final project

Aiying Huang

2023-12-01

Table 1: Data summary

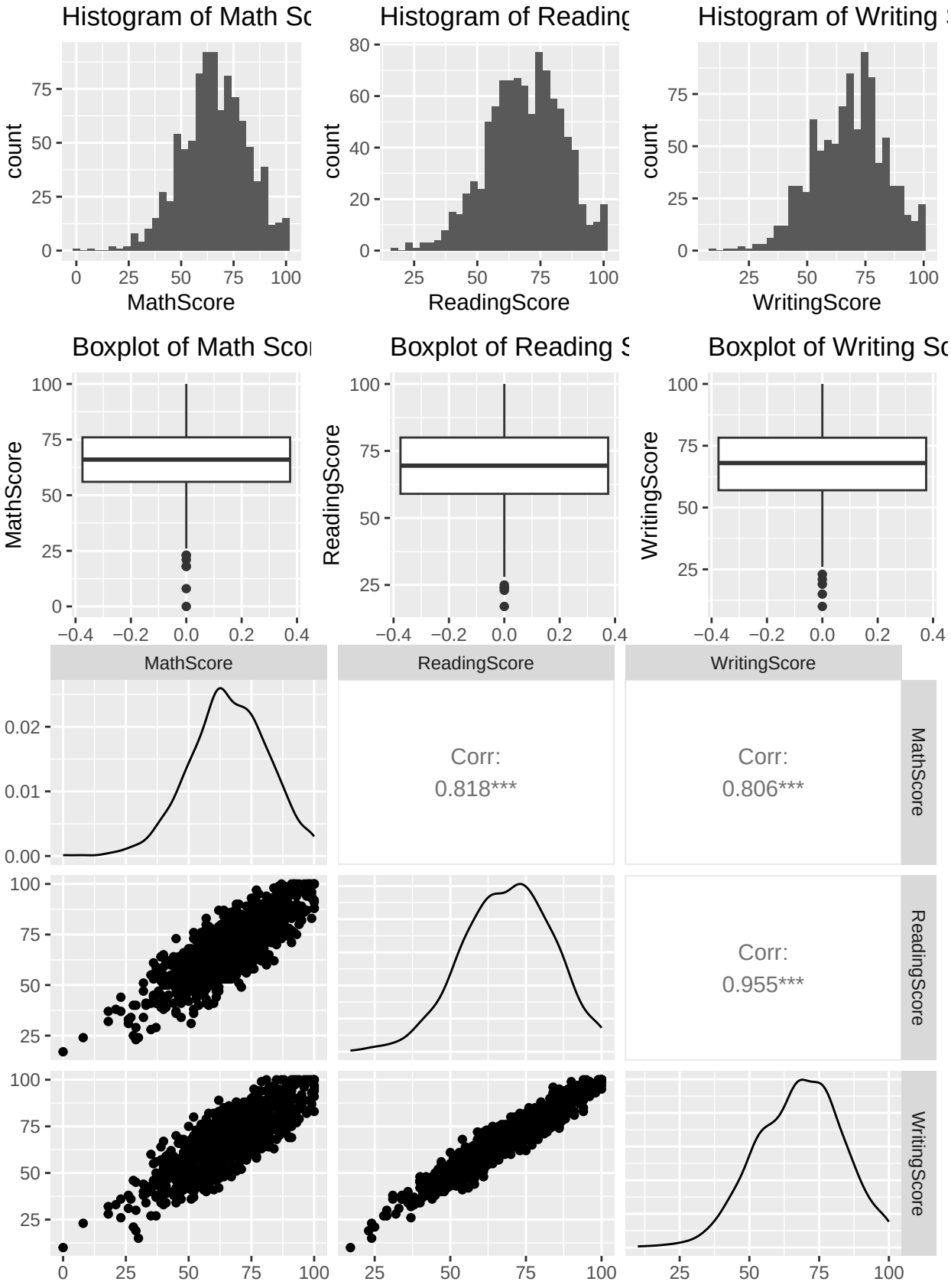
Name	data
Number of rows	948
Number of columns	14
Column type frequency:	
character	10
numeric	4
Group variables	None

Variable type: character

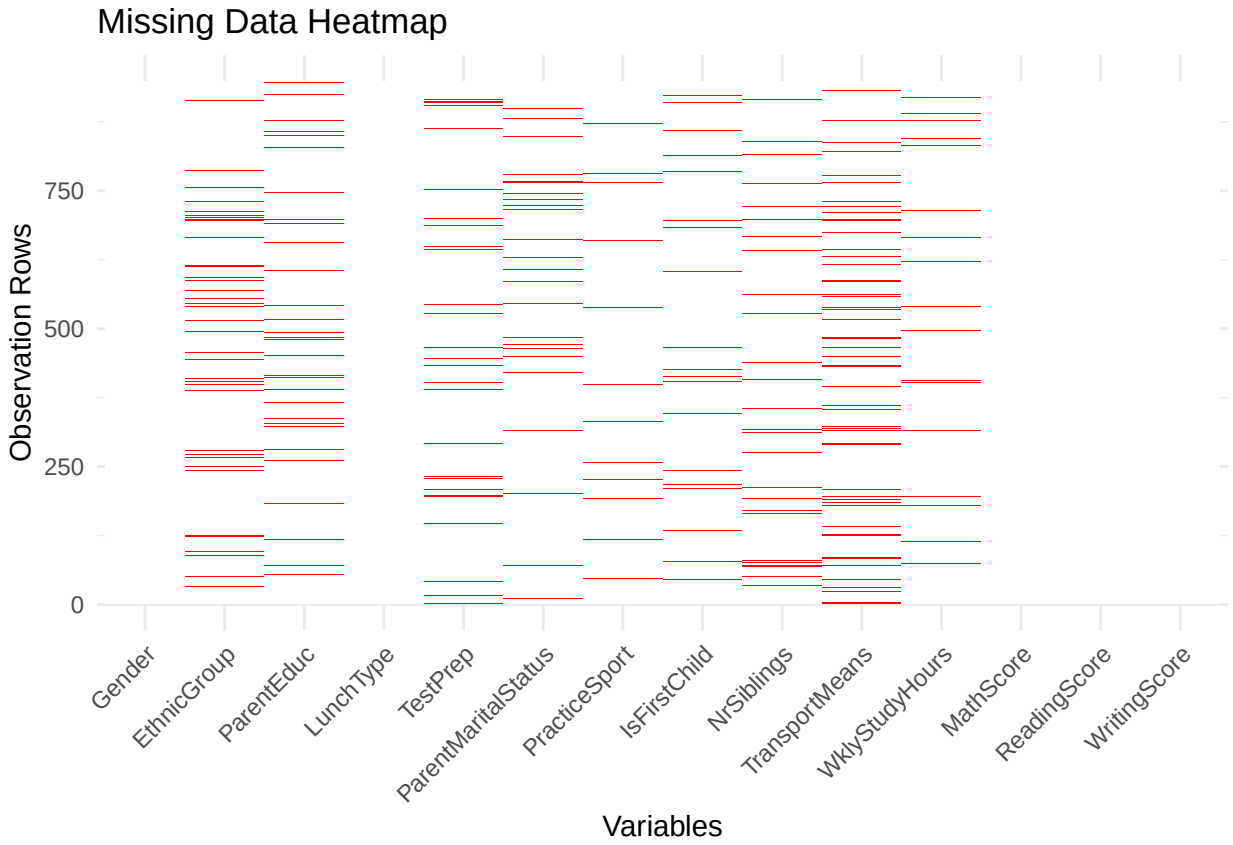
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1.00	4	6	0	2	0
EthnicGroup	59	0.94	7	7	0	5	0
ParentEduc	53	0.94	11	18	0	6	0
LunchType	0	1.00	8	12	0	2	0
TestPrep	55	0.94	4	9	0	2	0
ParentMaritalStatus	49	0.95	6	8	0	4	0
PracticeSport	16	0.98	5	9	0	3	0
IsFirstChild	30	0.97	2	3	0	2	0
TransportMeans	102	0.89	7	10	0	2	0
WklyStudyHours	37	0.96	3	6	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
NrSiblings	46	0.95	2.16	1.48	0	1	2.0	3.00	7	
MathScore	0	1.00	65.98	15.53	0	56	66.0	76.00	100	
ReadingScore	0	1.00	68.84	14.80	17	59	69.5	80.00	100	
WritingScore	0	1.00	67.93	15.41	10	57	68.0	78.25	100	


Histogram of Math Sc
Histogram of Reading
Histogram of Writing
count
MathScore
ReadingScore
WritingScore
Boxplot of Math Scor
Boxplot of Reading S
Boxplot of Writing Sc
MathScore
ReadingScore
WritingScore
Corr:
0.818***
Corr:
0.806***
Corr:
0.955***

```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



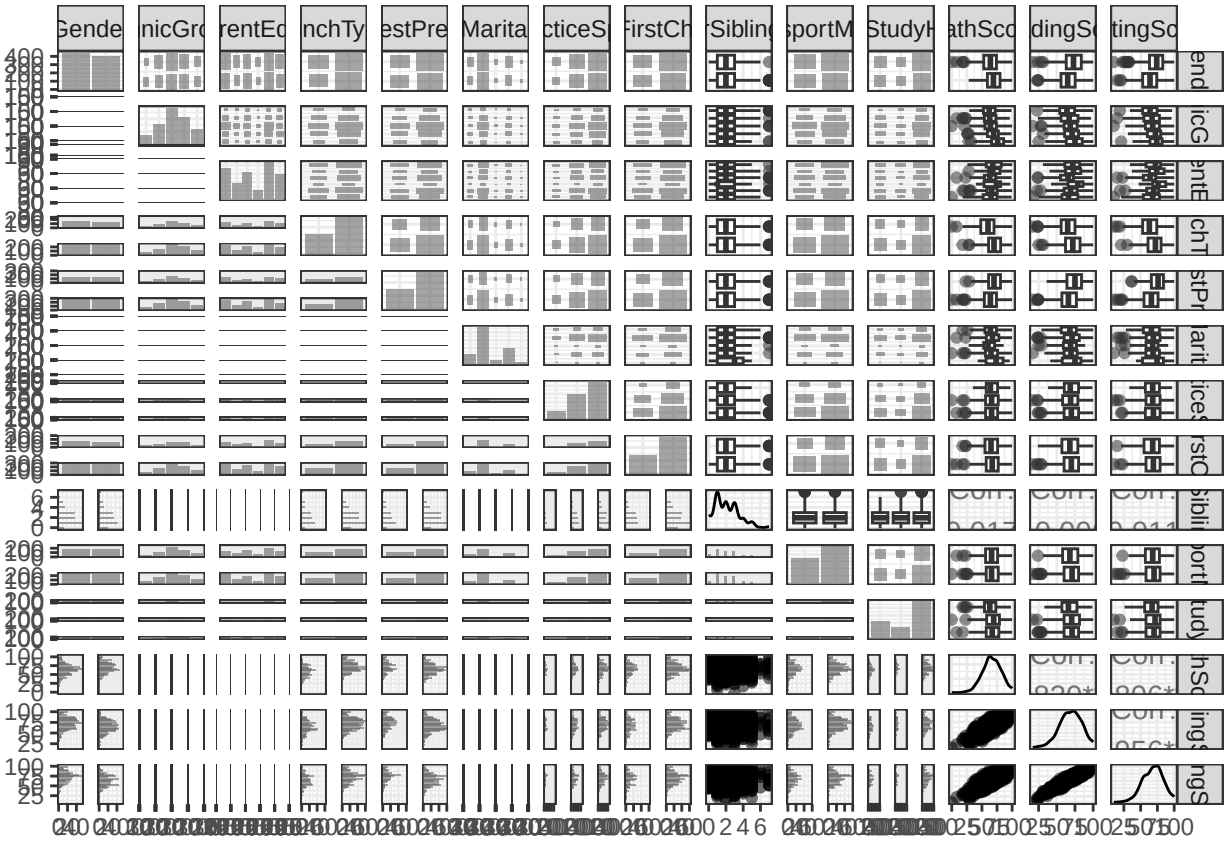

