

SVILUPPO E DEPLOY DI UNA PIPELINE DI MACHINE LEARNING.
PREVISIONE DELL' IMPORTO DELLA CORSA DI UN TAXI DI NEW YORK



PREDIZIONI TRA I TAXI DI NEW YORK

Presentato da Nicola Lo Surdo

TOPICS

SUMMARY

Assignment

Taxi a New York: esplorando i dati

Dati in Movimento.

I Modelli Predittivi

Il Modello ML Svelato

Largo agli Insights

Le strategie del taxi driver

Conclusioni





ASSIGNMENT

Con i dati di un mese di corse dei **taxi** a New York City

Bisogna sviluppare e simulare il deployment un modello di Machine Learning che preveda l'**importo totale della corsa** al momento del pickup.

Identificate le strategie che un taxista di NYC dovrebbe mettere in atto per:

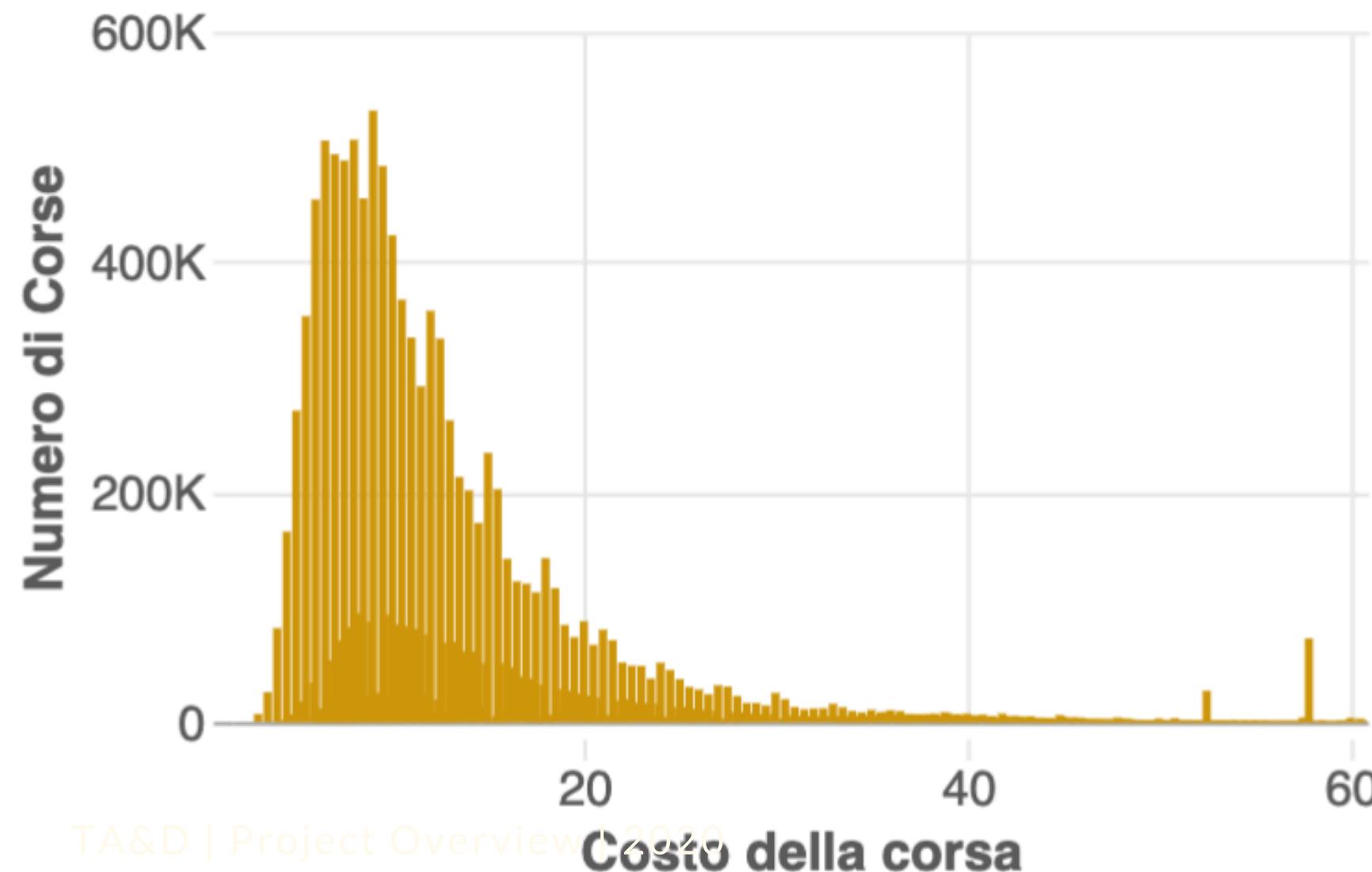
- Massimizzare il `total_amount` di una singola corsa.
- Massimizzare il `total_amount` di fine giornata

IL NOSTRO TARGET

IDATI FORNITI

Corse dei taxi di New York del mese di Aprile
2013. 15 Milioni di corse registrate.

Minimum	2.5
5-th percentile	5.5
Q1	8
median	11
Q3	16.5
95-th percentile	38.33
Maximum	628.1
Range	625.6
Interquartile range (IQR)	8.5



IL COSTO DELLA CORSA

Costo medio 14 \$. Costo in mediana 11 \$.

Il 95% delle corse è al di sotto dei 38 \$.

Nel percentile superiore la deviazione è molto elevata, con costi della corsa sino a 628 \$.

Il costo minimo registrato è di 2.5 \$.

SENZA PASSEGERO A BORDO

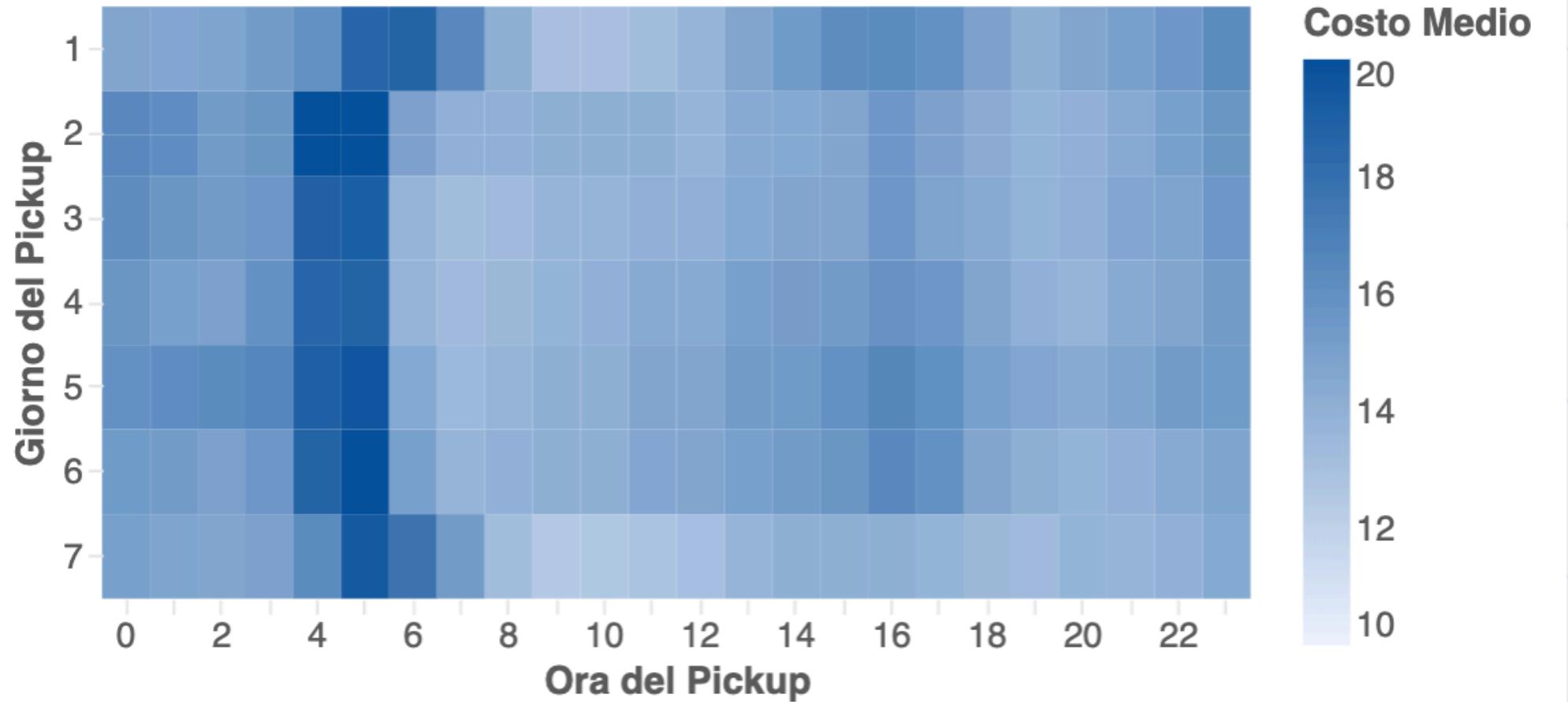
RATE_CODE	Codice tariffario del tipo di corsa: standard, aeroporto, fuori città ...
PICKUP_LONGITUDE	Longitudine del punto di partenza
PICKUP_LATITUDE	Latitudine del punto di partenza
DROPOFF_LONGITUDE	Longitudine del punto di arrivo.
DROPOFF_LATITUDE	Latitudine del punto di arrivo
SURCHARGE	Supplemento aggiunto al costo della corsa: causa traffico, orario ...
MTA_TAX	Tassa fissa applicata alle corse taxi a New York
VENDOR_ID	Identificativo del fornitore del servizio taxi

AL PICK-UP

TRIP_DISTANCE	Distanza percorsa durante la corsa
PASSENGER_COUNT	Numero di passeggeri trasportati
PAYMENT_TYPE	Modalità di pagamento: in contanti, con carta di credito/debito ...

CORSA FINITA

AMOUNTS	Costo netto della corsa, mance, costo pedaggi
TRIP_TIME_IN_SECS	Durata della corsa in secondi
STORE_AND_FWD_FLAG	Se i dati della corsa sono stati temporaneamente salvati e poi inviati



COSTO MEDIO PER GIORNO/ORA

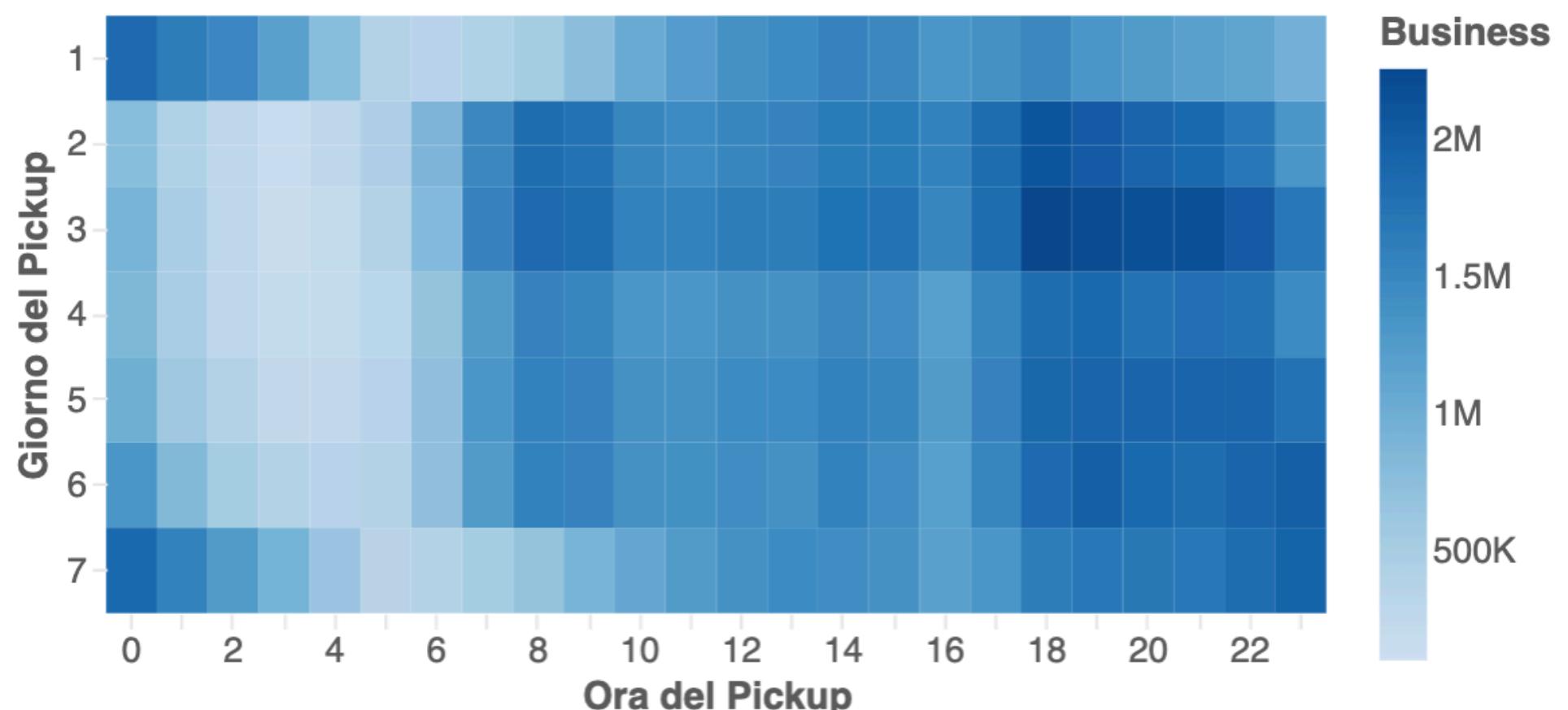
Durante la fascia oraria tardo-notturna, primo mattino, il costo medio aumenta di quasi il 200%

Le altre fascie notturne e la fascia pomeridiana 16-18 presentano valori sopra la media

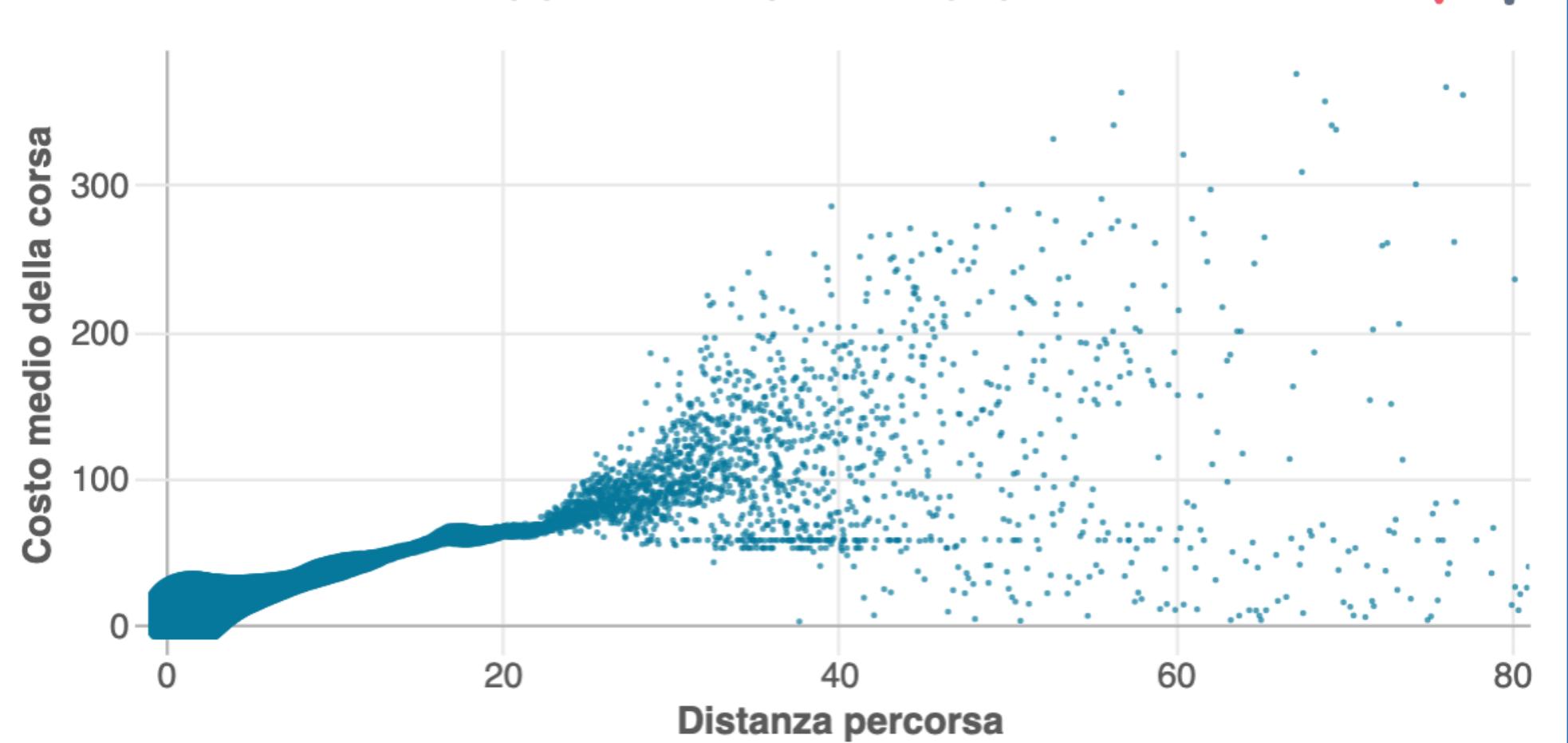
VOLUME TOTALE DEL BUSINESS

La domanda di taxi si presenta con volumi importanti dalle 18 alle 24 e dalle 8 alle 10 del mattino

Il peak è martedì dalle 18 alle 19 con un volume pari a 2.26 Mln



Relazione tra costo medio (\$) e distanza percorsa (mi)



COSTO MEDIO E DISTANZA

Il costo medio cresce linearmente con la distanza percorsa sino alle 18 miglia
Dopo le 18 miglia la correlazione si indebolisce,
Un sub-set di corse mostra valori anomali ed asimmetrici: distanze elevate e costi medi bassi

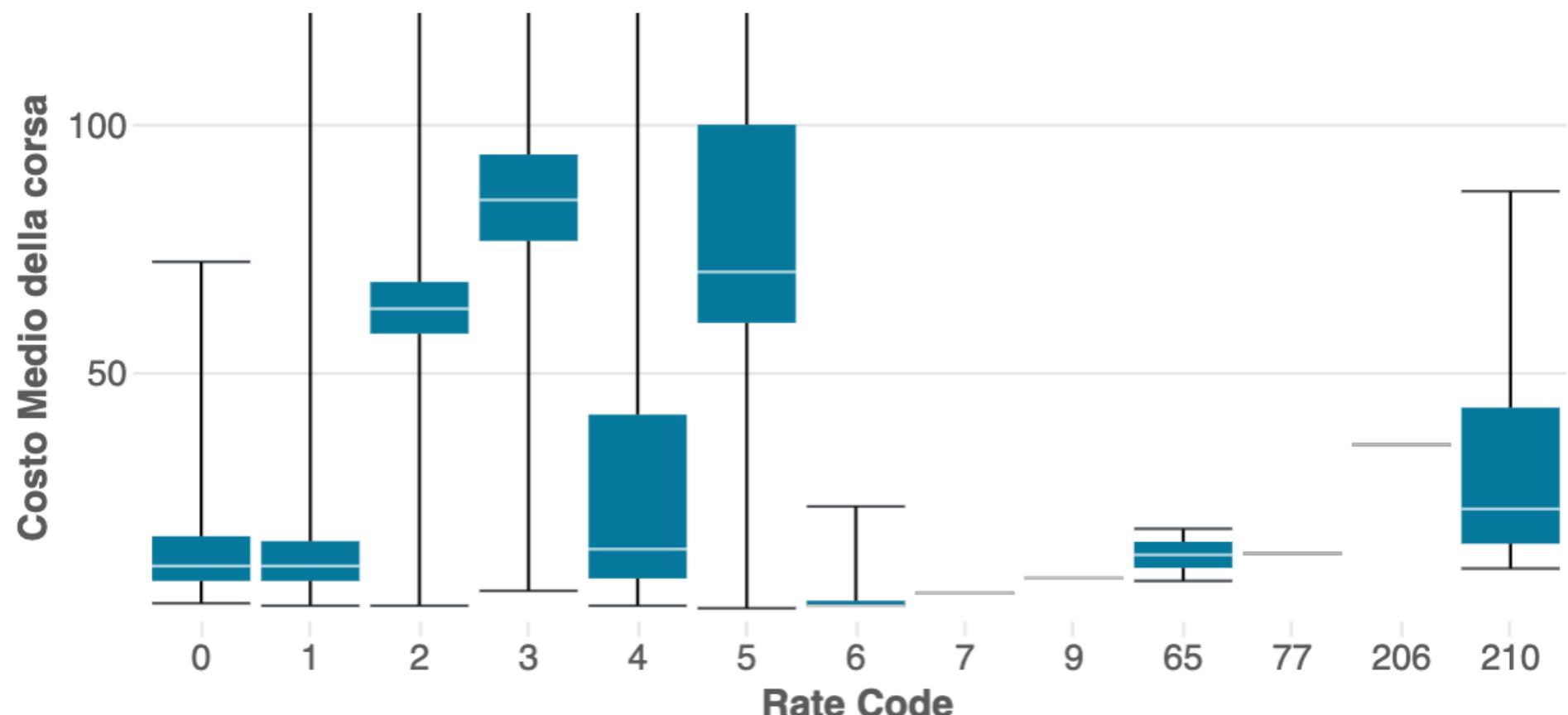
CODICE TARIFFA

La tipologia di tariffa applicata incide in modo determinante sul costo della corsa.

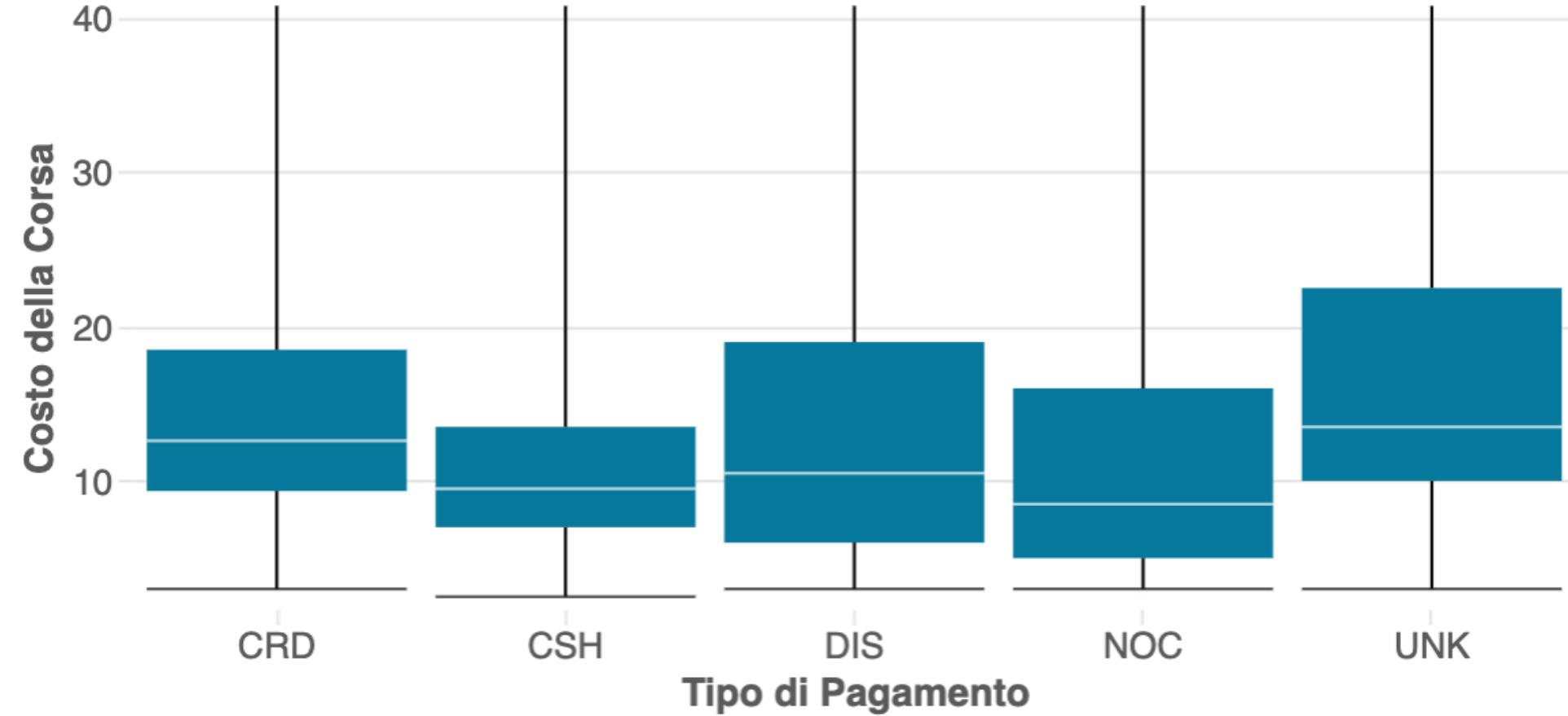
La capacità predittiva della variabile è indebolita dalla distribuzione di frequenza.

Il 98% delle corse è in rate code 1

Distribuzione del Costo medio (\$) per Codice Tariffa



Distribuzione del costo della corsa (\$) per tipologia di pagamento



TIPOLOGIE DI PAGAMENTO

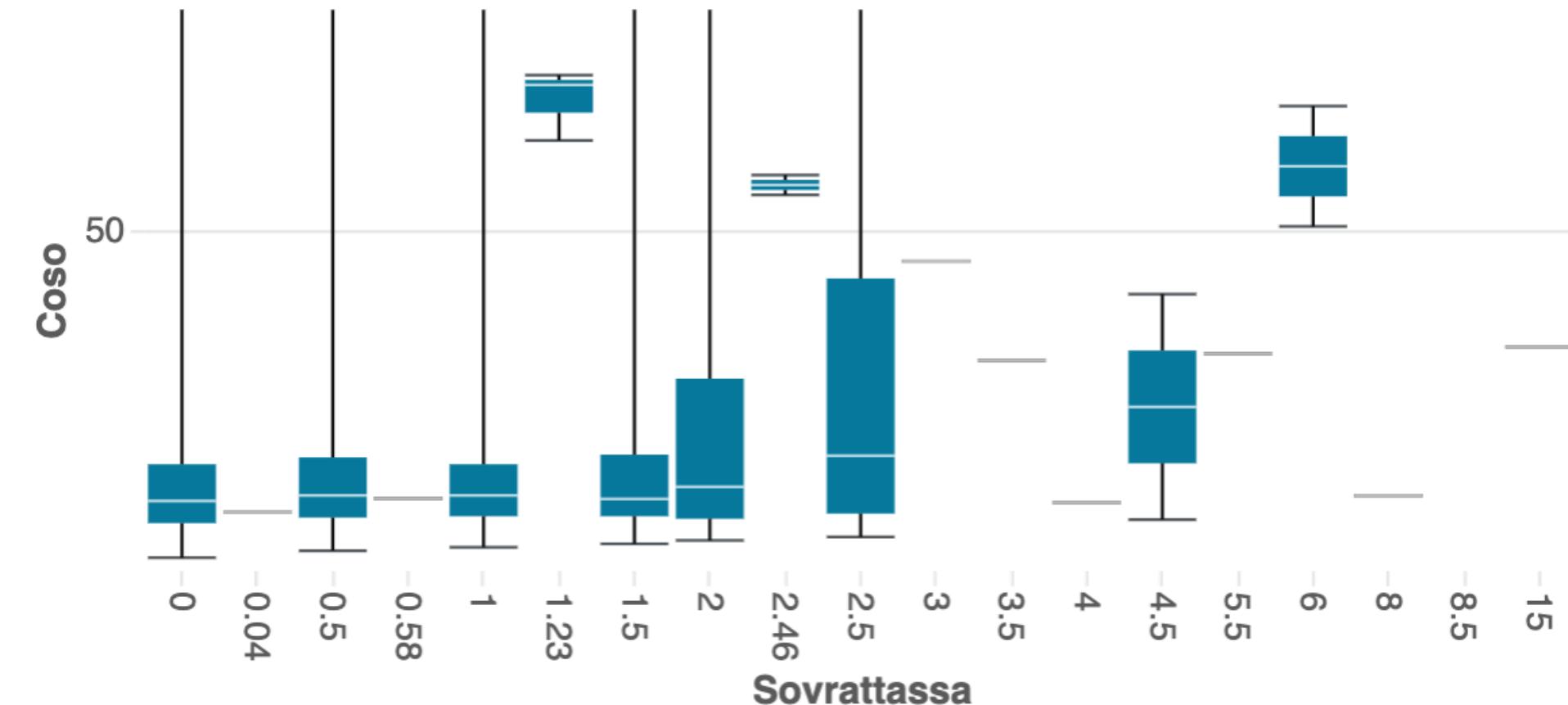
Il 53.7% delle corse è pagato con Carta di Credito/Debito, con un costo mediano di 12.6 \$. La seconda modalità di pagamento più utilizzata è il contante (46%). Le corse pagate in contanti risultano leggermente più economiche, con il costo più frequente pari a 9.5 \$.

SOVRATTASSA

Il costo della corsa varia in modo significativo con la sovrattassa applicata.

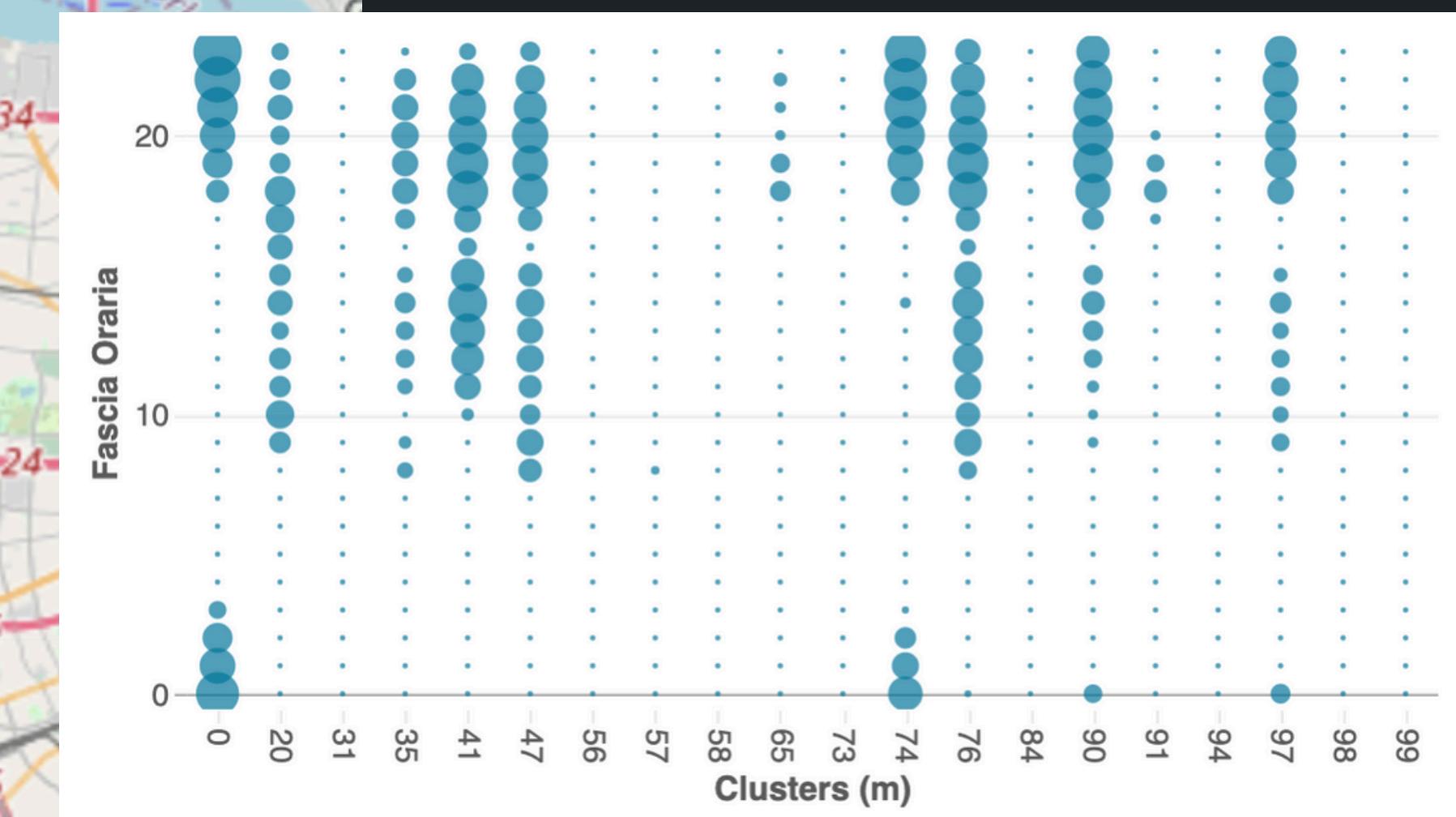
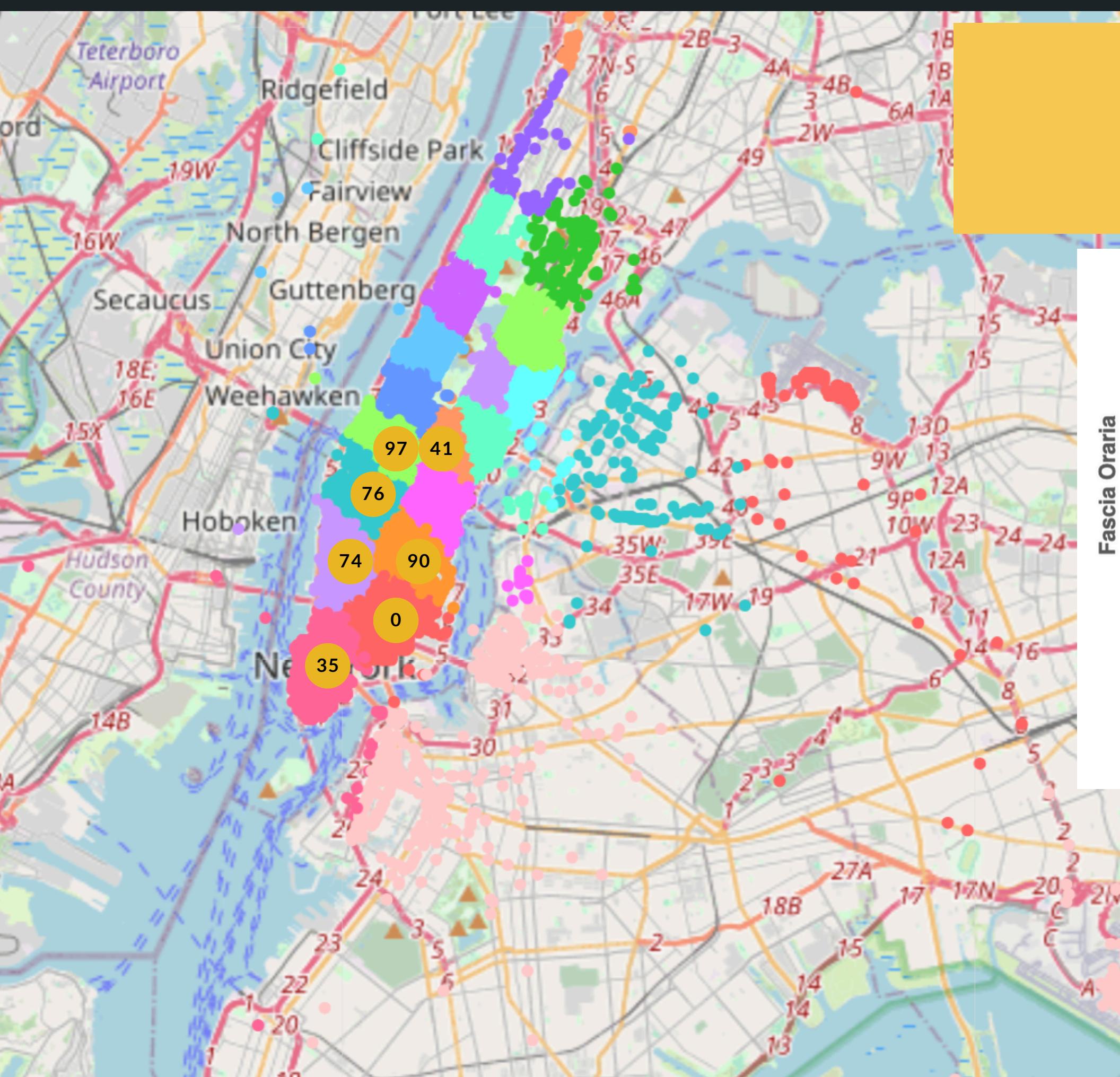
La distribuzione della variabile però risulta fortemente sbilanciata. Al 99% delle corse viene applicata una surcharge non superiore ad 1 \$.

Distribuzione del costo della corsa (\$) per sovrattassa



NEW YORK IN CLUSTERS

VOLUML DEL BUSINESS

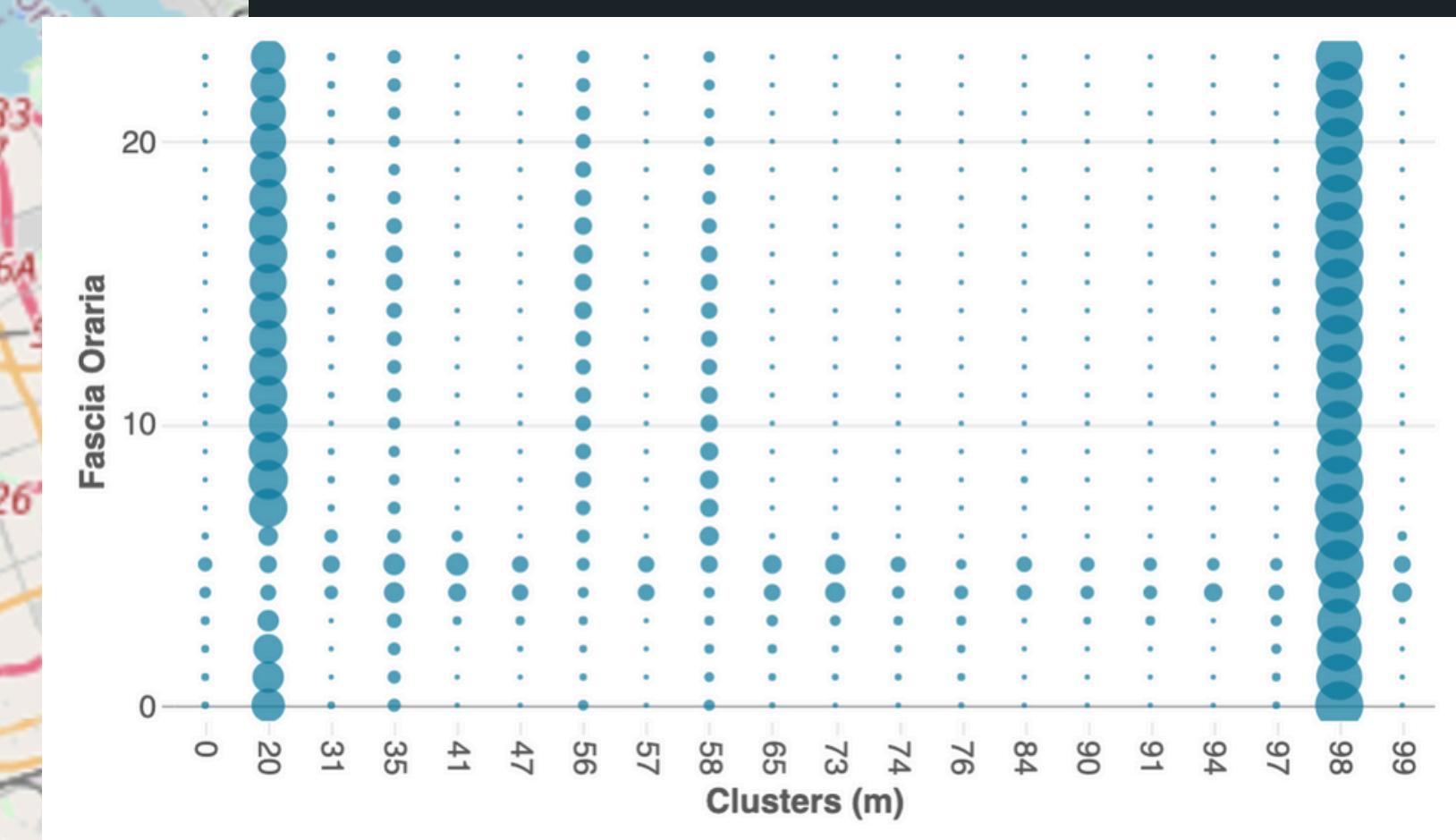
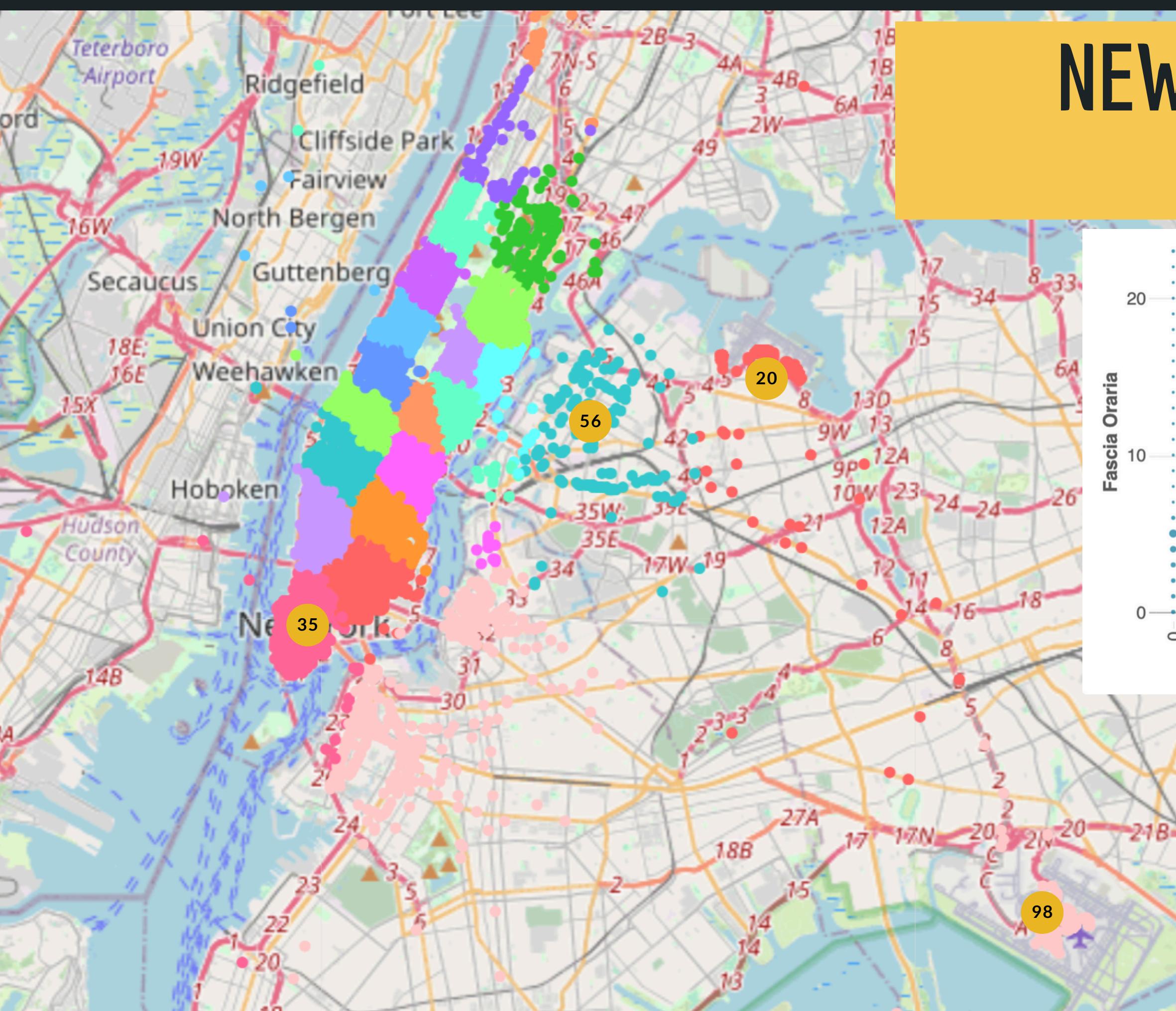


Cluster 0-35-41-74-76-90-97

Il volume del business si concentra in Central/Lower Manhattan con fasce orarie che superano i 1 mln \$ per ora.

NEW YORK IN CLUSTERS

COSTO MEDIO DELLA CORSA



Cluster 20-35-56-58-98

Gli aeroporti JFK/La Guardia dominano la scena delle tariffe, rispettivamente 50 \$ / 40 \$ il costo della corsa.

I Cluster 35/56/58 si attestano a circa 20 \$ a corsa, ma con distanze molto più brevi di quelle aeroportuali.

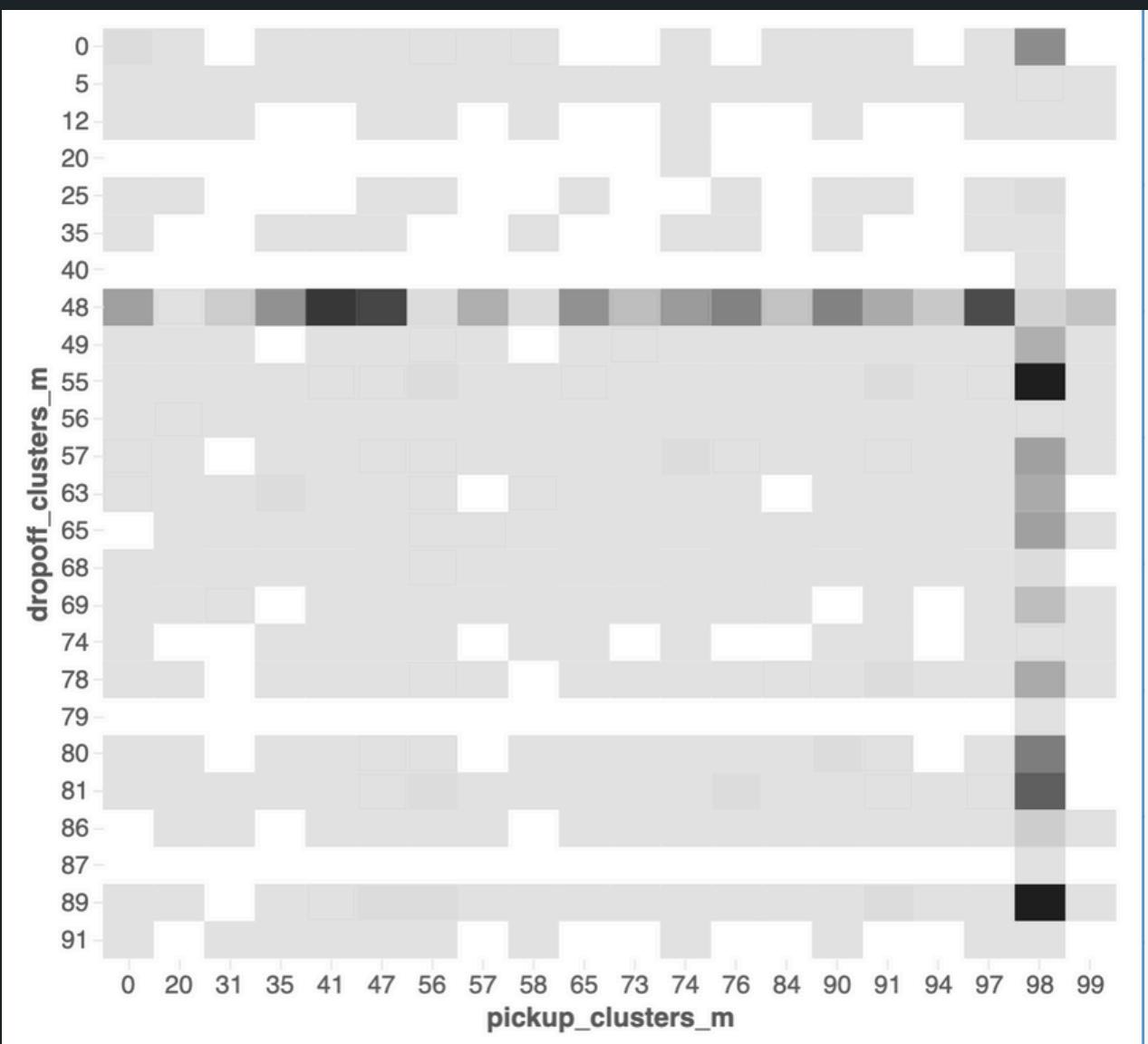
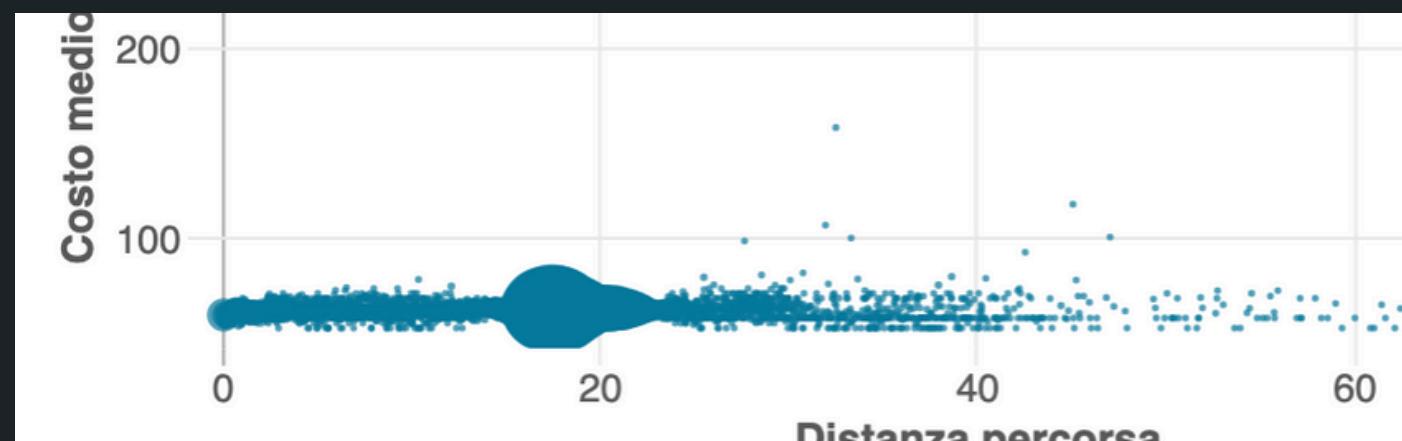
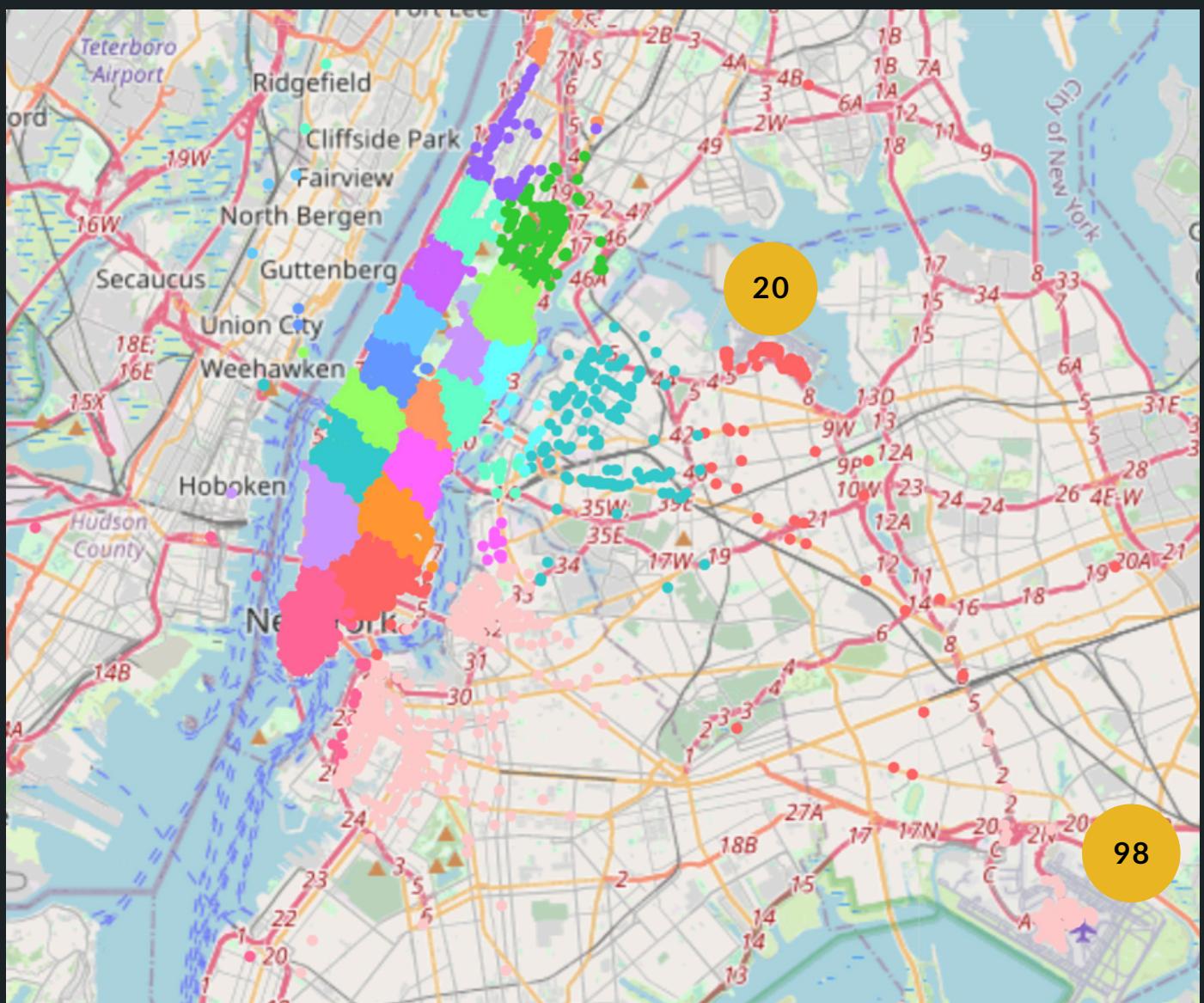
PATTERN DI DATI

Il rate code, la distanza percorsa e le connessioni tra le aree di pickup e drop off evidenziano chiari pattern nei comportamenti di viaggio dei taxi.

