Unit 5: Exploring Data Distributions in the Bike Ride Dataset

Student's Name: Avinash Bunga

Course: CIS621 Data Analysis for Business Analytics

Professor: Robert Kao

Date: 09.20.2023

Institution: Park University

Exploring Data Distributions in the Bike Ride Dataset

Introduction

Understanding data distributions is a crucial aspect of data analysis. In this assignment, I delve into a dataset containing information about bike rides. The objective is to uncover insights into the distribution of time and ride durations.

Methods

1. **Average Hour and Duration**

- To determine the average hour at which bike rides start and their average duration, I conducted the following analyses:
    - I extracted the hour component from the "start_time" field to calculate the average starting hour.
        - *(Formula Used =HOUR(D2)),* This will change all the hours into numbers from the "start_time" column
    - Utilising the "duration" field, I computed the average ride duration.

Results

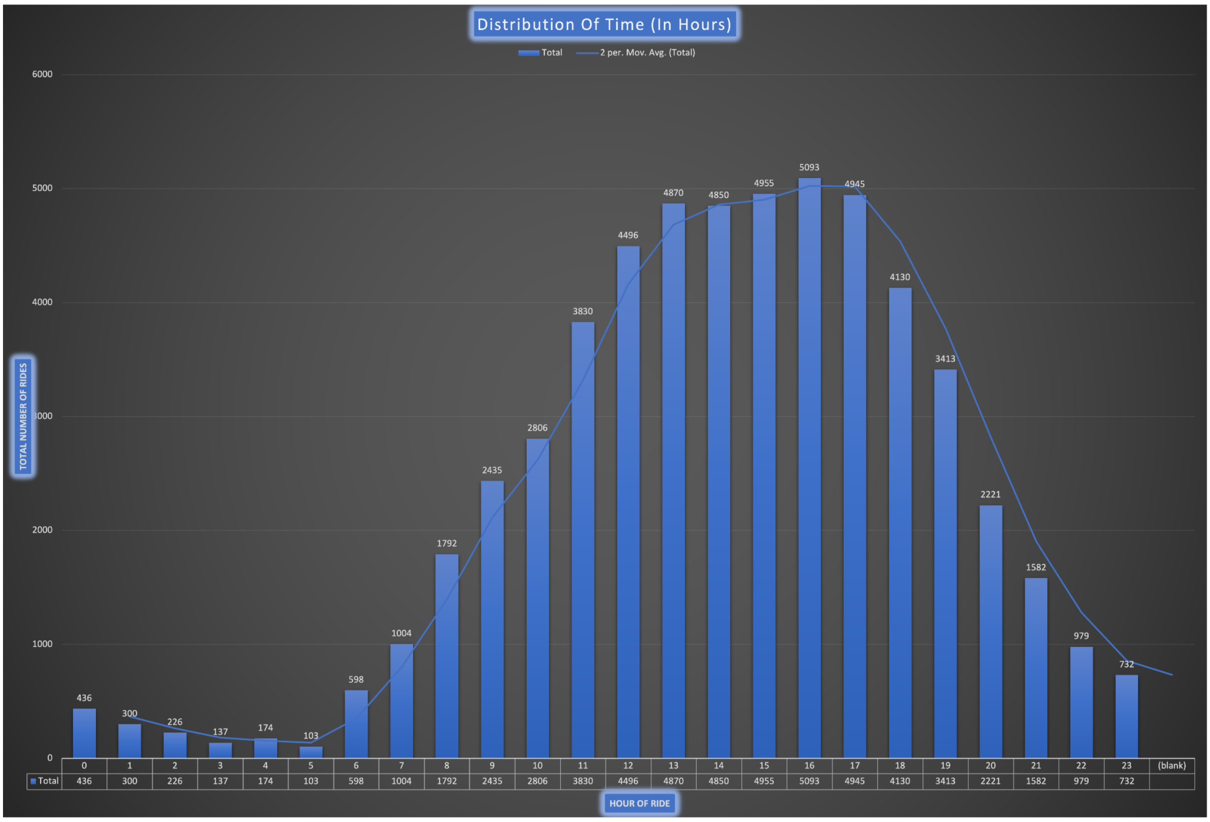
1. **Average Hour of Bike Rides**

- The dataset indicated an average starting hour for bike rides as 14.25 (2:15 PM). However, the histogram showed most rides occur around the 16th hour (4 PM), pointing towards a trend of late afternoon bike usage, possibly aligning with people commuting back from work or school.

