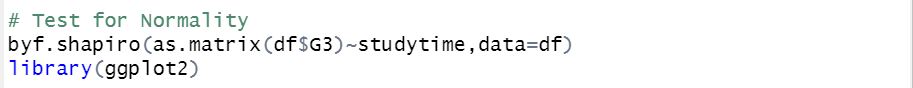
1. The variables taken into consideration are as follows:

G3 is the dependent variable and continuous.

studytime is the independent variable and categorical.

First, we perform the test for normality,



Text

Description automatically generated

To further support the numbers, we plot a histogram to confirm the not normal nature of the variables.

Chart, histogram

Description automatically generated

We can clearly see from the numbers and the chart that the p-values are less that 0.05 and hence the variables are not normal, and we decide to go with the non-parametric tests.

Hence, we perform, the Kruskal Wallis Ranked Test for the hypothesis testing.

Graphical user interface, text, application, email

Description automatically generated

Since, the p-value is 0.5557 which is greater than 0.05, we fail to reject the null hypothesis. Hence, there is no significant evidence to prove the relationship between G3 and studytime.

1. The variables taken into consideration are as follows:

G1 is the dependent variable and continuous.

absences is the independent variable and continuous.

First, we perform the test for normality,



Text

Description automatically generated

We can clearly see that the p-values for both the variables is less than 0.05. Hence, both G1 and absences are not normal and hence we proceed with non-parametric tests.

Since, the variables are not normal, we will be employing Spearman’s correlation test to obtain the relationship between G1 and absences.

Text

Description automatically generated

Text

Description automatically generated

From the above results, the p-value is 0.0293 which is greater than 0.05. So we fail to reject the null hypothesis. Hence, we can conclude that there is no significant evidence to prove the existence of relationship between G1 and absences.

1. The question asks to show and explain the relationship between multiple variables and G2. The best way to do this is to plot a correlation plot and analyze the visualization.

Text, letter

Description automatically generated

A picture containing chart

Description automatically generated

It is clear from the correlation plot that the given range of variables does not have a strong positive correlation with the G2 variable. Although it is interesting to note that gout which shows the going out aspect of the student has a negative correlation with G2. This means, the more the student goes out, lesser the grades are.

1. The audience of this visualization is the government, the healthcare bodies, and the governing bodies of schools and colleges.

Chart, bar chart, histogram

Description automatically generated

The above visualization is a grouped bar plot that has the age of the students on x-axis and the color signifies whether the student is in a romantic relationship or not. From the visualization, we can clearly see that there is an upward trend for students coming in relationship after the age of 15. This is the application insight that we get from the visualization.

The application insights make the audience realize that majority of the students get into relationship between the age of 15 and 18. This is the teenage phase of the lives of the students. This is important for the audience to realize as many teenagers are not ready for a relationship, and they still get into one and do not have idea of how to handle them. The schools and colleges can organize sex educations sessions, counselling sessions, and other workshops in order to teach the students about relationships. This can result in less psychological issues that is observed in teenagers due to unwanted complications.

1. Polynomial Regression is a regression algorithm that models the relationship between a dependent and independent variable as an nth degree polynomial, which means the relationship is non-linear and the value of dependent variable is related to the independent variable raised to a certain power. The equation for Polynomial regression is as follows:

y = b0 + b1x + b2x2 + b2x3 +... + bnxn .

Polynomial regression is also known as the special case of Multiple Linear Regression in Machine Learning. This is because some polynomial terms need to be added to the Multiple Linear regression equation to convert it into the equation for Polynomial Regression. It is a linear model with some modifications to increase its accuracy. It uses a linear regression model to fit the non-linear datasets. Hence, in Polynomial regression, the original values are converted into Polynomial values of required degree (2, 3, …, n) and then modelled using a linear model. Let’s take an example of a data where we get a low accuracy while using linear regression. This indicates that the data is polynomial in nature. In this case, we can see that the straight line is unable to capture the patterns in the data. This is an example of under-fitting. To overcome under-fitting, we need to increase the degree of the polynomial equation.

An interesting use of polynomial regression can be seen is Global Climate Models or GCM. GCM is used to simulate climate models and then study the effects of global warming and climate change. Polynomial regression is employed in climate modelling because the data used for climate model is very random and does not follow a pattern. Hence, to fit this random and uncertain pattern within the data, polynomial regression seems to be the perfect fit.

Shen, M., Chen, J., Zhuan, M., Chen, H., Xu, C.-Y., & Xiong, L. (2018). Estimating uncertainty and its temporal variation related to global climate models in quantifying climate change impacts on hydrology. *Journal of Hydrology*, *556*, 10–24. https://doi.org/10.1016/j.jhydrol.2017.11.004