Uber Customer Segmentation Analysis and Fare Prediction

https://www.linkedin.com/in/andi--pratama/

DATA OVERVIEW

- Uber, a global leader in the ride-sharing industry, operates in a complex and rapidly evolving transportation marketplace. The company leverages innovative technology to connect drivers with riders through a seamless digital platform, facilitating millions of rides daily across numerous countries.
- Uber operates in a highly competitive market, where it contends not only with traditional taxi services but also with other ride-sharing companies like Lyft.
- Customer retention, pricing strategy, and cost management are perennial challenges amid fluctuating demand patterns and variable operational costs.

Total Revenue

$ 1.419M

Total Trip

386K Km

Total User

327

Cost

$ 1.200 M

Churn Rate

38%

PROBLEM, OBJECTIVE AND HYPOTHESIS

1. Problem Statement

- How can Uber integrate to improve operational efficiency and reduce 5% costs ?
- How can Uber identify distinct customer segments based on riding patterns and preferences to tailor its marketing strategies?
- How can Uber dynamically adjust fare prices to maximize revenue while ensuring fairness and transparency to customers and reduce 5% Customer Churn rate?

2. Objective

- Explore and integrate ride-sharing solutions to improve operational efficiency and reduce costs
- Apply customer segmentation analysis to understand diverse user needs and preferences, facilitating personalized services and targeted marketing efforts.
- Develop and implement data-driven pricing models that adapt to various factors influencing fare amounts

3. Hypothesis

- Customer segments can be identified based on their riding patterns, such as average distance and typical fare amount.
- The fare amount is strongly influenced by ride distance, time of day, and day of the week.
- Ride demand varies predictably based on the time of day, day of the week, and location.

ANALYSIS FRAMEWORK

1. Business Understanding

- Define problem statement
- Define Objective
- Define hypothesis

2. Data Preparation

- Data Cleanning
- Data Transformation
- Outlier Handling

3. Exploratory Data Analysis

- Explore data distribution, etc

4. Modelling and Evaluation

- ANOVA Testing
- K-Means Clustering
- Random Forest Regressor & XGBoost

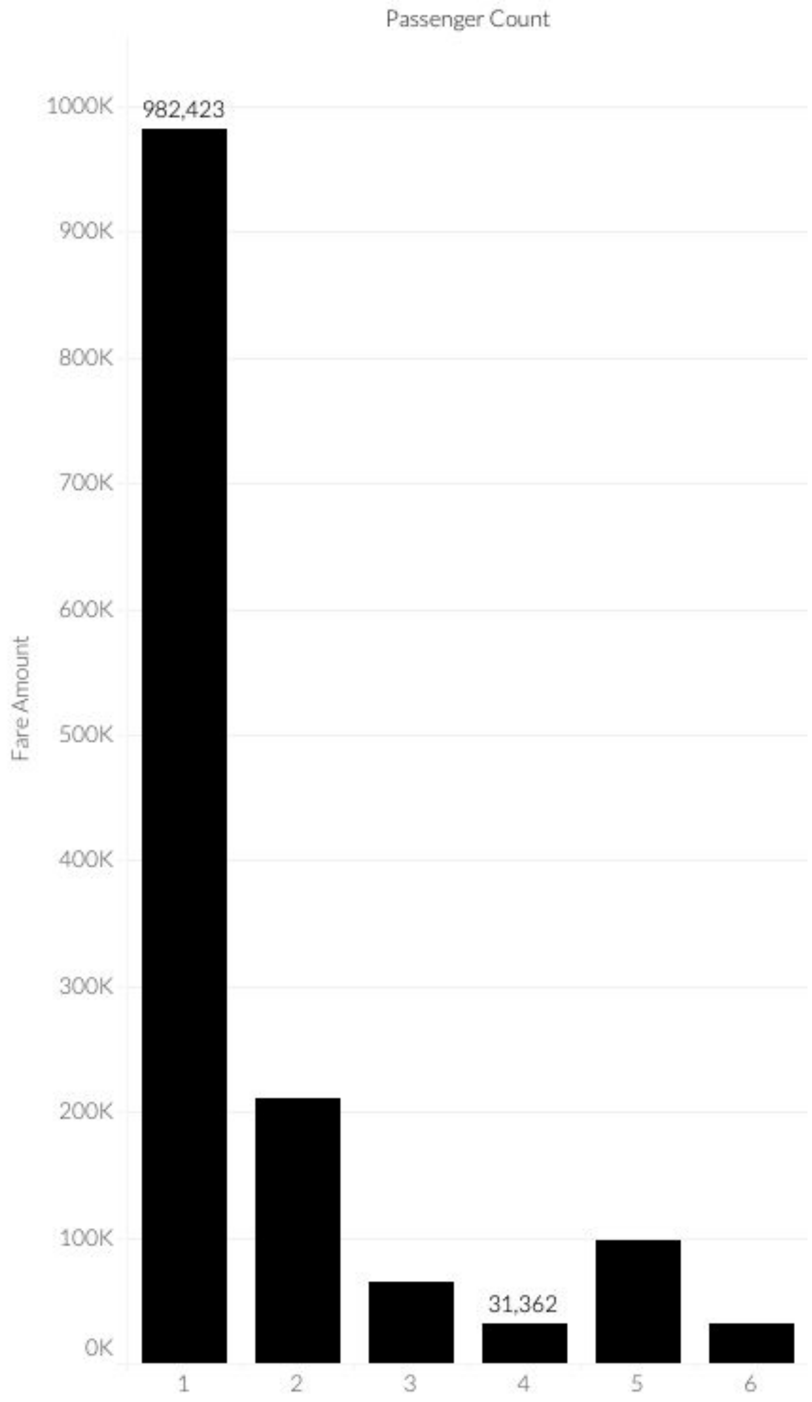
5. Insight and Recommendation

- Summarize Insight & Recommendation from findings

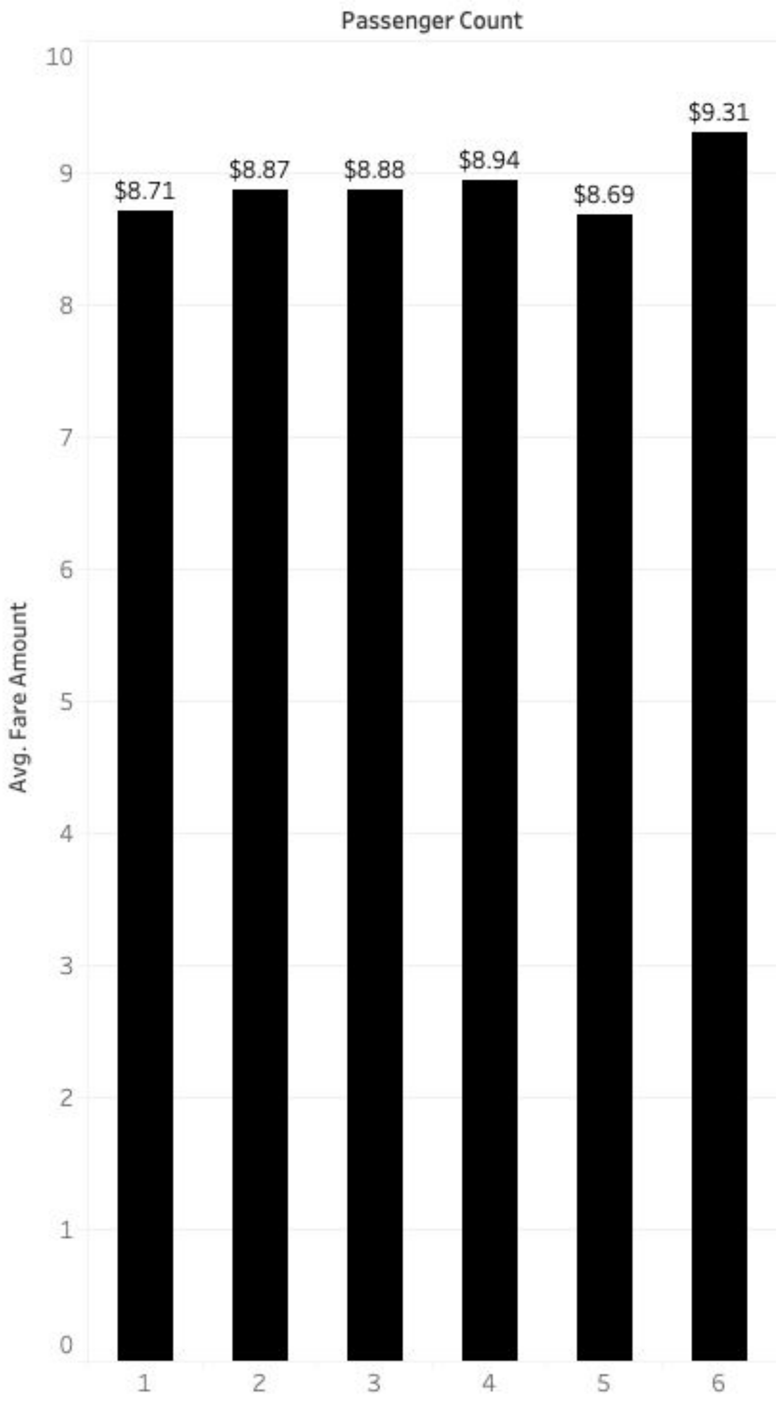
Passenger size and Seasonality effect Analysis

