Applied Linear Regression on The Insurance Dataset

2023-12-04

Group Members

Eric Wu, Derek Avery, Allison Schaedler, Xinyi Bao

Background

The Insurance dataset extracted from Kaggle.com describes the medical costs for over 1300 individuals based on their age, sex, BMI in the United States. The goal of the project is to examine how many children, Smoker or not, and which region in the U.S they lived. The methods of the project are using multiple linear regression to find the best model with the lowest MSE value and use LASSO regression to make the comparison. In addition, there will be a detail diagnostic on the linear regression model for the assumptions, outliers, and other possible issues.

The dataset come from here

Exploratory Data Analysis

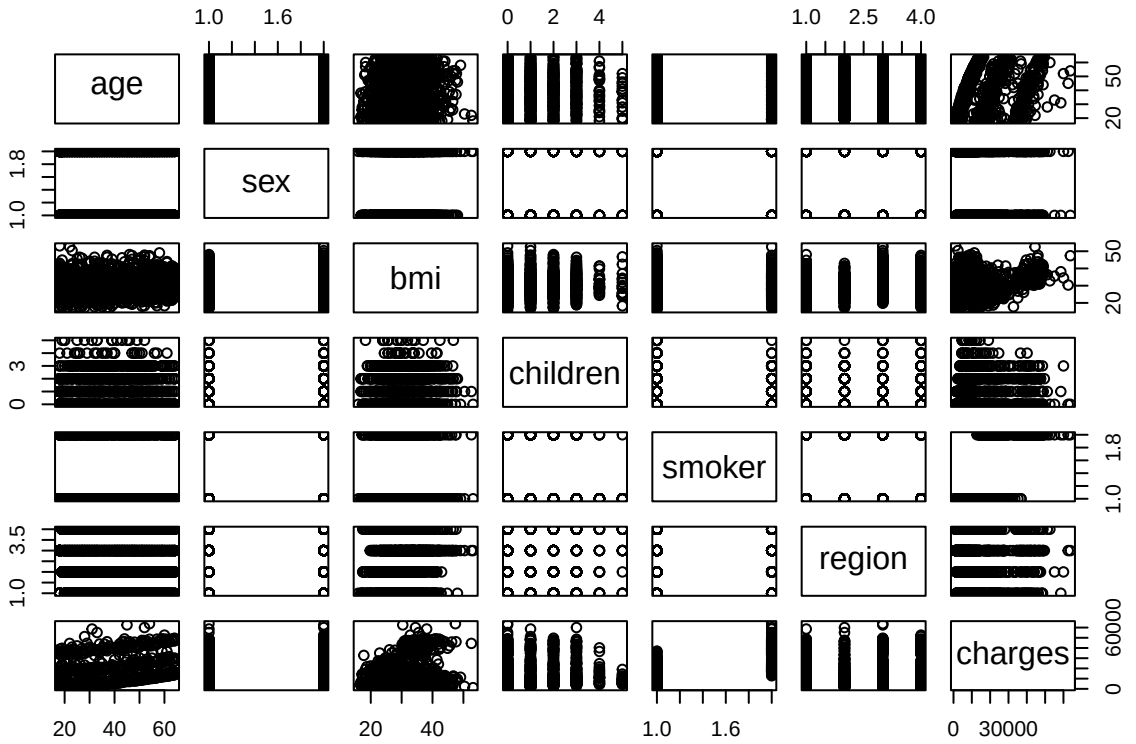
Data Statistics

Table 1: Basic statistics of quantitative variables

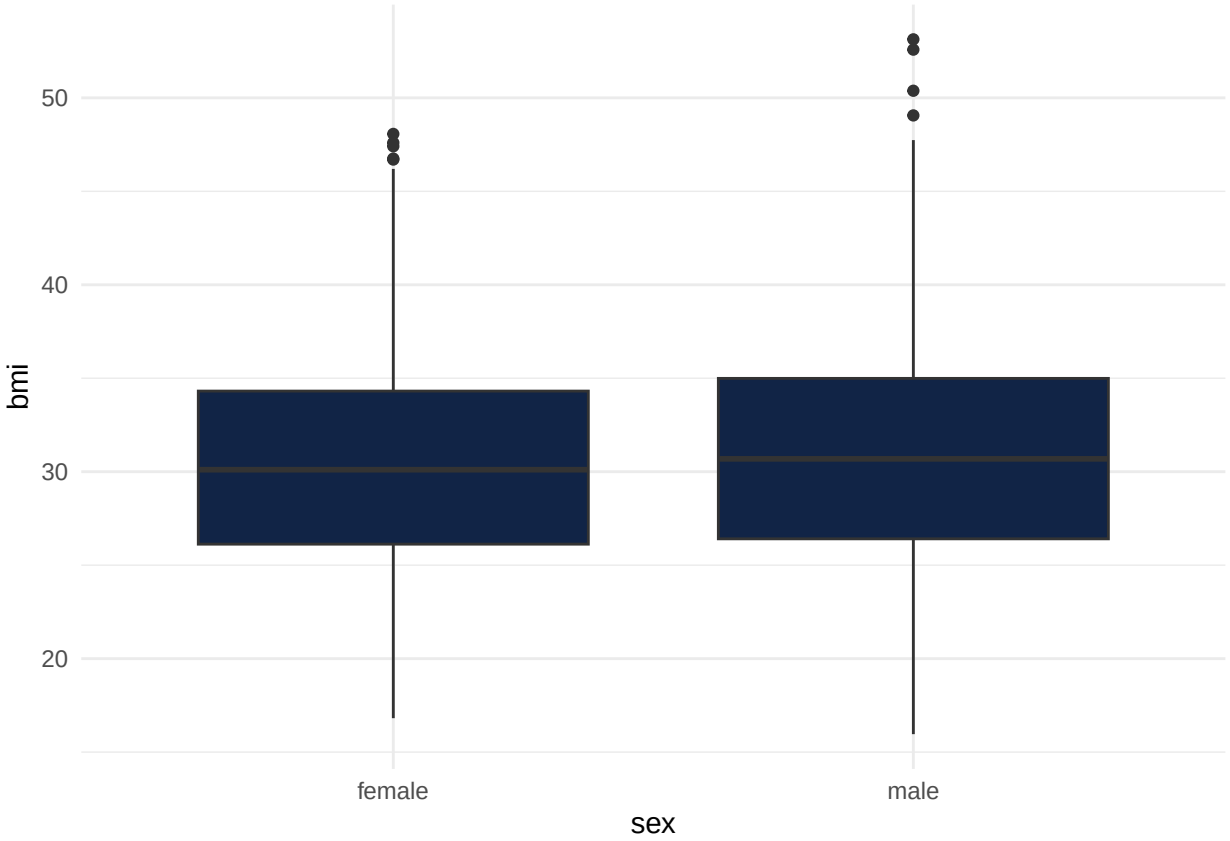
	vars	n	mean	sd	min	max	range	se
age	1	1338	39.21	14.05	18.00	64.00	46.00	0.38
bmi	2	1338	30.66	6.10	15.96	53.13	37.17	0.17
children	3	1338	1.09	1.21	0.00	5.00	5.00	0.03
charges	4	1338	13270.42	12110.01	1121.87	63770.43	62648.55	331.07

The statistic table presents some of the basic statistic summaries for each quantitative variables. Including the number of records, mean, standard deviation, minimum, maximum, range, and standard error of the corresponding variable. Some interesting finding including the average of BMI, as the average BMI category of the dataset is classified as obesity, this suggested that most people probably are classified as overweight or more.

