BA 820 - Project Milestone 2

- Fair Fare: A Deep Dive into Taxi Price Fluctuations and Congestion Pricing in New York City
- B1 Team 3
- Jooyeon Lee, Jeonghee (Christina) Son, Crystal Leatvanich, Courtney Vincent
- March 3rd, 2025

I. Proposal

New York City taxi fares vary by location, time, and other factors, but the lack of pricing transparency raises concerns about fairness, especially in congestion pricing. High fares have driven many riders to alternatives like Uber and Lyft[1]. This analysis explores how taxi fares fluctuate based on ride characteristics and identifies disproportionate pricing, particularly the impact of additional surcharges on different trip types. Based on these findings, we aim to improve fare transparency and passenger trust.

The findings will assess fare equity across neighborhoods and provide strategic insights into minimizing congestion-related costs while improving the rider experience. By highlighting fare inconsistencies and improving pricing transparency, this analysis aims to help passengers better anticipate their taxi fares, reduce unexpected costs, and enhance overall customer satisfaction through clearer pricing structures and better communication of fare components.

II. Exploratory Data Analysis (EDA) & Preprocessing

Building upon the findings from M1, we confirmed that the initial EDA and preprocessing steps remained valid. In M2, additional refinements such as a correlation heatmap and feature engineering were introduced to enhance numerical analysis and clustering insights.

The correlation matrix revealed key relationships between numerical features. Fare amount strongly correlates with total cost, indicating that base fare drives pricing, while additional charges play a minor role. Trip distance and tip amount moderately correlate with fare, which suggests longer trips cost more and receive higher tips. Congestion surcharge has a weak negative correlation (-0.16) with fare amount, suggesting higher fares don't always mean higher surcharges, possibly due to route choices or location. Passenger count shows little correlation with fare, confirming that NYC taxi fares are trip-based rather than passenger-based. These insights highlight the role of fare amount in total cost, the influence of non-distance-based fees, and potential surcharge inconsistencies.

For feature engineering, time-based features were created by categorizing timestamps into morning (6-12H), afternoon (12-18H), and night (18-6H) to identify trip patterns. Weekdays and weekends were classified separately to explore ride behavior. Trip duration was calculated, with negative values corrected and zero-duration, high-fare trips flagged as potential anomalies. Short and long trips were defined using the mean duration, and surcharges were classified as binary variables (no fee, fee). Initially, passengers were grouped into small (1-2) and large (3+), but after analyzing association rules, we refined them into single (1), couple (2), and group (3+), to prompt deeper analysis to uncover the most influential factors.

[1] CNBC, "New York City taxis fight for survival against Uber and Lyft", https://www.cnbc.com/2023/07/15/new-york-city-taxis-fight-for-survival-against-uber-and-lyft.html

Location-based transformations for pickup and drop-off locations were mapped to boroughs, and interborough trips were categorized. Borough pairings defined distance categories such as close (same/nearby boroughs), moderate (e.g., Brooklyn to Staten Island), and far (e.g., Manhattan to Staten Island or Newark Airport). These refinements in M2 improved interpretability while keeping M1's core preprocessing intact.

III. Analysis & Experiments

Given the implementation of congestion pricing and additional surcharges, concerns arise about whether these policies disproportionately impact specific ride characteristics. This analysis aims to address these concerns through data-driven analysis.

We used clustering because it efficiently groups unlabeled data into clusters based on similarity, which is ideal for identifying patterns in taxi trip characteristics (e.g., fare amount, trip distance, congestion surcharge). Its scalability suits the large NYC taxi dataset, and its ability to work with PCA-reduced dimensions ensures robust handling of high-dimensional data. Association rule mining was also used to identify frequent co-occurrences and relationships within each cluster. It is particularly suited for categorical data and can reveal actionable insights, such as which trip features are associated with higher fares or congestion surcharges, supporting our goal of assessing fare equity and pricing influences.

Once the data was preprocessed and features were added, we proceeded with Principal Component Analysis. We first applied standard scaling to normalize the variables before running the PCA model. The model showed that 4 principal components accounted for 90% of the variance in the dataset. After analyzing the reconstruction error, we noticed that approximately 5% of the data points were likely outliers. To address this, we filtered out values beyond three standard deviations. We then converted back to our original data frame, removed the rows corresponding to these outliers, and repeated the PCA process. With the removed outliers we observed that the data could now be explained using 3 principal components instead of 4. To interpret these results, we examined the loadings, which indicate the contribution of each original feature to each principal component. To capture the key trends in our data, here are the general insights from the first three principal components:

- PC1 has strong positive relationships with all variables, indicating that this component represents the main variance driver. The high positive loadings suggest that all these factors—distance, fare, tips, tolls, total cost, and duration—tend to increase together. Higher values in PC1 are associated with more expensive and longer rides due to higher fares, tips, tolls, and overall trip costs.

- PC2 shows a mix of positive and negative relationships. Positive for tip_amount and trip_duration_minutes. This component seems to reflect a trade-off between tips and tolls, higher tips are associated with lower tolls and vice versa. Higher PC2 values can indicate

rides with more tips but lower tolls

- PC3 captures a contrast between trip duration and additional costs. Negative for tip_amount and toll_amount suggests that longer durations might come with lower tips and tolls, or vice versa. Represents ride occupancy and congestion-related costs. Higher PC3 values indicate longer trips but lower additional costs like tips and tolls.

To identify meaningful patterns in the dataset, we applied K-Means clustering at first after performing PCA. The number of clusters (k) was determined using the Elbow Method, plotting inertia against k from 1 to 10. The "elbow" at k=2 indicated optimal clustering.[2] Next we visualized the clustering results using a pairplot to examine relationships between key variables. As expected we observed a strong correlation between, total amount vs. tip amount, fare amount vs. total amount, and fare amount vs. trip distance[3]. K-means cluster imbalance is evident, with Cluster 1 dominating the dataset, while Cluster 0 is significantly smaller. This indicates that the clustering model primarily captures short-distance, lower-cost trips (Cluster 1), while longer, more expensive trips (Cluster 0) form a minority group. The silhouette score averaged > 0.7, indicating good separation, though cluster imbalance suggested potential limitations. Now that we have two distinct clusters our team wants to find out how it relates to other categorical variables in our data, such as different fees, location, days, and time. The goal is to determine whether disproportionality influences specific ride characteristics based on the clusters of trip distance and pricing.

Hierarchical clustering with Ward's linkage was attempted as an alternative to K-means but failed due to memory constraints (dataset size: ~183k rows after filtering). The linkage computation crashed the system. We abandoned hierarchical clustering and focused on K-means, which is less memory-intensive and better suited for large datasets.

Association rules were incorporated as a complementary method to capture relationships that hierarchical clustering might have revealed. Rather than relying solely on cluster proportions, association rule mining was used to identify intra-cluster patterns and understand how different trip characteristics interact within each cluster. To balance granularity and computational feasibility, the minimum support was set to 0.05, ensuring that only patterns occurring in at least 5% of trips were considered. Additionally, lift > 1.0 ensured the identified rules reflected meaningful positive associations, while confidence > 0.8 and support > 0.1 filtered for strong and frequent patterns. Association rules were generated separately for each cluster to highlight cluster-specific pricing.

This approach added a new layer of insight beyond K-means clustering alone, allowing for a deeper understanding of taxi fare dynamics. By capturing subtle behavioral patterns within each cluster, association rules provided valuable takeaways for riders looking to anticipate fares

[2] Appendix - A2
[3] Appendix - A3

more accurately and optimize their travel choices to minimize unexpected costs.

IV. Challenges, Dead ends, and Adjustments

One of the biggest challenges was handling the large dataset and extensive code, which caused repeated crashes in Google Colab, even after attempting optimizations such as reducing the sample size, changing the runtime type, and adjusting the hardware accelerator. After exploring alternative solutions, we transitioned to Jupyter Notebook, which allowed the code to run more reliably. However, this shift introduced additional difficulties, as Colab and Jupyter require different commands for loading data, managing libraries, and handling missing values with MSNO, necessitating adjustments for smooth execution.

Another challenge arose when interpreting the K-means clustering results. Initially, it was unclear how to leverage the clusters to align with the project's objectives. To gain deeper insights, we applied association rule analysis, which helped identify trip distance and pickup/drop-off locations as the most influential factors, clarifying how to effectively utilize clustering results. However, the association rule analysis initially failed to produce meaningful differences between clusters, contradicting the pairplot insights observed earlier. This discrepancy complicated interpretation, prompting a deeper investigation into the dataset. By adjusting the rules and refining the antecedents and consequences used in the association rule analysis, we were able to pinpoint the features with the most significant impact, ultimately resolving the inconsistency and improving our understanding of the clustering results.

These challenges reinforced the importance of adapting methodologies, validating data relationships through multiple approaches, and ensuring that clustering results align with real-world patterns before drawing conclusions.

V. Findings and Interpretations

Our analysis identified two distinct trip clusters based on cost, distance, and congestion charges. Cluster 0 primarily consists of long-distance interborough trips, often from Queens to Manhattan, with weekday mornings being the most frequent travel time. While these trips tend to have lower average congestion charges, they impose a greater financial burden on passengers due to high toll and tip costs. The strong association between these factors suggests that many of these rides are likely taking Taxis for commuting purposes. Additionally, congestion surcharges often apply to these trips, likely due to traffic congestion at key transit points such as bridges and tunnels. The data also shows that these rides are frequently taken by single riders or pairs, reinforcing the idea that taxis serve as an alternative to public transportation for commuting. However, the presence of congestion surcharges in these longer rides may contribute to passenger frustration, potentially driving them toward more predictable ride-sharing alternatives.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(DO_Borough_Queens, PU_Borough_Manhattan)	(congestion_surcharge_type_Fee)	0.208123	0.78585	0.198784	0.955127	1.215406
(DO_Borough_Queens, PU_Borough_Manhattan, interborough_trip_1)	(congestion_surcharge_type_Fee)	0.208123	0.78585	0.198784	0.955127	1.215406
(day_type_Weekday, PU_Borough_Manhattan, DO_Borough_Queens)	(congestion_surcharge_type_Fee)	0.153608	0.78585	0.146386	0.952987	1.212683

Cluster 1, in contrast, is characterized by shorter, intra-Manhattan trips, with weekday nights being the most common travel time. These trips are strongly associated with entertainment and nightlife activity, where taxis are frequently used for social outings. Despite their shorter distances, these trips often incur higher congestion charges due to heavy traffic in densely populated areas. Unlike longer interborough rides, where surcharges may feel justified, passengers taking short trips within Manhattan may find these additional costs unexpected and frustrating. As a result, similar to Cluster 0, this group may also prefer ride-sharing services that offer more transparent pricing structures.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(day_type_Weekday, PU_Borough_Manhattan, time_of_day_Night, DO_Borough_Manhattan)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
(day_type_Weekday, PU_Borough_Manhattan, time_of_day_Night, interborough_trip_0)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
(day_type_Weekday, PU_Borough_Manhattan, interborough_trip_0, DO_Borough_Manhattan, time_of_day_Night)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834

A key finding from this analysis is that congestion surcharges do not always align with trip length. While interborough trips naturally incur these charges due to greater distances and toll road usage, short trips within Manhattan are also frequently subject to surcharges despite covering much shorter distances. This has potential fairness implications, as passengers taking short trips within Manhattan may face disproportionate pricing effects compared to those traveling longer distances across boroughs. This pricing structure may lead to a perception of unfairness, reducing customer satisfaction and usage. To improve fare transparency and enhance the overall rider experience, congestion pricing models should be reassessed. Adjusting surcharges for short intra-Manhattan trips or introducing fare incentives for interborough travel could help traditional NYC yellow taxis remain competitive in the evolving transportation industry.

VI. Appendix

- **Contribution**:

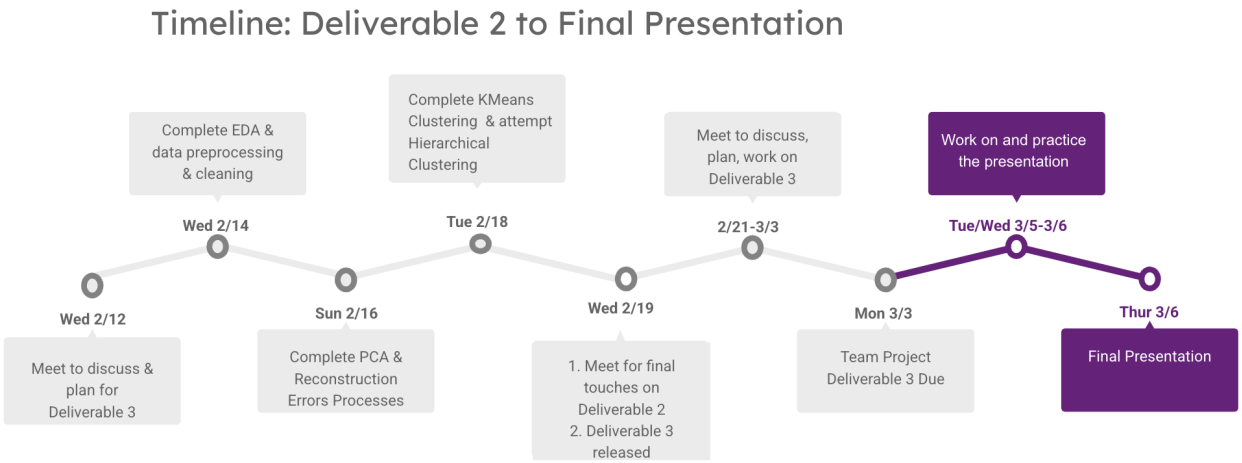
Name	Tasks & Contributions	Coding Contribution	Challenges & Dead Ends
Jooyeon Lee	K-Means Clustering: Elbow Method & Visualization, Silhouette Score, Written analysis plan and next steps(M1) and proposal and analysis(M2).	25%	Optimizing cluster number with long computation time, PCA & outlier discussion, Explored different K values and evaluated clustering quality
Christina Son	Organization project documents, Hierarchical Clustering, Skew Chart, Proposal, EDA, GitHub project creation, Association Rules, Written proposal and Findings(M2).	25%	Colab notebook crashing, Visualize extreme outliers (skew chart), PCA & outlier discussion, Initial Hierarchical clustering results not well defined, Association rules didn't run properly in Colab, so I had to use Jupyter Notebook instead.
Crystal Leatvanich	EDA, Pre-processing, Written EDA/Pre-processing & Preliminary Results and Analysis, Timeline, Feature engineering	25%	Difficulty handling 38.2 million rows (sampling), Using Python in command terminal for data extraction, PCA & outlier discussion
Courtney Vincent	PCA,Reconstruction Error, Written Preliminary Results	25%	Interpreting PCA results and Outliers, using PCA to see

	and Analysis Plan, Clustering, K-Means Analysis, Pairplot.		explained variance, PCA & outlier discussion

- **GitHub Project**: https://github.com/christinabrnn/NYC-Taxi-Project

- **References**:
  - We used AI tools to clarify ideas related to Principal Component Analysis (PCA) and outlier removal. The key prompts used and AI responses are recorded: https://chatgpt.com/share/67b65f38-2770-800e-a9c2-3bad8949e689
  - We used AI tools to help with feature engineering, transforming large volumes of text into a structured format, making it suitable for analysis and model development. The key prompts used and AI responses are recorded: https://chatgpt.com/share/67c63c20-6bbc-8012-abf9-f0e27a0f42aa

- **Timeline:**

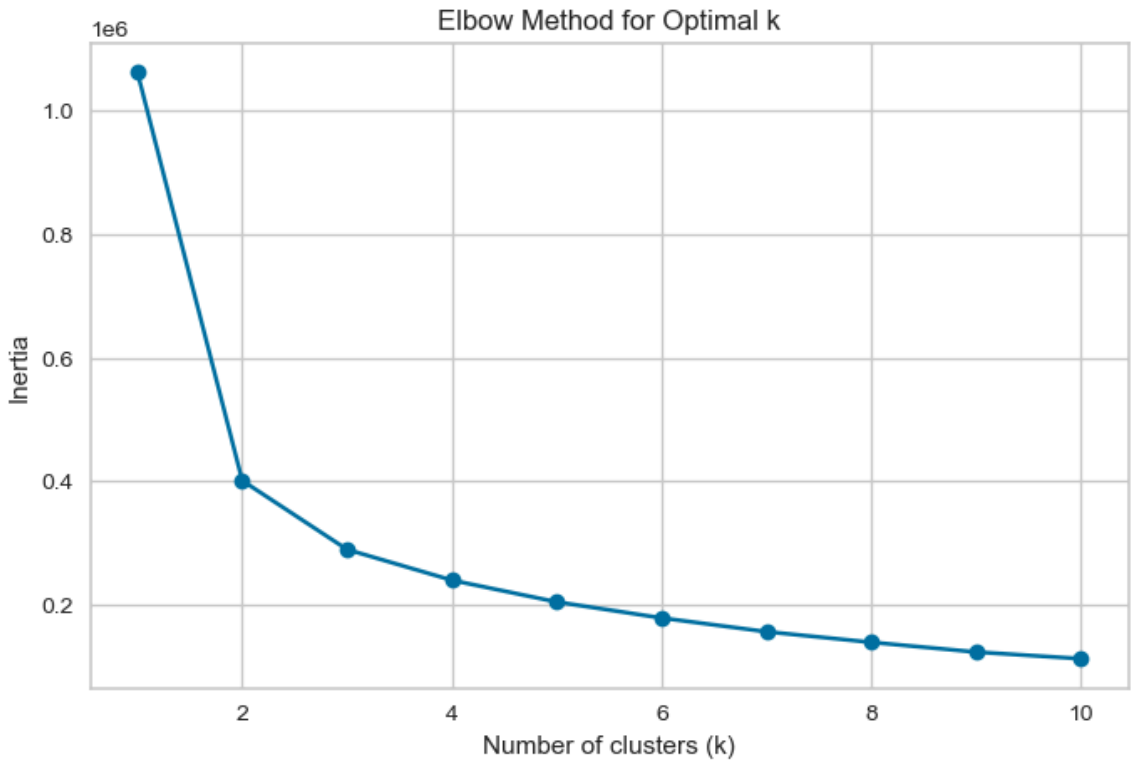


- **Data and Results**:

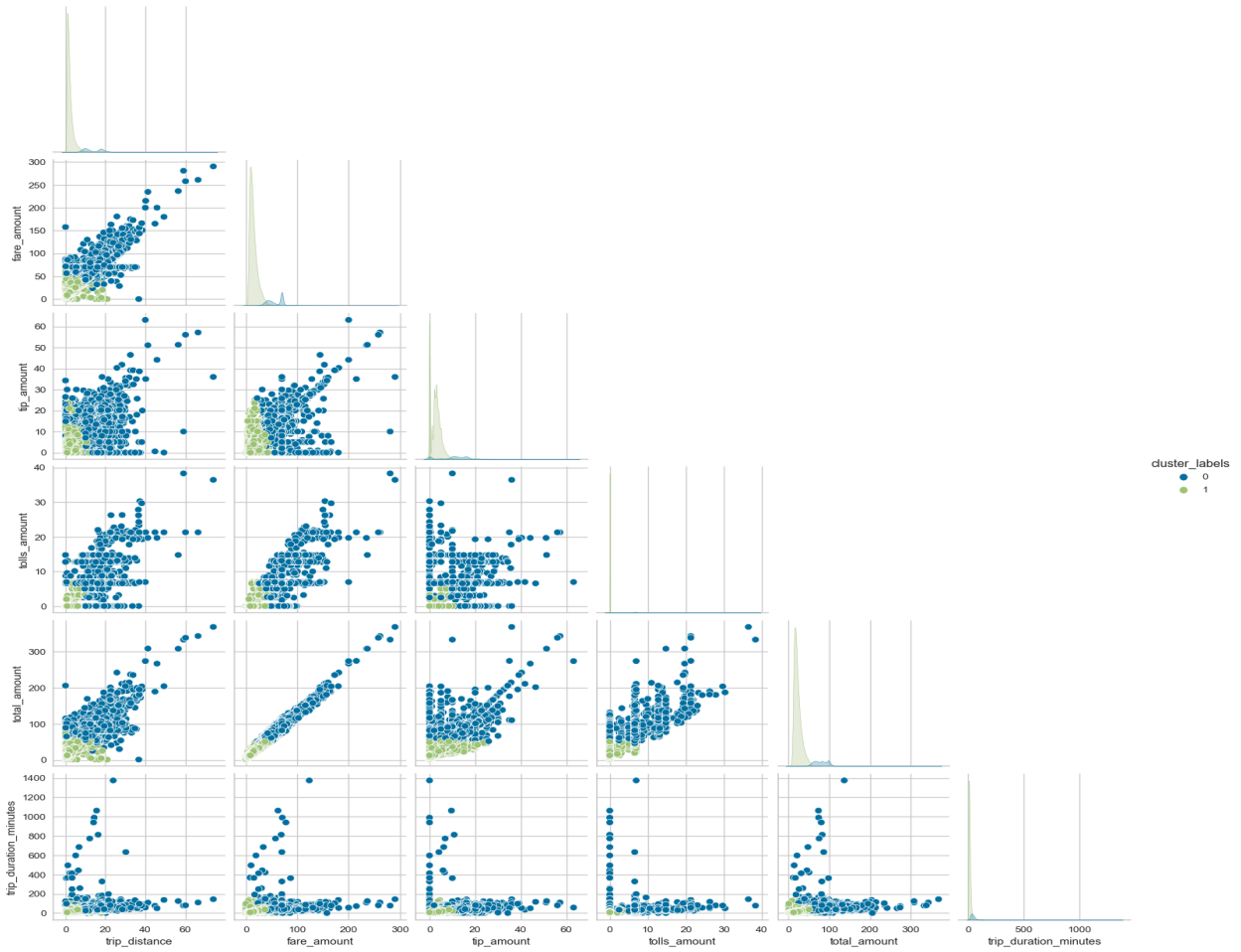
A1: Original Dataframe Describe Table

	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amount	extra
count	200000.000000	193101.000000	200000.000000	193101.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000
mean	1.738495	1.374866	4.785999	1.634005	165.369115	163.796390	1.184465	19.444202	1.554802
std	0.444883	0.896346	304.341505	7.374074	63.925262	69.928236	0.554237	18.830230	1.831698
min	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.000000	-500.000000	-7.500000
25%	1.000000	1.000000	1.040000	1.000000	132.000000	113.000000	1.000000	9.300000	0.000000
50%	2.000000	1.000000	1.790000	1.000000	162.000000	162.000000	1.000000	13.500000	1.000000
75%	2.000000	1.000000	3.400000	1.000000	234.000000	234.000000	1.000000	21.900000	2.500000
max	6.000000	7.000000	119576.840000	99.000000	265.000000	265.000000	4.000000	637.900000	12.750000

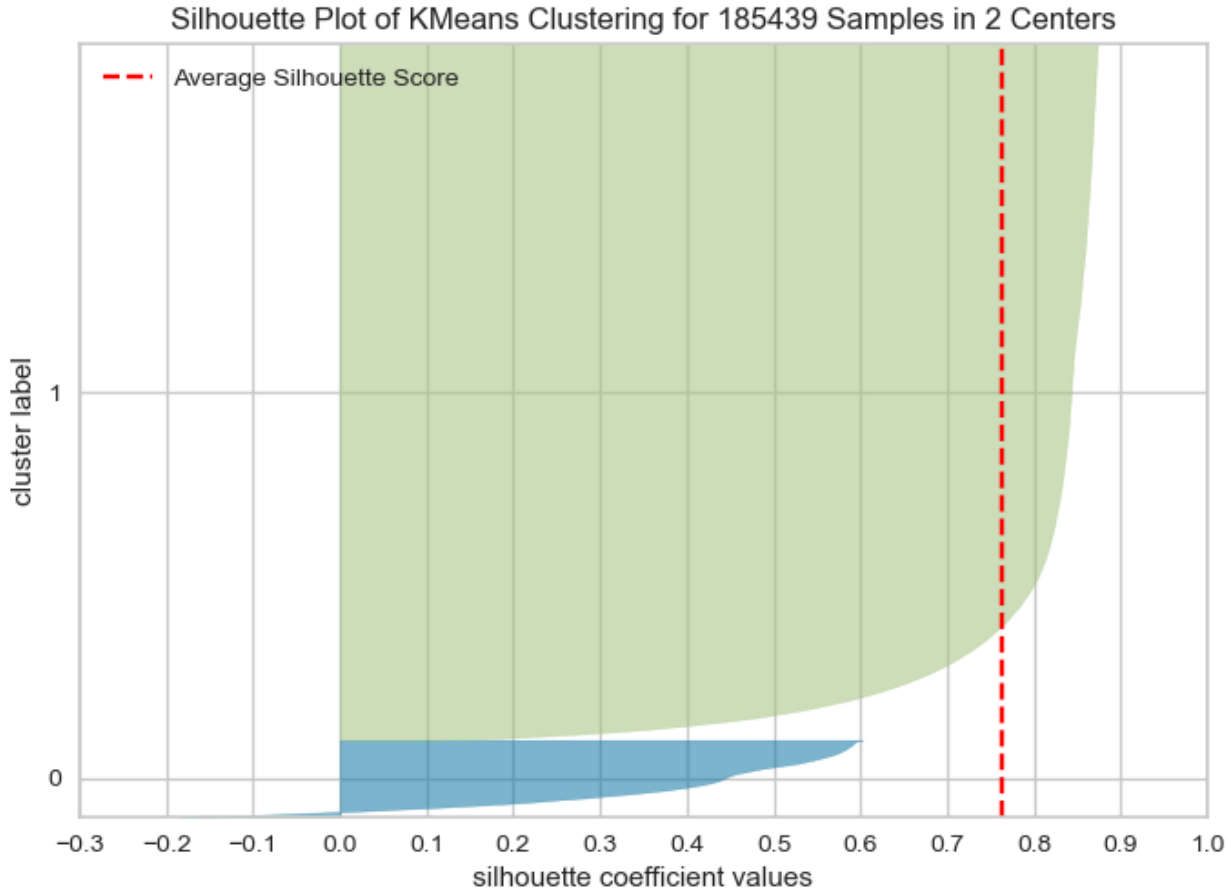
A2: Elbow Method for Optimal K for K-Means



A3: Pairplot for Two Clusters for Numerical Data



A4: Silhouette Plot of KMeans Clustering



A5: Top 10 Frequent Itemsets of Cluster 0 and 1

Top 10 Frequent Itemsets - Cluster 0

	support	itemsets
14	0.926318	(interborough_trip_1)
0	0.785850	(congestion_surcharge_type_Fee)
27	0.749362	(congestion_surcharge_type_Fee, interborough_trip_1)
15	0.740077	(day_type_Weekday)
7	0.708422	(passenger_type_Single)
104	0.686322	(interborough_trip_1, day_type_Weekday)
9	0.670902	(PU_Borough_Queens)
83	0.653852	(passenger_type_Single, interborough_trip_1)
92	0.634360	(PU_Borough_Queens, interborough_trip_1)
28	0.584514	(congestion_surcharge_type_Fee, day_type_Weekday)

Top 10 Frequent Itemsets - Cluster 1

	support	itemsets
7	0.962526	(PU_Borough_Manhattan)
9	0.959269	(interborough_trip_0)
0	0.956167	(congestion_surcharge_type_Fee)
18	0.946097	(PU_Borough_Manhattan, congestion_surcharge_type_Fee)
8	0.937427	(DO_Borough_Manhattan)
56	0.932284	(DO_Borough_Manhattan, PU_Borough_Manhattan)
167	0.932284	(interborough_trip_0, DO_Borough_Manhattan, PU_Borough_Manhattan)
57	0.932284	(interborough_trip_0, PU_Borough_Manhattan)
60	0.932284	(interborough_trip_0, DO_Borough_Manhattan)
20	0.922813	(interborough_trip_0, congestion_surcharge_type_Fee)

A6: Filtered Association Rules of Cluster 0

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
704	(DO_Borough_Queens, PU_Borough_Manhattan)	(congestion_surcharge_type_Fee)	0.208123	0.78585	0.198784	0.955127	1.215406
711	(DO_Borough_Queens, PU_Borough_Manhattan, interborough_trip_1)	(congestion_surcharge_type_Fee)	0.208123	0.78585	0.198784	0.955127	1.215406
1376	(day_type_Weekday, PU_Borough_Manhattan, DO_Borough_Queens)	(congestion_surcharge_type_Fee)	0.153608	0.78585	0.146386	0.952987	1.212683
1348	(day_type_Weekday, PU_Borough_Manhattan, DO_Borough_Queens, interborough_trip_1)	(congestion_surcharge_type_Fee)	0.153608	0.78585	0.146386	0.952987	1.212683
2665	(passenger_type_Single, PU_Borough_Manhattan, DO_Borough_Queens, day_type_Weekday)	(congestion_surcharge_type_Fee)	0.110496	0.78585	0.104957	0.949877	1.208726

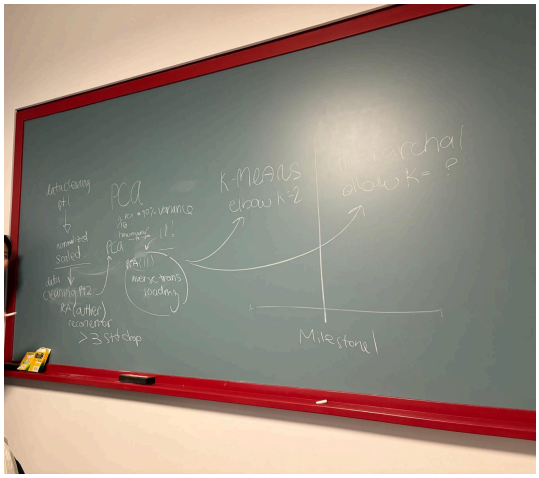
A7: Filtered Association Rules of Cluster 1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
1023	(day_type_Weekday, PU_Borough_Manhattan, time_of_day_Night, DO_Borough_Manhattan)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
1034	(day_type_Weekday, PU_Borough_Manhattan, time_of_day_Night, interborough_trip_0)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
983	(day_type_Weekday, PU_Borough_Manhattan, interborough_trip_0, DO_Borough_Manhattan, time_of_day_Night)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
1045	(day_type_Weekday, interborough_trip_0, time_of_day_Night, DO_Borough_Manhattan)	(congestion_surcharge_type_Fee)	0.267504	0.956167	0.264432	0.988518	1.033834
615	(interborough_trip_0, PU_Borough_Manhattan, time_of_day_Night)	(congestion_surcharge_type_Fee)	0.385332	0.956167	0.380854	0.988378	1.033687

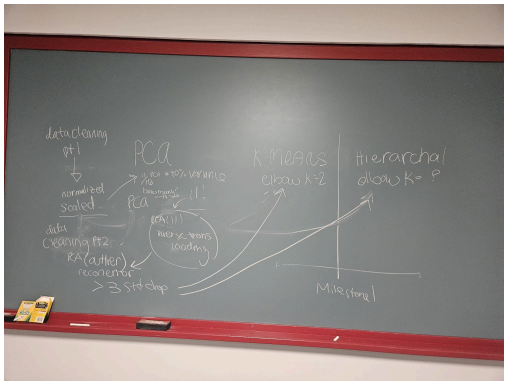
- **Collaborative Problem-Solving in Action**

Teamwork in action! Figuring out how the processes connect and bringing ideas to life on the board

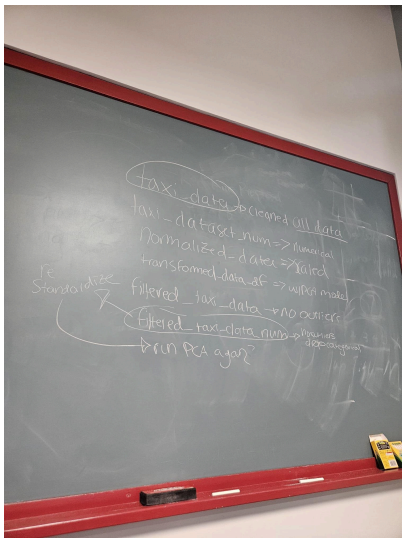
Mapping out the process flow:



Rethinking and updating the process flow:



Confirming on DataFrame naming conventions:



A brainstorming session captured on a whiteboard, mapping out potential meanings behind each cluster after running association rules.