**Instructions:** Download the **citibike** dataset from Canvas and import it into R (see the handout on importing data to review the method demonstrated in class.) Once loaded, check that the dataset has loaded properly by confirming that the object appears in your environment, and contains the correct number of observations (50,000 rows) and variables (18 columns).

Citi Bike is a bikesharing company based in New York City. Customers rent a bicycle from a station, and may ride the bike anywhere in the city for as long as they like. At the end of the trip, the customer deposits the bicycle at a designated Citi Bike station. Customers pay a small fixed fee for the rental session, plus a variable fee based on the duration of the trip. Customers may register as subscribers of the service, which allows them small discounts on trips and other special offers. The company is a predecessor to contemporary scooter-sharing companies like Bird or Lime. In this dataset, each row represents a rental session of a Citi Bike bicycle. The variables contained in the dataset are as follows:

|  |  |
| --- | --- |
| **name** | **description** |
| trip\_id | Primary key; a unique identifier of the rental session. |
| bike\_id | A code identifying the bike rented for the session. |
| weekday | The day of the week on which the session occurred. |
| start\_hour | The hour of the day (0-23) at which the session began. |
| start\_time | The date and time at which the rental session began. |
| start\_station\_id | The code identifying the station at which the rental session began. |
| start\_station\_name | The cross streets identifying the station at which the rental session began. |
| start\_station\_latitude | The latitudinal coordinates of the station at which the rental session began. |
| start\_station\_longitude | The longitudinal coordinates of the station at which the rental session bega. |
| end\_time | The date and time at which the rental session ended. |
| end\_station\_id | The code identifying the station at which the rental session ended. |
| end\_station\_name | The cross streets identifying the station at which the rental session ended. |
| end\_station\_latitude | The latitudinal coordinates of the station at which the rental session ended. |
| end\_station\_longitude | The longitudinal coordinates of the station at which the rental session ended. |
| trip\_duration | The length of the rental session, in seconds. |
| subscriber | An indicator of whether the customer who initiated the session was a subscriber to Citi Bike. |
| birth\_year | The year that the customer who initiated the rental session was born. |
| gender | The gender of the customer who initiated the rental session (0 = unknown, 1 = male, 2 = female) |

**Q1.** Identify the data type of each of the following variables: **(1/4 pt each, 3 pts total)**

1. trip\_id
2. bike\_id
3. weekday
4. start\_hour
5. start\_station\_id
6. start\_station\_name
7. start\_station\_latitude
8. start\_station\_longitude
9. trip\_duration
10. subscriber
11. birth\_year
12. gender

**Q2.** Write the command to generate a proportion table showing the proportion of sessions that were initiated by subscribers vs. non-subscribers. What proportion of trips were initiated by subscribers? **(1 pt)**

**Q3.** Write the command to create a new variable called trip\_minutes that converts the duration of the trip from seconds to minutes. What is the average length of a trip in minutes? **(2 pts)**

**Q4.** Using the aggregate() command, find the average trip length in minutes among subscribers vs. non- subscribers. **(2 pts)**

**Q5.** Write the command to create a new variable called weekend that flags all trips that occurred on either Saturday OR Sunday. What proportion of trips occur on the weekend? **(2 pts)**

**Q6.** Write the command to create a crosstable of subscriber status by weekend status. Express the crosstable as a proportion table, with proportions aggregated by row (you will need to include the margin parameter demonstrated in class.) Describe the patterns you see in the table: does there appear to be a difference in bike usage for weekdays vs. weekends among subscribers vs. non-subscribers? **(2 pts)**

**Q7.** Using the information found in Q4 and Q6, offer a possible explanation for why you’re observing the differences in ride length and weekend vs. not among subscribers vs. non-subscribers. Why do you think each group is using the service? **(2 pts)**

**Q8.** Write the command to create a crosstable of subscriber status by gender. Express the crosstable as a proportion table, with proportions aggregated by row (you will need to include the margin parameter demonstrated in class.) According to the table, does Citi Bike’s subscriber base appear to skew male or female? **(2 pts)**

Note: R often expresses decimals using scientific notation. As a reminder, the symbol e+01 indicates to move the decimal one place to the right, and the symbol e-01 indicates to move the decimal one place to the left.

**Q9.** Write the command to create a variable called age that subtracts the year the rider was born from the current year, and create a histogram of the age variable. Describe the distribution of ages shown in the data. Does anything strike you as odd? **(2 pts)**

**Q10.** Using the aggregate() command to find the average age by gender. Does there appear to be a meaningful difference in average rider age by gender? **(2 pts)**