Visualization of Yelp Dataset 2019: Interactive Analysis and Comparison of Catering Businesses

Yuzhe Yang

In this project, we performed various interactive analyses and comparisons using the Business.json and Review.json files of the open dataset from the Yelp Dataset Challenge by implementing data visualization techniques we learnt from this course through the semester.

Preprocessing data in Business.json

The Yelp dataset contains high-dimensional data. In order to extract the most relevant features, we first set our focus on the catering businesses so that we preprocessed the Yelp dataset to create the data frame which fits our goals accordingly. After the preprocessing, we filtered 59,371 restaurants from the original dataset for the subsequent data visualization and analysis (Fig.1).


```
{"business_id":"1SWheh84yJXfytovILXOAQ","name":"Arizona Biltmore Golf Club","address":"2818 E Camino Acequia Driv
e","city":"Phoenix","state":"AZ","postal_code":"85016","latitude":33.5221425,"longitude":-112.0184807,"stars":3.0,"re
view_count":5,"is_open":0,"attributes":{"GoodForKids":"False"},"categories":"Golf, Active Life","hours":null}
```


```
59,371 restaurants in the dataset.
                       address  \
0          30 Eglinton Avenue W
1     10110 Johnston Rd, Ste 15
2       2450 E Indian School Rd
3               5981 Andrews Rd
4  1775 E Tropicana Ave, Ste 29

                                    attributes             business_id  \
0  {'RestaurantsReservations': 'True', 'GoodForMe...  QXAEGFB4oINsVuTFxEYKFQ
1  {'GoodForKids': 'True', 'NoiseLevel': 'u'avera...  gnKjwL_1w79qoiV3IC_xQQ
2  {'RestaurantsTakeOut': 'True', 'BusinessParkin...  1Dfx3zM-rW4n-31KeC8sJg
3  {'RestaurantsPriceRange2': '2', 'BusinessAccep...  fweCYi8FmbJXHCqLnwuk8w
4  {'OutdoorSeating': 'False', 'BusinessAcceptsCr...  PZ-LZzSlhSe9utkQYU8pFg

                                    categories               city  \
0  Specialty Food, Restaurants, Dim Sum, Imported...        Mississauga
1                Sushi Bars, Restaurants, Japanese          Charlotte
2  Restaurants, Breakfast & Brunch, Mexican, Taco...            Phoenix
3        Italian, Restaurants, Pizza, Chicken Wings  Mentor-on-the-Lake
4                             Restaurants, Italian          Las Vegas

                                               hours  is_open   latitude  \
0  {'Monday': '9:0-0:0', 'Tuesday': '9:0-0:0', 'W...        1  43.605499
1  {'Monday': '17:30-21:30', 'Wednesday': '17:30-...        1  35.092564
2  {'Monday': '7:0-0:0', 'Tuesday': '7:0-0:0', 'W...        1  33.495194
3  {'Monday': '10:0-0:0', 'Tuesday': '10:0-0:0', ...        1  41.708520
4                                               None        0  36.100016

    longitude                         name postal_code  review_count  stars  \
0  -79.652289   Emerald Chinese Restaurant     L5R 3E7           128    2.5
1  -80.859132  Musashi Japanese Restaurant       28210           170    4.0
2 -112.028588                   Taco Bell       85016            18    3.0
3  -81.359556               Marco's Pizza       44060            16    4.0
```

Fig. 1 Data structures of the business.json file.

Q1. What are the distributions of the catering businesses collected in Yelp dataset at the world, country, and city levels respectively?

We know from the data preprocessing that there are 59,371 catering businesses in total. To further explore the distribution of these businesses, we extracted data by using keywords such as ‘country’, ‘state’, ‘city’, etc. to perform statistical analysis and visualization of the data to get more information.

The Yelp dataset contains data from USA and Canada. The state distribution includes states in both countries. The bar-plot (Fig. 2) shows the sorted list of the 16 states with the most counts of catering businesses. Most of the businesses are distributed in 2 Canadian states and 7 states in USA.

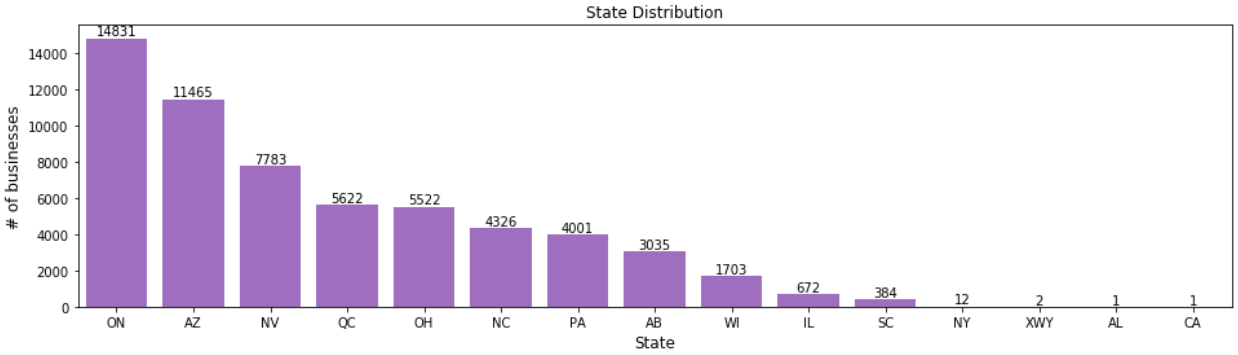

Fig. 2 Catering businesses distribution among states.

Next, we implemented GeoMap technique (Fig.3) to show the above distribution but only in the territory of USA. The color intensities correlate with the businesses counts. The GeoMap provides a more direct way to understand the counts and distributions.

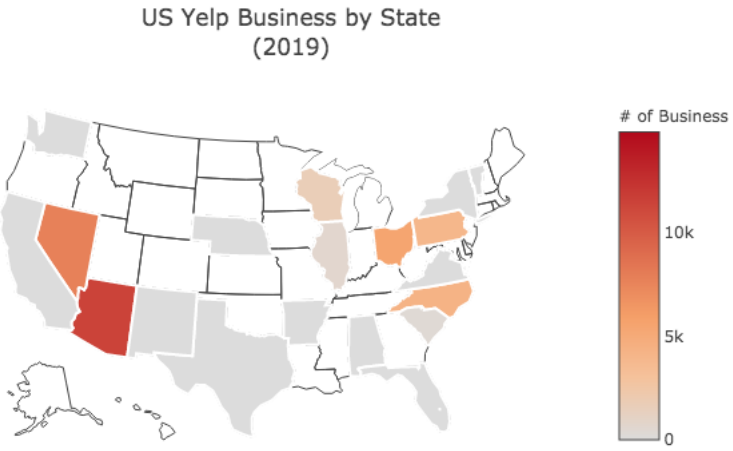

Fig. 3 Interactive GeoMap of catering businesses distribution among US states.

