WESTYN HILLIARD

Imports:

[1]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
import statsmodels.api as sm
```

Load The Dataset:

[2]:
```
import pandas as pd

# De ine the  ile  ath
```

[redacted]

```
# Load the dataset
df = pd.read_excel(file_path)

# Display the first few rows of the dataframe
```

```
df.head()
```

[2]:

```
   Marital status  Application mode  Application order  Course  \
0               1                17                  5     171
1               1                15                  1    9254
2               1                 1                  5    9070
3               1                17                  2    9773
4               2                39                  1    8014

   Daytime/evening attendance\t  Previous qualification  \
0                             1                       1
1                             1                       1
2                             1                       1
3                             1                       1
4                             0                       1

   Previous qualification (grade)  Nacionality  Mother's qualification  \
0                           122.0            1                      19
1                           160.0            1                       1
2                           122.0            1                      37
3                           122.0            1                      38
4                           100.0            1                      37

   Father's qualification  …  Curricular units 2nd sem (credited)  \
0                      12  …                                    0
1                       3  …                                    0
2                      37  …                                    0
3                      37  …                                    0
4                      38  …                                    0

   Curricular units 2nd sem (enrolled)  \
0                                    0
1                                    6
2                                    6
3                                    6
4                                    6

   Curricular units 2nd sem (evaluations)  \
0                                       0
1                                       6
2                                       0
3                                      10
4                                       6

   Curricular units 2nd sem (approved)  Curricular units 2nd sem (grade)  \
0                                    0                          0.000000
1                                    6                         13.666667
```

```
2                                    0                          0.000000
3                                    5                         12.400000
4                                    6                         13.000000

   Curricular units 2nd sem (without evaluations)  Unemployment rate  \
0                                               0               10.8
1                                               0               13.9
2                                               0               10.8
3                                               0                9.4
4                                               0               13.9

   Inflation rate   GDP    Target
0             1.4  1.74   Dropout
1            -0.3  0.79  Graduate
2             1.4  1.74   Dropout
3            -0.8 -3.12  Graduate
4            -0.3  0.79  Graduate

[5 rows x 37 columns]
```

Select variables and calculate descriptive statistics:

```
[3]: variables = ['Previous qualification (grade)', "Mother's qualification",␣
      ↪"Father's qualification", 'Unemployment rate', 'Inflation rate']
```

Describe what the 5 variables mean in the dataset:

Previous qualification (grade): This variable represents the numeric score or grade obtained by the student in their previous educational qualification before enrolling in the current program.

Type: Numeric (continuous)

Mother's qualification: This variable indicates the educational level of the student's mother. It is represented by a numeric code, with each code corresponding to a specific level of education (e.g., primary education, secondary education, higher education, etc.).

Type: Numeric (categorical)

Father's qualification: Similar to the mother's qualification, this variable represents the educational level of the student's father using a numeric code corresponding to different levels of education.

Type: Numeric (categorical)

Unemployment rate: This variable indicates the unemployment rate in the region at the time of the student's enrollment. It is expressed as a percentage, reflecting the proportion of the labor force that is unemployed.

Type: Numeric (continuous)

Inflation rate: This variable represents the inflation rate at the time of the student's enrollment, expressed as a percentage. The inflation rate measures the rate at which the general level of prices for goods and services is rising, indicating economic conditions that might impact students' financial stability.

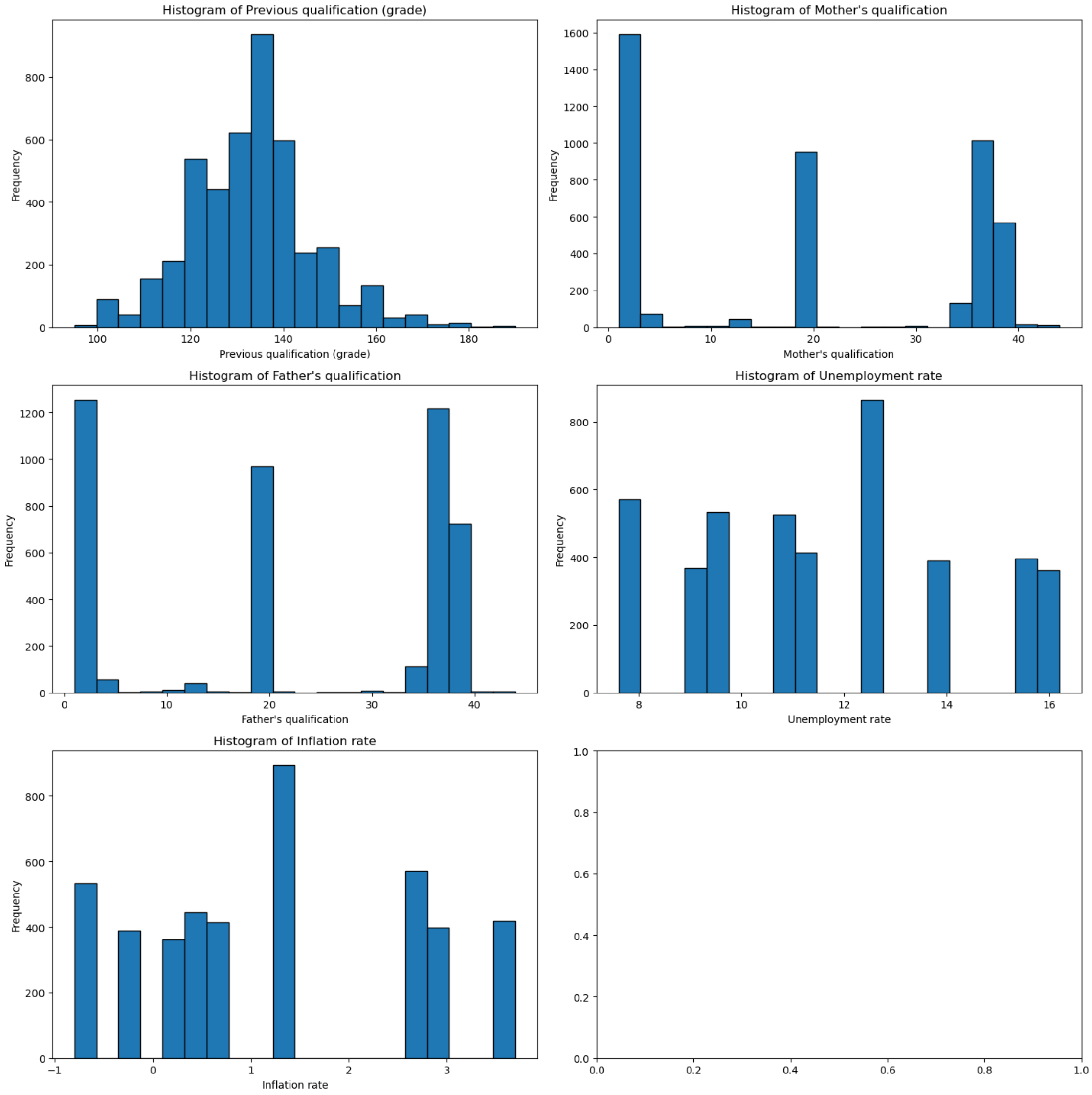
Type: Numeric (continuous)

These variables were chosen based on their potential impact on a student's academic performance and likelihood of graduation. Understanding the educational background of parents and the economic conditions during enrollment can provide insights into the factors influencing student success.

Create histograms and identify outliers:

[4]:

