SDS Datathon

Student Performance Dataset

Renita Kurian
PES1UG20CS331

Student Performance Dataset

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	114.0	128.0	118.0
1	male	group E	associate's degree	free/reduced	completed	111.0	146.0	132.0
2	male	group B	high school	free/reduced	completed	132.0	145.0	137.0
3	male	group B	some college	standard	completed	89.0	113.0	88.0
4	male	group A	associate's degree	standard	completed	118.0	134.0	119.0
...	...	...	...	...	...	...	...	...
995	male	group A	some high school	standard	none	130.0	149.0	139.0
996	male	group C	associate's degree	standard	completed	104.0	111.0	99.0
997	male	group C	bachelor's degree	free/reduced	completed	101.0	125.0	109.0
998	male	group C	bachelor's degree	free/reduced	completed	110.0	134.0	121.0
999	male	group C	some college	free/reduced	completed	119.0	142.0	130.0

	math score	reading score	writing score
count	997.000000	996.000000	996.000000
mean	107.664995	122.341365	111.401606
std	16.758326	17.490096	16.297959
min	50.000000	51.000000	54.000000
25%	97.000000	113.000000	101.000000
50%	108.000000	124.000000	113.000000
75%	119.000000	134.000000	123.000000
max	149.000000	150.000000	144.000000

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage
0	female	group B	bachelor's degree	standard	none	114.0	128.0	118.0	80.00
1	male	group E	associate's degree	free/reduced	completed	111.0	146.0	132.0	86.44
2	male	group B	high school	free/reduced	completed	132.0	145.0	137.0	92.00
3	male	group B	some college	standard	completed	89.0	113.0	88.0	64.44
4	male	group A	associate's degree	standard	completed	118.0	134.0	119.0	82.44
...	...	...	...	...	...	...	...	...	...
995	male	group A	some high school	standard	none	130.0	149.0	139.0	92.89
996	male	group C	associate's degree	standard	completed	104.0	111.0	99.0	69.78
997	male	group C	bachelor's degree	free/reduced	completed	101.0	125.0	109.0	74.44
998	male	group C	bachelor's degree	free/reduced	completed	110.0	134.0	121.0	81.11
999	male	group C	some college	free/reduced	completed	119.0	142.0	130.0	86.89

	math score	reading score	writing score	percentage
count	997.000000	996.000000	996.000000	992.000000
mean	107.664995	122.341365	111.401606	75.851653
std	16.758326	17.490096	16.297959	9.480590
min	50.000000	51.000000	54.000000	39.330000
25%	97.000000	113.000000	101.000000	69.110000
50%	108.000000	124.000000	113.000000	76.440000
75%	119.000000	134.000000	123.000000	82.670000
max	149.000000	150.000000	144.000000	96.890000

Adding Percentage Column

The percentage of each student is calculated and added to the dataframe. Since the maximum score is taken to be 150, percentage is calculated by –

percentage = (sum of scores /450) *100

It can be seen that that the mean percentage is 75.85%. The max and min percentages are 96.89% and 39.33%.

```
df.isnull().sum()
```

```
gender                       0
race                         0
parental level of education  3
lunch                        0
test preparation course      0
math score                   3
reading score                4
writing score                4
percentage                   8
dtype: int64
```

