

Cyence collects, curates and models a variety of data to quantify the financial impact of cyber risk. The idea of this exercise is to simulate some of day-to-day data problems we tackle at Cyence. We hope you enjoy exploring the dataset in this exercise!

Background:

- U.S. Department of Labor's (DoL) permanent labor certification program (PERM) allows a U.S. employer to **hire a foreign worker** to work permanently in the U.S.
- A PERM application collects various information on the **employer**, the **job offered** and the **employee** – rendering itself a rich dataset on **U.S.'s labor market**
- A subset of the PERM application is the subject of analysis in this exercise (**data_perm_take_home.csv**)

Goal:

- Identify the characteristics most predictive of a response variable, build a robust predictive model, and communicate the assumptions made as well as guidance on interpretation

Instruction:

- The exercise is broken down into 2 parts (further instructions in the following pages):
 - **Part 1:** 4 short questions, for which we expect just answers with no explanations needed
 - **Part 2:** an open-ended data analysis/modeling problem, with some prompts and for which comments are recommended
- Use either R or Python (*if you use python 3, you will have to read the csv using "ISO-8859-1" encoding*)
- When unsure, feel free to make and document any necessary assumptions (after all, data analysis / modeling is a series of assumptions!), and Google / consult Stack Overflow
- Return the results within **3 hours**:
 - The results should contain a **short summary** as well as the **code**, which could be a markdown file or a script that contains **enough comments to communicate interpretations/assumptions made**
 - The time constraint is intentional, and we understand candidates have other obligations; there will be more to explore than the time allows, so please feel free to **start off sufficiently simple**, and **outline any next steps if given more time**

This exercise will be evaluated for data analysis/modeling proficiency and prioritization, which includes:

- Experience with the programming language of choice (R or Python)
- **Informed interpretations**, **assumptions** and **modeling strategies** made on the data

Data Dictionary:

- **case_number**: identification of each case
- **case_received_date**: when the case was received
- **decision_date**: when a decision on the case was made
- **case_status**: the decision on the case
- **employer_name**: the employer applying for the case
- **employer_num_employees**: the number of employees employed by the employer
- **employer_yr_established**: the year the employer was established
- **job_education**: the education level required for the job offered
- **job_experience_num_month**: the experience required for the job offered, in months
- **job_state**: the location of the job
- **job_foreign_lang_req**: if a foreign language was required for the job
- **job_level**: the level of the job offered
- **employee_citizenship**: the citizenship of the employee
- **wage_offer**: the wage offered, or the lower-end of the wage offered
- **wage_unit**: the unit of the wage

Part 1: Exploratory Data Analysis / Warm Up

1. Which employer has the most entries in the dataset?
2. Which employer has the most "Certified-Expired" cases?
3. How many unique employers contain the character string "APPLE" that is not "APPLE INC."?
4. Write a function that plots a histogram with 50 bins of the duration between from **case_received_date** to **decision_date** for any given **employee_name**. Use the function to plot for "APPLE INC." and "GOOGLE INC."

Part 2: Data Modeling

For **certified cases (case_status = "Certified")**, build a model that **predicts wage_offered** by using **5 features**** with the most predictive power. Some considerations:

- How does the response variable look? How do the predictors look?
- How might **missing values** be handled?
- What transformations are relevant?
- Are there any **outliers**?
- What additional features could be engineered?
- What features are most predictive of **wage_offered**?
- Which model is picked? What advantages does the model offer against an alternative?
- What **do the errors/residuals and their distributions tell us**?
- How **well do you think this model will extrapolate to the entire U.S. population**?
- What are ways to **improve** this model?

** Also known as independent variables, covariates or predictors. A categorical feature can be counted as 1 feature. For example, X1 is a categorical variable with 3 categories A, B and C. X1_A, X1_B and X1_C can be counted as 1 feature, not 3 features.