Programming Exercise 4

Crime Analytics (CJUS 6106)

Instructions

For this assignment, you will be using the graphing functions reviewed in the Data Visualization Part I - Exploratory Graphics lecture and the NLSY97 data set I provided for an earlier exercise (the one you didn't need to create). To complete this assignment, you will need to use variables within this data frame (or newly created/recoded ones) to create a series of plots like those reviewed during lecture. You should submit both a .rmd file and knitted PDF by 11:59pm on the due date indicated on Canvas.

Question 1

Load the ggplot2 library and then make a copy of the nlsy97 data frame.

```
library(ggplot2)
nlsy97_copy <- nlsy97
```

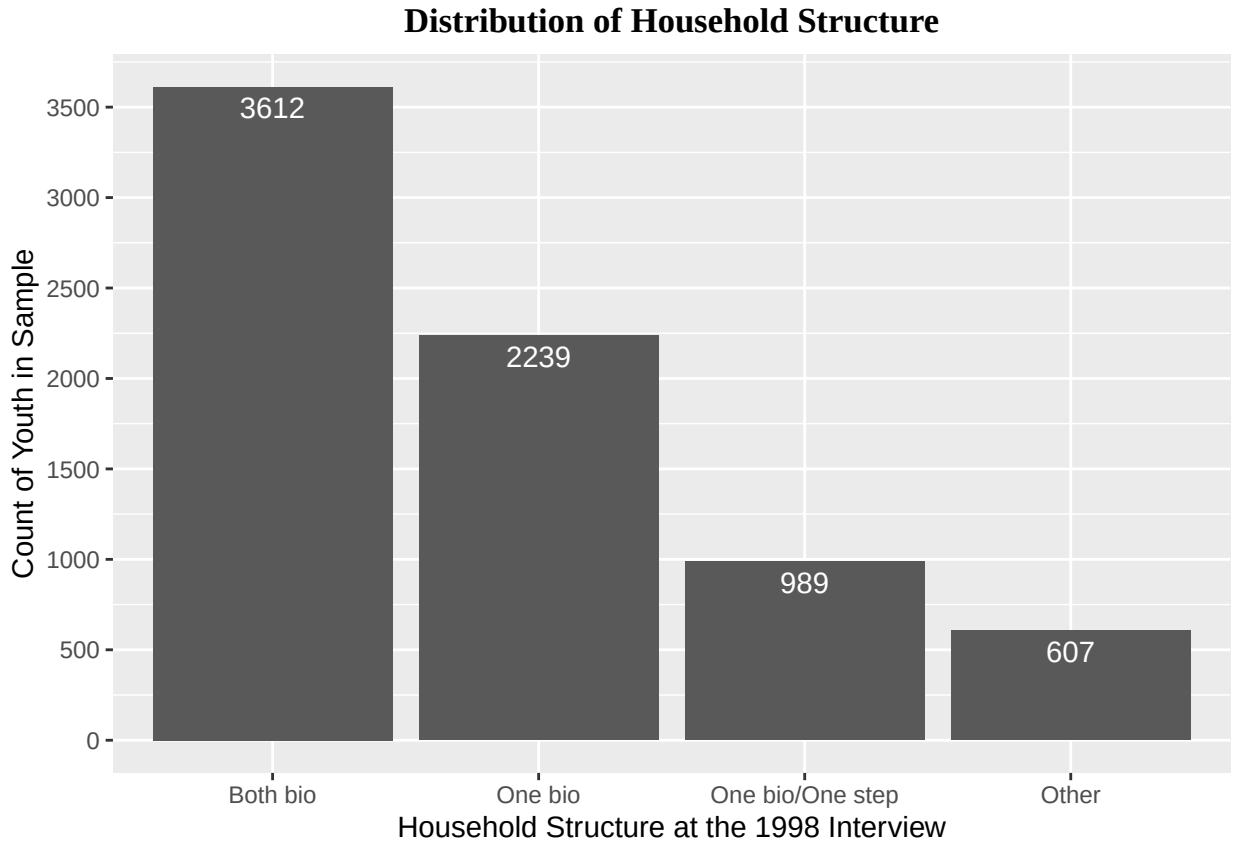
Then, make sure that all categorical variables in the copied data frame are stored as factors.

```
nlsy97_copy$hh_struc98 <- as.factor(nlsy97_copy$hh_struc98)
nlsy97_copy$race <- as.factor(nlsy97_copy$race)
nlsy97_copy$sex <- as.factor(nlsy97_copy$sex)
```

Question 2

Create a barplot that counts the number of youth who belong to each category of household structure. Be sure to provide appropriate axis and graph titles and to create bar labels for the total counts within each bar. Your final bar plot should look like the final graph for the transmission types in my lecture. **Note**: You do not need to create the graph iteratively like I do in lecture - only one code chunk is necessary to complete this part of the question.

