Notebook

February 17, 2020

Use the `head` command on your four files again. This time, describe at least one potential problem with the data you see. Consider issues with missing values and bad data.

Please write your answer in the markdown cell below. You may create new cells below your answer to run code, but **please never add cells between a question cell and the answer cell below it.**

One potential issue I see with the data is that the 'phone_number' field for the business data frame are not cleaned, some values are missing or contain more numbers than others. Some other columns also contain missing values that leave the data set empty or null. The 'score' column in ins.csv is confusing. What does the -1 mean, why are there so many -1's? This is something that needs to be cleanes as well.

1 6: Compute Something Interesting

Play with the data and try to compute something interesting about the data. Please try to use at least one of groupby, pivot, or merge (or all of the above).

Please show your work in the cell below and describe in words what you found in the same cell. This question will be graded leniently but good solutions may be used to create future homework problems.

Please have both your code and your explanation in the same one cell below. Any work in any other cell will not be graded.

In [416]:
```
def diff_of_ratings(scores):
    if (len(scores) >= 2):
        return max(scores) - min(scores)
    return 0

max_diff_id = ins_named.groupby('bid')['score'].agg(diff_of_ratings).sort_values(ascending = I
max_diff = ins_named.loc[ins_named["bid"] == max_diff_id]['name'].iloc[0]
max_diff

# I wanted to see which restaurant improved the most based on its rating or score.
# I am only going to consider restaurants with at least 2 ratings.
# San Francisco Champagne Society had the most improvement in its rating.
```

Out[416]:
```
'San Francisco Champagne Society'
```