DS5110 Homework 1

Kylie Ariel Bemis

14 January 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- Knitted PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–2 ask you to practice writing some basic R functions that may be useful for data management and processing. You may need to review commonly-used base R functions from the "Vocabulary" chapter of the *Advanced R* textbook.

Problem 1

Write a function of the following form:

`imputeNA(data, use.mean = FALSE)`

- `data`: A `data.frame` for which to impute the missing values
- `use.mean`: Use the mean instead of the median for imputing continuous values

The function should return a modified copy of `data` with missing values (NAs) imputed. Continuous variables (`numeric` types) should be imputed using the median or mean (according to `use.mean`) of the non-missing values. Categorical variables (`character` or `factor` types) should be imputed using the mode. (You may find it useful to first create a function for calculating the mode.)

Examples:

```
testdf <- data.frame(
  row.names=c("Jack", "Rosa", "Dawn", "Vicki", "Blake", "Guillermo"),
  age=c(24, 23, NA, 25, 32, 19),
```

```
  city=c("Harlem", NA, "Queens", "Brooklyn", "Brooklyn", NA),
  gpa=c(3.5, 3.6, 4.0, NA, 3.8, NA))

testdf
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23     <NA> 3.6
## Dawn       NA   Queens 4.0
## Vicki      25 Brooklyn  NA
## Blake      32 Brooklyn 3.8
## Guillermo  19     <NA>  NA
```

```
imputeNA(testdf)
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23 Brooklyn 3.6
## Dawn       24   Queens 4.0
## Vicki      25 Brooklyn 3.7
## Blake      32 Brooklyn 3.8
## Guillermo  19 Brooklyn 3.7
```

```
imputeNA(testdf, use.mean=TRUE)
```

```
##            age     city   gpa
## Jack      24.0   Harlem 3.500
## Rosa      23.0 Brooklyn 3.600
## Dawn      24.6   Queens 4.000
## Vicki     25.0 Brooklyn 3.725
## Blake     32.0 Brooklyn 3.800
## Guillermo 19.0 Brooklyn 3.725
```

Problem 2

Write a function of the following form:

`countNA(data, byrow = FALSE)`

- `data`: A `data.frame` for which to count the number of missing values
- `byrow`: Should missing values be counted by row (TRUE) or by column (FALSE)?

The function should return a named numeric vector giving the count of missing values (NAs) for each row or each column of `data` (depending on the value of `byrow`). The names of the result should be the `rownames()` or `colnames()` of `data`, whichever is appropriate.

Examples:

```
testdf <- data.frame(
  row.names=c("Jack", "Rosa", "Dawn", "Vicki", "Blake", "Guillermo"),
  age=c(24, 23, NA, 25, 32, 19),
  city=c("Harlem", NA, "Queens", "Brooklyn", "Brooklyn", NA),
  gpa=c(3.5, 3.6, 4.0, NA, 3.8, NA))

testdf
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
```

```
## Rosa       23     <NA> 3.6
## Dawn       NA   Queens 4.0
## Vicki      25 Brooklyn  NA
## Blake      32 Brooklyn 3.8
## Guillermo  19     <NA>  NA
```

```
countNA(testdf)
```

```
##  age city  gpa
##    1    2    2
```

```
countNA(testdf, byrow=TRUE)
```

```
##      Jack      Rosa      Dawn     Vicki     Blake Guillermo
##         0         1         1         1         0         2
```

Part B

Problems 3–5 use datasets from the `fivethirtyeight` package. Install the `fivethirtyeight` package from CRAN using `install.packages()`.

Problem 3

Using the `police_killings` dataset, we would like to visualize the distribution of Americans killed by police by race and income. First, use the `na.omit()` function to remove missing data from the dataset. Then, visualize the count of Americans killed of each race/ethnicity, broken out by national quintile of household income (use the `nat_bucket` column). Do you notice any differences in the distribution of police killings based on income level?

Problem 4

Using the `congress_age` dataset, we would like to visualize the distribution of ages in US Congress. Use box-and-whiskey plots to visualize the distribution of ages for each congress number (#80 through #113), broken out by the congress chamber (House and Senate). How does the median age of congress members change over time? Do you notice any differences between the two chambers?

Hint: Use `as.factor(congress)` to treat it as a categorical variable.

Problem 5

Using the `bechdel` dataset, we would like to investigate if there is a relationship between passing the Bechdel test and the amount of money spent and made from a movie. The Bechdel test is a basic set of criteria designed to reveal trends of gender bias in the movies. The test asks: does a movie (1) have at least two female characters (2) who talk to each other (3) about something other than a man? Plot the worldwide gross (in 2013 dollars) as the dependent variable against the movie budget (in 2013 dollars) as the independent variable, using color to indicate whether the movie passes the Bechdel test or not. Describe the relationship between movie budget and movie gross, and whether passing the Bechdel test seems to have an affect on this relationship.