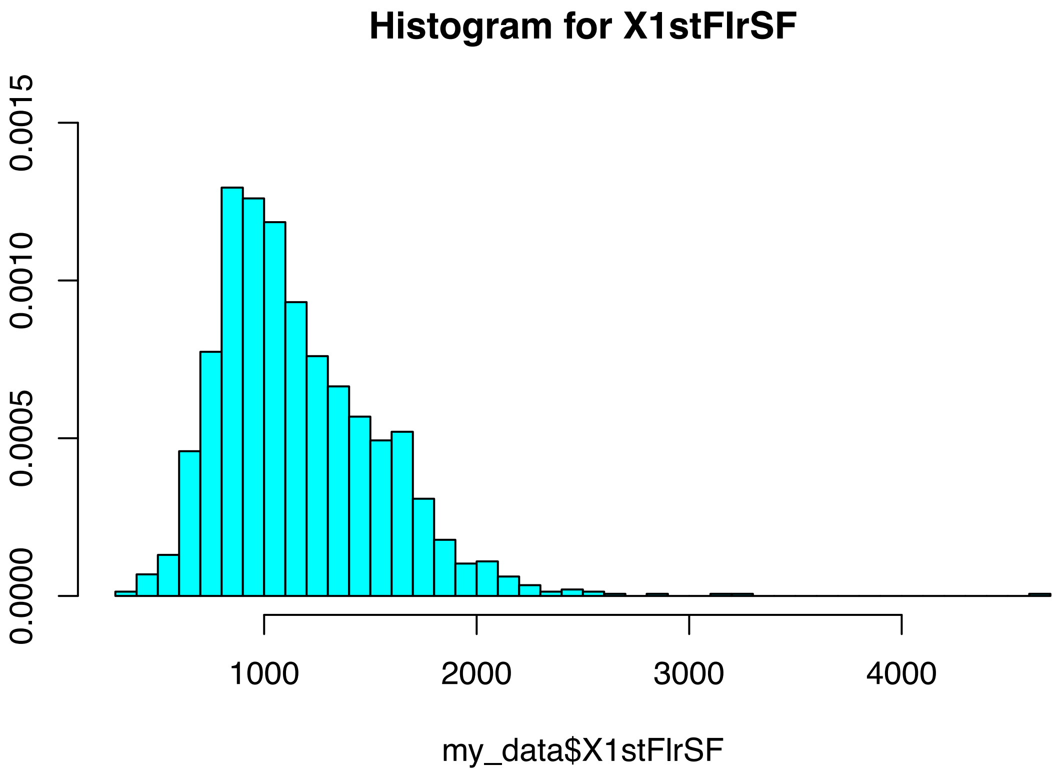
```
## [1] 323 543
```

```
#this is a qqplot confirming the exact observation from our histogram

hist(my_data$BsmtUnfSF, breaks="FD", freq=FALSE, ylim=c(0,.0015))
#Shows a bit of right skew in the distribution. I adjusted the limit and bin
#width to show this adequately. We will need to consider a transformation,
#but this data is more closely resembling a normal distribution than others.
lines(density(my_data$BsmtUnfSF,lwd = 2))
```

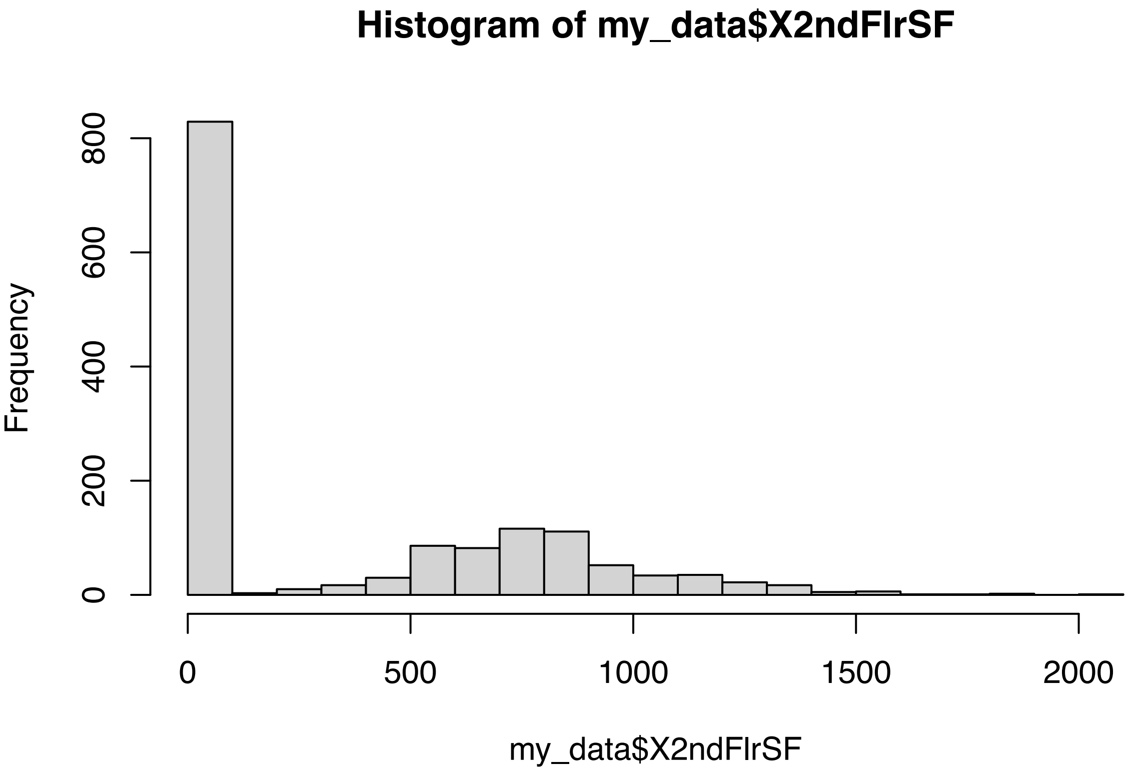
```
## Warning: In density.default(my_data$BsmtUnfSF, lwd = 2) :
##  extra argument ’lwd’ will be disregarded
```



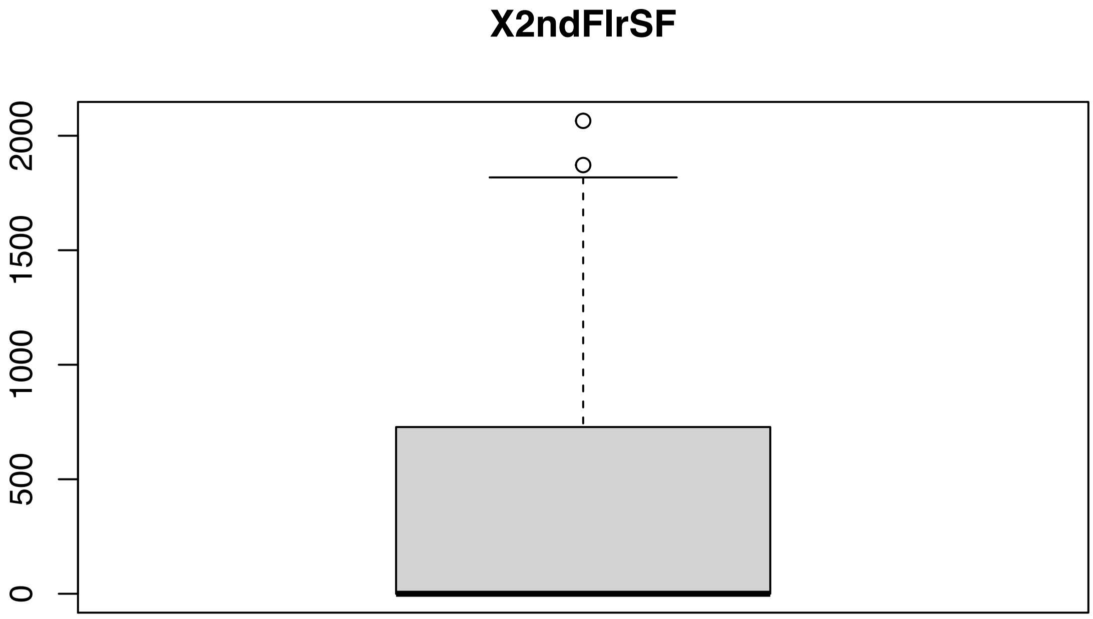
```
truehist(my_data$X1stFlrSF, ylim=c(0,.0015), main="Histogram for X1stFlrSF")
```



```
#similar to most of our data we see that this true histogram appears to be right
#skewed but not by nearly as much as other variables. In fact, this appears to
#follow a normal distribution more closely than any previous variables.

hist(my_data$X2ndFlrSF, breaks="FD")
```

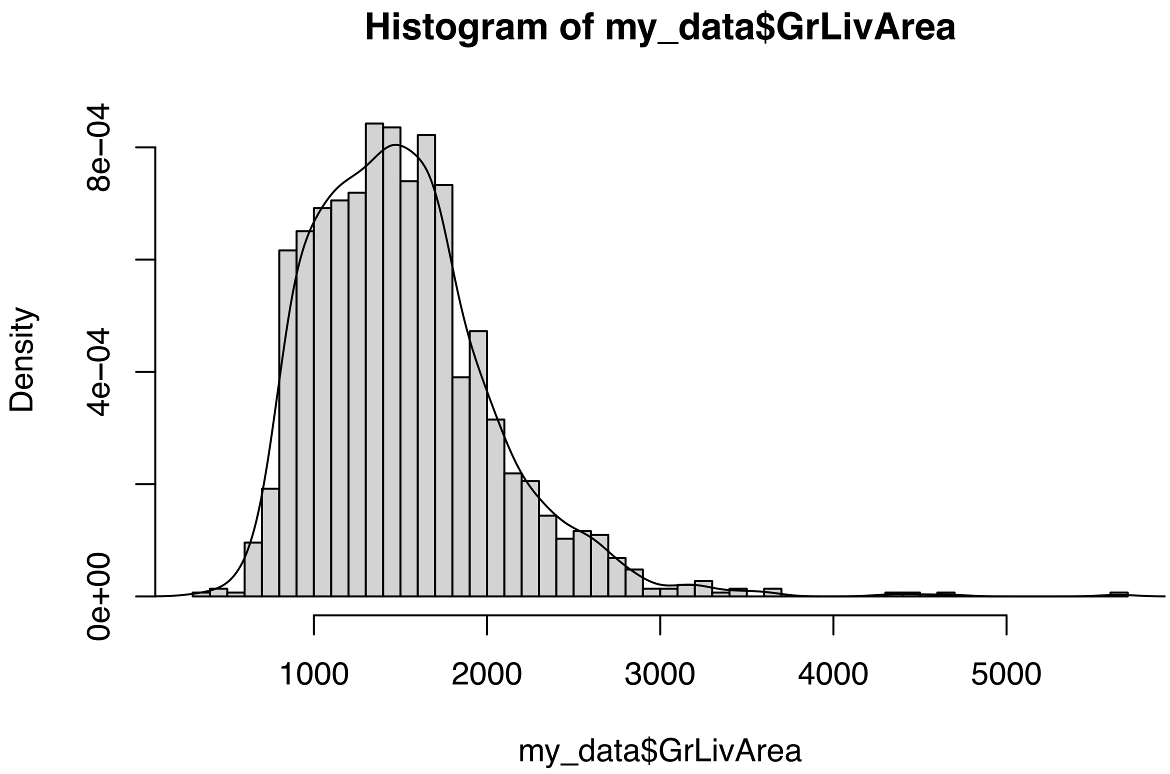


```
#This histogram shows a very unique distribution in that the vast majority of
#observations occur between 0-100, but then for all subsequent observations
#beyond this point we see a fairly normal distribution.
#We will need to consider a transformation.

boxplot(my_data$X2ndFlrSF, main="X2ndFlrSF")
```

