Statistical Analysis of Health Charges

Julia Sheriff

9/12/2018

An Overview of the Dataset

Health Variables:

Variable	Description
Age	individual's age in years
Sex	insurance contractor gender: female, male
BMI	Body mass index: weight in kg / heght in m^2
BMI_factor	Categories of BMI values: underweight, healthy weight, overweight, obese
Children	Number of children covered by health insurance, Number of dependents
Smoker	Smoker or Non-smoker
Region	Beneficiary's US residental area: northeast, southeast, northwest, southwest
Charges	Individual medical costs billed by health insurance

```r
health_charges_clean <- read.csv("health_charges_clean.csv", header=TRUE)

head(health_charges_clean)
```

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

Single Variable Analysis

An overview of each variable with anecdotal notes

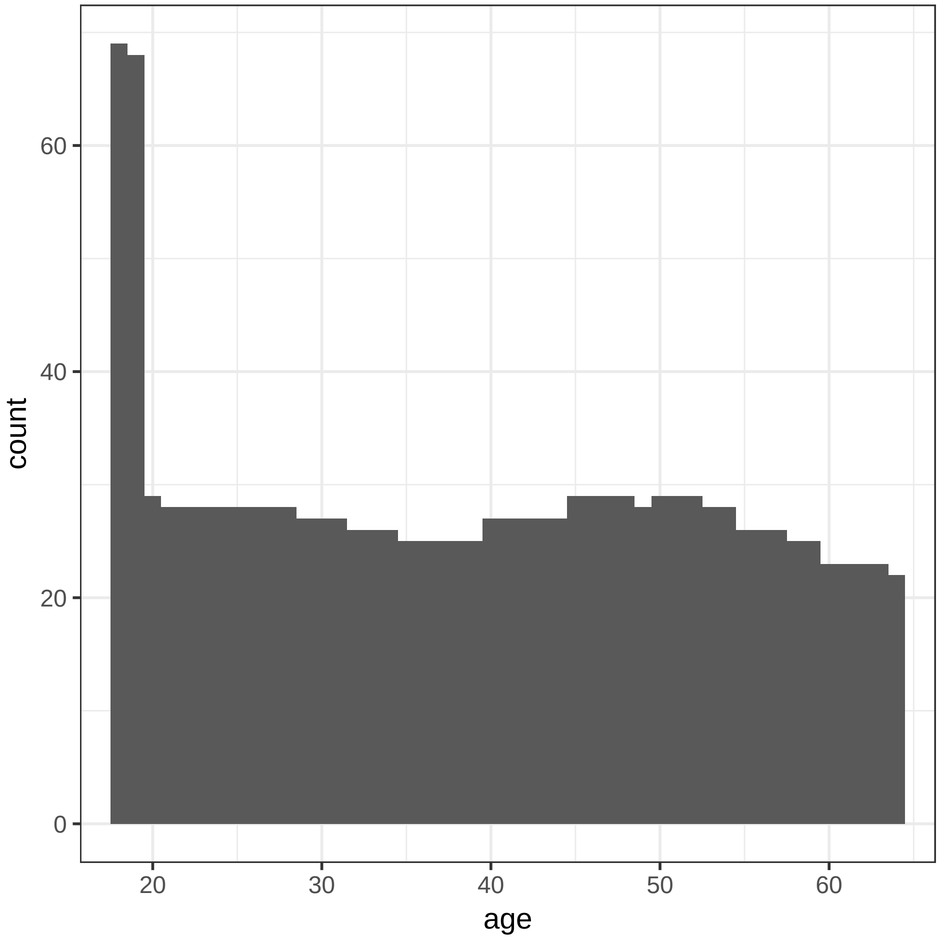
```
library(ggplot2)
library(dplyr)
```

```
##
## 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(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

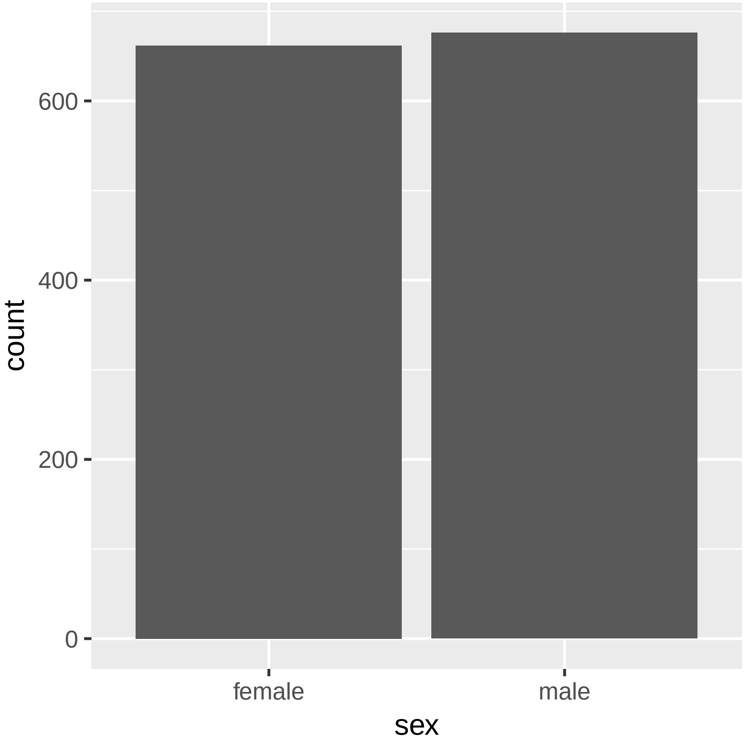
```
ggplot(health_charges_clean, aes(age))+
  geom_histogram(binwidth = 1)+
  coord_cartesian(xlim = c(18, 64))+
  theme_bw()
```



Age

- Disporportionately high number of 18-19 ages;
- Otherwise, even age distribution.

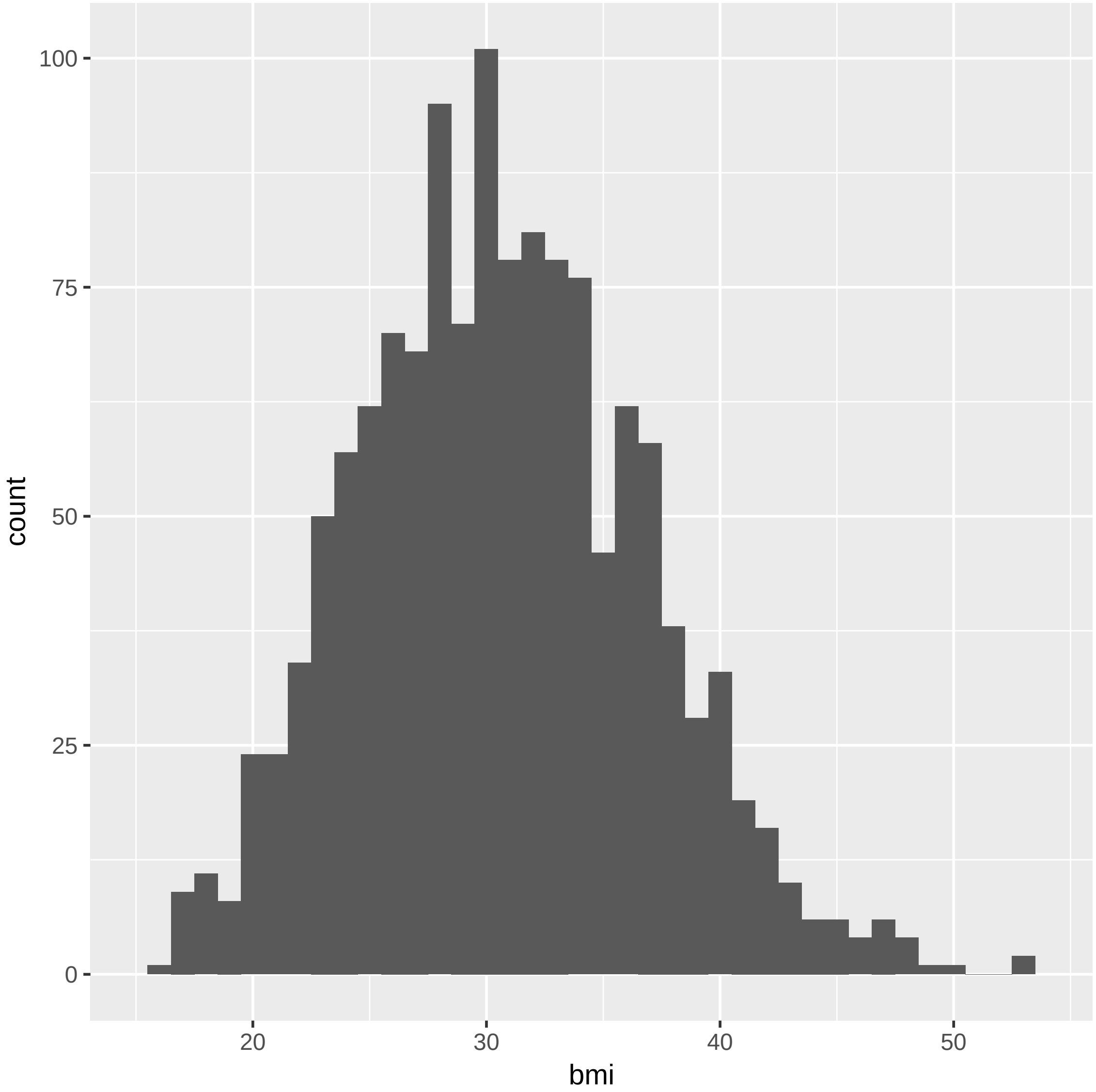
```r
ggplot(health_charges_clean, aes(sex))+
  geom_bar()
```



Sexes

- Even distribution

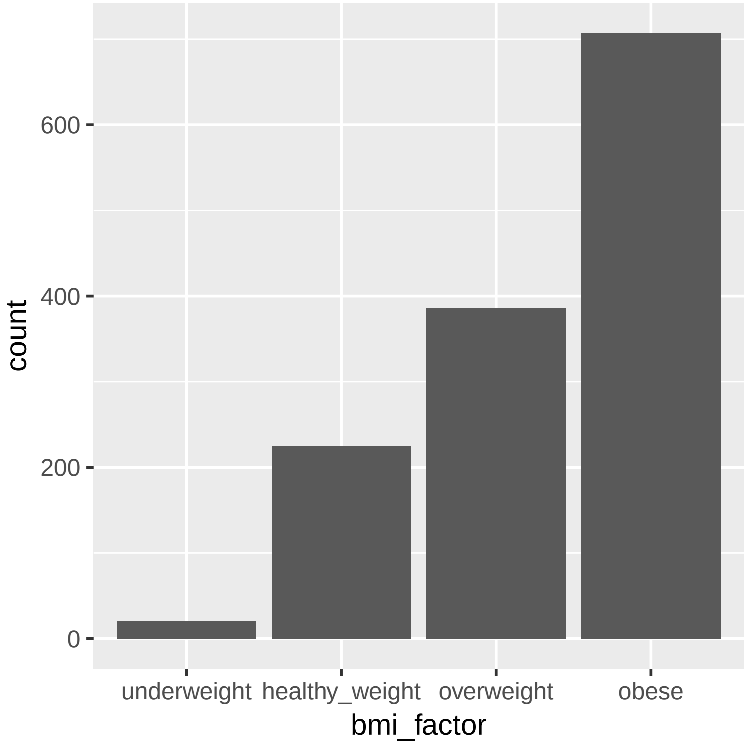
```r
ggplot(health_charges_clean, aes(bmi)) +
  geom_histogram(binwidth = 1) +
  coord_cartesian(xlim = c(15, 54))
```



BMI

- Normal distribution
- The mean of the data is approximately at the border of overweight and obese.
- The number of obese observations is approximately equal to the sum of the non-obese observations.

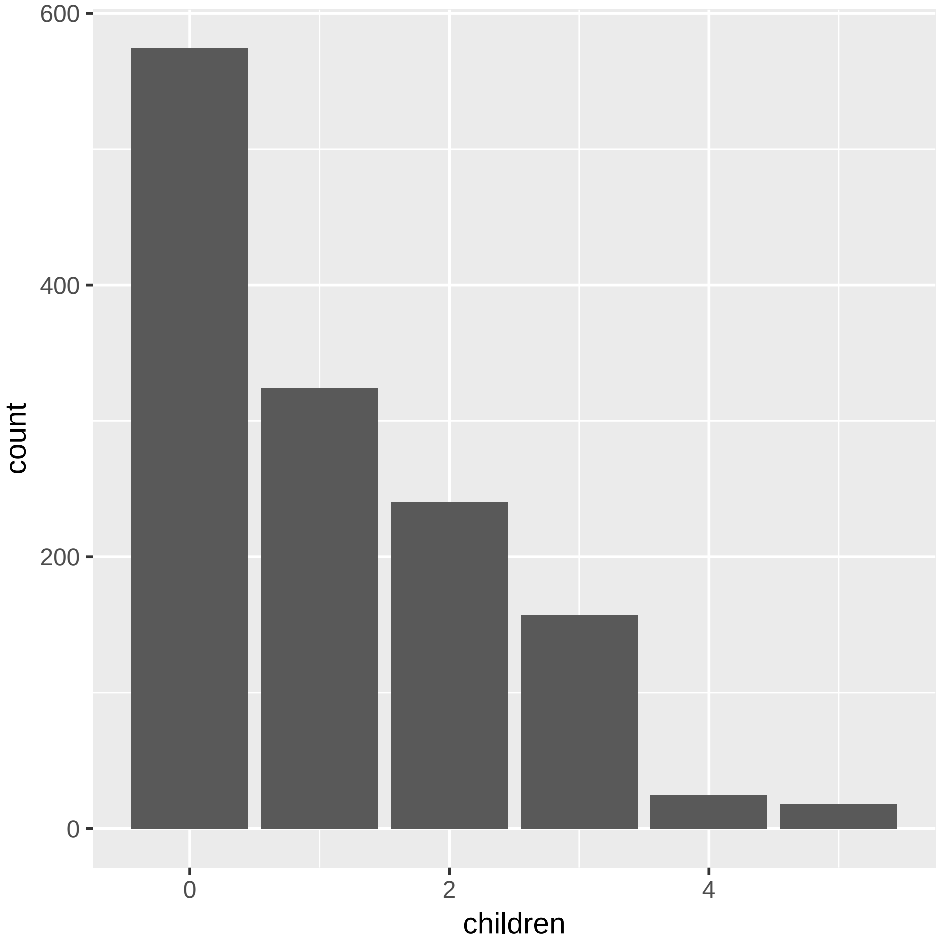
```
health_charges_clean$bmi_factor <- factor(health_charges_clean$bmi_factor,
      levels = c("underweight", "healthy_weight", "overweight", "obese"),
      ordered = TRUE)
ggplot(health_charges_clean, aes(bmi_factor)) +
  geom_bar()
```



BMI_factor

- More observations for higher BMI categories

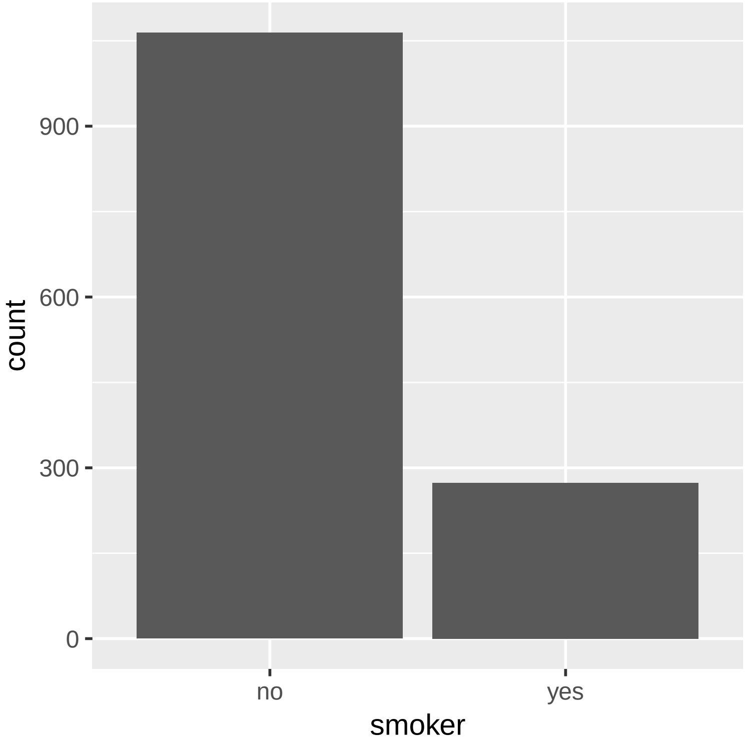
```
ggplot(health_charges_clean, aes(children))+
  geom_bar()
```



Children

- The data is skewed right.

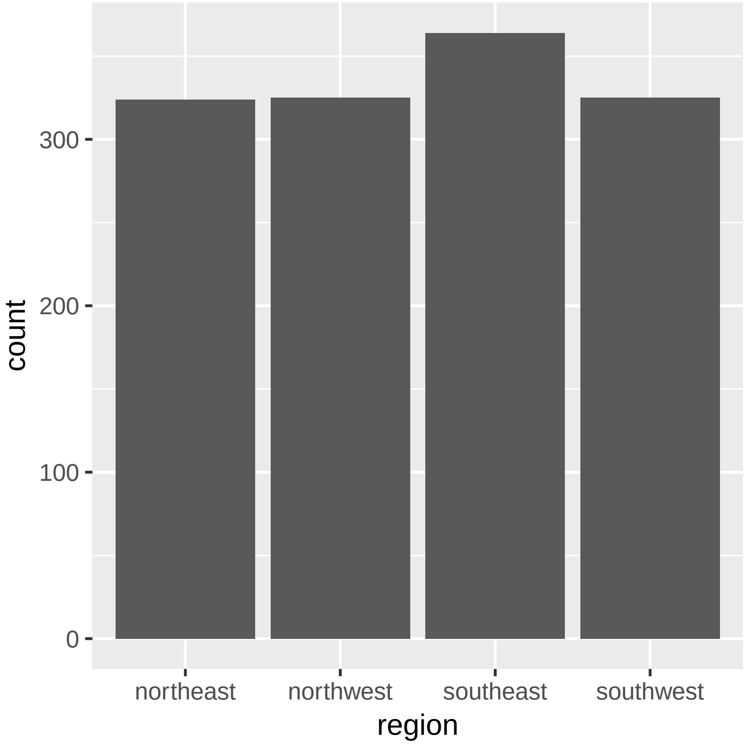
```r
ggplot(health_charges_clean, aes(smoker))+
  geom_bar()
```



Smoker

- The ratio of non-smokers to smokers is approximately $4:1$

