Can you accurately predict insurance costs?

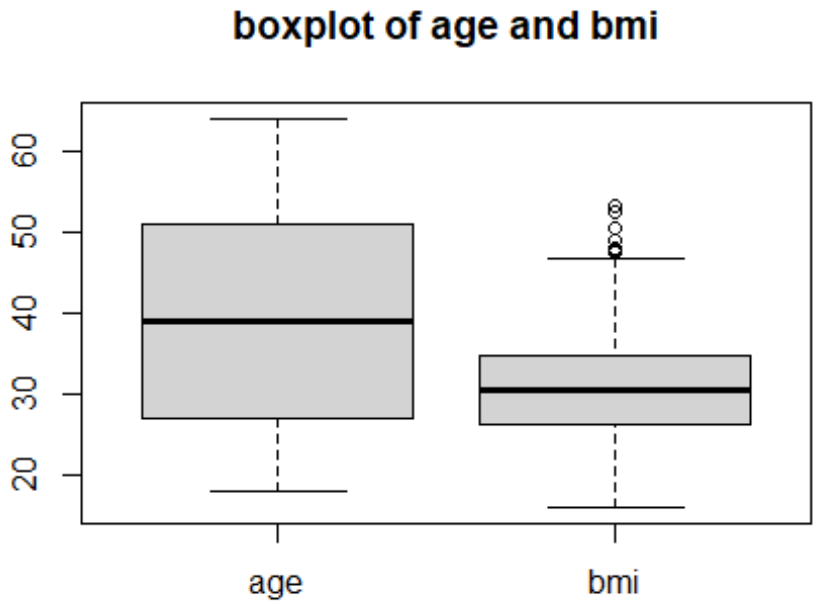
```
#importing data
library(readr)
insurancedata <- read_csv("C:/Users/Job/Downloads/insurance data
(1)/insurance data/insurance_data.csv")

## Rows: 1338 Columns: 7
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

head(insurancedata)

## # A tibble: 6 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.

#we are checking number of rows and columns in our dataset.
dim(insurancedata)

## [1] 1338    7

#checking the columns we have in our dataset
names(insurancedata)

## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"
"charges"

str(insurancedata)

## spc_tbl_ [1,338 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
```

```
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest"
...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

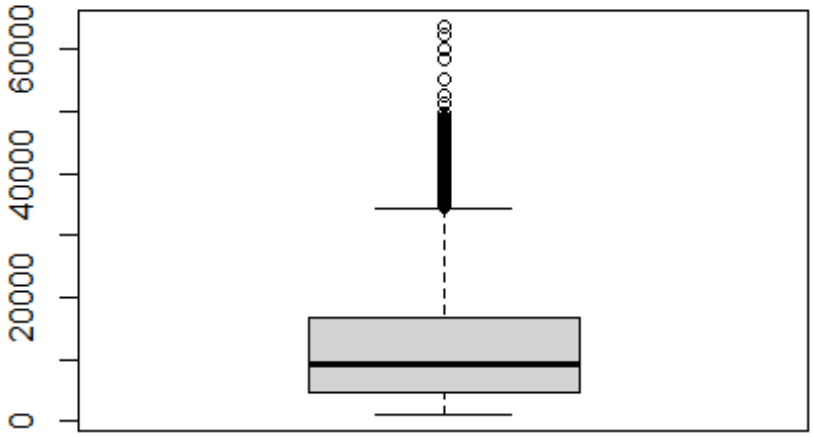
```
#we are checking missing values which is 0 in the case of our dataset
sum(is.na(insurancedata))
```

```
## [1] 0
```

```
#mainly we drew the boxplot  to check for the cases of outliers
name<-insurancedata$age
name1<-insurancedata$bmi
boxplot(name,name1,names=c("age","bmi"),main="boxplot of age and bmi")
```




```
boxplot(insurancedata$charges)
```




```r
unique(insurancedata$region)
```

```
## [1] "southwest" "southeast" "northwest" "northeast"
```

```r
summary(insurancedata)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```r
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ purrr     1.0.2
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
```

