Exploratory Data Analysis in R

Author:Pasquale Salomone

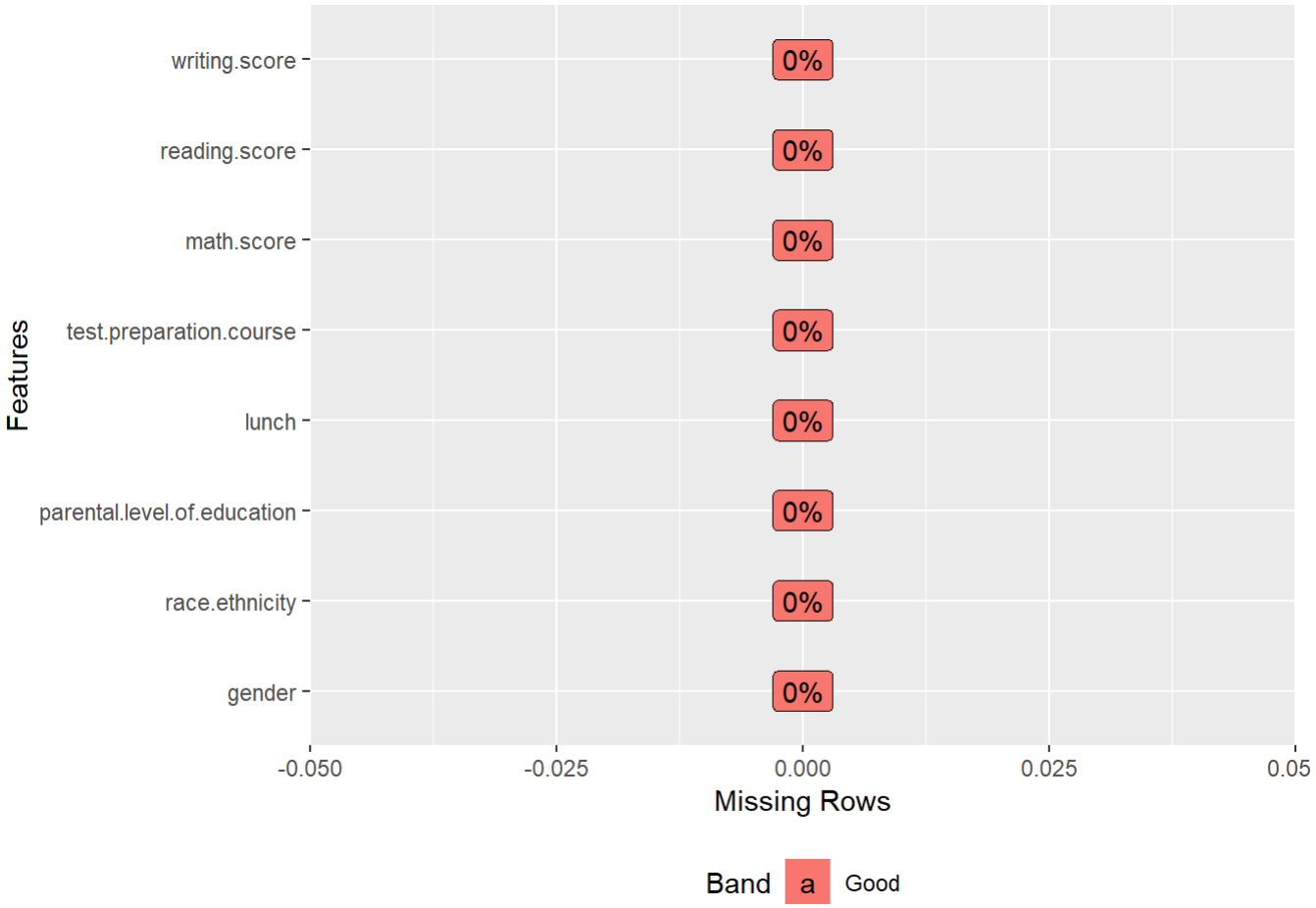
December 2018

Exploring the data set

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...
##  $ race.ethnicity             : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...
##  $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...
##  $ lunch                      : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...
##  $ test.preparation.course    : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...
##  $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
##  $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
##  $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

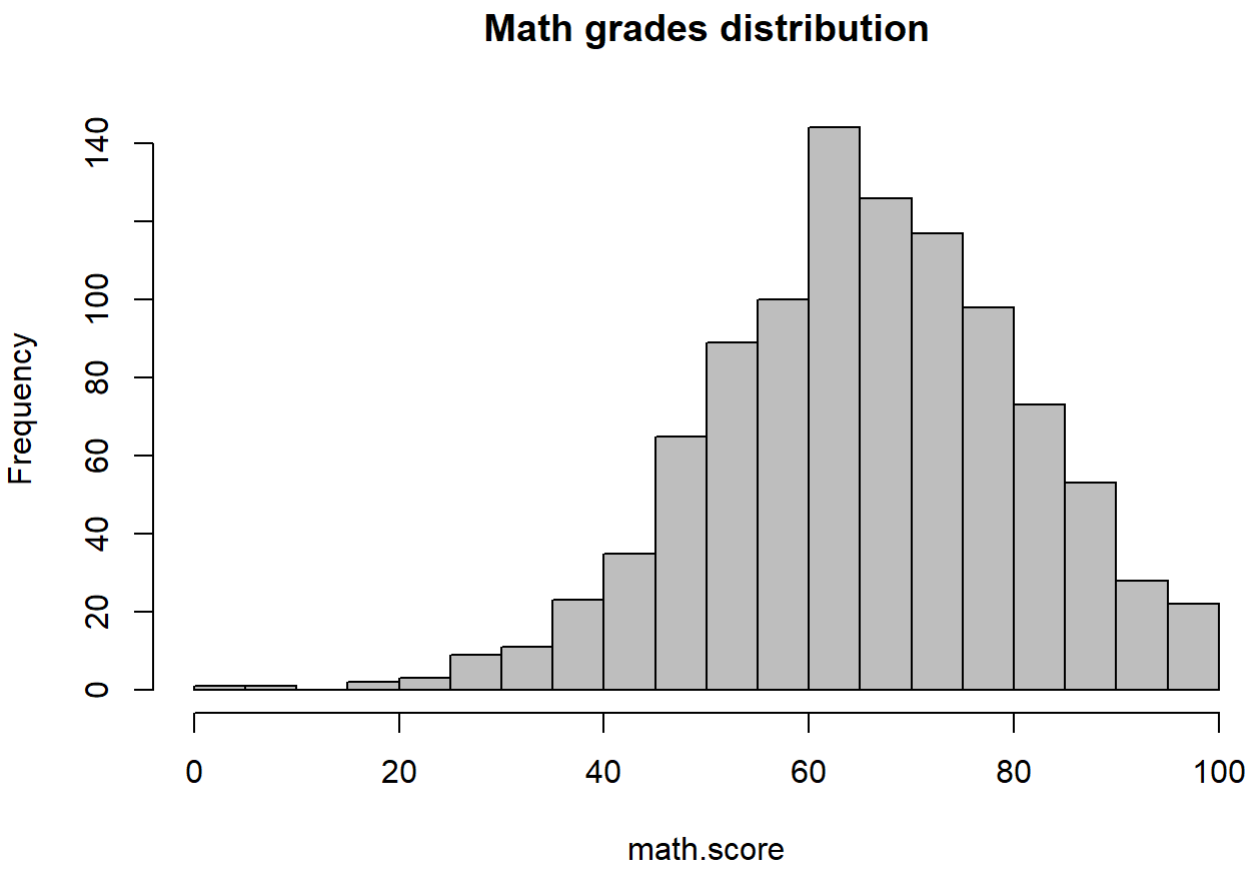
```
##     gender    race.ethnicity     parental.level.of.education
##  female:518   group A: 89    associate's degree:222
##  male  :482   group B:190    bachelor's degree :118
##               group C:319    high school       :196
##               group D:262    master's degree   : 59
##               group E:140    some college      :226
##                              some high school  :179
##           lunch     test.preparation.course   math.score
##  free/reduced:355   completed:358           Min.   :  0.00
##  standard    :645   none     :642           1st Qu.: 57.00
##                                             Median : 66.00
##                                             Mean   : 66.09
##                                             3rd Qu.: 77.00
##                                             Max.   :100.00
##  reading.score    writing.score
##  Min.   : 17.00   Min.   : 10.00
##  1st Qu.: 59.00   1st Qu.: 57.75
##  Median : 70.00   Median : 69.00
##  Mean   : 69.17   Mean   : 68.05
##  3rd Qu.: 79.00   3rd Qu.: 79.00
##  Max.   :100.00   Max.   :100.00
```