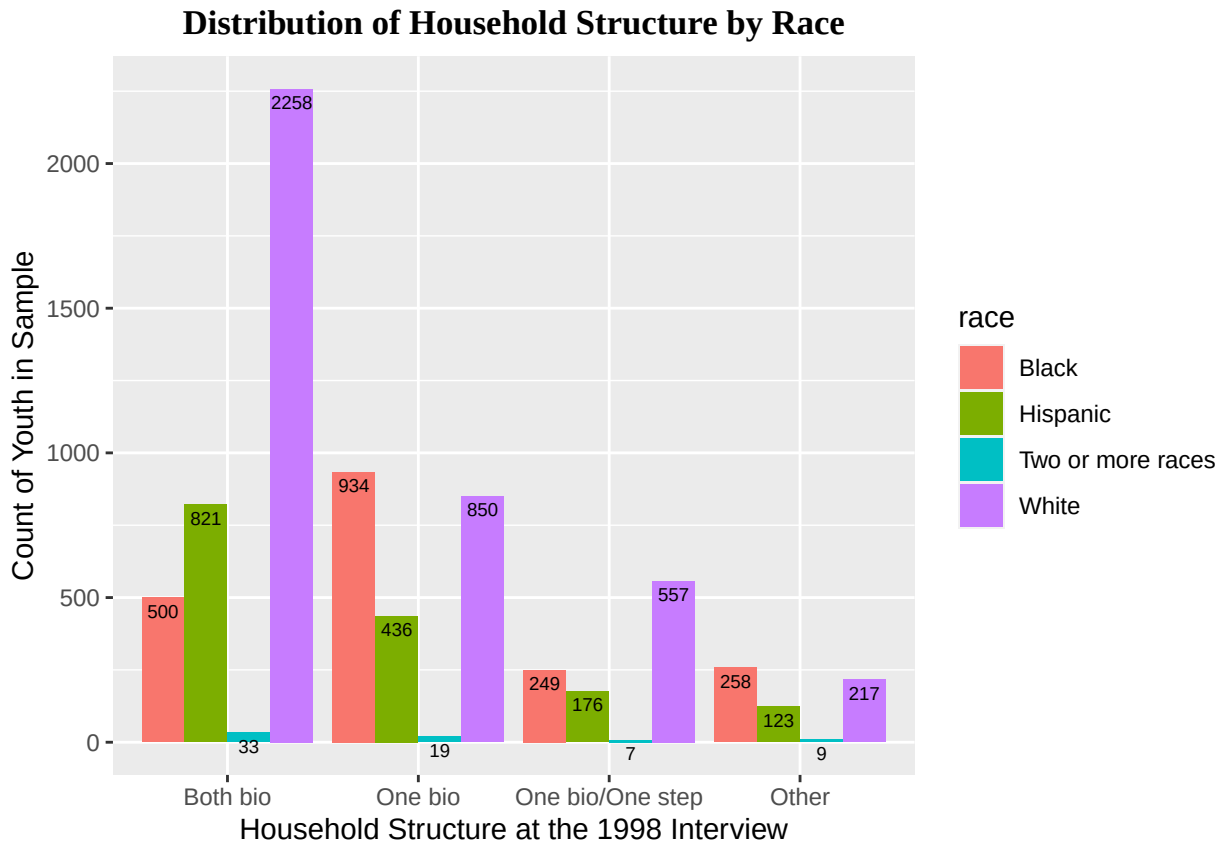
```
ggplot(nlsy97_copy, aes(x=hh_struc98)) + geom_bar(stat="count") +
  ggtitle("Distribution of Household Structure") +
  theme(plot.title=element_text(hjust=0.5,
        face="bold", family="serif")) +
  xlab("Household Structure at the 1998 Interview") +
  ylab("Count of Youth in Sample") +
  scale_y_continuous(breaks=seq(0, 4000, by=500),
                     minor_breaks=seq(0, 4000, by=250)) +
  geom_text(stat="count", aes(label=after_stat(count)),
            vjust=1.5, color="white")
```



Interpretation: By far the most common household structure reported by youth is having both biological parents in the household (n=3612). This is followed by having just one biological parent in the household (n=2239), having one biological parent and one step parent (n=989) and some other type of household structure (607), which includes living with other family members (grandparents, aunts/uncles) or living with no family members (among other structures).

