FinalProject

Ben Tankus

3/2/2021

```
df <- read.csv('OR_acs_house_occ.csv')
```

QUESTIONS TO ANSWER:

1. Do people living in apartments pay less on electricity than those living in houses? (adjust for number of bedrooms and number of occupants)
2. How much does each group pay?
3. What about extra heating cost using fuel or gas?

CREATE MODEL

Model is used to predict electricity costs for a house in Oregon

MISSINGNESS

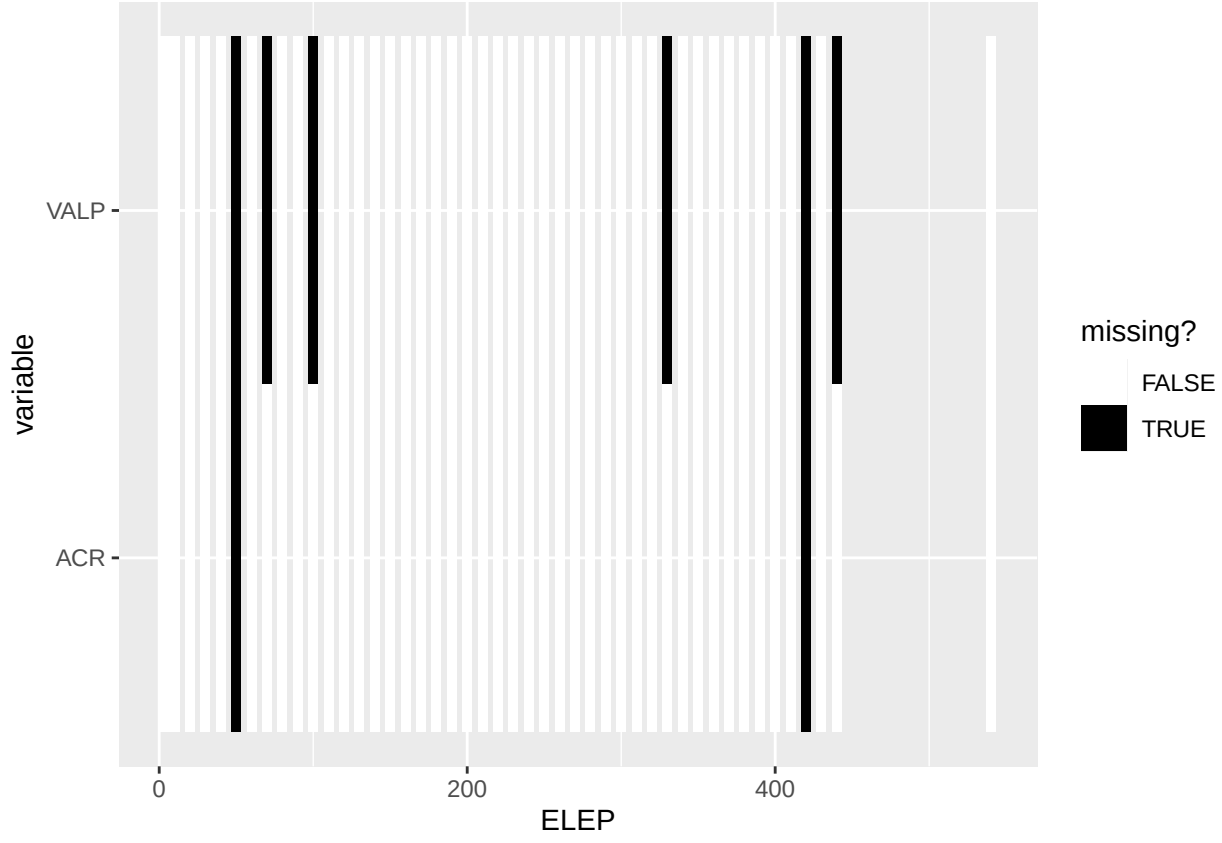
ACR and VALP both have many null values, 2586 and 4632 respectively. Analysis in the *TBD* section will determine how the missingness is distributed.

```
i = 1
names <- names(df)

for( col in df) {
  #print(paste(col.names(), sum(is.na(col))))
  print(paste(names[i], "Has", sum(is.na(col)), "nulls"))
  i <- i + 1
}
```

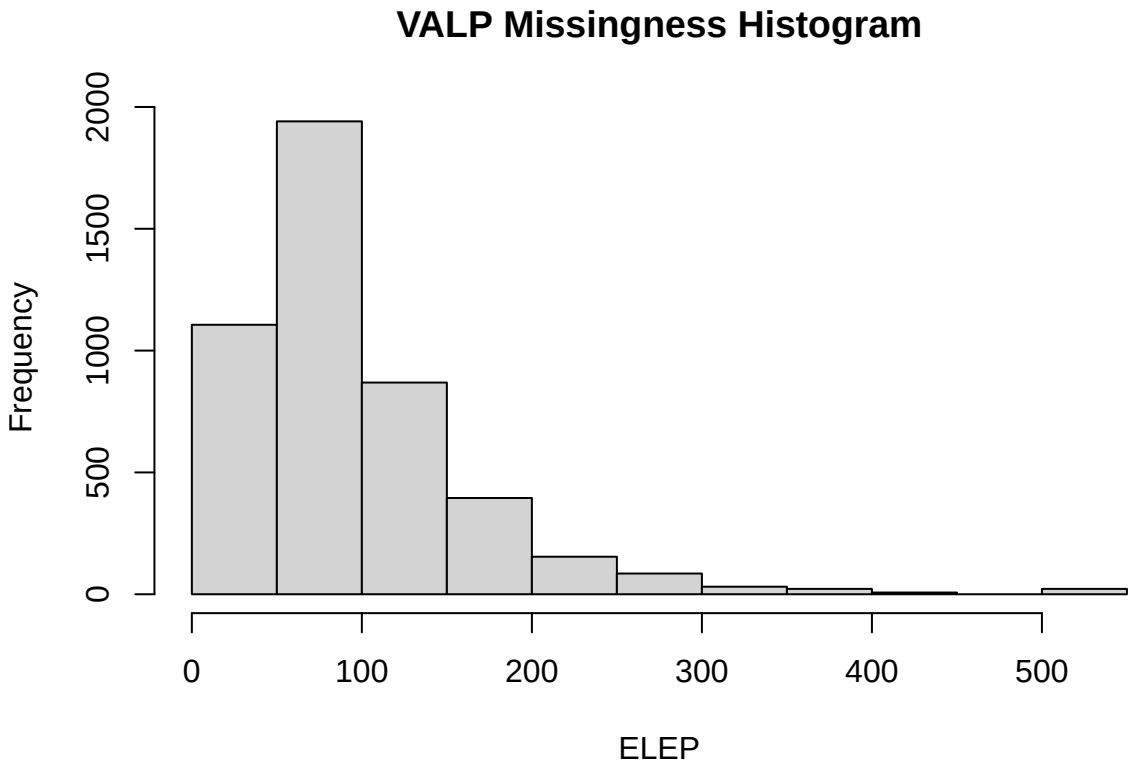
```
## [1] "SERIALNO Has 0 nulls"
## [1] "NP Has 0 nulls"
## [1] "TYPE Has 0 nulls"
## [1] "ACR Has 2586 nulls"
## [1] "BDSP Has 0 nulls"
## [1] "BLD Has 0 nulls"
## [1] "ELEP Has 0 nulls"
## [1] "FULP Has 0 nulls"
## [1] "GASP Has 0 nulls"
```

```
## [1] "HFL Has 0 nulls"
## [1] "RMSP Has 0 nulls"
## [1] "TEN Has 0 nulls"
## [1] "VALP Has 4632 nulls"
## [1] "YBL Has 0 nulls"
## [1] "R18 Has 0 nulls"
## [1] "R60 Has 0 nulls"
```

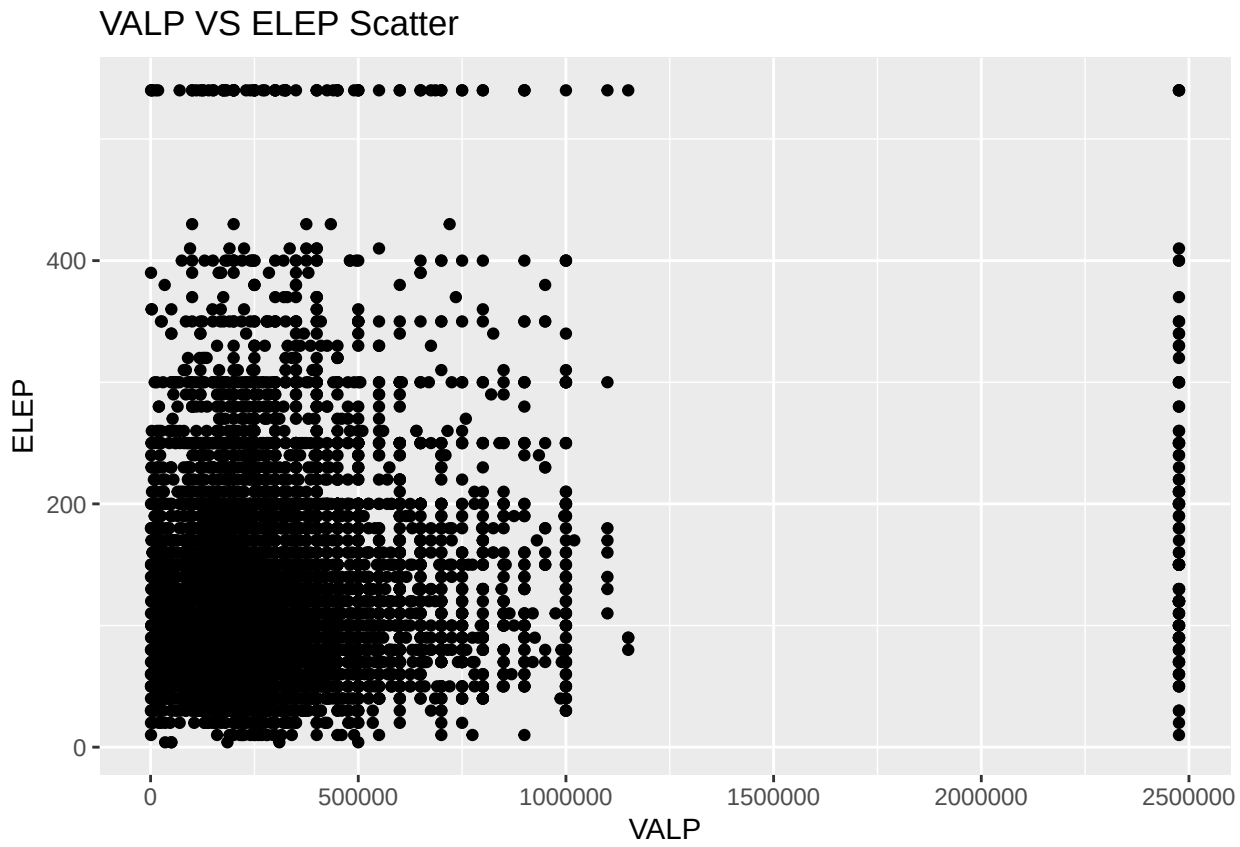
```
dfNulls <- df[, c('ACR','ELEP', 'VALP')]

dfNulls_long <- gather(dfNulls, variable, value, -ELEP)


qplot(ELEP, variable, data = dfNulls_long, geom = "tile",
fill = is.na(value)) +
scale_fill_manual("missing?",
values = c('TRUE' = "black", 'FALSE' = "white")) +
theme(axis.text.x = element_text(angle = 0))
```



