Project Problem 3

```
# Load libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

Load Medical Insurance Dataset

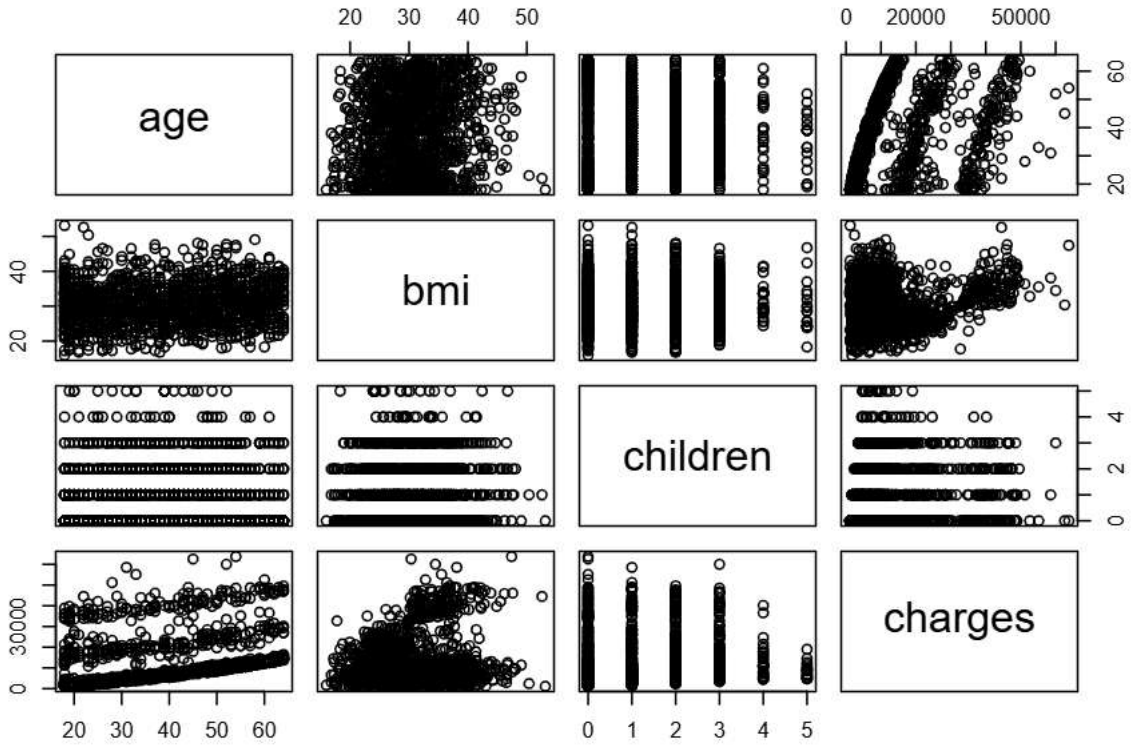
```
ins_df = read.csv("insurance.csv")
head(ins_df)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

