Notebook

September 24, 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.

Date is improperly formatted and we could fall into trouble while sorting by date. Multiple inspections per company. Missing data in Business data frame.

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, with it's id, address, PO Box, Lat, Long, Phone Number and City
2. Primary Key should be business_id
3. When we individually groupby each and try to see their sizes, business ids are unique, names of businesses aren't and neither are addresses. Some addresses are labeled 'OFF THE GRID'.

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 Data, Nominal
2. They are 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 of the restaurants are off the grid and don't have latitudes, longitudes or addresses. Only 57 of these places have latitudes and longitudes listed while others haven't listed them.

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- Hayward City, Russell City, Alameda County, one of the locations is at 94545. 94602- Oakland, CA 94602 appears because of an error in typing their postal code- 94102

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?

Routine and Complaint are the values for inspection type. 1 of complaint and 14221 of routine. I think we shouldn't use this for further analysis as there isn't much diversity in the values. 1 complaint out of 14222 types of inspections doesn't make a huge difference.

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.

Range of years are 2015 to 2018. There were roughly the same amount of inspections in 2016 and 2017 but less inspections in 2015 and 2018.

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 [212]: a = [1,5,10,15,20,25,30,35,40]
```

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?

Even values are more likely than odd scores. There are more >90 rated restaurants. The lowest score seems to be in the high 50's. Seems to be exponentially tailed. The scores are tending to progress in an exponential manner with a lot of people getting in the 90's and above while fewer people get 80 or below.

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.

DA Cafe has the lowest inspection scores. 3 star reviewed on yelp. Health score was 72/100.

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 [590]: a = scores_pairs_by_business.loc[:].values
          b = list(zip(*a))[0]
          x = np.asarray([b[i][0] for i in range(len(b))])
          y = np.asarray([b[i][1] for i in range(len(b))])
```

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:

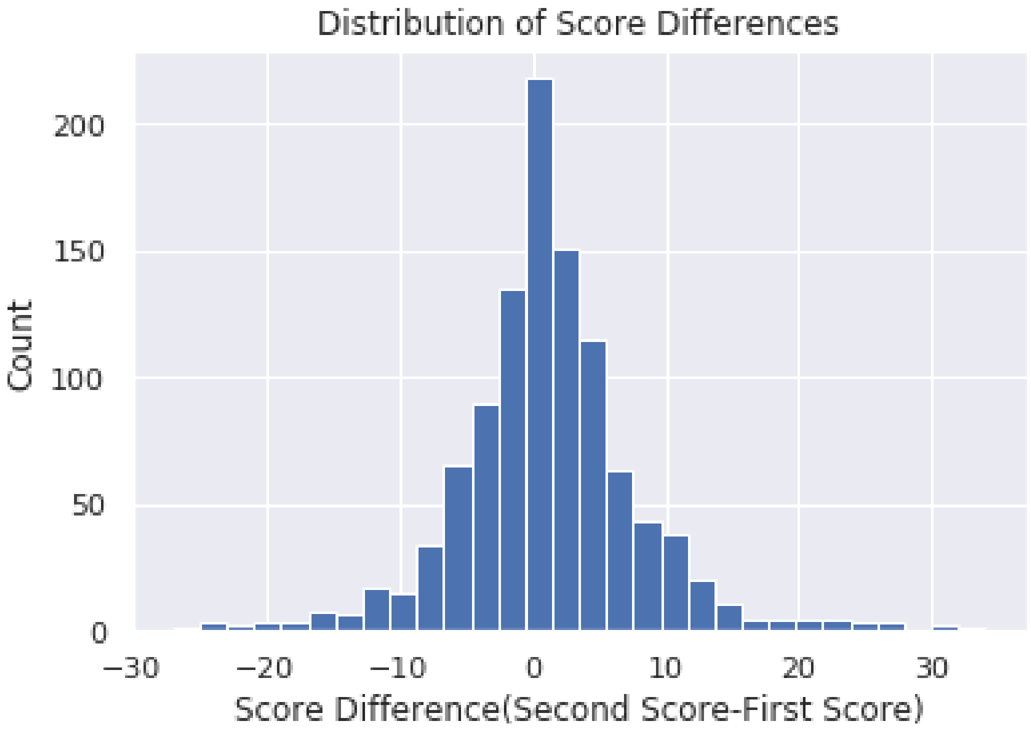
Hint: Use `second_score` and `first_score` created in the scatter plot code above.

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 [614]: plt.hist(y-x,30)
          plt.xlabel('Score Difference(Second Score-First Score)')
          plt.ylabel('Count')
          plt.title('Distribution of Score Differences')
```

```
Out[614]: 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 the score improves then the scatter plot point should lie above the y=x line as y, aka the second score is more than the first score, aka x. A lot of restaurants improved and we can see their points above the y=x line.

If the score improve, the histogram point will be in the bins above 0, as the difference in the scores is positive. A lot of restaurants improved by 1-10 points and we can see that in the histogram.