final project

2023-12-20

descriptive statistics

Distribution

```r
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
# Read the data
data <- read.csv("./Project_1_data.csv")
data[data == ""] <- NA
# 1. Descriptive statistics table for all variables
skimr::skim(data)
```

Table 1: Data summary

Name	data
Number of rows	948
Number of columns	14

Column type frequency:

character	10
numeric	4
Group variables	None

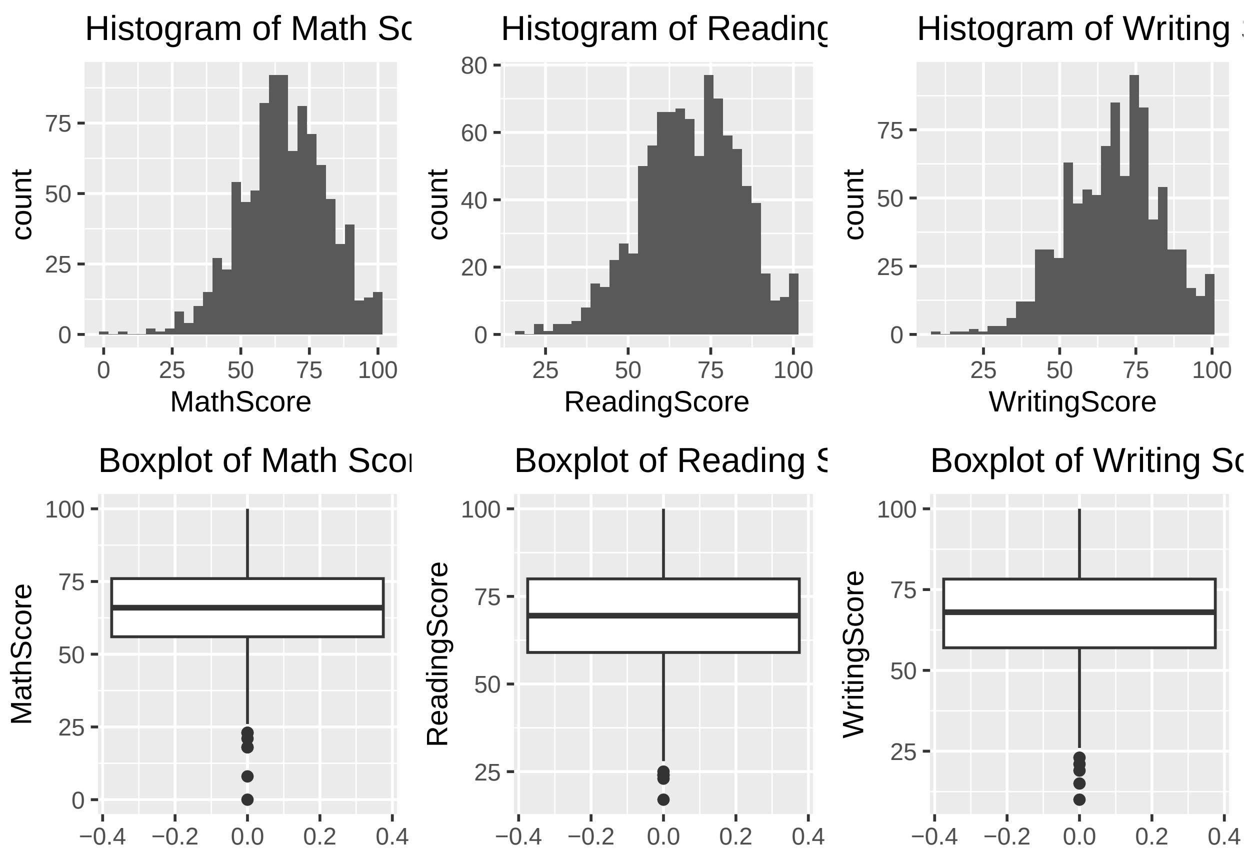
Variable type: character

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

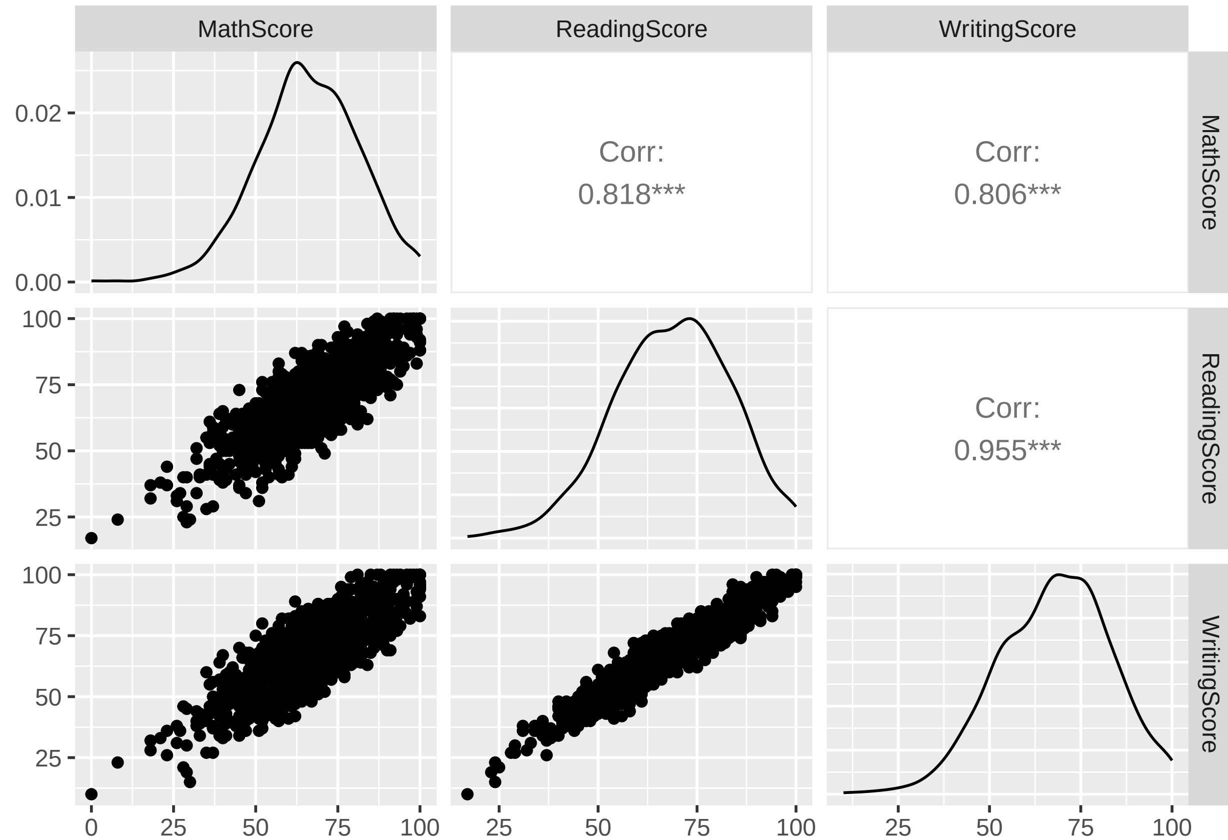
Variable type: numeric

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	

```
# 2. Explore distribution of results and consider potential transformations
# Histograms for continuous variables
hist_math <- ggplot(data, aes(x = MathScore)) + geom_histogram(bins = 30) + ggtitle("Histogram of Math S
hist_reading <- ggplot(data, aes(x = ReadingScore)) + geom_histogram(bins = 30) + ggtitle("Histogram of
hist_writing <- ggplot(data, aes(x = WritingScore)) + geom_histogram(bins = 30) + ggtitle("Histogram of

# Boxplots for continuous variables to check for outliers
box_math <- ggplot(data, aes(y = MathScore)) + geom_boxplot() + ggtitle("Boxplot of Math Scores")
box_reading <- ggplot(data, aes(y = ReadingScore)) + geom_boxplot() + ggtitle("Boxplot of Reading Score
box_writing <- ggplot(data, aes(y = WritingScore)) + geom_boxplot() + ggtitle("Boxplot of Writing Score

# Grid of plots
grid.arrange(hist_math, hist_reading, hist_writing, box_math, box_reading, box_writing, ncol = 3)
```




```r
# 3. Check for potential outliers or influential points
# Scatterplot matrix for continuous variables
ggpairs(data, columns = c("MathScore", "ReadingScore", "WritingScore"))
```



