Analysis of restaurant markets by city using TripAdvisor and Scrapy

David Corrigan – Web Scraping Project

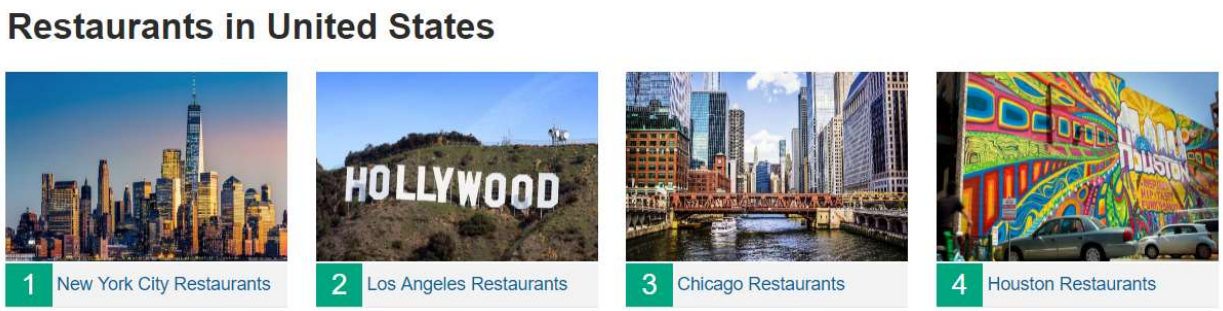
October 16, 2018

Restaurant Analysis by City -- Rationale

- Potential interest for two groups – consumers and developers
  - Consumers
    - Which cuisines/types of restaurants are successful and well-received in a specific city?
      - Vacation planning – what to avoid, try in a particular place
  - Restaurant Developers
    - What cuisines/restaurant types are thriving in a given city?
      - Likely high barriers to entry
    - What types of restaurants are well-rated but underrepresented?
      - Good opportunities for development
    - What types of restaurants are poorly-received in a city of interest?
      - Further research needed – possible cultural or geographical reasons, potentially an opportunity to develop a market

Web Scraping from TripAdvisor.com -- Scrapy

Top 20 Cities in the US on the first page:



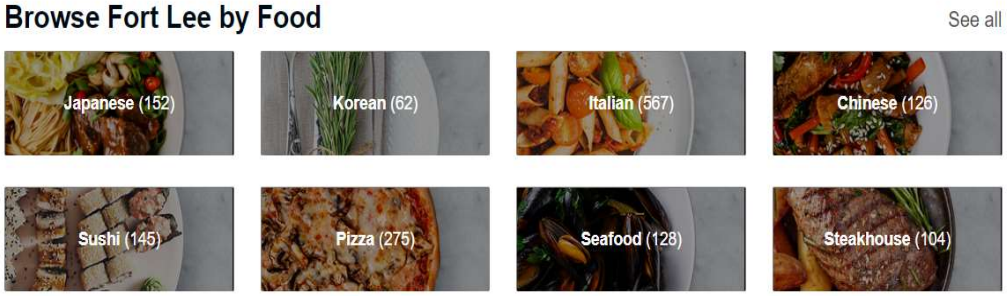
All subsequent pages (20 cities per page)

Obtain URLS, parse the first page of the city (general info and the top 30 restaurants), create parse function for subsequent pages

137,000 restaurants parsed among the top 60 US cities
- Cut to 50 (Bronx, Scottsdale, AZ, Oahu&Honolulu)

One technical issue – Is every restaurant on the city's page actually in the city?

Fort Lee, NJ



Menya Sandaime

49 reviews

#3 of 131 Restaurants in Fort Lee

$, Japanese, Asian, Soups

"Always been the best" 09/16/2018

"Best Ramen in town" 07/10/2018

SottoCasa Pizzeria New York City, NY 3.5 mi away

261 reviews

#2 of 10,751 Restaurants in New York City

$$ - $$$, Pizza, Vegetarian Friendly, Vegan Options, Gluten Free Options

"Pizza!" 10/13/2018

"Great place!" 10/13/2018

#18 Oahu

Sweet E's Cafe Honolulu

329 reviews

#1 of 3,040 Restaurants in Oahu

$$ - $$$, American, Cafe, Vegetarian Friendly, Vegan Options, Gluten Free Options