```
##                                Variable MissingValues
## Gender                           Gender             0
## EthnicGroup                 EthnicGroup            59
## ParentEduc                   ParentEduc            53
## LunchType                     LunchType             0
## TestPrep                       TestPrep            55
## ParentMaritalStatus ParentMaritalStatus            49
## PracticeSport           PracticeSport              16
## IsFirstChild             IsFirstChild              30
## NrSiblings                 NrSiblings              46
## TransportMeans         TransportMeans             102
## WklyStudyHours         WklyStudyHours              37
## MathScore                   MathScore              0
## ReadingScore             ReadingScore              0
## WritingScore             WritingScore              0
```

-correlation/pairwise-

Examine the marginal distributions and pairwise relationships between variables

```
# Load necessary libraries
library(tidyverse)
library(ggplot2)
library(GGally)

# draw the pariplot
ggpairs(data, columns=1:14, aes(alpha = 0.3))+
  theme_bw()
```



Correlation between variables

```
# Load necessary libraries
library(greybox)
```

```
## Package "greybox", v2.0.0 loaded.
```

```
##
## Attaching package: 'greybox'
```

```
## The following object is masked from 'package:lubridate':
##
##     hm
```

```
## The following object is masked from 'package:tidyr':
##
##     spread
```

```
library(tidyverse)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

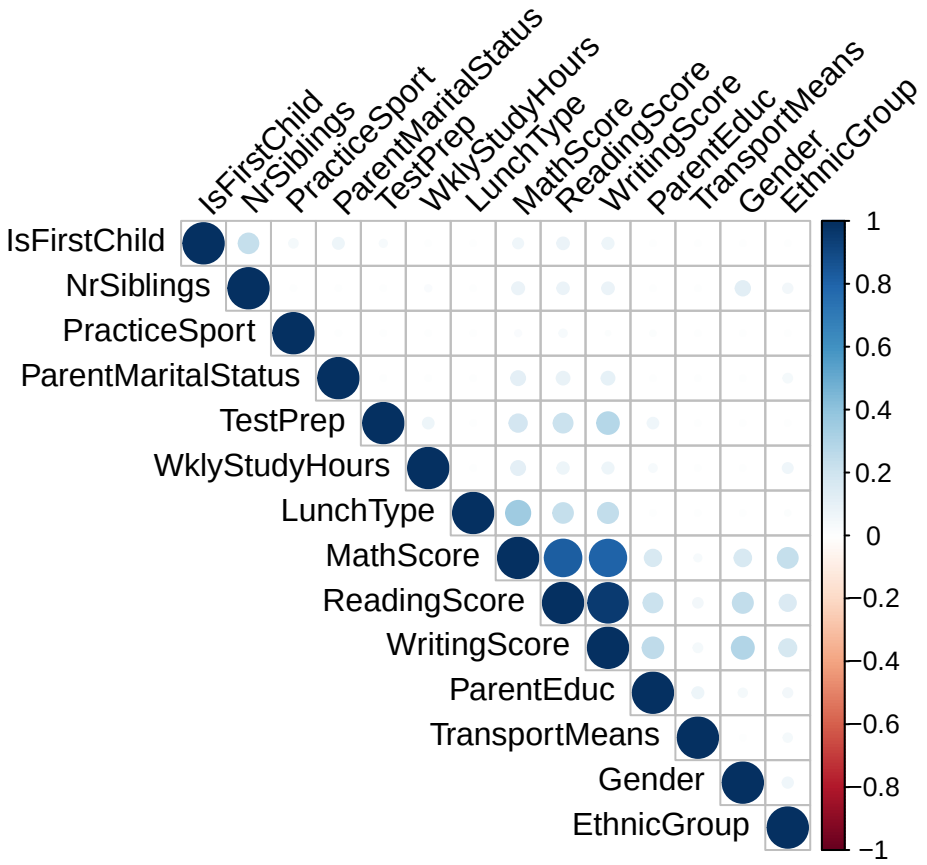
```
# Compute the Cramer's V correlation between variables
cramer_v_matrix <- assoc(data, method = "auto")

# Extract the matrix with Cramer's V values
cramer_v_values <- as.matrix(cramer_v_matrix$value)

# Print the correlation matrix results
knitr::kable(cramer_v_values, digits = 3)
```

	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	PracticeSport	IsFirstChild	NrSiblings	TransportMeans	WklyStudyHours	MathScore	ReadingScore	WritingScore
Gender	1.000	0.064	0.042	0.000	0.000	0.000	0.000	0.000	0.126	0.000	0.000	0.168	0.244	0.294
EthnicGroup	0.064	1.000	0.050	0.018	0.000	0.047	0.000	0.000	0.054	0.044	0.060	0.240	0.160	0.177
ParentEduc	0.042	0.050	1.000	0.000	0.069	0.000	0.018	0.000	0.000	0.074	0.036	0.163	0.217	0.260
LunchType	0.000	0.018	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.357	0.236	0.246
TestPrep	0.000	0.000	0.069	0.000	1.000	0.000	0.000	0.032	0.000	0.000	0.070	0.184	0.217	0.286
ParentMaritalStatus	0.018	0.047	0.000	0.000	0.000	1.000	0.000	0.074	0.000	0.000	0.000	0.118	0.099	0.100
PracticeSport	0.000	0.000	0.018	0.000	0.000	0.000	1.000	0.045	0.000	0.000	0.000	0.022	0.033	0.012
IsFirstChild	0.000	0.000	0.000	0.000	0.032	0.074	0.045	1.000	0.235	0.000	0.000	0.061	0.083	0.075
NrSiblings	0.126	0.054	0.000	0.000	0.000	0.000	0.000	0.235	1.000	0.000	0.024	0.088	0.081	0.084
TransportMeans	0.000	0.044	0.074	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.030	0.056	0.047
WklyStudyHours	0.000	0.060	0.036	0.000	0.070	0.000	0.000	0.000	0.024	0.000	1.000	0.119	0.079	0.075
MathScore	0.168	0.240	0.163	0.357	0.184	0.118	0.022	0.061	0.088	0.030	0.119	1.000	0.820	0.806
ReadingScore	0.244	0.160	0.217	0.236	0.217	0.099	0.033	0.083	0.081	0.056	0.079	0.820	1.000	0.956
WritingScore	0.294	0.177	0.260	0.246	0.286	0.100	0.012	0.075	0.084	0.047	0.075	0.806	0.956	1.000

```
# Create a heatmap
corrplot(cramer_v_values, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Cramér's V (for categorical variables) varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when each variable is completely determined by the other.

no relation predicators colinear auto correlation compare Strength of association is calculated for nominal vs nominal with a bias corrected Cramer's V, numeric vs numeric with Spearman (default) or Pearson correlation, and nominal vs numeric with ANOVA.

```
#########   ##########################################################
```

```
## # A tibble: 3 x 3
##   variable1    variable2       p_value
##   <chr>        <chr>             <dbl>
## 1 Gender       NrSiblings     0.00250
## 2 TestPrep     WklyStudyHours 0.0465
## 3 IsFirstChild NrSiblings     0.000500
```

```
##
## Call:
## lm(formula = Y_math_train ~ ., data = X_train)
##
## Coefficients:
##                  (Intercept)                   Gendermale
##                     50.10619                      4.76011
##          EthnicGroupgroup B          EthnicGroupgroup C
##                      1.22325                      1.48983
##          EthnicGroupgroup D          EthnicGroupgroup E
##                      5.34856                      9.87450
## ParentEducbachelor's degree        ParentEduchigh school
##                      1.46321                     -6.27372
```

```
##   ParentEducmaster's degree       ParentEducsome college
##                     1.83828                     -2.62638
##  ParentEducsome high school           LunchTypestandard
##                     -6.00375                    10.97098
##                TestPrepnone   ParentMaritalStatusmarried
##                     -5.62471                     4.69923
##    ParentMaritalStatusnone    ParentMaritalStatussingle
##                     1.64623                     2.01068
##  ParentMaritalStatuswidowed      PracticeSportregularly
##                     5.78004                     2.21718
##     PracticeSportsometimes             IsFirstChildyes
##                     1.03807                     1.74854
##                 NrSiblings1                 NrSiblings2
##                     -0.48258                     0.21839
##                 NrSiblings3                 NrSiblings4
##                     0.08025                    -0.04309
##                 NrSiblings5                 NrSiblings6
##                     1.91486                     5.65833
##                 NrSiblings7    TransportMeansschool_bus
##                     1.12613                     0.82894
##          WklyStudyHours> 10        WklyStudyHours10-May
##                     4.17235                     3.92089
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## Gender              1.064586  1        1.031788
## EthnicGroup         1.215358  4        1.024679
## ParentEduc          1.198109  5        1.018239
## LunchType           1.031755  1        1.015753
## TestPrep            1.049215  1        1.024312
## ParentMaritalStatus 1.178015  4        1.020690
## PracticeSport       1.065261  2        1.015930
## IsFirstChild        1.111657  1        1.054352
## NrSiblings          1.341818  7        1.021224
## TransportMeans      1.043894  1        1.021711
## WklyStudyHours      1.082693  2        1.020061
```

```
########################################
```

Model Selection

Despite the absence of discernible linear correlations among the variables, the inclusion of interaction terms is justified, guided by prior theoretical knowledge and practical considerations.

```
# Checking for interaction effects (example for math score)
full_model_math_interaction <- lm(Y_math_train ~  (.)^2, data = X_train)
full_model_reading_interaction <- lm(Y_reading_train ~  (.)^2, data = X_train)
full_model_writing_interaction <- lm(Y_writing_train ~  (.)^2, data = X_train)

# backward modeling(compare)
AICmodel_math_interaction =
  step(full_model_math_interaction, trace = 0, direction='backward')
BICmodel_math_interaction =
  step(full_model_math_interaction, scale = log(nrow(X_train)), trace = 0, direction='backward')

# show parameter numbers
```

```r
num_params_AICmodel <- length(coef(AICmodel_math_interaction))
num_params_BICmodel <- length(coef(BICmodel_math_interaction))

cat("AIC Model Parameters:", num_params_AICmodel, "\n")
```

```
## AIC Model Parameters: 140
```

```r
cat("BIC Model Parameters:", num_params_BICmodel, "\n")
```

```
## BIC Model Parameters: 387
```

Consequently, a comprehensive model was formulated, encompassing all 11 independent variables along with their respective pairwise interaction terms. In the ensuing stages of the analysis, a focus will be maintained on selecting a parsimonious subset of variables, with an aim to mitigate the risk of overfitting.