```
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ── Conflicts ─────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
require(plyr)
```

```
## Loading required package: plyr
```

```
## Warning: package 'plyr' was built under R version 4.3.2
```

```
## ------------------------------------------------------------------------
----
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first,
then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------
----
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following object is masked from 'package:purrr':
##
##     compact
```

```
insurance1<-insurancedata %>% select(sex,smoker,region) %>% filter(sex %in%
c("male","female")) %>% mutate(sex=recode(sex,male=1,female=0)) %>%
mutate(smoker=recode(smoker,yes =1 ,no=0)) %>%
mutate(region=recode(region,southwest=1,southeast=2 ,northwest=3
,northeast=4))
insurance1
```

```
## # A tibble: 1,338 × 3
##      sex smoker region
##    <dbl>  <dbl>  <dbl>
##  1     0      1      1
##  2     1      0      2
##  3     1      0      2
##  4     1      0      3
##  5     1      0      3
```

```
##  6     0      0      2
##  7     0      0      2
##  8     0      0      3
##  9     1      0      4
## 10     0      0      3
## # ℹ 1,328 more rows
```

```
insurance<-cbind(insurancedata,insurance1)
head(insurance)
```

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

```
head(insurance[,-c(2,5,6)])
```

```
##   age    bmi children   charges sex smoker region
## 1  19 27.900        0 16884.924   0      1      1
## 2  18 33.770        1  1725.552   1      0      2
## 3  28 33.000        3  4449.462   1      0      2
## 4  33 22.705        0 21984.471   1      0      3
## 5  32 28.880        0  3866.855   1      0      3
## 6  31 25.740        0  3756.622   0      0      2
```

```
#k-fold cross-validation,
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
control <- trainControl(method = "cv", number = 5)
model <- train(charges~age+sex+region+ bmi+children+smoker,data = insurance,
method = "lm", trControl = control)
model
```

```
## Linear Regression
##
## 1338 samples
##    6 predictor
##
```

```
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1070, 1070, 1070, 1071, 1071
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   6082.197  0.7488974  4204.662
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

