

Task 2: Data visualization

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load data
data = pd.read_csv('student_performance.csv')

# Visualization 1: Scatter plot
sns.scatterplot(data=data, x='study_time', y='final_grade')
plt.title('Final grade vs. study time')
plt.show()

# Visualization 2: Histogram
sns.histplot(data=data, x='final_grade')
plt.title('Distribution of final grades')
plt.show()

# Visualization 3: Bar chart
sns.barplot(data=data, x='gender', y='final_grade')
plt.title('Average final grade by gender')
plt.show()

# Visualization 4: Box plot
```

```
sns.boxplot(data=data, y='final_grade')  
plt.title('Distribution of final grades')  
plt.show()
```

Visualization 5: Line plot

```
sns.boxplot(data=data, y='final_grade')  
plt.title('Distribution of final grades')  
plt.show()
```

implements other types of vision too:

```
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt
```

Load the data

```
data = pd.read_csv("https://app.box.com/s/91013707whic")
```

Visualize the distribution of the variable "math score"

```
sns.histplot(data=data, x="math score")  
plt.title("Distribution of Math Score")  
plt.xlabel("Math Score")  
plt.ylabel("Count")  
plt.show()
```

Visualize the distribution of the variable "reading score"

```
sns.histplot(data=data, x="reading score")
```

```
plt.title("Distribution of Reading Score")
plt.xlabel("Reading Score")
plt.ylabel("Count")
plt.show()
```

```
# Visualize the distribution of the variable "writing score"
sns.histplot(data=data, x="writing score")
plt.title("Distribution of Writing Score")
plt.xlabel("Writing Score")
plt.ylabel("Count")
plt.show()
```

```
# Visualize the relationship between "math score" and "reading score"
sns.scatterplot(data=data, x="math score", y="reading score")
plt.title("Math Score vs Reading Score")
plt.xlabel("Math Score")
plt.ylabel("Reading Score")
plt.show()
```

