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
```

Regarding the bus file, the latitude, longitude and phone_number entires seem to be missing some values. Missing values and bad data are an issue because they could potentially affect the results from the data, further more it is unknown whether missing values will cause problems until the data is actually utilized.

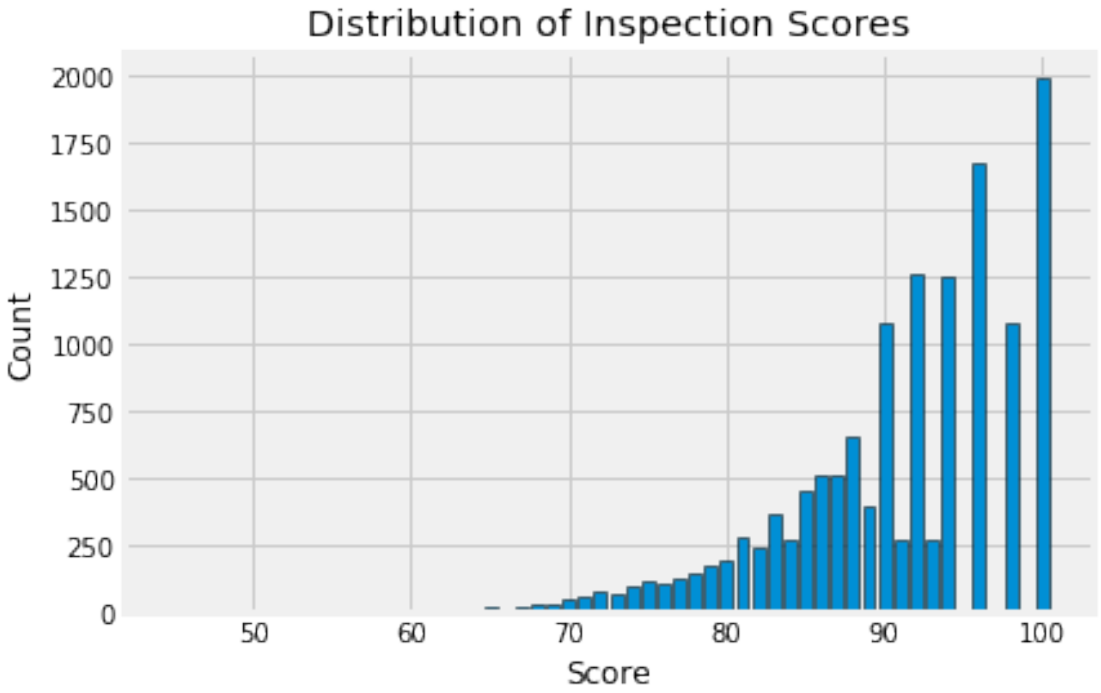
In the cell below, write the name of the restaurant with the lowest inspection scores ever. You can also head to yelp.com and look up the reviews page for this restaurant. Feel free to add anything interesting you want to share.

Lollipot

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.

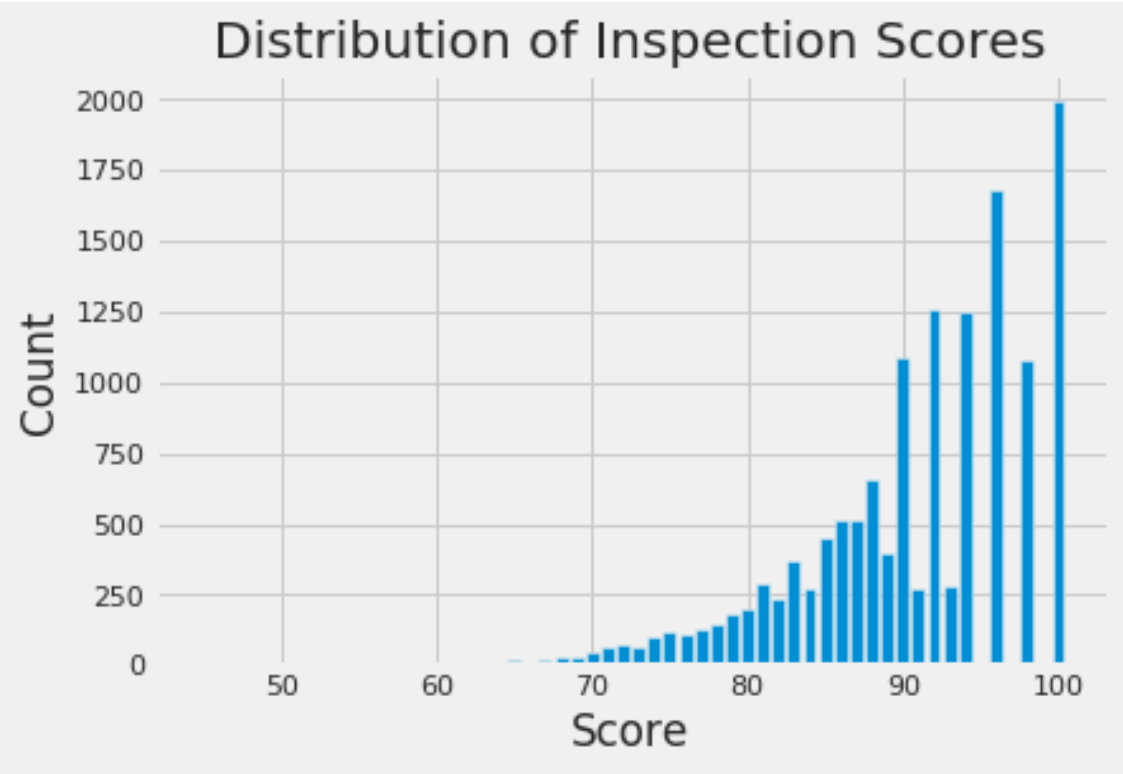


You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need:

```
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 [75]: score_counts = ins_named['score'].value_counts()
         plt.bar(score_counts.keys(), score_counts)
         plt.xlabel("Score")
         plt.ylabel("Count")
         plt.title("Distribution of Inspection Scores");
```

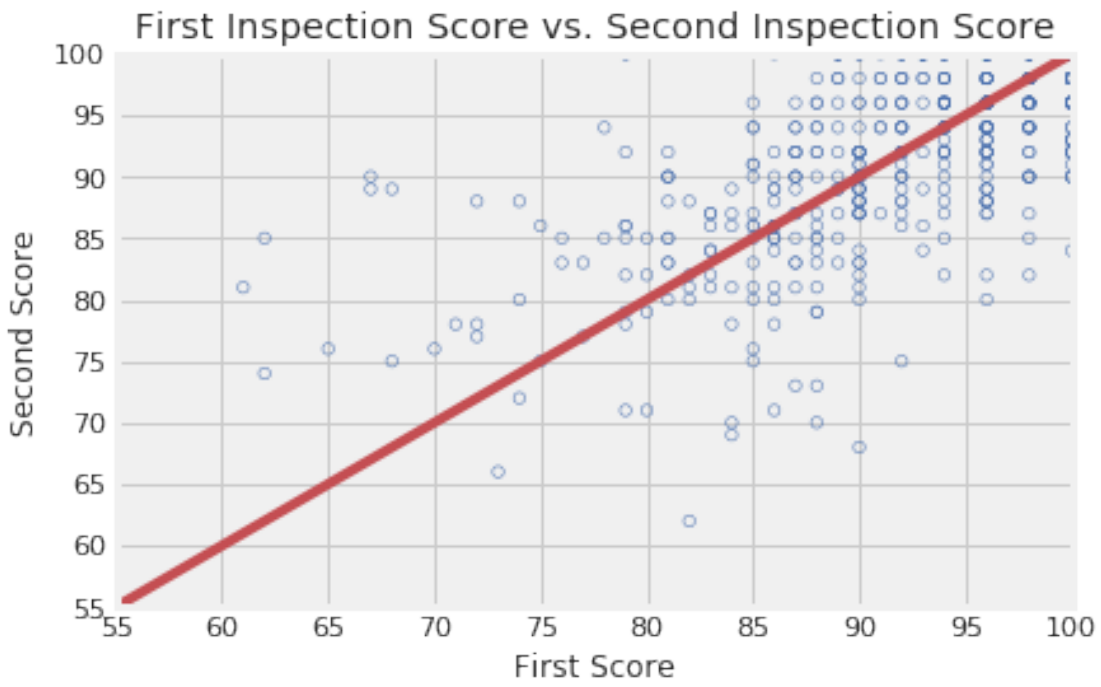


0.1.1 Question 6b

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?

The mode is 100, there does not seem to be any symmetry and the distribution seems to be skewed to the left. There does not to seem to be