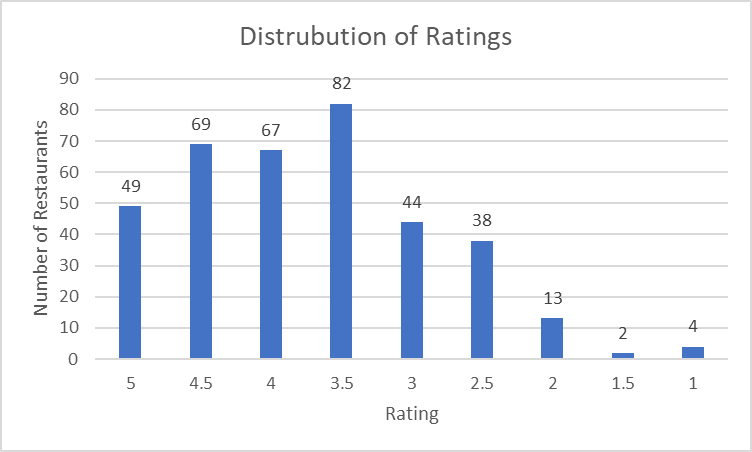
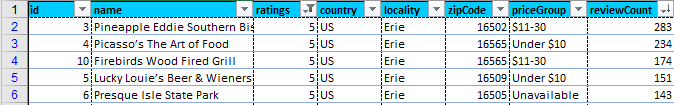
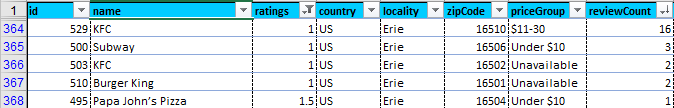
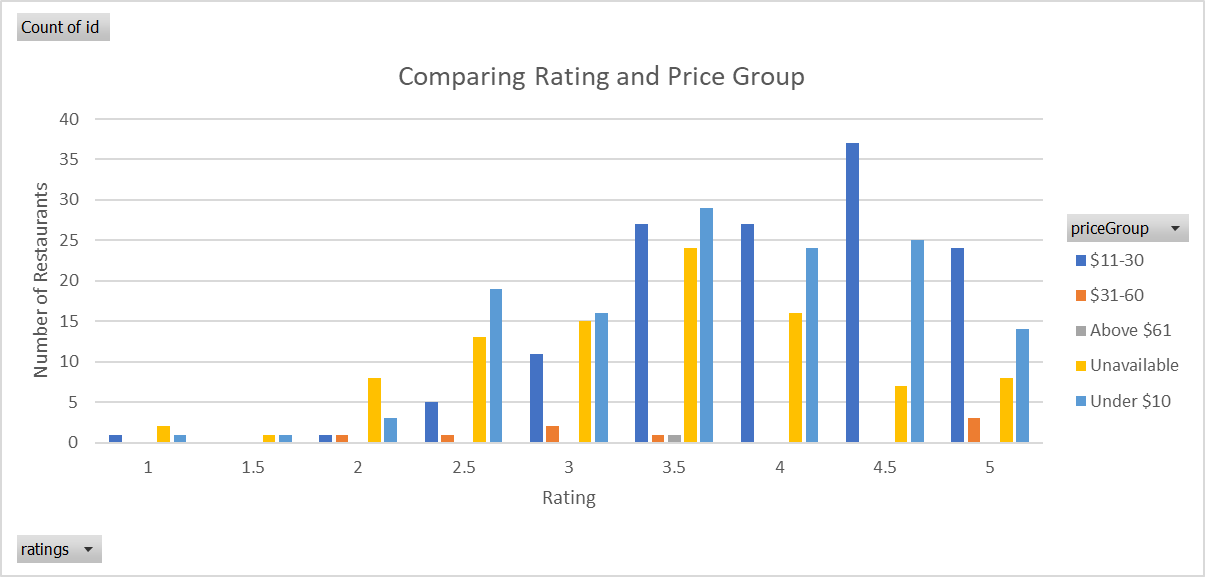
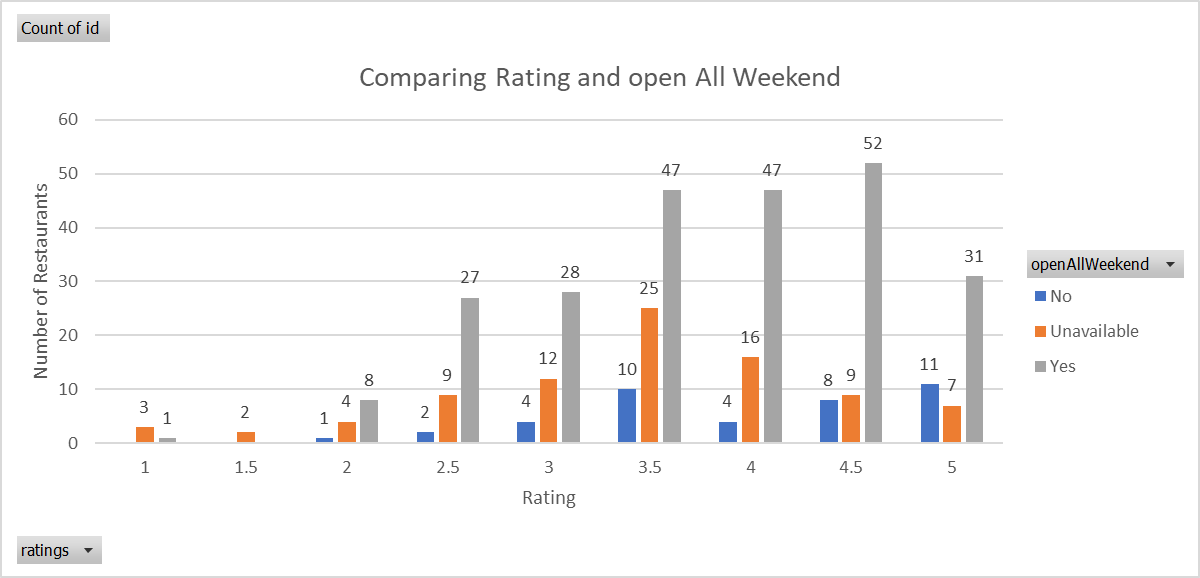
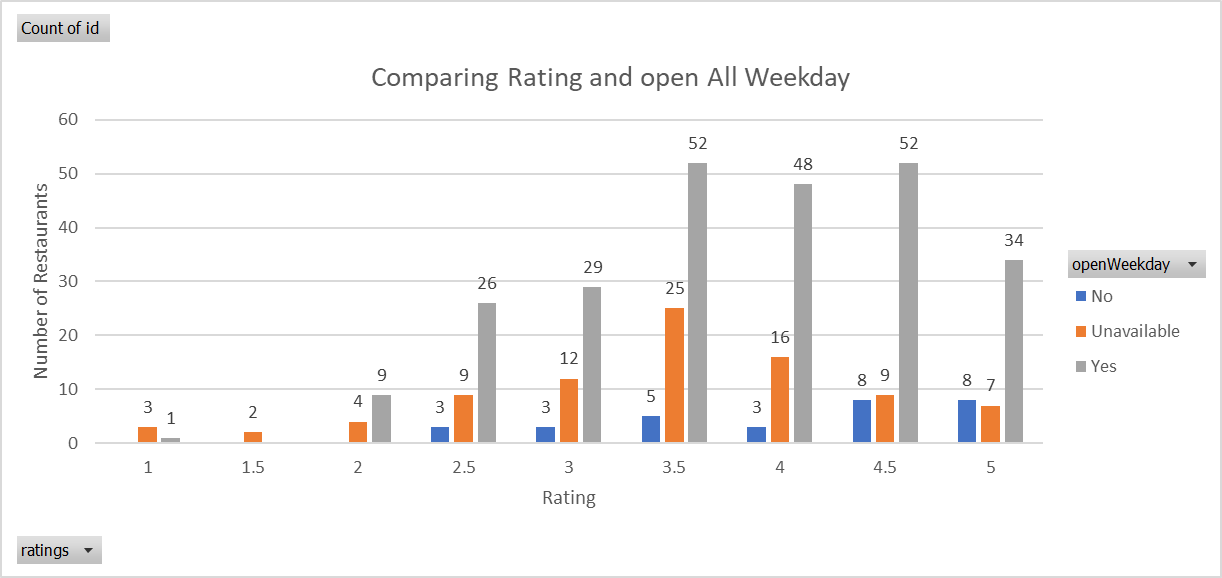
* 1. Translate the business requirements into data analytics problem to define the scope of your data analysis. *For this question, you need to state what analytical approach will be most suitable for addressing each business problem.* **[10 points]**

