Daniel Greiner
December 15, 2022
DS 210 Final Project

For my final project, I decided to use Dijkstra's shortest path algorithm to solve a six degrees of separation problem. I downloaded a csv file with over 1.2 million entries of famous people from kaggle and kept approximately 50,000 entries in a new csv file named famous.csv. Each row corresponds to a famous person throughout history, with a column for name, short description, gender, country, occupation, birth year, death year, manner of death, and age of death. For my project I mainly focused on creating edges based on equality between an entry's country or birth year. Drawing inspiration from Kontothanassis' Lecture 30, I built an Outedge struct and a Graph struct and implemented it with both a create_directed() function and a csv_to_edges() function. From here, I was able to create my directed graph in my main function using my famous.csv file.

I then implemented Dijkstra's well-known algorithm on this Graph struct. For this six degrees of separation problem, considering there are over 50,000 data points, it takes quite a while to run my program. If I could have redone this project, I would choose a lot less data in my csv file to begin with so I know that my algorithm is doing exactly what it should be doing. Overall, I learned a lot throughout this process, particularly how to work with csv files in a different language from python. This is a very basic example of six degrees of separation, but nonetheless it works to show how interconnected data points can be given the right parameters. This principle is extremely useful and important in social media platforms with the "friend of a friend" logic. It is great to learn how to compute the minimum distance between two nodes, as many Data Scientists utilize this skill throughout their careers. Looking forward, I wish to learn more about more efficient graph algorithms, and building other graphs with csv files.