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 mˆ2
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

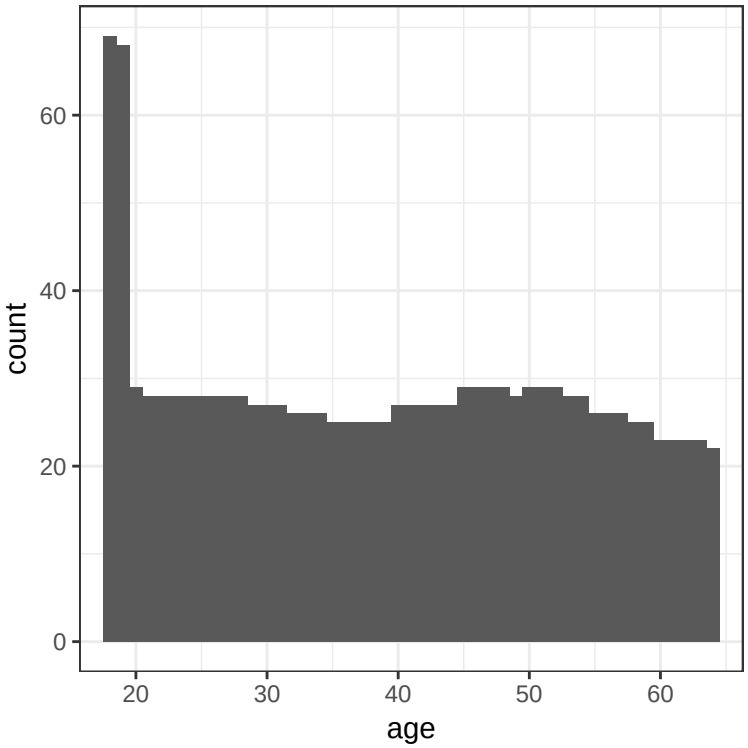
```
health_charges_clean <- read.csv("health_charges_clean.csv", header=TRUE)

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

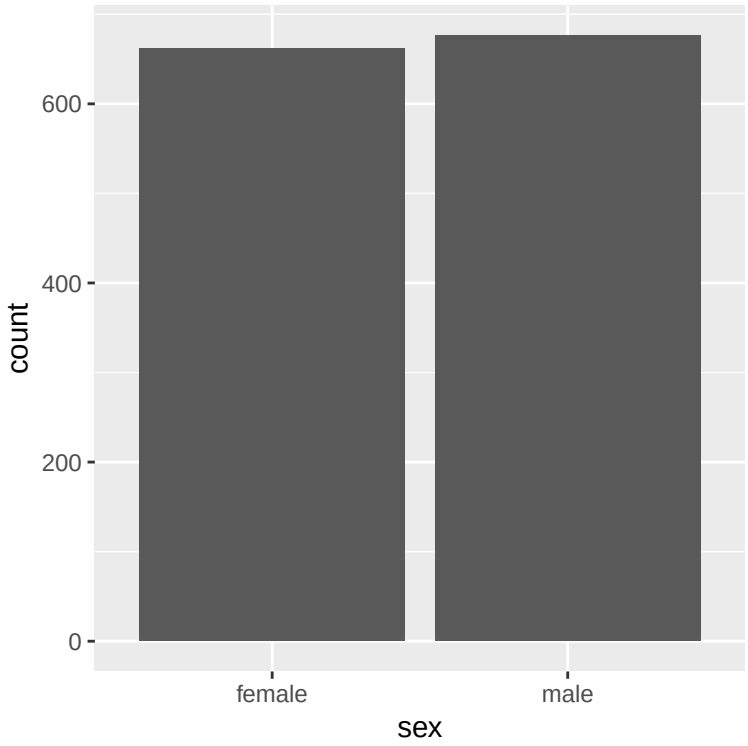
```
library(ggplot2)

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.

