# WEEK 4 OF CELLULA INTERNSHIP TASK4

## EDA ON RIDE FARE DATASET

# DATASET OVERVIEW

**Basic Information:**

- **500,000 rows × 26 columns**
- **Time Span: 2009–2015**

**Column Categories:**

- **Identifiers: user_id, user_name, driver_name, key**
- **Ride Context: car_condition, weather, traffic_condition**
- **Time Features: pickup_datetime, year, month, weekday, hour, day**
- **Location: pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude**
- **Distance: distance, bearing, jfk_dist, ewr_dist, lga_dist, sol_dist, nyc_dist**
- **Fare: fare_amount**

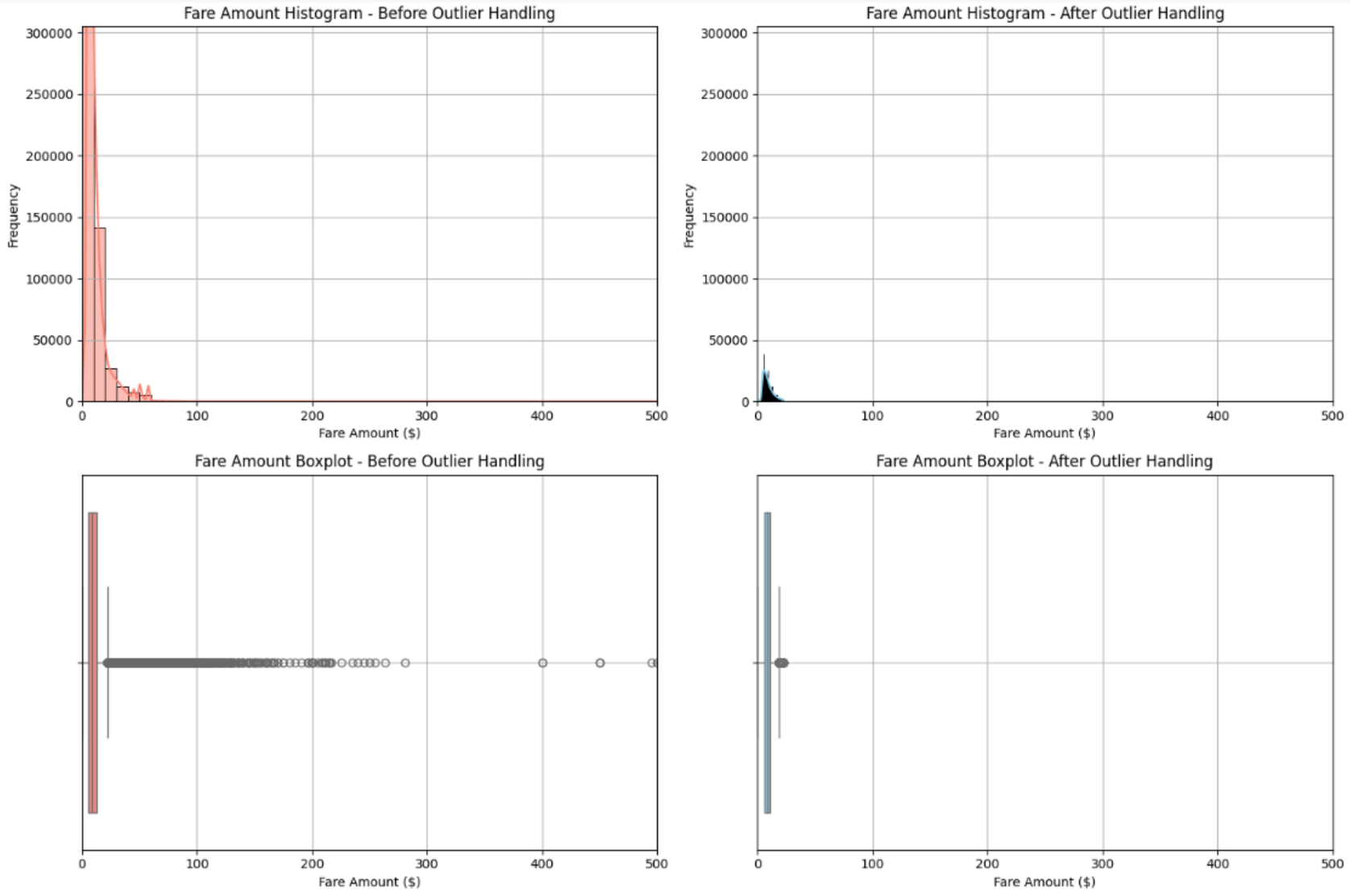
# 1. DATA CLEANING STEPS

- **Standardized column names :** Converted all column names to lowercase and replaced spaces with underscores for consistency
- **Checked for missing values :** 5 Rows Contains missing values - drop them
- **Checked for duplicate records :** No Duplicates

| | user_id | user_name | driver_name | car_condition | weather | traffic_condition | key | fare_amount | pickup_datetime | pickup_longitude | ... | month | weekday | year | jfk_dist | ewr_dist | lga_dist | sol_dist | nyc_dist | distance | bearing |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **120227** | BOLML7gg | Carol Kim | John Scott | Very Good | rainy | Dense Traffic | 2012-12-11 12:57:00.00000013 | 12.5 | 2012-12-11 12:57:00 | -1.291417 | ... | 12 | 1 | 2012 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **245696** | AA838qgm | Mark Jones | Christy Taylor | Bad | windy | Dense Traffic | 2013-03-21 18:07:07.0000001 | 86.5 | 2013-03-21 18:07:07 | -1.291397 | ... | 3 | 3 | 2013 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **340533** | BqrlHUr2 | Joshua Mullins | Sarah Khan | Excellent | cloudy | Dense Traffic | 2012-12-11 12:50:52.00000010 | 27.5 | 2012-12-11 12:50:52 | -1.291188 | ... | 12 | 1 | 2012 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **428108** | rL4WTHGq | Richard Brown | Rachel Miller | Very Good | cloudy | Flow Traffic | 2011-09-08 09:12:52.0000001 | 11.8 | 2011-09-08 09:12:52 | -1.291317 | ... | 9 | 3 | 2011 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **471472** | tKIOKS8Y | Larry Wade | Howard Jackson | Good | windy | Congested Traffic | 2012-12-11 12:34:20.0000006 | 7.8 | 2012-12-11 12:34:20 | 0.000000 | ... | 12 | 1 | 2012 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

