Dav Team Assignment

Samarth Banodia	**Roll No: 24B0392** **Ph No. : 7999842460**
Submitted on: May 13, 2025	**Webpage:** https://samarthbanodia.github.io **Email: 24b0392@iitb.ac.in**

Question 1 & 2

Question 1:

Handling Missing Data: - For Numerical Categories (Age , Studyhours , Attendance etc) I used linear interpolation for filling the gaps and Further used Forward Fill method for the left out starting missing data cells.
 - For Categorical Columns I simply filled the missing ones with Mode of that category.

Outliers : Retained the visually found outliers from plots as they might show genuine variations.

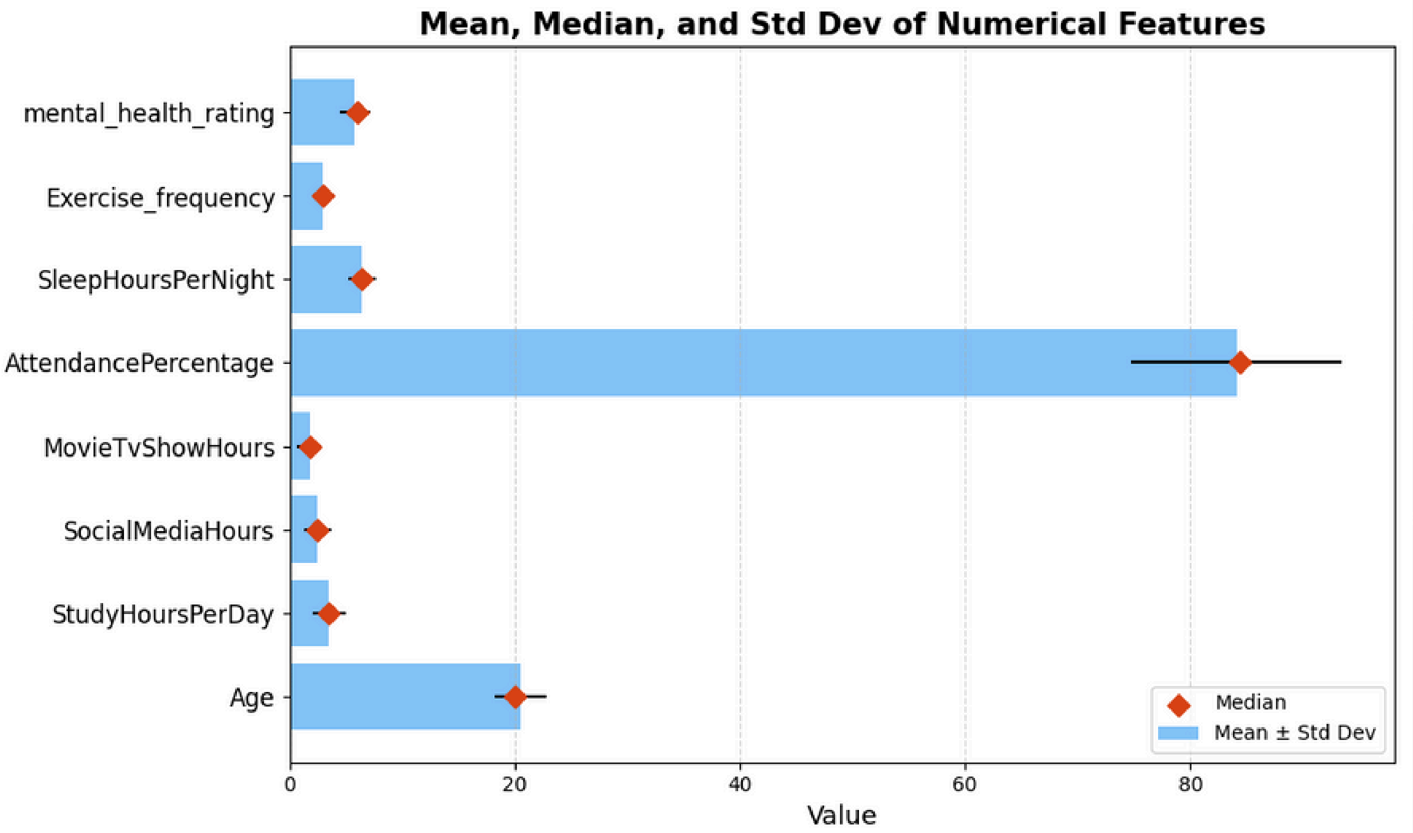
Setting Data Types : Numerical Columns are set to float or int.

My Approach-

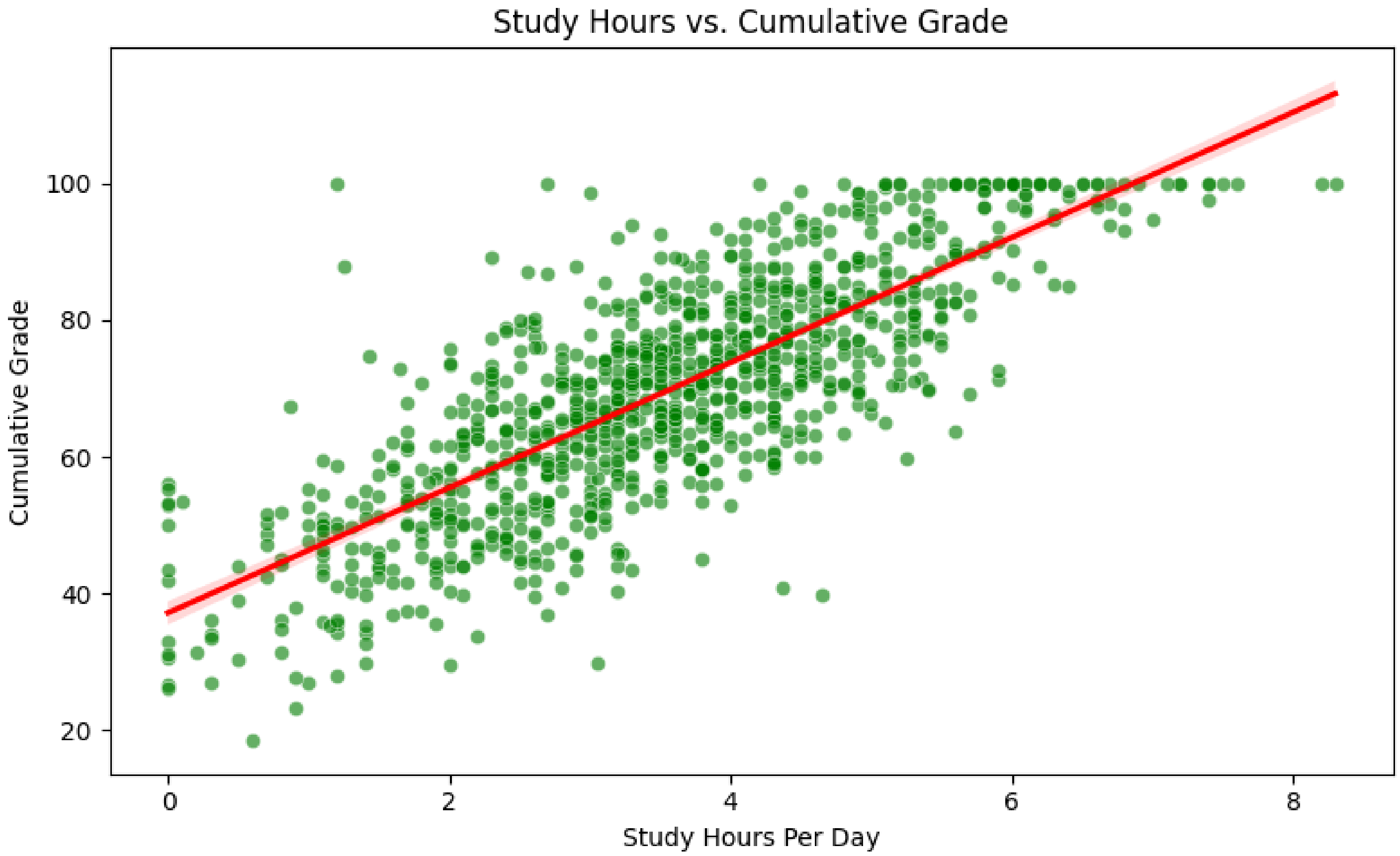
I used Seaborn library to show various plot as it generates more attractive plots and is easy to use . It also has several builtin themes. Then i searched various ways to handle missing data - I came across mean , median , mode methods , forward and backward filling , interpolation. After which i chose linear interpolation as i think that would be more accurate. for categorical columns i chose to fill with mode of the category as i believe it'll preserve the pattern. Then i had to choose various plots to help visualise the dataset , i wrote down in a paper what types of plots i want and what features to show.

