

# 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 <sup>2</sup>
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 residential 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 ...
```

---

## 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 an 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)
```

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

```

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

#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", "obese_3"),
  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 ...
## $ 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 2 ...
## $ age_factor     : Ord.factor w/ 6 levels "10s"<"20s"<"30s"<...: 1 1 2 3 3 3 4 3 3 6 ...
```

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

```
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 1 ...
## $ charges_factor_low      : int   0 1 1 0 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

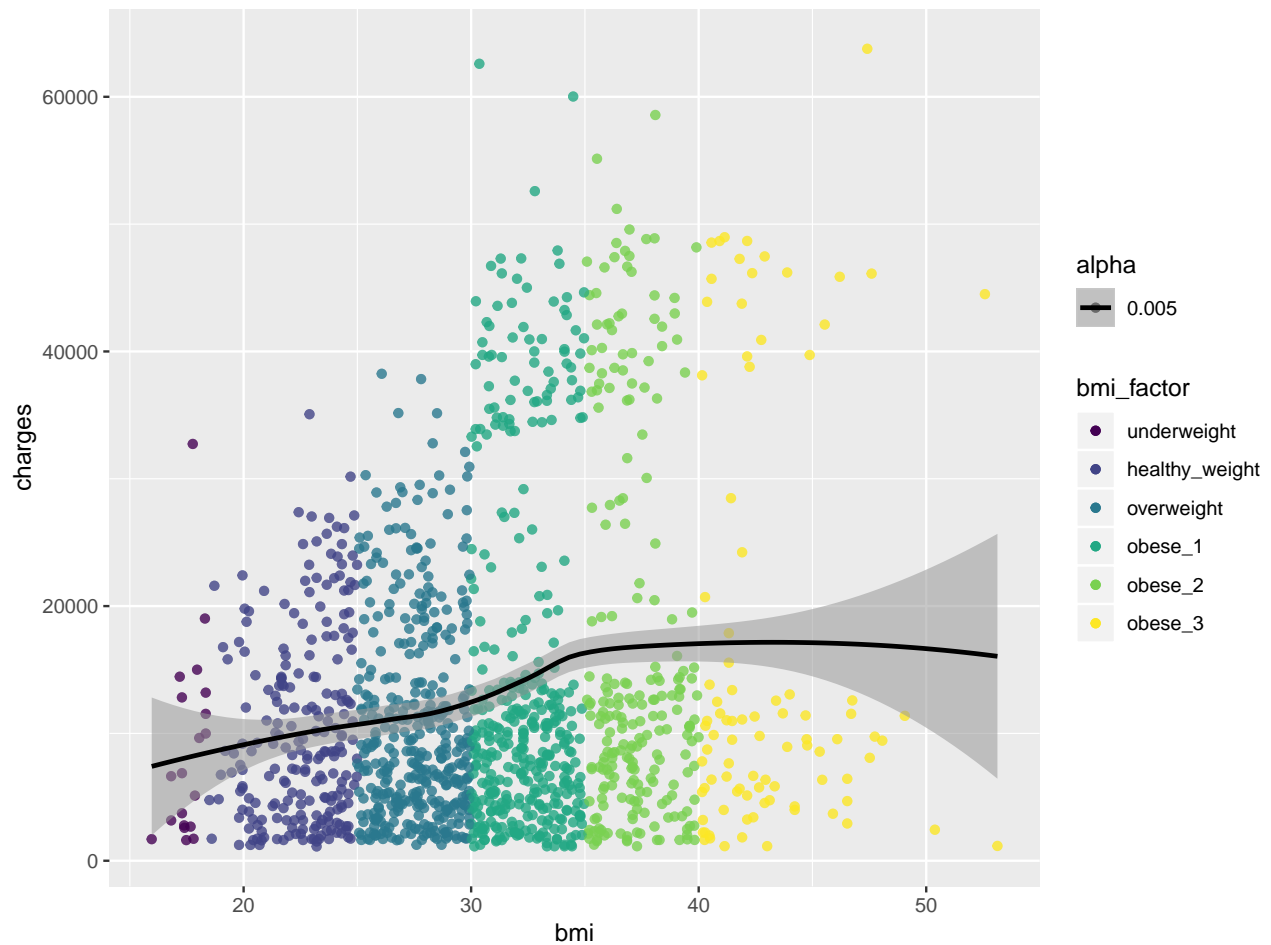
- + HO: Charges frequency follows a normal distribution.
- + HA: Charges frequency does not follow a normal distribution.
- + RESULTS:

- 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)
```

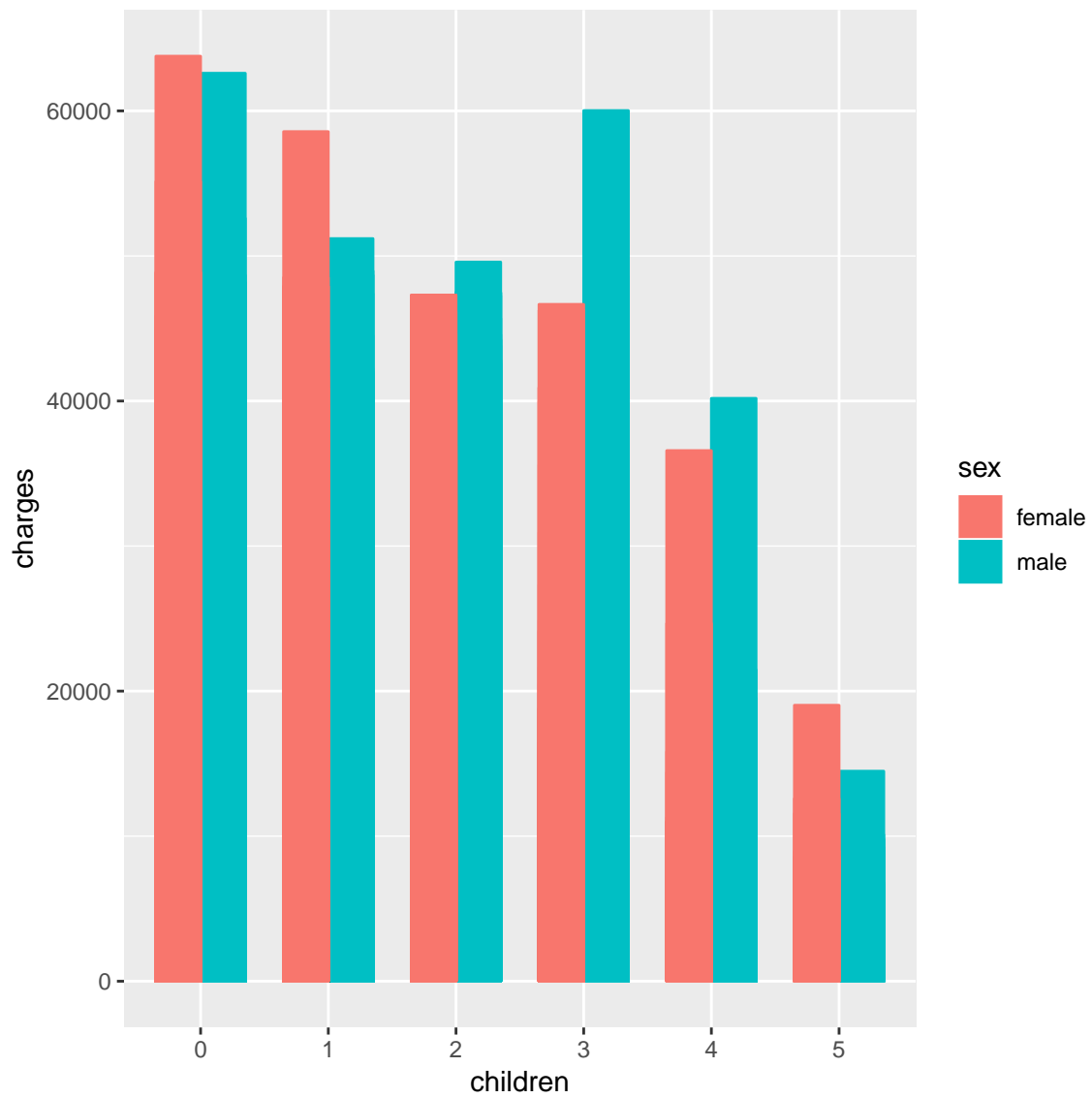
```
##
## Shapiro-Wilk normality test
##
## data: health_charges_clean$charges
## W = 0.81469, p-value < 2.2e-16
```

## MULTIVARIATE ANALYSIS



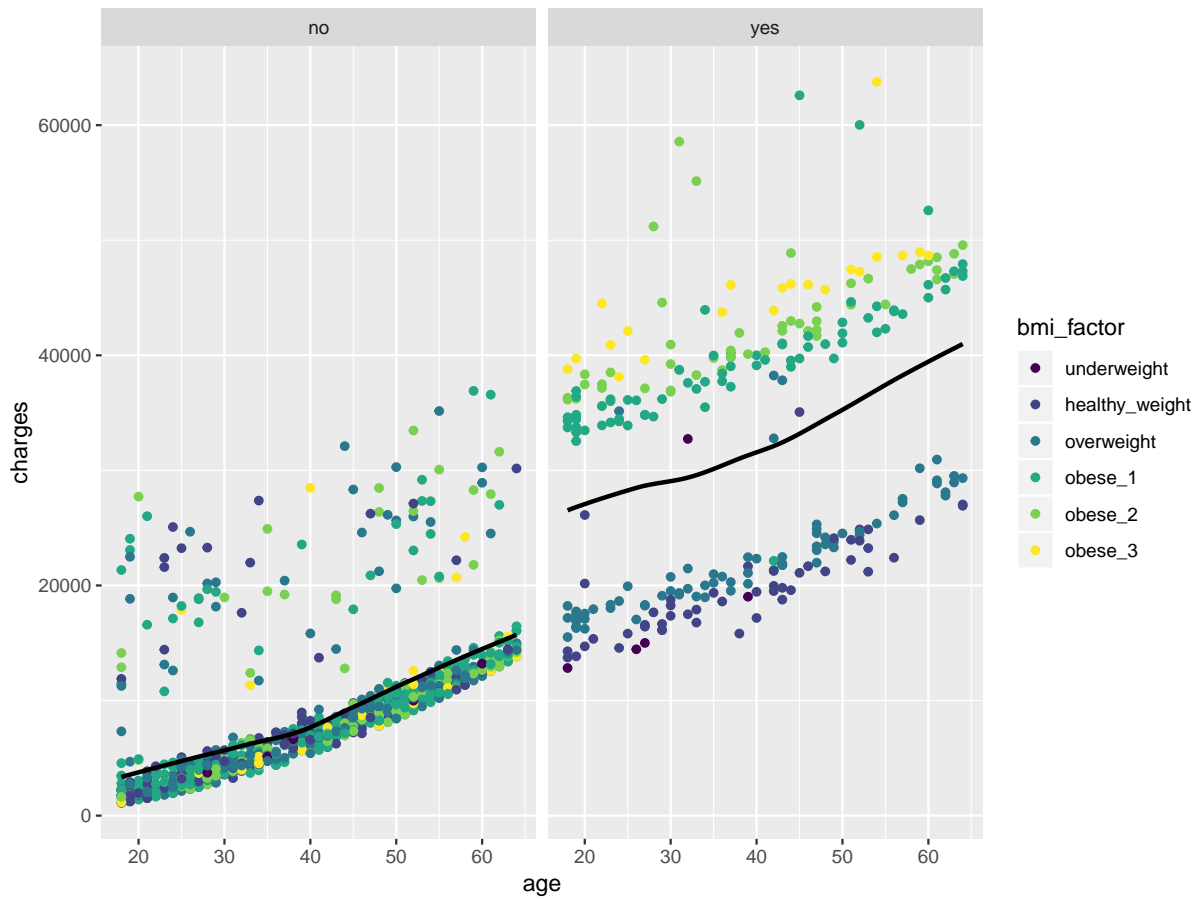
### Effect of BMI on charges

- 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

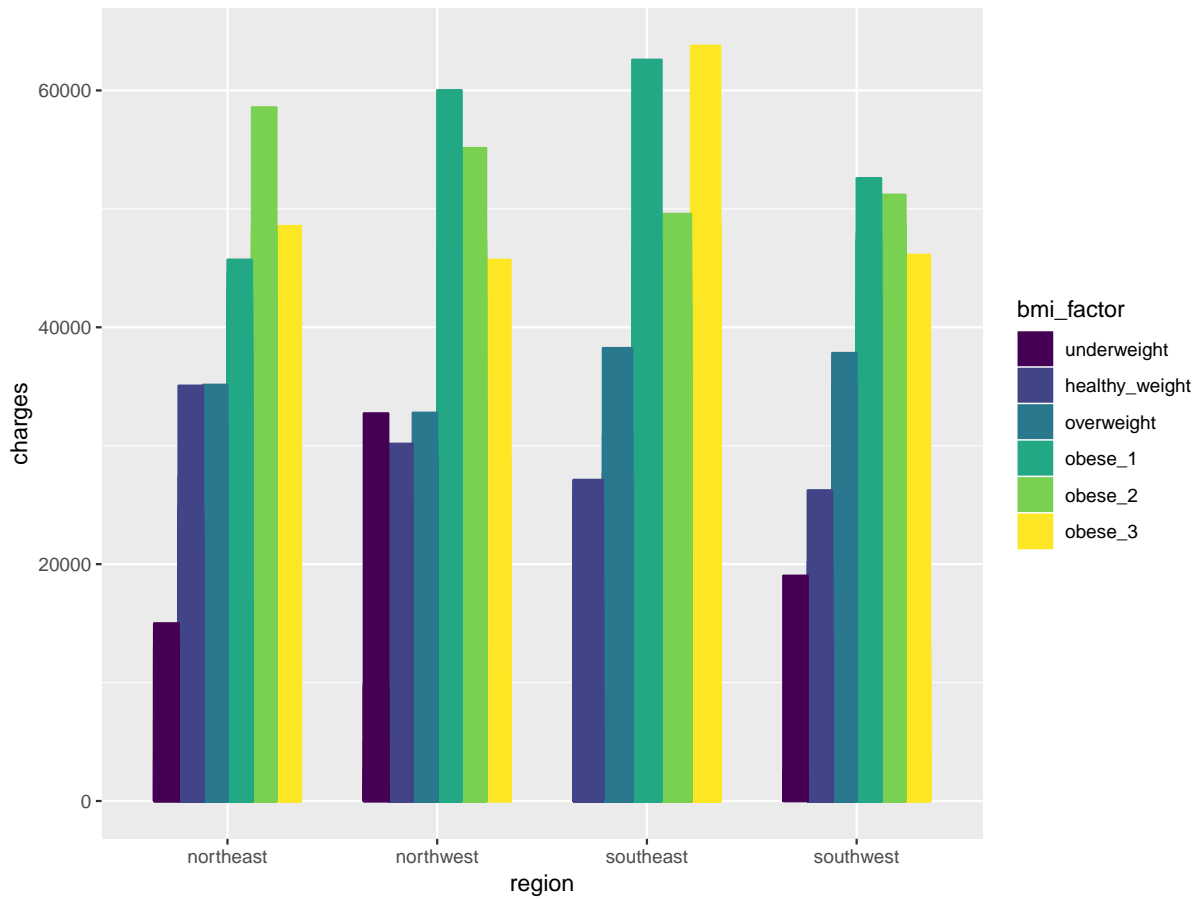
- Charges decrease with higher numbers of children.
- Women do not have higher health charges than men in regard to the number of children.



### 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.





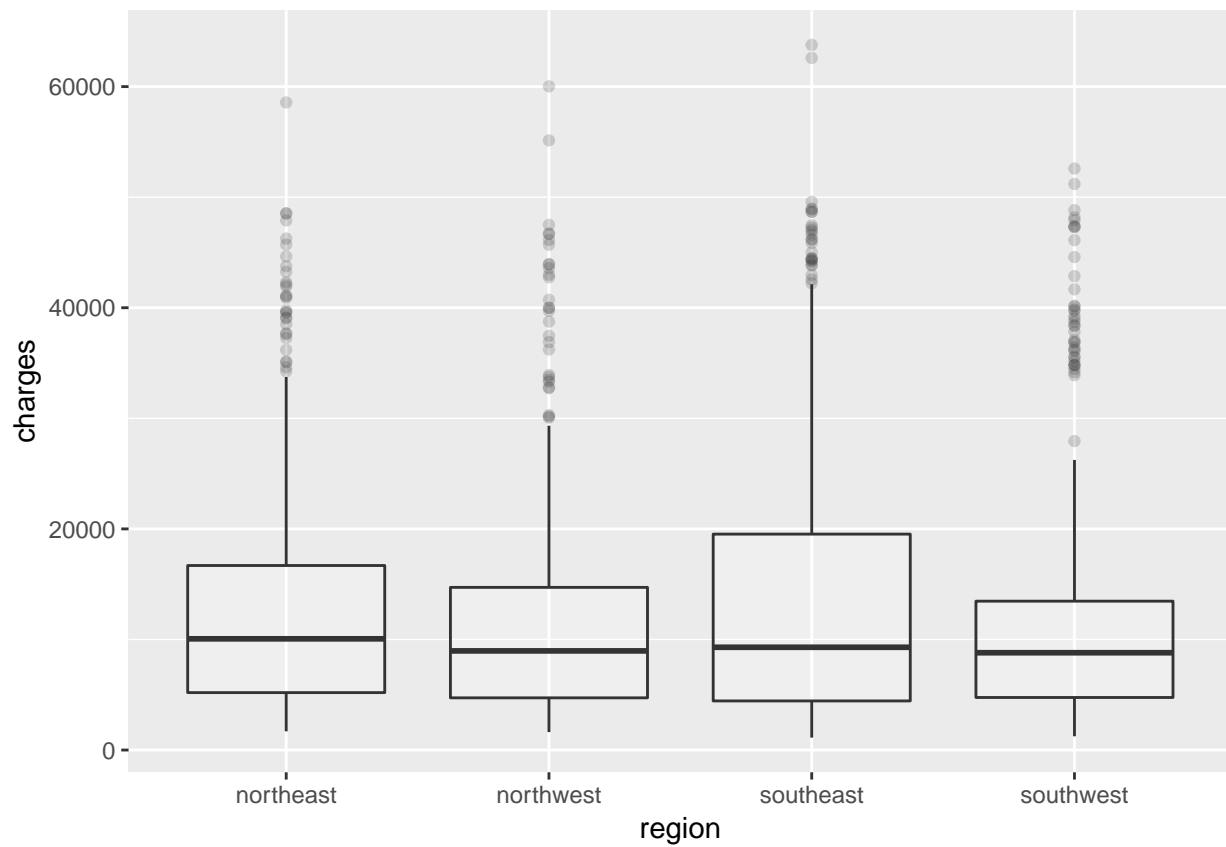
### Region's effect on charges, considering BMI

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

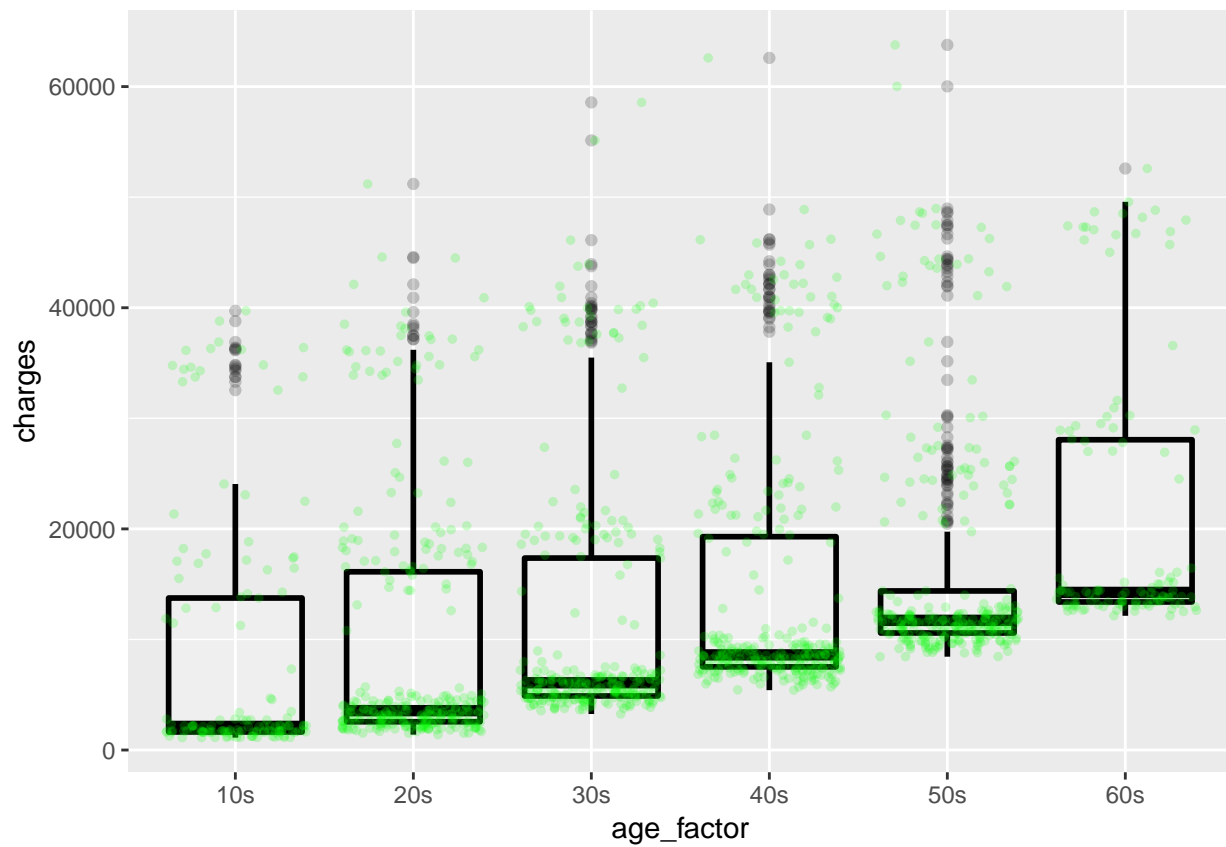
### OUTLIER EXPLORATION

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.

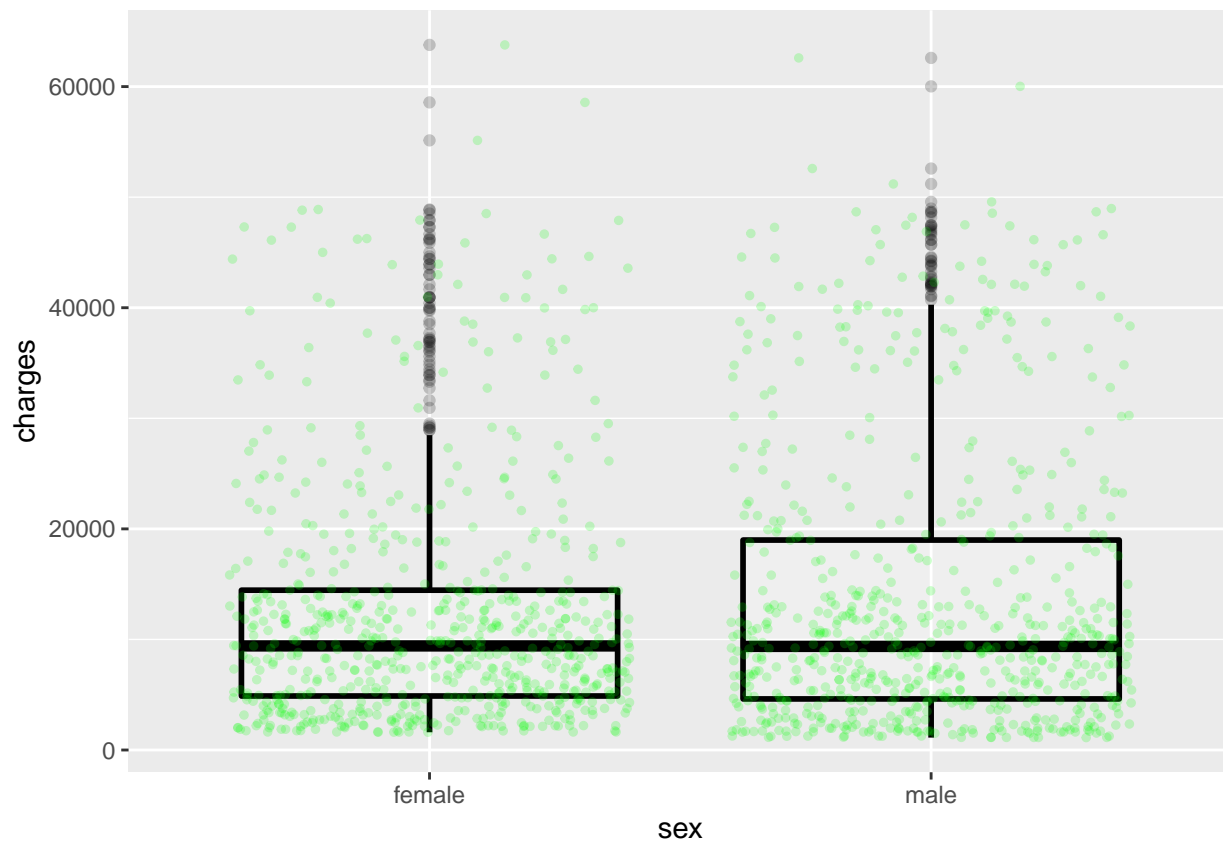
```
ggplot(health_charges_clean, aes(x = region, y = charges)) +
  geom_boxplot(alpha = .2)
```



```
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)
```



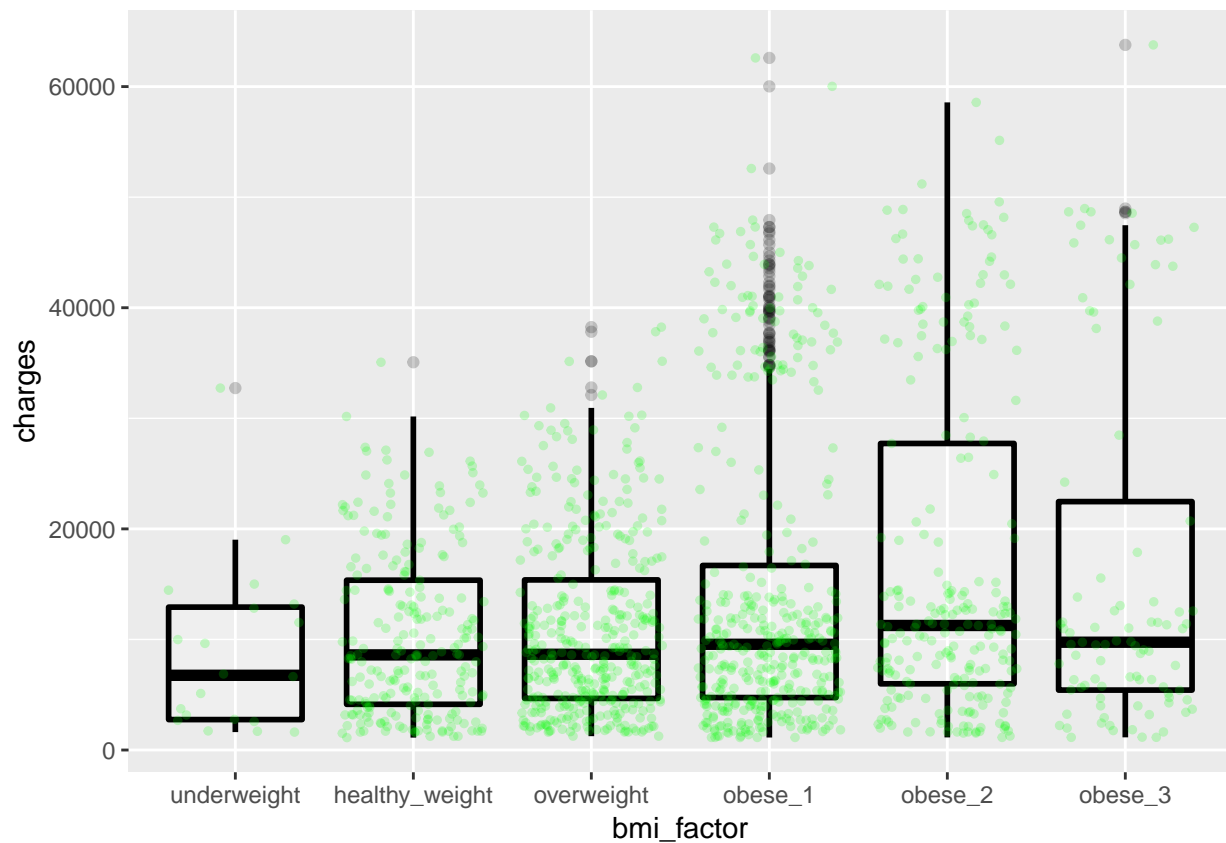
```
ggplot(health_charges_clean, aes(x = sex, y = charges)) +  
  geom_boxplot(color = "black", size = 1, alpha = .2) +  
  geom_jitter(color = "green", size = 1, alpha = .2)
```



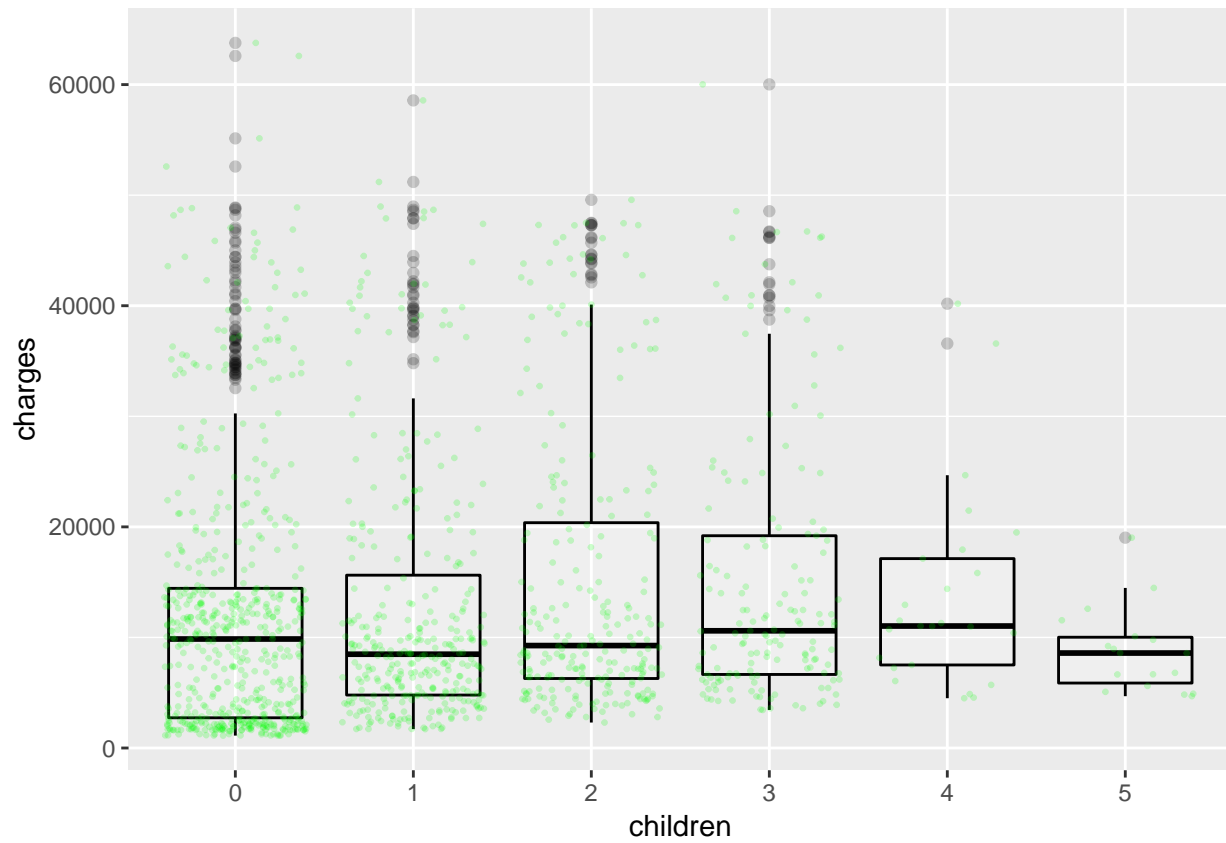
```
levels(health_charges_clean$bmi_factor)
```

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

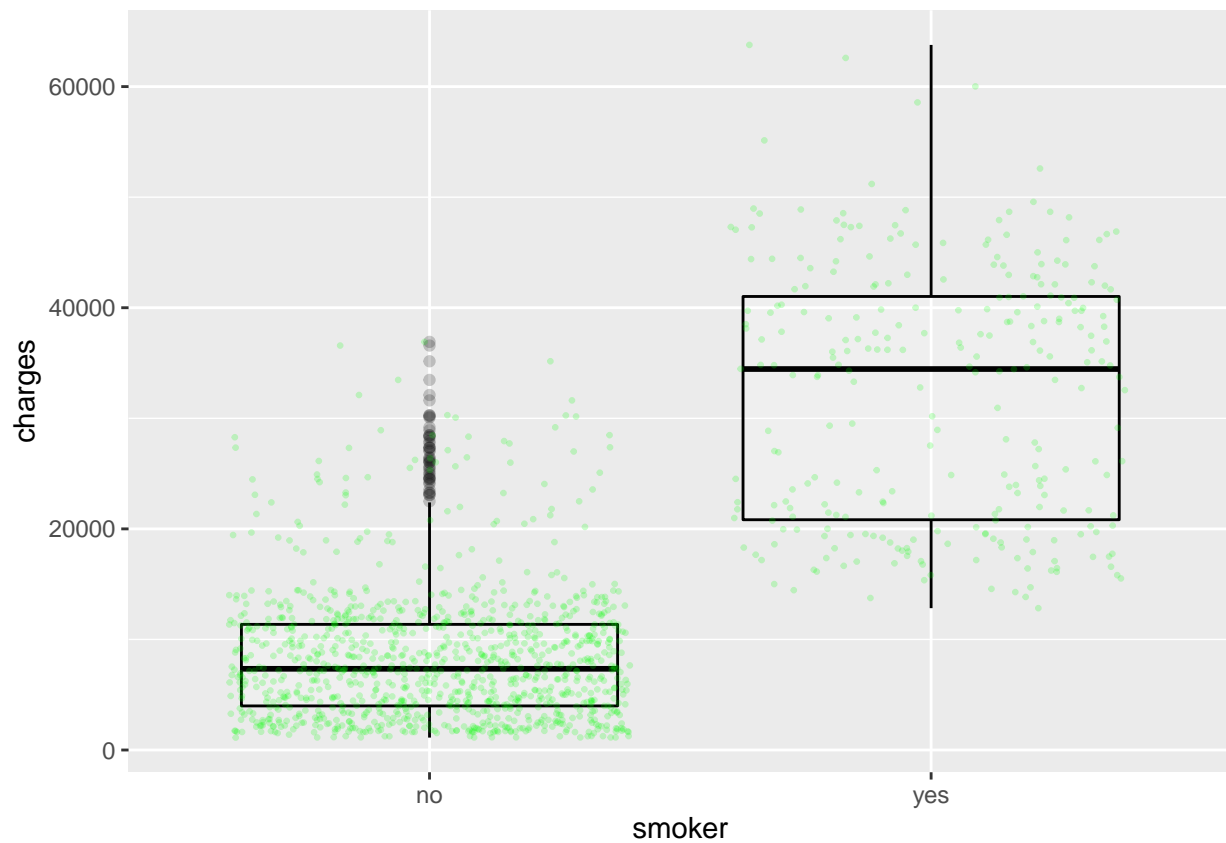
```
ggplot(health_charges_clean, aes(x = bmi_factor, y = charges))+
  geom_boxplot(color = "black", size = 1, alpha = .2) +
  geom_jitter(color = "green", size = 1, alpha = .2)
```



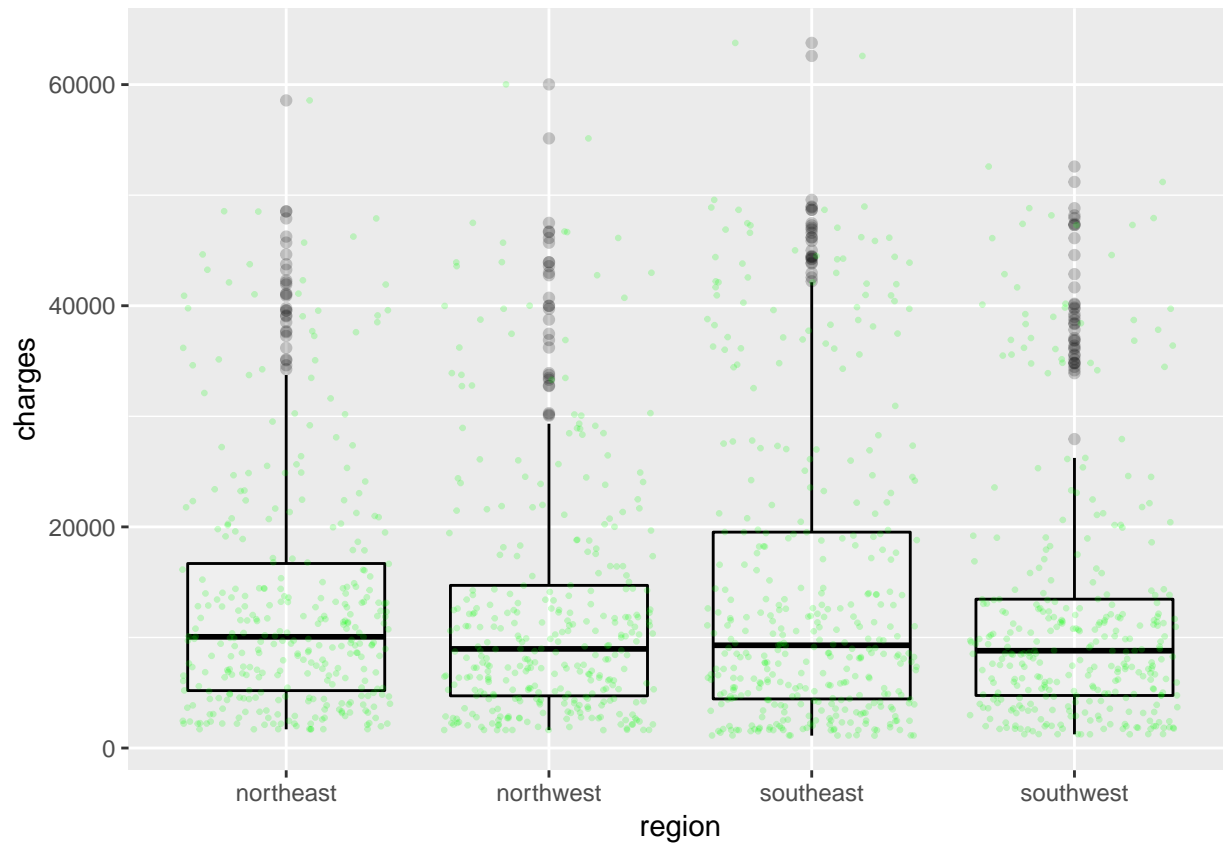
```
childrenf <- as.factor(health_charges_clean$children)
ggplot(health_charges_clean, aes(x = children, y = charges))+
  geom_boxplot(color = "black", alpha = .2) +
  geom_jitter(color = "green", size = .5, alpha = .2)
```



```
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 of full dataset:
  - Adjusted R-squared: .7494, so the 75% of the value of the charges can be attributed to these variables.
  - Significant variables:
    - being a smoker increases charges by \$23848.9
    - having higher numbers of children increases charges by \$475.5
    - higher bmi increases charges by \$339.2
    - higher age increases charges by \$256.9

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

```
##
## Call:
## lm(formula = charges ~ bmi + age + smoker + sex + children +
##     region, data = health_charges_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11689.4  -2902.6   -943.7   1492.2  30042.7
##
```



```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11927.17    993.66 -12.003 < 2e-16 ***
## bmi           336.91      28.61  11.775 < 2e-16 ***
## age          257.19      11.91  21.587 < 2e-16 ***
## smokeryes     23836.41    414.14  57.557 < 2e-16 ***
## sexmale       -128.16     332.83  -0.385 0.700254
## children1      390.98     421.35   0.928 0.353619
## children2     1635.78     466.67   3.505 0.000471 ***
## children3      964.34     548.10   1.759 0.078735 .
## children4     2947.37    1239.16   2.379 0.017524 *
## children5     1116.04    1456.02   0.767 0.443514
## regionnorthwest -380.04     476.56  -0.797 0.425318
## regionsoutheast -1033.14    479.14  -2.156 0.031245 *
## regionsouthwest -952.89     478.15  -1.993 0.046483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6059 on 1325 degrees of freedom
## Multiple R-squared:  0.7519, Adjusted R-squared:  0.7497
## F-statistic: 334.7 on 12 and 1325 DF, p-value: < 2.2e-16
```

- Linear model of “high” charges:
  - 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.

```
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
```

- 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

```
#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

Create testing and training set:

```
library(caTools)
set.seed(88)
hc1 <- health_charges_clean
split = sample.split(hc1$charges_factor, SplitRatio = .75 )
hc1train = subset(health_charges_clean, split == TRUE)
hc1test= subset(health_charges_clean, split == FALSE)
```

- Important variables:
  - Being a smoker.
  - Age.
  - BMI: overweight, obese1, obese2

```
lgall= glm(charges_factor ~ bmi_factor + age_factor + smoker + children + sex + region, data = hc1train)
summary(lgall)
```

```
##
## Call:
## glm(formula = charges_factor ~ bmi_factor + age_factor + smoker +
##      children + sex + region, family = binomial, data = hc1train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35982  -0.41472  -0.30250   0.05371   2.96045
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.5468     0.4122  -8.605  < 2e-16 ***
## bmi_factor.L    2.4226     0.8129   2.980  0.00288 **
## bmi_factor.Q   -1.8496     0.7191  -2.572  0.01011 *
## bmi_factor.C    0.3035     0.5288   0.574  0.56595
## bmi_factor^4   -0.2284     0.3793  -0.602  0.54720
## bmi_factor^5   -0.1148     0.2681  -0.428  0.66862
## age_factor.L    1.2642     0.4431   2.853  0.00433 **
## age_factor.Q   -0.5880     0.4303  -1.366  0.17185
## age_factor.C    0.6597     0.3420   1.929  0.05373 .
## age_factor^4   -0.6558     0.2991  -2.192  0.02835 *
## age_factor^5   -0.2382     0.2930  -0.813  0.41638
```

```
## smokeryes          6.0056      0.4159  14.440 < 2e-16 ***
## children1          0.4251      0.3477   1.223  0.22145
## children2          0.5925      0.3691   1.605  0.10847
## children3          0.2719      0.4358   0.624  0.53264
## children4          1.7382      0.6443   2.698  0.00698 **
## children5          0.3292      1.2520   0.263  0.79258
## sexmale            -0.1030      0.2555  -0.403  0.68694
## regionnorthwest    -0.2578      0.3753  -0.687  0.49225
## regionsoutheast     0.1401      0.3461   0.405  0.68563
## regionsouthwest    -0.5529      0.3859  -1.433  0.15199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 1128.6  on 1002  degrees of freedom
## Residual deviance:  471.2  on  982  degrees of freedom
## AIC: 513.2
##
## Number of Fisher Scoring iterations: 6
```

Prediction:

25% ability to predict high charges, and 92% ability to predict low charges.

```
predicttrain = predict(lgall, type = "response")
tapply(predicttrain, hc1train$charges_factor, mean)
```

```
##          low          high
## 0.08317371 0.75081025
```

Confusion matrix on training set with .5 threshold:

93.02% accuracy of predicting high health charges

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

```
##
##          FALSE TRUE
## low      739   13
## high     59  192
```

```
(193 + 740) / ( 193 + 58 + 12 + 740)
```

```
## [1] 0.9302094
```

Confusion matrix on testing set with .5 threshold:

Our model has 91.94% accuracy of predicting high health charges

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

```
##
##          FALSE TRUE
## low      246    5
## high     22   62
```

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

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

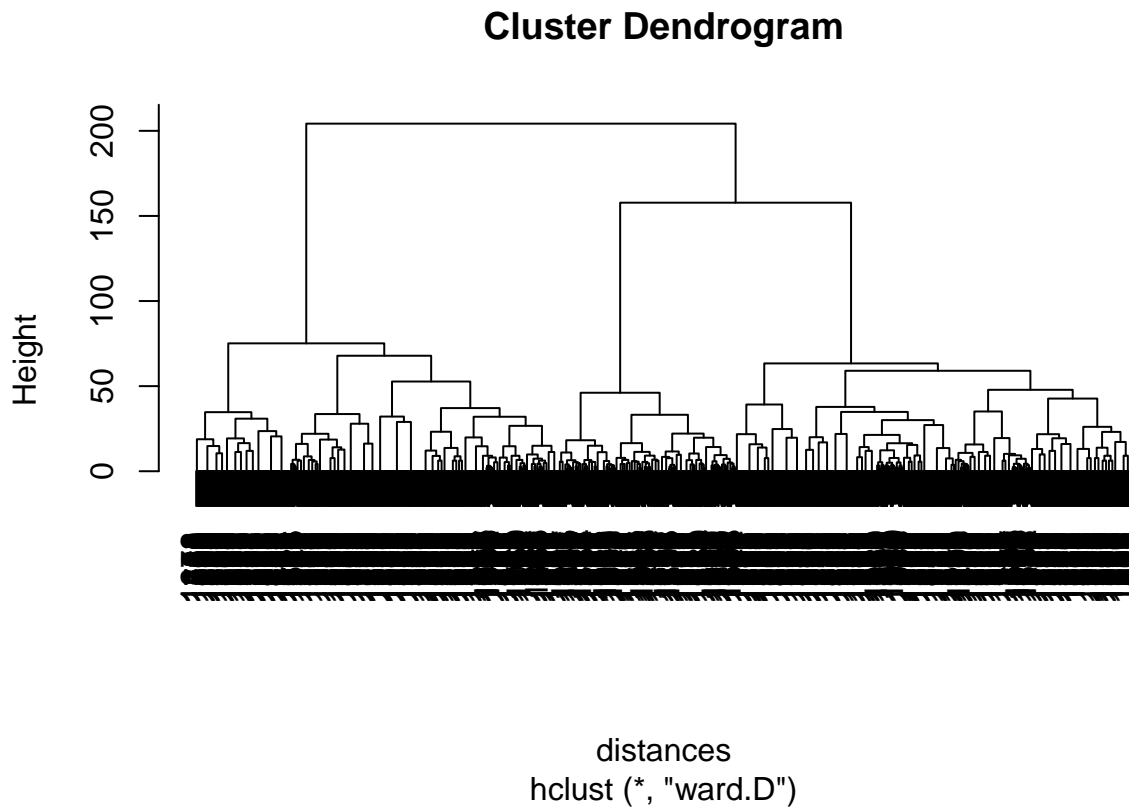
## CLUSTERING:

- Heigharchial clustering with dendrogram; all variables aside from charges.

```
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"
```

```
plot(cluster1)
```



- 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.

```
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(smokev)
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, ch2v, ch3v, ch4v, ch5v, nev, nwv, sev)

```

```

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
## [7,] 0.08148148 0.08148148 1.00000000 0.00000000 0.00000000 0.0000000
## [8,] 0.08854167 0.08854167 0.97395833 0.01562500 0.34375000 0.3020833
##           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
## [7,] 1.00000000 0.00000000 0.000000000 0.4592593 0.2074074 0.2148148
## [8,] 0.04166667 0.29687500 0.000000000 0.2760417 0.2760417 0.2291667
##           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
## [7,] 0.1185185 0.000000000 0.000000000 0.28148148 0.3185185 0.10370370
## [8,] 0.1510417 0.046875000 0.020833333 0.32812500 0.3229167 0.01041667
##          swv
## [1,] 0.1897233
## [2,] 0.4308300
## [3,] 0.1121495
## [4,] 0.0000000
## [5,] 0.3692308
## [6,] 0.1515152
## [7,] 0.2962963
## [8,] 0.3385417
```

---

## CONCLUSION:

By far, the most significant predictor of high health charges was being a smoker. Other important predictors were age and bmi. Children was a significant predictor for the general population's charges, but not a significant predictor for the data subset of "high" charges. 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. Collectively, our variables were impactful enough to be able to predict high charges at 91.94% accuracy using logistic regression. Clustering drew similar conclusions to the linear regression models.

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. As smoking is an addiction, it borders on something between a behavioral choice and a mental health condition.

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, caffeine consumption, alcohol consumption, and diet. Accurately collecting that data and having people report their status accurately would be challenging. Technology such as smart watches and more could eventually be used for this data collection. Perhaps people could receive reduced rates to incentivize cost-lowering behaviors and encourage data sharing.

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 of our predictive models.