5 rows × 26 columns

# 2. OUTLIER DETECTION & HANDLING



# 2. OUTLIER DETECTION & HANDLING

**Before Handling:**

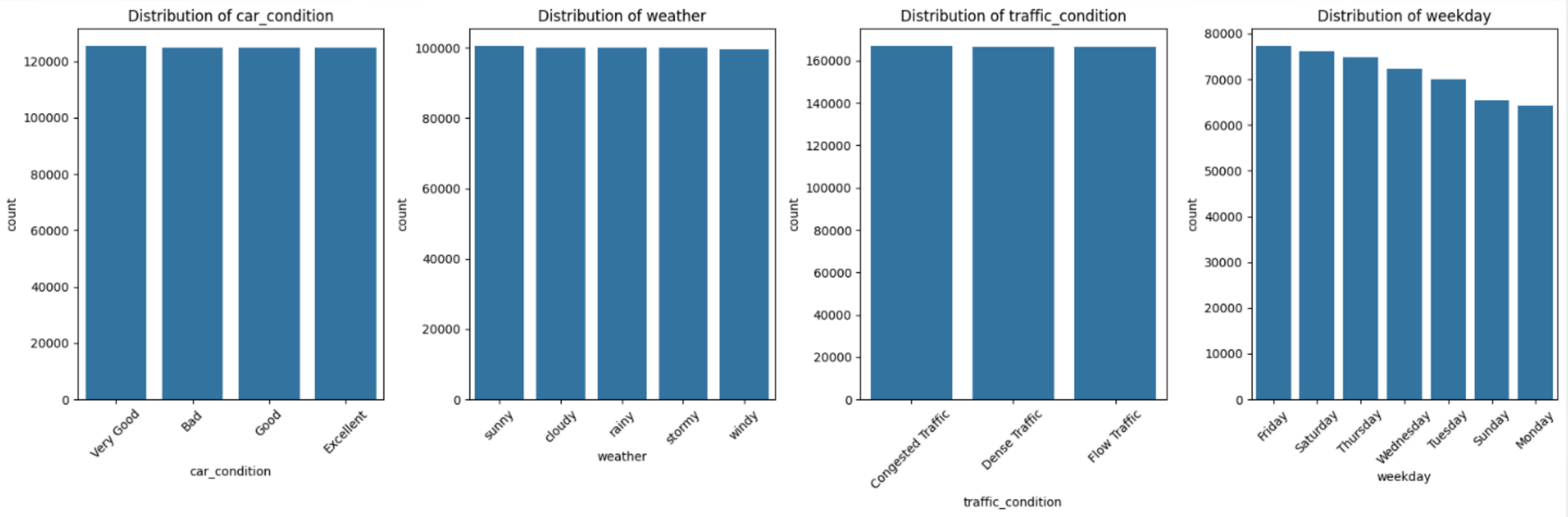
- Outliers: 43,329
- Fare range exceeded $100
- Data was highly skewed