Mean , Median and Standard Deviation

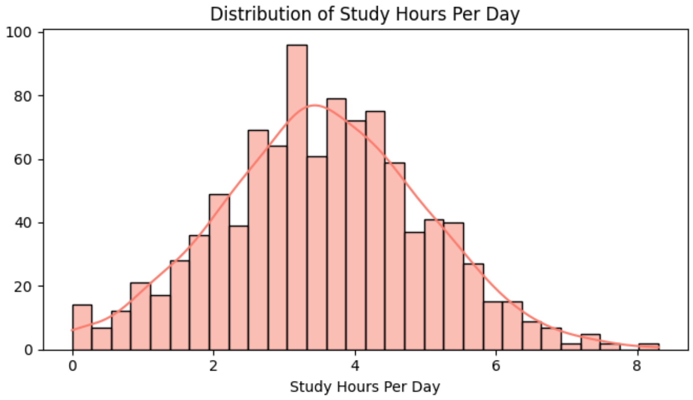
FEATURE	MEAN	MEDIAN	STD
Age	20.493	20	2.269
Study Hours per Day	3.543	3.5	1.455
Social Media Hours	2.510	2.5	1.170
Movie/Tv Show Hours	1.818	1.8	1.066
Attendance %	84.08	84.4	9.378
Sleep Hrs per Night	6.54	6.4	1.206
Exercise Frequency	3	3	0.930
Mental Health Rating	5.81	6	1.318



Cumulative Grade vs Study Hours

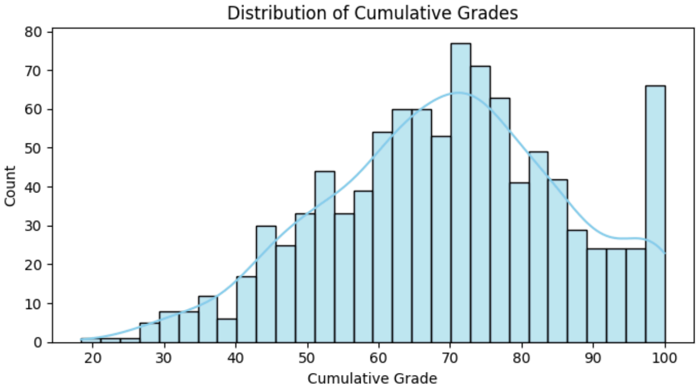


The scatter plot shows a clear positive relationship between Study hours per day and Cumulative Grade. Students who dedicate more hours to studying generally achieve better grades, as indicated by the plot and the fitted regression line. However, there is noticeable spread at each study hour level, suggesting that while study time is important other factors also influence academic performance.

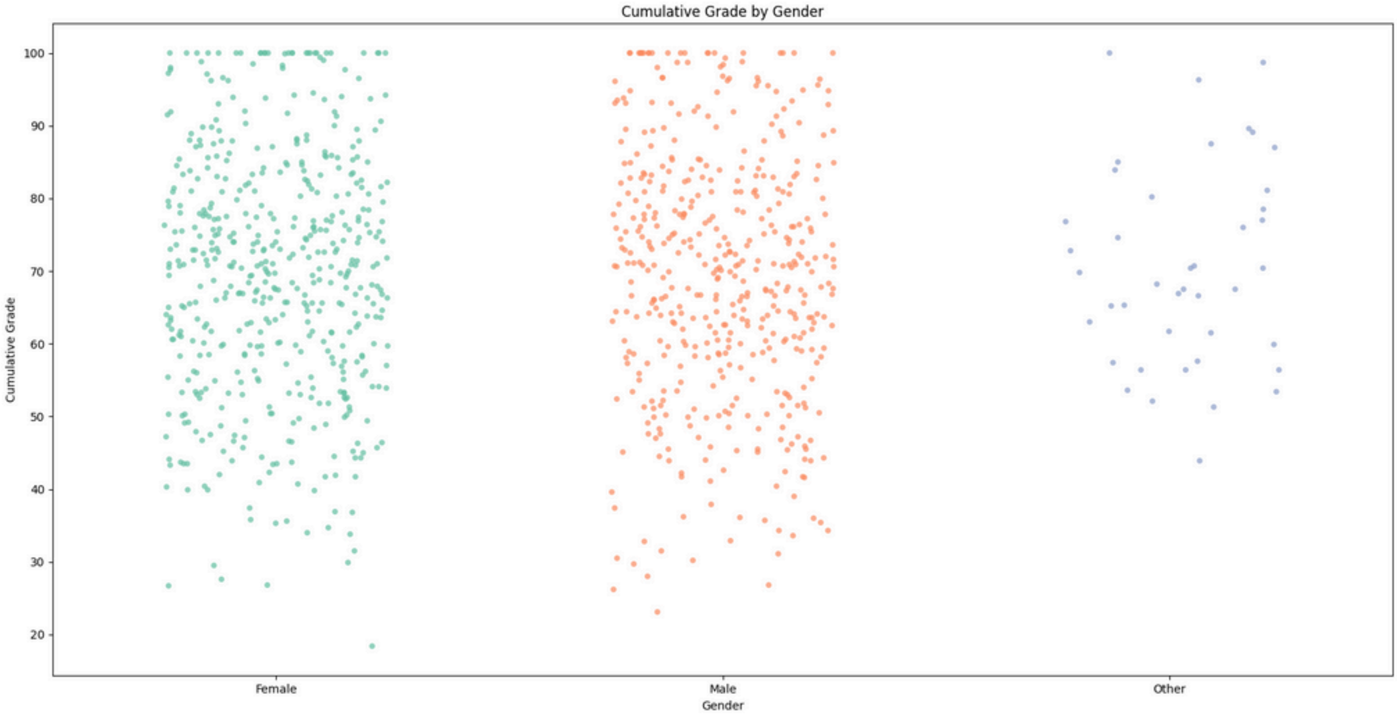


This plot reveals that study hours is normally distributed centered around 3 to 4 hours. Most students study between 2 and 6 hours daily, with relatively few studying very little or more than 7 hours.

The histogram of cumulative grades displays a right-skewed distribution, with most students scoring between 60 and 80. Notably, there is a spike at the 100, indicating a significant number of students achieving perfect scores.

