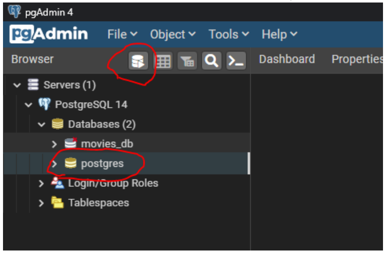
Master SQL for Data Analysis - Level 1

Project Solution

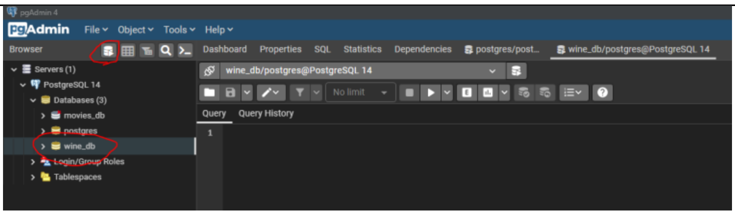
Phase 1 - Dataset Preparation

Welcome to our final project exercise. We are planning to load a dataset about **wines rating and prices** and then perform multiple queries while exploring and analyzing the dataset (data source – Kaggle/ Vivino.com).

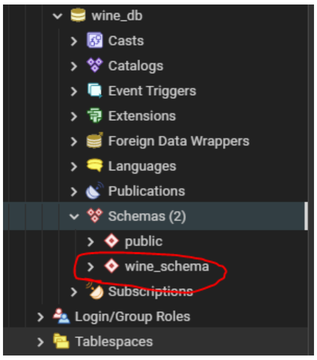
Please review the following steps as part of the initial data preparation:

<table>
<tr><th>Step</th><th>Description</th></tr>
<tr><td>1</td><td>Download the dataset file from the course resources. The CSV file is called “WineDataset”. You can open it using Excel and quickly review the dataset columns:
<table>
<tr><th>Attribute</th><th>Description</th></tr>
<tr><td>Type</td><td>Type of wine</td></tr>
<tr><td>Name</td><td>Name of wine</td></tr>
<tr><td>Country</td><td>Origin Country</td></tr>
<tr><td>Region</td><td>Origin region or province</td></tr>
<tr><td>Winery</td><td>Origin winery</td></tr>
<tr><td>Rating</td><td>Average rating</td></tr>
<tr><td>NumberOfRatings</td><td>Number of people who rated this wine</td></tr>
<tr><td>Price</td><td>Price in EUR</td></tr>
<tr><td>Year</td><td>Year of production</td></tr>
</table>
</td></tr>
<tr><td>2</td><td>Open PostgreSQL admin console (pgAdmin), select the default “postgres” database on the left side, and then open the query tool:


From the query tool, create the following database objects using the CREATE command:
<ul><li>Create a new database called “wine_db” → refresh the list of databases to view the new database:</li></ul>
<u>Answer:</u> CREATE DATABASE wine_db;
<ul><li>Select the new database and open a new query tool from the “wine_db”.</li></ul>
</td></tr>
</table>



- Create a new <span>database schema</span> called “wine_schema”, inside the “wine_db” database.



Answer: CREATE SCHEMA wine_schema;

- Create a new table called “wine_table” inside the “wine_schema” schema using the following list of attributes. Please note that there are constraints on some of the columns.

Attribute	Data Type	Constraint
WineIndex	Integer	PRIMARY
Type	varchar(10)	
Name	varchar(200)	
Country	varchar(50)	
Region	varchar(50)	
Winery	varchar(50)	
Rating	decimal(2,1)	
NumberOfRatings	Integer	
Price	decimal(5,2)	Price>0
Year	Integer	Year>=1950

Answer:

```
CREATE TABLE wine_schema.wine_table
(
   WineIndex integer PRIMARY KEY,
   Type varchar(10),
   Name varchar(200),
   Country varchar(50),
   Region varchar(50),
   Winery varchar(100),
   Rating decimal(2,1),
   NumOfRating integer,
   Price decimal(6,2) CHECK (Price>0),
   Year integer CHECK (Year>=1950)
)
```

3	Upload the wine dataset CSV file into the new table “wine_table” using the COPY command in PostgreSQL. Answer: COPY wine_schema.wine_table (WineIndex, Type, Name, Country, Region, Winery, Rating, NumOfRating, Price, Year) FROM 'c:\data\wine\WineDataset.csv' DELIMITER ',' CSV HEADER;

Great, now we are ready to move into data analysis!

****

Phase 2 - Data Analysis

Exercise #1 - Query all columns from the wine table with a limit of getting only 20 records.

Answer:

```
SELECT *

FROM wine_schema.wine_table

LIMIT 20
```

Exercise #2 - Query the following columns: Type, Name, Country, Rating from the wine table with a limit of 20 records.

Answer:

```
SELECT Type, Name, Country, Rating

FROM wine_schema.wine_table

LIMIT 20
```

Exercise #3 - What are the distinct wine types?

Answer:

```
SELECT DISTINCT type

FROM wine_schema.wine_table
```

Exercise #4 - Calculate the number of distinct wine types.

Answer:

```
SELECT COUNT(DISTINCT type) AS num_wine_types

FROM wine_schema.wine_table
```

Exercise #5 - Calculate the number of distinct countries producing Sparkling wines.

Answer:

```
SELECT COUNT(DISTINCT Country) AS distinct_countries

FROM wine_schema.wine_table

WHERE type = 'Sparkling'
```

Exercise #6 – List the number of wines produced per country in descending order.

Answer:

```
SELECT Country, COUNT(DISTINCT Name) AS distinct_wines

FROM wine_schema.wine_table

GROUP BY 1
```

```
ORDER BY 2 DESC
```

Exercise #7 – What is the average price per each wine type? Round the number to 2 decimal places and order the average price result in ascending order (tip – use the ROUND function).

Answer:

```
SELECT type, ROUND(AVG(Price),2) AS avg_price

FROM wine_schema.wine_table

GROUP BY 1

ORDER BY 2
```

Exercise #8 – What is the average price by year? Order the result in ascending order based on the Year. Exclude NULL values in the Year column from the group-level result.

Answer:

```
SELECT Year, ROUND(AVG(Price),2) AS avg_price

FROM wine_schema.wine_table

GROUP BY 1

HAVING Year IS NOT NULL

ORDER BY 1
```

Exercise #9 – What are the average price and average rating by country? Order by the Country name.

Answer:

```
SELECT Country, ROUND(AVG(Price),2) AS avg_price , ROUND(AVG(Rating),2) AS avg_rating

FROM wine_schema.wine_table

GROUP BY 1

ORDER BY 1
```

Exercise #10 – What are the average price and average rating by year for Italy? Exclude NULL values in the Year column from the raw table before grouping.

Answer:

```
SELECT Year, ROUND(AVG(Price),2) AS avg_price , ROUND(AVG(Rating),2) AS avg_rating

FROM wine_schema.wine_table

WHERE Country = 'Italy'

GROUP BY 1

ORDER BY 1
```

Exercise #11 – What is the average price by country and by region in each country for the following countries: Argentina, Canada, Italy, Greece? Order the result based on the Country ascending and secondly based on the average price in a region descending.

Answer:

```
SELECT Country, Region, ROUND(AVG(Price),2) AS avg_price_region

FROM wine_schema.wine_table

WHERE Country IN ('Argentina', 'Canada', 'Italy', 'Greece')

GROUP BY 1, 2

ORDER BY 1, 3 DESC
```

Exercise #12 – How many wines are available per each rating?

Answer:

```
SELECT Rating, COUNT(DISTINCT Name) AS amount_wines

FROM wine_schema.wine_table

GROUP BY 1

ORDER BY 1
```

Exercise #13 – How many wines of each wine type were produced in each country?

Answer:

```
SELECT Country, Type, COUNT(DISTINCT Name) AS amount_wines

FROM wine_schema.wine_table

GROUP BY 1, 2

ORDER BY 1, 2
```

Exercise #14 – What is the maximum price per each wine type excluding the following years – 2011, 2013, 2015, 2018)? Order by maximum price in descending order.

Answer:

```
SELECT Type, MAX(Price) AS max_price

FROM wine_schema.wine_table

WHERE Year NOT IN ('2011','2013','2015','2018')

GROUP BY 1

ORDER BY 2 DESC
```

Exercise #15 - What are the names and country locations of the top 10 red wines with the highest rating?

Answer:

```
SELECT Type, Name, Country, Rating

FROM wine_schema.wine_table

WHERE Type = 'Red'

ORDER BY Rating DESC

LIMIT 10
```

Exercise #16 – List the 10 top Wineries in France that have the highest rating excluding wines with a number of reviews below 200.

Answer:

```
SELECT Winery, Rating

FROM wine_schema.wine_table

WHERE (Country = 'France' AND numofrating >= 200)

ORDER BY Rating DESC

LIMIT 10
```

Exercise #17 – Which group of wine types has the highest average rating for wines that were produced between 2000 and 2010 or between 2015 and 2020.

Answer:

```
SELECT Type, ROUND(AVG(Rating),2) AS avg_rating

FROM wine_schema.wine_table

WHERE (Year BETWEEN 2000 AND 2010) OR (Year BETWEEN 2015 AND 2020)

GROUP BY 1

ORDER BY 2 DESC

LIMIT 1
```

Exercise #18 – What are the five top countries with the highest average rating for wines that are above the price of 20 Euro?

Answer:

```
SELECT Country, ROUND(AVG(Rating),2) AS avg_rating

FROM wine_schema.wine_table

WHERE PRICE > 20

GROUP BY 1

ORDER BY 2 DESC

LIMIT 5
```

Exercise #19 – What are the top 20 regions that produce the highest number of wines with a minimum of 50 wines, where the price of a wine is below 300 EURO, and the number of rating reviews for the wine is more than 100?

Answer:

```
SELECT Region, COUNT(Name) AS amount_wines

FROM wine_schema.wine_table

WHERE Price < 300 AND numofrating > 100

GROUP BY 1

HAVING COUNT(Name) > 100

ORDER BY 2 DESC

LIMIT 20
```

****

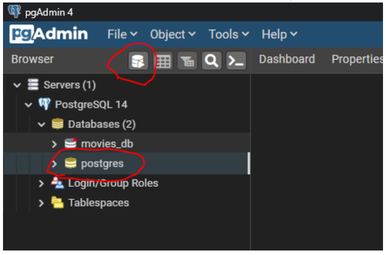
Master SQL for Data Analysis - Level 1

Your Final Project

Phase 1 - Dataset Preparation

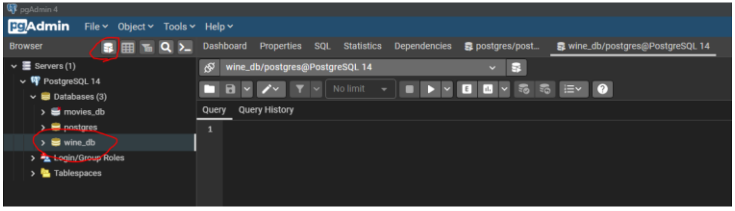
Welcome to our final project exercise. We are planning to load a dataset about **wines rating and prices** and then perform multiple queries while exploring and analyzing the dataset (data source – Kaggle/ Vivino.com).

Please review the following steps as part of the initial data preparation:

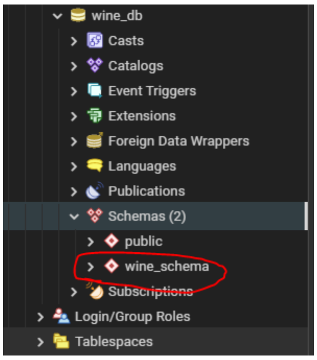
<table>
<tr><th>Step</th><th>Description</th></tr>
<tr><td>1</td><td>Download the dataset file from the course resources. The CSV file is called “WineDataset”. You can open it using Excel and quickly review the dataset columns:
<table>
<tr><th>Attribute</th><th>Description</th></tr>
<tr><td>Type</td><td>Type of wine</td></tr>
<tr><td>Name</td><td>Name of wine</td></tr>
<tr><td>Country</td><td>Origin Country</td></tr>
<tr><td>Region</td><td>Origin region or province</td></tr>
<tr><td>Winery</td><td>Origin winery</td></tr>
<tr><td>Rating</td><td>Average rating</td></tr>
<tr><td>NumberOfRatings</td><td>Number of people who rated this wine</td></tr>
<tr><td>Price</td><td>Price in EUR</td></tr>
<tr><td>Year</td><td>Year of production</td></tr>
</table>
</td></tr>
<tr><td>2</td><td>Open PostgreSQL admin console (pgAdmin), select the default “postgres” database on the left side, and then open the query tool:


From the query tool, create the following database objects using the CREATE command:
<ul><li>Create a new database called “wine_db” → refresh the list of databases to view the new database:</li></ul>
Answer:</td></tr>
</table>

Select the new database and open a new query tool from the “wine_db”.



- Create a new database schema called “wine_schema”, inside the “wine_db” database.



Answer:

- Create a new table called “wine_table” inside the “wine_schema” schema using the following list of attributes. Please note that there are constraints on some of the columns.

Attribute	Data Type	Constraint
WineIndex	Integer	PRIMARY
Type	varchar(10)	
Name	varchar(200)	
Country	varchar(50)	
Region	varchar(50)	
Winery	varchar(50)	
Rating	decimal(2,1)	
NumberOfRatings	Integer	
Price	decimal(5,2)	Price>0
Year	Integer	Year>=1950

Answer:

3 Upload the wine dataset CSV file into the new table “wine_table” using the COPY command in PostgreSQL.

Answer:

Great, now we are ready to move into data analysis!

****

Phase 2 - Data Analysis

Exercise #1 - Query all columns from the wine table with a limit of getting only 20 records.

Answer:

Exercise #2 - Query the following columns: Type, Name, Country, Rating from the wine table with a limit of 20 records.

Answer:

Exercise #3 - What are the distinct wine types?

Answer:

Exercise #4 - Calculate the number of distinct wine types.

Answer:

Exercise #5 - Calculate the number of distinct countries producing Sparkling wines.

Answer:

Exercise #6 – List the number of wines produced per country in descending order.

Answer:

Exercise #7 – What is the average price per each wine type? Round the number to 2 decimal places and order the average price result in ascending order (tip – use the ROUND function).

Answer:

Exercise #8 – What is the average price by year? Order the result in ascending order based on the Year. Exclude NULL values in the Year column from the group-level result.

Answer:

Exercise #9 – What are the average price and average rating by country? Order by the Country name.

Answer:

Exercise #10 – What are the average price and average rating by year for Italy? Exclude NULL values in the Year column from the raw table before grouping.

Answer:

Exercise #11 – What is the average price by country and by region in each country for the following countries: Argentina, Canada, Italy, Greece? Order the result based on the Country ascending and secondly based on the average price in a region descending.

Answer:

Exercise #12 – How many wines are available per each rating?

Answer:

Exercise #13 – How many wines of each wine type were produced in each country?

Answer:

Exercise #14 – What is the maximum price per each wine type excluding the following years – 2011, 2013, 2015, 2018)? Order by maximum price in descending order.

Answer:

Exercise #15 - What are the names and country locations of the top 10 red wines with the highest rating?

Answer:

Exercise #16 – List the 10 top Wineries in France that have the highest rating excluding wines with a number of reviews below 200.

Answer:

Exercise #17 – Which group of wine types has the highest average rating for wines that were produced between 2000 and 2010 or between 2015 and 2020.

Answer:

Exercise #18 – What are the five top countries with the highest average rating for wines that are above the price of 20 Euro?

Answer:

Exercise #19 – What are the top 20 regions that produce the highest number of wines with a minimum of 50 wines, where the price of a wine is below 300 EURO, and the number of rating reviews for the wine is more than 100?

Answer:

****