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?

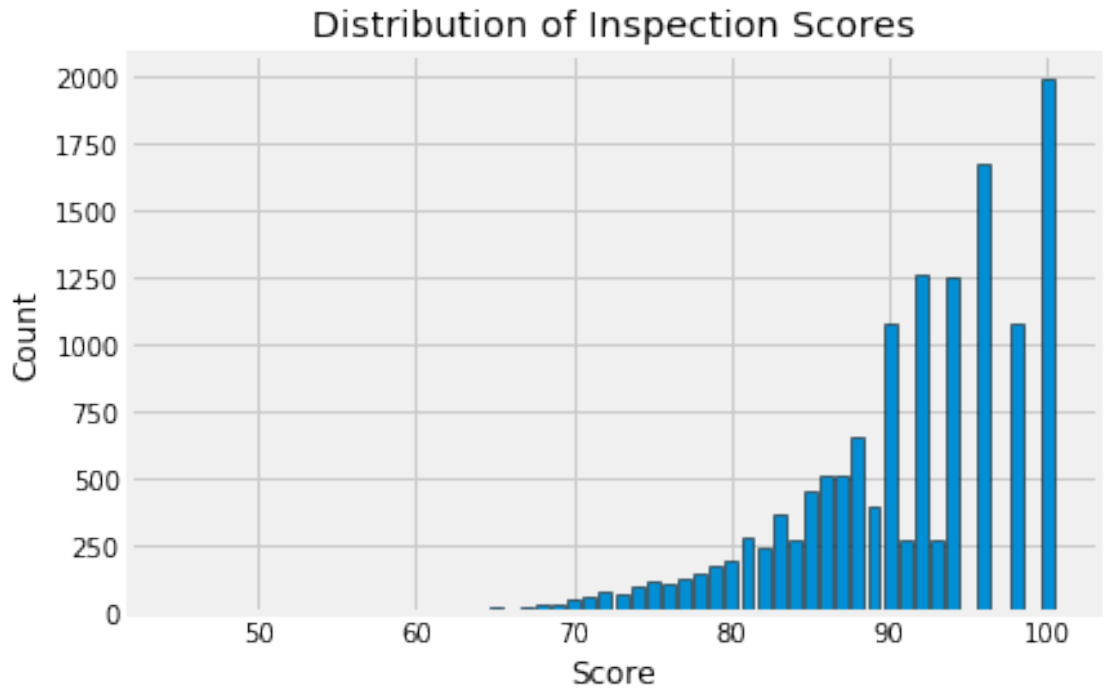
Hint: You may want to look at the names of the businesses with missing postal codes. Feel free to reuse parts of your code from 2d, but we will not be grading your code.

If we just look up some of these businesses we are indeed omitting some actual businesses that make a living by selling products/services on the move. For example we omit food trucks, usually selling food from location to location.

0.2 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
```

```
In [72]: ins['score'].index
```

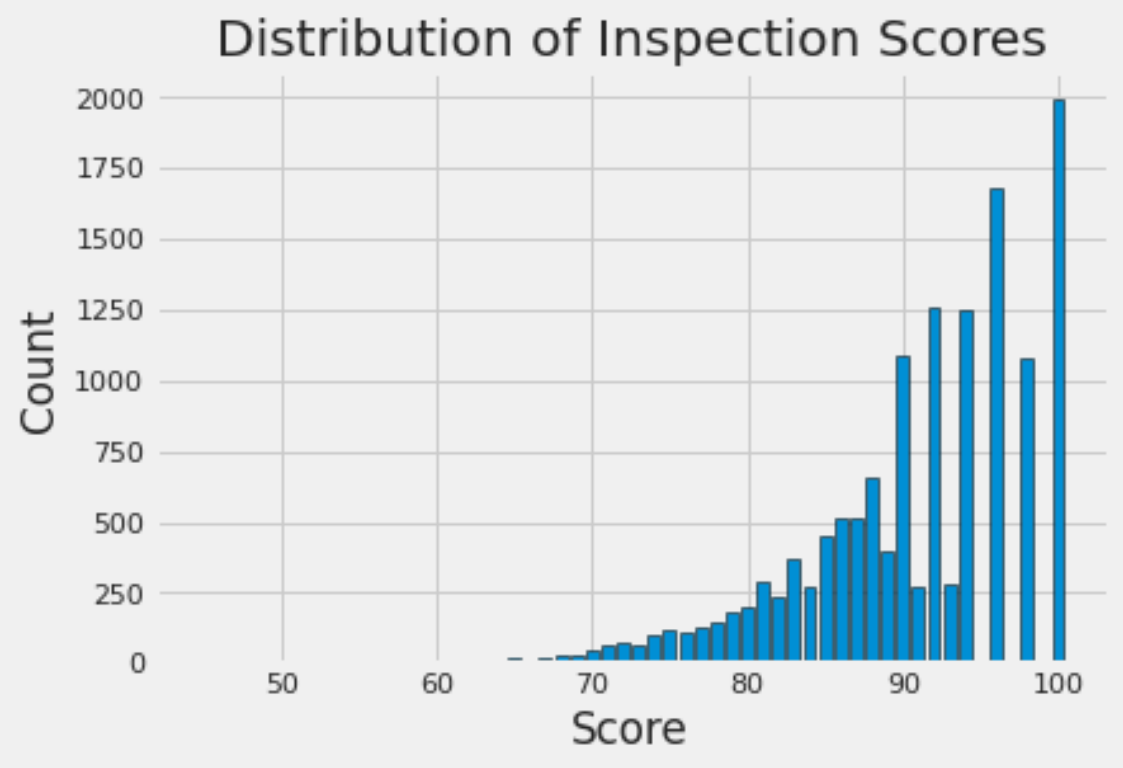
```
Out[72]: Int64Index([    1,     3,    15,    20,    21,    23,    30,    33,    34,
```

```
            38,
 …
 26628, 26635, 26644, 26646, 26648, 26654, 26656, 26660, 26661,
 26662],
dtype='int64', length=14031)
```

```
In [73]: ins['score'].value_counts().index
```

```
Out[73]: Int64Index([100,  96,  92,  94,  90,  98,  88,  86,  87,  85,  89,  83,  81,
                      93,  84,  91,  82,  80,  79,  78,  77,  75,  76,  74,  72,  73,
                      71,  70,  68,  69,  65,  67,  66,  64,  63,  62,  60,  61,  59,
                      58,  57,  55,  54,  46,  45,  51,  48],
                    dtype='int64')
```

```
In [74]: scores = ins['score'].value_counts()
         score_index = ins['score'].value_counts().index
         plt.bar(score_index,scores,edgecolor='black');
         plt.title("Distribution of Inspection Scores");
         plt.xlabel("Score");
         plt.ylabel("Count");
```

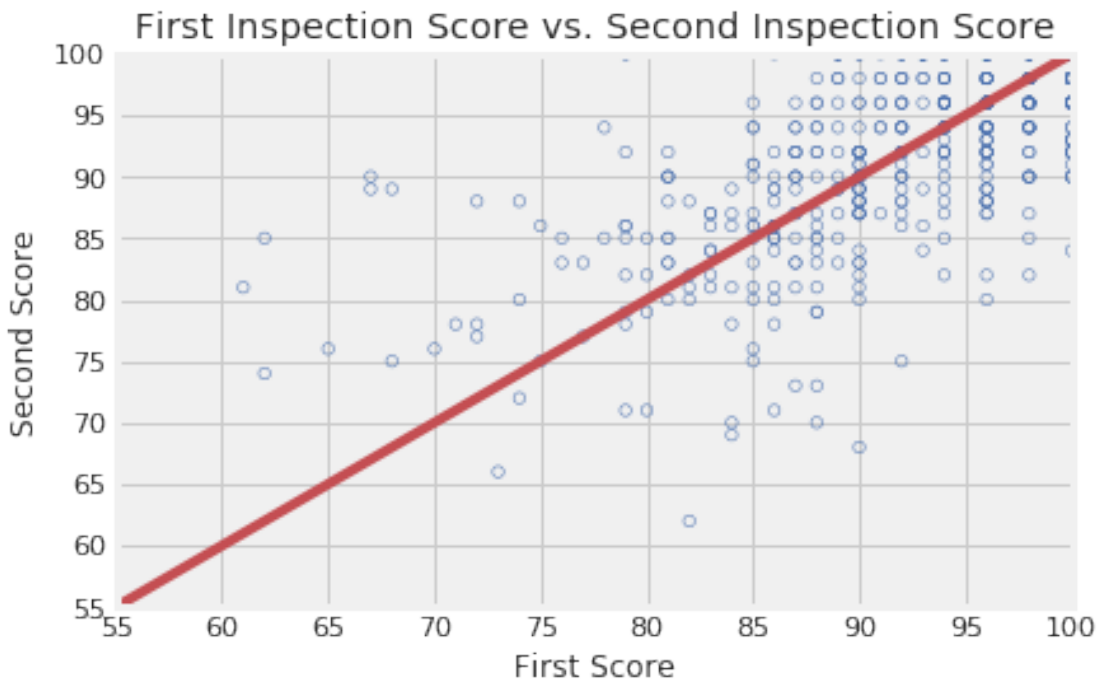


0.2.1 Question 5b

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 graph is not a normal curve and it is skewed to the left. We can see most of the scores are bunched towards the higher scores. This has a long left tail. There are gaps from 90-100 where we see that values 99,97 and 95 are missing. This is probably because health inspection scores are primary judged by even grading of points at least for higher scores, hence why we see no values for odd labeled scores. We know that 90-100 is incremented by 1, so each bar represents scores 90-100 by 1. We also see that the highest counts of scores are awarded to 100, where its count outnumbers most of the scores below 90, suggesting that resturants follow sanitiation guidelines.

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.



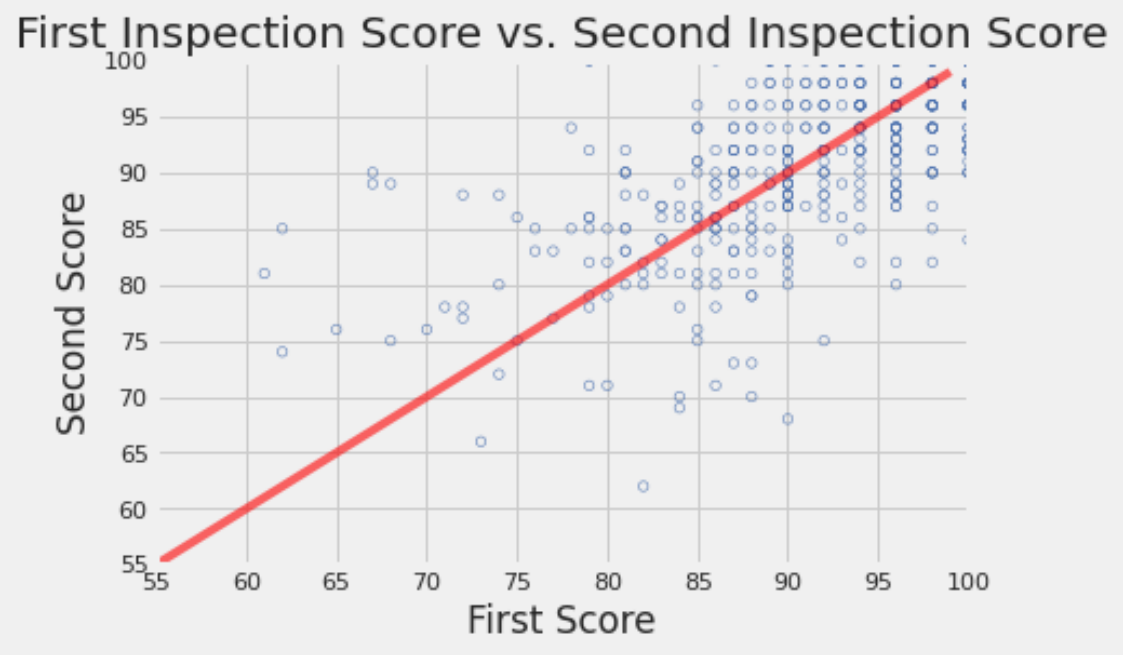
Key pieces of syntax you'll need:

`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`.

```
In [81]: x_linear = np.arange(0,100,1)
         y_linear = np.arange(0,100,1)
         plt.plot(x_linear,y_linear,color='red',linewidth=4,alpha=0.6);
         plt.scatter(scores_pairs_by_business["new_col"],scores_pairs_by_business["new_coll"],s=20,face
         plt.xlabel('First Score');
         plt.ylabel('Second Score');
         plt.title('First Inspection Score vs. Second Inspection Score');
         plt.xlim(55, 100);
         plt.ylim(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

0.2.2 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 5c? What do you oberve from the plot? Are your observations consistent with your expectations?

Hint: What does the slope represent?

When we look at the graph, we see that during the second inspection not many scores are above the regression line. We expected that the points lie above the regression line because we anticipate that second scores are higher than first scores. We also expected that the values difference from first to second inspection increase because we want scores to be higher second time around, yet we see in the graph that this is not the case, some values are worse second inspection. This signals that resturants tend to not see much change in their first and second score.