Data sources used

- Download source data from https://divvy-tripdata.s3.amazonaws.com/index.html
- License for using the data from https://ride.divvybikes.com/data-license-agreement
- List of raw data files:

    1. 202110-divvy-tripdata.csv
    2. 202111-divvy-tripdata.csv
    3. 202112-divvy-tripdata.csv
    4. 202201-divvy-tripdata.csv
    5. 202202-divvy-tripdata.csv
    6. 202203-divvy-tripdata.csv
    7. 202204-divvy-tripdata.csv
    8. 202205-divvy-tripdata.csv
    9. 202206-divvy-tripdata.csv
    10. 202207-divvy-tripdata.csv
    11. 202208-divvy-tripdata.csv
    12. 202209-divvy-publictripdata.csv
    13. 202210-divvy-tripdata.csv

- Create project root folder - /20221205-capstone_datascience/ - as ../ in this document
- Move raw, zipped data to ../099_original_raw_data/001_raw_compressed/
- Unzip data into ../099_original_raw_data/002_raw_csv/
- Copy data to ../002_data/001_csv/ as a working copy
- Rename file 202209-divvy-publictripdata.csv in ../002_data/001_csv/ to 202209-divvy-tripdata.csv in order to match filename schema

- Download quarterly formatted data from https://divvy-tripdata.s3.amazonaws.com/index.html to compare with monthly formatted data

- Copy dataset Divvy_Trips_2020_Q1.csv to ../099_original_raw_data/002_raw_csv/ and working copy to ../002_data/001_csv/

- Comparing data in Divvy_Trips_2020_Q1.csv to 202210-divvy-tripdata.csv as to similarity.

    + Using OpenOffice Calc V 4.1.7 for comparison
    + Column Names match
    + The *rideable_type* column contains more information in the 202210-divvy-tripdata.csv as opposed to the Divvy_Trips_2020_Q1.csv
    + In Divvy_Trips_2020_Q1.csv only one type: dockable_bike
    + In 202210-divvy-tripdata.csv either classic_bike or electric_bike
    + The columns *start_station_id* and *end_station_id* have different values
    + In Divvy_Trips_2020_Q1.csv the values are numerical
    + In 202210-divvy-tripdata.csv some values are numerical, others are a mix of uppercase letters and numbers a few also include a dash '-'
    + The columns *start_lat*, *start_lng*, *end_lat*, *end_lng* are formatted differently
    + In Divvy_Trips_2020_Q1.csv the values are truncated to 4 decimal places
    + In 202210-divvy-tripdata.csv the values contain up to 10 decimal places
    + Some records in 202210-divvy-tripdata.csv do not contain start or end station data, leading me to believe it includes data from bikes that were rented not from a station, but standing around in the city.

- Checking 202110-divvy-tripdata.csv, the oldest of the monthly datasets, to confirm that the differences persist

- I will be using the monthly datasets for the analysis, because:

    + The data is more current
    + It includes additional data as to the type of bike used
    + The data that is missing (station names) or not formatted nicely (gps coordinates) doesn't seem to impact the analysis

- Download weather data for the Chicago area from https://www.weather.gov/wrh/climate?wfo=lot
- Combine data into a spreadsheet climate_chicago_2110-2210.ods in ../099_original_raw_data/
- Export data for use in R Studio as climate_chicago_202110-202210.csv
- Copy dataset climate_chicago_202110-202210.csv to ../099_original_raw_data/002_raw_csv/ and working copy to ../002_data/001_csv/