TOTAL POINTS- 32 POINTS
-70 MINUTES-

Student Performance Challenge

Questions and Answers
Easy -identification
Medium -
Hard- analysis

PART 1

1. How many rows are present in the dataset? (1mrk)

   1000 rows

2. What is the mean and minimum writing score for Student Exam performance? (1mrk)

   68.05 , 10.00

3. What is the mean and minimum reading score for Reading Score performance? (1 mrk)

   69.16 , 17.00

4. What is the most and least frequent race/ethnicity of Students within the Dataset?

   Group C , group A

5. What is the missing value distribution within column features of the dataset? (1 mrk)

   Non-Existent (0.0%) NAN-Value distribution.

6. List the following columns with only numerical data within Dataset? (2 mrk)

   Math_Score, Reading Score, Writing Score

7. List the following minimum scores in maths, reading and writing respectively? (2 mrk)

Math_Score =0 , Reading_Score = 17 , Writing_Score = 10

8. List the following columns with only categorical data within the Dataset? (2mrk)

Gender, Race/Ethnicity, Parental_Level_of_Education, Lunch, Test_Preparation_Course.

9. List the following, students' lunch distribution by their Gender? (2mrk)

Gender	Lunch	Count
Female	standard	329
	Free/reduced	189
Male	standard	316
	Free/reduced	166

10. List the following, students' parental level education distribution by their Gender? (2 mrk)

Gender	Parental_Level_of_Education	Count
Female	some college	118
	associate college	116
	high school	94
	Some high school	91

	bachelor ‘s degree	63
	master’s degree	36
Female	some college	108
	Associate’s degree	106
	high school	102
	some high school	88
	bachelor’s degree	55
	master’s degree	23

11. List the following students’ lunch , parental level education by their gender? Infer any insights on analysis generated. (2 mrk)

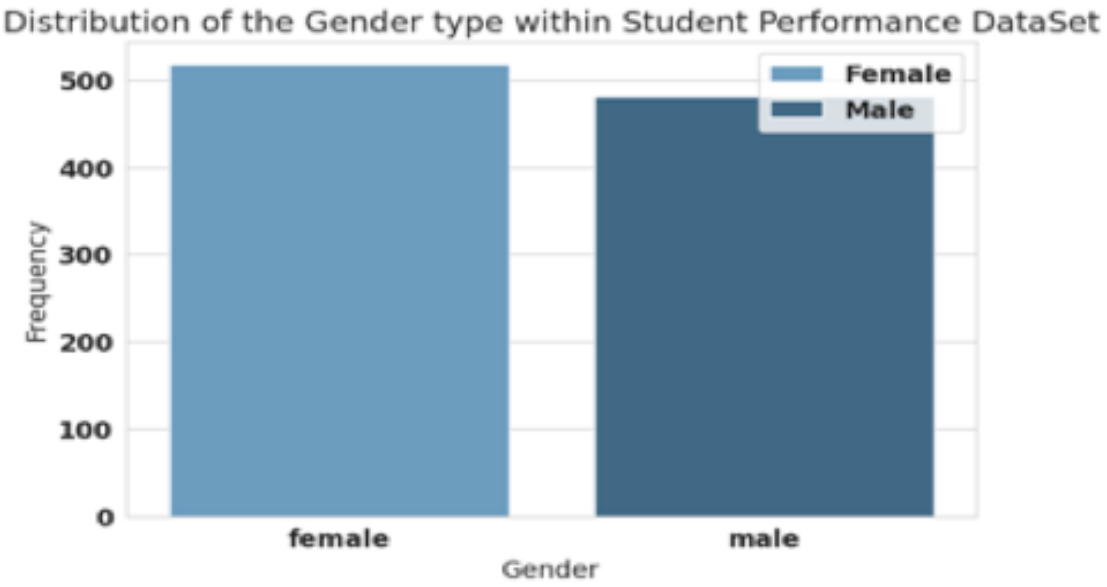
Gender	Parental_Level_of_Education Test	Preparation Course	Count
Female	Some college	none	76
	associate’s degree	none	74
	high school	none	65
	some high school	none	56
	Bachelor’s Degree	none	41
	Master’s Degree	none	22
	[redacted]		
	some college	completed	42
	associate’s degree	completed	42
	High school	completed	29
	some high school	completed	35
	Bachelor’s Degree	completed	22
	Master’s Degree	completed	14
Male	some college	none	73