We want to further explore the businesses distributions at the city-level. There are 782 cities in the dataset therefore it is not a good idea to show all the cities in one bar-plot. Hence, we only show the top 20 cities with the most business counts (Fig. 4). In order to show the information of all the cities in one graph, we again implemented the GeoMap. Yelp dataset provides the longitude and latitude of each business, so we utilized the coordinate information to map the cities with counts and ranking range information (Fig.5).

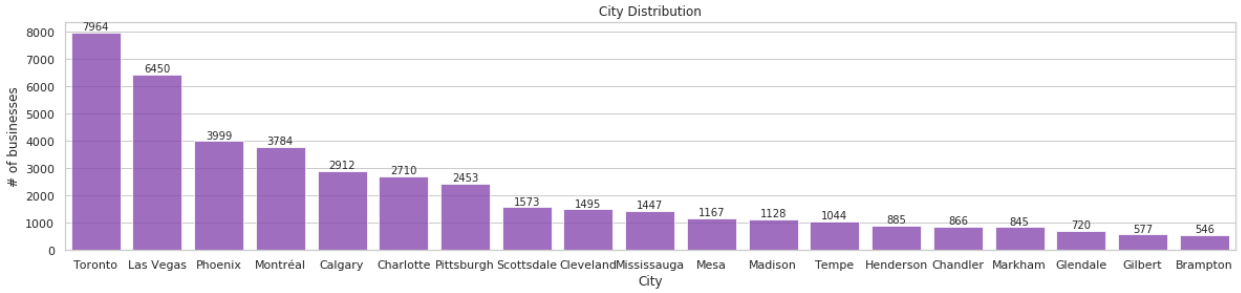

Fig. 4 Catering businesses distribution among cities.

In the GeoMap, we were able to show all the data of 782 cities. This viz also reveals a fact that Yelp collected 10 major metropolitan areas. Most of the cities are satellite cities. From the GeoMap we can also make estimation that if we merge the cities to metropolitan areas, Toronto and Phoenix may be the top 2 areas with most business counts (Fig. 5).

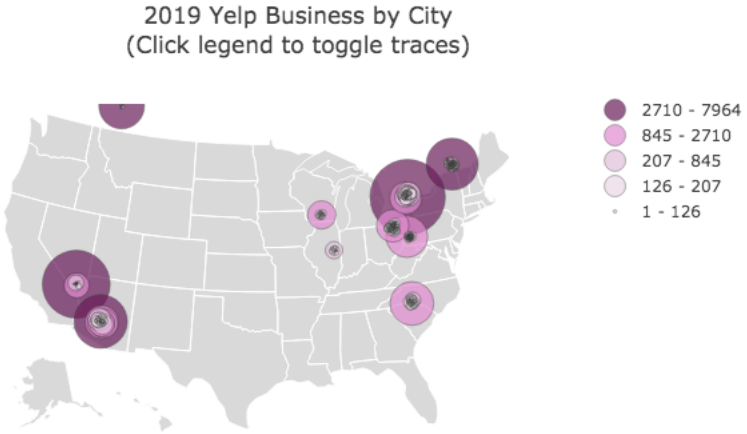

Fig. 5 Interactive GeoMap of catering businesses distribution among cities. Bubble sizes correlate with the business counts, colors show the range ranking of the represented city.

We then picked five cities in USA: Las Vegas, Phoenix, Charlotte, Pittsburgh, and Cleveland, and plot business distribution at the street-level. These visualizations also provide urban planning information: the western cities are block style while the eastern old cities are circle style (Fig. 6). Taking Las Vegas as an example, we can observe a dense area align with the Strip street which is the main street in the downtown Las Vegas. It makes sense that restaurants mostly locate around this area.

