7-2 Project Two: Part 1: Narrative

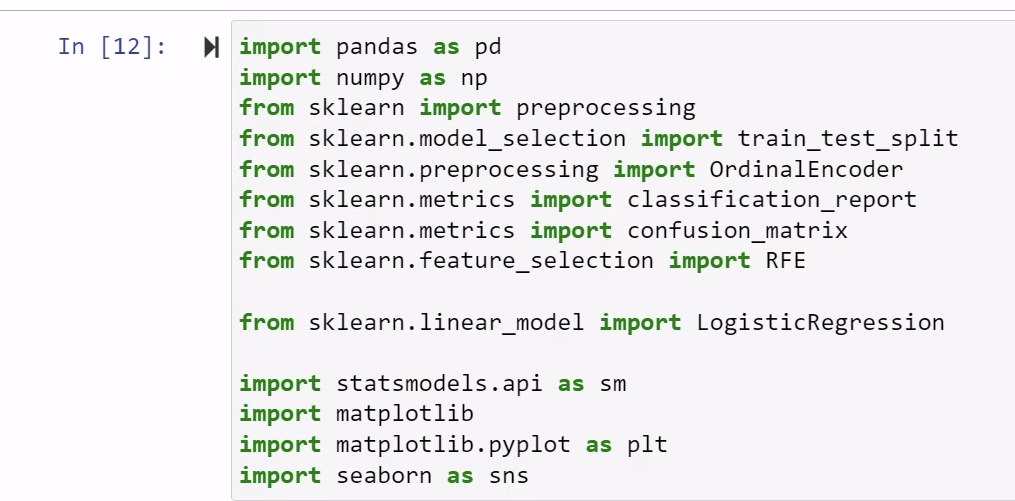
DAT-430

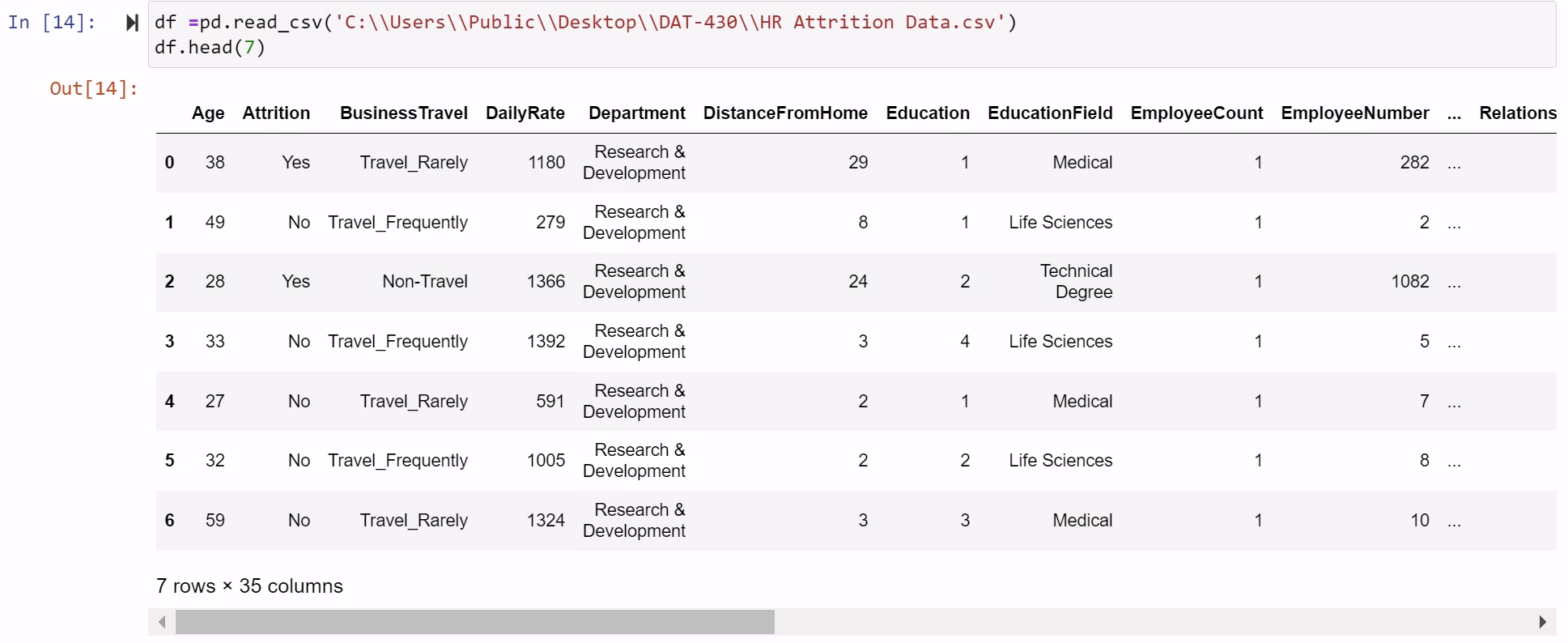
Ashley Littles

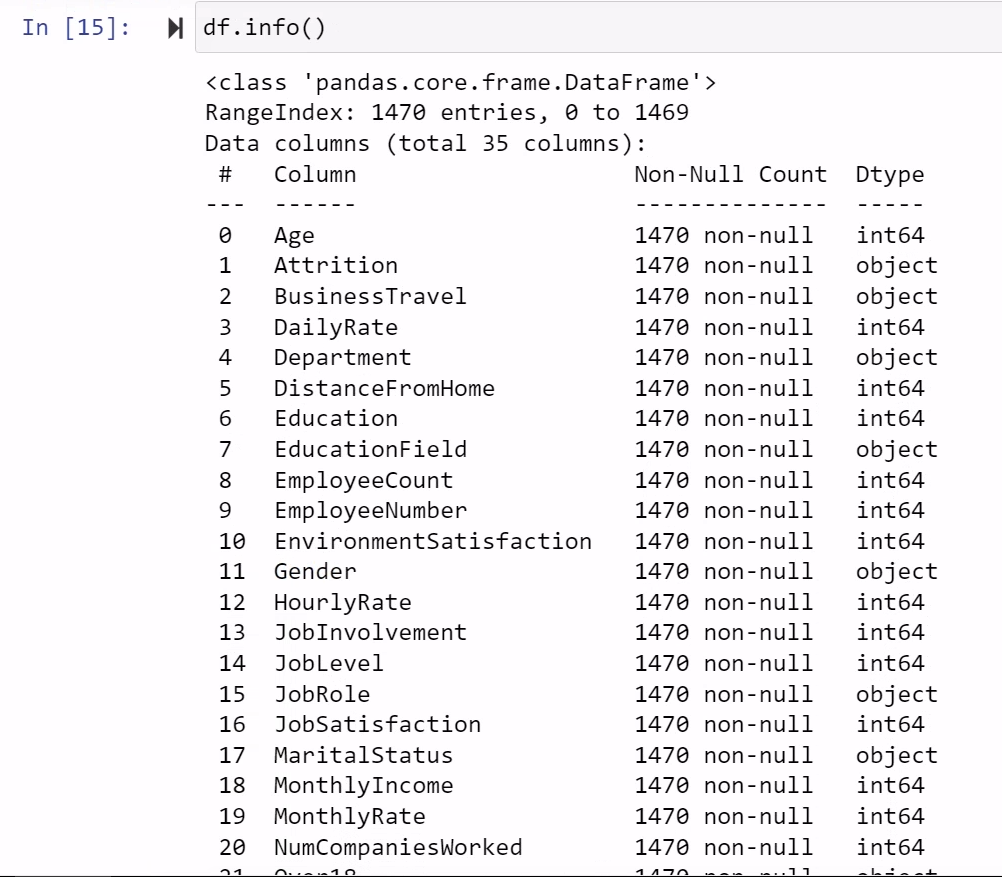
Southern New Hampshire University

Exploratory data analysis

An exploratory data analysis (EDA) is used to identify general patterns within a data set. The exploratory data analysis is the first step in any kind of data analysis. This type of analysis helps us understand where the outliers are and how our variables are related to them which in turn aids in designing a statistical analysis that gives meaningful results. There are a few different graphs that can be used with exploratory data analysis such as histograms, boxplots, cumulative distribution functions, and quantile-quantile plots. When creating an EDA, we will start by importing our libraries and data.