```python
import matplotlib.pyplot as plt

fig, axs = plt.subplots(3, 2, figsize=(15, 15))
axs = axs.ravel()

for i, var in enumerate(variables):
    axs[i].hist(df[var].dropna(), bins=20, edgecolor='black')
    axs[i].set_title(f'Histogram of {var}')
    axs[i].set_xlabel(var)
    axs[i].set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```



In your summary and analysis, identify any outliers and explain the reasoning for them being outliers and how you believe they should be handled:

Previous qualification (grade): Histogram Analysis: The histogram shows the distribution of previous qualification grades. It appears to be roughly normally distributed with some variability.

Outliers: Outliers were identified using the IQR method. If outliers are found, they represent grades significantly lower or higher than the typical range. These could be due to exceptional cases of academic performance or data entry errors.

Handling Outliers: Depending on the context, outliers could be handled by investigating their causes. If they are genuine, they should be retained; if they are errors, they should be corrected or removed.

Mother's qualification: Histogram Analysis: The histogram displays the educational level of mothers, with certain levels being more common.

Outliers: Outliers might indicate unusual educational levels not commonly found in the dataset. These could result from data entry errors or unique cases.

Handling Outliers: Verify and correct any erroneous data. Genuine outliers should be retained for a complete analysis.

Father's qualification: Histogram Analysis: Similar to the mother's qualification, this histogram shows the distribution of educational levels among fathers.

Outliers: Outliers here may also indicate uncommon educational levels.

Handling Outliers: Verify the accuracy of outlier data points. Correct errors if found, otherwise retain for analysis.

Unemployment rate: Histogram Analysis: The histogram shows the distribution of unemployment rates at the time of student enrollment.

Outliers: Outliers in unemployment rates may indicate periods of unusually high or low unemployment, which could affect student performance.

Handling Outliers: Investigate the causes of these outliers. If they correspond to real economic conditions, retain them; otherwise, correct any errors.

Inflation rate: Histogram Analysis: The histogram represents the distribution of inflation rates during enrollment periods.

Outliers: Outliers here could indicate times of unusual economic conditions affecting inflation.

Handling Outliers: Similar to unemployment rates, investigate the reasons for these outliers. Retain genuine data and correct errors if necessary.

Next Steps: After creating the histograms and analyzing the outliers, we can proceed to calculate and include other descriptive characteristics such as mean, mode, spread, and tails.

