Exploratory Data Analysis: DDS Case Study

Chapter 1, Lab 2: Solutions

OpenIntro Biostatistics

This lab presents the details of how to conduct the analysis discussed in Section 1.7.1 of *OpenIntro Biostatistics*. A reader interested in applied data analysis may benefit from working through this lab and reviewing the solutions instead of reading the section in the text.

Background information

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 California Department of Developmental Services (DDS); individuals receiving DDS funds are referred to as 'consumers'. The dataset `dds.discr` represents a sample of 1,000 DDS consumers (out of a total population of approximately 250,000), and includes information about age, gender, ethnicity, and the amount of financial support per consumer provided by the DDS. The dataset is available in the `oibiostat` package.

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.

Does this finding represent sufficient evidence of ethnic discrimination, or might there be more to the story? 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.

Distributions of single variables

To begin understanding a dataset and developing a sense of context, start by examining the distributions of single variables.

1. Load the `dds.discr` dataset into *RStudio*. Descriptions of the variables are provided in the documentation file. Produce a table of the first five rows in the data matrix.

```
#load the dataset
library(oibiostat)
data("dds.discr")

#produce table of the first five rows
dds.discr[1:5,]
```

```
##      id age.cohort age gender expenditures          ethnicity
## 1 10210      13-17  17 Female         2113 White not Hispanic
## 2 10409      22-50  37   Male        41924 White not Hispanic
```

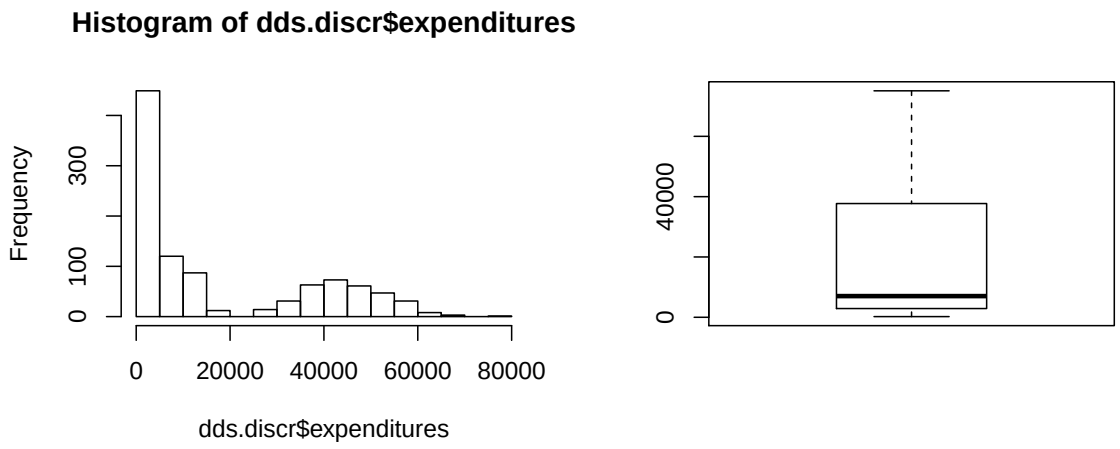
```
## 3 10486        0-5   3   Male         1454           Hispanic
## 4 10538      18-21  19 Female         6400           Hispanic
## 5 10568      13-17  13   Male         4412 White not Hispanic
```

2. Using appropriate numerical and graphical summaries, examine the distributions of each of the variables in the dataset and answer the following questions.

   a) Describe the distribution of annual expenditures. For most consumers, is the amount of financial support provided by the DDS relatively high or low?

   The distribution of annual expenditures exhibits right skew, indicating that for a majority of consumers, expenditures are relatively low; most are within the $0 - $5,000 range. There are some consumers for which expenditures are much higher, such as within the $60,000 - $80,000 range. The quartiles for expenditures are $2,899, $7,026, and $37,710.

```
#graphical summaries
par(mfrow = c(1, 2))  #displays plots as 1 row / 2 column layout
hist(dds.discr$expenditures)
boxplot(dds.discr$expenditures)
```




```
#numerical summaries
summary(dds.discr$expenditures)
```

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

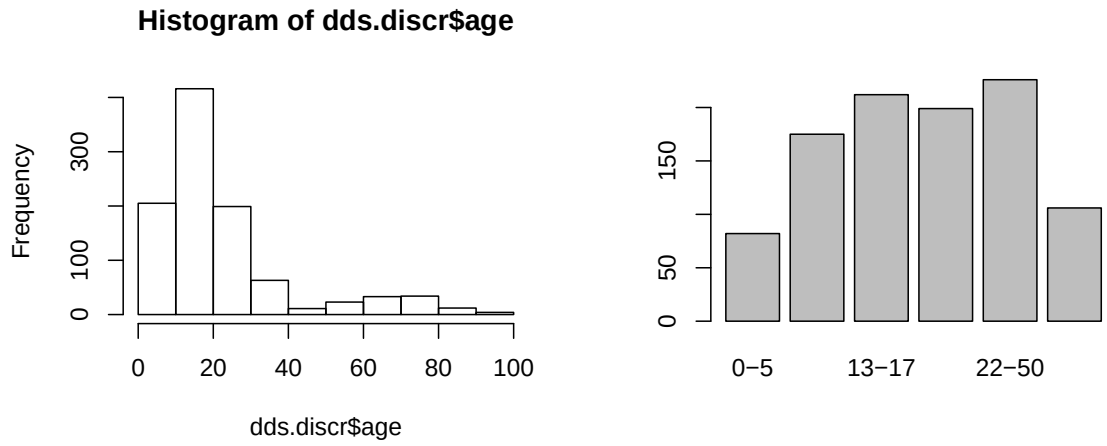
   b) The variable age directly records a consumer's age; in the `age.cohort` variable, consumers are assigned to one of six age cohorts. Describe the distribution of age in this sample of consumers. Do consumers tend to be older or younger?

   The cohorts are indicative of particular life phases. In the first three cohorts, consumers are still living with their parents as they move through preschool age, elementary/middle school age, and high school age. In the 18-21 cohort, consumers are transitioning from their parents' homes to living on their own or in supportive group homes. From ages 22-50, individuals are mostly no longer living with their parents but may still receive

some support from family. In the 51+ cohort, consumers often have no living parents and typically require the most amount of support.

As indicated in the histogram, there is right-skewing; most consumers are younger than 30 years old. The median age is 18 years. There are approximately 200 individuals in each of the middle four cohorts, and about 100 individuals in the other two cohorts.

```
#graphical summaries
par(mfrow = c(1, 2))  #displays the following plots as 1 row / 2 column layout
hist(dds.discr$age)
plot(dds.discr$age.cohort)
```




```
#numerical summaries
summary(dds.discr$age)
```

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

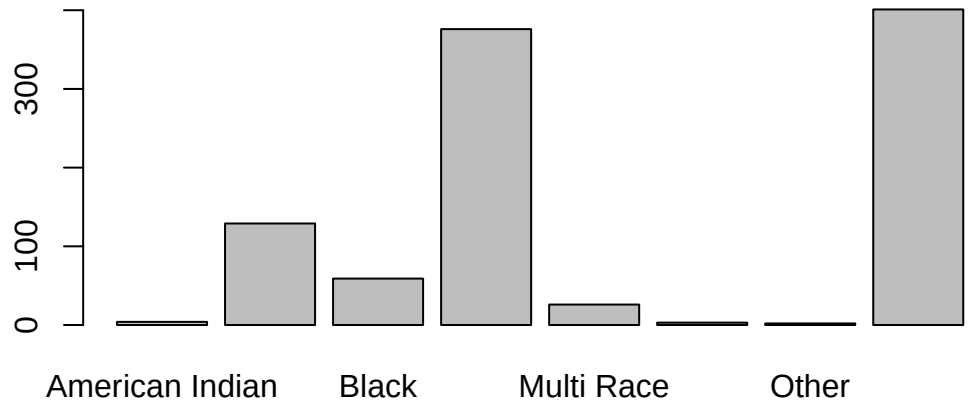
```
table(dds.discr$age.cohort)
```

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

c) Is there an equal representation of ethnic groups in this sample of consumers?

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.

```
#graphical summaries
plot(dds.discr$ethnicity)
```




```r
#numerical summaries
table(dds.discr$ethnicity)
```

```
##
##    American Indian              Asian              Black           Hispanic
##                  4                129                 59                376
##         Multi Race    Native Hawaiian              Other White not Hispanic
##                 26                  3                  2                401
```

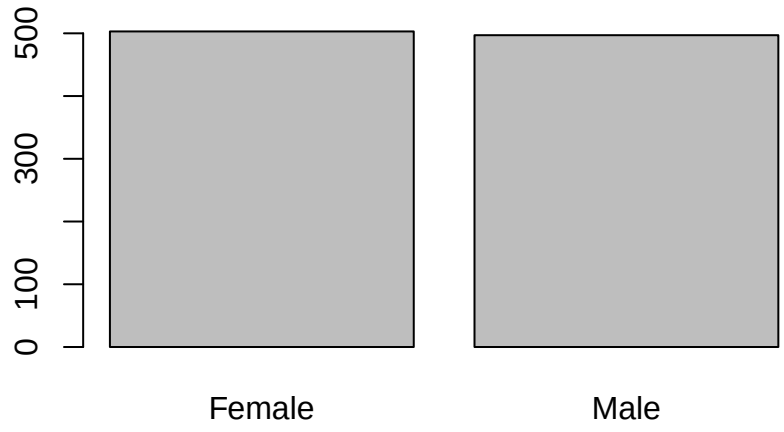
```r
prop.table(table(dds.discr$ethnicity))  #converts a table of counts to proportions
```

```
##
##    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
```

d) Does gender appear to be balanced in this sample of consumers?

Yes, approximately half the individuals are female and half are male.

```r
#graphical summaries
plot(dds.discr$gender)
```




```
#numerical summaries
table(dds.discr$gender)
```

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

Relationships between two variables

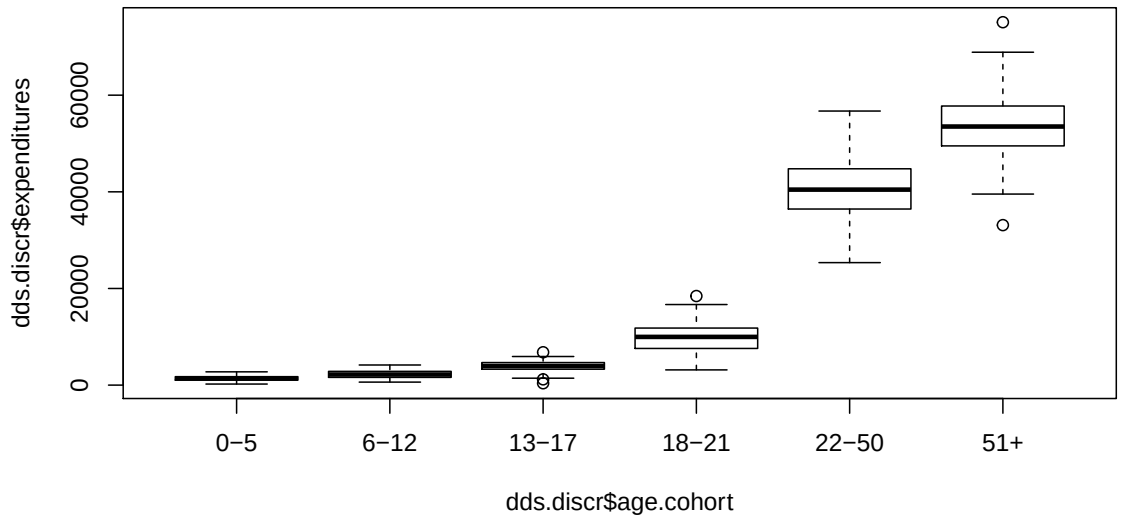
After examining variables individually, explore how variables are related to each other. It is often useful to start by investigating the relationships between two variables, particularly between the primary response variable of interest and the exploratory variables. For this case study, the response variable is `expenditures`, the amount of funds the California DDS allocates annually to each consumer.

3. How do annual expenditures vary by age? Is there a large amount of variation in expenditures between age cohorts? Use the `age.cohort` variable.

   There is a clear upward trend in expenditures as age increases; older individuals tend to receive more DDS funds. For the first three age cohorts, average expenditures ranges between $1,400 to $10,000. Average expenditures in the oldest two cohorts, respectively, are about $40,000 and $53,500. Some of the observed variation in `expenditures` can be attributed to the fact that the dataset includes a wide range of ages. If the data included only individuals from one age cohort, such as the 18-21 year cohort, the distribution would be less variable, and range between $3,000 and $20,000 rather than $0 and $75,000.

   The upward trend reflects the underlying context of the data. The purpose of providing funds to developmentally disabled individuals is to help them maintain a quality of life similar to those without disabilities; as individuals age, it is expected that their financial needs will increase.

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




```
#numerical summaries
summary(dds.discr$expenditures[dds.discr$age.cohort == "0-5"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     222    1034    1380    1415    1739    2750
```

```
summary(dds.discr$expenditures[dds.discr$age.cohort == "6-12"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     620    1602    2191    2227    2846    4163
```

```
summary(dds.discr$expenditures[dds.discr$age.cohort == "13-17"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     386    3306    3952    3923    4666    6798
```

```
summary(dds.discr$expenditures[dds.discr$age.cohort == "18-21"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3153    7588    9979    9889   11806   18435
```

```
summary(dds.discr$expenditures[dds.discr$age.cohort == "22-50"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25348   36447   40456   40209   44721   56716
```

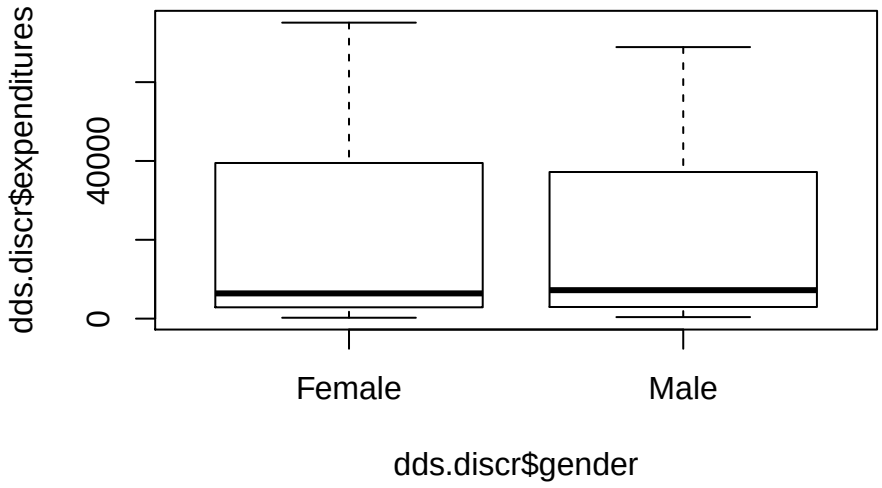
```
summary(dds.discr$expenditures[dds.discr$age.cohort == "51+"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33110   49515   53509   53522   57746   75098
```

4. Do annual expenditures seem to vary by gender?

   No, the distribution of expenditures within males and females is very similar; both are right skewed, with approximately equal median and interquartile range.

```
#graphical summaries
boxplot(dds.discr$expenditures ~ dds.discr$gender)
```



```
#numerical summaries
summary(dds.discr$expenditures[dds.discr$gender == "Male"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     386    2954    7219   18001   37201   68890
```

```
summary(dds.discr$expenditures[dds.discr$gender == "Female"])
```

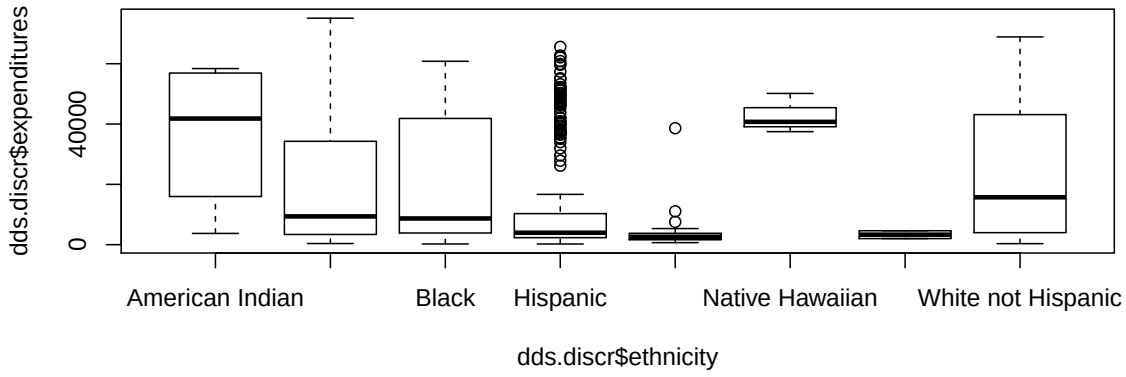
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     222    2872    6400   18130   39488   75098
```

5. 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?

   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 $40,000, versus medians of approximately $10,000 for Asian and Black consumers.

   Rather than using summary()) for each ethnicity group, the tapply() command allows for a convenient shortcut. The specific syntax of the tapply() function is explained in the lab notes.

```
#graphical summaries
boxplot(dds.discr$expenditures ~ dds.discr$ethnicity)
```




```
#numerical summaries
summary(dds.discr$expenditures[dds.discr$ethnicity == "American Indian"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3726   22085   41818   36438   56171   58392
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Asian"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     374    3382    9369   18392   34274   75098
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Black"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     240    3870    8687   20885   41857   60808
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Hispanic"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     222    2331    3952   11066   10292   65581
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Multi Race"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     669    1690    2622    4457    3750   38619
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Native Hawaiian"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37479   39103   40727   42782   45434   50141
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "Other"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2018    2667    3316    3316    3966    4615
```

```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "White not Hispanic"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     340    3977   15718   24698   43134   68890
```

```r
#bonus: using tapply( )
tapply(dds.discr$expenditures, dds.discr$ethnicity, summary)
```

```
## $`American Indian`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3726   22085   41818   36438   56171   58392
##
## $Asian
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     374    3382    9369   18392   34274   75098
##
## $Black
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     240    3870    8687   20885   41857   60808
##
## $Hispanic
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     222    2331    3952   11066   10292   65581
```

```
##
## $`Multi Race`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     669    1690    2622    4457    3750   38619
##
## $`Native Hawaiian`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37479   39103   40727   42782   45434   50141
##
## $Other
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2018    2667    3316    3316    3966    4615
##
## $`White not Hispanic`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     340    3977   15718   24698   43134   68890
```

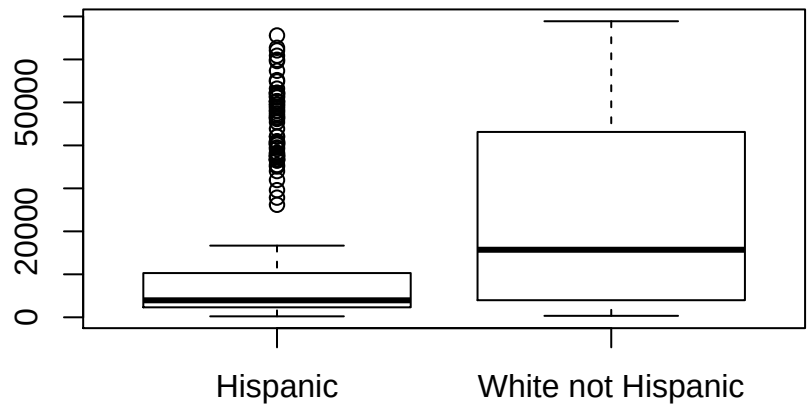
A closer look

As shown in Question 2c), two of the ethnic groups, Hispanic and White non-Hispanic, comprise the majority of the data; some ethnic groups represent less than 10% of the observations. For ethnic groups with relatively small sample sizes, it is possible that the observed samples are not representative of the larger populations. The rest of this analysis will focus on comparing how `expenditures` varies between the two largest ethnic groups.

6. Compare the distribution of `expenditures` between Hispanic and White non-Hispanic consumers, graphically and numerically. Do Hispanic consumers, on average, seem to receive less financial support from the California DDS than a White non-Hispanic consumer?

   Based on the boxplot, most Hispanic consumers receive between approximately $0 to $20,000 from the California DDS; individuals receiving amounts higher than this are upper outliers. However, for White non-Hispanic consumers, median `expenditures` is at $15,718, and the middle 50% of consumers receive between about $4,000 and $43,000. The mean expenditures for Hispanic consumers is $11,066, while the mean expenditures for White non-Hispanic consumers is over twice as high at $24,698. On average, a Hispanic consumer receives less financial support from the California DDS than a White non-Hispanic consumer.

```r
#graphical summaries
boxplot(dds.discr$expenditures[dds.discr$ethnicity == "Hispanic"],
    dds.discr$expenditures[dds.discr$ethnicity == "White not Hispanic"],
    names = c("Hispanic", "White not Hispanic"))
```




```
#numerical summaries
summary(dds.discr$expenditures[dds.discr$ethnicity == "Hispanic"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     222    2331    3952   11066   10292   65581
```

```
IQR(dds.discr$expenditures[dds.discr$ethnicity == "Hispanic"])
```

```
## [1] 7961.25
```

```
summary(dds.discr$expenditures[dds.discr$ethnicity == "White not Hispanic"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     340    3977   15718   24698   43134   68890
```

```
IQR(dds.discr$expenditures[dds.discr$ethnicity == "White not Hispanic"])
```

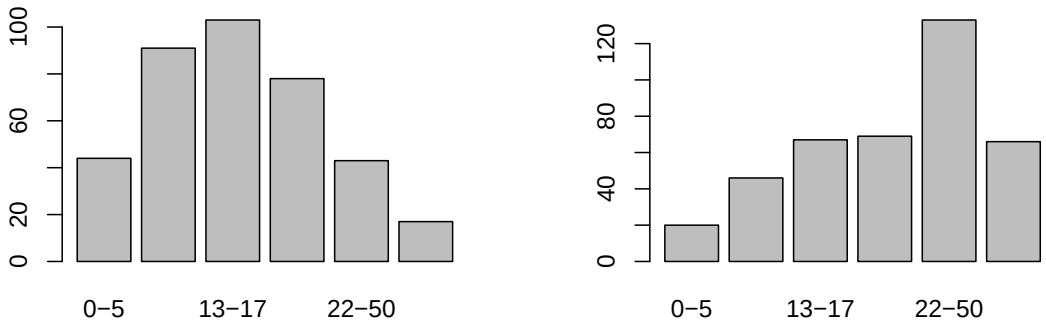
```
## [1] 39157
```

7. Recall that expenditures is strongly associated with age—older individuals tend to receive more financial support. Is there also an association between age and ethnicity, for these two ethnic groups? Examine the distribution of age within each group and describe your findings.

   When using data to investigate a question, it is important to explore not only how explanatory variables are related to the response variable(s), but also how explanatory variables influence each other.

   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.

```r
#graphical summaries
par(mfrow = c(1, 2))  #displays the following plots as 1 row / 2 column layout
plot(dds.discr$age.cohort[dds.discr$ethnicity == "Hispanic"])
plot(dds.discr$age.cohort[dds.discr$ethnicity == "White not Hispanic"])
```




```r
#numerical summaries
table(dds.discr$age.cohort[dds.discr$ethnicity == "Hispanic"])
```

```
##
##   0-5  6-12 13-17 18-21 22-50   51+
##    44    91   103    78    43    17
```

```r
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
```

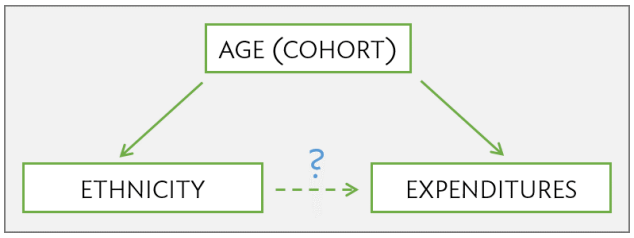
```r
table(dds.discr$age.cohort[dds.discr$ethnicity == "White not Hispanic"])
```

```
##
##   0-5  6-12 13-17 18-21 22-50   51+
##    20    46    67    69   133    66
```

```r
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
```

Recall that a confounding variable is a variable that is associated with the response variable and the exploratory variable under consideration; confounding was initially introduced in the context of sunscreen use and the incidence of skin cancer, where sun exposure is a confounder. In this setting, age is a confounder for the relationship between `expenditures` and `ethnicity`. Just as it would be incorrect to claim that sunscreen causes skin cancer, it is essential here to recognize that there is more to the story than the apparent association between `expenditures` and `ethnicity`.



8. For a closer look at the relationship between age, ethnicity, and expenditures, compare how average `expenditures` differs by ethnicity within each age cohort. If age is indeed the primary source of the observed variation in expenditures, then there should be little difference in average `expenditures` between individuals in different ethnic groups but the same age cohort. Is this the case? Describe your findings.

   When `expenditures` is compared within age cohorts, there are not large differences between mean `expenditures` for White non-Hispanics versu Hispanics. Comparing individuals of similar ages reveals that the association between ethnicity and expenditures is not nearly as strong as it seemed from the initial comparison of overall averages.

```
#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", ]

#calculate mean expenditures by age cohort for Hispanics
hisp.mean.0to5 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                 "0-5"])
hisp.mean.6to12 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                  "6-12"])
hisp.mean.13to17 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                   "13-17"])
hisp.mean.18to21 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                   "18-21"])
hisp.mean.22to50 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                   "22-50"])
hisp.mean.51 = mean(dds.hispanics$expenditures[dds.hispanics$age.cohort ==
                                                   "51+"])

#calculate mean expenditures by age cohort for White non Hispanics
nonhisp.mean.0to5 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
                                                     age.cohort == "0-5"])
nonhisp.mean.6to12 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
                                                     age.cohort == "6-12"])
nonhisp.mean.13to17 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
```

```
                                                        age.cohort == "13-17"])
nonhisp.mean.18to21 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
                                                        age.cohort == "18-21"])
nonhisp.mean.22to50 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
                                                        age.cohort == "22-50"])
nonhisp.mean.51 = mean(dds.white.non.hisp$expenditures[dds.white.non.hisp$
                                                        age.cohort == "51+"])

#calculate differences in mean expenditures between ethnicity groups
hisp.means = c(hisp.mean.0to5, hisp.mean.6to12, hisp.mean.13to17,
           hisp.mean.18to21, hisp.mean.22to50, hisp.mean.51)
hisp.means
```

```
## [1]  1393.205  2312.187  3955.282  9959.846 40924.116 55585.000
```

```
nonhisp.means = c(nonhisp.mean.0to5, nonhisp.mean.6to12, nonhisp.mean.13to17,
              nonhisp.mean.18to21, nonhisp.mean.22to50, nonhisp.mean.51)
nonhisp.means
```

```
## [1]  1366.900  2052.261  3904.358 10133.058 40187.624 52670.424
```

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

```
## [1]   -26.30455  -259.92594   -50.92334   173.21182  -736.49222 -2914.57576
```

```
#bonus: using tapply( )
hisp.means = tapply(dds.hispanics$expenditures, dds.hispanics$age.cohort, mean)
nonhisp.means = tapply(dds.white.non.hisp$expenditures, dds.white.non.hisp$age.cohort,
mean)
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
```

9. 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.

   There does not seem to be evidence of ethnic discrimination. Although the average annual expenditures is lower for Hispanics than for White non-Hispanics, this is due to the difference in age distributions 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. When individuals of similar ages are compared, there are not large differences in the average amount of financial support provided to a Hispanic consumer versus a White non-Hispanic consumer.

Simpson's paradox

Identifying confounding variables is essential for understanding data. Confounders are often context-specific; for example, age is not necessarily a confounder for the relationship between ethnicity and expenditures in a different population. Additionally, it is rarely immediately obvious which variables in a dataset are confounders; looking for confounding variables is an integral part of exploring a dataset.

These data represent an extreme example of confounding known as **Simpson's paradox**, in which an association observed in several groups may disappear or reverse direction once the groups are combined. In other words, an association between two variables X and Y may disappear or reverse direction once data are partitioned into subpopulations based on a third variable Z, the confounding variable.

Mean `expenditures` is higher for Hispanics than White non-Hispanics in all age cohorts except one. Yet, once all the data are aggregated, the average expenditures for White non-Hispanics is over twice as large as the average for Hispanics. This paradox can be explored from a mathematical perspective by using weighted averages, where the average expenditure for each cohort is weighted by the proportion of the population in that cohort.

10. Calculate the overall weighted average expenditures for Hispanics and for White non-Hispanics, using the proportions of individuals in each age cohort (Question 7) and the average expenditures for each Cohort (Question 8). How does the weighting lead to overall average expenditures for White non-Hispanics to be higher than for Hispanics?

    The weights for the youngest four cohorts, which have lower expenditures, are higher for the Hispanic population than the White non-Hispanic population; additionally, the weights for the oldest two cohorts, which have higher expenditures, are higher for the White non-Hispanic population. This leads to overall average expenditures for the White non-Hispanics to be higher than for Hispanics.

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

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

```
hisp.weights*hisp.means
```

```
##
##       0-5      6-12     13-17     18-21     22-50       51+
##  163.0346  559.5984 1083.4947 2066.1383 4680.1516 2513.1516
```

```
sum(hisp.weights*hisp.means)
```

```
## [1] 11065.57
```

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

```
nonhisp.weights
```

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

```
nonhisp.weights*nonhisp.means
```

```
##
##         0-5        6-12       13-17       18-21       22-50         51+
##    68.17456   235.42145   652.34913  1743.59352 13329.06234  8668.94763
```

```
sum(nonhisp.weights*nonhisp.means)
```

```
## [1] 24697.55
```