Pandas for Data Wrangling Quiz

1. Suppose you load in a Pandas dataframe with read_csv("df.csv") but you notice the first column is called "Unnamed: 0" with values going up in sequential order (0, 1, 2, 3, …). What could be a potential cause of this additional column?

2. Identify and explain two different methods of outlier removal for numerical data that can be used in Pandas or similar libraries.

3. Suppose you have Series object called 'counts' which is made of 100 integers, and you want to convert it into a dataframe of 10 rows and 10 columns. Give a line of code to do this using numpy and an equal method of doing so using just Pandas.

4. date = '12/20/2020'

We want to convert date (which is a string) into a datetime object. Give the function call as well as the format that will successfully convert this value. Do not use the function 'to_datetime' for this problem.

5. Labels are used when indexing with loc and positions are used for iloc. However, there are more subtle differences amongst the two, especially when it comes to Boolean indexing. Explain a key difference between loc and iloc in this respect and provide an example to support your claim.

6. Explain the differences between applying, aggregating, and transforming a groupby object.

7. Suppose you're given a dataset of rows consisting of houses in Berkeley and 150 columns for the price of each house for the past 150 years. Explain why you might want to transform this data and how.

8. How are categorical features one-hot encoded in Pandas? Does it always make sense to one-hot encode all your categorical features?

9. Name the four different types of time related concepts in Pandas and the differences between them.

10. The advantage of using Pandas over other libraries is its simplicity. However, without taking extra steps to optimize your code, its performance may suffer. This is especially true when working with large data and complex machine learning algorithms. Provide two ways one can optimize their code to improve computation speed and reduce overhead.