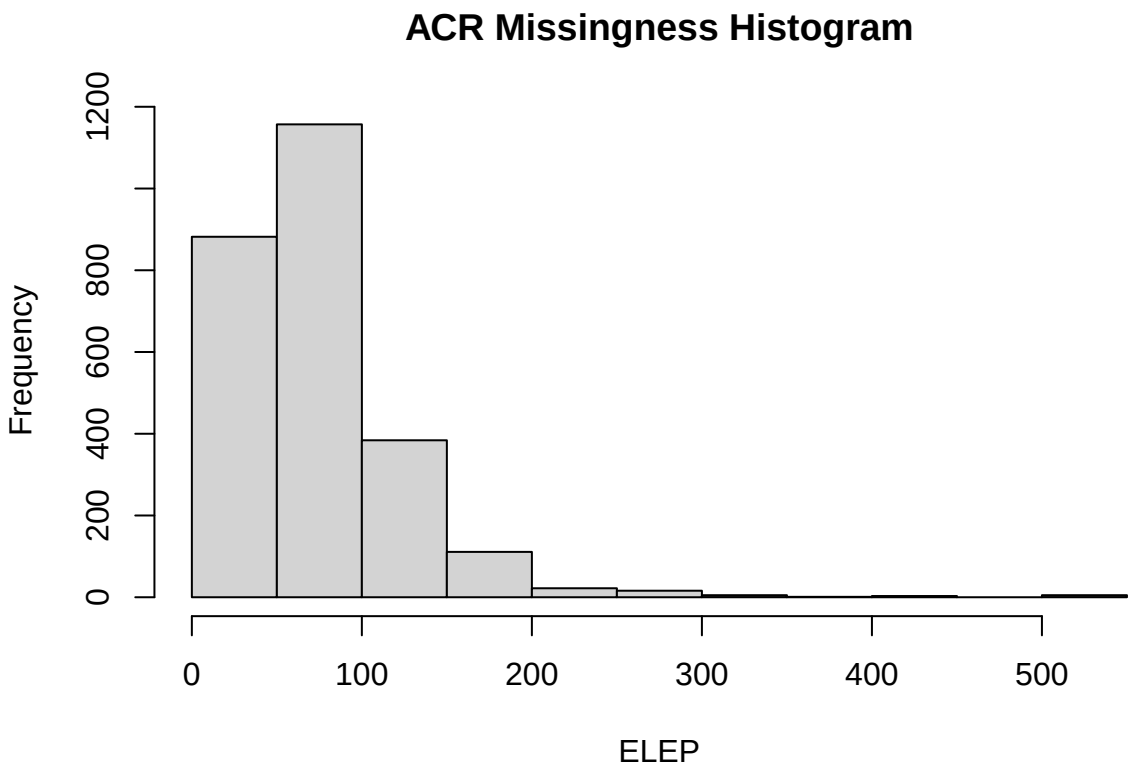

```
VALPNullsHist.df <- subset(df[,c('VALP', 'ELEP')], is.na(VALP) == TRUE)
hist(VALPNullsHist.df[,'ELEP'], main = 'VALP Missingness Histogram', xlab = 'ELEP')
```



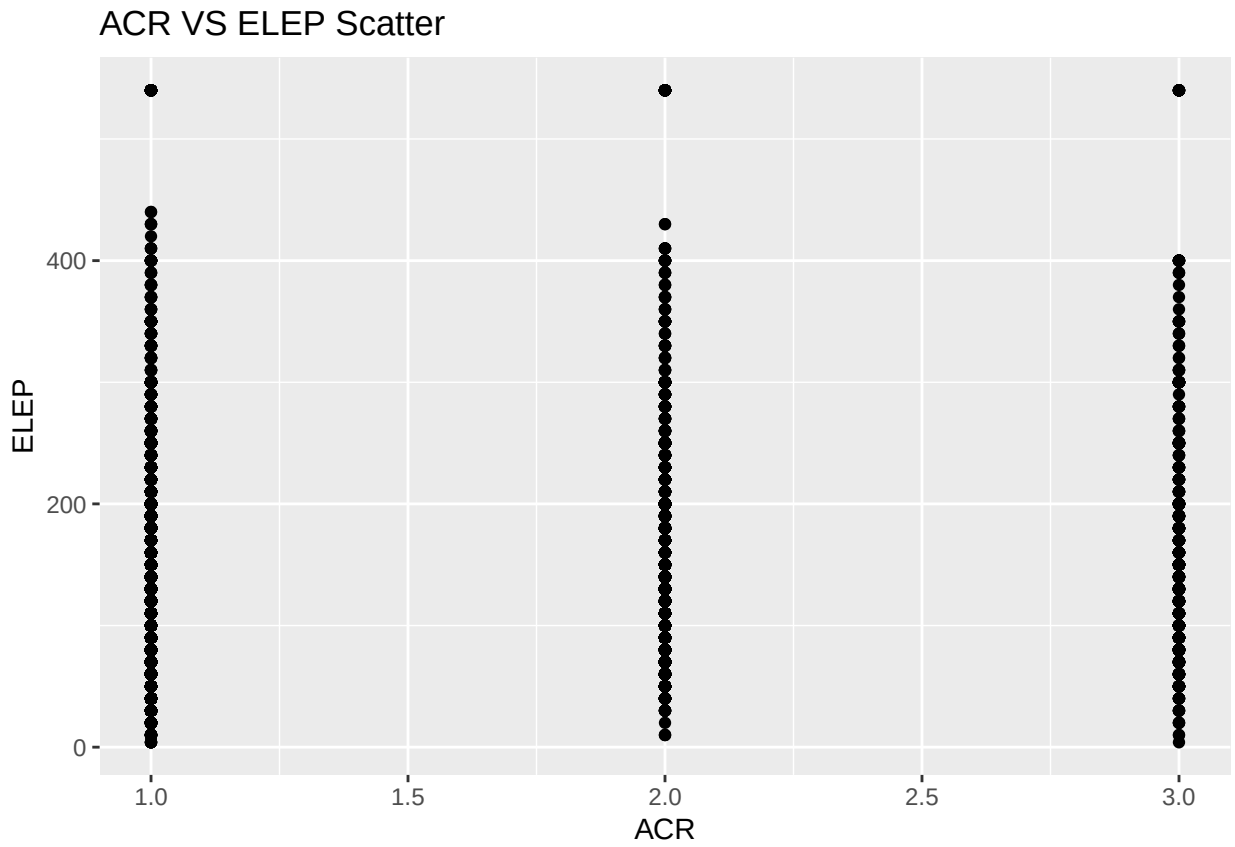

```
VALPNoNullsHist.df <- subset(df[,c('VALP', 'ELEP')], is.na(VALP) == FALSE)
qplot(VALPNoNullsHist.df[,'VALP'], VALPNoNullsHist.df[,'ELEP'], main = 'VALP VS ELEP Scatter', ylab = '
```




```
ACRNullsHist.df <- subset(df[,c('ACR', 'ELEP')], is.na(ACR) == TRUE)
hist(ACRNullsHist.df[,'ELEP'], main = 'ACR Missingness Histogram', xlab = 'ELEP')
```



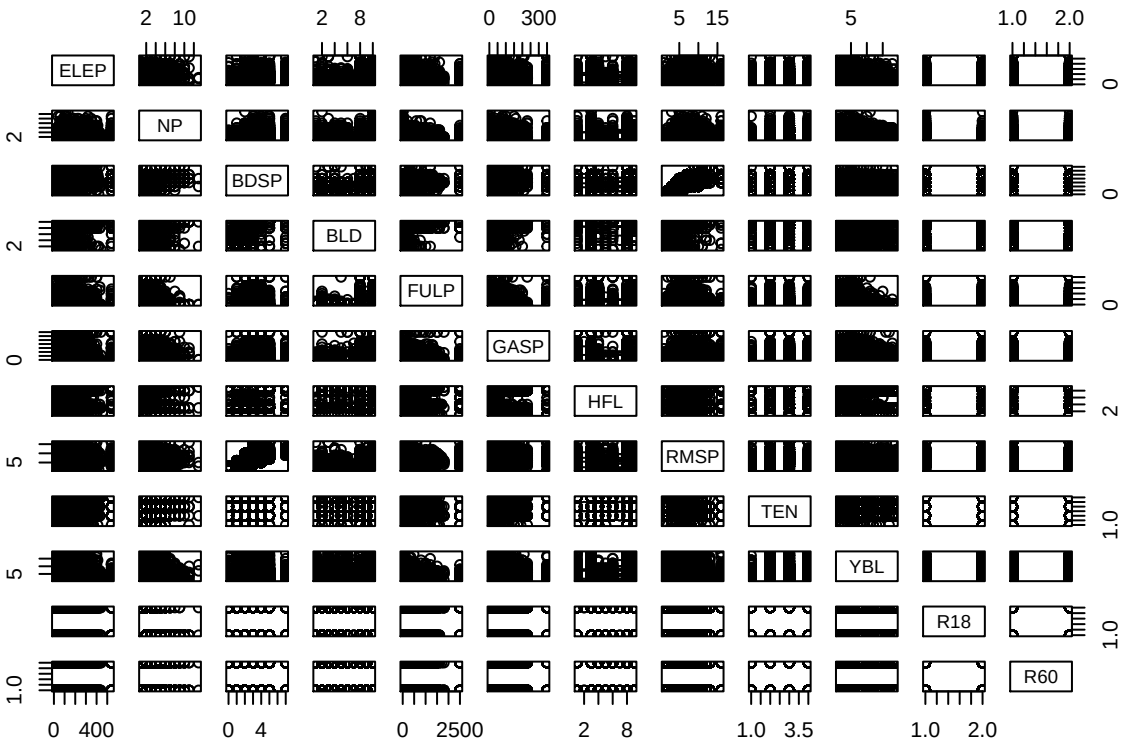

```
ACRNoNullsHist.df <- subset(df[,c('ACR', 'ELEP')], is.na(ACR) == FALSE)
qplot(as.numeric(factor(ACRNoNullsHist.df['ACR'])), ACRNoNullsHist.df['ELEP'], main = 'ACR VS ELEP Sca
```




```
panel.hist <- function(x, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks
  nB <- length(breaks)
  y <- h$counts
  y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "white", ...)
}


panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y, use = "complete.obs"))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste(prefix, txt, sep = "")
  if (missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex =  cex.cor * (1 + r) / 2)
}
```

