W203 - Statistics for Data Science - Unit 1 Homework

Key

Exercises

Load the dataset found in the file, cars.csv.

Use the read.csv() function to read in the data and head() function to view first 5 rows

```
Cars = read.csv("cars.csv")
head(Cars)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

1. What are the variables in the file?

Use objects() function to return the variable names.

```
vars = objects(Cars)
vars
```

```
##  [1] "am"   "carb" "cyl"  "disp" "drat" "gear" "hp"   "mpg"  "qsec" "vs"
## [11] "wt"
```

2. Find the mean, median, minimum, maximum, 1st quartile and 3rd quartile for the mpg variable.

Get mpg from Cars using $ and use the various inbuilt functions.

```
mpg = Cars$mpg
mpg
```

```
##  [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2
## [15] 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2
```

```
mean(mpg)
```

```
## [1] 19.492
```

```
median(mpg)
```

```
## [1] 18.7
```

```
min(mpg)
```

```
## [1] 10.4
```

```
max(mpg)
```

```
## [1] 33.9
```

```
quantile(mpg)[2]
```

```
##  25%
## 15.2
```

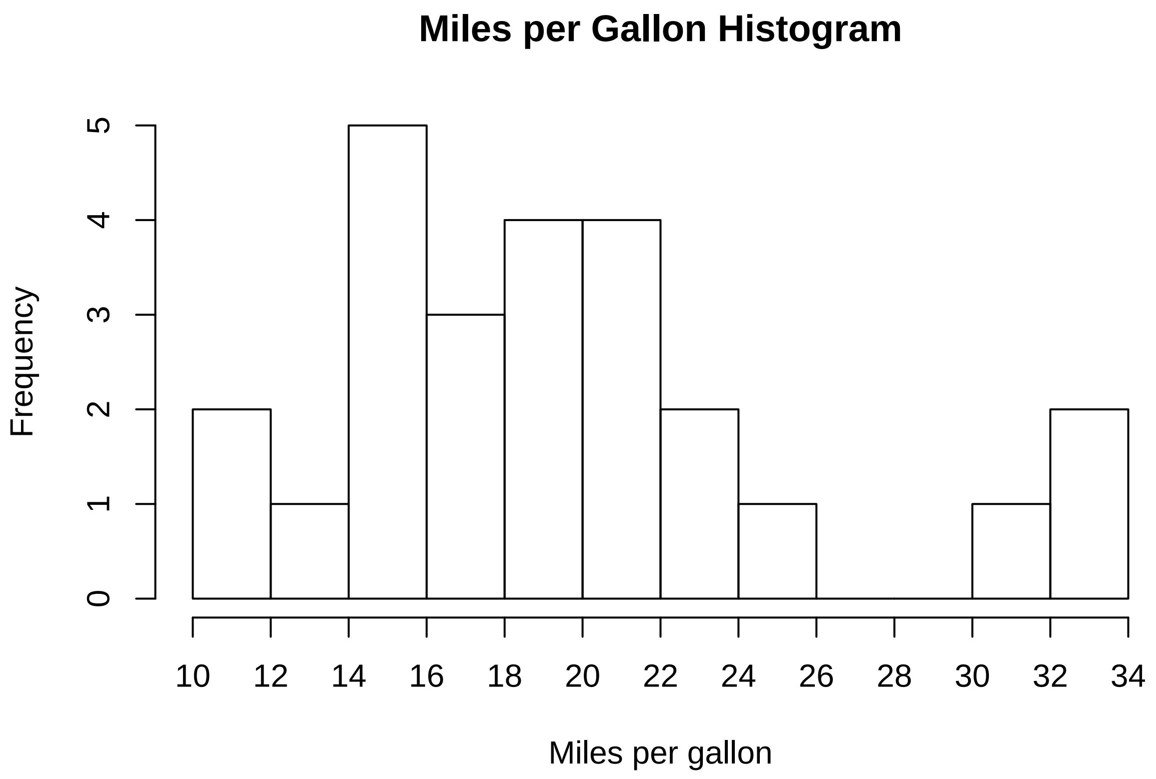
```
quantile(mpg)[4]
```

```
##  75%
## 21.5
```

3. Create a histogram of the mpg variable.

Use default hist() function to plot a histogram of the mpg

```
hist(mpg, breaks = 12, xlab = "Miles per gallon", xaxt='n',
     main = "Miles per Gallon Histogram")
axis(1, at = seq(10, 34, 2))
```



4. What is the standard deviation of mpg variable?

Use sd() function to calculate sample standard deviation.

```
sd(mpg)
```

```
## [1] 6.047446
```

5. What is the variance of mpg variable?

Use var() function to calculate sample variance.

```
var(mpg)
```

```
## [1] 36.5716
```

6. What is the relationship of the standard deviation to the variance? Why does the standard deviation and variance of the mpg variable differ?

Standard deviation (s) is equal to the square root of the variance (s ** 2), so they should differ from each other unless s = 0 or 1. Show that this relationship holds true:

```
x1 = var(mpg)
x1
```

```
## [1] 36.5716
```

```
x2 = sd(mpg) ** 2
x2
```

```
## [1] 36.5716
```

```
x1 == x2
```

```
## [1] TRUE
```

7. How many data points are there for the cyl variable?

Get cyl from Cars using $ and print. Number of data points is obtained using length() function

```
cyl = Cars$cyl
cyl
```

```
##  [1]  6  6  4  6  8  6  8  4  4  6  6  8  8  8  8  8  8  4  4  4  4  8  8
## [24] NA NA
```

```
length(cyl)
```

```
## [1] 25
```

However, there are two null values in cyl, so remove and recalculate

```
cyl_ex_na = na.omit(cyl)
cyl_ex_na
```

```
##  [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 8 4 4 4 4 8 8
## attr(,"na.action")
## [1] 24 25
## attr(,"class")
## [1] "omit"
```

```
length(cyl_ex_na)
```

```
## [1] 23
```

So the length of cyl allowing for nulls is 23.

An alternate way of removing the NA's is to create a boolean filter based on cyl whether each element is NA or not and reverse the filter.

```
cyl_ex_na1 = cyl[!is.na(cyl)]
cyl_ex_na1
```

```
##  [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 8 4 4 4 4 8 8
```

8. What is the mean of the cyl variable?

Showing the risk of working directly from cyl:

```
wrong_mean = sum(cyl, na.rm = TRUE) / length(cyl)
correct_mean = mean(cyl, na.rm = TRUE)
wrong_mean
```

```
## [1] 5.76
```

```
correct_mean
```

```
## [1] 6.26087
```

Better to remove NAs from cyl to start with, i.e. use cyl_ex_na

```
correct_mean1 = sum(cyl_ex_na) / length(cyl_ex_na)
correct_mean2 = mean(cyl_ex_na)
correct_mean1
```

```
## [1] 6.26087
```

```
correct_mean2
```

```
## [1] 6.26087
```