CS 520 - Data Integration, Warehousing and Provenance

Data Curation Project - Yelp Dataset

Group 12 - Fateen | Sahil | Himanshu

Topics to be Covered

Overview and Data Intro	Data Curation and Analysis	Conclusion

Understanding the Dataset

Domain

The dataset represents user-generated reviews and business details from the Yelp platform, providing a comprehensive insight into local businesses, primarily in the North American region.

Resources

The dataset was available on the Kaggle's platform via the following link: [Kaggle Yelp Dataset](https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data).

Dataset Characteristics

- Dataset Size: 10,000 Records
- Attributes: Diverse attributes such as `address`, `categories`, `city`, `latitude`, `longitude`, `name`, `postal_code`, `review_count`, `stars`, among others.
- Data Format: Data was in JSON format

Understanding the Dataset

```
LOAD DATASET Business_Data AS csv FROM yelp_academic_dataset_business_project.csv @ artifact file 41
```

Console Timing Datasets Charts

business_data (10000 rows)

Views

	business_id (string)	name (string)	address (string)	city (string)	state (string)	postal_code (int)	latitude (float)	
0	TacYUYhU3HpLHF9Rs6fW2w	Steps to Learning Montessori Preschool	6901 Phelps Rd	Goleta	CA	93117	34.423309326171875	-1
1	RnExalCvleXxFpbIKEqJsQ	Breeze Blow Dry Bar	9916 Clayton Rd	St. Louis	MO	63124	38.636714935302734	-9
2	pjtjBeZC3gvmtlilQt-DFA	Impact Guns	11655 W Executive Dr	Boise	ID	83713	43.608699798583984	-1
3	OMI5pTUBVUjW_jvDKLvYtw	Palms Primary Care	1615 Pasadena Ave S, Ste 430	Saint Petersburg	FL	33707	27.752653121948242	-8
4	MO-LTDfO843xaRtW0bx6jQ	J&Q Nails	9655 E US Hwy 36, Unit H	Avon	IN	46123	39.763057708740234	-8
5	jFq8QSWDwtAWMA1FaNvRhw	Candy Barrel	735 Dodecanese Blvd	Tarpon Springs	FL	34689	28.155075073242188	-8
6	t53MqkLTtWxJ7jMSqXYz-g	Luminosity	690 W Dekalb Pike	King of Prussia	PA	19406	40.08949661254883	-7
7	dY6rzL7Gw1U5afOskTlMmg	Nail Care Salon	12337 Olive Blvd	Creve Coeur	MO	63141	38.67354202270508	-9
8	VIJ4wKPf2TmKbaOTKYHROg	Architectural Antiques of Indianapolis	5000 W 96th St	Indianapolis	IN	46268	39.92622375488281	-8
9	IuKCyfSY7AKhRbRA1JPlPw	Aster's Floral Shop	41 Haddon Ave	Collingswood	NJ	8108	39.91556167602539	-7
10	6XOn1p3sbO22UjJGpmCgxg	China Wok	4319 Telegraph Rd	Saint Louis	MO	63129	38.486759185791016	-9
11	c8d8h47cogM_B_ZMC-h3zg	Brandon Family Medical Care	1218 Millennium Pkwy	Brandon	FL	33511	27.929439544677734	-8
12	YEwmI50bso_LYtuivsdwiA	7-Eleven	13151 Race Track Rd	Tampa	FL	33626	28.07027244567871	-8
13	6IG4SysBKyRnFW_e22q13A	Uber		Philadelphia	PA	19107	39.955928802490234	-7
14	bNBi-RVlx71bugXY0GRLtQ	Chestnut St. Cafe	4403 Chestnut St	Philadelphia	PA	19104	39.95672607421875	-7
15	rbPK4jSyFS10zhWYvo_Srg	Cafe Porche and snowbar	1625 Baronne St	New Orleans	LA	70113	29.939315795898438	-9
16	lrO5vOwa7gE0qAo0UQJYIA	Eyeglass World	13002 Seminole Blvd, Ste 10-11	Largo	FL	33778	27.891815185546875	-8
17	thpDDcdKLPzSFldLLBULLA	Callahan's Corner	914 Edwardsville Rd	Troy	IL	62294	38.732933044433594	-8
18	-EyRrBY1d-EQZrTlXWH4BQ	Spa Guy Dave		Pennsauken	NJ	8109	39.967105865478516	-7
19	EkthrfcRWCVYy-NuvfNmPg	FroYo Frozen Yogurt	4663 Maryland Ave	Saint Louis	MO	63108	38.64484405517578	-9

```
LOAD DATASET Reviews_Data AS csv FROM yelp_academic_dataset_checkin.csv @ artifact file 70
```

Console Timing Datasets Charts

reviews_data (28224 rows)

Views

	review_id (string)	user_id (string)	business_id (string)	stars (long)	useful (long)	funny (long)
0	FCXselWrkSUzC5wHZxVXrw	NDkwKnvjhBbjCh1cNlBoAw	t1qF12NdW8KvCqxqbvy-Hg	4	1	0
1	jTGp3mbMA8w_Prg-Ufxfkg	YAlp90pskaKL1_Wtn7kbKQ	kqAa2CtPGA-QsZhhbzpzUQ	5	0	0
2	Great to find a restaurant close by that's both healthy and affordable."	1454131476000				
3	RxJw4NR37bcVKQj_BoG7-w	pzfZBcILOqf5wMj0O5yLlg	MaYb7qMN6BomP1zQGj3Wjg	5	1	0
4						
5	Please don't go ""big-timer"" on us	Pi; you're far too awesome for that."	1294720909000			
6	zPt6CuKvx1v24BXMCwHzRA	p2PwR7oPLDo_vbo8gCCa7A	pSmOH4a3HNNpYM82J5ycLA	5	0	0
7	6g39Ku2yBsC4-l9N3prEaA	YsqK0URRY3oRwC3wTRtd7Q	h4C7s_Go9UKo5twaLIWc8A	5	0	0
8	I will be back!"	1488733986000				
9	qY11QD0JVNNuB9ObWDzBsQ	Hwzn3_MuTD_ZhY5-pjURUQ	8eDkw7CE0NKqMknPlu26fw	3	0	0
10	NbvR8lib1vdbXDMeXbhnVA	fEtWwhNKSoTqIWTSEwGlvg	KVFZ6yiM1DgiEGK-Jn80ig	2	0	0
11	UbhAJpJvtt4jaOv-dpSAyg	kMT7Hb8zRubKuissGbjcfw	0sr1EyOc6Td1C-962QW88w	1	3	1
12	vuS2mktaoBZo77XdLljjNA	kmXNP537i0dyYOFF-sxtdg	rG0UTjvbmVVsh9-kGzeliQ	4	0	0
13	Rt-BrvgR8k_mbmazrPoXbw	6cUt5rA5EY8-mH4q-v239Q	yeHLiKNp0hyR-ig4M6us-w	5	2	1
14	AGgFJ8VPX7kAwdy2XTh95w	3MYdpmHeNwC6FquRWi3YOg	uEWsfftrJ7ukPv1xyMVcrg	5	6	0
15	No3OGeXLydmnOYqG3FguUw	Wl30_VevEQt4lPbhqtNhrA	3SHw4aZW_muE1kRSLQQjjA	3	1	0
16	d7qA-vF3ICEYg0pzxP33sg	mleNRnJ7fdEEpUp8nvTSrw	jgcQGltZISMAwzo10XOjOw	5	2	1
17	ij3w_OPSeHZflospE4jLNQ	XSFkFLUC9dFvvyJPJHhg_A	2oav5QoWgnvTl2gO5xFMjw	3	1	0
18	xtoFKDQVvrQv6LRnmJF9Nw	6cFqRc7XOZlrQJ2f0pWvDw	vsrryxBR2Jykl71qa1H6SQ	4	1	0
19						

Understanding the Dataset

Attributes - Business Dataset

Attribute	Description
business_id	*Unique identifier for each business*
name	*Name of the business*
address	*Street address of the business*
city	*City where the business is located*
state	*State where the business is located*
postal_code	*Postal or ZIP code of the business location*
latitude	*Geographic latitude of the business*
longitude	*Geographic longitude of the business*
stars	*Average star rating of the business (e.g., 1 to 5 scale)*
review_count	*Number of reviews received by the business*
is_open	*Indicator if the business is currently open or closed*
attributes	*Various attributes of the business (e.g., WiFi, parking)*
categories	*Types or categories the business falls into*
hours	*Operating hours of the business*

Reviews Dataset

Attribute	Description
review_id	*Unique identifier for each review*
user_id	*Identifier for the user who posted the review*
business_id	*Unique identifier for each business reviewed*
stars	*Star rating given by the user (e.g., 1 to 5 scale)*
date	*Date when the review was posted*
text	*Text content of the review*
useful	*Number of users who found the review useful*
funny	*Number of users who found the review funny*
cool	*Number of users who found the review cool*

Preliminary Observations - Business Dataset

Attributes Data Types

```
The total number of rows are -  10000
The total number of features are -  14
The datatype of various features is -
business_id      object
name             object
address          object
city             object
state            object
postal_code     float64
latitude        float32
longitude       float32
stars           float32
review_count      int16
is_open            bool
attributes       object
categories       object
```

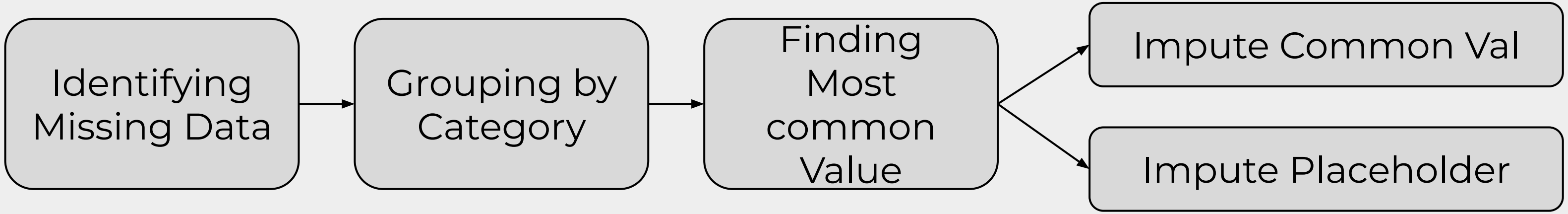
Null Values

```
Checking the Null/Missing values Feature-w
business_id       0
name              0
address         349
city              0
state             0
postal_code     372
latitude          0
longitude         0
stars             0
review_count      0
is_open           0
attributes      964
categories        4
hours           649
dtype: int64
```

Data Cleaning Strategies - Business Data

Handling the Missing Values for Column 'Attribute and Hours'

For each category c in our DataFrame, we found the most common value v in attributes and hours.
For all rows r where r.categories = c and r.attribute (or r.hours) is null, we assigned v to r.attribute (or r.hours). If no common value is found, assign 'Not specified'.

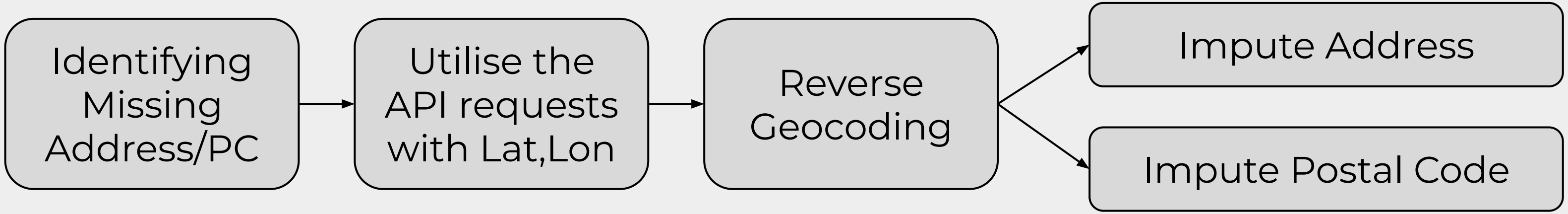


Different business categories might have distinct common attributes and operating hours. Grouping by category allows for a more accurate and contextually relevant imputation.

Data Cleaning Strategies - Business Data

Handling the Missing Values for Column 'Address and Postal Code'

For each row r in our DataFrame, if r.address or r.postal_code is null, use the OpenStreetMap Nominatim API to perform reverse geocoding based on r.latitude and r.longitude. Assign the obtained address to r.address and the postal code to r.postal_code.



For each identified row with missing data, send a request to the Nominatim API using the row's latitude and longitude to obtain the corresponding address and postal code.

Data Cleaning Strategies - Business Data

Ensuring Consistency and Dropping Missing 'Categories'

Given the low number of missing values (only 4) for the 'categories' column in the dataset, it is decided to drop these rows to maintain data integrity.

For all rows in our DataFrame that have a postal_code, ensure that the data type of postal_code is a string.

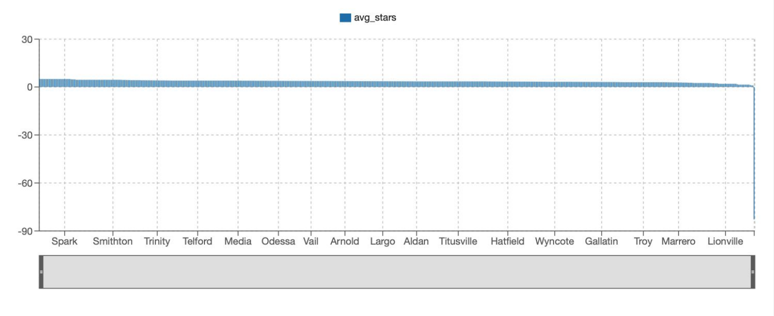
Saving as New Dataset

```
Remaining null values in each column:
business_id     0
name            0
address         0
city            0
state           0
postal_code     0
latitude        0
longitude       0
stars           0
review_count    0
is_open         0
attributes      0
categories      0
hours           0
dtype: int64

Number of rows assigned with new address: 348
Number of rows assigned with new postal code: 371
```

Analysis on Business Data

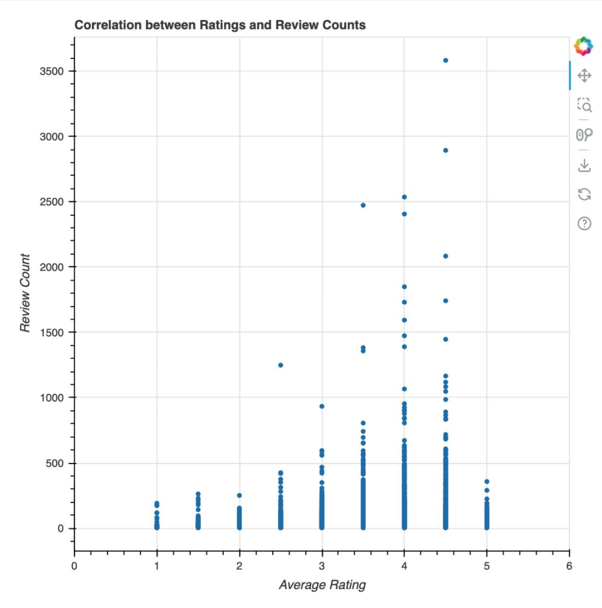
Average Star Ratings



Category Based Businesses

	categories (string)	city (string)	business_count
0	'BusinessAcceptsCreditCards': 'True'	Philadelphia	69
1	"'BusinessParking': """"""""{'garage': False"""	Philadelphia	65
2	'BusinessAcceptsCreditCards': 'True'	Tampa	61
3	'street': True	Philadelphia	59
4	'BusinessAcceptsCreditCards': 'True'	Indianapolis	53
5	'street': False	Philadelphia	49
6	'RestaurantsPriceRange2': '2'	Philadelphia	45
7	'street': False	Tucson	42
8	"'BusinessParking': """"""""{'garage': False"""	Tucson	41

Correlation Between Ratings and Review Counts



Preliminary Observations - Reviews Dataset

Attributes Data Types

```
The total number of rows are -  28224
The total number of features are -  9
The datatype of various features is -
review_id       object
user_id         object
business_id     object
stars          float64
useful         float64
funny          float64
cool           float64
text            object
date           float64
dtype: object

Computing the duplicate rows
8266
```

Null Values

```
Computing the Null/Missing values
review_id       7773
user_id        10459
business_id    13670
stars          17787
useful         18019
funny          18103
cool           18172
text           17808
date           21962
dtype: int64
```

Data Cleaning Strategies - Reviews Data

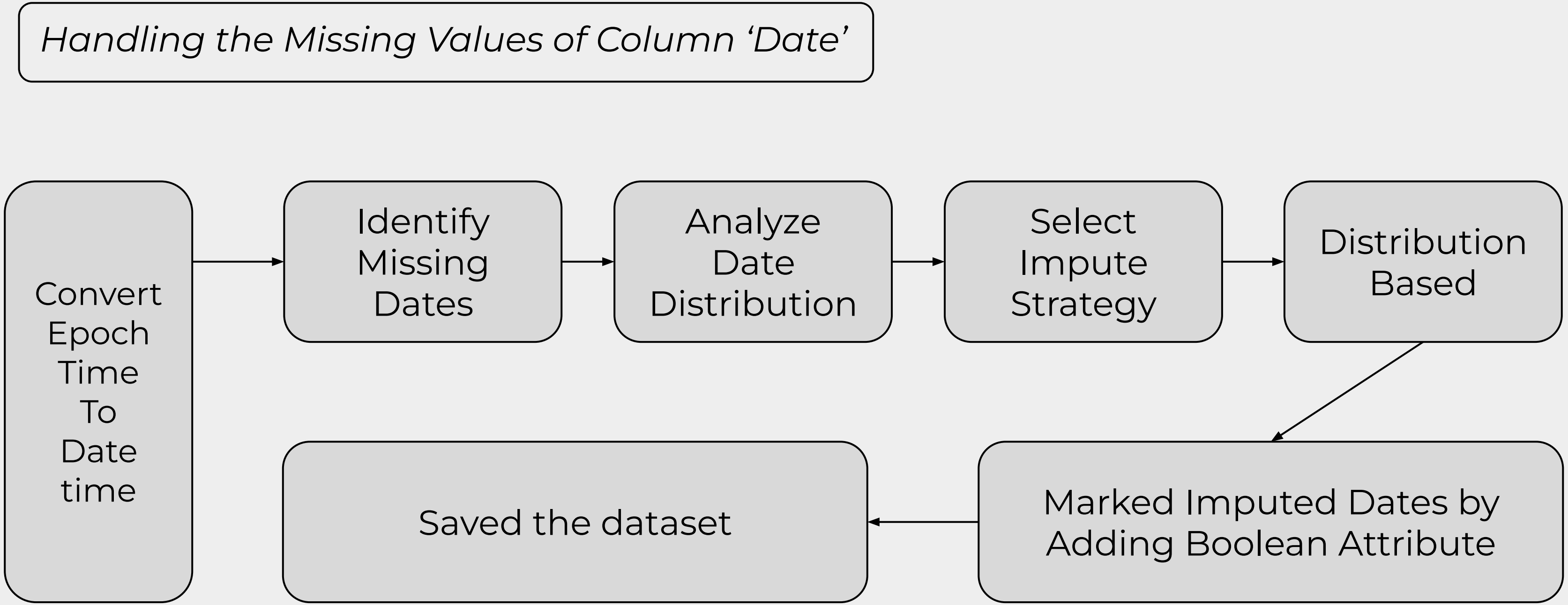
Handling the Duplicate and Incomplete Data

Records without IDs in reviews dataset may exist due to data entry errors or extraction issues. A filtering method was used to remove these records, ensuring data quality for analysis.

Check the Number of Missing Attributes → Compare with threshold → Drop the tuple

```
Initial number of rows: 28224
Number of rows after dropping missing IDs: 14553
Number of rows after filtering based on null count: 10000
Total number of rows removed: 18224
```

Data Cleaning Strategies - Reviews Data



Data Cleaning Strategies - Reviews Data

Why Use Distribution

Maintains Data Integrity: Imputing dates based on existing distributions maintains the integrity of the dataset, ensuring that the imputed dates are representative of the actual data patterns.

Avoiding Misleading Analysis: Placeholders can skew results and give a false impression of data trends. Time-based analysis preserves the natural variance and distribution of the data.

Integrating the Datasets

Combining the Datasets

Integration Process:
Step 1: Identify the common key ('business_id') between datasets.
Step 2: Merge datasets on 'business_id' using an inner join.
Step 3: Retain relevant columns from both datasets for comprehensive analysis.
Step 4: Review the integrated dataset to ensure accuracy and completeness.

Key Features for Analysis:

Business Dataset: Contains details about businesses (e.g., name, location, attributes).

Reviews Dataset: Contains customer reviews and ratings for businesses.

Integrating the Datasets

```
LOAD DATASET Integrated_Data AS csv FROM Integrated_Data.csv @ url 'Integrated_Data.csv'
```

Console | Timing | Datasets | Charts

integrated_data (688 rows)

Views

	_c0 (short)	business_id (string)	name (string)	address (string)	city (string)	state (string)	postal_code (float)	latitude (float)
0	9	IuKCyfSY7AKhRbRA1JPIPw	Aster's Floral Shop	41 Haddon Ave	Collingswood	NJ	8108	39.91556167602539
1	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
2	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
3	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
4	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
5	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
6	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
7	59	vD2jzpPv4iyOLKzITscGvA	Farmhaus Restaurant	3257 Ivanhoe Ave	Saint Louis	MO	63139	38.604129791259766
8	90	8OPE0UaZHf7vdYkgbIj8vg	Fox and Hound English Pub and Grille	4901 E 82nd St, Ste 900	Indianapolis	IN	46250	39.90424728393555
9	90	8OPE0UaZHf7vdYkgbIj8vg	Fox and Hound English Pub and Grille	4901 E 82nd St, Ste 900	Indianapolis	IN	46250	39.90424728393555
10	367	SRb3xScVOeYfhZP4U8LMdA	Paradise Bakery	8540 Castleton Corner Dr	Indianapolis	IN	46250	39.910667419433594
11	367	SRb3xScVOeYfhZP4U8LMdA	Paradise Bakery	8540 Castleton Corner Dr	Indianapolis	IN	46250	39.910667419433594
12	371	zjGHI3SlpU-N3ZPC3skhSg	Grapevine Wines & Spirits	309 S Kirkwood Rd	Saint Louis	MO	63122	38.578548431396484
13	376	QeO2pk5Bg_olt8-j7wg_uQ	Bayshore Pet Resort	4804 S Manhattan Ave	Tampa	FL	33611	27.895435333251953
14	394	zSEFzAbxnO44w8ACQ_XkAA	Akira Hibachi & Sushi Bar	13420 Boyette Rd	Riverview	FL	33569	27.852144241333008
15	394	zSEFzAbxnO44w8ACQ_XkAA	Akira Hibachi & Sushi Bar	13420 Boyette Rd	Riverview	FL	33569	27.852144241333008
16	422	-7GjicSH_rM8JeZGCXGcUg	Double Decker	1721 E 7th Ave	Tampa	FL	33605	27.960084915161133
17	445	0hIXH9jMdHov1VrLC8ujUg	Al's Finger Licking Good Bar-B-Que	1609 Angel Oliva Sr St	Tampa	FL	33605	27.95924949645996
18	445	0hIXH9jMdHov1VrLC8ujUg	Al's Finger Licking Good Bar-B-Que	1609 Angel Oliva Sr St	Tampa	FL	33605	27.95924949645996
19	445	0hIXH9jMdHov1VrLC8ujUg	Al's Finger Licking Good Bar-B-Que	1609 Angel Oliva Sr St	Tampa	FL	33605	27.95924949645996

Thank you