```r
ggplot(health_charges_clean, aes(region))+
  geom_bar()
```

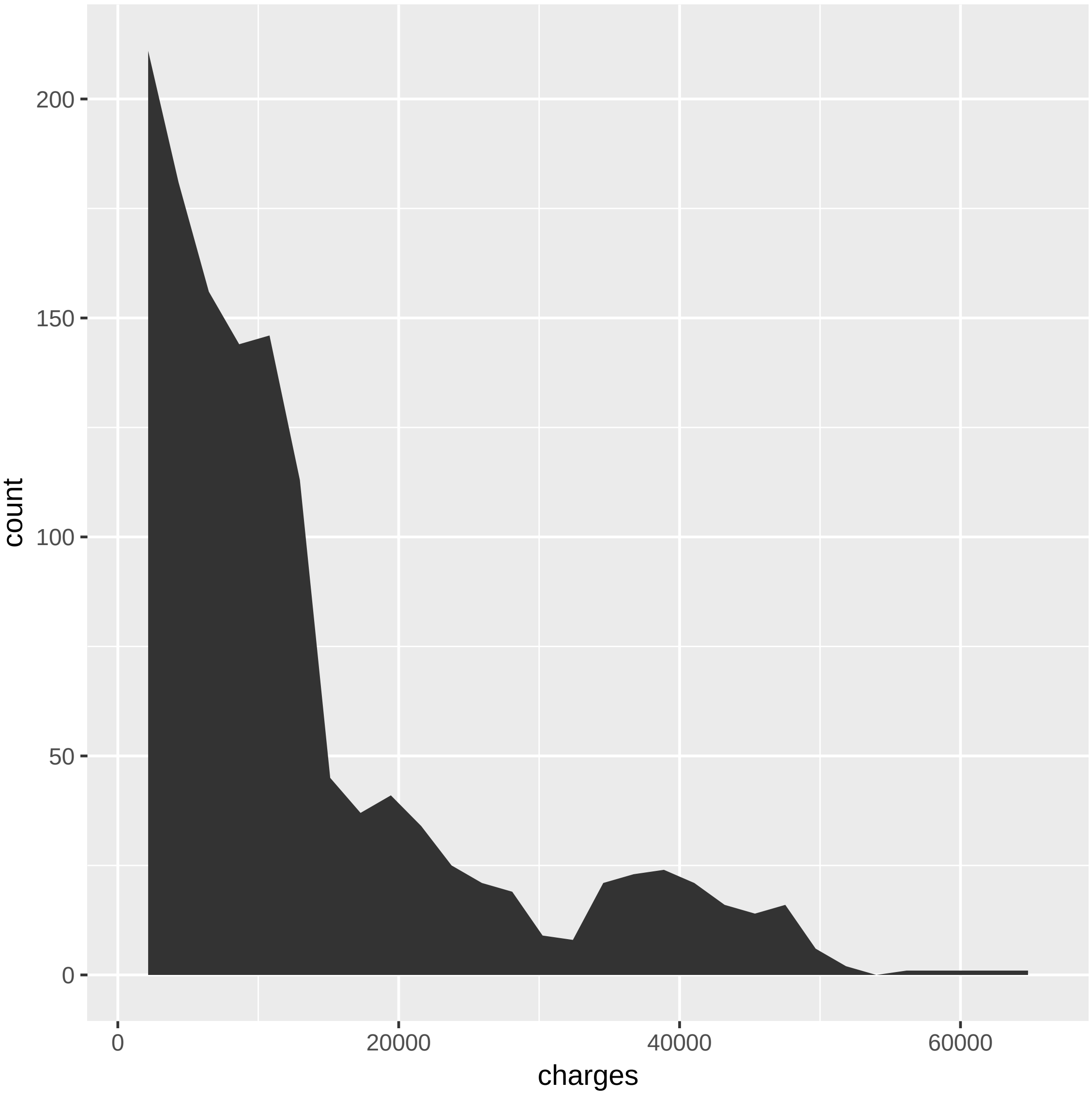


Region

- All regions except southeast had between 324-325 observations.
- Perhaps cluster sampling was used for data collection.

