Intro_to_Data_Viz

SMWadgymar

2022-08-15

Let's explore a real dataset that resulted in an allegation of ethnic discrimination against the California Department of Developmental Services (DDS).

In the United States, individuals with developmental disabilities typically receive services and support from state governments. The State of California allocates funds to developmentally-disabled residents through the DDC; individuals receiving DDS funds are referred to as 'consumers'. The dataset you will work with represents a sample of 1,000 DDS consumers out of a total of ~250,000 and includes information about their age, gender, ethnicity, and the amount of financial support per consumer provided by the DDS.

A team of researchers examined the mean annual expenditure on consumers by ethnicity and found that the mean annual expenditures on Hispanic consumers was approximately one-third of the mean expenditures on White non-Hispanic consumers. As a result, an allegation of ethnic discrimination was brought against the California DDS.

This lab provides a walkthrough to conducting an exploratory analysis that not only investigates the relationship between two variables of interest, but also considers whether other variables might be influencing that relationship.

1) Load the data from a package associated with our textbook. The library function below will load the dataset from a package that we have already installed and the data function loads the data into your workspace. You should see dds.discr show up in the environment tab on the top left panel.

```
library(oibiostat)
data("dds.discr")
```

2) Let's explore this dataset using the summary function.

```
summary(dds.discr)
```

```
##        id        age.cohort       age          gender     expenditures
##  Min.   :10210   0-5  : 82   Min.   : 0.0   Female:503   Min.   :  222
##  1st Qu.:31809   6-12 :175   1st Qu.:12.0   Male  :497   1st Qu.: 2899
##  Median :55384   13-17:212   Median :18.0                Median : 7026
##  Mean   :54663   18-21:199   Mean   :22.8                Mean   :18066
##  3rd Qu.:76135   22-50:226   3rd Qu.:26.0                3rd Qu.:37713
##  Max.   :99898   51+  :106   Max.   :95.0                Max.   :75098
##
##               ethnicity
##  White not Hispanic:401
##  Hispanic          :376
##  Asian             :129
##  Black             : 59
```

```
##  Multi Race        : 26
##  American Indian   :  4
##  (Other)           :  5
```

Q: What does the summary function do? What are the names and types of variables included in this dataset?

A: There are six columns. id: numeric, unique ID numbers for each consumer age.cohort: categorical ordinal, the age cohort of each individual consumer age: numerical: the age, in years, of each individual consumer gender: categorical nominal, actually reflects the sex of each consumer expenditures: numerical, the dollar amount in funding used for each consumer's care ethnicity: categorical nominal, the ethnicity of each individual consumer

3) Below is the code for some of the most common graphs and summaries people make. To make this as useful a reference as possible, replace the ??? with either 'continuous' or 'categorical' as appropriate.

To make a histogram: hist(continuous)

To make a boxplot: boxplot(continuous)

To make a set of boxplots across a categorical variable (Y~X): boxplot(continuous~categorical)

To make a barplot of a continuous variable: barplot(continuous)

To make a set of barplots across a categorical variable (Y~X): boxplot(continuous~categorical)

To make a scatterplot (X, Y) or (Y~X): plot(continuous, continuous)

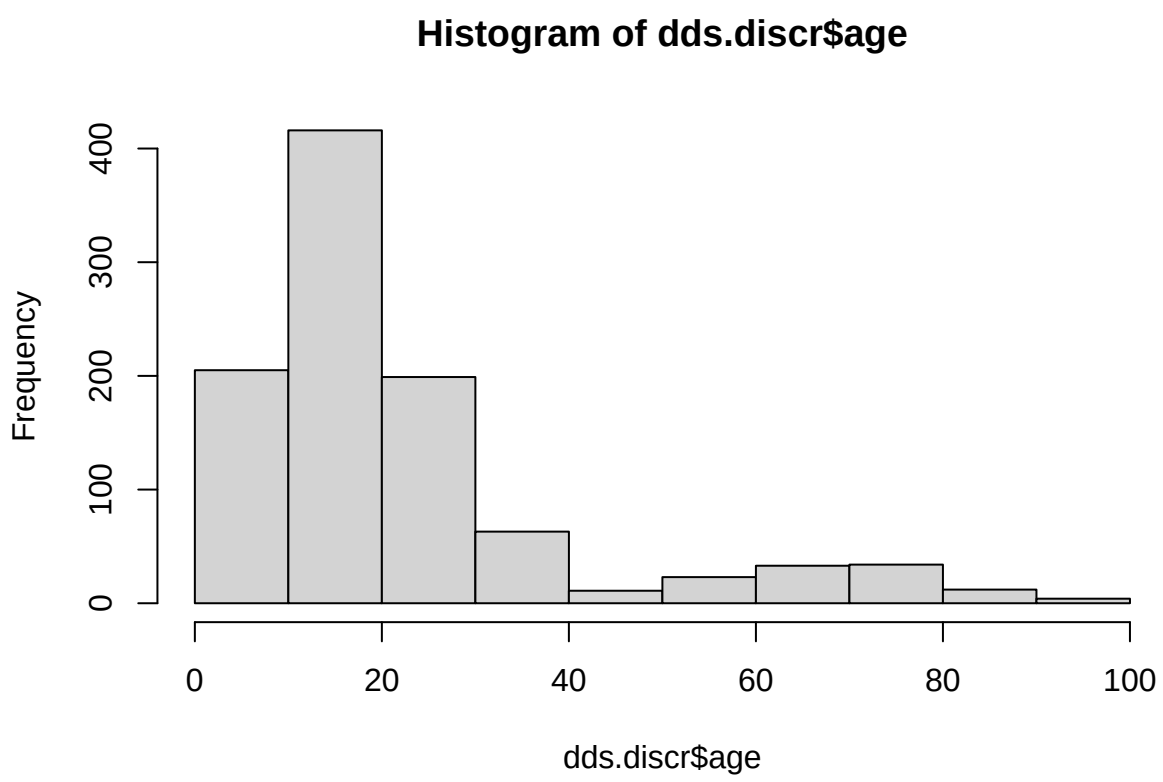
To make a barplot exhibiting counts across categories: plot(categorical)

