Data Story

Julia Sheriff

9/22/2018

INTRODUCTION

Health insurance companies must create plans that effectively ensure their clients and maximize profits. Because health charges vary from person to person, it is difficult for insurance companies to design insurance plans which collectively maximize profits. Collectively, the insurance company must charge clients more than the cost of covered health charges. We can therefore use individual health charges as an estimate for insurance charges.

We can begin to estimate individual charges by comparing charges between different clusters of individuals (sex, region, number of children, bmi, region, age).

By understanding the relationship between charges and these variables, insurance companies can do the following:
* Predict their charges as their population changes over time.
* Examine how to provide reimbursement for health services which could make their population less costly.
* Determine the most locations of the most profitable populations and how to increase clients from that area. Example: Choosing a location for an HMO.

OVERVIEW OF THE DATASET

Health Variables:

The dataset is available at https://www.kaggle.com/mirichoi0218/insurance/home.

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

```
health_charges <- read.csv("capstone_data.csv", header = TRUE)
head(health_charges)
```

```
##   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
```

```
str(health_charges)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

CONSIDERATIONS

While we have data on seven variables in our observations, there are other factors which could impact health charges.
* income of individual
* education level
* employment status
* location: urban, suburban, rural
* chronic health conditions
* muscle / fat ratio (in addition to BMI which just compares weight to height)

There are also other factors that would be useful in interpreting the charges themselves: * breakdown of charges
* charges from previous years

DATA CLEANING

I created a new variable, BMI_factor, which treats the numerical BMI variable as a factor variable of four categories: underweight, healthy weight, overweight, obese. These are standard categories used by agencies such as the CDC. I created this variable to give a general meaning to the numerical variable, bmi.

```
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
```

```
health_charges <- health_charges %>%
  mutate(bmi_factor = if_else ( bmi < 18.5, "underweight",
         if_else ( bmi >= 18.5 & bmi < 25, "healthy_weight",
         if_else ( bmi >= 25 & bmi < 30, "overweight",
         if_else ( bmi >= 30, "obese", NA_character_)))))

health_charges$bmi_factor <- factor(health_charges$bmi_factor,
```

```
                      levels = c("underweight", "healthy_weight", "overweight", "obese"),
                      ordered = TRUE)

health_charges <- health_charges[ , c(1:3, 8, 4:7)]

health_charges_clean <- health_charges

head(health_charges_clean)
```

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

I assessed the data for missing values and nonsensical outliers, and the data was clean. The data was tidy because each row represents and observation and each column represents a variable.

```
summary(health_charges == "")
summary(is.na.data.frame(health_charges))

unique(health_charges[,1])
unique(health_charges[,2])
unique(health_charges[,3])
unique(health_charges[,4])
unique(health_charges[,5])
unique(health_charges[,6])
unique(health_charges[,7])
unique(health_charges[,8])

head(sort(health_charges$bmi), n=25)
tail(sort(health_charges$bmi), n=25)
head(sort(health_charges$charges), n=25)
head(sort(health_charges$charges), n=25)
```

INITIAL FINDINGS

UNIVARIATE ANALYSIS

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

SEXES
* Even distribution

BMI and BMI_FACTOR
* 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.

CHILDREN
* The data is skewed right.

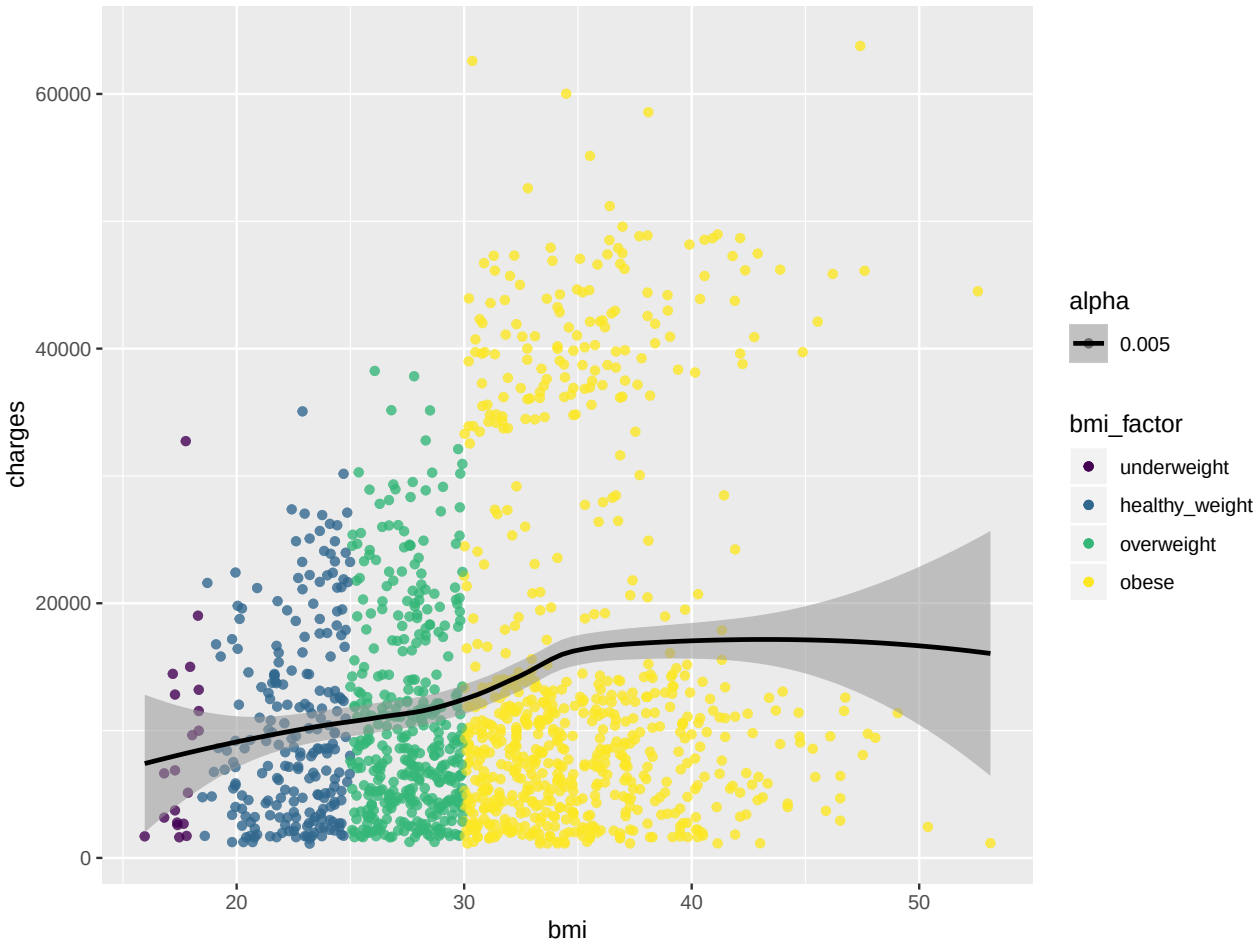
SMOKER
* The ratio of non-smokers to smokers is approximately 4 : 1
REGION
* All regions except southeast had between 324-325 observations. * Perhaps cluster sampling was used for data collection.
CHARGES
* SHAPIRO.TEST
+ HO: Charges frequency follows a normal distribution.
+ HA: Charges frequency does not follow a normal distribution.
+ RESULTS:
- P-Value: $<$ 2.2e-16 $<$ .05
- Reject HO.
- Evidence supports the claim that charges frequency does not follow a normal distribution.

