An Exploration of Yelp Restaurant Reviews

IST 652 - Scripting for Data Analysis Final Project | Audrey Crockett & Kelly Hwang

Introduction

Yelp is company designed around the collection of reviews for the users of its application to be able to make informed decisions about their consumer needs. As a part of an academic data challenge, Yelp has provided a subset of our businesses, reviews, and user data. The datasets provided were in JSON format and included over 1.4 million business attributes, 188,593 businesses, and 5,996,996 reviews. For the project, we wanted to examine the restaurant reviews with sentiment analysis. Below, we will discuss the preparation of the data, data exploration, analysis, results, and our final conclusions.

Techniques Utilized:

- Data Cleansing on Unstructured Data
- Statistical Analysis
- Data Visualization
- Sentiment Analysis

Data Preparation

Part 1 - Yelp Business Data
Data Source: [Yelp.com Dataset Challenge] (also available on Kaggle)
File Type: JSON
File Size: 1.1 GB
Modules Used: pandas

We first imported the JSON file into a readable dataframe format:

Out[4]:

	address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	name	neighborhood
0	1314 44 Avenue NE	{'BikeParking': 'False', 'BusinessAcceptsCredi...	Apn5Q_b6Nz61Tq4XzPdf9A	Tours, Breweries, Pizza, Restaurants, Food, Ho...	Calgary	{'Monday': '8:30-17:0', 'Tuesday': '11:0-21:0'...	1	51.091813	-114.031675	Minhas Micro Brewery	
1		{'Alcohol': 'none', 'BikeParking': 'False', 'B...	AjEbIBw6ZFfln7ePHha9PA	Chicken Wings, Burgers, Caterers, Street Vendo...	Henderson	{'Friday': '17:0-23:0', 'Saturday': '17:0-23:0...	0	35.960734	-114.939821	CK'S BBQ & Catering	
2	1335 rue Beaubien E	{'Alcohol': 'beer_and_wine', 'Ambience': '{'ro...	O8S5hYJ1SMc8fA4QBtVujA	Breakfast & Brunch, Restaurants, French, Sandw...	MontrÃ©al	{'Monday': '10:0-22:0', 'Tuesday': '10:0-22:0'...	0	45.540503	-73.599300	La Bastringue	Rosemont-La Petite-Patrie

For the purposes and scope of our project, we identified columns that were not needed and therefore dropped the following attributes to reduce data size:

- Business address
- Business attributes
- Business hours
- Is_open
- Latitude

- Longitude
- Neighborhood

We noticed odd postal codes in the postal_code category. Upon further investigation there were international businesses included in this dataset. We wanted to stick with US/domestic businesses only, and removed all rows containing a postal code that was not a 5 digit number.

Next, we tackled the problem of the sheer number of categories and businesses. There were about 188,000 businesses and over 1,200 different categories. To narrow down the scope of our project, we decided to focus only on restaurants. We eliminated any business that did not contain the general category 'restaurant' in its long list of possible categories.

The most challenging aspect of cleaning the data into a format necessary for analysis was transforming the categories column further after reducing its size. As seen in the screenshot above, categories is a multi-value column, with millions of possible value combinations. To produce any meaningful insight, we needed to isolate each value into its own cell. There were 2 possible methods that we could try:

1. Transforming each value into its own binary column (1 for True and 0 for False) while maintaining one row per business ID, or
2. Transforming each value into its own cell, creating multiple rows per business ID if a business ID had multiple categories (which they all did)

Because there are hundreds of different categories, we did not choose method 1 since we didn't want to create 500-1000 columns (and not knowing exactly how many categories there were at this point didn't help).