2. Average Duration of Bike Rides

- The average duration was approximated as 42.8 minutes. However, a notable outlier exists: some rides have a duration of 1440 minutes (24 hours), suggesting potential errors, perhaps where rides were not correctly ended in the system.

Histogram 1: Distribution of Time



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	(blank)
■ Total	436	300	226	137	174	103	598	1004	1792	2435	2806	3830	4496	4870	4850	4955	5093	4945	4130	3413	2221	1582	979	732	

The X-axis has the Hour of ride, and the Y-axis has the total number of rides.

Description: The histogram above represents the distribution of time (in hours) when bike rides start. It provides insights into the frequency of rides at different hours of the day.

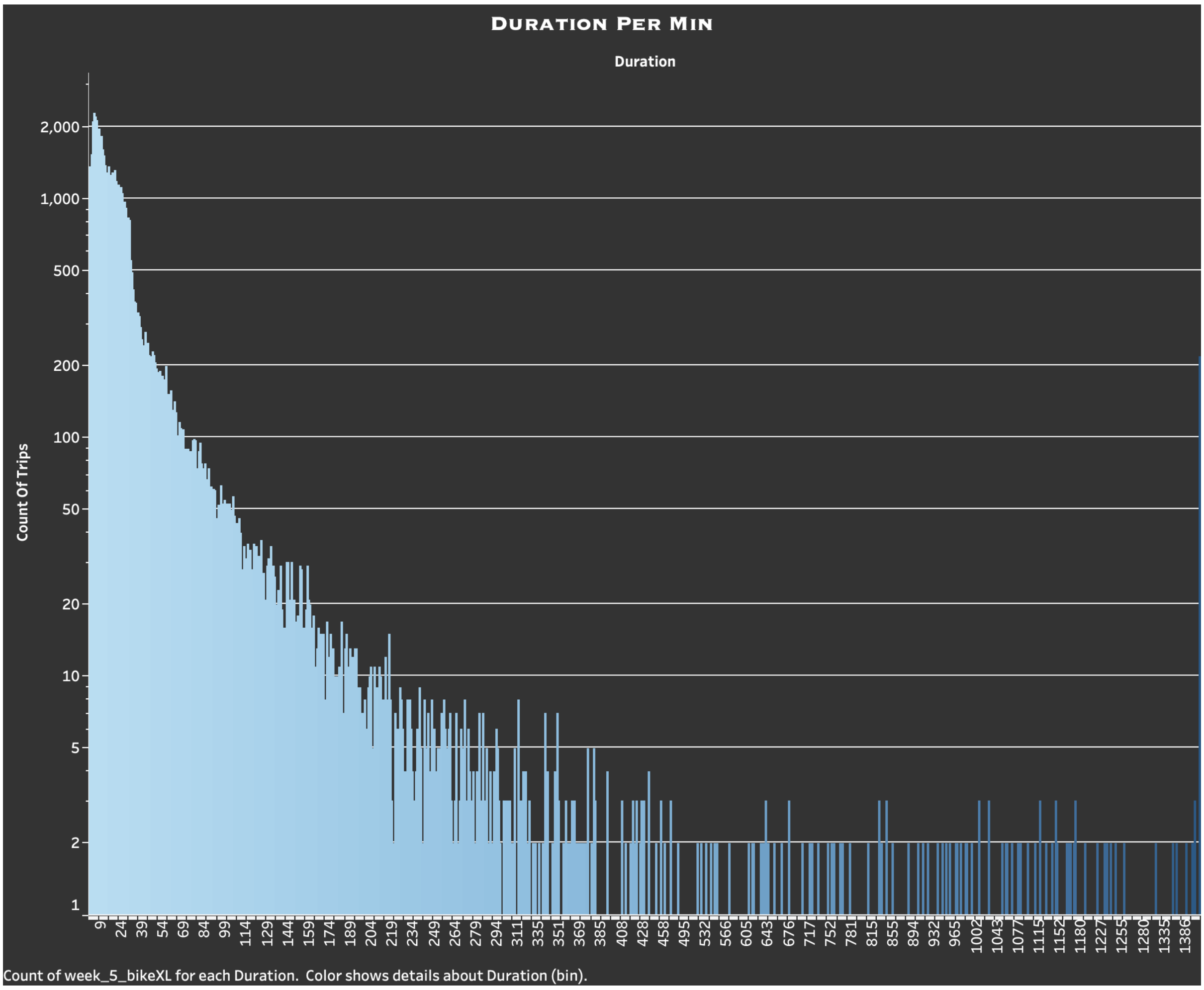
Interpreting the Distributions: My analysis of the time distribution revealed the following points