```
shapiro.test(health_charges_clean$charges)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  health_charges_clean$charges
## W = 0.81469, p-value < 2.2e-16
```

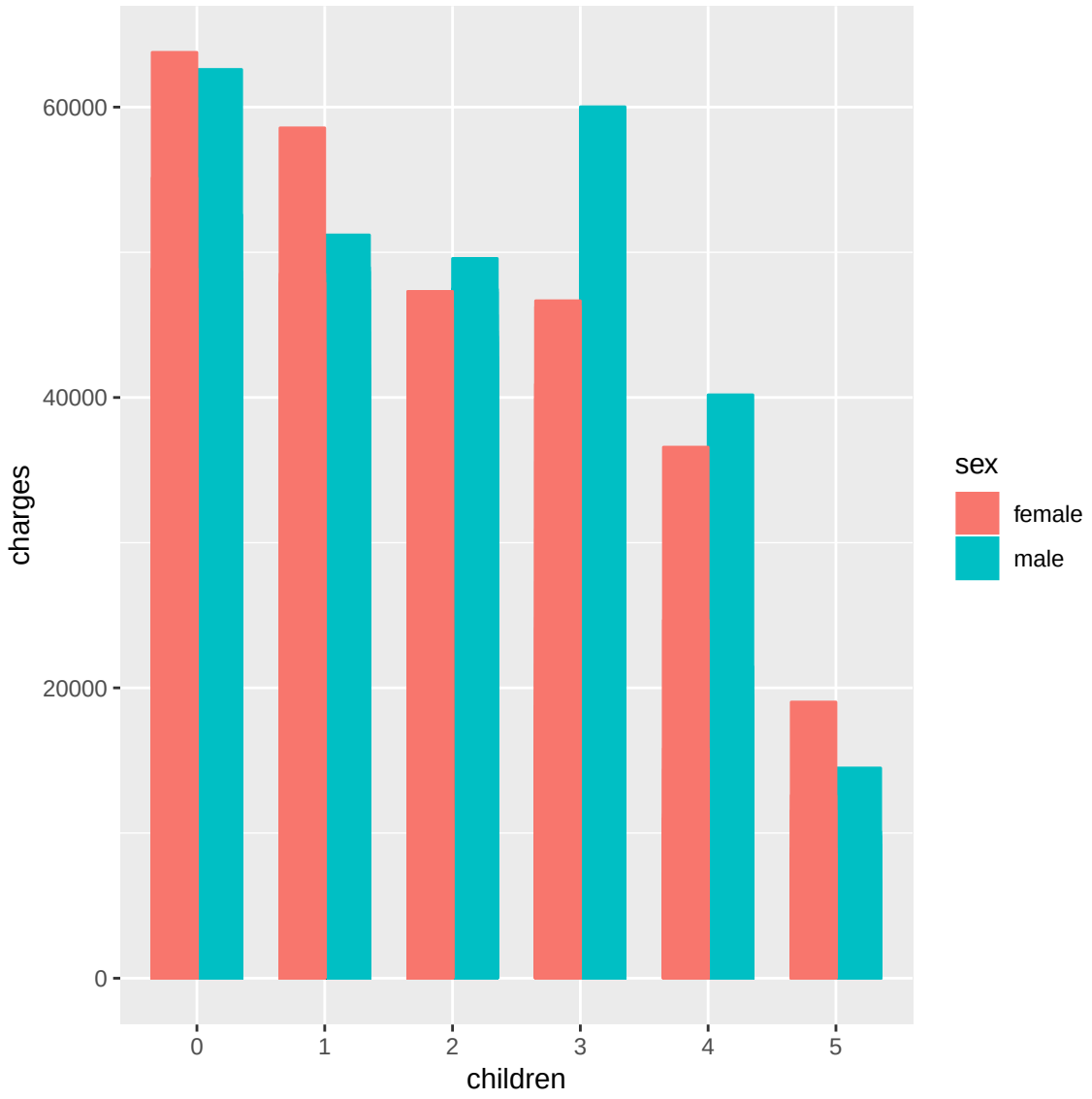
MULTIVARIATE ANALYSIS

Relationships between multiple variables with anecdotal notes



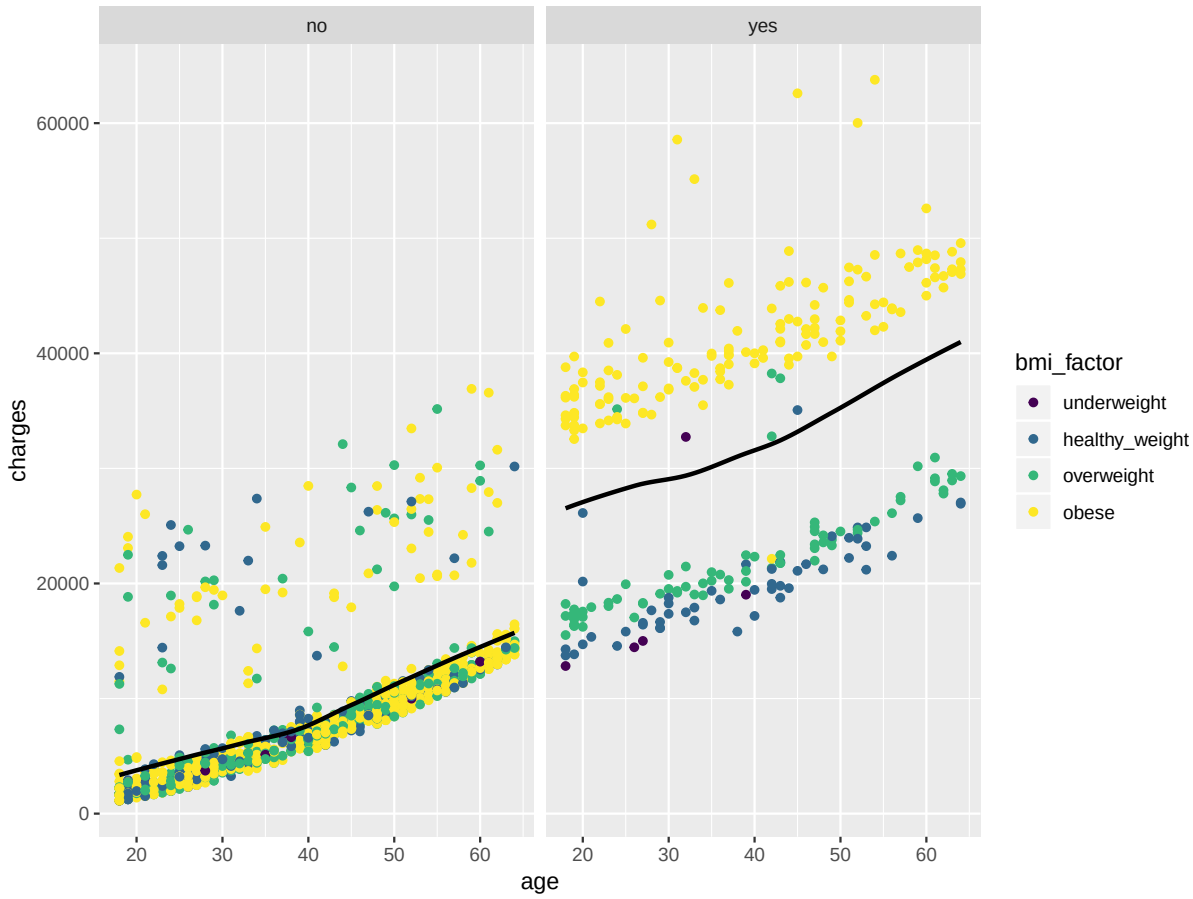
Effect of BMI on charges

- 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.



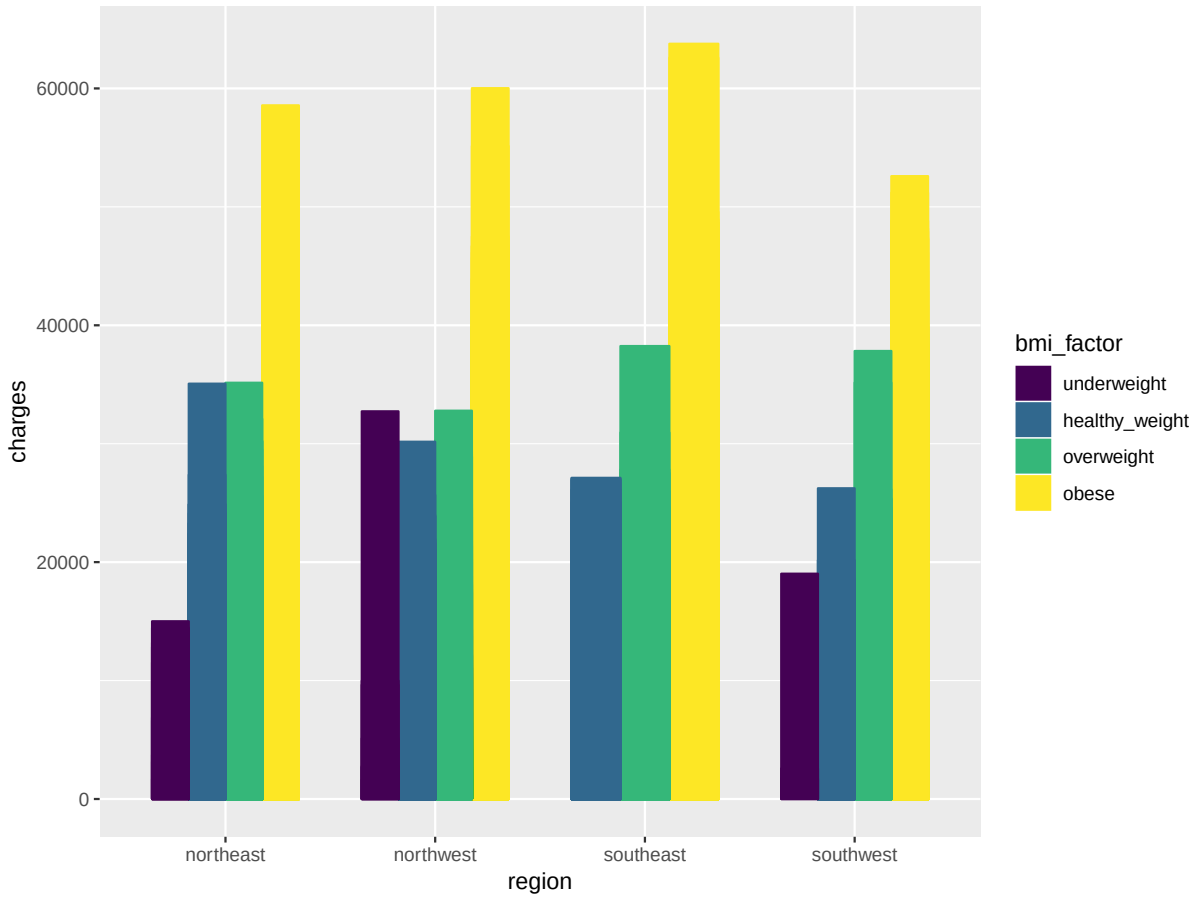
Effect of children on charges, considering sex

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



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.



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.

EXPLORATION OF OUTLIER REMOVAL