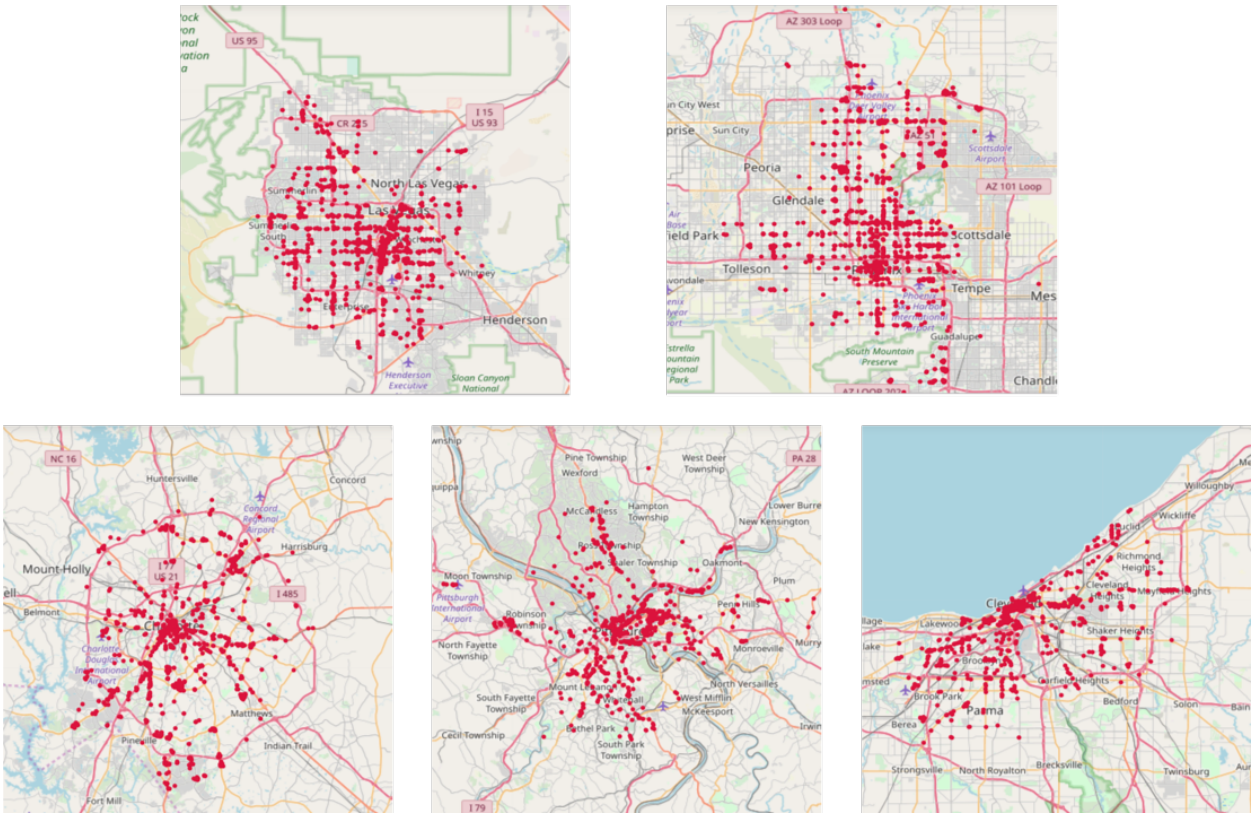

Fig. 6 Catering businesses distribution at city street level.

After adjusting the Q1, we have a clearer picture of how the catering businesses distribute as well as gotten more familiar with the dataset for further in-depth visualization and analyses of five cities in USA: Las Vegas, Phoenix, Charlotte, Pittsburgh, and Cleveland.

Q2. What are the ratings of businesses in the five cities?

To answer this question, we first visualize the overall star rating of all catering businesses in the Yelp dataset (Fig. 7). We explore the star rating counts and distributions of the whole dataset. In order to show clusters of good and bad catering business in a GeoMap presentation, we defined a color palette to assign different colors to each rating corresponding column. The warmer the higher rating, while the cooler the lower rating it has.

Fig. 7 Star rating distribution of all catering businesses.

We want to explore if there are differences in the rating trends among the 5 cities. We first plot the star rating distributions of each city to get a general idea of most of the ratings are in 3.5-4.0 star range. (Fig. 8).

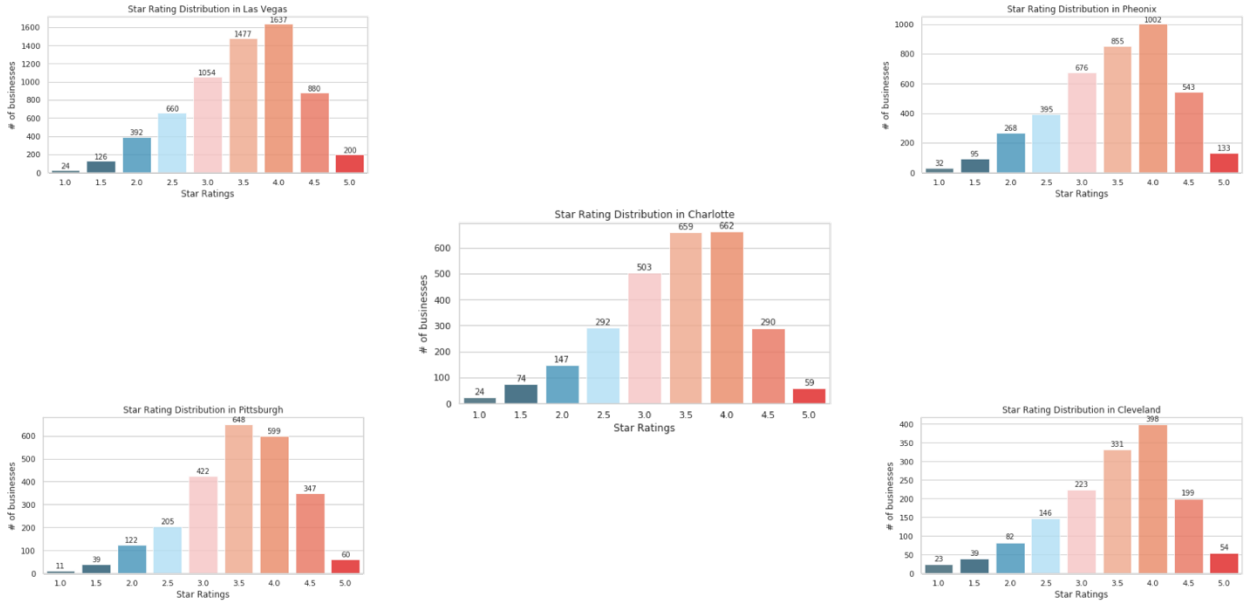

Fig. 8 Star rating distribution of each city.

We then performed statistical analysis by normalizing rating counts to total counts and found that people in Cleveland tend to give extreme high or low ratings. On the other hand, it seems that Charlotte and Pittsburgh do not have many great restaurants or people in those cities have stricter rating criteria (Fig. 9).

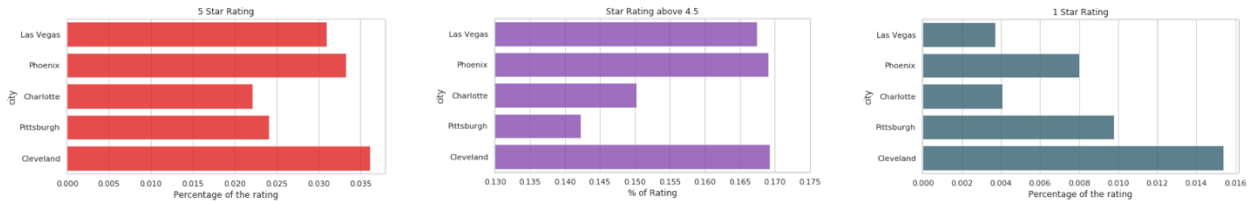

Fig. 9 Normalized star rating distribution of indicated star rating score.

We next visualized the restaurant distribution again at the street-level but this time with rating information embedded as circle colors (Fig. 10). This visualization provides a guide to the reader on which area of the city have more highly rated restaurants. In order to show the dots clearer, we use black and white map to enhance the contrast.