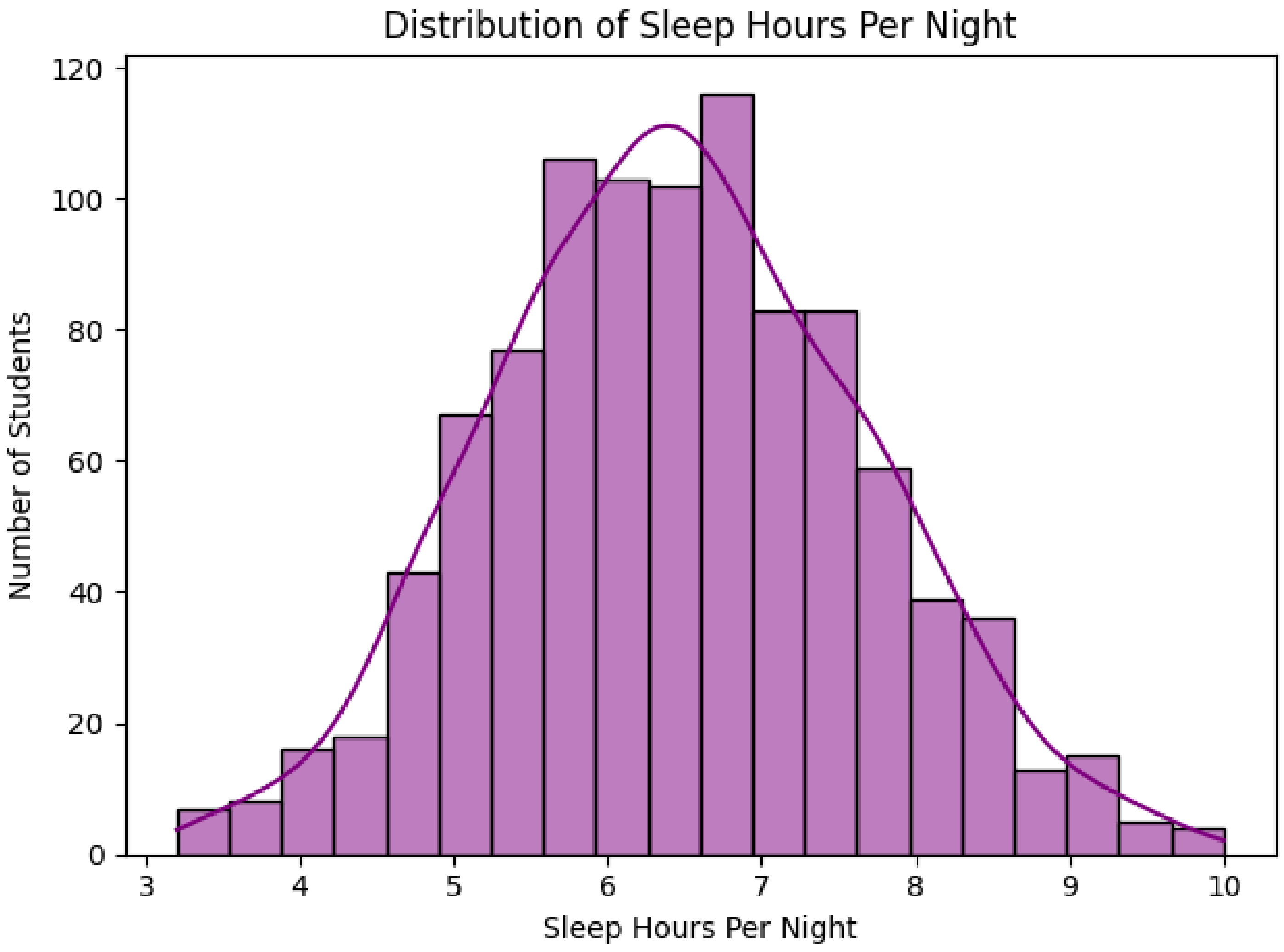


Cumulative Grade vs Gender



This strip plot shows the distribution of cumulative grades across different gender . The spread and central tendency of grades appear similar for both male and female students, with no significant difference in performance between these groups. Students in "Other" also display a comparable range of grades, though their sample size is relatively smaller. Overall, academic achievement shows no significant gender based disparity.

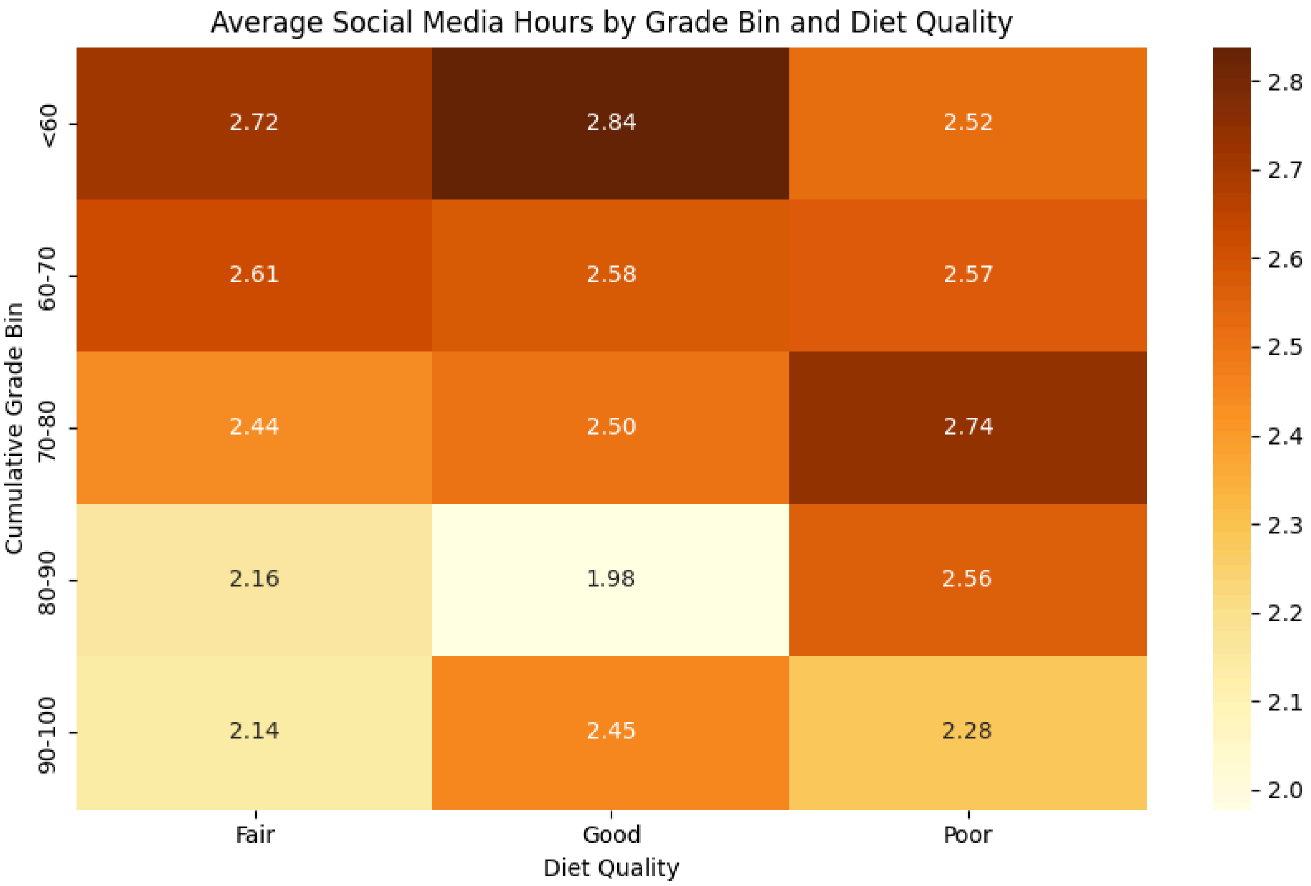
Distribution of Sleep Hours Per Night



The histogram reveals a roughly normal distribution of sleep patterns among students, centered around 6-7 hours per night. Most students maintain sleep durations within the 5-8 hour range, with significantly fewer students reporting either very short (less than 4 hours) or extended (more than 9 hours) sleep hours. This suggests that while the average student gets adequate sleep, there is still some variation in sleep habits across the students.