Correlation check before dropping potentially practically irrelevant columns

```
df.corr <-df[,c('ELEP', 'NP', 'BDSP','BLD' ,'FULP', 'GASP', 'HFL', 'RMSP', 'TEN', 'YBL', 'R18', 'R60')]

df.corr['BLD'] = as.numeric(factor(df.corr[,'BLD']))
df.corr['HFL'] = as.numeric(factor(df.corr[,'HFL']))
df.corr['TEN'] = as.numeric(factor(df.corr[,'TEN']))
df.corr['YBL'] = as.numeric(factor(df.corr[,'YBL']))
df.corr['R18'] = as.numeric(factor(df.corr[,'R18']))
df.corr['R60'] = as.numeric(factor(df.corr[,'R60']))



plot(df.corr)
```

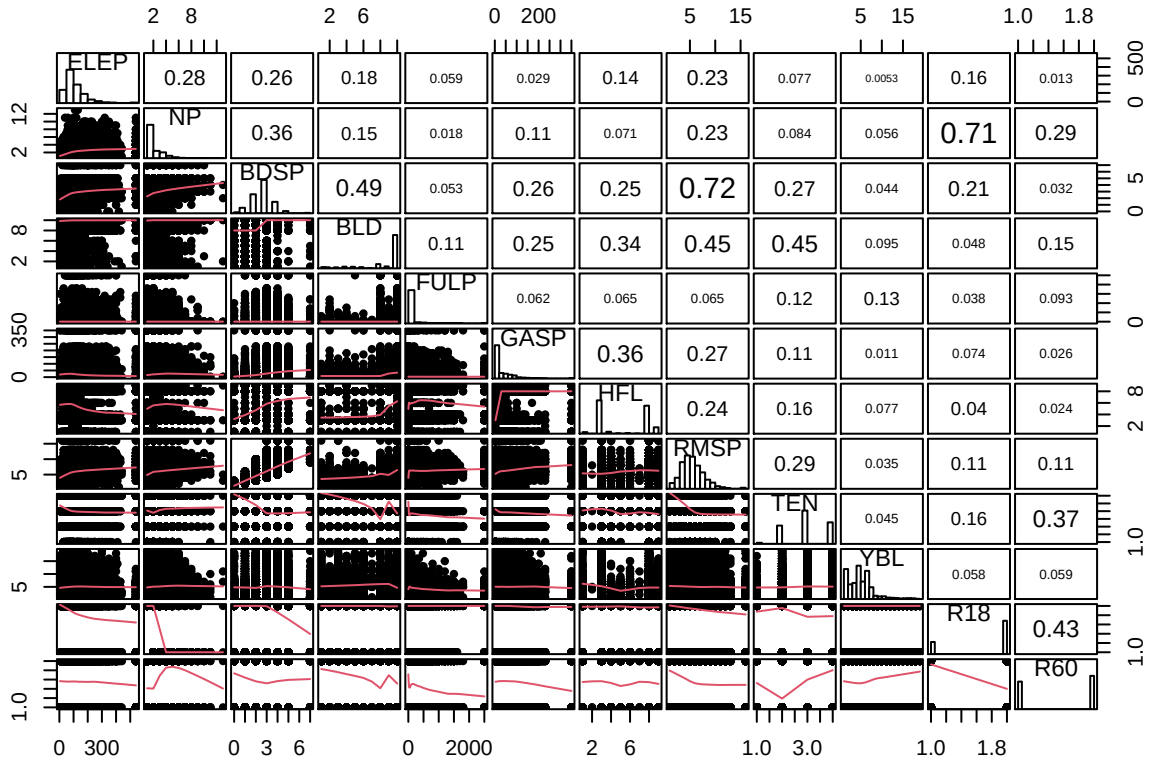



```
summary(df.corr)
```

```
##       ELEP             NP             BDSP            BLD
##  Min.   :  4.0   Min.   : 1.000   Min.   :0.000   Min.   : 1.000
##  1st Qu.: 70.0   1st Qu.: 1.000   1st Qu.:2.000   1st Qu.: 8.000
##  Median :100.0   Median : 2.000   Median :3.000   Median :10.000
##  Mean   :116.2   Mean   : 2.403   Mean   :2.789   Mean   : 8.695
##  3rd Qu.:150.0   3rd Qu.: 3.000   3rd Qu.:3.000   3rd Qu.:10.000
##  Max.   :540.0   Max.   :13.000   Max.   :7.000   Max.   :10.000
```

```
##       FULP              GASP             HFL             RMSP
##  Min.   :   1.00   Min.   :  3.00   Min.   :1.000   Min.   : 1.00
##  1st Qu.:   2.00   1st Qu.:  3.00   1st Qu.:3.000   1st Qu.: 4.00
##  Median :   2.00   Median :  3.00   Median :4.000   Median : 6.00
##  Mean   :  85.25   Mean   : 35.68   Mean   :5.488   Mean   : 6.01
##  3rd Qu.:   2.00   3rd Qu.: 50.00   3rd Qu.:8.000   3rd Qu.: 7.00
##  Max.   :2500.00   Max.   :350.00   Max.   :9.000   Max.   :16.00
##       TEN              YBL              R18             R60
##  Min.   :1.000   Min.   : 1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.: 3.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median : 5.00   Median :2.000   Median :2.000
##  Mean   :3.013   Mean   : 5.37   Mean   :1.733   Mean   :1.547
##  3rd Qu.:4.000   3rd Qu.: 7.00   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :19.00   Max.   :2.000   Max.   :2.000
```

```r
pairs(
  df.corr,
  upper.panel = panel.cor,
  diag.panel  = panel.hist,
  lower.panel = panel.smooth,
  gap = 1/5,
  pch = 20 #small dots
)
```



```
#DECLARE FINAL RELEVANT DF
df.relevant <- df[,c('ELEP', 'FULP', 'GASP', 'YBL', 'BLD', 'HFL', 'BDSP', 'NP')]
head(df.relevant)
```

```
##   ELEP FULP GASP             YBL                       BLD
## 1   70    2    3 1939 or earlier One-family house detached
## 2  100  600    3 1939 or earlier One-family house detached
## 3   60    2  110 1939 or earlier One-family house detached
## 4   80    2   20    1950 to 1959 One-family house detached
## 5  150    2    3    1990 to 1999    Mobile home or trailer
## 6  200    2   20 1939 or earlier One-family house detached
##                         HFL BDSP NP
## 1                      Wood    2  4
## 2 Fuel oil, kerosene, etc.    2  2
## 3              Utility gas    3  1
## 4              Utility gas    4  2
## 5              Electricity    2  1
## 6              Electricity    3  3
```

CLEANING FINAL COLUMN CHOICES

```
df.final <- df.relevant

# Drop rows with "Mobile home or trailer
df.final <- filter(df.final, BLD != 'Mobile home or trailer' & BLD != 'Boat, RV, van, etc.')

# Transform BLD
adjustVector <- c()
i <- 1

for (e in df.final[,'BLD'])
{

  if (grepl('house' ,e, fixed = TRUE)){
    adjustVector[i] <- "House"
    i <- i+1

  } else
  {
    adjustVector[i] <- "Apartment"
    i <- i+1
  }
}
df.final['BLDAdjusted'] = adjustVector


# Transform HFL
adjustVector <- c()
i <- 1

for (e in df.final[,'HFL'])
```

```
{

  if (grepl('Electricity' ,e, fixed = TRUE)){
    adjustVector[i] <- "Electricity"
    i <- i+1

  } else
  {
    adjustVector[i] <- "Not Electricity"
    i <- i+1
  }
}
df.final['HFLAdjusted'] = adjustVector


# TRANSFORM YBL
adjustVector <- c()
i <- 1

for (e in df.final[,'YBL'])
{

  if (e >= 2005){
    adjustVector[i] <- "2005 to 2015"
    i <- i+1

  } else
  {
    adjustVector[i] <- e
    i <- i+1
  }
}
df.final['YBLAdjusted'] = adjustVector
```

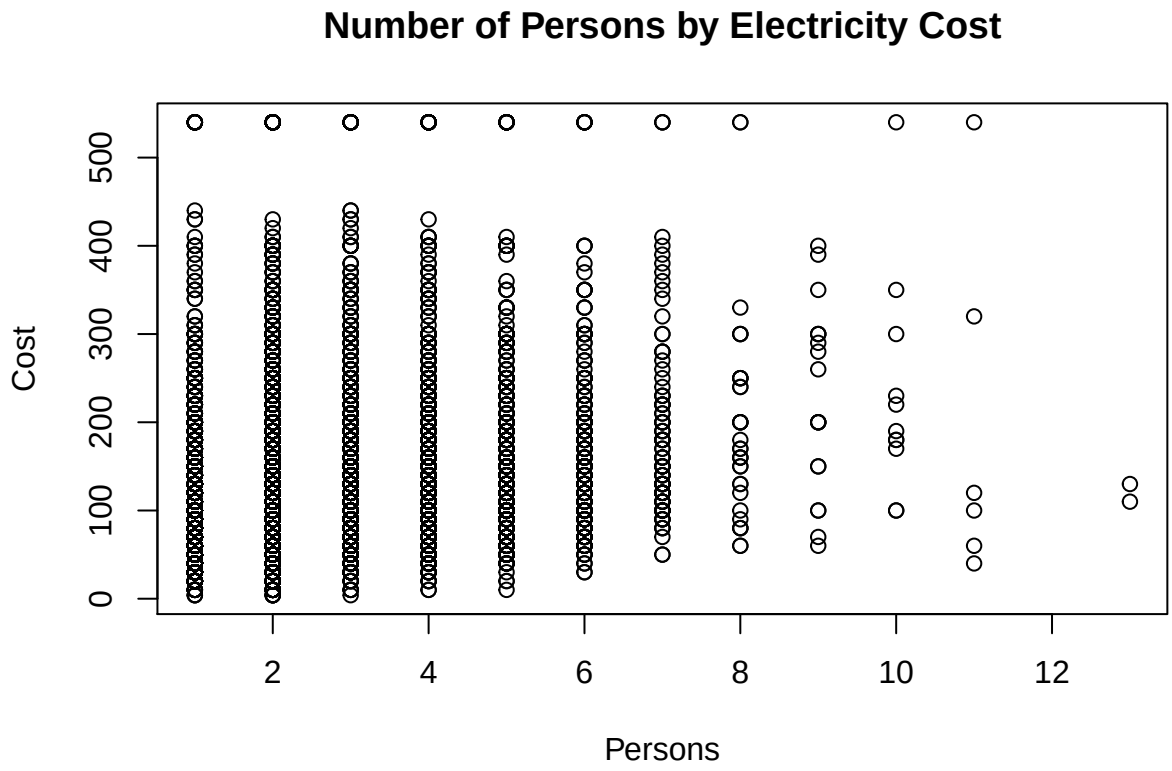
EXPLORE DATA

```
# NO SIGNIFICANT VISUAL CORRELATION BETWEEN COST AND NUMBER OF PERSONS
summary(df)
```

