Homework #5

Due Dec 20 at 11:59pm **Points** 100 **Questions** 16
Available Dec 2 at 12am - Dec 22 at 11:59pm 21 days **Time Limit** None

Instructions

This homework assignment is a bit different than the others. **You do not need to push any files to a homework repository.** Submission will be like an in-class exam - **you will answer a series of multiple choice questions.** Most of these questions will require you to write code, but you will not submit this code - you will use it to choose one of the answer options. **This assignment is due 12/20 by 11:59pm. You may use 2 late days if you need them and have any left. Assignments submitted after 12/22 by 11:59pm will not be accepted and will receive a grade of 0/100.**

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	21 minutes	100 out of 100

ⓘ Correct answers are no longer available.

Score for this quiz: **100** out of 100
Submitted Dec 20 at 6:03pm
This attempt took 21 minutes.

Question 1 — 5 / 5 pts

Describe what the `unite()` function does.

- Converts wide data into tidy data.
- Separates multiple variables in a column name into two or more columns.
- Converts wide data into tidy data.

- Unites two columns into one; the inverse of separate.

Question 2

6 / 6 pts

Read the following data set into R:
https://raw.githubusercontent.com/datasciencelabs/data/master/ny_airquality.csv

(https://raw.githubusercontent.com/datasciencelabs/data/master/ny_airquality.csv)

Split the `Date` variable into year, month and day. Compute the average wind speed for each month. What was the average wind speed in September?

- 8.79
- 10.27
- 10.18
- 11.62
- 8.94

Question 3

7 / 7 pts

For this question we will use the following two tables:

```
master <- read_csv("https://raw.githubusercontent.com/datasciencelabs/dat
a/master/Master.csv")
player_info <- master %>% select(playerID, nameFirst, nameLast, birthYea
r, height)
salaries <- read_csv("https://raw.githubusercontent.com/datasciencelabs/d
ata/master/Salaries.csv")
```

Create a table with one row for each player that shows their average salary. Use one of the dplyr join functions to add average salaries to the

`player_info` table. Of the players born after 1986, who had the highest average salary?

- ○ Bruce Jay
- ○ Buster Posey
- ○ Justin Upton
- ◉ Masahiro Tanaka

Question 4

6 / 6 pts

Load the Teams data set from the `Lahman` library of baseball statistics.

What was the observed correlation between team home runs (HR) and team base-on-balls (BB) in 1999?

- ○ Impossible to compute
- ○ 0.833
- ◉ 0.345
- ○ 0.023

Question 5

7 / 7 pts

Read the following data set into R:

```
ny_airquality <- read_csv("https://raw.githubusercontent.com/datasciencelabs/data/master/ny_airquality.csv")
```

Use the `separate` function to split the `Date` variable into year, month and day. Then use the summarize function to compute the average temperature

for each month. What was the average temperature in July?

- 88 degrees Fahrenheit
- 76 degrees Fahrenheit
- 65 degrees Fahrenheit
- (•) 84 degrees Fahrenheit

Question 6

6 / 6 pts

Read in this data:

https://raw.githubusercontent.com/datasciencelabs/data/master/midterm2-shoe-size.csv (https://raw.githubusercontent.com/datasciencelabs/data/master/midterm2-shoe-size.csv)

It includes results from a math test administered to a random sample of students in an elementary school. Explore the relationship between the columns, in particular, note the correlation between shoe size and math scores. Which of the following conclusions would you draw from this data?

- Shoe size and math ability are clearly not related. Thus, the observed correlation is 0.
- There must be a mistake in the data because the correlation between scores and shoe size is 0.95!
- Children with small feet get made fun of and the pressure makes them not test well.

Grade is a confounding factor and explains the observed high correlation. If we stratify by grade there is no signifiant correlation.

Question 7

5 / 5 pts

After the 2008 Olympics, a small country that will remain unnamed was very proud of the fact that they won more medals than any previous year.

However, Mr. Downer, the president of the Olympic committee, warned the athletes to not rest on their laurels (as he had seen many times).

He noted that the best performing small countries in any given year rarely matched their performance 4 years later.

Which statement best explains this observation?

To win all the medals they probably overspent and have no money left to train for the next Olympics.

Data shows that athletes that win 4 or more medals tend to win only about 3 in the next Olympics. Clearly they become overconfident and don't train as hard.

This is simply an example of the regression fallacy.

Other small countries are inspired to train harder, thus taking some of the medals that would otherwise go to Mr. Downer's team.

Question 8

7 / 7 pts

Load the `Lahman` library of baseball statistics. We want to estimate the effect of team bases on balls (BB) on team runs (R) using the following model:

$$R = \alpha + \beta(BB) + \epsilon$$

We want to estimate β using the data from just one year. Assume the model holds for each year starting in 1961. Which of the following best represents an approximate 95% confidence interval for a least squares estimate of β based on data from just one year?

- [0.3, 0.75]
- [0.47, 0.58]
- [0.1, 0.9]
- [0.6, 0.7]

Question 9

6 / 6 pts

Install and load the `babynames` package. This package contains 3 datatsets. The `babynames` dataset contains the number of children of each sex given each name for each year from 1880 to 2017. The information is contained in a data frame with five variables: `year`, `sex`, `name`, `n` and `prop` (n divided by total number of applicants in that year, which means proportions are of people of that sex with that name born in that year). You can read more about the dataset here: **https://cran.r-project.org/web/packages/babynames/babynames.pdf (https://cran.r-project.org/web/packages/babynames/babynames.pdf)**

Using the babynames dataset, how many boy names contain the string `ZZ`, `Zz`, `zz` or `zZ`?

- 588
- 50

- 589
- 49

Question 10

6 / 6 pts

How many girl names contain the string `mira` or `Mira`?

- 86
- 2009
- (•) 118
- 1448

Question 11

6 / 6 pts

How many girl names have at most 1 letter "a"? Specifically, how many have at most 1 uppercase or lowercase "a"?

- 1,139,293
- 65,021
- (•) 42,647
- 67,046

Question 12

6 / 6 pts

How many girl names contain 1 or more letter "a"s? Specifically, how many have 1 or more uppercase or lowercase "a"?

- 832,350
- (•) 53,431
- 51,219
- 871,697

Question 13

6 / 6 pts

How many boy names start with a vowel and end with a vowel?

- 17,351
- 28,325
- 42,321
- (•) 1,953

Question 14

7 / 7 pts

Low sodium levels, also known as hyponatremia, have emerged as a leading cause of race-related death among marathon runners. A fairly recent **study (https://www.nejm.org/doi/full/10.1056/NEJMoa043901)** investigated a cohort of marathon runners in the US to identify the principal risk factors for low sodium levels. All registered participants 18 years or older were eligible for inclusion, and subjects were approached at random during registration and invited to participate. The primary hypothesis of the study was that excessive consumption of fluids is associated with lower serum sodium levels in marathon runners. Researchers were also interested in assessing other

factors that may predict dangerously low levels, generally thought to be below 135 milli-equivalents per liter. Accordingly, independent variables analyzed for association with serum sodium level included weight change during the race, and self-report of fluid intake including volume and frequency. Other predictors considered a priori included: female gender (dichotomous), BMI, training pace, number of previous marathons, marathon duration, use of nonsteroidal anti-inflammatory medications (NSAID; dichotomous), age, and non-white race (dichotomous).

A sample of the original data have been saved in the file marathon.csv. This sample does not contain any missing data. The variable descriptions are below.

* sodium: (serum sodium level, in milli-equivalents per liter)
* female: (1 = female, 0 = male)
* age: (years)
* bmi: (pre-race wt/ht-squared, using the appropriate units)
* fluidfr3: (fluid frequency drank through the marathon, coded as 1=every mile, 2=every other mile, 3=every 3rd mile or less)
* howmany: (number of prior marathons run)
* lwobup01: (1 = reported NSAID use, 0 = not)
* runtime: (marathon running time, in minutes)
* trainpse: (training pace for a one-mile run, in seconds)
* urinat3p: (1 = urinated 3 or more times during the race, 0 = not)
* wateld01: (1 = reported water loading prior to the race, 0 = not)
* wtdiff: (weight change during the marathon in kilograms)

Load the readr package and read in the marathon data csv:
https://raw.githubusercontent.com/datasciencelabs/data/master/marathon.csv

(https://raw.githubusercontent.com/datasciencelabs/data/master/marathon.csv)

While the authors used logistic regression to study the variables associated with hyponatremia (dichotomous sodium level), we will use linear regression to investigate the variables associated with continuous sodium level.

Fit the following model:

$$sodium = \beta_0 + \beta_1 (\text{fluidfr3}) + \epsilon$$

What is the estimate for β_1?

- 0.78
- 1.55

- (•) 1.36
- ( ) 2.72

Question 15

7 / 7 pts

Is the coefficient for `fluidfr3` significantly different from 0 at the α = 0.05 level?

- ( ) No
- (•) Yes

Question 16

7 / 7 pts

Now fit the model $sodium = \beta_0 + \beta_1\ (fluidfr3) + \beta_2\ (wtdiff)$

Is the coefficient for `fluidfr3` significantly different from 0 at the α = 0.05 level?

- ( ) Yes
- (•) No

Quiz Score: **100** out of 100