Homework Assignment 3

STA 141A

Due Tuesday, November 14 by 5:00 pm

Description

Bike sharing programs have become popular in big cities over the last decade. In the most common arrangement, customers can rent a bike from an automated "dock" and return it later to any other dock in the city. The charge the customer is based on how long they used the bike.

The largest bike sharing program in the United States is in New York City, but bike shares have also recently formed in Los Angeles and the San Francisco Bay Area. The bike sharing programs in many cities publish anonymized data about trips and stations.

In this assignment, you will analyze open data provided by the Los Angeles and Bay Area bike shares. The data sets are provided in a ZIP compressed file. The ZIP file contains:

- 1 CSV file with Bay Area bike share stations.
- 1 CSV file with Bay Area bike share trips.
- 1 CSV file with Los Angeles (Metro) bike share stations.
- 5 CSV files with Los Angeles (Metro) bike share trips.

Questions

Use R to find answers to all of the following questions (that is, don't do any by hand or by point-and-click). Save your code in an R script.

1. Write a function that loads the Bay Area bike share trip data from a CSV file, converts the columns to appropriate data types, and then saves the tidied data frame to an RDS file. Your function should have arguments to set the path for the input CSV file and the output RDS file.

   Write a second function that does the same thing for the Bay Area bike share station data.

2. Create a map that shows the locations of the Bay Area bike share stations in San Francisco (only). Label each station with its name. Make the size of each point correspond to the number of trips started from that station. Discuss what you can conclude from the map.

3. Write a function that loads the Los Angeles bike share trip data from the 5 provided CSV files, binds them into one data frame, converts the columns to appropriate data types, and saves the tidied data frame to an RDS file. Your function should have arguments to set the path for the input directory and the output RDS file. Keep your function short and simple by using an apply function rather than repeating code.

   Write a second function that loads, tidies, and saves the Los Angeles bike share station data.

4. Create a map that shows the locations of the Los Angeles bike share stations near downtown Los Angeles (only). Label each station with its name. Make the size of each point correspond to the number of trips started from that station. Discuss what you can conclude from the map.

5. How do trip frequency, distance, and duration change at different times of day? Investigate for both the Bay Area bike share and the Los Angeles bike share. Compare your findings. The `geosphere::distGeo()`[1] function can compute distances for longitude and latitude coordinates.
6. For Bay Area bike share trips in San Francisco, how does bearing (angle) change at different times of day? What can you conclude about traffic patterns in the city? The `geosphere::bearing()` function can compute bearings for longitude and latitude coordinates.

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted[2] tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 5 pages** including graphics, but excluding code and citations.

What To Submit

Submit a digital copy on Canvas. The digital copy must contain your report (as a PDF) and your code (as one or more R scripts).

Additionally, submit a printed copy to the box in the statistics department office[3]. The printed copy must contain your report and your code (in an appendix). Please print double-sided to save trees. It is your responsibility to make sure the graphics are legible in the printed copy!

Data Documentation

The features in the Bay Area bike share trips data set are:

`trip_id`	unique id of trip
`duration_sec`	duration of trip (in seconds)
`start_date`	time trip started (in PST)
`start_station_name`	name of station where trip started
`start_station_id`	unique id of station where trip started
`end_date`	time trip ended (in PST)
`end_station_name`	name of station where trip ended
`end_station_id`	unique id of station where trip ended
`bike_number`	unique id of bike used
`zip_code`	home zip code of rider (potentially unreliable)
`subscriber_type`	Subscriber = annual or 30-day member; Customer = 24-hour or 3-day member

The features in the Bay Area bike share stations data set are:

`station_id`	unique id of station
`name`	name of station
`latitude`	latitude of station
`longitude`	longitude of station
`dockcount`	total number of docks at the station
`landmark`	city
`installation_date`	date the station was installed

A station row may be partially duplicated in the data set if the station was moved.

The features in the Metro bike share data sets are documented here: `https://bikeshare.metro.net/about/data/`

[1] The `::` is R notation for a function in a specific package. Read it as `PACKAGE::FUNCTION()`.

[2] See the graphics checklist on Canvas.

[3] 4th floor of Mathematical Sciences Building

Relevant Functions

All of the functions from previous assignments, as well as:

`cut()`, `saveRDS()`, `strptime()`, `as.data.frame()`, `lapply()`

The functions in the packages *ggmap*, *geosphere*, *ggplot2*, *lubridate*, and *readr* may also be useful, although you are not required to use these packages.