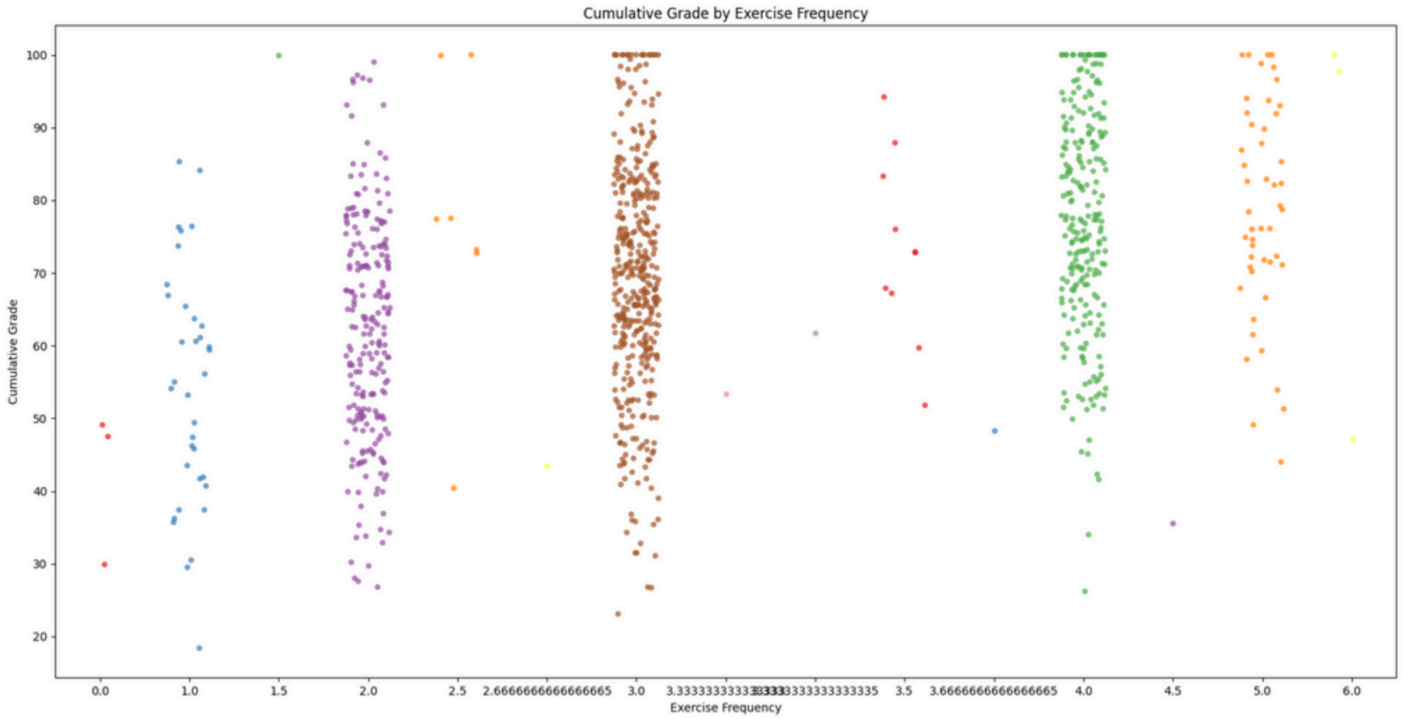
- Peak sleep duration: 6.5-7 hours (approximately 115-120 students)
- Range: 3 to 10 hours per night
- Mean sleep duration: ~6.5 hours
- Most common range: 5-8 hours (covers ~75% of students)
- Students with potentially insufficient sleep (<5 hours): ~10%
- Students with extended sleep (>8 hours): ~15%

Heat map of Avg Social Hours by Grade bin & Diet Quality



This heatmap shows how average social media usage varies across different grade and diet ranges. Students with lower cumulative grades (<70) tend to spend more time on social media, regardless of diet quality. Those in the highest grade bin (90–100) consistently report the lowest average social media hours across all diet groups. Additionally, students with both good diet quality and higher grades (80–100) have the lowest social media usage overall, suggesting a link between healthier habits, reduced social media time, and better grades.

Exercise Frequency vs Cumulative Grade



The scatter plot displays the distribution of cumulative grades across different exercise frequencies per week. Students who exercise more frequently (4–6 times per week) tend to cluster at higher grade levels, with a greater concentration of scores above 80. In contrast, those with little or no exercise show a wider spread and more lower grades.

- Mean grade for students exercising 5–6 times/week: ~80
- Mean grade for students with no exercise: ~65
- Highest grades (near 100) are most common among students exercising 4+ times per week.
- Lower grades (<60) are more prevalent among students who exercise less than twice a week

---

QUESTIONS & PATTERNS

How does exercise frequency impact academic performance? The exercise frequency plot shows clear grade distributions across different exercise levels, revealing whether more frequent physical activity correlates with better academic outcomes.

What relationship exists between social media usage and academic performance? The heatmap demonstrates how social media hours vary across grade bins, answering whether higher-performing students spend less time on social platforms.

How do diet quality and grade performance interact with social media habits? The social media/diet/grade heatmap reveals patterns across these three variables, showing whether healthier eating habits and better grades correlate with different digital behaviors.

Is there a gender gap in academic performance at IIT Bombay? The gender plot directly answers whether male, female, and other-identifying students show different grade distributions or similar academic outcomes.

What sleep duration is most common among students, and how does this compare to recommended healthy sleep ranges? The sleep hours histogram provides detailed distribution data to evaluate student sleep patterns against health recommendations.

Do students with combined positive habits (good diet, regular exercise, moderate social media) show superior academic performance? Cross-referencing across plots allows analysis of how multiple lifestyle factors together influence grades.

How significant is the correlation between lifestyle factors (sleep, diet, exercise, social media) and academic performance compared to gender differences? Comparing all plots together answers whether behavioral factors show stronger relationships with grades than demographic factors.

- Strongest Predictors of high Grades are Study hours, exercise frequency, and low social media usage.

- The grade spike at 100 may reflect lenient grading.

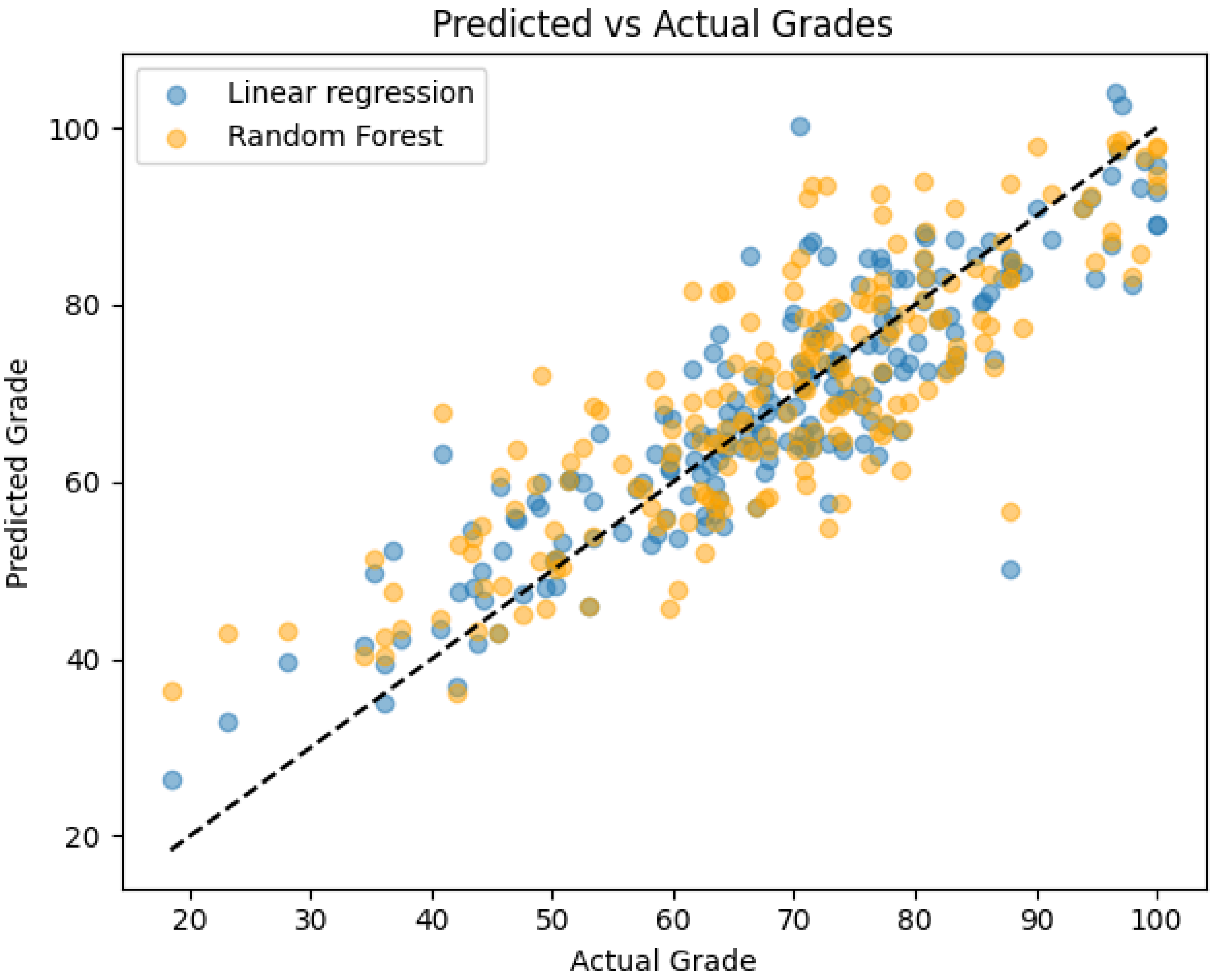
- While most patterns align with health-academic links, anomalies emphasize individual adaptability e.g., high achievers with poor sleep/diet.

Question 2

Made linear regresion and Random Forest Model with help of scikit learn , Train and Tested the model with 80-20 split. Below are the performance metrics of both the models.

Metrics	Linear Regression	Random Forest
R^2	0.77	0.68
RMSE	58.46	82.42
MAE	5.79	7.13

Error plots & Prediction vs Actual Grades

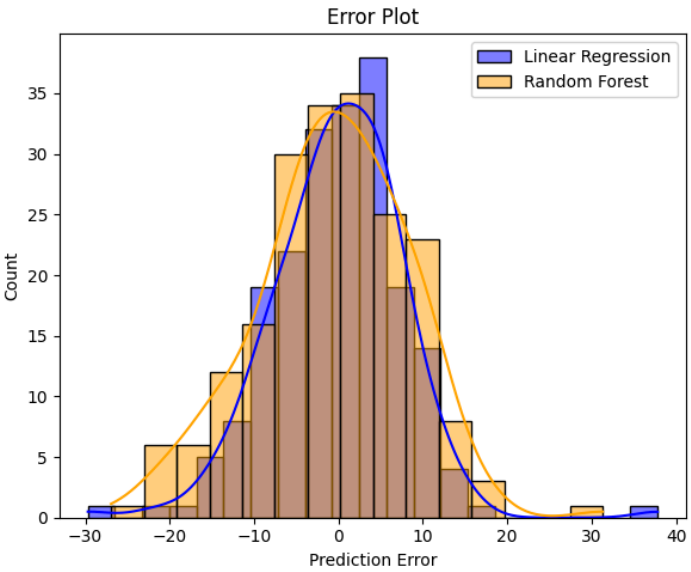


The middle line is for 100% accuracy that is all predicted grades are true, we can see many points fall near the line but also some are scattered away. It is not clear from this plot that which model is more accurate , for that we generate error plot which is a histogram of actual-predicted grades for each model.

- Random Forest (orange) shows a more concentrated error distribution with fewer extreme errors

- Linear Regression (blue) has a slightly wider spread of errors, particularly visible in the ends.

- Both are centered around zero, but Random Forest's errors clusters more tightly



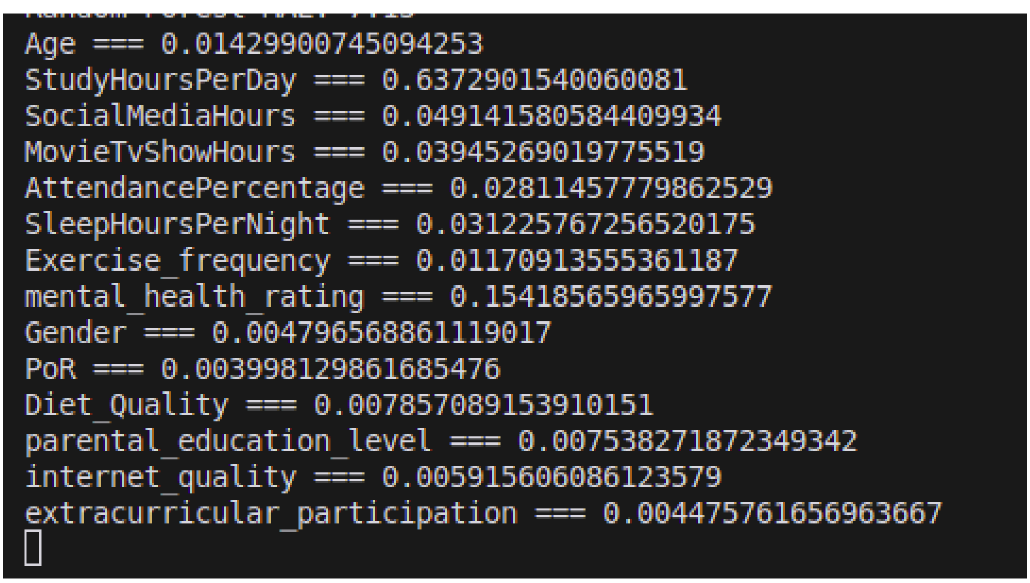
Prediction Accuracy	Why Random Forest Performs Better
In the scatter plot, both models follow the diagonal trend (perfect predictions)	**Random Forest can capture non-linear relationships between features like study hours and mental health**
Random Forest points appear slightly more concentrated along the ideal prediction line	**Random Forest helps reduce overfitting and handles outliers better**
Linear Regression shows more variance, particularly for grades in the middle range	**Educational data often contains complex interactions between variables that tree-based models can detect, while linear models cannot**

Feature Importance & Interpretation

Feature importance refers to assigning a score to each input feature based on how useful or influential it is in predicting the target variable.

