# Health Charges

*Julia Sheriff*

*11/6/2018*

## INTRODUCTION

With advances in health analytics, we are better to assess relationships between various health conditions, treatments, and costs. This dataset describes health charges, sex, region, number of children, bmi, region, and age over 1338 observations. Our variable of interest is health charges. 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 to design insurance plans which collectively maximize profits. Our goal is to estimate insurance charges base on an individual's various characteristics.

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.

---

## THE DATA

### 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 |

```r
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
```

```r
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 ...
```

## CAVEATS

*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 compares weight to height)

*There are also other factors that would be useful in interpreting the charges themselves:*
* breakdown of charges for the following: + urgent care + preventative care + medication

## DATA CLEANING

- 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.

- I created factor variables for age, bmi, and charges to allow for categorical studies on those continuous variables.

- I also created dummy variables for all facets in order to run clustering algorhithms on the dataset.

```r
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)
```

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

```r
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
```

```r
#age variables:
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 & bmi < 35, "obese_1",
         if_else (bmi >= 35 & bmi < 40, "obese_2",
         if_else (bmi >= 40, "obese_3", NA_character_)))))))

health_charges$bmi_factor <- factor(health_charges$bmi_factor,
                      levels = c("underweight", "healthy_weight", "overweight", "obese_1", "obese_2", "
                      ordered = TRUE)

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

#health charges split into "high" and "low"
vquantile <- as.vector(quantile(health_charges$charges))
hcut <- vquantile[c(4)]
health_charges <- health_charges %>% mutate(charges_factor = if_else (charges < hcut, "low",
                                            if_else (charges >= hcut, "high", NA_character_)))

#age variables:
health_charges <- health_charges %>% mutate(age_factor = if_else( age < 20, "10s",
                                                  if_else ( age >= 20 & age < 30, "20s",
                                                  if_else ( age >= 30 & age < 40, "30s",
                                                  if_else ( age >= 40 & age < 50, "40s",
                                                  if_else ( age >= 50 & age < 60, "50s",
                                                  if_else ( age >= 60, "60s", NA_character_))))))
health_charges$age_factor <- factor(health_charges$age_factor,
                              levels = c("10s", "20s", "30s", "40s", "50s", "60s"),
                              ordered = TRUE)
health_charges$charges_factor <- factor(health_charges$charges_factor,
                          levels = c("low","high"),
                          ordered = TRUE)

health_charges_clean <- health_charges
str(health_charges_clean)
```

```
## 'data.frame':    1338 obs. of  10 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 ...
##  $ bmi_factor    : Ord.factor w/ 6 levels "underweight"<..: 3 4 4 2 3 3 4 3 3 3 ...
##  $ 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 ...
##  $ charges_factor: Ord.factor w/ 2 levels "low"<"high": 2 1 1 2 1 1 1 1 1 2 ...
##  $ age_factor    : Ord.factor w/ 6 levels "10s"<"20s"<"30s"<..: 1 1 2 3 3 3 4 3 3 6 ...
```

```
binary_charges <- health_charges_clean
colnames(binary_charges)
```

```
##  [1] "age"            "sex"            "bmi"            "bmi_factor"
##  [5] "children"       "smoker"         "region"         "charges"
##  [9] "charges_factor" "age_factor"
```

```
library(fastDummies)

binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "charges_factor")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "age_factor")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "bmi_factor")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "children")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "smoker")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "region")
binary_charges <- fastDummies::dummy_cols(binary_charges, select_columns = "sex")

#deleted non-binary columns:
binary_charges <- binary_charges[ c("charges_factor_high",
                                    "charges_factor_low",
                                    "bmi_factor_overweight",
                                    "bmi_factor_obese_1",
                                    "bmi_factor_healthy_weight",
                                    "bmi_factor_obese_2",
                                    "bmi_factor_obese_3",
                                    "bmi_factor_underweight",
                                    "children_0",
                                    "children_1",
                                    "children_3",
                                    "children_2",
                                    "children_5",
                                    "children_4" ,
                                    "smoker_yes",
                                    "smoker_no",
                                    "region_southwest",
                                    "region_southeast",
                                    "region_northwest",
                                    "region_northeast",
                                    "sex_female",
                                    "sex_male") ]

str(binary_charges)
```

```
## 'data.frame':    1338 obs. of  22 variables:
##  $ charges_factor_high      : int  1 0 0 1 0 0 0 0 0 1 ...
##  $ charges_factor_low       : int  0 1 1 0 1 1 1 1 1 0 ...
##  $ bmi_factor_overweight    : int  1 0 0 0 1 1 0 1 1 1 ...
##  $ bmi_factor_obese_1       : int  0 1 1 0 0 0 1 0 0 0 ...
##  $ bmi_factor_healthy_weight: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ bmi_factor_obese_2       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmi_factor_obese_3       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmi_factor_underweight   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ children_0               : int  1 0 0 1 1 1 0 0 0 1 ...
##  $ children_1               : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ children_3               : int  0 0 1 0 0 0 0 1 0 0 ...
##  $ children_2               : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ children_5               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ children_4               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smoker_yes               : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ smoker_no                : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ region_southwest         : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ region_southeast         : int  0 1 1 0 0 1 1 0 0 0 ...
##  $ region_northwest         : int  0 0 0 1 1 0 0 1 0 1 ...
##  $ region_northeast         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ sex_female               : int  1 0 0 0 0 1 1 1 0 1 ...
##  $ sex_male                 : int  0 1 1 1 1 0 0 0 1 0 ...
```

---

# EXPLORATORY DATA ANALYSIS

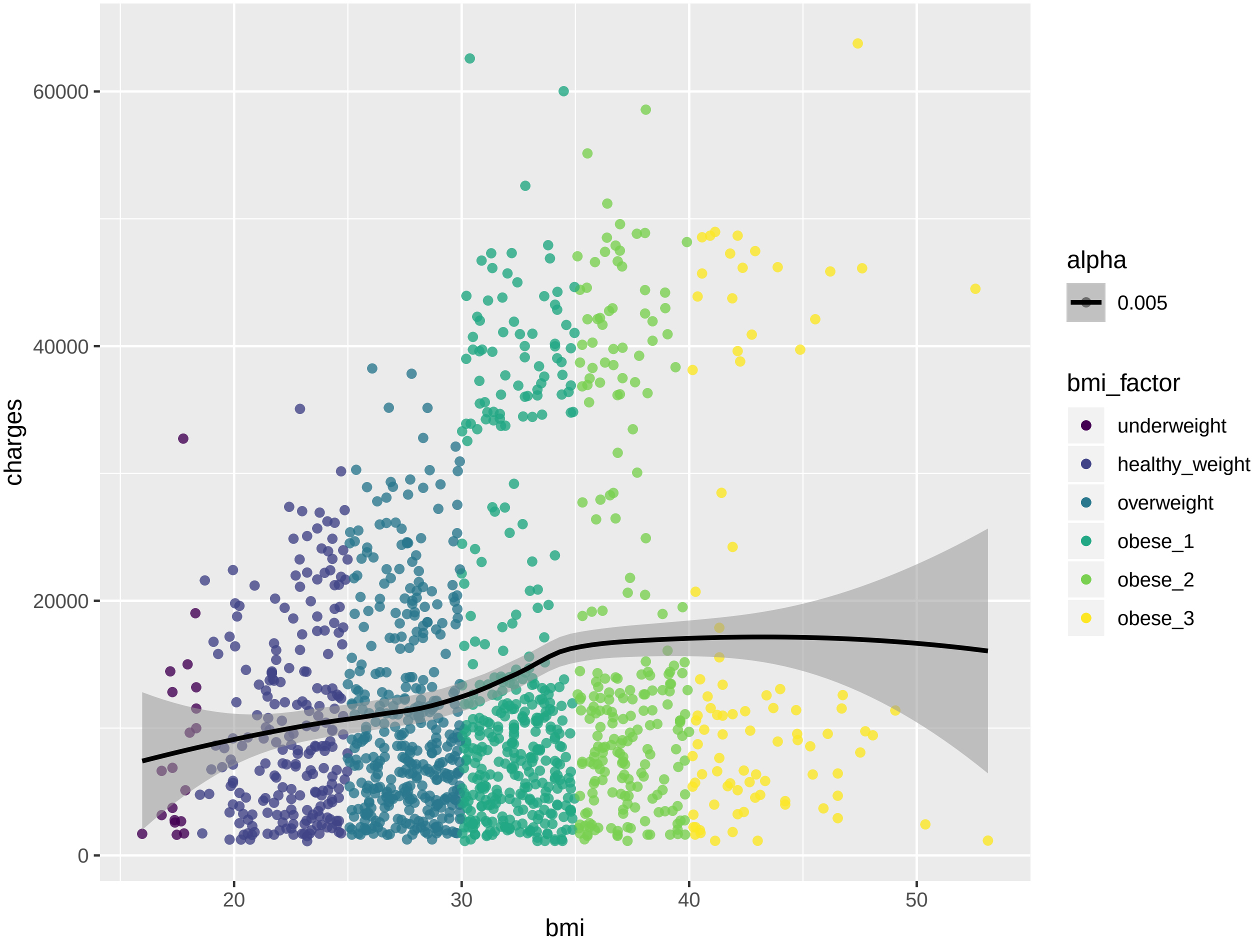
## UNIVARIATE ANALYSIS

**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)
```

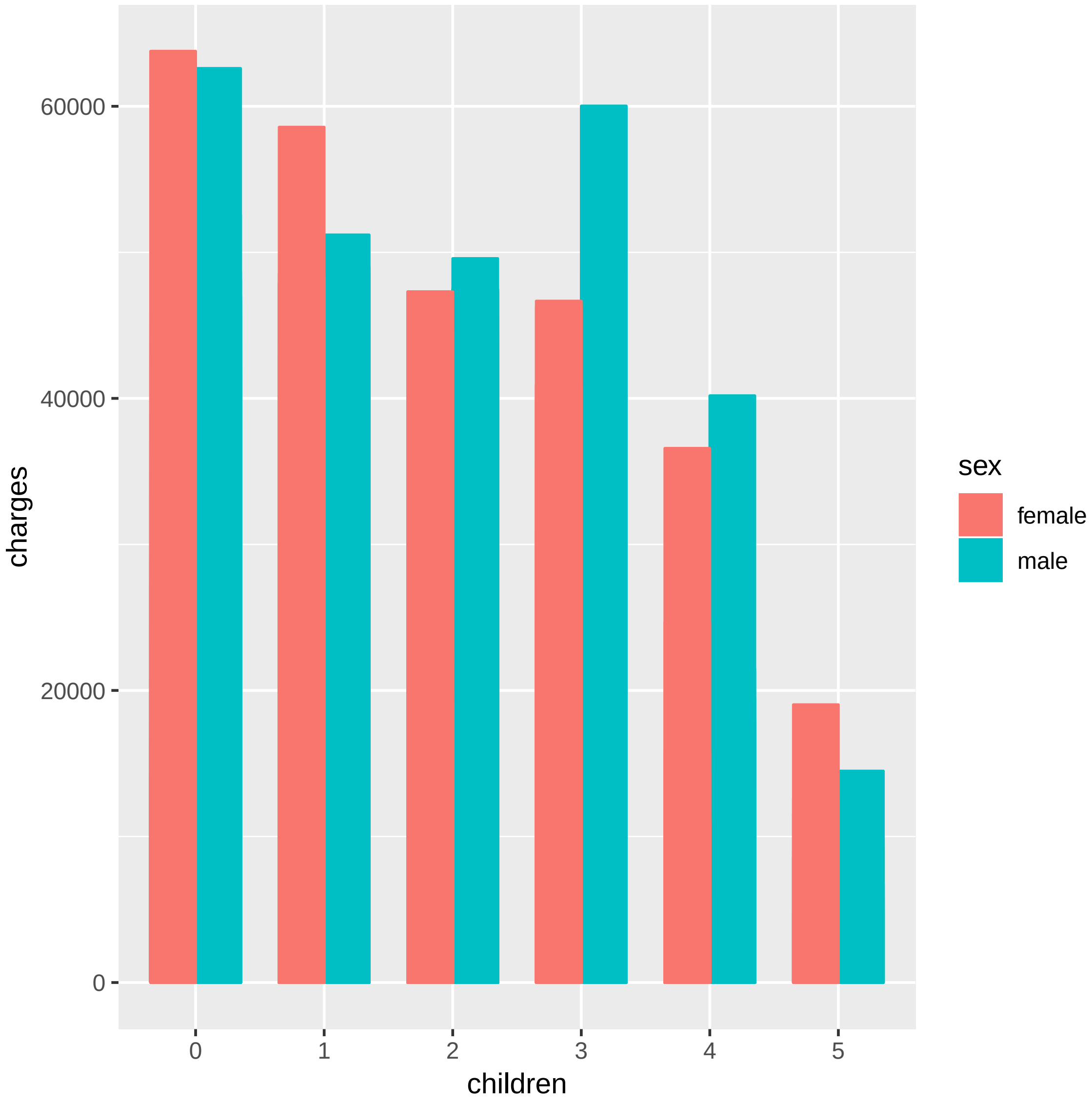
```
##
##  Shapiro-Wilk normality test
##
## data:  health_charges_clean$charges
## W = 0.81469, p-value < 2.2e-16
```

### MULTIVARIATE ANALYSIS



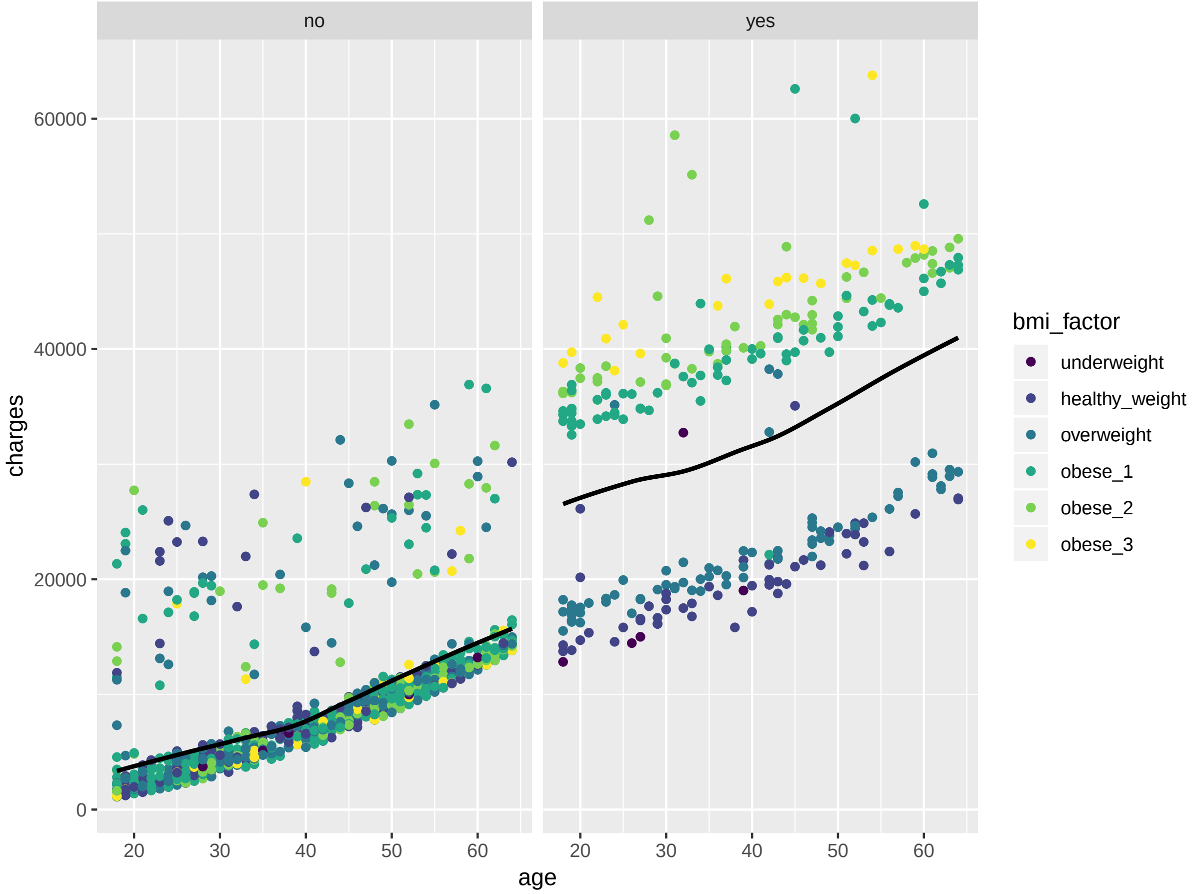
### Effect of BMI on charges

- 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.



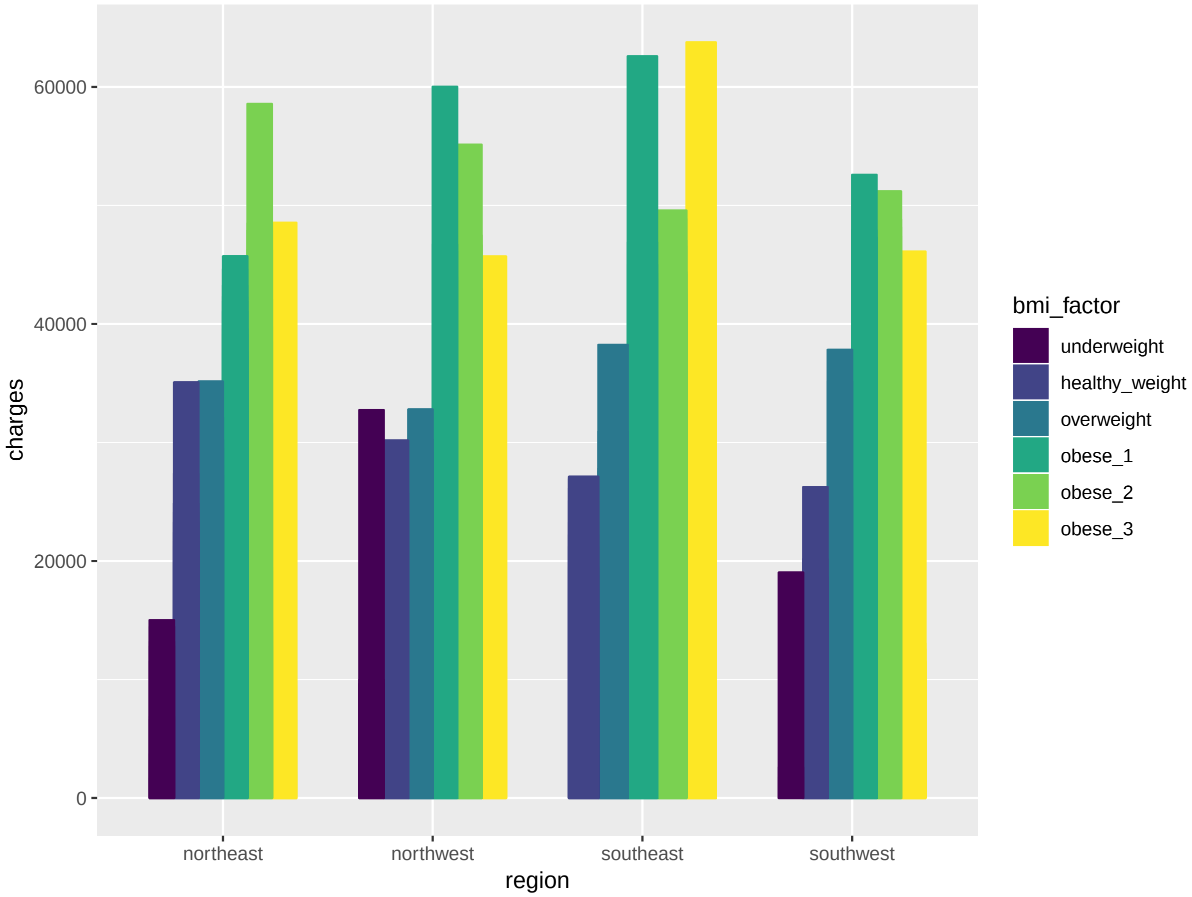
**Effect of children on charges, considering sex**

- 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.



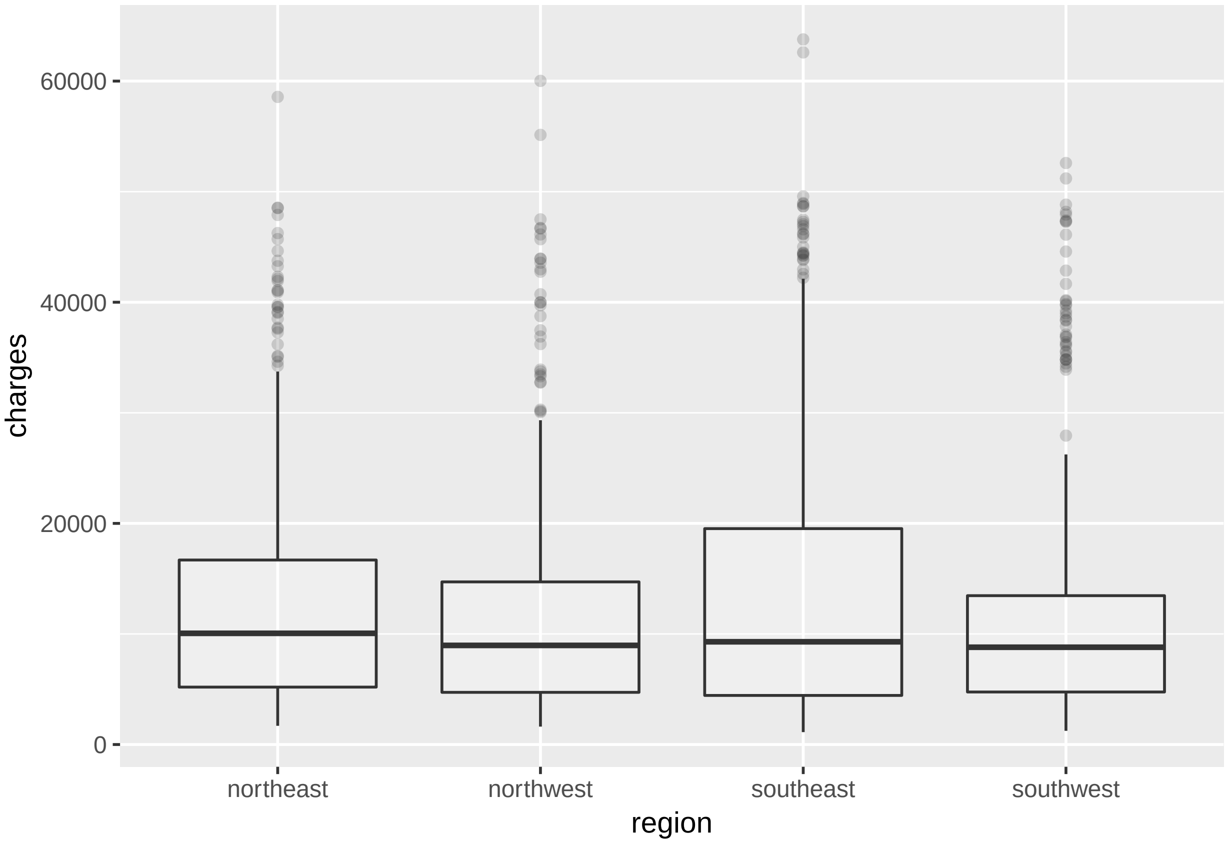
### Region's effect on charges, considering BMI

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

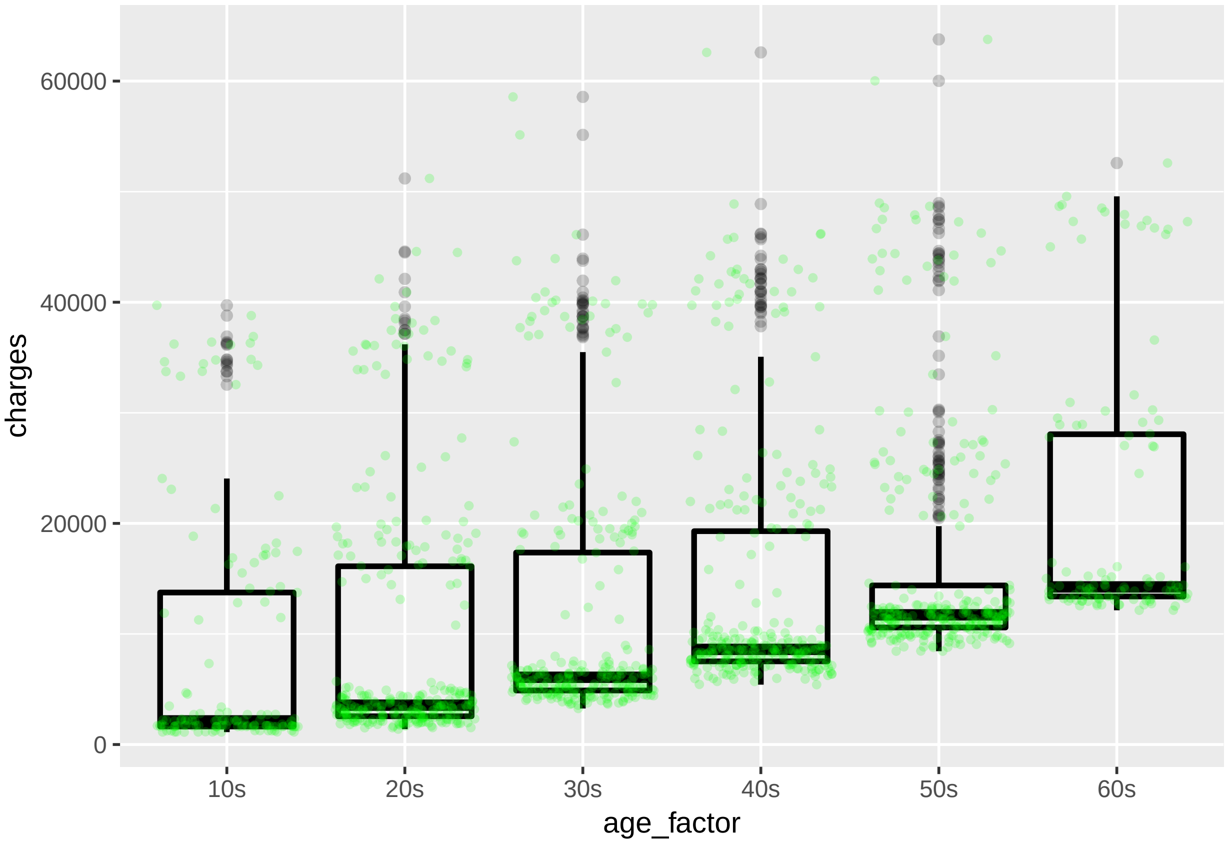
### OUTLIER EXPLORATION

I was curious to observe outliers for the dataset. None of the outliers seemed unreasonable. It is important to keep high charges in order to accurately assess the population charges as a whole.

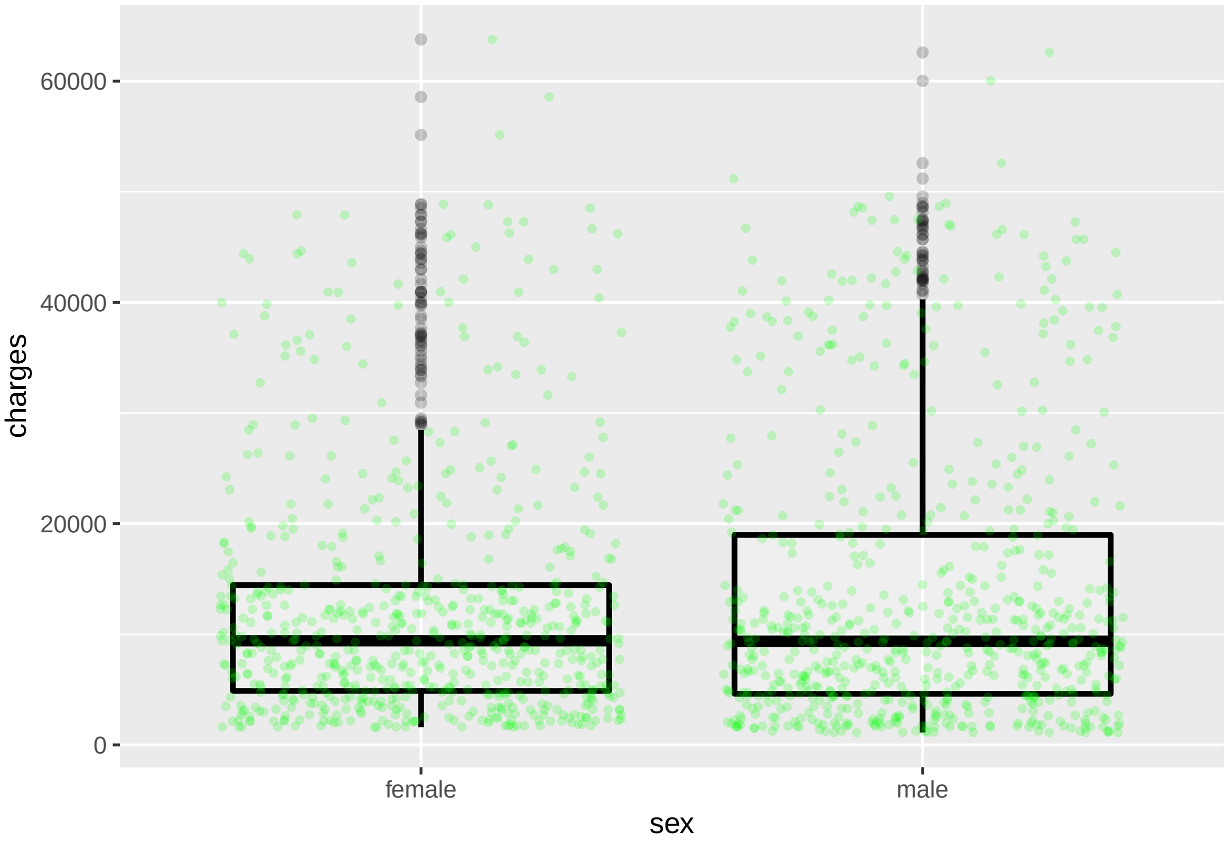
```r
ggplot(health_charges_clean, aes(x = region, y = charges))+
  geom_boxplot( alpha = .2)
```



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



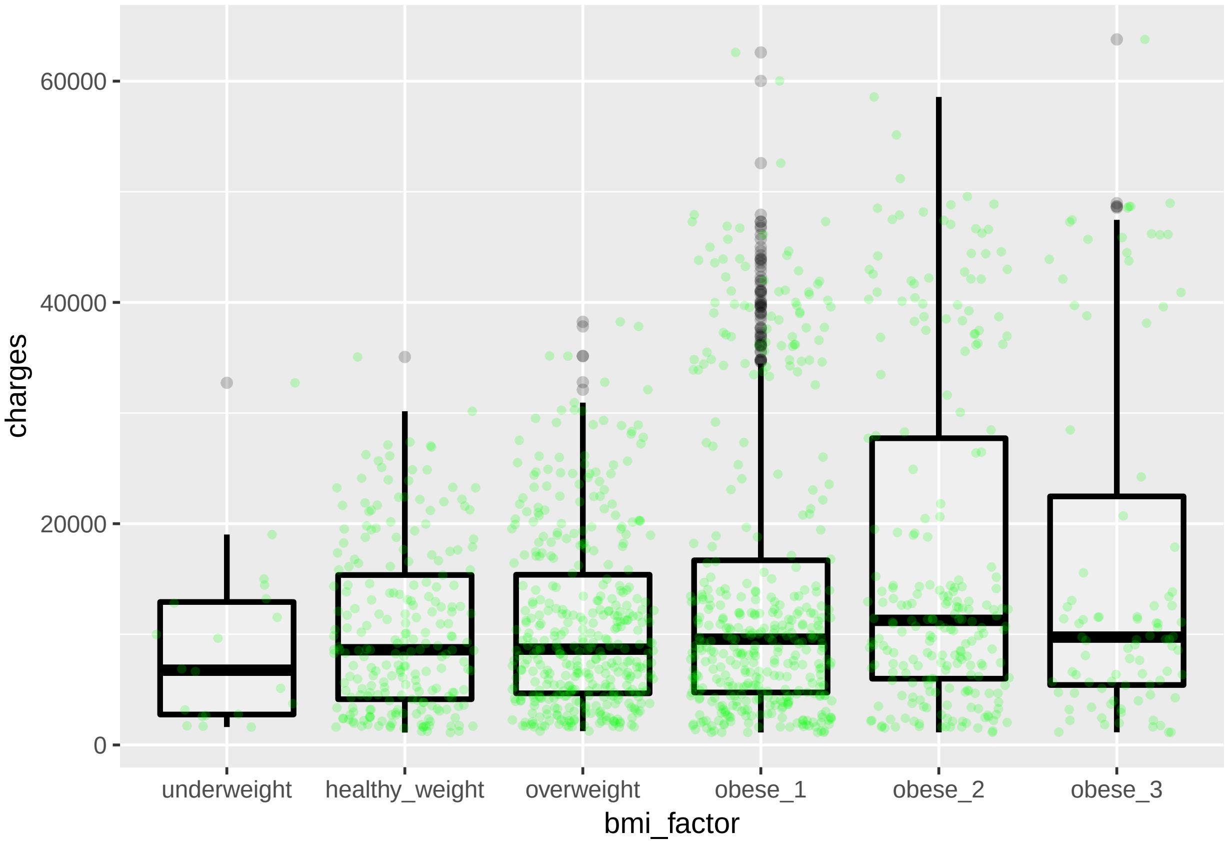
```
ggplot(health_charges_clean, aes(x = sex, y = charges))+
geom_boxplot(color = "black", size = 1, alpha = .2) +
geom_jitter(color = "green", size = 1, alpha = .2)
```



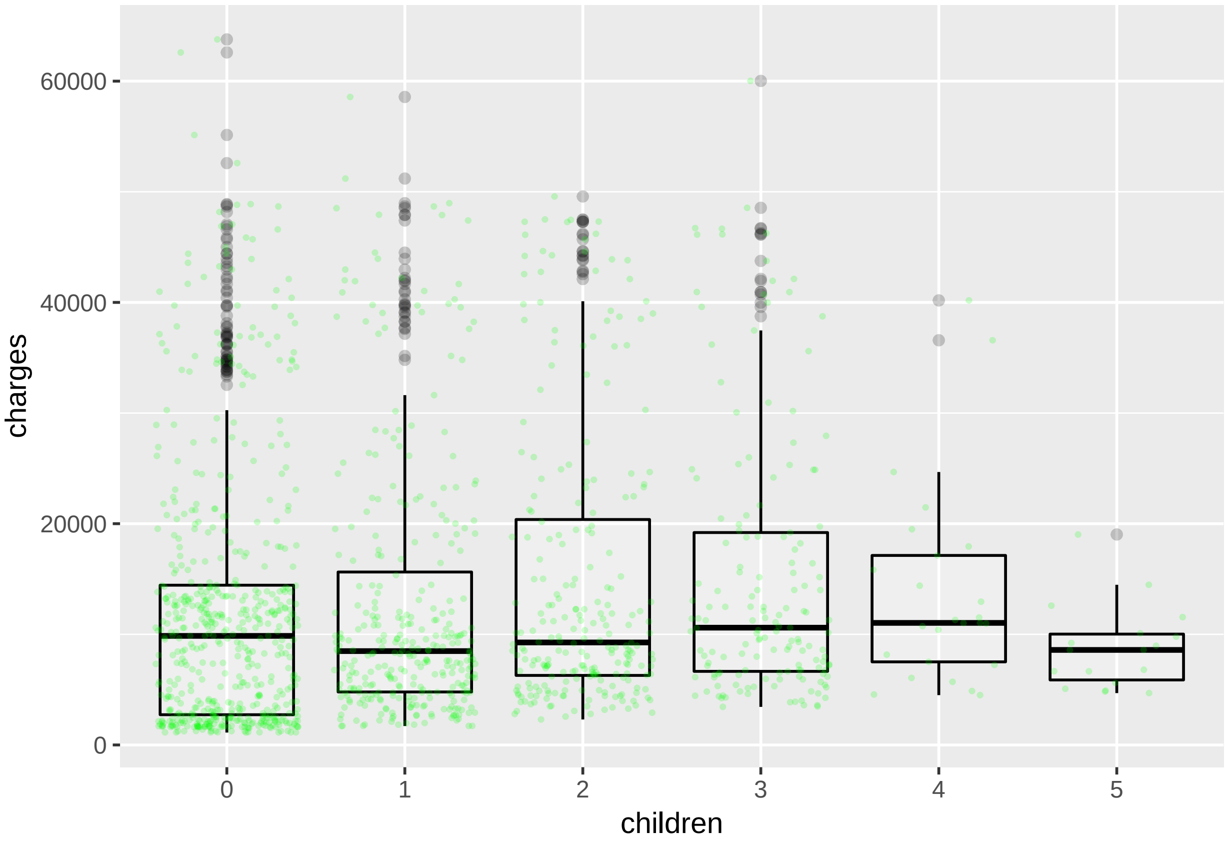
```
levels(health_charges_clean$bmi_factor)
```

```
## [1] "underweight"    "healthy_weight" "overweight"     "obese_1"
## [5] "obese_2"        "obese_3"
```

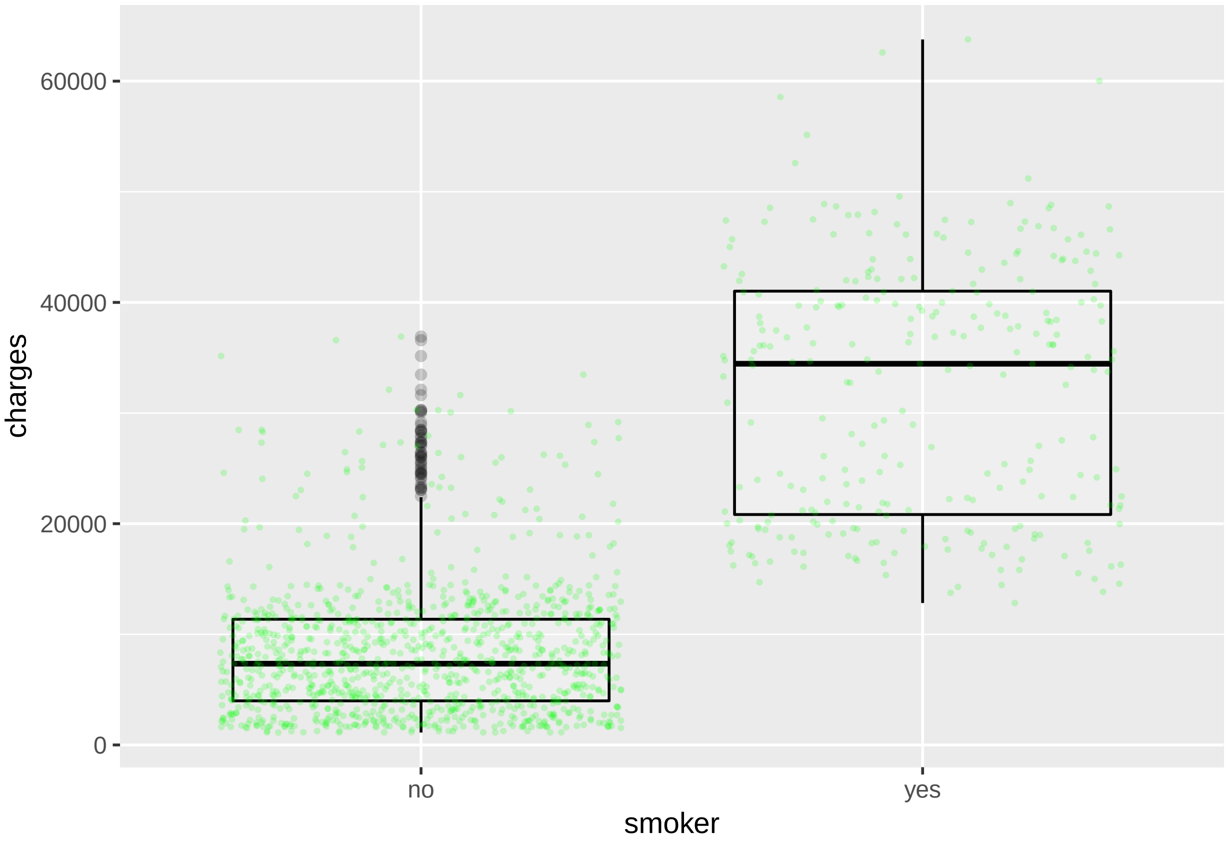
```
ggplot(health_charges_clean, aes(x = bmi_factor, y = charges))+
geom_boxplot(color = "black", size = 1, alpha = .2) +
geom_jitter(color = "green", size = 1, alpha = .2)
```



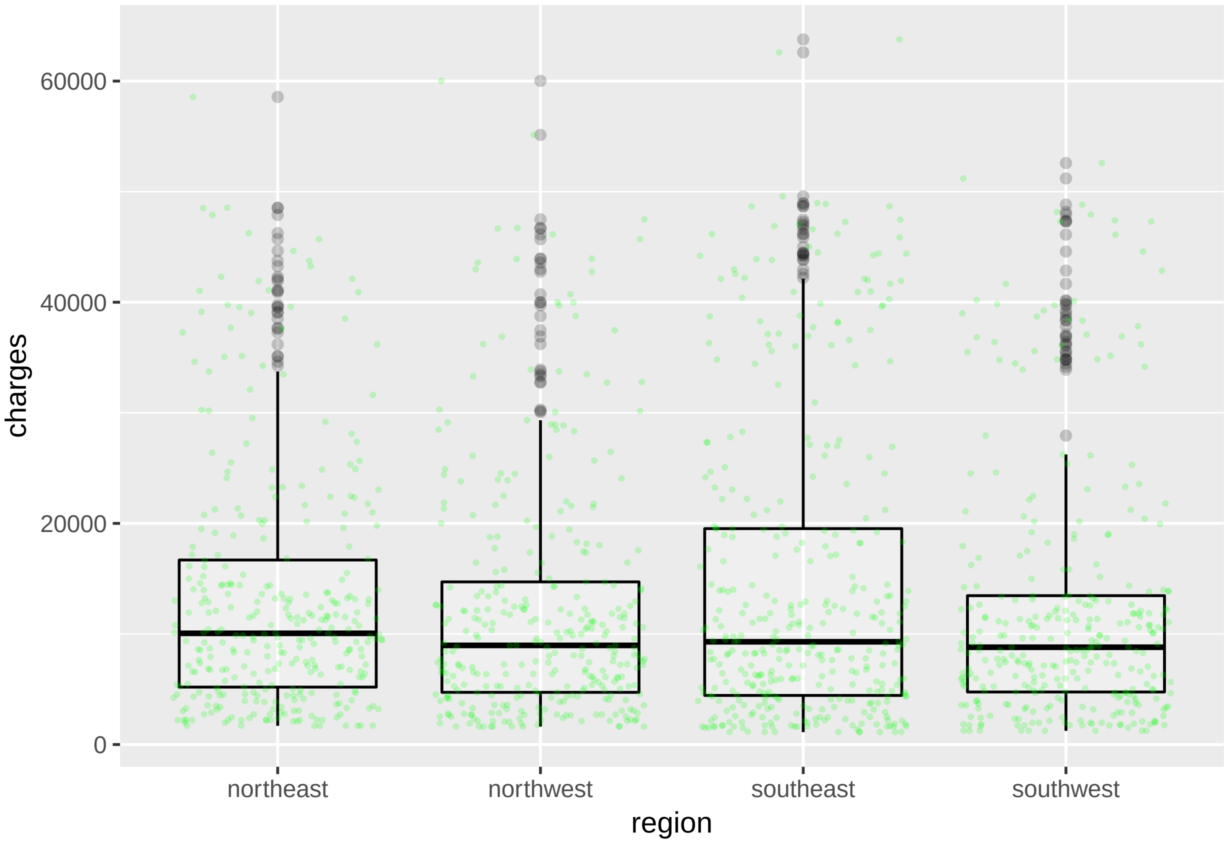
```
childrenf <- as.factor(health_charges_clean$children)
ggplot(health_charges_clean, aes(x = children, y = charges))+
geom_boxplot(color = "black", alpha = .2) +
geom_jitter(color = "green", size = .5, alpha = .2)
```



```
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)
```



## PREDICTIVE MODELS

### LINEAR REGRESSION

- *Linear model general description:*
  - No high collinearity between variables.
  - This model includes all variables aside from sex, because it was insignificant in initial models.
  - This model has 3 outliers removed, as determined by the initial linear model with all variables aside from sex.
  - Adjusted R-squared: .7536, so the 75% of the value of the charges can be attributed to these variables.
- *Linear model variables:*
  - Sex was insignificant in initial models, so the final model below does not include sex.
  - Significant variables:
  - Being a smoker increases charges by $23754.01
  - Children:
    - Having 2 children increases charges by 1633.53.
    - Having 3 children increases charges by 963.67.
  - Higher bmi increases charges by $331.84
  - Higher age increases charges by $257.43
  - Region:
    - Living in the southeast decreased charges by $941.98.
    - Living in the southwest decreased charges by $809.73.

```
lmall <- lm ( charges ~ bmi + age + smoker + children + region + sex, data = health_charges_clean)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
car::vif(lmall)
```

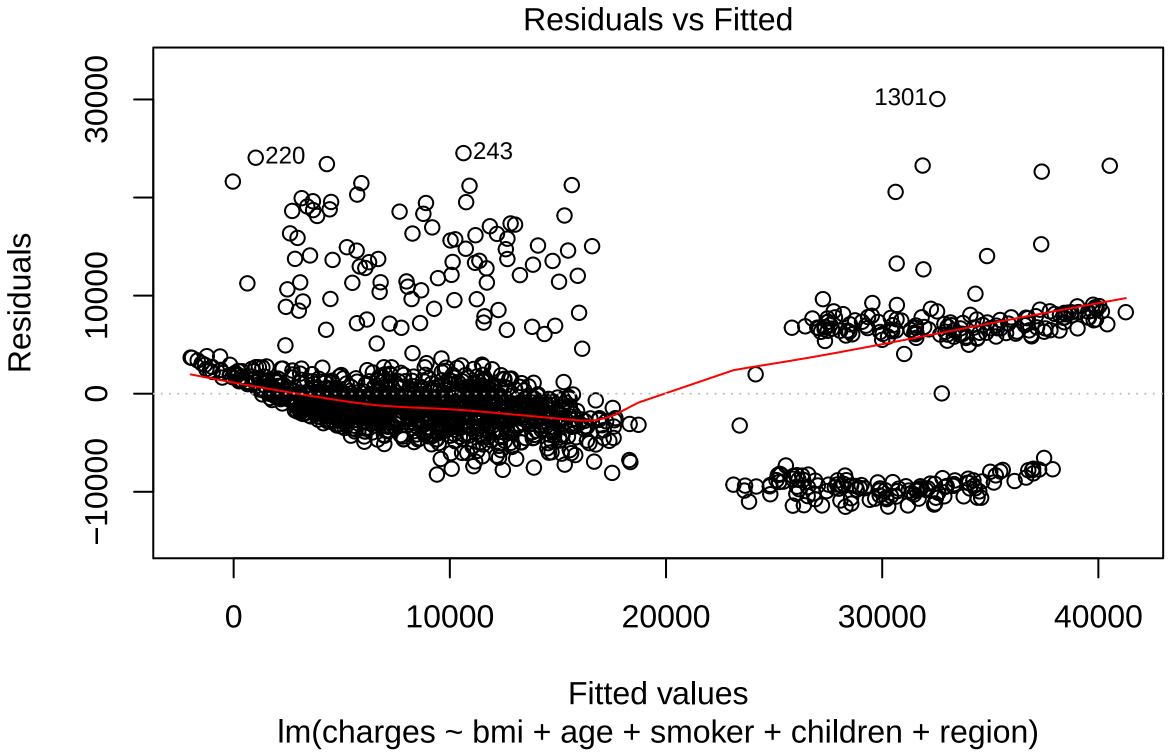
```
##              GVIF Df GVIF^(1/(2*Df))
## bmi      1.108836  1        1.053013
## age      1.020607  1        1.010251
## smoker   1.018024  1        1.008972
## children 1.024871  5        1.002460
## region   1.109919  3        1.017533
## sex      1.009334  1        1.004656
```

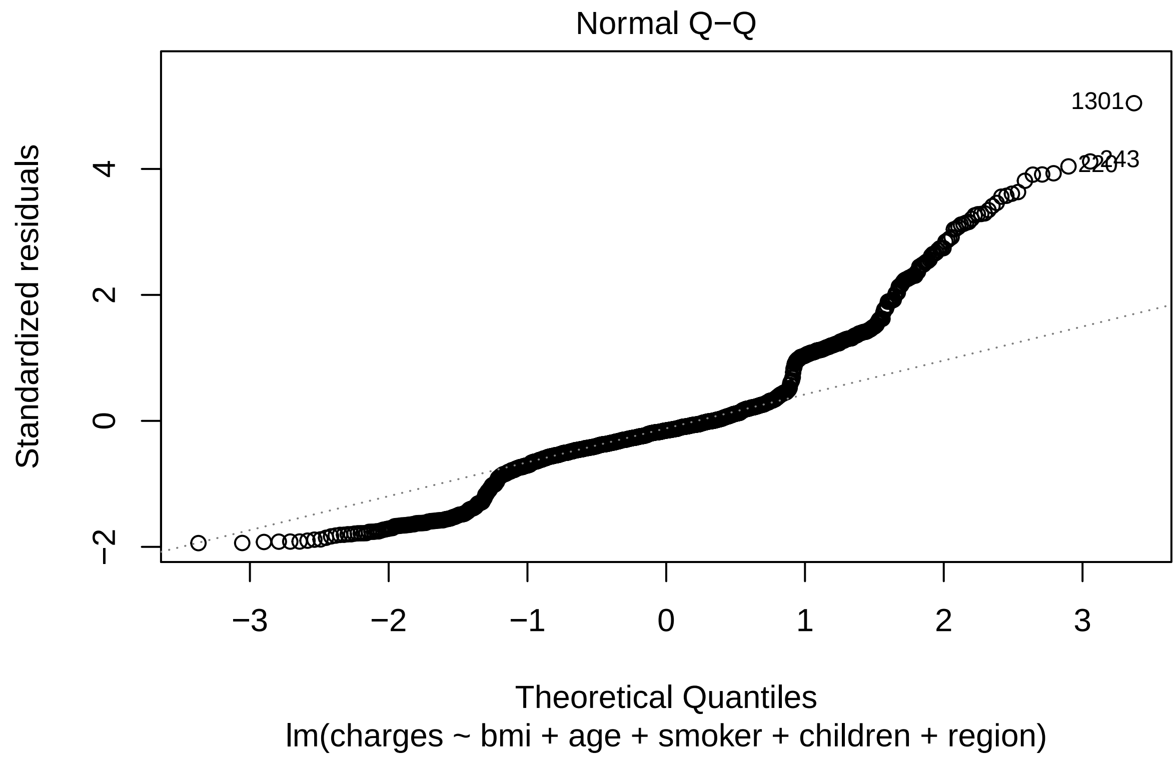
```
summary(lmall)
lmall <- lm ( charges ~ bmi + age + smoker + children + region, data = health_charges_clean)
plot(lmall)
nrow(health_charges_clean)
plot(lmall)
#outliers: 322, 578, 1013
```

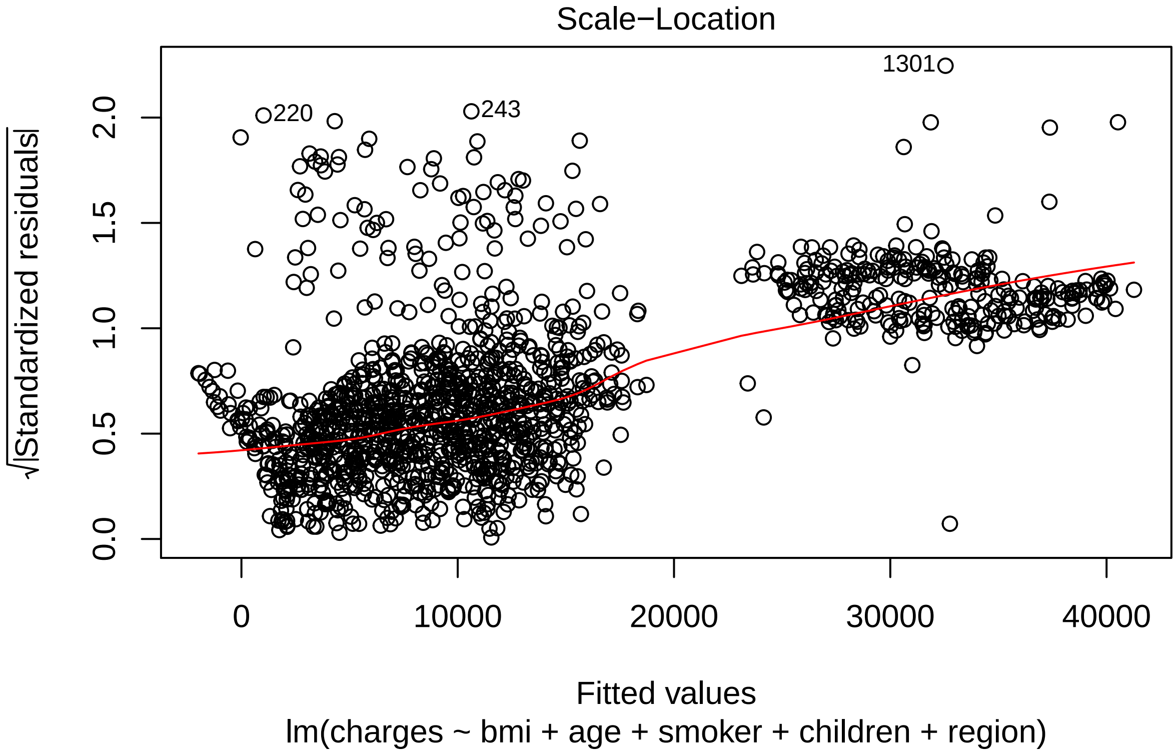
```
chargesout <- health_charges_clean[ c(1:321, 323:577, 579:1012, 1014:1338), ]
lmallout <- lm ( charges ~ bmi + age + smoker + children + region, data = chargesout)
summary(lmallout, method = lm)
```

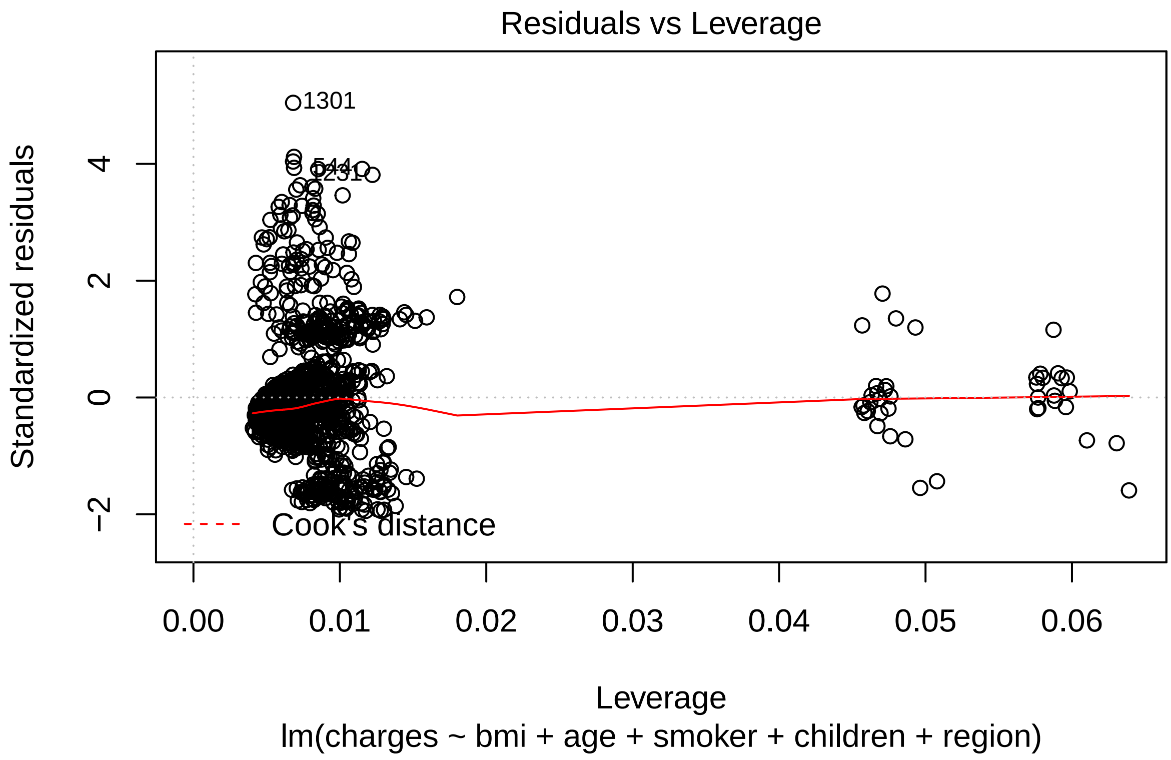
```
##
## Call:
## lm(formula = charges ~ bmi + age + smoker + children + region,
##     data = chargesout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11525.1  -2861.1   -910.5   1459.8  30039.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11917.56     972.26 -12.258   <2e-16 ***
## bmi                331.84      28.22  11.759   <2e-16 ***
## age                257.43      11.77  21.879   <2e-16 ***
## smokeryes        23754.01     407.95  58.228   <2e-16 ***
## children1          306.80     416.00   0.738   0.4609
## children2         1633.53     460.28   3.549   0.0004 ***
## children3          963.67     540.58   1.783   0.0749 .
## children4         1333.68    1272.03   1.048   0.2946
## children5         1080.46    1435.88   0.752   0.4519
## regionnorthwest   -249.81     470.72  -0.531   0.5957
## regionsoutheast   -941.98     473.98  -1.987   0.0471 *
## regionsouthwest   -809.73     472.40  -1.714   0.0867 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5976 on 1323 degrees of freedom
## Multiple R-squared:  0.7556, Adjusted R-squared:  0.7536
## F-statistic: 371.9 on 11 and 1323 DF,  p-value: < 2.2e-16
```

```r
plot(lmallout)
```




Normal Q−Q
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
1301
220
243
Theoretical Quantiles
lm(charges ~ bmi + age + smoker + children + region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10000
20000
30000
40000
220
243
1301
Fitted values
lm(charges ~ bmi + age + smoker + children + region)



### LINEAR REGRESSION OF "HIGH" CHARGES

- The purpose of this model is to see which facets are most significant in this subset of the population.
- Adjusted R-squared: 0.6135, so 61% of the values of charges can be attributed to these variables. "High" charges are less predictable than charges in general.
- Age, smoker, and bmi are significant in both.
- Notable differences in predictive nature of "smoker" and "bmi" between the general population and "high charge" population:
  - Smoker: stronger predictor for the general population
    - Difference in charges: $23616.0 versus $9586.85
    - Difference in R^2: 0.6195 versus 0.1342
  - Bmi: stronger predictor for population with "high" charges
    - Difference in charges: $393.87 versus $1166.60
    - Difference in R^2: 0.03862 versus 0.4016

```
vquantile <- as.vector(quantile(health_charges$charges))
hcut <- vquantile[c(4)]
hcut
```

```
## [1] 16639.91
```

```
cut <- health_charges_clean[ c(health_charges_clean$charges > 16639.91), ]

str(cut)
```

```
## 'data.frame':    335 obs. of  10 variables:
##  $ age          : int  19 33 60 62 27 30 34 31 22 28 ...
##  $ sex          : Factor w/ 2 levels "female","male": 1 2 1 1 2 2 1 2 2 2 ...
##  $ bmi          : num  27.9 22.7 25.8 26.3 42.1 ...
##  $ bmi_factor   : Ord.factor w/ 6 levels "underweight"<..: 3 2 3 3 6 5 4 5 5 5 ...
##  $ children     : Factor w/ 6 levels "0","1","2","3",..: 1 1 1 1 1 1 2 3 1 2 ...
##  $ smoker        : Factor w/ 2 levels "no","yes": 2 1 1 2 2 2 2 2 2 2 ...
##  $ region        : Factor w/ 4 levels "northeast","northwest",..: 4 2 2 3 3 4 1 4 4 4 ...
##  $ charges       : num  16885 21984 28923 27809 39612 ...
##  $ charges_factor: Ord.factor w/ 2 levels "low"<"high": 2 2 2 2 2 2 2 2 2 2 ...
##  $ age_factor    : Ord.factor w/ 6 levels "10s"<"20s"<"30s"<..: 1 3 6 6 2 3 3 3 2 2 ...
```

```
lmcut <- lm ( charges ~ bmi + age + smoker + sex + children + region, data = cut)
summary(lmcut, method = lm)
```

```
## 
## Call:
## lm(formula = charges ~ bmi + age + smoker + sex + children + 
##     region, data = cut)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -17274.6  -4975.9    305.1   4271.3  30197.9 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -21373.05    2503.50  -8.537 5.52e-16 ***
## bmi               1163.47      64.43  18.058  < 2e-16 ***
## age                222.01      26.50   8.377 1.70e-15 ***
## smokeryes         9564.87     870.54  10.987  < 2e-16 ***
## sexmale           -162.64     742.27  -0.219    0.827    
## children1         -125.62     953.96  -0.132    0.895    
## children2         1075.98     981.91   1.096    0.274    
## children3         -168.00    1148.16  -0.146    0.884    
## children4        -2005.29    2610.48  -0.768    0.443    
## children5          556.25    6756.36   0.082    0.934    
## regionnorthwest    258.29    1079.31   0.239    0.811    
## regionsoutheast   -947.76     985.86  -0.961    0.337    
## regionsouthwest    325.18    1116.71   0.291    0.771    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6628 on 322 degrees of freedom
## Multiple R-squared:  0.6228, Adjusted R-squared:  0.6087 
## F-statistic:  44.3 on 12 and 322 DF,  p-value: < 2.2e-16
```

```
#smoker:
lmsmoke <- lm ( charges ~ smoker, data = health_charges_clean)
summary(lmsmoke, method = lm)
```

```
## 
## Call:
## lm(formula = charges ~ smoker, data = health_charges_clean)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -19221  -5042   -919   3705  31720 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   8434.3      229.0   36.83   <2e-16 ***
## smokeryes    23616.0      506.1   46.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7470 on 1336 degrees of freedom
## Multiple R-squared:  0.6198, Adjusted R-squared:  0.6195
## F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

```
lmsmokecut <- lm ( charges ~ smoker, data = cut)
summary(lmsmokecut, method = lm)
```

```
##
## Call:
## lm(formula = charges ~ smoker, data = cut)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -16642  -8532   1211   6753  30470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24122       1102  21.885  < 2e-16 ***
## smokeryes       9178       1263   7.265 2.66e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9858 on 333 degrees of freedom
## Multiple R-squared:  0.1368, Adjusted R-squared:  0.1342
## F-statistic: 52.78 on 1 and 333 DF,  p-value: 2.663e-12
```

```
#bmi
lmbmi <- lm ( charges ~ bmi, data = health_charges_clean)
summary(lmbmi, method = lm)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = health_charges_clean)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1192.94    1664.80   0.717    0.474
## bmi           393.87      53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

```
lmbmicut <- lm ( charges ~ bmi, data = cut)
summary(lmbmicut, method = lm)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = cut)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24765  -5352    657   4658  32577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4946.68    2444.45  -2.024   0.0438 *
## bmi          1151.61      76.75  15.004   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8196 on 333 degrees of freedom
## Multiple R-squared:  0.4033, Adjusted R-squared:  0.4016
## F-statistic: 225.1 on 1 and 333 DF,  p-value: < 2.2e-16
```

### LOGISTIC REGRESSION

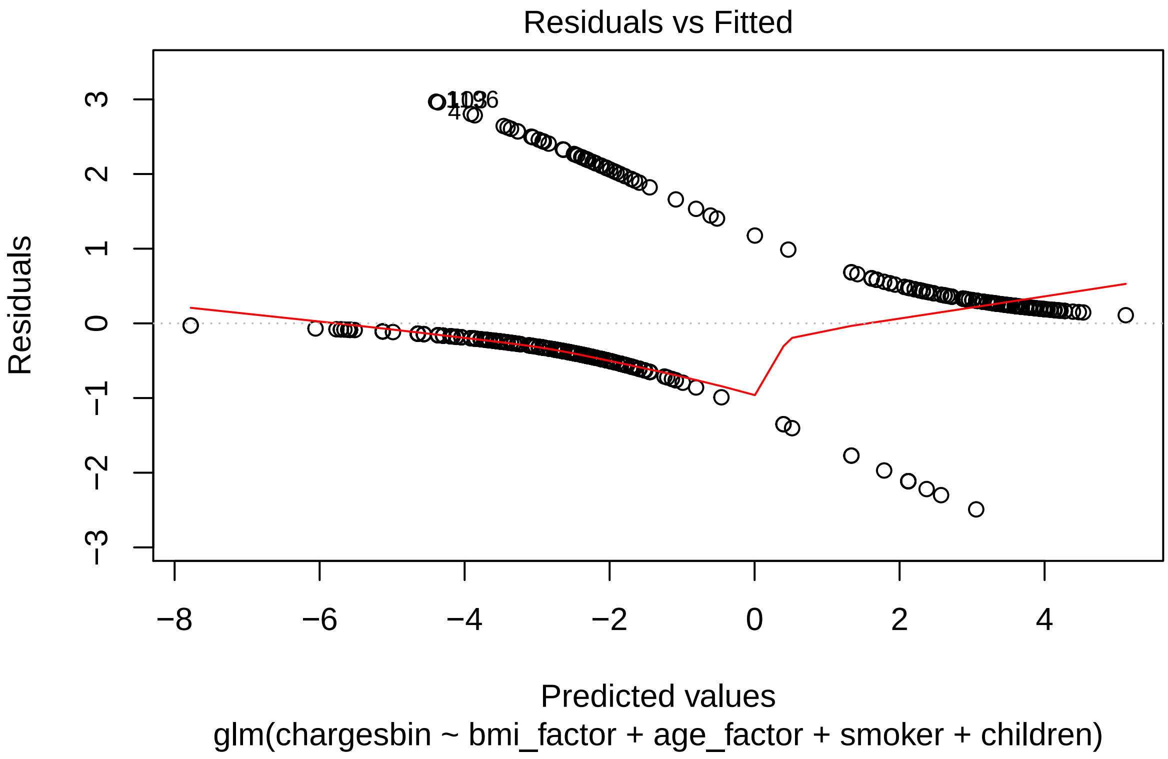
- Important variables:
  - Being a smoker.
  - Age.
  - BMI: overweight, obese1, obese2
  - Children
- The initial model included region and sex, but those were insignificant. They are not inluded in the model below.

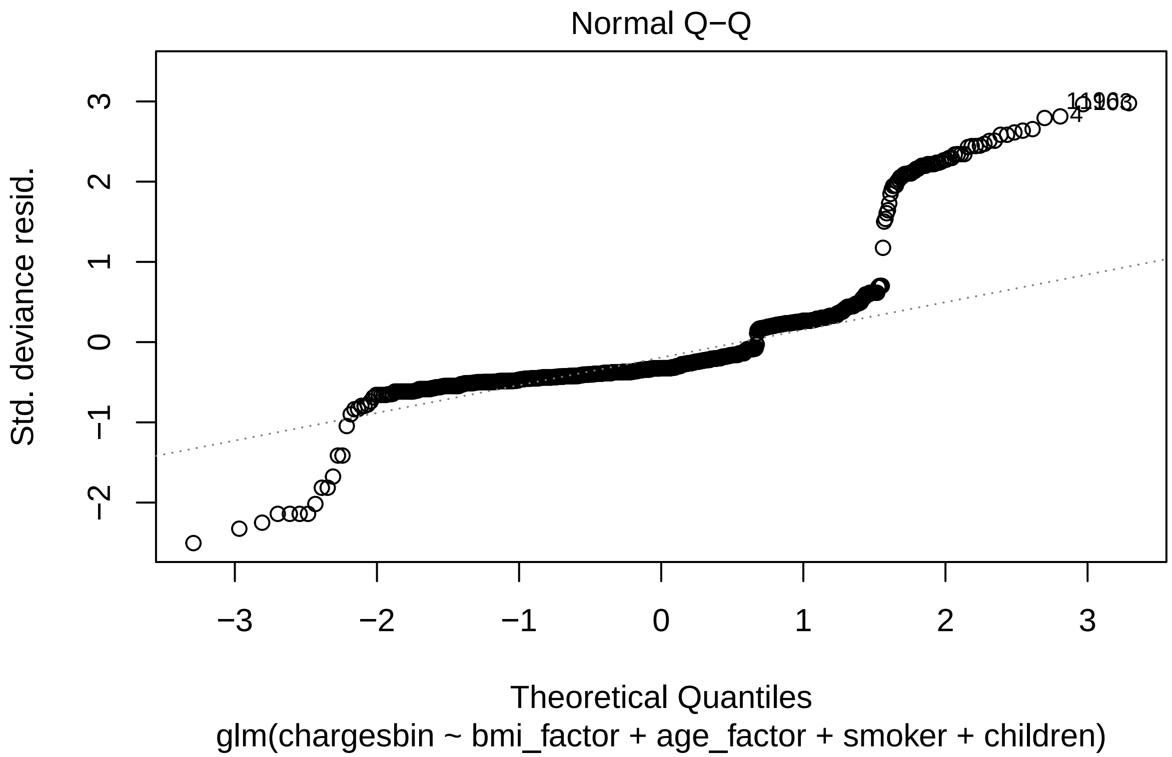
```r
#binary variable for charges_factor = high
library(dplyr)
hc1 <- health_charges_clean %>% mutate(chargesbin = if_else (charges_factor == "high", 1, 0))
library(caTools)
set.seed(88)
split  = sample.split(hc1$chargesbin, SplitRatio = .75 )
hc1train = subset(hc1, split == TRUE)
hc1test= subset(hc1, split == FALSE)
```

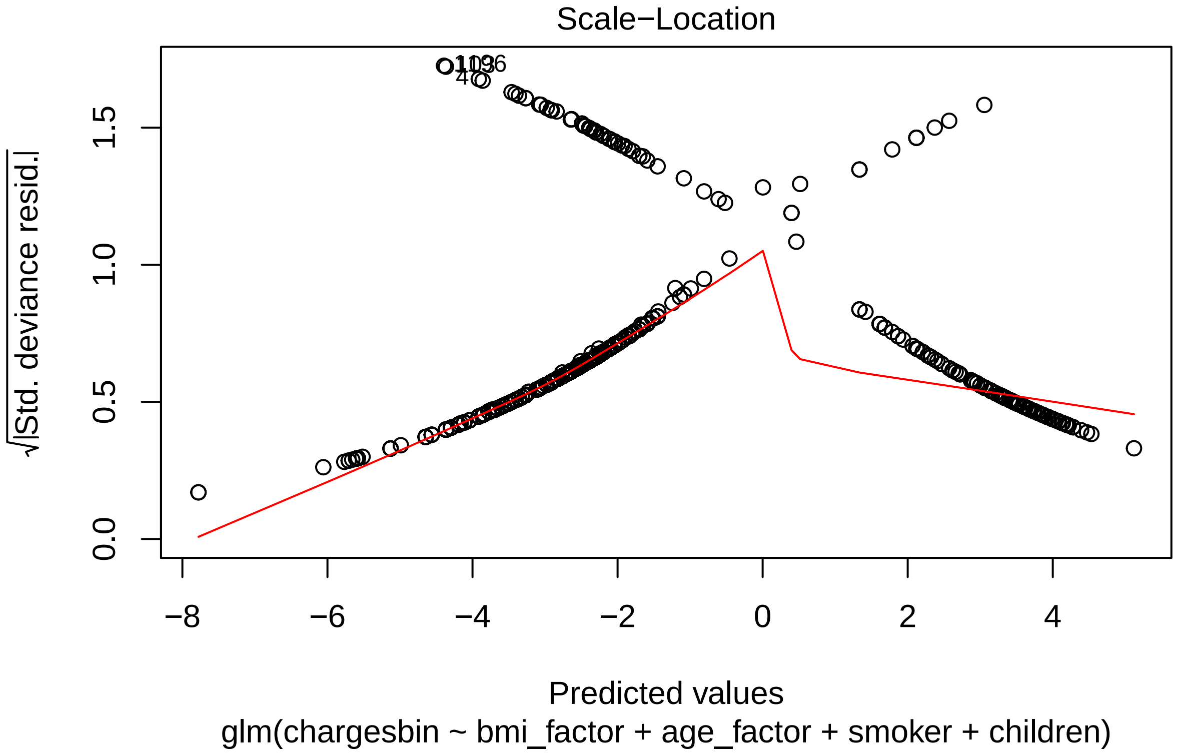
```r
lgall= glm(chargesbin ~ bmi_factor + age_factor + smoker + children,  data = hc1train, family = binomial
predicttrain = predict(lgall, type = "response")
summary(lgall)
```

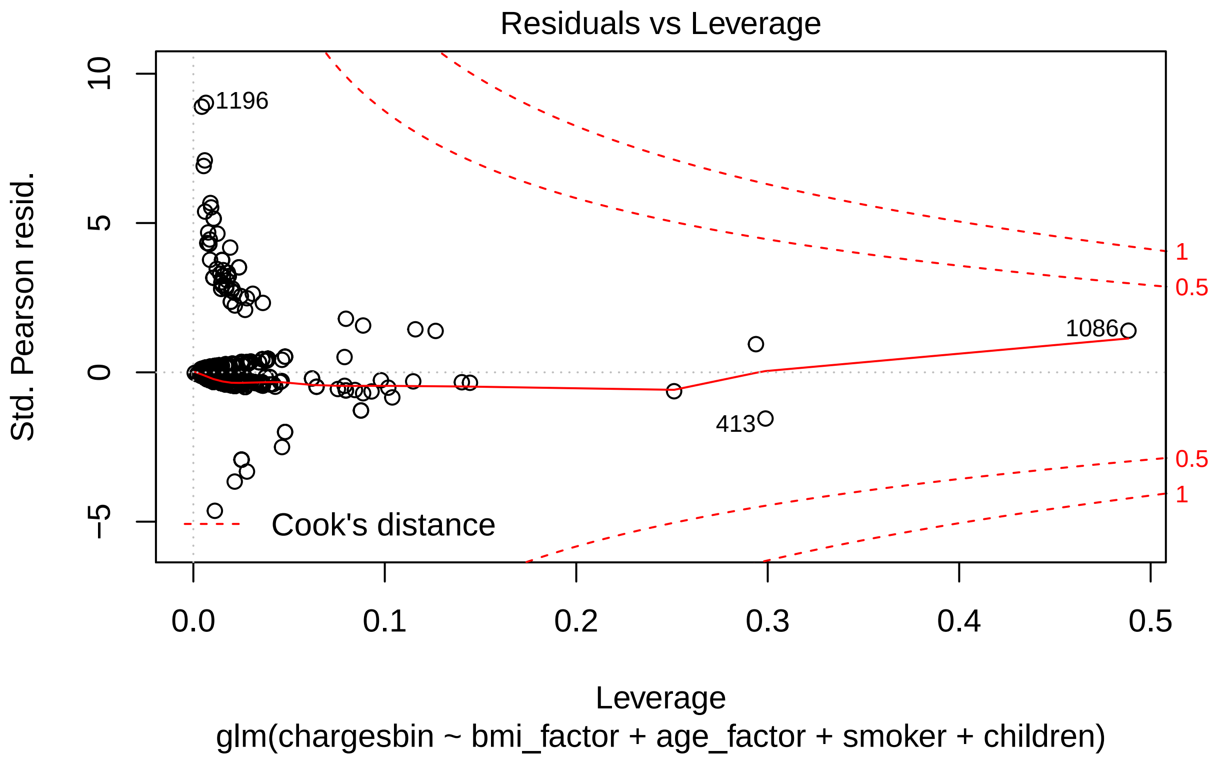
```
##
## Call:
## glm(formula = chargesbin ~ bmi_factor + age_factor + smoker +
##     children, family = binomial, data = hc1train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.49101  -0.42168  -0.32352   0.04012   2.96831
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.71500    0.34393 -10.802  < 2e-16 ***
## bmi_factor.L  2.48224    0.79504   3.122  0.00180 **
## bmi_factor.Q -1.74149    0.71458  -2.437  0.01481 *
## bmi_factor.C  0.34913    0.52528   0.665  0.50628
## bmi_factor^4 -0.24168    0.37606  -0.643  0.52045
## bmi_factor^5 -0.08409    0.26646  -0.316  0.75231
## age_factor.L  1.28139    0.44481   2.881  0.00397 **
## age_factor.Q -0.60643    0.43149  -1.405  0.15990
## age_factor.C  0.68103    0.34296   1.986  0.04706 *
## age_factor^4 -0.65399    0.29820  -2.193  0.02830 *
## age_factor^5 -0.22463    0.29183  -0.770  0.44146
## smokeryes     5.98044    0.41298  14.481  < 2e-16 ***
## children1     0.45212    0.34533   1.309  0.19045
## children2     0.59403    0.36778   1.615  0.10627
## children3     0.25262    0.43305   0.583  0.55966
## children4     1.83728    0.63850   2.877  0.00401 **
## children5     0.13342    1.27472   0.105  0.91664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1128.59  on 1002  degrees of freedom
## Residual deviance:  475.57  on  986  degrees of freedom
## AIC: 509.57
##
## Number of Fisher Scoring iterations: 6
```

```
plot(lgall)
```




Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
11963
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(chargesbin ~ bmi_factor + age_factor + smoker + children)


Scale−Location
√|Std. deviance resid.|
1.5
1.0
0.5
0.0
1036
4
−8
−6
−4
−2
0
2
4
Predicted values
glm(chargesbin ~ bmi_factor + age_factor + smoker + children)



### CONFUSION MATRIX ON TRAINING SET

- 93.02% accuracy of predicting high health charges.

### CONFUSION MATRIX ON TESTING SET

- Our model has 91.94% accuracy of predicting high health charges
- Sensitivity: true positive rate, 0.7380952
- Specificity: false positive rate, 0.01992032

```
table(hc1train$chargesbin, predicttrain > .5)
```

```
##
##     FALSE TRUE
##   0   739   13
##   1    58  193
```

```
(192 + 739) / (192 + 59 + 13 + 739)
```

```
## [1] 0.9282154
```

```
#true positive:
192 / (192 + 59)
```

```
## [1] 0.7649402
```

```
#false positive:
13 / (739 + 13)
```

```
## [1] 0.01728723
```

```
predicttest = predict(lgall, type = "response", newdata = hc1test)
table(hc1test$chargesbin, predicttest > .5)
```

```
##
##     FALSE TRUE
##   0   246    5
##   1    22   62
```

```
(62 + 246) / ( 62 + 22 + 5 + 246)
```

```
## [1] 0.919403
```

```
#true positive:
62 / (22 + 62)
```

```
## [1] 0.7380952
```

```
#false positive:
5 / (246 + 5)
```

```
## [1] 0.01992032
```

*ROC Curve:* Area under the curve: 0.8706. We have a good model.

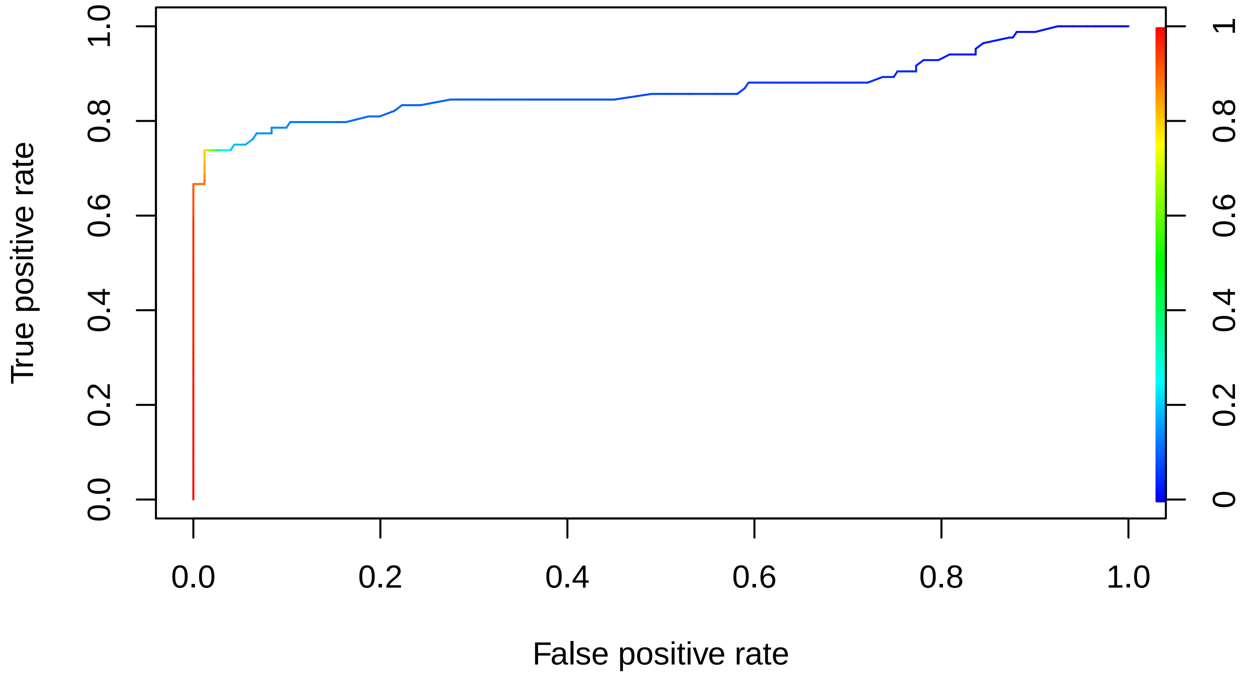
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
ROCRpred <- prediction(predicttest, hc1test$chargesbin)
ROCRperf <- performance(ROCRpred, "tpr", "fpr")
plot(ROCRperf, colorize = TRUE)
```



```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
roc_obj <- roc(hc1test$chargesbin, predicttest)
auc(roc_obj)
```

```
## Area under the curve: 0.868
```

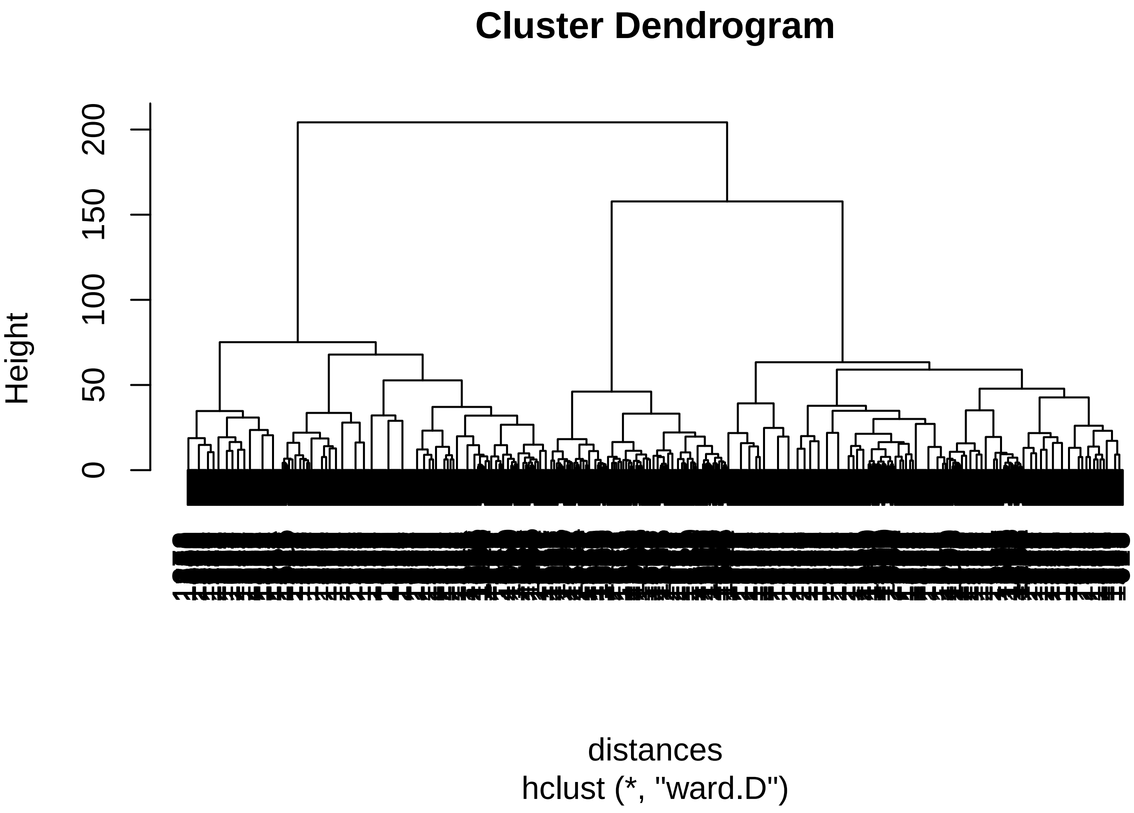
### CLUSTERING

- Heigharchial clustering with dendrogram; all variables aside from charges.
- Eight clusters: groups with high charges:
  - Group 1 was a predictor of high charges at 91.7%
  - The percentage of high charges and the percent of smokers within each cluster were equal across all clusters.
  - Smoking and high charges were most powerful in the clustering algorithm.

```r
hc2 <- binary_charges
distances = dist(hc2[c(-1, -2)], method = "euclidian")
cluster1 = hclust(distances, method = "ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```r
plot(cluster1)
```



```
clustergroups = cutree(cluster1, k = 8)
str(clustergroups)
```

```
##  int [1:1338] 1 2 2 3 3 4 4 3 3 5 ...
```

```
highv <- tapply(hc2$charges_factor_high, clustergroups, mean)
highv <- as.vector(highv)
smokev <- tapply(hc2$smoker_yes, clustergroups, mean)
smokev <- as.vector(highv)
nev <- tapply(hc2$region_northeast, clustergroups, mean)
nev <- as.vector(nev)
nwv <- tapply(hc2$region_northwest, clustergroups, mean)
nwv <- as.vector(nwv)
swv <- tapply(hc2$region_southwest, clustergroups, mean)
swv <- as.vector(swv)
sev <- tapply(hc2$region_southeast, clustergroups, mean)
sev <- as.vector(sev)
sexfv <- tapply(hc2$sex_female, clustergroups, mean)
sexfv <- as.vector(sexfv)
ch0v <- tapply(hc2$children_0, clustergroups, mean)
ch0v <- as.vector(ch0v)
ch1v <- tapply(hc2$children_1, clustergroups, mean)
ch1v <- as.vector(ch1v)
ch2v <- tapply(hc2$children_2, clustergroups, mean)
ch2v <- as.vector(ch2v)
ch3v <- tapply(hc2$children_3, clustergroups, mean)
ch3v <- as.vector(ch3v)
ch4v <- tapply(hc2$children_4, clustergroups, mean)
ch4v <- as.vector(ch4v)
ch5v <- tapply(hc2$children_5, clustergroups, mean)
ch5v <- as.vector(ch5v)
bmiuv <- tapply(hc2$bmi_factor_underweight, clustergroups, mean)
bmiuv <- as.vector(bmiuv)
bmihv <- tapply(hc2$bmi_factor_healthy_weight, clustergroups, mean)
bmihv <- as.vector(bmihv)
bmiov <- tapply(hc2$bmi_factor_overweight, clustergroups, mean)
bmiov <- as.vector(bmiov)
bmio1v <- tapply(hc2$bmi_factor_obese_1, clustergroups, mean)
bmio1v <- as.vector(bmio1v)
bmio2v <- tapply(hc2$bmi_factor_obese_2, clustergroups, mean)
bmio2v <- as.vector(bmio2v)
bmio3v <- tapply(hc2$bmi_factor_obese_3, clustergroups, mean)
bmio3v <- as.vector(bmio3v)

clusterframe <- cbind(highv, smokev, sexfv, bmiuv, bmihv, bmiov, bmio1v, bmio2v, bmio3v, ch0v, ch1v, ch

head(clusterframe)
```

```
##           highv     smokev      sexfv      bmiuv      bmihv     bmiov
## [1,] 0.91699605 0.91699605 0.41897233 0.01976285 0.19762846 0.2806324
## [2,] 0.11067194 0.11067194 0.15019763 0.00000000 0.09486166 0.1383399
## [3,] 0.07476636 0.07476636 0.07009346 0.05607477 0.33177570 0.4859813
## [4,] 0.13385827 0.13385827 0.91338583 0.00000000 0.11023622 0.4173228
## [5,] 0.04615385 0.04615385 1.00000000 0.00000000 0.00000000 1.0000000
## [6,] 0.11111111 0.11111111 0.00000000 0.00000000 0.00000000 0.0000000
##          bmio1v     bmio2v      bmio3v      ch0v      ch1v      ch2v
## [1,] 0.24901186 0.16996047 0.083003953 0.4545455 0.2332016 0.1897233
## [2,] 0.18972332 0.31620553 0.260869565 0.3517787 0.3241107 0.2015810
## [3,] 0.09345794 0.02803738 0.004672897 0.3551402 0.2476636 0.1542056
## [4,] 0.14173228 0.30708661 0.023622047 0.3779528 0.3070866 0.1811024
## [5,] 0.00000000 0.00000000 0.000000000 1.0000000 0.0000000 0.0000000
## [6,] 1.00000000 0.00000000 0.000000000 0.6666667 0.1010101 0.1212121
##           ch3v        ch4v        ch5v        nev       nwv        sev
## [1,] 0.1106719 0.011857708 0.000000000 0.26482213 0.2134387 0.33201581
## [2,] 0.1067194 0.007905138 0.007905138 0.07905138 0.0513834 0.43873518
## [3,] 0.1355140 0.051401869 0.056074766 0.47663551 0.3831776 0.02803738
## [4,] 0.1338583 0.000000000 0.000000000 0.00000000 0.0000000 1.00000000
## [5,] 0.0000000 0.000000000 0.000000000 0.32307692 0.3076923 0.00000000
## [6,] 0.1111111 0.000000000 0.000000000 0.13131313 0.5151515 0.20202020
##            swv
## [1,] 0.1897233
## [2,] 0.4308300
## [3,] 0.1121495
## [4,] 0.0000000
## [5,] 0.3692308
## [6,] 0.1515152
```

---

## CONCLUSION

By far, the most significant predictor of high health charges was being a smoker. Other important predictors were age and bmi.

For our linear regression, all variables were significant aside from sex in developing a model, with the following variables did have significant effects on charges: being a smoker (+$23754.01), having two childern (+$1633.53), having 3 children (+$963.67), higher bmi (+$331.84), higher age (+$257.43), living in the southeast (-$941.98), and living in the southwest (-$809.73). Age, number of children, and bmi were most significant. The linear regressions including all variables had an R^2 of 75% for the general population, and R^2 of 61% for the high charges population. The linear regression of only "high" charges showed that it was more difficult to predict the cause of the charges without the comparison of "low" charges. The significance of certain facets was different between the general model and the "high" charges model. Smoker was a stronger predictor for the general population, with a difference in charges of $23616.0 versus $9586.85, and a difference in R^2 of 0.6195 versus 0.1342. Bmi was a stronger predictor for population with "high" charges with a difference in charges of $393.87 versus $1166.60, and a difference in R^2 of 0.03862 versus 0.4016. Children was a significant predictor for the general population's charges, but not a significant predictor for the data subset of "high" charges.

For the logistic regression, the significant variables were smoker, age, bmi, and number of children. Our model has 91.94% accuracy of predicting high health charges, with a true positive rate of 0.7380952 and false positive rate of 0.01992032. The area under the ROCR curve was 0.8706.

Clustering drew similar conclusions to the linear regression models; smoking was the only significant variable in clusters with high health charges.

It's interesting to consider how to handle charging smokers for health insurance, when this is a behavioral cause of high charges. It is illegal to charge more for insurance for individuals with pre-existing conditions, but insurance companies do charge more for people who don't attest to non-smoking status.

It could be useful to have more data on behavioral habits to use as predictive measures; it is legal to adjust insurance charges for individual behaviors. Examples could include exercise level, sleep, and diet. Technology such as smart watches and more could eventually be used for this data collection. Perhaps people could receive reduced rates for providing data to incentivize the provision of data. People engaging in behaviors that generally reduce charges could receive lower rates. This also becomes an ethics question, as some individuals may have significant obstacles to engaging in cost-lowering behaviors due to their living location, profession, and income level, among other things.

It would be interesting to see how socioeconomic variables impact health charges (education level, income, marital status, and housing). It could also be valuable to study the breakdown of charges themselves (medication, urgent care, preventative care) in respect to overall charges. From this study, it is clear that personal attributes can predict health charges to a notable degree. Increasing the scope of our data collection and the specificity of charges breakdown could improve the accuracy and scope of our predictive models.