MAHARISHI MARKANDESHWAR DEEMED TO BE UNIVERSITY , MULLANA , AMBALA , HARYANA

MAHARISHI MARKANDESHWAR
(DEEMED TO BE UNIVERSITY)
Mullana-Ambala, Haryana
(Established under Section 3 of the UGC Act, 1956)
(Accredited by NAAC with Grade 'A++')

An Internship Program

On

Data Science and Analytics

Under the Guidance :

INVECAREER

Submitted by

Ishan Gupta

PROJECTS

1. **Student Performance Analysis**

2. **E-commerce Sales Analysis**

3. **Weather Data Analysis**

Acknowledgment :

I would like to express my sincere gratitude to everyone who has supported and guided me throughout the completion of this project, "Student Performance Analysis."

First and foremost, I would like to thank Bikash Bashyal, my project advisor, for their invaluable guidance, continuous support, and encouragement throughout this project. Their expertise and insights have been instrumental in shaping this project and ensuring its successful completion.

I am deeply grateful to INVECAREER for providing the necessary resources and a conducive environment for conducting this research. The access to academic resources and computational facilities has been crucial for the analysis and visualization work involved in this project.

I would also like to thank my family and friends for their unwavering support and understanding during this period. Their encouragement has been a constant source of motivation for me.

Finally, I extend my appreciation to all my professors and colleagues whose feedback and suggestions have greatly enhanced the quality of this project.

Thank you all for your invaluable contributions and support.

PROJECT: 1

Student Performance Analysis

Declaration :

I hereby declare that the project report entitled **"Student Performance Analysis"** submitted by me for the partial fulfillment of the requirements for the degree of Btech(CSE) at INVECAREER is my original work. This work has not been submitted elsewhere for the award of any other degree or diploma.

I have performed a comprehensive analysis of the dataset containing student exam scores, demographic information, and study habits. I have utilized various statistical and visualization techniques to investigate trends, correlations, and distributions within the data. The insights derived from this analysis are based on the dataset provided and the methodologies applied during this project.

I have adhered to ethical guidelines and academic standards while performing this analysis. All sources of information, including books, articles, and online resources, have been duly acknowledged in the report.

I understand that any form of plagiarism or academic dishonesty will be subject to disciplinary action as per the policies of INVECAREER.

I acknowledge the support and guidance of my INVECAREER, without which this project would not have been possible.

PROBLEM STATEMENT:

Student Performance Analysis

Utilize a dataset containing student exam scores, demographic information, and study habits. Analyze the distribution of exam scores and identify trends.

Investigate correlations between study time, demographic factors, and exam performance. Visualize the data using bar charts, scatter plots, and histograms

Provide recommendations for improving student performance based on the analysis.

Project Goals

The primary goals of this project, "Student Performance Analysis," are as follows:

1. **Data Collection and Preparation**:
    - Collect and load the dataset containing student exam scores, demographic information, and study habits.
    - Clean and preprocess the dataset to ensure it is suitable for analysis.
2. **Descriptive Analysis**:
    - Explore the dataset to understand its structure and main features.
    - Analyze the distribution of exam scores to identify patterns and outliers.
3. **Correlation Analysis**:
    - Investigate correlations between study time, demographic factors, and exam performance.
    - Identify key factors that have a significant impact on students' exam scores.
4. **Data Visualization**:
    - Use various visualization techniques such as histograms, scatter plots, bar charts, and heatmaps to represent the data and findings effectively.
    - Provide visual insights into the relationships between different variables.
5. **Predictive Modeling (if applicable)**:
    - Develop predictive models to estimate student performance based on demographic and study habit factors.
    - Evaluate the accuracy and reliability of the models.

6. **Insights and Recommendations**:
    - Draw actionable insights from the analysis to understand what factors contribute most to student success.
    - Provide recommendations for students, educators, and policymakers to improve student performance based on the analysis.
7. **Documentation and Reporting**:
    - Document the entire process, including data collection, preparation, analysis, and findings.
    - Prepare a comprehensive project report detailing the methodology, analysis, results, and recommendations.
8. **Ethical Considerations**:
    - Ensure the analysis is conducted ethically, respecting the privacy and confidentiality of the students' data.
    - Acknowledge all sources of information and support received during the project.

Introduction

Background :

Education plays a crucial role in shaping the future of individuals and societies. Understanding the factors that influence student performance can provide valuable insights for educators, policymakers, and students themselves. In recent years, there has been growing interest in analyzing educational data to identify trends, uncover correlations, and develop strategies to enhance learning outcomes. This project aims to contribute to this field by conducting a comprehensive analysis of student performance data.

Motivation :

The motivation for this project stems from the increasing recognition that academic success is influenced by a multitude of factors, including demographic characteristics, study habits, and socio-economic conditions. By analyzing these factors, we can gain a deeper understanding of how they interact and impact student performance. This knowledge is essential for designing effective interventions and policies that can help improve educational outcomes for all students.

Objectives :

The primary objectives of this project are as follows:

1. **Data Collection and Preparation**: To gather and preprocess a dataset that includes student exam scores, demographic information, and study habits.
2. **Descriptive Analysis**: To explore the dataset and understand the distribution of exam scores, identifying any notable patterns or outliers.
3. **Correlation Analysis**: To investigate the relationships between study time, demographic factors, and exam performance, and identify the key factors that significantly impact student success.
4. **Data Visualization**: To use various visualization techniques to effectively represent the data and the findings from the analysis.
5. **Predictive Modeling (if applicable)**: To develop predictive models that can estimate student performance based on demographic and study habit factors, and evaluate the models' accuracy and reliability.
6. **Insights and Recommendations**: To draw actionable insights from the analysis and provide recommendations for improving student performance.
7. **Documentation and Reporting**: To document the entire process, including data collection, preparation, analysis, and findings, in a comprehensive project report.

