# Clustering NYC Taxi Trips to Identify Travel Patterns

## Uncovering High-Demand Zones

**Sunaina Jain**
**Instructor: Jonathan Agyeman**
**Data Source: NYC TLC 2016 Yellow Taxi Trips**

# Why Clustering Taxi Trips?

- NYC sees millions of taxi rides monthly → huge opportunity for optimization
- Clustering helps identify hidden patterns in complex data without labels.
- Applications:
  - Fleet management
  - Fare modeling
  -City planning and congestion control
- Project goal: Understand spatial, temporal, and fare-based behaviors

# Data Loading and Exploration

Merged 3 months of 2016 data

- **January**
- **February**
- **March**

Ensured uniform structure by checking column names across datasets

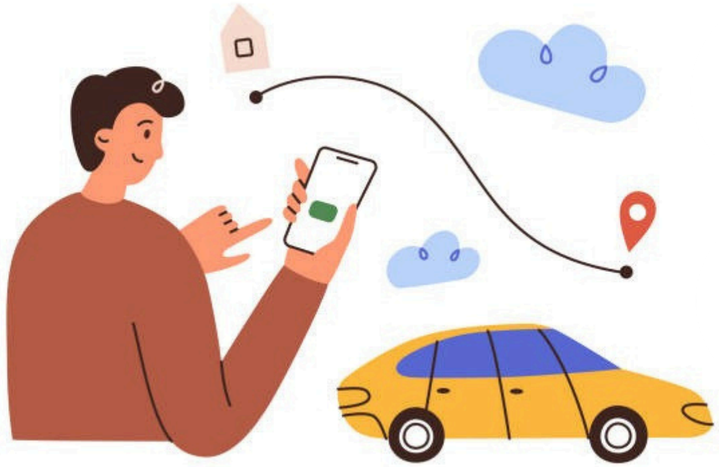
**Used a presence matrix to verify that all column names match before merging**

Explanatory Data Analysis

**Cleaned for:**

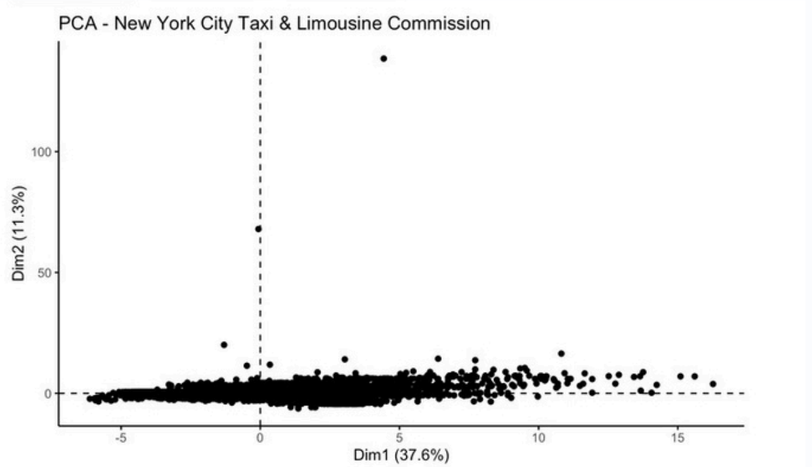
- **Missing values**
- **Duplicates**
- **Outliers**

# Clustering Validation



01

## PCA (Principal Component Analysis) done for dimensionality reduction




PC1 + PC2 ≈ 49% variance explained

02

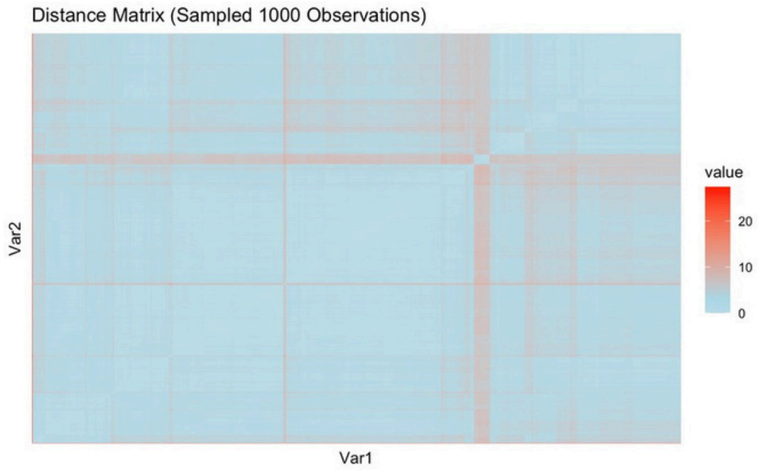
## Hopkins Statistic: 0.8497

Strong clustering tendency

03

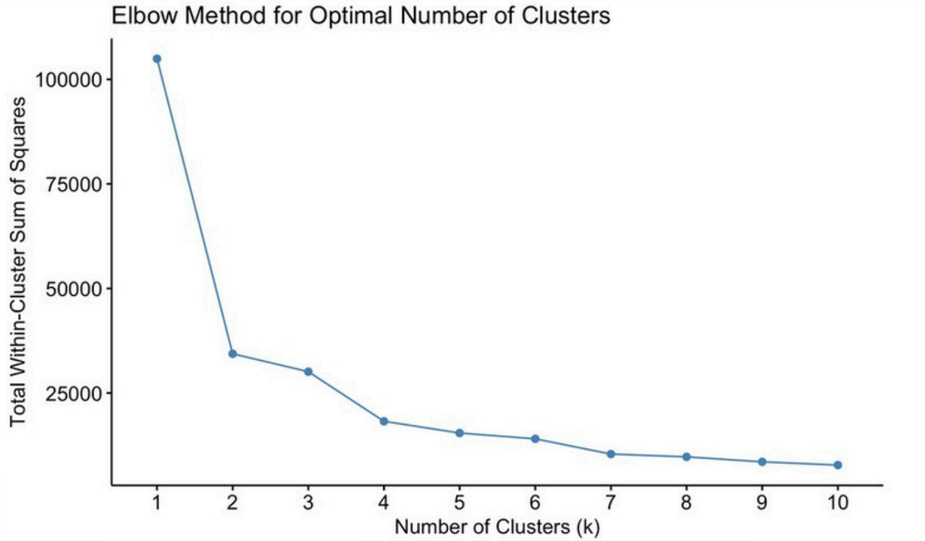
## Distance Matrix Heatmap

Showed dense clusters and clear anomalies

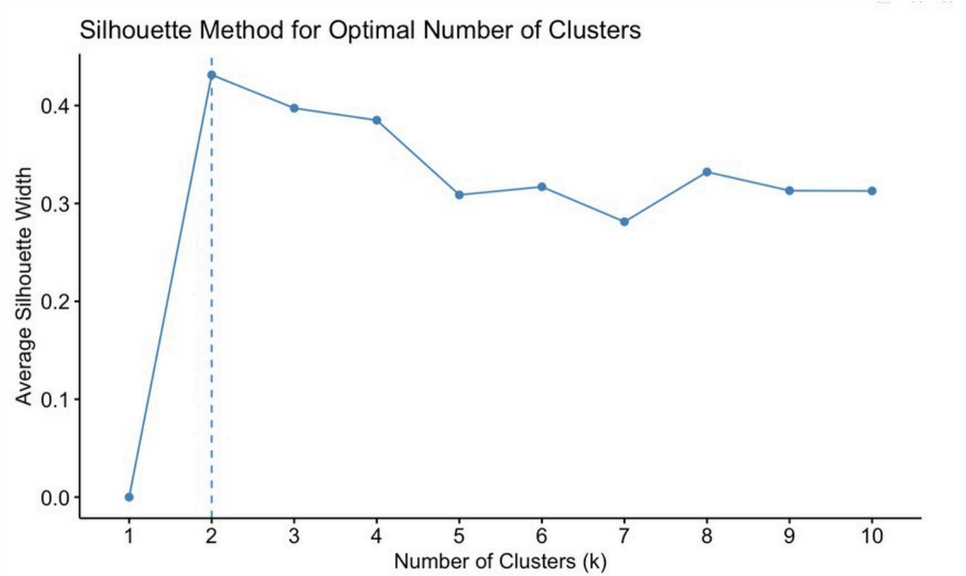


# Choosing the Optimal Clusters (k)

- **Elbow Method: Plotted total WSS for k = 1 to 10**
  **Elbow point at k = 4**




- **Silhouette Analysis: Peak at k = 2, still strong for k = 4**




**Final choice: k = 4 for balance between complexity and interpretability**

# Comparing Clustering Algorithms

## Evaluated 3 methods

- K-Means,
- PAM (K-Medoids)
- CLARA

## Internal Validation

K-Means scored best on DBI (0.9075) CH (1305.96) Dunn (0.0012)

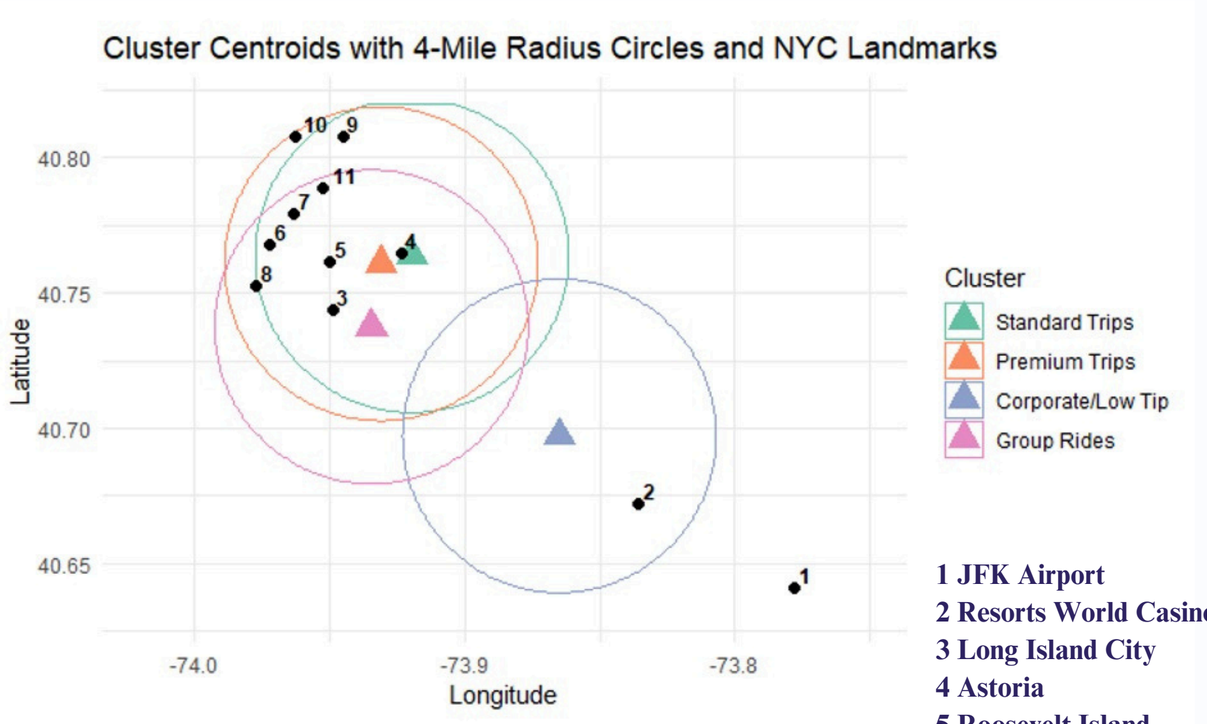
## External Validation

K-Means ↔ CLARA had highest Adjusted Rand Index (0.9286)

## Stability Metrics

K-Means had lowest APN and ADM

# Geospatial Clustering – Hotspots

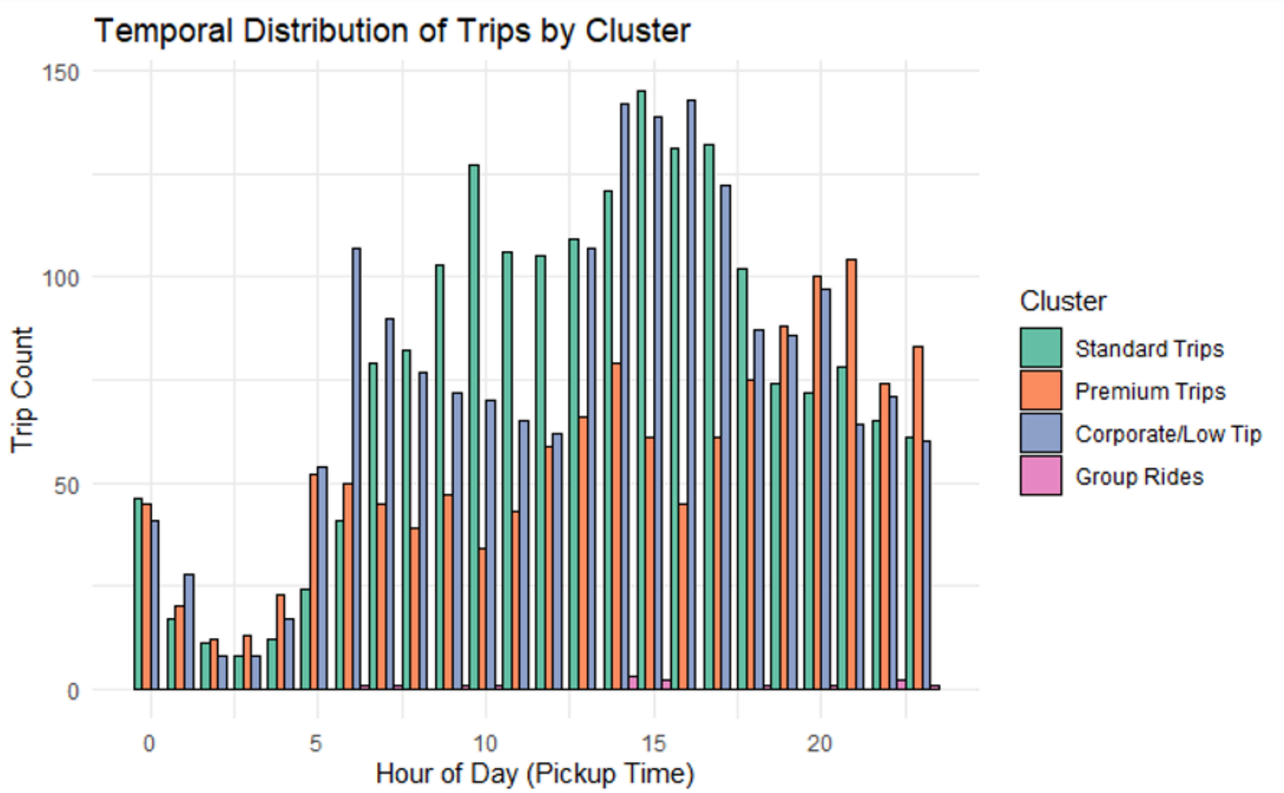



1 JFK Airport
2 Resorts World Casino
3 Long Island City
4 Astoria
5 Roosevelt Island
6 Central Park
7 The Met (Museum)
8 Grand Central Terminal
9 Harlem (125th St)
10 Columbia University
11 Mount Sinai Hospital

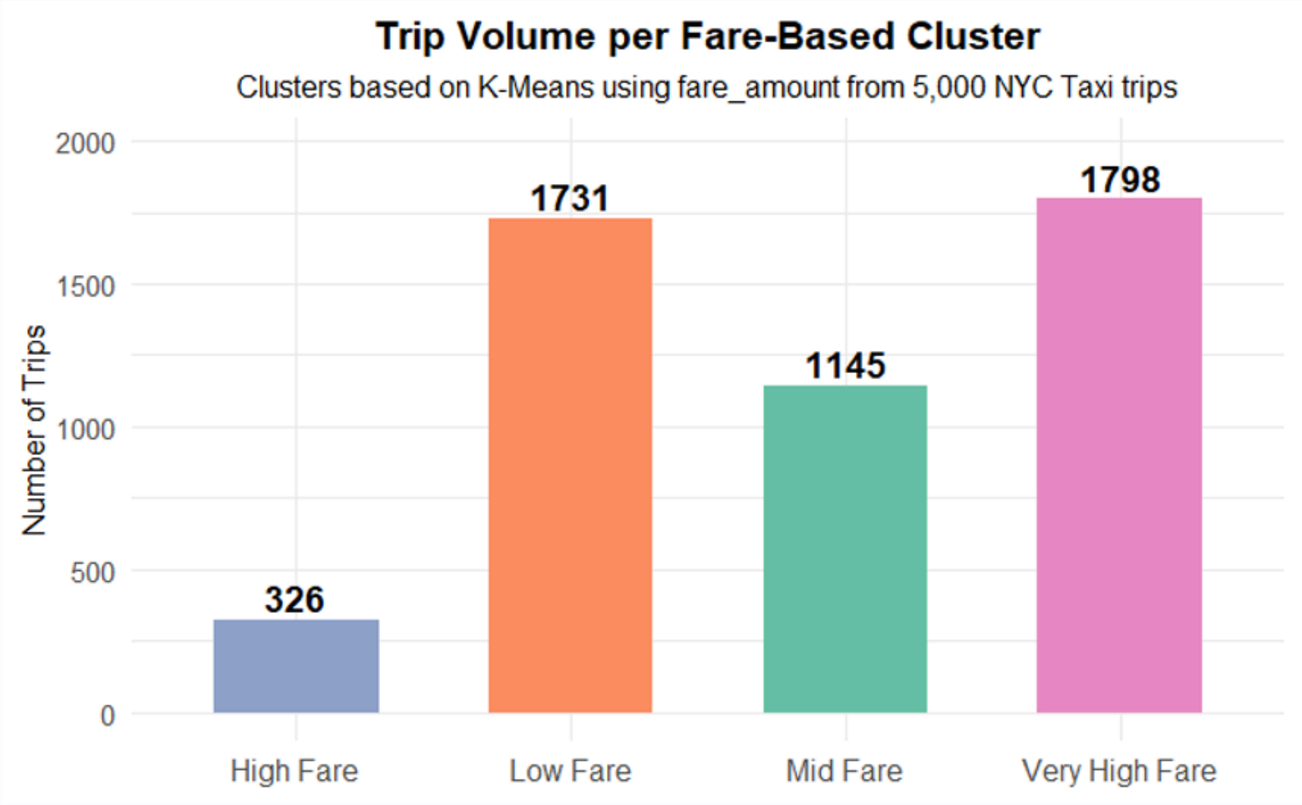
**The scatter plot illustrating pickup locations categorized by cluster demonstrates clear spatial distributions associated with different trip types in New York City. The "Corporate/Low Tip" cluster, represented in blue, is predominantly found around JFK Airport and surrounding areas, implying that these trips are likely to cover greater distances or pertain to airport services. Conversely, the "Standard Trips" cluster (green) and the "Premium Trips" cluster (orange) are primarily concentrated in central Manhattan, reflecting a high level of activity and demand within the city's commercial hub. The "Group Rides" cluster, depicted in pink, shows a more scattered distribution, indicating a range of pickup locations that may be influenced by ride-sharing or arrangements involving multiple passengers.**

# Temporal Trip Behaviors

**The bar chart illustrates that the demand for taxis reaches its highest levels from 12:00 PM to 5:00 PM, as all categories demonstrate increased activity during this timeframe. Standard Trips and Corporate/Low Tip trips are particularly prevalent during the afternoon peak, indicating a significant occurrence of work-related or habitual personal travel. In contrast, Premium Trips display a more uniform distribution across the day, with a marked increase in the late evening, which may indicate a correlation with nightlife or the use of higher-end services. Group Rides maintain a relatively steady trend but exhibit a lower overall volume, with slight increases coinciding with the afternoon and early evening periods.**



# Fare-Based Segmentation




- Very Low Fare (< $10) – short city hops
- Mid Fare ($10–$30) – typical urban rides
- Very High Fare (> $60) – airports, outer boroughs

**The bar chart depicts the categorization of NYC taxi trips according to fare amounts, utilizing K-Means clustering. A significant portion of the trips is categorized into the Very High Fare (1,798 trips) and Low Fare (1,731 trips) clusters. This trend implies that passengers predominantly opt for either short, economical rides or longer, pricier journeys, potentially to or from airports or between boroughs. The Mid Fare cluster encompasses 1,145 trips, which signifies rides of moderate distance typical of travel within the city. Conversely, the High Fare cluster is the least populated, with only 326 trips, suggesting that this fare range may represent anomalies or less frequent trip types.**

# Conclusion

**Following comprehensive data cleansing and dimensionality reduction, K-Means clustering was determined to be the most effective algorithm, surpassing PAM and CLARA in terms of internal, external, and stability validation metrics.**

**Utilizing K = 4 clusters, the analysis uncovered significant insights: specific pickup locations, such as JFK Airport and central Manhattan, were identified as high-demand areas; temporal patterns indicated that taxi usage peaked in the afternoon; and fare-based segmentation revealed notable variations in customer behavior and trip types.**

**In summary, the results underscore the practical significance of clustering in uncovering latent structures within intricate transportation data. By pinpointing distinct trip patterns and demand trends, this methodology can facilitate data-informed decision-making aimed at optimizing service distribution, improving passenger experiences, and informing future urban transportation planning efforts.**

# THANK-YOU