ICFAI UNIVERSITY DEHRADUN

Project Synopsis Report
On
EDA

Submitted in partial fulfillment of the requirements of the degree of

Master of Computer Application

By

Hardik Pratap Singh

For the Session
2022-2024

ACKNOWLEDGEMENT

I Would like to acknowledge all the people who have motivated and help me throughout my dissertation. Firstly, Mr. Mohit Kumar Arya suggested us a unique project topic. Our project guide Mr. Mohit Kumar Arya motivates and helps us for the completion of our project. The members of group work very fluently and honestly that is why our project is completed in given time and successfully. All group members have given their best for successful completion of project. With the help of group members and under guidance of our guide and other teachers also we have done the project as expected and in given time.

Signature:
Mr. Mohit Kumar Arya

ABSTRACT

EDA is a fundamental step in data analysis that aims to understand the characteristics and relationships between variables in a data set. It involves visualizing, summarizing, and exploring data to gain insights and identify patterns. This synopsis highlights the importance of EDA, its main techniques and tools, and how it can be used to improve decision-making processes in various fields such as business, health care, and research. Additionally, it discusses the challenges and limitations of EDA and provides some best practices for effective data exploration.

INDEX

Chapter 1: Introduction ……...1-5

1.1 Introduction of the Project ……

1.2 Problem Statement ……

1.3 Motivation ……..…

1.4 Objectives ……..…

1.5 Scope of the Project Works ……...

1.6 Limitations of Study …….

1.7 Expected outcomes ……...……

Chapter 2: Literature Review …….……6-9

2.1 Theoretical Support …….…….……

2.2 Details regarding work done by various other persons ……

2.3 Method established for EDA ……

2.4 Existing methodologies in EDA ……

2.5 Existing methodologies in EDA ……

2.6 New approached introduced for EDA ……

Chapter 3: Proposed Work …….10-31

3.1 Work carried out ……...

Chapter 4: Development plan/ Schedule/ Flow Chart for Completion of Project …….…….32- 33

4.1 Development plan ……

4.2 Schedule ……...

Chapter 5: Conclusion ……. 34

Chapter 6: References and Bibliography ……... 35-36

CHAPTER 1:

INTRODUCTION

1.1 Introduction of the Project Work

Welcome to the introduction of the project work on EDA! EDA stands for Exploratory Data Analysis, which is an important step in the data analysis process. The goal of EDA is to understand the data and uncover insights by summarizing its main characteristics, such as its distribution, central tendency, variability, and relationships between variables.

In this project, we will be performing EDA on a dataset that we have collected or have been provided with. We will begin by cleaning and preparing the data for analysis, which may include handling missing values, transforming variables, and removing outliers. Then, we will explore the data using various statistical techniques and visualization tools, such as histograms, scatterplots, and correlation matrices.

Our ultimate goal in this project is to gain a better understanding of the data, identify any patterns or trends, and generate hypotheses that can be tested in subsequent analyses. We may also use EDA to identify potential issues or limitations with the data, which could inform decisions about future data collection or analysis.

Throughout the project, we will document our methods and findings in a report or presentation format, which will help us communicate our results effectively to others. We may also use programming languages such as Python or R to perform our analyses and generate visualizations.

Overall, this project will provide us with a valuable opportunity to apply our data analysis skills and gain experience with EDA techniques, which are essential for many fields, including business, health-care, and social sciences.

1.2 Problem Statement

The problem statement for our EDA project will depend on the specific dataset that we are working with. However, in general, the goal of our project will be to gain a better understanding of the data and uncover any insights or patterns that may be useful for decision-making.
In our project, Data on Students Performance in Exams.

Problem statement for Data on Students Performance in Exams:

The purpose of this exploratory data analysis (EDA) project is to analyse and understand the performance of students in exams based on different parameters. The dataset used in this project includes records of students' performance in a particular exam, covering various parameters such as gender, ethnicity, parental education level, test preparation, and subject-wise scores, among others.
The primary objectives of this project are to identify the following:

1. The overall performance of students in the exam and how it varies across different parameters.
2. The relationship between the various parameters and the performance of students, such as the effect of parental education level on the students' scores.
3. The effectiveness of test preparation strategies used by students in improving their scores.
4. The impact of demographic factors such as gender and ethnicity on the performance of students.

By analyzing and visualizing the data, this project aims to provide insights that can help educators and policymakers in developing strategies to improve the overall academic performance of students. Additionally, this project can also help in identifying the factors that hinder students' performance and developing targeted interventions to address them.

1.3 Motivation

The motivation for our EDA project can stem from several reasons. One of the main reasons is that EDA can help us gain insights and make informed decisions based on data. By exploring and understanding the characteristics and patterns of our data, we can make more accurate predictions, identify potential issues or opportunities, and develop strategies that are more likely to succeed.

In addition, EDA is a fundamental step in the data analysis process. It provides us with a better understanding of our data, which can help us select appropriate statistical methods and models for subsequent analysis. Without EDA, we risk making incorrect assumptions about our data or missing important insights that could impact our results.

Furthermore, EDA is a valuable skill that is highly sought after in many industries, including business, health care, and social sciences. By gaining experience with EDA techniques, we can improve our data analysis skills and increase our value as data analysts.

Ultimately, the motivation for our EDA project is to gain a deeper understanding of our data, identify insights that can inform decision-making, and improve our data analysis skills.

1.4 Objectives

The objectives of our EDA project will depend on the specific dataset that we are working with. However, in general, the objectives of our project will be to:

1. Clean and prepare the data for analysis: This will involve handling missing values, transforming variables, and removing outliers to ensure that the data is ready for exploration.
2. Explore the data using various EDA techniques: We will use statistical techniques and visualization tools to identify patterns, trends, and relationships between variables in the data.
3. Identify any issues or limitations with the data: Through EDA, we will identify any potential issues with the data, such as biases, errors, or missing information, that could impact our analysis.

4. Generate hypotheses that can be tested in subsequent analyses: Based on our EDA, we will develop hypotheses about the relationships between variables in the data, which can be tested using more advanced statistical methods.
5. Document our methods and findings: We will record our EDA methods and findings in a report or presentation format, which will help us communicate our results effectively to others.
6. Apply programming skills: We may use programming languages such as Python or R to perform our analyses and generate visualizations.
Our objectives for this EDA project are to gain a better understanding of our data, identify any issues or limitations, and generate insights that can inform decision-making. By achieving these objectives, we can develop a deeper understanding of the data and make more informed decisions based on our analysis.

1.5 Scope of the Project Works

The scope of our project works for EDA will include the following:
1. Data Collection: We will collect the data from various sources, which could include public datasets or data provided by an organization.
2. Data Cleaning and Preparation: We will clean and prepare the data for analysis, which may include handling missing values, transforming variables, and removing outliers.
3. Exploratory Data Analysis: We will explore the data using various statistical techniques and visualization tools to identify patterns, trends, and relationships between variables in the data.
4. Identification of issues and limitations: We will identify any potential issues with the data, such as biases, errors, or missing information, that could impact our analysis.
5. Hypothesis generation: Based on our EDA, we will develop hypotheses about the relationships between variables in the data, which can be tested using more advanced statistical methods.
6. Communication of results: We will record our EDA methods and findings in a report or presentation format, which will help us communicate our results effectively to others.
7. Programming skills: We may use programming languages such as Python or R

to perform our analyses and generate visualizations.

The scope of our project works will be focused on performing a thorough EDA of the data and generating insights that can inform decision-making. The scope may be expanded or narrowed depending on the specific objectives of the analysis and the complexity of the dataset.

1.6 Limitations of study

There are several limitations of our EDA study that we should consider:

1. Sampling Bias: If the data was collected using a biased sampling technique, the results of our EDA may not accurately represent the population that the data is meant to describe.

2. Data Quality: The accuracy and completeness of the data may be limited by measurement errors, missing values, or inconsistencies, which can affect the reliability of our EDA.

3. Scope of Analysis: The scope of our EDA may be limited by the available data or the time constraints of the project. This may prevent us from exploring all relevant variables or identifying more complex relationships between them.

4. Data Privacy: Depending on the nature of the data, there may be privacy concerns that limit the level of detail that we can include in our EDA or prevent us from sharing our findings publicly.

5. Statistical Significance: While EDA can provide valuable insights into the data, the relationships and patterns that we identify may not be statistically significant, and further testing may be required to confirm our findings.

It is important to be aware of these limitations when interpreting the results of our EDA and to communicate these limitations clearly in our report or presentation. By acknowledging these limitations, we can avoid making inappropriate conclusions and provide a more accurate assessment of the data.

1.7 Expected outcomes

The expected outcomes of our EDA project will depend on the specific dataset that

we are working with and the objectives of our analysis. However, in general, we can expect to achieve the following outcomes:

1. Gain a better understanding of the data: Through EDA, we will gain insights into the distribution of variables, the relationships between variables, and the presence of any outliers or missing data. This understanding will help us make more informed decisions based on the data.

2. Identify potential issues with the data: EDA will help us identify any potential issues with the data, such as biases, errors, or missing information, which can impact our analysis. By identifying these issues, we can take steps to mitigate their effects and improve the accuracy of our analysis.

3. Generate hypotheses for further testing: Based on our EDA, we may develop hypotheses about the relationships between variables in the data, which can be tested using more advanced statistical methods.

4. Communicate results effectively: We will record our EDA methods and findings in a report or presentation format, which will help us communicate our results effectively to others. This communication will help ensure that our insights are understood and can be acted upon.

5. Improve decision-making: By gaining a deeper understanding of the data and identifying potential issues, we can make more informed decisions based on our analysis. This can lead to improved outcomes and better decision-making in various fields, including business, health care, and policy-making.

The expected outcomes of our EDA project are to gain a better understanding of the data, identify any potential issues, generate insights that can inform decision-making, and communicate our findings effectively to others. By achieving these outcomes, we can make better use of the data and achieve more impactful results.

CHAPTER 2:

LITERATURE REVIEW

2.1 Theoretical Support

Theoretical support for an EDA project may depend on the specific dataset and objectives of the analysis. However, in general, EDA is based on a few fundamental principles from statistical theory, such as:

1. Descriptive statistics: EDA relies heavily on descriptive statistics to summarize and visualize data. Descriptive statistics include measures of central tendency, such as the mean and median, and measures of variability, such as the standard deviation and range.

2. Data distribution: EDA aims to explore the distribution of data, which is important for identifying any patterns or outliers that may be present. Theoretical support for data distribution can come from probability theory, which provides a framework for understanding the behaviour of random variables and the likelihood of certain outcomes.

3. Hypothesis testing: EDA can generate hypotheses about the relationships between variables in the data, which can be tested using more advanced statistical methods. Theoretical support for hypothesis testing can come from statistical inference, which provides a framework for making conclusions about a population based on a sample of data.

4. Data visualization: EDA often involves the use of data visualization techniques, such as histograms, scatter plots, and box plots. Theoretical support for data visualization can come from the field of data visualization, which focuses on the effective communication of information through graphical means.

Theoretical support for an EDA project may draw on various areas of statistical theory,

including descriptive statistics, probability theory, statistical inference, and data visualization. By using these principles, we can gain a deeper understanding of the data, identify patterns and outliers, and generate hypotheses for further testing.

2.2 Details regarding work done by various other persons

Peoples who had major role in development of EDA:
John Tukey: John Tukey, an American mathematician, is credited with coining the term "Exploratory Data Analysis" in his book of the same name. Tukey's work on EDA emphasized the importance of visualization and graphical methods for exploring data, as well as the need to be open to unexpected findings
William Cleveland: William Cleveland is an American statistician who is known for his work on graphical methods for data analysis. He developed the Cleveland dot plot, which is a variation of the scatter plot that emphasizes the distribution of the data, and also introduced the concept of brushing and linking, which allows users to interactively explore multiple views of the same data
Ben Shneiderman: Ben Shneiderman is an American computer scientist who is known for his work on information visualization and human-computer interaction. He developed the concept of "overview and detail," which allows users to interactively explore data at multiple levels of detail, and also introduced the concept of "tree maps" which use nested rectangles to represent hierarchical data.

2.3 Methods Established In EDA

Exploratory Data Analysis (EDA) involves a variety of methods and techniques for exploring and understanding a dataset. Following are the few of them.

- **Descriptive statistics:** EDA often starts with calculating and summarizing descriptive statistics, such as mean, median, mode, range, standard deviation, and variance. These statistics help in understanding the central tendency, variability, and distribution of the data.
- **Data visualization:** EDA heavily relies on data visualization techniques to explore the patterns and relationships in the data. Some commonly used visualization methods include histograms, box plots, scatter plots, heat maps, and bar charts.

Visualization helps in identifying outliers, trends, clusters, and other patterns in the data.

- **Correlation analysis:** Correlation analysis helps in identifying the strength and direction of the relationship between two variables. Commonly used methods for correlation analysis include Pearson correlation coefficient, Spearman rank correlation, and Kendall tau correlation.
- **Dimensional reduction:** EDA often deals with high-dimensional data, which can be difficult to visualize and analyse. Dimensional reduction techniques, such as Principal Component Analysis (PCA) and t-SNE, can help in reducing the dimensional of the data while retaining most of the variation and structure.
- **Clustering:** Clustering helps in identifying groups or clusters of similar data points in the dataset. Commonly used clustering techniques include k-means clustering, hierarchical clustering, and DBSCAN.
- **Association rules:** Association rule mining helps in identifying the frequent patterns and associations between different variables in the dataset. Commonly used methods for association rule mining include Apriori and FP-growth algorithms.

2.4 Existing Methodologies For EDA

- **Tukey's Exploratory Data Analysis:** This methodology, proposed by John Tukey, involves a cyclical process of graphical and numerical analysis. The steps include calculating summary statistics, creating plots to visualize the data, identifying, and dealing with outliers, assessing the distribution of the data, and making conclusions based on the analysis.
- **Cleveland's Data Analysis Pipeline:** William Cleveland proposed a methodology that involves a series of steps to transform, explore, and model the data. The steps include data cleaning, transforming, and mapping, exploratory data analysis, model building, and model validation and refinement.
- **CRISP-DM:** CRISP-DM (Cross-Industry Standard Process for Data Mining)is a widely used methodology for data mining, which includes EDA as one of its stages. The steps in CRISP-DM include business understanding, data understanding, data preparation, modelling, evaluation, and deployment.
- **Wickham's Tidy Data and Tidyverse:** Hadley Wickham's methodology

emphasizes the importance of structuring the data in a tidy format, where each variable has its own column and each observation has its own row. The Tidyverse collection of R packages provides a suite of tools for transforming and visualizing tidy data.

- **Agile Data Science:** Agile Data Science is a methodology that combines agile software development practices with data science. The methodology involves iterative and incremental development, with an emphasis on collaboration, feedback, and continuous improvement.

2.5 Existing Algorithms For EDA

There are several algorithms that are commonly used in Exploratory Data Analysis (EDA) to identify patterns, relationships, and anomalies in the data.

- **Clustering algorithms:** Clustering algorithms such as k-means clustering, hierarchical clustering, and DBSCAN are commonly used in EDA to group similar data points together based on their characteristics. Clustering can help identify patterns and similarities in the data and can be useful for segmenting customers, grouping products, or identifying outliers.
- **Principal Component Analysis (PCA):** PCA is a dimensionality reduction algorithm that is used to identify the most important variables in the data. PCA can help reduce the dimensionality of the data and can be used to identify the variables that are most strongly correlated with each other.
- **Association Rule Mining:** Association rule mining algorithms such as Apriori algorithm and FP-growth algorithm are used to identify frequent patterns and associations between variables in the data. These algorithms can help identify which variables tend to occur together and can be useful for market basket analysis or identifying customer preferences.
- **Anomaly Detection Algorithms:** Anomaly detection algorithms such as Isolation Forest and Local Outlier Factor are used to identify outliers or anomalies in the data. These algorithms can be useful for detecting fraud, errors, or other unusual behaviour in the data.
- **Regression Analysis:** Regression analysis is a statistical technique that is use to

identify the relationship between variables in the data. Regression can be used to identify which variables are most strongly correlated with the outcome variable and can be useful for predicting future outcomes.

- **Decision Trees:** Decision trees are a type of machine learning algorithm that can be used for classification and regression analysis. Decision trees can be used to identify which variables are most important in predicting the outcome variable and can be useful for identifying customer segments or predicting customer behaviour.

2.6 New Approaches introduced for EDA

EDA is a constantly evolving field, and new approaches are being introduced all the time. Here are some of the recent approaches that have been introduced for EDA:

1. Machine learning-based EDA: Machine learning algorithms can be used to identify patterns and relationships within the data, and to automate the process of EDA. Recent studies have shown that machine learning-based EDA can be more effective and efficient than traditional EDA methods.

