Data Wrangling Plan

Spring 2026

Objectives: Before you begin coding, it's important to have a plan for data wrangling. In this activity, you will be asked to obtain a specific data frame fitting one or more requirements. You will need to determine which `dplyr` verbs can be applied to obtain which variables in which order to obtain the desired data. You only need to worry about the **verbs**, the **variables**, and the **order**, not the particular syntax in R.

Assignment Organize yourself into groups of 2 or 3. **Work together** to answer the following questions. Write your answers on a whiteboard in the classroom.

The Data

In Spring 2022 and Spring 2023, students in intro statistics classes were asked to complete a survey containing a series of serious and not-so-serious questions. A key explaining each variable in this survey is printed on the reverse side of this page; additionally, to help orient yourself with the data, a small sample of 10 observations is available on the last pages of this document.

The Tasks

For this task, you do not need to create the actual data frames or actually compute any values. Just described the verbs, variables, and order needed to obtain the desired data frame.

1. Sort students by division, and within division, by year.
2. For each division, determine the mean and standard deviation in the number of hours spent studying.
3. Calculate the median number of college applications submitted by students who play Wordle.
4. Determine the number of students who list each of the different locations as their preferred study place.
5. Identify students whose social views are more liberal than their economic views.
6. Create a data set consisting of two columns: height in centimeters and height in meters.
7. Count how many students think both that dogs should wear pants on their back legs and that hot dogs are sandwiches.
8. Create two new variables: one, a binary categorical indicating if their distance from home is less than 500 miles and a second binary categorical indicating if their primary mode of transportation is a car. Do these variables appear to be associated? What are some ways we could tell?
9. Create your own data wrangling question based on this survey data frame.

Variable Key

Header	Explanation
Timestamp	Time when survey was submitted (Month / Day / Year Hour: Minute)
height_cm	Height in centimeters
applications	Number of colleges applied to
distance_home_miles	How far (in miles) the student's hometown is from the college.
academic_year	Year in college
weekly_study_hours	typical number of hours spent each week studying and going to class
major_division	Division of the college that major belongs to (Undecided / Science / Social Studies / Humanities / Interdisciplinary)
study_place	Typical location student studies in (Dorm commons / My room / The library / Other)
social	Personal political views on social issues (1 is very conservative, 5 is very liberal)
economic	Personal political views on economic issues (1 is very conservative, 5 is very liberal)
roommates	Number of roommates
six_month_books	Number of books read in the past 6 months
transportation	Primary mode of transportation (Car / Walk / Bike / Bus / Other)
dog_pants	How a dog would wear pants (All four legs / back legs / front legs)
hotdog	Are hotdogs sandwiches? (Yes / No)
coffee_tea	What is your preferred caffeinated beverage? (Coffee / Tea / Energy Drink / Soda / Other)
bedtime	What is your typical bedtime?
diet	Do you have any dietary restrictions?
play_wordle	Do you play Wordle? (Yes / No / What is wordle?)

dplyr Verbs

- **select:** Extracts the specified columns from the data frame.
- **filter:** Extracts rows meeting certain conditions from the data frame.
- **mutate:** Add columns to the data frame according to a formula.
- **arrange:** Sort the data frame according to the values of one or more variables
- **group_by:** Prepare the data frame to be summarized within values of a certain variable.
- **summarize:** Compute specified summary statistics for specified columns of the data frame.

A subset of 10 responses to a 2022 and 2023 Intro Stats survey are printed below:

ID	Timestamp	height_cm	applications	distance_home_miles	academic_year	weekly_study_hours
1	2/7/22 22:16	170.00	14	2904	Sophomore	47
2	2/8/22 21:27	165.10	5	1161	Sophomore	40
3	2/7/22 10:57	165.00	10	925	Sophomore	50
4	2/7/22 11:01	178.00	4	1747	Sophomore	60
5	2/7/22 11:21	175.00	7	2672	First year	40
6	2/7/22 15:29	165.10	12	1312	Senior	75
7	2/7/22 15:58	177.80	2	2000	Sophomore	38
8	2/7/22 16:18	166.00	2	48	Senior	45
9	2/7/22 16:39	167.64	1	2135	Sophomore	35
10	2/7/22 18:37	159.00	15	1596	First year	30

ID	major_division	social_views	economic_views	roommates	study_place	six_month_books
1	Social Studies	4	4	1	Dorm commons	1
2	Science	4	5	16	Dorm commons	5
3	Social Studies	5	4	0	My room	25
4	Humanities	4	5	4	My room	12
5	Science	4	3	1	My room	5
6	Interdisciplinary	4	4	0	My room	10
7	Undecided	5	5	1	My room	30
8	Science	3	5	0	My room	2
9	Science	3	1	0	My room	30
10	Science	5	5	0	My room	1

ID	transportation	dog_pants	hotdog	coffee_tea	bedtime	diet	play_wordle
1	Bike	Front legs	No	Coffee	1:30:00 AM	Vegan	Yes
2	Walk	Back legs	No	Coffee	12:30:00 AM	None	No
3	Walk	Back legs	No	Tea	9:30:00 PM	None	Yes
4	Bike	Back legs	No	Other	1:00:00 AM	Fish allergy	Yes
5	Walk	All four legs	No	Coffee	1:00:00 AM	Vegetarian	No
6	Walk	All four legs	Yes	Tea	1:30:00 AM	Pescatarian	No
7	Car	Back legs	Yes	Other	6:00:00 AM	None	No
8	Car	All four legs	Yes	Coffee	1:00:00 AM	None	Yes
9	Walk	Back legs	No	Other	12:30:00 AM	None	What's wordle?
10	Walk	Back legs	Yes	Tea	1:30:00 AM	None	Yes