Created by: Dan MacCarthy Aspiring Data Scientist

Date: 11th August 2021

Course: UCD Professional Academy

Specialist Certificate in Data Analytics Essentials

Assignment: Find a dataset on Kaggle

Clean the dataset using Python and Pandas

Prepare the dataset for data analysis

Repository: DanMacCarthy/Practice\_Cleaning\_Data

Source Dataset: <https://www.kaggle.com/kewagbln/absenteeism-at-work-uci-ml-repositiory>

The database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil

With acknowledgement and thanks to Kerstin Wagner Research Assistant at Beuth University of Applied Sciences Berlin, Berlin, Germany. Dataset Owner

Number of Instances (Rows) = 740

Number of Attributes (Columns) = 21

**Elements from Source Dataset**

Absenteeism\_at\_work.csv is the raw dataset from Kaggle

Attribute Information is a guide to the dataset

**Elements from the Data Cleaning Exercise**

CSV file ‘Simpler Column Names’ is an initial work file mapping more concise column names for clarity.

This CSV file was exported into PyCharm and called in the Python script absenteeism.py

CSV file ‘Key for causes of absence’ is an initial work file mapping the source file’s codes for medical causes of absence (International Disease Codes IDC) with simpler descriptive names for simplicity and readability.

Absenteeism.py is the script that I developed to clean and prepare this dataset for exploratory data analysis

**Summary of Cleaning work**

# 696 observations over 18 columns remaining after cleaning

# Several columns changed from integer to category datatype

# All category columns have categorical values in correct range

# 3 columns not of interest were dropped

# rows with absenteeism\_time\_hours of zero were dropped

# Column names were replaced with clearer, more concise descriptors

# ICD codes for medical causes for absence were replaced with concise descriptors

# a single categorical value 'routine\_medical\_appt' replaced 7 category values

# All column datatypes correct