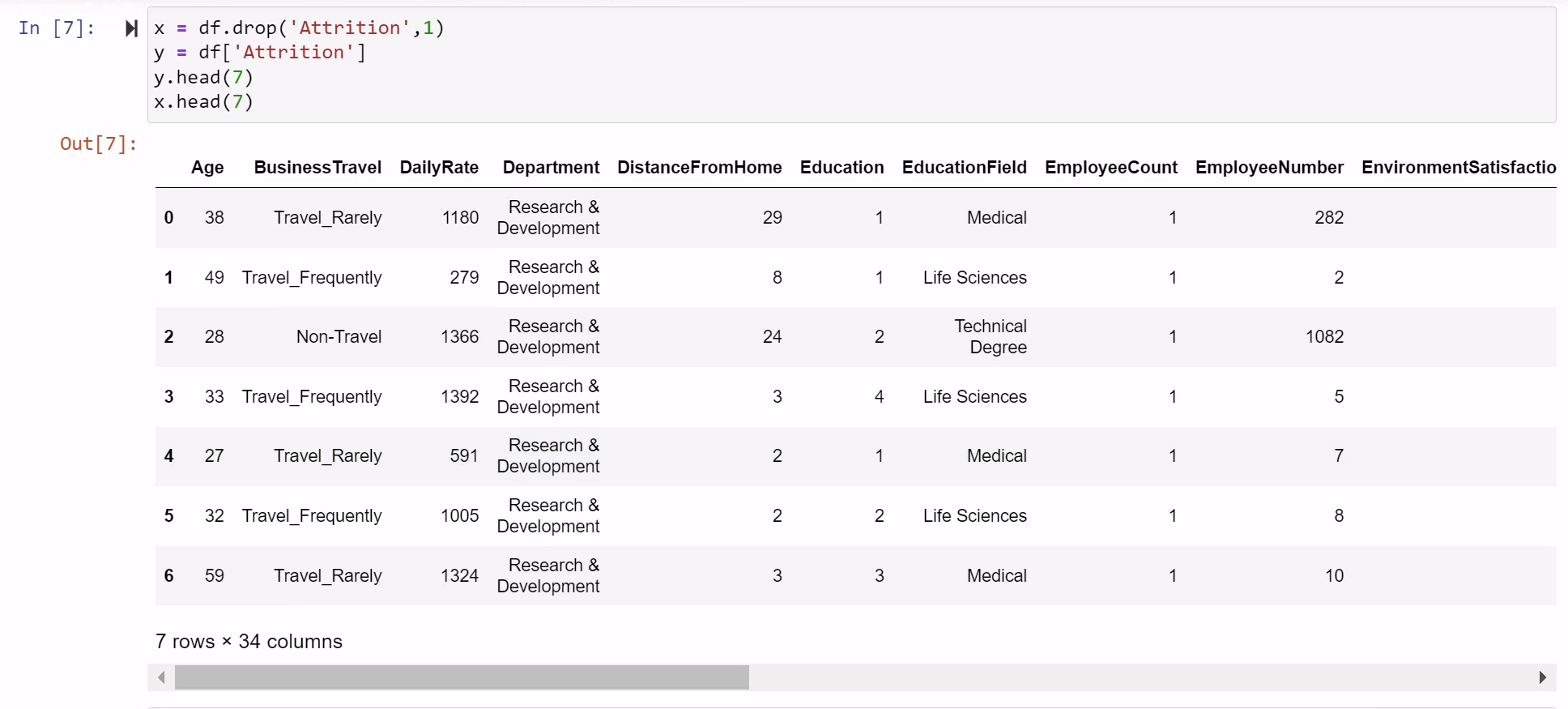


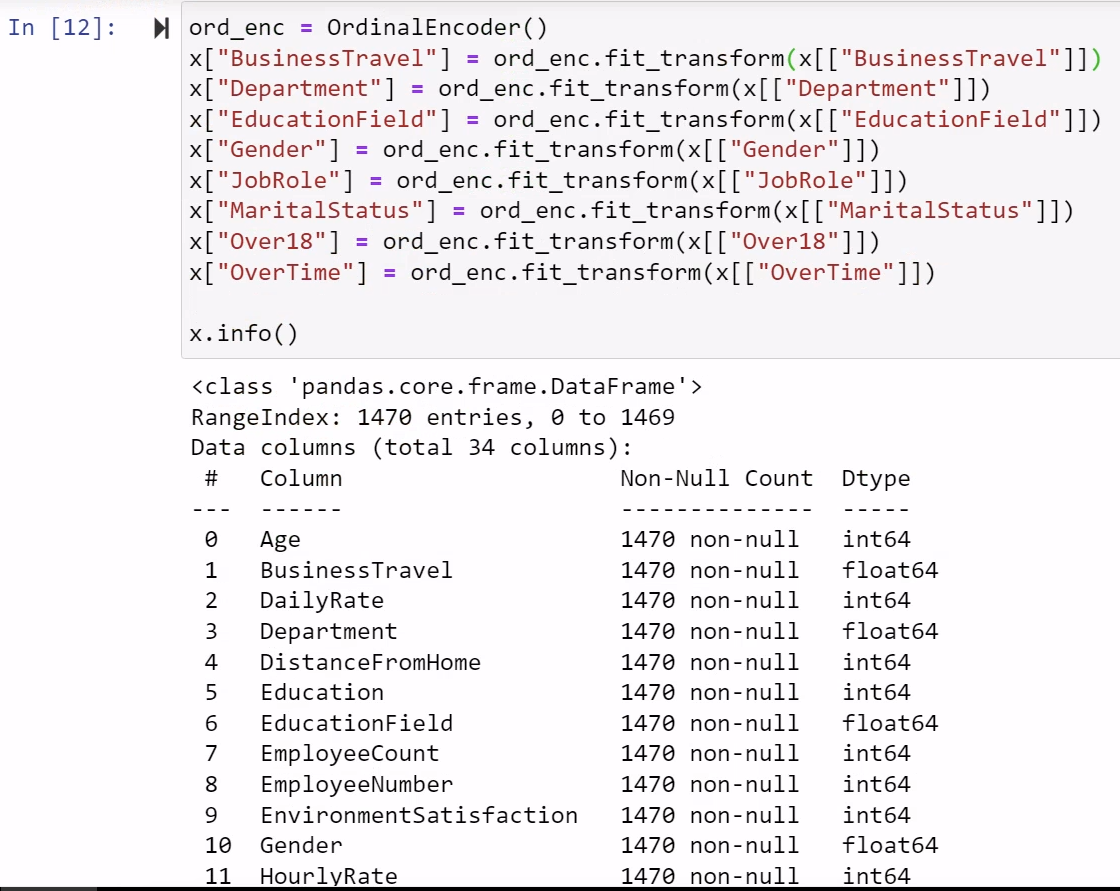
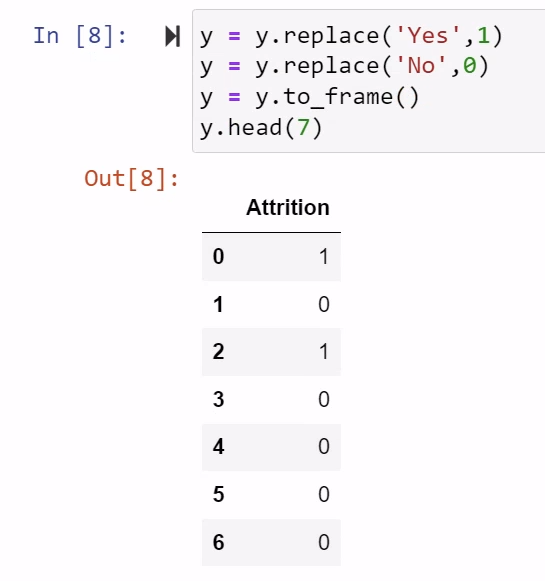


