Lab 03 - Categorical Data - Key

PSC-012Y

Jonathan J. Park

10/17/2024

Points

This assignment has 15 total questions thus 15 points and one bonus.

Part 1.

1. Read in both the `BirthTab.RDS` and `BirthDF.RDS` files using the `readRDS()` function and save them as objects called `btab` and `bdf`, respectively

```
btab = readRDS("C:/Users/MyComputer/Desktop/BirthTab.RDS")
bdf = readRDS("C:/Users/MyComputer/Desktop/BirthDF.RDS")
```

2. Use the `print()` function to view the `btab` variable you just created

```
print(btab)
```

```
##                Jan      Feb      Mar      Apr      May      Jun      Jul
## Frequency 69.79322 65.11697 71.47368 68.50733 71.75367 71.22586 75.27284
##                Aug      Sep      Oct      Nov      Dec
## Frequency 76.24626 74.40659 72.93822 68.84173 71.55101
```

3. Use the `str()` function on the variable `bdf$Month`

```
str(bdf$Month)
```

```
##  int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
```

4. Is `bdf$Month` coded as a categorical variable?

No.

5. Use the `factor()` function to turn `bdf$Month` into a factor variable:

```
bdf$Month = factor(bdf$Month,
                   levels = c(1:12),
                   labels = c("Jan", "Feb", "Mar", "Apr",
                              "May", "Jun", "Jul", "Aug",
                              "Sep", "Oct", "Nov", "Dec"),
                   ordered = FALSE)
```

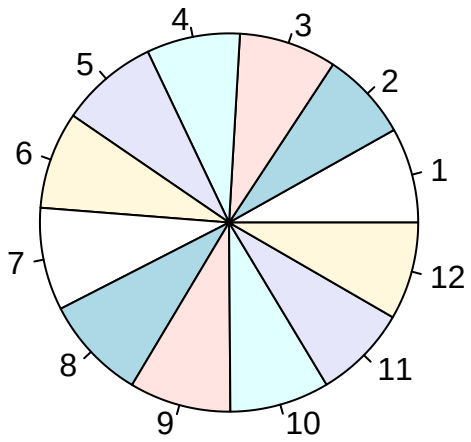
6. Check that your recoding worked by using `print()` to view `bdf`

```
print(bdf)
```

```
##    Month Frequency
## 1    Jan  69.79322
## 2    Feb  65.11697
## 3    Mar  71.47368
## 4    Apr  68.50733
## 5    May  71.75367
## 6    Jun  71.22586
## 7    Jul  75.27284
## 8    Aug  76.24626
## 9    Sep  74.40659
## 10   Oct  72.93822
## 11   Nov  68.84173
## 12   Dec  71.55101
```

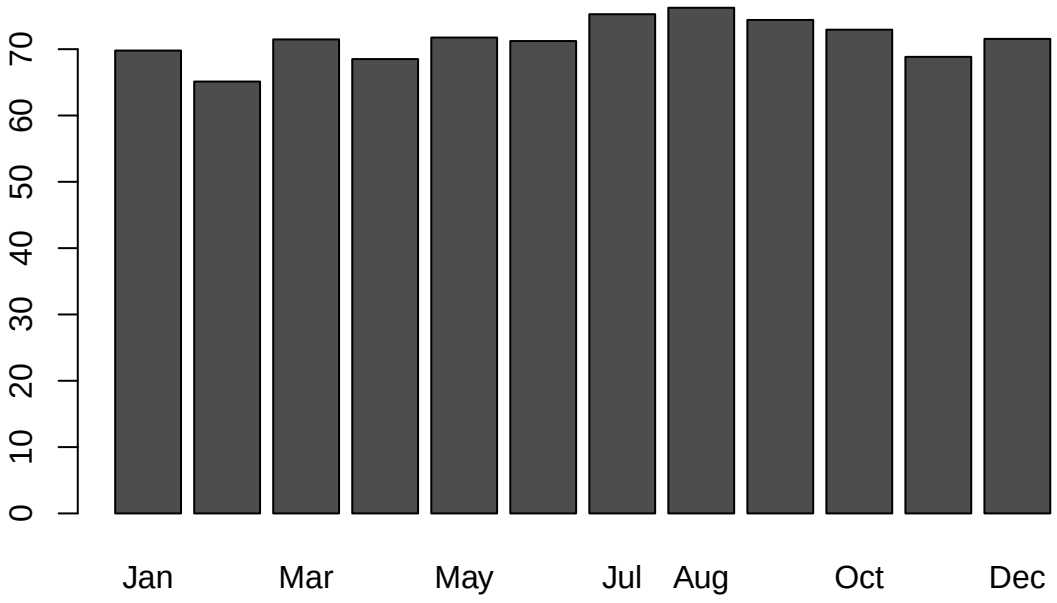
7. Create a pie chart using the `pie()` function using the `btab` data

```
pie(btab)
```