"Delicious Food & Friendly Service" 10/04/2018

"Great breakfast!" 10/02/2018

#26 Honolulu

Sweet E's Cafe

329 reviews

#1 of 2,000 Restaurants in Honolulu

$$ - $$$, American, Cafe, Vegetarian Friendly, Vegan Options, Gluten Free Options

"Delicious Food & Friendly Service" 10/04/2018

"Great breakfast!" 10/02/2018

Extra Xpath element in restaurants outside of the target city that can be dropped using a Pipeline function

- Saved Scrapy output as a JSONL file, which was read into Python as text, converted null>None and false>False

- Literal_eval could then understand this as a dictionary, converted list of dictionaries to Pandas DF


```
'{"CityCode": "g35805", "CityName": "Chicago", "StateName": "Illinois (IL)", "RegionName": null, "CountryName": "United State
s", "IsOutofCity": false, "NumReviews": "5,454", "Price": "$$ - $$$", "CuisinesList": ["Mediterranean", "Vegetarian Friendly",
"Vegan Options", "Gluten Free Options", "Restaurants"], "AvgRating": "4.5", "RestaurantName": "The Purple Pig", "RestaurantLin
k": "/Restaurant_Review-g35805-d1647641-Reviews-The_Purple_Pig-Chicago_Illinois.html", "RankBlurb": "\\n#24 of 8,601 Restaurant
s in Chicago\\n"}\n'
```



```
AvgRating                                                  4.5
CityCode                                                 g35805
CityName                                                Chicago
CountryName                                       United States
CuisinesList      [Mediterranean, Vegetarian Friendly, Vegan Opt...
IsOutofCity                                              False
NumReviews                                               5,454
Price                                                  $$ - $$$
RankBlurb                    \n#24 of 8,601 Restaurants in Chicago\n
RegionName                                                None
RestaurantLink     /Restaurant_Review-g35805-d1647641-Reviews-The...
RestaurantName                                  The Purple Pig
StateName                                        Illinois (IL)
Name: 77, dtype: object
```


Cleaned the data and removed some of the columns not needed for this analysis

Finally, I “unlisted” the CuisinesList column from this data frame, creating a new line for each type of cuisine found at a restaurant (multiple lines per restaurant now). I also trimmed Cuisines that were very obscure.

- Careful to use either the “collapsed” DF or “expanded” DF, depending on which question being asked

118,000 length data frame – “Collapsed”

