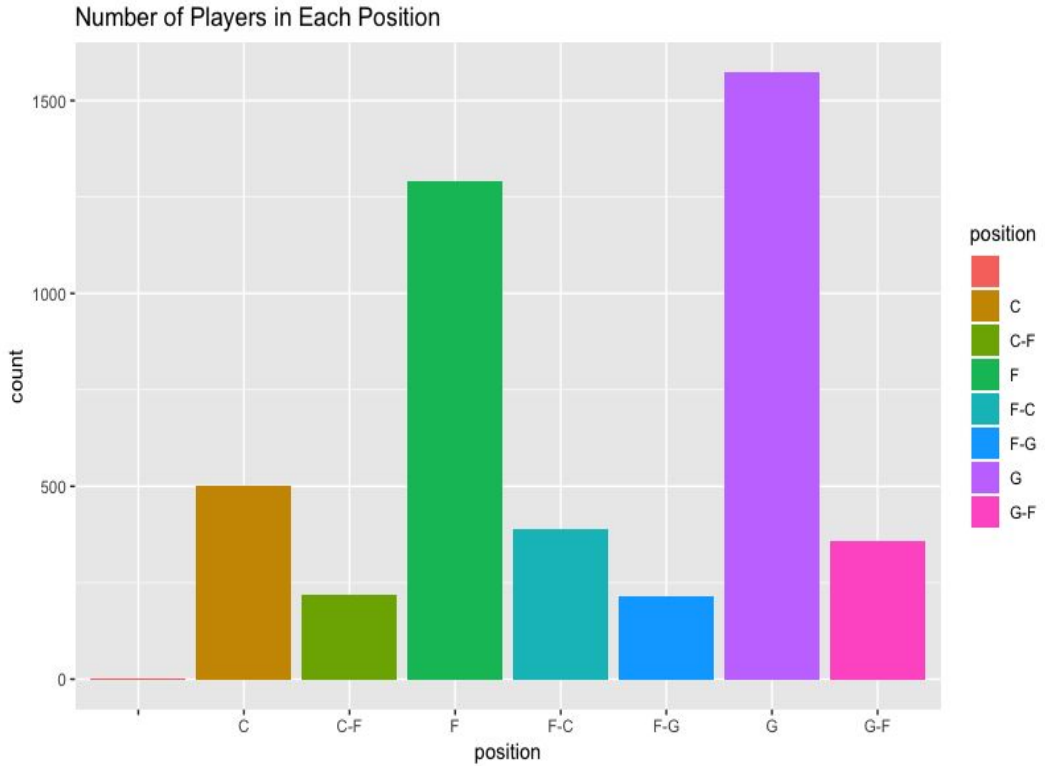
Unit 2: For Live Session

DS6306
Garrity

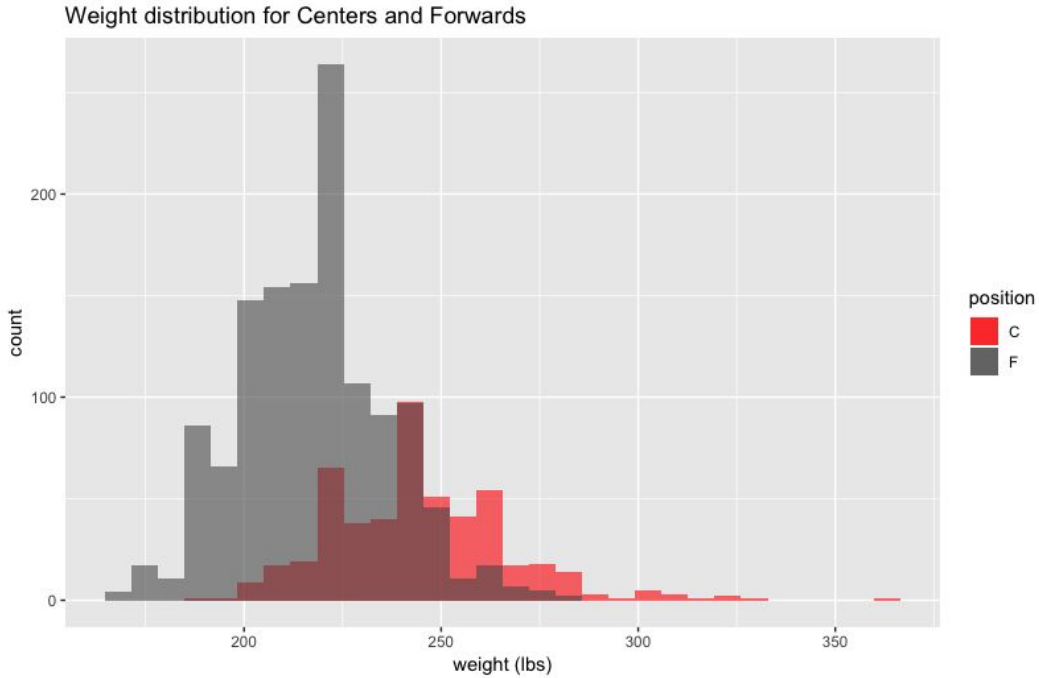
Question 1: Visually represent (summarize) the number of players in each position.




```
playerbb %>% ggplot(aes(x=position, fill=position)) + geom_bar() + ggtitle("Number of Players in Each Position")
```

Question 2: Use the dataset to visually investigate the distribution of the weight of centers (C) is greater than the distribution of the weight of forwards (F).

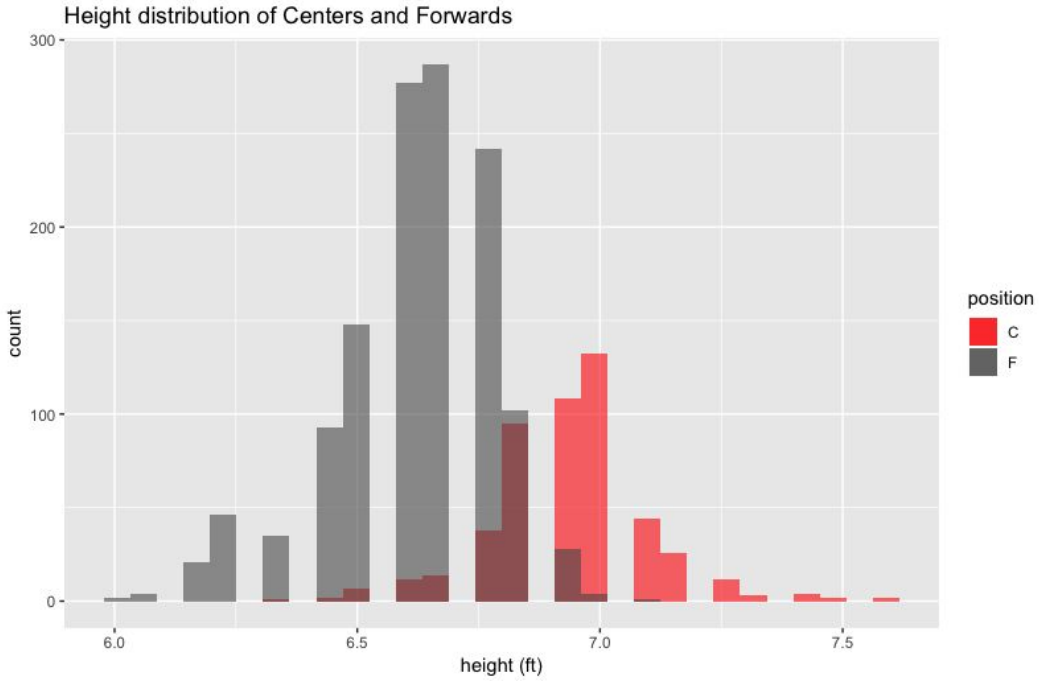
It appears that centers tend to weigh more than forwards.




```
ggplot(weights_C_F,aes(x=weight))+
   geom_histogram(data=subset(weights_C_F,position=='C'),aes(fill=position),alpha=0.6)+
   geom_histogram(data=subset(weights_C_F,position=='F'),aes(fill=position),alpha=0.6)+
   scale_fill_manual(name="position", values=c("red","gray35"),labels=c("C","F"))+
 ggtitle("Weight distribution for Centers and Forwards") +  xlab("weight (lbs)")
```

Question 3: Use the dataset to visually investigate if the distribution of the height of centers (C) is greater than the distribution of the height of forwards (F).

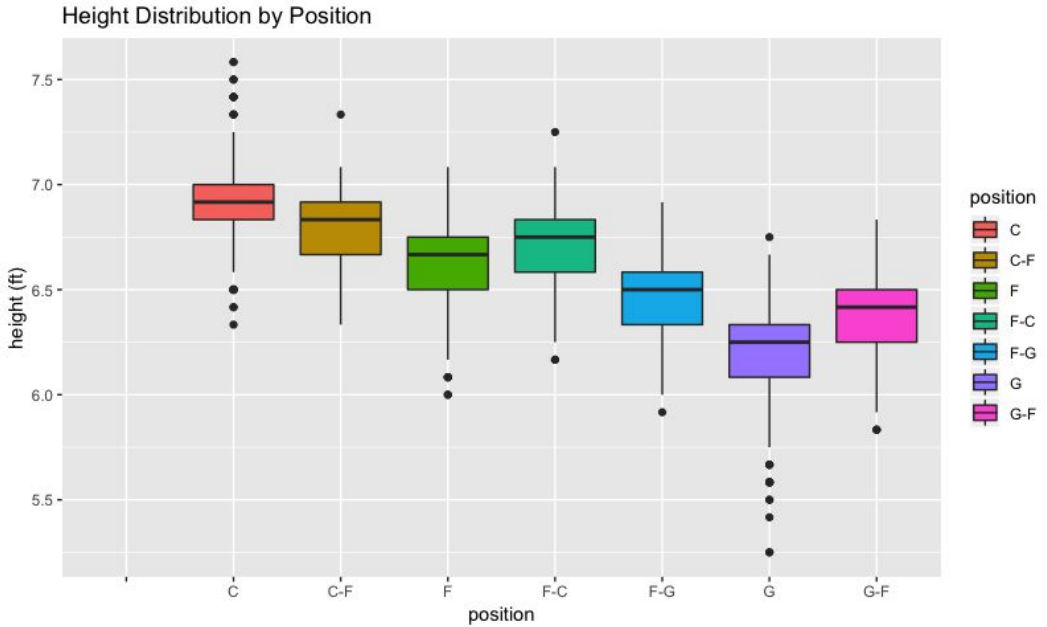
It appears that centers tend to be taller than forwards. The distribution (range?) appears to be similar. For both positions, the range of heights is approximately 2.0 feet.




```
ggplot(playerbb,aes(x=feet_height))+
   geom_histogram(data=subset(playerbb,position=='C'),aes(fill=position),alpha=0.6)+
   geom_histogram(data=subset(playerbb,position=='F'),aes(fill=position),alpha=0.6)+
   scale_fill_manual(name="position", values=c("red","gray35"),labels=c("C","F"))+
  ggtitle("Height distribution of Centers and Forwards") +  xlab("height (ft)")
```

Question 4: Visually investigate if the distribution of height is different between any of the positions.

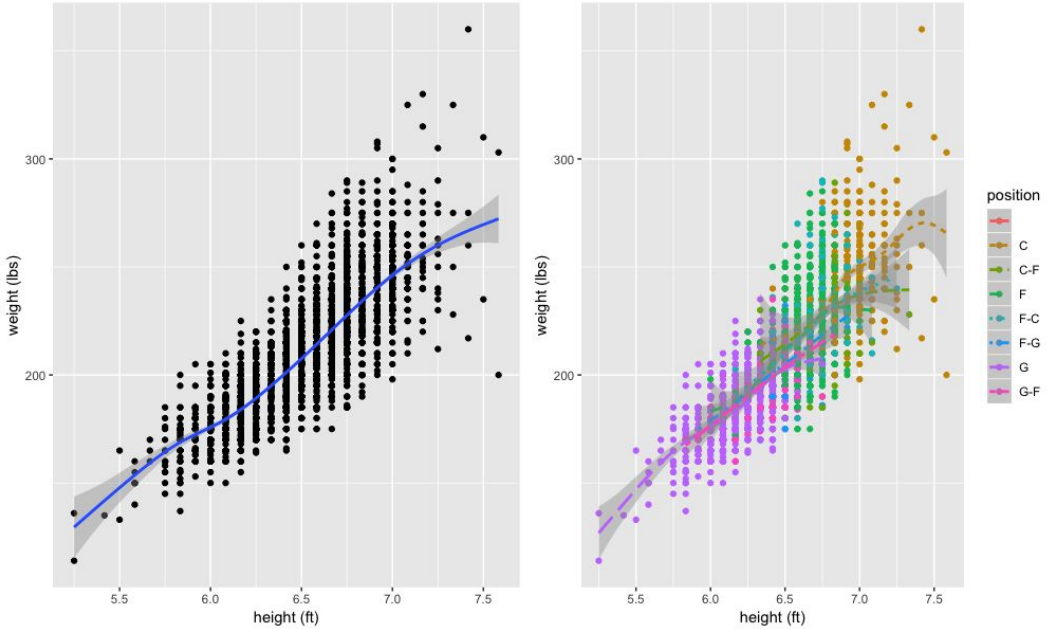
Height distribution varies by position. Centers tend to be the tallest players and guards tend to be the shortest players.




```
playerbb %>% ggplot(aes(x = position, y=feet_height, fill=position)) + geom_boxplot() +
  ylab("height (ft)") + ggtitle("Height Distribution by Position")
```

Question 5/6: Use the dataset to investigate how the player's height is related to the player's weight. How does height change as the weight changes? Are height and weight related differently for different positions?

Height and weight are positively correlated, i.e., as one increases, so does the other. The relationship appears to be fairly consistent across all positions. Weight variability appears to increase somewhat with increasing height.



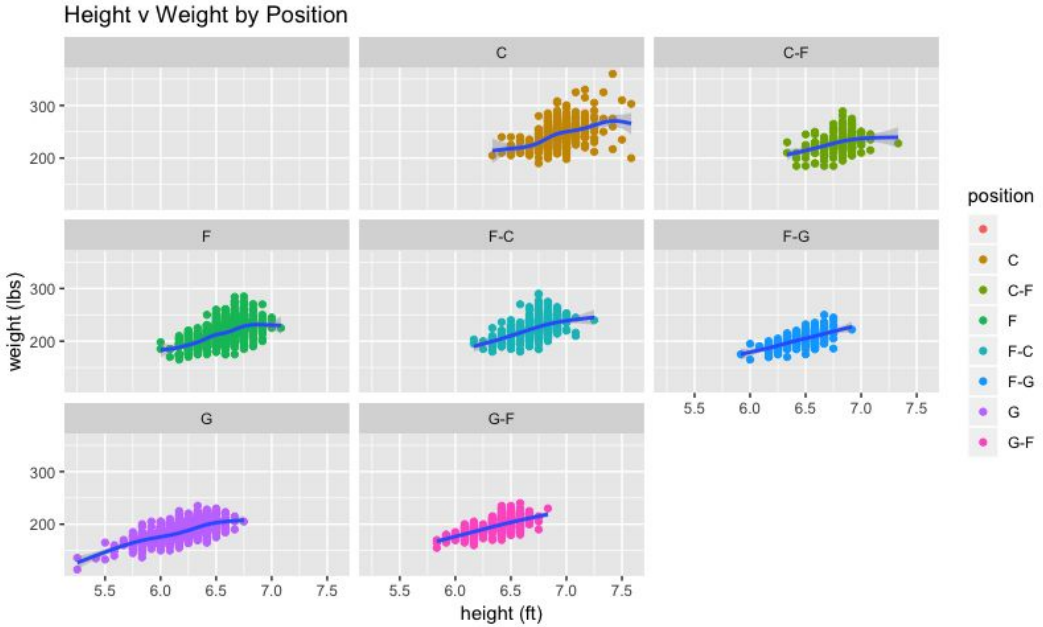

```
library(gridExtra)
p1 <- ggplot(data = playerbb, mapping = aes(x=feet_height, y=weight)) +   geom_point() +
  geom_smooth() +   xlab("height (ft)") +  ylab("weight (lbs)")

p2 <- ggplot(data = playerbb, mapping = aes(x = feet_height, y = weight, linetype = position, color = position)) + geom_point() + geom_smooth()
+  xlab("height (ft)") +  ylab("weight (lbs)")

grid.arrange(p1, p2, nrow = 1)
```

Question 6 cont.: Are height and weight related differently for different positions?

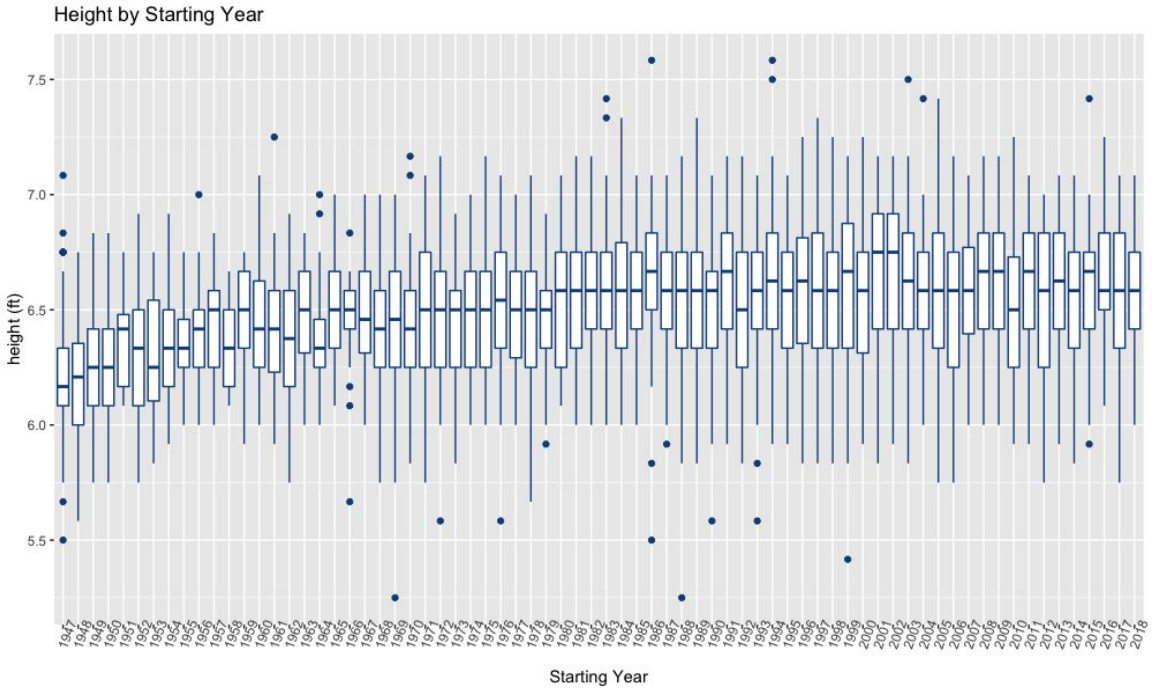
The relationship between height and weight is fairly consistent across different positions. Centers tend to approach the limits of human height, so there is a bit more scatter in the weight vs. height relationship.




```
ggplot(data = playerbb) + geom_point(mapping = aes(x = feet_height, y = weight, color = position)) +
geom_smooth(mapping = aes(x = feet_height, y = weight)) + facet_wrap(~position) +  ylab("weight (lbs)") +
xlab("height (ft)") + ggtitle("Height v Weight by Position")
```

Question 7: A historian would like to investigate the claim that the heights of players have increased over the years. Analyze this claim graphically / visually.

