Final Project: Insurance

By: Rebecca Nguyen
The purpose of this final project is to explore the health insurance data set and predict insurance charges based on several determinants.

Load appropriate packages/csv file

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
insurance_df <- read_csv('insurance.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.
```

Data Exploration

```
head(insurance_df)
```

```
## # A tibble: 6 x 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.
```

```
str(insurance_df)
```

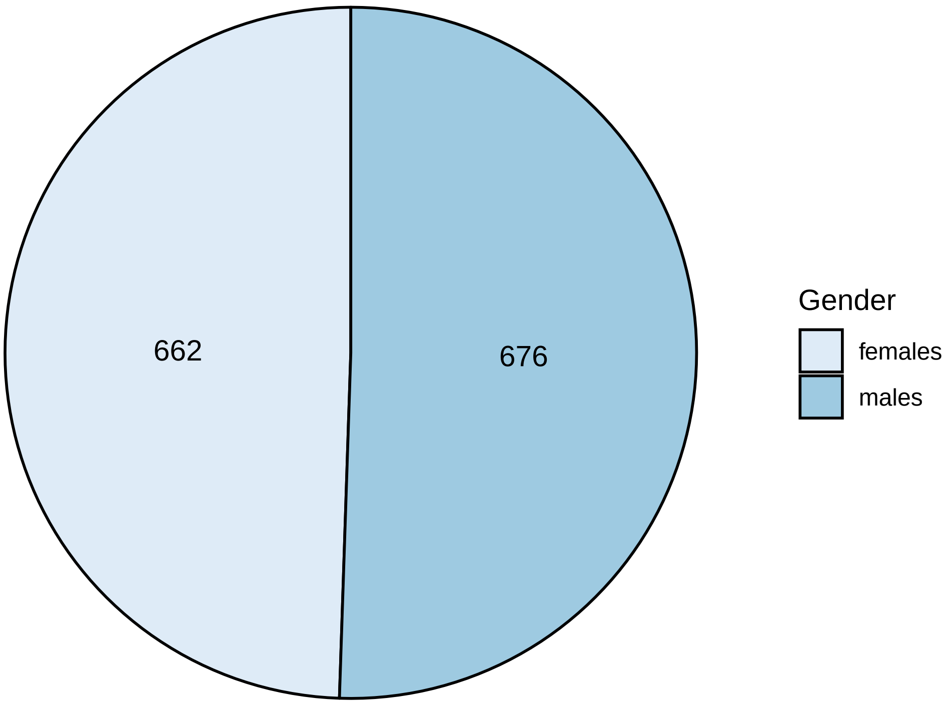
```
## spec_tbl_df [1,338 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ 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>
```

1338 observations
7 variables:
numerical: age, bmi, children, charges | character: sex, smoker, region

```
table(insurance_df$sex)
```

```
##
## female   male
##    662    676
```

```
gender_num <- data.frame(value = c(662, 676),
                        Gender = c("females", "males"))
ggplot(gender_num, aes(x = "", y = value, fill = Gender)) +
  geom_col(color = "black") +
  geom_text(aes(label = value),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_brewer() +
  theme_void()
```



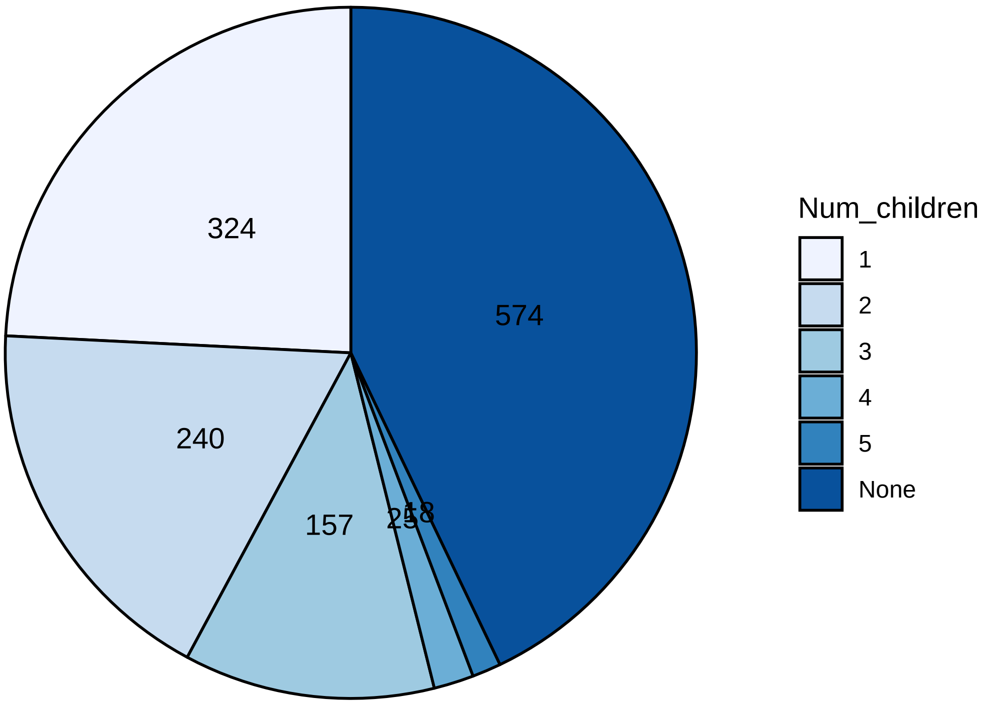
```
insurance_df %>%
  group_by(sex) %>%
  summarize(avg = mean(age, na.rm = TRUE), sd = sd(age, na.rm = TRUE))
```

```
## # A tibble: 2 x 3
##   sex      avg    sd
##   <chr>  <dbl> <dbl>
## 1 female  39.5  14.1
## 2 male    38.9  14.1
```