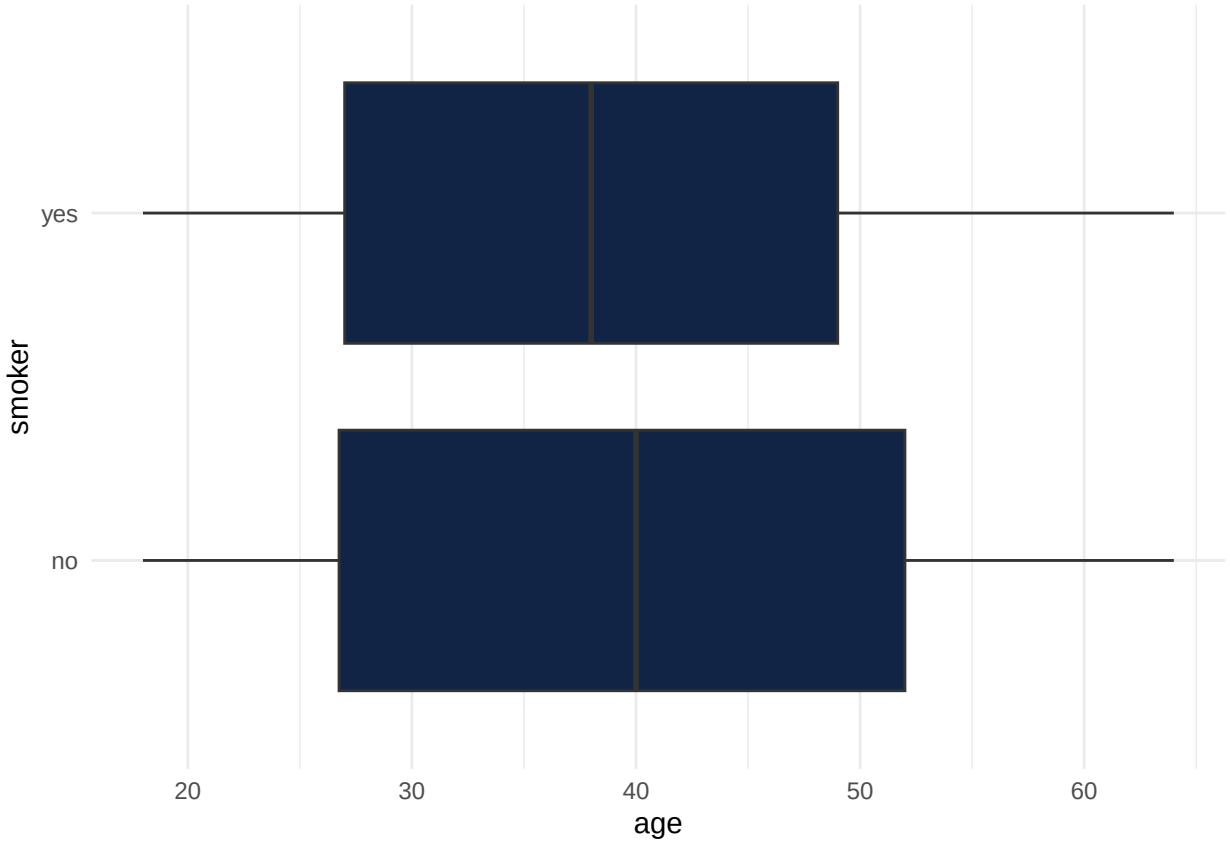
Data Visualization



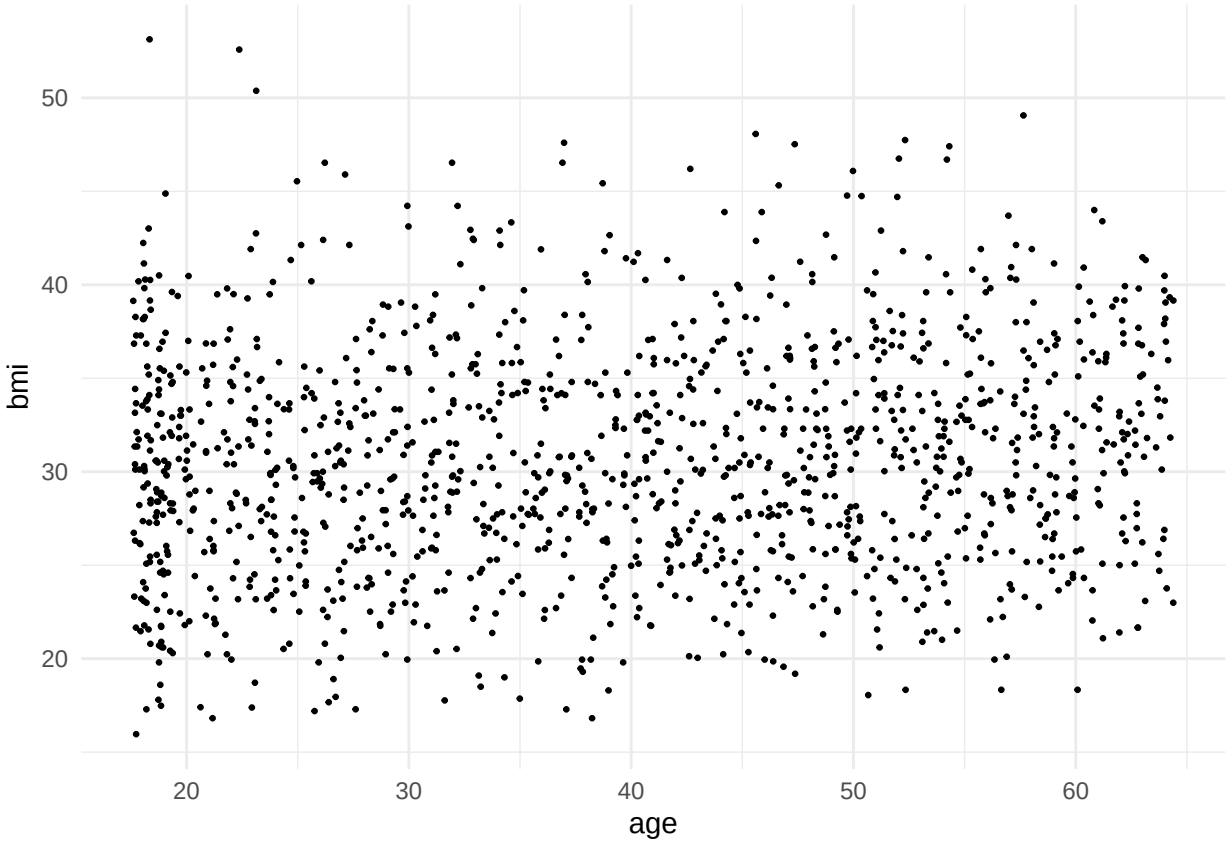
These plots display the relationships between each of the variables in the dataset.