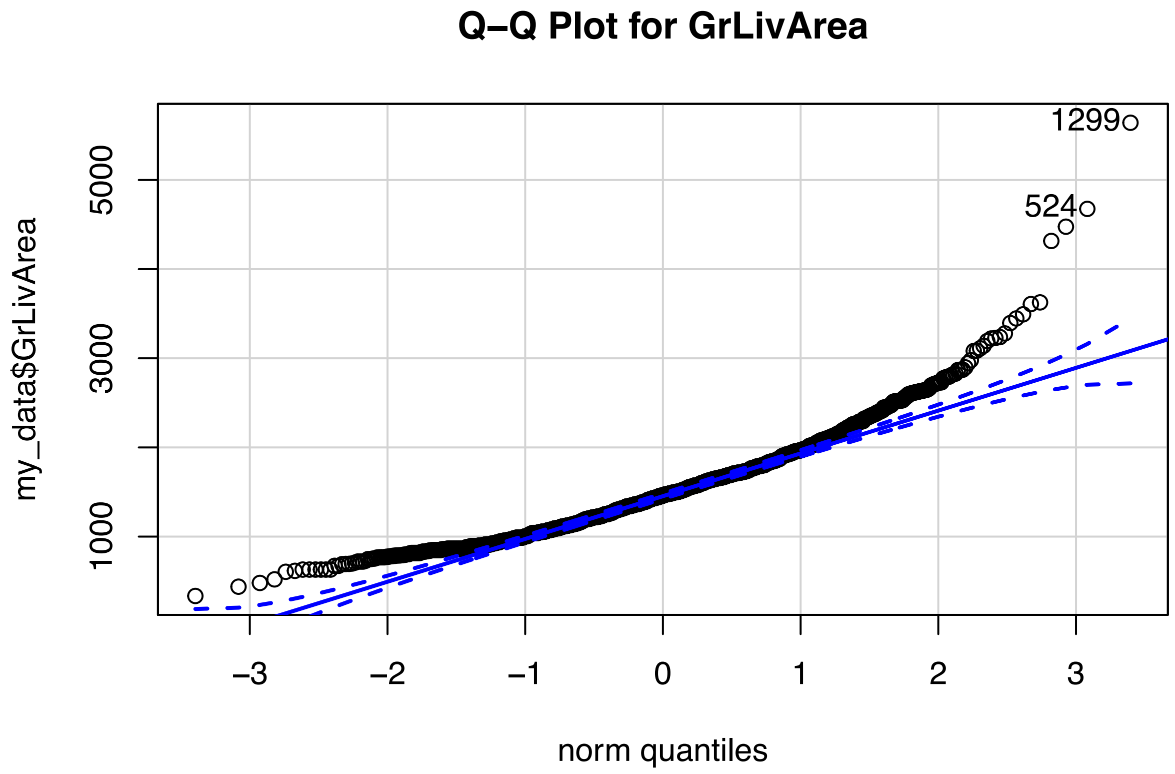


```
#this boxplot confirms the fact that we have an extremely unusual distribution
#and will need some sort of transformation to make this variable workable.

hist(my_data$GrLivArea, breaks="FD", freq=FALSE)
#this density histogram shows another right skewed distribution, but in this
#case we see a much more equitable spread around the peak of 1,500.
#Although it is not terribly skewed we will still consider transforming.

lines(density(my_data$GrLivArea,lwd = 2))
```

```
## Warning: In density.default(my_data$GrLivArea, lwd = 2) :
##  extra argument 'lwd' will be disregarded
```

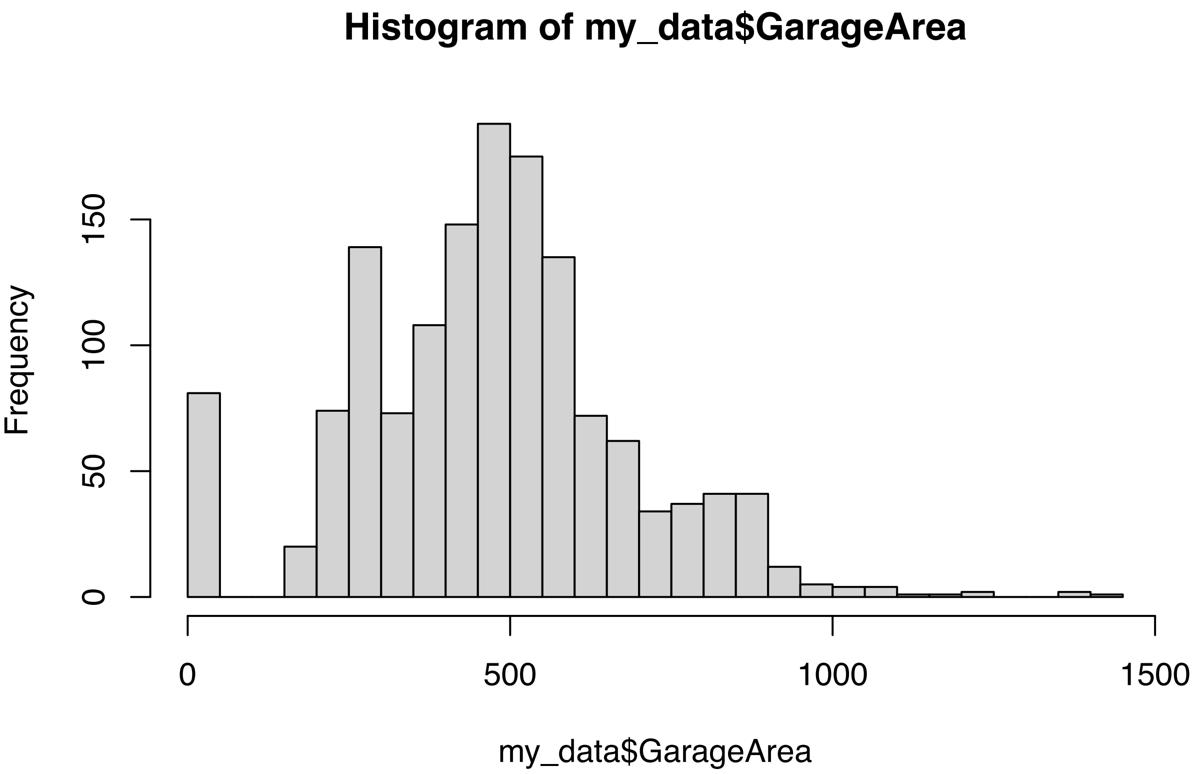


```
qqPlot(my_data$GrLivArea, main="Q-Q Plot for GrLivArea")
```



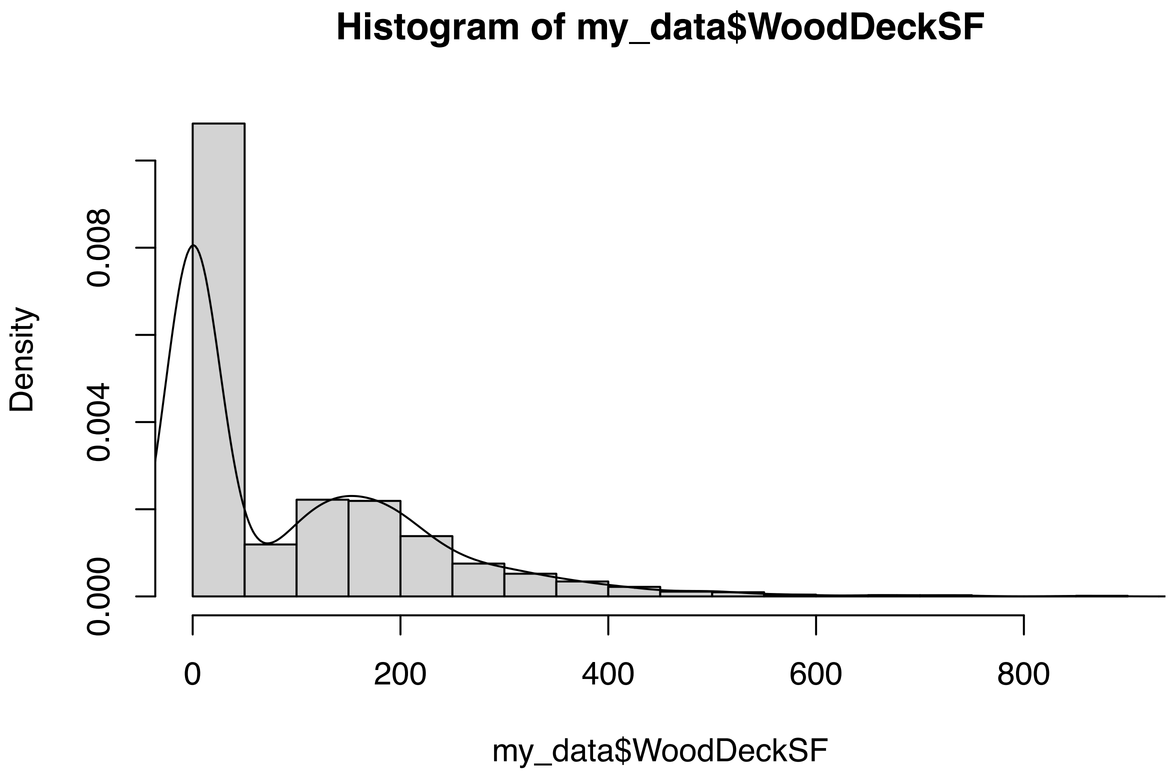
```
## [1] 1299  524
```

```
#this qqplot shows similar trends on both the lower and upper tail, which is
#an upward trend away from the normally distributed reference line .

hist(my_data$GarageArea, breaks="FD")
```