```
table(insurance_df$children)
```

```
##
##   0   1   2   3   4   5
## 574 324 240 157  25  18
```

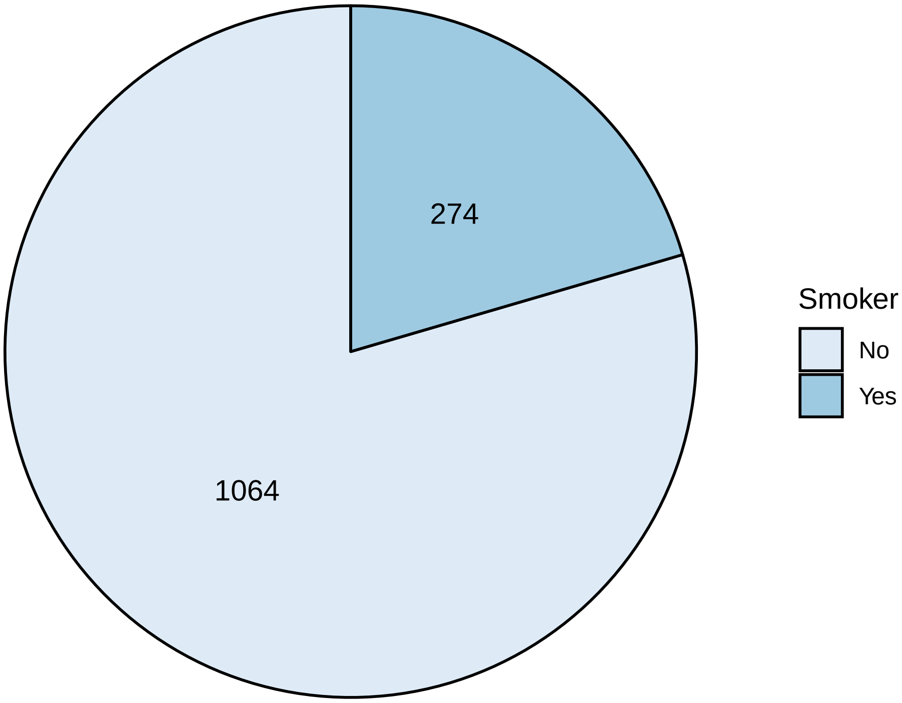
```
children_num <- data.frame(value = c(574, 324, 240, 157, 25, 18),
                           Num_children = c("None", 1, 2, 3, 4, 5))
ggplot(children_num, aes(x = "", y = value, fill = Num_children)) +
  geom_col(color = "black") +
  geom_text(aes(label = value),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_brewer() +
  theme_void()
```



```
table(insurance_df$smoker)
```

```
##
##   no  yes
## 1064  274
```

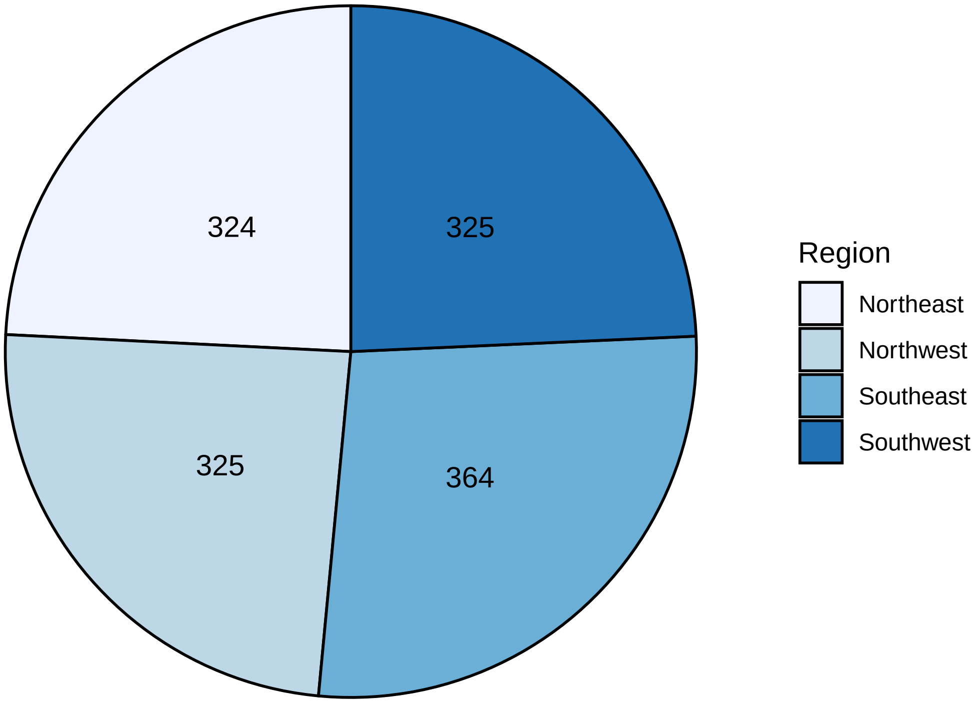
```
smoke_num <- data.frame(value = c(1064, 274),
                        Smoker = c("No", "Yes"))
ggplot(smoke_num, aes(x = "", y = value, fill = Smoker)) +
  geom_col(color = "black") +
  geom_text(aes(label = value),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_brewer() +
  theme_void()
```



```
table(insurance_df$region)
```

```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

```
region_num <- data.frame(value = c(324, 325, 364, 325),
                         Region = c("Northeast", "Northwest", "Southeast",
                                    "Southwest"))
ggplot(region_num, aes(x = "", y = value, fill = Region)) +
  geom_col(color = "black") +
  geom_text(aes(label = value),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_brewer() +
  theme_void()
```



