**Project One: Part 2 – Python Code**

Kayla Sacks

Southern New Hampshire University

DAT-430 Leverage Data for Org Results

Professor Aubrey Condor

September 29, 2024

## Merging Data Sets

To start merging, I imported the two libraries that I will use, pandas and glob. After important the libraries, I created a list of the 15 files to be merged using glob then removed the ‘HR Attrition Data.csv’ file. These steps can be seen below.

A screenshot of a computer

Description automatically generated

After creating the list of csv files, I used .read\_csv() to read all files into a list and then created a data frame list of all files that will be merged. Before merging the files, I changed the columns named “TrainingTimesLastYear” to “training” and checked that the renaming was successful to ensure consistency when merging the data.

A screenshot of a computer

Description automatically generated

Next, I merged the data frames within the list and viewed the merged data frame to ensure it merged successfully before saving the new merged data set to a csv file as seen below.

A screenshot of a computer

Description automatically generated



## Preprocessing the Data

To create a structure that can be used by the junior analysts, the data set will first need to be cleaned. To start with I first checked for missing values. During this, I found that there were 64 rows that were missing values for ‘Attrition’. Since this is an important variable that is needed, I removed these rows from the data set as seen below.

A screenshot of a computer

Description automatically generated

The next step was to check for duplicated rows. We can see from the screenshot below that there are 1470 duplicated rows.

A screenshot of a computer

Description automatically generated

I removed the 1470 rows as they can affect the accuracy of an analysis.

A screenshot of a computer

Description automatically generated

This leaves 5506 rows of the original 7040. I then dropped variables that are redundant or have errors, such as ‘MonthlyIncome’, as seen below. This leaves 28 of the 35 rows to use for analysis.



A screenshot of a computer

Description automatically generated

Once I was done, I saved the data frame as a new data set. I left more variables in the data set then I planned in case they are needed for additional analysis patterns change due to updated information being added.