IDS REPORT

ABOUT THE DATA SET

We had been given the zomato dataset(zomato.csv)

This dataset has information of various restaurants stored under Zomato, a restaurant aggregator and food delivery start up.

It contains 17 columns, with 51717 rows.

The columns described in brief:

1. Url - contains the url of the restaurant in the zomato website

2. Address - contains the address of the restaurant in Bengaluru

3. Name - contains the name of the restaurant

4. Online_order- whether online ordering is available in the restaurant or not

5. Book_table - table book option available or not

6. Rate - contains the overall rating of the restaurant out of 5

7. Votes - contains total number of rating for the restaurant as of the above mentioned date

8. Phone - contains the phone number of the restaurant

9. Location- contains the neighborhood in which the restaurant is located

10. Rest_type- restaurant type

11. Dish_liked - dishes people liked in the restaurant

12. Cuisines - food styles, separated by comma

13. Approx_cost - (for two people)contains the approximate cost for meal for two people

14. Reviews_list - list of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer.

15. Menu_item - contains list of menus available in the restaurant

16. listed_in(type) - type of meal

17. listed_in(city) - contains the neighborhood in which the restaurant is listed

ABSTRACT

Using the dataset, we need to find factors which may affect the rating of restaurants by plotting visuals and applying logic. We may pose several questions in an attempt to find insights. Some of them include finding out what restaurant type is most popular, what cuisine is more popular, which locations have the most number of restaurants(and if that correlates to ratings), how online order and book table options affect ratings, votes vs ratings, popular meal types etc.

EXPLORATORY ANALYSIS

Cleaning:

- Url and Phone columns dropped as one cannot do analysis with such information.
- Name-
  Upon inspection, we saw that a bunch of names were repeated. We had to make sure they weren't just names of different branches of the same restaurant, which is allowed. Hence, we compared names with addresses. If name and address pairs were not unique, that meant the same restaurant was repeated.

```
In [139]: len(df)

Out[139]: 51717
```

```
In [143]: df.duplicated(subset=["name","address"]).sum()

Out[143]: 39218
```

Upon running the above code, we found that 39218 restaurants were repeated.
To make sure only a single copy of each unique name, address pair was kept, we used a dictionary of the form
dict={name1:[address1,address2,---], name2:[address1,address2,---],-----}
If the name, address pair alreasy exists, drop that row. In this way, we reduced the population size to 12499 after getting rid of duplicates.

- Online_order and book_table were already clean
- Rate-
  Renamed to "ratings" for convenience.
  There were 2045 nan values.
  6 values with "-".
  772 with "NEW".

All 9316 others were of string type, of format(x/5). (converted to float after dropping denominator part)

Dropped “NEW” values as they’d clash with analysis(6.18%)
Dropped nan and “-“ values as it is a very high percentage(19.29%, nearly a fifth of the data) and if we replaced it with anything and strived for insights on ratings, outcome would be biased.

Even though we lost some data, we wanted insights to be as accurate as Possible.

- Votes-
  Type int.
  No nan values.

  No. of 0 values=6
  Replaced with median(53)

  Didn’t do check for outliers as even though there are extreme values, they are still very much plausible and getting rid of them would lead to inaccuracy.

- Location-
  Renamed to “city_specifics”

  No missing values.

  92 categories.

- Rest_type-
  37 missing values.

  Replaced with mode.

  87 categories.

- Cuisines-
  3 missing values, type float.
  All other values=str

  replaced missing values with mode of cuisines. (there are multiple cuisines in each row

as some restaurants have more than one cuisine.
Therefore, have to split each cuisine for each row at comma and append to cuisines list. We can then find mode)

mode=North Indian

- Approx._cost(for two people)-
  Renamed to “approx._cost”

  Again, extreme, but plausible values.

  32 missing values, type float(nan)
  others all str
  replaced with mean(536)

  4 digit values had “,”. Had to remove it before converting to int type.

- listed_in(type)- already cleaned

  renamed to 'meal_type'

