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.

1.'bus'dataframe: 1)the formats of address are not consistent. Some mention which floor they locate in, while others don't mention, such as'1200 VAN NESS AVE, 3RD FLOOR' and '2801 LEAVENWORTH ST'.2)there are missing values in the column 'phone_number' and they are shown in 'NaN'.3)missing value in the column 'postal_code' and other abnormal formats of postcode, such as 'CA','Ca' and 9-digit postcode.

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 and its relative information, such as business_id, name, address, location (city, state, lattitude and longtitude), postcode,and phone number
2. the primary key is the 'business_id'
3. 1)when grouping by 'business_id', the size of the id is 6406, meaning there are 6406 restaurants in the dataframe. Also, all the id are unique. 2)when grouping by 'name', the size of the restaurant names is 5758, smaller than the size of the id, however, because there are no missing values in the 'name' column and some names have been used by more than 1 restaurants, therefore, the amount of groups is less than the amount of id. 3)when grouping by 'address', the size of the address column is 5626, smaller than the size of the id, that's because some restaurants have the same address;also,some address don't specify its spccific location but say "various 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.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.

1.Some restaurants have unclear address.1)'OFF THE GRID' for 'PASSION PIZZA', 'JACKRABBIT', 'SENOR SISIG (#6)', 'GAGA'S ROLLIN DINER', 'BAAGAN', 'COLETTA GELATO', 'MOONRAKER MOBILE (#2)', 'THE GAME DAY TRUCK', 'DA POKE MAN', 'THE WAFFLE ROOST', 'MOBI MUNCH, INC.' and 'BAHN MI ZON'. 'OFF THE GRID' has lots of branches with different zipcodes in San Francisco, therefore, it's unclear withou specifying which location it belongs to.2)'OTG' for 'KOME SUSHI BURRITO' and 'MAI THAI KITCHEN'. 'OTG' should not be the correct address of these two restaurants, therefore, their zipcodes are unknown.2.Some restaurants have various locations, such as 'CANTEEN VENDING COMPANY' and 'WEST COAST VENDING & FOOD SERVICE'.3.One restaurant has a 'PRIVATE LOCATIONS', the 'LAMAS PERUVIAN FOOD TRUCK'.

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-Alameda county (Hayward and Russell) 94602-Oakland These three cities are all very close to San Francisico, with distances from 12 miles to 30 miles. Therefore, it's easy to put these zipcodes in the category of San Francisco's zipcodes. Restaurant: ORBIT ROOM, Address: 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.

1.The benefit(s) performing SRS over stratified sampling:it's easy to use for extracting a research sample of business names from the whole; There's no need to divide the whole dataframe into groups and each element in the dataframe has an equal probability to be selected. 2.The benefit(s) performing stratified sampling over cluster sampling: it considers all the groups of business names grouped by postcodes, making it more precise. Since the homogeneity within each group is increased after stratification, the variation is reduced, and the sampling error of each group is reduced. Instead, the cluster sampling might not contain all groups from the whole; when the sample amount is constant, because the sample unit is not widely dispersed in the whole, the sampling error of the cluster sampling is generally larger than the SRS. 3.Depending on the purpose of my project, I will consider the method that fits it. SRS: 1>Strengths:ease of use, and the accuracy of representation 2>Weaknesses:When dealing with a large population of data, it's hard to number them; and it's difficult to implement in actual work; Stratified sampling: 1>Strenghs: 1)Since the homogeneity within the group is increased after stratification, the variation is reduced, and the sampling error of each group is reduced; 2)Different sampling methods can be applied to different groups based on the difference of each group;3)Independent analysis of each group become feasible. 2>Weaknesses:If the grouping features are not properly selected, the variation inside thhe group might become large, and the inter-group variation might become small. Then the sampling error will still be large. Cluster sampling: 1>Strenghs:it's easy to organize surveys; it's easy for quality control; It's less expensive in survey costs; 2>Weaknesses: when the sample amount is constant, because the sample unit is not widely dispersed in the whole, the sampling error of the cluster sampling is generally larger than the SRS.

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?

1. What other values does the inspection type take? 'complaint'.
2. How many occurrences of each value is in ins? 'complaint': 1 'routine': 14221
3. What can we tell about these values? The inspections on these restaurants are mostly routine inspections, while only one inspection was specially a complaint inspection, which means a complaint inspection is very rare.
4. Can we use them for further analysis? If so, how? Yes. We can combine the 'description' of the vio dataframe with the 'type' of the ins and the location information of the restaurants from the bus dataframe to analyze the inspection results based on different locations.

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.

1. What range of years is covered in this data set? From 2015 to 2018.
2. Are there roughly the same number of inspections each year? The number of inspections in 2016 and 2017 are rougnly the same, where both were around 5000 times. The number of inspections in 2015 was around 3000 times. However, the number of inspections in 2018 was only 308.

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.