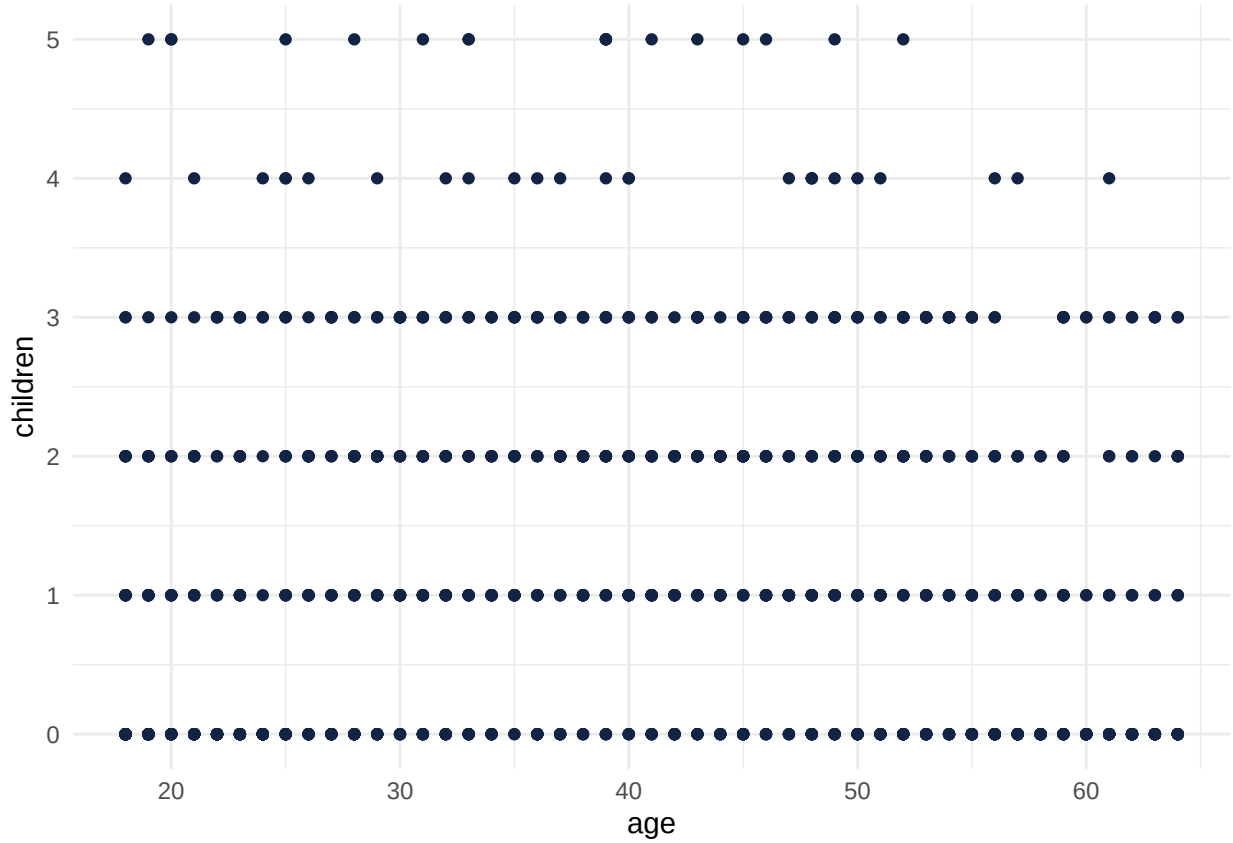
These boxplots display the relationship between the variables sex and BMI. There are a few outliers for each, but they are not significantly far from the rest of the data. The average male BMI is slightly higher than the average female BMI.



These boxplots show the relationship between the variables age and smoker. The average age of smokers is slightly lower than the average age of non-smokers.There are no outliers visible on the plot.



This plot shows the distribution of the variables BMI and age. The data appear very random and evenly distributed across the plot. There is considerable variation in the datapoints.



This plot shows the relationship between age and amount of children. Most of the data is concentrated between 0 to 3 children across the whole spectrum of ages.

