Exploratory analysis

```
library(ggplot2)
library(gridExtra)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-6
```

```
# getwd()
#setwd('DATA 583/583-project')

# Load the data from the CSV file
data <- read.csv("data_whole.csv", header = TRUE, check.names = FALSE)
dim(data) # 2857, 33
```

```
## [1] 2857   33
```

```
## data cleaning
# remove y2, x6, x8, x9, x22, x29 because of missing values for some rows
data <- data[, -c(3, 9, 11, 12, 25, 32)]

# remove rows with missing values, obtain only weekly data, rather than daily
data <- na.omit(data)
dim(data) # 352, 27
```

```
## [1] 352  27
```

```
# renme y1 to y
names(data)[2] <- "y"

#remove date
data_no_date <- data[, -1]

head(data)
```

```
##           date    y    x1    x2  x3         x4     x5         x7      x10
## 1079 2016/1/15 4380 29.42 28.94 370  -47.12009 -14.08 -1313.2016 312.1795
## 1084 2016/1/22 4500 32.19 32.18 370  -83.91724 -30.60 -1019.3778 312.1795
## 1089 2016/1/29 4715 33.62 34.74 375 -226.87681 -43.40  -637.3231 397.6496
## 1094  2016/2/5 4710 30.89 34.06 380 -214.94086 -46.42  -470.3440 397.6496
## 1099 2016/2/12 4710 29.44 33.36 380 -180.44046 -42.46  -466.3399 372.6496
## 1104 2016/2/19 5075 29.64 33.01 380  -72.41137 -34.74  -261.4895 408.1197
##         x11    x12    x13    x14 x15    x16  x17  x18 x19 x20 x21  x23  x24
```

```
## 1079 -173.8   26.2  186.2  211.2  90  111.2 70.7 66.4  18  57  49 70.3 15.3
## 1084 -274.8  -34.8   60.2   95.2  75   10.2 68.4 65.9  18  48  38 64.6 15.6
## 1089 -256.5   -6.5   -1.5   13.5  55   33.5 65.7 60.3  15   9   7 60.0 14.1
## 1094 -246.2    3.8    8.8   23.8  55   43.8 65.7 60.3  15   9   7 60.0 14.1
## 1099 -246.2    3.8    8.8   23.8  55   43.8 62.5 60.3  15   9   7 60.0 14.1
## 1104 -377.8 -127.8 -142.8 -132.8  95 -127.8 63.5 59.8  15  23  37 53.9 18.4
##       x25  x26  x27 x28     x30
## 1079 21.0  7.2 60.1 7.0 8.72712
## 1084 21.7  7.8 60.6 7.7 8.90368
## 1089 21.8  7.8 51.6 7.6 9.24000
## 1094 21.8  7.8 51.6 6.9 8.91000
## 1099 21.8  7.8 51.6 7.3 7.68750
## 1104 30.0 16.5 79.7 7.4 6.90140
```

As the raw data has missing values, we need to remove the rows with missing values. After removing the rows with missing values, we have 352 observations and 27 variables. The response variable is y, which is . The explanatory variables are x1 to x27, which are the .

The data structure is shown below.

```
# The structure of the data,
# including the number of variables, their types, and the first few observations.
str(data)
```

```
## 'data.frame':    352 obs. of  27 variables:
##  $ date: chr  "2016/1/15" "2016/1/22" "2016/1/29" "2016/2/5" ...
##  $ y   : num  4380 4500 4715 4710 4710 ...
##  $ x1  : num  29.4 32.2 33.6 30.9 29.4 ...
##  $ x2  : num  28.9 32.2 34.7 34.1 33.4 ...
##  $ x3  : num  370 370 375 380 380 380 380 385 390 390 ...
##  $ x4  : num  -47.1 -83.9 -226.9 -214.9 -180.4 ...
##  $ x5  : num  -14.1 -30.6 -43.4 -46.4 -42.5 ...
##  $ x7  : num  -1313 -1019 -637 -470 -466 ...
##  $ x10 : num  312 312 398 398 373 ...
##  $ x11 : num  -174 -275 -256 -246 -246 ...
##  $ x12 : num  26.2 -34.8 -6.5 3.8 3.8 ...
##  $ x13 : num  186.2 60.2 -1.5 8.8 8.8 ...
##  $ x14 : num  211.2 95.2 13.5 23.8 23.8 ...
##  $ x15 : int  90 75 55 55 55 95 80 255 135 160 ...
##  $ x16 : num  111.2 10.2 33.5 43.8 43.8 ...
##  $ x17 : num  70.7 68.4 65.7 65.7 62.5 63.5 69.9 74.3 78 80.7 ...
##  $ x18 : num  66.4 65.9 60.3 60.3 60.3 59.8 67.6 74 78.9 80.7 ...
##  $ x19 : num  18 18 15 15 15 15 20 30 36 39 ...
##  $ x20 : num  57 48 9 9 9 23 61 76 82 82 ...
##  $ x21 : num  49 38 7 7 7 37 63 74 79 79 ...
##  $ x23 : num  70.3 64.6 60 60 60 53.9 60.8 65.4 69.2 69 ...
##  $ x24 : num  15.3 15.6 14.1 14.1 14.1 18.4 18.1 17 13.2 13.7 ...
##  $ x25 : num  21 21.7 21.8 21.8 21.8 30 29.6 28.6 25.2 25 ...
##  $ x26 : num  7.2 7.8 7.8 7.8 7.8 16.5 17.5 15.5 10.1 9.9 ...
##  $ x27 : num  60.1 60.6 51.6 51.6 51.6 79.7 79.8 69.7 53.5 54.5 ...
##  $ x28 : num  7 7.7 7.6 6.9 7.3 7.4 9 8.3 7.7 11.1 ...
##  $ x30 : num  8.73 8.9 9.24 8.91 7.69 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2505] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:2505] "1" "2" "3" "4" ...
```

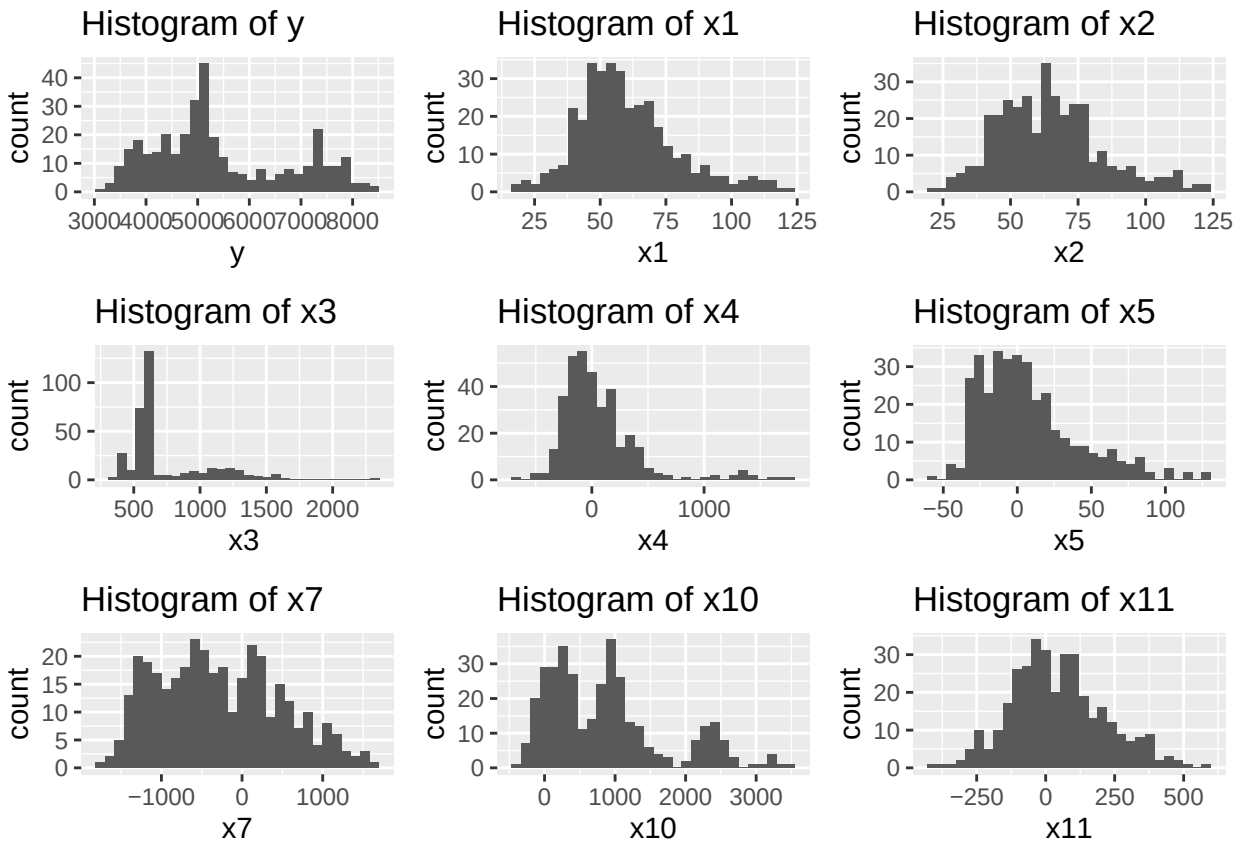
Here is the summary statistics of the data.