Before Data Cleaning

After Data Cleaning

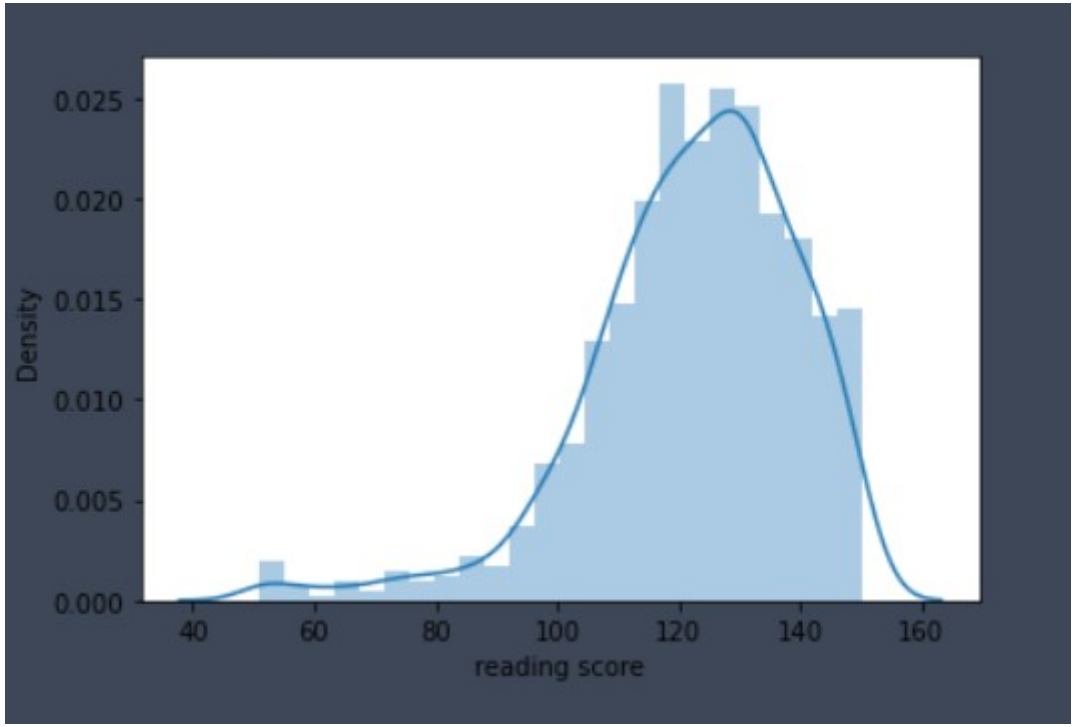
```
df.isnull().sum()
```

```
gender                       0
race                         0
parental level of education  0
lunch                        0
test preparation course      0
math score                   0
reading score                0
writing score                0
percentage                   0
dtype: int64
```

Cleaning the Dataset

From the given data, it can be seen that the given dataset has 3 null values in parent level of education and 3,4 and 4 null values in math, reading and writing scores. These null values must either be replaced with average value or must be dropped.

After applying the fillna() and dropna() functions the null values are removed.



Reading Score Distribution

The reading score distribution is left skewed.

Formula for left skewed:

