SDS DATATHON

NAME: Sukruth Keshava Gowda

SECTION: H

SRN: PES1UG20CS443

ROLL NO.: 22

Introductory tasks:

1.Have added percentage column

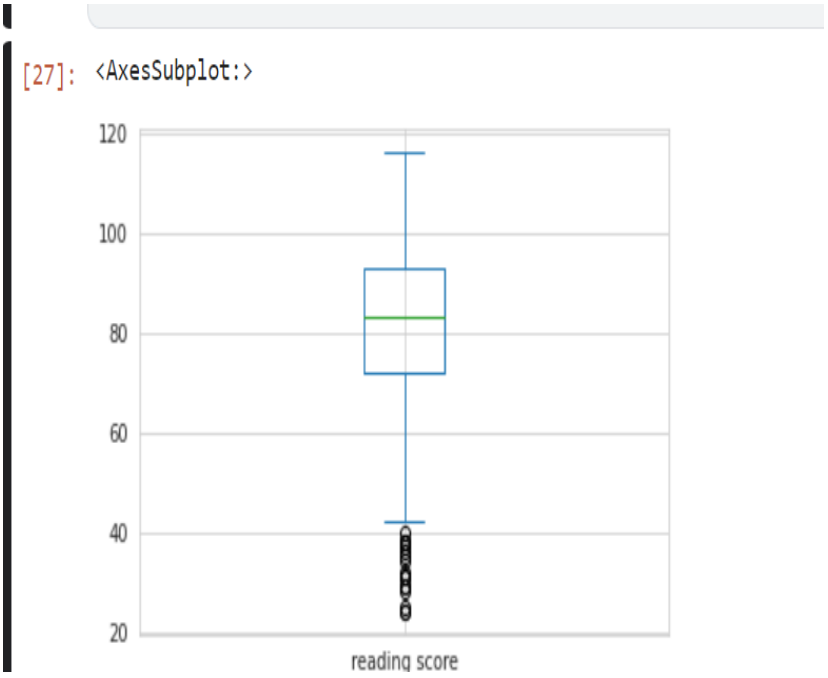
```
data["percentage"]=((data["math score"]+data["reading score"]+data["writing score"])/300)*100
data
```

	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	86.0	86.0	87.0	86.333333
1	male	group A	master's degree	standard	completed	83.0	104.0	118.0	101.666667
2	male	group E	high school	standard	completed	104.0	109.0	106.0	106.333333
3	male	group C	associate's degree	standard	none	61.0	71.0	57.0	63.000000
4	male	group A	associate's degree	standard	completed	90.0	92.0	110.0	97.333333
...	...	...	...	...	...	...	...	...	...
995	female	group C	some college	standard	none	102.0	113.0	108.0	107.666667

2. Performed data cleaning by first filling in the missing numerical values using fillna function with their mode

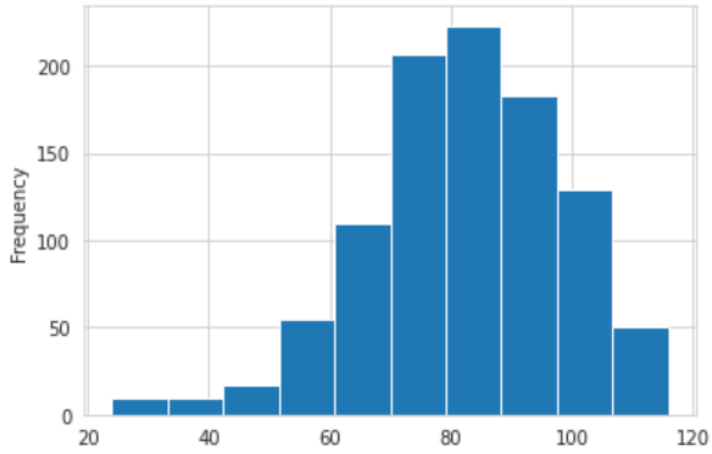
value and removed rows with missing categorical values using dropna function .

3.




```
data["reading score"].plot.hist()
```

```
<AxesSubplot:ylabel='Frequency'>
```



As seen in the graphs the distribution is left skewed in nature (mean < median < mode)

The boxplot shows most of the students have scored above 72 and below 93

Inter-Quartile range(IQR) = Q3-Q1=93-72=21

MIN = Q1 - 1.5*IQR = 40.5

MAX = Q3 + 1.5*IQR = 124.5

Values below 41 are considered outliers

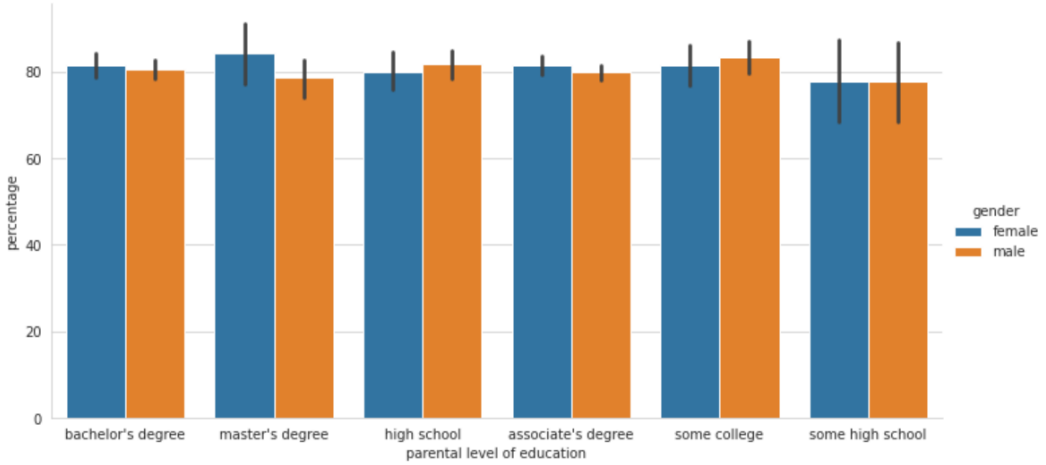
4.Have added grades using lambda function.

```
data['Grades'] = data.apply(lambda x : 'F' if x['percentage']<40 else ('D' if x['percentage']>=40 and x['
data
```

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	Percentage	Grades
0	female	group B	bachelor's degree	standard	none	86.0	86.0	87.0	86.333333	87.0	A
1	male	group A	master's degree	standard	completed	83.0	104.0	118.0	101.666667	118.0	S
2	male	group E	high school	standard	completed	104.0	109.0	106.0	106.333333	106.0	S
3	male	group C	associate's degree	standard	none	61.0	71.0	57.0	63.000000	57.0	C
4	male	group A	associate's degree	standard	completed	90.0	92.0	110.0	97.333333	110.0	S

5. Have plotted graph visualizing the distribution of percentage across parental level of education split by gender.

]: <seaborn.axisgrid.FacetGrid at 0x7ffa1bbf2610>



TASK QUESTIONS:

1.A) Have taken a random sample of 100 people.

```
subset = data.sample(100)
subset
```

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	Percentage	Grades
186	male	group A	bachelor's degree	free/reduced	completed	94.0	90.0	78.0	87.333333	78.0	A
985	female	group D	bachelor's degree	standard	none	71.0	65.0	67.0	67.666667	67.0	C
213	male	group B	associate's degree	standard	none	74.0	65.0	69.0	69.333333	69.0	C
128	female	group C	associate's degree	standard	completed	96.0	96.0	87.0	93.000000	87.0	S
99	female	group E	master's degree	standard	none	79.0	81.0	75.0	78.333333	75.0	B

1.B)Have taken a stratified sample of 100 people using race as the strata

[68]:

```
def stratified_sample(data, color,n):
    n1 = min(n, data[color].value_counts().min())
    df_ = data.groupby(color).apply(lambda x: x.sample(n1))
    df_.index = df_.index.droplevel(0)
    return df_
strata = stratified_sample(data,'race',100)
strata
```

[68]:

	gender	race	parental level of education	lunch	test preparation course	math score	reading score	writing score	percentage	Percentage	Grades
966	male	group A	associate's degree	standard	completed	80.0	82.0	77.0	79.666667	77.0	B
207	female	group A	associate's degree	standard	completed	115.0	94.0	89.0	99.333333	89.0	S
116	male	group A	associate's degree	standard	none	89.0	99.0	95.0	94.333333	95.0	S
692	female	group A	associate's degree	free/reduced	none	80.0	88.0	94.0	87.333333	94.0	A

Console

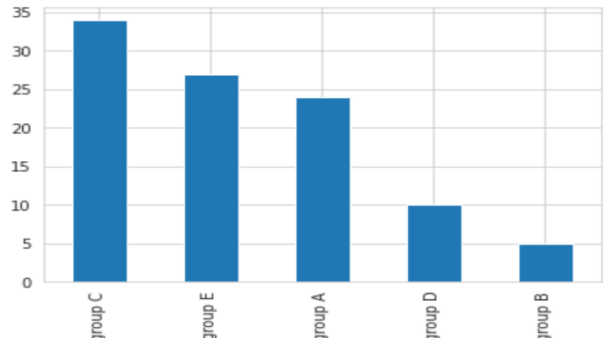
1.C)Calculated mean and sampling error of both the samples . Mean of random sampling is found to be the highest while mean of stratified sampling is found close to mean of population.The sampling error of stratified sampling is found to be lesser than that of random sampling.

```
random_mean = subset["math score"].mean()
print(random_mean)
strata_mean = strata["math score"].mean()
print(strata_mean)
mean = data["math score"].mean()
print(mean)
a = subset["math score"].std() #standard deviation of random sample
b = strata["math score"].std() #standard deviation of stratified sample
z_random = random_mean - mean
z_strata = strata_mean - mean
error_random = (z_random*a)/10 # root n is replaced by 10 since sample size is 100
error_strata = (z_strata*b)/10
print(error_random)
print(error_strata) #the error in stratified sampling is lower
```

```
80.38
79.29230769230769
79.29447236180904
1.7811516603488553
-0.0035233846863393183
```

```
#random
subset['race'].value_counts().plot(kind='bar')
```

