Student Scores

Exploratory Analysis with R

Morgan Hardin, mhardin5@bellarmine.edu
Zoe Mecklenburg, zmecklenburg@bellarmine.edu
Gavin Trumbull, gtrumbull@bellarmine.edu

1. INTRODUCTION

The dataset chosen for the analysis is called students_scores and it shows how students scored in various subjects. The dataset also contains information on how many hours the students studied and what careers they want to enter. Here is a link to the dataset: https://www.kaggle.com/datasets/mexwell/student-scores The dataset was found on Kaggle and was chosen because it contained a mix of numerical and categorical variables. The scores are the dependent variables for our comparisons. The goal is to see if all the scores correlate with one another and if the study hours and absent days impact student scores. Along with this, another goal is to see if these scores mirror a student's career aspirations. This means that if a student wants to go into a specific field that requires science or math, their scores should reflect their interests.

2. DATA SET DESCRIPTION

The student scores dataset consists of 1 file with 17 columns and 2001 rows. Within the dataset, there are no null entries in any of the rows or columns. There are 10 numerical columns, 5 categorical columns, and 2 boolean columns. For all of the columns, 0% of the data is missing.

The columns 'id', 'absence_days', 'weekly_self_study_hours', 'math_score', 'history_score', 'physics_score', 'chemistry_score', 'biology_score', 'english_score', and 'geography_score' are all numerical data types and integers. They are all ratio data since there is an absolute zero. The column 'id' is a continuous variable while the other numerical variables are all discrete since they are whole number values that are specific numbers that are counted.

The columns 'first_name', 'last_name', 'email', 'gender', and 'career_aspiration' are all categorical data types and strings. They are all nominal variables since there is no order.

The columns 'part_time_job' and 'extracurricular_activities' are logical values and boolean. Their values are true and false or 1 and 0.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
id	Numerical (int)	0%
first_name	Categorical (string)	0%
last_name	Categorical (string)	0%
email	Categorical (string)	0%
gender	Categorical (string)	0%
part_time_job	Logical (Boolean)	0%
absence_days	Numerical (int)	0%
extracurricular_activities	Logical (Boolean)	0%
weekly_self_study_hours	Numerical (int)	0%
career_aspiration	Categorical (string)	0%
math_score	Numerical (int)	0%
history_score	Numerical (int)	0%
physics_score	Numerical (int)	0%
chemistry_score	Numerical (int)	0%
biology_score	Numerical (int)	0%
english_score	Numerical (int)	0%
geography_score	Numerical (int)	0%

3. **DATA SET SUMMARY STATISTICS**

The student scores dataset has 2,001 rows and 17 columns, each with no null values. There was good distribution across all the columns. For the logical columns 'part_time_job' and 'extracurricular_activities', the values were skewed and there were more false values than true. There was a decent range in the scores with 40 being the lowest and 100 being the highest. The statistics for the categorical values were not helpful, but did show that there were no missing values, and they were distributed evenly. The correlation table showed that all of the scores correlated with one another positively, although it was extremely weak. It also showed that 'weekly_self_study_hours' positively correlated with all the scores. This means that the higher the hours studied for any of the classes, the higher the scores will be in that class. The table also showed that 'absence_days' had a negative impact on all the scores. This means that the more days of class missed, the lower the scores will be.

Table 2: Summary Statistics for Student Scores Dataset – Numerical Columns

Variable Name	*Min*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max*
Id	1	500.8	1000.5	1000.5	1500.2	2000
Absence Days	0	2	3	3.666	5	10
Weekly Self Study Hours	0	5	18	17.76	28	50
Math Score	40	77	87	83.45	93	100
History Score	50	69.75	82	80.33	91	100
Physics Score	50	71	83	81.34	92	100
Chemistry Score	50	69	81	80	91	100
Biology Score	30	69	81	79.58	91	100
English Score	50	72	83	81.28	91	99
Geography Score	60	71	81	80.89	91	100

