

Code

Read and Clean Data

```
data =  
read_csv("./data.csv") |>  
janitor::clean_names() |>  
mutate(  
  gender = factor(case_when(  
    gender == "male" ~ 0,  
    gender == "female" ~ 1,  
  )),  
  ethnic_group = factor(case_when(  
    ethnic_group == "group A" ~ 0,  
    ethnic_group == "group B" ~ 1,  
    ethnic_group == "group C" ~ 2,  
    ethnic_group == "group D" ~ 3,  
    ethnic_group == "group E" ~ 4,  
  )),  
  parent_educ = factor(case_when(  
    parent_educ == "some highschool" ~ 0,  
    parent_educ == "some college" ~ 1,  
    parent_educ == "associate's degree" ~ 2,  
    parent_educ == "bachelor's degree" ~ 3,  
    parent_educ == "master's degree" ~ 4,  
  )),  
  lunch_type = factor(case_when(  
    lunch_type == "standard" ~ 0,  
    lunch_type == "free/reduced" ~ 1,  
  )),  
  test_prep = factor(case_when(  
    test_prep == "none" ~ 0,  
    test_prep == "completed" ~ 1,  
  )),  
  parent_marital_status = factor(case_when(  
    parent_marital_status == "married" ~ 0,  
    parent_marital_status == "single" ~ 1,  
    parent_marital_status == "widowed" ~ 2,  
    parent_marital_status == "divorced" ~ 3,  
  )),  
  practice_sport = factor(case_when(  
    practice_sport == "never" ~ 0,  
    practice_sport == "sometimes" ~ 1,  
    practice_sport == "regularly" ~ 2,  
  )),  
  is_first_child = factor(case_when(  

```

```

    is_first_child == "no" ~ 0,
    is_first_child == "yes" ~ 1,
  )),
  transport_means = factor(case_when(
    transport_means == "school_bus" ~ 0,
    transport_means == "private" ~ 1,
  )),
  wkly_study_hours = factor(case_when(
    wkly_study_hours == "< 5" ~ 0,
    wkly_study_hours == "10-May" ~ 1,
    wkly_study_hours == "> 10" ~ 2,
  ))
) |>
mutate(nr_siblings = factor(nr_siblings))

```

```

## Rows: 948 Columns: 14
## -- Column specification -----
## Delimiter: ","
## chr (10): Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMarita...
## dbl (4): NrSiblings, MathScore, ReadingScore, WritingScore
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```

```

'
data <- read_csv("./data.csv") |>
janitor::clean_names() |>
mutate(
  gender = case_when(
    gender == "male" ~ 0,
    gender == "female" ~ 1,
  ),
  ethnic_group = case_when(
    ethnic_group == "group A" ~ 0,
    ethnic_group == "group B" ~ 1,
    ethnic_group == "group C" ~ 2,
    ethnic_group == "group D" ~ 3,
    ethnic_group == "group E" ~ 4,
  ),
  parent_educ = case_when(
    parent_educ == "some highschool" ~ 0,
    parent_educ == "some college" ~ 1,
    parent_educ == "associate" ~ 2,
    parent_educ == "bachelor" ~ 3,
    parent_educ == "master" ~ 4,
  ),
  lunch_type = case_when(
    lunch_type == "standard" ~ 0,
    lunch_type == "free/reduced" ~ 1,
  ),
  test_prep = case_when(
    test_prep == "none" ~ 0,
    test_prep == "completed" ~ 1,
  )
)

```

```

    ),
    parent_marital_status = case_when(
      parent_marital_status == "married" ~ 0,
      parent_marital_status == "single" ~ 1,
      parent_marital_status == "widowed" ~ 2,
      parent_marital_status == "divorced" ~ 3,
    ),
    practice_sport = case_when(
      practice_sport == "never" ~ 0,
      practice_sport == "sometimes" ~ 1,
      practice_sport == "regularly" ~ 2,
    ),
    is_first_child = case_when(
      is_first_child == "no" ~ 0,
      is_first_child == "yes" ~ 1,
    ),
    transport_means = case_when(
      transport_means == "school_bus" ~ 0,
      transport_means == "private" ~ 1,
    ),
    wkly_study_hours = case_when(
      wkly_study_hours == "< 5" ~ 0,
      wkly_study_hours == "10-May" ~ 1,
      wkly_study_hours == "> 10" ~ 2,
    )
  ) |>
  drop_na()

```

```
## [1] "\ndata <- read_csv(\"./data.csv\") |>\n  janitor::clean_names() |>\n  mutate(\n    gender = case
```

```

'
# Deal with NA -- Calculate the column mean (round to integer) and plug it into NA cell
column_means <- round(colMeans(data, na.rm = TRUE), digits = 0)
for (col in names(data)) {
  data[[col]][is.na(data[[col]])] <- column_means[col]
}

head(data)
'

```

```
## [1] "\n# Deal with NA -- Calculate the column mean (round to integer) and plug it into NA cell\nncolum
```

```

# Another data set for EDA
data_long <- data |>
  pivot_longer(cols = c(math_score, reading_score, writing_score),
    names_to = "test", values_to = "score")

```

Summary

```

'
continuous_vars <- data[, c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)]
summary_df <- data.frame(
  Min = sapply(continuous_vars, min, na.rm = TRUE),
  Q1 = sapply(continuous_vars, function(x) quantile(x, probs = 0.25, na.rm = TRUE)), Median = sapply(conti
  Mean = sapply(continuous_vars, mean, na.rm = TRUE),
  Q3 = sapply(continuous_vars, function(x) quantile(x, probs = 0.75, na.rm = TRUE)), Max = sapply(continuo
)
kable(summary_df, caption = "Summary Statistics of Data", digits = 1)
'

```

```
## [1] "\ncontinuous_vars <- data[, c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)]\nsummary_df <- da
```

```

sum_data_fct =
  data |>
  dplyr::select(1:11) |>
  skimr::skim() |>
  dplyr::select(skim_variable, n_missing, complete_rate, factor.n_unique, factor.top_counts)

colnames(sum_data_fct) = c("Variable", "Missing", "Complete Rate", "Unique", "Top Counts")

knitr::kable(x = sum_data_fct, caption = "Categorical Variables pre-analysis", digits = 1)

```

Table 1: Categorical Variables pre-analysis

Variable	Missing	Complete Rate	Unique	Top Counts
gender	0	1.0	2	1: 488, 0: 460
ethnic_group	59	0.9	5	2: 277, 3: 237, 1: 171, 4: 124
parent_educ	392	0.6	4	1: 199, 2: 198, 3: 104, 4: 55
lunch_type	0	1.0	2	0: 617, 1: 331
test_prep	55	0.9	2	0: 571, 1: 322
parent_marital_status	49	0.9	4	0: 516, 1: 213, 3: 146, 2: 24
practice_sport	16	1.0	3	1: 477, 2: 343, 0: 112
is_first_child	30	1.0	2	1: 604, 0: 314
nr_siblings	46	1.0	8	1: 245, 2: 213, 3: 198, 0: 101
transport_means	102	0.9	2	0: 509, 1: 337
wkly_study_hours	37	1.0	3	1: 508, 0: 253, 2: 150

```

data =
  data |>
  drop_na()

sum_data_score =
  data |>
  dplyr::select(12:14) |>
  skimr::skim() |>
  dplyr::select(skim_variable, numeric.mean, numeric.sd, numeric.p0, numeric.p25, numeric.p50, numeric.p75, numeric.max)

colnames(sum_data_score) = c("Variable", "Mean", "SD", "Min", "Q1", "Median", "Q3", "Max")

knitr::kable(x = sum_data_score, caption = "Continuous Variables pre-analysis", digits = 1)

