**1.**

[5, 20, 1, 6, 13, 8, 9, 11, 17, 7, 2, 12]

1. **Equal-frequency binning**

**Sorted**:

[1, 2, 5, 6, 7, 8, 9, 11, 12, 13, 17, 20]

**Bins**

Bin1: 1, 2, 5, 6

Bin2: 7, 8, 9, 11

Bin3: 12, 13, 17, 20

1. **Smoothing by bin boundaries**

**Before Bin Boundaries:**

Bin1: 1, 2, 5, 6

Bin2: 7, 8, 9, 11

Bin3: 12, 13, 17, 20

**After bin boundaries:**

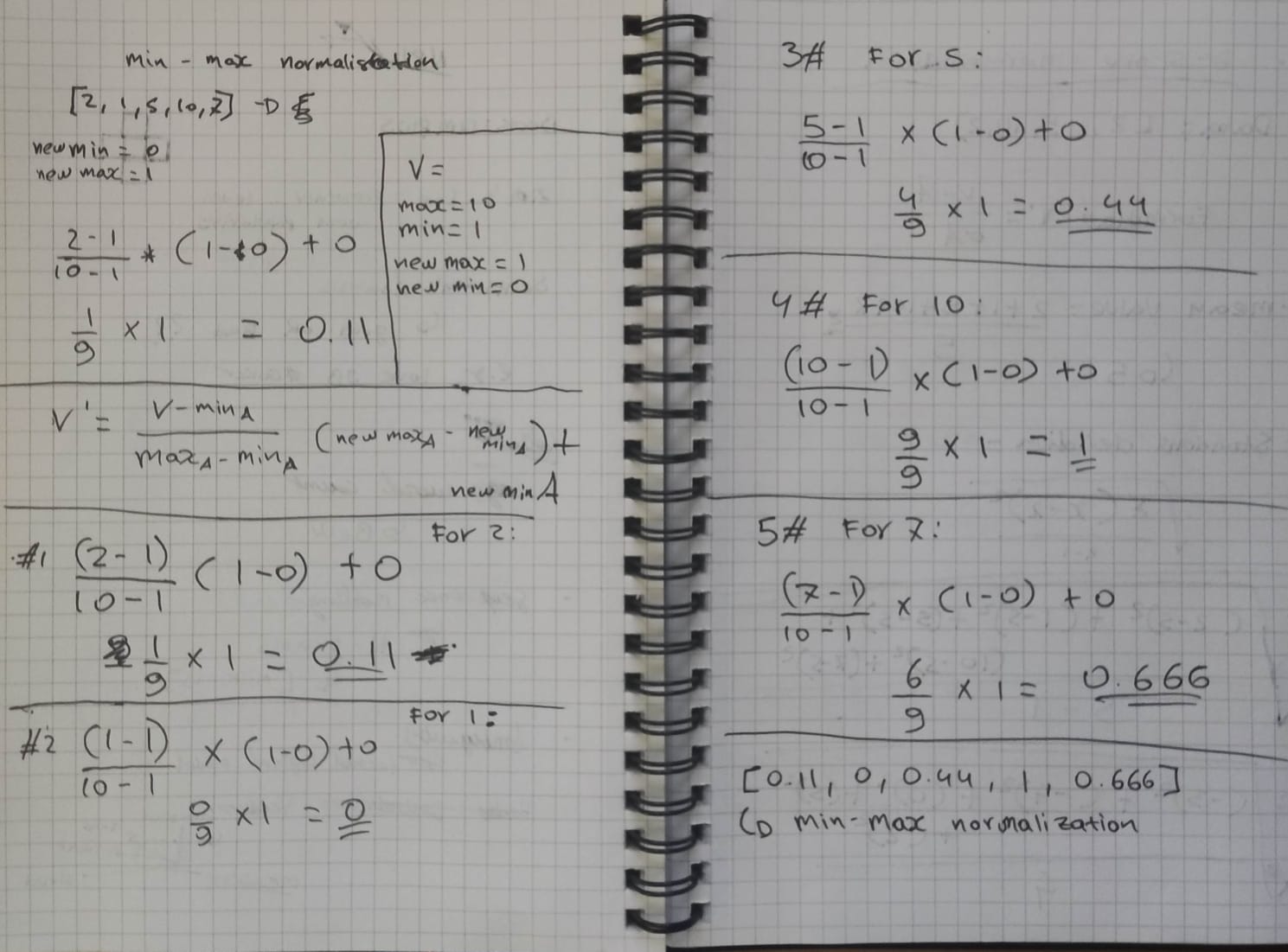
Bin1: 1, 1, 6, 6

Bin2: 7, 7, 7, 11

Bin3: 12, 12, 20, 20

**2- Use the below methods to normalize the following data: [2, 1, 5, 10, 7]:**

1. **min-max normalization with min=0 and max=1.**

**[0.11, 0, 0.44, 1, 0.666]**

1. **z-score normalization**