To make a table of counts: table(categorical)

To make a table reported in proportions: prop.table(table(categorical))

4) Make an appropriate graph and/or table for age in the first code chunk and another for age.cohort in the second code chunk and describe their distributions. Do consumers tend to be older or younger?

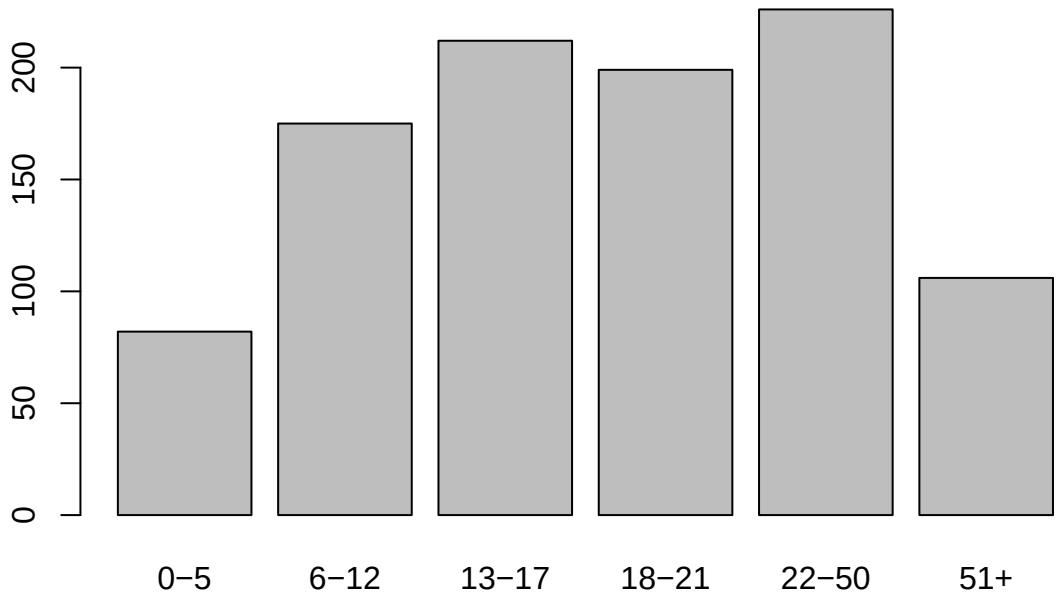
```
hist(dds.discr$age)
```




```
summary(dds.discr$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    12.0    18.0    22.8    26.0    95.0
```

```
plot(dds.discr$age.cohort)
```