  0 missing values

  mode=delivery

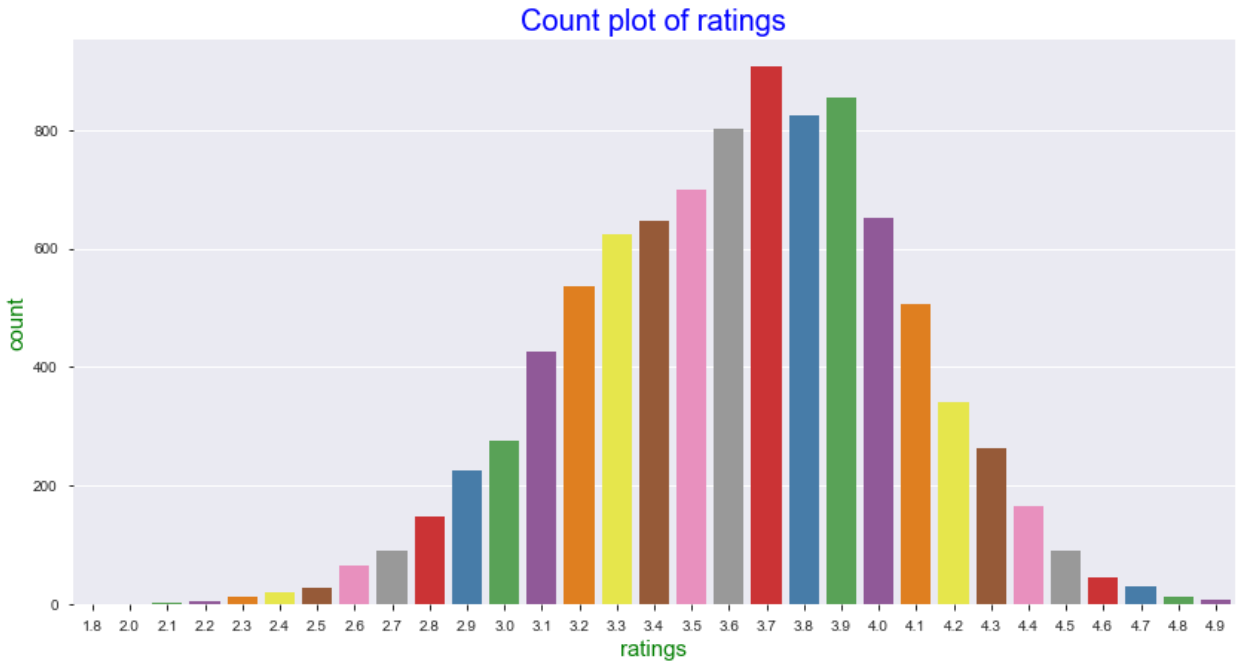
  7 categories

****

After cleaning, we tried to gather insights.

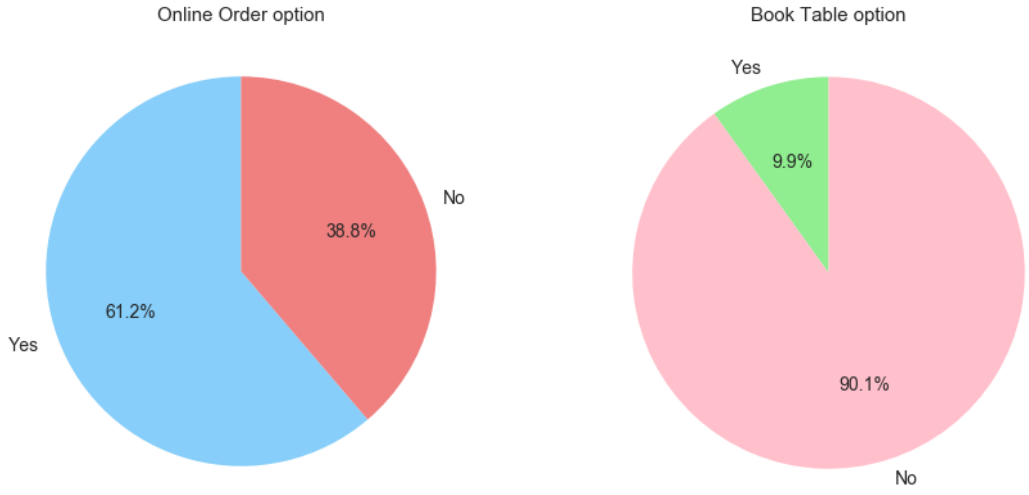
Plotting basic graphs, we see that

Count vs Ratings seems to show a near normal distribution with mean=3.7



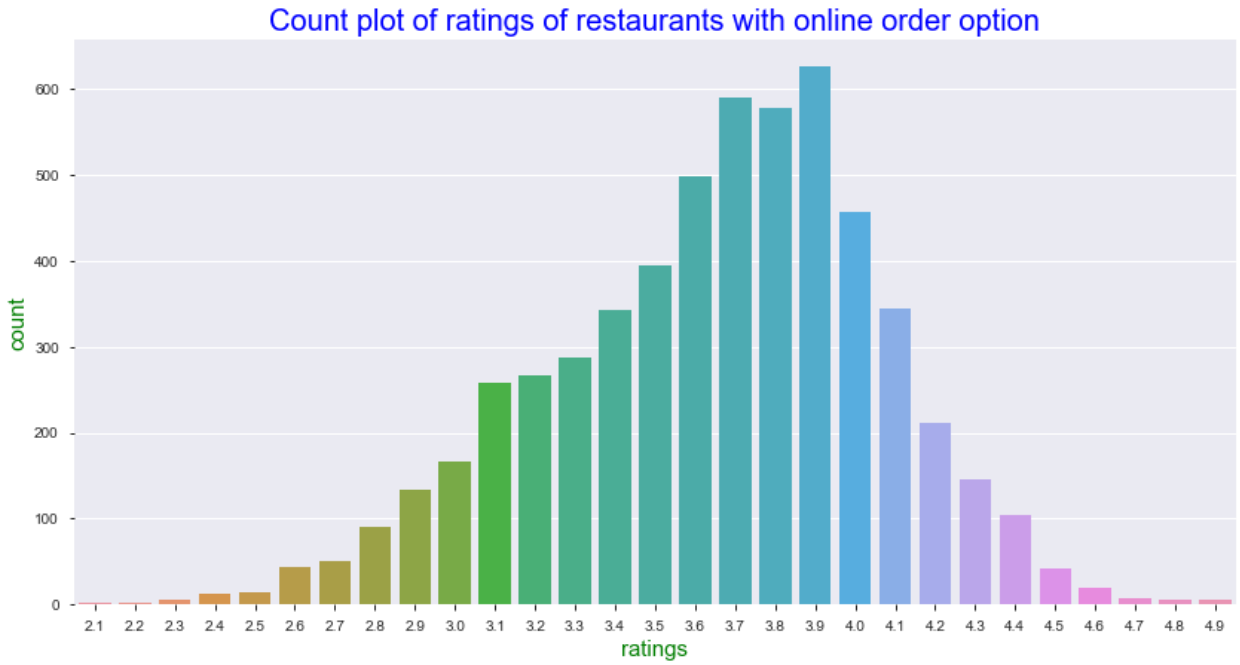
Percentage of restaurants with the online order option is more.

Percentage of restaurants with the book table option is less.



Diving in deeper, let's see if online order affects ratings.

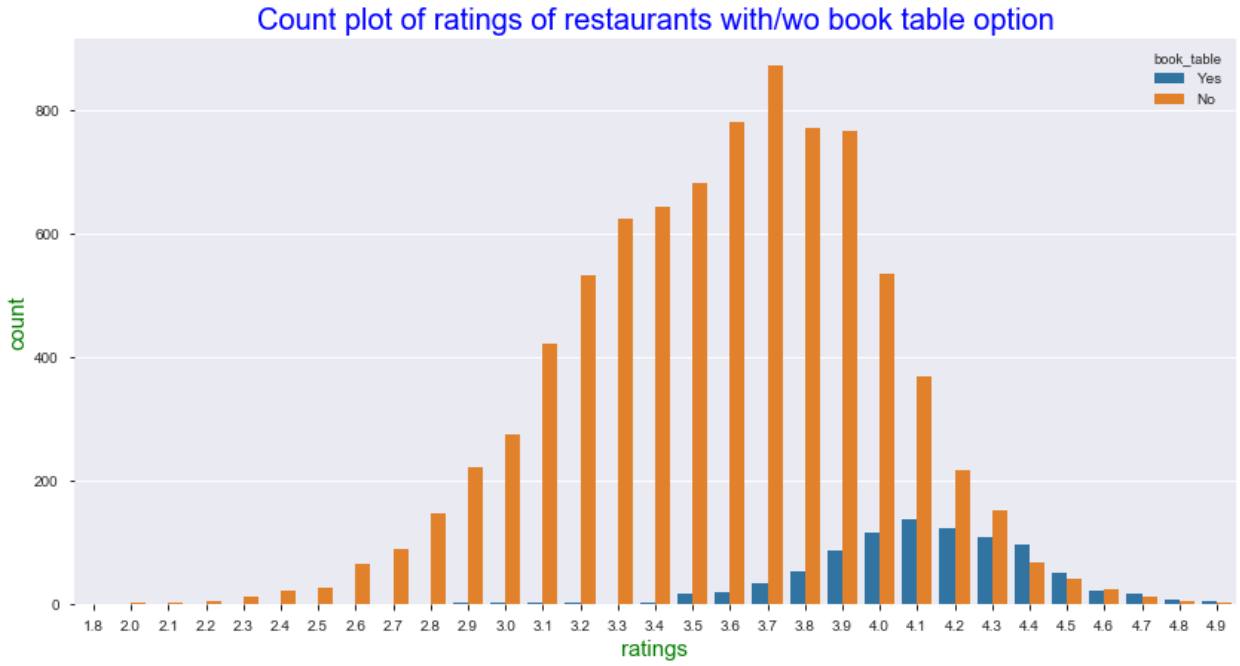
We can construct a count plot of restaurants with the online order option.



We can see that the mean doesn't shift too far away from what it was overall. Therefore, online order doesn't seem to have an effect on ratings.

What about book table?

Let us see the count plot for ratings restaurants with and without this feature.



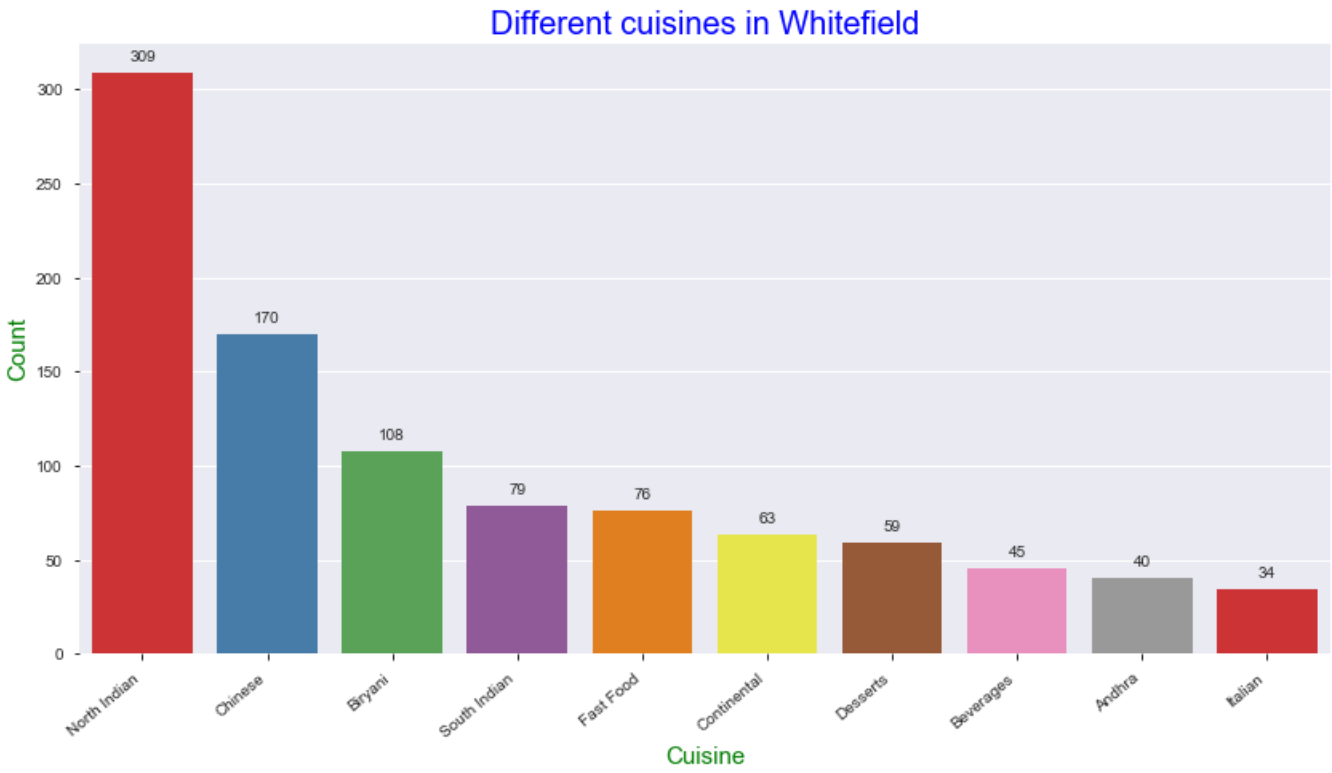
We can clearly see that there is a positive shift in the mean ratings when the book table option is there.

This could be due to the fact that in the rush of the 21st century, people do not have sufficient time to wait for tables, and the booking feature would be more convenient, hence the higher rating.

Now, which locations in the city have the most number of restaurants?

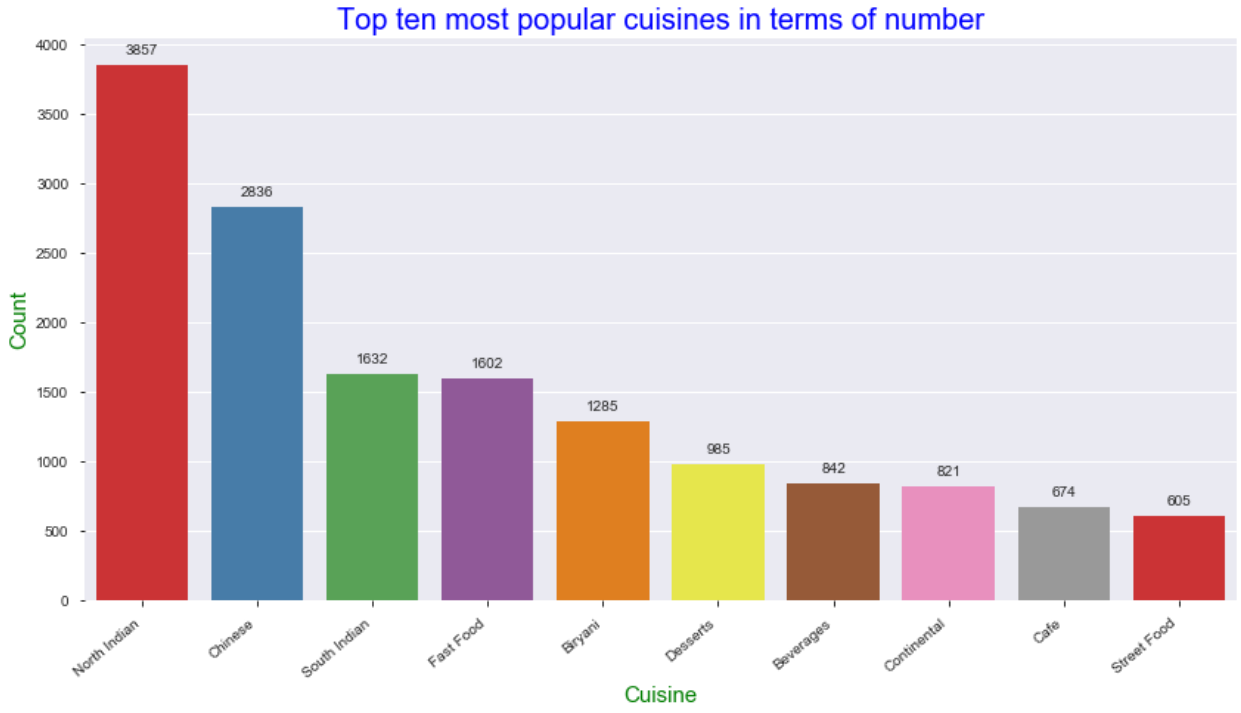
From this graph, we can see that Whitefield has the most number of restaurants. Makes sense, as it is a major IT hub, and hence, would be densely populated with people from all over.

Now, Whitefield has the most number of restaurants. What sort of cuisines do these restaurants have?



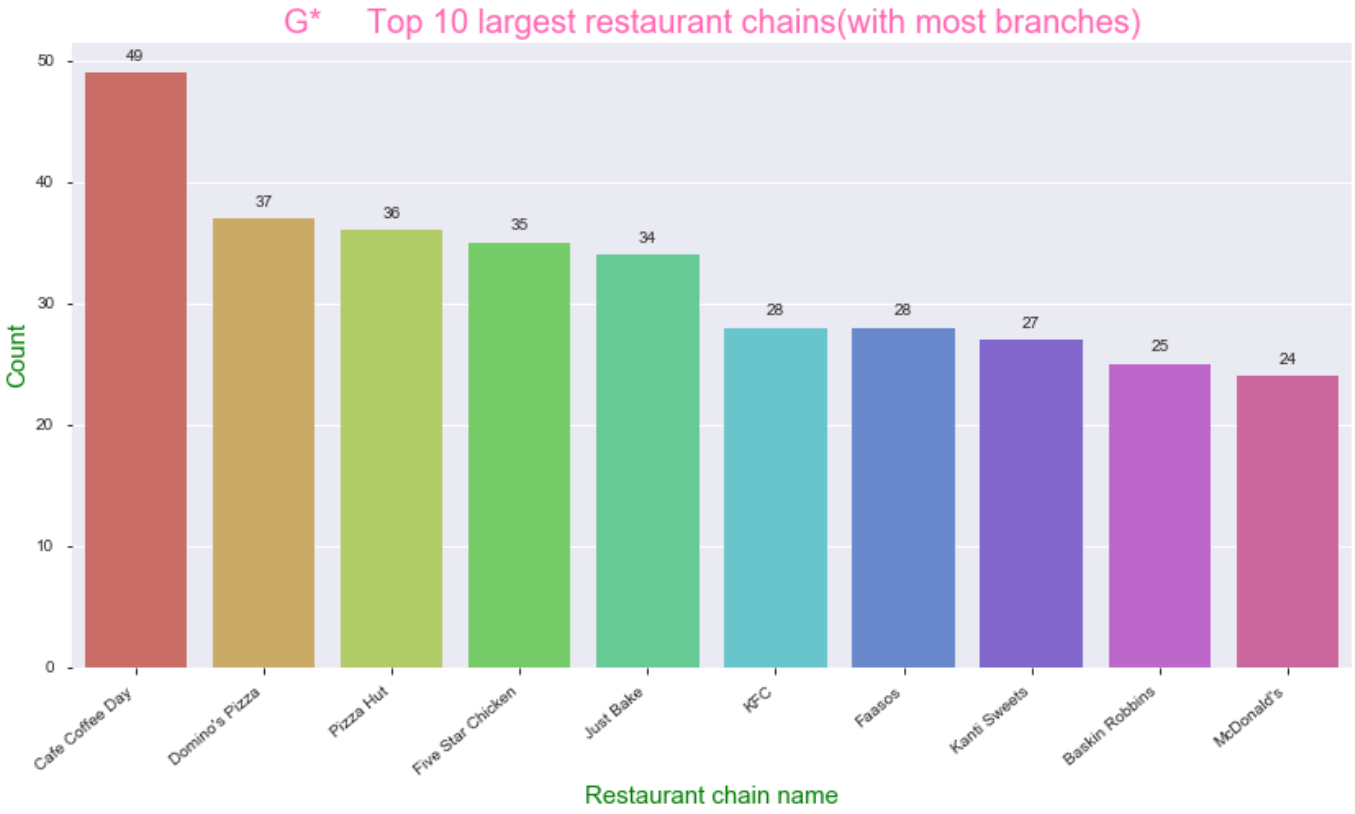
We can see that the most popular cuisine in Whitefield is North Indian. Again, makes sense as Bangalore is a cosmopolitan IT hub with many of our neighbours from the north living with us.