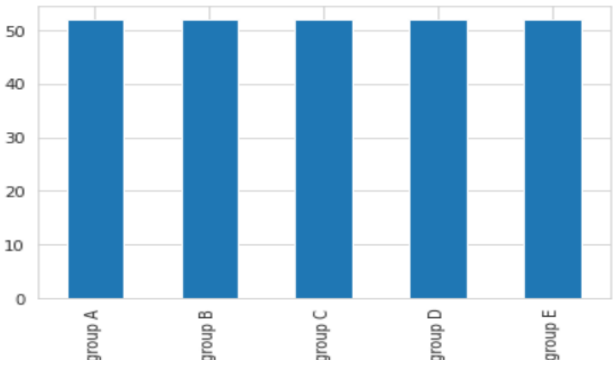
```
<AxesSubplot:>
```



1.D)

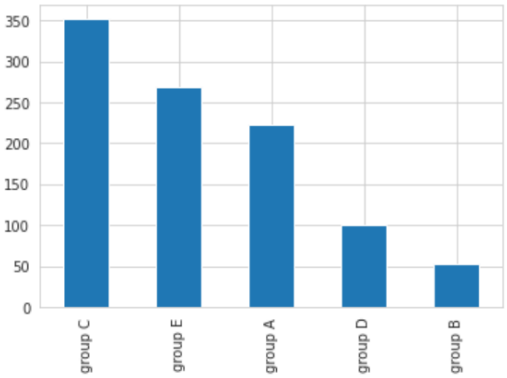
```
#stratified
strata['race'].value_counts().plot(kind='bar') #
```

```
<AxesSubplot:>
```




```
#population
data['race'].value_counts().plot(kind='bar')
```

```
<AxesSubplot:>
```



+ Code + Markdown

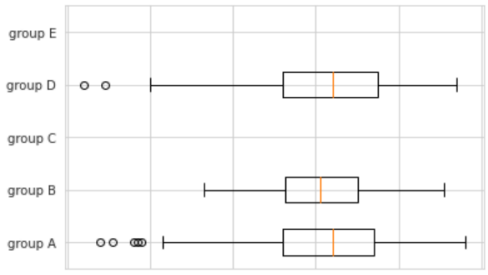
It is found that the distribution of random sampling is similar to the population since random elements were

picked while the distribution of stratified sampling is evenly distributed since race was used as strata.

2. Based on the boxplots in maths group A has the most outliers, in reading group A again has the most outliers which is due to huge population size of group A and in writing group C has the most outliers . Altogether Group A has the most number of outliers.

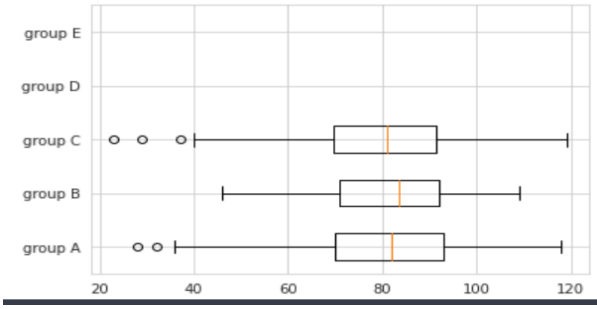
```
rew=data[data['race']=='group E']['reading score']
values1=[raw,rbw,rcw,rdw,rew]
plt.boxplot(values1, vert=False,labels=['group A','group B','grou
plt.show
#group a as the most outliers
```

```
]: <function matplotlib.pyplot.show(close=None, block=None)>
```



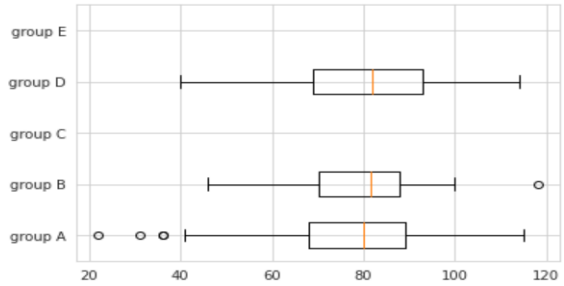
#reading

```
bdw=data[data[ race ]== group D ][ writing score ]
bew=data[data['race']=='group E']['writing score']
values1=[baw,bbw,bcw,bdw,bew]
plt.boxplot(values1, vert=False,labels=['group A','grou
plt.show()
#group c has the most outliers
```



#writing

```
bem=data[data['race']=='group E']['math score']
values=[bam,bbm,bcm,bdm,bem]
plt.boxplot(values, vert=False,labels=['group A','group B','g
plt.show()
#group a has the most outliers
```



#maths