Table 3: Summary Statistics for Student Scores Dataset – Categorical Columns

Variable Name	*Length*	*Class*	*Mode*
First Name	2000	Character	Character
Last Name	2000	Character	Character
Email	2000	Character	Character
Gender	2000	Character	Character
Career Aspiration	2000	Character	Character

Table 4: Summary Statistics for Student Scores Dataset – Logical Columns

Variable Name	*Mode*	*False*	*True*
Part Time Job	Logical	1684	316
Extracurricular Activities	Logical	1592	408

Table 4: Correlation Table/Tables Part 1

	Id	Absence Days	Weekly Self Study Hours	Math Score	History Score
Id	1.000000000	-0.017787761	0.008524265	-0.01404200	0.02195925
Absence Days	-0.017787761	1.000000000	-0.286085623	-0.23707179	-0.12815879
Weekly Self Study Hours	0.008524265	-0.286085623	1.000000000	0.39356930	0.27623076
Math Score	-0.014041999	-0.237071786	0.393569298	1.000000000	0.14724747
History Score	0.021959252	-0.128158794	0.276230761	0.14724747	1.000000000
Physics Score	-0.002002633	-0.136419207	0.202119849	0.11571874	0.04847843
Chemistry Score	0.004725661	-0.084028985	0.201340203	0.12713149	0.12149786
Biology Score	0.045436647	-0.090553768	0.190480823	0.08129806	0.08850197

English Score	-0.001907527	-0.084861109	0.247796015	0.13483078	0.14719288
Geography Score	-0.005969168	-0.002941598	0.153622443	0.04967233	0.06575135

Table 4: Correlation Table/Tables Part 2

	Physics Score	Chemistry Score	Biology Score	English Score	Geography Score
Id	-0.002002633	0.004725661	0.04543665	-0.001907527	-0.005969168
Absence Days	-0.136419207	-0.084028985	-0.09055377	-0.084861109	-0.002941598
Weekly Self Study Hours	0.202119849	0.201340203	0.19048082	0.247796015	0.153622443
Math Score	0.115718743	0.127131491	0.08129806	0.134830783	0.049672327
History Score	0.048478427	0.121497862	0.08850197	0.147192883	0.065751351
Physics Score	1.000000000	0.126162795	0.13227971	0.054313698	0.103125921
Chemistry Score	0.126162795	1.000000000	0.11999168	0.068340579	0.065430082
Biology Score	0.132279708	0.119991683	1.000000000	0.074226910	0.106525909
English Score	0.054313698	0.068340579	0.07422691	1.000000000	0.072249792
Geography Score	0.103125921	0.065430082	0.10652591	0.072249792	1.000000000

4. DATA SET GRAPHICAL EXPLORATION

Through exploring the dataset, it was found that time spent studying generally improved student scores. It was also found that gender and extracurriculars had almost no impact on the student's test scores, so the focus was on how people did on each subject relative to what profession they aspired to be. For this, the goal was to choose four different careers that students said they wanted to do. The chosen career aspirations were Accounting, Game Development, Teaching, and Writing. These four were chosen because of the belief that their results would be significantly different. They were also chosen because of the different areas of focus. For accounting, the main area is expected to be math. For game development, the main areas expected are math and physics. For teaching, all subjects are expected to be in the middle range. Finally, for writing, the main area is expected to be English. The goal is to determine if our hypothesis was true and what areas and scores indicate a student's interest and their chosen career aspirations.

