Notebook

February 7, 2021

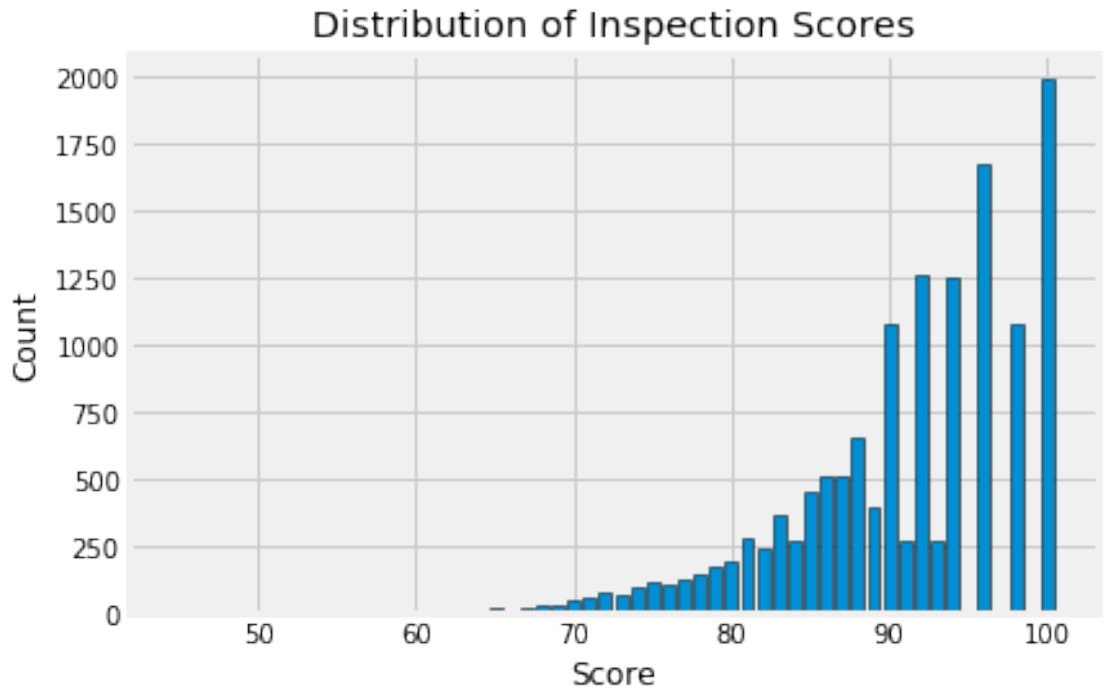
Use the `head` command on your three files again. This time, describe at least one potential problem with the data you see. Consider issues with missing values and bad data.

The values in the longtitude, latitude, postal_code and phone_number columns of the bus dataframe include nonsensical data. Namely, that latitudes should range from -90 to 90 and longtitudes should range from -180 to 180. However, some entires have -9999.000000 as their coordinates. Further, there are some entries of -9999 as the phone number and postal code as well. These bad data can cause bias and lead to inaccurate analysis.

0.1 Question 5a

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 bar plot 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.