Most revenue gained from **solo traveller. Diversification** vehicle type is mandatory to **optimizing operational efficiency and reduce cost**

Revenue by passenger count

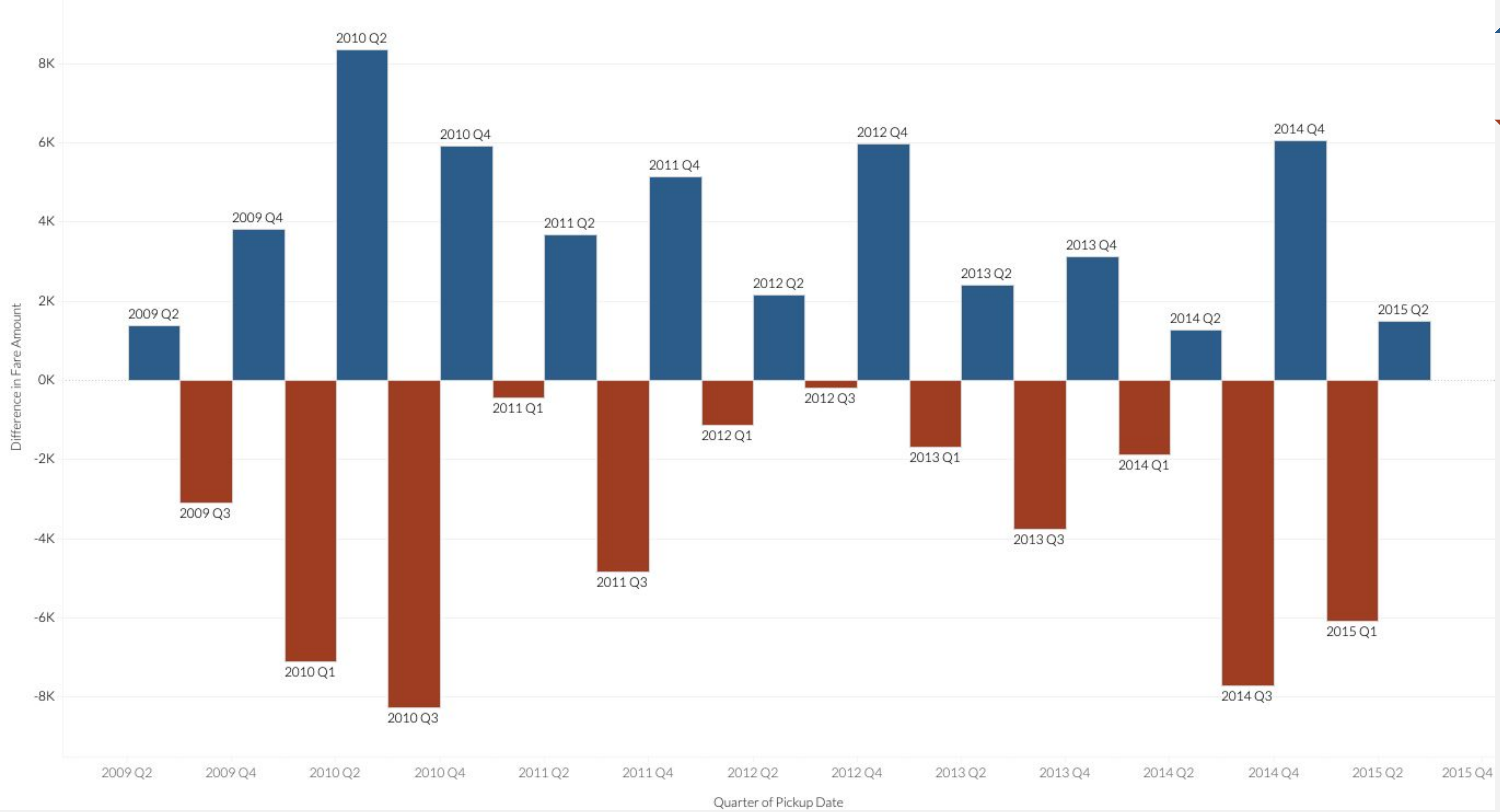


Average fare by passenger count



- A vast majority of the revenue comes from solo travelers, reaffirming that they are the core customer base for the service. Focusing on solo traveler preferences and behavior could lead to further revenue optimization.
- Diversification of vehicle types suggests the potential for cost savings. A mixed fleet with a variety of vehicle sizes can match the demand for different passenger counts more closely.
- Revenue from higher passenger counts is lower, the average fares are higher, which could offset the costs of larger vehicles.

There is **quarterly seasonality** effect , every **even quarter** is **increase** while **odds quarter** is **decrease**

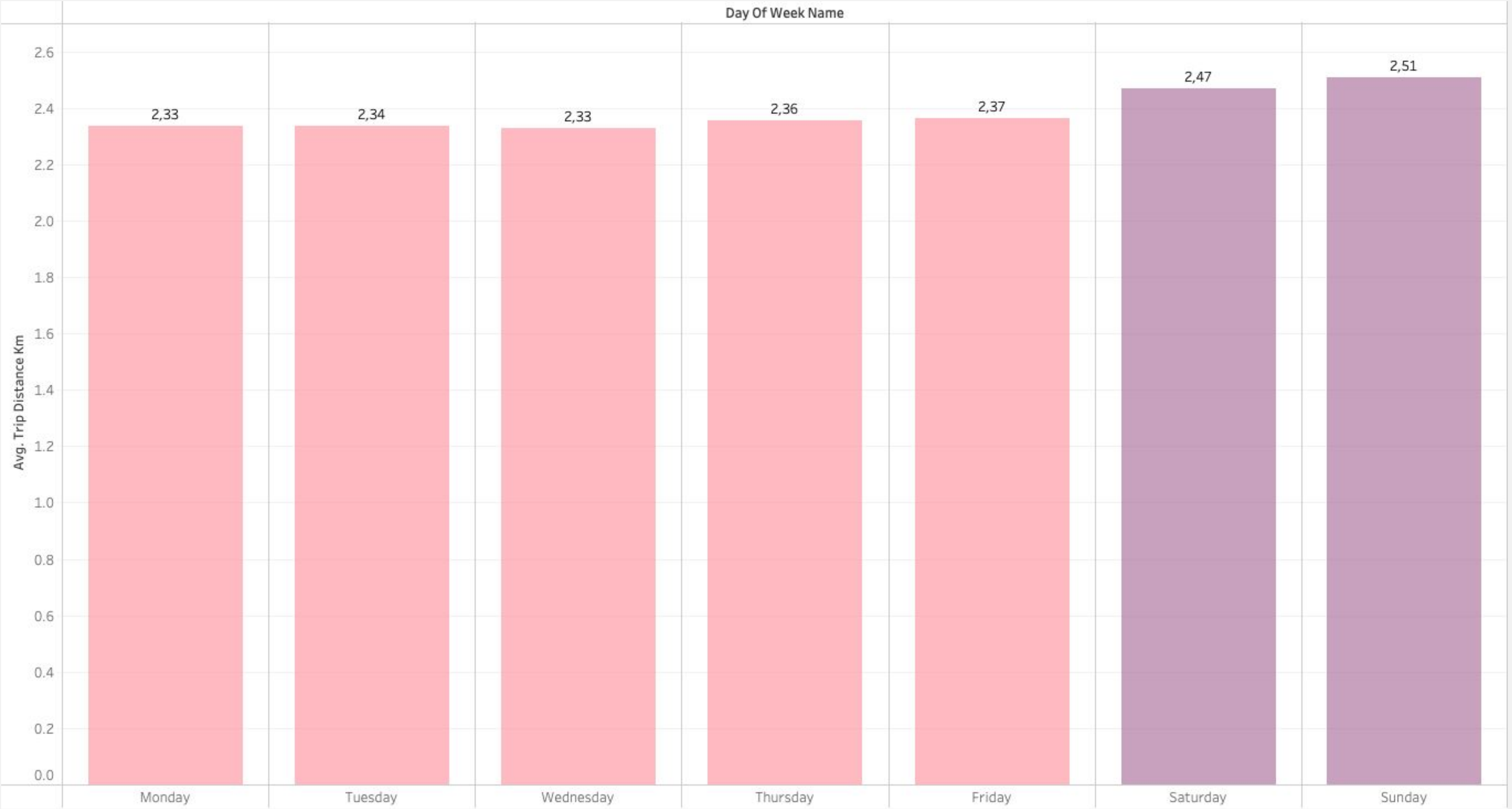


Q2 and Q4 Increase

Q1 and Q3 Decrease

- Cyclical pattern of rises and falls that suggests business operations are significantly impacted by seasonal factors. These could be related to weather, holidays, or industry-specific cycles.
- Seasonality can be used for forecasting and planning purposes. During expected peak quarters (Q2 and Q4), the business might need to scale up resources to meet demand, while during the slower quarters (Q1 and Q3), it could scale back or focus on maintenance and improvement activities.

Weekdays and **weekends** affect user commuting distances **behavior**.



Weekday

Weekend

- During weekdays, users will travel less distance. This pattern explains that holiday has a significant effect on user behavior.

ANOVA testing, Clustering and Predictive Model analysis

ANOVA testing shows **Weekend** is not significantly affect to fare amount. **Pickup city** and **dropoff city** are **multicollinearity**

ANOVA Testing

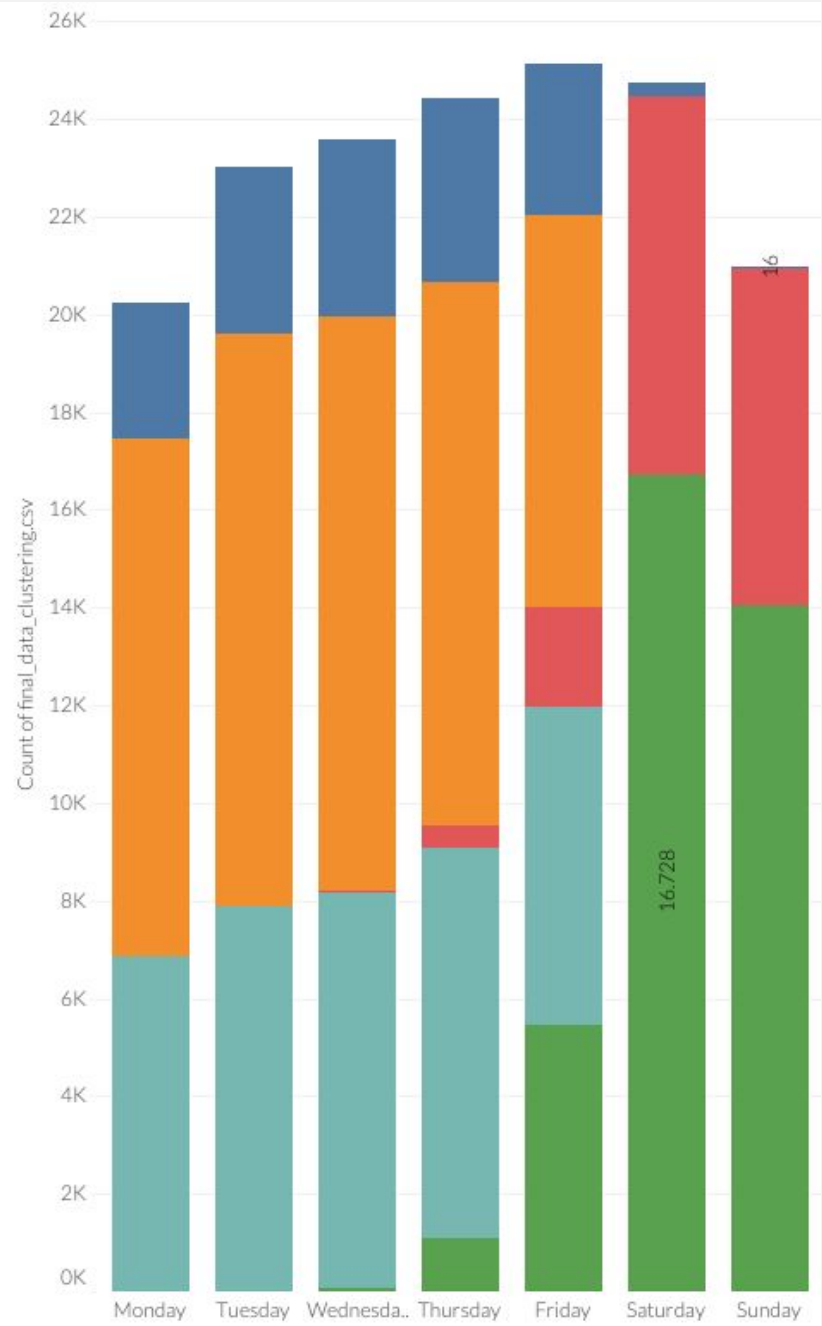
No	Variable	F Statistic	P Value	Intepretation
1	pickup city	36.39	0.0	Extremely significant
2	dropoff city	93.45	0.0	Extremely significant
3	trip distance	81.09	0.0	Extremely significant
4	part of day	41.82	5.24×10−27	Highly significant
5	day of week	49.73	4.08×10−32	Highly significant
6	passenger count	22.29	2.35×10−6	Highly significant
7	pickup hour	8.20	1.88×10−5	Highly significant
8	month	73.51	1.67×10−47	Highly significant
9	**weekend**	**0.86**	**0.35**	**Not significant**

Variable Inflation Factor (VIF)

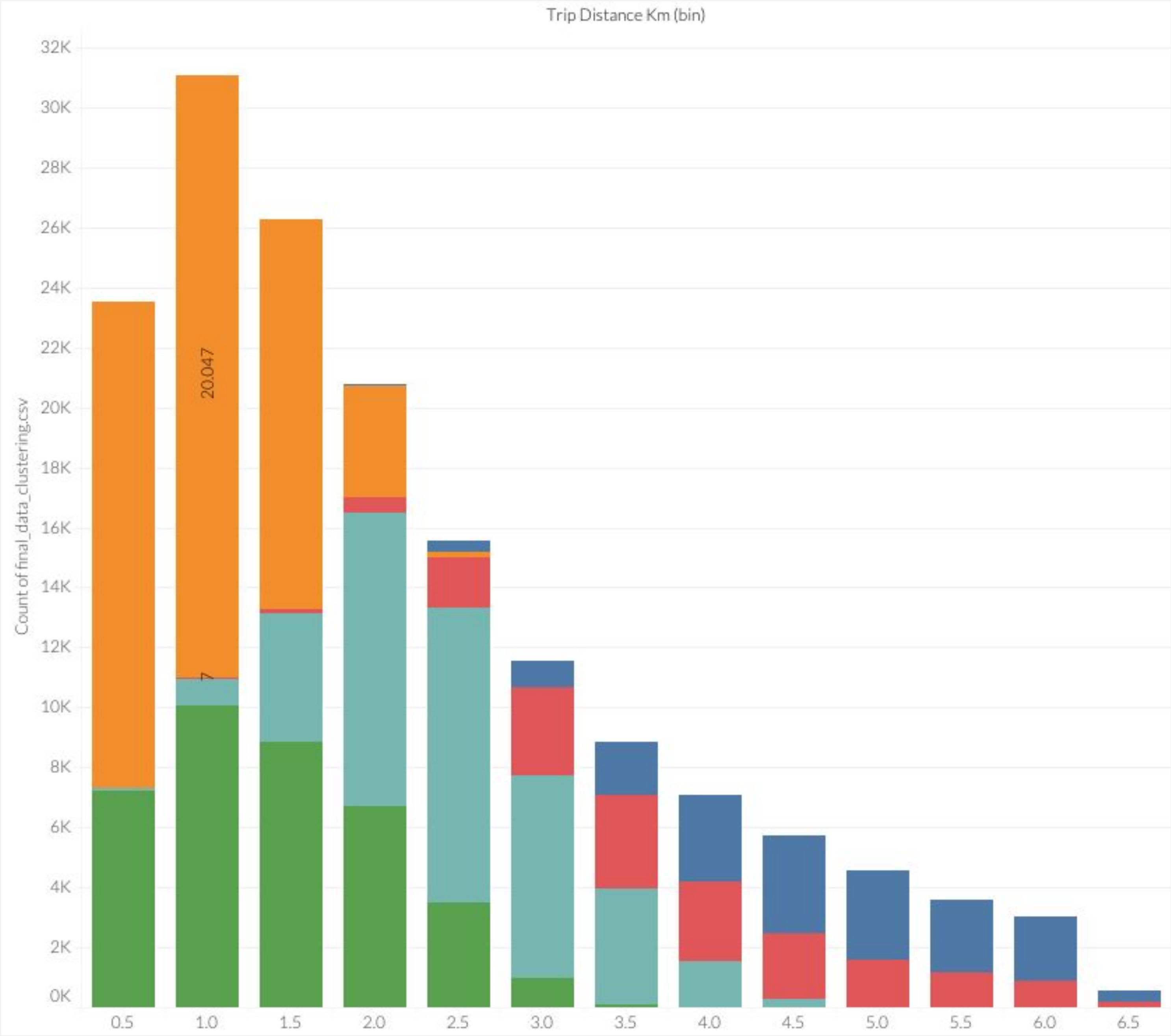
No	Variable	VIF
1	**pickup city**	**6.92**
2	**dropoff city**	**6.54**
3	day of week	3.12
4	part of day	2.81
5	pickup latitude	2.27
6	dropoff latitude	2.14
7	pickup longitude	2.06
8	dropoff longitude	2.00
9	day of week	1.07