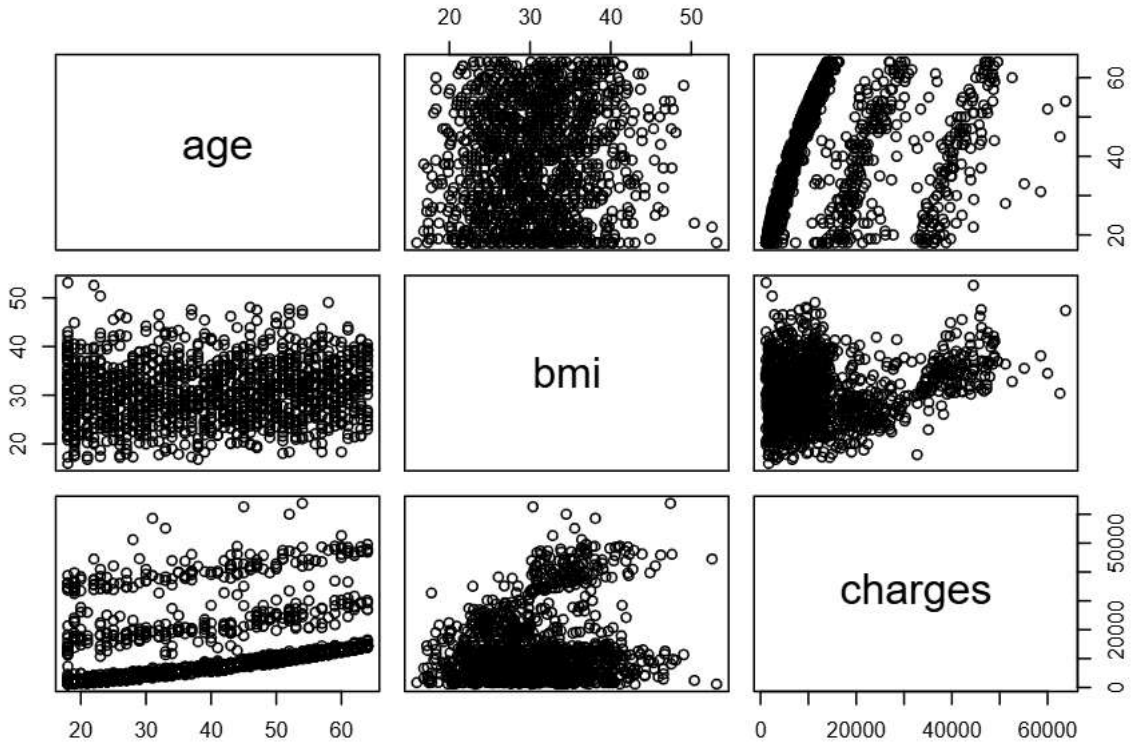
Pairs Plot

There appears to be a positive linear relationship between BMI and insurance charges.

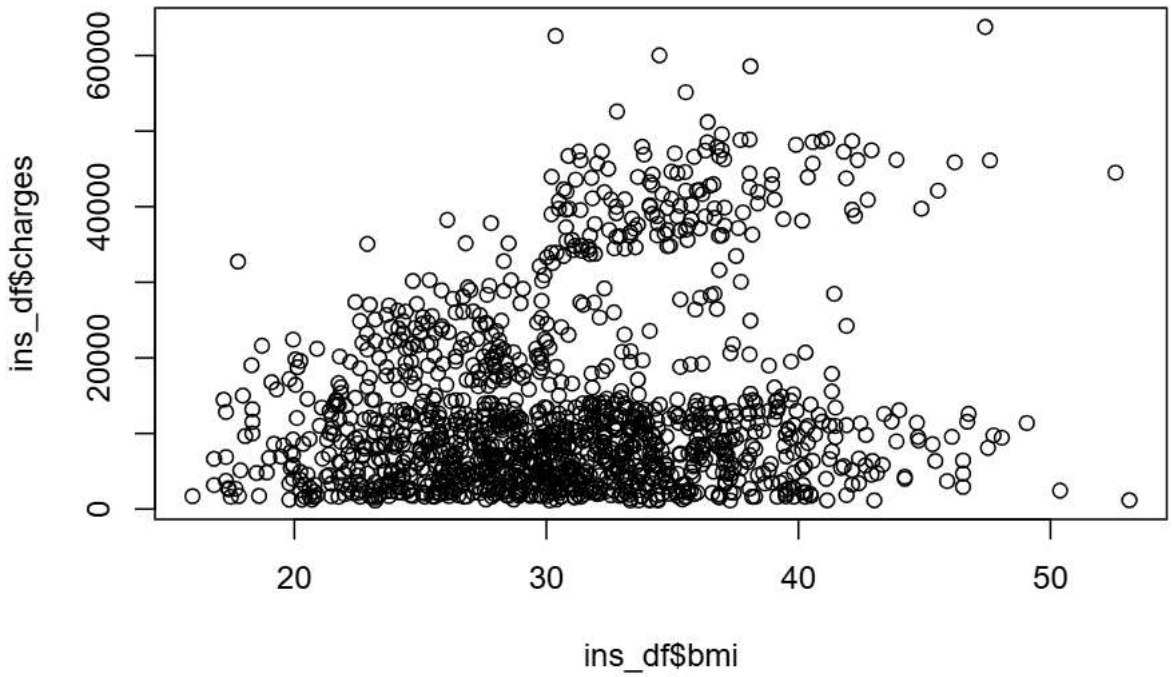
```
# Remove categorical variables
subset_ins_df <- subset(ins_df, select = -c(sex, region, smoker))
pairs(subset_ins_df) # Not very informative, further analysis needed (below)
```




```
# Remove discrete predictors
subset_ins_df <- subset(ins_df, select = -c(sex, region, smoker, children))
pairs(subset_ins_df) #Slightly more informative
```