Model Building

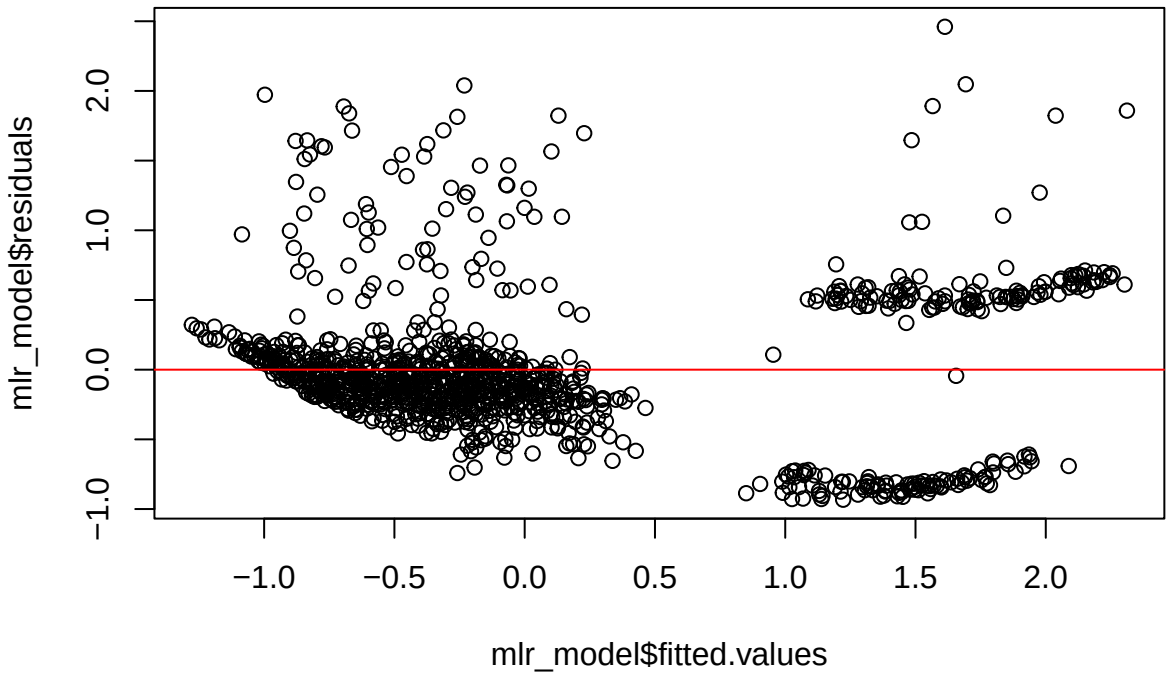
Multiple Linear Regression

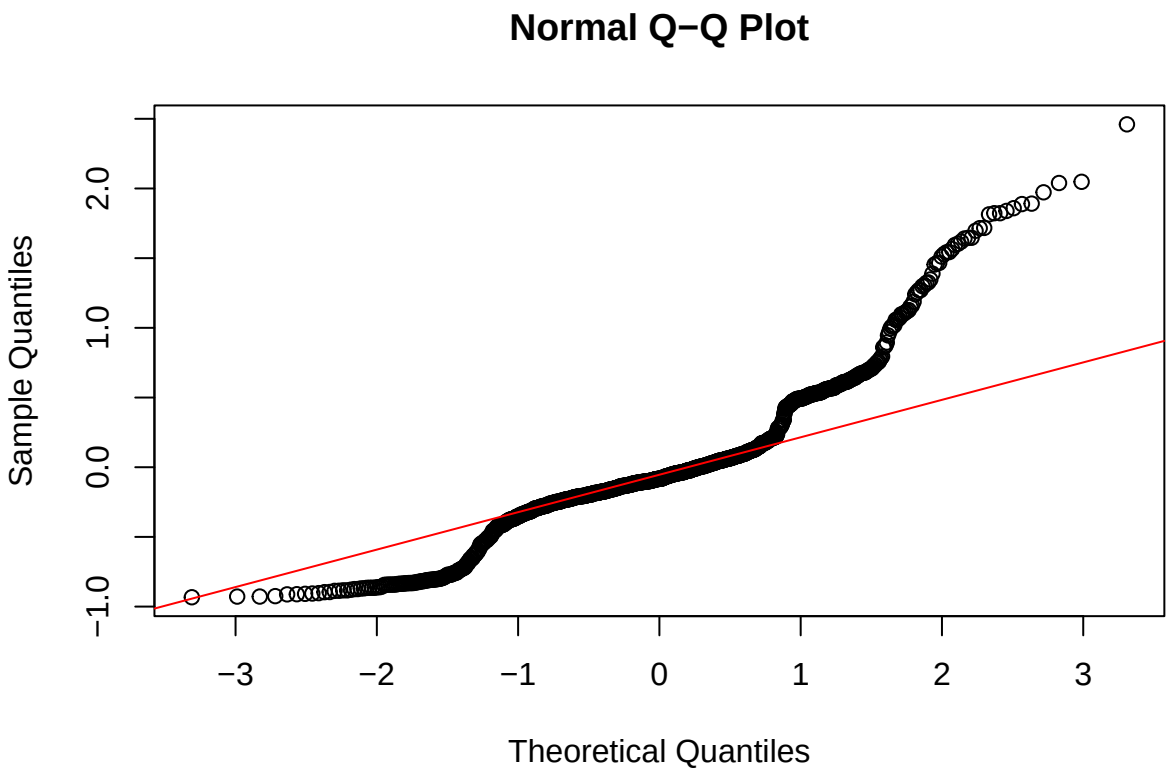
```
##
## Call:
## lm(formula = charges ~ ., data = scaled_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93288 -0.23503 -0.08298  0.12737  2.46005
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.001665   0.015529  -0.107 0.914651
## age           0.282430   0.015837  17.834  < 2e-16 ***
## bmi           0.178795   0.016427  10.884  < 2e-16 ***
## children      0.055729   0.015557   3.582 0.000356 ***
## is_male      -0.008957   0.015633  -0.573 0.566792
## is_smoker     0.805797   0.015333  52.552  < 2e-16 ***
## is_southwest -0.034285   0.019199  -1.786 0.074421 .
## is_southeast -0.031476   0.019915  -1.581 0.114281
```

```
## is_northwest -0.019715   0.019115  -1.031 0.302594
## is_northeast        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5074 on 1061 degrees of freedom
## Multiple R-squared:  0.7533, Adjusted R-squared:  0.7514
## F-statistic: 404.9 on 8 and 1061 DF,  p-value: < 2.2e-16
```

```
## Warning in predict.lm(mlr_model, newdata = scaled_test): prediction from a
## rank-deficient fit may be misleading
```

```
## [1] "MSE for Multiple Linear Regression:  33216601.9547535"
```






```
## Model :
## charges ~ (age + bmi + children + is_male + is_smoker + is_southwest +
##     is_southeast + is_northwest + is_northeast) - is_northeast

##          age          bmi     children      is_male    is_smoker is_southwest
##     1.016119     1.113290     1.002896     1.013515     1.008439     1.526926
## is_southeast is_northwest
##     1.646481     1.513548
```

Model Summary:

The model explains a significant amount of variance in the charges, with an R-squared value of 0.7533. This indicates that approximately 75.33% of the variability can be explained by the model.
The Adjusted R-squared value is 0.7514, which suggests that the model fits the data well.
The F-statistic is highly significant ($p < 2.2e\text{-}16$), indicating that the model as a whole is statistically significant.

Coefficients:

age, bmi, children, and is_smoker are significant predictors of charges, with p-values less than 0.05. age has a coefficient of 0.282, suggesting that each additional year of age is associated with an increase in charges by a factor of 0.282, holding other variables constant.
bmi is also a significant predictor, with each unit increase in BMI associated with an increase in charges by a factor of 0.178. children has a positive coefficient (0.0557), indicating that having more children is associated with higher charges.
is_smoker has the largest coefficient of 0.8058, which means being a smoker is associated with a substantial increase in charges. The variables is_male, is_southwest, is_southeast, and is_northwest are not statistically

significant at the 0.05 level.
However, is_southwest has a p-value close to the significance level, which may warrant further investigation. The is_northeast variable is not defined due to singularities, likely due to it being perfectly collinear with the other region variables.

Residual Analysis:

The Residuals vs Fitted plot shows a random scatter of residuals, which is good as it suggests that the variance of the residuals is constant (homoscedasticity). However, there seems to be a slight pattern with the residuals fanning out for larger fitted values, which could indicate potential issues with non-constant variance.
The Normal Q-Q Plot shows that the residuals deviate from the line at the two ends, indicating that the residuals may not be normally distributed. This could affect the model's confidence intervals and hypothesis tests.

