Exercise 2 – Central Tendency and Similarity Measures

Part 1

In this exercise you are provided with data (Salary Data – Ex2.csv) about salary, education, and other attributes collected from various people for the purposes of a study. As a first step, dealing with data, a data analyst must measure the central tendency of the numerical attributes. Check the file provided on D2L to calculate the following using either Python or R:

1. For **salary**, **education**, and **prestige** please calculate and report the following:
    a. **Minimum value**
    b. **1st quartile**
    c. **Median**
    d. **Mean**
    e. **3rd quartile**
    f. **Maximum value**

2. Plot the **histogram** for *prestige*
3. Plot the **histogram** for *education*
4. Plot a **scatter plot** for *salary* and *education*
5. Plot a **scatter plot** for *education and prestige*
6. Calculate and report the **variance** and the **standard deviation** for *salary.*
7. Explain what does variance measure.
8. What does a high variance value mean?
9. As a data analysts, if you calculate the salary variance and it is 0 (zero). What does that mean? Is it good or bad?
10. **Submit** a MS Word file that contains all the plots and the answers for the questions. Also, submit **the Python or R script** that you used to do the required steps in a .R or .PY file.

Part 2

It is important to define or select similarity measures in data analysis. However, there is no commonly- accepted subjective similarity measure.

Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

One of the methods to calculate the dissimilarity score for nominal attributes is simple matching. The equation of simple matching as follows:

$$d(i,j) = \frac{number\ of\ all\ attributes\ (p) - number\ of\ matching\ attribites\ (m)}{Number\ of\ all\ attributes\ (p)}$$

Examples are provided in D2L slides.

Table 1 shows a set of video games and their nominal attributes. Please **calculate and report the dissimilarity score** between two pairs of data objects

a. d (Witcher 3, Mortal Kombat 11)
b. d (Super Mario Bros., Super Sonic)

Table 1: Games that Rule

Name	Platform	Genre	Publisher
Super Mario Bros.	NES	Adventure	Nintendo
Wii Sports Resort	Wii	Sport	Nintendo
Super Sonic	SEGA	Adventure	Nintendo
Mortal Kombat 11	PlayStation 4	Fighting	NetherRealm Studios
The Witcher 3: Wild Hunt	PlayStation 4	Adventure	CD Projekt