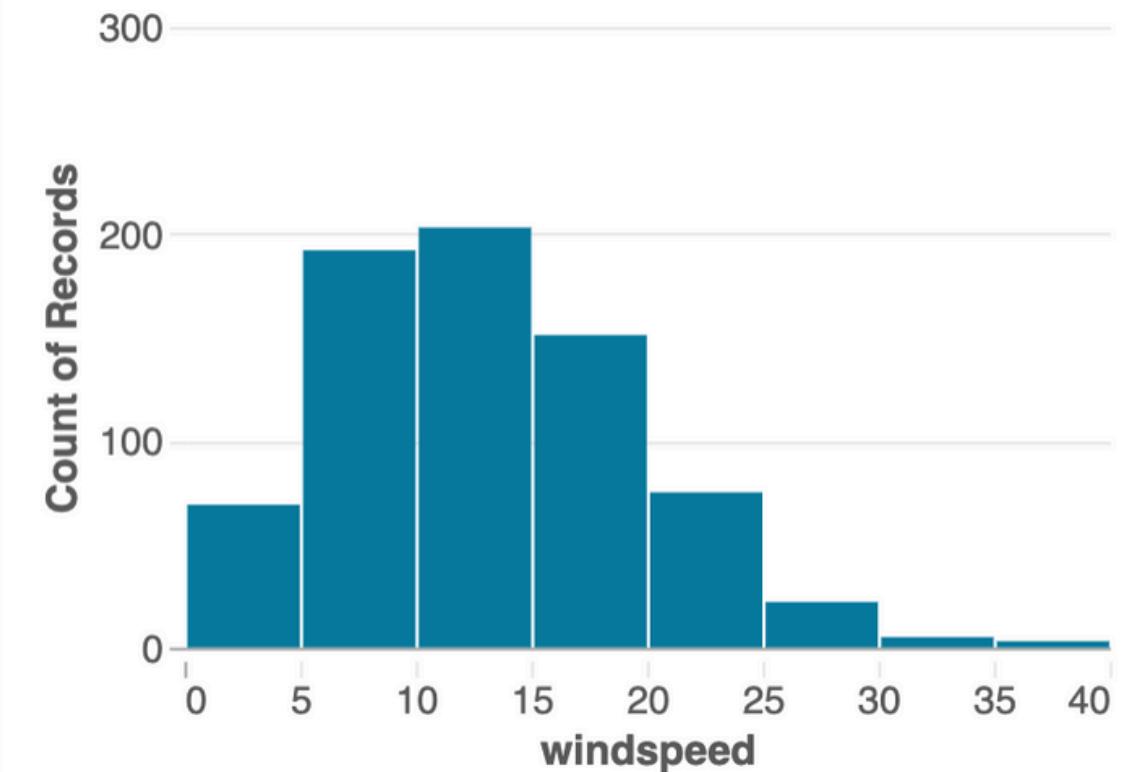
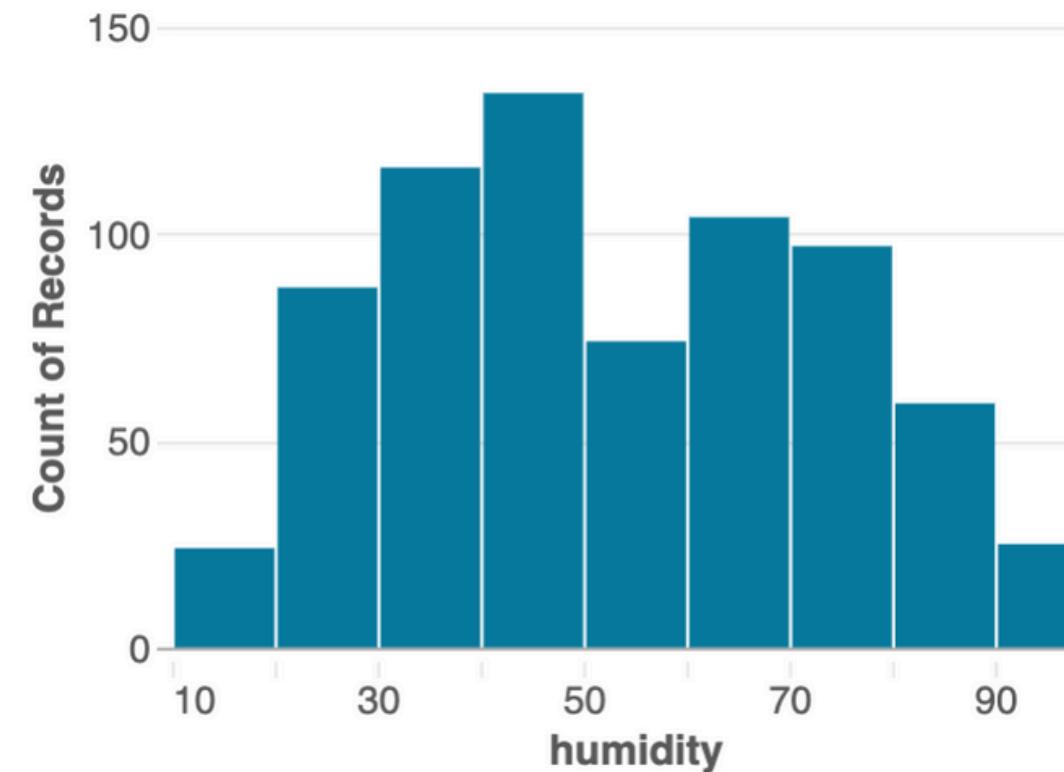
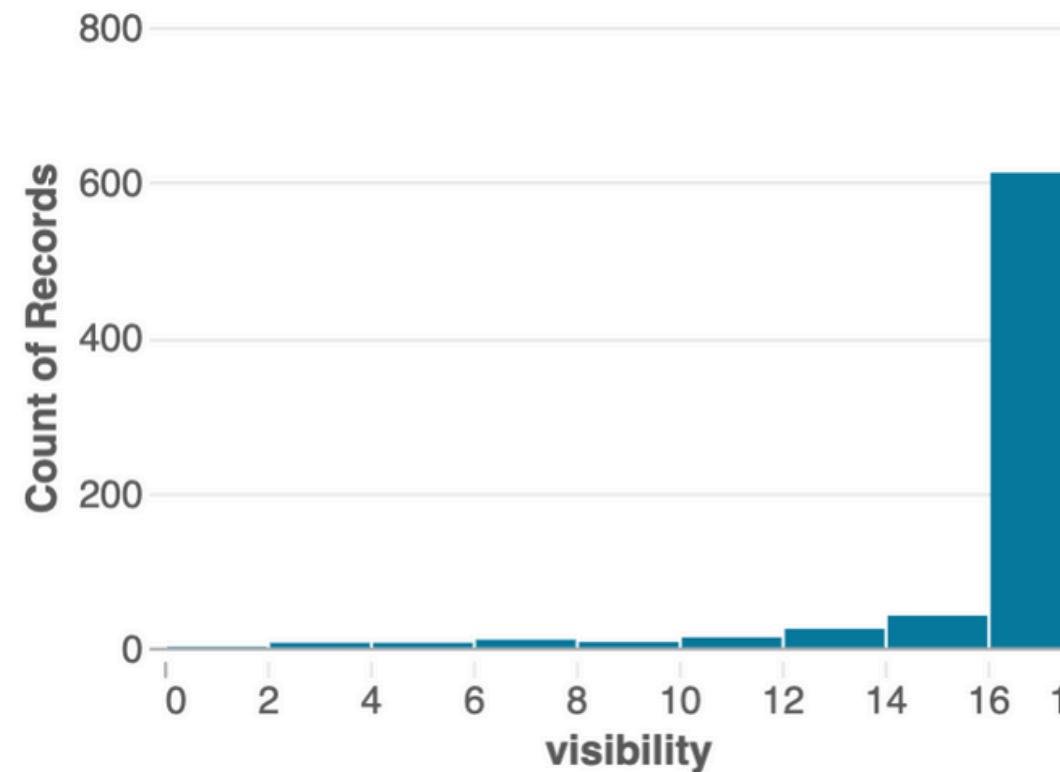
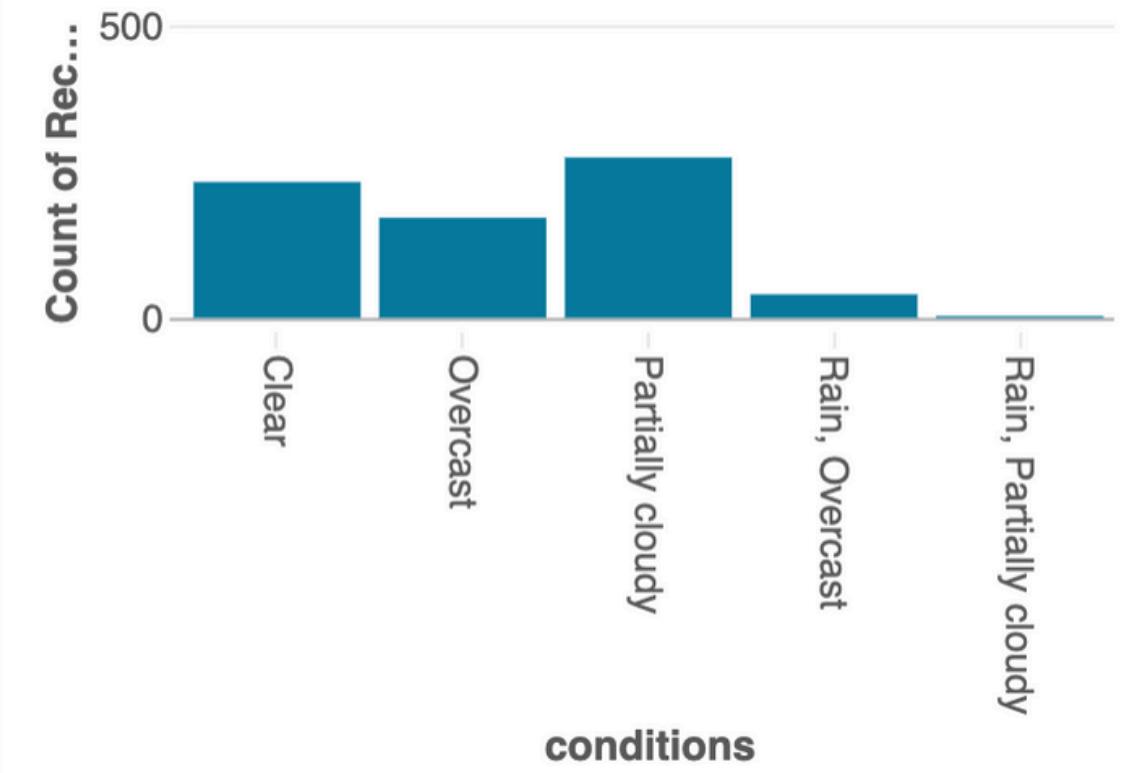
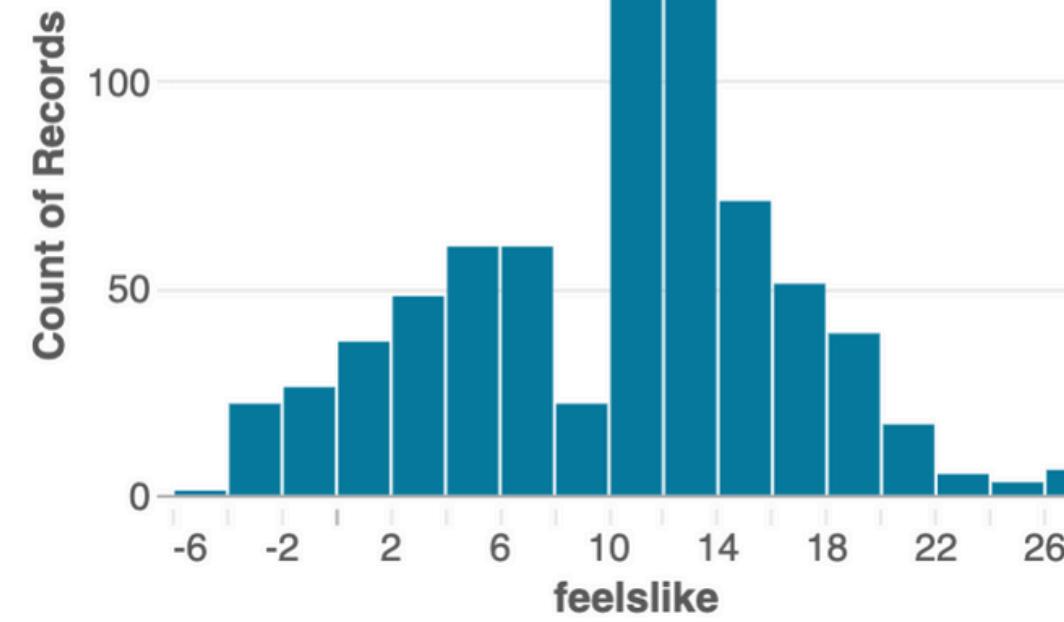
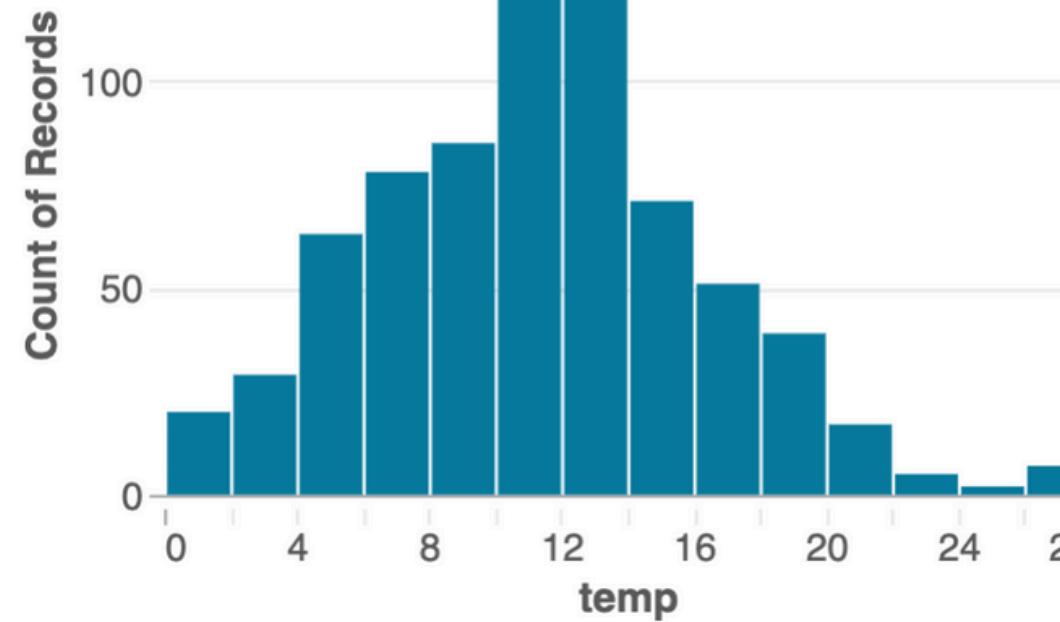
TRAGITTO NEW YORK CITY - AEROPORTO JFK

Un esempio: il tragitto città - aeroporto JFK, Rate code 2, cluster di pickup 98, cluster di drop off 48 e tariffa fissa media di circa 50 \$.

