

# **MEDICAL INSURANCE DATA ANALYSIS REPORT**

## **USING SQL**

### **Introduction**

SQL (Structured Query Language) is a powerful programming language designed for managing and manipulating relational databases. It is widely used for querying, updating, and managing data stored in relational database management systems (RDBMS). SQL provides a robust framework for handling large datasets, performing complex queries, and extracting valuable insights from the data.

### **Key Features of SQL:**

- **Data Retrieval:** SQL allows users to retrieve specific data from large datasets using SELECT statements.
- **Data Manipulation:** Users can insert, update, and delete records in the database.
- **Data Definition:** SQL provides commands to define and modify the structure of database objects.
- **Data Control:** Users can control access to data and manage database transactions.

SQL is widely used in various domains, including finance, healthcare, marketing, and human resources, to analyze data and make informed business decisions.

### **Introduction to MySQL**

MySQL is an open-source relational database management system that uses SQL. It is one of the most popular databases due to its reliability, ease of use, and performance. MySQL is suitable for a wide range of applications, from small to large-scale enterprise applications.

### **Key Features of MySQL:**

- **Scalability and Flexibility:** MySQL can handle large datasets and complex queries.
- **High Performance:** Optimized for high-speed transactions and quick query processing.

- **Open-Source:** MySQL is open-source and has a large community for support and development.
- **Cross-Platform:** Compatible with various operating systems, including Windows, Linux, and macOS.
- **Security:** Offers robust security features to protect data.

MySQL is extensively used for data analysis in various industries, including healthcare, where it helps manage and analyze medical insurance data, patient records, and healthcare metrics.

## Objectives


The objective of this project is to analyze medical insurance data using SQL to derive insights that can help improve decision-making in healthcare management. The specific problem statements addressed in this project are:

1. Select all columns for all patients.
2. Display the average claim amount for patients in each region.
3. Select the maximum and minimum BMI values in the table.
4. Select the PatientID, age, and BMI for patients with a BMI between 40 and 50.
5. Select the number of smokers in each region.
6. Determine the average claim amount for patients who are both diabetic and smokers.
7. Retrieve all patients who have a BMI greater than the average BMI of patients who are smokers.
8. Select the average claim amount for patients in each age group.
9. Retrieve the total claim amount for each patient, along with the average claim amount across all patients.
10. Retrieve the top 3 patients with the highest claim amount, along with their respective claim amounts greater than the average claim amount for their region.
11. Select the details of patients who have a claim amount greater than the average claim amount for their region.
12. Retrieve the rank of each patient based on their claim amount.
13. Select the details of patients along with their claim amount, and their rank based on claim amount within their region.

## PROBLEM STATEMENTS AND SOLUTIONS

### 1. Select all columns for all patients.

sql


 Copy code

```
SELECT * FROM insurance_data;
```

- **Output:** This query retrieves all columns and all records from the insurance\_data table, providing a complete view of the dataset.

### 2. Display the average claim amount for patients in each region.

sql

 Copy code


```
SELECT region, AVG(claim) AS avg_claim  
FROM insurance_data  
GROUP BY region;
```

- **Output:** This query calculates the average claim amount for patients in each region. It helps in understanding the claim distribution across different regions.

	A	B
1	region	avg_claim
2	southeast	13058.52266
3	northwest	11612.72142
4	southwest	12686.72721
5	northeast	16889.04472
6		

### 3. Select the maximum and minimum BMI values in the table.

sql

 Copy code


```
SELECT MAX(bmi) AS max_bmi, MIN(bmi) AS min_bmi
FROM insurance_data;
```

- **Output:** This query retrieves the maximum and minimum BMI values from the insurance\_data table, providing insights into the range of BMI values in the dataset.

	A	B
1	max_bmi	min_bmi
2	53.1	16

### 4. Select the PatientID, age, and BMI for patients with a BMI between 40 and 50.

sql

 Copy code

```
SELECT PatientID, age, bmi
FROM insurance_data
WHERE bmi BETWEEN 40 AND 50;
```

- **Output:** This query selects the PatientID, age, and BMI for patients whose BMI is between 40 and 50. It helps identify patients within this specific BMI range.

