Emmanuel Pedernal MSDS

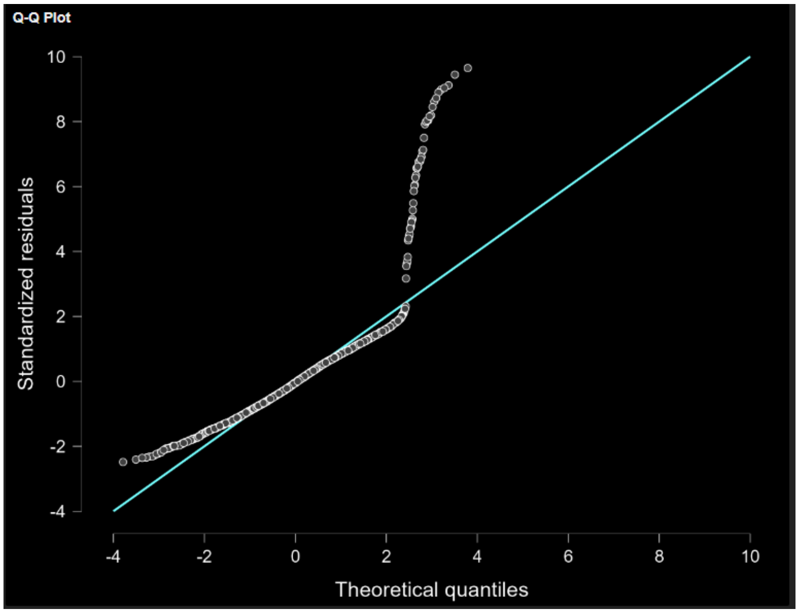
ANCOVA: Analyze the attached data set. Test the hypothesis that the students' performance (score) is affected by the income (low, medium, high) while accounting for the hours spent in studying.

Assumption check

Independence of observation (PASSED)

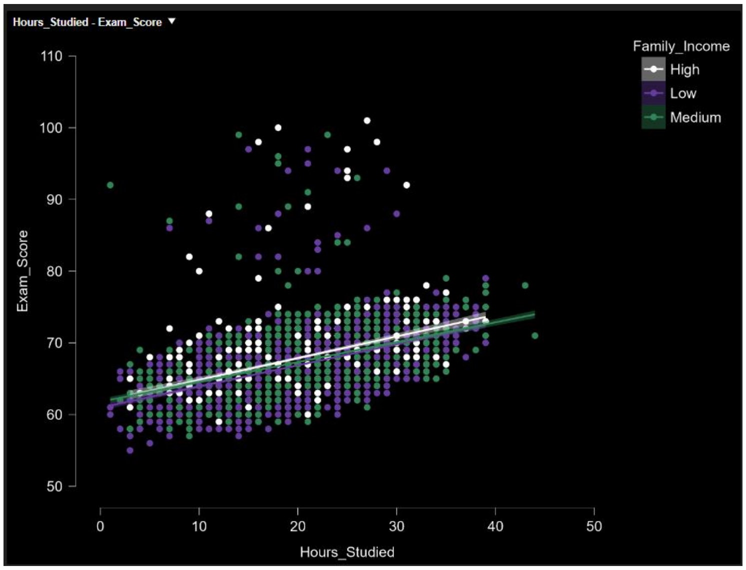
Unique observations in dataset

Normality of Residuals (FAILED) none linear residuals

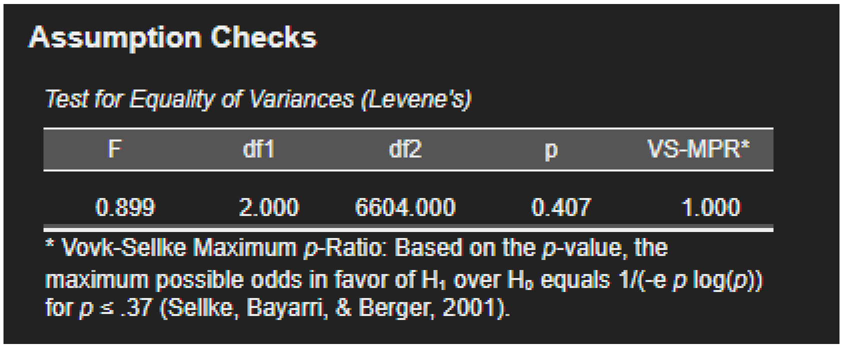


Linearity between Covariate and Dependent Variable (Passed)