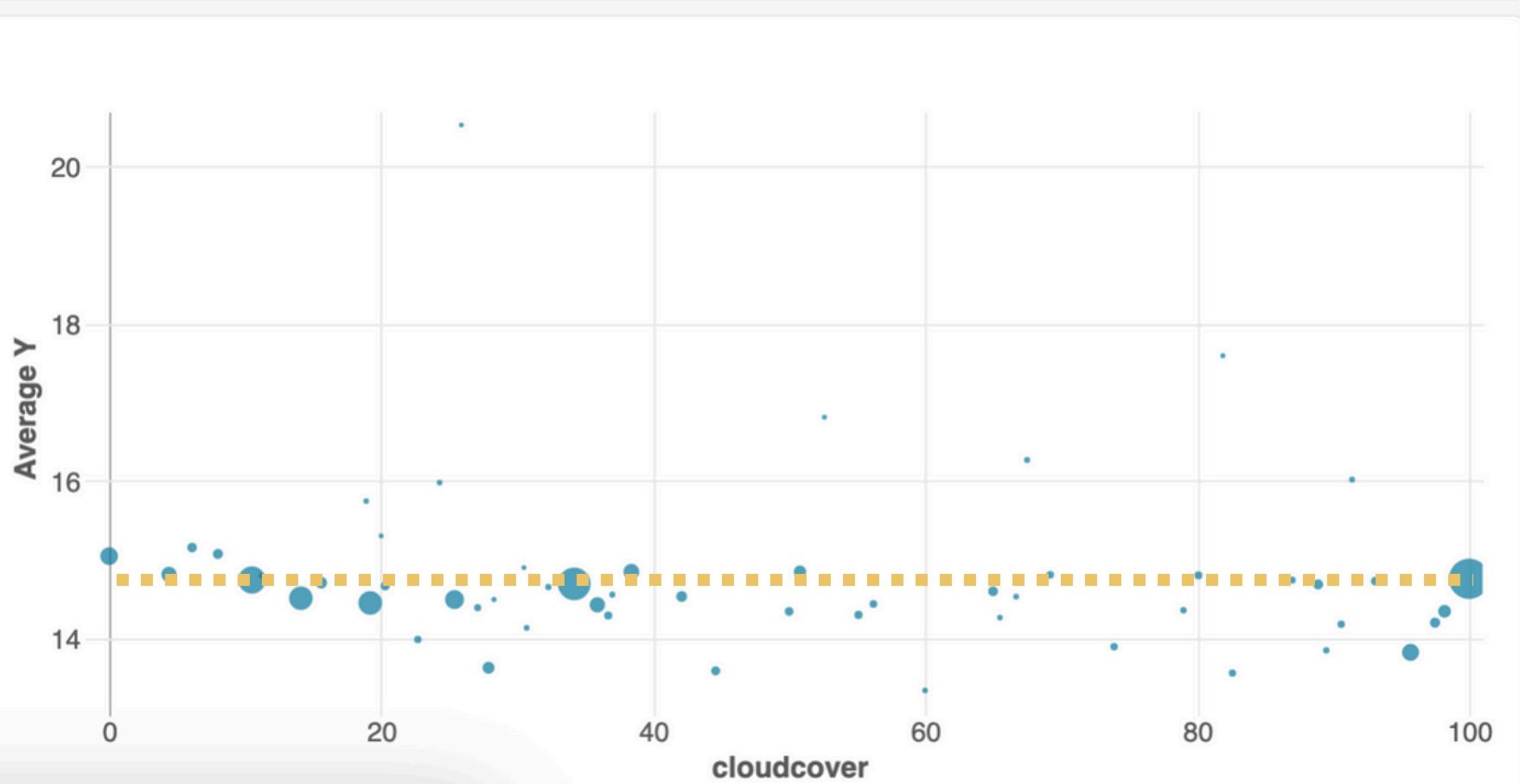
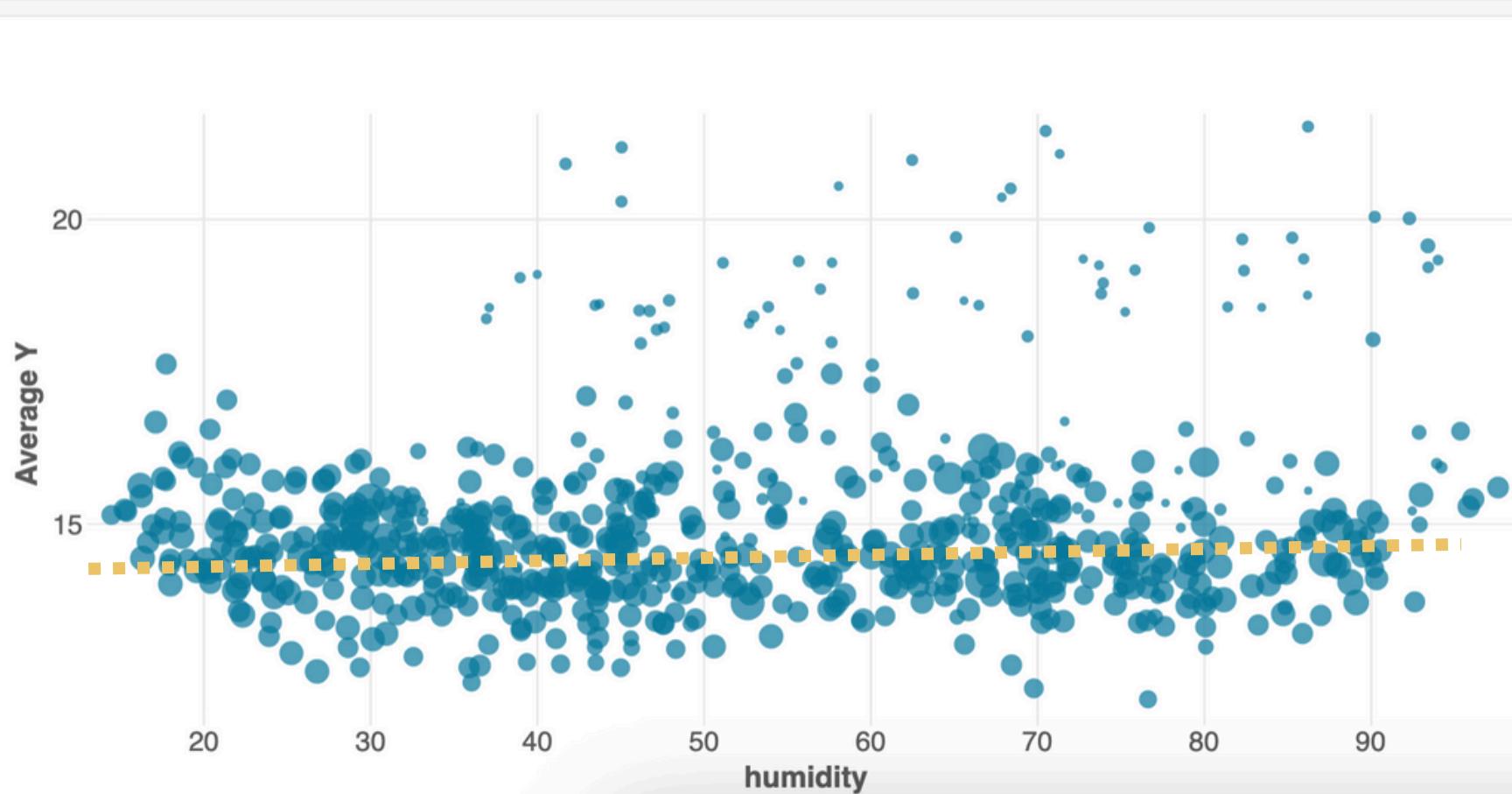
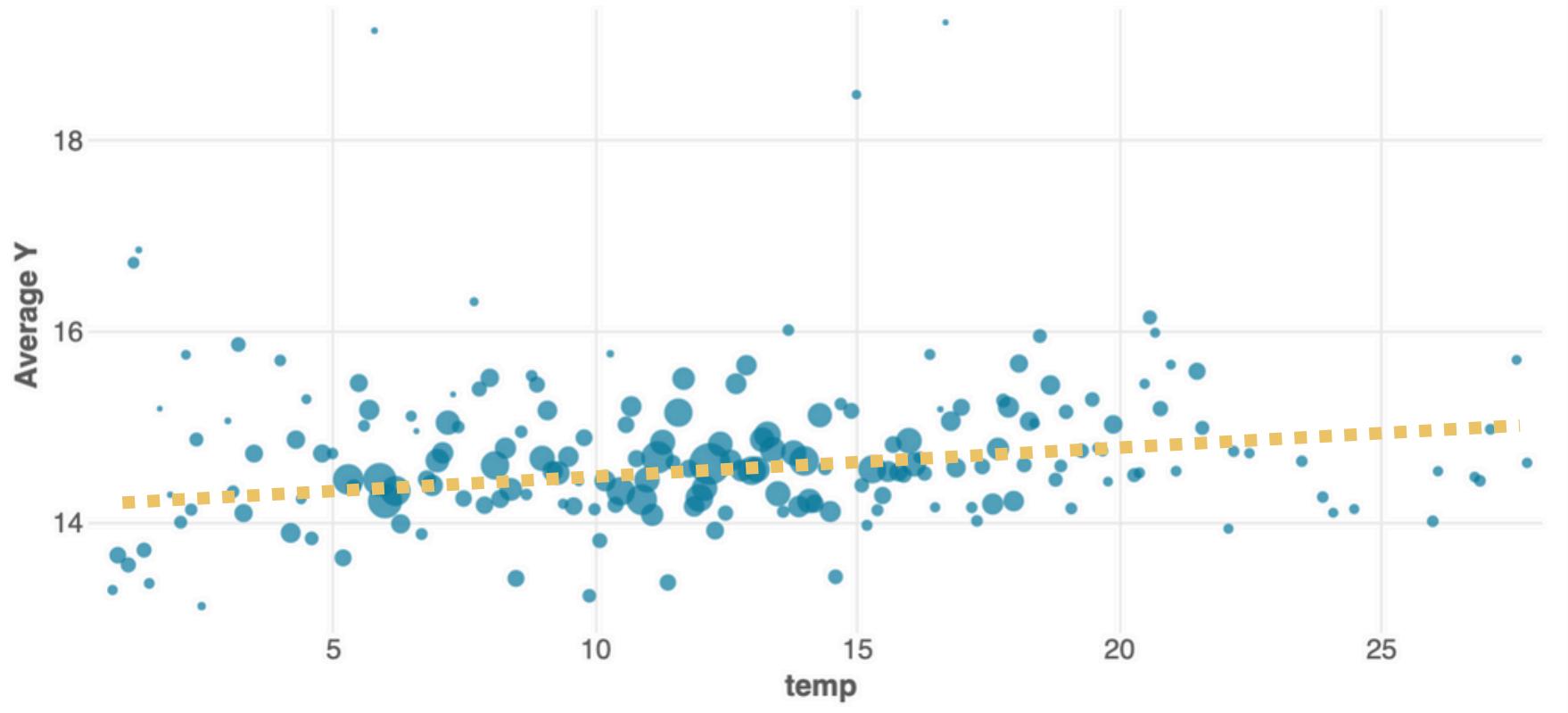
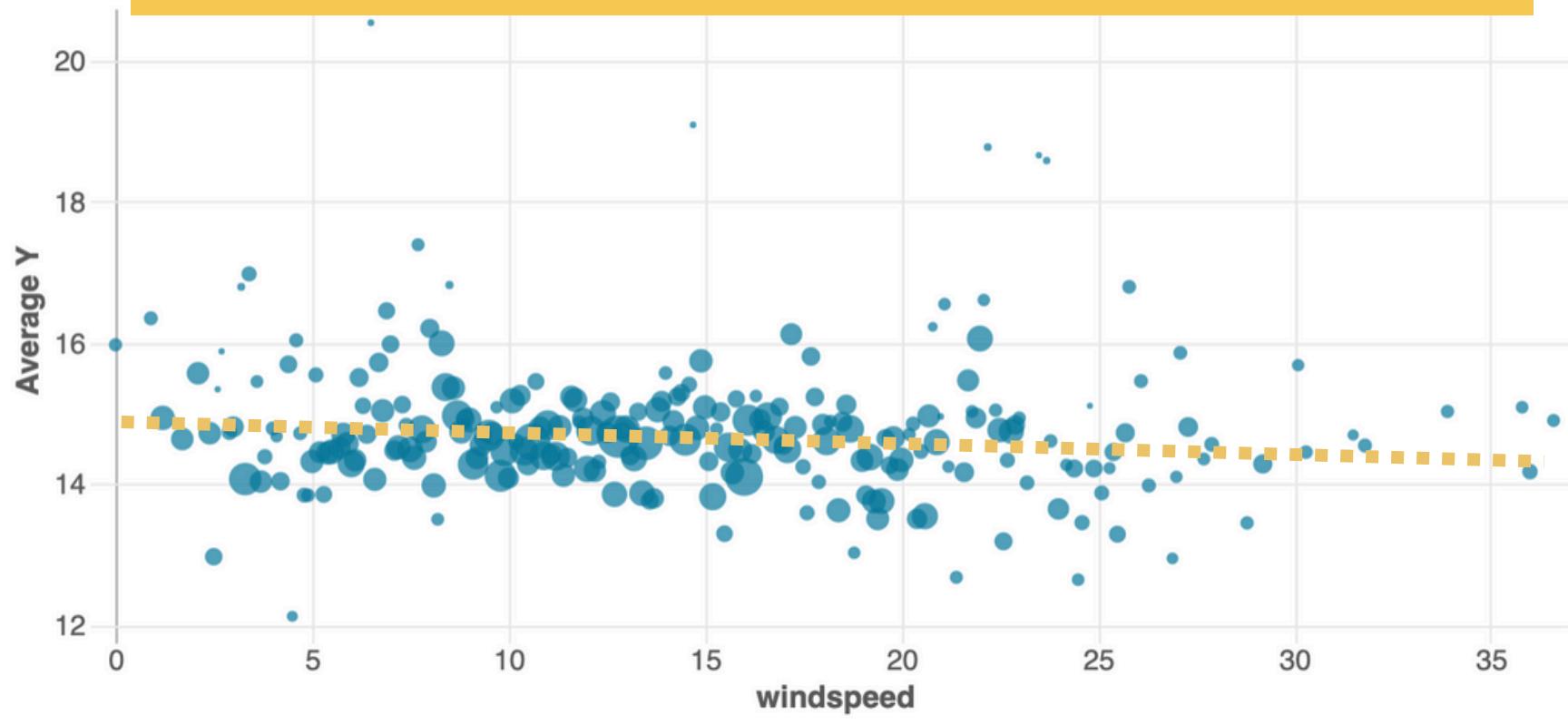
Il rate code 2, a tariffa fissa, spezza in modo completo la correlazione esistente tra la distanza percorsa ed il costo della corsa.



METEO A NEW YORK



CORRELAZIONI METEO



MACHINE LEARNING MODELS

REDUCED MODEL

Uso di variabili pre-boarding
del passeggero

Obiettivo: guidare la strategia
del taxi driver nell'allocazione
degli spostamenti

BASE MODEL

Uso di variabili pre e post
boarding del passeggero

Obiettivo: Predizioni real-time
per ottenere il costo della corsa
al pick-up

ENRICHED MODEL

Uso di variabili base model +
variabili climatiche

Obiettivo: migliorare le
performance del base model

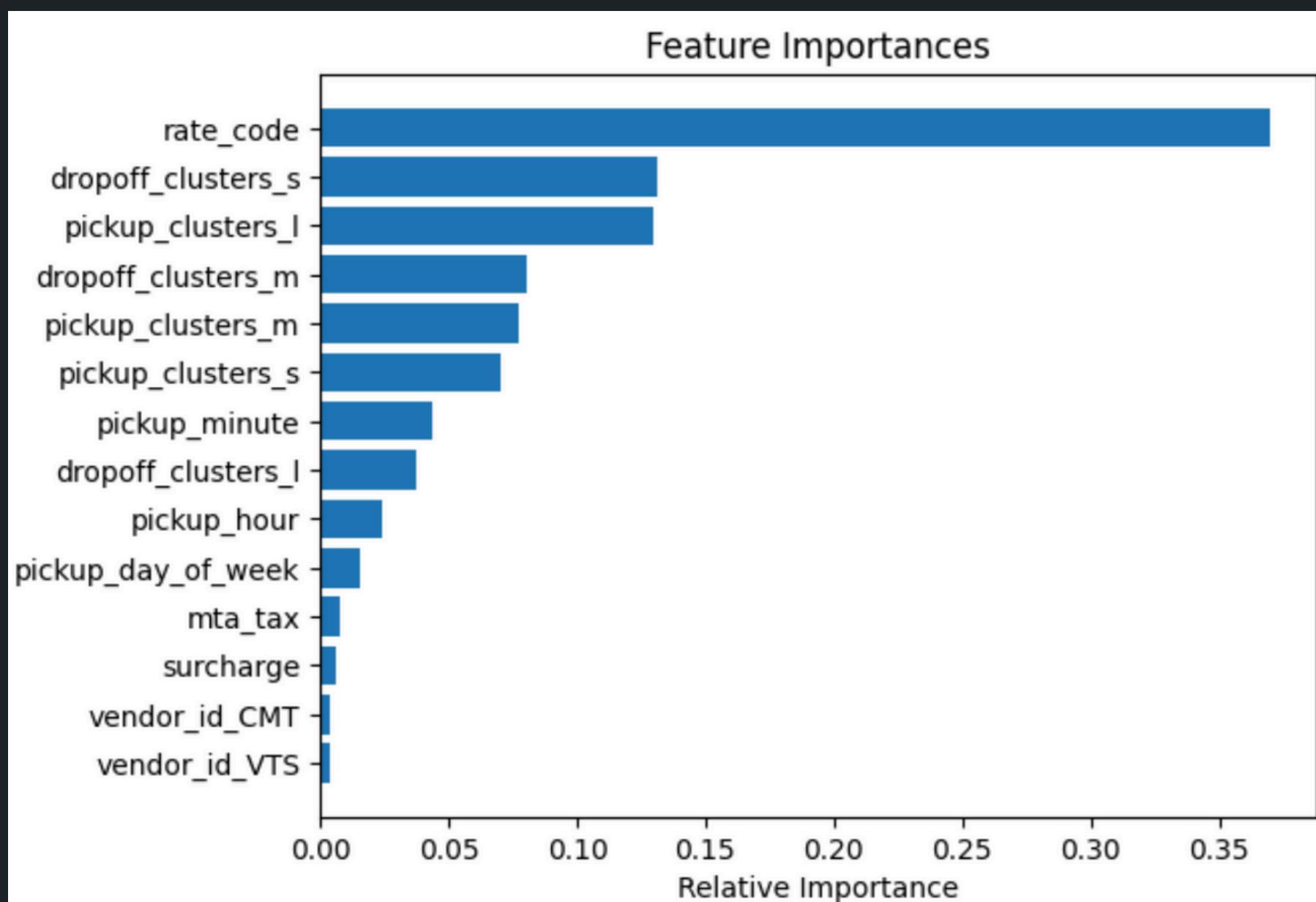
REDUCED MODELS

Il miglior modello ridotto **XGBoost** presenta un **MAE di 2.77 \$ (~20%)**

L'**RMSE** è significativamente più alto del MAE, suggerendo che il modello potrebbe avere difficoltà a gestire gli **outliers**.

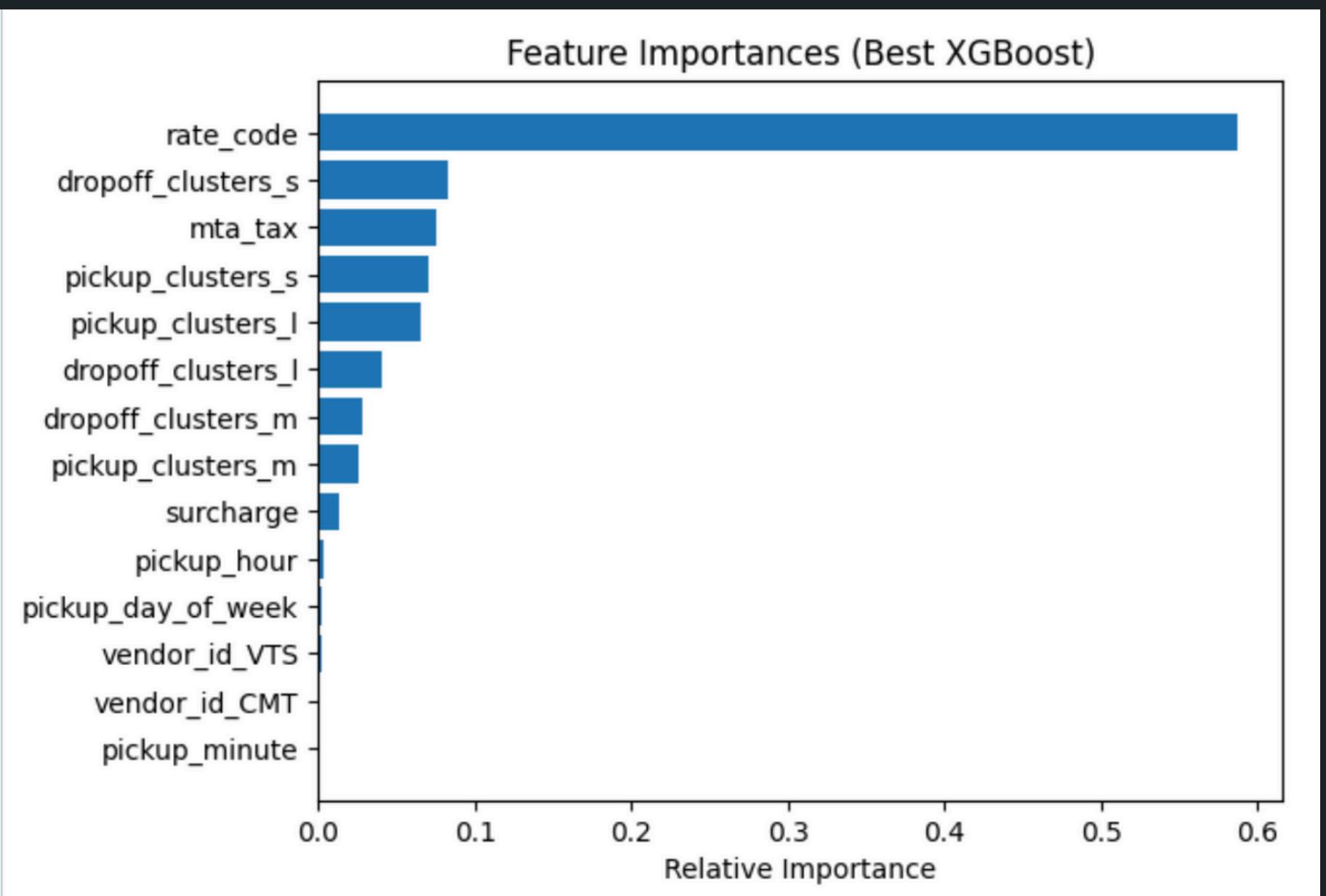
Il modello utilizza in modo preponderante il codice tariffa (fi ~60%) e le tratte di percorrenza (~fi 35% come somma delle variabili clusters)

DecisionTreeRegressor



Compare by	DecisionTreeRegressor	Hypertuned XGBoos
R2	0.78	0.86
mae	3.27	2.77
rmse	5.43	4.36

Hypertuned XGBoost

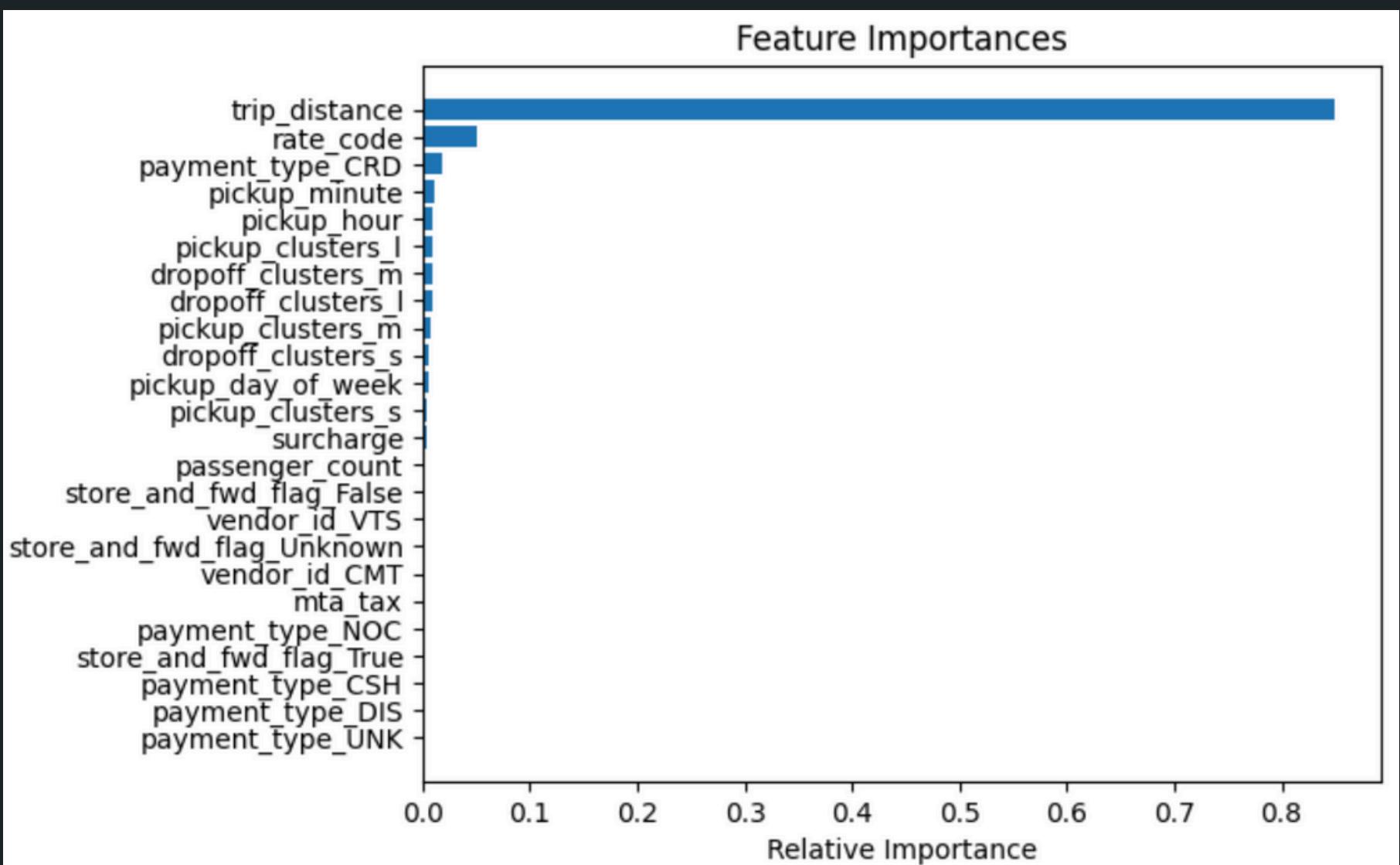


BASE MODELS

Il miglior modello base **XGBoost** riporta un **MAE di 1.51 \$ (~11%)** ed un **R2 del 94%**, l'**RMSE** presenta invece lo stesso problema dei reduced models, faticando sugli outliers.

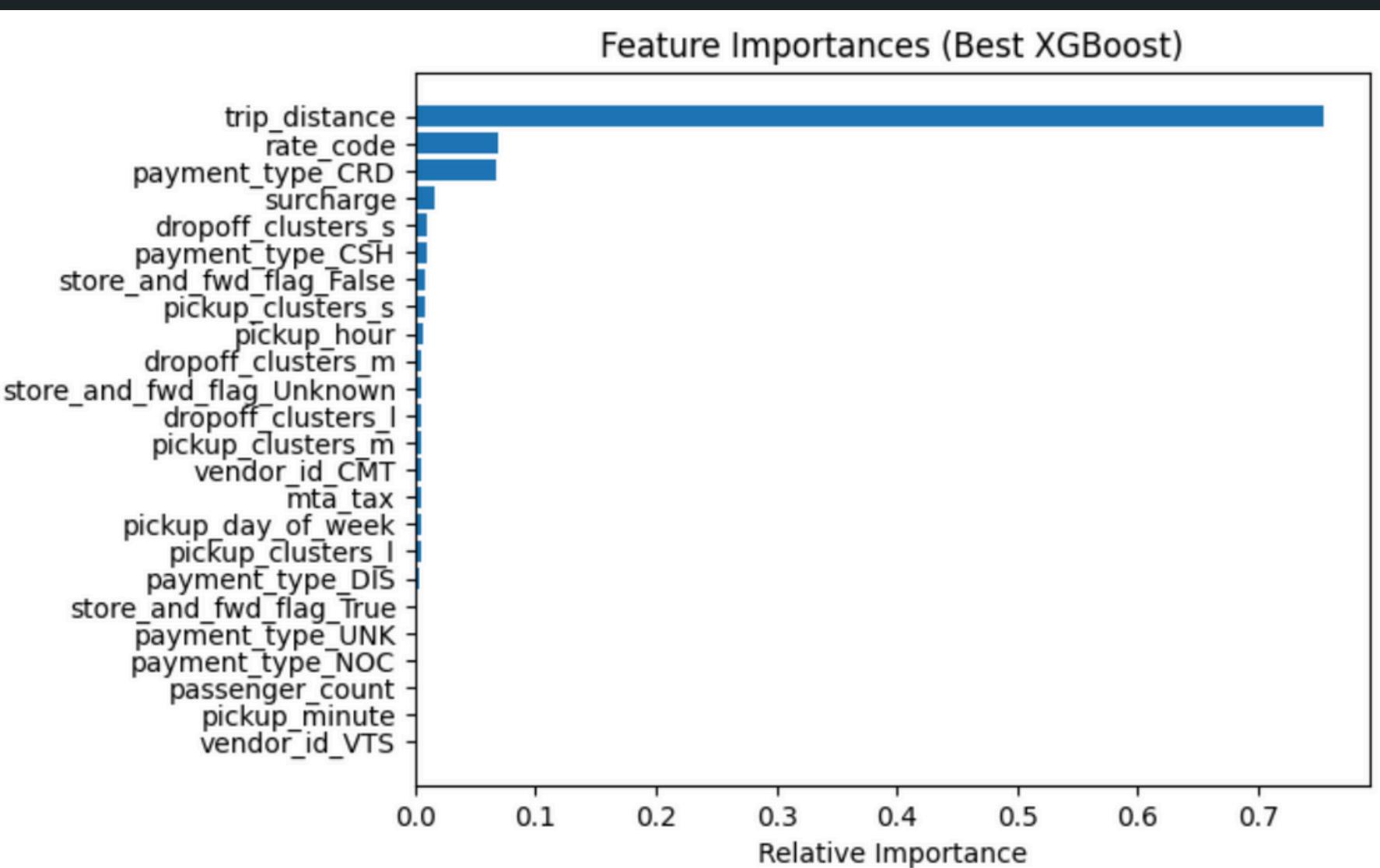
Domina il ranking di feature importante la distanza del tragitto (fi ~80%), che insieme a codice tariffa e metodo di pagamento sommano a ~95% del totale.

DecisionTreeRegressor



Compare by	DecisionTreeRegressor	Hypertuned XGBoost
R2	0.89	0.94
mae	1.96	1.51
rmse	3.75	2.82

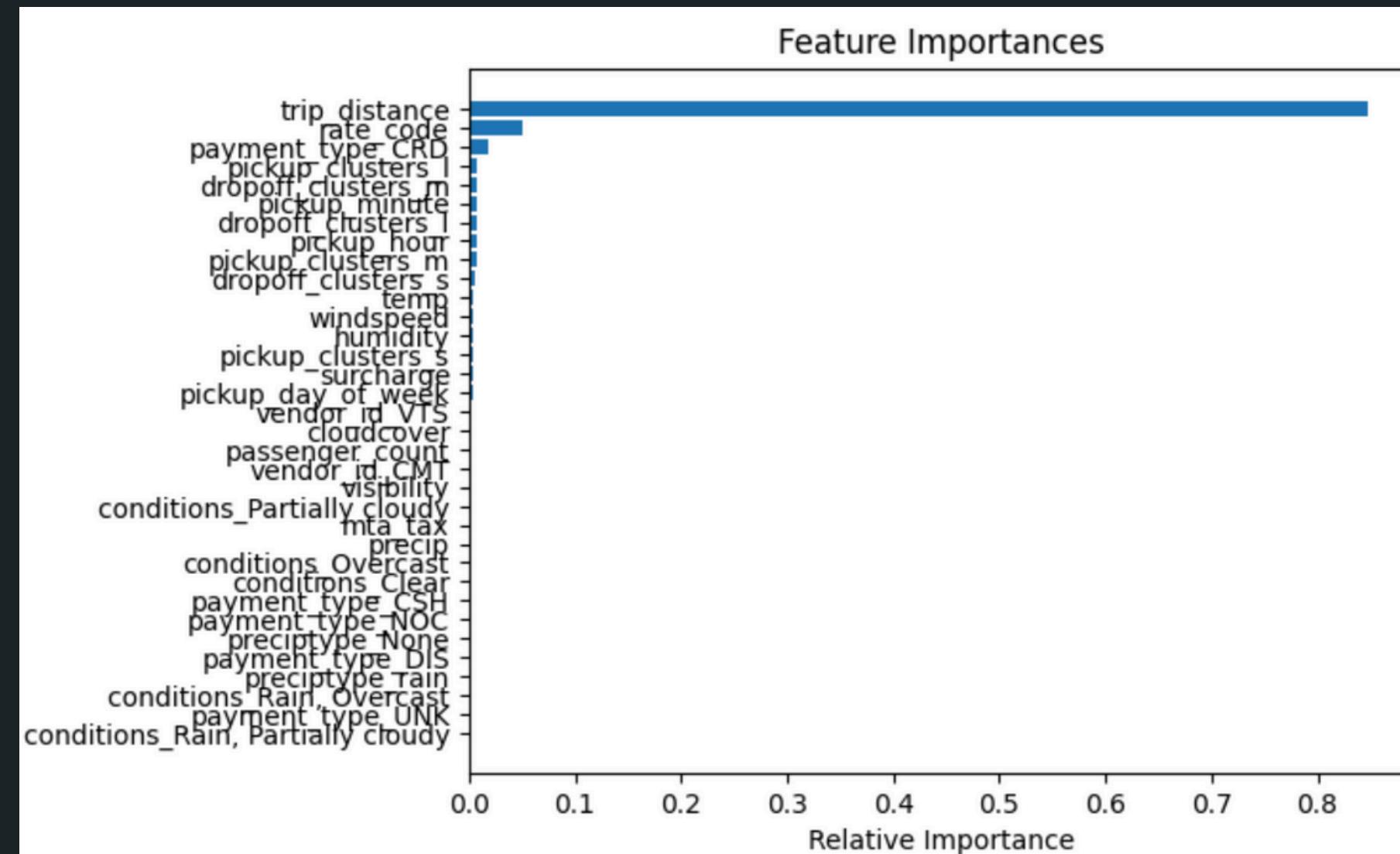
Hypertuned XGBoost



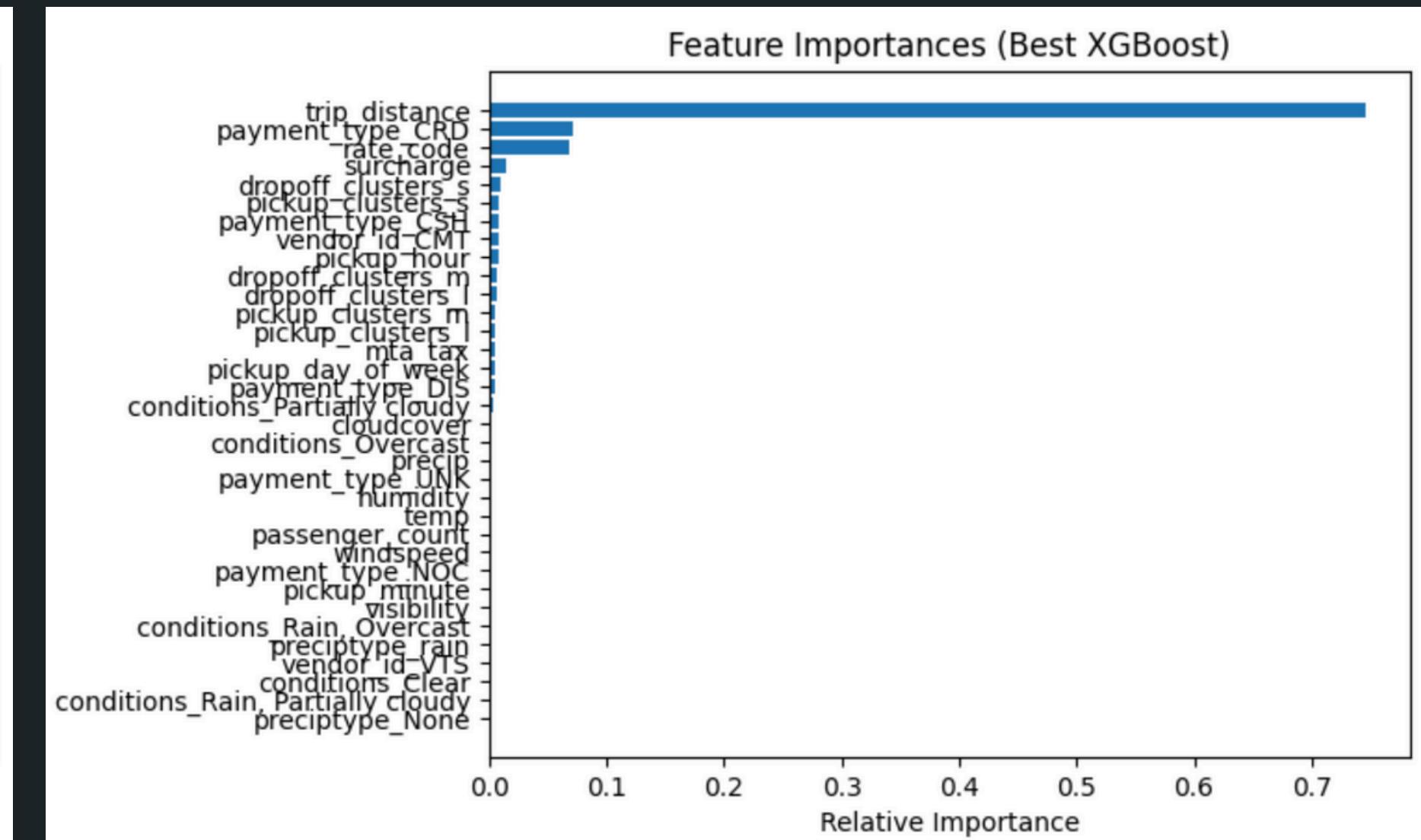
ENRICHED MODELS

Il miglior modello enriched **XGBoost** riporta valori metrici paritetici al base models. Le variabili atmosferiche per quanto utilizzate dal modello non riescono a catturare varianza addizionale e ad aumentare la capacità predittiva complessiva del modello.

DecisionTreeRegressor



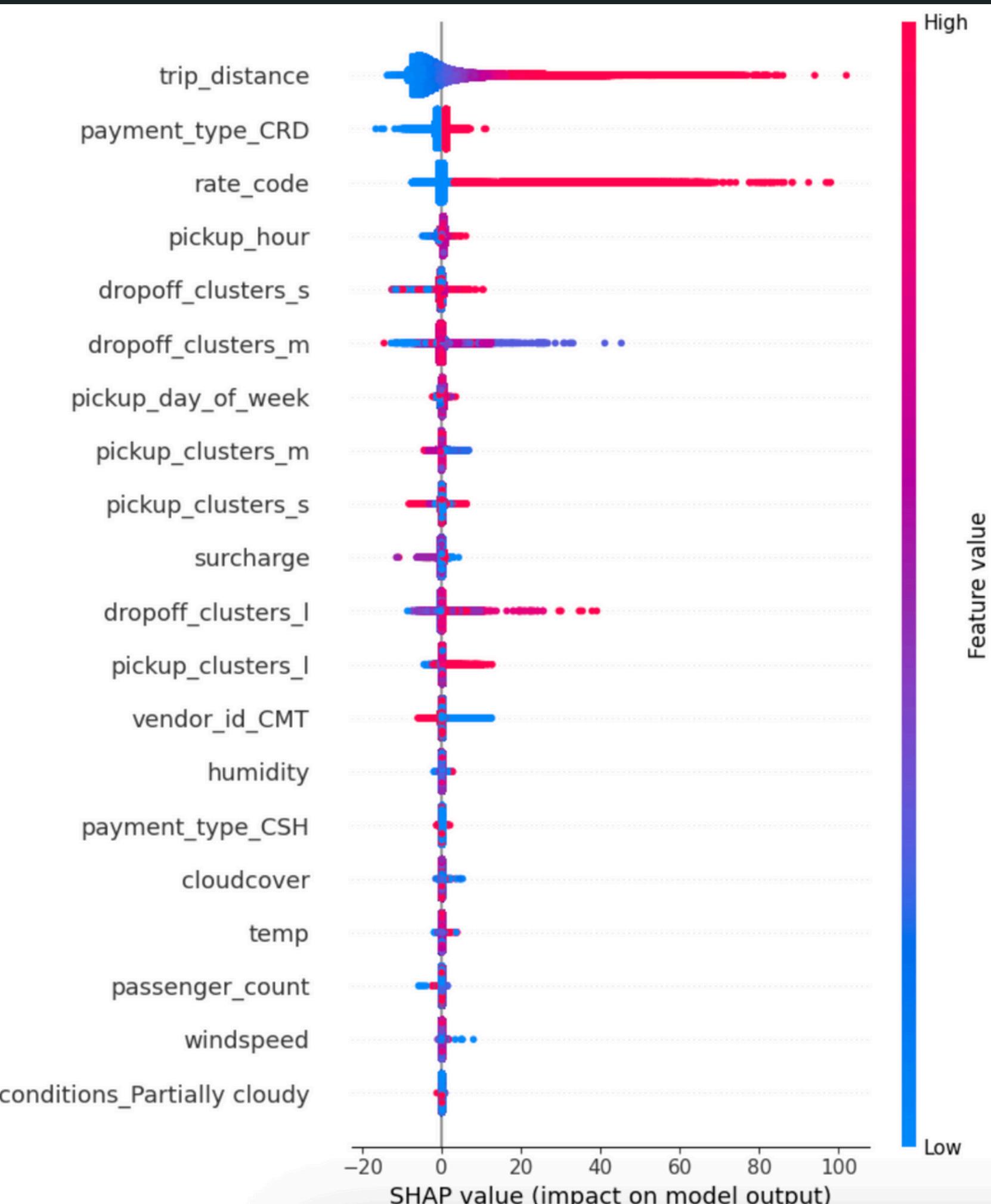
Hypertuned XGBoost

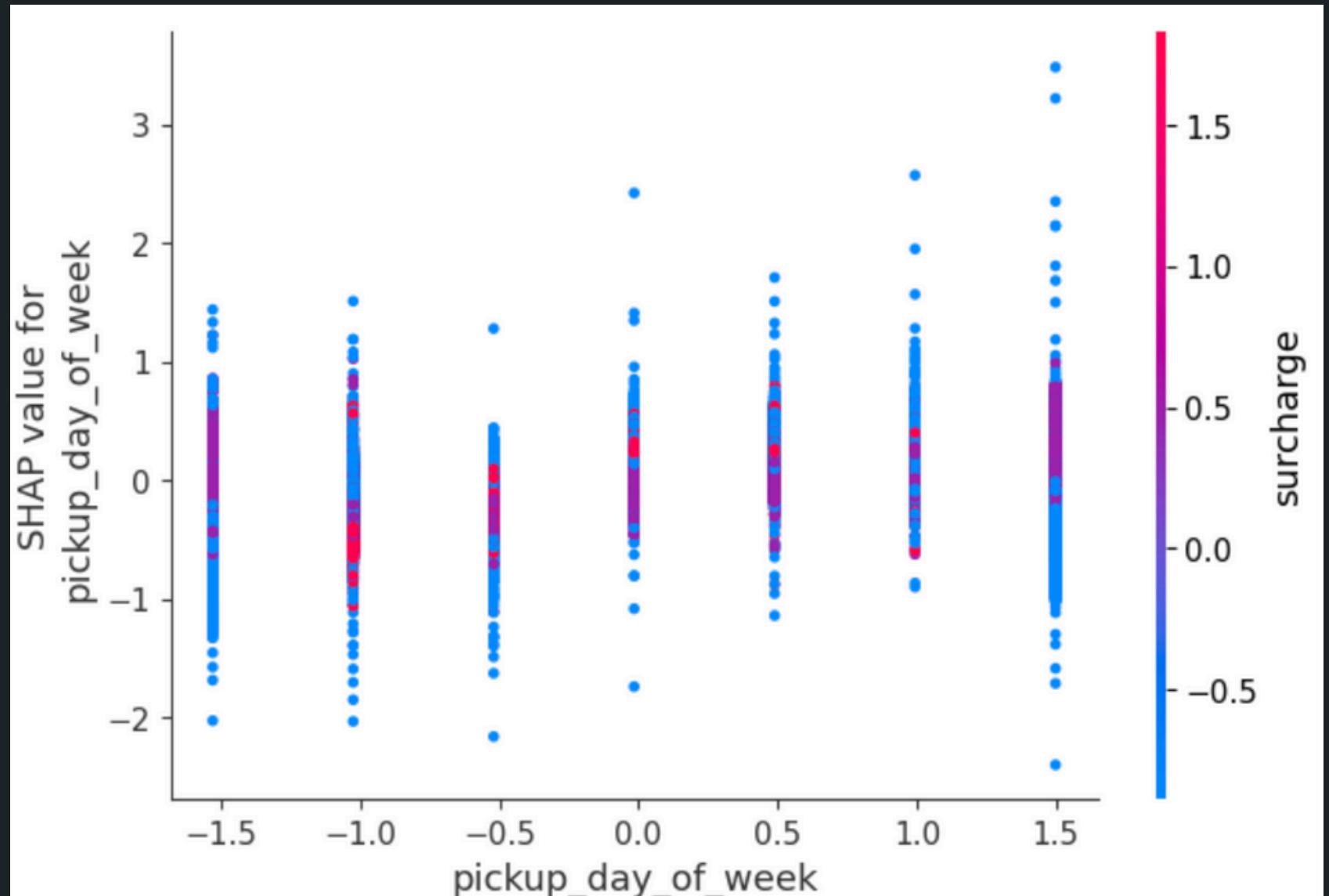


IL MODELLO SVELATO

BEST ENRICHED HYPER-XGBOOST MODEL

- **Distanze di tragitto** molto elevate e **codici tariffa** differenti dallo standard, impattano in modo importante sulla predizione finale del modello. Tuttavia, da notare che la quota maggiore di osservazioni si concentra su piccoli valori negativi prossimi allo 0, quindi con contribuzioni minori.
- **Le tratte percorse, combinazioni di cluster di pickup e dropoff**, seguono nel ranking, con contribuzioni sensibilmente più basse rispetto alle variabili precedenti e con spinte direzionali non sempre correlabili al valore assunto dalla variabile stessa.



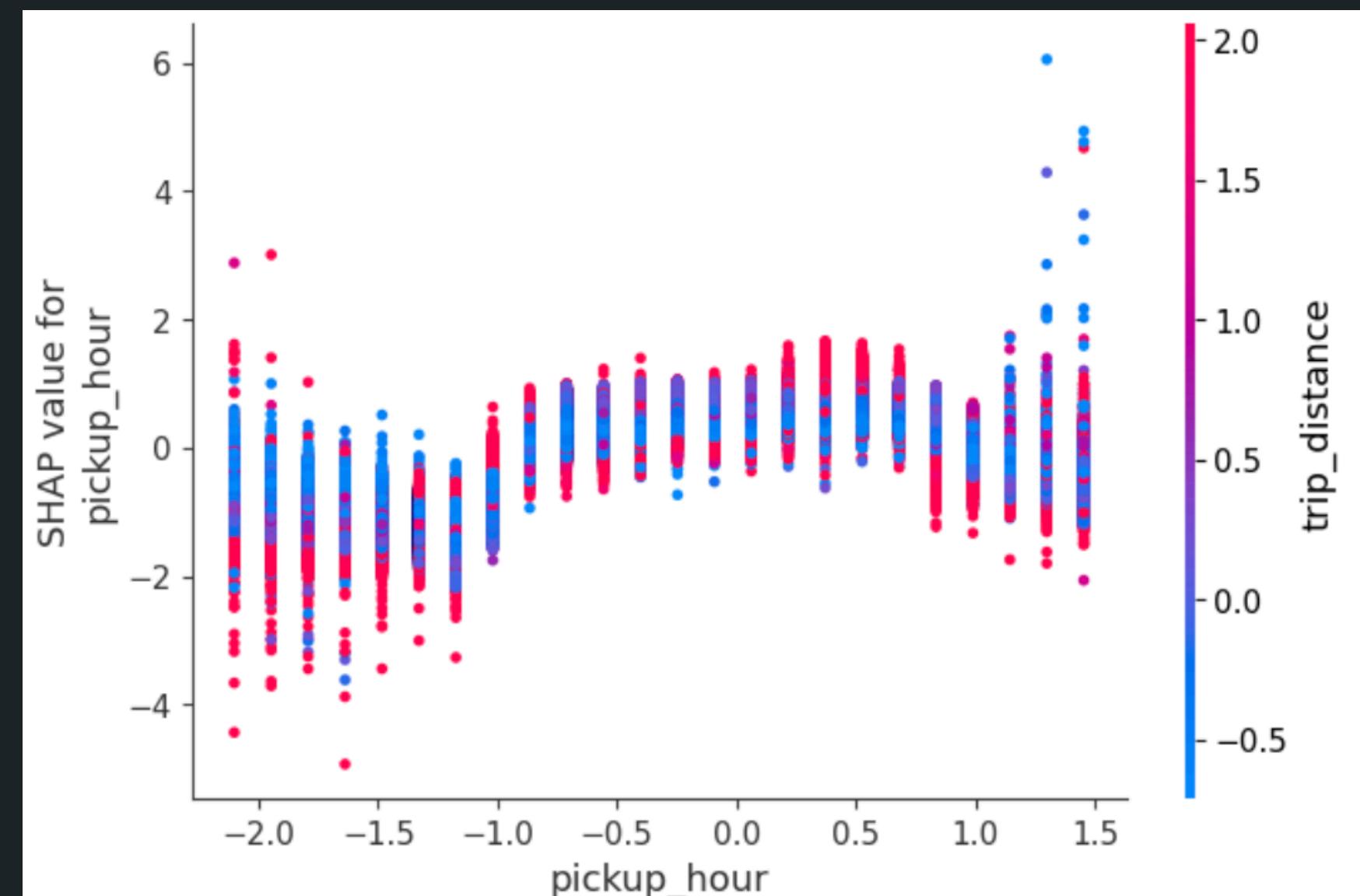


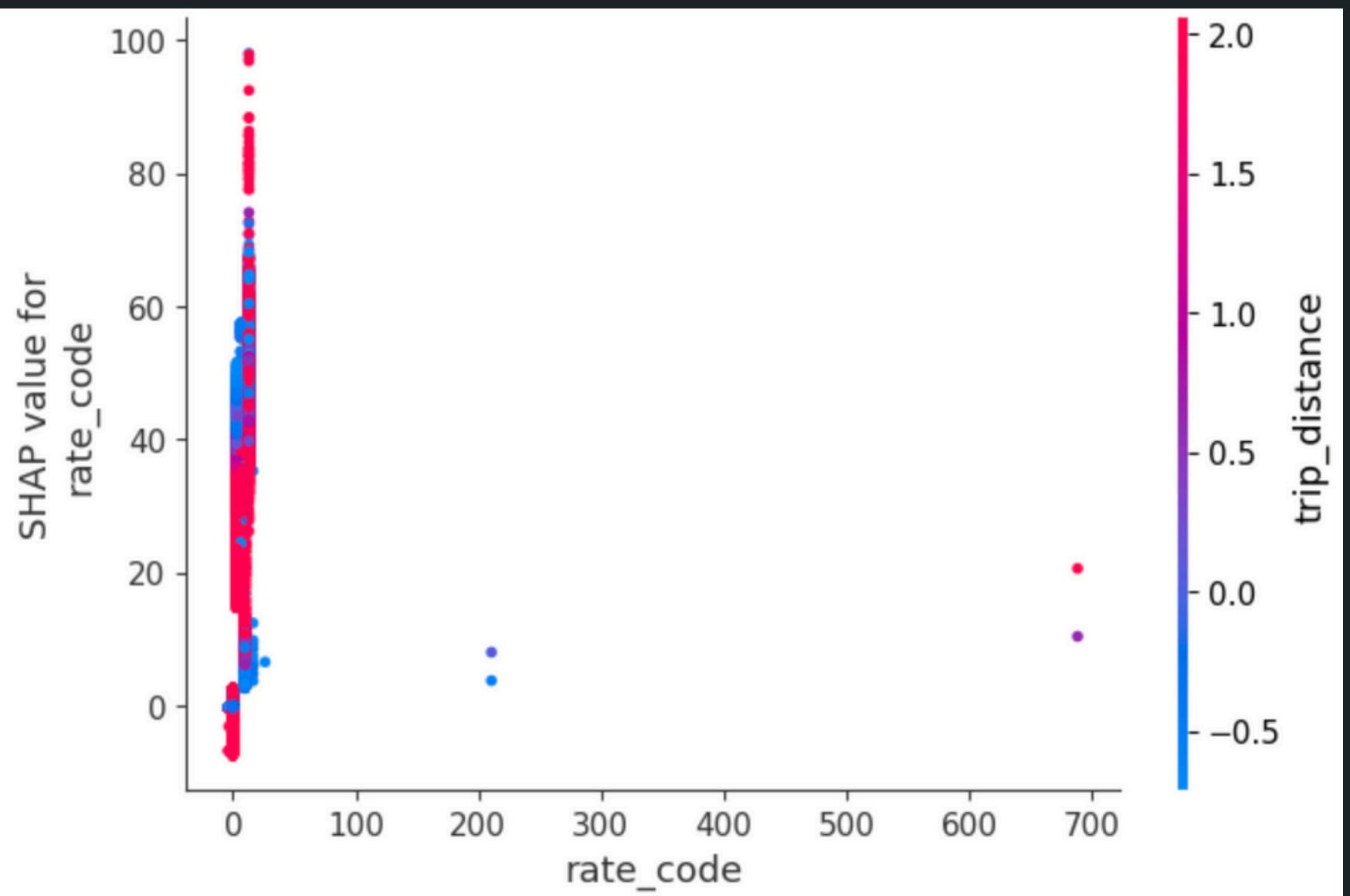
PICKUP - FASCIA ORARIA

Il pattern per la fascia oraria è più evidente. Transitando dalle 00:00 alle 6:00 la predizione di costo si dirige verso valori più bassi. Il trend si capovolge dalle 7:00 alle 19:00 aumentando in modo sensibile, infine per appiattirsi nella fascia serale.

PICKUP - GIORNO SETTIMANALE

Si nota un leggero pattern. Più si entra nella settimana più la predizione del costo della corsa aumenta.



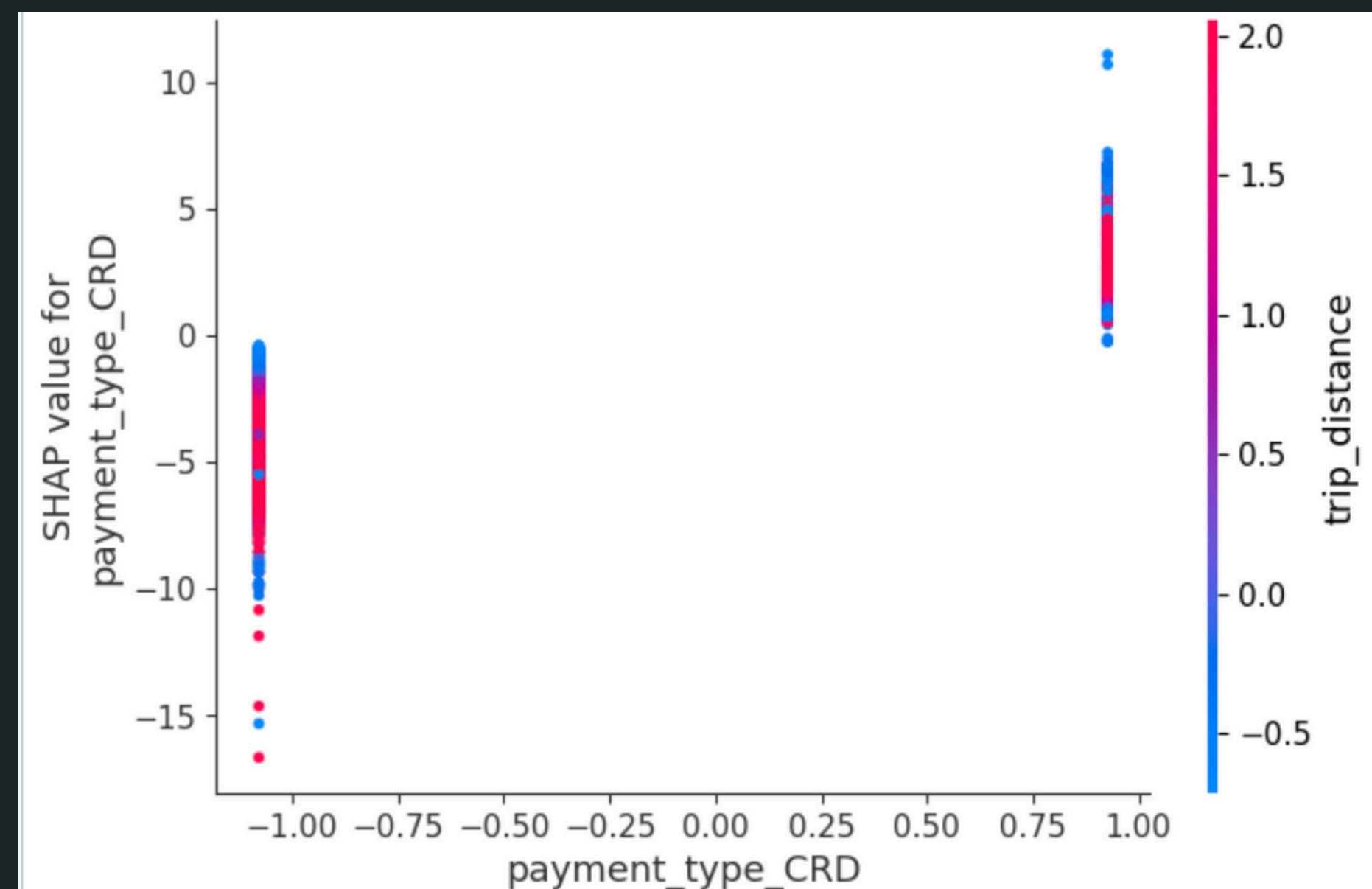


RATE CODE

Il codice tariffa 2 (spostamenti da/verso aeroporti è il vero fattore determinante alla predizione del modello, con spinte estremamente elevate.

PAYMENT TYPE

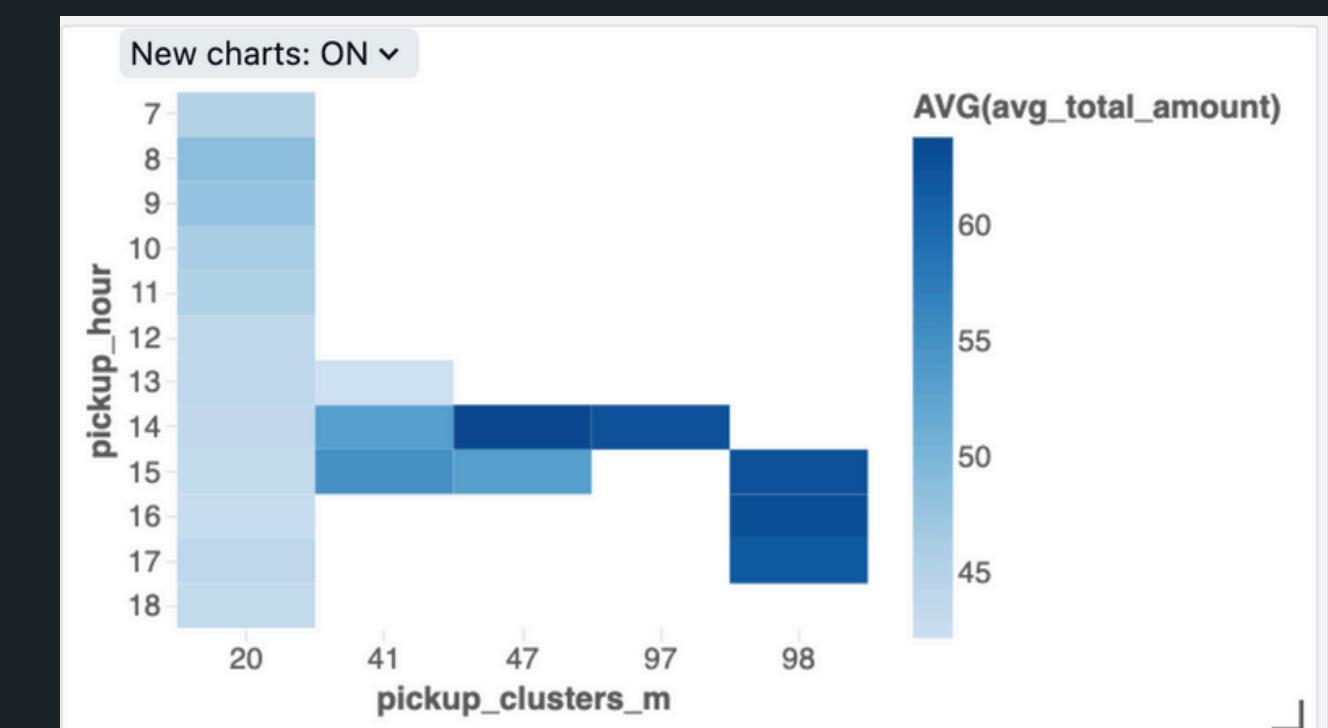
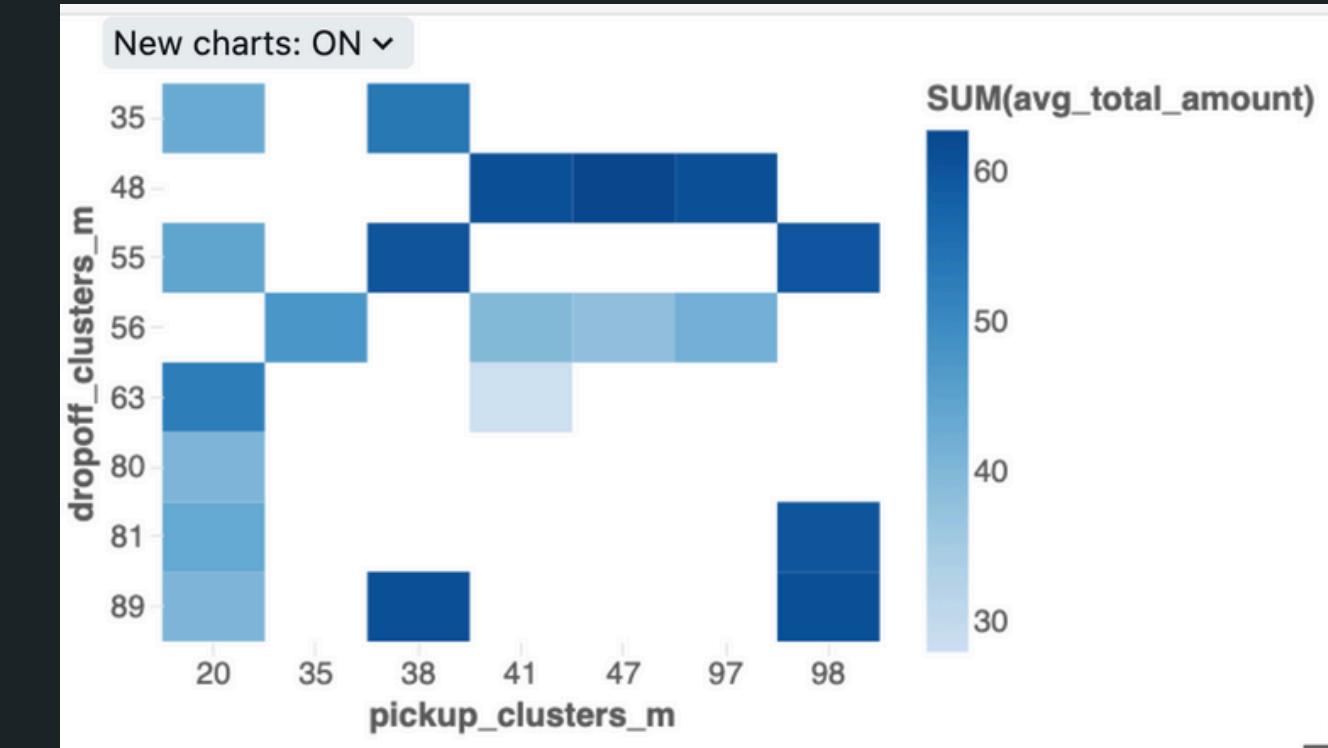
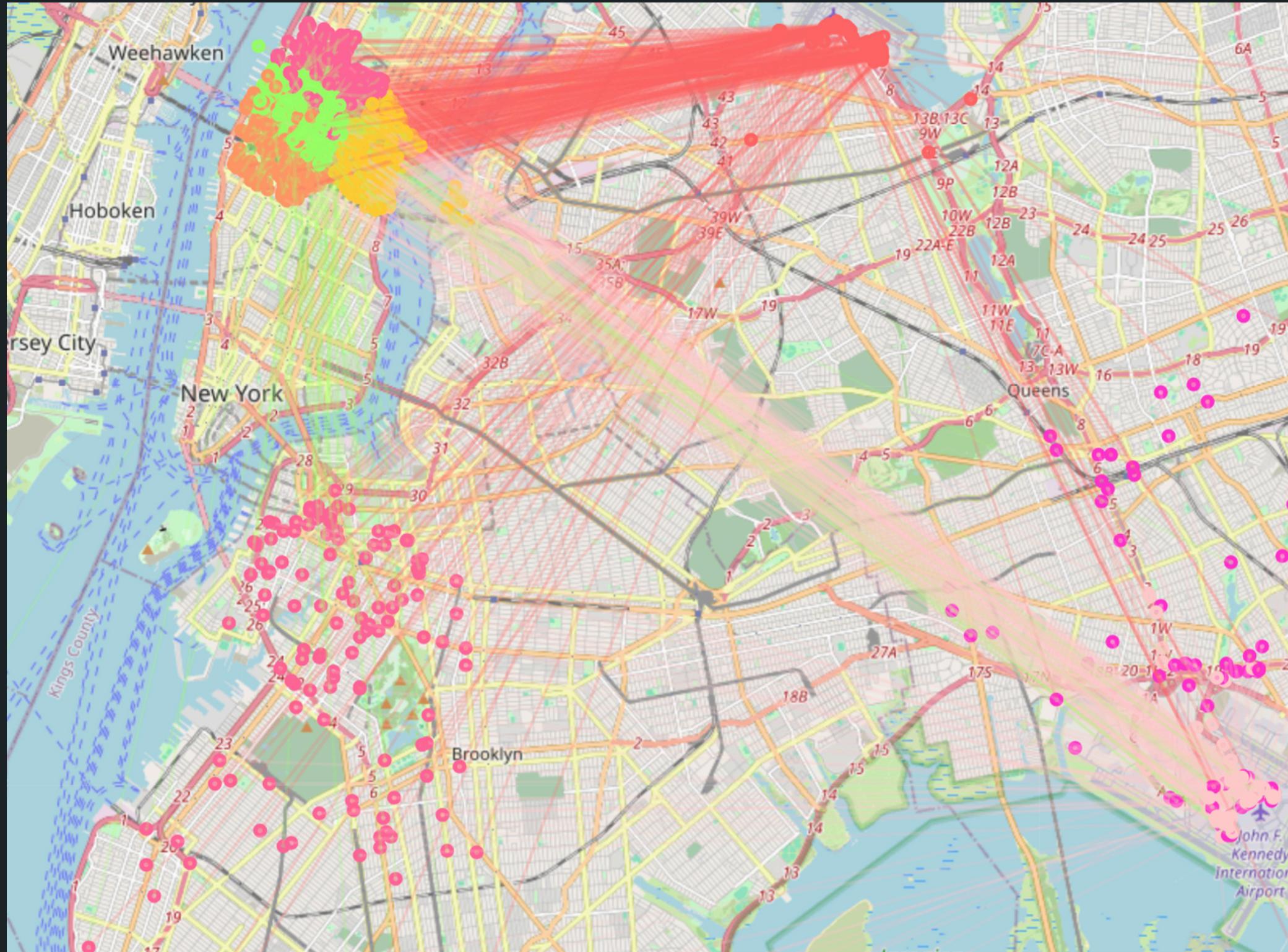
Il modello presenta un pattern predittivo ben chiaro con riferimento alla modalità di pagamento adottata. Il pagamento con Carta porta ad un costo di corsa maggiore, rispetto al pagamento in contanti, con una direzione ad intensità simmetrica ed opposta.



Le tratte che presentano costi di corsa più alti rispetto alla media, sono principalmente da e verso gli aeroporti la Guardia e JFK.

GEO ANALISI A SERVIZIO DEL TAXI DRIVER

Il taxi driver che vuole ottimizzare i costi medi di corsa dovrebbe prediligere i cluster di percorso evidenziati in mappa



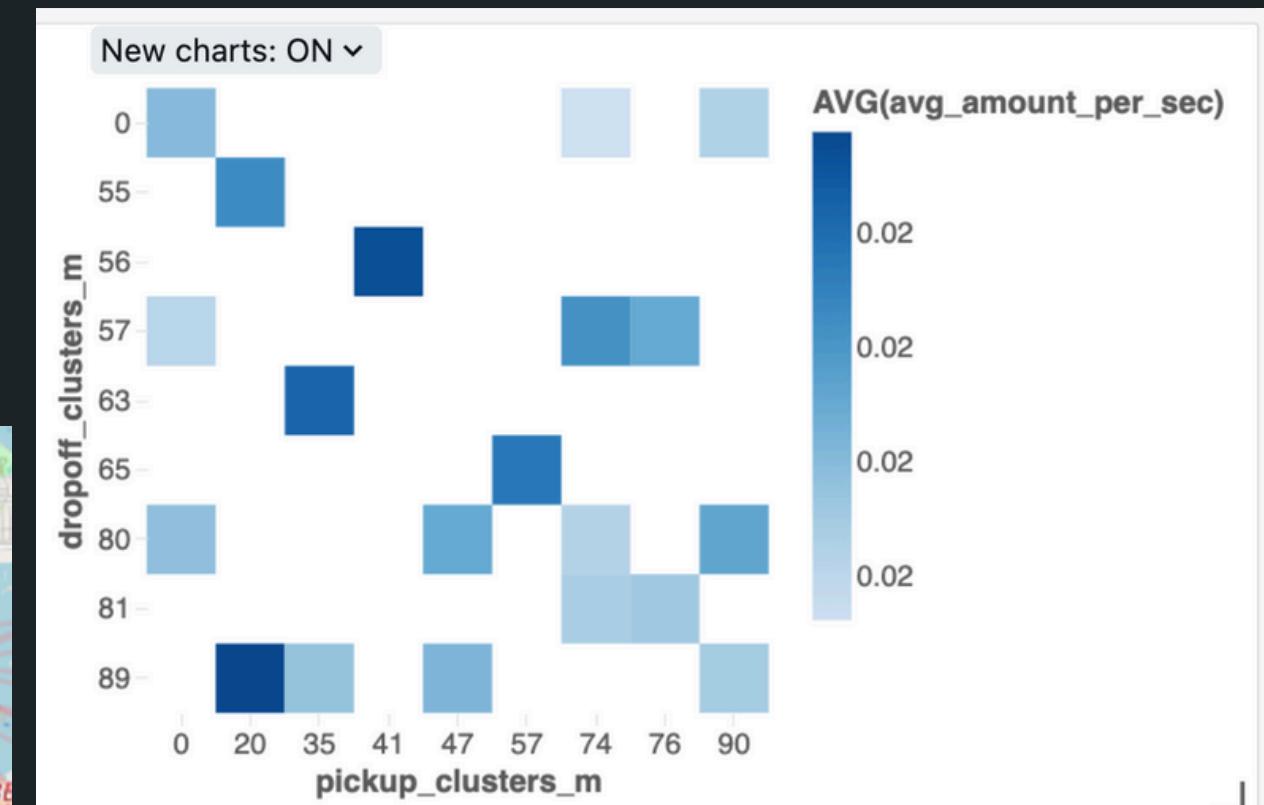
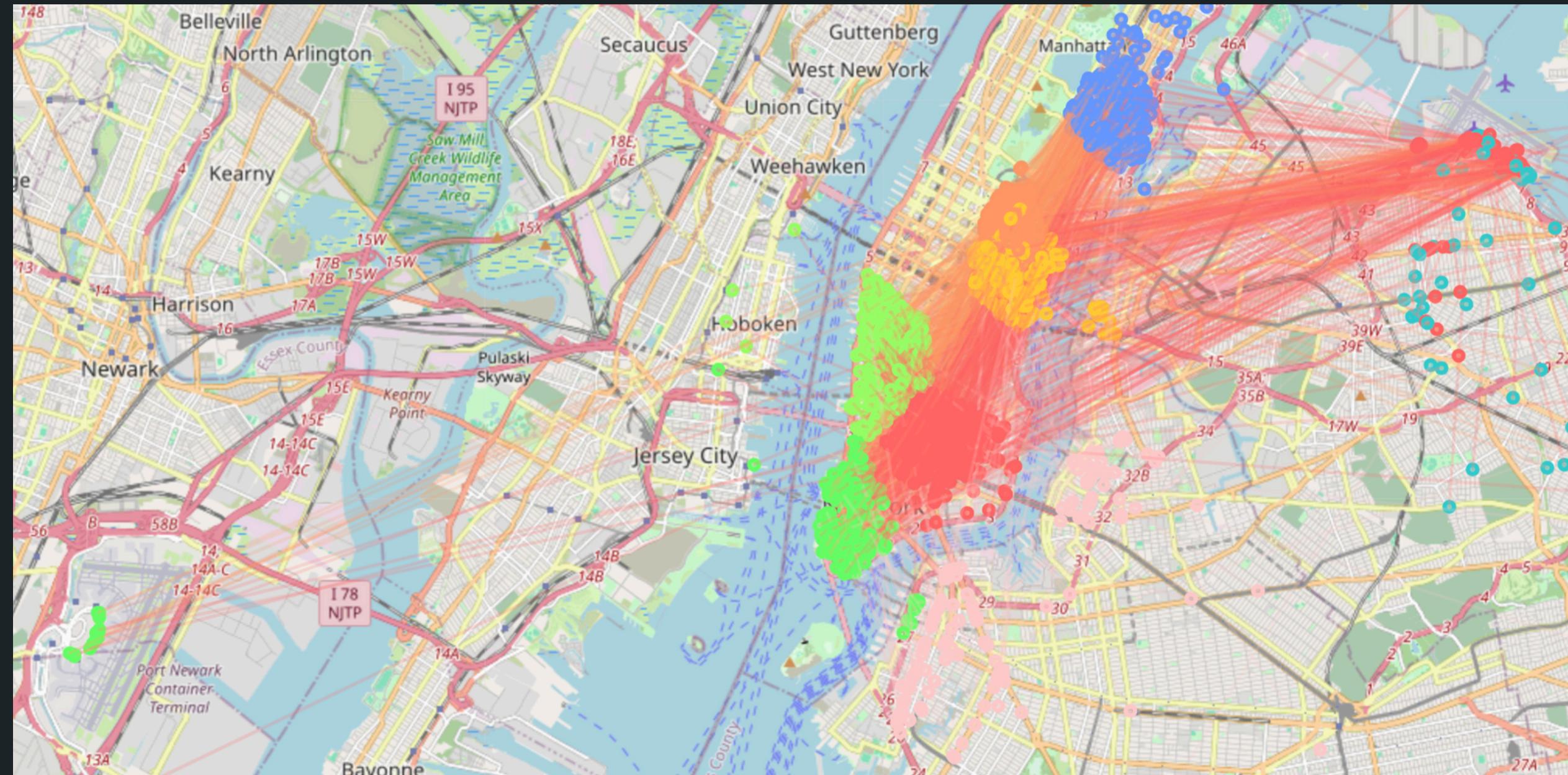
Massimizzare il costo medio di corsa significa massimizzare il profitto di fine giornata?

GEO ANALISI A SERVIZIO DEL TAXI DRIVER

Per rispondere bisogna introdurre una ulteriore variabile. Il tempo reale della corsa. Corse con tempi minore di percorrenza consentono di realizzare un numero maggiore di corse nella giornata.

Le evidenze cambiano.

Le tratte a miglior rapporto tempi/costi sono da/verso l'aeroporto della Guardia ed in parte dal Port Network. Il JFK viene escluso dalle risultanze.



1

COLLOCAZIONE

Aeroporto la Guardia per il pickup

2

PAGAMENTO

Non accettare pagamento in contanti

3

QUANDO

Dalle 14 alle 22

4

CHI FAR SALIRE

Chiedere a chi è in fila dove va e far salire chi si dirige all'Empire State Building, trascurando gli altri!!!



CONCLUSIONI

"Il risultato arriva a chi non teme di sacrificarsi e sperare."

Anonimo

....Mi sbaglio??

Io