dillhoffuta / assignment-2-NARESHBABU-CLOUD

Code Issues Pull requests Actions Projects Security Insights

assignment-2-NARESHBABU-CLOUD created by GitHub Classroom

MIT license

0 stars 0 forks 1 watching Activity

Private repository

main

Branches Tags

github-classroom[bot] Initial commit ... last month 1

View code

README.md

DASC 5300 Assignment 2

Objective

The objective of this assignment is to equip you with the practical skills necessary to handle, process, and analyze large datasets efficiently. You should use libraries such as Pandas and Matplotlib for data manipulation and visualization, and investigate the performance implications of different sorting algorithms and data structures.

Dataset

Utilize the **New York City Taxi Trip Duration** dataset available on Kaggle. This dataset contains information on taxi trips in New York City over a specific period, with features such as pickup time, drop-off time, trip duration, passenger count, and geographical coordinates.

Tasks

1. **Data Pre-processing**:

   For the first part of the assignment, you will be working with a subset of the dataset containing 100,000 randomly sampled rows. Use the last 4 digits of your student ID as the random seed to ensure consistency across submissions. You can use the following code to load the dataset and perform the sampling:

```
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('train.csv')

# Set the random seed
np.random.seed(1234)

# Sample 100,000 rows
df = df.sample(n=100000)
```

   With the samples available, you should perform some preliminary analysis to understand the data. Some suggestions are

   - Provide summary statistics for the dataset.
   - Check the data types of the features.
   - Handle missing or erroneous data.
   - Convert categorical data to numerical values where necessary.
   - Normalize or standardize numerical features if required.
   - Look for errors or outliers in the data.

2. **Discovering Relationships**:

   - Employ correlation analysis to discover relationships between different features (`pandas` supports this with `corr`).
   - Apply regression analysis to identify significant predictors for trip duration.
   - Explore the possibility of creating new features that might be relevant for analysis.

   In your report, discuss the relationships discovered and provide visualizations to support your findings.

   - How might these relationships be useful for predicting trip duration?
   - What are some of the limitations of the analysis?

3. **Data Visualization**:
    - Create visualizations to represent the distribution of key features.
    - Visualize the relationship between different features and the trip duration.
    - Employ geographic visualizations to represent pickup and drop-off locations.

    In your report, discuss the visualizations used and explain how they either were or were not helpful in understanding the data.

4. **Algorithm and Data Structure Efficiency**:
    - Justify the choice of data structure (from arrays, stacks, queues, linked lists, and hash tables) to store the data for each of the following scenarios:
        - The dataset should be sorted and viewed in ascending order of trip duration. New data is added to the dataset frequently, where these new trips should show up at the end of the sorted list.
        - A new field is added to the data representing the passenger's phone number. It will be used to quickly filter out the trips made by a specific passenger.

5. **Final Analysis**:

    Based on your analysis from tasks 1-4, discuss which factors might have the greatest impact on trip duration. For a company that is looking to optimize the trip duration, what are some of the recommendations that you can provide?

Submission

Submit a report based on the requested items in the tasks above. Your submission should also include the code used for analysis (Python scripts and Jupyter Notebook files).

Evaluation Criteria

- Clarity and completeness of the data preprocessing and analysis.
- Insightfulness of the relationship discoveries and visualizations.
- Rationality and justification of the chosen sorting algorithm and data structure.
- Quality of the report presentation, including the organization, clarity, and aesthetics of visualizations, and code documentation.

Releases

No releases published

Create a new release

Packages

No packages published
Publish your first package