Dataset Analysis Report

Uploaded Dataset

August 11, 2025

Introduction

This report presents a preliminary analysis of a dataset containing information on road user involvement across various categories and genders for 7 different States/Union Territories. The dataset aims to provide insights into the distribution of individuals affected by road incidents, categorized by their mode of transport (e.g., Pedestrian, Two Wheelers, Cars) and gender. The analysis utilizes descriptive statistics to identify key trends and patterns within this sample.

Methods

The analysis commenced by assessing the dataset's structure, including its dimensions (rows and columns) and data types. A thorough check for missing values was performed across all columns. Subsequently, basic descriptive statistics (mean, standard deviation, min, max, quartiles) were computed for all numerical columns to understand the central tendency, spread, and range of values. Categorical columns were examined for unique entries and frequency. Due to the small size of the dataset (7 rows), the analysis focused on high-level averages and distributions rather than inferential statistics.

Executive Summary

- The dataset is relatively small, comprising 7 rows and 29 columns, with each row representing a unique State/Union Territory.
- There are no missing values across any of the columns, indicating a clean and complete dataset for the sampled entries.
- Data types are appropriate, with 'States/Uts' as an object and all numerical counts as integers.
- Two-wheelers consistently show the highest average number of affected individuals (mean total: 30.71), followed by Pedestrians (mean total: 21.0).
- Across almost all categories, males are significantly more affected than females. For instance, in Two-wheelers, the average male count is 28.14 compared to 2.57 for females; for Pedestrians, it's 16.43 (male) vs 4.57 (female).
- Categories such as Bicycles, Auto Rickshaws, Cars/Taxies/LMV, Trucks/Lorries, and Buses show lower average involvement numbers compared to Two-wheelers and Pedestrians.
- The 'Other non Motor vehicles(E-Rickshaw)' category shows zero affected individuals across all sampled States/UTs, for both males and females, suggesting either no incidents or no data captured for this category within the dataset's scope.
- The 'Bicycles - Female' and 'Buses - Female' and 'Others - Female' categories also show an average of zero affected individuals, indicating very low or no reported female involvement in these specific categories within the given data.

Data Cleaning Notes

The dataset was remarkably clean with no missing values detected in any column. All columns were of appropriate data types, requiring no type conversions. Column names are descriptive, although some are quite long; no renaming was deemed strictly necessary for this preliminary analysis. Therefore, no significant data cleaning or preprocessing steps were required before proceeding with descriptive analysis.

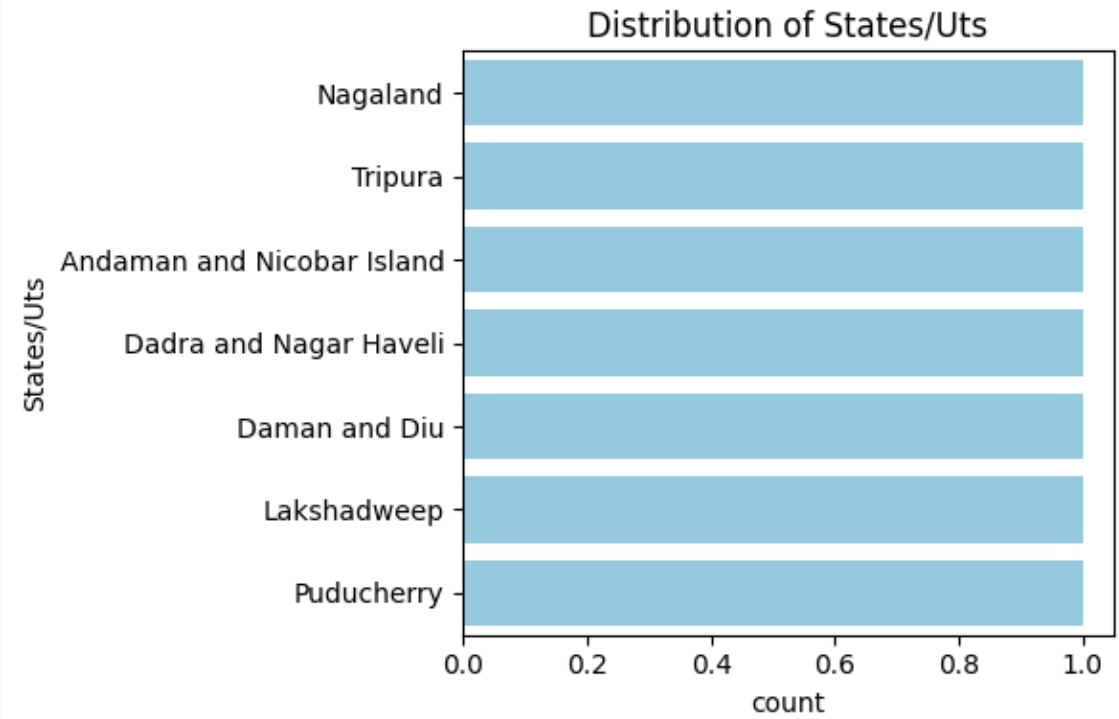
Statistical Summary

Variable	Mean	SE	MoE (95% CI)
Pedestrian - Male	16.43	7.6	14.9
Pedestrian - Female	4.57	2.14	4.19
Pedestrian - Total	21.0	9.64	18.9
Bicycles - Male	3.14	1.86	3.64
Bicycles - Female	0.0	0.0	0.0
Bicycles - Total	3.14	1.86	3.64
Two Wheelers - Male	28.14	12.03	23.58
Two Wheelers - Female	2.57	0.97	1.91
Two Wheelers - Total	30.71	12.97	25.42
Two Wheelers - Rank	30.0	1.7	3.34
Auto Rickshaws - Male	2.29	1.51	2.95
Auto Rickshaws - Female	0.43	0.43	0.84
Auto Rickshaws - Total	2.71	1.92	3.77
Cars, taxies Vans andLMV - Male	7.14	3.69	7.24
Cars, taxies Vans andLMV - Female	0.86	0.7	1.38
Cars, taxies Vans andLMV - Total	8.0	4.36	8.55
Trucks/Lorries - Male	2.71	1.69	3.31
Trucks/Lorries - Female	0.29	0.29	0.56
Trucks/Lorries - Total	3.0	1.95	3.83
Buses - Male	1.86	1.1	2.16
Buses - Female	0.0	0.0	0.0
Buses - Total	1.86	1.1	2.16
Other non Motor vehicles(E-Rickshaw) - Male	0.0	0.0	0.0

Other non Motor vehicles(E-Rickshaw) - Female	0.0	0.0	0.0
Other non Motor vehicles(E-Rickshaw) - Total	0.0	0.0	0.0
Others - Male	2.29	1.08	2.13
Others - Female	0.0	0.0	0.0
Others - Total	2.29	1.08	2.13

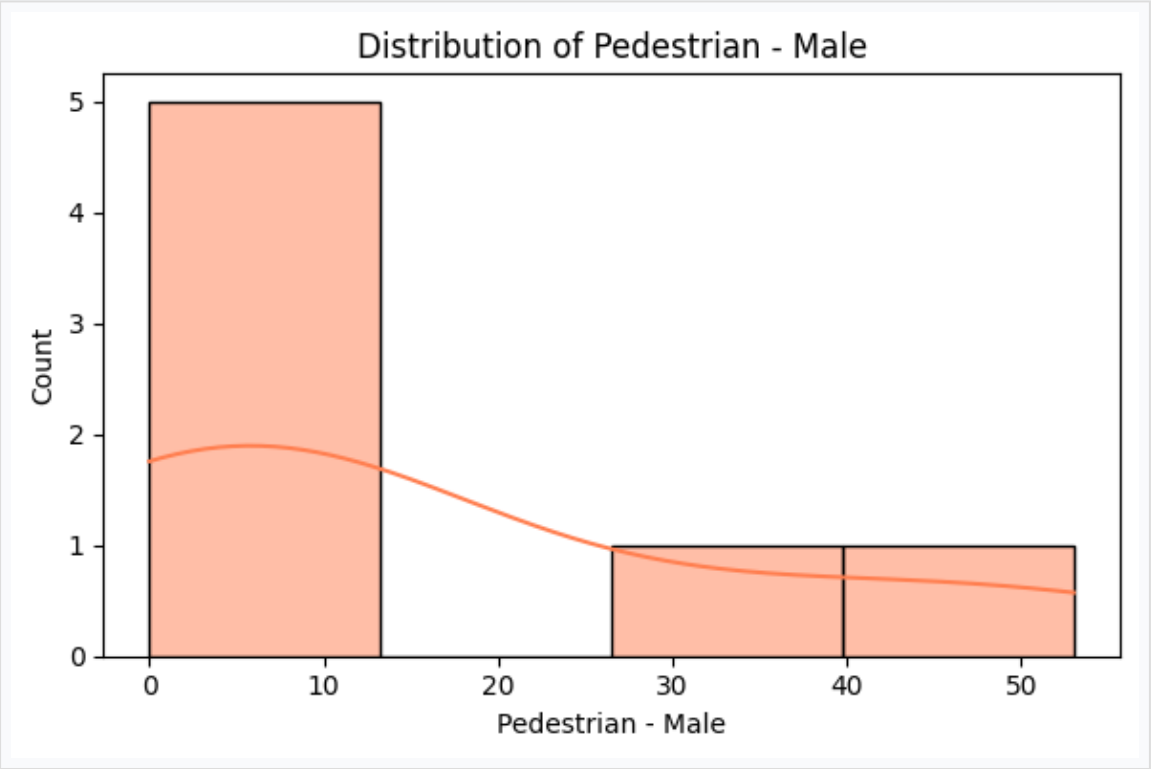
Results & Insights

Distribution: States/Uts



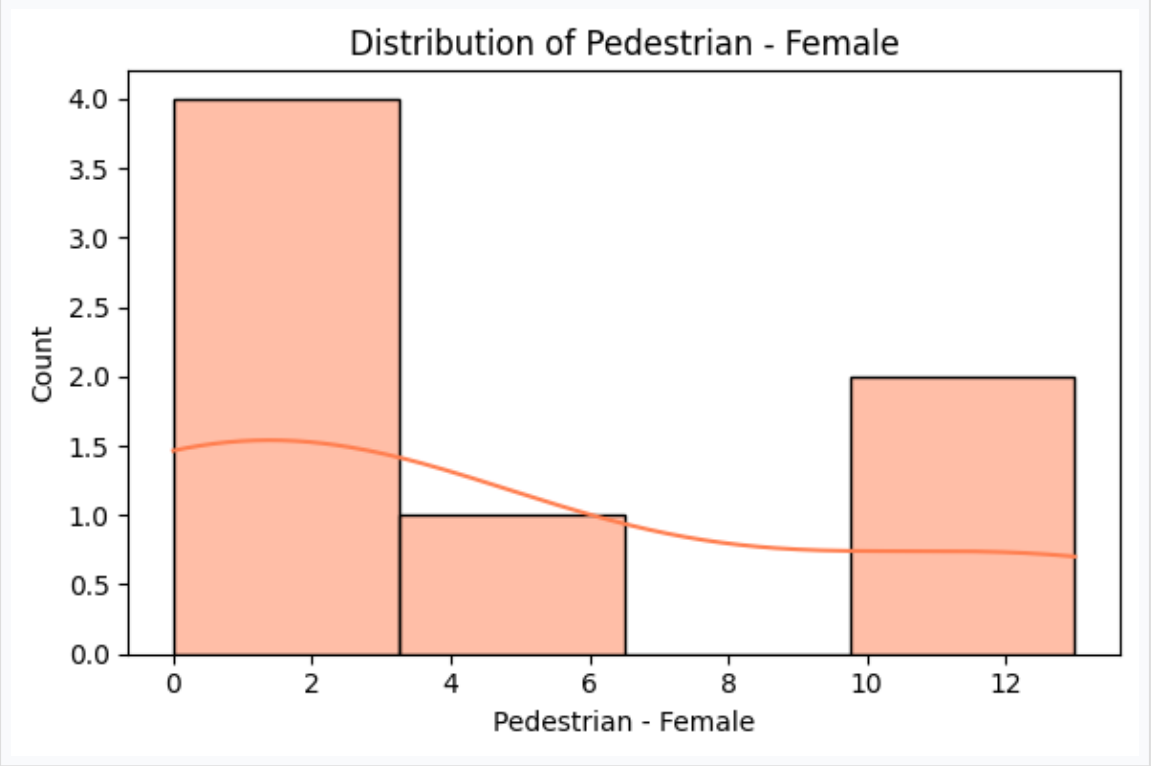
The most frequent category in States/Uts is 'Nagaland' (14.3%).

Distribution: Pedestrian - Male



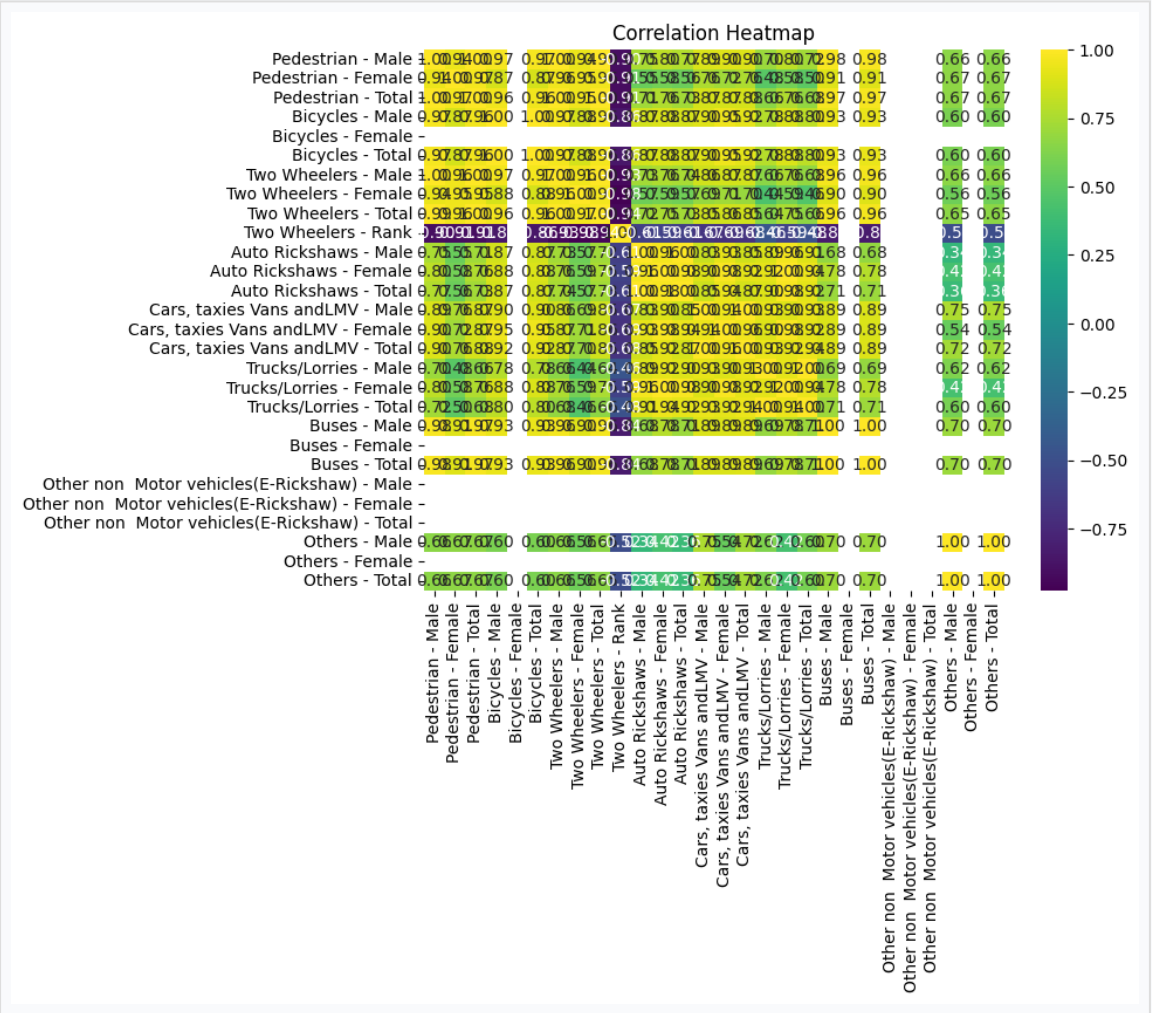
The average Pedestrian - Male is 16.43 with a spread of 20.11.

Distribution: Pedestrian - Female

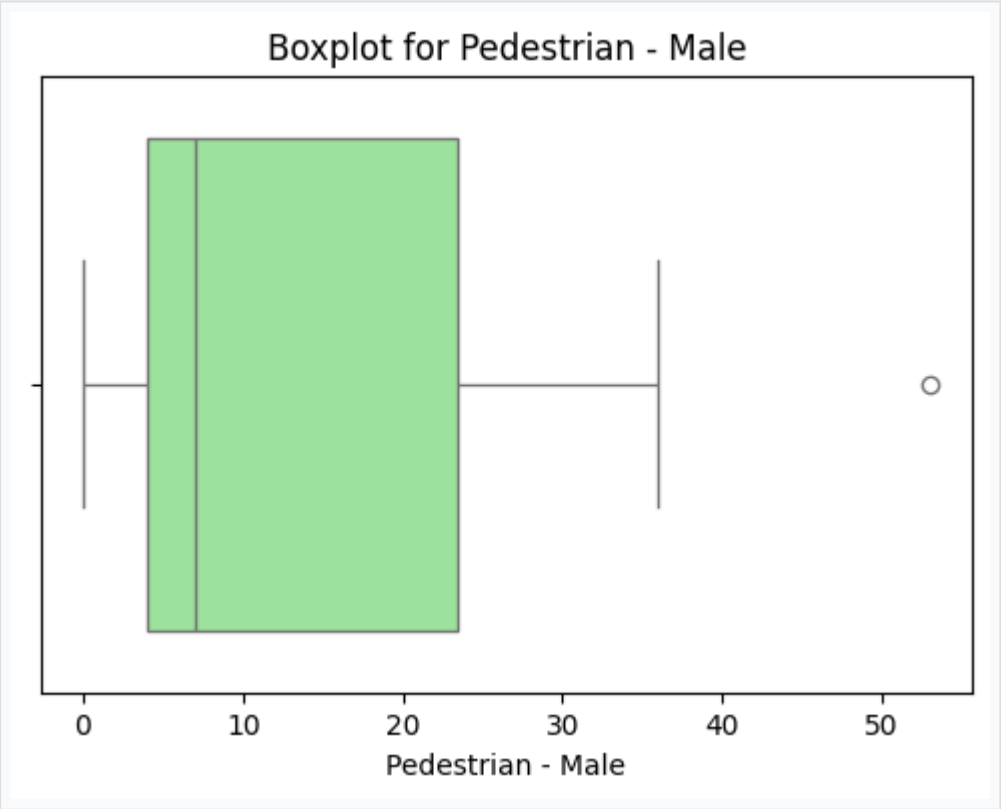


The average Pedestrian - Female is 4.57 with a spread of 5.65.

Correlation Heatmap

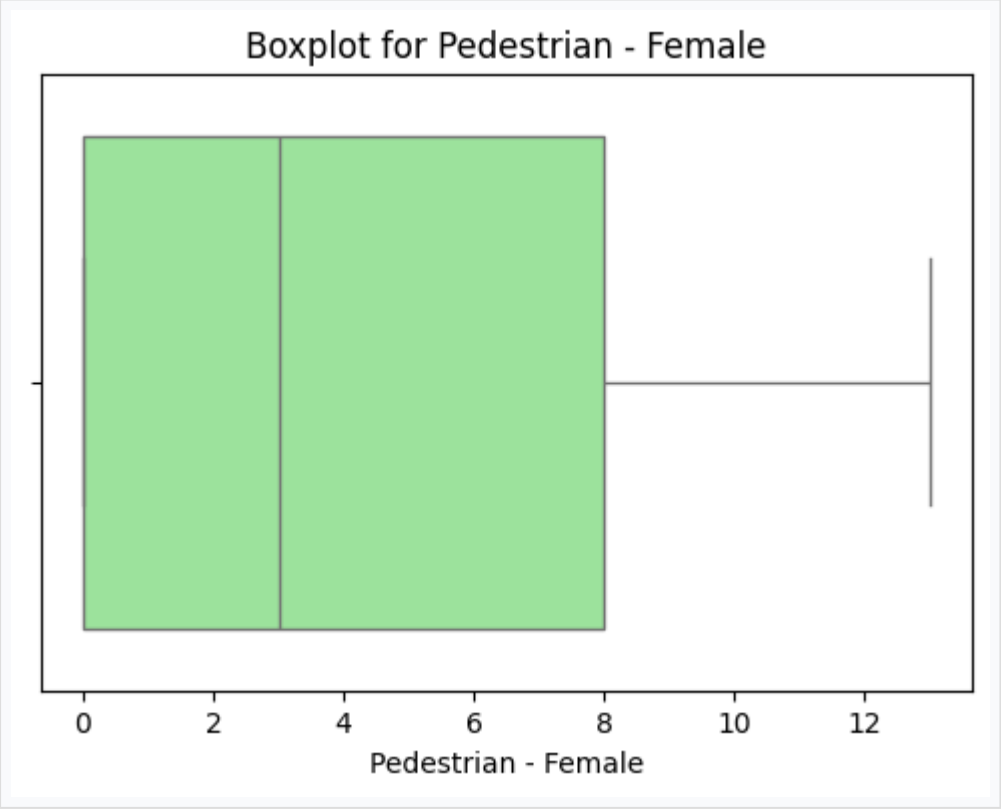


Boxplot: Pedestrian - Male



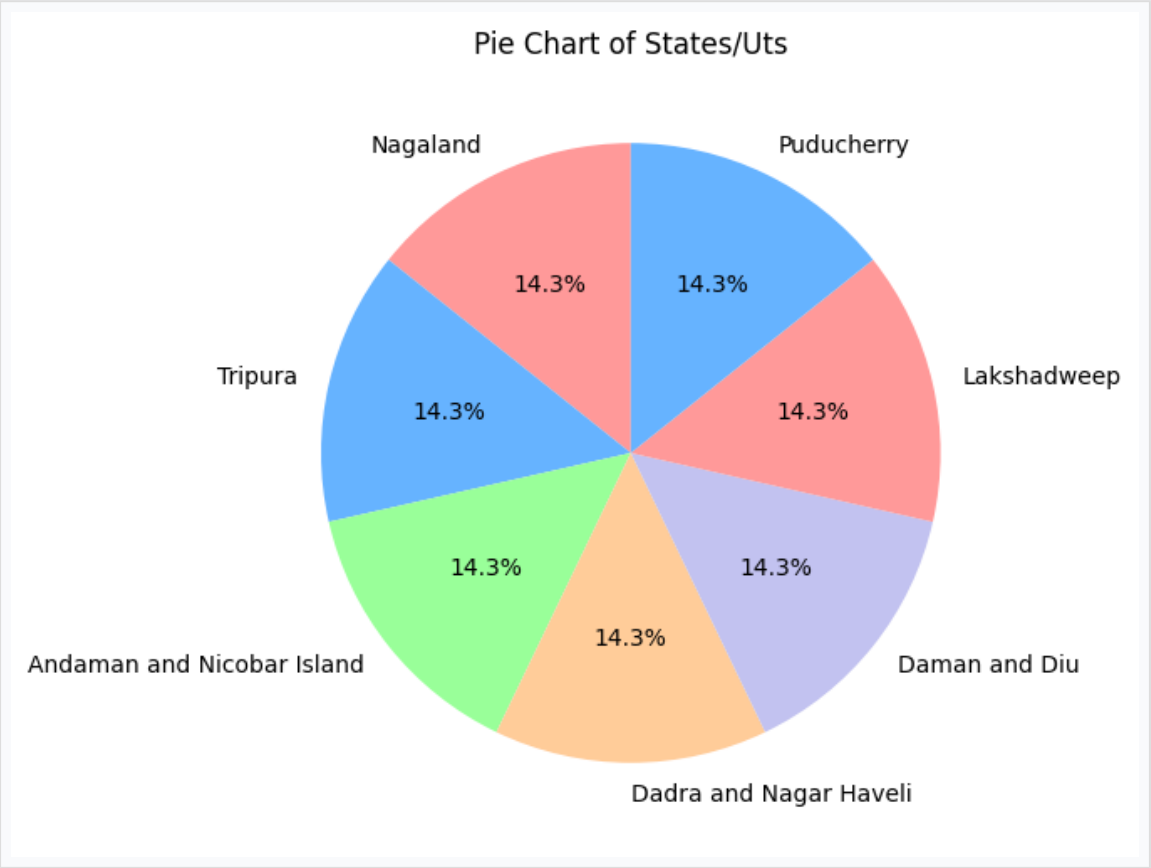
The average Pedestrian - Male is 16.43 with a spread of 20.11.

Boxplot: Pedestrian - Female



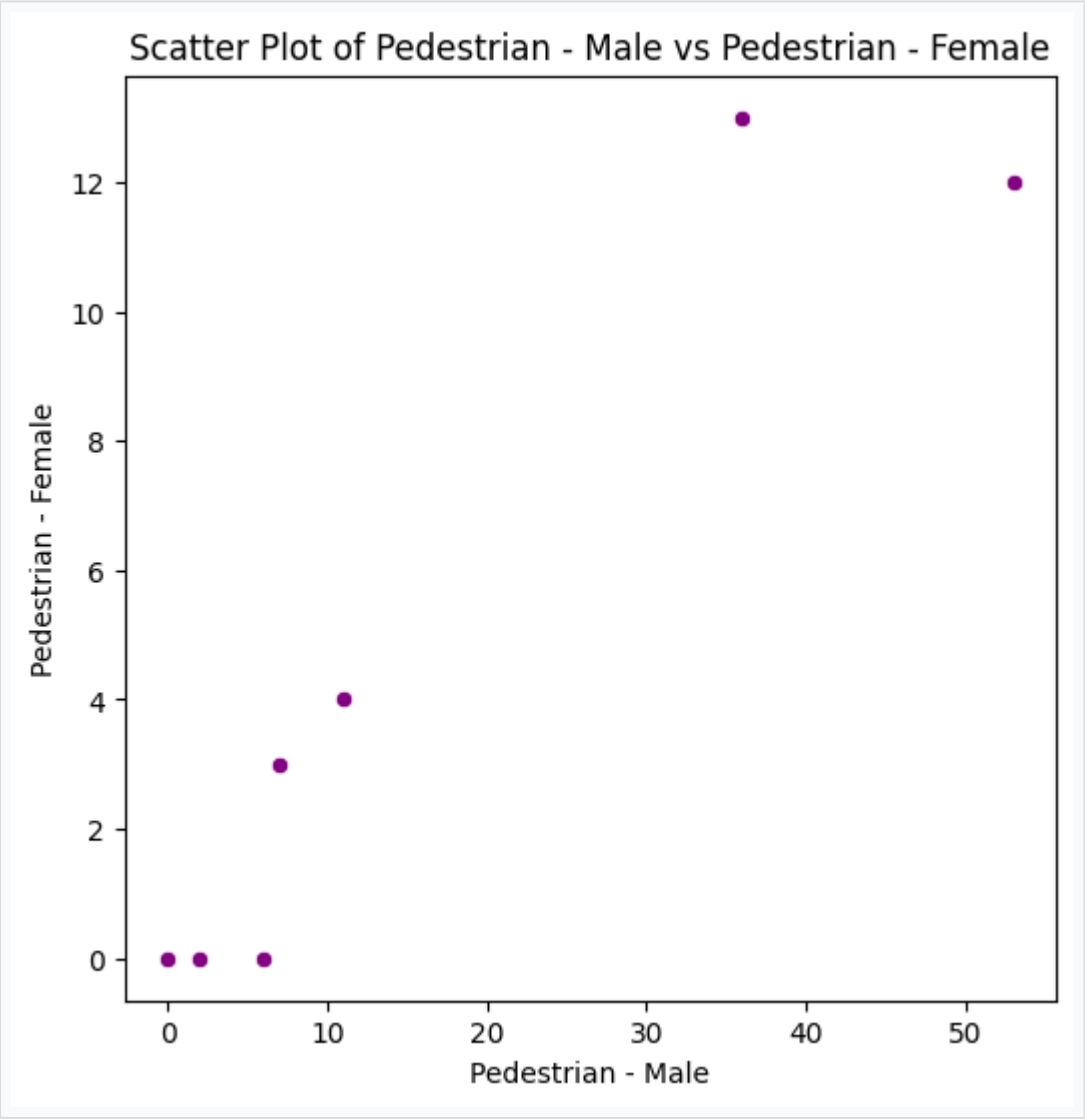
The average Pedestrian - Female is 4.57 with a spread of 5.65.

Pie Chart: States/Uts



The most frequent category in States/Uts is 'Nagaland' (14.3%).

Scatter Plot: Pedestrian - Male vs Pedestrian - Female



Conclusion & Recommendations

['The analysis of this dataset, albeit limited by its small size, highlights distinct patterns in road user involvement across the sampled States/UTs. Two-wheeler users and pedestrians appear to be the most frequently affected groups. A pervasive trend across nearly all vehicle categories is the disproportionately higher number of affected males compared to females. The absence of data for E-Rickshaws and minimal female involvement in several categories (Bicycles, Buses, Others) could indicate specific regional dynamics, reporting practices, or a true lack of incidents in these areas/demographics.']

- **Expand Data Coverage:** Obtain a larger dataset encompassing more States/Union Territories and potentially over multiple years to validate and generalize these preliminary findings.
- **Investigate High-Impact Categories:** Conduct deeper investigations into factors contributing to the high involvement of Two-wheelers and Pedestrians, such as infrastructure, traffic patterns, and user behavior.
- **Analyze Gender Disparity:** Research the underlying reasons for the significant gender imbalance in incidents. This could involve exploring differences in travel patterns, occupations, risk exposure, and road safety awareness among genders.
- **Clarify 'Two Wheelers - Rank':** Understand the methodology and significance of the 'Two Wheelers - Rank' column to integrate it into future, more comprehensive analyses.
- **Examine Zero-Entry Categories:** If expected to contain data, investigate why 'Other non Motor vehicles (E-Rickshaw)', 'Bicycles - Female', 'Buses - Female', and 'Others - Female' categories show zero affected individuals. This could be due to data collection limitations or actual absence of incidents.
- **Geo-spatial Analysis:** If location data were available, a geo-spatial analysis could pinpoint high-risk areas within states.