Problem Set #1 Answers

Name(s): _____

Q1

Use the "cars" dataset (openintro) to answer the following questions:

a) Create a histogram and boxplot (box and whiskers) for the 'mpgCity' variable (attach)

```
library(openintro)
```

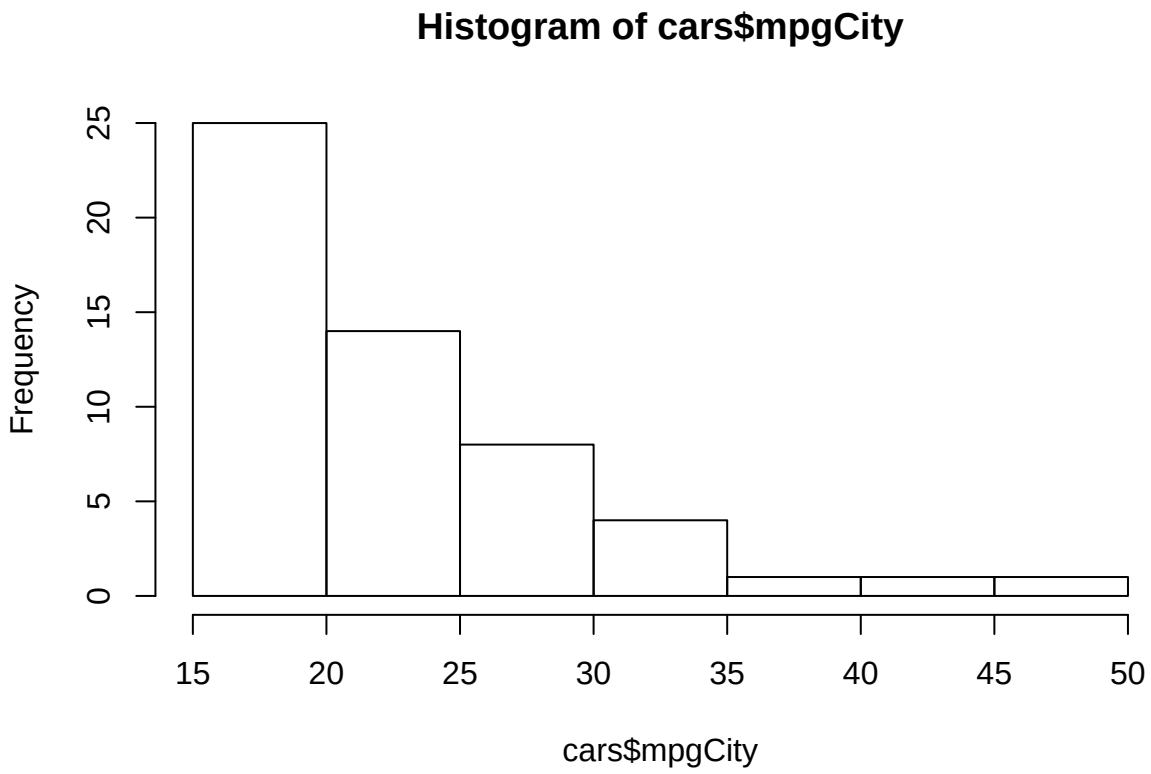
```
## Warning: package 'openintro' was built under R version 3.2.3
```

```
## Please visit openintro.org for free statistics materials
```

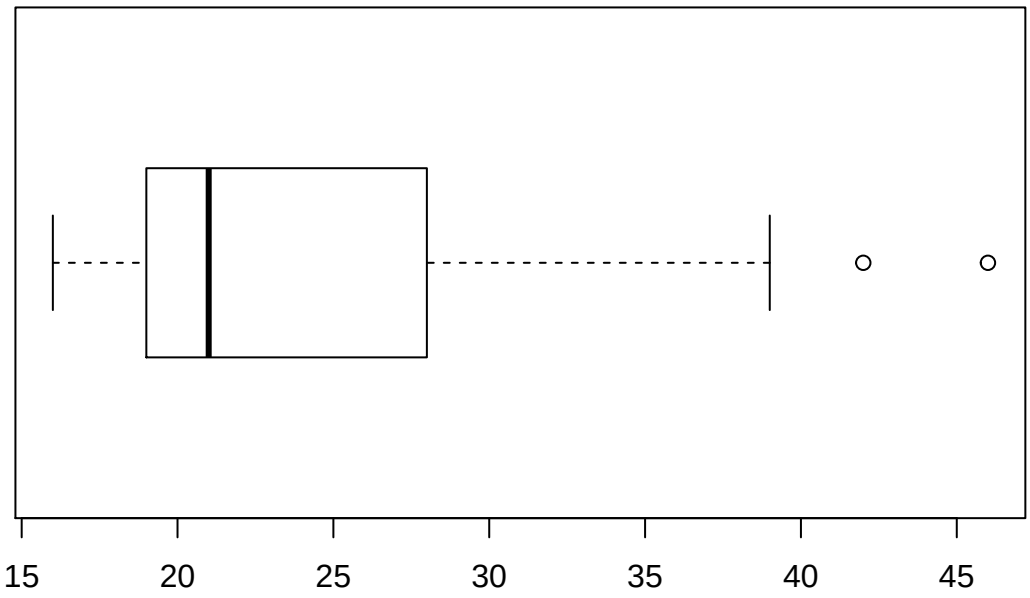
```
##
## Attaching package: 'openintro'
```

```
## The following object is masked from 'package:datasets':
##
##     cars
```

```
hist(cars$mpgCity)
```




```
boxplot(cars$mpgCity, horizontal = T)
```



b) Determine the skewness of the 'mpgCity' distribution (actual skewness value)

- You will need to do a library statement before using the "skewness" function - library(moments)

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 3.2.3
```

```
skewness(cars$mpgCity)
```

```
## [1] 1.450431
```

c) Calculate the mean and the median for the 'mpgCity' variable.

```
mean(cars$mpgCity)
```

```
## [1] 23.31481
```

```
median(cars$mpgCity)
```

```
## [1] 21
```

d) Does each of the previous tasks remain consistent in terms of skewness? Does the boxplot and histogram demonstrate the direction of skewness that the skewness measurement provides. Does the relationship between the mean and median support such skewness? Elaborate...

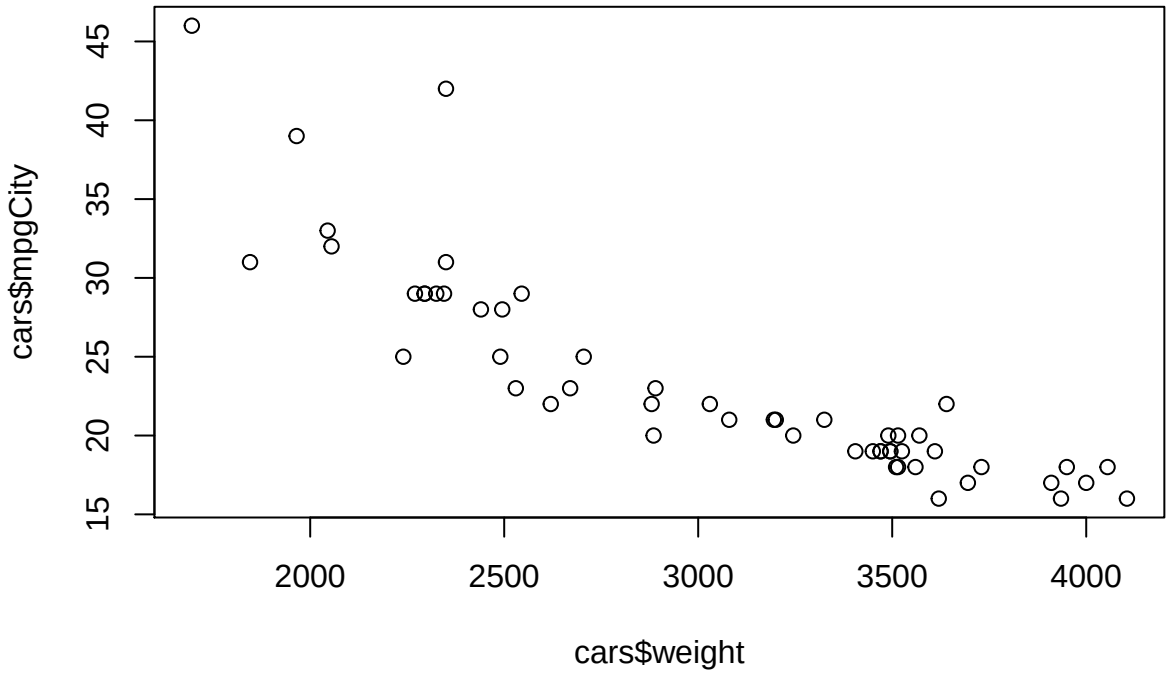
- Boxplot and histogram for 'mpgCity' exhibit positive skewness (right)
- The skewness calculation is 1.45. The sign of this calculation is *positive*, therefore the variable is positively skewed
- The mean is greater than the median. Another indication of positive skewness

Q2

Use the "cars" dataset (openintro) to answer the following questions:

a) Create a scatterplot for the 'weight' (x) and 'mpgCity' (y) variable. Use the "plot" function (attach)