After category purging and transformation, we were left with 649 unique categories and 142k unique business ids.

```
In [153]: #category basic statistics
          businesses_final['category'].describe()

Out[153]: count         141880
          unique           649
          top      Restaurants
          freq           34332
          Name: category, dtype: object

In [161]: #restaurants id basic statistics
          businesses_final['business_id'].describe()

Out[161]: count                    141880
          unique                    34332
          top      IjsLANGkmAqCsF6-zgIA8w
          freq                         37
          Name: business_id, dtype: object
```

This is our final dataframe for this dataset after cleansing:

```
In [198]: businesses_final.head()
```

Out[198]:

new category column

original category column

	business_id	name	category	city	state	postal_code	review_count	stars	instance	categories
0	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	Chicken Wings	Henderson	NV	89002	3	4.5	0	Chicken Wings, Burgers, Caterers, Street Vendo...
1	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	Burgers	Henderson	NV	89002	3	4.5	1	Chicken Wings, Burgers, Caterers, Street Vendo...
2	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	Caterers	Henderson	NV	89002	3	4.5	2	Chicken Wings, Burgers, Caterers, Street Vendo...
3	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	Street Vendors	Henderson	NV	89002	3	4.5	3	Chicken Wings, Burgers, Caterers, Street Vendo...
4	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	Barbeque	Henderson	NV	89002	3	4.5	4	Chicken Wings, Burgers, Caterers, Street Vendo...

A new column named 'instance' was created to track the number of categories per business.

Part 2 - Yelp Review Data
Data Source: [Yelp.com Dataset Challenge]() (also available on Kaggle)
File Type: JSON
File Size: 4.6 GB
Modules Used: json, pandas, re, textblob

The Yelp Review data was also in JSON format and included nearly six million reviews. This set was difficult to read in due to its enormous size. Using the subsetted Business dataset that was subsetted into restaurants, we were able to use the business ids to filter the review dataset as we brought it in. This was a very time consuming process. When the JSON Yelp reviews were read in they were in the form of a list of dictionaries. We transformed the data from a list to a data frame, but the dictionaries were still embedded with in our our data frame. To solve this problem we had to write a function, as seen in the figure below.

```python
def strip_dict_values(x):
    if isinstance(x,dict):
        v = list(x.values())
        return v[0]
    else:
        return x

cols_with_dict = list(yelp_reviews_df)

for col in cols_with_dict:
 yelp_reviews_df[col] = yelp_reviews_df[col].apply(strip_dict_values)
```

The function strip_dict_values returns the value from the dictionary key value pair. Then with an apply function, we can apply our function to all the columns in our dataframe.

After the review dataset was cleaned, it was ready to be joined with the business data set. We were able to accomplish this through a merge using the business_id. The end result of our merge was a dataframe with 17 columns and 69, 429 rows. A glimpse of the data set can be seen in the figure below.

business_id	categories	city	name	postal_code	review_count	stars_x	state	_id	cool	date	funny
_c3ixq9jYKxhLUB0czi0ug	Bars, Sports Bars, Dive Bars, Burgers, Nightli...	Phoenix	Original Hamburger Works	85007	277	4.0	AZ	5bfb30c90d7c4d55f05c518e	0	2015-05-11	0
_c3ixq9jYKxhLUB0czi0ug	Bars, Sports Bars, Dive Bars, Burgers, Nightli...	Phoenix	Original Hamburger Works	85007	277	4.0	AZ	5bfb30c90d7c4d55f05c592c	0	2010-08-14	0
_c3ixq9jYKxhLUB0czi0ug	Bars, Sports Bars, Dive Bars, Burgers, Nightli...	Phoenix	Original Hamburger Works	85007	277	4.0	AZ	5bfb30ca0d7c4d55f05c6dc2	1	2014-07-29	0

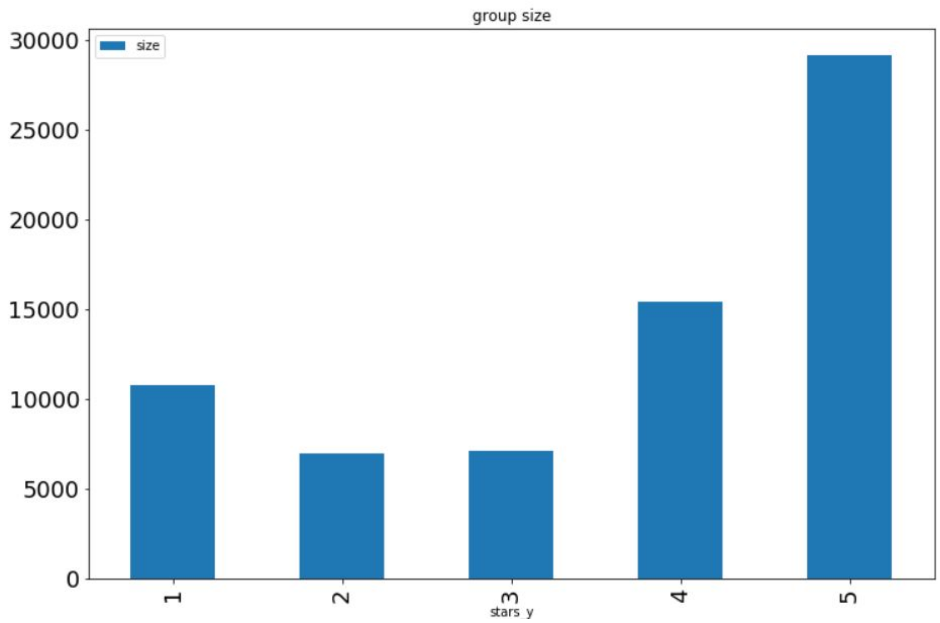
review_id	stars_y	text	useful	user_id
IM_XM7e1nD7d7NJ815inuA	4	Cozy neighborhood sports bar w good burgers. L...	2	UfUFjbwLpYCeJrWUWdMYVA
ACW_G1G0PG0GNyGUPfI3UA	4	The bad: the bar closes at 10pm. It seems lik...	0	VWDL0VgQ2ivpN3oYhL_WsA
KKJa4pRGwq8eO6YCdhjNoA	5	Good food, good vibes, good service. All adds ...	1	SoL4ToJdvxWpGGzxYVA86A

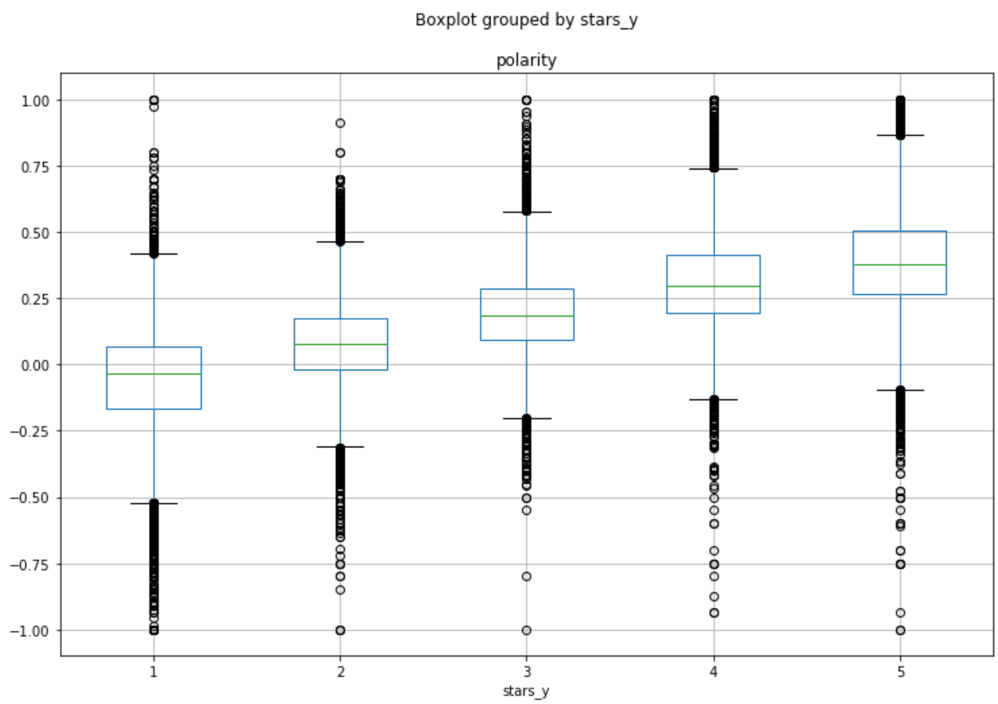
To save on time from loading the larger datasets while performing analysis, we lastly wrote our cleaned and merged data frame to CSV.