many gaps until the 95 - 100 range. My observations imply that the scores are on the higher side, but the average is lower than the median.

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:

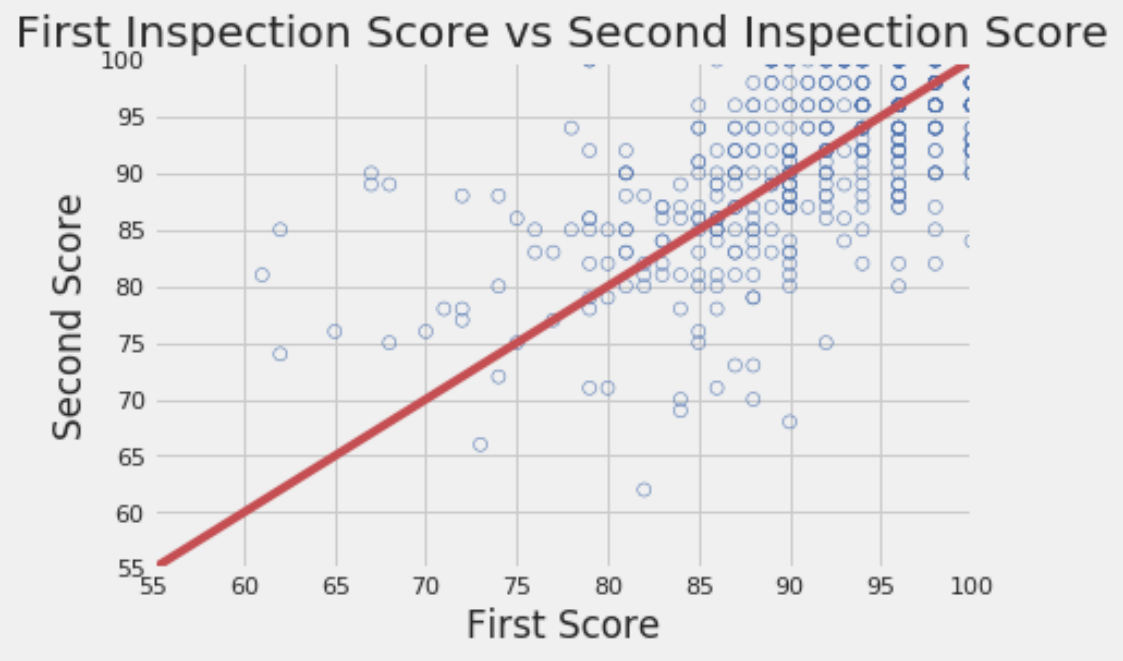
`plt.scatter` plots a set of points. Use `facecolors='none'` and `edgecolors=b` to make circle markers with blue borders.

`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.

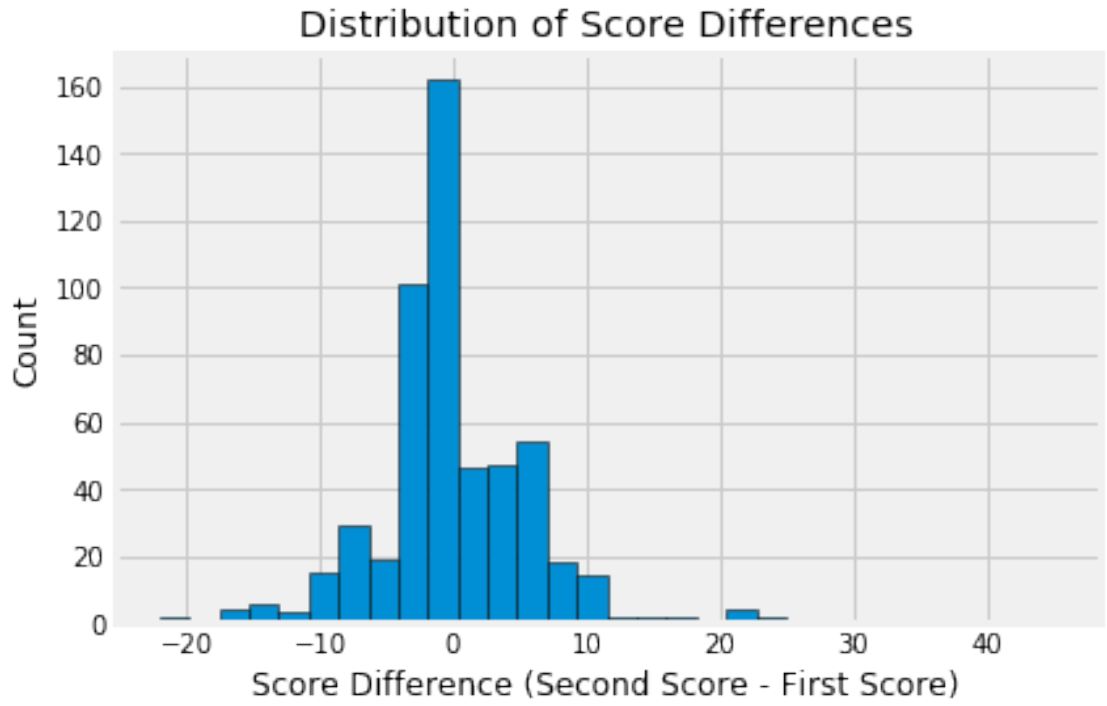
```
In [84]: first_score, second_score = zip(*scores_pairs_by_business['score_pair'])

         plt.scatter(first_score, second_score, facecolors= 'none', edgecolors= 'b')
         plt.plot([55,100], [55,100], 'r')
         plt.xlabel('First Score')
         plt.ylabel('Second Score')
         plt.axis([55,100,55,100])
         plt.title('First Inspection Score vs Second Inspection Score');
```


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:

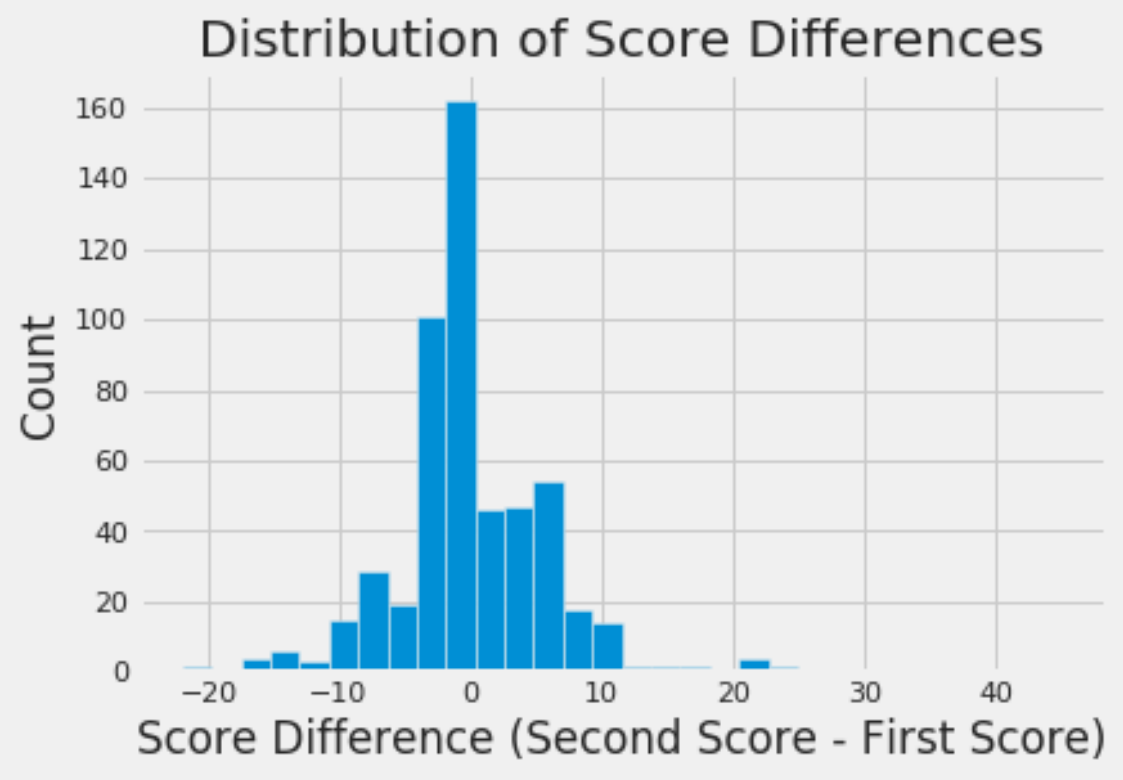


Hint: Use `second_score` and `first_score` created in the scatter plot code above.

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 [85]: diff_in_scores = np.array(second_score) - np.array(first_score)
         plt.hist(diff_in_scores, bins = 30)
         plt.xlabel('Score Difference (Second Score - First Score)')
         plt.ylabel('Count')
         plt.title('Distribution of Score Differences');
```


Distribution of Score Differences
Count
160
140
120
100
80
60
40
20
0
−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?

If restaurants' scores tend to improve from the first to the second inspection, I would expect to see the points above the line, because the line has a slope of 1. The points seem to represent that for most restrauants there seem to be little change between the first and second inspections. Yes, the observation is consistent with my expectations because I expected the restaurants to be more or less consistent in their scores throughout the two inspections.

0.1.4 Question 7f

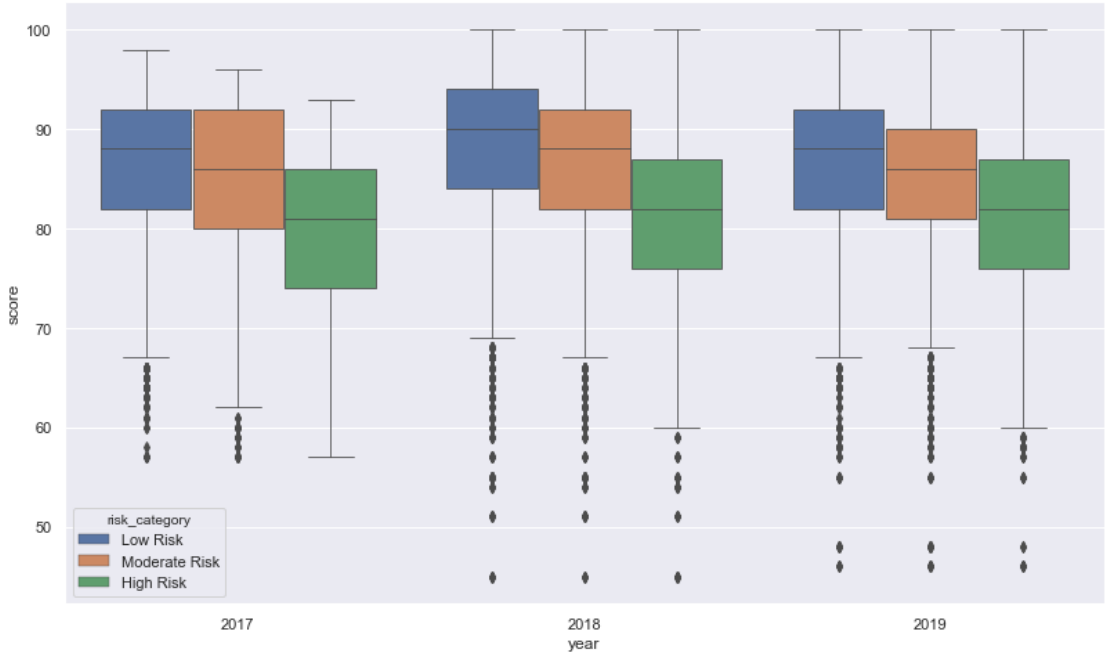
If a restaurant's score improves from the first to the second inspection, how would this be reflected in the histogram of the difference in the scores that you made in question 7d? What do you oberve from the plot? Are your observations consistent with your expectations? Explain your observations in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

If a restaurant's score improves from the first to the second inspection we would expect its data to be towards the positive values. Overall, it seems that there were little to no changes for most restaurants as shown by 0 being the mode. The center is also 0 and the spread is mostly between -10 and 10. I expected there to be little to no change in the restaurants between the first and second. Therefore, my observatoins were consistent with my expectations.

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.

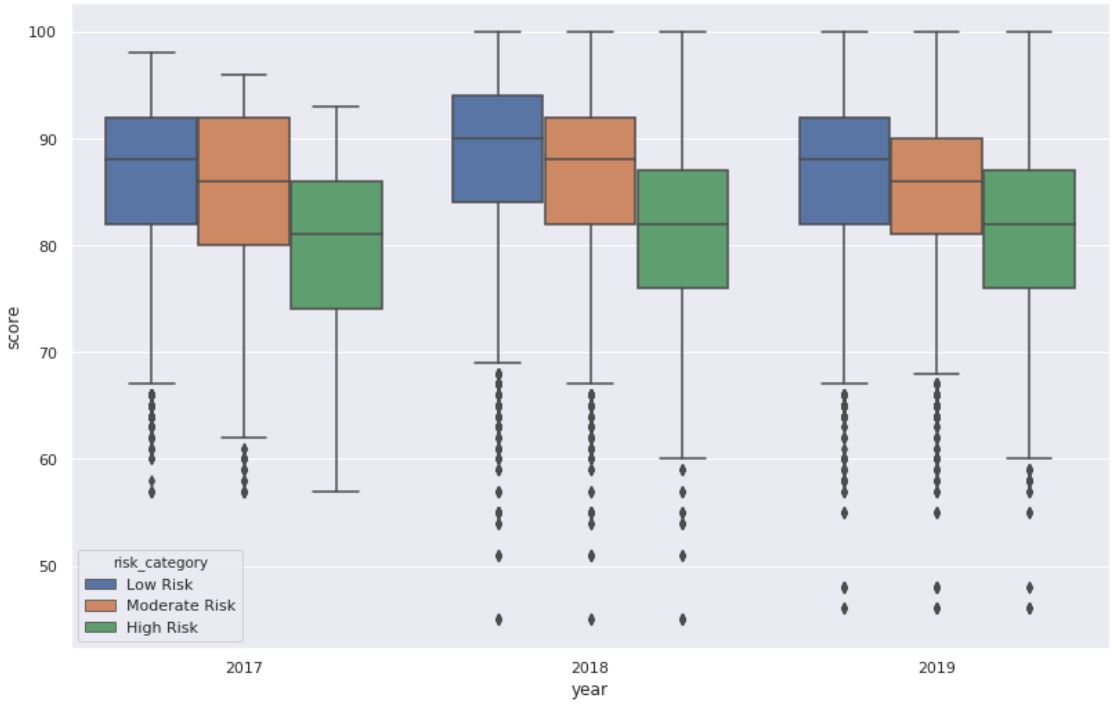
The boxplot should look similar to the sample below. Make sure the boxes are in the correct order!



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 [86]: # Do not modify this line
         vio_w_iid = vio.merge(ins2vio, how = 'right', left_on = 'vid', right_on = 'vid')
         vio_w_score = vio_w_iid.merge(ins_named, how = 'right', left_on = 'iid', right_on = 'iid')
         vio_w_score2 = vio_w_score[vio_w_score['year'] >= 2017]
         sns.set()
         plt.figure(figsize = (12,8))
         sns.boxplot(x = 'year', y = 'score', hue = 'risk_category', hue_order = ['Low Risk', 'Moderate
```


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.

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 [139]:
```
#YOUR CODE HERE

risk_category = vio.merge(ins2vio, how = 'right', left_on = 'vid', right_on = 'vid')
risk_w_score = risk_category.merge(ins_named, how = 'right', left_on = 'iid', right_on = 'iid
five_ins = risk_w_score.groupby('bid').filter(lambda x : len(x) == 5)
five_ins = five_ins.loc[:,['description', 'risk_category', 'score']].dropna()
ten_ins = risk_w_score.groupby('bid').filter(lambda x : len(x) == 10)
ten_ins = ten_ins.loc[:,['description', 'risk_category', 'score']].dropna()
twenty_ins = risk_w_score.groupby('bid').filter(lambda x : len(x) == 20)
twenty_ins = twenty_ins.loc[:,['description', 'risk_category', 'score']].dropna()
```

```
display(five_ins, ten_ins, twenty_ins)

#YOUR EXPLANATION HERE (in a comment)
#Used merge to get descriptions, risk and score in one table.
#Wanted to see if businesses with higher scores,
#had higher scores just because they were inspected less or
#because it was due to the fact that they continued to make the same violations
#and were inspected more with no or little improvement leading to lower scores.
```

	description	risk_category	score
14	Inadequately cleaned or sanitized food contact...	Moderate Risk	76
15	Unapproved or unmaintained equipment or utensils	Low Risk	76
16	Improper thawing methods	Moderate Risk	76
17	Improper cooling methods	High Risk	76
18	Unclean hands or improper use of gloves	High Risk	76
...	...	...	...
39449	Improper food storage	Low Risk	78
39450	Inadequate food safety knowledge or lack of ce...	Moderate Risk	78
39451	Unclean or unsanitary food contact surfaces	High Risk	78
39452	Inadequate dressing rooms or improper storage ...	Low Risk	78
39453	Other high risk violation	High Risk	78

[2344 rows x 3 columns]

	description	risk_category	score
1124	High risk vermin infestation	High Risk	79
1125	Improper or defective plumbing	Low Risk	79
1126	Inadequate and inaccessible handwashing facili...	Moderate Risk	79
1127	Improper food storage	Low Risk	79
1128	Wiping cloths not clean or properly stored or ...	Low Risk	79
...	...	...	...
39364	Sewage or wastewater contamination	High Risk	80
39365	Inadequately cleaned or sanitized food contact...	Moderate Risk	88
39366	Moderate risk food holding temperature	Moderate Risk	88
39367	Unapproved or unmaintained equipment or utensils	Low Risk	88
39368	Unclean or degraded floors walls or ceilings	Low Risk	88

[2333 rows x 3 columns]

	description	risk_category	score
1257	Unapproved or unmaintained equipment or utensils	Low Risk	80
1258	Improper thawing methods	Moderate Risk	80
1259	Improper reheating of food	High Risk	80
1260	Unclean hands or improper use of gloves	High Risk	80
1262	Moderate risk food holding temperature	Moderate Risk	81
...	...	...	...
32380	Moderate risk food holding temperature	Moderate Risk	83
32381	Low risk vermin infestation	Low Risk	83

```
32382  Unapproved or unmaintained equipment or utensils       Low Risk     83
32383                    Unclean nonfood contact surfaces       Low Risk     83
32384         Unclean or unsanitary food contact surfaces      High Risk     83

