NYC Taxi Fare Prediction - Exploratory Data Analysis

This presentation delves into the analysis of NYC taxi fare trends, employing tools like Python, Pandas, and Seaborn to uncover insights.

Soumyajit Chakraborty
Presenter Designation

Essential Steps in Data Cleaning

Preprocessing Steps for Accurate Data Analysis

- **Removed Negative Fare Entries**

  Eliminated any fare entries with negative values to ensure data integrity.

- **Dropped Rows with Missing Values**

  Removed rows that had missing values to maintain a complete dataset.

- **Filtered Invalid Coordinates**

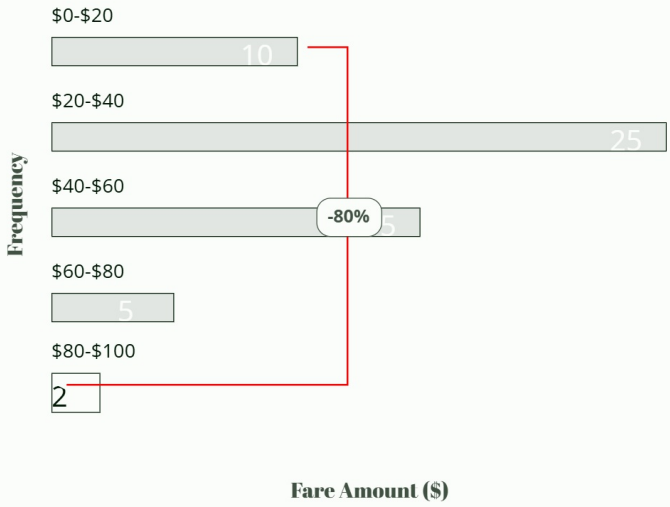
  Excluded latitudes and longitudes that fell outside the NYC boundaries.

- **Ensured Clean Data for Modeling**

  Ensured only valid and clean data is used in the modeling pipeline for accuracy.

Distribution of NYC Taxi Fare Amounts

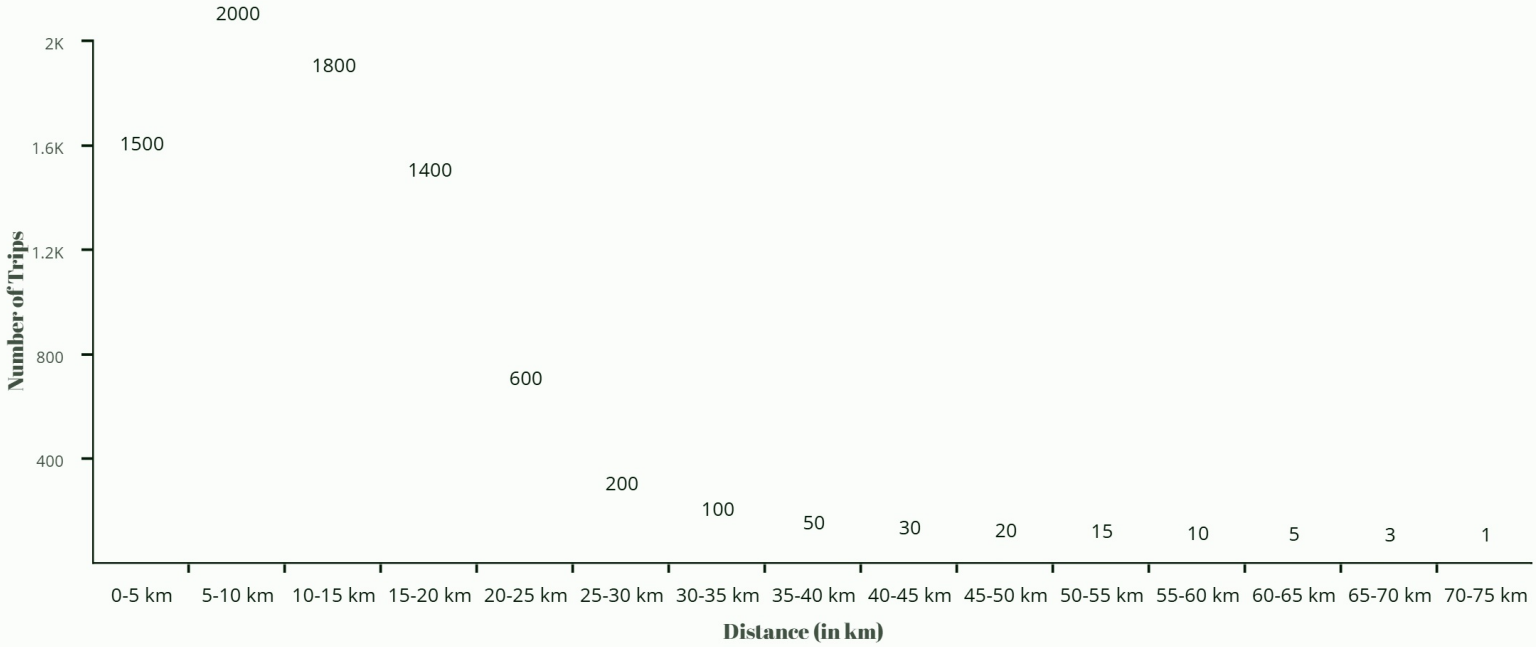
Analyzing the Cost Trends of Taxi Rides



- The histogram shows a long tail distribution of taxi fares.
  Description of a primary heading
- There are noticeable small spikes in fares between $40 and $60.
  Description of a primary heading
- Outliers have been trimmed to improve model performance.
  Description of a primary heading

NYC Taxi Fare Dataset

Analyzing Ride Distances in NYC

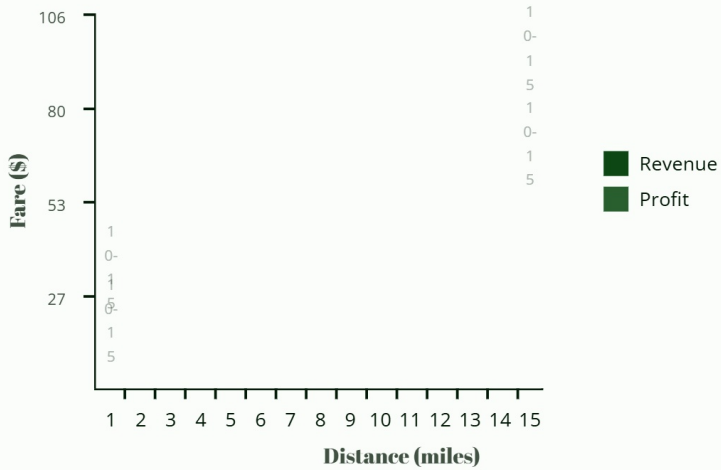


Source: Companies Market Cap

Exploring Distance and Fare Correlation

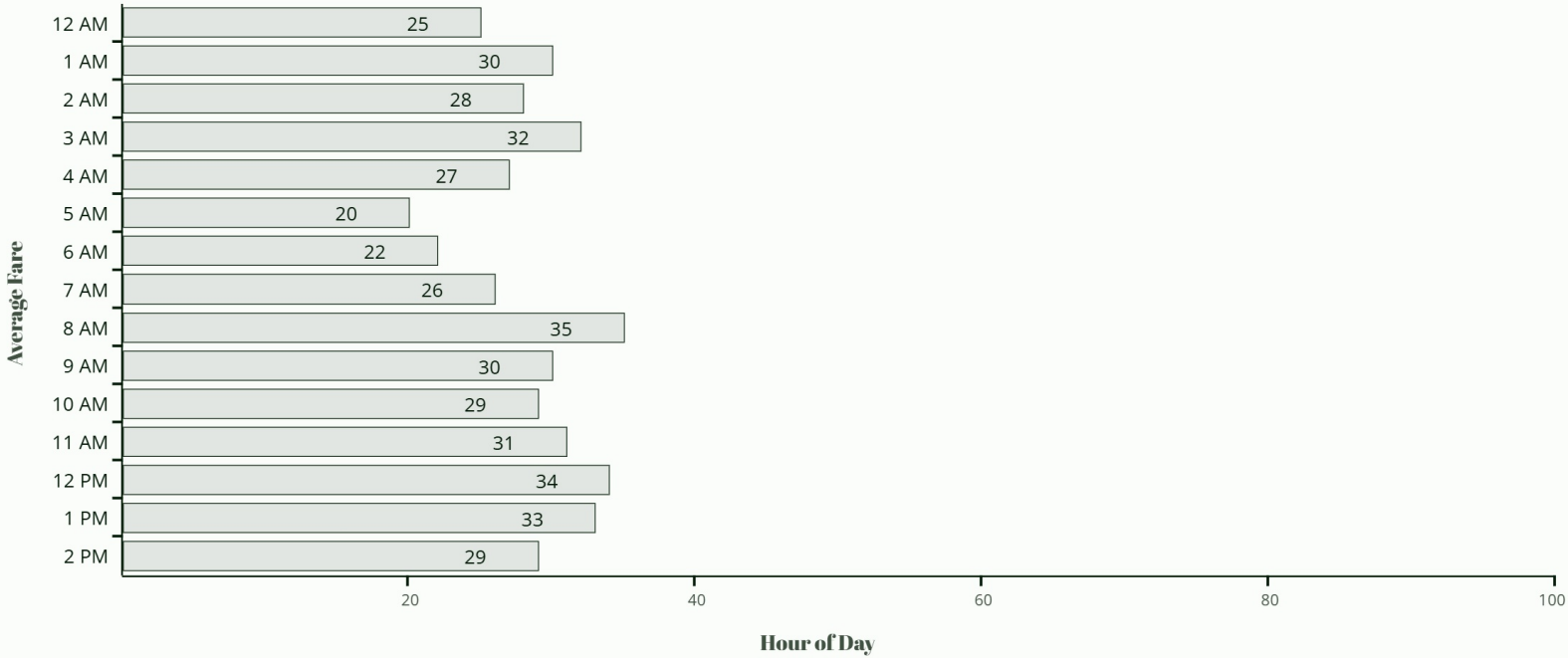
Analyzing the relationship between distance and fare

- Fare tends to increase with distance traveled
- Scatter plot indicates a strong linear relationship
- Fixed fare rides create horizontal clusters



NYC Taxi Fare Data (2023)

Exploring Time-Based Fare Variations



Source: Companies Market Cap

Impact of Passenger Count on Fares

1 **Majority of rides are solo or duo**

Most taxi rides are undertaken by 1 or 2 passengers, highlighting a common usage pattern.

2 **Fare independence from passenger count**

The total fare amount does not significantly fluctuate based on how many people are riding.

3 **Slight fare changes for large groups**

Larger groups may see minor fare adjustments, but these are not substantial.

Key Observations and Outliers in Data

Identifying Key Anomalies in Taxi Fare Data

- **Zero-distance trips with fares**

  These trips show non-zero fares, indicating possible data entry errors or unique circumstances.

- **Long trips with low fares**

  Instances of extended trips costing significantly less than expected, suggesting irregular pricing structures.

- **Fixed-price clusters**

  Data reveals horizontal lines indicating fixed-price fare clusters, calling for detailed analysis.

- **Need for custom handling**

  Identified anomalies may require tailored treatments or exclusions in predictive modeling.

Comprehensive Pivot Table Analysis

Insights into taxi fare structures and usage patterns

Distance Bucket (km)	Avg. Fare ($)	Passenger Count	Ride Frequency
0-1	$2.50	150	200
1-2	$4.00	200	300
2-3	$6.00	250	350
3-4	$8.00	300	400
4-5	$10.00	350	450

Exploring Key Insights on NYC Taxi Fares

Summary of Findings on Fare Dynamics and Patterns

- **Short-Distance Rides Predominate**

  Most NYC taxi rides are short, highlighting commuter patterns.

- **Fare Increases with Distance**

  The fare amount generally increases linearly with the distance traveled.

- **Fixed Fare Patterns for Airports**

  Certain rides, like those to airports, have fixed fare structures.

- **Importance of Data Cleaning**

  Data cleaning is crucial for enhancing model performance.

- **Time of Day Influences Fares**

  The time of day significantly affects taxi fare dynamics.