- We can see that most rides were taken in the 16th Hour, around 4 PM.
- This can be because students and the working class travel back home, which might create a spike in demand in that time zone.

Histogram 2

Duration Graphs: Three unique histograms present the distribution of ride durations:

1) **Duration Graph Divided into Minutes:** A granular perspective detailing exact ride counts per minute.

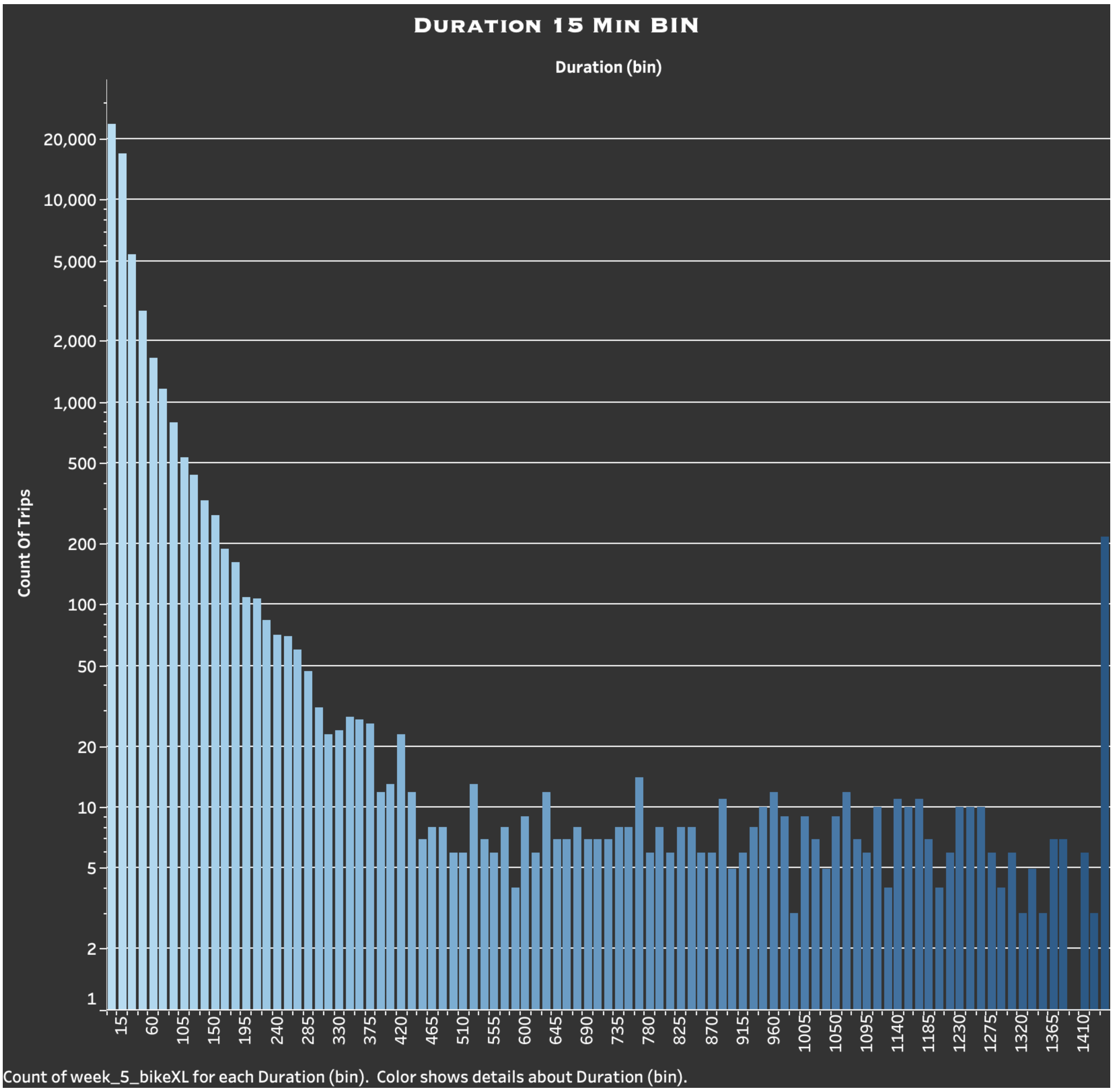


Count of week_5_bikeXL for each Duration. Color shows details about Duration (bin).

The X-axis has the Duration of rides in minutes, and the Y-axis has the count of rides.

2) **Duration Graph with 15-Minute Bins:** Balances detail with clarity by grouping durations in 15-minute intervals.

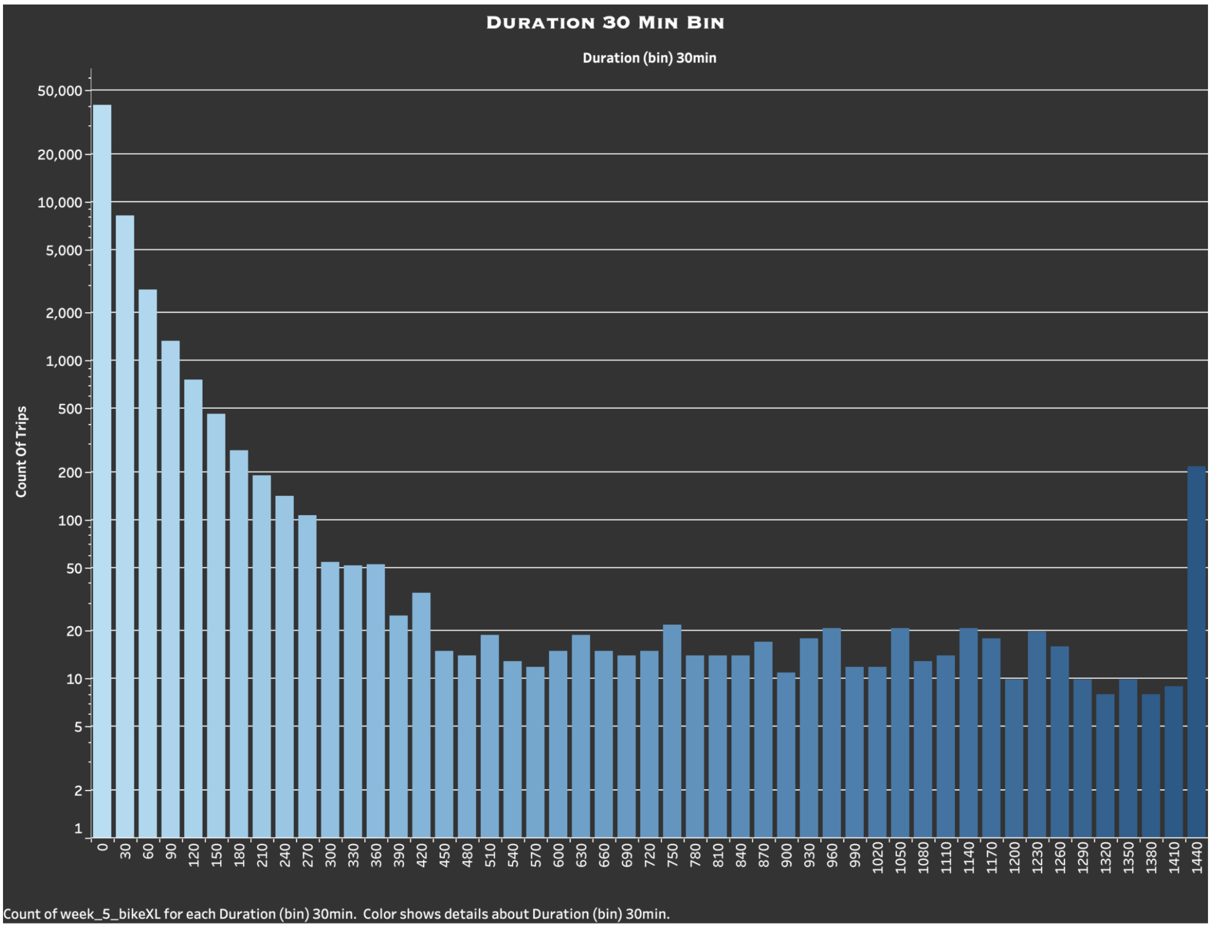
Histogram 3



The X-axis has the Duration of rides in 15 minutes bin size, and the Y-axis has the count of rides

3) **Duration Graph with 30-Minute Bins:** Provides a broad overview, ideal for discerning general patterns.

Histogram 4