No pattern just linear



Homogeneity of Variance (Homoscedasticity) (PASSED) p val > 0.05



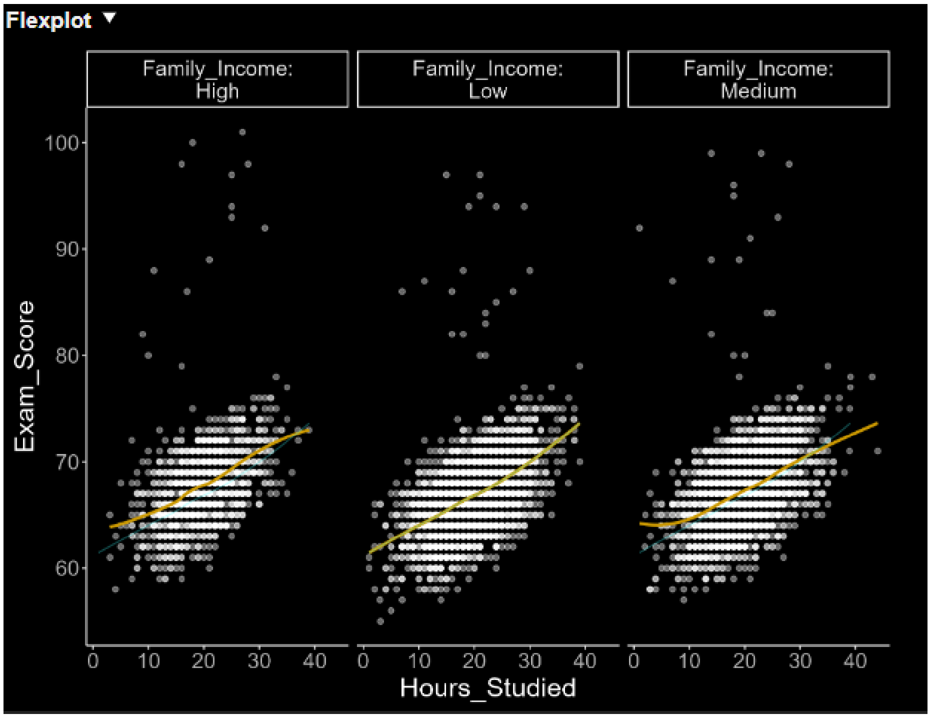
Assumption Checks

Test for Equality of Variances (Levene's)

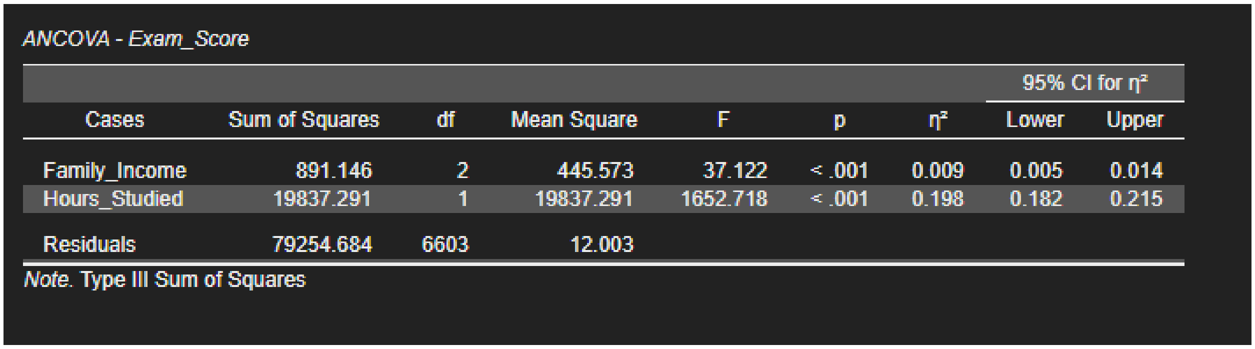
F	df1	df2	p	VS-MPR*
0.899	2.000	6604.000	0.407	1.000

* Vovk-Sellke Maximum *p*-Ratio: Based on the *p*-value, the maximum possible odds in favor of H_1 over H_0 equals $1/(-e\ p\ \log(p))$ for $p \le .37$ (Sellke, Bayarri, & Berger, 2001).