```
library(ggplot2)

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

swdf <- health_charges_clean[which(health_charges_clean$region == "southwest") , ]

swmoddf <- swdf
swmod <- swmoddf$charges
qnt <- quantile(swmod, probs=c(.25, .75), na.rm = T)
H <- 1.5 * IQR(swmod, na.rm = T)
swmod[swmod < (qnt[1] - H)] <- (qnt[1] - H)
swmod[swmod > (qnt[2] + H)] <- (qnt[2] + H)
swmoddf$charges <- swmod

min(swdf$charges)
```

```
## [1] 1241.565
```

```
min(swmoddf$charges)
```

```
## [1] 1241.565
```

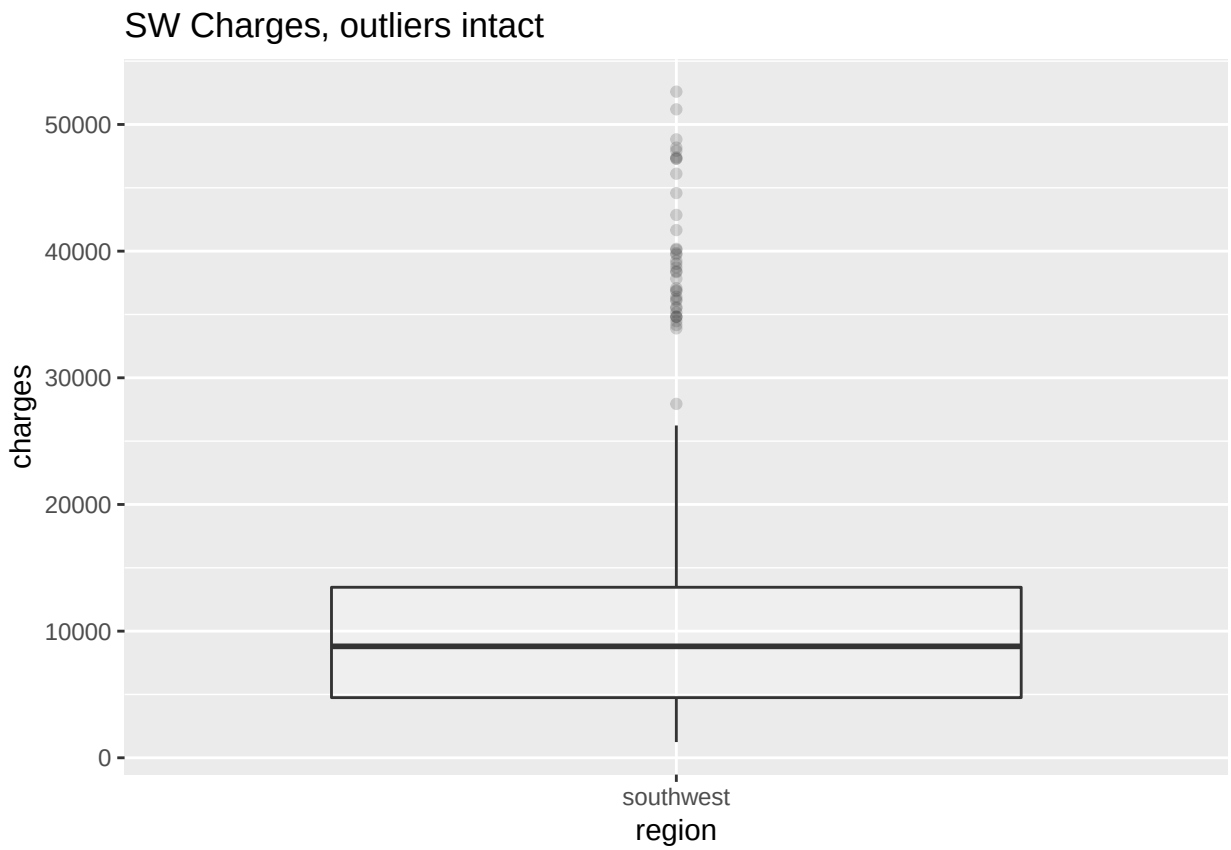
```
max(swdf$charges)
```

