Cyclistic Data Cleaning Log

1/27:

- All fields are consistent between tables with correct data type
- Confirmed all ride_id rows are unique in all 12 tables
- Number of observations with NULL (checking all fields)
  - A significant fractions of observations have a NULL value in some field

Month	Total Unique Rides	Total Nulls in Observations	Fraction of Nulls
01	144873	31065	0.21
02	223164	38428	0.17
03	301687	71409	0.24
04	415025	117227	0.28
05	609493	167325	0.27
06	710721	216395	0.3
07	748962	208031	0.28
08	755639	214424	0.28
09	821276	284042	0.35
10	616281	167167	0.27
11	335075	89104	0.27
12	178372	47642	0.27

- If this were a project where I were interacting with the company directly, I would seek out the missing data since it hovers around 30% for most of the months
  - Any rows that contain a NULL value will be excluded
- NULLs per field for entire yearly data

Field	NULL count
ride_id	0
rideable_type	0
started_at	0
ended_at	0

start_station_name	1073951
start_station_id	1073951
end_station_name	1104653
end_station_id	1104653
start_lat	0
start_lng	0
end_lat	7232
end_lng	7232
member_casual	0

  - Most of the NULL counts comes from the station names and id
  - There are several lat/lng data close in value that the same stations share, making it unlikely to correct the missing name and id values
- Rideable_type data has three possible values
  - classic_bike
  - electric_bike
  - electric_scooter
- Combined all monthly datasets to create a single yearly set with total observations: 5,860,568
- Number of starting stations in yearly dataset is 1,808
  - Starting station names were trimmed and made lowercase
- Number of ending stations in yearly dataset is 1,815
  - Ending station names were trimmed and made lowercase
- Comparing the stations not included in starting or ending columns after filtering for distinct, trimmed, and lowercase station names
  - Stations in starting not in ending (8)
    - public rack - strohacker park
    - public rack - tuley (murray) park
    - public rack - artesian & 71st
    - public rack - prairie ave & 78th st
    - public rack - northwest hwy & highland ave
    - public rack - ellis ave & 132nd pl
    - public rack - pittsburgh ave & irving park
    - oketo ave & addison
  - Stations in ending not in starting (15)

  - base - 2132 w hubbard
  - public rack - champlain ave & 134th st
  - public rack - exchange ave & 131st st
  - kedzie ave & 38th pl
  - public rack - western & 79th
  - public rack - kedzie & 73rd
  - public rack - baltimore ave & 134th st
  - public rack - pulaski & 84th
  - west - chi - cassette repair
  - public rack - langley ave & 87th st
  - scooters - 2132 w hubbard st
  - w. chicago warehouse
  - public rack - eberhart ave & 131st st
  - public rack - columbus & 79th
  - public rack - northwest hwy & overhill ave
- While there is a difference in starting and ending stations, there are no duplicates and no misspelled stations
- Possible reason for difference is the NULL values in dataset

1/28:

- Station ID values trimmed to remove extra spaces
  - Unique start_station_id: 1,763
  - Unique end_station_id: 1,768
  - Mismatch of station names and IDs in the dataset
- Created a field called ride_length_s and ride_length_hh:mm:ss to track total time of each ride
  - ride_length_s is an integer value
  - ride_length_hh:mm:ss field is a time value
  - Longest ride: 93,596 seconds; 01:59:56 hours minutes seconds
  - Shortest ride:
- Start times recorded after end times are also filtered out to prevent negative/zero ride_length values
  - 1,600 observations with trip times less than or equal to 0
  - 4,646 observations with 1 second trip times
  - 26,434 observations with trip times between [1,10) seconds
  - We will keep only trips that were at least 1 second in duration as any value smaller would not have meaningful impact on the analysis and may have to do with a system issue or cancelation
- Created a field called day_of_week that returns a day name string

- Finished investigation/cleaning, removing all observations that contain a NULL in the start/end time as this will be the metric most used
    - The station names/ids have too many NULL values to make them useful in the dataset as it is, if we were able to collect the missing data and ensure it is all standardized (the company would most likely have to do this since they have all of this data) to make it usable to trace common routes and hotspot stations
    - The latitude/longitude have much fewer NULL values, but it was found that the same station did not correspond to an identical lat/long value and would need to be checked if the data could be standardized as well
- Final cleaned yearly dataset has 5,858,968 unique observations with the following cleaning procedures applied:
    - Found unique ride_id that were trimmed and made lowercase to standardize
    - Created a field that lists name of day called day_of_week
    - Created a field that lists month called month
    - Created field that records length of ride in seconds called ride_length_s
    - Created field that records length of ride in hh:mm:ss format called ride_length_hh:mm:ss
    - Applied the following filters even though NULL values do not exist in the fields to keep query applicable to other datasets
        - Filters out observations with NULL values in ride_id
        - Filters out observations with NULL values in started_at
        - Filters out observations with NULL values in ended_at
        - Filters out observations with NULL values in trimmed mamber_casual
    - Filtered out any ride_length_s that were less than 1 second

1/30

- During analysis, found cases where observation that started in prior month but ended in current month is included in the latter, making a difference when looking at month data or labeled yearly data.