Skew = 3 * (Mean – Median) / Standard Deviation

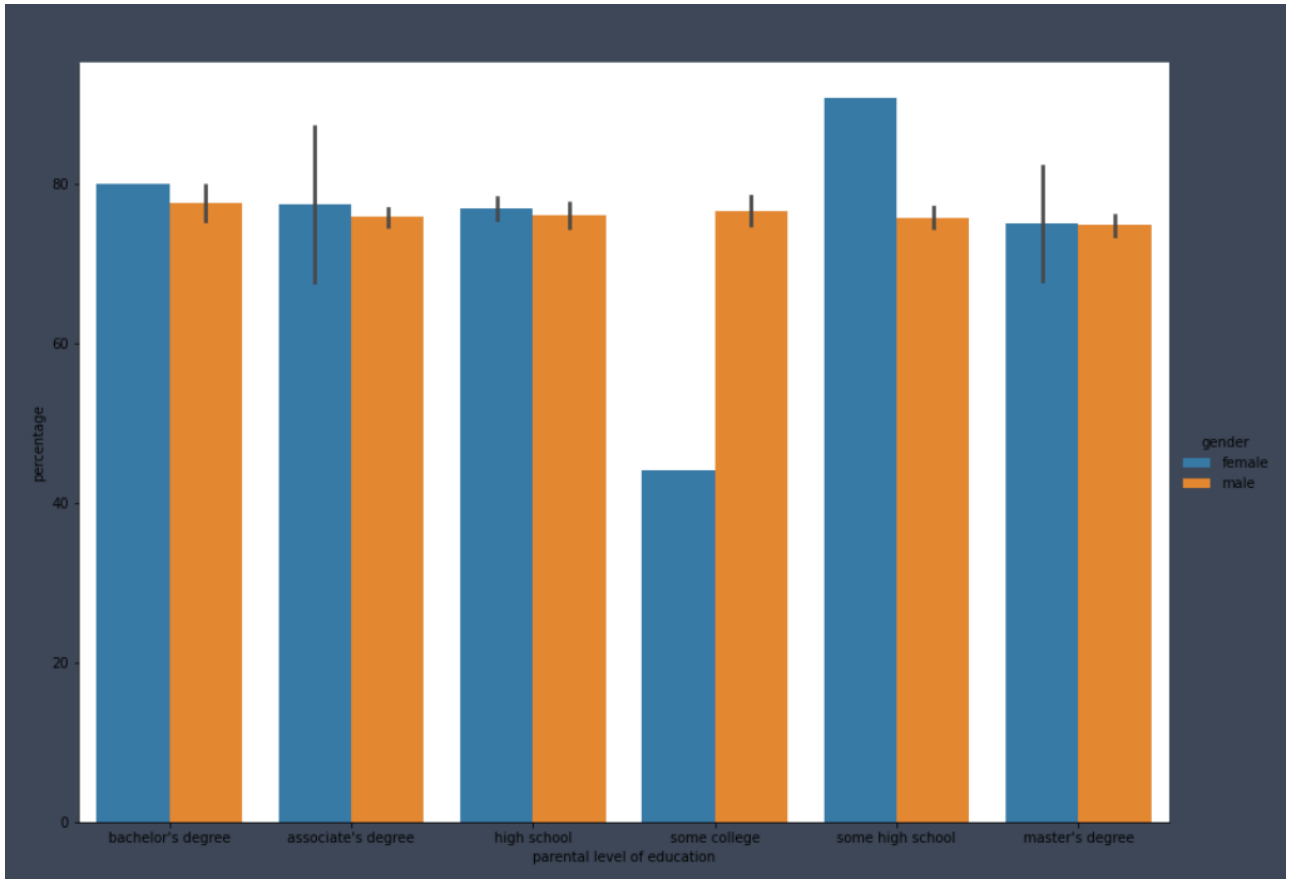
•Left skewed graphs are also called negatively skewed

•Large number of data values occur on the right side with a fewer number of data values on the left side.

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	grade
0	female	group B	bachelor's degree	standard	none	114.0	128.0	118.0	80.00	A
1	male	group E	associate's degree	free/reduced	completed	111.0	146.0	132.0	86.44	A
2	male	group B	high school	free/reduced	completed	132.0	145.0	137.0	92.00	S
3	male	group B	some college	standard	completed	89.0	113.0	88.0	64.44	C
4	male	group A	associate's degree	standard	completed	118.0	134.0	119.0	82.44	A
...	...	...	...	...	...	...	...	...	...	...
995	male	group A	some high school	standard	none	130.0	149.0	139.0	92.89	S
996	male	group C	associate's degree	standard	completed	104.0	111.0	99.0	69.78	C
997	male	group C	bachelor's degree	free/reduced	completed	101.0	125.0	109.0	74.44	B
998	male	group C	bachelor's degree	free/reduced	completed	110.0	134.0	121.0	81.11	A
999	male	group C	some college	free/reduced	completed	119.0	142.0	130.0	86.89	A

Grades

The students have been assigned grades based on the percentages that have been calculated previously. A student scoring above 90 gets S grade, 80-90 - A, 70-80 gets B and so on. The lowest grade, F, is given to students scoring less than 40. The grades are added to dataframe using map function.