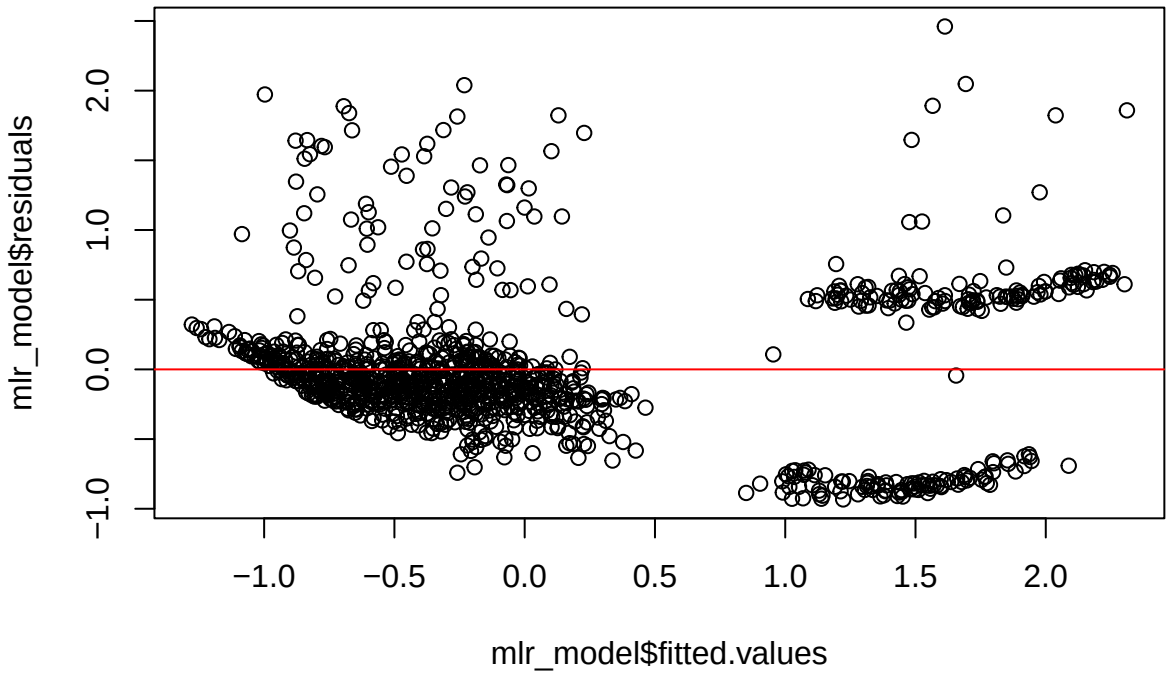
Other Analysis:

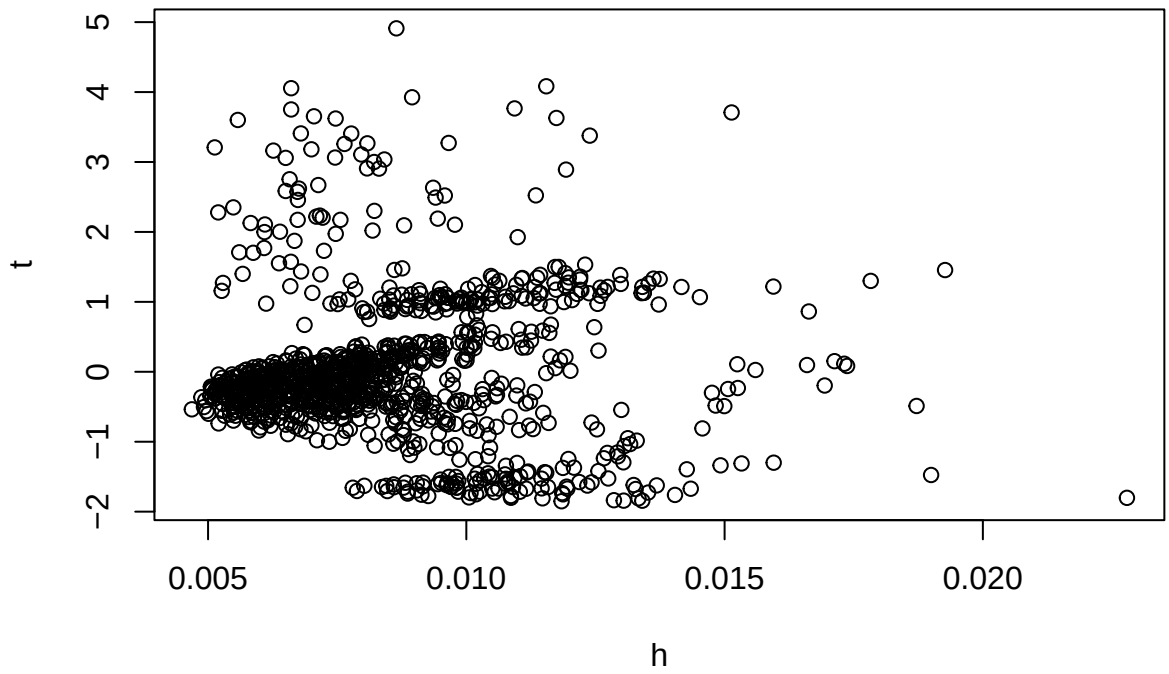
Key variables like age and smoking status are strong predictors of charges.
The Mean Squared Error (MSE) for the model is 0.2265, which is relatively low and indicates good predictive performance.