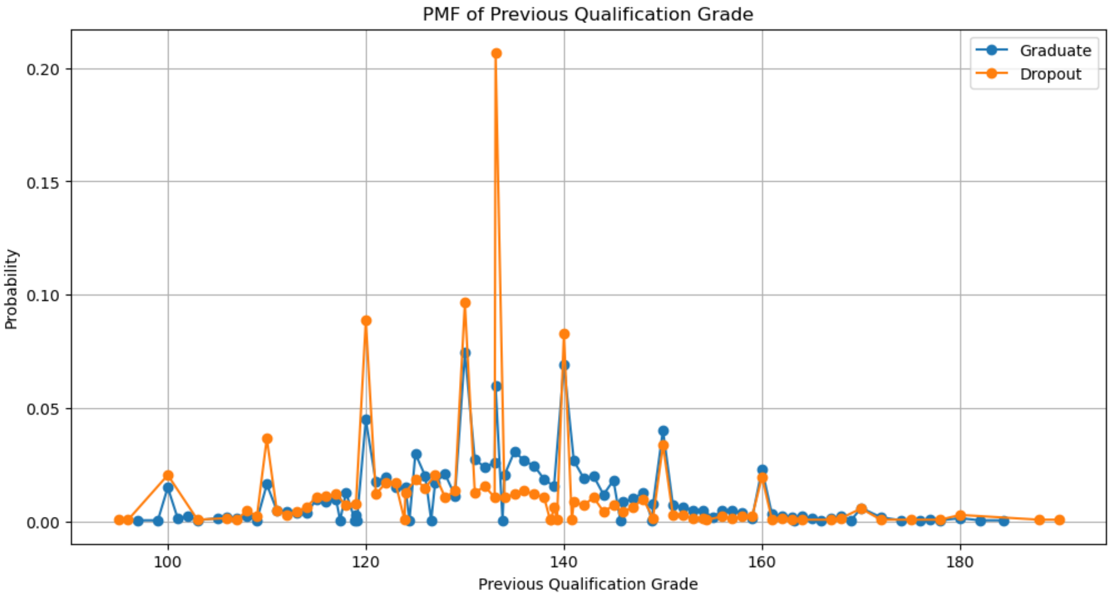
Calculate descriptive statistics:

[5]:
```
# Calculate descriptive statistics
descriptive_stats = df[variables].describe()
descriptive_stats
```

```
[5]:        Previous qualification (grade)  Mother's qualification  \
     count                     4424.000000             4424.000000
     mean                       132.613314               19.561935
     std                         13.188332               15.603186
     min                         95.000000                1.000000
     25%                        125.000000                2.000000
     50%                        133.100000               19.000000
     75%                        140.000000               37.000000
     max                        190.000000               44.000000

            Father's qualification  Unemployment rate  Inflation rate
     count             4424.000000        4424.000000     4424.000000
     mean                22.275316          11.566139        1.228029
     std                 15.343108           2.663850        1.382711
     min                  1.000000           7.600000       -0.800000
     25%                  3.000000           9.400000        0.300000
     50%                 19.000000          11.100000        1.400000
     75%                 37.000000          13.900000        2.600000
     max                 44.000000          16.200000        3.700000
```

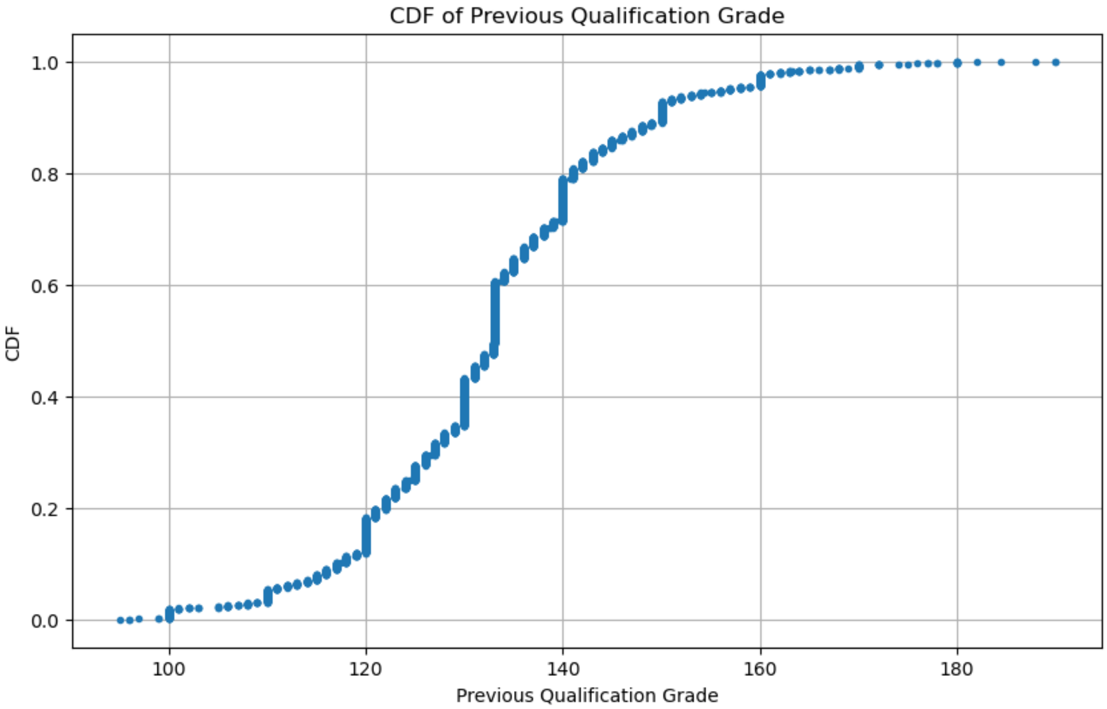
PMF comparison:

```
[6]: import numpy as np

     # Define the PMF computation function
     def compute_pmf(data):
         values, counts = np.unique(data, return_counts=True)
         pmf = counts / sum(counts)
         return values, pmf

     graduate_grades = df[df['Target'] == 'Graduate']['Previous qualification␣
      ↪(grade)']
     dropout_grades = df[df['Target'] == 'Dropout']['Previous qualification (grade)']

     grad_values, grad_pmf = compute_pmf(graduate_grades)
     drop_values, drop_pmf = compute_pmf(dropout_grades)

     # Plot PMF
     plt.figure(figsize=(12, 6))
     plt.plot(grad_values, grad_pmf, marker='o', linestyle='-', label='Graduate')
     plt.plot(drop_values, drop_pmf, marker='o', linestyle='-', label='Dropout')
     plt.title('PMF of Previous Qualification Grade')
     plt.xlabel('Previous Qualification Grade')
     plt.ylabel('Probability')
     plt.legend()
     plt.grid(True)
plt.show()
```