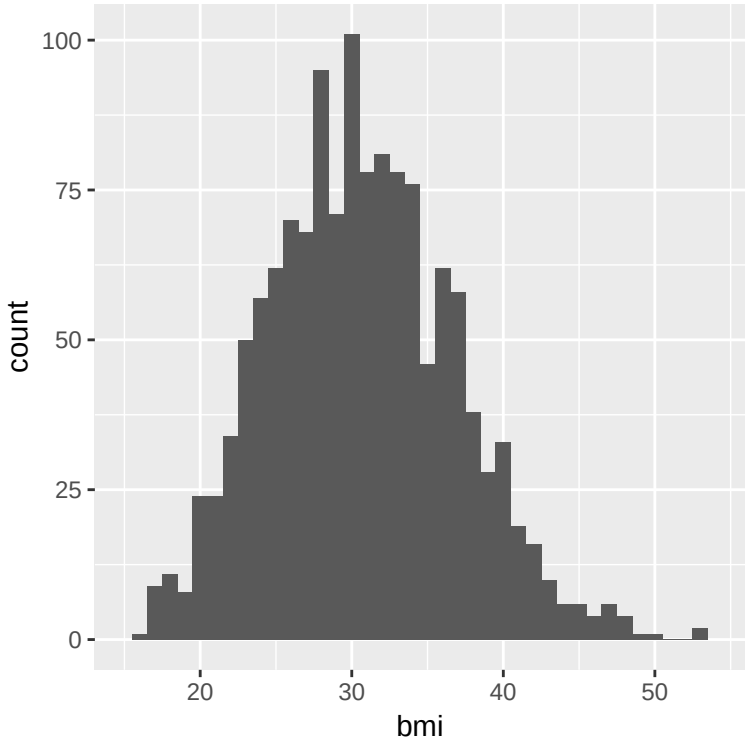
```r
ggplot(health_charges_clean, aes(sex))+
  geom_bar()
```



Sexes

- Even distribution

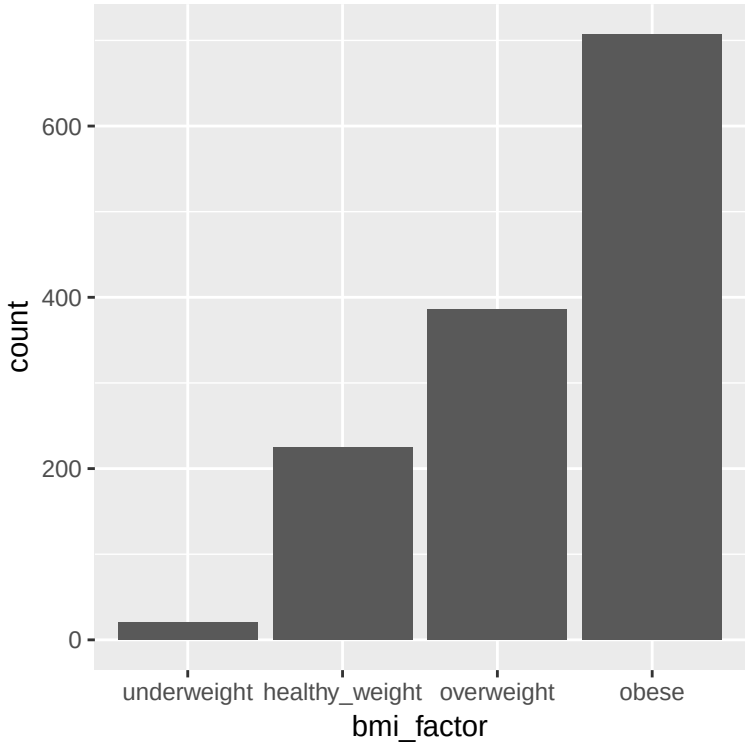
```r
ggplot(health_charges_clean, aes(bmi)) +
  geom_histogram(binwidth = 1) +
  coord_cartesian(xlim = c(15, 54))
```



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.

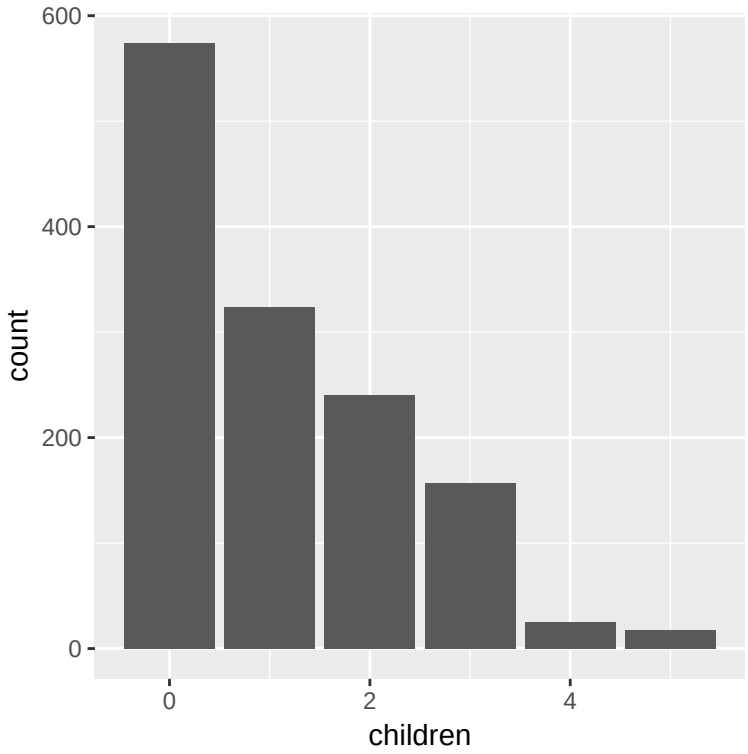
```r
health_charges_clean$bmi_factor <- factor(health_charges_clean$bmi_factor,
      levels = c("underweight", "healthy_weight", "overweight", "obese"),
      ordered = TRUE)
ggplot(health_charges_clean, aes(bmi_factor)) +
  geom_bar()
```



BMI_factor

- More observations for higher BMI categories

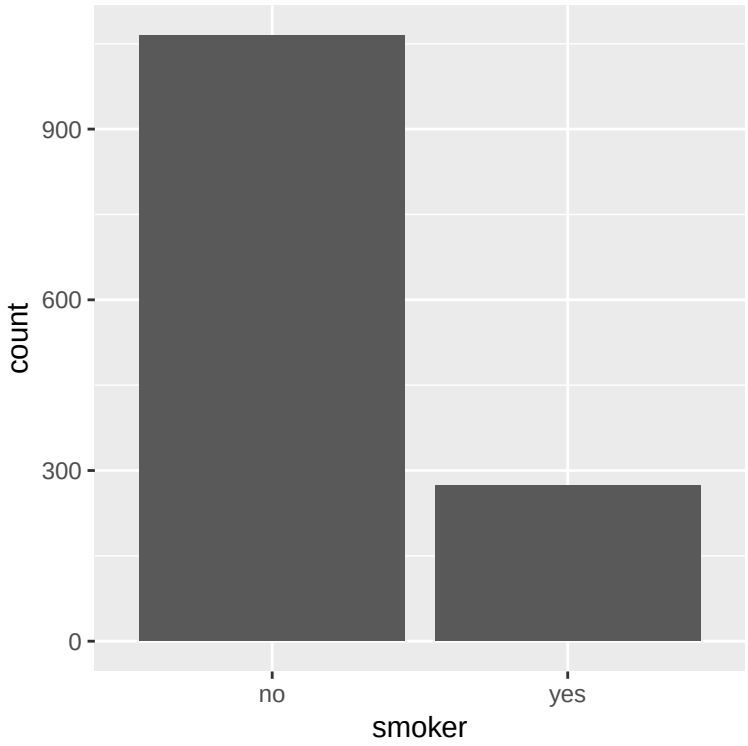
```
ggplot(health_charges_clean, aes(children))+
  geom_bar()
```



Children

- The data is skewed right.

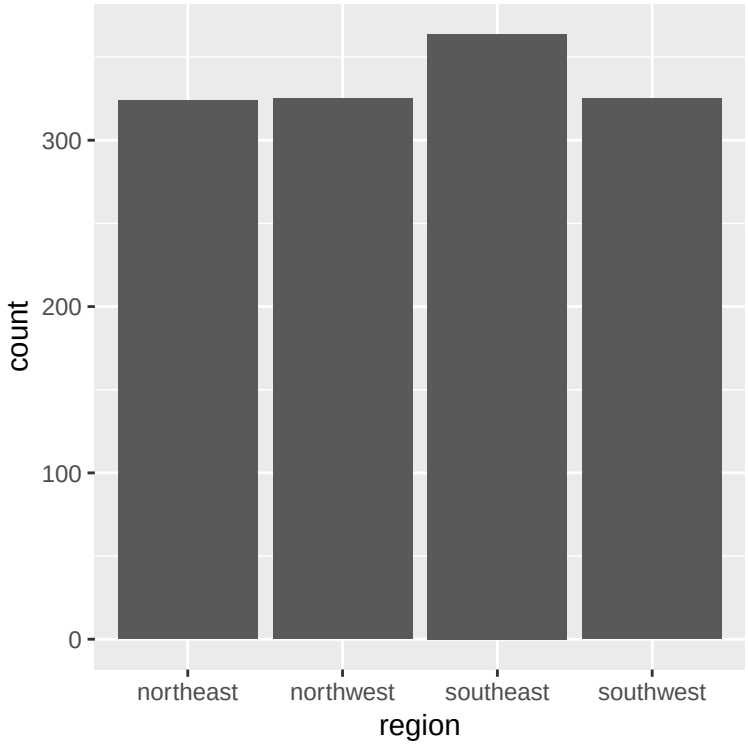
```r
ggplot(health_charges_clean, aes(smoker))+
  geom_bar()
```