Distribution of Percentage across Parental Level Of Education and Gender

Blue – female
Orange – male

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	grade
445	male	group D	associate's degree	standard	completed	109.0	130.0	121.0	80.00	A
288	male	group C	associate's degree	standard	completed	130.0	131.0	120.0	84.67	A
463	male	group C	high school	free/reduced	completed	119.0	146.0	135.0	88.89	A
914	male	group B	associate's degree	free/reduced	completed	88.0	117.0	99.0	67.56	C
387	male	group E	bachelor's degree	standard	completed	99.0	134.0	117.0	77.78	B
...	...	...	...	...	...	...	...	...	...	...
938	male	group C	high school	free/reduced	completed	127.0	137.0	129.0	87.33	A
598	male	group E	master's degree	standard	none	111.0	133.0	117.0	80.22	A
478	male	group E	associate's degree	free/reduced	completed	97.0	120.0	114.0	73.56	B
167	male	group A	some high school	standard	none	100.0	88.0	122.0	68.89	C
805	male	group A	associate's degree	standard	completed	117.0	137.0	118.0	82.67	A

Simple Random Sampling

A random sample is generated

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	grade
304	male	group A	some college	standard	completed	116.0	131.0	127.0	83.11	A
848	male	group A	some high school	standard	completed	101.0	128.0	112.0	75.78	B
215	male	group A	some college	free/reduced	completed	126.0	139.0	119.0	85.33	A
184	male	group A	associate's degree	free/reduced	completed	87.0	108.0	93.0	64.00	C
854	male	group A	high school	free/reduced	completed	104.0	120.0	99.0	71.78	B
...	...	...	...	...	...	...	...	...	...	...
620	male	group E	associate's degree	standard	completed	77.0	117.0	98.0	64.89	C
927	male	group E	associate's degree	standard	completed	107.0	117.0	115.0	75.33	B
954	male	group E	master's degree	standard	completed	104.0	125.0	83.0	69.33	C
182	male	group E	associate's degree	free/reduced	none	92.0	106.0	91.0	64.22	C
745	male	group E	bachelor's degree	standard	completed	114.0	135.0	118.0	81.56	A

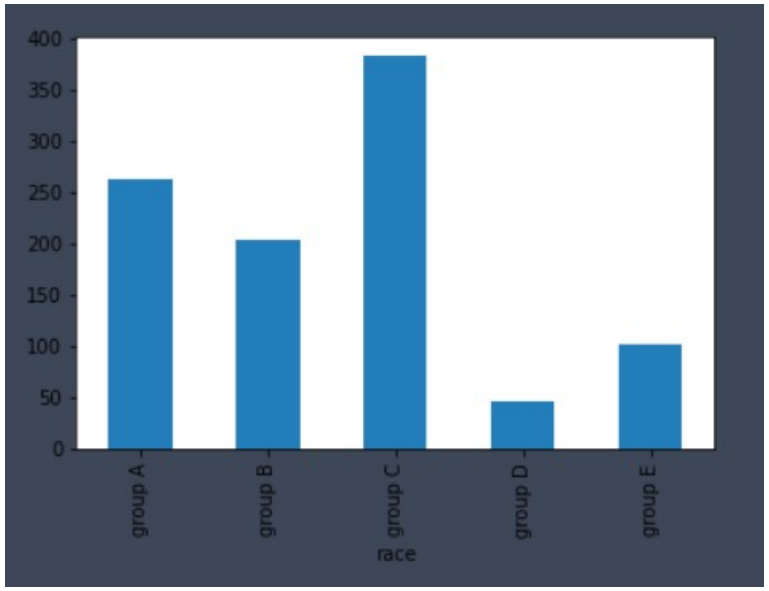
Stratified Random Sampling

This dataset has been generated by stratified sampling. The number of records for each race in sample is in proportion to the population. This has been done by first obtaining 5 different dataframes for each race. Then a random sample is collected from each dataframe. These samples are then merged to get a stratified sample. The groups A,B,C,D and E have 26,20,39,5 and 10 records respectively.

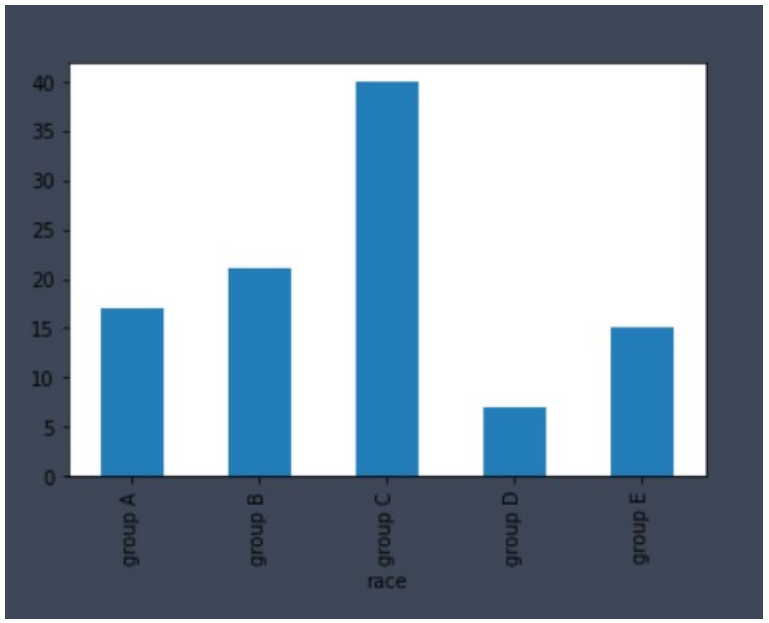
```
Population Mean:  107.66

Mean for Sample 1:  107.96
Sampling Error for Sample 1:  0.27 %
Mean for Sample 2:  109.22
Sampling Error for Sample 2:  1.44 %
```

Sampling Means and Errors

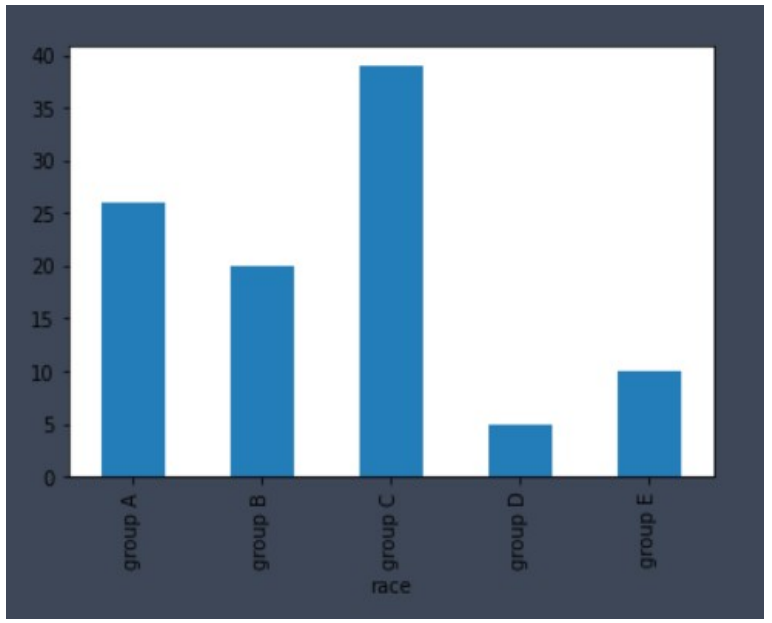
The sample means for math scores for a and b are 107.96 and 109.22. The mean for population was calculated to be 107.66. The sampling error for a and b are 0.27% and 1.44%. Hence, a is a better sample than b since it has a lower sampling error.