Missing Value

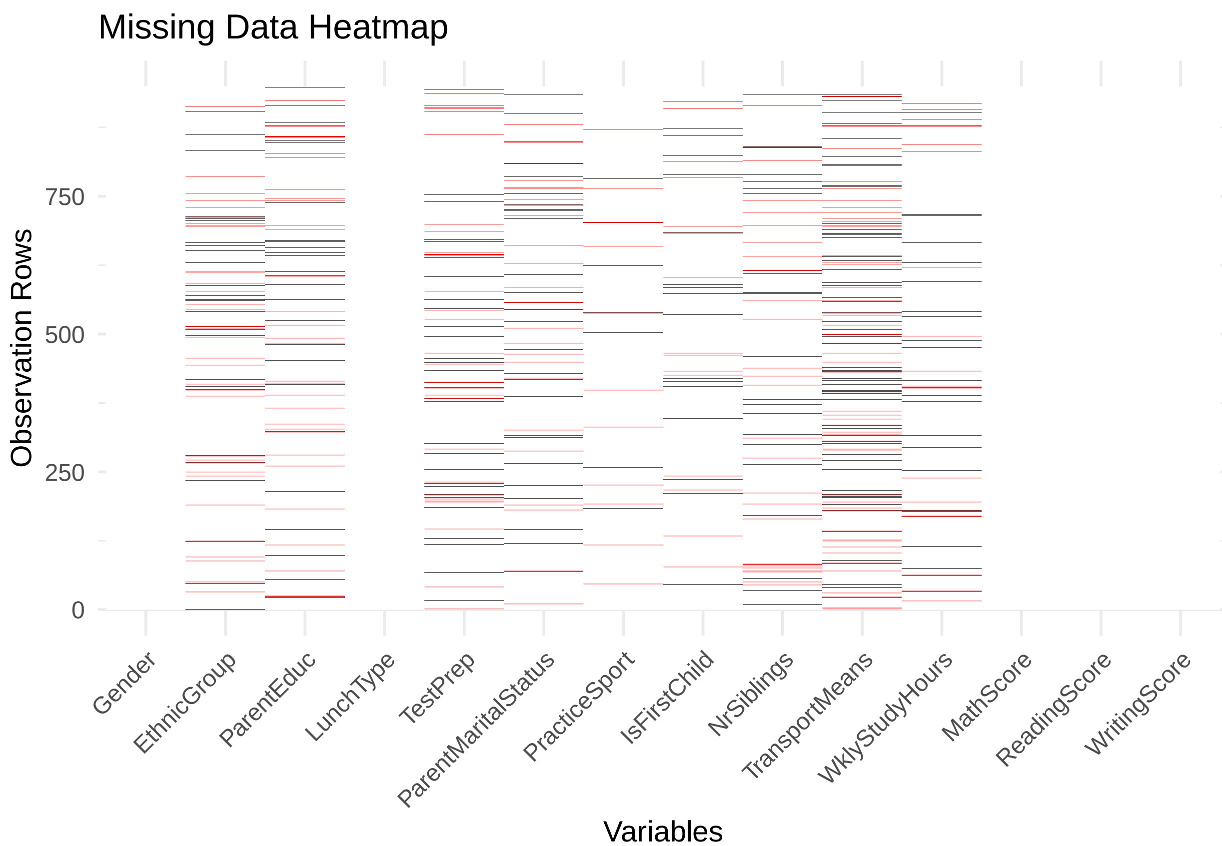
```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

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

```
# Creating a function to count NA and empty strings as missing values
count_missing <- function(x) sum(is.na(x) | x == "")
# Calculating the missing values
missing_values <- sapply(data, function(x) count_missing(x))

# Creating a dataframe for missing values
missing_data_frame <- data.frame(Variable = names(missing_values), MissingValues = missing_values)

# Convert empty strings to NA
data[data == ""] <- NA

# Melt the data for visualization
melted_data <- melt(data.frame(row = 1:nrow(data), data), id.vars = 'row')

# Creating the heatmap
ggplot(melted_data, aes(x = variable, y = row)) +
```

```r
  geom_tile(aes(fill = is.na(value))) +
  scale_fill_manual(values = c('white', 'red'), guide = FALSE) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = 'Variables', y = 'Observation Rows', title = 'Missing Data Heatmap')
```

```
## 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.
```




```r
missing_data_frame
```

```
##                                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
```

Data Preprocessing

Filling Missing Value

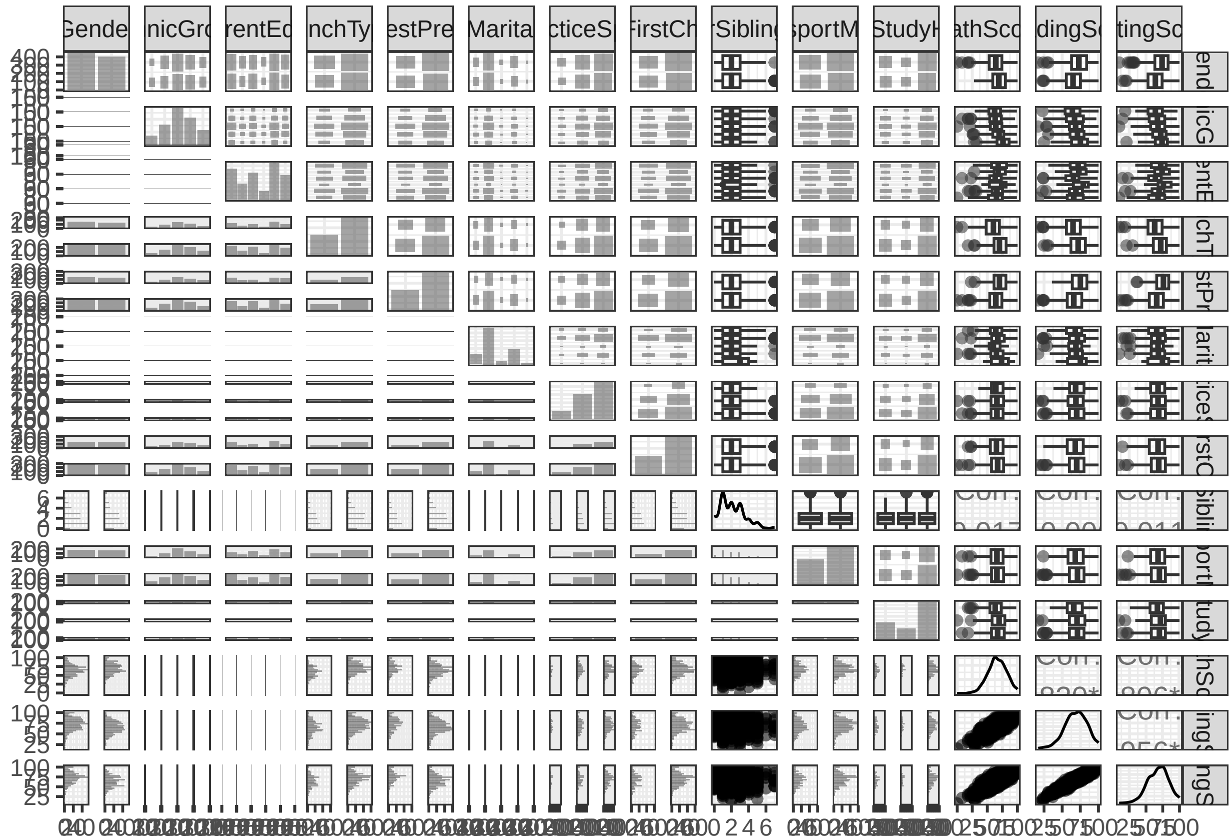
```r
# Imputing missing values
# For columns with fewer missing values, replace with mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

data$PracticeSport[is.na(data$PracticeSport)] <- get_mode(data$PracticeSport)
data$IsFirstChild[is.na(data$IsFirstChild)] <- get_mode(data$IsFirstChild)