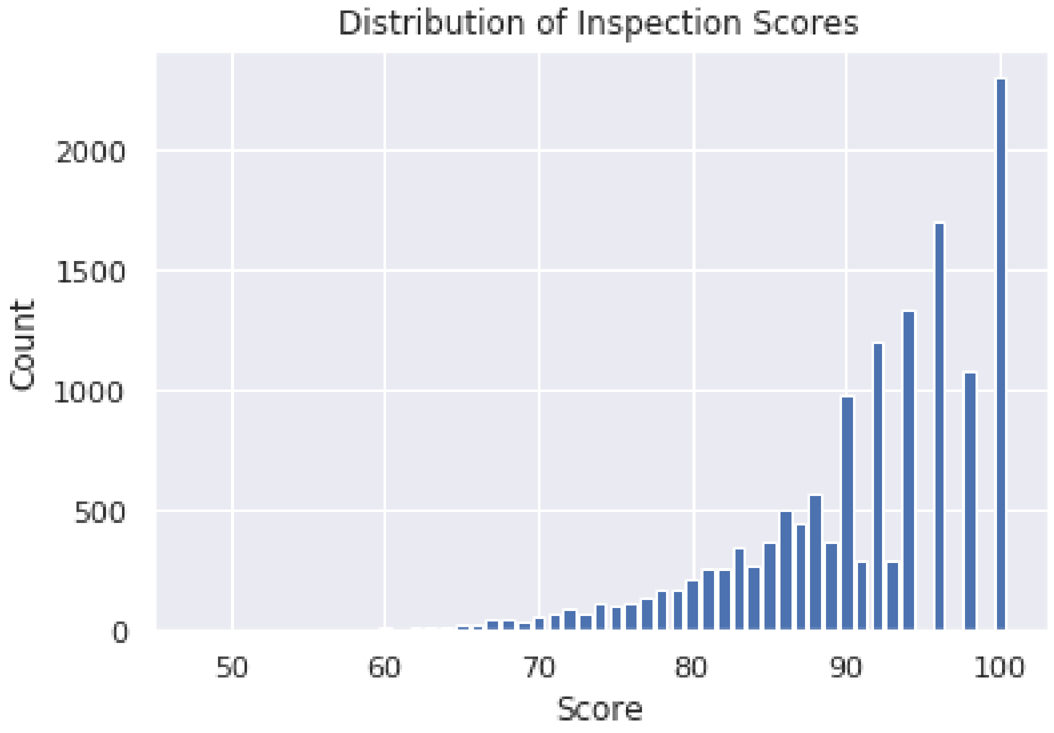
It should look like the image below. It does not need to look exactly the same (e.g., no grid), but make sure that all labels and axes are correct.

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 [637]: a = ins['type'].groupby(ins['score']).count()
          a.index
          plt.bar(a.index,a)
          plt.title('Distribution of Inspection Scores')
          plt.xlabel('Score')
          plt.ylabel('Count')

