TASK-1

Starting-

1.First I changed the column names to make them understandable.Then made another dataframe 2 to store some columns with yes/no as 0 and 1
2.Checked the range of values in feature 1 ,2 and 3.
3.Then Plotted the correlation map for the hidden features versus other features

Hidden feature identification:

Feature_1-From the graphs and heatmap:

Feature 1 has positive correlation with the number of failures and the value of feature_1 is in the range of the ages of the secondary school students so if we assume that the feature_1 is age then with increasing age the number of failures is increasing which is very evident that's why the feature one is the AGE of the students
Feature_2-From the graphs and heatmap:

Feature_2 and positive correlation with grades and preference for higher studies.So this feature is indicating more studious students.So this feature can be a measure on some scale of STUDIOUSNESS of students or how much THEY STUDY

Feature_3-From the graphs and heatmap:

Feature_3 has strong Positive correlation with weekday alcohol consumption and friends hang out frequency,so this feature is indicative of some type the social activeness of the student.It might be related to the parting nature of the students.It can also be the weakend alcohol consumption as that column is not there.It may be the tendency of the student to go to in some scale.So Feature_3 is WEEKEND ALCOHOL CONSUMPTION OR PARTYING FREQUENCY

Handling missing values:

1.Father Education: From the graphs plotted father education had very much relationship with mother education and I also noticed that the fathers whose job is at home have generally a low level of Education. So in the missing values when the father job was at home I filled the values with 1 and at rest of the values I filled the father's education value same as the mother.

2.Term_2 grade-Had strong correlation with other grades so filled the value as average of other grades.

3.Absence count-Students with health score 1 had median absence count as 3.5 And students with other health states had absence median 2, so I filled the students with health status 1 as 3 and others with value 2.

4.Feature_1-As it is the age it had relation with failures.And most of the missing values had failures 0.

```
failures
0    34
3     2
2     1
1     1
```

I grouped the feature_1 values according to failures and then saw the description for failures==0,==1,==2,==3.
From that i filled the students with 0 and 1 failure with 17 age and other with 18.Most of the students were with 0 failures.

5.Higher Education:It had correlation with failures','term_1_grade','term_2_grade','final_grade','Feature_1','Feature_

2','father_education','mother_education','health_status','family_education_support','weekday_alcohol_consumption','friends_hangout_frequency','absence_count'
Made a Random Forest model and fitted the values

6.Feature_3-First tried random forest.But after that i saw that filling with average of weekday alcohol consumption and handout frequency gave good result so filled with the mean

7.Feature_2-Had correlation with ['term_1_grade','term_2_grade','final_grade','weekday_alcohol_consumption','friends_hangout_frequency','extra_classes','extra_activities']
SO made random forest model.

8.Travel Time- Filled the students coming from urban areas as time one and coming from rural areas as time 2 from observations and graphs.

9.Family Size: Observe that the parents whose status was apart had a lower family size than the parents whose status was together so filled the values accordingly

10.Free Time:Free time depended on 'friends_hangout_frequency','weekday_alcohol_consumption','extra_activities','Feature_3','Feature_2','term_1_grade','term_2_grade','final_grade','free_time' from heatmap.Filled the values withKNN Imputer.

EDA

Asked these 19 Questions:

Q1. What is the distribution of absence count?
Q2. What is Number of students going to each school

Q3. Students of which school perform better

Q4. Students from which area perform better
Q5. Popular school according to address
Q6. Students of which gender have better grade
Q7. what is the popular job among the parents
Q8. Which parents are the most likely to be guardian
Q9. What is the major reason of choosing the school
Q10. Which sex has most alcohol consumption
Q11. Does performing good in one paper mean performing good in others
Q12. Is there any impact of parent education on the grade of students
Q13. Does studying more mean better grades
Q14. Does alcohol consumption affect the Grades
Q15. Which age group consumes more alcohol
Q16. Does having more failures affect the grades
Q17. Does friends hangout frequency affect the grades
Q18. Does having home internet mean higher grades
Q19. Is there any correlation between parents education vs internet access at home

Answers to then with graphs are in the notebook

Relationship Prediction:

Plotted the correlation heat map to see the relation with other features.
Most of the students were not in romantic relationship.
Did categorical column encoding.
Then first trained Random Forest
Did Hyperparameter tuning Using grid search cv for many combinations a sin notebook

At the end I selected 2 models -
Model-1 accuracy-0.569
Model-2 accuracy-0.546

Did same with logistic regression and got
Model_3 accuracy-0.584

Did same with XGBoost classifier and got
Model_4 accuracy-0.52

As from hyperparameter training and 4 models still the accuracy is not much high. We can infer that maybe predicting the relationship status from these parameters is not fully possible.This fact is also visible from Correlation heatmap

DID SHAP ANALYSIS

From model_3:TOP FEATURES

1.Being male decreases the chances of being in relationship
2.Choosing school for reputation decreases the chances of being in relationship

From model_2:TOP FEATURES

1.Being female increases the chances of being in relationship
2.High absence count also has negative impact

From model_4:TOP FEATURES

1.Being of more age increases the chances of being in relationship
2.Being female increases the chances of being in relationship

FROM THESE MODELS ONE THING TO NOTE IS THAT BEING FEMALES FAVOURS THE CHANCES OF BEING IN RELATIONSHIP

THEN TOOK FEATURE PAIRS AND PLOTTED DECISION BOUNDARY PLOTS FOR IMPORTANT FEATURES.

Selecting two predicted values one positive and one negative for logistic regression model and exploring

TOOK THE VALUES WITH HIGH PROBABILITY OF BEING 0 AND 1 BY Model

FOR IN RELATION-more absence count and alcohol consumption favoured the chances of being in relation.

FOR not IN RELATION-Choosing school for reputation and being male unfavoured the chances of being in relation

Selecting two predicted values one positive and one negative for Random Forest model and exploring

FOR IN RELATION-Being female,more marks and greater value of age is favouring being in a relationship

FOR NOT IN RELATION-Very low absent count is pushing the prediction to not in a relationship

From the analysis and decision boundary plots some features like age ,being female,not choosing school for reputation are some features positively affecting the chances of being in relationship but there is uncertainty in predicting the outcome exactly