**After Handling:**

- Removed negative fares
- Capped max fare at $22.25
- Outliers reduced to 0

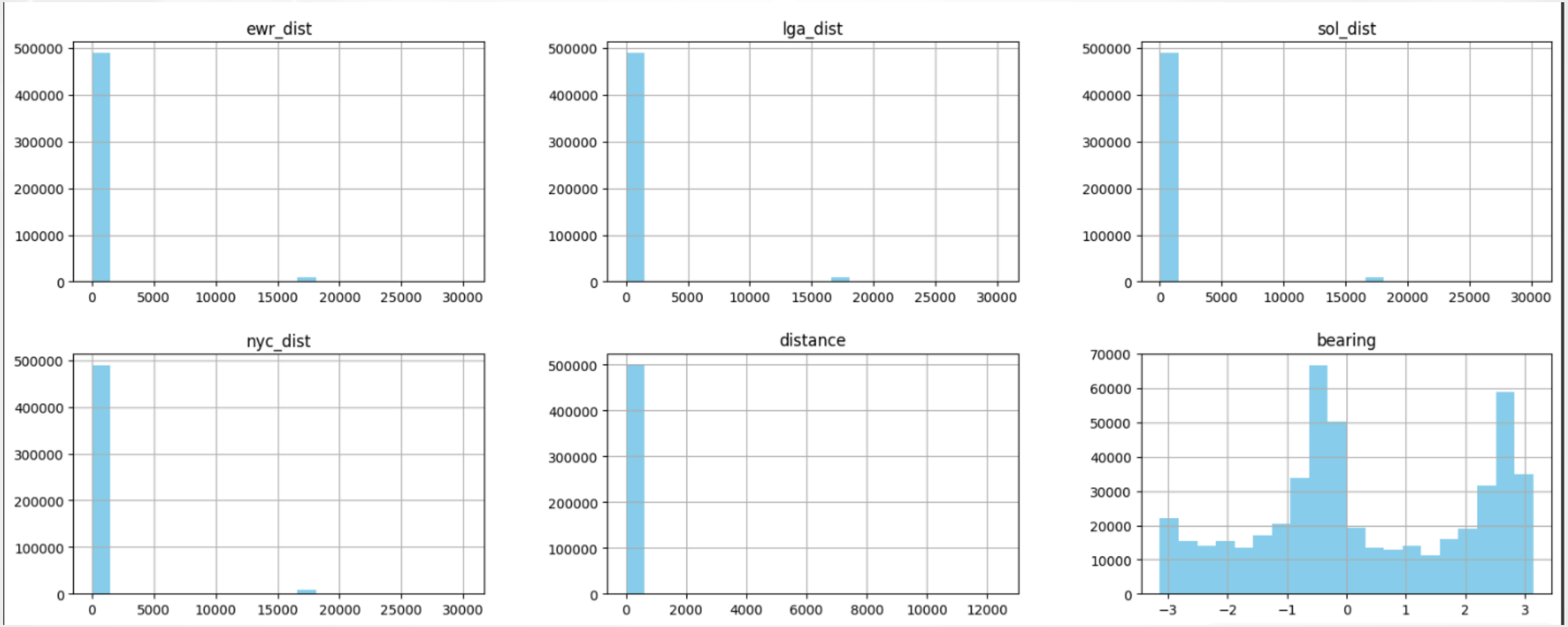
# 3. CATEGORICAL DATA

**Distribution of Categorical Features**



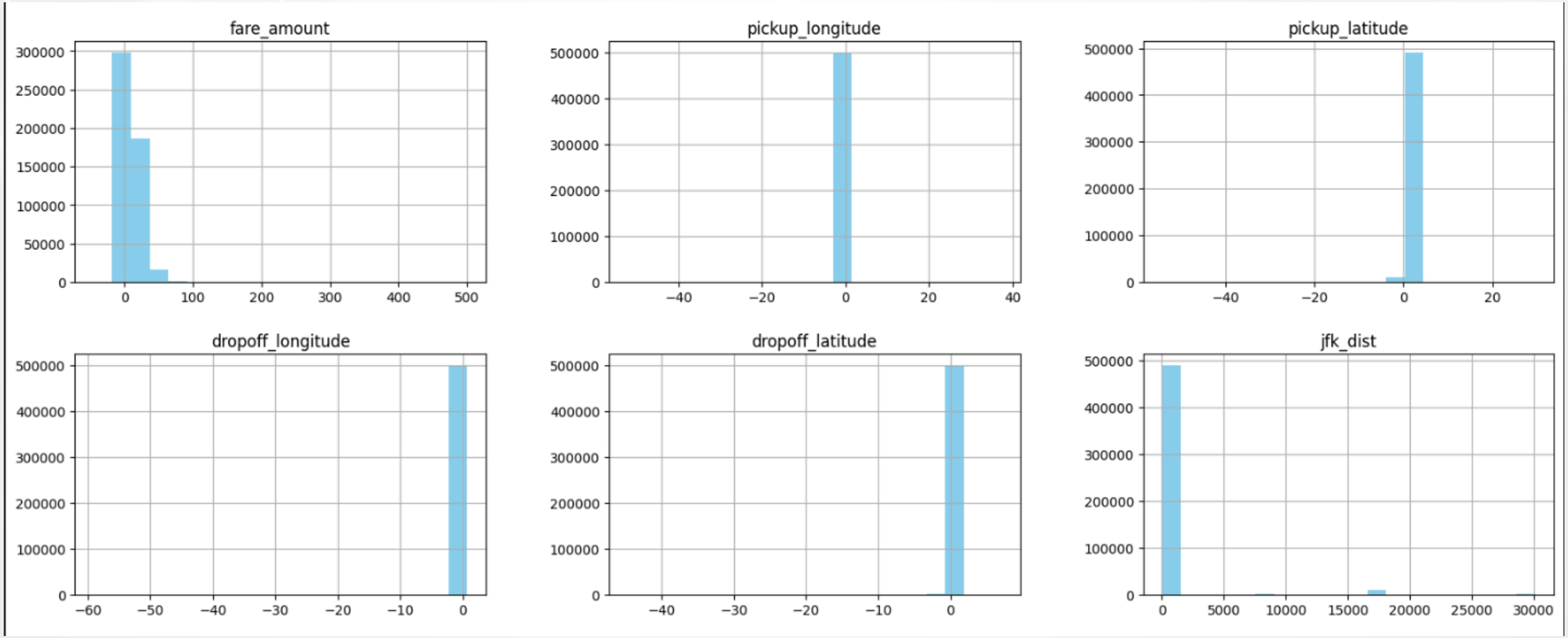
