0.1 Question 2e

If we were to drop businesses with MISSING postal code values, what specific types of businesses would we be excluding? In other words, is there a commonality among businesses with missing postal codes?

Hint: You may want to look at the names of the businesses with missing postal codes. Feel free to reuse parts of your code from 2d, but we will not be grading your code.

Most of the MISSING postal code values are food trucks or restaurants that don't have a permanent location.

0.2 Question 5

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 should use the `ins` dataframe, and should ignore any score that is less than 0.

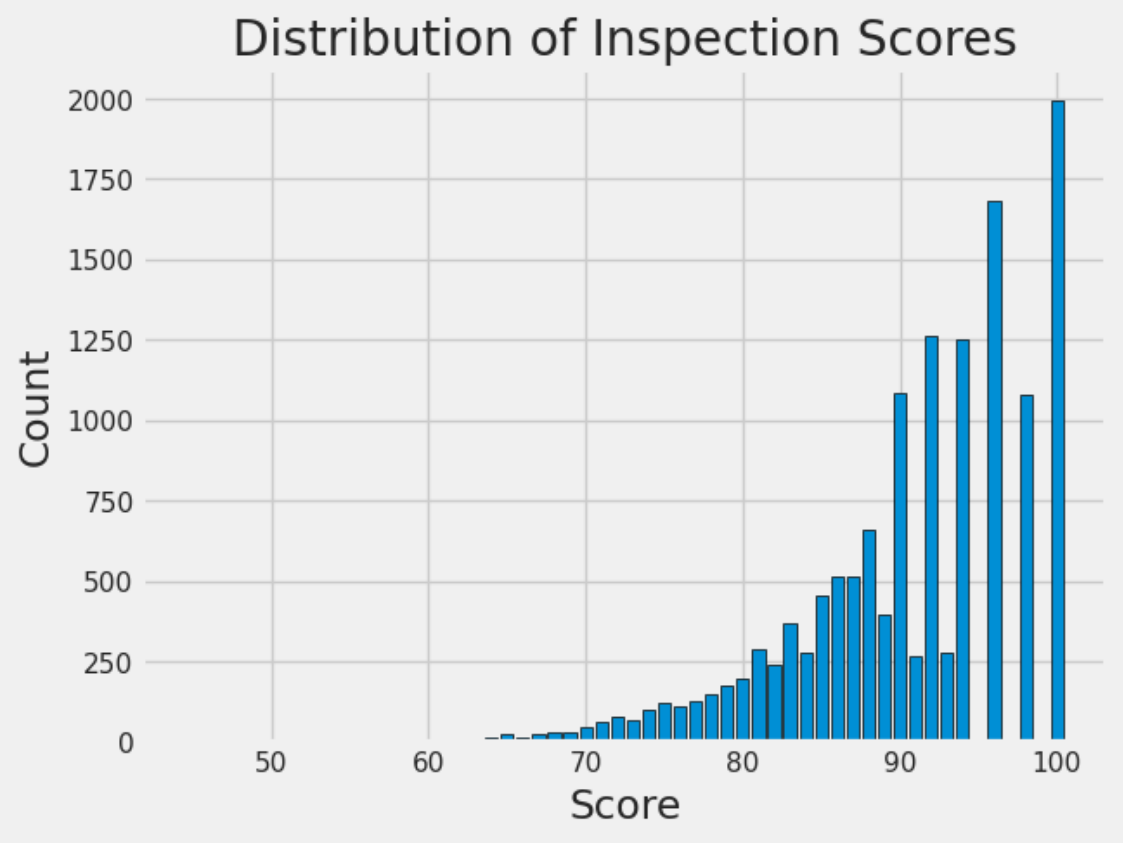
You might find this matplotlib.pyplot tutorial useful. Key syntax that you'll need:

```
plt.bar
plt.xlabel
plt.ylabel
plt.title
```

To set the color of the edges for your bars, include 'edgecolor = 'black'.

```
In [131]: score_count = ins.groupby('score').size()
          plt.bar(score_count.index, score_count, edgecolor='black')
          plt.xlabel("Score")
          plt.ylabel("Count")
          plt.title("Distribution of Inspection Scores")
```

```
Out[131]: Text(0.5, 1.0, 'Distribution of Inspection Scores')
```


Distribution of Inspection Scores
Count
2000
1750
1500
1250
1000
750
500
250
0
50
60
70
80
90
100
Score

0.3 Question 6b

Now let's make a scatter plot to display these pairs of scores. Include on the plot a reference line with slope 1 and y-intercept 0. Since restaurant scores bottom out at 45 points, we'll only focus on ratings between 45 and 100. Thus your reference line should start at `[45, 45]` and go up to `[100, 100]`.

