**Background:**

The Lead Data Scientist in the office has asked to identify “spells” on Social Assistance (SA), measure their length and provide basic summary statistics about them. A “spell” is defined as a consecutive series of months during which a client is eligible for social assistance, with the exception that a gap in eligibility that is only 1 month long does not end a spell. A gap of 2 or more months does constitute an end to a spell. A change in the case identifier (case\_id\_anon) for a given client, even if there is no gap in time, also ends a spell.

You have been provided with a data extract from the database. The extract contains the following four variables:

* extract\_month - a code indicating the month of record in YYYYMM format
* case\_unit\_count - indicating that the client was eligible for assistance that month
* integrated\_case\_id\_anon - identifying a particular social assistance client
* case\_id\_anon – an identifier for a particular social assistance case; each client may have several cases over time

**Assignment:**

* Please provide code to create an output file with the following format: Each row should correspond to a single spell and include the integrated\_case\_id and case\_id\_anon corresponding to the spell, the spell start month, and the spell end month (last month in the spell).
* What would be the best way to describe the distribution of spell lengths? Are there any inclusion or exclusion criteria related to the data extract that may affect the interpretation of your results or that would be advisable to use? Consider that the beginning of the extract may contain spells already underway. What considerations are relevant near the end of the extract?
* Is there a statistical difference between the length of first-time spells vs. repeat spells?
* Please describe the distribution of gaps between spells (remember, 1-month-long gaps are not considered spell gaps). Is there a difference between gaps that result in a new case\_id\_anon and gaps that do not?
* What distribution could best describe the number of spells per client?

For the interview, prepare a 10-minute presentation to the Lead Data Scientist that highlights your findings from the data. In that time, you are also asked to provide a run-through of your code.