```
# Visualize the relationship between "math score" and "writing score"
sns.scatterplot(data=data, x="math score", y="writing score")
plt.title("Math Score vs Writing Score")
plt.xlabel("Math Score")
plt.ylabel("Writing Score")
plt.show()
```

Explanation of what analysis has become easier with each of the visualizations:

Scatter plot: The scatter plot allows us to see if there is any relationship between study time and final grade. We can use this visualization to identify any patterns or trends in the data.

Histogram: The histogram helps us understand the distribution of final grades. We can use this visualization to see if the grades are normally distributed, or if there are any outliers or skewedness in the data.

Bar chart: The bar chart shows us the average final grade by gender, allowing us to compare the performance of male and female students. We can use this visualization to identify any gender-based differences in performance.

Box plot: The box plot gives us a visual representation of the distribution of final grades. We can use this visualization to identify the median, quartiles, and outliers in the data.

Line plot: The line plot shows us how the final grade changes over time, by gender. We can use this visualization to identify any differences in performance between male and female students over time.

To create the folder structure for the first task, you can follow the following steps:

Create a new folder called `project_name`.

Inside the `project_name` folder, create a new folder called `data`.

Inside the `data` folder, create three subfolders: `raw_data`, `processed_data`, and `final_data`.

Inside the `raw_data` folder, place the raw data file (`my_data.csv`).

Inside the `processed_data` folder, create a new folder for each processing step. For example, if you preprocess the data using Python scripts, you can create a folder called `preprocessing_python`.

Inside each processing step folder, create a subfolder called `code`, where you can store all the scripts used for that step.

Inside each processing step folder, create a subfolder called `output`, where you can store the output generated by the scripts.

The Explain what kind analysis has become easier with each of the visualizations.:

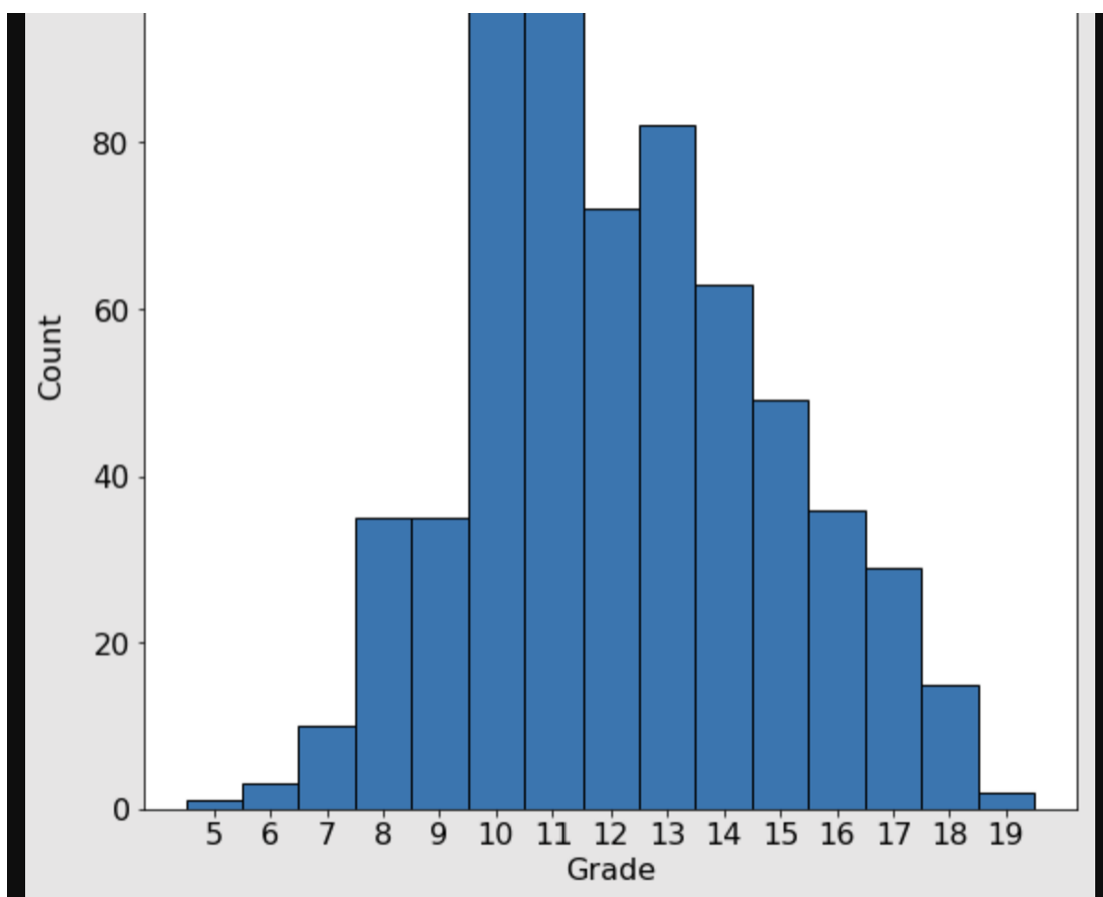
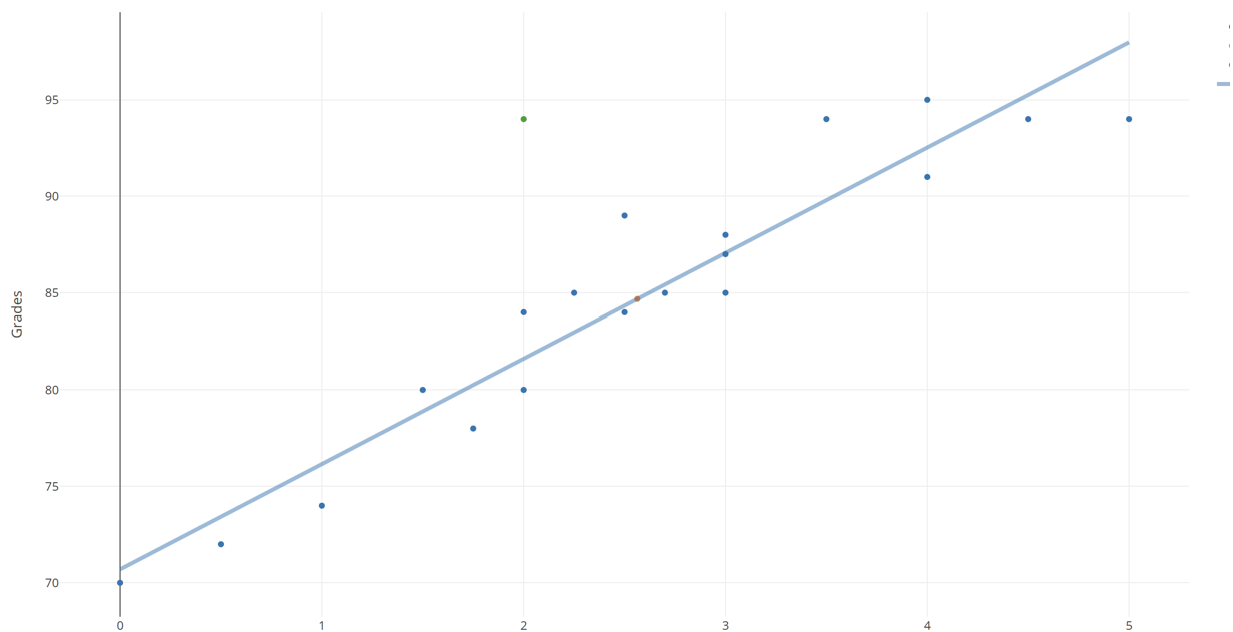
