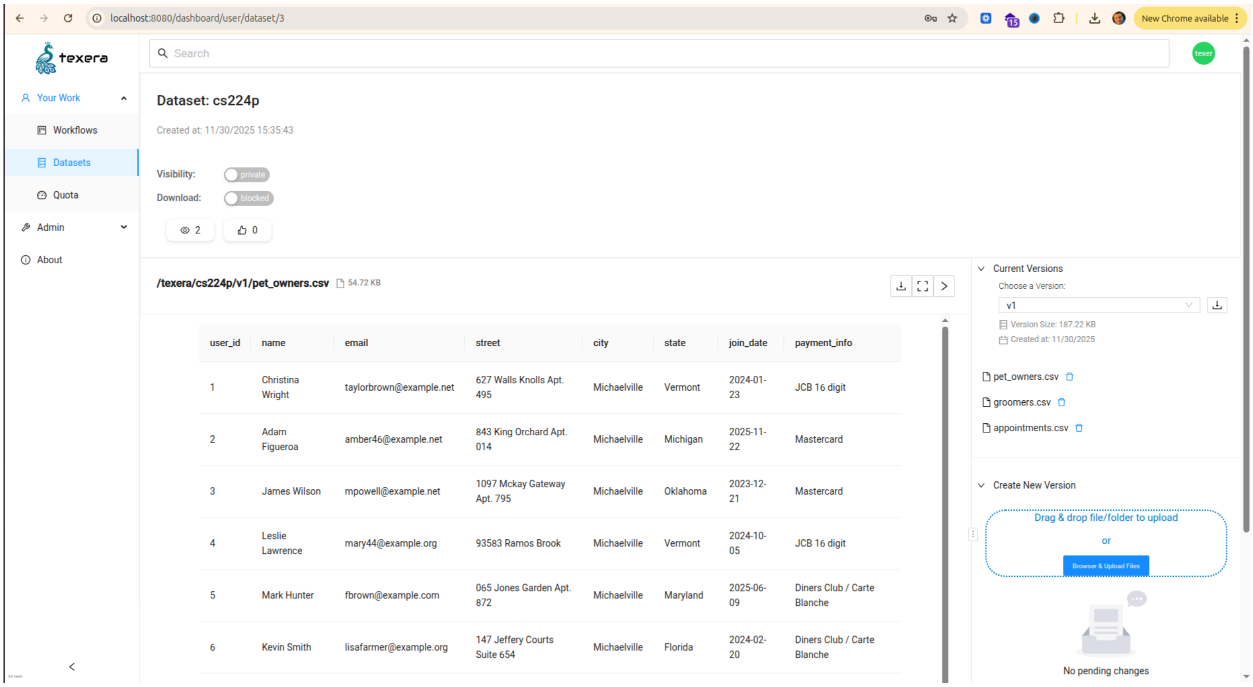
Homework Assignment #7

Deadline: Dec 7 (Sunday) 11:45PM

Setup

Please follow the instructions in this link. The following is a screenshot of the working environment of Texera you should expect.



Dataset: cs224p

Created at: 11/30/2025 15:35:43

Visibility: private

Download: blocked

/texera/cs224p/v1/pet_owners.csv 54.72 KB

user_id	name	email	street	city	state	join_date	payment_info
1	Christina Wright	taylorbrown@example.net	627 Walls Knolls Apt. 495	Michaelville	Vermont	2024-01-23	JCB 16 digit
2	Adam Figueroa	amber46@example.net	843 King Orchard Apt. 014	Michaelville	Michigan	2025-11-22	Mastercard
3	James Wilson	mpowell@example.net	1097 Mckay Gateway Apt. 795	Michaelville	Oklahoma	2023-12-21	Mastercard
4	Leslie Lawrence	mary44@example.org	93583 Ramos Brook	Michaelville	Vermont	2024-10-05	JCB 16 digit
5	Mark Hunter	fbrown@example.com	065 Jones Garden Apt. 872	Michaelville	Maryland	2025-06-09	Diners Club / Carte Blanche
6	Kevin Smith	lisafarmer@example.org	147 Jeffery Courts Suite 654	Michaelville	Florida	2024-02-20	Diners Club / Carte Blanche

Current Versions

Choose a Version: v1

Version Size: 187.22 KB

Created at: 11/30/2025

- pet_owners.csv
- groomers.csv
- appointments.csv

Create New Version

No pending changes

Questions

1. In this task, you will create a comprehensive workflow in the Texera system to analyze appointment data. You will work with three datasets: `pet_owners.csv`, `appointments.csv`, and `groomers.csv`.

Your workflow must branch into two distinct analytical paths to answer the following business questions:

1. **Michaelville Analysis:** Identify how many appointments were made by pet owners living in the city of `Michaelville`.
2. **State Spending Analysis:** Calculate the total amount spent on appointments in each state. You must join all three tables, aggregate the total cost by state, identify the top 5 states with the highest spending, and finally visualize the distribution.

Equivalent SQL Queries: Your workflow should logically represent the following SQL operations:

Part A (Michaelville Count):

```
SQL
SELECT COUNT(*)
FROM appointments A
JOIN pet_owners PO ON A.pet_owner_id = PO.user_id
WHERE PO.city = 'Michaelville';
```

Part B (Top 5 States & Visualization):

```
SQL
SELECT PO.state, SUM(A.cost) as total_cost
FROM appointments A
JOIN pet_owners PO ON A.pet_owner_id = PO.user_id
JOIN groomers G ON A.groomer_id = G.user_id
GROUP BY PO.state
ORDER BY total_cost DESC
LIMIT 5;
-- Result is then fed into a Pie Chart
```

Submission: Submit a screenshot of your workflow with the result of the pie chart.

2. The operations team wants to implement a "Smart Priority" system for handling future appointments. Simple filtering isn't enough; they need a calculated **Priority Score** that balances the **value** of the appointment (Cost) against its **urgency** (Time remaining).

Standard database operators can do basic arithmetic (`+`, `-`, `*`, `/`), but they cannot handle complex mathematical functions like logarithms or square roots, nor can they easily calculate time deltas between dates. You will use a **Python UDF** with the **NumPy** library to solve this.

The Logic: For every appointment scheduled **after January 1, 2025**, you must calculate a `priority_score` using the following non-linear formula:

$$\text{Priority Score} = \frac{\text{Cost}}{\ln(\text{DaysUntil} + 2)} + \sqrt{\text{Cost}}$$

Next, sort the result based on the priority score and get top 10 scores and show them in a bar chart.

Submission: Submit a screenshot of workflow with the final bar chart result and also add your Python code.

3. Build a Machine Learning workflow to predict the cost of an appointment based on the groomer's rating.

Hypothesis: We believe there is a linear relationship between a groomer's rating and their service cost. The formula we are approximating is:

Cost ≈ Base_Price + (Coefficient × Rating).

Workflow Requirements. You are required to construct a Texera workflow that performs the following logical steps. You must select the appropriate operators to achieve these goals:

1. **Data Preparation**
   - Ingest the appointments.csv and groomers.csv datasets.
   - Join them to associate every appointment with its groomer's rating.
2. **Training & Testing Split**
   - Split the joined data into two distinct sets:
     - Training Set: Appointments with appointment_id < 2500.
     - Test Set: Appointments with appointment_id >= 2500.
3. **Model Training**
   - Train a **Linear Regression** model using the **Training Set**.
   - Target: cost
   - Feature: rating
4. **Prediction**
   - Apply the trained model to the **Test Set** to generate predicted costs.
5. **Visualization**
   - Generate a Line Chart (or Scatter Plot) to visually compare the results.
   - Plot the Actual Cost (dots) and the Predicted Cost (line) against the rating.

- *Hint: Your visualization should show that the regression line fits the actual data points.*

Submission:

1. A screenshot of your complete ML workflow.
2. A screenshot of the final **Line Chart** visualization showing the comparison between actual and predicted values.