Methodology :

The methodology of this project involves several key steps:

1. **Data Collection and Preparation**: The dataset will be sourced from reliable educational databases. The data will be cleaned and preprocessed to handle any missing values, outliers, or inconsistencies.
2. **Descriptive Analysis**: Initial exploratory data analysis (EDA) will be performed to understand the main features of the dataset. Statistical measures such as mean, median, and standard deviation will be calculated to summarize the data.
3. **Correlation Analysis**: Pearson correlation coefficients will be computed to quantify the relationships between different variables. Scatter plots and heatmaps will be used to visualize these correlations.
4. **Data Visualization**: Various charts and plots, including histograms, bar charts, and scatter plots, will be created to visually represent the findings. These visualizations will help in identifying trends and patterns in the data.
5. **Predictive Modeling (if applicable)**: Machine learning algorithms, such as linear regression and decision trees, will be employed to develop predictive models. The performance of these models will be evaluated using metrics such as R-squared and mean absolute error.
6. **Insights and Recommendations**: Based on the analysis, key insights will be extracted, and recommendations will be formulated to help improve student performance. These recommendations will be aimed at students, educators, and policymakers.

Expected Outcomes :

The expected outcomes of this project include:

1. A detailed understanding of the distribution of student exam scores and the factors that influence them.
2. Identification of key demographic and study habit factors that significantly impact student performance.
3. Visual representations of the data that clearly illustrate the relationships between different variables.
4. Predictive models that can estimate student performance based on input factors (if applicable).
5. Actionable insights and recommendations to improve educational outcomes for students.

Significance :

This project is significant for several reasons:

1. **Educational Impact**: By identifying the factors that influence student performance, this project can help educators and policymakers design more effective educational interventions and policies.
2. **Data-Driven Decision Making**: The insights derived from this analysis can inform data-driven decision-making processes in educational institutions.
3. **Student Support**: The findings can help students understand how their study habits and demographic characteristics affect their academic performance, enabling them to make informed choices about their learning strategies.

Structure of the Report :

The report is structured as follows:

1. **Introduction**: Provides an overview of the project, including its background, motivation, objectives, methodology, expected outcomes, and significance.
2. **Literature Review**: Reviews relevant literature and previous studies on student performance analysis.
3. **Data Collection and Preparation**: Describes the dataset used in the analysis and the preprocessing steps taken.
4. **Descriptive Analysis**: Presents the results of the initial exploratory data analysis.
5. **Correlation Analysis**: Discusses the correlations between different variables and their impact on student performance.
6. **Data Visualization**: Shows the visual representations of the data and findings.
7. **Predictive Modeling (if applicable)**: Details the development and evaluation of predictive models.
8. **Insights and Recommendations**: Provides key insights from the analysis and recommendations for improving student performance.
9. **Conclusion**: Summarizes the main findings of the project and suggests directions for future research.

Data Collection

[1]:
```
pip install seaborn==0.13.2
```

[2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
sns.set_style('darkgrid')
sns.set_palette('pastel')
```

[4]:
```
data = pd.read_csv("/kaggle/input/student-performance-data/student_data.csv")
data.head()
```

[4]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10

5 rows × 33 columns

Exploring Data

```
[5]: print(data.columns)
```

```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

1. **school:** School attended
2. **sex:** Gender
3. **age:** Age of student
4. **address:** Type of address (urban or rural)
5. **famsize:** Family size
6. **Pstatus:** Parent's cohabitation status
7. **Medu:** Mother's education level
8. **Fedu:** Father's education level
9. **Mjob:** Mother's job
10. **Fjob:** Father's job
11. **reason:** Reason for choosing school
12. **guardian:** Student's guardian
13. **traveltime:** Travel time to school
14. **studytime:** Weekly study time
15. **failures:** Number of past class failures
16. **schoolsup:** Extra educational support
17. **famsup:** Family educational support
18. **paid:** Extra paid classes
19. **activities:** Extra-curricular activities
20. **nursery:** Attended nursery school
21. **higher:** Wants to pursue higher education
22. **internet:** Internet access at home
23. **romantic:** In a romantic relationship
24. **Famrel:** Quality of family relationships
25. **freetime:** Free time after school
26. **goout:** Going out with friends
27. **Dalc:** Workday alcohol consumption
28. **Walc:** Weekend alcohol consumption
29. **health:** Current health status
30. **absences:** Number of school absences
31. **G1:** Grade 1
32. **G2:** Grade 2
33. **G3:** Final grade

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
```

```
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```

[7]:
```
data.isnull().sum()
```

[7]:
```
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
```

```
freetime     0
goout        0
Dalc         0
Walc         0
health       0
absences     0
G1           0
G2           0
G3           0
dtype: int64
```