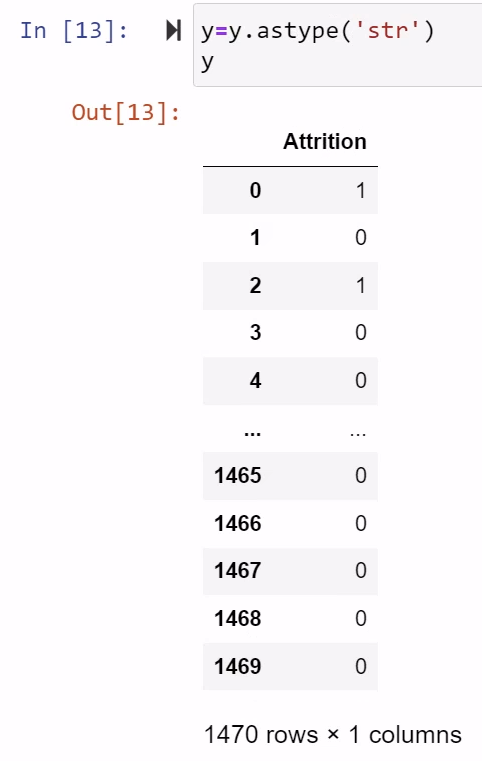


Feature Engineering

To select and create the right features for my predictive model I labeled the x and y variables. The x variables are all attributes in the data set besides the attrition attribute. The y variable is the attrition attribute. I displayed the first seven rows of the data set to show these variables. By utilizing the y.replace() command I was able to designate yes to one and no to zero. We can see the results with the attrition from the output. I then used the ordinal encoder feature to update the variables' data types and convert our y variable to a string.