Cluster **Silhouette score is 0.62**, **Calinski harabasz 268065,25 and Davies Bouldin 0.60**. Score shows a good result to divide customer segmentation

Total of order by day



Total order by trip distance



All-Week Long Hauls (0)

- With trips distributed throughout the weekday, longer distances, and higher fares

Early Week Short Hops (1)

- Short distances and low fares, actively order all week with trips concentrated early in the week,

Weekend Midday Outings (2)

- Activity peaks on Saturday, with midday pickups and moderate fares and distances

Pre-Weekend Longer Rides (3)

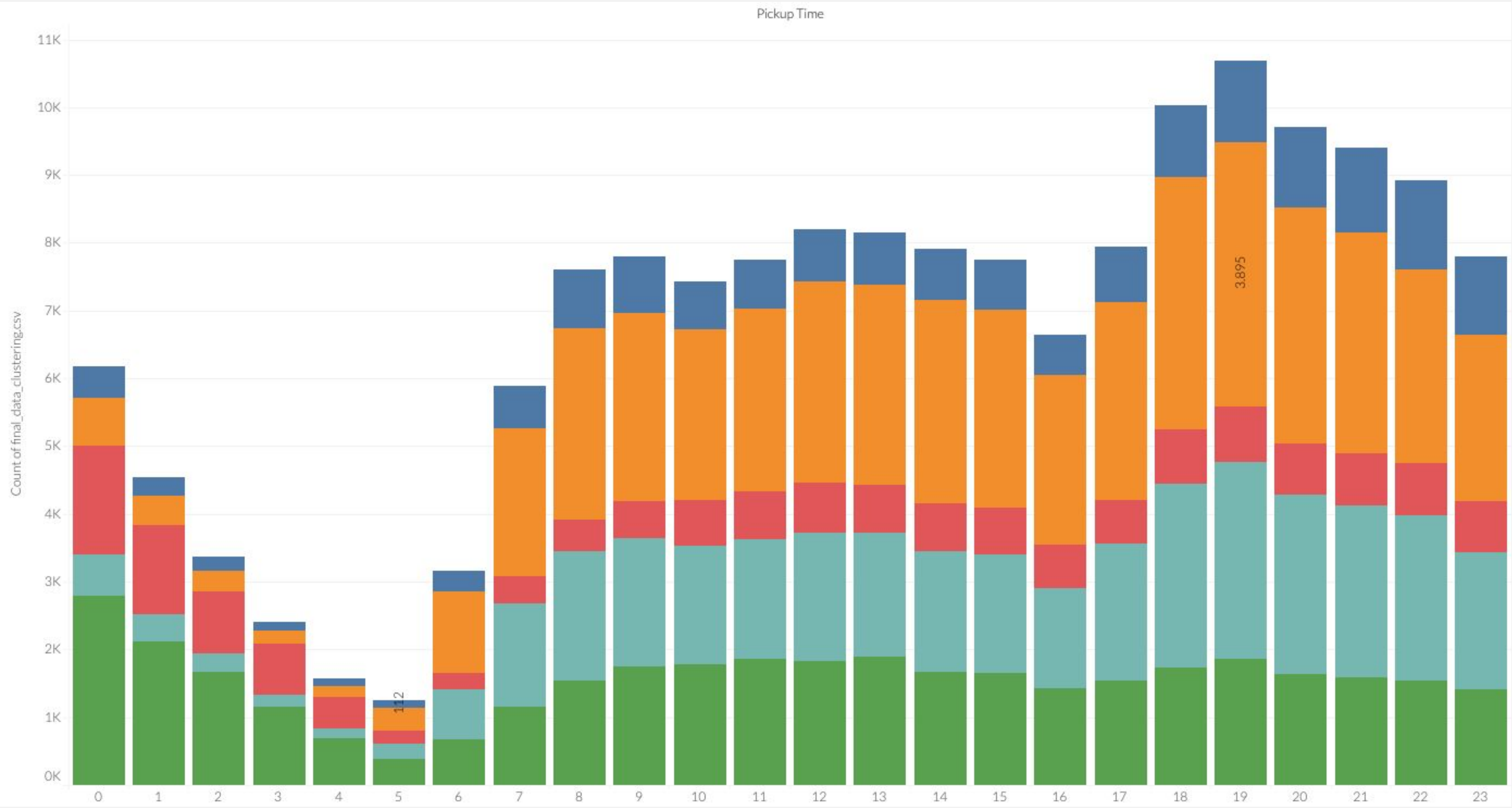
- Similar to Cluster 2 but with earlier pickups and longer distances

Versatile Afternoon Commuters (4)

- A versatile cluster with moderate trip lengths and fares, active throughout the early to mid-week afternoons

Each cluster behavior reveals by the **Trip distance, Day, and Order Hour**

Total order by hourly each cluster



All-Week Long Hauls (0)

- With trips distributed throughout the weekday, longer distances, and higher fares

Early Week Short Hops (1)

- Short distances and low fares, actively order all week with trips concentrated early in the week,

Weekend Midday Outings (2)

- Activity peaks on Saturday, with midday pickups and moderate fares and distances

Pre-Weekend Longer Rides (3)

- Similar to Cluster 2 but with earlier pickups and longer distances

Versatile Afternoon Commuters (4)

- A versatile cluster with moderate trip lengths and fares, active throughout the early to mid-week afternoons

Cluster-Based Insights and Recommendations:

1. All-Week Long Hauls (Cluster 0):

- Insight: Distributed trips throughout the weekday with longer distances and higher fares could indicate a mix of commute and leisure trips.
- Recommendation: Introduce dynamic pricing strategies to maximize profits during peak hours, and promote shared rides for cost-conscious travelers on longer trips.

2. Early Week Short Hops (Cluster 1):

- Insight: Frequent short-distance trips, primarily early in the week, suggest routine commutes or errands.
- Recommendation: Offer loyalty discounts or a subscription model for regular commuters to ensure retention.

3. Weekend Midday Outings (Cluster 2):

- Insight: Peak activity on Saturday midday with moderate distances and fares suggests leisure or shopping trips.
- Recommendation: Partner with shopping centers or tourist attractions for promotions to encourage weekend use.

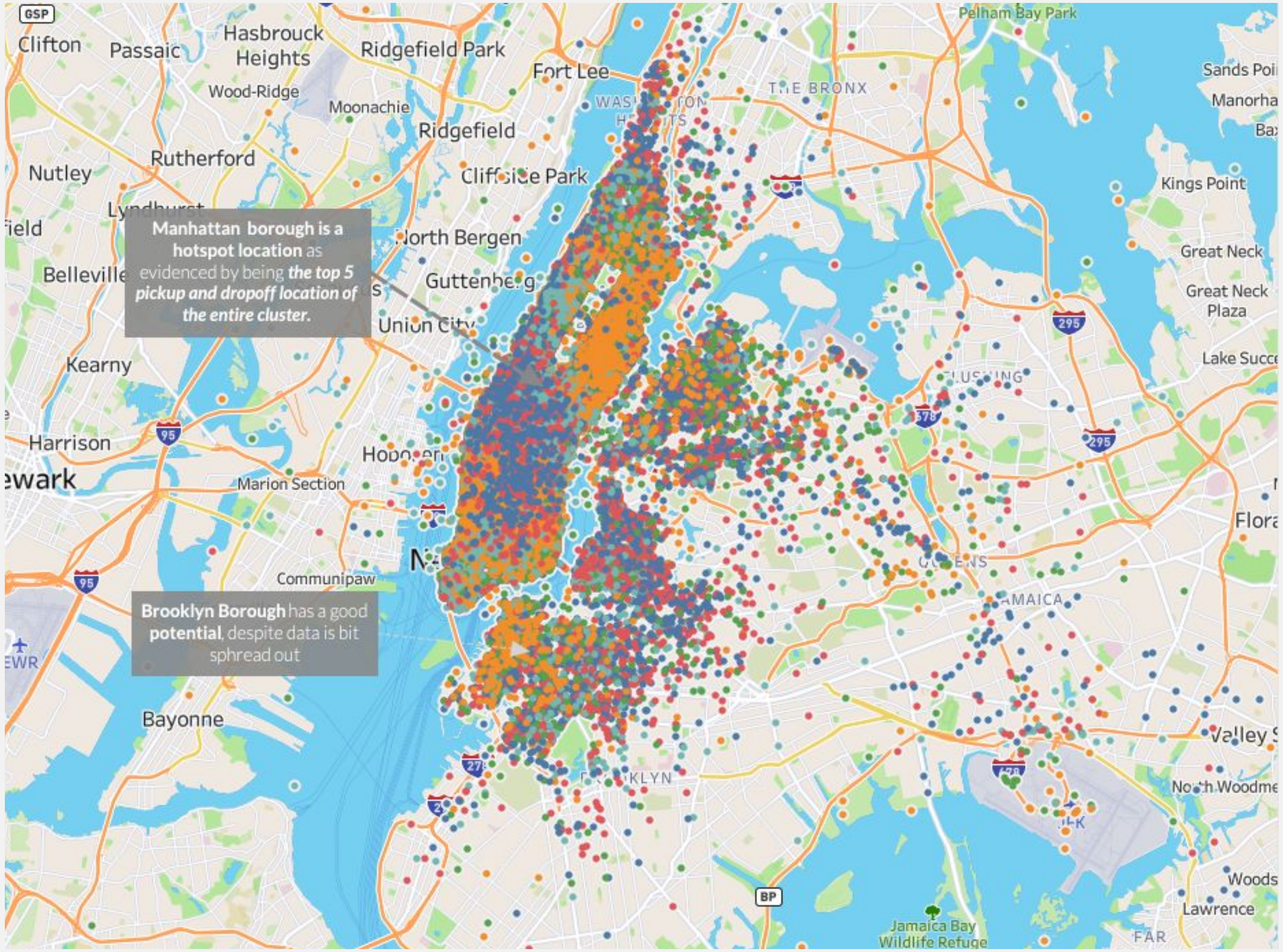
4. Pre-Weekend Longer Rides (Cluster 3):

