Data analysis exercise information sheet

You have been provided with two data sets (.csv format).
Data set **applicants.csv** contains information collected about job applicants applying to one or more job positions at a specified organization. Each row corresponds to information on an individual applying to a single job position. The fields and their definitions are listed in Table 1.

Data set **hires.csv** contains information collected about a subset of these applicants that were hired to fill a job position at one of the specified organizations. Each row corresponds to an individual employee. Table 2 lists the fields contained in this data set.

The field *User_Id* is contained in both files and can be used to join applicants to their associated hire information. Note that some of the individuals contained in the **hires.csv** file are currently employed, thus the stated tenure length represents tenure length at the time of the most recent data extract. A few individuals have tenure lengths that are not known, i.e. their tenure_length value is missing.

Instructions

Use these two data sets, and whatever tools you are most comfortable with, to answer the following questions to the best of your ability.

1. Data Exploration
    a. Which organizations have the greatest applicant-to-hire ratio?
    b. Which organizations have the greatest challenge with attrition in the first 3 months of employment?
2. Data Analysis
    a. Do 6 month attrition rates vary significantly across different job categories?
    b. How does the type of device used to take the questionnaire impact an applicant's tenure length, if at all?
3. Predictive Modeling
    a. Develop a model that can be used to predict whether an applicant to one of the given organizations and job positions is likely to remain employed for less than 6 months, between 6 and 12 months, or greater than 12 months. Provide information on any assumptions you have made, as well as steps you may have taken to address potential issues in the data.
    b. Using a method of your choosing, evaluate the performance of the model and provide an assessment of whether your model(s) would be useful for making hiring recommendations to HR managers at the specified organizations.

Please submit a brief writeup (3-4 pages at most) that summarizes your findings in a form that could be understood by a non-technical colleague, along with the code used to conduct your

analysis, via e-mail to bonnie@peggedsoftware.com. The report can be submitted in the format of your choosing, e.g. Word/PPT, Google Doc/Slides, PDF, R markdown report, Jupyter notebook, whatever works, as long as it is readable and contains explanations of problem/approach/summary in English (not just code).

The results of your analysis will be assessed based on the following criteria:
Report:

- Readability
- Statement of assumptions and methods used
- Appropriateness of the approach(s)
- Discussion of results

Code

- Structure
- Readability (includes some basic documentation)
- Executability (can we get it to run?)

We are not necessarily interested in having you spend most of your time coming up with the absolute best predictive model for the stated problem. Our interest is more in gaining an understanding of how you approach different types of questions, i.e. your analysis *process*. We suggest that you spend no more than half of your time on the first two questions, saving sufficient time to work through Question 3.

Table 1: Field definitions for data contained in **applicants.csv**

Column name	Data type	Description
user_id	int	Unique user_id for each applicant
client_name	string	client organization to which the candidate applied
answer1:answer25	int	answers to 25 questions included in questionnaire completed by applicants (encoded as integers) **Note:** most of the questions required selecting an answer from among a set of four or five possible choices
log_total_time	float	the logarithm of the amount of time the applicant took to complete the screening (in seconds)
device	string	the device that the applicant used most often to complete the screening

Table 2: Field definitions for data contained in **hires.csv**

Column name	Data type	Description
user_id	int	Unique user_id for each employee at time of application
client	string	Client organization for whom the employee works
tenure_length	int	tenure length of the employee (in days)
currently_employed	string	Yes/No indicator
hire_job_category	string	job category into which the employee was hired