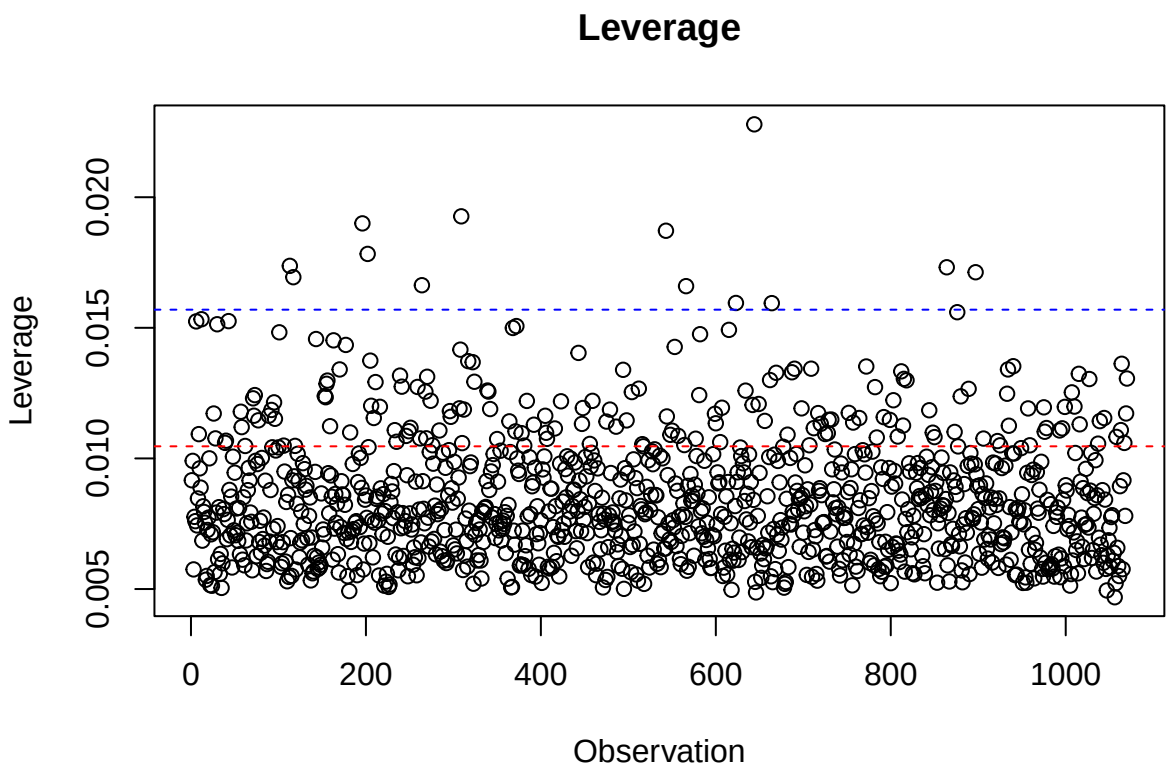
LINE Assumption

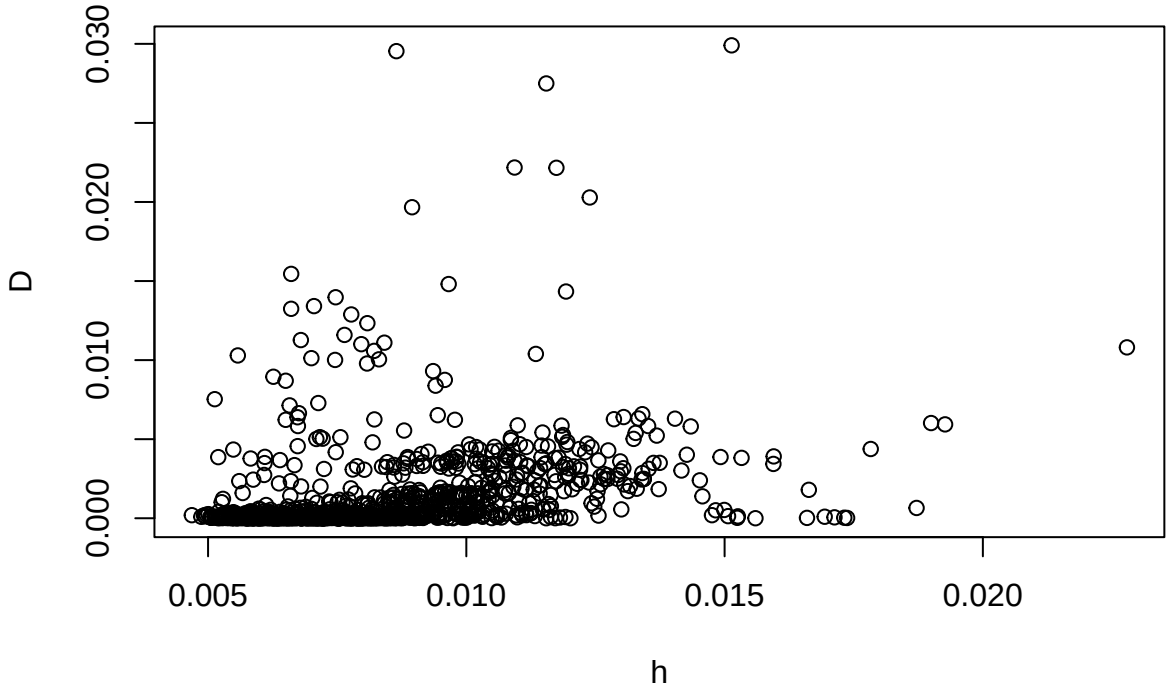
The basic assumptions for a linear regression are: Linearity, Independence, Normality, and Equivariance. If any of these assumptions are violated, there are two concerns: incorrect conclusions and interpretations, or a lack of power for discovery.

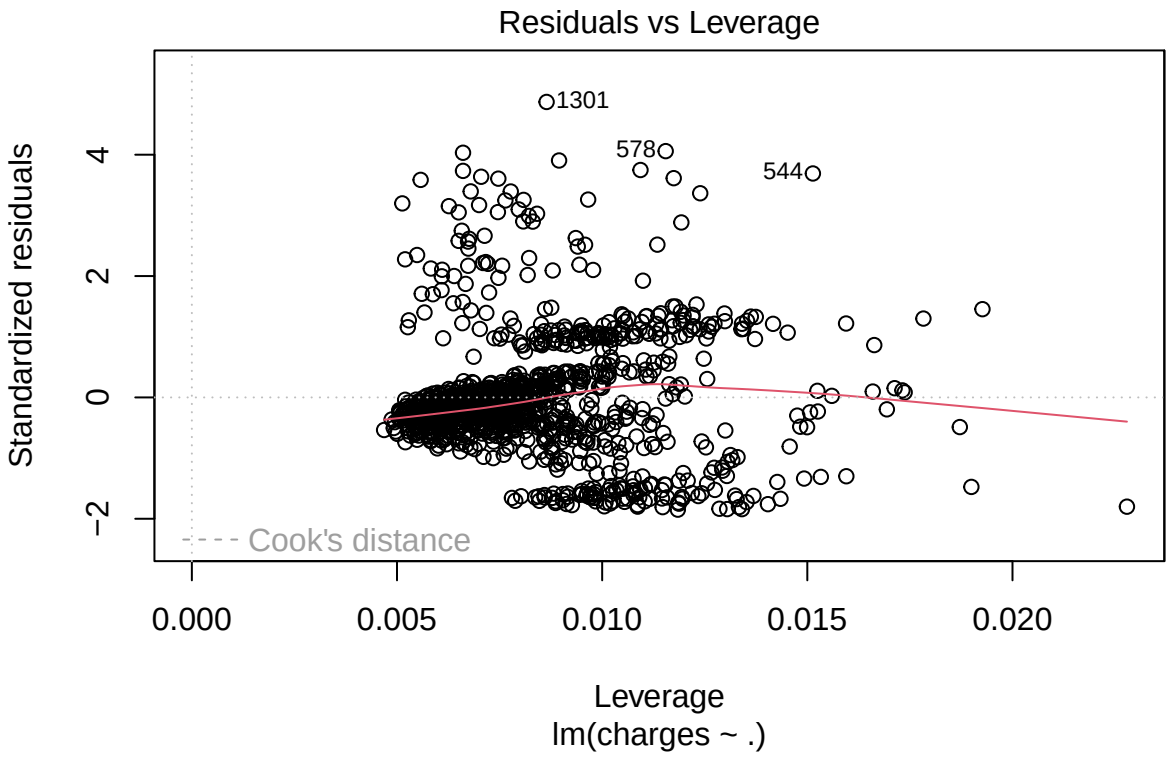
We analyze these assumptions by using diagnostic plots on our full model.