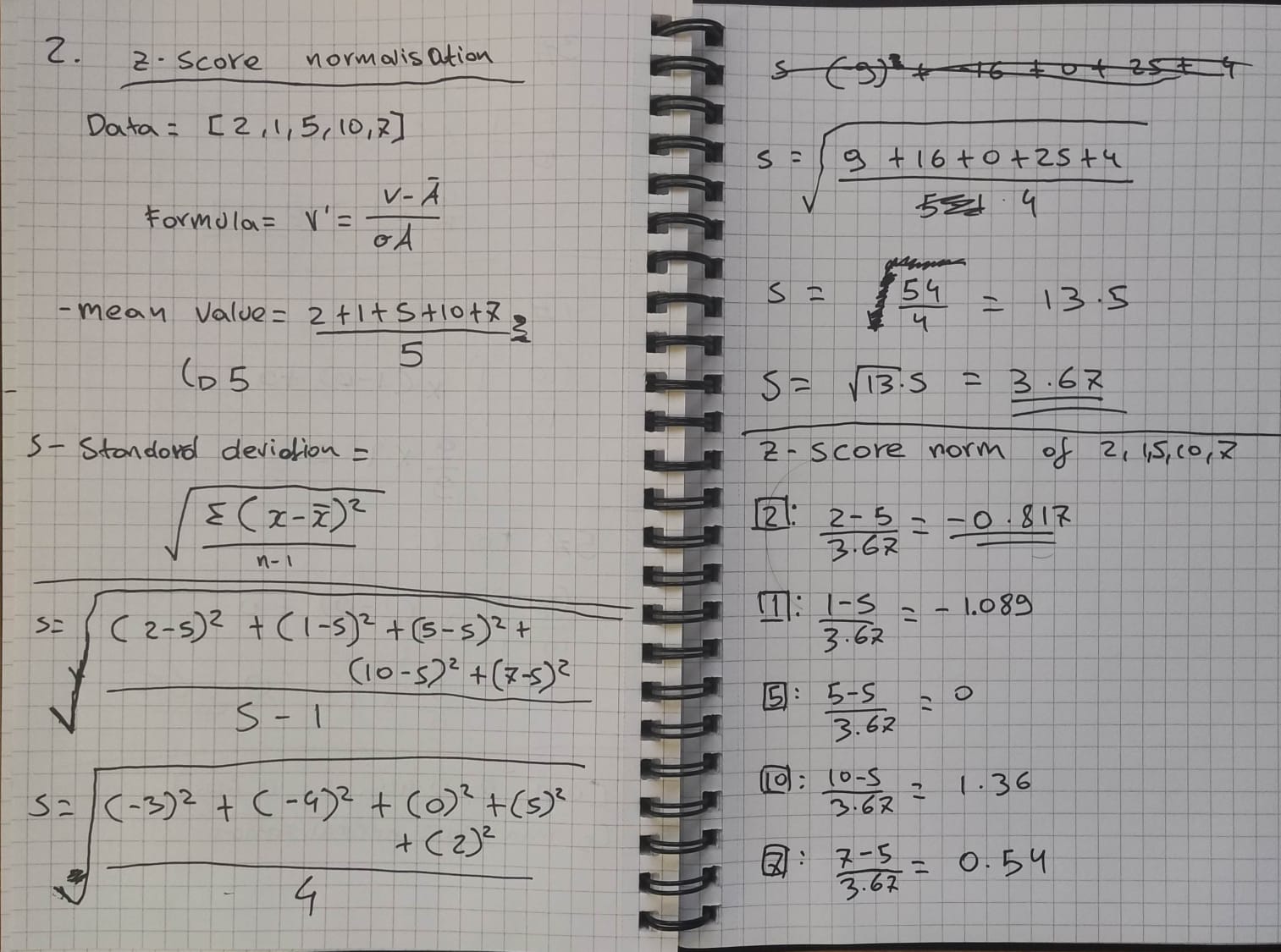
**2 = -0.817**

**1 = -1.089**

**5 = 0**

**10 = 1.36**

**7 = 0.54**

****

1. **Students at two universities, University A and University B, have been provided with feedback forms on student satisfaction, with the below responses recorded. Is student satisfaction correlated with a specific university? Use a chi-square test to find out, assuming a significance level of 0.001 and a corresponding chi-square significance value of 10.828. [1 mark out of 5]**

Table

Description automatically generated

No there isn’t correlation between student satisfaction with a specific university.