```
## [1] 52590.83
```

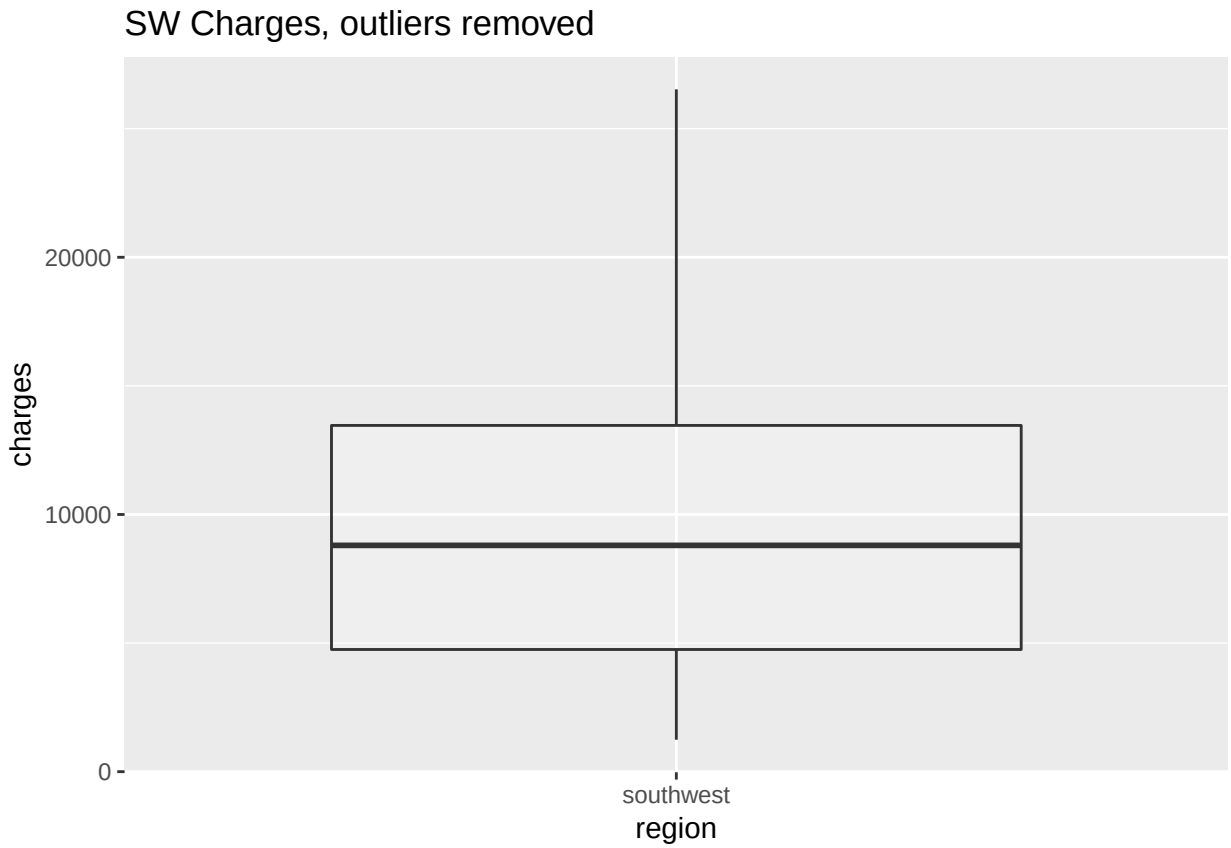
```
max(swmoddf$charges)
```

```
## [1] 26529.69
```

```
ggplot(swdf, aes(x = region, y = charges)) +
  geom_boxplot( alpha = .2) +
  ggtitle("SW Charges, outliers intact")
```




```
ggplot(swmoddf, aes(x = region, y = charges)) +
         geom_boxplot( alpha = .2) +
         ggtitle("SW Charges, outliers removed")
```




```
nwdf <- health_charges_clean[which(health_charges_clean$region == "northwest") , ]

nwmoddf <- nwdf
nwmod <- nwmoddf$charges
qnt <- quantile(nwmod, probs=c(.25, .75), na.rm = T)
H <- 1.5 * IQR(nwmod, na.rm = T)
nwmod[nwmod < (qnt[1] - H)] <- (qnt[1] - H)
nwmod[nwmod > (qnt[2] + H)] <- (qnt[2] + H)
nwmoddf$charges <- nwmod

min(nwdf$charges)
```

```
## [1] 1621.34
```

```
min(nwmoddf$charges)
```

```
## [1] 1621.34
```

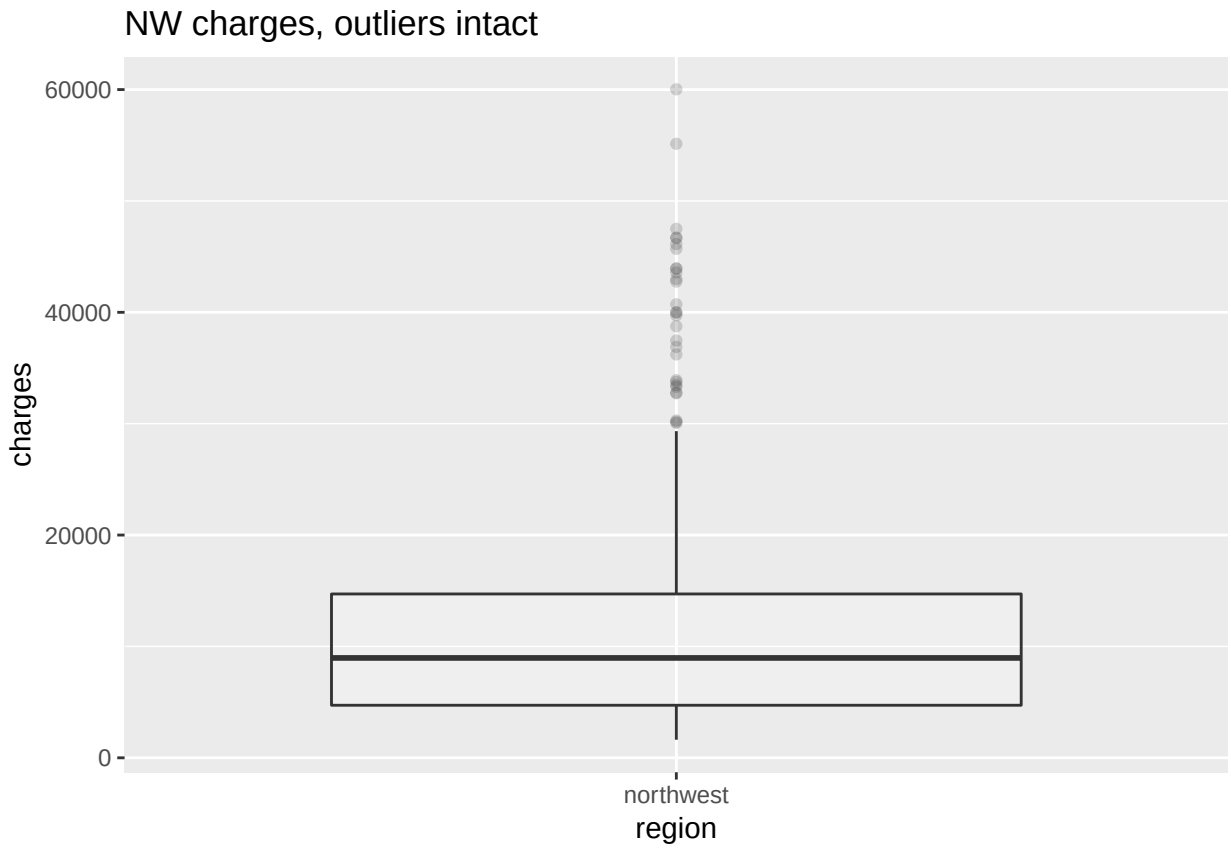
```
max(nwdf$charges)
```

