Notebook

July 2, 2019

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

- besiness_id is not index of data frame, so three table cannot be joined or merged.
- 'city' and 'state' in 'bus' are redundant because all the values are the same.
- there are NaNs or named missing values.
- 'date' in 'ins' and 'vio' is not in datetime format.
- 'description' in 'vio' is hard to generate information.

0.0.1 Question 2b

With this information, you can address the question of granularity. Answer the questions below.

1. What does each record represent (e.g., a business, a restaurant, a location, etc.)?

2. What is the primary key?
3. What would you find by grouping by the following columns: `business_id`, `name`, `address` each individually?

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.**

1. Each record represents a restaurant
2. The primary key is business_id
3. The number of records is not changed when grouping by 'business_id' because all the values of 'business_id' are unique and 'business_id' is primary key. As for grouping by 'name' and 'address', the number of records is reduced because some restaurants have same name and some are in the same location.

0.1 3: Zip Codes

Next, let's explore some of the variables in the business table. We begin by examining the postal code.

0.1.1 Question 3a

Answer the following questions about the `postal code` column in the `bus` data frame?
1. Are ZIP codes quantitative or qualitative? If qualitative, is it ordinal or nominal? 1. What data type is used to represent a ZIP code?
Note: ZIP codes and postal codes are the same thing.

1. ZIP codes are qualitative and nominal.
2. string

0.1.2 Question 3c : A Closer Look at Missing ZIP Codes

Let's look more closely at records with missing ZIP codes. Describe why some records have missing postal codes. Pay attention to their addresses. You will need to look at many entries, not just the first five.

Hint: The `isnull` method of a series returns a boolean series which is true only for entries in the original series that were missing.

In 'address', some are off the grid, some are private lovations, some are various locations, so they don't have postal codes. Maybe other's addresses are not clear so they also don't have postal codes.

If we were doing very serious data analysis, we might indivdually look up every one of these strange records. Let's focus on just two of them: ZIP codes 94545 and 94602. Use a search engine to identify what cities these ZIP codes appear in. Try to explain why you think these two ZIP codes appear in your dataframe. For the one with ZIP code 94602, try searching for the business name and locate its real address.

94545 is in Hayward, California, 94602 is in Oakland, California.

I think maybe those restaurants have moved to San Francisco but not update the zip codes. orbit room's current location is 1900 Market St, San Francisco, CA 94102

0.1.3 Question 4g

In the context of this question, what are the benefit(s) you can think of performing SRS over stratified sampling? what about stratified sampling over cluster sampling? Why would you consider performing one sampling method over another? Compare the strengths and weaknesses of these three sampling techniques.

- To do stratified sampling, people need to divide dataset to several subsets with no overlap and it takes time.It can only be used to some of studies that categories are known. While SRS is simple and easy to conduct.
- Stratified sampling reflects population which makes sure points from all subsets are involved. Since clusters chosen by cluster sampling are random, there may be bias in clusters.
- The reason why we need to consider performing one sampling method over another: one is to better understand the advantages and disadvantages of those three sampling method and know which kind of sampling method is better to be used on different datasets. Because one method may suitable for one dataset but not the other.
- **SRS**: strengths:it is easy to carry out. people select samples from population so it is generalized and unbiased; weaknesses: all population are needed. **Stratified sampling**: strengths: it reduces bias. weakness: it is not useful when the data cannot be divided into subsets with no overlap. **Cluster sampling**: strengths:variabiliy can be reduced. weakness: there may be bias.

0.1.4 Question 6b

Next, let us examine the Series in the `ins` dataframe called `type`. From examining the first few rows of `ins`, we see that `type` takes string value, one of which is `'routine'`, presumably for a routine inspection. What other values does the inspection `type` take? How many occurrences of each value is in `ins`? What can we tell about these values? Can we use them for further analysis? If so, how?

'type' only has two unique values, one is 'routine', the other is 'complaint'. There are 14221 occurrences of 'routine', and only one occurrence of 'complaint'. The occurrences of 'complaint' is just one, it is too little and almost has no influence in more than 10000 data, so that it is meaningless to use them for further analysis.

Now that we have this handy `year` column, we can try to understand our data better.

What range of years is covered in this data set? Are there roughly the same number of inspections each year? Provide your answer in text only in the markdown cell below. If you would like show your reasoning with codes, make sure you put your code cells **below** the markdown answer cell.

The year range is 3. There are 3305 inspections in 2015, 5443 inspections in 2016, 5166 inspections in 2017, 308 inspections in 2018. So that each year has different number of inspections.

0.1.5 Question 7a

Let's look at the distribution of inspection scores. As we saw before when we called `head` on this data frame, inspection scores appear to be integer values. The discreteness of this variable means that we can use a barplot to visualize the distribution of the inspection score. Make a bar plot of the counts of the number of inspections receiving each score.

