Lab 1: Intro to Data Solutions

Statistics 139

September 08, 2023

Topics

- Basic Data Cleaning
- Numerical and Graphical Summaries
- Subsetting Data

The data in the file 'dds_discrimination.csv' represent a sample of 1,000 residents of California who receive funds from the California Department of Developmental Services (DDS); individuals receiving funds are referred to as 'consumers'.

A study team 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?

The following variables are included in the dataset.

- `ID`: consumer ID number
- `Age.Cohort`: age group, where `1` refers to 0 - 5 years, `2` refers to 51+ years, `3` refers to 13 - 17 years, `4` refers to 18 - 21 years, `5` refers to 22 - 50 years, and `6` refers to 6 - 12 years.
- `Age`: age in years
- `Gender`: gender, recorded as `1` for female and `2` for male
- `Expenditures`: annual expenditure in dollars
- `Ethnicity`: ethnicity, recorded as either `1` for American Indian, `2` for Asian, `3` for Black, `4` for Hispanic, `5` for Multi Race, `6` for Native Hawaiian, `7` for Other, and `8` for White not Hispanic.

Problem 1: A little data clean-up

First, a bit of data cleaning will be helpful.

a) Read the file into R as the `dds.discr` dataframe, and examine the first few observations.

```
#read file into R
dds.discr <- read.csv("data/dds_discrimination.csv")
head(dds.discr)
```

```
##   X    ID Age.Cohort Age Gender Expenditures Ethnicity
## 1 1 10210          3  17      1         2113         8
## 2 2 10409          5  37      2        41924         8
## 3 3 10486          1   3      2         1454         4
## 4 4 10538          4  19      1         6400         4
## 5 5 10568          3  13      2         4412         8
## 6 6 10690          3  15      1         4566         4
```

b) The first column contains a 'variable' `X` that is just the row number as carried over from the CSV file. Run the following to eliminate `X`.

```
#remove first column
dds.discr[,1] <- NULL

#alternatively, we can give an extra argument into the read.csv function
dds.discr2 <- read.csv("data/dds_discrimination.csv", row.names=1)
head(dds.discr2)
```

```
##      ID Age.Cohort Age Gender Expenditures Ethnicity
## 1 10210          3  17      1         2113         8
## 2 10409          5  37      2        41924         8
## 3 10486          1   3      2         1454         4
## 4 10538          4  19      1         6400         4
## 5 10568          3  13      2         4412         8
## 6 10690          3  15      1         4566         4
```

c) Datasets can sometimes have variables with long or messy names. For the sake of practice, read the documentation for the `colnames( )` function and change the names of the variables to ones you find more convenient. To access the R help files for a function, type `?` before the function name.

```
# your work here

names(dds.discr) = c("id","age.cohort" ,"age","gender","expenditures","ethnicity")
names(dds.discr)
```

```
## [1] "id"           "age.cohort"   "age"          "gender"       "expenditures"
## [6] "ethnicity"
```

d) Let's look again at the dataset, this time using the str() function. What looks strange about the Gender and Expenditures variables? Hint: You've seen this in class already...

```
#maybe better to use str()
str(dds.discr)
```

```
## 'data.frame':    1000 obs. of  6 variables:
##  $ id          : int  10210 10409 10486 10538 10568 10690 10711 10778 10820 10823 ...
##  $ age.cohort  : int  3 5 1 4 3 3 3 3 3 3 ...
##  $ age         : int  17 37 3 19 13 15 13 17 14 13 ...
##  $ gender      : chr  "1" "2" "2" "1" ...
##  $ expenditures: chr  "2113" "41924" "1454" "6400" ...
##  $ ethnicity   : int  8 8 4 4 8 4 8 3 8 4 ...
```

Explain how the following two lines are designed to find the problems with the Gender and Expenditures variables, and explain what those problems are. You might have to read the documentation by typing `?` before any function names you don't know.

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

```
##
##      1      2 female FEMALE   male   Male   MALE
##    500    494      2      1      1      1      1
```

```
dds.discr$expenditures[which(is.na(as.numeric(dds.discr$expenditures)))]
```

```
## Warning in which(is.na(as.numeric(dds.discr$expenditures))): NAs introduced by
## coercion
```

```
## [1] "$46,571 " "$42,192 " "$54,616 " "$60,871 " "$3,673 "
```

Challenge: Explain how the following four lines of code fix the problem with the Expenditures variable. It's ok if you can't figure it out, this one is hard. You can just run the code for now.

```
dds.discr.uppercase <- toupper(dds.discr$gender)
dds.discr.uppercase[which(dds.discr.uppercase == "MALE")] <- "2"
dds.discr.uppercase[which(dds.discr.uppercase == "FEMALE")] <- "1"
dds.discr$gender <- as.numeric(dds.discr.uppercase)
```

Challenge: Explain how the following two lines of code fix the problem with the Expenditures variable. It's ok if you can't figure it out, this one is hard. You can just run the code for now.

```
dds.discr$expenditures[which(is.na(as.numeric(dds.discr$expenditures)))] <- gsub("\\$","",dds.d
```

```
## Warning in which(is.na(as.numeric(dds.discr$expenditures))): NAs introduced by
## coercion

## Warning in which(is.na(as.numeric(dds.discr$expenditures))): NAs introduced by
## coercion
```

```
dds.discr$expenditures[which(is.na(as.numeric(dds.discr$expenditures)))] <- gsub(",","",dds.dis
```

```
## Warning in which(is.na(as.numeric(dds.discr$expenditures))): NAs introduced by
## coercion

## Warning in which(is.na(as.numeric(dds.discr$expenditures))): NAs introduced by
## coercion
```

```
dds.discr$expenditures <- as.numeric(dds.discr$expenditures)
```

e) The categorical variables (age cohort, gender, and ethnicity) should be converted to factor variables. Read the documentation for `factor( )` and recode these three variables. Note that age cohort is an ordered categorical variable.

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

```
##        id          age.cohort         age           gender       expenditures
##  Min.   :10210   Min.   :1.000   Min.   : 0.0   Min.   :1.000   Min.   :  222
##  1st Qu.:31809   1st Qu.:3.000   1st Qu.:12.0   1st Qu.:1.000   1st Qu.: 2899
##  Median :55384   Median :4.000   Median :18.0   Median :1.000   Median : 7026
##  Mean   :54663   Mean   :3.906   Mean   :22.8   Mean   :1.497   Mean   :18066
##  3rd Qu.:76135   3rd Qu.:5.000   3rd Qu.:26.0   3rd Qu.:2.000   3rd Qu.:37713
##  Max.   :99898   Max.   :6.000   Max.   :95.0   Max.   :2.000   Max.   :75098
##    ethnicity
##  Min.   :1.000
##  1st Qu.:4.000
##  Median :4.000
##  Mean   :5.313
##  3rd Qu.:8.000
##  Max.   :8.000
```

```
dds.discr$gender =  factor(dds.discr$gender, levels=1:2, labels=c("F","M"))

dds.discr$age.cohort = factor(dds.discr$age.cohort, levels = c(1,6,3,4,5,2),
                              labels=c("0-5","6-12","13-17","18-21","22-50","51+"))
dds.discr$ethnicity =  factor(dds.discr$ethnicity, levels=1:8, labels=c("American Indian","Asia
```

f) Save the clean version of the dataframe as an `.Rdata` file. (Alternatively, you could use `write.csv( )` to write the dataframe to a CSV file.)

```
#save the file
save(dds.discr, file = "dds_discr.Rdata")
```