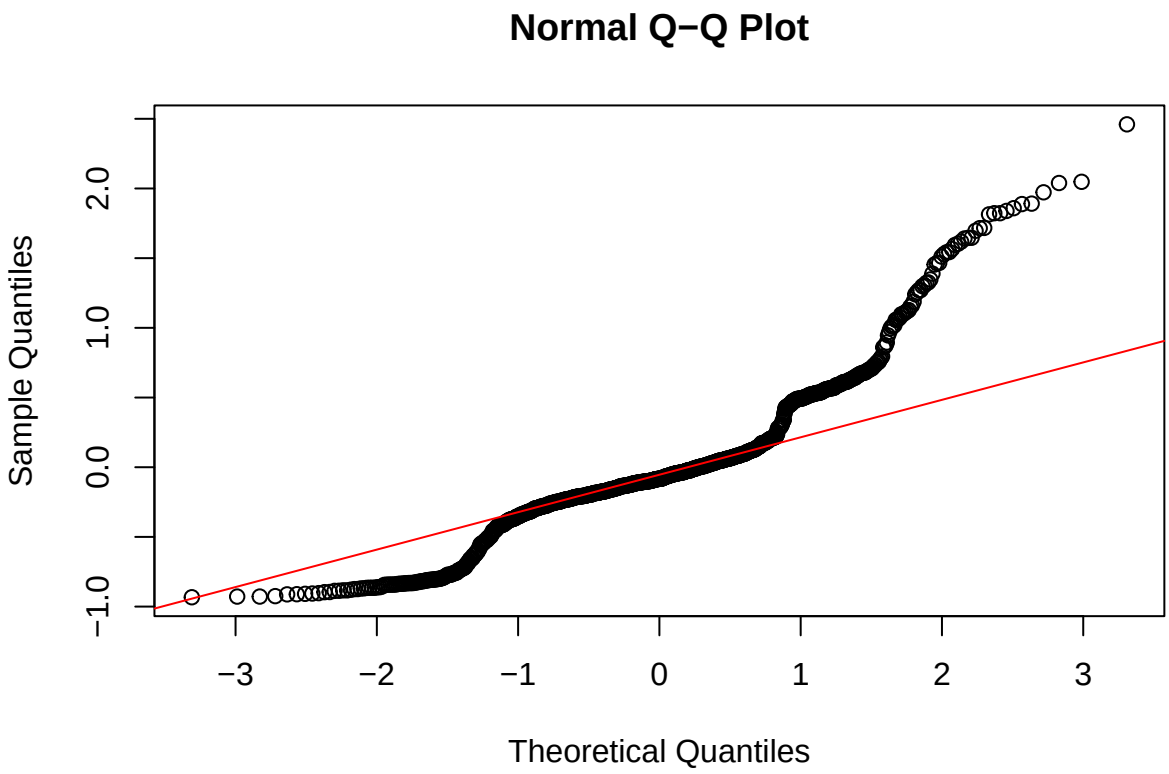



t
5
4
3
2
1
0
−1
−2
0.005
0.010
0.015
0.020
h


Leverage
Leverage
0.005
0.010
0.015
0.020
0
200
400
600
800
1000
Observation


D
0.000
0.010
0.020
0.030
0.005
0.010
0.015
0.020
h


Residuals vs Leverage
1301
578
544
Cook's distance
Standardized residuals
4
2
0
−2
0.000
0.005
0.010
0.015
0.020
Leverage
lm(charges ~ .)



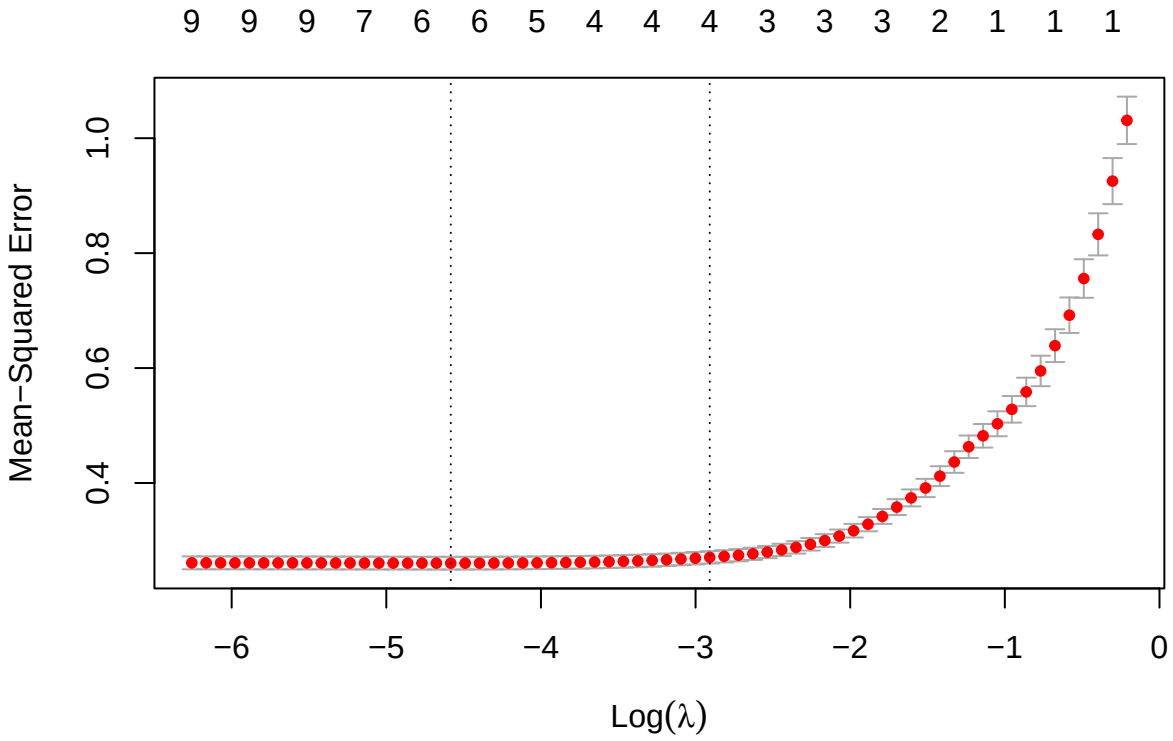

```
##
##  Shapiro-Wilk normality test
##
## data:  mlr_model$residuals
## W = 0.9048, p-value < 2.2e-16
```

Based on the Shapiro-Wilk test, the data is not normally distributed. However, due to the fact our dataset is quite large, we can reasonably ignore the issue. This is because (n - p) >= 30 where (1338 - 10) >= 30.

```
## Model :
## charges ~ (age + bmi + children + is_male + is_smoker + is_southwest +
##     is_southeast + is_northwest + is_northeast) - is_northeast
```

```
##          age          bmi     children      is_male    is_smoker is_southwest
##     1.016119     1.113290     1.002896     1.013515     1.008439     1.526926
## is_southeast is_northwest
##     1.646481     1.513548
```

LASSO Regression




```
## [1] 33212733
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                            s1
## (Intercept)  -0.001053016
## age           0.273576379
## bmi           0.165102982
## children      0.045557391
## is_male       .
## is_smoker     0.794871464
## is_southwest -0.001841697
## is_southeast  .
## is_northwest  .
## is_northeast  0.015205565
```

Model Comparison

Discussion

Several data visualization techniques were utilized for exploratory analysis of the data. The boxplots of smokers stratified by age show that the average age of smokers is slightly lower than that of non-smokers.