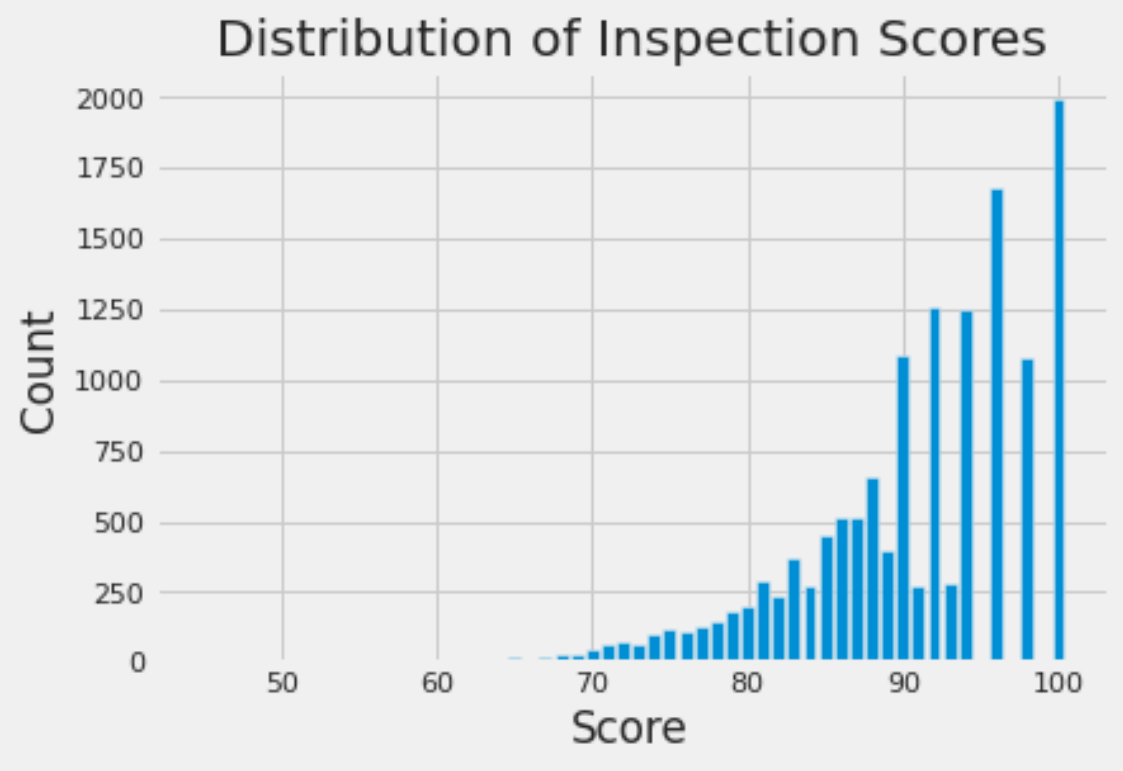
You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need:

```
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.

[71]:
```
ins_grouped = ins.groupby("score").size()