	A	B	C				
1	PatientID	age	bmi	8	92	20	40.5
2	8	19	41.1	9	124	55	40.2
3	9	20	43	10	125	33	40.3
4	26	23	43	11	191	28	46.5
5	41	29	40.3	12	208	48	40.2
6	69	26	40.5	13	213	46	42.1
7	76	30	41.9	14	228	34	42.4
				15	247	20	45.9

## 5. Select the number of smokers in each region.

```

sql
Copy code

SELECT region, COUNT(PatientID) AS smoker_count
FROM insurance_data
WHERE smoker = 'Yes'
GROUP BY region;

```

- **Output:** This query counts the number of smokers in each region, grouped by region. It helps understand the distribution of smokers across different regions.

	A	B
1	region	count(PatientID)
2	northeast	67
3	southwest	58
4	northwest	58
5	southeast	91

**6. Determine the average claim amount for patients who are both diabetic and smokers.**

```
sql Copy code  
  
SELECT AVG(claim) AS avg_claim  
FROM insurance_data  
WHERE diabetic = 'Yes' AND smoker = 'Yes';
```

- **Output:** This query calculates the average claim amount for patients who are both diabetic and smokers, providing insights into the claim behavior of this specific group.

	A
1	avg_claim
2	31277.5501

**7. Retrieve all patients who have a BMI greater than the average BMI of patients who are smokers.**


```
sql Copy code  
  
SELECT AVG(bmi) AS avg_bmi  
FROM insurance_data  
WHERE smoker = 'Yes'; -- Average BMI of smokers: 30.71  
  
SELECT *  
FROM insurance_data  
WHERE bmi > (SELECT AVG(bmi) FROM insurance_data WHERE smoker = 'Yes');
```

- **Output:** The first query calculates the average BMI of patients who are smokers, and the second query retrieves all patients whose BMI is greater than this average value, helping identify patients with relatively higher BMI among smokers.

	A	B	C	D	E	F	G	H	I	J	K
1	index	PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim
2	1192	1193	47	male	31.7	129	Yes	0	Yes	northeast	33732.69
3	1193	1194	46	male	31.9	82	Yes	0	Yes	northwest	33750.29
4	1196	1197	36	female	31.4	136	Yes	0	Yes	southwest	34166.27
5	1197	1198	30	male	31.1	136	No	0	Yes	northeast	34254.05
6	1198	1199	22	male	31.7	115	No	2	Yes	southeast	34303.17
7	1199	1200	30	female	33.1	93	Yes	0	Yes	southeast	34439.86
8	1200	1201	40	male	32.7	98	No	0	Yes	southwest	34472.84
9	1201	1202	45	male	33.5	81	No	0	Yes	northeast	34617.84
10	1202	1203	32	male	31.7	125	No	0	Yes	southeast	34672.15
11	1203	1204	26	male	34.8	94	Yes	0	Yes	southwest	34779.62
12	1204	1205	32	male	31.1	114	Yes	1	Yes	southeast	34806.47
13	1205	1206	20	male	34.9	124	Yes	0	Yes	southwest	34828.65
14	1206	1207	46	female	31.4	111	No	0	Yes	southwest	34838.87
15	1210	1211	25	male	30.8	140	Yes	0	Yes	southwest	35491.64
16	1211	1212	22	male	35.6	97	Yes	0	Yes	southwest	35585.58
17	1212	1213	56	female	31	129	Yes	3	Yes	southeast	35595.59
18	1213	1214	30	female	32.8	98	No	2	Yes	southeast	36021.01
19	1214	1215	40	male	32.9	87	Yes	2	Yes	southwest	36085.22
20	1215	1216	22	male	33.3	117	No	2	Yes	southeast	36124.57
21	1216	1217	34	female	36.9	131	No	0	Yes	southeast	36149.48

## 8. Select the average claim amount for patients in each age group.

sql

 Copy code


```
SELECT
  CASE
    WHEN age < 18 THEN 'Under 18'
    WHEN age BETWEEN 18 AND 30 THEN '18-30'
    WHEN age BETWEEN 31 AND 50 THEN '31-50'
    ELSE 'Over 50'
  END AS age_group,
  ROUND(AVG(claim), 2) AS avg_claim
FROM insurance_data
GROUP BY age_group;
```

- **Output:** This query calculates the average claim amount for patients in different age groups, providing insights into claim behavior across age demographics.