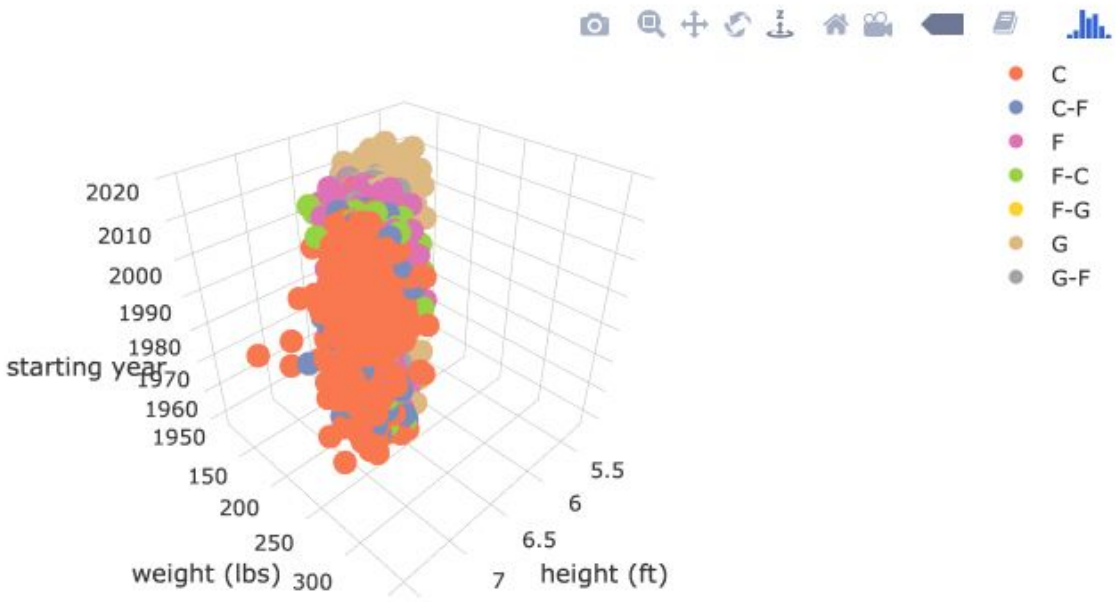
It appears that median player height increased between 1947 and the late 1960s. Median player height and distribution of heights have remained relatively stable since that time.




```
playerbb %>% ggplot(aes(x = factor(year_start), y = feet_height)) + geom_boxplot(color = "dodgerblue4") + theme(axis.text.x = element_text(angle = 70)) + ylab("height (ft)") + xlab("Starting Year") + ggtitle("Height by Starting Year")
```

Question 8: Create a 3D plot of height vs. weight vs. year and color code the points by position.

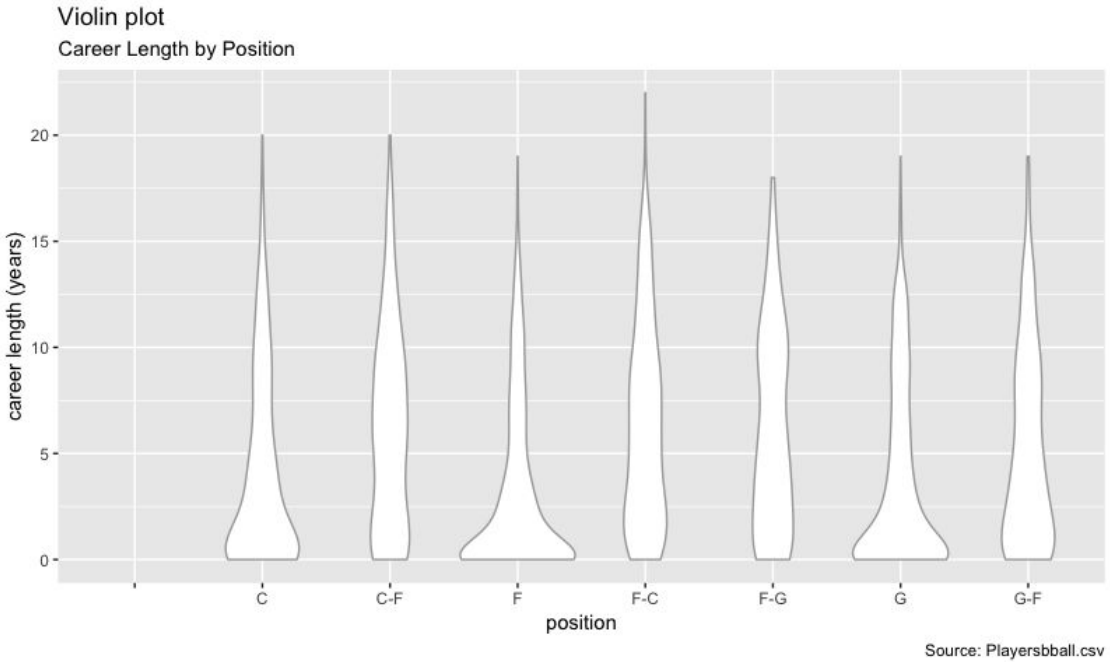
Rather ugly.




```
library(ggthemes)
library(plotly)
p <- plot_ly(playerbb, x = ~feet_height, y = ~weight, z = ~year_start, color = ~position) %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = 'height (ft)'),
                      yaxis = list(title = 'weight (lbs)'),
                      zaxis = list(title = 'starting year')))
p
```

Question 9: Go to this website and use one of the 50 best plots to visualize some aspect of the data and provide at least one insight. You will present your work in breakout! http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

I wanted to see if career length was different among the positions. It appears that career length is very right skewed for C, F, and G positions, but less skewed for C-F, F-C, F-G, and G-F positions.