plt.xlabel("Score")
plt.ylabel("Count")
plt.title("Distribution of Inspection Scores")
plt.bar(ins_grouped.index, ins_grouped.values)
```

[71]:
```
<BarContainer object of 47 artists>
```

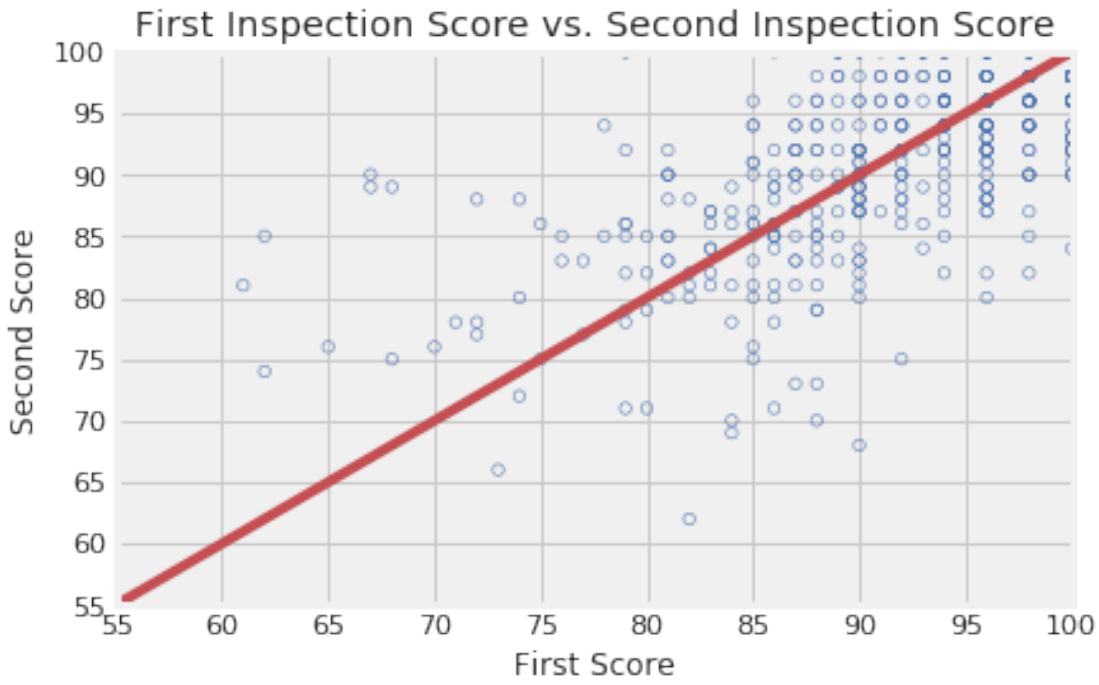


0.1.1 Question 5b

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

The distribution of the inspections scores do not have any symmetries, but it generally follows an exponential trend. Thus, the scores that have at least 1000 counts are all >90, and the scores that have less than 10 counts are all <62. The mode is 100, minimum score is 45, and tail is at the left end of the distribution at around score 72. Some gaps exist, and they fall in the 90-100 score range. There are also some anomalous values, which are the extremely low counts for scores around 91 and 94 as compared to sorrounding scores' counts. An unusual feature is that these two unusually low frequency bins share around the same counts. My observations imply that most places would have a "passing" inspection score unless they have extreme inadequate sanitation.

Now, create your scatter plot in the cell below. It does not need to look exactly the same (e.g., no grid) as the sample below, 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'` and `edgecolors='b'` to make circle markers with blue borders.

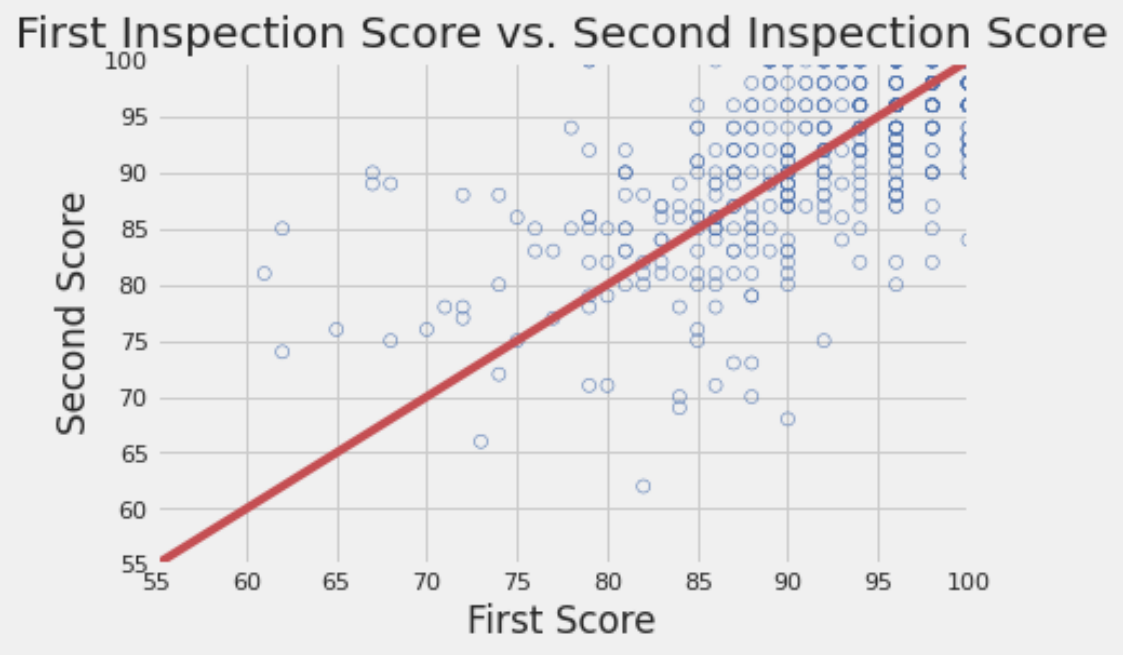
`plt.plot` for the reference line.

`plt.xlabel`, `plt.ylabel`, `plt.axis`, and `plt.title`.

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

[76]:
```
score1 = [scores_pairs_by_business["score_pair"].values[i][0] for i in␣
 ↪range(len(scores_pairs_by_business))]
score2 = [scores_pairs_by_business["score_pair"].values[i][1] for i in␣
 ↪range(len(scores_pairs_by_business))]

plt.scatter(score1, score2, facecolors='none', edgecolors='b')
plt.xlabel("First Score")
plt.ylabel("Second Score")
plt.axis([55, 100, 55, 100])
plt.title("First Inspection Score vs. Second Inspection Score")
plt.plot(plt.xlim(), plt.ylim(), color = 'r')
```

[76]:
```
[<matplotlib.lines.Line2D at 0x7f2c0f7d1430>]
```



0.1.2 Question 6c

If restaurants' scores tend to improve 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 oberve from the plot? Are your observations consistent with your expectations?

Hint: What does the slope represent?

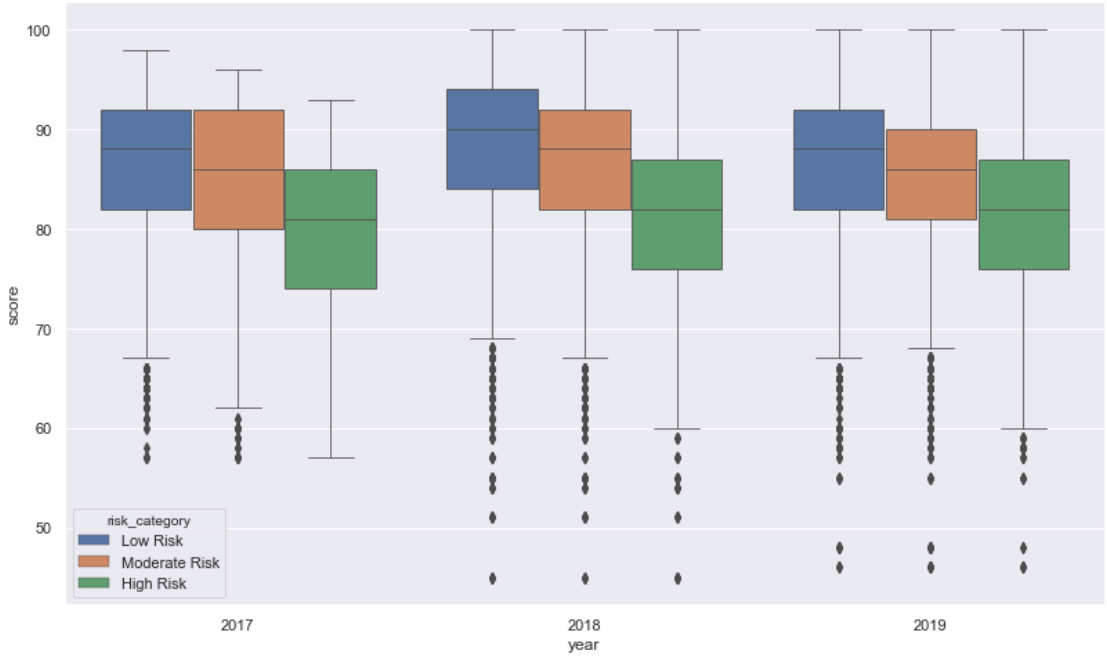
The reference line of slope 1 represents the "border" where the first score is the same as the second score: if the data is below the reference line, then it shows that the restaurant score did not improve because first score > second score; likewise, if the data is above the reference line, then it shows that the restaurant score improved because first core < second score. Thus, if restaurants' scores tend to improve from the first to the second inspection, then I would expect more data to be above the reference line. From the plot, there seems to be about the same distribution above and below the reference line, with slightly more data above the line. This indicates that only a few more restaurants tend to improve than those that did not improve from the first to the second inspection. My observation is not consistent with my expectation since I would expect most restaurants to learn from their first inspection and try to improve on their second expectation.

0.1.3 Question 6d

To wrap up our analysis of the restaurant ratings over time, one final metric we will be looking at is the distribution of restaurant scores over time. Create a side-by-side boxplot that shows the

distribution of these scores for each different risk category from 2017 to 2019. Use a figure size of at least 12 by 8.

The boxplot should look similar to the sample below. Make sure the boxes are in the correct order!

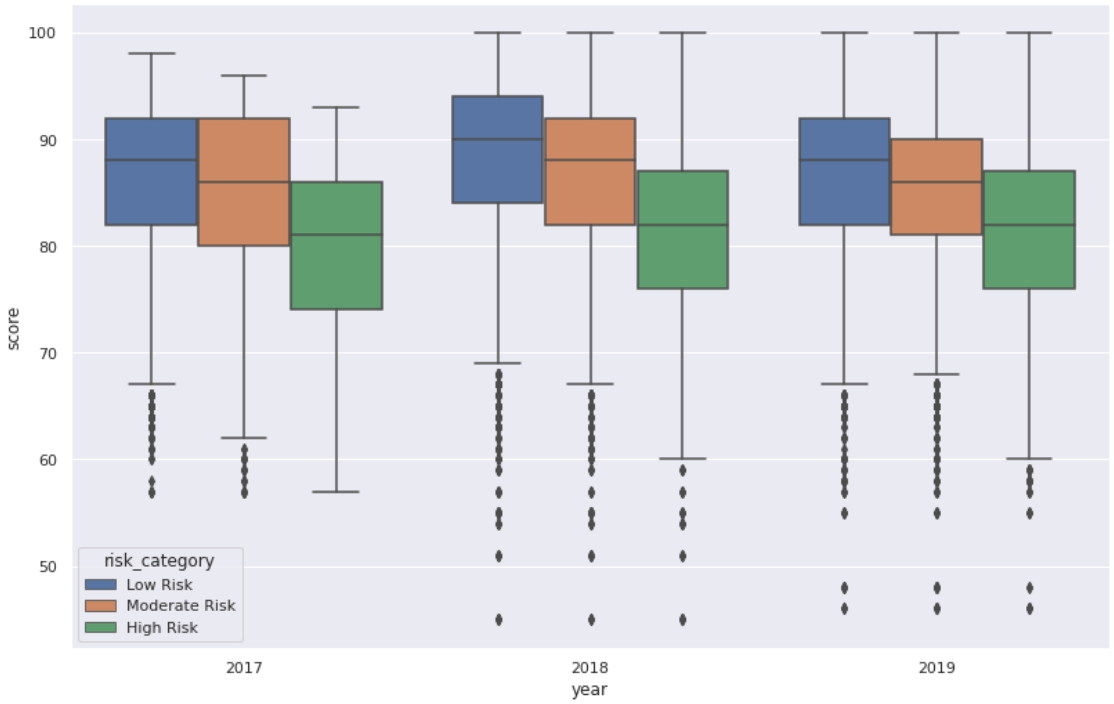


Hint: Use `sns.boxplot()`. Try taking a look at the first several parameters. The documentation is linked here!

Hint: Use `plt.figure()` to adjust the figure size of your plot.

[77]:
```
# Do not modify this line
sns.set()

ins_merge = ins.merge(ins2vio, how = "inner", left_on = "iid", right_on = "iid")
ins_merge_vio = ins_merge.merge(vio, how = "inner", left_on = "vid", right_on =␣
↪"vid")
ins_merge_vio_year = ins_merge_vio[ins_merge_vio["year"] >= 2017]
plt.figure(figsize = (12, 8))
sns.boxplot(x = "year", y = "score", hue = "risk_category", data =␣
↪ins_merge_vio_year, hue_order = ["Low Risk","Moderate Risk", "High Risk"])
```

[77]:
```
<AxesSubplot:xlabel='year', ylabel='score'>
```



0.1.4 Grading

Since the question is more open ended, we will have a more relaxed rubric, classifying your answers into the following three categories:

- **Great** (4 points):
  - For a dataframe, a combination of pandas operations (such as groupby, pivot, merge) is used to answer a relevant question about the data. The text description provides a reasonable interpretation of the result.
  - For a visualization, the chart is well designed and the data computation is correct. The conclusion based on the visualization articulates a reasonable metric and correctly describes the relevant insight and answer to the question you are interested in.
- **Passing** (1-3 points):
  - For a dataframe, computation is flawed or very simple. The conclusion doesn't fully address the question, but reasonable progress has been made toward answering it.
  - For a visualization, a chart is produced but with some flaws such as bad encoding. The conclusion based on the visualization is incomplete but makes some sense.
- **Unsatisfactory** (0 points):
  - For a dataframe, no computation is performed, or the conclusion does not match what is computed at all.
  - For a visualization, no chart is created, or a chart with completely wrong results.

We will lean towards being generous with the grading. We might also either discuss in discussion or post on Piazza some exemplary analysis you have done (with your permission)!

You should have the following in your answers: * a question you want to explore about the data.

* either of the following: * a few computed dataframes. * a few visualizations. * a few sentences summarizing what you found based on your analysis and how that answered your question (not too long please!)

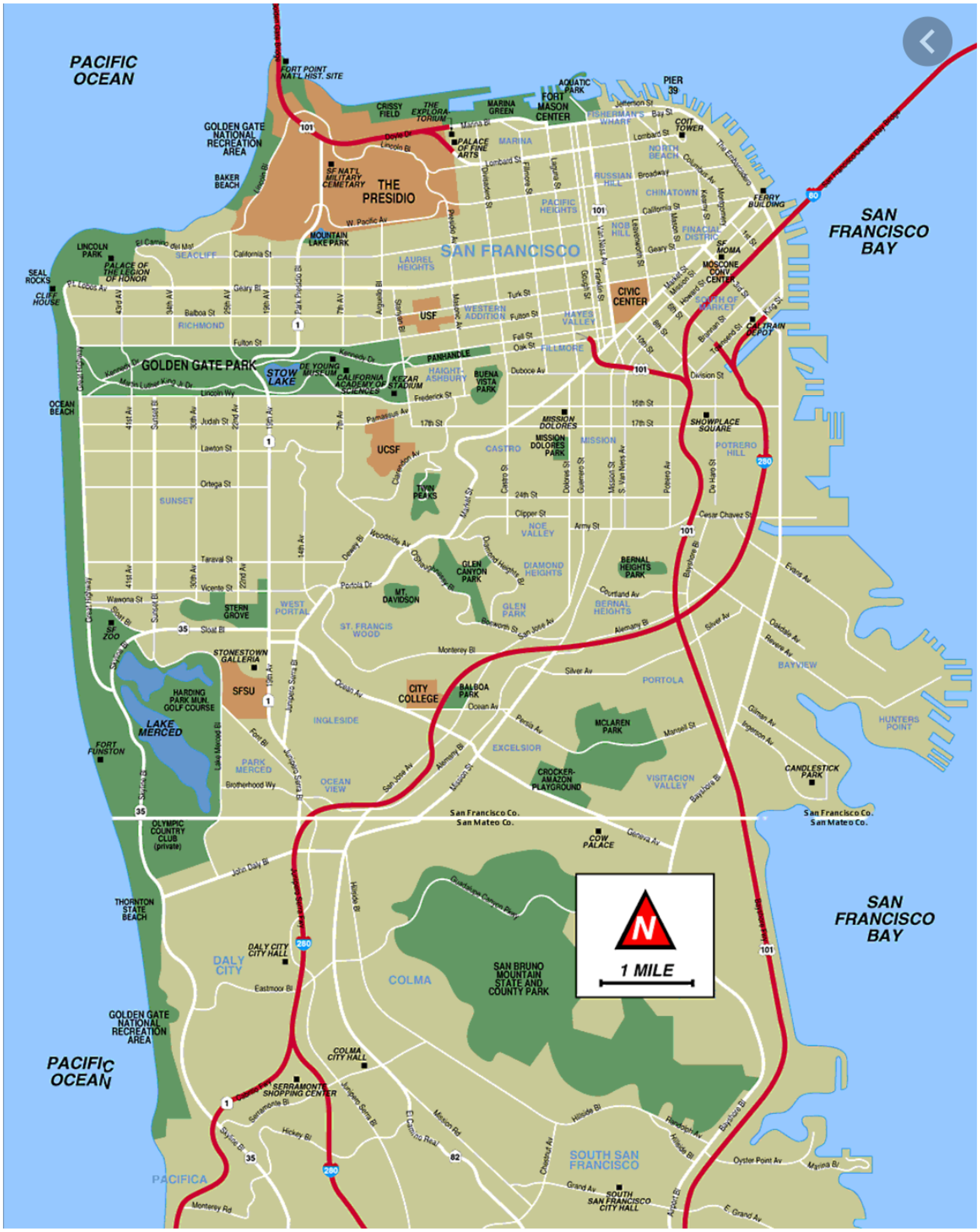
Please limit the number of your computed dataframes and visualizations **you plan on showing** to no more than 5.

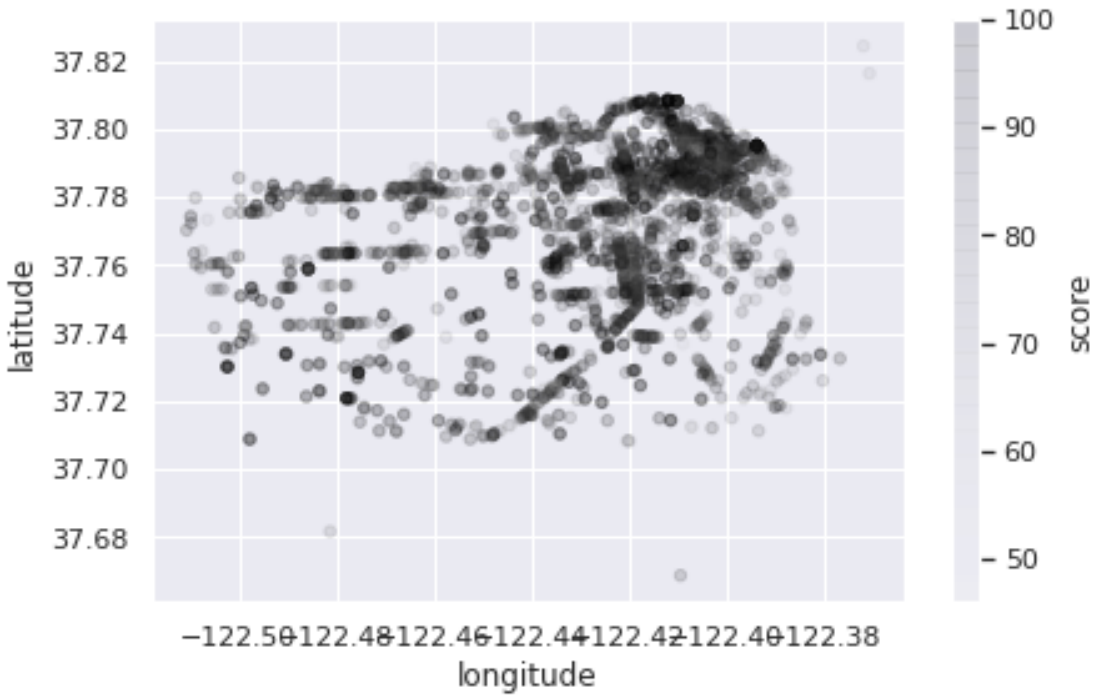
Please note that you will only receive support in OH and Piazza for Matplotlib and seaborn questions. However, you may use some other Python libraries to help you create you visualizations. If you do so, make sure it is compatible with the PDF export (e.g., Plotly does not create PDFs properly, which we need for Gradescope).

[78]:
```
# YOUR QUESTION HERE (in a comment)
#How do the inspection scores relate to the geolocation (latitude, longitude)␣
 ↪of a restaurant?

# YOUR DATA PROCESSING AND PLOTTING HERE
from mpl_toolkits.basemap import Basemap
from matplotlib.axes._axes import _log as matplotlib_axes_logger
matplotlib_axes_logger.setLevel('ERROR')

from IPython.display import Image
Image(filename='sf.png')

#Construct a DataFrame containing only the businesses which DO NOT have valid␣
 ↪ZIP codes.
valid_zip_bus = bus[bus["postal_code"].isin(valid_zips)]

#filter latitude and longtitude based on valid ranges
lat_filter = valid_zip_bus[(valid_zip_bus["latitude"] >= -90) &␣
 ↪(valid_zip_bus["latitude"] <= 90)]
long_lat_filter = valid_zip_bus[(valid_zip_bus["longitude"] >= -180) &␣
 ↪(valid_zip_bus["longitude"] <= 180)]
merge_score = long_lat_filter.merge(ins_named, how = "inner", on = "bid")

locations = merge_score[['latitude', 'longitude']]
locationlist = locations.values.tolist()

#color density of data based on score
merge_score = merge_score[['longitude','latitude', 'score']]
scatterplot = merge_score.plot(kind='scatter', x="longitude", y="latitude",␣
 ↪alpha=1/15, c=2)
scatterplot

from IPython.display import Image
Image(filename='sf.png', width=300, height=600)

# YOUR SUMMARY AND CONCLUSION HERE (in a comment)
```

```
#Based on the scatterplot, there seems to be an association between the␣
 ↪inspection count and the inspection score.
#Namely, that the upper right corner of the scatterplot have higher inspection␣
 ↪frequencies and higher inspection scores.
#This trend makes sense because from the San Francisco map displayed, we can␣
 ↪see that the corresponding area has a
#larger longitude and latitude degree. These regions are popular tourist spots␣
 ↪(fisherman's wharf, civic center,
#Chinatown) and financial districts of SF. The further away from these areas,␣
 ↪the fewer the inspection frequency,
#and the lower the inspection score, as illustrated in the lighter distribution␣
 ↪around the middle and left-end of
#the scatterplot. One inference for this observation is that restaurants in␣
 ↪these areas typically attract more people,
#and thus might have comparatively more competition. This competition might␣
 ↪push restaurants to maintain a higher
#standard and therefore, a higher inspection score.
```

[78]:


PACIFIC OCEAN
SAN FRANCISCO BAY
SAN FRANCISCO
THE PRESIDIO
GOLDEN GATE PARK
LAKE MERCED
DALY CITY
COLMA
SAN BRUNO MOUNTAIN STATE AND COUNTY PARK
SOUTH SAN FRANCISCO
PACIFICA
N
1 MILE


37.82
37.80
37.78
37.76
37.74
37.72
37.70
37.68
latitude
−122.50
122.48
122.46
122.44
122.42
122.40
122.38
longitude
100
90
80
70
60
50
score

[ ]: