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

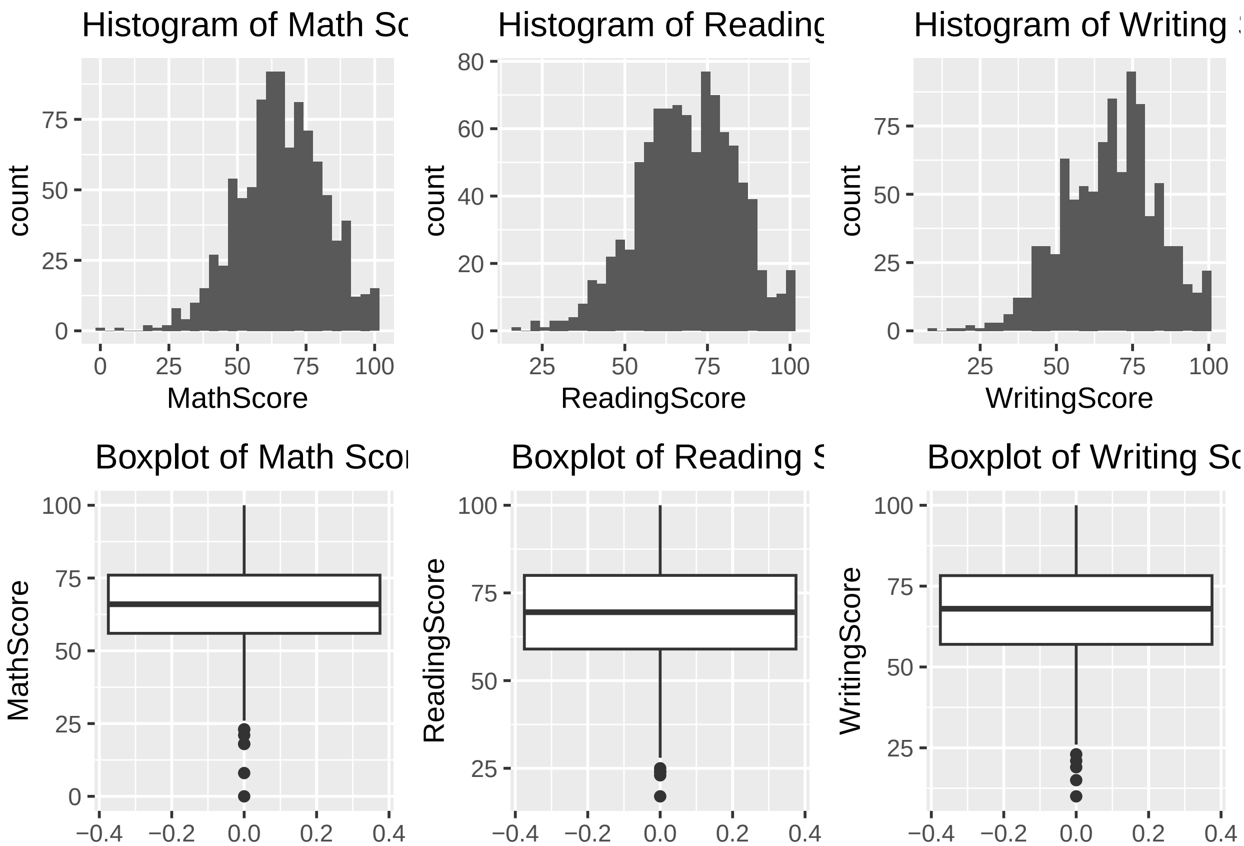
Variable type: character

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