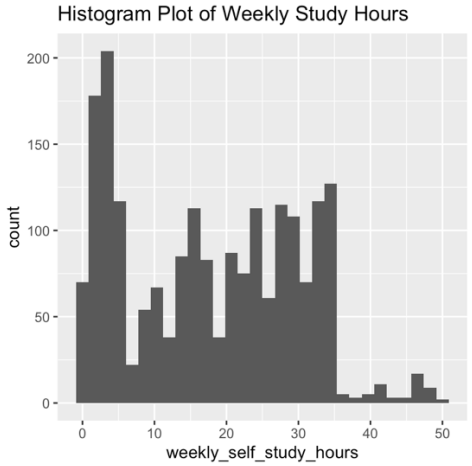

Figure 1: Histogram Plot of Weekly Study Hours / Count from Student Scores dataset (single plot)

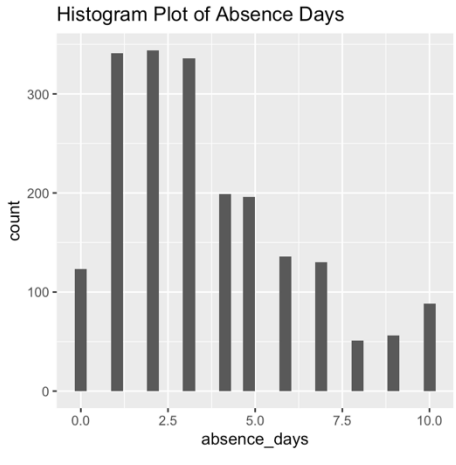

Figure 2: Histogram Plot of Absence Days / Count from Student Scores dataset (single plot)

Figures 1 and 2 Analysis: These histograms show the distribution of the weekly study hours and the absence days. Both are right skewed, meaning most of their data lies on the left side of the graph and slopes downward to the right side. This shows that most of the study hours and absence days fall in a smaller range from 0 to 35 study hours and 0 to 5 absence days. These plots show that many of the students do not study more than 35 hours a week and miss less than 5 days of school.

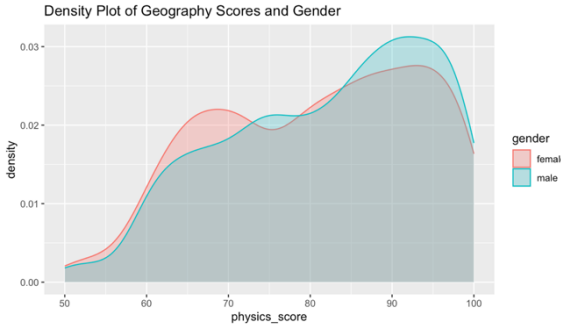

Figure 3: Density Plot of Physics Score / Density with Gender from Student Scores dataset (single plot)

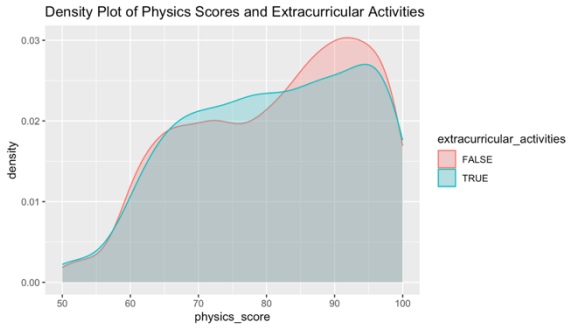

Figure 4: Density Plot of Physics Score / Density with Extracurricular Activities from Student Scores dataset (single plot)

Figures 3 and 4 Analysis: These density plots show the physics score alongside both gender and extracurricular activities. Figure 3 shows that females typically do worse in physics than males. This can be seen through the females peaking over the males at around 68% while the males almost consistently stay above the females from 68% and above and peak at around 93%. This goes hand in hand with Figure 4 because the students who do not participate in extracurricular activities typically score higher than students who are in extracurricular activities. This could mean they have more time to study and dedicate to physics than those who participate in other activities. This shows that there is a correlation between extracurricular activities and physics scores because if a person does not participate in any activities, then they are likely to score higher in physics.

Figure 5: Scatter Plot of History Score / Weekly Study Hours from Student Scores dataset (single plot)

Figure 5 Analysis: This scatter plot depicts the history score compared to the weekly study hours. There are clear sections that show the more a student studies during the week, the higher their history score will be. From the score range 50% to 60%, no students studied over 5 hours. For the score range 60% to 85%, no students studied over 35 hours. For the score range 85% to 100%, students studied up to 50 hours. This shows that there is a positive correlation between weekly study hours and history scores. This also proves the correlation table also saying that there was a positive correlation between the two areas.