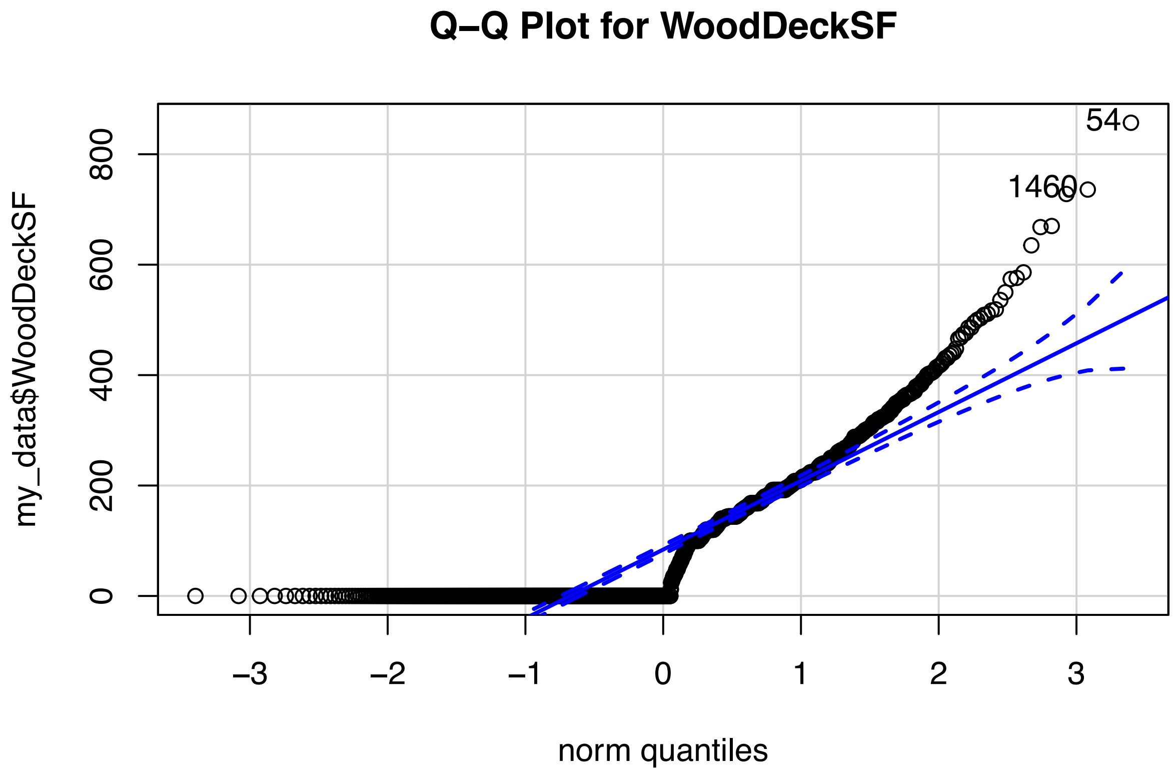


```
#this histogram shows a little right skew, but also has a partial semblance of
#a normal distribution and therefore we could possibly maybe be okay without
#a transformation.

hist(my_data$WoodDeckSF, breaks="FD", freq=FALSE)
#This histogram shows a large cluster of observations occurring right around
#0-50, which makes sense because it is likely that many individuals do not
#have a wooddeck and if they do it could be relatively small. We will therefore
#consider a transformation.

lines(density(my_data$WoodDeckSF,lwd = 2))
```

```
## Warning: In density.default(my_data$WoodDeckSF, lwd = 2) :
##  extra argument 'lwd' will be disregarded
```



```
qqPlot(my_data$WoodDeckSF, main="Q-Q Plot for WoodDeckSF")
```



```
## [1]   54 1460
```

```
#This is a very odd looking qqplot and it shows that the data only follows the
#reference line in quantile 1. We also notice a slight s-shaped in the
#observation cluster.
```

(b)

```
summary(a3 <- powerTransform(cbind(LotFrontage, LotArea, BsmtFinSF1, BsmtFinSF2,
                                BsmtUnfSF, X1stFlrSF, X2ndFlrSF, GrLivArea,
                                GarageArea, WoodDeckSF) ~ 1, data=my_data,
                             family="yjPower"))
```

```
## yjPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## LotFrontage    0.7860        0.79       0.7425       0.8295
## LotArea        0.0193        0.00      -0.0248       0.0634
## BsmtFinSF1     0.2991        0.30       0.2725       0.3258
## BsmtFinSF2    -1.5213       -1.52      -1.6002      -1.4424
## BsmtUnfSF      0.6510        0.65       0.6179       0.6842
## X1stFlrSF      0.1877        0.19       0.1291       0.2463
## X2ndFlrSF      0.0218        0.00      -0.0103       0.0539
## GrLivArea      0.1601        0.16       0.1092       0.2109
## GarageArea     0.8632        0.86       0.8197       0.9066
## WoodDeckSF    -0.0718       -0.07      -0.1044      -0.0391
##
##  Likelihood ratio test that all transformation parameters are equal to 0
##                                              LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0 0 0 0) 10096.49 10 < 2.22e-16
```

```
#Checking for transformations of each variable

transformedY <- yjPower(with(my_data, cbind(LotFrontage, LotArea, BsmtFinSF1,
                                            BsmtFinSF2, BsmtUnfSF, X1stFlrSF,
                                            X2ndFlrSF, GrLivArea, GarageArea,
                                            WoodDeckSF)),
coef(a3, round=TRUE))
#This is transforming the variables based on what was recommended transformation

#Saving Transformed, indexed variables so we can put into new data frame
transformedYLotFrontage<-transformedY[, 1]
transformedYLotArea<-transformedY[, 2]
transformedYBsmtFinSF1<-transformedY[, 3]
transformedYBsmtSF2<-transformedY[, 4]
transformedYBsmtUnfSF<-transformedY[, 5]
transformedYX1stFlrSF<-transformedY[, 6]
transformedYX2ndFlrSF<-transformedY[, 7]
transformedYGrLivArea<-transformedY[, 8]
transformedYGarageArea<-transformedY[, 9]
transformedYWoodDeckSF<-transformedY[, 10]

transformedYSalePrice <- log(my_data$SalePrice) #Log transforming SalePrice
```

```
df_transformedY <- data.frame(cbind(transformedYSalePrice,transformedYLotFrontage,transformedYLotArea,t
#Saving transformed variables into new data frame
```

Remark: I chose to apply the suggested transformations to all my variables except I hard-coded a log transformation to my dependent variable SalePrice. It is worth noting that I could have also hard-coded a log transformation for the variables LotArea & X2ndFlrSF as the suggested transformation for these two variables was close to 0. Applying the exact power transformation to the independent variables will increase accuracy, but likely hurt our ability to easily interpret the estimates.

(c)

```
base.mod <- lm(transformedYSalePrice ~ transformedYLotFrontage +
                 transformedYLotArea + transformedYBsmtFinSF1 +
                 transformedYBsmtSF2 + transformedYBsmtUnfSF +
                 transformedYX1stFlrSF + transformedYX2ndFlrSF +
                 transformedYGrLivArea + transformedYGarageArea +
                 transformedYWoodDeckSF, data=my_data)
S(base.mod)
```

```
## Call: lm(formula = transformedYSalePrice ~ transformedYLotFrontage +
##          transformedYLotArea + transformedYBsmtFinSF1 + transformedYBsmtSF2 +
##          transformedYBsmtUnfSF + transformedYX1stFlrSF + transformedYX2ndFlrSF +
##          transformedYGrLivArea + transformedYGarageArea + transformedYWoodDeckSF, data =
##          my_data)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.297e+00  1.194e-01  69.505  < 2e-16 ***
## transformedYLotFrontage -3.005e-04  3.589e-04  -0.837 0.402492
## transformedYLotArea      3.269e-02  1.230e-02   2.658 0.007949 **
## transformedYBsmtFinSF1   1.058e-02  8.231e-04  12.850  < 2e-16 ***
## transformedYBsmtSF2     -3.307e-02  2.755e-02  -1.201 0.230111
## transformedYBsmtUnfSF    1.307e-03  1.662e-04   7.863 7.25e-15 ***
## transformedYX1stFlrSF   -5.801e-02  1.700e-02  -3.412 0.000662 ***
## transformedYX2ndFlrSF   -3.079e-02  6.150e-03  -5.006 6.25e-07 ***
## transformedYGrLivArea    2.759e-01  2.158e-02  12.790  < 2e-16 ***
## transformedYGarageArea   1.184e-03  7.005e-05  16.898  < 2e-16 ***
## transformedYWoodDeckSF   2.373e-02  2.626e-03   9.036  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard deviation: 0.2077 on 1449 degrees of freedom
## Multiple R-squared: 0.7315
## F-statistic: 394.8 on 10 and 1449 DF,  p-value: < 2.2e-16
##     AIC     BIC
## -433.20 -369.77
```

```
Confint(base.mod)
```

```
##                               Estimate         2.5 %        97.5 %
## (Intercept)               8.2971934009  8.0630269651  8.5313598367
## transformedYLotFrontage -0.0003005373 -0.0010045233  0.0004034487
## transformedYLotArea      0.0326932559  0.0085649249  0.0568215868
## transformedYBsmtFinSF1   0.0105777522  0.0089630727  0.0121924317
## transformedYBsmtSF2     -0.0330747626 -0.0871149367  0.0209654115
## transformedYBsmtUnfSF    0.0013070237  0.0009809722  0.0016330753
## transformedYX1stFlrSF   -0.0580120849 -0.0913612154 -0.0246629543
## transformedYX2ndFlrSF   -0.0307874859 -0.0428522001 -0.0187227716
## transformedYGrLivArea    0.2759470091  0.2336235977  0.3182704205
## transformedYGarageArea   0.0011836982  0.0010462891  0.0013211073
## transformedYWoodDeckSF   0.0237271099  0.0185762246  0.0288779953
```