Gender	Parental_Level_of_Education Test	Preparation Course	Count
Female	Some college	none	76
	associate’s degree	none	74
	high school	none	65
	some high school	none	56
	Bachelor’s Degree	none	41
	Master’s Degree	none	22
	[redacted]		
	some college	completed	42
	associate’s degree	completed	42
	High school	completed	29
	some high school	completed	35
	Bachelor’s Degree	completed	22
	Master’s Degree	completed	14
	associate’s degree	none	66
	High school	none	75
	some high school	none	46
	bachelor’s degree	none	31
	master’s degree	none	17
	[redacted]		
	some college	completed	35
	associate’s degree	completed	40
	high school	completed	27
	Some high school	completed	42
	bachelor’s degree	completed	24
	master’s Degree	completed	6

Part 2

12. Create a bar plot showcasing the gender distribution of Student Performance in Exams? (2 mrk)

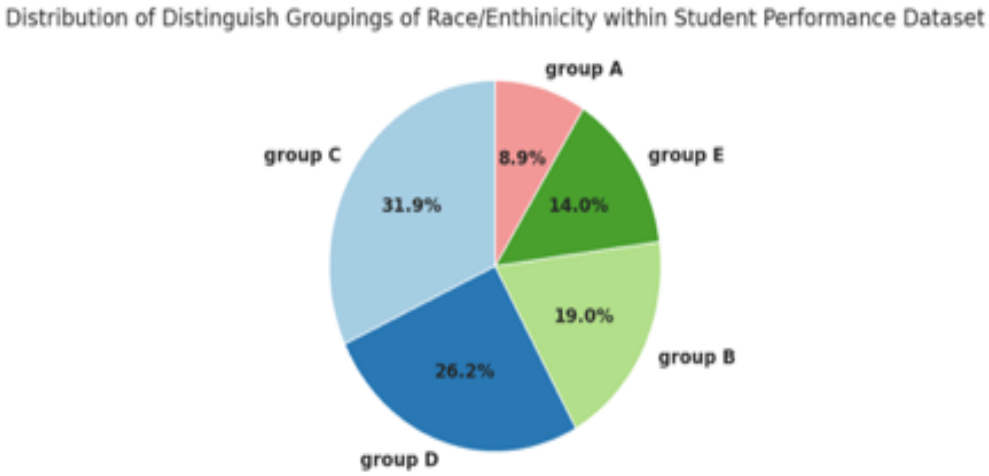
```
sns.barplot(x= Student_Perf['gender'].value_counts().index, y=Student_Perf['gender'].value_counts(),palette="Blues_d", hue=['female','male'])
plt.xlabel('Gender')
plt.ylabel('Frequency')

plt.title('Distribution of the Gender type within Student Performance DataSet')
plt.show()
```



13. Create a pie chart showcasing the distribution of diverse race/ethnicity grouping within the Student Performance Dataset? (2 mrk)

(i)