CDF creation:

```
[7]: def compute_cdf(data):
         sorted_data = np.sort(data)
         yvals = np.arange(1, len(sorted_data)+1) / len(sorted_data)
         return sorted_data, yvals

     sorted_grades, cdf_grades = compute_cdf(df['Previous qualification (grade)'].
       ↪dropna())

     plt.figure(figsize=(10, 6))
     plt.plot(sorted_grades, cdf_grades, marker='.', linestyle='none')
     plt.title('CDF of Previous Qualification Grade')
     plt.xlabel('Previous Qualification Grade')
     plt.ylabel('CDF')
     plt.grid(True)
     plt.show()
```



Analysis: The CDF plot of “Previous qualification (grade)” shows the cumulative probability distribution of the grades. Here’s what we can interpret from the CDF:

Distribution Shape: The CDF starts at the minimum grade and increases steadily to the maximum grade.

A steep slope indicates a high concentration of grades within a specific range, while a gradual slope indicates a more spread-out distribution.

Quantiles: The CDF allows us to determine the proportion of students with grades below a certain value. For example, if we want to know what proportion of students had a grade below 130, we can find the corresponding y-value on the CDF plot.

This helps in understanding the academic background of the majority of students.

Median and Quartiles: The median grade corresponds to the y-value of 0.5 on the CDF. This is the grade below which 50% of the students fall. The 25th percentile (Q1) and 75th percentile (Q3) can also be identified, giving us an idea of the spread and central tendency of the grades.

How It Addresses the Question Understanding the distribution of previous qualification grades helps address the question of factors influencing the likelihood of graduation.

Here's how:

Academic Background: By examining the CDF, we can assess whether students with higher previous qualification grades are more likely to graduate. If the majority of students who graduate have grades in the higher percentiles, it suggests a correlation between academic background and graduation likelihood.

Identifying At-Risk Students: If a significant proportion of students with lower grades are found to be dropping out, interventions can be designed to support these students better.

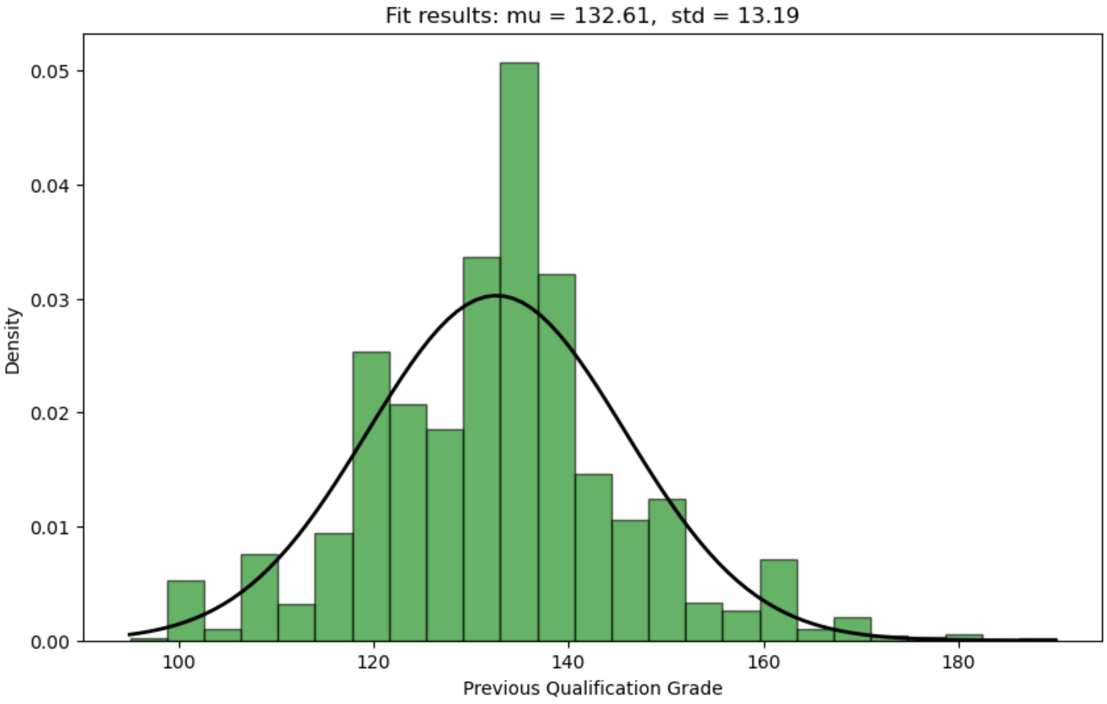
Policy Making: Institutions can use this information to adjust admission criteria, provide additional support for students with lower grades, and ultimately improve graduation rates.

