S5110 Homework 1 - Solutions

Kylie riel Bemis

14 January 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. ll of your code should run (assuming the data is available). ll 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. ny 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.

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

```
# lo d ggplot2
library(ggplot2)
```

Part

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 *dvanced R* textbook.

Pro lem 1

Write a function of the following form:

```
imputeN (data, use.mean = F LSE)
```

- `data`: 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 (N s) 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.)

```
Mode <- fun tion(x, na.rm = F LSE) {
    if ( na.rm )
```

```
        x <- x[!is.na(x)]
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
```

```
imputeN  <- fun tion(data, use.mean = F LSE) {
  for ( i in seq_along(data) ) {
    x <- data[[i]]
    if ( is.factor(x) || is.character(x) ) {
      data[[i]][is.na(x)] <- Mode(x, na.rm=TRUE)
    } else if ( is.numeric(x) ) {
      if ( use.mean ) {
        data[[i]][is.na(x)] <- mean(x, na.rm=TRUE)
      } else {
        data[[i]][is.na(x)] <- median(x, na.rm=TRUE)
      }
    } else {
      stop("unexpected data type")
    }
  }
  data
}
```

Now we test our function on the example data frame:

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

testdf
```

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

```
imputeN (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
```

```
imputeN (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
```

Pro lem 2

Write a function of the following form:

```
countN (data, byrow = F LSE)
```

- `data:` `data.frame` for which to count the number of missing values

- `byrow`: Should missing values be counted by row (TRUE) or by column (F LSE)?

The function should return a named numeric vector giving the count of missing values (N s) 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.

```
countN  <- fun tion(data, byrow = F LSE) {
    if ( byrow ) {
       countN _row <- fun tion(ic) sum(is.na(data[ic,,drop=F LSE]))
       result <- vapply(seq_len(nrow(data)), countN _row, numeric(1))
       names(result) <- rownames(data)
    } else {
       countN _col <- fun tion(ir) sum(is.na(data[[ir]]))
       result <- vapply(seq_len(ncol(data)), countN _col, numeric(1))
       names(result) <- colnames(data)
    }
  result
}
```

Now we test our function on the example data frame:

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

testdf
```

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

```
countN (testdf)
```

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

```
countN (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?

```
library(fivethirtyeight)
```

```
police_killings2 <- na.omit(police_killings)

ggplot(police_killings2, aes(x=raceethnicity,
                             fill=raceethnicity)) +
  geom_bar(na.rm=TRUE, show.legend=FALSE) +
  facet_wrap(~nat_bucket) +
  labs(x="Race/Ethnicity", y="Count of deceased",
       title="Police killings by income quintile (1-5)") +
  coord_flip() +
  theme_minimal()
```
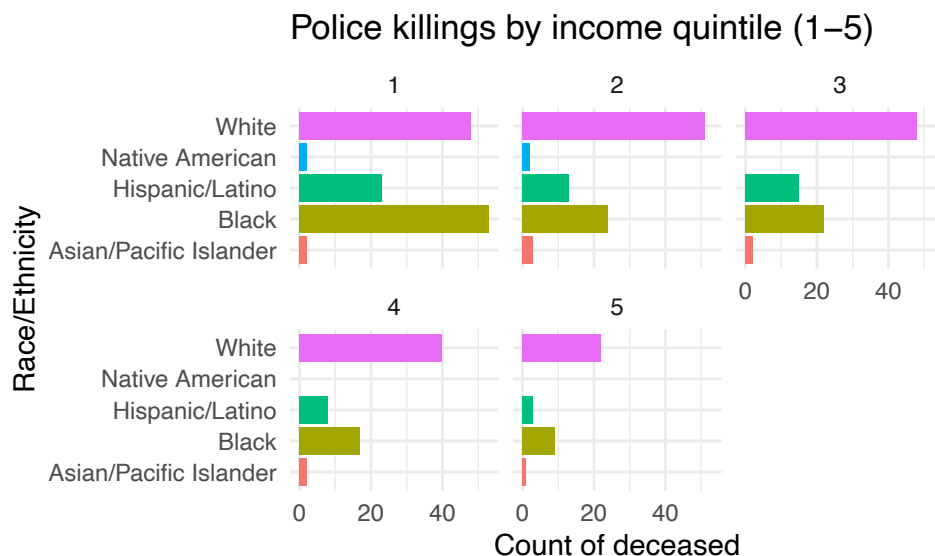

White and Black Americans make up the largest categories of Americans killed by police by race. For the lowest income quintile, Black Americans are killed by the police most often of all races. For all other quintiles, white Americans are killed by the police most often of all races. Local racial demographics are not accounted for in this graphic, so some of this relationship may be explained by differences in racial demographics in poorer versus richer neighborhoods.

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 `s.f ctor(congress)` to treat it as a categorical variable.

```
ggplot(congress_age, aes(x=as.factor(congress), y=age, color=chamber)) +
  geom_boxplot(show.legend=F LSE) +
  facet_wrap(~chamber) +
  labs(x="Congress #", y=" ge (years)",
       title="Increasing age of US Congress") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```
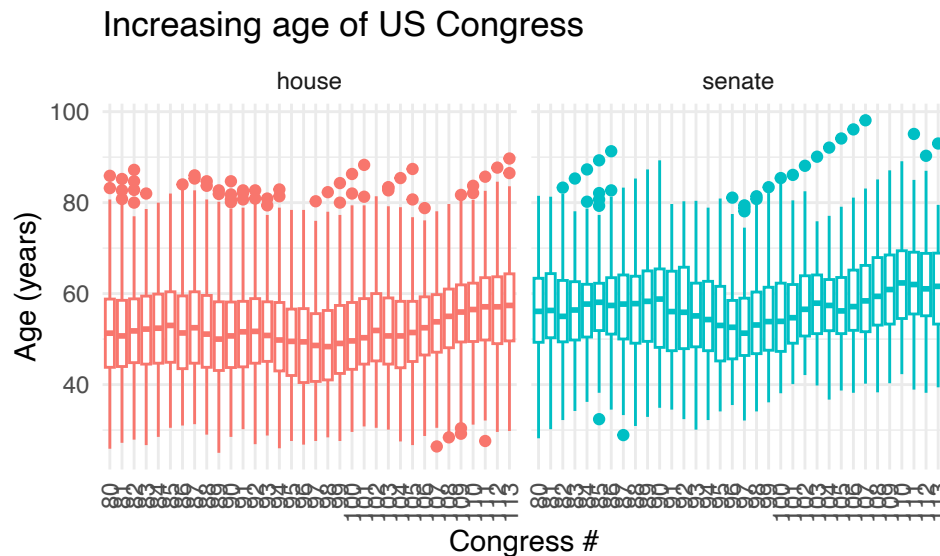

For both chambers, the median age of congress members held relatively constant until a slight dip in age around the (roughly) 97th congress. Since the 100th congress (or so), there has been a steady increase in the median age of congress members in both chambers. On average, congress members seem to be older in the Senate than in the House.

Pro lem 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.

```
ggplot(bechdel, aes(x=budget_2013,
                    y=intgross_2013,
                    color=binary)) +
  geom_point() +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  labs(x="Movie Budget in US$", y="Worldwide Gross in US$",
       title="Bechdel test has no effect on movie profitability") +
```

5

```
theme_minimal()
```

Bechdel test has no effect on movie profitability
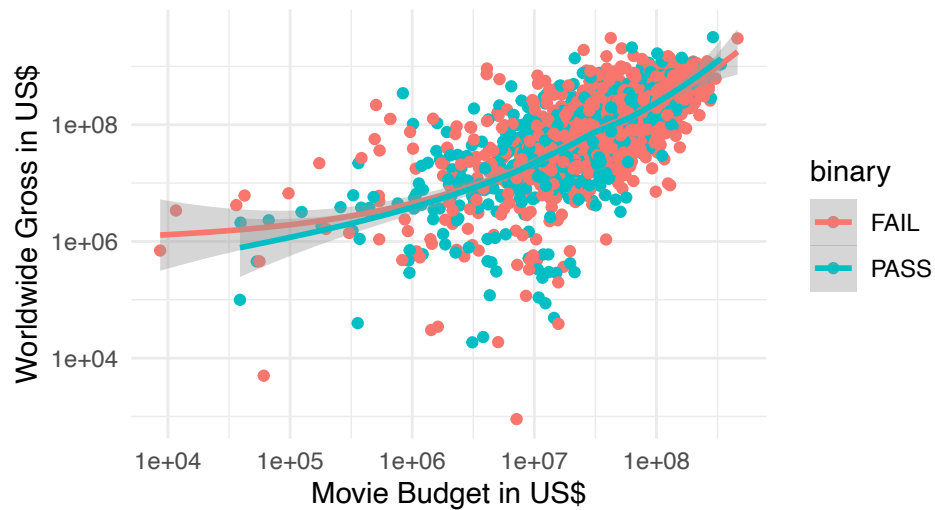
There is a positive relationship between a movie's budget and its worldwide gross in US dollars. However, whether a movie passes the Bechdel test seems to have little to no impact on the relationship between a movie's budget and its worldwide gross. Note, however, if we (optionally) investigate further, movies that pass the Bechdel test still tend to have smaller budgets than movies that fail it:

```
ggplot(bechdel, aes(x=binary,
                    y=budget_2013,
                    color=binary)) +
  geom_boxplot(show.legend=F LSE) +
  scale_y_log10() +
  labs(x="Bechdel test", y="Movie Budget in US$",
       title="Movies that fail the Bechdel test have higher budgets") +
  theme_minimal()
```

Movies that fail the Bechdel test have higher budgets