Homework 4 – BBUS 301 100 points

Material covered: 9. Joining Data
10. Special Data

Follow the directions below to complete the assignment. Once you have completed the assignment, submit your answer on Canvas before the due date.

Submission Guidelines:

Due date: August 5, 2025 at 11:15 am

Submission means: You will submit an .R script containing all code, comments and explanations (remember to use # at the beginning of your comment) for each question in order to receive credit for your work.

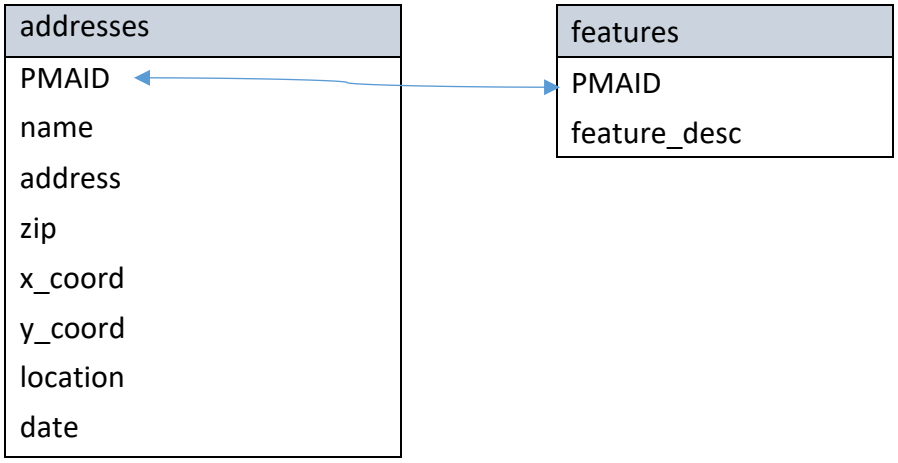
Submission details: Students must make sure that their answers are readable, complete, and submitted before the deadline. Late submission for ANY reason, whether in part or in whole, or unreadable files will trigger the late penalties. The time when I receive the last part of your answers will count when computing late penalties, not the time when you sent the answers.

Late Penalties: If an assignment is turned immediately after the due date, your score will be reduced by 10% of the maximum assignment grade. Additional 10% will be deducted from your score for every 24 hours after that.

Academic Integrity: This is an individual assignment. Students need to submit their own answers and understand those answers. I reserve the right to ask you to walk me through the code for your assignment. If you fail to explain the code in detail and the choices you made, you will receive 0 points for the part of the assignment. Students who refuse to provide the explanation for their work will receive 0 points for the entire assignment.

Using code not discussed in class: Students who are submitting code not discussed in the class will temporary receive 0 points for that part of the assignment until they meet with me and they walk me through their code. Students need to meet with me within one week of returning the graded assignment, no later than August 19. If you fail to meet with me or if you cannot explain some of the code & the choices you made, you will keep the 0 points for that part of the assignment.

Data Description:



You will download two CSV files: *Seattle_Park_Addresses.csv* and *Seattle_Park_Features.csv* from Canvas. See the diagram above, highlighting the key PMAID variable that allows you to connect the files. The park features dataset contains a list of amenities at each park, while the park addresses dataset includes name, location, and other identifying details.

(Select) Variables for the **addresses** data:

PMAID	unique park ID
name	name of the park
address	street address of the park
zip	zip code of the park
x_coord, y_coord	spatial coordinates
location	geolocation data
date	approximate opening date of the park

(Select) Variables for the **features** data:

PMAID	unique park ID
feature_desc	a list of features at each park, separated by commas

Homework Questions to Answer:

Part 1: Joining Data (15 pts)

1. (0 pts) Load the `tidyverse` package and the `lubridate` package in your library.
2. (5 pts) Import the Seattle_Park_Addresses.csv and create a data frame called **addresses**. Import the Seattle_Park_Features.csv and create a data frame called **features**.
3. (10 pts) Use the pipe operator to:
    a. inner_join the **addresses** data with the **features** data
    b. save the newly joined data frame as **parks_df**

Part 2: Special Data (85 pts)

4. (10 pts) How many parks contain the feature "Play Area"? What percentage of all parks does this represent?
5. (10 pts) Which parks offer both "Swimming Beaches" and "Grills"? Hint: Use str_detect() for both features.
6. (10 pts) Some parks have features like basketball courts, tennis courts, etc. Which park mentions the word "courts" the most times in feature_desc? How many times is it mentioned? Hint: R is case sensitive, use str_to_lower() and str_count()..
7. (10 pts) Use the pipe operator to change the variable type of variable *zip* from numeric to factor. Add a new variable called *area* to collapse the levels of *zip* as described below. Save /overwrite your data frame
    i. North: zip codes 98103, 98105, 98107, 98115, 98117, 98125, 98133, 98177
    ii. South: 98118, 98178, 98108
    iii. West: 98106, 98126, 98116, 98136, 98146
    iv. Eastside: 98102, 98122, 98112, 98144
    v. Downtown: 98101, 98121, 98104, 98109, 98119, 98199

    Hint: within the mutate function you will use both: as.factor() and fct_collapse().

8. (15) Draw a bar plot showing the number of parks in each area. Choose an appropriate title and label your axis. Mention two insights that we learn from this visual representation of our data.
9. (15 pts) Which area has the most feature-rich parks on average? Why might this be the case?
    Hints:
    - The feature_desc values are a list of features for each park separated by commas. Let's say a park has two features: Views, Trails. If we count the commas in the text and add the value 1, we find the number of features (1 + 1 = 2).
    - Create a new variable n_features by counting the number of commas in feature_desc and adding 1.

- Group, summarise, slice

10. (5 pts) Parse the *date* variable into a date format using the mutate() and the appropriate lubridate function. Save /overwrite your data frame.
11. (5 pts) Create two new variables: *year* which extracts the opening year from the *date* variable, and the second variable *age* that calculates the number of years since the park opened (from 2025). Save /overwrite your data frame.
12. (5 pts) What is the oldest park in our data (based on the age variable)? Display the park name and year it was opened.