AMS578_project_116125547

Ajeetkumar Rai (Id : 116125547)

2025-04-30

#Library

```
required_packages <- c("tidyverse", "caret", "neuralnet", "ggplot2", "glmnet","rpart", "rattle", "facto

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)}
  library(pkg, character.only = TRUE)}
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error

## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift

## Warning: package 'neuralnet' was built under R version 4.3.3

##
## Attaching package: 'neuralnet'
##
## The following object is masked from 'package:dplyr':
##
##     compute

## Warning: package 'glmnet' was built under R version 4.3.3

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8

## Warning: package 'rpart' was built under R version 4.3.3

## Warning: package 'rattle' was built under R version 4.3.3

## Loading required package: bitops

## Warning: package 'bitops' was built under R version 4.3.3

##
## Attaching package: 'bitops'
##
## The following object is masked from 'package:Matrix':
##
##     %&%
##
## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

## Warning: package 'factoextra' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
## Warning: package 'gridExtra' was built under R version 4.3.3

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine

## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.95 loaded

## Warning: package 'lmtest' was built under R version 4.3.3

## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
##
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Warning: package 'forecast' was built under R version 4.3.3

## Warning: package 'e1071' was built under R version 4.3.3
```

Data

```
df <- read.csv("C:/Users/Ajeet Rai/OneDrive/Desktop/SBU/Academics/Sem III/AMS 578 Regression Theory/Pro
head(df)
```

```
##     State  County TotalPop   Men Women Hispanic White Black Native Asian
## 1 Alabama Autauga    55221 26745 28476      2.6  75.8  18.5    0.4   1.0
## 2 Alabama Baldwin   195121 95314 99807      4.5  83.1   9.5    0.6   0.7
## 3 Alabama Barbour    26932 14497 12435      4.6  46.2  46.7    0.2   0.4
## 4 Alabama    Bibb    22604 12073 10531      2.2  74.5  21.4    0.4   0.1
## 5 Alabama  Blount    57710 28512 29198      8.6  87.9   1.5    0.3   0.1
## 6 Alabama Bullock    10678  5660  5018      4.4  22.2  70.7    1.2   0.2
##   Pacific Citizen Income IncomeErr IncomePerCap IncomePerCapErr Poverty
## 1       0   40725  51281      2391        24974            1080    12.9
## 2       0  147695  50254      1263        27317             711    13.4
## 3       0   20714  32964      2973        16824             798    26.7
## 4       0   17495  38678      3995        18431            1618    16.8
## 5       0   42345  45813      3141        20532             708    16.7
## 6       0    8057  31938      5884        17580            2055    24.6
##   ChildPoverty Professional Service Office Construction Production Drive
## 1         18.6         33.2    17.0   24.2          8.6       17.1  87.5
## 2         19.2         33.1    17.7   27.1         10.8       11.2  84.7
## 3         45.3         26.8    16.1   23.1         10.8       23.1  83.8
## 4         27.9         21.5    17.9   17.8         19.0       23.7  83.2
## 5         27.2         28.5    14.1   23.9         13.5       19.9  84.9
## 6         38.4         18.8    15.0   19.7         20.1       26.4  74.9
##   Carpool Transit Walk OtherTransp WorkAtHome MeanCommute Employed PrivateWork
## 1     8.8     0.1  0.5         1.3        1.8        26.5    23986        73.6
## 2     8.8     0.1  1.0         1.4        3.9        26.4    85953        81.5
## 3    10.9     0.4  1.8         1.5        1.6        24.1     8597        71.8
## 4    13.5     0.5  0.6         1.5        0.7        28.8     8294        76.8
## 5    11.2     0.4  0.9         0.4        2.3        34.9    22189        82.0
## 6    14.9     0.7  5.0         1.7        2.8        27.5     3865        79.5
##   PublicWork SelfEmployed FamilyWork Unemployment
## 1       20.9          5.5        0.0          7.6
## 2       12.3          5.8        0.4          7.5
## 3       20.8          7.3        0.1         17.6
## 4       16.1          6.7        0.4          8.3
## 5       13.5          4.2        0.4          7.7
## 6       15.1          5.4        0.0         18.0
```

Summary

Shape

```
dim(df)
```

```
## [1] 3220   36
```

Datatypes

```
str(df)
```

```
## 'data.frame':    3220 obs. of  36 variables:
##  $ State          : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ County         : chr  "Autauga" "Baldwin" "Barbour" "Bibb" ...
##  $ TotalPop       : int  55221 195121 26932 22604 57710 10678 20354 116648 34079 26008 ...
##  $ Men            : int  26745 95314 14497 12073 28512 5660 9502 56274 16258 12975 ...
##  $ Women          : int  28476 99807 12435 10531 29198 5018 10852 60374 17821 13033 ...
##  $ Hispanic       : num  2.6 4.5 4.6 2.2 8.6 4.4 1.2 3.5 0.4 1.5 ...
##  $ White          : num  75.8 83.1 46.2 74.5 87.9 22.2 53.3 73 57.3 91.7 ...
##  $ Black          : num  18.5 9.5 46.7 21.4 1.5 70.7 43.8 20.3 40.3 4.8 ...
##  $ Native         : num  0.4 0.6 0.2 0.4 0.3 1.2 0.1 0.2 0.2 0.6 ...
##  $ Asian          : num  1 0.7 0.4 0.1 0.1 0.2 0.4 0.9 0.8 0.3 ...
##  $ Pacific        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Citizen        : int  40725 147695 20714 17495 42345 8057 15581 88612 26462 20600 ...
##  $ Income         : int  51281 50254 32964 38678 45813 31938 32229 41703 34177 36296 ...
##  $ IncomeErr      : int  2391 1263 2973 3995 3141 5884 1793 925 2949 1710 ...
##  $ IncomePerCap   : int  24974 27317 16824 18431 20532 17580 18390 21374 21071 21811 ...
##  $ IncomePerCapErr: int  1080 711 798 1618 708 2055 714 489 1366 1556 ...
##  $ Poverty        : num  12.9 13.4 26.7 16.8 16.7 24.6 25.4 20.5 21.6 19.2 ...
##  $ ChildPoverty   : num  18.6 19.2 45.3 27.9 27.2 38.4 39.2 31.6 37.2 30.1 ...
##  $ Professional   : num  33.2 33.1 26.8 21.5 28.5 18.8 27.5 27.3 23.3 29.3 ...
##  $ Service        : num  17 17.7 16.1 17.9 14.1 15 16.6 17.7 14.5 16 ...
##  $ Office         : num  24.2 27.1 23.1 17.8 23.9 19.7 21.9 24.2 26.3 19.5 ...
##  $ Construction   : num  8.6 10.8 10.8 19 13.5 20.1 10.3 10.5 11.5 13.7 ...
##  $ Production     : num  17.1 11.2 23.1 23.7 19.9 26.4 23.7 20.4 24.4 21.5 ...
##  $ Drive          : num  87.5 84.7 83.8 83.2 84.9 74.9 84.5 85.3 85.1 83.9 ...
##  $ Carpool        : num  8.8 8.8 10.9 13.5 11.2 14.9 12.4 9.4 11.9 12.1 ...
##  $ Transit        : num  0.1 0.1 0.4 0.5 0.4 0.7 0 0.2 0.2 0.2 ...
##  $ Walk           : num  0.5 1 1.8 0.6 0.9 5 0.8 1.2 0.3 0.6 ...
##  $ OtherTransp    : num  1.3 1.4 1.5 1.5 0.4 1.7 0.6 1.2 0.4 0.7 ...
##  $ WorkAtHome     : num  1.8 3.9 1.6 0.7 2.3 2.8 1.7 2.7 2.1 2.5 ...
##  $ MeanCommute    : num  26.5 26.4 24.1 28.8 34.9 27.5 24.6 24.1 25.1 27.4 ...
##  $ Employed       : int  23986 85953 8597 8294 22189 3865 7813 47401 13689 10155 ...
##  $ PrivateWork    : num  73.6 81.5 71.8 76.8 82 79.5 77.4 74.1 85.1 73.1 ...
##  $ PublicWork     : num  20.9 12.3 20.8 16.1 13.5 15.1 16.2 20.8 12.1 18.5 ...
##  $ SelfEmployed   : num  5.5 5.8 7.3 6.7 4.2 5.4 6.2 5 2.8 7.9 ...
##  $ FamilyWork     : num  0 0.4 0.1 0.4 0.4 0 0.2 0.1 0 0.5 ...
##  $ Unemployment   : num  7.6 7.5 17.6 8.3 7.7 18 10.9 12.3 8.9 7.9 ...
```