```
plot(cars$weight, cars$mpgCity)
```



b) Determine the covariance and correlation measures between variables of interest in part (a)

```
cov(cars$weight, cars$mpgCity)
```

```
## [1] -3820.3
```

```
cor(cars$weight, cars$mpgCity)
```

```
## [1] -0.8769183
```

c) Explain the connection between the scatterplot and metrics calculated in part (b)

- The plot indicates a negative relationship between 'weight' and 'mpgCity'
- Such observed relationship is consistent with the covariance calculated. The covariance has a negative sign... proves a negative relationship
- The correlation is exhibits a strong negative relationship between variables. -0.8769 is close to -1... proving a strong negative connection

d) Does such findings in the previous tasks make "intuitive" sense? Elaborate...

- Such findings are consistent with what is expected. The heavier the car, the worse gas mileage in the city

Q3

Construct a box and whiskers plot with the following data (x):

```
x <- c(1, 3, 4, 5, 7, 4, 5, 4, 2, 9, 10, 4, 3, 5, 11, 1, 3, 3, 5, 4)
```

You must provide the R code or mathematical approach for each of the following questions (a-f) in your response. Just do not provide the answer!

a) What is the median?

```
median(x)
```

```
## [1] 4
```

b) What is Q3 (75% Percentile)?

```
Q3 <- quantile(x, probs = c(0.75))
Q3
```

```
## 75%
##   5
```

c) What is Q1 (25% Percentile)?

```
Q1 <- quantile(x, probs = c(0.25))
Q1
```

```
## 25%
##   3
```

d) What is the IQR of the data set?

```
iqr <- Q3 - Q1
```

e) What is the lower fence?

```
lf <- Q1 - 1.5*iqr
lf
```

```
## 25%
##   0
```

f) What is the upper fence?

```
uf <- Q3 + 1.5*iqr
uf
```

```
## 75%
##   8
```

g) How many "extreme" values are present in the data set?

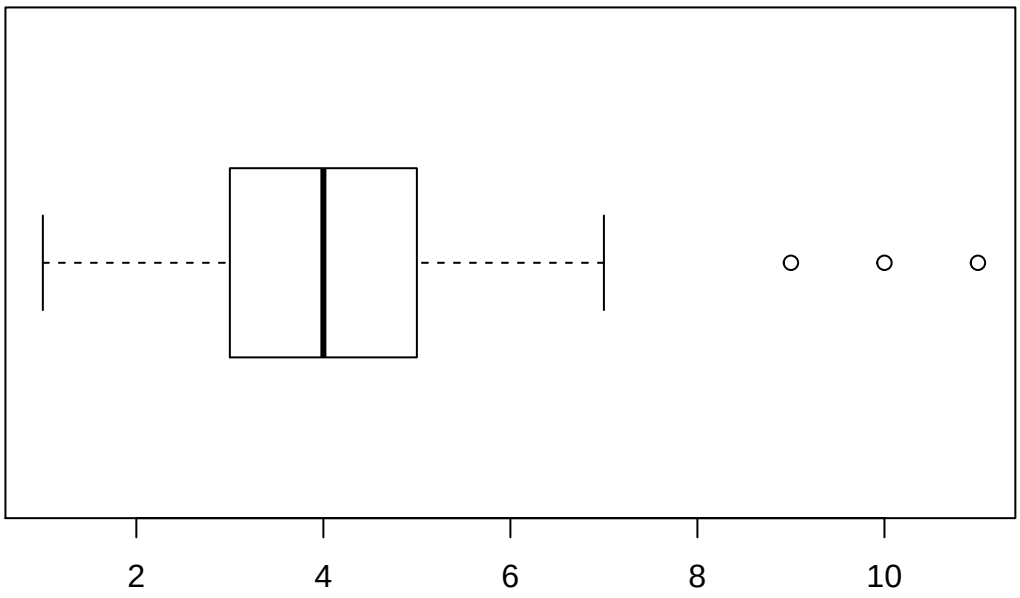
```
x > uf | x < lf
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
## [12] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