[8]:
```
cat_cols = data.select_dtypes(include=[object]).columns
cat_cols
```

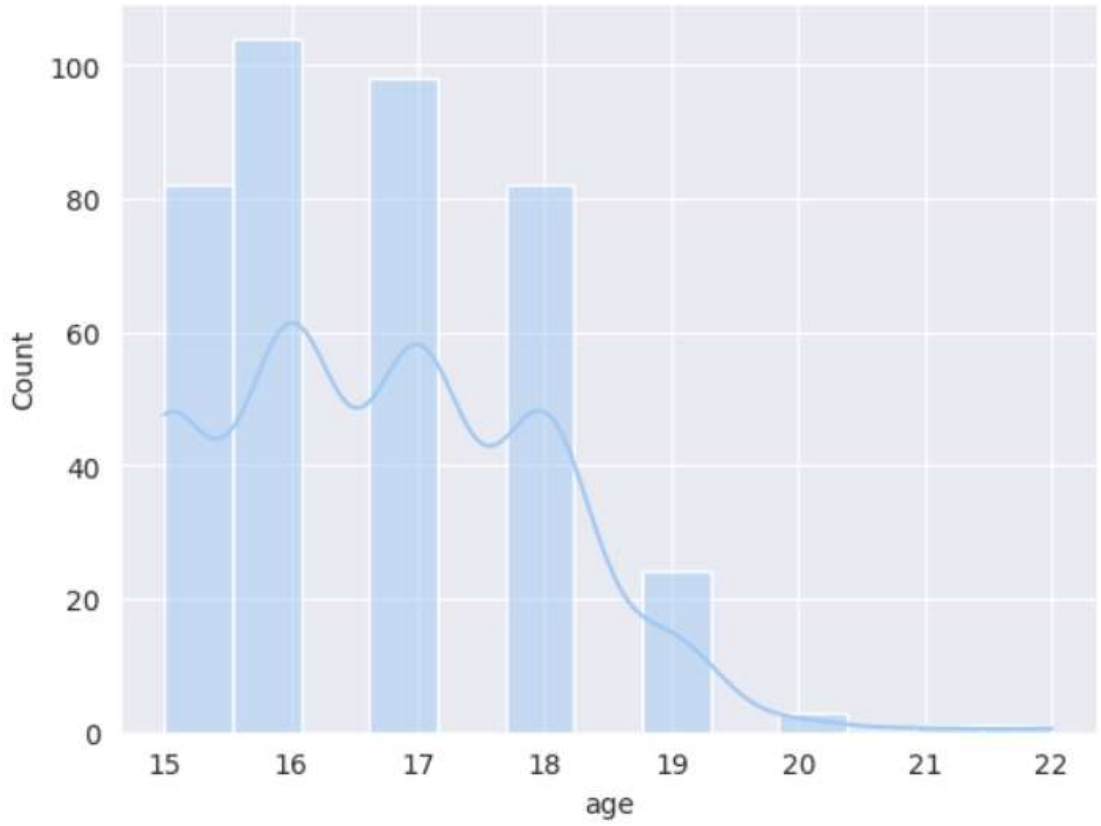
[8]:
```
Index(['school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob',
       'reason', 'guardian', 'schoolsup', 'famsup', 'paid', 'activities',
       'nursery', 'higher', 'internet', 'romantic'],
      dtype='object')
```

[9]:
```
num_cols = data.select_dtypes(include=['number']).columns
num_cols
```

[9]:
```
Index(['age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'famrel',
       'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'G1', 'G2',
       'G3'],
      dtype='object')
```

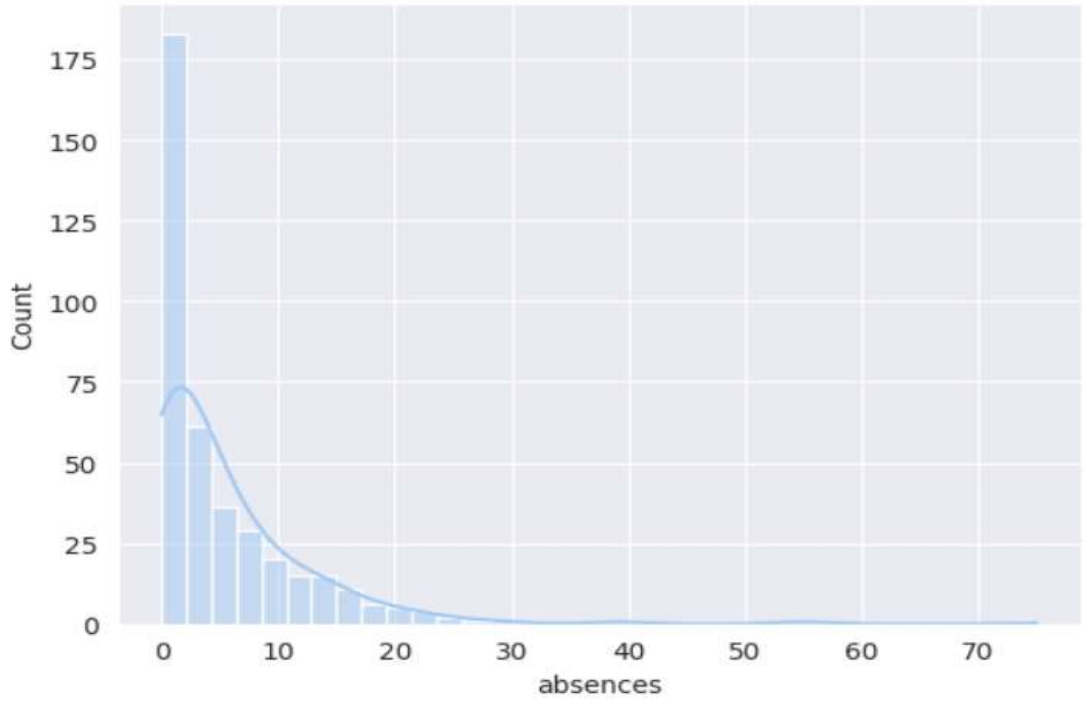
[10]:

```
sns.histplot(x=data['age'], kde=True);
plt.savefig("age.png")
```



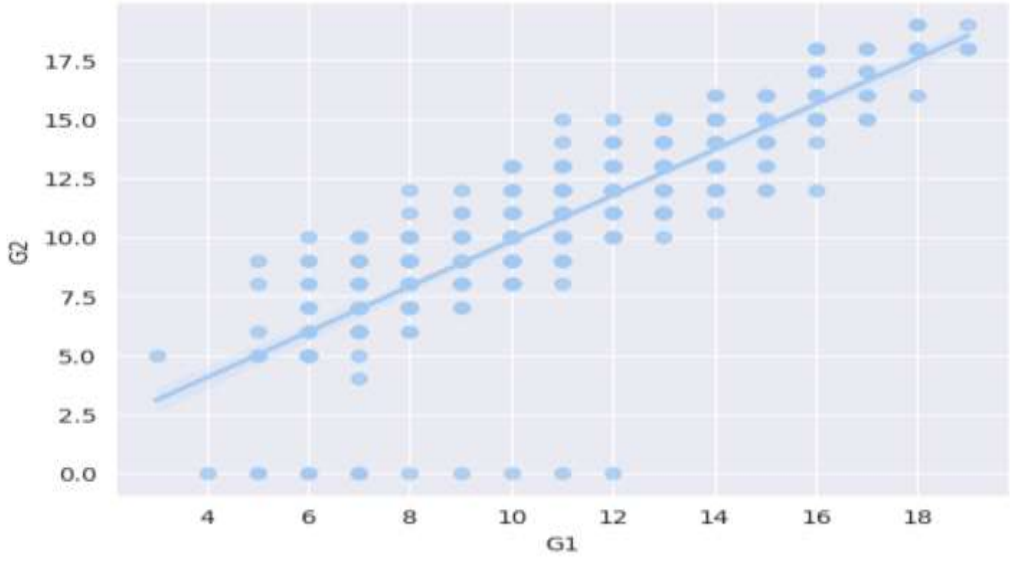
[11]:

```
sns.histplot(x=data['absences'], kde=True);
plt.savefig("absence.png")
```



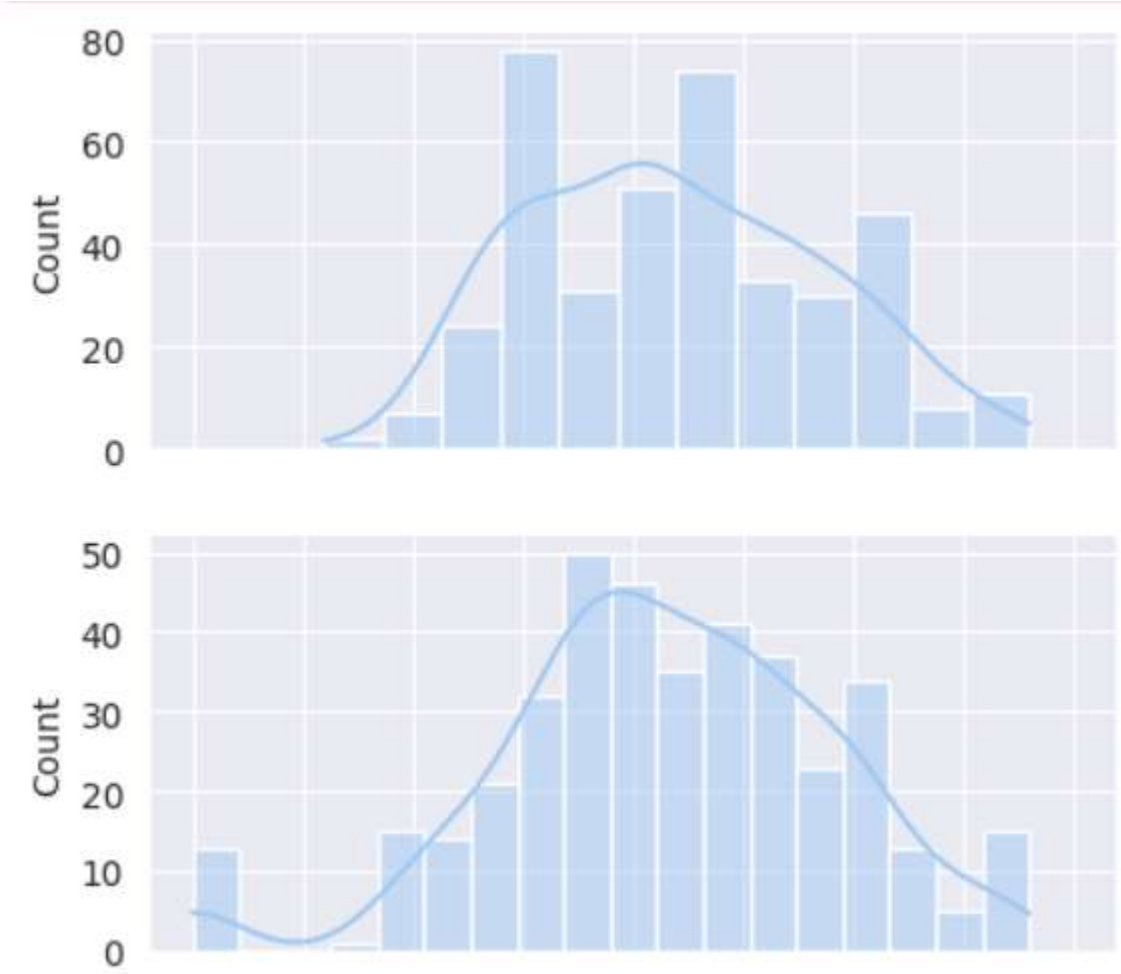
[12]:

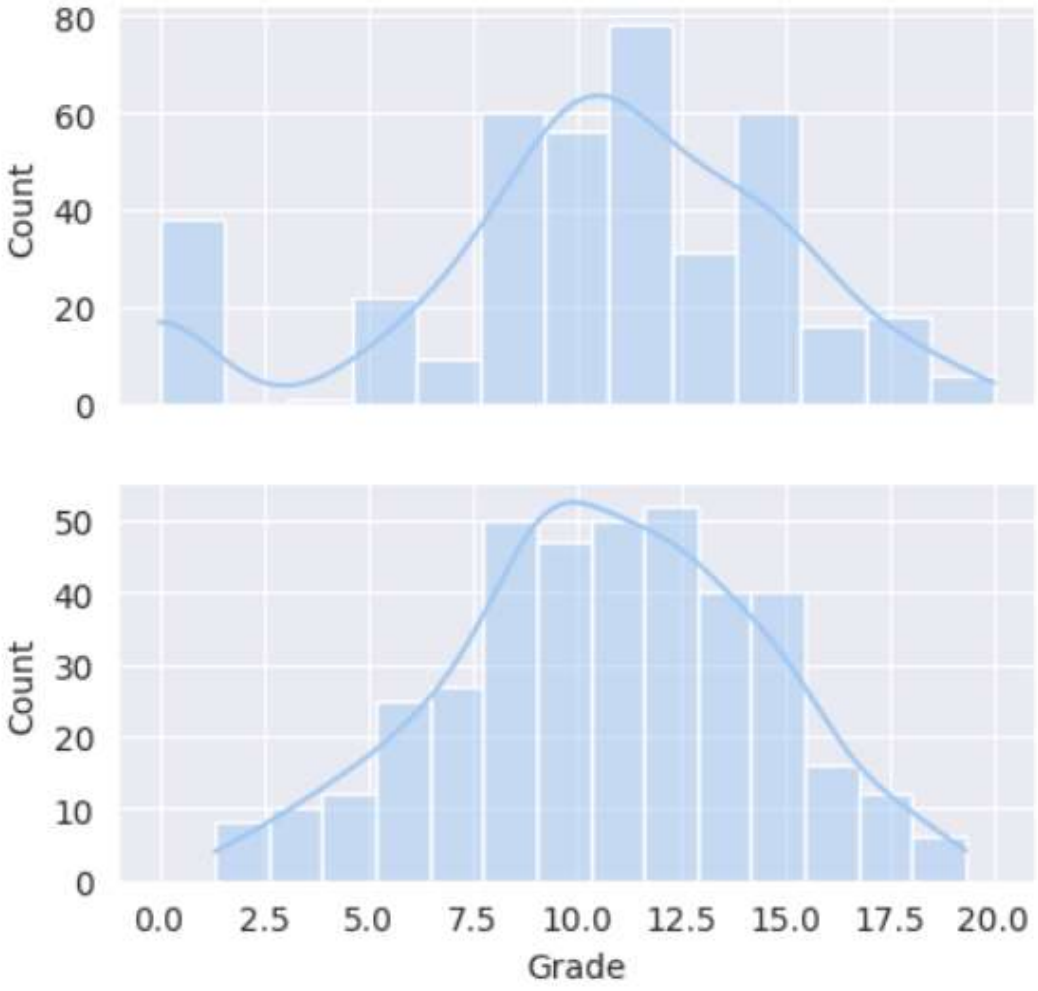
```
sns.regplot(data=data, x='G1', y='G2')
plt.savefig("G1G2 regplot.png")
```



[13]:
```
data['Grade'] = (data["G1"]+data["G2"]+data["G3"])/3
```

[14]:
```
fig, axes = plt.subplots(4, 1, sharex=True, figsize=(5, 10))
for col, ax in zip(['G1', 'G2', 'G3', 'Grade'], axes):
    sns.histplot(x=data[col], ax=ax, kde=True);
plt.savefig("Grades' distribution.png")
```



