

Università degli Studi Milano Bicocca, corso di Laurea Magistrale in Data Science

Stelle Michelin e TripAdvisor

Valutazioni ai migliori ristoranti italiani

Galimberti Dario, Nardi Sara, Stoffa Giacomo

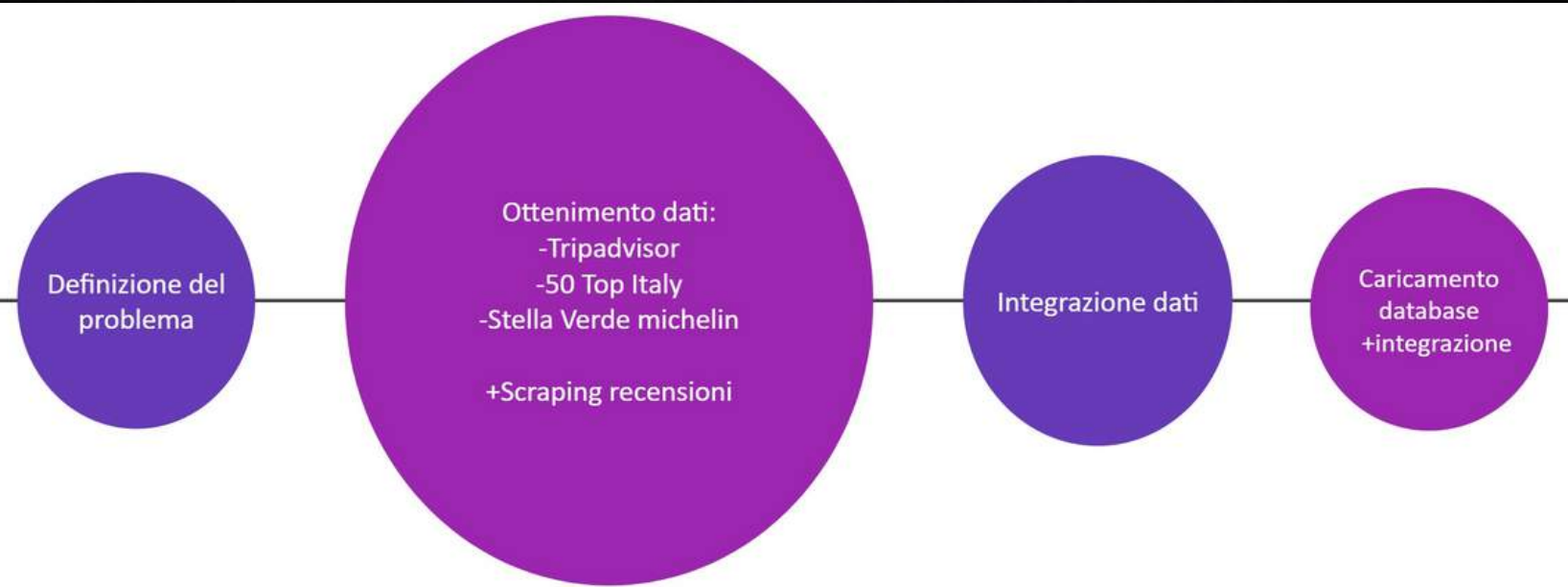
INTRODUZIONE

“Cosa sarebbe il mondo senza la cucina italiana?”

Corriere della Sera, Dott. Danilo Taino

L'opinione dei clienti comuni è la stessa di quella degli esperti gastronomi?

E gli strumenti a disposizione per effettuare la scelta relativa a dove sedersi a tavola sono soddisfacenti?



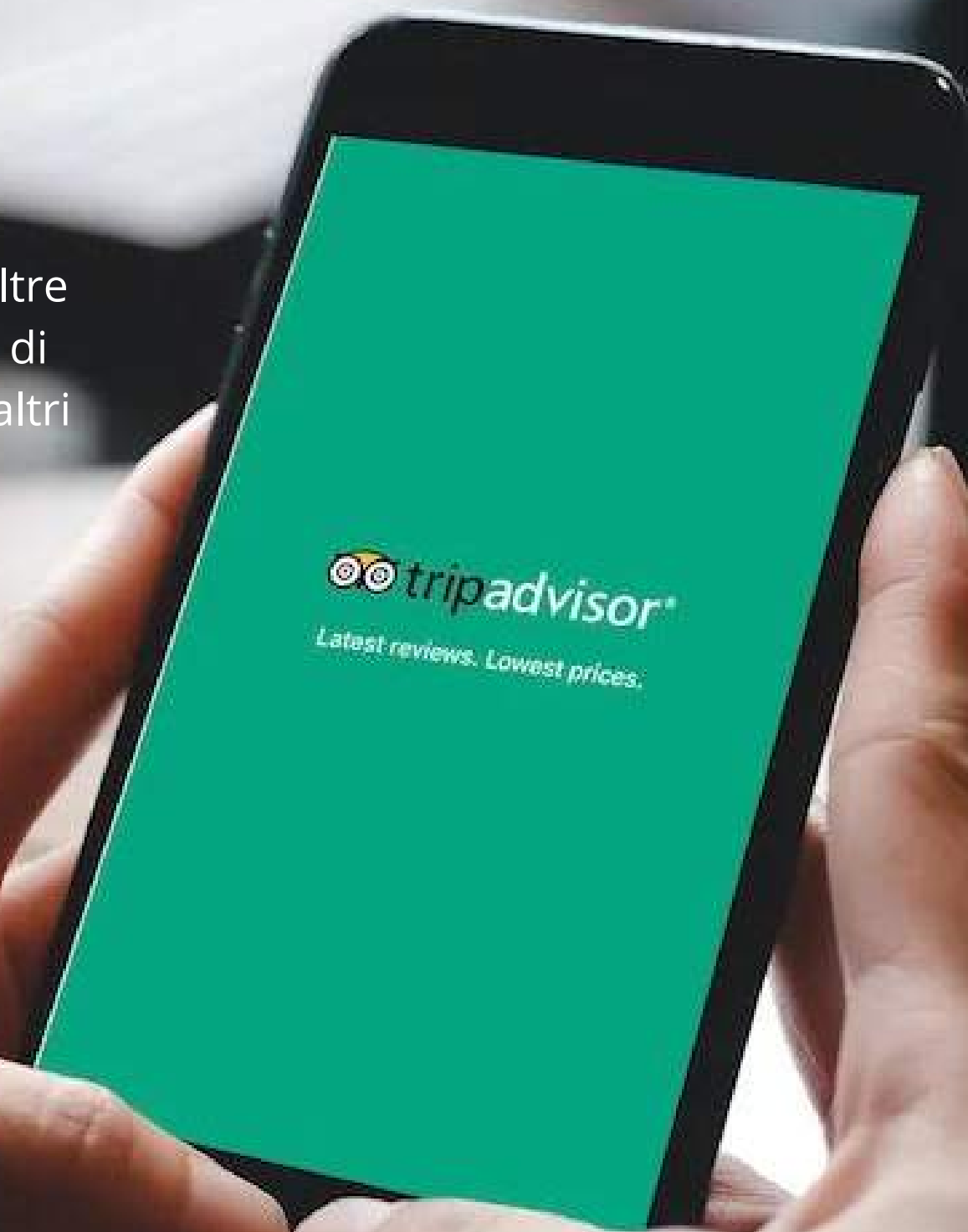
Strumenti utilizzati: Python, Tableau Public

Sorgenti dati

1. Tripadvisor

TripAdvisor è un sito web internazionale di recensioni di ristoranti e di altre tipologie di strutture ricettive. Fondato nel 2000, esso offre la possibilità di consultare in qualsiasi momento e a costo zero le opinioni rilasciate da altri utenti.

181.951 attività di ristorazione italiane su un totale di
196.000 attività presenti sul nostro territorio



Web scraping

Effettuato servendosi di una piattaforma online specializzata: "Apify".



Data Preprocessing

Concretizzato in vari tipi di elaborazioni eseguite sui dati grezzi provenienti dal web scraping, quali:

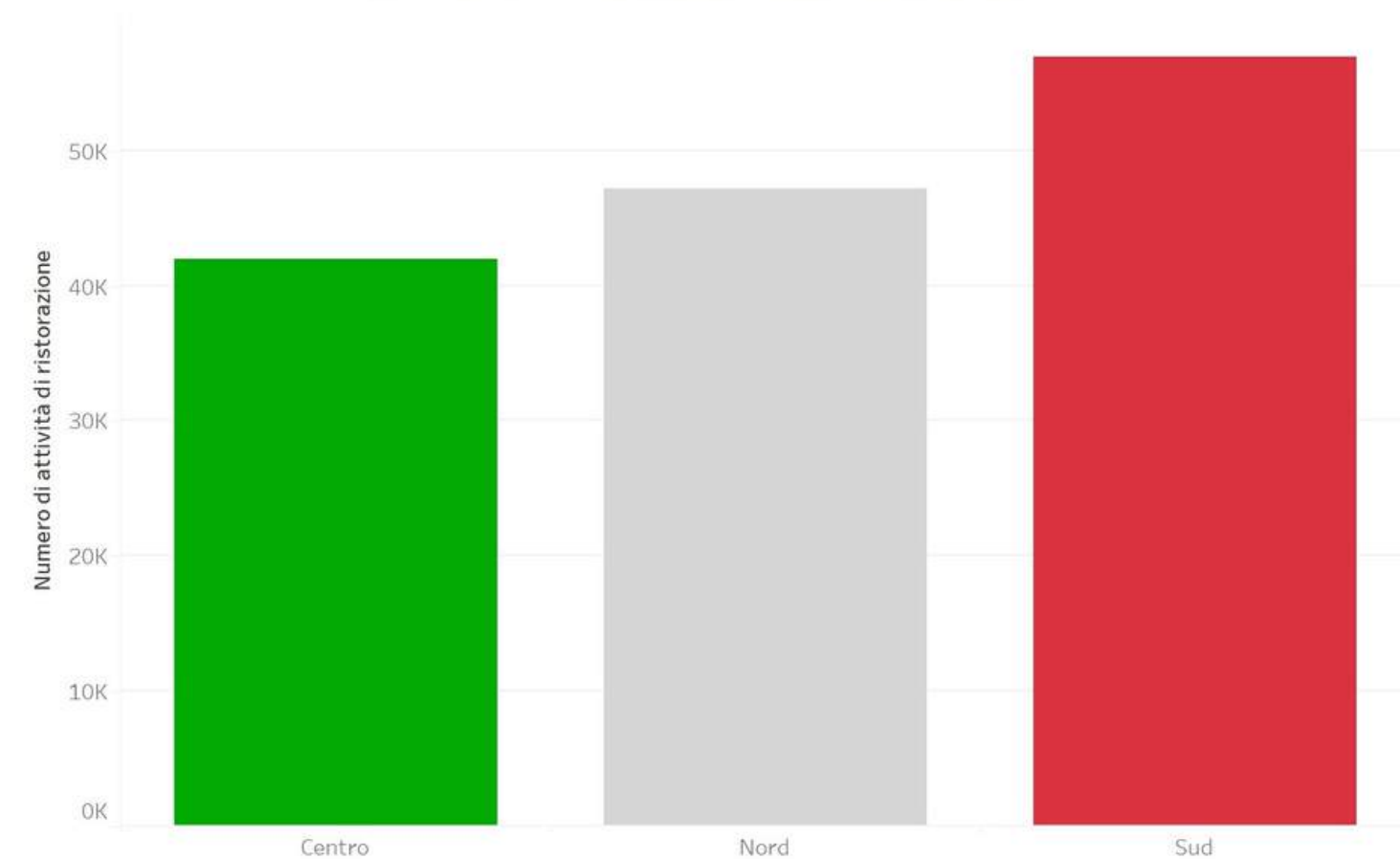
- Pulizia da attività di ristorazione con un numero di recensioni inferiore a 5 e conseguente rating non valido;
- Elisione di quei campi considerati irrilevanti ai fini della ricerca;
- Rimozione delle variabili ritenute ridondanti;
- Unione in un'unica colonna delle features aventi lo stesso contenuto informativo;
- Ridefinizione della variabile "priceLevel", in modo da renderla maggiormente comprensibile;
- Creazione della variabile "Comuni", al fine di aumentare il contenuto informativo;
- Ridefinizione dell'ordine dei campi del data frame e dell'indice di quest'ultimo.



Tripadvisor

(Dataset in seguito alla Data Preprocessing)
nr. 181.951 record e
nr. 16 colonne
vs
nr. 120 colonne
e **nr. 242.400 record**
(Dataset generato dal web scraping)

Distribuzione delle attività di ristorazione in Italia



In Italia la maggior quantità di attività di ristoro è rinvenibile nel Mezzogiorno (barra rossa), mentre la minor concentrazione è nel Centro Italia (barra verde).

Data Exploration



Istantanea di una Heat Map interattiva che consente di geolocalizzare le attività di ristorazione in Italia, permettendo di osservare agevolmente la concentrazione di tali attività e offre un maggior grado di dettaglio.



2. 50 TOP ITALY



50 TOP
ITALY



Le classifiche 50 TOP Italy sono delle guide gastronomiche pensate sia per gli italiani sia per i turisti amanti della buona tavola.

In particolare, tali classifiche sono affidate a 150 esperti gastronomi, ai quali viene richiesto di effettuare visite in anonimato in ogni ristorante e trattoria potenzialmente includibili nelle stesse.

Classifiche di interesse: 50 Top Italy ristoranti e 50 Top Italy trattorie
(2020-2021-2022)

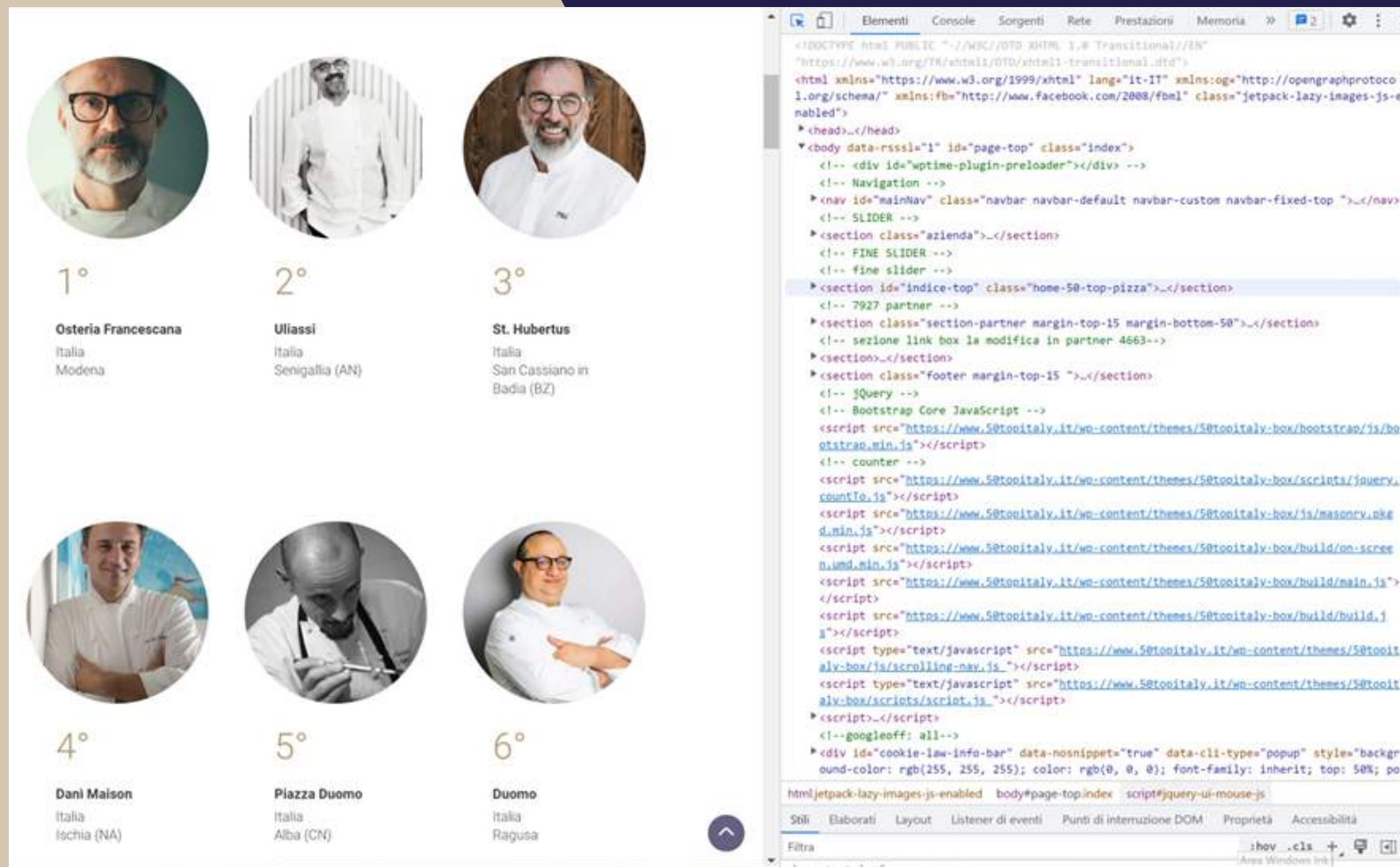
Web scraping



Data Preprocessing



Il processo di web scraping relativo alle diverse classifiche 50 Top Italy è stato effettuato utilizzando la libreria Selenium di Python.



Csv ristoranti e trattorie Classifiche Top 50 Italy (**2020**):

- Ridefinizione campo "città", separando le città dalle relative provincie e conseguente creazione di un nuovo campo, denominato "Provincia";
- Creazione del campo "Regione";
- Ridefinizione dell'ordine delle variabili e dell'indice, per entrambi i data frame.

Csv ristoranti e trattorie Classifiche Top 50 Italy (**2021-2022**):

- Riorganizzazione dei campi "Città" e "Regione";
- Creazione del campo "Nazione";
- Ridefinizione dell'ordine dei campi e dell'indice dei vari dataset.

In seguito, i dataset così ripuliti sono stati concatenati, in modo da formare un unico dataset relativo a tutti i ristoranti e trattorie presenti nelle classifiche 50 Top Italy, anno 2020, 2021 e 2022.

3. Stella Verde Michelin

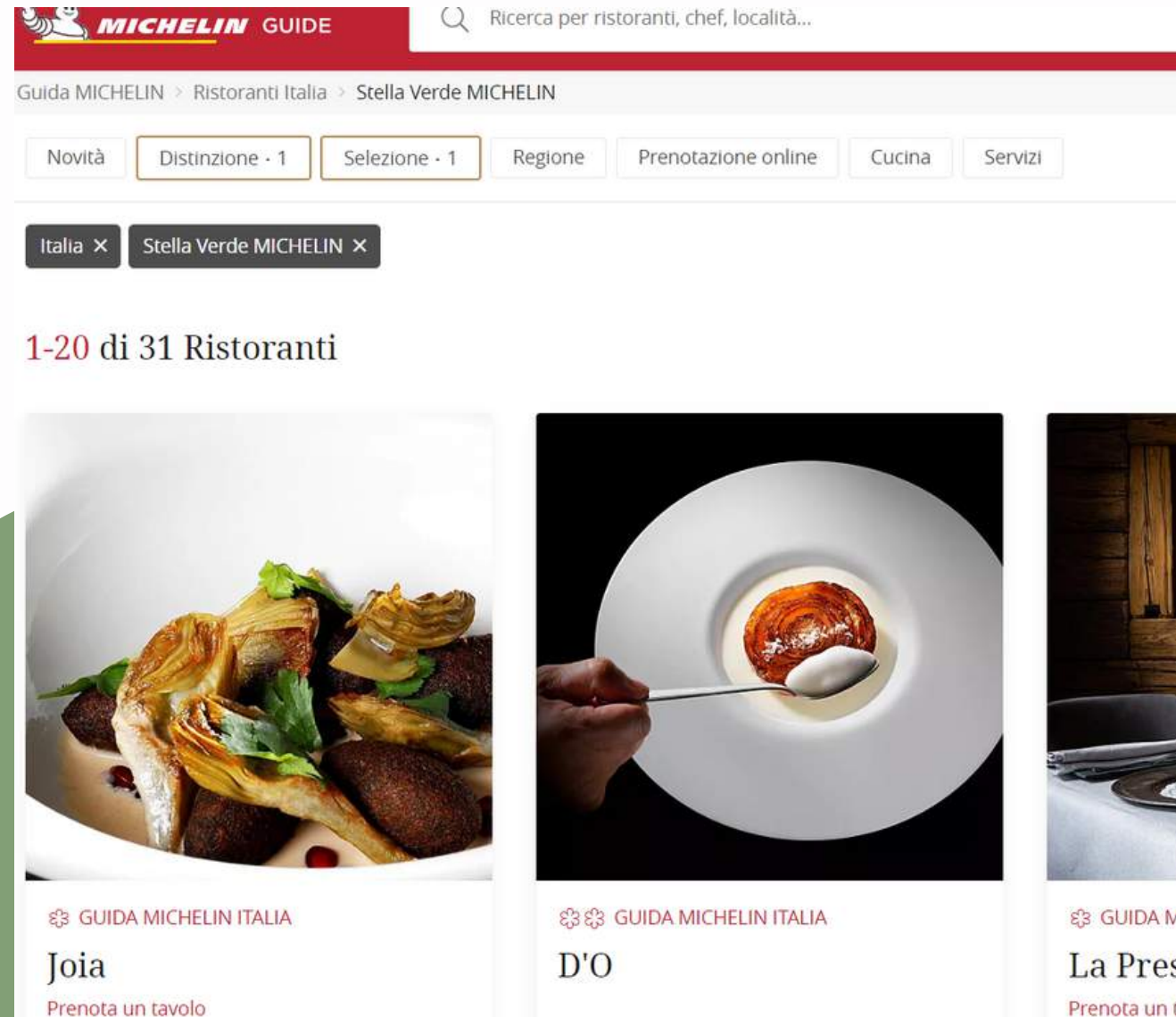
La Stella Verde Michelin è un riconoscimento assegnato dalla stessa Guida Michelin ai ristoranti e alle trattorie presenti nelle proprie selezioni, che si sono distinti per il loro impegno verso l'ambiente.

anno di interesse: 2022

Metodo, filosofia, stagionalità...



Web Scrapping



L'acquisizione dei dati relativi ai 31 ristoranti premiati è stata conseguita utilizzando la *libreria Selenium* di Python.
Il data frame così ottenuto contiene il nome di ogni ristorante e rispettivo comune di ubicazione.



Data Preprocessing

- Aggiunta del campo "Regione".

Quest'ultimo è stato ottenuto mediante tecniche di **web scraping**, associando al comune in cui sono collocati i ristoranti premiati, la rispettiva regione.

```
# Ottenimento della regione, a partire dal comune.  
reg="Null"  
location = "briosco"  
  
url = f'https://nominatim.openstreetmap.org/search?q={location}&format=json&accept=  
result = requests.get(url=url)  
result_json = result.json()  
  
string=result_json[0]["display_name"]  
print(string.split(','))  
  
for x in string.split(','):  
    if x in regioni:  
        reg=x  
print(reg)  
  
['Briosco', ' Monza e della Brianza', ' Lombardia', ' 20836', ' Italia']  
Lombardia
```


Data frame ottenuto dal web scraping

	Nome	Comune
0	Lanterna Verde	Villa di Chiavenna
1	Gardenia	Caluso
2	Casa Format	Orbassano
3	La Preséf	Mantello
4	D'O	Cornaredo
5	Joia	Milano
6	Terra	Sarentino
7	Antica Locanda al Cervo - Landgasthof zum Hirs...	San Genesio Atesino
8	1908	Soprabolzano
9	Lerchner's In Runggen	San Lorenzo di Sebato
10	St. Hubertus	San Cassiano
11	Agritur El Mas	Moena
12	SanBrite	Cortina d'Ampezzo
13	El Brite di Larieto	Cortina d'Ampezzo

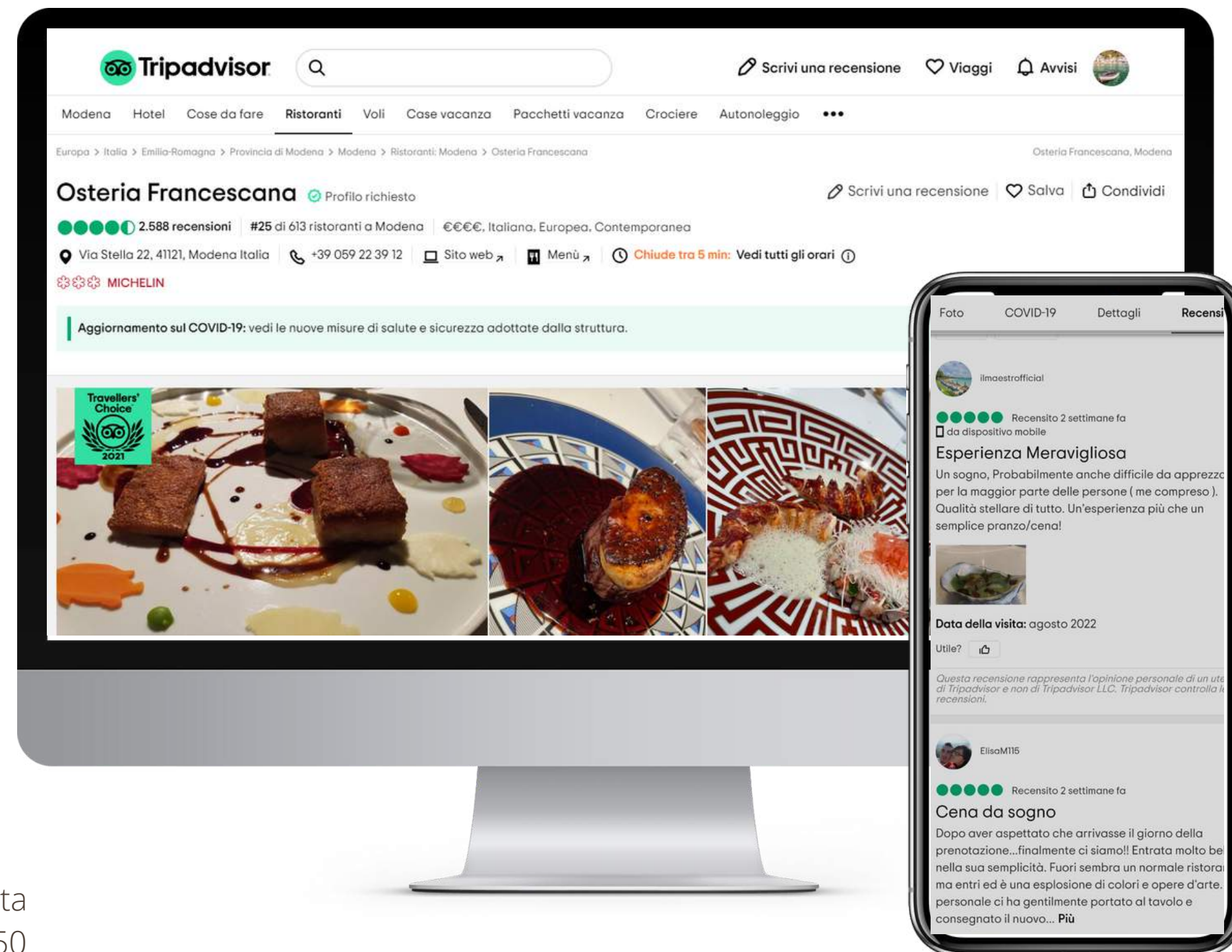
Data frame risultante dalle operazioni di Data Preprocessing

	Name	City	Region
0	Lanterna Verde	Villa di Chiavenna	Lombardia
1	Gardenia	Caluso	Piemonte
2	Casa Format	Orbassano	Piemonte
3	La Preséf	Mantello	Lombardia
4	D'O	Cornaredo	Lombardia
5	Joia	Milano	Lombardia
6	Terra	Sarentino	Trentino-Alto Adige
7	Antica Locanda al Cervo - Landgasthof zum Hirs...	San Genesio Atesino	Trentino-Alto Adige
8	1908	Soprabolzano	Trentino-Alto Adige
9	Lerchner's In Runggen	San Lorenzo di Sebato	Trentino-Alto Adige
10	St. Hubertus	San Cassiano	Trentino-Alto Adige
11	Agritur El Mas	Moena	Trentino-Alto Adige
12	SanBrite	Cortina d'Ampezzo	Veneto
13	El Brite di Larieto	Cortina d'Ampezzo	Veneto
14	Caffè La Crepa	Isola Dovarese	Lombardia
15	La Cru	Romagnano	Emilia-Romagna
16	Osteria Enoteca Gambrinus	San Polo di Piave	Veneto
17	Osteria Francescana	Modena	Emilia-Romagna
18	Venissa	Mazzorbo	Veneto
19	Lazzaro 1915	Pontelongo	Veneto
20	PS Ristorante	Cerreto Guidi	Toscana



Recensioni TripAdvisor di ristoranti e trattorie 50 Top Italy e Stella Verde Michelin.

