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 [17]: bus.head(50)
         #ins2vio.head()
         #ins.head()
         #vio.head()
```

```
Out[17]:     business id column                                     name  \
         0                 1000                    HEUNG YUEN RESTAURANT
         1               100010                    ILLY CAFFE SF_PIER 39
         2               100017              AMICI'S EAST COAST PIZZERIA
         3               100026                           LOCAL CATERING
         4               100030                         OUI OUI! MACARON
         5               100036                          Hula Truck (#2)
         6               100039                 GENKI CREPES & MINI MART
         7               100041                           UNCLE LEE CAFE
         8               100055                            Twirl and Dip
         9               100058                              SF PITA HUB
         10              100059                           DUMPLING ALLEY
         11              100069                             Mission Blue
         12              100072                  SUBWAY SANDWICHES #7307
         13              100079                           POSITIVE FOODS
         14              100081     THE MATTERHORN RESTAURANT AND BAKERY
         15              100082                                SLN CTRNG
         16              100083                     THE EPICUREAN TRADER
         17              100084                            FRJTZ KITCHEN
         18              100096  THE LITTLE CHIHUAHUA MEXICAN RESTAURANT
         19              100097                              GANGNAM BBQ
         20              100098                    ZHONG SHAN RESTAURANT
         21              100099                             KEN KEE CAFE
         22              100126                Lamas Peruvian Food Truck
         23              100135        Hotel Whitcomb - Employee Kitchen
         24              100137                        100137 Cloud Club
         25              100142               Multi Service Center South
         26              100145                               Conchinita
         27                1002                        BIG MOUTH BURGERS
         28              100202                            FACEBOOK INC.
         29              100203                           FACEBOOK, INC.
         30              100204                CUIA ACAI & POSITIVE FOOD
         31              100205                             HUNAN EMPIRE
         32              100210                KING OF THAI NOODLE HOUSE
         33              100211                           THE EAGLE CAFE
         34              100212                         BELCAMPO MEAT CO
         35              100214                               REAL KABOB
         36              100215                  CHICKEN N WAFFLES PLACE
         37              100216                                BUNN MIKE
         38              100219                                NAYA CAFE
         39              100238                      PEACHES PATTIES LLC
         40              100239           HUNTINGTON HOTEL SAN FRANCISCO
         41              100240                         U :DESSERT STORY
```

```
42              100241                      Taqueria San Marcos
43              100252                       BITE ME SANDWICHES
44              100253                      BISTRO LOVESSY, LLC
45              100255                            H&M FOOD MART
46              100274                         THE OLYMPIC CAFE
47              100275         LITTLE CREATURES BREWING COMPANY
48              100277          COAST TO COAST ACAI AND GRANOLA
49              100278                     KINARA KITCHEN, INC.

                               address           city state postal_code  \
0                         3279 22nd St  San Francisco    CA       94110
1                     PIER 39  K-106-B  San Francisco    CA       94133
2                          475 06th St  San Francisco    CA       94103
3                     1566 CARROLL AVE  San Francisco    CA       94124
4               2200 JERROLD AVE STE C  San Francisco    CA       94124
5                        2 Marina Blvd  San Francisco    CA       94123
6                       330 CLEMENT ST  San Francisco    CA       94118
7                       3608 BALBOA ST  San Francisco    CA       94121
8        335 Martin Luther King Jr. Dr  San Francisco    CA       94118
9                          475 06TH ST  San Francisco    CA       94103
10                     2512 CLEMENT ST  San Francisco    CA       94121
11                      144 Leland Ave  San Francisco    CA       94134
12                      2375 MARKET ST  San Francisco    CA       94114
13                         475 06TH ST  San Francisco    CA       94103
14                   2323 VAN NESS AVE  San Francisco    CA       94109
15                       103 HORNE Ave  San Francisco    CA       94124
16                        465 HAYES ST  San Francisco    CA       94102
17                 475 06TH ST UNIT 15  San Francisco    CA       94103
18                     475 06TH ST K16  San Francisco    CA       94103
19                  3251 20TH AVE 250B  San Francisco    CA       94132
20                     2237 TARAVAL ST  San Francisco    CA       94116
21                     2109 CLEMENT ST  San Francisco    CA       94121
22                    Private Location  San Francisco    CA       -9999
23                      1231 Market St  San Francisco    CA       94103
24   24 Willie Mays Plaza Suites Level  San Francisco    CA       94107
25                         525 05th St  San Francisco    CA       94107
26           2 Marina Blvd  Fort Mason  San Francisco    CA       94123
27                        3392 24th St  San Francisco    CA       94110
28               181 FREMONT ST FL 5TH  San Francisco    CA       94105
29               181 FREMONT ST FL 6TH  San Francisco    CA       94105
30                   1 MARKET ST STE 8  San Francisco    CA       94105
31                  2001 UNION ST #107  San Francisco    CA       94123
32                    184 O'FARRELL ST  San Francisco    CA       94102
33                        39 PIER A201  San Francisco    CA       94133
34                         475 06TH ST  San Francisco    CA       94103
35                      475 06TH St 23  San Francisco    CA       94103
36                     1968 LOMBARD ST  San Francisco    CA       94123
37                      300 DE HARO ST  San Francisco    CA       94103
38                     5338 geary BLVD  San Francisco    CA       94121
39                      2948 FOLSOM ST  San Francisco    CA       94110
40                  1075 CALIFORNIA ST  San Francisco    CA       94108
41                   2120 GREENWICH ST  San Francisco    CA       94123
42                  2380 San Bruno Ave  San Francisco    CA       94134
43                         701 COLE ST  San Francisco    CA       94117
```

```
44                  832 CLEMENT ST  San Francisco    CA       94118
45              2400 SAN BRUNO AVE  San Francisco    CA       94134
46                    555 GEARY ST  San Francisco    CA       94102
47                  1000 A 03RD St  San Francisco    CA       94158
48                 160 14TH STREET  San Francisco    CA       94103
49                    607 GEARY ST  San Francisco    CA       94102

       latitude    longitude  phone_number
0     37.755282  -122.420493         -9999
1  -9999.000000 -9999.000000   14154827284
2  -9999.000000 -9999.000000   14155279839
3  -9999.000000 -9999.000000   14155860315
4  -9999.000000 -9999.000000   14159702675
5  -9999.000000 -9999.000000         -9999
6  -9999.000000 -9999.000000   14155376414
7  -9999.000000 -9999.000000         -9999
8  -9999.000000 -9999.000000   14155300260
9  -9999.000000 -9999.000000   14155642006
10 -9999.000000 -9999.000000         -9999
11 -9999.000000 -9999.000000         -9999
12 -9999.000000 -9999.000000   14155981866
13 -9999.000000 -9999.000000   14155397209
14 -9999.000000 -9999.000000   14155474029
15 -9999.000000 -9999.000000   14155965620
16 -9999.000000 -9999.000000   14155606092
17 -9999.000000 -9999.000000   14155868272
18 -9999.000000 -9999.000000         -9999
19 -9999.000000 -9999.000000   14150494183
20 -9999.000000 -9999.000000   14155806898
21 -9999.000000 -9999.000000   14155699118
22 -9999.000000 -9999.000000         -9999
23 -9999.000000 -9999.000000         -9999
24 -9999.000000 -9999.000000         -9999
25 -9999.000000 -9999.000000         -9999
26 -9999.000000 -9999.000000         -9999
27    37.752158  -122.420362         -9999
28 -9999.000000 -9999.000000   14150799045
29 -9999.000000 -9999.000000   14150799045
30 -9999.000000 -9999.000000   14158609815
31 -9999.000000 -9999.000000   14155774735
32 -9999.000000 -9999.000000   14155821999
33 -9999.000000 -9999.000000   14155985872
34 -9999.000000 -9999.000000   14157800656
35 -9999.000000 -9999.000000   14158705851
36 -9999.000000 -9999.000000   14156425140
37 -9999.000000 -9999.000000   14155299775
38 -9999.000000 -9999.000000   14155995527
39 -9999.000000 -9999.000000         -9999
40 -9999.000000 -9999.000000   14155342803
41 -9999.000000 -9999.000000   14155333435
42 -9999.000000 -9999.000000         -9999
43 -9999.000000 -9999.000000   14155665282
44 -9999.000000 -9999.000000   14155827593
45 -9999.000000 -9999.000000   14159277470
```

```
46 -9999.000000 -9999.000000   14155718182
47 -9999.000000 -9999.000000   14153334433
48 -9999.000000 -9999.000000   14159230622
49 -9999.000000 -9999.000000         -9999
```

- In the bus dataframe we see that the address column has different fromatting. For example, 475 06th St vs 1566 CARROLL AVE
- The phone number column also has missing values denoted as -9999
- the latitude and longitude columns are missing

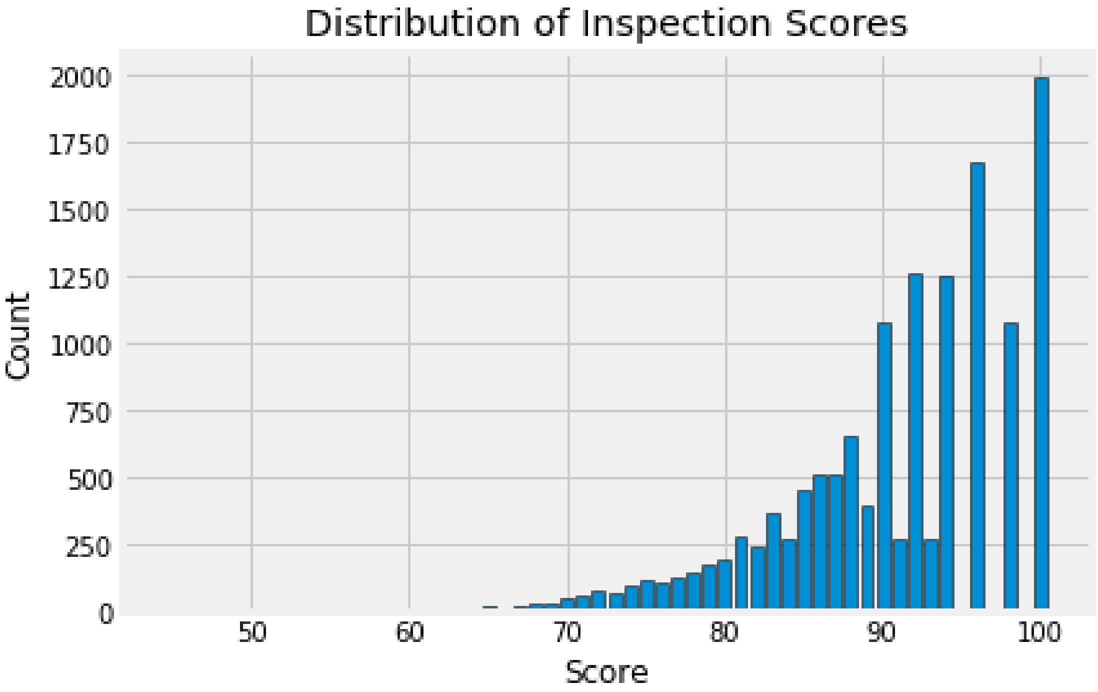
In the cell below, write the name of the restaurant with the lowest inspection scores ever. You can also head to yelp.com and look up the reviews page for this restaurant. Feel free to add anything interesting you want to share.

’Lollipot” this is true because it shows the WORST score. I could not find this on yelp, maybe it was closed because the restaurant had the worst score.

0.1 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.



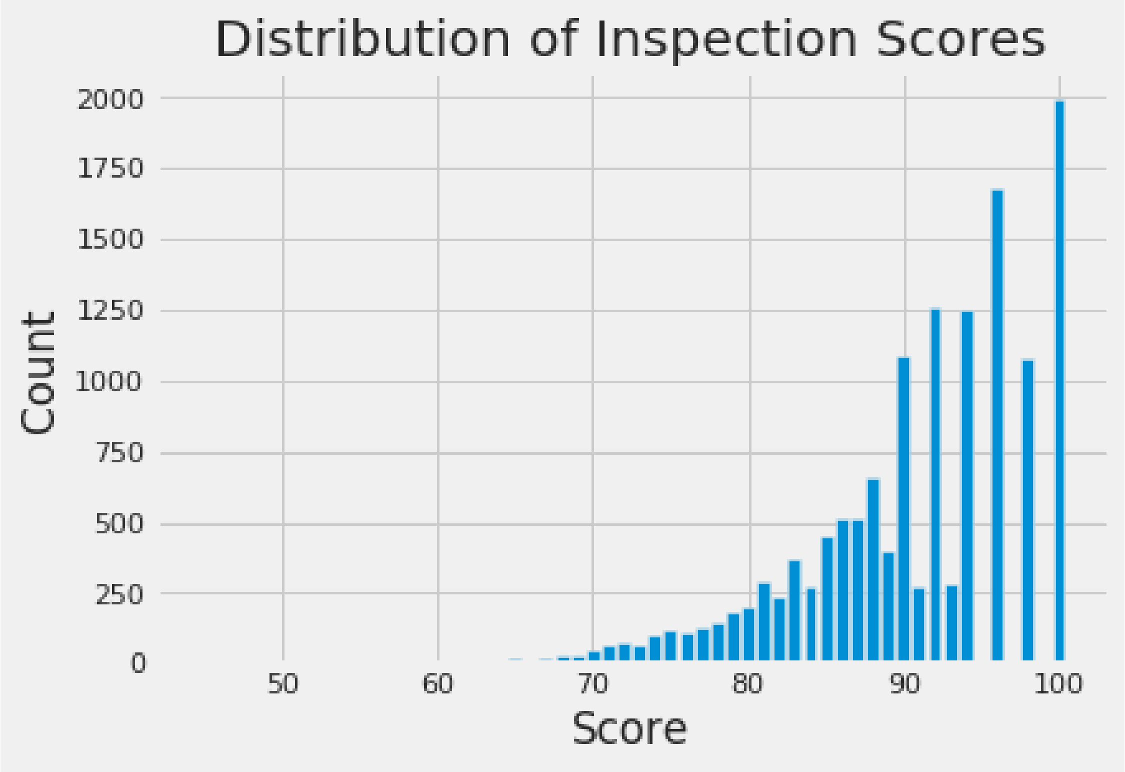
You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need:

```
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 [78]: x = ins['score'].value_counts().keys()
         y = ins['score'].value_counts()
         plt.bar(x,y)
         plt.xlabel('Score')
         plt.ylabel('Count')
         plt.title('Distribution of Inspection Scores')
```

```
Out[78]: Text(0.5, 1.0, 'Distribution of Inspection Scores')
```

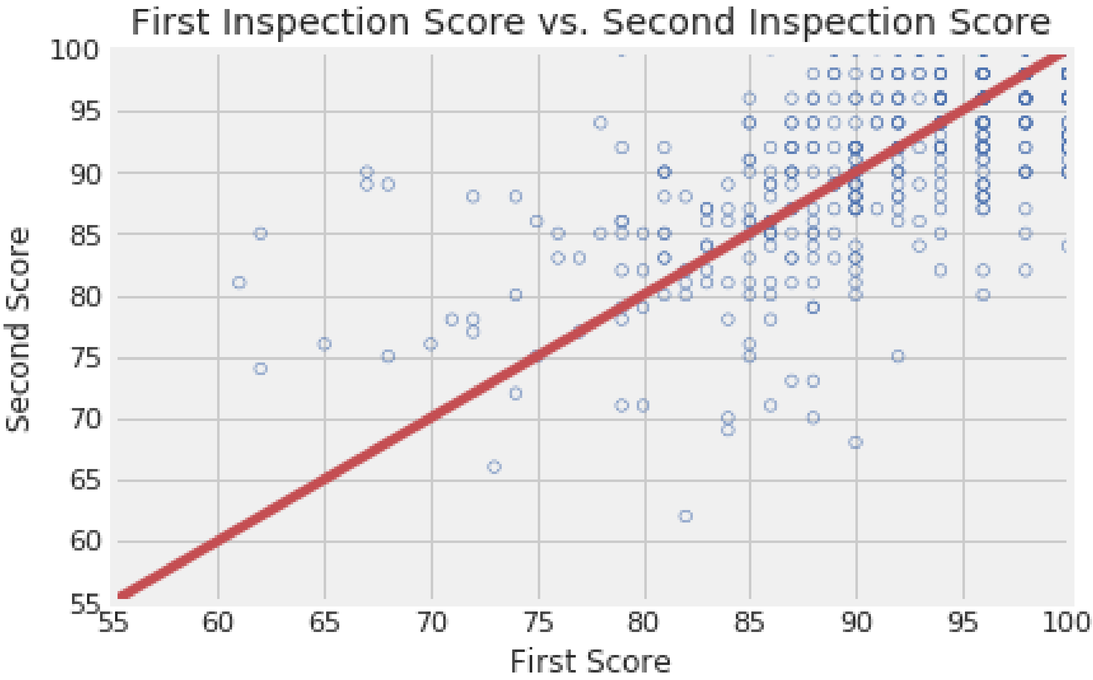


0.1.1 Question 6b

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

- The mode is the most frequent score and in the distribution above, the most frequent score is 100. We also have some other high scores that include 97,95,93,91.
- There is no evidence of symmetry in this graph. Instead this is a negatively skewed graph, the peak is on the right side with a relatively long left negative tail.
- We see gaps in three places: x=94, 96, 98.
- It seems that there is no scores that equal to the gaps of 94, 96, 98 which makes me wonder how the subcategories are calculated. Maybe there is no way to get those scores mathematically. I am pleasantly surpised that the highest scores are 100/100. This means that most restaurants are following the insepction codes and also the other majority of the scores seem to be in the upper 90s in terms of score out of 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 sample below, but make sure that all labels, axes and data itself are correct.



Key pieces of syntax you'll need:

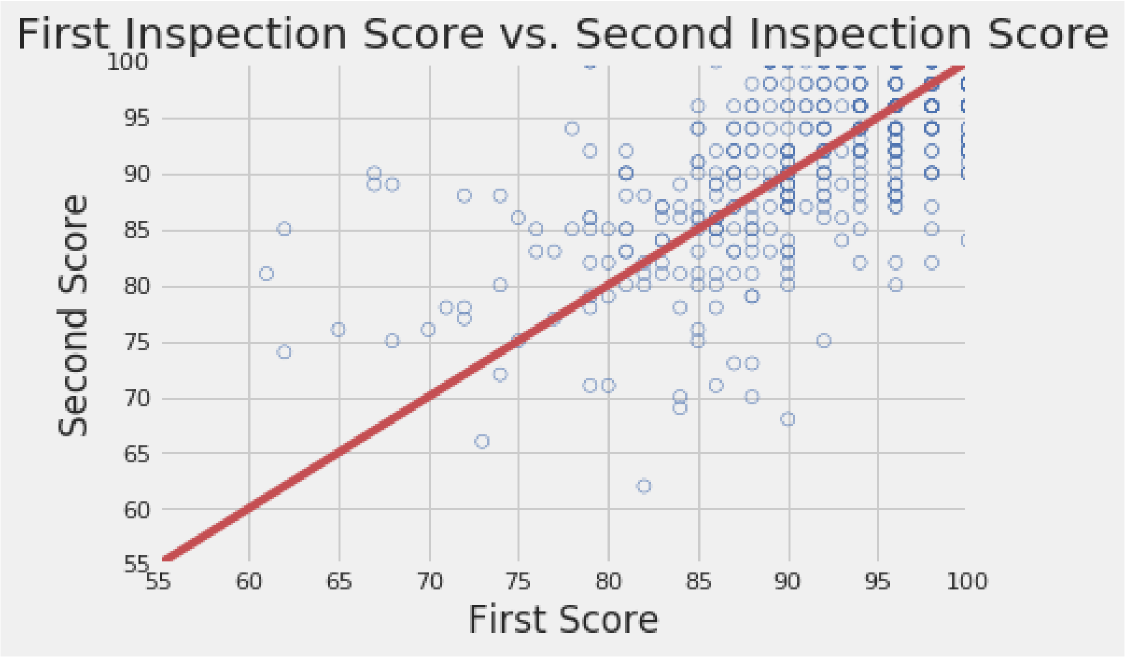
`plt.scatter` plots a set of points. Use `facecolors='none'` and `edgecolors=b` to make circle markers with blue borders.

`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.

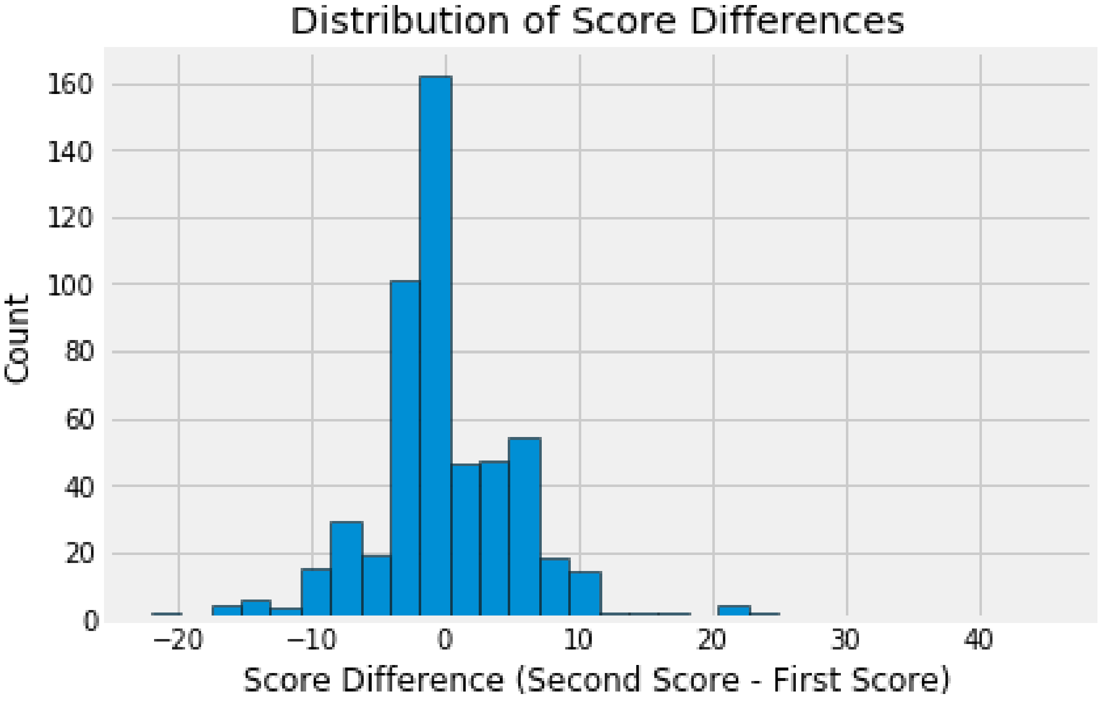
```
In [87]: x = scores_pairs_by_business['score_pair'].agg(lambda x: x[0]).to_list()
         y = scores_pairs_by_business['score_pair'].agg(lambda x: x[1]).to_list()
         plt.scatter(x, y, facecolors = 'none', edgecolors='b')
         plt.axis([55, 100, 55, 100])
         plt.plot([55,100],[55,100], c='r')
         plt.xlabel('First Score')
         plt.ylabel('Second Score')
         plt.title('First Inspection Score vs. Second Inspection Score');
```


First Inspection Score vs. Second Inspection Score
Second Score
100
95
90
85
80
75
70
65
60
55
55
60
65
70
75
80
85
90
95
100
First Score

0.1.2 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:

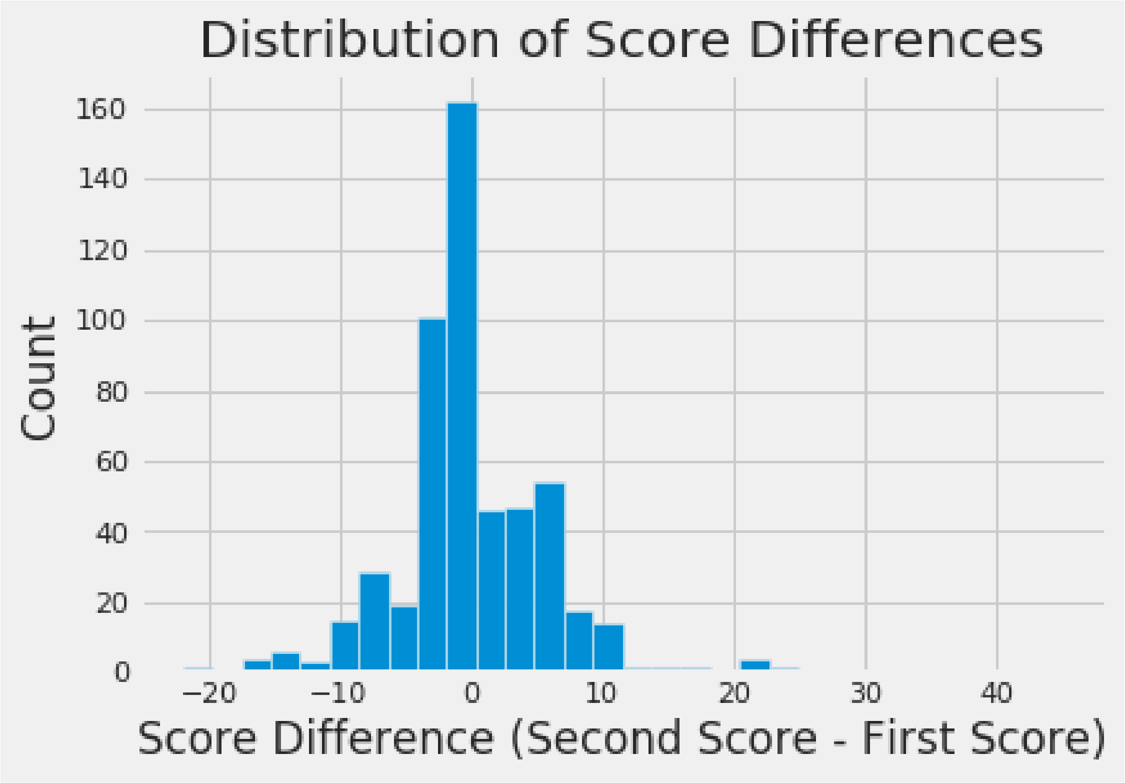


Hint: Use `second_score` and `first_score` created in the scatter plot code above.

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 [88]: diff = [score_pair[1] - score_pair[0] for t in scores_pairs_by_business.values for score_pair
         plt.hist(diff, bins=30);
         plt.xlabel('Score Difference (Second Score - First Score)')
         plt.ylabel('Count')
         plt.title('Distribution of Score Differences');
```


Distribution of Score Differences
Count
0
20
40
60
80
100
120
140
160
−20
−10
0
10
20
30
40
Score Difference (Second Score - First Score)

0.1.3 Question 7e

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?

From the scatter plot we can see that the graph has a positive relationship between the First and Second Scores. The red line is the best fit line that is computed with all the blue points. It looks like most of the data is clustered around 100 for both the first and second scores.

0.1.4 Question 7f

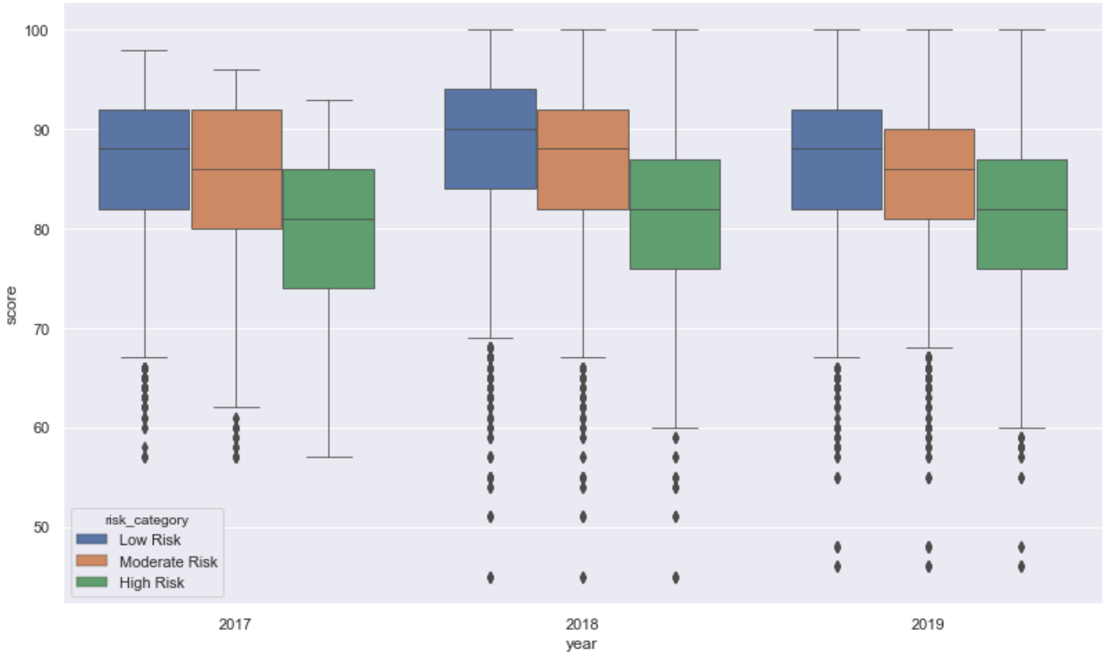
If a restaurant's score improves from the first to the second inspection, how would this be reflected in the histogram of the difference in the scores that you made in question 7d? What do you oberve from the plot? Are your observations consistent with your expectations? Explain your observations in the language of Statistics: for instance, the center, the spread, the deviation etc.

Yes the idea behind the histogram is that your second score should be higher than the first. This would result in a bigger remainder. If difference is negative then that means the second inspection was worst than the first. my observations are consistent with my expectation becuase most of the data was clusered in the high 80s to 100 and so I would assume that the score would be better the second time around. I think the one part im confused on is why the highest count is a negative vlaue on the difference scale however, I would assume that the original score was very close to the high 90s.

0.1.5 Question 7g

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!

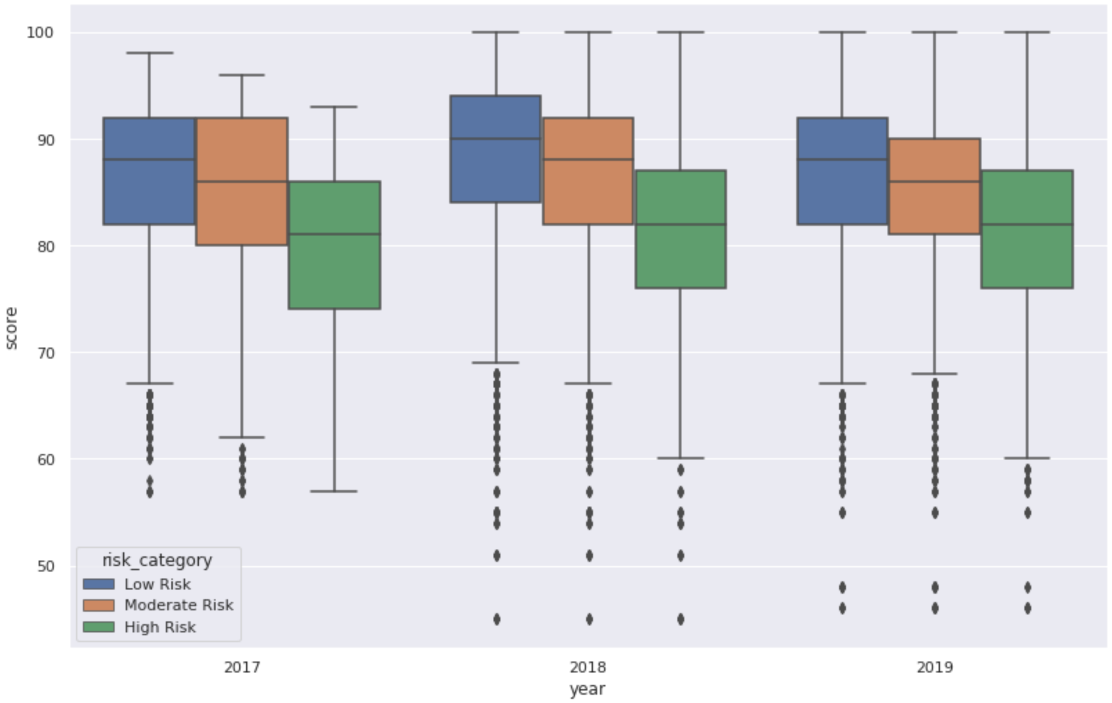


Hint: Use `sns.boxplot()`. Try taking a look at the first several parameters. The documentation is linked here!

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

```
In [89]: # Do not modify this line
         sns.set()
         plt.figure(figsize = (12,8))
         desmerge = vio.merge(ins2vio, how='left', on='vid')
         insmerge = ins.merge(desmerge, on='iid')
         merge = insmerge[insmerge['year']>= 2017]

         sns.boxplot(x='year', y='score', data = merge, hue='risk_category', hue_order=['Low Risk', "Mo
```


100
90
80
70
60
50
score
2017
2018
2019
year
risk_category
Low Risk
Moderate Risk
High Risk

1 8: Open Ended Question

1.1 Question 8a

1.1.1 Compute Something Interesting

Play with the data and try to compute something interesting about the data. Please try to use at least one of groupby, pivot, or merge (or all of the above).

Please show your work in the cell below and describe in words what you found in the same cell. This question will be graded leniently but good solutions may be used to create future homework problems.

1.1.2 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): Uses a combination of pandas operations (such as groupby, pivot, merge) to answer a relevant question about the data. The text description provides a reasonable interpretation of the result.
- **Passing** (1-3 points): Computation is flawed or very simple. The text description is incomplete but makes some sense.
- **Unsatisfactory** (0 points): No computation is performed, or a computation with completely wrong results.

Please have both your code and your explanation in the same one cell below. Any work in any other cell will not be graded.

```
In [112]: q8a = vio.drop(columns=['vid']).groupby(by='risk_category').count()
          q8a['% total'] = (q8a['description']/sum(q8a['description']))*100
          q8a

          #ins.iloc[0]

          # The first observation that I see from the output below is the inspections have less than
          # half with low risk category score. I also noticed that if you add the high risk and moderat
          # risk you get majority of the data (24.61+32.30 = 56.81%). From the data analytic standpoint
          # this is concerning that most inspections were barely passing/high risk for many violations.
```

```
Out[112]:                description    % total
          risk_category
          High Risk               16  24.615385
          Low Risk                28  43.076923
          Moderate Risk           21  32.307692
```

1.1.3 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): The chart is well designed, and the data computation is correct. The text written articulates a reasonable metric and correctly describes the relevant insight and answer to the question you are interested in.
- **Passing** (1-3 points): A chart is produced but with some flaws such as bad encoding. The text written is incomplete but makes some sense.
- **Unsatisfactory** (0 points): No chart is created, or a chart with completely wrong results.

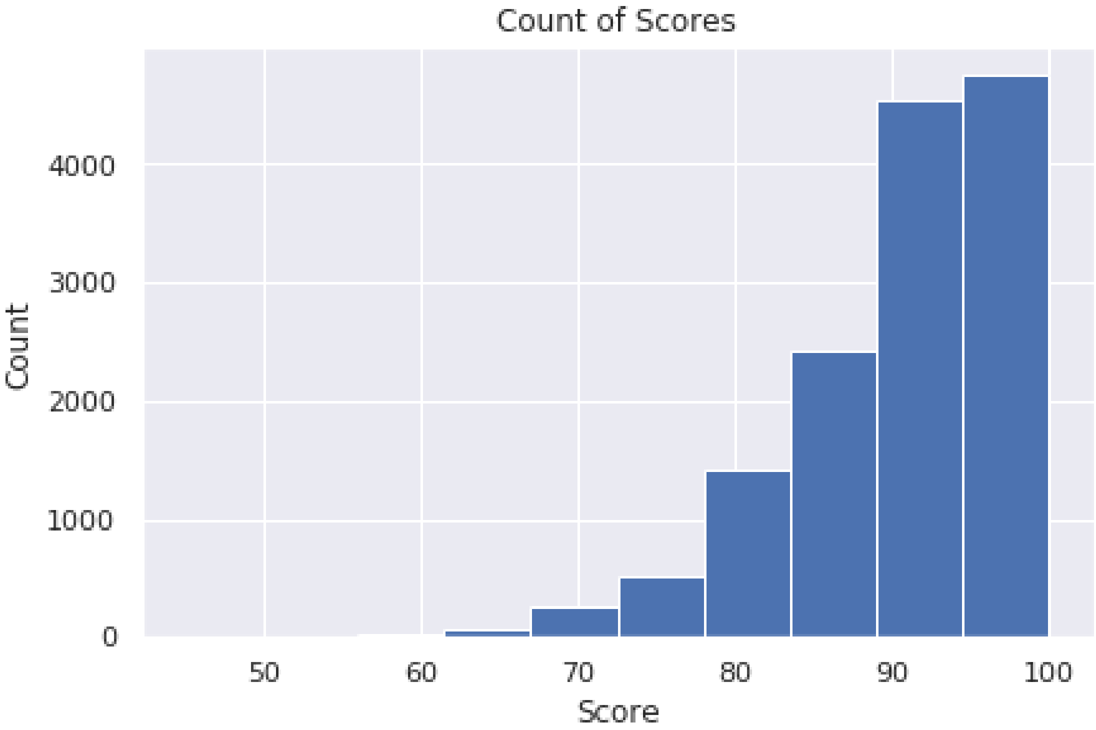
We will lean towards being generous with the grading. We might also either discuss in discussion or post on Piazza some examplar analysis you have done (with your permission)!

You should have the following in your answers: * a few visualizations; Please limit your visualizations to 5 plots. * a few sentences (not too long please!)

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).

```
In [ ]: 0,70,"Poor"
            71,85,"Needs Improvement"
            86,90,"Adequate"
            91,100,"Good"
```

```
In [177]: x_ins= ins['year']
          y_ins = ins['score']
          plt.hist(y_ins)
          #plt.axis([2016, 2019, 0, 100])
          plt.xlabel('Score')
          plt.ylabel('Count')
          plt.title('Count of Scores');
          # This is a great representation that majority of the inspections that take place are
          # in good standing. They seem to stay in the upper 80s to 100. I think what is interesting
          # about this graph is that there are a good amount of restaurants
```


Count of Scores
Count
4000
3000
2000
1000
0
50
60
70
80
90
100
Score