# 4. NUMERICAL DATA

**Histogram of Numerical Features**



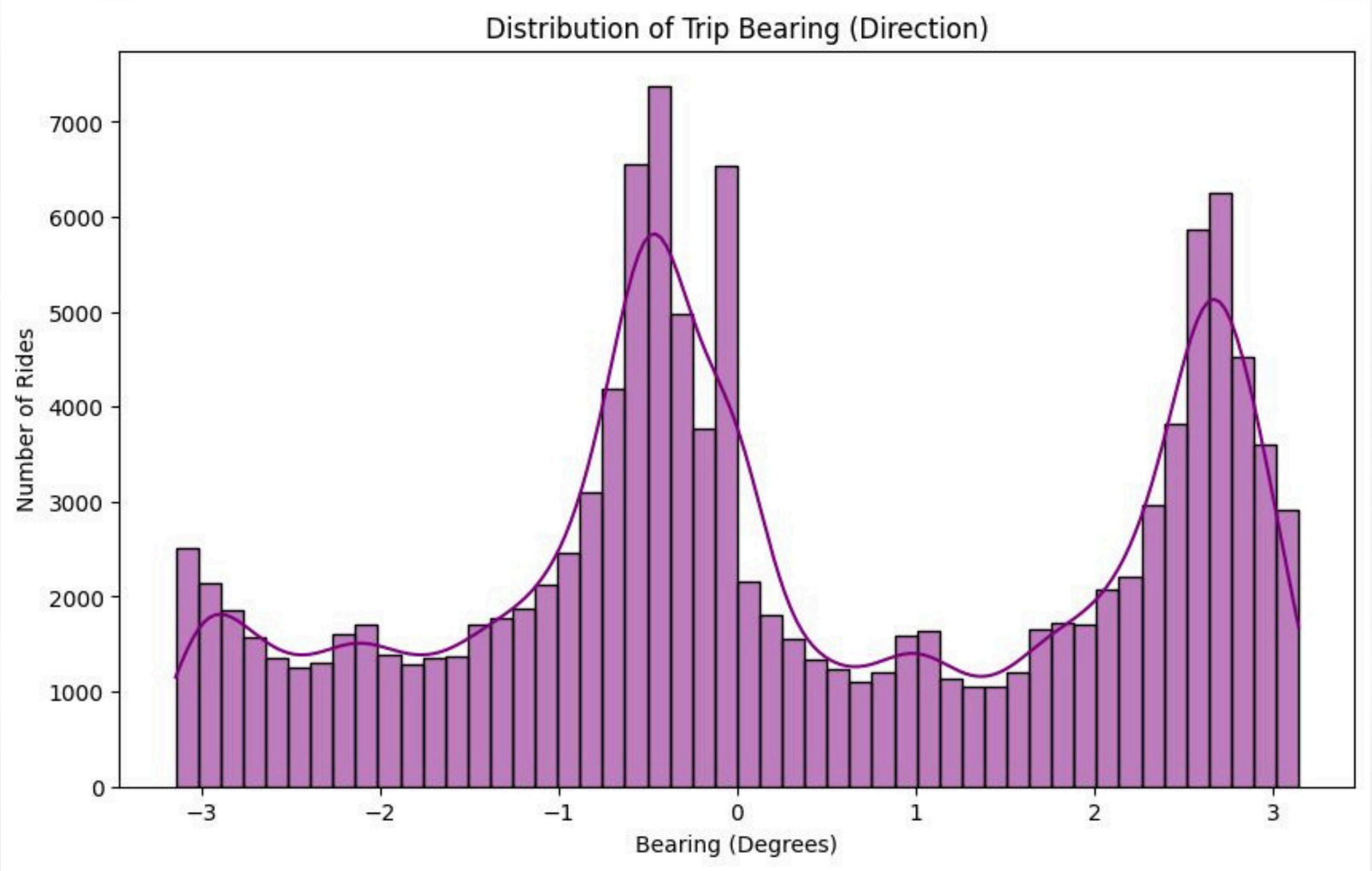
# 4. NUMERICAL DATA

**Histogram of Numerical Features**



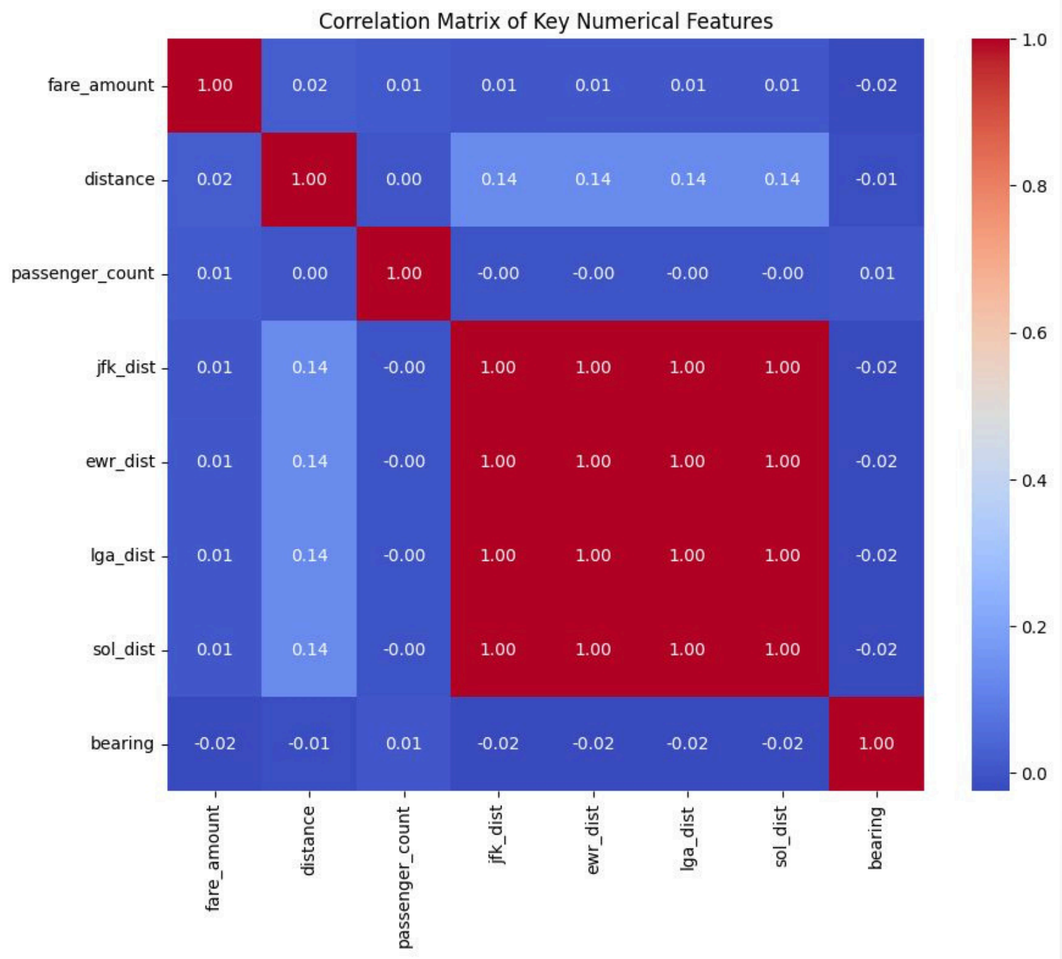