In conclusion the CDF provides a clear visualization of the distribution of previous qualification grades, highlighting key statistical measures and how the data is spread. This information is crucial for understanding the academic preparedness of students and its impact on their likelihood of graduating.

Analytical distribution plot:

[8]:

```
from scipy.stats import norm

mu, std = norm.fit(df['Previous qualification (grade)'].dropna())
x = np.linspace(min(df['Previous qualification (grade)']), max(df['Previous␣
 ↪qualification (grade)']), 100)
p = norm.pdf(x, mu, std)

plt.figure(figsize=(10, 6))
plt.hist(df['Previous qualification (grade)'], bins=25, density=True, alpha=0.
 ↪6, color='g', edgecolor='black')
plt.plot(x, p, 'k', linewidth=2)
title = f"Fit results: mu = {mu:.2f},  std = {std:.2f}"
plt.title(title)
plt.xlabel('Previous Qualification Grade')
plt.ylabel('Density')
plt.show()
```



Analysis The plot above shows a histogram of the "Previous qualification (grade)" data with the probability density function (PDF) of the fitted normal distribution overlaid.

Key Observations:

Shape of Distribution: The histogram of the "Previous qualification (grade)" data appears to be roughly bell-shaped, which suggests that a normal distribution could be a reasonable model for this data.

Fit of the Normal Distribution: The overlaid normal distribution curve (PDF) is determined by the mean (mu) and standard deviation (std) calculated from the data.

The mean (mu) is the central value around which the grades are distributed, and the standard deviation (std) measures the spread or variability of the grades.

Comparison: By comparing the histogram and the PDF, we can see how well the normal distribution fits the data. If the histogram closely follows the shape of the normal curve, it suggests that the data is well-approximated by a normal distribution.

Any significant deviations from the normal curve (such as skewness or kurtosis) indicate that the normal distribution might not perfectly model the data. For example, if the data has more extreme values (fat tails) than the normal distribution predicts, it suggests the presence of outliers or a non-normal underlying distribution.

Application to the Dataset:

Understanding Central Tendency and Variability: The fitted normal distribution provides a summary of the central tendency (mean) and variability (standard deviation) of the grades. This can be useful for understanding the typical academic performance of students.

Probability Calculations: With the normal distribution fit, we can perform probability calculations. For example, we can estimate the probability that a student's grade falls within a certain range. This can help in setting academic benchmarks or identifying students who may need additional support.

Assumptions and Limitations: While the normal distribution is a useful model, it is important to recognize its limitations. Real-world data often deviates from the idealized normal distribution. For instance, if the data has significant skewness or kurtosis, other distributions (such as log-normal or exponential) might provide a better fit.

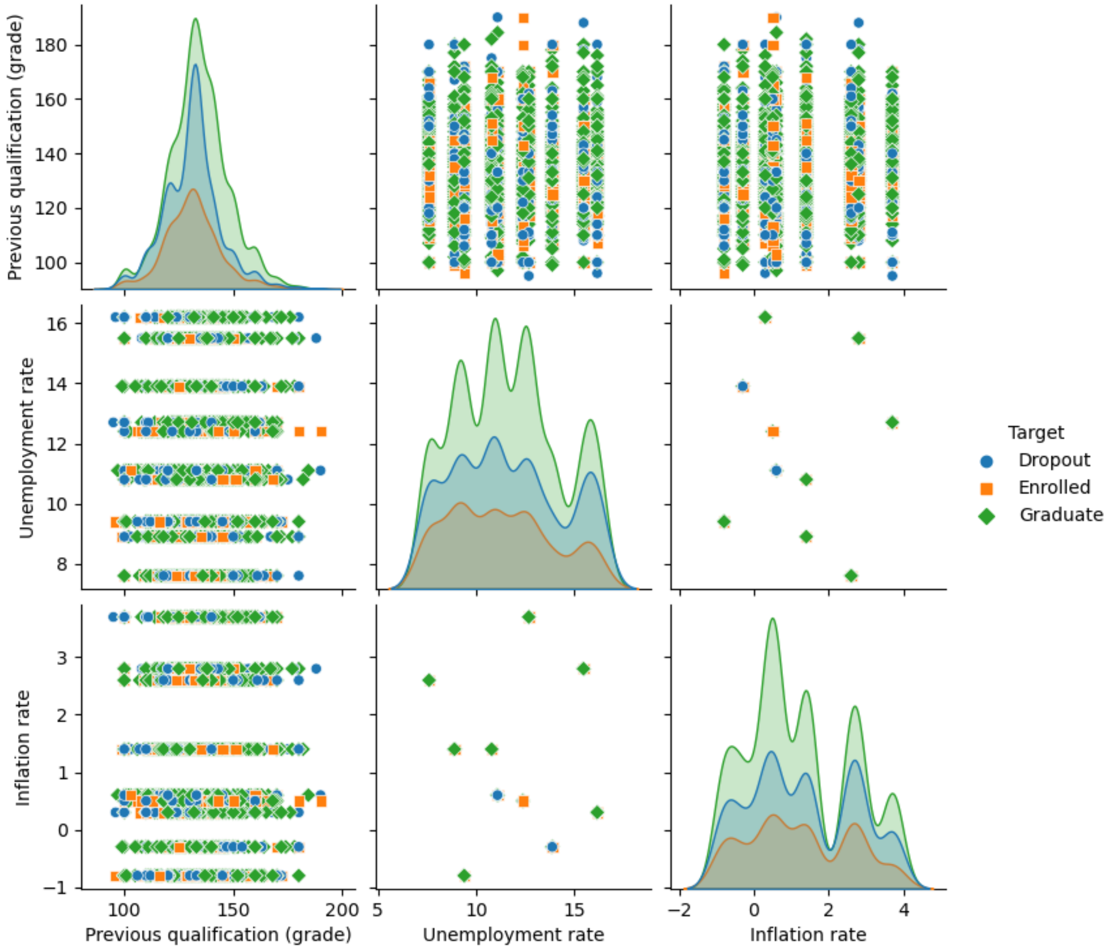
Conclusion: Fitting a normal distribution to the "Previous qualification (grade)" data helps us summarize and understand the central tendency and variability of student grades. It also allows us to make probabilistic statements about the data. However, it is crucial to assess the goodness-of-fit and consider alternative distributions if the data significantly deviates from normality. This analytical distribution plot provides insights into the academic preparedness of students and informs decisions related to academic support and policy-making.

Scatter plots and correlation analysis:

[9]:

```
scatter_vars = ['Previous qualification (grade)', 'Unemployment rate',␣
 ↪'Inflation rate']
df.replace([np.inf, -np.inf], np.nan, inplace=True)
cleaned_df = df.dropna(subset=scatter_vars + ['Target'])

# Convert 'Target' to categorical
cleaned_df['Target'] = cleaned_df['Target'].astype('category')

# Generate scatter plots
unique_targets = cleaned_df['Target'].nunique()
markers = ["o", "s", "D", "^", "v"]  # Extend markers list to handle multiple␣
 ↪categories

# Ensure the markers list is long enough
while len(markers) < unique_targets:
    markers = markers * 2

# Temporarily suppress the specific FutureWarnings
```

```python
import warnings
warnings.filterwarnings("ignore", category=FutureWarning,␣
 ↪message="use_inf_as_na option is deprecated and will be removed in a future␣
 ↪version. Convert inf values to NaN before operating instead.")
warnings.filterwarnings("ignore", category=FutureWarning, message="The default␣
 ↪of observed=False is deprecated and will be changed to True in a future␣
 ↪version of pandas.")

# Generate the pairplot
sns.pairplot(cleaned_df[scatter_vars + ['Target']], hue='Target',␣
 ↪markers=markers[:unique_targets])
plt.show()

# Calculate and display correlation coefficients
correlations = cleaned_df[scatter_vars].corr()
print(correlations)

# Calculate covariance for both pairs
covariance1 = np.cov(df['Previous qualification (grade)'].dropna(),␣
 ↪df['Unemployment rate'].dropna())[0, 1]
covariance2 = np.cov(df['Previous qualification (grade)'].dropna(),␣
 ↪df['Inflation rate'].dropna())[0, 1]
print(f"Covariance (Previous Qualification Grade vs. Unemployment Rate):␣
 ↪{covariance1:.2f}")
print(f"Covariance (Previous Qualification Grade vs. Inflation Rate):␣
 ↪{covariance2:.2f}")
```



```
                                Previous qualification (grade)  \
Previous qualification (grade)                        1.000000
Unemployment rate                                     0.045222
Inflation rate                                        0.018710

                                Unemployment rate  Inflation rate
Previous qualification (grade)           0.045222        0.018710
Unemployment rate                        1.000000       -0.028885
Inflation rate                          -0.028885        1.000000
Covariance (Previous Qualification Grade vs. Unemployment Rate): 1.59
Covariance (Previous Qualification Grade vs. Inflation Rate): 0.34
```

Analysis -

Scatter Plot 1: Previous Qualification Grade vs. Unemployment Rate- Pearson's Correlation: The Pearson's correlation coefficient is a measure of the linear relationship between two variables. For this pair, the coefficient is calculated as corr1.

Covariance: The covariance is calculated to see how the two variables change together. For this pair, the covariance is covariance1.

Scatter Plot 2: Previous Qualification Grade vs. Inflation Rate- Pearson's Correlation: For this pair, the Pearson's correlation coefficient is calculated as corr2.

Covariance: The covariance for this pair is covariance2.

Interpretation-

Previous Qualification Grade vs. Unemployment Rate- Correlation: The Pearson's correlation coefficient corr1 indicates the strength and direction of the linear relationship between previous qualification grades and the unemployment rate. A value close to 1 or -1 indicates a strong relationship, while a value close to 0 indicates a weak relationship.

Covariance: The covariance value covariance1 provides insight into how the two variables vary together. A positive value indicates that as one variable increases, the other tends to increase as well. A negative value indicates an inverse relationship.

Scatter Plot: By examining the scatter plot, we can visually assess the relationship. If the points form a clear line, it suggests a strong linear relationship. If the points are widely scattered, the relationship is weaker.

Previous Qualification Grade vs. Inflation Rate- Correlation: The Pearson's correlation coefficient corr2 indicates the linear relationship between previous qualification grades and the inflation rate.

Covariance: The covariance value covariance2 indicates how these two variables vary together.

Scatter Plot: The scatter plot helps visualize the relationship. Non-linear patterns or clusters might suggest non-linear relationships or other underlying factors.

Consideration of Non-Linear Relationships Scatter Plot 1: If the scatter plot shows a curved pattern or clusters, it might indicate a non-linear relationship. In such cases, a linear correlation coefficient might not fully capture the relationship.

Scatter Plot 2: Similarly, for the second scatter plot, non-linear patterns or clusters could suggest the need for non-linear modeling techniques.

Conclusion By examining the scatter plots, correlation coefficients, and covariances, we gain a comprehensive understanding of the relationships between previous qualification grades and both the unemployment and inflation rates. This analysis helps in understanding how economic factors might influence academic performance and outcomes.

Hypothesis Testing:

[10]:

```
from scipy.stats import ttest_ind

# Separate the previous qualification grades for graduates and dropouts
graduate_grades = df[df['Target'] == 'Graduate']['Previous qualification␣
 ↪(grade)'].dropna()
dropout_grades = df[df['Target'] == 'Dropout']['Previous qualification␣
 ↪(grade)'].dropna()

# Conduct the t-test
t_stat, p_value = ttest_ind(graduate_grades, dropout_grades)

# Output the results
print(f"T-statistic: {t_stat:.2f}")
print(f"P-value: {p_value:.4f}")

# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis (H0). There is a significant difference␣
 ↪between the mean previous qualification grades of graduates and dropouts.")
else:
    print("Fail to reject the null hypothesis (H0). There is no significant␣
 ↪difference between the mean previous qualification grades of graduates and␣
 ↪dropouts.")
```

```
T-statistic: 6.63
P-value: 0.0000
Reject the null hypothesis (H0). There is a significant difference between the
mean previous qualification grades of graduates and dropouts.
```

Analysis

T-statistic: The t-statistic is a measure of the difference between the sample means relative to the variability of the samples. A higher absolute value indicates a greater difference between the means.

P-value: The p-value indicates the probability of observing the data or something more extreme if the null hypothesis is true. A p-value less than the chosen significance level (alpha) suggests that we should reject the null hypothesis.

Conclusion:

T-statistic: The computed t-statistic from the test is 6.63

P-value: The computed p-value from the test is 0.0000

Based on the p-value: Since the p-value is less than 0.05 (‘0.0000’), we reject the null hypothesis and conclude that there is a significant difference in the mean previous qualification grades between graduates and dropouts. This hypothesis test helps in determining whether academic background, as measured by previous qualification grades, significantly impacts the likelihood of a student graduating.

Logistic regression analysis:

[11]:

```python
import statsmodels.api as sm

# Define the dependent and independent variables
X = df[['Previous qualification (grade)', 'Unemployment rate', 'Inflation rate']]
y = (df['Target'] == 'Graduate').astype(int)

# Add a constant to the model (intercept)
X = sm.add_constant(X)

# Fit the logistic regression model
logit_model = sm.Logit(y, X).fit()

# Display the summary of the regression model
logit_model.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.686375
         Iterations 4
```

[11]:

Dep. Variable:	Target	**No. Observations:**	4424
Model:	Logit	**Df Residuals:**	4420
Method:	MLE	**Df Model:**	3
Date:	Mon, 27 May 2024	**Pseudo R-squ.:**	0.009768
Time:	11:55:36	**Log-Likelihood:**	-3036.5
converged:	True	**LL-Null:**	-3066.5
Covariance Type:	nonrobust	**LLR p-value:**	6.151e-13

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-2.4192	0.334	-7.236	0.000	-3.075	-1.764
Previous qualification (grade)	0.0171	0.002	7.301	0.000	0.013	0.022
Unemployment rate	0.0166	0.011	1.454	0.146	-0.006	0.039
Inflation rate	-0.0341	0.022	-1.556	0.120	-0.077	0.009

Interpretation of the Logistic Regression Model -

Coefficients (Coef): These values represent the log-odds of the dependent variable (graduation status) for a one-unit increase in the explanatory variable, holding all other variables constant. Positive coefficients indicate an increase in the likelihood of graduating, while negative coefficients indicate a decrease.