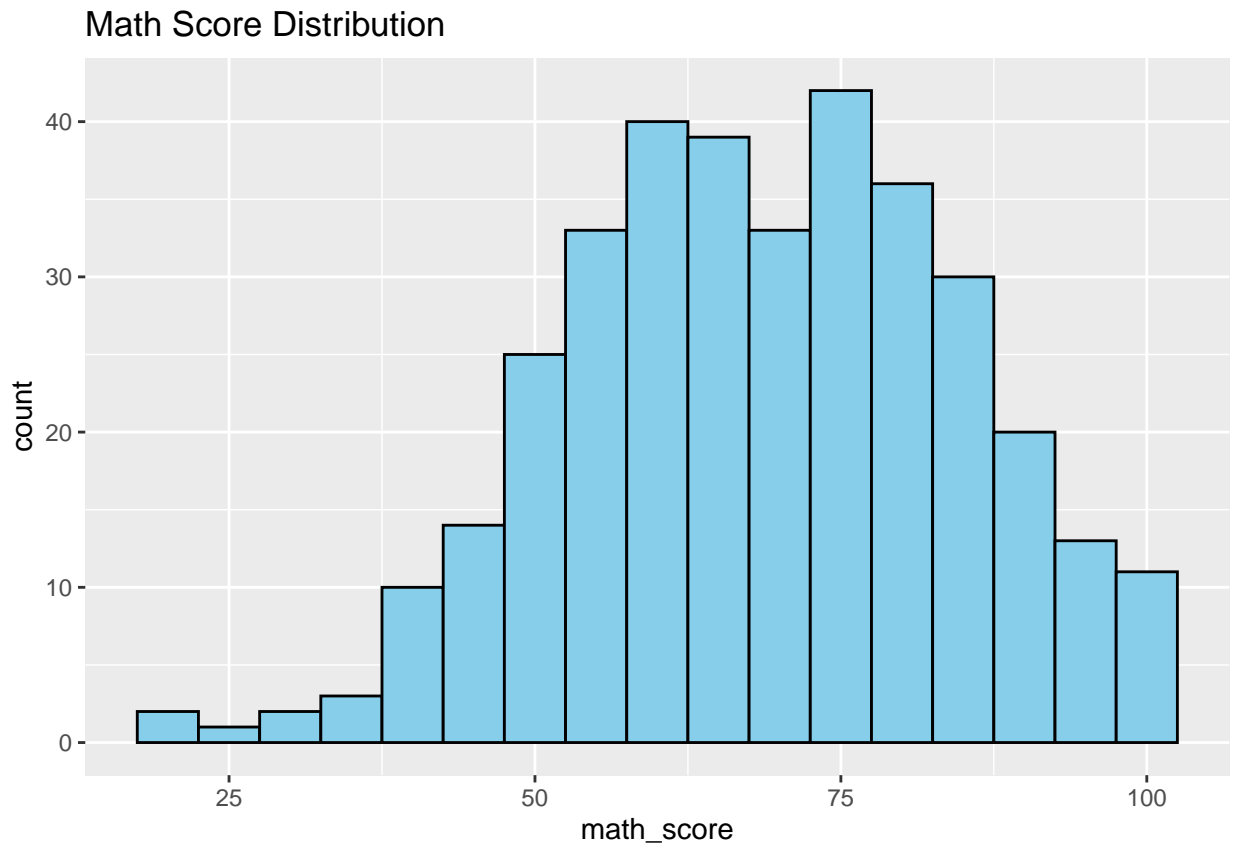
```

Table 2: Continuous Variables pre-analysis

Variable	Mean	SD	Min	Q1	Median	Q3	Max
math_score	68.7	15.9	18	57	69.0	81	100
reading_score	72.3	14.8	23	61	73.0	84	100
writing_score	72.0	15.2	19	62	72.5	84	100

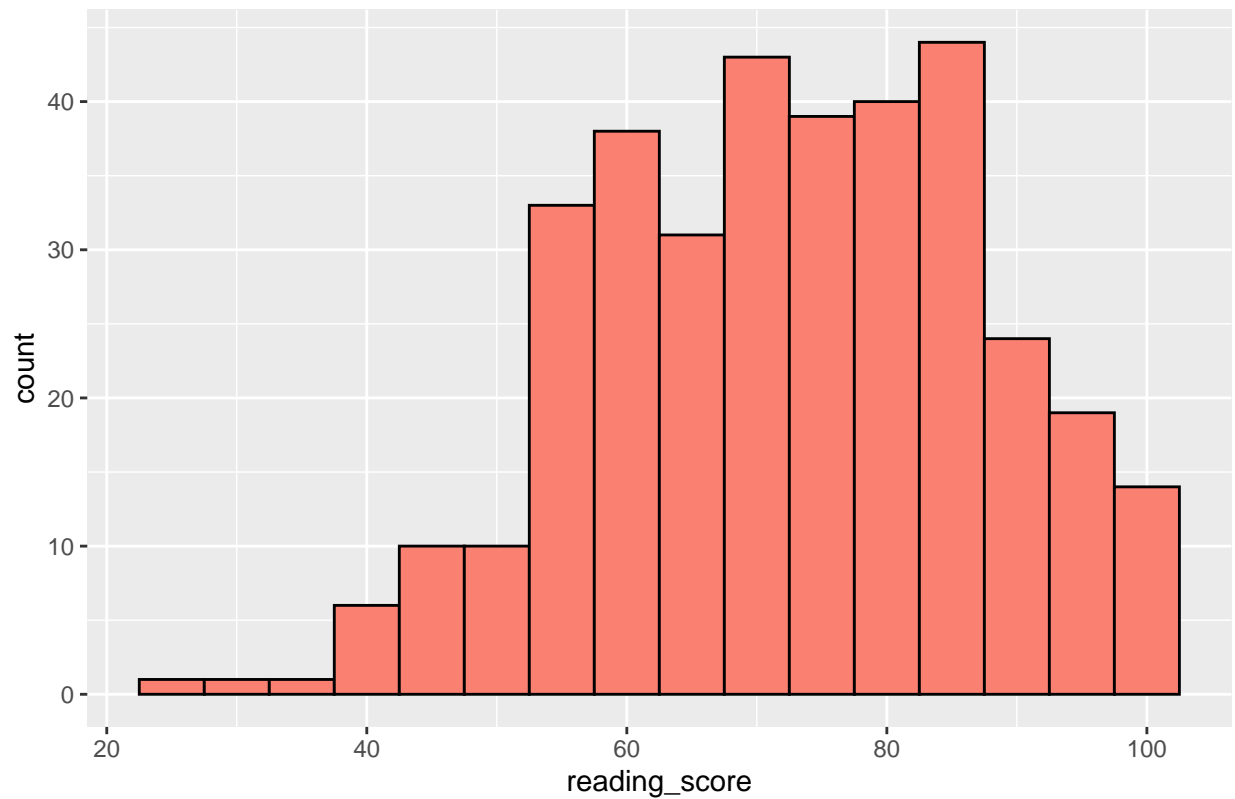
Histograms

```
ggplot(data, aes(x = math_score)) +  
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +  
  labs(title = "Math Score Distribution")
```



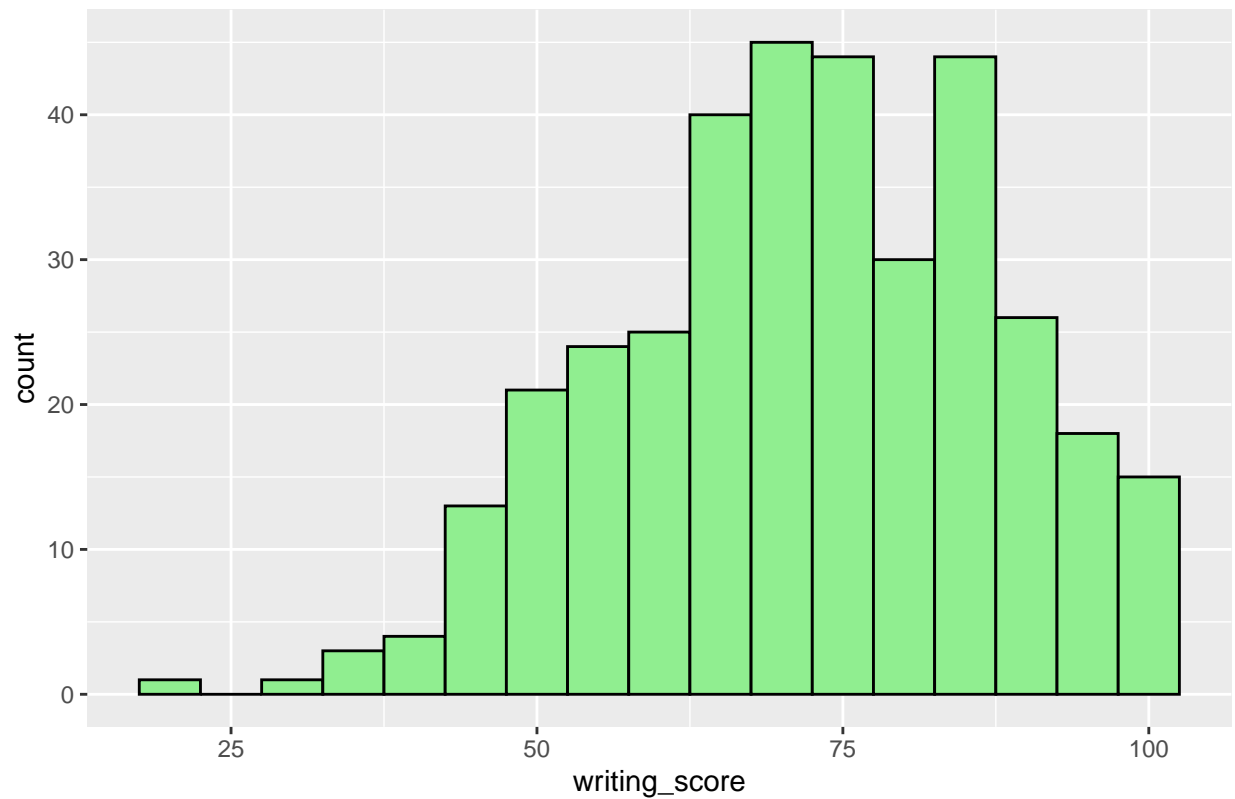
```
ggplot(data, aes(x = reading_score)) +  
  geom_histogram(binwidth = 5, fill = "salmon", color = "black") +  
  labs(title = "Reading Score Distribution")
```

Reading Score Distribution



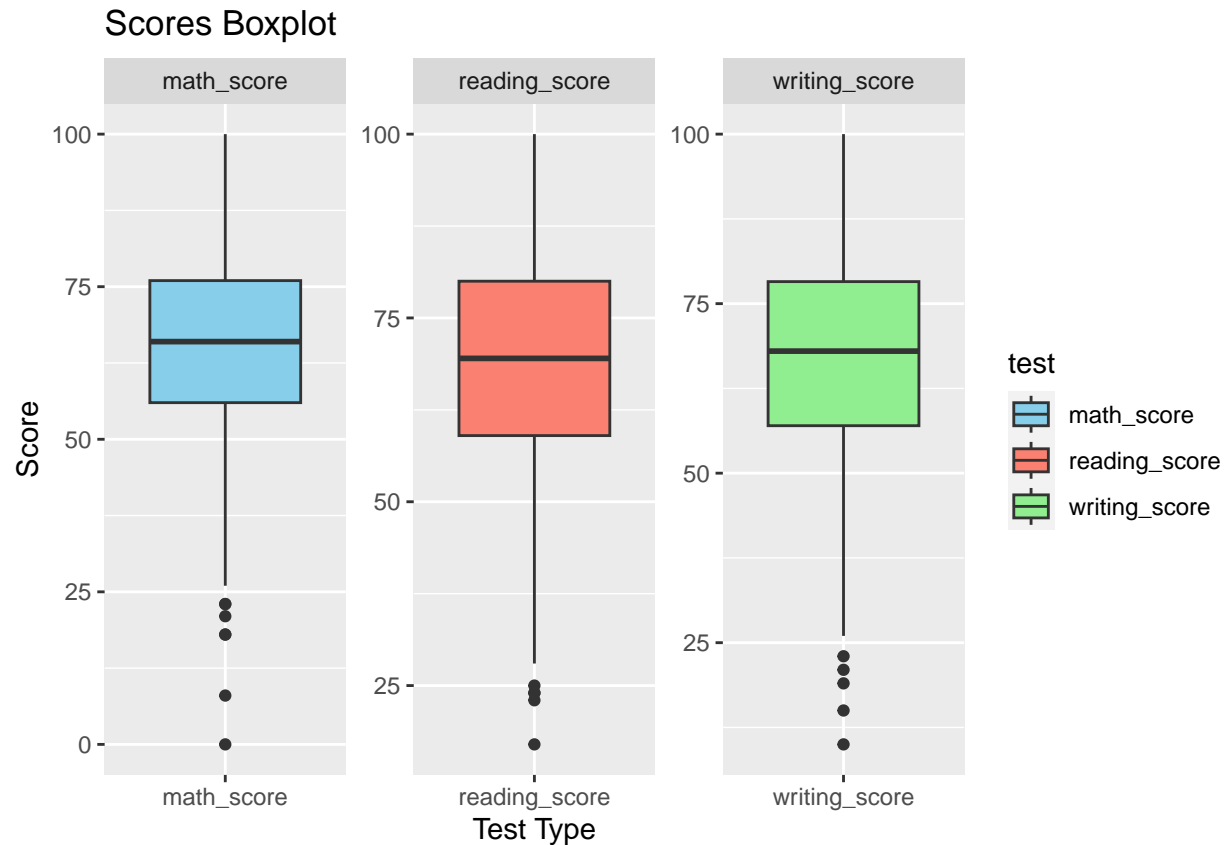
```
ggplot(data, aes(x = writing_score)) +  
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +  
  labs(title = "Writing Score Distribution")
```

Writing Score Distribution



Boxplots

