STUDENT_NAME: GHADAH ABDULLAH ALAMOUDI

STUDENT_ID: 21520734

CO.NO: TM351

Q2:

Report Title: The traffic violations

Summary

There are about 65,000 traffic-related violation reports in the traffic infractions dataset which has been published. This dataset contains a wealth of information regarding several aspects of moving offenses, such as:

- Information Date and Time: Specifies the exact time the infraction happened.
- Country Name: Indicates the nation in which the transgression occurred.
- Gender of Violators: Indicates the gender (M for Male, F for Female) of the people who committed the infractions.
- Age of Violators: Indicates how old the violators are.
- Race of Violators: Provides information about the violators' race.
- Category of Violation: Groups violations according to their type.
- Search Conducted: This indicates if there was a search done during the traffic stop.
- Violation Result: Indicates the result of the infraction (warning, citation, etc.)
- Arrest Information: Indicates if the someone was taken into custody.
- Detained Time: Documents the amount of time that infractions take to cease.
- Drug-Related Activities: Emphasizes any connection with drug-related activities.

Data_Quality_Assessment

We must take into account a number of elements in order to evaluate the data's quality:

- **VALIDITY**: Although the dataset seems to have a legitimate structure, more validation procedures are required to verify the accuracy of specific entries.
- **ACCURACY**: While accuracy is normally good, certain verifications are necessary to verify sure all inputs make sense, such as those pertaining to age and timings.
- **COMPLETENESS**: The analysis may be impacted by a few missing values. It is essential to make sure all fields are filled in or to handle missing information properly.

- **CONSISTENCY**: Although the format on the data appears consistent, uniformity tests are necessary, particularly for categorical data.
- **UNIFORMITY**: To keep the dataset uniform, standardizing units and formats is essential.

Impact_of_Data_Dirtiness

- Inconsistencies and missing values can greatly distort the analysis and produce false conclusions.
- Inaccurate age or time data can have an impact on how patterns in violator demographics and violation timing are understood.
- It may be more difficult to aggregate and compare records if the data is not uniform.
- We should handle missing values in the dataset, confirm the accuracy of the entries, and make sure the data formats are consistent and uniform in order to prepare it for a through analysis.

Q3

The investigation of the relationship between gender and the category of traffic violations

Executive Synopsis

Using the Traffic offenses Dataset, this project examines the association between gender and the type of traffic offenses. Our goal in examining this dataset is to determine whether the kinds of violations committed and the gender of the violators are significantly correlated., in order to arrive at insightful conclusions, the analysis involves computing correlation coefficients, producing visuals , and using pivot tables to summarize the data.

Aims and Objectives

General Aims

- To Examine the correlation between the sex of transgressors and the types of transgressions committed.
- To find out if the kinds of infractions that men and women commit differ noticeably from one another.

Detailed Objectives

- To examine how gender affects the distribution of moving infractions.
- To determine the relationship between the category of infraction and gender.
- To use data visualization to draw attention to any noteworthy trends.
- To offer a critical analysis of the findings and recommendations.

THE RESEARCH QUESTION

Identifying Variables:

- **Dependent variable**: Category of violation
- **Independent variable**: violators' gender

ANALYSIS AND FINDINGS

Correlation Analysis:

We will apply the Chi-Square test for independence, appropriate for categorical data, to evaluate the association between gender and the category of violation.

Interpretation of Correlation Results:

If a statistically significant association exists between gender and the category of violation, it is indicated the Chi-Square test result. A Substantial correlation is suggested by a p-value of less than 0.05.

Visualizations

1. Bar Plot of Violations by Gender.
2. Pie Chart Gender Distribution in Violations.

Interpretation of Visualizations:

- The Bar Plot illustrates how males and females differ in the frequency of infraction categories.
- The Pie Chart illustrates the entire distribution of traffic infractions by gender.

Final Answer to the Research Question

The analysis reveals whether there is a significant correlation between gender and category of traffic violation. If a significant correlation is found, it suggests that gender influences the type of traffic violations committed.

Critical Comment on Conclusions

The analysis's insights can be used to better understand the demographic trends in traffic infractions. Which can guide the development of targeted programs and policies.

Reflection

Experience with the Project

- The project demonstrated the value of data quality and preprocessing by giving participants practical experience with data analysis and visualization.
- Discovered the use of statistical testing in figuring out how variable relate to one another.

What Went Well

- Successful data loading and cleaning process.
- Making good use of visuals to draw attention to important patterns.

What Went Wrong

- Initial issues with missing values and data formatting.
- Inconsistencies in some of the entries made it difficult to interpret some of the categorical variables.

Future Benefits

- Improved understanding of data analysis techniques.
- Better preparation for handling real-world data issues in future projects.

References

1.Shubamsumbria.Traffic Violations Dataset. Retrieved from Kaggle: Link, Traffic and Drugs Related Violations Dataset (kaggle.com)

2. McKinney, W.(2017). Python for Data Analysis. O'Reilly Media.

3. Seabold, S.,&Perktold, J. (2010). Stats models: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference.

Q4

Study the ‘Predict Online Course Engagement Dataset’

Developing the Python code to abstract some aspects around this dataset.

The code was publishing on Jupiter notebook as will as by the IDLE python a script, all these files have been attached with the TMA351 folder.

Link of the dataset:

https://www.kaggle.com/datasets/rabieelkharoua/predict-online-course-engagement-dataset

coursesnolinedataset.py - C:/Users/galam/AppData/Local/Programs/Python/Python312/cou...

File Edit Format Run Options Window Help

```
#online_course_engagement_data.csv
import pandas as pd

# Correct file path
file_path = 'C:\\Users\\galam\\Downloads\\archive (1)\\online_course_engagement_

# Load the dataset
try:
    df = pd.read_csv(file_path)

    # a. Write its shape
    shape = df.shape
    print(f"Shape of the dataset: {shape}")

    # b. Provide information about the dataset
    print("\nInformation about the dataset:")
    info = df.info()

    # c. Provide statistical summary of the dataset
    print("\nStatistical summary of the dataset:")
    summary = df.describe()
    print(summary)

    # d. Find the pairwise correlation of all columns in the data frame
    # Exclude non-numeric columns
    numeric_df = df.select_dtypes(include=['float64', 'int64'])
    correlation = numeric_df.corr()

    print("\nPairwise correlation of all columns:")
    print(correlation)

except FileNotFoundError:
    print(f"File not found: {file_path}")
```