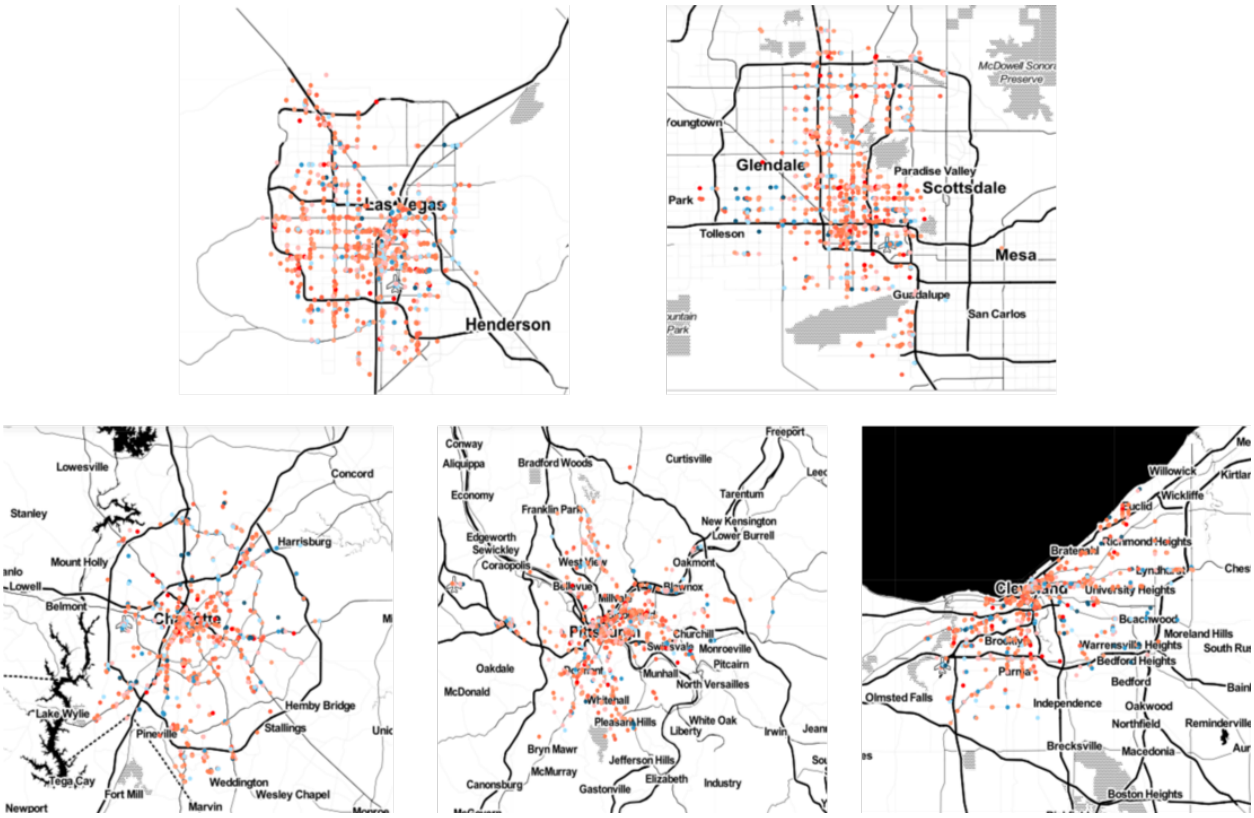

Fig. 10 Catering businesses city distribution with rating information as represented by dot color.

Q3 What are the top categories of catering businesses in different cities? Do categories have correlations with ratings and counts of reviews?

We first looked at the overall categories of the whole dataset. There are 761 different types/categories of catering businesses in Yelp dataset. The bar-plot shows the top 20 categories of businesses (Fig. 11).

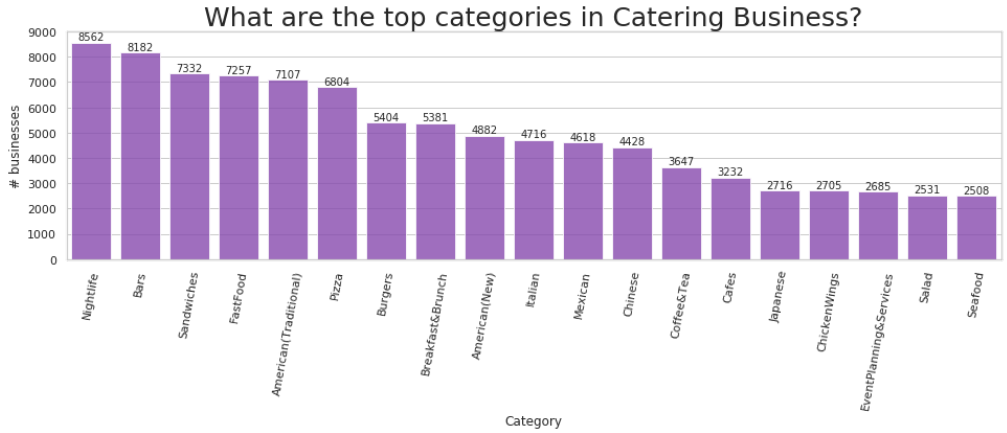

Fig. 11 Top 20 categories in catering businesses.

We seek to further explore if the top categories vary in different cities. The bar-plot data visualization (Fig. 12) reveals that in the western cities of the US (Las Vegas and Phoenix) the largest non-American food business are Mexican food, while in the eastern/New England area cities of US (Pittsburgh and Cleveland) the largest non-American food business is Italian food. This observation might due to in the western cities there are more Mexican immigrants whereas in the eastern cities indicated above there are many descants of earlier European immigrants or simply because the cities in the east are closer to Europe. In order to test our hypothesis, we viz two more Canadian cities – Toronto and Montreal. In Toronto, Chinese restaurants are next to the traditional Canadian food. Toronto does have a lot of Chinese immigrants. In Montreal, it is clearer that French food counts for the largest number of businesses, as Montreal is actually a French spoken district of Canada! Therefore, we can conclude that the ancestral homelands of the population play a critical role in determining the popularity of restaurant style.

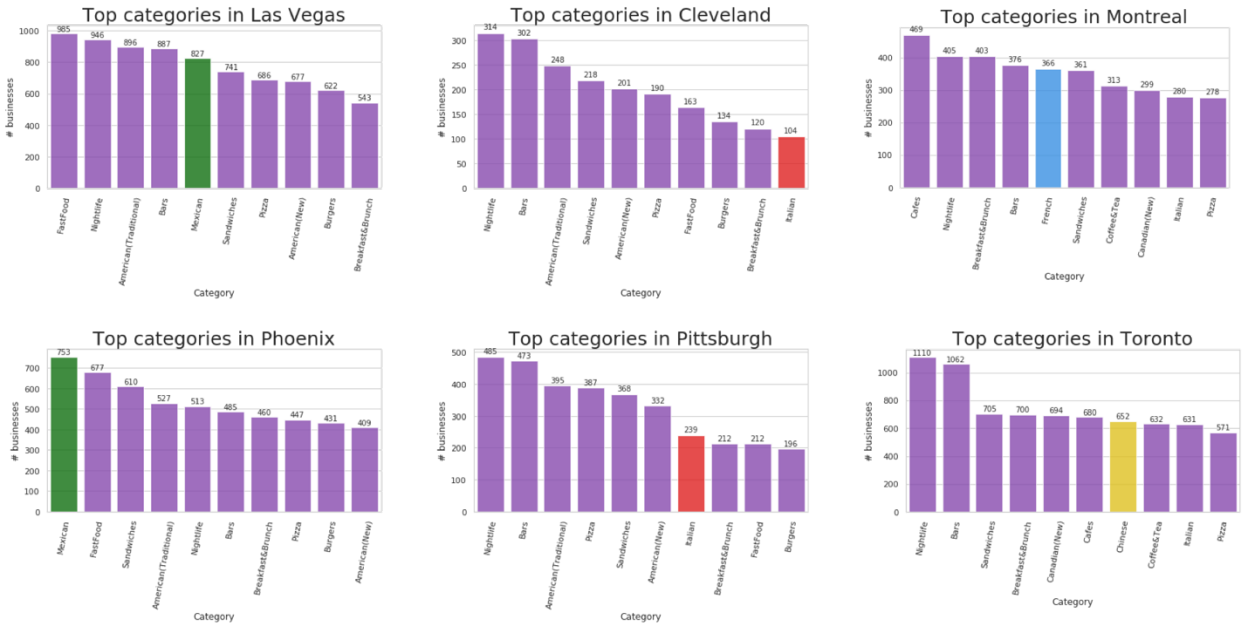

Fig. 12 Top 10 categories in catering business of indicated cities.

We also seek to find if the star rating has correlation with business categories by plotting average star ratings against business categories (Fig. 13). Fine-dining categories such as Japanese, sushi bar, and seafood categories tend to have higher average ratings while fast food including burgers, chicken wings, and pizza tend to have lower average ratings.

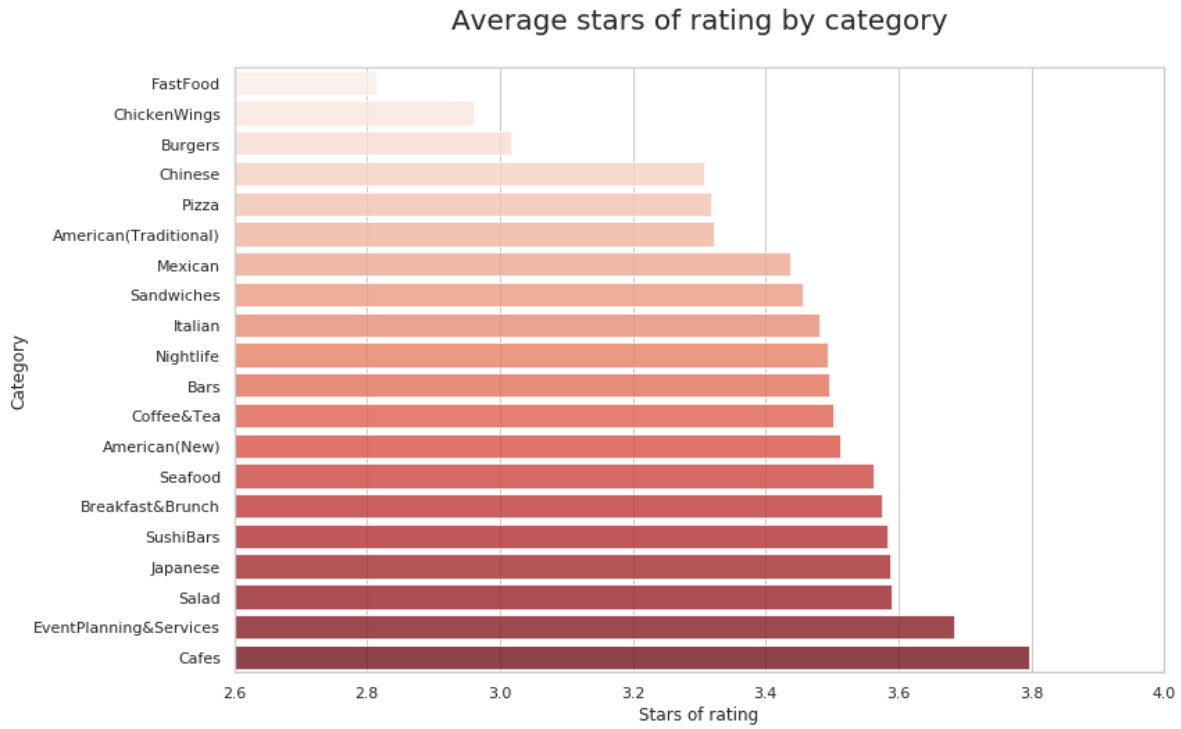

Fig. 13 Average star rating vs businesses categories.

We also implemented a bubble plot to embed rating counts into the presentation to check the confidence of the average star rating result (Fig.14). Some of the highly rated businesses such as seafood and SushiBars have significantly fewer rating counts than fast food. In general, fewer people can afford going to fine dining restaurants and the riches don't have time to leave ratings on Yelp App.

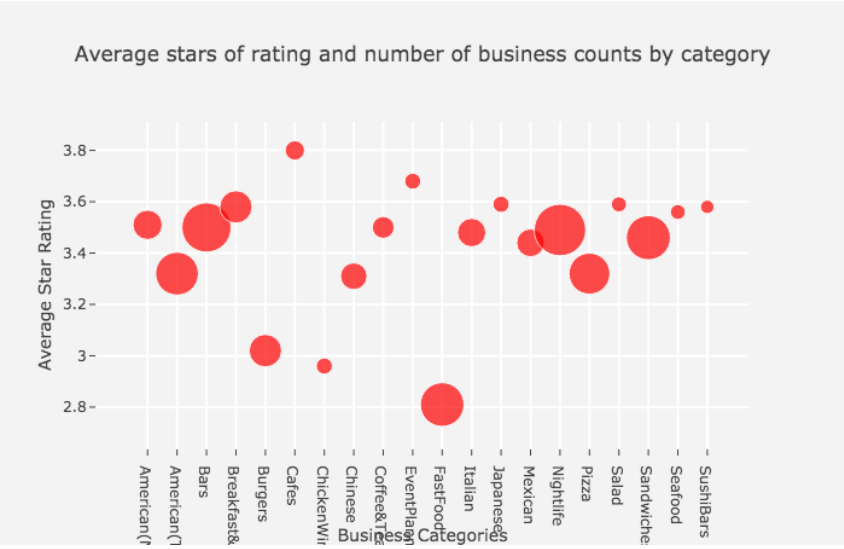

Fig. 14 Average star rating vs businesses categories with average counts of ratings represented by size of bubble.

If we take a closer look at the five selected cities, we can find that each city has different preference of the best rating categories however the worse rating category is always fast food related catering businesses. People in Las Vegas love Asian fusions since this is a tourism city and probably is open to exotic food. In the rest 4 cities, café is consistently one of the top-rated types of restaurant (Fig.15). Take the rating counts into to our scope of view (Fig.16), it indicates the similar pattern as in Fig.14. The rating counts might be positively correlated to the number of catering businesses.

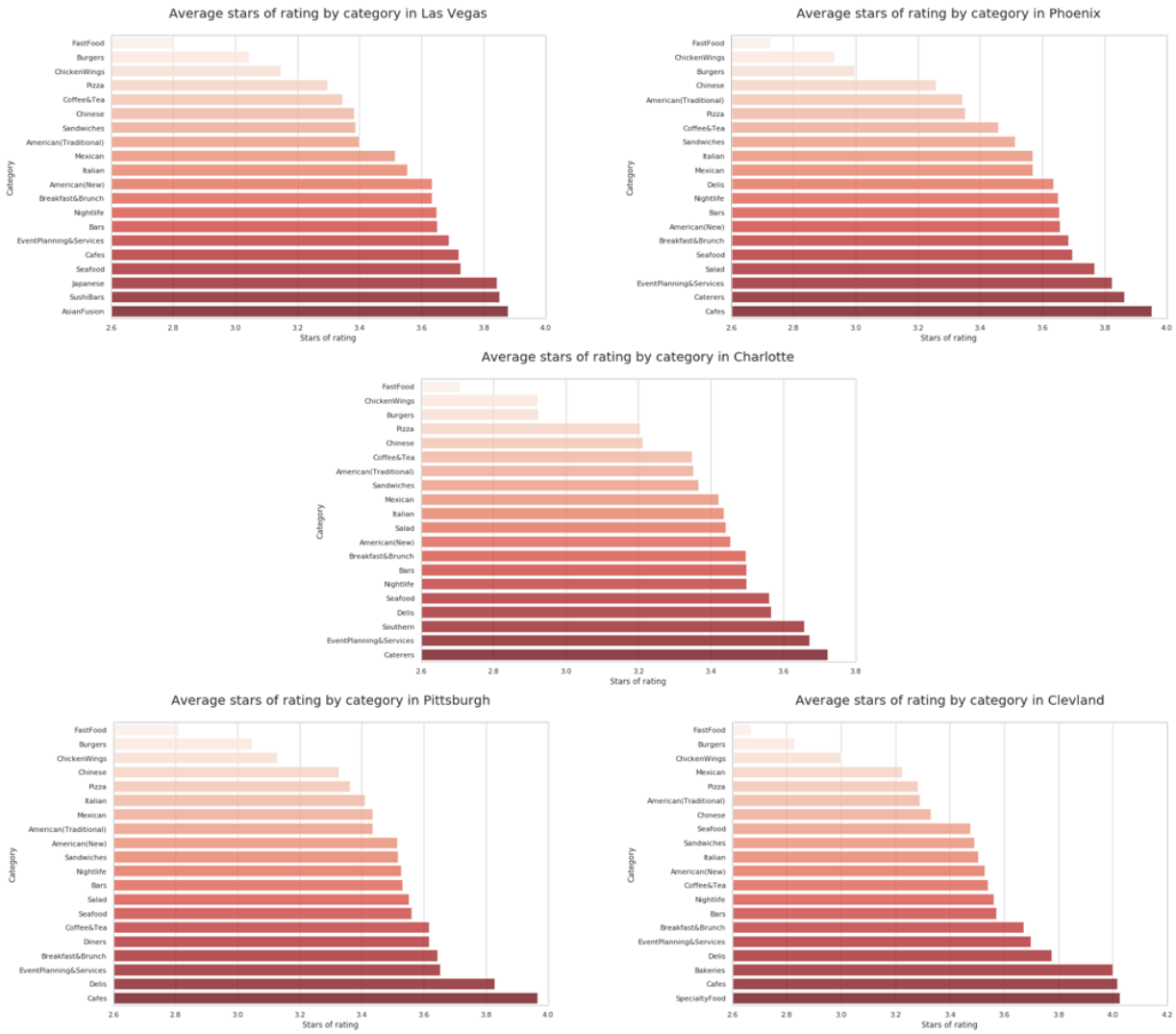

Fig. 15 Average star rating vs businesses categories at city level.

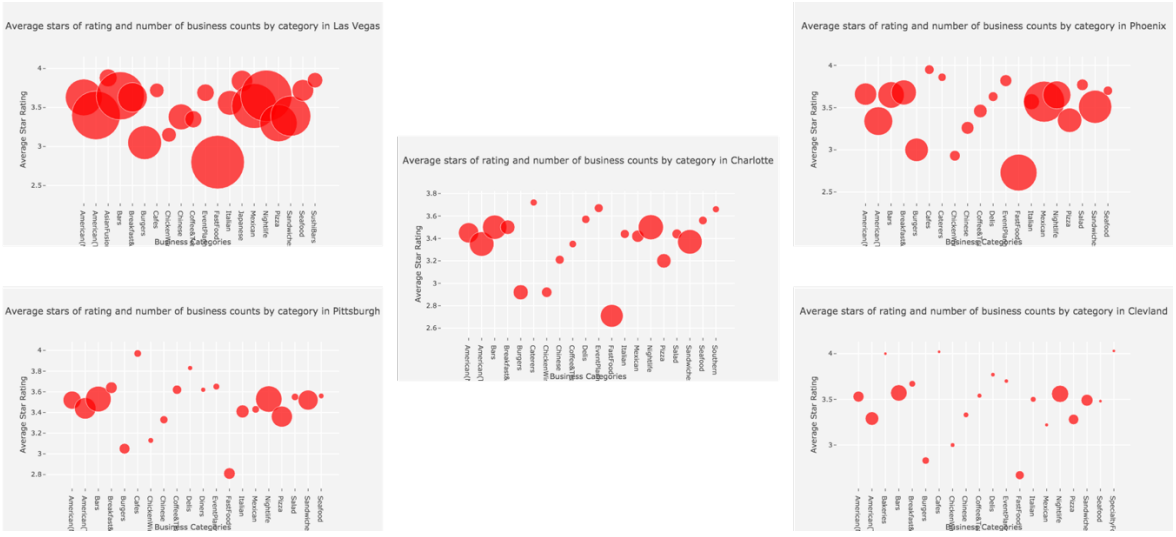

Fig. 16 Fig. 14 Average star rating vs businesses categories at city level with average counts of ratings represented by size of bubble.

Q4 Does average star rating have correlation with number of reviews? Do people tend to leave review to certain star rated restaurants?

To answer these two questions, we plot review count against average star rating of the top counted business categories and compare the bar-plot (Fig.17) with Fig. 13. The visualization indicated that people tend to give high ratings while leave some comments; however, for the bad rated business people usually do not leave a word. To get even better visualization for this correlation, we did bubble plot as shown in Fig. 18.

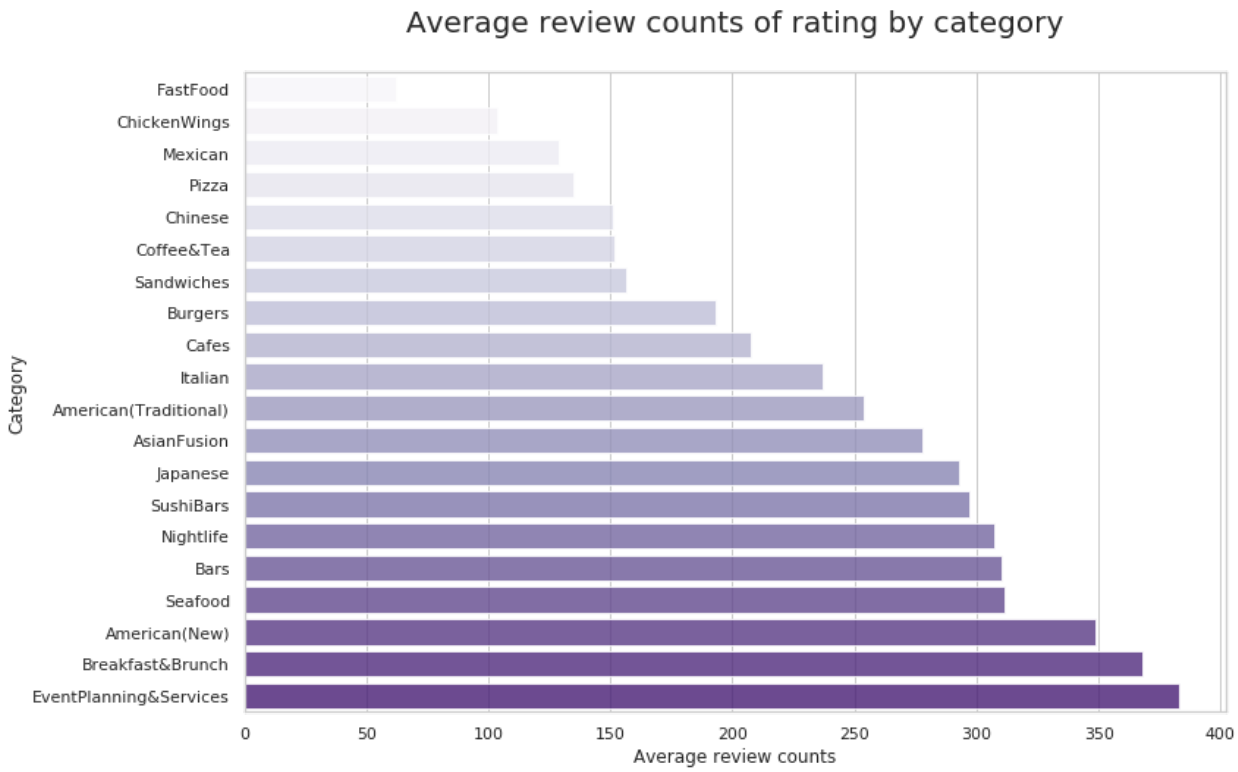

Fig. 17 Average review counts vs business categories.

We also took a deeper look at the data of the five cities we selected for the visualization project. The absolute number of averaged review counts seem to have positive correlation with the total business counts across the five selected cities (Fig.19). We can learn from the bar-plot that people won't spend time to leave comments on fast food related businesses. The observations are more straightforward if we look at the bubble plot showed in Fig.20.

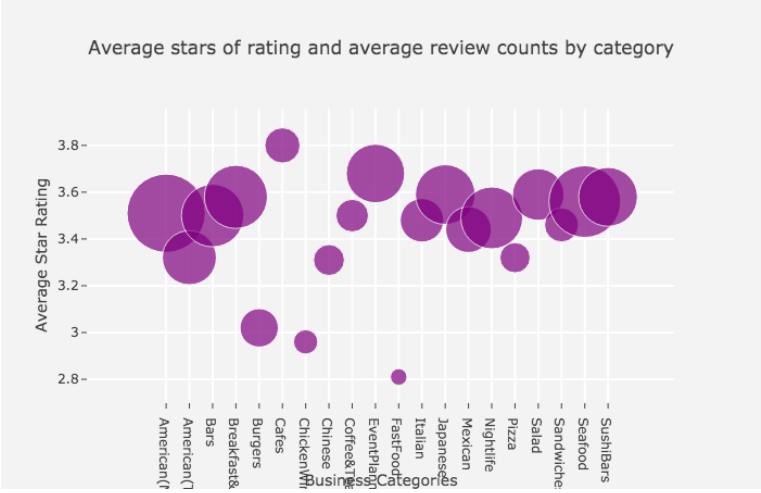

Fig. 18 Average star rating vs businesses categories with average counts of reviews represented by size of bubble.

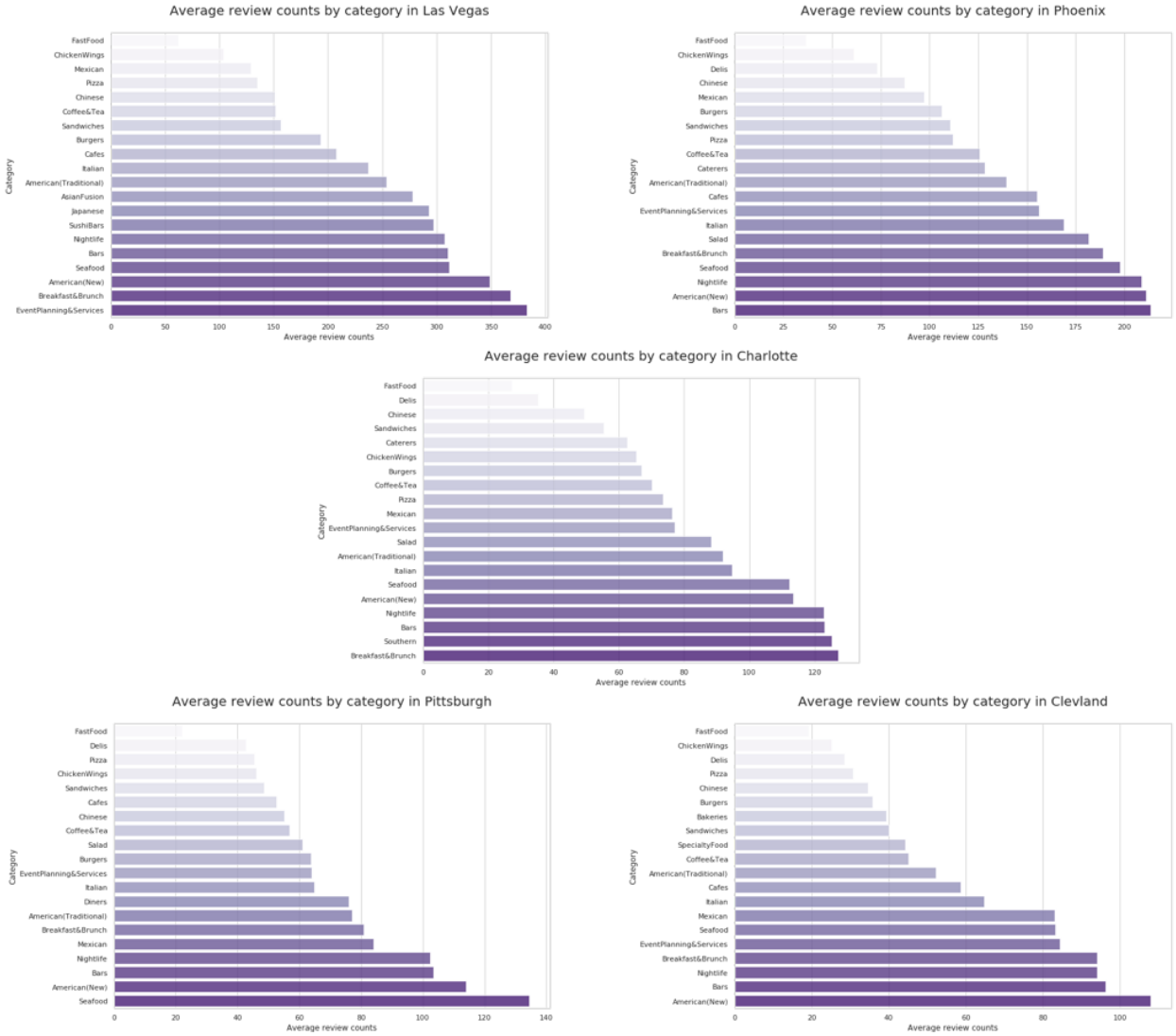

Fig. 19 Average review counts vs business categories at city level.

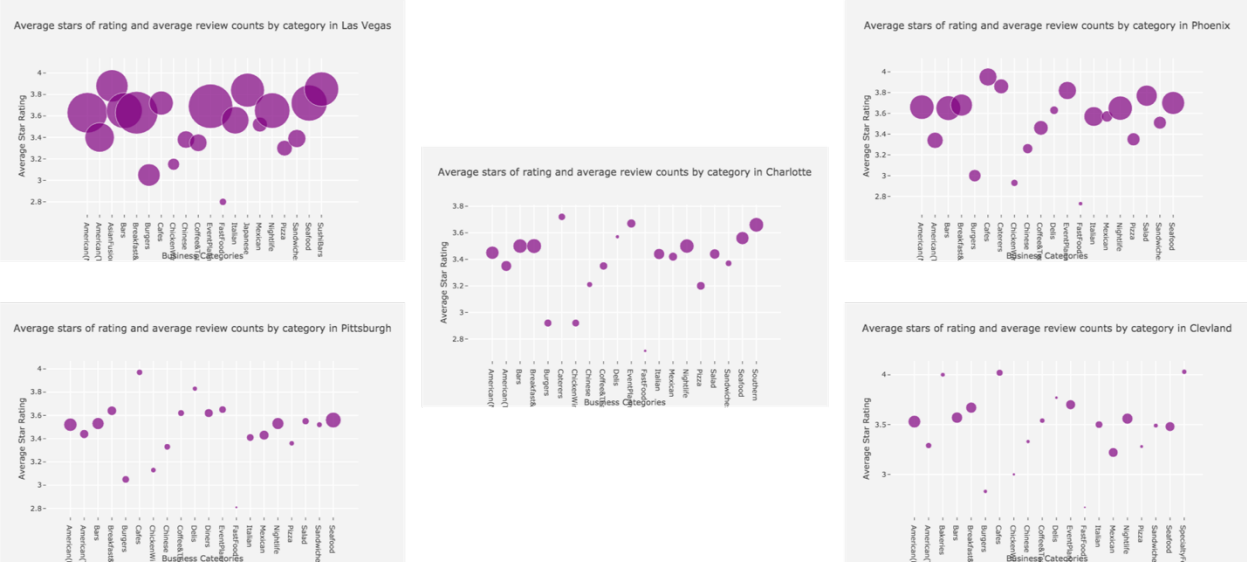

Fig.20 Average star rating vs businesses categories at city level with average counts of reviews represented by size of bubble.