```
ggplot(data_long, aes(x = test, y = score, fill = test)) +  
  geom_boxplot() +  
  labs(title = "Scores Boxplot", x = "Test Type", y = "Score") +  
  facet_wrap(~ test, scales = "free") +  
  scale_fill_manual(values = c("skyblue", "salmon", "lightgreen"))
```



Diagnostics

```
# Math
model_math_full = lm(math_score ~ . - reading_score - writing_score, data = data)

summary(model_math_full)
```

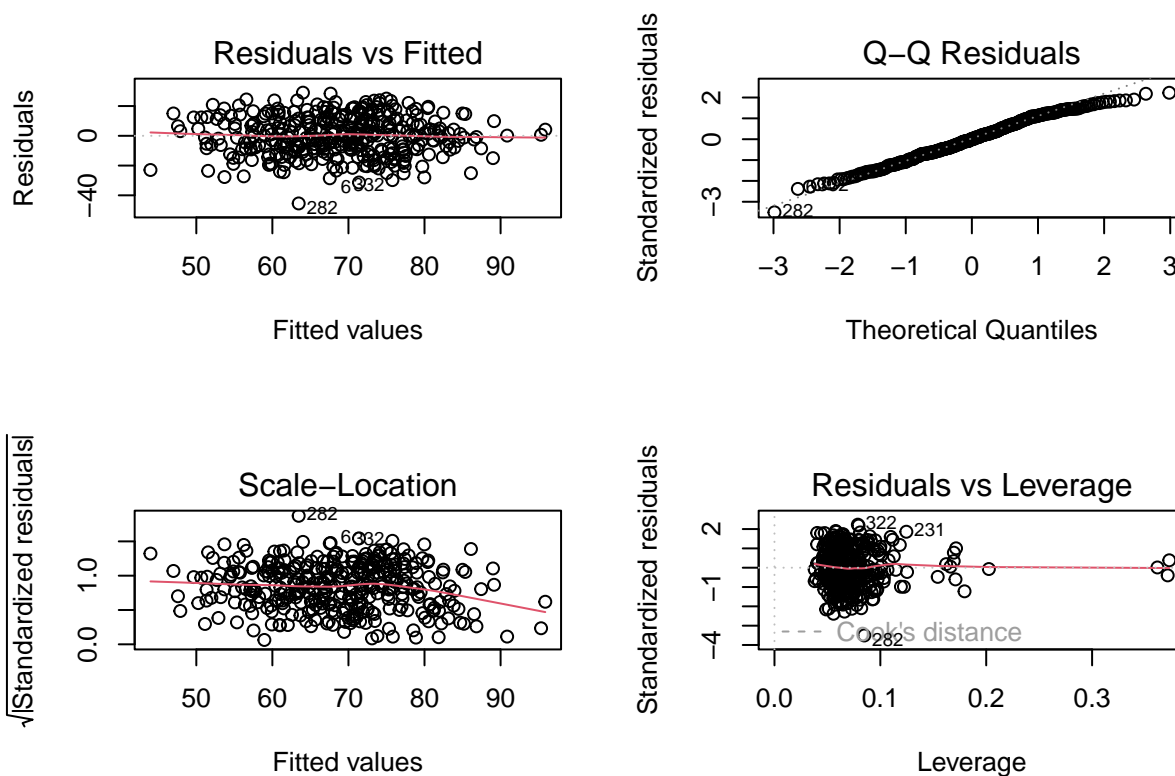
```
##
## Call:
## lm(formula = math_score ~ . - reading_score - writing_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.458  -8.961   0.089   9.800  28.981
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    62.3523     4.9540  12.586 < 2e-16 ***
## gender1        -3.6522     1.4958  -2.442  0.015150 *
## ethnic_group1    1.8120     3.2790   0.553  0.580912
## ethnic_group2   -1.1247     3.1319  -0.359  0.719748
## ethnic_group3    3.0342     3.1826   0.953  0.341109
```



```
## ethnic_group4      8.7423      3.3555      2.605 0.009598 **
## parent_educ2       1.8031      1.7975      1.003 0.316545
## parent_educ3       3.1775      2.0927      1.518 0.129886
## parent_educ4       4.0051      2.5782      1.553 0.121282
## lunch_type1      -12.1275      1.5423     -7.863 5.49e-14 ***
## test_prep1        5.7990      1.5706      3.692 0.000260 ***
## parent_marital_status1 -4.2006      1.8079     -2.323 0.020770 *
## parent_marital_status2  7.0930      4.7226      1.502 0.134083
## parent_marital_status3 -4.8362      2.1726     -2.226 0.026694 *
## practice_sport1     3.0566      2.3818      1.283 0.200295
## practice_sport2     3.2296      2.4896      1.297 0.195466
## is_first_child1    -0.3254      1.6378     -0.199 0.842638
## nr_siblings1      -0.1780      2.7665     -0.064 0.948739
## nr_siblings2      -1.1446      2.8721     -0.399 0.690507
## nr_siblings3       3.1546      2.8049      1.125 0.261548
## nr_siblings4       2.8587      3.3920      0.843 0.399963
## nr_siblings5       2.4937      3.9289      0.635 0.526071
## nr_siblings6      14.5158     13.9723      1.039 0.299617
## nr_siblings7       9.5593      8.3433      1.146 0.252735
## transport_means1    1.0585      1.5640      0.677 0.499003
## wkly_study_hours1   6.4822      1.7525      3.699 0.000254 ***
## wkly_study_hours2   4.2523      2.2536      1.887 0.060065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 327 degrees of freedom
## Multiple R-squared:  0.3256, Adjusted R-squared:  0.272
## F-statistic: 6.073 on 26 and 327 DF, p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model_math_full)
```

```
## Warning:      :
## 186
```



```
# Reading
model_reading_full = lm(reading_score ~ . - math_score - writing_score, data = data)

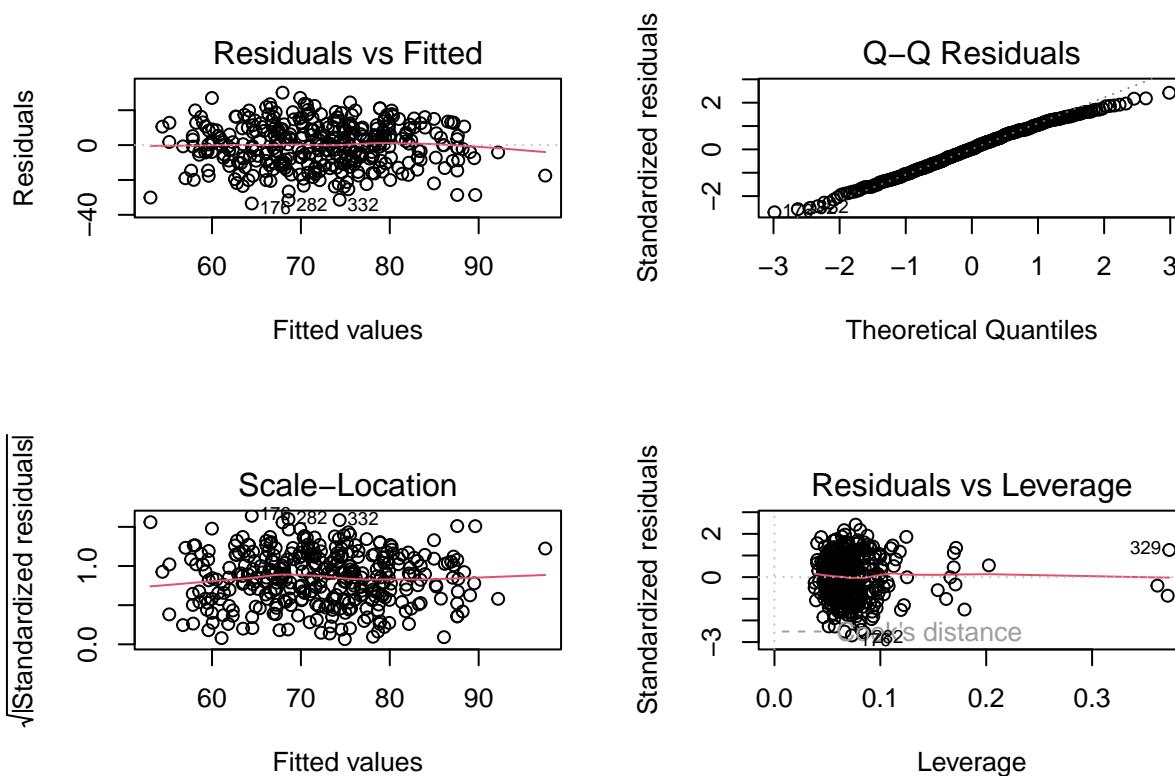
summary(model_reading_full)
```

```
##
## Call:
## lm(formula = reading_score ~ . - math_score - writing_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -33.470  -8.942   0.403   9.553  30.063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    59.3627     4.7169  12.585 < 2e-16 ***
## gender1         8.2587     1.4242   5.799 1.57e-08 ***
## ethnic_group1    1.4533     3.1220    0.466  0.64188
## ethnic_group2   -0.5044     2.9819   -0.169  0.86578
## ethnic_group3    2.8080     3.0302    0.927  0.35479
## ethnic_group4    4.7359     3.1949    1.482  0.13921
## parent_educ2     2.6502     1.7114    1.549  0.12246
## parent_educ3     4.5816     1.9925    2.299  0.02211 *
## parent_educ4     6.4240     2.4548    2.617  0.00929 **
```

```
## lunch_type1          -7.8783      1.4685   -5.365  1.54e-07 ***
## test_prep1           7.6036      1.4954    5.085  6.21e-07 ***
## parent_marital_status1 -4.6412      1.7214   -2.696  0.00738 **
## parent_marital_status2  4.6364      4.4966    1.031  0.30325
## parent_marital_status3 -4.2660      2.0686   -2.062  0.03997 *
## practice_sport1       1.9156      2.2678    0.845  0.39890
## practice_sport2       1.2989      2.3705    0.548  0.58408
## is_first_child1       0.6384      1.5594    0.409  0.68252
## nr_siblings1          0.4794      2.6341    0.182  0.85569
## nr_siblings2         -1.4869      2.7347   -0.544  0.58700
## nr_siblings3          1.8958      2.6706    0.710  0.47830
## nr_siblings4          2.3345      3.2296    0.723  0.47028
## nr_siblings5         -1.4797      3.7408   -0.396  0.69269
## nr_siblings6         11.7473     13.3034    0.883  0.37787
## nr_siblings7          7.7275      7.9439    0.973  0.33139
## transport_means1      0.5365      1.4891    0.360  0.71890
## wkly_study_hours1     5.3310      1.6686    3.195  0.00154 **
## wkly_study_hours2     1.1401      2.1458    0.531  0.59557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.9 on 327 degrees of freedom
## Multiple R-squared:  0.2971, Adjusted R-squared:  0.2412
## F-statistic: 5.315 on 26 and 327 DF,  p-value: 6.451e-14
```

```
par(mfrow = c(2,2))
plot(model_reading_full)
```

```
## Warning:           :
## 186
```



```
# Writing
model_writing_full = lm(writing_score ~ . - reading_score - math_score, data = data)

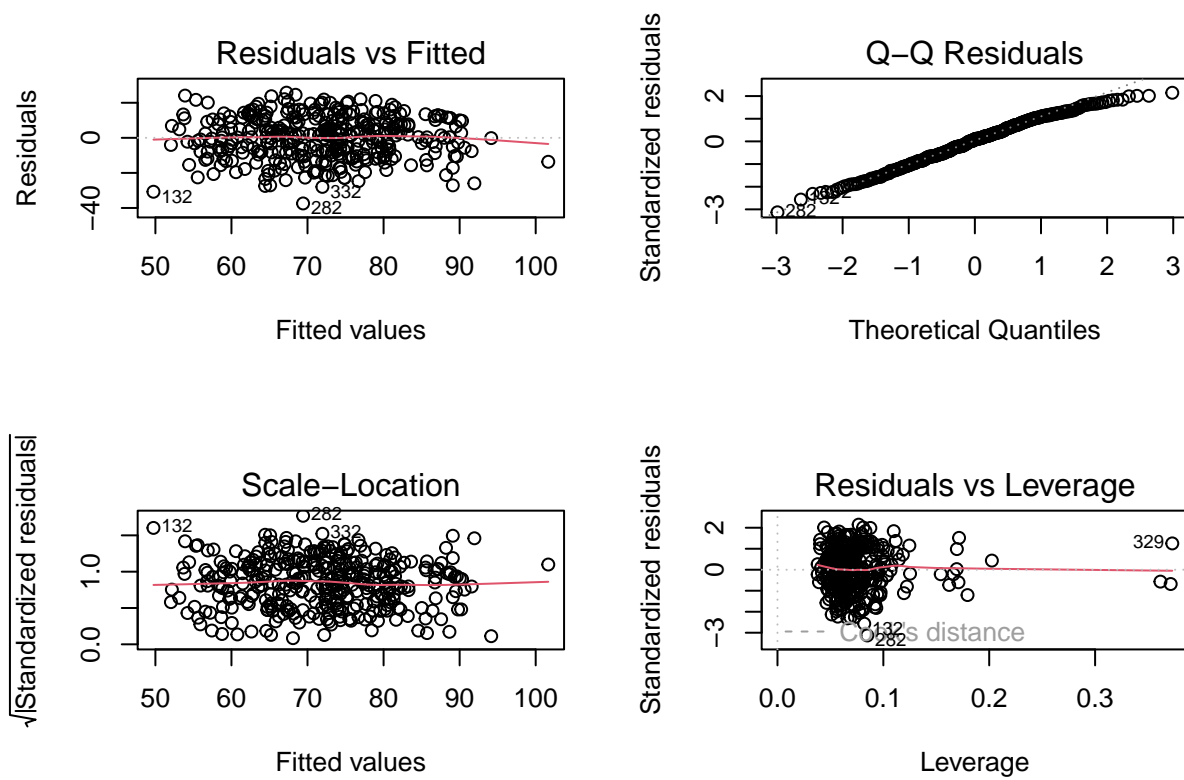
summary(model_writing_full)
```

```
##
## Call:
## lm(formula = writing_score ~ . - reading_score - math_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -37.416  -8.131   1.123   9.165  25.765
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    55.1871     4.5675  12.083  < 2e-16 ***
## gender1         10.0433     1.3791   7.283 2.46e-12 ***
## ethnic_group1     1.7982     3.0232   0.595 0.552382
## ethnic_group2     0.7708     2.8875   0.267 0.789684
## ethnic_group3     5.5577     2.9343   1.894 0.059101 .
## ethnic_group4     5.5666     3.0937   1.799 0.072893 .
## parent_educ2      2.0224     1.6572   1.220 0.223203
## parent_educ3      4.5673     1.9294   2.367 0.018507 *
## parent_educ4      7.5525     2.3771   3.177 0.001629 **
```

```
## lunch_type1          -8.9424      1.4220   -6.289 1.03e-09 ***
## test_prep1           9.6428      1.4480    6.659 1.16e-10 ***
## parent_marital_status1 -4.5781      1.6669   -2.747 0.006356 **
## parent_marital_status2  5.2451      4.3542    1.205 0.229221
## parent_marital_status3 -4.4305      2.0031   -2.212 0.027669 *
## practice_sport1        3.3011      2.1960    1.503 0.133746
## practice_sport2        3.0186      2.2954    1.315 0.189415
## is_first_child1       -0.2525      1.5100   -0.167 0.867295
## nr_siblings1          0.3186      2.5507    0.125 0.900665
## nr_siblings2         -1.2993      2.6481   -0.491 0.624008
## nr_siblings3          2.2515      2.5860    0.871 0.384594
## nr_siblings4          2.9536      3.1273    0.944 0.345630
## nr_siblings5         -0.5419      3.6224   -0.150 0.881167
## nr_siblings6         14.3830     12.8821    1.117 0.265024
## nr_siblings7          8.0232      7.6923    1.043 0.297708
## transport_means1       0.9938      1.4420    0.689 0.491208
## wkly_study_hours1      5.4344      1.6157    3.363 0.000861 ***
## wkly_study_hours2      2.0335      2.0778    0.979 0.328454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.5 on 327 degrees of freedom
## Multiple R-squared:  0.3762, Adjusted R-squared:  0.3266
## F-statistic: 7.586 on 26 and 327 DF,  p-value: < 2.2e-16
```

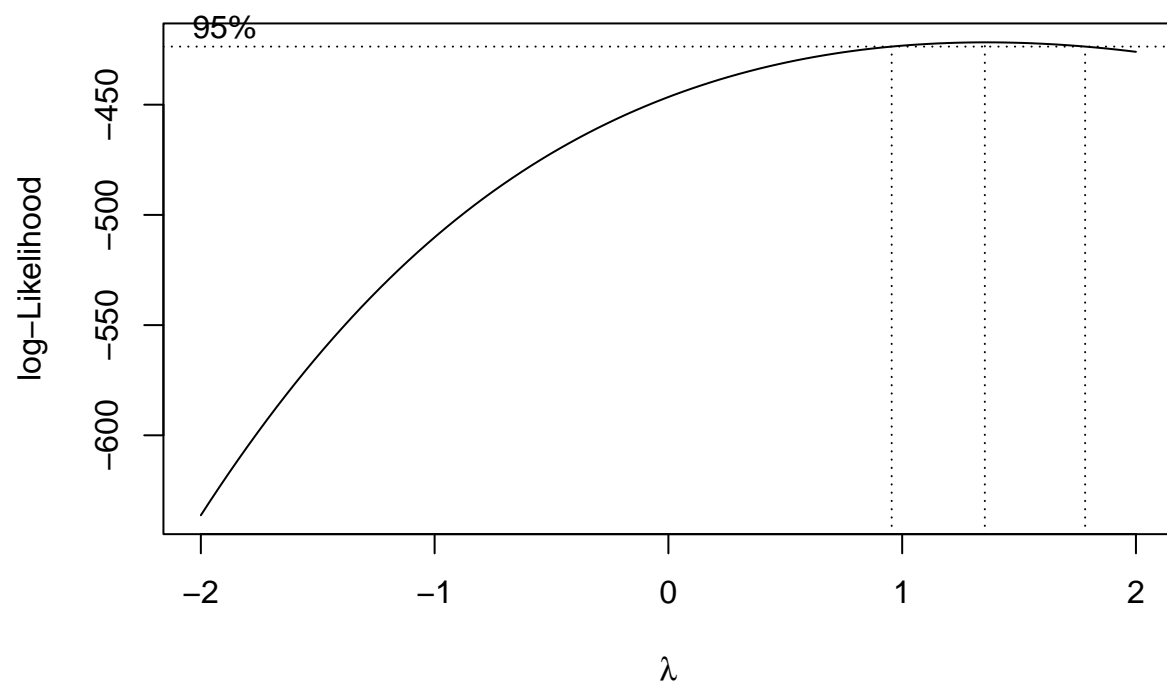
```
par(mfrow = c(2,2))
plot(model_writing_full)
```

```
## Warning:           :
## 186
```



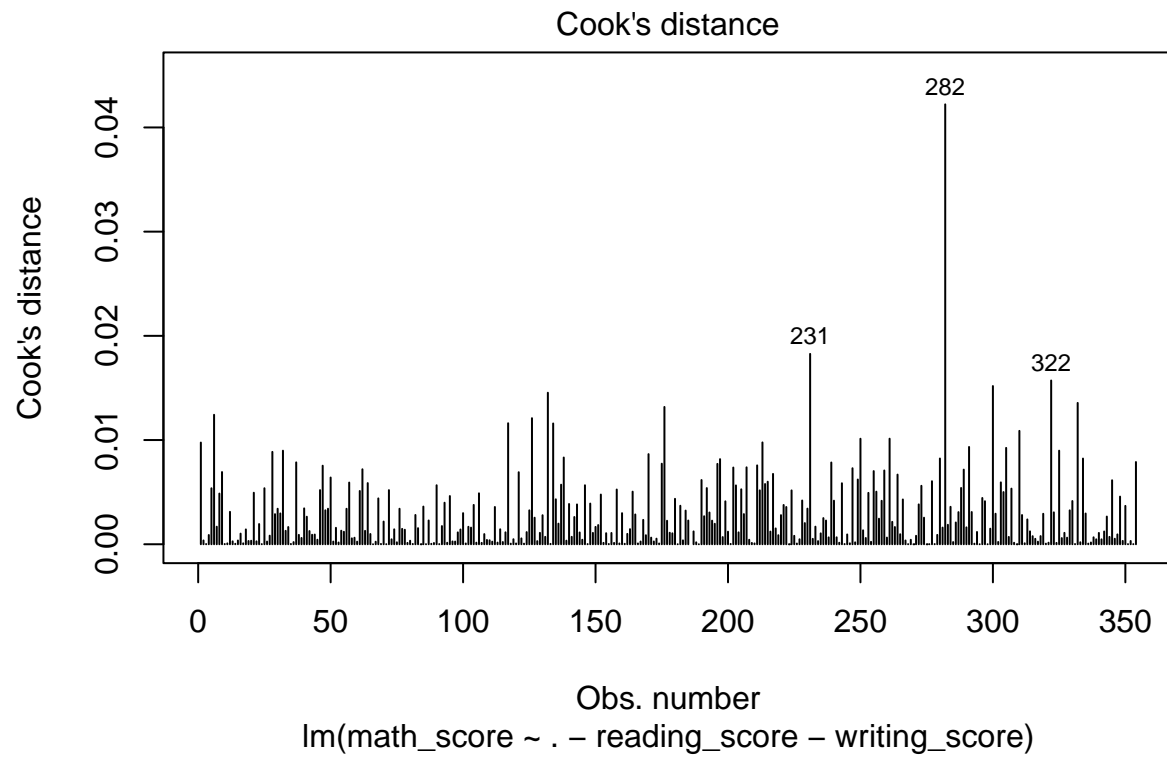
Transformation

```
boxcox(model_reading_full)
```

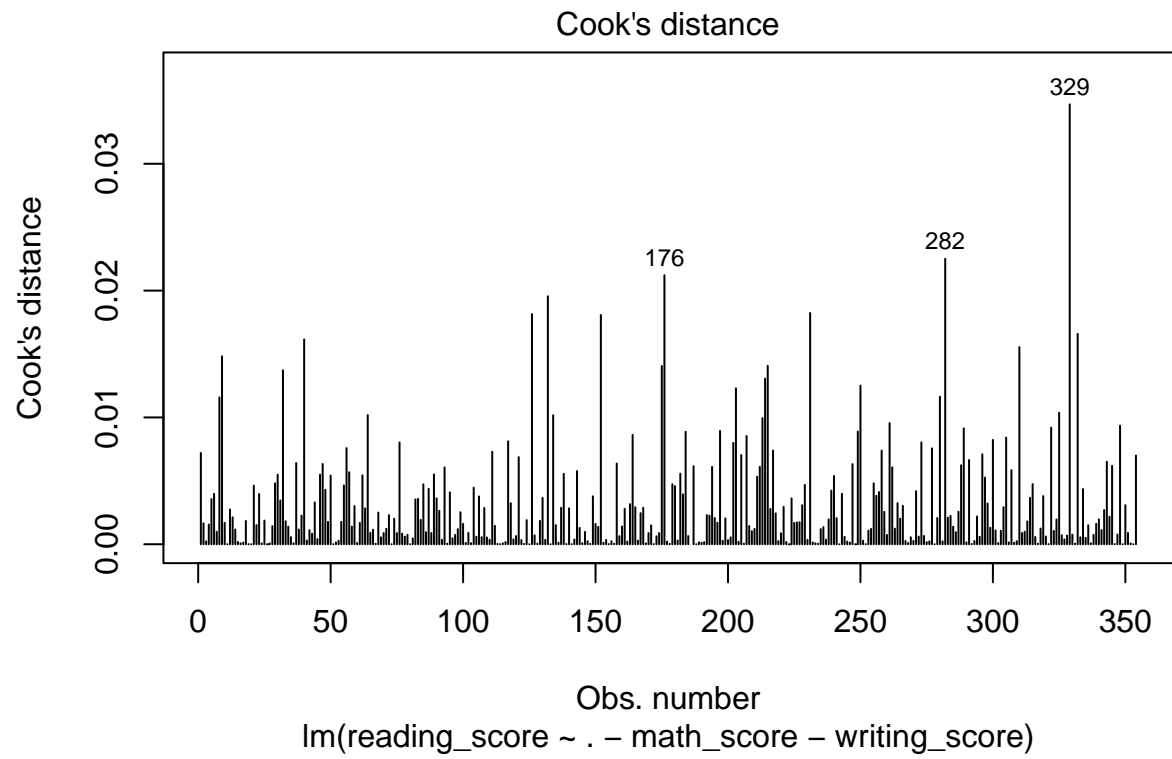


Outlier and influence points

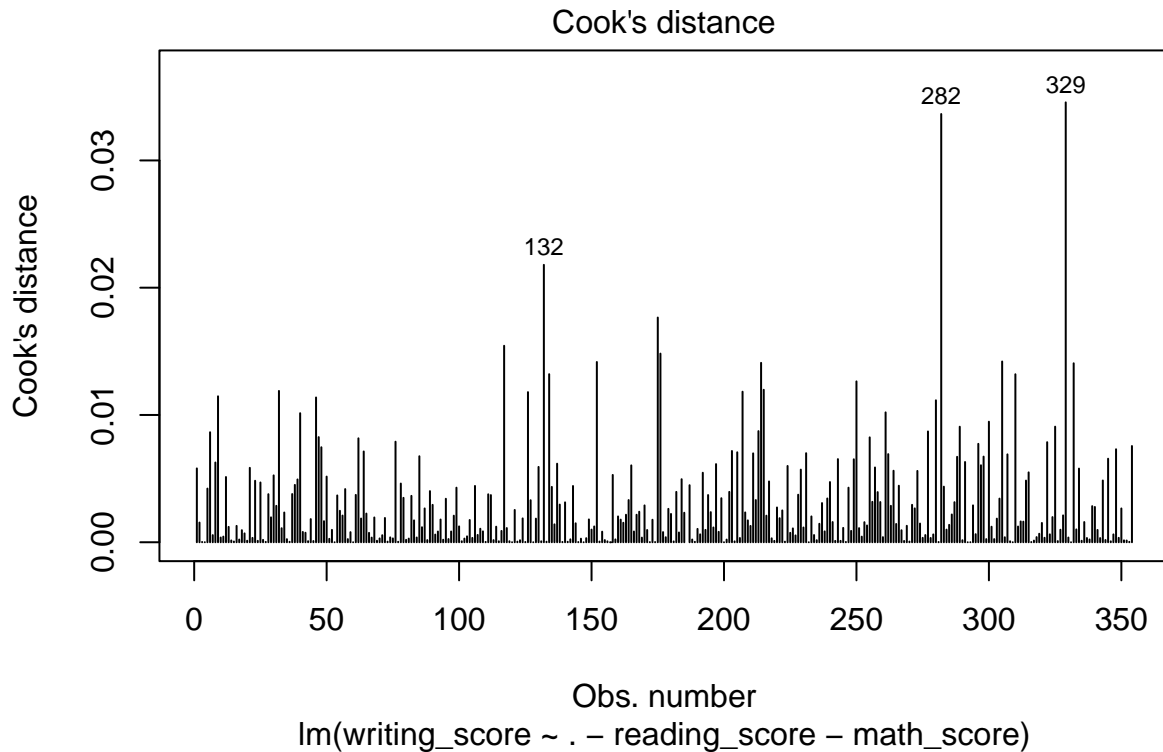
```
plot(model_math_full, which = 4)
```



```
plot(model_reading_full, which = 4)
```

```
plot(model_writing_full, which = 4)
```



Multicollinearity

```
# check VIF
performance::check_collinearity(model_math_full)
```

```
## # Check for Multicollinearity
```

```
##
```

```
## Low Correlation
```

```
##
```

	Term	VIF	VIF 95% CI	Increased SE	Tolerance
##	gender	1.06	[1.01, 1.35]	1.03	0.94
##	ethnic_group	1.24	[1.13, 1.43]	1.11	0.81
##	parent_educ	1.22	[1.12, 1.41]	1.10	0.82
##	lunch_type	1.05	[1.01, 1.40]	1.03	0.95
##	test_prep	1.09	[1.03, 1.31]	1.05	0.91
##	parent_marital_status	1.17	[1.08, 1.36]	1.08	0.86
##	practice_sport	1.17	[1.08, 1.36]	1.08	0.86
##	is_first_child	1.15	[1.07, 1.35]	1.07	0.87
##	nr_siblings	1.54	[1.38, 1.78]	1.24	0.65
##	transport_means	1.11	[1.04, 1.32]	1.05	0.90
##	wkly_study_hours	1.14	[1.06, 1.33]	1.07	0.88

```
## Tolerance 95% CI
```

```
## [0.74, 0.99]
```

```
##      [0.70, 0.88]
##      [0.71, 0.89]
##      [0.71, 0.99]
##      [0.76, 0.97]
##      [0.74, 0.93]
##      [0.74, 0.93]
##      [0.74, 0.94]
##      [0.56, 0.73]
##      [0.76, 0.97]
##      [0.75, 0.95]
```

```
performance::check_collinearity(model_reading_full)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##           Term  VIF   VIF 95% CI Increased SE Tolerance
##           gender 1.06 [1.01, 1.35]         1.03      0.94
##           ethnic_group 1.24 [1.13, 1.43]         1.11      0.81
##           parent_educ 1.22 [1.12, 1.41]         1.10      0.82
##           lunch_type 1.05 [1.01, 1.40]         1.03      0.95
##           test_prep 1.09 [1.03, 1.31]         1.05      0.91
## parent_marital_status 1.17 [1.08, 1.36]         1.08      0.86
##           practice_sport 1.17 [1.08, 1.36]         1.08      0.86
##           is_first_child 1.15 [1.07, 1.35]         1.07      0.87
##           nr_siblings 1.54 [1.38, 1.78]         1.24      0.65
##           transport_means 1.11 [1.04, 1.32]         1.05      0.90
##           wkly_study_hours 1.14 [1.06, 1.33]         1.07      0.88
## Tolerance 95% CI
##      [0.74, 0.99]
##      [0.70, 0.88]
##      [0.71, 0.89]
##      [0.71, 0.99]
##      [0.76, 0.97]
##      [0.74, 0.93]
##      [0.74, 0.93]
##      [0.74, 0.94]
##      [0.56, 0.73]
##      [0.76, 0.97]
##      [0.75, 0.95]
```

```
performance::check_collinearity(model_writing_full)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##           Term  VIF   VIF 95% CI Increased SE Tolerance
##           gender 1.06 [1.01, 1.35]         1.03      0.94
##           ethnic_group 1.24 [1.13, 1.43]         1.11      0.81
##           parent_educ 1.22 [1.12, 1.41]         1.10      0.82
##           lunch_type 1.05 [1.01, 1.40]         1.03      0.95
```

```
##          test_prep 1.09 [1.03, 1.31]          1.05          0.91
## parent_marital_status 1.17 [1.08, 1.36]          1.08          0.86
##          practice_sport 1.17 [1.08, 1.36]          1.08          0.86
##          is_first_child 1.15 [1.07, 1.35]          1.07          0.87
##          nr_siblings 1.54 [1.38, 1.78]          1.24          0.65
##          transport_means 1.11 [1.04, 1.32]          1.05          0.90
##          wkly_study_hours 1.14 [1.06, 1.33]          1.07          0.88
## Tolerance 95% CI
## [0.74, 0.99]
## [0.70, 0.88]
## [0.71, 0.89]
## [0.71, 0.99]
## [0.76, 0.97]
## [0.74, 0.93]
## [0.74, 0.93]
## [0.74, 0.94]
## [0.56, 0.73]
## [0.76, 0.97]
## [0.75, 0.95]
```

Model building for math

```
# backward model
step(model_math_full, direction='backward')

## Start:  AIC=1871.41
## math_score ~ (gender + ethnic_group + parent_educ + lunch_type +
##          test_prep + parent_marital_status + practice_sport + is_first_child +
##          nr_siblings + transport_means + wkly_study_hours + reading_score +
##          writing_score) - reading_score - writing_score
##
##          Df Sum of Sq  RSS    AIC
## - nr_siblings      7   1388.4 61456 1865.5
## - parent_educ       3    667.0 60735 1869.3
## - practice_sport    2    344.1 60412 1869.4
## - is_first_child    1      7.3 60075 1869.5
## - transport_means    1     84.1 60152 1869.9
## <none>                  60068 1871.4
## - gender            1   1095.1 61163 1875.8
## - parent_marital_status 3   2192.4 62260 1878.1
## - wkly_study_hours   2   2514.1 62582 1881.9
## - test_prep          1   2504.3 62572 1883.9
## - ethnic_group        4   3792.0 63860 1885.1
## - lunch_type         1  11357.8 71425 1930.7
##
## Step:  AIC=1865.5
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##          test_prep + parent_marital_status + practice_sport + is_first_child +
##          transport_means + wkly_study_hours
##
##          Df Sum of Sq  RSS    AIC
```

```

## - parent_educ      3      654.6 62111 1863.2
## - is_first_child   1        0.9 61457 1863.5
## - practice_sport    2      373.0 61829 1863.6
## - transport_means    1       57.9 61514 1863.8
## <none>                61456 1865.5
## - gender            1     1188.3 62644 1870.3
## - parent_marital_status 3     2362.9 63819 1872.9
## - wkly_study_hours   2     2350.7 63807 1874.8
## - test_prep          1     2571.8 64028 1878.0
## - ethnic_group       4     4102.7 65559 1880.4
## - lunch_type        1    12401.9 73858 1928.6
##
## Step: AIC=1863.25
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + practice_sport + is_first_child +
##   transport_means + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - practice_sport      2      306.8 62417 1861.0
## - is_first_child       1        2.1 62113 1861.3
## - transport_means      1       31.6 62142 1861.4
## <none>                  62111 1863.2
## - gender              1     1164.4 63275 1867.8
## - parent_marital_status 3     2366.7 64477 1870.5
## - wkly_study_hours     2     2220.2 64331 1871.7
## - test_prep            1     2823.2 64934 1877.0
## - ethnic_group         4     4148.6 66259 1878.1
## - lunch_type           1    12325.1 74436 1925.3
##
## Step: AIC=1861
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + is_first_child + transport_means +
##   wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - is_first_child       1        7.6 62425 1859.0
## - transport_means      1       26.3 62444 1859.1
## <none>                  62417 1861.0
## - gender              1     1163.9 63581 1865.5
## - parent_marital_status 3     2337.1 64755 1868.0
## - wkly_study_hours     2     2207.5 64625 1869.3
## - test_prep            1     2830.8 65248 1874.7
## - ethnic_group         4     4111.9 66529 1875.6
## - lunch_type           1    12239.5 74657 1922.4
##
## Step: AIC=1859.04
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + transport_means + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - transport_means      1       27.7 62453 1857.2
## <none>                  62425 1859.0
## - gender              1     1158.1 63583 1863.5
## - parent_marital_status 3     2347.6 64773 1866.1

```

```

## - wkly_study_hours      2      2201.8 64627 1867.3
## - test_prep             1      2826.4 65251 1872.7
## - ethnic_group          4      4105.4 66531 1873.6
## - lunch_type            1     12233.8 74659 1920.4
##
## Step: AIC=1857.2
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## <none>                    62453 1857.2
## - gender                  1     1160.8 63614 1861.7
## - parent_marital_status   3     2320.8 64774 1864.1
## - wkly_study_hours        2     2192.8 64646 1865.4
## - test_prep               1     2920.0 65373 1871.4
## - ethnic_group            4     4097.1 66550 1871.7
## - lunch_type              1     12211.5 74664 1918.4

##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + lunch_type +
##   test_prep + parent_marital_status + wkly_study_hours, data = data)
##
## Coefficients:
##      (Intercept)              gender1          ethnic_group1
##           67.3260             -3.7049              2.4461
##      ethnic_group2          ethnic_group3          ethnic_group4
##           0.3026              4.1687              10.1791
##      lunch_type1          test_prep1 parent_marital_status1
##          -12.3773              6.0788             -4.0821
## parent_marital_status2 parent_marital_status3      wkly_study_hours1
##           6.7982             -5.2507              5.9171
##      wkly_study_hours2
##           3.8301

```

```

model_math_fit_back = lm(formula = math_score ~ gender + ethnic_group + parent_educ +
  lunch_type + test_prep + parent_marital_status + practice_sport +
  is_first_child + wkly_study_hours, data = data)

```

```

summary(model_math_fit_back)

```

```

##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##   lunch_type + test_prep + parent_marital_status + practice_sport +
##   is_first_child + wkly_study_hours, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -42.641  -9.388   0.444  10.841  29.060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)

```

```
## (Intercept)          63.3058      4.0723 15.545 < 2e-16 ***
## gender1             -3.7768      1.4786 -2.554 0.011080 *
## ethnic_group1        2.0233      3.2739  0.618 0.536983
## ethnic_group2       -0.1921      3.1097 -0.062 0.950767
## ethnic_group3        3.5985      3.1572  1.140 0.255191
## ethnic_group4        9.8452      3.3254  2.961 0.003289 **
## parent_educ2         1.6680      1.7628  0.946 0.344724
## parent_educ3         3.1571      2.0672  1.527 0.127641
## parent_educ4         3.7243      2.5498  1.461 0.145058
## lunch_type1        -12.4609      1.5198 -8.199 5.22e-15 ***
## test_prep1          5.9501      1.5447  3.852 0.000140 ***
## parent_marital_status1 -4.1882      1.7844 -2.347 0.019505 *
## parent_marital_status2  7.3458      4.7089  1.560 0.119707
## parent_marital_status3 -4.9516      2.1536 -2.299 0.022104 *
## practice_sport1      3.1345      2.3452  1.337 0.182276
## practice_sport2      3.2766      2.4641  1.330 0.184500
## is_first_child1     -0.1431      1.5713 -0.091 0.927481
## wkly_study_hours1     6.1263      1.7189  3.564 0.000418 ***
## wkly_study_hours2     4.2272      2.2378  1.889 0.059752 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 335 degrees of freedom
## Multiple R-squared:  0.3094, Adjusted R-squared:  0.2723
## F-statistic: 8.338 on 18 and 335 DF,  p-value: < 2.2e-16
```

```
# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

cv_object_math = cv.glmnet(as.matrix(data[1:11]), data$math_score,
                           lambda = lambda_seq,
                           nfolds = 5)

model_math_lasso = glmnet(as.matrix(data[1:11]), data$math_score, lambda = cv_object_math$lambda.min, a
coef(model_math_lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  62.7158706
## gender      -3.4172517
## ethnic_group 2.0740949
## parent_educ  0.9804808
## lunch_type  -11.7678104
## test_prep    5.0255504
## parent_marital_status -1.0446103
## practice_sport 0.4391390
## is_first_child .
## nr_siblings   0.7146589
## transport_means .
## wkly_study_hours 2.4395500
```

```
model_math_lasso$dev.ratio
```

```
## [1] 0.2622201
```

Model building for reading

```
# backward model
step(model_reading_full, direction='backward')

## Start:  AIC=1836.68
## reading_score ~ (gender + ethnic_group + parent_educ + lunch_type +
##   test_prep + parent_marital_status + practice_sport + is_first_child +
##   nr_siblings + transport_means + wkly_study_hours + math_score +
##   writing_score) - math_score - writing_score
##
##           Df Sum of Sq  RSS    AIC
## - nr_siblings      7    887.9 55342 1828.4
## - practice_sport    2    123.8 54578 1833.5
## - transport_means    1     21.6 54476 1834.8
## - is_first_child    1     27.9 54482 1834.9
## - ethnic_group      4   1227.5 55682 1836.6
## <none>                      54454 1836.7
## - parent_educ       3   1558.4 56013 1840.7
## - parent_marital_status 3   1908.7 56363 1842.9
## - wkly_study_hours   2   2004.0 56459 1845.5
## - test_prep          1   4305.6 58760 1861.6
## - lunch_type         1   4793.1 59248 1864.5
## - gender             1   5599.8 60054 1869.3
##
## Step:  AIC=1828.41
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##   test_prep + parent_marital_status + practice_sport + is_first_child +
##   transport_means + wkly_study_hours
##
##           Df Sum of Sq  RSS    AIC
## - practice_sport    2    145.3 55488 1825.3
## - transport_means    1     11.4 55354 1826.5
## - is_first_child    1     40.1 55382 1826.7
## <none>                      55342 1828.4
## - ethnic_group      4   1318.8 56661 1828.7
## - parent_educ       3   1681.4 57024 1833.0
## - parent_marital_status 3   1924.1 57267 1834.5
## - wkly_study_hours   2   1969.7 57312 1836.8
## - test_prep          1   4222.4 59565 1852.4
## - lunch_type         1   5437.8 60780 1859.6
## - gender             1   5693.8 61036 1861.1
##
## Step:  AIC=1825.34
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##   test_prep + parent_marital_status + is_first_child + transport_means +
##   wkly_study_hours
##
##           Df Sum of Sq  RSS    AIC
## - transport_means    1      5.8 55493 1823.4
## - is_first_child    1    40.9 55529 1823.6
## <none>                      55488 1825.3
```



```

## - ethnic_group      4      1294.8 56782 1825.5
## - parent_educ       3      1654.8 57143 1829.7
## - parent_marital_status 3      1902.9 57391 1831.3
## - wkly_study_hours  2      1959.0 57447 1833.6
## - test_prep         1      4316.3 59804 1849.8
## - lunch_type        1      5421.7 60909 1856.3
## - gender            1      5678.5 61166 1857.8
##
## Step: AIC=1823.37
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + is_first_child + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - is_first_child      1      39.4 55533 1821.6
## <none>                  55493 1823.4
## - ethnic_group        4      1295.8 56789 1823.5
## - parent_educ         3      1649.4 57143 1827.7
## - parent_marital_status 3      1899.1 57393 1829.3
## - wkly_study_hours    2      1958.5 57452 1831.7
## - test_prep           1      4422.7 59916 1848.5
## - lunch_type          1      5422.5 60916 1854.4
## - gender              1      5674.9 61168 1855.8
##
## Step: AIC=1821.62
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## <none>                  55533 1821.6
## - ethnic_group        4      1305.9 56839 1821.8
## - parent_educ         3      1654.8 57188 1826.0
## - parent_marital_status 3      1899.5 57432 1827.5
## - wkly_study_hours    2      1974.9 57508 1830.0
## - test_prep           1      4531.6 60064 1847.4
## - lunch_type          1      5440.2 60973 1852.7
## - gender              1      5644.2 61177 1853.9
##
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##      lunch_type + test_prep + parent_marital_status + wkly_study_hours,
##      data = data)
##
## Coefficients:
##      (Intercept)                gender1                ethnic_group1
##             61.6474                 8.1816                 1.8945
##      ethnic_group2                ethnic_group3                ethnic_group4
##             0.3778                 3.3789                 5.6870
##      parent_educ2                parent_educ3                parent_educ4
##             2.3964                 4.6728                 6.4917
##      lunch_type1                test_prep1 parent_marital_status1
##             -8.2631                 7.6175                 -4.5976
##      parent_marital_status2 parent_marital_status3                wkly_study_hours1
##             4.1841                 -4.3042                 5.1565

```

```
##      wkly_study_hours2
##              1.0458

model_reading_back = lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
  lunch_type + test_prep + parent_marital_status + is_first_child +
  transport_means + wkly_study_hours, data = data)
summary(model_reading_back)

##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##      lunch_type + test_prep + parent_marital_status + is_first_child +
##      transport_means + wkly_study_hours, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -32.522  -9.335   0.253   9.491  29.948
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.0959     3.4189  17.870 < 2e-16 ***
## gender1           8.2151     1.4010   5.864 1.08e-08 ***
## ethnic_group1      1.8440     3.0962   0.596 0.55187
## ethnic_group2      0.3221     2.9318   0.110 0.91257
## ethnic_group3      3.3272     2.9801   1.116 0.26502
## ethnic_group4      5.6186     3.1503   1.784 0.07540 .
## parent_educ2       2.4730     1.6822   1.470 0.14248
## parent_educ3       4.7430     1.9674   2.411 0.01645 *
## parent_educ4       6.4579     2.4012   2.689 0.00751 **
## lunch_type1       -8.2690     1.4432  -5.730 2.24e-08 ***
## test_prep1        7.5208     1.4711   5.112 5.35e-07 ***
## parent_marital_status1 -4.5595     1.6944  -2.691 0.00748 **
## parent_marital_status2  4.3781     4.4330   0.988 0.32405
## parent_marital_status3 -4.3645     2.0421  -2.137 0.03330 *
## is_first_child1     0.7327     1.4725   0.498 0.61910
## transport_means1     0.2718     1.4551   0.187 0.85195
## wkly_study_hours1     5.1383     1.6296   3.153 0.00176 **
## wkly_study_hours2     1.0442     2.1217   0.492 0.62294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 336 degrees of freedom
## Multiple R-squared:  0.2837, Adjusted R-squared:  0.2475
## F-statistic: 7.829 on 17 and 336 DF,  p-value: < 2.2e-16

# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

cv_object_reading = cv.glmnet(as.matrix(data[1:11]), data$reading_score,
  lambda = lambda_seq,
  nfolds = 5)
cv_object_reading$lambda.min

## [1] 0.5011872
```

```
model_reading_lasso = glmnet(as.matrix(data[1:11]), data$reading_score, lambda = cv_object_reading$lambda1,
coef(model_reading_lasso))
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)                63.0047330
## gender                      6.8714456
## ethnic_group                1.0191726
## parent_educ                 1.6822432
## lunch_type                  -7.2445118
## test_prep                   6.2890596
## parent_marital_status      -0.7735146
## practice_sport              .
## is_first_child              .
## nr_siblings                 .
## transport_means             .
## wkly_study_hours            0.4772919
```

```
model_reading_lasso$dev.ratio
```

```
## [1] 0.2302132
```

Model building for writing

```
# backward model
step(model_writing_full, direction = "backward", )
```

```
## Start:  AIC=1813.9
## writing_score ~ (gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + practice_sport + is_first_child +
##      nr_siblings + transport_means + wkly_study_hours + math_score +
##      reading_score) - reading_score - math_score
##
##              Df Sum of Sq  RSS    AIC
## - nr_siblings    7    1019.1 52079 1806.9
## - is_first_child    1         4.4 51064 1811.9
## - practice_sport    2     361.2 51421 1812.4
## - transport_means    1      74.2 51134 1812.4
## <none>                        51060 1813.9
## - ethnic_group      4     1779.1 52839 1818.0
## - parent_educ        3     1940.3 53000 1821.1
## - parent_marital_status  3     1991.7 53052 1821.4
## - wkly_study_hours    2     1901.4 52961 1822.8
## - lunch_type          1     6175.3 57235 1852.3
## - test_prep           1     6924.6 57985 1856.9
## - gender              1     8281.3 59341 1865.1
##
## Step:  AIC=1806.89
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
```

```

##      test_prep + parent_marital_status + practice_sport + is_first_child +
##      transport_means + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - is_first_child      1      1.2 52080 1804.9
## - transport_means      1     52.4 52132 1805.2
## - practice_sport       2    404.8 52484 1805.6
## <none>                  52079 1806.9
## - ethnic_group         4    1870.2 53949 1811.4
## - parent_marital_status 3    2027.5 54107 1814.4
## - parent_educ          3    2069.1 54148 1814.7
## - wkly_study_hours     2    1830.3 53910 1815.1
## - test_prep            1    6879.5 58959 1848.8
## - lunch_type           1    6955.3 59035 1849.3
## - gender               1    8444.0 60523 1858.1
##
## Step:  AIC=1804.9
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + practice_sport + transport_means +
##      wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - transport_means      1     53.0 52133 1803.3
## - practice_sport       2    408.3 52489 1803.7
## <none>                  52080 1804.9
## - ethnic_group         4    1869.4 53950 1809.4
## - parent_marital_status 3    2028.9 54109 1812.4
## - parent_educ          3    2068.7 54149 1812.7
## - wkly_study_hours     2    1829.2 53910 1813.1
## - test_prep            1    6907.4 58988 1847.0
## - lunch_type           1    6954.4 59035 1847.3
## - gender               1    8463.2 60544 1856.2
##
## Step:  AIC=1803.26
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + practice_sport + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## - practice_sport       2    397.6 52531 1802.0
## <none>                  52133 1803.3
## - ethnic_group         4    1901.9 54035 1808.0
## - parent_marital_status 3    1986.8 54120 1810.5
## - parent_educ          3    2041.4 54175 1810.9
## - wkly_study_hours     2    1821.0 53954 1811.4
## - lunch_type           1    6905.0 59038 1845.3
## - test_prep            1    7190.9 59324 1847.0
## - gender               1    8443.2 60577 1854.4
##
## Step:  AIC=1801.95
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + wkly_study_hours
##
##              Df Sum of Sq  RSS    AIC
## <none>                  52531 1802.0

```

```
## - ethnic_group      4      1950.7 54482 1806.9
## - parent_educ       3      1925.8 54457 1808.7
## - parent_marital_status 3      1962.6 54494 1808.9
## - wkly_study_hours  2      1804.0 54335 1809.9
## - lunch_type        1      6837.1 59368 1843.3
## - test_prep         1      7210.3 59741 1845.5
## - gender            1      8486.0 61017 1853.0

##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + wkly_study_hours,
##     data = data)
##
## Coefficients:
##             (Intercept)                gender1                ethnic_group1
##                   58.522                   10.032                   2.213
##             ethnic_group2                ethnic_group3                ethnic_group4
##                   1.850                   6.338                   6.617
##             parent_educ2                parent_educ3                parent_educ4
##                   1.789                   4.598                   7.212
##             lunch_type1                test_prep1 parent_marital_status1
##                   -9.263                   9.609                   -4.417
## parent_marital_status2 parent_marital_status3                wkly_study_hours1
##                   4.668                   -4.644                   5.168
##             wkly_study_hours2
##                   1.893
```

```
model_writing_back = lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
    lunch_type + test_prep + parent_marital_status + practice_sport +
    is_first_child + transport_means + wkly_study_hours, data = data)
summary(model_writing_back)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + transport_means + wkly_study_hours, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -35.016  -8.347   0.861   9.431  25.920
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    55.6226     3.7915  14.670 < 2e-16 ***
## gender1        10.0283     1.3627   7.359 1.45e-12 ***
## ethnic_group1    1.9857     3.0171   0.658  0.51089
## ethnic_group2    1.3766     2.8687   0.480  0.63164
## ethnic_group3    5.7836     2.9166   1.983  0.04819 *
## ethnic_group4    6.4017     3.0645   2.089  0.03747 *
## parent_educ2     1.8930     1.6347   1.158  0.24769
## parent_educ3     4.7742     1.9128   2.496  0.01305 *
```

```
## parent_educ4          7.5674      2.3506   3.219  0.00141 **
## lunch_type1          -9.3729      1.4034  -6.679  1.01e-10 ***
## test_prep1           9.5404      1.4363   6.642  1.25e-10 ***
## parent_marital_status1 -4.5162      1.6470  -2.742   0.00643 **
## parent_marital_status2  5.4329      4.3399   1.252   0.21150
## parent_marital_status3 -4.4594      1.9914  -2.239   0.02579 *
## practice_sport1       3.4669      2.1647   1.602   0.11020
## practice_sport2       3.0695      2.2715   1.351   0.17751
## is_first_child1      -0.1246      1.4489  -0.086   0.93152
## transport_means1      0.8261      1.4256   0.580   0.56263
## wkly_study_hours1     5.2430      1.5846   3.309   0.00104 **
## wkly_study_hours2     2.0645      2.0654   1.000   0.31826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.49 on 334 degrees of freedom
## Multiple R-squared:  0.3638, Adjusted R-squared:  0.3276
## F-statistic: 10.05 on 19 and 334 DF,  p-value: < 2.2e-16
```

```
# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

cv_object_writing = cv.glmnet(as.matrix(data[1:11]), data$writing_score,
                             lambda = lambda_seq,
                             nfolds = 5)
cv_object_writing$lambda.min
```

```
## [1] 0.5011872
```

```
model_writing_lasso = glmnet(as.matrix(data[1:11]), data$writing_score, lambda = cv_object_writing$lambda.min,
                             coef(model_writing_lasso))
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  59.3844759
## gender       8.7384396
## ethnic_group 1.4955961
## parent_educ  1.8826016
## lunch_type   -8.1037819
## test_prep    8.0886240
## parent_marital_status -0.8123378
## practice_sport .
## is_first_child .
## nr_siblings   0.1133873
## transport_means .
## wkly_study_hours 0.7539334
```

```
model_writing_lasso$dev.ratio
```

```
## [1] 0.3119987
```