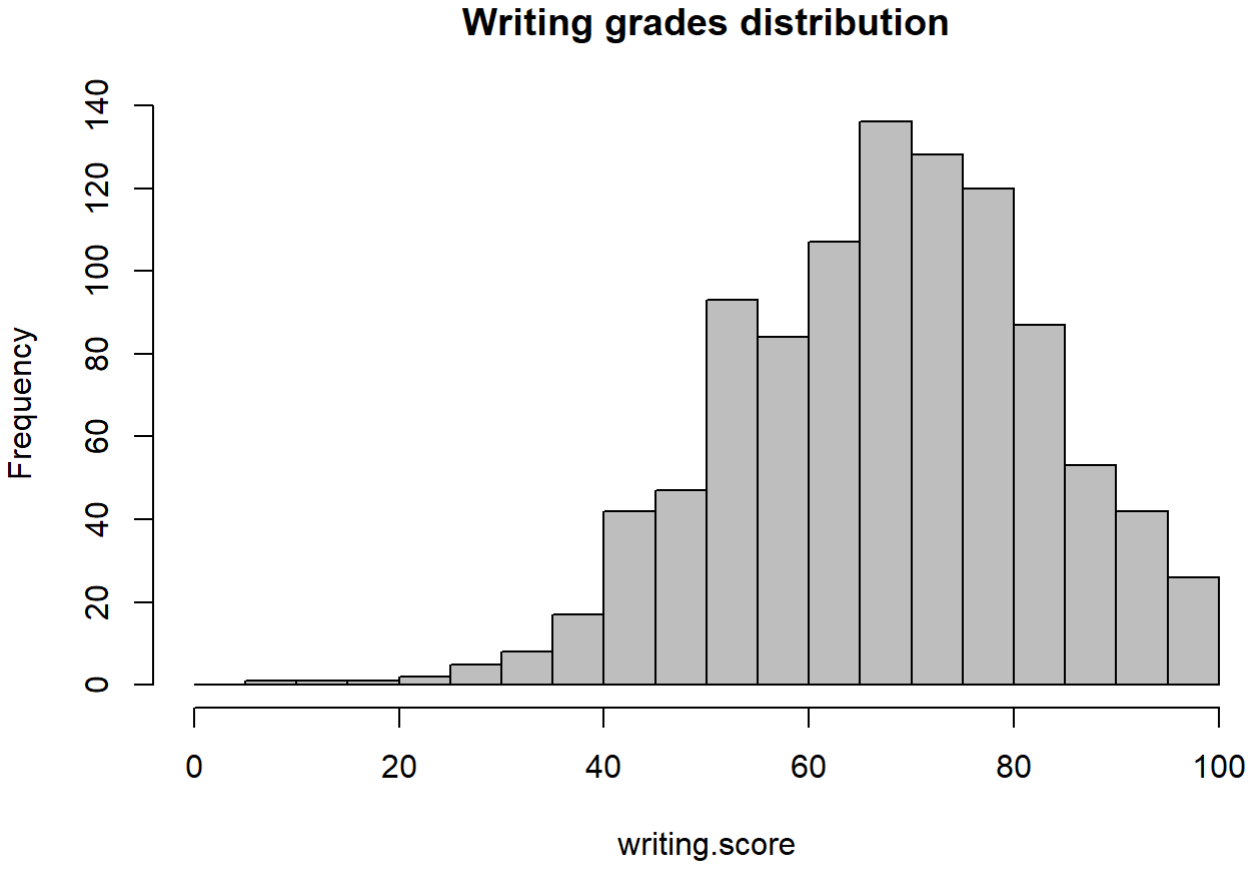
The summary above doesn't show the presence of missing/incorrect values.

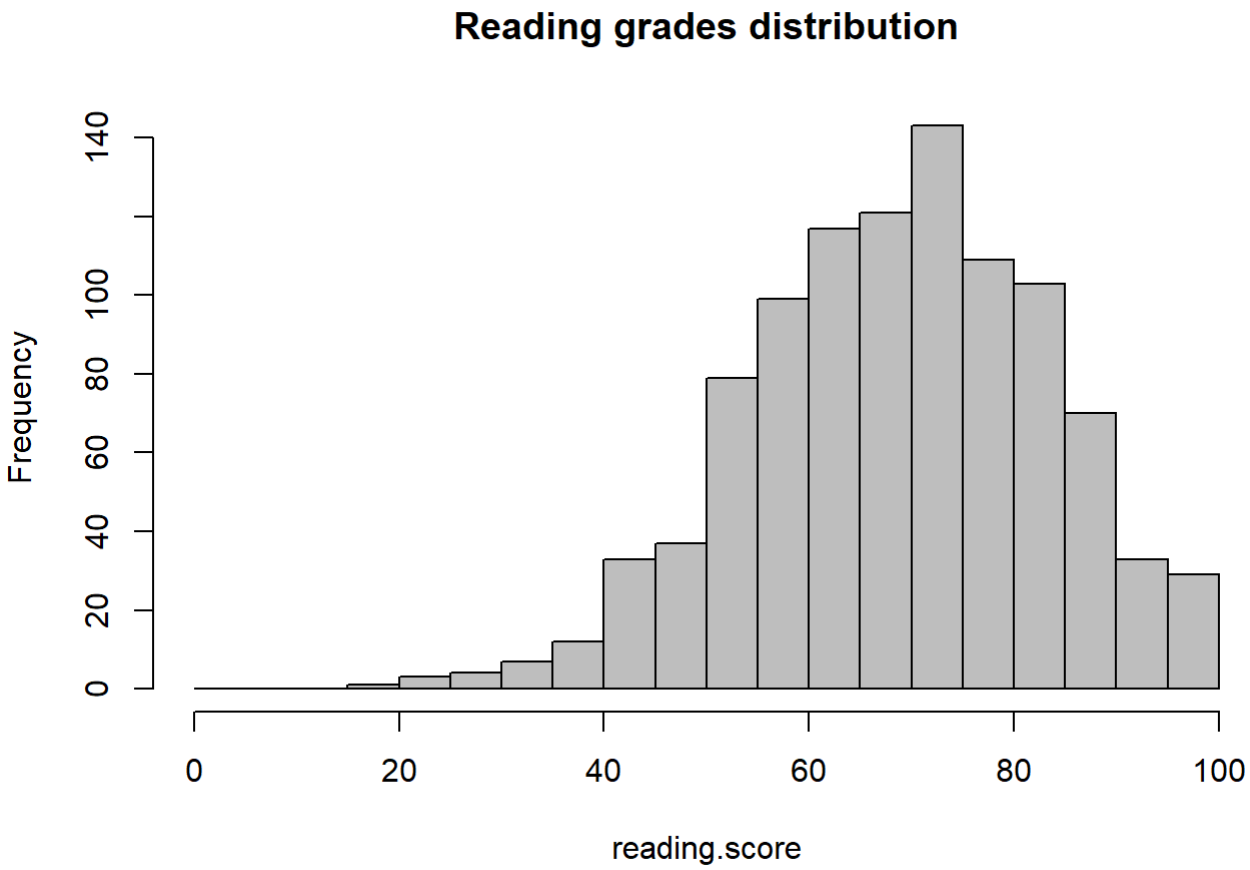


A visual analysis confirms the absence of missing values.

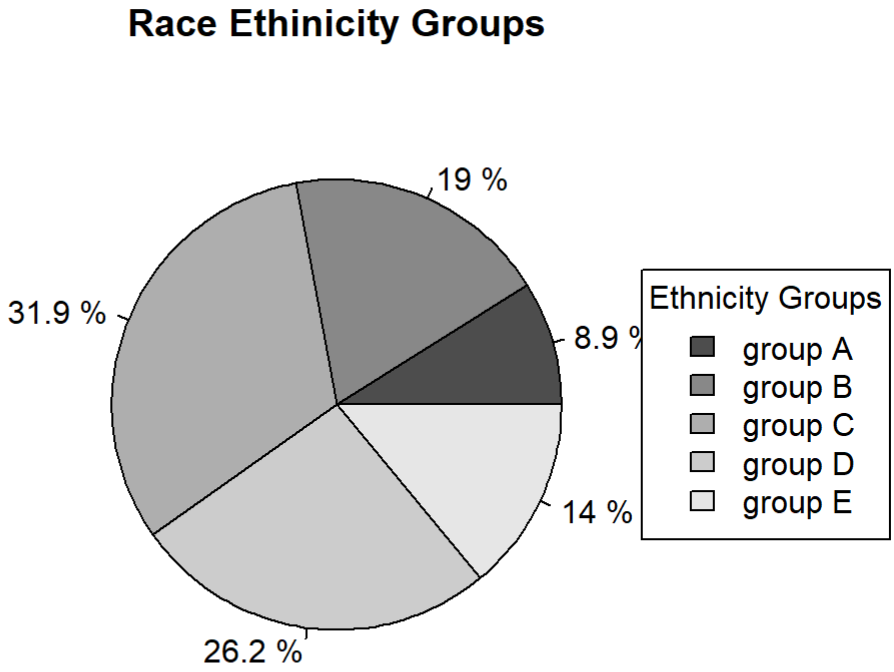
Univariate Plot Section



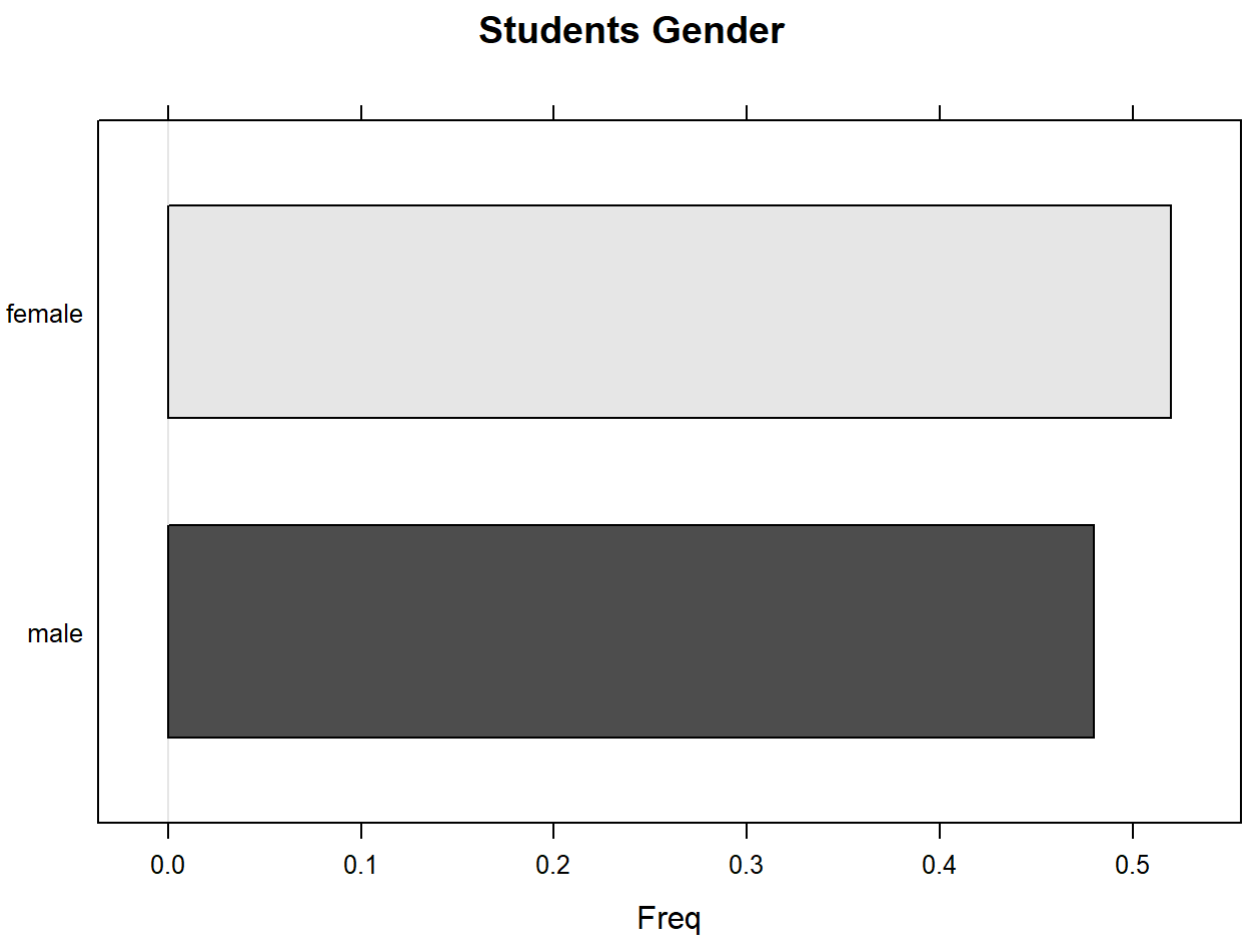




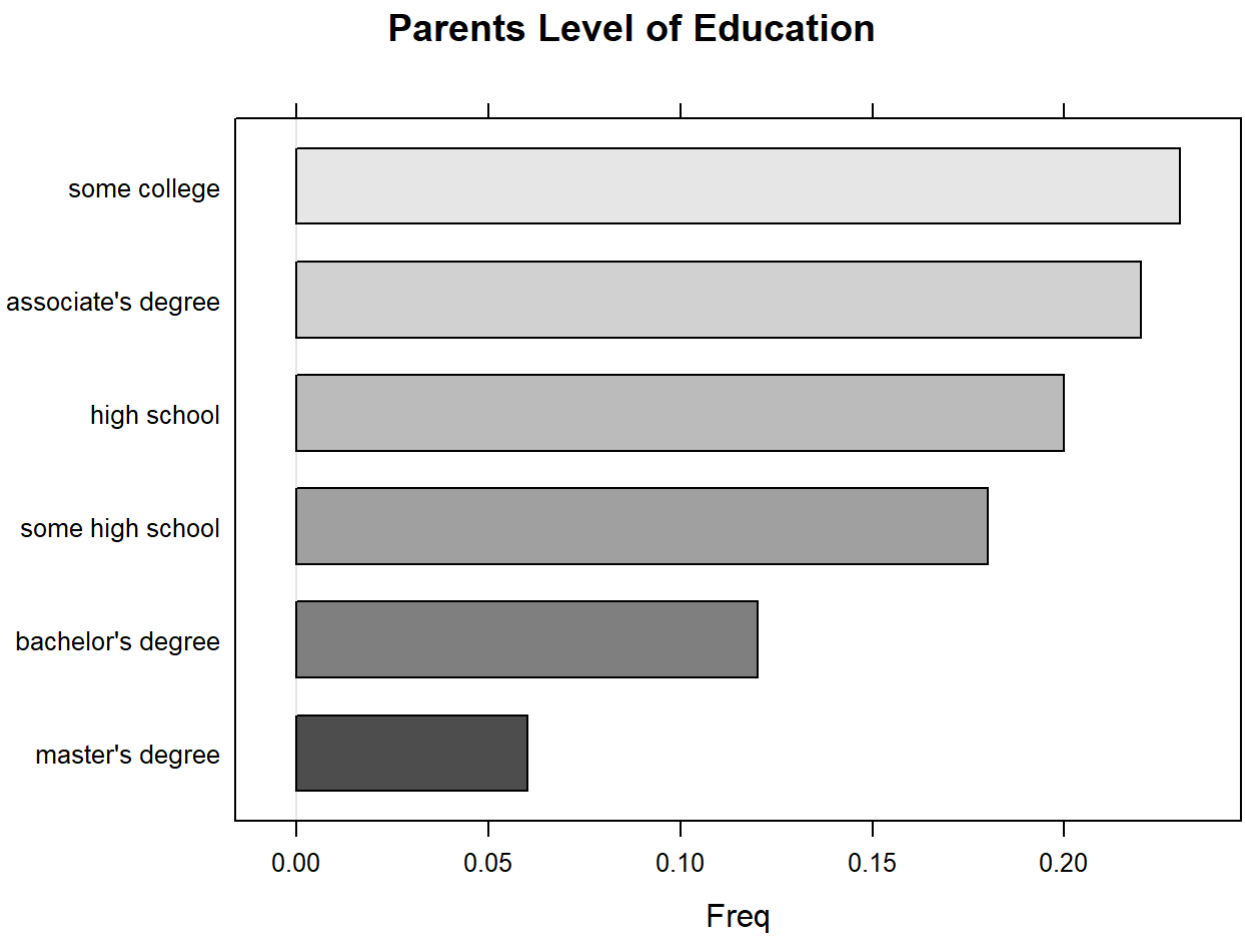
The histograms above display the distribution of math, writing, and reading scores. Math, writing, and reading scores all appear to follow an almost normal distribution; there is a slight left skew possibly due to the presence of outliers.



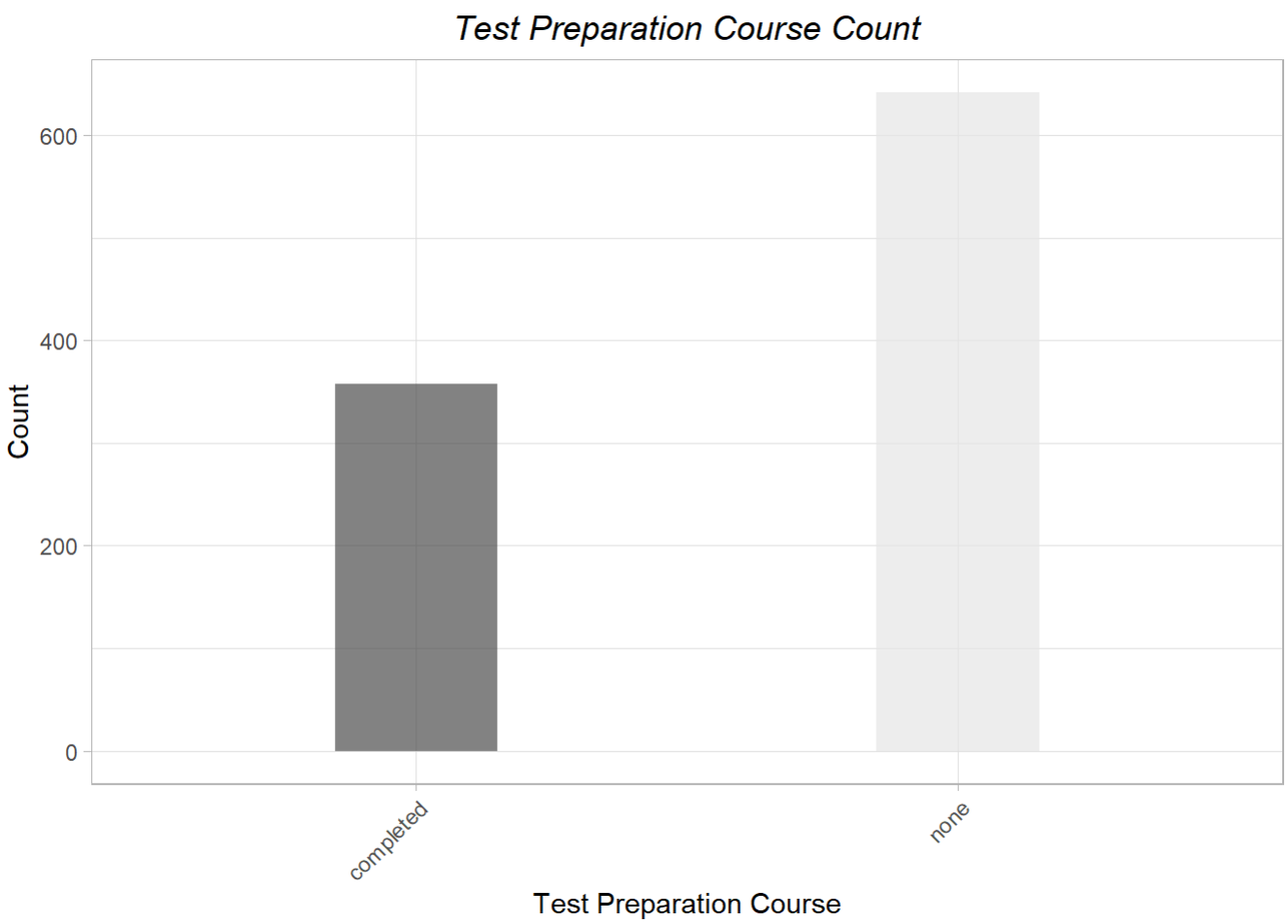
The pie chart displays the percentage of ethnicity groups. The data set does not provide an explanation for the ethnicity groups coding.



The plot shows the percentage of female students in the data set is slightly greater than male students.



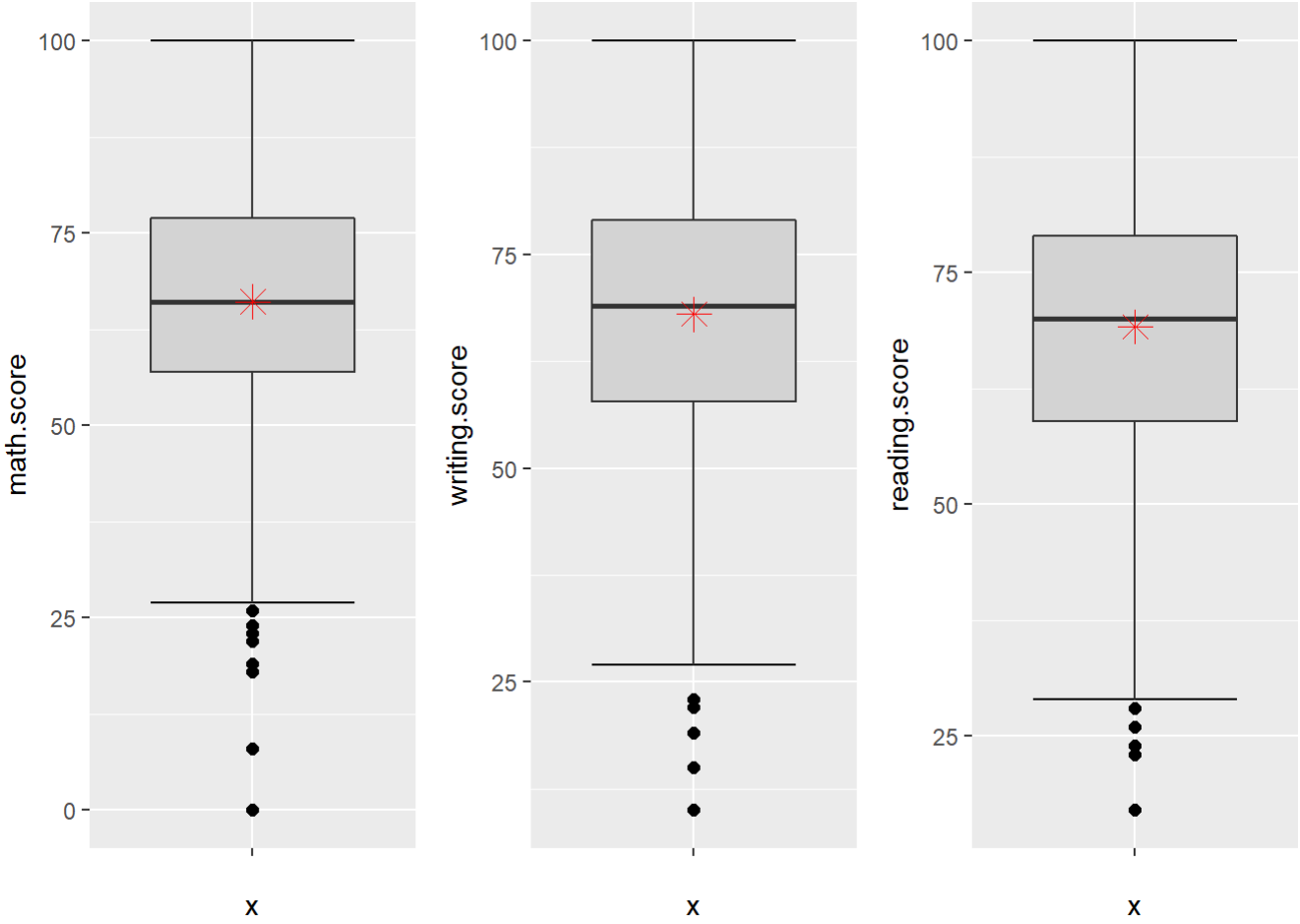
The plot shows that master level education is the least common among parents.



The plot shows that the count of students who did not complete a test preparation course is greater than students who did.

The plot shows most students did not qualify for a free/reduced lunch type.

Outliers analysis



*All grading scores present outliers. Outliers can drastically bias/change the fit estimates and predictions.For a given continuous variable, outliers are those observations that lie outside 1.5*IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. We can see outliers as the points outside the whiskers in the box plot. Also, mean values are displayed with a red asterisk.*

Univariate analysis

The data set has 1000 observation with 8 features. There are no missing values in the data set.

The ordered factor variables have the following orders:

```
## $gender
## [1] "female" "male"
##
## $race.ethnicity
## [1] "group A" "group B" "group C" "group D" "group E"
##
## $parental.level.of.education
## [1] "associate's degree" "bachelor's degree"  "high school"
## [4] "master's degree"    "some college"       "some high school"
##
## $lunch
## [1] "free/reduced" "standard"
##
## $test.preparation.course
## [1] "completed" "none"
```

Other observations:

- In all three variables the mean and median are the same or very close which confirms that they are normally distributed
- About 52% of the students in the data-sets are female and 48% are males
- There are 5 ethnicity group in the data set
- About 36% of students completed a test preparation course while 64% did not complete one.

- The highest level of parental education is “some college”

What is/are the main feature(s) of interest in your dataset?

The main features in data set are the grading scores. The analysis tries to determine if there is a correlation among grading scores.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

To investigate if gender and test.preparation.course features weaken or strengthen reading.score vs writing.score correlation.

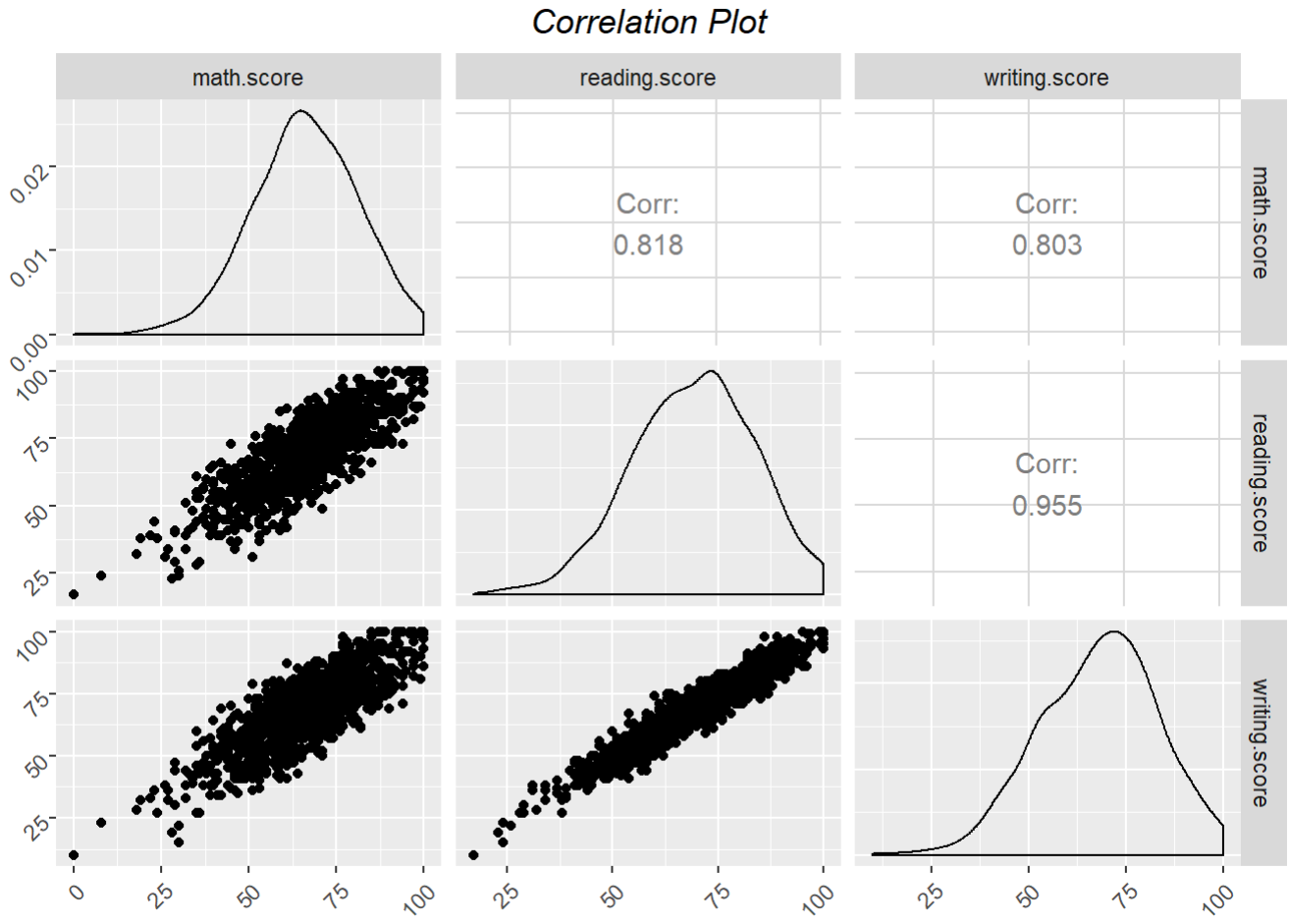
Did you create any new variables from existing variables in the dataset?

Categorical variables were converted to dummy variables using the dummies library.

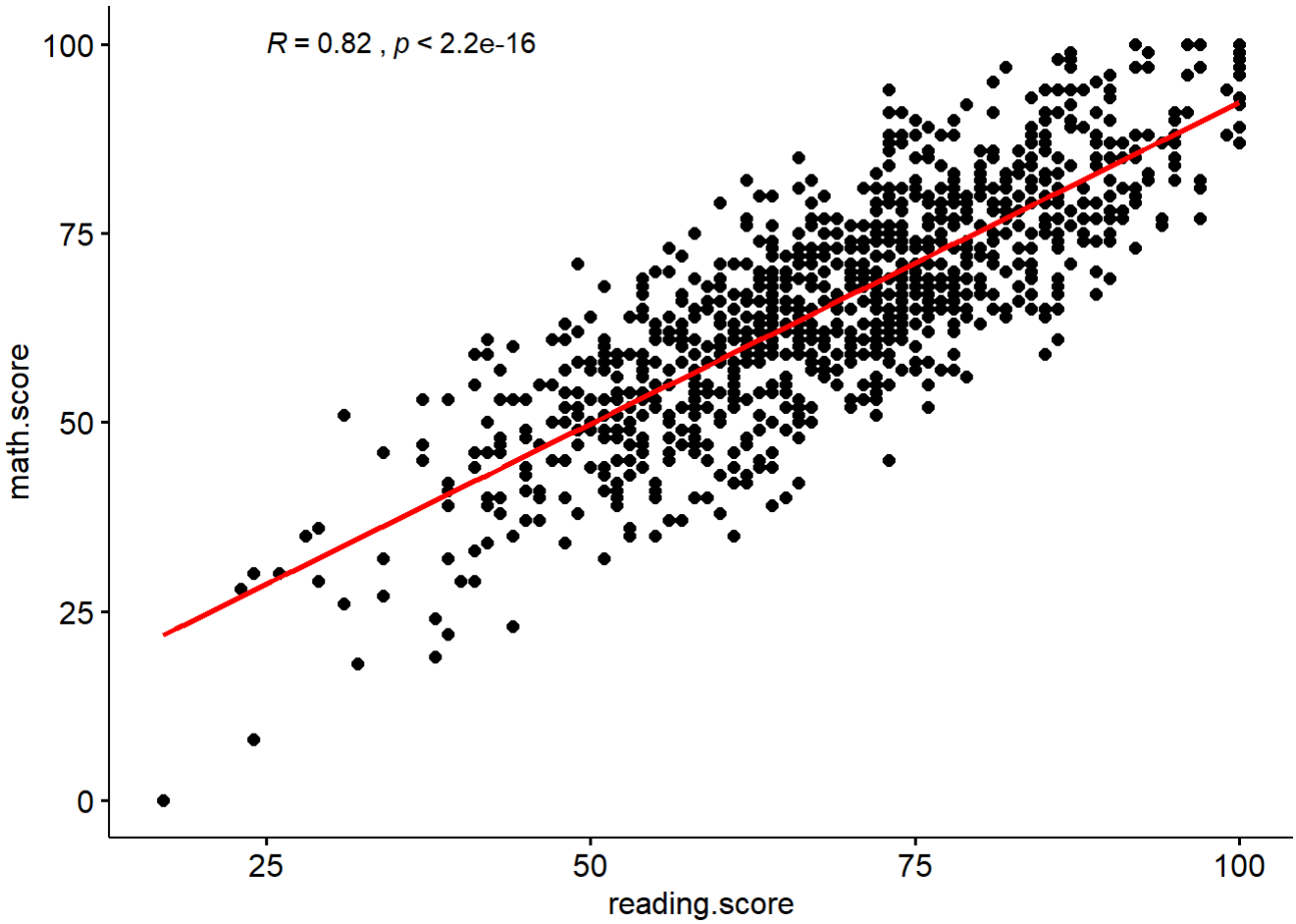
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

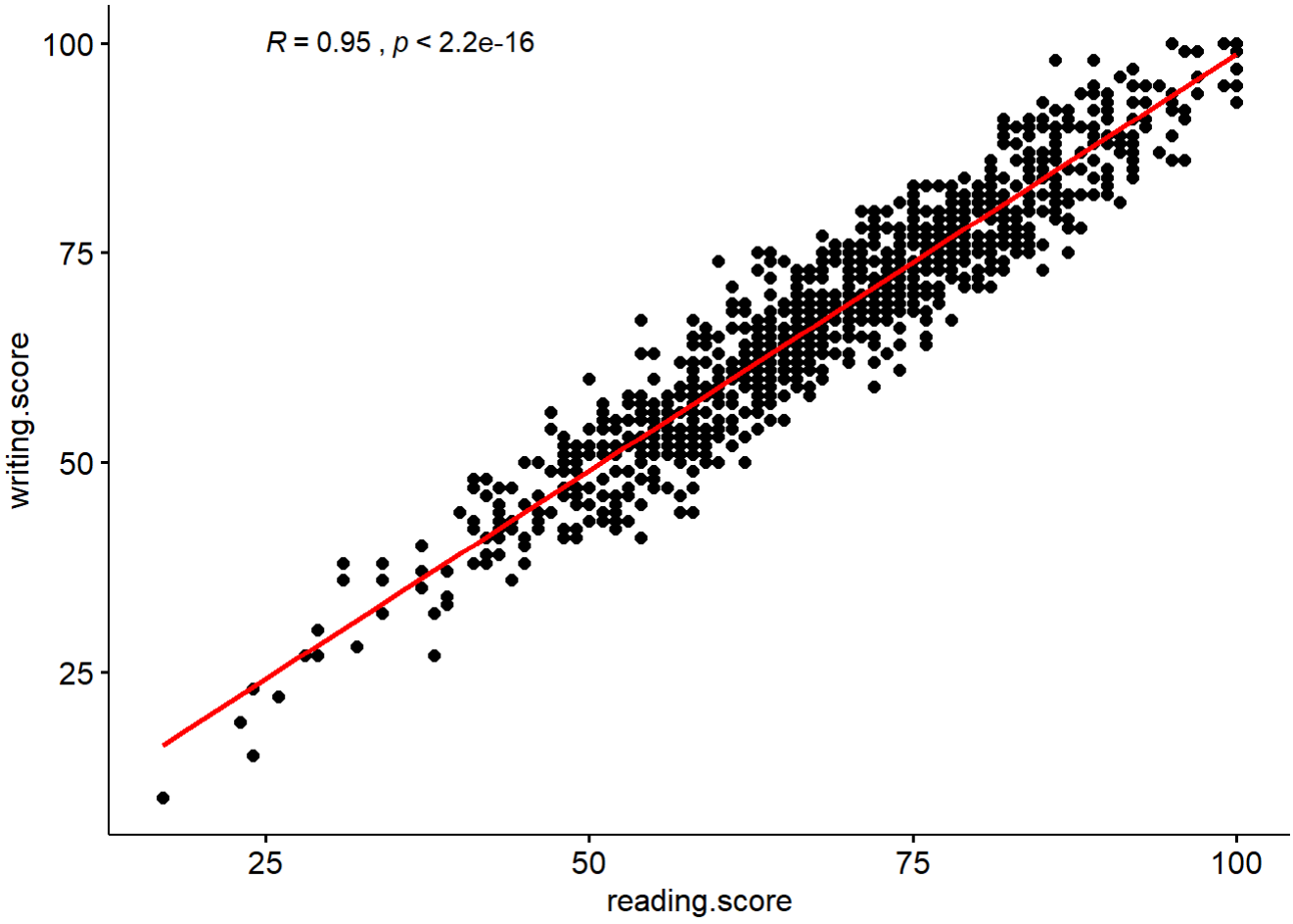
The data set was downloaded from https://www.kaggle.com/spscientist/students-performance-in-exams (https://www.kaggle.com/spscientist/students-performance-in-exams). No operations were performed to tidy the data set.

Bivariate Plots Section



The plot above shows the correlation among the variables of interest.





The scatter plots above show the correlation between math.score and reading.score and writing.score and reading.score.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Correlation is a statistical measure that suggests the level of linear dependence between two variables, that occur in pair. Correlation can take values between -1 to +1.The grading score variables are all positively and strongly correlated as shown in the correlation table.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Gender and test.preparation.course features were analyzed for possible relationship but nothing significant was observed.

Female Test Preparation Rates :

```
##
## completed      none
## 0.3552124 0.6447876
```

Male Test Preparation Rates :

```
##
## completed      none
## 0.3609959 0.6390041
```

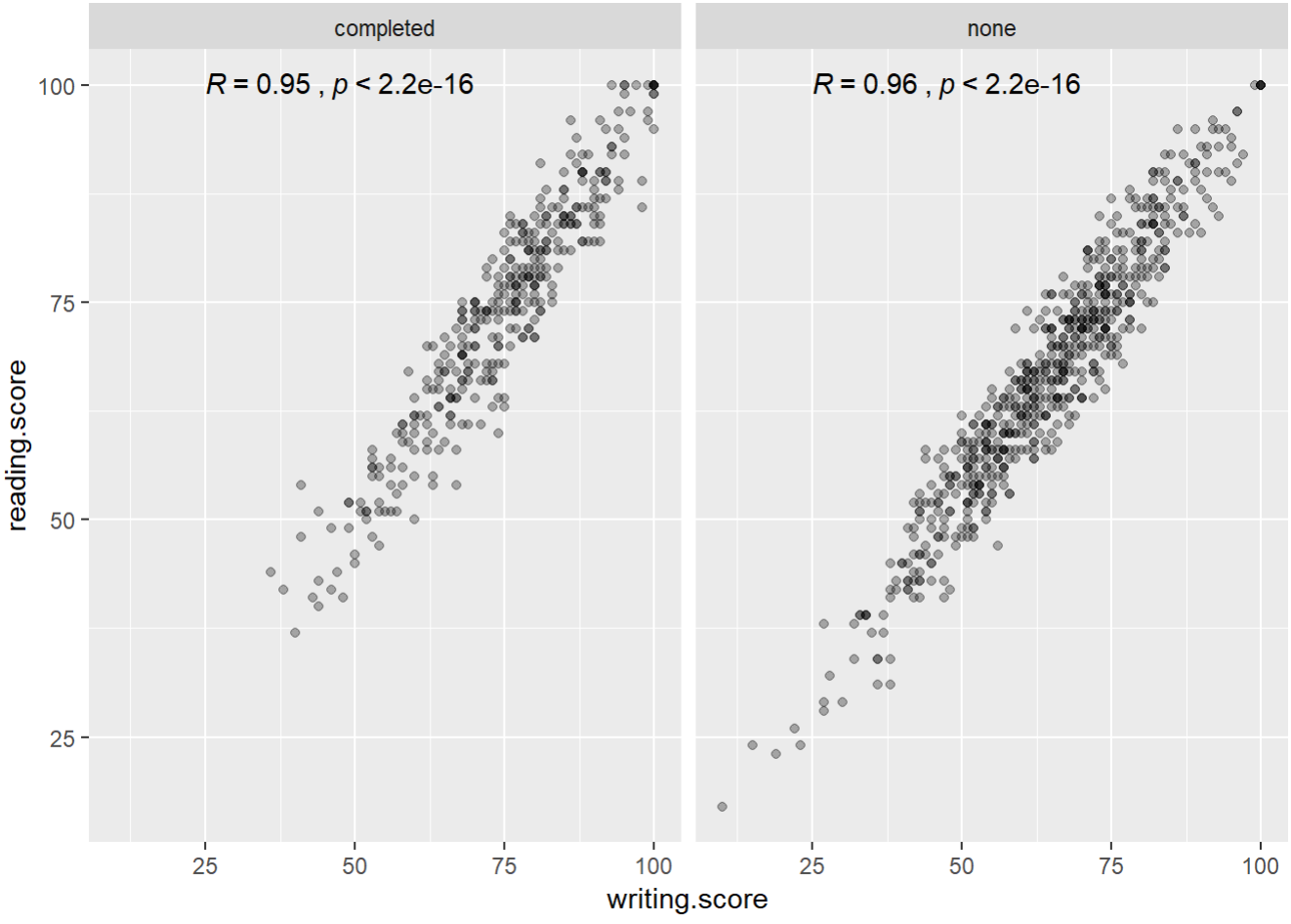
What was the strongest relationship you found?

The reading score is strongly correlated with the writing.score with a correlation coefficient of 0.95

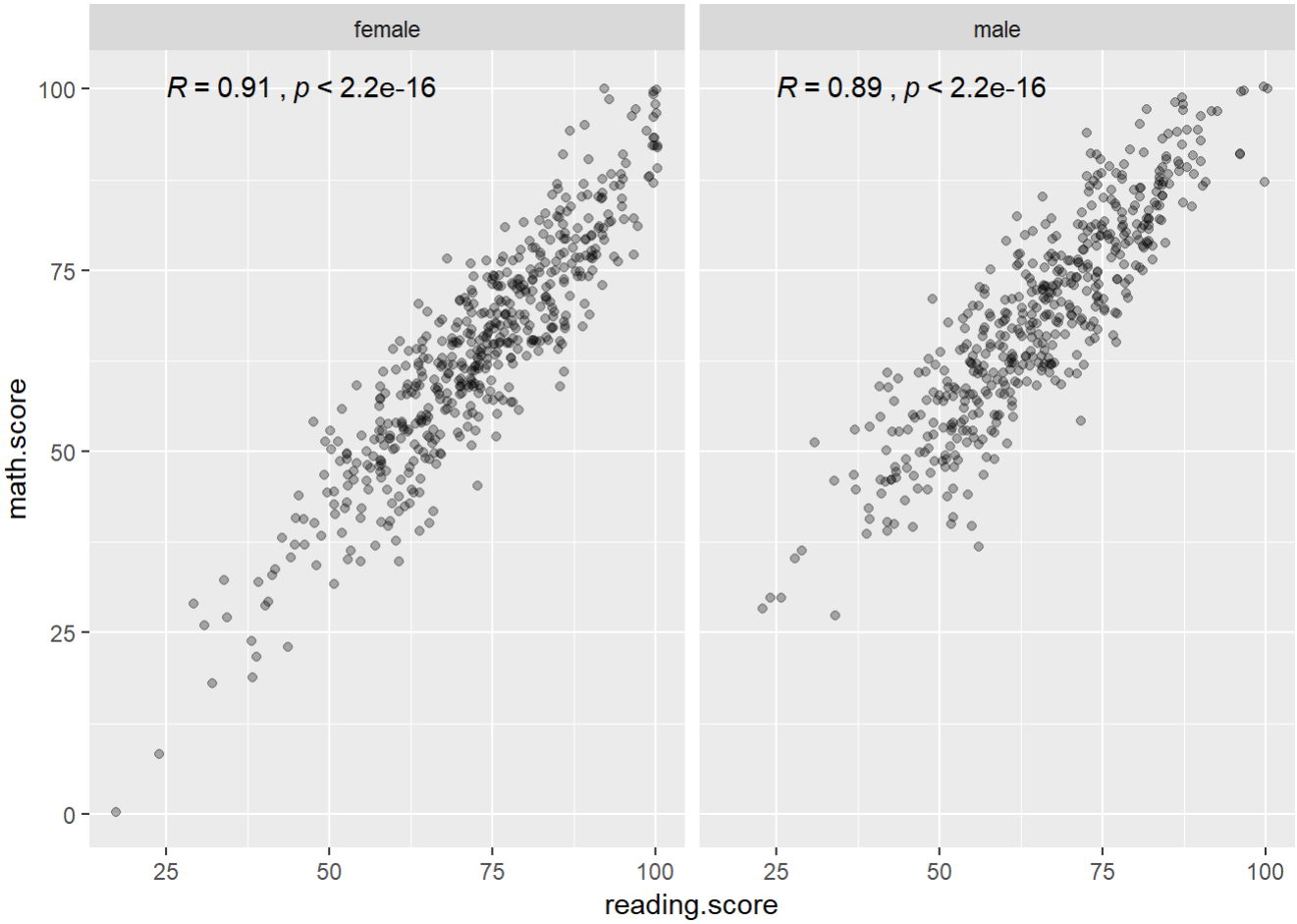
Correlation Table:

```
##               math.score writing.score reading.score
## math.score          1.00          0.80          0.82
## writing.score       0.80          1.00          0.95
## reading.score       0.82          0.95          1.00
```

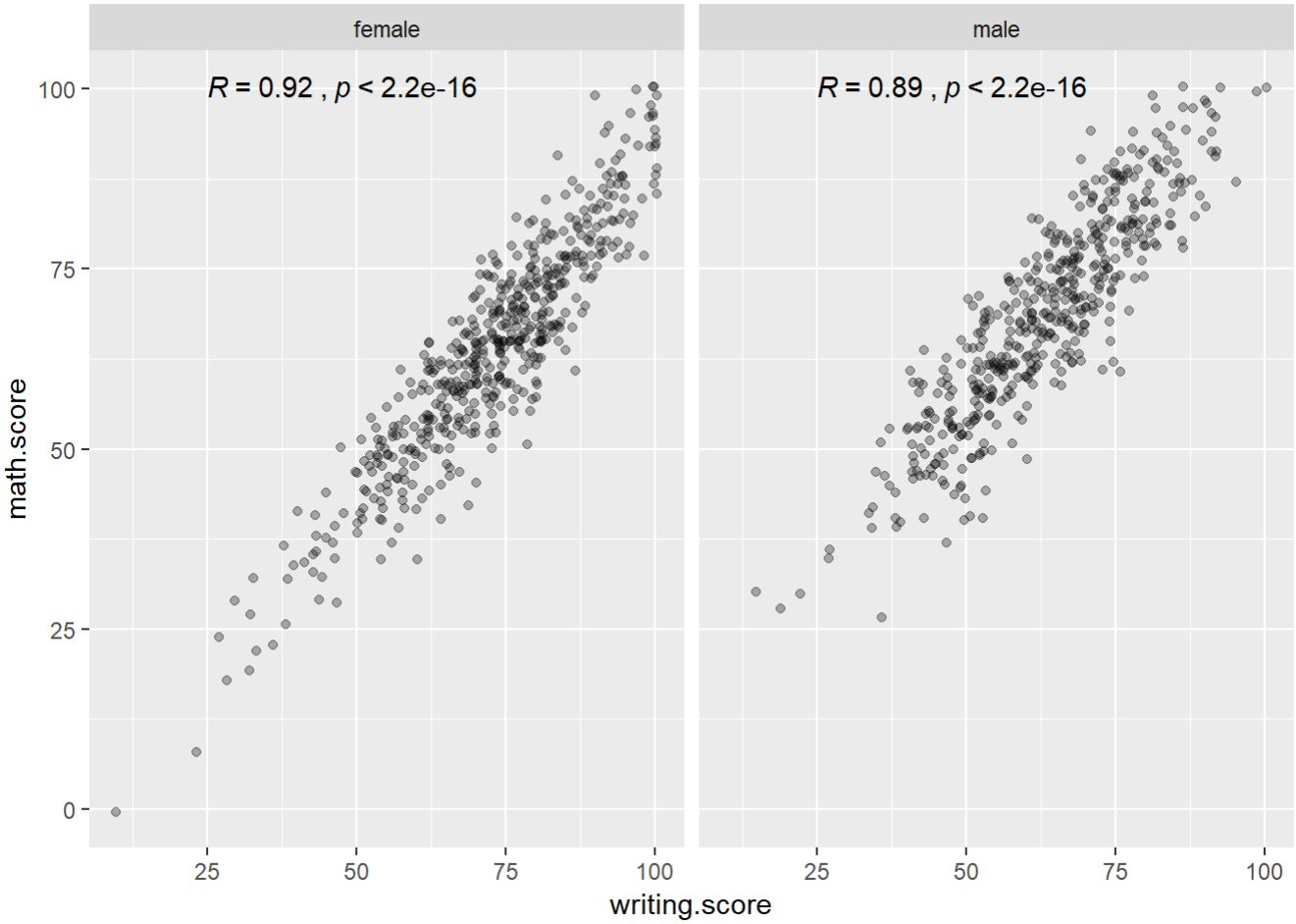
Multivariate Plots Section



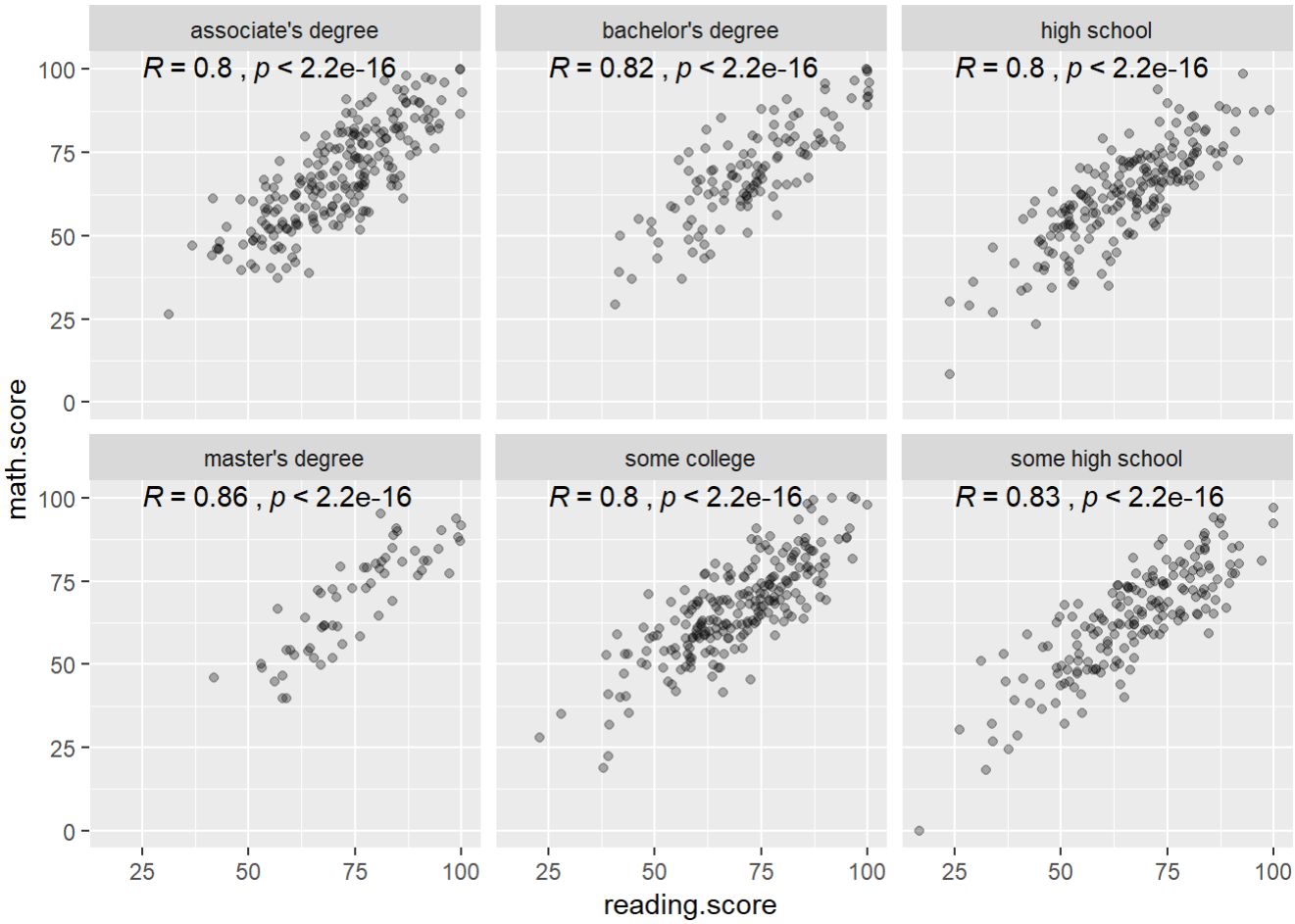
Scatter plots to investigate the reading score by writing.score and test.preparation.course to see if a third variable has any impact on the features of interest correlation.



Scatter plots to investigate the math.score by reading.score and gender to see if a third variable has any impact on the features of interest correlation.



Scatter plots to investigate the math.score by writing.score and gender to see if a third variable has any impact on the features of interest correlation.



Scatter plots to investigate the math.score by reading.score and parental.level.of.education to see if a third variable has any impact on the features of interest correlation.

Multivariate Analysis

Creating Dummy Variables and avoiding Trap in Regression Models

```
library(dummies)
new_data <- dummy.data.frame(students_data)
new_data <- select(new_data,-c(1,3,8,14,16)) # removing m-1 in the dataframe (m number of categories)
```

The Dummy Variable trap is a scenario in which the independent variables are multicollinear - a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others. The solution to the dummy variable trap is to drop one of the categorical variables (or alternatively, drop the intercept constant) - if there are m number of categories, use m-1 in the model, the value left out can be thought of as the reference value and the fit values of the remaining categories represent the change from this reference.

Creating a training and testing data set.

```
#loading package
library(caTools)

#use caTools function to split, SplitRatio for 70%:30% splitting
data1= sample.split(new_data,SplitRatio = 0.3)

#subsetting into Train data
train =subset(new_data, data1==TRUE)

#subsetting into Test data
test =subset(new_data,data1==FALSE)
```

Creating a first linear model which includes higly correlated predictors.

```
##      RMSE        R2
## 1 5.60072 0.8663273
```

The table above assess model_1 performance metrics.

Using R function vif() to detect multicollinearity in a regression model_1.

```
##                                     gendermale
##                                       1.240198
##                          `race.ethnicitygroup B`
##                                       2.538108
##                          `race.ethnicitygroup C`
##                                       3.442796
##                          `race.ethnicitygroup D`
##                                       2.996622
##                          `race.ethnicitygroup E`
##                                       2.390213
## `parental.level.of.educationbachelor's degree`
##                                       1.377580
##       `parental.level.of.educationhigh school`
##                                       1.565188
##   `parental.level.of.educationmaster's degree`
##                                       1.288317
##      `parental.level.of.educationsome college`
##                                       1.709445
##  `parental.level.of.educationsome high school`
##                                       1.658429
##                                  lunchstandard
##                                       1.209631
##                     test.preparation.coursenone
##                                       1.305132
##                                  reading.score
##                                      13.134485
##                                  writing.score
##                                      15.290106
```

The VIF score for the predictor variable reading.score and writing.score are very high (VIF = 12.72, 14.69) which may be problematic.

In multiple regression two or more predictor variables might be correlated with each other. This situation is referred as collinearity. There is an extreme situation, called multicollinearity, where collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. This means that there is redundancy between predictor variables. In the presence of multicollinearity, the solution of the regression model becomes unstable. For a given predictor (p), multicollinearity can assessed by computing a score called the variance inflation factor (or VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. When faced to multicollinearity, the concerned variables should be removed, since the presence of multicollinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables (James et al. 2014,P. Bruce and Bruce (2017)).

Creating a second linear model which has removed reading.score.

```
##       RMSE        R2
## 1 6.126628 0.8378872
```

The table above assess model_2 performance metrics.

Using R function vif() to detect multicollinearity in a regression model_2.

```
##                                     gendermale
##                                       1.115015
##                           `race.ethnicitygroup B`
##                                       2.532324
##                           `race.ethnicitygroup C`
##                                       3.428536
##                           `race.ethnicitygroup D`
##                                       2.979063
##                           `race.ethnicitygroup E`
##                                       2.387425
## `parental.level.of.educationbachelor's degree`
##                                       1.376443
##       `parental.level.of.educationhigh school`
##                                       1.545330
##   `parental.level.of.educationmaster's degree`
##                                       1.287457
##      `parental.level.of.educationsome college`
##                                       1.709286
##  `parental.level.of.educationsome high school`
##                                       1.629302
##                                  lunchstandard
##                                       1.158598
##                     test.preparation.coursenone
##                                       1.123915
##                                  reading.score
##                                       1.335288
```

The analysis of VIF scores doesn't present any highly correlated predictor variable.

Using both linear models to predict math.score based on all predictors.

Model	Model Predicted Math Score
Model_1	89.39751
Model_2	92.36569

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Stratifying the correlation by a third variable such as gender resulted in a slightly better correlation coefficients for the female gender; stratification by test preparation course resulted in a slightly better correlation coefficient for students who did not take a test preparation course. This is only a descriptive report which may be the result of data set real and random effects. Further statistical analysis would be needed to make any inference.

Were there any interesting or surprising interactions between features?

Stratifying the correlation by a third variable such as parental.level.of.education resulted in a higher correlation coefficient for the variables of interest(reading.score, math.score) for master's degree parental level of education .

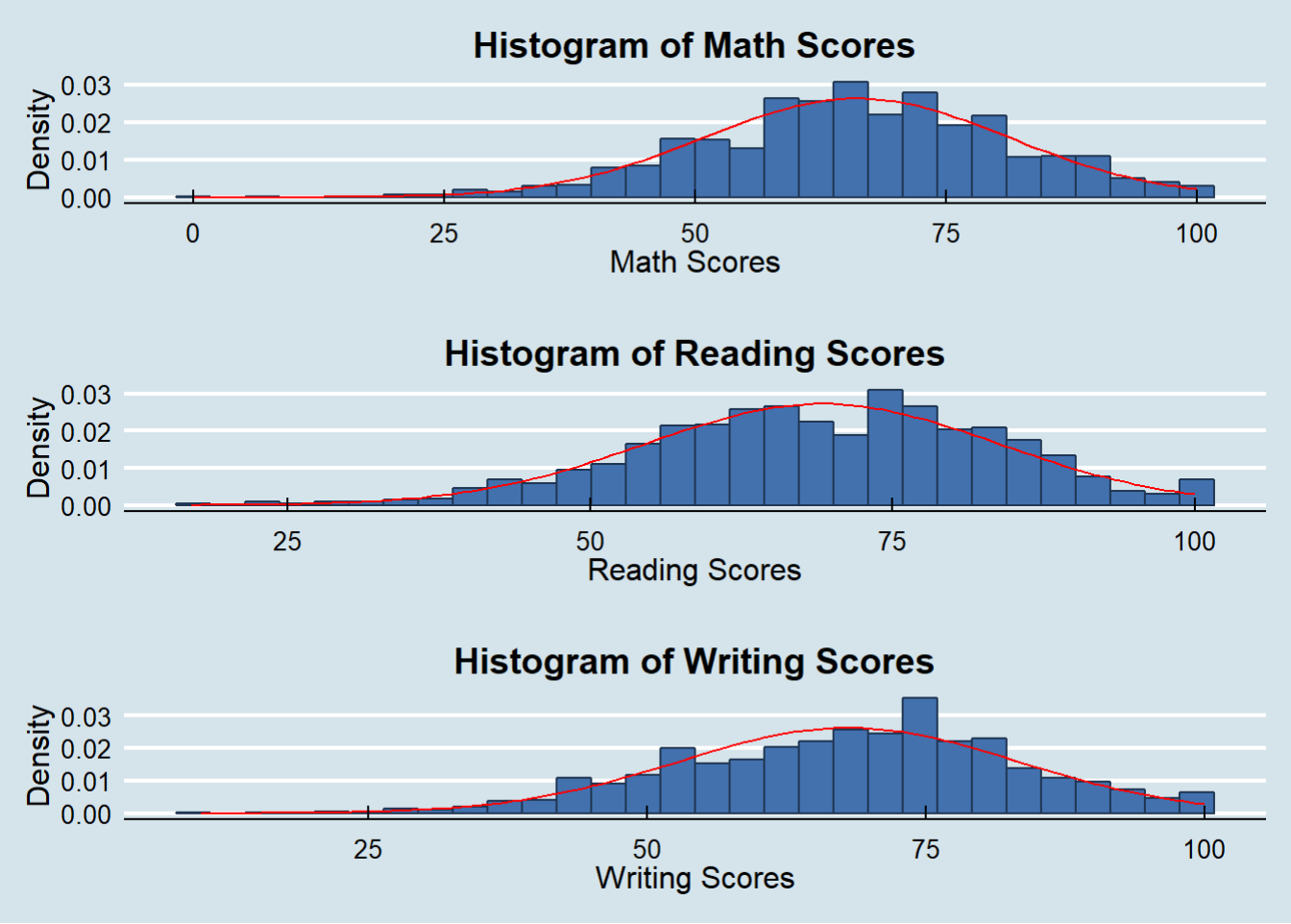
```
## # A tibble: 6 x 2
##   parental.level.of.education   cor
##   <fct>                       <dbl>
## 1 associate's degree           0.8
## 2 bachelor's degree            0.82
## 3 high school                  0.8
## 4 master's degree              0.86
## 5 some college                 0.8
## 6 some high school             0.83
```

Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Two linear models were created to predict a math scores; model_1 had slightly better performance metrics due to multicollinearity. Model_2 removed the writing.score variable to address co-variance and model performance metrics were not much affected.

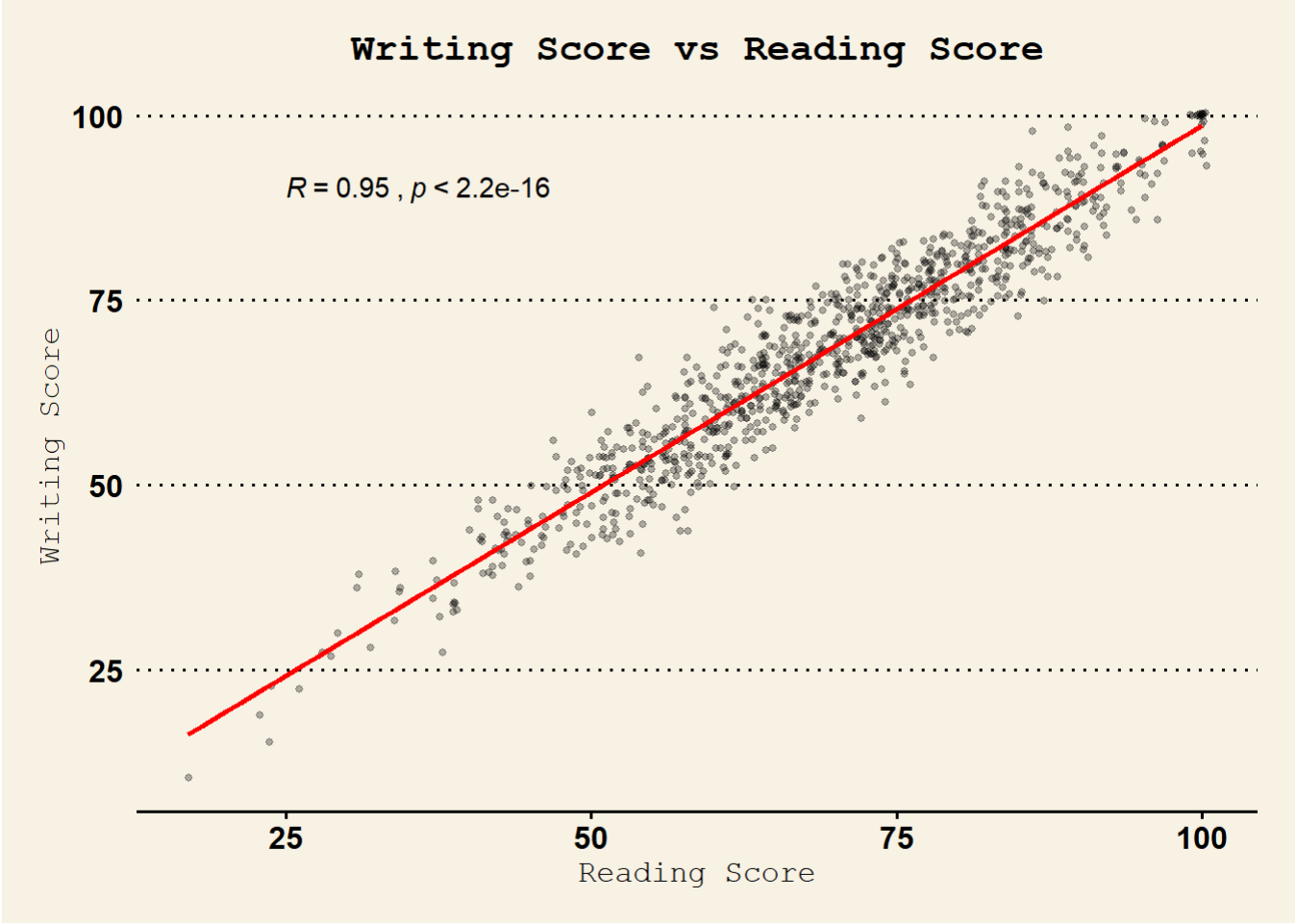
Final Plots and Summary

Plot one



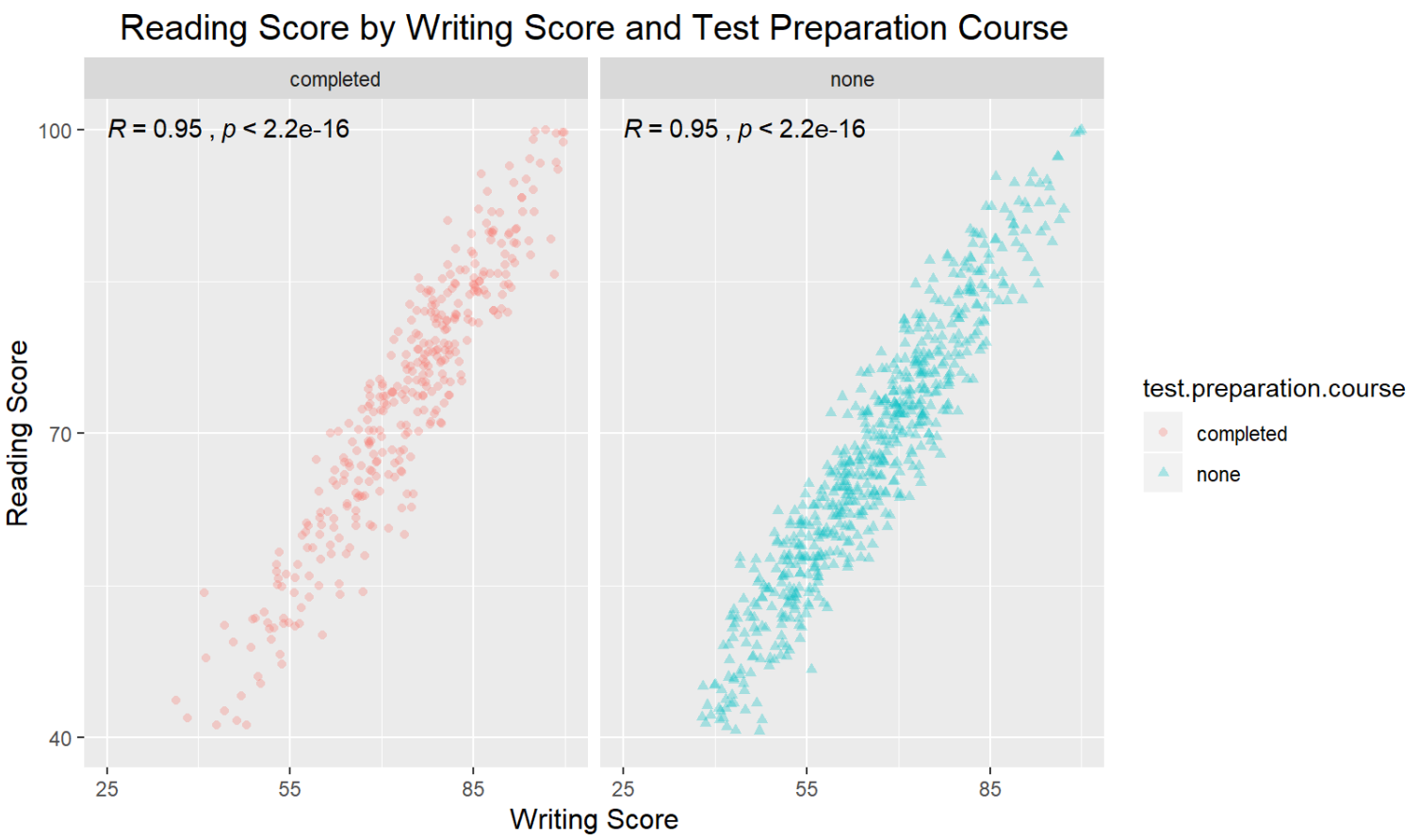
We overlay a normal density function curve on top of our histogram to see how closely (or not) it fits a normal distribution.

Plot Two



The plot indicates that of the variable of interests writing.score and reading.score display the strongest correlation.Used jitter to lower the impact of over plotting and a lower alpha coefficient.

Plot Three



Plots created with scale_x_continuous and scale_y_continuous to zoom in on the distributions.The plots suggest that gender and test preparation do not significantly strengthen/weaken the reading.score vs writing.score correlation.

Summary

The students performances data set contains information on 1000 data points across 8 variables. The analysis started with understanding the individual variables in the data set, and then proceeded to a bi-variate and multivariate analysis. Eventually, two linear models with math.score as the dependent variable were created; the first model included all predictors. The second model did not include the reading.score predictor.Multicollinearity was assessed by computing variance inflation factor (VIF) score. Both models included predictors such as gender and race ethnicity which may not be used in certain countries due to legal constrains. Also, in a large data set presenting multiple correlated predictor variables, you can perform principal component regression and partial least square regression strategies.

There was a strong correlation between reading.score and writing.score.A third level variable was added to the analysis;however but both gender and test preparation course did not have a significant impact on reading.score vs writing.score.

Some limitations of this model include the source of the data which was downloaded from kaggle.com. It is not known how the data was collected. Furthermore, we don't know how representative is the sample. The author of the data set doesn't provide any coding for the ethnicity groups.

References

- http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/ (http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/)
- https://www.algosome.com/articles/dummy-variable-trap-regression.html (https://www.algosome.com/articles/dummy-variable-trap-regression.html)
- http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/ (http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/)