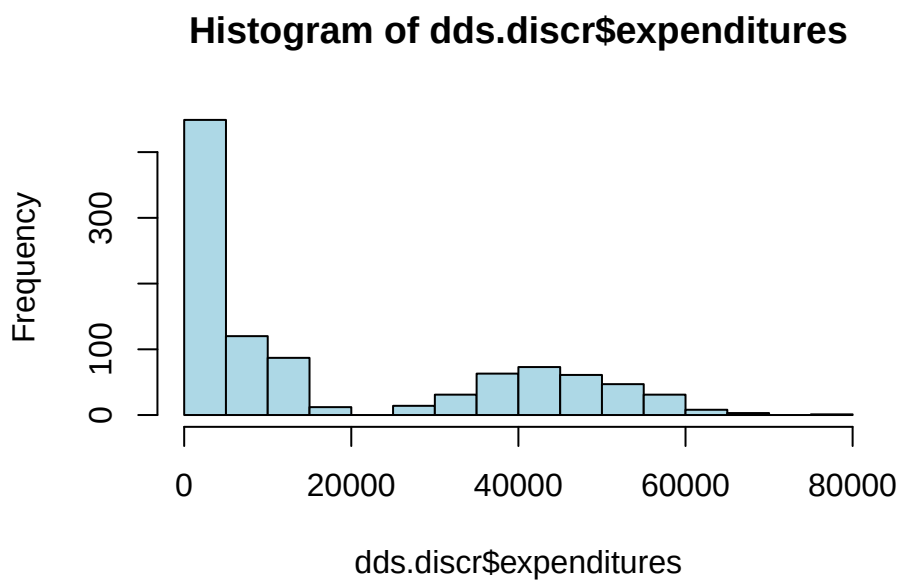
Problem 2: Univariate Explorations (aka, distributions)

To develop a sense of{ context for the data, start by examining the distributions of single variables. Create numerical and graphical summaries as appropriate.

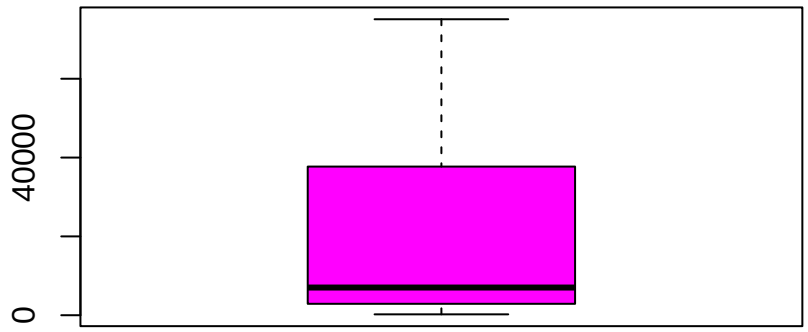
a) Describe the distribution of annual expenditures. For most consumers, is the amount of financial support provided by the DDS relatively high or low?

The distribution is bimodal (i.e., there are two "bumps"). Half the consumers have relatively low expenditures, less than about $7,000. The other half follow a bell-shaped distribution centered around $40k. The box and whisker plot does not suggest what we would classify as outliers. We might say there is evidence that there are two relevant subgroups of consumers (i.e., were are observing a "mixture distribution").

```
#graphical summaries
hist(dds.discr$expenditures,col="lightblue")
```




```
boxplot(dds.discr$expenditures, col="magenta")
```



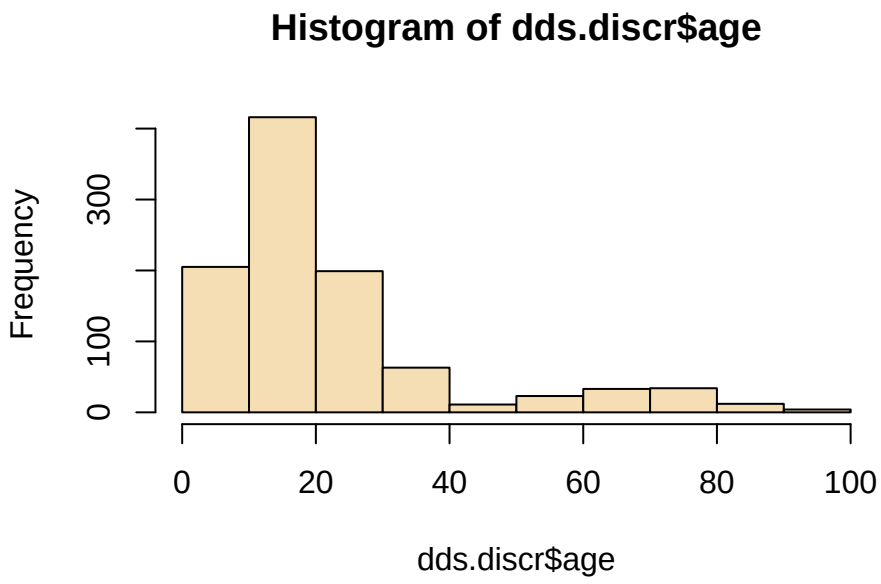

```
#numerical summaries
summary(dds.discr$expenditures)
```