Homogeneity of Regression Slopes (PASSED) same linear plots for each class



Insights



ANCOVA - Exam_Score

							95% CI for η²	
Cases	Sum of Squares	df	Mean Square	F	p	η²	Lower	Upper
Family_Income	891.146	2	445.573	37.122	< .001	0.009	0.005	0.014
Hours_Studied	19837.291	1	19837.291	1652.718	< .001	0.198	0.182	0.215
Residuals	79254.684	6603	12.003					

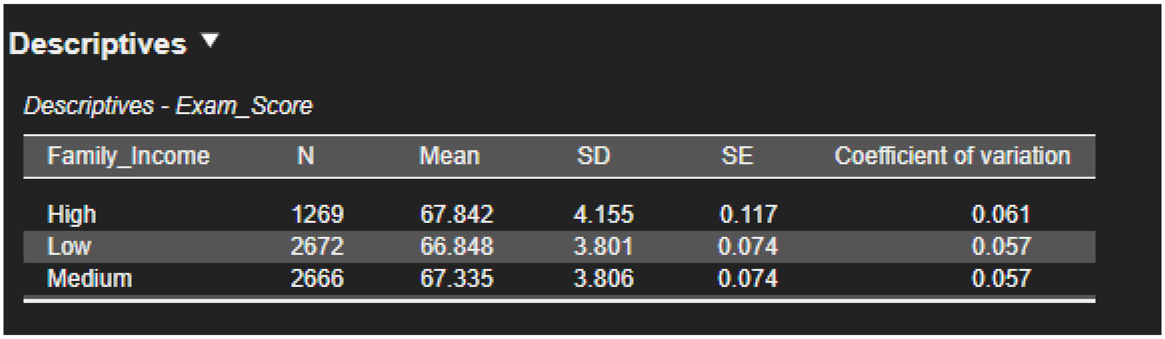
Note. Type III Sum of Squares

RESULTS

Stat	Family_Income	Hours_studied
Sum of Square	891.146 score difference (variation) because of income class	With a high (19,837.291) sum of sqrs, Hours_studied variable has the high influence in scores
df	3 income classes	1 single predictor (linear effect)

Mean Square	mean square error of 445.573 or ~21(root), is how much the avg exam changes due to income class	Amount of variance on exam scores per study hour
F	37.122, huge f statistic, shows clear difference between income classes	(1652.718) Based from the data, student who study more has higher score
p-value	Income is statistically significant in affecting scores	Also, a significant feature
Eta^2	.9% (small) of score score variability is due to income	Hours_studied variable explains ~20% of score variability
C.I.	Small confidence on income effect on scores	95% confident that portion of score variance by the data lies between 0.182 - 0.215

Residuals showed large sum of squares indicating that the model can explain the variance, a huge amount of variation in exam scores are unexplained mainly due to other factors not included in the model



Descriptives ▼

Descriptives - Exam_Score

Family_Income	N	Mean	SD	SE	Coefficient of variation
High	1269	67.842	4.155	0.117	0.061
Low	2672	66.848	3.801	0.074	0.057
Medium	2666	67.335	3.806	0.074	0.057

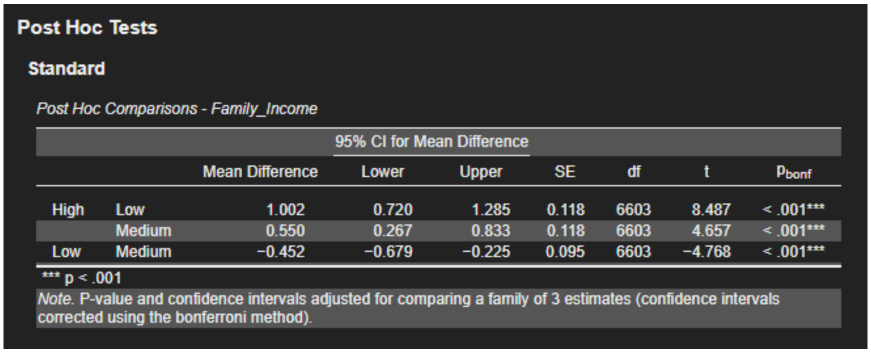
N is the amount of data for each class

Mean – average exam scores, a bit of upward trend. Raw scores (small) increase as income increase.

