**Higher Applications of Mathematics: R studio Tutorial – Correlation and Predictions**.

For the following you will need the testscores.csv, sportsday.csv and rainforest.csv data sets.

The following questions should be done using the testscore.csv data set

1. Create a histogram for each of the columns of data in testscores.csv (Age, Maths, Physics, and English).

|  |  |
| --- | --- |
|  |  |
|  |  |

2a. Plot a scatter graph of Maths against Age. Do you see any correlation (Type your answer below)

|  |
| --- |
|  |

b. Plot a scatter graph of Maths against English. Do you see any correlation?

|  |
| --- |
|  |

c. Plot a scatter graph of Maths against Physics. Do you see any correlation?

|  |
| --- |
|  |

3. To test for correlation you can use the following

*cor.test(X, Y)* for two different variables

Find which of the following are shown to have a correlation.

a) Age and Maths

|  |
| --- |
|  |

b) Maths and English

|  |
| --- |
|  |

c) English and Physics

|  |
| --- |
|  |

d) Maths and Physics

|  |
| --- |
|  |

4. Create the scatter plot of Maths and Physics again.

a) Use the following function to calculate the linear regression between Maths and Physics

Write as a formula as well.

*lm(Y ~ X)* Calculates the linear regression between two variables.

|  |
| --- |
|  |

b) Add a line of best fit to your diagram.

*abline(lm(Y~X))* Adds a line of best fit to the diagram.

|  |
| --- |
|  |

c) You can predict someone’s results if there is a correlation.

*predict(lm(Y ~ X), newdata=data.frame(X=C),interval = "pred")*

Give an estimation for the following

i) If they had a Physics score of 75 what is their lowest likely Maths Score?

|  |
| --- |
|  |

ii) If they have a Physics score of 10 what is their highest likely Maths score

|  |
| --- |
|  |

The following should be completed using the data set *sportsday.csv*

Remember to remove the testscores data file

detach(Scores)

remove(Scores)

5. Calculate the mean and standard deviation for each of the categories and rank from lowest to highest for mean and standard deviation.

a) 100m (Note: It is X100m in R studio because there is a number in the beginning of the variable)

|  |
| --- |
|  |

b) High Jump

|  |
| --- |
|  |

c) Hurdles

|  |
| --- |
|  |

d) Swim25m

|  |
| --- |
|  |

6. Is the data for those who swam 25m normally distributed? Show this graphically and by doing a test.

*shapiro.test(X,Y)*

7. Create a scatterplot of100m and Hurdles.

a) What is the linear equation between these? Write it in the form y = mx +c.

|  |
| --- |
|  |

b) Add a line of best fit to your equation.

|  |
| --- |
|  |

c) Make a prediction for if they completed the hurdles in 48seconds how long would it take them to do the 100m?

|  |
| --- |
|  |

8. Using 100m and Swim create a linear model as you did in Question 7.

Write the formula and predict if they swam 25 in 31 seconds how long would it take for them to run 100m?

|  |
| --- |
|  |

9. Why will this not work for a linear model of hurdles and High Jump?

|  |
| --- |
|  |

For the following use the rainforest.csv file (remember to detach and remove the sports day file)

10. Show that the number of trees in the data are not normally distributed.

|  |
| --- |
|  |

11. Create a scatterplot to show the number of Trees in a Rainforest compared to the population of humans.

|  |
| --- |
|  |

12. Create a linear regression model for this and write down the formula.

|  |
| --- |
|  |

13. Predict how many Trees would be in the rainforest if there were 2500 people living in it.

|  |
| --- |
|  |