# 4. NUMERICAL DATA

**Histogram of Numerical Features**



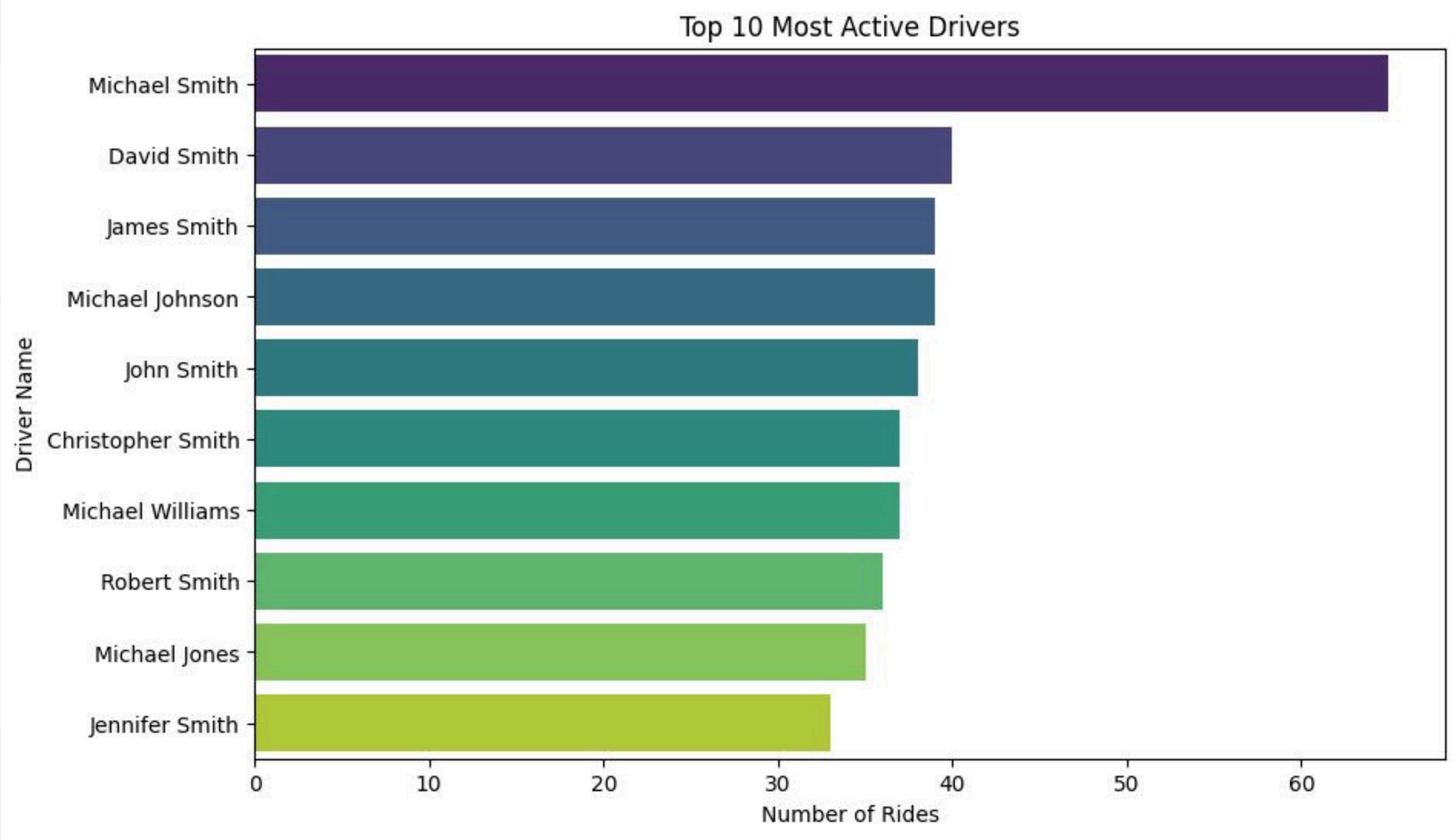
# 4. NUMERICAL DATA

**Heatmap for Numerical Features**



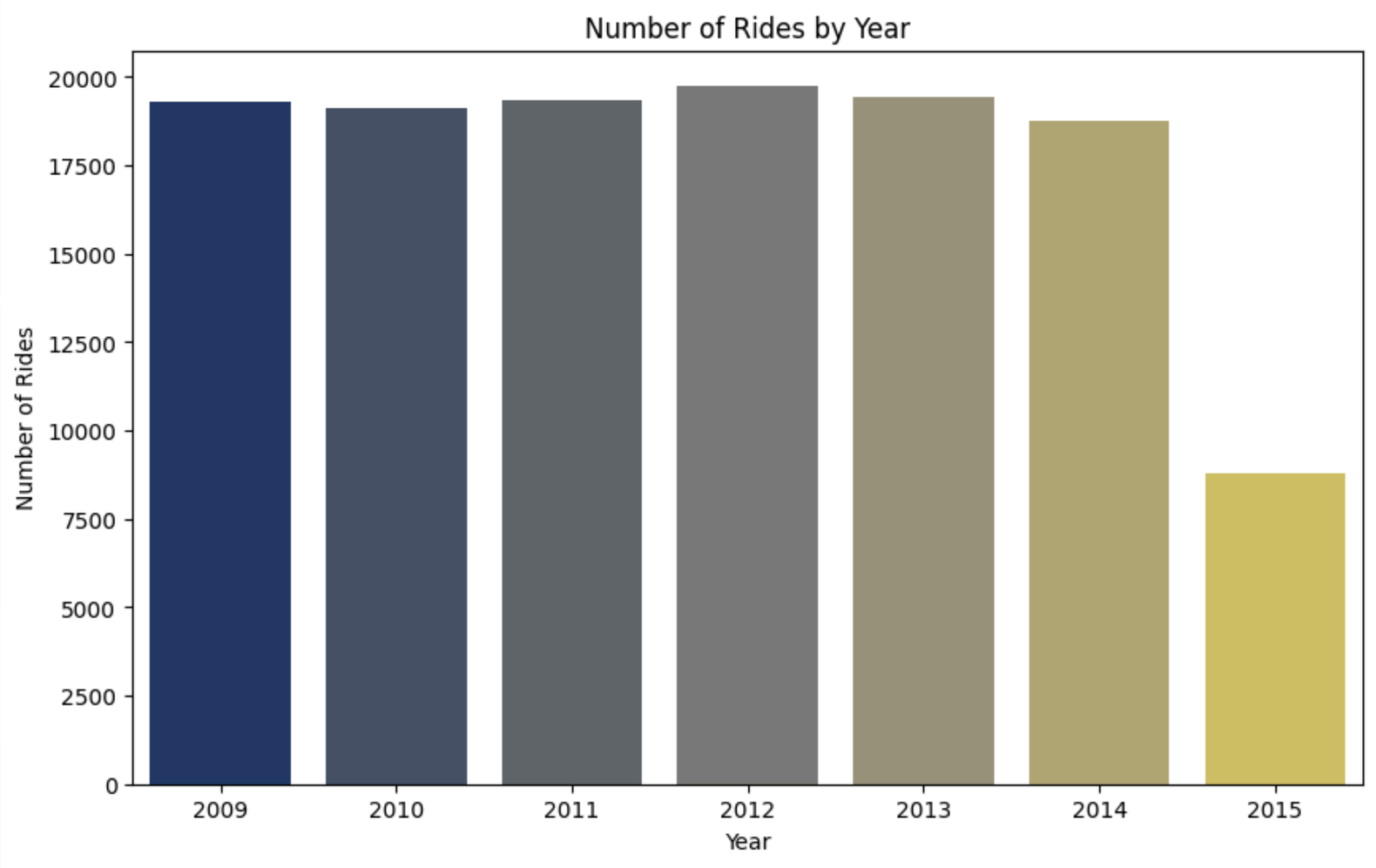