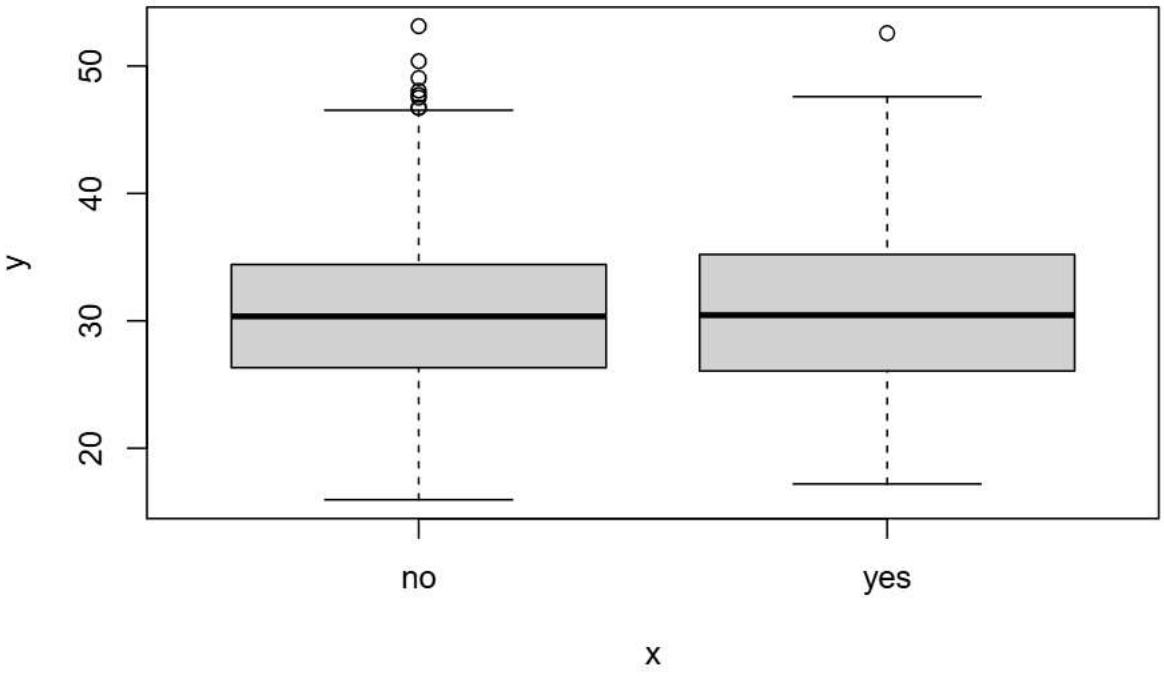

```
plot(ins_df$bmi, ins_df$charges)
```



###

Viewing the relationships between other variables There does not seem to be a relationship between smoking and BMI. The median BMI is about the same for both smokers and non-smokers.

```
ins_df$smoker = as.factor(ins_df$smoker)
plot(ins_df$smoker, ins_df$bmi)
```




```
# Chi square test between categorical variables
chisq.test(ins_df$sex, ins_df$smoker, correct=FALSE) # Appears to be a statistically significant relati
```

```
##
##  Pearson's Chi-squared test
##
## data:  ins_df$sex and ins_df$smoker
```

```
## X-squared = 7.7659, df = 1, p-value = 0.005324
```

```
# Chi square test between categorical variables
chisq.test(ins_df$smoker, ins_df$children, correct=FALSE) # Does not appear to be a statistically signi
```

```
## Warning in chisq.test(ins_df$smoker, ins_df$children, correct = FALSE):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  ins_df$smoker and ins_df$children
## X-squared = 6.8877, df = 5, p-value = 0.2291
```

```
# Chi square test between categorical variables
chisq.test(ins_df$sex, ins_df$children, correct=FALSE) # Does not appear to be a statistically signific
```

```
##
##  Pearson's Chi-squared test
##
## data:  ins_df$sex and ins_df$children
## X-squared = 0.73521, df = 5, p-value = 0.981
```

Checking for outliers in response variable (139 rows)

```
summary(ins_df$charges) # Q1 = 4740 Q3 = 16640
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```
IQR <- 16640 - 4740
lower_bound <- 4740 - (1.5 * IQR)
upper_bound <- 16640 + (1.5 * IQR)

ins_df %>% filter(charges > upper_bound | charges < lower_bound)
```

```
##     age    sex    bmi children smoker    region  charges
## 1    27   male 42.130        0    yes southeast 39611.76
## 2    30   male 35.300        0    yes southwest 36837.47
## 3    34 female 31.920        1    yes northeast 37701.88
## 4    31   male 36.300        2    yes southwest 38711.00
## 5    22   male 35.600        0    yes southwest 35585.58
## 6    28   male 36.400        1    yes southwest 51194.56
## 7    35   male 36.670        1    yes northeast 39774.28
## 8    60   male 39.900        0    yes southwest 48173.36
## 9    36   male 35.200        1    yes southeast 38709.18
## 10   36   male 34.430        0    yes southeast 37742.58
## 11   58   male 36.955        2    yes northwest 47496.49
## 12   22   male 37.620        1    yes southeast 37165.16
## 13   37 female 34.800        2    yes southwest 39836.52
## 14   57 female 31.160        0    yes northwest 43578.94
## 15   64 female 31.300        2    yes southwest 47291.06
## 16   63   male 35.090        0    yes southeast 47055.53
## 17   44   male 31.350        1    yes northeast 39556.49
## 18   46   male 30.495        3    yes northwest 40720.55
## 19   30   male 35.530        0    yes southeast 36950.26
## 20   18 female 36.850        0    yes southeast 36149.48
```