	A	B
1	age_group	avg_claim
2	31-50	12981.09
3	18-30	14004.48
4	Over 50	12744.96

**9. Retrieve the total claim amount for each patient, along with the average claim amount across all patients.**

sql

 Copy code

```
SELECT *,
    SUM(claim) OVER (PARTITION BY PatientID) AS total_claim,
    AVG(claim) OVER () AS avg_claim
FROM insurance_data;
```


- **Output:** This query retrieves the total claim amount for each patient and the average claim amount across all patients, giving a comprehensive view of individual and overall claim amounts.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	index	PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim	total_claim	avg_claim
2	0	1	39	male	23.2	91	Yes	0	No	southeast	1121.87	1121.87	13252.74564
3	1	2	24	male	30.1	87	No	0	No	southeast	1131.51	1131.51	13252.74564
4	2	3	23	male	33.3	82	Yes	0	No	southeast	1135.94	1135.94	13252.74564
5	3	4	34	male	33.7	80	No	0	No	northwest	1136.4	1136.4	13252.74564
6	4	5	30	male	34.1	100	No	0	No	northwest	1137.01	1137.01	13252.74564
7	5	6	29	male	34.4	96	Yes	0	No	northwest	1137.47	1137.47	13252.74564
8	6	7	25	male	37.3	86	Yes	0	No	northwest	1141.45	1141.45	13252.74564
9	7	8	19	male	41.1	100	No	0	No	northwest	1146.8	1146.8	13252.74564
10	8	9	20	male	43	86	No	0	No	northwest	1149.4	1149.4	13252.74564
11	9	10	30	male	53.1	97	No	0	No	northwest	1163.46	1163.46	13252.74564
12	10	11	36	male	19.8	88	Yes	0	No	northwest	1241.57	1241.57	13252.74564
13	11	12	37	male	20.3	90	Yes	0	No	northwest	1242.26	1242.26	13252.74564
14	12	13	19	male	20.7	81	No	0	No	northwest	1242.82	1242.82	13252.74564
15	13	14	32	male	27.6	100	No	0	No	northwest	1252.41	1252.41	13252.74564
16	14	15	40	male	28.7	81	Yes	0	No	northwest	1253.94	1253.94	13252.74564
17	15	16	32	male	30.4	86	Yes	0	No	southwest	1256.3	1256.3	13252.74564
18	16	17	35	male	34.1	90	No	0	No	southwest	1261.44	1261.44	13252.74564
19	17	18	41	male	34.4	84	No	0	No	southwest	1261.86	1261.86	13252.74564
20	18	19	49	male	35.4	97	Yes	0	No	southwest	1263.25	1263.25	13252.74564



**10. Retrieve the top 3 patients with the highest claim amount, along with their respective claim amounts and the total claim amount for all patients.**

sql

 Copy code


```
SELECT PatientID, claim, SUM(claim) OVER () AS total_claim
FROM insurance_data
ORDER BY claim DESC
LIMIT 3;
```

- **Output:** This query retrieves the top 3 patients with the highest claim amounts, along with their respective claim amounts and the total claim amount for all patients.

	A	B	C
1	PatientID	claim	total_claim
2	1340	63770.4	17758679.2
3	1339	62592.9	17758679.2
4	1338	60021.4	17758679.2

**11. Select the details of patients who have a claim amount greater than the average claim amount for their region.**

sql

 Copy code


```
SELECT *
FROM (SELECT *, AVG(claim) OVER (PARTITION BY region) AS avg_claim
      FROM insurance_data) AS subquery
WHERE claim > avg_claim;
```

- **Output:** This query selects the details of patients whose claim amounts are greater than the average claim amount for their region, providing insights into patients with higher-than-average claims within their regions.

	A	B	C	D	E	F	G	H	I	J	K	L
1	index	PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim	avg_claim
2	1012	1013	35	male	33.6	90	Yes	4	No	northeast	17128.43	16889.04472
3	1013	1014	45	male	27.4	83	No	1	Yes	northeast	17178.68	16889.04472
4	1023	1024	51	female	41.3	98	No	0	No	northeast	17878.9	16889.04472
5	1024	1025	22	male	24.8	100	No	0	Yes	northeast	17904.53	16889.04472
6	1036	1037	38	male	29.8	87	Yes	0	Yes	northeast	18648.42	16889.04472
7	1042	1043	26	female	32.4	81	No	1	No	northeast	18903.49	16889.04472
8	1049	1050	27	female	35.7	96	Yes	2	No	northeast	19144.58	16889.04472
9	1056	1057	21	male	39.7	107	Yes	4	No	northeast	19496.72	16889.04472
10	1060	1061	53	female	20.2	87	No	1	Yes	northeast	19594.81	16889.04472
11	1064	1065	28	female	20	96	Yes	2	Yes	northeast	19798.05	16889.04472
12	1066	1067	43	female	23.4	82	Yes	0	Yes	northeast	19964.75	16889.04472
13	1068	1069	19	male	26.4	87	Yes	0	Yes	northeast	20149.32	16889.04472
14	1072	1073	20	male	29.6	91	Yes	1	No	northeast	20277.81	16889.04472
15	1073	1074	46	female	25.6	100	Yes	1	Yes	northeast	20296.86	16889.04472
16	1077	1078	19	male	40.3	110	No	0	No	northeast	20709.02	16889.04472
17	1079	1080	49	male	28	106	No	1	Yes	northeast	20773.63	16889.04472
18	1081	1082	42	female	33.3	107	No	0	No	northeast	20878.78	16889.04472
19	1082	1083	30	male	27.7	89	Yes	2	Yes	northeast	20984.09	16889.04472
20	1088	1089	33	male	24.6	87	Yes	2	Yes	northeast	21259.38	16889.04472
21	1089	1090	45	female	30.1	83	Yes	0	No	northeast	21344.85	16889.04472
22	1093	1094	59	female	24.9	108	No	3	Yes	northeast	21659.93	16889.04472
23	1095	1096	38	female	25.3	104	Yes	1	Yes	northeast	21771.34	16889.04472
24	1104	1105	60	female	28.1	84	No	1	Yes	northeast	22331.57	16889.04472
25	1105	1106	45	female	24.2	103	No	2	No	northeast	22395.74	16889.04472

## 12. Retrieve the rank of each patient based on their claim amount.

sql

 Copy code


```
SELECT *, RANK() OVER (ORDER BY claim DESC) AS claim_rank
FROM insurance_data;
```

- **Output:** This query retrieves the rank of each patient based on their claim amount, helping identify the relative claim positions of patients.

	A	B	C	D	E	F	G	H	I	J	K	L
1	index	PatientID	age	gender	bmi	bloodpress	diabetic	children	smoker	region	claim	rank() over
2	1339	1340	30	female	47.4	101	No	0	Yes	southeast	63770.43	1
3	1338	1339	37	male	30.4	106	No	0	Yes	southeast	62592.87	2
4	1337	1338	30	male	34.5	91	Yes	3	Yes	northwest	60021.4	3
5	1336	1337	59	female	38.1	120	No	1	Yes	northeast	58571.07	4
6	1335	1336	44	female	35.5	88	Yes	0	Yes	northwest	55135.4	5
7	1334	1335	43	male	32.8	125	No	0	Yes	southwest	52590.83	6
8	1333	1334	44	male	36.4	127	No	1	Yes	southwest	51194.56	7
9	1332	1333	26	male	37	120	No	2	Yes	southeast	49577.66	8
10	1331	1332	18	male	41.1	104	No	1	Yes	southeast	48970.25	9
11	1330	1331	25	female	38.1	111	No	0	Yes	southeast	48885.14	10
12	1329	1330	52	female	37.7	109	Yes	0	Yes	southwest	48824.45	11
13	1328	1329	45	male	42.1	117	No	1	Yes	southeast	48675.52	12
14	1327	1328	49	male	40.9	107	No	0	Yes	southeast	48673.56	13
15	1326	1327	26	male	40.6	113	Yes	3	Yes	northeast	48549.18	14
16	1325	1326	52	female	36.4	133	Yes	1	Yes	northeast	48517.56	15
17	1324	1325	39	male	39.9	115	No	0	Yes	southwest	48173.36	16
18	1323	1324	49	female	33.8	107	No	1	Yes	southwest	47928.03	17
19	1322	1323	33	female	36.8	117	Yes	1	Yes	northeast	47896.79	18
20	1321	1322	26	male	37	81	No	2	Yes	northwest	47496.49	19
21	1320	1321	19	male	42.9	104	Yes	2	Yes	southeast	47462.89	20
22	1319	1320	40	male	36.3	94	Yes	1	Yes	southwest	47403.88	21
23	1318	1319	27	female	32.2	115	Yes	2	Yes	southwest	47305.31	22
24	1317	1318	49	female	31.3	111	No	2	Yes	southwest	47291.06	23
25	1316	1317	41	male	41.8	109	Yes	2	Yes	southeast	47269.85	24

**13. Select the details of patients along with their claim amount, and their rank based on claim amount within their region.**

sql

 Copy code

```
SELECT *, RANK() OVER (PARTITION BY region ORDER BY claim DESC) AS region_claim_rank
FROM insurance_data;
```

- **Output:** This query selects the details of patients along with their claim amount and their rank based on claim amount within their region, providing a region-specific ranking of claim amounts.

	A	B	C	D	E	F	G	H	I	J	K	L
1	index	PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim	rank() over(partition by region order by claim desc)
2	1336	1337	59	female	38.1	120	No	1	Yes	northeast	58571.07	1
3	1326	1327	26	male	40.6	113	Yes	3	Yes	northeast	48549.18	2
4	1325	1326	52	female	36.4	133	Yes	1	Yes	northeast	48517.56	3
5	1322	1323	33	female	36.8	117	Yes	1	Yes	northeast	47896.79	4
6	1310	1311	26	female	37.1	95	No	3	Yes	northeast	46255.11	5
7	1304	1305	42	male	32	83	Yes	0	Yes	northeast	45710.21	6
8	1301	1302	60	female	35	92	Yes	2	Yes	northeast	44641.2	7
9	1290	1291	47	male	41.9	140	Yes	3	Yes	northeast	43753.34	8
10	1288	1289	22	male	34.1	108	No	0	Yes	northeast	43254.42	9
11	1282	1283	42	male	30.7	117	Yes	0	Yes	northeast	42303.69	10
12	1278	1279	35	female	35.5	135	No	0	Yes	northeast	42111.66	11
13	1275	1276	37	male	32.3	115	No	1	Yes	northeast	41919.1	12
14	1272	1273	31	male	31.8	95	No	0	Yes	northeast	41097.16	13
15	1271	1272	35	male	35	109	No	1	Yes	northeast	41034.22	14
16	1267	1268	43	female	42.8	105	Yes	1	Yes	northeast	40904.2	15
17	1257	1258	43	male	36.7	139	Yes	1	Yes	northeast	39774.28	16
18	1252	1253	37	male	30.8	125	Yes	3	Yes	northeast	39597.41	17
19	1251	1252	21	male	31.4	89	Yes	1	Yes	northeast	39556.49	18
20	1249	1250	20	male	32.8	99	Yes	1	Yes	northeast	39125.33	19
21	1248	1249	29	male	34.2	101	Yes	1	Yes	northeast	39047.29	20
22	1242	1243	27	female	36.7	140	No	2	Yes	northeast	38511.63	21
23	1234	1235	47	female	31.9	89	Yes	1	Yes	northeast	37701.88	22
24	1233	1234	20	male	33.6	103	No	1	Yes	northeast	37607.53	23
25	1230	1231	58	female	30.8	139	No	0	Yes	northeast	37270.15	24

## CONCLUSION

The Medical Insurance Data Analysis project using SQL provides valuable insights into various aspects of medical insurance claims. By analyzing key metrics such as average claim amounts, BMI values, and the distribution of smokers across regions, the project helps in understanding the factors influencing medical claims. The detailed SQL queries enable a comprehensive analysis of the dataset, offering actionable insights for healthcare management and decision-making.