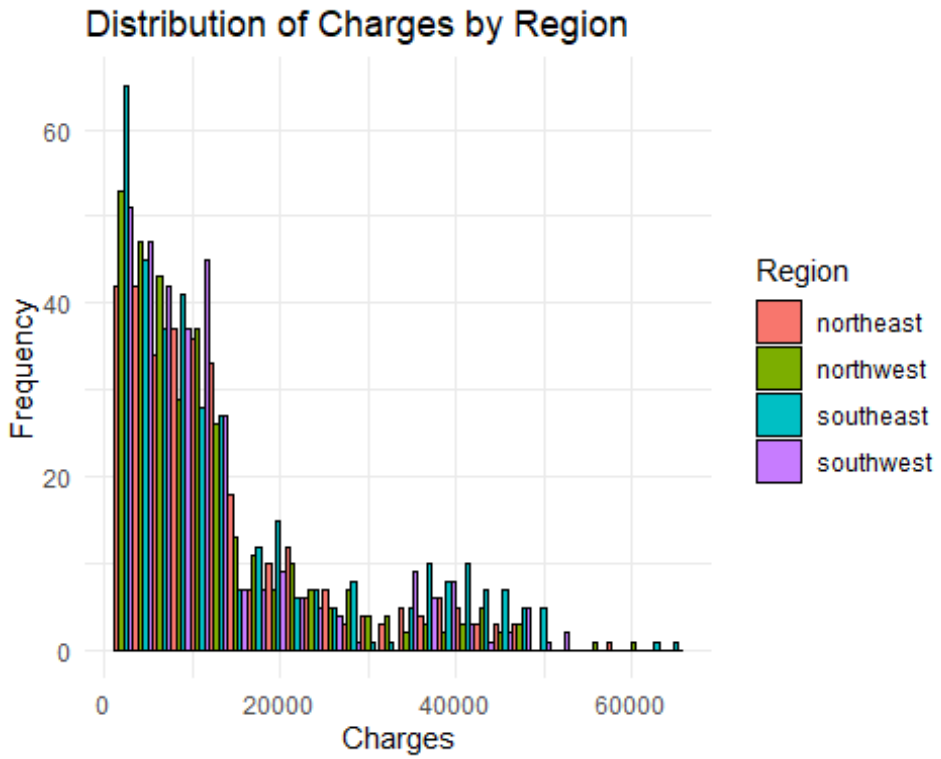
##RMSE (Root Mean Squared Error)the RMSE is 6082.311. This means that, on average, the model's predictions of the insurance cost are off by 6082.311. This value is relatively high, indicating that the model may not be accurately .
##he R-squared value is 0.7473437. This means that approximately 74.7% of the variance in the insurance cost can be explained by the independent variables in the linear regression model. This value indicates that the model has a moderate fit to the data, but there may be other factors that influence the insurance cost that are not captured by the independent variables in the model
##MAE is 4205.732. This means that, on average, the model's predictions of the insurance cost are off by 4205.732. This value is