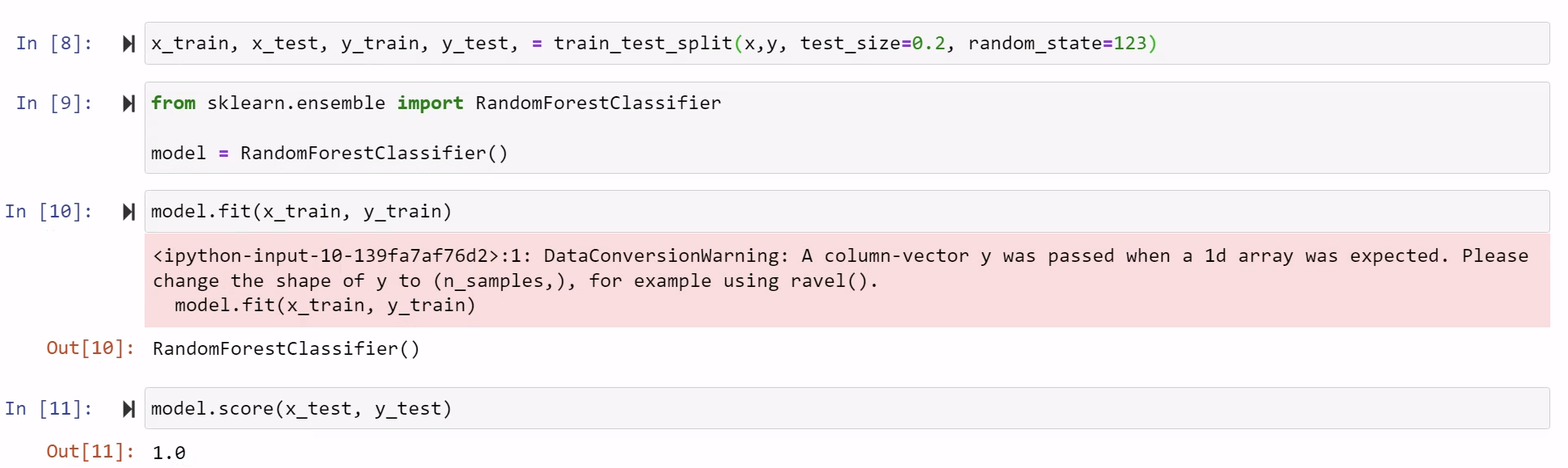




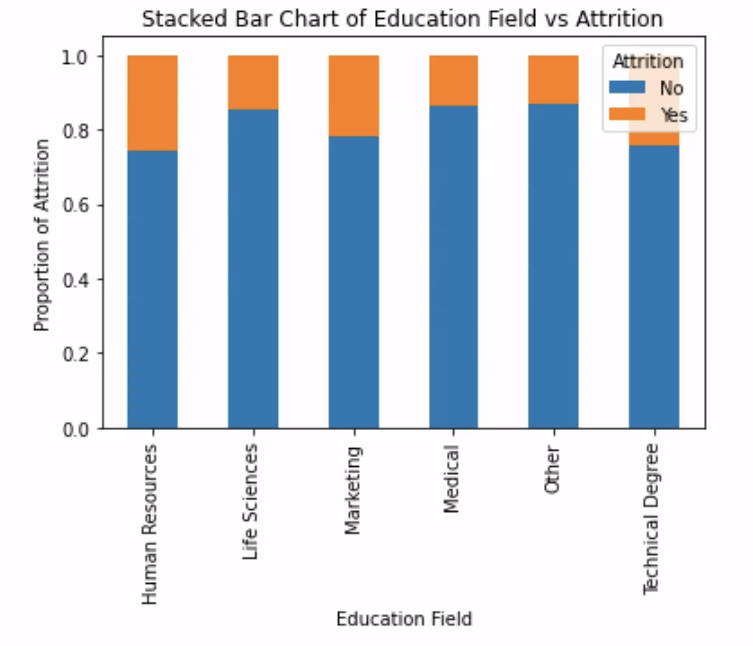


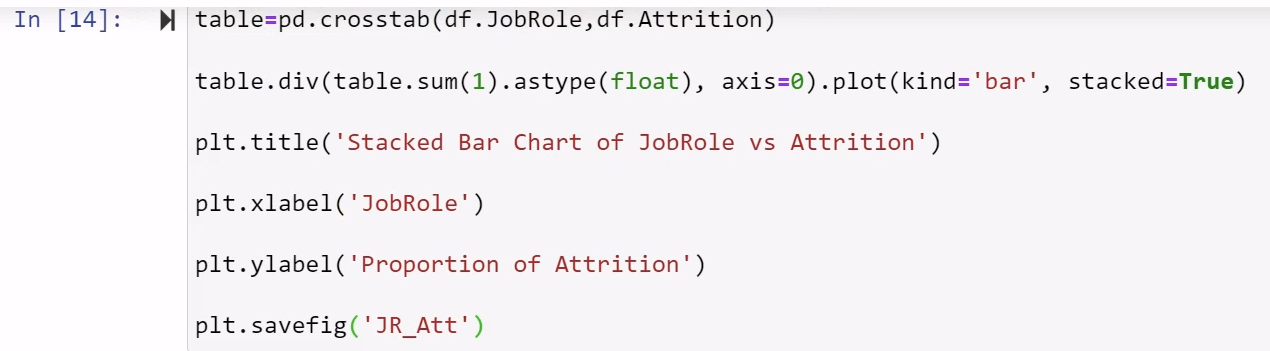
Predictive model selections

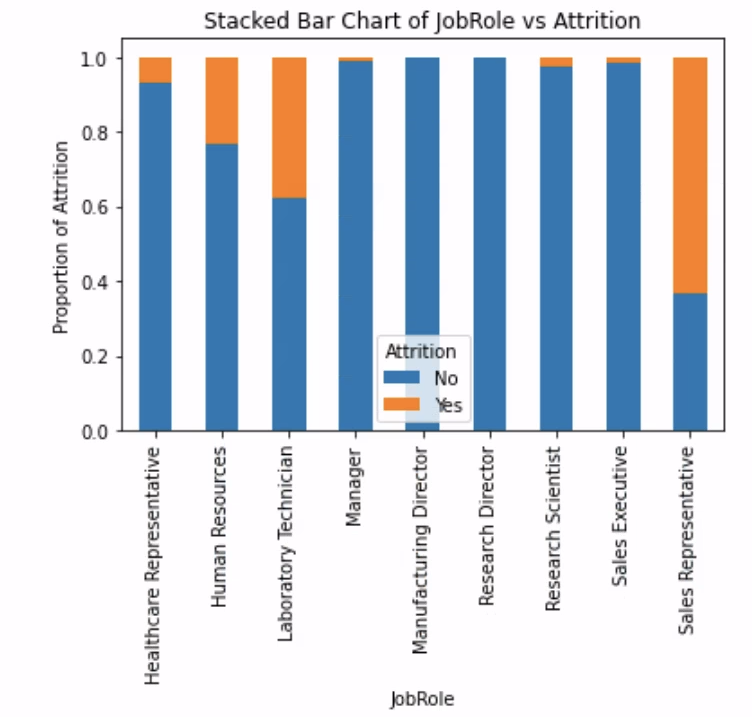
The two predictive models that I selected to show the variation in predictability are the random forests model and the logistic regression model. I began by importing the machine learning algorithms and testing the data to ensure they are accurate. I then created the logistic regression models with two bar charts showing education field compared to attrition and job role compared to attrition. Lastly, I added in a histogram model to show age and frequency.