- Insight: Similar to Cluster 2 but with earlier pickups and longer distances, possibly reflecting weekend getaway preparations and possibly out of town..
- Recommendation: Create weekend packages or offer tailored services for group outings to capture this market segment.

5. Versatile Afternoon Commuters (Cluster 4):

- Insight: Trips are moderate in both length and fare, consistent throughout early to mid-week afternoons, possibly from varied customer needs.
- Recommendation: Enhance mobile app features to cater to the diverse needs of this group, like scheduling rides in advance or choosing ride types.

Manhattan is a **hotspot** location, **Brooklyn** potential for **tourist** and **shopping location**



- All-Week Long Hauls
- Early Week Short Hops
- Weekend Midday Outings
- Pre-Weekend Longer Rides
- Versatile Afternoon Commuters

- All cluster actively on Manhattan Borough
- Some Early week short hops and weekend midday outings active near airport
- Coordinates in the Brooklyn area are more spread out, but there are points that are concentrated in an area which indicates possible tourist area.

Geospatial based Insight and Recommendation:

1. Hotspot Utilization:

- Insight: Manhattan is a major hotspot, indicating high demand.
- Recommendation: Increase fleet availability in Manhattan to reduce wait times and improve customer satisfaction.

2. Brooklyn Potential:

- Insight: Activity in Brooklyn is more spread out, with certain concentrated areas potentially indicating tourist locations or underserved areas.
- Recommendation: Conduct targeted marketing campaigns in these specific Brooklyn areas to capture potential tourist or local commuter markets. Consider partnership opportunities with businesses in these tourist areas to offer exclusive deals or discounts.

3. Airport Rides:

- Insight: Given the cluster activities near airports, there is significant demand for airport rides.
- Recommendation: Offer fixed rates or discounts for airport rides to attract more customers looking for reliable transportation to and from airports

4. Expansion Opportunities:

- Insight: Some areas outside of Manhattan have less activity, which may indicate underserved markets or potential for expansion.
- Recommendation: Explore expansion into these areas by offering introductory rates or service guarantees to build the customer base.

Random Forest and **XGBoost** have **good** scores for predicting fares, **XGBoost with tuning** outperform with R2 score of **0.77.**

Evaluation Metrics

No	Algorithm	MSE	MAE	R2(Squared)
1	Random Forest	3.55	1.38	0.74
2	Random Forest Tuning	3.27	1.31	0.76
3	XGBoost	3.40	1.35	0.75
4	**XGBoost Tuning**	**3.15**	**1.28**	**0.77**

- The lower MSE for XGBoost Tuning suggests it is generally more reliable for predicting fares, making smaller errors in the squared term, which can significantly impact fare estimation accuracy. This is crucial in fare prediction as large errors can lead to dissatisfaction for both drivers (overestimates) and customers (underestimates).
- MAE provides a clear measure of average error.With a lower MAE, XGBoost Tuning again shows it can predict fare values closer to the actual charges. On average, the model's predictions are about $1.35 off from the actual fare amounts.
- R2 of 0.7737 means that about 77.37% of the variability in Uber fares can be explained by the XGBoost Tuning model.

Summary Insight and Recommendation:

1. Identifying Customer Segments:
    - Insight: Customer Segments reveal by riding pattern like fare, distance, week, hour and part of day.
    - Recommendation: Improving service offerings to meet the specific needs of different user segments, targeting daily commuters with monthly flat rates, offering weekend discounts to casual riders, and partnering with hotels or businesses for tourists and business travelers.
2. Improving Operational Efficiency and Cost Reduction:
    - Insight: The majority of trips have a pattern tied to time and location.Higher demands during mid-day to night and lower demands in early morning hours.
    - Recommendation: To improve operational efficiency and reduce costs by 5%, Uber could:
        - Deploy dynamic scheduling where driver deployment aligns with demand patterns to reduce idle time.
        - Diversify vehicle type to match passenger size.
        - Optimize ride-sharing opportunities, especially for hotspot location to and from high-demand locations like Manhattan and Brooklyn borough.

Summary Insight and Recommendation:

3. Dynamically Adjusting Fare Prices:
   - Insight: There is a correlation between fare prices, demand elasticity, and customer churn. Price-sensitive customers may be more prone to churn if fares frequently surge.
   - Recommendation: To maximize revenue while maintaining fairness and transparency:
     - Implement a XGBoost model to adjust prices dynamically, taking into account customer price sensitivity.
     - Clearly communicate how dynamic pricing works within the app, giving customers insights into peak times and potential fare changes.
     - Offer fare estimates with a maximum cap to build trust and avoid surprises.
4. Reducing Customer Churn Rate:
   - Insight: Churn can be influenced by customer experience, fare discrepancies, and the availability of alternatives.
   - Recommendation: To reduce the churn rate by 5%:
     - Personalize customer experiences based on individual travel history, such as preferred temperature and vehicle type.
     - Implement a feedback loop to resolve issues quickly and improve service quality.
     - Introduce features that increase stickiness to the app, such as ride scheduling, subscription models, or family accounts.
     - Engage with customers through targeted offers and discounts that make frequent use of the service more rewarding.
     - Develop a loyalty program where frequent riders can earn points to lock in lower rates

Thanks!

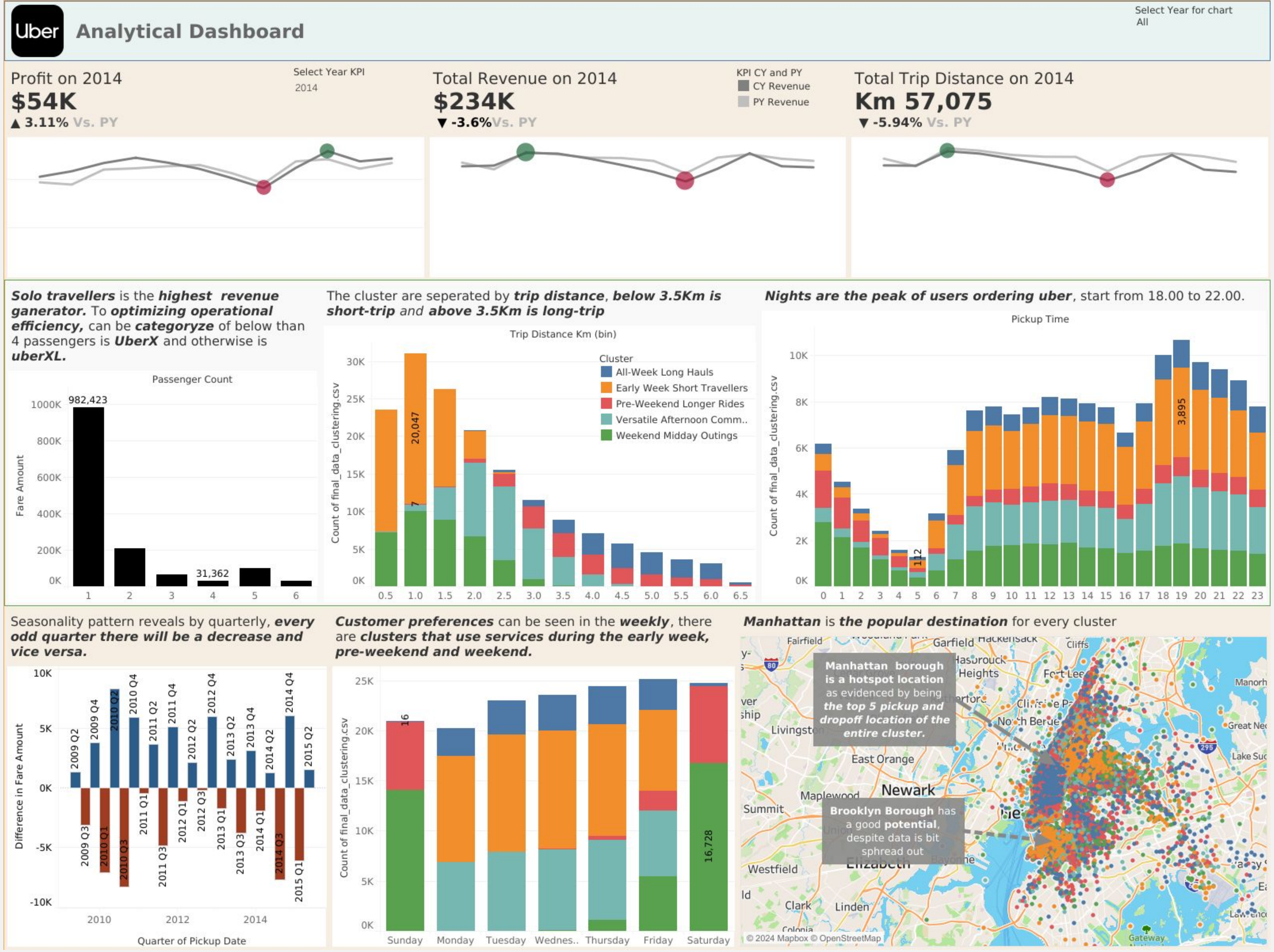
andi_tama@outlook.com
https://www.linkedin.com/in/andi--pratama/

Appendix

Dataset

Jupyter Notebook

Dashboard


Uber
Analytical Dashboard
Select Year for chart
All
Profit on 2014
$54K
▲ 3.11% Vs. PY
Select Year KPI
2014
Total Revenue on 2014
$234K
▼ -3.6% Vs. PY
KPI CY and PY
CY Revenue
PY Revenue
Total Trip Distance on 2014
Km 57,075
▼ -5.94% Vs. PY
Solo travellers is the highest revenue ganerator. To optimizing operational efficiency, can be categoryze of below than 4 passengers is UberX and otherwise is uberXL.
Passenger Count
Fare Amount
982,423
31,362
The cluster are seperated by trip distance, below 3.5Km is short-trip and above 3.5Km is long-trip
Trip Distance Km (bin)
Count of final_data_clustering.csv
Cluster
All-Week Long Hauls
Early Week Short Travellers
Pre-Weekend Longer Rides
Versatile Afternoon Comm..
Weekend Midday Outings
20,047
Nights are the peak of users ordering uber, start from 18.00 to 22.00.
Pickup Time
3,895
Seasonality pattern reveals by quarterly, every odd quarter there will be a decrease and vice versa.
Difference in Fare Amount
Quarter of Pickup Date
Customer preferences can be seen in the weekly, there are clusters that use services during the early week, pre-weekend and weekend.
Sunday
Monday
Tuesday
Wednes..
Thursday
Friday
Saturday
16,728
Manhattan is the popular destination for every cluster
Manhattan borough is a hotspot location as evidenced by being the top 5 pickup and dropoff location of the entire cluster.
Brooklyn Borough has a good potential, despite data is bit sphread out
Newark
© 2024 Mapbox © OpenStreetMap

Appendix 1

Data Transformation to get important data

BEFORE

No	Variable
1	key
2	fare amount
3	pickup datetime
4	passenger count
5	pickup longitude
6	pickup latitude
7	dropoff longitude
8	dropoff latitude

AFTER

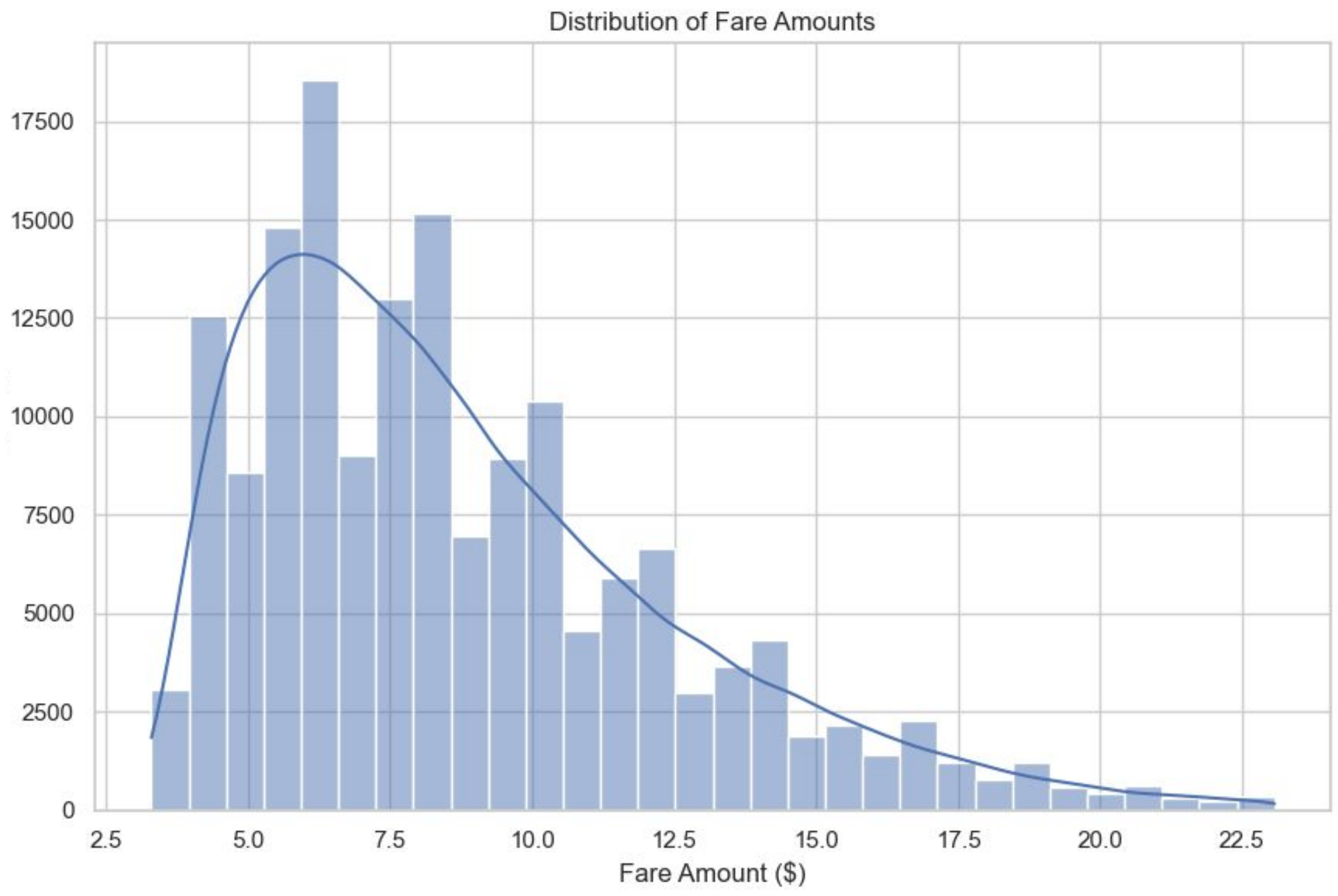
No	Variable
1	fare amount
2	pickup datetime
3	passenger count
4	pickup longitude
5	pickup latitude
6	dropoff longitude
7	dropoff latitude
8	key id