La selezione delle recensioni redatte dagli utenti sul sito di TripAdvisor è stata effettuata sia per i ristoranti e le trattorie presenti nelle rispettive classifiche 50 Top Italy (2022), sia per quelli che hanno ricevuto l'assegnazione della Stella Verde Michelin nell'anno corrente.



La metodologia utilizzata per l'integrazione è stata la stessa per entrambe le fonti dato, 50 Top Italy e Stella Verde Michelin. Pertanto, viene spiegato il procedimento attuato soltanto per i ristoranti presenti nella classifica 50 Top Italy (2022).

Ricerca dei ristoranti presenti nella classifica 50 Top Italy (2022) all'interno del data frame relativo a tutta la ristorazione italiana indicata da TripAdvisor, mediante l'associazione del link TripAdvisor di ciascun ristorante in classifica.

In particolare, la corrispondenza si è effettuata per:

- nome del ristorante (match unico);
- regione (comune spesso non combacia);
- distanza di Lehvenstein (metodo alternativo).
 - sort by:
 - "lev"
 - "ratings"
 - numberReviews

```
# Prova di funzionamento
Levenshtein.ratio('Ristorante Cinzia Christian & Manuel', 'Christian & Manuel Ristorante')

0.5538461538461539
```



Creazione di un **dizionario** avente come:

- chiave* → il nome del ristorante
- valore* → il link TripAdvisor associato al ristorante indicato dalla chiave

```
{'Osteria Francescana': 'https://www.tripadvisor.it/Restaurant_Review-g187803-d1160010-Reviews-Osteria_Francescana-Modena_Province_of_Modena_Emia_Romagna.html',
'Uliassi': 'https://www.tripadvisor.it/Restaurant_Review-g194914-d1100666-Reviews-Ristorante_Uliassi-Senigallia_Province_of_Ancona_Marche.html',
'Ristorante Enrico Bartolini': 'https://www.tripadvisor.it/Restaurant_Review-g187849-d12432575-Reviews-Ristorante_Enrico_Bartolini-Milan_Lombardy.html',
'Dani Meison': 'https://www.tripadvisor.it/Restaurant_Review-g590114-d10376317-Reviews-Dani_Meison-Israel_Israel_Israel_Province_of_Naples_Campania.html'}
```



name	address	latitude	longitude	Regione	cuisine	dietaryRestrictions	rating	Certificati_Eccellenza	Pricelevel	email	numberOfReviews	phone	webUrl
Passion Como	Via Adamo del Pero 35, 22100, Como Italia	45.808900	9.083953	Lombardia	[Italiana, Cibo di strada, Per vegetariani, Op...	[Per vegetariani, Opzioni senza glutine]	5.0	[2021.0, 2020.0, 2019.0]	Bassa	NaN	242	+39 339 533 7419	https://www.tripadvisor.it/Restaurant_Review-g...
BLUE M Milano - Bottega archigiana	Via Lanzzone 27 Sant'Ambrogio - Università Catt...	45.460682	9.176967	Lombardia	[Italiana, Bar, Salutistica, Pub]	[]	5.0	[]	Bassa	bluemmilano@gmail.com	105	+39 02 9738 6209	https://www.tripadvisor.it/Restaurant_Review-g...
Marea Seafood & Beverage	Via Papa Gregorio XIV, 16, 20123 Milano Italia	45.459140	9.182726	Lombardia	[Italiana, Pesce, Contemporanea, Mediterranea,...	[]	5.0	[]	Medio-Alta	info@mareamilano.it	202	+39 02 8457 0315	https://www.tripadvisor.it/Restaurant_Review-g...
Pizzaut	Via Don Verderio, 1, 20051 Cassina de' Pecchi ...	45.518390	9.352924	Lombardia	[Italiana, Pizze, Mediterranea, Per vegetarian...	[Per vegetariani, Opzioni vegane]	5.0	[]	Medio-Alta	prenotazioni@pizzaut.it	238	NaN	https://www.tripadvisor.it/Restaurant_Review-g...
Cesarino	Via Pattari 2, 20122 Milano Italia	45.464493	9.193562	Lombardia	[Italiana, Salutistica, Cibo di strada, Per ve...	[Per vegetariani, Opzioni vegane]	5.0	[2021.0]	Bassa	NaN	130	+39 347 948 7039	https://www.tripadvisor.it/Restaurant_Review-g...



Recensioni (2.591)

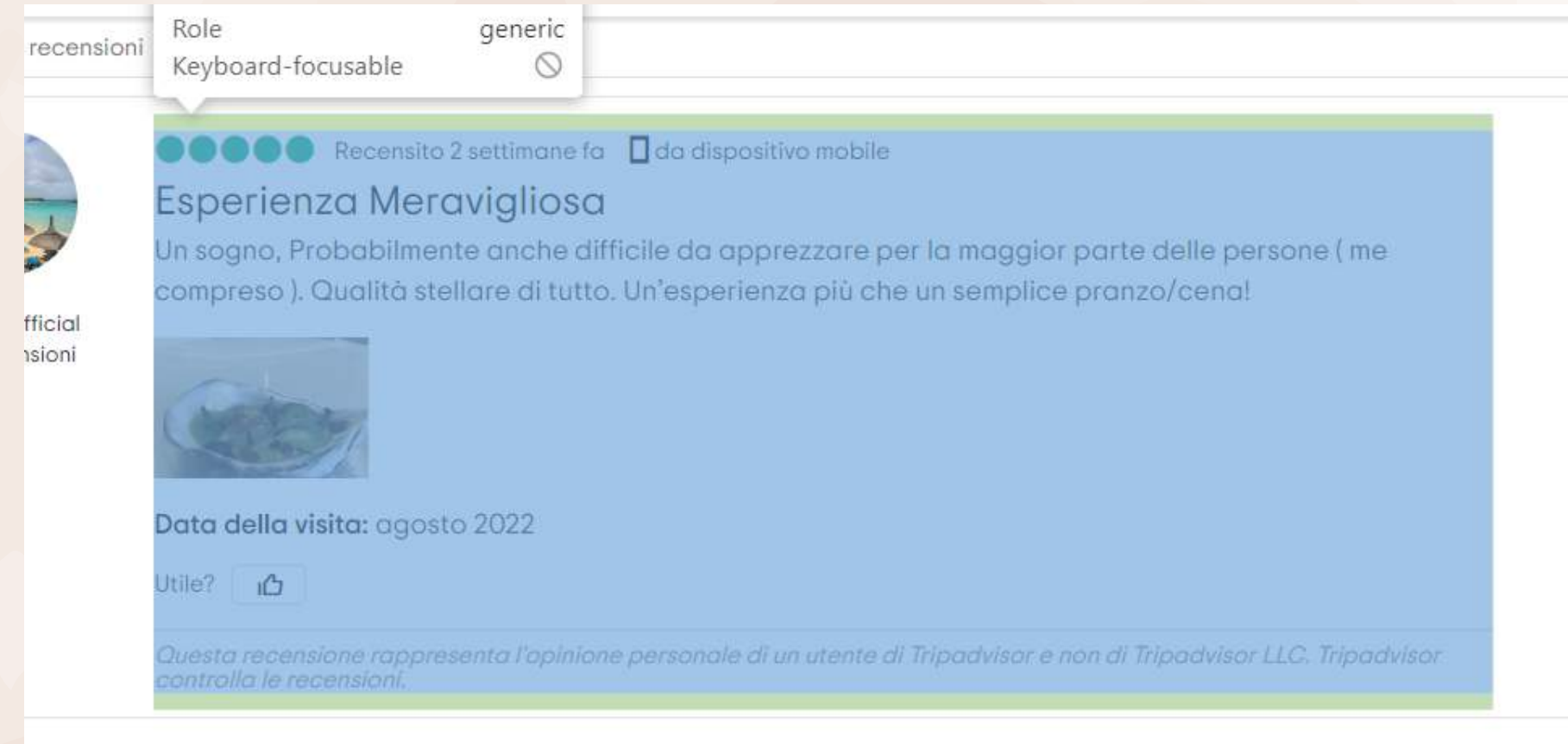
Controlliamo le recensioni ⓘ

Web scraping

Attuato utilizzando la libreria *Selenium* di Python, tenendo conto della presenza di 10 recensioni per pagina TripAdvisor che, una volta terminate, rendono necessaria una variazione automatica dell'url del sito, per cercare nella pagina successiva.