# 4. NUMERICAL DATA

**Top Drivers**



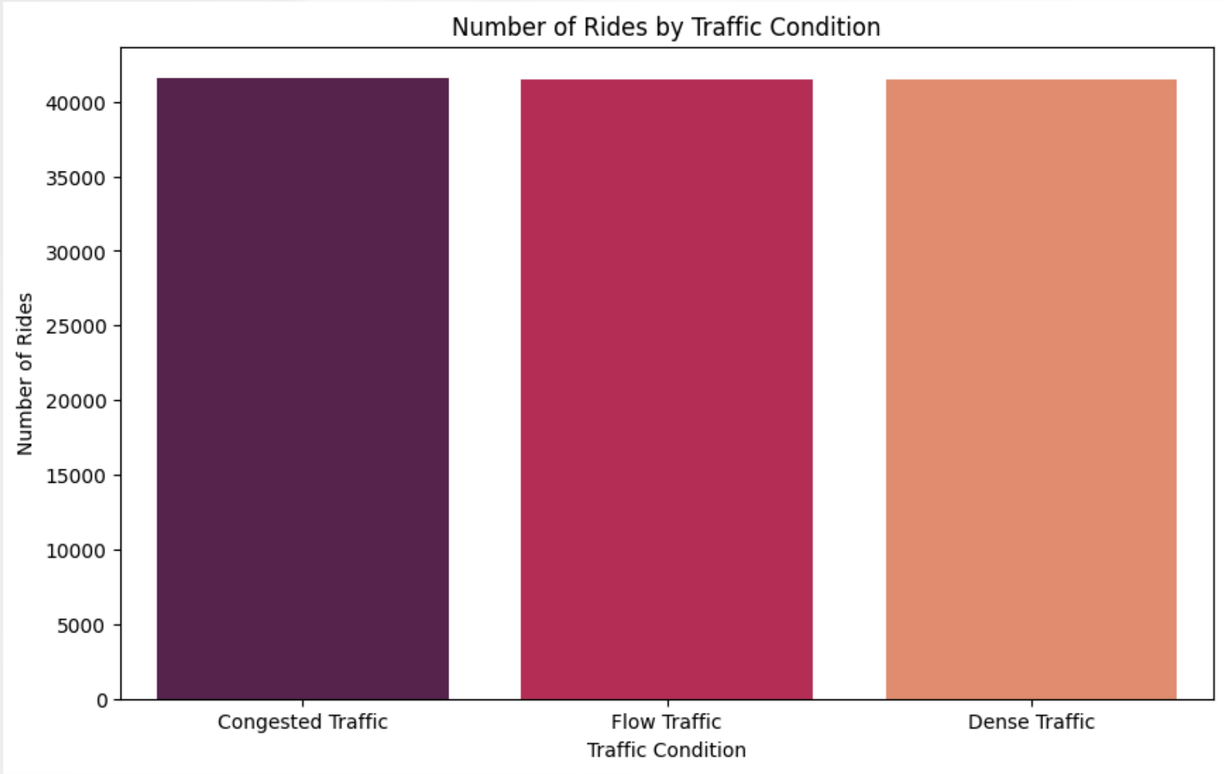
# 4. NUMERICAL DATA

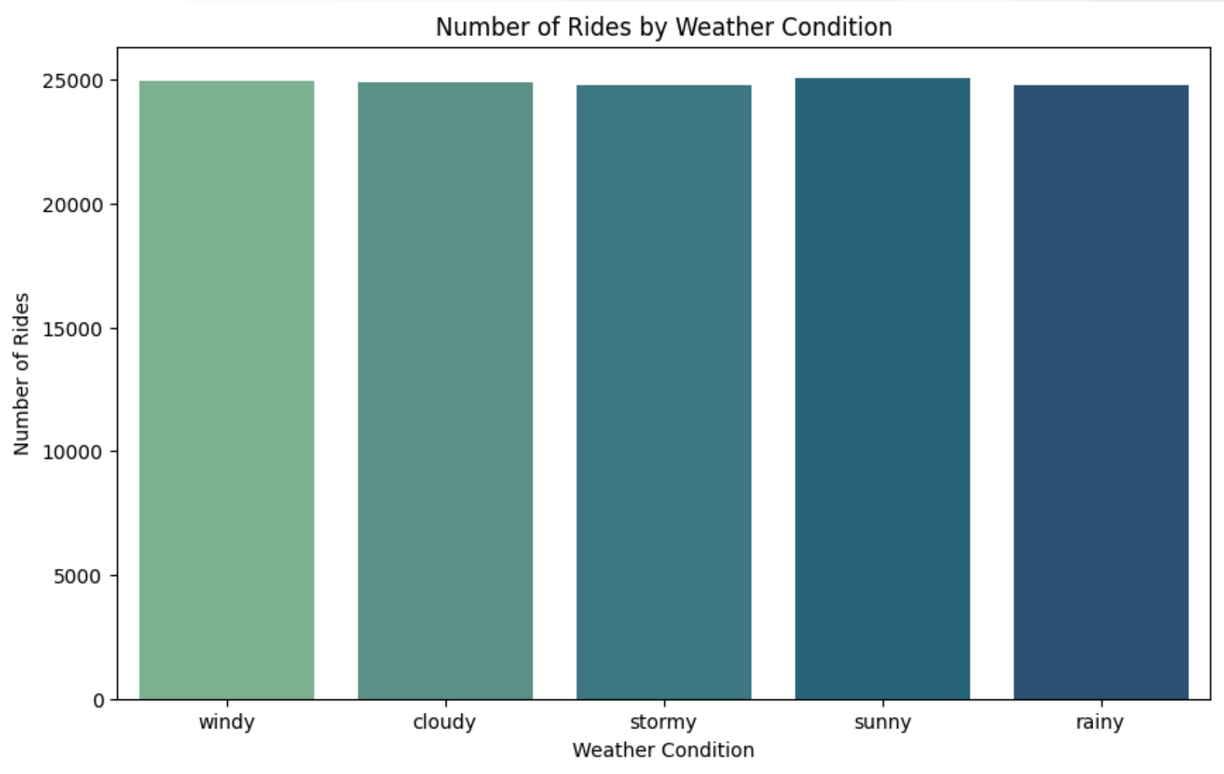
**Rides per year**



# 4. NUMERICAL DATA

**Rides per traffic conditions and weather conditons**

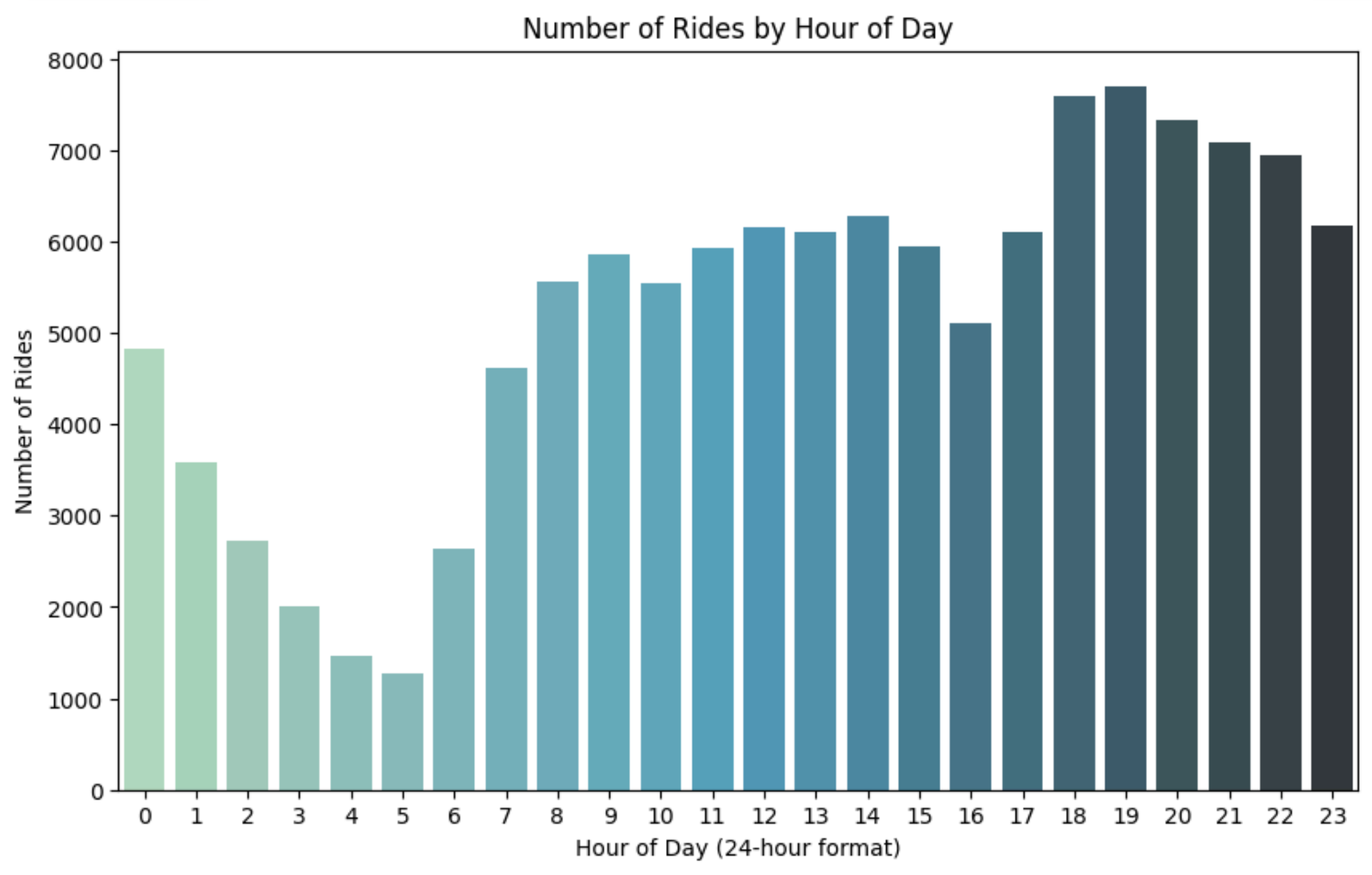


