FINAL PROJECT – GROUP 2

TAXI NY LLC

**DOMINIO DEL POBLEMA:**

**La Città di New York**

**Approfondimenti**

* **NEW YORK (Stato)**
* **NEW YORK City (grande metropolitana) formata da cinque distretti: Manhattan, Brooklyn, Queens, Bronx e Staten Island**

**Descrizione del problema:**

New York è una città trafficata e affollata, con una grande richiesta di servizi taxi perché più veloce rispetto ai mezzi di trasporto locali.

La gestione di una grande flotta di taxi a New York può essere un compito complesso e impegnativo. Per massimizzare le entrate e minimizzare i costi, le compagnie di taxi devono ottimizzare il percorso e la programmazione dei loro taxi.

Tuttavia, questo può essere difficile senza avere accesso a dati precisi e aggiornati sulle corse dei taxi e sulla domanda dei passeggeri. Per affrontare questi problemi, proponiamo di utilizzare tecniche di ingegneria dei dati per analizzare i dati delle corse dei taxi e identificare le aree di miglioramento nella gestione della flotta di taxi.

**DOMINIO DEL POBLEMA:**

**Obiettivo:**

L'obiettivo di questo progetto di ingegneria dei dati è analizzare i dati relativi alle corse dei taxi a New York per ottimizzare la gestione della flotta di taxi per una compagnia di Taxi (“Taxi NY”).

Il progetto si propone di utilizzare le tecniche di ingegneria dei dati per elaborare e analizzare grandi volumi di dati sulle corse dei taxi per fornire approfondimenti e raccomandazioni per migliorare la gestione della flotta di taxi a New York.

L'ingegneria dei dati può di fatti anche essere utilizzata per migliorare l'esperienza del cliente, fornendo raccomandazioni personalizzate e migliorando l'accuratezza delle stime delle corse.

**Fonte dati:**

Per analizzare i dati sulle corse dei taxi abbiamo utilizzato un dataset train.csv (dal sito Kaggle).

I dati comprendono le località (in coordinate geografiche), la data/ora di partenza e di arrivo, il conteggio dei passeggieri, la durata della tratta in secondi le distanze, gli importi delle tariffe, i tipi di pagamento e altri dettagli relativi alle corse in taxi effettuate nella città di New York nell’anno 2016.

In locale che lo richiamaiamo nel

Lo tiriamo fuori taxi\_ny

Prima di caricare screamo lo schema

Facciamo i test e popoliamo la tabella

**Caricamento dei dati:**

Il dataset train.csv originale contiene più di un milione di righe e quindi abbiamo deciso di fare un random sample per estrarre un campione casuale dal dataset e abbiamo scaricato il file (nominandolo “trainsampled.csv”) in locale per poi richiamarla nello script Python.

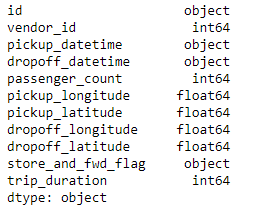
**Esplorazione dei dati:**

Ecco la descrizione di tutte le variabili/caratteristiche disponibili nel data set di dati che ci aiuteranno a comprendere meglio:

|  |  |  |
| --- | --- | --- |
| **VARIABILI** | **DESCRIZIONE** | **FORMATO** |
| **id** | un identificatore univoco per ogni viaggio | String |
| **vendor\_id** | un codice che indica il fornitore associato al record del viaggio | Numeric (int) |
| **pickup\_datetime** | data e ora in cui il contatore è stato attivato | Date\_Time |
| **dropoff\_datetime** | data e ora in cui il tassametro è stato disattivato | Date\_Time |
| **passenger\_count** | il numero di passeggeri nel veicolo (valore inserito dal conducente) | Numeric (int) |
| **pickup\_longitude** | la longitudine in cui il tassametro è stato attivato | Numeric (double) |
| **pickup\_latitude** | la latitudine in cui il tassametro è stato attivato | Numeric (double) |
| **dropoff\_latitude** | la longitudine in cui il tassametro è stato disattivato | Numeric (double) |
| **store\_and\_fwd\_flag** | questo flag indica se il record di viaggio è stato conservato nella memoria del veicolo prima di essere inviato al fornitore perché il veicolo non aveva una connessione al server (Y=store and forward; N=non un viaggio store and forward) | String |
| **trip\_duration** | la durata del viaggio in secondi | Numeric |
|  |  |  |

((((((((((df.dtypes))))))))))

(((((((((df.describe))))))))



(((((((((((( missing values o meno valori duplicati e numero di righe )))))))

Visto la mancanza delle informazioni di nostro interesse come la **distanza** e la **tariffa del viaggio** ci siamo appoggiati a un API di geocodifica di Geoapify.

