**Exercise 2 – Central Tendency and Similarity Measures**

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

**Answer:**

**For Salary:**

Salary Minimum Value 7.0

Salary 1st Quartile 21.0

Salary Median 42.0

Salary Mean 44.86666666666667

Salary 3rd Quartile 64.0

Salary Maximum value 197.

A black screen with white text

Description automatically generated

**For Education:**

education Minimum Value 7.0

education 1st Quartile 26.0

education Median 45.0

education Mean 55.55555555555556

education 3rd Quartile 84.0

education Maximum value 221.0

**A black screen with white text

Description automatically generated**

**For Prestige:**

prestige Minimum Value 3.0

prestige 1st Quartile 16.0

prestige Median 41.0

prestige Mean 50.68888888888889

prestige 3rd Quartile 81.0

prestige Maximum value 217.0

**A black screen with white text

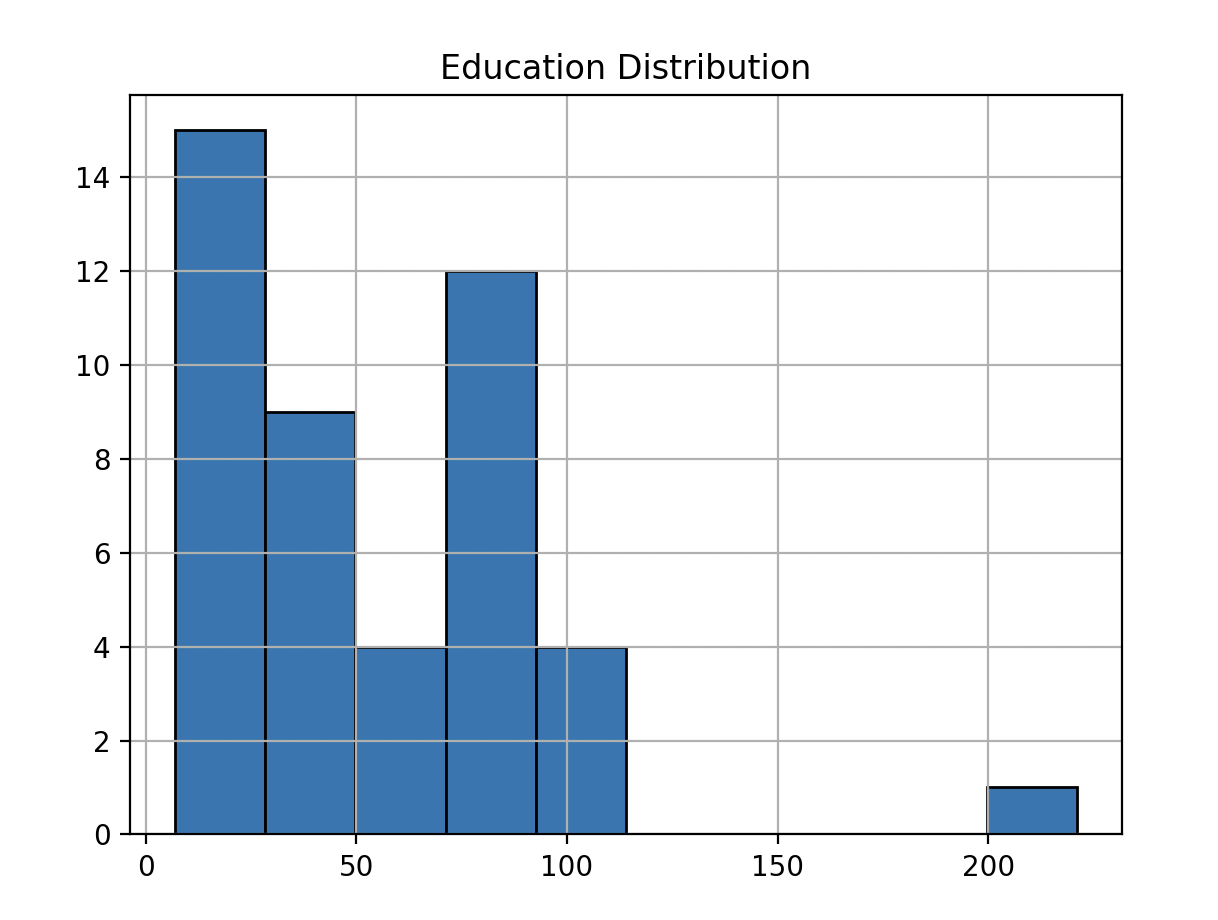
Description automatically generated**

2. Plot the histogram for prestige

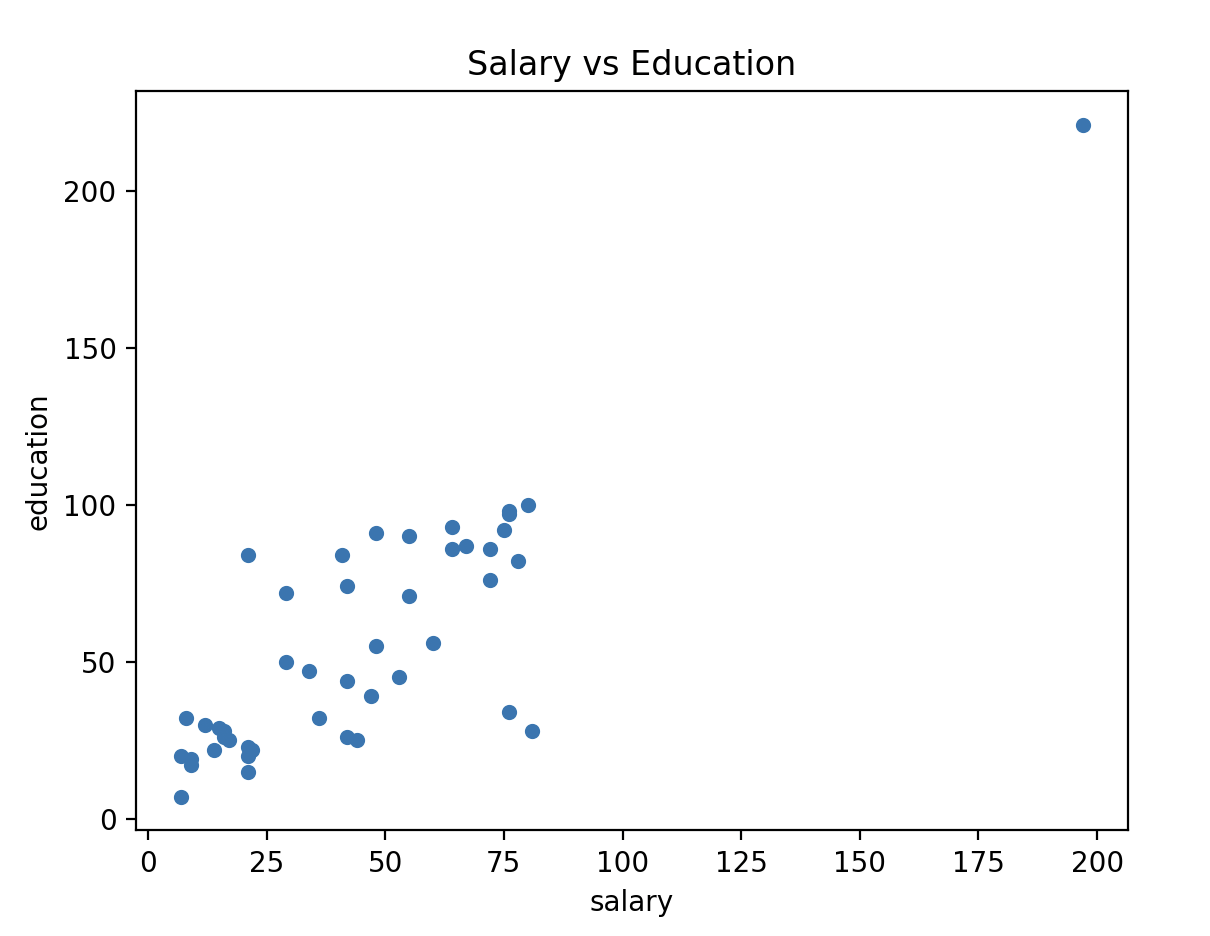
A graph with blue squares

Description automatically generated

3. Plot the histogram for education



4. Plot a scatter plot for salary and education

  
5. Plot a scatter plot for education and prestige

A graph with blue dots

Description automatically generated

6. Calculate and report the variance and the standard deviation for salary.

Salary Variance: 1125.6181818181817

Salary Standard deviation: 33.55023370735563

A black background with white numbers

Description automatically generated

7. Explain what does variance measure.

**Answer**: Variance is a statistical measure that describes the spread or dispersion of a set of data points. Specifically, it quantifies how much the numbers in a data set deviate from the mean (average) of the set. Variance is calculated as the average of the squared differences from the mean.

Variance provides a measure of how much each number in a data point is likely to differ from the mean of the data points. It is equal to the square of the standard deviation and gives a more **exaggerated** **value** as it squares the standard deviation.

8. What does a high variance value mean?

**Answer:** The high variance value indicates that the data points are widely dispersed from the mean. The values in the dataset are more spread out and less consistent. Moreover, it also often implies less predictability and higher uncertainty, as the values significantly deviate from the average.

9. As a data analyst, if you calculate the salary variance and it is 0 (zero).  
What does that mean? Is it good or bad?

**Answer:**

As a data analyst, if the calculated salary variance is 0 (zero), it means that there is no variability in the salary data among the individuals or entities being analyzed. In other words, there is no spread or variation at all; every individual has the same salary.

**Good or Bad?**

In some scenarios, such as wage equality or uniform compensation across a group, a zero variance might be seen as a positive indicator. For example, despite the experiences, performance, and skills, some IT organizations might give the same salary to all the software engineers who are software engineer level 2.

However, in other contexts, it could indicate a lack of diversity or differentiation in compensation, which might be undesirable, particularly in roles where experience, performance, or responsibilities vary significantly.

In summary, variance is a key measure in statistics for understanding the dispersion of data points. A high variance indicates a wide spread in data, and a variance of zero indicates no spread or dispersion at all. The interpretation of these values as good and bad can vary greatly depending on the specific context and the nature of the data being analyzed.

**Part 2**

**A table with text on it

Description automatically generated**

**Calculating dissimilarity:**

1. d (Witcher 3, Mortal Kombat 11)

Solution:

we know that,

𝑑(𝑖, 𝑗) = 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑎𝑙𝑙 𝑎𝑡𝑡𝑟𝑖𝑏𝑢𝑡𝑒𝑠 (𝑝) − 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑚𝑎𝑡𝑐ℎ𝑖𝑛𝑔 𝑎𝑡𝑡𝑟𝑖𝑏u𝑡𝑒𝑠 (𝑚)

𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑎𝑙𝑙 𝑎𝑡𝑡𝑟𝑖𝑏𝑢𝑡𝑒𝑠 (𝑝)

Here,

number of all attributes(p) = 3

Number of matching attributes (m) = 1

d (Witcher 3, Mortal Kombat 11) = (3-1)/3 = 2/3 = 0.66

1. d (Super Mario Bros., Super Sonic)

Solution:

we know that,

𝑑(𝑖, 𝑗) = 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑎𝑙𝑙 𝑎𝑡𝑡𝑟𝑖𝑏𝑢𝑡𝑒𝑠 (𝑝) − 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑚𝑎𝑡𝑐ℎ𝑖𝑛𝑔 𝑎𝑡𝑡𝑟𝑖𝑏u𝑡𝑒𝑠 (𝑚)

𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑎𝑙𝑙 𝑎𝑡𝑡𝑟𝑖𝑏𝑢𝑡𝑒𝑠 (𝑝)

Here,

Number of all attributes(p) = 3

Number of matching attributes (m) = 2

d (Super Mario Bros., Super Sonic) = (3-2)/3 = 1/3 = 0.33