

**Capsule8 Data Science Challenge**

Greetings, fellow data scientist. Thank you for accepting to take the Capsule8 data science challenge !

This challenge includes a set of exercises for you to work through, at your own pace. Your answers will help us determine your fit for our growing data science team, so take as much time as you need and feel free to use whatever resources you like. When you have completed the challenge, you then have a couple of options. You can either send us the results or, if you like, you can give a short 30-45 minute presentation at our office, or via video conference.

As you take the challenge, you might have questions. By all means, feel free to ask us questions. Send your questions to [george@capsule8.com](mailto:george@capsule8.com) and we’ll get back to you as soon as we can.

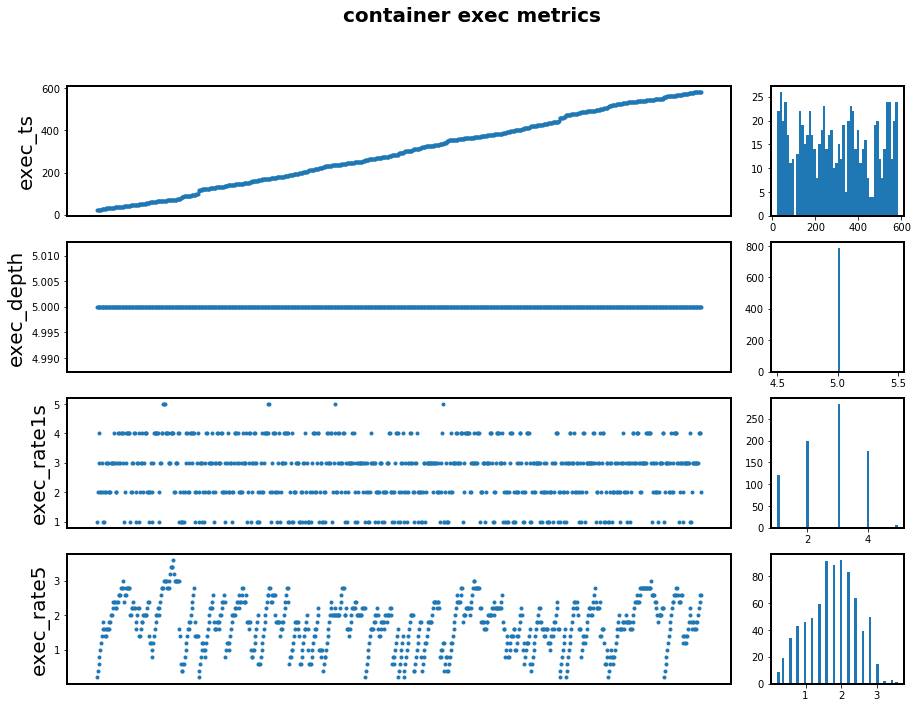
Good luck with the challenge, and thank you for your time !

**Data Wrangling and Exploration**

As you well know, the job of a data scientist involves taking data from various types of data sources and exploring it using various tools like IPython or R. We have included a sample dataset in this challenge for you to explore. What’s really cool about this dataset is that it includes a database of raw data gathered via our Capsule8 probes and sensors !

The dataset file is called “sqlite.db” and is included in this challenge. Go ahead and load the database using your favorite tools and create the 4 graphs described below. I’ve gone ahead and did the first one for you. I chose one way to show it and you should feel free to be as creative as you like. If you choose to explore different types of graphs, please explain the relative benefits of each type of graph.

If you are having trouble loading the data, please send us an email and we’ll provide some suggestions for you.

1. Locate the table called “exec” and plot the following data points as a time-series chart: “depth”, “rate1s”, and “rate5s”. The time value is in the “ts” column. I’ve done an example below. Note that the values for “rate1s” and “rate5s” are embedded in a JSON dictionary in the “rate” column of the table. I did an example visualization for you below. Please to do your own version and show your work ! 
2. Plot also the histogram for all values of each of these columns. I did an example above, next to the time series chart.
3. Locate the table called “tcplife” and plot the following data columns as time series plots: “rx”, “tx”, “dur”, “lport”, and “rport”. Plot also the histogram as you did in 2 above. Please show your work and code.

**Data Statistics**

Mark Twain has an awesome quote about lying He said that three kinds of lies: “there are lies, damn lies, and statistics.” Very funny don’t you think ? Anyway, we won’t be dealing with lies here, but we will be dealing with statistics. I hope you enjoy these statistics challenges !

1. For the “dur” values in 3, perform a study of its *descriptive statistics* and show your work and code.
2. Do you believe any of the values of “dur” are *anomalous* ? Which ones ? Please use concepts of *robust statistics* to demonstrate which values, if any, are anomalous. If you have more than answer, please explain why. Again, please show your work and code.

**Data Modeling**

Now, all your work has led up to the final part ! At this point, you probably have a good “feel” for the data so this part won’t be difficult, and it should be fun :)

1. Take the time series data for “dur” in 3 above. Predict the next 25 values in the time series based on the previous data. Justify and explain your predictions and your approach. Again, please show your work and code.
2. Take the time series data for “lports” in the “tcplife” datatable. Predict the next 25 values in the time series based on the previous data. Justify and explain your predictions and your approach. Again, please show your work and code.

**Extra Credit**

This is tough one, so don’t worry about not being able to give a confident answer, but feel free to take a wild guess !

1. What do you think the data in the “tcplife” and “exec” tables represent ?