Homework Assignment 2

Parker DeBruyne - V00837207

09/02/2022

Stat 123 Homework Assignment 2 Due Friday February 18th by 9:00pm Using R Markdown, please complete the following assignment. If an answer does not require any R code, you can type the answer to the question outside of a chunk. Make sure that your assignment is well labelled so that it is clear where each question's answer begins. Your assignment should be submitted as a pdf (whether you knit directly to PDF, or knit to HTML or Word and then convert the file to a pdf).

1. The built-in Titanic data set is a 4-dimensional array that contains the following information:

• Dimension 1: Class of the passenger (1 = 1st, 2 = 2nd, 3 = 3rd, 4 = Crew member) • Dimension 2: Sex of the passenger (1 = male, 2 = female) • Dimension 3: Age of the passenger (1 = child, 2 = adult) • Dimension 4: Survival of the passenger (1 = died, 2 = survived) If you wanted to determine, for example, how many male, adult, crew members survived, you could type in Titanic[4,1,2,2] to get this value. If you wanted to create a table with how many 1st class passengers (of all genders and ages) died, you could type Titanic[1, , ,1].

(a) Create (and print out) a table which contains the adult passengers (of all classes and genders) who survived.

```
Titanic_adult_survivors = as.table(Titanic[,,2,2])
Titanic_adult_survivors
```

```
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

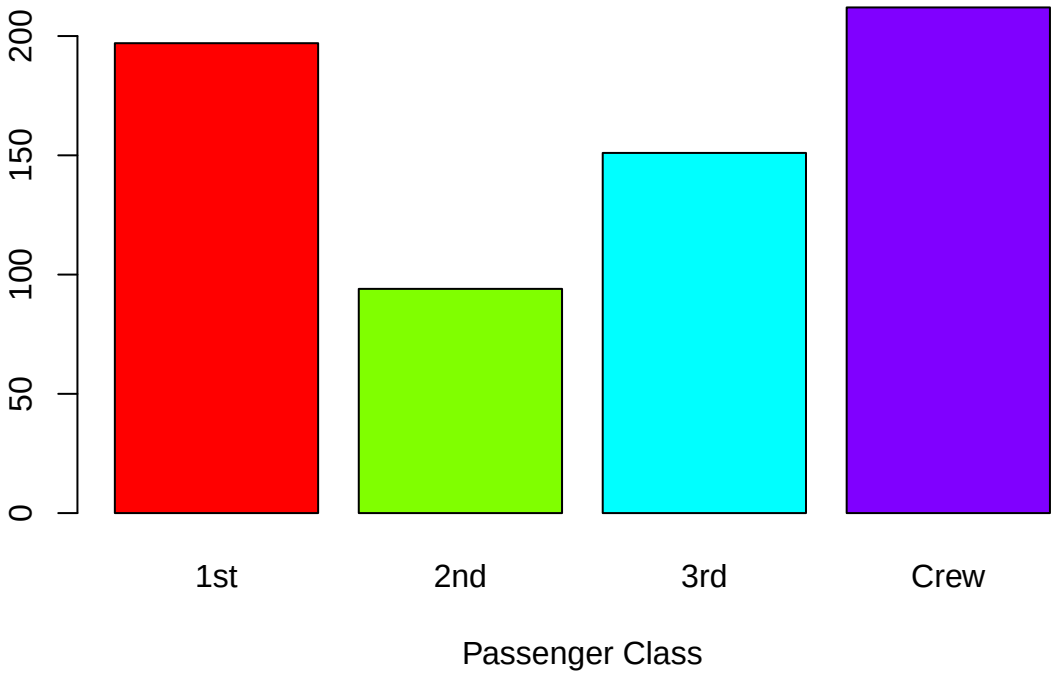
(b) Create (and print out) a vector called survived which contains the adult passengers who survived. Hint: You may need to use rowSums() on your answer from part (a).

```
survived = rowSums(Titanic_adult_survivors)
print(survived)
```

```
##  1st  2nd  3rd Crew
##  197   94  151  212
```

(c) Create a barplot displaying the survived vector. Make sure to include a main title and to label your x-axis. Also, make sure that each bar is a different colour.