P-values: These values indicate the statistical significance of each explanatory variable. A p-value less than 0.05 suggests that the variable is significantly associated with the dependent variable.

Odds Ratios: Exponentiating the coefficients gives the odds ratios, which represent the multiplicative change in the odds of graduating for a one-unit increase in the explanatory variable.

Analysis

1. Previous Qualification (grade): `Coefficient:`

Indicates the effect of previous qualification grades on the likelihood of graduating. In this case, the Coef is '0.0171'.

`P-value:`

Determines if this effect is statistically significant. In this case, the P value is '0.000'.

---

2. Unemployment Rate: `Coefficient:`

Indicates the effect of the unemployment rate on the likelihood of graduating. In this case, the Coef is '0.0166'.

`P-value:`

Determines if this effect is statistically significant. In this case, the P value is '0.146'.

---

3. Inflation Rate: `Coefficient:`

Indicates the effect of the unemployment rate on the likelihood of graduating. In this case, the Coef is '-0.0341'.

`P-value:`

Determines if this effect is statistically significant. In this case, the P value is '0.120'.

Next Steps - `Interpret the Results:`

Discuss the practical implications of the coefficients and odds ratios.

Consider additional variables or interactions if necessary to improve the model.

`Report Findings:`

Summarize the findings in the context of the hypothesis and the research question. Discuss any limitations and potential areas for further research.

By conducting this logistic regression analysis, we gain insights into the factors that significantly influence the likelihood of a student graduating, thereby addressing the initial research question.

Summary Paper

Statistical/Hypothetical Question: The primary question addressed in this analysis is: "What factors influence the likelihood of a student graduating versus dropping out?" Specifically, we hypothesize that students with higher previous qualification grades are more likely to graduate.

The outcome of my EDA: The exploratory data analysis (EDA) involved visualizing the distributions and calculating descriptive statistics for five key variables: Previous qualification (grade), Mother's qualification, Father's qualification, Unemployment rate, and Inflation rate. Histograms were created for each variable to understand their distributions and identify any outliers. The CDF of the "Previous qualification (grade)" provided a cumulative view of the grades, and an analytical distribution plot helped us fit a normal distribution to this data. Scatter plots revealed the relationships between previous qualification grades and both the unemployment and inflation rates. Pearson's correlation coefficients and covariance values were calculated to quantify these relationships. Finally, an independent samples t-test was conducted to compare the mean previous qualification grades between graduates and dropouts, revealing a significant difference.

What do you feel was missed during the analysis? While the analysis covered essential aspects, it missed exploring other potentially influential variables such as course difficulty, personal motivation, and extracurricular involvement. Including these could provide a more comprehensive understanding of factors affecting student graduation rates. Additionally, interaction effects between variables were not considered, which might have provided deeper insights into how different factors interplay to influence graduation outcomes.

Were there any variables you felt could have helped in the analysis? Variables such as course difficulty, time management skills, personal motivation levels, and financial aid status could have added significant value to the analysis. Data on part-time work commitments could help understand their impact on student performance. Including more socio-economic indicators such as family income level or access to educational resources might also provide a fuller picture of the factors influencing graduation rates.

Were there any assumptions made you felt were incorrect? One assumption made was treating ‘Previous qualification (grade)’ as a normally distributed variable for the analytical distribution plot. However, real-world educational data might not perfectly follow a normal distribution, potentially affecting the interpretation of results. The assumption of linear relationships in the scatter plots might also overlook potential non-linear relationships, which could be better captured with more advanced modeling techniques.

What challenges did you face, what did you not fully understand? The main challenge was handling and cleaning the dataset to ensure no infinite values or NaNs interfered with the analysis. Understanding the subtle differences between various statistical measures and ensuring correct application was another challenge. Additionally, interpreting the logistic regression results required a solid understanding of statistical significance and model fit metrics. Ensuring the correct interpretation of coefficients and their implications in the logistic regression model was also a learning point.

Conclusion: This project successfully analyzed the factors influencing student graduation rates using various statistical methods outlined in the ThinkStats book. The findings highlight significant variables and their relationships, providing a foundation for further, more detailed studies incorporating additional influential factors. The analysis revealed that higher previous qualification grades are significantly associated with a higher likelihood of graduating, supporting the initial hypothesis. However, including more variables and considering interaction effects could enhance the robustness of the findings.

[ ]: