STAT.614.01

Course Project

Abstract

To get hands on experience analysing student performance dataset using techniques learned in this course

Megha Gupta

mg9428@rit.edu

Contents

1. Introduction:

1.1. Description of the dataset: This dataset contains the score of the students about math, reading and writing. Along with this, the data includes different categorical variables of the students such as ('gender', 'race/ethnicity', 'parental level of education', 'lunch', 'test preparation course').

StudentPerformace.csv

1.2. Objective: To determine the most significant factors involved in affecting the scores of the students and to explore if some of the urban myths ('gender', 'race/ethnicity', 'parental level of education', 'lunch', 'test preparation course') has statistical influence or not.

1.3. Predictor Variables: Test preparation of the course, parental level of education, gender, race, lunch.

1.4. Response Variables: math score, reading score, writing score.

2. Statistical Tests Used: 2.1. Multi-factor ANOVA
2.2. Linear regression

2.1. Multi-factor ANOVA
Factor A: Gender

Levels for factor A: 2 (Male/Female)

Factor B: Race/ethnicity

Levels for factor B: 5 (A, B, C, D, E)

Factor C: Parental level of education

Levels for factor C: 6 (associate's degree/ bachelor's degree/ high school/ master's degree/ some college/some high school)

Factor D: Lunch

Levels for factor D: 2 (standard/ free-reduced)

Factor E: Test preparation course

Levels for factor E: 2 (none/completed)

2.1.1. Analysis for Math score: From the fig. 2 we can conclude that since the p-values for gender, race/ethnicity, parental level of education, lunch and test preparation course is small. We can conclude that all the factors affect the math score of students on the test.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	62	63371.80	1022.13	5.7585
Error	937	166317.28	177.50	**Prob > F**
C. Total	999	229689.08		0.000000

Fig. 1

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
gender	1	1	4884.382	27.5177	<.0001*
race/ethnicity	4	4	6000.365	8.4512	<.0001*
parental level of education	5	5	5365.201	6.0453	<.0001*
lunch	1	1	16425.931	92.5406	<.0001*
test preparation course	1	1	5935.268	33.4382	<.0001*
gender*race/ethnicity	4	4	212.358	0.2991	0.8786
gender*parental level of education	5	5	1268.127	1.4289	0.2112
gender*lunch	1	1	333.485	1.8788	0.1708
gender*test preparation course	1	1	13.270	0.0748	0.7846
race/ethnicity*parental level of education	20	20	1333.759	0.3757	0.9944
race/ethnicity*lunch	4	4	340.744	0.4799	0.7505
race/ethnicity*test preparation course	4	4	481.645	0.6784	0.6070
parental level of education*lunch	5	5	788.610	0.8886	0.4880
parental level of education*test preparation course	5	5	323.503	0.3645	0.8730
lunch*test preparation course	1	1	4.650	0.0262	0.8715

Fig. 2

2.1.2. Practical Interpretation: From the interaction profiles we can conclude the following for the math score.

1. Male performed better compared to females
2. Group D students performed better.
3. Students with parental level of education with Master's degree performed better.
4. Students with standard lunch performed better.
5. Students who completed their test preparation before the taking the test performed better.

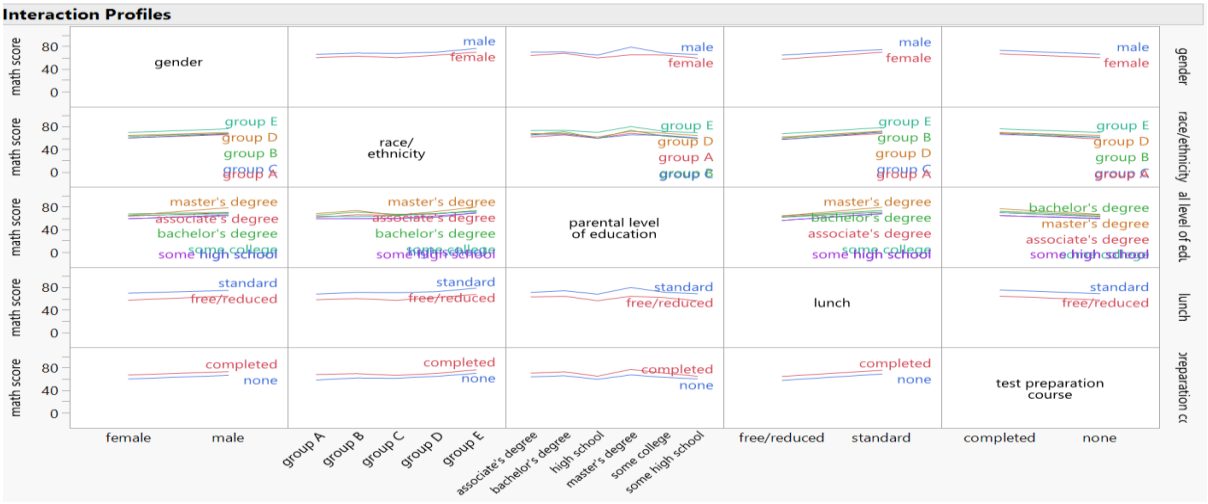

Fig. 3

2.1.3. Analysis for reading score: From the fig. 5, we can conclude that since the p-values for gender, race/ethnicity, parental level of education, lunch and test preparation course is small. We can conclude that all the factors affect the reading score of students on the test.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	62	53918.41	869.652	5.1238
Error	937	159034.03	169.727	**Prob > F**
C. Total	999	212952.44		0.000000

Fig. 4

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
gender	1	1	4881.8492	28.7630	<.0001*
race/ethnicity	4	4	1704.3193	2.5104	0.0404*
parental level of education	5	5	6485.8749	7.6427	<.0001*
lunch	1	1	6584.5885	38.7952	<.0001*
test preparation course	1	1	8389.6650	49.4304	<.0001*
gender*race/ethnicity	4	4	319.5168	0.4706	0.7573
gender*parental level of education	5	5	980.8597	1.1558	0.3292
gender*lunch	1	1	187.2351	1.1032	0.2938
gender*test preparation course	1	1	0.0086	0.0001	0.9943
race/ethnicity*parental level of education	20	20	1876.1783	0.5527	0.9437
race/ethnicity*lunch	4	4	350.1013	0.5157	0.7242
race/ethnicity*test preparation course	4	4	644.4186	0.9492	0.4347
parental level of education*lunch	5	5	814.6108	0.9599	0.4414
parental level of education*test preparation course	5	5	730.0189	0.8602	0.5073
lunch*test preparation course	1	1	13.8248	0.0815	0.7754

Fig. 5

2.1.4. Practical Interpretation: From the interaction profiles fig. 6 we can conclude the following for the reading score.

1. Female performed better compared to females
2. Group E students performed better.
3. Students with parental level of education with Master's degree performed better.
4. Students with standard lunch performed better.
5. Students who completed their test preparation before the taking the test performed better.

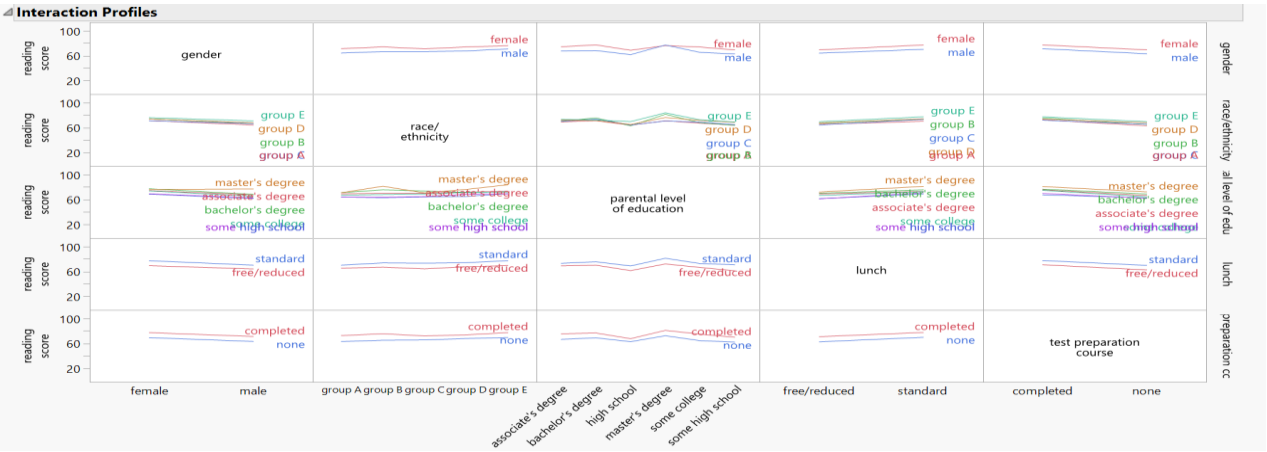

Fig. 6

2.1.5. Analysis for writing score: From the fig. 8, we can conclude that since the p-values for gender, race/ethnicity, parental level of education, lunch and test preparation course is small. We can conclude that all the factors affect the writing score of students on the test.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	62	82026.58	1323.01	8.3394
Error	937	148650.50	158.65	Prob > F
C. Total	999	230677.08		0.000000

Fig. 7

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
gender	1	1	7844.635	49.4477	<.0001*
race/ethnicity	4	4	2575.184	4.0581	0.0029*
parental level of education	5	5	9994.831	12.6002	<.0001*
lunch	1	1	8252.008	52.0155	<.0001*
test preparation course	1	1	15526.235	97.8677	<.0001*
gender*race/ethnicity	4	4	171.939	0.2709	0.8968
gender*parental level of education	5	5	1236.619	1.5590	0.1691
gender*lunch	1	1	342.339	2.1579	0.1422
gender*test preparation course	1	1	14.966	0.0943	0.7588
race/ethnicity*parental level of education	20	20	1596.808	0.5033	0.9662
race/ethnicity*lunch	4	4	337.303	0.5315	0.7126
race/ethnicity*test preparation course	4	4	288.693	0.4549	0.7688
parental level of education*lunch	5	5	873.084	1.1007	0.3584
parental level of education*test preparation course	5	5	317.944	0.4008	0.8484
lunch*test preparation course	1	1	112.058	0.7063	0.4009

Fig. 8

2.1.6. Practical Interpretation: From the interaction profiles fig. 9, we can conclude the following for the writing score.

1. Female performed better compared to females
2. Group E students performed better.
3. Students with parental level of education with bachelor's degree performed better.
4. Students with standard lunch performed better.
5. Students who completed their test preparation before the taking the test performed better.

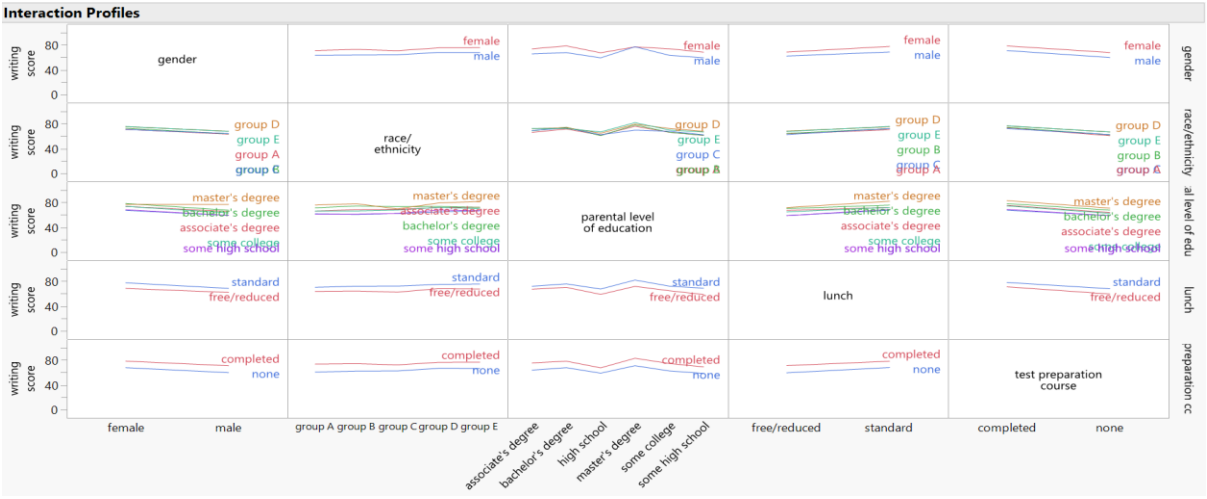

Fig. 9

2.2. Linear Regression: To find if a student who performs good at reading is likely to perform good at reading as well.

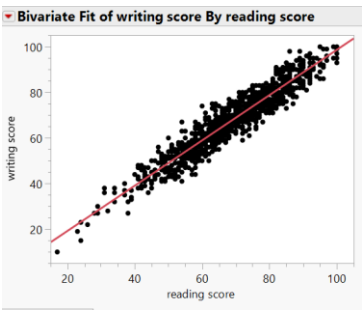

Fig. 10

The fitted regression equation is

writing score = -0.667554 + 0.9935311*reading score

$b_0 = -0.667554$

$b_1 = 0.9935311$

Hypothesis Test in Simple Linear Regression

Hypothesis: H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Test statistic: $t_0 = \frac{0.9935311}{0.009814} = 101.23$

p-value: 2. P ($t_{998} > 101.23$) < 0.001

2.2.1. Conclusion: To reject H_0.
There is strong evidence to support the claim that the regression model reading score is statistically significant to the writing score. Thus, a student who performs good at reading is likely to do well in writing as well.

2.3. Recommendation: Thus, from the above statistical analysis, we can say that standard lunch and test preparation for the test are the two factors which can be standardized to improve the student's score. Also, the statistical test gives us the evidence that reading and writing are statistically significant to each other. A student who performs good in reading is likely to score good in the writing section as well.