```r
# try AIC and BIC
model_math_interaction = AICmodel_math_interaction
model_reading_interaction =
  step(full_model_reading_interaction, trace = 0, direction='backward')
model_writing_interaction =
  step(full_model_writing_interaction, trace = 0, direction='backward')

# results
# r.squared
glance_math = broom::glance(model_math_interaction) |>
  mutate(model = "Math") |>
  select(model, r.squared, adj.r.squared, p.value, AIC, BIC)

glance_reading = broom::glance(model_reading_interaction) |>
  mutate(model = "Reading") |>
  select(model, r.squared, adj.r.squared, p.value, AIC, BIC)

glance_writing = broom::glance(model_writing_interaction) |>
  mutate(model = "Writing") |>
  select(model, r.squared, adj.r.squared, p.value, AIC, BIC)

bind_rows(glance_math, glance_reading, glance_writing) |>
  knitr::kable()
```

model	r.squared	adj.r.squared	p.value	AIC	BIC
Math	0.4866905	0.3559779	0	5482.060	6109.811
Reading	0.2490318	0.2213464	0	5453.040	5570.461
Writing	0.3599282	0.3249243	0	5407.924	5575.023

```r
# coef
broom::tidy(model_math_interaction) |>
  knitr::kable(caption = "Math")
```

Table 6: Math

term	estimate	std.error	statistic	p.value
(Intercept)	61.4088176	11.046889	5.5589240	0.0000000
Gendermale	9.3164343	6.000885	1.5525100	0.1211283

term	estimate	std.error	statistic	p.value
EthnicGroupgroup B	3.1725096	6.067085	0.5229050	0.6012556
EthnicGroupgroup C	-1.1855436	5.791732	-0.2046959	0.8378872
EthnicGroupgroup D	-0.1604340	6.116623	-0.0262292	0.9790843
EthnicGroupgroup E	3.9653103	6.306522	0.6287634	0.5297708
ParentEducbachelor's degree	13.0835066	11.351808	1.1525482	0.2496077
ParentEduchigh school	-10.0867244	10.035671	-1.0050872	0.3153067
ParentEducmaster's degree	-18.6347424	16.131638	-1.1551674	0.2485347
ParentEducsome college	-16.5654032	8.822173	-1.8777010	0.0609628
ParentEducsome high school	0.2328669	9.665795	0.0240919	0.9807883
LunchTypestandard	3.9162442	3.629378	1.0790401	0.2810534
TestPrepnone	-13.8234639	4.256595	-3.2475403	0.0012366
ParentMaritalStatusmarried	7.4579774	6.573034	1.1346324	0.2570347
ParentMaritalStatusnone	24.8719704	11.983226	2.0755655	0.0384087
ParentMaritalStatussingle	19.9856912	8.778502	2.2766630	0.0231983
ParentMaritalStatuswidowed	46.6926001	17.376320	2.6871397	0.0074297
PracticeSportregularly	-23.3229950	9.413486	-2.4776151	0.0135323
PracticeSportsometimes	-10.4722488	9.009083	-1.1624101	0.2455843
IsFirstChildyes	6.7591381	4.526718	1.4931654	0.1359801
NrSiblings1	1.1198173	6.386274	0.1753475	0.8608726
NrSiblings2	10.7048146	7.172073	1.4925690	0.1361362
NrSiblings3	2.9509080	6.968085	0.4234891	0.6721077
NrSiblings4	-6.8647778	10.101360	-0.6795895	0.4970567
NrSiblings5	-9.5054553	8.647917	-1.0991612	0.2721894
NrSiblings6	2.5509272	9.715463	0.2625636	0.7929874
NrSiblings7	-2.6654559	6.200122	-0.4299038	0.6674378
TransportMeansschool_bus	-4.0415193	2.999921	-1.3472086	0.1784802
WklyStudyHours> 10	1.9296494	6.179432	0.3122697	0.7549565
WklyStudyHours10-May	1.1938396	4.639791	0.2573046	0.7970420
Gendermale:EthnicGroupgroup B	-1.6360392	4.814066	-0.3398456	0.7341054
Gendermale:EthnicGroupgroup C	4.7965139	4.570706	1.0494033	0.2944637
Gendermale:EthnicGroupgroup D	-1.3061363	4.672370	-0.2795447	0.7799343
Gendermale:EthnicGroupgroup E	-1.6021336	4.984668	-0.3214123	0.7480228
Gendermale:ParentMaritalStatusmarried	-7.8006112	3.189957	-2.4453652	0.0147901
Gendermale:ParentMaritalStatusnone	-4.0299685	6.643769	-0.6065786	0.5443864
Gendermale:ParentMaritalStatussingle	-3.6734702	3.745811	-0.9806876	0.3271876
Gendermale:ParentMaritalStatuswidowed	-2.7625008	7.930574	-0.3483356	0.7277245
Gendermale:PracticeSportregularly	3.7038254	3.962396	0.9347438	0.3503397
Gendermale:PracticeSportsometimes	-2.6660003	3.833323	-0.6954801	0.4870548
EthnicGroupgroup B:ParentEducbachelor's degree	11.5380591	10.366792	1.1129826	0.2662129
EthnicGroupgroup C:ParentEducbachelor's degree	10.7608882	9.778678	1.1004441	0.2716308
EthnicGroupgroup D:ParentEducbachelor's degree	10.6288519	10.070354	1.0554596	0.2916889
EthnicGroupgroup E:ParentEducbachelor's degree	17.3963097	10.810495	1.6092057	0.1081580
EthnicGroupgroup B:ParentEduchigh school	-8.5301281	7.224639	-1.1806996	0.2382438
EthnicGroupgroup C:ParentEduchigh school	-0.4617453	6.801476	-0.0678890	0.9458992
EthnicGroupgroup D:ParentEduchigh school	2.2186359	7.163699	0.3097054	0.7569049
EthnicGroupgroup E:ParentEduchigh school	7.4590517	7.653258	0.9746244	0.3301846
EthnicGroupgroup B:ParentEducmaster's degree	20.6784704	13.668673	1.5128367	0.1309083
EthnicGroupgroup C:ParentEducmaster's degree	11.5414327	12.368751	0.9331122	0.3511806
EthnicGroupgroup D:ParentEducmaster's degree	29.1142305	12.268333	2.3731203	0.0179889
EthnicGroupgroup E:ParentEducmaster's degree	24.9938880	13.842312	1.8056151	0.0715373
EthnicGroupgroup B:ParentEducsome college	0.5422308	6.931302	0.0782293	0.9376748
EthnicGroupgroup C:ParentEducsome college	2.9097536	6.443238	0.4515980	0.6517406

term	estimate	std.error	statistic	p.value
EthnicGroupgroup D:ParentEducsome college	11.1132332	6.739593	1.6489473	0.0997422
EthnicGroupgroup E:ParentEducsome college	4.4467878	7.018584	0.6335734	0.5266285
EthnicGroupgroup B:ParentEducsome high school	-8.6299471	7.387235	-1.1682242	0.2432338
EthnicGroupgroup C:ParentEducsome high school	-5.4554825	7.328833	-0.7443862	0.4569679
EthnicGroupgroup D:ParentEducsome high school	-1.1207948	7.256886	-0.1544457	0.8773162
EthnicGroupgroup E:ParentEducsome high school	10.6125942	8.338515	1.2727199	0.2036672
ParentEducbachelor's degree:TestPrepnone	-3.5520227	4.100230	-0.8662984	0.3867127
ParentEduchigh school:TestPrepnone	0.0829263	3.613702	0.0229477	0.9817005
ParentEducmaster's degree:TestPrepnone	-6.2625004	5.962210	-1.0503656	0.2940216
ParentEducsome college:TestPrepnone	4.1245569	3.305755	1.2476898	0.2126875
ParentEducsome high school:TestPrepnone	5.7480449	3.643417	1.5776520	0.1152333
ParentEducbachelor's degree:ParentMaritalStatusmarried	-20.1804590	7.418390	-2.7203286	0.0067327
ParentEduchigh school:ParentMaritalStatusmarried	1.7971003	4.737930	0.3793007	0.7046142
ParentEducmaster's degree:ParentMaritalStatusmarried	-7.7891767	8.397266	-0.9275849	0.3540389
ParentEducsome college:ParentMaritalStatusmarried	-6.0507973	4.513566	-1.3405801	0.1806226
ParentEducsome high school:ParentMaritalStatusmarried	-9.4606586	5.385427	-1.7567148	0.0795353
ParentEducbachelor's degree:ParentMaritalStatusnone	-21.3091996	14.310216	-1.4890900	0.1370496
ParentEduchigh school:ParentMaritalStatusnone	-5.7308943	8.488951	-0.6751004	0.4999020
ParentEducmaster's degree:ParentMaritalStatusnone	-4.4589100	16.160311	-0.2759173	0.7827175
ParentEducsome college:ParentMaritalStatusnone	-7.7864746	8.010173	-0.9720732	0.3314510
ParentEducsome high school:ParentMaritalStatusnone	-11.3667469	11.604328	-0.9795265	0.3277602
ParentEducbachelor's degree:ParentMaritalStatussingle	-28.4811963	7.960590	-3.5777744	0.0003778
ParentEduchigh school:ParentMaritalStatussingle	-1.0722393	5.546451	-0.1933199	0.8467814
ParentEducmaster's degree:ParentMaritalStatussingle	-18.1609892	9.547631	-1.9021462	0.0576858
ParentEducsome college:ParentMaritalStatussingle	-13.8510720	5.314170	-2.6064413	0.0094025
ParentEducsome high school:ParentMaritalStatussingle	-12.1454163	6.067053	-2.0018643	0.0458009
ParentEducbachelor's degree:ParentMaritalStatuswidowed	-21.1198120	17.613866	-1.1990446	0.2310384
ParentEduchigh school:ParentMaritalStatuswidowed	-29.6781998	18.925354	-1.5681715	0.1174291
ParentEducmaster's degree:ParentMaritalStatuswidowed	-70.4602104	29.625244	-2.3783841	0.0177368
ParentEducsome college:ParentMaritalStatuswidowed	-16.3238533	16.737683	-0.9752756	0.3298619
ParentEducsome high school:ParentMaritalStatuswidowed	-42.8527054	17.124777	-2.5023803	0.0126317
ParentEducbachelor's degree:PracticeSportregularly	8.3680970	7.981818	1.0483949	0.2949275
ParentEduchigh school:PracticeSportregularly	-4.1603053	6.347495	-0.6554247	0.5124745
ParentEducmaster's degree:PracticeSportregularly	-3.2032371	9.735132	-0.3290389	0.7422543
ParentEducsome college:PracticeSportregularly	5.3482401	5.468121	0.9780764	0.3284761
ParentEducsome high school:PracticeSportregularly	-0.8172625	6.225220	-0.1312825	0.8956009
ParentEducbachelor's degree:PracticeSportsometimes	-1.9374238	7.710948	-0.2512562	0.8017119
ParentEduchigh school:PracticeSportsometimes	-0.2283046	6.141239	-0.0371757	0.9703587
ParentEducmaster's degree:PracticeSportsometimes	13.9149051	8.105402	1.7167446	0.0866014
ParentEducsome college:PracticeSportsometimes	9.9330248	5.320216	1.8670341	0.0624405
ParentEducsome high school:PracticeSportsometimes	0.9212007	5.973698	0.1542095	0.8775024
ParentEducbachelor's degree:IsFirstChildyes	-4.9833985	4.447708	-1.1204419	0.2630254

term	estimate	std.error	statistic	p.value
ParentEduchigh school:IsFirstChildyes	7.6841592	3.568022	2.1536188	0.0317137
ParentEducmaster's degree:IsFirstChildyes	8.8598397	5.513945	1.6068059	0.1086837
ParentEducsome college:IsFirstChildyes	8.3688353	3.273490	2.5565482	0.0108451
ParentEducsome high school:IsFirstChildyes	1.8272436	3.729396	0.4899570	0.6243641
LunchTypestandard:PracticeSportregularly	7.1452804	4.117244	1.7354521	0.0832334
LunchTypestandard:PracticeSportsometimes	9.1654660	3.943823	2.3240054	0.0204968
TestPrepnone:PracticeSportregularly	8.4956384	4.282605	1.9837546	0.0477910
TestPrepnone:PracticeSportsometimes	7.4450778	4.082624	1.8236011	0.0687671
ParentMaritalStatusmarried:PracticeSportregularly	10.0380401	5.342114	1.8790390	0.0607795
ParentMaritalStatusnone:PracticeSportregularly	-7.1349365	9.513131	-0.7500093	0.4535770
ParentMaritalStatussingle:PracticeSportregularly	-5.0377188	7.368460	-0.6836868	0.4944673
ParentMaritalStatuswidowed:PracticeSportregularly	-2.8371839	16.044581	-0.1768313	0.8597075
ParentMaritalStatusmarried:PracticeSportsometimes	3.0773083	5.070675	0.6068833	0.5441844
ParentMaritalStatusnone:PracticeSportsometimes	-12.9471008	8.806823	-1.4701217	0.1421134
ParentMaritalStatussingle:PracticeSportsometimes	-10.2709987	7.144038	-1.4377022	0.1510999
ParentMaritalStatuswidowed:PracticeSportsometimes	-2.8495634	16.810904	-0.1695069	0.8654617
ParentMaritalStatusmarried:IsFirstChildyes	-6.2453634	3.926687	-1.5904920	0.1123114
ParentMaritalStatusnone:IsFirstChildyes	-12.4793139	7.714099	-1.6177279	0.1063073
ParentMaritalStatussingle:IsFirstChildyes	-1.1855340	4.401450	-0.2693508	0.7877630
ParentMaritalStatuswidowed:IsFirstChildyes	10.0658977	12.508267	0.8047396	0.4213255
ParentMaritalStatusmarried:TransportMeansschool_bus	7.8797761	3.339323	2.3596928	0.0186462
ParentMaritalStatusnone:TransportMeansschool_bus	7.9984563	6.641455	1.2043230	0.2289943
ParentMaritalStatussingle:TransportMeansschool_bus	2.8278397	3.781858	0.7477383	0.4549447
ParentMaritalStatuswidowed:TransportMeansschool_bus	-19.6445027	11.606791	-1.6925008	0.0911295
PracticeSportregularly:NrSiblings1	0.5088957	7.157308	0.0711015	0.9433434
PracticeSportsometimes:NrSiblings1	-3.4973706	6.961731	-0.5023708	0.6156122
PracticeSportregularly:NrSiblings2	-6.9387451	7.910053	-0.8772059	0.3807663
PracticeSportsometimes:NrSiblings2	-17.2640507	7.763011	-2.2238860	0.0265705
PracticeSportregularly:NrSiblings3	1.3902158	7.764139	0.1790560	0.8579611
PracticeSportsometimes:NrSiblings3	-6.0851146	7.491617	-0.8122565	0.4170037
PracticeSportregularly:NrSiblings4	3.5073502	10.922535	0.3211114	0.7482507
PracticeSportsometimes:NrSiblings4	7.3654394	10.579080	0.6962269	0.4865874
PracticeSportregularly:NrSiblings5	16.6383891	9.836623	1.6914737	0.0913255
PracticeSportsometimes:NrSiblings5	10.4190439	9.348883	1.1144694	0.2655755
PracticeSportregularly:NrSiblings6	14.8266072	12.732168	1.1644998	0.2447376
PracticeSportsometimes:NrSiblings6	NA	NA	NA	NA
PracticeSportregularly:NrSiblings7	13.6357274	10.879764	1.2533110	0.2106370
PracticeSportsometimes:NrSiblings7	NA	NA	NA	NA
PracticeSportregularly:WklyStudyHours> 10	1.6935685	6.138330	0.2759005	0.7827304
PracticeSportsometimes:WklyStudyHours> 10	4.9818703	5.946600	0.8377678	0.4025331
PracticeSportregularly:WklyStudyHours10-May	11.3965731	4.657203	2.4470852	0.0147205
PracticeSportsometimes:WklyStudyHours10-May	5.8204031	4.493068	1.2954184	0.1957315
IsFirstChildyes:WklyStudyHours> 10	-1.0834857	3.598399	-0.3011022	0.7634529
IsFirstChildyes:WklyStudyHours10-May	-7.7061288	2.775496	-2.7764872	0.0056865

```
broom::tidy(model_reading_interaction) |>
  knitr::kable(caption = "Reading")
```

Table 7: Reading

term	estimate	std.error	statistic	p.value
(Intercept)	62.9639016	3.885243	16.2059119	0.0000000
Gendermale	-11.2814161	1.837147	-6.1407249	0.0000000
EthnicGroupgroup B	0.7322924	2.170119	0.3374434	0.7358914
EthnicGroupgroup C	1.0972551	2.044481	0.5366914	0.5916641
EthnicGroupgroup D	4.3368085	2.084041	2.0809610	0.0378278
EthnicGroupgroup E	5.8223627	2.307244	2.5235141	0.0118557
ParentEducbachelor's degree	1.1717604	1.940551	0.6038286	0.5461680
ParentEduchigh school	-5.9564479	1.631873	-3.6500674	0.0002832
ParentEducmaster's degree	3.1905460	2.374783	1.3435103	0.1795749
ParentEducsome college	-3.4572525	1.515029	-2.2819717	0.0228129
ParentEducsome high school	-6.3687036	1.732866	-3.6752434	0.0002572
LunchTypestandard	9.6783918	1.952202	4.9576801	0.0000009
TestPrepnone	-7.1729691	1.128634	-6.3554435	0.0000000
ParentMaritalStatusmarried	10.6131092	3.213924	3.3022279	0.0010117
ParentMaritalStatusnone	9.8107851	6.279111	1.5624481	0.1186684
ParentMaritalStatussingle	4.2137616	3.489714	1.2074804	0.2276857
ParentMaritalStatuswidowed	9.0417409	6.287990	1.4379383	0.1509321
IsFirstChildyes	7.9880898	3.719306	2.1477365	0.0321032
TransportMeansschool_bus	1.7322034	1.073741	1.6132408	0.1071768
Gendermale:IsFirstChildyes	5.0520373	2.231088	2.2643830	0.0238787
LunchTypestandard:IsFirstChildyes	-3.9410694	2.356205	-1.6726344	0.0948798
ParentMaritalStatusmarried:IsFirstChildyes	-8.5763774	3.697258	-2.3196586	0.0206678
ParentMaritalStatusnone:IsFirstChildyes	-10.2742988	7.021450	-1.4632731	0.1438753
ParentMaritalStatussingle:IsFirstChildyes	-2.4343962	4.078524	-0.5968817	0.5507939
ParentMaritalStatuswidowed:IsFirstChildyes	-4.1273385	7.575946	-0.5447951	0.5860810

```
broom::tidy(model_writing_interaction) |>
  knitr::kable(caption = "Writing")
```

Table 8: Writing

term	estimate	std.error	statistic	p.value
(Intercept)	56.7216183	4.670924	12.1435548	0.0000000
Gendermale	-5.1023135	2.969769	-1.7180842	0.0862649
EthnicGroupgroup B	-0.3609326	2.104296	-0.1715218	0.8638677
EthnicGroupgroup C	0.8447501	1.979834	0.4266772	0.6697579
EthnicGroupgroup D	5.7911650	2.014806	2.8743036	0.0041836
EthnicGroupgroup E	4.5399762	2.239547	2.0271852	0.0430574
ParentEducbachelor's degree	6.1984741	3.353272	1.8484853	0.0649932
ParentEduchigh school	-10.3655894	2.769891	-3.7422374	0.0001988
ParentEducmaster's degree	5.2500895	4.060929	1.2928296	0.1965365
ParentEducsome college	-8.7289152	2.451715	-3.5603306	0.0003978
ParentEducsome high school	-9.2257524	2.785427	-3.3121506	0.0009779
LunchTypestandard	8.0305724	1.058835	7.5843453	0.0000000
TestPrepnone	-9.4788974	1.088616	-8.7072939	0.0000000
ParentMaritalStatusmarried	11.9633720	3.122530	3.8313075	0.0001400
ParentMaritalStatusnone	6.9763575	6.065634	1.1501448	0.2505138
ParentMaritalStatussingle	5.8741390	3.408781	1.7232375	0.0853286
ParentMaritalStatuswidowed	6.7746978	6.117637	1.1074043	0.2685352

term	estimate	std.error	statistic	p.value
PracticeSportregularly	3.5334573	2.311655	1.5285400	0.1268726
PracticeSportsometimes	4.8986231	2.208763	2.2178127	0.0269180
IsFirstChildyes	8.9439025	4.191695	2.1337196	0.0332453
TransportMeansschool_bus	1.7049106	1.032210	1.6517092	0.0990844
WklyStudyHours> 10	3.1632615	2.681651	1.1795947	0.2385997
WklyStudyHours10-May	7.1506197	2.108350	3.3915718	0.0007377
Gendermale:PracticeSportregularly	-2.6986774	3.416209	-0.7899625	0.4298423
Gendermale:PracticeSportsometimes	-6.2654174	3.292604	-1.9028763	0.0575052
ParentEducbachelor's degree:IsFirstChildyes	-5.2894474	4.038891	-1.3096288	0.1907914
ParentEduchigh school:IsFirstChildyes	4.4165403	3.346881	1.3195988	0.1874409
ParentEducmaster's degree:IsFirstChildyes	0.2763606	4.901774	0.0563797	0.9550569
ParentEducsome college:IsFirstChildyes	8.3587252	3.046358	2.7438418	0.0062427
ParentEducsome high school:IsFirstChildyes	2.7994433	3.448495	0.8117869	0.4172158
ParentMaritalStatusmarried:IsFirstChildyes	-8.7730971	3.590119	-2.4436788	0.0148074
ParentMaritalStatusnone:IsFirstChildyes	-6.8052717	6.790204	-1.0022190	0.3166166
ParentMaritalStatussingle:IsFirstChildyes	-3.8134655	3.973008	-0.9598433	0.3374965
ParentMaritalStatuswidowed:IsFirstChildyes	-2.4713661	7.353749	-0.3360688	0.7369291
IsFirstChildyes:WklyStudyHours> 10	-1.8374437	3.351393	-0.5482627	0.5837027
IsFirstChildyes:WklyStudyHours10-May	-6.2461410	2.574453	-2.4262013	0.0155331

Initially, we performed a approach combining automated procedures and criterion-based with both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for model selection. It was observed that the application of the AIC criterion resulted in a model with fewer variables. Thus, we utilized the AIC criterion for backward elimination.

```
# try LASSO
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```
X_math <- model.matrix(~ Gender + EthnicGroup + ParentEduc +
                   LunchType + TestPrep + ParentMaritalStatus +
                   PracticeSport + IsFirstChild + NrSiblings +
                   TransportMeans + WklyStudyHours +
                   Gender:LunchType + Gender:PracticeSport +
                   EthnicGroup:ParentEduc + EthnicGroup:IsFirstChild +
                   ParentEduc:TestPrep + ParentEduc:ParentMaritalStatus +
                   ParentEduc:PracticeSport + ParentEduc:IsFirstChild +
                   LunchType:PracticeSport + LunchType:TransportMeans +
                   TestPrep:WklyStudyHours + ParentMaritalStatus:PracticeSport + ParentMaritalStatus:Isl
                   data = X_train)
# cv
cv_model <- cv.glmnet(X_math, Y_math_train, alpha = 1)
best_lambda <- cv_model$lambda.min
lasso_model <- glmnet(X_math, Y_math_train, alpha = 1, lambda = best_lambda)
```

```r
coef(lasso_model)
```

```
## 133 x 1 sparse Matrix of class "dgCMatrix"
##                                                          s0
## (Intercept)                                      58.72876548
## (Intercept)                                       .
## Gendermale                                        2.83505490
## EthnicGroupgroup B                                .
## EthnicGroupgroup C                                .
## EthnicGroupgroup D                                .
## EthnicGroupgroup E                                5.79861629
## ParentEducbachelor's degree                       .
## ParentEduchigh school                            -1.27311272
## ParentEducmaster's degree                         .
## ParentEducsome college                            .
## ParentEducsome high school                       -1.81196054
## LunchTypestandard                                 8.56139638
## TestPrepnone                                     -4.56439629
## ParentMaritalStatusmarried                        .
## ParentMaritalStatusnone                           .
## ParentMaritalStatussingle                         .
## ParentMaritalStatuswidowed                        .
## PracticeSportregularly                            .
## PracticeSportsometimes                            .
## IsFirstChildyes                                   .
## NrSiblings1                                       .
## NrSiblings2                                       .
## NrSiblings3                                       .
## NrSiblings4                                       .
## NrSiblings5                                       .
## NrSiblings6                                       .
## NrSiblings7                                       .
## TransportMeansschool_bus                          .
## WklyStudyHours> 10                                .
## WklyStudyHours10-May                              0.04365133
## Gendermale:LunchTypestandard                      .
## Gendermale:PracticeSportregularly                 1.30000312
## Gendermale:PracticeSportsometimes                 .
## EthnicGroupgroup B:ParentEducbachelor's degree    .
## EthnicGroupgroup C:ParentEducbachelor's degree    .
## EthnicGroupgroup D:ParentEducbachelor's degree    .
## EthnicGroupgroup E:ParentEducbachelor's degree    .
## EthnicGroupgroup B:ParentEduchigh school         -4.26752647
## EthnicGroupgroup C:ParentEduchigh school          .
## EthnicGroupgroup D:ParentEduchigh school          .
## EthnicGroupgroup E:ParentEduchigh school          .
## EthnicGroupgroup B:ParentEducmaster's degree      0.37864772
## EthnicGroupgroup C:ParentEducmaster's degree      .
## EthnicGroupgroup D:ParentEducmaster's degree      4.71152500
## EthnicGroupgroup E:ParentEducmaster's degree      .
## EthnicGroupgroup B:ParentEducsome college         .
## EthnicGroupgroup C:ParentEducsome college         .
## EthnicGroupgroup D:ParentEducsome college         4.38132298
```

```
## EthnicGroupgroup E:ParentEducsome college               .
## EthnicGroupgroup B:ParentEducsome high school          -2.35204701
## EthnicGroupgroup C:ParentEducsome high school          -2.24203316
## EthnicGroupgroup D:ParentEducsome high school           .
## EthnicGroupgroup E:ParentEducsome high school           2.54296473
## EthnicGroupgroup B:IsFirstChildyes                      .
## EthnicGroupgroup C:IsFirstChildyes                      .
## EthnicGroupgroup D:IsFirstChildyes                      .
## EthnicGroupgroup E:IsFirstChildyes                      .
## ParentEducbachelor's degree:TestPrepnone                .
## ParentEduchigh school:TestPrepnone                     -0.53694610
## ParentEducmaster's degree:TestPrepnone                  .
## ParentEducsome college:TestPrepnone                     .
## ParentEducsome high school:TestPrepnone                 .
## ParentEducbachelor's degree:ParentMaritalStatusmarried  .
## ParentEduchigh school:ParentMaritalStatusmarried        .
## ParentEducmaster's degree:ParentMaritalStatusmarried    .
## ParentEducsome college:ParentMaritalStatusmarried       .
## ParentEducsome high school:ParentMaritalStatusmarried   .
## ParentEducbachelor's degree:ParentMaritalStatusnone    -4.08240978
## ParentEduchigh school:ParentMaritalStatusnone          -1.51473842
## ParentEducmaster's degree:ParentMaritalStatusnone       .
## ParentEducsome college:ParentMaritalStatusnone          .
## ParentEducsome high school:ParentMaritalStatusnone      .
## ParentEducbachelor's degree:ParentMaritalStatussingle   .
## ParentEduchigh school:ParentMaritalStatussingle         0.05244680
## ParentEducmaster's degree:ParentMaritalStatussingle     .
## ParentEducsome college:ParentMaritalStatussingle       -4.39926620
## ParentEducsome high school:ParentMaritalStatussingle    .
## ParentEducbachelor's degree:ParentMaritalStatuswidowed  5.82516248
## ParentEduchigh school:ParentMaritalStatuswidowed        .
## ParentEducmaster's degree:ParentMaritalStatuswidowed    .
## ParentEducsome college:ParentMaritalStatuswidowed       6.82700370
## ParentEducsome high school:ParentMaritalStatuswidowed   .
## ParentEducbachelor's degree:PracticeSportregularly      7.08378357
## ParentEduchigh school:PracticeSportregularly            .
## ParentEducmaster's degree:PracticeSportregularly       -0.86538663
## ParentEducsome college:PracticeSportregularly          -0.81980287
## ParentEducsome high school:PracticeSportregularly       .
## ParentEducbachelor's degree:PracticeSportsometimes      .
## ParentEduchigh school:PracticeSportsometimes            .
## ParentEducmaster's degree:PracticeSportsometimes        2.03284914
## ParentEducsome college:PracticeSportsometimes           .
## ParentEducsome high school:PracticeSportsometimes       .
## ParentEducbachelor's degree:IsFirstChildyes             .
## ParentEduchigh school:IsFirstChildyes                   .
## ParentEducmaster's degree:IsFirstChildyes               .
## ParentEducsome college:IsFirstChildyes                  .
## ParentEducsome high school:IsFirstChildyes              .
## LunchTypestandard:PracticeSportregularly                .
## LunchTypestandard:PracticeSportsometimes                2.41706843
## LunchTypestandard:TransportMeansschool_bus              .
## TestPrepnone:WklyStudyHours> 10                         .
## TestPrepnone:WklyStudyHours10-May                       .
```

```
## ParentMaritalStatusmarried:PracticeSportregularly       0.98955649
## ParentMaritalStatusnone:PracticeSportregularly          .
## ParentMaritalStatussingle:PracticeSportregularly        .
## ParentMaritalStatuswidowed:PracticeSportregularly       .
## ParentMaritalStatusmarried:PracticeSportsometimes       .
## ParentMaritalStatusnone:PracticeSportsometimes          .
## ParentMaritalStatussingle:PracticeSportsometimes        .
## ParentMaritalStatuswidowed:PracticeSportsometimes       .
## ParentMaritalStatusmarried:IsFirstChildyes              .
## ParentMaritalStatusnone:IsFirstChildyes                 .
## ParentMaritalStatussingle:IsFirstChildyes               0.53533289
## ParentMaritalStatuswidowed:IsFirstChildyes              .
## ParentMaritalStatusmarried:TransportMeansschool_bus     2.17345991
## ParentMaritalStatusnone:TransportMeansschool_bus        .
## ParentMaritalStatussingle:TransportMeansschool_bus      .
## ParentMaritalStatuswidowed:TransportMeansschool_bus     .
## PracticeSportregularly:WklyStudyHours> 10               .
## PracticeSportsometimes:WklyStudyHours> 10               .
## PracticeSportregularly:WklyStudyHours10-May             2.96677716
## PracticeSportsometimes:WklyStudyHours10-May             .
## IsFirstChildyes:NrSiblings1                             .
## IsFirstChildyes:NrSiblings2                             .
## IsFirstChildyes:NrSiblings3                             .
## IsFirstChildyes:NrSiblings4                             .
## IsFirstChildyes:NrSiblings5                             1.17886779
## IsFirstChildyes:NrSiblings6                             0.58565614
## IsFirstChildyes:NrSiblings7                             .
## IsFirstChildyes:TransportMeansschool_bus                .
## IsFirstChildyes:WklyStudyHours> 10                      2.36047609
## IsFirstChildyes:WklyStudyHours10-May                    .
```

```
model_math_best = lm(Y_math_train ~  Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + ParentMa
```

```
# reading LASSO

X_reading <- model.matrix(~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours +
    Gender:IsFirstChild + LunchType:PracticeSport + LunchType:IsFirstChild +
    TestPrep:NrSiblings + TestPrep:TransportMeans + ParentMaritalStatus:PracticeSport + ParentMaritalSta

# cv
cv_model <- cv.glmnet(X_reading, Y_reading_train, alpha = 1)
best_lambda <- cv_model$lambda.min
lasso_model <- glmnet(X_reading, Y_reading_train, alpha = 1, lambda = best_lambda)
coef(lasso_model)
```

```
## 73 x 1 sparse Matrix of class "dgCMatrix"
##                                                       s0
## (Intercept)                                   69.8355115
## (Intercept)                                    .
## Gendermale                                    -6.8808132
## EthnicGroupgroup B                             .
## EthnicGroupgroup C                             .
## EthnicGroupgroup D                             1.5614107
```

```
## EthnicGroupgroup E                                    3.0720371
## ParentEducbachelor's degree                           1.1808026
## ParentEduchigh school                                -3.6896705
## ParentEducmaster's degree                             3.4044096
## ParentEducsome college                               -0.8911333
## ParentEducsome high school                           -3.4545397
## LunchTypestandard                                     5.2741723
## TestPrepnone                                         -6.1721822
## ParentMaritalStatusmarried                            1.7869496
## ParentMaritalStatusnone                               .
## ParentMaritalStatussingle                             .
## ParentMaritalStatuswidowed                            .
## PracticeSportregularly                                .
## PracticeSportsometimes                                .
## IsFirstChildyes                                       0.5528251
## NrSiblings1                                           .
## NrSiblings2                                           .
## NrSiblings3                                           .
## NrSiblings4                                           .
## NrSiblings5                                           .
## NrSiblings6                                           .
## NrSiblings7                                           .
## TransportMeansschool_bus                              0.8420014
## WklyStudyHours> 10                                    .
## WklyStudyHours10-May                                  .
## Gendermale:IsFirstChildyes                            .
## LunchTypestandard:PracticeSportregularly              .
## LunchTypestandard:PracticeSportsometimes              1.7542088
## LunchTypestandard:IsFirstChildyes                     .
## TestPrepnone:NrSiblings1                              .
## TestPrepnone:NrSiblings2                              .
## TestPrepnone:NrSiblings3                              .
## TestPrepnone:NrSiblings4                              .
## TestPrepnone:NrSiblings5                             -0.5200176
## TestPrepnone:NrSiblings6                              .
## TestPrepnone:NrSiblings7                              .
## TestPrepnone:TransportMeansschool_bus                 .
## ParentMaritalStatusmarried:PracticeSportregularly     0.4986706
## ParentMaritalStatusnone:PracticeSportregularly        .
## ParentMaritalStatussingle:PracticeSportregularly     -0.9848865
## ParentMaritalStatuswidowed:PracticeSportregularly     .
## ParentMaritalStatusmarried:PracticeSportsometimes     .
## ParentMaritalStatusnone:PracticeSportsometimes        .
## ParentMaritalStatussingle:PracticeSportsometimes      .
## ParentMaritalStatuswidowed:PracticeSportsometimes     1.3236864
## ParentMaritalStatusmarried:IsFirstChildyes            .
## ParentMaritalStatusnone:IsFirstChildyes               .
## ParentMaritalStatussingle:IsFirstChildyes             0.8768162
## ParentMaritalStatuswidowed:IsFirstChildyes            .
## PracticeSportregularly:WklyStudyHours> 10             .
## PracticeSportsometimes:WklyStudyHours> 10             .
## PracticeSportregularly:WklyStudyHours10-May           2.0636849
## PracticeSportsometimes:WklyStudyHours10-May           .
## NrSiblings1:WklyStudyHours> 10                        .
```

