Data cleaning log - Cyclistic Bike Share

Data cleaning steps that have been taken:

Excel:

- Made sure that the columns for each file for the 12 months of data are the same.
- Made sure that there were no duplicate rows in any of the datasets.
- Checked that the information is consistent among the datasets, by analyzing the unique values in each column and that there are no spelling errors or misfielded values.
  Since the data is collected by the geo tracking device instead of being manually entered, the probability of spelling errors or misfielded values is very rare. That's the reason why misfielded values, spelling errors and trailing or leading spaces weren't found in the data.

SQL:

- Removed the TEST station names, using a WHERE clause.
- Trimmed the columns rideable_type, start_station_name, end_station_name and member_casual, using the TRIM function.
- Formatted the different fields in the spreadsheet to make sure that fields have the same data type and that the dates have the same format across the different spreadsheets:
  - ride_id: string
  - rideable_type: string
  - started_at: datetime
  - ended_at: datetime
  - start_station_name: string
  - start_station_id: string
  - end_station_name: string
  - end_station_id: string
  - start_lat: integer
  - start_lng: integer
  - end_lat: integer
  - end_lng: integer
  - member_casual: string
- Appended the datasets for the 12 months in order to create a consolidated file, and name the consolidated table Cyclistic.

Tableau:

- Changed the column headers in the Cyclistic table.
- Corrected ride length negative values to positive values.

New columns created:

- day_of_week: Calculates the day of the week that each ride started, formatted as the day's name, from Sunday to Saturday (created in Excel).
- ride_length: The difference between the ride start time and end time (created in SQL).