[898 rows x 3 columns]
```

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.

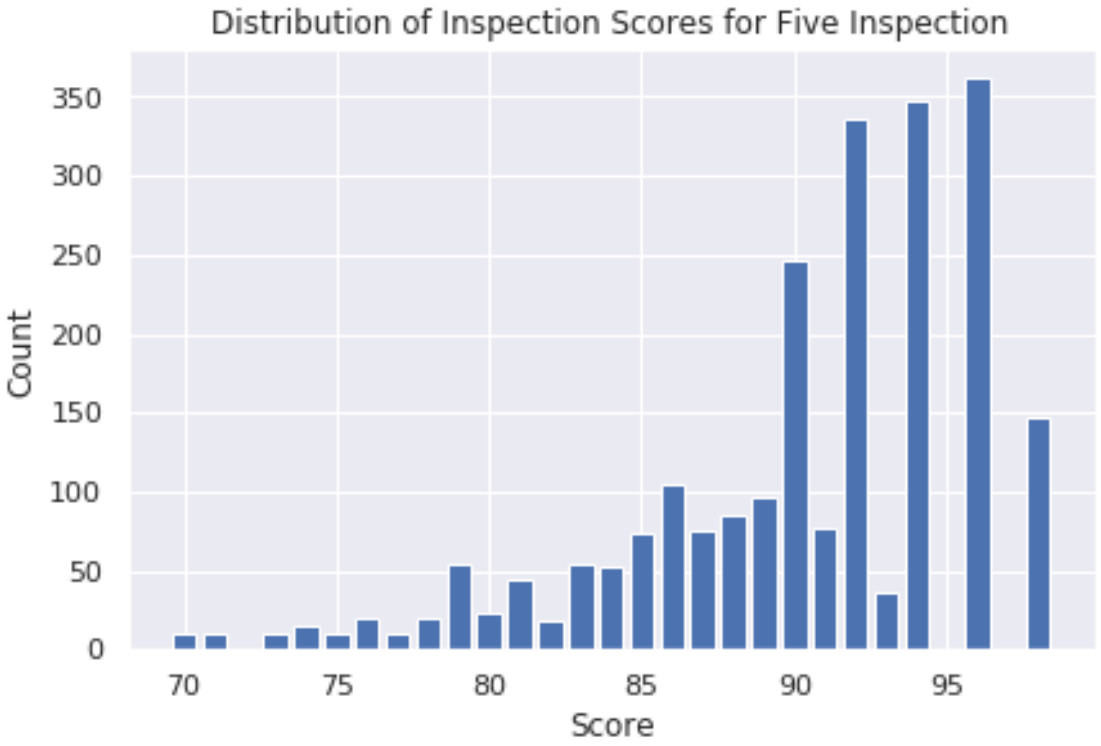
We will lean towards being generous with the grading. We might also either discuss in discussion or post on Piazza some examplar analysis you have done (with your permission)!

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 [134]:
```
# YOUR DATA PROCESSING AND PLOTTING HERE
score_counts_five = five_ins['score'].value_counts()
plt.bar(score_counts_five.keys(), score_counts_five)
plt.xlabel("Score")
plt.ylabel("Count")
plt.title("Distribution of Inspection Scores for Five Inspection");
plt.show()

score_counts_ten = ten_ins['score'].value_counts()
plt.bar(score_counts_ten.keys(), score_counts_ten)
plt.xlabel("Score")
plt.ylabel("Count")
plt.title("Distribution of Inspection Scores for Ten Inspections");
plt.show()

score_counts_twenty = twenty_ins['score'].value_counts()
plt.bar(score_counts_twenty.keys(), score_counts_twenty)
plt.xlabel("Score")
plt.ylabel("Count")
plt.title("Distribution of Inspection Scores for Twenty Inspections");
plt.show()

# YOUR EXPLANATION HERE (in a comment)
# Used three bar graphs, one for each number of inspections in order to see the distributions
```

```
# Used these graphs to supplement the tables in 8A and to see if businesses with lower scores
# Although the number of businesses in each group are different and thus might affect
# the distributions of scores, it is clear that business with less inspections on average had
# With 5 inspections having a mode of 95 and having a left skew towards higher scores.
# Businesses with 10 inspections had a mode of 90, still with a left skew but their was more
# around the 80 - 90 range compared to the 5 inspections
# Businesses with 20 inspections, seems to be bimodal at 75 and 87. Most of their scores were
# on average lower than the other businesses.
# Referring back to question 7 on how the scores of business overall seemed to be consistent
# throughout their inspections, it might be safer to suggest that businesses with more inspec
# had lower scores not because they were inspected more but because they repeated the same of
```




Distribution of Inspection Scores for Ten Inspections
Count
Score
0
50
100
150
200
250
300
60
70
80
90
100


Distribution of Inspection Scores for Twenty Inspections
Count
Score
0
10
20
30
40
50
60
70
60
70
80
90
100