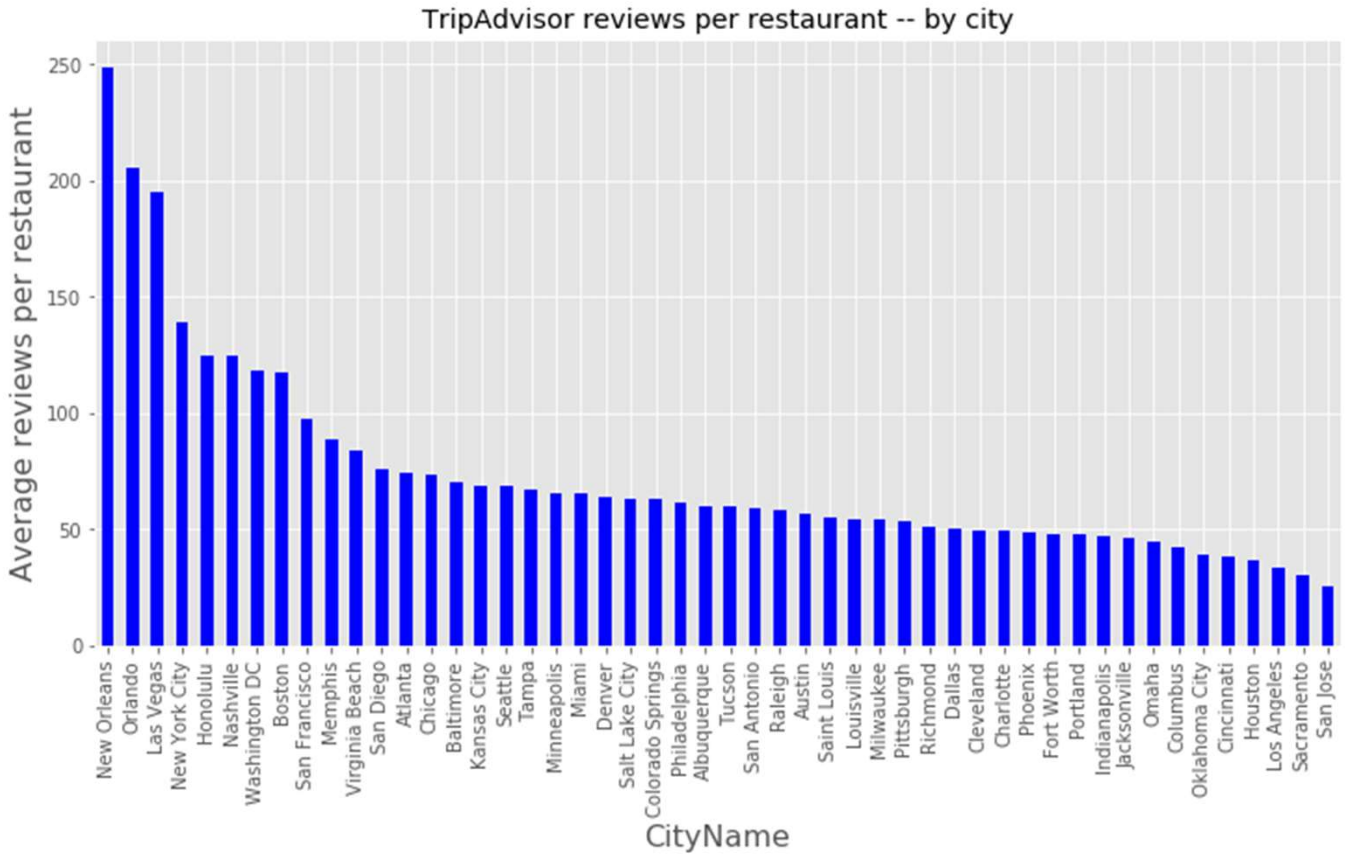
	AvgRating	CityName	CuisinesList	NumReviews	Price	RestaurantName	StateName
0	5.0	New York City	[Pizza, Vegetarian Friendly, Vegan Options, Gl...	251.0	2.0	SottoCasa Pizzeria	New York (NY)
1	5.0	New York City	[French, Vegetarian Friendly, Vegan Options, G...	273.0	3.0	Boucherie Park Avenue South	New York (NY)
2	4.5	New York City	[Italian, Fast Food, Vegetarian Friendly, Vega...	847.0	1.0	Pisillo Italian Panini	New York (NY)
3	4.5	New York City	[American, Steakhouse, Gluten Free Options, Re...	3149.0	3.0	Club A Steakhouse	New York (NY)
4	4.5	New York City	[French, Vegetarian Friendly, Vegan Options, G...	3032.0	3.0	Daniel	New York (NY)

300,000+ length data frame – “Expanded”

	index	AvgRating	CityName	NumReviews	Price	RestaurantName	StateName	Cuisine
0	0	5.0	New York City	251.0	2.0	SottoCasa Pizzeria	New York (NY)	Pizza
1	0	5.0	New York City	251.0	2.0	SottoCasa Pizzeria	New York (NY)	Vegetarian Friendly
2	0	5.0	New York City	251.0	2.0	SottoCasa Pizzeria	New York (NY)	Vegan Options
3	0	5.0	New York City	251.0	2.0	SottoCasa Pizzeria	New York (NY)	Gluten Free Options
5	1	5.0	New York City	273.0	3.0	Boucherie Park Avenue South	New York (NY)	French