Tale operazione ha condotto all'ottenimento delle seguenti informazioni:

- Rating
- Titolo
- Testo
- Data recensione



```
# Cambiamento dell'url, corrispondente al cambiamento di pagina
url=url[:h] + "-or"+str(j)+"0"+ url[h:]
wd.get(url)

# Apertura del file per salvare le recensioni
csvFile = open(path_to_file, 'a', encoding="utf-8")
csvWriter = csv.writer(csvFile)

time.sleep(2)

wd.find_element("xpath", "//span[@class='taLnk ulBlueLinks']")#.click()

# Definizione del blocco recensioni per poi prendere ogni recensione singolarmente
container=wd.find_elements("xpath", ".//div[@class='review-container']")

for t in range(len(container)):
    title = container[t].find_element("xpath", ".//span[@class='noQuotes']").text
    date = container[t].find_element("xpath", ".//span[contains(@class, 'ratingDate')]").get_attribute("title")
    rating = container[t].find_element("xpath", ".//span[contains(@class, 'ui_bubble_rating bubble_')]").get_att
    review = container[t].find_element("xpath", ".//p[@class='partial_entry']").text.replace("\n", " ")

    csvWriter.writerow([date, rating, title, review,name,i])
# Cambiamento di pagina
```


	Date	Ratings	Title	Text	Name
0	2022-08-25	50	Applausi!	Un bel regalo per il mio compleanno. Cucina ch...	Casa Format
1	2022-06-18	40	Bel posto	Buon pranzo. Mangiato e bevuto bene. Servizio ...	Casa Format
2	2022-06-02	50	Una conferma al 360° !	Casa Format è sempre un grande certezza ! Cuci...	Casa Format
3	2022-05-08	30	Purtroppo non mi ha convinto	Posto incantevole, arredamento e atmosfera (av...	Casa Format
4	2022-04-06	50	un dolce anniversario	ci ritorniamo purtroppo solo dopo due anni ma ...	Casa Format
...
16458	2012-02-20	40	OTTIMO RISTORANTE	sono stato più volte al dattilo si mangia da D...	Dattilo
16459	2012-01-30	50	rilassante	Per una serata rilassante,piacevole,e all'inse...	Dattilo
16460	2011-11-04	20	da non tornarci	Sono stato invitato da colleghi i quali mi ave...	Dattilo
16461	2011-10-07	30	inferiore alle aspettative	Dopo aver tanto sentito parlare di questo rist...	Dattilo
16462	2011-09-03	50	Ottimo agriturismo	Il servizio è il cibo sono stato ottimi. L'atm...	Dattilo

16463 rows × 6 columns

	Date	Ratings	Title	Text	Name	Position
0	2022-09-05	50	Un bel mix	Nonostante ricordi e famiglia ci porti spessis...	Sora Maria e Arcangelo	0
1	2022-08-14	50	FlashBack	La sua fama la precede ... ristorante con la R g...	Sora Maria e Arcangelo	0
2	2022-08-08	50	Cena strepitosa	Sono stato qualche giorno fà a cena in questo ...	Sora Maria e Arcangelo	0
3	2022-08-07	50	Qui il tempo non si ferma. Sempre unici e bont...	Come sempre Cibo buonissimo di altissima quali...	Sora Maria e Arcangelo	0
4	2022-07-10	50	Bella esperienza	Sicuramente un ambiente caldo e accogliente co...	Sora Maria e Arcangelo	0
...
33652	2015-03-11	50	La purezza della tradizione	In un angolo splendido del centro storico di P...	Botteghe Antiche	49
33653	2015-03-09	30	Buono!!!!!!!!! Ma con qualche errore.	Andati a pranzo io, mia moglie mio figlio. Dov...	Botteghe Antiche	49
33654	2015-03-08	50	Che dire... Ottimo.	Bellissima serata. Tutto molto buono e accoppi...	Botteghe Antiche	49
33655	2015-03-07	50	Semplice, deliziosa, tradizione	Eravamo curiosi di provare questo ristorante, ...	Botteghe Antiche	49
33656	2015-03-06	40	che sorpresa!	con un collega ci sono andato per una pausa pr...	Botteghe Antiche	49

33657 rows × 6 columns

	Date	Ratings	Title	Text	Name	Position
0	2022-08-25	10	Delusione totale	Onestamente se questo è' stato il miglior rist...	Osteria Francescana	1
1	2022-08-01	50	Eccellente	Finalmente siamo riuscite a prenotare presso l...	Osteria Francescana	1
2	2022-08-01	50	Un viaggio straordinario	Era da tempo che desideravo provare questa esp...	Osteria Francescana	1
3	2022-07-12	50	Perfezione	Perfezione - L'Osteria Francescana non ha biso...	Osteria Francescana	1
4	2022-06-28	20	Grandissima delusione	Abbiamo atteso così tanti mesi per poter cenar...	Osteria Francescana	1
...
38508	2015-08-13	50	Superbo ristorante situato vicino al Castello	Questo ristorante è un must se visitate Bari. ...	Pashà Ristorante	50
38509	2015-08-05	50	QUESTO È da asporto	Non avrei mai trovato questo posto se i miei a...	Pashà Ristorante	50
38510	2015-05-10	40	Trattoria - non il ristorante	Pasha dispone di un ristorante elegante e una ...	Pashà Ristorante	50
38511	2015-04-25	40	Eccellente	Siamo stati accolti calorosamente il restaurant...	Pashà Ristorante	50
38512	2015-04-24	50	PERFEZIONE!!!!	Non c'è niente di meglio in TUTTE LE CATEGORIE...	Pashà Ristorante	50

38513 rows × 6 columns

Data frame relativi alle recensioni TripAdvisor di ristoranti e trattorie 50 Top Italy e Stella Verde Michelin.

