Kindergarten Children Score Analysis

Chenyang Pan
1005131554

1. Introduction

In recent years, the importance of education has been paid much attention not only in elementary school and high school but also starting from kindergarten. The general knowledge test in kindergarten is one of the best ways to assess the current learning status of young children. The reading test is another important assessment to evaluate the conceptual understanding of children. By analyzing the score of the general knowledge test and reading test, we can acquire an understanding of how well the young children interact with the external world and how well their conceptual understanding is. However, since each child grows from a different family background, there can be factors affecting their general knowledge and reading ability. The central goal of this analysis is to take into consideration the factor income of families and to see whether it is significant in affecting the students' reading ability and general knowledge.

In order to provide additional help with children's reading ability and general knowledge understanding for children who are weak in these areas, we focus on the factor of family income and want to know how different income group family children may have different abilities.

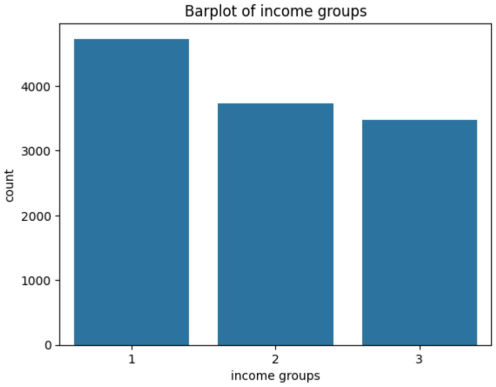
To explore this question, we will perform an in depth analysis based on child longitudinal study data.

2. Data cleaning

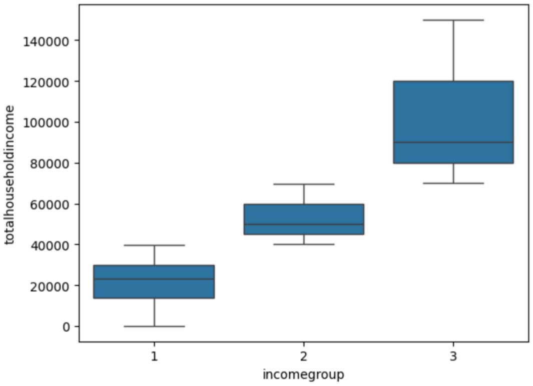
We first had a preview of the children scores dataset. There are 9 features in total with 8 numerical variables and 1 categorical variable. Since our analysis will majorly focus on the general knowledge and reading score, we will keep only the relevant features which are the 1998 fall and 1999 spring general knowledge and reading scores (numerical), the income groups(categorical) and the income amount(numerical) of each student. Then, we drop the NA terms in the data and performed exploratory data analysis (EDA)

3. Exploratory Data Analysis

As the first step of EDA, we first took a look at the distribution of students in different income groups, both numerically, showing the number of them in each income group, and visually, using barplot. In terms of numerical observation, income group 1 has the highest number of students of 4729 and group 3 has lowest of 3478. Shown in the barplot on the right. We can see that in general, the distribution of three groups is constant.



Then, I also made a side by side boxplot of the total incomes of families by income group. The plot is shown on the right. I can acquire an important



information from this box plot which is that the income group 1, 2 and 3 are ordered from lowest to highest, which means that income group 1 indicates the families with lowest income and group 3 indicate the families with highest income.

Next, I then conducted two summary statistics presented in the form of a table.

The table below describes the summary statistics of general knowledge scores of students in fall and spring grouped by the income groups. We can see that the mean and median of the general knowledge test score for both fall and spring terms shows to be from lowest to highest as the income group from low income to high income. This implies that the family income may be a factor affecting test scores. The spread, standard deviation and IQR seems similar across all groups. One thing to notice is that in all three groups, the students' spring score is higher than fall score.

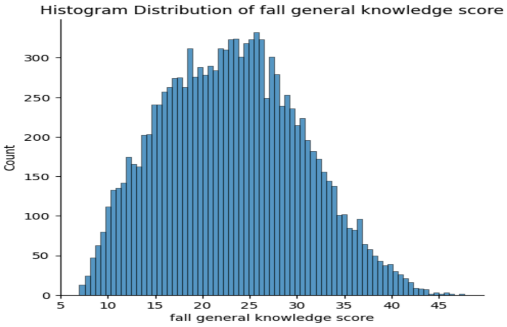
	Min	Mean	Max	Spread	Median	Standard deviation	IQr
Fall / group 1	6.985	19.948	43.508	36.523	19.298	6.717	9.803
Fall / group 2	7.115	23.888	46.12	39.0	23.869	6.87	9.783
Fall / group 3	7.504	26.452	47.691	40.19	26.505	7.102	9.761
Spring /group 1	8.124	25.069	45.581	37.457	25.065	7.248	10.447
Spring group 2	7.858	29.144	48.055	40.197	29.476	6.965	9.623
Spring group 3	9.513	31.568	48.345	38.832	32.082	6.928	9.514

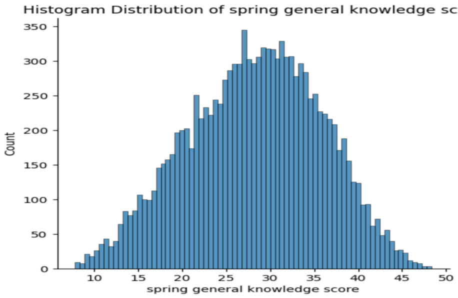
The next table, shown below describes the summary statistics of reading scores of students in fall and spring grouped by the income groups. Overall, the summary statistics of the reading test score is similar to the general knowledge scores, we observe similar patterns that test score for both fall and spring terms shows to be from lowest to highest as the income group from low income to high income, which gives us similar implication. However, the standard deviation and IQR across groups are quite different

	Min	Mean	Max	Spread	Median	Standard deviation	IQr
Fall / group 1	21.01	32.787	118.29	97.28	31.66	8.092	8.95
Fall / group 2	22.19	36.293	138.51	116.32	34.53	9.994	9.728

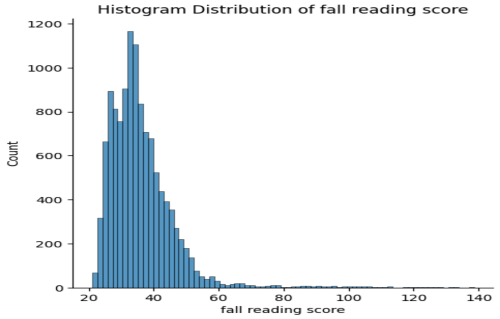
Fall / group 3	23.01	39.898	133.56	110.55	37.575	12.289	11.45
Spring /group 1	22.35	43.665	142.49	120.14	41.97	12.005	12.43
Spring group 2	23.93	48.009	142.49	118.56	46.065	13.505	12.296
Spring group 3	24.54	52.207	156.85	132.31	48.74	16.45	13.458

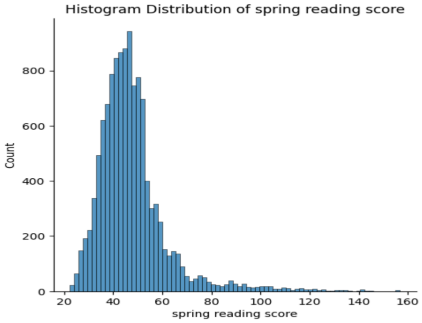
Then, the histogram distribution of fall and spring general scores shown between also provide us some comparison and insight. The histogram of fall general knowledge score on left below and spring general knowledge on the right below are both symmetric and unimodal. Observing carefully, we can find that the spring score histogram slightly shifts to the right compared to the fall scores. This tells us that the students are improving as time goes.





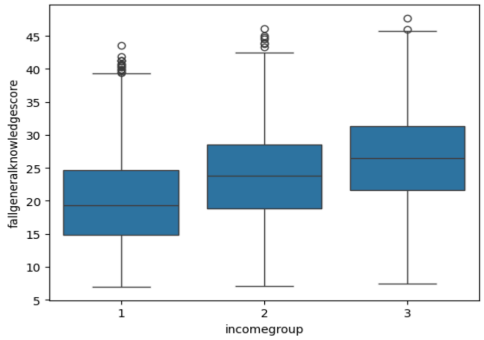
Next, the next two histograms are fall and spring reading scores shown below. We can observe a similar comparison between fall and spring reading scores as the general knowledge. One thing that is different is that the two histograms of reading score are both a bit right skewed but overall symmetric .

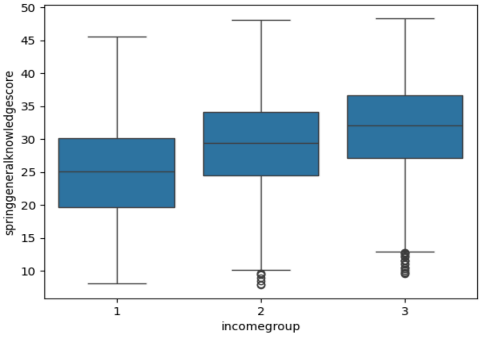


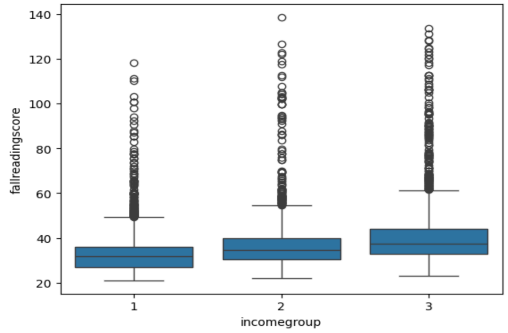


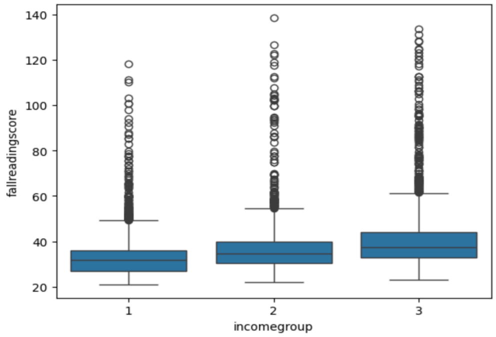
Four side- by-side box plots are also conducted to compare the fall and spring general knowledge score by income groups and fall and spring reading score by income groups shown below. The top left and top right box plots are fall and spring general knowledge. We can see visually that the IQR is similar across all groups and the median is increasing as the income group goes from lowest to highest. Also there are outliers on the right end of the whisker of the fall boxplots and on the left end of the spring boxplots. The bottom left and bottom right box plots are fall and spring reading scores by income groups.

We can observe similar observations as the former. One thing different is that there are lots of extreme values on the right end of the whisker in both fall and spring boxplots.

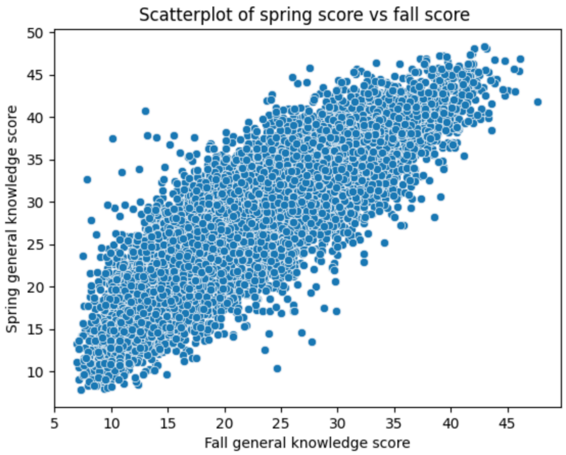


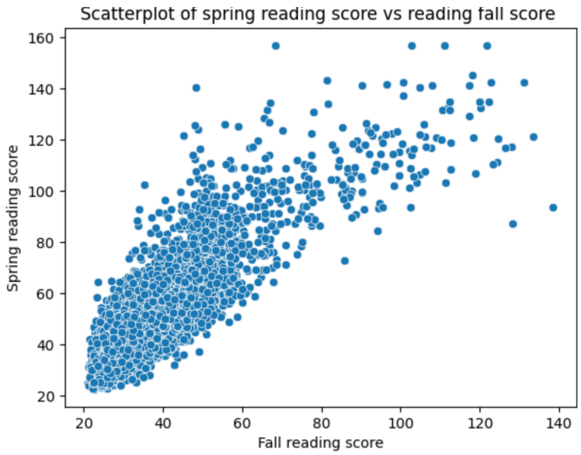






Lastly, two scatter plots below are made to observe the correlation between fall general knowledge score and spring general knowledge score, and also fall reading score and spring reading score. On the scatterplot on the left we can find a positive linear correlation between fall and spring general knowledge score and the scatterplot between fall and spring reading score also shows a positive correlation but weaker. This correlation implies that the fall scores may be a covariate that needs to be controlled when we want to detect the effect of income groups on spring scores.





4.One Way ANCOVAs

4.1 General Knowledge Test Score

Null hypothesis: the average score of the spring general knowledge test is the same across different income groups aftering taking control of the fall general test score as covariate .

Alternative hypothesis: the average score of the spring general knowledge test is not the same across different income groups aftering taking control of the fall general test score as covariate .

4.1.1 Model Result

The result of the model is presented in the following tables. There are some findings and conclusions can be made from the results.

R squared	Adjusted R squared	F-statistic	Prob (F-stat)	Log-likelihood	AIC	BIC
0.731	0.731	1.621e+04	< 0.001	-33263	6.653e+04	6.656e+04

From the above table, we can see that the R square is 0.731 indicating about 73.1% or variability is explained and the adjusted R square is similar to the R square indicating the proper fit of the model. The p value is smaller than 0.001 which is smaller than the significance level of 0.05. This means that we have strong evidence to reject the null hypothesis that states the average score of the spring general knowledge test is the same across different income groups aftering taking control of the fall general knowledge test score as covariate.

	Coefficients	Standard error	T statistic	Pvalue
Intercept	7.5991	0.126	60.237	< 0.001
income group	0.4803	0.047	10.222	< 0.001
Fall general knowledges score	0.8549	0.005	163.605	< 0.001

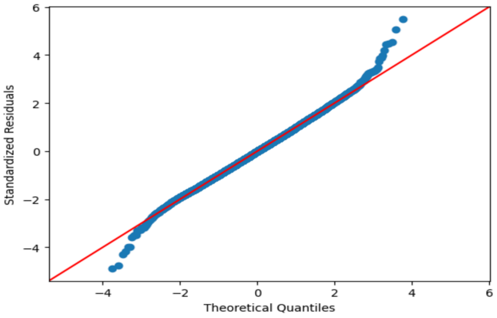
The coefficients from the above table tells use that for every unit increase of the income group while the other variable remain unchanged, the spring general knowledge score increases by 0.4803, and for every unit increase of fall general knowledge score, while the other remain unchanged, the spring score increases by 0.8549. Moreover, we can see that all coefficient p value are < 0.001 which are significant and indicate the each of the factors affect the response variable significantly

4.1.2 Model Assumptions

The assumptions of normality and homogeneity of variance are check based on below:

The QQ plot based on standardized residual is shown on the right. We can see that most of the dots stick to the straight line with some dots slightly off on two ends of the line. Thus the the normal assumption followed

The homogeneity of variance assumption is tested through Bartlett's test since the sample distribution is normal. We found from the test result that the p value is smaller than 0.001 that means the homogeneity of variance assumption is violated



4.2 Reading test Score

Null hypothesis: The average score of the spring reading test score is the same across different income groups aftering taking control of the fall reading test score as covariate .

Alternative hypothesis: The average score of the spring reading test score is not the same across different income groups aftering taking control of the fall reading test score as covariate .

4.2.1 Model Result

The result of the model is presented in the following tables. There are some findings and conclusions can be made from the results.

R squared	Adjusted R squared	F-statistic	Prob (F-stat)	Log-likelihood	AIC	BIC
0.692	0.692	1.339e+04	< 0.001	-41675	8.336e+04	8.338e+04

Similar to the first model, a R squared of 0.692 indicates around 69.2% of variability is explained and the adjusted R square is similar to R squared indicates a proper fit of model. We get a similar p value as the previous test so we reject the null hypothesis

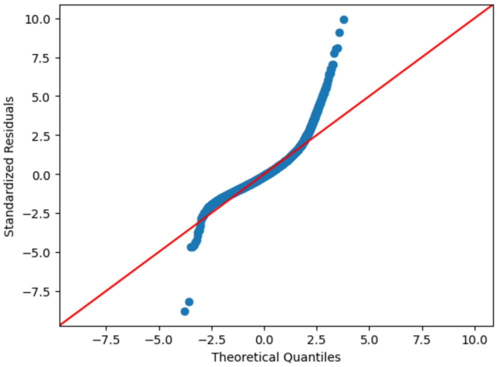
In terms of the coefficients, for every unit increase of the income group while the other variable remains unchanged, the spring reading score increases by 0.2512, and for every unit increase of fall reading score, while the other remains unchanged, the spring score increases by 1.1312. Moreover, we can see again that all coefficient p value are very small which are significant and indicate the each of the factors affect the response variable significantly

4.1.2 Model Assumptions

The assumptions of normality and homogeneity of variance are check based on below:

The QQ plot based on standardized residual is shown on the right. We can see that the dots are quite off from the straight line indicating the violation of normality assumption.

Also The homogeneity of variance assumption is tested through Bartlett's test since the sample distribution is normal. We found from the test result that the p value is smaller than 0.001 that means the homogeneity of variance assumption is violated.



5 Conclusion

All in all, we can conclude from the test result that income families cna indeed affect the children's reading ability and amount of general knowledge. I suggest that kindergarten can offer additional help sessions for children so that children who have weak ability in these skills can get additional help. However, since there exists assumption violation in both models, the accuracy of the result is still questionable.