Insurance Costs | EDA, Clustering, and Regression

Antonio Pano

10/6/2022

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(patchwork)
library(reshape2)
library(tidyselect)
library(fmsb)
library(tibble)
```

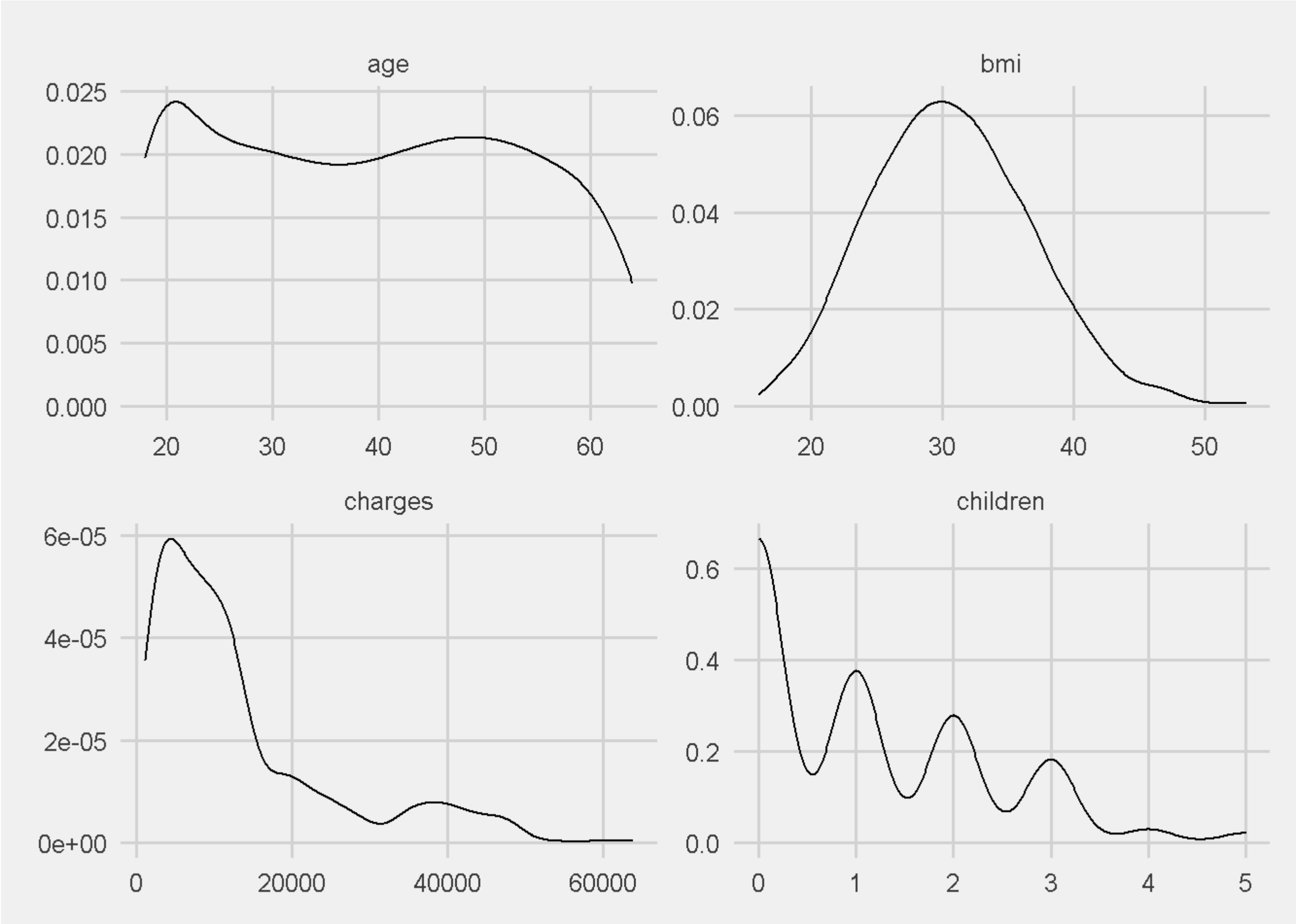
Variables Used:

Age: Age of insured patient.
Sex: Sex of insured patient.
BMI: Body Mass Index of patient
Smoker: Whether the patient smokes or not.
Region: Region of residence.
Charges: Total Insurance Cost.

```
PC <- read.csv("patient_charges.csv",stringsAsFactors = T)
# any(is.na(PC))
```

- Noticing that charges and children count are right skewed.
- Age is somewhat uniform.
- BMI is definitely normally distributed.

```
PC %>%
  purrr::keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density() + ggthemes::theme_fivethirtyeight()
```



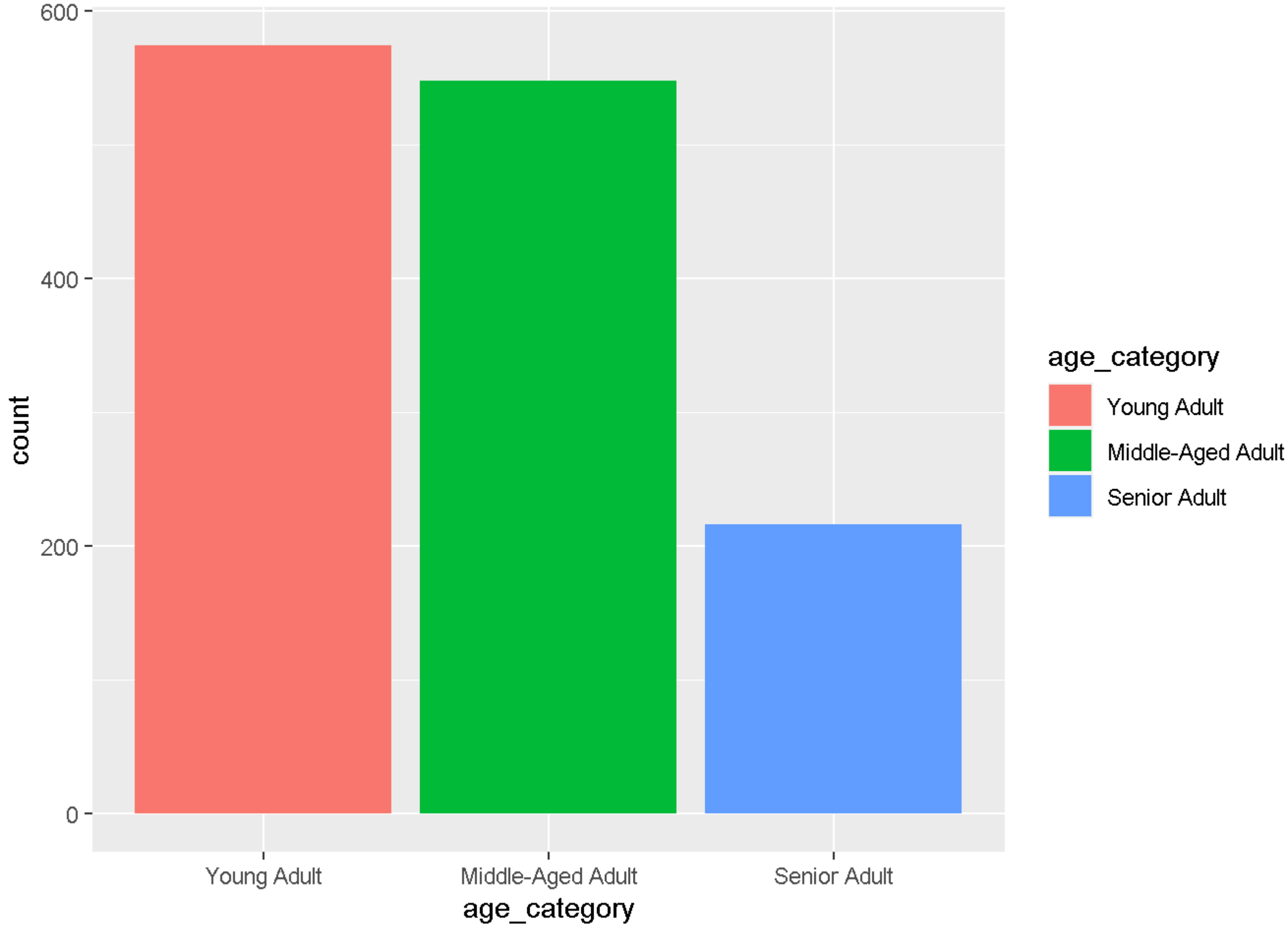
- Feature Engineering categories for `bmi` and `age` .
- Will be useful for visualization.