```
barplot(survived, main="Adult Survivors vs. Passenger Class", xlab="Passenger Class", col=rainbow(lengt
```



(d) What does the bar graph imply about the survival of adult passengers based on class?

That most of the survivors were from either 1st class or Crew.

(e) Create (and print out) a vector called died which contains the adult passengers who did not survive.

```
Titanic_adult_survivors = as.table(Titanic[,,2,1])
died = rowSums(Titanic_adult_survivors)
print(died)
```

```
##  1st  2nd  3rd Crew
##  122  167  476  673
```

(f) Create (and print out) a vector called percent.Survived which contains the percentage of adult passengers who survived in each class.

```
total = sum(survived)

percent.Survived = round(((survived / total) * 100), 0)
print(percent.Survived)
```

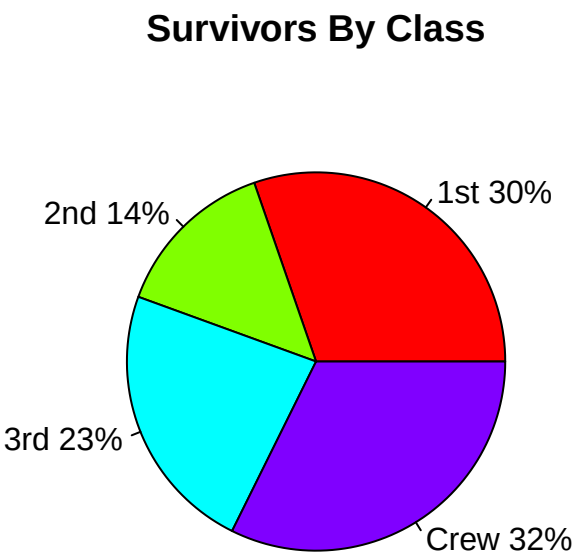
```
##  1st  2nd  3rd Crew
##   30   14   23   32
```

```
percent.Survived.ch = paste(percent.Survived, "%", sep="")

percent.Survived.ch = paste(names(percent.Survived), percent.Survived.ch)

percent.Survived.ch
```

```
## [1] "1st 30%"  "2nd 14%"  "3rd 23%"  "Crew 32%"
```

(g) Create a pie chart that displays the percent.Survived data. Be sure to include a main title for your pie chart.

```
pie(percent.Survived, main="Survivors By Class", labels=percent.Survived.ch, col=rainbow(length(percent
```



(h) What does the pie chart imply about the survival of the adult passengers based on class? Does this imply something different than the bar graph did? If yes, why?

It implies that 1st class passengers were most likely to survive. In the bar graph it looks like the Crew and 1st Class had equal survivors; this is because the total survivors from each were almost equal, but the total Crew passengers was greater than total 1st Class passengers, leading to better odds for the 1st class.

2. The following question deals with the data set NHLData.csv which you will need to download from the assignment page.

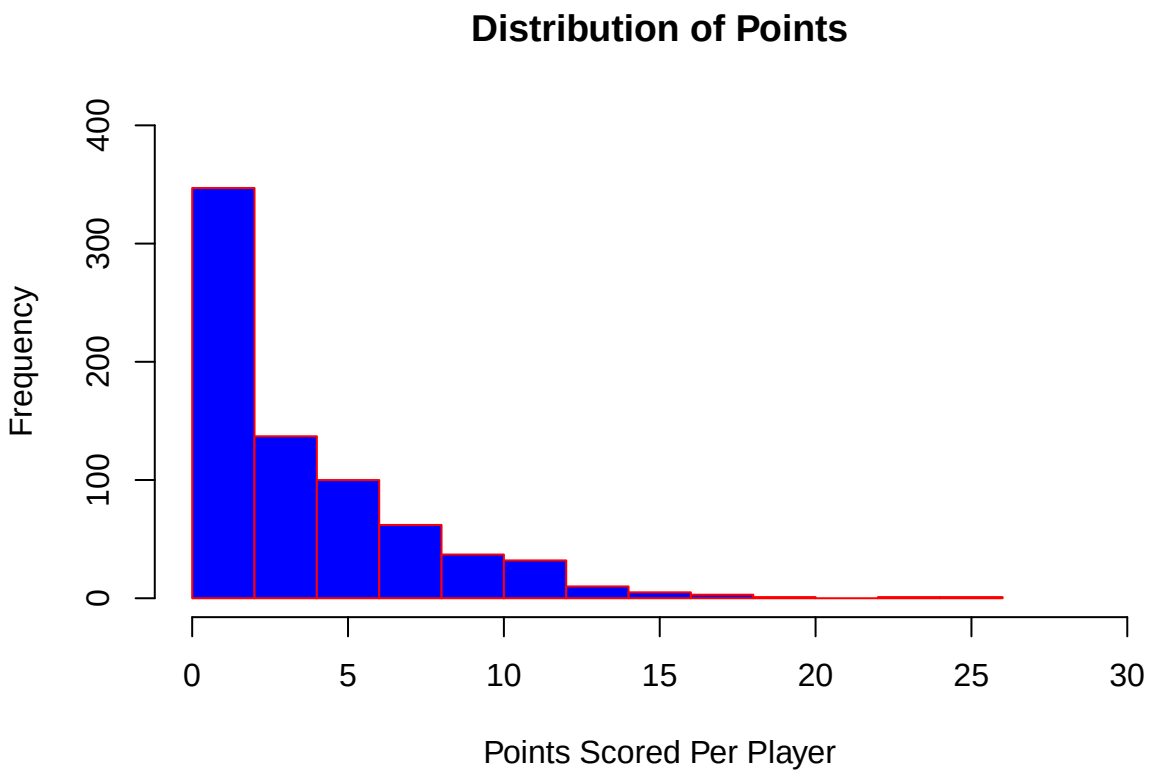
```
NHLData = read.csv("NHLData.csv")
```

(a) Create (but do not print) a vector called points containing the number of points for each player (the variable P in the data set).