```
summary(data)
```

```
##      date                y              x1              x2        
##  Length:352         Min.   :3130   Min.   : 16.94   Min.   : 21.44  
##  Class :character   1st Qu.:4428   1st Qu.: 47.73   1st Qu.: 49.97  
##  Mode  :character   Median :5118   Median : 56.70   Median : 63.34  
##                     Mean   :5427   Mean   : 60.18   Mean   : 64.51  
##                     3rd Qu.:6511   3rd Qu.: 69.64   3rd Qu.: 75.33  
##                     Max.   :8440   Max.   :121.37   Max.   :122.95  
##        x3               x4               x5               x7        
##  Min.   : 370.0   Min.   :-688.89   Min.   :-56.480   Min.   :-1742.1  
##  1st Qu.: 575.0   1st Qu.:-151.02   1st Qu.:-17.983   1st Qu.: -908.7  
##  Median : 604.8   Median : -17.50   Median :  0.600   Median : -348.5  
##  Mean   : 747.8   Mean   :  58.92   Mean   :  7.387   Mean   : -287.3  
##  3rd Qu.: 900.0   3rd Qu.: 174.37   3rd Qu.: 22.125   3rd Qu.:  244.5  
##  Max.   :2350.0   Max.   :1754.92   Max.   :128.740   Max.   : 1658.5  
##       x10              x11               x12               x13        
##  Min.   :-372.6   Min.   :-420.80   Min.   :-585.23   Min.   :-745.23  
##  1st Qu.: 204.4   1st Qu.: -73.16   1st Qu.:  30.88   1st Qu.: -29.79  
##  Median : 767.6   Median :  26.89   Median : 170.20   Median : 168.22  
##  Mean   : 861.3   Mean   :  43.85   Mean   : 315.46   Mean   : 201.25  
##  3rd Qu.:1166.6   3rd Qu.: 145.03   3rd Qu.: 472.56   3rd Qu.: 430.65  
##  Max.   :3516.2   Max.   : 573.15   Max.   :2370.97   Max.   :1188.15  
##       x14              x15               x16              x17        
##  Min.   :-715.2   Min.   :-180.0   Min.   :-234.5   Min.   :59.50  
##  1st Qu.:-114.4   1st Qu.: 145.0   1st Qu.: 107.7   1st Qu.:82.90  
##  Median : 138.7   Median : 270.0   Median : 270.2   Median :87.60  
##  Mean   : 163.6   Mean   : 282.9   Mean   : 316.9   Mean   :86.13  
##  3rd Qu.: 418.9   3rd Qu.: 410.0   3rd Qu.: 507.5   3rd Qu.:91.20  
##  Max.   :1246.7   Max.   : 920.0   Max.   :1313.1   Max.   :97.70  
##       x18              x19              x20              x21        
##  Min.   :47.50   Min.   :13.30   Min.   : 4.00   Min.   : 1.00  
##  1st Qu.:78.50   1st Qu.:41.00   1st Qu.:68.00   1st Qu.:59.00  
##  Median :85.00   Median :47.00   Median :78.00   Median :70.00  
##  Mean   :82.72   Mean   :47.27   Mean   :72.82   Mean   :65.35  
##  3rd Qu.:89.12   3rd Qu.:54.92   3rd Qu.:85.00   3rd Qu.:78.00  
##  Max.   :97.10   Max.   :84.00   Max.   :95.00   Max.   :93.00  
##       x23              x24              x25              x26        
##  Min.   : 42.50   Min.   : 2.00   Min.   : 7.30   Min.   : 0.80  
##  1st Qu.: 78.90   1st Qu.:11.65   1st Qu.:16.68   1st Qu.: 7.50  
##  Median : 83.60   Median :16.10   Median :21.70   Median :11.80  
##  Mean   : 83.16   Mean   :16.79   Mean   :21.99   Mean   :13.07  
##  3rd Qu.: 89.05   3rd Qu.:22.23   3rd Qu.:26.90   3rd Qu.:16.73  
##  Max.   :100.40   Max.   :31.10   Max.   :38.60   Max.   :33.40  
##       x27              x28              x30        
##  Min.   : 10.70   Min.   : 2.500   Min.   : 6.77  
##  1st Qu.: 40.75   1st Qu.: 6.100   1st Qu.:10.39  
##  Median : 54.75   Median : 7.700   Median :11.92  
##  Mean   : 58.56   Mean   : 8.263   Mean   :11.97  
##  3rd Qu.: 70.92   3rd Qu.:10.000   3rd Qu.:13.42  
##  Max.   :114.00   Max.   :19.000   Max.   :17.38
```

Let's explore the histogram of the response variable y and the explanatory variables.

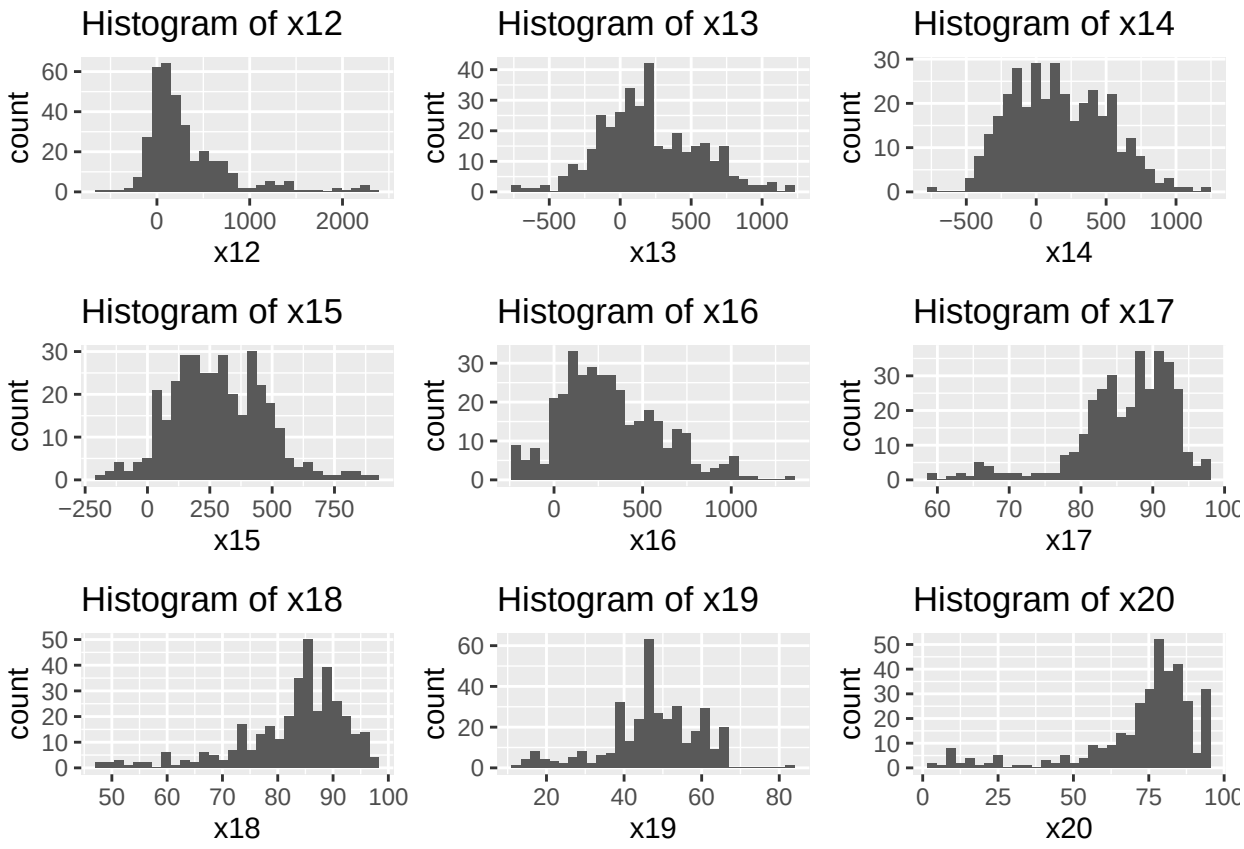
```
# Create a histograms of the variables
hist_y <- ggplot(data, aes(x = y)) + geom_histogram() + labs(title = "Histogram of y")
hist_x1 <- ggplot(data, aes(x = x1)) + geom_histogram() + labs(title = "Histogram of x1")
hist_x2 <- ggplot(data, aes(x = x2)) + geom_histogram() + labs(title = "Histogram of x2")
hist_x3 <- ggplot(data, aes(x = x3)) + geom_histogram() + labs(title = "Histogram of x3")
hist_x4 <- ggplot(data, aes(x = x4)) + geom_histogram() + labs(title = "Histogram of x4")
hist_x5 <- ggplot(data, aes(x = x5)) + geom_histogram() + labs(title = "Histogram of x5")
hist_x7 <- ggplot(data, aes(x = x7)) + geom_histogram() + labs(title = "Histogram of x7")
hist_x10 <- ggplot(data, aes(x = x10)) + geom_histogram() + labs(title = "Histogram of x10")
hist_x11 <- ggplot(data, aes(x = x11)) + geom_histogram() + labs(title = "Histogram of x11")

grid.arrange(hist_y, hist_x1, hist_x2, hist_x3, hist_x4,
             hist_x5, hist_x7, hist_x10, hist_x11, ncol = 3)
```




```
# Create a histogram of the response variable x12, x13, x14, x15, x16, x17, x18, x19, x20
hist_x12 <- ggplot(data, aes(x = x12)) + geom_histogram() + labs(title = "Histogram of x12")
hist_x13 <- ggplot(data, aes(x = x13)) + geom_histogram() + labs(title = "Histogram of x13")
hist_x14 <- ggplot(data, aes(x = x14)) + geom_histogram() + labs(title = "Histogram of x14")
hist_x15 <- ggplot(data, aes(x = x15)) + geom_histogram() + labs(title = "Histogram of x15")
hist_x16 <- ggplot(data, aes(x = x16)) + geom_histogram() + labs(title = "Histogram of x16")
hist_x17 <- ggplot(data, aes(x = x17)) + geom_histogram() + labs(title = "Histogram of x17")
hist_x18 <- ggplot(data, aes(x = x18)) + geom_histogram() + labs(title = "Histogram of x18")
hist_x19 <- ggplot(data, aes(x = x19)) + geom_histogram() + labs(title = "Histogram of x19")
hist_x20 <- ggplot(data, aes(x = x20)) + geom_histogram() + labs(title = "Histogram of x20")
```

```
grid.arrange(hist_x12, hist_x13, hist_x14, hist_x15, hist_x16,
             hist_x17, hist_x18, hist_x19, hist_x20,
             nrow = 3, ncol = 3)
```

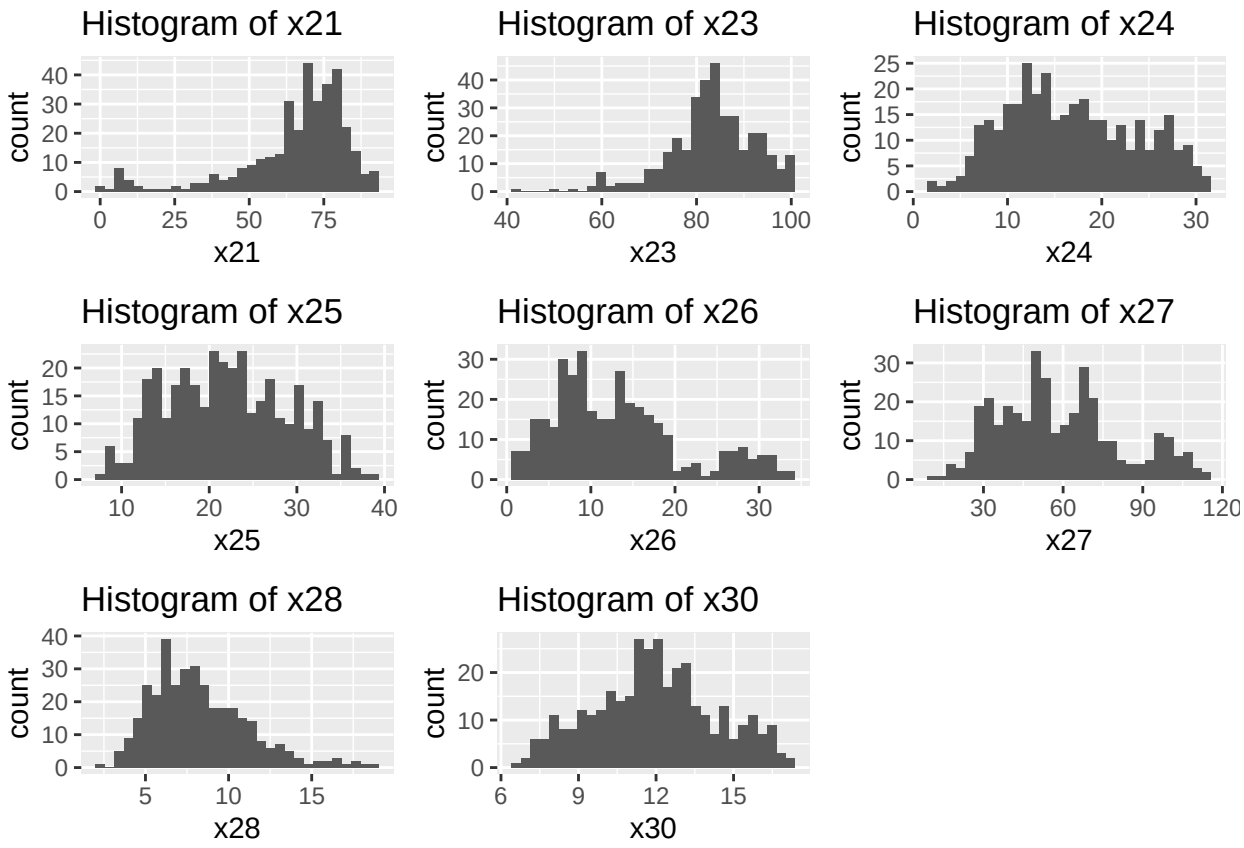
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Create a histogram of the response variable x21, x23, x24, x25, x26, x27, x28, x30
hist_x21 <- ggplot(data, aes(x = x21)) + geom_histogram() + labs(title = "Histogram of x21")
hist_x23 <- ggplot(data, aes(x = x23)) + geom_histogram() + labs(title = "Histogram of x23")
hist_x24 <- ggplot(data, aes(x = x24)) + geom_histogram() + labs(title = "Histogram of x24")
hist_x25 <- ggplot(data, aes(x = x25)) + geom_histogram() + labs(title = "Histogram of x25")
hist_x26 <- ggplot(data, aes(x = x26)) + geom_histogram() + labs(title = "Histogram of x26")
hist_x27 <- ggplot(data, aes(x = x27)) + geom_histogram() + labs(title = "Histogram of x27")
hist_x28 <- ggplot(data, aes(x = x28)) + geom_histogram() + labs(title = "Histogram of x28")
hist_x30 <- ggplot(data, aes(x = x30)) + geom_histogram() + labs(title = "Histogram of x30")
```

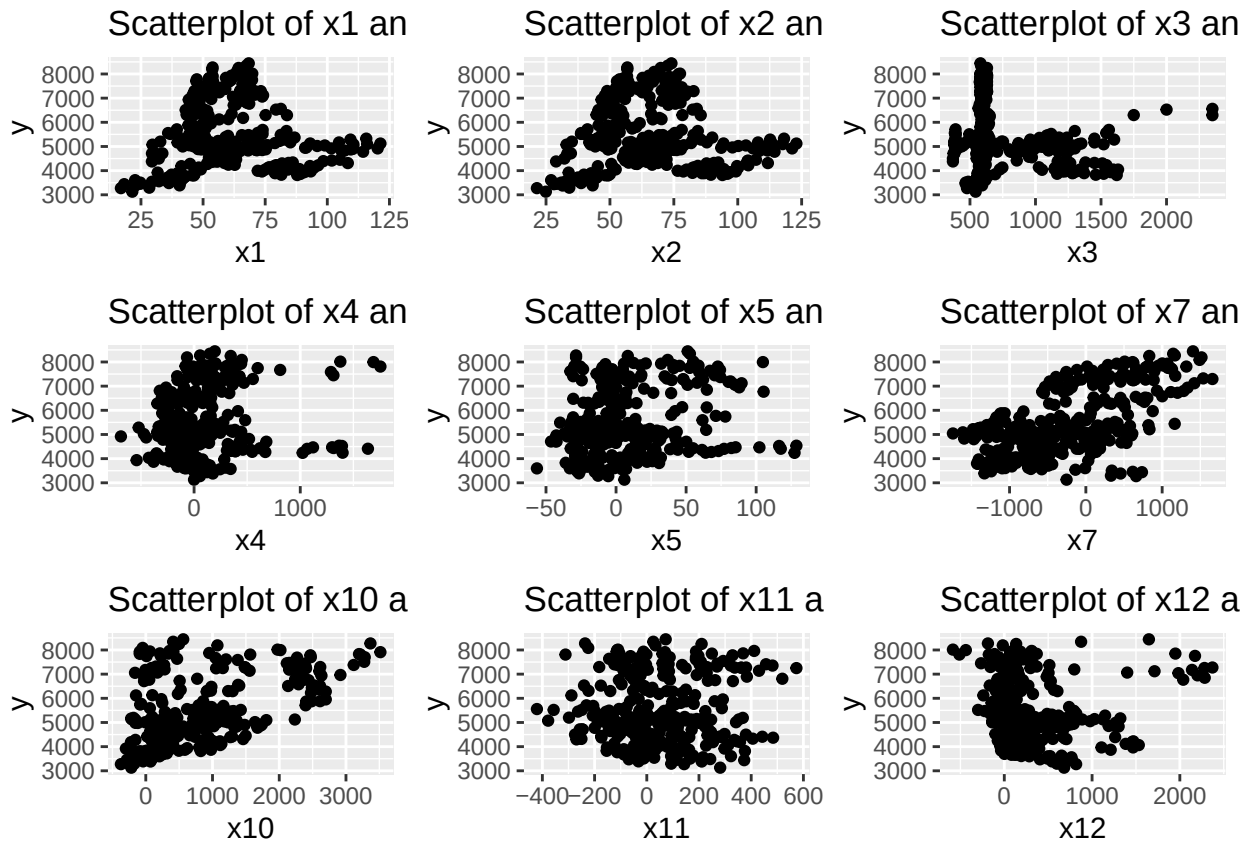
```
grid.arrange(hist_x21, hist_x23, hist_x24, hist_x25, hist_x26,
             hist_x27, hist_x28, hist_x30,
             nrow = 3, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

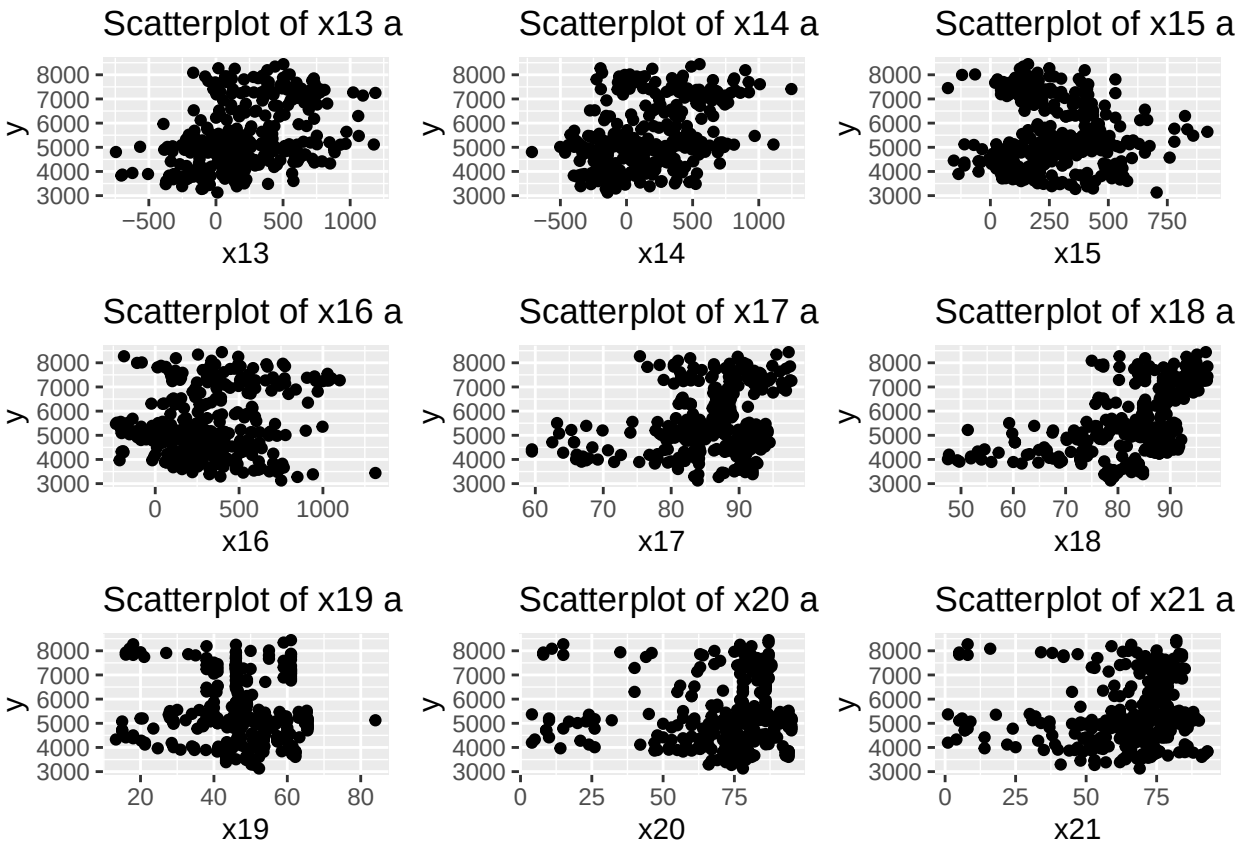



```
# create scatterplots of the variables x1, x2, x3, x4, x5, x7, x10, x11, x12
scatter_x1 <- ggplot(data = data, aes(x = x1, y = y)) + geom_point() + labs(title = "Scatterplot of x1 a
scatter_x2 <- ggplot(data = data, aes(x = x2, y = y)) + geom_point() + labs(title = "Scatterplot of x2 a
scatter_x3 <- ggplot(data = data, aes(x = x3, y = y)) + geom_point() + labs(title = "Scatterplot of x3 a
scatter_x4 <- ggplot(data = data, aes(x = x4, y = y)) + geom_point() + labs(title = "Scatterplot of x4 a