Out[637]: Text(0, 0.5, 'Count')
```



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?

1. The shape of the distribution-Skewed left It has distributions with fewer observations on the left. It mostly distributed from 80 to 95. From the shape of it, we can see that there are more restaurants with higher scores among the the ins dataframe
2. Gaps. The figure above has three small gaps between the score from 95 to 100.
3. Spread. The variability of the score is large because it covers a wide range.
4. Center. The center of a distribution is around the score of 89, where about half of the observations are on either side.
5. Mode. Shows a mode when score = 100.

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 scores 48:DA CAFE Good reviews: 1. 4.0 star rating 3/13/2019 I was very surprised with this Boston Market. I had not been to this location for more than 10 years when I used to live in SF. My buddy wanted to stop by here before going to Safeway; I was slightly hesitant, but figured...oh well.

We came here at approximately 12:00 pm. I figured it was going to be slammed, but to my surprise, it wasn't busy at all. I ordered the boneless chicken breast with stuffing and potatoes. The food and service was good. I was really shocked. I was even more shocked to learn they still had stuffing because last December, a Boston Market in Littleton, Colorado stopped serving stuffing and told me it was now seasonal. That was so upsetting.

The service at this location was fast and good. Definitely a change from what I recall 10 years ago. 2. 4.0 star rating 1/29/2019 Wow it's been 25yrs since I've been here. 1st of all customer service was great the restaurant was clean the food was hot and fresh and they have more hot food options better than they did 25 years ago. lol I even asked have a sample so they let me tried their Brussels sprouts what the hell had chili sauce did you put some on there? He replied there's no hot sauce Ummmmm dude yes it did was too spicy for me. Guess what if u Sign up with texting or email you can get a coupon I would've got one if the website was working today so I called the Manager she said just come in next Monday or Tuesday and I'll take care of you. (Thank you Mildred!)

Mixed reviews: 1. 3.0 star rating 6/8/2019 My boyfriend and I ordered Boston Market from DoorDash and we were pretty satisfied - I would order again (except not the spinach!). The portions were very generous for the price and the food was all pretty good. He got the half white meat chicken in garlic herb with steamed veggies and mashed potatoes and I got the turkey dinner with mashed potatoes and creamed spinach. The creamed spinach was HORRIBLE - completely inedible. It tasted like they mixed canned spinach with cream cheese and sour cream, it was super thick, gloppy and extremely sour. It was so sour that I wondered it it had gone bad. The steamed veggies were not very fresh, but they were edible. The meats were all good and the potatoes and gravy hit the spot. BM forgot to pack the side of sweet potatoes we ordered, so that was a bummer. Overall, good food for the price, just definitely skip the spinach! 2. 3.0 star rating 4/25/2019 The food was ok. Service was ok. Portions are good. Lemonade was great. Location is in a shady area of sf. But overall was good.

Bad reviews: 1. 2.0 star rating 1/25/2019 The chicken pot pies have no flavor at all. Just terrible. The soft drinks were good so was the chocolate cake. 2. 1.0 star rating 9/19/2018 I usually go to the Boston Market in Gellert, but since I was in the city I decided to give this location a try. I placed a pick up order and two of the items I ordered were missing -_- I called them and asked if I could get credited for the two items since I wasn't willing to drive back but I don't think they understood me and just kept telling me to come back so they can give me a refund. I know it's partially my fault for not checking if everything was there, but there were only two people behind me in line and it wasn't crazy busy.The food is good, but because of this experience, I won't be coming back for a while.

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.

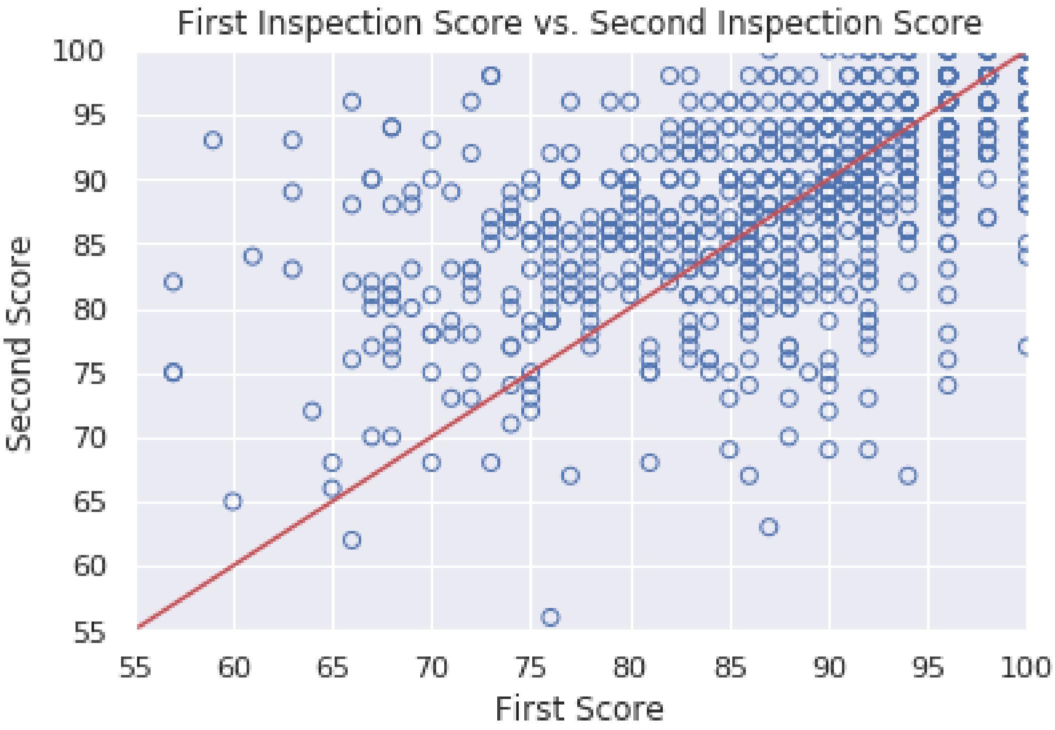
Hint: Use `plt.plot()` for the reference line, if you are using matplotlib.

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

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

```
In [649]: first = scores_pairs_by_business['score_pair'].agg(lambda x: x[0]).tolist()
          second = scores_pairs_by_business['score_pair'].agg(lambda x: x[1]).tolist()
          plt.scatter(first,second,facecolors='none',edgecolors='b')
          plt.axis([55,100,55,100])
          plt.title('First Inspection Score vs. Second Inspection Score')
          plt.plot([55,100],[55,100],c='r')
          plt.xlabel('First Score')
          plt.ylabel('Second Score')
```

```
Out[649]: Text(0, 0.5, 'Second Score')
```



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 [650]: newlist.reset_index().head()
```

```
Out[650]:   business_id score_pair  first  second  diff
          0          24   [96, 98]     96      98     2
          1          45   [78, 84]     78      84     6
          2          66  [98, 100]     98     100     2
          3          67   [87, 94]     87      94     7
          4          76  [100, 98]    100      98    -2
```

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?

1. 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? I expect to see that the datapoints should be mostly above the reference line. From the graph above, the amount of datapoints on the both side of the reference line is similar.
2. 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? The count of the positive score difference should increase and the height of that bin would be higher. From the graph, half of the restaurants had lower scores during their second inspection than their first inspection. The mode is at diff=0, meaning 200 restaurants didn't change their scores from the first to the second inspection.