Report for Reading Week Project:

Introduction

In this report we analyse a subset of 13000 plus pupils taken from both the Next Steps study and National Pupil Database (NPD) between 2005 and 2006. The study starts when the cohort are aged 14 in the Next Steps study. The NPD has information on exams scores, while the Next Steps study has information on the pupil's private lives. We are interested to know which variables have significant explanatory power for GCSE grades (ks4score).

The two studies together contain many variables. In this report we focus on the relationship between GCSE score (ks4score) and 9 significant explanatory variables: Test scores from age 14 (sum_ks3score), gender (genderMale), their attitude towards school (attitudelow), whether they want to continue studies after GCSE's (pupaspYes), how many evenings they do homework a week (homeworkin45), whether they have been excluded (excludeYes), whether they are truant (truancyYes) if their school is in a high FSM band (FSMbandhigh) and whether they come from a single parent household (singleparyes).

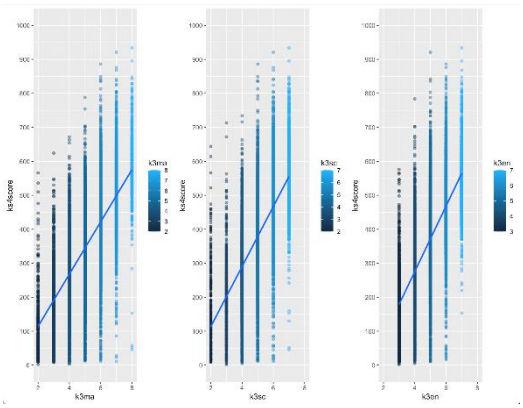
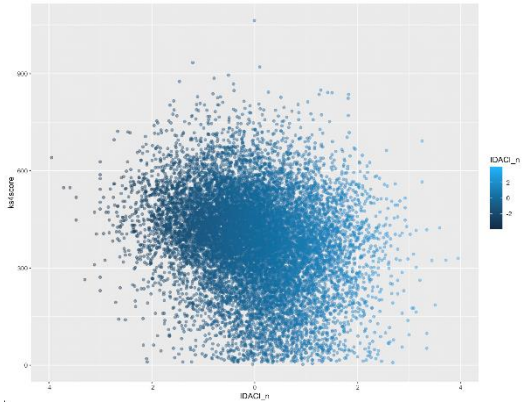
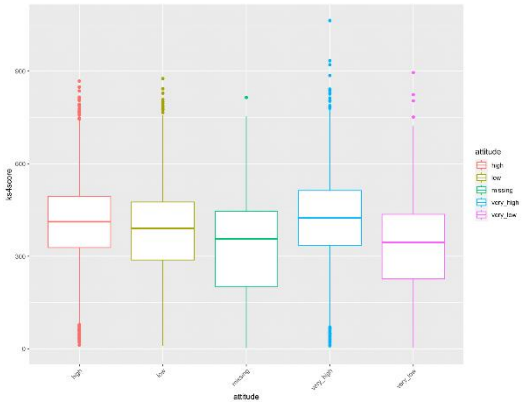
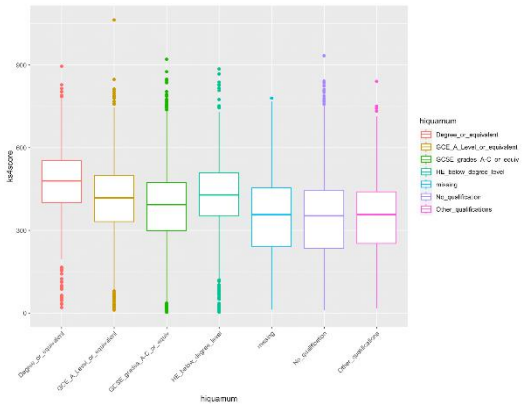
Our results show that the test scores at age 14 (sum_k3score), had the greatest magnitude of effect on the results, whilst the others are statistically significant, but smaller in the magnitude of their effect.

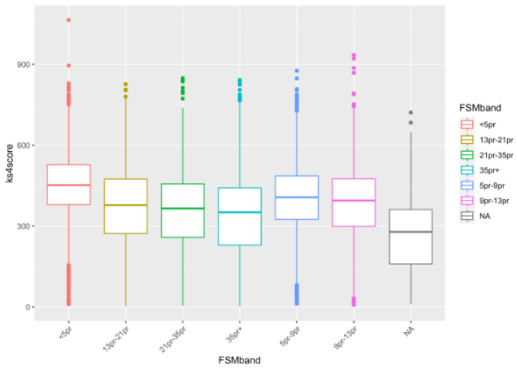
Exploratory Analysis and Full to Final Model:

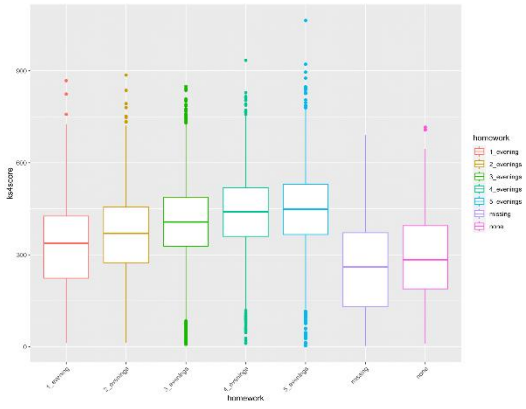
From the data set, there are four numerical predictors: k3en, k3ma, k3sc, and IDACI_n. When creating scatter plots for the numerical variables of to ks4score, it is observed that there exists a strong positive linear correlation between each of the k3en, k3ma, k3sc, and ks4score. However, there was no clear relationship found between ks4score against IDACI_n. To analyse the categorical predictors, we created box plots, but most of the predictors have numerous levels resulting in similar medians for each box. It is difficult to tell which predictors ks4score is dependent on. Therefore, we included all categorical predictors into our initial linear model. Also, we found that the box plot generated for ks4score based on gender and attitude indicates that female students scored higher than male students. Furthermore, the box plots also demonstrate that as the attitude score decreases, the corresponding ks4score also tends to decrease.

We put some key plots in Graph 1 to justify our decision, which could also be used later in the merging parts.

Graph 1:

k3en, k3ma, k3sc	IDACI_n
	
Attitude	Hiqumum
	
FSMband	Homework





Within the categorical predictors, there are two predictors called fiveac and fiveem, which mean ‘5 or more GCSE grades A*-C', excluding and including maths and English respectively. These two predictors are directly linked to ks4score, both indicating the exam data from the same tests. We have chosen to exclude these two predictors as they are as dependent on the remaining predictors as ks4score. After analysis of scatter and box plots, we started running a linear regression model. Our initial idea was running all the variables given except fiveac and fiveem as explained above, the evidence for deleting here is their relatively high p-value. By evaluating through the charts, it was clear that over half of the variables had a significance level of 99%. Subsequently, we decided to use this significance level as a benchmark to exclude any predictors in further models. Thereby we deleted those variables whose significance of all subvariable’s is lower than 1% which are [SECshort, fsm, computer, tuition, parasp, absent, IDACI_n], and obtained 14 variables [k3en, k3ma, k3sc, gender, hiquamum, pupasp, homework, attitude, sen, truancy, exclude, FSMband, singlepar, house]. With these remaining variables, we ran the regression to check and deleted sen based on the significance level.

All predictors in this linear regression model were significant if we consider predictors which have multiple levels as a single variable. Previously, we had included all of the values indicated as ‘missing’ or ‘NA’ within our regression. Here, we chose to delete all the missing values since we did not know the meaning of the missing value. For example, around a thousand observed values are missing in the predictor ‘Truancy’, some students have been truant, but choose not to tell us. It will be a higher bias if we choose to combine the missing values with other levels, so can't merge it with the other levels. To manipulate, we analysed the data using summary, converted ‘NA’ in FSMband into ‘missing’, and found [hiquamum, homework, attitude, truancy, exclude, FSMband, singlepar] have missing values. Then we

created a new data frame without missing values. After this step, we reran the regression model and deleted the predictor house based on the significance level.

For the merging part, we use plots and data to evaluate and found:

1. k3en, k3ma, k3sc show similar effects with high significance level

2. Homework has 6 answers where some showed similar results

3. Attitude has 4 answers where some showed similar results

4. Hiqumum has 6 answers where some showed similar results

5. FSMband has 6 answers where some showed similar results

After analysis of the plots in Graph 1 and data of linear regression, it is clear that some predictors have different significance levels across their subgroups, for example, attitude very low has <0.1%, low has 5%. Combining this point with our key reasons to merge below, we reached merged variables in Table 1.

1. For the predictors - k3en, k3ma, k3sc, just according to our analysis above, we decided to merge them all into one stronger predictor -- sum_k3score.

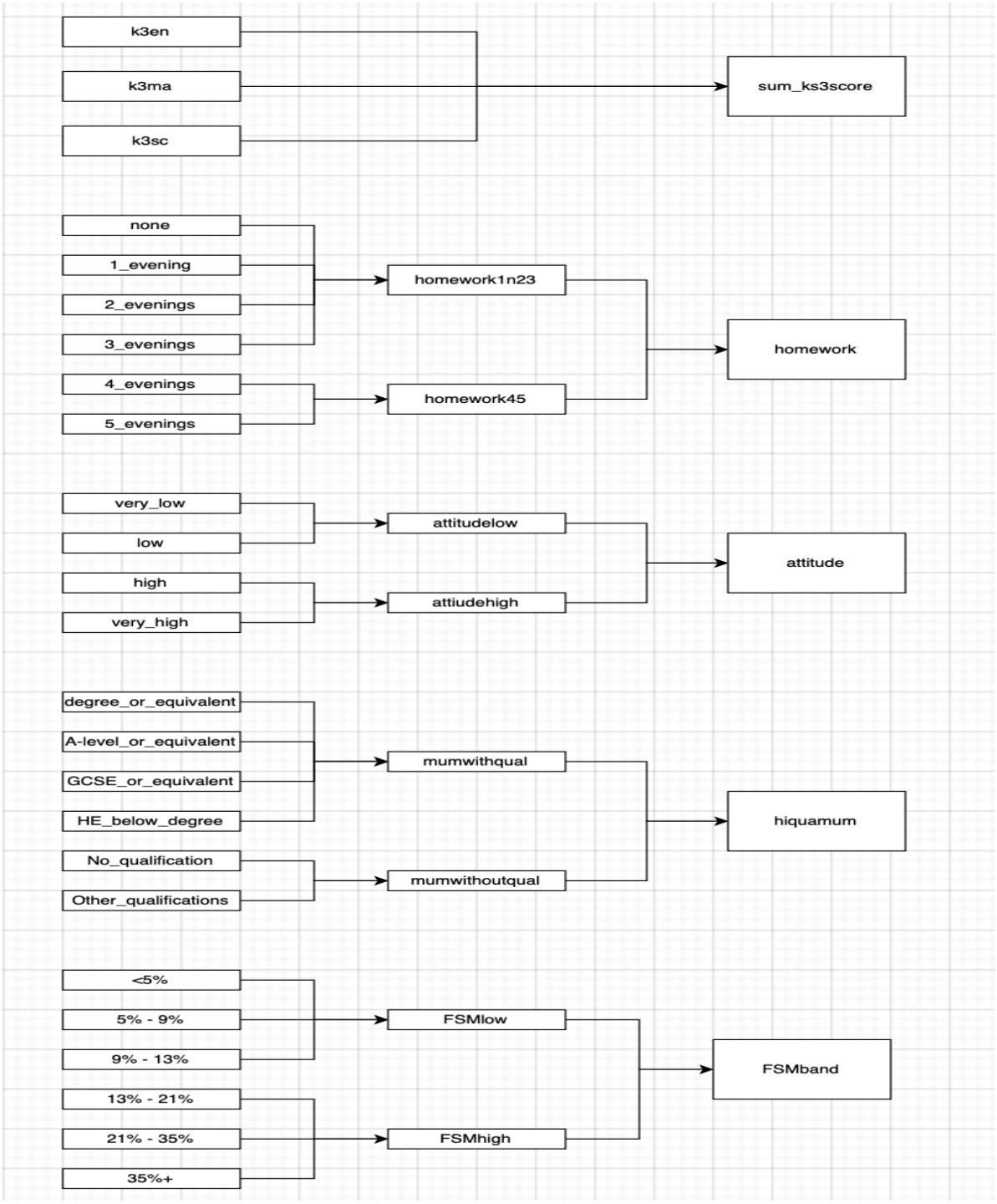
2. For the predictor homework, we thought that students who do their homework 4 or 5 times a week can be referred to as always doing homework, so these two levels can be merged into one level homework45, which has a strong positive correlation with ks4score. And for the other four levels, then can be a new level called homework1n23.

3. For the predictor attitude, when setting the baseline of attitude low, we found that obviously the attitude high and very high had positive intercepts, while very low had a negative one. So we decided to create two new levels, attitudelow and attitudehigh.

4. For the predictor hiquamum, we combined the 'No_qualification' and 'Other_qualifications' into one level named mumwithoutqual, while the remaining four categories were grouped as mumwithqual.

5. For the predictor FSMband, we reassign the values accordingly as what we did in attitude and created FSMbandlow, FSMbandhigh.

Table 1: merged variables



After performing these procedures, we reran the linear model and observed that the significance of the predictor hiquamum decreased. As a result, we decided to eliminate it from the model. As hiquamum is not a predictor anymore, we added the missing cases from the hiquamum back into the data and ran the model again. We then got the final model.

We present our final model below:

ks4score= -107.1771 + 31.2296 * sum_k3score - 14.9499 * genderMale + 24.8807 * pupaspYes + 10.7055 * homeworkin45 -16.0126 * attitudelow - 23.7605* truancyYes - 50.4809 * excludeYes + 11.4987 * FSMbandhigh -18.6255 * singleparyes + e

Table 2: Final Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-107.1771	5.7726	-18.566	< 2e-16	***
sum_k3score	31.2296	0.3214	97.177	< 2e-16	***
genderMale	-14.9499	1.8234	-8.199	2.73E-16	***
pupaspYes	24.8807	2.6684	9.324	< 2e-16	***
homeworkin45	10.7055	1.9783	5.411	6.40E-08	***
attitudelow	-16.0126	1.8675	-8.574	< 2e-16	***
truancyYes	-23.7605	2.9103	-8.164	3.63E-16	***
FSMbandhigh	11.4987	1.9155	6.003	2.01E-09	***
excludeYes	-50.4809	3.7025	-13.634	< 2e-16	***
singleparyes	-18.6255	2.211	-8.424	< 2e-16	***

The goodness of fit statistics for the final model are shown in Table 2 above, whose coefficients are sum_k3score, gender, pupasp, homework, attitude, truancy, exclude, FSMband and singlepar with significance level less than 0.1% level, which is relatively high. The R2 and Adj R2 were about all 0.598, which is around 0.6, so satisfactory. At the same time we could see that both the F-statistic and p-value showed the model is a good fit. Also, the residual standard error is 89.02 on 9847 degrees of freedom, which is also valuable to indicate a good model.

Results

After excluding a large number of the initial variables provided, as was described in the Exploratory Analysis and Full to Final sections of the model, we arrived at our final regression model, in which regressors are highly statistically significant.

The results of our final regression are summarized in the table below:

Variable	**Final Model 1**	**P-Values**
Intercept	-107.1771	<2e-16
sum_k3score	31.2296	< 2e-16
genderMale	-14.9499	2.73E-16
pupaspYes	24.8807	< 2e-16
homeworkin45	10.7055	6.40E-08
attitudelow	-16.0126	< 2e-16
truancyYes	-23.7605	3.63E-16
FSMbandhigh	11.4987	2.01E-09
excludeYes	-50.4809	< 2e-16
singleparyes	-18.6255	< 2e-16

All the variables in our results table are statistically significant to at least the 0.1% level.

The most statistically significant variables are sum_k3score, pupaspYes, attitudelow, excludeYes and singleparyes. It makes sense for sum_k3score to be highly statistically significant, as sum_k3score likely acts as a good measure of natural ability. Furthermore, it may measure a students' test-taking and exam preparation skills, since the grades which comprise sum_k3score were from age 14, which was only 2 years prior to writing GCSE's.

Both attitudelow and pupaspYes capture a student's motivation, so it's obvious that they should be significant. Furthermore, some students with a good attitude would not want to go to university, so they are not too similar, which would lead to issues of low significance.

excludeYes and singleparyes' Strong explanatory power also makes sense, since excludeYes likely functions as a good proxy for how much a student cares about school, as well as how good or bad their home situation is, while singleparyes is highly related to how much parental attention and assistance a student receives.

genderMale, pupaspYes, homeworkin45 truancyYes, attitudelow, FSMbandhigh and singleparyes all have effects which are fairly small in magnitude, whilst excludeYes and sum_k3score have effects of the greatest magnitudes (especially when considering that sum_k3score takes on a wide range of values, as opposed to being binary like many of the other variables). In fact, the R2 of a regression using just sum_k3score as an independent variable is 0.56, thus meaning that the bulk of our final R2 can be attributed to sum_k3score.

The intercept of the regression was negative. This clearly doesn't make sense, as if all the variables were zero, it would predict a negative ks4score, which is impossible. However, a student with a value of zero for all variables would arguably get a score close to zero (they

would have obtained zero in the exams which comprise sum_k3score, would do little homework and would have a bad attitude). However, although the difference between the intercept value and zero is highly statistically significant, the magnitude of the difference is small relative to ks4score's typical values, which leads us to conclude that the intercept value is somewhat reasonable given the data.

Comments about data/analysis

Much of the data we were provided with was self-reported, this may lead to issues of bias when missing data was excluded. For example, data may be missing when students with an extremely bad attitude did not fill in certain information, leading to the worst students being excluded from the data set, thus causing the model to indicate that having a bad attitude has a smaller negative effect relative to what it does in reality.

Furthermore, certain variables are prone to having been lied about (even in anonymous surveys, people will occasionally lie). For example, individuals may lie and say that they've handed in more homework than they actually have in order to feel better about themselves. This would result in a decrease in the change in GCSE grades attributed to doing more homework, as some people will report doing more homework who did not and thus did not gain the additional benefit of having done more homework. It could also go the other way, in that students may report doing less homework than they did in reality in order to feel better about their natural ability, thus making the direction of bias difficult to predict in this scenario.

There are certainly also flaws in our model because of the time frame in which the data were captured. The data we used only spanned the years 2005 to 2006, but certain factors may have become much more or less important since then. For example, free online learning resources have become much more freely available since then, so the effect of going to a better school, or hiring private tutors may be diminished, whilst the effect of having access to a computer may be increased.

In terms of flaws in our model due to a lack of expertise, the mild heteroskedasticity present in our model is cause for concern, however at the current point in time we do not have the appropriate expertise to deal with the potential issues caused by this.

A counterintuitive result we came across in the model was that all other factors held equal, going to a school in a higher FSM band increased GCSE grades. This is counterintuitive, as typically students in schools in higher free school meal bands are from backgrounds with less access to resources, academic help etc. than those who are not. Our counterintuitive result may be due to flaws in the data, or confounders which aren't included in our regression.

Interpretations and Conclusions for a lay audience

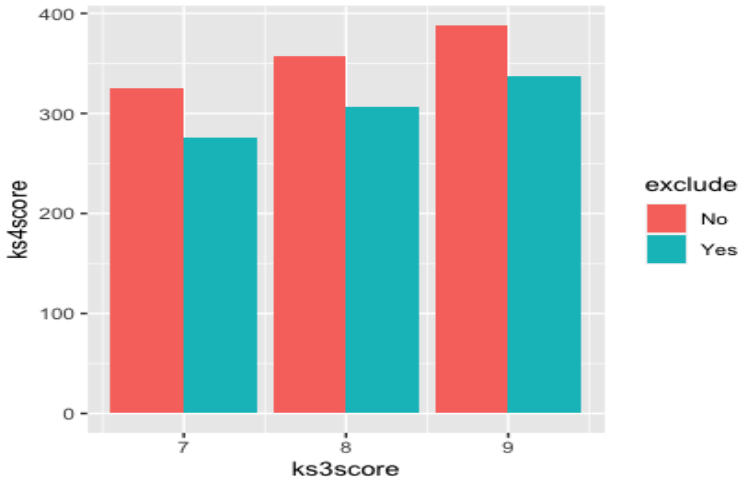
In our analysis, we attempted to understand the effect of nine variables on a students' GCSE grades. We obtained the data for our analysis from a combination of data taken from the Next Steps Study and the National Pupils Database between 2005 and 2006. The nine variables we used were: the sum of the students' scores on tests that they took at age 14, the gender of the student, whether or not the student wants to continue full time education after age 16, whether or not the student completed homework four or more evenings per week, Whether or not the student had a good attitude, whether or not the student was truant in the last 12 months, whether or not the student was excluded from school between ages 11 and 14, whether or not the student is in a high FSM band and whether the student comes from a single parent household

We ran a regression analysis to determine the relationships between the above variables and a student's GCSE scores.

Using our regression, we found that the most important factor was the sum of the students' grades on the exams they took at age 14. In fact, an increase in the sum of their exam scores by 1 was associated with an expected increase in the students' total GCSE score of 31.22. This is especially impressive when considering that the sum of the students' scores ranges from 7 all the way to 22.

Another significant factor was whether or not the student had been excluded from school in the last year. A male student who got a sum of 10 on the exams they took at 14, wants to continue full time education, did homework at least 4 times per week, had a good attitude, was not in a high a high free school meal band, did not commit truancy in the last 12 months and did not get excluded from school at any point from ages 11 to 14, with two parents in the household, would have an estimated GCSE score of 225 according to the model. On the other hand, a student with the exact same stats listed above, but with the sole change of being excluded from school would have a predicted GCSE score of 175 in the model, a change of 50 points.

The below bar chart displays the expected grades for two students who were and weren't excluded in the last 12 months for various scores obtained on exam taken at age 14, other factors held constant.



As can be seen, both exclusion and the exam scores have a significant impact on the subsequent GCSE scores of students.

Based on our model and data, we are hesitant to make policy recommendations based on our data and model. We have three main reasons for this:

1. The data we used are outdated and the importance of different variables may have changed significantly since 2006 (e.g. having access to a computer may be more important for grades now than it was than).
2. There was a significant amount of missing data. We are concerned that the missing data is not missing at random, but is instead correlated with other factors, such as attitude. This would result in bias in our analysis, because we may have excluded the people with the worst attitudes from the study (due to their refusal to fill in data).
3. We had access to limited data and there may be other factors which are relevant for the model which we could not include.

If forced to make recommendations, we would recommend that the government invest more heavily in early education and teaching students study skills at a younger age. The reason we recommend this is that in our analysis we found that results on test scores taken at 14 were very important for final GCSE grades. Furthermore, we would recommend that the government pay special attention to students who have been excluded from the school and provides them with additional help. Perhaps encouraging them to go through counselling and offering them homework assignments so that they can keep up to date with school work whilst excluded would help. The reason for this recommendation is again that the effect of being excluded from school on GCSE grades is quite large.