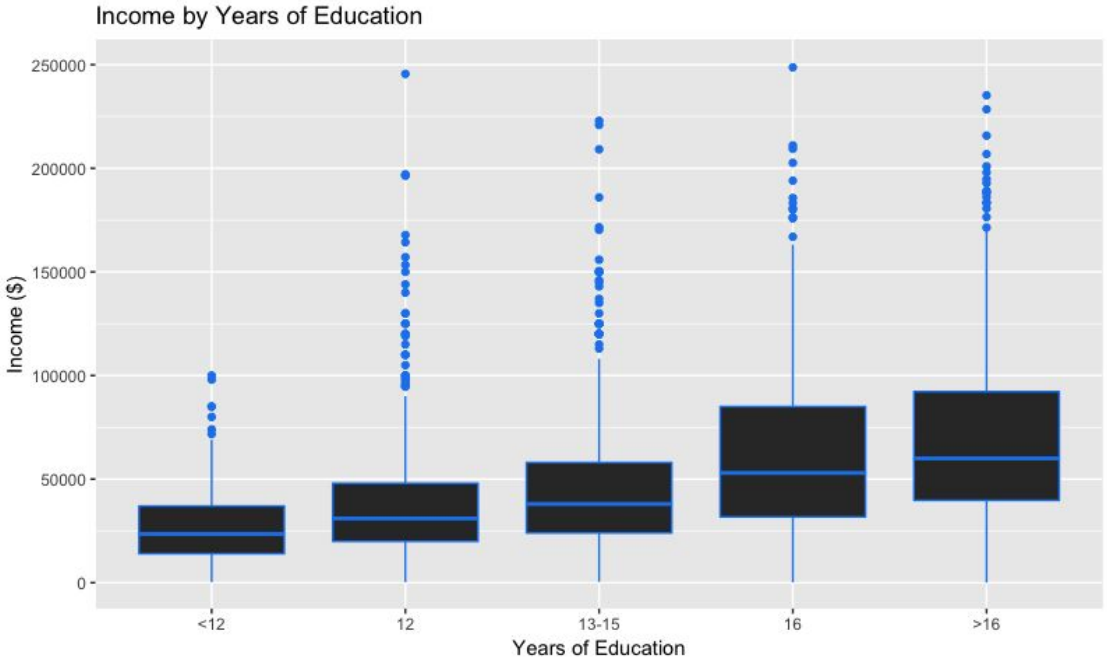



```
playerbb$careerlength = playerbb$year_end - playerbb$year_start

playerbb %>% ggplot(aes(x=position, y=careerlength)) +  geom_violin(color="darkgray") + labs(title="Violin plot",
      subtitle="Career Length by Position",  caption="Source: Playersbball.csv",  x="position", y="career length (years)")
```

Question 10: Visually test the claim that the distribution of incomes increase (mean or median) as the education level rises.

Median income appears to increase as the number of years of education increases.




```
edincome$EducR <- factor(edincome$Educ, c("<12","12","13-15","16",">16"))

edincome %>% ggplot(aes(x = EducR, y = Income2005)) + geom_boxplot(color = "dodgerblue2", fill = "gray20") +
ylab("Income ($)") + xlab("Years of Education") + ggtitle("Income by Years of Education") +  ylim(0,250000)
```

Takeaways and Questions

- I underestimated this homework. A bit more exposure to data manipulation than I was expecting. I have a better understanding of “factors” and picked up a few techniques for data filtering. I look forward to diving deeper.
- I figured out how to reorder factors manually, but would like to find a more efficient way or doing so in the event that I encounter a factor with numerous levels. Any suggestions?
- ggplot2 is really powerful. I’ve never worked with a program in Python or Matlab that made it so efficient to construct really pleasing graphics. Seems like there are a multitude of different ways of producing the same thing, though, which had me confused on more than one occasion.
- I used droplevels to remove unused factors in the weight distribution by position problem. However, it didn’t work for me when I tried something similar on the whole dataset. There appears to be an empty factor for position, but I couldn’t figure out how to remove it. Any suggestions?