Author: Katelyn Lindsey

Date: 5/27/2021

Viewpoint Data Engineer Answers

(All linked code files can also be found at <u>my github repository</u>)

1.a.) See *https://github.com/katecoded/viewpoint_questions/blob/main/question_1a.sql*.

b.) See the upperLower() function at:
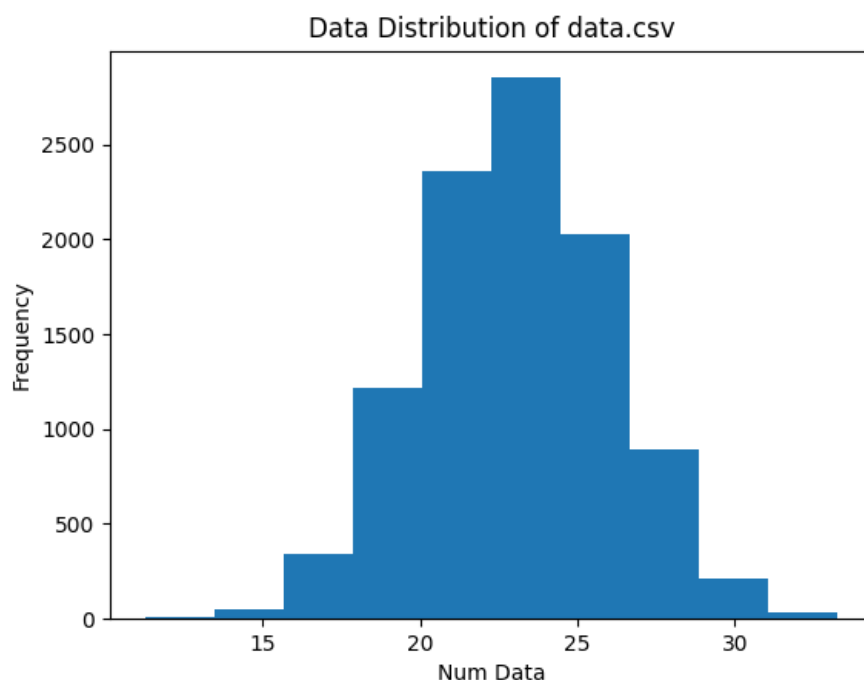*https://github.com/katecoded/viewpoint_questions/blob/main/question_1bc.py*.

c.) See the markAverages() function at:
https://github.com/katecoded/viewpoint_questions/blob/main/question_1bc.py.

2.) For the white house data set, I decided to first fix the data from the csv file so that all "POSITION TITLE" entries were correct after noticing a few of the entries were mistakenly in the "PAY BASIS" column. Then, I realized that the "PAY BASIS" column was not very useful in terms of data - there was no variety at all - therefore, I went ahead and removed that column. I also removed a small number of outliers from the data based on salary - there were a few white house staffers with a salary of $0.00 due to not accepting a salary, which are not a good representation of the salaries of a white house staffer. All of this can be seen in the dataTasks() function at:
https://github.com/katecoded/viewpoint_questions/blob/main/question_2.py.

For the data.csv file, the data ranges from about 11.25 to 33.29, with an average of 23.04 and a median of 23.02. The standard deviation (or variance) of the data is about 3.00. Here is a graph of the distribution:

Data Distribution of data.csv

As seen in the graph, the data is roughly symmetric in its distribution with a very slight skew to the left.

The calculations for this can be seen in the describeData() function at:

https://github.com/katecoded/viewpoint_questions/blob/main/question_2.py.

3.) I would use either an SQL query, Python (or another language such as JavaScript), or formulas in Excel to update all of the necessary columns at once. With that large of a file, it would be ridiculous to enter that by hand, so I would absolutely not manually enter null values with a text editor or use Excel without formulas.

4.) Realistically, I would likely refuse if given the option to unless it was only once or twice - that would be a bad precedent to set for maintaining a quality work-life balance unless I was working nights. However, if I was working nights and used one of the above tools/strategies to update them all at once, it would only take a minute, which would be completely fine.

5.) I definitely admire the work of Sir Isaac Newton simply because of my love of physics and space - I actually nearly pursued physics in college, and he was (probably) the first one that invented calculus (which is something I have mostly forgotten by now but used to greatly enjoy). However, I wouldn't necessarily say that he is my favorite mathematician

or that I even have a favorite. Most of my modern day inspiration for computer science pursuits comes from programmers of open-source projects and passion projects, especially since I use open-source software on a day-to-day basis (and I unfortunately do not know the names of the people who created most of these projects).