Next, add the race variable to the plot like I add the number of cylinders in my example. Be sure to adjust the graph title as appropriate. **Note** - you can adjust the size of the bar labels using the size= option within the geom_text() function (I used a value of 2.5). Some bars will be very small and their labels won't have space to appear - change the color of the labels to "black" to partially account for this.

```
ggplot(nlsy97_copy, aes(x=hh_struc98, fill=race)) +
  geom_bar(position="dodge", stat="count") +
  ggtitle("Distribution of Household Structure by Race") +
  theme(plot.title=element_text(hjust=0.5,
        face="bold", family="serif")) +
  scale_y_continuous(breaks=seq(0, 2500, by=500),
                     minor_breaks=seq(0, 2500, by=250)) +
  xlab("Household Structure at the 1998 Interview") +
  ylab("Count of Youth in Sample") +
  geom_text(stat="count", aes(label=after_stat(count)),
            vjust=1.5, color="black",
            position=position_dodge(0.9),
            size=2.5)
```



Interpretation: Across race, there is some variety in the modal type of family structure. For all but Black youth, most respondents indicated living with both biological parents. By contrast, Black youth more commonly reported living with one biological parent. White youth also report living with one biological and one step parent more often than youth from other races/ethnicities. Finally, Black youth more often report living in other types of household structures as compared to other races/ethnicities.

Provide a short, 2-3 sentence interpretation for each graph.

Question 3

Create a variable within the copied data frame called crime_freq that is the sum of all the crime count variables measured at the 1999 interview (five variables, in total).

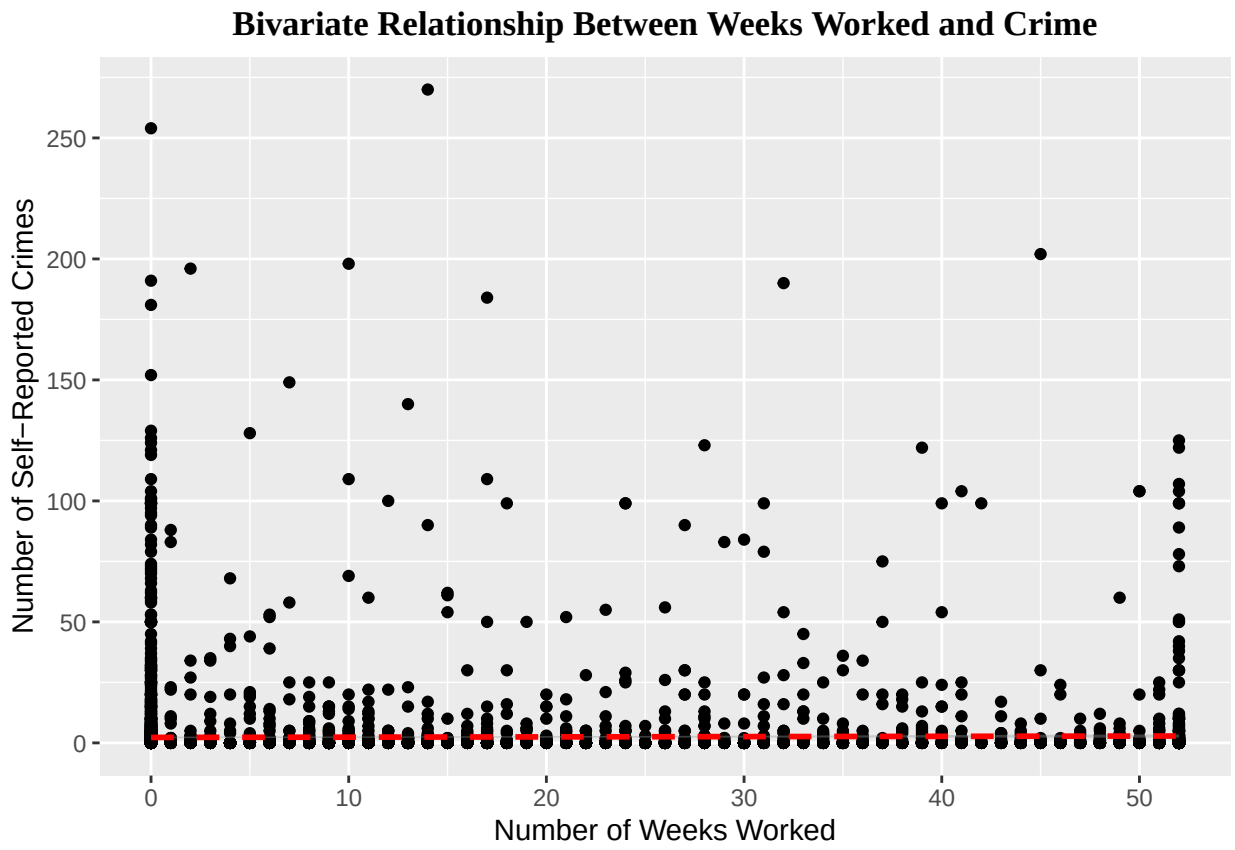
Create a scatterplot with weeks employed along the x-axis and crime_freq along the y-axis. Be sure to apply the appropriate graph and axis titles and include a best fit line (with proper formatting) using the geom_smooth() function.

```
nlsy97_copy$crime_freq <- nlsy97_copy$destprop_num99 +
                          nlsy97_copy$stllt50_num99 +
                          nlsy97_copy$stlgt50_num99 +
                          nlsy97_copy$othprop_num99 +
                          nlsy97_copy$selldrugs_num99
```

```
ggplot(nlsy97_copy, aes(x=empweeks_num98, y=crime_freq)) +
  geom_point() +
  ggtitle("Bivariate Relationship Between Weeks Worked and Crime") +
```

```
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Number of Weeks Worked") +
  ylab("Number of Self-Reported Crimes") +
  scale_y_continuous(breaks = seq(0, 300, by=50),
                     minor_breaks = seq(0, 300, by=25)) +
  geom_smooth(method="lm", color="red", linetype="dashed")
```