In particolare, conoscendo le coordinate geografiche della partenza e dell’arrivo del tragitto del taxi possiamo risalire all’**indirizzo**, la **città**, lo **stato** e il **quartiere** utilizzando Reverse Geocoding API. Di seguito il link

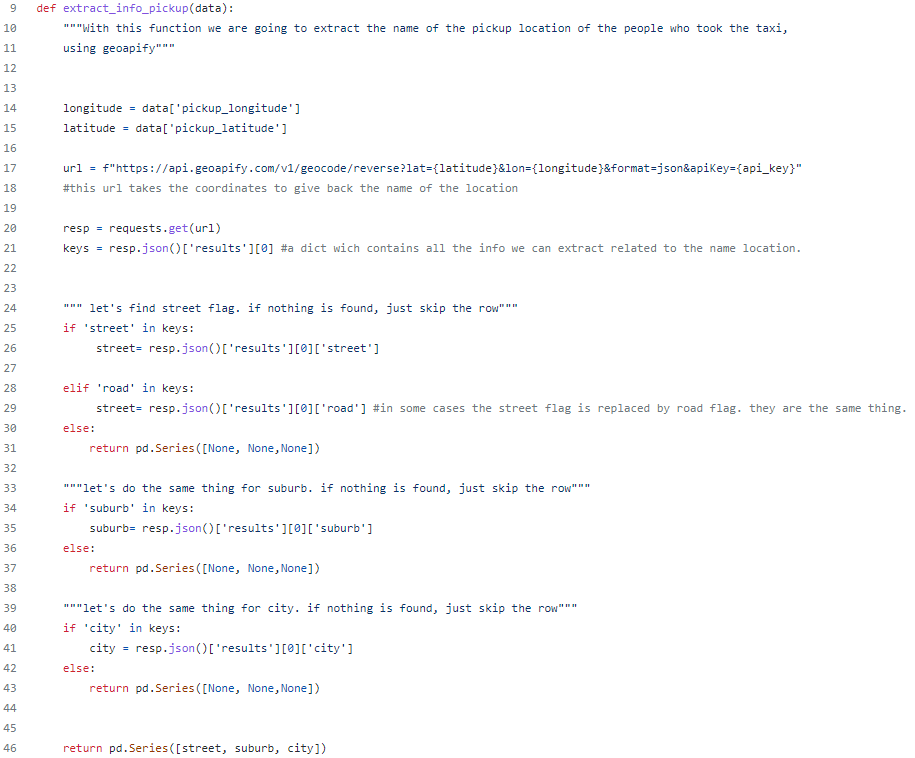
[**https://apidocs.geoapify.com/docs/geocoding/reverse-geocoding/#about**](https://apidocs.geoapify.com/docs/geocoding/reverse-geocoding/#about)

Per questo abbiamo creato 4 funzioni:

1. extract\_info\_pickup(data)
2. extract\_extra\_dropoff(data)
3. calculate\_distance(data)
4. calculate\_fare\_2016(data)
5. La funzione extract\_info\_pickup(data) date le coordinate di partenza restituisce solo le informazioni che ci interessano dal file .json dell’API Geoapify

street, suburb, city

di partenza.



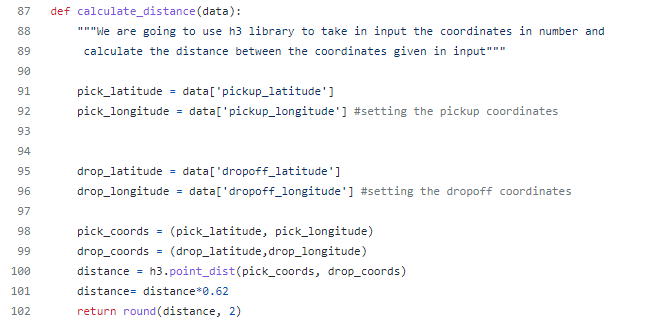
1. La funzione extract\_extra\_dropoff(data) date le coordinate d’arrivo restituisce solo le informazioni utili

street, suburb, city

dell’arrivo.



1. La funzione calculate\_distance(data) restituisce la distanza in miglia date le coordinate di ogni tragitto presente nel dataset



1. calculate\_fare\_2016(data)

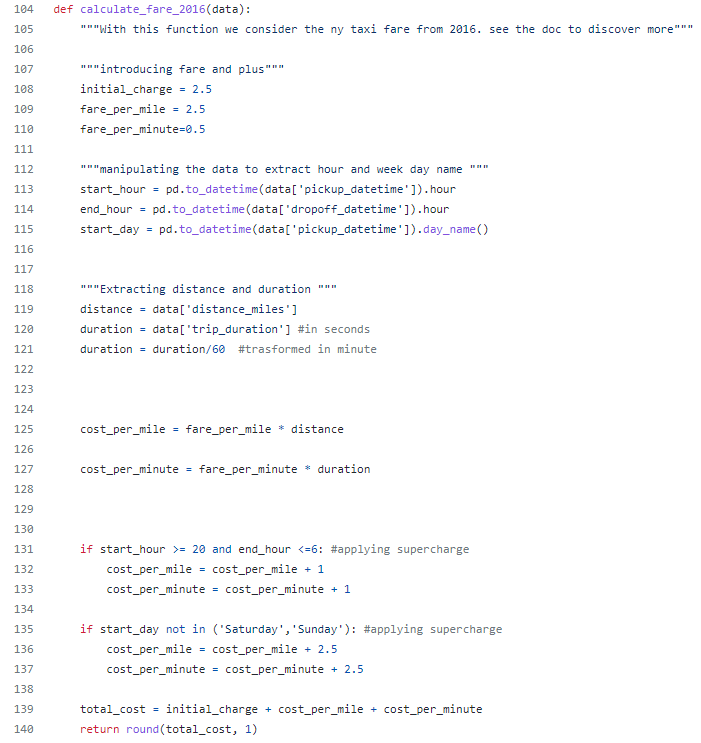
Per quanto riguarda la tariffa del tragitto del taxi ci siamo informati sul metodo di tariffazione dei taxi di NY del 2016 e abbiamo ipotizzato un sistema di tariffazione che tenesse conto della durata (in minuti) e la distanza percorsa (in miglia) e del giorno della settimana e dell’orario in cui avviene il tragitto.