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

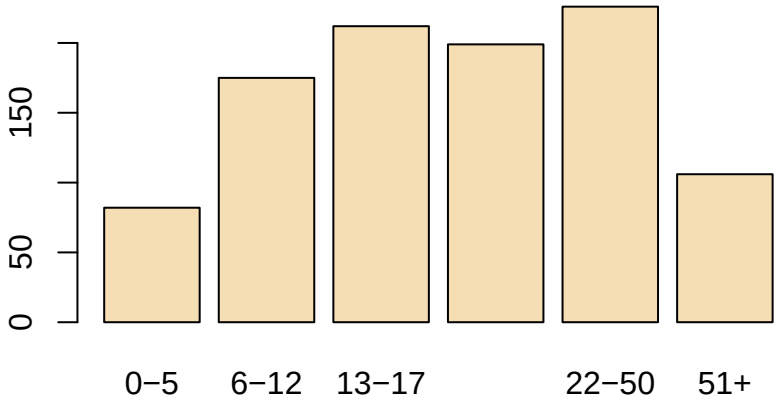
b) Do consumers tend to be older or younger?

Most consumers are less than 18 (note this is much younger than the US population in general).

```
#graphical summaries
hist(dds.discr$age, col = "wheat")
```




```
plot(dds.discr$age.cohort, col = "wheat")
```




```
#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
```

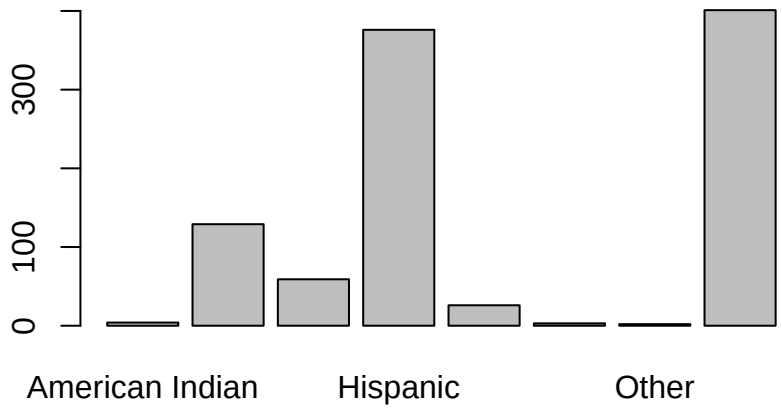
```
table(dds.discr$age.cohort)
```

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

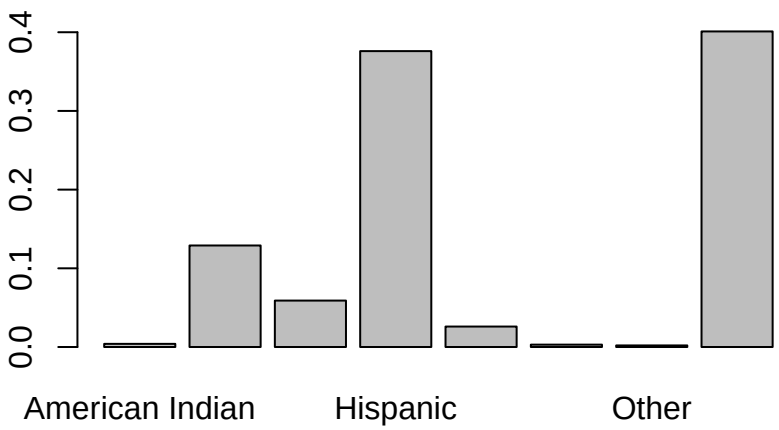
c) Is there an equal representation among ethnic groups?