2. Interactive EDA: Interactive EDA allows users to manipulate and explore the data in real-time using interactive tools and interfaces. This approach can help identify patterns and relationships more quickly and effectively than traditional EDA methods.

3. Network analysis: Network analysis can be used to explore the relationships between variables in the data, and to identify clusters and subgroups. This approach is particularly useful for exploring complex, interconnected datasets.

4. Deep learning-based EDA: Deep learning algorithms can be used to identify patterns and relationships within the data, and to generate new insights and hypotheses. Recent studies have shown that deep learning-based EDA can be more effective and efficient than traditional EDA methods.

5. Big data EDA: With the rise of big data, new approaches are being developed for EDA that can handle large, complex datasets. These approaches often involve distributed computing and parallel processing techniques to speed up the analysis.

CHAPTER 3:

PROPOSED WORK

3.1 Work carried out

In order to get more familiar with EDA, in this project we have carried out EDA on two separate datasets.
1. Data on Criminal Activities in India (2019 – 2021)
(https://ncrb.gov.in/en/crime-india)

2. Data on Students Performance in Exams.
With the help of different methodologies mention in previous chapter, we'll perform EDA on the above datasets.

Resources used:

To collect Data:
Kaggle

Programming Language:
Python

IDE:

Jupyter Notebook

Code Implementation:

Data on Students Performance in Exams

Importing libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Import the CSV Data as Pandas DataFrame

In [2]:
```
df = pd.read_csv('StudentsPerformance.csv')
```

Show Top 5 Records

In [3]:
```
df.head(5)
```

Out[3]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Shape of the dataset

In [4]:
```
df.shape
```

Out[4]:
```
(1000, 8)
```

Dataset information

- gender : sex of students -> (Male/female)
- race/ethnicity : ethnicity of students -> (Group A, B,C, D,E)
- parental level of education : parents' final education ->(bachelor's degree,some college,master's degree,associate's degree,high school)
- lunch : having lunch before test (standard or free/reduced)
- test preparation course : complete or not complete before test
- math score
- reading score
- writing score

In [5]:
```
df.info
```

Out[5]:
```
<bound method DataFrame.info of      gender race/ethnicity parental level of education
lunch  \
0    female        group B           bachelor's degree      standard
1    female        group C                some college      standard
2    female        group B             master's degree      standard
3      male        group A          associate's degree  free/reduced
4      male        group C                some college      standard
..      ...            ...                         ...           ...
995  female        group E             master's degree      standard
996    male        group C                 high school  free/reduced
997  female        group C                 high school  free/reduced
998  female        group D                some college      standard
999  female        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      ...         ...            ...            ...
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]>
```

Data Checks to perform

- Check Missing values
- Check Duplicates
- Check data type
- Check the number of unique values of each column
- Check statistics of data set
- Check various categories present in the different categorical column

Check Missing values

In [6]:
```
df.isna().sum()
```

Out[6]:
```
gender                       0
race/ethnicity               0
parental level of education  0
lunch                        0
test preparation course      0
math score                   0
reading score                0
writing score                0
dtype: int64
```

There are no missing values in the data set

Check Duplicates

In [7]:
```
df.duplicated().sum()
```

Out[7]:
```
0
```

There are no duplicates values in the data set

Check data types

Check Null and Dtypes

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64 
 6   reading score                1000 non-null   int64 
 7   writing score                1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Checking the number of unique values of each column

In [9]:
```
df.nunique()
```

Out[9]:
```
gender                         2
race/ethnicity                 5
parental level of education    6
lunch                          2
test preparation course        2
math score                    81
reading score                 72
writing score                 77
dtype: int64
```

Check statistics of data set

In [10]:
```
df.describe()
```

Out[10]:

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

Insight

- From above description of numerical data, all means are very close to each other - between 66 and 68.05;
- All standard deviations are also close - between 14.6 and 15.19;
- While there is a minimum score 0 for math, for writing minimum is much higher = 10 and for reading higher = 17

Exploring Data

In [11]:
```
df.head()
```

Out[11]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

In [12]:
```
print("Categories in 'gender' variable:     ",end=" " )
print(df['gender'].unique())

print("Categories in 'race/ethnicity' variable:  ",end=" ")
print(df['race/ethnicity'].unique())

print("Categories in'parental level of education' variable:",end=" " )
print(df['parental level of education'].unique())

print("Categories in 'lunch' variable:     ",end=" " )
print(df['lunch'].unique())

print("Categories in 'test preparation course' variable:     ",end=" " )
print(df['test preparation course'].unique())
```

```
Categories in 'gender' variable:      ['female' 'male']
Categories in 'race/ethnicity' variable:   ['group B' 'group C' 'group A' 'group D' 'gro
up E']
Categories in'parental level of education' variable: ["bachelor's degree" 'some college'
"master's degree" "associate's degree"
 'high school' 'some high school']
Categories in 'lunch' variable:      ['standard' 'free/reduced']
Categories in 'test preparation course' variable:      ['none' 'completed']
```

define numerical & categorical columns

In [13]:
```
numeric_features = [feature for feature in df.columns if df[feature].dtype != 'O']
categorical_features = [feature for feature in df.columns if df[feature].dtype == 'O']
```

print columns

In [14]:
```
print('We have {} numerical features : {}'.format(len(numeric_features), numeric_features
print('\nWe have {} categorical features : {}'.format(len(categorical_features), categori
```

```
We have 3 numerical features : ['math score', 'reading score', 'writing score']

We have 5 categorical features : ['gender', 'race/ethnicity', 'parental level of educati
on', 'lunch', 'test preparation course']
```

Adding columns for "Total Score" and "Average"

In [15]:
```
df['total score'] = df['math score'] + df['reading score'] + df['writing score']
df['average'] = df['total score']/3
df.head()
```

Out[15]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	total score	average
0	female	group B	bachelor's degree	standard	none	72	72	74	218	72.666667
1	female	group C	some college	standard	completed	69	90	88	247	82.333333
2	female	group B	master's degree	standard	none	90	95	93	278	92.666667
3	male	group A	associate's degree	free/reduced	none	47	57	44	148	49.333333
4	male	group C	some college	standard	none	76	78	75	229	76.333333

In [16]:
```
reading_full = df[df['reading score'] == 100]['average'].count()
writing_full = df[df['writing score'] == 100]['average'].count()
math_full = df[df['math score'] == 100]['average'].count()
print(f'Number of students with full marks in Maths: {math_full}')
print(f'Number of students with full marks in Writing: {writing_full}')
print(f'Number of students with full marks in Reading: {reading_full}')
```

```
Number of students with full marks in Maths: 7
Number of students with full marks in Writing: 14
Number of students with full marks in Reading: 17
```

In [17]:
```
reading_less_20 = df[df['reading score'] <= 20]['average'].count()
writing_less_20 = df[df['writing score'] <= 20]['average'].count()
math_less_20 = df[df['math score'] <= 20]['average'].count()
print(f'Number of students with less than 20 marks in Maths: {math_less_20}')
print(f'Number of students with less than 20 marks in Writing: {writing_less_20}')
print(f'Number of students with less than 20 marks in Reading: {reading_less_20}')
```

```
Number of students with less than 20 marks in Maths: 4
Number of students with less than 20 marks in Writing: 3
Number of students with less than 20 marks in Reading: 1
```

Insights

- From above values we get students have performed the worst in Maths
- Best performance is in reading section

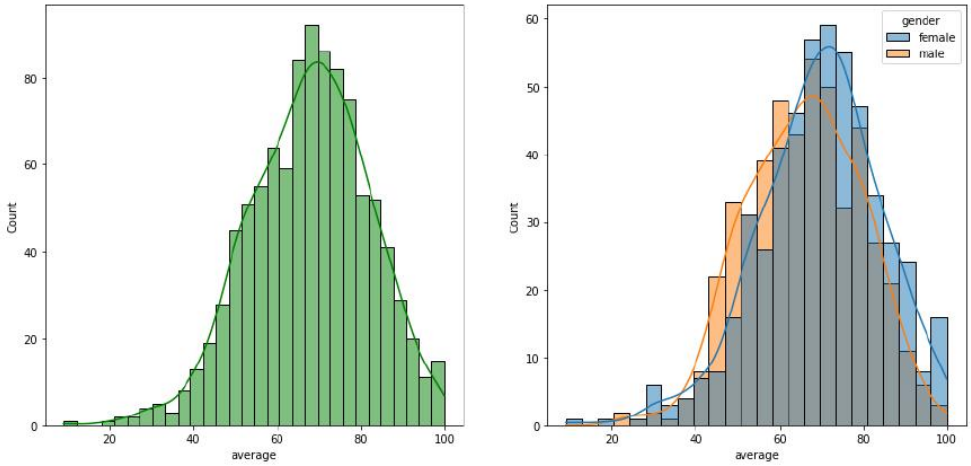
Exploring Data (Visualization)