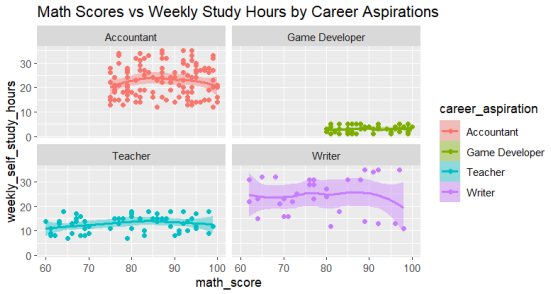

Figure 6: Facet Grid of Math Score / Weekly Study Hours by Career Aspirations from Student Scores dataset (multiple plots)

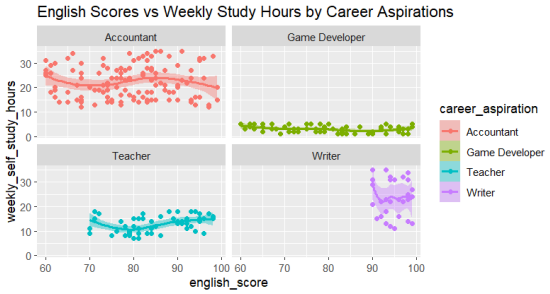

Figure 7: Facet Grid of English Score / Weekly Study Hours Career Aspirations from Student Scores dataset (multiple plots)

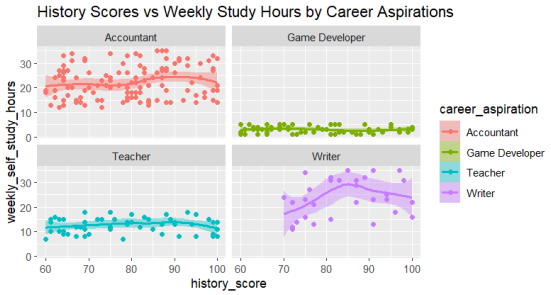

Figure 8: Facet Grid of History Score / Weekly Study Hours by Career Aspirations from Student Scores dataset (multiple plots)

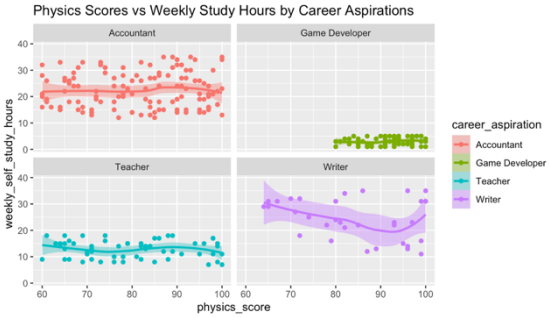

Figure 9: Facet Grid of Physics Score / Weekly Study Hours Career Aspirations from Student Scores dataset (multiple plots)

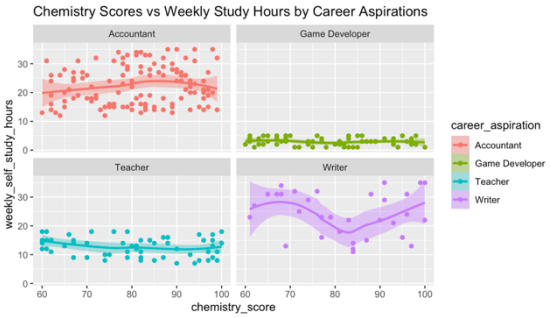

Figure 10: Facet Grid of Chemistry Score / Weekly Study Hours by Career Aspirations from Student Scores dataset (multiple plots)