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



Interpretation: There does not appear to be a relationship between the number of weeks worked and the total number of self-reported crimes as the best fit line is flat near a value of 0. However, there are spikes in the number of slef-reported crimes for youth who do not report working any weeks and those who report working all 52 weeks of the last year.

Next, create a new version of this graph where the color of the points is represented by the variable sex.

```
ggplot(nlsy97_copy, aes(x=empweeks_num98, y=crime_freq, color=sex)) +
  geom_point() +
  ggtitle("Bivariate Relationship Between Weeks Worked and Crime") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Number of Weeks Worked") +
  ylab("Number of Self-Reported Crimes") +
  scale_y_continuous(breaks = seq(0, 300, by=50),
```

```
                    minor_breaks = seq(0, 300, by=25)) +
  geom_smooth(method="lm", color="red", linetype="dashed")
```

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

Interpretation: As with the prior graph, the relationship between number of weeks worked and total self-reported crimes appears to be non-existent or very weak. When the dots are coded based upon gender, we do see that substantially more male youth report higher values of total crimes as compared to female youth. Only three female youth report values at or above 100 crimes, while there are more than a dozen male youth at or above that value.

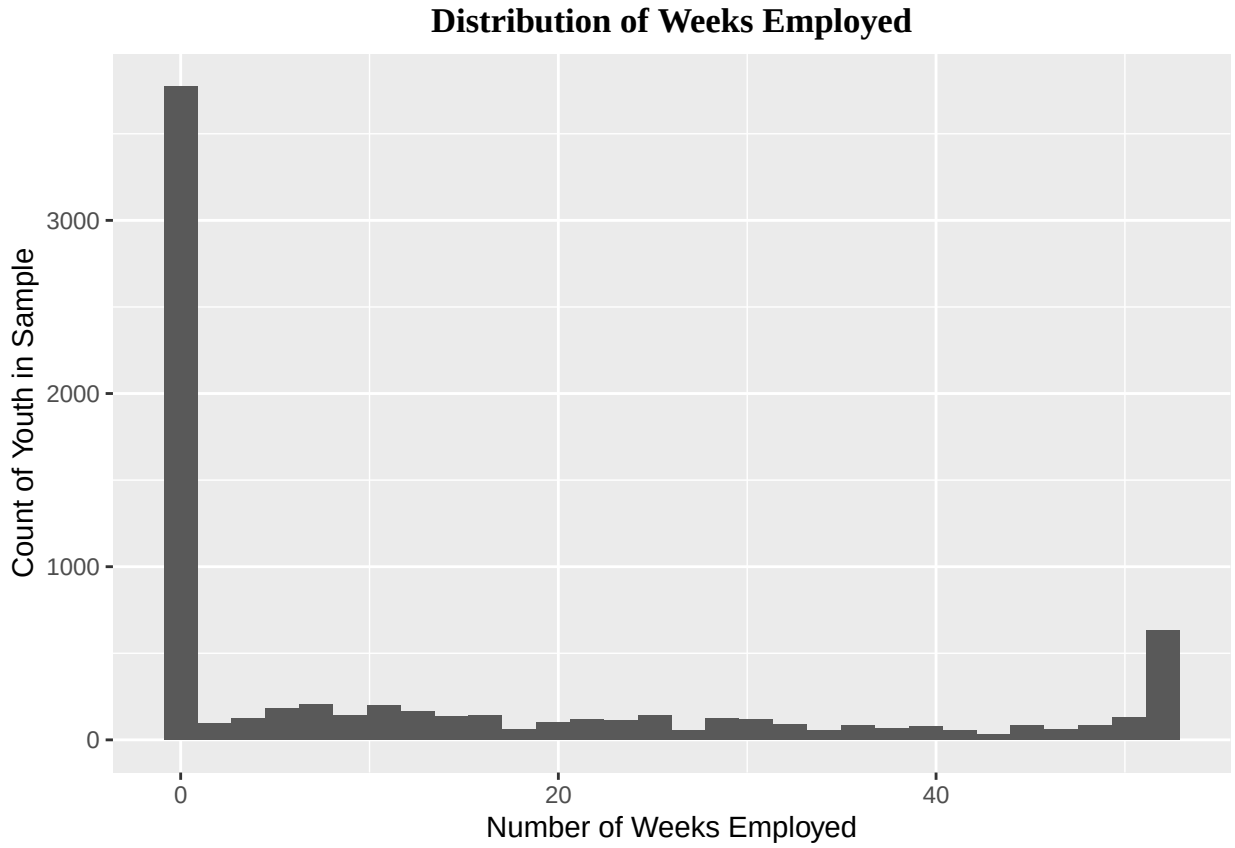
Provide a short, 2-3 sentence interpretation for each graph.

Question 4

Create a histogram that depicts the distribution of weeks employed. Be sure to provide the appropriate formatting (i.e., graph and axis titles).

```
ggplot(nlsy97_copy, aes(x=empweeks_num98)) + geom_histogram() +
   ggtitle("Distribution of Weeks Employed") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Number of Weeks Employed") +
  ylab("Count of Youth in Sample")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

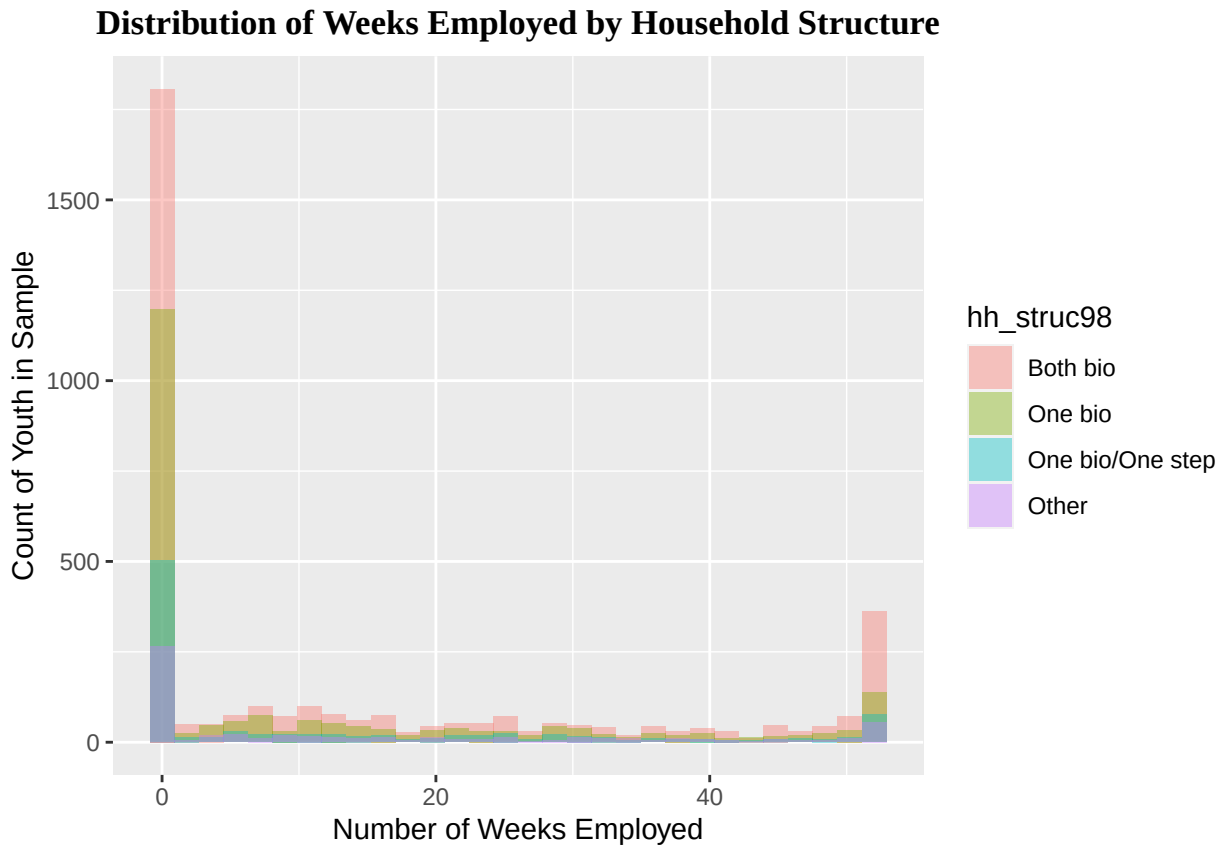


Interpretation: As we could see from the prior graph, the majority of youth do not report working at all since the date of their last interview. Values from one to fifty-one are relatively flat and then there is a spike at fifty-two weeks worked where slightly over 500 youth belong to said category.

Next, change the colors of the bars to coincide with the household structure variable. Be sure to adjust the appearance of the bars so they are all visible instead of stacked on top of one another in solid colors.

```
ggplot(nlsy97_copy, aes(x=empweeks_num98, fill=hh_struc98)) +
  geom_histogram(position="identity", alpha=0.4) +
   ggtitle("Distribution of Weeks Employed by Household Structure") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Number of Weeks Employed") +
  ylab("Count of Youth in Sample")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Interpretation: This graph shows a fairly similar distribution to the prior graph, but we now see that youth in households with both biological parents make up the majority of youth that report working zero or fifty-two weeks since the date of the last interview. These youth also make up the majority in just about every other category, as well. They are followed by youth with one biological parent in the household, one bio/one step parent, and then the other household structure category.

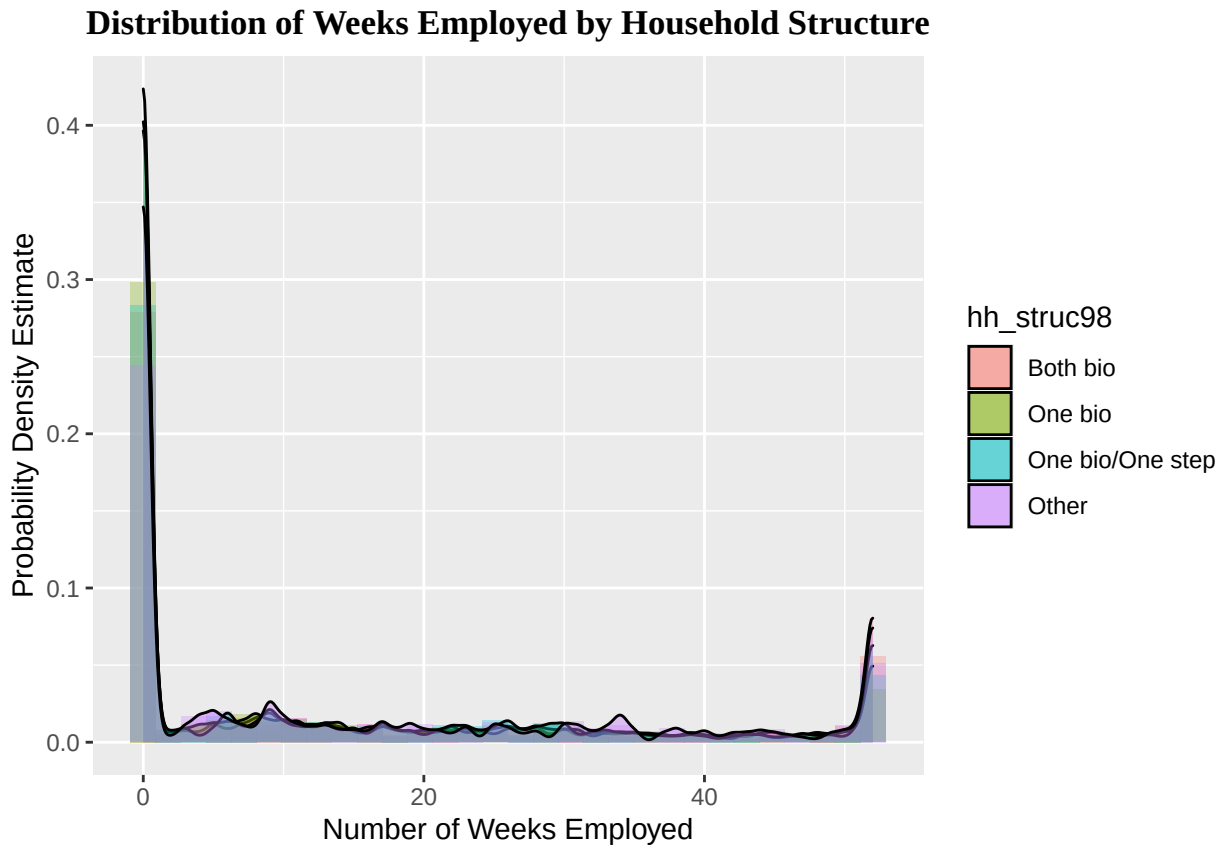
Provide a short, 2-3 sentence interpretation for each graph.

Question 5

Take the last graph you completed in Question 4 and add a density plot to it. Be sure to adjust the appearance of the plot as necessary (including the appearance of bars and titles).

```
ggplot(nlsy97_copy, aes(x=empweeks_num98, y=after_stat(density),
                        fill=hh_struc98)) +
  geom_histogram(position="identity", alpha=0.3) +
  geom_density(bw=0.5, alpha=0.4) +
   ggtitle("Distribution of Weeks Employed by Household Structure") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Number of Weeks Employed") +
  ylab("Probability Density Estimate")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



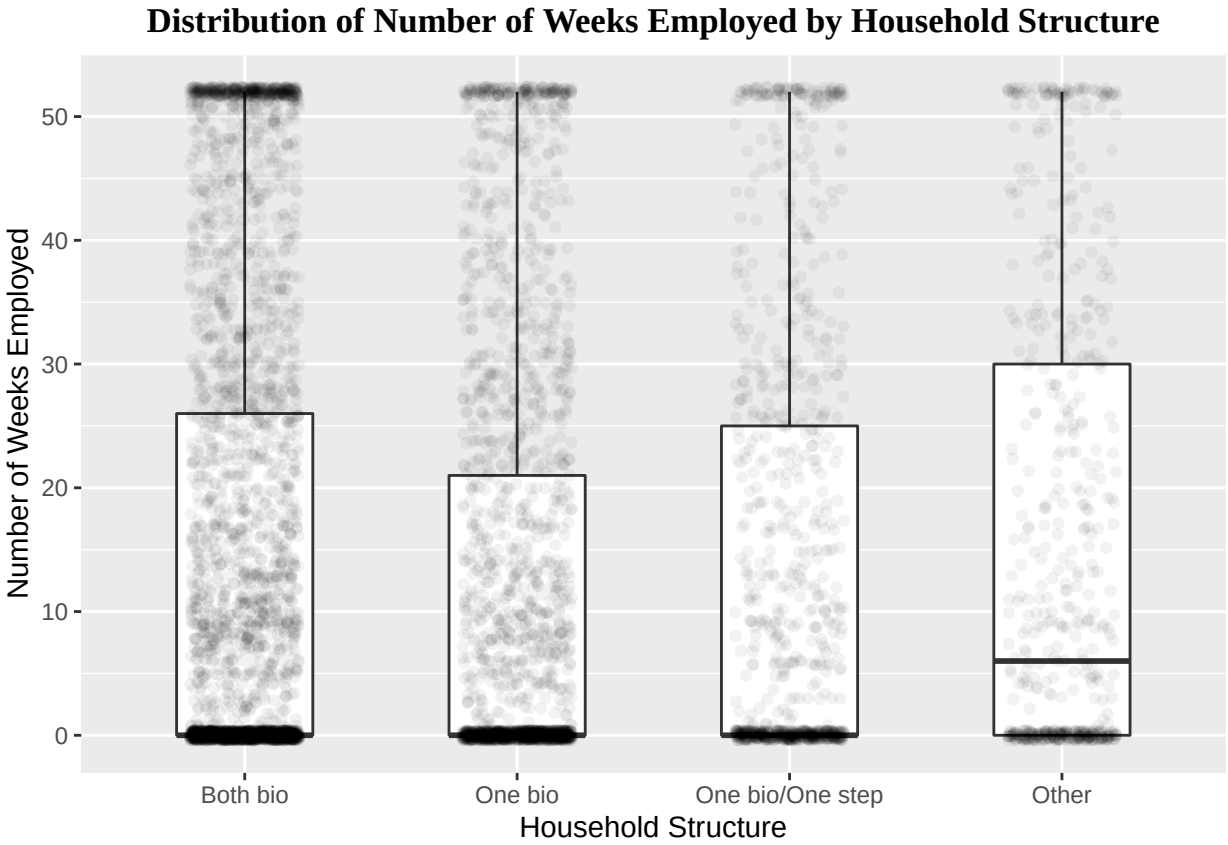
Provide a short, 2-3 sentence interpretation for this graph.

Interpretation: There does not appear to be much change from the prior plots when we add density estimates. The slight exception is that the bar at zero is now highest for youth in households with one biological parent as opposed to households with both biological parents.

Question 6

Create a boxplot where the x-axis is household structure and the y-axis is number of weeks employed. Be sure to provide the appropriate formatting options for this figure and to include the data points as dots as I do in lecture (use jitter=0.2 alpha=0.05).

```
ggplot(nlsy97_copy, aes(x=hh_struc98, y=empweeks_num98)) +
  geom_boxplot(width=0.5) +
   geom_jitter(position=position_jitter(0.2), alpha=0.05) +
  ggtitle("Distribution of Number of Weeks Employed by Household Structure") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Household Structure") +
  ylab("Number of Weeks Employed")
```

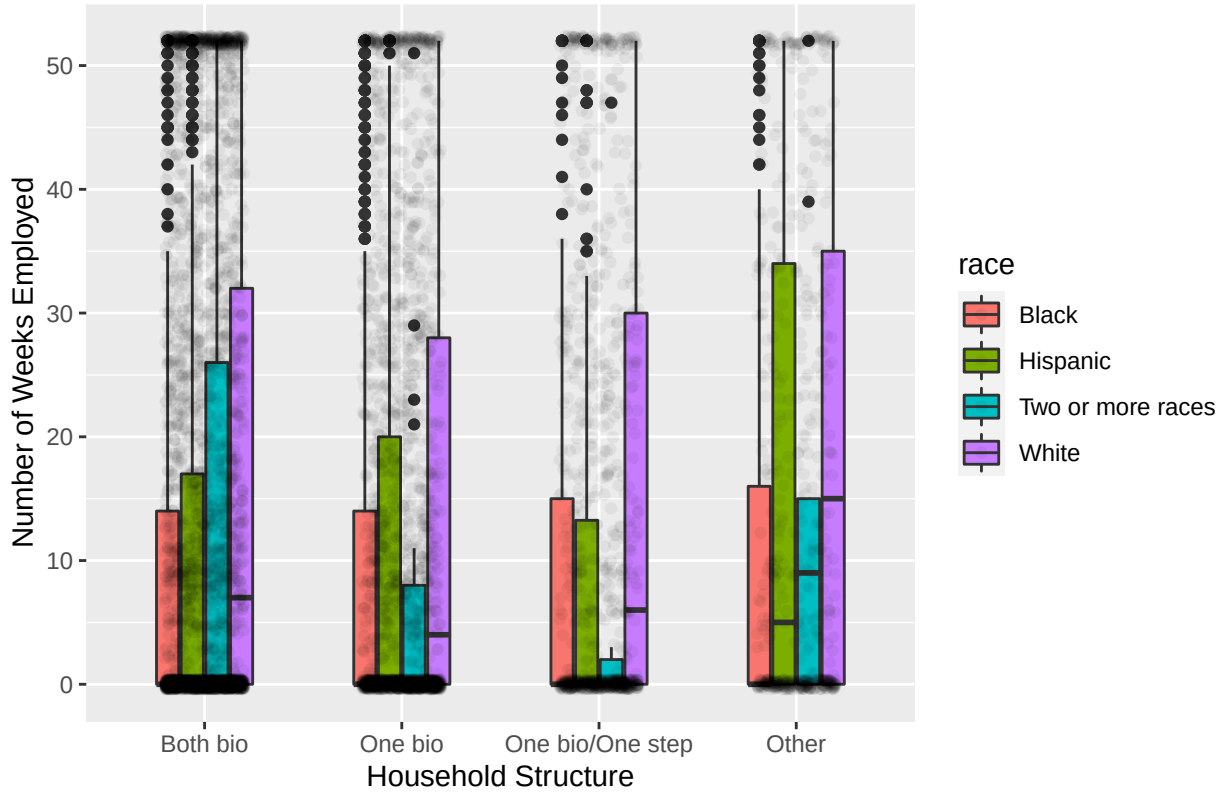


Finally, add the race variable as the fill for the boxes. This is a little different from lecture, but the process is like the other graphs (this time, it just adds a box for each household structure and race combination).

Interpretation: The distributions for weeks worked across the different household structures are fairly similar to one another, though there are slight differences in the 75th percentile values. Specifically, one biological parent households has the lowest 75th percentile value around 21 weeks worked, both bio and one bio/one step parent households have values higher than this at around 25 weeks worked, and other household structures are even higher, at 30 weeks worked. Finally, the median weeks worked for youth in other types of household structures is higher than all other categories.

```
ggplot(nlsy97_copy, aes(x=hh_struc98, y=empweeks_num98, fill=race)) +
  geom_boxplot(width=0.5) +
   geom_jitter(position=position_jitter(0.2), alpha=0.05) +
  ggtitle("Distribution of Weeks Employed by Household Structure and Race") +
  theme(plot.title=element_text(hjust=0.5,
                                face="bold", family="serif")) +
  xlab("Household Structure") +
  ylab("Number of Weeks Employed")
```

Distribution of Weeks Employed by Household Structure and Race



Provide a short, 2-3 sentence interpretation for each graph.

Interpretation: With respect to Black youth, household structure does not appear to change the distribution for their weeks worked in the last year. For Hispanic youth, the distributions for both bio, one bio, and one bio/one step are all comparable, but the distribution for other household structures implies that Hispanic youth in these households work more frequently, with some of these youth working about 30 weeks a year at the 75th percentile. Youth of two or more races have very dissimilar distributions across all categories, working the least often in one bio or one bio/one step households, working more often in other types of household structures, and working most often in households with both biological parents. Finally, White youth exhibit similar distributions for all household structures in terms of 75th percentile values, but they do vary by median weeks worked, with the median White youth in other household structures working about 15 weeks a year.

Note: Your interpretation does not need to be this long - so long as you mention some of these differences in 2-3 sentences, that is sufficient.