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.**

Some potential problems with the 'bus' data are as follows: 1. There are a lot of missing values for the latitude and longitude which is denoted by: '-9999.000000' 2. There are also many missing values for 'phone_number' column which is denoted by: '-9999' Some potential problems with the 'ins' data are as follows:
1. There are a lot of score missing values in the 'score' column which are denoted as '-1'. 2. In the 'date' column it does not seem like the timing is accurate but a fixed abitrary time seen as '12:00:00 AM' One potential problem with 'vio' data is as follows: 1. The 'description' column seems to be different for every room which could cause difficulties if one wants to group my the 'description problem'. One potential problem with 'ins2vio' data is as follows: 1. The 'iid' which represents the business ID provides extra numerical values that are not part of the business ID. In other words after the '_' those numbers do no represent the actual business ID

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 [385]: #YOUR CODE HERE
          busrisk = pd.merge(vio, ins2vio, how = 'left')
          ins_named = pd.merge(ins, bus, how = 'left')
          busrisk = pd.merge(busrisk, ins_named, how = 'left')
          busrisk = busrisk[['name', 'risk_category']]
          busrisk.groupby('name')['risk_category'].apply(list).to_frame()

          #YOUR EXPLANATION HERE (in a comment)
          #This is a dataframe consisting of two columns: 'name' which is a representation of the name
          #of the restaurant and 'risk_category' representing the various risk categories that particul
```

```
Out[385]:                                                        risk_category
          name
          111 Minna Gallery       [Low Risk, Low Risk, Moderate Risk, Low Risk, …
          … Omitting 8 lines …
          iNoodles                [Moderate Risk, Low Risk, Low Risk, Low Risk, …

          [5047 rows x 1 columns]
```