```r
ggplot(health_charges_clean, aes(charges)) +
  geom_area(stat = "bin")
```

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



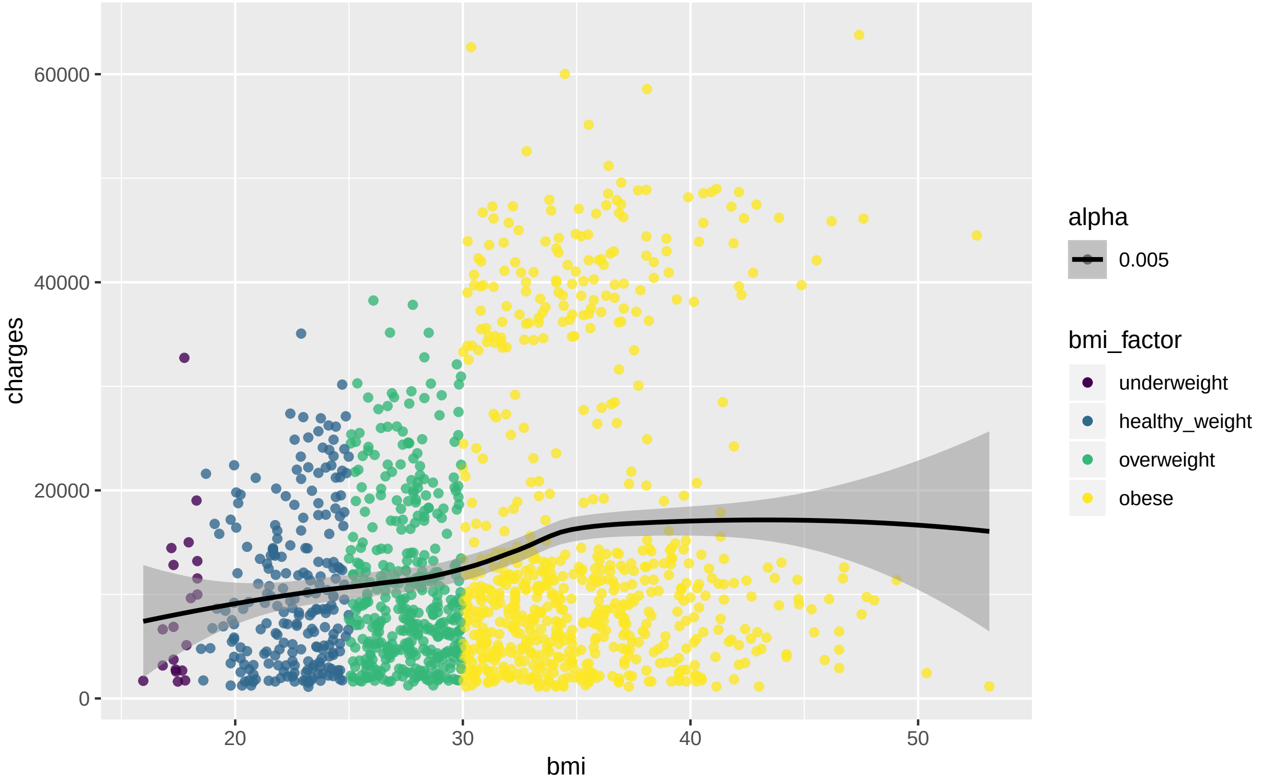
Charges

- Skewed right

Multivariable analysis

Relationships between multiple variables with anecdotal notes

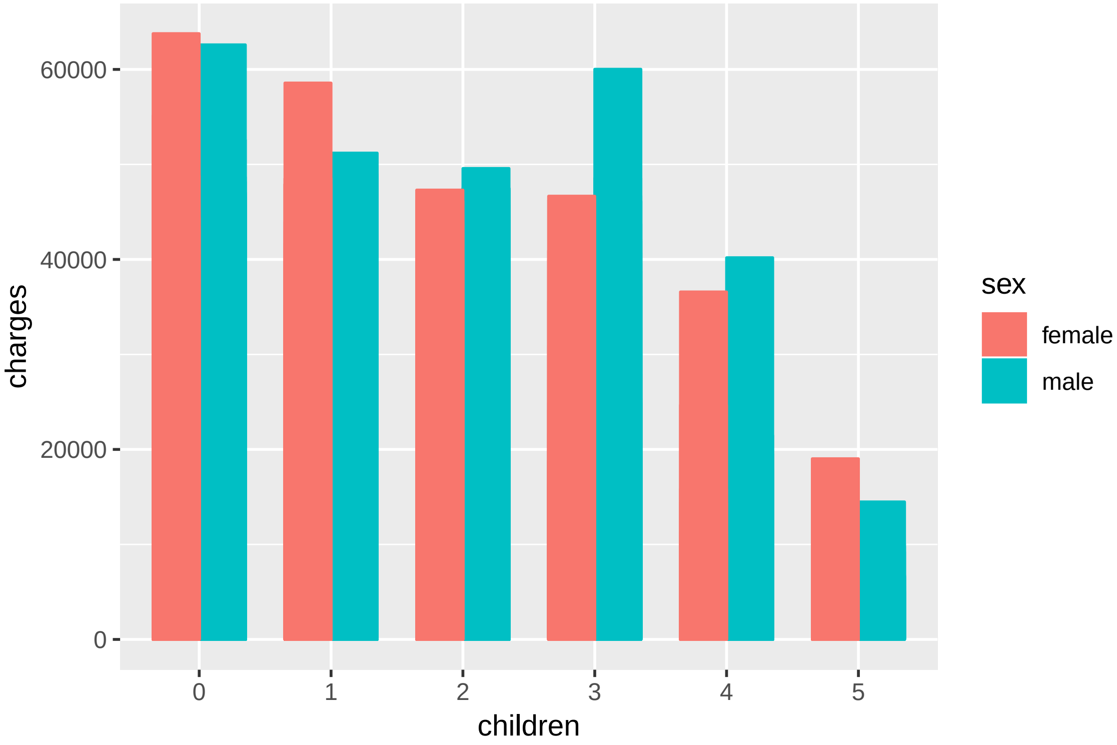
```r
ggplot(health_charges_clean,
       aes(x = bmi, y = charges, color = bmi_factor, alpha = .005 ))+
  geom_point() +
  geom_jitter() +
  geom_smooth (method = "loess", color = "black")
```



Effect of BMI on charges

- Charges increase with higher BMIs.
- There is a positive linear correlation between charges and bmi less than 35.
- There is no meaningful correlation between charges and bmi above 35.

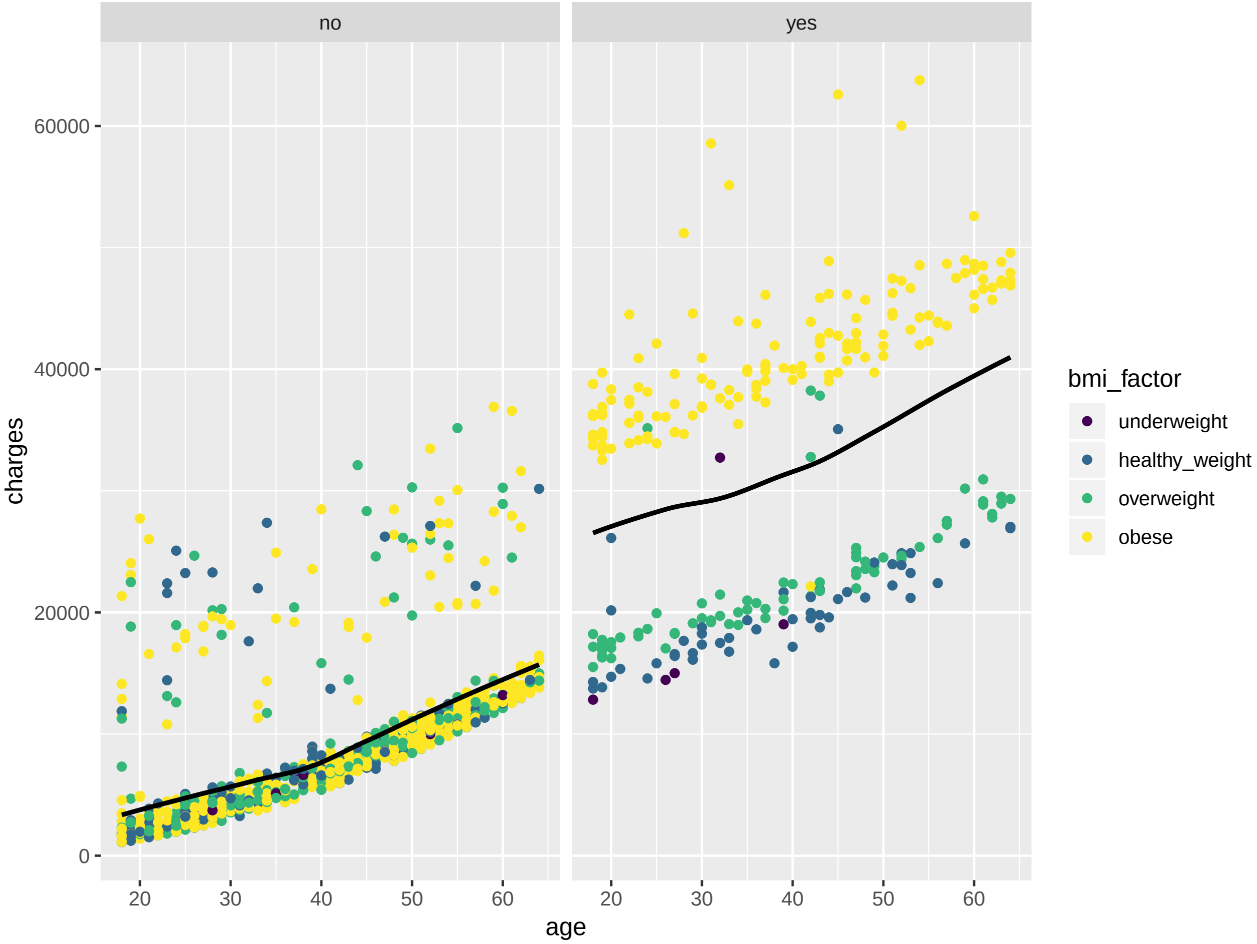
```r
health_charges_clean$children <- as.factor(health_charges_clean$children)