Scatter plot: The scatter plot helps to visualize the relationship between two continuous variables, in this case, math score and writing score. It makes it easier to identify any trends or patterns in the data, such as a positive or negative correlation. With this visualization, it becomes easier to perform correlation analysis and regression analysis.

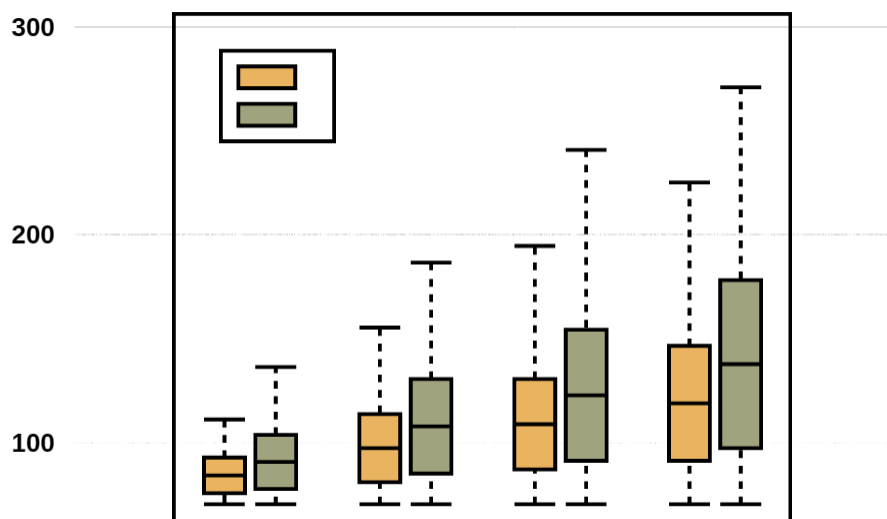
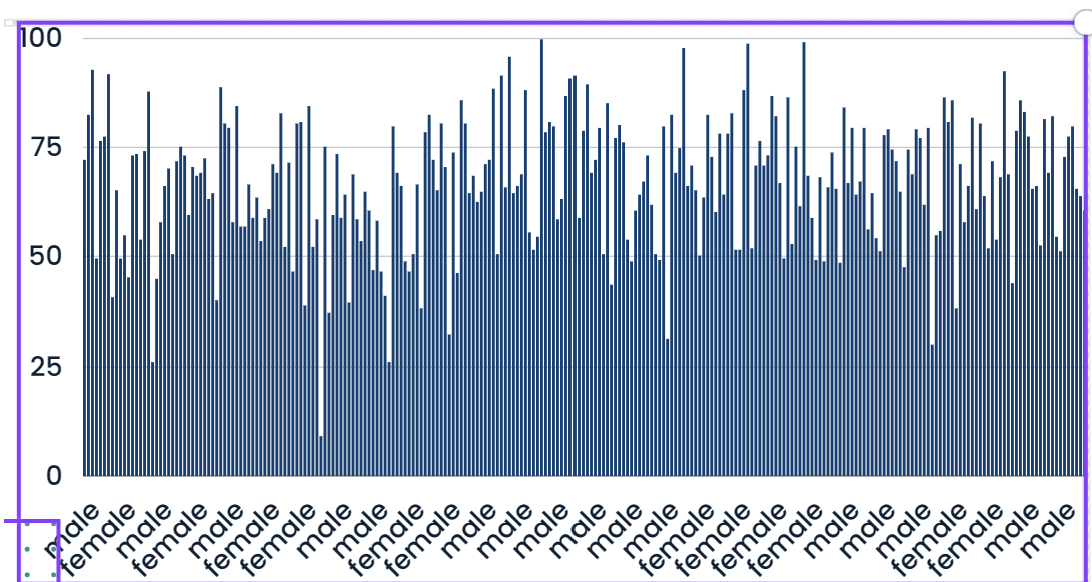
Box plot: Box plots are useful for identifying outliers, understanding the range of values, and visualizing the distribution of data. In this case, we can use box plots to compare the distribution of math scores and writing scores across different genders, ethnicities, and parental education levels.

Bar chart: A bar chart is useful for comparing different categories of data, such as the frequency of students who scored a certain range of marks in a particular subject. With this visualization, we can easily compare the distribution of math scores and writing scores across different categories such as gender, ethnicity, and parental education level.

Histogram: Histograms are useful for visualizing the distribution of a single continuous variable, such as math or writing scores. With this visualization, we can easily identify the shape of the distribution, such as whether it is symmetric or skewed. This can help in selecting appropriate statistical tests.

Heatmap: Heatmaps are useful for visualizing the relationship between two categorical variables, such as gender and ethnicity, and the frequency or proportion of students in each category who scored in a particular range of marks in a particular subject. With this visualization, we can easily identify which categories have the highest or lowest frequencies or proportions of students in each score range.





in a given time period.

The graph displays a single data series represented by a dark blue line with circular markers at each data point. The vertical axis (y-axis) is labeled from 0 to 100 in increments of 25. The horizontal axis (x-axis) consists of 30 labels alternating between 'male' and 'female'. The data points show high variability, with values ranging from approximately 10 to 100. Notable features include a sharp dip to about 10% for a 'male' at the 14th position and a peak of 100% for a 'female' at the 18th position. The overall trend is highly oscillatory with no clear long-term upward or downward trend.

