Patel’s 02 DDS Case Study

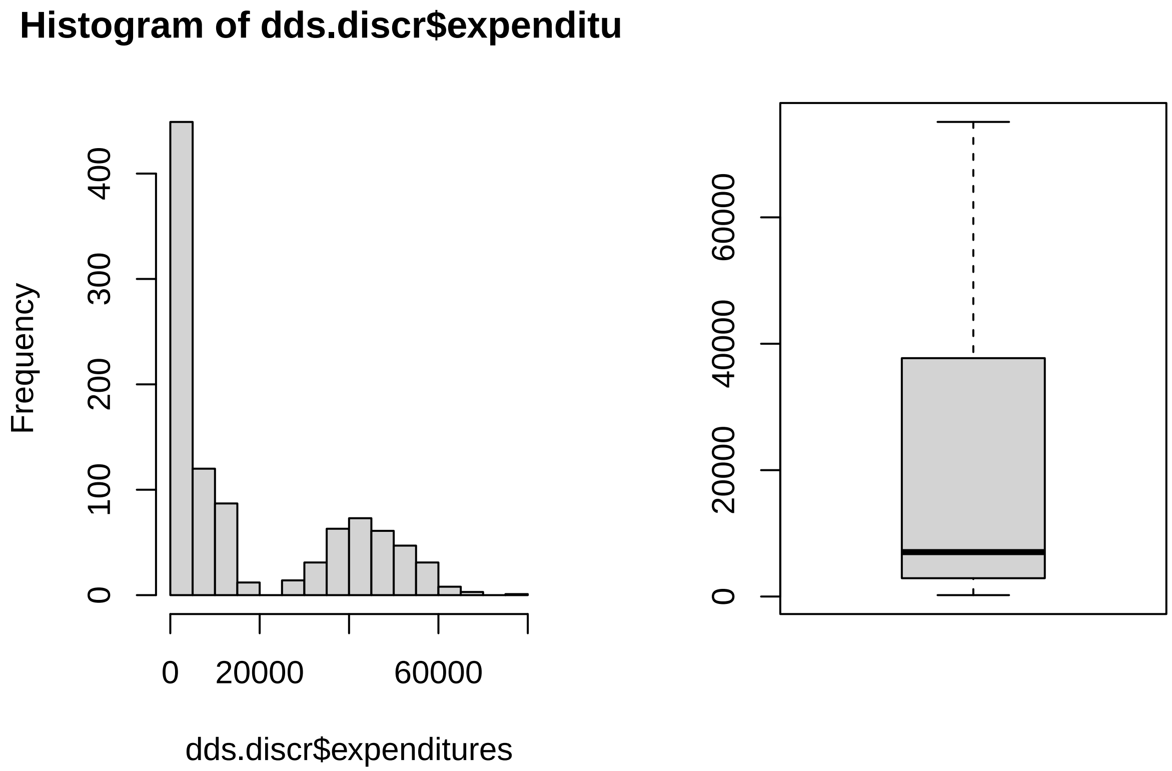
Harsh Patel

2024-09-06

```
# Distributions of single variables

#1.

#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
```