It should look like the image below. It does not need to look exactly the same (e.g., no grid), but make sure that all labels and axes are correct.

Hint: Use `plt.bar()` for plotting. See PyPlot tutorial from Lab01 for other references, such as labeling.

Note: If you use seaborn `sns.countplot()`, you may need to manually set what to display on xticks.

```
In [131]: ins.head()
```

Out[131]:

	business_id	score	date	type	new_date	year
0	19	94	20160513	routine	2016-05-13	2016
1	19	94	20171211	routine	2017-12-11	2017
2	24	98	20171101	routine	2017-11-01	2017
3	24	98	20161005	routine	2016-10-05	2016
4	24	96	20160311	routine	2016-03-11	2016

0.1.6 Question 7b

Describe the qualities of the distribution of the inspections scores based on your bar plot. Consider the mode(s), symmetry, tails, gaps, and anamolous values. Are there any unusual features of this distribution? What do your observations imply about the scores?

As we can see in this bar plot, the mode of all scores is 100. It is asymmetry and has a tail towards left. Some scores are not observed between 90 and 100. Scores around 65 is anamolous. One unusual feature is that most inspection scores are between 90 and 100 while the scores are concentrated on around 5 scores. Unlikely, there are relatively less scores range from 75 to 90 while roughly all integer scores are given. This observation implies that most of restaurants are good and got high score, some restaurants didn't reach the standards due to several reasons. Different standards have different measurements, if a restaurant didn't reach a number of standards, the score is low and more likely to be different to each other. For the restaurants that hardly loss scores, they may be more likely to get similar scores with only one or two weakness.

Using this data frame, identify the restaurant with the lowest inspection scores ever. Head to yelp.com and look up the reviews page for this restaurant. Copy and paste anything interesting you want to share.

The restaurant with the lowest inspection score ever is 'DA CAFE'. 'DA CAFE' got 48 score in 2016 for 9 violations most of which are uncleanness and inadequate food safety(https://www.yelp.com/inspections/d-and-a-cafe-san-francisco-5).

Now, create your scatter plot in the cell below. It does not need to look exactly the same (e.g., no grid) as the above sample, but make sure that all labels, axes and data itself are correct.

Hint: Use `plt.plot()` for the reference line, if you are using matplotlib.

Hint: Use `facecolors='none'` to make circle markers.

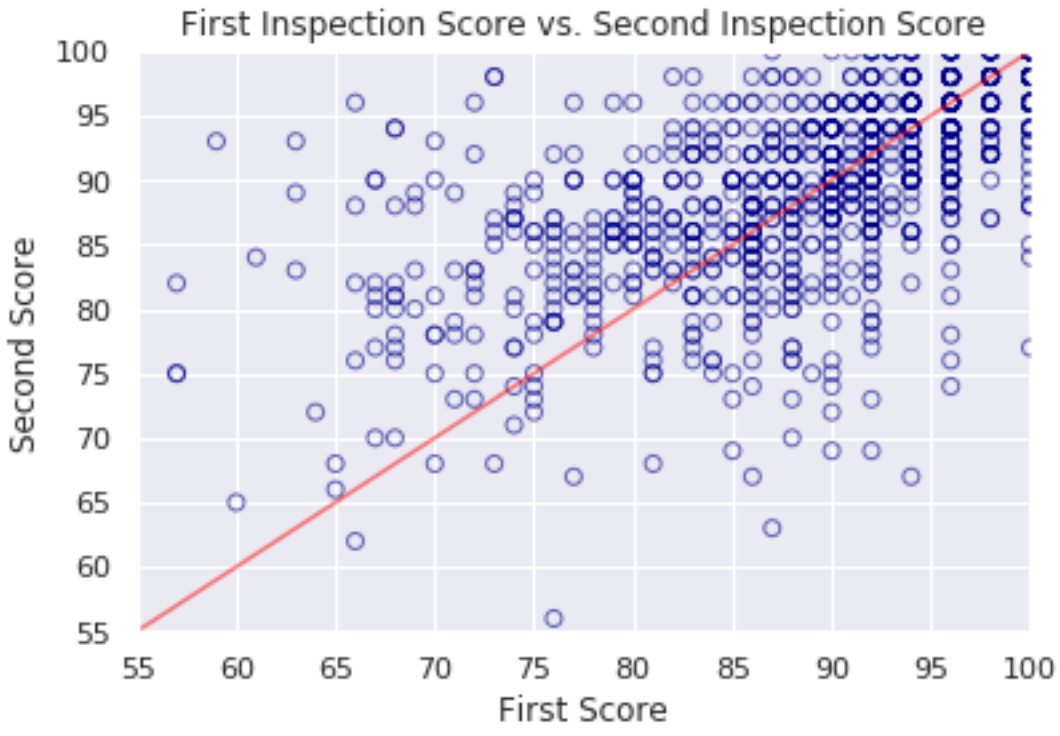
Hint: Use `zip()` function to unzip scores in the list.

```
In [144]: xs = np.arange(55, 101)
          ys = xs
          plt.plot(xs,ys,color = 'red',alpha = 0.5)

          plt.scatter(*zip(*scores_pairs_by_business['score_pair']),
                      edgecolor='darkblue',
                      facecolors='none',
                      alpha = 0.6)
          plt.xlabel('First Score')
          plt.ylabel('Second Score')
          plt.title('First Inspection Score vs. Second Inspection Score')

          plt.ylim((55,100))
          plt.xlim((55,100))

Out[144]: (55, 100)
```



