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]: display(bus.head(), ins.head(), vio.head())
```

```
   business id column                          name                  address  \
0                1000         HEUNG YUEN RESTAURANT             3279 22nd St
1              100010         ILLY CAFFE SF_PIER 39         PIER 39  K-106-B
2              100017   AMICI'S EAST COAST PIZZERIA              475 06th St
3              100026                LOCAL CATERING         1566 CARROLL AVE
4              100030              OUI OUI! MACARON   2200 JERROLD AVE STE C

            city state postal_code     latitude    longitude  phone_number
0  San Francisco    CA       94110    37.755282  -122.420493         -9999
1  San Francisco    CA       94133 -9999.000000 -9999.000000   14154827284
2  San Francisco    CA       94103 -9999.000000 -9999.000000   14155279839
3  San Francisco    CA       94124 -9999.000000 -9999.000000   14155860315
4  San Francisco    CA       94124 -9999.000000 -9999.000000   14159702675
```

```
               iid                    date  score                    type
0  100010_20190329  03/29/2019 12:00:00 AM     -1       New Construction
1  100010_20190403  04/03/2019 12:00:00 AM    100  Routine - Unscheduled
2  100017_20190417  04/17/2019 12:00:00 AM     -1          New Ownership
3  100017_20190816  08/16/2019 12:00:00 AM     91  Routine - Unscheduled
4  100017_20190826  08/26/2019 12:00:00 AM     -1  Reinspection/Followup
```

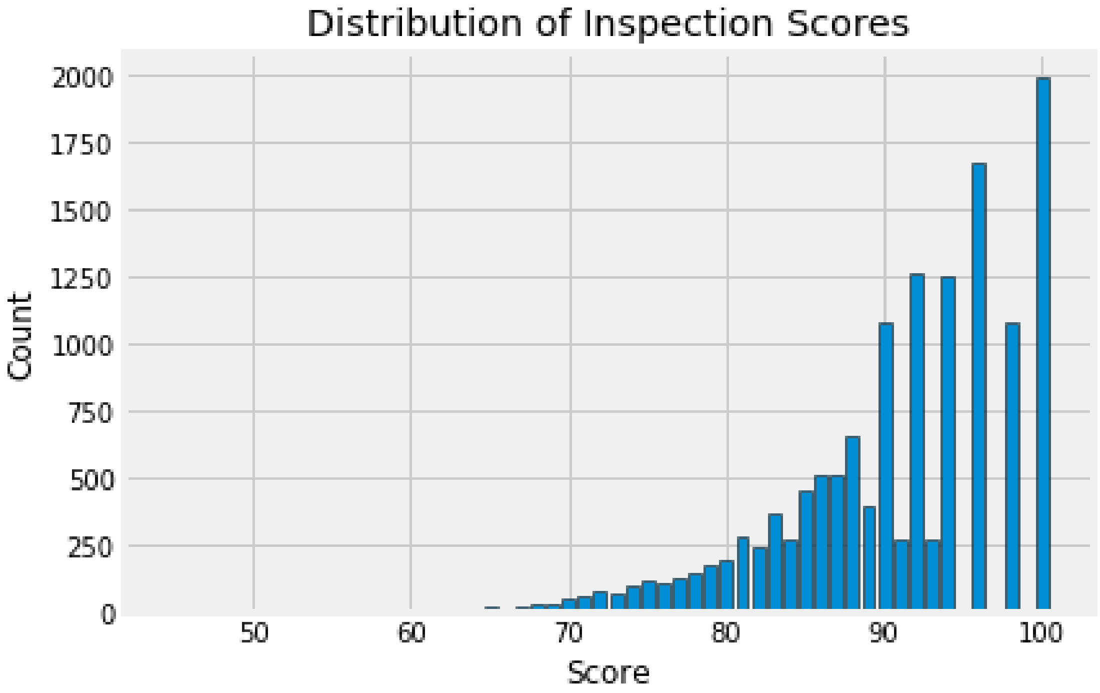
```
                                         description  risk_category     vid
0  Consumer advisory not provided for raw or unde...  Moderate Risk  103128
1                   Contaminated or adulterated food      High Risk  103108
2          Discharge from employee nose mouth or eye  Moderate Risk  103117
3                         Employee eating or smoking  Moderate Risk  103118
4                             Food in poor condition  Moderate Risk  103123
```

One potential problem is an incorrect value, such as the phone number value for the first entry in the business table; the value for number is -9999, which is an invalid phone number. Another issue may be the inconsistent address formatting in the business table. The last issue I see are the longitude and latitude values in the business column – -9999.000000 is not a valid value for langitude and longitude coordinates.

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.

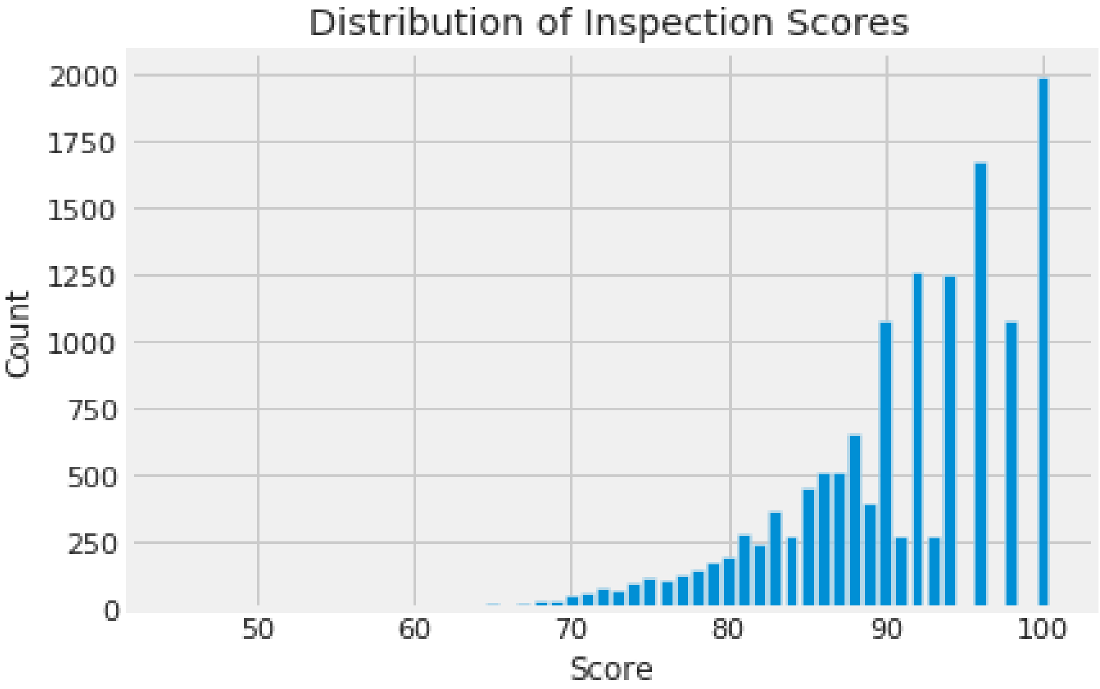


You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need:

```
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 [74]: ins_pos = ins[ins["score"] > 0]
         plt.bar(ins_pos['score'].value_counts().keys(), ins_pos['score'].value_counts())
         plt.xlabel('Score')
         plt.ylabel('Count')
         plt.title('Distribution of Inspection Scores');
```



0.1.1 Question 6b

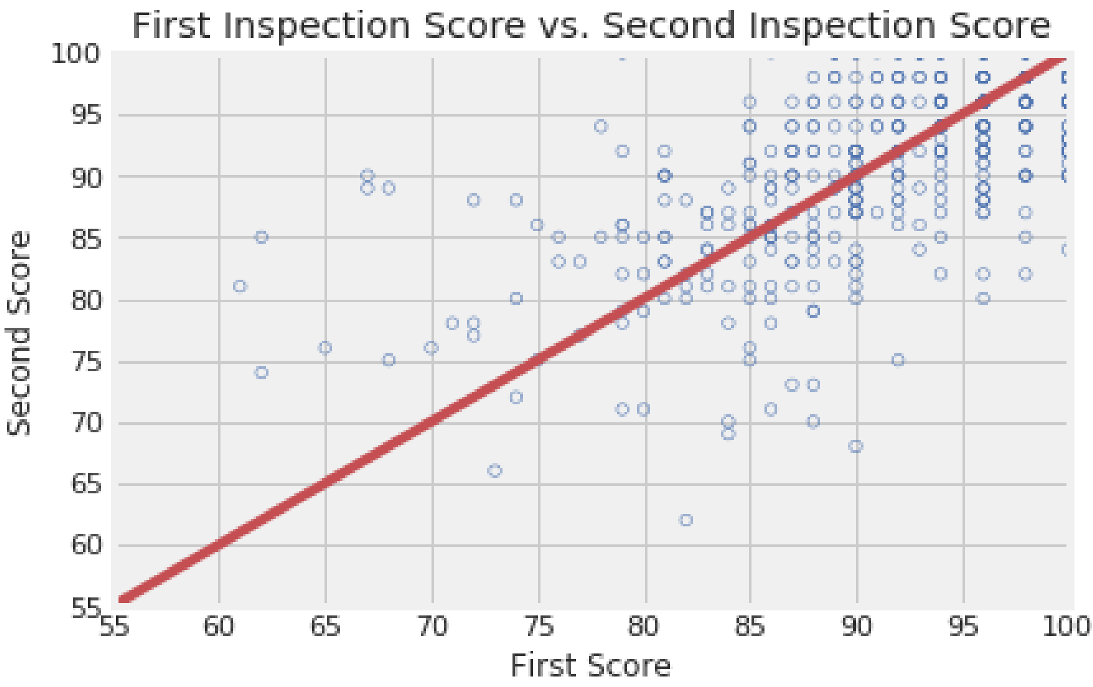
Describe the qualities of the distribution of the inspections scores based on your bar plot. Consider the mode(s), symmetry, tails, gaps, and anomalous values. Are there any unusual features of this distribution? What do your observations imply about the scores?

There seems to be one mode, with its peak at 100. It is not symmetrical, but instead skewed to the left, with the majority of scores located to the far right of the distribution. There are a few gaps between 90-100, meaning no restraurants received the scores represented at the gap, and this might imply that these scores are either not given out or extremely hard to receive. Lower scores have lower counts and as the scores decrease so do the counts. There are also virtually no restraurants with a score that is less than 60 which may imply that restraurants generally don't score that low or are closed due to a low score. Overall, it seems like most restraurants have high scores.

Use the cell above to identify the restaurant with the lowest inspection scores ever. Be sure to include the name of the restaurant as part of your answer in the cell below. 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

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.

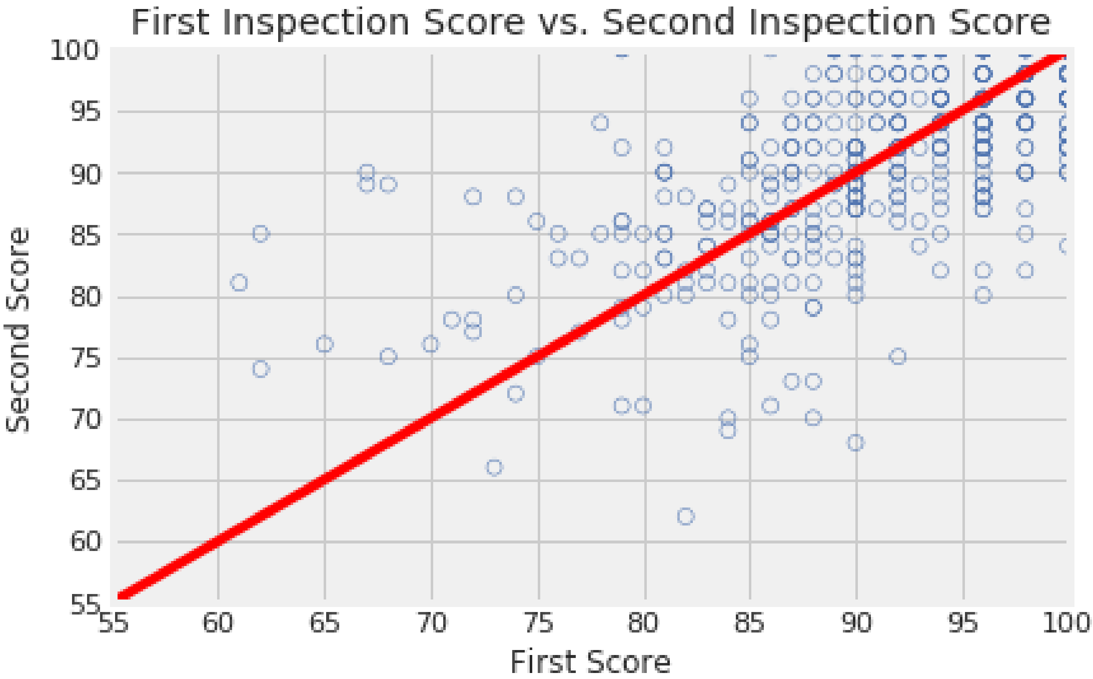
`plt.plot` for the reference line.

`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 [85]: x = [i[0] for i in scores_pairs_by_business['score_pair']]
         y = [i[1] for i in scores_pairs_by_business['score_pair']]
         plt.scatter(x, y, facecolors = 'none', edgecolors = 'b')
         plt.plot([55,100], [55,100], color = 'red')
         plt.xlabel('First Score')
         plt.ylabel('Second Score')
         plt.title('First Inspection Score vs. Second Inspection Score')
         plt.axis(xmin = 55 , xmax = 100, ymin = 55, ymax = 100)
```

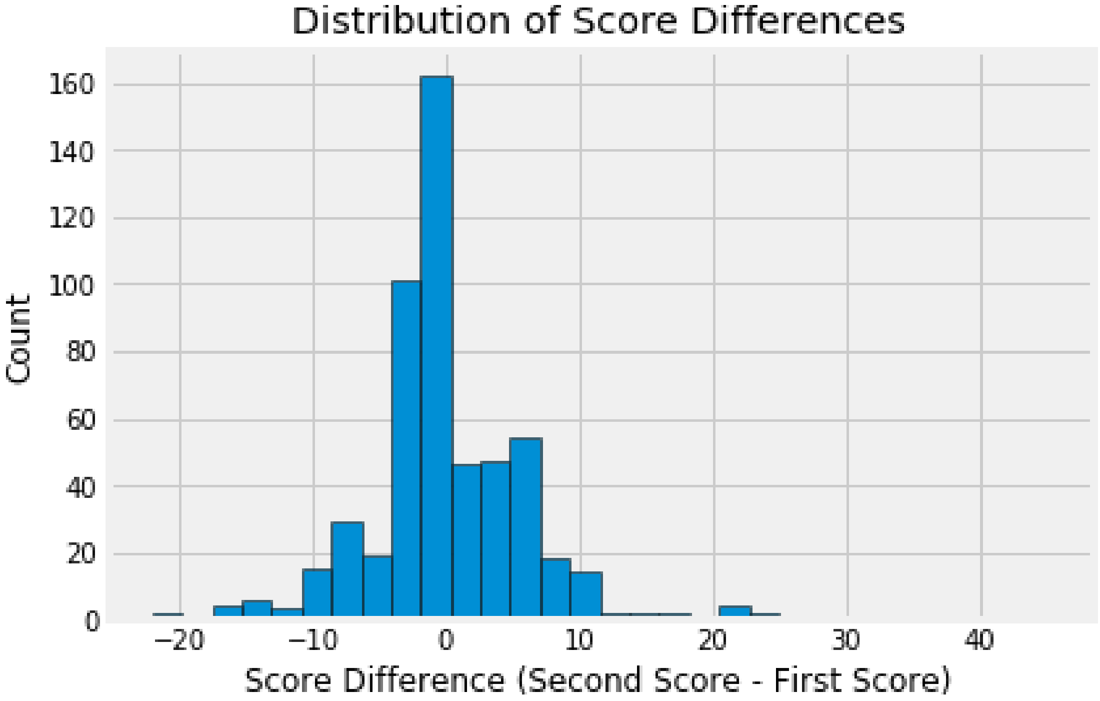
```
Out[85]: (55, 100, 55, 100)
```


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

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:

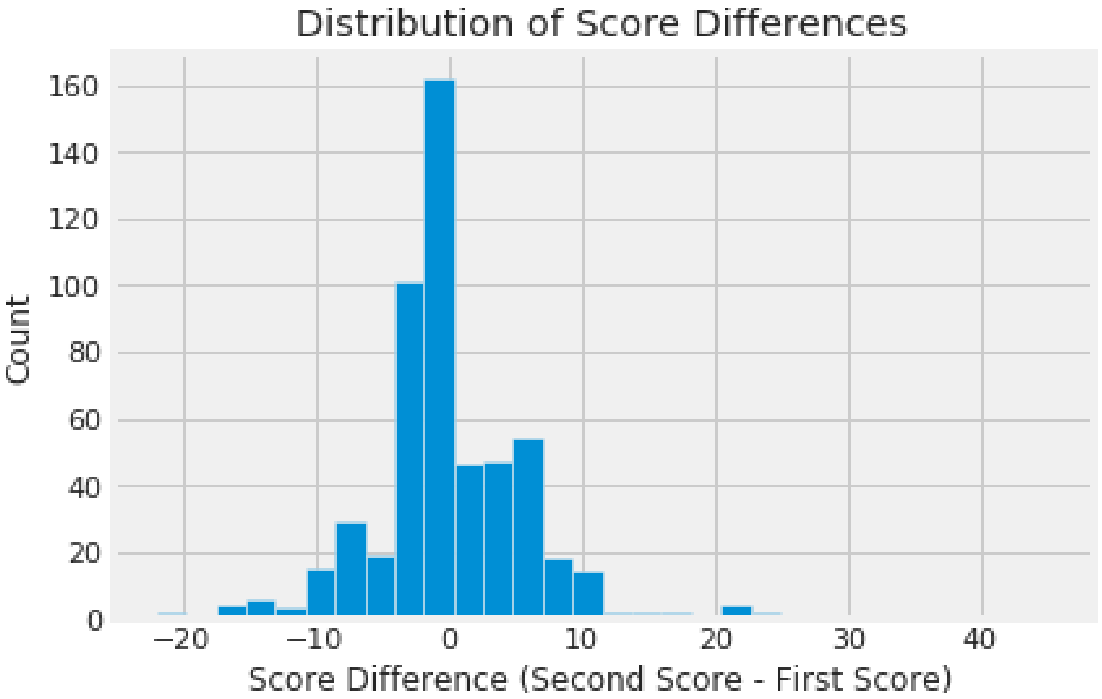


Hint: Use `second_score` and `first_score` created in the scatter plot code above.

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 [86]: score_difference = np.array([i[1] - i[0] for i in scores_pairs_by_business['score_pair']])
         plt.hist(score_difference, 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 2c? What do you oberve from the plot? Are your observations consistent with your expectations?

Hint: What does the slope represent?

Since the red reference line indicates where the points would be located if the first score were to be equal to the second score, any points above the line indicate that the second score was higher than the first and any points below the line indicate that the second scores was lower than the first. If we wanted to check for improvement from the first to the second inspection, we would expect to see more points above the reference line. From the plot, it seems like half of the points are above the line while half of the points are below the line, indicating that half of the restraurants saw improvements while the other half saw worse scores.

0.1.4 Question 7f

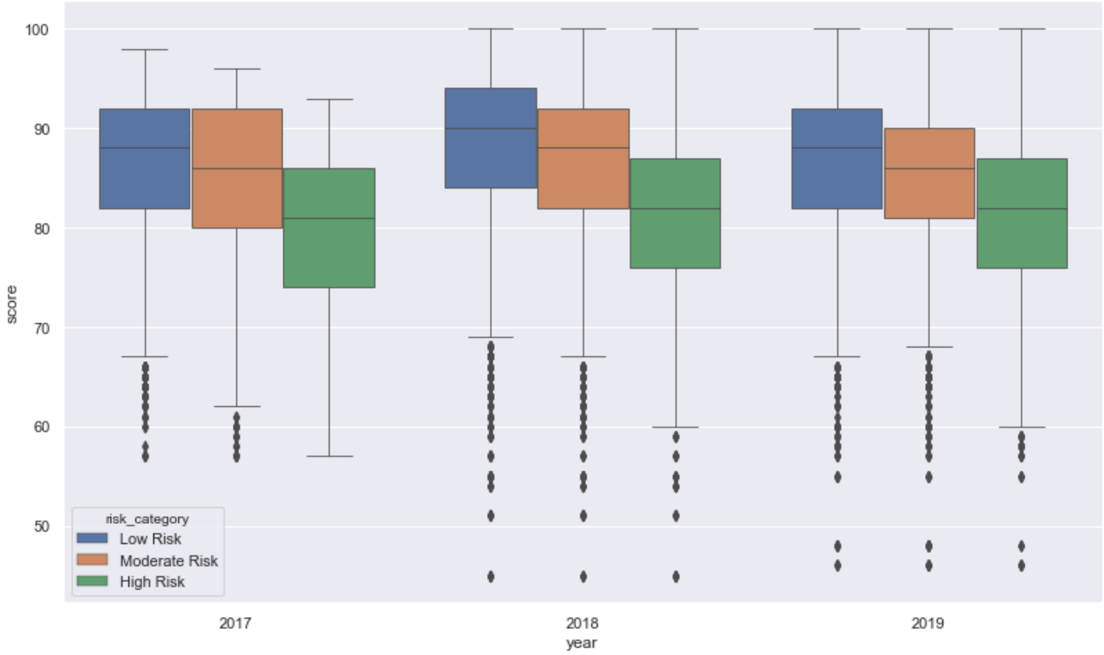
If a restaurant's score improves from the first to the second inspection, how would this be reflected in the histogram of the difference in the scores that you made in question 8d? What do you oberve from the plot? Are your observations consistent with your expectations? Explain your observations in the language of Statistics: for instance, the center, the spread, the deviation etc.

If a restraurant's score were to improve from the first to second inspection, the histogram should be either centered at a positive difference value or skewed to the left with the center at a positive value. If this were the case, this would indicate that most restraurants saw an improvement in their inspection scores because a positive difference means the second score was lower than the first. What we see in the histogram is a mean that is centered at zero, with a large spike in frequency at a bin that is just below zero. Aside from this spike, the data approximately ranges from -20 to 20 with a relatively wide deviation. It seems like half of the differences are below the center and half are above, which I expected from the analysis of the scatter plot. Half of the restraurants saw an improvement (positive difference) while the other half did not (negative difference).

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.

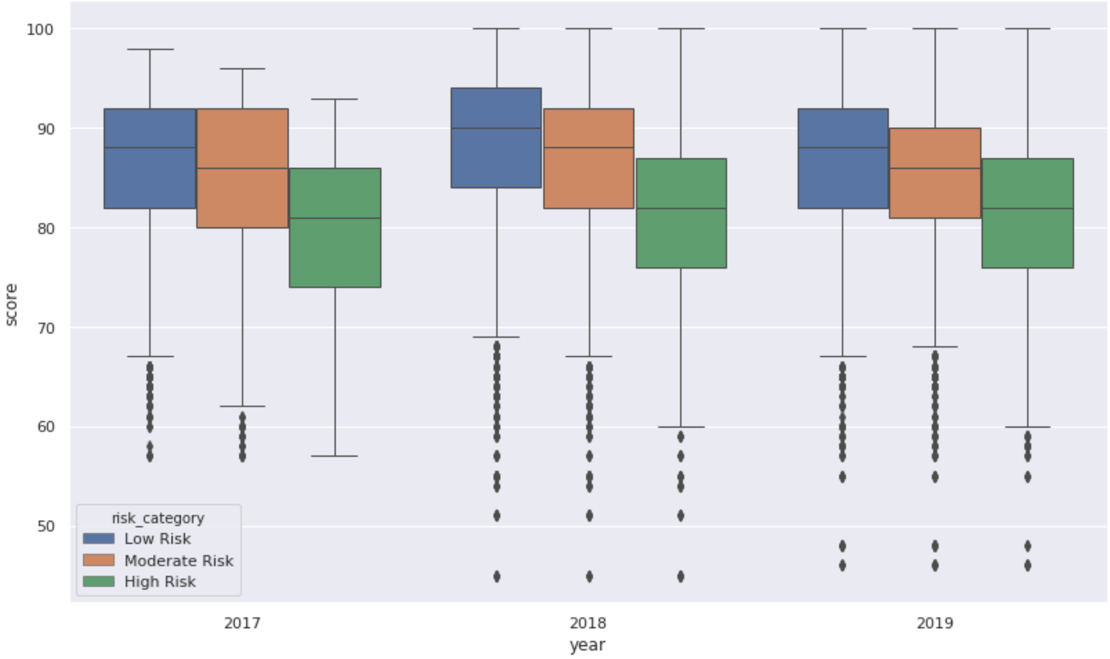
The boxplot should look similar to the sample below. Make sure the boxes are in the correct order!



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

         ins171819 = ins[(ins['year'] >= 2017) & (ins['year'] <= 2019) & (ins['Missing Score'] == False
         insvio = ins171819.merge(ins2vio, how = "left", left_on = "iid", right_on = "iid")
         insvio = insvio.merge(vio, how = "left", left_on = "vid", right_on = "vid")
         insvio = insvio[(insvio['risk_category'] == "Low Risk") | (insvio['risk_category'] == "High Ri

         plt.figure(figsize = (12,8))
         sns.boxplot(x="year", y="score", hue="risk_category", data=insvio, hue_order = ["Low Risk", "M
Out[87]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdd8cd7d850>
```


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

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.

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 [88]: #YOUR CODE HERE

         ins_bus = ins[ins["score"] > 0].merge(ins2vio, how = "left", left_on = "iid", right_on = "iid"
         ins_bus = ins_bus.merge(vio, how = "left", left_on = "vid", right_on = "vid")
         ins_bus1 = ins_bus.groupby("description").mean()[["score", "vid"]]
         ins_bus2 = ins_bus.groupby("description").count()[['iid']]
         ins_bus = ins_bus1.merge(ins_bus2, how = "left", left_on = "description", right_on = "descript

         #YOUR EXPLANATION HERE (in a comment)
         """I wanted to see if certain violations were associated with higher or lower scores. To do th
         and ins2vio table where the scores were above 0 and then merged that table to the vio table so
         would be on the same table with the violation descriptions. I then found grouped by the descri
         the mean scores for every violation. I also added a column called count that I got from the gr
         this column to see if the mean scores were based off of few inspections. I found that having "
         within 200 feet of mobile food facility" had the lowest average score but this score was only
         Violations like "contaminated or adulterated foor" or "High risk vermin infestation" had reall
         high count values. Violations like "Mobile food facility not operating with an approved commis
         charge of food facility" had higher mean scores. To conclude whether these violations are/are
         with lower inspections score, I would have to run a deeper analysis.
         """

         ins_bus.head(20)
```

```
Out[88]:                                                     mean score       vid  \
         description
         No restroom facility within 200 feet of mobile …   62.000000  103171.0
         Noncompliance with Gulf Coast oyster regulation      71.750000  103126.0
```

```
Improper cooking time or temperatures                72.000000  103106.0
Contaminated or adulterated food                     74.183333  103108.0
Mobile food facility with unapproved operating …   75.400000  103172.0
Unapproved food source                               78.200000  103110.0
Sewage or wastewater contamination                   78.241379  103113.0
Other high risk violation                            78.629630  103115.0
High risk vermin infestation                         78.714870  103114.0
Unclean hands or improper use of gloves              79.199346  103102.0
Unapproved  living quarters in food facility         80.137931  103155.0
Improperly displayed mobile food permit or signage   80.750000  103175.0
Improper thawing methods                             81.084986  103132.0
Unclean or unsanitary food contact surfaces          81.100173  103109.0
Improper cooling methods                             81.142518  103105.0
Mobile food facility stored in unapproved location   81.400000  103173.0
High risk food holding temperature                   81.487788  103103.0
Unauthorized or unsafe use of time as a public …   81.548387  103104.0
No person in charge of food facility                 81.700000  103135.0
Mobile food facility not operating with an appr…   82.130435  103170.0

                                                    count
description
No restroom facility within 200 feet of mobile …      1
Noncompliance with Gulf Coast oyster regulation         4
Improper cooking time or temperatures                   5
Contaminated or adulterated food                      120
Mobile food facility with unapproved operating …     10
Unapproved food source                                 15
Sewage or wastewater contamination                     29
Other high risk violation                              27
High risk vermin infestation                          733
Unclean hands or improper use of gloves               612
Unapproved  living quarters in food facility           29
Improperly displayed mobile food permit or signage      4
Improper thawing methods                              706
Unclean or unsanitary food contact surfaces          1158
Improper cooling methods                              842
Mobile food facility stored in unapproved location      5
High risk food holding temperature                   1474
Unauthorized or unsafe use of time as a public …     31
No person in charge of food facility                   40
Mobile food facility not operating with an appr…     23
```

1.1.2 Grading

Since the assignment 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)!

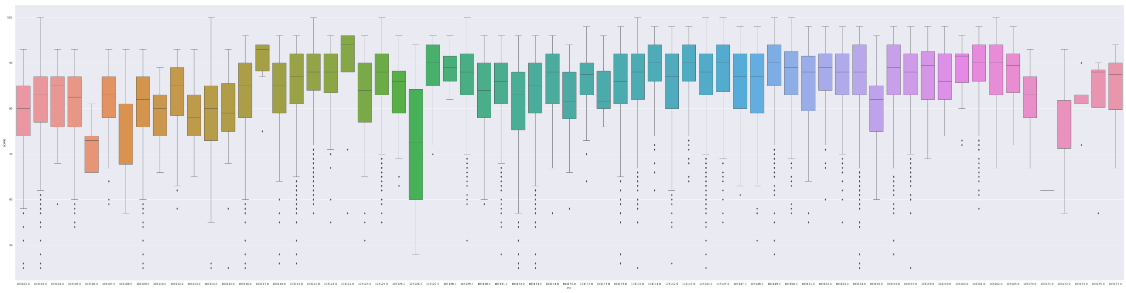
You should have the following in your answers: * a few visualizations; Please limit your visualizations to 5 plots. * a few sentences (not too long please!)

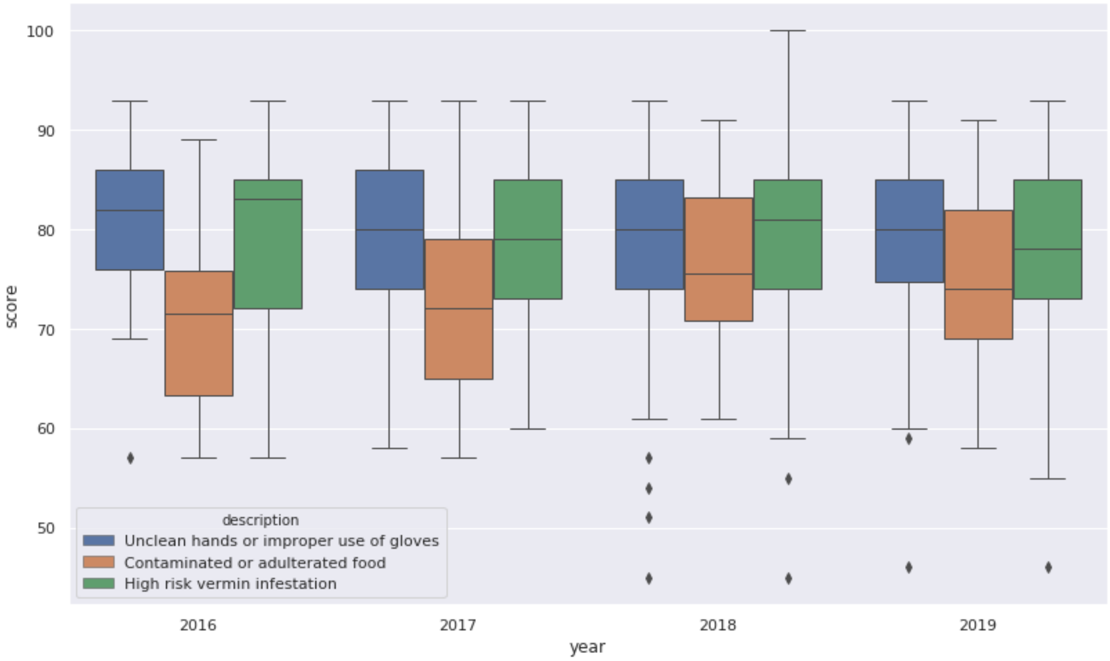
Please note that you will only receive support in OH and Piazza for Matplotlib and seaborn questions. However, you may use some other Python libraries to help you create you visualizations. If you do so, make sure it is compatible with the PDF export (e.g., Plotly does not create PDFs properly, which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

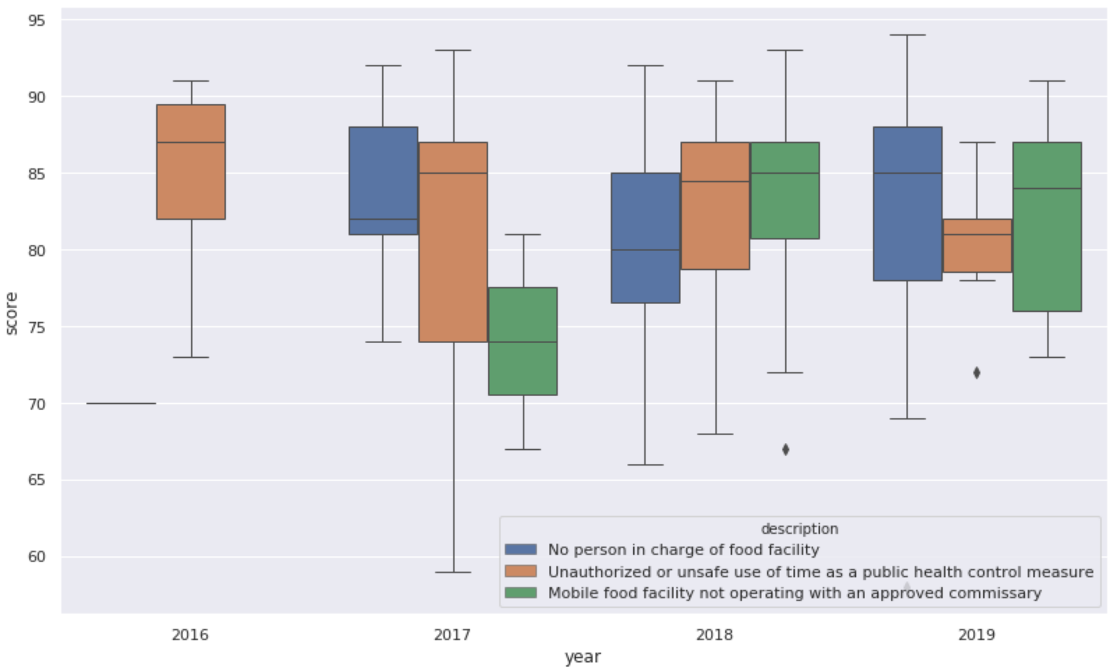
```
In [89]: # YOUR DATA PROCESSING AND PLOTTING HERE
         insvio = ins[ins["score"] > 0].merge(ins2vio, how = "left", left_on = "iid", right_on = "iid")
         insvio = insvio.merge(vio, how = "left", left_on = "vid", right_on = "vid")
         plt.figure(figsize = (70, 20))
         sns.boxplot(x="vid", y="score", data=insvio, linewidth = 1.0)

         insvio1 = insvio[(insvio["description"] == "Contaminated or adulterated food") | (insvio["desc
         plt.figure(figsize = (12, 8))
         sns.boxplot(x="year", y="score", hue="description", data=insvio1, linewidth = 1.0)

         insvio2 = insvio[(insvio["description"] == "Mobile food facility not operating with an approve
         plt.figure(figsize = (12, 8))
         sns.boxplot(x="year", y="score", hue="description", data=insvio2, linewidth = 1.0)



         # YOUR EXPLANATION HERE (in a comment)
         """I constructed three boxplot graphs: the first one is the distribution of scores for all vio
         represents a violation), the second one is the distribution of scores for three violations tha
         a low average score and high count throughout the years, the third is the same as the second e
         the top three highest average scores. I constructed the first graph to see if there were any w
         any of the plots were significantly lower/higher than the rest and if this makes sense with th
         distribution for each violation makes sense with the mean scores I found above. The one anomol
         #103126, which is the Noncompliance with Gulf Coast oyster regulation violation, where the box
```

seems like a very specific violation and it also had a count value of 4 so I would have to do the IQR is so big. With the second and third graph, I wanted to see how the distribution of th throughout the years. For the second graph, 'High risk vermin infestation' and 'Unclean hands seems to have the same distribution of scores throughout the years whereas the distribution sc adulterated food' improved over the years. For the third graph, the widths of the box plots fl explained by more collected data for any given violation that would've decreased the IQR. In t of scores for 'Unauthorized or unsafe use of time as a public health control measure' seemed t the years, while the opposite is true for'Mobile food facility not operating with an approved

```
Out[89]: "I constructed three boxplot graphs: the first one is the distribution of scores for all viola
```






95
90
85
80
75
70
65
60
score
2016
2017
2018
2019
year
description
No person in charge of food facility
Unauthorized or unsafe use of time as a public health control measure
Mobile food facility not operating with an approved commissary