Descriptions

```
summary(df)
```

```
##     State              County            TotalPop             Men          
##  Length:3220        Length:3220        Min.   :      85   Min.   :     42  
##  Class :character   Class :character   1st Qu.:   11218   1st Qu.:   5637  
##  Mode  :character   Mode  :character   Median :   26035   Median :  12932  
```

```
##                                        Mean   :   99409   Mean   :  48897
##                                        3rd Qu.:   66430   3rd Qu.:  32993
##                                        Max.   :10038388   Max.   :4945351
##
##      Women            Hispanic          White           Black
##  Min.   :     43   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:   5572   1st Qu.: 1.900   1st Qu.:64.10   1st Qu.: 0.500
##  Median :  13057   Median : 3.900   Median :84.10   Median : 1.900
##  Mean   :  50512   Mean   :11.012   Mean   :75.43   Mean   : 8.665
##  3rd Qu.:  33488   3rd Qu.: 9.825   3rd Qu.:93.20   3rd Qu.: 9.600
##  Max.   :5093037   Max.   :99.900   Max.   :99.80   Max.   :85.900
##
##      Native           Asian           Pacific            Citizen
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00000   Min.   :     80
##  1st Qu.: 0.100   1st Qu.: 0.200   1st Qu.: 0.00000   1st Qu.:   8450
##  Median : 0.300   Median : 0.500   Median : 0.00000   Median :  19643
##  Mean   : 1.724   Mean   : 1.229   Mean   : 0.08273   Mean   :  69935
##  3rd Qu.: 0.600   3rd Qu.: 1.200   3rd Qu.: 0.00000   3rd Qu.:  49920
##  Max.   :92.100   Max.   :41.600   Max.   :35.30000   Max.   :6046749
##
##      Income         IncomeErr      IncomePerCap   IncomePerCapErr
##  Min.   : 10499   Min.   :  270   Min.   : 5878   Min.   :  113
##  1st Qu.: 38192   1st Qu.: 1635   1st Qu.:20239   1st Qu.:  755
##  Median : 44749   Median : 2406   Median :23460   Median : 1096
##  Mean   : 46130   Mean   : 2850   Mean   :23982   Mean   : 1363
##  3rd Qu.: 52074   3rd Qu.: 3446   3rd Qu.:27053   3rd Qu.: 1631
##  Max.   :123453   Max.   :21355   Max.   :65600   Max.   :15266
##  NA's   :1        NA's   :1
##     Poverty       ChildPoverty    Professional      Service
##  Min.   : 1.40   Min.   : 0.00   Min.   :13.50   Min.   : 5.00
##  1st Qu.:12.10   1st Qu.:16.30   1st Qu.:26.70   1st Qu.:16.00
##  Median :16.15   Median :22.70   Median :29.90   Median :18.10
##  Mean   :17.49   Mean   :24.18   Mean   :30.99   Mean   :18.35
##  3rd Qu.:20.70   3rd Qu.:30.00   3rd Qu.:34.40   3rd Qu.:20.30
##  Max.   :64.20   Max.   :81.60   Max.   :74.00   Max.   :38.20
##                  NA's   :1
##      Office       Construction     Production        Drive
##  Min.   : 4.10   Min.   : 1.70   Min.   : 0.00   Min.   : 5.20
##  1st Qu.:20.20   1st Qu.: 9.80   1st Qu.:11.50   1st Qu.:76.60
##  Median :22.40   Median :12.10   Median :15.25   Median :80.70
##  Mean   :22.22   Mean   :12.71   Mean   :15.73   Mean   :79.18
##  3rd Qu.:24.40   3rd Qu.:14.90   3rd Qu.:19.32   3rd Qu.:83.70
##  Max.   :35.40   Max.   :40.30   Max.   :55.60   Max.   :94.60
##
##     Carpool         Transit             Walk         OtherTransp
##  Min.   : 0.00   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 8.40   1st Qu.: 0.1000   1st Qu.: 1.400   1st Qu.: 0.900
##  Median : 9.90   Median : 0.4000   Median : 2.400   Median : 1.300
##  Mean   :10.28   Mean   : 0.9718   Mean   : 3.324   Mean   : 1.613
##  3rd Qu.:11.80   3rd Qu.: 0.8000   3rd Qu.: 4.000   3rd Qu.: 1.900
##  Max.   :29.90   Max.   :61.7000   Max.   :71.200   Max.   :39.100
##
##    WorkAtHome      MeanCommute       Employed        PrivateWork
##  Min.   : 0.000   Min.   : 4.90   Min.   :     62   Min.   :25.00
```

```
##  1st Qu.: 2.700   1st Qu.:19.50   1st Qu.:   4551   1st Qu.:70.50
##  Median : 3.900   Median :23.00   Median :  10508   Median :75.70
##  Mean   : 4.632   Mean   :23.28   Mean   :  45594   Mean   :74.22
##  3rd Qu.: 5.600   3rd Qu.:26.80   3rd Qu.:  28633   3rd Qu.:79.70
##  Max.   :37.200   Max.   :44.00   Max.   :4635465   Max.   :88.30
##
##    PublicWork     SelfEmployed      FamilyWork      Unemployment
##  Min.   : 5.80   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:13.10   1st Qu.: 5.400   1st Qu.:0.1000   1st Qu.: 5.500
##  Median :16.20   Median : 6.900   Median :0.2000   Median : 7.600
##  Mean   :17.56   Mean   : 7.932   Mean   :0.2881   Mean   : 8.094
##  3rd Qu.:20.50   3rd Qu.: 9.400   3rd Qu.:0.3000   3rd Qu.: 9.900
##  Max.   :66.20   Max.   :36.600   Max.   :9.8000   Max.   :36.500
##
```

Imputing missing values

```
dim(df)
```

```
## [1] 3220   36
```

```
df = na.omit(df)
dim(df)
```

```
## [1] 3218   36
```

Independent vs dependent

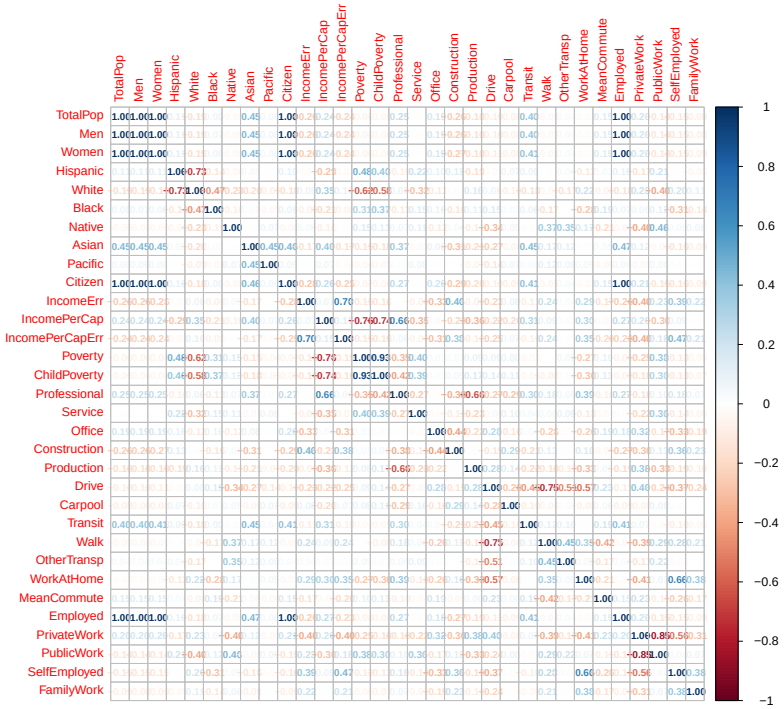
```
Y1 <- df$Income
Y2 <- df$Unemployment
Y=Y1
X <- df[, !(names(df) %in% c("Income", "Unemployment","State","County"))]
```

Exploratory data analysis

Correlation

As we can see collinearity between features, we'll need check VIF carefully.

```
correlation_matrix <- cor(X)
corrplot(correlation_matrix, method = "number", number.cex = 0.7, tl.cex = 0.8)
```



Distributin (Dependent)

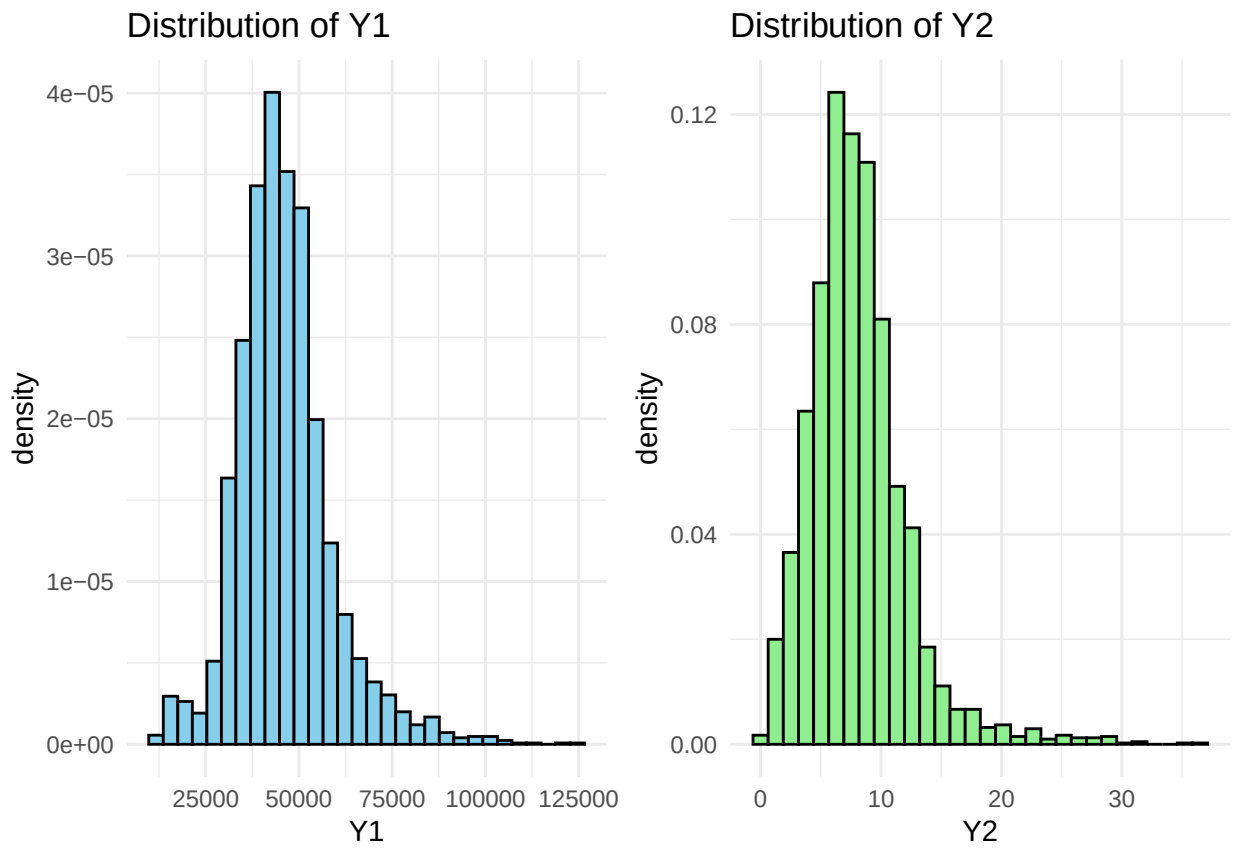
Y1 and Y2 are slightly skewed, and transformation is needed to prevent this.

```
p1 <- ggplot(data = data.frame(Y1), aes(x = Y1)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'skyblue', color = 'black') +
  ggtitle('Distribution of Y1') +
  theme_minimal()

p2 <- ggplot(data = data.frame(Y2), aes(x = Y2)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightgreen', color = 'black') +
  ggtitle('Distribution of Y2') +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



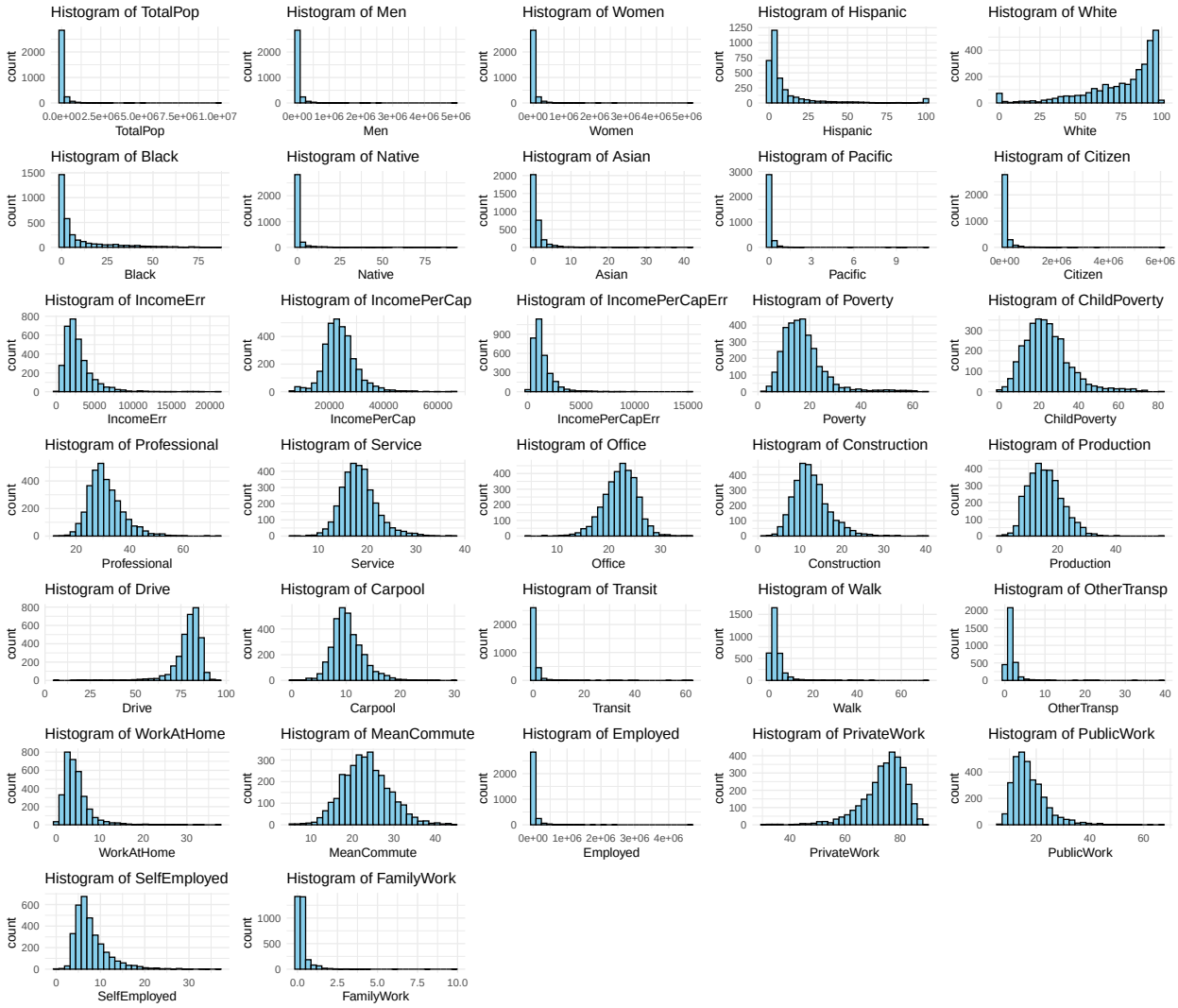
Distributin (Independent)

Independent features are skewed, and transformation is needed to prevent this.

```
plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(x = colname)) +
      geom_histogram(bins = 30, fill = "skyblue", color = "black") +
      ggtitle(paste("Histogram of", colname)) +
      theme_minimal() } else {
    ggplot(X, aes_string(x = colname)) +
      geom_bar(fill = "lightgreen", color = "black") +
      ggtitle(paste("Bar Plot of", colname)) +
      theme_minimal()}})
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
grid.arrange(grobs = plots, ncol = 5)
```

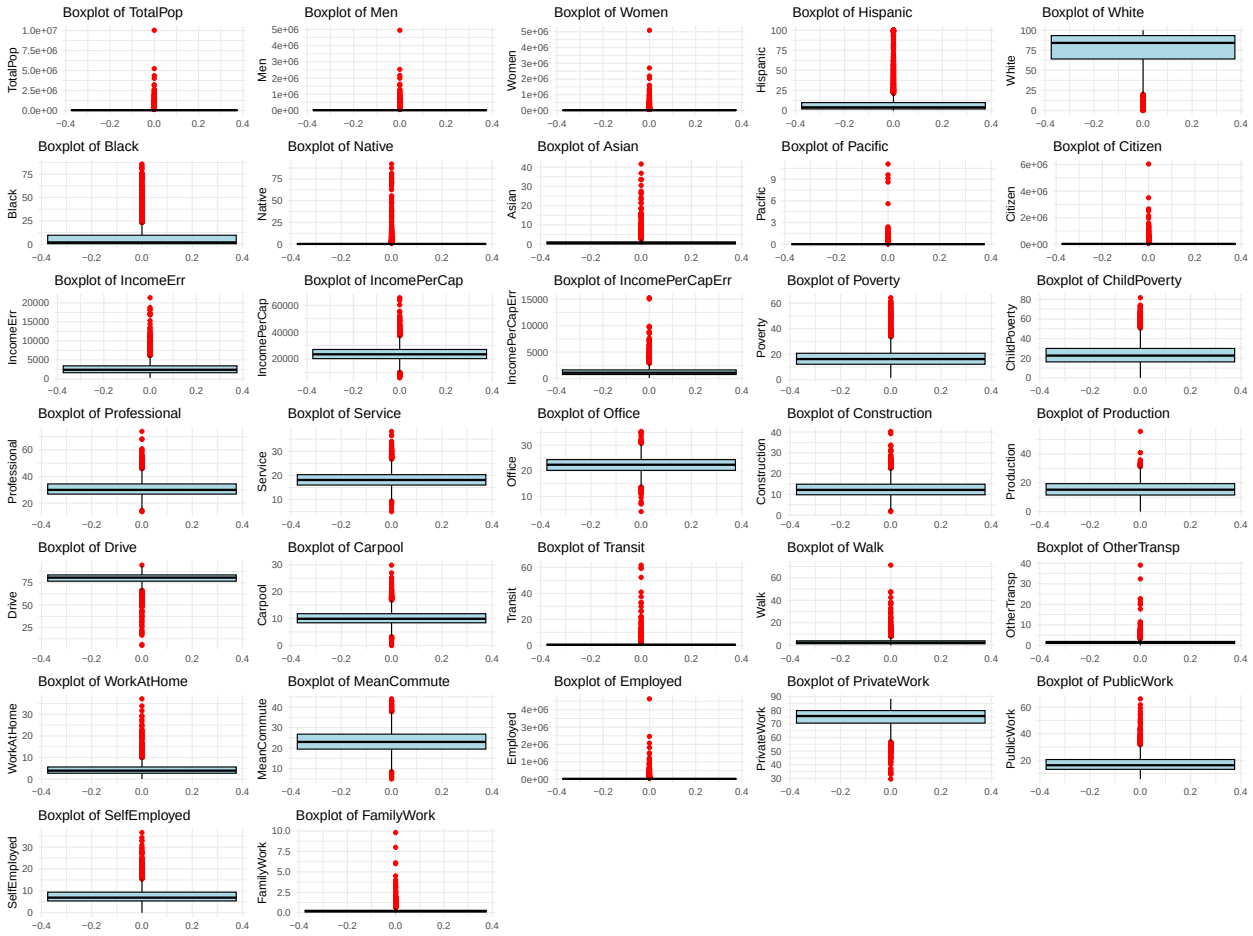


Outliers

Need careful treatment for outliers as it will effect cook's distance

```
plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(y = colname)) +
      geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
      ggtitle(paste("Boxplot of", colname)) +
      theme_minimal()
  } else {
    NULL
  }
})
plots <- Filter(Negate(is.null), plots)
```

```
grid.arrange(grobs = plots, ncol = 5)
```



Feature selection

As we have seen above so many features are correlated and among 34 features many of them are not contributing in Y1/Y2.

So, we will use STEP wise model in both direction to selected only meaningful features.

```
full_model <- lm(Y ~ ., data = X)
stepwise_model_both <- step(full_model, direction = "both")
```

```
## Start:  AIC=53496.15
## Y ~ TotalPop + Men + Women + Hispanic + White + Black + Native +
##     Asian + Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Service + Office +
##     Construction + Production + Drive + Carpool + Transit + Walk +
```

```
##     OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##     PublicWork + SelfEmployed + FamilyWork
##
##
## Step:  AIC=53496.15
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Service + Office +
##     Construction + Production + Drive + Carpool + Transit + Walk +
##     OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##     PublicWork + SelfEmployed + FamilyWork
##
##                   Df  Sum of Sq        RSS   AIC
## - Service          1 1.9808e+05 5.2321e+10 53494
## - Office           1 9.5063e+05 5.2322e+10 53494
## - Production       1 1.2568e+06 5.2322e+10 53494
## - Transit          1 1.4236e+06 5.2322e+10 53494
## - OtherTransp      1 1.5730e+06 5.2323e+10 53494
## - Construction     1 1.8211e+06 5.2323e+10 53494
## - SelfEmployed     1 2.1011e+06 5.2323e+10 53494
## - Hispanic         1 2.2760e+06 5.2323e+10 53494
## - Professional     1 2.6270e+06 5.2324e+10 53494
## - Drive            1 2.9825e+06 5.2324e+10 53494
## - Walk             1 3.0939e+06 5.2324e+10 53494
## - Carpool          1 3.5828e+06 5.2325e+10 53494
## - WorkAtHome       1 4.3709e+06 5.2325e+10 53494
## - FamilyWork       1 4.4875e+06 5.2326e+10 53494
## - PrivateWork      1 8.5938e+06 5.2330e+10 53495
## - PublicWork       1 9.1306e+06 5.2330e+10 53495
## - Black            1 2.0583e+07 5.2342e+10 53495
## - Employed         1 3.1169e+07 5.2352e+10 53496
## <none>                          5.2321e+10 53496
## - Native           1 4.6610e+07 5.2368e+10 53497
## - White            1 5.4779e+07 5.2376e+10 53498
## - Pacific          1 1.2380e+08 5.2445e+10 53502
## - TotalPop         1 3.3719e+08 5.2658e+10 53515
## - ChildPoverty     1 3.8681e+08 5.2708e+10 53518
## - Citizen          1 5.2462e+08 5.2846e+10 53526
## - Asian            1 5.5916e+08 5.2880e+10 53528
## - Men              1 6.0032e+08 5.2921e+10 53531
## - IncomeErr        1 1.1968e+09 5.3518e+10 53567
## - MeanCommute      1 3.8026e+09 5.6124e+10 53720
## - Poverty          1 3.9329e+09 5.6254e+10 53727
## - IncomePerCapErr  1 4.5101e+09 5.6831e+10 53760
## - IncomePerCap     1 3.5643e+10 8.7964e+10 55166
##
## Step:  AIC=53494.16
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Transit + Walk + OtherTransp +
##     WorkAtHome + MeanCommute + Employed + PrivateWork + PublicWork +
##     SelfEmployed + FamilyWork
##
```

```
##                   Df  Sum of Sq        RSS   AIC
## - Transit          1 1.4335e+06 5.2323e+10 53492
## - OtherTransp      1 1.5835e+06 5.2323e+10 53492
## - SelfEmployed     1 2.0763e+06 5.2323e+10 53492
## - Hispanic         1 2.2555e+06 5.2324e+10 53492
## - Drive            1 2.9977e+06 5.2324e+10 53492
## - Walk             1 3.1100e+06 5.2324e+10 53492
## - Carpool          1 3.6008e+06 5.2325e+10 53492
## - WorkAtHome       1 4.3882e+06 5.2326e+10 53492
## - FamilyWork       1 4.4590e+06 5.2326e+10 53492
## - PrivateWork      1 8.5465e+06 5.2330e+10 53493
## - PublicWork       1 9.0820e+06 5.2330e+10 53493
## - Black            1 2.0644e+07 5.2342e+10 53493
## - Employed         1 3.1151e+07 5.2352e+10 53494
## <none>                          5.2321e+10 53494
## - Native           1 4.6544e+07 5.2368e+10 53495
## - White            1 5.4881e+07 5.2376e+10 53496
## + Service          1 1.9808e+05 5.2321e+10 53496
## - Pacific          1 1.2381e+08 5.2445e+10 53500
## - Office           1 2.9460e+08 5.2616e+10 53510
## - TotalPop         1 3.3759e+08 5.2659e+10 53513
## - ChildPoverty     1 3.8676e+08 5.2708e+10 53516
## - Citizen          1 5.2499e+08 5.2846e+10 53524
## - Asian            1 5.5896e+08 5.2880e+10 53526
## - Men              1 6.0104e+08 5.2922e+10 53529
## - Production       1 8.8543e+08 5.3207e+10 53546
## - Construction     1 1.1263e+09 5.3447e+10 53561
## - IncomeErr        1 1.1971e+09 5.3518e+10 53565
## - Professional     1 2.3568e+09 5.4678e+10 53634
## - MeanCommute      1 3.8039e+09 5.6125e+10 53718
## - Poverty          1 3.9334e+09 5.6255e+10 53725
## - IncomePerCapErr  1 4.5102e+09 5.6831e+10 53758
## - IncomePerCap     1 3.5663e+10 8.7984e+10 55165
##
## Step:  AIC=53492.25
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + OtherTransp + WorkAtHome +
##     MeanCommute + Employed + PrivateWork + PublicWork + SelfEmployed +
##     FamilyWork
##
##                   Df  Sum of Sq        RSS   AIC
## - OtherTransp      1 1.0917e+06 5.2324e+10 53490
## - SelfEmployed     1 2.1242e+06 5.2325e+10 53490
## - Hispanic         1 2.2783e+06 5.2325e+10 53490
## - FamilyWork       1 4.5138e+06 5.2327e+10 53491
## - PrivateWork      1 8.6434e+06 5.2331e+10 53491
## - PublicWork       1 9.1840e+06 5.2332e+10 53491
## - Black            1 2.0603e+07 5.2343e+10 53492
## - Employed         1 3.0905e+07 5.2354e+10 53492
## <none>                          5.2323e+10 53492
## - Native           1 4.6584e+07 5.2369e+10 53493
## - White            1 5.4788e+07 5.2377e+10 53494
```

```
## + Transit          1 1.4335e+06 5.2321e+10 53494
## + Service          1 2.0800e+05 5.2322e+10 53494
## - Pacific          1 1.2415e+08 5.2447e+10 53498
## - Walk             1 1.7202e+08 5.2495e+10 53501
## - Drive            1 2.7629e+08 5.2599e+10 53507
## - Carpool          1 2.9082e+08 5.2614e+10 53508
## - Office           1 2.9562e+08 5.2618e+10 53508
## - TotalPop         1 3.3863e+08 5.2661e+10 53511
## - WorkAtHome       1 3.7853e+08 5.2701e+10 53513
## - ChildPoverty     1 3.8679e+08 5.2709e+10 53514
## - Citizen          1 5.2464e+08 5.2847e+10 53522
## - Asian            1 5.6018e+08 5.2883e+10 53525
## - Men              1 6.0217e+08 5.2925e+10 53527
## - Production       1 8.8666e+08 5.3209e+10 53544
## - Construction     1 1.1256e+09 5.3448e+10 53559
## - IncomeErr        1 1.1968e+09 5.3519e+10 53563
## - Professional     1 2.3562e+09 5.4679e+10 53632
## - MeanCommute      1 3.8029e+09 5.6126e+10 53716
## - Poverty          1 3.9346e+09 5.6257e+10 53724
## - IncomePerCapErr  1 4.5089e+09 5.6832e+10 53756
## - IncomePerCap     1 3.5664e+10 8.7987e+10 55163
##
## Step:  AIC=53490.32
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##     Employed + PrivateWork + PublicWork + SelfEmployed + FamilyWork
##
##                    Df  Sum of Sq        RSS   AIC
## - SelfEmployed      1 2.2034e+06 5.2326e+10 53488
## - Hispanic          1 2.2763e+06 5.2326e+10 53488
## - FamilyWork        1 4.6060e+06 5.2328e+10 53489
## - PrivateWork       1 8.8086e+06 5.2333e+10 53489
## - PublicWork        1 9.3541e+06 5.2333e+10 53489
## - Black             1 2.0543e+07 5.2344e+10 53490
## - Employed          1 3.1636e+07 5.2355e+10 53490
## <none>                           5.2324e+10 53490
## - Native            1 4.7236e+07 5.2371e+10 53491
## - White             1 5.4828e+07 5.2379e+10 53492
## + OtherTransp       1 1.0917e+06 5.2323e+10 53492
## + Transit           1 9.4177e+05 5.2323e+10 53492
## + Service           1 2.0842e+05 5.2324e+10 53492
## - Pacific           1 1.2329e+08 5.2447e+10 53496
## - Walk              1 1.7772e+08 5.2501e+10 53499
## - Office            1 3.0476e+08 5.2629e+10 53507
## - Carpool           1 3.3661e+08 5.2660e+10 53509
## - TotalPop          1 3.8389e+08 5.2708e+10 53512
## - Drive             1 3.8598e+08 5.2710e+10 53512
## - ChildPoverty      1 3.8676e+08 5.2711e+10 53512
## - WorkAtHome        1 4.3877e+08 5.2763e+10 53515
## - Citizen           1 5.2444e+08 5.2848e+10 53520
## - Asian             1 5.6132e+08 5.2885e+10 53523
## - Men               1 6.8535e+08 5.3009e+10 53530
```

```
## - Production       1 8.9190e+08 5.3216e+10 53543
## - Construction     1 1.1322e+09 5.3456e+10 53557
## - IncomeErr        1 1.1965e+09 5.3520e+10 53561
## - Professional     1 2.3631e+09 5.4687e+10 53630
## - MeanCommute      1 3.8397e+09 5.6163e+10 53716
## - Poverty          1 3.9461e+09 5.6270e+10 53722
## - IncomePerCapErr  1 4.5089e+09 5.6833e+10 53754
## - IncomePerCap     1 3.5812e+10 8.8136e+10 55166
##
## Step:  AIC=53488.46
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##     Employed + PrivateWork + PublicWork + FamilyWork
##
##                   Df  Sum of Sq        RSS   AIC
## - Hispanic         1 2.2615e+06 5.2328e+10 53487
## - Black            1 2.0616e+07 5.2347e+10 53488
## - FamilyWork       1 2.2413e+07 5.2348e+10 53488
## - Employed         1 3.1971e+07 5.2358e+10 53488
## <none>                          5.2326e+10 53488
## - Native           1 4.7146e+07 5.2373e+10 53489
## - White            1 5.4941e+07 5.2381e+10 53490
## + SelfEmployed     1 2.2034e+06 5.2324e+10 53490
## + OtherTransp      1 1.1709e+06 5.2325e+10 53490
## + Transit          1 1.0129e+06 5.2325e+10 53490
## + Service          1 1.8240e+05 5.2326e+10 53490
## - Pacific          1 1.2444e+08 5.2450e+10 53494
## - Walk             1 1.7731e+08 5.2503e+10 53497
## - Office           1 3.0640e+08 5.2632e+10 53505
## - Carpool          1 3.3684e+08 5.2663e+10 53507
## - TotalPop         1 3.8395e+08 5.2710e+10 53510
## - Drive            1 3.8518e+08 5.2711e+10 53510
## - ChildPoverty     1 3.8641e+08 5.2712e+10 53510
## - WorkAtHome       1 4.3756e+08 5.2764e+10 53513
## - Citizen          1 5.2344e+08 5.2849e+10 53518
## - Asian            1 5.6189e+08 5.2888e+10 53521
## - Men              1 6.8555e+08 5.3012e+10 53528
## - Production       1 8.9403e+08 5.3220e+10 53541
## - Construction     1 1.1337e+09 5.3460e+10 53555
## - IncomeErr        1 1.1949e+09 5.3521e+10 53559
## - Professional     1 2.3681e+09 5.4694e+10 53629
## - MeanCommute      1 3.8414e+09 5.6167e+10 53714
## - PublicWork       1 3.9045e+09 5.6230e+10 53718
## - Poverty          1 3.9543e+09 5.6280e+10 53721
## - PrivateWork      1 4.1319e+09 5.6458e+10 53731
## - IncomePerCapErr  1 4.5067e+09 5.6833e+10 53752
## - IncomePerCap     1 3.5834e+10 8.8160e+10 55165
##
## Step:  AIC=53486.59
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##     Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##     ChildPoverty + Professional + Office + Construction + Production +
```

```
##     Drive + Carpool + Walk + WorkAtHome + MeanCommute + Employed +
##     PrivateWork + PublicWork + FamilyWork
##
##                    Df  Sum of Sq        RSS   AIC
## - FamilyWork        1 2.2334e+07 5.2351e+10 53486
## - Employed          1 3.1472e+07 5.2360e+10 53487
## <none>                           5.2328e+10 53487
## + Hispanic          1 2.2615e+06 5.2326e+10 53488
## + SelfEmployed      1 2.1887e+06 5.2326e+10 53488
## + OtherTransp       1 1.1685e+06 5.2327e+10 53489
## + Transit           1 1.0095e+06 5.2327e+10 53489
## + Service           1 1.6287e+05 5.2328e+10 53489
## - Walk              1 1.7866e+08 5.2507e+10 53496
## - Pacific           1 1.8011e+08 5.2508e+10 53496
## - Office            1 3.0625e+08 5.2634e+10 53503
## - Carpool           1 3.3557e+08 5.2664e+10 53505
## - TotalPop          1 3.8267e+08 5.2711e+10 53508
## - Drive             1 3.8662e+08 5.2715e+10 53508
## - ChildPoverty      1 3.8951e+08 5.2718e+10 53508
## - WorkAtHome        1 4.4268e+08 5.2771e+10 53512
## - Citizen           1 5.2955e+08 5.2858e+10 53517
## - Native            1 5.6386e+08 5.2892e+10 53519
## - Men               1 6.8406e+08 5.3012e+10 53526
## - Production        1 9.0010e+08 5.3228e+10 53539
## - Construction      1 1.1369e+09 5.3465e+10 53554
## - IncomeErr         1 1.1992e+09 5.3527e+10 53558
## - Asian             1 1.3174e+09 5.3646e+10 53565
## - Black             1 2.1749e+09 5.4503e+10 53616
## - Professional      1 2.3720e+09 5.4700e+10 53627
## - MeanCommute       1 3.8514e+09 5.6180e+10 53713
## - PublicWork        1 3.9022e+09 5.6230e+10 53716
## - Poverty           1 3.9583e+09 5.6287e+10 53719
## - PrivateWork       1 4.1380e+09 5.6466e+10 53730
## - IncomePerCapErr   1 4.5049e+09 5.6833e+10 53750
## - White             1 7.4249e+09 5.9753e+10 53912
## - IncomePerCap      1 3.5842e+10 8.8170e+10 55164
##
## Step:  AIC=53485.97
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##     Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##     ChildPoverty + Professional + Office + Construction + Production +
##     Drive + Carpool + Walk + WorkAtHome + MeanCommute + Employed +
##     PrivateWork + PublicWork
##
##                    Df  Sum of Sq        RSS   AIC
## - Employed          1 3.0990e+07 5.2382e+10 53486
## <none>                           5.2351e+10 53486
## + FamilyWork        1 2.2334e+07 5.2328e+10 53487
## + SelfEmployed      1 1.9943e+07 5.2331e+10 53487
## + Hispanic          1 2.1822e+06 5.2348e+10 53488
## + OtherTransp       1 8.4910e+05 5.2350e+10 53488
## + Transit           1 7.2473e+05 5.2350e+10 53488
## + Service           1 2.8674e+05 5.2350e+10 53488
## - Pacific           1 1.7971e+08 5.2530e+10 53495
```

```
## - Walk             1 1.8257e+08 5.2533e+10 53495
## - Office           1 3.0845e+08 5.2659e+10 53503
## - Carpool          1 3.3318e+08 5.2684e+10 53504
## - TotalPop         1 3.8241e+08 5.2733e+10 53507
## - Drive            1 3.8771e+08 5.2738e+10 53508
## - ChildPoverty     1 3.9695e+08 5.2748e+10 53508
## - WorkAtHome       1 4.5945e+08 5.2810e+10 53512
## - Citizen          1 5.2627e+08 5.2877e+10 53516
## - Native           1 5.6844e+08 5.2919e+10 53519
## - Men              1 6.8260e+08 5.3033e+10 53526
## - Production       1 9.0088e+08 5.3251e+10 53539
## - Construction     1 1.1621e+09 5.3513e+10 53555
## - IncomeErr        1 1.2108e+09 5.3561e+10 53558
## - Asian            1 1.3207e+09 5.3671e+10 53564
## - Black            1 2.1566e+09 5.4507e+10 53614
## - Professional     1 2.3701e+09 5.4721e+10 53626
## - MeanCommute      1 3.8370e+09 5.6188e+10 53712
## - Poverty          1 3.9444e+09 5.6295e+10 53718
## - PublicWork       1 3.9764e+09 5.6327e+10 53720
## - PrivateWork      1 4.2727e+09 5.6623e+10 53736
## - IncomePerCapErr  1 4.5545e+09 5.6905e+10 53752
## - White            1 7.4025e+09 5.9753e+10 53910
## - IncomePerCap     1 3.5820e+10 8.8171e+10 55162
##
## Step:  AIC=53485.87
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##     Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##     ChildPoverty + Professional + Office + Construction + Production +
##     Drive + Carpool + Walk + WorkAtHome + MeanCommute + PrivateWork +
##     PublicWork
##
##                   Df  Sum of Sq        RSS   AIC
## <none>                          5.2382e+10 53486
## + Employed         1 3.0990e+07 5.2351e+10 53486
## + FamilyWork       1 2.1852e+07 5.2360e+10 53487
## + SelfEmployed     1 1.9355e+07 5.2362e+10 53487
## + Hispanic         1 1.6973e+06 5.2380e+10 53488
## + OtherTransp      1 1.5088e+06 5.2380e+10 53488
## + Transit          1 1.3551e+06 5.2380e+10 53488
## + Service          1 2.6332e+05 5.2381e+10 53488
## - Pacific          1 1.6759e+08 5.2549e+10 53494
## - Walk             1 1.9161e+08 5.2573e+10 53496
## - Office           1 3.1372e+08 5.2695e+10 53503
## - Carpool          1 3.4274e+08 5.2724e+10 53505
## - ChildPoverty     1 3.8845e+08 5.2770e+10 53508
## - Drive            1 4.0174e+08 5.2783e+10 53508
## - TotalPop         1 4.1628e+08 5.2798e+10 53509
## - WorkAtHome       1 4.7718e+08 5.2859e+10 53513
## - Native           1 5.6097e+08 5.2943e+10 53518
## - Citizen          1 5.6475e+08 5.2946e+10 53518
## - Men              1 6.6787e+08 5.3049e+10 53525
## - Production       1 8.7883e+08 5.3260e+10 53537
## - Construction     1 1.1431e+09 5.3525e+10 53553
## - IncomeErr        1 1.1974e+09 5.3579e+10 53557
```

```
## - Asian            1 1.2899e+09 5.3671e+10 53562
## - Black            1 2.1790e+09 5.4561e+10 53615
## - Professional     1 2.3395e+09 5.4721e+10 53624
## - MeanCommute      1 3.8441e+09 5.6226e+10 53712
## - PublicWork       1 3.9779e+09 5.6359e+10 53719
## - Poverty          1 4.0267e+09 5.6408e+10 53722
## - PrivateWork      1 4.2693e+09 5.6651e+10 53736
## - IncomePerCapErr  1 4.5413e+09 5.6923e+10 53751
## - White            1 7.5118e+09 5.9893e+10 53915
## - IncomePerCap     1 3.6404e+10 8.8786e+10 55182
```

```
summary(stepwise_model_both)
```

```
##
## Call:
## lm(formula = Y ~ TotalPop + Men + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##     PrivateWork + PublicWork, data = X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20911.8  -2378.7   -120.8   2101.6  22020.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4.567e+04  4.513e+03 -10.118  < 2e-16 ***
## TotalPop        -9.203e-02  1.827e-02  -5.037 4.98e-07 ***
## Men              2.214e-01  3.470e-02   6.381 2.02e-10 ***
## White           -1.121e+02  5.237e+00 -21.398  < 2e-16 ***
## Black           -7.532e+01  6.536e+00 -11.525  < 2e-16 ***
## Native           7.301e+01  1.249e+01   5.848 5.49e-09 ***
## Asian            3.879e+02  4.374e+01   8.867  < 2e-16 ***
## Pacific         -6.946e+02  2.173e+02  -3.196  0.00141 **
## Citizen         -2.879e-02  4.907e-03  -5.867 4.88e-09 ***
## IncomeErr        4.851e-01  5.679e-02   8.543  < 2e-16 ***
## IncomePerCap     1.300e+00  2.759e-02  47.107  < 2e-16 ***
## IncomePerCapErr -1.767e+00  1.062e-01 -16.638  < 2e-16 ***
## Poverty         -4.786e+02  3.055e+01 -15.667  < 2e-16 ***
## ChildPoverty    -9.329e+01  1.917e+01  -4.866 1.19e-06 ***
## Professional     3.235e+02  2.709e+01  11.942  < 2e-16 ***
## Office           1.508e+02  3.448e+01   4.373 1.26e-05 ***
## Construction     2.517e+02  3.016e+01   8.347  < 2e-16 ***
## Production       1.861e+02  2.542e+01   7.319 3.14e-13 ***
## Drive            1.344e+02  2.717e+01   4.949 7.86e-07 ***
## Carpool          1.764e+02  3.860e+01   4.571 5.04e-06 ***
## Walk             1.474e+02  4.312e+01   3.418  0.00064 ***
## WorkAtHome       2.333e+02  4.326e+01   5.393 7.42e-08 ***
## MeanCommute      2.325e+02  1.519e+01  15.308  < 2e-16 ***
## PrivateWork      4.546e+02  2.818e+01  16.132  < 2e-16 ***
## PublicWork       4.831e+02  3.102e+01  15.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4050 on 3193 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.9015
## F-statistic:  1228 on 24 and 3193 DF,  p-value: < 2.2e-16
```

```
selected_features_step <- names(coef(stepwise_model_both))[-1]
selected_features_step
```

```
##  [1] "TotalPop"        "Men"             "White"           "Black"
##  [5] "Native"          "Asian"           "Pacific"         "Citizen"
##  [9] "IncomeErr"       "IncomePerCap"    "IncomePerCapErr" "Poverty"
## [13] "ChildPoverty"    "Professional"    "Office"          "Construction"
## [17] "Production"      "Drive"           "Carpool"         "Walk"
## [21] "WorkAtHome"      "MeanCommute"     "PrivateWork"     "PublicWork"
```

Filtering out these features

```
X=X[selected_features_step]
```

Base Model with main effect before diagnosis

```
fit1 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit1)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared
```

Base Model diagnosis with main effect

```
cat("R2:", r2, "\n")
```

```
## R2: 0.9022807
```

```
cat("Adjusted R2:", adj_r2, "\n")
```

```
## Adjusted R2: 0.9015461
```

```
cat("AIC:", AIC(fit1), "\n")
```

```
## AIC: 62620.16
```

```
cat("BIC:", BIC(fit1), "\n")
```

```
## BIC: 62778.15
```

```r
print(dwtest(fit1))
```

```
##
##  Durbin-Watson test
##
## data:  fit1
## DW = 1.7407, p-value = 4.733e-14
## alternative hypothesis: true autocorrelation is greater than 0
```

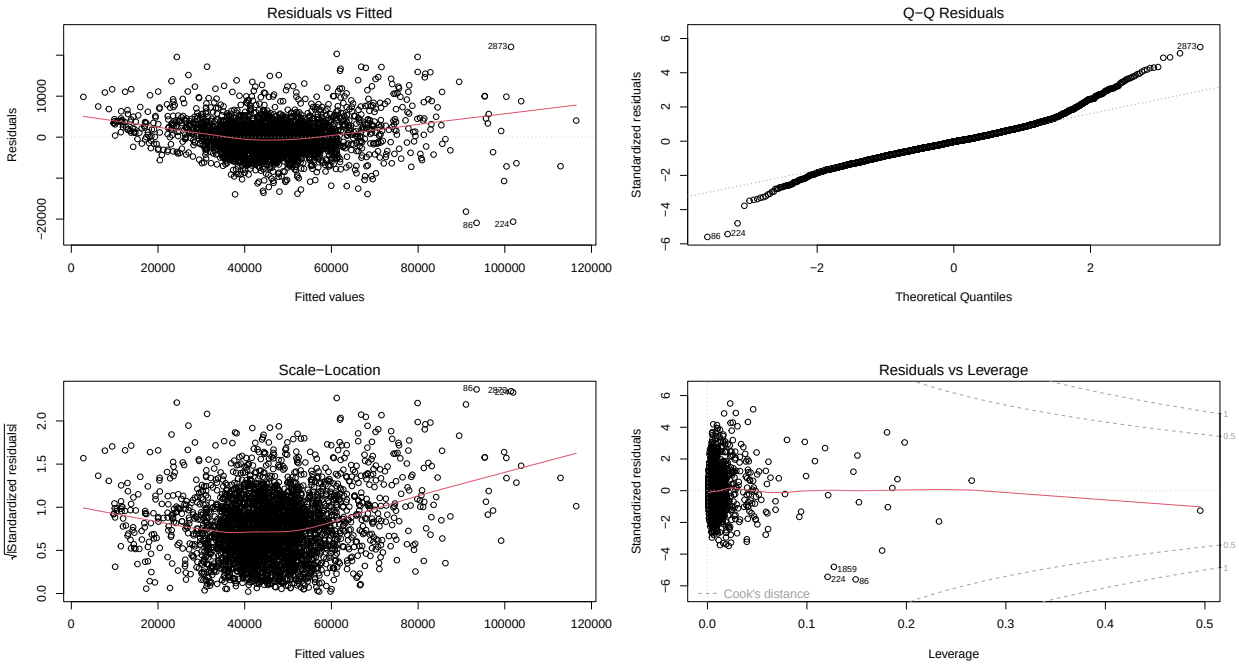
```r
print(shapiro.test(residuals(fit1)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit1)
## W = 0.96419, p-value < 2.2e-16
```

```r
print(vif(fit1))
```

```
##        TotalPop             Men           White           Black          Native
##     6677.425309     5799.286884        2.825381        1.708575        1.607888
##           Asian         Pacific         Citizen       IncomeErr    IncomePerCap
##        2.556547        1.433631      198.749750        2.319333        5.726323
## IncomePerCapErr         Poverty    ChildPoverty    Professional          Office
##        2.386229       12.661735        9.854505        5.837453        2.380646
##    Construction      Production           Drive         Carpool            Walk
##        3.166153        4.171999        8.392551        2.471069        4.989173
##      WorkAtHome     MeanCommute     PrivateWork      PublicWork
##        3.695681        1.416821        9.517031        7.870448
```

```r
par(mfrow = c(2, 2))
plot(fit1)
```



As we can see we got good R2 but assumptions failed.

Treatment in main effect

Dropping features with VIF<5

```
vif_values <- vif(fit1)
low_vif_features <- names(vif_values)[vif_values < 5]
print(low_vif_features)
```

```
##  [1] "White"           "Black"           "Native"          "Asian"
##  [5] "Pacific"         "IncomeErr"       "IncomePerCapErr" "Office"
##  [9] "Construction"    "Production"      "Carpool"         "Walk"
## [13] "WorkAtHome"      "MeanCommute"
```

```
X <- X[low_vif_features]
dim(X)
```

```
## [1] 3218   14
```

Tranformation

```
Y <- bestNormalize::yeojohnson(Y)$x.t
X[abs(apply(X, 2, e1071::skewness)) > 1] <- lapply(X[abs(apply(X, 2, e1071::skewness)) > 1], log1p)  # A
head(X)
```

```
##      White     Black    Native      Asian Pacific IncomeErr IncomePerCapErr
## 1 4.341205 2.9704145 0.3364722 0.69314718       0  7.779885        6.985642
## 2 4.432007 2.3513753 0.4700036 0.53062825       0  7.142037        6.568078
## 3 3.854394 3.8649314 0.1823216 0.33647224       0  7.997663        6.683361
## 4 4.324133 3.1090610 0.3364722 0.09531018       0  8.293049        7.389564
## 5 4.487512 0.9162907 0.2623643 0.09531018       0  8.052615        6.563856
## 6 3.144152 4.2724907 0.7884574 0.18232156       0  8.680162        7.628518
##   Office Construction Production Carpool      Walk WorkAtHome MeanCommute
## 1   24.2          8.6       17.1     8.8 0.4054651  1.0296194        26.5
## 2   27.1         10.8       11.2     8.8 0.6931472  1.5892352        26.4
## 3   23.1         10.8       23.1    10.9 1.0296194  0.9555114        24.1
## 4   17.8         19.0       23.7    13.5 0.4700036  0.5306283        28.8
## 5   23.9         13.5       19.9    11.2 0.6418539  1.1939225        34.9
## 6   19.7         20.1       26.4    14.9 1.7917595  1.3350011        27.5
```

Base model after diagnosis(VIF filter+transformation) with main effect

```r
fit2 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit2)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared
```

```r
cat("R2:", r2, "\n")
```

```
## R2: 0.5565732
```

```r
cat("Adjusted R2:", adj_r2, "\n")
```

```
## Adjusted R2: 0.554635
```

```r
cat("AIC:", AIC(fit2), "\n")
```

```
## AIC: 6546.338
```

```r
cat("BIC:", BIC(fit2), "\n")
```

```
## BIC: 6643.563
```

```r
print(dwtest(fit2))
```

```
##
##  Durbin-Watson test
##
## data:  fit2
## DW = 1.5695, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

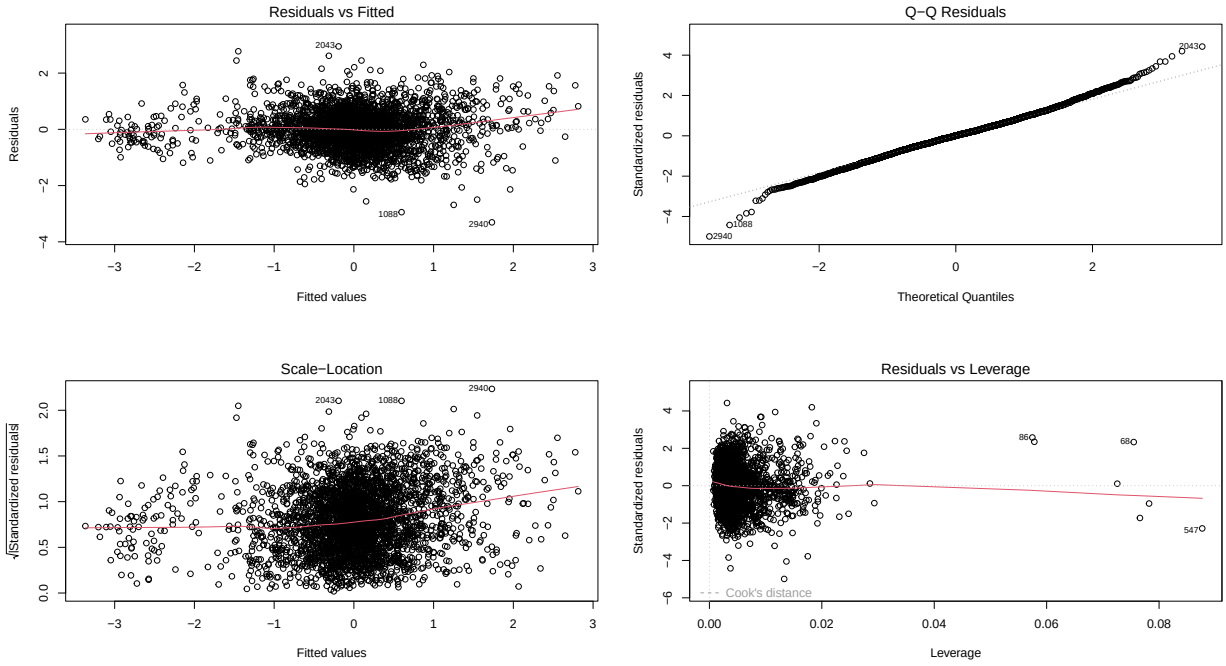
```
print(shapiro.test(residuals(fit2)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit2)
## W = 0.99474, p-value = 2.442e-09
```

```
print(vif(fit2))
```

```
##          White          Black         Native          Asian        Pacific
##       1.263248       1.429725       1.205693       1.942369       1.190704
##       IncomeErr IncomePerCapErr         Office   Construction     Production
##       3.537887       3.732609       1.657800       1.889478       1.648585
##        Carpool           Walk     WorkAtHome     MeanCommute
##       1.156902       1.781497       1.778833       1.423280
```

```
par(mfrow = c(2, 2))
plot(fit2)
```



Dropping outliers for cooks distance

```
dim(X)
```

```
## [1] 3218   14
```

```r
cooks_dist <- cooks.distance(fit2)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
X <- X[-influential_points, ]
Y <- Y[-influential_points]
```

Base model after diagnosis(Influential points) with main effect

```r
fit3 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit3)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared
```

```r
cat("R2:", r2, "\n")
```

```
## R2: 0.6333641
```

```r
cat("Adjusted R2:", adj_r2, "\n")
```

```
## Adjusted R2: 0.6316416
```

```r
cat("AIC:", AIC(fit3), "\n")
```

```
## AIC: 4965.617
```

```r
cat("BIC:", BIC(fit3), "\n")
```

```
## BIC: 5061.692
```

```r
print(dwtest(fit3))
```

```
##
##  Durbin-Watson test
##
## data:  fit3
## DW = 1.6749, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

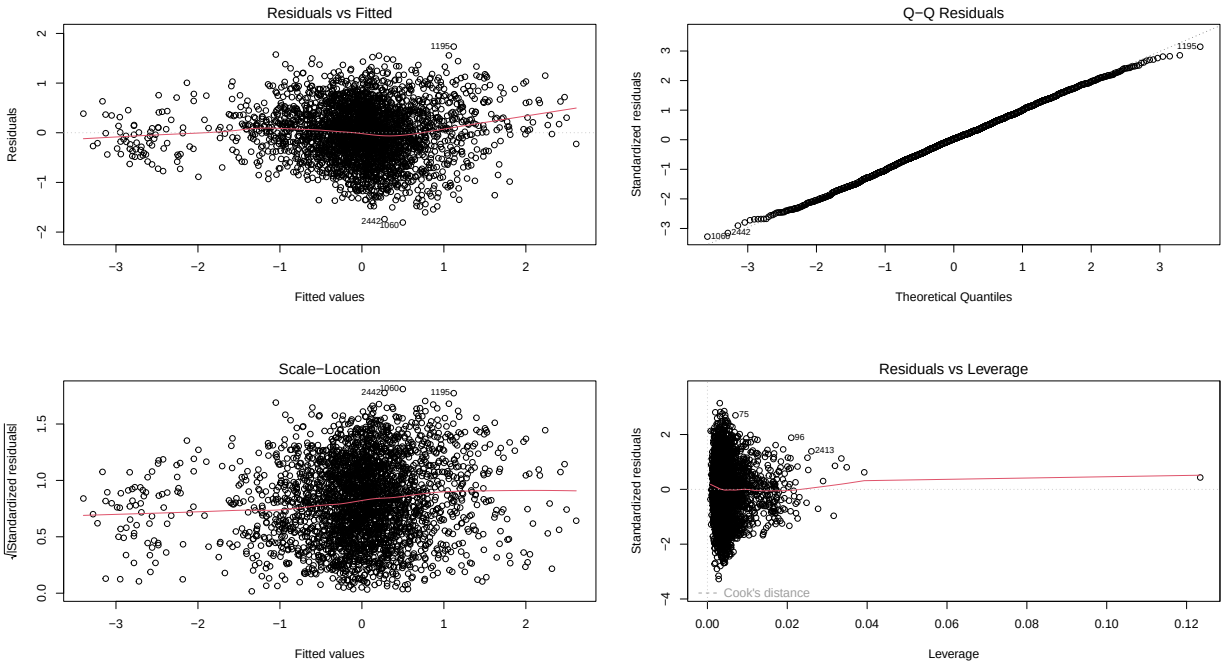
```r
print(shapiro.test(residuals(fit3)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit3)
## W = 0.9989, p-value = 0.05253
```

```
print(vif(fit3))
```

```
##           White           Black          Native           Asian         Pacific
##        1.298041        1.460077        1.186241        2.010300        1.127789
##       IncomeErr IncomePerCapErr          Office    Construction      Production
##        3.482675        3.742256        1.725720        1.901318        1.793523
##         Carpool            Walk      WorkAtHome      MeanCommute
##        1.154905        1.820614        1.875744        1.391618
```

```
par(mfrow = c(2, 2))
plot(fit3)
```



Only multicollineariy has been passed, but plot has been improved.

two main effect

All combination of two main effect has been implemented and the best features extracted using STEP.

```
full_model <- lm(Y ~ (.)^2, data = X)
step_model <- step(full_model, direction = "both", trace = FALSE)
final_formula <- formula(step_model)
final_fit <- lm(final_formula, data = X)
cooks_dist <- cooks.distance(final_fit)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
```

```r
X <- X[-influential_points, , drop = FALSE]
Y <- Y[-influential_points]
final_fit <- lm(final_formula, data = X)
model_summary <- summary(final_fit)
r2 <- model_summary$r.squared
adj_r2 <- model_summary$adj.r.squared

cat("R2:", r2, "\n")
```

```
## R2: 0.7121303
```

```r
cat("Adjusted R2:", adj_r2, "\n")
```

```
## Adjusted R2: 0.705694
```

```r
cat("AIC:", AIC(final_fit), "\n")
```

```
## AIC: 3820.55
```

```r
cat("BIC:", BIC(final_fit), "\n")
```

```
## BIC: 4201.359
```

```r
print(final_formula)
```

```
## Y ~ White + Black + Native + Asian + Pacific + IncomeErr + IncomePerCapErr +
##     Office + Construction + Production + Carpool + Walk + WorkAtHome +
##     MeanCommute + White:Black + White:Native + White:Asian +
##     White:IncomePerCapErr + White:Office + White:Construction +
##     White:Production + White:Carpool + White:Walk + White:WorkAtHome +
##     Black:Asian + Black:Pacific + Black:IncomeErr + Black:IncomePerCapErr +
##     Black:Production + Black:WorkAtHome + Black:MeanCommute +
##     Native:IncomePerCapErr + Native:Construction + Native:WorkAtHome +
##     Native:MeanCommute + Asian:IncomeErr + Asian:Office + Asian:Production +
##     Asian:Carpool + Asian:Walk + Asian:WorkAtHome + Asian:MeanCommute +
##     Pacific:Production + Pacific:Carpool + Pacific:MeanCommute +
##     IncomeErr:IncomePerCapErr + IncomeErr:Office + IncomeErr:Walk +
##     IncomePerCapErr:Construction + IncomePerCapErr:Walk + Construction:Production +
##     Construction:Carpool + Construction:Walk + Construction:WorkAtHome +
##     Construction:MeanCommute + Production:Carpool + Production:Walk +
##     Production:WorkAtHome + Production:MeanCommute + Carpool:Walk +
##     Walk:MeanCommute + WorkAtHome:MeanCommute
```

```r
print(dwtest(final_fit))
```

```
##
##  Durbin-Watson test
##
## data:  final_fit
## DW = 1.7113, p-value = 2.99e-15
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
print(shapiro.test(residuals(final_fit)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(final_fit)
## W = 0.99932, p-value = 0.3903
```

```r
par(mfrow = c(2, 2))
plot(final_fit)
```