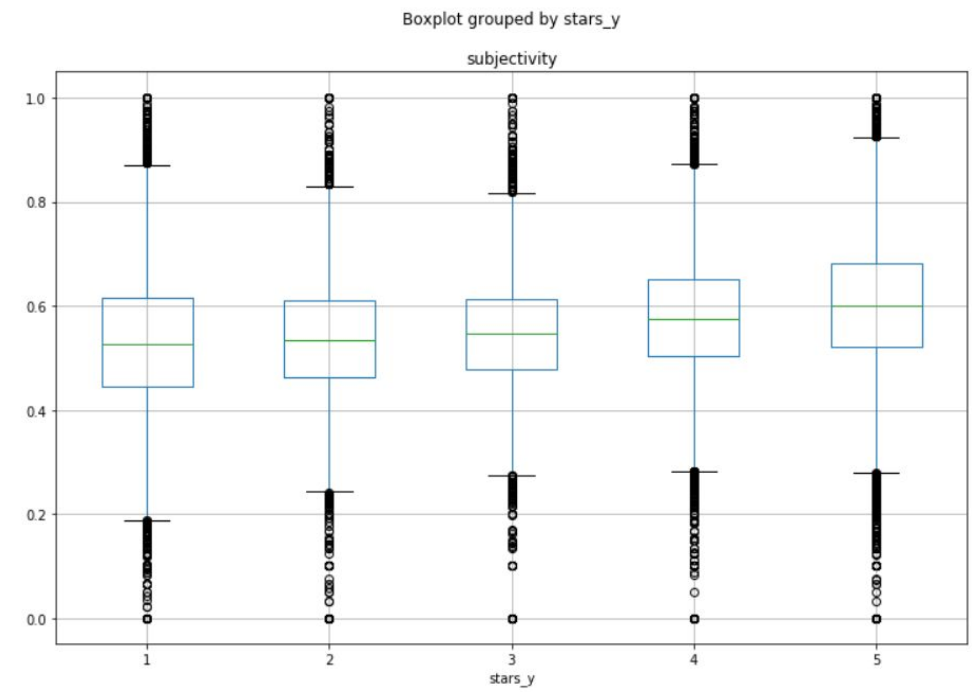
Exploration of Data

Using correlations and visualization, we were able to look at the relationships among attributes and the distributions of certain attributes. Below are outputs of the visualizations. For details, please view the corresponding Jupyter Notebook.

	polarity	subjectivity
polarity	1.000000	0.286414
subjectivity	0.286414	1.000000


group size
size
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
stars_y


Boxplot grouped by stars_y
polarity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
1
2
3
4
5
stars_y



Analysis and Results

The goal of our analysis was to find the sentiment of the reviews for restaurants. We used to TextBlob to calculate the polarity and subjectivity of each review.The authors of TextBlob assign these values as a named tuple of the form Sentiment(polarity,subjectivity). The polarity score has a range of -1.0-1.0. Polarity is the positive or negative sentiment given by the words in a chunk of text. The subjectivity score has a range of 0 to 1 with 0 being very objective and 1 being very subjective. Using the sentiment function, we were able to apply the sentiment analysis to all reviews within our dataframe.

stars_y	text	useful	user_id	sentiment
4	Cozy neighborhood sports bar w good burgers. L...	2	UfUFjbwLpYCeJrWUWdMYVA	(0.2833333333333334, 0.525)
4	The bad: the bar closes at 10pm. It seems lik...	0	VWDL0VgQ2ivpN3oYhL_WsA	(0.16313131313131313, 0.5383838383838384)
5	Good food, good vibes, good service. All adds ...	1	SoL4ToJdvxWpGGzxYVA86A	(0.4402777777777777, 0.5597222222222222)

As seen above, the sentiment column is tuple within the dataframe, which will make analysis difficult. Next we parse the values into to two columns, polarity and subjectivity.

text	useful	user_id	sentiment	polarity	subjectivity
Cozy neighborhood sports bar w good burgers. L...	2	UfUFjbwLpYCeJrWUWdMYVA	(0.28333333333333334, 0.525)	0.283333	0.525000
The bad: the bar closes at 10pm. It seems lik...	0	VWDL0VgQ2ivpN3oYhL_WsA	(0.16313131313131313, 0.5383838383838384)	0.163131	0.538384
Good food, good vibes, good service. All adds	1	SoL4ToJdvxWpGGzxYVA86A	(0.44027777777777777, 0.5597222222222222)	0.440278	0.559722

Using user defined functions called category_polarity and normalize polarity, we were able to put polarity on a 5 point likert scale (one with categorical values and the other with numeric, respectively). The output is seen below.

polarity	subjectivity	polarity_category	polarity_normalized
0.283333	0.525000	Positive	4
0.163131	0.538384	Neutral	3
0.440278	0.559722	Positive	4

Following this, we were able to take the delta of polarity_normalized and the user evaluated star rating. The average delta between stars and normalized polarity was .761, indicated that on average the sentiment analysis was in within one star of the true customer rating. Our average overestimate was 1.056 and our underestimate 0.799, meaning sentiment analysis was more likely to skew positive when compared to star ratings. We then compared these by zip code and restaurant.

		stars_y	polarity_normalized	star_delta
postal_code	name			
6502	Altes Bootshaus	4.000000	4.000000	0.000000
	Athos	4.333333	3.333333	1.000000
	Berghotel Rosstrappe	1.000000	3.000000	-2.000000
	Eisvilla	5.000000	3.000000	2.000000
	GaststÃ¤tte KÃ¶nigsruhe	1.000000	3.000000	-2.000000
	Restaurant Forelle	4.000000	5.000000	-1.000000
6632	CafÃ© Merle	3.000000	3.000000	0.000000
	MÃ¼hle	3.500000	3.000000	0.500000
	Restaurant am Unstrut- Wehr Donath Stefan	1.000000	3.000000	-2.000000
6917	Restaurant zur Alten Brauerei	1.000000	4.000000	-3.000000
	Restaurante Pizzeria	5.000000	3.000000	2.000000
11290	Auberge du Dominicain	4.000000	3.000000	1.000000
12923	Filion's Diner	4.500000	3.500000	1.000000
15003	Ambridge Italian Villa	1.000000	2.000000	-1.000000
	Bridgetown Taphouse	5.000000	3.500000	1.500000
	Frank's Pizzeria	2.500000	3.000000	-0.500000
	K & N Restaurant	5.000000	4.000000	1.000000

Another analysis we conducted is one way ANOVA, which compares the means between the groups and whether or not any of those means are statistically significantly different from eachother. One way ANOVA was the methodology chosen because we have one categorical independent variable (star rating) and one continuous variable (polarity).

The null hypothesis for one way ANOVA:
$H_o = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

If one way ANOVA shows there is a statistically significant result (P < Alpha 0.05) then we reject the Null and accept the Alternative, meaning that there are at least 2 group means that are significantly different from each other.

We conducted one way ANOVA on the polarity values for each star rating group:

```
In [222]: #additional one way ANOVA view
          import statsmodels.api as sm
          from statsmodels.formula.api import ols
```

```
In [218]: model = ols('polarity ~ stars_y', data = yelp_stars).fit()
```

```
In [223]: anova = sm.stats.anova_lm(model, typ=2)
```

```
In [224]: print(anova)

                sum_sq       df             F  PR(>F)
stars_y    1937.368282      1.0  52746.814126     0.0
Residual   2550.061218  69428.0           NaN     NaN
```

Showing P value of 0, this is less than our alpha 0.05. We can conclude that there are at least 2 group means that are statistically different from each other. However, one way ANOVA is limited in where it cannot tell us which groups are the ones that are different from each other. For this, additional analysis must be conducted. Due to the time limitations of our project, this is something we would have liked to do but will not have time to complete.