```
PC <- PC %>%
  mutate(age_category = cut(PC$age, breaks = c(18, 35, 55, Inf),
          labels = c("Young Adult", "Middle-Aged Adult", "Senior Adult"),
          include.lowest = TRUE),
         weight_condition = cut(PC$bmi, breaks = c(0, 18.4, 24.9, 29.9, Inf),
          labels = c("Underweight", "Normal Weight", "Overweight", "Obese")))
```

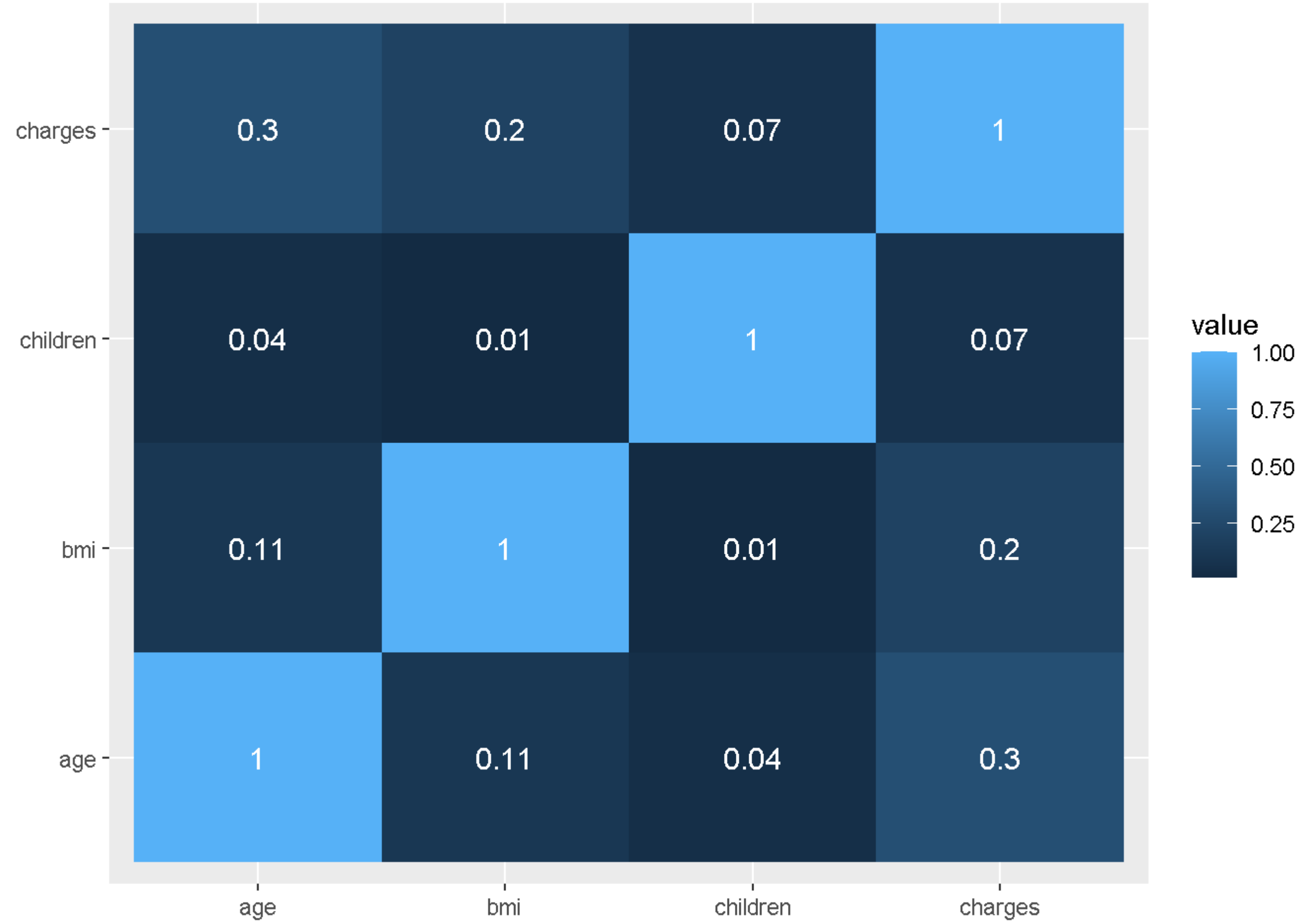
Exploratory Data Analysis

- Even number of patients for those young and middle-aged.
- Age and BMI are the most correlated with charges, numerically.

