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.

There are missing values in postal code, longitude, latitude, and phone numbers variables.

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 the "business id".
3. When you groupby "business_id" it replaces the default index, when you groupby "name" it replaces the default index, when you groupby "address" it replaces the default index.

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.

Question 1: ZIP Codes are qualitative and nominal. Question 2: A ZIP code is represented as strings.

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.

Some restaurants are off the grid according to their addresses. This means that they have no postal code to report.

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.

For ZIP code 94545 it appears in Hayward and Russell City. For ZIP code 94602 it appears in Oakland. It appears that the ZIP code is associated to a vending machine that appears in many locations and the main location is located somewhere in Hayward and Oakland

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.

The benefits of performing SRS over stratified sampling is that we have ZIP code data that is compere-hensive and we know it is mostly representative of the population of restaurants. In comparison, stratified sampling might be better than cluster sampling in this case because there are missing values in ZIP code data so since stratified sampling doesn't require some known information from prior data it could be better.

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?

The other values that the inspection type takes on are "complaint" and "routine" where "routine" has 14221 occurances and "complaint" only occurs once. These values tells us that there are almost no complaints, and we should find out why that is the case.

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.

This data set covers the years ranging from 2015 to 2018. In year 2015, there were 3305 inspections; however, during years 2016 and 2017, there were roughly the same amount of inspections (approximately 5000 each). Then in 2018, there was only 308 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 [305]: import seaborn as sns
          import matplotlib.pyplot as plt
          import numpy as np

          sns.set(style='darkgrid')
          ax=sns.countplot(x=ins['score'])
          ax.set_xticklabels(ax.get_xticklabels())
          ax.set_xlabel('Score')
          ax.set_ylabel('Count')
          ax.set_title('Distribution of Inspection Scores');
```

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?

We can see from the plot that the mode appears to the restaurants that received the highest rating. And most restaurants receive a high score for the inspections. There are an unsually high number of restaurants with high scores; however, that can be partially explained by the idea that good score implies safer restaurant. This implies that the scores are reflective of restaurant's quality of health.

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 inspect scores ever is DA CAFE. One customer said "Wipes counter. Wipes nose. Handles cash. Puts a straw in your drink. Not just one staff member but all 3 ladies at the counter did this. Not sure they could earn their 72 inspection score on a regular day."

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.

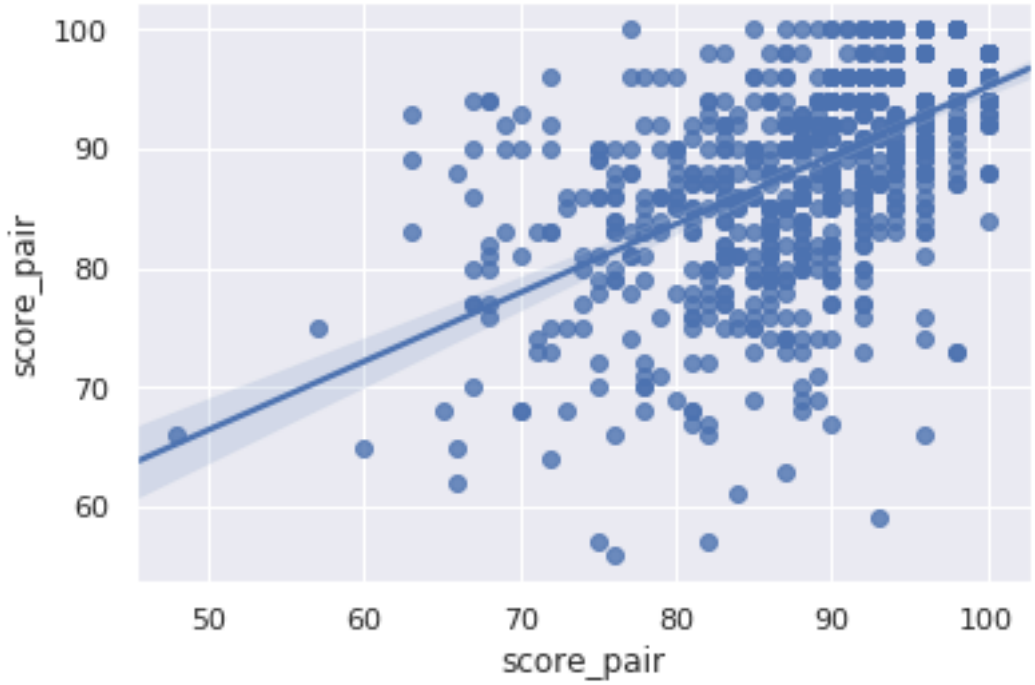
Hint: Use `plt.plot()` for the reference line, if you are using matplotlib.

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

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

