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	488	51.476793
gender	male	460	48.523207
ethnic_group	group A	80	8.998875
ethnic_group	group B	171	19.235096
ethnic_group	group C	277	31.158605
ethnic_group	group D	237	26.659168
ethnic_group	group E	124	13.948256
parent_educ	associate's degree	198	22.122905
parent_educ	bachelor's degree	104	11.620112

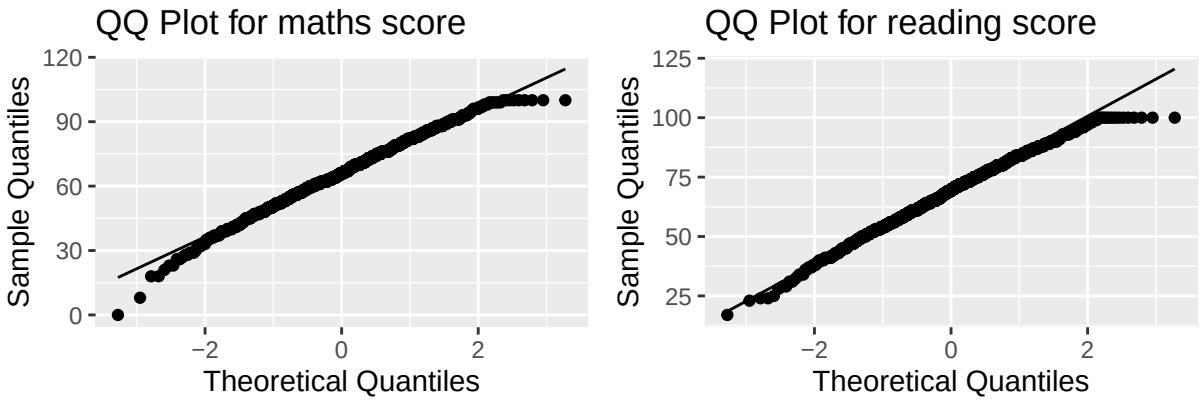
variable	category	count	percent
parent_educ	high school	176	19.664804
parent_educ	master's degree	55	6.145251
parent_educ	some college	199	22.234637
parent_educ	some high school	163	18.212291
lunch_type	free/reduced	331	34.915612
lunch_type	standard	617	65.084388
test_prep	completed	322	36.058231
test_prep	none	571	63.941769
parent_marital_status	divorced	146	16.240267
parent_marital_status	married	516	57.397108
parent_marital_status	single	213	23.692992
parent_marital_status	widowed	24	2.669633
practice_sport	never	112	12.017167
practice_sport	regularly	343	36.802575
practice_sport	sometimes	477	51.180258
is_first_child	no	314	34.204793
is_first_child	yes	604	65.795207
transport_means	private	337	39.834515
transport_means	school_bus	509	60.165485
wkly_study_hours	< 5	253	27.771680
wkly_study_hours	> 10	150	16.465423
wkly_study_hours	5-10	508	55.762898

Numeric Variables

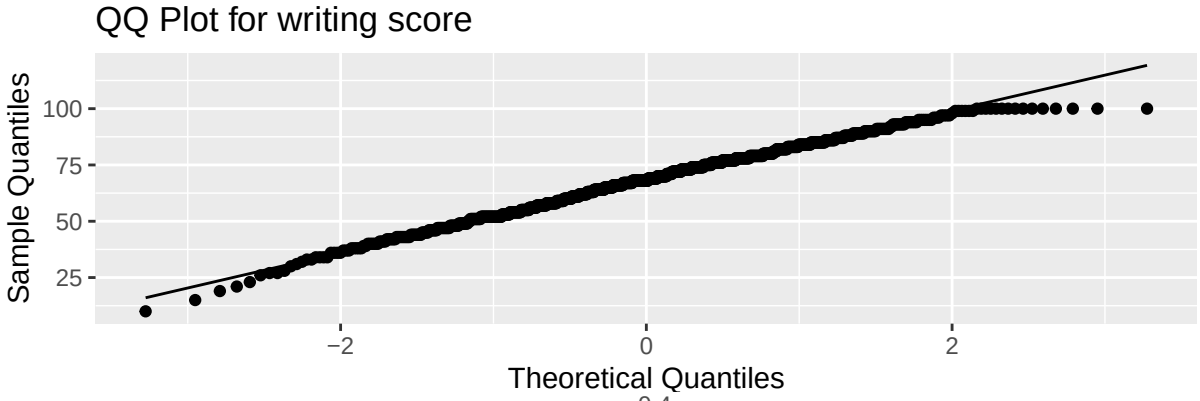
variable	mean	median	sd	minimum	maximum	q1	q3
nr_siblings	2.155211	2.0	1.483266	0	7	1	3.00
math_score	65.982067	66.0	15.530353	0	100	56	76.00
reading_score	68.841772	69.5	14.799621	17	100	59	80.00
writing_score	67.929325	68.0	15.412603	10	100	57	78.25

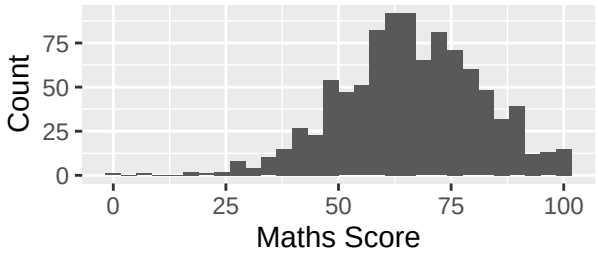
Distribution of the outcomes

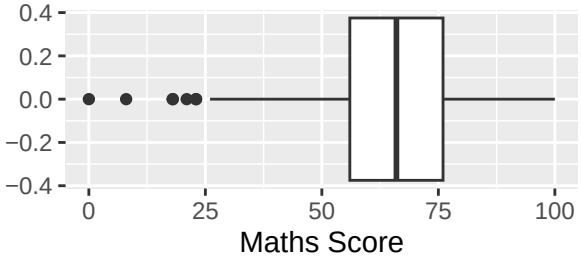
The outcome of this study includes the following variables: maths scores, reading scores, and writing scores. QQplots of the outcome variables are created to explore the distribution of each score. QQplot compares the quantiles of the data against the quantiles of a normal distribution. According the 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.

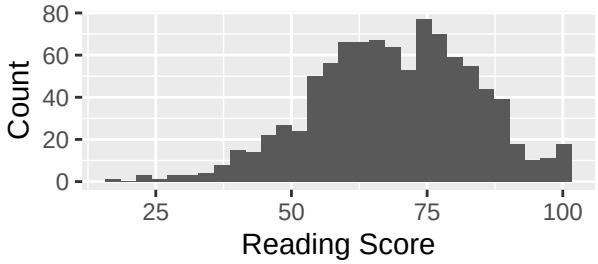

QQ Plot for maths score
Sample Quantiles
Theoretical Quantiles
120
90
60
30
0
−2
0
2

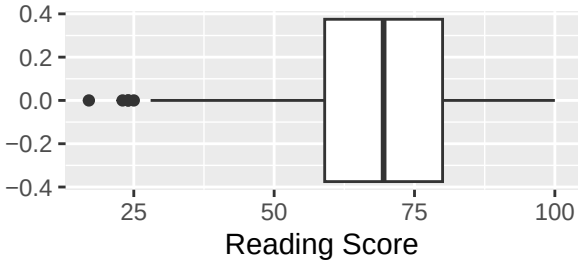
QQ Plot for reading score
Sample Quantiles
Theoretical Quantiles
125
100
75
50
25
−2
0
2

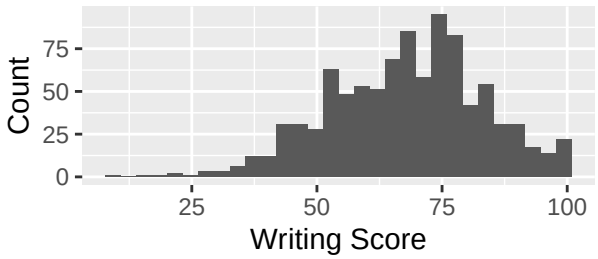

QQ Plot for writing score
Sample Quantiles
Theoretical Quantiles
100
75
50
25
−2
0
2

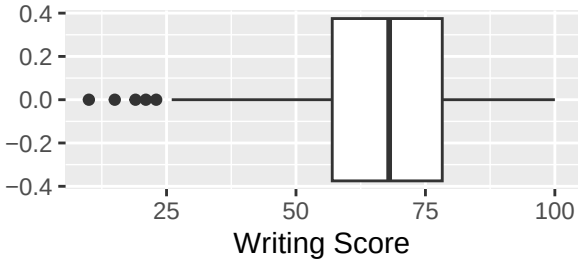

Count
Maths Score
75
50
25
0
0
25
50
75
100


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


Count
Reading Score
80
60
40
20
0
25
50
75
100

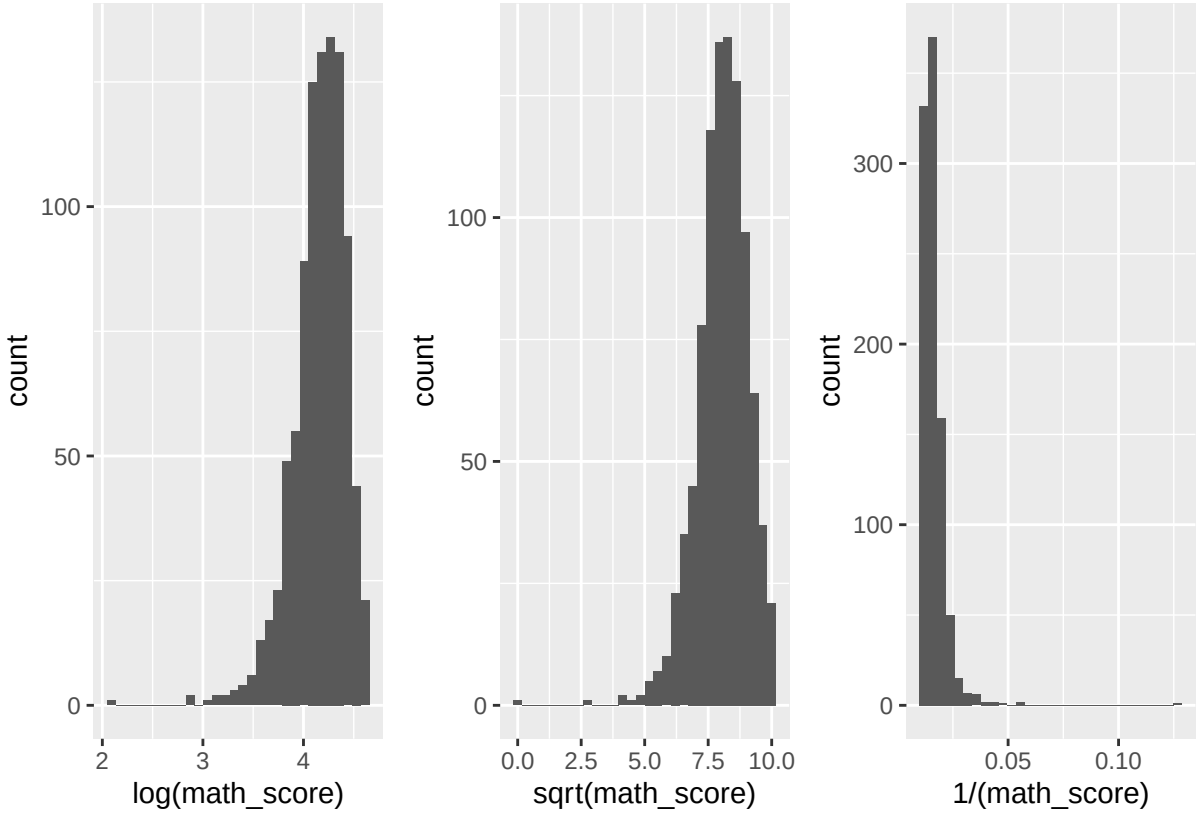

Reading Score
0.4
0.2
0.0
−0.2
−0.4
25
50
75
100

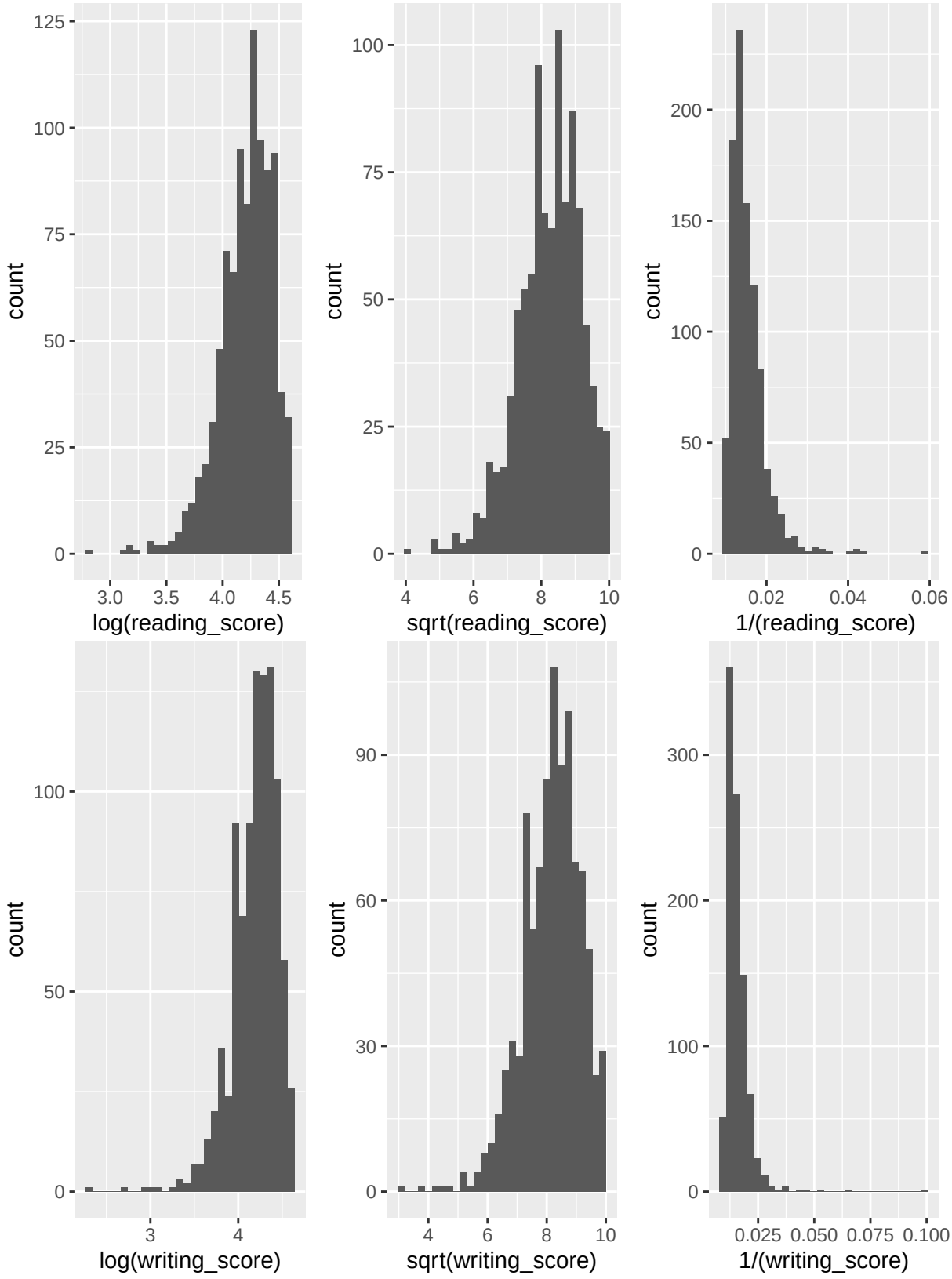

Count
Writing Score
75
50
25
0
25
50
75
100


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)