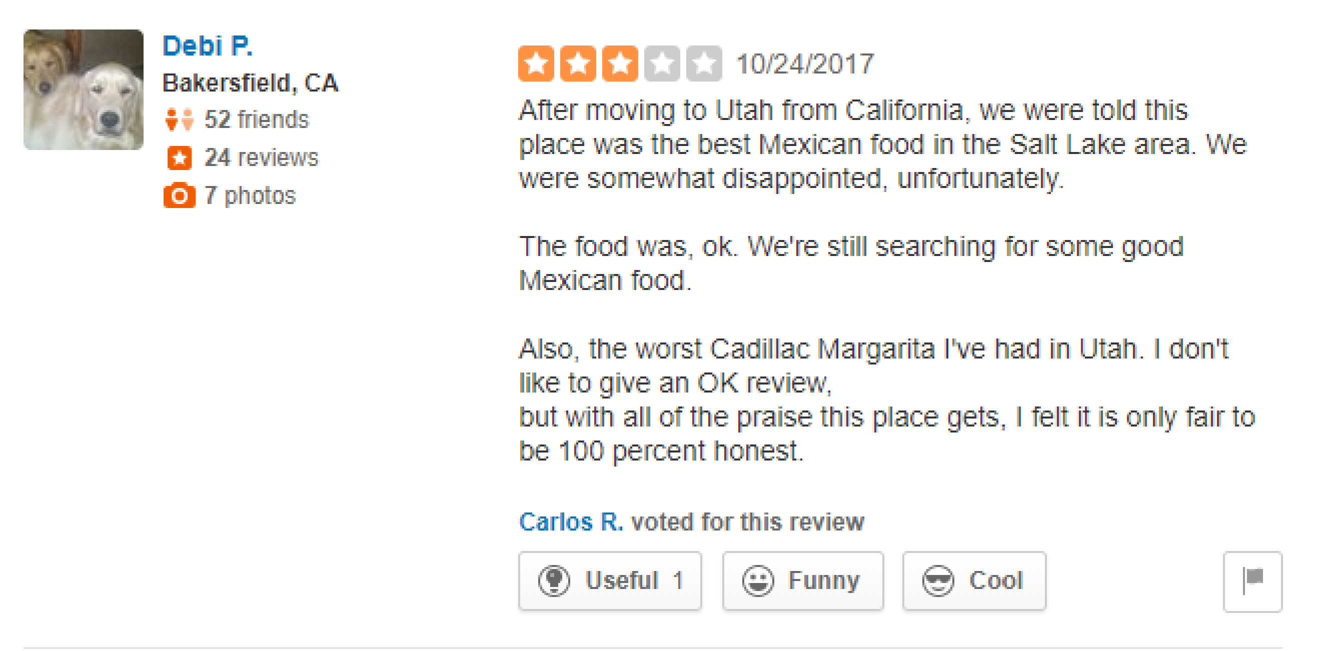
Analyzing Yelp Reviews

NAME: JAIMIE CHOI

EMAIL: JCHEV95@HOTMAIL.COM

Star Rating vs. Actual Reviews

- How reliable are the star ratings for the restaurants?
- Here is an example:



Debi P.
Bakersfield, CA
52 friends
24 reviews
7 photos

★★★☆☆ 10/24/2017

After moving to Utah from California, we were told this place was the best Mexican food in the Salt Lake area. We were somewhat disappointed, unfortunately.

The food was, ok. We're still searching for some good Mexican food.

Also, the worst Cadillac Margarita I've had in Utah. I don't like to give an OK review,
but with all of the praise this place gets, I felt it is only fair to be 100 percent honest.