Visualize average score distribution to make some conclusion

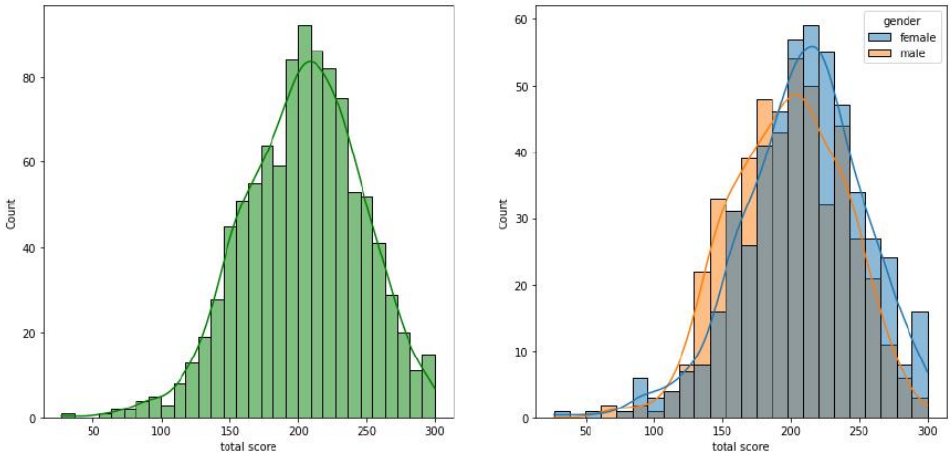
- Histogram
- Kernel Distribution Function (KDE)

Histogram & KDE

In [18]:
```
fig, axs = plt.subplots(1, 2, figsize=(15, 7))
plt.subplot(121)
sns.histplot(data=df,x='average',bins=30,kde=True,color='g')
plt.subplot(122)
sns.histplot(data=df,x='average',kde=True,hue='gender')
plt.show()
```



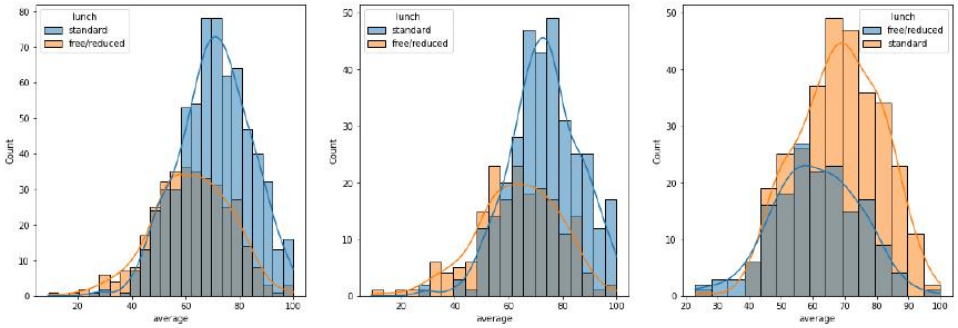
In [19]:
```
fig, axs = plt.subplots(1, 2, figsize=(15, 7))
plt.subplot(121)
sns.histplot(data=df,x='total score',bins=30,kde=True,color='g')
plt.subplot(122)
sns.histplot(data=df,x='total score',kde=True,hue='gender')
plt.show()
```



Insights

- Female students tend to perform well then male students

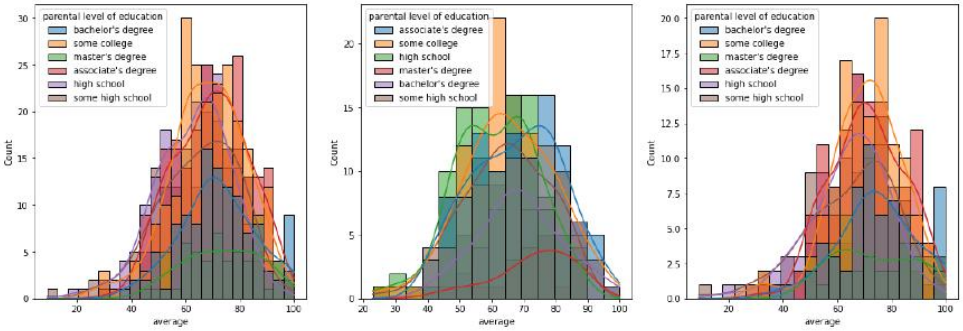
In [20]:
```
plt.subplots(1,3,figsize=(25,6))
plt.subplot(141)
sns.histplot(data=df,x='average',kde=True,hue='lunch')
plt.subplot(142)
sns.histplot(data=df[df.gender=='female'],x='average',kde=True,hue='lunch')
plt.subplot(143)
sns.histplot(data=df[df.gender=='male'],x='average',kde=True,hue='lunch')
plt.show()
```



Insights

- Standard lunch helps perform well in exams.
- Standard lunch helps perform well in exams be it a male or a female.

In [21]:
```
plt.subplots(1,3,figsize=(25,6))
plt.subplot(141)
ax =sns.histplot(data=df,x='average',kde=True,hue='parental level of education')
plt.subplot(142)
ax =sns.histplot(data=df[df.gender=='male'],x='average',kde=True,hue='parental level of e
plt.subplot(143)
ax =sns.histplot(data=df[df.gender=='female'],x='average',kde=True,hue='parental level of
plt.show()
```

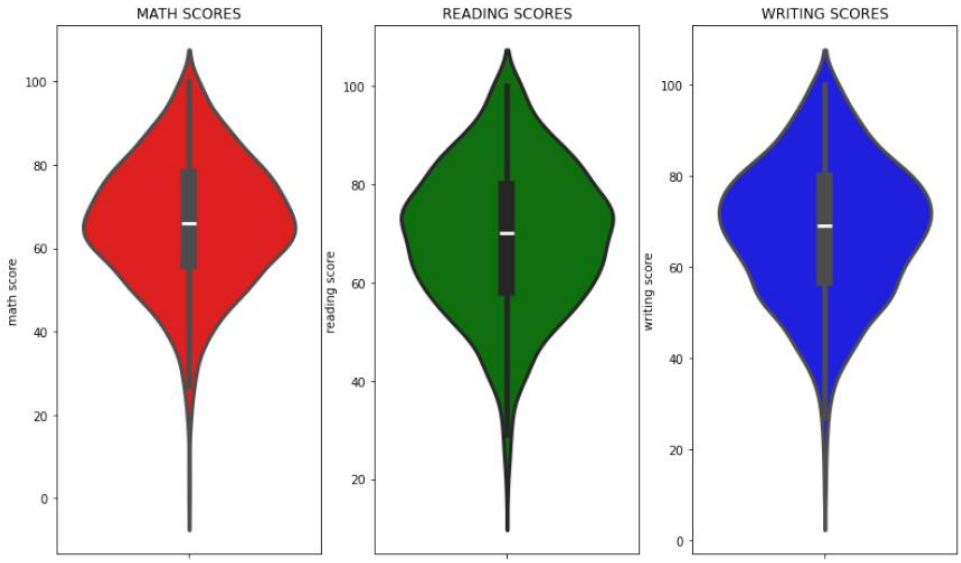


Insights

- Students of group A and group B tends to perform poorly in exam.
- Students of group A and group B tends to perform poorly in exam irrespective of whether they are male or female

Maximumum score of students in all three subjects

In [22]:
```
plt.figure(figsize=(18,8))
plt.subplot(1, 4, 1)
plt.title('MATH SCORES')
sns.violinplot(y='math score',data=df,color='red',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('READING SCORES')
sns.violinplot(y='reading score',data=df,color='green',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('WRITING SCORES')
sns.violinplot(y='writing score',data=df,color='blue',linewidth=3)
plt.show()
```



Insights

• From the above three plots its clearly visible that most of the students score inbetween 60-80 in Maths whereas in reading and writing most of them score from 50-80

