Decoding NYC: Unveiling Insights Through Data Analysis

SCALABLE DATABASE MIDTERM

Group Members:

Abhrajit Das

Sai Bhere

Ved Aralkar

Project Objectives:

Objective 1: Uncovering demographic insights within NYC, exploring population demographics, and their correlations with various factors.

Objective 2: Identifying crime trends and patterns across different neighbourhoods and boroughs within NYC.

Objective 3: Assessing transportation accessibility and its impact on crime rates or demographic variations.

Data Sources:

Utilized three primary datasets sourced from NYC Open Data:

NYPD Incident Data: Offering insights into crime rates, types, locations, and related attributes.

Holiday List: Providing information on holidays and their occurrences within NYC.

NYC Weather Data: Furnishing details about weather conditions across different time frames.

Data Compression:

To manage extensive datasets comprising over 30 million entries, totalling 32 GB, we employed a data compression technique involving random sampling. Through this process, we effectively condensed the dataset to a significantly smaller size, reducing it to a mere 38 MB.

Version Control (Git):

Employed Git version control within the project workflow:

Set up and managed a Git repository enabling collaboration, branching, and tracking project development.

Facilitated collaborative efforts among team members, aiding in code management and project documentation.

Data Exploration and Cleaning:

Leveraged Python's panda's library for comprehensive exploration and cleaning of datasets:

Addressed missing values, outliers, and inconsistencies through data preprocessing techniques.

Ensured data integrity and quality by handling data anomalies effectively.

SQL Queries:

Employed SQL queries for data retrieval, filtering, aggregation, and potential data joins:

Submitted and utilized SQL query snippets (not explicitly visible in the provided code) to extract and manipulate data from the datasets.

Utilized SQL as a key tool for data transformation and retrieval alongside Python's analysis functionalities.

Python for Data Analysis:

Employed Python libraries such as pandas, matplotlib, and seaborn:

Conducted detailed data analysis, generating various visualizations (e.g., bar charts, and scatter plots) to illustrate trends and patterns within the datasets.

Collaborative Development:

Utilized Git version control for collaborative development:

Collaborated with multiple team members, implemented branching strategies, and resolved potential conflicts during collaborative coding.

Challenges and Solutions:

Data Quality Challenges:

Challenge: Handling missing or inconsistent data within datasets.

Solution: Employing data imputation methods and domain knowledge for inference, ensuring data quality.

Managing Large Dataset:

Challenge: Initially dealt with a large dataset of 30 million records.

Solution: Applied random sampling to create a smaller representative subset for more efficient analyses.

Complexities in Data Merging:

Challenge: Faced complexities during merging due to varied formats and inconsistencies.

Solution: Prioritized data cleaning and standardization, ensuring accurate integration across datasets.

Future Steps:

Integration of Additional Datasets:

Expansion Opportunity: Incorporate socio-economic, transportation, or real estate data for a more comprehensive analysis.

Enhanced Visualization and Reporting: Interactive Dashboards: Develop interactive dashboards for stakeholders to explore crime trends, demographic correlations, and insights dynamically.