50 Top Italy 2022 trattorie

50 Top Italy 2022 ristoranti

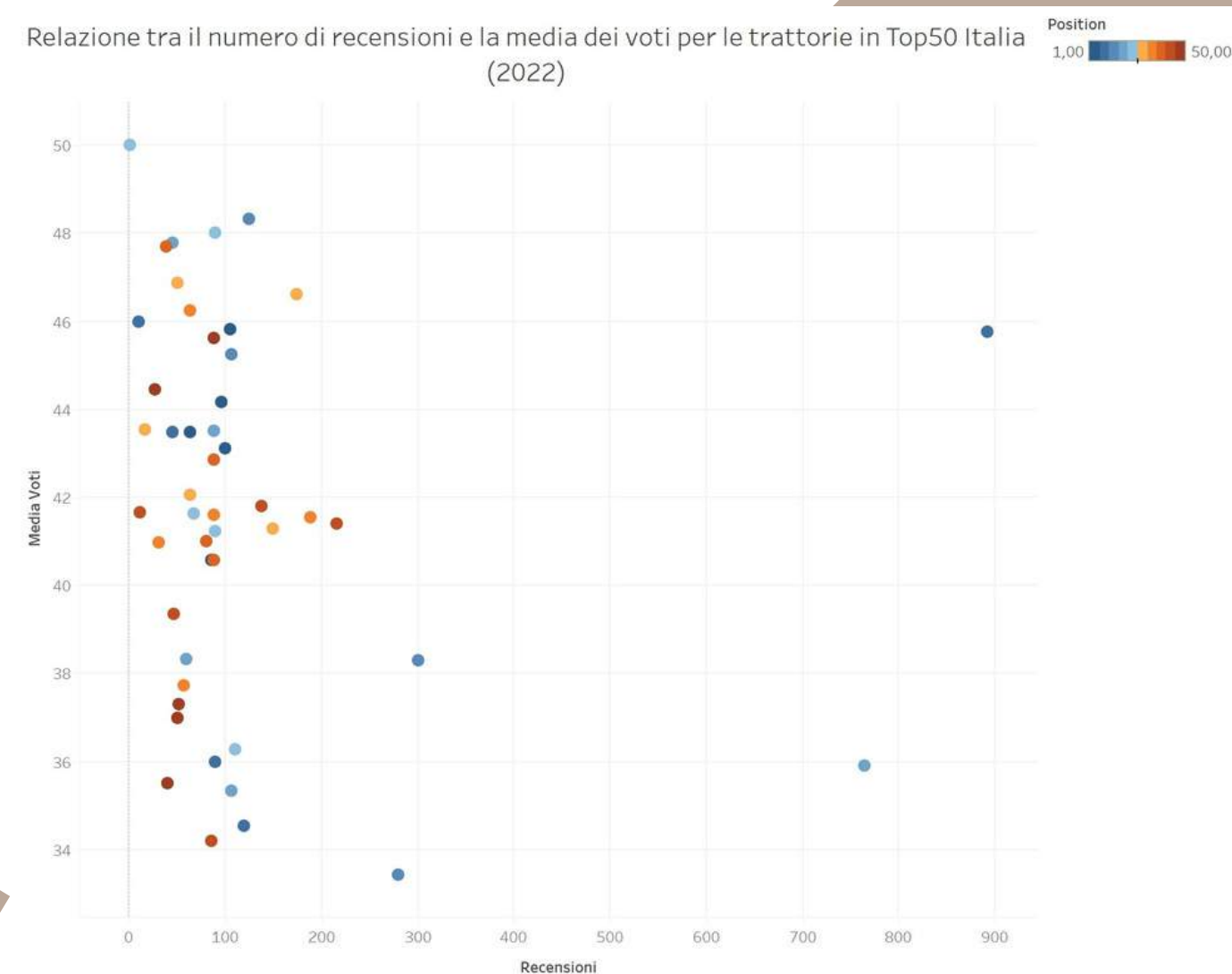
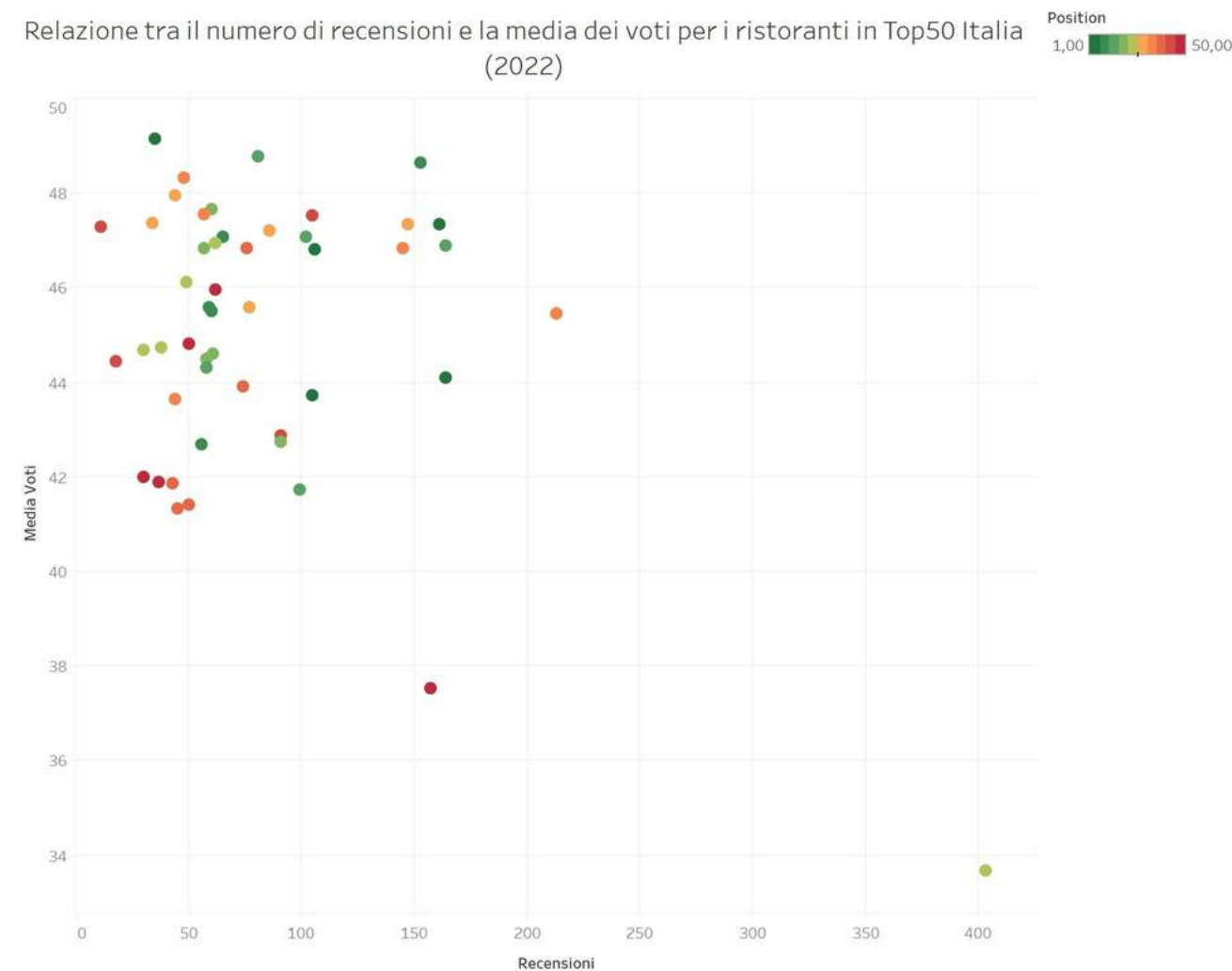


Data Exploration

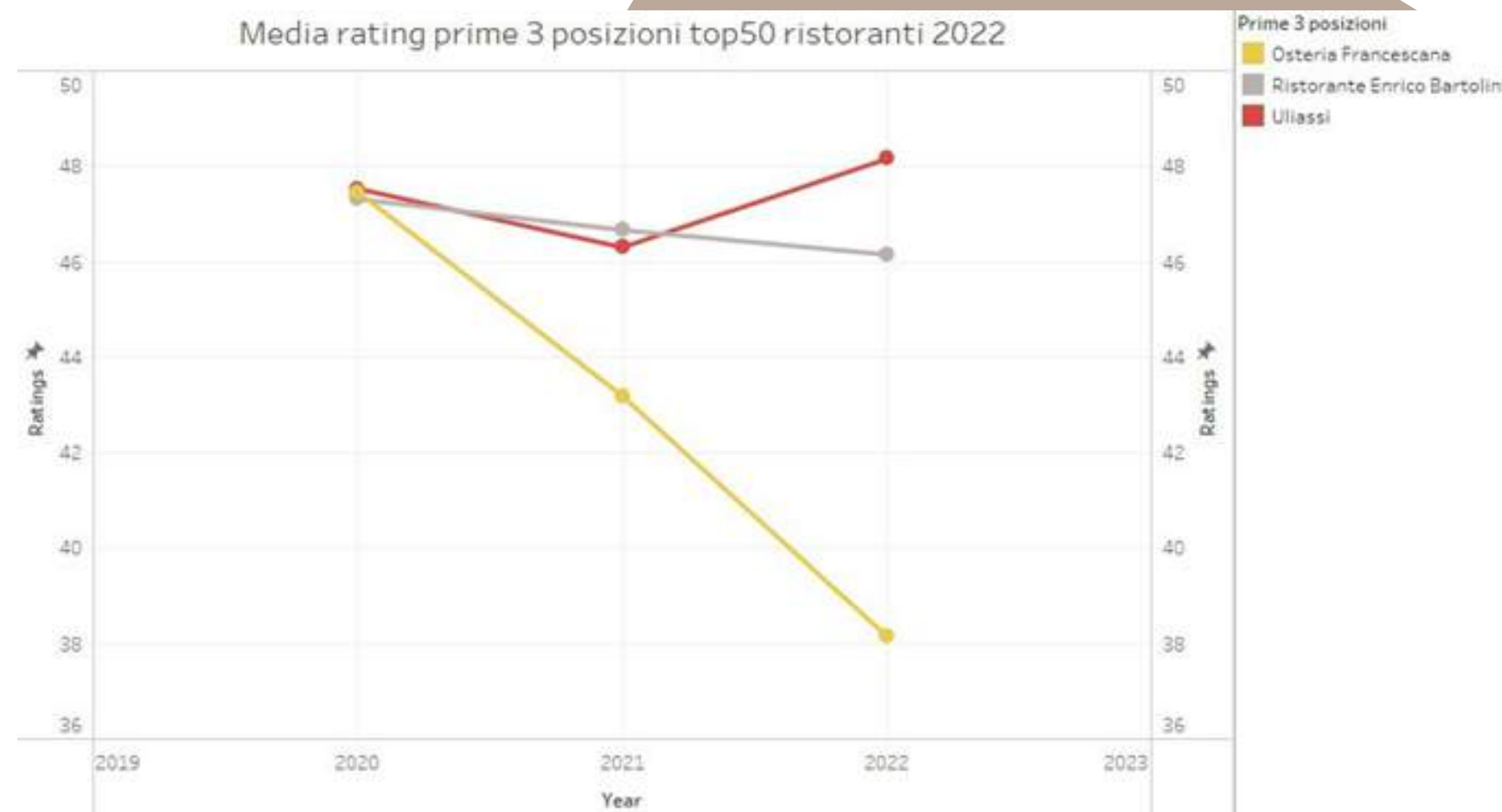
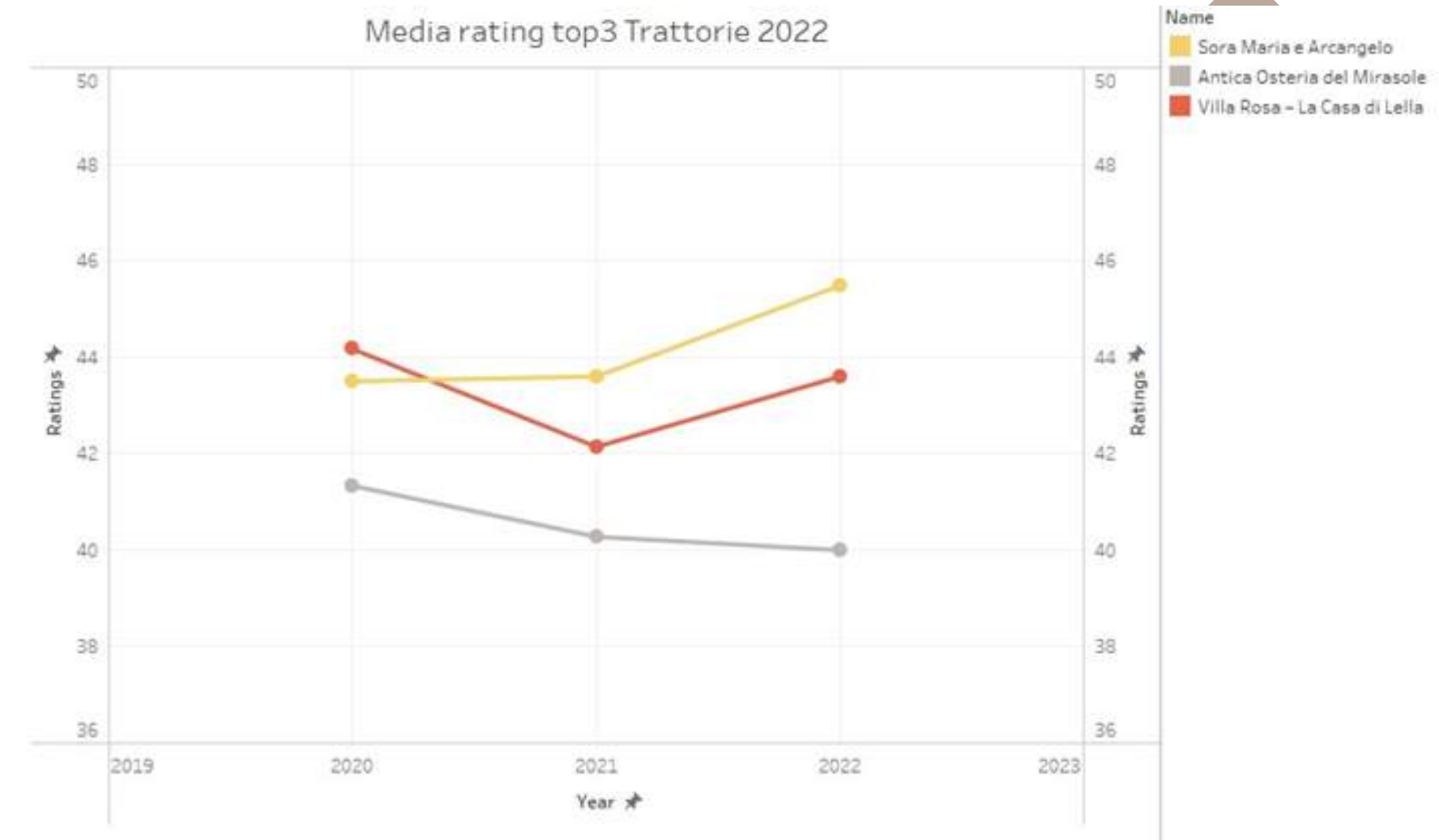
50 Top Italy



Effettuata considerando solo le recensioni TripAdvisor dal 2020 all'anno corrente.



Gli Scatter Plot mostrano che non è rinvenibile un legame lineare tra il rating medio e il numero delle recensioni e dimostrano come la classifica stilata dagli esperti gastronomi sia ininfluente per gli utenti di recensioni di TripAdvisor.



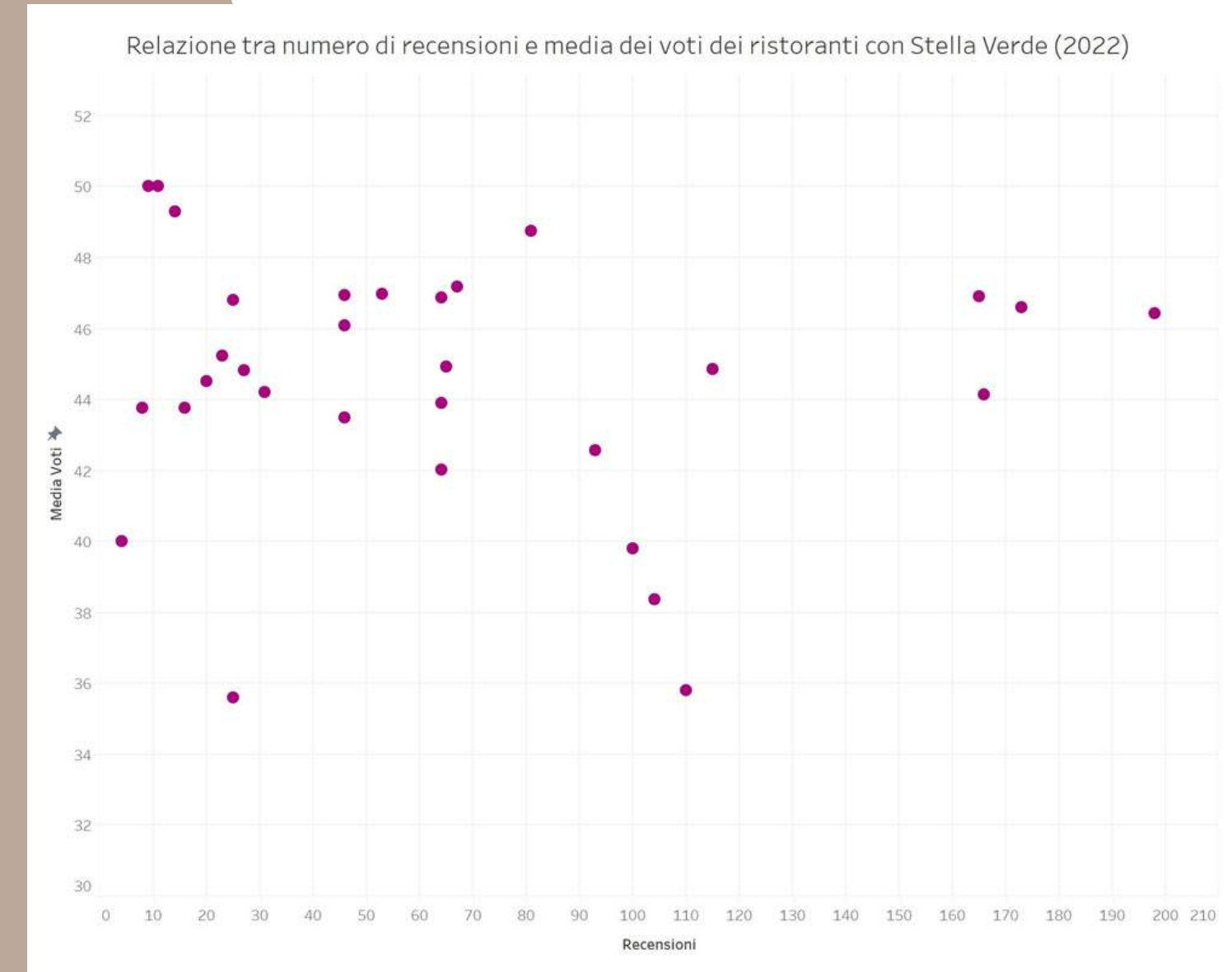
I Line Plot consentono di asserire nuovamente che l'opinione dei critici non è pienamente condivisa dai clienti.

Tale affermazione è particolarmente evidente in relazione al caso dell'Osteria Franceseana di Modena, la quale è posizionata nella prima posizione della classifica 50 Top Italy 2022, nonostante il calo considerevole del suo rating annuale dal 2020 ad oggi.

Data Exploration Stella Verde Michelin




Scatter Plot che sebbene consenta di indagare la presenza o meno di un legame tra il numero di recensioni e la media dei voti, per la sua caratteristica intrinseca di non collocare i ristoranti e le trattorie in una classifica, consente solamente di notare anche in questo caso assenza di legame lineare tra le variabili messe in relazione.



Data Quality






Le dimensioni di qualità trattate sono prevalentemente:

completezza

consistenza

accuratezza

La valutazione di qualità è stata attuata sui dataset interessati da un successivo caricamento nel database prescelto, quindi:

- dataset relativo alle attività di ristorazione italiane presenti su TripAdvisor;
 - quelli relativi ai ristoranti e alle trattorie presenti nelle classifiche 50 Top Italy 2022 e relative recensioni;
 - quello inerente ai ristoranti che nel 2022 hanno ricevuto la Stella Verde Michelin e relative recensioni.
- 

Per ogni dataset valutato:



**Verifica ed eventuale eliminazione di
valori nulli**

**Controllo ed eliminazione di
duplicati**


```
id 0.000000
name 0.000000
address 0.000000
latitude 0.574904
longitude 0.574904
Regione 0.000000
cuisine 0.000000
dietaryRestrictions 0.000000
rating 0.000000
Certificati_Eccellenza 0.000000
PriceLevel 10.490589
email 44.250112
numberOfReviews 0.000000
phone 6.009386
webUrl 0.000000
website 20.607885
dtype: float64
```

I valori mancanti del dataset relativo alle attività di ristorazione italiane presenti su TripAdvisor sono circa:

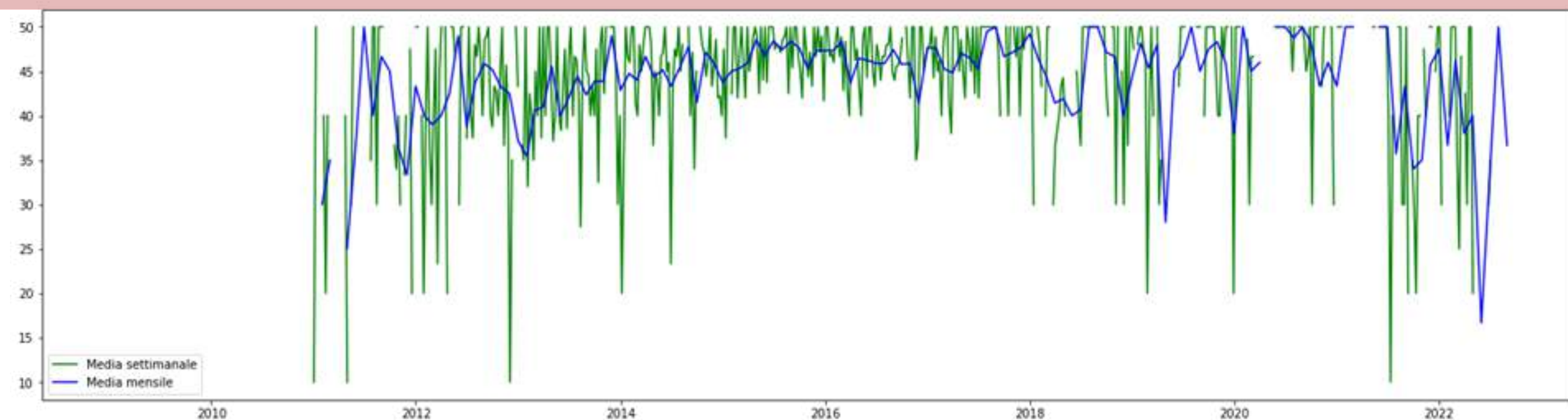
- il 44.37% nel campo "email";
- il 20.61% nella variabile "website";
- il 10.49% nel campo "PriceLevel";
- il 6% nella feature "phone".

Inoltre per il suddetto dataset, si è indagata la presenza o meno di tutti i possibili casi di dominio, per alcune variabili.

Per esempio, si è notato che il campo "Price Level" assume solo 3 dei 6 possibili valori.



Spostando il focus sui data frame relativi alle recensioni di ristoranti e trattorie presenti in 50 Top Italy 2022 e in Stella Verde Michelin 2022, si è analizzata la buona riuscita dell'acquisizione e l'esistenza del testo delle recensioni.



Attuando delle analisi specifiche, si è potuta osservare la presenza o meno di archi temporali in cui tali opinioni risultano mancanti. L'esempio mostrato in figura permette di evidenziare che nel periodo interessato dal Covid 19 non è stata elaborata alcuna recensione per il ristorante analizzato.



DATABASE





I csv relativi alle recensioni e quello inerente alle attività di ristorazione indicate da TripAdvisor sono stati memorizzati nel database MongoDB denominato “DataMan”, previa conversione in formato JSON.

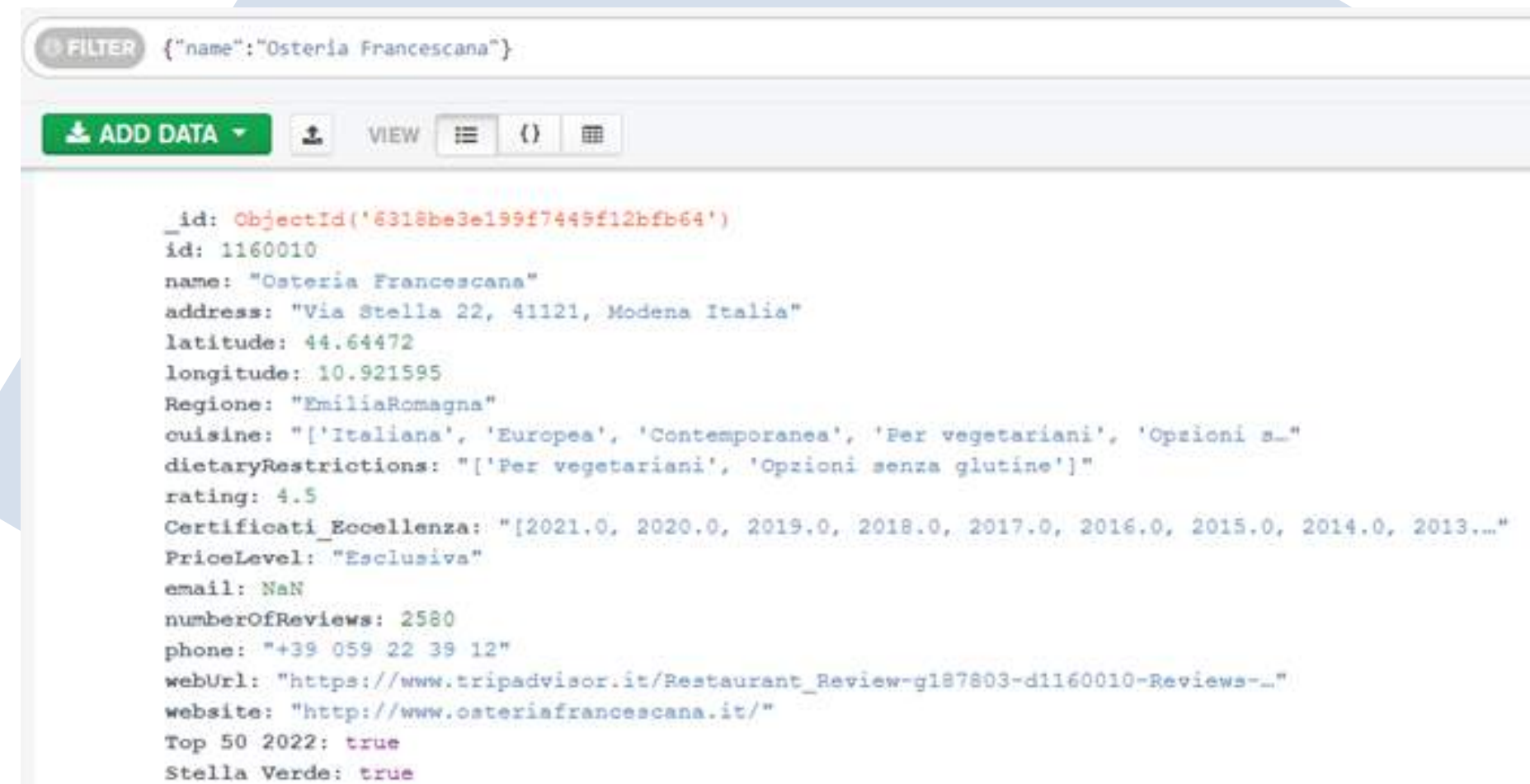
HOST localhost:27017	CLUSTER Standalone	EDITION MongoDB 6.0.1 Community	My Queries	Databases	Filter your data	DataMan	+	-
Osterie_reviews								
Storage size: 46.68 MB	Documents: 182 K	Avg. document size: 640.00 B	Indexes: 1	Total index size: 1.81 MB				
Top50_reviews								
Storage size: 8.59 MB	Documents: 39 K	Avg. document size: 403.00 B	Indexes: 1	Total index size: 409.60 kB				
Osterie_reviews								
Storage size: 7.00 MB	Documents: 34 K	Avg. document size: 383.00 B	Indexes: 1	Total index size: 352.26 kB				
Stella_Verde_reviews								
Storage size: 3.37 MB	Documents: 16 K	Avg. document size: 412.00 B	Indexes: 1	Total index size: 167.94 kB				
Top50_ranking								
Storage size: 28.67 kB	Documents: 300	Avg. document size: 185.00 B	Indexes: 1	Total index size: 20.48 kB				

MongoDB è stato utilizzato in locale e l’interazione con lo stesso è avvenuta servendosi della libreria Python denominata *Pymongo*.



Integrazione delle informazioni possedute, creando una variabile dummy nel dataset relativo alle attività di ristoro presenti su TripAdvisor. Tale variabile indica l'assegnazione o meno di ognuna attività alla classifica 50 Top Italy 2022 e della Stella Verde Michelin.

Ad esempio, è possibile notare che Osteria Francescana ha ottenuto sia la Stella Verde Michelin, sia una posizione in classifica 50 Top Italy 2022.



Interrogazione database



Query effettuate:

1. quali ristoranti hanno ottenuto rispettivamente i singoli riconoscimenti nell'anno 2022 (ricoprire una posizione nella classifica 50 Top Italy o assegnazione della Stella Verde Michelin);
2. ristoranti ai quali hanno contemporaneamente attribuito sia un collocamento in classifica, sia il conferimento della Stella Verde Michelin.

L'interrogazione del database mostra che sono sette i ristoranti dotati di entrambi i riconoscimenti.

Risultato della seconda query trasformato in un file csv

Stella Verde	OsterieTop 50 2022	Top 50 2022	name	address
true		true	D'O	Piazza della Chiesa, 14, 20010 Cornaredo
true	true		La crepa	Piazza Giacomo Matteotti 13, 26031 Isola Dovarese
true		true	Joia	Via Panfilo Castaldi 18, 20124 Milano
true		true	Ristorante Don Alfonso 1890	11 Corso Sant' Agata, 80061 Massa Lubrense
true		true	Osteria Francescana	Via Stella 22, 41121, Modena
true	true		Osteria di Suvereto da l'Ciocio	Piazza Dei Giudici 1, 57028 Suvereto
true		true	Ristorante Gourmet Virtuoso - Tenuta Le Tre Virtù	Vai di Lucigliano 13, 50038, Scarperia e San Piero

CONCLUSIONI

Obiettivi preposti in fase iniziale sono stati efficacemente soddisfatti !

Sviluppi futuri:

Algoritmi di text analysis ("text summarization", "sentiment analysis" o "topic modeling") ;

Influenza delle recensioni sulla pubblicità di un determinato locale;

Wordcloud o network graph interattivi.

Rispetto per l'ambiente !



Grazie