Project – Data Science Skills

This is a project for your entire class section to work on together, since being able to work effectively on a virtual team is a key "soft skill" for data scientists. Please note especially the requirement about making a presentation during our first meetup after the project is due.

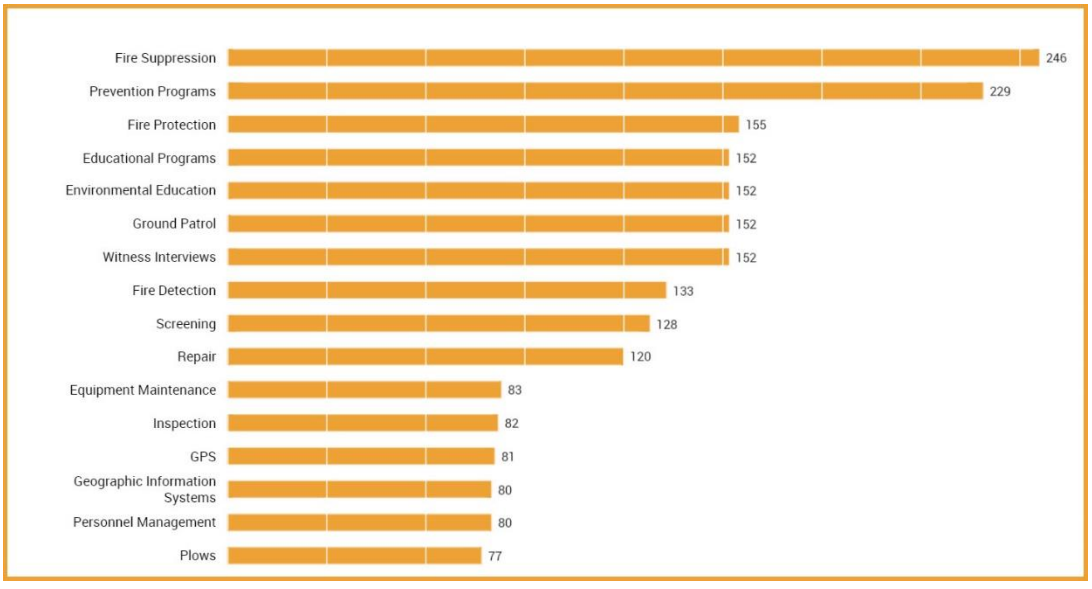
●

W. Edwards Deming said, "In God we trust, all others must bring data." **Please use data to answer the question, "Which are the most valued data science skills?"** Consider your work as an exploration; there is not necessarily a "right answer."

Grading rubric:

- You will need to determine what tool(s) you'll use as a group to effectively collaborate, share code and any project documentation (such as motivation, approach, findings).
- You will have to determine what data to collect, where the data can be found, and how to load it.
- The data that you decide to collect should reside in a relational database, in a set of normalized tables.
- You should perform any needed tidying, transformations, and exploratory data analysis in R.
- Your deliverable should include all code, results, and documentation of your motivation, approach, and findings.
- As a group, you should appoint (at least) three people to lead parts of the presentation.
- While you are strongly encouraged (and will hopefully find it fun) to try out statistics and data models, ***your grade will not be affected by the statistical analysis and modeling performed*** (since this is a semester one course on Data Acquisition and Management).
- Every student must be prepared to explain how the data was collected, loaded, transformed, tidied, and analyzed for outliers, etc. in our Meetup. This is the only way I'll have to determine that everyone actively participated in the process, so you need to hold yourself responsible for understanding what your class-size team did! If you are unable to attend the meet up, then you need to either present to me one-on-one *before* the meetup presentation, or post a 3 to 5 minute video (e.g. on YouTube) explaining the process. Individual students will not be responsible for explaining any forays into statistical analysis, modeling, data mining, regression, decision trees, etc.

You are encouraged to start early, ask many questions, actively post on the provided discussion forum, etc.



One *example* graph: Top Forest Ranger Skills (based on number of resumes with specified skills). You are encouraged to come up with your own approach that may use different kinds of data sources.