Lab 04 - Multivariate Categorical Data - Key

PSC-012Y

Jonathan J. Park

10/24/2024

Part 1. Coding and Visualization

1. Read in the `Dating.csv` Dataset and assign it into the object, `date`. Use the `head()` function to display the first 6-rows of the data

```
date = read.csv("Dating.csv")
head(date)
```

```
##        Age    Frequency                   Goal            AppLength
## 1 24 to 34   Frequently Long-term relationship           1-6 months
## 2 24 to 34   Frequently Long-term relationship     more than 1 year
## 3 24 to 34       Rarely Long-term relationship     more than 1 year
## 4 24 to 34       Rarely       I don't know yet     more than 1 year
## 5 18 to 24 Occasionally Long-term relationship approximately 1 year
## 6 18 to 24       Rarely Long-term relationship           1-6 months
##     BeenGhosted
## 1          Both
## 2          Both
## 3 Being ghosted
## 4 Being ghosted
## 5          Both
## 6          Both
```

2. Turn the following variables into factors with the `factor()` function. Decide if they should be ordered or not and use the `str()` function on `date` to show your work.

- Age ("24 to 34", "18 to 24")
- Frequency ("Frequently", "Rarely", "Occasionally")
- Goal ("Long-term relationship", "I don't know yet", "Short-term or casual relationship")
- AppLength ("1-6 months", "more than 1 year", "approximately 1 year")

```
# Nominal
  date$Goal = factor(date$Goal,
                     levels = c("Long-term relationship",
                                "I don't know yet",
                                "Short-term or casual relationship"),
                     ordered = FALSE)
# Ordinal
  date$Age = factor(date$Age,
                    levels = c("24 to 34",
```

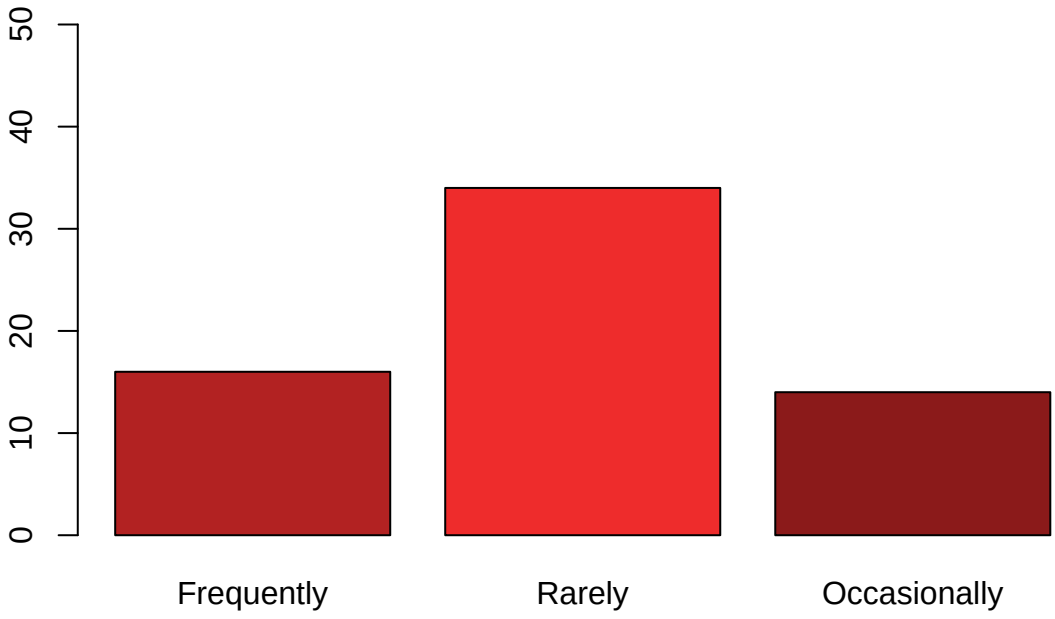
```
                                "18 to 24"),
                     ordered = TRUE)
  date$Frequency = factor(date$Frequency,
                          levels = c("Frequently",
                                     "Rarely",
                                     "Occasionally"),
                          ordered = TRUE)
  date$AppLength = factor(date$AppLength,
                          levels = c("1-6 months",
                                     "approximately 1 year",
                                     "more than 1 year"),
                          ordered = TRUE)

# Showing Proof:
  str(date)
```

```
## 'data.frame':    64 obs. of  5 variables:
##  $ Age        : Ord.factor w/ 2 levels "24 to 34"<"18 to 24": 1 1 1 1 2 2 1 1 1 1 ...
##  $ Frequency  : Ord.factor w/ 3 levels "Frequently"<"Rarely"<..: 1 1 2 2 3 2 3 2 1 2 ...
##  $ Goal       : Factor w/ 3 levels "Long-term relationship",..: 1 1 1 2 1 1 1 2 3 1 ...
##  $ AppLength  : Ord.factor w/ 3 levels "1-6 months"<"approximately 1 year"<..: 1 3 3 3 2 1 3 1 3 3 .
##  $ BeenGhosted: chr  "Both" "Both" "Being ghosted" "Being ghosted" ...
```

3. Create a univariate visualization of your choice to show how often people in our sample use dating apps (`date$Frequency`)

