**Secondary School Performance: Math Class**

Brashon Ford

National University

**Secondary School Performance: Math Class**

Hello my name is Brashon and I am the lead data scientist, our client is wanting use to analyze the leading cause of student absences in a math class. We are given the data that the number of student absences will range from 0 to 93. Our client hypothesizes that the reason that students are missing their math class is because they are a romantic relationship. With this information we will be using an explanatory model to test this research question. When we get the basic definition of Statistics and Data Generating Process, DGP, it all sums up to the rules we must follow when we want to find out the process that can cause data to occur as they do. To put DGP into a story we would say when students enter romantic relationships, they start to accumulate more absences especially in math class. Even though the prediction is that relationship status could be a factor in absences, but there could be variation in the outcome. Students who aren’t in a relationship could also be having the same number of absences as the students that are.

Our research will be used to help our client explain why they are seeing a high number of absences with students that are in a romantic relationship. We want to study this data so that we can provide feedback on why and how the client can approach this situation. Taking it from an approach of agreeing with the hypothesis I think that romantic relationships would explain variation in absences because students become too distracted. They could be missing class because they skip, or just not being prepared for class due to bad time management. In a word equation it would look like: Absences = Romantic Relationship + Error.

**Data Description**

Now we can get started with diving deeper into this data and find out if our hypothesis is correct. With this dataset there are 395 observations of 33 variables, and we want to start with finding the frequencies of the categorical demographic variables. This type of variable is described as a data grouped into categories instead of being measure numerically. The categorical data that I used include: age, sex, and romantic status; to find our frequency of these variables we can use our data to see how many students answered yes and no plus the age range. First we can review our age category, looking at the data we can see that the ages range from 15-22 years old. Students that are 16 had the highest frequency with 104 participants; even though we do have participants that are between 19-22 I believe we can just regard those as outliers. Next our sex, or gender, frequency was separated into two factors: female and male. We found that out of 395 observations, 208 of them are female and 187 are male. Now we can dive into the data to see how many participants are in romantic relationships. We found that 263 participants said that they are not in a relationship, while 132 students said they are in a relationship. To provide information on our quantative demographic variables we used a five number summary. Using a five number summary will give us the lower and upper quartiles, the maximum and minimum, and the median. First, we found the five number summary for absences, and we gathered that: our minimum and first quartile are 0, the median is 4, mean is 5.7, the third quartile is 8, and the maximum of absences is 75. Let’s investigate some descriptive statistics for our outcome and explanatory variable. An outcome variable is what we are trying to explain, while an explanatory variable is what we are manipulating or observing changes in; in the case of what we are studying the explanatory variable is relationship status and the outcome variable is absences. Analyzing the descriptive statistics, we have a lot of options to use, we’ve gotten a hint into that statistics with finding the five number summary, but we can use some more. Variance is a type of descriptive statistics that can show the measurement of spread between numbers in a data set. The variance in our explanatory variable is going to be 64 and we can only use variance for numeric values. To do some more digging into the data, I also wanted to find the out how much time the participants studied. I was able to find out that they are studying for a minimum of 1 hour to a maximum of 4 hours, the average study time came out to be 2 hours. Let’s show data visualization to show our absences and if romantic relationships are the cause of that. To show this visualization I have constructed a histogram to show the absences; a histogram is a graph that shows the frequency of numerical data using rectangles. As we look at the first bar in Figure.1 we can see that it has the highest number of participants that had absences between 0-15 days. The shape of our histogram does not look like it is normal or bell-shaped, so we can conclude that it’s skewed to the right. The center of our histogram is right around the mean, 5.7 .

Figure.1

A graph of a bar graph

Description automatically generated

**Research Question: Absences explained by Romantic**

The first task in analyzing our research question more in detail; I used an empty model which will only give us the intercept. By using the empty model, we can first off get the intercept, that value is 5.709. The equation that I used to get the empty model in our programming software looks like this: lm(absences ~ NULL , data = data.frame), and that is how the intercept was generated. Once we move away from the empty model, or baseline, we can add in romantic value we find out that 2.599 are in romantic relationships. With reviewing our empty model and the numbers we got, we will call them our parameter estimates. These estimates are called b0 and b1; b0 equals 4.840 which is close to our mean of absences, and b1 equals 2.599. Once we added our explanatory variable to the model, romantic, it no longer was an empty model. In our model equation, we will have Xi as the explanatory variable which is romantic relationships and that will generate an Explanatory Model. Now we can fit our estimates into a model equation: Absences = 4.840 + 2.599Xi + ei. In word form we can say that the absences equal the intercept of 4.840 plus 2.599, relationship status, and add in the error.

**Comparing The Two Models**

We have gotten our important data, and now we want to visualize a table that shows us different outcomes.

**Table.1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **SS** | **df** | **MS** | **F** | **PRE** | **p** |
| **Model (error reduced)** | **593.711** | **1** | **593.711** | **9.469** | **.0235** | **.0022** |
| **Error (from model)** | **24641.808** | **393** | **62.702** |  |  |  |
| **Total (empty model)** | **25235.519** | **394** | **64.050** |  |  |  |

Let’s discuss Table.1 in more detail, we see that there are sections on the side and at the top with data that matches with each model. On the left corner of Table.1 we can observe that there are three sections: Model (Error reduced), Error (from model), and Total(empty model). The Error (from model) row tells us how much error is left over after fitting the romantic relationship into our analysis. Looking at Model (error reduced) and error(from model), we see that the values actually total and give us the Total(empty model). Our first variable that I want to discuss is the F ratio. This ration is defined as the amount of error reduced by a model that adjusts for the number of parameters it takes to realize what the reduction in error, it is also considered a sample statistic. To calculate the F Ratio, we would use the MS Model divided by the MS Error. In our data the F Ratio is 9.469, the larger the F ratio the more effect it has on our model. Let’s discuss the df, or degrees of freedom, that is in our Table.1; degrees of freedomis an estimate for the number of independent pieces, n-1. In our Table.1 we see that the df for Model (error reduced) equals 1, while the df for Error ( from model) equals 393. The third column is labeled Mean Square (MS), this value is calculated by dividing the SS Total by the degree of freedom.