0.1.7 Question 8d

Another way to compare the scores from the two inspections is to examine the difference in scores. Subtract the first score from the second in `scores_pairs_by_business`. Make a histogram of these differences in the scores. We might expect these differences to be positive, indicating an improvement from the first to the second inspection.

The histogram should look like this:

Hint: Use `second_score` and `first_score` created in the scatter plot code above.

Hint: Convert the scores into numpy arrays to make them easier to deal with.

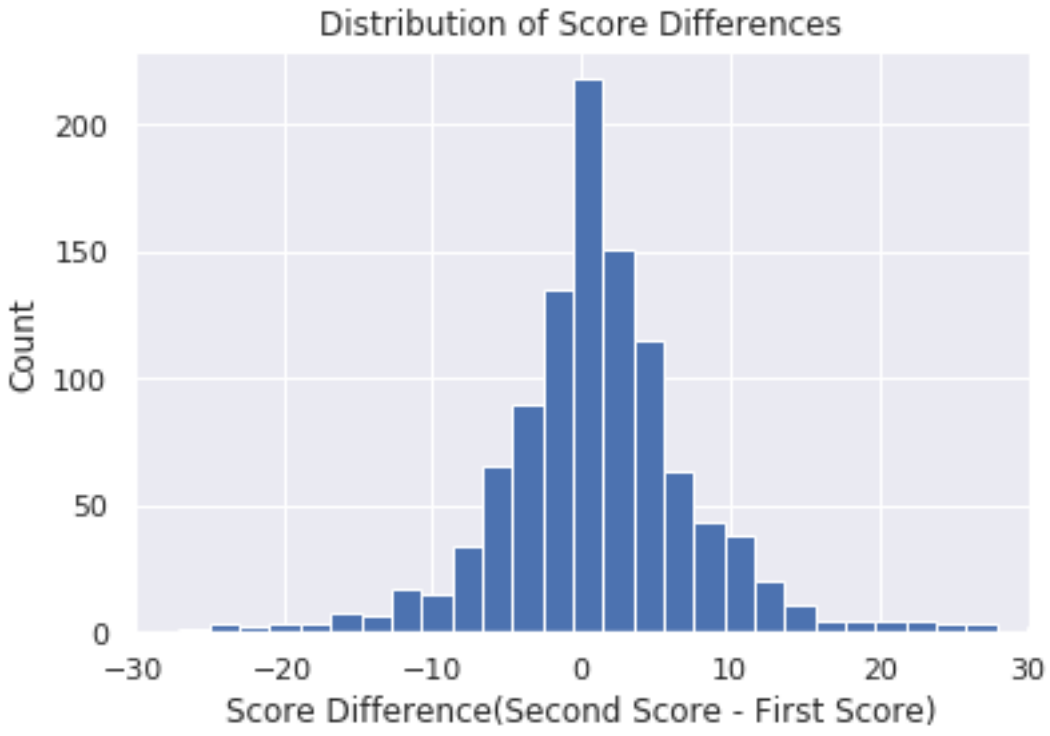
Hint: Use `plt.hist()` Try changing the number of bins when you call `plt.hist()`.

```
In [145]: def score1(x):
              return x.iloc[0]
          def score2(x):
              return x.iloc[1]

          newframe = myframe2['score'].groupby(myframe2['business_id']).agg([score1,score2])
          newframe1 = newframe.assign(diff = lambda x: x.score2 - x.score1).sort_values('diff')
          mydiff = newframe1['diff'].to_frame()

          mydiff.hist(bins = 30)
          plt.xlim((-30,30))
          plt.title('Distribution of Score Differences')
          plt.xlabel('Score Difference(Second Score - First Score)')
          plt.ylabel('Count')

Out[145]: Text(0, 0.5, 'Count')
```



0.1.8 Question 8e

If a restaurant's score improves from the first to the second inspection, what do you expect to see in the scatter plot that you made in question 8c? What do you see?

If a restaurant's score improves from the first to the second inspection, how would this be reflected in the histogram of the difference in the scores that you made in question 8d? What do you see?

If a restaurant's score improves from the first to the second inspection, the point would be in the upper triangle of the scatter plot. The score difference is positive so it would be in the right part of histogram.