```
import pandas as pd

# Load the dataset
file_path = '/mnt/data/pone.0269022.s002.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
data.head()
```

Task

Tell me about this dataset.

Here is all the data you need: "pone.0269022.s002.csv"

Data loading

Subtask:

Load the dataset "pone.0269022.s002.csv" into a pandas DataFrame.

Reasoning: Load the dataset "pone.0269022.s002.csv" into a pandas DataFrame and display the first few rows.

```
import pandas as pd

try:
    df = pd.read_csv('pone.0269022.s002.csv')
    display(df.head())
except FileNotFoundError:
    print("Error: File 'pone.0269022.s002.csv' not found.")
    df = None
except pd.errors.ParserError:
    print("Error: Unable to parse the CSV file.")
    df = None
except Exception as e:
    print(f"An unexpected error occurred: {e}")
    df = None
```


```
{"summary":"{\n  \"name\": \"    df = None\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"ID\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": 1,\n
\"min\": 1,\n        \"max\": 5,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          2,\n          5,\n          3\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Month\",\n      \"properties\": {\
n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\":
1,\n        \"max\": 1,\n        \"num_unique_values\": 1,\n
```

\"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Death\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 0,\n   \"max\": 1,\n   \"num_unique_values\": 2,\n   \"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Roadway_Class\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 1,\n   \"min\": 1,\n   \"max\": 4,\n   \"num_unique_values\": 3,\n   \"samples\": [\n   4\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Road_surface\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 1,\n   \"max\": 1,\n   \"num_unique_values\": 1,\n   \"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Road_section\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 1,\n   \"max\": 3,\n   \"num_unique_values\": 3,\n   \"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Weather_condition\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 1,\n   \"max\": 1,\n   \"num_unique_values\": 1,\n   \"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Light_condition\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 1,\n   \"max\": 3,\n   \"num_unique_values\": 3,\n   \"samples\": [\n   1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe"}

Data exploration

Subtask:

Explore the loaded dataset to understand its characteristics.

Reasoning: Examine the shape, data types, missing values, and distributions of numerical and categorical features.

```
# Examine the shape of the DataFrame
print("Shape of the DataFrame:", df.shape)

# Check data types and identify numerical and categorical columns
print("\nData Types:\n", df.dtypes)

# Investigate missing values
```

```
print("\nMissing Values:\n", df.isnull().sum())

# Analyze numerical features
numerical_cols = df.select_dtypes(include=['number']).columns
print(f"\nNumerical Features: {numerical_cols.tolist()}")
print("\nSummary Statistics for Numerical Features:\n",
df[numerical_cols].describe())

# Analyze categorical features
categorical_cols = df.select_dtypes(include=['object',
'category']).columns
print(f"\nCategorical Features: {categorical_cols.tolist()}")
for col in categorical_cols:
    print(f"\nUnique values and frequencies for {col}:\
n{df[col].value_counts()}")

# Visualize numerical features with histograms (if applicable)
import matplotlib.pyplot as plt
plt.figure(figsize=(12, 8))  # Adjust figure size
for i, col in enumerate(numerical_cols):
  plt.subplot(2, 4, i + 1)  # Adjust subplots as needed
  plt.hist(df[col], bins=10, color='skyblue', edgecolor='black')
  plt.title(f'Distribution of {col}')
  plt.xlabel(col)
  plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

```
Shape of the DataFrame: (20229, 8)

Data Types:
 ID                  int64
Month               int64
Death               int64
Roadway_Class       int64
Road_surface        int64
Road_section        int64
Weather_condition   int64
Light_condition     int64
dtype: object

Missing Values:
 ID                  0
Month               0
Death               0
Roadway_Class       0
Road_surface        0
Road_section        0
Weather_condition   0
Light_condition     0
```

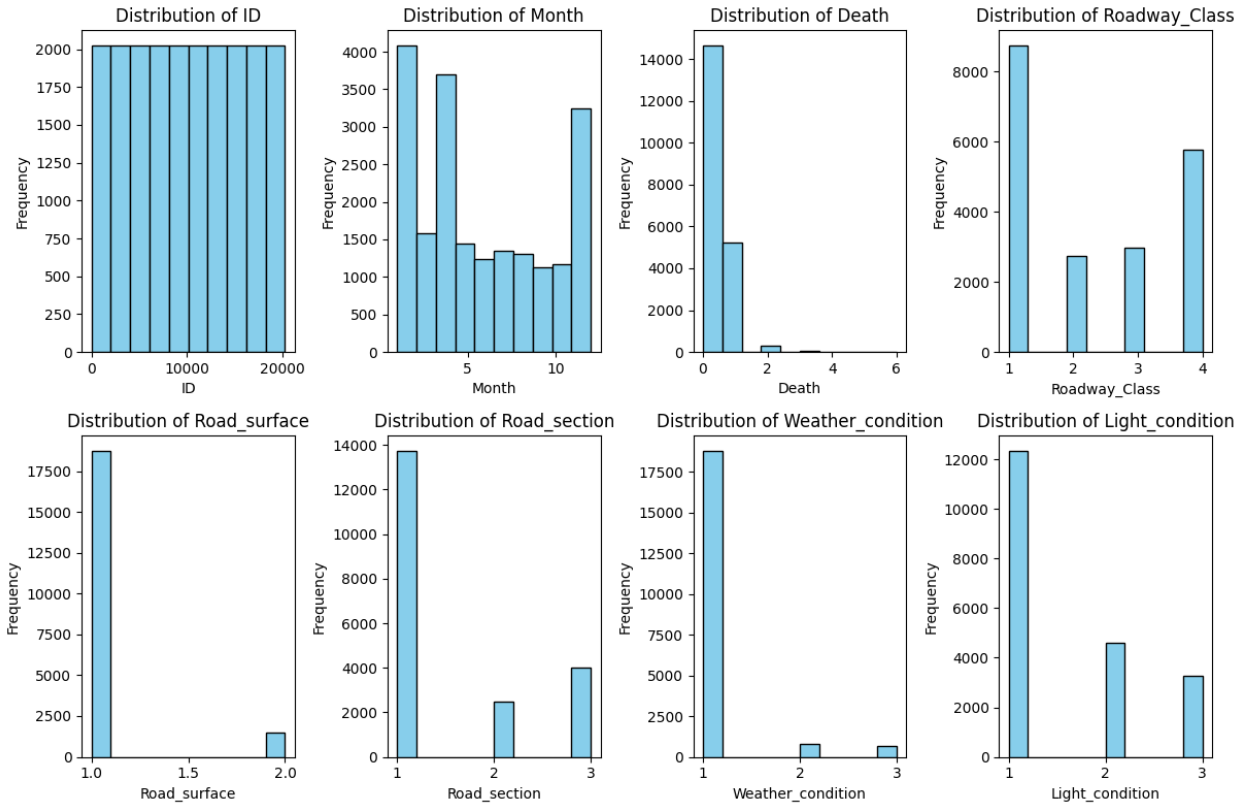
```
dtype: int64

Numerical Features: ['ID', 'Month', 'Death', 'Roadway_Class',
'Road_surface', 'Road_section', 'Weather_condition',
'Light_condition']

Summary Statistics for Numerical Features:
                  ID         Month         Death  Roadway_Class
Road_surface  \
count  20229.000000  20229.000000  20229.000000   20229.000000
20229.000000
mean   10115.000000      5.885857      0.301992       2.286420
1.073014
std     5839.753634      3.572849      0.535940       1.280404
0.260166
min        1.000000      1.000000      0.000000       1.000000
1.000000
25%     5058.000000      3.000000      0.000000       1.000000
1.000000
50%    10115.000000      5.000000      0.000000       2.000000
1.000000
75%    15172.000000      9.000000      1.000000       4.000000
1.000000
max    20229.000000     12.000000      6.000000       4.000000
2.000000

       Road_section  Weather_condition  Light_condition
count  20229.000000       20229.000000     20229.000000
mean       1.518068           1.102427         1.549854
std        0.803291           0.393978         0.754493
min        1.000000           1.000000         1.000000
25%        1.000000           1.000000         1.000000
50%        1.000000           1.000000         1.000000
75%        2.000000           1.000000         2.000000
max        3.000000           3.000000         3.000000

Categorical Features: []
```



Data analysis

Subtask:

Perform detailed data analysis on the loaded dataset.

Reasoning: Calculate descriptive statistics, analyze correlations between numerical features, and investigate relationships between numerical and categorical variables. Visualize the correlation matrix using a heatmap.

```
# Calculate descriptive statistics
numerical_cols = df.select_dtypes(include=['number']).columns
print("Descriptive Statistics:\n", df[numerical_cols].describe())

# Analyze correlations between numerical features
correlation_matrix = df[numerical_cols].corr()
print("\nCorrelation Matrix:\n", correlation_matrix)

import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```

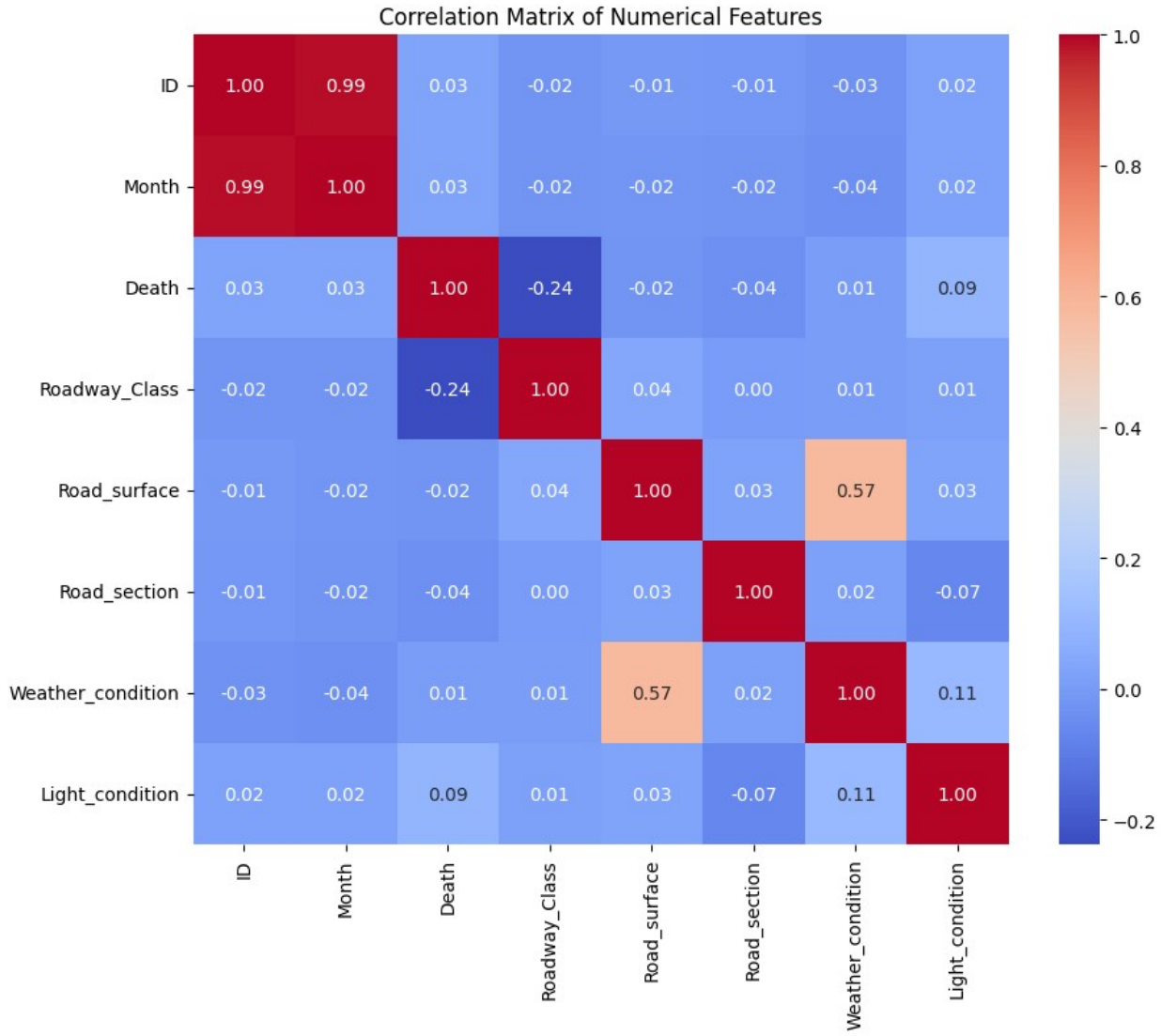
```python
# Analyze relationships between numerical and (potential) categorical features
categorical_cols = df.select_dtypes(include=['object', 'category']).columns
if len(categorical_cols) > 0:
    for cat_col in categorical_cols:
        print(f"\nRelationship between numerical features and {cat_col}:")
        for num_col in numerical_cols:
          if num_col != 'ID': # Exclude ID from the analysis.
            print(f"  Mean of {num_col} for each category in {cat_col}:")
            print(df.groupby(cat_col)[num_col].mean())
            print(f"  Median of {num_col} for each category in {cat_col}:")
            print(df.groupby(cat_col)[num_col].median())
else:
    print("\nNo categorical features found.")
```

```
Descriptive Statistics:
                 ID         Month         Death  Roadway_Class  \
count  20229.000000  20229.000000  20229.000000   20229.000000
mean   10115.000000      5.885857      0.301992       2.286420
std     5839.753634      3.572849      0.535940       1.280404
min        1.000000      1.000000      0.000000       1.000000
25%     5058.000000      3.000000      0.000000       1.000000
50%    10115.000000      5.000000      0.000000       2.000000
75%    15172.000000      9.000000      1.000000       4.000000
max    20229.000000     12.000000      6.000000       4.000000

       Road_surface  Road_section  Weather_condition  Light_condition
count  20229.000000  20229.000000       20229.000000     20229.000000