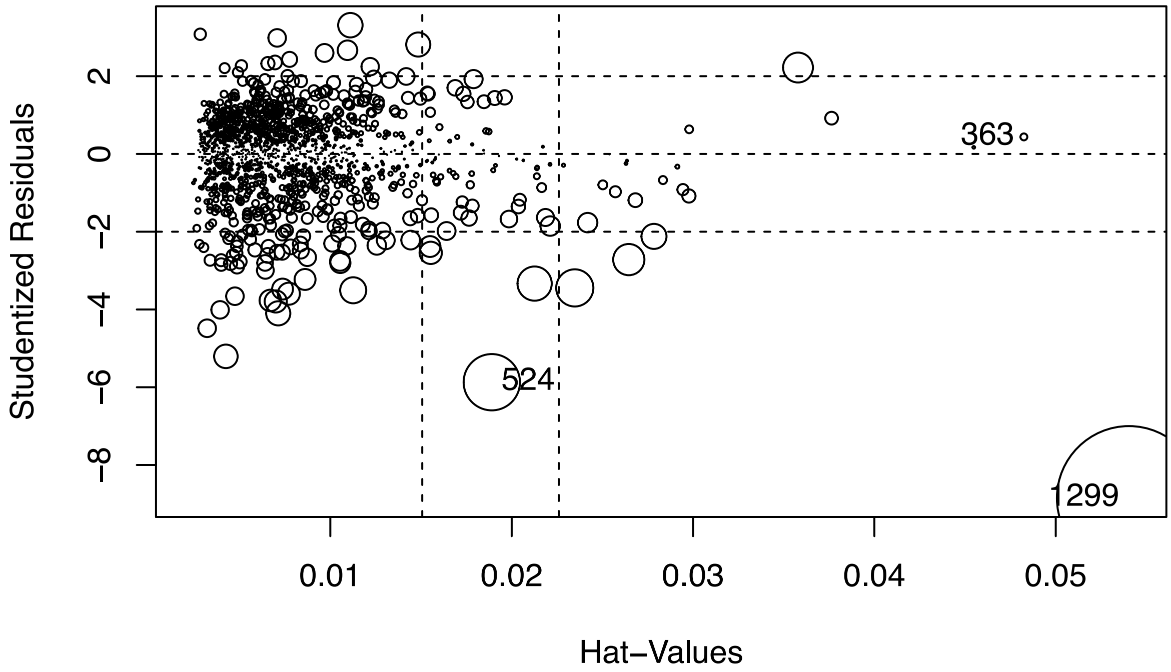
```
#1st floor square footage has confidence interval that crosses 0 bad for
#regression because there is no clear trend
```

Remark: We see that transformedYLotFrontage has a p-value that is slightly above the significance level of 0.05 and transformedYBsmtSF2 has a p-value that is extremely high. Both indicating that they are likely statistically insignificant predictors. We also confirm that their confidence intervals cross the zero value indicating no statistical significance. All other estimates show statistical significance at all levels. I chose to leave in the predictors with insignificant effects (high p-values) because we have a relatively large sample. By applying the transformations above we improve accuracy, but sacrifice the interpretability of our coefficients. Therefore, my below interpretations do not include the exact transformation interpretations, but rather just reference the transformed variables as is.

Interpretations: A unit change in our transformedYLotFrontage will lead to approximately on average a -0.03% change in transformedYSalePrice. A unit change in our transformedYLotArea will lead to approximately on average a 3.27% change in transformedYSalePrice. A unit change in our transformedYBsmtFinSF1 will lead to approximately on average a 1.06% change in transformedYSalePrice. A unit change in our transformedYBsmtFinSF2 will lead to approximately on average a -3.31% change in transformedYSalePrice. A unit change in our transformedYBsmtUnfSF will lead to approximately on average a 0.13% change in transformedYSalePrice. A unit change in our transformedYX1stFlrSF will lead to approximately on average a -5.80% change in transformedYSalePrice. A unit change in our transformedYX2ndFlrSF will lead to approximately on average a -3.08% change in transformedYSalePrice. A unit change in our transformedYGrLivArea will lead to approximately on average a 27.59% change in transformedYSalePrice. A unit change in our transformedYGarageArea will lead to approximately on average a 0.12% change in transformedYSalePrice. A unit change in our transformedYWoodDeckSF will lead to approximately on average a 2.37% change in transformedYSalePrice.

(d)

```
influencePlot(base.mod) #testing for influential observations
```



```
##         StudRes        Hat        CookD
## 363   0.4386193 0.04824320 0.0008870233
## 524  -5.8705730 0.01890311 0.0590029790
## 1299 -8.8511083 0.05403789 0.3862280030
```

```
base.mod.nout<-update(base.mod, subset=-c(524, 1299))
#Removing the 4 influential observations
summary(base.mod.nout)
```

```
##
## Call:
## lm(formula = transformedYSalePrice ~ transformedYLotFrontage +
##     transformedYLotArea + transformedYBsmtFinSF1 + transformedYBsmtSF2 +
##     transformedYBsmtUnfSF + transformedYX1stFlrSF + transformedYX2ndFlrSF +
##     transformedYGrLivArea + transformedYGarageArea + transformedYWoodDeckSF,
##     data = my_data, subset = -c(524, 1299))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05359 -0.09279  0.01875  0.12326  0.64719
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.1883211  0.1152565  71.044  < 2e-16 ***
## transformedYLotFrontage 0.0001371  0.0003478   0.394 0.693374
## transformedYLotArea     0.0351866  0.0118313   2.974 0.002988 **
## transformedYBsmtFinSF1  0.0109566  0.0007924  13.828  < 2e-16 ***
## transformedYBsmtSF2    -0.0401462  0.0264989  -1.515 0.129987
## transformedYBsmtUnfSF   0.0012824  0.0001599   8.020 2.16e-15 ***
## transformedYX1stFlrSF  -0.0328793  0.0165127  -1.991 0.046652 *
## transformedYX2ndFlrSF  -0.0223835  0.0059651  -3.752 0.000182 ***
## transformedYGrLivArea    0.2527561  0.0208627  12.115  < 2e-16 ***
## transformedYGarageArea   0.0011967  0.0000674  17.754  < 2e-16 ***
## transformedYWoodDeckSF   0.0238021  0.0025251   9.426  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1997 on 1447 degrees of freedom
## Multiple R-squared:  0.7521, Adjusted R-squared:  0.7504
## F-statistic:   439 on 10 and 1447 DF,  p-value: < 2.2e-16
```

Remark: I utilized the influencePlot() function which printed 3 influential observations and from the plot we can see that they are all outliers and high leverage. All the observations except for #363 were shown to have a large influence on the model, which is noted by the small density of that observations circle. As a result, I decided to remove all the observations except for #363. Once these outliers were removed we also saw an increase in the predictive power of our model.

(e)

```
#Creating mallows cp function
mallows_cp = function(model1, model2){
  n = nrow(model1$model)
  p1 = length(coef(model1))
  p2 = length(coef(model2))
  if(p2<p1)
    stop('You have interchanged the full model and the subset model',
         call. = FALSE)
  sum(resid(model1)**2) / sum(resid(model2)^2) *(n-p2) + 2 * p1 -n
}

#Creating function to replicate step(), but use mallows cp instead of AIC
mystep = function(object){
  reduced_object = object
  old_mcp = mallows_cp(object, object)
  while(TRUE){
    nms = attr(terms(reduced_object),"term.labels")
    u = lapply(nms, function(x) update(reduced_object, paste0(".~ .-", x)))
    mcp = sapply(u, mallows_cp, object)
    # same as sapply(u, function(x) mallows_cp(x, object))
    if(min(mcp) > old_mcp) break
    old_mcp = min(mcp)
    reduced_object = u[[which.min(mcp)]]
  }
  reduced_object
}

final.model<-mystep(base.mod.nout) #Saving new model with removed variable

summary(final.model)
```

```
##
## Call:
```

```
## lm(formula = transformedYSalePrice ~ transformedYLotArea + transformedYBsmtFinSF1 +
##     transformedYBsmtSF2 + transformedYBsmtUnfSF + transformedYX1stFlrSF +
##     transformedYX2ndFlrSF + transformedYGrLivArea + transformedYGarageArea +
##     transformedYWoodDeckSF, data = my_data, subset = -c(524,
##     1299))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05373 -0.09247  0.01764  0.12303  0.64894
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.184e+00  1.146e-01  71.403  < 2e-16 ***
## transformedYLotArea     3.581e-02  1.172e-02   3.055 0.002291 **
## transformedYBsmtFinSF1  1.095e-02  7.919e-04  13.826  < 2e-16 ***
## transformedYBsmtSF2    -4.028e-02  2.649e-02  -1.521 0.128559
## transformedYBsmtUnfSF   1.286e-03  1.596e-04   8.062 1.56e-15 ***
## transformedYX1stFlrSF  -3.286e-02  1.651e-02  -1.991 0.046685 *
## transformedYX2ndFlrSF  -2.243e-02  5.962e-03  -3.761 0.000176 ***
## transformedYGrLivArea   2.529e-01  2.085e-02  12.130  < 2e-16 ***
## transformedYGarageArea  1.198e-03  6.727e-05  17.813  < 2e-16 ***
## transformedYWoodDeckSF  2.373e-02  2.518e-03   9.425  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1997 on 1448 degrees of freedom
## Multiple R-squared:  0.7521, Adjusted R-squared:  0.7505
## F-statistic:   488 on 9 and 1448 DF,  p-value: < 2.2e-16
```

```r
vif(final.model)
```

```
##    transformedYLotArea transformedYBsmtFinSF1    transformedYBsmtSF2
##               1.325297               2.114189               1.123829
##  transformedYBsmtUnfSF  transformedYX1stFlrSF  transformedYX2ndFlrSF
##               2.309901              13.717548              14.080624
##  transformedYGrLivArea transformedYGarageArea transformedYWoodDeckSF
##              17.750993               1.468960               1.081088
```

```r
#looking for multicollinearity between the variables in the final model without transformedYBsmtFinSF2

mcollin1<-lm(transformedYSalePrice~transformedYLotFrontage +
               transformedYLotArea + transformedYBsmtFinSF1 +
               transformedYBsmtUnfSF + transformedYGrLivArea +
               transformedYGarageArea + transformedYWoodDeckSF,
             subset=-c(524, 1299), data=my_data)
vif(mcollin1)
```