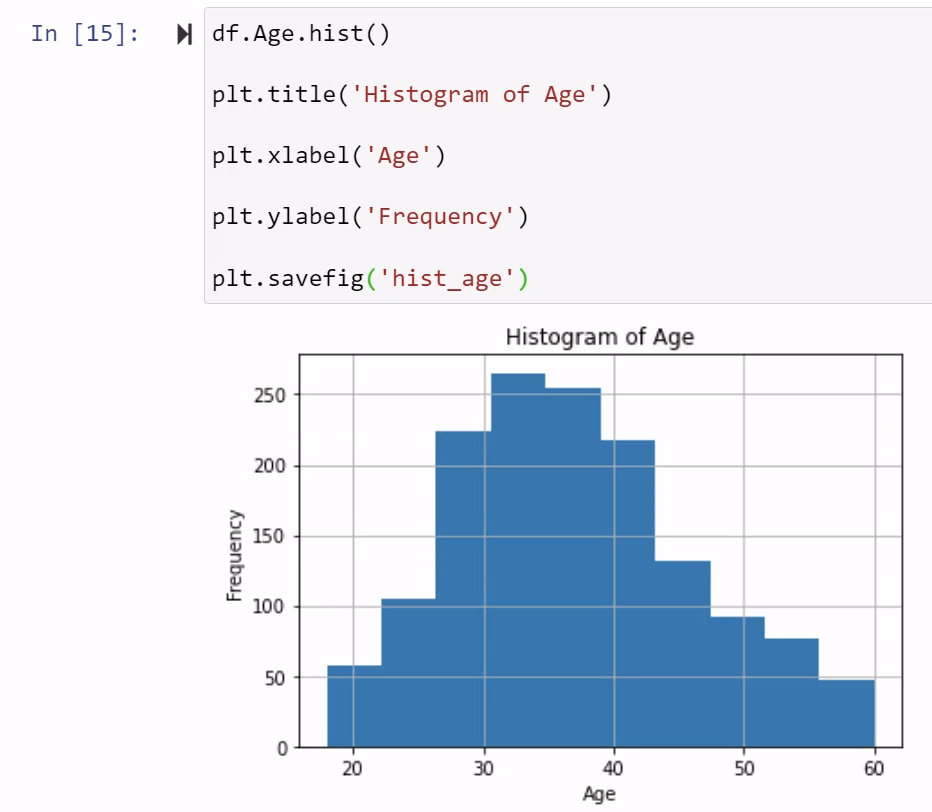






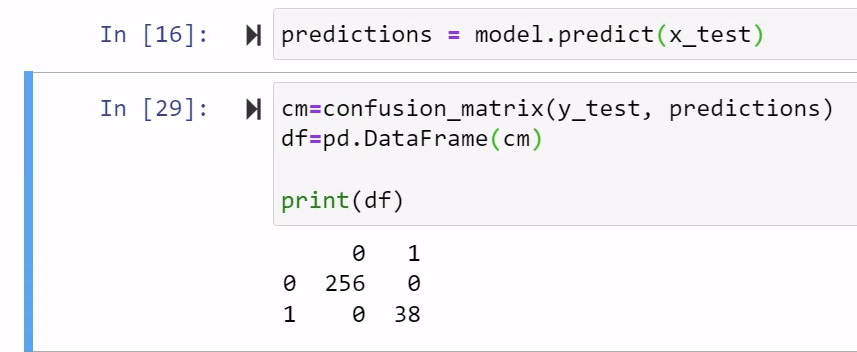


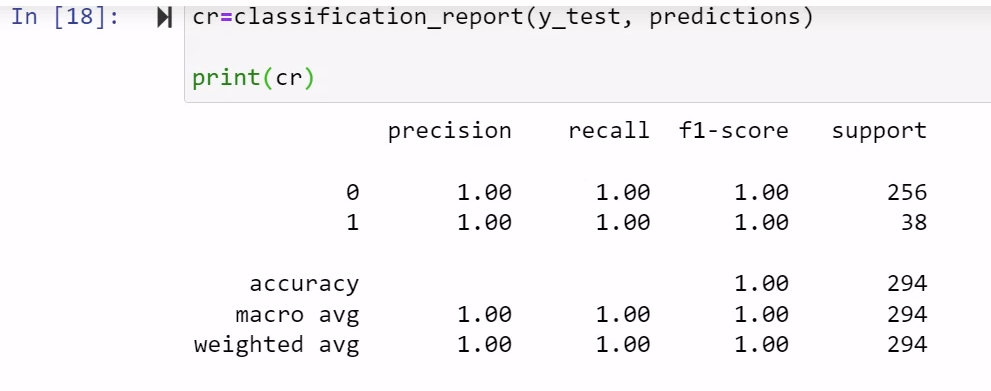


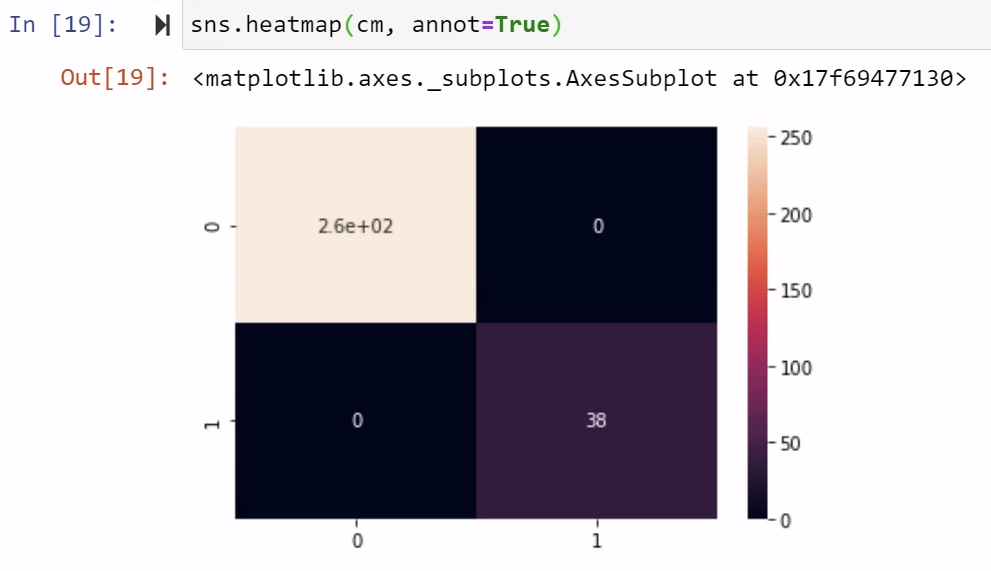


Predicting Outcomes

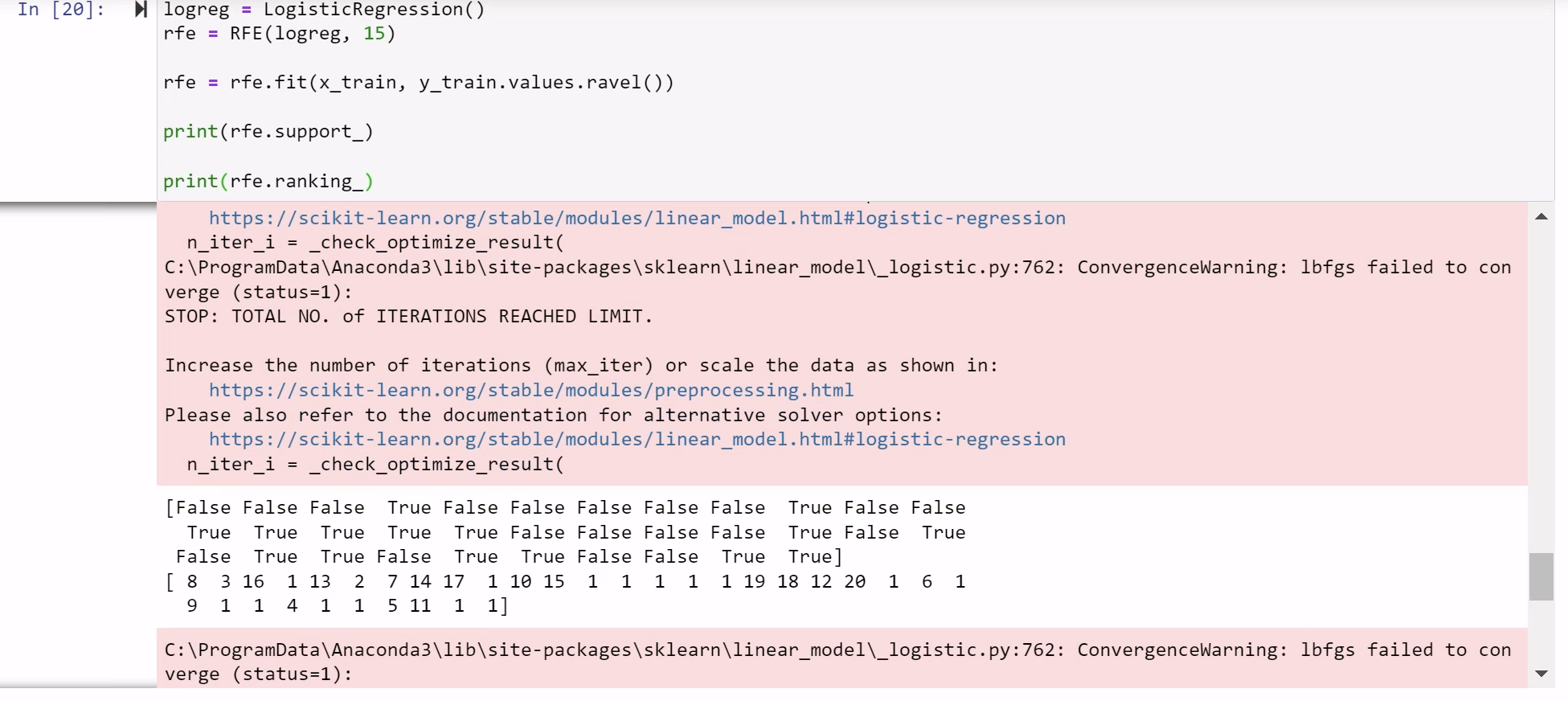
To accurately predict outcomes based on the data we have I used the confusion matrix and a classification report with our test data. The results from these commands show tables with the exact numbers of precision.