1. **Data Cleaning**

Graphical user interface, text, application, Teams

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

1. **Plot scatterplot of shoe size…**

Graphical user interface, application

Description automatically generated with medium confidence

**For Male:**

Application

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

**For Female:**

Chart, scatter chart

Description automatically generated

Scatter plot and Pearson Correlation Coefficient tells us that height of male has more direct relation with shoe size than the height of female.

**Q6.**

Graphical user interface, application

Description automatically generated with medium confidence

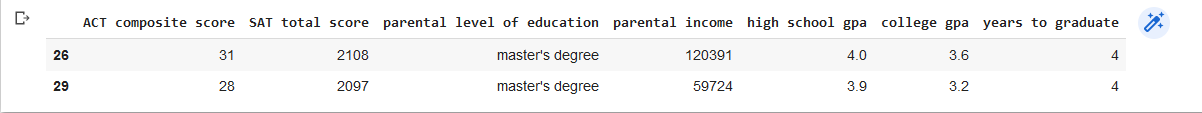
**Assignment 1 [part 2 of 2]**

1. In Section 1, what kind of relationship can be inferred from summary statistics regarding ACT composite score and SAT total score? Which visualisations make this relationship apparent? [0.5 marks out of 5]

As the ACT composite score increases, the SAT total score also increases. Students who have a median SAT score of 2000 will have more than 25+ in ACT composite score. Scatter plots can make the visualisation apparent.

1. Based on the box plots presented in Section 1, what is the relationship between parental level of education and parental income? Using table visualisation, find and show the entire rows that correspond to the outliers regarding parental income whose parents have a master's degree. [0.5 marks out of 5]

The median of the parental level of education increased as along side the parental income. Each level of education had higher income that it’s lower-level counterpart.

****

1. Using an example, explain the importance of scaling features so that their magnitudes are comparable when computing distances. [0.5 marks out of 5]

Scaling features are essential machine learning algorithms for calculating distances between data. As an example, if there is a vast difference in the range of a dataset between 10s and 1000s and it makes assumption of higher values to be superior. It can affect the training of a model as it would start to put these higher values to have significant roles in the model.

As example, someone had a college GPA of 3.7 and the years to graduate was 3 years. Both numbers are similar in range but represent different things but the model as a feature it treats it same. Considering the GPA and the graduation time as features the algorithm would make it, so the GPA has a higher value than the graduation time thus the GPA being more important. However, if we make the graduation time (years) into months it becomes 36 months which is more dominant than the GPA.

For this reason, feature scaling is required to bring every feature in the similar context or equality to have better scalability for the range of data.

1. In Section 1, the distance matrix visualization is not very informative. However, it is still possible to infer that the average distance between students whose parents only have some high school education and students whose parents have a master's degree is larger than the average distance between students whose parents only have some high school education. Explain how this inference is possible from the visualization. [0.5 marks out of 5]

A picture containing graphical user interface

Description automatically generated

From the distance matrix visualization, the distance for some high school education is far left bottom which contrasts with whose parents have a master’s degree which is much higher coming far right top of the matrix. The density in the lower left part is much greater suggesting the average for some high school education. As for the masters the density can be seem much greater at the highest point of 10 at average.

1. In Section 2, increase the number of evenly spaced numbers from 10 to 100 for both axes and observe the corresponding heat map created through nearest neighbor interpolation. Read about this interpolation method and explain what you observed. [0.5 marks out of 5]

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Nearest Neighbor interpolation is one of simplest type of interpolation which requires little calculations allowing it to be one of fastest algorithms. The Nearest Neighbor interpolation determines its pixel based on the closest data value in that position.

From the observation, the distribution of pixel is about equal for both axes. This is because it’s evenly spaced.

1. Wine Data
   1. Load the wine dataset. Compute the frequency of each value of the 'target' feature. [0.5 marks out of 5]

Table

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated