Data Exploration

Miao Fu

2023-12-03

Descriptive summary statistics for all variables

Two table with summary information on the descriptive statistics of all variables are listed below. The frequency and percentage of each categories in each categorical variable is listed out. For each numeric variable, the table includes values of mean, median, standard deviation, minimum, maximum, Q1 and Q3 values.

Categorical Variables

variable	category	count	percent
gender	female	315	53.662692
gender	male	272	46.337308
ethnic_group	group A	50	8.517888
ethnic_group	group B	123	20.954003
ethnic_group	group C	174	29.642249
ethnic_group	group D	155	26.405451
ethnic_group	group E	85	14.480409
parent_educ	associate's degree	128	21.805792
parent_educ	bachelor's degree	71	12.095400

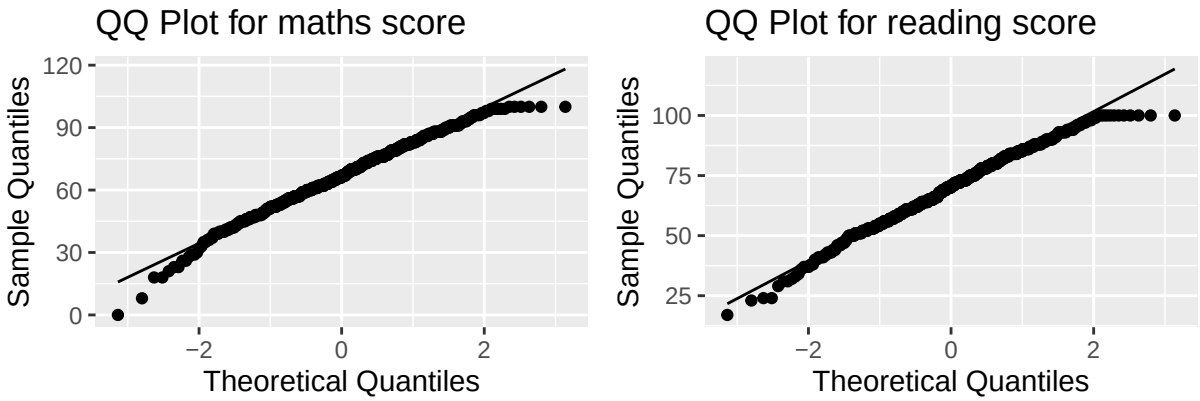
variable	category	count	percent
parent_educ	high school	122	20.783646
parent_educ	master's degree	39	6.643952
parent_educ	some college	116	19.761499
parent_educ	some high school	111	18.909710
lunch_type	free/reduced	206	35.093697
lunch_type	standard	381	64.906303
test_prep	completed	208	35.434412
test_prep	none	379	64.565588
parent_marital_status	divorced	92	15.672913
parent_marital_status	married	343	58.432709
parent_marital_status	single	137	23.339012
parent_marital_status	widowed	15	2.555366
practice_sport	never	68	11.584327
practice_sport	regularly	218	37.137990
practice_sport	sometimes	301	51.277683
is_first_child	no	192	32.708688
is_first_child	yes	395	67.291312
transport_means	private	229	39.011925
transport_means	school_bus	358	60.988075
wkly_study_hours	< 5	154	26.235094
wkly_study_hours	> 10	104	17.717206
wkly_study_hours	5-10	329	56.047700

Numeric Variables

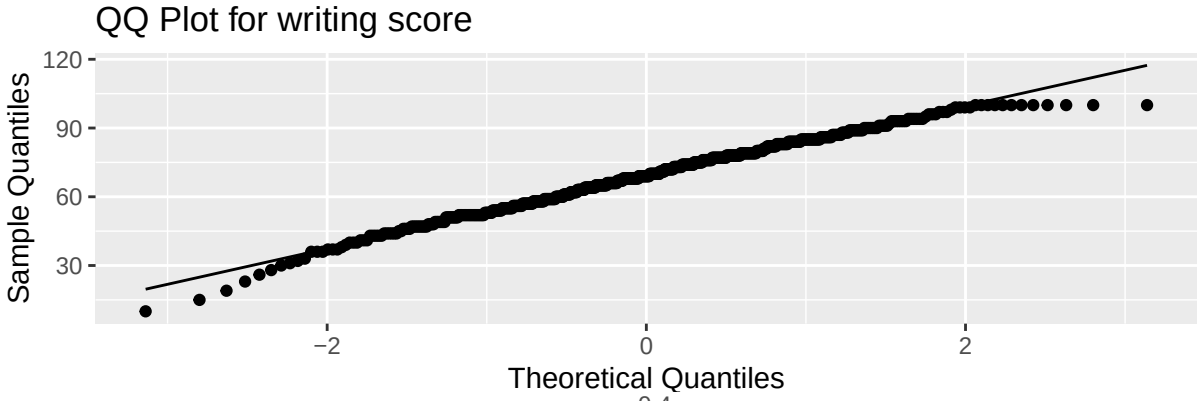
variable	mean	median	sd	minimum	maximum	q1	q3
nr_siblings	2.139693	2	1.481712	0	7	1	3
math_score	66.676320	67	16.113744	0	100	56	78
reading_score	69.846678	70	15.166662	17	100	60	81
writing_score	68.901192	69	15.550000	10	100	58	79

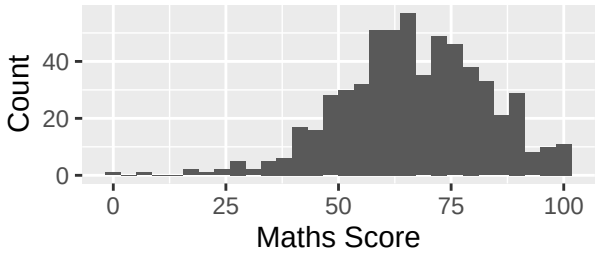
Distribution of the outcomes

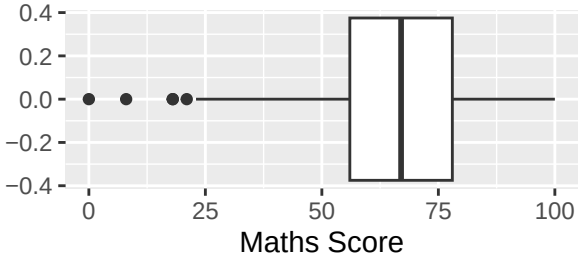
The outcome of this study includes the following variables: maths scores, reading scores, and writing scores. QQplots of the outcome variables are created to explore the distribution of each score. QQplot compares the quantiles of the data against the quantiles of a normal distribution. According the plots, majority of the data points of all three scores follow the straight qqline, which indicates they follow the normal distribution. However, there are some deviations from the line on the two ends of the distribution, which indicates the distributions might have heavier tails than normal distribution. Or, there might be skewness or outliers in the dataset. To further explore the distribution of outcomes, histograms and boxplots for the scores were incorporated. As suggested by the histograms and boxplots, all three scores are left-skewed with outliers on the left side of the distribution.

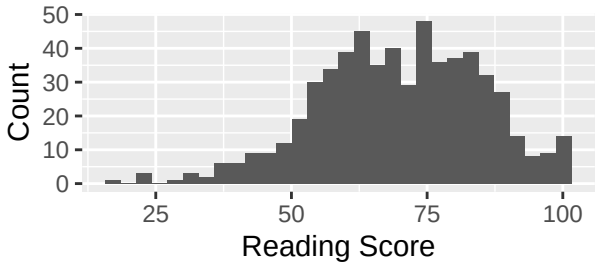

QQ Plot for maths score
Sample Quantiles
Theoretical Quantiles
120
90
60
30
0
−2
0
2

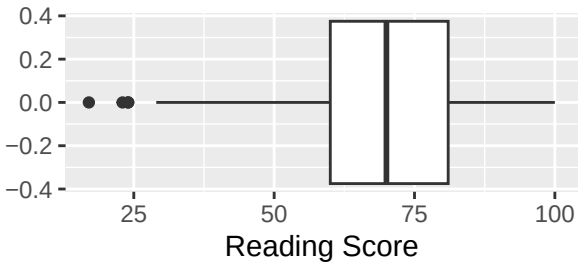
QQ Plot for reading score
Sample Quantiles
Theoretical Quantiles
100
75
50
25
−2
0
2

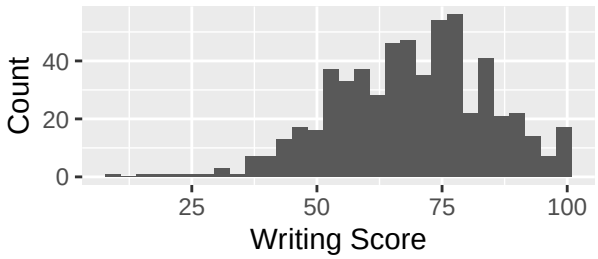

QQ Plot for writing score
Sample Quantiles
Theoretical Quantiles
120
90
60
30
−2
0
2

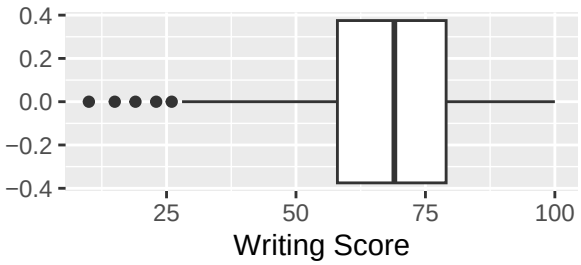

Count
Maths Score
40
20
0
0
25
50
75
100


Maths Score
0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100


Count
Reading Score
50
40
30
20
10
0
25
50
75
100

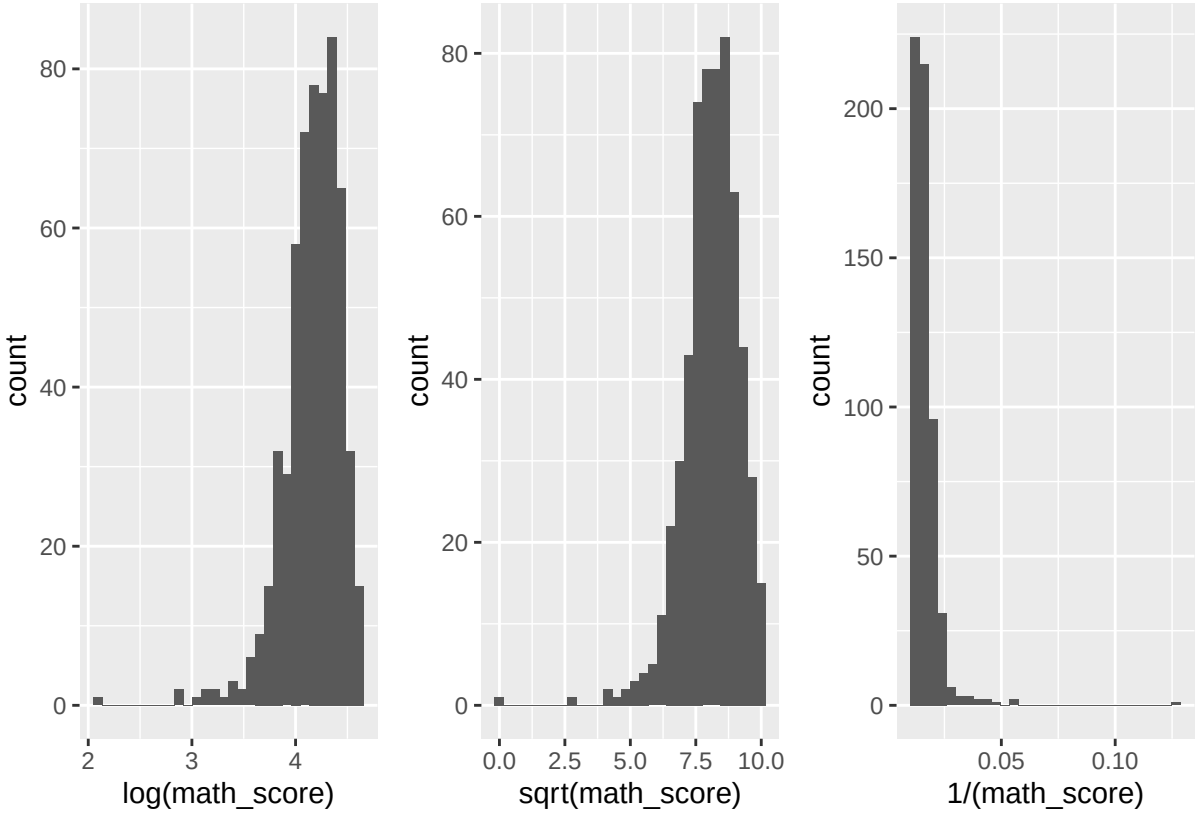

Reading Score
0.4
0.2
0.0
−0.2
−0.4
25
50
75
100

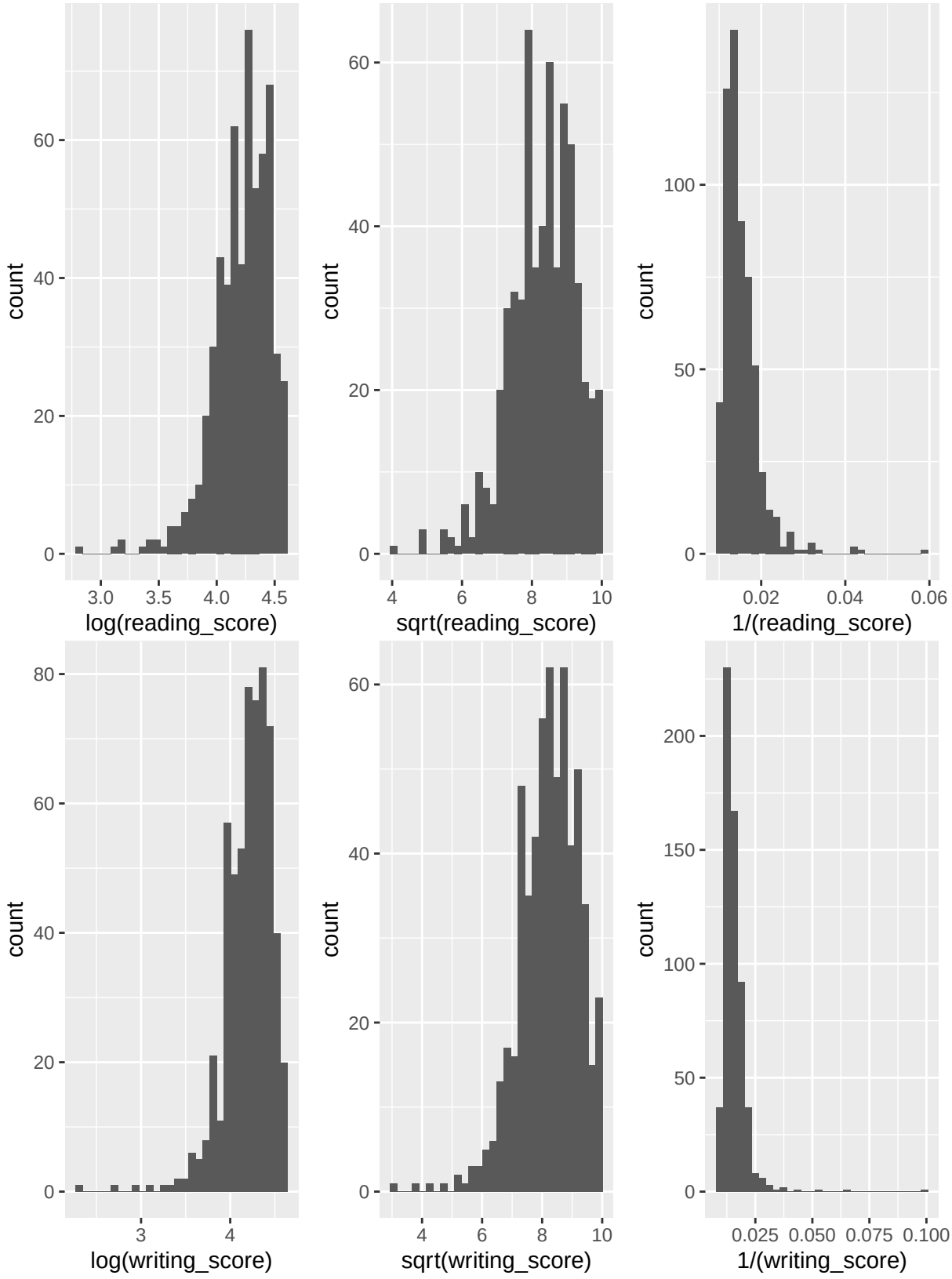

Count
Writing Score
40
20
0
25
50
75
100


Writing Score
0.4
0.2
0.0
−0.2
−0.4
25
50
75
100

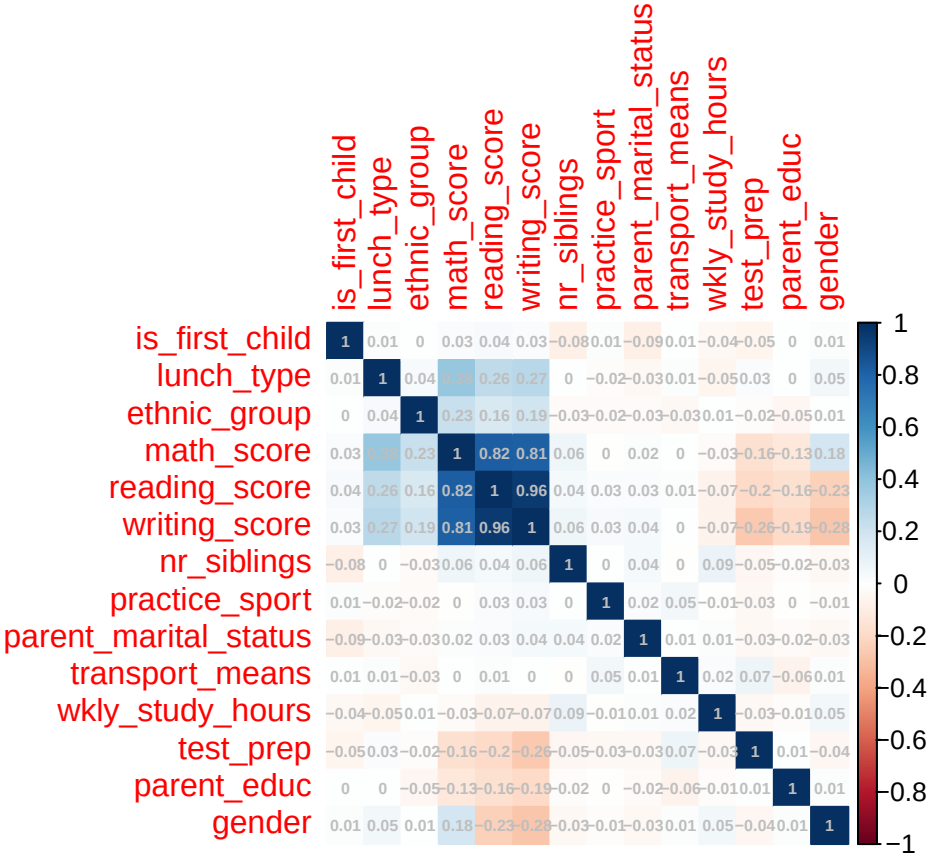
Potential transformations

Potential transformations that may help further prepare the variables for later analysis were tested. With the expectation to normalize distribution and minimize skewness and impact of outliers, three types of transformations were tested: 1) Natural logarithm 2) Square Root 3) Inverse. The resulting plots are plotted in histograms shown below. There is no apparent improvement on the distribution of the outcome through the three transformations mentioned. Thus, orignial outcome data were chosen to be used in following statistical modeling steps.




count
log(reading_score)
sqrt(reading_score)
1/(reading_score)
log(writing_score)
sqrt(writing_score)
1/(writing_score)

Pairwise relationships



By plotting our the pairwise correlation between variables, there is apparent linearity among the three scores. Other correlation coefficients are relatively small, indicating weak linear relationship between the variables.