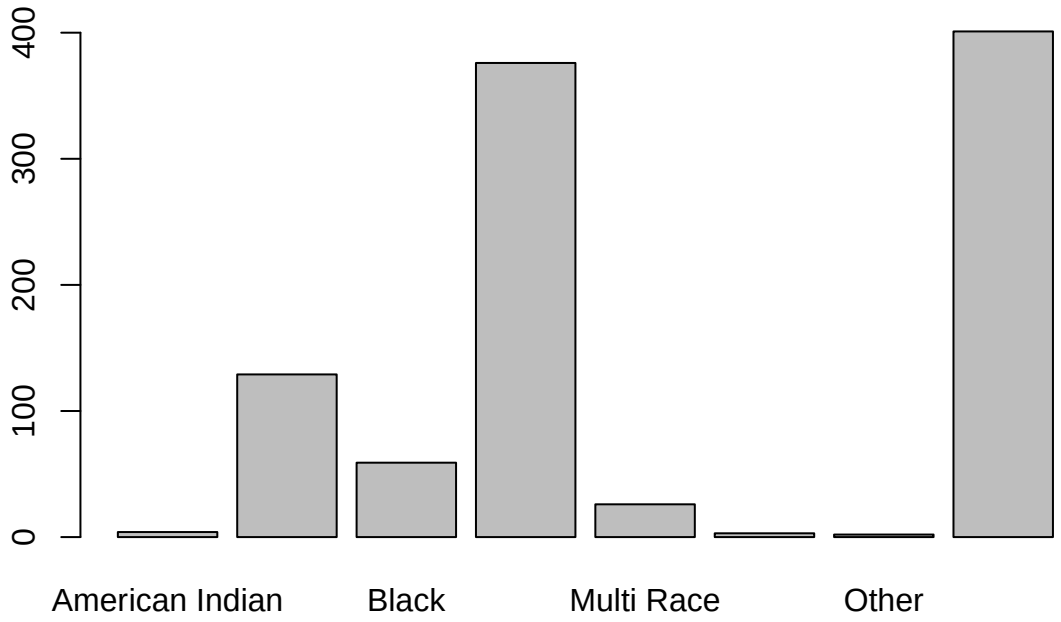

```
summary(dds.discr$age.cohort)
```

```
##   0-5  6-12 13-17 18-21 22-50   51+
##    82   175   212   199   226   106
```

Description: As indicated in the histogram, there is a right skew to this distribution; most consumers are younger than 30 years old. Using the summary() function, we see that the median age is 18 years old.

5) Make an appropriate graph and/or table for ethnicity and describe its distribution. Is there equal representation of ethnic groups in this sample of consumers?

```
plot(dds.discr$ethnicity)
```



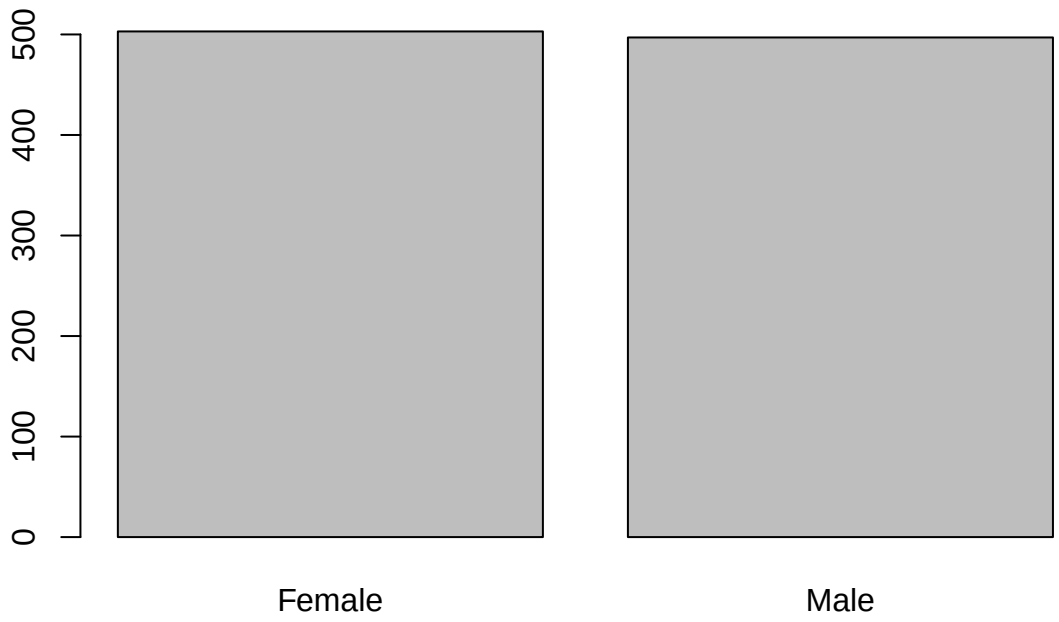

```
prop.table(table(dds.discr$ethnicity))
```

```
##
##    American Indian              Asian              Black           Hispanic
##              0.004              0.129              0.059              0.376
##         Multi Race    Native Hawaiian              Other White not Hispanic
##              0.026              0.003              0.002              0.401
```

Description: There are eight ethnic groups represented in the data, however there is not equal representation. The two largest groups, Hispanics and White non-Hispanics, together represent about 80% of the consumers.