Smoker

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

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

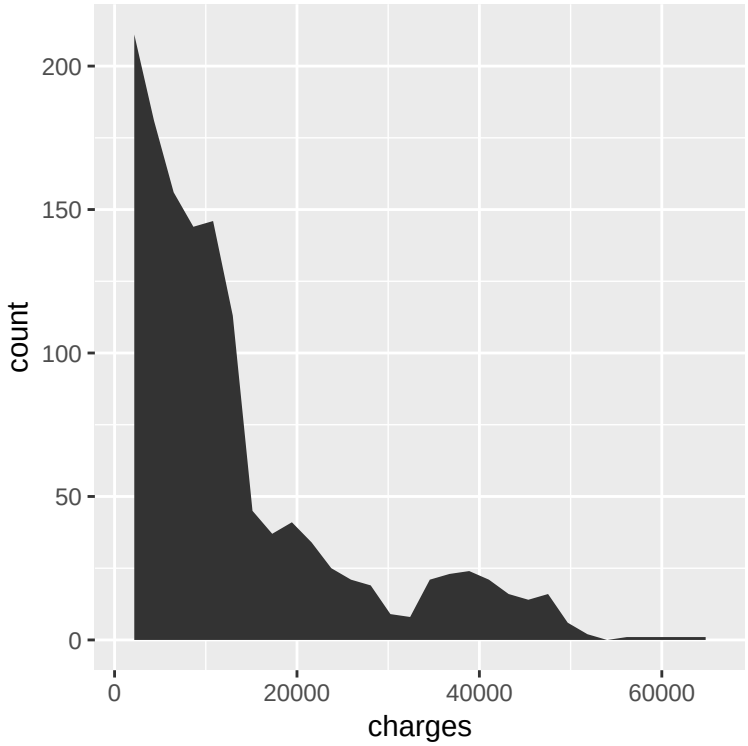


Region

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

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

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



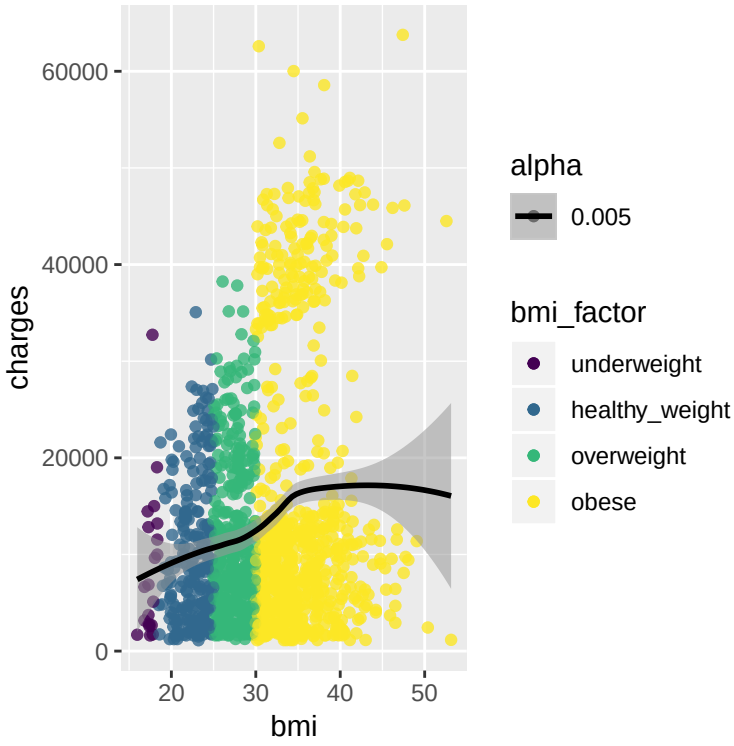
Charges

- Skewed right

Multivariable analysis

Relationships between multiple variables with anecdotal notes

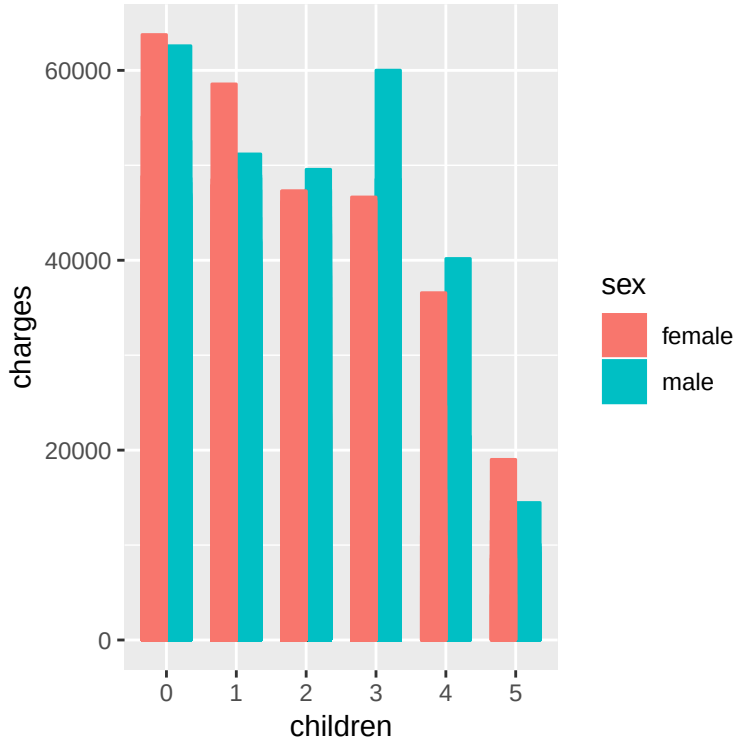
```
ggplot(health_charges_clean,
       aes(x = bmi, y = charges, color = bmi_factor, alpha = .005 ))+
  geom_point() +
  geom_jitter() +
  geom_smooth (method = "loess", color = "black")
```



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.