# 4. NUMERICAL DATA

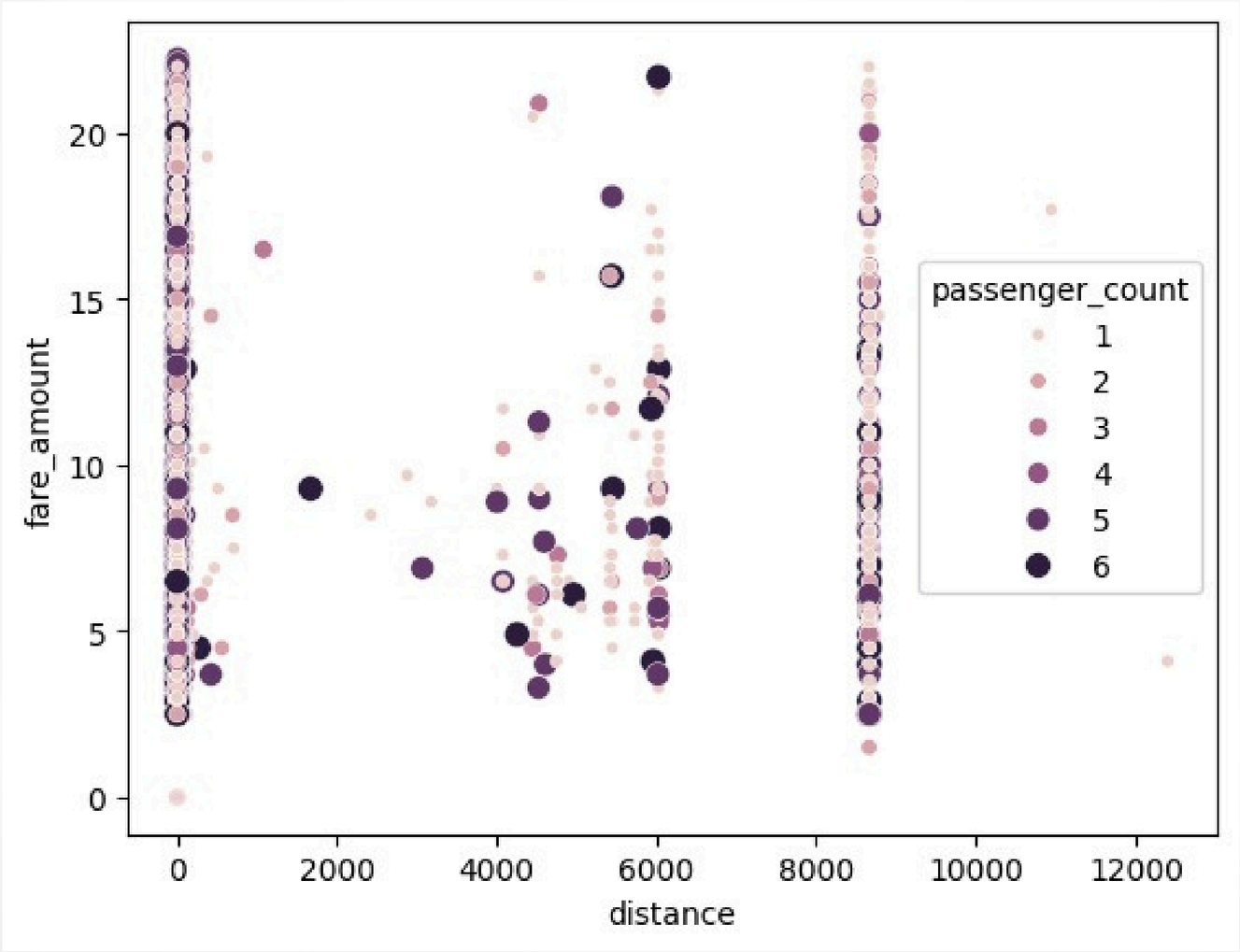
**Rides per time**

- **Number of Rides by days of week was consistent**
- **Number of rides per month of year was mostly similar**