```
ggplot(PC, aes(x=age_category, fill=age_category)) + geom_bar()
```

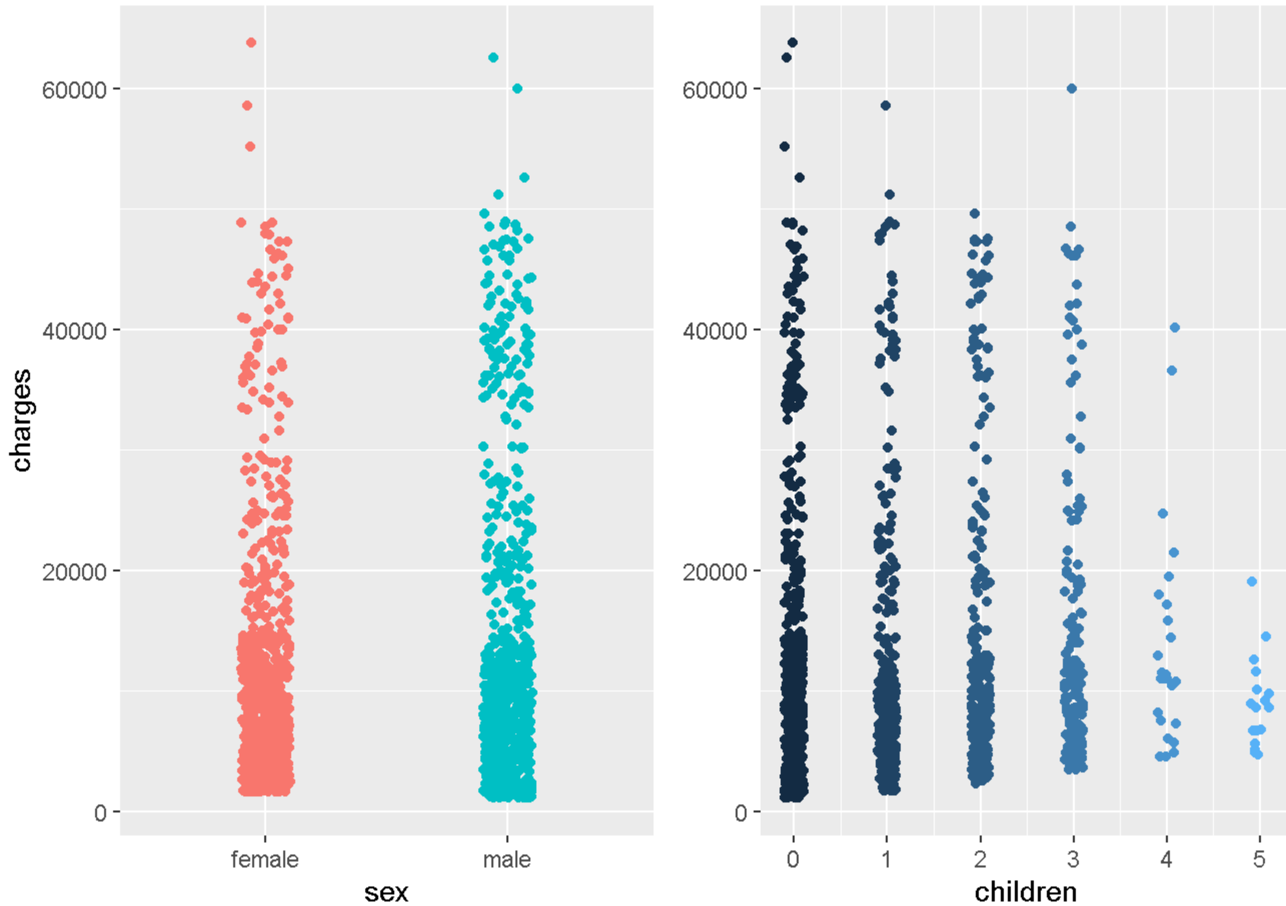



```
cormat <- PC %>% select(age, bmi, children, charges) %>% cor() %>% round(2)
melted_cormat <- reshape2::melt(cormat)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value), color = "white", size = 4) +
   theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        )
```



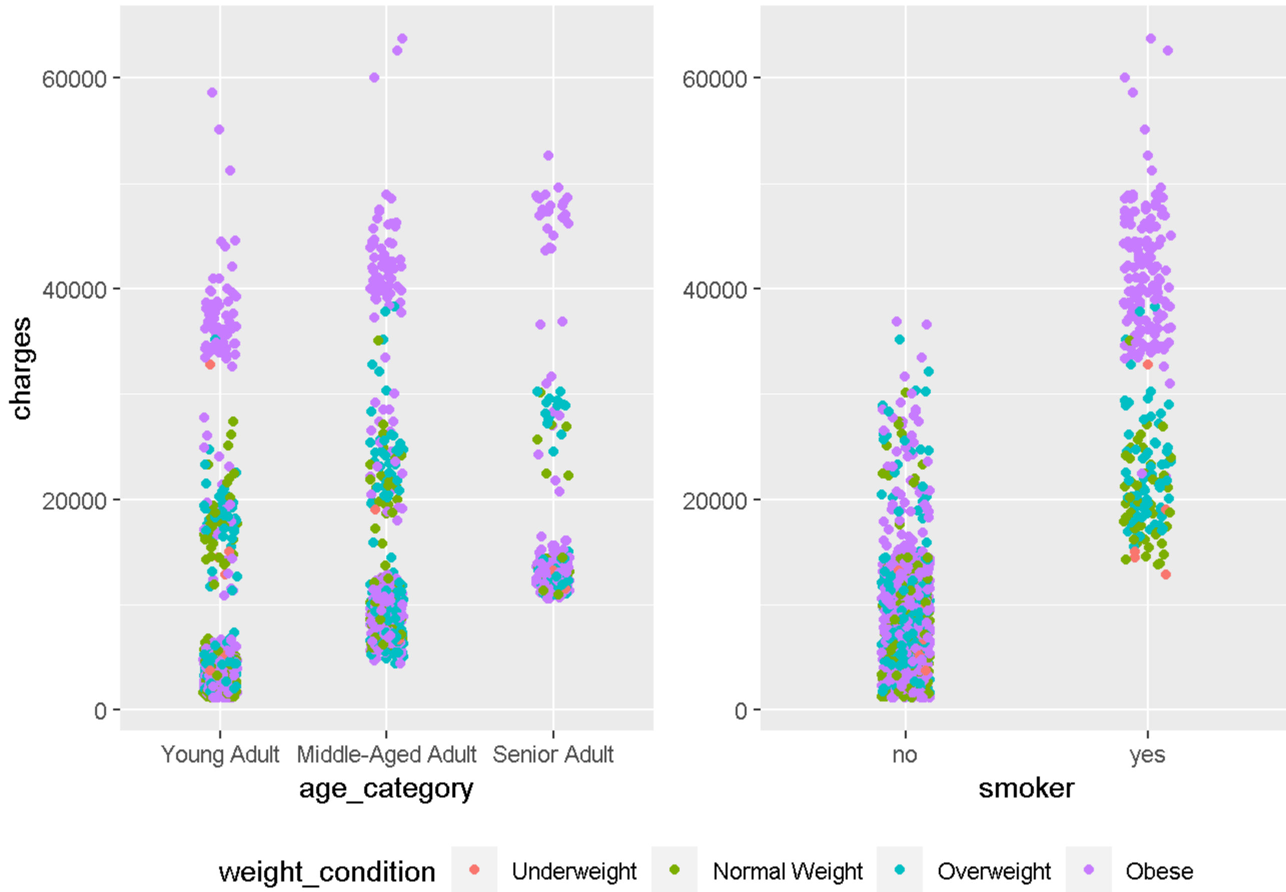
- Moving on to visualizations against charges, our target variable.
- Not noticing any clusters within the ‘sex on charges’ graph. Meaning, either being male/female doesn’t seem to have a noticeable effect.
- Children on Charges also does not have protruding data points or clusters. Will further analyze these two fields.
  - No clusters based on weight_condititio were found, either, with these fields.

```
g1 <- ggplot(PC, aes(x=sex, y=charges, color=sex)) +
  geom_jitter(width=.1) + theme(legend.position="none")

g2 <- ggplot(PC, aes(x=children, y=charges, color=children)) +
  geom_jitter(width=.1) + theme(axis.title.y=element_blank()) + theme(legend.position="none")

combined <- g1 + g2
combined
```



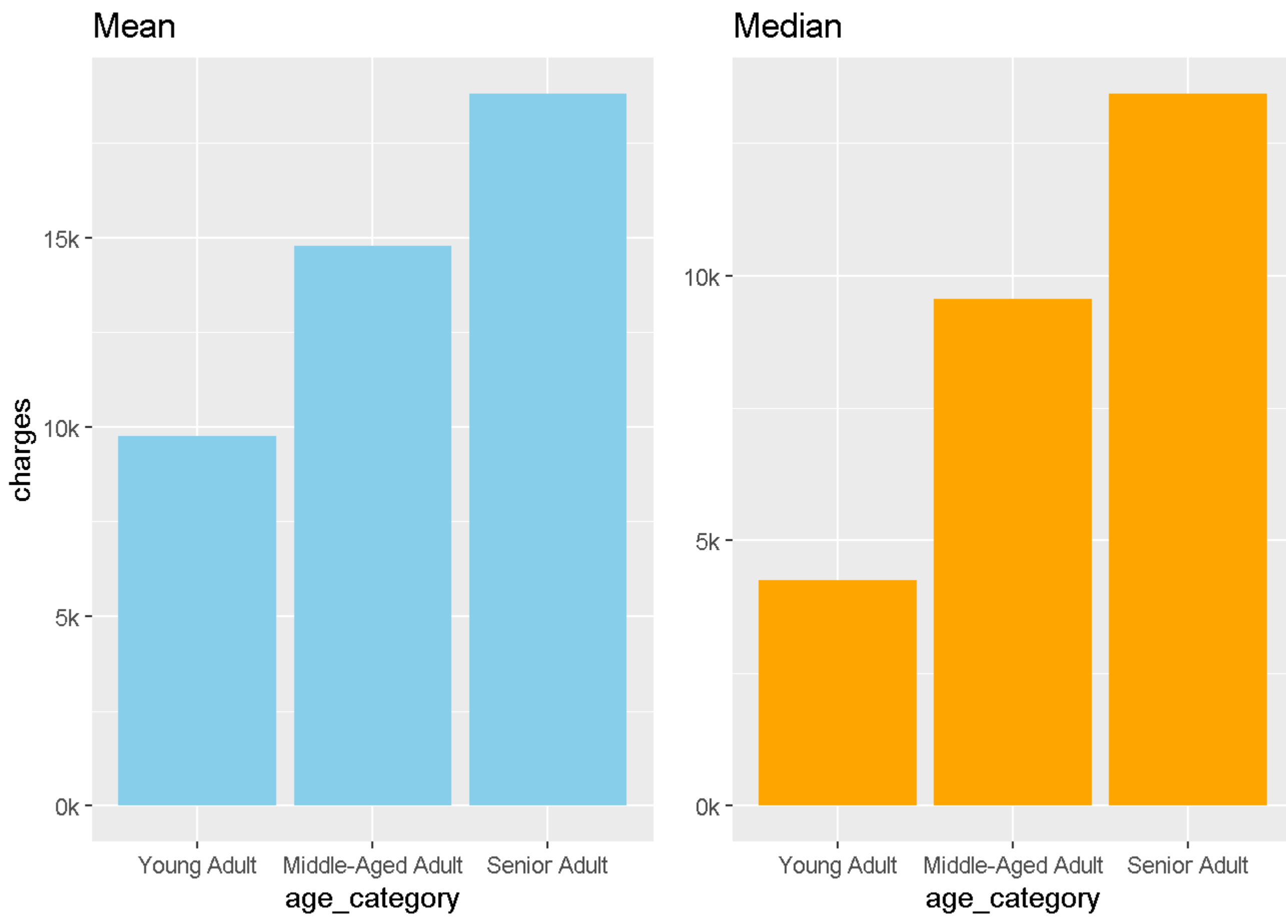
- There is a pronounced correlation between charges and smokers with the graph on the right-hand side.
- As patients rise in age, there is also an increase in charges.
- Most notably, Obese patients seem to be charged the greatest within both `age_category` & `smoker` .

```
p1 <- ggplot(PC, aes(x=age_category, y=charges, color=weight_condition)) +
  geom_jitter(width=.1)

p2 <- ggplot(PC, aes(x=smoker, y=charges, color=weight_condition)) +
  geom_jitter(width=.1) + theme(axis.title.y=element_blank())

combined <- p1 + p2 & theme(legend.position = "bottom")
combined + plot_layout(guides = "collect")
```



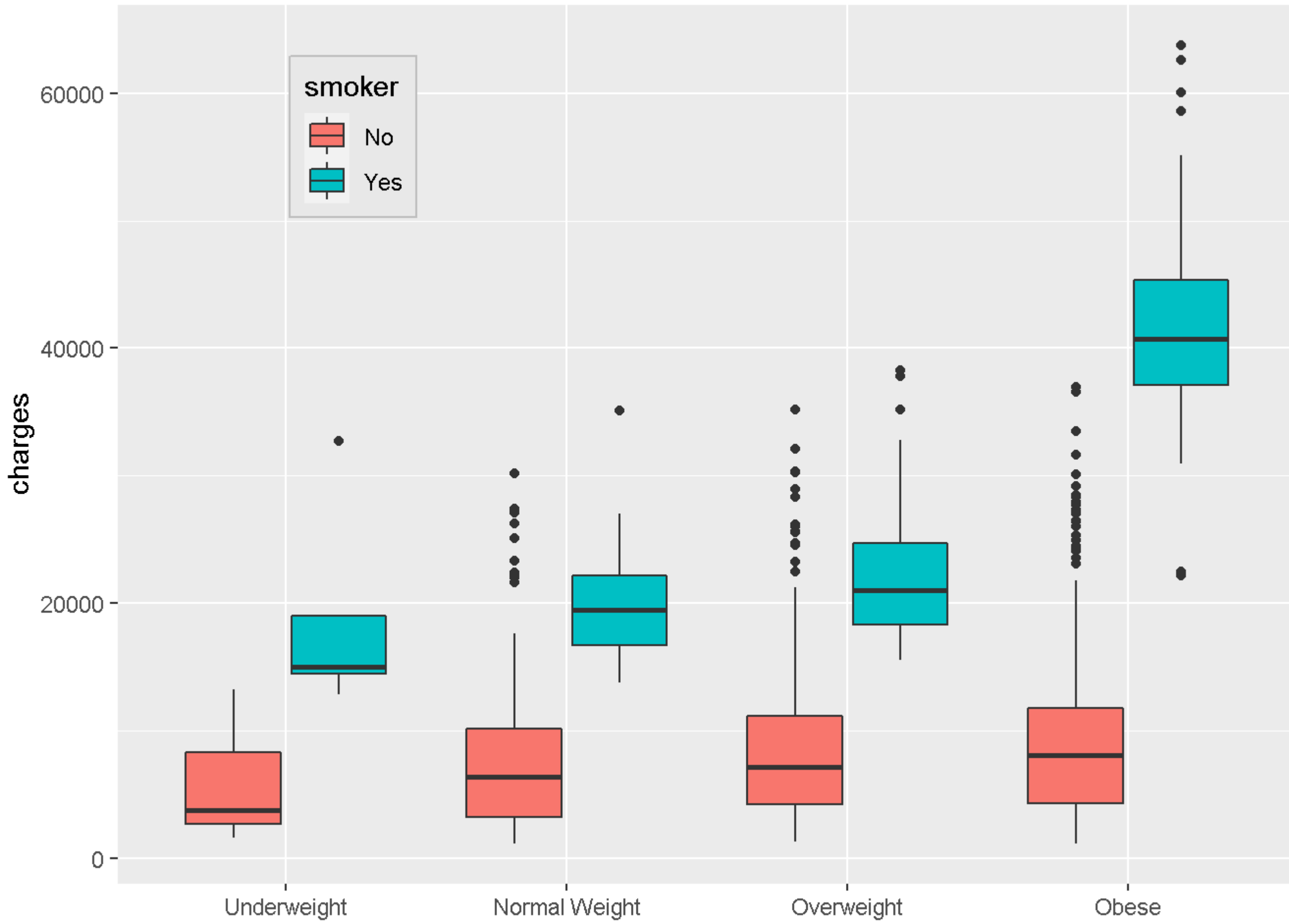
- Since there were some outliers in the last visualizations, I create bar plots to measure the mean and median for `age_category` as some charges could be being influenced too much by those outliers.

```
a1 <- ggplot(data=PC, aes(x=age_category, y=charges)) +
  geom_bar(stat="summary", fun="mean", fill="skyblue") +
  scale_y_continuous(labels = scales::label_number(suffix = "k", scale = 1e-3)) +
  ggtitle(label="Mean")

a2 <- ggplot(data=PC, aes(x=age_category, y=charges)) +
  geom_bar(stat="summary", fun="median", fill="orange") +
  scale_y_continuous(labels = scales::label_number(suffix = "k", scale = 1e-3)) +
  ggtitle(label="Median") +
  theme(axis.title.y=element_blank())


combined <- a1 + a2
combined
```



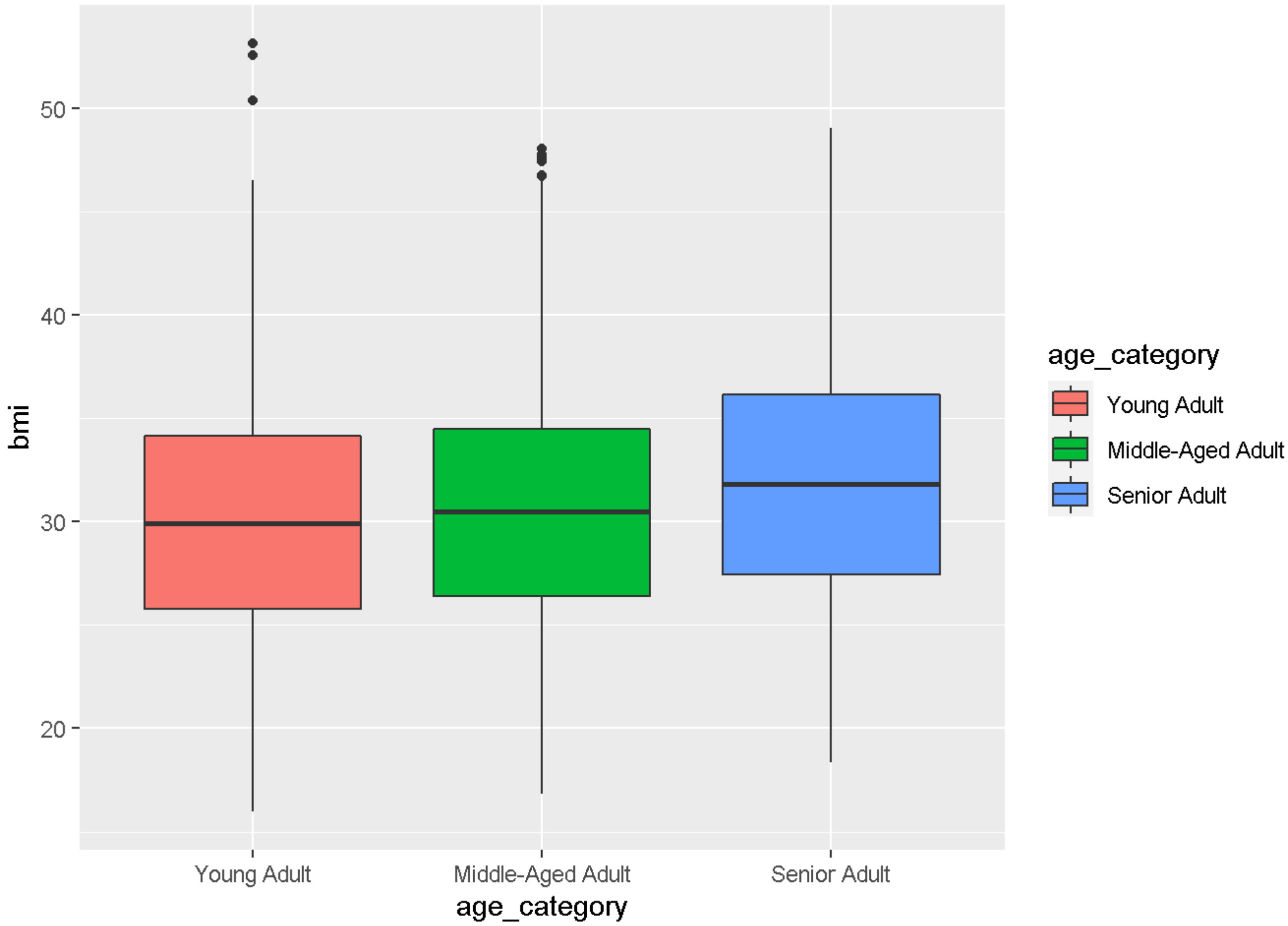
- Obese and smoking seemed to be important contributing factors to charges from earlier graphs. Let's see how charges look for obese smokers against obese non-smokers.
- It's clear that smoking is a great predictor of a higher insurance charge for all weight conditions and not just Obese.

```
PC %>%
  ggplot(aes(x=weight_condition,y=charges,fill=smoker)) +
  geom_boxplot(position="dodge2") +
  scale_fill_discrete(labels=c("No","Yes")) +
  theme(axis.title.x = element_blank(),
        legend.position = c(0.2,0.85),
        legend.background = element_rect(fill="gray91", color="gray"))
```



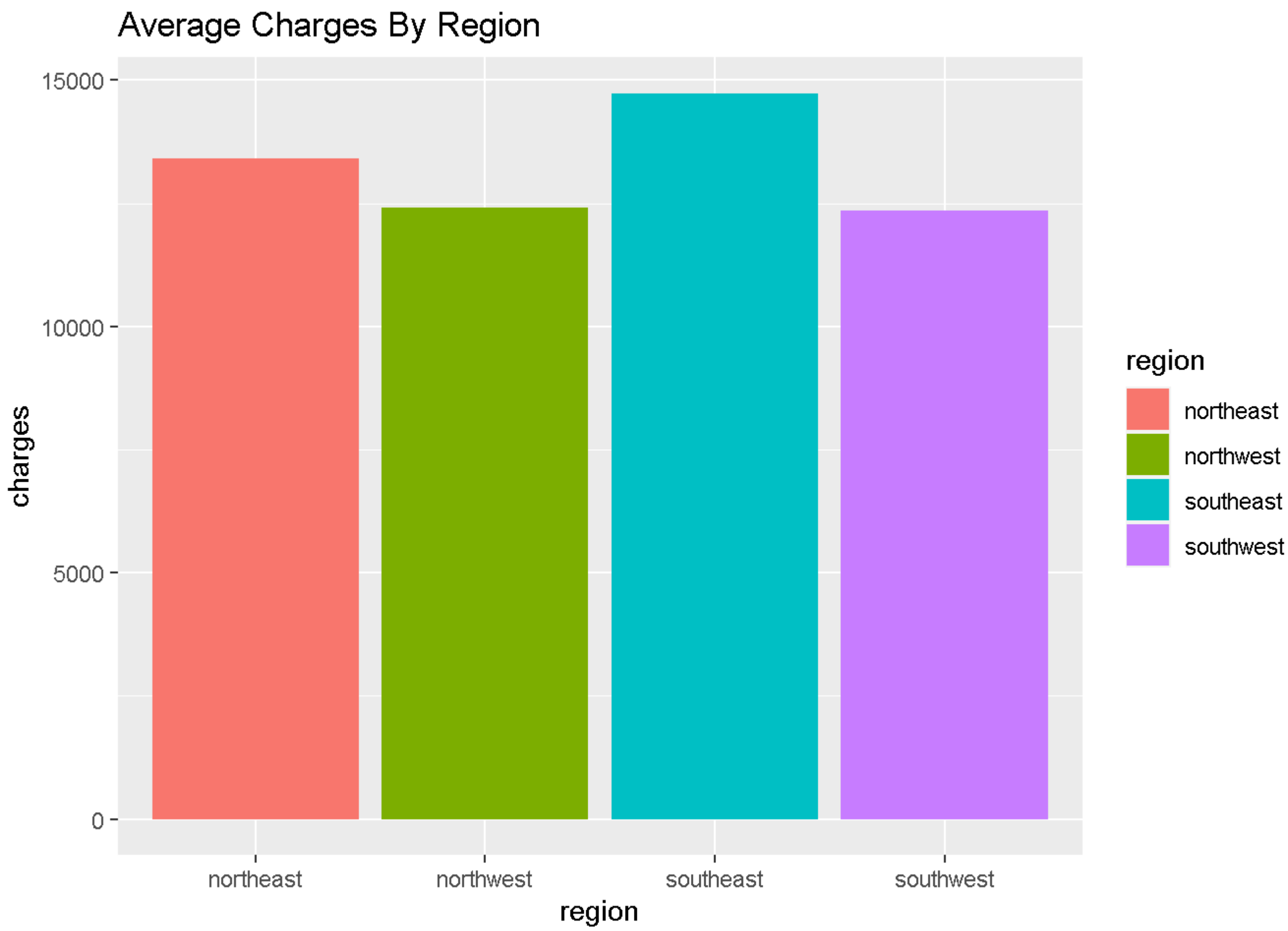
- BMI steadily increases with age, as suspected.

```
PC %>%
  ggplot(aes(x=age_category,y=bmi, fill=age_category)) +
  geom_boxplot()
```



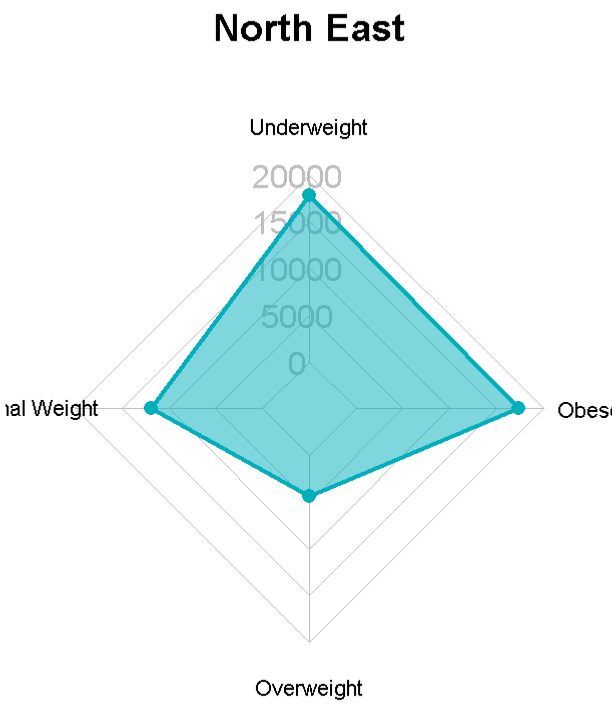
- *Incorporating Regions.:*
- Noticing that the `South East` region has highest charges, on average.
- `South West` with the lowest charges, on average.

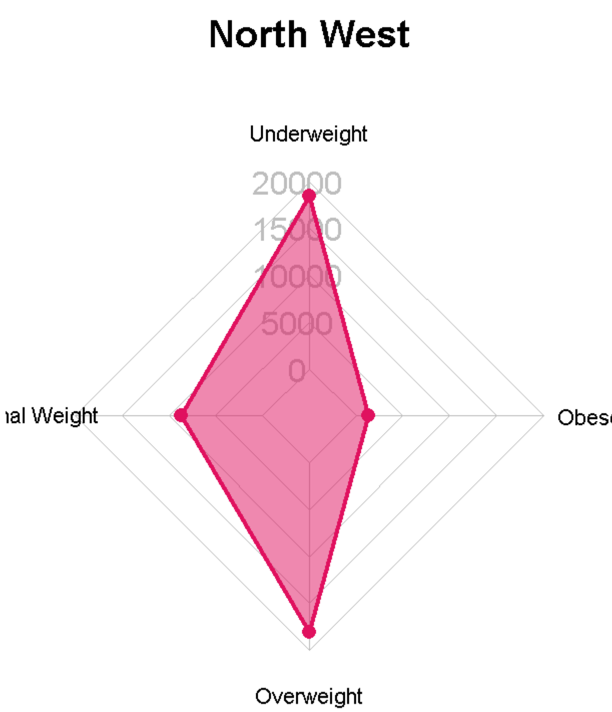
```
ggplot(PC, aes(x=region,y=charges, fill=region)) +
  geom_bar(stat="summary", fun="mean") +
  ggtitle("Average Charges By Region")
```

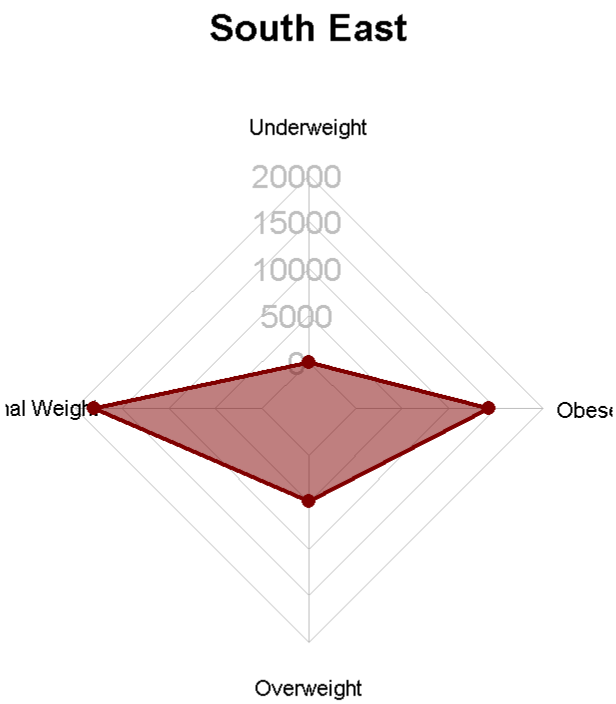


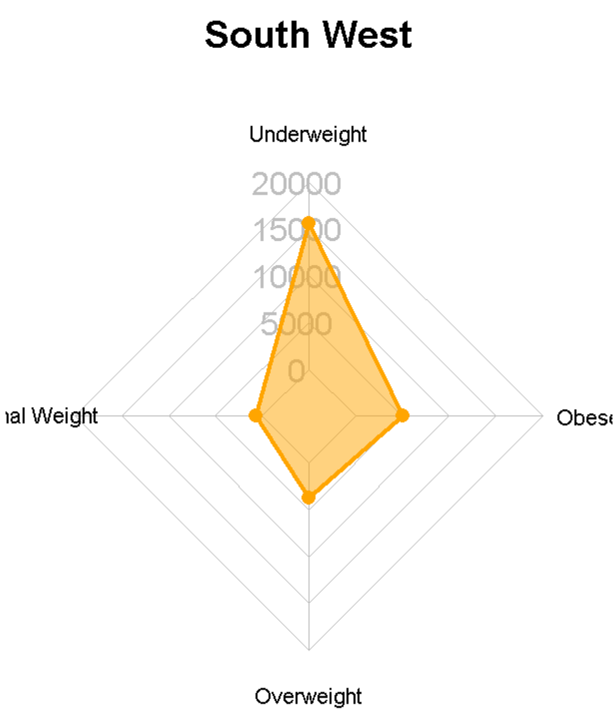
- Radar Charts picturing the average charging prices for each weight condition in each region.
    - From these charts, we can tell that the East Coast has more residents falling under the Overweight and Obese categories as opposed to the West Coast.

```
region_on_weight <- PC %>%
  select(region, charges, weight_condition) %>%
  group_by(region, weight_condition) %>%
  summarize(mean = mean(charges), .groups="drop") %>%
  pivot_wider(names_from = "weight_condition", values_from = mean) %>%
  remove_rownames %>%
  column_to_rownames(var="region") %>% # simultaneously removing non-numeric variable
  mutate(across(everything(), ~ replace_na(., 0)))


create_beautiful_radarchart <- function(data, color = "#800000",
                                        vlabels = colnames(data), vlcex = 0.7,
                                        caxislabels = NULL, title = NULL, ...){
  radarchart(
    data, axistype = 1,
    pcol = color, pfcol = scales::alpha(color, 0.5), plwd = 2, plty = 1,
    cglcol = "grey", cglty = 1, cglwd = 0.8,
    axislabcol = "grey",
    vlcex = vlcex, vlabels = vlabels,
    caxislabels = caxislabels, title = title, ...
  )
}


range <- as.data.frame(lapply(region_on_weight, function(x) rev(range(pretty(x)))))
colnames(range) <- colnames(region_on_weight)

colors <- c("#00AFBB", "#E0115F", "#800000", "orange")
titles <- c("North East", "North West", "South East", "South West")



for(i in 1:4){
  create_beautiful_radarchart(
    data = rbind(range, region_on_weight[i,]), caxislabels = c(0,5000,10000,15000,20000),
    color = colors[i], title = titles[i],
    seg=4)
}
```


North East
Underweight
20000
15000
10000
5000
0
al Weight
Obes
Overweight

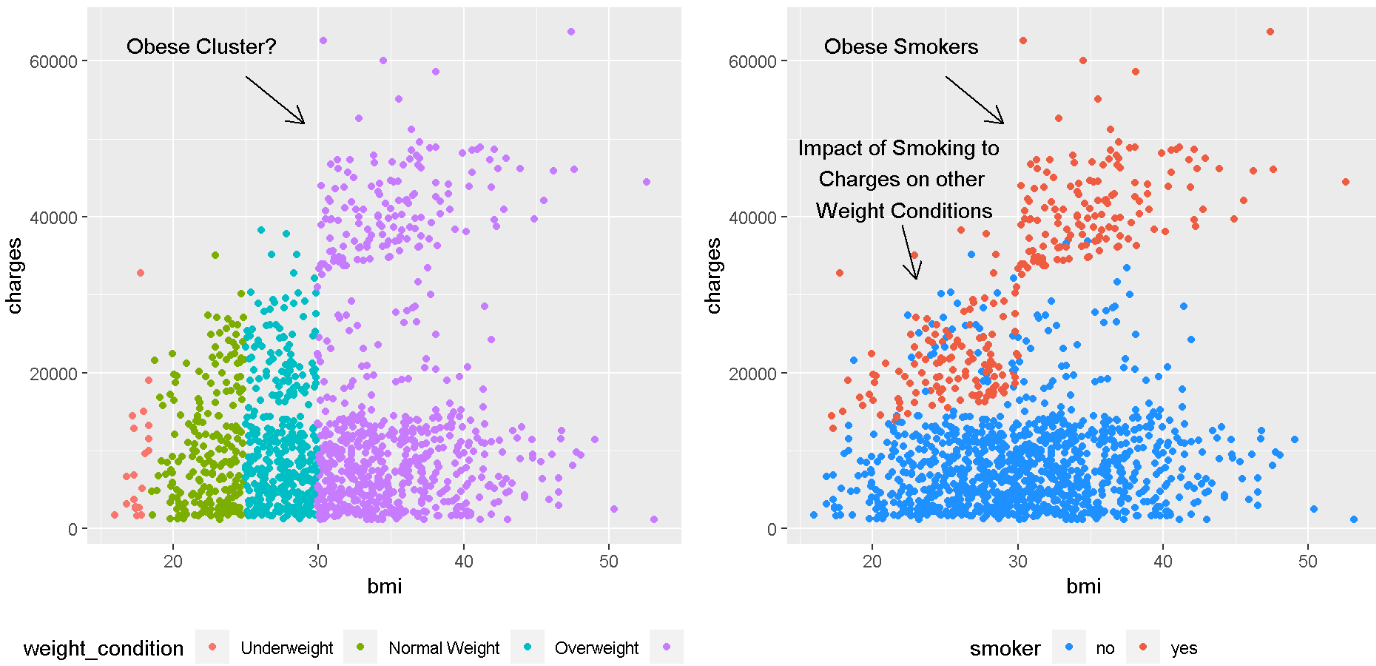

North West
Underweight
20000
15000
10000
5000
0
al Weight
Obes
Overweight





- Performing manual clustering on BMI.

```
b1 <- PC %>%
  ggplot(aes(x=bmi,y=charges,color=weight_condition)) + geom_point() +
  geom_segment(aes(x = 25,
                   y = 58000,
                   xend = 29,
                   yend = 52000),
               arrow = arrow(length = unit(0.35, "cm")),
               color = "black") +
  annotate("text", x=22, y=62000, label= "Obese Cluster?")

b2 <- PC %>%
  ggplot(aes(x=bmi,y=charges,color=smoker)) + geom_point() +
  scale_color_manual(values=c("dodgerblue", "tomato2")) +
  geom_segment(aes(x = 25,
                   y = 58000,
                   xend = 29,
                   yend = 52000),
               arrow = arrow(length = unit(0.35, "cm")),
               color = "black") +
  annotate("text", x=22, y=62000, label= "Obese Smokers") +
  geom_segment(aes(x = 22,
                   y = 39000,
                   xend = 23,
                   yend = 32000),
               arrow = arrow(length = unit(0.35, "cm")),
               color = "black") +
  annotate("text", x=22, y=45000, label= "Impact of Smoking to \n Charges on other \n Weight Condi
tions")



combined <- b1 + b2 & theme(legend.position = "bottom")
combined
```



LINEAR REGRESSION

- Constructing a linear regression using all fields and incorporating step-wise reduction using the AIC.
- Then, performing ANOVA to ensure a better model has been found.

```
# Modeling and performing stepwise reduction.
firstMod <- lm(charges ~ . , data=PC)

stepfirst <- step(firstMod, direction="both", trace=FALSE)

reducedMod <- lm(as.formula(stepfirst), data=PC)


nobs(firstMod) # Once again making sure there were no NA values by counting the number of observat
ions the model used.
```

```
## [1] 1338
```

```
nobs(reducedMod)
```

```
## [1] 1338
```

```
# With the second model's p-value greater than an alpha of 0.05, we discard the first model. The p
-value doesn't showcase any significant difference between the two models and thus, use the model
 with less variables.
anova(firstMod, reducedMod, test = "Chi")
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + sex + bmi + children + smoker + region + age_category +
##     weight_condition
## Model 2: charges ~ age + bmi + children + smoker + age_category + weight_condition
##   Res.Df        RSS Df  Sum of Sq Pr(>Chi)
## 1   1324 4.7820e+10
## 2   1328 4.8025e+10 -4 -205278282   0.2241
```

```
# Looking at the AIC values for each mode to compare them, we see that the reduced model also has
 a lower AIC by roughly 3 points further confirming a better model.
stats::extractAIC(firstMod)
```

```
## [1]    14.0 23298.2
```

```
stats::extractAIC(reducedMod)
```

```
## [1]    10.00 23295.93
```

- Our three most extreme outliers according to the final model's Residuals vs Leverage plot were all female, smokers, and had extreme weight conditions.

```
PC[c(544,413,1086),]
```

```
##      age    sex    bmi children smoker    region  charges      age_category
## 544   54 female 47.410        0    yes southeast 63770.43 Middle-Aged Adult
## 413   26 female 17.195        2    yes northeast 14455.64       Young Adult
## 1086  39 female 18.300        5    yes southwest 19023.26 Middle-Aged Adult
##      weight_condition
## 544             Obese
## 413       Underweight
## 1086      Underweight
```

Hypothetical Patient

```
Amelia <- data.frame(age = 32,
                   bmi = 24.35,
                   children = 2,
                   smoker = "no",
                   region = "southeast",
                   age_category = "Young Adult",
                   weight_condition = "Normal Weight")

print(paste0("Health care charges for Amelia: ", round(predict(reducedMod, Amelia), 2)))
```

```
## [1] "Health care charges for Amelia: 4739.24"
```

Conclusion

The optimal linear regression is portrayed by: charges ~ age + bmi + children + smoker + age_category + weight_condition