Self-Assessment 1

Signal Data Science

Today, you'll be completing a short assessment so that we can get a better sense of your individual level of progress.

- Type your answers in a new R script file with comments indicating where the answer to each question begins.
- When you finish, email `signaldatascience@gmail.com` with your R script attached along with the amount of time you spent on the self assessment.
- Work individually. You can however consult R documentation, look at old assignments, use the Internet, etc., but don't copy and paste code verbatim.
- Make your code as clear, compact, and efficient as possible. Use everything that you've learned! **Please comment and organize your code so we can easily tell how parts of your R script correspond to specific problems.**

Packages you will find useful: `ggplot2`, `psych`, `dplyr`.

Part 1: R and Probability

Here's an interview question from *Euclid Analytics*:

> Suppose that X is uniformly distributed over $[0, 1]$. Now choose $X = x$ and let Y be uniformly distributed over $[0, x]$. Is it possible for us to calculate the "expected value of X given $Y = y$", *i.e.*, $\mathbb{E}(X \mid Y = y)$?

(If you don't know what *expected value* is, you can think of it as the mean of each possible outcome weighted by its probability.)

Now, we don't know the answer yet, but maybe we can get some sense of what it might look like by doing some Monte Carlo simulations. To that end:

- A *single trial* of the process described in the problem will yield a pair of values (x, y), where the probability distribution which y is drawn from depends on the value of x. Simulate k trials of this process for $k = 10000$. (You may find `runif()` helpful.)

  - Plot the simulated values with `qplot()`, putting values of y on the vertical axis and values of x on the horizontal axis.

- Since we're interested in the *expected value* of X given some $Y = y$, we can approximate this by separating our values of Y into *bins* (of equal width) and taking the *mean* of X within each bin.
  - Write code that does the above for a bin width of w, setting $w = 0.01$ (equivalently, setting the number of bins to $1/0.01 = 100$).
  - Use `qplot()` to view the results. Do they make sense?

The correct theoretical answer is in fact

$$\mathbb{E}(X \mid Y = y) = \frac{y-1}{\ln y}.$$

In light of this new knowledge, you want to verify your computational results from earlier. To that end:

- Choose many different values of y in the interval $[0, 1]$ and calculate the corresponding values of $\mathbb{E}(X \mid Y = y)$ according to the equation above. Your choice of y shouldn't matter much, but one common practice is to use the midpoints of the bins in which you computed the mean of X, so as to make the comparison between theoretical and simulated values of $\mathbb{E}(X \mid Y = y)$ as fair as possible.
  - Graph the results using `qplot()`, putting y on the horizontal axis and $\mathbb{E}(X \mid Y = y)$ on the vertical axis. Does this graph match your simulated results?
- Make a *single* dataframe with both your Monte Carlo-simulated results *and* your direct calculation of the theoretical result.
  - Make a single graph with (1) a scatterplot of the Monte Carlo-simulated results and (2) a smooth line connecting the points corresponding to the theoretical values. This should just basically be both of your previous graphs plotted at the same time.

Part 2: Data Analysis

In the next part, we'll be looking at psychological test data.

Note: When running a linear regression, you can pass into `lm()` the formula `y ~ .` to regress against all other variables (instead of, *e.g.*, `y ~ a + b + c + ...`).

- Call `help(msq)` to read about the `msq` dataset (loaded from the `psych` package). For convenience, set `df = msq`. We'll be working with this dataset in the following problems.
- Compute the fraction of missing values for each feature, sorted in descending order. The first of these should be 0.528747 for `"kindly"`.
- Make a new dataframe with columns `"active"` through `"scornful"` as well as Extraversion and Neuroticism.
- Replace each missing value with the mean of the column which the missing value is in.
- Create each of the following plots:
  - histograms for Extraversion and Neuroticism (`geom_histogram()`),
  - density plots for Extraversion and Neuroticism (`geom_density()`), and
  - a scatterplot of Extraversion scores vs. Neuroticism scores, with a smoothed nonlinear fit to the points overlaid on top (`geom_smooth()`).
- Run linear regressions of Extraversion and Neuroticism against all the other features. Neuroticism should *not* be one of the predictors for Extraversion and vice versa, but `"active"` through `"scornful"` should be included simultaneously. (**Read the note above** to learn how to easily regress one variable against all others without having to type out the name of every predictor.)
- Use `coef()` on each of the two linear models to view the coefficients associated with each linear fit.
  - Print out the top 10 coefficients, *not including the intercept*, for each of the two linear fits. They should be ordered by absolute value (largest to smallest). Interpret the results.

Part 3: SQL Queries

Finally, we'll finish with three SQL questions! You can write your answers to these questions as comments in your R script file.

- What's the difference between `WHERE` and `HAVING`? (Just write 1-2 sentences.)
- Suppose you are given a table `Employees` with a single column `Salary` of integers. Write **two** different SQL queries to determine the second highest distinct salary.

If you've gotten to the SQLZoo section about `JOIN` operations:

- What's the difference between `LEFT JOIN`, `RIGHT JOIN`, and `INNER JOIN`? (Feel free to illustrate the differences with examples if that helps you communicate the differences.)
- Given a `COURSES` table with columns `course_id` and `course_name`, a `FACULTY` table with columns `faculty_id` and `faculty_name`, and a `COURSE_FACULTY` table with columns `faculty_id` and `course_id`, how would you return a list of the names of faculty members who teach a course given the name of a course?