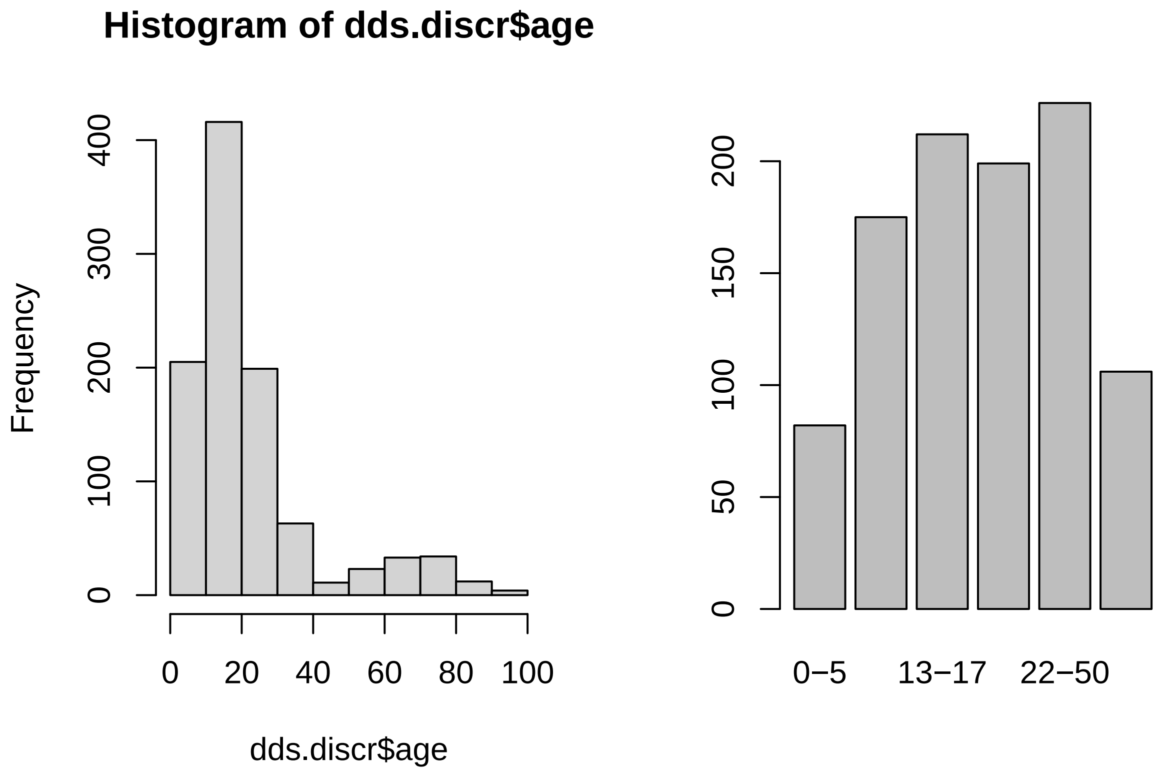
```
#2.

#a)
#The distribution of annual expenditures is right-skewed.
#With most consumers spending between $0 and $5,000.
#While a few spend $60,000 to $80,000. Quartiles 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)

#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
```

```
#The histogram shows right-skewing, with most consumers under 30 years old.
#The median age is 18, with around 200 people in the middle four age groups and about 100 in the other

#c)

#graphical summaries
plot(dds.discr$ethnicity)

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

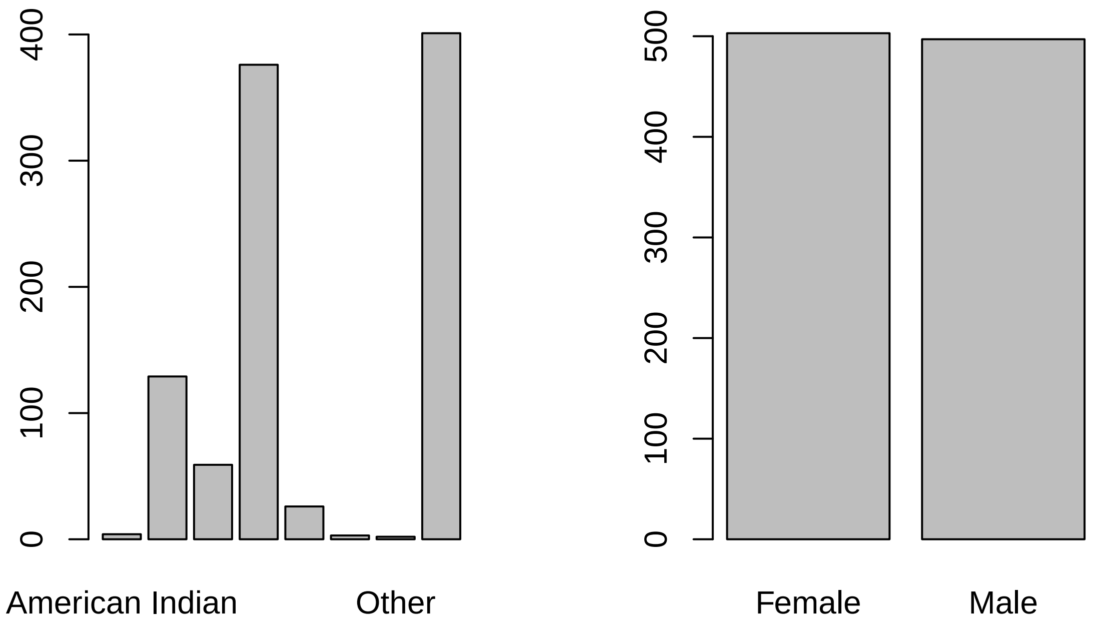
```
##
```

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

```
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
```

```
#The data includes eight ethnic groups, but with unequal representation.
#Hispanics and White non-Hispanics make up about 80% of the consumers.

#d)

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



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

```
##
```

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

```
#Yes, the gender distribution is roughly equal, with about half female and half male.

#Relationships between two variables

#3.

#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"])
```

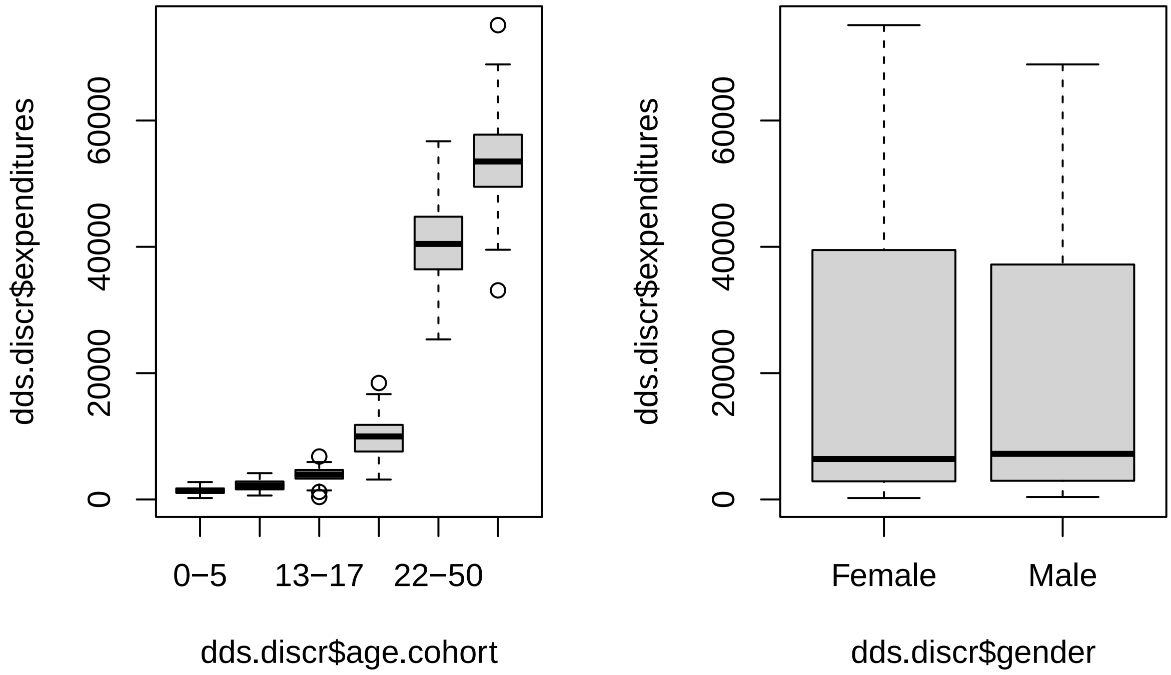
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25348   36447   40456   40209   44721   56716
```

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

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

```
#Expenditures rise with age, with older individuals receiving more DDS funds.
#Average expenditures range from $1,400 to $10,000.
#For the youngest cohorts and increase to about $40,000 and $53,500 for the oldest.
#The data’s broad age range explains the variation.
#A dataset limited to one age group, like 18-21 years, would show less variability.
#This trend aligns with the goal of DDS funds to support increasing financial needs as individuals age.
```

```r
#4.

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



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

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

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

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

```r
#Expenditures for both males and females are similarly right-skewed.
#With comparable medians and interquartile ranges.

#5.

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

```r
#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
```

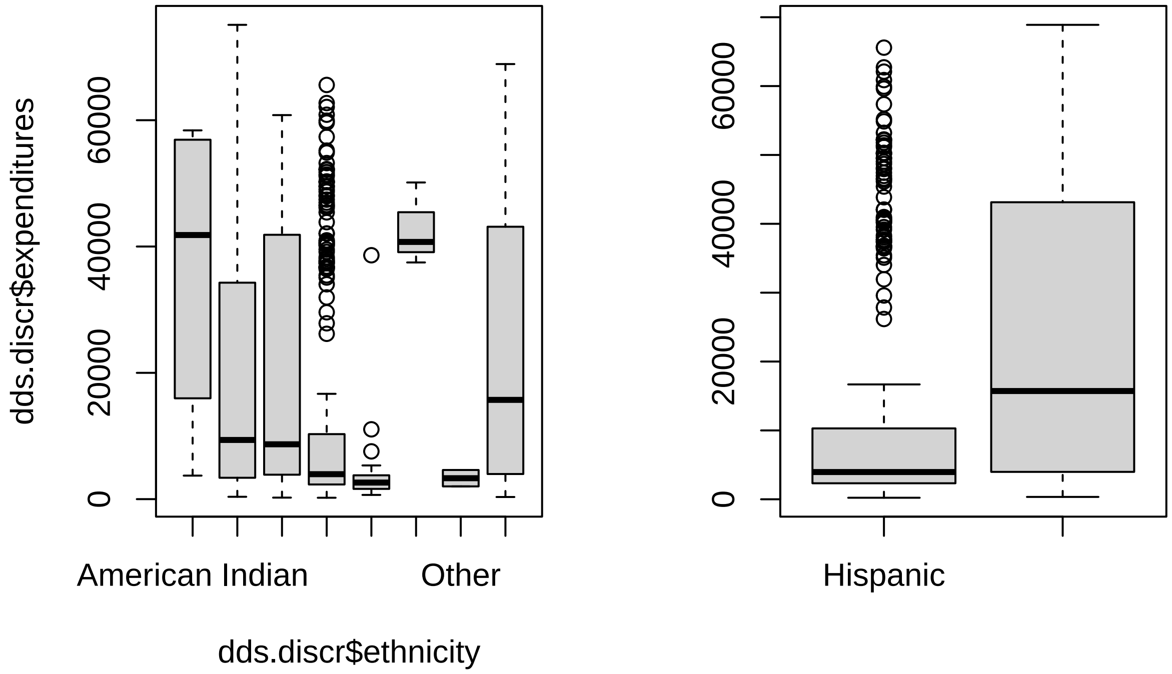
```r
summary(dds.discr$expenditures[dds.discr$ethnicity == "White not Hispanic"])
```

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

```r
#bonus:usingtapply( )
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
```

```
#Expenditure distributions vary by ethnicity. Multi Race, Native Hawaiian.
#And Other groups show little variation.
#While groups like White non-Hispanics have a wider range.
#American Indian and Native Hawaiian groups have a median annual support of about $40,000.
#Compared to $10,000 for Asian and Black consumers.
#The tapply() function can summarize these differences more efficiently than summary().

#A closer look

#6.

#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"])
```

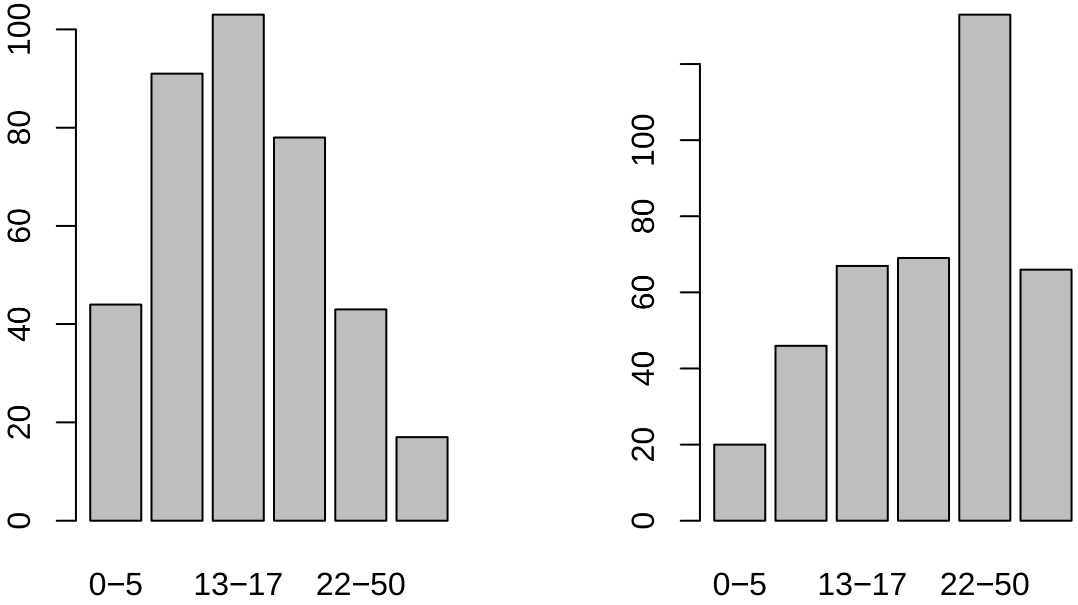
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     340    3977   15718   24698   43134   68890
```

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

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

```
#The boxplot shows that most Hispanic consumers receive between $0 and $20,000.
#From California DDS, with higher amounts being upper outliers.
#In contrast, White non-Hispanic consumers have a median expenditure of $15,718.
```

```
#With the middle 50% receiving between $4,000 and $43,000.
#Hispanic consumers average $11,066, while White non-Hispanics average $24,698.
#Indicating that Hispanics receive less financial support on average.

#7.

#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"])
```



```
#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
```

```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
```

```r
#Hispanics are generally younger, with most in the 6-12, 13-17, and 18-21 age groups.
#In contrast, White non-Hispanics are older, with the majority in the 22-50 age group.
#And a higher proportion in the 51+ group.

#8.

#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
```

```
#Within age cohorts, mean expenditures for White non-Hispanics and Hispanics are similar.
#This suggests that the initial observed difference in overall averages is less pronounced.
#When comparing individuals of the same age.

#9.

#There is no evidence of ethnic discrimination.
#Lower average expenditures for Hispanics are due to their younger age.
#Compared to White non-Hispanics, younger individuals typically receive less support.
#When comparing individuals of similar ages.
#Expenditure differences between Hispanics and White non-Hispanics are minimal.

#Simpson's paradox

#10.

#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
```