Project Data Proposal

Group Members: Yue Wang, Zoe Chen, Tiger Zhang
Dataset: IBM HR Analytics Employee Attrition & Performance
Dataset Link: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

This dataset contains HR data from IBM. It has 1470 observations. Variables of interest are listed below. The list below is not finalized and is subject to change as the research moves on.

Variable Name	Definition	Range
MonthlyIncome	Monthly Income of an employee	[1009,19999]
Attrition	Whether the employee left the company	1 or 0 | Yes or No
WorkLifeBalance	Work-life balance - the higher the better	1,2,3,4
Department	Which department the employee was/is at	HR, R&D, Sales
College(from Education)	Whether the employee has a bachelor degree or above	1 or 0 | Yes or N
Age	Age of the employee	[18,60]
Gender	Gender of the employee	Female or Male
Married	Marital status of the employee	1 or 0 | Yes or No
TrainingTimesLastYear	# of training taken in the past year	[0,6]
YearsInCurrentRole	# of years in the current role	[0,18]
StockOptionLevel	Stock and option level - the higher the better	0,1,2,3
NumCompaniesWorked	# of companies the employee worked at	[0,9]
DistanceFromHome	Distance Level from home to office	[1,29]

Research Questions:

1. What factors impact an employee's monthly income?
2. What factors contribute to an employee's attrition?

We are interested in exploring these questions as employee retention and fair offer offering are two important topics for companies' human resources departments.

Types of Analysis:

1. Linear Regression to model question #1.
2. Generalized Linear Model(i.e. Logit Model) to answer question #2.

Comments, Questions, or Concerns:

1. Can we still run OLS regression when predictors are in the unit of levels? Do we need to transform them? We believe we can still interpret the coefficient estimates by concluding a level increase in the predictor will increase/decrease the dependent variable by the coefficient amount unit.