The relationship between the variables age and children were also plotted in a scatterplot. There is no visible pattern or correlation in the data. Most of the data is concentrated between 0 and 3 children across all ages. The points are random and evenly distributed.

Conclusion

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
# Load necessary Libraries
library(tidyverse)
library(ggplot2)
library(janitor)
library(kableExtra)
library(glmnet)
library(faraway)

# Load Dataset
data <- read.csv("insurance.csv")
# Check for Null Value
na <- sum(is.na(data))

# Create dummy variable
dummy_data <- data %>%
  mutate(
    is_male = if_else(sex == "male", 1, 0),
    is_smoker = if_else(smoker == "yes", 1, 0),
    is_southwest = if_else(region == "southwest", 1, 0),
    is_southeast = if_else(region == "southeast", 1, 0),
    is_northwest = if_else(region == "northwest", 1, 0),
    is_northeast = if_else(region == "northeast", 1, 0)
  ) %>%
  select(-sex, -smoker, -region)

# Create standardized dataset for ML that requires scaling
xvars <- names(dummy_data)
scaled_data <- dummy_data
scaled_data[ , xvars] <- scale(scaled_data[ , xvars],
                               center = TRUE,
                               scale = TRUE)

scaled_data <- scale(dummy_data, center = TRUE, scale = TRUE)
scaleList <- list(scale = attr(scaled_data, "scaled:scale"),
                  center = attr(scaled_data, "scaled:center"))

scaled_data <- as.data.frame(scaled_data)
# Create a basic statistic table using the quantitative variables of the dataset.
quant <- c("age", "bmi", "children", "charges")

summary_stat <- psych::describe(data[quant], skew = FALSE) %>%
  round(2)
```

```r
summary_stat %>%
  kable(
    caption = 'Basic statistics of quantitative variables',
    booktabs = TRUE,
    align = c('l', rep('c', 8))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 10,
    latex_options = "hold_position"
  )
plot(data)
ggplot(data) +
  aes(x = sex, y = bmi) +
  geom_boxplot(fill = "#112446") +
  theme_minimal()
ggplot(data) +
  aes(x = age, y = smoker) +
  geom_boxplot(fill = "#112446") +
  theme_minimal()
ggplot(data) +
  aes(x = age, y = bmi) +
  geom_jitter(size = 0.5) +
  theme_minimal()
ggplot(data) +
  aes(x = age, y = children) +
  geom_point(shape = "circle", size = 1.5, colour = "#112446") +
  theme_minimal()
# Doing a 8/2 training, testing splits
set.seed(123)
train_ind <- sample(1:nrow(dummy_data), floor(0.8*nrow(dummy_data)))
set.seed(NULL)

train <- dummy_data[train_ind, ]
test <- dummy_data[-train_ind, ]
scaled_train <- scaled_data[train_ind, ]
scaled_test <- scaled_data[-train_ind, ]
# Building the Multiple Linear Regression model
mlr_model <- lm(charges ~ ., data = scaled_train)

# Viewing the summary of the model
summary(mlr_model)

# Predicting with the test data
pred_mlr <- predict(mlr_model, newdata = scaled_test)

# Calculating Mean Squared Error (MSE) for the test set
actual <- scaled_test$charges * scaleList$scale["charges"] + scaleList$center["charges"]
pred <- pred_mlr * scaleList$scale["charges"] + scaleList$center["charges"]

MSE_mlr <- mean((actual - pred)^2)

# Displaying the MSE
```

```r
print(paste("MSE for Multiple Linear Regression: ", MSE_mlr))
# Diagnostic plot - Residuals vs Fitted values
plot(mlr_model$fitted.values, mlr_model$residuals)
abline(h = 0, col = "red")
# Diagnostic plot - QQ plot for residuals
qqnorm(mlr_model$residuals)
qqline(mlr_model$residuals, col = "red")
# Check for multicollinearity using Variance Inflation Factor (VIF)
mlr_model_1 <- lm(charges ~ . -is_northeast, data = train)
# Using alias to check for the coefficients
print(alias(mlr_model_1))
# Calculate VIF
vif(mlr_model_1)
# Diagnostic plot - Residuals vs Fitted values
plot(mlr_model$fitted.values, mlr_model$residuals)
abline(h = 0, col = "red")

# Scatter plot with studentized deleted residuals
h <- hatvalues(mlr_model)
sum <- summary(mlr_model)
sig2hat <- sum$sigma^2
r <- mlr_model$residuals/sqrt(sig2hat*(1-h))
n <- 1338
p <- 7
t <- r*((n-p-1)/(n-p-r^2))^(1/2)
plot(h, t)

# Large leverage points
plot(h,xlab='Observation',ylab='Leverage',main='Leverage')
abline(h=2*p/n,lty=2,col="red")
abline(h=3*p/n,lty=2,col="blue")

# Cook's distance
D = (1/p)*r^2*h/(1-h)
plot(h,D)
plot(mlr_model, which=5)

# Diagnostic plot - QQ plot for residuals
qqnorm(mlr_model$residuals)
qqline(mlr_model$residuals, col = "red")

# Shapiro - Wilk Test for normality
shapiro.test(mlr_model$residuals)
# Check for multicollinearity using Variance Inflation Factor (VIF)
mlr_model_1 <- lm(charges ~ . -is_northeast, data = train)

# Using alias to check for the coefficients
print(alias(mlr_model_1))

# Calculate VIF
vif(mlr_model_1)
Xmat <- model.matrix(charges ~ . , data=scaled_data)[ ,-1]
y <- scaled_data$charges
```

```r
set.seed(123)
train_ind <- sample(1:nrow(Xmat), floor(0.8*nrow(Xmat)))
set.seed(NULL)

X_mat_train <- Xmat[train_ind,]
X_mat_test <- Xmat[-train_ind,]
y_train <- y[train_ind]
y_test <- y[-train_ind]
set.seed (123)
cv.out <- cv.glmnet(x= X_mat_train, y = y_train,
                    alpha = 1, standardize = TRUE,
                    nfolds=10)
plot(cv.out)
bestlam <- cv.out$lambda.min
pred_lasso <- predict(cv.out, s = bestlam,
                      newx = X_mat_test)
coef_lasso <- predict(cv.out, s = bestlam,
                      type = "coefficients")

#actual <- scaled_test$charges * scaleList$scale["charges"] + scaleList$center["charges"]
pred <- pred_lasso * scaleList$scale["charges"] + scaleList$center["charges"]

RSS_lasso <- mean((actual - pred)^2)
RSS_lasso
coef_lasso
```