scatter_x5 <- ggplot(data = data, aes(x = x5, y = y)) + geom_point() + labs(title = "Scatterplot of x5 a
scatter_x7 <- ggplot(data = data, aes(x = x7, y = y)) + geom_point() + labs(title = "Scatterplot of x7 a
scatter_x10 <- ggplot(data = data, aes(x = x10, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x11 <- ggplot(data = data, aes(x = x11, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x12 <- ggplot(data = data, aes(x = x12, y = y)) + geom_point() + labs(title = "Scatterplot of x
```

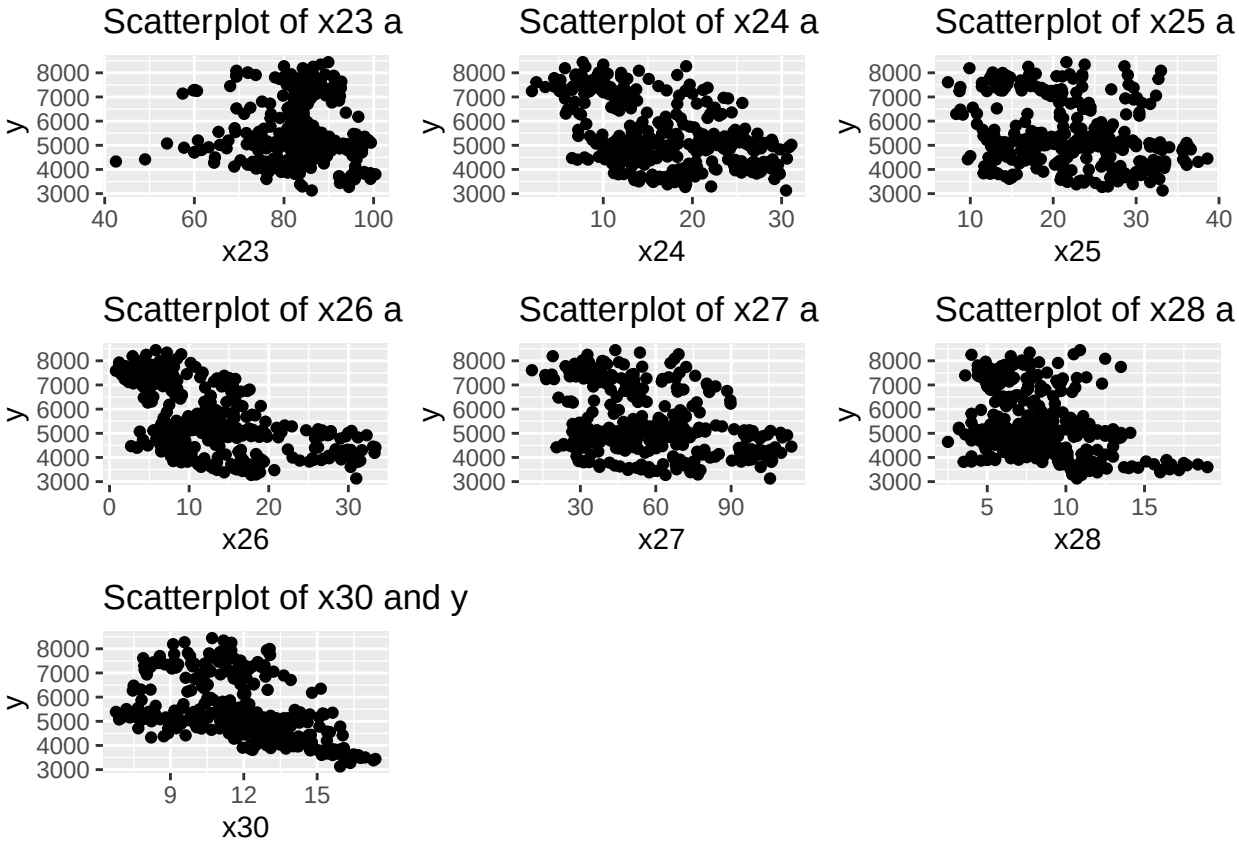
```
grid.arrange(scatter_x1, scatter_x2, scatter_x3, scatter_x4, scatter_x5,
          scatter_x7, scatter_x10, scatter_x11, scatter_x12, ncol = 3, nrow = 3)
```



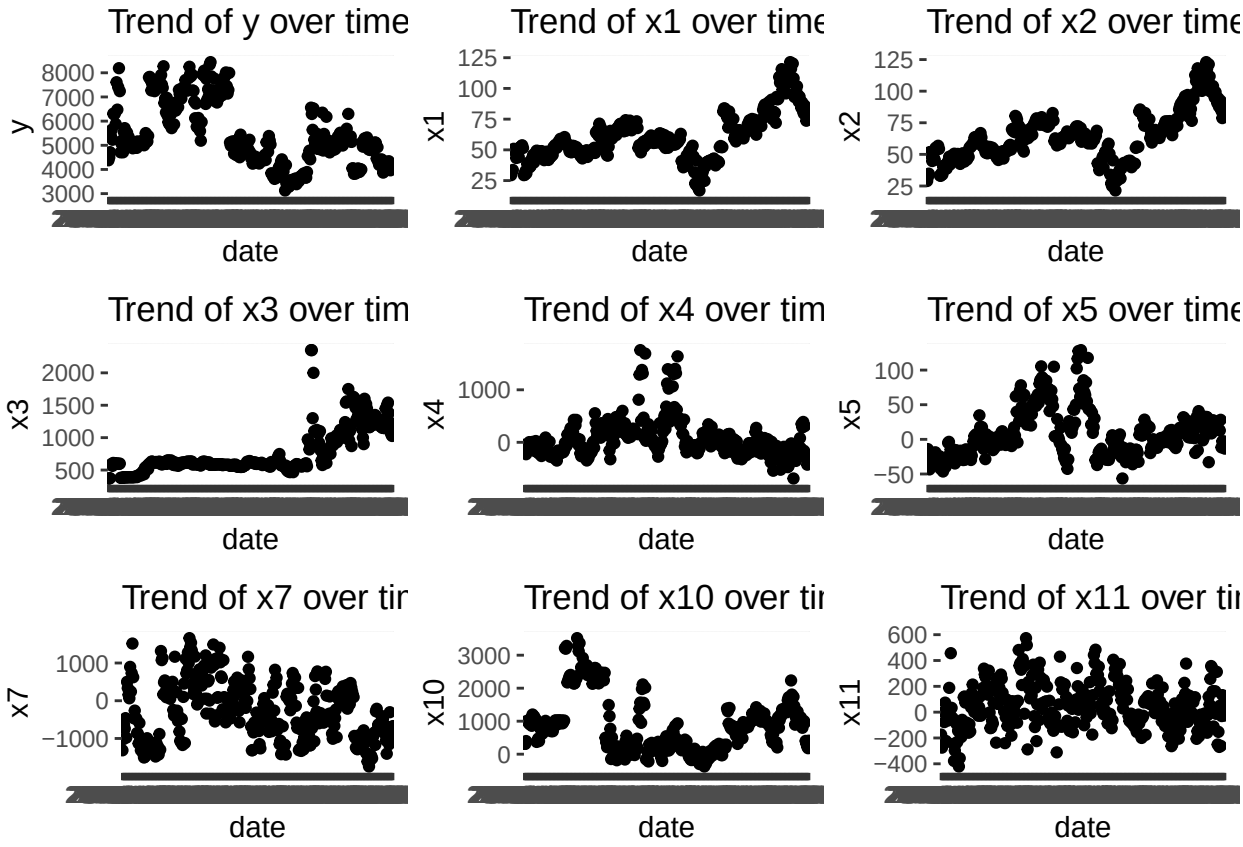

```
# create scatterplots of the variables x13, x14, x15, x16, x17, x18, x19, x20, x21
scatter_x13 <- ggplot(data = data, aes(x = x13, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x14 <- ggplot(data = data, aes(x = x14, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x15 <- ggplot(data = data, aes(x = x15, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x16 <- ggplot(data = data, aes(x = x16, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x17 <- ggplot(data = data, aes(x = x17, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x18 <- ggplot(data = data, aes(x = x18, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x19 <- ggplot(data = data, aes(x = x19, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x20 <- ggplot(data = data, aes(x = x20, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x21 <- ggplot(data = data, aes(x = x21, y = y)) + geom_point() + labs(title = "Scatterplot of x

grid.arrange(scatter_x13, scatter_x14, scatter_x15, scatter_x16, scatter_x17,
          scatter_x18, scatter_x19, scatter_x20, scatter_x21, ncol = 3, nrow = 3)
```




```r
# create scatterplots of the variables x23, x24, x25, x26, x27, x28, x30
scatter_x23 <- ggplot(data = data, aes(x = x23, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x24 <- ggplot(data = data, aes(x = x24, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x25 <- ggplot(data = data, aes(x = x25, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x26 <- ggplot(data = data, aes(x = x26, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x27 <- ggplot(data = data, aes(x = x27, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x28 <- ggplot(data = data, aes(x = x28, y = y)) + geom_point() + labs(title = "Scatterplot of x
scatter_x30 <- ggplot(data = data, aes(x = x30, y = y)) + geom_point() + labs(title = "Scatterplot of x

grid.arrange(scatter_x23, scatter_x24, scatter_x25, scatter_x26, scatter_x27,
          scatter_x28, scatter_x30, ncol = 3, nrow = 3)
```




```
# plot the variables versus date
trend_y <- ggplot(data = data, aes(x = date, y = y)) + geom_point() + labs(title = "Trend of y over tim
trend_x1 <- ggplot(data = data, aes(x = date, y = x1)) + geom_point() + labs(title = "Trend of x1 over
trend_x2 <- ggplot(data = data, aes(x = date, y = x2)) + geom_point() + labs(title = "Trend of x2 over
trend_x3 <- ggplot(data = data, aes(x = date, y = x3)) + geom_point() + labs(title = "Trend of x3 over
trend_x4 <- ggplot(data = data, aes(x = date, y = x4)) + geom_point() + labs(title = "Trend of x4 over
trend_x5 <- ggplot(data = data, aes(x = date, y = x5)) + geom_point() + labs(title = "Trend of x5 over
trend_x7 <- ggplot(data = data, aes(x = date, y = x7)) + geom_point() + labs(title = "Trend of x7 over
trend_x10 <- ggplot(data = data, aes(x = date, y = x10)) + geom_point() + labs(title = "Trend of x10 ov
trend_x11 <- ggplot(data = data, aes(x = date, y = x11)) + geom_point() + labs(title = "Trend of x11 ov

grid.arrange(trend_y, trend_x1, trend_x2, trend_x3, trend_x4, trend_x5,
          trend_x7, trend_x10, trend_x11, ncol = 3, nrow = 3)
```




```