# For columns with more missing values, you can choose to impute or drop
# Imputing with mode (as an example)
data$EthnicGroup[is.na(data$EthnicGroup)] <- get_mode(data$EthnicGroup)
data$ParentEduc[is.na(data$ParentEduc)] <- get_mode(data$ParentEduc)
data$TestPrep[is.na(data$TestPrep)] <- get_mode(data$TestPrep)
data$ParentMaritalStatus[is.na(data$ParentMaritalStatus)] <- get_mode(data$TestPrep)
data$WklyStudyHours[is.na(data$WklyStudyHours)]<- get_mode(data$WklyStudyHours)
data$NrSiblings[is.na(data$NrSiblings)] <- get_mode(data$NrSiblings)

# Alternatively, to drop rows with NA values in these columns-TransportMeans
data <- data %>% drop_na(TransportMeans)
```

```r
# Creating a function to count NA and empty strings as missing values
count_missing <- function(x) sum(is.na(x) | x == "")
# Calculating the missing values
missing_values <- sapply(data, function(x) count_missing(x))

# Creating a dataframe for missing values
missing_data_frame <- data.frame(Variable = names(missing_values), MissingValues = missing_values)
```

Examine correlation/pairwise

Examine the marginal distributions and pairwise relationships between variables

```r
# 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

```r
# 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
```

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

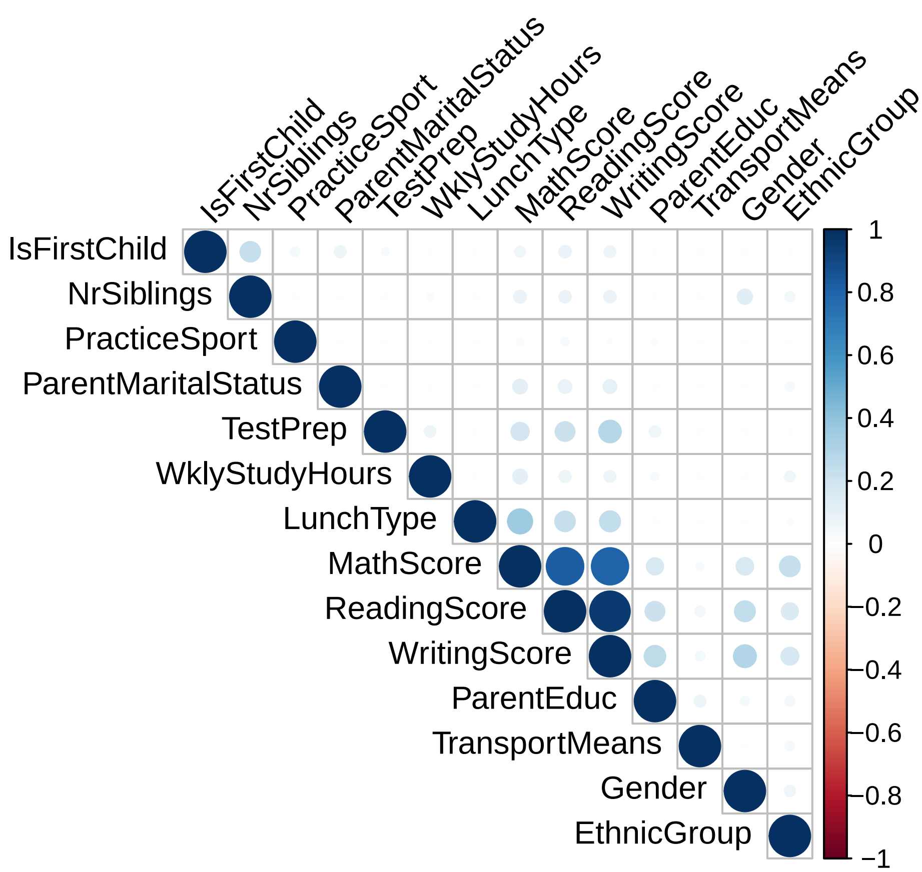
```
## corrplot 0.92 loaded
```

```r
# 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)
```

```r
# 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	0.000	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

```r
# 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.

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. There should be a lot of no relation, and no two of the predictors are colinearity. If auto, it will automatically select the compare method for these correlation:

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
set.seed(123)
splitRatio <- 0.8

trainIndex <- sample(seq_len(nrow(data)), size = floor(splitRatio * nrow(data)))
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]

# Splitting the train dataset into independent variables (X) and dependent variables (Y)
X_train <- trainData %>% select(-c(MathScore, ReadingScore, WritingScore))
Y_math_train <- trainData$MathScore
Y_reading_train <-trainData$ReadingScore
Y_writing_train <- trainData$WritingScore
```

Even if two variables are statistically correlated, it does not necessarily mean that they lead to severe multicollinearity. For example, two variables may be statistically related in some categories, but their overall linear relationship may not be strong. So both are included in the model.

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

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

```
## AIC Model Parameters: 120
```

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

```
## BIC Model Parameters: 246
```

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')
```

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.

```r
# 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
```

```r
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:Is
                   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)

coef(lasso_model)
```

```
## 121 x 1 sparse Matrix of class "dgCMatrix"
##                                                          s0
```

```
## (Intercept)                                              58.4368761
## (Intercept)                                               .
## Gendermale                                                2.8960783
## EthnicGroupgroup B                                        .
## EthnicGroupgroup C                                        .
## EthnicGroupgroup D                                        .
## EthnicGroupgroup E                                        5.8071333
## ParentEducbachelor's degree                               .
## ParentEduchigh school                                    -1.3483799
## ParentEducmaster's degree                                 .
## ParentEducsome college                                    .
## ParentEducsome high school                               -1.7073150
## LunchTypestandard                                         8.5356902
## TestPrepnone                                             -4.5652098
## ParentMaritalStatusmarried                                .
## ParentMaritalStatusnone                                   .
## ParentMaritalStatussingle                                 .
## ParentMaritalStatuswidowed                                .
## PracticeSportregularly                                    .
## PracticeSportsometimes                                    .
## IsFirstChildyes                                           .
## NrSiblings                                                .
## TransportMeansschool_bus                                  .
## WklyStudyHours> 10                                        0.2125310
## WklyStudyHours10-May                                      0.0202314
## Gendermale:LunchTypestandard                              .
## Gendermale:PracticeSportregularly                         1.2942875
## 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                 -4.2614010
## EthnicGroupgroup C:ParentEduchigh school                  .
## EthnicGroupgroup D:ParentEduchigh school                  .
## EthnicGroupgroup E:ParentEduchigh school                  .
## EthnicGroupgroup B:ParentEducmaster's degree              0.3791516
## EthnicGroupgroup C:ParentEducmaster's degree              .
## EthnicGroupgroup D:ParentEducmaster's degree              4.9106200
## EthnicGroupgroup E:ParentEducmaster's degree              .
## EthnicGroupgroup B:ParentEducsome college                 .
## EthnicGroupgroup C:ParentEducsome college                 .
## EthnicGroupgroup D:ParentEducsome college                 4.4099481
## EthnicGroupgroup E:ParentEducsome college                 .
## EthnicGroupgroup B:ParentEducsome high school            -2.4117233
## EthnicGroupgroup C:ParentEducsome high school            -2.3144843
## EthnicGroupgroup D:ParentEducsome high school             .
## EthnicGroupgroup E:ParentEducsome high school             2.4631429
## EthnicGroupgroup B:IsFirstChildyes                        .
## EthnicGroupgroup C:IsFirstChildyes                        .
## EthnicGroupgroup D:IsFirstChildyes                        .
## EthnicGroupgroup E:IsFirstChildyes                        .
## ParentEducbachelor's degree:TestPrepnone                  .
## ParentEduchigh school:TestPrepnone                       -0.5221445
```

```
## 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  -3.6751603
## ParentEduchigh school:ParentMaritalStatusnone        -1.5043643
## ParentEducmaster's degree:ParentMaritalStatusnone             .
## ParentEducsome college:ParentMaritalStatusnone                .
## ParentEducsome high school:ParentMaritalStatusnone            .
## ParentEducbachelor's degree:ParentMaritalStatussingle         .
## ParentEduchigh school:ParentMaritalStatussingle       0.2274941
## ParentEducmaster's degree:ParentMaritalStatussingle           .
## ParentEducsome college:ParentMaritalStatussingle     -4.2160673
## ParentEducsome high school:ParentMaritalStatussingle          .
## ParentEducbachelor's degree:ParentMaritalStatuswidowed  6.0875539
## ParentEduchigh school:ParentMaritalStatuswidowed              .
## ParentEducmaster's degree:ParentMaritalStatuswidowed          .
## ParentEducsome college:ParentMaritalStatuswidowed     6.4288698
## ParentEducsome high school:ParentMaritalStatuswidowed         .
## ParentEducbachelor's degree:PracticeSportregularly    6.9475182
## ParentEduchigh school:PracticeSportregularly                  .
## ParentEducmaster's degree:PracticeSportregularly     -0.9271534
## ParentEducsome college:PracticeSportregularly        -0.9034811
## ParentEducsome high school:PracticeSportregularly             .
## ParentEducbachelor's degree:PracticeSportsometimes            .
## ParentEduchigh school:PracticeSportsometimes                  .
## ParentEducmaster's degree:PracticeSportsometimes      1.8605900
## 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              2.4229902
## LunchTypestandard:TransportMeansschool_bus                    .
## TestPrepnone:WklyStudyHours> 10                               .
## TestPrepnone:WklyStudyHours10-May                             .
## ParentMaritalStatusmarried:PracticeSportregularly     1.1072670
## ParentMaritalStatusnone:PracticeSportregularly                .
## ParentMaritalStatussingle:PracticeSportregularly              .
## ParentMaritalStatuswidowed:PracticeSportregularly             .
## ParentMaritalStatusmarried:PracticeSportsometimes             .
## ParentMaritalStatusnone:PracticeSportsometimes                .
## ParentMaritalStatussingle:PracticeSportsometimes              .
## ParentMaritalStatuswidowed:PracticeSportsometimes             .
## ParentMaritalStatusmarried:IsFirstChildyes                    .
## ParentMaritalStatusnone:IsFirstChildyes                       .
## ParentMaritalStatussingle:IsFirstChildyes             0.2873802
```

```
## ParentMaritalStatuswidowed:IsFirstChildyes               .
## ParentMaritalStatusmarried:TransportMeansschool_bus     2.1148289
## ParentMaritalStatusnone:TransportMeansschool_bus        .
## ParentMaritalStatussingle:TransportMeansschool_bus      .
## ParentMaritalStatuswidowed:TransportMeansschool_bus     .
## PracticeSportregularly:WklyStudyHours> 10               .
## PracticeSportsometimes:WklyStudyHours> 10               .
## PracticeSportregularly:WklyStudyHours10-May             2.9426880
## PracticeSportsometimes:WklyStudyHours10-May             .
## IsFirstChildyes:NrSiblings                              0.2857360
## IsFirstChildyes:TransportMeansschool_bus                .
## IsFirstChildyes:WklyStudyHours> 10                      1.9358455
## 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)
```

```
## 49 x 1 sparse Matrix of class "dgCMatrix"
##                                                  s0
## (Intercept)                             69.24438978
## (Intercept)                              .
## Gendermale                              -9.64456022
## EthnicGroupgroup B                       .
## EthnicGroupgroup C                       0.12187154
## EthnicGroupgroup D                       2.73804550
## EthnicGroupgroup E                       4.32714531
## ParentEducbachelor's degree              1.00843155
## ParentEduchigh school                   -5.16634609
## ParentEducmaster's degree                3.61997993
## ParentEducsome college                  -2.23041408
## ParentEducsome high school              -5.17395739
## LunchTypestandard                        6.79219962
## TestPrepnone                            -6.21291827
## ParentMaritalStatusmarried               2.37055212
## ParentMaritalStatusnone                  0.41689791
## ParentMaritalStatussingle                .
## ParentMaritalStatuswidowed               1.74285608
## PracticeSportregularly                  -2.52686071
## PracticeSportsometimes                   .
## IsFirstChildyes                          1.15235055
## NrSiblings                               .
## TransportMeansschool_bus                 0.08017577
```

```
## WklyStudyHours> 10                                     .
## WklyStudyHours10-May                                   .
## Gendermale:IsFirstChildyes                             2.62766037
## LunchTypestandard:PracticeSportregularly               .
## LunchTypestandard:PracticeSportsometimes               3.07255306
## LunchTypestandard:IsFirstChildyes                     -1.88926071
## TestPrepnone:NrSiblings                               -0.91045866
## TestPrepnone:TransportMeansschool_bus                  2.10087739
## ParentMaritalStatusmarried:PracticeSportregularly  3.63317210
## ParentMaritalStatusnone:PracticeSportregularly    -1.03469273
## ParentMaritalStatussingle:PracticeSportregularly  -0.95977110
## ParentMaritalStatuswidowed:PracticeSportregularly -0.40510097
## ParentMaritalStatusmarried:PracticeSportsometimes  .
## ParentMaritalStatusnone:PracticeSportsometimes     .
## ParentMaritalStatussingle:PracticeSportsometimes  -1.35869930
## ParentMaritalStatuswidowed:PracticeSportsometimes  3.33778366
## ParentMaritalStatusmarried:IsFirstChildyes        -0.41359962
## ParentMaritalStatusnone:IsFirstChildyes            .
## ParentMaritalStatussingle:IsFirstChildyes          3.11304653
## ParentMaritalStatuswidowed:IsFirstChildyes         1.11954328
## PracticeSportregularly:WklyStudyHours> 10          .
## PracticeSportsometimes:WklyStudyHours> 10          .
## PracticeSportregularly:WklyStudyHours10-May        2.99309704
## PracticeSportsometimes:WklyStudyHours10-May       -0.84120400
## NrSiblings:WklyStudyHours> 10                      0.88322964
## NrSiblings:WklyStudyHours10-May                    0.94407262
```

```
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)
```

```
## 52 x 1 sparse Matrix of class "dgCMatrix"
##                                                   s0
## (Intercept)                               69.5913009
## (Intercept)                                .
## Gendermale                                -9.1466566
## EthnicGroupgroup B                        -0.8588264
## EthnicGroupgroup C                         .
## EthnicGroupgroup D                         3.9530918
## EthnicGroupgroup E                         2.6507802
## ParentEducbachelor's degree                2.0339330
```

```
## ParentEduchigh school                              -5.5986108
## ParentEducmaster's degree                           5.7036126
## ParentEducsome college                             -2.9655360
## ParentEducsome high school                         -5.5165771
## LunchTypestandard                                   6.0671040
## TestPrepnone                                       -8.6298117
## ParentMaritalStatusmarried                          2.4165951
## ParentMaritalStatusnone                             .
## ParentMaritalStatussingle                           .
## ParentMaritalStatuswidowed                          0.5886266
## PracticeSportregularly                              .
## PracticeSportsometimes                              .
## IsFirstChildyes                                     .
## NrSiblings                                          0.3821740
## TransportMeansschool_bus                            1.2730919
## WklyStudyHours> 10                                  .
## WklyStudyHours10-May                                0.4208346
## ParentEducbachelor's degree:IsFirstChildyes         .
## ParentEduchigh school:IsFirstChildyes               .
## ParentEducmaster's degree:IsFirstChildyes           .
## ParentEducsome college:IsFirstChildyes              2.4844072
## ParentEducsome high school:IsFirstChildyes          .
## LunchTypestandard:PracticeSportregularly            .
## LunchTypestandard:PracticeSportsometimes            2.8821110
## LunchTypestandard:IsFirstChildyes                   .
## TestPrepnone:NrSiblings                            -0.3665883
## ParentMaritalStatusmarried:PracticeSportregularly  2.1468214
## ParentMaritalStatusnone:PracticeSportregularly    -2.1837752
## ParentMaritalStatussingle:PracticeSportregularly  -0.6445970
## ParentMaritalStatuswidowed:PracticeSportregularly  .
## ParentMaritalStatusmarried:PracticeSportsometimes  .
## ParentMaritalStatusnone:PracticeSportsometimes     .
## ParentMaritalStatussingle:PracticeSportsometimes   .
## ParentMaritalStatuswidowed:PracticeSportsometimes  1.9964773
## ParentMaritalStatusmarried:IsFirstChildyes          .
## ParentMaritalStatusnone:IsFirstChildyes             .
## ParentMaritalStatussingle:IsFirstChildyes           1.4918099
## ParentMaritalStatuswidowed:IsFirstChildyes          0.2210626
## PracticeSportregularly:WklyStudyHours> 10           .
## PracticeSportsometimes:WklyStudyHours> 10           .
## PracticeSportregularly:WklyStudyHours10-May         3.6295404
## PracticeSportsometimes:WklyStudyHours10-May         .
## IsFirstChildyes:WklyStudyHours> 10                  1.2960564
## 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.

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

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

glance_writing = broom::glance(model_writing_best) |>
  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	0.3896522	0.3040798	0	5491.110	5874.986
Reading	0.2822946	0.2334634	0	5460.414	5663.643
Writing	0.3841167	0.3359085	0	5409.882	5640.208

```r
png(file = "math.png", width = 800, height = 800)
par(mfrow = c(2, 2))
plot(model_math_best)
mtext("Math Model Diagnostic", outer = TRUE, cex = 1.5, line = -1)
dev.off()
```

```
## pdf
##   2
```

```r
png(file = "reading.png", width = 800, height = 800)
par(mfrow = c(2, 2))
plot(model_reading_best)
mtext("Reading Model Diagnostic", outer = TRUE, cex = 1.5, line = -1)
dev.off()
```

```
## pdf
##   2
```

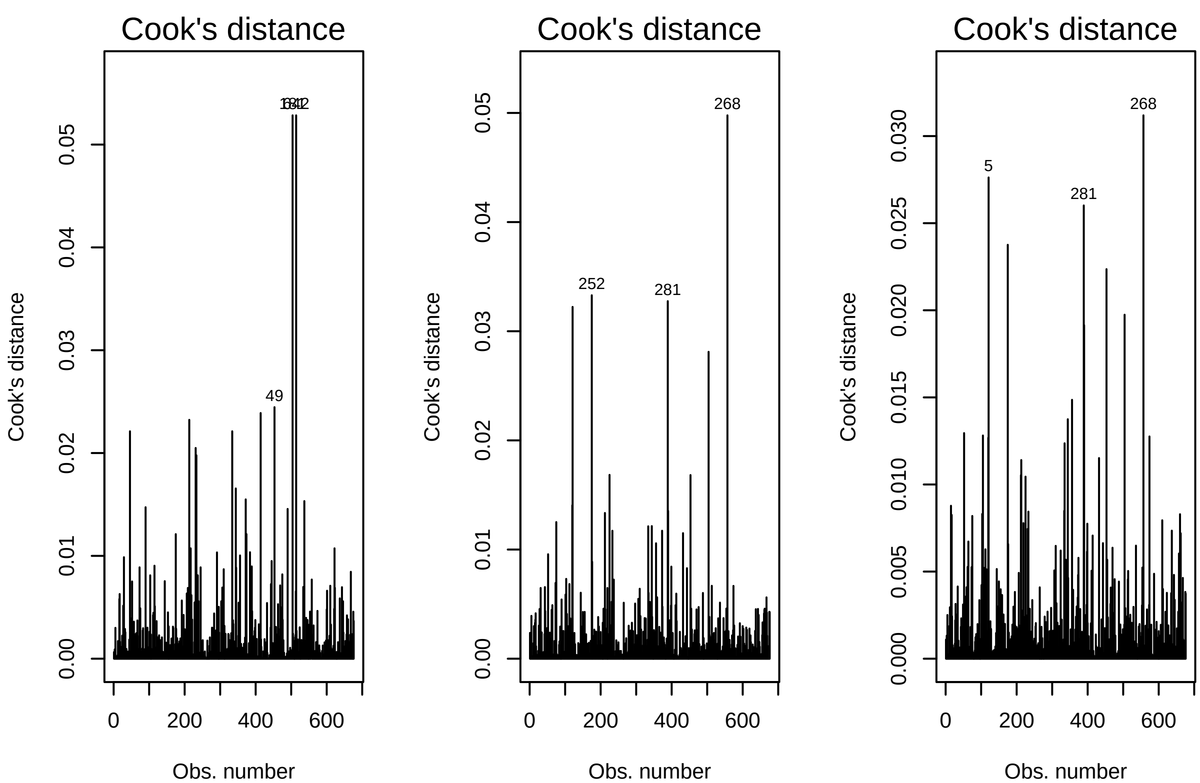
```r
png(file = "writing.png", width = 800, height = 800)
par(mfrow = c(2, 2))
plot(model_writing_best)
mtext("Writing Model Diagnostic", outer = TRUE, cex = 1.5, line = -1)
dev.off()
```

```
## pdf
##   2
```

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

```r
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

```r
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, Paren
## TransportMeans                                Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, 
## WklyStudyHours                          Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, Paren
```

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

```
##                             GVIF Df GVIF^(1/(2*Df))
## Gender              1.086453e+00  1        1.042331
## EthnicGroup         1.384742e+00  4        1.041528
## ParentEduc          2.474226e+01 11        1.157013
## LunchType           1.582338e+02  5        1.659319
## TestPrep            1.270161e+00  3        1.040662
## ParentMaritalStatus 6.007068e+02 19        1.183376
## PracticeSport       4.482670e+03 23        1.200553
## IsFirstChild        6.978027e+05 23        1.339793
## NrSiblings          1.270161e+00  3        1.040662
## TransportMeans      1.069186e+00  1        1.034014
## WklyStudyHours      2.010883e+03 11        1.413038
##                                                         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, P
## ParentMaritalStatus                                         Gender, EthnicGroup, ParentEduc, Lunch
## PracticeSport                                                            Gender, EthnicGroup, Pare
## IsFirstChild                                                             Gender, EthnicGroup, Lunc
## NrSiblings                       Gender, EthnicGroup, ParentEduc, LunchType, ParentMaritalStatus, P
## TransportMeans             Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatu
## WklyStudyHours                                         Gender, EthnicGroup, ParentEduc, LunchType,
```

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

```
##                            GVIF Df GVIF^(1/(2*Df))
## Gender                 1.073508  1        1.036102
## EthnicGroup            1.364765  4        1.039638
## ParentEduc             1.374557  5        1.032325
## LunchType            147.832518  5        1.648075
## TestPrep               1.091363  1        1.044683
## ParentMaritalStatus   68.897268 19        1.117825
## PracticeSport       4319.902647 23        1.199588
## IsFirstChild        5843.077251  9        1.619041
## NrSiblings           115.835734  5        1.608364
## TransportMeans         1.069289  1        1.034064
## WklyStudyHours       148.368681 11        1.255155
##                                                  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, Parent
## TransportMeans             Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatu
## WklyStudyHours                                       Gender, EthnicGroup, ParentEduc, LunchType, T
```

model validation

cross validation

```r
library(caret)
```

```
## Loading required package: lattice

## Registered S3 method overwritten by 'lava':
##   method     from
##   print.pcor greybox

##
## Attaching package: 'caret'

## The following object is masked from 'package:greybox':
##
##     MAE

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
control <- trainControl(method = "cv", number = 10)
set.seed(123)
math_model_data <- cbind(X_train, Y_math_train)
math_model_cv <- train( Y_math_train ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + Parent
    data = math_model_data, method = "lm", trControl = control)

set.seed(124)
reading_model_data <- cbind(X_train, Y_reading_train)
reading_model_cv <- train(Y_reading_train ~ Gender + EthnicGroup + ParentEduc +
    LunchType + TestPrep + ParentMaritalStatus + PracticeSport +
    IsFirstChild + NrSiblings + TransportMeans + WklyStudyHours + LunchType:PracticeSport + ParentMarit
    method = "lm", trControl = control)

set.seed(125)
writing_model_data <- cbind(X_train, Y_writing_train)
writing_model_cv <- train(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,
    method = "lm", trControl = control)


print(math_model_cv)
```

```
## 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
```

```r
print(reading_model_cv)
```

```
## 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
##   13.7283  0.2021777  11.19904
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
print(writing_model_cv)
```

```
## 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
##   13.15398  0.299929  10.62044
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
library(readr)
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
```

```r
library(modelr)
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
```

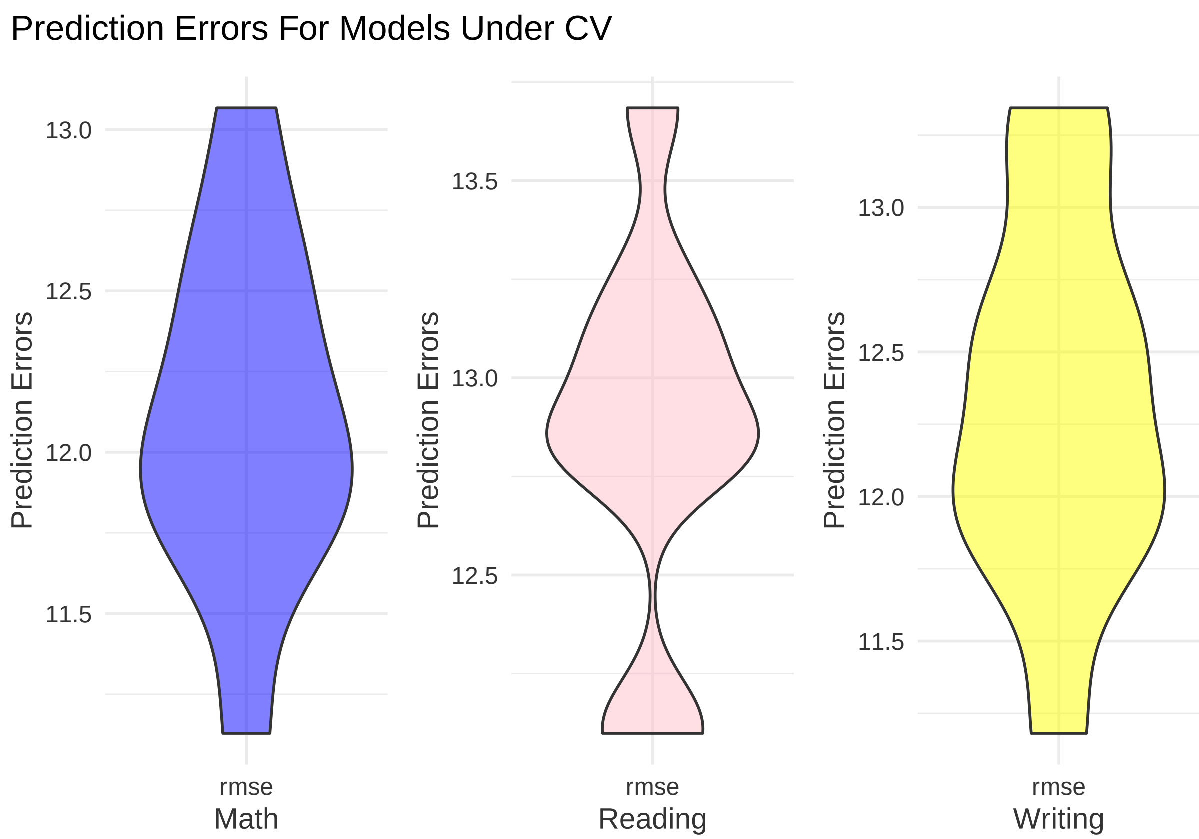
```r
library(boot)
```

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

```r
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 = "Math",
    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)))

# 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 = "Reading",
    y = "Prediction Errors"
  ) +
  theme_minimal() +
  theme(
```

```r
    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 = "Writing",
    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 6: MSPE Values for Different Subjects

Subject	MSPE
Math	198.3466
Reading	152.9267
Writing	142.8281

Take a look of coefficents. Try to understand model in more practical way.

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

Table 7: Math

term	estimate	std.error	statistic	p.value
(Intercept)	59.8236540	7.824529	7.6456552	0.0000000
Gendermale	4.9989664	3.390164	1.4745500	0.1408652
EthnicGroupgroup B	2.1158513	5.390584	0.3925087	0.6948236
EthnicGroupgroup C	-0.3917127	5.103840	-0.0767486	0.9388495
EthnicGroupgroup D	-0.2697026	5.456409	-0.0494286	0.9605944
EthnicGroupgroup E	4.8561321	5.555826	0.8740613	0.3824393
ParentEducbachelor's degree	13.2158169	10.456616	1.2638713	0.2067737
ParentEduchigh school	-1.3788436	8.702785	-0.1584370	0.8741665
ParentEducmaster's degree	-11.1824725	13.312940	-0.8399702	0.4012642
ParentEducsome college	1.2690513	7.903915	0.1605598	0.8724949
ParentEducsome high school	2.8839705	8.905244	0.3238508	0.7461654
LunchTypestandard	2.5820328	3.430405	0.7526903	0.4519352
TestPrepnone	-5.4668672	1.164944	-4.6928142	0.0000034
ParentMaritalStatusmarried	4.4767797	3.717787	1.2041517	0.2290122
ParentMaritalStatusnone	5.1486153	6.329960	0.8133725	0.4163316
ParentMaritalStatussingle	7.1051335	4.343620	1.6357630	0.1024207
ParentMaritalStatuswidowed	32.1825946	13.767636	2.3375542	0.0197426
PracticeSportregularly	-7.0323692	6.092899	-1.1541910	0.2488876
PracticeSportsometimes	-6.2033803	5.900179	-1.0513884	0.2935092
TransportMeansschool_bus	-2.9955588	2.957436	-1.0128904	0.3115263
WklyStudyHours> 10	2.1744283	5.306108	0.4097972	0.6821029
WklyStudyHours10-May	-3.0158131	3.856450	-0.7820179	0.4345167
Gendermale:PracticeSportregularly	2.4088354	3.826334	0.6295413	0.5292376
Gendermale:PracticeSportsometimes	-3.1094030	3.682194	-0.8444429	0.3987631
EthnicGroupgroup B:ParentEducbachelor's degree	12.3301744	10.103598	1.2203746	0.2228088
EthnicGroupgroup C:ParentEducbachelor's degree	15.6483238	9.442126	1.6572881	0.0979910
EthnicGroupgroup D:ParentEducbachelor's degree	11.1026820	9.857639	1.1263023	0.2604939
EthnicGroupgroup E:ParentEducbachelor's degree	17.5929375	10.628403	1.6552757	0.0983986
EthnicGroupgroup B:ParentEduchigh school	-6.6339532	7.086337	-0.9361611	0.3495720
EthnicGroupgroup C:ParentEduchigh school	1.0022039	6.695100	0.1496922	0.8810585
EthnicGroupgroup D:ParentEduchigh school	1.0241333	7.063144	0.1449968	0.8847628
EthnicGroupgroup E:ParentEduchigh school	4.1683624	7.576404	0.5501769	0.5824056
EthnicGroupgroup B:ParentEducmaster's degree	23.2799497	13.355971	1.7430369	0.0818464
EthnicGroupgroup C:ParentEducmaster's degree	15.3169367	12.185689	1.2569611	0.2092634
EthnicGroupgroup D:ParentEducmaster's degree	26.8573940	11.968517	2.2440036	0.0252009
EthnicGroupgroup E:ParentEducmaster's degree	21.6878571	13.395444	1.6190473	0.1059697
EthnicGroupgroup B:ParentEducsome college	0.9014048	6.923716	0.1301909	0.8964596

term	estimate	std.error	statistic	p.value
EthnicGroupgroup C:ParentEducsome college	3.7712301	6.460460	0.5837402	0.5596175
EthnicGroupgroup D:ParentEducsome college	11.1836587	6.773641	1.6510556	0.0992576
EthnicGroupgroup E:ParentEducsome college	2.5687923	7.049554	0.3643907	0.7156964
EthnicGroupgroup B:ParentEducsome high school	-5.2322199	7.387119	-0.7082897	0.4790442
EthnicGroupgroup C:ParentEducsome high school	-0.6264665	7.267877	-0.0861966	0.9313392
EthnicGroupgroup D:ParentEducsome high school	1.5059296	7.255510	0.2075567	0.8356465
EthnicGroupgroup E:ParentEducsome high school	9.7108636	8.278557	1.1730140	0.2412619
ParentEducbachelor's degree:ParentMaritalStatusmarried	-21.6431778	7.246291	-2.9867939	0.0029358
ParentEduchigh school:ParentMaritalStatusmarried	0.2578684	4.637813	0.0556013	0.9556782
ParentEducmaster's degree:ParentMaritalStatusmarried	-5.6890480	8.429506	-0.6748969	0.5000047
ParentEducsome college:ParentMaritalStatusmarried	-7.4587586	4.427521	-1.6846354	0.0925860
ParentEducsome high school:ParentMaritalStatusmarried	-6.8910904	5.205234	-1.3238771	0.1860548
ParentEducbachelor's degree:ParentMaritalStatusnone	-23.7628624	13.278842	-1.7895282	0.0740408
ParentEduchigh school:ParentMaritalStatusnone	-8.7603507	8.413185	-1.0412645	0.2981778
ParentEducmaster's degree:ParentMaritalStatusnone	0.2601949	14.410477	0.0180560	0.9856003
ParentEducsome college:ParentMaritalStatusnone	-5.3820540	7.548785	-0.7129696	0.4761455
ParentEducsome high school:ParentMaritalStatusnone	-8.1299851	11.423571	-0.7116851	0.4769402
ParentEducbachelor's degree:ParentMaritalStatussingle	-27.9975589	7.995804	-3.5015315	0.0004974
ParentEduchigh school:ParentMaritalStatussingle	1.1510539	5.420215	0.2123632	0.8318968
ParentEducmaster's degree:ParentMaritalStatussingle	-10.4975236	9.553394	-1.0988266	0.2722904
ParentEducsome college:ParentMaritalStatussingle	-13.8823053	5.170519	-2.6848962	0.0074585
ParentEducsome high school:ParentMaritalStatussingle	-8.4415130	5.906359	-1.4292244	0.1534672
ParentEducbachelor's degree:ParentMaritalStatuswidowed	-14.5660804	14.065360	-1.0355995	0.3008118
ParentEduchigh school:ParentMaritalStatuswidowed	-22.3538476	12.996728	-1.7199596	0.0859624
ParentEducmaster's degree:ParentMaritalStatuswidowed	-32.9776873	21.234213	-1.5530450	0.1209468
ParentEducsome college:ParentMaritalStatuswidowed	-5.9891743	12.354242	-0.4847869	0.6280069
ParentEducsome high school:ParentMaritalStatuswidowed	-31.1843496	13.996685	-2.2279811	0.0262567
ParentEducbachelor's degree:PracticeSportregularly	2.4921887	7.600560	0.3278954	0.7431067
ParentEduchigh school:PracticeSportregularly	-4.7479893	5.952541	-0.7976408	0.4253988
ParentEducmaster's degree:PracticeSportregularly	-11.5834986	9.179462	-1.2618930	0.2074842
ParentEducsome college:PracticeSportregularly	-2.8096597	5.125701	-0.5481513	0.5837947
ParentEducsome high school:PracticeSportregularly	-3.9740663	5.911862	-0.6722190	0.5017065
ParentEducbachelor's degree:PracticeSportsometimes	-9.8379934	7.378795	-1.3332791	0.1829531
ParentEduchigh school:PracticeSportsometimes	-3.2966005	5.802557	-0.5681289	0.5701629
ParentEducmaster's degree:PracticeSportsometimes	3.4760708	7.539298	0.4610603	0.6449247
ParentEducsome college:PracticeSportsometimes	0.5729083	4.983034	0.1149718	0.9085065
ParentEducsome high school:PracticeSportsometimes	-1.7933565	5.768358	-0.3108955	0.7559895
LunchTypestandard:PracticeSportregularly	7.8623509	3.902305	2.0147967	0.0443776
LunchTypestandard:PracticeSportsometimes	10.9083551	3.774567	2.8899616	0.0039940
ParentMaritalStatusmarried:TransportMeansschool_bus	5.9984008	3.307985	1.8133093	0.0702904
ParentMaritalStatusnone:TransportMeansschool_bus	4.1673088	6.034708	0.6905568	0.4901148
ParentMaritalStatussingle:TransportMeansschool_bus	1.9494148	3.747047	0.5202536	0.6030813

term	estimate	std.error	statistic	p.value
ParentMaritalStatuswidowed:TransportMeansschool_bus	-13.8706862	9.598489	-1.4450906	0.1489615
PracticeSportregularly:WklyStudyHours> 10	1.7873962	6.016529	0.2970810	0.7665089
PracticeSportsometimes:WklyStudyHours> 10	2.8131491	5.807567	0.4843938	0.6282856
PracticeSportregularly:WklyStudyHours10-May	9.5741323	4.391825	2.1799895	0.0296514
PracticeSportsometimes:WklyStudyHours10-May	4.8408421	4.247024	1.1398198	0.2548223

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

Table 8: Reading

term	estimate	std.error	statistic	p.value
(Intercept)	66.7294950	6.5129906	10.2455998	0.0000000
Gendermale	-8.3451224	1.0593384	-7.8776741	0.0000000
EthnicGroupgroup B	0.9255796	2.2073855	0.4193103	0.6751318
EthnicGroupgroup C	1.7676168	2.0668039	0.8552417	0.3927415
EthnicGroupgroup D	4.3694159	2.1234404	2.0577059	0.0400280
EthnicGroupgroup E	5.5691293	2.3447971	2.3751008	0.0178416
ParentEducbachelor's degree	0.9593230	1.9583347	0.4898667	0.6243982
ParentEduchigh school	-5.9700940	1.6444629	-3.6304217	0.0003059
ParentEducmaster's degree	3.0739419	2.4233342	1.2684763	0.2050949
ParentEducsome college	-3.4228951	1.5292774	-2.2382434	0.0255526
ParentEducsome high school	-6.2272681	1.7534277	-3.5514826	0.0004116
LunchTypestandard	1.5465446	3.2543084	0.4752299	0.6347873
TestPrepnone	-6.7112750	1.1369352	-5.9029530	0.0000000
ParentMaritalStatusmarried	12.9982171	5.3710658	2.4200443	0.0157996
ParentMaritalStatusnone	25.0534251	9.6932917	2.5846148	0.0099718
ParentMaritalStatussingle	17.1900325	6.2980878	2.7294050	0.0065215
ParentMaritalStatuswidowed	20.3651926	9.6806792	2.1036946	0.0357997
PracticeSportregularly	-8.4725203	6.2381933	-1.3581689	0.1748947
PracticeSportsometimes	-2.8648163	5.9105655	-0.4846941	0.6280613
IsFirstChildyes	10.0828700	3.4087261	2.9579584	0.0032125
NrSiblings	-1.4325336	0.7269284	-1.9706669	0.0491980
TransportMeansschool_bus	2.0276616	1.0777747	1.8813410	0.0603850
WklyStudyHours> 10	-4.0765631	5.5292118	-0.7372774	0.4612273
WklyStudyHours10-May	-8.7804016	4.1184180	-2.1319841	0.0333931
LunchTypestandard:PracticeSportregularly	4.2752122	3.7133490	1.1513090	0.2500405
LunchTypestandard:PracticeSportsometimes	7.4223767	3.5916114	2.0665868	0.0391799
ParentMaritalStatusmarried:PracticeSportregularly	2.1171894	5.1522455	0.4109256	0.6812664
ParentMaritalStatusnone:PracticeSportregularly	-13.2554099	8.4242250	-1.5734872	0.1161065
ParentMaritalStatussingle:PracticeSportregularly	-10.5446824	6.2293309	-1.6927472	0.0909964
ParentMaritalStatuswidowed:PracticeSportregularly	-15.7467113	10.7332090	-1.4671019	0.1428458
ParentMaritalStatusmarried:PracticeSportsometimes	-3.4069333	4.8619672	-0.7007314	0.4837285
ParentMaritalStatusnone:PracticeSportsometimes	-14.4055236	7.9601015	-1.8097161	0.0708147
ParentMaritalStatussingle:PracticeSportsometimes	-12.9439931	5.9797313	-2.1646446	0.0307886
ParentMaritalStatuswidowed:PracticeSportsometimes	-10.3871025	10.9192492	-0.9512653	0.3418334
ParentMaritalStatusmarried:IsFirstChildyes	-10.4573902	3.7160407	-2.8141215	0.0050433
ParentMaritalStatusnone:IsFirstChildyes	-15.2781783	7.2881211	-2.0963124	0.0364515
ParentMaritalStatussingle:IsFirstChildyes	-5.0872414	4.1352014	-1.2302282	0.2190694
ParentMaritalStatuswidowed:IsFirstChildyes	-4.5530850	7.9795899	-0.5705914	0.5684795
PracticeSportregularly:WklyStudyHours> 10	3.5682387	5.7478515	0.6207952	0.5349582

term	estimate	std.error	statistic	p.value
PracticeSportsometimes:WklyStudyHours> 10	2.8699791	5.5660851	0.5156190	0.6063009
PracticeSportregularly:WklyStudyHours10-May	11.5013304	4.3089635	2.6691640	0.0077993
PracticeSportsometimes:WklyStudyHours10-May	5.4826677	4.1395741	1.3244521	0.1858317
NrSiblings:WklyStudyHours> 10	1.5599932	1.1177197	1.3956927	0.1632972
NrSiblings:WklyStudyHours10-May	2.0696810	0.8583462	2.4112427	0.0161825

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

Table 9: Writing

term	estimate	std.error	statistic	p.value
(Intercept)	63.3709846	6.6409190	9.5425023	0.0000000
Gendermale	-9.8669217	1.0224029	-9.6507172	0.0000000
EthnicGroupgroup B	-0.2067857	2.1153119	-0.0977566	0.9221569
EthnicGroupgroup C	1.3435451	1.9866591	0.6762836	0.4991104
EthnicGroupgroup D	5.7975931	2.0367028	2.8465582	0.0045643
EthnicGroupgroup E	4.4379024	2.2613889	1.9624676	0.0501506
ParentEducbachelor's degree	3.8014422	3.3859652	1.1227056	0.2619929
ParentEduchigh school	-11.4098557	2.7739139	-4.1132696	0.0000442
ParentEducmaster's degree	4.5006017	4.0524117	1.1105983	0.2671677
ParentEducsome college	-8.8143126	2.4578101	-3.5862464	0.0003616
ParentEducsome high school	-8.9386870	2.8010267	-3.1912180	0.0014876
LunchTypestandard	1.8520503	3.1483239	0.5882655	0.5565663
TestPrepnone	-7.3456304	1.8869030	-3.8929561	0.0001097
ParentMaritalStatusmarried	11.6927213	5.1736308	2.2600610	0.0241603
ParentMaritalStatusnone	17.0801063	9.3437675	1.8279678	0.0680302
ParentMaritalStatussingle	13.9820111	6.0620775	2.3064719	0.0214100
ParentMaritalStatuswidowed	18.4272852	9.2741784	1.9869453	0.0473639
PracticeSportregularly	-8.3652226	6.0499866	-1.3826845	0.1672547
PracticeSportsometimes	-3.4537144	5.7240259	-0.6033715	0.5464801
IsFirstChildyes	8.7899099	4.2419170	2.0721551	0.0386598
NrSiblings	1.0067232	0.5727539	1.7576891	0.0792891
TransportMeansschool_bus	2.1796409	1.0339310	2.1081106	0.0354183
WklyStudyHours> 10	-0.7970647	5.5864483	-0.1426783	0.8865902
WklyStudyHours10-May	-0.3997597	4.1749272	-0.0957525	0.9237478
ParentEducbachelor's degree:IsFirstChildyes	-2.4899524	4.0774623	-0.6106623	0.5416448
ParentEduchigh school:IsFirstChildyes	6.1861884	3.3680045	1.8367518	0.0667205
ParentEducmaster's degree:IsFirstChildyes	1.5275660	4.8873450	0.3125554	0.7547226
ParentEducsome college:IsFirstChildyes	8.6738003	3.0536356	2.8404831	0.0046510
ParentEducsome high school:IsFirstChildyes	2.9574964	3.4677705	0.8528524	0.3940673
LunchTypestandard:PracticeSportregularly	5.4421952	3.6009281	1.5113313	0.1312087
LunchTypestandard:PracticeSportsometimes	7.8378604	3.4747701	2.2556486	0.0244371
TestPrepnone:NrSiblings	-1.0363150	0.7099890	-1.4596212	0.1448958
ParentMaritalStatusmarried:PracticeSportregularly	4.3272509	4.9576748	0.8728388	0.3830856
ParentMaritalStatusnone:PracticeSportregularly	-11.6096660	8.1309654	-1.4278336	0.1538384
ParentMaritalStatussingle:PracticeSportregularly	-6.1061216	6.0318785	-1.0123085	0.3117817
ParentMaritalStatuswidowed:PracticeSportregularly	-16.4866756	10.3318618	-1.5957120	0.1110578
ParentMaritalStatusmarried:PracticeSportsometimes	-1.6078719	4.6764775	-0.3438212	0.7310962
ParentMaritalStatusnone:PracticeSportsometimes	-10.5375101	7.6623981	-1.3752235	0.1695541
ParentMaritalStatussingle:PracticeSportsometimes	-9.1583320	5.7693200	-1.5874197	0.1129226

term	estimate	std.error	statistic	p.value
ParentMaritalStatuswidowed:PracticeSportsometimes	-11.4398918	10.5454267	-1.0848202	0.2784189
ParentMaritalStatusmarried:IsFirstChildyes	-9.6971320	3.5961881	-2.6965030	0.0071957
ParentMaritalStatusnone:IsFirstChildyes	-8.6878943	7.0146922	-1.2385282	0.2159845
ParentMaritalStatussingle:IsFirstChildyes	-5.1212907	4.0082739	-1.2776798	0.2018359
ParentMaritalStatuswidowed:IsFirstChildyes	-1.6510229	7.6782636	-0.2150256	0.8298174
PracticeSportregularly:WklyStudyHours> 10	3.1918173	5.5479415	0.5753156	0.5652846
PracticeSportsometimes:WklyStudyHours> 10	3.9406467	5.4195912	0.7271114	0.4674295
PracticeSportregularly:WklyStudyHours10-May	11.0302513	4.1694733	2.6454783	0.0083624
PracticeSportsometimes:WklyStudyHours10-May	5.5098814	3.9942868	1.3794406	0.1682515
IsFirstChildyes:WklyStudyHours> 10	-0.8801575	3.3630648	-0.2617129	0.7936289
IsFirstChildyes:WklyStudyHours10-May	-5.4687471	2.5802518	-2.1194626	0.0344444