```
## [1] 60021.4
```

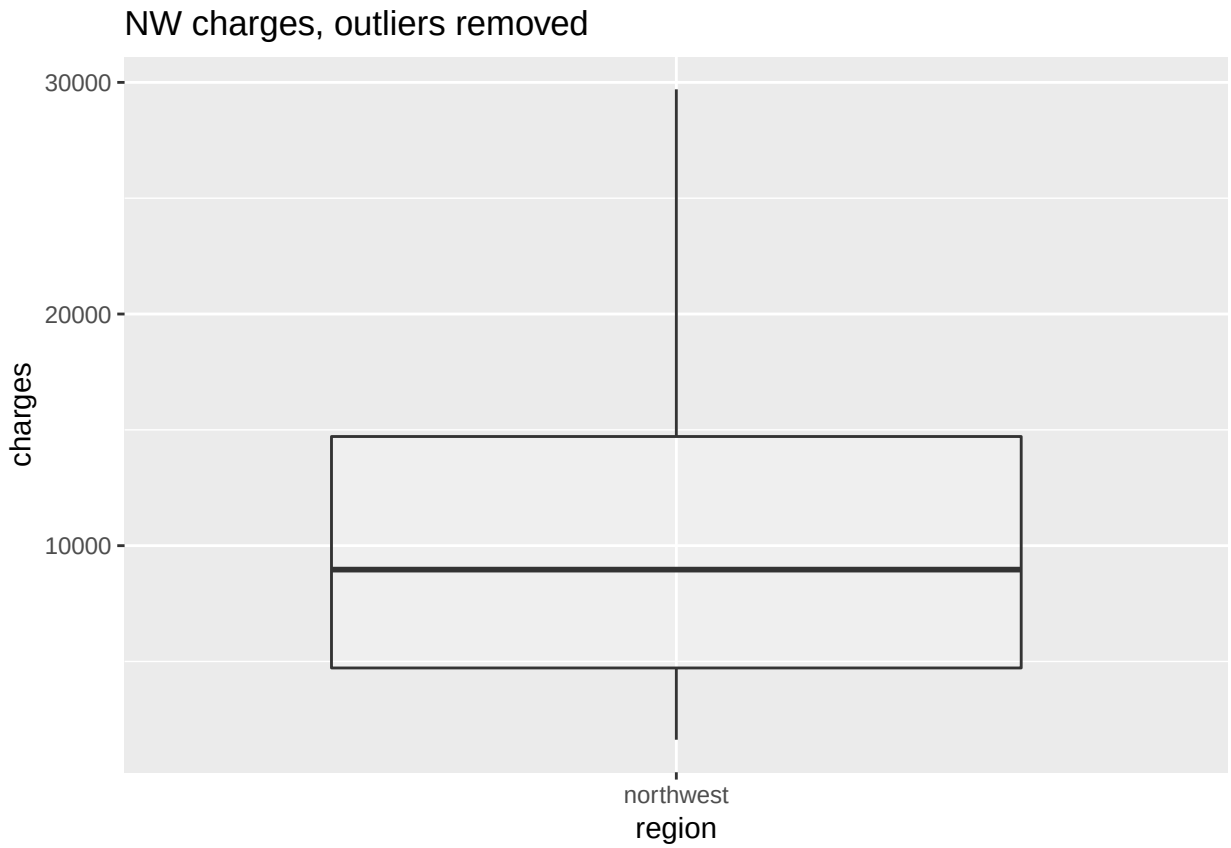
```
max(nwmoddf$charges)
```

```
## [1] 29699.75
```

```
ggplot(nwdf, aes(x = region, y = charges)) +
  geom_boxplot( alpha = .2) +
  ggtitle("NW charges, outliers intact")
```

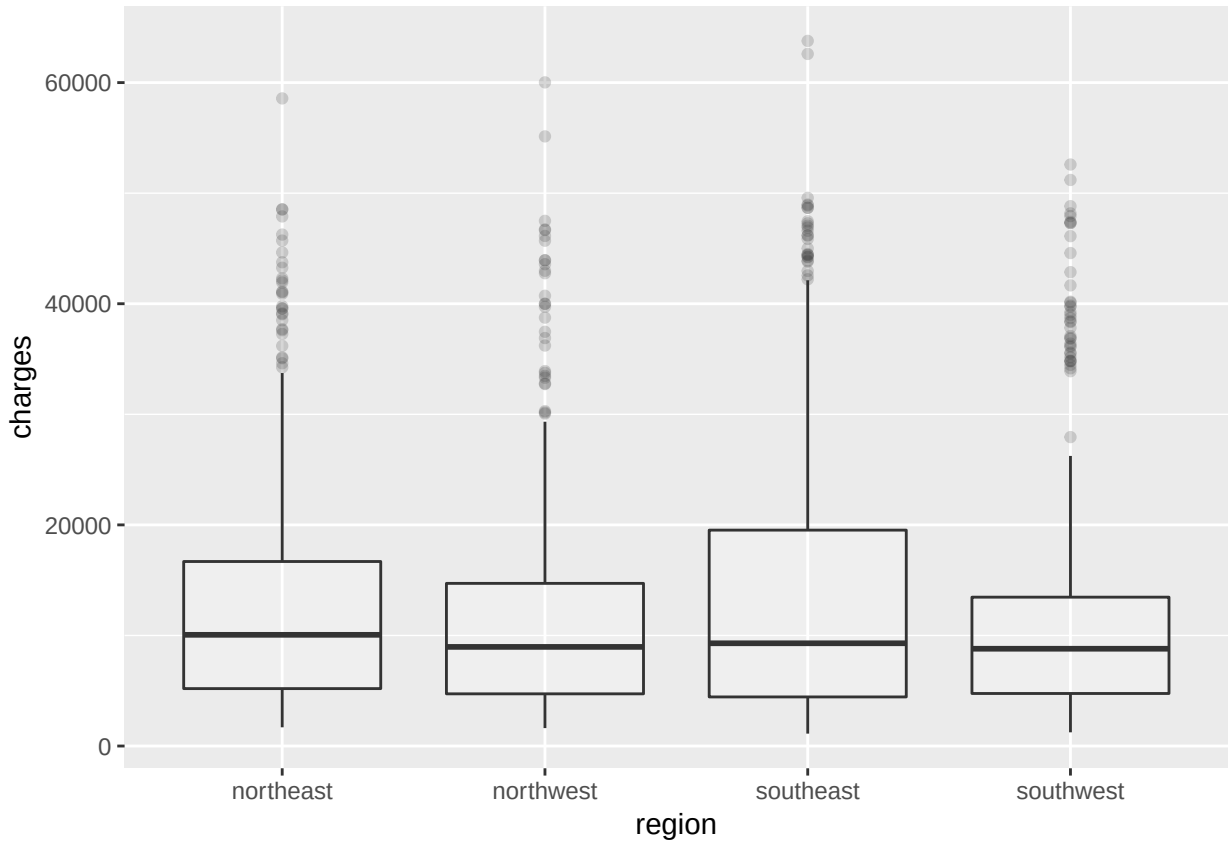



```
ggplot(nwmoddf, aes(x = region, y = charges)) +
         geom_boxplot(alpha = .2) +
         ggtitle("NW charges, outliers removed")
```

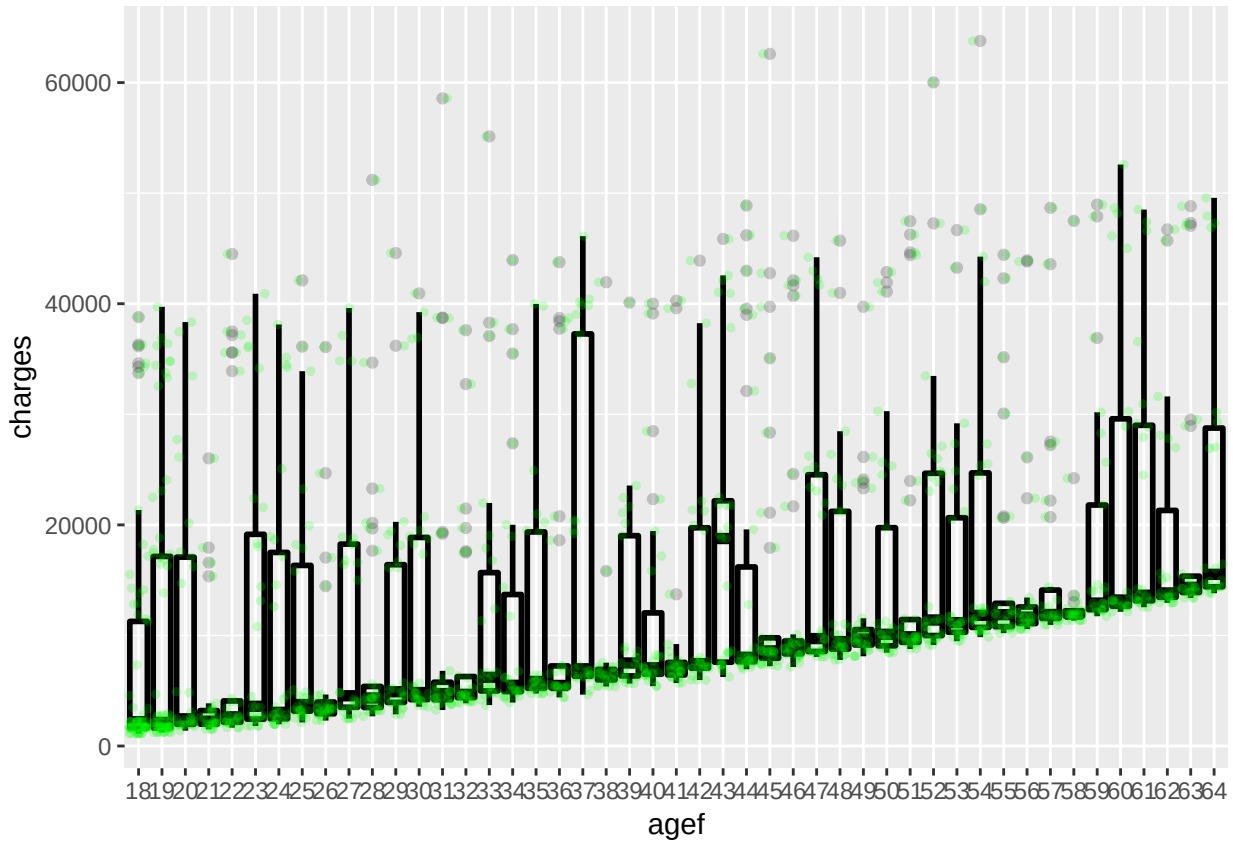


****BOXPLOTS SHOWING OUTLIERS**

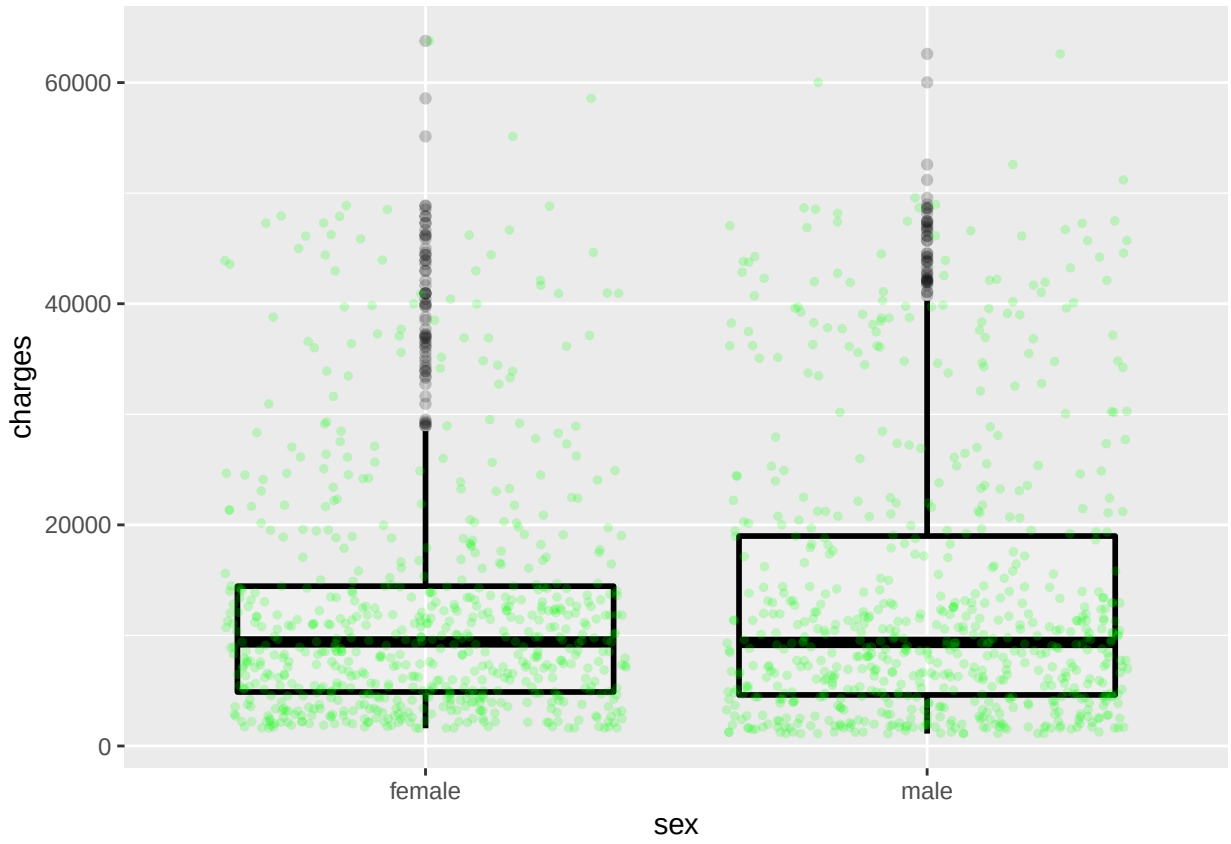
```
ggplot(health_charges_clean, aes(x = region, y = charges))+
  geom_boxplot( alpha = .2)
```



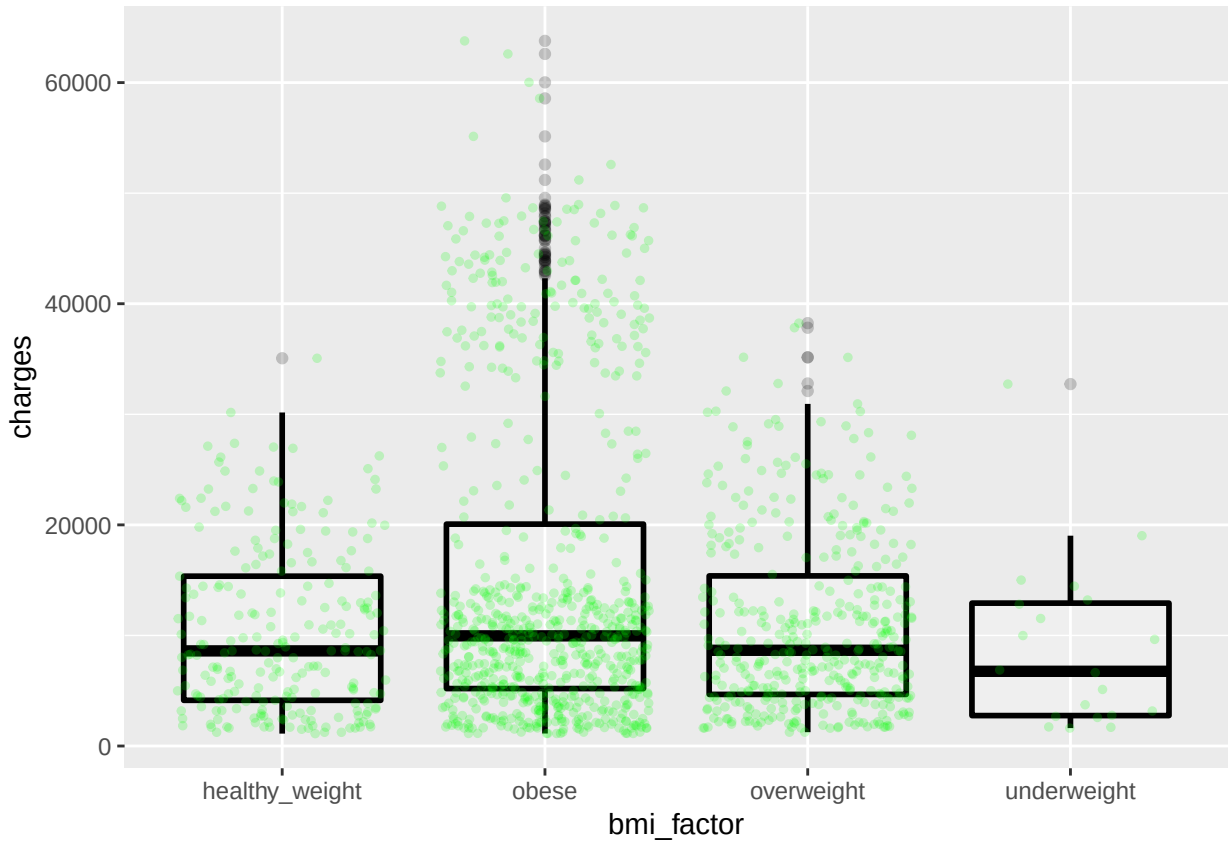

```
agef <- as.factor(health_charges_clean$age)
ggplot(health_charges_clean, aes(x = agef, y = charges))+
geom_boxplot(color = "black", size = 1, alpha = .2) +
geom_jitter(color = "green", size = 1, alpha = .2)
```




```
ggplot(health_charges_clean, aes(x = sex, y = charges))+
geom_boxplot(color = "black", size = 1, alpha = .2) +
geom_jitter(color = "green", size = 1, alpha = .2)
```

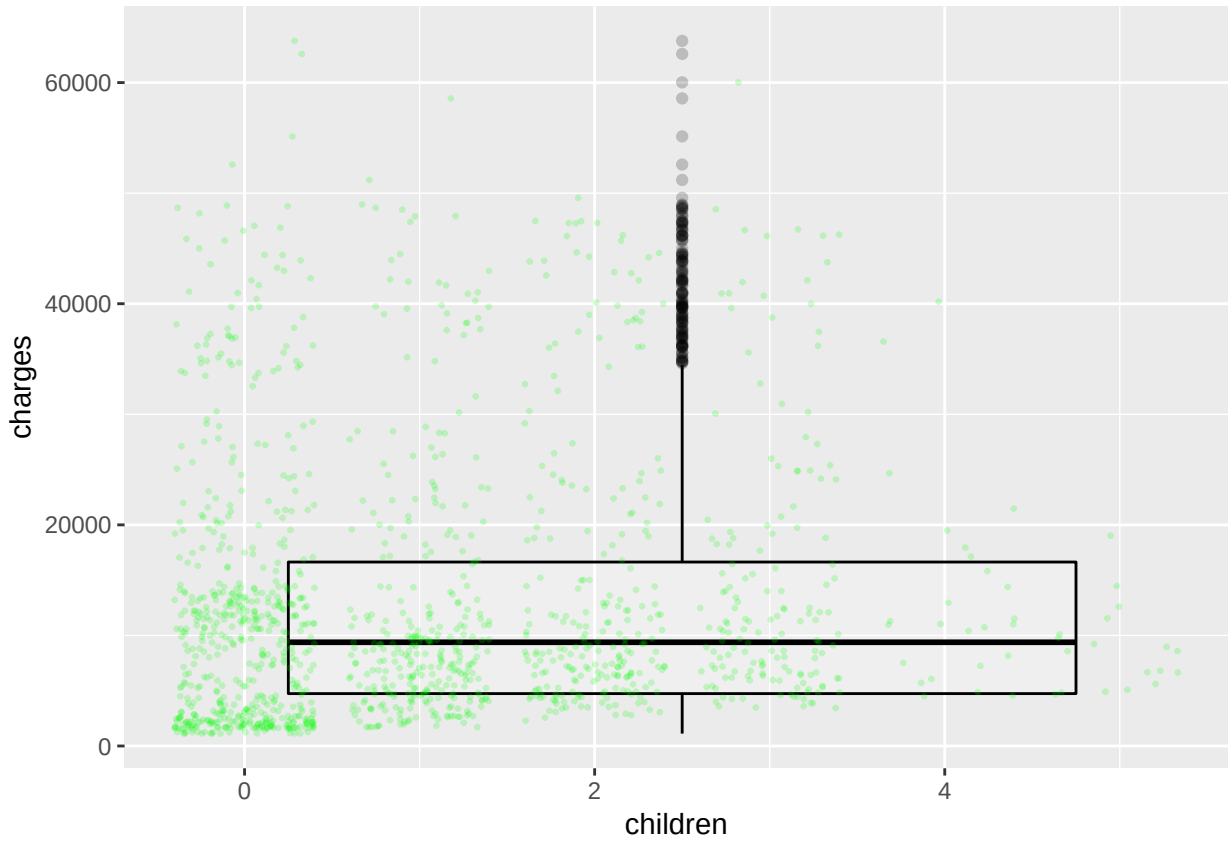



```
ggplot(health_charges_clean, aes(x = bmi_factor, y = charges))+
geom_boxplot(color = "black", size = 1, alpha = .2) +
geom_jitter(color = "green", size = 1, alpha = .2)
```

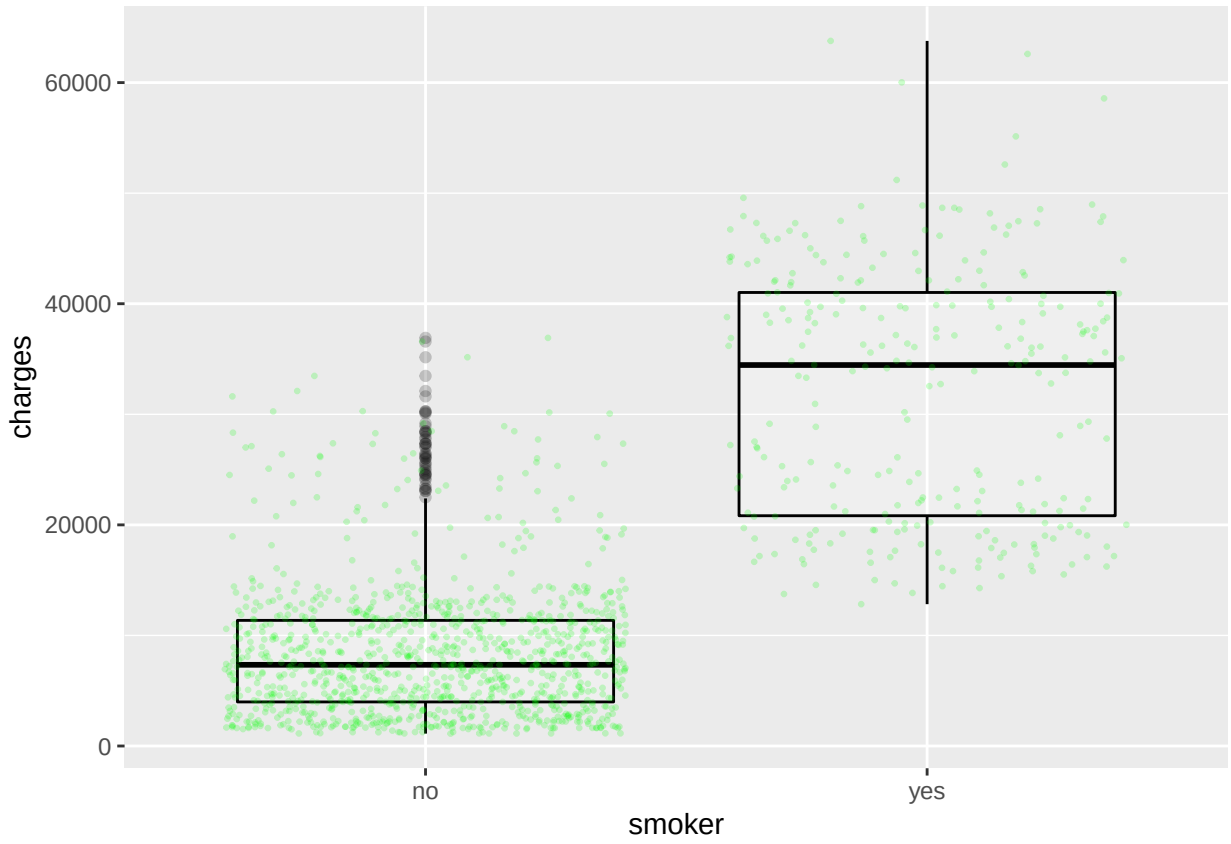



```
childrenf <- as.factor(health_charges_clean$children)
ggplot(health_charges_clean, aes(x = children, y = charges))+
geom_boxplot(color = "black", alpha = .2) +
geom_jitter(color = "green", size = .5, alpha = .2)
```

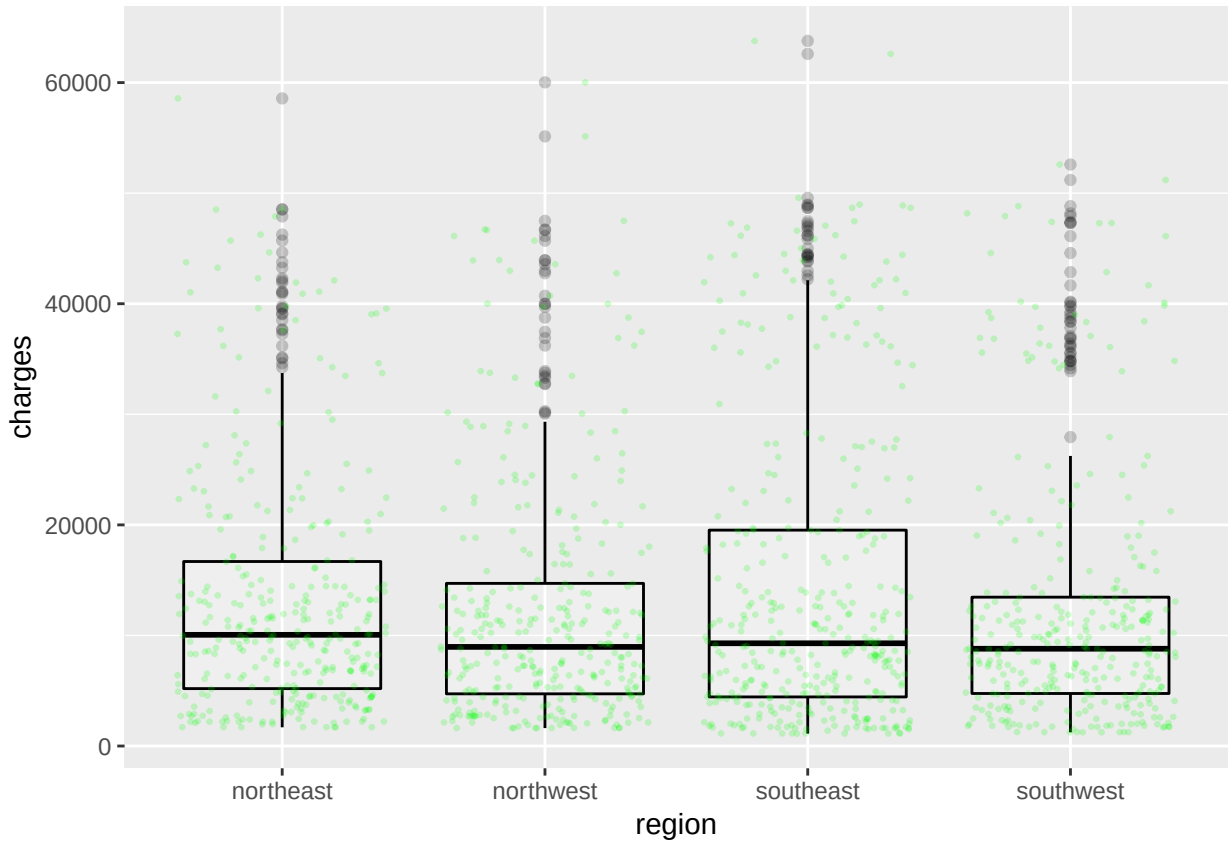
```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



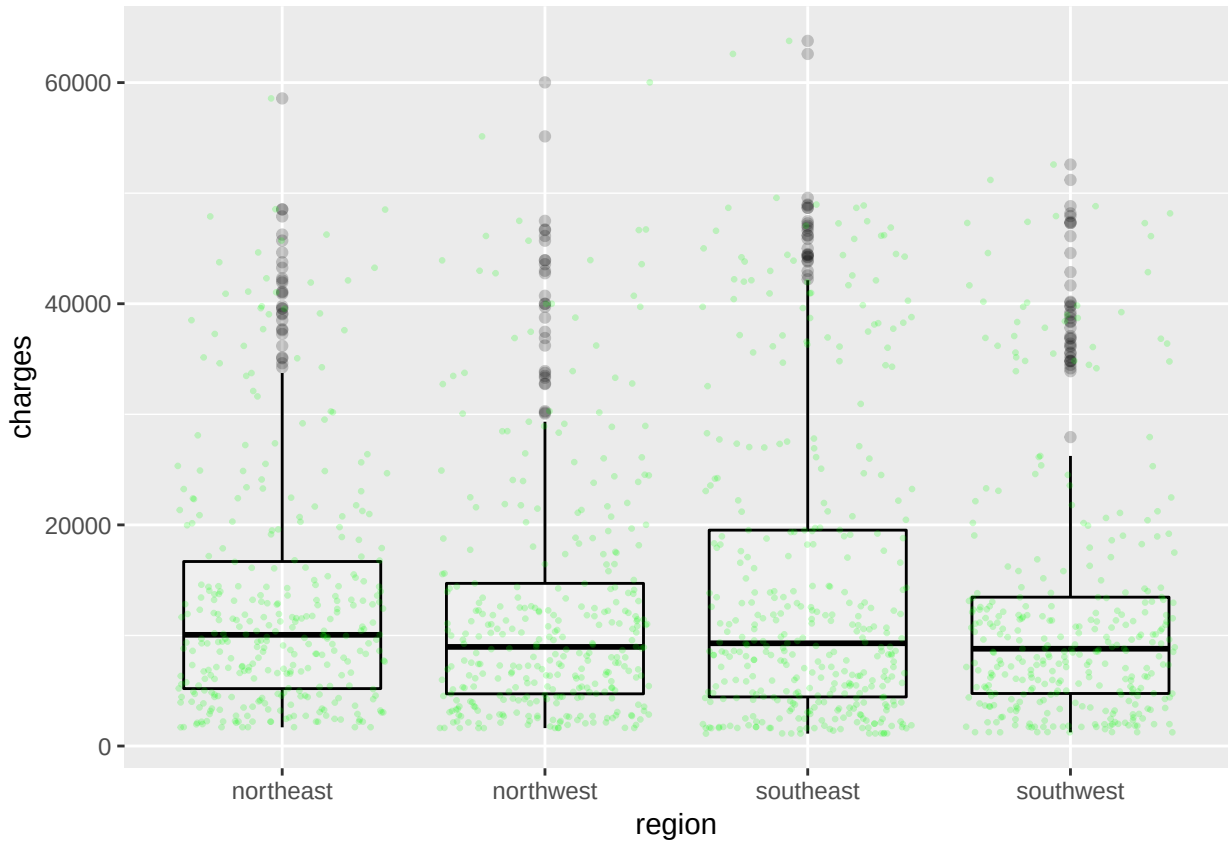

```
ggplot(health_charges_clean, aes(x = smoker, y = charges))+
geom_boxplot(color = "black", alpha = .2) +
geom_jitter(color = "green", size = .5, alpha = .2)
```




```
ggplot(health_charges_clean, aes(x = region, y = charges))+
geom_boxplot(color = "black", alpha = .2) +
geom_jitter(color = "green", size = .5, alpha = .2)
```




```
ggplot(health_charges_clean, aes(x = region, y = charges))+
geom_boxplot(color = "black", alpha = .2) +
geom_jitter(color = "green", size = .5, alpha = .2)
```



APPROACH

- The approach has not changed. I will apply statistical analysis and machine learning to the data to study the relationship of different variables on charges.