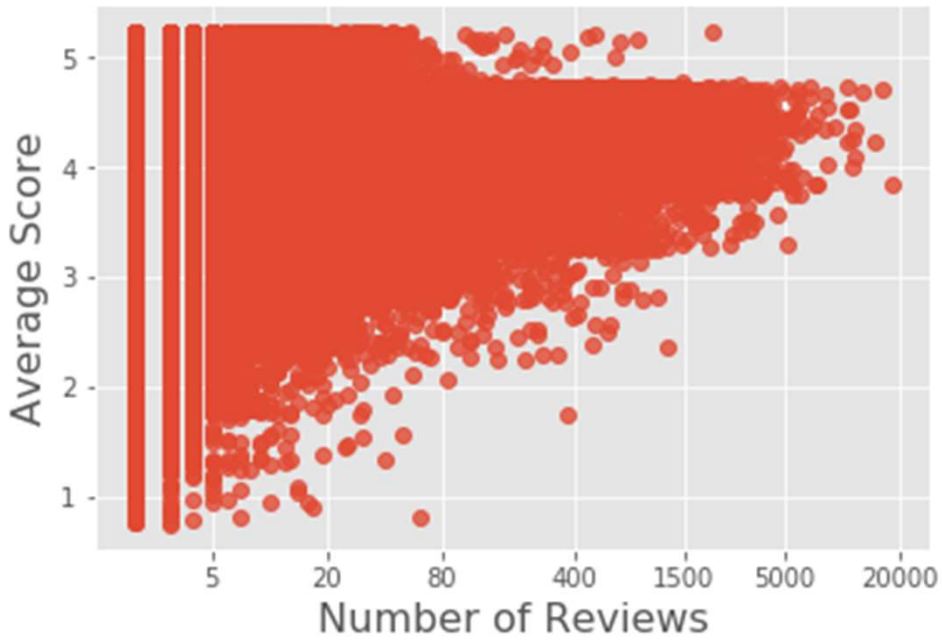
CityName	Cuisine	index	AvgRating
Albuquerque	American	330	3.990909
	Asian	121	4.111570
	Bar	95	3.868421
	Barbecue	23	4.021739
	Cafe	51	4.264706
	Caribbean	5	4.400000
	Chinese	67	3.805970
	Deli	22	4.000000
	Diner	24	3.979167
	Fast Food	91	4.049451
	French	6	4.250000
	Gluten Free Options	120	4.233333
	Greek	14	4.071429
	Indian	9	4.222222
	Italian	61	3.959016
	Japanese	43	4.046512
	Latin American	30	4.333333
	Mediterranean	22	4.340909
	Mexican/Central American	202	4.059406
	Middle Eastern	5	4.100000
	Pizza	103	3.927184
	Pub	78	3.839744
	Seafood	26	3.942308
	Soups	25	4.160000
	South American	5	4.600000
	Southwestern	125	4.140000
	Spanish	40	4.325000
	Steakhouse	17	3.852941
	Sushi	45	4.044444

Does every city have a similar number of reviews per restaurant?



Larger cities – more restaurants – same average # of reviews per restaurant, right? No. It would appear that more “touristy” cities have more reviews per restaurant. May be particularly true of Tripadvisor.com

