nextleap

SQL 2 Data Analyst Fellowship

Saksham Arora

Software Engineer Microsoft

Saksham Arora - Microsoft | LinkedIn

1. Exploring Data with SQL

- Exploratory Data Analysis using SQL
- Functions (MIN, MAX, AVG, SUM, STDDEV, TRUNCATE, ROUND)

2. Summarising Data (Session 6)

- Correlation function

3. Working with strings/ data cleaning

- LOWER, UPPER
- LIKE, _, % wildcards
- TRIM, BTRIM, RTRIM, LTRIM
- SUBSTRING, LEFT, RIGHT
- SPLIT_PART, CONCAT

4. Working with Dates and Timestamps

- Date, Timestamp, Intervals
- Date Comparisons
- Date Arithmetic
- Extracting Fields
- Date Functions
- Truncating Times
- Aggregation with Date/Time Series

Exploring Data Functions

STDDEV

1. Standard deviation is the square root of the average of squared deviations of the items from their mean.
2. SELECT STDDEV(SALARY) From employees;

TRUNCATE

1. The TRUNCATE TABLE command deletes the data inside a table, but not the table itself.
2. TRUNCATE TABLE tableName

ROUND(number, decimals)

1. SELECT ROUND(amount, 3) FROM payment
2. SELECT ROUND(235.415, 1) AS RoundValue;

Summarising Data

Youtube videos for Summarising Data

Correlation - What Is Correlation? | Types of Correlation | Correlation Coefficient | Statistics | Simplilearn

Median & Percentile -
Median, Mean, Mode, Percentile | Math, Statistics for data science, machine learning

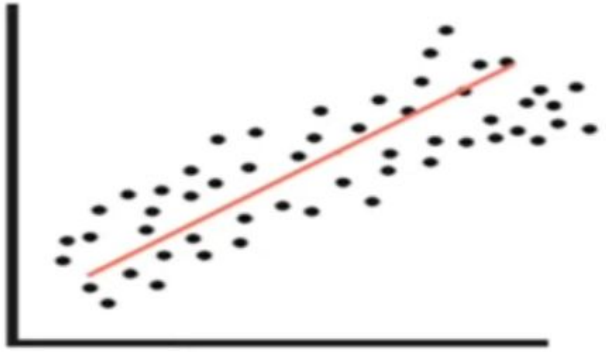
Percentiles And Quartiles In Statistics | Percentiles And Quartiles Explained | Simplilearn

Correlation is a statistical technique used to determine the degree to which two variables are related

For example, the price and demand of a product is in correlation

Correlation lies between -1 to +1

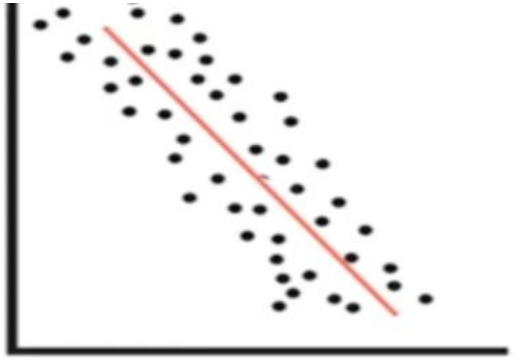
It is a relationship between two variables where if one variable increases, the other one also increases and vice-versa



Positive Correlation

As the number of trees cut down increase, the probability of erosion increases

Negative correlation means there is an inverse relationship between two variables. When one variable decreases, other increases.



Negative Correlation

If a car decreases speed, the time taken to reach the destination increases

SELECT corr(staff_id, amount) FROM payment

Working with Strings

LOWER

1. Convert the text to lower-case
2. SELECT LOWER(columnName) FROM tableName

UPPER

1. Convert the text to upper-case
2. SELECT UPPER(columnName) FROM tableName

TRIM

1. **TRIM** - the TRIM function is used to remove specific characters (or whitespace by default) from the beginning, end, or both sides of a string
2. SELECT TRIM(‘character’ from columnName) FROM tableName
3. SELECT TRIM(LEADING 'A' from district) FROM address
4. SELECT TRIM(TRAILING 'A' from district) FROM address
5. SELECT TRIM('A' from district) FROM address

BTRIM, LTRIM, RTRIM

1. BTRIM - Removes both leading and trailing characters from a string
2. LTRIM - Removes leading characters from a string
3. RTRIM - Removes trailing characters from a string
4. SELECT BTRIM('---Hello World---', '-'); -- Output: 'Hello World'
5. If any character is not provided it will trim spaces.

