Introduction to Data Science (CSCI 4341)
Data and Distributions

Problem 1. For this exercise, we will use two datasets that are provided with the assignment:

• The file “airport routes.csv” contains the number of available routes of 3409 airports all around the world (as of February 2017). Each row indicates an airport (identified with a 3-letter code) and the number of routes. For example, ”MFE, 6” indicates that McAllen International Airport has outgoing flights to 6 different airports. See data source for more information.

• The file “movie votes.csv” contains the average rating (between 1 and 10) of 4392 movies in TMDb database sorted in descending order. Each row contains a movie name and the average TMDb vote of that movie. For example, "The Godfather", 8.4, "Interstellar",8.1 etc. See data source for more information.

For each of these datasets, consider the following models:

(a) Suppose the given data points follow a power law distribution. Estimate the corresponding α parameter.
(b) Suppose the given data points follow an exponential distribution. Estimate the corresponding λ parameter.
(c) Suppose the given data points follow a uniform distribution. Estimate the corresponding range parameters [a, b] of the uniform distribution.
(d) Suppose the given data points follow a normal distribution. Estimate the corresponding μ and σ parameters.

For each these dataset separately, compare the models you estimated in parts (a) to (d). Which distribution do you think the data follows and why? Explain. For each model, generate random data samples drawn from the respective distribution. Use visualizations of the empirical data and the data you generate to support your conclusions. (You can use Probability Density Function, Cumulative Density Function or QQ Plot)

Problem 2. Analyze the age dataset provided by IMDB Wiki, which contains the birth date of actors and actresses whose photos are present in the IMDB database. The data contains 460718 birth date. Use the date to calculate the age values in years, pick the ones ranging from 1 to 99 years. Investigate the distribution of the age, and also the first and last digits of the ages in this sample (For example, if the data has [34 65], plot the histogram/distribution of these numbers and also histogram/distribution of [3,6] as the first digit and [4 ,5] as last digit separately). Does any of these digits follow a uniform distribution? Is this expected? If one of them does not follow a uniform distribution, what distribution does it follow? Can you explain why?