The X-axis has the Duration of rides in 30 minutes bin size, and the Y-axis has the count of rides

Binning & Visualization:

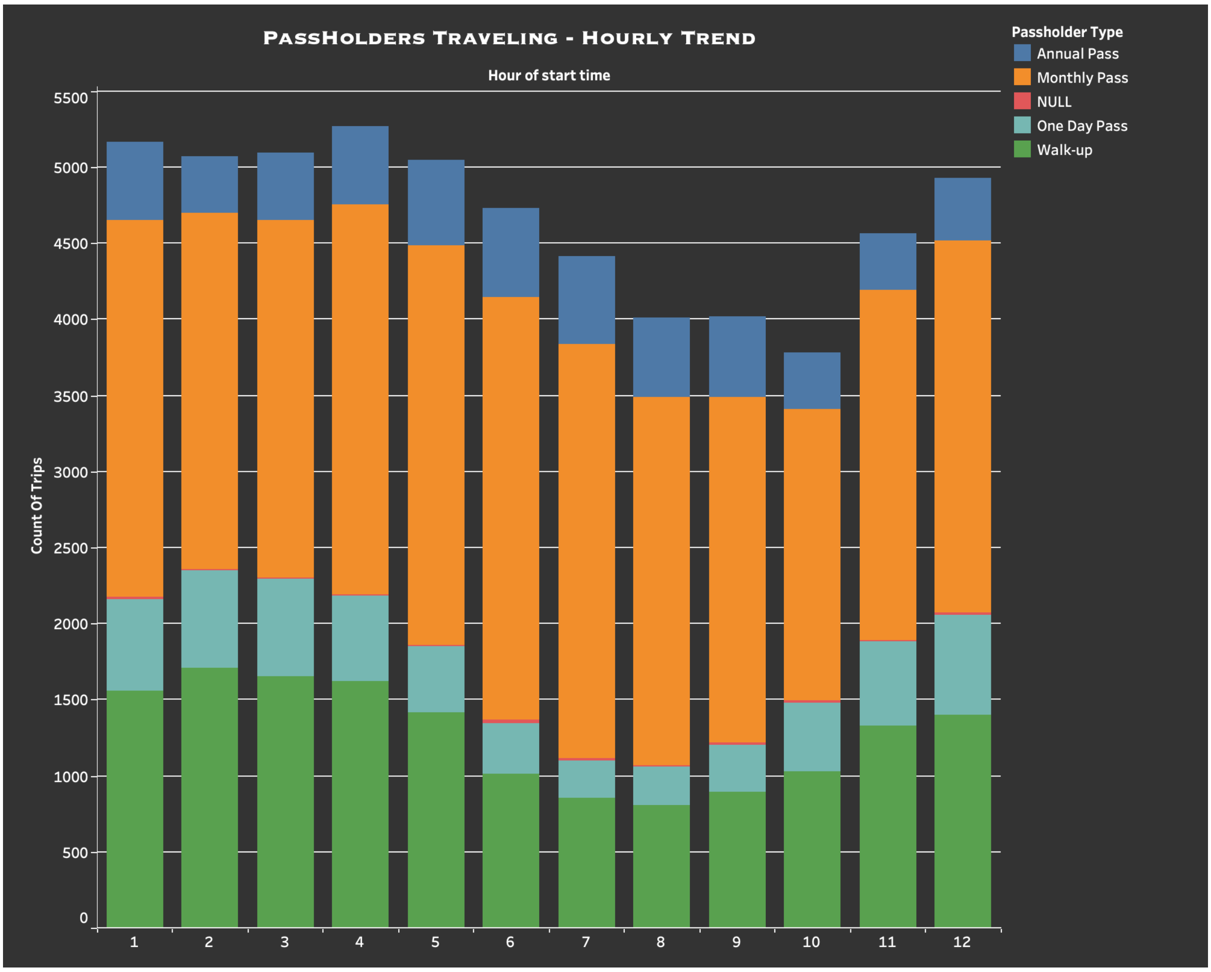
Binning, particularly when handling expansive data ranges, provides numerous visualization benefits:

- Clarity & Simplicity: Binned histograms simplify views, making patterns more evident.
- Outlier Management: Binning mitigates the visual impact of outliers.
- Comparative Analysis: They facilitate easy comparative evaluations.

Dissecting Dataset Using the PassHolders Category:

PassHolders Traveling - Hourly Trend: This segment analyzes the hourly bike ride trends based on different passholder types. A dedicated graph provides a visual representation.

Histogram 5



The X-axis has the Hours 1 to 12, and the Y-axis has the count of rides

Analyzing the given data, several observations arise:

- **Annual Pass holders** consistently use the service, peaking during the early hours and the late afternoon.
- **Monthly Pass holders** depict the highest numbers, indicating a large regular user base, with peaks during commuting hours.

- **One Day Pass and Walk-up users** have varied patterns, likely representing tourists or occasional riders. Their usage tends to peak around midday, perhaps indicating leisurely rides.
- There's a minimal count under the **'NULL' category**, suggesting potential data entry errors or unregistered users.

The graph aids in identifying trends and drawing comparisons between different pass holder types, providing deeper insights into hourly user preferences.

Interpreting the Distributions

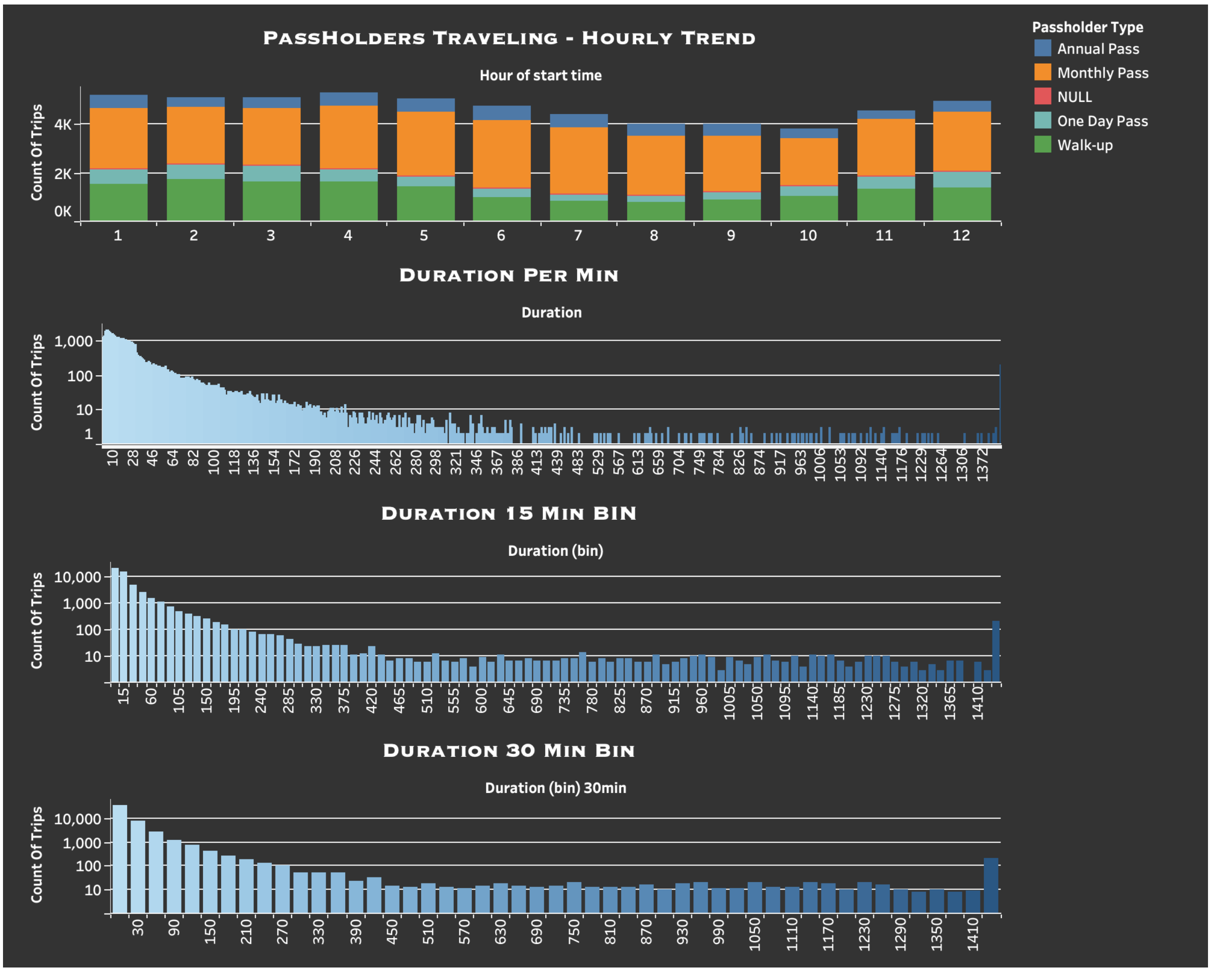
- We now understand that most of the rides are under 30 Minutes.
- Shorter rides are preferred than the longer ones.
- This demand is primarily due to the people using short bike rides to places where they can access much faster public transportation.
- As stated above, the data is right-skewed, with most rides being shorter, while a few are much longer.

Conclusion:

Unraveling data distributions is foundational for data analysis. This exploration has granted invaluable insights into the nuances of the bike ride dataset. While average values are illuminating, they must be evaluated in tandem with the distribution's shape and specific attributes for a holistic understanding.

Of note, rides with a 24-hour duration might indicate system errors, such as rides not being accurately terminated. It underscores the need for data cleansing and verification in analytics.

Histogram 6



The dashboard images attached highlight the discussed distributions and patterns, presenting a visual summary of our findings. Utilizing different bin sizes, from minute-wise details to broader 30-minute intervals, provides a panoramic view of ride durations, ensuring both detail-oriented and summarized perspectives are available for analysis.

Reference

wikiHow. (2023, May 24). *How to make bar graphs: 6 steps (with pictures)*. https://www.wikihow.com/Make-Bar-Graphs