relatively high, indicating that the model may not be accurately capturing the relationship between the independent variables and the dependent variable.

```
summary(model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
```

```
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

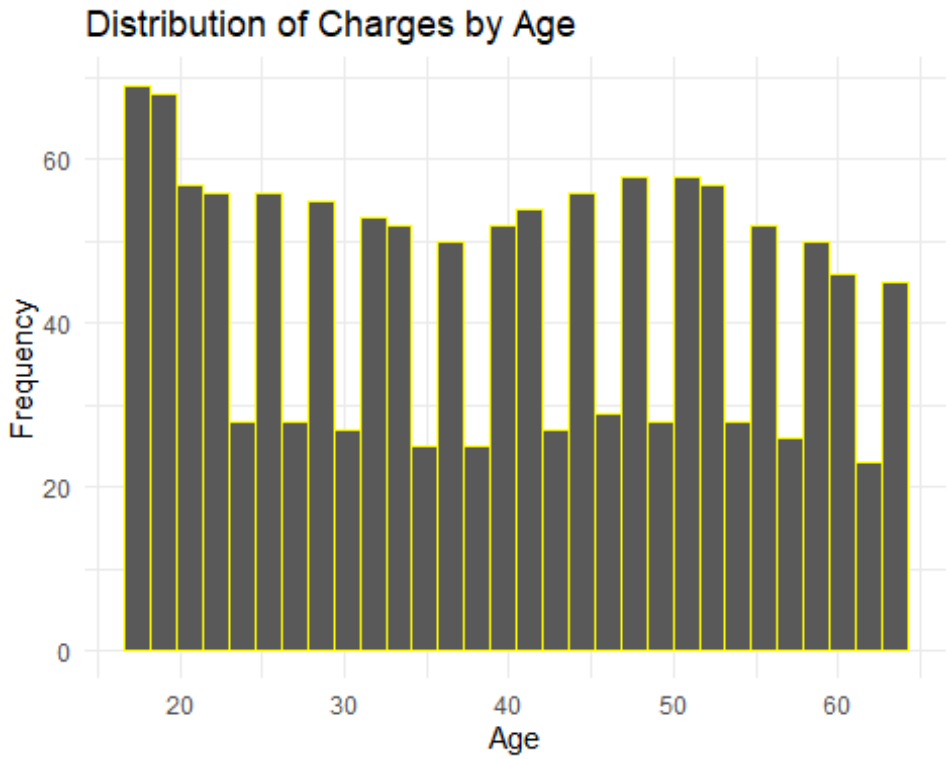
```
##Distribution of Charges by Region,we had the highest charges at southeast.
library(ggplot2)
library(ggplot2)
insurance <- insurance[, !duplicated(names(insurance))]
ggplot(insurance, aes(x = charges, fill = region)) +
  geom_histogram(position = "dodge", bins = 30, color = "black") +
  labs(title = "Distribution of Charges by Region", x = "Charges", y =
"Frequency", fill = "Region") +
  theme_minimal()
```



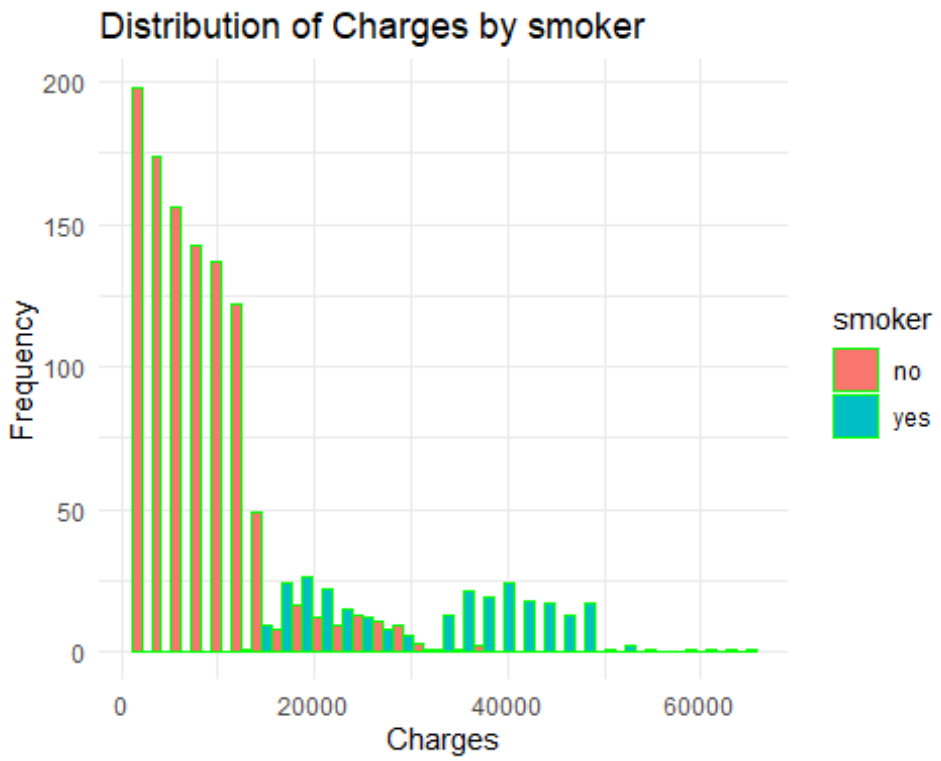