```
## NrSiblings2:WklyStudyHours> 10                    0.6746785
## NrSiblings3:WklyStudyHours> 10                    .
## NrSiblings4:WklyStudyHours> 10                    .
## NrSiblings5:WklyStudyHours> 10                    .
## NrSiblings6:WklyStudyHours> 10                    .
## NrSiblings7:WklyStudyHours> 10                    .
## NrSiblings1:WklyStudyHours10-May                  .
## NrSiblings2:WklyStudyHours10-May                  .
## NrSiblings3:WklyStudyHours10-May                  0.8601131
## NrSiblings4:WklyStudyHours10-May                  1.2404940
## NrSiblings5:WklyStudyHours10-May                  .
## NrSiblings6:WklyStudyHours10-May                  .
## NrSiblings7:WklyStudyHours10-May                  .
```

```
model_reading_best = lm(Y_reading_train ~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours + LunchType:PracticeSport + ParentMarita
```

```
X_writing <- model.matrix(~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours +
    ParentEduc:IsFirstChild + LunchType:PracticeSport + LunchType:IsFirstChild +
    TestPrep:NrSiblings + ParentMaritalStatus:PracticeSport +
    ParentMaritalStatus:IsFirstChild + PracticeSport:WklyStudyHours +
    IsFirstChild:WklyStudyHours, data = X_train)

# cv
cv_model <- cv.glmnet(X_writing, Y_writing_train, alpha = 1)
best_lambda <- cv_model$lambda.min
lasso_model <- glmnet(X_writing, Y_writing_train, alpha = 1, lambda = best_lambda)
coef(lasso_model)
```

```
## 64 x 1 sparse Matrix of class "dgCMatrix"
##                                                      s0
## (Intercept)                                  70.6829405
## (Intercept)                                   .
## Gendermale                                   -9.0596356
## EthnicGroupgroup B                           -0.8604185
## EthnicGroupgroup C                            .
## EthnicGroupgroup D                            3.7618498
## EthnicGroupgroup E                            2.4755467
## ParentEducbachelor's degree                   1.9470067
## ParentEduchigh school                        -5.3944225
## ParentEducmaster's degree                     5.5773248
## ParentEducsome college                       -2.3933416
## ParentEducsome high school                   -5.3377261
## LunchTypestandard                             6.0308668
## TestPrepnone                                 -9.1017130
## ParentMaritalStatusmarried                    2.2355329
## ParentMaritalStatusnone                       .
## ParentMaritalStatussingle                     .
## ParentMaritalStatuswidowed                    0.2953984
## PracticeSportregularly                        .
## PracticeSportsometimes                        .
## IsFirstChildyes                               .
```

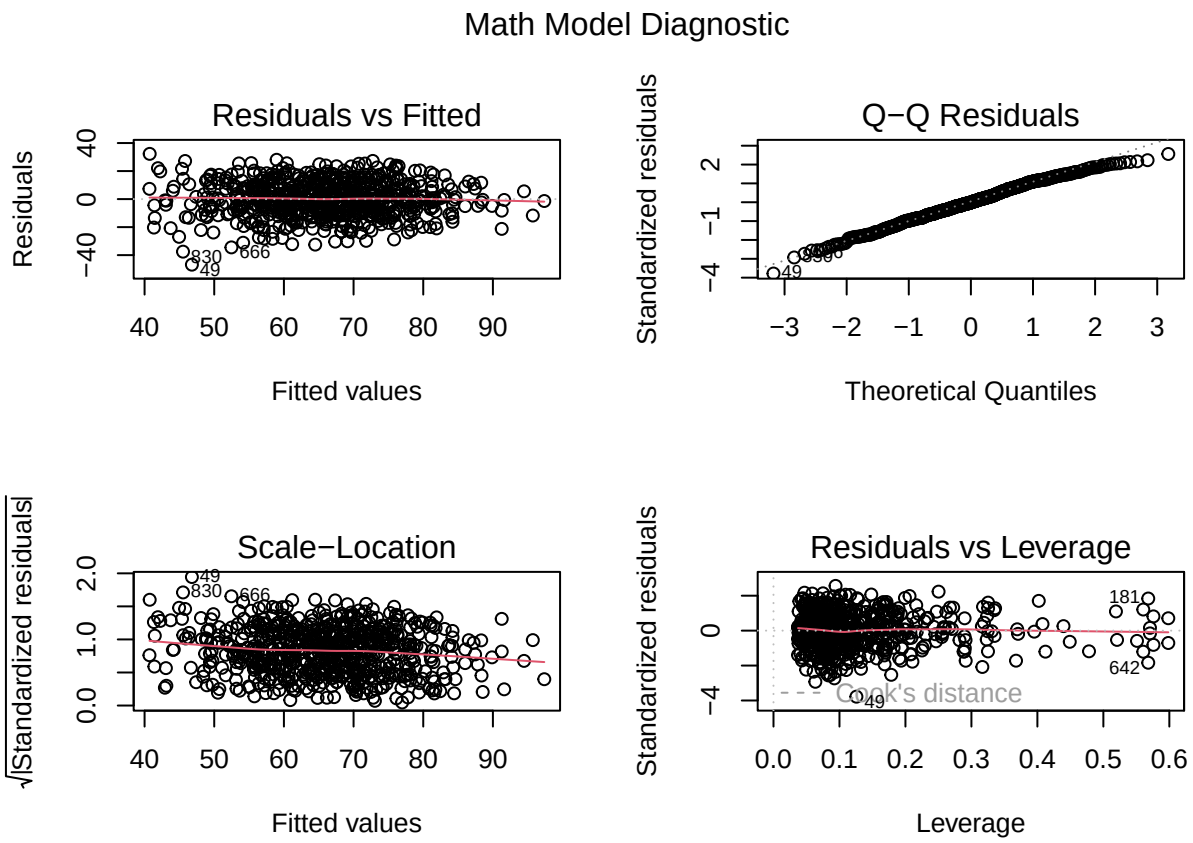
```
## NrSiblings1                                          .
## NrSiblings2                                          .
## NrSiblings3                                          .
## NrSiblings4                                          .
## NrSiblings5                                          .
## NrSiblings6                                          .
## NrSiblings7                                          1.9601335
## TransportMeansschool_bus                             1.2023389
## WklyStudyHours> 10                                   .
## WklyStudyHours10-May                                 0.3212997
## ParentEducbachelor's degree:IsFirstChildyes          .
## ParentEduchigh school:IsFirstChildyes                .
## ParentEducmaster's degree:IsFirstChildyes            .
## ParentEducsome college:IsFirstChildyes               1.8421498
## ParentEducsome high school:IsFirstChildyes           .
## LunchTypestandard:PracticeSportregularly             .
## LunchTypestandard:PracticeSportsometimes             2.6494205
## LunchTypestandard:IsFirstChildyes                    .
## TestPrepnone:NrSiblings1                            -0.3741817
## TestPrepnone:NrSiblings2                             .
## TestPrepnone:NrSiblings3                             .
## TestPrepnone:NrSiblings4                            -0.4739880
## TestPrepnone:NrSiblings5                            -1.2460926
## TestPrepnone:NrSiblings6                             2.1969158
## TestPrepnone:NrSiblings7                             .
## ParentMaritalStatusmarried:PracticeSportregularly    1.9239985
## ParentMaritalStatusnone:PracticeSportregularly      -1.9169982
## ParentMaritalStatussingle:PracticeSportregularly    -0.5778886
## ParentMaritalStatuswidowed:PracticeSportregularly    .
## ParentMaritalStatusmarried:PracticeSportsometimes    .
## ParentMaritalStatusnone:PracticeSportsometimes       .
## ParentMaritalStatussingle:PracticeSportsometimes     .
## ParentMaritalStatuswidowed:PracticeSportsometimes    1.4693036
## ParentMaritalStatusmarried:IsFirstChildyes           .
## ParentMaritalStatusnone:IsFirstChildyes              .
## ParentMaritalStatussingle:IsFirstChildyes            1.2289395
## ParentMaritalStatuswidowed:IsFirstChildyes           0.5142420
## PracticeSportregularly:WklyStudyHours> 10            .
## PracticeSportsometimes:WklyStudyHours> 10            .
## PracticeSportregularly:WklyStudyHours10-May          3.4575598
## PracticeSportsometimes:WklyStudyHours10-May          .
## IsFirstChildyes:WklyStudyHours> 10                   1.1518098
## IsFirstChildyes:WklyStudyHours10-May                 .
```

```
model_writing_best = lm(Y_writing_train ~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours +
    ParentEduc:IsFirstChild + LunchType:PracticeSport +
    TestPrep:NrSiblings + ParentMaritalStatus:PracticeSport +
    ParentMaritalStatus:IsFirstChild + PracticeSport:WklyStudyHours +
    IsFirstChild:WklyStudyHours, data = X_train)
```

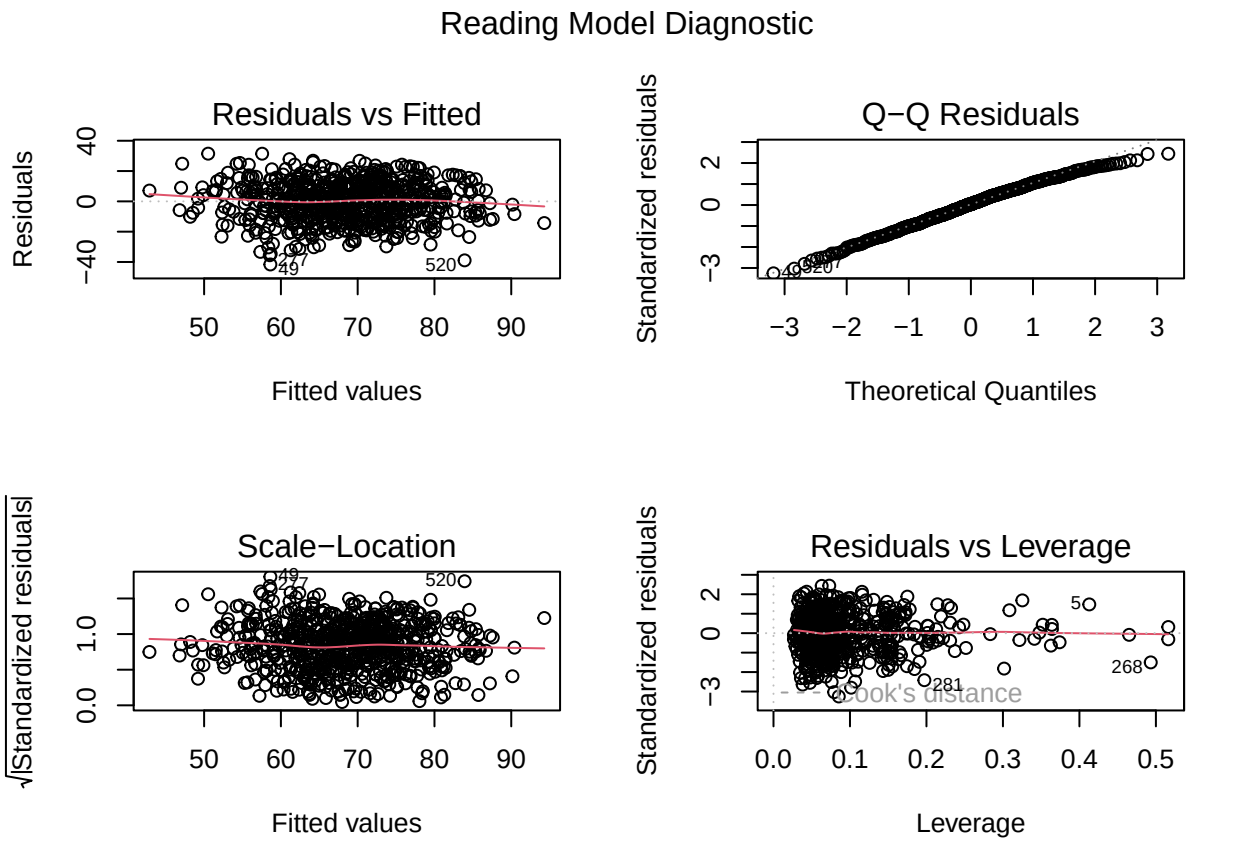
However, the initial process leaving a considerable number of variables, we applied the LASSO (Least Absolute Shrinkage and Selection Operator) method for penalization. Utilizing cross-validation (cv), we identified the optimal lambda value. Subsequently, all interaction terms with shrinkage coefficients (s0) below 0.5 were

eliminated. This refined approach resulted in the derivation of three models that were not only more efficient but also nested.

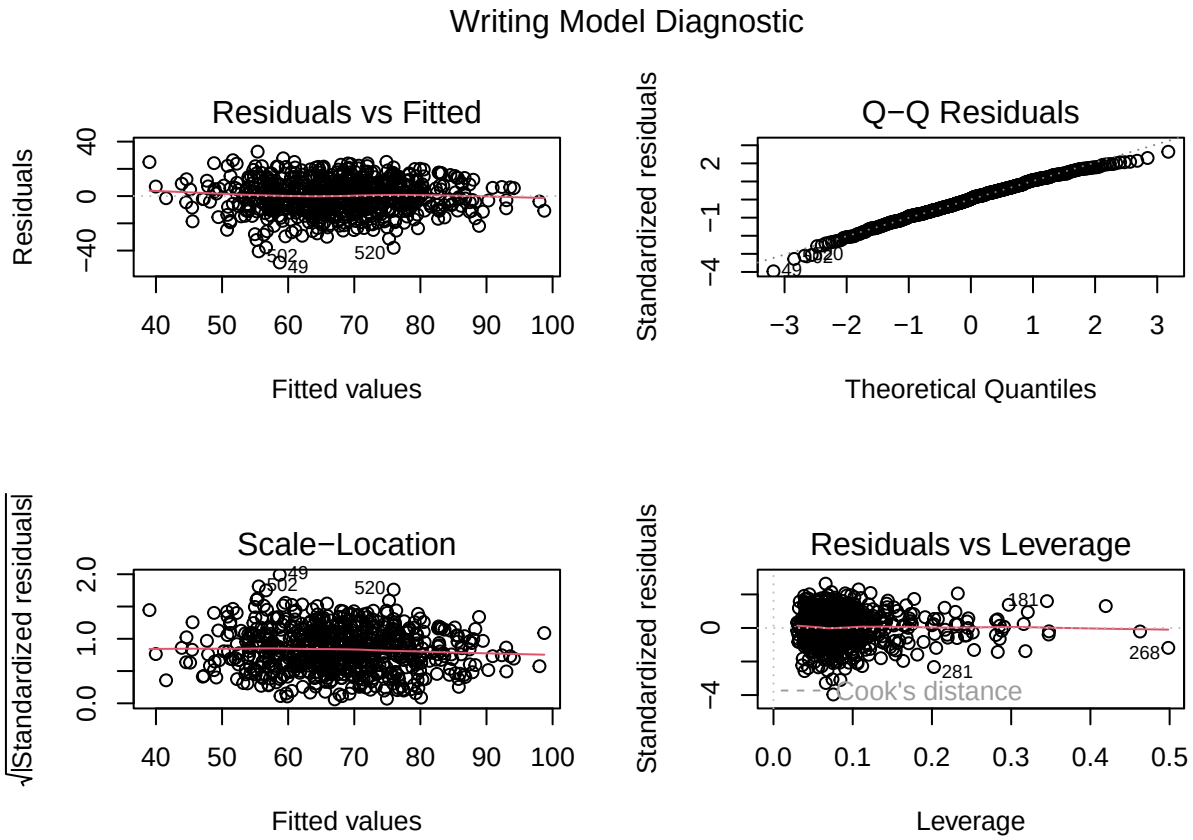
```
par(mfrow=c(2,2))
plot(model_math_best)
mtext("Math Model Diagnostic", outer = TRUE, cex = 1, line = -1)
```




```
plot(model_reading_best)
mtext("Reading Model Diagnostic", outer = TRUE, cex = 1, line = -1)
```



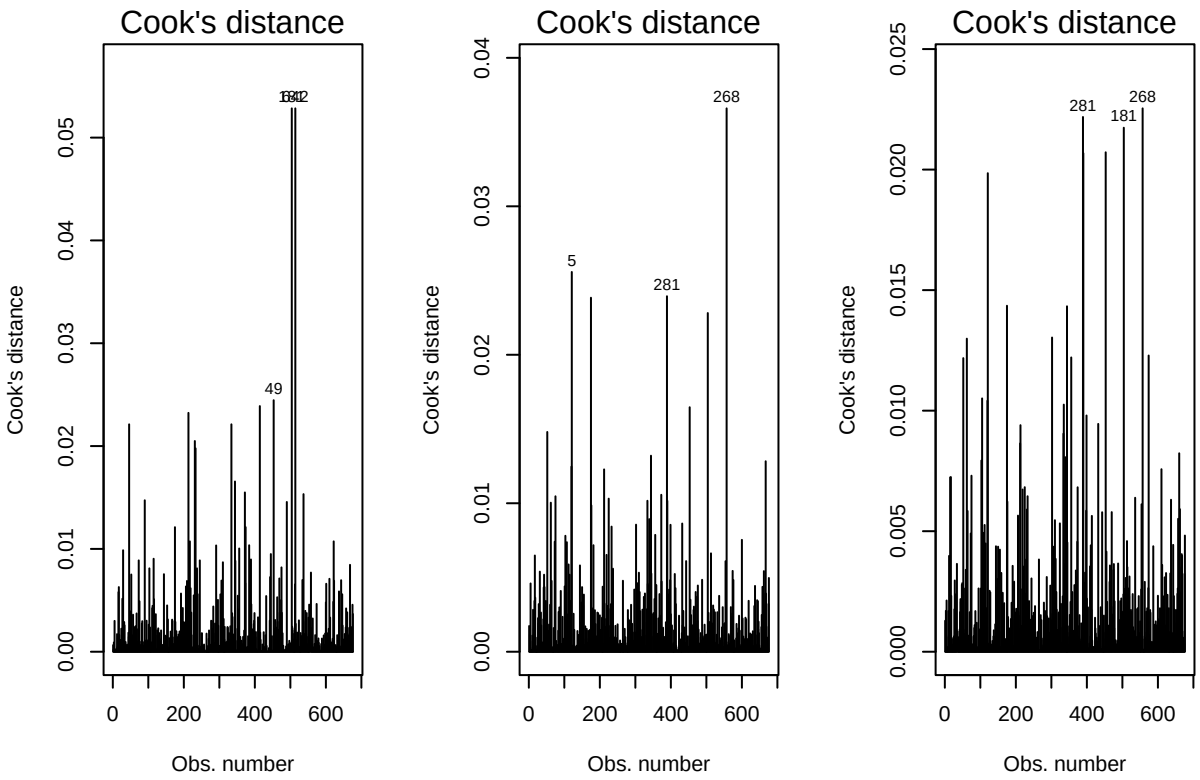

```
plot(model_writing_best)
mtext("Writing Model Diagnostic", outer = TRUE, cex = 1, line = -1)
```



In the diagnostic analysis of our linear regression model, the Residuals versus Fitted values plot exhibited a stochastic distribution of residuals, devoid of any systematic patterns, thereby conforming to the assumptions of homoscedasticity and linearity. The Quantile-Quantile (QQ) plot demonstrated a close alignment of residuals with the theoretical normal distribution, as evidenced by the linear arrangement of data points. Furthermore, the Scale-Location plot revealed a uniform dispersion of residuals around a central horizontal axis, indicative of consistent variance across the spectrum of fitted values. Finally, the examination of the Residuals versus Leverage plot revealed an absence of high-leverage observations, thus suggesting that the model is not unduly influenced by outlier data points.

Influential observations

```
par(mfrow=c(1,3))
plot(model_math_best, which = 4)
plot(model_reading_best, which = 4)
plot(model_writing_best, which = 4)
```



From the analysis of the plots, we identified a few points that appeared to be potential outliers or high-influence observations. However, upon examination, the Cook's distance values for these points were not significantly large. Additionally, when these points were excluded and the model was re-estimated, there was no substantial change in the model's performance. Upon further investigation of these specific data points, no anomalies were detected. Consequently, the final model was retained with these data points included.

multicolinearity-

```
vif_values_math <- vif(model_math_best , type = 'predictor')
print(vif_values_math)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## Gender                1.542005e+02  5        1.655040
## EthnicGroup           4.185669e+07 29        1.353349
## ParentEduc            2.600420e+04 65        1.081339
## LunchType             1.433646e+02  5        1.643025
## TestPrep              1.154486e+00  1        1.074470
## ParentMaritalStatus 2.805551e+08 34        1.331176
## PracticeSport         5.013426e+07 29        1.357566
## TransportMeans        1.794224e+03  9        1.516250
## WklyStudyHours        1.477385e+02  8        1.366449
##                                                  Interacts With
## Gender                                            PracticeSport
## EthnicGroup                                          ParentEduc
## ParentEduc          EthnicGroup, ParentMaritalStatus, PracticeSport
## LunchType                                         PracticeSport
## TestPrep                                                     --
## ParentMaritalStatus                        ParentEduc, TransportMeans
## PracticeSport          Gender, ParentEduc, LunchType, WklyStudyHours
```

```
## TransportMeans                                    ParentMaritalStatus
## WklyStudyHours                                         PracticeSport
##
## Gender                              EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus,
## EthnicGroup                           Gender, LunchType, TestPrep, ParentMaritalStatus, PracticeSport,
## ParentEduc                                                                Gender, LunchType, TestPrep,
## LunchType                              Gender, EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus,
## TestPrep             Gender, EthnicGroup, ParentEduc, LunchType, ParentMaritalStatus, PracticeSport,
## ParentMaritalStatus                                          Gender, EthnicGroup, LunchType, TestPrep,
## PracticeSport                                                        EthnicGroup, TestPrep, Parent
## TransportMeans                                    Gender, EthnicGroup, ParentEduc, LunchType, TestPrep,
## WklyStudyHours                                 Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, Parent
```