- Therefore, 3 extreme values

h) Generate the box and whiskers plot using the "boxplot" R function (attach). Ensure the boxplot created is consistent with the results derived in the previous tasks.

```
boxplot(x, horizontal = T)
```



- The results calculated seems to be the same as the boxplot created. It shows the three extreme observations outside the upper fence

Q4

Use the "cars" dataset (openintro) to answer the following questions:

You must provide the R code for each of the following questions (a-c) in your response.

a) How many 'midsize' (count) cars are in the data set "cars"? Use the "table" function with the variable of interest, 'type'

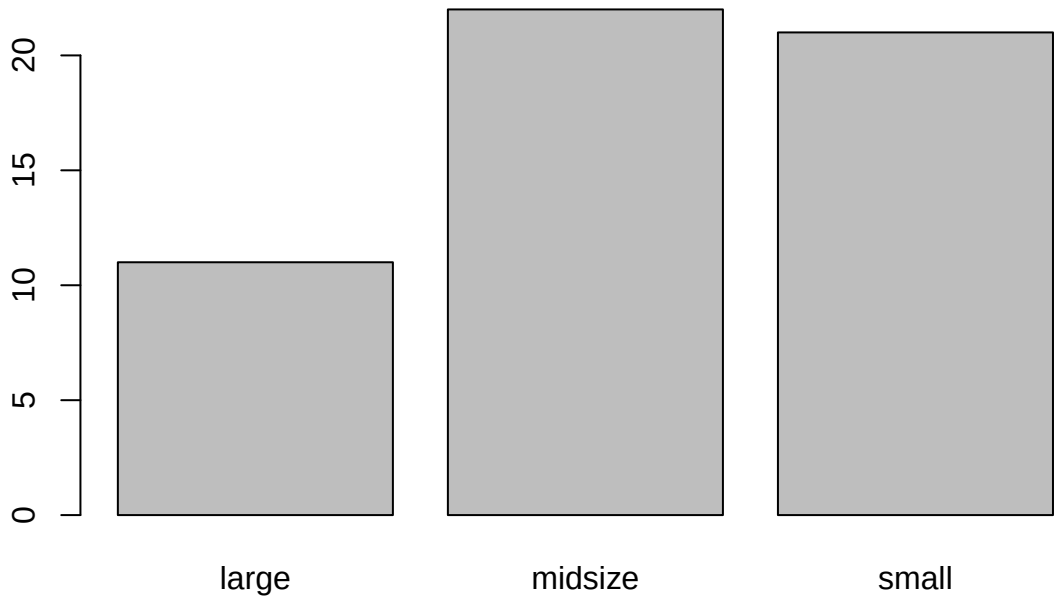
```
table(cars$type)
```

```
##
##   large midsize   small
##      11      22      21
```

- 22 Midsize Cars

b) Create a barplot to confirm your answer in part (a) and attach

```
barplot(table(cars$type))
```



c) Build a contingency table with the variables 'type' and 'driveTrain'. How many front wheel drive small cars exist in the data set?

```
table(cars$type, cars$driveTrain)
```

```
##
##           4WD front rear
##   large     0     7    4
##   midsize   0    17    5
##   small     2    19    0
```

- 19 front wheel drive small cars

Q5

Use the "beaver1" dataset (datasets) to answer the following questions:

Run the following code to generate two samples of 'temp' from the "beaver1" dataset. Each will be assigned a new name, 'temp1' and 'temp2'. Each sample will consist of 15 observations.

```
set.seed(1986)
temp1 <- sample(beaver1$temp, 15)
temp2 <- sample(beaver1$temp, 15)
```

You must provide the R code for each of the following questions (a-b) in your response.

a) Calculate the sample variance, sample stadard deviation, sample IQR for each sample

```r
var(temp1)
```

```
## [1] 0.04898095
```

```r
var(temp2)
```

```
## [1] 0.01752667
```

```r
sd(temp1)
```

```
## [1] 0.2213164
```

```r
sd(temp2)
```

```
## [1] 0.1323883
```

```r
IQR(temp1)
```

```
## [1] 0.175
```

```r
IQR(temp2)
```

```
## [1] 0.12
```

b) Since both samples were taken from the same population, compare your findings for each measure of spread between samples. Elaborate...

- The spread of data for the sample, temp2 is less than that of the sample, temp1. This is evident in all measures for spread, variance, standard deviation, and interquartile range. This proves that regardless the measure of spread you use, given the samples can be compared (coming from the same population), you should get an accurate comparison.