**Data Description:**

There are three csv files that you have access to. These files contain Yelp and Point of Interest data for restaurants in Indianapolis. As it is very rare to have access to actual sales/transactions data, please use visits to a location, in the poi.csv, as a proxy for demand. Please find a description of each file below:

**poi.csv:**

This csv file has the following columns.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Type** | **Example** |
| location\_name | The name of the place of interest. | String | Boston Market |
| street\_address | Street address of the place of interest. | String | 1244 W 86th St |
| city | The city of the point of interest. | String | Indianapolis |
| region | The state, province or county of the place of interest. | String | IN |
| postal\_code | The postal code of the place of interest. | String | 46260 |
| date\_range\_start | Start time for measurement period in ISO 8601 format of YYYY-MM-DDTHH:mm:SS±hh:mm (local time with offset from GMT). | String | 2020-03-01T00:00:00-06:00 |
| date\_range\_end | End time for measurement period in ISO 8601 format of YYYY-MM-DDTHH:mm:SS±hh:mm (local time with offset from GMT). The end time will be the last day of the month at 12 a.m. local time. | String | 2020-03-31T00:00:00-06:00 |
| raw\_visit\_counts | Number of visits to this POI during the date range. | Integer | 1542 |
| raw\_visitor\_counts | Number of unique visitors to this POI during the date range. | Integer | 1221 |
| top\_category | The business category that the place belongs to | String | Restaurants and Other Eating Places |
| sub\_category | The sub business category that the place belongs to | string | Limited-Service Restaurants |
| open\_hours | Open hours of the place | string | { "Mon": [["8:00", "21:00"]], "Tue": [["8:00", "21:00"]], "Wed": [["8:00", "21:00"]], "Thu": [["8:00", "21:00"]], "Fri": [["8:00", "21:00"]], "Sat": [["8:00", "21:00"]], "Sun": [["8:00", "21:00"]] } |
| opened\_on | When the business started | Year-month | 2020-02 |
| closed\_on | When the business ended (if applicable) | Year-month | 2020-08 |

**Yelp\_business.csv**

Contains Yelp business data including location data, attributes, and categories.

{

// string, 22 character unique string business id

"business\_id": "tnhfDv5Il8EaGSXZGiuQGg",

// string, the business's name

"name": "Garaje",

// string, the full address of the business

"address": "475 3rd St",

// string, the city

"city": "San Francisco",

// string, 2 character state code, if applicable

"state": "CA",

// string, the postal code

"postal code": "94107",

// float, latitude

"latitude": 37.7817529521,

// float, longitude

"longitude": -122.39612197,

// float, star rating, rounded to half-stars

"stars": 4.5,

// integer, number of reviews

"review\_count": 1198,

// integer, 0 or 1 for closed or open, respectively

"is\_open": 1,

// object, business attributes to values. note: some attribute values might be objects

"attributes": {

"RestaurantsTakeOut": true,

"BusinessParking": {

"garage": false,

"street": true,

"validated": false,

"lot": false,

"valet": false

},

},

// an array of strings of business categories

"categories": [

"Mexican",

"Burgers",

"Gastropubs"

],

// an object of key day to value hours, hours are using a 24hr clock

"hours": {

"Monday": "10:00-21:00",

"Tuesday": "10:00-21:00",

"Friday": "10:00-21:00",

"Wednesday": "10:00-21:00",

"Thursday": "10:00-21:00",

"Sunday": "11:00-18:00",

"Saturday": "10:00-21:00"

}

}

**Yelp\_review.csv**

Contains full review text data including the user\_id that wrote the review and the business\_id the review is written for.

{

// string, 22 character unique review id

"review\_id": "zdSx\_SD6obEhz9VrW9uAWA",

// string, 22 character unique user id, maps to the user in user.json

"user\_id": "Ha3iJu77CxlrFm-vQRs\_8g",

// string, 22 character business id, maps to business in business.json

"business\_id": "tnhfDv5Il8EaGSXZGiuQGg",

// integer, star rating

"stars": 4,

// string, date formatted YYYY-MM-DD

"date": "2016-03-09",

// string, the review itself

"text": "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks.",

// integer, number of useful votes received

"useful": 0,

// integer, number of funny votes received

"funny": 0,

// integer, number of cool votes received

"cool": 0

}

Description (Q4): Using review text to understand restaurant performance:

* + 1. Microsoft Azure sentiment analysis API to identify the sentiment labels for the review text in the reviews. Make sure to go through the Azure documentation to understand how best to go about sending requests so as not to deplete your free funding. Keep in mind that you do not necessarily have to process all available textual reviews. Please explain and justify the set of reviews you decide to process.
    2. Compare the distribution of sentiment labels for high/low- performance restaurants. You will be graded based upon the rigor and methodology with which you use to determine the characteristics. Clearly state rational for the approach(es) you adopt.