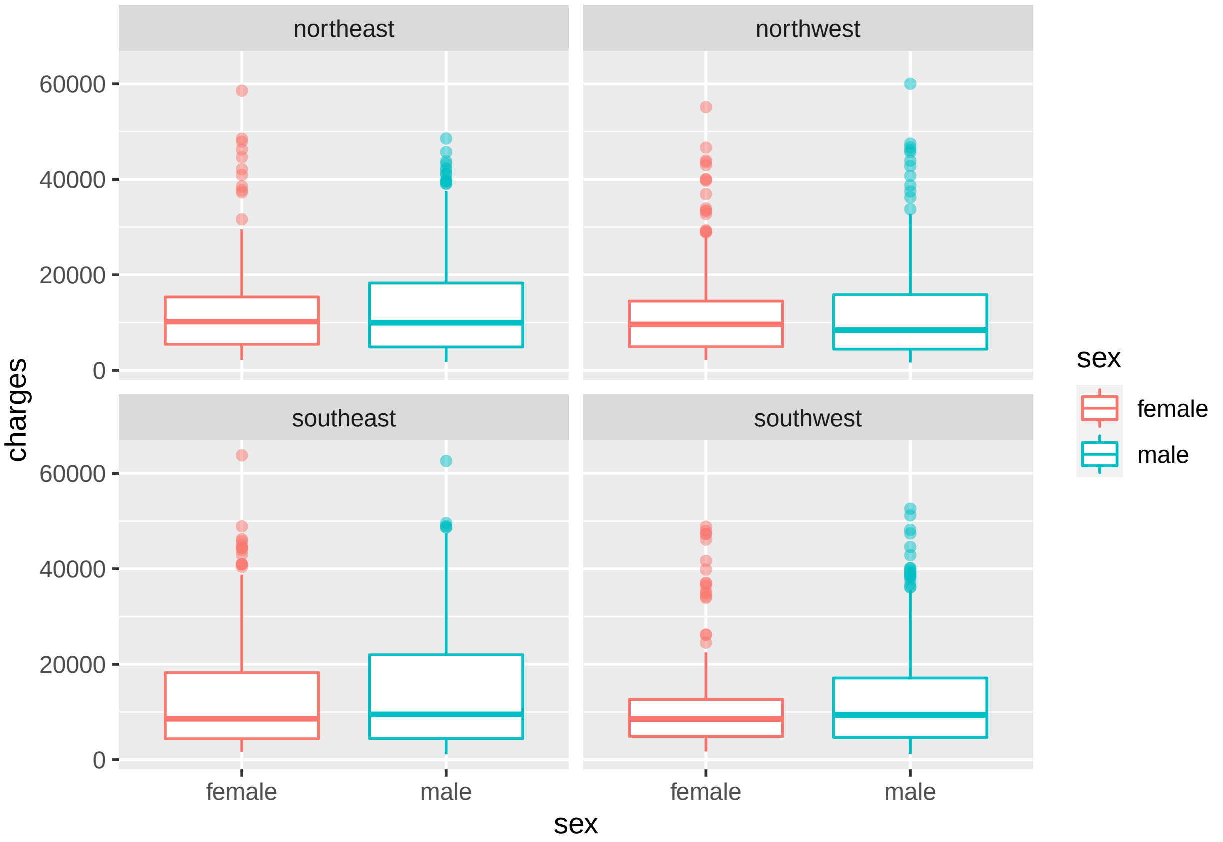
```
mean(insurance_df$charges)
```

```
## [1] 13270.42
```

The overall average insurance charge is $13,270.42

Sex vs. Charges Significance

```
sex_plot <- insurance_df %>%
  ggplot(aes(x = sex, y = charges, color = sex)) +
  facet_wrap(~ region) +
  geom_boxplot(outlier.alpha = 0.5)
sex_plot
```



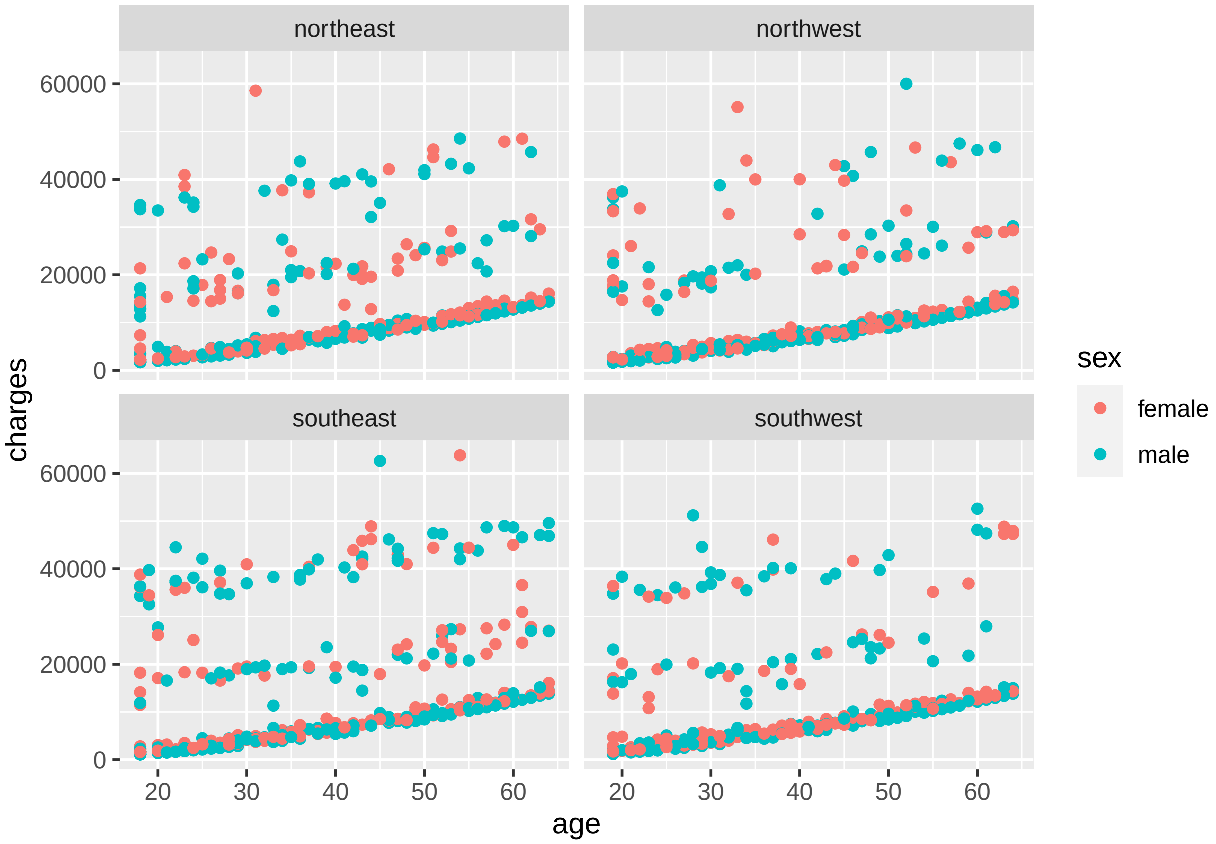
```
sex_sum <- insurance_df %>%
  group_by(sex) %>%
  summarize(avg = mean(charges, na.rm = TRUE), sd = sd(charges, na.rm = TRUE))
sex_sum
```

```
## # A tibble: 2 x 3
##   sex        avg     sd
##   <chr>    <dbl>  <dbl>
## 1 female 12570. 11129.
## 2 male   13957. 12971.
```

This histogram illustrates how there is no apparent bias between gender and insurance charges. Regardless of gender, the mean insurance charges are relatively similar as are the median charges per region.

Age vs. Charges Significance

```
age_plot <- insurance_df %>%
  ggplot(aes(x = age, y = charges, color = sex)) +
  facet_wrap(~ region) +
  geom_point()
age_plot
```



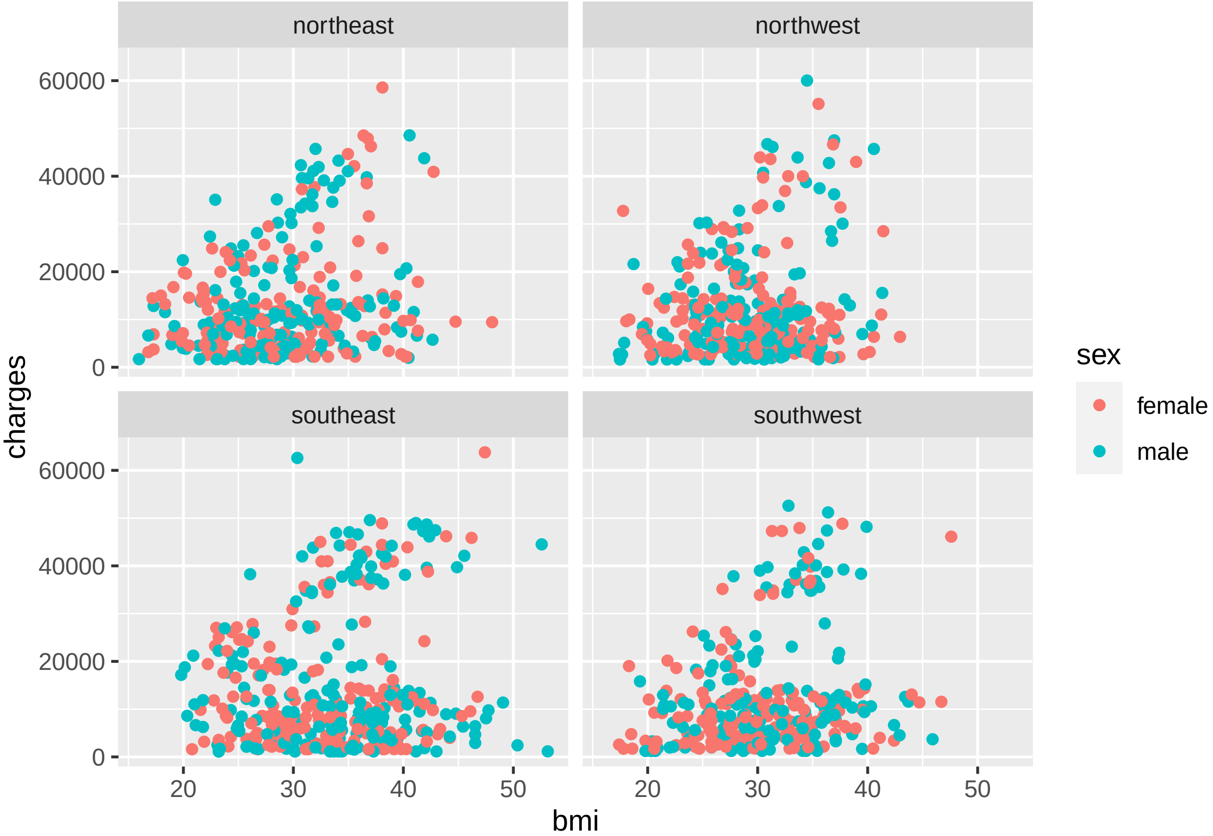
```
cor(insurance_df$age, insurance_df$charges)
```

```
## [1] 0.2990082
```

Based on the graphs, there seems to be a positive, linear relationship between age and insurance charges. As age increases, so does the health insurance. There also seems to be 2 distinct levels above the baseline, which require additional information, such as insurance type, to explore this significance. However, the correlation coefficient indicates a weak to moderate linear relation between age and insurance charges.

BMI vs. Charges Significance

```
BMI_plot <- insurance_df %>%
  ggplot(aes(x = bmi, y = charges, color = sex)) +
  facet_wrap(~ region) +
  geom_point()
BMI_plot
```



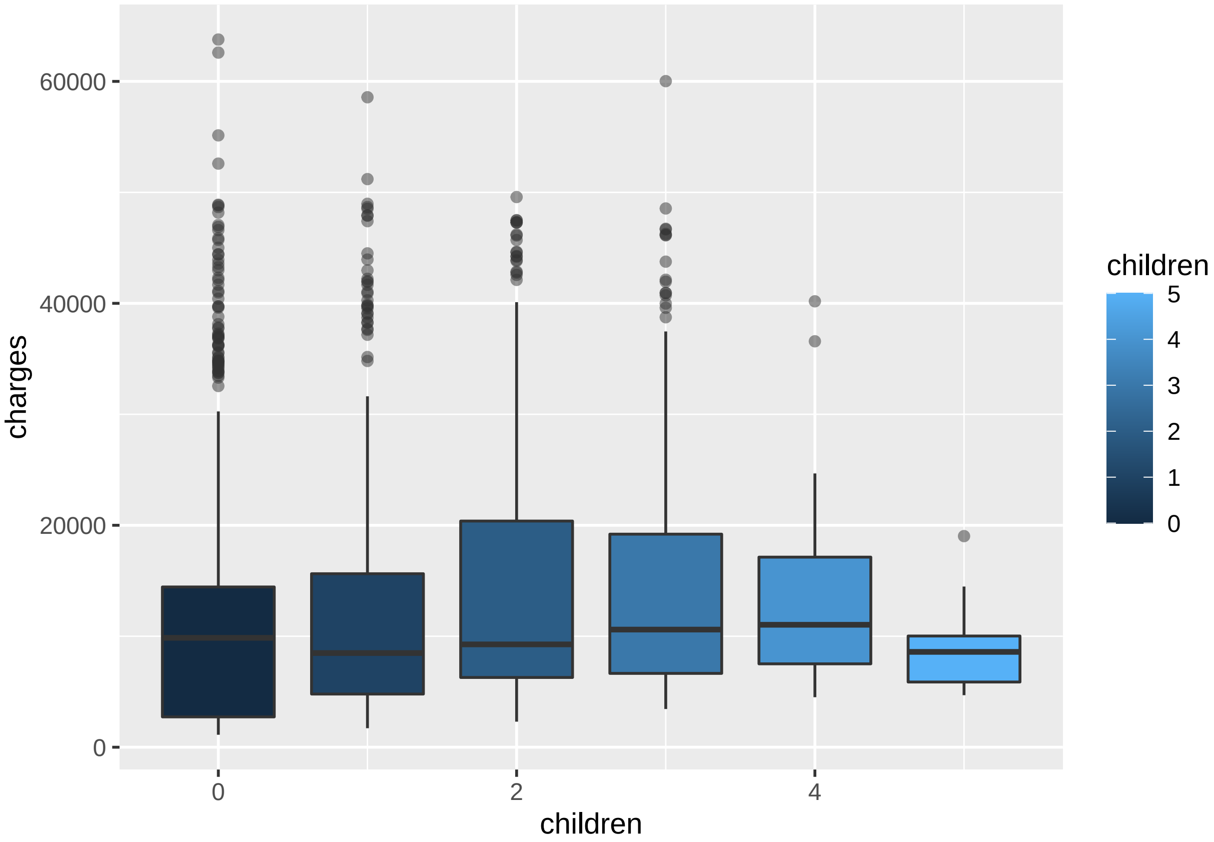
```
cor(insurance_df$bmi, insurance_df$charges)
```

```
## [1] 0.198341
```

Based on the graphs and correlation coefficient, there appears to be a weak, positive linear relationship between BMI and insurance charges. However, NE & NW look similar to each other as well as SE & SW.

Children vs. Charges Significance

```
child_plot <- insurance_df %>%
  ggplot(aes(x = children, y = charges, fill = children, group = children)) +
  geom_boxplot(outlier.alpha = 0.5)
child_plot
```



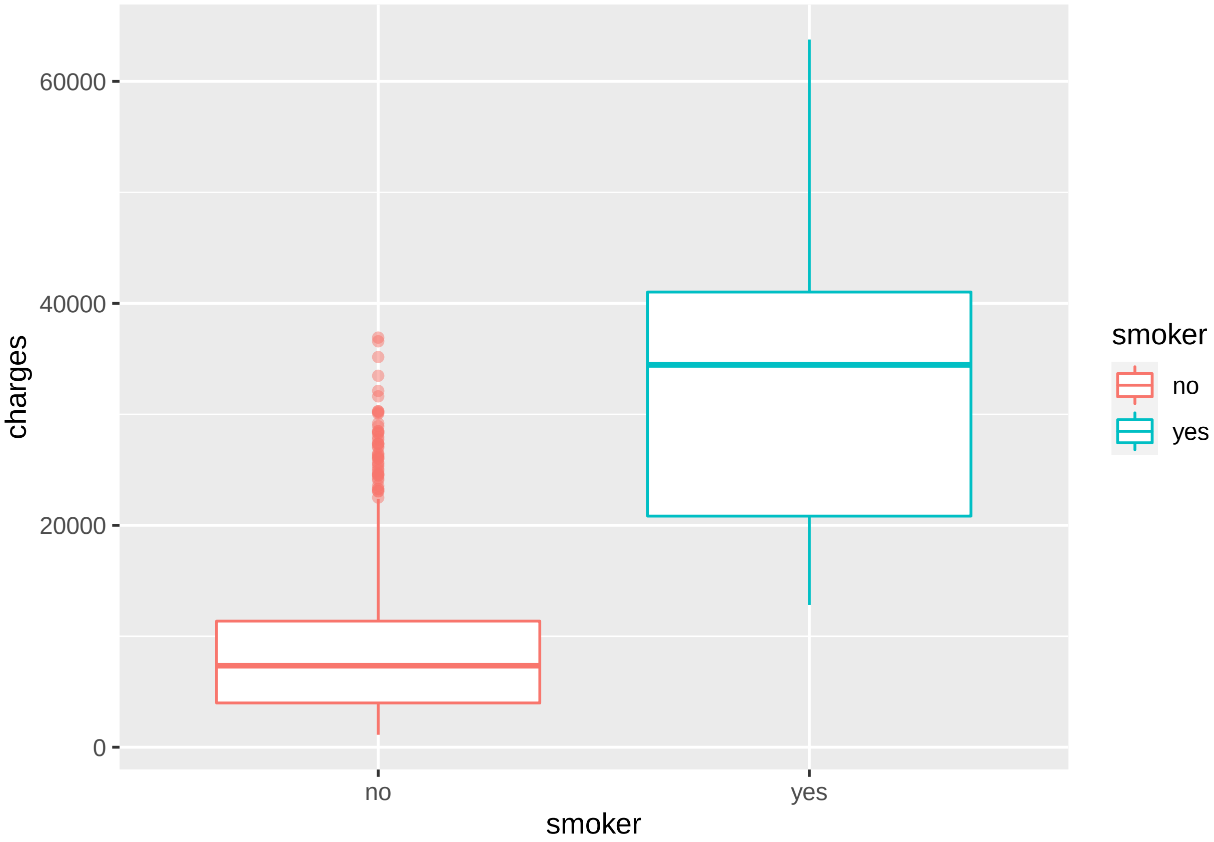
```
cor(insurance_df$children, insurance_df$charges)
```

```
## [1] 0.06799823
```

Based on the graph & correlation coefficient, there is no correlation between the number of offsprings & insurance charges.

Smoker vs. Charges Significance

```
smoke_plot <- insurance_df %>%
  ggplot(aes(x = smoker, y = charges, color = smoker)) +
  geom_boxplot(outlier.alpha = 0.5)
smoke_plot
```

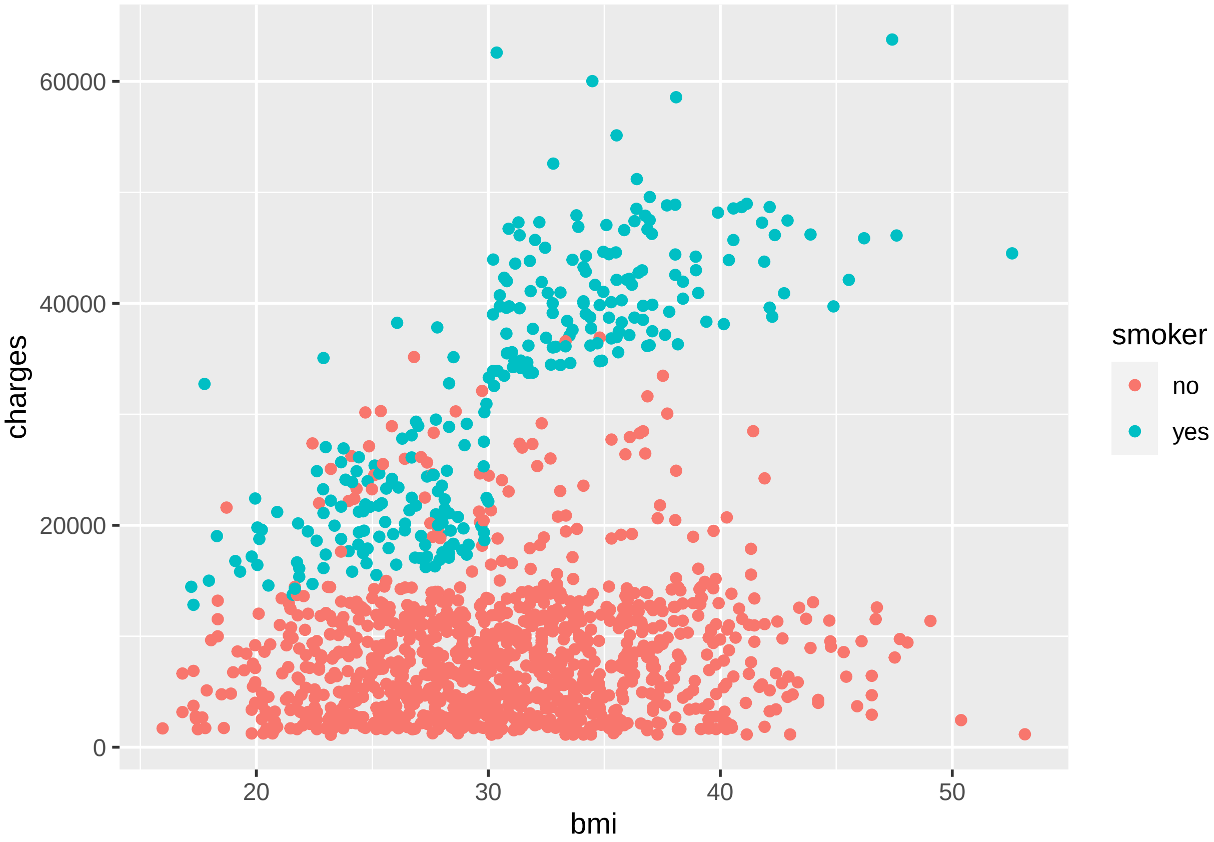


```
smoke_sum <- insurance_df %>%
  group_by(smoker) %>%
  summarize(avg = mean(charges, na.rm = TRUE), sd = sd(charges, na.rm = TRUE))
smoke_sum
```

```
## # A tibble: 2 x 3
##   smoker    avg     sd
##   <chr>   <dbl>  <dbl>
## 1 no      8434.  5994.
## 2 yes    32050. 11542.
```

This graph illustrates a startling, significant difference between smoker status and health insurance charges. A smoker is expected to pay on average $32,050, whereas a non-smoker is expected to pay on average $8,434. That is a $23,616 difference!

```
smokebmi_plot <- insurance_df %>%
  ggplot(aes(x = bmi, y = charges, color = smoker)) +
  geom_point()
smokebmi_plot
```



This graph indicates a positive, moderate linear relation between being a smoker as well as obese (> 30 BMI) and health insurance charges. Health insurance increases as the BMI of a smoker increases. It is apparent that the base charges of smokers are higher than their counterparts.

```
bmismc <- insurance_df %>%
  filter(smoker == "yes") %>%
  select(bmi, charges)
bmismc
```

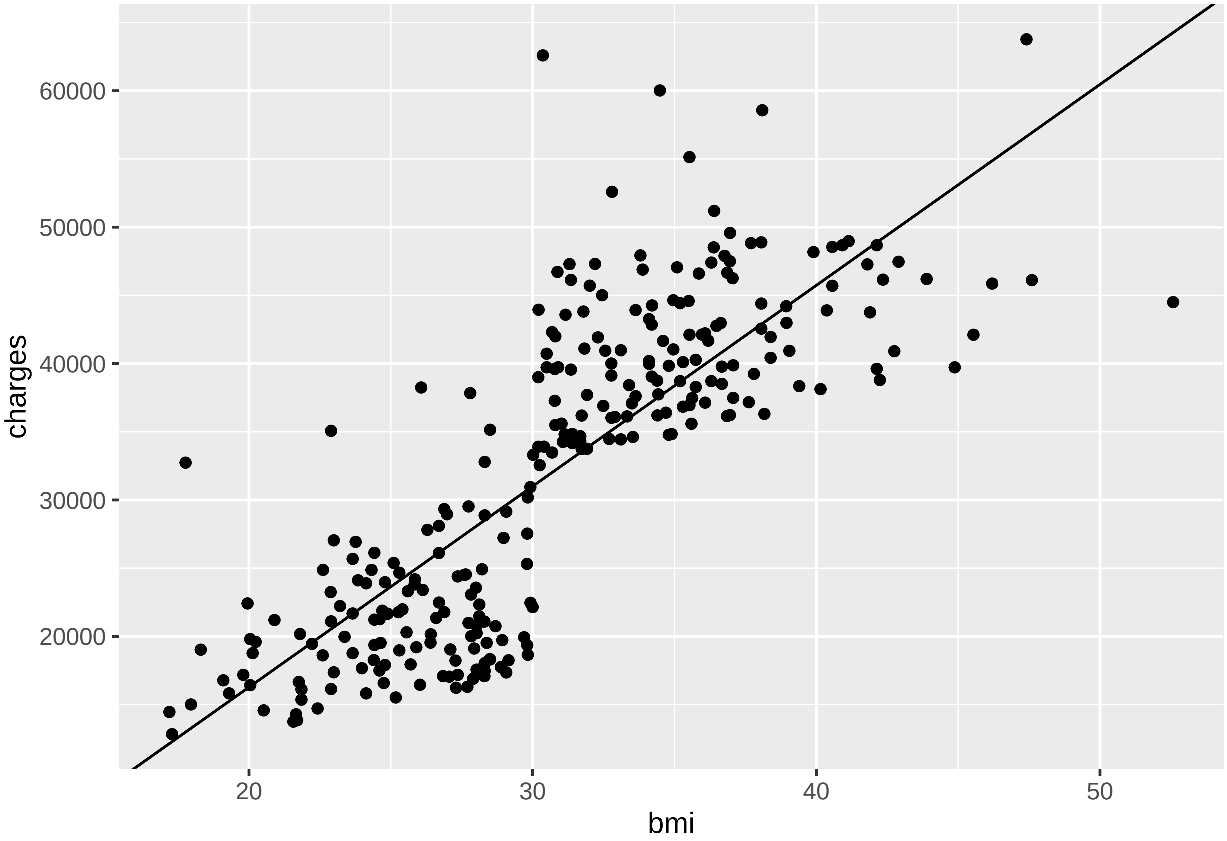
```
## # A tibble: 274 x 2
##      bmi charges
##    <dbl>   <dbl>
##  1  27.9  16885.
##  2  26.3  27809.
##  3  42.1  39612.
##  4  35.3  36837.
##  5  31.9  37702.
##  6  36.3  38711
##  7  35.6  35586.
##  8  36.4  51195.
##  9  36.7  39774.
## 10  39.9  48173.
## # ... with 264 more rows
```

```
lm(charges~bmi, bmismc)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = bmismc)
```

```
##
## Coefficients:
## (Intercept)          bmi
##      -13187         1473
```

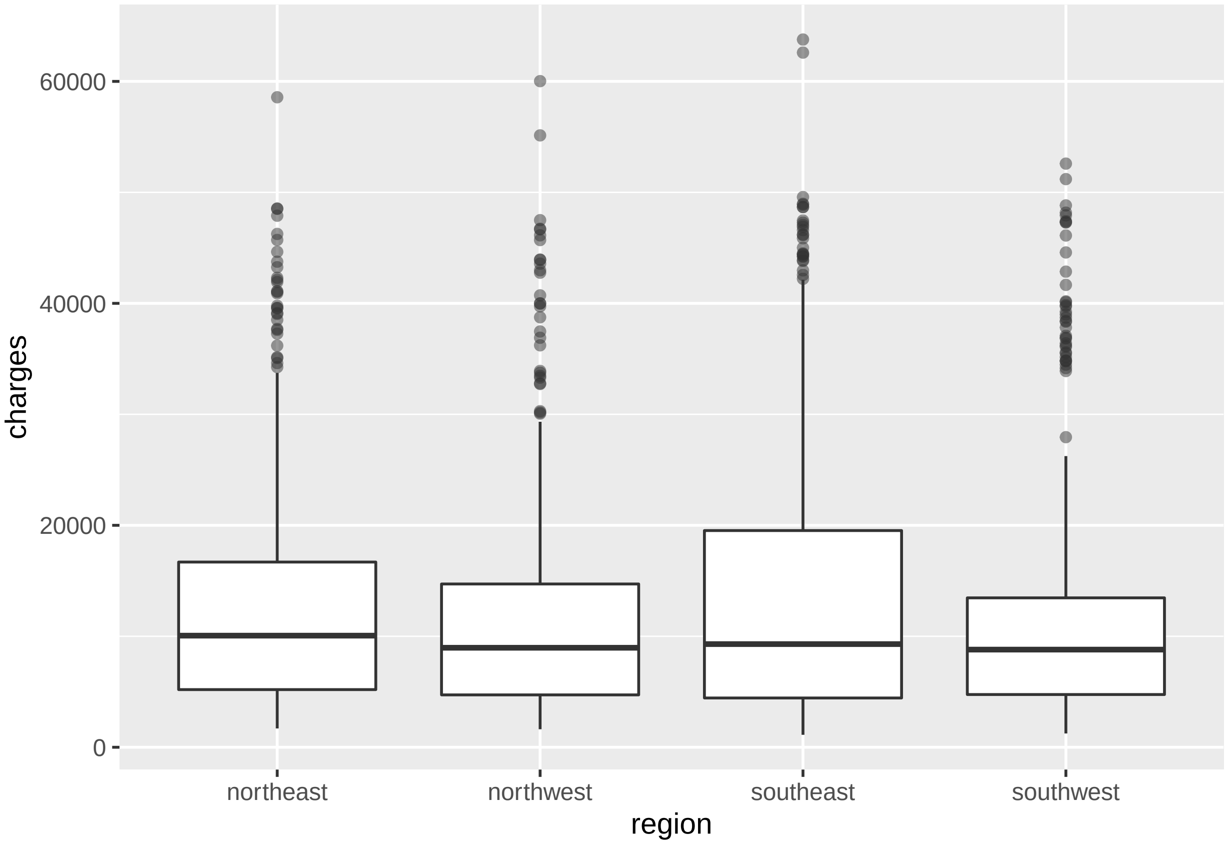
```
lm_bmismcplot <- bmismc %>%
  ggplot(aes(bmi, charges)) +
  geom_point() +
  geom_abline(intercept = -13187, slope = 1473)
lm_bmismcplot
```



As a smoker by default, one unit increase in BMI results in 1473 increase in charges.

Region vs. Charges Significance

```
region_plot <- insurance_df %>%
  ggplot(aes(x = region, y = charges)) +
  geom_boxplot(outlier.alpha = 0.5)
region_plot
```



```
region_sum <- insurance_df %>%
  group_by(region) %>%
  summarize(avg = mean(charges, na.rm = TRUE), sd = sd(charges, na.rm = TRUE))
region_sum
```

```
## # A tibble: 4 x 3
##   region       avg     sd
##   <chr>      <dbl>  <dbl>
## 1 northeast 13406. 11256.
## 2 northwest 12418. 11072.
## 3 southeast 14735. 13971.
## 4 southwest 12347. 11557.
```

Based on the box plot and summary, there is very weak relation between region and insurance. Regardless of location, the mean & median insurance charges are relatively similar. However, the east seems to charge a bit more than the west. The southeast being the most expensive.

From this data exploration, we can conclude that age and BMI alongside smoker play a significant role in health insurance charges. Sex, children, and region do not significantly affect health insurance charges.

Linear Regression Model

```
linear_model_insurance <- lm(charges~., insurance_df)
linear_model_insurance
```

```
##
```

```
## Call:
## lm(formula = charges ~ ., data = insurance_df)
##
## Coefficients:
##     (Intercept)              age          sexmale              bmi
##        -11938.5            256.9           -131.3            339.2
##        children        smokeryes  regionnorthwest  regionsoutheast
##           475.5          23848.5           -353.0          -1035.0
## regionsouthwest
##          -960.1
```

```
summary(linear_model_insurance)
```

```
##
## Call:
## lm(formula = charges ~ ., data = insurance_df)
##
## 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
## 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 ***
## 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 *
## ---
## 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
```

Based on the summary, the most significant factors are age, BMI, children, & smoker. The least significant is the regions. Sex holds no significance in relation to charges. The p-value for sexmale is big, so let's take it out

```
linear_model_insurance2 <- lm(charges~.-sex, insurance_df)
linear_model_insurance2
```

```
##
## Call:
## lm(formula = charges ~ . - sex, data = insurance_df)
##
## Coefficients:
##     (Intercept)              age              bmi         children
##        -11990.3            257.0            338.7            474.6
##       smokeryes  regionnorthwest  regionsoutheast  regionsouthwest
##         23836.3           -352.2          -1034.4           -959.4
```

```
summary(linear_model_insurance2)
```

```
##
## Call:
## lm(formula = charges ~ . - sex, data = insurance_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
## age                256.97      11.89  21.610  < 2e-16 ***
## bmi                338.66      28.56  11.858  < 2e-16 ***
## children           474.57     137.74   3.445 0.000588 ***
## smokeryes        23836.30     411.86  57.875  < 2e-16 ***
## regionnorthwest   -352.18     476.12  -0.740 0.459618
## regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
## regionsouthwest   -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

The multiple R-squared: 0.7509 tells us this linear regression model decently explains the data.

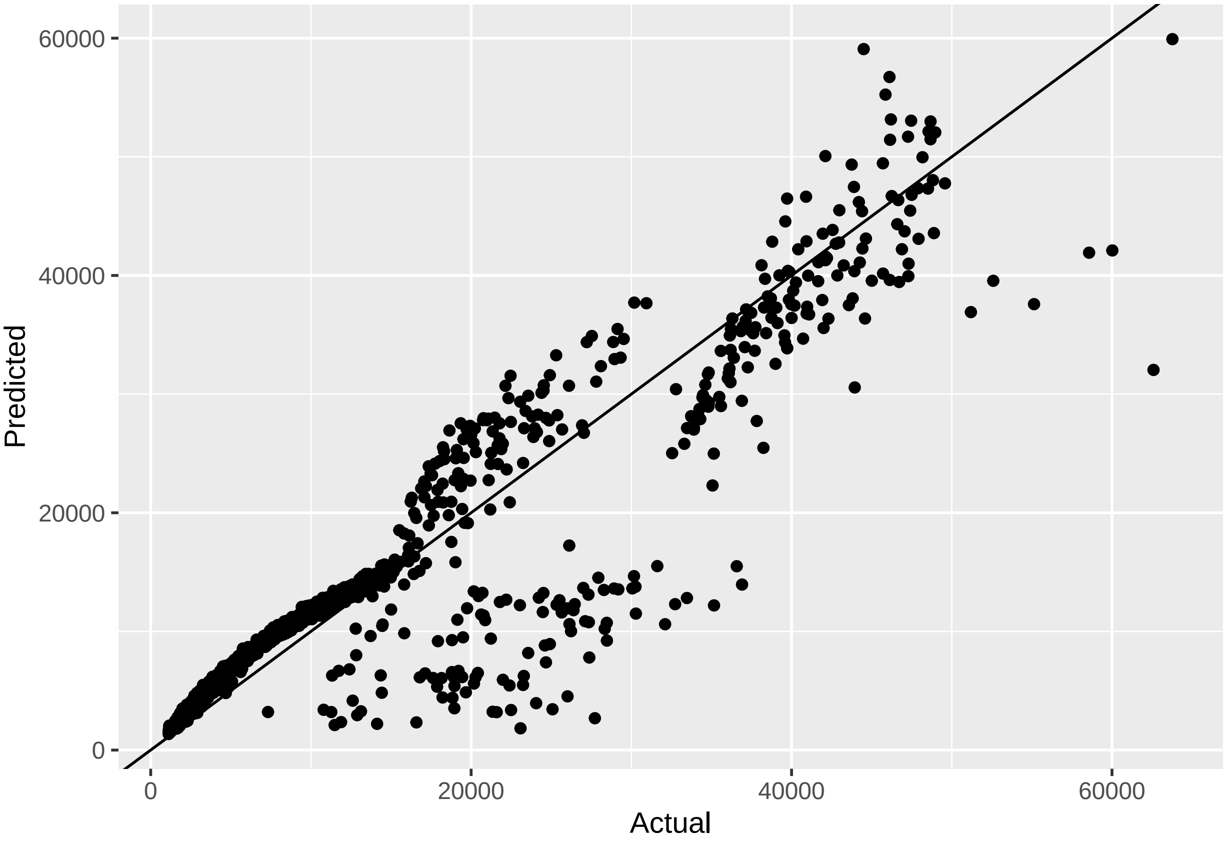
```
lm_test <- lm(charges ~.+bmi*smoker, insurance_df)
summary(lm_test)
```

```
##
## Call:
## lm(formula = charges ~ . + bmi * smoker, data = insurance_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14580.7  -1857.2  -1360.8   -475.7  30552.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2223.454    865.611  -2.569  0.01032 *
## age                263.620      9.516  27.703  < 2e-16 ***
## sexmale           -500.146    266.518  -1.877  0.06079 .
## bmi                 23.533     25.601   0.919  0.35814
## children           516.403    110.179   4.687 3.06e-06 ***
## smokeryes       -20415.611   1648.277 -12.386  < 2e-16 ***
## regionnorthwest   -585.478    380.859  -1.537  0.12447
## regionsoutheast  -1210.131    382.750  -3.162  0.00160 **
## regionsouthwest  -1231.108    382.218  -3.221  0.00131 **
## bmi:smokeryes     1443.096     52.647  27.411  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4846 on 1328 degrees of freedom
## Multiple R-squared:  0.8409, Adjusted R-squared:  0.8398
## F-statistic:   780 on 9 and 1328 DF,  p-value: < 2.2e-16
```

By including the synergistic effect of BMI & smoker to LM, the new multiple R-squared increased to 0.840, which directly translates to a better model.

```
lm_predict <- predict(lm_test, insurance_df)
test <- data.frame("Predicted" = lm_predict, "Actual" = insurance_df$charges)
ggplot(test, aes(Actual, Predicted)) + geom_point() + geom_abline()
```



After adjusting the LM, we can see the model being the best fit for the data.

Prediction Application

```
age = 60
sex = "female"
bmi = 35
children = 2
smoker = "yes"
region = "northeast"
single_obs <- data.frame(age = 60,
                         sex = "female",
                         bmi = 35,
                         children = 2,
                         smoker = "yes",
                         region = "northeast")
single_obs
```

```
##   age    sex bmi children smoker    region
## 1  60 female  35        2    yes northeast
```

```
predict(lm_test, single_obs)
```

```
##        1
## 45542.98
```

$45,503.18 is our prediction.