Data Cleaning and Analysis Report for 2024 Cyclistics Ride Data

1. Data Collection and Initial Setup:

- The dataset consisted of 12 CSV files, each representing a month of ride data for the year 2024.
- A Python function was developed to merge all monthly data into a single DataFrame, joining on the start datetime to consolidate the dataset for analysis.
- After merging, the combined dataset contained **5,860,568 entries** across **13 columns**.

```python
import os
import pandas as pd

# Function to load each csv from folder, subfolders
def load_csvs_as_trip_variables(main_folder):
    trip_data = {}
    sorted_folders = sorted(os.listdir(main_folder))
    for i, folder in enumerate(sorted_folders, start=1):
        folder_path = os.path.join(main_folder, folder)

        if os.path.isdir(folder_path):
            for file in os.listdir(folder_path):
                if file.endswith(".csv"):
                    file_path = os.path.join(folder_path, file)
                    df = pd.read_csv(file_path)
                    var_name = f"trip_{str(i).zfill(2)}
                    trip_data[var_name] = df

                    print(f"Loaded {file} into {var_name}")

    return trip_data

main_folder = r"Case Study Cyclistics"


trips = load_csvs_as_trip_variables(main_folder)
print(trips['trip_01'].head())

# Loop through the dictionary and assign each DataFrame to a variable
for key, df in trips.items():
    globals()[key] = df

df = pd.concat([trip_01, trip_03, trip_05, trip_07, trip_09, trip_11,
                trip_13, trip_15, trip_17, trip_19, trip_21, trip_23],
               ignore_index=True)

```

2. Initial Data Exploration:

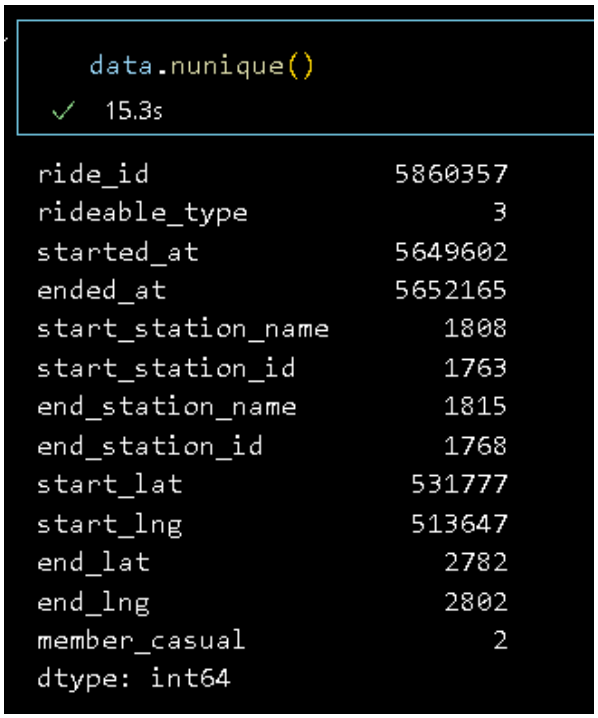
- The data included start and end dates, start and end locations, rideable types, and membership categories.
- This provided the foundation for **time-based** and **location-based** analyses.

Following were the columns:

```
RangeIndex: 5860568 entries, 0 to 5860567
Data columns (total 13 columns):
 #   Column              Dtype
---  ------              -----
 0   ride_id             object
 1   rideable_type       object
 2   started_at          object
 3   ended_at            object
 4   start_station_name  object
 5   start_station_id    object
 6   end_station_name    object
 7   end_station_id      object
 8   start_lat           float64
 9   start_lng           float64
 10  end_lat             float64
 11  end_lng             float64
 12  member_casual       object
dtypes: float64(4), object(9)
memory usage: 581.3+ MB
```

3. Key Findings and Data Anomalies:

- **Rideable Types:** 3 unique rideable types.
- **Membership Types:** 2 categories (member and casual).
- **Station Data Inconsistencies:**
    - Start station names: **1,808 unique values**
    - Start station IDs: **1,763 unique values**
    - End station names: **1,815 unique values**
    - End station IDs: **1,768 unique values**
    - This discrepancy raised questions about whether stations had sub-docks or naming inconsistencies.



4. Handling Duplicates:

- **211 duplicate ride IDs** were found, primarily due to millisecond-level timestamp issues.
- Duplicates were exported ”**duplicated_rows.csv”** and removed, and milliseconds were truncated for accurate datetime conversion.
- After resolving duplicates, the remaining rows were **4,207,975**.

5. Dealing with Data Quality Issues:

- **Negative and Zero Durations:**
    - **227 rows** with negative durations were exported **“negative_ride_duration.csv”** and removed.

    - **496 rows** with zero durations were also exported **“zer_ride_duration.csv”** and removed.
- **Null Values in Location Data:**
    - **1,651,749 rows** had nulls in station names or IDs but not in other fields.
    - These rows were exported **“null_feilds_data.csv”** for review and removed from the final dataset.

6. Correcting Location ID and Name Mismatches:

- **Shared IDs and Names:**
    - **Start & End** ids for shared location names were exported **“start_id_shared.csv”** & **“end_id_shared.csv”.**
    - **26,884 rows** had shared station IDs for multiple names.
    - **21,447 public rack entries** were corrected and updated and exported **“public_rack_data.csv”**.
    - **37 test records** were identified and removed, exported as **“testing_cycle_data.csv”**.
    - **21,540 rows** with repeated TA1305000030 IDs were dropped and exported **“ta_id_data.csv”**.
    - **6,323 rows** with minor location name typos (e.g., "avenue" vs. "ave") were corrected and exported **“minor_name_error_id_data.csv”**
- **Non-Unique IDs:**
    - **260,638 rows** had location IDs used across multiple locations, which were exported **“loc_id_dstnct_loc_name.csv”** and dropped to preserve data integrity.

7. Final Data Preparation:

- **Rows Remaining After Cleaning: 3,675,066**
- **Exported Clean Data:** "v8_data_for_visualization.csv"

8. Feature Engineering:

- **New Columns Created:**
    - `month`: Extracted month as an integer.
    - `quarter`: Calculated quarter based on the month.
    - `ride_distance_km`: Estimated ride distance in kilometers.
    - `ride_hour`: Extracted the hour from the start datetime.
    - `time_bucket`: Grouped rides into time-based buckets (e.g., 6AM-9AM, 9AM-12AM).

9. Final Insights:

- **Average Ride Duration:**
  - Casual riders: **25 minutes**
  - Members: **12 minutes**

10. Tools and Techniques:

- **Python Libraries Used:**
  - `pandas`, `numpy`, `datetime`, `geopy` (for distance calculations), `SQL` (for shared id for location analysis).
- **Data Exports:**
  - Intermediate files were exported for each cleaning stage, preserving traceability.
  - Final data exported **"v8_data_for_visualization.csv"**

This thorough data cleaning process ensured that the final dataset was well-structured, free of critical inconsistencies, and ready for advanced analytics and visualization.