6) Make an appropriate graph and/or table for gender and describe its distribution [note that what they are really reporting here is sex]. Is there equal representation of sexes in this sample of consumers?

```
plot(dds.discr$gender)
```



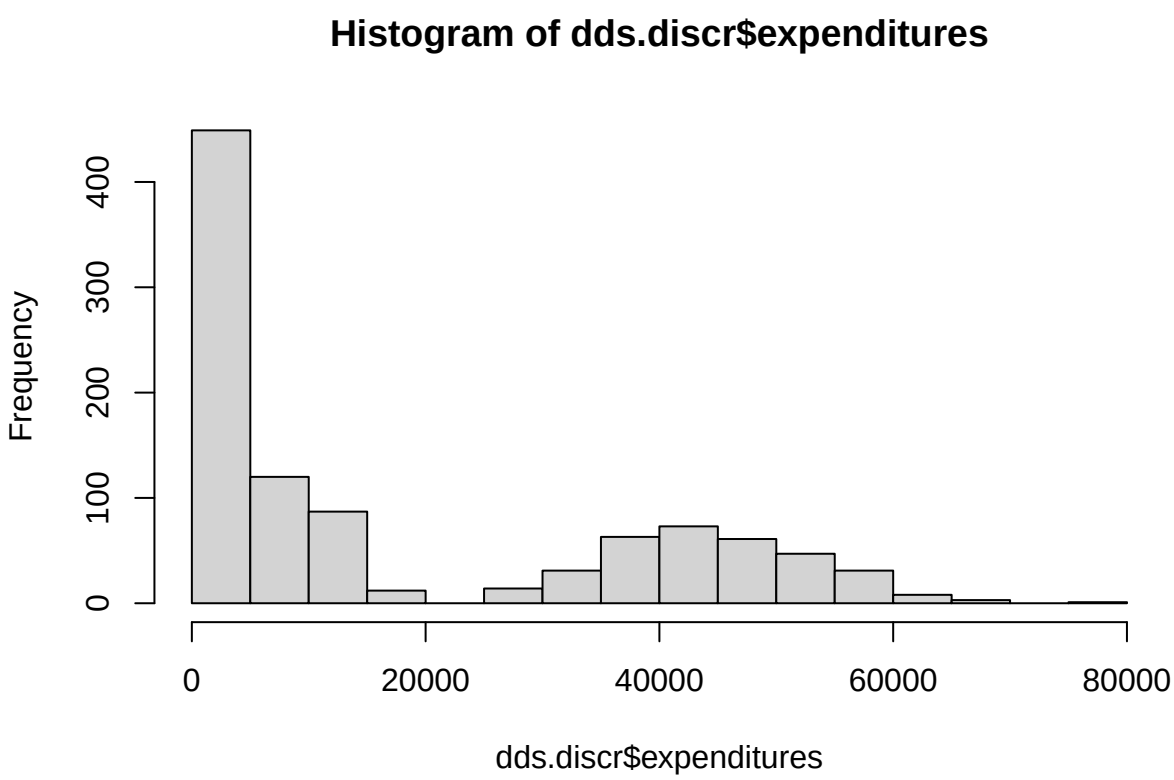

```
table(dds.discr$gender)
```

```
##
## Female   Male
##    503    497
```

Description: Yes, approximately half of the individuals are male and half are female.

7) Make an appropriate graph and/or table for expenditures and describe its distribution. For most of the consumers, is the amount of financial support provided by the DDS relatively high or low?

```
hist(dds.discr$expenditures)
```




```
#Could also plot a boxplot using boxplot(dds.discr$expenditures)

summary(dds.discr$expenditures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     222    2899    7026   18066   37713   75098
```

Description: The distribution of annual expenditures exhibits right skew, indicating that for a majority of consumers, expenditures are relatively low; most are within the 0-5000 dollar range. There are some consumers for which expenditures are much higher, such as within the 60-80K range.

To extract a specific percentile of a continuous variable, you can use the quantile function like this: quantile(variable, decimal), where decimal is the percentile you are interested listed as a decimal.

How would you extract the IQR of expenditure?

```
quantile(dds.discr$expenditures, 0.25)
```

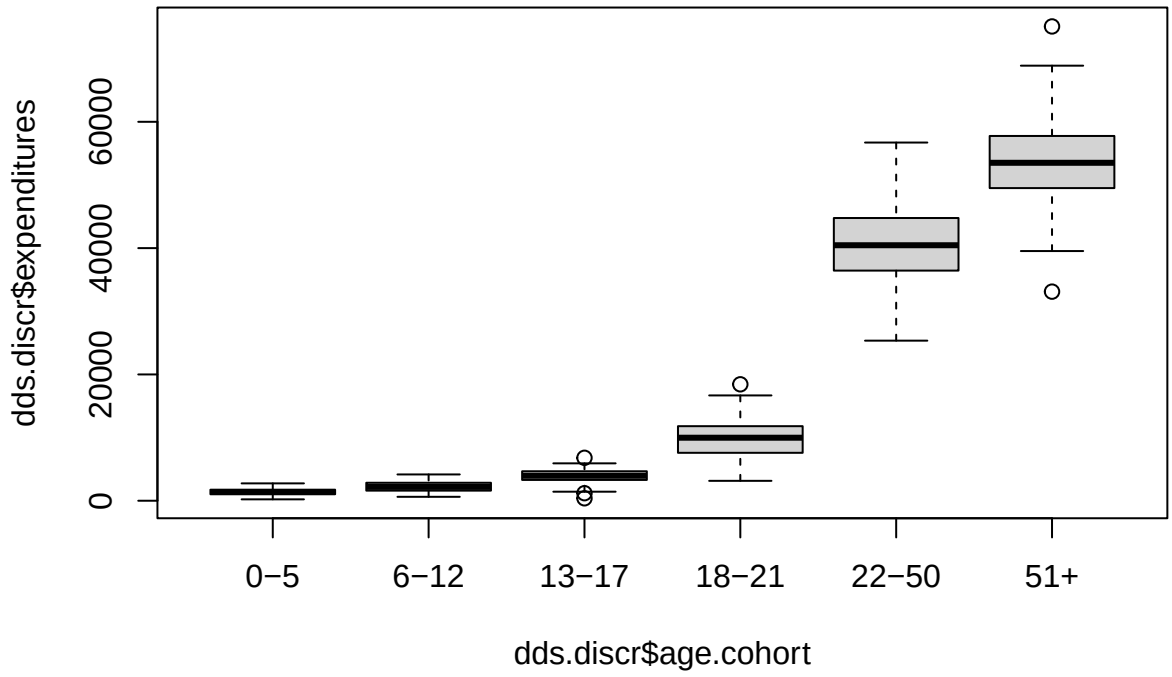
```
##     25%
## 2898.75
```

```
quantile(dds.discr$expenditures, 0.75)
```

```
##      75%
## 37712.75
```

8) How do annual expenditures vary by age? Is there a large amount of variation in expenditures between age cohorts? Make a graph to explore this relationship.

```
boxplot(dds.discr$expenditures~dds.discr$age.cohort)
```

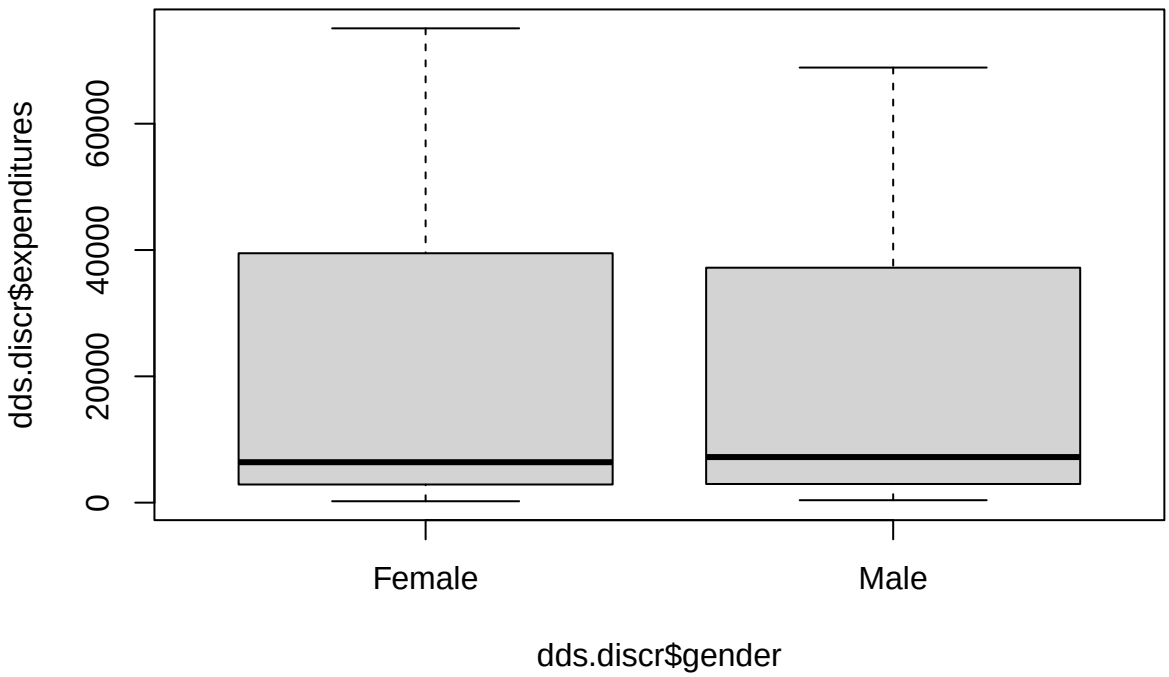



```
#Could also have made a scatterplot
#plot(dds.discr$expenditures~dds.discr$age.cohort)
```

Description: There is a clear upward trend in expenditures as age increases; older individuals tend to receive more DDS funds. For the first three age chorts, average expendistures range between 1400 and 10000 dollars. Average expenditures in the oldest two cohorts, respectively, are about 40 and 54k. Some of the observed variation in expenditures can be attributed to the fact that the dataset includes a wide range of ages.

9) Do annual expenditures seem to vary by gender? Note that the dataset uses the term gender here, but what they are actually recording is sex.

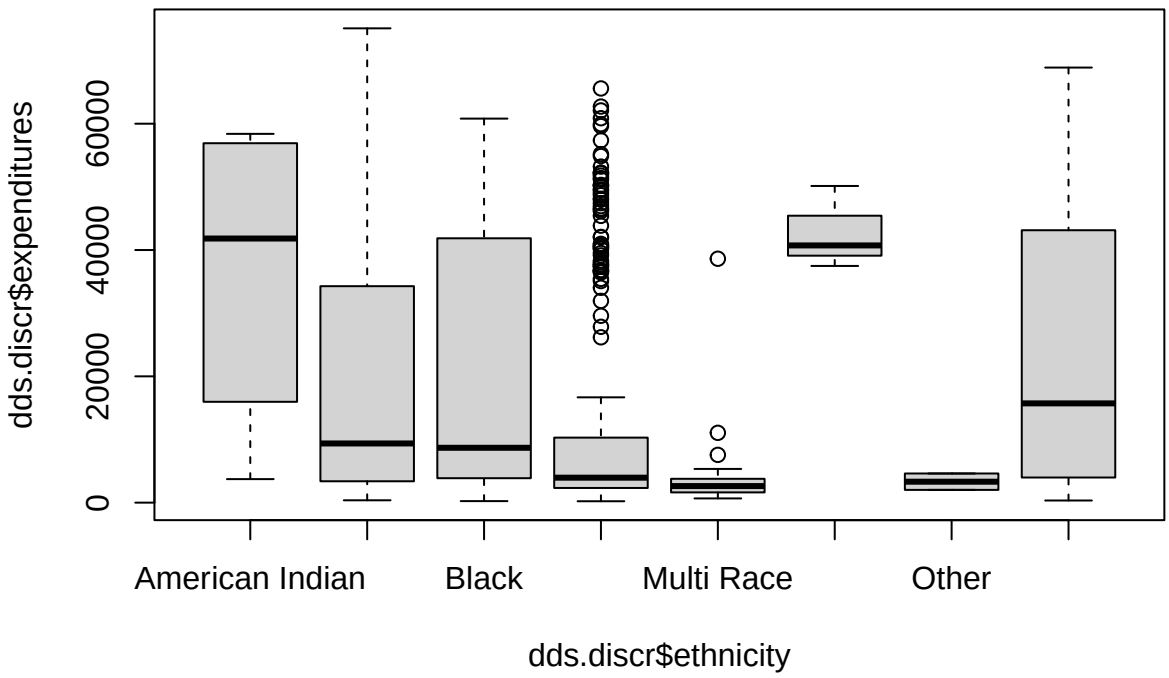
```
boxplot(dds.discr$expenditures~dds.discr$gender)
```



Description: No, the expenditures are the same for both genders (sexes)

10) How does the distribution of expenditures vary by ethnic group? Does there seem to be a difference in the amount of funding that a person receives, on average, between different ethnicities?

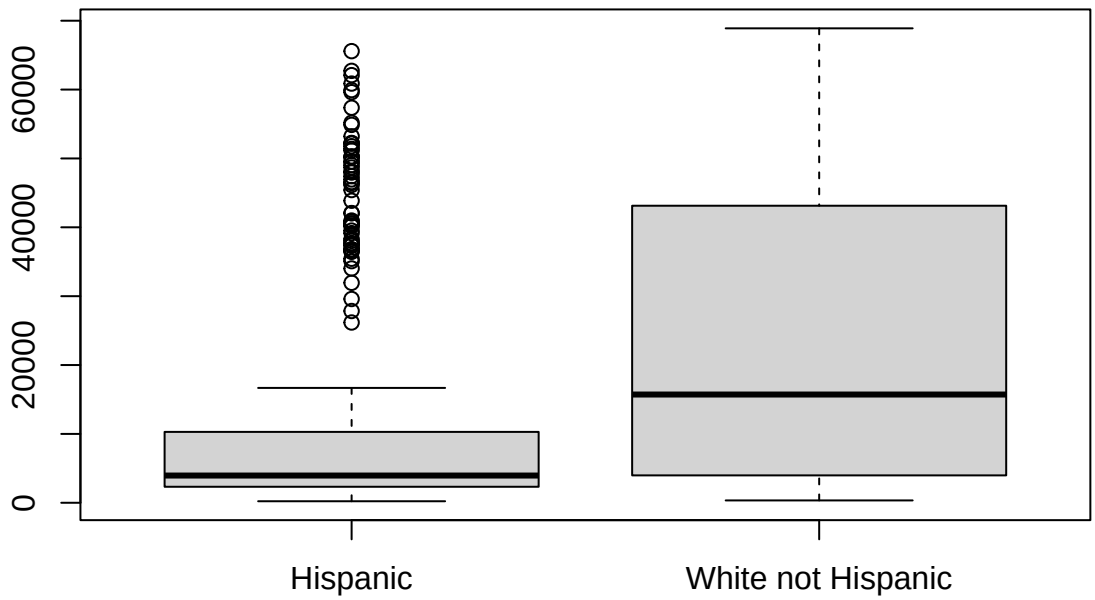
```
boxplot(dds.discr$expenditures~dds.discr$ethnicity)
```