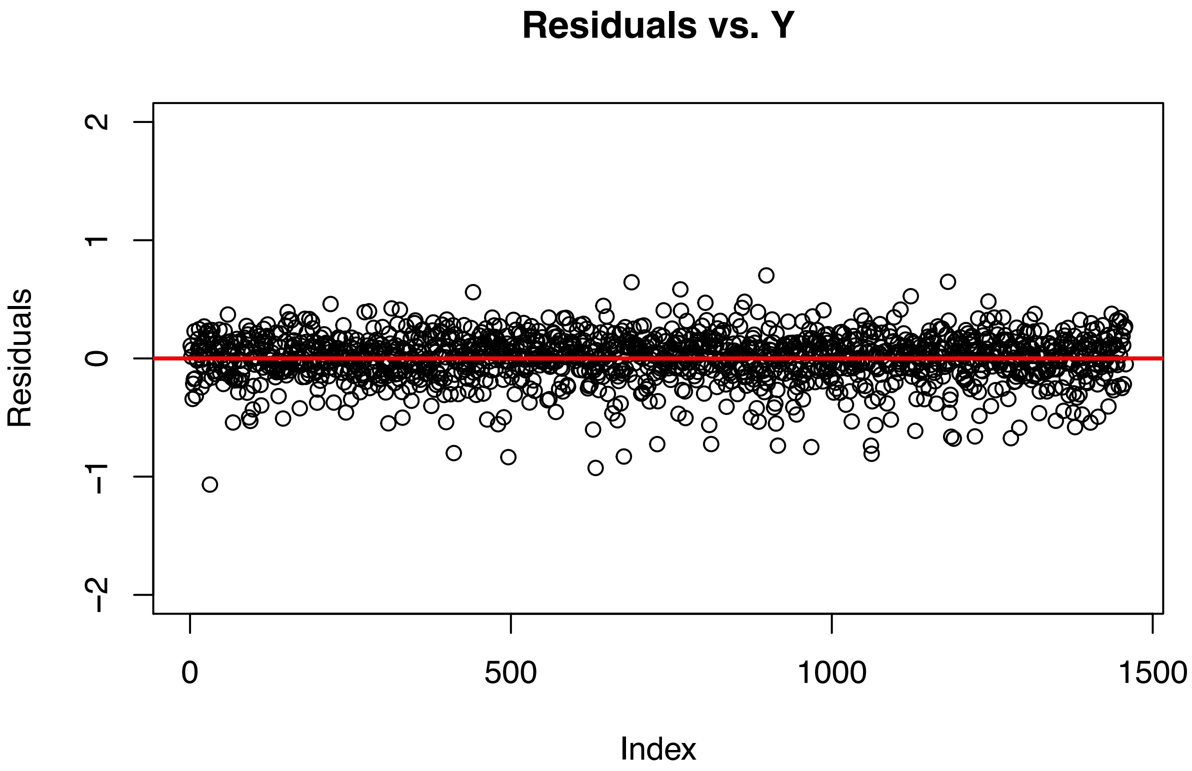
```
## transformedYLotFrontage     transformedYLotArea  transformedYBsmtFinSF1
##                1.076237                1.240547                1.677384
##   transformedYBsmtUnfSF   transformedYGrLivArea  transformedYGarageArea
##                1.747810                1.481480                1.411894
##  transformedYWoodDeckSF
##                1.080737
```

```
#testing multicollinearity for model without transformedYBsmtFinSF2,
#transformedYX1stFlrSF, transformedY2ndFlrSF
final.model<-mcollin1
#saving the model that removed 3 variables and outliers as my final.model.
```

Remark: Using this created mystep() function gives us the best model using Mallows Cp, instead of AIC which the built-in step() function uses. This analysis identified that transformedYBsmtFinSF2 was not a necessary predictor in our model. This estimate does not account for multicollinearity though and it appears that there is very high multicollinearity between some variables when we run our first VIF. When I remove the variables transformedYX1stFlrSF and transformedYX2ndFlrSF, then run VIF again we see that this multicollinearity is eliminated and all the results are substantially below a value of 4. Intuitively, I removed both the 1st and 2nd floor square footage predictors because I realized they would of course be highly colinear with above ground square footage (GrLivArea) since they are essentially measuring the same thing. As I moved along in this modeling process, I realized how poor my variable selection was, but it allowed me to adequately piece together where my model was flawed and justify the removal of unnecessary variables which was a good exercise. Although I was apprehensive to remove predictors, I realized due to this extreme redundancy I could explain SalePrice equally as well with fewer predictors in the model.

(f)

```
#plotting residuals against y
plot(final.model$residuals,pch=1, ylab="Residuals", ylim=c(-2,2),
     main="Residuals vs. Y")

abline(h=0 ,lwd=2, col="red")
```



Remark: These results are promising in that they shows a fairly even spread of residuals around a value of

0.

(g)

```
AIC(base.mod, final.model) #Checking AIC of base.mod vs final.model
```

```
## Warning in AIC.default(base.mod, final.model): models are not all fitted to the
## same number of observations

##             df       AIC
## base.mod    12 -433.2013
## final.model  9 -525.3328
```

```
BIC(base.mod, final.model) #Checking BIC of base.mod vs. final.model
```

```
## Warning in BIC.default(base.mod, final.model): models are not all fitted to the
## same number of observations

##             df       BIC
## base.mod    12 -369.7670
## final.model  9 -477.7694
```

Remark: My final model is a better model since we observe smaller AIC and BIC values, which makes sense considering we removed influential observations and the transformed variables BsmtFinSF2, X1stFlrSF, and X2ndFlrSF. That being said, it is clear this choice of predictors was quite poor.

(h)

```
model_interactions = lm(transformedYSalePrice ~ (transformedYLotFrontage
                                + transformedYLotArea + transformedYBsmtFinSF1
                                + transformedYBsmtUnfSF + transformedYGrLivArea
                          + transformedYGarageArea + transformedYWoodDeckSF)^2,
                        data=my_data, subset=-c(524, 1299))

final.model_interactions<-mystep(model_interactions)
#using mystep() function to eliminate any unneeded terms from our model
vif(final.model_interactions)
```

```
##                          transformedYLotFrontage
##                                       321.582557
##                              transformedYLotArea
##                                        17.091433
##                           transformedYBsmtFinSF1
##                                        17.406215
##                            transformedYBsmtUnfSF
##                                        18.145036
##                           transformedYGarageArea
##                                      399.741213
##    transformedYLotFrontage:transformedYLotArea
##                                      350.505240
## transformedYLotFrontage:transformedYBsmtFinSF1
##                                       12.663355
##  transformedYLotFrontage:transformedYBsmtUnfSF
##                                       14.326400
## transformedYLotFrontage:transformedYGarageArea
##                                       20.603638
##      transformedYLotArea:transformedYGrLivArea
##                                       17.562312
##     transformedYLotArea:transformedYGarageArea
##                                      502.679301
##     transformedYLotArea:transformedYWoodDeckSF
##                                        4.258043
##   transformedYBsmtFinSF1:transformedYBsmtUnfSF
##                                        4.316070
##  transformedYBsmtFinSF1:transformedYGarageArea
##                                       17.522940
##   transformedYBsmtUnfSF:transformedYGarageArea
##                                       23.155748
##   transformedYBsmtUnfSF:transformedYWoodDeckSF
##                                        5.478834
##   transformedYGarageArea:transformedYGrLivArea
##                                      237.160475
```

```
#testing for multicollinearity within this final model including interactions
summary(final.model_interactions)
```

```
##
## Call:
## lm(formula = transformedYSalePrice ~ transformedYLotFrontage +
##     transformedYLotArea + transformedYBsmtFinSF1 + transformedYBsmtUnfSF +
##     transformedYGarageArea + transformedYLotFrontage:transformedYLotArea +
##     transformedYLotFrontage:transformedYBsmtFinSF1 + transformedYLotFrontage:transformedYBsmtUnfSF +
##     transformedYLotFrontage:transformedYGarageArea + transformedYLotArea:transformedYGrLivArea +
##     transformedYLotArea:transformedYGarageArea + transformedYLotArea:transformedYWoodDeckSF +
##     transformedYBsmtFinSF1:transformedYBsmtUnfSF + transformedYBsmtFinSF1:transformedYGarageArea +
##     transformedYBsmtUnfSF:transformedYGarageArea + transformedYBsmtUnfSF:transformedYWoodDeckSF +
##     transformedYGarageArea:transformedYGrLivArea, data = my_data,
##     subset = -c(524, 1299))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07905 -0.09037  0.02202  0.12037  0.67608
##
## Coefficients:
##                                                   Estimate Std. Error t value
## (Intercept)                                      1.034e+01  2.713e-01  38.121
## transformedYLotFrontage                         -2.495e-02  5.825e-03  -4.283
## transformedYLotArea                             -8.439e-02  4.085e-02  -2.066
## transformedYBsmtFinSF1                           1.163e-02  2.205e-03   5.273
## transformedYBsmtUnfSF                            9.552e-04  4.340e-04   2.201
## transformedYGarageArea                           3.536e-03  1.077e-03   3.283
## transformedYLotFrontage:transformedYLotArea     2.220e-03  6.393e-04   3.473
## transformedYLotFrontage:transformedYBsmtFinSF1  9.554e-05  4.703e-05   2.031
## transformedYLotFrontage:transformedYBsmtUnfSF   1.287e-05  8.575e-06   1.501
## transformedYLotFrontage:transformedYGarageArea  7.352e-06  4.353e-06   1.689
## transformedYLotArea:transformedYGrLivArea       1.621e-02  1.522e-03  10.650
## transformedYLotArea:transformedYGarageArea     -6.046e-04  1.260e-04  -4.798
## transformedYLotArea:transformedYWoodDeckSF      1.197e-03  5.259e-04   2.277
## transformedYBsmtFinSF1:transformedYBsmtUnfSF   -9.131e-05  1.313e-05  -6.953
## transformedYBsmtFinSF1:transformedYGarageArea   2.738e-05  7.341e-06   3.729
## transformedYBsmtUnfSF:transformedYGarageArea    4.085e-06  1.443e-06   2.831
## transformedYBsmtUnfSF:transformedYWoodDeckSF    1.233e-04  4.862e-05   2.536
## transformedYGarageArea:transformedYGrLivArea    1.627e-04  5.459e-05   2.981
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## transformedYLotFrontage                        1.97e-05 ***
## transformedYLotArea                            0.039033 *
## transformedYBsmtFinSF1                         1.55e-07 ***
## transformedYBsmtUnfSF                          0.027904 *
## transformedYGarageArea                         0.001050 **
## transformedYLotFrontage:transformedYLotArea    0.000530 ***
## transformedYLotFrontage:transformedYBsmtFinSF1 0.042391 *
## transformedYLotFrontage:transformedYBsmtUnfSF  0.133630
## transformedYLotFrontage:transformedYGarageArea 0.091465 .
## transformedYLotArea:transformedYGrLivArea       < 2e-16 ***
## transformedYLotArea:transformedYGarageArea     1.77e-06 ***
## transformedYLotArea:transformedYWoodDeckSF     0.022947 *
## transformedYBsmtFinSF1:transformedYBsmtUnfSF   5.42e-12 ***
## transformedYBsmtFinSF1:transformedYGarageArea  0.000199 ***
## transformedYBsmtUnfSF:transformedYGarageArea   0.004701 **
## transformedYBsmtUnfSF:transformedYWoodDeckSF   0.011332 *
## transformedYGarageArea:transformedYGrLivArea   0.002917 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1938 on 1440 degrees of freedom
## Multiple R-squared:  0.7677, Adjusted R-squared:  0.765
## F-statistic:   280 on 17 and 1440 DF,  p-value: < 2.2e-16
```

```
resettest(final.model_interactions, power=2, type="regressor") #ResetTest
```

```
##
##  RESET test
##
## data:  final.model_interactions
## RESET = 7.1449, df1 = 7, df2 = 1433, p-value = 2.03e-08
```

Remark: Given my choice of predictors in this case, it is clear my model is quite bad at predicting SalePrice. In regard to the final model with interactions included, we see it has an extremely high level of multicollinearity between many of the variables which makes sense given the closeness in my predictors and the inclusion of interaction. That being said, the adjusted R-squared value is higher in this final model with interactions, than compared to my two previous models. Also, we know that multicollinearity doesn’t directly affect the predictive power of my model, but rather the ability to isolate effects. The RESET test shows that my model is misspecified, and that there may be some value in having higher power terms. That said, we know that if we have to raise terms to anything beyond a quadratic, lets say a cubic, it means that our variables are likely the issue, which in my case makes sense given previous analysis of my predictors.

(i)

```
#Using AIC and BIC to pick model
AIC(base.mod, base.mod.nout, final.model, model_interactions,
    final.model_interactions)
```

```
## Warning in AIC.default(base.mod, base.mod.nout, final.model,
## model_interactions, : models are not all fitted to the same number of
## observations

##                          df       AIC
## base.mod                 12 -433.2013
## base.mod.nout            12 -546.7947
## final.model               9 -525.3328
## model_interactions       30 -614.6761
## final.model_interactions 19 -627.8350
```

```
BIC(base.mod, base.mod.nout, final.model, model_interactions,
    final.model_interactions)
```

```
## Warning in BIC.default(base.mod, base.mod.nout, final.model,
## model_interactions, : models are not all fitted to the same number of
## observations

##                          df       BIC
## base.mod                 12 -369.7670
## base.mod.nout            12 -483.3769
## final.model               9 -477.7694
## model_interactions       30 -456.1315
## final.model_interactions 19 -527.4234
```

Remark: We pick the final model with interactions based on these results.

```
#Cross-Validation
#We aren't given any Sale Price values in test.data
set.seed(1)
training.samples <- df_transformedY$transformedYSalePrice %>%
  createDataPartition(p=0.8, list=FALSE)
train.data <- df_transformedY[training.samples, ]
test.data <- df_transformedY[-training.samples, ]

train_model <- lm(transformedYSalePrice ~., data = train.data)

predictions <- train_model %>% predict(test.data)
data.frame(
  RMSE=RMSE(predictions, test.data$transformedYSalePrice),
  R2=R2(predictions, test.data$transformedYSalePrice)
  )
```

```
##        RMSE        R2
## 1 0.2383234 0.6599338
```

Remark: Overall we find that this model does a moderate job at predicting sale price for our data set, but that it is likely overfit for our specific data set. We know that this final.model_interactions performs better than the previously identified models, given the low AIC and BIC values comparatively to other models and a higher R-squared. That being said, I realize I could have picked much better predictor variables that aren't so similar, which is why after all my analysis so many of the initial variables were excluded (not necessarily interactions between such variables though). It is worth noting that I had quite a small RMSE, which is promising as it means my model fits well to this data.

II. Problem 2

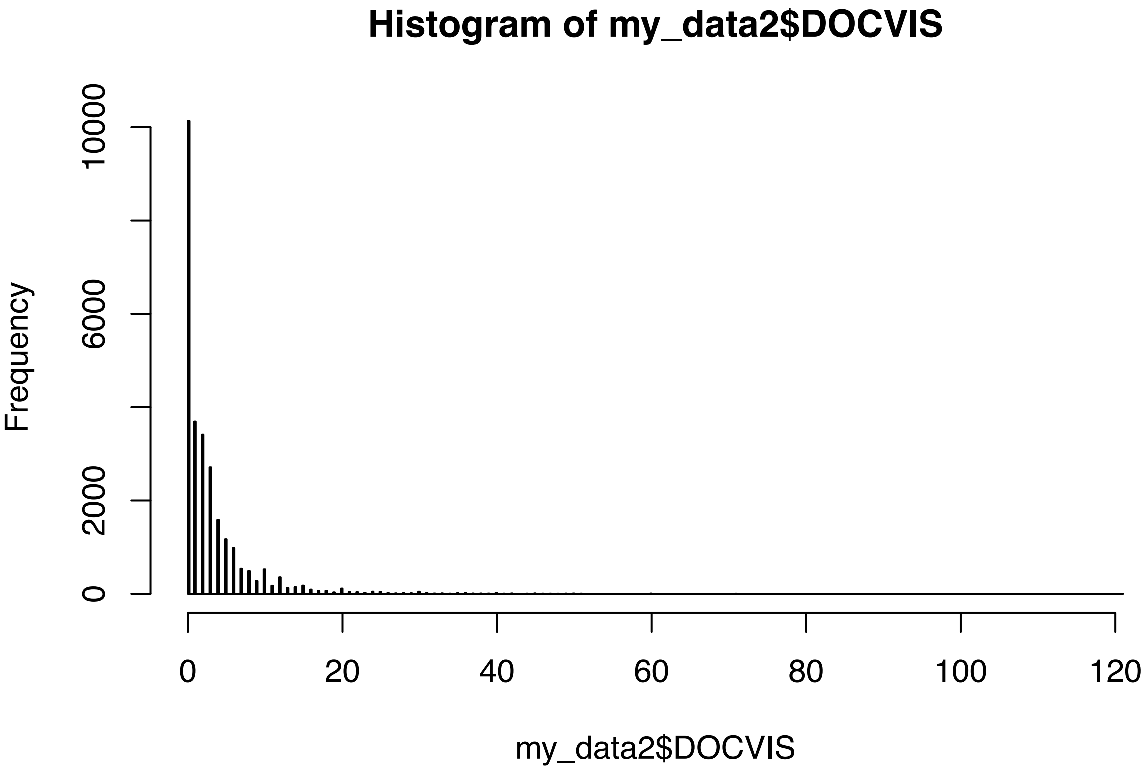
```
data2<-read.csv("german_healthcare_usage.csv")
#data2
my_data2 <- data2[ , c("DOCVIS", "FEMALE", "AGE", "UNEMPLOY", "MARRIED",
                      "HANDPER", "EDUC", "HHNINC", "HOSPVIS", "PRESCRIP", "YEAR")]
attach(my_data2)
#my_data2
```

(a)