mean       1.073014      1.518068           1.102427         1.549854
std        0.260166      0.803291           0.393978         0.754493
min        1.000000      1.000000           1.000000         1.000000
25%        1.000000      1.000000           1.000000         1.000000
50%        1.000000      1.000000           1.000000         1.000000
75%        1.000000      2.000000           1.000000         2.000000
max        2.000000      3.000000           3.000000         3.000000
```

```
Correlation Matrix:
                            ID     Month     Death  Roadway_Class
Road_surface  \
ID                    1.000000  0.985688  0.030253      -0.023192     -
0.007628
Month                 0.985688  1.000000  0.029259      -0.020529     -
0.018689
Death                 0.030253  0.029259  1.000000      -0.238079     -
0.023416
Roadway_Class        -0.023192 -0.020529 -0.238079       1.000000
0.039766
Road_surface         -0.007628 -0.018689 -0.023416       0.039766
1.000000
Road_section         -0.014849 -0.022527 -0.038914       0.001649
0.026450
Weather_condition    -0.029205 -0.035700  0.007790       0.006423
0.574773
Light_condition       0.019323  0.022660  0.094131       0.013978
0.028425

                    Road_section  Weather_condition  Light_condition
ID                     -0.014849          -0.029205         0.019323
Month                  -0.022527          -0.035700         0.022660
Death                  -0.038914           0.007790         0.094131
Roadway_Class           0.001649           0.006423         0.013978
Road_surface            0.026450           0.574773         0.028425
Road_section            1.000000           0.016646        -0.066924
Weather_condition       0.016646           1.000000         0.110547
Light_condition        -0.066924           0.110547         1.000000
```




```
No categorical features found.
```

Data visualization

Subtask:

Visualize the data distributions and relationships.

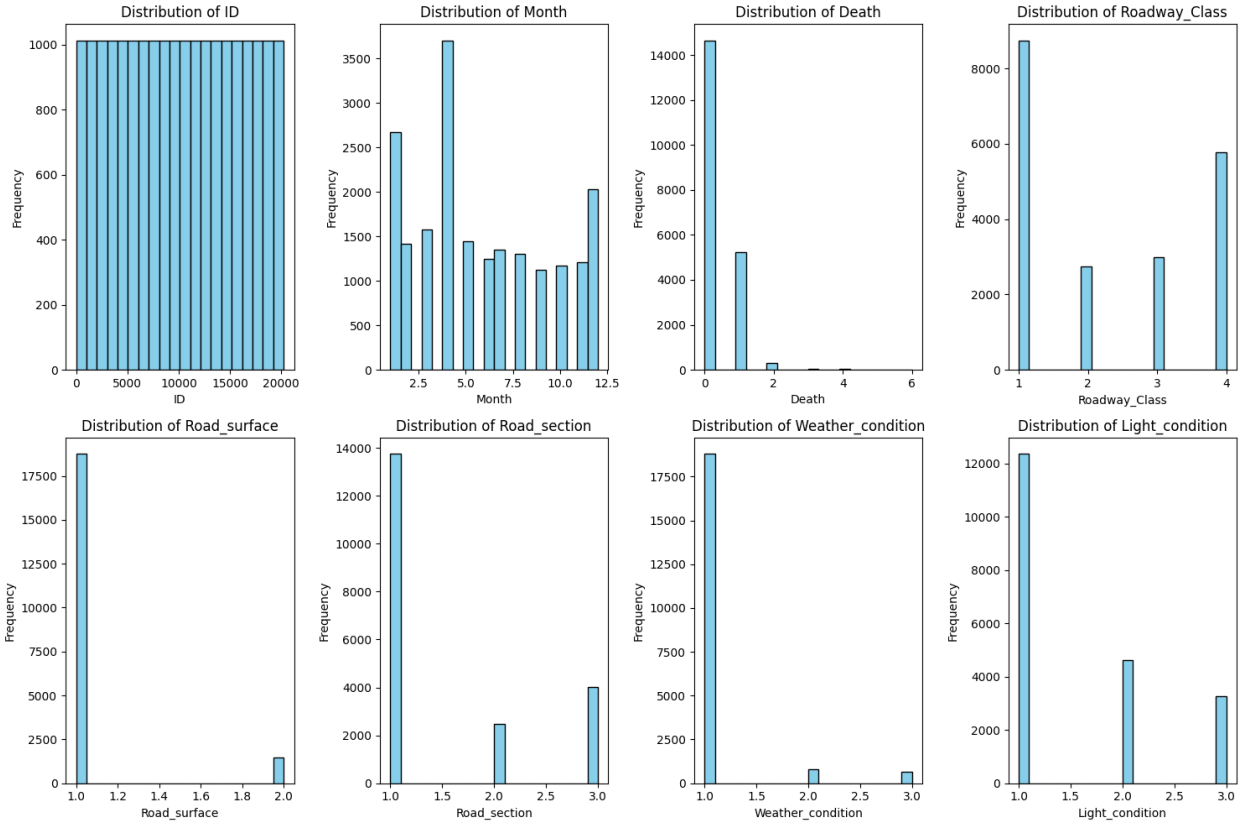
Reasoning: Create histograms for numerical features, a correlation matrix heatmap, and bar charts for categorical features (if any).

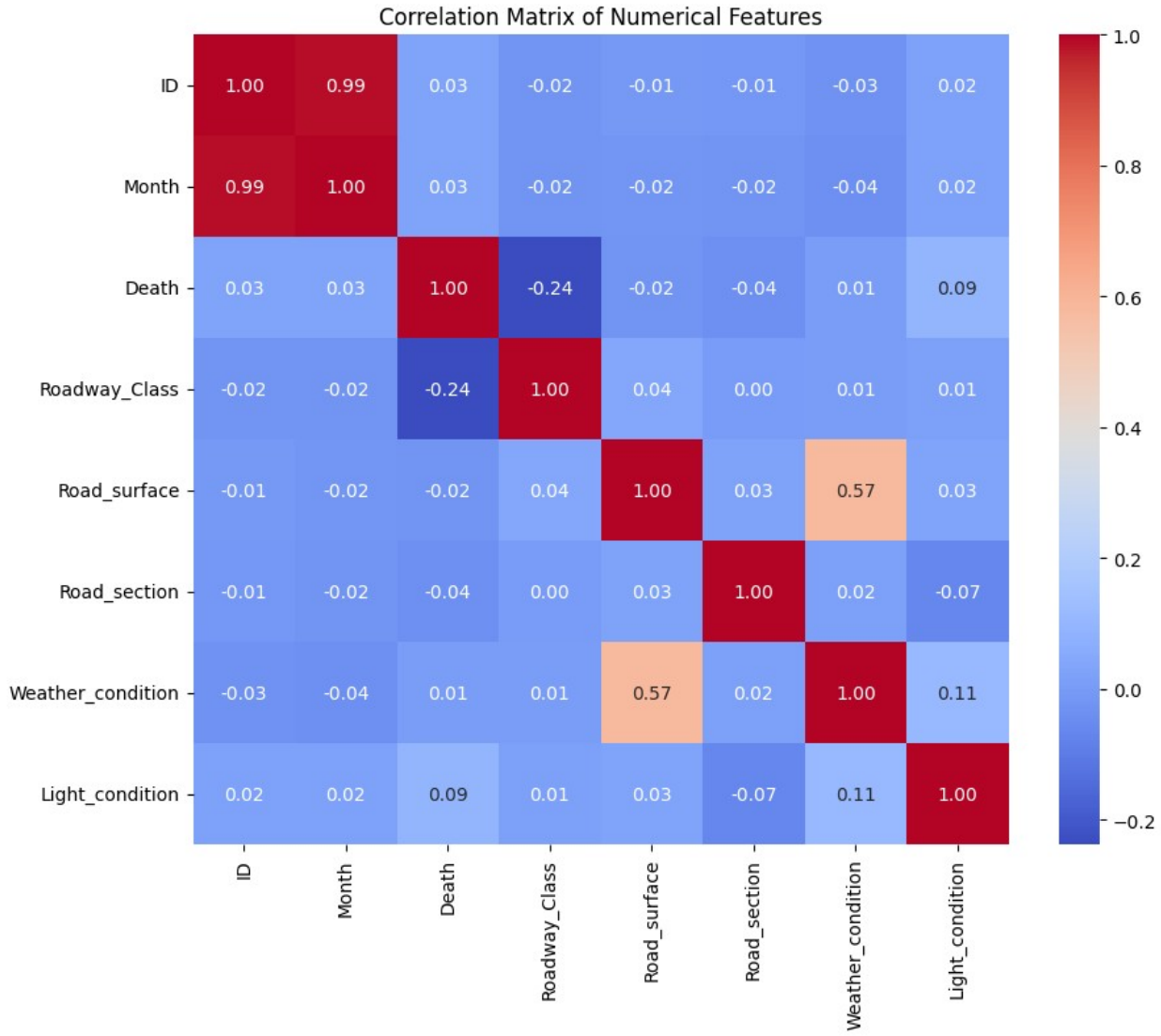
```
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Histograms for numerical features
plt.figure(figsize=(15, 10))
numerical_cols = df.select_dtypes(include=['number']).columns
for i, col in enumerate(numerical_cols):
    plt.subplot(2, 4, i + 1)
    plt.hist(df[col], bins=20, color='skyblue', edgecolor='black')
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

# Correlation matrix heatmap
plt.figure(figsize=(10, 8))
correlation_matrix = df[numerical_cols].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Correlation Matrix of Numerical Features')
plt.show()

# Bar charts for categorical features (if any)
categorical_cols = df.select_dtypes(include=['object',
'category']).columns
if len(categorical_cols) > 0:
    plt.figure(figsize=(12, 6))
    for i, col in enumerate(categorical_cols):
        plt.subplot(1, len(categorical_cols), i + 1)
        df[col].value_counts().plot(kind='bar', color='lightgreen')
        plt.title(f'Distribution of {col}')
        plt.xlabel(col)
        plt.ylabel('Frequency')
    plt.tight_layout()
    plt.show()
else:
    print("No categorical features found.")
```


Distribution of ID
Frequency
ID
Distribution of Month
Frequency
Month
Distribution of Death
Frequency
Death
Distribution of Roadway_Class
Frequency
Roadway_Class
Distribution of Road_surface
Frequency
Road_surface
Distribution of Road_section
Frequency
Road_section
Distribution of Weather_condition
Frequency
Weather_condition
Distribution of Light_condition
Frequency
Light_condition




```
No categorical features found.
```

Summary:

Q&A

There were no explicit questions asked in the task. However, the task implicitly asks for a description of the dataset "pone.0269022.s002.csv".

Data Analysis Key Findings

- The dataset "pone.0269022.s002.csv" contains 20,229 rows and 8 columns.
- All columns are numerical. No categorical features were identified.
- There are no missing values in the dataset.
- The numerical features include 'ID', 'Month', 'Death', 'Roadway_Class', 'Road_surface', 'Road_section', 'Weather_condition', and 'Light_condition'.

- A correlation matrix was computed to analyze the relationships between the numerical features. A heatmap visualization of the correlation matrix was also generated.

Insights or Next Steps

- Investigate the potential relationships between the numerical features based on the correlation matrix and consider applying dimensionality reduction techniques.
- While no categorical features were explicitly identified, revisit the data dictionary or explore the features more thoroughly to see if any numerical columns could represent categorical data in disguise. If found, these features could reveal more interesting insights.