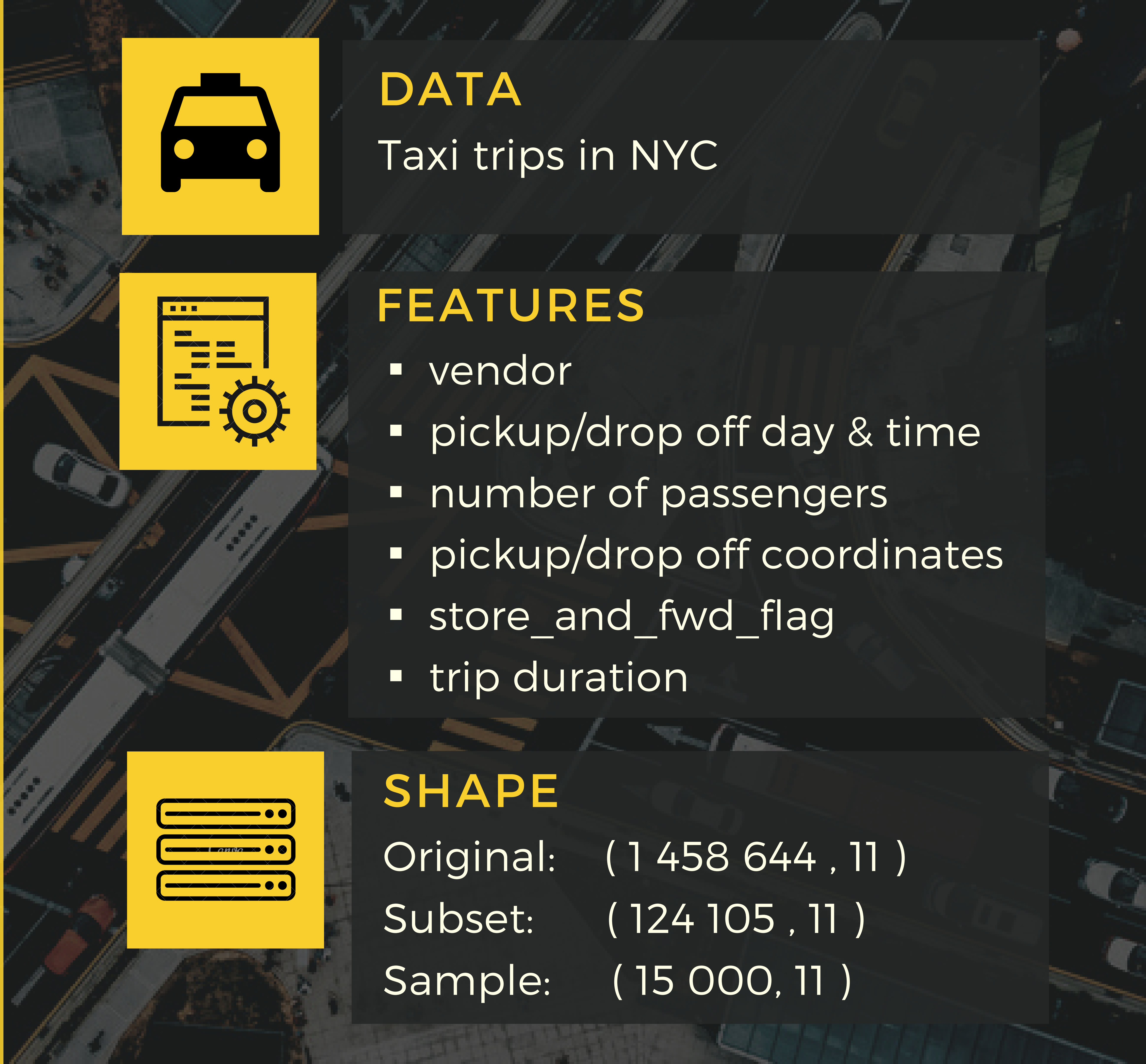
UNSUPERVISED ML PROJECT

NYC taxi trip duration

LAUMA USTUPA

THE DATA



DATA

Taxi trips in NYC

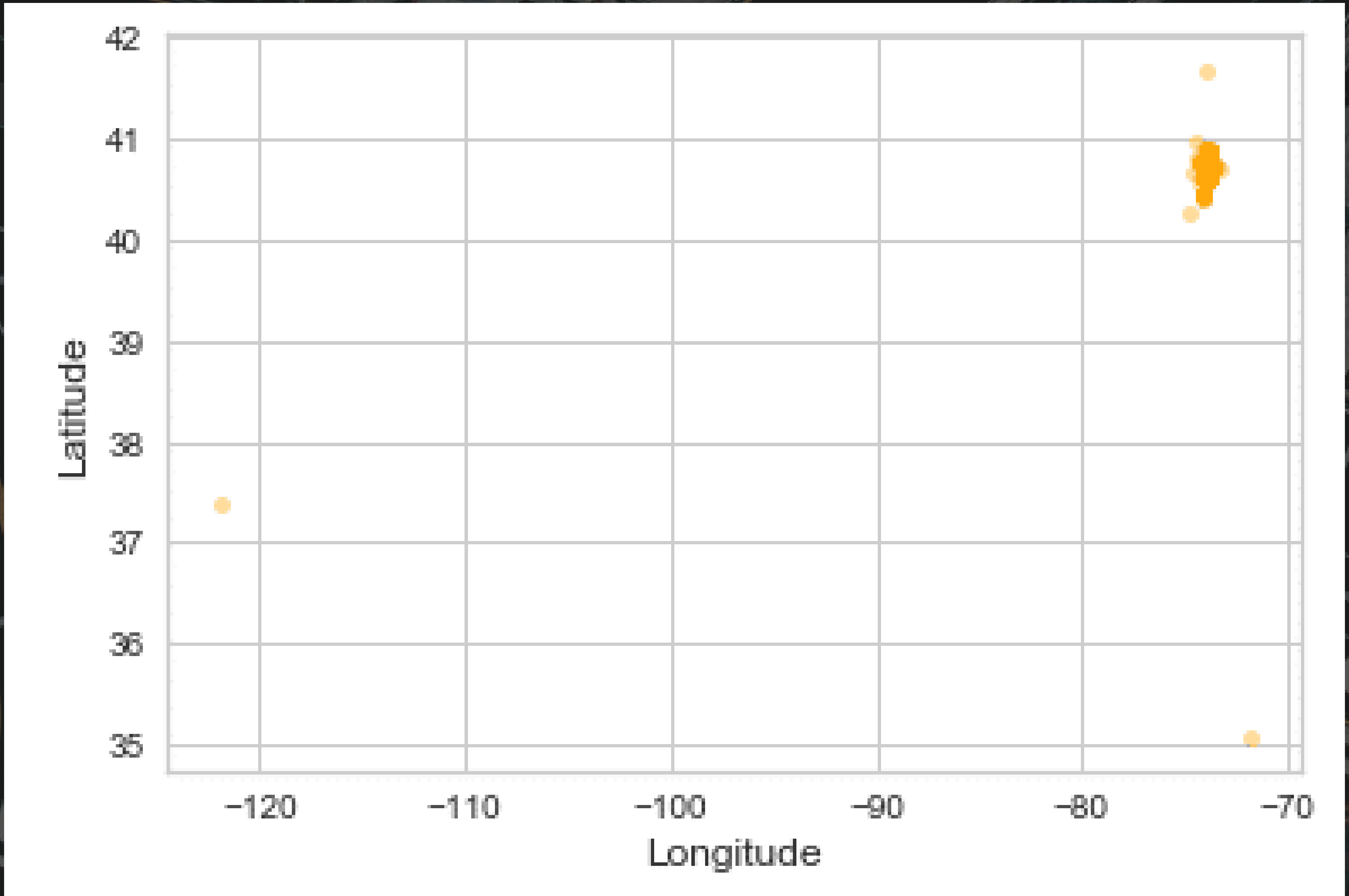
FEATURES

- vendor
- pickup/drop off day & time
- number of passengers
- pickup/drop off coordinates
- store_and_fwd_flag
- trip duration

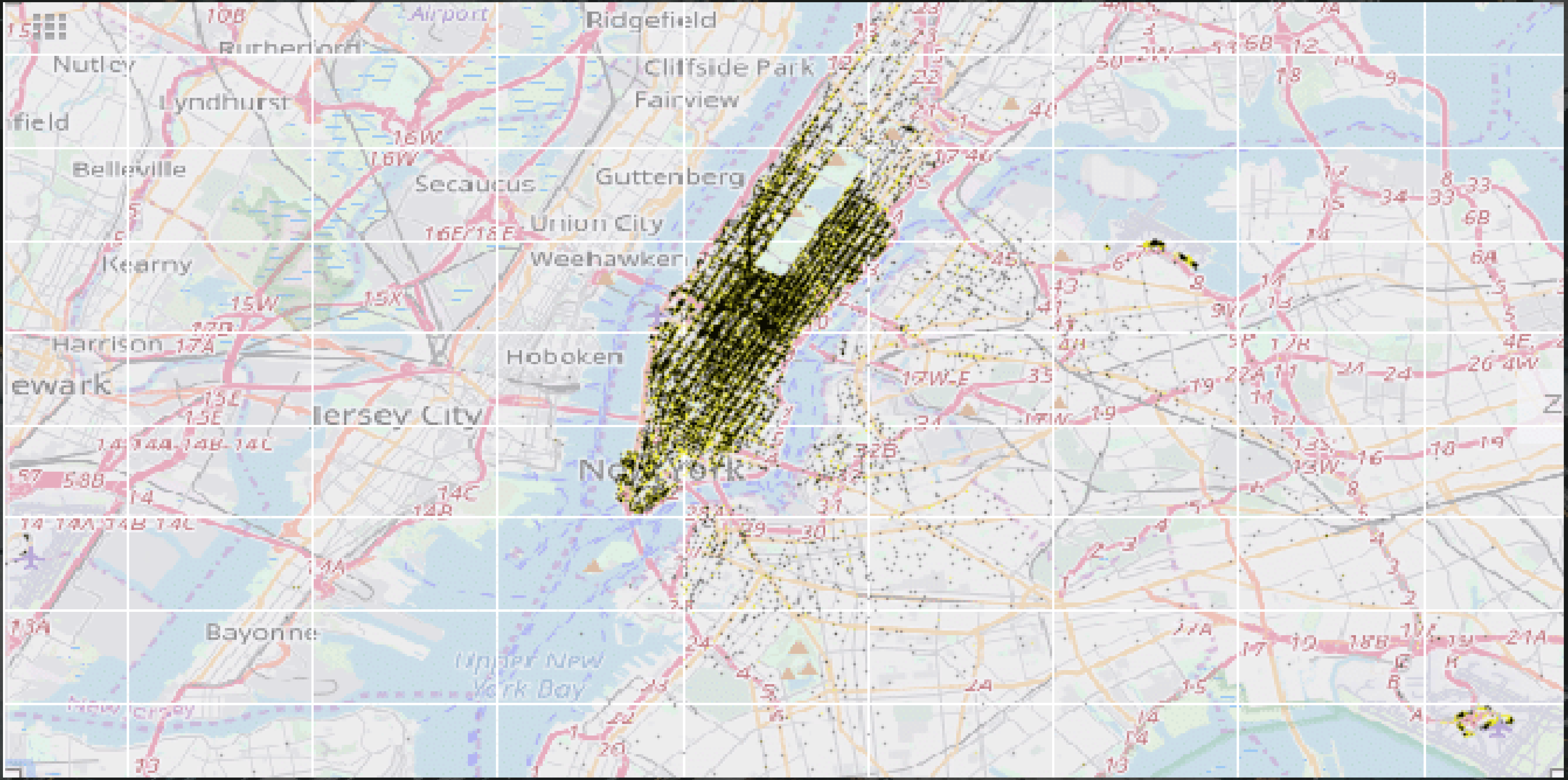
SHAPE

Original:	(1 458 644 , 11)
Subset:	(124 105 , 11)
Sample:	(15 000, 11)

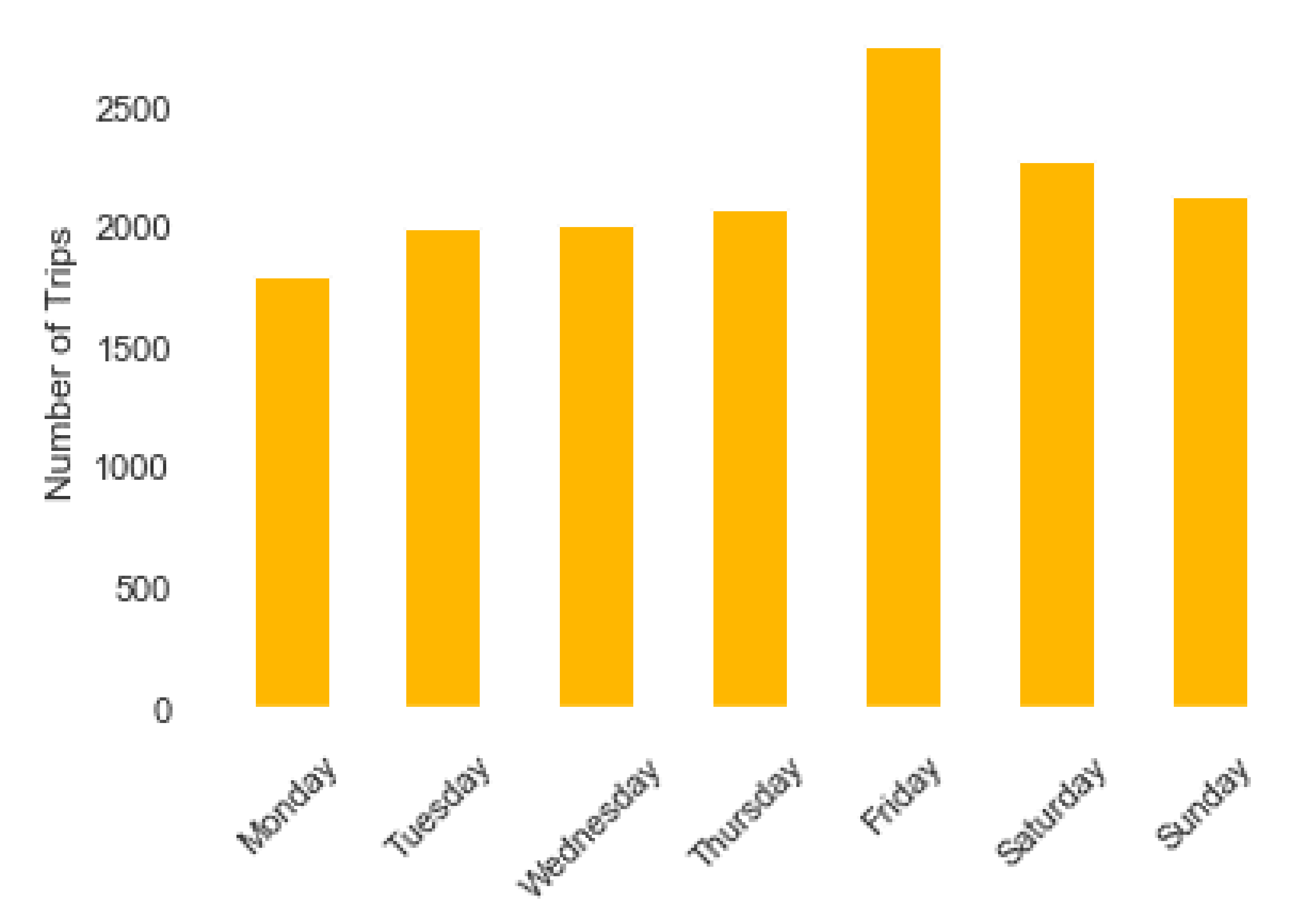
PICKUP & DROP OFF LOCATIONS

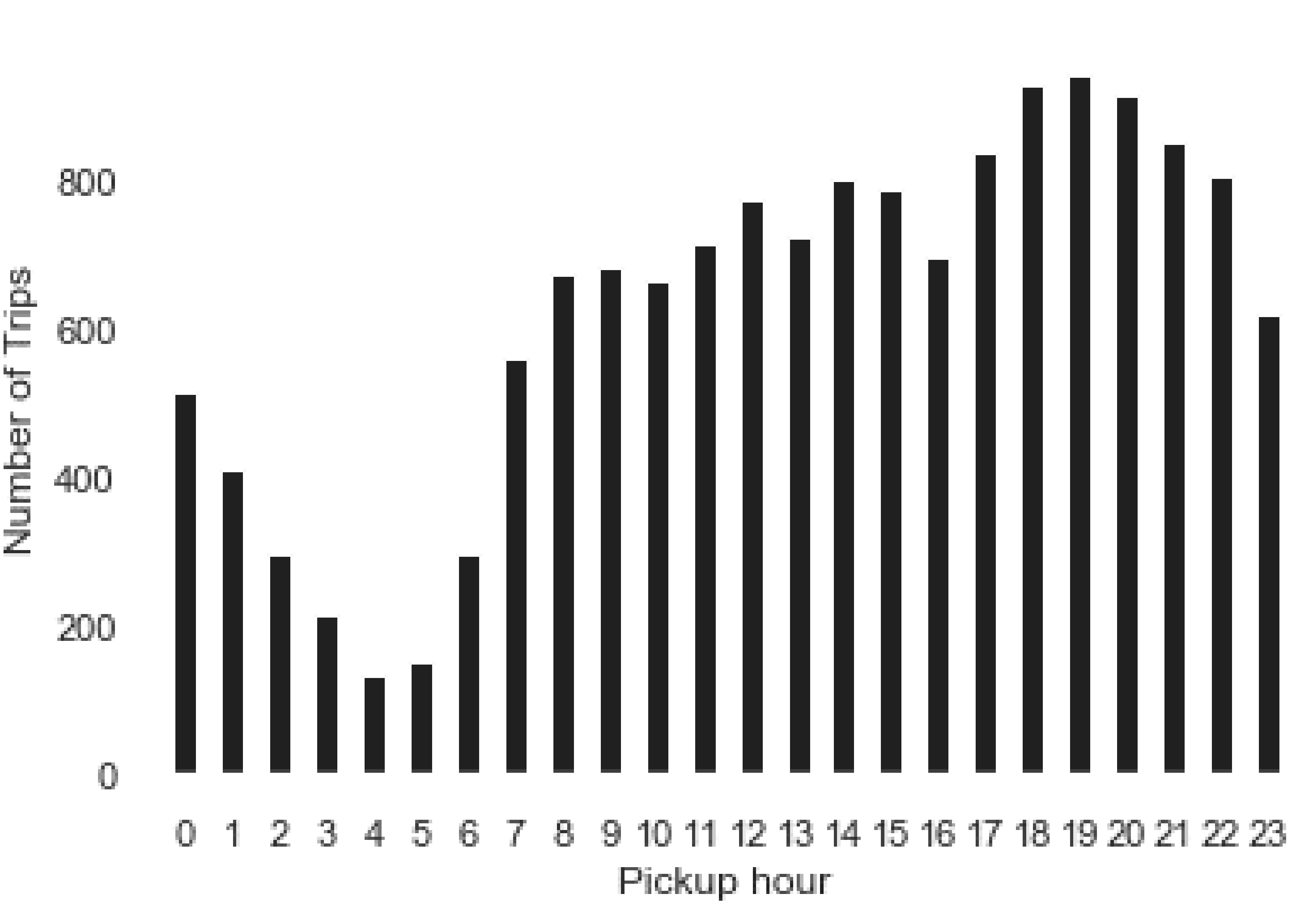


PICKUP & DROP OFF LOCATIONS

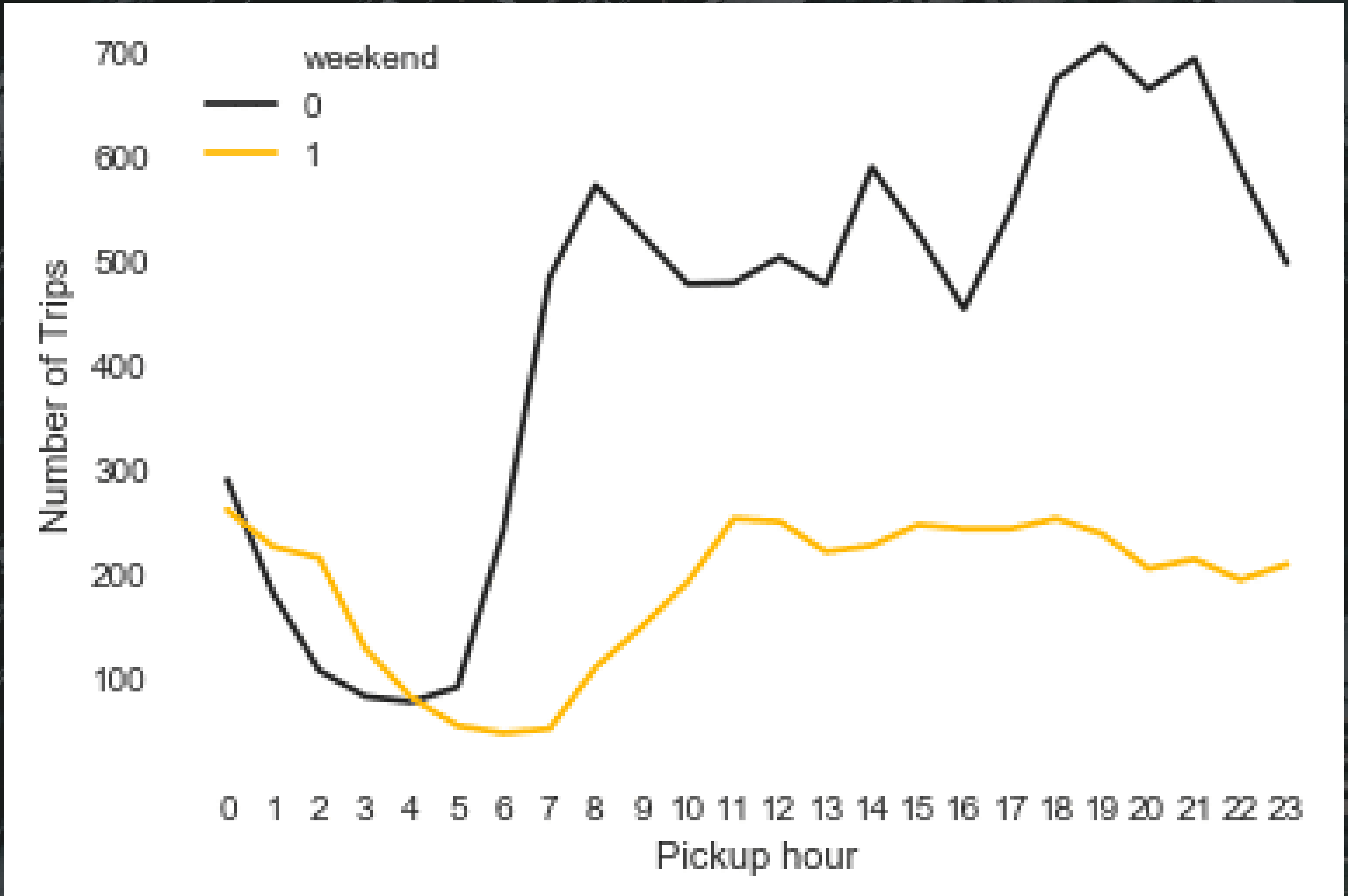


TRIPS PER DAY & PER HOUR

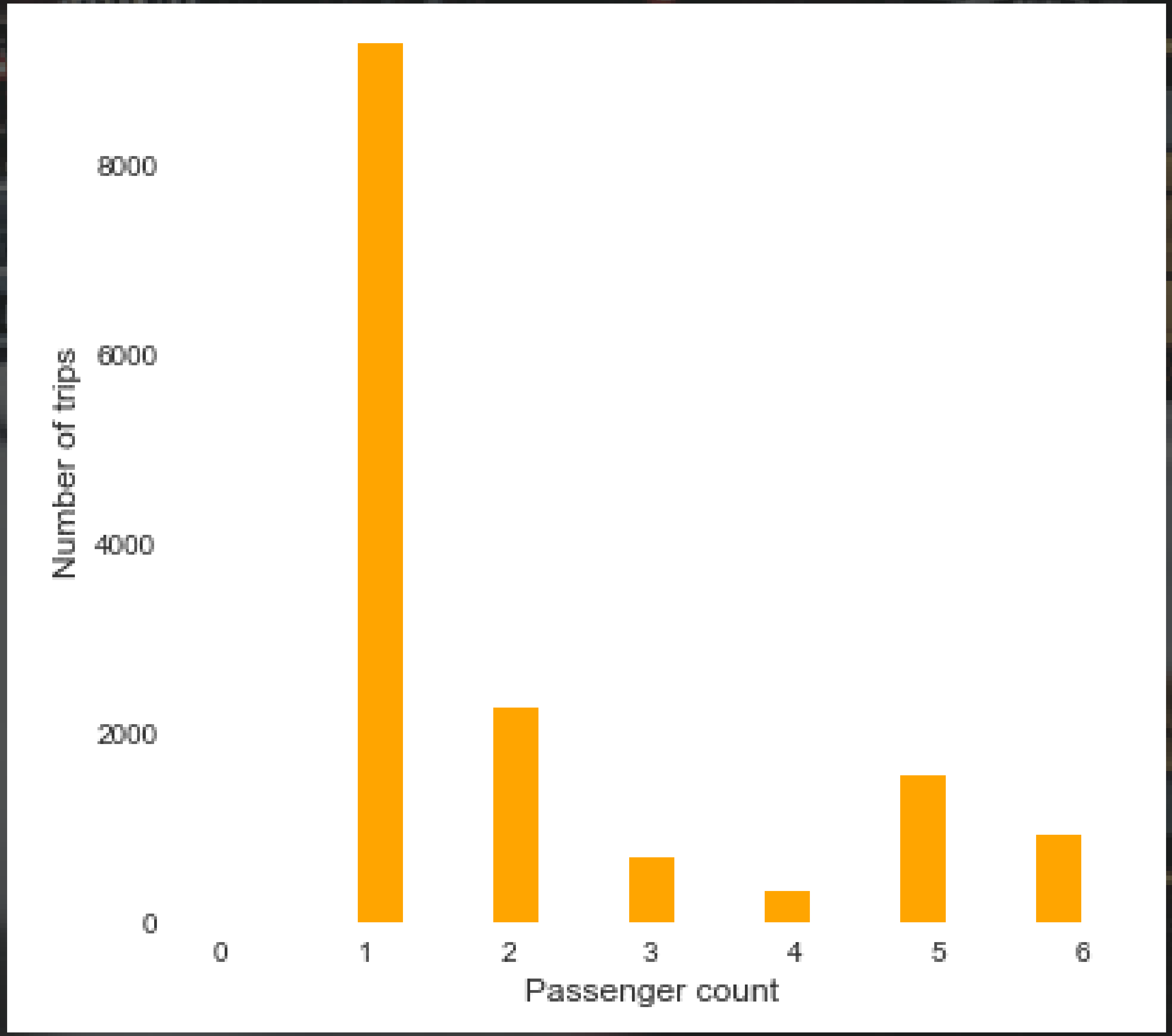


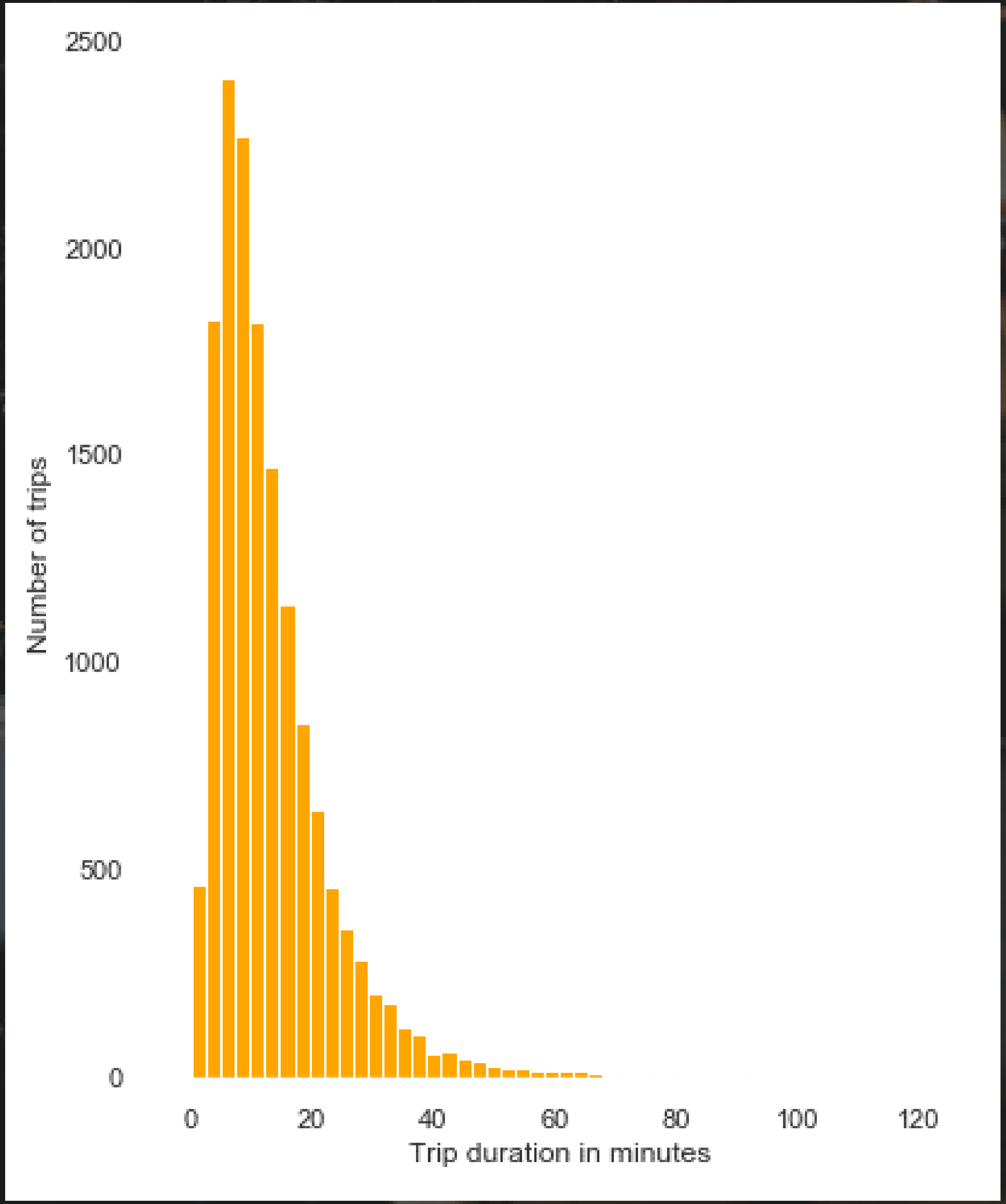


TRIPS DURING WORK WEEK VS WEEKEND



Passenger count & trip distribution