```
players = NHLData$Player
points = NHLData$P
names(points) = players
```

(b) Create a histogram displaying the distribution of this variable. Be sure to have both a main title and a title on your x-axis. Also, make sure that the scale on the x axis goes to 30 and the scale on the y axis goes to 400.

```
hist(points, main="Distribution of Points", xlab="Points Scored Per Player", xlim=c(0,30), ylim=c(0,400
```



(c) Describe the shape of the distribution (symmetric, left-skewed, right-skewed).

The shape of this distribution is left-skewed.

(d) What is an appropriate measure of the center of the distribution (mean or median), why?

The appropriate measure of the distribution is the median. Using the median allows us to compare outliers to the rest of the data.

(e) Compute the appropriate center value and the corresponding measures of variability.

```
med_str = paste("Median: ", as.character(median(points), sep=""))
print(med_str)
```

```
## [1] "Median:  3"
```

```
var_str = paste("Variance: ", as.character(round(var(points), digits=2), sep=""))
print(var_str)
```

```
## [1] "Variance:  14.78"
```

```
sd_str = paste("Standard Deviation: ", as.character(round(sd(points), digits=2), sep=""))
print(sd_str)
```

```
## [1] "Standard Deviation:  3.84"
```

```
rng_str = paste("Range: ", as.character(round(range(points), digits=2), sep=""))
print(rng_str)
```

```
## [1] "Range:  0"  "Range:  26"
```

```
IQR_str = paste("Inter Quartile Range: ", as.character(round(IQR(points), digits=2), sep=""))
print(IQR_str)
```

```
## [1] "Inter Quartile Range:  5"
```

```
summary_str = paste("Summary: ")
print(summary_str)
```

```
## [1] "Summary: "
```

```
print(summary(points))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   3.826   6.000  26.000
```

3. (You may wish to do this question by hand) Consider the following sample of points from the NHL data set:

```
19
20  sample(nhl$P,20)
21
22 ```
```

```
 [1]  3  1  8  5  3  1  2  0  5  2  0  1  3  3  2  1 11  0
[19]  6  1
```

(a) Create a stemplot of the distribution of the sample.

```
0 | 0001111112223333
0 | 5568
1 | 1
1 |
```

(b) Does the distribution resemble the one seen in question 2? Explain why there might be some differences.

Yes, it does. There may be some differences however because this is only a sample of the population while the histogram includes all individuals.