Multivariate analysis using pieplot

In [23]:
```
plt.rcParams['figure.figsize'] = (30, 12)

plt.subplot(1, 5, 1)
size = df['gender'].value_counts()
labels = 'Female', 'Male'
color = ['red','green']

plt.pie(size, colors = color, labels = labels,autopct = '.%2f%%')
plt.title('Gender', fontsize = 20)
plt.axis('off')

plt.subplot(1, 5, 2)
size = df['race/ethnicity'].value_counts()
labels = 'Group C', 'Group D','Group B','Group E','Group A'
color = ['red', 'green', 'blue', 'cyan','orange']

plt.pie(size, colors = color,labels = labels,autopct = '.%2f%%')
plt.title('Race_Ethnicity', fontsize = 20)
plt.axis('off')

plt.subplot(1, 5, 3)
size = df['lunch'].value_counts()
labels = 'Standard', 'Free'
color = ['red','green']

plt.pie(size, colors = color,labels = labels,autopct = '.%2f%%')
plt.title('Lunch', fontsize = 20)
plt.axis('off')

plt.subplot(1, 5, 4)
size = df['test preparation course'].value_counts()
labels = 'None', 'Completed'
color = ['red','green']

plt.pie(size, colors = color,labels = labels,autopct = '.%2f%%')
plt.title('Test Course', fontsize = 20)
plt.axis('off')
```

```
plt.subplot(1, 5, 5)
size = df['parental level of education'].value_counts()
labels = 'Some College', "Associate's Degree",'High School','Some High School',"Bachelor'
color = ['red', 'green', 'blue', 'cyan','orange','grey']

plt.pie(size, colors = color,labels = labels,autopct = '.%2f%%')
plt.title('Parental Education', fontsize = 20)
plt.axis('off')

plt.tight_layout()
plt.grid()

plt.show()
```

Insights

- Number of Male and Female students is almost equal
- Number students are greatest in Group C
- Number of students who have standard lunch are greater

- Number of students who have not enrolled in any test preparation course is greater
- Number of students whose parental education is "Some College" is greater followed closely by "Associate's Degree"

Feature Wise Visualization

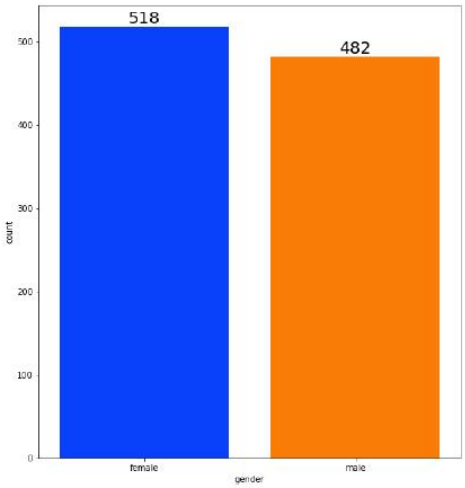
GENDER COLUMN

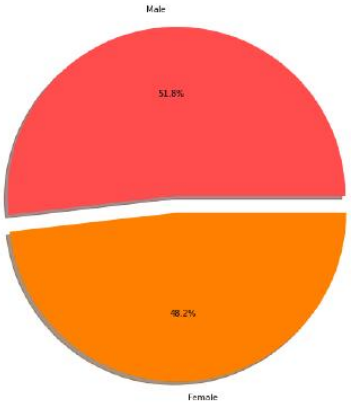
- How is distribution of Gender ?
- Is gender has any impact on student's performance ?

UNIVARIATE ANALYSIS (How is distribution of Gender ?)

In [24]:
```
f,ax=plt.subplots(1,2,figsize=(20,10))
sns.countplot(x=df['gender'],data=df,palette ='bright',ax=ax[0],saturation=0.95)
for container in ax[0].containers:
    ax[0].bar_label(container,color='black',size=20)
plt.pie(x=df['gender'].value_counts(),labels=['Male','Female'],explode=[0,0.1],autopct='%
plt.show()
```

```
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\3407978415.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x=df['gender'],data=df,palette ='bright',ax=ax[0],saturation=0.95)
```





Insights

- Gender has balanced data with female students are 518 (48%) and male students are 482 (52%)

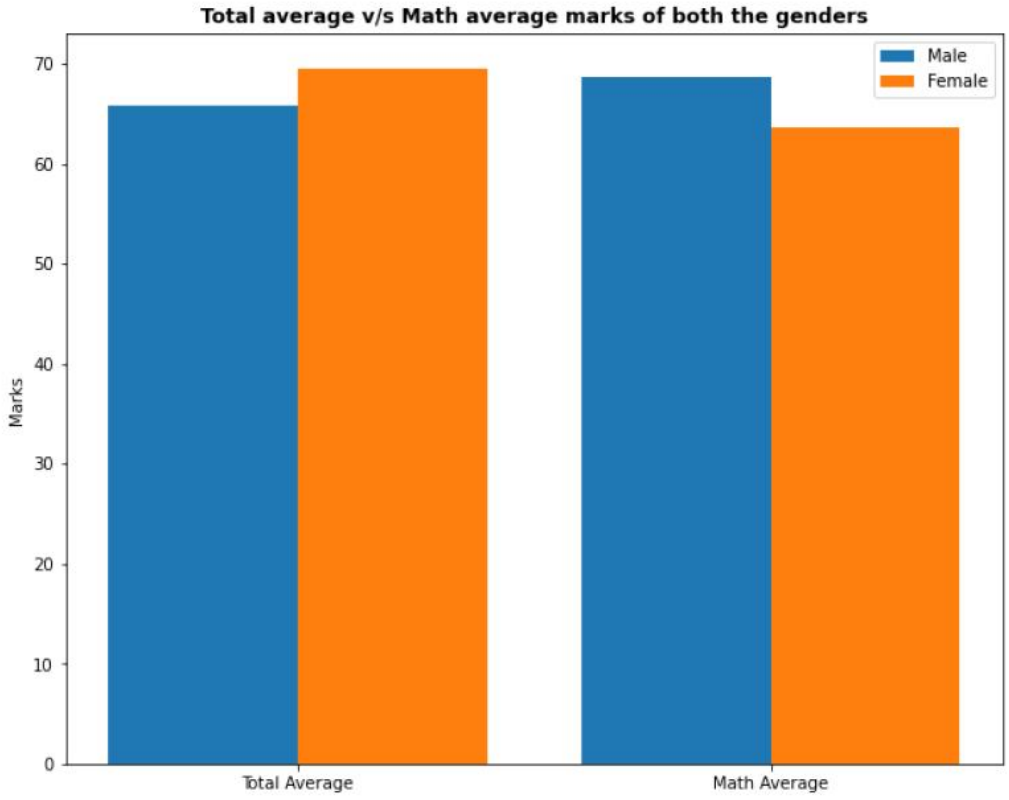
BIVARIATE ANALYSIS (Is gender has any impact on student's performance ?)

In [25]:
```
gender_group = df.groupby('gender').mean()
gender_group
```

Out[25]:

gender	math score	reading score	writing score	total score	average
female	63.633205	72.608108	72.467181	208.708494	69.569498
male	68.728216	65.473029	63.311203	197.512448	65.837483

In [26]:
```
plt.figure(figsize=(10, 8))

X = ['Total Average','Math Average']

female_scores = [gender_group['average'][0], gender_group['math score'][0]]
male_scores = [gender_group['average'][1], gender_group['math score'][1]]

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_scores, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_scores, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.ylabel("Marks")
plt.title("Total average v/s Math average marks of both the genders",fontweight='bold')
plt.legend()
plt.show()
```



Insights

- On an average females have a better overall score than men.
- whereas males have scored higher in Maths.

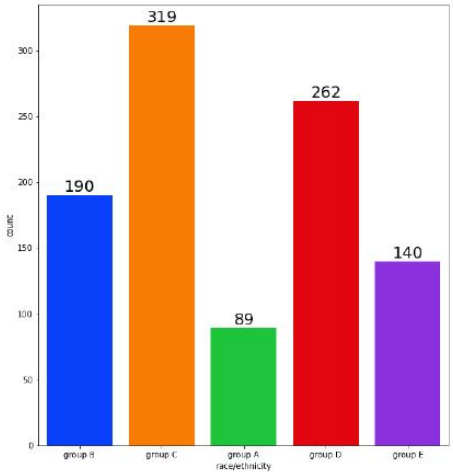
RACE/EHNICITY COLUMN

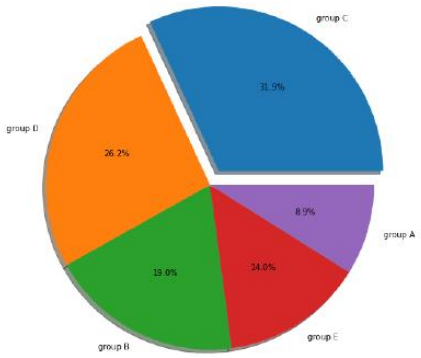
- How is Group wise distribution ?
- Is Race/Ehnicity has any impact on student's performance ?

UNIVARIATE ANALYSIS (How is Group wise distribution ?)

In [27]:
```
f,ax=plt.subplots(1,2,figsize=(20,10))
sns.countplot(x=df['race/ethnicity'],data=df,palette = 'bright',ax=ax[0],saturation=0.95)
for container in ax[0].containers:
    ax[0].bar_label(container,color='black',size=20)
plt.pie(x = df['race/ethnicity'].value_counts(),labels=df['race/ethnicity'].value_counts(
plt.show()
```

```
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\667979199.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x=df['race/ethnicity'],data=df,palette = 'bright',ax=ax[0],saturation=0.
95)
```





Insights

- Most of the student belonging from group C /group D.
- Lowest number of students belong to groupA.

BIVARIATE ANALYSIS (Is Race/Ehnicity has any impact on student's performance ?)

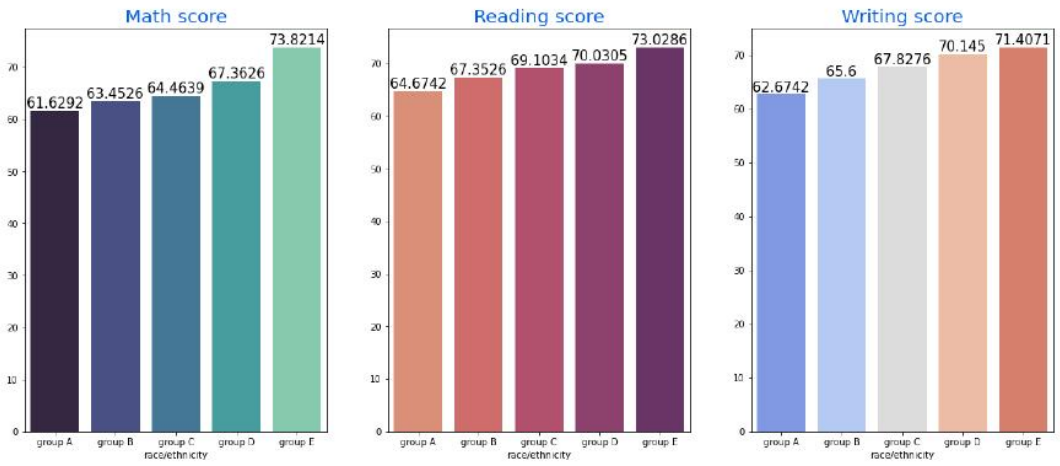
Insights

- Most of the student belonging from group C /group D.
- Lowest number of students belong to groupA.

BIVARIATE ANALYSIS (Is Race/Ehnicity has any impact on student's performance ?)

In [28]:
```
Group_data2=df.groupby('race/ethnicity')
f,ax=plt.subplots(1,3,figsize=(20,8))
sns.barplot(x=Group_data2['math score'].mean().index,y=Group_data2['math score'].mean().v
ax[0].set_title('Math score',color='#005ce6',size=20)

for container in ax[0].containers:
    ax[0].bar_label(container,color='black',size=15)

sns.barplot(x=Group_data2['reading score'].mean().index,y=Group_data2['reading score'].me
ax[1].set_title('Reading score',color='#005ce6',size=20)

for container in ax[1].containers:
    ax[1].bar_label(container,color='black',size=15)

sns.barplot(x=Group_data2['writing score'].mean().index,y=Group_data2['writing score'].me
ax[2].set_title('Writing score',color='#005ce6',size=20)

for container in ax[2].containers:
    ax[2].bar_label(container,color='black',size=15)
```

```
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\1485709615.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=Group_data2['math score'].mean().index,y=Group_data2['math score'].mean
().values,palette = 'mako',ax=ax[0])
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\1485709615.py:9: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=Group_data2['reading score'].mean().index,y=Group_data2['reading scor
e'].mean().values,palette = 'flare',ax=ax[1])
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\1485709615.py:15: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=Group_data2['writing score'].mean().index,y=Group_data2['writing scor
e'].mean().values,palette = 'coolwarm',ax=ax[2])
```



Insights

- Group E students have scored the highest marks.
- Group A students have scored the lowest marks.

- Students from a lower Socioeconomic status have a lower avg in all course subjects

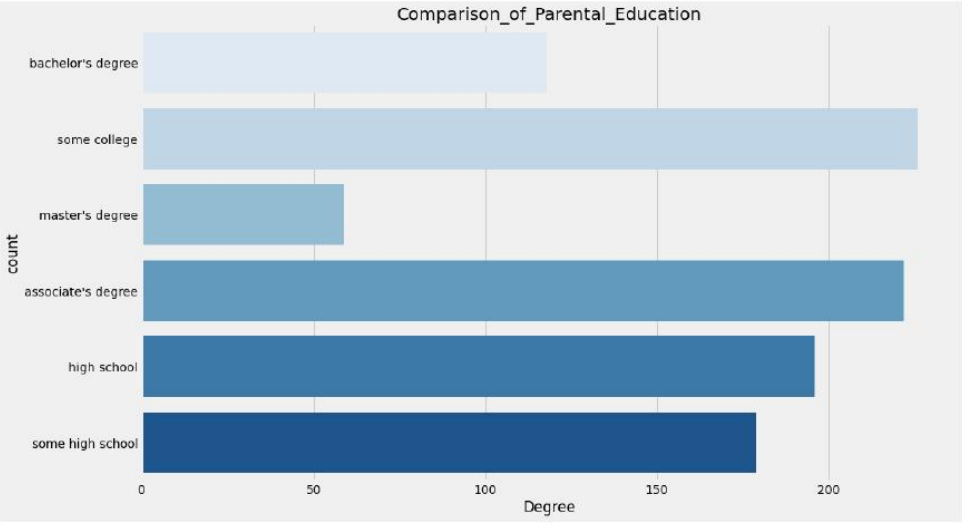
PARENTAL LEVEL OF EDUCATION COLUMN

- What is educational background of student's parent ?
- Is parental education has any impact on student's performance ?

UNIVARIATE ANALYSIS (What is educational background of student's parent ?)

In [29]:
```
plt.rcParams['figure.figsize'] = (15, 9)
plt.style.use('fivethirtyeight')
sns.countplot(df['parental level of education'], palette = 'Blues')
plt.title('Comparison_of_Parental_Education', fontweight = 30, fontsize = 20)
plt.xlabel('Degree')
plt.ylabel('count')
plt.show()
```

```
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\3632211931.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(df['parental level of education'], palette = 'Blues')
```

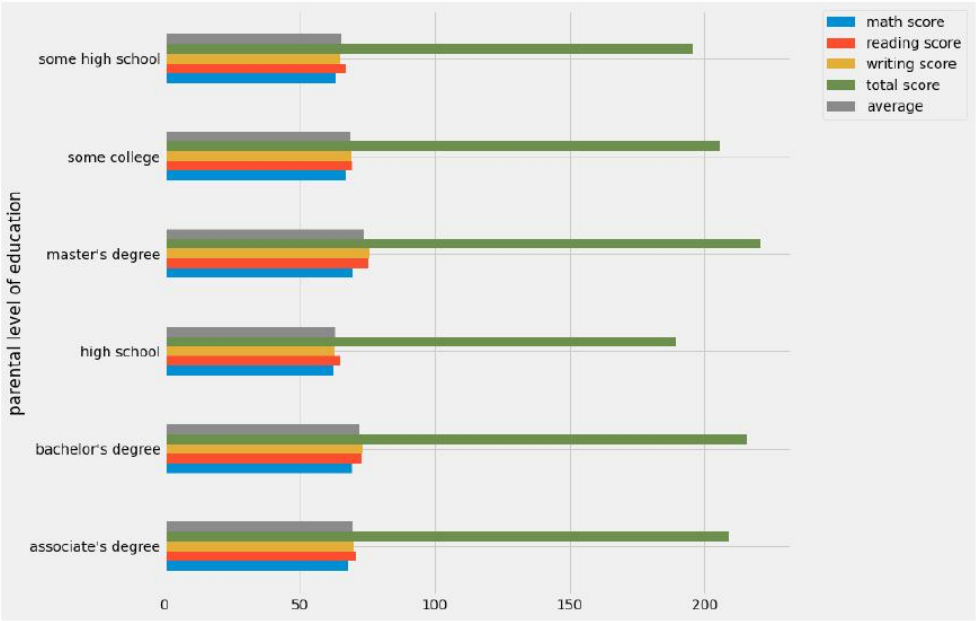


Insights

- Largest number of parents are from some college.

BIVARIATE ANALYSIS (Is parental education has any impact on student's performance ?)

In [30]:
```
df.groupby('parental level of education').agg('mean').plot(kind='barh',figsize=(10,10))
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```



Insights

• The score of student whose parents possess master and bachelor level education are higher than others

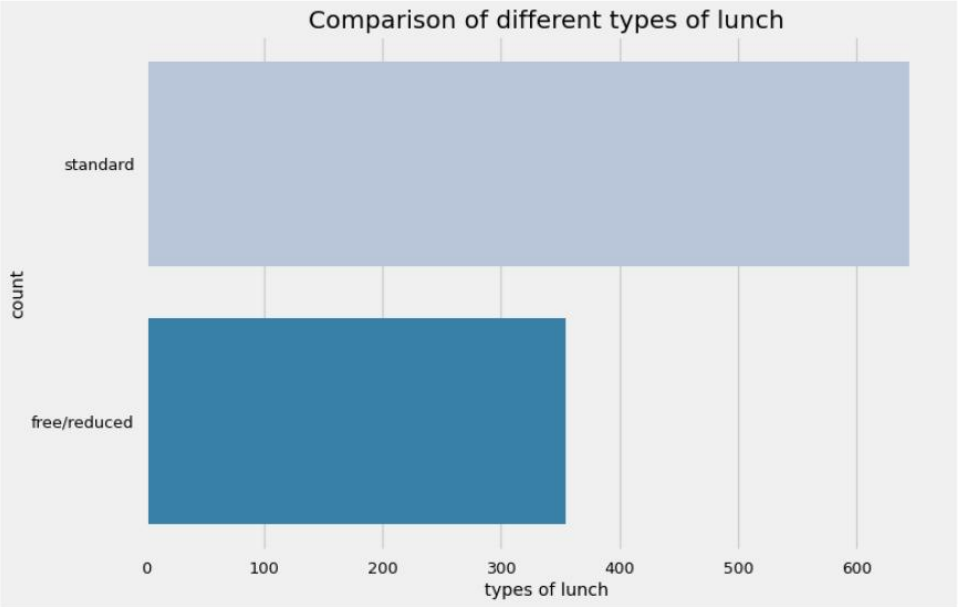
Lunch Colomn

- Which type of lunch is most common amoung students ?
- What is the effect of lunch type on test results?

UNIVARIATE ANALYSIS (Which type of lunch is most common amoung students ?)

In [31]:
```
plt.rcParams['figure.figsize'] = (15, 9)
plt.style.use('seaborn-talk')
sns.countplot(df['lunch'], palette = 'PuBu')
plt.title('Comparison of different types of lunch', fontweight = 30, fontsize = 20)
plt.xlabel('types of lunch')
plt.ylabel('count')
plt.show()
```

```
C:\Users\hardi\AppData\Local\Temp\ipykernel_29684\3431367200.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(df['lunch'], palette = 'PuBu')
```

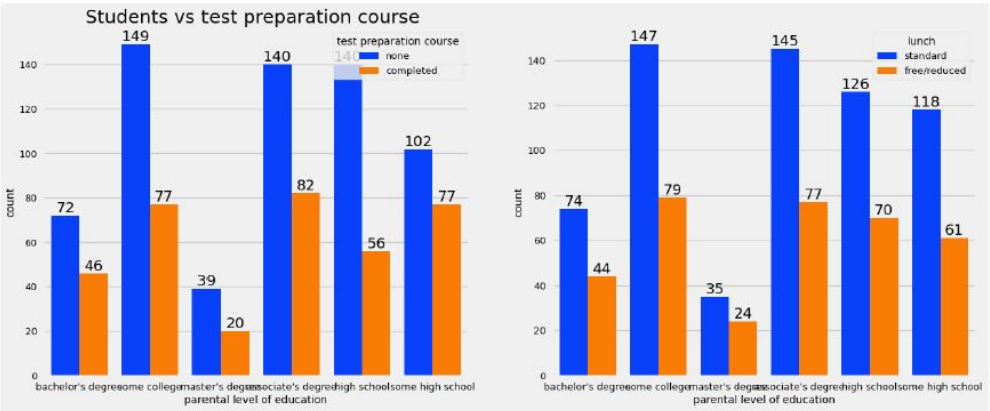


Insights

• Students being served Standard lunch was more than free lunch

BIVARIATE ANALYSIS (Is lunch type intake has any impact on student's performance ?)

In [32]:
```
f,ax=plt.subplots(1,2,figsize=(20,8))
sns.countplot(x=df['parental level of education'],data=df,palette = 'bright',hue='test pr
ax[0].set_title('Students vs test preparation course ',color='black',size=25)

for container in ax[0].containers:
    ax[0].bar_label(container,color='black',size=20)

sns.countplot(x=df['parental level of education'],data=df,palette = 'bright',hue='lunch',
for container in ax[1].containers:
    ax[1].bar_label(container,color='black',size=20)
```



Insights

- Students who get Standard Lunch tend to perform better than students who got free/reduced lunch

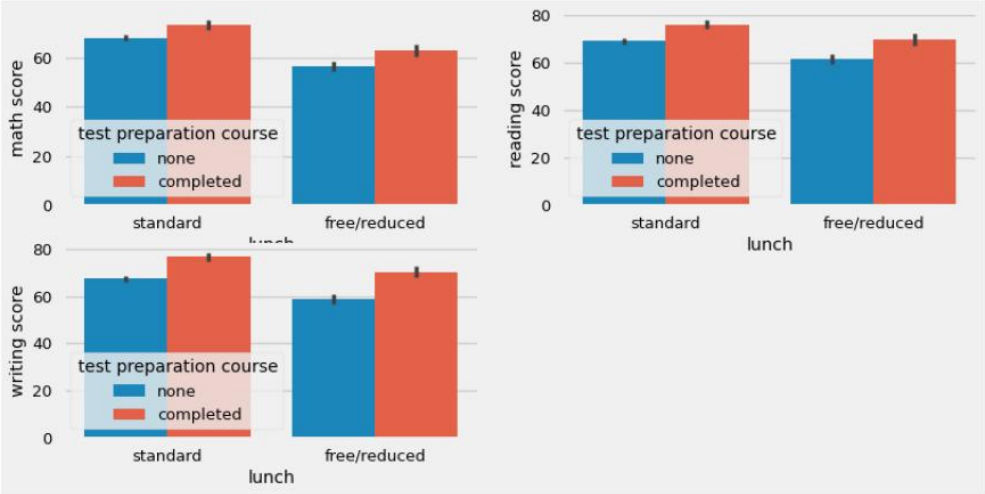
TEST PREPARATION COURSE COLUMN

- Which type of lunch is most common amoung students ?
- Is Test prepration course has any impact on student's performance ?

BIVARIATE ANALYSIS (Is Test prepration course has any impact on student's performance ?)

In [33]:
```
plt.figure(figsize=(12,6))
plt.subplot(2,2,1)
sns.barplot (x=df['lunch'], y=df['math score'], hue=df['test preparation course'])
plt.subplot(2,2,2)
sns.barplot (x=df['lunch'], y=df['reading score'], hue=df['test preparation course'])
plt.subplot(2,2,3)
sns.barplot (x=df['lunch'], y=df['writing score'], hue=df['test preparation course'])

# <AxesSubplot:xlabel='lunch', ylabel='writing score'>
```

Out[33]: <AxesSubplot:xlabel='lunch', ylabel='writing score'>

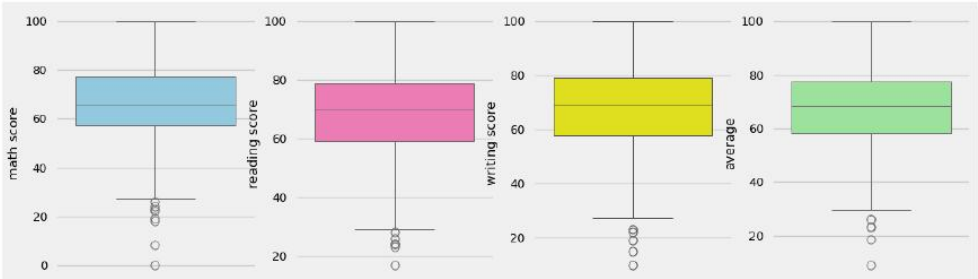


Insights

• Students who have completed the Test Prepration Course have scores higher in all three categories than those who haven't taken the course

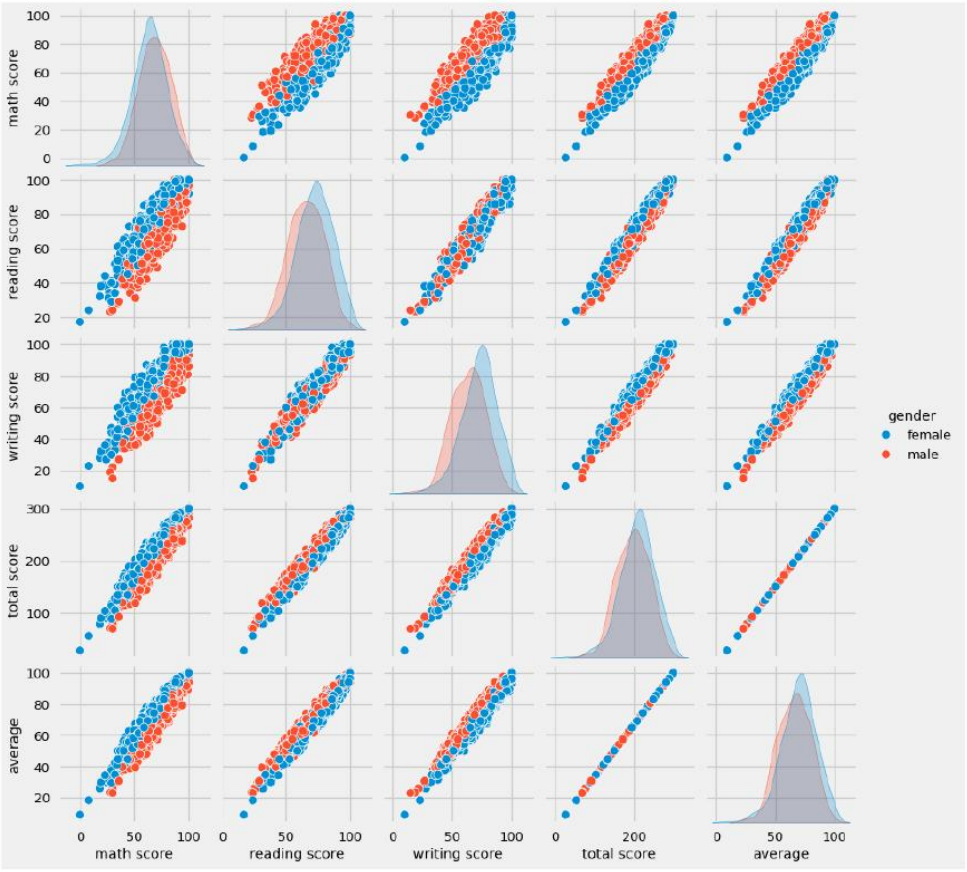
CHECKING OUTLIERS

In [34]:
```
plt.subplots(1,4,figsize=(16,5))
plt.subplot(141)
sns.boxplot(df['math score'],color='skyblue')
plt.subplot(142)
sns.boxplot(df['reading score'],color='hotpink')
plt.subplot(143)
sns.boxplot(df['writing score'],color='yellow')
plt.subplot(144)
sns.boxplot(df['average'],color='lightgreen')
plt.show()
```



MUTIVARIATE ANALYSIS USING PAIRPLOT

In [35]:
```
sns.pairplot(df,hue = 'gender')
plt.show()
```



Insights

• From the above plot it is clear that all the scores increase linearly with each other

Conclusions

- Student's Performance is related with lunch, race, parental level education
- Females lead in pass percentage and also are top-scorers
- Student's Performance is not much related with test preparation course
- Finishing preparation course is beneficial

CHAPTER 4:

DEVELOPMENT PLAN/ SCHEDULE/ FLOW CHART FOR COMPLETION OF PROJECT

4.1 Development Plan

Here is a development plan for an exploratory data analysis (EDA) project:

1) Define the problem statement: The first step is to define the problem statement for the EDA project. It could be to explore a specific research question or hypothesis related to a particular dataset.

In this project, we defined the problem statement in the First Chapter as per the datasets.

2) Collect the data: The next step is to collect the data. This can be done by searching for publicly available datasets on sources such as Kaggle or data.gov.in, or by obtaining data from relevant stakeholders.

We collected data from various sites like ncrb.gov.in/en/crime-india and Kaggle.

3) Data cleaning: Once the data has been collected, it needs to be cleaned and pre-processed. This includes removing duplicates, handling missing data, and converting data types.

With the help of Python (programming language) and Jupyter notebook (IDE), we imported the data and performed the operation i.e., data cleaning.

4) Exploratory data analysis: In this step, you will analyse the data to understand its characteristics, identify patterns, and visualize the data. You can use various techniques such as summary statistics, histograms, scatter plots, and box plots to explore the dataset. The focus here is to identify any trends, relationships, or anomalies in the data.

Performed EDA on imported datasets.

4.2 Schedule

Here is a schedule for our EDA project:

Week 1 - **Define the problem statement and research question/hypothesis**

- Collect the data

- Understand the data sources and data types
- Start data cleaning and per-processing

Week 2 - Complete data cleaning and per-processing

- Perform exploratory data analysis
- Identify any trends, patterns, or anomalies in the data
- Plan feature engineering tasks

Week 3 - Perform feature engineering tasks

- Create and test different models, if applicable
- Identify the most relevant insights from the data
- Start drafting the learning report

Week 4 - Finalize the learning report

- Review the report and refine as necessary
- Create visualizations and tables to support the findings
- Prepare the report for presentation or publication

CHAPTER 5:

CONCLUSION

In conclusion, the EDA (Exploratory Data Analysis) project was a valuable learning experience that allowed us to apply our knowledge of statistical analysis and data visualization to a real-world dataset. Through this project, we were able to:
Identify and clean the data to prepare it for analysis Perform various statistical analyses to gain insights into the dataset Create meaningful visualizations to communicate the results of our analyses. One of the key takeaways from this project is the importance of exploratory data analysis in the data science field. Before applying any modelling techniques or making predictions, it is crucial to understand the data and uncover any patterns or trends that may be present. EDA provides a foundation for further analysis and can help guide decision-making in a variety of industries.
Basically, the EDA project provided a hands-on learning experience that allowed us to practice and improve our data analysis skills. We gained valuable insights into the dataset and the importance of EDA in the data science workflow. This project has prepared us well for future data analysis projects and real-world applications.

CHAPTER 6:

REFERENCES AND BIBLIOGRAPHY

1. Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.

2. John T. Behrens, "Principles and Procedures of Exploratory Data Analysis," Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.

3. Chokey Wangmo, "An Exploratory Study On Bank Lending To SME Sector In Bhutan," International Journal of Scientific & Technology Research, volume 6, issue 11, November 2017, pp. 47-51.

4. Matthew Ntow-Gyamfi and Sarah Serwaa Boateng, "Credit Risk and Loan Default among Ghanaian Banks: An Exploratory Study," Management Science Letters, Vol. 3, 2013, pp.753–762.

5. X. Francis Jency, V. P. Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients," International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018, pp.176-179.

6. K. Ulaga Priya, S. Pushp, K. Kalaivani, A. Sartiha, "Exploratory Analysis on Prediction of Loan Privilege for Customers using Random Forest," International Journal of Engineering & Technology, Vol. 7, Issue 2.21, 2018, pp. 339-341.

7. Bogumil M. Konopka, Felicja Lwow, Magdalena Owczarz, Łukasz Łaczmański, "Exploratory Data Analysis of a Clinical Study Group: Development of a Procedure for Exploring Multidimensional Data," PLOS ONE, [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC6107146/pdf/pone.0201950.pdf, August23, 2018, pp. 1-21.

8. Introduction To Machine Learning using Python [Online], Available: https://www.geeksforgeeks.org/introduction-machine-learning-using-python/

9. Exploratory data analysis – From Wikipedia, the free encyclopedia [Online], Available: https://en.wikipedia.org/wiki/Exploratory_data_ analysis

6.2 Bibliography

1. Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

2. Bruce, P. C., Bruce, A., & Gedeck, P. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media.

3. Grolemund, G., & Wickham, H. (2017). R for Data Science: Import, Tidy, Transform,
Visualize, and Model Data. O'Reilly Media.

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

5. Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). Analysing Ecological Data. Springer