TDS Mock ROE 4

Doc Created: 28/02/2025

Please find the dataset `mock_roe_4.zip` in the link below.

Mock ROE 4 link

It has:

- A set of HTML files `biz-*.html` that has a list of restaurants in San Francisco.
- A violations table in `violations.db` SQLite DB with food violations found in restaurant inspections.

Question 1:

Of the restaurants present in postal code `94110`, how many had a **Moderate Risk** violation on a **Monday**?

Steps to answer:

1. From the violations table, count the violations grouped by `business_id` where the `risk_category` is **Moderate Risk** and `date` is on a **Monday**.
2. Scrape the `postal_code` and `business_id` from each HTML file.
3. Add the violations count for all the restaurants where the `postal_code` is `94110`.

Correct Answer: `249`

Question 2:

What is the highest average inspection score in the month `2015-05` that any latitude-longitude grid (rounded off to 2 decimal places) has received?

Steps to answer:

1. Scrape the `business_id`, `latitude`, and `longitude` from each HTML file. Drop missing values.
2. Scrape the `business_id` and `score` from the PDF file `inspections-2015-05.pdf`. Drop missing values.
3. Join these two datasets on `business_id`.

4. Round off the latitude and longitude to 2 decimal places.
5. Find the average `score` for each rounded-off latitude-longitude combination.
6. Pick the highest of these averages.

Correct Answer: `94.44`

Question 3:

Among these postal codes, which postal code has a restaurant furthest away from the centroid of the restaurants?

Postal Codes: `94116, 94127, 94117, 94114, 94115, 94132, 94108, 94111, 94105, 94134, 94131, 94109, 94121, 94107, 94133`

Steps to answer:

1. Scrape the `postal_code`, `latitude`, and `longitude` from each HTML file having your postal code. Drop missing values.
2. For each postal code:
   - Calculate the average of the latitude and longitude values for all rows in each `postal_code`. This is (roughly) the centroid.
   - Calculate the **Pythagorean** distance between each restaurant and the centroid.
   - Calculate the average of these distances.
   - Pick the postal code with the restaurant having the highest **AVERAGE_DISTANCE_FROM_CENTROID**.

Correct Answer: `94121`

Question 4:

Find the number of violations that were reported without an inspection (same business, same date) on or after `2016-02-23` in the **Moderate Risk** risk category.

Steps to answer:

1. Scrape the `business_id` and `date` from the `inspections-*.pdf` files. Drop missing values.
2. Extract the `business_id` and `date` from the `violations` table where the `risk_category` is **Moderate Risk** and `date` is on or after `2016-02-23`.
3. Count the number of violations where the `business_id` and `date` are NOT in the `inspections-*.pdf` files.

Correct Answer: `11`

Question 5:

How many businesses in postal code `94110` had a violation that contained one or more of the words **water, unapproved, moderate, facilities, unsanitary** and an associated inspection (same date, same `business_id`) with a score of **80 or more**?

Steps to answer:

1. Scrape the `business_id` and `postal_code` from each HTML file. Drop missing values.
2. Scrape the `business_id`, `date`, and `score` from the PDF file `inspections-*.pdf`. Drop missing values.
3. Extract the `business_id`, `date`, and `description` from the `violations` table where the `description` contains one or more of the words **water, unapproved, moderate, facilities, unsanitary**.
4. Join the `biz-*.html`, the `inspections-*.pdf`, and the `violations` table data by matching `business_id` across all three datasets and the `date` across `inspections-*.pdf` and the `violations` table.
5. Filter the joined data where the description contains one or more of the words **water, unapproved, moderate, facilities, unsanitary** and the `score` is **80 or more**.
6. Find the postal code of these businesses and filter those matching `94110`.

Correct Answer: `256`

Question 6:

Within the latitude-longitude bounds of `37.7, 37.900000000000006, -122.4, and -122.2`, count the businesses with the most dissimilar description.

Steps to answer:

1. Scrape the `business_id`, `latitude`, `longitude`, and `description` from each HTML file. Drop missing values.
2. Also drop zero values for latitude or longitude.
3. Extract the `business_id` and `description` from the `violations` table.
4. Join the data from the HTML files and violations table on `business_id`.
5. Filter the joined data where the latitude is between `37.7` and `37.900000000000006` and the longitude is between `-122.4` and `-122.2`.

6. Calculate the vector embeddings of all the descriptions using `text-embedding-3-small`.
7. Find the centroid of the embeddings by averaging all the vector embeddings.
8. Find the most dissimilar embeddings (highest Pythagorean distance from the centroid).
9. Count the number of **UNIQUE** `business_id`s that have this most dissimilar embedding.

Correct Answer: `7`

Question 7:

Using **linear regression**, predict the inspection score of a restaurant in these postal codes: `94121, 94133, 94116, 94103, 94117` on `2016-10-10`.

Steps to answer:

1. Scrape the `business_id`, `date`, and `score` from the PDF files `inspections-*.pdf`. Drop missing values.
2. Scrape the `business_id` and `postal_code` from the HTML files `biz-*.html` for the postal codes mentioned above. Drop missing values.
3. Join the inspections data with the HTML data on `business_id`, combining data across all the above postal codes.
4. Calculate the regression slope of the inspection scores (`Y`) against the date (`X`).
5. Predict the inspection score for the date `2016-10-10`.

Correct Answer: `87.94`