**Welcome!**

This is a practical quiz of your data science, visualization, and communication skills. It covers a broad range of topics, and 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 123 possible points on the quiz. Good luck!

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

You’re running an analysis job on your favorite distributed big data platform (Spark, Hadoop, Presto, etc…) and 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.

**Answer:**

2. **Significance Testing (15 points)**

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

1. Describe what you expect the yearly medical cost data to look like.
2. How would you determinewhether the experimental group has higher medical costs, including certainty?
3. You also have access to 100 other features computed on these patients (weight, pharma cost, 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?

**Answer:**

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?

**Answer:**

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

Take a look at TABLE\_III.\_Deaths\_in\_122\_U.S.\_cities.csv from data.gov. Imagine that your job is to quickly create an executive summary of the dataset for the CEO of a major nonprofit. Create a few plots for it showing a high level overview of the data with a sentence or two explaining each plot. Also create plots highlighting one or two insights in the data that you think are particularly interesting. For bonus points, create a simple interactive plot exploring some aspect of the data. Shoot for about an hour on this problem. Don’t worry too much about finer aesthetic points like colors or perfect axis labels, we’d rather see you finish quickly than slowly with perfect plots.

You can insert plot images & text right here, or attach a pdf / html / ipython notebook / whatever document if you prefer. Please also include your code in whatever format is most convenient.

**Answer:**

**Code:**

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

Take a look at timeseries\_users.csv and timeseries\_events.csv. This problem will be working with those files, and you can use whatever tools you think are most appropriate.

Plot a [histogram](https://en.wikipedia.org/wiki/Histogram) of total number of events per user for all male users who are 30+ years old.

**Plot:**

**Code:**

For each user, compute the list of inter-event intervals in days. An inter-event interval is the period of time between an event and the one directly before it in time for the same user. Once you have a list of all the inter-event intervals across all users, plot a histogram of them below:

**Plot:**

**Code:**

**6. Storage (6 points)**

Sort these by how long it takes to read one random byte of data: SSD, HDD, CPU L2 cache, S3 (accessed from your laptop), redis\*, RAM.

\*Assume that the redis server is running on a separate machine in the same building as your client, and that the server and client have a wired ethernet connection between them.

**Answer:**

**7. 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 different treatments on a diabetic cohort over the last 4 years. The model surfaces an unexpected result: Apple Watch users are significantly more likely to use TreatmentX.

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

**Answer:**