## Evidation Data Scientist Quiz

© 2015-2021 Evidation Health, Inc.

This document may not be distributed or published without express written consent.

**Welcome!**

This is a practical quiz of your data science, visualization, and communication skills. It covers a broad range of topics, and we don’t expect you to get every question 100% right. Feel free to use external resources like Google or a calculator. This quiz will be timed, so try to set aside three hours to take it. We’re not strict about how long you take to complete the quiz, but every extra 15 minutes that you take beyond 3 hours will subtract 5 points from your score. There are 117 possible points on the quiz. Good luck!

**1. Performance Issues (12 points)**

You’re running an analysis job that you’ve run many times before on your favorite distributed big data platform (Spark, Hadoop, Presto, etc…). You notice that the job is taking far more time than it usually does.

What could have happened at the software or system level? List at least 3 possibilities, and the tools/strategies you’d use to determine if each is the cause.

**2. Statistical Certainty (15 points)**

You have yearly medical costs for 200 patients in an experimental group and for 200 in a matched control group.

1. Describe what you expect the distribution of yearly medical costs to look like.
2. How would you determine whether the experimental group has higher medical costs, including certainty?
3. You also have access to 100 other features computed on these patients (weight, medication class, number of refills,... ) and have been asked by a client to determine which of these features differ between the two groups. How do you go about this task?

**3. Machine Learning (20 points)**

You are trying to build a predictor for a rare disease from features computed on a large labeled population. There are 1K positive cases out of 100K patients and we have 400 features computed for each patient.

1. What considerations go into building your prediction pipeline?
2. Propose a specific pipeline (model and training/testing setup).
3. How would you report the results of your predictive algorithm to a clinical researcher?

**4. Freeform Exploration (30 points)**

Suppose you are a data scientist working for an organization that is investigating trends in mortality across various geographic areas within the United States. During the preliminary phase of investigation you have been supplied with a dataset that records the total number of death certificates issued across various cities, as well as the total number of deaths in which the cause of death was pneumonia or influenza. This dataset has been supplied and is named TABLE\_III.\_Deaths\_in\_122\_U.S.\_cities.csv. You may find more background knowledge on the dataset [here.](https://catalog.data.gov/dataset/table-iii-deaths-in-122-u-s-cities)

1. Your boss has asked you to conduct an initial analysis of the dataset. While there is no specific research question you are trying to answer with this analysis, you should focus on (i) using appropriate plots to summarize interesting aspects of the dataset and (ii) using methods you find most useful to highlight interesting findings in the data. Save your Jupyter notebook to the shared folder or alternatively paste your analysis below.
2. After conducting your analysis write a short (< 200 words) email to your boss (pretend their name is Luca) summarizing the results of your analysis. Your email should focus on clearly summarizing the most interesting aspects of the analysis. We would like to see how well you can articulate the results of an analysis in a semi-technical manner.

You will be graded on the clarity and correctness of your code, the creativity of your analysis, as well as your ability to clearly explain the results of your analysis.

**5. Time Series Exploration (25 points)**

Take a look at the two supplied files: (1) timeseries\_users.csv and (2) timeseries\_events.csv. Table (1) Corresponds to a table of users with their respective ids and Table (2) corresponds to the time stamped events recorded for each user.

Suppose you want to do some preliminary analysis surrounding user behaviors.

1. Plot the distribution of “event counts” for all users who are male and 30 years old. An “event count” is defined as the number of observed events for a given user.
2. Suppose you are interested in understanding how long the gap is between observed events across users. Compute the “inter-event interval” for all events (defined as the difference in time between two sequential events associated with the same user). Once you have computed these intervals, plot the distribution of these intervals. Note: Certain users may contribute more inter-event intervals to the distribution than others- this is okay.

You will be graded on the correctness and reasonableness of your code. We also expect that you choose a reasonable method for visualizing the distributions above.

**6. Observational Studies (15 points)**

We have a dataset of patients for which we have medical data (e.g., what conditions they have been diagnosed with, what medication they are on) and lifestyle (e.g., whether they are using a tracker and which one, how many steps they take per day, etc.)

We’re running a regression model to find variables correlated with two different treatments, A and B, on a diabetic cohort over the last 4 years. The model surfaces an unexpected result: Apple Watch users are statistically much more likely (than users of other wearable sensors) to use treatment A than B.

What next steps would you take before drawing any conclusions on the nature of the association discovered?