Standard Deviation – Scores have low SD, exam scores are somewhat consistent for this dataset

Stadard Error – small errors (more accurate estimate) for bigger N than small N

CoV – Income classes are consistent (similar) exam score



Post Hoc Tests

Standard

Post Hoc Comparisons - Family_Income

			95% CI for Mean Difference					
		Mean Difference	Lower	Upper	SE	df	t	p_{bonf}
High	Low	1.002	0.720	1.285	0.118	6603	8.487	< .001***
	Medium	0.550	0.267	0.833	0.118	6603	4.657	< .001***
Low	Medium	−0.452	−0.679	−0.225	0.095	6603	−4.768	< .001***

*** p < .001

Note. P-value and confidence intervals adjusted for comparing a family of 3 estimates (confidence intervals corrected using the bonferroni method).

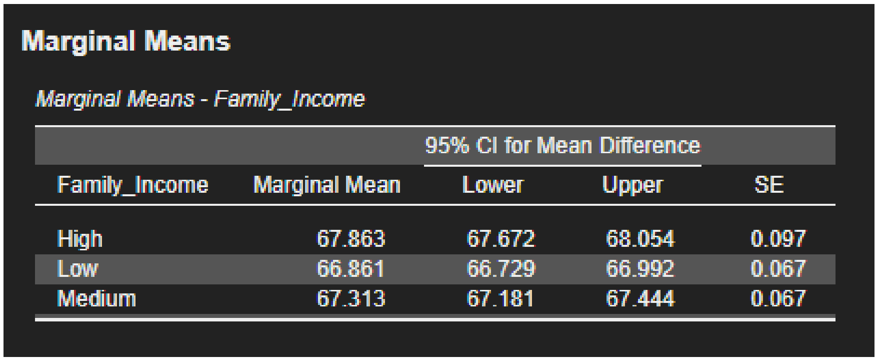
On average High vs low scores has a 1.002 difference assuming same hours of studying. High vs med has 0.55 points difference and low vs med has -0.452 score difference all differences are statistically significant.

95% C.I. states that for High vs Low the true score difference are between 0.720 and 1.285, 0.267 – 0.8333 for High to med, Low vs Medium of -0.679 to -0.225 high difference among classes.

Df – same number of independent information

High vs low has the highest observed difference with 8.5 standard errors from 0

Bonferroni p val – all comparisons are significant, all evidence points that exam_scores are different for each income class



Marginal Means

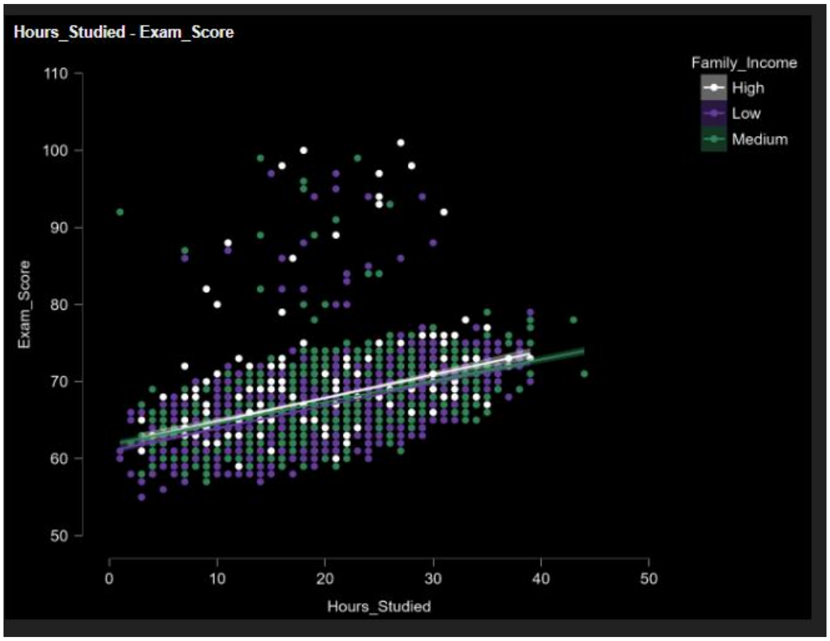
Marginal Means - Family_Income

		95% CI for Mean Difference		
Family_Income	Marginal Mean	Lower	Upper	SE
High	67.863	67.672	68.054	0.097
Low	66.861	66.729	66.992	0.067
Medium	67.313	67.181	67.444	0.067

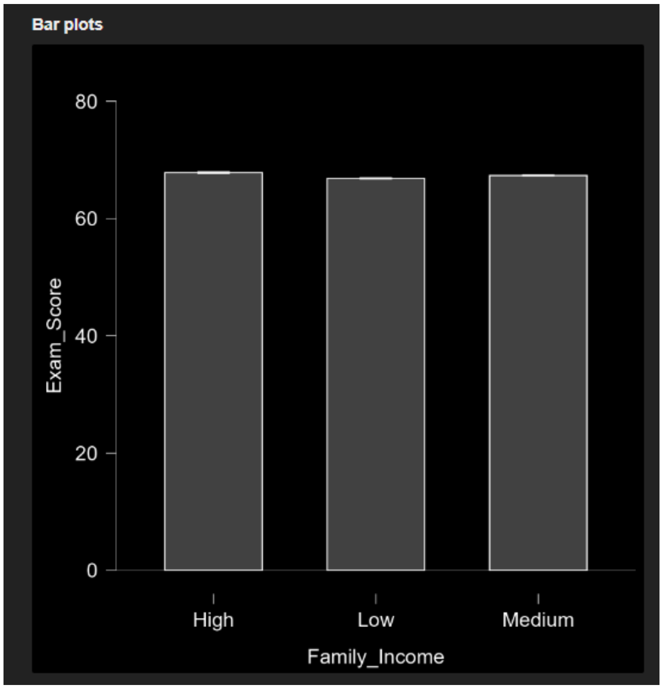
Marginal means shows adjusted mean, but has same results that high income families tend to do better (even with small difference of 1.002 vs low and 0.550 vs medium)

ANSWER: Family income has statistically significant effect on exam scores but those scores does not have huge difference among income classes, while time spent on studying has greater impact on exam scores.

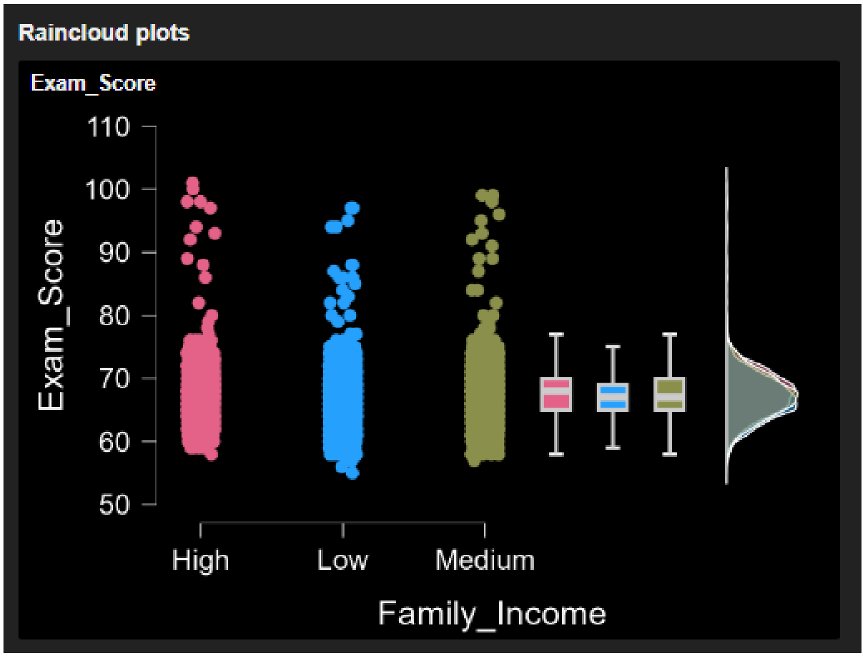
PLOTS



The scatter graph shows exam score based on hours studied separated by income classes, based on the graph hours studied affect the exam score and income as little effect on it.



Bar plots showed little difference with exam scores with family income as confirmed by our Ancova analysis.



Raincloud plot shows the data distribution for each class where we have a lot of outliers, most scores are within 60 to 75, and scores are mostly the same with small difference for each class