Pipeline

DATA CLEANING

- Outliers: far away trips
- Exclude trip duration

2 STEP APPROACH FOR CLUSTERING

NEIGHBOURHOOD CLUSTERING

- K-means

DATA ENGINEERING

- OHE for neighbourhood, day of week & hour
- Calculate trip distance
- Final number of features: 85

FURTHER CLUSTERING

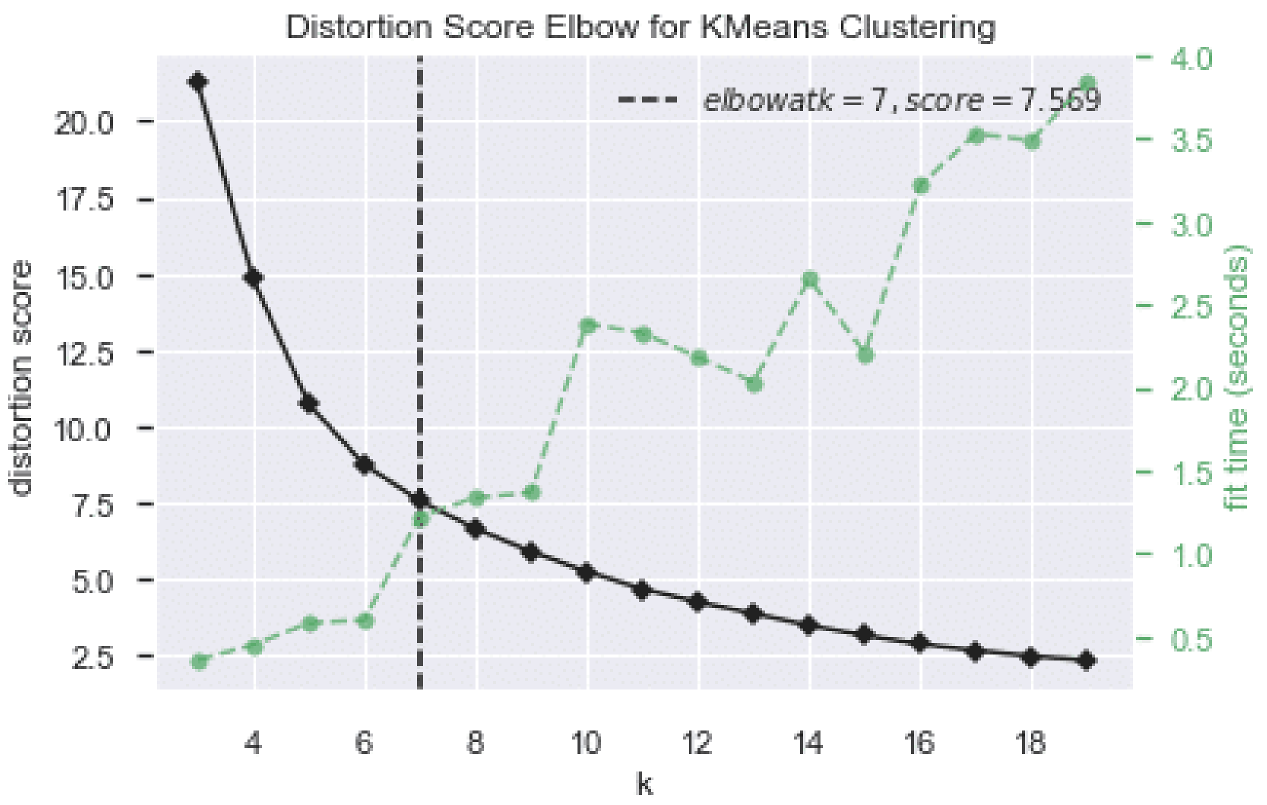
- PCA
- K-means/ DBSCAN/ GMM/ Hierarchical clustering

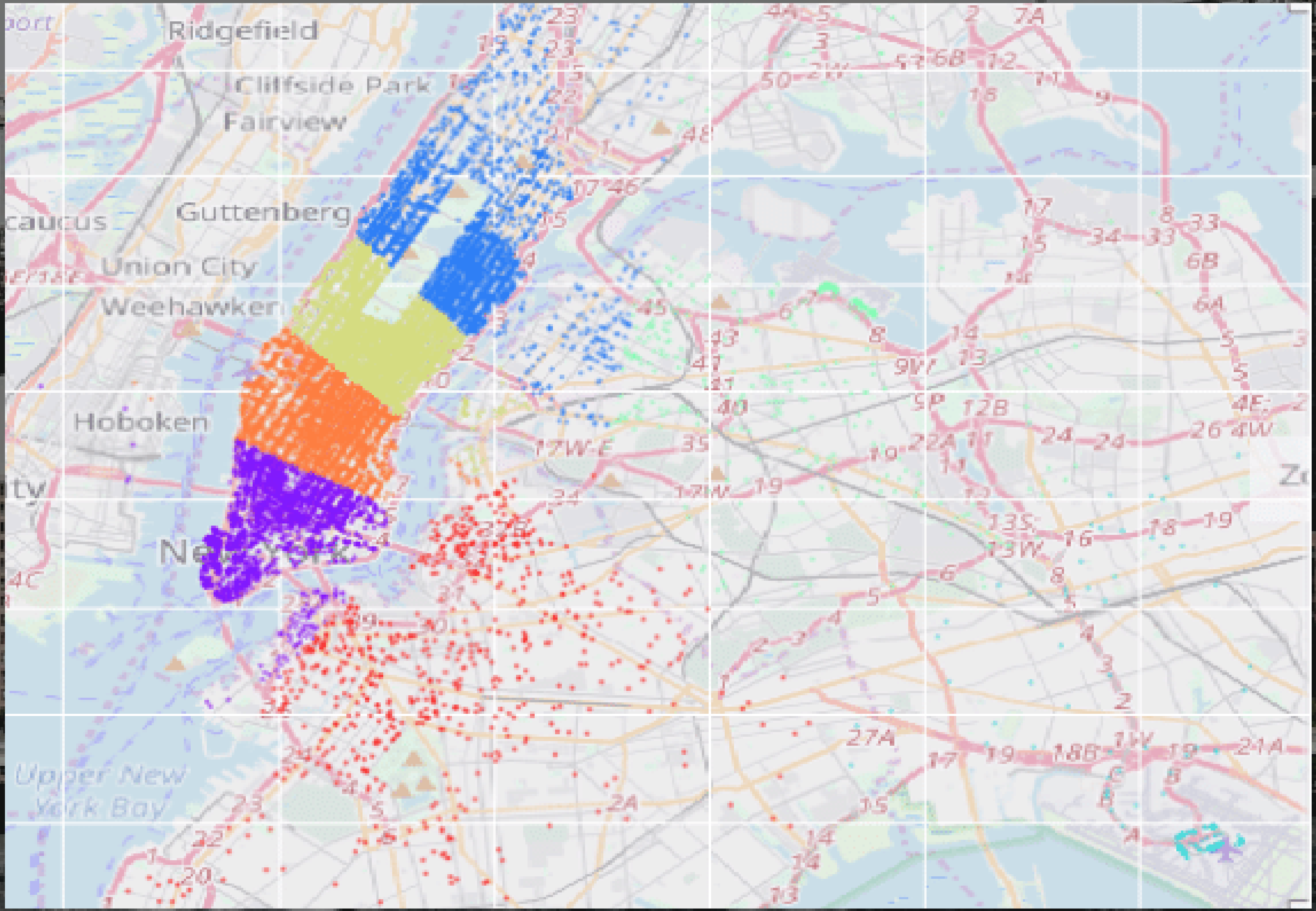
MODEL EVALUATION

- Silhouette score & Plotting

NEIGHBOURHOOD CLUSTERING

K-MEANS & ELBOW CURVE



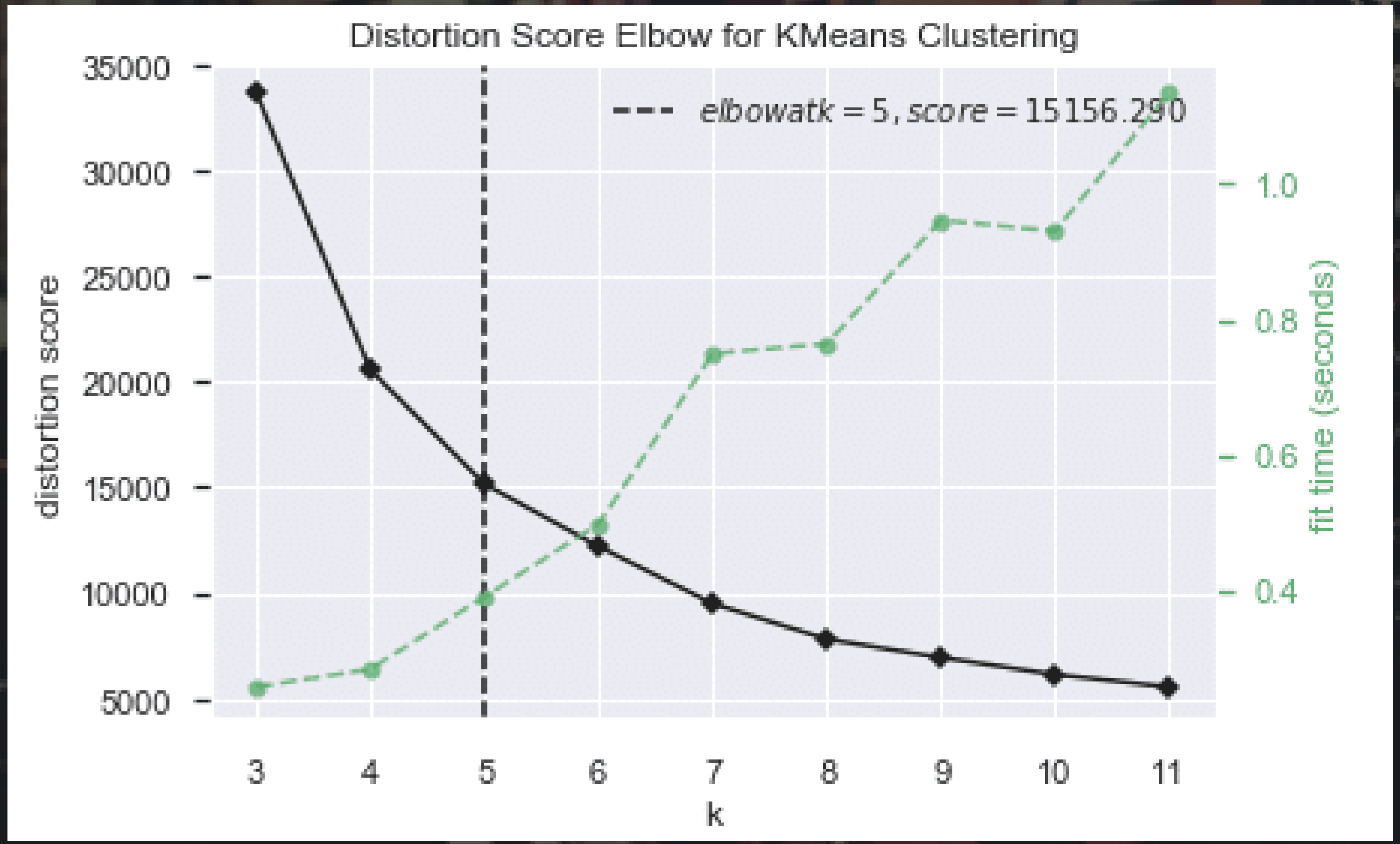


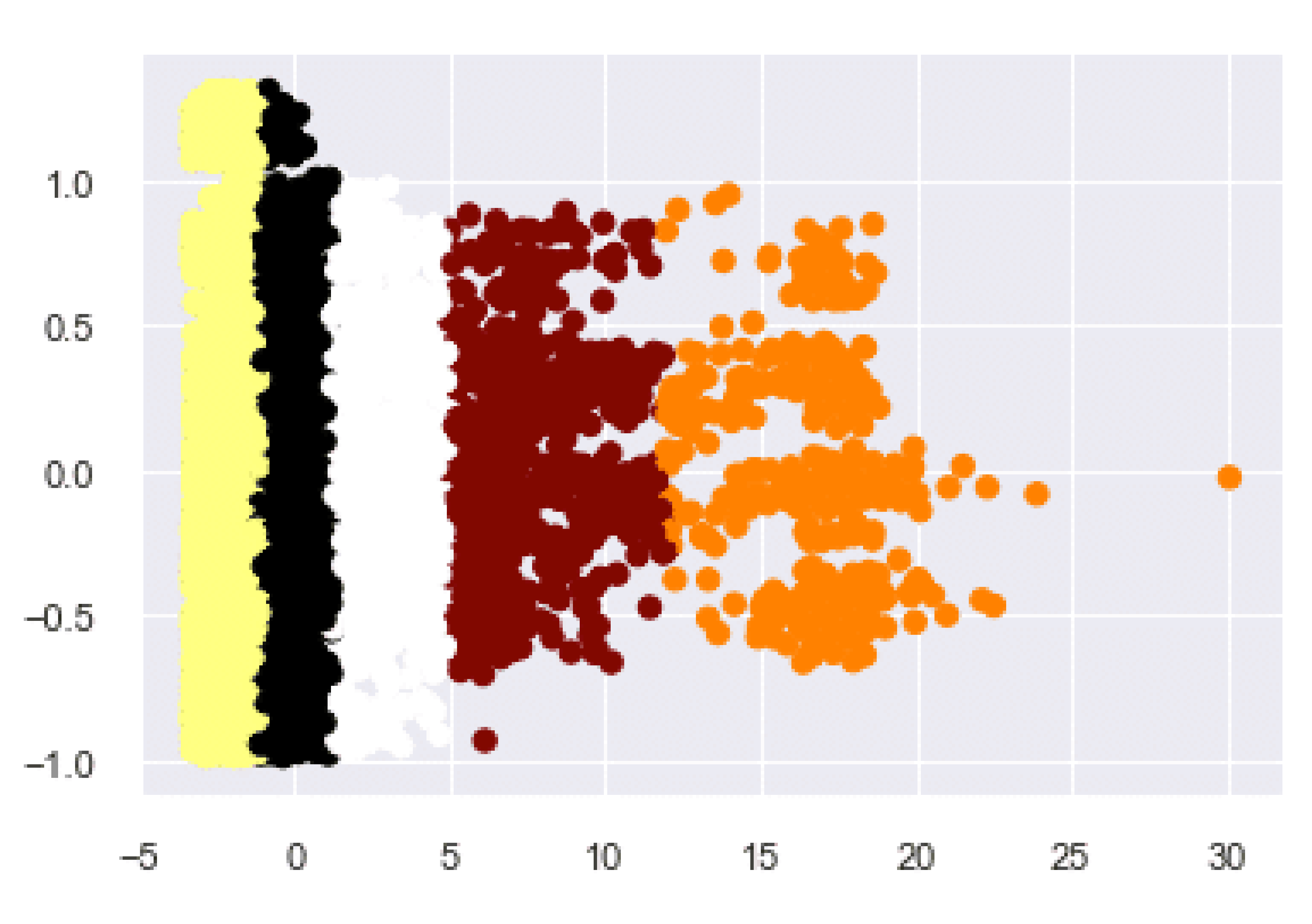
FUTHER CLUSTERING

K-MEANS
DBSCAN
GAUSIAN MIXTURE MODELS
HIERARCHICAL CLUSTERING

K-MEANS

Elbow curve



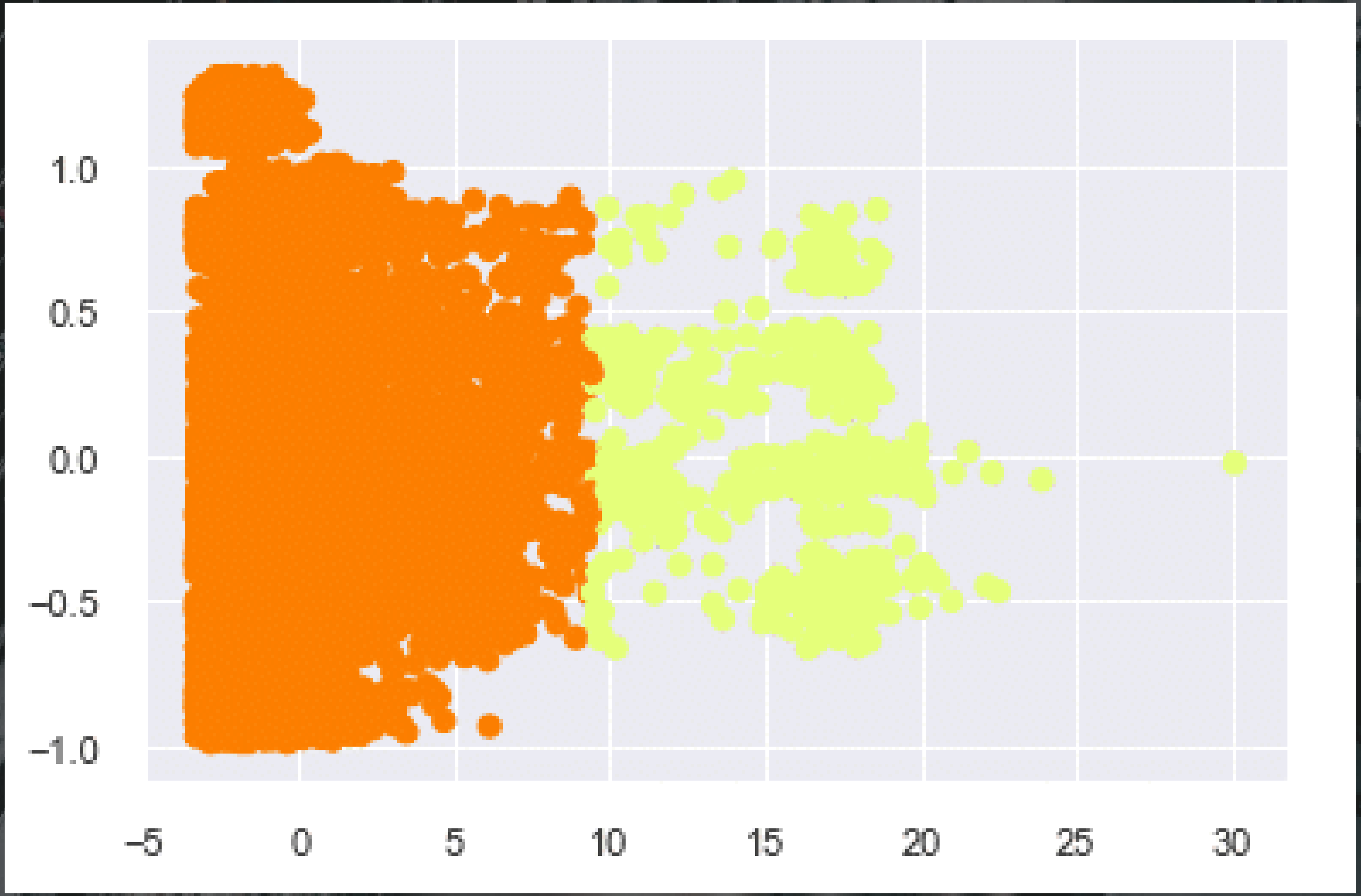


Silhouette score: 0.4438

DBSCAN

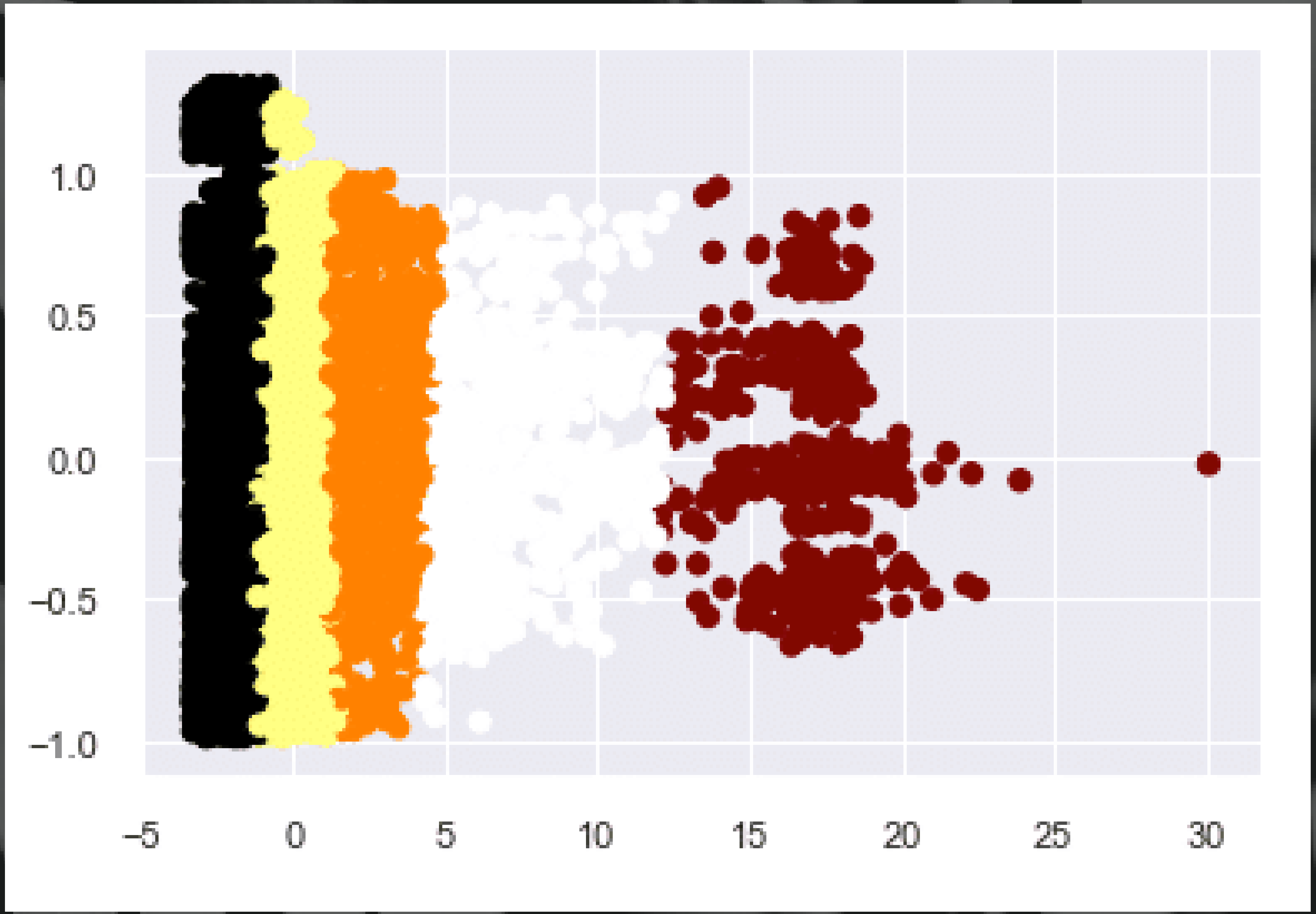
DBSCAN (eps=1.5, min_samples=300)