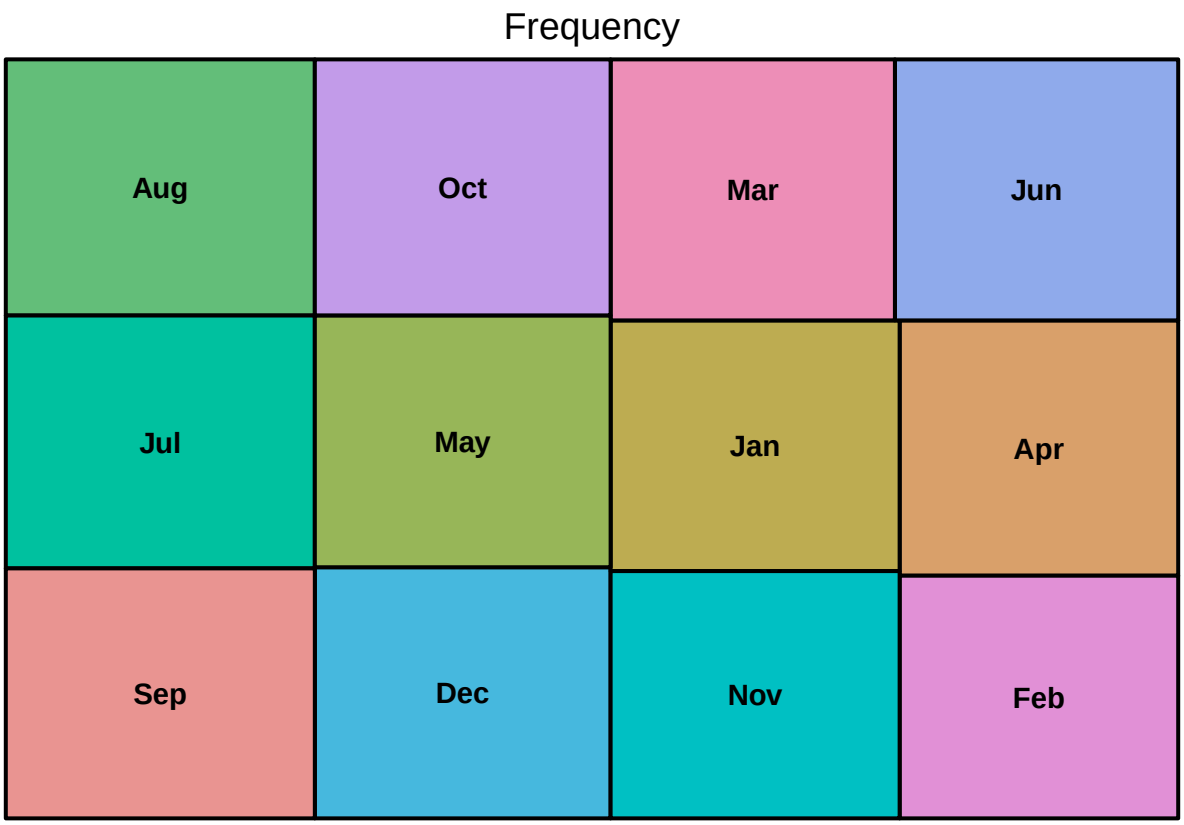
8. Create a bar chart using the `barplot()` function using the `btab` data

```
barplot(btab)
```



9. Create a tree map using the `treemap()` function in `R` using the `bdf` data

```
treemap(bdf, index = "Month", vSize = "Frequency")
```



10. Which visualization of the data is clearest in your opinion? Why?

Personally, I feel like the bar chart will be the clearest. If their justification is decent for another plot do give them credit anyway.

11. Use the function `chisq.test()` on the `btab` data and evaluate whether we have sufficient evidence to conclude that births occured more often in any one month than we would expect by chance (i.e., equally distributed across the 12-months of the year)

```
chisq.test(btab)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  btab
## X-squared = 1.4985, df = 11, p-value = 0.9996
```

There's not enough evidence to suggest any one month diverged from any other months, $\chi^2(11) = 1.49, p = 0.99$

Part 2. Course Knowledge

1. Which visualization–generally–would be better with many categories?

Bar Charts/Plots

2. Provide a research hypothesis relating two binary variables to one another:

_____ relates to greater/lower _____; your discretion

3. What is the defining difference between a nominal and ordinal variable?

Ordinal variables have a ranking. Nominal do not.

4. What does the "average" of a binary variable tell us?

Percentages or proportions

5. For nominal data, which metric would be best: Mean, Median, or Mode

Mode

6. The χ^2 goodness-of-fit is a statistical tool which helps us establish confidence in what?

Whether an observed result differs from what we expected