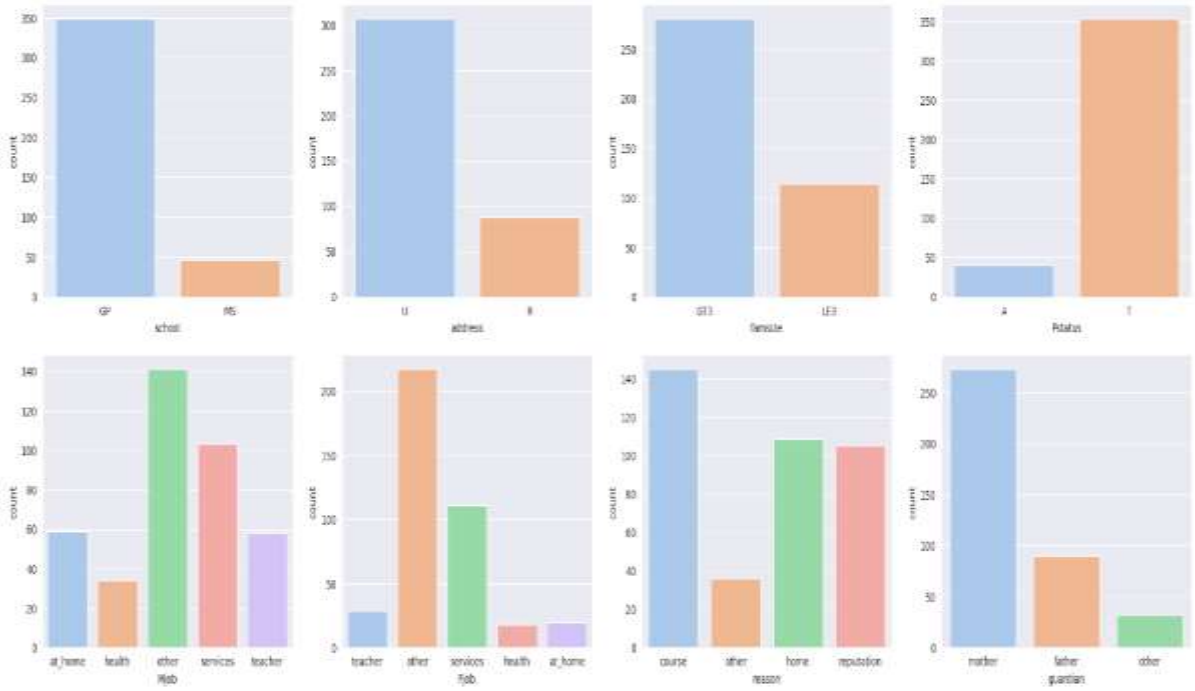


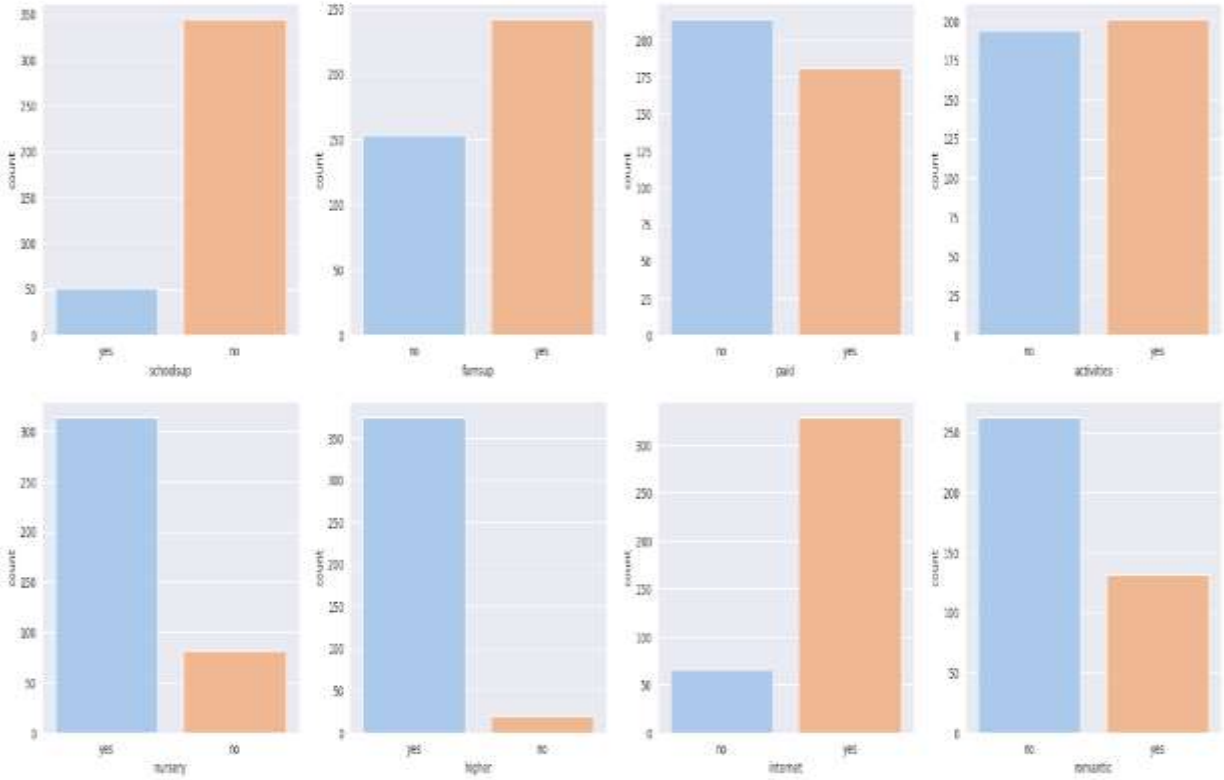
[15]:
```
data = data.drop(["G1", "G2", "G3"], axis=1)
```

[16]:
```
data[cat_cols].nunique()
```

[16]:
```
school        2
sex           2
address       2
famsize       2
Pstatus       2
Mjob          5
Fjob          5
reason        4
guardian      3
schoolsup     2
famsup        2
paid          2
activities    2
nursery       2
higher        2
internet      2
romantic      2
dtype: int64
```