```
In [315]: xscores = scores_pairs_by_business['score_pair'].str[0]
          yscores = scores_pairs_by_business['score_pair'].str[1]
          sns.regplot(xscores,yscores)
```

```
Out[315]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc58f9a16d8>
```



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.

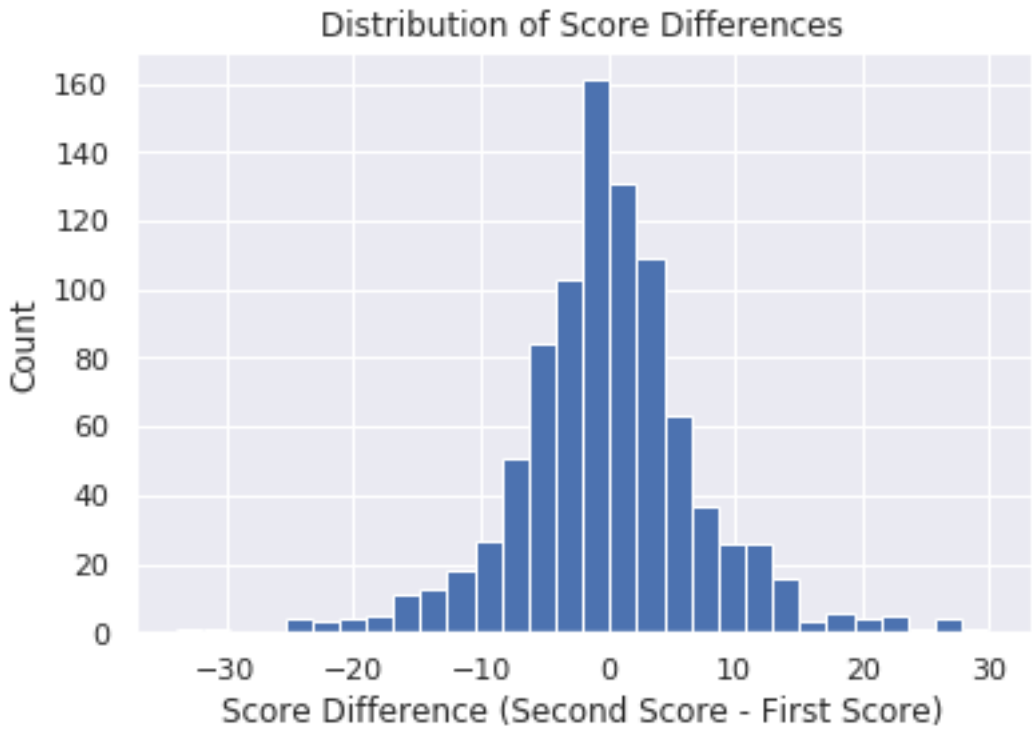
Hint: Use `plt.hist()` Try changing the number of bins when you call `plt.hist()`.

```
In [316]: xpair = [i[0] for i in scores_pairs_by_business['score_pair']]
          ypair = [i[1] for i in scores_pairs_by_business['score_pair']]

          score_difference = [ypair[i] - xpair[i] for i in range(len(xpair))]

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

Out[316]: Text(0.5, 1.0, 'Distribution of Score Differences')
```



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, we should expect to see most points clustered around the 90 ticks. And we see in the scatterplot that there are many points still laying around 70-80.

For the histogram, we should see distribution of scores be in the positive range from 10-30. However, we see there are mostly 400 restaurants that did not improve their scores.