Population



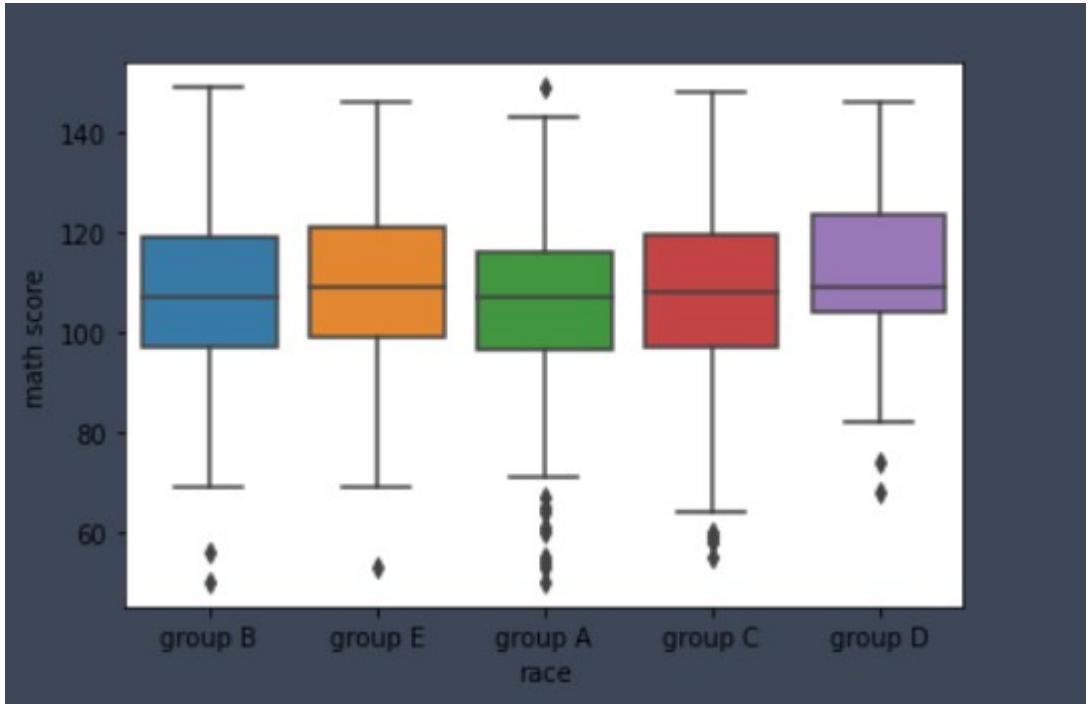
Sample a



Sample b

Race Distribution

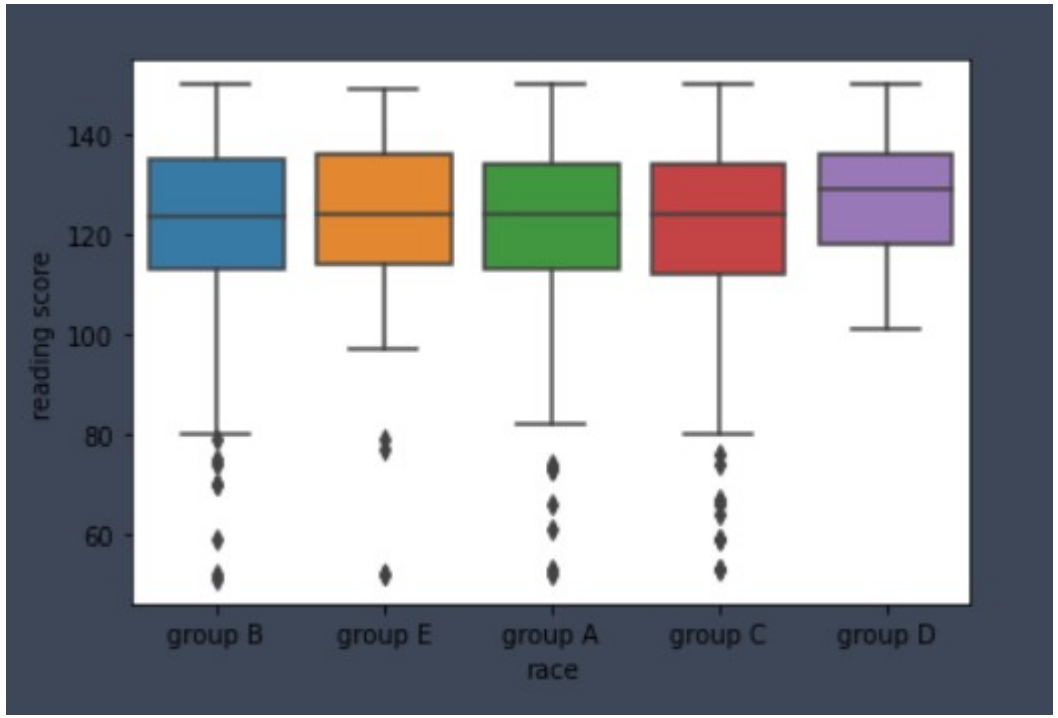
The given bar charts show the distribution of race in population and samples a and b. It can be seen that the distribution of race is similar in sample b and population and varies slightly in sample a. However, from the data calculated in previous steps sample a has a lower error compared to b and is hence a better sample. It can thus be inferred that race does not play a significant role in student perfomance.



Math Score Boxplot

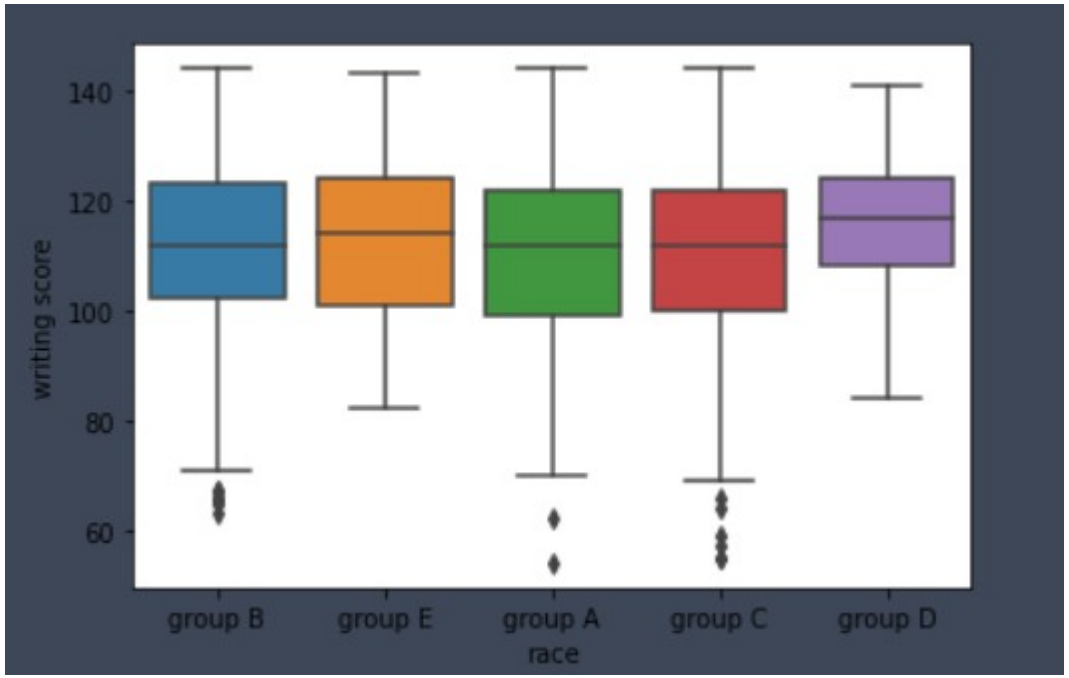
Group A has the highest number of outliers

Reading Score Boxplot



It is seen that Group C has the highest number of outliers. Group A and B also has a lot of outlier values

Writing Score Boxplot



Group C has the highest number of outliers