Project 2: Second Proposal

Michael Langmayr (ml8vp), Jonathan Gomez (jag2j),   
Yihnew Eshetu (yte9pc), Nathan England (nle4bz)

# Objective and Questions

We plan to investigate employee retention and monthly income.

**Question**: What factors influence attrition?

*Example model:* Attrition ~ MonthlyIncome + JobInvolvement + YearsInCurrentRole

*Example model*: Attrition ~ DistanceFromHome + RelationshipSatisfaction + WorkLifeBalance

**Question**: Is income related to education and job experience?

*Example model*: MonthlyIncome ~ Education + JobLevel + JobRole + TotalWorkingYears

# About the Data

We will use the *IBM HR Analytics Employee Attrition & Performance* dataset available on Kaggle ([link](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)). According to the documentation on Kaggle, this dataset was synthesized at IBM for experimentation in data analytic methods. We were unable to find the original source on IBM’s website.

The data offers a chance to demonstrate the techniques we studied in class. Observations describe hypothetical individuals, featuring information on the following: attrition, frequency of business travel, income, distance of workplace from home, education level, field of study, satisfaction with work environment, gender, job role, marital status, employment history, years at company, years in current role, and years with current manager.

We overcame four initial reservations about this data set.

* First, we were concerned about the fictitious nature of the data. We reasoned, however, that we could still tell a story around the data. If the data were to prove unusual, we could still potentially express the divergence from standard models.
* Second, we were initially concerned that many of the predictors were categorical rather than quantitative. Ultimately, we feel this is a good opportunity to work with indicator variables in our models.
* Third, some features must be interpreted at face value. For example, WorkLifeBalance ranges from “Bad” to “Best”, without any specific objective criterion. In fact, the values are likely hypothetically subjective responses to surveys. However, as long as we are circumspect in how we discuss them, we can navigate this.
* Lastly, it is worth noting that one of the fields, EmployeePerformance, includes four categories: “Bad”, “Good”, “Excellent”, and “Outstanding” according to the documentation on Kaggle. However, when looking at the data, it appears that all employees fall into either the “Excellent” or “Outstanding” categories. For this reason, the scope of this analysis is limited to high performing employees.

# Data Cleansing

Being made for analysis, the data does not require high levels of normalisation. However, the nature of the columns must be correctly identified. For example, MaritalStatus should be loaded as a factor rather than a character vector.

# Methods of Analysis

We anticipate this project to use several modeling methods learned from the course including multiple linear regression and logistic regression.

Multiple linear regression will be used to answer questions with quantitative outcomes (response variables). These questions include the objective questions pertaining to employee monthly income.

Logistic regression will be used to answer questions with binary outcomes (response variables). The questions include the objective questions pertaining to employee attrition (i.e predicting whether or not an employee is likely to leave the company).

Specific modeling techniques we plan on utilizing include (but are not limited to):

* model selection through the use of automatic search procedures
* model validation by assessing model fit, coefficient interpretation, and utilization of training and test sets
* Performing model diagnostics and implementing remedial measures through the use of residuals analysis, looking at leverage and measures of influence, and transforming the response and/or predictor variables when necessary