```
#Distribution of Charges by Age, we had the highest distribution of charges
at 20 year.
library(ggplot2)
insurance <- insurance[, !duplicated(names(insurance))]
ggplot(insurance, aes(x = age, fill = charges)) +
  geom_histogram(position = "dodge", bins = 30, color = "yellow") +
  labs(title = "Distribution of Charges by Age", x = "Age", y = "Frequency",
fill = "Charges") +
  theme_minimal()
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
```

```
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

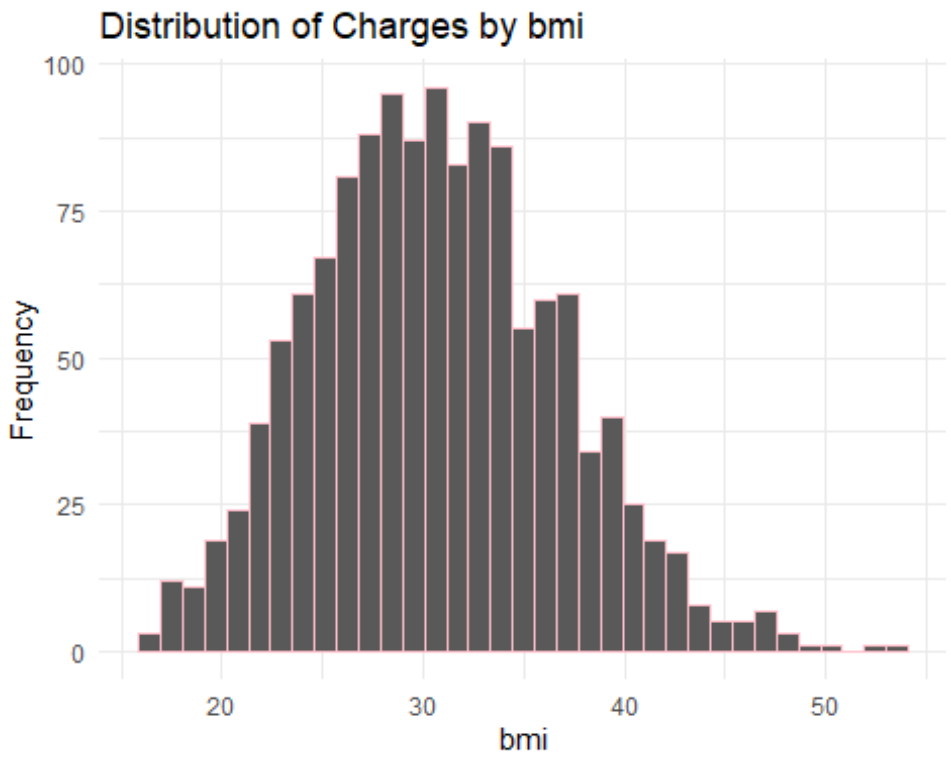



```
##those who did not smoke had the highest frequency of charges
library(ggplot2)
ggplot(insurance, aes(x = charges, fill = smoker)) +
  geom_histogram(position = "dodge", bins = 31, color = "green") +
  labs(title = "Distribution of Charges by smoker", x = "Charges", y =
"Frequency", fill = "smoker") +
  theme_minimal()
```

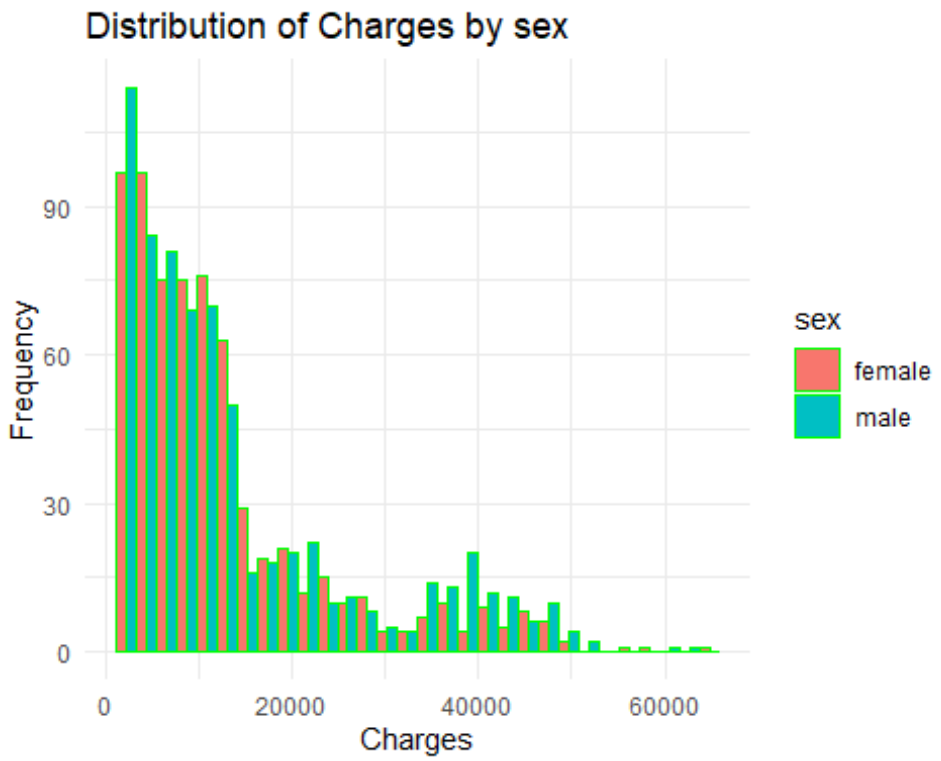



```
ggplot(insurance, aes(x =bmi , fill =charges)) +
  geom_histogram(position = "dodge", bins = 35, color = "pink") +
  labs(title = "Distribution of Charges by bmi", x = "bmi", y = "Frequency",
fill = "age") +
  theme_minimal()
```