Silhouette score: 0.8159



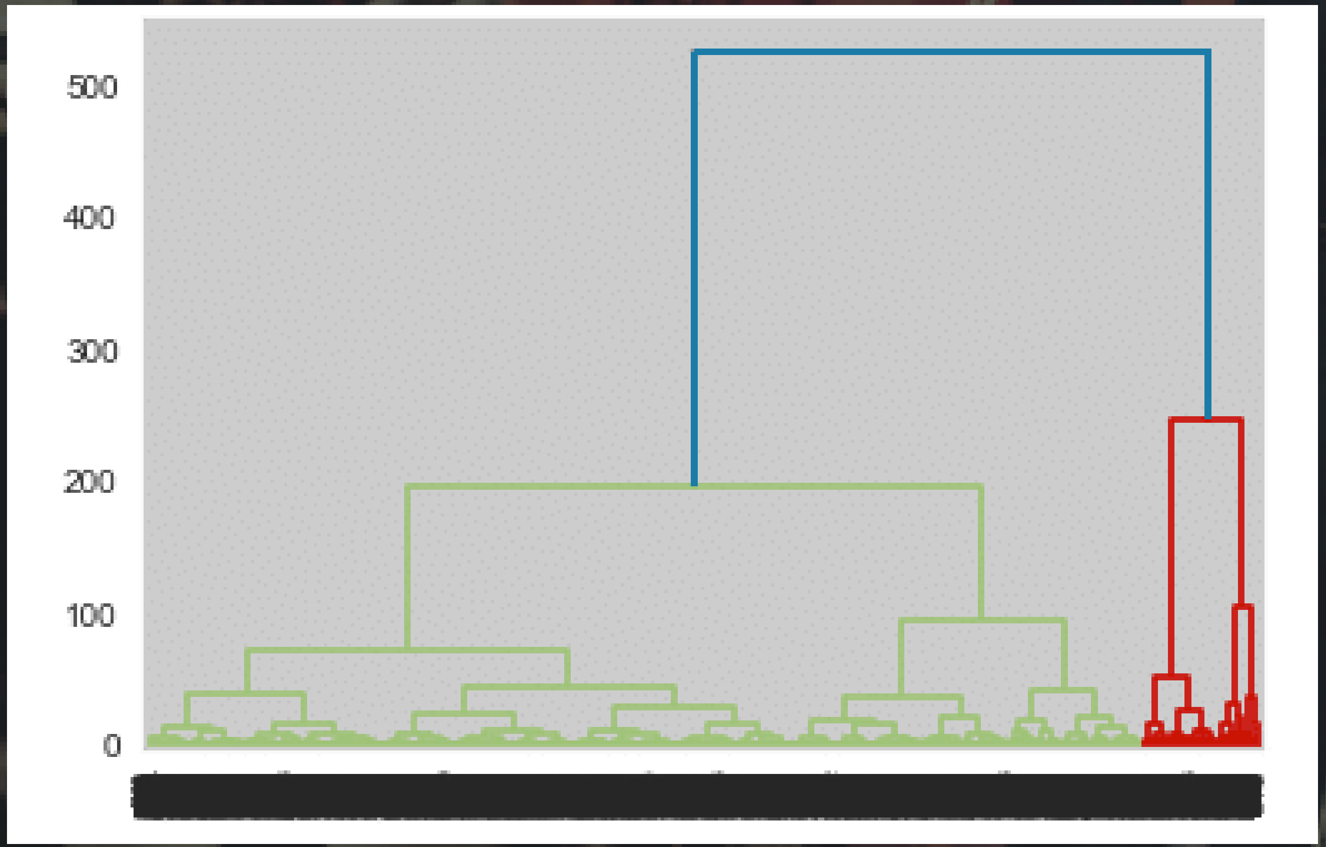
GAUSSIAN MIXTURE MODELS

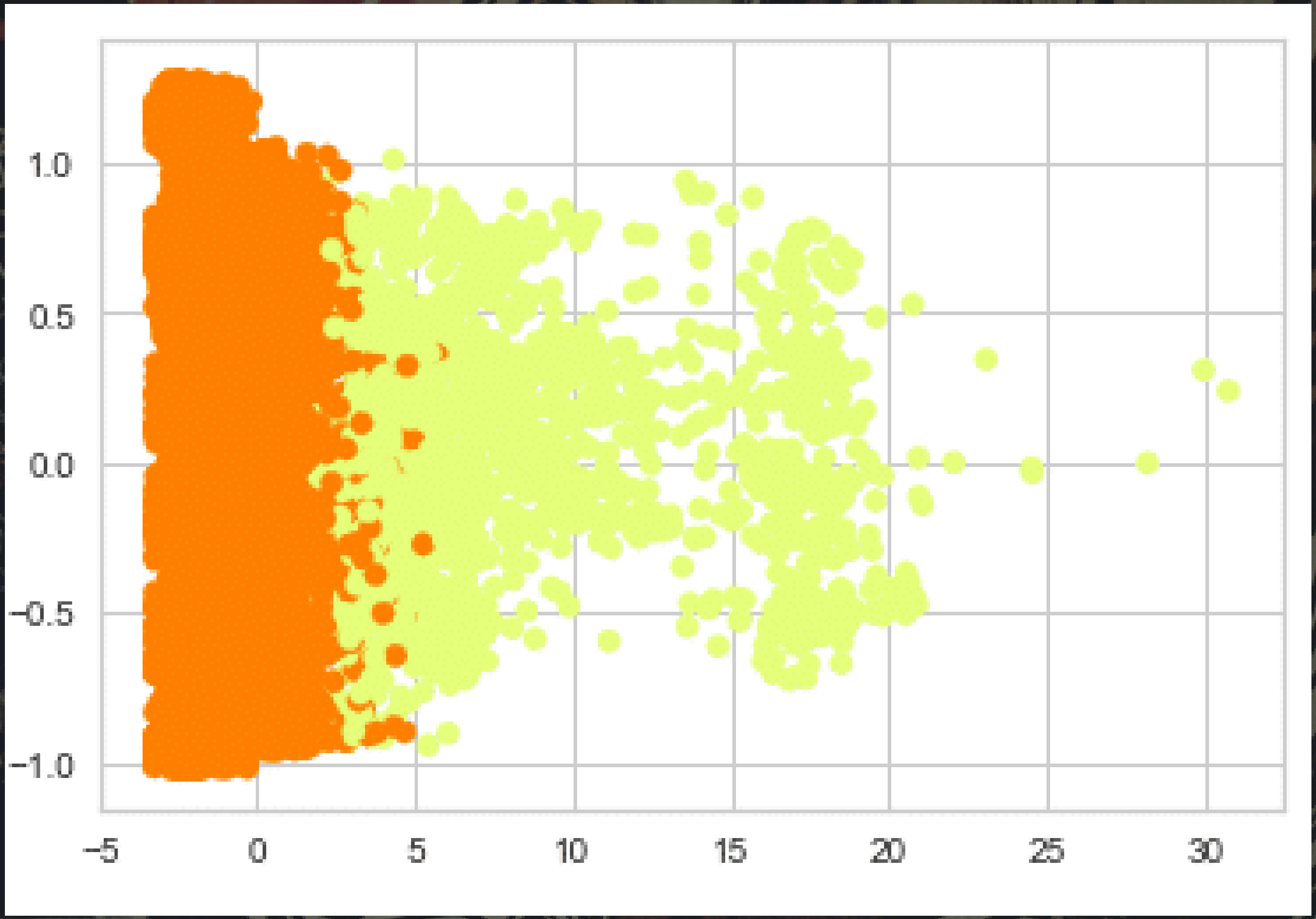
Silhouette score: 0.4407



HIERACHICAL CLUSERING

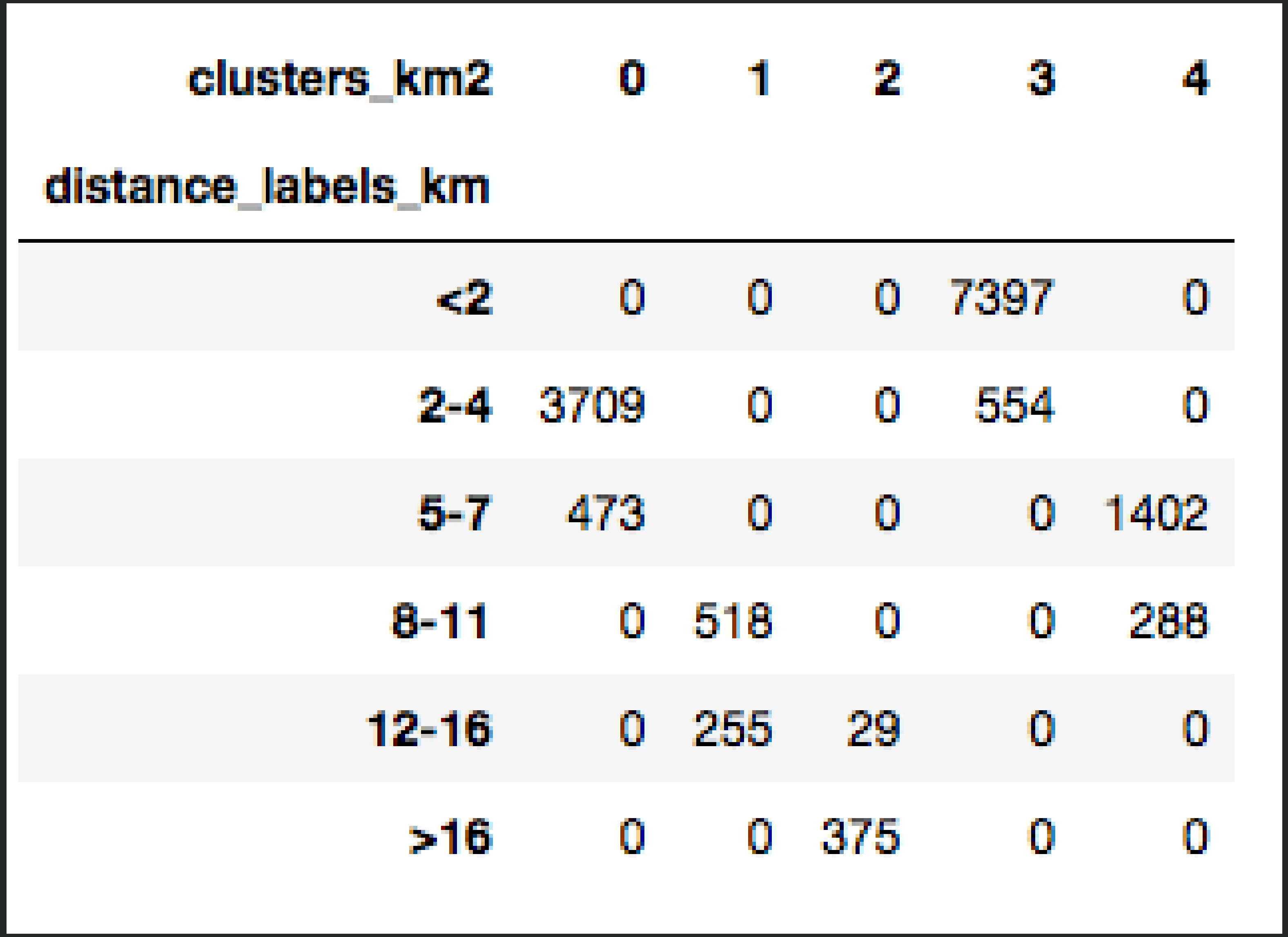
Silhouette score: 0.6551



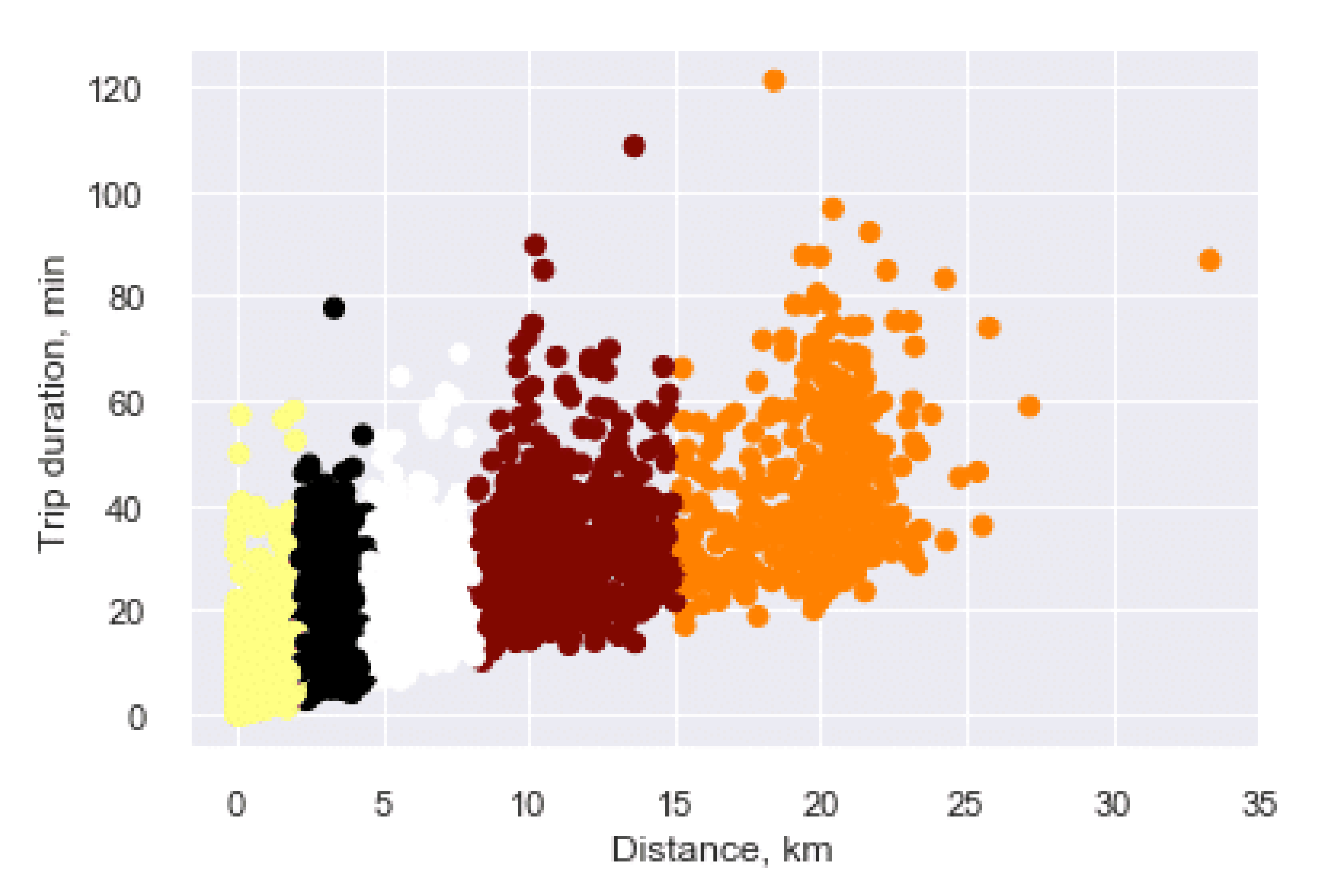


EXPLORING BEST PERFORMING MODELS

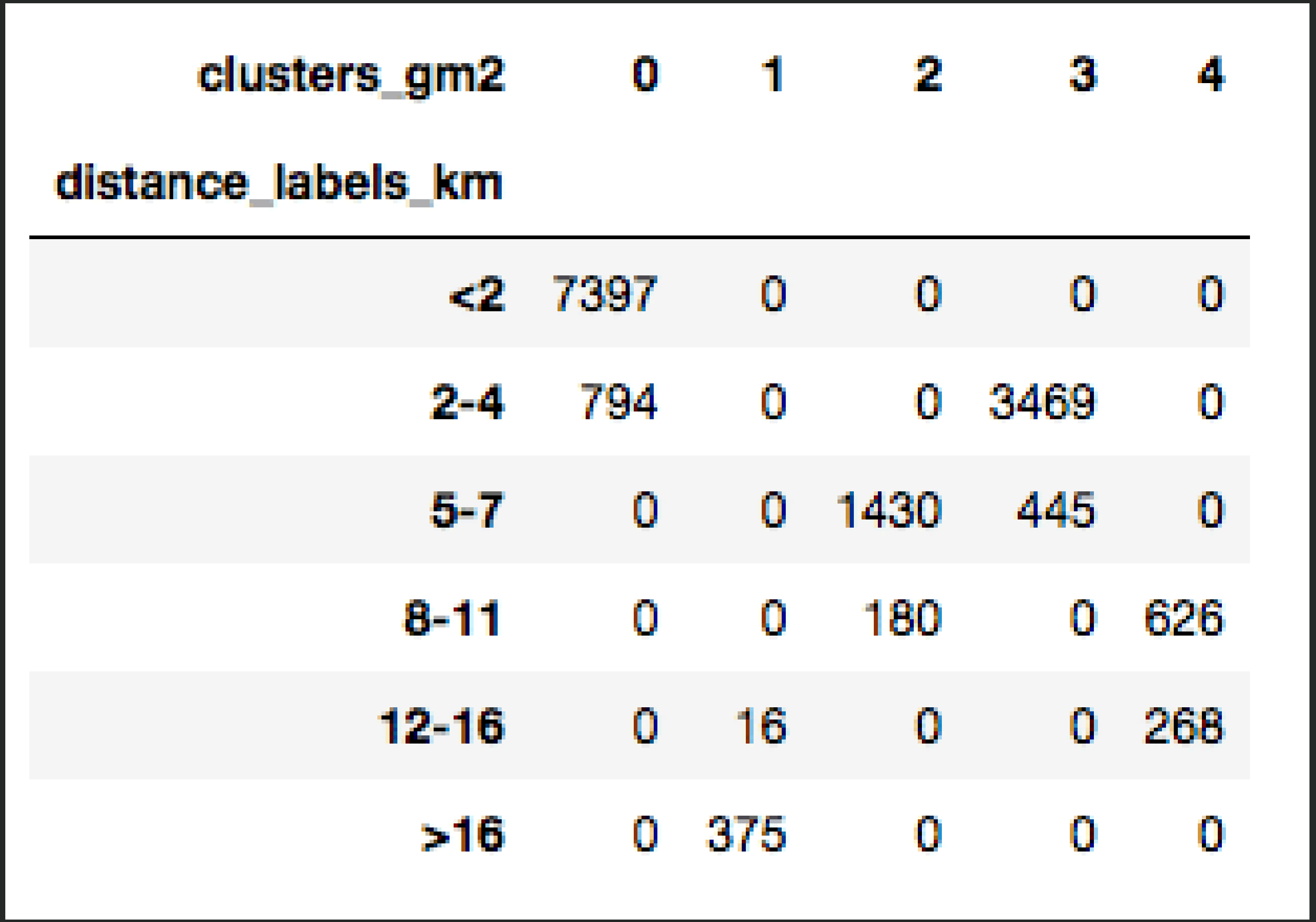
K-means



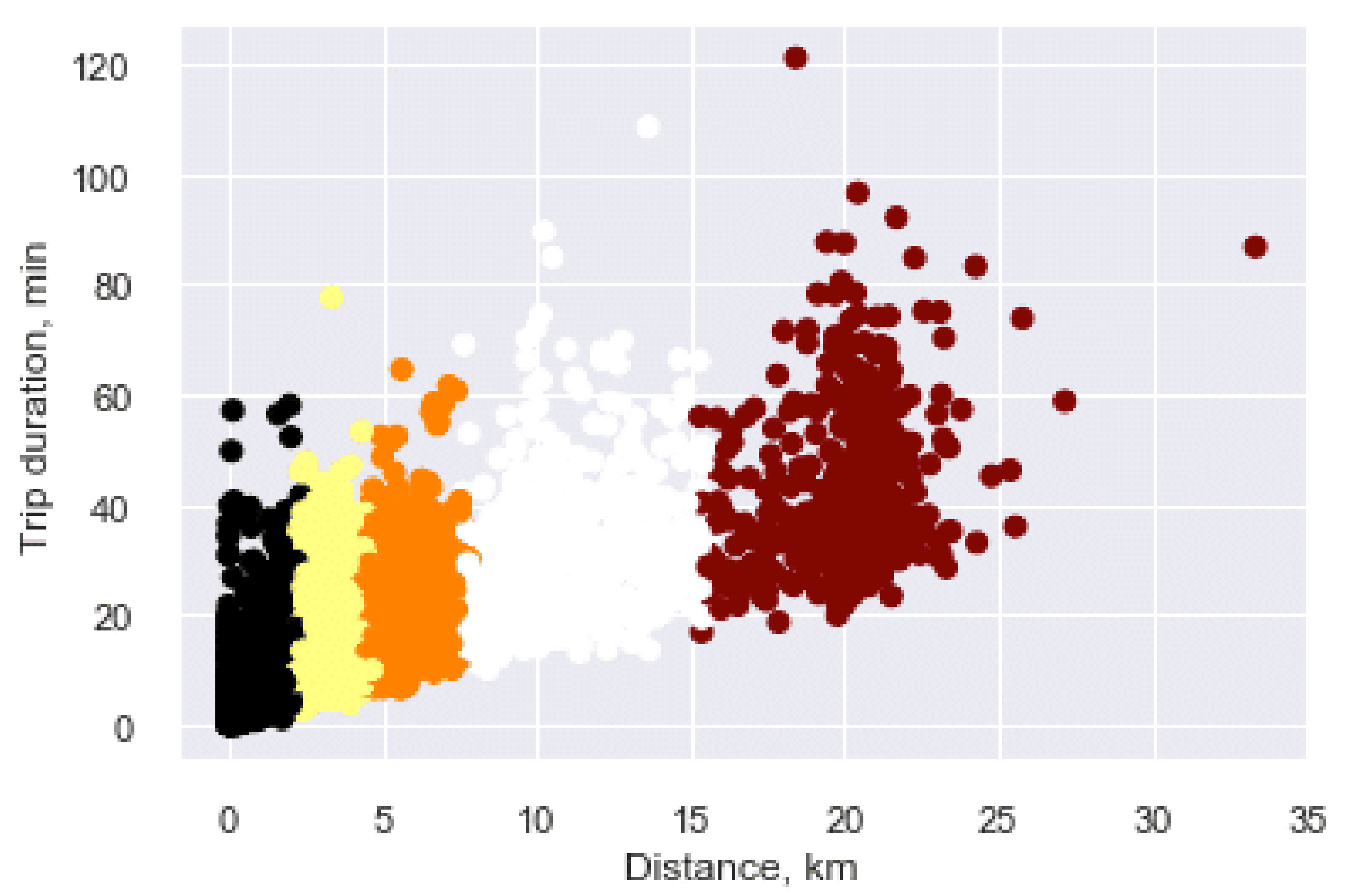
clusters_km2	0	1	2	3	4
distance_labels_km					
<2	0	0	0	7397	0
2-4	3709	0	0	554	0
5-7	473	0	0	0	1402
8-11	0	518	0	0	288
12-16	0	255	29	0	0
>16	0	0	375	0	0



GAUSIAN Mixture Models



clusters_gm2	0	1	2	3	4
distance_labels_km					
<2	7397	0	0	0	0
2-4	794	0	0	3469	0
5-7	0	0	1430	445	0
8-11	0	0	180	0	626
12-16	0	16	0	0	268
>16	0	375	0	0	0



Conclusion

Adding features & creating dummy variables can drive the model in one direction or another

Further improvements: refine the grouping to have clearer definition of clusters

Further improvements: do Supervised ML to predict Trip duration

UNSUPERVISED ML PROJECT

Thank you !

LAUMA USTUPA