ggplot(health_charges_clean, aes(x = children, y = charges, color = sex)) +
  geom_bar(stat = "identity", aes(color = sex, fill = sex),
           width = .7, position = "dodge")
```



Effect of children on charges, considering sex

- Charges decrease with higher numbers of children.
- Women do not have higher health charges than men in regard to the number of children.

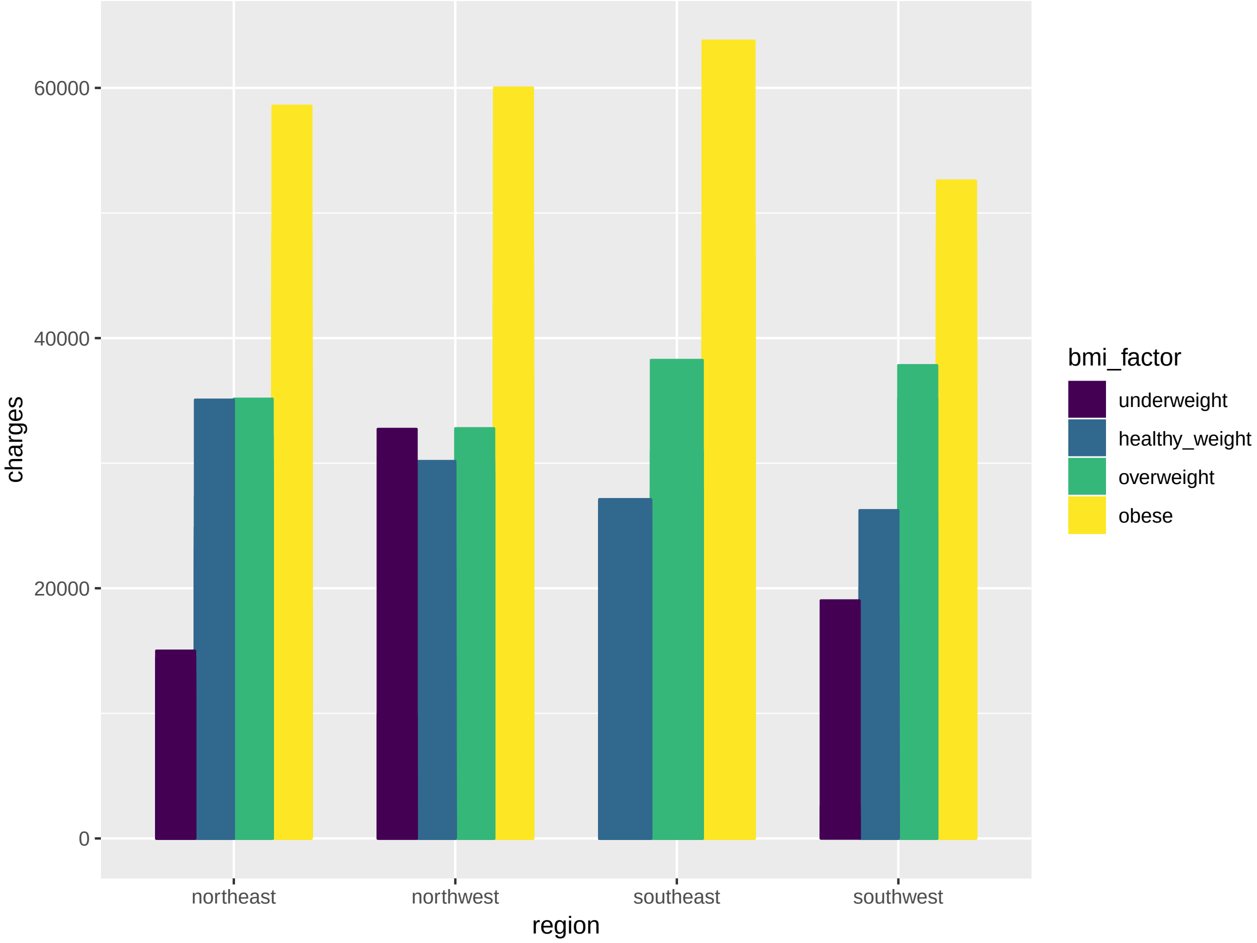
```
ggplot(health_charges_clean, aes(x = age, y = charges, color = bmi_factor), alpha = .02, size = .02) +
  geom_point(aes(color = bmi_factor, fill = bmi_factor))+
  facet_grid( . ~ smoker)+
  geom_smooth(se = FALSE, method = "loess", weight = .005, color = "black", alpha = .02 )
```



Timeseries of charges, considering BMI and smoking

- Smokers have higher charges than non-smokers.
- Smokers see a strong positive correlation between a higher BMI and charges.
- Obese smokers have higher charges than most non-smokers of all BMIs.

```r
ggplot(health_charges_clean, aes(x = region, y = charges, color = bmi_factor))+
  geom_bar(stat = "identity", position = "dodge",
           aes(color = bmi_factor, fill = bmi_factor), width = .7)
```



Region's effect on charges, considering BMI

- There were no underweight observations in the southeast region.
- BMI is a stronger indicator for charges in the south than in the north.

Statistical tests

Parametric and non-parametric tests with graphical representations.

ANOVA test, comparing the true mean BMI of adults with different numbers of children

- HO:The true mean BMI for adults with different numbers of children is uniform, at a .05 significance level.

- HA:The true mean BMI for adults with different numbers of children is not uniform, at .05 significance level.

- RESULT:
  - P = .883 > .05.

  - Fail to reject HO.

  - There is not enough evidence to support that the true mean BMI for adults with different numbers of children is not uniform, at .05 significance level.

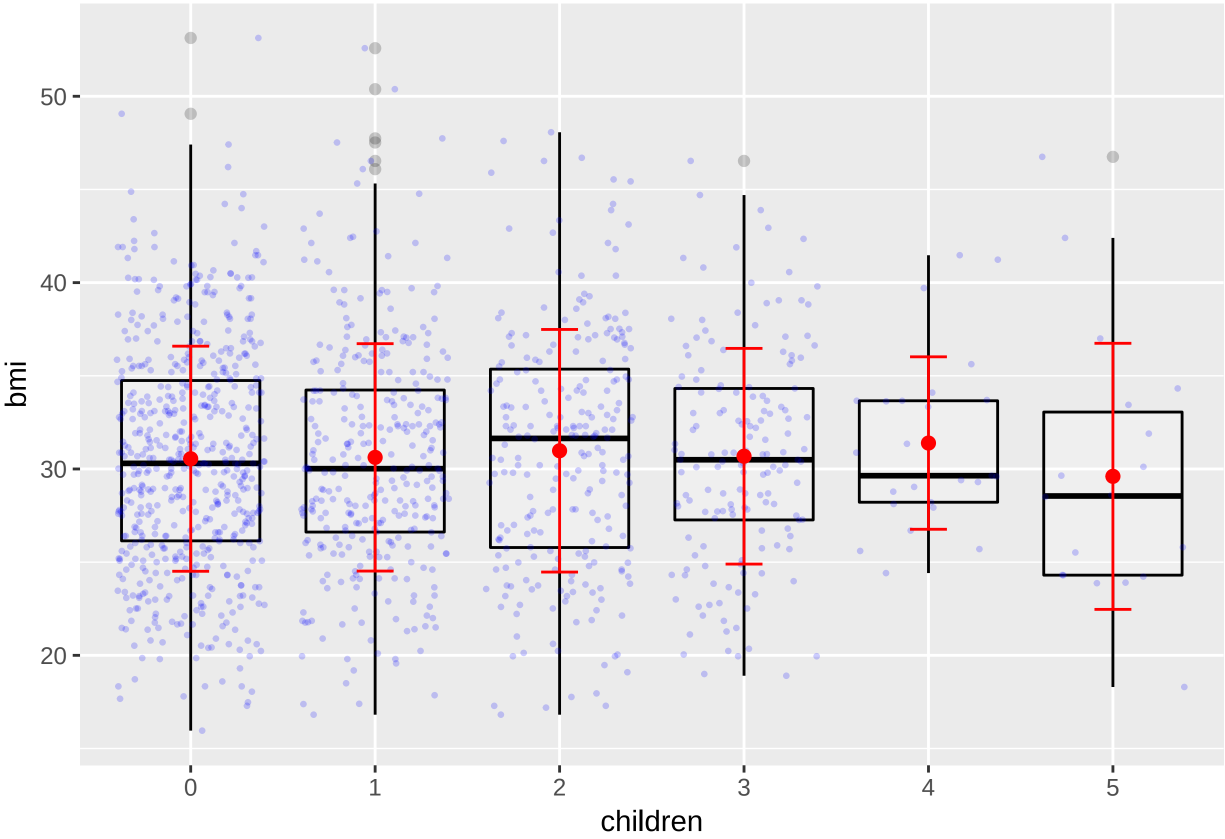
```
group_by(health_charges_clean, children) %>%
  summarise(
    count = n(),
    mean = mean(bmi, na.rm = TRUE),
    sd = sd(bmi, na.rm = TRUE)
    )
```

```
## # A tibble: 6 x 4
##   children count  mean    sd
##   <fct>    <int> <dbl> <dbl>
## 1 0          574  30.6  6.04
## 2 1          324  30.6  6.10
## 3 2          240  31.0  6.51
## 4 3          157  30.7  5.79
## 5 4           25  31.4  4.63
## 6 5           18  29.6  7.14
```

```
aov_childrenbmi <- aov(bmi ~ children, data = health_charges_clean)
summary(aov_childrenbmi)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## children       5     65   13.02   0.349  0.883
## Residuals   1332  49655   37.28
```

```
ggplot(health_charges_clean, aes(x=children, y=bmi)) +
  geom_boxplot(color = "black", alpha = .2) +
  geom_jitter(color = "blue", size = .5, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),        geom = "errorbar", color = "red",
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "point", color = "red", size = 2)
```



Kruskal-Wallis Test, comparing median health charges for adults with different numbers of children

- HO: The median health charges between adults with different numbers of children are equal, at a .05 significance level.

- HA: The median health charges between adults with different numbers of children are unequal, at .05 significance level.

- RESULT:
  - P = 1.86e-05 < .05.

  - Reject HO.

  - Evidence supports that the median health charges between adults with different numbers of children are unequal.

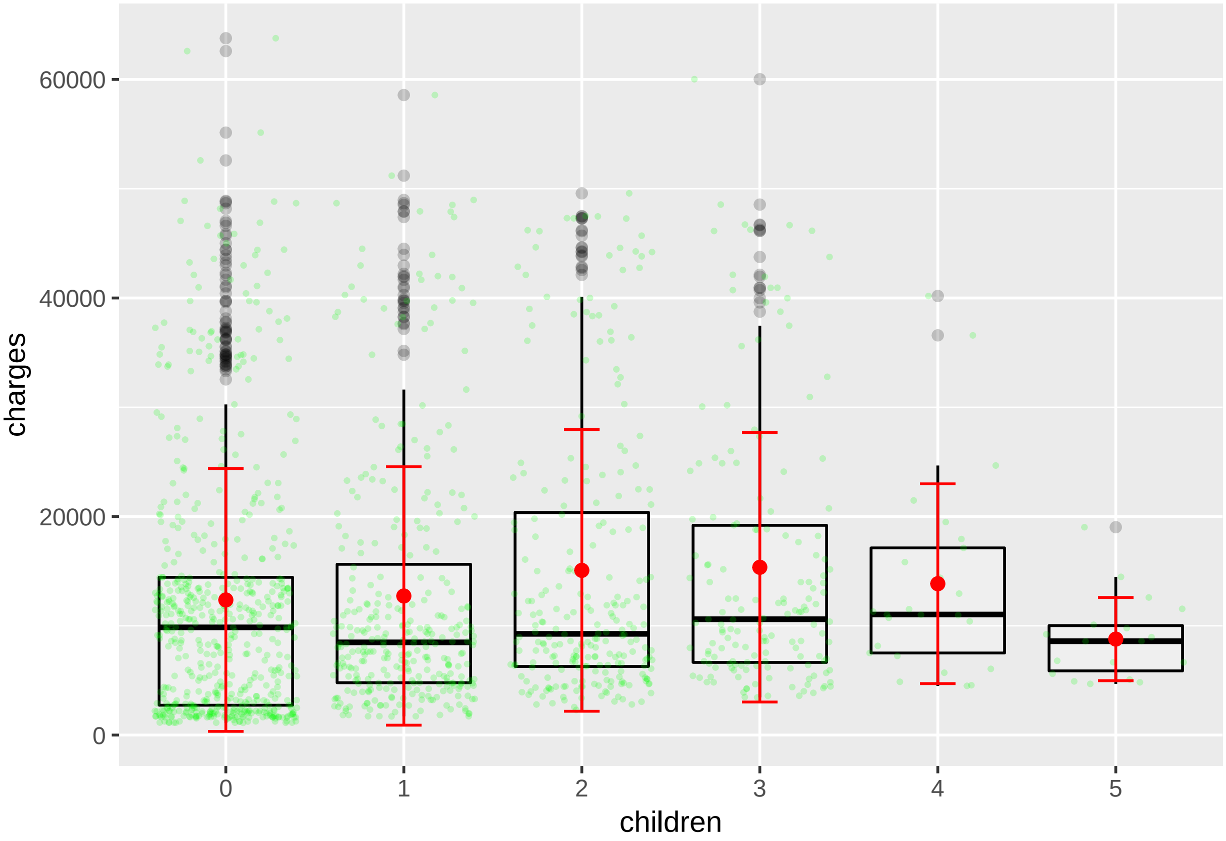
```r
group_by(health_charges_clean, children) %>%
  summarise(
    count = n(),
    mean = mean(charges, na.rm = TRUE),
    sd = sd(charges, na.rm = TRUE)
    )
```

```
## # A tibble: 6 x 4
##   children count   mean     sd
##   <fct>    <int>  <dbl>  <dbl>
## 1 0          574 12366. 12023.
## 2 1          324 12731. 11824.
## 3 2          240 15074. 12891.
## 4 3          157 15355. 12331.
## 5 4           25 13851.  9139.
## 6 5           18  8786.  3808.
```

```r
kruskal.test(charges ~ children, data = health_charges_clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  charges by children
## Kruskal-Wallis chi-squared = 29.487, df = 5, p-value = 1.86e-05
```

```r
ggplot(health_charges_clean, aes(x=children, y=charges)) +
  geom_boxplot(color = "black", alpha = .2) +
  geom_jitter(color = "green", size = .5, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),
    geom = "errorbar", color = "red", width = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),
    geom = "point", color = "red", size = 2)
```



Independent T-Test, comparing mean bmi between sexes

- HO: Both sexes have the same true mean bmi, at a .05 significance level. .
- HA: Sexes have a different true mean bmi, at a .05 significance level.

- RESULTS:
  - P = .08992 > .05.

  - Fail to reject HO.

  - There is not enough evidence to support that sexes have a different true mean bmi, at a .05 significance level.

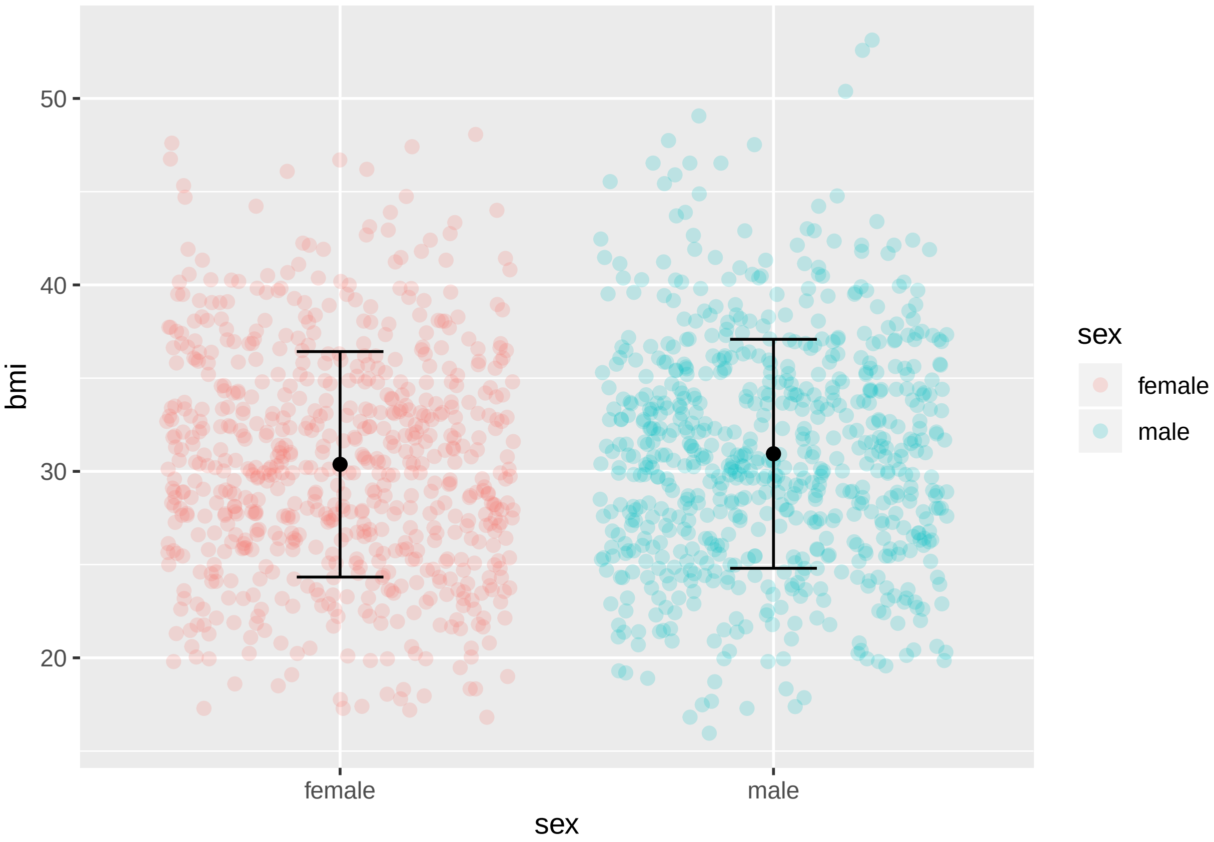
```
group_by(health_charges_clean, sex) %>%
  summarise(
    count = n(),
    mean = mean(charges),
    sd = sd(charges)
  )
```

```
## # A tibble: 2 x 4
##   sex    count   mean     sd
##   <fct>  <int>  <dbl>  <dbl>
## 1 female   662 12570. 11129.
## 2 male     676 13957. 12971.
```

```
t.test(bmi ~ sex, data = health_charges_clean)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by sex
## t = -1.697, df = 1336, p-value = 0.08992
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.21895043  0.08819153
## sample estimates:
## mean in group female   mean in group male
##             30.37775             30.94313
```

```
ggplot(health_charges_clean, aes(x=sex, y=bmi, color =sex)) +
  geom_jitter(size = 2, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),        geom = "errorbar", color = "black
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "point", color = "black", size = 
```



Independent T-Test, comparing mean bmi between smokers and non-smokers

- HO: Smokers and non-smokers have the same true mean bmi, at a .05 significance level.

- HA: Smokers and non-smokers have a different true bmi, at a .05 significance level.

- RESULTS:
  - P = 0.8938 > .05.

  - Failt to reject HO.

  - There is not enough evidence to support that smokers and non-smokers have a different true bmi, at a .05 significance level.

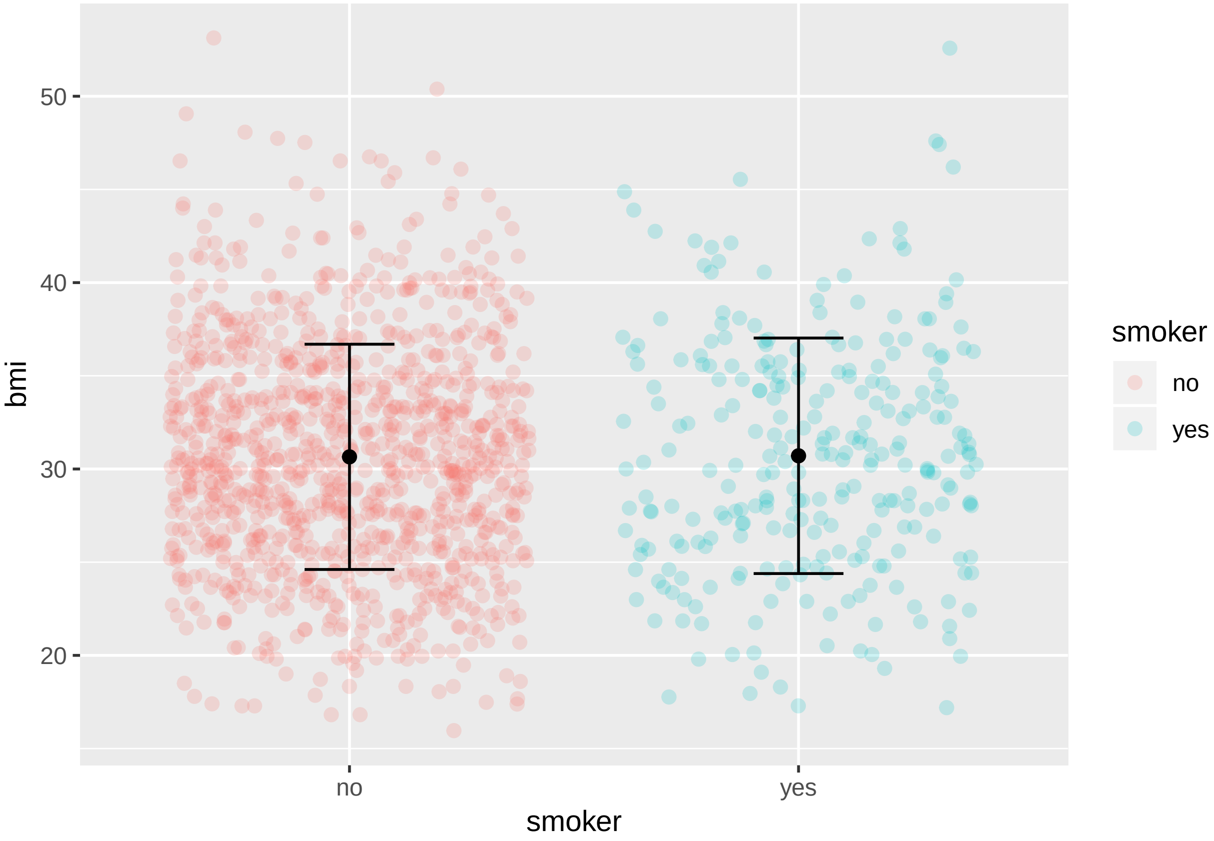
```
group_by(health_charges_clean, smoker) %>%
  summarise(
    count = n(),
    mean = mean(charges),
    sd = sd(charges)
  )
```

```
## # A tibble: 2 x 4
##   smoker count   mean     sd
##   <fct>  <int>  <dbl>  <dbl>
## 1 no      1064  8434.  5994.
## 2 yes      274 32050. 11542.
```

```
  t.test(bmi ~ smoker, data = health_charges_clean)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by smoker
## t = -0.13352, df = 410.9, p-value = 0.8938
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8907341  0.7774265
## sample estimates:
##  mean in group no mean in group yes
##          30.65180          30.70845
```

```
library(ggplot2)
  ggplot(health_charges_clean, aes(x=smoker, y=bmi, color =smoker)) +
  geom_jitter(size = 2, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),         geom = "errorbar", color = "black
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "point", color = "black", size = 2
```



MANN-WHITNEY-WILCOXON TEST, comparing charges between the sexes

- HO: The charges of females and males have identical distributions of charges at a .05 significance level.

- HA: The charges of females and males have different distributions of charges at a .05 significance level.

- RESULTS:
  - P=.7287 > .05.

  - Fail to reject HO.

  - There is not enough evidence to prove that the charges of females and males have different distributions of charges at a .05 significance level

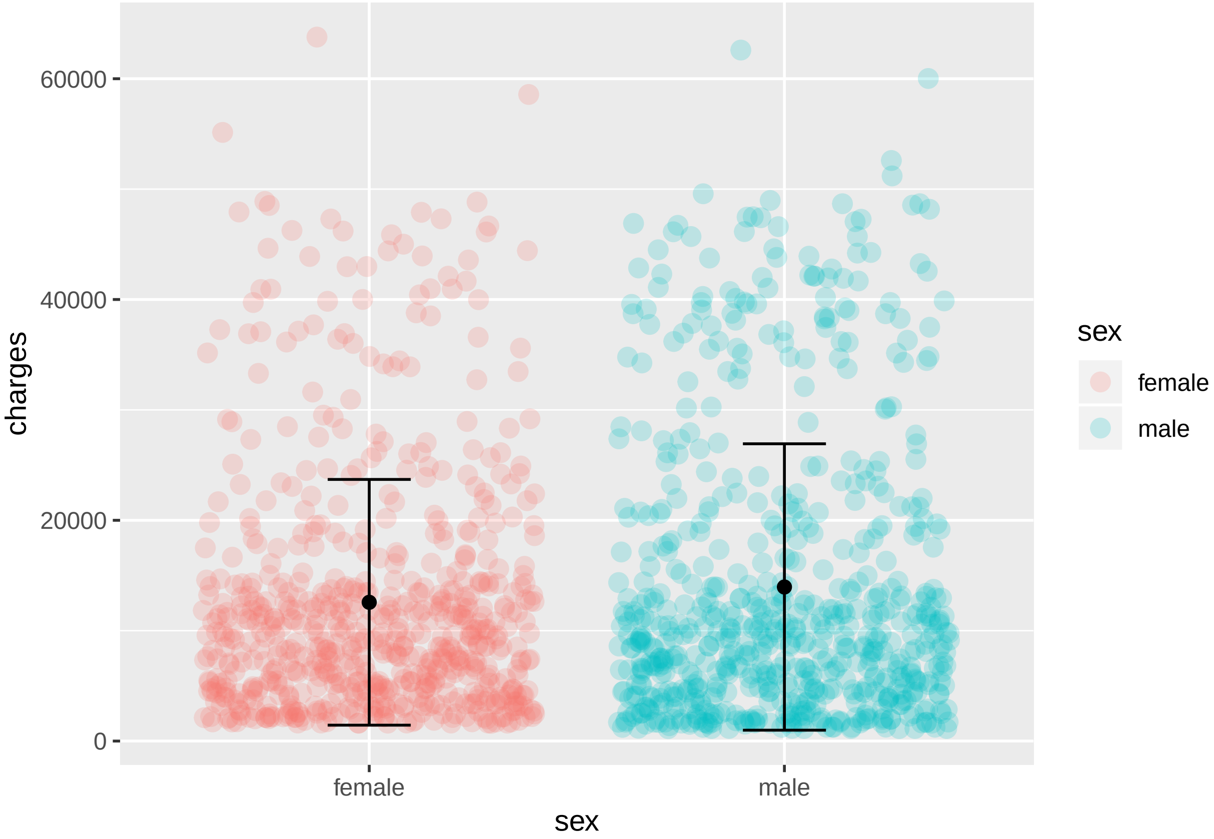
```r
group_by(health_charges_clean, sex) %>%
  summarise(
    count = n(),
    mean = mean(charges),
    sd = sd(charges)
  )
```

```
## # A tibble: 2 x 4
##   sex    count  mean     sd
##   <fct>  <int>  <dbl>  <dbl>
## 1 female   662 12570. 11129.
## 2 male     676 13957. 12971.
```

```r
wilcox.test( charges ~ sex, data = health_charges_clean)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  charges by sex
## W = 221300, p-value = 0.7287
## alternative hypothesis: true location shift is not equal to 0
```

```r
ggplot(health_charges_clean, aes(x=sex, y=charges, color =sex)) +
  geom_jitter(size = 3, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),         geom = "errorbar", color = "black
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "point", color = "black", size = 2
```



MANN-WHITNEY-WILCOXON TEST, comparing charges between smokers and non-smokers

- HO: The charges of smokers and non-smokers have identical distributions of charges at a .05 significance level.

- HA: The charges of smokers and non-smokers have different distributions of charges at a .05 significance level.

- RESULTS:
  - P < 2.2e-16 < .05.

  - Reject HO.

  - Evidence supports that the charges of smokers and non-smokers have different distributions of charges at a .05 significance level

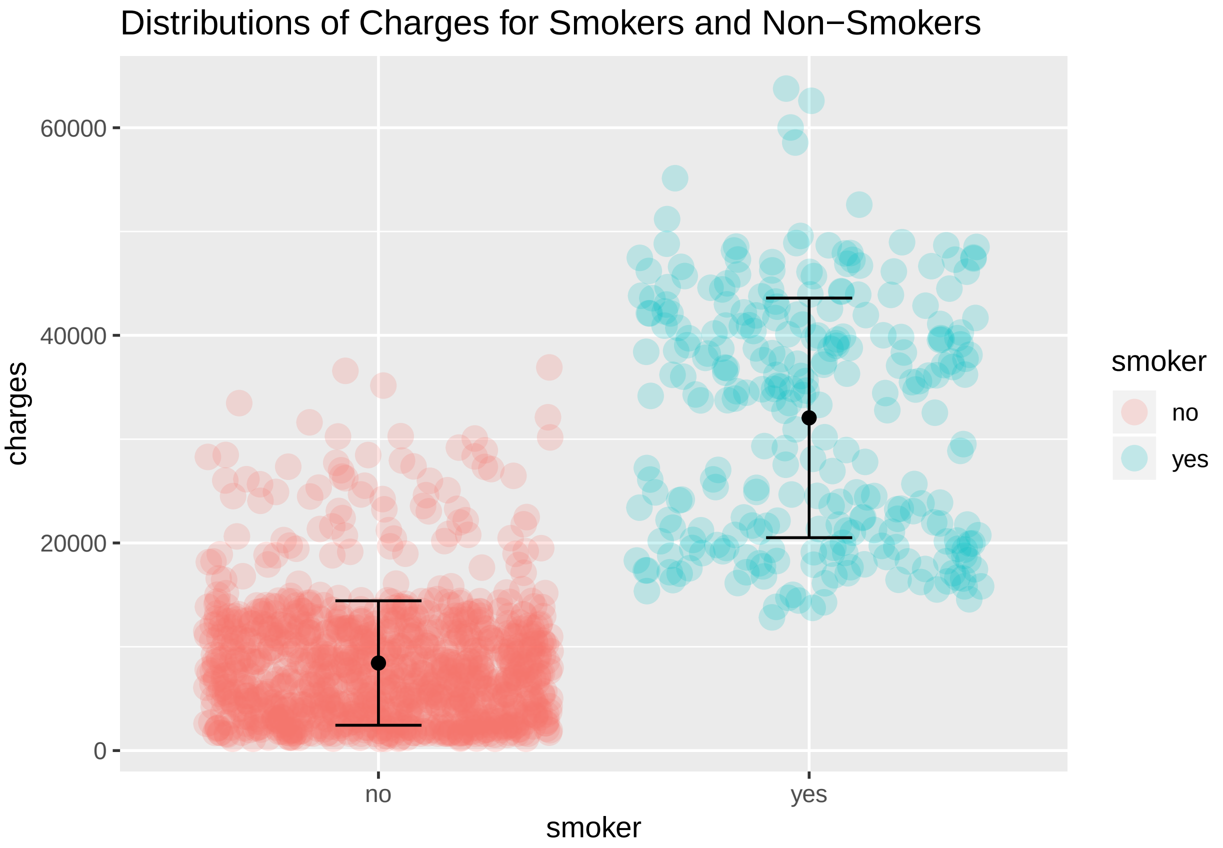
```
library(dplyr)
group_by(health_charges_clean, smoker) %>%
  summarise(
    count = n(),
    mean = mean(charges),
    sd = sd(charges)
  )
```

```
## # A tibble: 2 x 4
##   smoker count   mean     sd
##   <fct>  <int>  <dbl>  <dbl>
## 1 no      1064  8434.  5994.
## 2 yes      274 32050. 11542.
```

```
wilcox.test( charges ~ smoker, data = health_charges_clean)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  charges by smoker
## W = 7403, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
ggplot(health_charges_clean, aes(x=smoker, y=charges, color =smoker)) +
  geom_jitter(size = 4, alpha = .2) +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1),         geom = "errorbar", color = "black
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "point", color = "black", size =
  labs(title = "Distributions of Charges for Smokers and Non-Smokers") +
  scale_fill_manual(name = "Status", labels = c("Nonsmoker", "Smoker"))
```


Distributions of Charges for Smokers and Non−Smokers

PEARSON'S LINEAR REGRESSION, describing the linear relationship between bmi and charges

- HO: The true correlation between bmi and charges is equal to 0 at a .05 significance level.

- HA: The true correlation between bmi and charges is not equal to 0 at a .05 significance level.

- RESULTS:
  - P-Value = 2.459e-13 < .05.

  - Reject HO.

  - Evidence supports that the true correlation between bmi and charges is not equal to 0 at a .05 significance level.

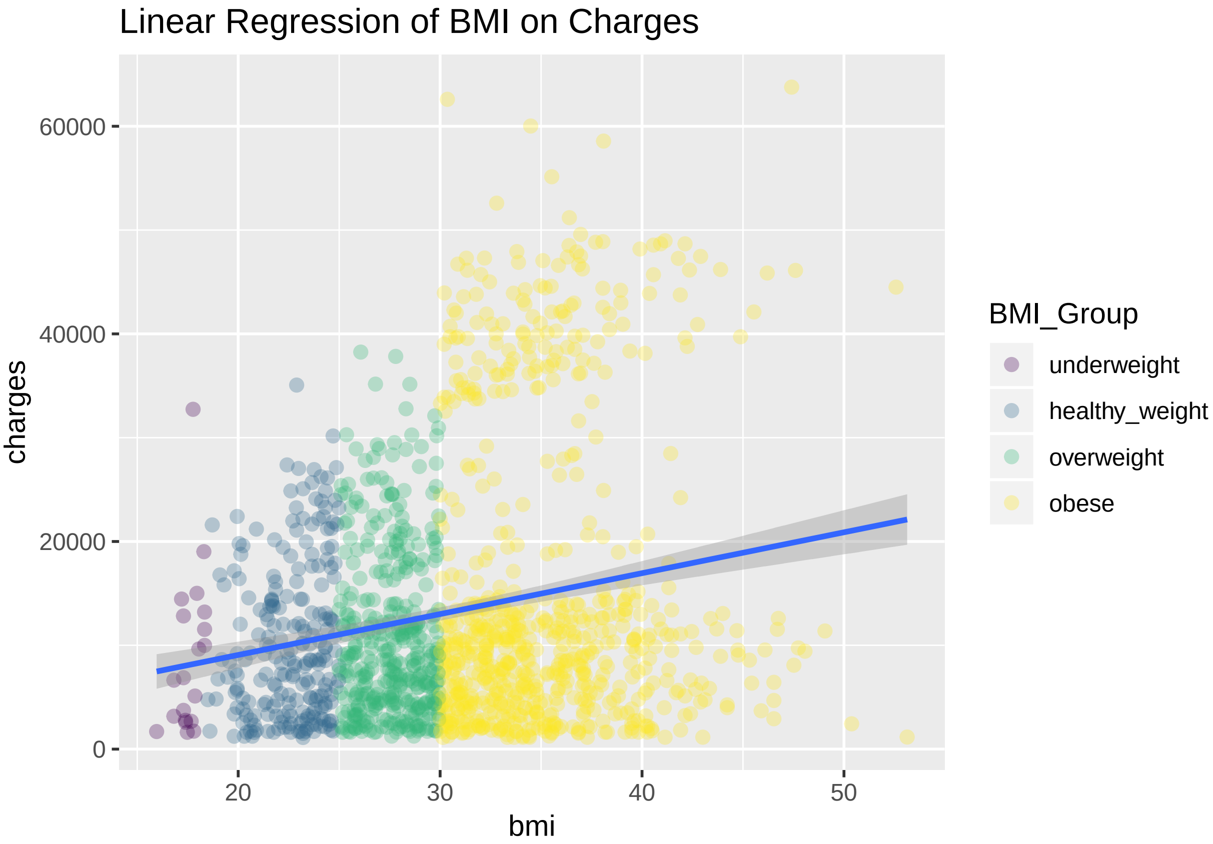
  - The true correlation between bmi and charges is .198341, with CI = 0.1463052, 0.2492822.

  - There is a weak positive correlation between bmi and charges.

```r
cor.test(health_charges_clean$bmi, health_charges_clean$charges, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  health_charges_clean$bmi and health_charges_clean$charges
## t = 7.3966, df = 1336, p-value = 2.459e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1463052 0.2492822
## sample estimates:
##      cor
## 0.198341
```

```
BMI_Group <- health_charges_clean$bmi_factor
ggplot(health_charges_clean, aes(x = bmi, y = charges))+
  geom_point(size = 2, alpha = .3, aes(color = BMI_Group))+
  geom_smooth(aes(x = bmi, y = charges), method = lm) +
  labs(title = "Linear Regression of BMI on Charges")
```



PEARSON'S LINEAR REGRESSION, describing the linear relationship between bmi and charges, subset by bmi_factor
HO: The true correlation between bmi and charges is equal to 0 at a .05 significance level.

- HA: The true correlation between bmi and charges is not equal to 0 at a .05 significance level.
- RESULTS:
- Underweight:
  - p-value = 0.071 > .05, fail to reject HO.

  - There is not enough evidence to support the claim that the true correlation between bmi and charges is not equal to 0 at a .05 significance level.

  - 95 percent confidence interval: (-0.03721726, 0.72280204).

  - Coefficient: 0.4120904

  - Moderate positive correlation between underweight bmi and charges.

- Healthy_weight:
  - p-value = 0.006103 < .05, reject HO.

  - Evidence supports the claim that the true correlation between bmi and charges is not equal to 0 at a .05 significance level.

  - 95 percent confidence interval: (0.05276277, 0.30579513).

  - Coefficient: 0.1822954

  - Weak positive correlation between underweight bmi and charges.

- Overweight:
  - p-value = 0.839 > .05, fail to reject HO.

  - There is not enough evidence to support the claim that the true correlation between bmi and charges is not equal to 0 at a .05 significance level.

  - 95 percent confidence interval: (-0.11007646, 0.08953425).

  - Coefficient: -0.01037446

  - There is a negligiblely weak correlation between overweight bmi and charges.
- Obese:
  - p-value = 0.09527 > .05, fail to reject HO.

  - There is not enough evidence to support the claim that the true correlation between bmi and charges is not equal to 0 at a .05 significance level.

  - 95 percent confidence interval: (-0.01099551, 0.13589593).

  - Coefficient: 0.06279025

  - There as a negligbly weak positive correlation between overweight bmi and charges.

```r
underweight <- subset(health_charges_clean, bmi < 18.5, select = c(bmi))
ucharges <- subset(health_charges_clean, bmi < 18.5, select = c(charges))
cor.test(underweight[ ,1], ucharges[ ,1], method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  underweight[, 1] and ucharges[, 1]
## t = 1.9189, df = 18, p-value = 0.071
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03721726  0.72280204
## sample estimates:
##       cor
## 0.4120904
```

```r
healthy_weight <- subset(health_charges_clean, bmi >= 18.5 & bmi < 25, select = c(bmi))
hwcharges <- subset(health_charges_clean, bmi >= 18.5 & bmi < 25, select = c(charges))
cor.test(healthy_weight[ ,1], hwcharges[ ,1], method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  healthy_weight[, 1] and hwcharges[, 1]
## t = 2.7686, df = 223, p-value = 0.006103
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05276277 0.30579513
## sample estimates:
##       cor
## 0.1822954
```

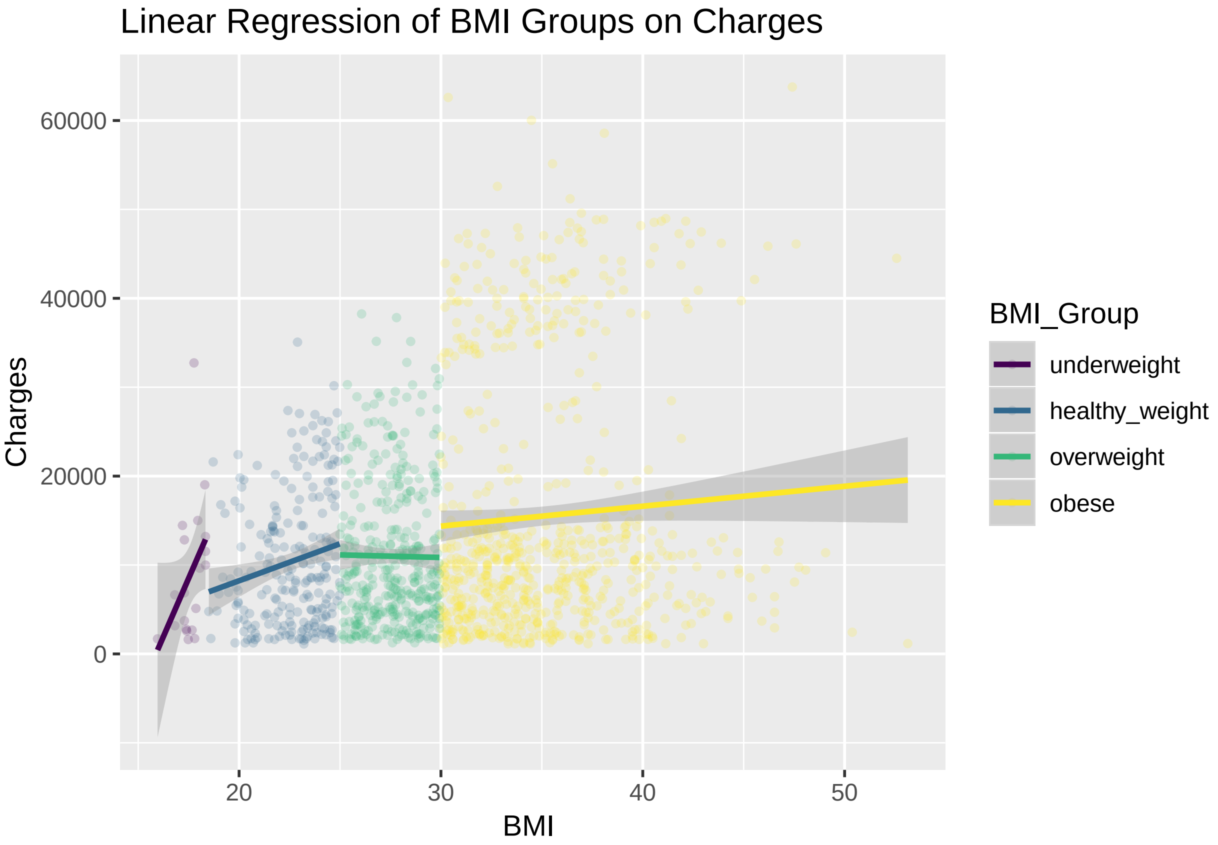
```r
overweight <- subset(health_charges_clean, bmi >= 25 & bmi < 30, select = c(bmi))
ovcharges <- subset(health_charges_clean, bmi >= 25 & bmi < 30, select = c(charges))
cor.test(overweight[ ,1], ovcharges[ ,1], method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  overweight[, 1] and ovcharges[, 1]
## t = -0.20331, df = 384, p-value = 0.839
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.11007646  0.08953425
## sample estimates:
##         cor
## -0.01037446
```

```r
obese <- subset(health_charges_clean, bmi >= 30, select = c(bmi))
obcharges <- subset(health_charges_clean, bmi >= 30, select = c(charges))
cor.test(obese[ ,1], obcharges[ ,1], method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  obese[, 1] and obcharges[, 1]
## t = 1.6705, df = 705, p-value = 0.09527
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01099551  0.13589593
## sample estimates:
##        cor
## 0.06279025
```

```
BMI_Group <- health_charges_clean$bmi_factor
ggplot(health_charges_clean, aes(x = bmi, y = charges, color = BMI_Group))+
  geom_point(size = 1, alpha = .2)+
  geom_smooth(aes(x = bmi, y = charges), method = lm)+
  labs(title = "Linear Regression of BMI Groups on Charges", y = "Charges", x = "BMI ")+
  guides(colorbar = "BMI Groups")
```



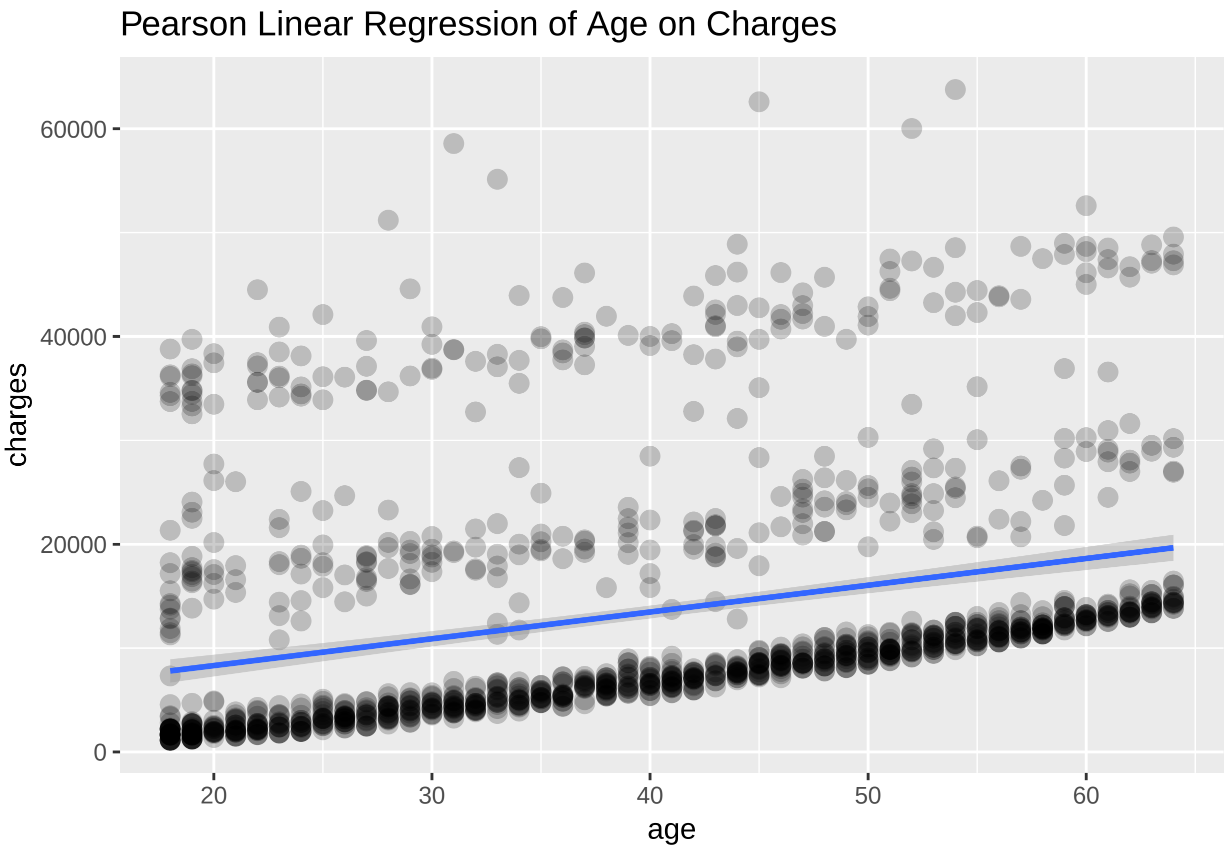
PEARSON'S LINEAR REGRESSION, describing the linear relationship between age and charges

- HO: The true correlation between age and charges is equal to 0 at a .05 significance level.
- HA: The true correlation between age and charges is not equal to 0 at a .05 significance level.
- RESULTS:
  - P-Value: < 2.2e-16 < .05.
  - Reject HO.
  - Evidence supports the claim that the true correlation between age and charges is not equal to 0 at a .05 significance level.
  - The true correlation between bmi and charges is 0.2990082, with CI = 0.2494139, 0.3470381.
  - There is a weak positive correlation between bmi and charges.

```
cor.test(health_charges_clean$age, health_charges_clean$charges, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  health_charges_clean$age and health_charges_clean$charges
## t = 11.453, df = 1336, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2494139 0.3470381
## sample estimates:
##       cor
## 0.2990082
```

```
ggplot(health_charges_clean, aes(x = age, y = charges))+
  geom_point(size = 3, alpha = .2)+
  geom_smooth(method = lm) +
  ggtitle("Pearson Linear Regression of Age on Charges")
```



CHI-SQUARED TEST FOR INDEPENDENCE, between bmi group and region
HO: Bmi group is independent of region at a .05 significance level.

- HA: Bmi group is dependent on region at a .05 significance level.
- RESULTS:
- P-Value: 4.015e-09 < .05
- Reject HO.
- Evidence supports that bmi group is dependent on region at a .05 significance level.

```
chisq.test(health_charges_clean$region, health_charges_clean$bmi_factor)
```

```
## Warning in chisq.test(health_charges_clean$region,
## health_charges_clean$bmi_factor): Chi-squared approximation may be
## incorrect
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  health_charges_clean$region and health_charges_clean$bmi_factor
## X-squared = 57.521, df = 9, p-value = 4.015e-09
```

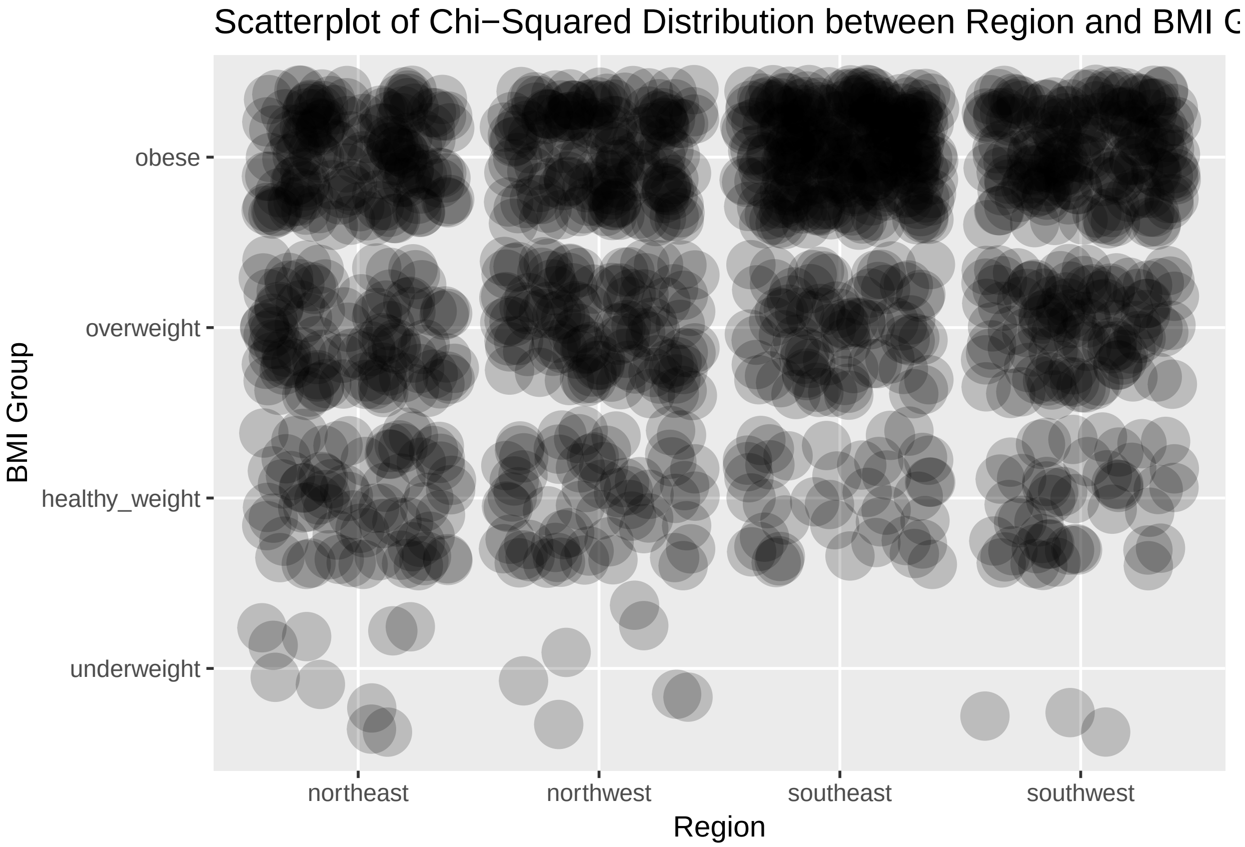
```
chisq <- chisq.test(health_charges_clean$region, health_charges_clean$bmi_factor)
```

```
## Warning in chisq.test(health_charges_clean$region,
## health_charges_clean$bmi_factor): Chi-squared approximation may be
## incorrect
```

```
chisq$observed
```

```
##                            
## health_charges_clean$region underweight healthy_weight overweight obese
##                   northeast          10             73         98   143
##                   northwest           7             63        107   148
##                   southeast           0             41         80   243
##                   southwest           3             48        101   173
```

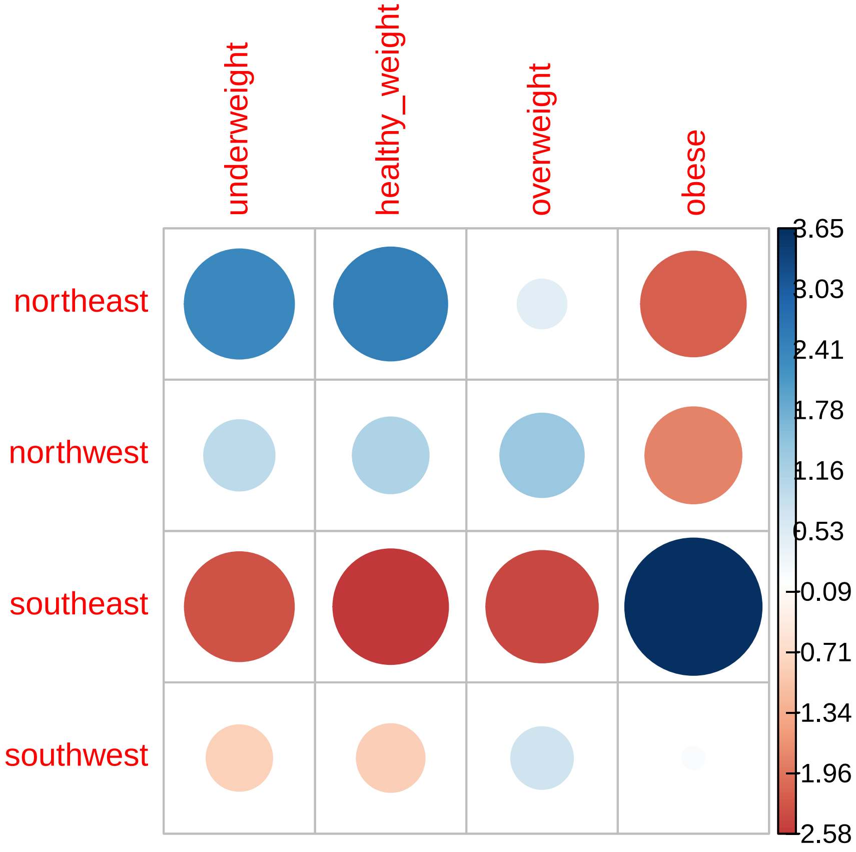
```r
ggplot(health_charges_clean, aes(x = region, y = bmi_factor)) +
  geom_jitter(alpha = .2, size = 8) +
  ggtitle("Scatterplot of Chi-Squared Distribution between Region and BMI Group") +
  ylab("BMI Group") +
  xlab("Region")
```



```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
corrplot(chisq$residuals, is.cor = FALSE)
```


underweight
healthy_weight
overweight
obese
northeast
northwest
southeast
southwest
3.65
3.03
2.41
1.78
1.16
0.53
-0.09
-0.71
-1.34
-1.96
-2.58