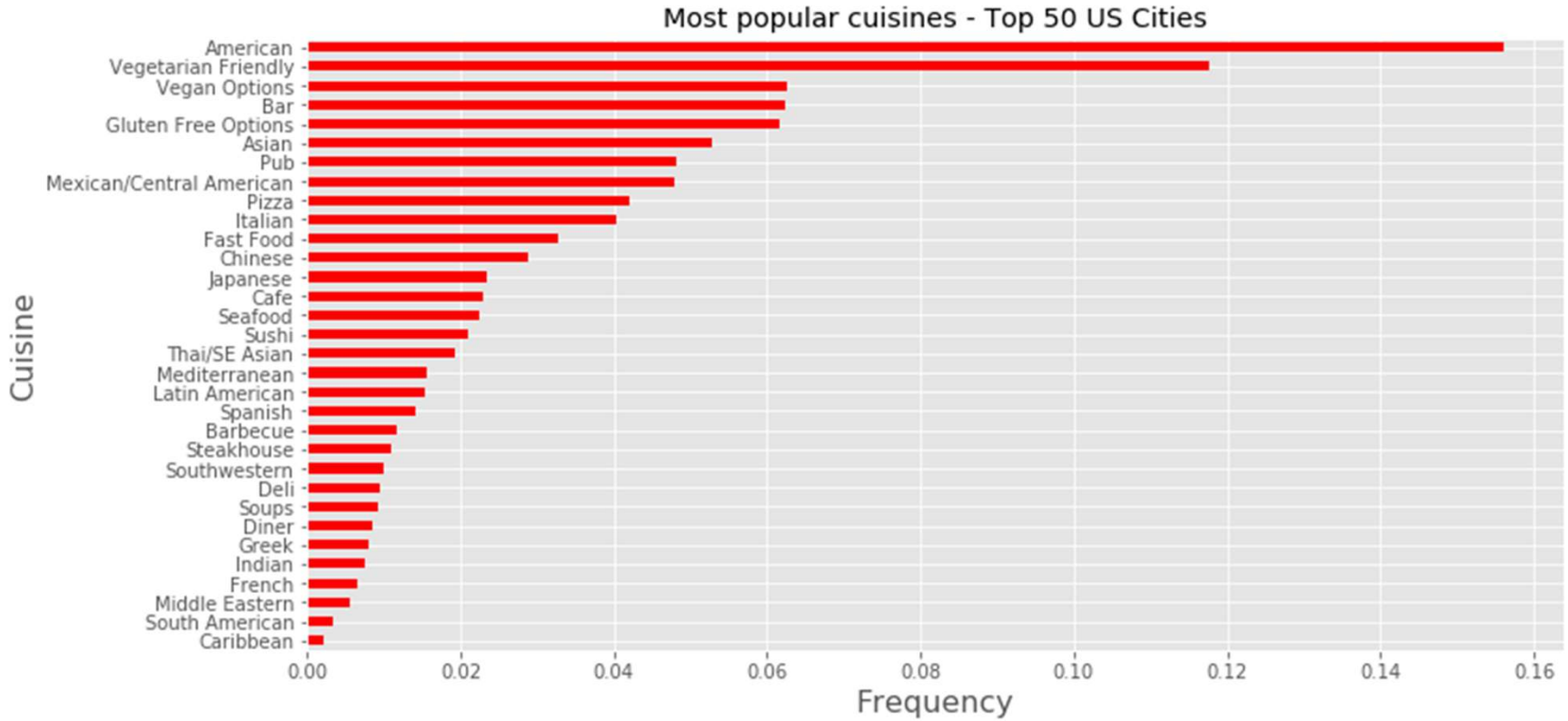
Does the average restaurant score correlate with the number of reviews per restaurant?



Linregress Result
Slope = 0.00011
Intercept = 4.043
Rvalue = 0.0596
Pvalue = 2.56e-74
Stderr = 6/27e-6

Yes, which seems like a logical conclusion, as the most popular restaurants (getting the most reviews) should be good enough to stay open and popular

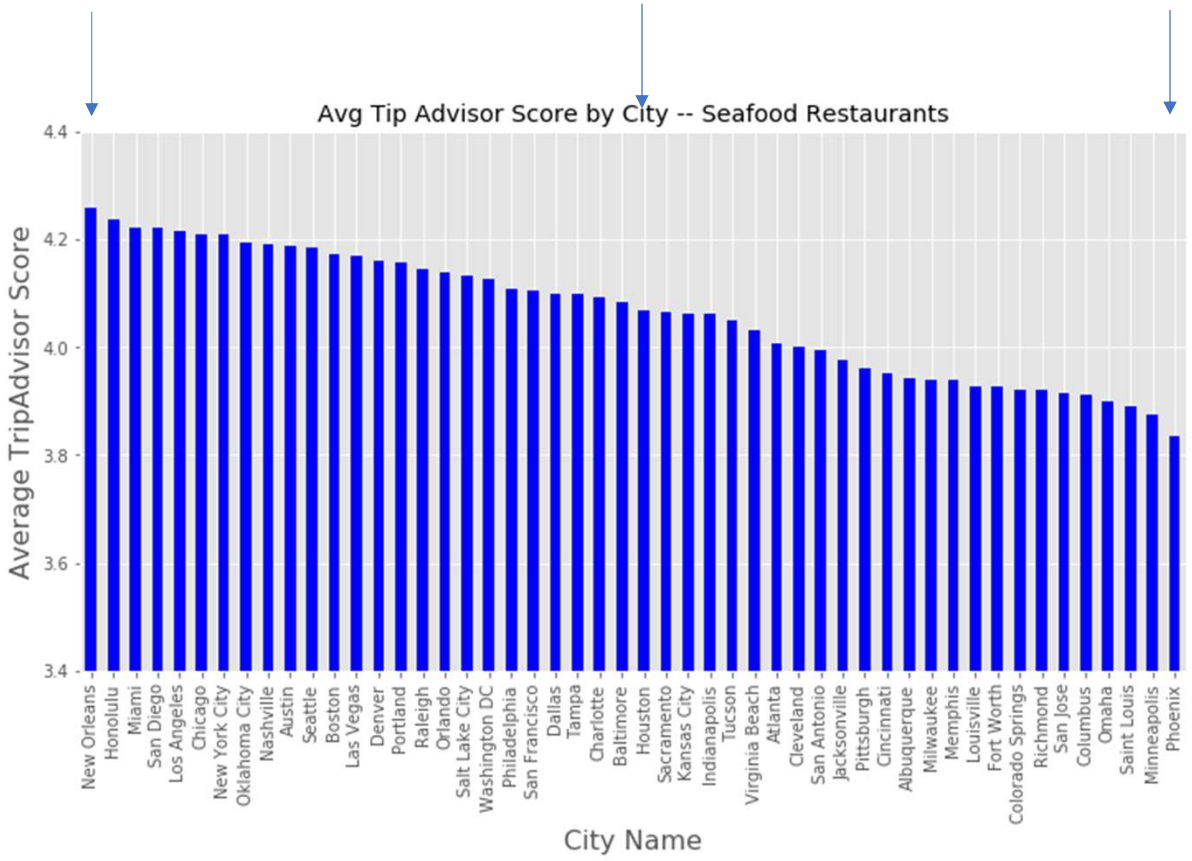
Which restaurant cuisines are most common overall?

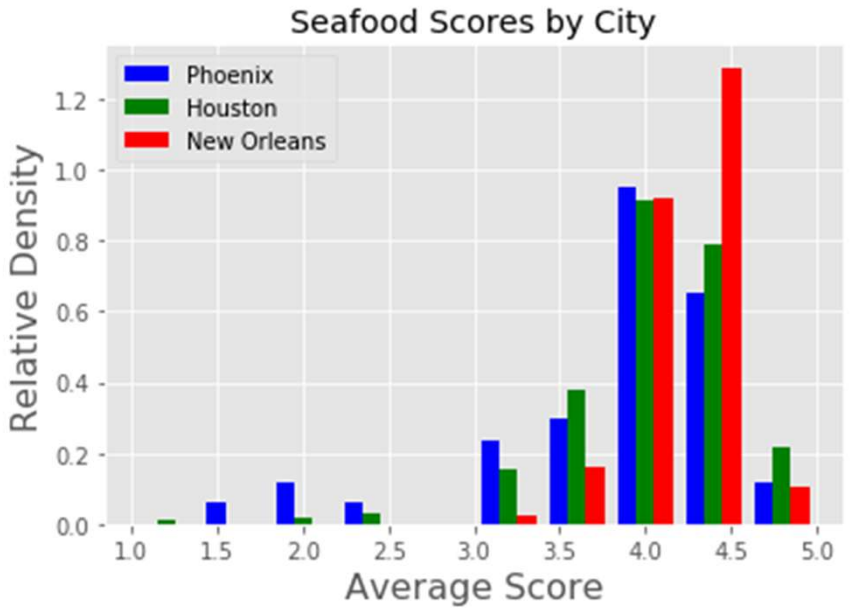


Trimmed DF – less common cuisines not represented

This information used to compile a list of relative cuisine frequencies in each city

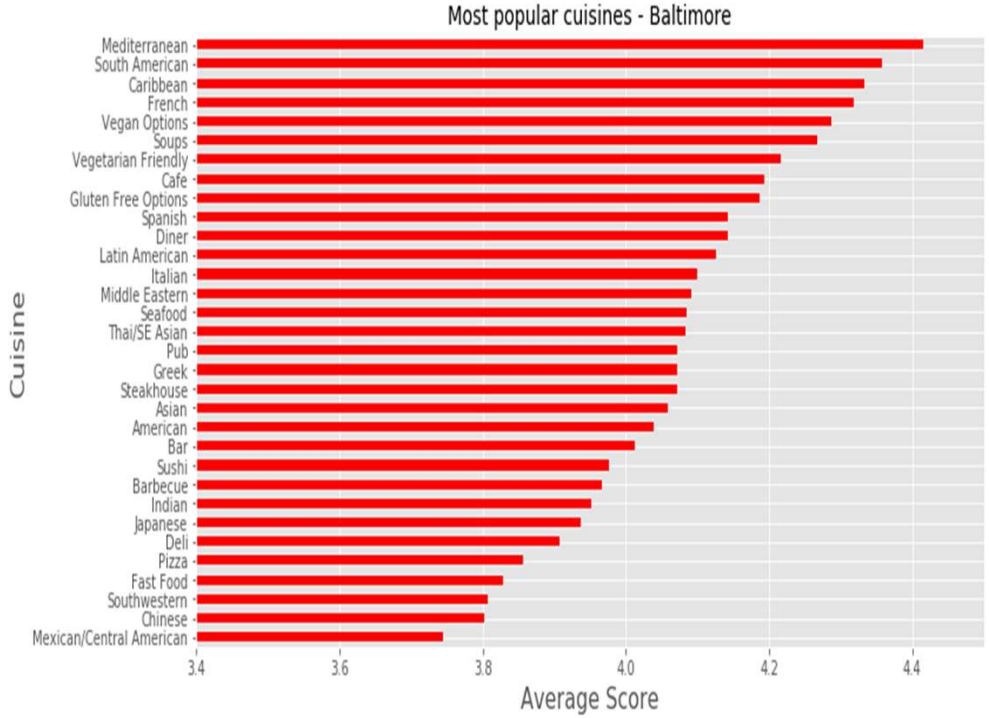
Using these modified DF, we can easily visualize the distribution of cuisines in a given city, or among all cities for a given cuisine – ex. Seafood



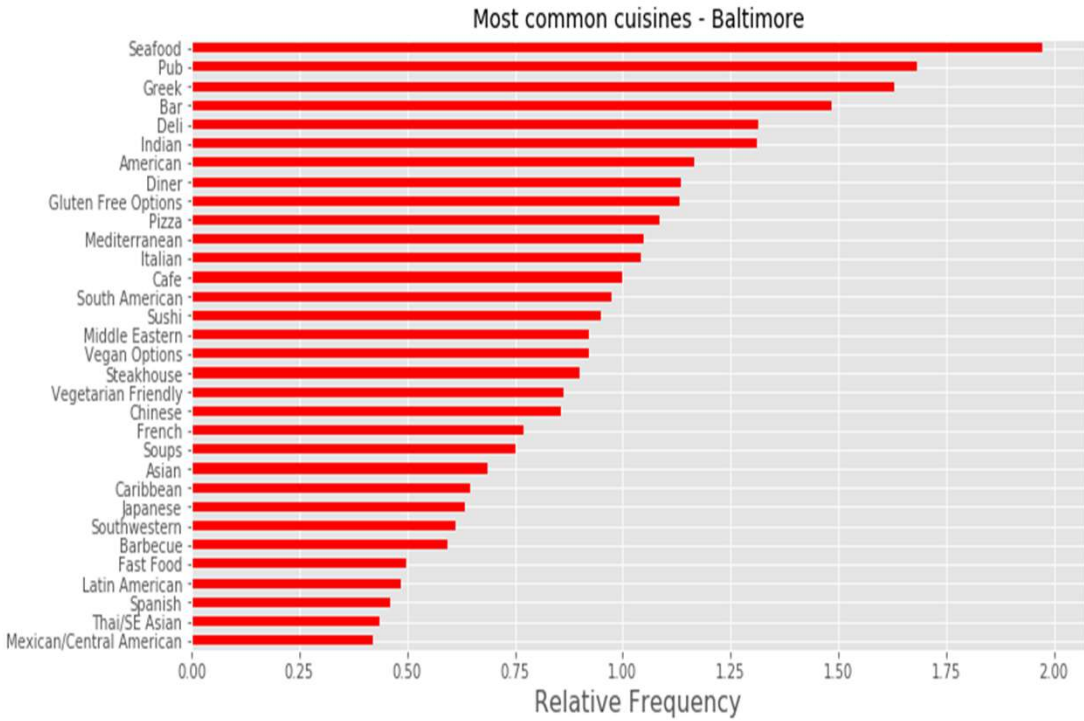


Importantly, we can also visualize the relative abundance and relative popularity of cuisines within a particular city