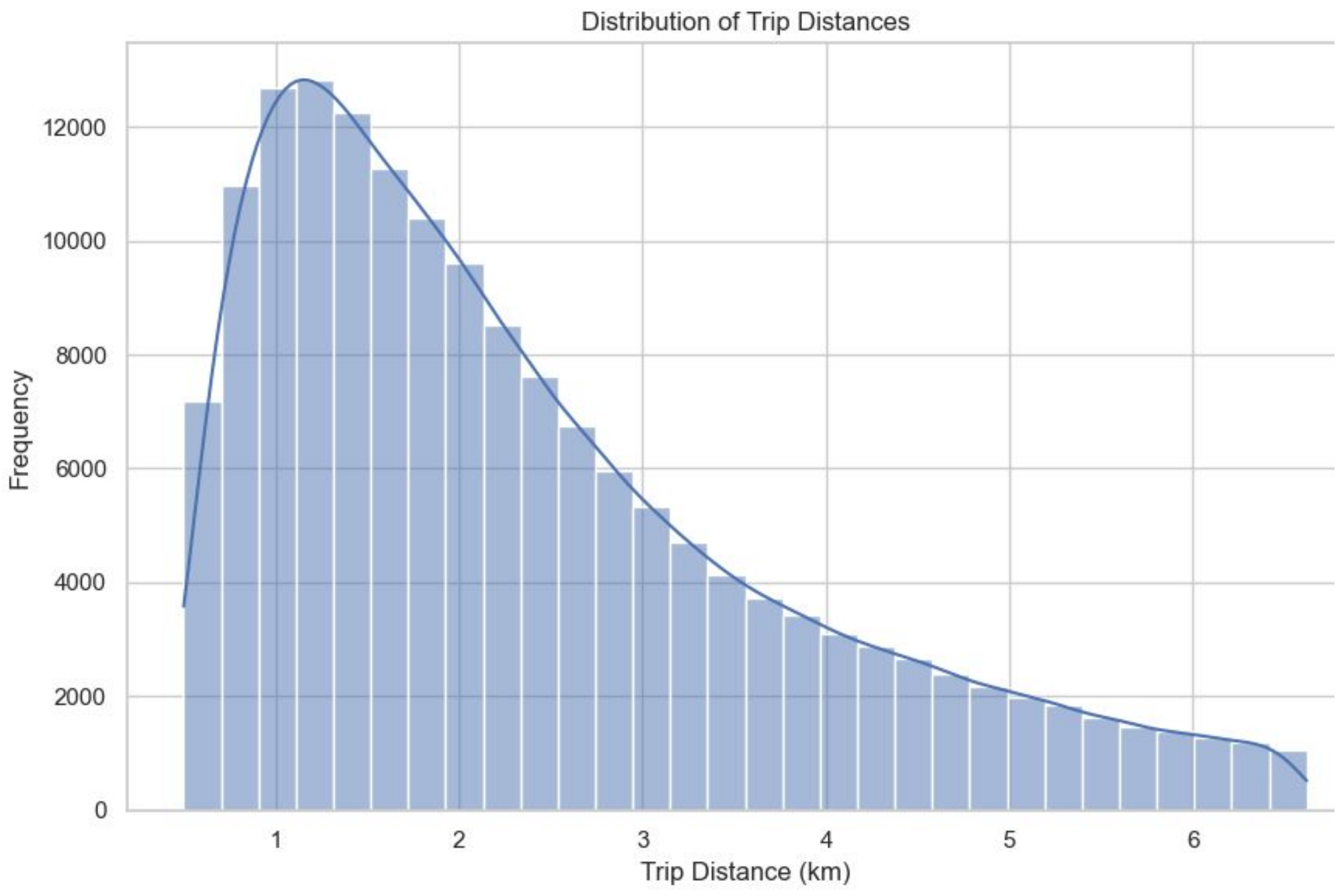
No	Variable
9	trip distance
10	pickup date
11	pickup time
12	pickup city
13	dropoff city
14	pickup hour
15	part of day
16	part of week

No	Variable
17	part of week name
18	month
19	weekend
20	cluster
21	cost
22	profit
23	pickup last time
24	churned

Appendix 2

Distribution of fare amount and trip distance are **right skew (positive skew)**,therefore **linear regression isn‘t match** to predict fare.





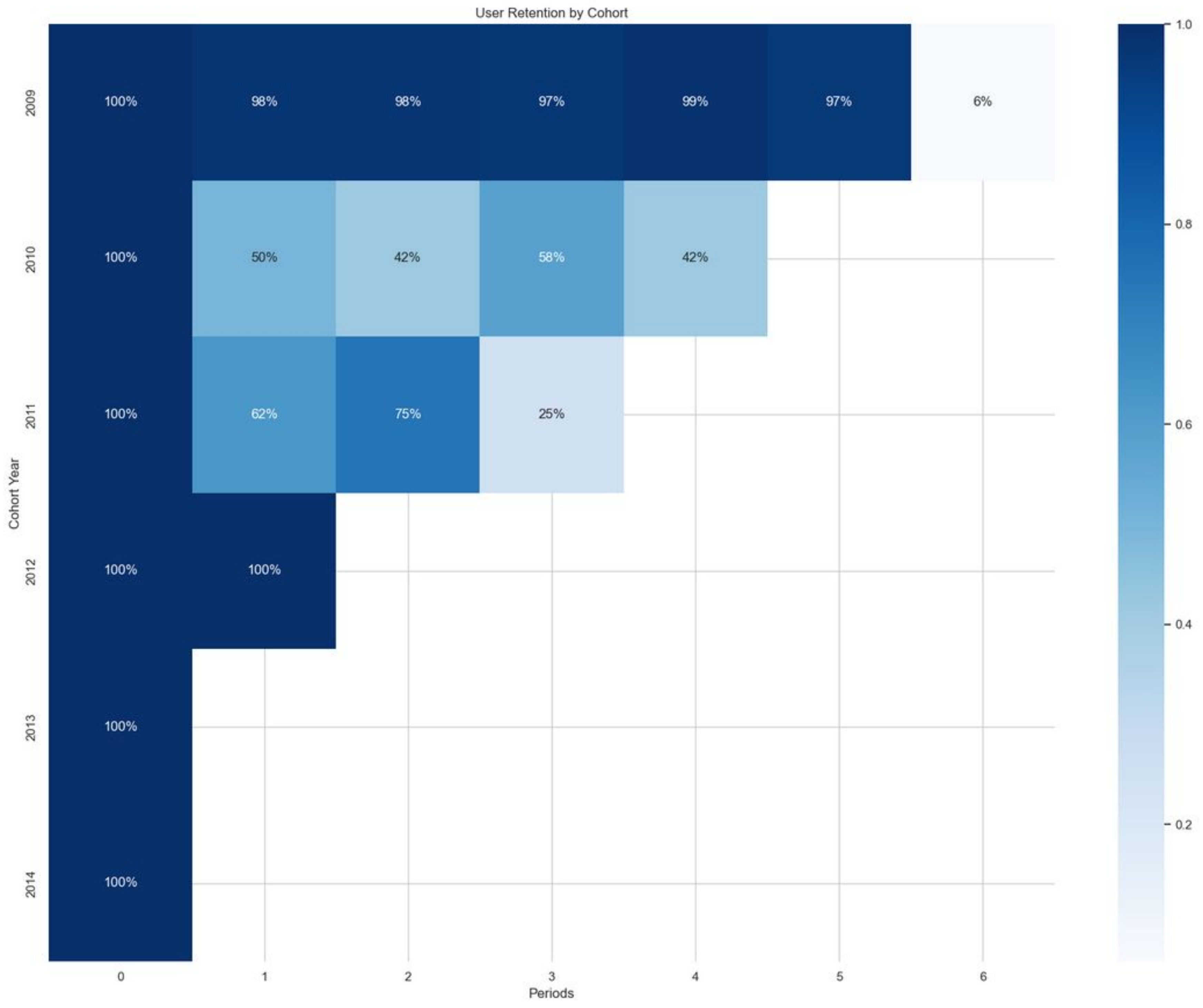
Appendix 3

Cluster **Silhouette score** is **0.62**, **Calinski harabasz 268065,25** and **Davies Bouldin 0.60.** Score shows a good result to divide customer segmentation

	day of week				trip distance				fare amount			
Cluster	Count	Mean	Min	Max	Count	Mean	Min	Max	Count	Mean	Min	Max
0	53237	2	0.5	4	53237	1	1	3	53237	6	3	12
1	17006	2	0.5	6	17006	5	2	7	17006	15	9	23
2	37367	5	2	6	37367	2	1	4	37367	7	3	14
3	17103	5	2	6	17103	4	1	7	17103	13	7	23
4	37399	2	0.5	4	37399	3	1	5	37399	10	6	17

Appendix 4

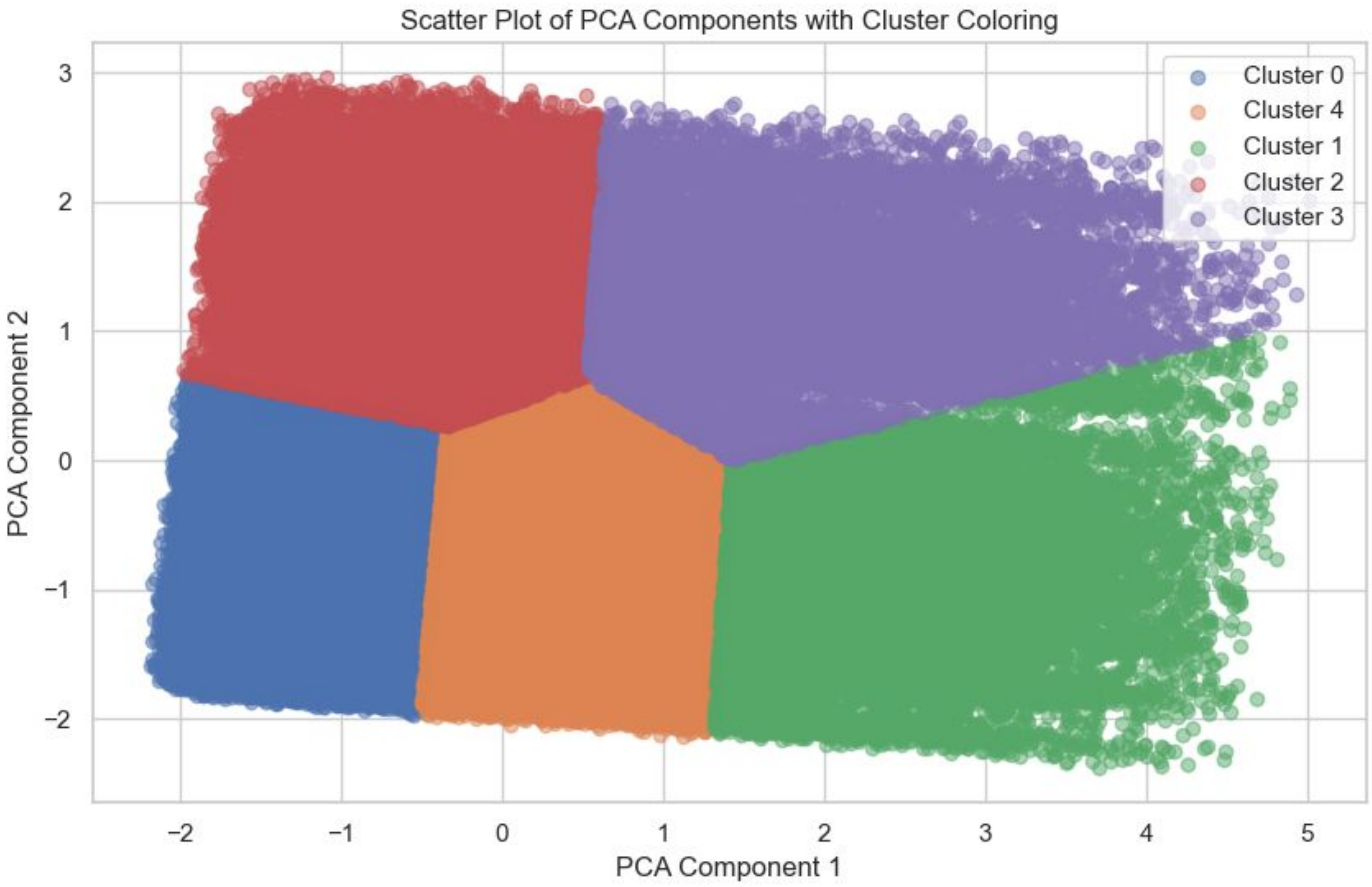
The yearly retention rate **looks good**, although there is a decline but in the following year there is an **recovery**.



- The retention rate in 2009 was very good every periods.
- Retention in 2011 period 1 declined, but recover on next year

Appendix 5

Principal Component Analysis (PCA) perform well to define each cluster, with clear boundaries between of them

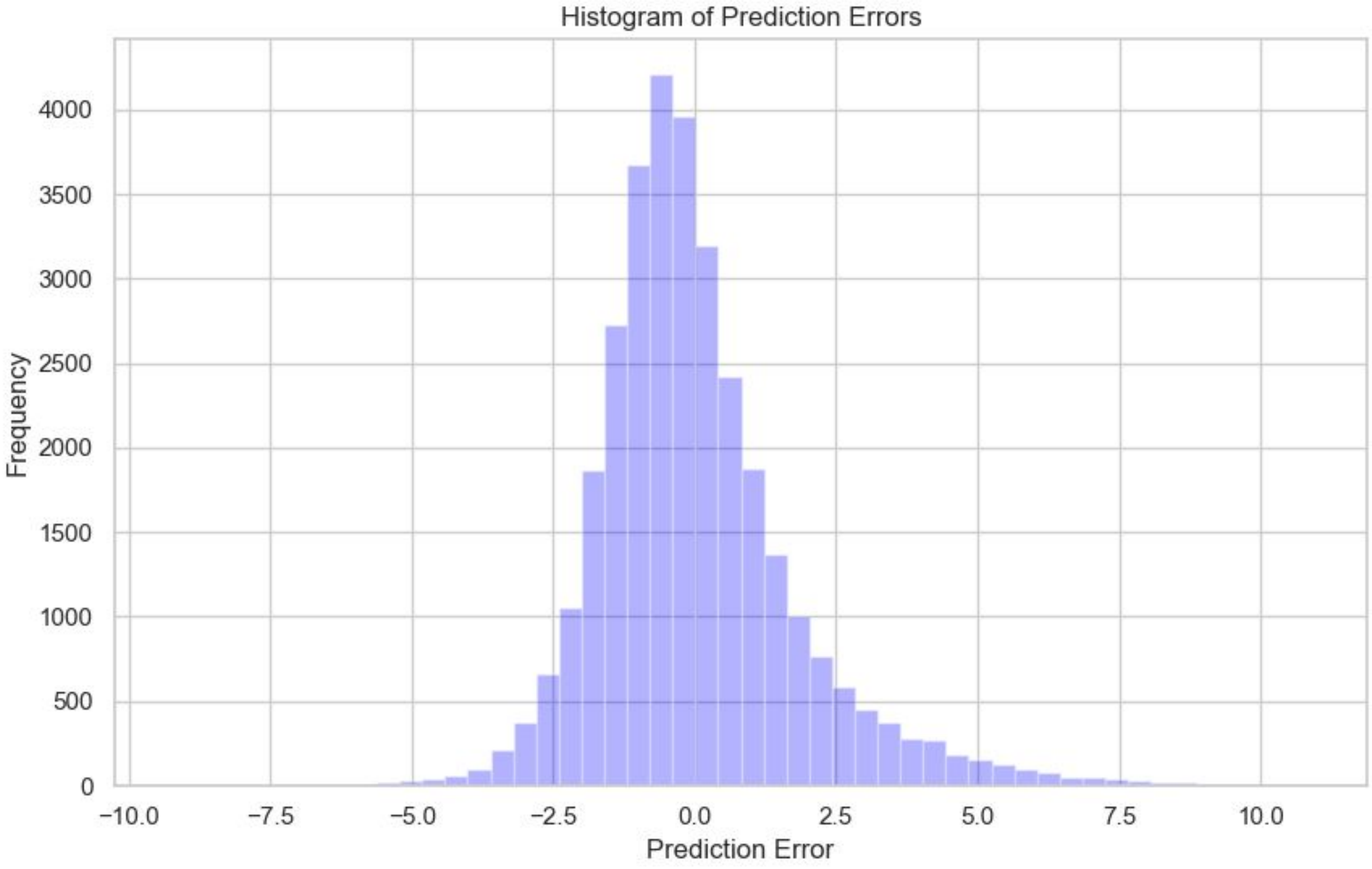


- Clusters 4 (orange) and 1 (green) seem to be positioned adjacent to each other with a relatively clear boundary.
- Cluster 3 (purple) is above clusters 1 and 4, sharing a border with both.
- Cluster 2 (red) is to the left, separated quite distinctly from clusters 1 and 4 by PCA Component 1.
- Cluster 0 (blue) is below with clusters 1 and 4.

Actual fare Vs. XGBoost Tuning Vs. Random Forest Tuning

No	Actual Fare	XGBoost Tuning	Random Forest Tuning
1	4.50	4.69	5.30
2	4.50	6.27	6.08
3	10.50	6.75	6.73
4	8.10	7.88	7.90
5	6.90	6.78	6.23
6	10.10	14.39	14.58
7	7.00	7.34	7.19
8	6.50	5.63	6.14
9	4.00	4.85	5.32
10	5.30	6.39	6.32

Distribution of prediction XGBoost with hyperparameter tuning is **bell shaped** and **mostly near to 0**



- Histogram is centered around a prediction error of zero, which is desirable as it indicates that on average the model's predictions are close to the actual values.
- Distribution is approximately normal (bell-shaped), which is common in prediction errors for well-fitting models.
- The errors range from approximately -5 to about 7.5.
- Most of the prediction errors are clustered around the center, between -2.5 and 2.5, indicating that the model is often quite accurate, with the highest frequency of errors being very small.