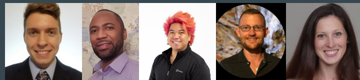
Dallas Restaurant Food Inspection Ratings

…

SMU Data Science Bootcamp: Project 1

Presented by:
Austin Potts, Kirpatrick Dorsey, Sean Kendrick Del Alcazar, Shane Gatenby, Stephanie Smith

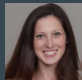


Explore Dallas Restaurants

Is there a correlation between Health Inspection Score and Customer Rating for Dallas restaurants?

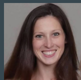
Secondary Question to Explore:

Is there a relationship between Yelp and Google ratings?



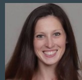
Refine Question and Purpose

- **What do we expect to find?**
  - **Null-hypothesis:** There is no expected relationship between the review a restaurant receives on Yelp or Google and the health inspection score that they receive.
  - **Alternative-hypothesis**: There is an expected positive relationship between the review a restaurant receives on Yelp or Google and the health inspection score that they receive.
- **Task**
  - Rejecting, or failing to reject, Null-hypothesis (H0). Tested using OLS regression.
  - p-value <0.05
- **Explore additional questions and considerations**
  - How consistent are Google and Yelp ratings?
  - Other questions - abandoned due to low confidence in data quality and other challenges



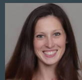
The Process

- Gather data
- Clean data
- Merge data
- Create visualizations
- Analyze findings
- Conclusions



Data Gathering - 3 Main Sources

- **Dallas Open Data**
  - Initial list of restaurants and inspection scores
  - https://www.dallasopendata.com/City-Services/Restaurant-and-Food-Establishment-Inspections-Octo/dri5-wcct
- **Google APIs**
  - Retrieved Lat and Long based on address and zip code
  - Using Lat and Long, retrieve restaurant reviews, review count, and type
  - https://developers.google.com/maps/documentation/geolocation/intro
  - https://developers.google.com/places/web-service/search
- **Yelp Fusion API**
  - Yelp restaurant reviews pulled through API call
  - Using restaurant name, street address and Business Search endpoint
  - https://www.yelp.com/developers

