ChatGPT

Separating Exploratory Data Analysis (EDA) from data cleaning into two different scripts can be a very effective approach, especially in complex data analysis projects like your crime analysis project. Here are some reasons why this approach is beneficial:

Clarity and Organization: By dividing the tasks into two scripts, you maintain a clear distinction between data preparation and analysis. This separation helps in organizing the workflow, making it easier to understand, maintain, and modify the code.

Reusability and Efficiency: The data cleaning script can be used as a standalone module that prepares data for various types of analysis, not just EDA. Once the data is cleaned and saved (e.g., as pickle files), it can be reused without the need to repeat the cleaning process, saving time and computational resources.

Error Handling and Debugging: Separating the scripts allows for more targeted error handling and debugging. If there's an issue in the data cleaning process, it can be addressed independently of the analysis, and vice versa. This modularity can be particularly helpful in large projects or when working in teams.

Scalability and Adaptability: As your analysis evolves or if new data is added, it's easier to adapt or scale your project. For instance, if new data sources are introduced or if the structure of the data changes, you would primarily need to update the data cleaning script.

Collaboration and Version Control: In team settings, different team members can work on the data cleaning and EDA scripts simultaneously. This separation also makes version control more manageable, as changes to one part of the project (e.g., data cleaning) don't directly impact the other (e.g., analysis).

Focus on Specific Objectives: By separating the scripts, each can be tailored to specific objectives—data cleaning for ensuring data quality and EDA for uncovering insights and patterns. This focus enhances the effectiveness of each phase.

However, it's important to note that this approach works best when there is clear communication between the two phases. The data cleaning script should prepare the data in a way that is suitable for the type of analysis you plan to perform in the EDA. Regular reviews and updates may be necessary to ensure alignment between the data preparation and the analysis objectives.

Vancouver Crime Analysis Project Documentation (Updated with Script Division)

Introduction

This document describes the approach and methodology for the Vancouver Crime Analysis project, focusing on crime data from 2003 to 2023. The project is dedicated to performing an exploratory data analysis (EDA) to identify trends, neighborhood differences, crime types, and the influence of factors like dates, holidays, weekdays, population growth, and the Consumer Price Index (CPI).

Objectives

The EDA objectives include:

* Trends Over Time Analysis: Investigating changes in crime rates from 2003 to 2023.
* Neighborhood Analysis: Analyzing crime rates across different neighborhoods.
* Crime Type Analysis: Examining various crime types.
* Date, Holiday, and Weekday Impact: Exploring the correlation with specific dates, holidays, and weekdays.
* Correlation with Population Growth and CPI: Studying the relationship between crime rates and external factors like population growth and CPI.

Data Source

The dataset, provided by the Vancouver Police, covers crime records from 2003 to 2023. It is updated weekly on Sundays and is available to the public at Vancouver Police GeoDash Open Data.

<https://geodash.vpd.ca/opendata/#>

Script Division for Data Preparation and EDA

* Data Preparation Script:

Libraries for data handling were imported.

Initial dataset exploration conducted for identifying key trends.

* Data Cleaning:

Managed missing data, duplicates, and outliers.

Ensured the accuracy of vital columns like 'NEIGHBOURHOOD'.

* Feature Engineering:

Added new columns for date, holiday, weekday, and crime type categories.

Integrated yearly aggregated data with population and CPI data.

* Export of Prepared Data:

Prepared data was exported as pickle files for subsequent EDA.

EDA Script:

This separate script will use the cleaned and prepared data in pickle format from the data preparation script.

It will focus on exploring and visualizing the trends, patterns, and correlations identified in the objectives section.

Approach and Methodology

The project starts with an initial review and cleaning of the dataset, followed by a detailed data preparation phase. The data preparation and EDA processes are divided into separate scripts for efficiency and clarity. The first script handles data cleaning and preparation, outputting pickle files, which are then used in the EDA script.

Approach and Methodology

The project's approach is structured around two distinct scripts:

Data Preparation Script:

Involves importing data, cleaning, and feature engineering.

Focuses on preparing the dataset for analysis by handling missing values, duplicates, and adding relevant features.

The prepared data is exported as pickle files for use in the EDA script.

EDA Script:

This script, separate from the data preparation, utilizes the cleaned and prepared data in pickle format.

It focuses exclusively on conducting the exploratory data analysis, aligning with the project objectives.

This bifurcated approach of using two different scripts ensures a clean separation of data preparation from analysis, facilitating a more organized and focused workflow.