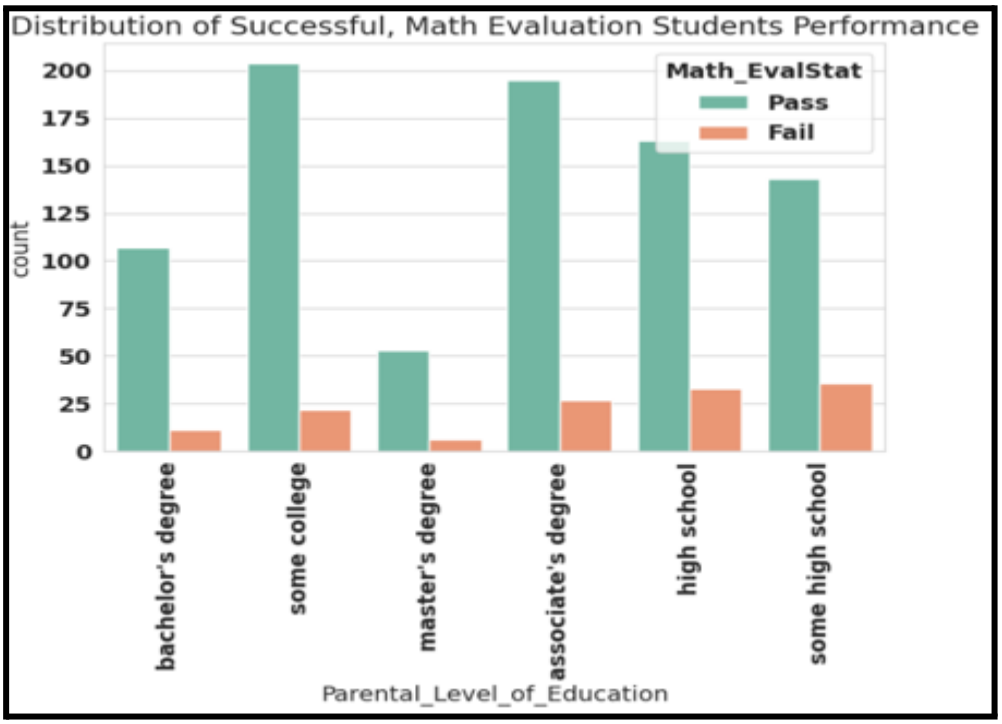
```
# Value Counts for 'Race/Ethnicity' column
race_ethnicity_value_count = Student_Perf['Race/Ethnicity'].value_counts()

#Plotting a pie chart
plt.figure(figsize=(8,6))
plt.pie(race_ethnicity_value_count, labels=race_ethnicity_value_count.index, colors=plt.cm.Paired.colors, autopct='%1.1f%%', startangle=90)
plt.title("Distribution of Distinguish Groupings of Race/Enthinicity within Student Performance Dataset")
plt.show()
```



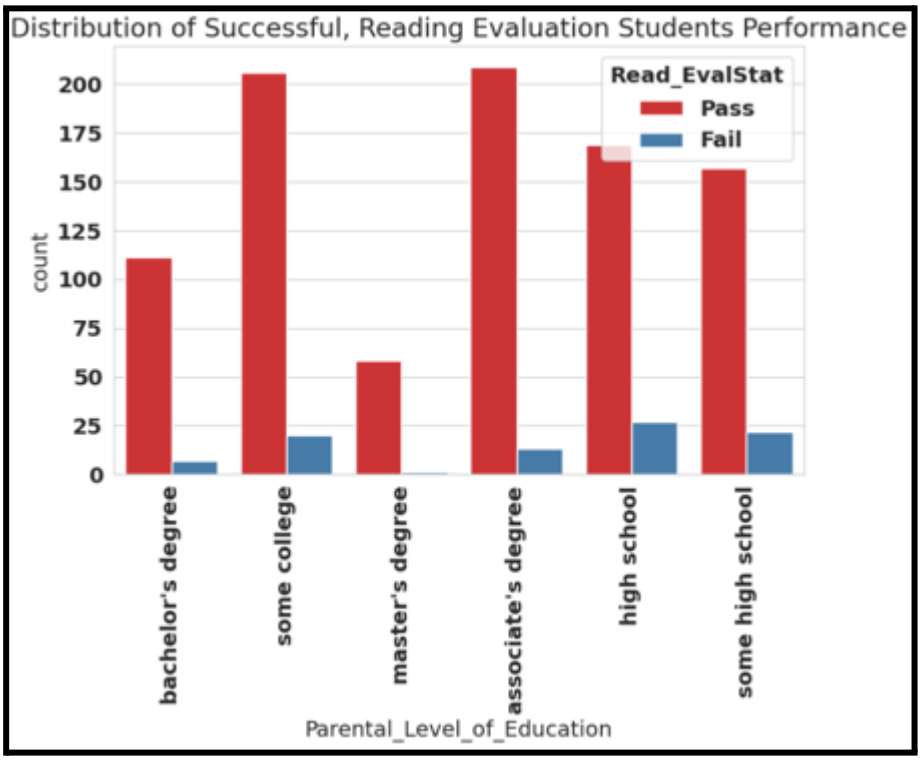
14. Create a graph illustrating the evaluation distribution of successful students passing their Maths, Writing and Reading subjects? (2 mrk)

i) Determine the evaluation of successful students who passed all the their subjects Maths, Writing and Reading accordingly. (3mrk)

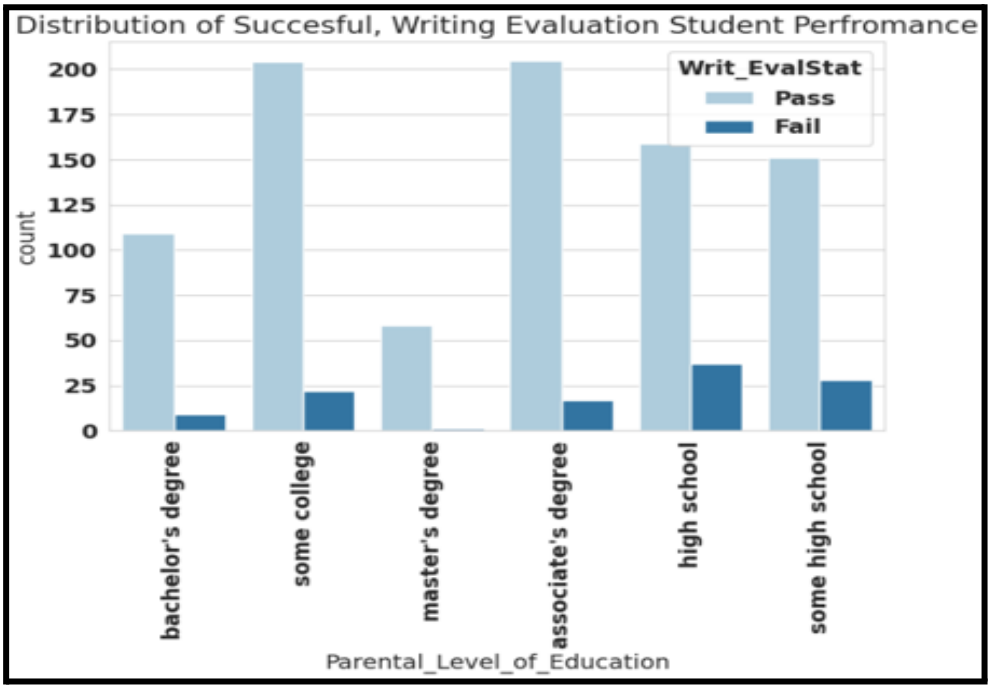
```
p= sns.countplot(x='Parental_Level_of_Education', data = Student_Perf, hue = 'Math_EvalStat', palette='Set2')
_ = plt.setp(p.get_xticklabels(), rotation =90)
plt.title('Distribution of Successful, Math Evaluation Students Performance')
```



```
p = sns.countplot(x ='Parental_Level_of_Education', data=Student_Perf, hue ='Read_EvalStat', palette='Set1')
_ = plt.setp(p.get_xticklabels(), rotation=90)
plt.title('Distribution of Successful, Reading Evaluation Students Performance')
```




```
p = sns.countplot(x ='Parental_Level_of_Education',data=Student_Perf , hue='Writ_EvalStat', palette='Paired')
_ = plt.setp(p.get_xticklabels(), rotation=90)
plt.title('Distribution of Succesful, Writing Evaluation Student Perfromance')
```



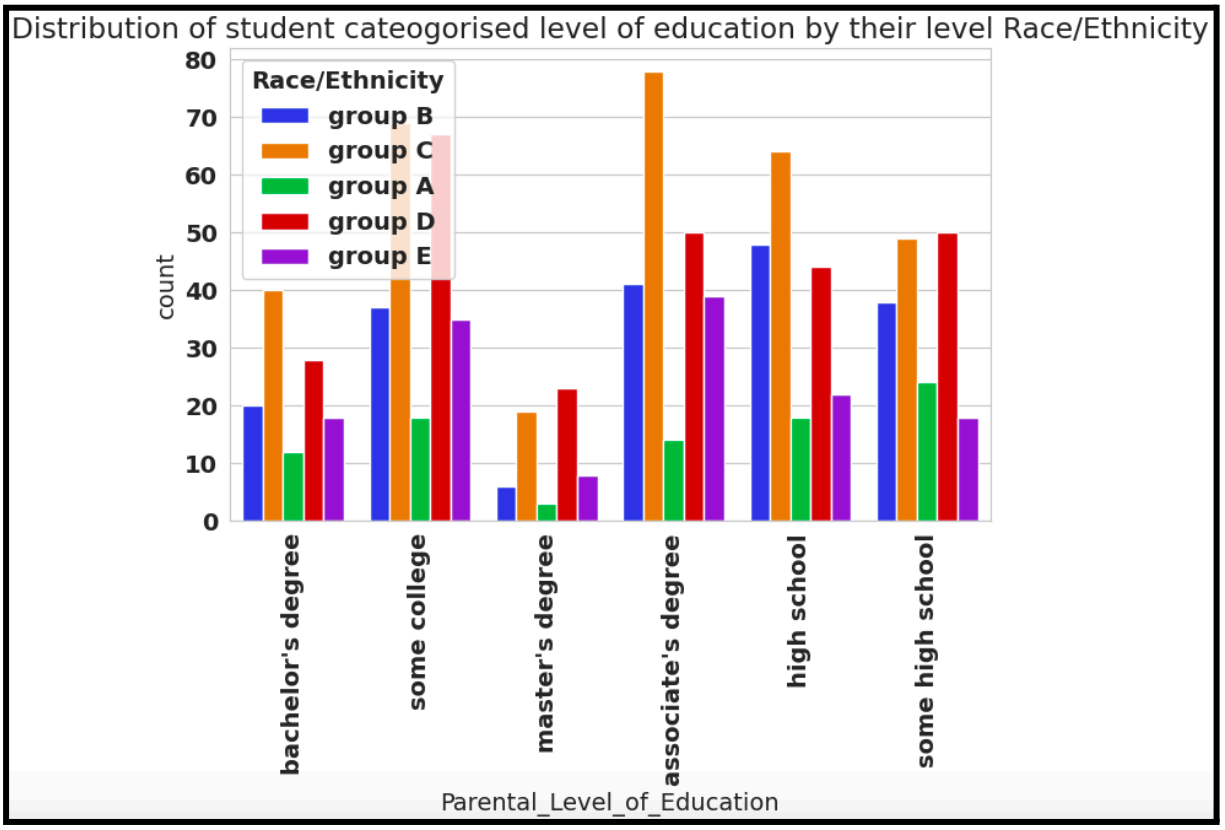
(i)

```
Student_Perf['Overall_StuPer_Stat'] = Student_Perf.apply(lambda x:'Pass' if x['Math_EvalStat']=='Pass' and
                                                         x['Read_EvalStat'] == 'Pass' and x['Writ_EvalStat']=='Pass'
                                                         else 'Fail',  axis = 1 )
Student_Perf['Overall_StuPer_Stat'].value_counts()
```