Moving on, which cuisines are popular overall?



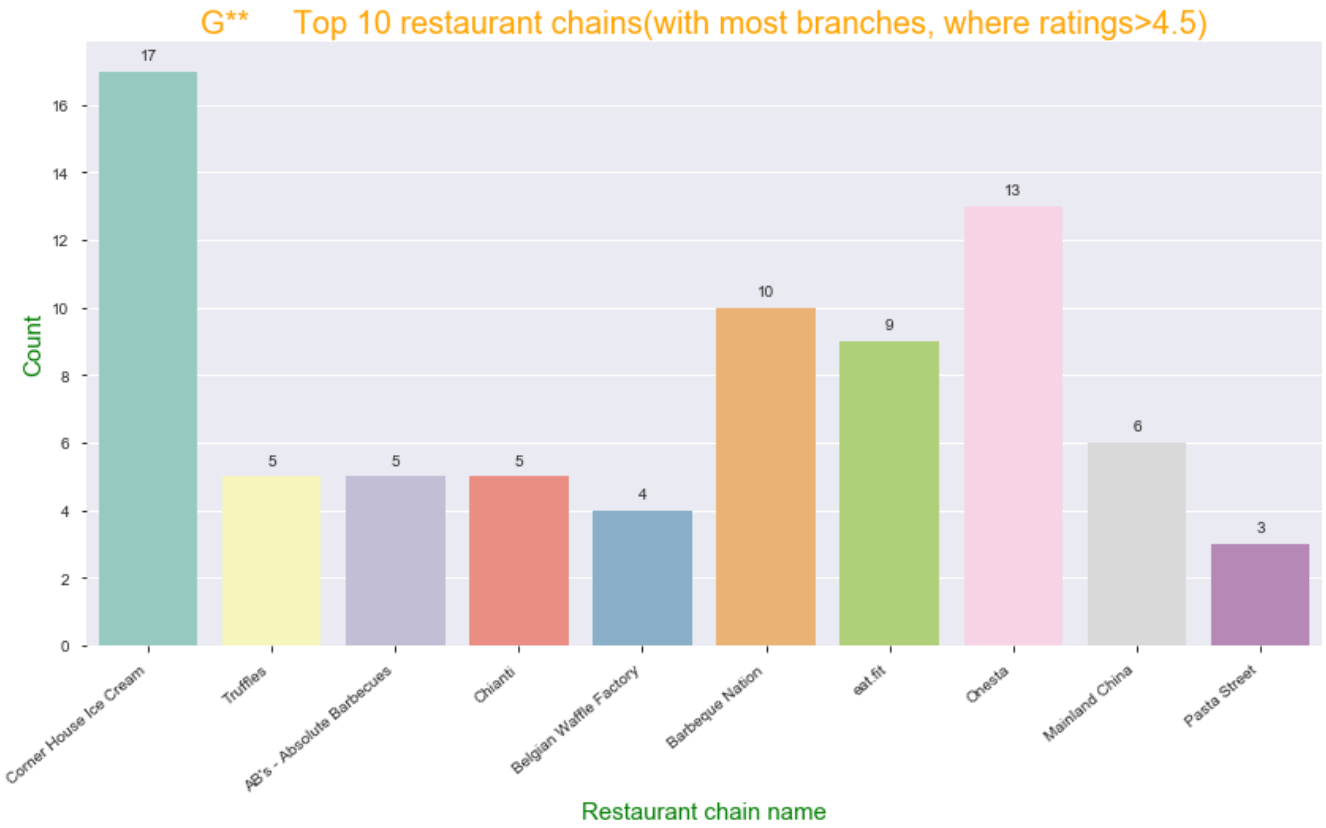
Once again, North Indian food is the winner. It's followed by Chinese and South Indian food.

Let's move on to another angle: Restaurant chains. Does number of branches of a restaurant affect rating? One would think so, as having more of a spread of a chain would make it more accessible to a larger demographic. That would make it convenient, and ratings should shoot up. Is it the case?



We can see that Café Coffee Day has the highest number of branches, namely 49. It's followed by Domino's Pizza and Pizza Hut.

Let's plot a graph which focusses on rating over number of restaurants.