No. The two largest groups, Hispanics and White non-Hispanics, together represent nearly 80% of the consumers (77.7%).

```
#graphical summaries
barplot(table(dds.discr$ethnicity))
```




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




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

```
##
## American Indian           Asian           Black        Hispanic      Multi Race
##               4             129              59             376              26
## Native Hawaiian           Other           White
##               3               2             401
```

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

```
##
## American Indian           Asian           Black        Hispanic      Multi Race
##           0.004           0.129           0.059           0.376           0.026
## Native Hawaiian           Other           White
##           0.003           0.002           0.401
```

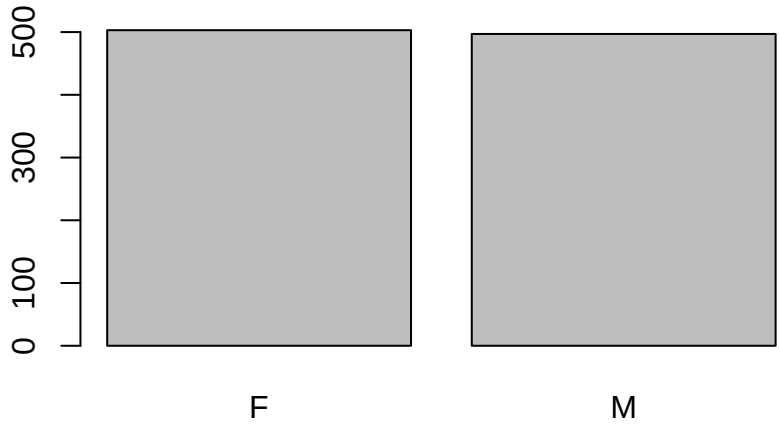
```
mean(dds.discr$ethnicity == "White" | dds.discr$ethnicity == "Hispanic")
```

```
## [1] 0.777
```

d) Does gender appear to be balanced?

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

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




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

```
##
##   F   M
## 503 497
```

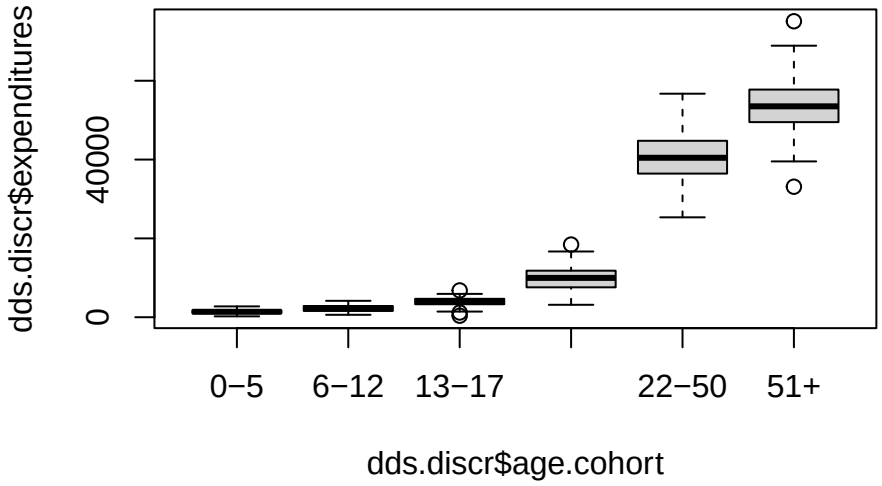
Problem 3: Bivariate Explorations (aka, relationships)

Explore how variables are related to each other.

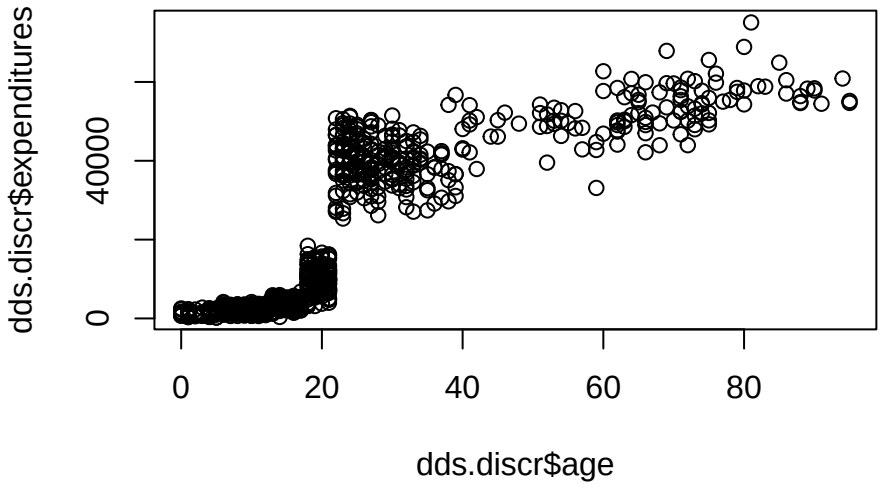
a) How do annual expenditures vary by age? Explore this using each of the quantitative version and categorical version of age separately. Conjecture a reason for the trend in the data.

