Influence of Age and BMI on US Insurance Costs

Phia Rau Halleen

12/12/2021

#Introduction: For this project, I am using the data set "insurance.csv" to examine the potential relationship between an individuals age and the insurance charges they receive, as well as the potential relationship between an individuals BMI (body mass index) and their insurance charges. I am focusing on these variables (of the seven in the data set) because of the noted presence of ageism and fat phobic discrimination in the medical field, and the insurance industry as an extension of that. Insurance is not cheap or free in the United States. With this, there is a certain misnomer that insurance costs more for individuals with preexisting conditions — many of which do not have to be medical — the age of the main insurer, number of dependents and individuals BMI are all charted to determine the cost of insurance charges. This means that individuals with preexisting conditions, such as age and BMI, are going to be charged for the same care as people who "cost less" in the eyes of the insurance company. In the US currently, individuals with a BMI above 30 (30<= considered obese, 25<= X >= 28 considered overweight) are expected to pay 25% to 50% more for insurance than individuals with a BMI of less than 30. About 42.4% of the US population currently has a BMI of 30 or above. That means that about 139.7 million Americans are paying more for insurance, and likely receive less quality care too. The same goes for individuals 55 and over, however, their percentage of the US population is . between the cost "charges" of US insurance, and an individuals age, as well as an individuals BMI.

Because of the presence of discrimination in the medical field, I am carrying out this analysis to see if there is a positive correlation relationship between the age/bmi of the individual and the charges they incurred. I am predicting that a as the age or BMI of the individual increases, as will the charges.

```
library(readr)
```

#The dataset: "insurance.csv" was pulled off of kaggle "https://www.kaggle.com/mirichoi0218/insurance". It was initally used in the textbook "Machine Learning with R" by Brett Lantz. There is currently not a known year for the dataset.

From first glance, the dataset has 1338 observations across seven variables. The seven variables are "age", "sex", "bmi", "children" (i.e. the number of dependents under 18), "smoker", "region" (of the United States: broken down into Northeast, Northwest, Southwest or Southeast), and lastly "charges" (amount of money an individual is billed by their insurance company). The variables I am specifically interested in for this analysis are the variables of "age", "bmi" and "charges".

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
library(readxl)
insurance_data<-read.csv("insurance.csv", header = TRUE)
str(insurance_data)
```

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

```
summary(insurance_data)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

#Visualizations: For the visualizations, I first used tidyverse to filter out the specific variables I wanted. For the bmi variable, I filtered out for a bmi more than or equal to 30, as well as less than or equal to 30

```
bmi_over_30.1<-insurance_data%>%
  filter(bmi >=30,)



bmi_under_30<-insurance_data%>%
  filter(bmi <=30,)
```

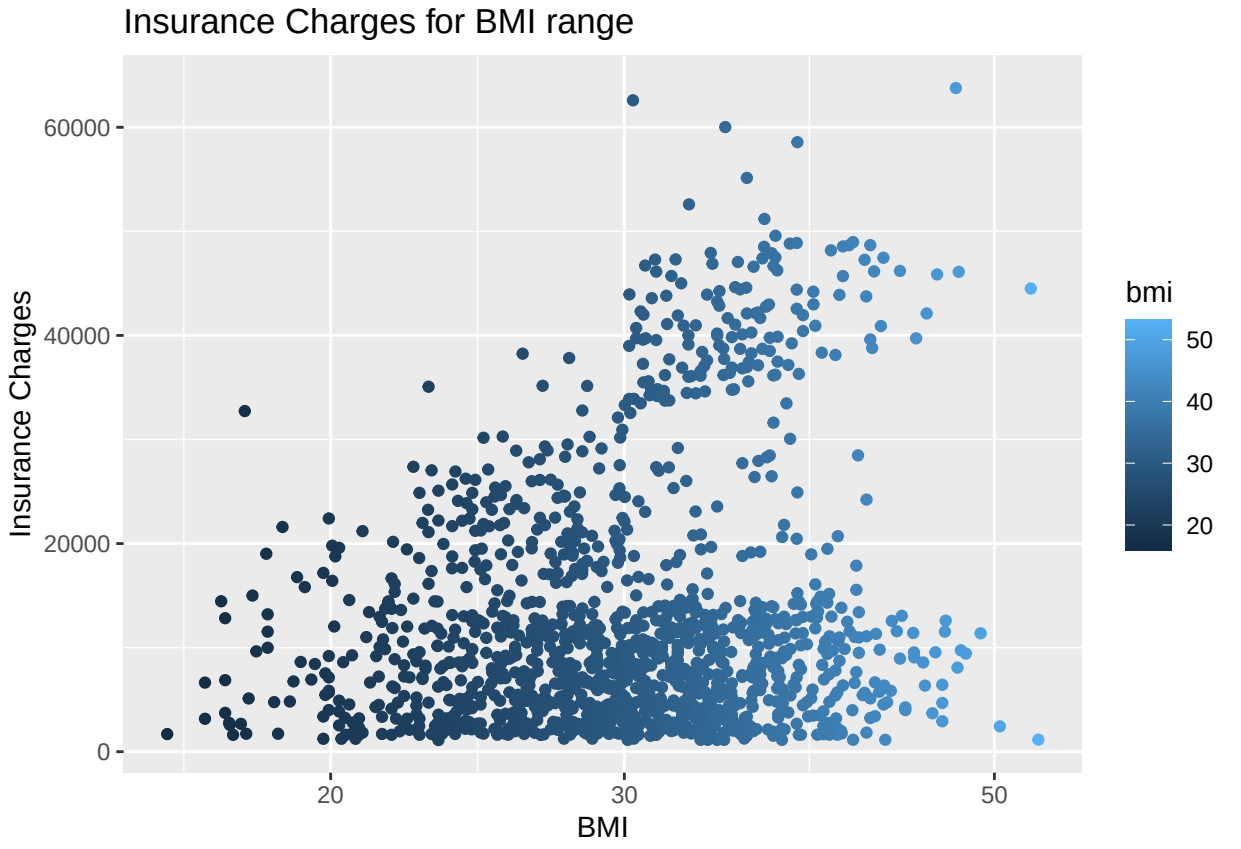
I also filtered out what I was looking for for the age variable, specifically less than or equal to 55, and more than or equal to 55. I specified the age 55 to be the cut off because usually 55 is the age that insurees can expect to pay more compared to their younger counterparts.

```
age_under_55<-insurance_data%>%
  filter(age <=55, na.rm = TRUE)
```

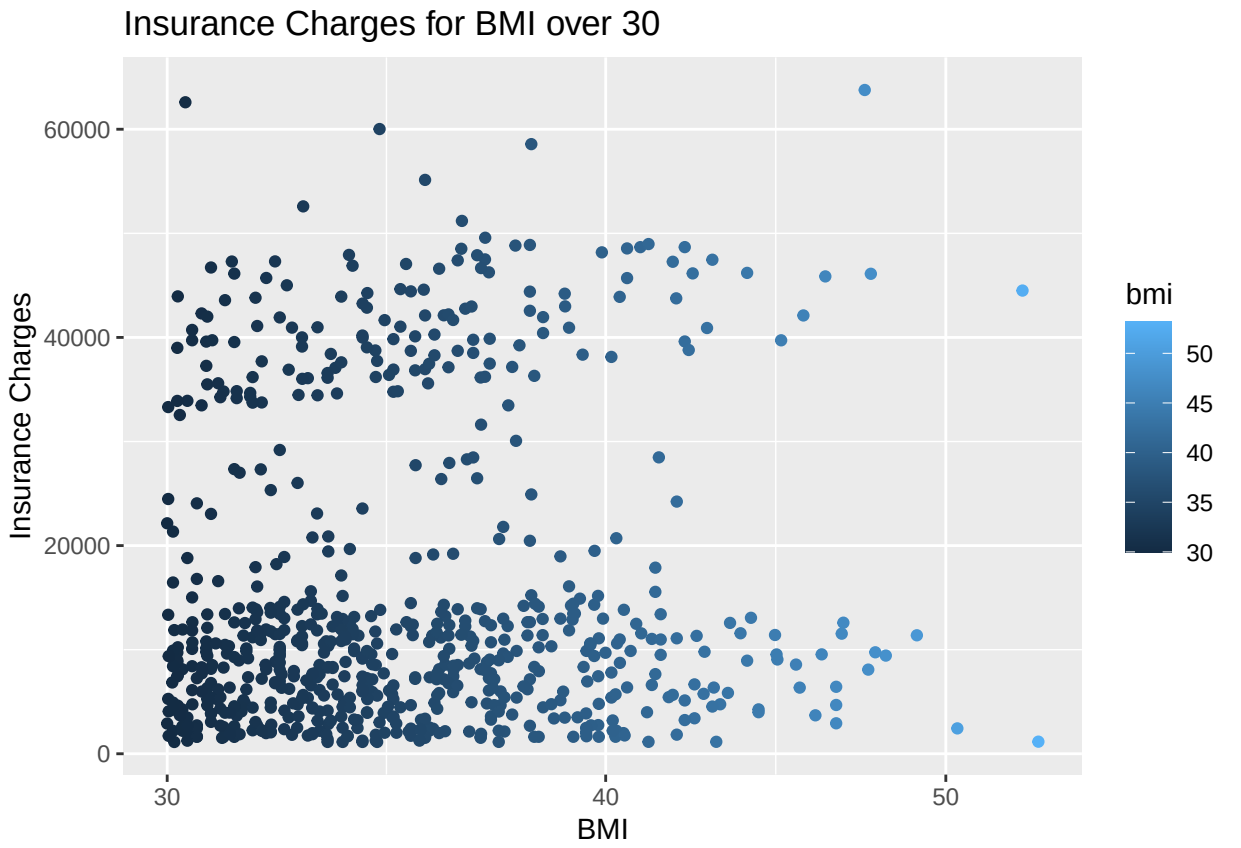
```
age_over_55<-insurance_data%>%
  filter(age >=55,)
```

For all of the visualizations, I chose charges as the y or "response" variable because I was interested in checking the potential influence of age or bmi on the charges. For the visualizations, I also stuck with a scatter plot because of both my experience with this visualization, and because I wanted to physically see the potential outliers outside of the line of best fit. First, I visualized BMI and charges, and then age and charges without the best fit line. The graphs I made were of: the whole BMI sample, individuals with a BMI of 30 or above, and 30 or below. I then applied the best fit line to this graphs, in that order.

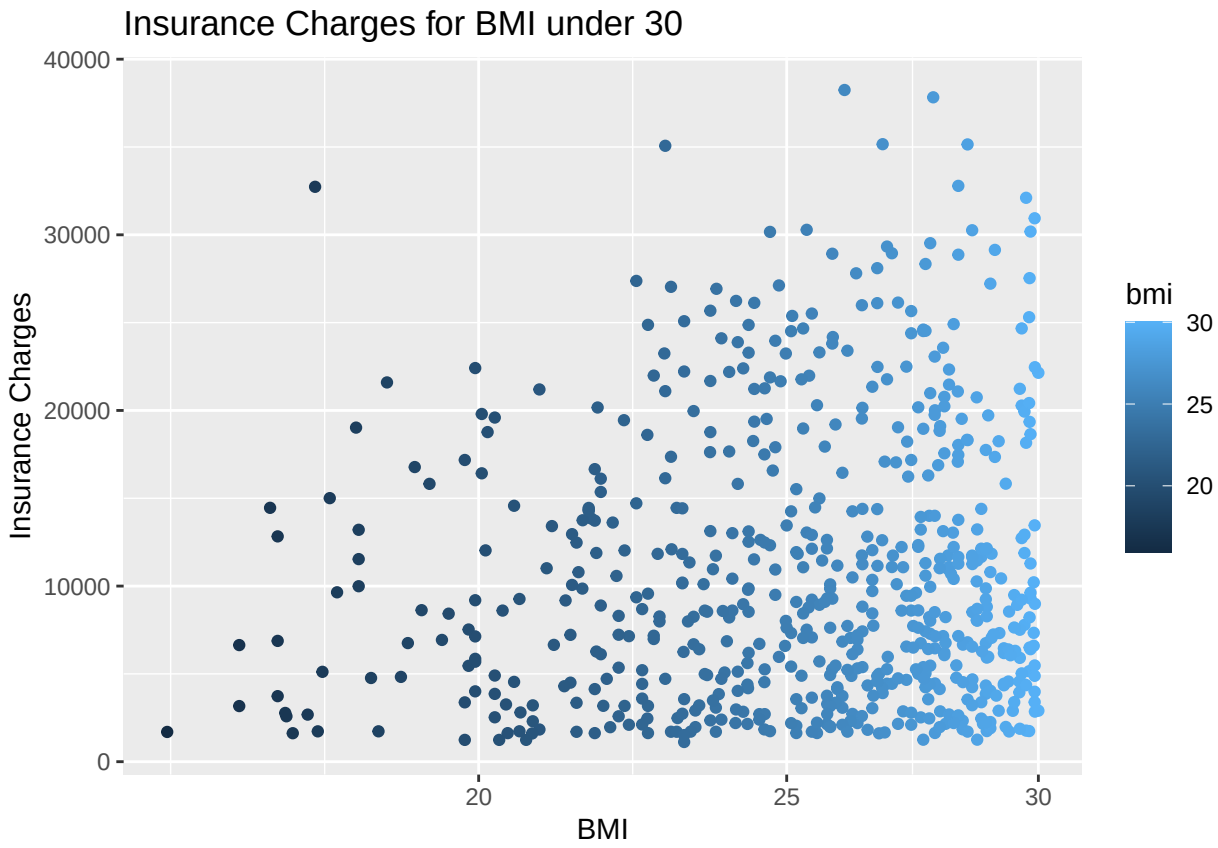
```
ggplot(data=insurance_data, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
  xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI range")
```




```
ggplot(bmi_over_30.1, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
  xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI over 30")
```

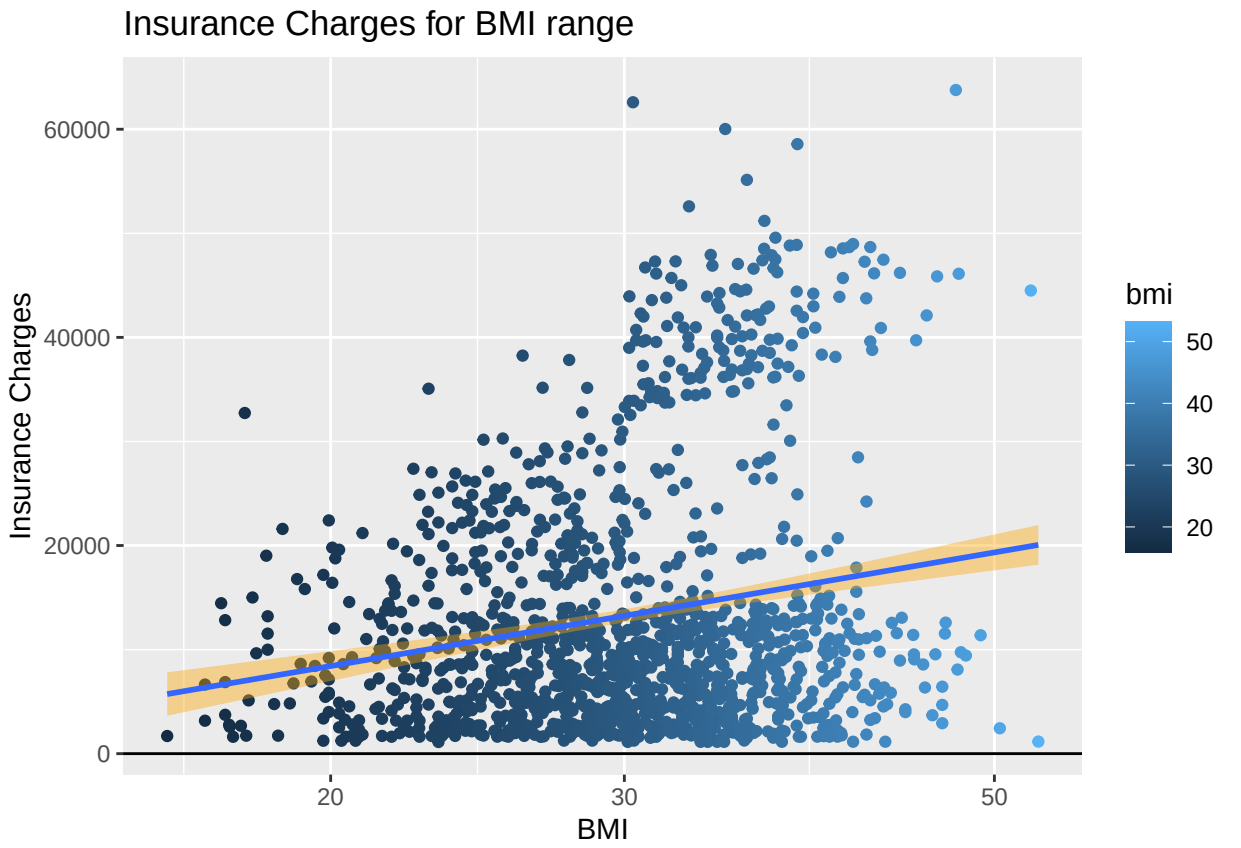



```
ggplot(data=bmi_under_30, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
  xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI under 30")
```



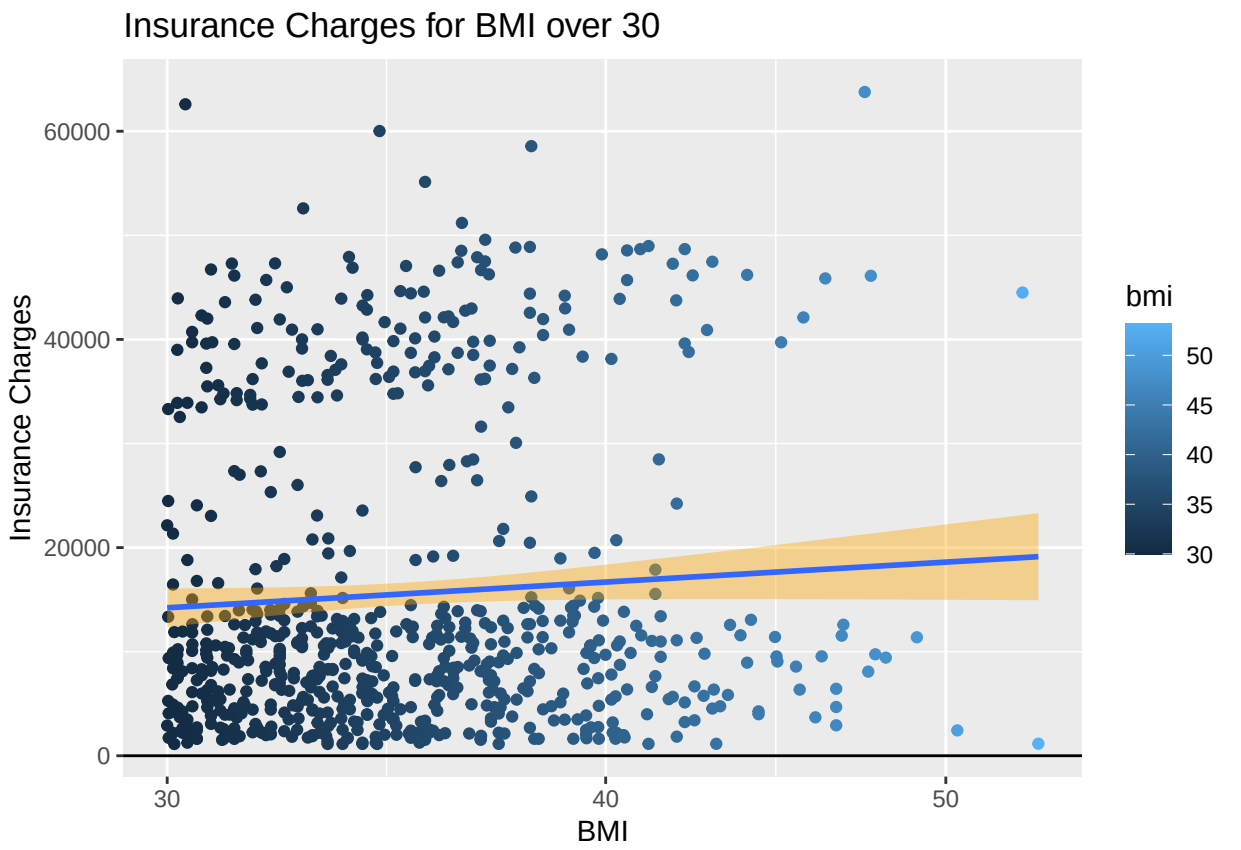

```
ggplot(data=insurance_data, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0,) +
  geom_smooth(method = "lm", se = TRUE, fill="orange")+
    xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI range")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



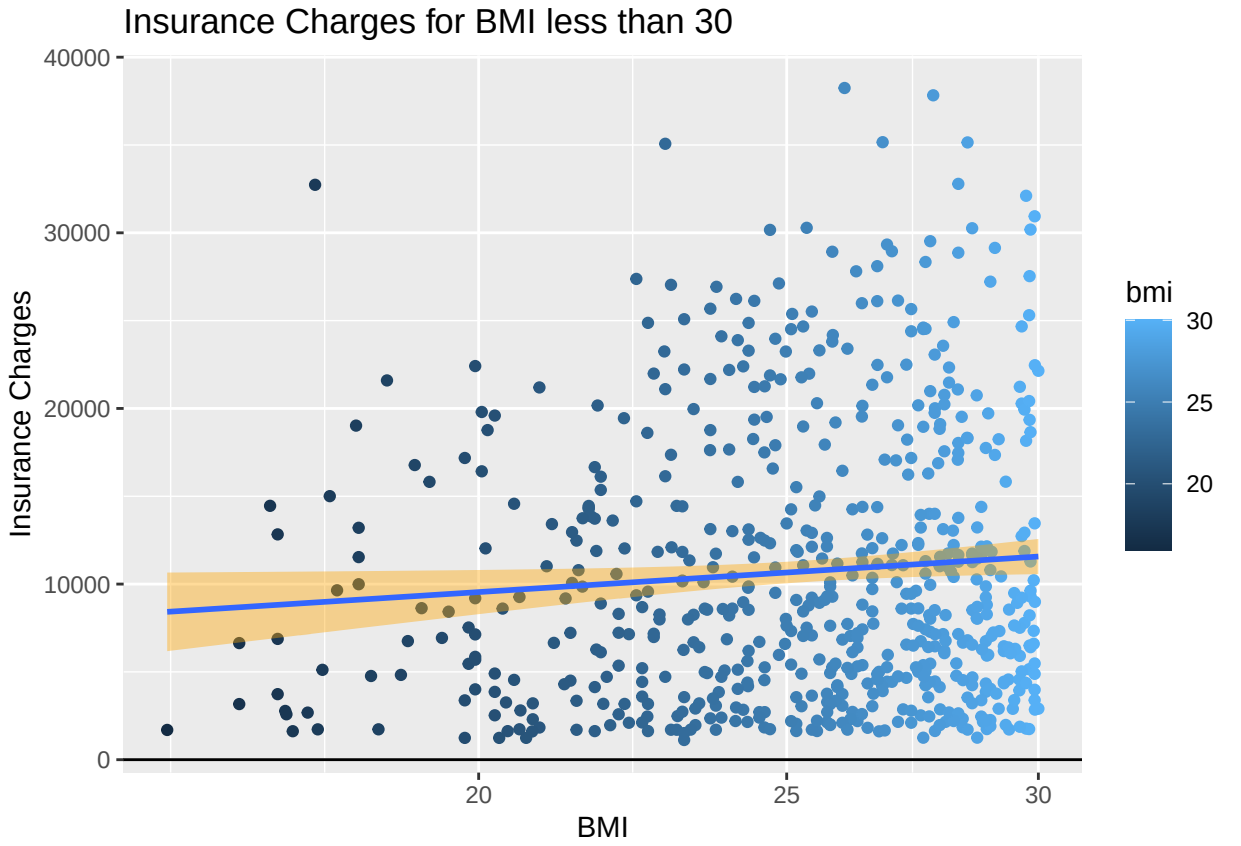

```
ggplot(data=bmi_over_30.1, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = "lm", se = TRUE, fill = "orange")+
    xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI over 30")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



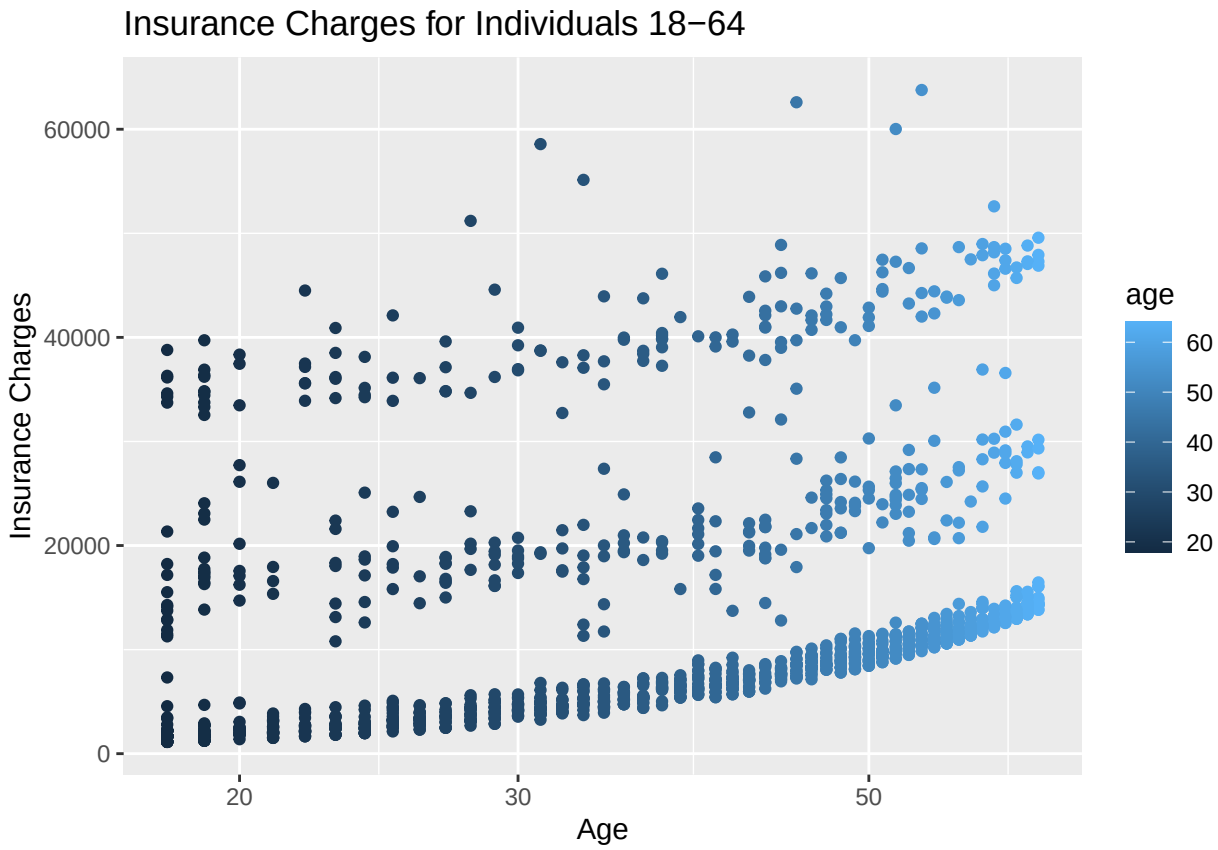

```
ggplot(data=bmi_under_30, aes(x=bmi, y=charges, color=bmi,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = "lm", se = TRUE, fill = "orange")+
    xlab("BMI")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for BMI less than 30")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

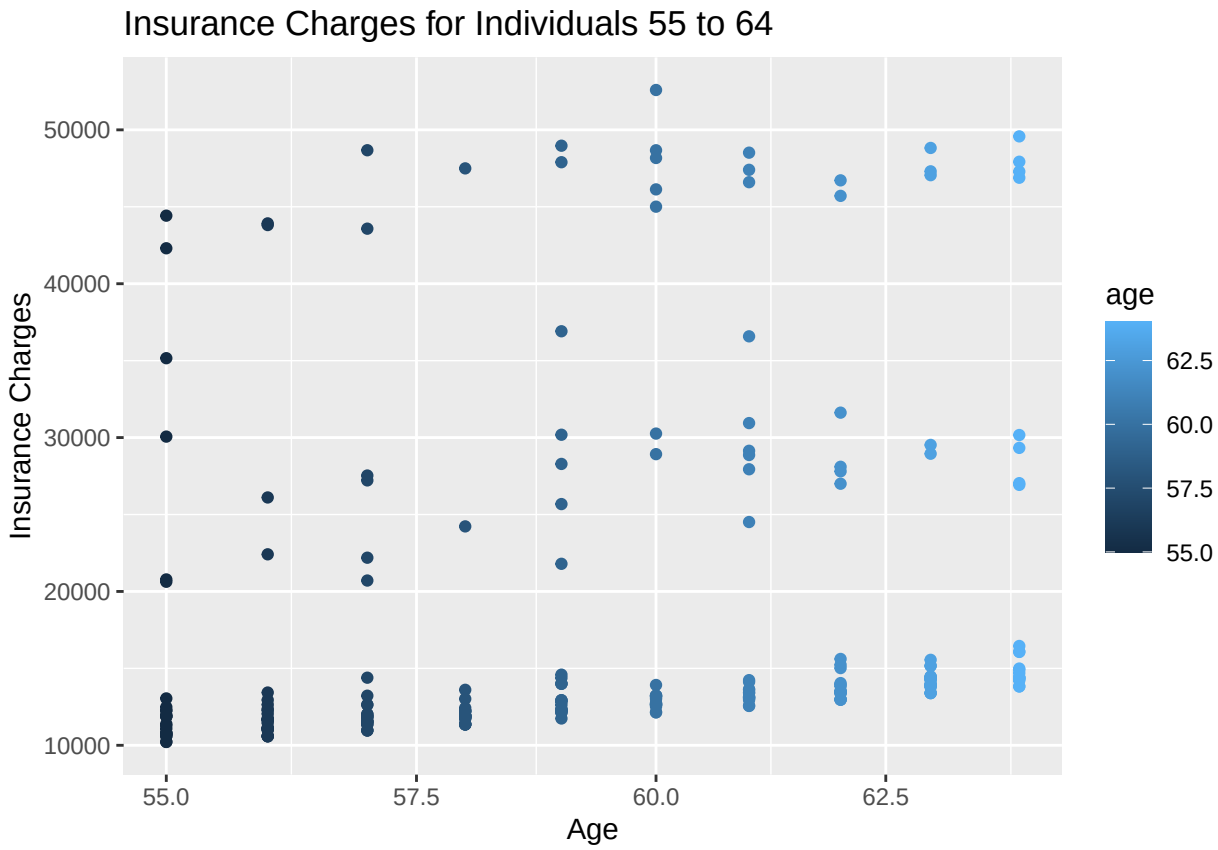


I then visualized the age and charges variable. I first graphed the whole sample (ages 18-64), then individuals 55 and above, and finally individuals 55 and below. Afterward I applied the best fit line to these graphs.

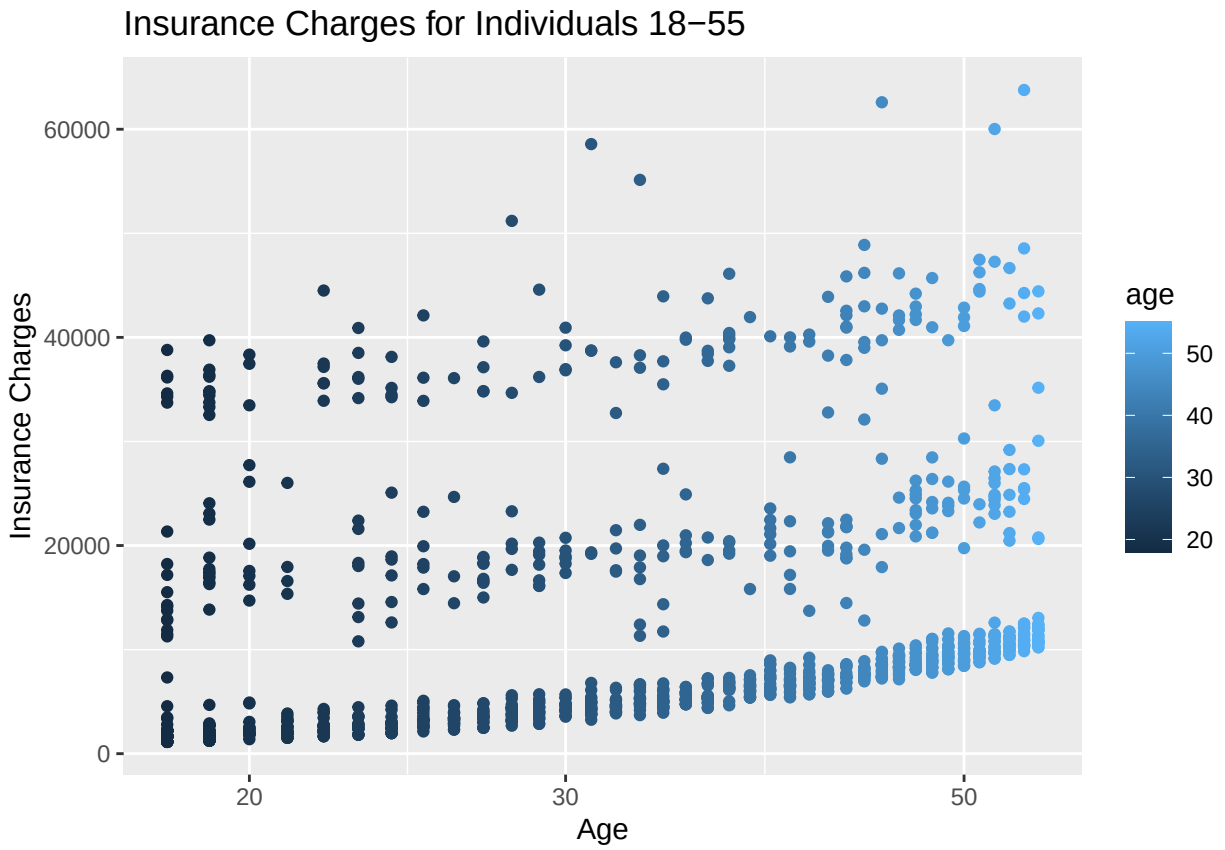
```
ggplot(data=insurance_data, aes(x=age, y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
  xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals 18-64")
```




```
ggplot(data=age_over_55, aes(x=age, y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
  xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals 55 to 64")
```

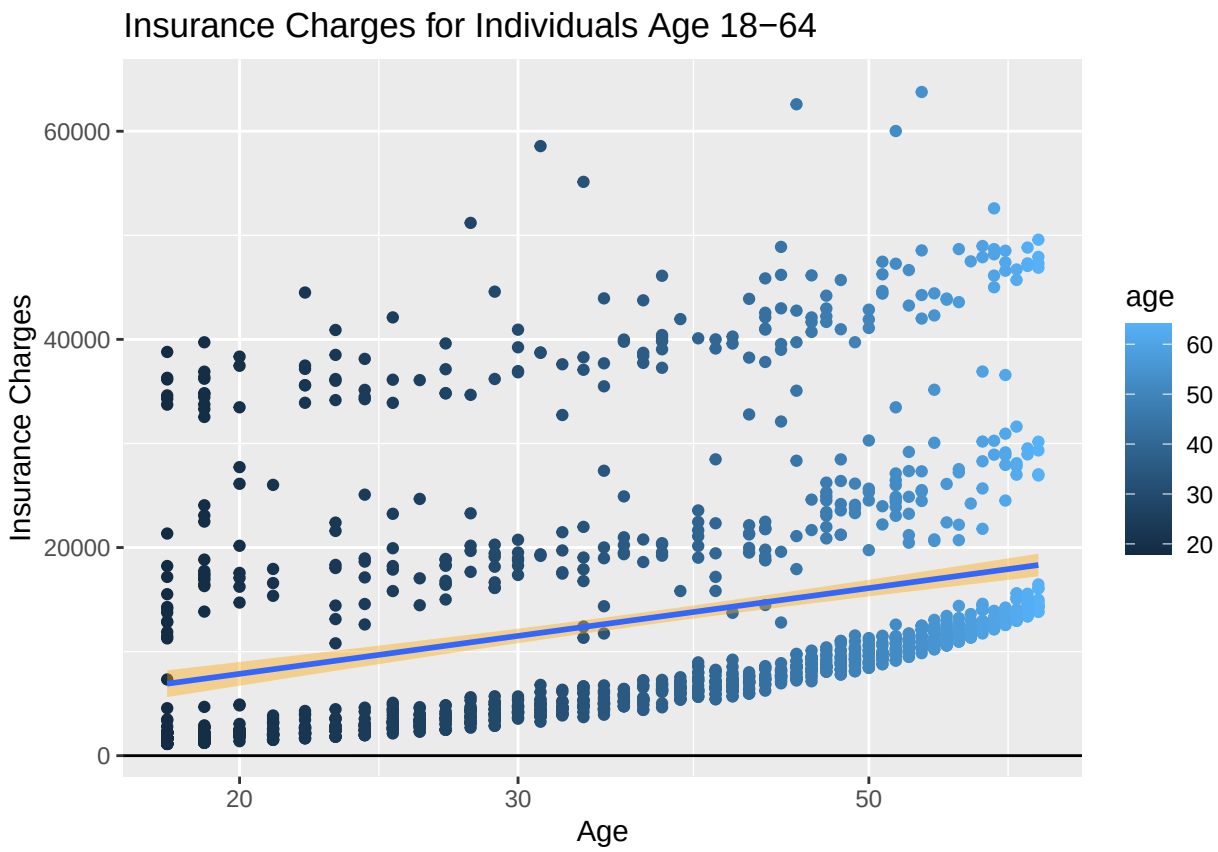



```
ggplot(age_under_55,aes(x=age,y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
  xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals 18-55")
```



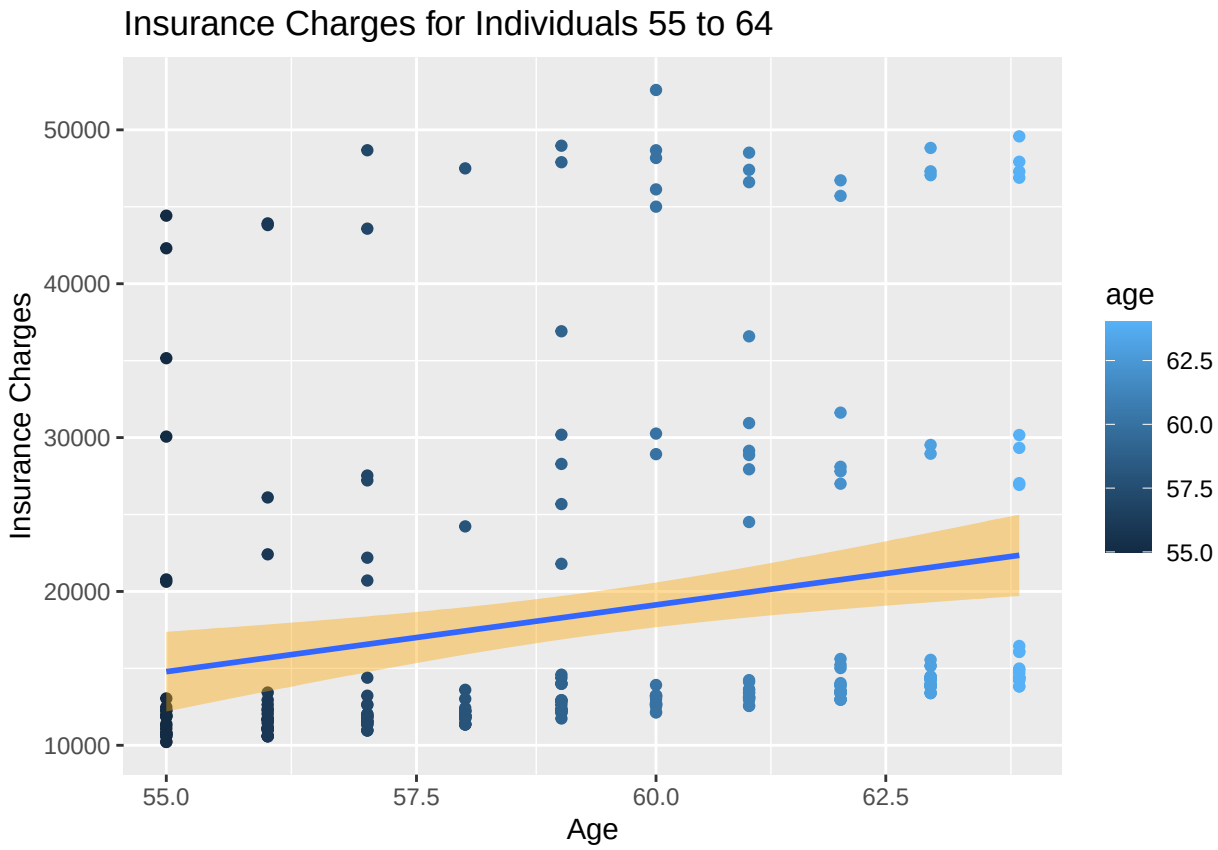

```
ggplot(data=insurance_data, aes(x=age, y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = "lm", se = TRUE, fill="orange")+
 xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals Age 18-64")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



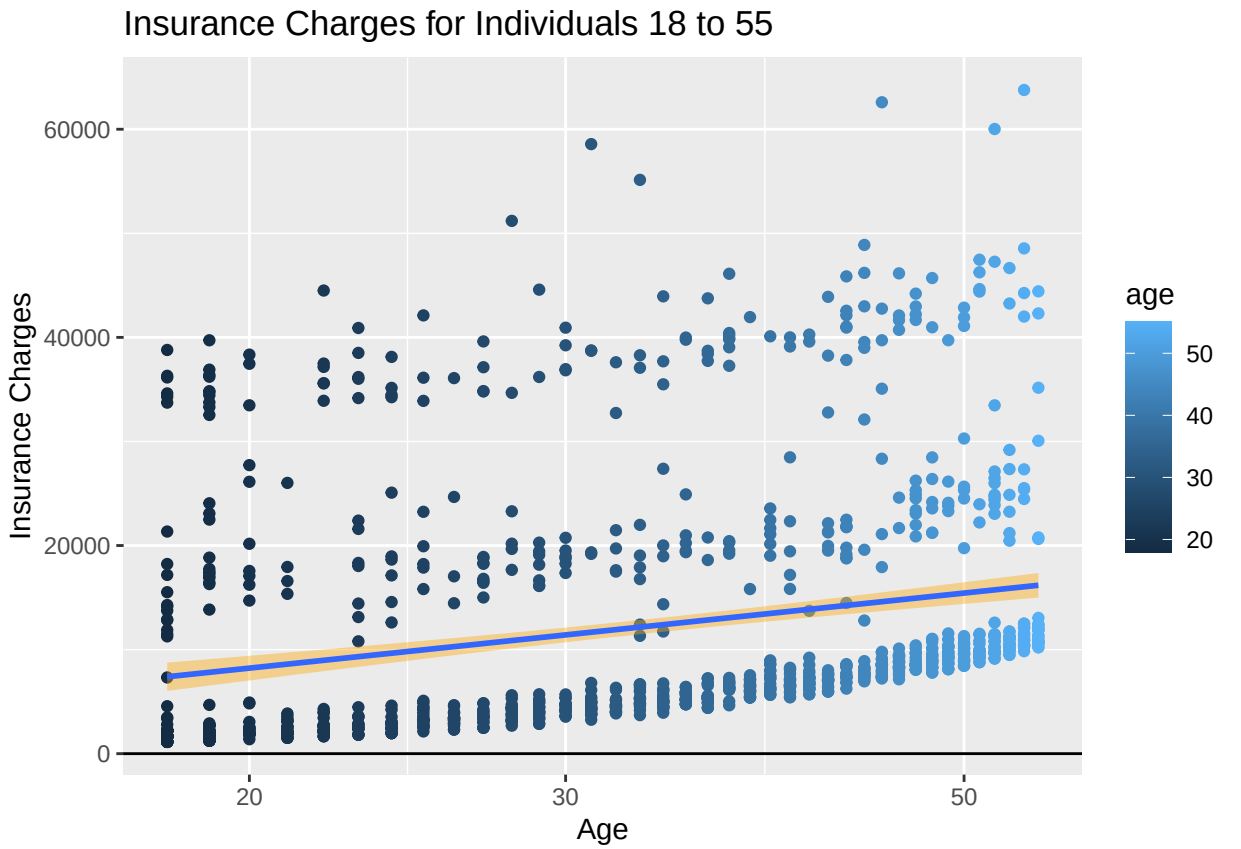

```
ggplot(data=age_over_55, aes(x=age, y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = "lm", se = TRUE, fill="orange")+
 xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals 55 to 64")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(data=age_under_55, aes(x=age, y=charges, color=age,))+
  geom_point()+
  scale_x_log10()+
 geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = "lm", se = TRUE, fill="orange")+
 xlab("Age")+
  ylab("Insurance Charges")+
  ggtitle("Insurance Charges for Individuals 18 to 55")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



#Linear regression I tested the correlation and linear regression between the variable (age or bmi) and the response variable (charges)

First for correlation.

```
cor(insurance_data$bmi,insurance_data$charges)
```

```
## [1] 0.198341
```

```
cor(bmi_over_30.1$bmi,bmi_over_30.1$charges)
```

```
## [1] 0.06279025
```

```
cor(bmi_under_30$bmi, bmi_under_30$charges)
```

```
## [1] 0.07837246
```

Of the three correlations tested, correlation for bmi versus charges, bmi above 30 and bmi under 30, the correlation was the highest between bmi and charges was largest for the whole sample, versus the filtered sample. Overall, the corrrelation was lower than expected.

```
cor(insurance_data$age,insurance_data$charges)
```

```
## [1] 0.2990082
```

```
cor(age_over_55$age,age_over_55$charges)
```

```
## [1] 0.2130481
```

```
cor(age_under_55$age,age_under_55$charges)
```

```
## [1] 0.2384561
```

The correlation shows that for correlation was largest for the general age sample, then the sample of individuals 18-55. This went against my assumption.

```
age_and_charges_model<-lm(formula = charges ~ age, data = insurance_data)
age_and_charges_model
```

```
##
## Call:
## lm(formula = charges ~ age, data = insurance_data)
##
## Coefficients:
## (Intercept)          age
##      3165.9        257.7
```

```
summary(age_and_charges_model)
```

```
##
## Call:
## lm(formula = charges ~ age, data = insurance_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8059  -6671  -5939   5440  47829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3165.9      937.1   3.378 0.000751 ***
## age            257.7       22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

```
bmi_and_charges_model<-lm(formula = charges ~ bmi, data = insurance_data)
bmi_and_charges_model
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = insurance_data)
##
## Coefficients:
## (Intercept)          bmi
##      1192.9        393.9
```

```r
summary(bmi_and_charges_model)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = insurance_data)
##
## 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
```

```r
single_obs_bmi<-tibble(age = 39, bmi = 31,)
single_obs_bmi
```

```
## # A tibble: 1 x 2
##     age   bmi
##   <dbl> <dbl>
## 1    39    31
```

```r
predict(bmi_and_charges_model, single_obs_bmi)
```

```
##     1
## 13403
```

```r
single_obs_age<-tibble(age = 55, bmi = 31)
single_obs_age
```

```
## # A tibble: 1 x 2
##     age   bmi
##   <dbl> <dbl>
## 1    55    31
```

```r
predict(age_and_charges_model, single_obs_age)
```

```
##        1
## 17340.63
```

```r
single_obs_age_and_sex_and_bmi<-tibble(age = 39, bmi = 31, sex = "female",)
single_obs_age_and_sex_and_bmi
```

```
## # A tibble: 1 x 3
##     age   bmi sex
##   <dbl> <dbl> <chr>
## 1    39    31 female
```

```
predict(age_and_charges_model, single_obs_age_and_sex_and_bmi)
```

```
##        1
## 13217.07
```

```
predict(bmi_and_charges_model, single_obs_age_and_sex_and_bmi)
```

```
##     1
## 13403
```

```
single_obs_older_female<-tibble(age = 62, bmi = 31, sex = "female",)
single_obs_older_female
```

```
## # A tibble: 1 x 3
##     age   bmi sex
##   <dbl> <dbl> <chr>
## 1    62    31 female
```

```
predict(age_and_charges_model, single_obs_older_female)
```

```
##        1
## 19144.69
```

#Conclusions: One of the biggest takeaways from this analysis was that the relationship I predicted between age/bmi and charges was not as large as I was expecting. Because of the literature around ageism and fat-discrimination in the medical and social work fields, I assumed that the data may reflect that, at least to a moderate level, however it did not. However, when including the prediction model for both age and bmi, the prediction does show that older adults did pay more, as well as those with a higher bmi. When combining the two, age and bmi, the charges were significantly higher. When then including the variable "sex" this increased the charges for the individual as well, especially when filtering for "female". This implies that this issue of ageism and fatphobia needs to be studied more interdisciplinary in nature, as including sex fit more with my inital hypothesis. In the future, studying this issue should be taken from an interdisciplinary lens.

Questions for further analysis How does the year of the charge inform the amount individuals are charged? How would including race and class impact the outcome of the study? What further information should be collected to give a Can this data set be generalized to the whole US population, and if not what data needs to be collected in order to make this possible? (e.g. class, income, race, sexual orientation, marital status, citizenship status?)