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:

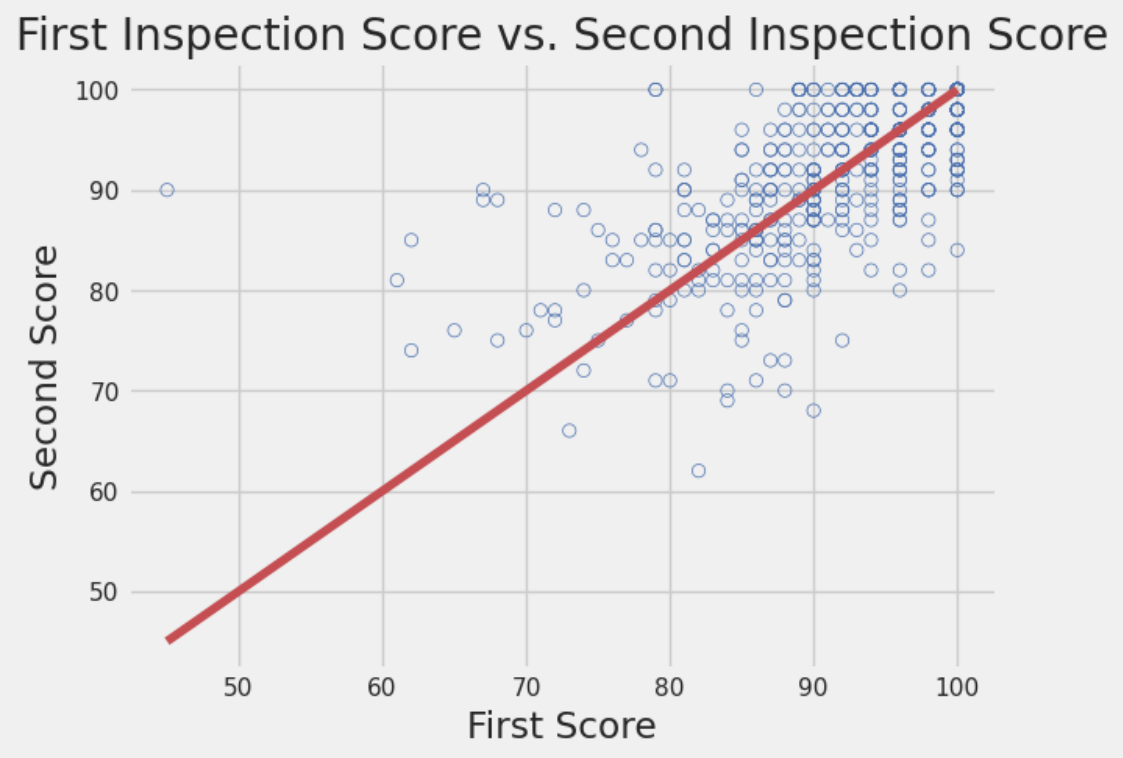
`plt.scatter` plots a set of points. Use `facecolors='none'` and `edgecolors='b'` to make circle markers with blue borders.

`plt.plot` for the reference line. Using the argument `r` will make the line red.

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

```
In [135]: plt.scatter(scores['first score'], scores['second score'], facecolors='none', edgecolors='b')
          plt.plot([45, 100], [45, 100], color='r')
          plt.xlabel('First Score')
          plt.ylabel('Second Score')
          plt.title('First Inspection Score vs. Second Inspection Score')
```

```
Out[135]: Text(0.5, 1.0, 'First Inspection Score vs. Second Inspection Score')
```


First Inspection Score vs. Second Inspection Score
Second Score
First Score
100
90
80
70
60
50
50
60
70
80
90
100

0.4 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 6b? What do you oberve from the plot? Are your observations consistent with your expectations?

Hint: What does the slope represent?

The scatter plot represents the first and second scores of each restaurant, and the reference line represents when the scores are equal. If the points fall below the line, the corresponding restaurants got a lower score, so they got worse. If the points fall on the line, the restaurants didn't change. If the points fall above, then the restaurants improved. If restaurants tend to improve form the first to second inspection, the scatterplot should show most of the points above the reference line. Since there is roughly the same amount of points above as below, the observations are not consistent with my expectations.