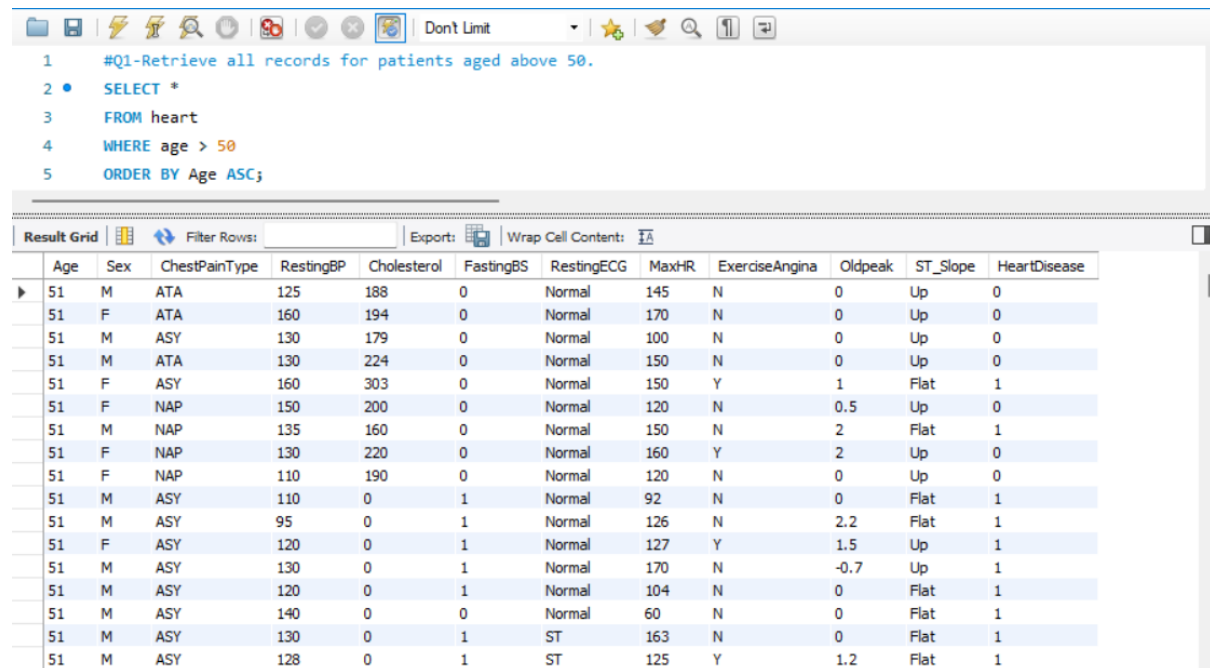


Heart Failure Prediction

SQL Queries:

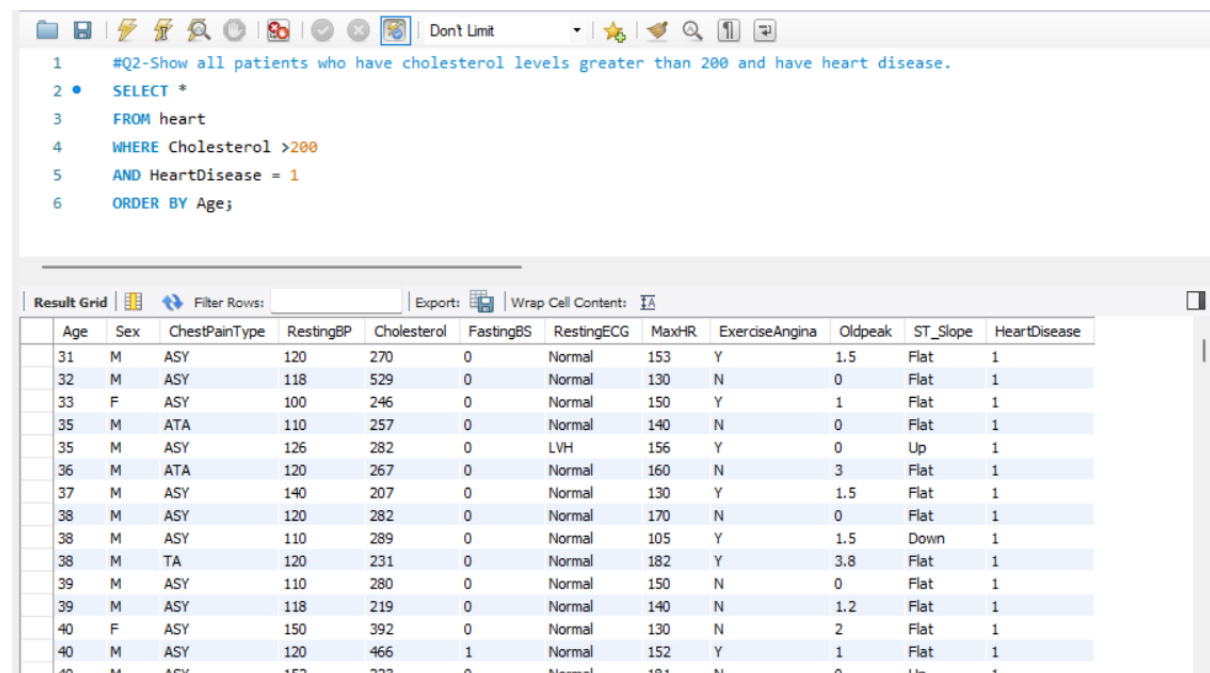
1) Retrieve all records for patients aged above 50.



```
1 #Q1-Retrieve all records for patients aged above 50.
2 • SELECT *
3 FROM heart
4 WHERE age > 50
5 ORDER BY Age ASC;
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
▶	51	M	ATA	125	188	0	Normal	145	N	0	Up	0
	51	F	ATA	160	194	0	Normal	170	N	0	Up	0
	51	M	ASY	130	179	0	Normal	100	N	0	Up	0
	51	M	ATA	130	224	0	Normal	150	N	0	Up	0
	51	F	ASY	160	303	0	Normal	150	Y	1	Flat	1
	51	F	NAP	150	200	0	Normal	120	N	0.5	Up	0
	51	M	NAP	135	160	0	Normal	150	N	2	Flat	1
	51	F	NAP	130	220	0	Normal	160	Y	2	Up	0
	51	F	NAP	110	190	0	Normal	120	N	0	Up	0
	51	M	ASY	110	0	1	Normal	92	N	0	Flat	1
	51	M	ASY	95	0	1	Normal	126	N	2.2	Flat	1
	51	F	ASY	120	0	1	Normal	127	Y	1.5	Up	1
	51	M	ASY	130	0	1	Normal	170	N	-0.7	Up	1
	51	M	ASY	120	0	1	Normal	104	N	0	Flat	1
	51	M	ASY	140	0	0	Normal	60	N	0	Flat	1
	51	M	ASY	130	0	1	ST	163	N	0	Flat	1
	51	M	ASY	128	0	1	ST	125	Y	1.2	Flat	1

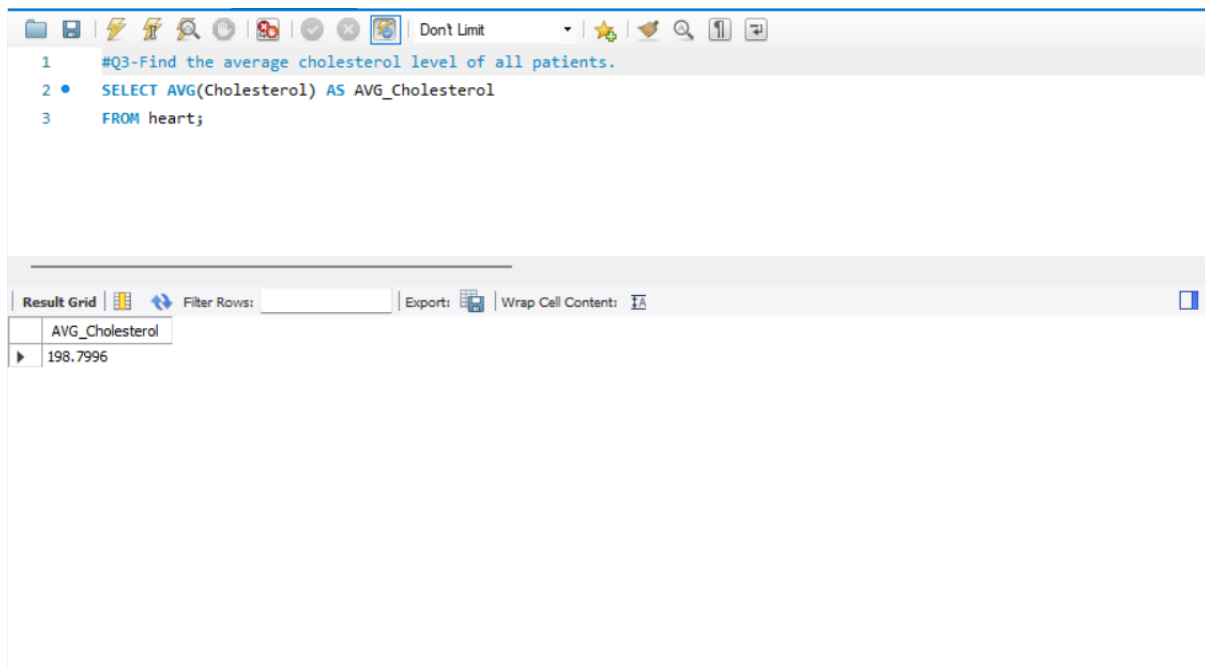
2) Show all patients who have cholesterol levels greater than 200 and have heart disease.



```
1 #Q2-Show all patients who have cholesterol levels greater than 200 and have heart disease.
2 • SELECT *
3 FROM heart
4 WHERE Cholesterol >200
5 AND HeartDisease = 1
6 ORDER BY Age;
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
	31	M	ASY	120	270	0	Normal	153	Y	1.5	Flat	1
	32	M	ASY	118	529	0	Normal	130	N	0	Flat	1
	33	F	ASY	100	246	0	Normal	150	Y	1	Flat	1
	35	M	ATA	110	257	0	Normal	140	N	0	Flat	1
	35	M	ASY	126	282	0	LVH	156	Y	0	Up	1
	36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
	38	M	ASY	120	282	0	Normal	170	N	0	Flat	1
	38	M	ASY	110	289	0	Normal	105	Y	1.5	Down	1
	38	M	TA	120	231	0	Normal	182	Y	3.8	Flat	1
	39	M	ASY	110	280	0	Normal	150	N	0	Flat	1
	39	M	ASY	118	219	0	Normal	140	N	1.2	Flat	1
	40	F	ASY	150	392	0	Normal	130	N	2	Flat	1
	40	M	ASY	120	466	1	Normal	152	Y	1	Flat	1
	40	M	ASY	152	223	0	Normal	181	N	0	Lin	1

3) Find the average cholesterol level of all patients.

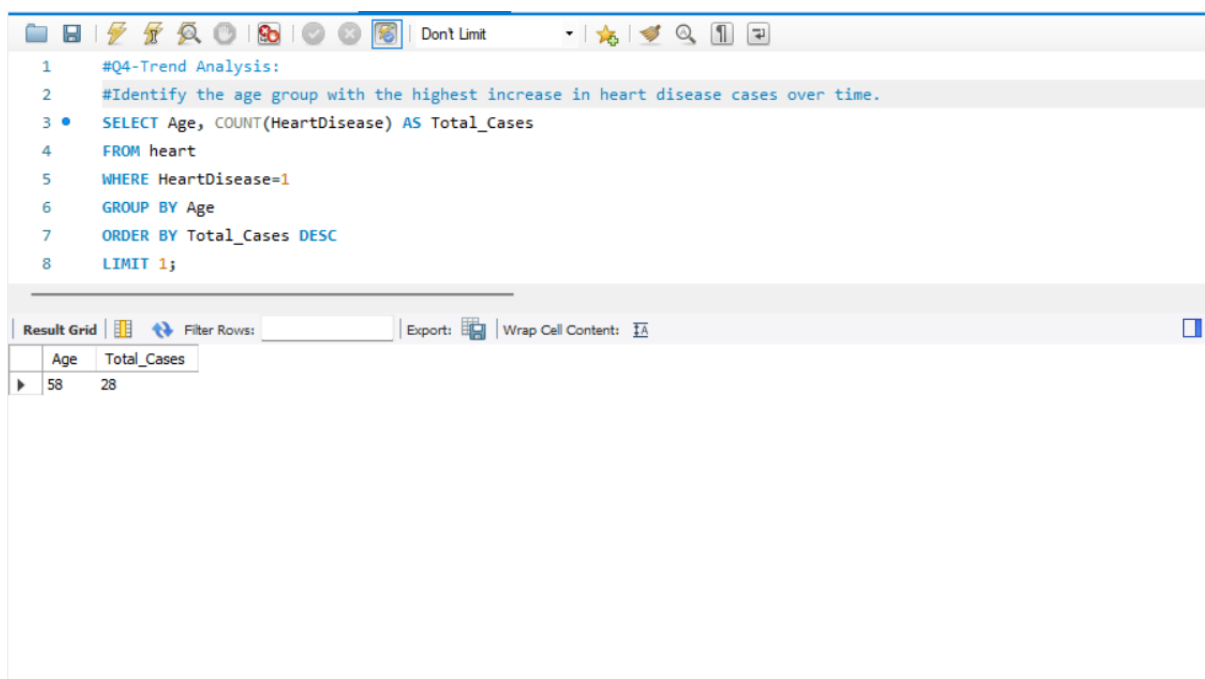


```
1 #Q3-Find the average cholesterol level of all patients.
2 • SELECT AVG(Cholesterol) AS AVG_Cholesterol
3 FROM heart;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: |

AVG_Cholesterol
198.7996

4) **Trend Analysis:** Identify the age group with the highest increase in heart disease cases over time.



```
1 #Q4-Trend Analysis:
2 #Identify the age group with the highest increase in heart disease cases over time.
3 • SELECT Age, COUNT(HeartDisease) AS Total_Cases
4 FROM heart
5 WHERE HeartDisease=1
6 GROUP BY Age
7 ORDER BY Total_Cases DESC
8 LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: |

Age	Total_Cases
58	28

5) Group data by gender to calculate the average resting blood pressure for each gender.

The screenshot shows a SQL query editor with the following code:

```
1 #Q5-Group data by gender to calculate the average resting blood pressure for each gender.
2 • SELECT Sex, AVG(RestingBP) AS AVG_RestingBP
3 FROM heart
4 GROUP BY Sex;
5
```

Below the query editor is a 'Result Grid' with the following data:

	Sex	AVG_RestingBP
▶	M	132.4455
	F	132.2124

6) **Ranking:** Rank patients by their cholesterol levels within each gender.

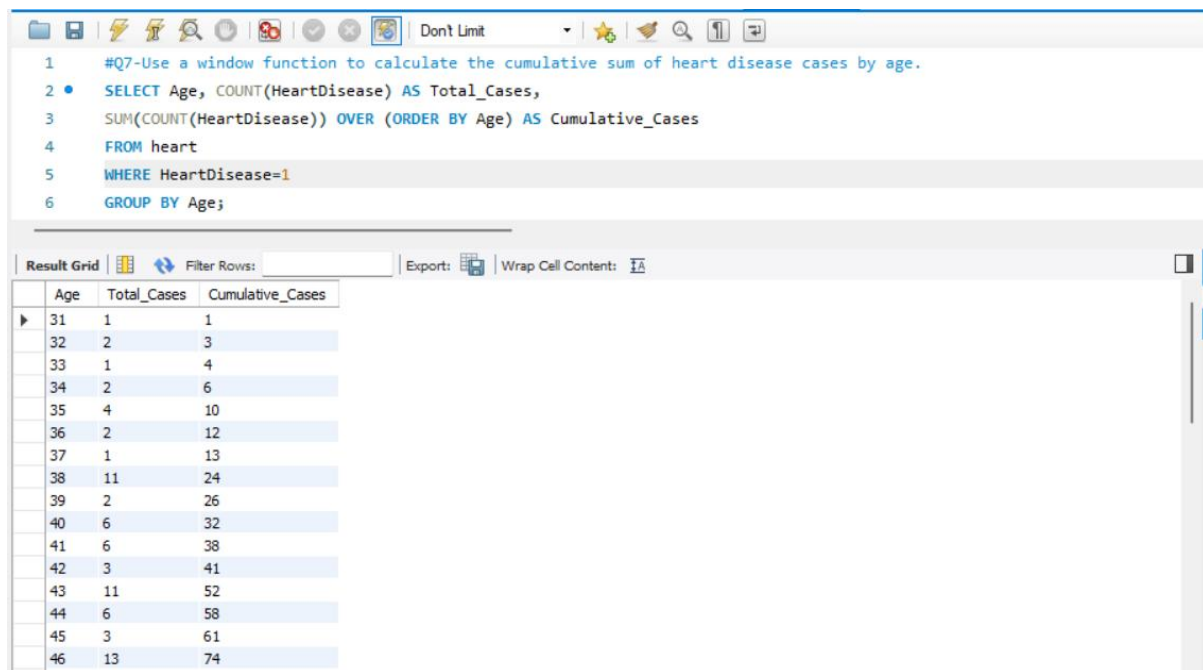
The screenshot shows a SQL query editor with the following code:

```
1 #Q6-Ranking: Rank patients by their cholesterol levels within each gender.
2 • SELECT Age, Sex, Cholesterol,
3 RANK() OVER (PARTITION BY Sex ORDER BY Cholesterol DESC) AS Cholesterol_Rank
4 FROM heart;
5
```

Below the query editor is a 'Result Grid' with the following data:

	Age	Sex	Cholesterol	Cholesterol_Rank
▶	67	F	564	1
	53	F	468	2
	65	F	417	3
	56	F	409	4
	63	F	407	5
	55	F	394	6
	62	F	394	6
	58	F	393	8
	40	F	392	9
	65	F	360	10
	57	F	354	11
	55	F	344	12
	55	F	344	13
	55	F	342	14
	43	F	341	15
	58	F	340	16
	59	F	338	17
	64	F	333	18

7) Use a window function to calculate the cumulative sum of heart disease cases by age.



The screenshot shows a SQL IDE window with a query editor and a result grid. The query uses a window function to calculate the cumulative sum of heart disease cases by age.

```

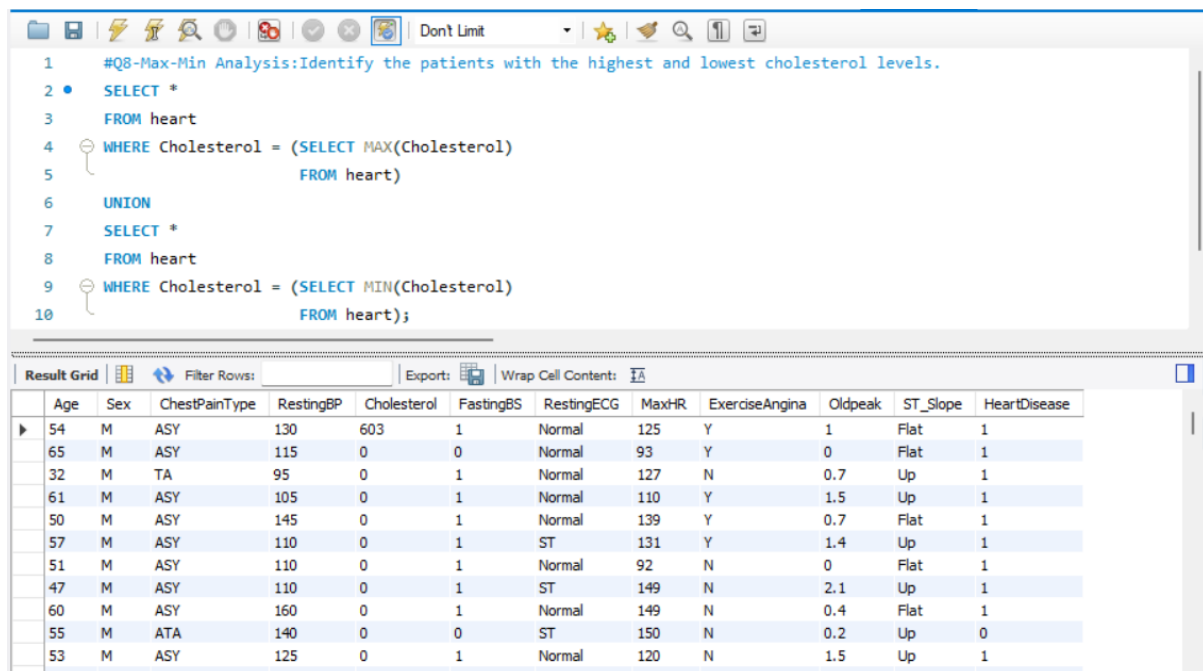
1 #Q7-Use a window function to calculate the cumulative sum of heart disease cases by age.
2 • SELECT Age, COUNT(HeartDisease) AS Total_Cases,
3    SUM(COUNT(HeartDisease)) OVER (ORDER BY Age) AS Cumulative_Cases
4 FROM heart
5 WHERE HeartDisease=1
6 GROUP BY Age;

```

The result grid displays the following data:

Age	Total_Cases	Cumulative_Cases
31	1	1
32	2	3
33	1	4
34	2	6
35	4	10
36	2	12
37	1	13
38	11	24
39	2	26
40	6	32
41	6	38
42	3	41
43	11	52
44	6	58
45	3	61
46	13	74

8) **Max-Min Analysis:** Identify the patients with the highest and lowest cholesterol levels.



The screenshot shows a SQL IDE window with a query editor and a result grid. The query identifies patients with the highest and lowest cholesterol levels using a subquery and UNION.

```

1 #Q8-Max-Min Analysis:Identify the patients with the highest and lowest cholesterol levels.
2 • SELECT *
3 FROM heart
4 WHERE Cholesterol = (SELECT MAX(Cholesterol)
5                      FROM heart)
6 UNION
7 SELECT *
8 FROM heart
9 WHERE Cholesterol = (SELECT MIN(Cholesterol)
10                     FROM heart);

```

The result grid displays the following data:

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
54	M	ASY	130	603	1	Normal	125	Y	1	Flat	1
65	M	ASY	115	0	0	Normal	93	Y	0	Flat	1
32	M	TA	95	0	1	Normal	127	N	0.7	Up	1
61	M	ASY	105	0	1	Normal	110	Y	1.5	Up	1
50	M	ASY	145	0	1	Normal	139	Y	0.7	Flat	1
57	M	ASY	110	0	1	ST	131	Y	1.4	Up	1
51	M	ASY	110	0	1	Normal	92	N	0	Flat	1
47	M	ASY	110	0	1	ST	149	N	2.1	Up	1
60	M	ASY	160	0	1	Normal	149	N	0.4	Flat	1
55	M	ATA	140	0	0	ST	150	N	0.2	Up	0
53	M	ASY	125	0	1	Normal	120	N	1.5	Up	1

9) **Subquery:** Find the patients who have a cholesterol level higher than the average cholesterol of all patients.

1 #Q9-Subquery: Find the patients who have a cholesterol level higher than the average cholesterol of all patients.
2 • SELECT *
3 FROM heart
4 WHERE Cholesterol > (SELECT AVG(Cholesterol)
5 FROM heart);

Result Grid Filter Rows: Export: Wrap Cell Content:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
▶	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
	37	M	ATA	130	283	0	ST	98	N	0	Up	0
	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
	48	F	ATA	120	284	0	Normal	120	N	0	Up	0
	39	F	NAP	130	211	0	Normal	142	N	0	Up	0
	37	M	ATA	120	204	0	Normal	145	N	0	Up	0
	49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
	42	F	NAP	115	211	0	ST	137	N	0	Up	0
	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
	43	F	ATA	120	201	0	Normal	165	N	0	Up	0
	60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
	36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
	43	F	TA	100	223	0	Normal	142	N	0	Up	0

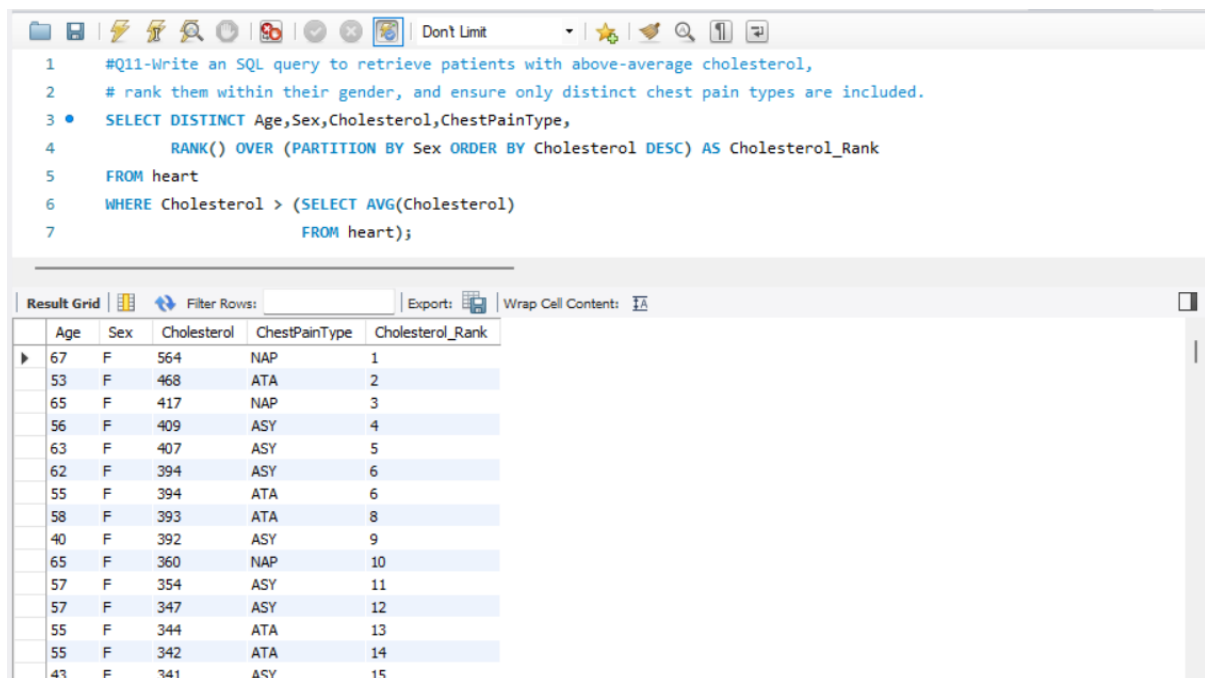
10) **Multi-Condition Query:** List all patients where both cholesterol levels and resting blood pressure are above 150.

1 #Q10-Multi-Condition Query: List all patients where both cholesterol levels and resting blood pressure are above 150.
2 • SELECT *
3 FROM heart
4 WHERE Cholesterol > 150 AND RestingBP > 150;
5

Result Grid Filter Rows: Export: Wrap Cell Content:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
▶	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
	51	F	ATA	160	194	0	Normal	170	N	0	Up	0
	52	M	ASY	160	246	0	ST	82	Y	4	Flat	1
	52	M	ATA	160	196	0	Normal	165	N	0	Up	0
	65	M	ASY	170	263	1	Normal	Normal	Y	2	Flat	1
	48	M	ASY	160	329	0	Normal	92	Y	1.5	Flat	1
	39	M	ATA	190	241	0	Normal	106	N	0	Up	0
	58	F	ATA	180	393	0	Normal	110	Y	1	Flat	1
	56	M	ASY	170	388	0	ST	122	Y	2	Flat	1
	52	M	ASY	160	331	0	Normal	94	Y	2.5	Flat	1
	47	M	ASY	160	291	0	ST	158	Y	3	Flat	1
	56	M	ASY	155	342	1	Normal	150	Y	3	Flat	1
	54	M	ATA	160	195	0	ST	130	N	1	Up	0
	47	M	ATA	160	263	0	Normal	174	N	0	Up	0
	54	F	ATA	160	312	0	Normal	130	N	0	Up	0
	58	M	NAP	160	211	1	ST	92	N	0	Flat	1
	53	M	ASY	180	285	0	ST	120	Y	1.5	Flat	1
	46	M	ASY	180	280	0	ST	120	N	0	Up	0

11) Write an SQL query to retrieve patients with above-average cholesterol, rank them within their gender, and ensure only distinct chest pain types are included.



The screenshot shows a SQL IDE interface. The top pane contains an SQL query. The bottom pane shows the 'Result Grid' with 15 rows of data. The query is as follows:

```
1 #Q11-Write an SQL query to retrieve patients with above-average cholesterol,  
2 # rank them within their gender, and ensure only distinct chest pain types are included.  
3 • SELECT DISTINCT Age,Sex,Cholesterol,ChestPainType,  
4     RANK() OVER (PARTITION BY Sex ORDER BY Cholesterol DESC) AS Cholesterol_Rank  
5 FROM heart  
6 WHERE Cholesterol > (SELECT AVG(Cholesterol)  
7     FROM heart);
```

The 'Result Grid' displays the following data:

	Age	Sex	Cholesterol	ChestPainType	Cholesterol_Rank
▶	67	F	564	NAP	1
	53	F	468	ATA	2
	65	F	417	NAP	3
	56	F	409	ASY	4
	63	F	407	ASY	5
	62	F	394	ASY	6
	55	F	394	ATA	6
	58	F	393	ATA	8
	40	F	392	ASY	9
	65	F	360	NAP	10
	57	F	354	ASY	11
	57	F	347	ASY	12
	55	F	344	ATA	13
	55	F	342	ATA	14
	43	F	341	ASY	15

Executive Summary

Objective

The primary goal of this project is to develop a predictive model for heart failure using Linear Regression, Logistic Regression and k -NN which includes data preprocessing, exploratory data analysis (EDA), model training, and evaluation.

Data Preprocessing

The dataset, stored as heart.csv, is loaded and examined. The preprocessing steps include:

- Checking data types of columns.
- Determining the number of records.
- Summary statistics (describe()) to understand variable distributions.
- Handling missing values (isnull().sum()).

Exploratory Data Analysis (EDA)

Several visualizations and insights are provided, including:

- **Heart Disease by Gender:** A **countplot** is used to analyze the distribution of heart disease across genders.
- Other potential exploratory insights such as bar plot, box plot (for detecting outliers) scatter plot, pie chart are included.

Model Training

The model development process includes:

1. *Feature Engineering:*
 - Data is split into training and testing sets using train_test_split().
 - Standardization and normalization is performed for numerical features.
2. *Model Selection:*
 - Linear Regression, Logistic Regression, k -NN is chosen as the predictive model.
 - The models are trained on the dataset using from scikit-learn.

3. *Performance Metrics:*

- The performance is evaluated using:
 - Accuracy Score
 - Precision Score
 - Recall Score
 - Confusion Matrix
 - Regression Summary (dmba.regressionSummary)
 - Adjusted R² Score

Key Findings & Inferences

- The model attempts to predict heart failure based on various health-related features.
- Insights from **EDA** may help identify key risk factors for heart disease.
- Performance metrics indicate how well the models generalize to unseen data.

Limitations & Future Scope

- **Regression Models and *k*-NN** may not be the best model for classification problems like heart disease prediction. A more suitable alternative could be used i.e **machine learning algorithms (Random Forest, XGBoost, Neural Networks, etc.)**.
- Additional **feature selection techniques** and **hyperparameter tuning** could improve the model's predictive accuracy.
- Further **data augmentation or external datasets** could be used to enhance model robustness.