## applications of statistical analysis techniques
## fitting models

# fit a linear regression model
model_ls <- lm(y ~ ., data = data_no_date)
summary(model_ls)
```

```
##
## Call:
## lm(formula = y ~ ., data = data_no_date)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1312.07  -318.68   -29.43   297.13  1434.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4008.76678  609.74607   6.574 1.94e-10 ***
## x1            17.36457   18.32233   0.948 0.343970
## x2            12.94083   17.85504   0.725 0.469112
## x3            -0.06382    0.18669  -0.342 0.732669
## x4            -0.16446    0.16950  -0.970 0.332639
## x5             5.40081    2.01929   2.675 0.007859 **
## x7             0.55400    0.04726  11.722  < 2e-16 ***
## x10            0.29464    0.05128   5.746 2.10e-08 ***
## x11           -1.12067    0.29935  -3.744 0.000214 ***
```

```
## x12            0.22659    0.08225   2.755 0.006201 **
## x13            0.11809    0.18233   0.648 0.517661
## x14            0.19558    0.17717   1.104 0.270440
## x15            0.44650    0.22016   2.028 0.043361 *
## x16            0.78960    0.19006   4.154 4.17e-05 ***
## x17          -84.72796   23.68640  -3.577 0.000400 ***
## x18           90.76171   14.62744   6.205 1.66e-09 ***
## x19          -41.71468    6.37786  -6.541 2.37e-10 ***
## x20           -3.35478    5.69031  -0.590 0.555894
## x21           17.14220    5.39011   3.180 0.001613 **
## x23            6.71208    7.30612   0.919 0.358936
## x24           -1.84148   17.30676  -0.106 0.915329
## x25           37.25362   16.60945   2.243 0.025574 *
## x26          -53.34944   15.29655  -3.488 0.000554 ***
## x27           -3.38515    9.73191  -0.348 0.728185
## x28          -11.08148   13.53817  -0.819 0.413649
## x30          -39.06404   25.10080  -1.556 0.120610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 505.5 on 326 degrees of freedom
## Multiple R-squared:  0.8645, Adjusted R-squared:  0.8541
## F-statistic: 83.16 on 25 and 326 DF,  p-value: < 2.2e-16
```

```r
# fit a stepwise regression model
model_step <- step(model_ls, direction = "both")
```

```
## Start:  AIC=4407.72
## y ~ x1 + x2 + x3 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 +
##     x15 + x16 + x17 + x18 + x19 + x20 + x21 + x23 + x24 + x25 +
##     x26 + x27 + x28 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x24   1      2893  83292418 4405.7
## - x3    1     29860  83319385 4405.8
## - x27   1     30912  83320437 4405.8
## - x20   1     88803  83378328 4406.1
## - x13   1    107168  83396693 4406.2
## - x2    1    134207  83423732 4406.3
## - x28   1    171178  83460703 4406.4
## - x23   1    215632  83505157 4406.6
## - x1    1    229477  83519002 4406.7
## - x4    1    240516  83530041 4406.7
## - x14   1    311352  83600877 4407.0
## <none>               83289525 4407.7
## - x30   1    618802  83908327 4408.3
## - x15   1   1050893  84340419 4410.1
## - x25   1   1285285  84574810 4411.1
## - x5    1   1827647  85117172 4413.4
## - x12   1   1938996  85228521 4413.8
## - x21   1   2584111  85873636 4416.5
## - x26   1   3107745  86397270 4418.6
## - x17   1   3269099  86558624 4419.3
## - x11   1   3580731  86870256 4420.5
```

```
## - x16   1   4409462  87698987 4423.9
## - x10   1   8435236  91724761 4439.7
## - x18   1   9836522  93126047 4445.0
## - x19   1  10929512  94219037 4449.1
## - x7    1  35108414 118397939 4529.5
##
## Step:  AIC=4405.73
## y ~ x1 + x2 + x3 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 +
##     x15 + x16 + x17 + x18 + x19 + x20 + x21 + x23 + x25 + x26 +
##     x27 + x28 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x3    1     32308  83324726 4403.9
## - x27   1     66073  83358491 4404.0
## - x20   1     96405  83388823 4404.1
## - x13   1    114034  83406452 4404.2
## - x2    1    141403  83433821 4404.3
## - x28   1    168304  83460722 4404.4
## - x23   1    214121  83506539 4404.6
## - x1    1    226621  83519039 4404.7
## - x4    1    243304  83535721 4404.8
## - x14   1    343226  83635643 4405.2
## <none>               83292418 4405.7
## - x30   1    628812  83921229 4406.4
## + x24   1      2893  83289525 4407.7
## - x15   1   1122481  84414899 4408.4
## - x25   1   1408599  84701016 4409.6
## - x5    1   1880967  85173385 4411.6
## - x12   1   2046599  85339016 4412.3
## - x21   1   2799980  86092397 4415.4
## - x26   1   3115093  86407510 4416.7
## - x17   1   3314162  86606580 4417.5
## - x11   1   3595054  86887471 4418.6
## - x16   1   4407435  87699853 4421.9
## - x10   1   8494995  91787412 4437.9
## - x18   1  10039439  93331857 4443.8
## - x19   1  11246360  94538778 4448.3
## - x7    1  35143217 118435634 4527.6
##
## Step:  AIC=4403.87
## y ~ x1 + x2 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 +
##     x16 + x17 + x18 + x19 + x20 + x21 + x23 + x25 + x26 + x27 +
##     x28 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x20   1     96932  83421657 4402.3
## - x27   1    106996  83431721 4402.3
## - x28   1    141129  83465855 4402.5
## - x13   1    141586  83466312 4402.5
## - x2    1    176750  83501476 4402.6
## - x1    1    196236  83520961 4402.7
## - x23   1    244821  83569547 4402.9
## - x4    1    259444  83584169 4403.0
## - x14   1    319417  83644142 4403.2
```

```
## <none>               83324726 4403.9
## - x30   1    725542  84050268 4404.9
## + x3    1     32308  83292418 4405.7
## + x24   1      5340  83319385 4405.8
## - x15   1   1094149  84418875 4406.5
## - x25   1   1795424  85120149 4409.4
## - x5    1   1942297  85267022 4410.0
## - x12   1   2102488  85427214 4410.6
## - x21   1   2774699  86099425 4413.4
## - x26   1   3083201  86407927 4414.7
## - x11   1   3563301  86888027 4416.6
## - x17   1   3613080  86937805 4416.8
## - x16   1   4375168  87699893 4419.9
## - x10   1   8735817  92060543 4437.0
## - x19   1  11321413  94646139 4446.7
## - x18   1  11583815  94908540 4447.7
## - x7    1  36034806 119359532 4528.4
##
## Step:  AIC=4402.28
## y ~ x1 + x2 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 +
##     x16 + x17 + x18 + x19 + x21 + x23 + x25 + x26 + x27 + x28 +
##     x30
##
##        Df Sum of Sq       RSS    AIC
## - x27   1    119399  83541056 4400.8
## - x28   1    138312  83559969 4400.9
## - x1    1    161914  83583572 4401.0
## - x2    1    229980  83651637 4401.2
## - x4    1    232325  83653982 4401.3
## - x13   1    261271  83682928 4401.4
## - x23   1    263761  83685419 4401.4
## - x14   1    276320  83697977 4401.4
## <none>               83421657 4402.3
## - x30   1    822205  84243863 4403.7
## + x20   1     96932  83324726 4403.9
## + x3    1     32835  83388823 4404.1
## + x24   1     14762  83406896 4404.2
## - x15   1   1197185  84618842 4405.3
## - x25   1   1899130  85320788 4408.2
## - x5    1   1963455  85385112 4408.5
## - x12   1   2062697  85484354 4408.9
## - x26   1   3048942  86470599 4412.9
## - x17   1   3596055  87017712 4415.1
## - x11   1   3820017  87241675 4416.0
## - x16   1   4527861  87949518 4418.9
## - x21   1   5543478  88965136 4422.9
## - x10   1   8832371  92254029 4435.7
## - x18   1  11487696  94909354 4445.7
## - x19   1  15484009  98905666 4460.2
## - x7    1  35987138 119408795 4526.5
##
## Step:  AIC=4400.78
## y ~ x1 + x2 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 +
##     x16 + x17 + x18 + x19 + x21 + x23 + x25 + x26 + x28 + x30
```

```
##
##        Df Sum of Sq       RSS    AIC
## - x28   1    106839  83647895 4399.2
## - x4    1    174707  83715764 4399.5
## - x1    1    178785  83719842 4399.5
## - x13   1    187512  83728569 4399.6
## - x2    1    207867  83748923 4399.7
## - x23   1    268051  83809108 4399.9
## - x14   1    380593  83921650 4400.4
## <none>               83541056 4400.8
## + x27   1    119399  83421657 4402.3
## + x20   1    109335  83431721 4402.3
## + x24   1     96366  83444690 4402.4
## + x3    1     77173  83463884 4402.5
## - x30   1    930306  84471362 4402.7
## - x15   1   1087260  84628317 4403.3
## - x5    1   1856788  85397845 4406.5
## - x12   1   2058887  85599944 4407.4
## - x25   1   3444233  86985289 4413.0
## - x17   1   3483291  87024348 4413.2
## - x11   1   3941751  87482808 4415.0
## - x16   1   5381509  88922565 4420.8
## - x21   1   5425428  88966485 4420.9
## - x10   1   8713365  92254421 4433.7
## - x26   1   9151984  92693040 4435.4
## - x18   1  11368332  94909388 4443.7
## - x19   1  15468584  99009641 4458.6
## - x7    1  36441641 119982697 4526.2
##
## Step:  AIC=4399.23
## y ~ x1 + x2 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 +
##     x16 + x17 + x18 + x19 + x21 + x23 + x25 + x26 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x1    1    121423  83769318 4397.7
## - x4    1    218437  83866332 4398.1
## - x13   1    233327  83881222 4398.2
## - x2    1    319702  83967597 4398.6
## - x23   1    353502  84001397 4398.7
## - x14   1    367501  84015396 4398.8
## <none>               83647895 4399.2
## + x28   1    106839  83541056 4400.8
## + x20   1    104929  83542966 4400.8
## + x27   1     87926  83559969 4400.9
## + x24   1     53029  83594866 4401.0
## + x3    1     28084  83619811 4401.1
## - x15   1   1075487  84723382 4401.7
## - x30   1   1209101  84856996 4402.3
## - x12   1   1954333  85602229 4405.4
## - x5    1   2014457  85662352 4405.6
## - x25   1   3337481  86985376 4411.0
## - x17   1   3581457  87229352 4412.0
## - x11   1   4037929  87685824 4413.8
## - x16   1   5321256  88969151 4418.9
```

```
## - x21   1   5337011  88984906 4419.0
## - x10   1   8921714  92569609 4432.9
## - x26   1   9067749  92715644 4433.5
## - x18   1  11494374  95142269 4442.6
## - x19   1  15411967  99059862 4456.8
## - x7    1  37139465 120787360 4526.6
##
## Step:  AIC=4397.74
## y ~ x2 + x4 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 + x16 +
##     x17 + x18 + x19 + x21 + x23 + x25 + x26 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x4    1    250555  84019873 4396.8
## - x14   1    313908  84083226 4397.1
## - x13   1    345333  84114651 4397.2
## - x23   1    393064  84162382 4397.4
## <none>               83769318 4397.7
## + x1    1    121423  83647895 4399.2
## + x27   1    108165  83661153 4399.3
## + x20   1     73890  83695428 4399.4
## + x28   1     49476  83719842 4399.5
## + x24   1     45142  83724176 4399.6
## + x3    1      9943  83759375 4399.7
## - x15   1   1152865  84922183 4400.6
## - x30   1   1281146  85050464 4401.1
## - x5    1   1894923  85664241 4403.6
## - x12   1   2109125  85878443 4404.5
## - x25   1   3397548  87166866 4409.7
## - x17   1   3475590  87244908 4410.1
## - x11   1   4090529  87859847 4412.5
## - x16   1   5284310  89053628 4417.3
## - x21   1   5582401  89351719 4418.4
## - x26   1   9059260  92828578 4431.9
## - x10   1   9914581  93683899 4435.1
## - x18   1  11374190  95143508 4440.6
## - x19   1  17407181 101176499 4462.2
## - x2    1  25351544 109120862 4488.8
## - x7    1  37496562 121265880 4526.0
##
## Step:  AIC=4396.79
## y ~ x2 + x5 + x7 + x10 + x11 + x12 + x13 + x14 + x15 + x16 +
##     x17 + x18 + x19 + x21 + x23 + x25 + x26 + x30
##
##        Df Sum of Sq       RSS    AIC
## - x14   1    287632  84307505 4396.0
## - x13   1    411277  84431150 4396.5
## - x23   1    438936  84458809 4396.6
## <none>               84019873 4396.8
## + x4    1    250555  83769318 4397.7
## + x1    1    153541  83866332 4398.1
## + x28   1     76211  83943662 4398.5
## + x27   1     43728  83976145 4398.6
## + x20   1     42972  83976901 4398.6
## + x24   1     19840  84000033 4398.7
```

```
