**Week 15 Assignment:**

**VERA MPENGULA**

**1. Data Dive:**

Working with the Netflix Shows dataset presented a few challenges during the data import and analysis process. One of the main difficulties was handling missing values, especially in columns crucial for analysis. This required careful consideration of data imputation techniques or exclusion of incomplete records to maintain data integrity. Additionally, ensuring consistency in data types and formats across columns was another aspect that required attention to avoid issues during querying and visualization tasks. One fascinating aspect observed in the dataset was the diversity of genres and categories represented in Netflix shows. From classic genres like drama and comedy to niche categories like stand-up comedy, reality TV, and documentaries, the dataset reflects the wide range of content available on the platform. This diversity not only showcases Netflix's extensive content library but also highlights the evolving preferences and interests of viewers worldwide

**2. Data Fun:**

There’s a total of one hundred shows and the average duration for each show is 56.46 minutes.  After analyzing the data, I have concluded that Toshiya Shinohara has directed the most shows.

This is the code I used:

**SELECT**

**COUNT(\*) AS total\_shows,**

**AVG(rating) AS average\_rating,**

**AVG(duration) AS average\_duration**

**FROM**

**Netflixshows\_csv.netflixshows;**

**-- Fact 1: The Year with the Most Releases**

**SELECT release\_year, COUNT(\*) AS release\_count**

**FROM netflixshows\_csv.netflixshows**

**GROUP BY release\_year**

**ORDER BY release\_count DESC**

**LIMIT 1;**

**-- Fact 2: Top 5 Directors with the Most Shows**

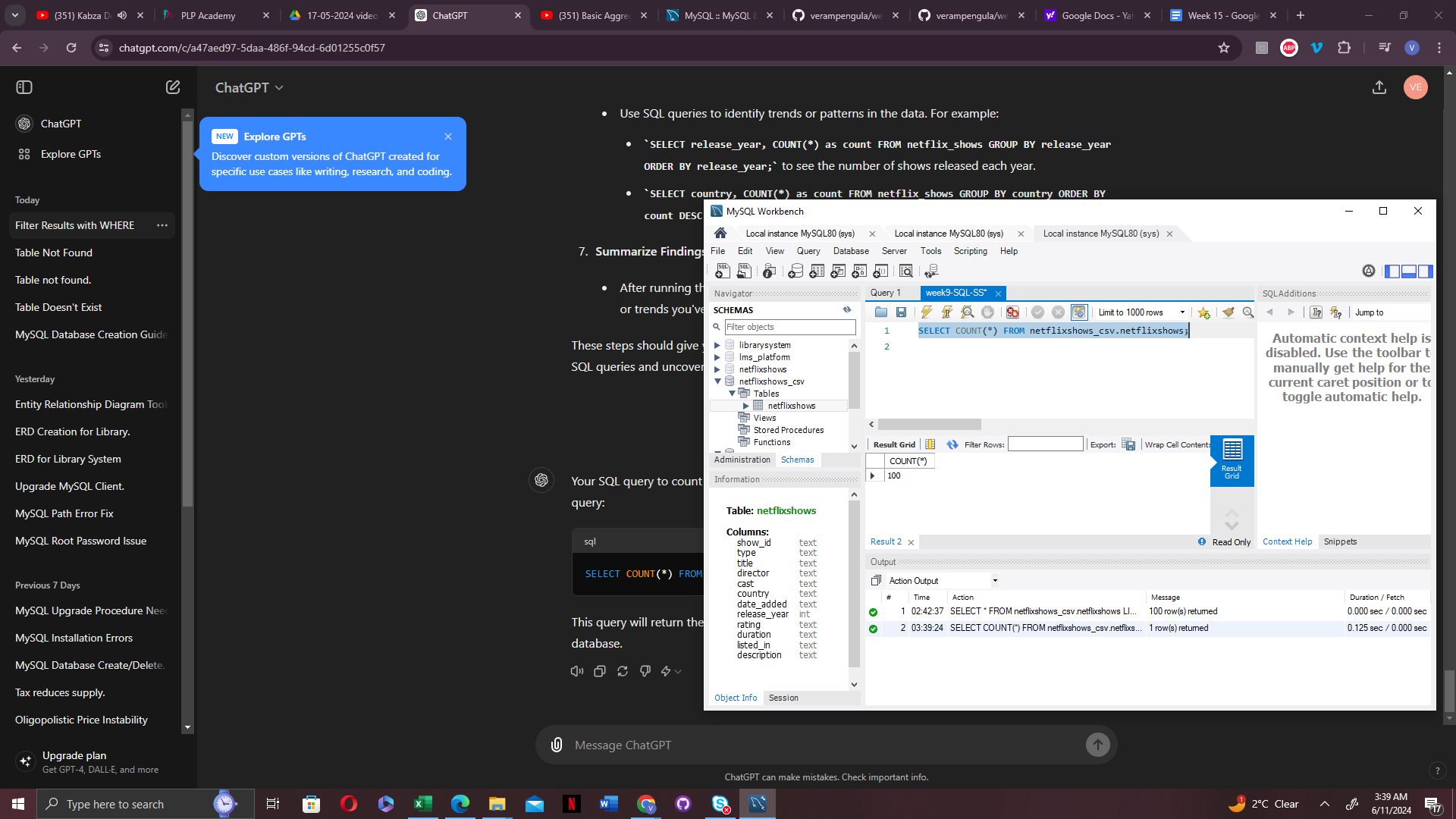
**SELECT director, COUNT(\*) AS show\_count**

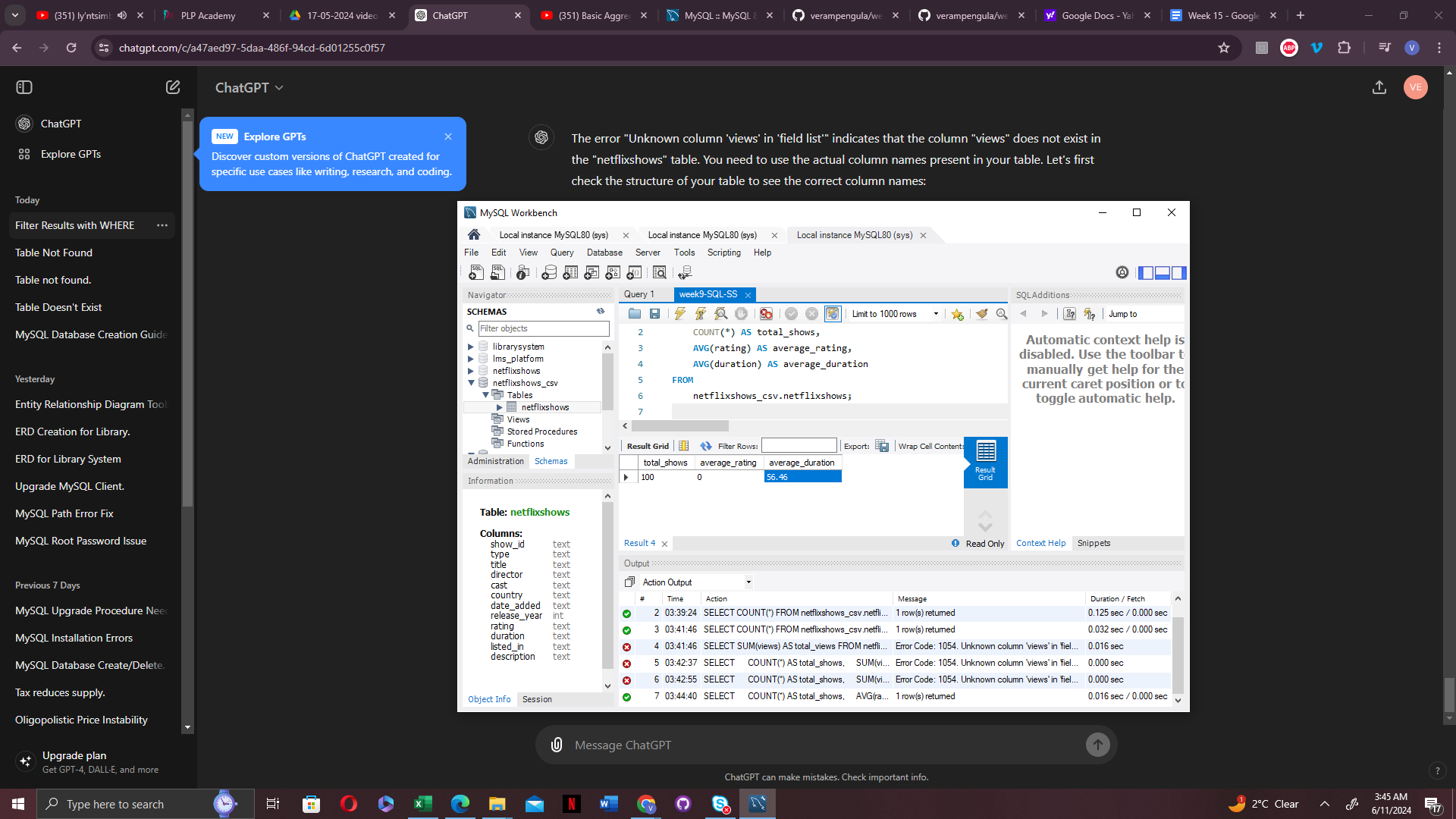
**FROM netflixshows\_csv.netflixshows**

**GROUP BY director**

**ORDER BY show\_count DESC**

**LIMIT 5;**





**3. Ask Away:**

1. **What are the top 5 highest rated shows?**

**The sql query I used to answer this question:**

**SELECT title, rating**

**FROM netflixshows\_csv.netflixshows**

**ORDER BY rating DESC**

**LIMIT 5;**

Here is the output:

A screenshot of a computer

Description automatically generated

1. How many movies and tv shows are there in the data set?

**SELECT**

**CASE**

**WHEN type = 'TV Show' THEN 'TV Show'**

**ELSE 'Movie'**

**END AS show\_type,**

**COUNT(\*) AS show\_count,**

**AVG(duration) AS average\_duration**

**FROM**

**netflixshows\_csv.netflixshows**

**GROUP BY**

**show\_type;**

The output is screenshoted below. there was a total of 100 shows, which can be broken down into 55 movies and 45 tv shows