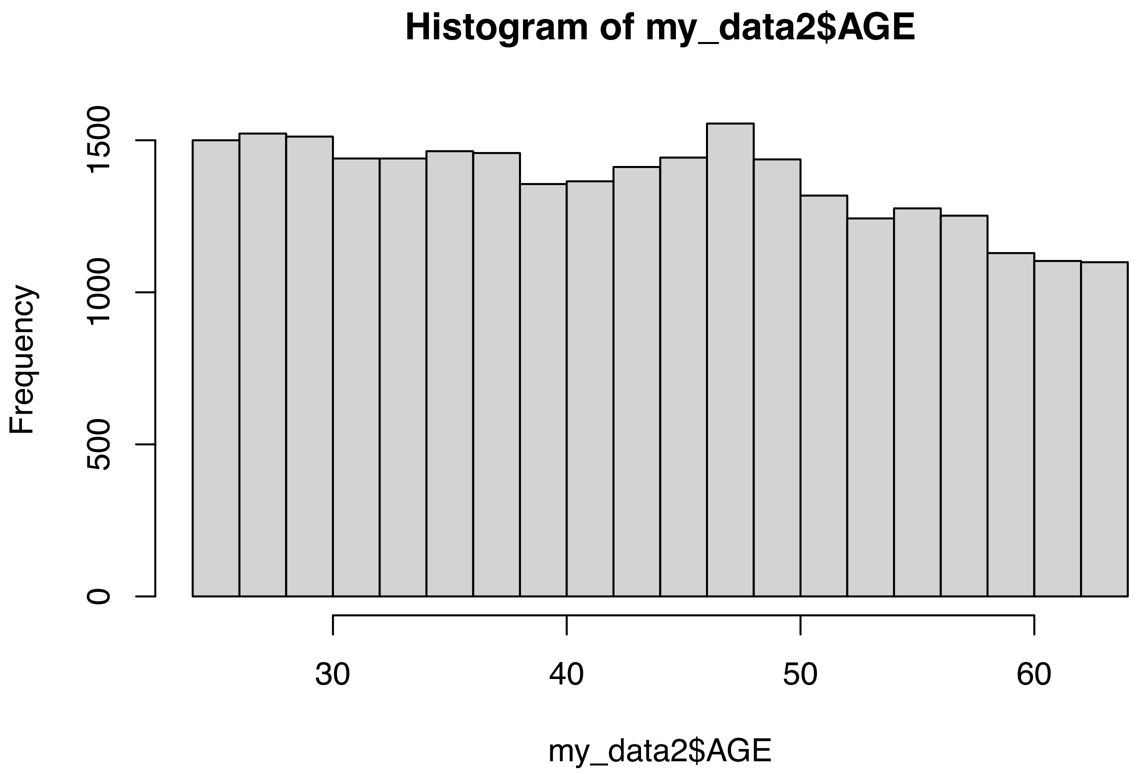
```
#Creating a baseline model without transformations
my_model <- lm(DOCVIS ~ AGE+UNEMPLOY+MARRIED+HANDPER+HHNINC+
                 HOSPVIS+EDUC+FEMALE, data=my_data2)

summary(my_model)
```

```
##
## Call:
## lm(formula = DOCVIS ~ AGE + UNEMPLOY + MARRIED + HANDPER + HHNINC +
##     HOSPVIS + EDUC + FEMALE, data = my_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.113  -2.511  -1.438   0.653 114.934
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.423356   0.237997   5.981 2.25e-09 ***
## AGE          0.042031   0.003160  13.300  < 2e-16 ***
## UNEMPLOY     0.240515   0.081176   2.963 0.003050 **
## MARRIED     -0.138211   0.080376  -1.720 0.085525 .
## HANDPER      0.052894   0.001817  29.118  < 2e-16 ***
## HHNINC      -1.062825   0.201285  -5.280 1.30e-07 ***
## HOSPVIS      0.772312   0.037383  20.659  < 2e-16 ***
## EDUC        -0.057124   0.015262  -3.743 0.000182 ***
## FEMALE       1.045614   0.072982  14.327  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.45 on 27310 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.08247,    Adjusted R-squared:  0.0822
## F-statistic: 306.8 on 8 and 27310 DF,  p-value: < 2.2e-16
```

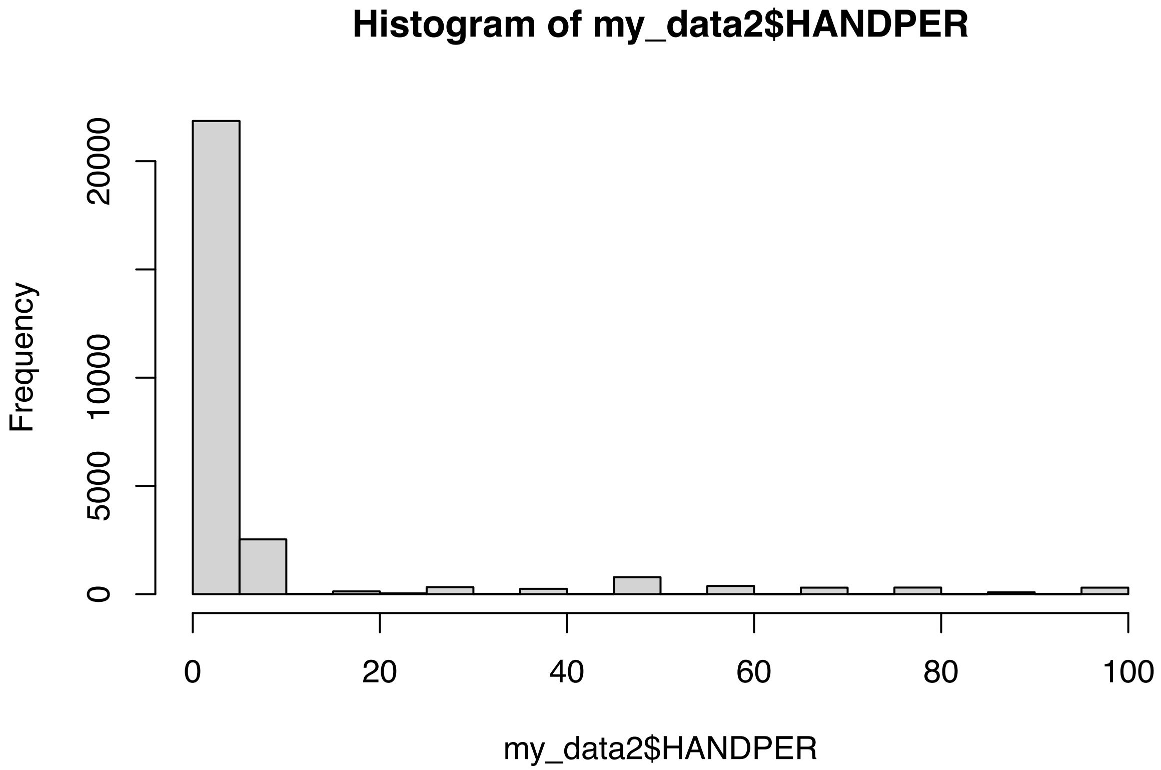
```
#CHECKING FOR TRANSFORMATIONS
hist(my_data2$DOCVIS, breaks="FD") #Very right skewed so likely need to transform
```



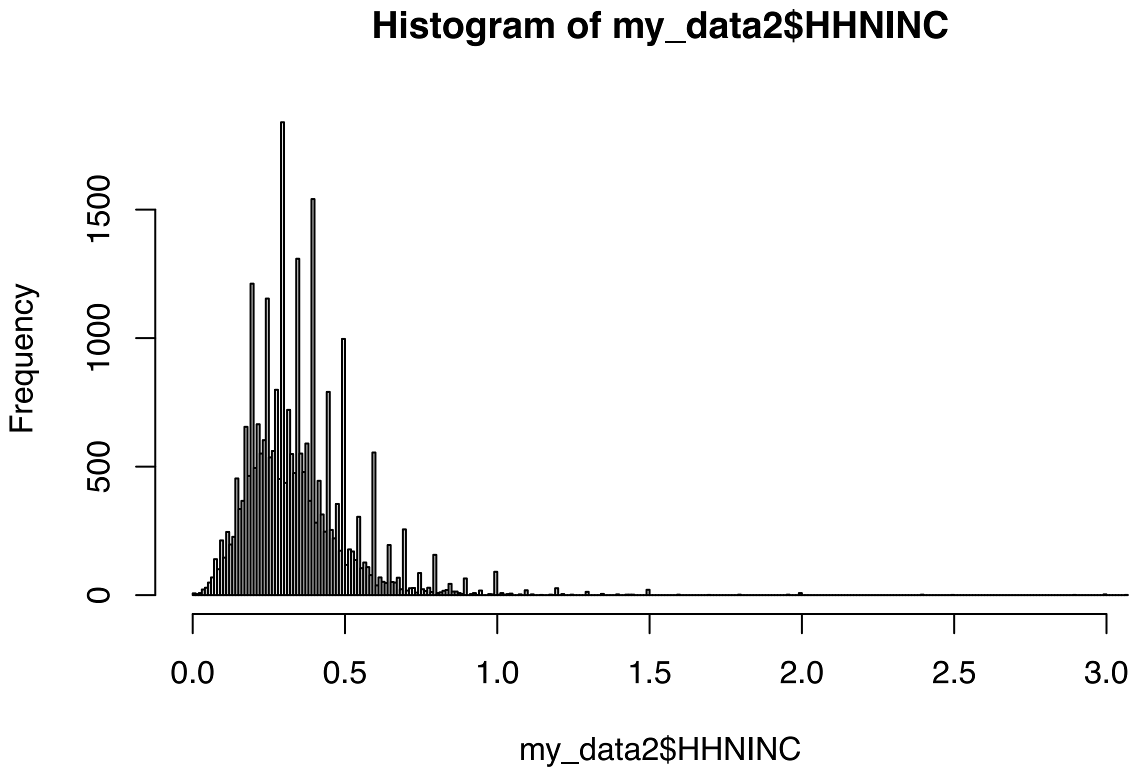
```
hist(my_data2$AGE) #Likely need to transform
```