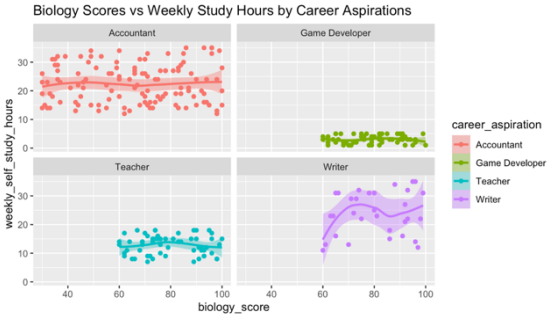

Figure 11: Facet Grid of Biology Score / Weekly Study Hours Career Aspirations from Student Scores dataset (multiple plots)

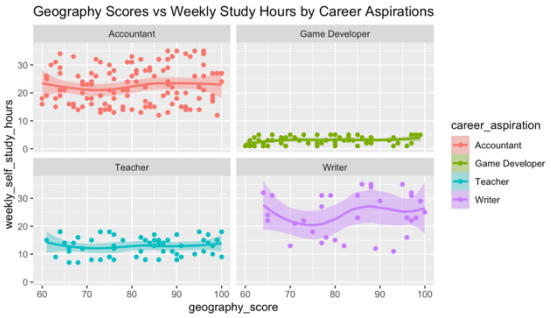

Figure 12: Facet Grid of Geography Score / Weekly Study Hours by Career Aspirations from Student Scores dataset (multiple plots)

Figures 6 to 12 Analysis: These facet grids compare each score to the weekly study hours based on the four career aspirations accountant, game developer, teacher, and writer. The original hypothesis was that each of these careers would have higher scores if the career is centered around a specific area. The belief was that accountants should have high math scores, game developers should have high physics and math scores, teachers should have decent scores for every area, and writers should have high english scores. Along with this and from the previous plots, if the scores are high, the weekly study hours for the higher scores should also be high. If the scores are low, the study hours should also be low. It was discovered that the hypothesis about accountants was correct. They scored and studied high in math but did not score high consistently in the other subjects. For game developers, it was found that they did not study much for any subject, but also scored high in math, physics, and biology. The hypothesis was somewhat correct that math and physics would be high, but biology was also high and they did not follow the pattern of studying more to get higher scores. Game developers did not study hardly at all for any subject. For teachers, our hypothesis was correct and they scored decently and studied decently for every subject. They tended to do better in biology and english which was an interesting find. Finally, for writers, the hypothesis was correct and the english scores were high, as well as biology and history which was a shocking find. Writers also tended to study more in each subject, even if the scores were not consistent. Overall, the majority of the hypothesis was correct, but there were a few outliers that were found.

5. SUMMARY

Overall, the original hypothesis was mostly correct. The findings showed that there was a positive correlation between weekly study hours and every subject score as well as a negative correlation between absence days and every subject score. This means that the more a student studies during the week, the higher their scores will be, and the more absence days they have, the lower their scores will be. With this being said and most of the findings supporting this hypothesis, there were a few outliers. Game developers tended to not study at all for every subject but still scored well in most subjects. The career aspirations also lined up well with the scores and areas of focus. Students with a career aspiration of being an accountant tended to study more during the week and had higher scores in math. This supports our original hypothesis of the aspirations and scores being positively correlated.

Being involved in extracurricular activities seemed to negatively impact scores as well. The plot of physics scores showed that being involved in extracurriculars led to lower physics scores. With this, gender also played a small role with males scoring higher in physics than females. This also could show that females tend to be involved in more extracurricular activities than males. The scores also were weakly, but positively correlated with one another. The original hypothesis was expecting more of a correlation between similar subjects, but the correlation table showed that there was a weak, positive correlation between each area. Overall, the findings show that the more hours a student studied, the higher they would score in each class. Students also tended to score higher in areas that correspond to their career aspirations. Missing days of school and extracurricular activities negatively impacted student scores in every subject. These findings show that our hypothesis was correct.

6. REFERENCES

Mexwell. (2024, March 5). *Student Scores.* Kaggle. https://www.kaggle.com/datasets/mexwell/student-scores