Notebook

September 25, 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.

```
In [78]: display(ins.head())
         display(bus.head())
         display(vio.head())
```

	business_id	score	date	type
0	19	94	20160513	routine
1	19	94	20171211	routine
2	24	98	20171101	routine
3	24	98	20161005	routine
4	24	96	20160311	routine

	business_id	name
0	19	NRGIZE LIFESTYLE CAFE
1	24	OMNI S.F. HOTEL - 2ND FLOOR PANTRY
2	31	NORMAN'S ICE CREAM AND FREEZES
3	45	CHARLIE'S DELI CAFE
4	48	ART'S CAFE

	address	city	state	postal_code	latitude
0	1200 VAN NESS AVE, 3RD FLOOR	San Francisco	CA	94109	37.786848
1	500 CALIFORNIA ST, 2ND FLOOR	San Francisco	CA	94104	37.792888
2	2801 LEAVENWORTH ST	San Francisco	CA	94133	37.807155
3	3202 FOLSOM ST	San Francisco	CA	94110	37.747114
4	747 IRVING ST	San Francisco	CA	94122	37.764013

	longitude	phone_number
0	-122.421547	+14157763262
1	-122.403135	+14156779494
2	-122.419004	NaN
3	-122.413641	+14156415051
4	-122.465749	+14156657440

	business_id	date	description
0	19	20171211	Inadequate food safety knowledge or lack of ce...
1	19	20171211	Unapproved or unmaintained equipment or utensils
2	19	20160513	Unapproved or unmaintained equipment or utensi...
3	19	20160513	Unclean or degraded floors walls or ceilings ...
4	19	20160513	Food safety certificate or food handler card n...

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 an individual business.
2. A primary key is a column in a table that is used to uniquely define all records. The primary key is the business ID.
3. Grouping by the following columns: business_id, name, address each individually would not change the rows because the combinations of these these factors are distinct.

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, because they are used to describe location and function as names rather than ordinal figures.
2. ZIP codes are represented by 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.

Missing ZIP codes as a result of food trucks, shared spaces/food halls with multiple restaurant inhabitants in one area, or restaurants housed within different establishments (e.g. Bon Appetit @ Airbnb, Chipotle within AMC, restaurant in Conservatory of Flowers).

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 Alameda County 94602 is in Oakland. The restaurant is Orbit Room. The address on 1900 Market Street does not match the zipcode that is currently in the table. The zip code is one digit off - "94602" vs. "94102." This could be a simple documentation error made during the data input process.

0.1.3 Question 5b

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 inspection type takes on two string values, but all except for one record takes on the value "routine". There is only one occurrence of the second value. As a result, the "type" column is not necessarily useful 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 data set ranges from 2015 to 2018. There are 3305 inspections in 2015, higher volumes in 2016 and 2017, and only 308 in 2018. Essentially, the majority of inspections are in 2016 and 2017.

0.1.4 Question 6a

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.

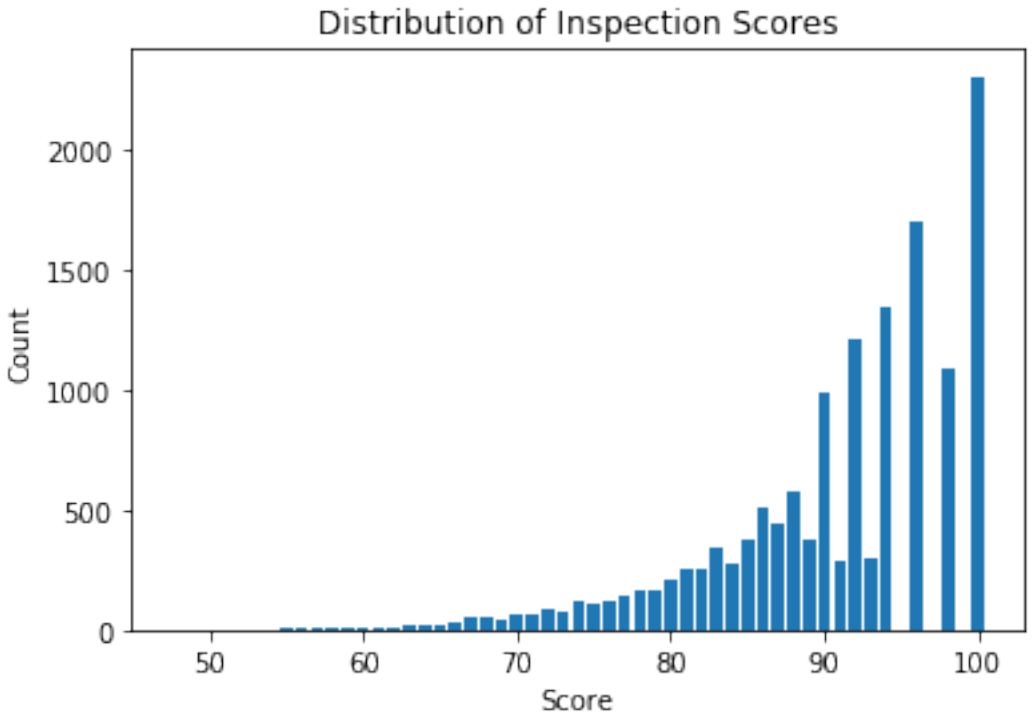
It should look like the image below. It does not need to look exactly the same (e.g., no grid), but make sure that all labels and axes are correct.

You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need: + `plt.bar` + `plt.xlabel` + `plt.ylabel` + `plt.title`

Note: If you want to use another plotting library for your plots (e.g. `plotly`, `sns`) you are welcome to use that library instead so long as it works on DataHub. If you use seaborn `sns.countplot()`, you may need to manually set what to display on xticks.

```
In [115]: scores = ins.groupby(['score']).size()
          scores.head()
          plt.bar(scores.index, scores)
          plt.title("Distribution of Inspection Scores")
          plt.xlabel("Score")
          plt.ylabel("Count")
```

```
Out[115]: Text(0, 0.5, 'Count')
```



0.1.5 Question 6b

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?

There are gaps and missing values. The distribution is unimodal and skewed much to the left, with a long tail to the left since some restaurants recieved very low scores. The gaps and bumps could be a product of the way that penalties are levied, such as in multiples of points, which result in certain deductions occurring.

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.

In my opinion, Yelp harms small businesses and is not necessarily a constructive tool for looking at restaurants. Although health inspections are valuable for public health reasons, Yelp reviews as a whole tend to highlight "Instagram-friendly" spots that actively contribute to gentrification, especially in a city like San Francisco (see the Mission district and its exodus of queer Latinx people).

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.

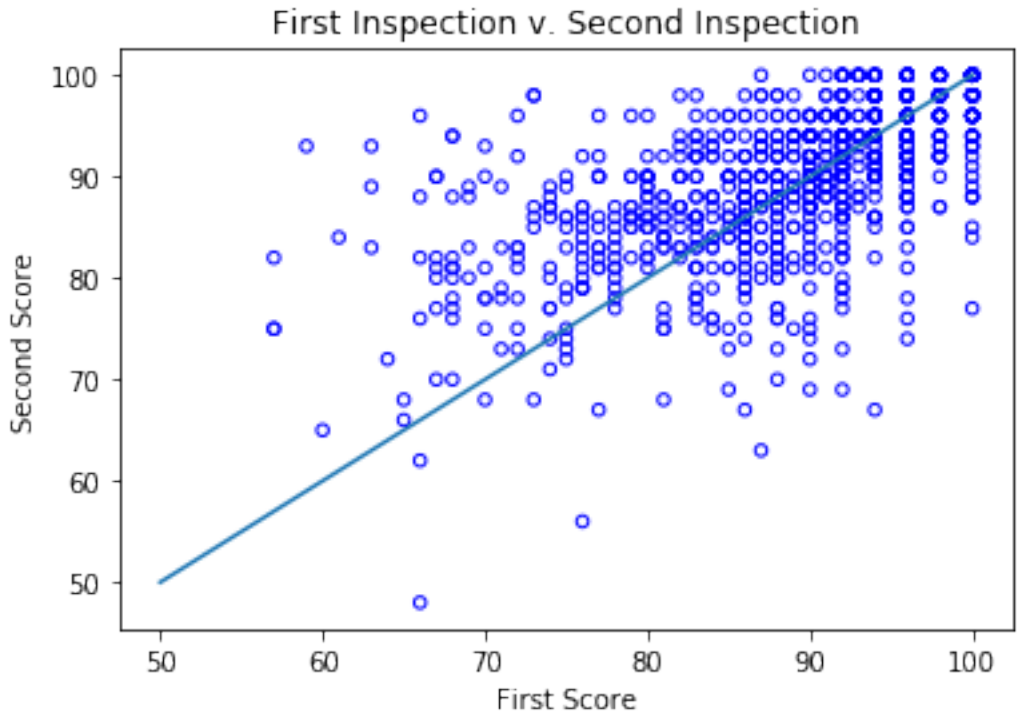
Key pieces of syntax you'll need: + `plt.scatter` plots a set of points. Use `facecolors='none'` to make circle markers. + `plt.plot` for the reference line. + `plt.xlabel`, `plt.ylabel`, `plt.axis`, and `plt.title`.

Note: If you want to use another plotting library for your plots (e.g. `plotly`, `sns`) you are welcome to use that library instead so long as it works on DataHub.

Hint: You may find it convenient to use the `zip()` function to unzip scores in the list.

```
In [152]: first_score, second_score = zip(*scores_pairs_by_business['score_pair'])
          plt.scatter(first_score,second_score,s=20,facecolors='none',edgecolors='b')
          plt.title("First Inspection v. Second Inspection")
          plt.plot([50,100],[50,100])
          plt.xlabel("First Score")
          plt.ylabel("Second Score")

Out[152]: Text(0, 0.5, 'Second Score')
```



0.1.6 Question 7d

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.

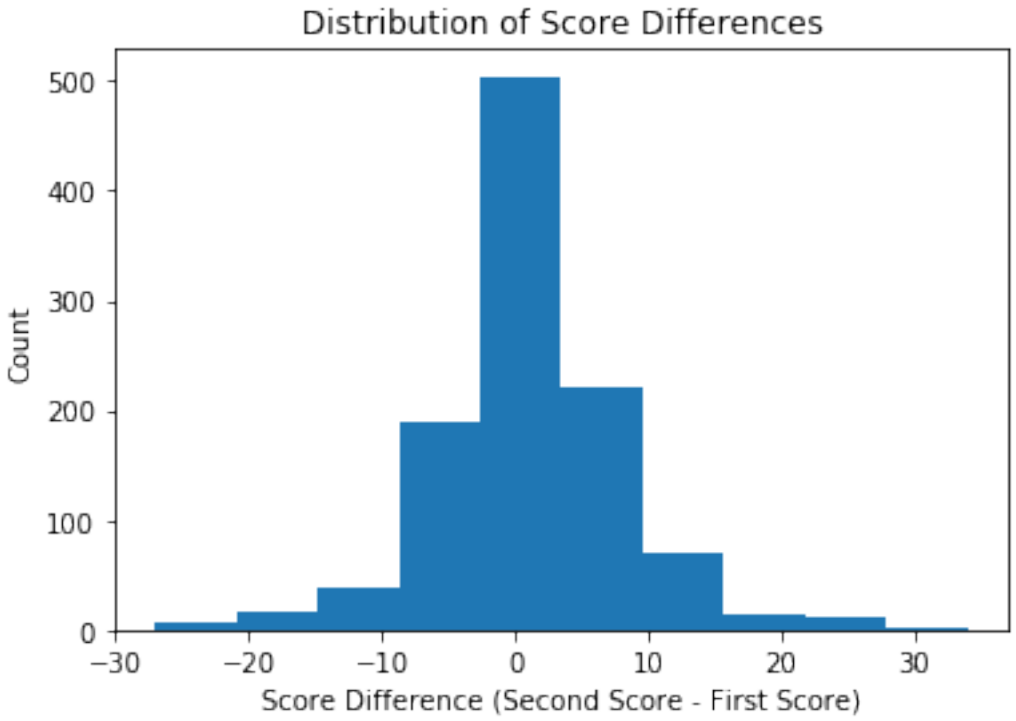
Hint: Use `plt.hist()` Try changing the number of bins when you call `plt.hist()`.

```
In [157]: biz = np.array(second_score) - np.array(first_score)

          plt.hist(biz)

          plt.xlabel("Score Difference (Second Score - First Score)")
          plt.ylabel("Count")
          plt.title("Distribution of Score Differences")
```

```
Out[157]: Text(0.5, 1.0, 'Distribution of Score Differences')
```



0.1.7 Question 7e

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 7c? 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 7d? What do you see?

If restaurant's score improves from the first to second inspection, the scatter plot points should fall above the line of y=mx+b, with m=1. The histogram of differences would also be shifted positively.

What we actually see in 7d is a unimodal distribution at 0 with long tails on either direction. This probably means that there usually isn't a big jump between first and second inspection scores.