SUBSTRING

1. SUBSTRING extracts a specific part of a string based on position and length.
2. SELECT SUBSTRING('Hello World' FROM 7);
-- Output: 'World' (starts from position 7)
3. SELECT SUBSTRING('Hello World' FROM 7 FOR 3);
-- Output: 'Wor' (starts from position 7, and extracts 3 characters)

LEFT, RIGHT

1. LEFT extracts a specific number of characters from the left (beginning) of the string.

2. SELECT LEFT('Hello World', 5); —> Output: 'Hello'

3. RIGHT extracts a specific number of characters from the right (end) of the string

4. SELECT RIGHT('Hello World', 5); —> Output: 'World'

SPLIT_PART

SPLIT_PART is used to split a string into parts based on a delimiter and then extract a specific part. SPLIT_PART(string, delimiter, field)

string: The string to be split.

delimiter: The character or string used to split the input string.

field: The position (1-based index) of the part to extract after splitting

```
SELECT SPLIT_PART('2023-09-14', '-', 3);
-- Output: '14' (extracts the third part)

SELECT SPLIT_PART('apple,banana,orange', ',', 2);
-- Output: 'banana'
```

CONCAT is used to concatenate (join) multiple strings into one.

```
SELECT CONCAT('I', ' have ', '5', ' apples.');
-- Output: 'I have 5 apples.'

SELECT CONCAT('PostgreSQL', NULL, ' is great');
-- Output: 'PostgreSQL is great' (NULL is ignored)
```

Working with Date and TimeStamps

- **DATE**: Stores dates in the format 'YYYY-MM-DD'.
- **TIME**: Stores times in the format 'HH:MM:SS'.
- In PostgreSQL, there is no explicit DATETIME data type. Instead, PostgreSQL uses the TIMESTAMP type, which is equivalent to DATETIME in other databases (like MySQL).
- **DATE()**: Extracts the date part from a DATETIME expression.
- For **TIMESTAMP**:

  TIMESTAMP 'YYYY-MM-DD HH:MI:SS'
- For **TIMESTAMPTZ** (Timestamp with time zone):

  TIMESTAMP WITH TIME ZONE 'YYYY-MM-DD HH:MI:SS+TZ'

CAST

CAST is a PostgreSQL function that defines how to convert data types. In cases where data needs to be converted from one type to another, casting allows you to achieve this seamlessly.
SELECT CAST(payment_date as TIME) FROM payment

EXTRACT

The EXTRACT() function extracts a field from a date/time value.

SELECT EXTRACT(minutes from order_purchase_timestamp) from orders
SELECT EXTRACT(seconds from order_purchase_timestamp) from orders
SELECT EXTRACT(hours from order_purchase_timestamp) from orders

5. Date Functions

- **NOW()**: Returns the current date and time with timezone.
- **CURRENT_DATE**: Returns the current date.
- **AGE**: Calculates the interval between two dates.

sql

Copy code

```sql
-- Current date and time
SELECT NOW();

-- Only current date
SELECT CURRENT_DATE;

-- Age between two dates
SELECT AGE(TIMESTAMP '2024-11-08', TIMESTAMP '2000-01-01');  -- Returns interval i
```

6. Truncating Times

Use `DATE_TRUNC` to truncate a timestamp to a specific precision (e.g., to the nearest day, hour).

sql Copy code

```sql
-- Truncate to the beginning of the day
SELECT DATE_TRUNC('day', TIMESTAMP '2024-11-08 12:30:45');  -- Returns 2024-11-08 (

-- Truncate to the beginning of the month
SELECT DATE_TRUNC('month', TIMESTAMP '2024-11-08 12:30:45');  -- Returns 2024-11-01
```

PostgreSQL can perform aggregations based on time intervals, which is useful for time-series data.

sql Copy code

```sql
-- Example data table
CREATE TEMP TABLE sales (sale_date DATE, amount NUMERIC);
INSERT INTO sales (sale_date, amount) VALUES
    ('2024-11-01', 100),
    ('2024-11-02', 150),
    ('2024-11-03', 120),
    ('2024-11-08', 130);

-- Aggregate by month
SELECT DATE_TRUNC('month', sale_date) AS month, SUM(amount) AS total_sales
FROM sales
GROUP BY DATE_TRUNC('month', sale_date);

-- Aggregate by week
SELECT DATE_TRUNC('week', sale_date) AS week, SUM(amount) AS total_sales
FROM sales
GROUP BY DATE_TRUNC('week', sale_date);
```

Q&A

What’s on your mind?


SKELLYSCOO
I have several questions.

Feed us back!