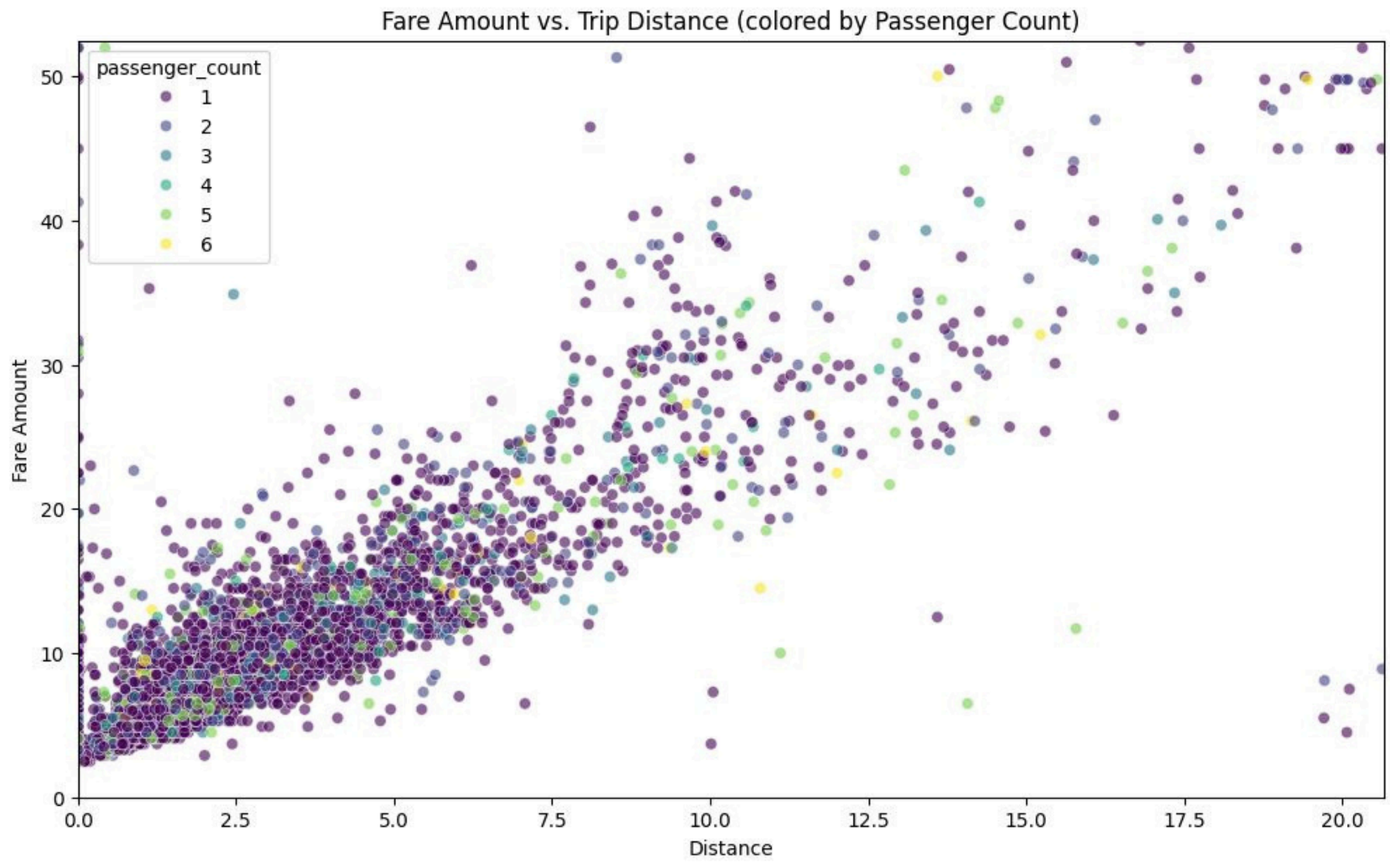
# 4. NUMERICAL DATA

**Relationship between Fare Amount and distance**



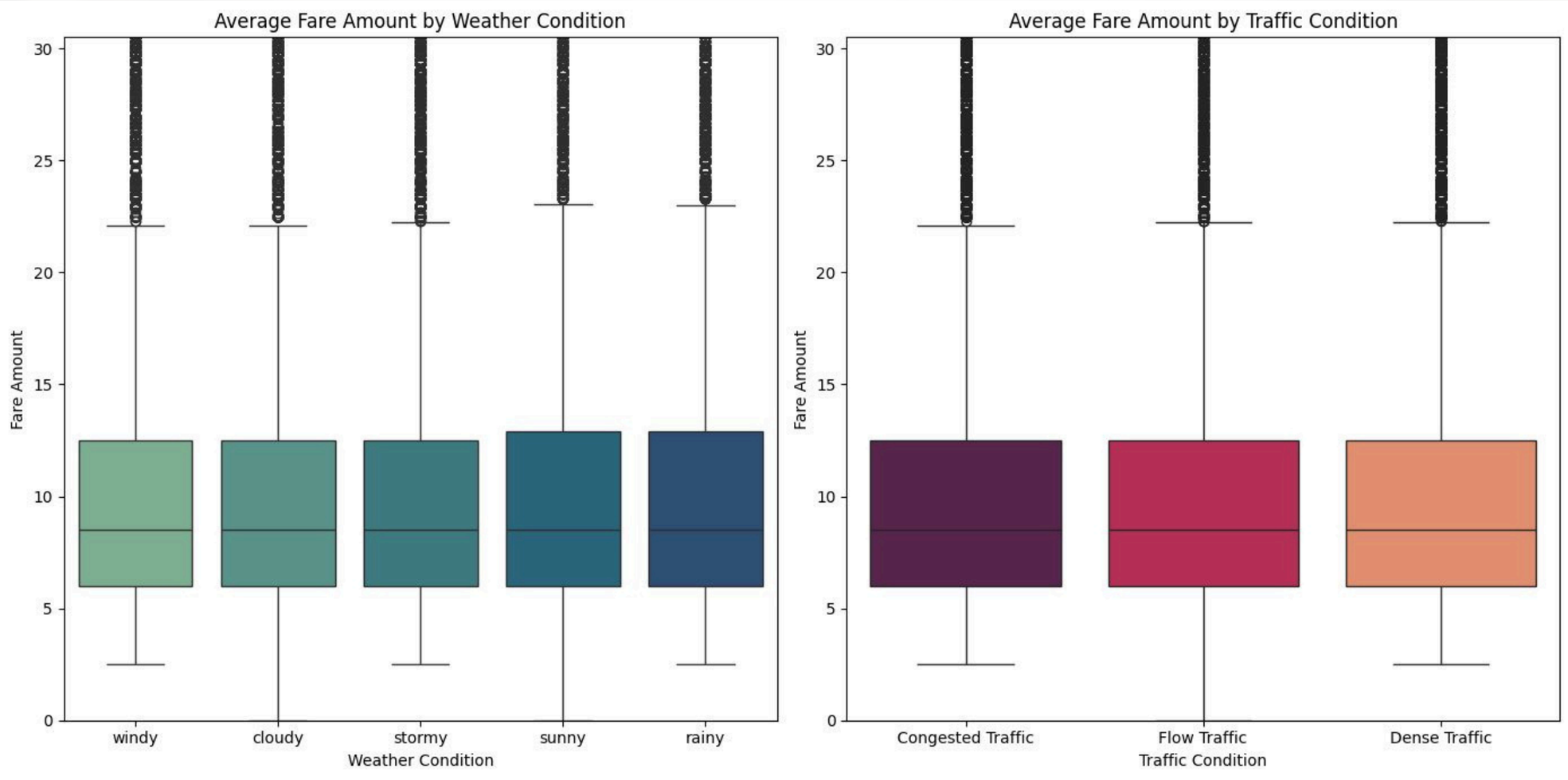
# 4. NUMERICAL DATA

**Relationship between Fare Amount and distance**



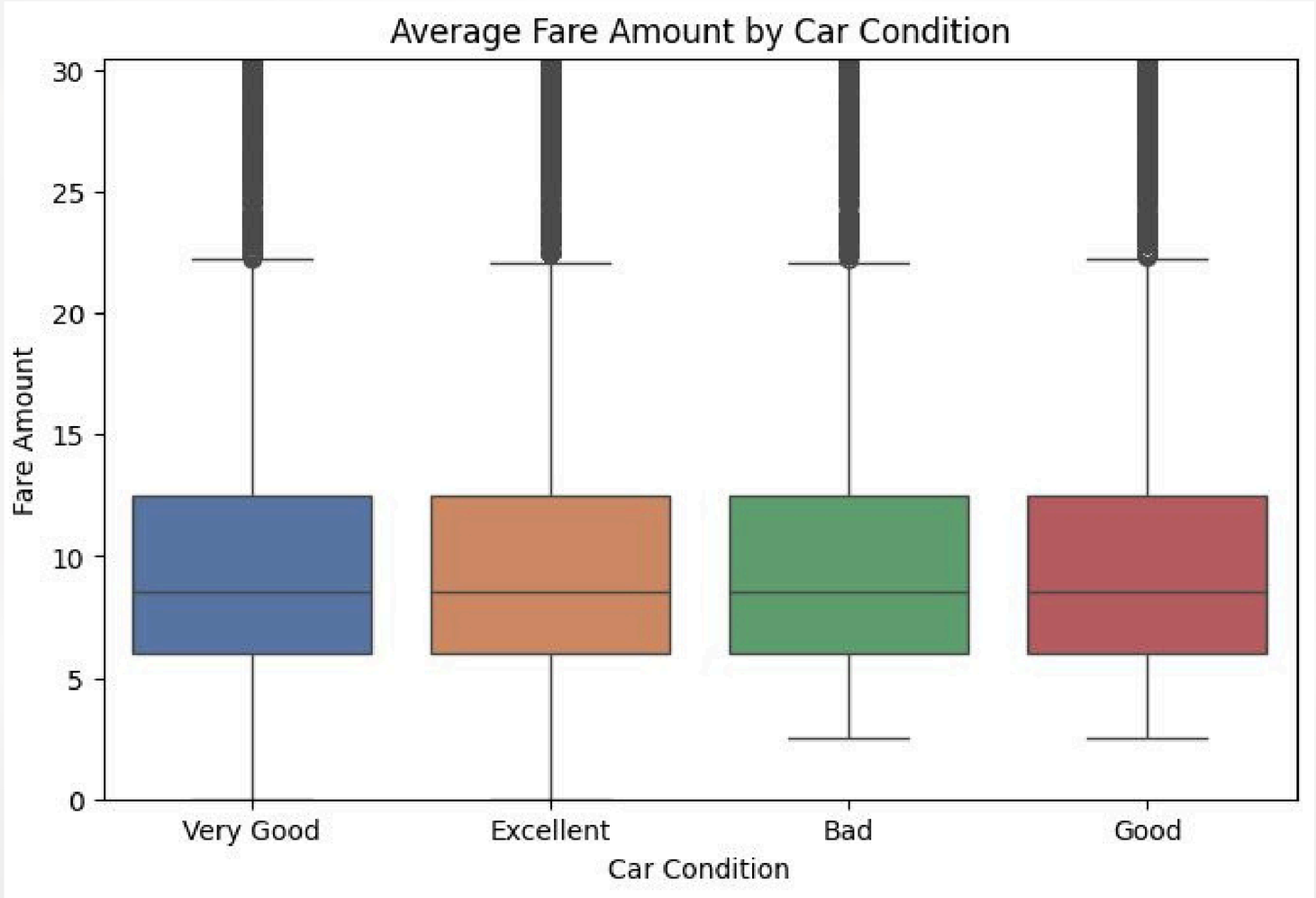
# 5. FARE AMOUNT RELATIONSHIP WITH EXTERNAL FACTORS

**Relationship between Fare Amount and Weather or Traffic Conditions**



# 5. FARE AMOUNT RELATIONSHIP WITH EXTERNAL FACTORS

**Relationship between Fare Amount and Car Condition**



# THANK YOU