## + x3    1      5865  84014008 4398.8
## - x30   1   1781094  85800967 4402.2
## - x15   1   1865292  85885165 4402.5
## - x5    1   2435420  86455293 4404.9
## - x12   1   2670633  86690506 4405.8
## - x17   1   3450284  87470157 4409.0
## - x11   1   3973129  87993003 4411.1
## - x25   1   4612614  88632488 4413.6
## - x16   1   5546323  89566196 4417.3
## - x21   1   5759529  89779402 4418.1
## - x26   1   9495506  93515379 4432.5
## - x10   1   9680256  93700129 4433.2
## - x18   1  11181952  95201825 4438.8
## - x19   1  17312003 101331876 4460.7
## - x2    1  31768956 115788830 4507.7
## - x7    1  37254109 121273982 4524.0
##
## Step:  AIC=4396
## y ~ x2 + x5 + x7 + x10 + x11 + x12 + x13 + x15 + x16 + x17 +
##     x18 + x19 + x21 + x23 + x25 + x26 + x30
##
##        Df Sum of Sq       RSS    AIC
## <none>              84307505 4396.0
## - x23   1    582261  84889766 4396.4
## + x14   1    287632  84019873 4396.8
## + x4    1    224279  84083226 4397.1
## + x27   1    102373  84205132 4397.6
## + x1    1     92804  84214701 4397.6
## + x24   1     86834  84220671 4397.6
## + x28   1     76785  84230720 4397.7
## + x20   1     22051  84285454 4397.9
## + x3    1      3146  84304359 4398.0
## - x13   1   1388601  85696106 4399.7
## - x15   1   2089062  86396567 4402.6
## - x5    1   2397180  86704685 4403.9
## - x30   1   2419466  86726971 4404.0
## - x12   1   3346479  87653984 4407.7
## - x11   1   3685618  87993123 4409.1
## - x17   1   3737082  88044587 4409.3
## - x25   1   5563514  89871019 4416.5
## - x16   1   6005477  90312982 4418.2
## - x21   1   6916701  91224206 4421.8
## - x10   1   9395862  93703367 4431.2
## - x26   1  10904521  95212026 4436.8
## - x18   1  11289976  95597481 4438.2
## - x19   1  18945127 103252632 4465.3
## - x2    1  32970463 117277968 4510.2
## - x7    1  37105081 121412586 4522.4
```

```
summary(model_step)
```

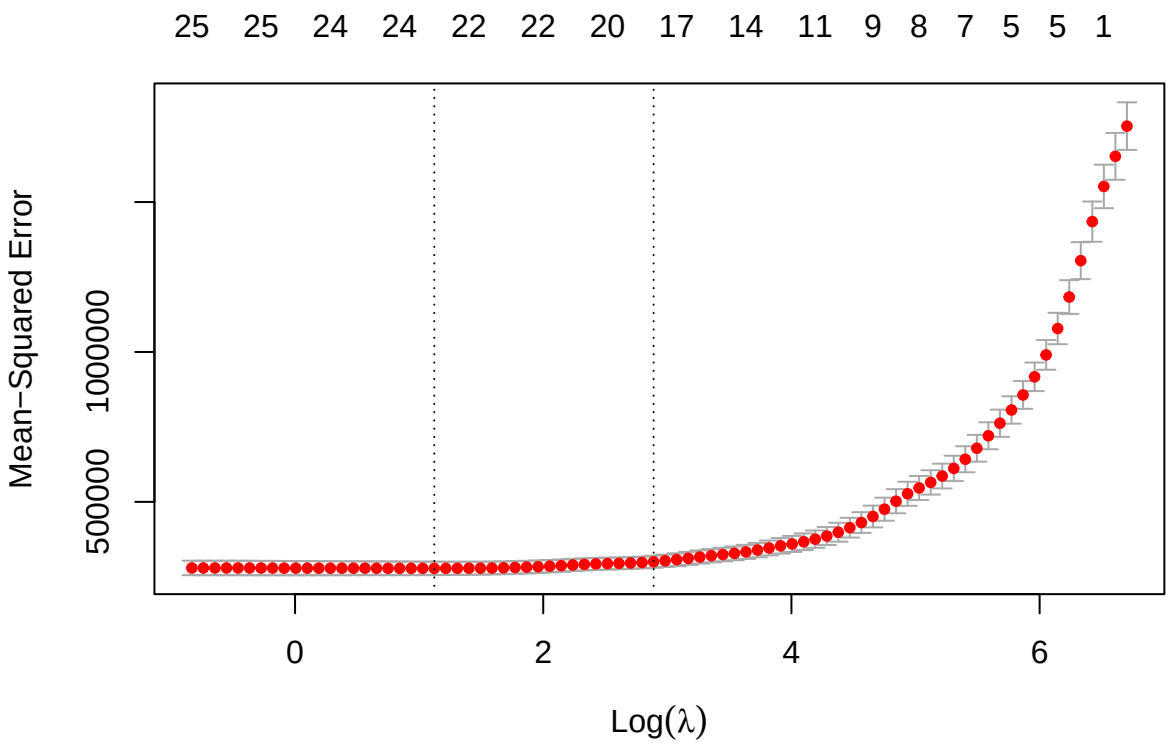
```
##
## Call:
## lm(formula = y ~ x2 + x5 + x7 + x10 + x11 + x12 + x13 + x15 +
```

```
##     x16 + x17 + x18 + x19 + x21 + x23 + x25 + x26 + x30, data = data_no_date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1363.0  -322.9   -29.9   296.4  1461.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3840.65818  535.96123   7.166 4.98e-12 ***
## x2            30.97228    2.71001  11.429  < 2e-16 ***
## x5             3.54262    1.14957   3.082 0.002229 **
## x7             0.54449    0.04491  12.124  < 2e-16 ***
## x10            0.27595    0.04523   6.101 2.92e-09 ***
## x11           -1.06544    0.27883  -3.821 0.000158 ***
## x12            0.26457    0.07266   3.641 0.000315 ***
## x13            0.26367    0.11242   2.345 0.019588 *
## x15            0.52555    0.18268   2.877 0.004275 **
## x16            0.83147    0.17046   4.878 1.66e-06 ***
## x17          -85.91303   22.32814  -3.848 0.000143 ***
## x18           89.69031   13.41093   6.688 9.54e-11 ***
## x19          -44.51743    5.13856  -8.663  < 2e-16 ***
## x21           15.12377    2.88915   5.235 2.93e-07 ***
## x23           10.45120    6.88125   1.519 0.129760
## x25           36.82319    7.84344   4.695 3.90e-06 ***
## x26          -63.54198    9.66757  -6.573 1.90e-10 ***
## x30          -64.60483   20.86725  -3.096 0.002128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 502.4 on 334 degrees of freedom
## Multiple R-squared:  0.8628, Adjusted R-squared:  0.8558
## F-statistic: 123.6 on 17 and 334 DF,  p-value: < 2.2e-16
```

```
# fit a lasso regression model
model_lasso <- cv.glmnet(x = as.matrix(data_no_date[, -1]), y = data_no_date[, 1], alpha = 1)
model_lasso
```

```
##
## Call:  cv.glmnet(x = as.matrix(data_no_date[, -1]), y = data_no_date[,      1], alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min  3.071    61  278141 22335      22
## 1se 17.985    42  300242 20023      19
```

```
# plot the lasso regression model
plot(model_lasso)
```


25 25 24 24 22 22 20 17 14 11 9 8 7 5 5 1
Mean−Squared Error
500000
1000000
0
2
4
6
Log(λ)