```
vif_values_writing <- vif(model_writing_best, type = 'predictor')
print(vif_values_writing)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## Gender                1.117122e+00  1        1.056940
## EthnicGroup           1.504424e+00  4        1.052377
## ParentEduc            2.965818e+01 11        1.166583
## LunchType             1.699906e+02  5        1.671254
## TestPrep              2.845571e+00 15        1.035473
## ParentMaritalStatus   8.226773e+02 19        1.193209
## PracticeSport         6.381940e+03 23        1.209808
## IsFirstChild          1.080258e+06 23        1.352582
## NrSiblings            2.845571e+00 15        1.035473
## TransportMeans        1.086213e+00  1        1.042216
## WklyStudyHours        2.384397e+03 11        1.424024
##                                                  Interacts With
## Gender                                                       --
## EthnicGroup                                                  --
## ParentEduc                                          IsFirstChild
## LunchType                                          PracticeSport
## TestPrep                                             NrSiblings
## ParentMaritalStatus                     PracticeSport, IsFirstChild
## PracticeSport          LunchType, ParentMaritalStatus, WklyStudyHours
## IsFirstChild          ParentEduc, ParentMaritalStatus, WklyStudyHours
## NrSiblings                                                 TestPrep
## TransportMeans                                               --
## WklyStudyHours                       PracticeSport, IsFirstChild
##
## Gender                EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport
## EthnicGroup                Gender, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport
## ParentEduc                      Gender, EthnicGroup, LunchType, TestPrep, ParentMaritalStatus,
## LunchType                       Gender, EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus
## TestPrep                  Gender, EthnicGroup, ParentEduc, LunchType, ParentMaritalStatus, Pr
## ParentMaritalStatus                                  Gender, EthnicGroup, ParentEduc, Lunch
## PracticeSport                                                  Gender, EthnicGroup, Parer
## IsFirstChild                                                   Gender, EthnicGroup, Lunch
## NrSiblings                Gender, EthnicGroup, ParentEduc, LunchType, ParentMaritalStatus, Pr
## TransportMeans         Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus
## WklyStudyHours                                  Gender, EthnicGroup, ParentEduc, LunchType,
```

```
vif_values_reading <- vif(model_reading_best, type = 'predictor')
print(vif_values_reading)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Gender                 NA  1              NA
## EthnicGroup            NA  4              NA
## ParentEduc             NA  5              NA
## LunchType              NA  5              NA
## TestPrep               NA  1              NA
## ParentMaritalStatus    NA 19              NA
## PracticeSport          NA 23              NA
## IsFirstChild           NA  9              NA
## NrSiblings             NA 23              NA
## TransportMeans         NA  1              NA
## WklyStudyHours         NA 29              NA
##                                                Interacts With
## Gender                                                     --
## EthnicGroup                                                --
## ParentEduc                                                 --
## LunchType                                         PracticeSport
## TestPrep                                                   --
## ParentMaritalStatus                    PracticeSport, IsFirstChild
## PracticeSport       LunchType, ParentMaritalStatus, WklyStudyHours
## IsFirstChild                                  ParentMaritalStatus
## NrSiblings                                        WklyStudyHours
## TransportMeans                                             --
## WklyStudyHours                          PracticeSport, NrSiblings
##
## Gender              EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport
## EthnicGroup              Gender, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport
## ParentEduc              Gender, EthnicGroup, LunchType, TestPrep, ParentMaritalStatus, PracticeSport
## LunchType                          Gender, EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus
## TestPrep              Gender, EthnicGroup, ParentEduc, LunchType, ParentMaritalStatus, PracticeSport
## ParentMaritalStatus                                          Gender, EthnicGroup, ParentEduc, Lunch
## PracticeSport                                                           Gender, EthnicGroup, Pare
## IsFirstChild                                  Gender, EthnicGroup, ParentEduc, LunchType, TestPrep,
## NrSiblings                              Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, Parentl
## TransportMeans             Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus
## WklyStudyHours                                      Gender, EthnicGroup, ParentEduc, LunchType, Te
```

model validation

cross validation

```
## Linear Regression
##
## 676 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 609, 608, 608, 608, 608, 610, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   14.34509  0.2210548  11.58918
##
```

```
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 676 samples
##  11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 609, 609, 606, 609, 609, 607, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   14.09604  0.1709726  11.52002
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 676 samples
##  11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 608, 608, 609, 608, 608, 610, ...
## Resampling results:
##
##   RMSE    Rsquared  MAE
##   13.261  0.289927  10.68384
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
library(readr)
library(dplyr)
library(ggplot2)
library(caret)
library(purrr)
library(tidyverse)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(modelr)
library(tidyr)
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:gridExtra':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(boot)
```

```
##
## Attaching package: 'boot'
## The following object is masked from 'package:lattice':
##
##     melanoma
## The following object is masked from 'package:car':
##
##     logit
```

```
library(patchwork)

set.seed(123)
# generate a cv dataframe
cv_df_math =
  crossv_mc(math_model_data, 10) %>%
  mutate(
    train = map(train, as_tibble),
    test = map(test, as_tibble))

# fit the model to the generated CV dataframe
cv_df_math =
  cv_df_math |>
  mutate(
    model  = map(train, ~lm( Y_math_train ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + I
    data = math_model_data)),
    rmse = map2_dbl(model, test, ~rmse(model = .x, data = .y)))

# plot the prediction error
plot_math <- cv_df_math |>
  select(rmse) |>
  pivot_longer(
    everything(),
    names_to = "model",
    values_to = "rmse") %>%
```

```
  ggplot(aes(x = model, y = rmse)) +
  geom_violin(fill = "blue", alpha = 0.5) +
  labs(
    x = "MLR",
    y = "Prediction Errors"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(color = "grey20"),
    axis.title = element_text(color = "grey20")
  )


set.seed(123)
# generate a cv dataframe
cv_df_reading =
  crossv_mc(reading_model_data, 10) %>%
  mutate(
    train = map(train, as_tibble),
    test = map(test, as_tibble))

# fit the model to the generated CV dataframe
cv_df_reading =
  cv_df_reading |>
  mutate(
    model  = map(train, ~lm(Y_reading_train ~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours + LunchType:PracticeSport + ParentMarita
    rmse = map2_dbl(model, test, ~rmse(model = .x, data = .y)))
```

```
## Warning: There were 10 warnings in `mutate()`.
## The first warning was:
## i In argument: `rmse = map2_dbl(model, test, ~rmse(model = .x, data = .y))`.
## Caused by warning in `predict.lm()`:
## ! prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
## i Run `dplyr::last_dplyr_warnings()` to see the 9 remaining warnings.
```

```
# plot the prediction error
plot_reading <- cv_df_reading |>
  select(rmse) |>
  pivot_longer(
    everything(),
    names_to = "model",
    values_to = "rmse") %>%
  ggplot(aes(x = model, y = rmse)) +
  geom_violin(fill = "pink", alpha = 0.5) +
  labs(
    x = "MLR",
    y = "Prediction Errors"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
```

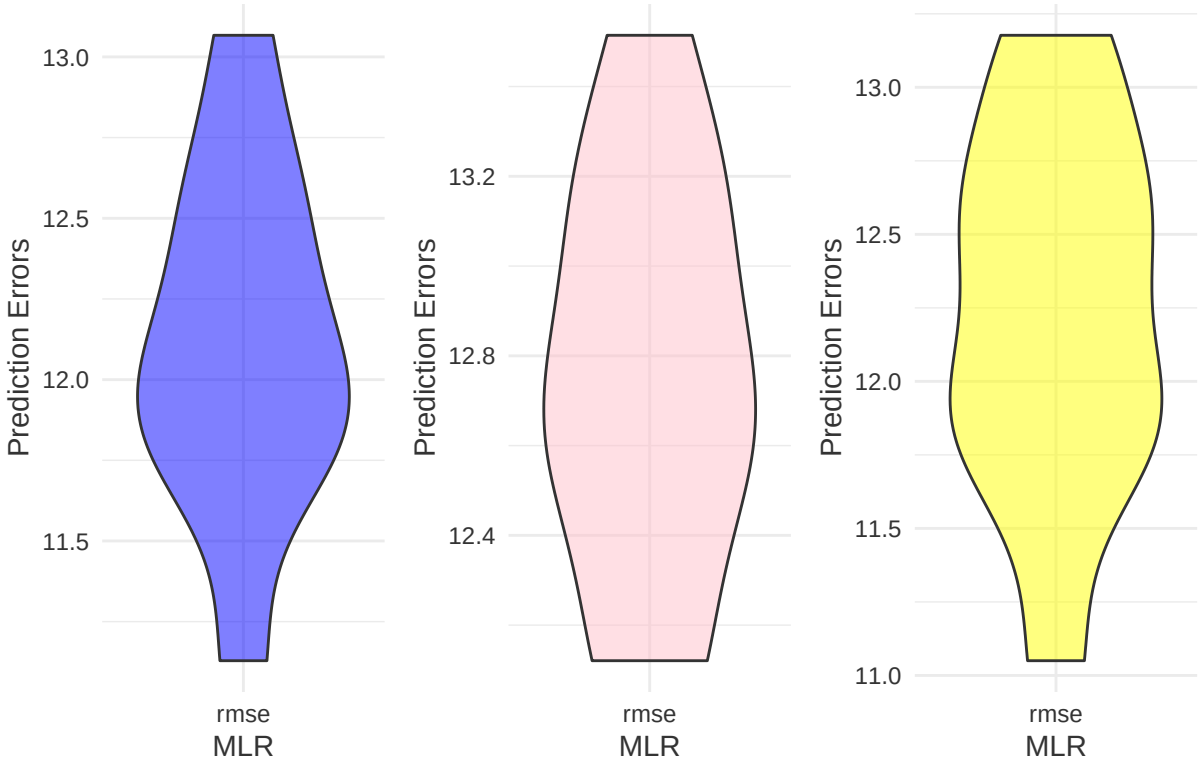
```
    axis.text = element_text(color = "grey20"),
    axis.title = element_text(color = "grey20")
  )

set.seed(123)
# generate a cv dataframe
cv_df_writing =
  crossv_mc(writing_model_data, 10) %>%
  mutate(
    train = map(train, as_tibble),
    test = map(test, as_tibble))

# fit the model to the generated CV dataframe
cv_df_writing =
  cv_df_writing |>
  mutate(
    model  = map(train, ~lm(Y_writing_train ~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours +
    ParentEduc:IsFirstChild + LunchType:PracticeSport +
    TestPrep:NrSiblings + ParentMaritalStatus:PracticeSport +
    ParentMaritalStatus:IsFirstChild + PracticeSport:WklyStudyHours +
    IsFirstChild:WklyStudyHours, data = writing_model_data)),
    rmse = map2_dbl(model, test, ~rmse(model = .x, data = .y)))

# plot the prediction error
plot_writing <-cv_df_writing |>
  select(rmse) |>
  pivot_longer(
    everything(),
    names_to = "model",
    values_to = "rmse") %>%
  ggplot(aes(x = model, y = rmse)) +
  geom_violin(fill = "yellow", alpha = 0.5) +
  labs(
    x = "MLR",
    y = "Prediction Errors"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_text(color = "grey20"),
    axis.title = element_text(color = "grey20")
  )

plot_math + plot_reading +
  plot_writing+plot_annotation(title="Prediction Errors For Models Under CV")
```



prediction

```
# Splitting the train dataset into independent variables (X) and dependent variables (Y)
X_test<- testData %>% select(-c(MathScore, ReadingScore, WritingScore))
Y_math_test <- testData$MathScore
Y_reading_test <-testData$ReadingScore
Y_writing_test <- testData$WritingScore
```

```
math_predictions <- predict(model_math_best, newdata = X_test)
reading_predictions <- predict(model_reading_best, newdata = X_test)
writing_predictions <- predict(model_writing_best, newdata = X_test)
```

```
math_mspe <- mean((Y_math_test - math_predictions)^2)
reading_mspe <- mean((Y_reading_test - reading_predictions)^2)
writing_mspe <- mean((Y_writing_test - writing_predictions)^2)
mspe_values <- data.frame(
  Subject = c("Math", "Reading", "Writing"),
  MSPE = c(math_mspe, reading_mspe, writing_mspe)
)
library(knitr)

kable(mspe_values, col.names = c("Subject", "MSPE"), caption = "MSPE Values for Different Subjects")
```

Table 9: MSPE Values for Different Subjects

Subject	MSPE
Math	198.3466
Reading	152.8595
Writing	146.8089