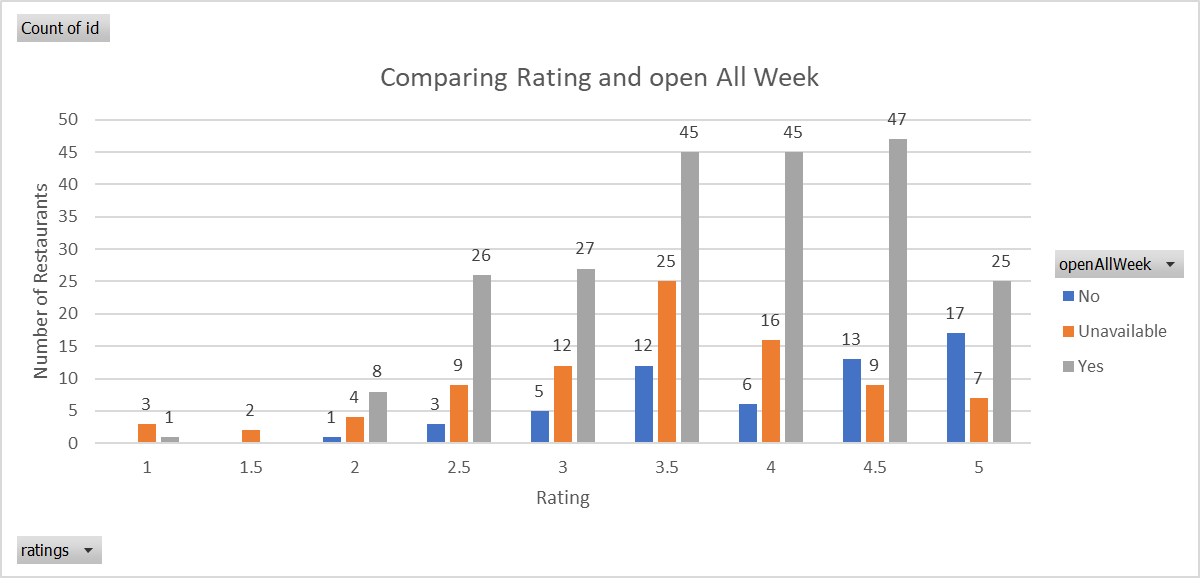
Since there is a technique that can be used to address all three questions, we can conclude that there can be an analytical approach to this question. In order to produce a list of the top 5 best performing and lowest 5 worst performing restaurants, the rating score will be used to perform an analysis. The ratings will be sorted from highest (5) to lowest (1), considering the number of reviews since more reviews mean a more credible and accurate view of the data. In order to know the characteristics of the restaurants captured in the dataset, the price group and opening hours will be used to perform an analysis. A pivot table will be used to sort the data that are common between the restaurants with high ratings and low ratings to determine the related characteristics.

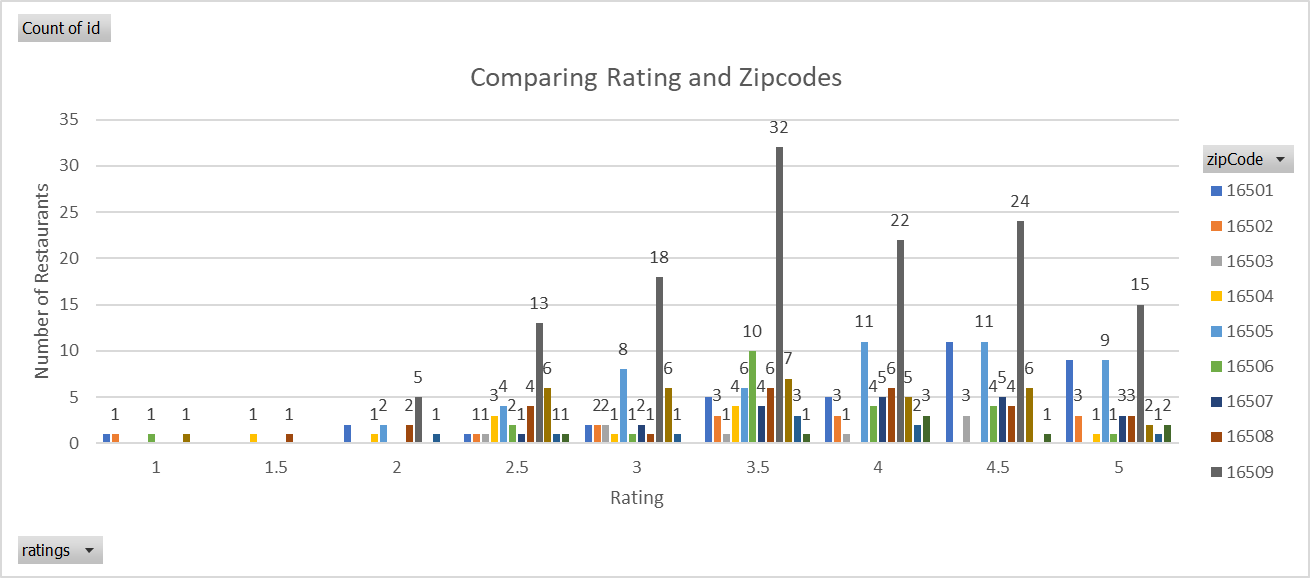
* 1. Describe the data type for each variable in your dataset. **[10 points]**

|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| ID | identifier |
| Name | identifier |
| reviewCount | Numerical discrete time-series data |
| ratings | Categorical ordinal time-series data |
| country | Categorical nominal cross-sectional data |
| locality | Categorical nominal cross-sectional data |
| zipCode | Categorical nominal cross-sectional data |
| priceGroup | Categorical ordinal cross-sectional data |
| region | Categorical ordinal cross-sectional data |
| openAllWeek | Categorical ordinal cross-sectional data |
| openAllWeekend | Categorical ordinal cross-sectional data |
| openWeekday | Categorical ordinal cross-sectional data |

* 1. Show the distribution of the star rating score of each restaurant using **at least two** different types of charts (e.g., you can use bar chart, histogram or pie chart as you see fit). **[20 points]**  
     For the bar chart, the COUNTIF function is used to collect the total number of restaurants with a certain rating. For the pie chart, the total number of restaurants with a certain rating will be divided by the total number of restaurants in the dataset to produce a percentage. 
  2. Conduct what you consider to be the appropriate descriptive analysis to answer the following questions:
     1. List the top 5 restaurants around Erie, PA, based on their star rating score. **[10 points]**I filtered the ratings to show those that have “5”, which is the highest possible rating, then I sorted the results by highest to lowest reviewCount to get the top 5 restaurants that have the highest rating of 5 from the most reviews.
     2. List the bottom 5 restaurants around Erie, PA, based on their star rating score. **[10 points]**I did the same as above for the bottom 5 restaurants, except I filtered ratings for “1.5” and “1” since there were not 5 restaurants when I only filtered for “1”.
  3. Conduct diagnostic analysis and using outputs from your analysis, report on how the characteristics of restaurants are ***related*** to their star rating score. (**Hint**: you may need to use all the diagnostic techniques discussed in class including crosstabs, pivot tables, and correlation analysis) **[30 points]**Through comparing the ratings to the price groups, one can assume that the lower the price is for the restaurant, the higher its rating will be, indicating a negative relationship where the price drops while the ratings rises. This can be seen from the overall higher ratings for lower price groups in the bar chart.

Through comparing the ratings and the opening days of the restaurants, one can assume that the restaurants that are open all weekday, all weekend, and all week respectively have higher ratings than those that do not. This can be seen in the bar graphs, where the “Yes” bars for all three criteria (open all weekday, open all weekend, open all week) are significantly larger than the “No” bars for each respectively graph.   




Initially, I was going to compare the locality, the region, and the country of the restaurants to their ratings, however, given that these three criteria are all consistent throughout the dataset, I opted for using the zip code instead since the zip codes are offer a general understanding of where the restaurants are located at within the same region and locality.

From the bar graph comparing ratings and zip codes, one can assume there is a correlation between the ratings and the zip codes. Overall, the restaurants in the region with the zip codes 16503 and 16505 have a higher rating than restaurants in other regions.

* 1. What recommendations will you include in your report based on your results? **[10 points]**Based on my results, I would recommend restaurants to open up in the regions with zip code 16503 and 16506, to be in the price range of around $11 - $30, and to be open either all weekdays, all weekend, or all week.