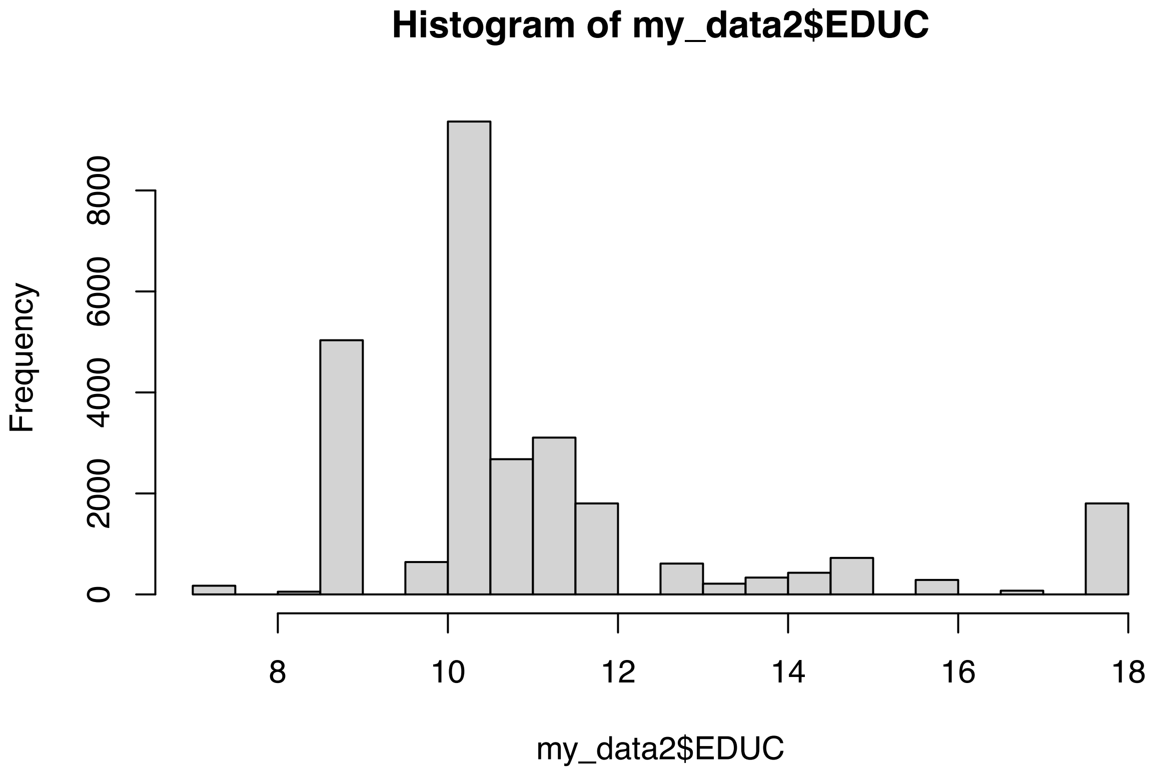
```
hist(my_data2$HANDPER) #Likely need to transform
```



```
hist(my_data2$HHNINC, breaks="FD") #No need to transform
```



```
hist(my_data2$EDUC) #No need to transform
```



```
#Checking appropriate transformations for variables
summary(a4 <- powerTransform(cbind(AGE, HANDPER)~1, my_data2, family="yjPower"))
```

```
## yjPower Transformations to Multinormality
##         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## AGE        0.6007        0.60       0.5512       0.6502
## HANDPER   -1.3141       -1.31      -1.3334      -1.2948
##
##  Likelihood ratio test that all transformation parameters are equal to 0
##                               LRT df       pval
## LR test, lambda = (0 0) 32439.72  2 < 2.22e-16
```

```
#Applying appropriate transformations
trans_my_data2 <- yjPower(with(my_data2, cbind(AGE,HANDPER))
                          ,coef(a4, round=TRUE))
trans_AGE <- trans_my_data2[, 1]
trans_HANDPER <- trans_my_data2[, 2]

trans_DOCVIS <- log(DOCVIS+1) #Adding constant so we get no undefined values
trans.df_my_data2<- data.frame(cbind(trans_DOCVIS, trans_AGE, trans_HANDPER,
                                     AGE, UNEMPLOY, HOSPVIS, EDUC, FEMALE,
                                     my_data2$MARRIED))
#Creating data frame with transformed variables and untransformed variables

trans_my_model <- lm(trans_DOCVIS ~ trans_AGE+trans_HANDPER+my_data2$MARRIED+HHNINC+HOSPVIS+EDUC+FEMALE

vif(my_model)
```

```
##      AGE UNEMPLOY  MARRIED  HANDPER   HHNINC  HOSPVIS     EDUC   FEMALE
## 1.179003 1.325002 1.087896 1.126374 1.165456 1.005311 1.157877 1.222386
```

```
#Checking for multicollinearity in baseline model
vif(trans_my_model)
```

```
##        trans_AGE    trans_HANDPER my_data2$MARRIED           HHNINC
##         1.148593         1.057805         1.089311         1.167568
##          HOSPVIS             EDUC           FEMALE         UNEMPLOY
##         1.003014         1.156497         1.206947         1.295141
```

```
#Checking for multicollinearity in model with transformations
AIC(trans_my_model, my_model)
```

```
##                df       AIC
## trans_my_model 10  69957.15
## my_model       10 170187.10
```

```
BIC(trans_my_model, my_model)
```

```
##                df      BIC
## trans_my_model 10  70039.3
## my_model       10 170269.3
```

Remark: Firstly, I chose this set of predictors because in learning from my mistakes in variable selection from problem 1, I realized the variables of choice should try and encompass a wide variety of factors effecting doctor visits instead of trying to isolate specific characteristics that are similar to one another (this leads to multicollinearity and other issues). That said, these variables seemed to encompass important factors of doctor visits. Also, all my predictors showed statistical significance at some level at or below 10%. The variable "MARRIED" is the one variable that showed only significance at the 10% level which we might intuitively expect when comparing it to the strength of our other predictors in estimating the number of doctor visits. Also, it was promising that there were no levels of multicollinearity shown that were even close to 4, in fact every variable was below a value of 2 when running VIF on my model. I also performed a log transformation to the variable "DOCVIS", and power transformations to "AGE" & "HANDPER", which strengthened the models performance, and actually didn't sacrifice all to much in terms of interpretability. Lastly, it is important to mention the low R^2 (multiple & adjusted) because typically it would be alarming but since we are working with panel data we expect this to be the case because of the immense heterogeneity of cross sections.

(b)

```r
#Differences in Differences: In 1987 the German Government passed a series of
#legislation to improve healthcare access for unemployed people and women.
#i. Determine whether or not the policy worked for women.
my_data2$YEAR <- ifelse(my_data2$YEAR >= 1987, 1, 0)
#Creating dummy variables: 1 is YEAR 1987+, 0 is anything before YEAR 1987

#creating difference in difference estimator
DiD <- my_data2$FEMALE*my_data2$YEAR
#Creating model with interaction (DiD)
my_model2 <- lm(trans_DOCVIS ~ trans_AGE+trans_HANDPER+my_data2$MARRIED+HHNINC+HOSPVIS+EDUC+FEMALE+
                  UNEMPLOY+DiD, data=trans.df_my_data2)
summary(my_model2)
```

```
##
## Call:
## lm(formula = trans_DOCVIS ~ trans_AGE + trans_HANDPER + my_data2$MARRIED +
##     HHNINC + HOSPVIS + EDUC + FEMALE + UNEMPLOY + DiD, data = trans.df_my_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4214 -0.7360 -0.0543  0.5862  3.6558
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.217179   0.045007   4.825 1.40e-06 ***
## trans_AGE         0.048253   0.002251  21.437  < 2e-16 ***
## trans_HANDPER     0.415035   0.018072  22.965  < 2e-16 ***
## my_data2$MARRIED -0.027572   0.012851  -2.145 0.031923 *
## HHNINC           -0.116526   0.032460  -3.590 0.000331 ***
## HOSPVIS           0.112578   0.005964  18.878  < 2e-16 ***
## EDUC             -0.011306   0.002436  -4.641 3.48e-06 ***
## FEMALE            0.252204   0.015171  16.624  < 2e-16 ***
## UNEMPLOY          0.080083   0.012861   6.227 4.83e-10 ***
## DiD              -0.023079   0.015911  -1.451 0.146923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8704 on 27309 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.09148,    Adjusted R-squared:  0.09119
## F-statistic: 305.5 on 9 and 27309 DF,  p-value: < 2.2e-16
```

Remark: Here we see that the policy yielded approximately on average a -2.31% change in doctor visits for females post 1987, but this result was not statistically significant at any level. Therefore we fail to conclude a statistically significant effect on the policy for women.

```
#ii. Determine whether or not the policy worked for unemployed.
DiD_2 <- my_data2$UNEMPLOY*my_data2$YEAR
my_model3 <- lm(trans_DOCVIS ~ trans_AGE+trans_HANDPER+my_data2$MARRIED+HHNINC+HOSPVIS+EDUC+FEMALE+UNEMP
summary(my_model3)
```

```
##
## Call:
## lm(formula = trans_DOCVIS ~ trans_AGE + trans_HANDPER + my_data2$MARRIED +
##     HHNINC + HOSPVIS + EDUC + FEMALE + UNEMPLOY + DiD_2, data = trans.df_my_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4037 -0.7365 -0.0553  0.5864  3.6744
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.217509   0.044999   4.834 1.35e-06 ***
## trans_AGE         0.048204   0.002251  21.419  < 2e-16 ***
## trans_HANDPER     0.416768   0.017969  23.193  < 2e-16 ***
## my_data2$MARRIED -0.027929   0.012850  -2.173 0.029756 *
## HHNINC           -0.114516   0.032364  -3.538 0.000403 ***
## HOSPVIS           0.112568   0.005963  18.878  < 2e-16 ***
## EDUC             -0.011335   0.002436  -4.654 3.28e-06 ***
## FEMALE            0.237346   0.011584  20.489  < 2e-16 ***
## UNEMPLOY          0.106202   0.016565   6.411 1.47e-10 ***
## DiD_2            -0.044376   0.018951  -2.342 0.019204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8704 on 27309 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.0916, Adjusted R-squared:  0.0913
## F-statistic:   306 on 9 and 27309 DF,  p-value: < 2.2e-16
```

Remark: Here we see that after the policy there was approximately on average a -4.44% change in doctor visits for those who are unemployed. This result was statistically significant at the 5% level and therefore we can conclude (at that level) there is a negative effect on unemployed individuals from this policy as there was a 4.44% decrease in number of doctor visits after 1987.

(c)

```r
my_data3<-na.exclude(my_data2) #excluding NA's
#Test the hypothesis that the number of doctor visits a patient
#has over a 3 month period is greater for women than for men.
model_4 <- lm(DOCVIS~FEMALE, my_data3)
#Creating model with doctor visits regressed on gender
print(model_4)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE, data = my_data3)
##
## Coefficients:
## (Intercept)       FEMALE
##       2.625        1.166
```

```r
anova(update(model_4,.~.-FEMALE),model_4)
```

```
## Analysis of Variance Table
##
## Model 1: DOCVIS ~ 1
## Model 2: DOCVIS ~ FEMALE
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  27318 884189
## 2  27317 874915  1    9273.6 289.55 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remark: Since the p-value is less than 0.05 we can conclude that there is a difference between the number of doctor visits for females comparatively to males. We see that our coefficient states there is approximately on average a 1.17 unit increase in doctor visits for females.

(d)

```r
#Based on your findings propose and test your own hypothesis of interest.
#My Hypothesis: The Number of doctor visits a patient has over a 3 month
#period is greater for unemployed than employed.
model_5 <- lm(DOCVIS~UNEMPLOY, my_data3)
#Creating model with doctor visits regressed on employment condition
print(model_5)
```

```
##
## Call:
## lm(formula = DOCVIS ~ UNEMPLOY, data = my_data3)
##
## Coefficients:
## (Intercept)     UNEMPLOY
##       2.703        1.487
```

```r
anova(update(model_5,.~.-UNEMPLOY),model_5)
```

```
## Analysis of Variance Table
##
## Model 1: DOCVIS ~ 1
## Model 2: DOCVIS ~ UNEMPLOY
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  27318 884189
## 2  27317 870983  1     13206 414.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Remark: Since the p-value is less than 0.05 we can conclude that there is a difference between the number doctor visits for those who are unemployed compared to those who are employed. We see that our coefficient shows that there is approximately on average a 1.49 unit increase in doctor visits for unemployed individuals.