Hint: This may require that you create a `table()`

```
barplot(table(date$Frequency),
        col = c("firebrick",
                "firebrick2",
                "firebrick4"),
        ylim = c(0, 50))
```



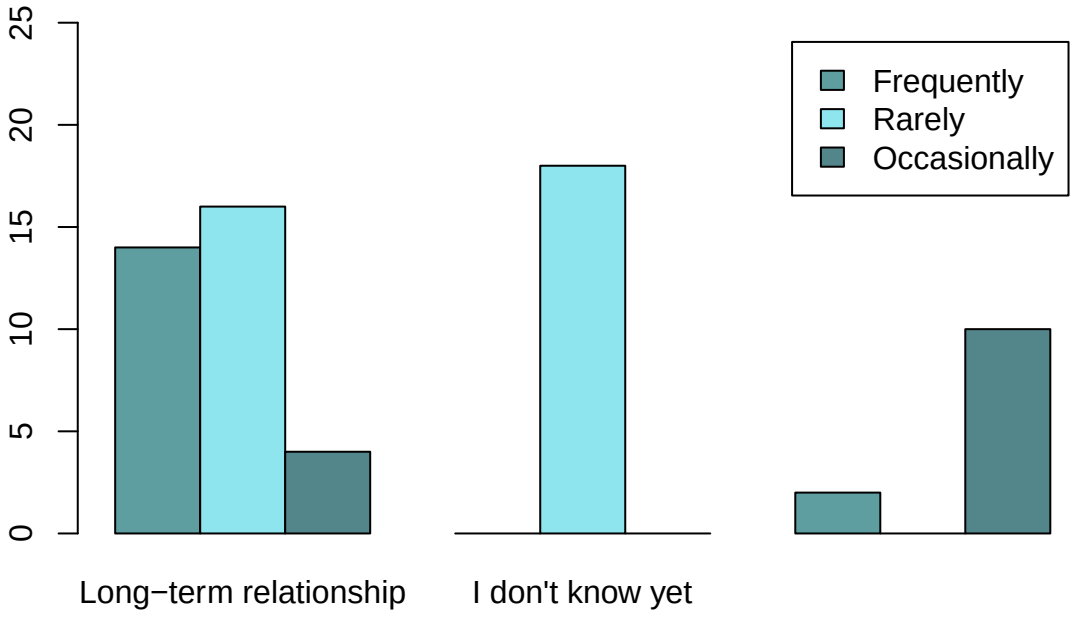
4. Based on your figure above, what can you conclude about our sample and their dating app usage?

Most people rarely use dating apps

5. Create a multivariate visualization of course choice to show any possible relationship between dating app frequency (`date$Frequency`) and dating goals (`date$Goal`).

Hint: This may require that you create a `table()`

```
barplot(table(date$Frequency, date$Goal),
        beside = TRUE,
        col = c("cadetblue", "cadetblue2", "cadetblue4"),
        ylim = c(0, 25),
        legend = TRUE)
```



6. Based on the visualization above, does there to be a relationship between dating app frequency and dating goals? Do frequent users seem to have a preference? What about occasional users?

Yes. Frequent users proportionally want long-term relationships. Occasional users tend to prefer short-term relationships.

7. Conduct a χ^2 Test of independence on dating goals and dating app frequency. Ignore any `warning` messages that appear.

```
chisq.test(date$Frequency, date$Goal)
```

```
## Warning in chisq.test(date$Frequency, date$Goal): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  date$Frequency and date$Goal
## X-squared = 48.694, df = 4, p-value = 6.763e-10
```

8. Based on the above test answer the following:

- Does there appear to be a relationship between dating app frequency and dating goals?

  Yes.

- Report your χ^2 test result using the format: $\chi^2(df = ??) = ?.???, p = ?.???$

  $\chi^2(df = 4) = 48.694, p = 6.76e - 10$

- Do these results align with your visual assessment in Question 6? Why or why not?

  Yes because frequent users tend to be searching for long-term relationships and occasional users tend to be searching for short-term relationships in greater proportion.

Part 2. Course Knowledge

1. Would you use the χ^2 Goodness-of-Fit test or the Test of Independence if you wanted to test whether there was an association between 2 categorical variables?

Test of Independence

2. The degrees of freedom for a χ^2 test of independence is calculated as (r-1) x (c-1) what are r and c, respectively?

Rows and Columns, alternatively, the levels of each categorical variable

3. If I have two categorical variables, VariableA has 3-levels and VariableB has 4 levels, what are my degrees of freedom in a χ^2 test of independence?

(3-1) x (4-1) = (2) x (3) = 6

4. You are conducting a research study on whether there is an association between Age ("Young Adult", "Adult", "Older Adult") and News Medium Preference ("TV", "Phone", "Radio"). What are the null and alternative hypotheses for this research question?

H_0: There is no association h_1: There is an association between age and news medium preference

5. You conduct your study and do a χ^2 test of independence and observe a χ^2_{crit} value is 3.84 and your $\chi^2_{observed}$ is 4.37. What do you conclude? Is your result statistically significant? Do you reject your null hypothesis or fail to reject it?

It's statisticall significant. Reject the null.