Dalle 20.00 alle 6:00 e nei giorni Sabato e Domenica viene applicato una tariffa maggiore (supercharge)

Il **costo totale in dollari** è la somma del **costo iniziale (2.5) + costo per miglio + costo al minuto**

🡪 Dal sito del governo <https://www.nyc.gov/site/tlc/passengers/taxi-fare.page>

La tariffa viene calcolata come segue:



Sviluppo futuro: è difficile prevedere la durata poiché il

Prevedere la tariffa della tratta sapendo la durata passata di una tratta, i giorni della settimana in cui c’è più affluenza

Il nuovo dataset (dopo integrazione con API e l’aggiunta delle colonne e l’eliminazione di colonne superflue per il dominio d’interesse) si presenta come segue:

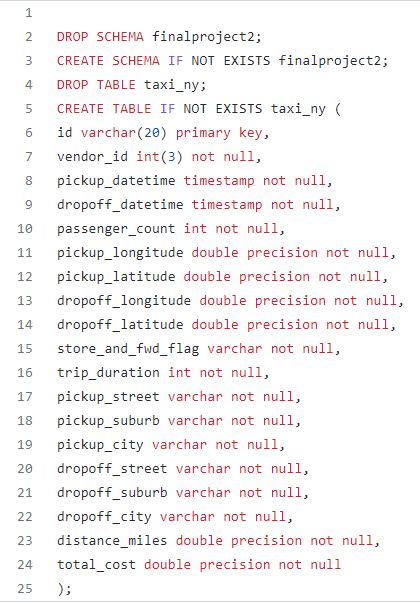




((((((((((((( sneak peack della nuova tabella con tutte le colonne di interesse )))))))))))))))

**POPOLIAMO IL NOSTRO DATABASE**

Creazione di un nuovo schema **finalproject2**



**Metriche chiave:**

Utilizzeremo le seguenti metriche per valutare il successo del progetto:

* Aumento delle entrate per taxi
* Riduzione del tempo medio di attesa per i passeggeri
* Miglioramento dell'efficienza delle rotte e della programmazione dei taxi
* Aumento del grado di soddisfazione dei clienti

**Data pipeline:**

Utilizzeremo una pipeline di dati per elaborare e analizzare i dati delle corse dei taxi e strumenti di visualizzazione dei dati per creare grafici, diagrammi e mappe per presentare i nostri risultati. La pipeline di dati consisterà nelle seguenti fasi:

* Ingestione e pulizia dei dati
* Trasformazione e aggregazione dei dati
* Archiviazione e recupero dei dati
* Visualizzazione e analisi dei dati

(((((((( disegno del power point )))))))

**Tecnologie utilizzate:**

Per implementazione e analisi della pipeline di dati utilizzeremo le seguenti tecnologie:

* Linguaggio di programmazione Python per l’elaborazione e l’analisi dei dati
* AWS
* PostgreSQL per l'archiviazione e il recupero dei dati
* PowerBi per la visualizzazione dei dati

**Risultati:**

Ci aspettiamo di ottenere i seguenti risultati dall'analisi dei dati:

* Fornire raccomandazioni per la regolazione del numero di taxi nelle aree ad alta domanda e quindi aumentare gli investimenti sull’acquisto di più taxi
* Identificare le aree di miglioramento nella gestione della flotta di taxi e , come:
* Le tratte di viaggio che durano maggiormente in modo da ottimizzare i percorsi dei taxi
* Ma anche l’ottimizzazione dei taxi in base al numero di passeggieri, in base al periodo (mattina/pomeriggio/sera) e in base alle città.
* Il ricavo che ogni la tratta più frequente genera all’azienda

**Conclusioni:**

In conclusione, questo progetto di ingegneria dei dati mira a migliorare la gestione della flotta di taxi per una compagnia di taxi di New York City attraverso l'uso di tecniche di analisi dei dati e di ingegneria. Dall'analisi dei dati ci aspettiamo di trarre preziose indicazioni e raccomandazioni, che aiuteranno la compagnia di taxi a ottimizzare la gestione della flotta e a fornire un servizio taxi migliore ai suoi clienti.

**Grazie per il vostro tempo e la vostra attenzione.**