Used the feature_importances_ method from scikitlearn

```
importances = rf.feature_importances_

for i in range(14):

    print(f'{features[i]} === {importances[i]}')
```



```
Age === 0.01429900745094253
StudyHoursPerDay === 0.6372901540060081
SocialMediaHours === 0.049141580584409934
MovieTvShowHours === 0.03945269019775519
AttendancePercentage === 0.02811457779862529
SleepHoursPerNight === 0.031225767256520175
Exercise_frequency === 0.01170913555361187
mental_health_rating === 0.15418565965997577
Gender === 0.004796568861119017
PoR === 0.003998129861685476
Diet_Quality === 0.007857089153910151
parental_education_level === 0.007538271872349342
internet_quality === 0.005915606086123579
extracurricular_participation === 0.004475761656963667
```

Here the top two Features are **Study Hours Per Day** and **Mental Health Rating.**

Interpretation of Top Two Features:

- **Study Hours Per Day :** Students who dedicate more hours to studying generally achieve higher grades. This strong, direct relationship is consistently observed in both the data and educational research.

- **Mental_health_rating :** Students with better mental health ratings tend to perform better academically. Good mental health supports concentration, motivation.

Question 3&4

My Approach:

For this Question i Decided to use Random Forest for the first part and Clustering for the later part both implemented using the scikitlearn lib. as for the first part we have labelled data which allows for supervised learning methods(Random forest, linear logistic reg, NNs etc) but for the Bonus Question we dont have a labelled dataset so i have to opt for unsupervised learning thats why i chose K means Clustering for that. I used pandas to create two different dataframes one for the Training csv file and other for the testing file.

Here is the flow of the Problem,

- **Handling Missing Data**
- **Label Encoding for the model**
- **Training and Testing the Model**

Handling Missing Data:

For Numerical columns (Age, Family_Size, Work_Experience) were filled with the mean value of each column.

Then For Categorical columns (Gender, Ever_Married, Profession, Graduated, Energy_Consumption, Preferred_Renewable) were filled with the mode of each column.

I Ensured that no missing data gaps were left by print the .isnul().sum() method for each column for Both The Training and the Testing Csv file.

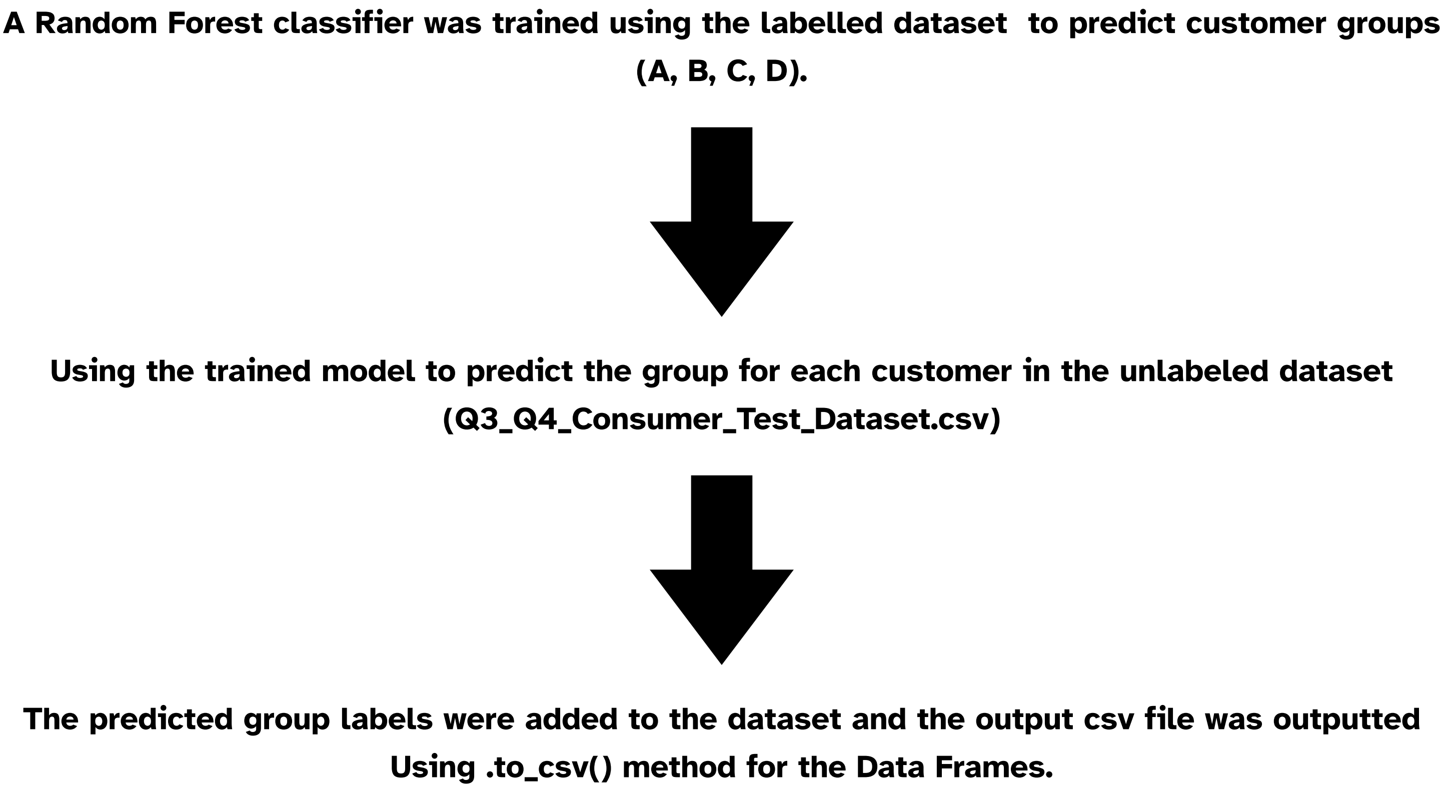
Label Encoding for the Model:

Categorical columns were label encoded to convert them into numerical format, which is required for machine learning algorithms as they can not directly understand the categories.

Used the LabelEncoder Package offerred by the Scikit learn library.

Training and Testing the Model:

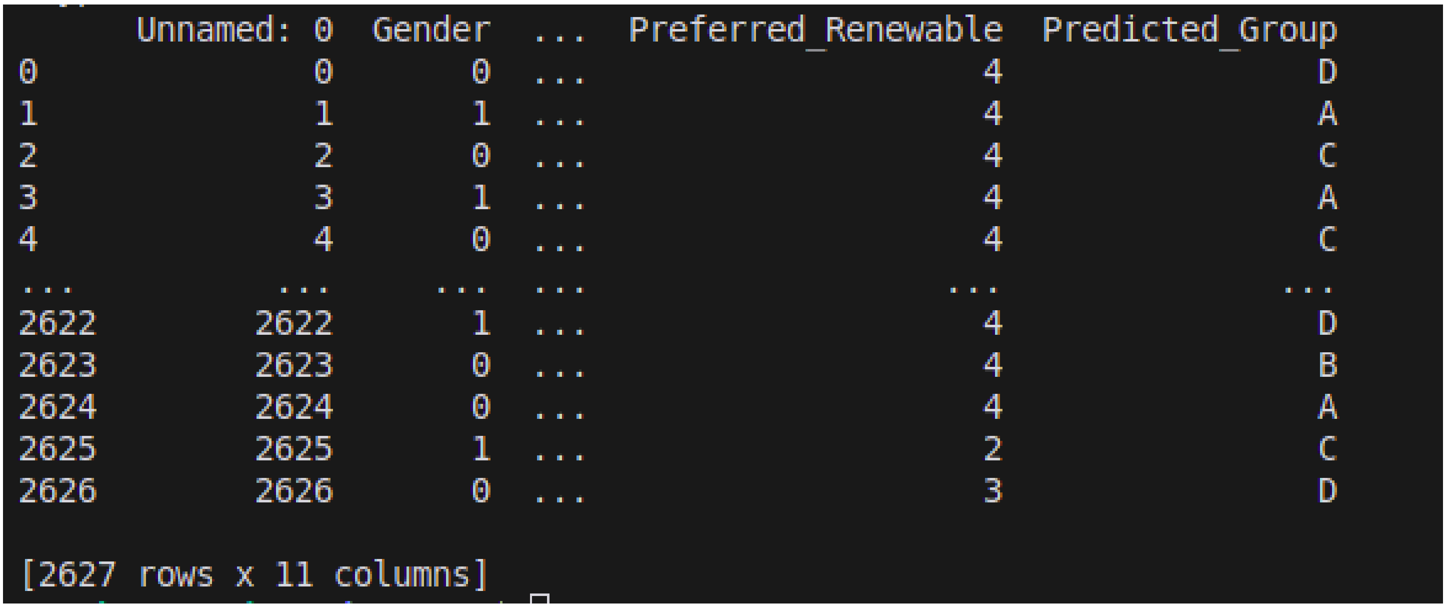
We chose Random Forest for customer group prediction because it is a highly accurate and robust supervised machine learning algorithm that combines multiple decision trees to make final predictions.



A Random Forest classifier was trained using the labelled dataset to predict customer groups (A, B, C, D).

Using the trained model to predict the group for each customer in the unlabeled dataset (Q3_Q4_Consumer_Test_Dataset.csv)

The predicted group labels were added to the dataset and the output csv file was outputted Using .to_csv() method for the Data Frames.



```
      Unnamed: 0  Gender  ...  Preferred_Renewable  Predicted_Group
0              0       0  ...                    4                D
1              1       1  ...                    4                A
2              2       0  ...                    4                C
3              3       1  ...                    4                A
4              4       0  ...                    4                C
...          ...     ...  ...                  ...              ...
2622        2622       1  ...                    4                D
2623        2623       0  ...                    4                B
2624        2624       0  ...                    4                A
2625        2625       1  ...                    2                C
2626        2626       0  ...                    3                D

[2627 rows x 11 columns]
```

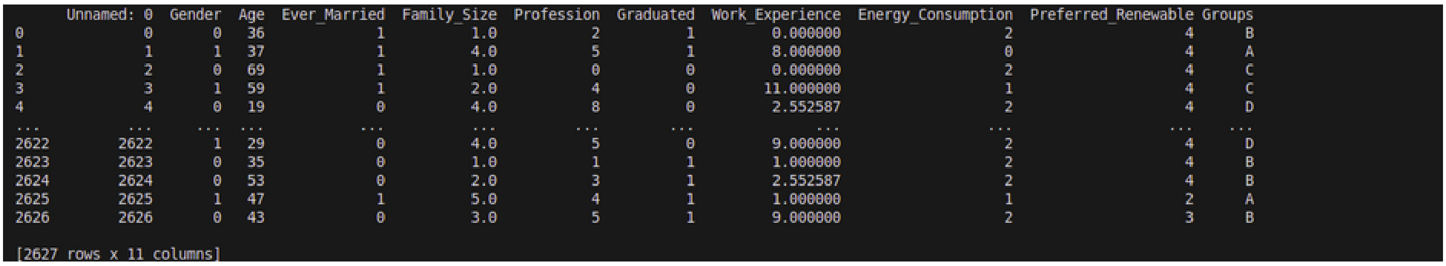
output from printing the final dataframe.

BONUS QUESTION

As we had to assume there was no labelled dataset for the given Customers, This problem represents need for unsupervised learning as there is no way you can see whether your classification is correct or not . So here i decided to use **Kmeans clustering** with help of sklearn lib, This Method basically forms different clusters among the dataset based on different feature similarirty. You can specify how many cluster you want thats why this method is apt in this problem

- **First Step : Scaling features with StandardScaler() :** Kmeans measures distance between data points. If one feature has a large range for e.g Age vs. Family_Size, it dominates the distance calculation. Scaling ensures all features contribute equally.

- **Second Step : Applying Kmeans Clustering :** Used the KMeans Package from sklearn. put n_cluster = 4 as we have four groups A, B,C,D also put random_state =42 as a common practice - i have used random_state =42 wherever necessary in this assignment as maintaining uniformity.



```
      Unnamed: 0  Gender  Age  Ever_Married  Family_Size  Profession  Graduated  Work_Experience  Energy_Consumption  Preferred_Renewable Groups
0              0       0   36             1          1.0           2          1         0.000000                   2                    4      B
1              1       1   37             1          4.0           5          1         8.000000                   0                    4      A
2              2       0   69             1          1.0           0          0         0.000000                   2                    4      C
3              3       1   59             1          2.0           4          0        11.000000                   1                    4      C
4              4       0   19             0          4.0           8          0         2.552587                   2                    4      D
...          ...     ...  ...           ...          ...         ...        ...              ...                 ...                  ...    ...
2622        2622       1   29             0          4.0           5          0         9.000000                   2                    4      D
2623        2623       0   35             0          1.0           1          1         1.000000                   2                    4      B
2624        2624       0   53             0          2.0           3          1         2.552587                   2                    4      B
2625        2625       1   47             1          5.0           4          1         1.000000                   1                    2      A
2626        2626       0   43             0          3.0           5          1         9.000000                   2                    3      B

[2627 rows x 11 columns]
```

output for printing the dataframe. I mapped the cluster 0,1,2,3 as A,B,C , D.

Question 5

My Approach:

After exploring some NLP concepts i divided this problem into three sections -

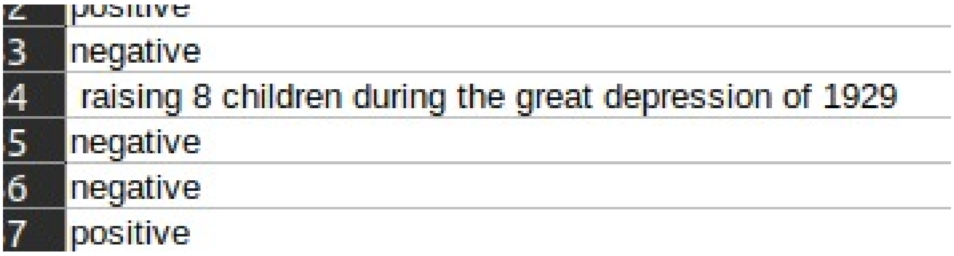
- **Data representation and Prepartion**
- **Learning Model Choice**
- **Evaluation of the Model**

Data Representation & Prep

There are Several method to prep the data like Tokenisation, stopword removal , steming , lemmatization. My strategy for cleaning the data is ,

- Converts text to **lowercase** for uniformity
- **Removes Punctuation** and non alpha-numeric characters (used Regex)
- Removes English **stopwords (** like 'the' , 'is' etc) using NLTK library
- Join the remaining words back into a cleaned string

After this, I removed the sentiment labels which were apart from 'negative' and 'positive'.



2	positive
3	negative
4	raising 8 children during the great depression of 1929
5	negative
6	negative
7	positive

Text Representation : Methods such as TFIDF, Word2Vec, Tranformers etc exist for text representation, I used TF-IDF(Term frequency inverse document frequency) - This is suitable as this method captures the importance of words in each movie review relative to the whole dataset, making it suitable for text classification tasks as it reduces the influence of common but less informative words.

ML Model Choice

There are many options for selection a model such as Logistic Regression , Naive Bayes, Random Forest, SVM , Neural Networks etc.

↓

Here I chose **Logistic Regression** as its more effective for binary classification (just 2 classes - negative and positive) and works well with TF-IDF.

↓

Training : The dataset is split into training and test sets (80/20 split) to evaluate performance on unseen data.

Evaluation Report

I used the Classfication Report offered by the scikit learn library itself for the evaluation of the model.

Classification Report

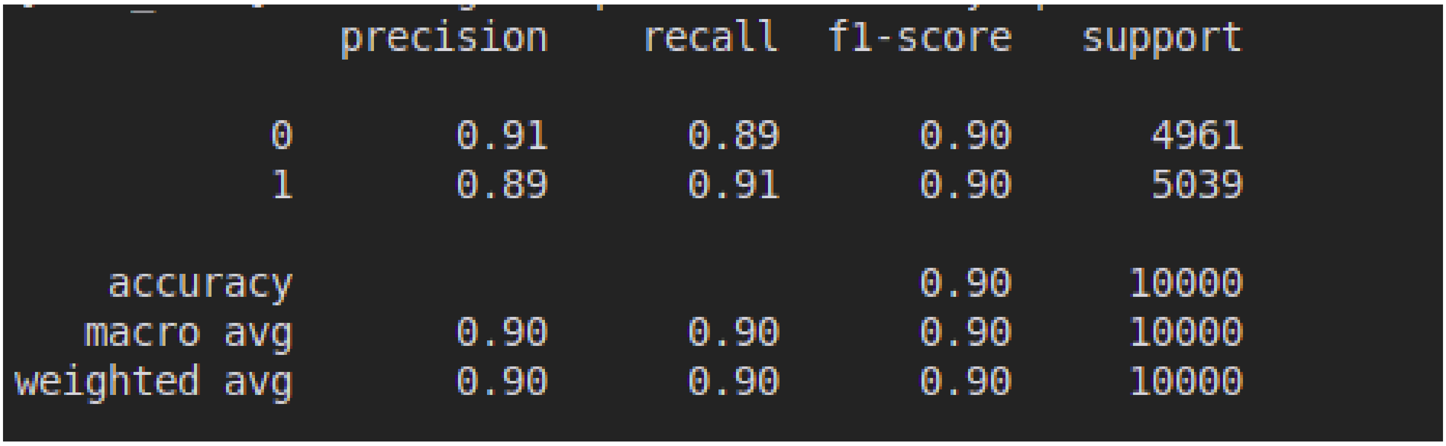
Label	Precision	Recall	F1-Score	Support	$150,000
0 (Negative)	0.91	0.89	0.90	4961	$45,000
1 (Positive)	0.89	0.91	0.90	5039	$55,000

- **Precision** : For negative, 91% of the reviews predicted as negative were actually negative. Similarly For class positive 89% of the reviews predicted as positive were actually positive.

- **Recall** : For negative, 89% of all actual negative reviews were correctly identified. For positive 91% of all actual positive reviews were correctly identified.

- **F1-score** : The harmonic mean of precision and recall is 0.90 for both classes, indicating a good balance between precision and recall.

- **Support** : There are 4961 negative and 5039 positive reviews in the test set.

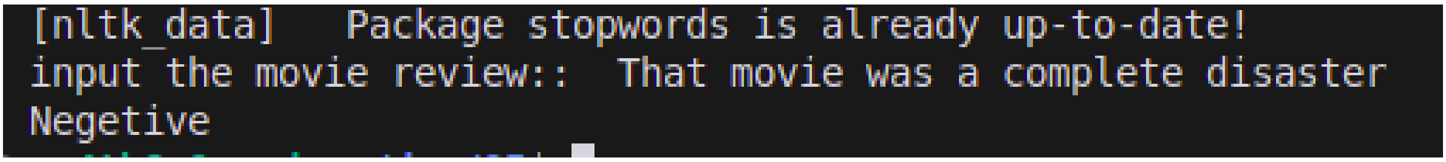


```
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      4961
           1       0.89      0.91      0.90      5039

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000
```

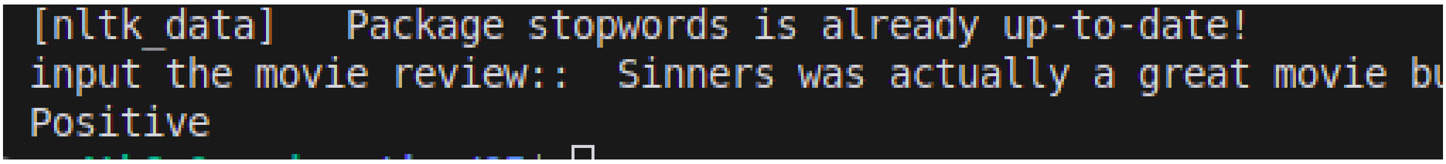
Overall Accuracy for the Model is 90%

Classifying Inputed string as negative or positive review:

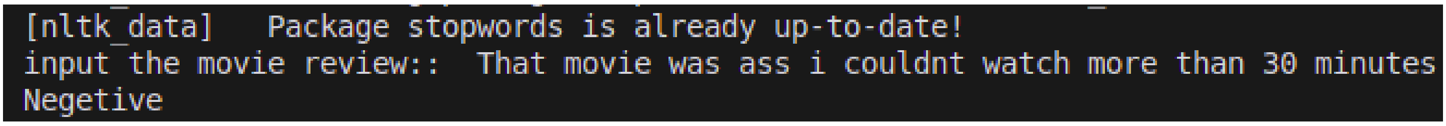
Used the Model trained on 80% dataset for this . The user inputed string is cleaned , vectorized and predicted by the model.



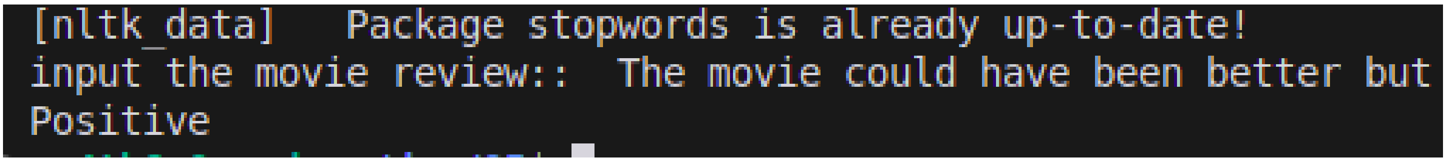
```
[nltk_data]   Package stopwords is already up-to-date!
input the movie review::  That movie was a complete disaster
Negetive
```



```
[nltk_data]   Package stopwords is already up-to-date!
input the movie review::  Sinners was actually a great movie bu
Positive
```



```
[nltk_data]   Package stopwords is already up-to-date!
input the movie review::  That movie was ass i couldnt watch more than 30 minutes
Negetive
```



```
[nltk_data]   Package stopwords is already up-to-date!
input the movie review::  The movie could have been better but
Positive
```