Description: The distribution of expenditures is quite different between ethnic groups. For example, there is very little variation in expenditures within the Multi Race, Native Hawaiian, and Other groups; in other groups, such as the White not Hispanic group, there is a greater range in expenditures. Additionally, there seems to be a difference in the amount of funding that a person receives, on average, between different ethnicities. The median amount of annual support received for individuals in the American Indian and Native Hawaiian groups is about 40k versus medians of approximately 10k for Asian and Black consumers.

11) The following code will let you examine the data for the two largest ethnic groups, Hispanic and White non-Hispanic, which comprise the majority of the data. In prior exercises, we used the subset() function to examine portions of the data independently. Here, we will subset data using brackets that specify specific criteria of interest. What do each of the components of this line of code accomplish? Do Hispanic consumers, on average, seem to receive less financial support from the California DDS than a White non-Hispanic consumer?

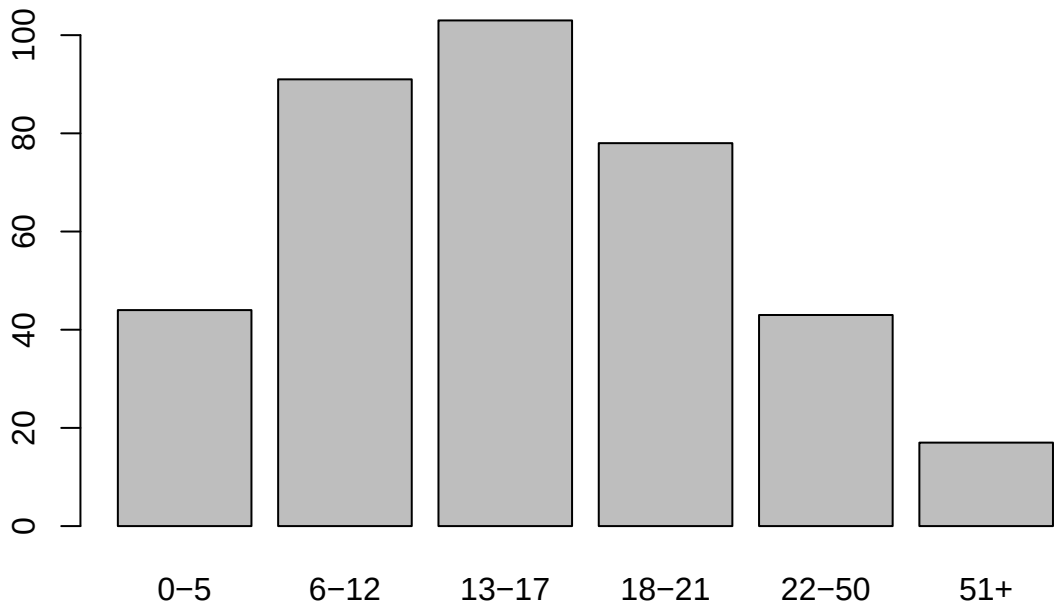
```
boxplot(dds.discr$expenditures[dds.discr$ethnicity=="Hispanic"], dds.discr$expenditures[dds.discr$ethni
```



Description: Based on the boxplot, most Hispanic consumers receive between approximately 0-20K from the California DDS; individuals receiving amounts higher than this are upper outliers. However, for White non-Hispanic consumers, median expenditures are over 15k and the middle 50% of consumers receive between about 4 and 43K. The mean expenditures for Hispanic consumers is ~11K while for White non-Hispanic consumers it is ~24.7k. On average, a Hispanic consumer receives less financial support from the California DDS than a White non-Hispanic consumer.

12) Recall that expenditures are strongly associated with age - older individuals tend to receive more financial support. Is there also an association between age and ethnicity? Examine the distribution of age within each of the two largest ethnic groups and describe your findings.

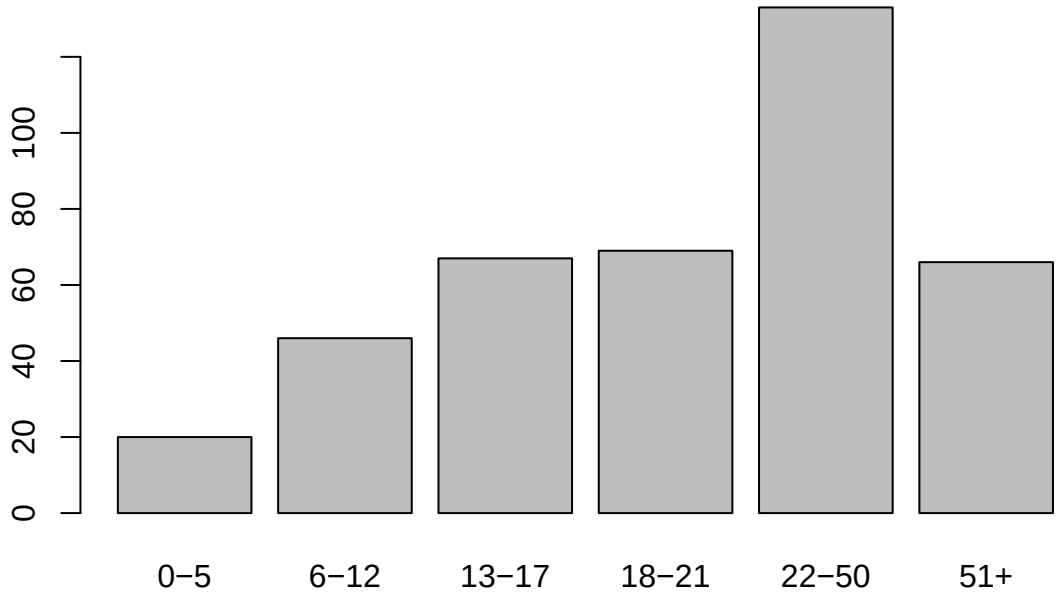
```
plot(dds.discr$age.cohort[dds.discr$ethnicity=="Hispanic"])
```




```
prop.table(table(dds.discr$age.cohort[dds.discr$ethnicity=="Hispanic"]))
```

```
##
##        0-5       6-12      13-17      18-21      22-50        51+
## 0.11702128 0.24202128 0.27393617 0.20744681 0.11436170 0.04521277
```

```
plot(dds.discr$age.cohort[dds.discr$ethnicity=="White not Hispanic"])
```




```
prop.table(table(dds.discr$age.cohort[dds.discr$ethnicity=="White not Hispanic"]))
```

```
##
##        0-5       6-12      13-17      18-21      22-50        51+
## 0.04987531 0.11471322 0.16708229 0.17206983 0.33167082 0.16458853
```

Description:Hispanics tend to be younger, with most Hispanic consumers falling into the 6-12, 13-17, and 18-21 age cohorts. In contrast, White non-Hispanics tend to be older; most consumers in this ethnic group are in the 22-50 age cohort, and relatively more White non-Hispanic consumers are in the 51+ age cohort as compared to Hispanics.

13) A confounding variable is one that is associated with the response variable and the exploratory variable under consideration. In this case, age is a confounding variable for the relationship between expenditures and ethnicity.

For a closer look at the relationship between age, ethnicity, and expenditures, compare how average expenditures differ by ethnicity within each age cohort.

First, subset data into two ethnicity groups:

```
dds.hispanics = dds.discr[dds.discr$ethnicity == "Hispanic",]

dds.white.non.hisp = dds.discr[dds.discr$ethnicity == "White not Hispanic",]
```

Second, calculate mean expenditures by age cohort for Hispanics using the tapply() function:

```
hisp.means = tapply(dds.hispanics$expenditures, dds.hispanics$age.cohort, mean)
hisp.means
```

```
##       0-5      6-12     13-17     18-21     22-50       51+
##  1393.205  2312.187  3955.282  9959.846 40924.116 55585.000
```

```
nonhisp.means = tapply(dds.white.non.hisp$expenditures, dds.white.non.hisp$age.cohort, mean)
nonhisp.means
```

```
##       0-5      6-12     13-17     18-21     22-50       51+
##  1366.900  2052.261  3904.358 10133.058 40187.624 52670.424
```

Q: What does the tapply function seem to do?

A:

Lastly, calculate the difference between Hispanic and White non Hispanic average expenditures across age cohorts. Based on this exploratory analysis, does there seem to be evidence of ethnic discrimination in the amount of financial support provided by the California DDS? Summarize your findings in language accessible to a non-statistician.

```
nonhisp.means - hisp.means
```

```
##         0-5        6-12       13-17       18-21       22-50         51+
##   -26.30455  -259.92594   -50.92334   173.21182  -736.49222 -2914.57576
```

Description: Although on average annual expenditures is lower for Hispanics than for White non-Hispanics, this is due to the difference in age distrigbutions between the two ethnic groups. The population of Hispanic consumers is relatively young compared to the population of White non-Hispanic consumers, and the amount of expenditures for younger consumers tends to be lower than for older consumers. Thus, when expenditures is compared within age cohorts, the differences between mean expenditures for White non Hispanics versus Hispanics are minor. Comparing individuals of similar ages reveals that the association between ethnicity and expenditures is not nearly as strong as it seemed without accounting for the distribution of ages within each ethnic group.