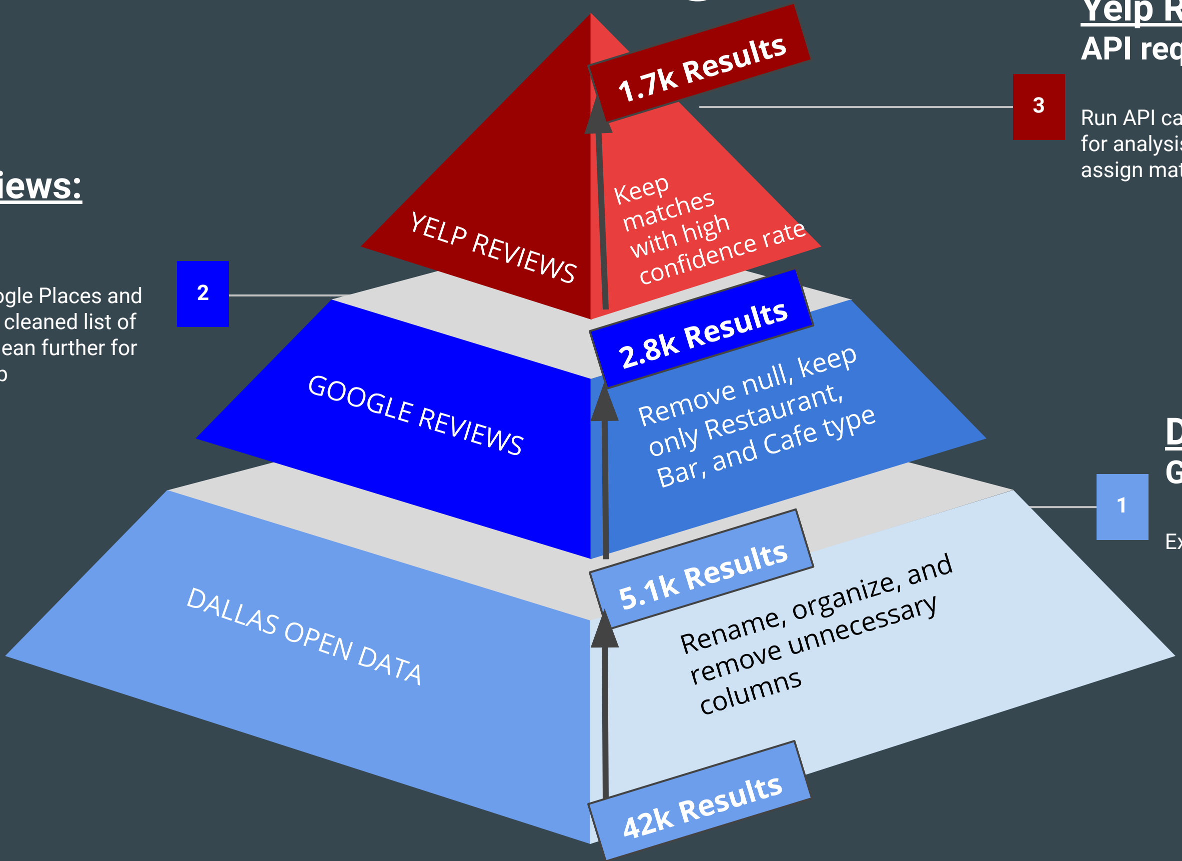


Flow of Data: Gather, Clean, Merge

<u>Google Reviews:</u>
API requests

2

Run API calls for Google Places and Geocoding based on cleaned list of Dallas restaurants, clean further for comparison with Yelp



<u>Yelp Reviews:</u>
API requests

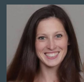
3

Run API calls and clean final data set for analysis. Use FuzzyWuzzy to assign match confidence.

<u>Dallas Open Data:</u>
Get initial restaurant list

1

Export CSV and clean data



Dallas Open Data[1] - Initial Dataset

Dallas OpenData

- Over 42k rows
  - Each row is a facility inspection. Types: Routine, Follow-up, Complaint, Temporary, Mobile
  - A facility can have multiple inspections per year.
- 114 columns
  - Location related - Name, full and segmented address fields, latitude / longitude
  - Time related - Inspection Date, Month, Year
  - Inspection score
    - Range 51 - 100. Excluded two outliers (0, -5)
  - Violation Information (25 type options per visit) - Description, Points, Detail, Memo

Inspection Score	Interpretation
>90	Good
86-90	Adequate
71-85	Needs improvement
<= 70	poor

Dallas Open Data - Clean

- Challenges
  - Low confidence in facility latitude / longitude values
  - 114 columns!
  - Filtering out non places of interest. E.g. School & hospital cafeterias, concession stands, convenience stores, etc.
- Approach
  - Download (raw) inspections as .csv and import to Jupyter Notebook as a dataframe
  - Clean and extract columns and rows of interest
    - Retain 1 unique facility inspection per year using **pd.drop_duplicates(subset=['street_address'])**
    - Keep location columns, facility name, datetime columns, inspection type, inspection score

Dallas Open Data - Clean (6k rows)

Out[8]:

Restaurant Name	Inspection Type	Inspection Date	Inspection Score	Street Number	Street Name	Street Unit	Zip	Street Address	Inspection Month	Inspecti Year
FRESHII	Routine	10/31/2018	96	2414	VICTORY PARK		75219	2414 VICTORY PARK LN	10/1/18	FY2019
MICKLE CHICKEN	Routine	10/30/2019	100	3203	CAMP WISDOM		75237	3203 W CAMP WISDOM RD	10/1/19	FY2020
WORLD TRADE CENTER MARKET	Routine	11/03/2016	100	2050	STEMMONS		75207	2050 N STEMMONS FRWY	11/1/16	FY2017
DUNKIN DONUTS	Routine	10/30/2019	99	8008	HERB KELLEHER	C2174	75235	8008 HERB KELLEHER WAY STE# C2174	10/1/19	FY2020
CANVAS HOTEL - 6TH FLOOR	Routine	06/11/2018	100	1325	LAMAR		75215	1325 S LAMAR ST	06/1/18	FY2018
...	...	...	...	...	...	...	...	...	...	...