TripAdvisor Rating

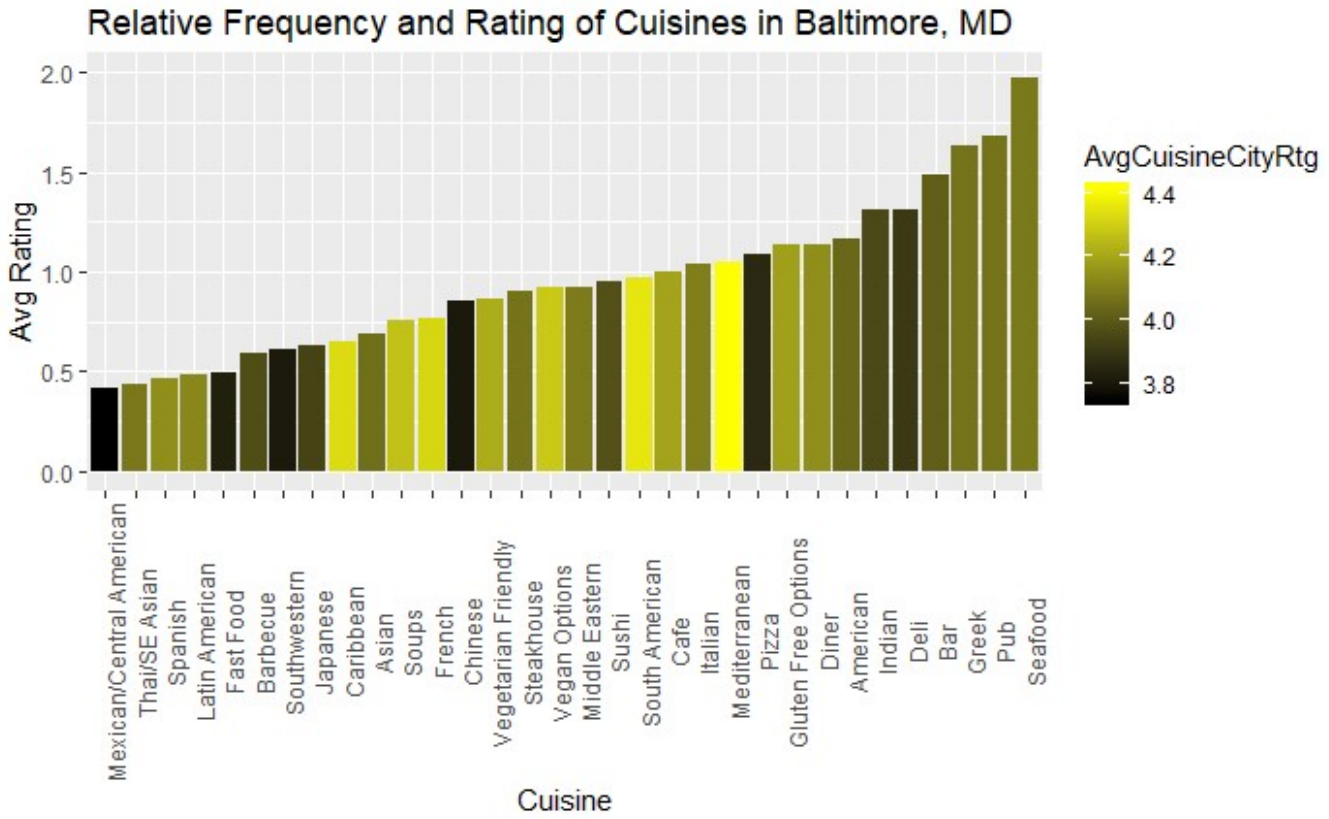


Relative Abundance



Can we combine these two pieces of data to get a more succinct perspective?

Ggplot can tell us in one glance both the relative abundance, and average TripAdvisor ratings of a cuisine within a particular city



Conclusions

- Scrape of TripAdvisor data from 50 major US cities
  - 50 cities, 110K restaurants
  - Cuisine type, opportunities to better understand a market from either a consumer or entrepreneurial perspective
  - Expandable to other geographies
    - Middlesex County, NJ – 1M people, plenty of restaurants
    - Analysis would need to be two-fold – scraping data from individual locations then joining to a larger “region” using another data source (the towns in Middlesex county, for example)
  - Able to visualize data either by cuisine, or city name, quickly using Python or R
  - In some instances (vegan options, i.e.), may be a question of marketing, or an indication that restaurants in that city could do a better job of informing TripAdvisor about their dietary options