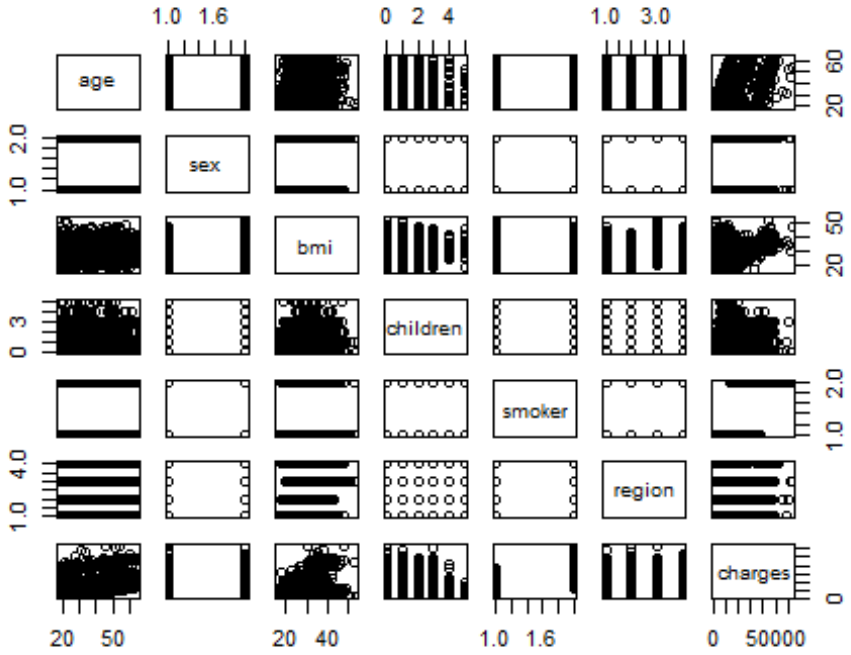
```
## Warning: The following aesthetics were dropped during statistical
transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
##male had the highest frequency of charges
ggplot(insurance, aes(x = charges, fill = sex)) +
  geom_histogram(position = "dodge", bins = 30, color = "green") +
  labs(title = "Distribution of Charges by sex", x = "Charges", y =
"Frequency", fill = "sex") +
  theme_minimal()
```




```
# Calculate the correlation coefficient between age and charges
cor(insurance$age, insurance$charges)

## [1] 0.2990082

##correlation between different variables of the data.
for (i in 1:ncol(insurance)) {
  ggplot(insurance, aes(x = charges, y = i)) +
    geom_point() +
    labs(title = paste("Charges vs.", names(insurance)[i]), x = "Charges", y
= names(insurance)[i])
}
plot(insurance )
```




```
#age and charges are weakly correlated also all other variables are weakly
correlated except children and charges  which are strongly correlated
indicating a strong relation.
correlation_matrix <- cor(insurance[, c("age", "charges","bmi","children")])
correlation_matrix
```

```
##                 age    charges       bmi   children
## age       1.0000000 0.29900819 0.1092719 0.04246900
## charges   0.2990082 1.00000000 0.1983410 0.06799823
## bmi       0.1092719 0.19834097 1.0000000 0.01275890
## children 0.0424690 0.06799823 0.0127589 1.00000000
```