Carlos R. voted for this review

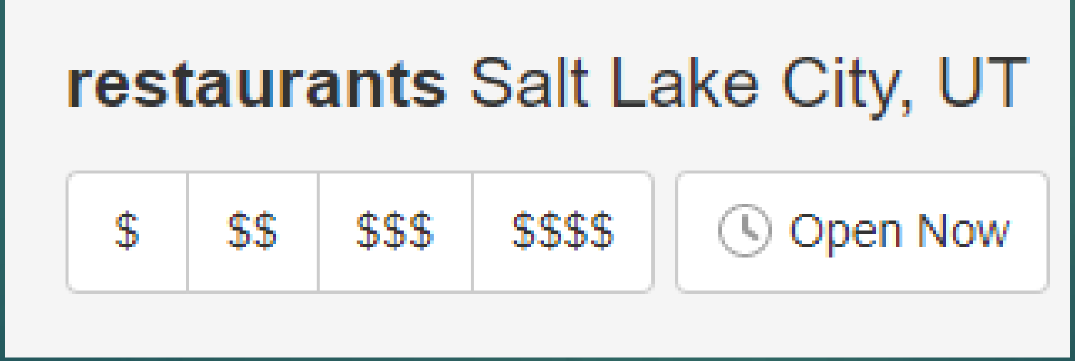
Review us on... yelp

Rating Score: 3/5
Polarity Score: 0.9306

"somewhat disappointed"
"unfortunately"
"ok"
"worst"
"I don't like to give an OK review"

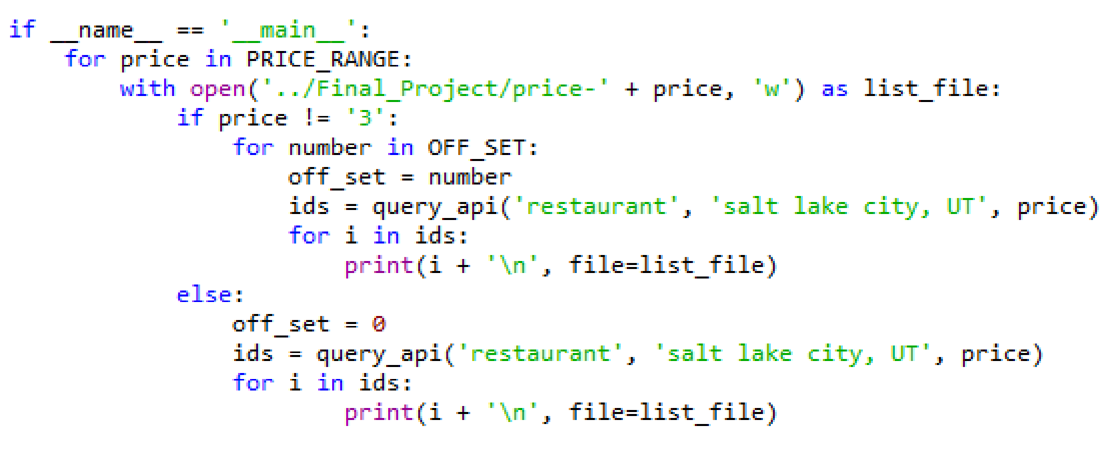
Price Level

- Specifically, does the price range of the restaurant affect the reliability of reviews?
- $ - cheap , $$ - Medium, $$$ expensive, $$$$ ultra-high end

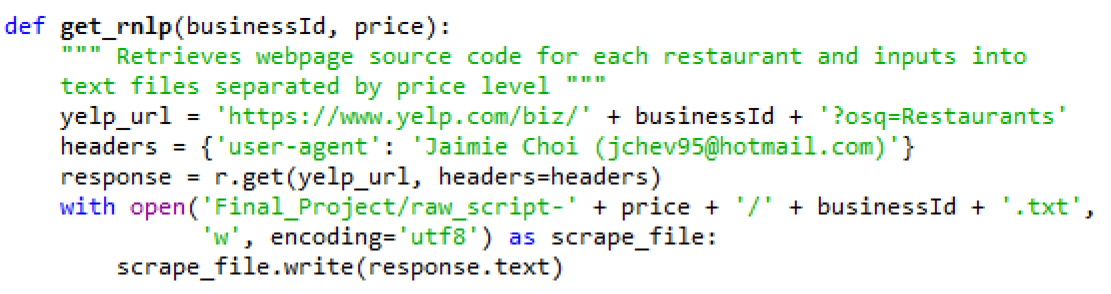


restaurants Salt Lake City, UT

Collecting Data



```python
if __name__ == '__main__':
    for price in PRICE_RANGE:
        with open('../Final_Project/price-' + price, 'w') as list_file:
            if price != '3':
                for number in OFF_SET:
                    off_set = number
                    ids = query_api('restaurant', 'salt lake city, UT', price)
                    for i in ids:
                        print(i + '\n', file=list_file)
            else:
                off_set = 0
                ids = query_api('restaurant', 'salt lake city, UT', price)
                for i in ids:
                        print(i + '\n', file=list_file)
```



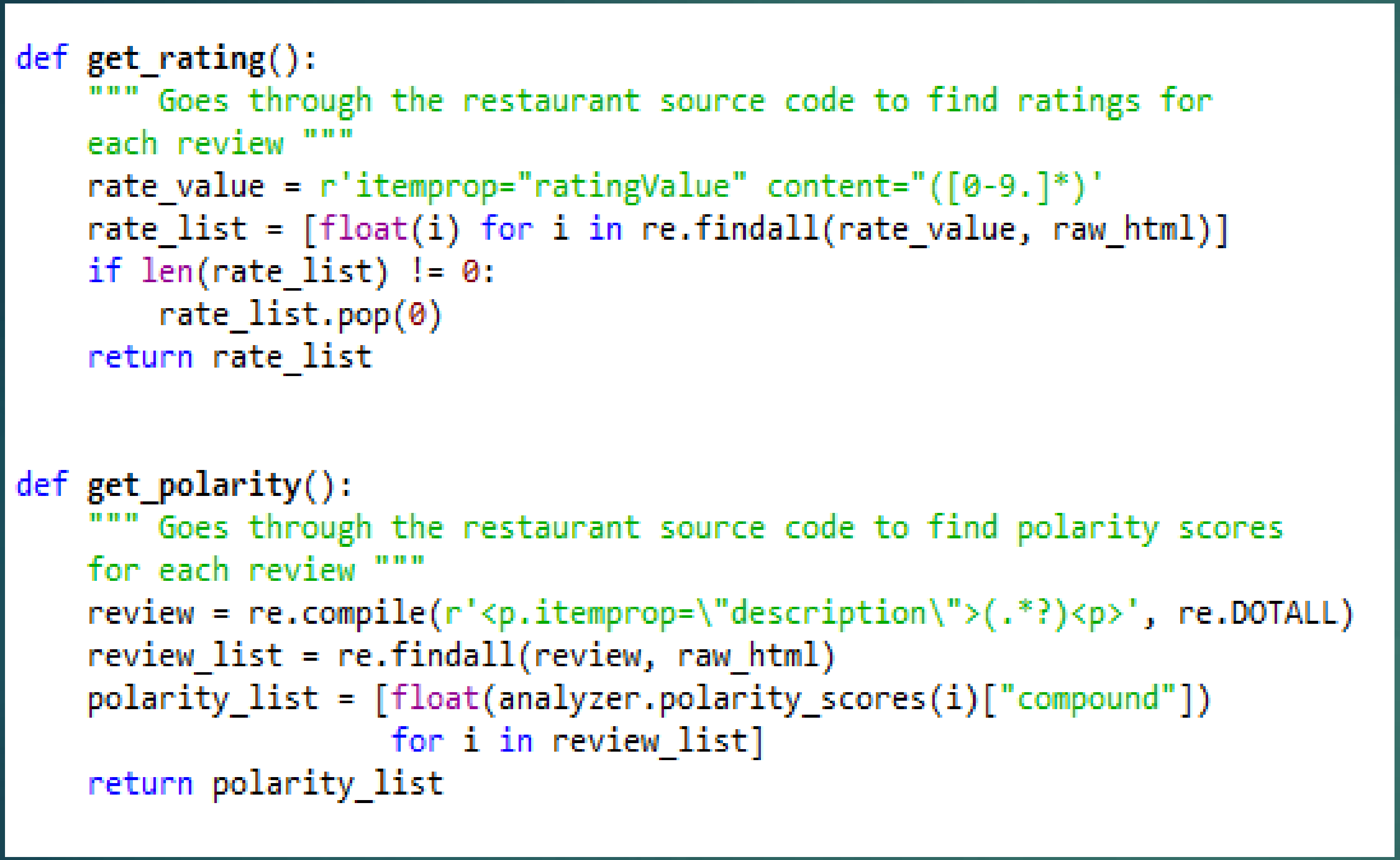
```python
def get_rnlp(businessId, price):
    """ Retrieves webpage source code for each restaurant and inputs into
    text files separated by price level """
    yelp_url = 'https://www.yelp.com/biz/' + businessId + '?osq=Restaurants'
    headers = {'user-agent': 'Jaimie Choi (jchev95@hotmail.com)'}
    response = r.get(yelp_url, headers=headers)
    with open('Final_Project/raw_script-' + price + '/' + businessId + '.txt',
              'w', encoding='utf8') as scrape_file:
        scrape_file.write(response.text)
```

1) Collected business ID for restaurants in Salt Lake City from Yelp API

2) Used the restaurant names to scrape each restaurant's page source

3) Separated the text files by the restaurant's price range

Data Cleaning



```python
def get_rating():
    """ Goes through the restaurant source code to find ratings for
    each review """
    rate_value = r'itemprop="ratingValue" content="([0-9.]*)'
    rate_list = [float(i) for i in re.findall(rate_value, raw_html)]
    if len(rate_list) != 0:
        rate_list.pop(0)
    return rate_list


def get_polarity():
    """ Goes through the restaurant source code to find polarity scores
    for each review """
    review = re.compile(r'<p.itemprop=\"description\">(.*?)<p>', re.DOTALL)
    review_list = re.findall(review, raw_html)
    polarity_list = [float(analyzer.polarity_scores(i)["compound"])
                     for i in review_list]
    return polarity_list
```

1) Used regex to find the rating scores and reviews for each restaurant

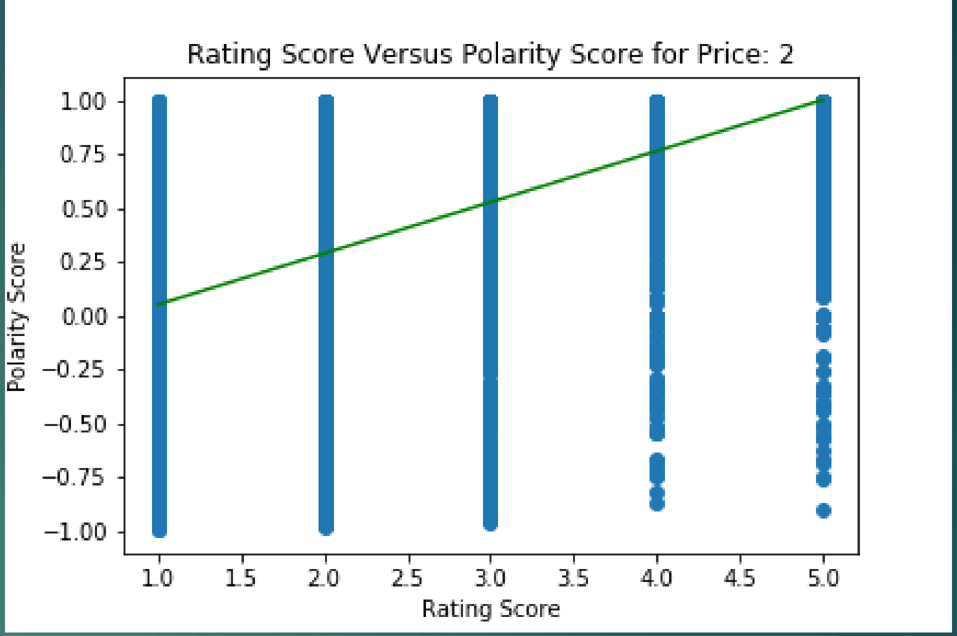
2) Used NLTK polarity score analyzer to find polarity scores for each restaurant

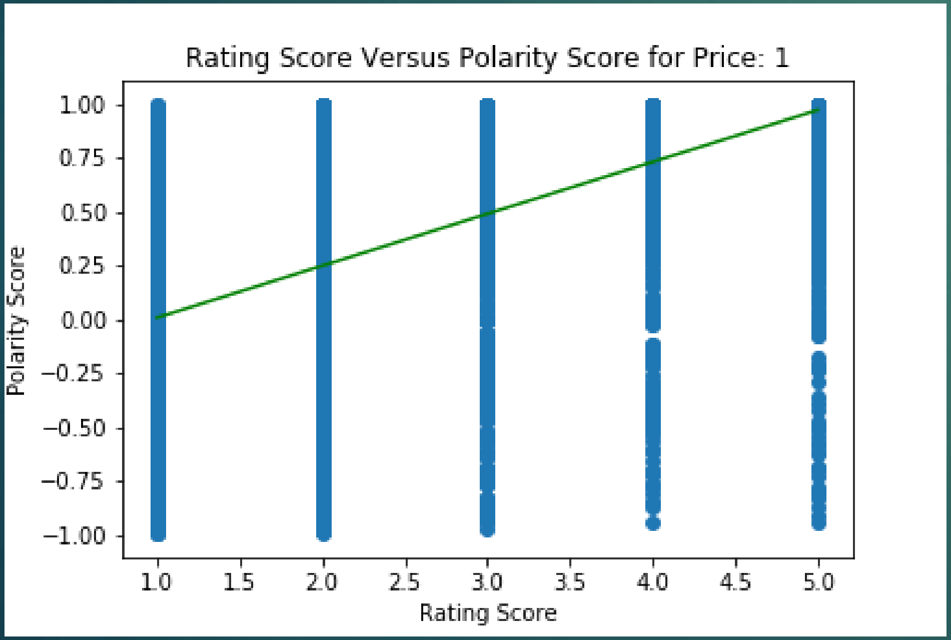
Statistical Results

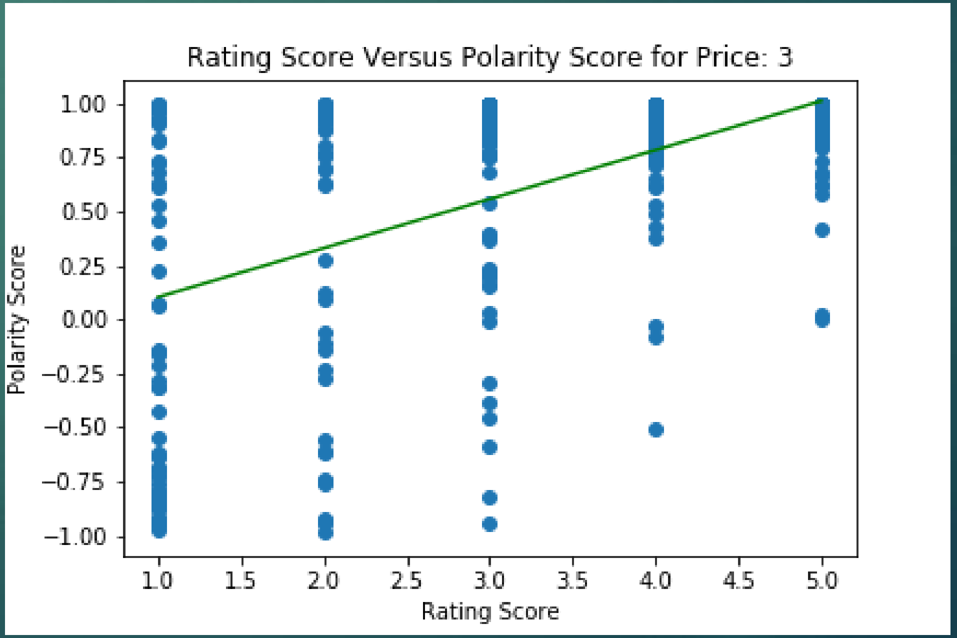
R-Value

1) 0.6292***
2) 0.6241***
3) 0.6275

*** statistically significant ***

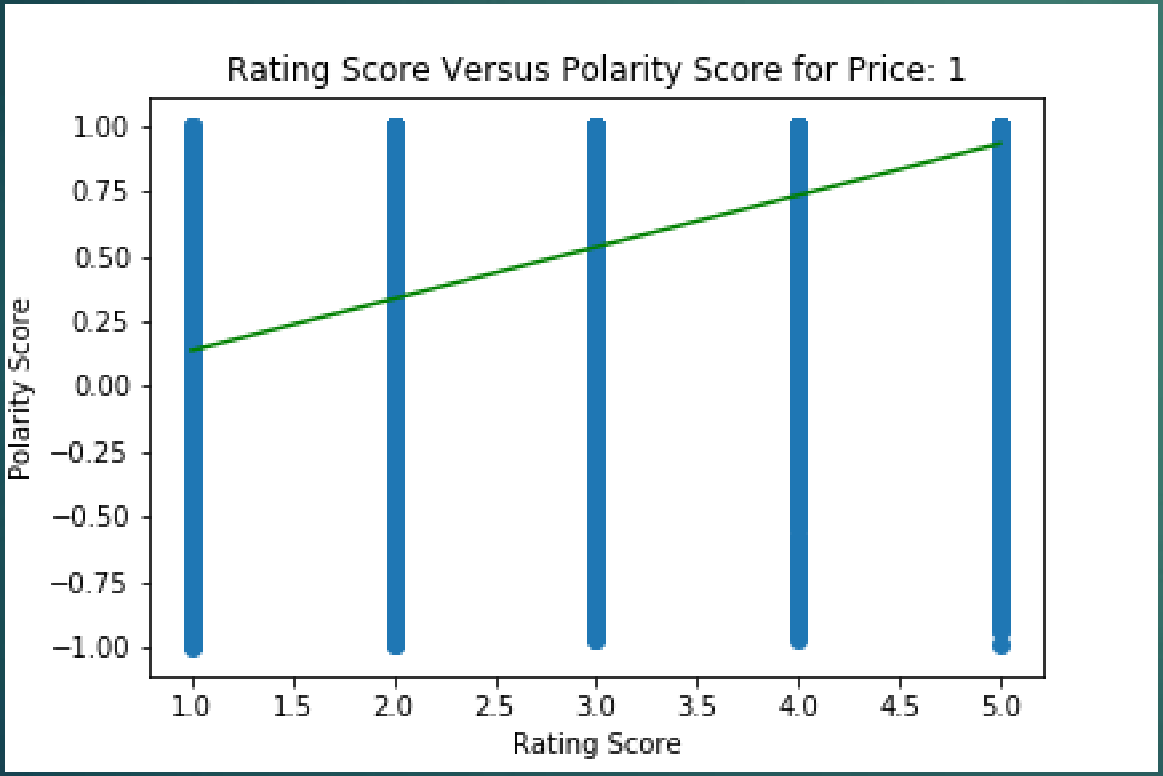




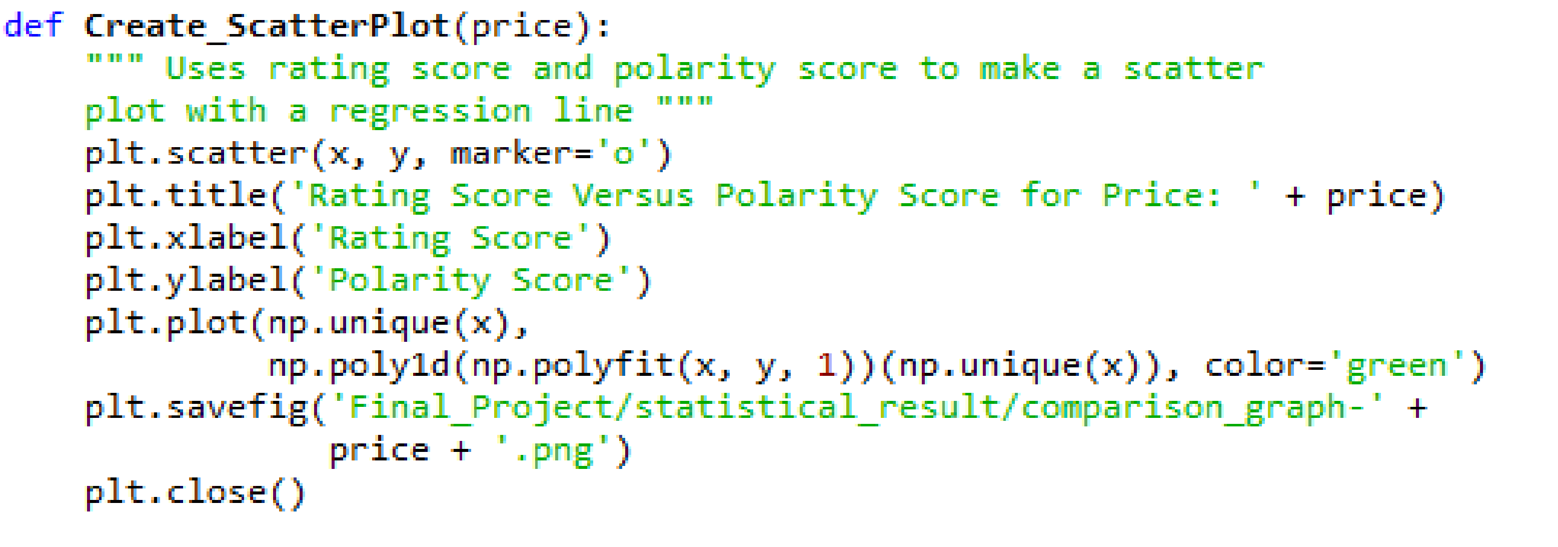


Statistical Result
Price 1 for 3 cities

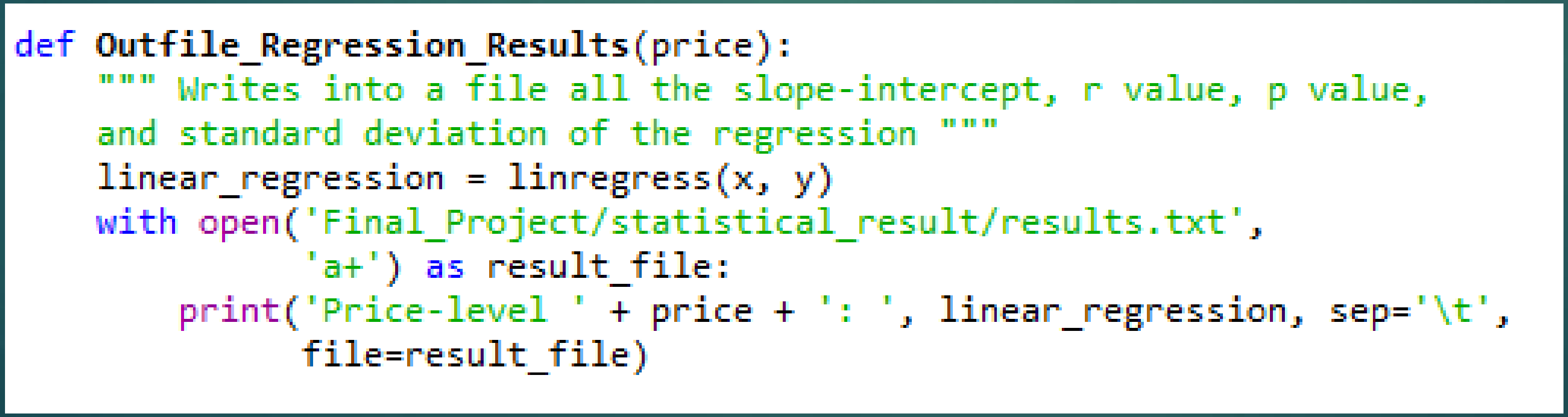
- Los Angeles, Chicago, New York City (0.524)***



Code for the Statistical Result



```python
def Create_ScatterPlot(price):
    """ Uses rating score and polarity score to make a scatter
    plot with a regression line """
    plt.scatter(x, y, marker='o')
    plt.title('Rating Score Versus Polarity Score for Price: ' + price)
    plt.xlabel('Rating Score')
    plt.ylabel('Polarity Score')
    plt.plot(np.unique(x),
             np.poly1d(np.polyfit(x, y, 1))(np.unique(x)), color='green')
    plt.savefig('Final_Project/statistical_result/comparison_graph-' +
                price + '.png')
    plt.close()
```



```python
def Outfile_Regression_Results(price):
    """ Writes into a file all the slope-intercept, r value, p value,
    and standard deviation of the regression """
    linear_regression = linregress(x, y)
    with open('Final_Project/statistical_result/results.txt',
              'a+') as result_file:
        print('Price-level ' + price + ': ', linear_regression, sep='\t',
              file=result_file)
```

Conclusion

- Significant positive relationship between rating score and polarity score
- The restaurant price level does not seem to affect the relationship
- How to extend this research:
- Collect yelp review data from all over United States for a more
- reliable statistic result