Stats 15 - Homework 3

Haojie Liu

Academic Integrity Statement

By modifying this statement, I, Haojie Liu, declare that all of the work in this assignment is my own original work. At no time did I look at the code of other students nor did I search for code solutions online. I understand that plagiarism on any single part of this assignment will result in a 0 for the entire assignment and that I will be referred to the dean of students.

```
library(knitr)
library(tidyverse)
```

Part 1: Date conversions for Wikipedia

```
library(rvest) # package used to extract information from websites
library(stringr) # package used to process text and strings
library(lubridate)
library(timetk)
library(dplyr)
library(tidyr)
```

```
url <- "https://en.wikipedia.org/wiki/List_of_U.S._states_by_date_of_admission_to_the_Union"

# the following code looks for a table on the page and stores it into list_of_tables
list_of_tables <-
  read_html(url) %>%
  html_nodes("table") %>%
  html_table()

# We want the first table in the list, so we extract it with [[1]]
table_admission_dates <- list_of_tables[[1]]

# the dates are stored in the third column.
# We select it, and pull() it out. We store the results in string_dates
string_dates <-
  table_admission_dates %>%
  select(3) %>%
  pull()

# the following code cleans up the text and
# > removes anything between an opening and closing parenthesis ( )
# > removes anything between an opening and closing square bracket [ ]
cleaned_dates <-
  string_dates %>%
```

```
  str_remove("\\(.*\\)") %>%
  str_remove("\\[.*\\]")

df <- data_frame(cleaned_dates)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
```

```
c <- table_admission_dates %>%
  select(2) %>%
  pull()

State <- data_frame(c)


df <- df %>%
  mutate(State = c,
         cleaned_dates = mdy(cleaned_dates)) %>%
  rename(Date = cleaned_dates)

head(df)
```

```
## # A tibble: 6 x 2
##   Date       State
##   <date>     <chr>
## 1 1787-12-07 Delaware
## 2 1787-12-12 Pennsylvania
## 3 1787-12-18 New Jersey
## 4 1788-01-02 Georgia
## 5 1788-01-09 Connecticut
## 6 1788-02-06 Massachusetts
```

```
tail(df)
```

```
## # A tibble: 6 x 2
##   Date       State
##   <date>     <chr>
## 1 1896-01-04 Utah
## 2 1907-11-16 Oklahoma
## 3 1912-01-06 New Mexico
## 4 1912-02-14 Arizona
## 5 1959-01-03 Alaska
## 6 1959-08-21 Hawaii
```

Your job is to now turn the string dates into a date object using lubridate.

Use `mutate()` or `cbind()` to create a new table with two columns: State, Date. Print the head and tail of the resulting table.

Part 2: Textbook, Section 6.6, Exercise 6

The HELPfull data within the mosaicData package contains information about the Health Evaluation and Linkage to Primary Care (HELP) randomized trial in tall format.

a. Generate a table of the data for subjects (ID) 1, 2, and 3 that includes the ID variable, the TIME variable, and the DRUGRISK and SEXRISK variables (measures of drug and sex risk-taking behaviors, respectively). (I have done Part a. for you.)

```
library(mosaicData)
data(HELPfull)
```

```
first3 <- HELPfull %>%
  filter(ID %in% c(1:3)) %>%
  select(ID, TIME, DRUGRISK, SEXRISK)
first3
```

```
##   ID TIME DRUGRISK SEXRISK
## 1  1    0        0       4
## 2  1    6        0       1
## 3  1   18        0       1
## 4  1   24        0       3
## 5  2    0        0       7
## 6  2    6        0       0
## 7  3    0       20       2
## 8  3    6       13       4
## 9  3   24       19       4
```

b. The HELP trial was designed to collect information at 0, 6, 12, 18, and 24 month intervals. At which timepoints were measurements available on the *RISK variables for subject 3?

At 0, 6 or 24 month

c. Let's restrict our attention to the data from the baseline (TIME = 0) and 6-month data. Use the pivot_wider() function to recreate the table seen in the textbook instructions.

```
first3 %>%
  filter(TIME == c(0,6)) %>%
  pivot_wider(names_from = TIME, values_from = c(DRUGRISK,SEXRISK))
```

```
## Warning in TIME == c(0, 6): longer object length is not a multiple of shorter
## object length
```

```
## # A tibble: 3 x 5
##      ID DRUGRISK_0 DRUGRISK_6 SEXRISK_0 SEXRISK_6
##   <int>      <int>      <int>     <int>     <int>
## 1     1          0          0         4         1
## 2     2          0          0         7         0
## 3     3         20         13         2         4
```

d. Skip part d.

Part 3: Textbook, Section 7.9, Exercise 1

Use the `HELPrct` data from the `mosaicData` library.

Follow the examples from section 7.2 that use `across()` and calculate the mean of all numeric variables (be sure to exclude missing values).

```
library(mosaicData)
data(HELPrct)

HELPrct %>%
  summarize(across(where(is.numeric), mean, na.rm = TRUE))
```

```
##        age anysubstatus     cesd       d1 daysanysub dayslink drugrisk      e2b
## 1 35.65342    0.7723577 32.84768 3.059603   75.30738 255.6056 1.887168 2.504673
##      female       i1       i2       id   indtot linkstatus      mcs      pcs
## 1 0.2362031 17.90728 24.54746 233.4018 35.72848  0.3781903 31.67668 48.04854
##     pss_fr  sexrisk avg_drinks max_drinks hospitalizations
## 1 6.706402 4.642384   17.90728   24.54746         3.059603
```

Part 4: Textbook, Section 7.9, Exercise 2

See textbook. Answer the question.

1. Step 1 create a list that contains the airport code of all seven airports.
2. Step 2 create a new table that contains the mean of delay time that gathered from each flight tibble, by using summarize() and across() to find out the mean() of the delay time.

Part 5: Textbook, Section 7.9, Exercise 4

See textbook.

Look at the example in section 7.4.2 for guidance. You will need to write a small function that will count the number of seasons for a team

```
library(Lahman)
library(purrr)
count_seasons <- function(data, team) {
  data %>%
    filter(teamID == team) %>%
    map_int(NROW)
}
```

Part 6: Textbook, Section 8.12, Exercise 1

A researcher is interested in the relationship of weather to sentiment (positivity or negativity of posts) on Twitter. They want to scrape data from https://www.wunderground.com and join that to Tweets in that geographic area at a particular time. One complication is that Weather Underground limits the number of data points that can be downloaded for free using their API (application program interface). The researcher sets up six free accounts to allow them to collect the data they want in a shorter time-frame. What ethical guidelines are violated by this approach to data scraping?

According to principle 9, "Consider carefully the ethical implications of choices we make when using data, and the impacts of our work on individuals and society." The way that the researcher collecting data is against the rule of the data source, which the main purpose of the rule to set up limit access is to share the data equally to the society.

Part 7: Textbook, Section 8.12, Exercise 2

A data scientist compiled data from several public sources (voter registration, political contributions, tax records) that were used to predict sexual orientation of individuals in a community. What ethical considerations arise that should guide use of such data sets?

First all of, the data scientist should protect the privacy of the data, "Protect the privacy and security of individuals represented in our data." After, to decrease the bias of those data, the scientist should consider not categorize the data by ethnicity or age.

Part 8: Textbook, Section 8.12, Problem 6

A Slate article (http://tinyurl.com/slate-ethics) discussed whether race/ethnicity should be included in a predictive model for how long a homeless family would stay in homeless services. Discuss the ethical considerations involved in whether race/ethnicity should be included as a predictor in the model.

If consider race as a factor or categories of the model, this be consider as a case of discrimination, which against the principle of, "Consider carefully the ethical implications of choices we make when using data, and the impacts of our work on individuals and society."