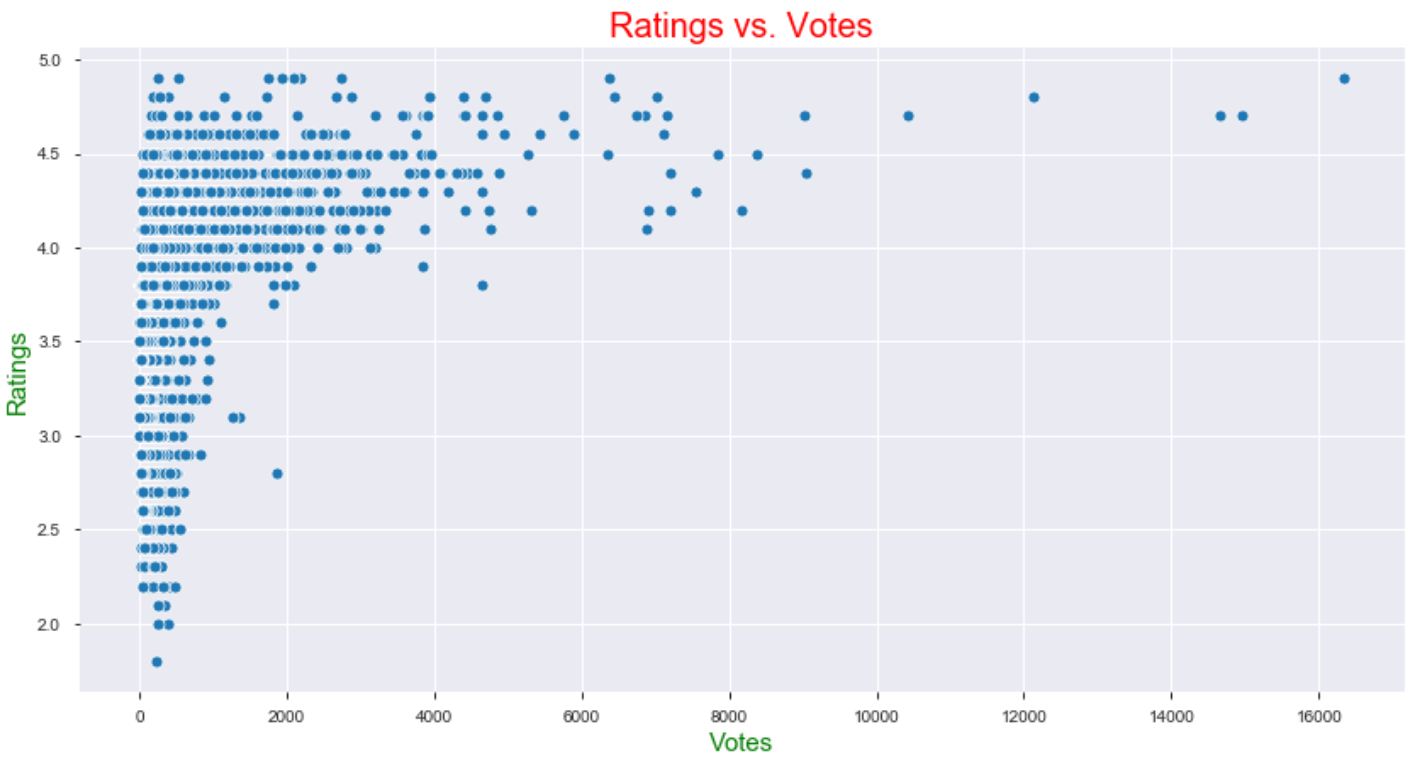


We can see that none of the places which were in G* showed up in G**. This means that none of the restaurants with a great number of branches have a high rating.

Pasta Street, which has 3 branches, out-rates CCD, which has 49.

Hence, our hypothesis is wrong, and just because a restaurant has more branches does not mean its ratings are better.

Now, is there any relationship between ratings and votes?



There’s no clear relationship, but we can see that

low rated restaurants(<3.5) do not have more than 1500 votes

most restaurants with higher ratings have votes lying bw 0-2500

Now, what is the one outlier in the top right corner?

It seems to have a high rating and a great number of votes. Let’s look into it.

```
In [180]: df.sort_values(by=['votes'],ascending=False)
```

Out[180]:

	Unnamed_0	name	address	online_order	book_table	ratings	votes	city_specifics	rest_type	dish_liked	cuisines	approx
1757	3921	Byg Brewski Brewing Company	Behind MK Retail, Sarjapur Road, Bangalore	Yes	Yes	4.9	16345	Sarjapur Road	Microbrewery	Cocktails, Dahi Kebab, Rajma Chawal, Butter Ch...	Continental, North Indian, Italian, South Indi...	
6430	18643	Toit	298, Namma Metro Pillar 62, 100 Feet Road, Ind...	No	No	4.7	14956	Indiranagar	Microbrewery	Beer, Pesto Pizza, Nachos, Cocktails, Beef Las...	Italian, American, Pizza	
3834	8330	Truffles	28, 4th 'B' Cross, Koramangala 5th Block, Bang...	No	No	4.7	14654	Koramangala 5th Block	Cafe, Casual Dining	Burgers, Pasta, Cocktails, American Cheese Bur...	Cafe, American, Burger, Steak	
8206	40506	AB's - Absolute Barbecues	90/4, 3rd Floor, Outer Ring Road, Munnekollaly...	No	Yes	4.8	12121	Marathahalli	Casual Dining	Raj Kachori, Paan Kulfi, Churros, Butter Chick...	European, Mediterranean, North Indian, BBQ	
3798	8268	The Black Pearl	105, 1st A Cross Road, Jyothi Nivas College Ro...	No	Yes	4.7	10413	Koramangala 5th Block	Casual Dining, Bar	Chocolate Lollipop, Chocolate Biscuit, Fire Sh...	North Indian, European, Mediterranean	

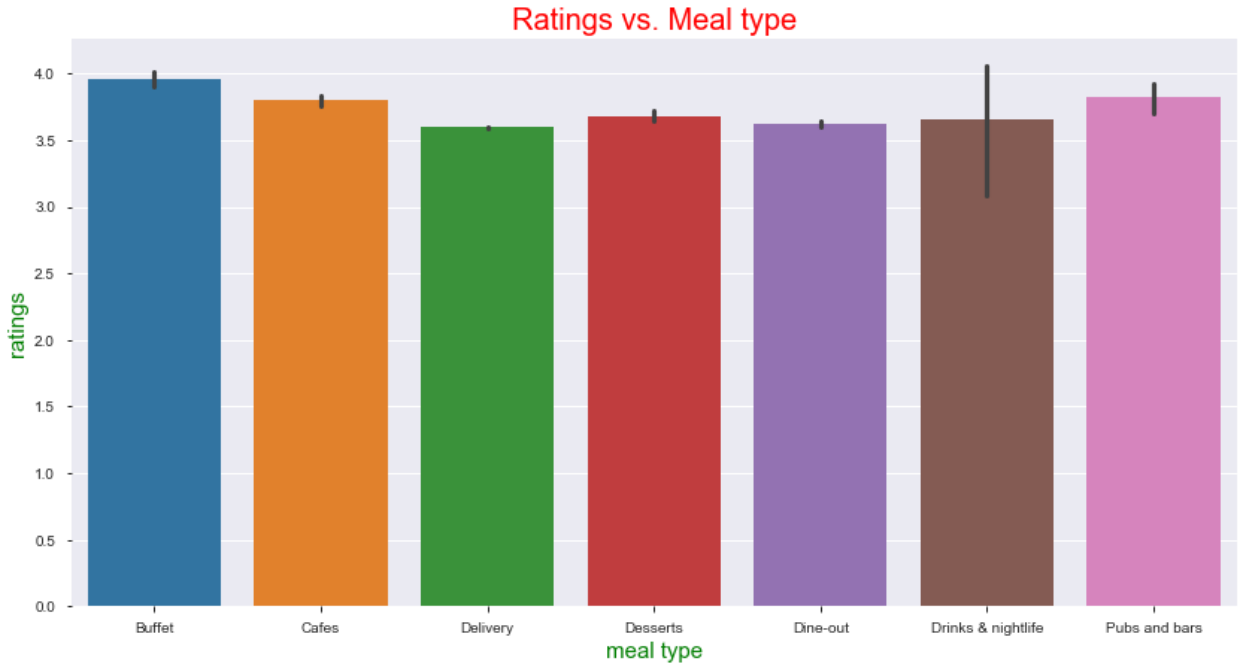
We can see that the outlier is 'Byg Brewski Brewing Company', followed by Toit and Truffles.

What makes these restaurants so popular?

If we look into the “city_specifics” column, we can see that really popular restaurants (votes>10000) all lie in IT/Business hubs (Sarjapur, Indiranagar, Marathahalli, Koramangala)

Furthermore, the rest_type of these are all perfect places to go for work lunches.

Next, let us plot rating vs meal type.



Hmm…let us analyse Drinks & nightlife. It seems to have the most deviation in ratings. Why is this so? The locations of places where meal type is Drinks & nightlife:

```
Shanti Nagar
MG Road
Old Airport Road
Vijay Nagar
Whitefield
```

One angle of looking at it is the despicable events of New Year’s 2017, which no doubt brought MG Road a bit of a bad name. Maybe that’s why there is such a wide range.

CONCLUSION

The takeaway from this exercise is that through EDA, one may figure out tiny details that may contradict one's assumptions. For example, chain restaurants with more branches need not do better than restaurants with less branches. Even though number of restaurants with book table feature is less, they tend to do better than restaurants with no book table option, etc.