Google Data - Gather

- Complete Geocode API requests with scrubbed data from Dallas Open Data
    - API parameters: Street Address, Zip code
- Append existing data from Dallas Open Data with Google results:
    - Latitude and Longitude
- Complete Places API requests using NearbySearch with updated dataframe
    - API parameters: Latitude and Longitude 50 meter radius/.03 miles
- Append existing data from Dallas Open Data with Google results:
    - Pull in Google Rating, Review Count, and filter by Type: Restaurant

Google Data - Clean

- Challenges
  - EXPENSIVE when running through thousands of Zips
    - Geocoding API and Places both had separate pricing
  - Results would be grouped due to a large radius
- What we did
  - Rotated API keys
  - Keep only Google Restaurant Type, which returned:
    - Restaurant
    - Cafe
    - Bar
  - Reduce dataset based on conditional values for “Type”
  - Set radius to 50 or .03 miles

Google - Final

- Final Dallas + Google data set
  - 2,850 records for Restaurant, Bar, and Cafe types
  - Includes some null ratings to be further cleaned during

```
g_lookup_pd_filter
```

pection Type	Inspection Date	Inspection Score	Street Number	Street Name	Street Unit	Zip	Street Address	Inspection Month	Inspection Year	Lat	Long	Type	Rating	Rating Count
Routine	10/31/2018	96	2414	VICTORY PARK		75219	2414 VICTORY PARK LN	10/1/18	FY2019	32.7879	-96.8092	Restaurant	4.1	100
Routine	04/27/2017	100	4142	CEDAR SPRINGS		75219	4142 CEDAR SPRINGS RD	04/1/17	FY2017	32.8134	-96.8121			
Routine	05/11/2017	98	3878	OAK LAWN	#314	75219	3878 OAK LAWN #314	05/1/17	FY2017	32.8155	-96.8007	Restaurant	4.2	334
Routine	10/10/2017	99	2821	TURTLE CREEK		75219	2821 TURTLE CREEK BLVD	10/1/17	FY2018	32.8041	-96.8073	Bar	4.6	433
Routine	05/23/2019	96	2827	THROCKMORTON		75219	2827 THROCKMORTON ST	05/1/19	FY2019	32.8102	-96.8131			
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Routine	02/02/2019	92	4006	CEDAR SPRINGS		75219	4006 CEDAR SPRINGS RD	02/1/19	FY2019	32.8112	-96.8113	Night_Club	3.6	81
Routine	03/24/2017	89	3030	OLIVE	#103	75219	3030 OLIVE ST #103	03/1/17	FY2017	32.7897	-96.8093	Bar	3.9	168
Routine	07/11/2018	92	3211	OAK LAWN	#C	75219	3211 OAK LAWN AVE #C	07/1/18	FY2018	32.8103	-96.8083	Restaurant	4.4	277
Routine	03/14/2019	93	3888	OAK LAWN	#106	75219	3888 OAK LAWN AVE #106	03/1/19	FY2019	32.816	-96.8014			
Routine	02/27/2019	91	2400	HENDERSON	#B	75219	2400 N HENDERSON #B	02/1/19	FY2019	32.8152	-96.7784	Bar	4.1	2017

Yelp Data - Gather

- Complete Business Search endpoint API request with scrubbed data from Dallas Open Data and Google
- API parameters:
  - Search term: restaurant name
  - Location: restaurant street number, street name, city, state and zip code
  - Radius: 4,000 meters (~2.5 miles)
  - Sort by: best match
- Append existing data from other two sources with Yelp results:
  - Yelp ratings and review counts
  - Various other fields that could potentially aid in further cleaning or analysis of the data

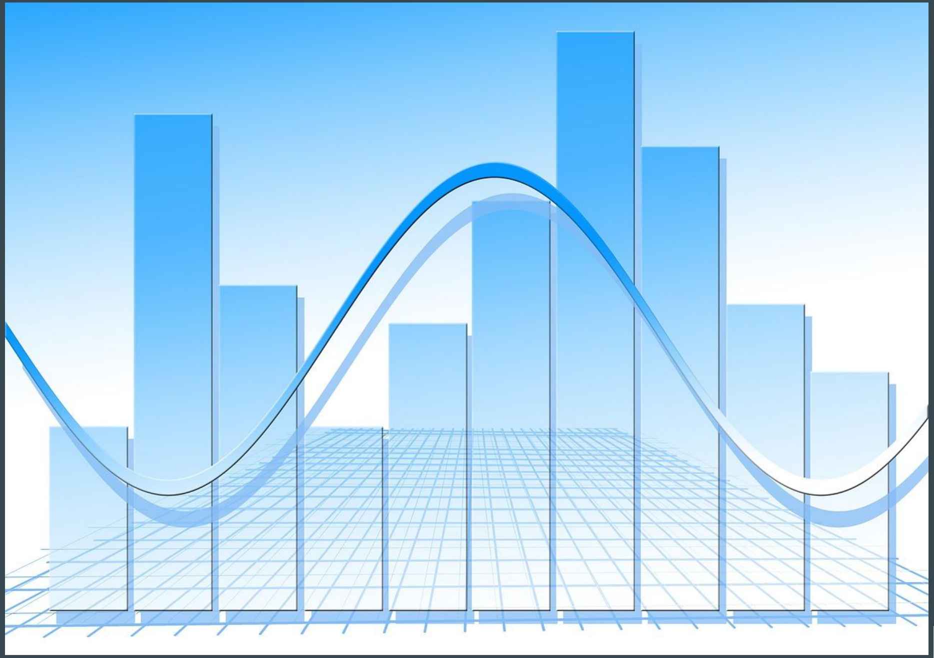
Yelp Data - Clean

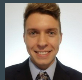
- What we did
  - Used the first result (top hit for “best match” search) returned from Yelp
  - Compared it to the input name and street address to try to determine if it was a match
- How we did it
  - FuzzyWuzzy (Python library; uses the Levenshtein Distance to compare string values)
  - Assign match confidence score (WRatio) to:
    - restaurant name
    - street address

Yelp - Final

- Final **Dallas** + **Google** + **Yelp** data set
  - For analysis, only uses records with
    - Yelp ratings
    - Restaurant name match scores (≥80)
    - Restaurant address match scores (≥80)
- Challenges:
  - Yelp results didn’t always match Dallas Open Data and Google
  - Results didn’t consistently include all objects (i.e., Price, Category 2, Category 3, etc...)
  - Broad range of categories and verbiage used inconsistently across three different category fields

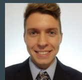
Analysis and Visualizations





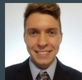
Inspection Score Count by Month

- **Findings:**
  - Greater occurrence of health inspections in second half of the year
  - December had highest occurance of inspections overall
- **Challenges/Concerns**:
  - Smaller dataset for prior years (< 2019), potentially skewing results
  - Broader dataset might show more even distribution across months



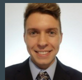
Is there a relationship between Yelp/Google rating and health inspection score?

- We fail to reject the null hypothesis, or we cannot conclude with any level of accuracy that there is any relationship between the review a restaurant receives and their health inspection score based on the P-value: 0.23522 for Yelp, P Value is: 0.53842 for Google, Sample Size = 1703



Is there a relationship between Google and Yelp reviews? Are reviews consistent across platforms?

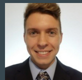
- We fail to reject the Null Hypothesis, it is possible that there is a relationship between the rating a restaurant receive on google and yelp, determined by our P-value: 4.84 X 10^104, and Sample size = 1703, R2 value is: 0.241113



Did we reject the Null-hypothesis?

Health Inspection vs Rating: Yes

Yelp vs Google: Potentially Not



Questions?