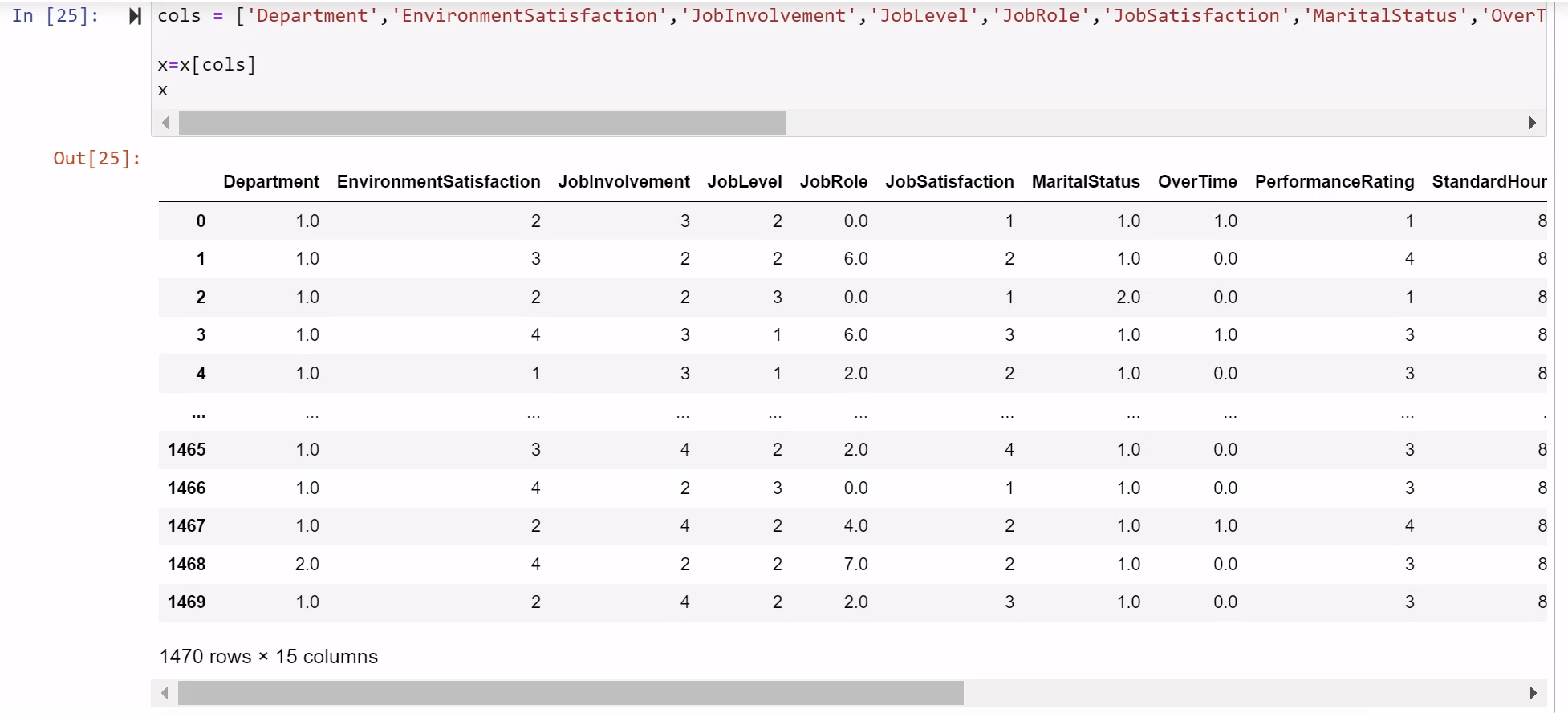






I then proceeded to create a heat map to help visualize the results of the confusion matrix displaying the actual values compared to the predicted values.





Lastly, for feature selection I used the Recursive Feature Elimination (RFE) command to show the fifteen features that are relevant from the data set. Knowing all of these features can help us accurately predict any outcomes based on the data. I am being asked to figure out what the cause of the employee turnover is within the organization using the employee data. I can answer this by analyzing the employee data to create regression models to determine what factors are leading to attrition.

References

EPA (February 29, 2024) *“Exploratory Data Analysis”* <https://www.epa.gov/caddis/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,step%20in%20any%20data%20analysis>.

Geeksforgeeks (March 06, 2024) *“Confusion Matrix in Machine Learning”* <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>