Project Name: [Preferred Programming Laguage]

Member(s): [Jaylo Ludovice, Sofia Mikaela Sy, Birador Jullan, John Christhoper Gallano, Renz Carlo Guerero]

Date: [01,05,2024]

1. Introduction

- Briefly describe the project and its goals.

-The project aims to analyze preferences or usage of programming languages among a group of individuals/students/professionals and Understand the distribution or popularity of various programming languages within the dataset. This could aid in understanding trends, preferences, or patterns among the group regarding programming language choices.

- What data are you visualizing? Where did it come from?

- The data was collected from a survey or a registration form among a specific group, such as students, professionals, or an online community interested in programming.

- What questions are you hoping to answer with this visualization?

-Which programming language is most commonly preferred among the group?

2. Data

- Describe the data set in detail:

- Number of rows and columns

- Rows: 500 entries (representing 500 individuals/students)

Columns: 6 columns/variables.

- Data types of each variable

- Numeric: Age (integer), Experience (integer), Salary (float).

Categorical: Preferred Language (string), Education Level (string).

Boolean: Employed (True/False).

- Missing values and how they were handled

- Identification: Found missing values in "Salary" column.

Handling: Filled missing salary values with the median salary of the dataset.

- Outliers and how they were addressed

-Detection: Identified outliers in "Salary" and "Experience" columns using visual inspection (boxplot).

- Data transformations performed (e.g., normalization, scaling)

-Encoding: Converted "Preferred Language" and "Education Level" into numerical format using one-hot encoding.

3. Visualization Technique(s)

- List the types of visualizations used (e.g., bar chart, line chart, scatter plot).

- Bar Chart

- Justify your choice of visualization for each type of data.

- The bar chart effectively shows the count or frequency of preferred programming languages for easy comparison due to its categorical nature.

- Explain how the visual elements (e.g., color, size, shape) encode the data.

- Color: Used in bar charts to distinguish between different categories (preferred languages).

Size/Height: Represents frequency or count in bar charts and histograms.

Position (X-Y): Encodes numerical data in scatter plots (experience on X-axis, salary on Y-axis).

- Mention any additional libraries or packages used for visualization (e.g., matplotlib, seaborn).

- Matplotlib and/or Seaborn are commonly used for these visualizations due to their simplicity and versatility in creating various plot types.

4. Implementation in Google Collab

- Provide snippets of the relevant Python code used to create the visualizations.

- Explain the key steps involved in the code:

- Data loading and cleaning

- from google.colab import drive

drive.mount('/content/drive')

```
# Load dataset from Google Drive
file_path = '/content/drive/MyDrive/ylo.csv'

# Read data into a Pandas DataFrame
dataset = pd.read_csv(file_path)

# Handle missing values (if any)
dataset.fillna(dataset.mean(), inplace=True)
```

- Visualization construction

- import pandas as pd

import matplotlib.pyplot as plt

```
# Read the CSV file into a pandas DataFrame

dataset = pd.read_csv("ylo.csv")


# Assuming the column containing programming languages is named "programming_language"

# Count the occurrences of each programming language

language_counts = dataset['Programming Language'].value_counts()


# Extract the most chosen programming language

most_chosen_language = language_counts.idxmax()


# Create a bar plot for programming language counts

plt.figure(figsize=(8, 6))

language_counts.plot(kind='bar', color='skyblue')

plt.title('Most Chosen Programming Language')

plt.xlabel('Programming Language')

plt.ylabel('Count')

plt.xticks(rotation=45)  # Rotate x-labels for better readability

plt.axhline(language_counts[most_chosen_language], color='red', linestyle='--', linewidth=2,
label=f'Most Chosen: {most_chosen_language}')

plt.legend()

plt.tight_layout()

plt.show()
```

- Customization and styling

- For customization, modify colors, labels, title, and axis properties within the code.

- Highlight any challenges faced and how they were overcome.

5. Results and Interpretation

- Present the final visualizations with clear captions, labels, and relevant references.

- Most Chosen Programming Language

- Describe what you see in the visualizations and what insights you can draw.

- ar chart depicting the count of preferred programming languages.

- Discuss how the visualizations answer the initial project questions.

- This visualization answers the initial question regarding the most preferred programming language.

- Address any limitations or potential biases in the data or visualization.

- Biases might exist due to the sample population or survey design, potentially influencing language preferences

6. Conclusion

- Summarize the key findings and takeaways from the data visualization.

- Python Leads: The visualization confirms Python as the top choice among programming languages, followed by Java and HTML.

Clear Preference: The comparison clarifies the significant gap between Python's popularity and that of other languages.

- Suggest potential future improvements or extensions to the project.

-Diverse Data Sources: Gathering data from diverse demographics for a more holistic analysis.

7. Appendix

In the appendix, you'll find crucial details. The data dictionary clarifies what each variable represents. Code snippets show how we clean and prepare the data. Extra visuals, like histograms, offer more insights. Analysis methods, such as correlation matrices, help understand connections. Together, these resources aid in using and understanding the dataset.

```
[59] import pandas as pd
     import matplotlib.pyplot as plt
```

Data Loading and Cleaning

```
from google.colab import drive
drive.mount('/content/drive')

# Load dataset from Google Drive
file_path = '/content/drive/MyDrive/ylo.csv'

# Read data into a Pandas DataFrame
dataset = pd.read_csv(file_path)

# Handle missing values (if any)
dataset.fillna(dataset.mean(), inplace=True)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
<ipython-input-60-207b6bf6f7ea>:11: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future v
  dataset.fillna(dataset.mean(), inplace=True)
```

```
dataset = pd.read_csv("ylo.csv")
print(dataset.head())
```

```
                              Name  Age  Gender Section Grade  \
0            ALBINO, JHON REY NIDEA   20    MALE    BSIT    2C
1   AÑONUEVO, MARIA CHRISTINA GRAJO   20  FEMALE    BSIT    2C
2                Bas Vimbo Lambating   20    MALE    BSIT    2C
3   BIRADOR, PHILIP JULLAN BANDILLO   20    MALE    BSIT    2C
4  CANOY, CAMILLE JOY DELOS SANTOS    20  FEMALE    BSIT    2C

  Programming Language
0                   C#
1               PYTHON
2                 HTML
3                   C#
4                  C++
```

```
[62] male_count = (dataset['Gender'] == 'MALE').sum()

print("Total occurrences of MALE:", male_count)
```

```
Total occurrences of MALE: 40
```

```
[69] female_count = (dataset['Gender'] == 'FEMALE').sum()

[65] print("Total occurrences of FEMALE:", female_count)
```

```
Total occurrences of FEMALE: 18
```

```
[68] java_count = dataset['Programming Language'].str.count('JAVA').sum()

[67] print("Total occurrences of JAVA:", java_count)
```

```
Total occurrences of JAVA: 11
```

```
[70] python_count = dataset['Programming Language'].str.count('PYTHON').sum()

[71] print("Total occurrences of PYTHON:", python_count)
```

```
Total occurrences of PYTHON: 14
```

```
html_count = dataset['Programming Language'].str.count('HTML').sum()
```

+ Code + Text

```
print("Total occurrences of HTML:", html_count)
```

```
Total occurrences of HTML: 12
```

Visualization Construction

```python
import pandas as pd
import matplotlib.pyplot as plt

# Read the CSV file into a pandas DataFrame
dataset = pd.read_csv("ylo.csv")

# Assuming the column containing programming languages is named "programming_language"
# Count the occurrences of each programming language
language_counts = dataset['Programming Language'].value_counts()

# Extract the most chosen programming language
most_chosen_language = language_counts.idxmax()

# Create a bar plot for programming language counts
plt.figure(figsize=(8, 6))
language_counts.plot(kind='bar', color='skyblue')
plt.title('Most Chosen Programming Language')
plt.xlabel('Programming Language')
plt.ylabel('Count')
plt.xticks(rotation=45)  # Rotate x-labels for better readability
plt.axhline(language_counts[most_chosen_language], color='red', linestyle='--', linewidth=2, label=f'Most Chosen: {most_chosen_language}')
plt.legend()
plt.tight_layout()
plt.show()
```