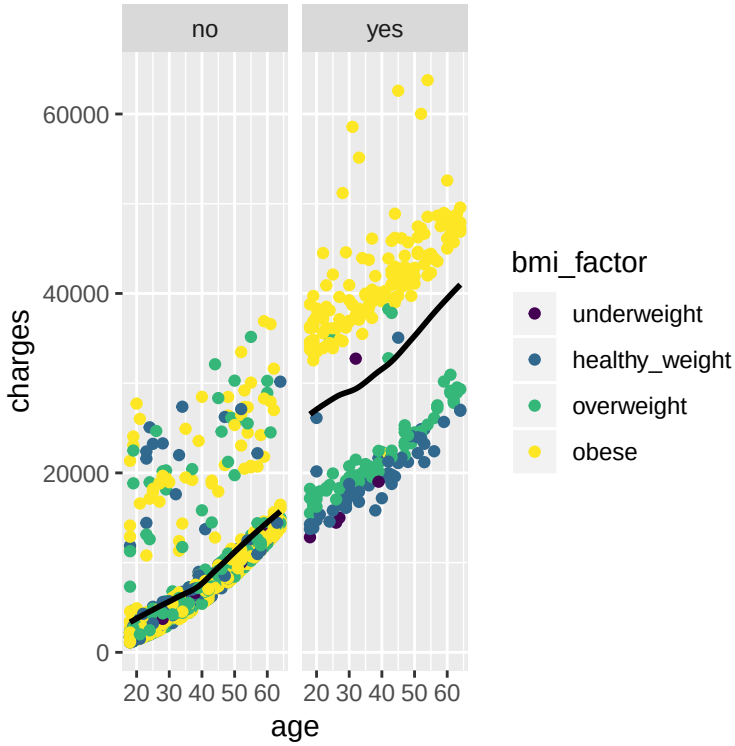
```r
health_charges_clean$children <- as.factor(health_charges_clean$children)

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.

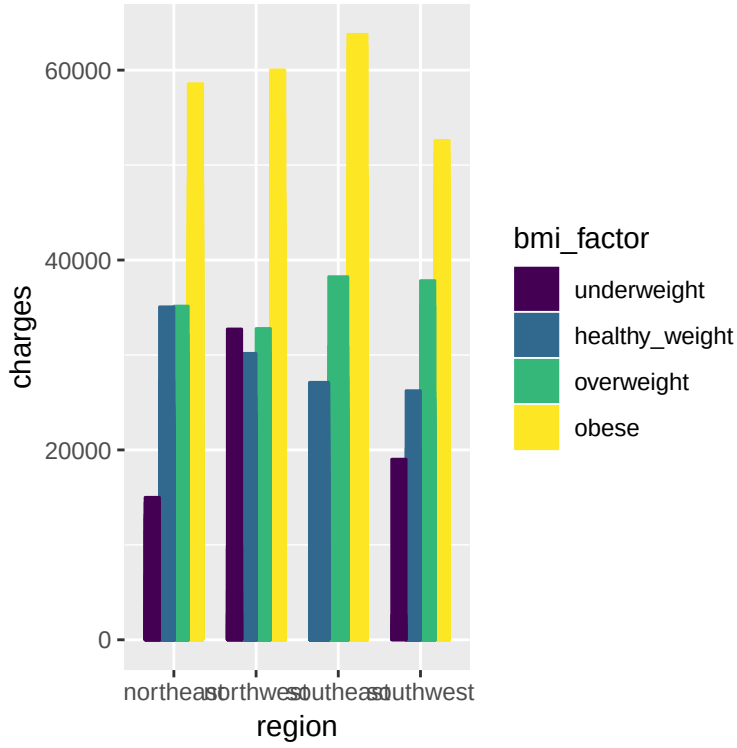
```
ggplot(health_charges_clean, aes(x = age, y = charges, color = bmi_factor), alpha = .02, size = .02) +
  geom_point(aes(color = bmi_factor, fill = bmi_factor))+
  facet_grid( . ~ smoker)+
  geom_smooth(se = FALSE, method = "loess", weight = .005, color = "black", alpha = .02 )
```



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.