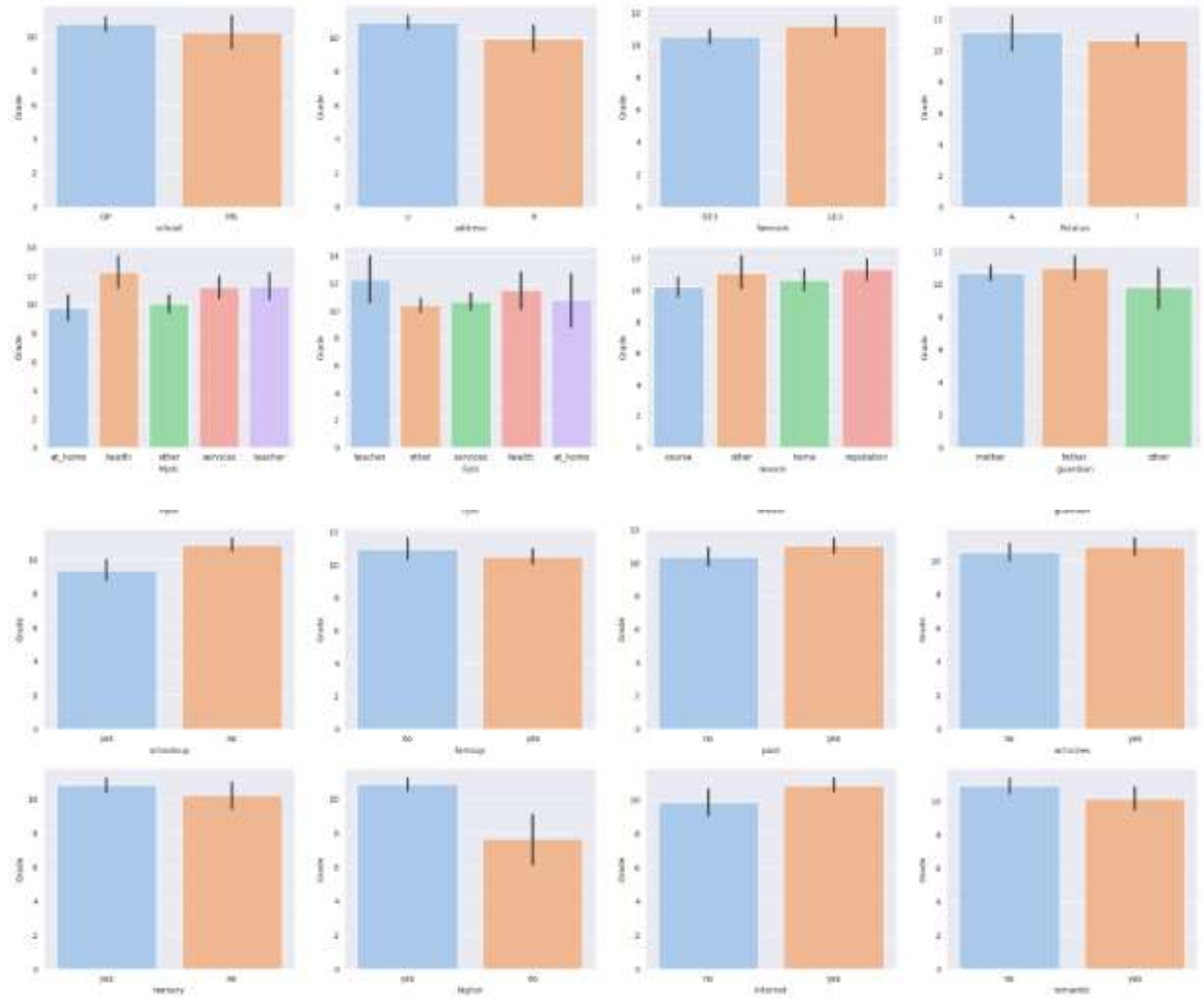
[17]:
```
nrows, ncols = 4, 4
fig, axes = plt.subplots(nrows, ncols, figsize=(25, 20))
i, j = 0, 0
for col in cat_cols:
    if col == "sex":
        continue
    ax = axes[i // ncols][j % ncols]
    sns.countplot(data=data, x=col, ax=ax)
    i += 1
    j += 1
plt.savefig("categories.png")
```


count
GP
MS
school
U
R
address
LE3
GT3
famsize
A
T
Pstatus
at_home
health
other
services
teacher
Mjob
Fjob
reason
course
home
reputation
guardian
mother
father


count
yes
no
schoolsup
famsup
paid
activities
nursery
higher
internet
romantic

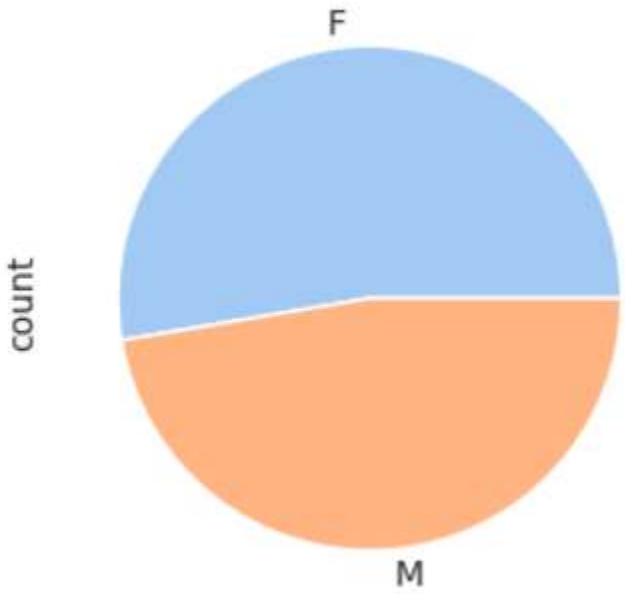
[18]:

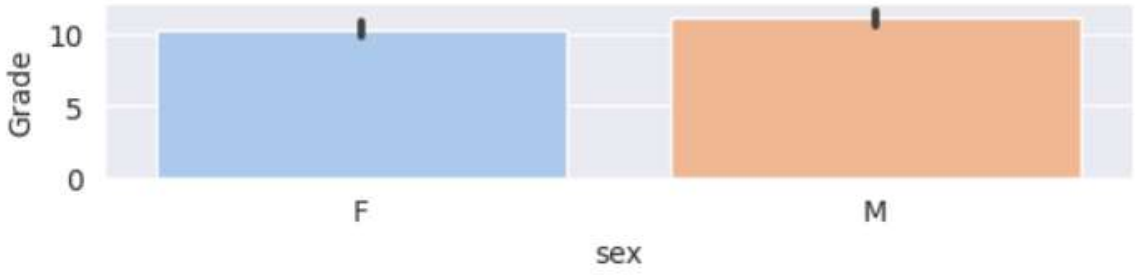
```
nrows, ncols = 4, 4
fig, axes = plt.subplots(nrows, ncols, figsize=(25, 20))
i, j = 0, 0
for col in cat_cols:
    if col == "sex":
        continue
    ax = axes[i // ncols][j % ncols]
    sns.barplot(data=data, x=col, y='Grade', ax=ax)
    i += 1
    j += 1
plt.savefig("barplot categories.png")
```



[19]:

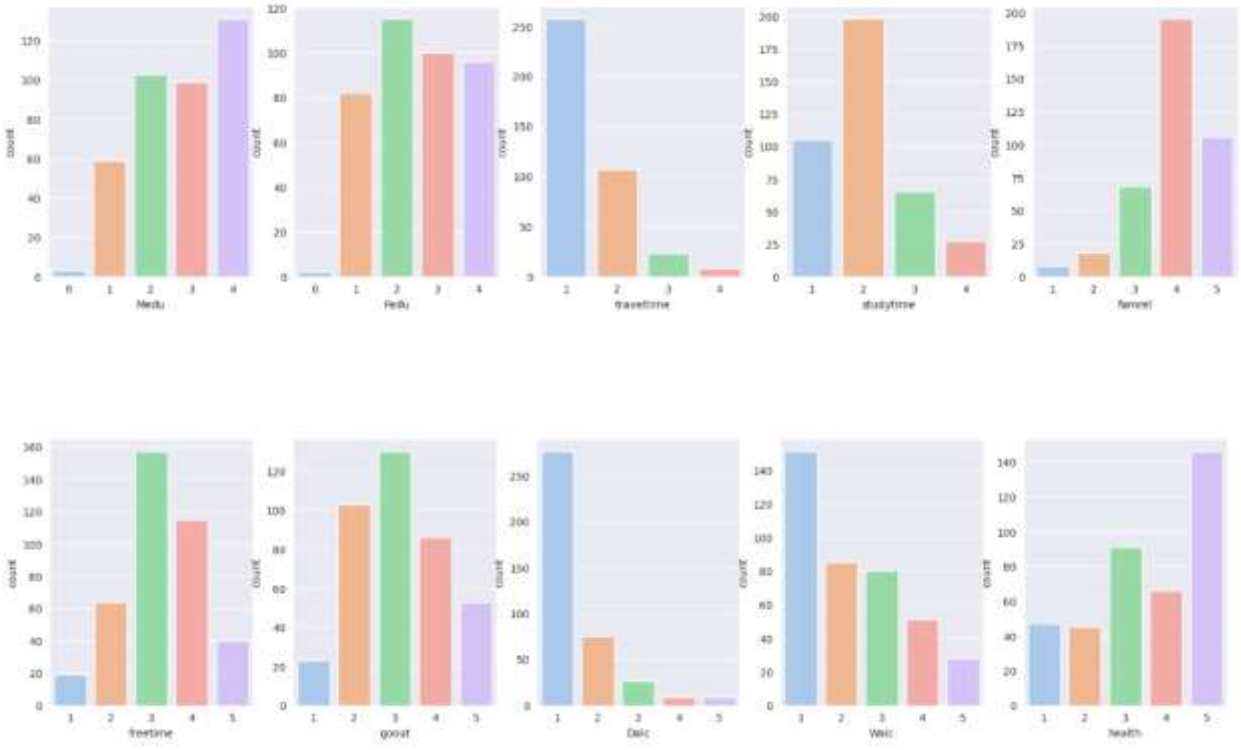
```
fig, axes = plt.subplots(2, 1, height_ratios=[0.75, 0.25])
data["sex"].value_counts().plot(kind='pie', ax=axes[0])
sns.barplot(data=data, x='sex', y='Grade', ax=axes[1])
plt.savefig("sex.png");
```





[20]:
```
num_cat_cols = ['Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'famrel',
       'freetime', 'goout', 'Dalc', 'Walc', 'health']
```

[21]:
```
nrows, ncols = 2, 5
i, j = 0, 0
fig, axes = plt.subplots(nrows, ncols, figsize=(20, 10))
for col in num_cat_cols:
    if col == 'failures':
        continue
    ax = axes[i//ncols][j%ncols]
    sns.countplot(x=data[col], ax=ax);
    i += 1
    j += 1
plt.savefig("numerical categories.png")
```



[22]:

```
nrows, ncols = 2, 5
i, j = 0, 0
fig, axes = plt.subplots(nrows, ncols, figsize=(20, 10))
for col in num_cat_cols:
    if col == 'failures':
        continue
    ax = axes[i//ncols][j%ncols]
    sns.barplot(data=data, x=col, y="Grade", ax=ax);
    i += 1
    j += 1
plt.savefig("barplot numerical categories")
```