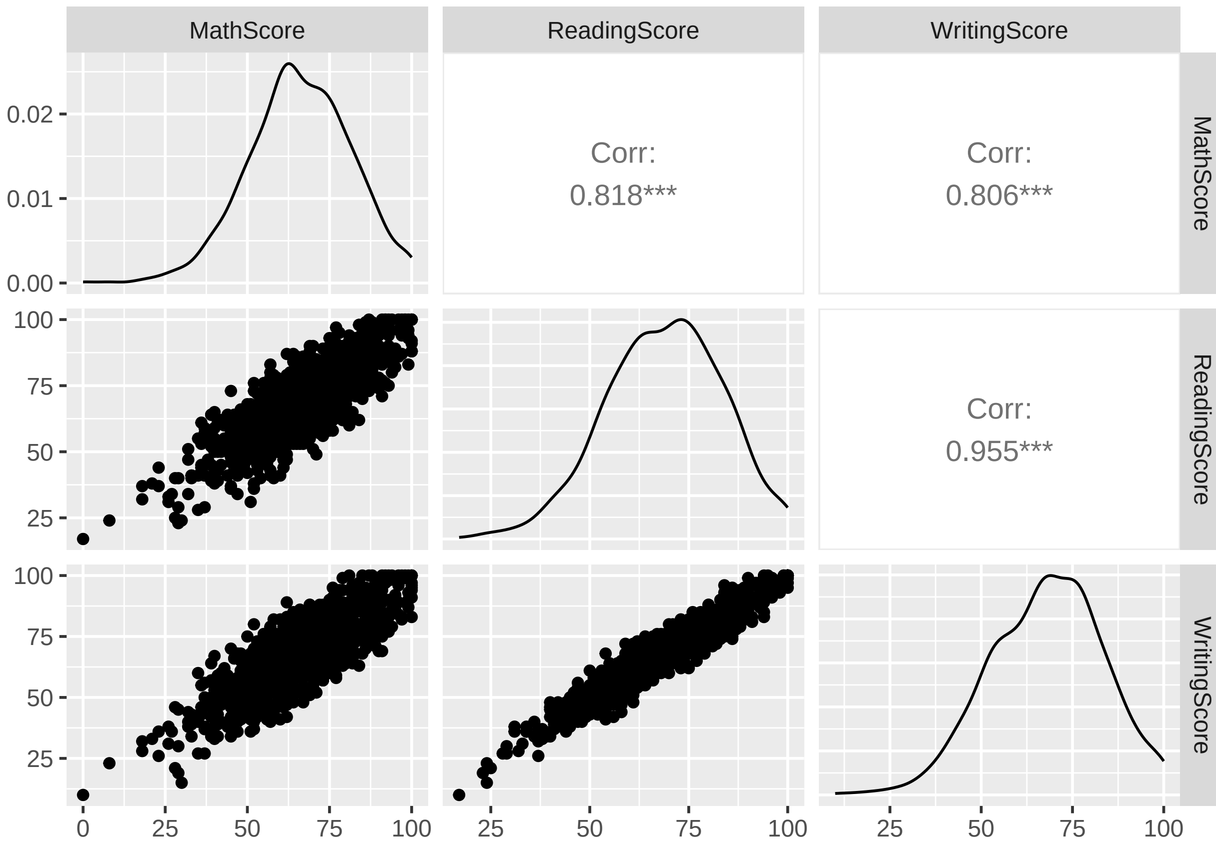
Variable type: numeric

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	

```r
# 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

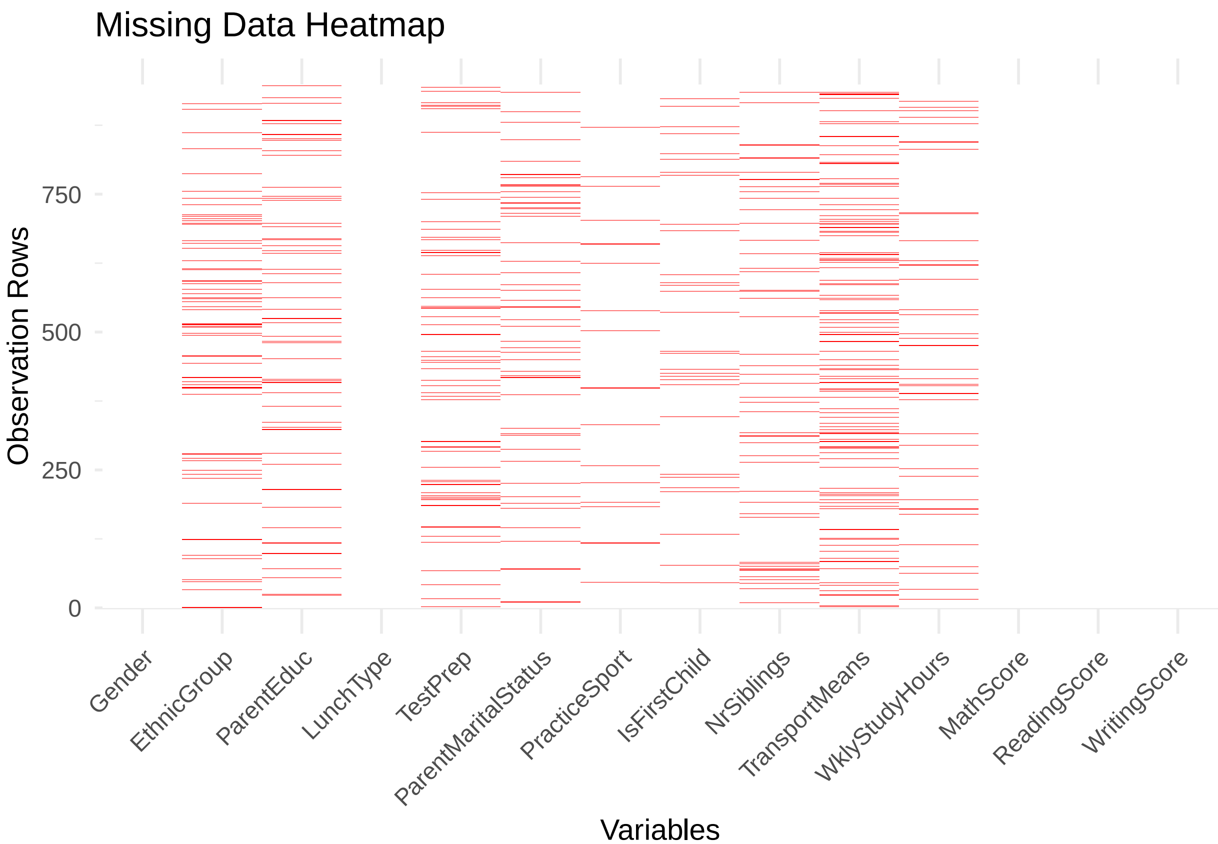
```
library(reshape2)
```

```
##
## 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)) +
  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.
```




```
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

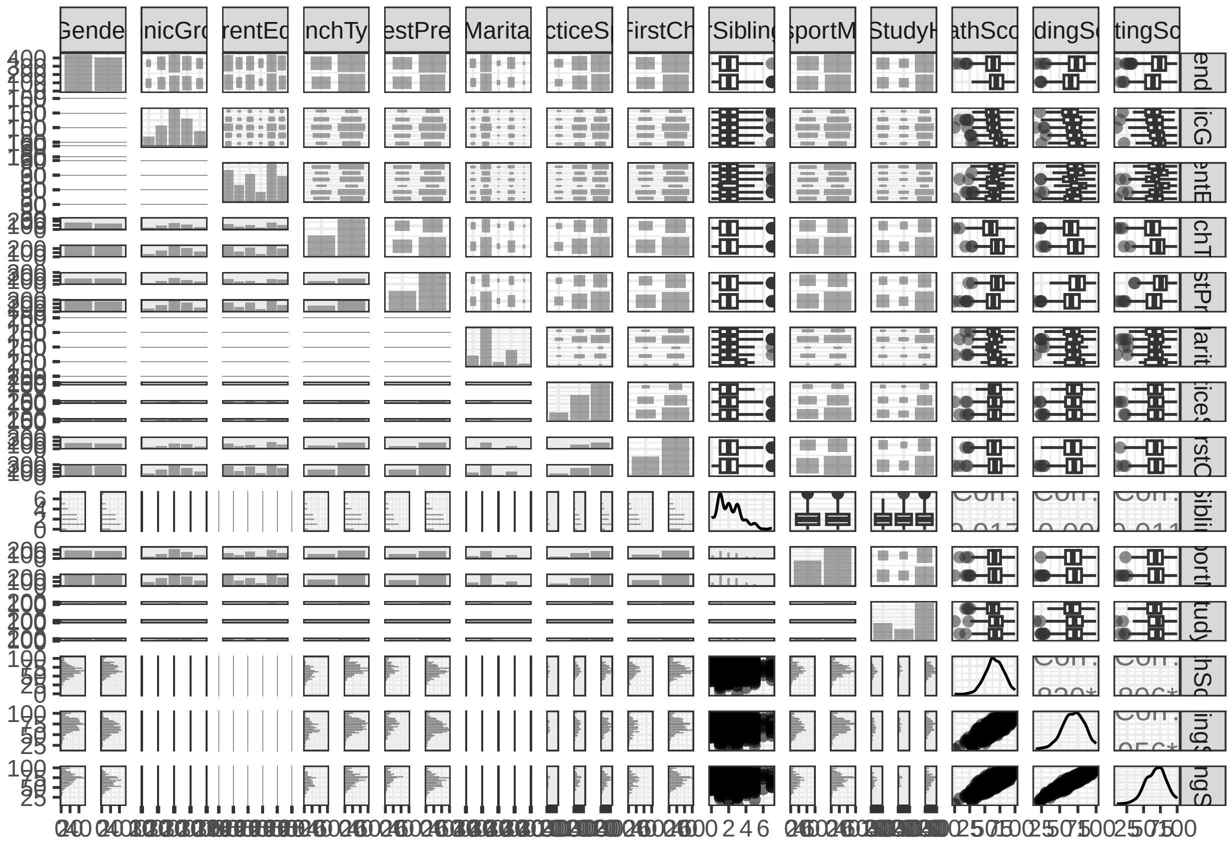
```r
# Imputing missing values
# For columns with fewer missing values, replace with mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

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)
```

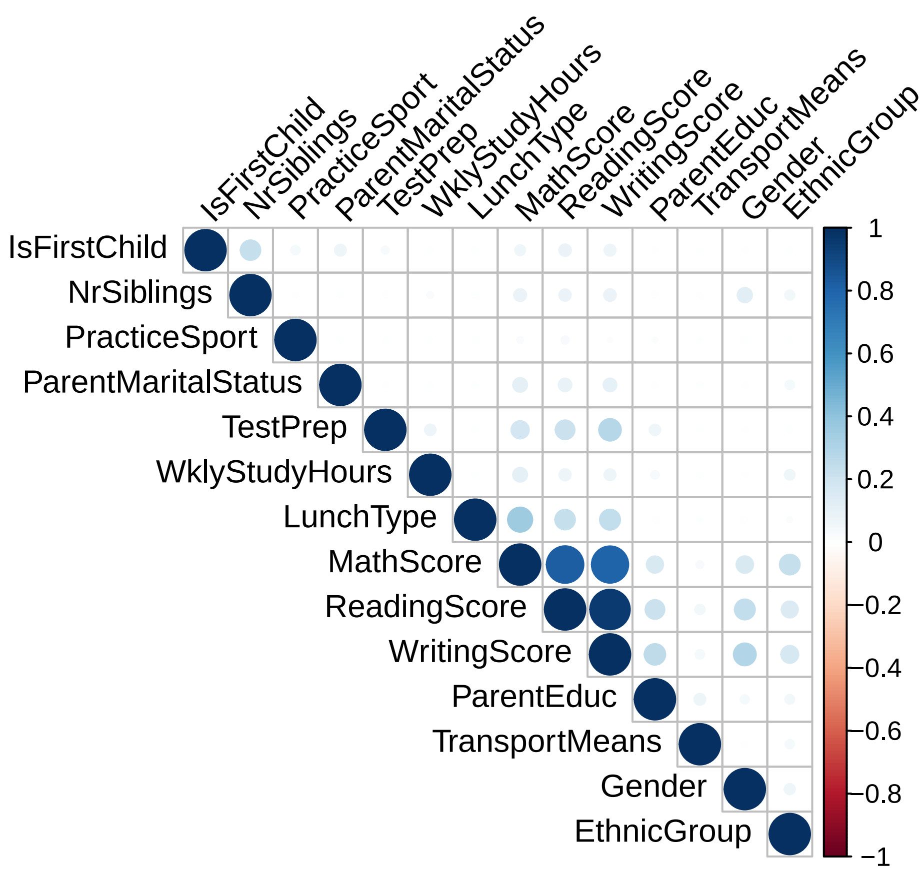
```
## corrplot 0.92 loaded
```

```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:IsI
                   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                    .
```

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

```r
# 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                  .
```

```r
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.

```
# 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

```
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
```

```
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
```

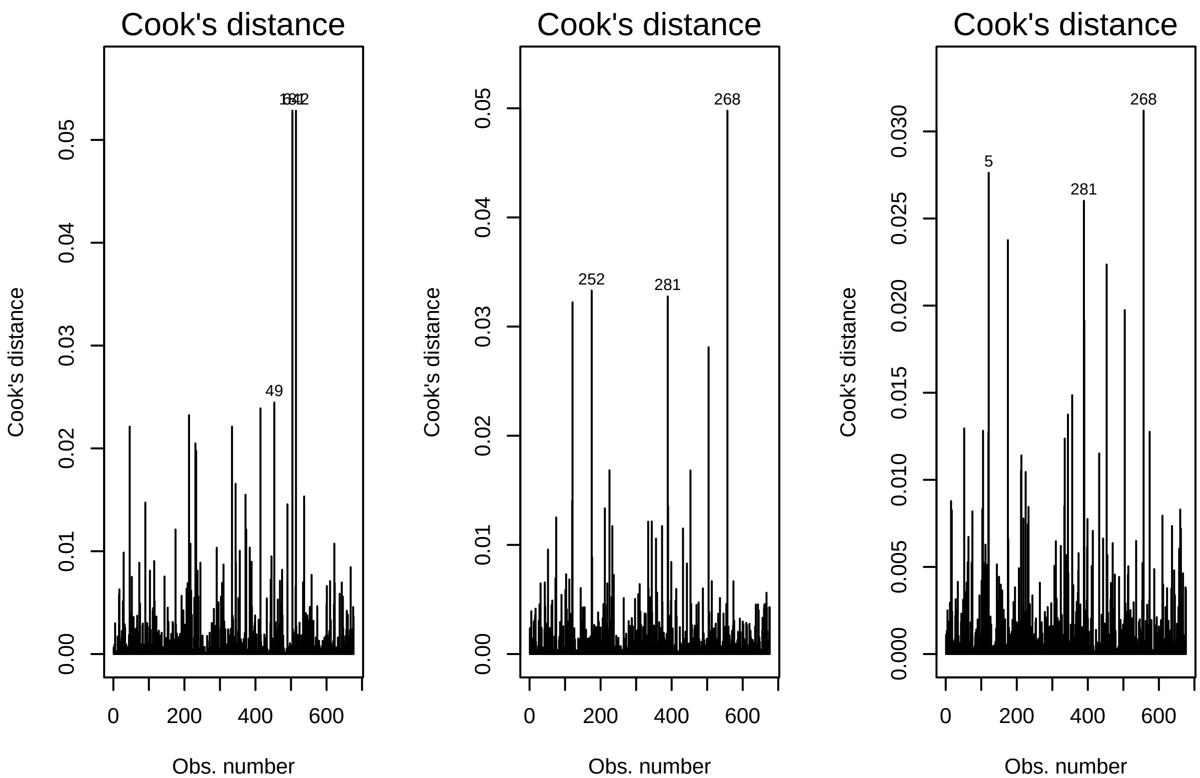
```
png(file = "writing.png", width = 800, height = 800)
par(mfrow = c(2, 2))
plot(model_writing_best)
mtext("Writing Model Diagnostic", outer = TRUE, cex = 1.5, line = -1)
dev.off()
```

```
## 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, Pr
## ParentMaritalStatus                                         Gender, EthnicGroup, ParentEduc, LunchT
## PracticeSport                                                            Gender, EthnicGroup, Paren
## 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                  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, LunchT
## PracticeSport                                                            Gender, EthnicGroup, Paren
## IsFirstChild                                 Gender, EthnicGroup, ParentEduc, LunchType, TestPrep,
## NrSiblings                             Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentM
## TransportMeans              Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus
## WklyStudyHours                                        Gender, EthnicGroup, ParentEduc, LunchType, Te
```

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 + Paren
    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
```

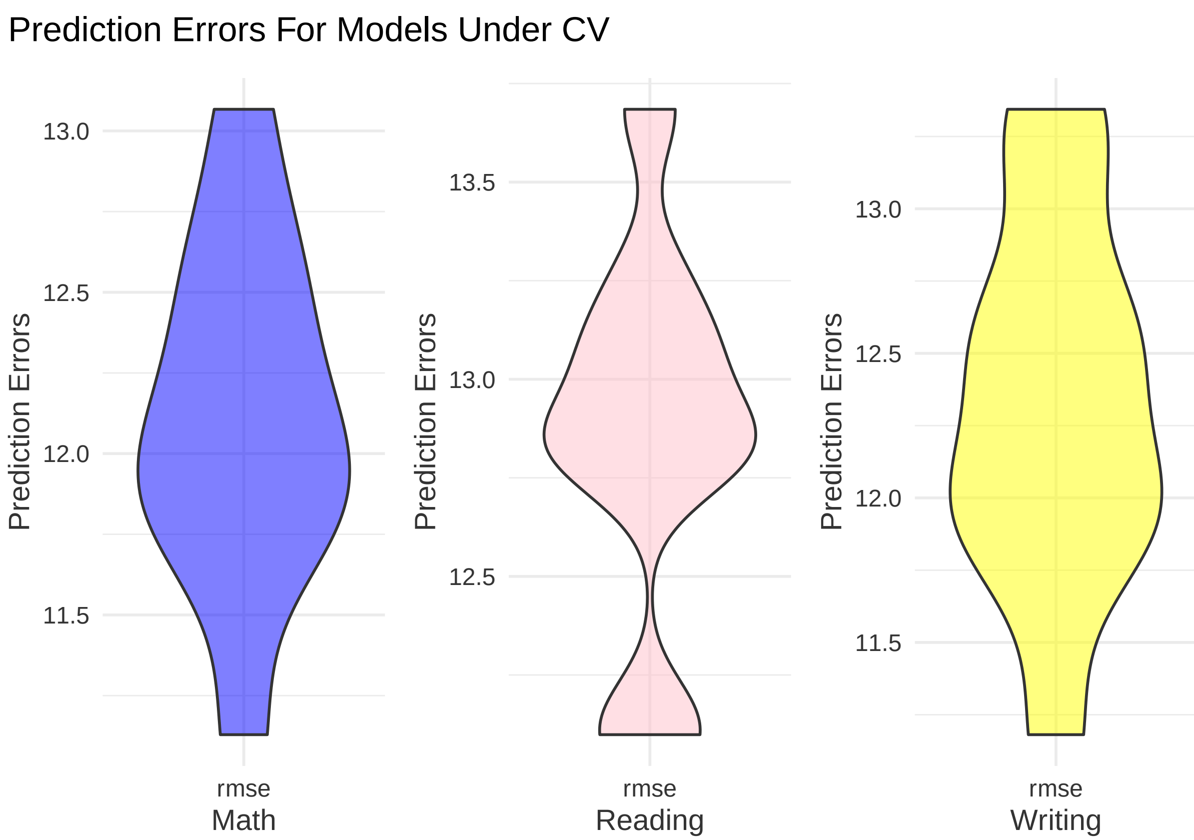
```r
library(boot)
```

```
##
## 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)))
```

```r
# 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(
```

```
    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

```r
# 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
```

```r
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)
```

```r
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 coeffcients. Try to understand model in more practical way.

```
# Save the results
broom::tidy(model_math_best) |>
  saveRDS("math_table.rds")
broom::tidy(model_reading_best) |>
  saveRDS("reading_table.rds")
broom::tidy(model_writing_best) |>
  saveRDS("writing_table.rds")
```