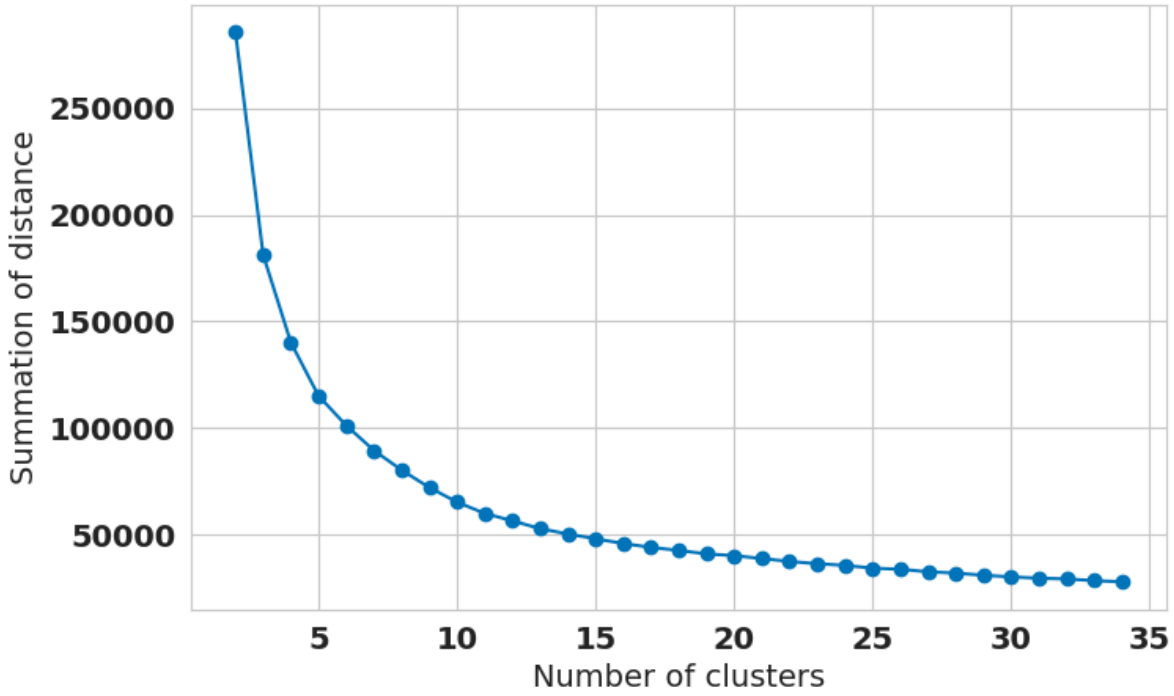
```
Overall_StuPer_Stat
Pass     812
Fail     188
Name: count, dtype: int64
```

15) Create a graph illustrating the student categorised parental distribution of Education by their Race/Ethnicity. (2 mrk)

```
p= sns.countplot(x='Parental_Level_of_Education', data = Student_Perf,hue=('Race/Ethnicity'), palette='bright' )
_ = plt.setp(p.get_xticklabels(), rotation=90)
plt.title('Distribution of student cateogorised level of education by their level Race/Ethnicity')
```



16) Comment on the importance of k-means cluster analysis, its crucial features/ highlights on the graph below and overall application. [3mrk]



Background:

Label Encoding:

Label Encoding permits the execution for algorithms to interpret and comprehend the data by giving each category within a variable a distinct number. It must be noted that variables which reflect distinct categories such as e.g.product types, colours known as categorical variables. Therefore label encoding is suitable for ordinal data as evidently observed with the correlation relationship between Student Test Subject Performance and their Lunch respectively. [Note: Lunch is distinctly categorised into standard , free/reduced also Test Subject Evaluation (Maths, Reading, Writing)] .

Cluster Analysis:

In general Cluster analysis has extensive wide range of applications of various fields including marketing, biology, finance and social sciences:

i) Detect anomalies in financial transactions, ii) Classify social media users into different categories based on their interest and behaviours, iii) Identify genetic markers with specific diseases.

Specifically in regards **K-means Clustering**:

- K-means clustering method proceeds in grouping data points into predetermined number(k) of clusters based on their distance to the centroid of each cluster.

- This specialised iterative algorithm focuses on minimising the sum of squared distances between data points and their assigned cluster centroids. Its evaluation metrics **WCSS** quantifies the sum of squared distance between each data point and the centroid of its respective cluster[: indicates proximity of grouped data point around their cluster centre.

PinPointing the Elbow Point:

- This is the associated k-value optimal number of clusters for your dataset capturing ideal data pattern. This is evident as WCSS is optimal/favourable in relation to pattern trend distribution. We can decipher that the k-value is cautious at the **k=8 cluster** as referred to as Elbow **Point.**

- Beyond this point, adding more clusters may not provide substantial improvements in capturing data patterns, and the WCSS tends to decrease at a gentler rate.

- Bonus Round

Classif_Met	Math_Score	Reading_Score	Writing_Score
0	60.092857	58.057143	56.364286
1	82.328671	81.762238	80.349650
2	32.135135	35.972973	33.594595
3	88.463918	92.793814	92.690722
4	57.926174	67.832215	67.993289
5	73.175182	68.416058	66.802920
6	68.337500	78.062500	77.731250
7	47.759124	51.437956	49.043796