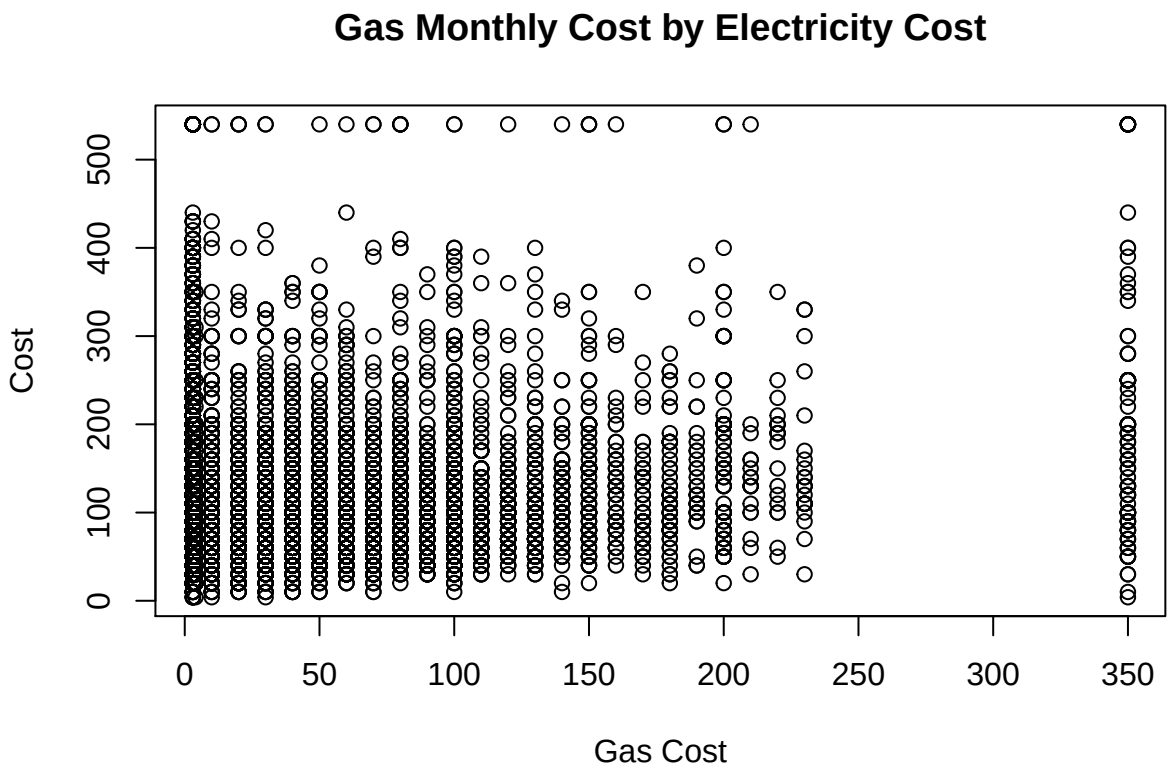
```
##     SERIALNO             NP              TYPE       ACR
##  Min.   :     70   Min.   : 1.000   Min.   :1   Length:15166
##  1st Qu.: 368628   1st Qu.: 1.000   1st Qu.:1   Class :character
##  Median : 748326   Median : 2.000   Median :1   Mode  :character
##  Mean   : 749620   Mean   : 2.403   Mean   :1
##  3rd Qu.:1126788   3rd Qu.: 3.000   3rd Qu.:1
##  Max.   :1513284   Max.   :13.000   Max.   :1
##
##       BDSP           BLD                ELEP            FULP
##  Min.   :0.000   Length:15166       Min.   :  4.0   Min.   :   1.00
##  1st Qu.:2.000   Class :character   1st Qu.: 70.0   1st Qu.:   2.00
##  Median :3.000   Mode  :character   Median :100.0   Median :   2.00
```

```
##  Mean   :2.789                    Mean   :116.2   Mean   :  85.25
##  3rd Qu.:3.000                    3rd Qu.:150.0   3rd Qu.:   2.00
##  Max.   :7.000                    Max.   :540.0   Max.   :2500.00
##
##       GASP            HFL                RMSP           TEN
##  Min.   :  3.00   Length:15166       Min.   : 1.00   Length:15166
##  1st Qu.:  3.00   Class :character   1st Qu.: 4.00   Class :character
##  Median :  3.00   Mode  :character   Median : 6.00   Mode  :character
##  Mean   : 35.68                      Mean   : 6.01
##  3rd Qu.: 50.00                      3rd Qu.: 7.00
##  Max.   :350.00                      Max.   :16.00
##
##       VALP             YBL                R18                R60
##  Min.   :   1000   Length:15166       Length:15166       Length:15166
##  1st Qu.: 160000   Class :character   Class :character   Class :character
##  Median : 250000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 301966
##  3rd Qu.: 360000
##  Max.   :2476000
##  NA's   :4632
```

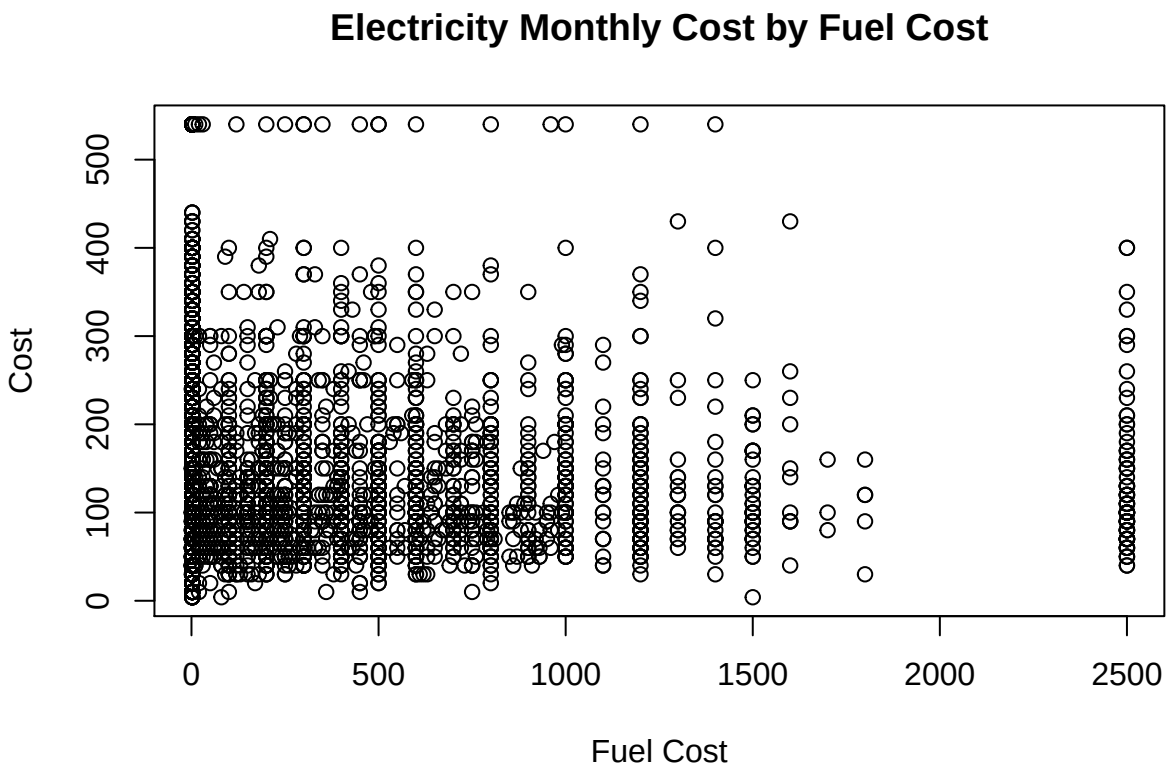
```r
plot(df$NP, df$ELEP, main = 'Number of Persons by Electricity Cost', xlab =  'Persons', ylab = 'Cost')
```



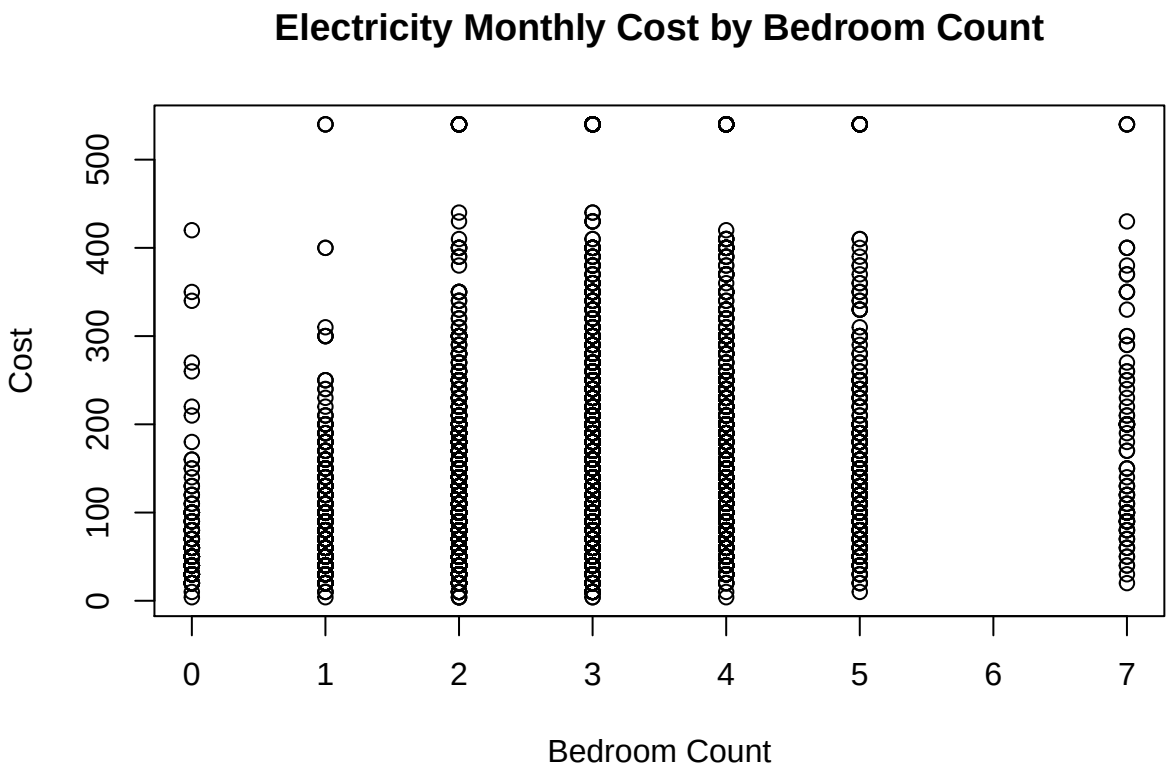

```r
# NO SIGNIFICANT VISUAL CORRELATION BETWEEN COST AND GAS PRICS
plot(df$GASP, df$ELEP, main = 'Gas Monthly Cost by Electricity Cost', xlab =  'Gas Cost', ylab = 'Cost')
```




```
# MAYBE CORRELATION
plot(df$FULP, df$ELEP, main = 'Electricity Monthly Cost by Fuel Cost', xlab =  'Fuel Cost', ylab = 'Cos
```




```
# MAYBE CORRELATION
plot(df$BDSP, df$ELEP, main = 'Electricity Monthly Cost by Bedroom Count', xlab =  'Bedroom Count', ylab
```




```
summary(df.final)
```

```
##       ELEP             FULP              GASP            YBL
##  Min.   :  4.0   Min.   :   1.00   Min.   :  3.00   Length:13774
##  1st Qu.: 70.0   1st Qu.:   2.00   1st Qu.:  3.00   Class :character
##  Median :100.0   Median :   2.00   Median : 10.00   Mode  :character
##  Mean   :114.1   Mean   :  82.38   Mean   : 37.87
##  3rd Qu.:140.0   3rd Qu.:   2.00   3rd Qu.: 60.00
##  Max.   :540.0   Max.   :2500.00   Max.   :350.00
##      BLD                HFL                 BDSP             NP
##  Length:13774       Length:13774       Min.   :0.000   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Median :3.000   Median : 2.000
##                                        Mean   :2.811   Mean   : 2.417
##                                        3rd Qu.:3.000   3rd Qu.: 3.000
##                                        Max.   :7.000   Max.   :13.000
##  BLDAdjusted        HFLAdjusted        YBLAdjusted
##  Length:13774       Length:13774       Length:13774
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
# DECLARE HOUSE MODEL
fit.basic <-  lm(ELEP ~ BLDAdjusted + NP + BDSP - 1, data = df.final)

# Add HFL and compare
fit.basic.HFL <-  lm(ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted, data = df.final)
anova(fit.basic, fit.basic.HFL) #SIGNIF DIFFERENCE, THEREFORE INCLUDE HFL
```

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ BLDAdjusted + NP + BDSP - 1
## Model 2: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  13770 66813068
## 2  13769 62027869  1   4785199 1062.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Add YBL and compare
fit.basic.HFLYBL <-  lm(ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted), data = d
anova(fit.basic.HFL, fit.basic.HFLYBL) # Lower RSS, therefore include YBL
```

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted
## Model 2: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted)
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  13769 62027869
## 2  13761 61675311  8    352558 9.8328 9.995e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Add GASP and compare
fit.basic.HFLYBLGASP <-  lm(ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted) + GAS
anova(fit.basic.HFLYBL, fit.basic.HFLYBLGASP) # Lower but not by much, therefore DONT include FULP
```

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted)
## Model 2: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted) +
##     GASP
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  13761 61675311
## 2  13760 61417716  1    257596 57.712 3.231e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Add FULP and compare
fit.basic.HFLYBLFULP <-  lm(ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted) + FUL
anova(fit.basic.HFLYBL, fit.basic.HFLYBLFULP) # Lower RSS therefore include FULP
```

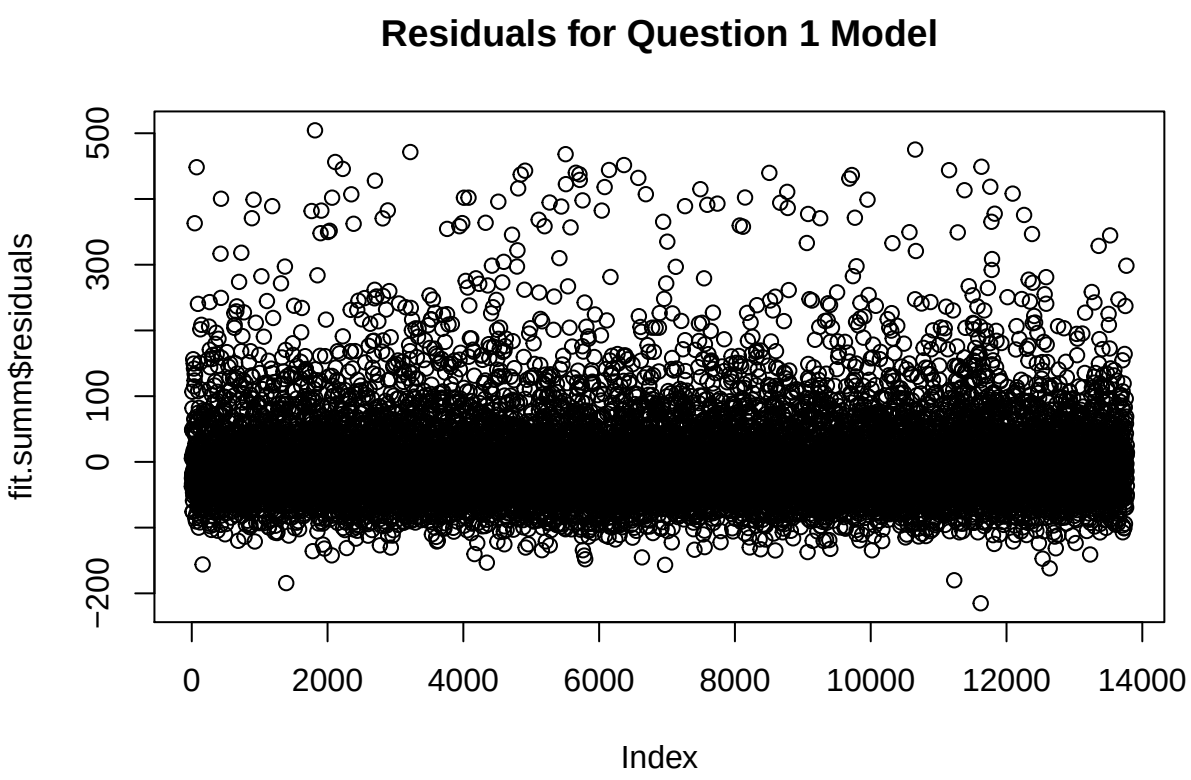
```
## Analysis of Variance Table
```

```
##
## Model 1: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted)
## Model 2: ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted + factor(YBLAdjusted) +
##     FULP
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  13761 61675311
## 2  13760 61323951  1    351360 78.839 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.summ<- summary(fit.basic.HFLYBLFULP)

round(fit.summ$coefficients,2)
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## BLDAdjustedApartment               39.80       2.35   16.90     0.00
## BLDAdjustedHouse                   76.06       2.68   28.41     0.00
## NP                                 11.89       0.45   26.43     0.00
## BDSP                               12.91       0.70   18.40     0.00
## HFLAdjustedNot Electricity        -41.96       1.29  -32.65     0.00
## factor(YBLAdjusted)1940 to 1949     6.91       2.77    2.50     0.01
## factor(YBLAdjusted)1950 to 1959     3.63       2.41    1.51     0.13
## factor(YBLAdjusted)1960 to 1969     2.47       2.39    1.03     0.30
## factor(YBLAdjusted)1970 to 1979     7.21       2.07    3.49     0.00
## factor(YBLAdjusted)1980 to 1989     6.15       2.42    2.53     0.01
## factor(YBLAdjusted)1990 to 1999    -0.59       2.14   -0.27     0.78
## factor(YBLAdjusted)2000 to 2004    -5.07       2.66   -1.91     0.06
## factor(YBLAdjusted)2005 to 2015    -7.02       2.43   -2.89     0.00
## FULP                                0.02       0.00    8.88     0.00
```

```
plot(fit.summ$residuals, main = 'Residuals for Question 1 Model')
```




```
round(confint(fit.basic.HFLYBLFULP),2)
```

```
##                                   2.5 % 97.5 %
## BLDAdjustedApartment              35.18  44.41
## BLDAdjustedHouse                  70.81  81.31
## NP                                11.01  12.77
## BDSP                              11.53  14.28
## HFLAdjustedNot Electricity      -44.48 -39.44
## factor(YBLAdjusted)1940 to 1949   1.49  12.34
## factor(YBLAdjusted)1950 to 1959  -1.08   8.35
## factor(YBLAdjusted)1960 to 1969  -2.22   7.15
## factor(YBLAdjusted)1970 to 1979   3.16  11.26
## factor(YBLAdjusted)1980 to 1989   1.39  10.90
## factor(YBLAdjusted)1990 to 1999  -4.79   3.62
## factor(YBLAdjusted)2000 to 2004 -10.28   0.13
## factor(YBLAdjusted)2005 to 2015 -11.78  -2.25
## FULP                              0.01   0.02
```

```
fit.summ
```

```
##
## Call:
## lm(formula = ELEP ~ BLDAdjusted + NP + BDSP - 1 + HFLAdjusted +
##     factor(YBLAdjusted) + FULP, data = df.final)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.04  -39.86  -13.57   22.63  504.42
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## BLDAdjustedApartment             39.797472   2.354597  16.902  < 2e-16 ***
## BLDAdjustedHouse                 76.059576   2.676958  28.413  < 2e-16 ***
## NP                               11.888388   0.449882  26.426  < 2e-16 ***
## BDSP                             12.907523   0.701427  18.402  < 2e-16 ***
## HFLAdjustedNot Electricity      -41.957828   1.285197 -32.647  < 2e-16 ***
## factor(YBLAdjusted)1940 to 1949   6.914673   2.765557   2.500 0.012421 *
## factor(YBLAdjusted)1950 to 1959   3.633474   2.406605   1.510 0.131119
## factor(YBLAdjusted)1960 to 1969   2.465777   2.391645   1.031 0.302561
## factor(YBLAdjusted)1970 to 1979   7.205902   2.065953   3.488 0.000488 ***
## factor(YBLAdjusted)1980 to 1989   6.145440   2.424971   2.534 0.011280 *
## factor(YBLAdjusted)1990 to 1999  -0.586210   2.144124  -0.273 0.784548
## factor(YBLAdjusted)2000 to 2004  -5.073202   2.655760  -1.910 0.056120 .
## factor(YBLAdjusted)2005 to 2015  -7.015364   2.429638  -2.887 0.003890 **
## FULP                              0.018342   0.002066   8.879  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.76 on 13760 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7598
## F-statistic:  3113 on 14 and 13760 DF,  p-value: < 2.2e-16
```

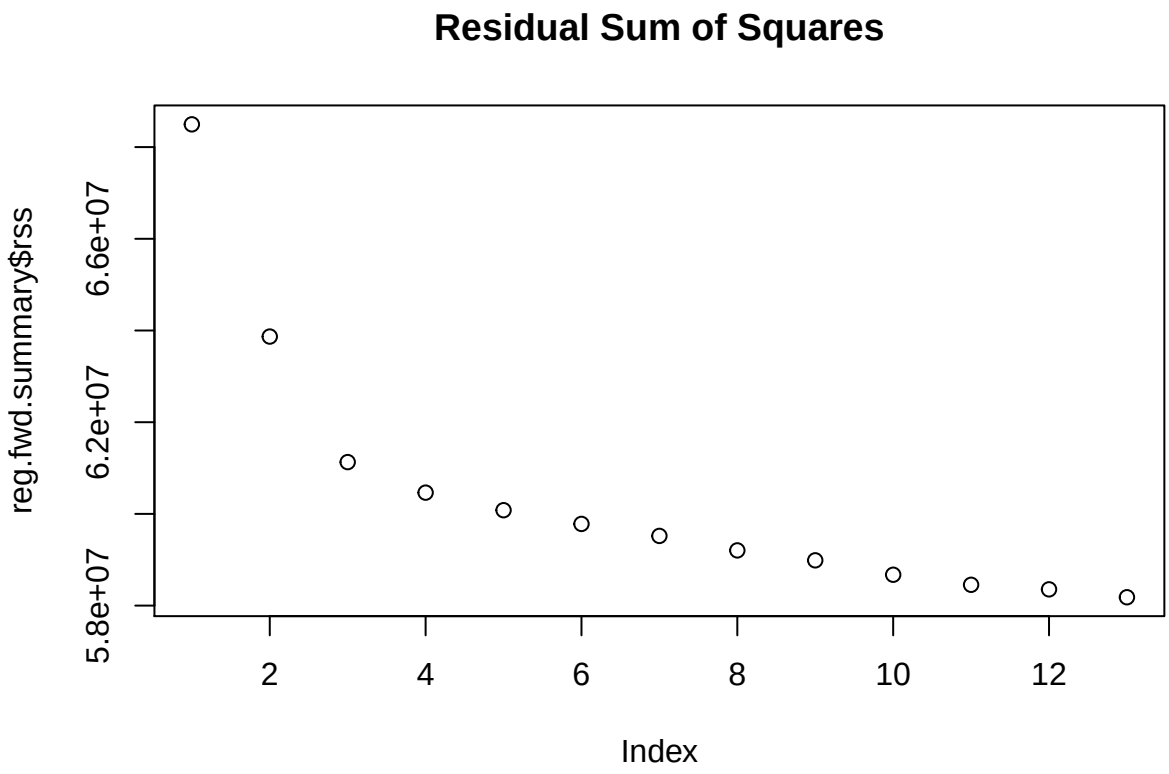
QUESTION 2

FORWARDS REGSPLIT METHOD

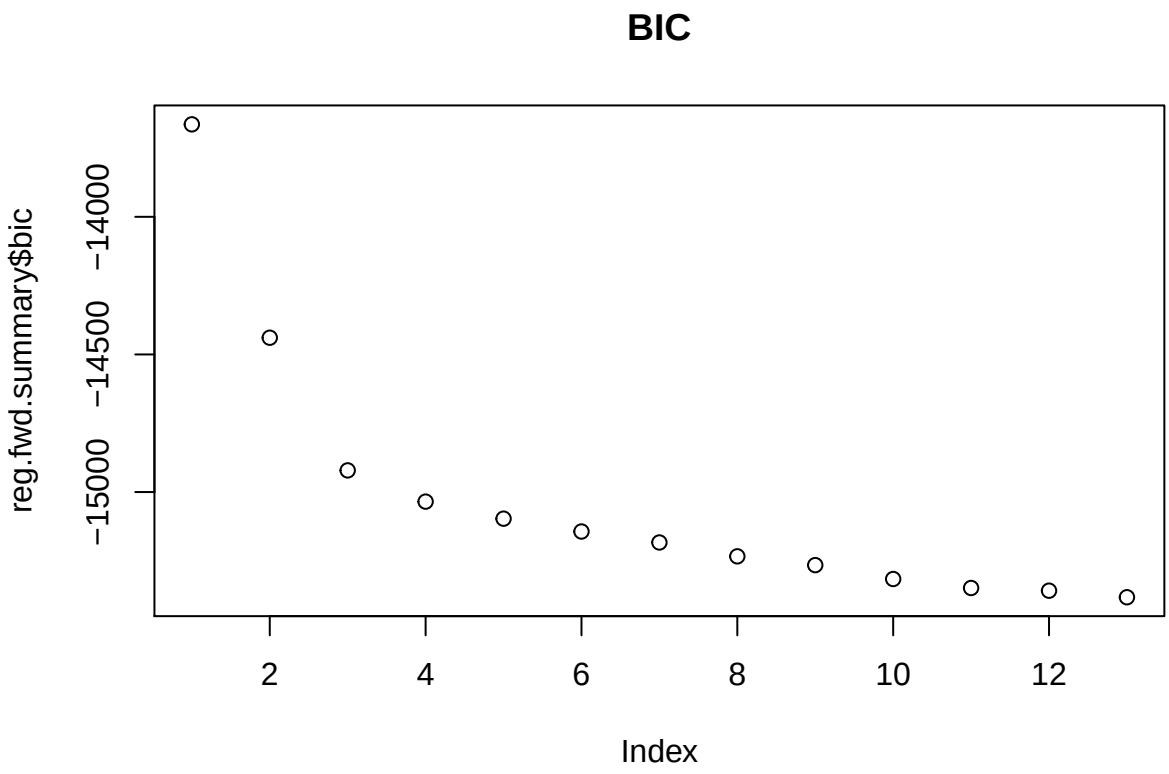
```
#Forwards Stepwise Selection
df.predict.house <- subset(df.final, BLDAdjusted == 'House', select = -c(BLDAdjusted, BLD, HFL, YBL))
regfit.fwd <- regsubsets(ELEP ~ ., data=df.predict.house, nvmax=19, intercept = FALSE, method = 'forwar
reg.fwd.summary <- summary(regfit.fwd)


names(reg.fwd.summary)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

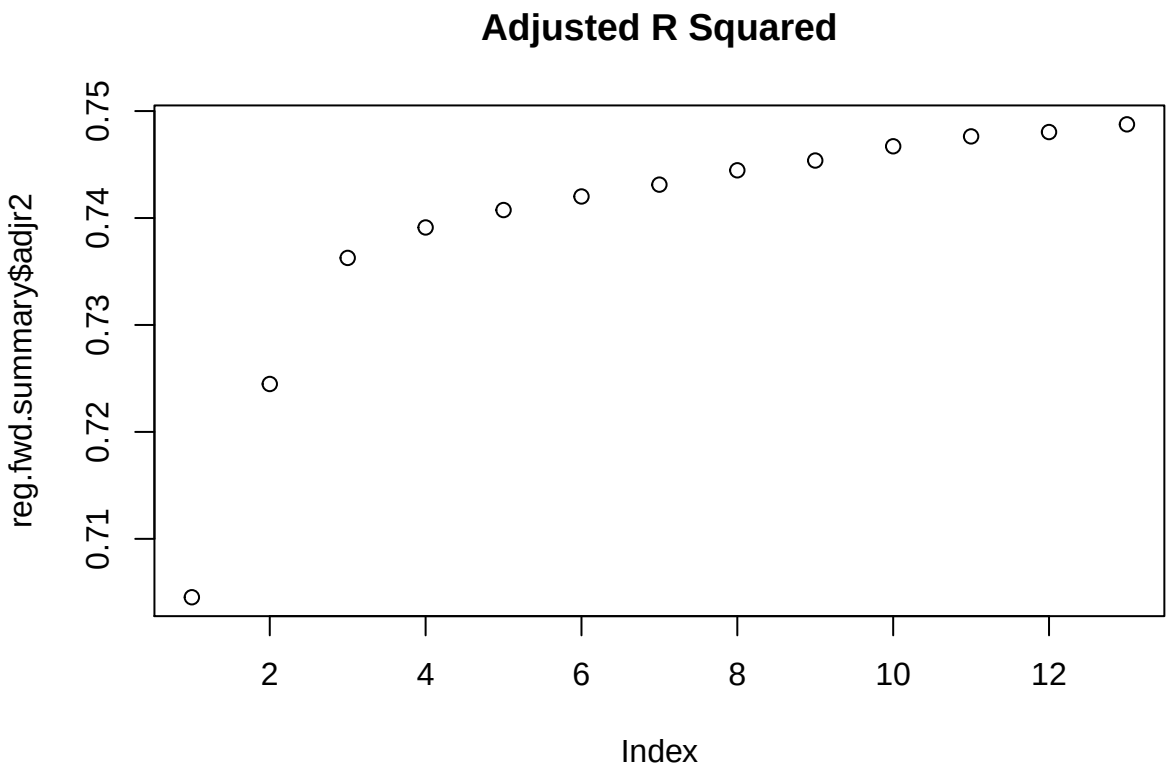
```
plot(reg.fwd.summary$rss, main = 'Residual Sum of Squares') #13 factors
```



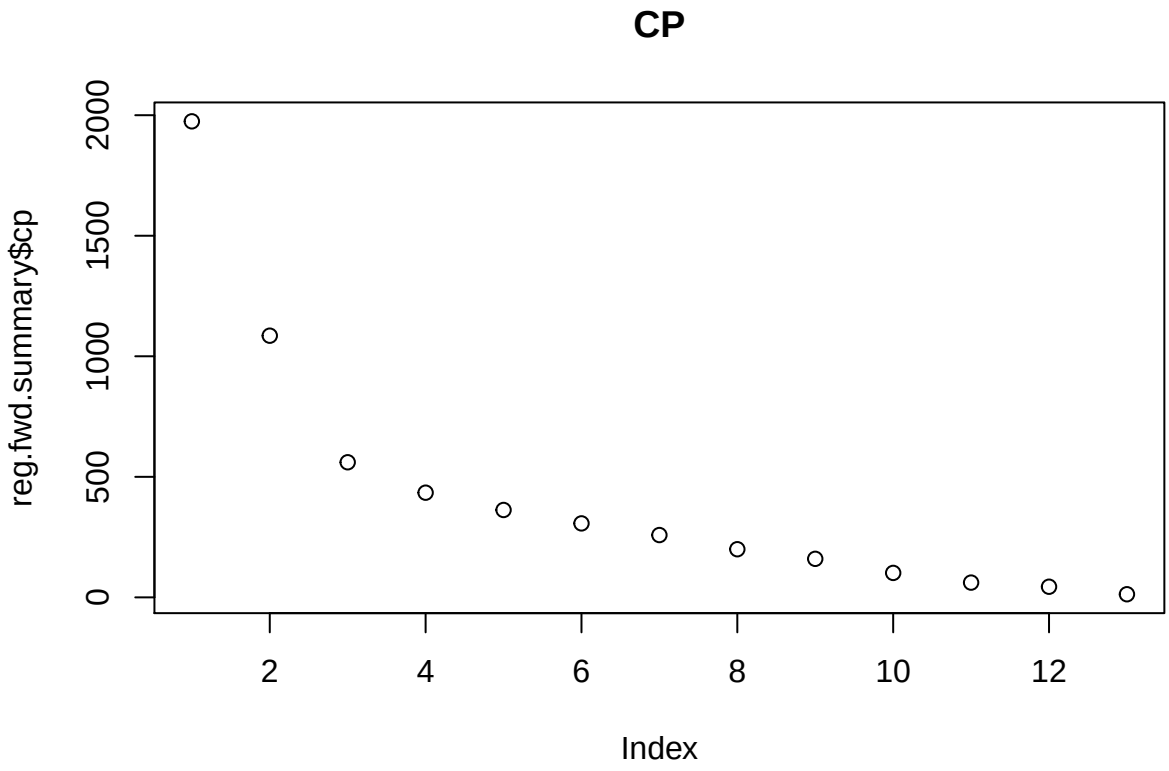

```
plot(reg.fwd.summary$bic, main = 'BIC') # 13 factors
```




```
plot(reg.fwd.summary$adjr2, main = 'Adjusted R Squared') # 13 factors
```




```
plot(reg.fwd.summary$cp, main = 'CP')# 13 factors
```




```
df.coef <- round(data.frame(coef(regfit.fwd, 13)),2)

reg.fwd.summary
```

```
## Subset selection object
## Call: regsubsets.formula(ELEP ~ ., data = df.predict.house, nvmax = 19,
##     intercept = FALSE, method = "forward")
## 13 Variables
##                             Forced in Forced out
## GASP                            FALSE      FALSE
## BDSP                            FALSE      FALSE
## NP                              FALSE      FALSE
## HFLAdjustedNot Electricity      FALSE      FALSE
## YBLAdjusted1940 to 1949         FALSE      FALSE
## YBLAdjusted1950 to 1959         FALSE      FALSE
## YBLAdjusted1960 to 1969         FALSE      FALSE
## YBLAdjusted1970 to 1979         FALSE      FALSE
## YBLAdjusted1980 to 1989         FALSE      FALSE
## YBLAdjusted1990 to 1999         FALSE      FALSE
## YBLAdjusted2000 to 2004         FALSE      FALSE
## YBLAdjusted2005 to 2015         FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: forward
##          FULP GASP BDSP NP  HFLAdjustedNot Electricity YBLAdjusted1940 to 1949
## 1  ( 1 )  " "  " "  "*"  " " " " "                          " "
```

```
## 2  ( 1 )  " "  " "  "*"  "*" " "                          " "
## 3  ( 1 )  " "  " "  "*"  "*" "*"                          " "
## 4  ( 1 )  "*"  " "  "*"  "*" "*"                          " "
## 5  ( 1 )  "*"  " "  "*"  "*" "*"                          " "
## 6  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 7  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 8  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 9  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 10  ( 1 ) "*"  " "  "*"  "*" "*"                          "*"
## 11  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
## 12  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
## 13  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
##           YBLAdjusted1950 to 1959 YBLAdjusted1960 to 1969
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  " "                     " "
## 6  ( 1 )  " "                     " "
## 7  ( 1 )  " "                     " "
## 8  ( 1 )  "*"                     " "
## 9  ( 1 )  "*"                     "*"
## 10  ( 1 ) "*"                     "*"
## 11  ( 1 ) "*"                     "*"
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
##           YBLAdjusted1970 to 1979 YBLAdjusted1980 to 1989
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  "*"                     " "
## 6  ( 1 )  "*"                     " "
## 7  ( 1 )  "*"                     "*"
## 8  ( 1 )  "*"                     "*"
## 9  ( 1 )  "*"                     "*"
## 10  ( 1 ) "*"                     "*"
## 11  ( 1 ) "*"                     "*"
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
##           YBLAdjusted1990 to 1999 YBLAdjusted2000 to 2004
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  " "                     " "
## 6  ( 1 )  " "                     " "
## 7  ( 1 )  " "                     " "
## 8  ( 1 )  " "                     " "
## 9  ( 1 )  " "                     " "
## 10  ( 1 ) "*"                     " "
## 11  ( 1 ) "*"                     " "
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
```

```
##           YBLAdjusted2005 to 2015
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
## 11  ( 1 ) " "
## 12  ( 1 ) " "
## 13  ( 1 ) "*"
```

```
print(paste("RSS:", round(reg.fwd.summary$rss[13], 2)))
```

```
## [1] "RSS: 58182444.45"
```

```
print(paste("BIC:", round(reg.fwd.summary$bic[13], 2)))
```

```
## [1] "BIC: -15382.14"
```

```
print(paste("Adjusted R Squared:", round(reg.fwd.summary$adjr2[13], 2)))
```

```
## [1] "Adjusted R Squared: 0.75"
```

```
print(paste("CP:", round(reg.fwd.summary$cp[12], 2)))
```

```
## [1] "CP: 44.13"
```

```
regfit.fwd$ress
```

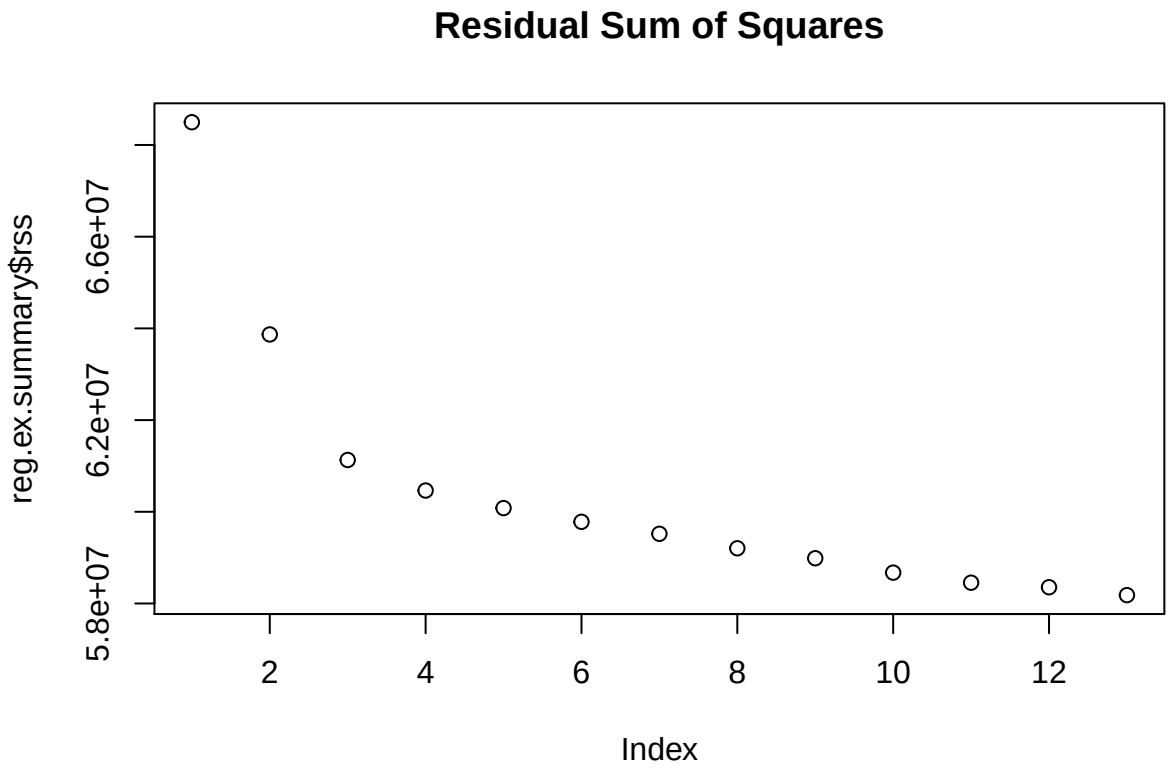
```
##           [,1]
##  [1,] 68495797
##  [2,] 63867596
##  [3,] 61129483
##  [4,] 60463945
##  [5,] 60080784
##  [6,] 59781503
##  [7,] 59519998
##  [8,] 59204121
##  [9,] 58986972
## [10,] 58673070
## [11,] 58453850
## [12,] 58354510
## [13,] 58182444
```

EXHAUSTIVE REGSPLIT METHOD

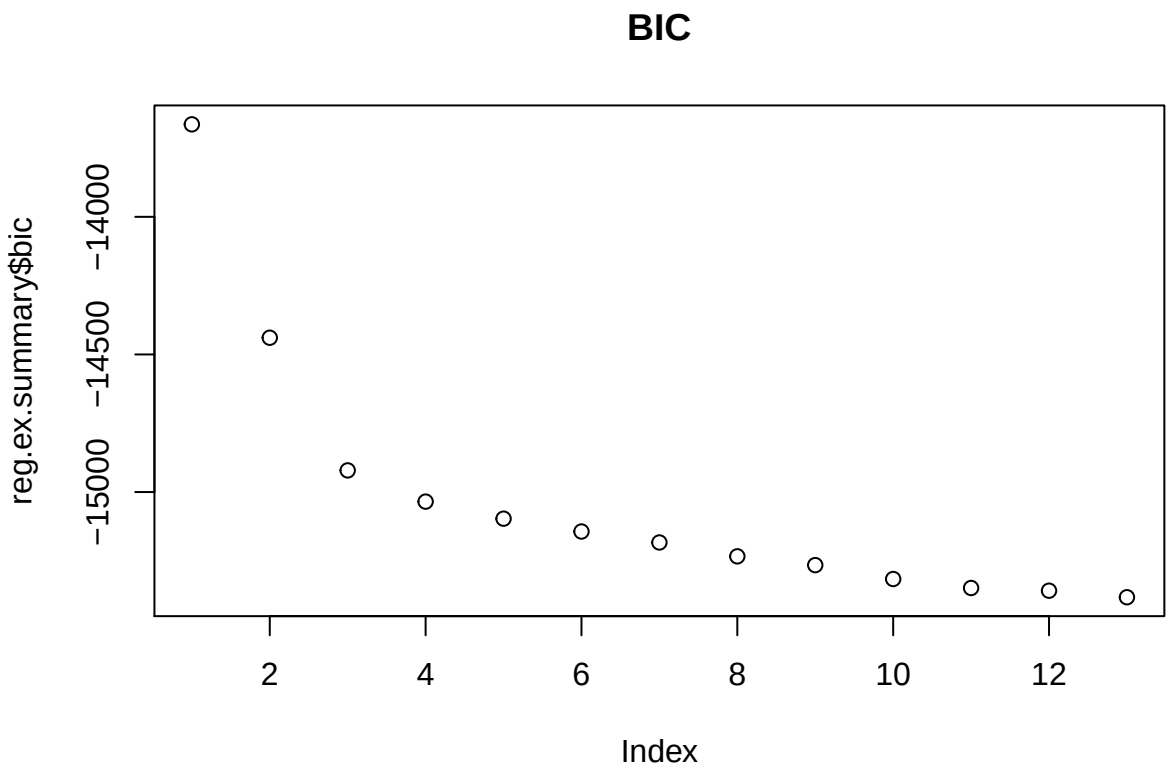
```r
#Exhaustive stepwise selection
regfit.ex <- regsubsets(ELEP ~., data = df.predict.house,  nvmax = 19, method = 'exhaustive', intercept
reg.ex.summary <- summary(regfit.ex)

names(reg.ex.summary)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

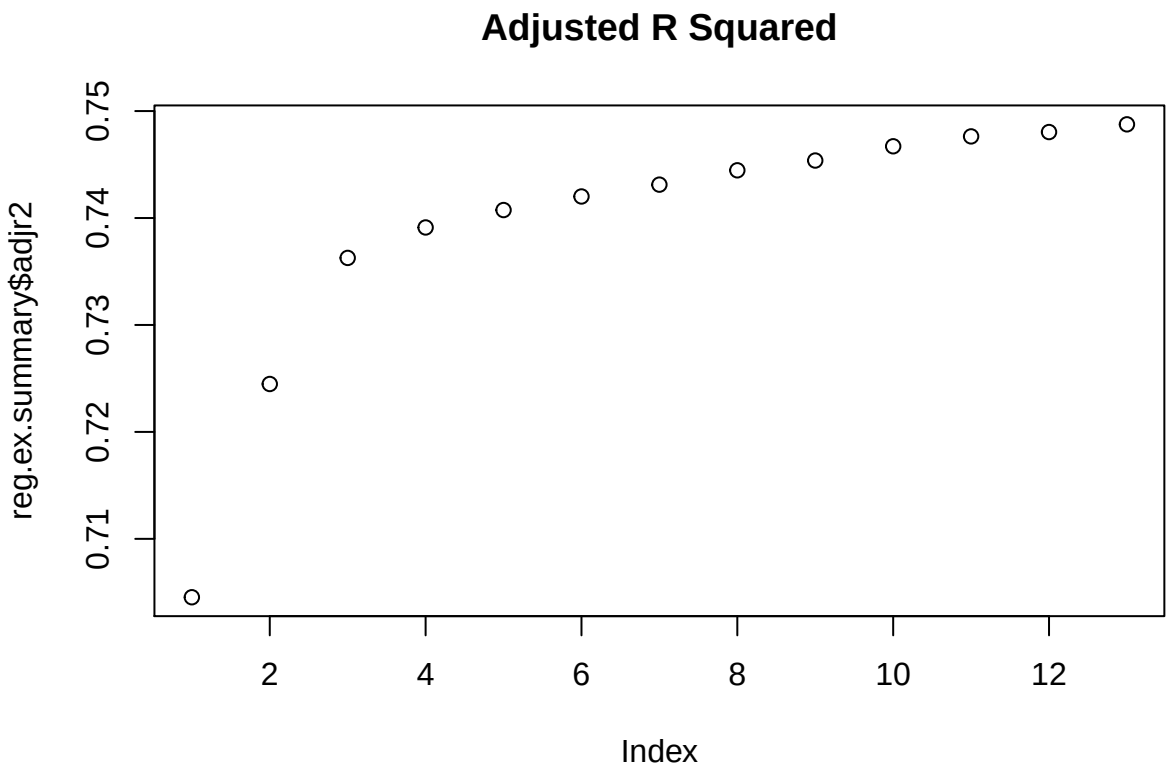
```r
plot(reg.ex.summary$rss, main = 'Residual Sum of Squares') # Best at 13
```



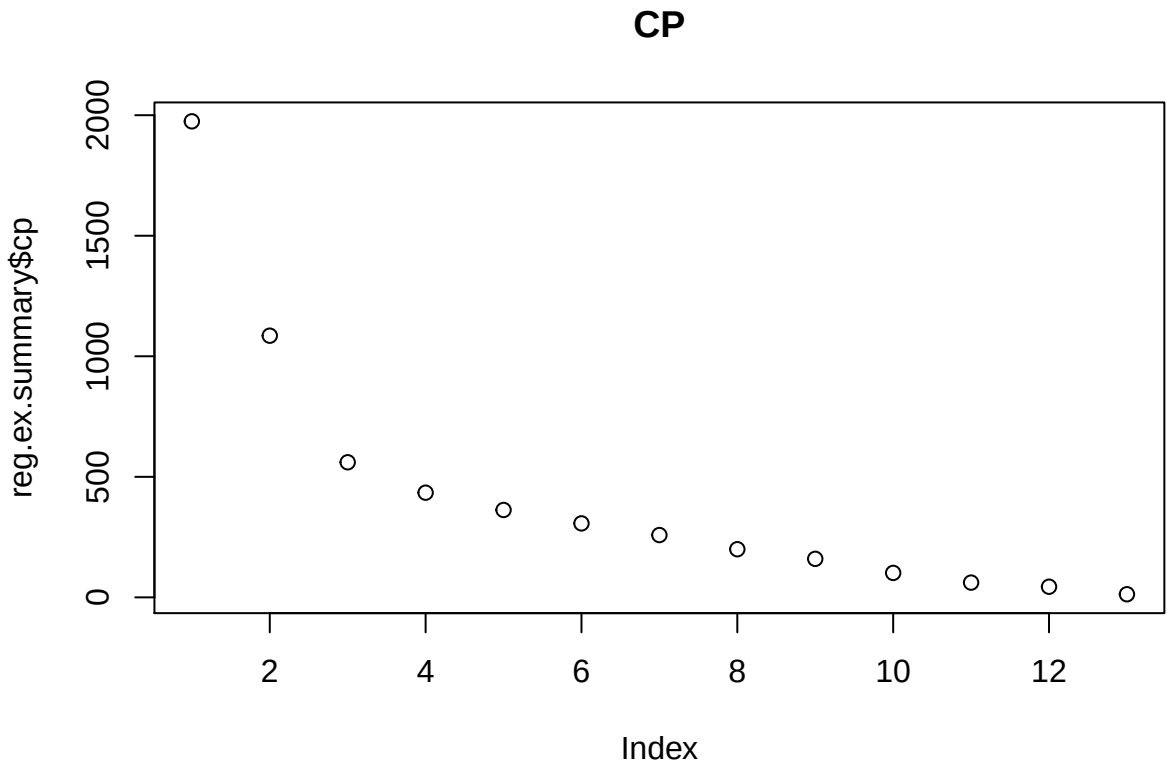

```r
plot(reg.ex.summary$bic, main = 'BIC') # Best at 13
```




```
plot(reg.ex.summary$adjr2, main = 'Adjusted R Squared') # Best at 13
```




```
plot(reg.ex.summary$cp, main = 'CP') # Best at 13
```




```
df.coef.ex <- round(data.frame(coef(regfit.ex, 13)),2)

reg.ex.summary
```

```
## Subset selection object
## Call: regsubsets.formula(ELEP ~ ., data = df.predict.house, nvmax = 19,
##     method = "exhaustive", intercept = FALSE)
## 13 Variables
##                             Forced in Forced out
## GASP                            FALSE      FALSE
## BDSP                            FALSE      FALSE
## NP                              FALSE      FALSE
## HFLAdjustedNot Electricity      FALSE      FALSE
## YBLAdjusted1940 to 1949         FALSE      FALSE
## YBLAdjusted1950 to 1959         FALSE      FALSE
## YBLAdjusted1960 to 1969         FALSE      FALSE
## YBLAdjusted1970 to 1979         FALSE      FALSE
## YBLAdjusted1980 to 1989         FALSE      FALSE
## YBLAdjusted1990 to 1999         FALSE      FALSE
## YBLAdjusted2000 to 2004         FALSE      FALSE
## YBLAdjusted2005 to 2015         FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##          FULP GASP BDSP NP  HFLAdjustedNot Electricity YBLAdjusted1940 to 1949
## 1  ( 1 )  " "  " "  "*"  " " " " "                          " "
```

```
## 2  ( 1 )  " "  " "  "*"  "*" " "                          " "
## 3  ( 1 )  " "  " "  "*"  "*" "*"                          " "
## 4  ( 1 )  "*"  " "  "*"  "*" "*"                          " "
## 5  ( 1 )  "*"  " "  "*"  "*" "*"                          " "
## 6  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 7  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 8  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 9  ( 1 )  "*"  " "  "*"  "*" "*"                          "*"
## 10  ( 1 ) "*"  " "  "*"  "*" "*"                          "*"
## 11  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
## 12  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
## 13  ( 1 ) "*"  "*"  "*"  "*" "*"                          "*"
##           YBLAdjusted1950 to 1959 YBLAdjusted1960 to 1969
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  " "                     " "
## 6  ( 1 )  " "                     " "
## 7  ( 1 )  " "                     " "
## 8  ( 1 )  "*"                     " "
## 9  ( 1 )  "*"                     "*"
## 10  ( 1 ) "*"                     "*"
## 11  ( 1 ) "*"                     "*"
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
##           YBLAdjusted1970 to 1979 YBLAdjusted1980 to 1989
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  "*"                     " "
## 6  ( 1 )  "*"                     " "
## 7  ( 1 )  "*"                     "*"
## 8  ( 1 )  "*"                     "*"
## 9  ( 1 )  "*"                     "*"
## 10  ( 1 ) "*"                     "*"
## 11  ( 1 ) "*"                     "*"
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
##           YBLAdjusted1990 to 1999 YBLAdjusted2000 to 2004
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  " "                     " "
## 5  ( 1 )  " "                     " "
## 6  ( 1 )  " "                     " "
## 7  ( 1 )  " "                     " "
## 8  ( 1 )  " "                     " "
## 9  ( 1 )  " "                     " "
## 10  ( 1 ) "*"                     " "
## 11  ( 1 ) "*"                     " "
## 12  ( 1 ) "*"                     "*"
## 13  ( 1 ) "*"                     "*"
```

```
##           YBLAdjusted2005 to 2015
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
## 11  ( 1 ) " "
## 12  ( 1 ) " "
## 13  ( 1 ) "*"
```

```
print(paste("RSS:", round(reg.ex.summary$rss[13], 2)))
```

```
## [1] "RSS: 58182444.45"
```

```
print(paste("BIC:", round(reg.ex.summary$bic[13], 2)))
```

```
## [1] "BIC: -15382.14"
```

```
print(paste("Adjusted R Squared:", round(reg.ex.summary$adjr2[13], 2)))
```

```
## [1] "Adjusted R Squared: 0.75"
```

```
print(paste("CP:", round(reg.ex.summary$cp[12], 2)))
```

```
## [1] "CP: 44.13"
```

```
#shinypairs(df.predict.house) # computationally expensive, only comment out when necessary
```