Data Science and OCS

# Data Science and The Office of Career Services

Demands on *OCS’* data are growing rapidly. In addition to the pressures of paperless training registration, OCS is also charged with answering increasingly granular questions regarding trainings delieverd by officer or contractor, hours per unit or by category, as well as evaluating the effectiveness and retention of trainings delivered. It is not difficult to imagine that, along with these descriptive figures, the Court will want more *inferential* statitsics as well. Put differently, OCS will have to describe the kind of work that is done, but also demonstrate that training has a positive effect on probation outcomes. While this report will clearly showcase how data science can improve OCS’ data maturity, the tools and methods described here can easily shifted towards other research projects.

# Tools and Methods

This report uses [R](https://www.r-project.org) and [R Studio](https://www.rstudio.com) for the statisitcal analysis as well as the body of the report. The 2017 data was captured by OCS staff into an Excel spreadsheet, commonly referred to as the “Training Database.” R allows indepth analysis and reporitng of the data that is captured, even when the data is in a messy state[[1]](#footnote-25). This tool will also allow for finding insights in the 2017 data and creating new ways to capture and store OCS data.

## Excel and Moving Forward

The 2017 Excel spreadsheet is the same spreadsheet that has been have been used to capture data in for years. Excel was chosen to the software’s ubiquity and a department-wide move away from Acess; however, given the amount of data that OCS is required to track and information it is requried to produce, Excel is no longer the best option.

For example, there currently two 2017 databases. The sworn office document contains 3 distinct workbooks tracking: Total Hours by Officer, Trainings offered in 2017, Trainer Data, and a pivot table. The 2017 Training sheet has a total of 5155 (which represents officers) and 9, which *should* represent a rough number of trainings offered. In total, there 46395 records in one sheet. The Support Proffessionals has a similar issue: Two sheets, Total hours and the trainings offered in 2017. The support staff report has 514 rows–which again represents support staff– and 7 which should represent number of trainings. This sheet has a total of 3598 records.

Both totals showcase the problems with the way Excel has been used to track our data: We do not employ 5669 staff, and we delivered far more than 16 trainings. In fact, the records suggest that 587 trainings were delivered. The way in which the data has been stored works for simple queries releated to the data such as: How many hours did each officer recieve, when were trainings offered, who conducted the trainings. Even questions about the kinds of training are relatively easy to determine. Additional questions, such as trainings per month or quarter, determining how many hours of training were delievered by each OCS officer, or which category of training was delievered by which OCS officer requires either significant knowledge of Excel or a complete rework of the way data is entered into the spreadsheet.

The future of OCS data is a full Access database that captures data by forms and produces reports by pre-written queries. Using forms and reports will ensure the integreity of the data base by limiting how data is entered and transformed. This will also allow for data to be tracked year over year, without fear of duplication or input errors.

# Trainings by Code

1. Messy, in this instance, is a reference to tidy data and data principles, and is not in any way a reflection of the capabilties of the staff in OCS. [↑](#footnote-ref-25)