There is a clear upward trend in expenditures as age increases; older individuals tend to receive more DDS funds. 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)
```




```
plot(dds.discr$expenditures ~ dds.discr$age)
```




```
#numerical summaries
tapply(dds.discr$expenditures, dds.discr$age.cohort, summary)
```

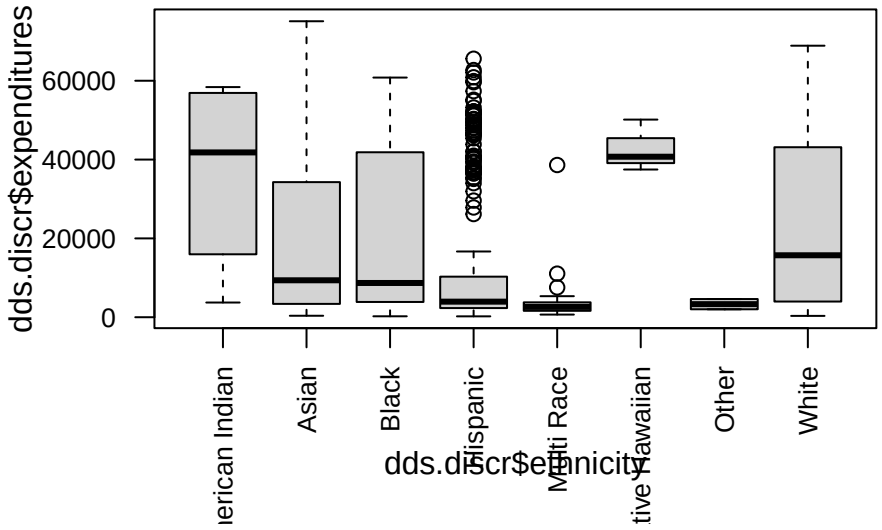
```
## $`0-5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     222    1034    1380    1415    1739    2750
##
## $`6-12`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     620    1602    2191    2227    2846    4163
##
## $`13-17`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     386    3306    3952    3923    4666    6798
##
## $`18-21`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3153    7588    9979    9889   11806   18435
##
## $`22-50`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25348   36447   40456   40209   44721   56716
##
## $`51+`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33110   49515   53509   53522   57746   75098
```

b) How does the distribution of expenditures vary by ethnic group?

The distribution is quite different between ethnic groups. For example, there is very little variation within the Multi Race, Native Hawaiian, and Other groups; there is a greater range in groups such as White not Hispanic. There also seems to be a difference in the amount of funding that a person receives, on average, between ethnic groups.

```
#graphical summaries
par(cex.axis = 0.8, las=2)
boxplot(dds.discr$expenditures ~ dds.discr$ethnicity)
```



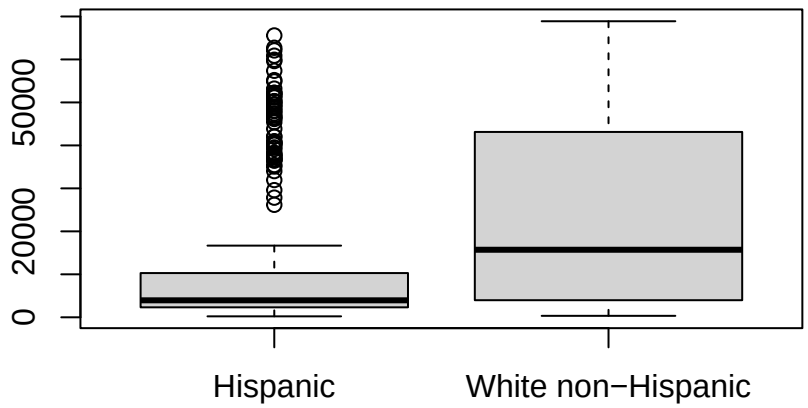
Problem 4: Exploring Evidence of Discrimination

Hispanic and White non-Hispanic individuals comprise the majority of the data. The rest of this analysis will focus on comparing how expenditures vary between these two groups.

a) 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. However, for White non-Hispanic consumers, the middle 50% of consumers receive between about \$4,000 and \$43,000. On average, a Hispanic consumer receives less financial support from the California DDS than a White non-Hispanic consumer.

```
#graphical summaries
boxplot(dds.discr$expenditures[dds.discr$ethnicity == "Hispanic"],
        dds.discr$expenditures[dds.discr$ethnicity == "White"],
        names=c("Hispanic", "White non-Hispanic"))
```




```
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"])
```

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

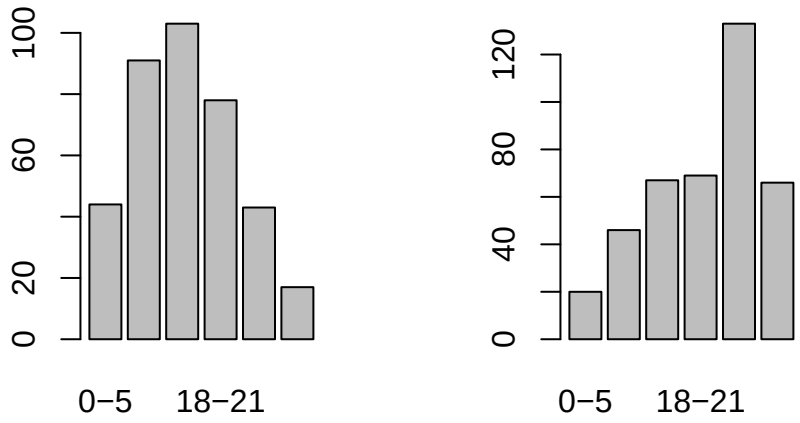
```
IQR(dds.discr$expenditures[dds.discr$ethnicity == "White"])
```

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

b) Recall that expenditures is strongly associated with age. Is there also an association between age and ethnicity, for these two ethnic groups?

Hispanics tend to be younger, with most Hispanic consumers falling into the youngest three age cohorts. In contrast, White non-Hispanics stend 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.

```
#graphical summaries
par(mfrow = c(1,2))
plot(dds.discr$age.cohort[dds.discr$ethnicity == "Hispanic"])
plot(dds.discr$age.cohort[dds.discr$ethnicity == "White"])
```




```
#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
```

```
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
```

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

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

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

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

c) For a closer look at the relationship between age, ethnicity, and expenditures, compare how average expenditures differs by ethnic group within each age cohort. Describe your findings.

When expenditures are compared within age cohorts, there are not large differences between mean expenditures for White non-Hispanics versus 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.

```
#first subset the data into two ethnicity groups
dds.hisp <- dds.discr[dds.discr$ethnicity == "Hispanic",]
dds.white <- dds.discr[dds.discr$ethnicity == "White",]

#calculate and compare means by group
hisp.means <- tapply(dds.hisp$expenditures, dds.hisp$age.cohort, mean)
white.means <- tapply(dds.white$expenditures, dds.white$age.cohort, mean)
data.frame(white.means, hisp.means, diff=white.means-hisp.means)
```

```
##       white.means hisp.means        diff
## 0-5       1366.900   1393.205   -26.30455
## 6-12      2052.261   2312.187  -259.92594
## 13-17     3904.358   3955.282   -50.92334
## 18-21    10133.058   9959.846   173.21182
## 22-50    40187.624  40924.116  -736.49222
## 51+      52670.424  55585.000 -2914.57576
```

d) Does there seem to be evidence of ethnic discrimination in the amount of financial support provided by the California DDS? Explain why the bivariate analysis conducted by the study team was misleading (bonus for remembering the specific term for the responsible phenomenon, which was covered in Stat 110!).

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—age is a confounder for the relationship between ethnicity and expenditures. 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.