Explore data using SQL

A two-day workshop at SNEHA, Mumbai

CivicDataLab

14th - 15th November, 2022

View the slides at https://bit.ly/sql-sneha

Our Journey (Workshop Overview)

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**
2. **What makes a dataset more accesible**

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**
2. **What makes a dataset more accesible**
3. **Analysing data in Excel**

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**
2. **What makes a dataset more accesible**
3. **Analysing data in Excel**
4. **Database Tools**

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**
2. **What makes a dataset more accesible**
3. **Analysing data in Excel**
4. **Database Tools**
5. **Reading and Writing Structured Query Language (SQL)**

Our Journey (Workshop Overview)

Over the next two days, we will learn about:

1. **A dataset**
2. **What makes a dataset more accesible**
3. **Analysing data in Excel**
4. **Database Tools**
5. **Reading and Writing Structured Query Language (SQL)**
6. **Analysing data using SQL**

Learning Objectives

Learning Objectives

A good session will be if by the end you:

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features
3. Are aware about the **ways in which each data point can be stored in a file**

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features
3. Are aware about the **ways in which each data point can be stored in a file**
4. Can evaluate the **data quality** of any data set

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features
3. Are aware about the **ways in which each data point can be stored in a file**
4. Can evaluate the **data quality** of any data set
5. Have a basic understanding about **databases**

Learning Objectives

A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features
3. Are aware about the **ways in which each data point can be stored in a file**
4. Can evaluate the **data quality** of any data set
5. Have a basic understanding about **databases**
6. Can **read and write basic SQL queries**

Learning Objectives

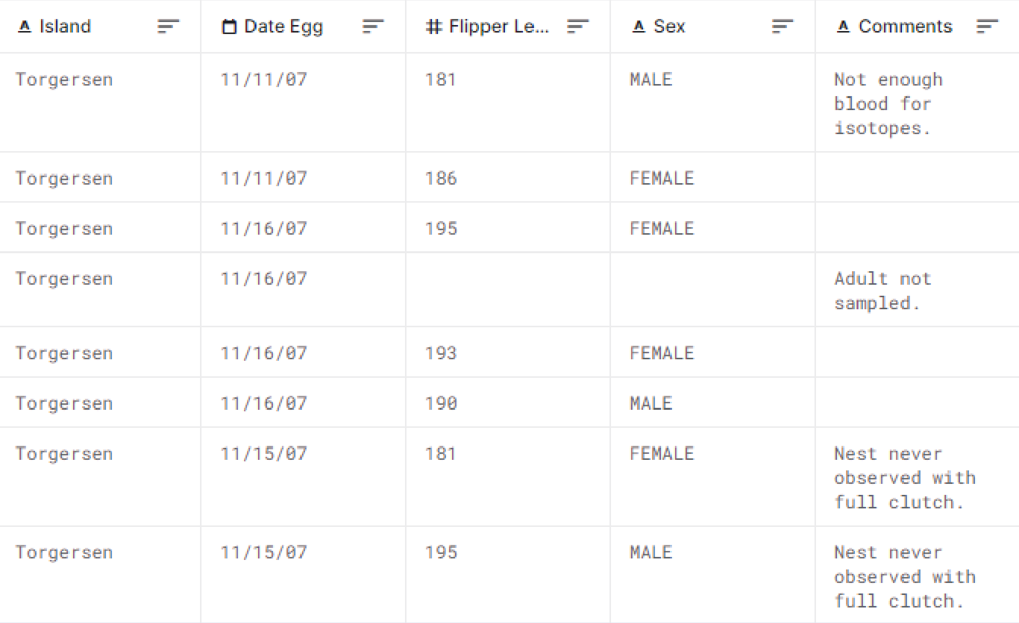
A good session will be if by the end you:

1. Are aware about the **basic structure of a dataset**
2. Can describe any **tabular dataset** in terms of its features
3. Are aware about the **ways in which each data point can be stored in a file**
4. Can evaluate the **data quality** of any data set
5. Have a basic understanding about **databases**
6. Can **read and write basic SQL queries**
7. Have a **pathway to develop your skills**

A basic dataset

palmerpenguins

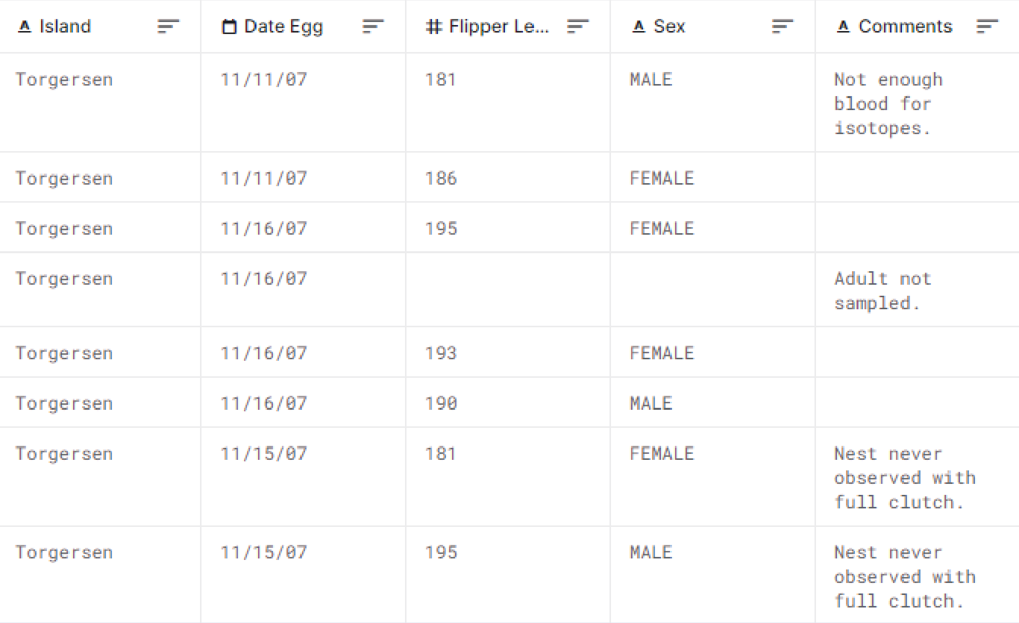
Features of dataset



Island	Date Egg	# Flipper Le...	Sex	Comments
Torgersen	11/11/07	181	MALE	Not enough blood for isotopes.
Torgersen	11/11/07	186	FEMALE	
Torgersen	11/16/07	195	FEMALE	
Torgersen	11/16/07			Adult not sampled.
Torgersen	11/16/07	193	FEMALE	
Torgersen	11/16/07	190	MALE	
Torgersen	11/15/07	181	FEMALE	Nest never observed with full clutch.
Torgersen	11/15/07	195	MALE	Nest never observed with full clutch.

A tabular dataset

Features of dataset



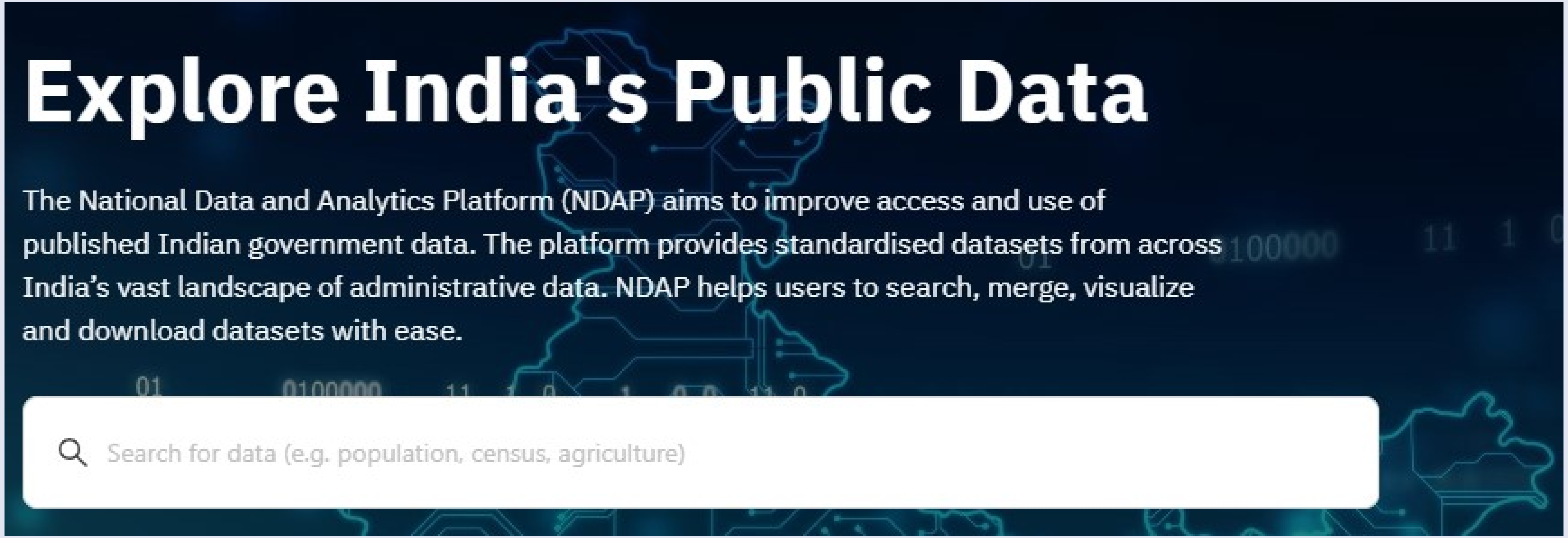
Island	Date Egg	Flipper Le...	Sex	Comments
Torgersen	11/11/07	181	MALE	Not enough blood for isotopes.
Torgersen	11/11/07	186	FEMALE	
Torgersen	11/16/07	195	FEMALE	
Torgersen	11/16/07			Adult not sampled.
Torgersen	11/16/07	193	FEMALE	
Torgersen	11/16/07	190	MALE	
Torgersen	11/15/07	181	FEMALE	Nest never observed with full clutch.
Torgersen	11/15/07	195	MALE	Nest never observed with full clutch.

A tabular dataset

Features of a dataset:

1. Total Rows
2. Total Columns
3. Variables
4. Type of variables (Data Types)
    1. Categorical
    2. Numeric
    3. Text
    4. Date

Quiz - Identify the features of a dataset



Explore India's Public Data

The National Data and Analytics Platform (NDAP) aims to improve access and use of published Indian government data. The platform provides standardised datasets from across India's vast landscape of administrative data. NDAP helps users to search, merge, visualize and download datasets with ease.

National Data and Analytics Platform (or NDAP)

Dataset: **Statewise Reproductive Child Health (RCH) Report Indicator Related to Maternal Health Antenatal Care (ANC)**

Evaluating a dataset (Dataset Quality)

How to create a **good quality** dataset

1. Be consistent.

Evaluating a dataset (Dataset Quality)

How to create a **good quality** dataset

1. Be consistent.
2. Formatting dates.

Evaluating a dataset (Dataset Quality)

How to create a **good quality** dataset

1. Be consistent.
2. Formatting dates.
3. Fill in all of the cells.

Evaluating a dataset (Dataset Quality)

How to create a **good quality** dataset

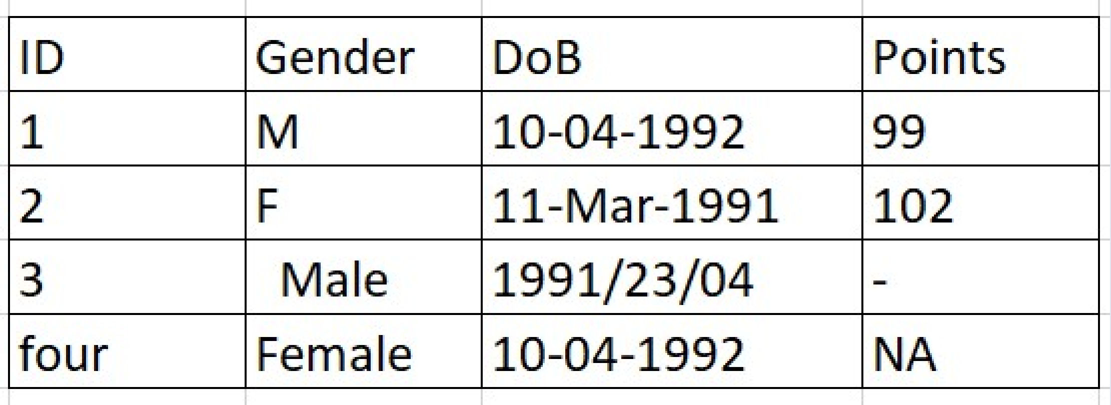
1. Be consistent.
2. Formatting dates.
3. Fill in all of the cells.
4. Don’t use font color or highlighting as data.

Evaluating a dataset (Dataset Quality)

How to create a **good quality** dataset

1. Be consistent.
2. Formatting dates.
3. Fill in all of the cells.
4. Don’t use font color or highlighting as data.
5. Choose good names for things.

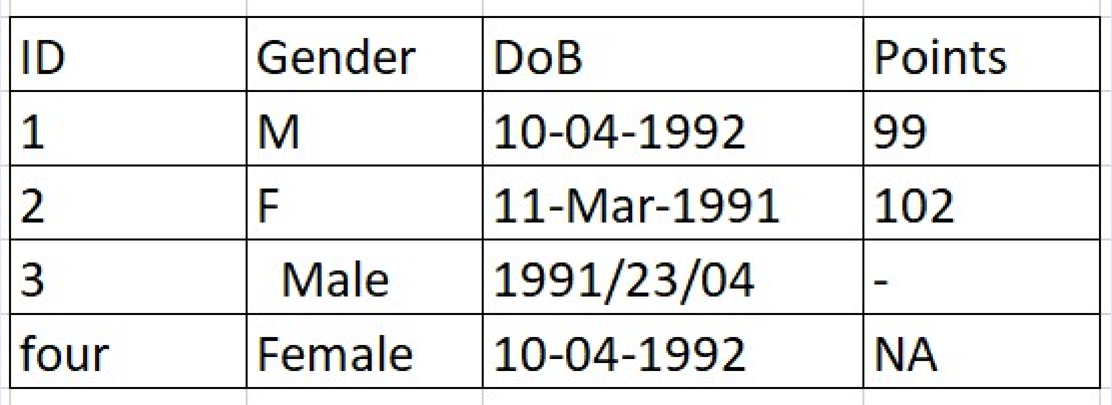
Be Consistent



ID	Gender	DoB	Points
1	M	10-04-1992	99
2	F	11-Mar-1991	102
3	Male	1991/23/04	-
four	Female	10-04-1992	NA

Sample Table

Be Consistent



ID	Gender	DoB	Points
1	M	10-04-1992	99
2	F	11-Mar-1991	102
3	Male	1991/23/04	-
four	Female	10-04-1992	NA

Sample Table

Do you see any issues with this table ?

Be Consistent - Principles

Be Consistent - Principles

1. Consistent codes for categorical variables

Be Consistent - Principles

1. Consistent codes for categorical variables
2. Single fixed code for any missing values

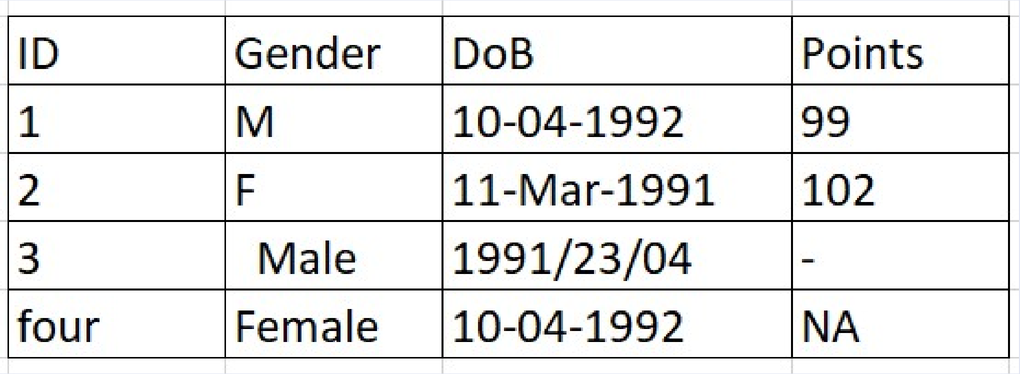
Be Consistent - Principles

1. Consistent codes for categorical variables
2. Single fixed code for any missing values
3. Single common format for all dates

Be Consistent - Principles

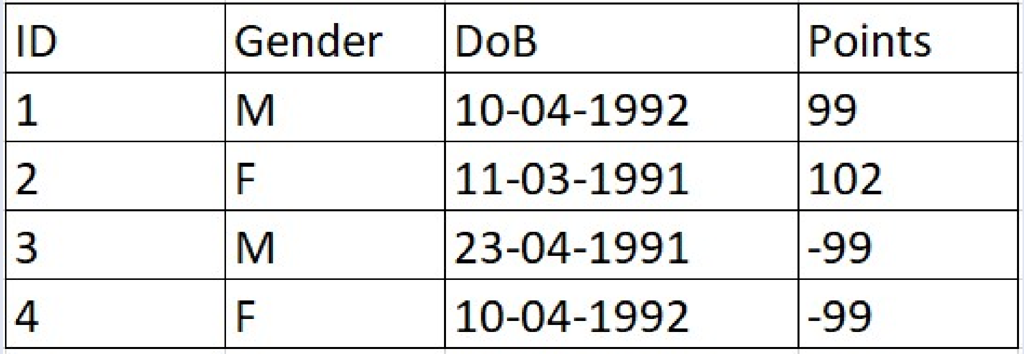
1. Consistent codes for categorical variables
2. Single fixed code for any missing values
3. Single common format for all dates
4. Extra spaces within cells

Be Consistent - The difference



ID	Gender	DoB	Points
1	M	10-04-1992	99
2	F	11-Mar-1991	102
3	Male	1991/23/04	-
four	Female	10-04-1992	NA

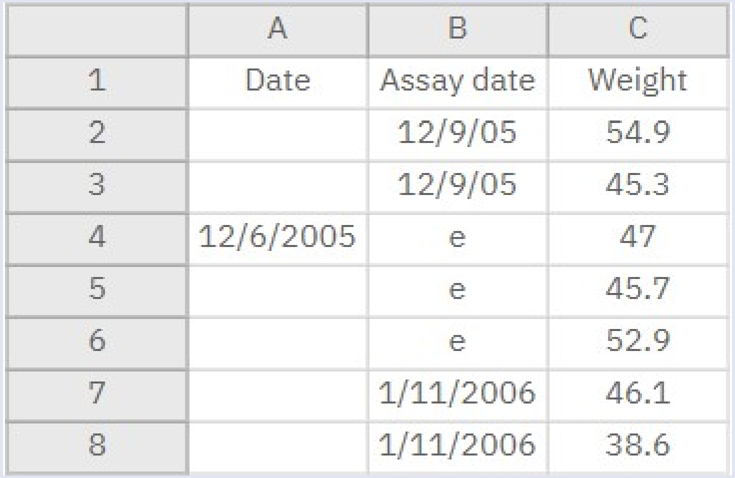
Sample Table



ID	Gender	DoB	Points
1	M	10-04-1992	99
2	F	11-03-1991	102
3	M	23-04-1991	-99
4	F	10-04-1992	-99

Formatted Table

Formatting dates



	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

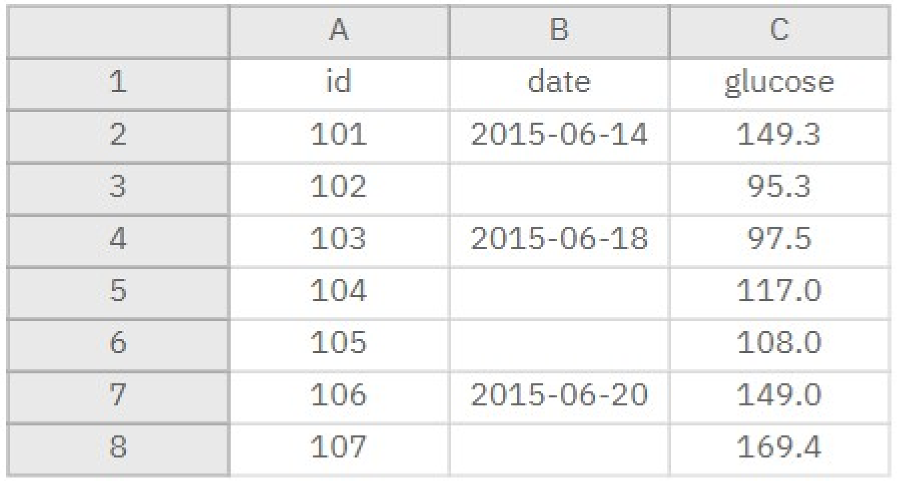
Be consistent in the way in which you write dates. And always use the YYYY-MM-DD format (or put the year, month, and day in separate columns). [1]

[1] Dates as Data

No empty cells

Missing Values

Better alternative

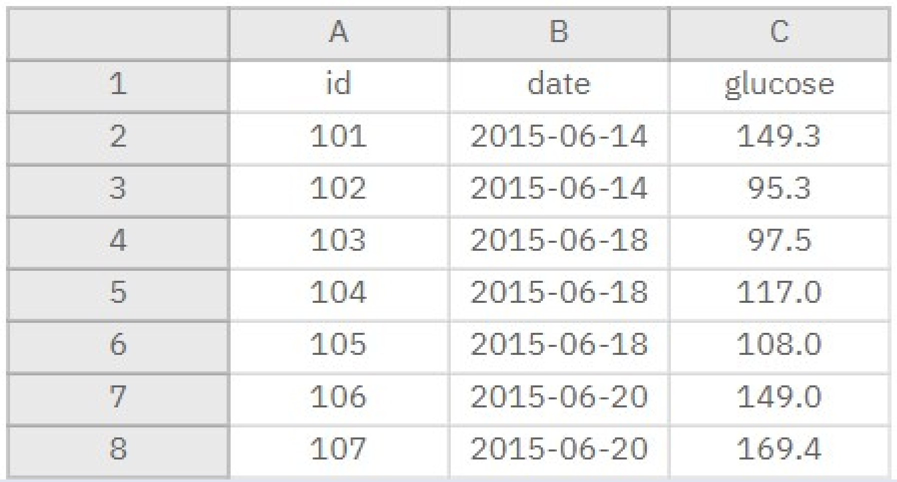


	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

No empty cells

Missing Values

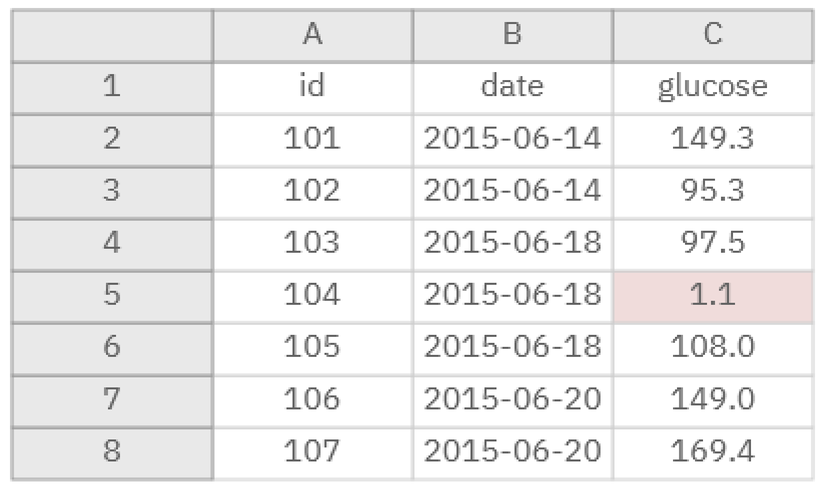
Better alternative



	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	117.0
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4

Formatting data within files

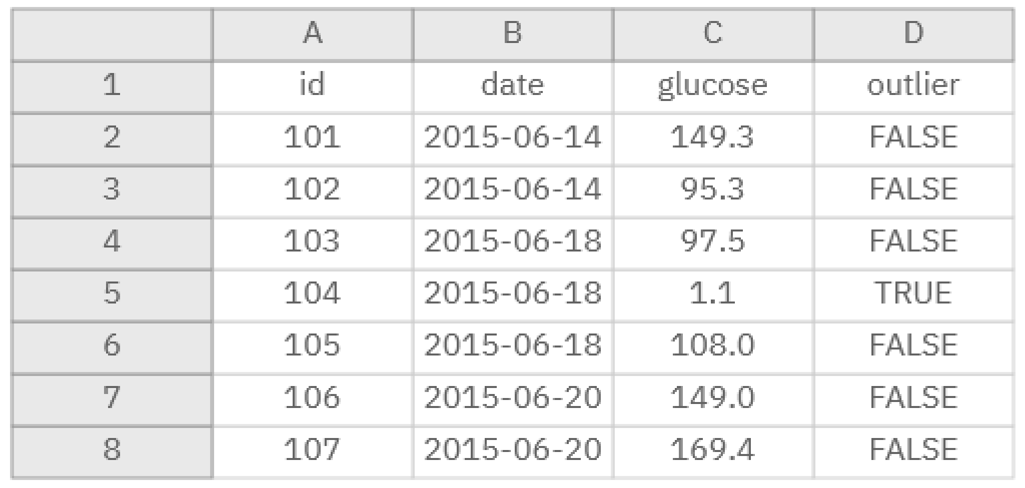
Formatted Table   Better alternative



	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	1.1
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4

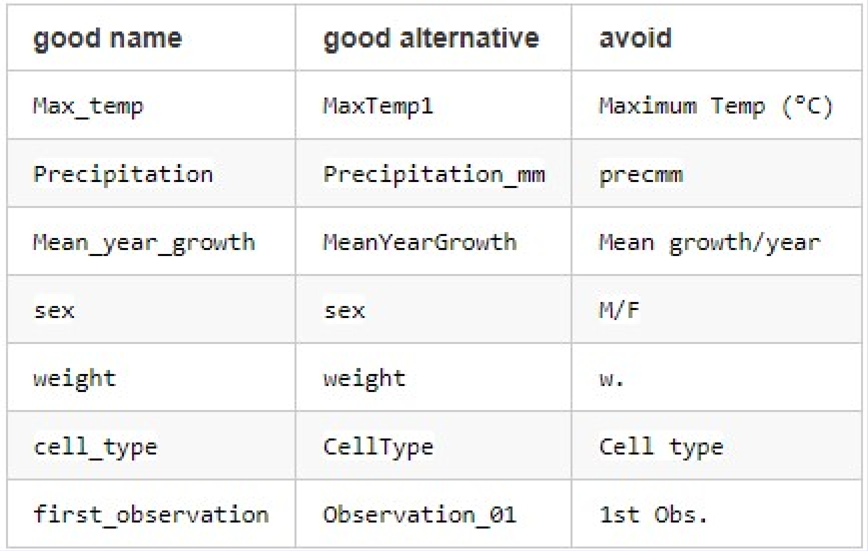
Formatting data within files

Formatted Table Better alternative



	A	B	C	D
1	id	date	glucose	outlier
2	101	2015-06-14	149.3	FALSE
3	102	2015-06-14	95.3	FALSE
4	103	2015-06-18	97.5	FALSE
5	104	2015-06-18	1.1	TRUE
6	105	2015-06-18	108.0	FALSE
7	106	2015-06-20	149.0	FALSE
8	107	2015-06-20	169.4	FALSE

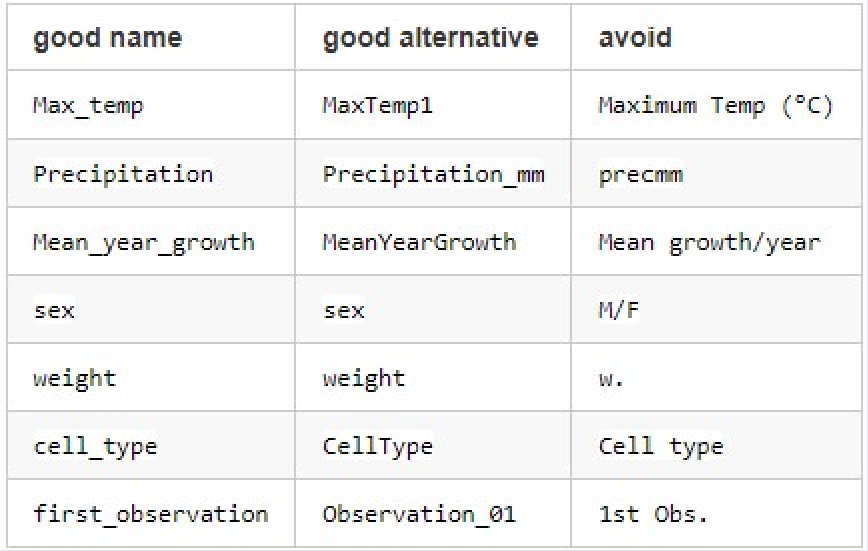
Naming things



good name	good alternative	avoid
Max_temp	MaxTemp1	Maximum Temp (°C)
Precipitation	Precipitation_mm	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
first_observation	Observation_01	1st Obs.

Variable Names

Naming things



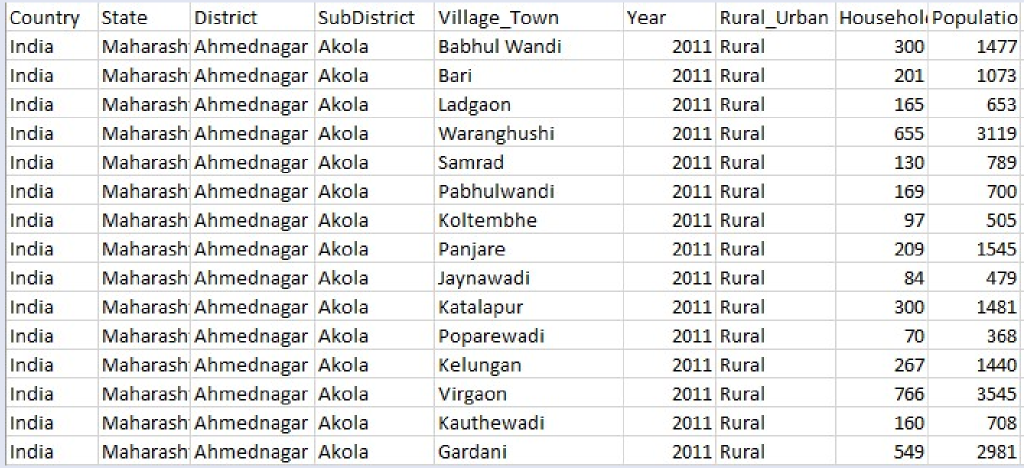
good name	good alternative	avoid
Max_temp	MaxTemp1	Maximum Temp (°C)
Precipitation	Precipitation_mm	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
first_observation	Observation_01	1st Obs.

Variable Names

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc

File Names

Analysing data in Excel



Country	State	District	SubDistrict	Village_Town	Year	Rural_Urban	Househol	Populatio
India	Maharash	Ahmednagar	Akola	Babhul Wandi	2011	Rural	300	1477
India	Maharash	Ahmednagar	Akola	Bari	2011	Rural	201	1073
India	Maharash	Ahmednagar	Akola	Ladgaon	2011	Rural	165	653
India	Maharash	Ahmednagar	Akola	Waranghushi	2011	Rural	655	3119
India	Maharash	Ahmednagar	Akola	Samrad	2011	Rural	130	789
India	Maharash	Ahmednagar	Akola	Pabhulwandi	2011	Rural	169	700
India	Maharash	Ahmednagar	Akola	Koltembhe	2011	Rural	97	505
India	Maharash	Ahmednagar	Akola	Panjare	2011	Rural	209	1545
India	Maharash	Ahmednagar	Akola	Jaynawadi	2011	Rural	84	479
India	Maharash	Ahmednagar	Akola	Katalapur	2011	Rural	300	1481
India	Maharash	Ahmednagar	Akola	Poparewadi	2011	Rural	70	368
India	Maharash	Ahmednagar	Akola	Kelungan	2011	Rural	267	1440
India	Maharash	Ahmednagar	Akola	Virgaon	2011	Rural	766	3545
India	Maharash	Ahmednagar	Akola	Kauthewadi	2011	Rural	160	708
India	Maharash	Ahmednagar	Akola	Gardani	2011	Rural	549	2981

Primary Population Census 2011

To-Do

Link to the file - Download from here

1. Open the file in excel
2. Count the total number of districts
3. Find the district with the highest number of sub districts
4. Find the village with the highest number of households
5. Find the top 10 villages (**having at-least 50 households**) with highest percentage of:
    1. Female population
    2. Female literate population
    3. Female working population

Database Tools

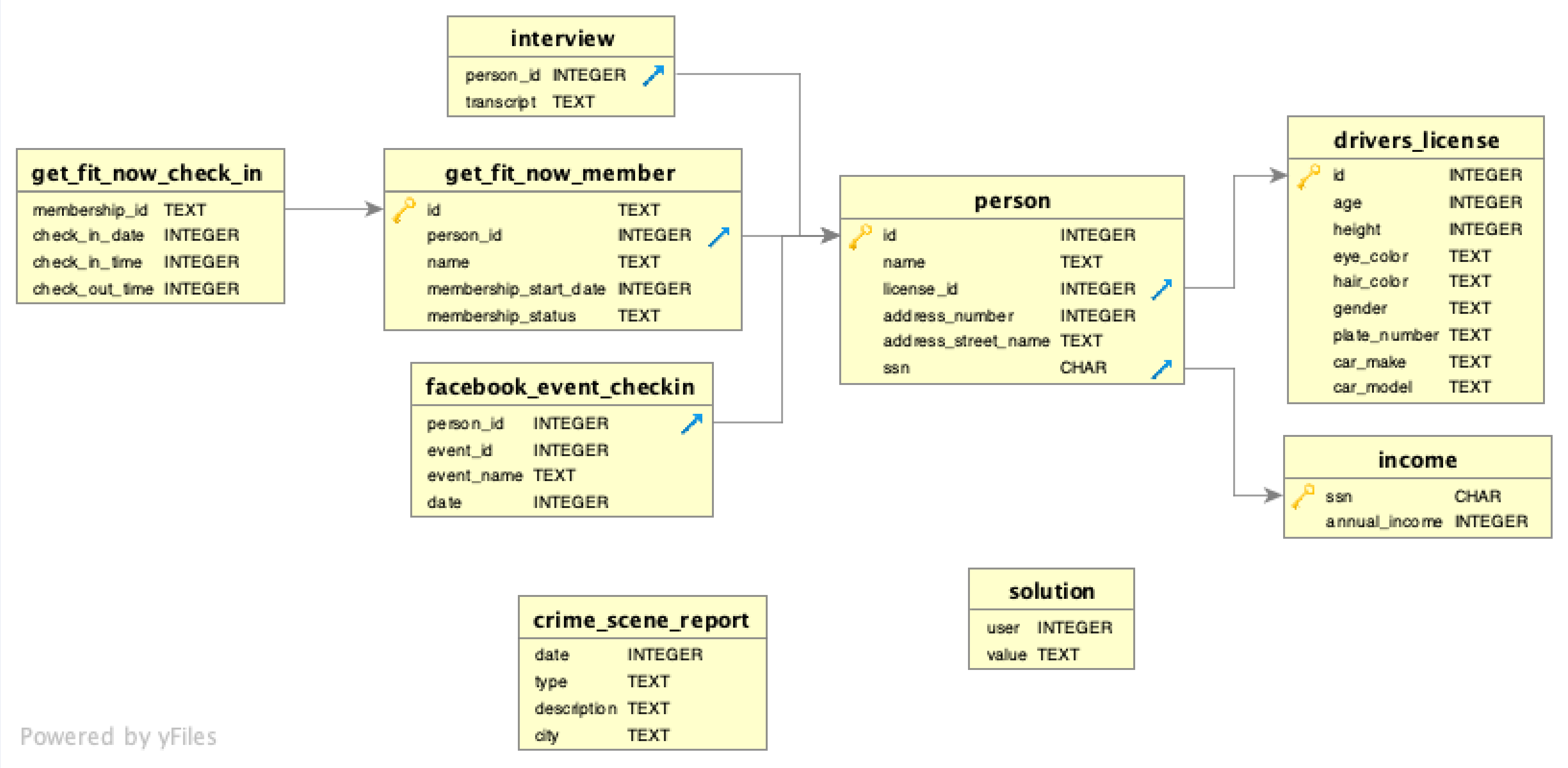
Database

PostgreSQL

Database Manager

pgAdmin

Database Schema



interview	
person_id	INTEGER
transcript	TEXT

get_fit_now_check_in	
membership_id	TEXT
check_in_date	INTEGER
check_in_time	INTEGER
check_out_time	INTEGER

get_fit_now_member	
id	TEXT
person_id	INTEGER
name	TEXT
membership_start_date	INTEGER
membership_status	TEXT

facebook_event_checkin	
person_id	INTEGER
event_id	INTEGER
event_name	TEXT
date	INTEGER

person	
id	INTEGER
name	TEXT
license_id	INTEGER
address_number	INTEGER
address_street_name	TEXT
ssn	CHAR

drivers_license	
id	INTEGER
age	INTEGER
height	INTEGER
eye_color	TEXT
hair_color	TEXT
gender	TEXT
plate_number	TEXT
car_make	TEXT
car_model	TEXT

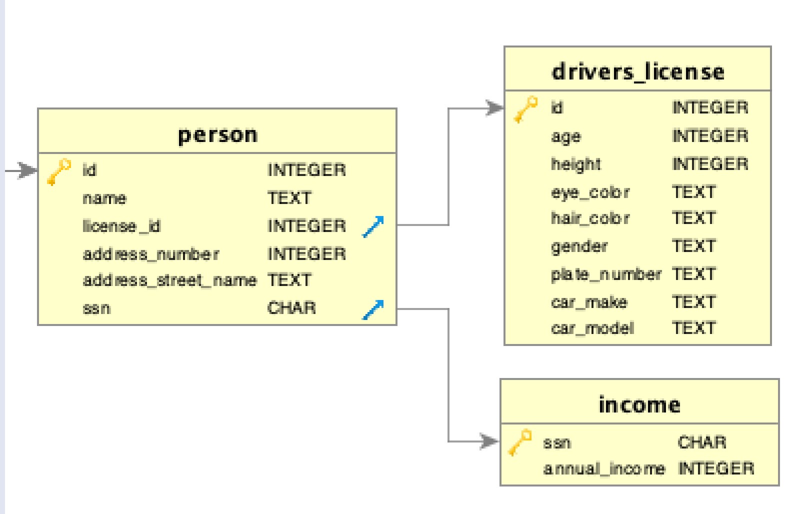
income	
ssn	CHAR
annual_income	INTEGER

crime_scene_report	
date	INTEGER
type	TEXT
description	TEXT
city	TEXT

solution	
user	INTEGER
value	TEXT

A schema diagram

Database Tables



person

id	INTEGER	
name	TEXT	
license_id	INTEGER	↗
address_number	INTEGER	
address_street_name	TEXT	
ssn	CHAR	↗

drivers_license

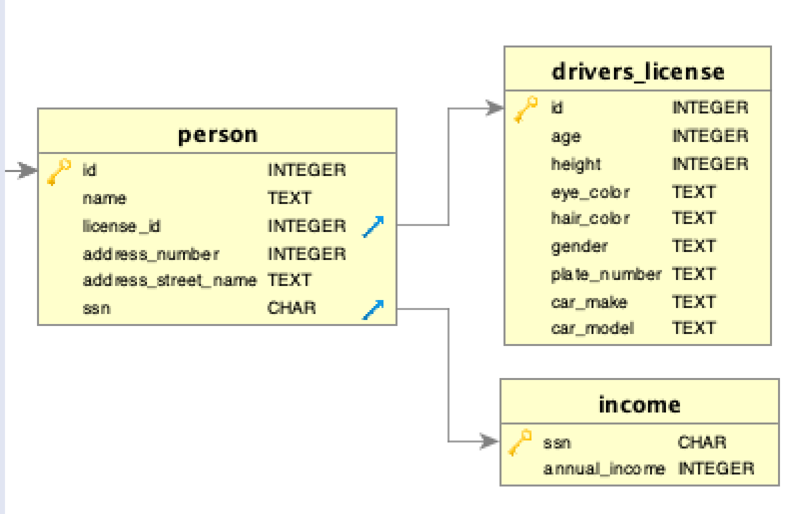
id	INTEGER
age	INTEGER
height	INTEGER
eye_color	TEXT
hair_color	TEXT
gender	TEXT
plate_number	TEXT
car_make	TEXT
car_model	TEXT

income

ssn	CHAR
annual_income	INTEGER

Tables in a database

Database Tables



person

id	INTEGER	
name	TEXT	
license_id	INTEGER	↗
address_number	INTEGER	
address_street_name	TEXT	
ssn	CHAR	↗

drivers_license

id	INTEGER
age	INTEGER
height	INTEGER
eye_color	TEXT
hair_color	TEXT
gender	TEXT
plate_number	TEXT
car_make	TEXT
car_model	TEXT

income

ssn	CHAR
annual_income	INTEGER

Tables in a database

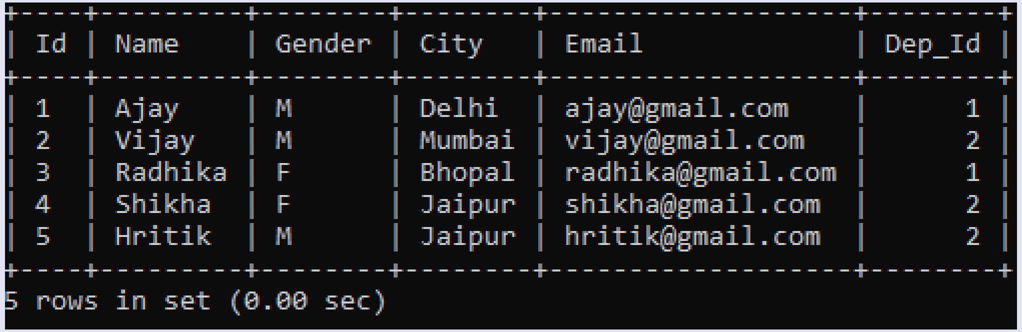
Table Names

1. `person`
2. `drivers_license`
3. `income`

Structured Query Language (SQL)

SQL is the most commonly used language to access data from a database.

SQL Query

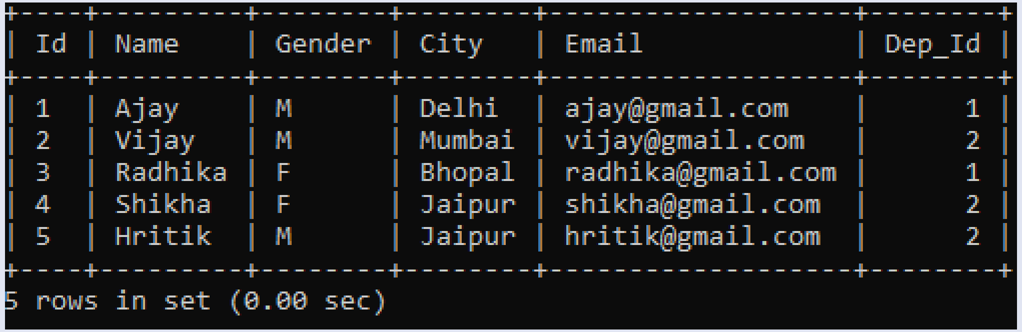


```
+----+---------+--------+--------+--------------------+--------+
| Id | Name    | Gender | City   | Email              | Dep_Id |
+----+---------+--------+--------+--------------------+--------+
| 1  | Ajay    | M      | Delhi  | ajay@gmail.com     |      1 |
| 2  | Vijay   | M      | Mumbai | vijay@gmail.com    |      2 |
| 3  | Radhika | F      | Bhopal | radhika@gmail.com  |      1 |
| 4  | Shikha  | F      | Jaipur | shikha@gmail.com   |      2 |
| 5  | Hritik  | M      | Jaipur | hritik@gmail.com   |      2 |
+----+---------+--------+--------+--------------------+--------+
5 rows in set (0.00 sec)
```

Table Name: **employee**

SELECT * FROM employee

SQL Query - SELECT

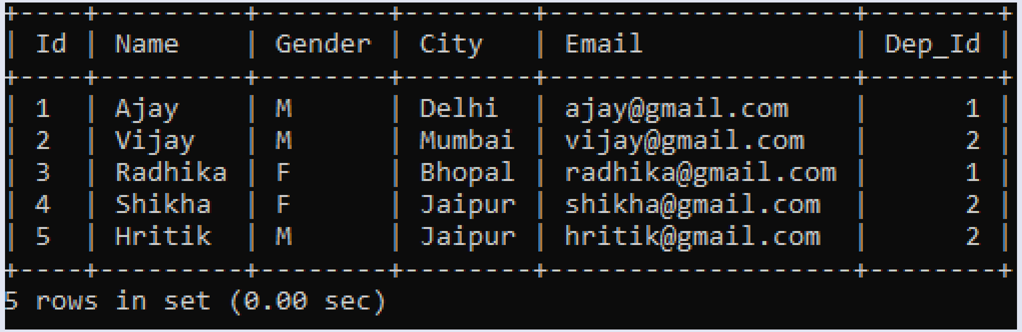


```
+----+---------+--------+--------+-------------------+--------+
| Id | Name    | Gender | City   | Email             | Dep_Id |
+----+---------+--------+--------+-------------------+--------+
| 1  | Ajay    | M      | Delhi  | ajay@gmail.com    |      1 |
| 2  | Vijay   | M      | Mumbai | vijay@gmail.com   |      2 |
| 3  | Radhika | F      | Bhopal | radhika@gmail.com |      1 |
| 4  | Shikha  | F      | Jaipur | shikha@gmail.com  |      2 |
| 5  | Hritik  | M      | Jaipur | hritik@gmail.com  |      2 |
+----+---------+--------+--------+-------------------+--------+
5 rows in set (0.00 sec)
```

Table Name: **employee**

SELECT * FROM employee

SQL Query - All

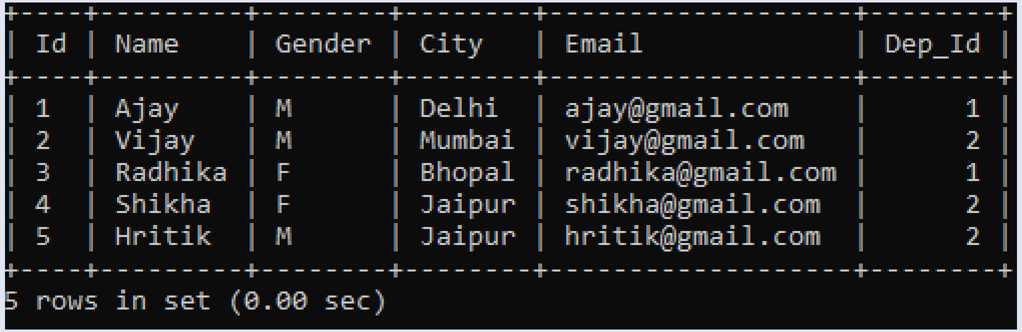


```
+----+---------+--------+---------+-------------------+--------+
| Id | Name    | Gender | City    | Email             | Dep_Id |
+----+---------+--------+---------+-------------------+--------+
| 1  | Ajay    | M      | Delhi   | ajay@gmail.com    |      1 |
| 2  | Vijay   | M      | Mumbai  | vijay@gmail.com   |      2 |
| 3  | Radhika | F      | Bhopal  | radhika@gmail.com |      1 |
| 4  | Shikha  | F      | Jaipur  | shikha@gmail.com  |      2 |
| 5  | Hritik  | M      | Jaipur  | hritik@gmail.com  |      2 |
+----+---------+--------+---------+-------------------+--------+
5 rows in set (0.00 sec)
```

Table Name: **employee**

SELECT * FROM employee

SQL Query - FROM



```
+----+---------+--------+--------+--------------------+--------+
| Id | Name    | Gender | City   | Email              | Dep_Id |
+----+---------+--------+--------+--------------------+--------+
| 1  | Ajay    | M      | Delhi  | ajay@gmail.com     |      1 |
| 2  | Vijay   | M      | Mumbai | vijay@gmail.com    |      2 |
| 3  | Radhika | F      | Bhopal | radhika@gmail.com  |      1 |
| 4  | Shikha  | F      | Jaipur | shikha@gmail.com   |      2 |
| 5  | Hritik  | M      | Jaipur | hritik@gmail.com   |      2 |
+----+---------+--------+--------+--------------------+--------+
5 rows in set (0.00 sec)
```

Table Name: **employee**

SELECT * **FROM** employee

SQL Query - Select variable[s]

SELECT **Name, Gender** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Select variable[s]

SELECT **Name, Gender** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Name	Gender
Ajay	M
Vijay	M
Radhika	F
Shikha	F
Hrithik	M

SQL Query - WHERE (Filter Table)

SELECT **Name** FROM employee **WHERE gender='M'**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - WHERE (Filter Table)

SELECT **Name** FROM employee **WHERE gender='M'**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Name
Ajay
Vijay
Hrithik

SQL Query - Sort Rows - Ascending

```
SELECT Name, Dep_Id FROM employee ORDER BY Dep_Id
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Sort Rows - Ascending

SELECT **Name, Dep_Id** FROM employee **ORDER BY** Dep_Id

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Name	Dep_Id
Ajay	1
Radhika	1
Vijay	2
Shikha	2
Hrithik	2

SQL Query - Sort Rows - Descending

```
SELECT Name, Dep_Id FROM employee ORDER BY Dep_Id DESC
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Sort Rows - Descending

SELECT **Name, Dep_Id** FROM employee **ORDER BY** Dep_Id **DESC**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Name	Dep_Id
Vijay	2
Shikha	2
Hrithik	2
Ajay	1
Radhika	1

SQL Query - Limit Rows

SELECT **Name, Gender** FROM employee **LIMIT 1**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

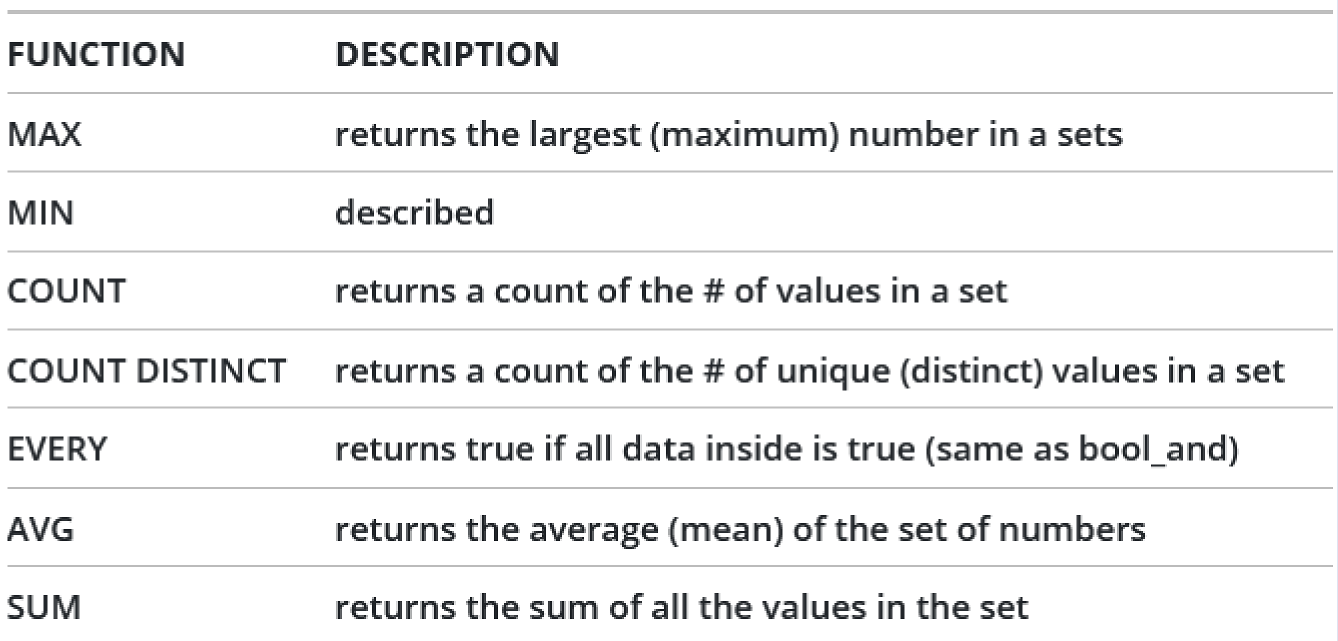
SQL Query - Limit Rows

SELECT **Name, Gender** FROM employee **LIMIT 1**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Name	Gender
Ajay	M

SQL Functions



FUNCTION	DESCRIPTION
MAX	returns the largest (maximum) number in a sets
MIN	described
COUNT	returns a count of the # of values in a set
COUNT DISTINCT	returns a count of the # of unique (distinct) values in a set
EVERY	returns true if all data inside is true (same as bool_and)
AVG	returns the average (mean) of the set of numbers
SUM	returns the sum of all the values in the set

SQL Query - Count all rows

```
SELECT COUNT(*) FROM employee
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Count all rows

```
SELECT COUNT(*) FROM employee
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

$$TotalRows \rightarrow 5$$

SQL Query - Count unique rows

```
SELECT COUNT(DISTINCT Gender) FROM employee
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Count unique rows

```
SELECT COUNT(DISTINCT Gender) FROM employee
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

$$TotalRows \rightarrow 2$$

SQL Query - Calculate SUM

SELECT **SUM(Points)** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Calculate SUM

SELECT **SUM(Points)** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

$$TotalSum \rightarrow 65$$

SQL Query - Find Maximum value

SELECT **MAX(Points)** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - Find Maximum value

SELECT **MAX(Points)** FROM employee

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

$$Ma\xi\mu mP\oint s = 25$$

SQL Quiz

Find the row with maximum number of points **without using MAX**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Quiz

Find the row with maximum number of points **without using MAX**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Hint: **Use ORDER BY and LIMIT**

SQL Quiz

Find the row with maximum number of points **without using MAX**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Hint: **Use ORDER BY and LIMIT**

```
SELECT * FROM employee ORDER BY points LIMIT 1
```

SQL Query - GROUP BY

SELECT **Gender, COUNT(*) as Total** FROM employee **GROUP BY Gender**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - GROUP BY

```
SELECT Gender, COUNT(*) as Total FROM employee GROUP BY Gender
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Gender	Total
M	3
F	2

SQL Query - GROUP BY (SUM)

```
SELECT Dep_Id, sum(Points) as Total_Points FROM employee GROUP BY Dep_Id
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - GROUP BY (SUM)

```
SELECT Dep_Id, sum(Points) as Total_Points FROM employee GROUP BY Dep_Id
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

Dep_Id	Total_Points
1	25
2	40

SQL Query - GROUP BY (MAX) + ORDER BY

SELECT **City, MAX(Points) as max_Points** FROM employee **GROUP BY City ORDER BY max_Points DESC**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

SQL Query - GROUP BY (MAX) + ORDER BY

SELECT **City, MAX(Points) as max_Points** FROM employee **GROUP BY City ORDER BY max_Points DESC**

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

City	max_Points
Jaipur	25
Bhipal	15
Delhi	10
Mumbai	5

SQL Query - GROUP BY + HAVING (Group Filter)

```
SELECT City, sum(Points) as total_Points FROM employee GROUP BY city HAVING sum(points) > 10
ORDER BY total_points desc;
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

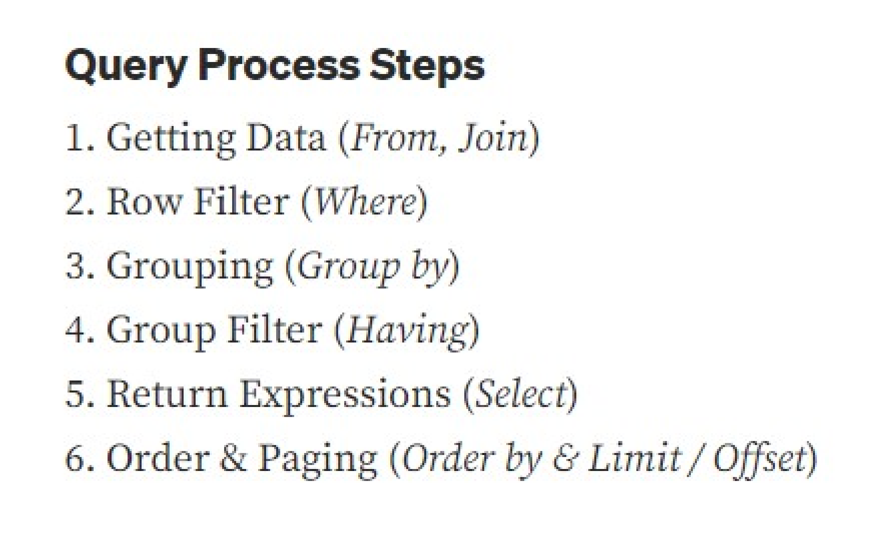
SQL Query - GROUP BY + HAVING (Group Filter)

```
SELECT City, sum(Points) as total_Points FROM employee GROUP BY city HAVING sum(points) > 10
ORDER BY total_points desc;
```

Id	Name	Gender	City	Dep_Id	Points
1	Ajay	M	Delhi	1	10
2	Vijay	M	Mumbai	2	5
3	Radhika	F	Bhipal	1	15
4	Shikha	F	Jaipur	2	25
5	Hrithik	M	Jaipur	2	10

City	total_Points
Jaipur	35
Bhipal	15

Order of SQL commands

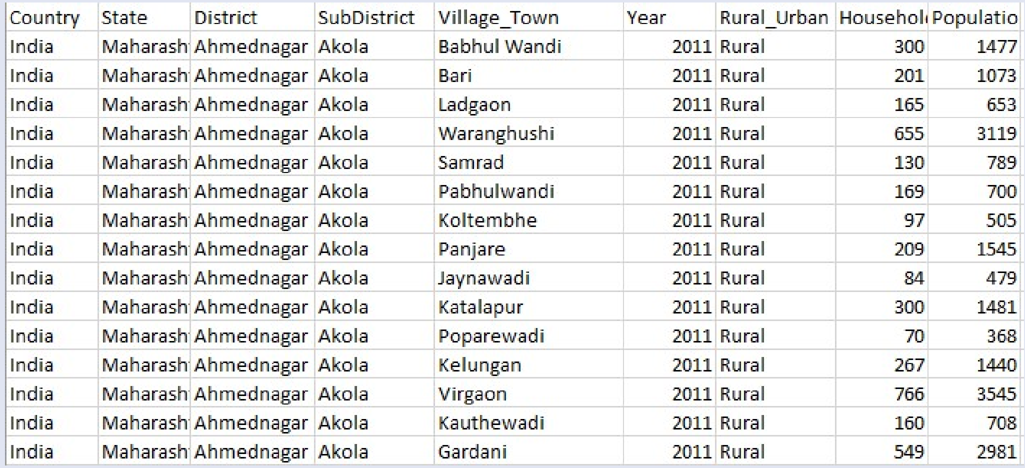


Query Process Steps

1. Getting Data (*From, Join*)
2. Row Filter (*Where*)
3. Grouping (*Group by*)
4. Group Filter (*Having*)
5. Return Expressions (*Select*)
6. Order & Paging (*Order by & Limit / Offset*)

The 6 Steps of a SQL Select Statement Process

Analysing data using SQL



Country	State	District	SubDistrict	Village_Town	Year	Rural_Urban	Househol	Populatio
India	Maharash	Ahmednagar	Akola	Babhul Wandi	2011	Rural	300	1477
India	Maharash	Ahmednagar	Akola	Bari	2011	Rural	201	1073
India	Maharash	Ahmednagar	Akola	Ladgaon	2011	Rural	165	653
India	Maharash	Ahmednagar	Akola	Waranghushi	2011	Rural	655	3119
India	Maharash	Ahmednagar	Akola	Samrad	2011	Rural	130	789
India	Maharash	Ahmednagar	Akola	Pabhulwandi	2011	Rural	169	700
India	Maharash	Ahmednagar	Akola	Koltembhe	2011	Rural	97	505
India	Maharash	Ahmednagar	Akola	Panjare	2011	Rural	209	1545
India	Maharash	Ahmednagar	Akola	Jaynawadi	2011	Rural	84	479
India	Maharash	Ahmednagar	Akola	Katalapur	2011	Rural	300	1481
India	Maharash	Ahmednagar	Akola	Poparewadi	2011	Rural	70	368
India	Maharash	Ahmednagar	Akola	Kelungan	2011	Rural	267	1440
India	Maharash	Ahmednagar	Akola	Virgaon	2011	Rural	766	3545
India	Maharash	Ahmednagar	Akola	Kauthewadi	2011	Rural	160	708
India	Maharash	Ahmednagar	Akola	Gardani	2011	Rural	549	2981

Primary Population Census 2011

To-Do

1. Locate the table in the database and print the first 10 rows
2. Count the total number of districts
3. Select the top 10 districts with the highest number of sub districts
4. Select the top 10 villages with the highest number of households
5. Find the top 10 villages (**having at-least 50 households**) with highest percentage of:
    1. Female population
    2. Female literate population
    3. Female working population

Query - 1

Locate the table in the database and print the first 10 rows

Query - 1

Locate the table in the database and print the first 10 rows

```
SELECT * FROM census11 LIMIT 10
```

Query - 1

Locate the table in the database and print the first 10 rows

```
SELECT * FROM census11 LIMIT 10
```

Id	Country	State	District	SubDistrict	Village_Town	Year	Rural_Urban
1	India	Maharashtra	Ahmednagar	Akola	Babhul Wandi	2011	Rural
2	India	Maharashtra	Ahmednagar	Akola	Bari	2011	Rural
3	India	Maharashtra	Ahmednagar	Akola	Ladgaon	2011	Rural
4	India	Maharashtra	Ahmednagar	Akola	Waranghushi	2011	Rural
5	India	Maharashtra	Ahmednagar	Akola	Samrad	2011	Rural

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

Count the total number of districts

Query - 2

Count the total number of districts

Query - 2

Count the total number of districts

```
SELECT COUNT(DISTINCT district) as Total_Districts FROM census11
```

Query - 2

Count the total number of districts

```
SELECT COUNT(DISTINCT district) as Total_Districts FROM census11
```

Total_Districts
35

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

~~2. Count the total number of districts~~

Select the top 10 districts with the highest number of sub districts

Query - 3

Select the top 10 district with the highest number of sub districts

Query - 3

Select the top 10 district with the highest number of sub districts

```
SELECT district, COUNT(DISTINCT subdistrict) as Total_SubDistricts FROM census11 GROUP BY district ORDER BY Total_SubDistricts DESC, district LIMIT 10
```

Query - 3

Select the top 10 district with the highest number of sub districts

```
SELECT district, COUNT(DISTINCT subdistrict) as Total_SubDistricts FROM census11 GROUP BY district ORDER
BY Total_SubDistricts DESC, district LIMIT 10
```

District	Total_SubDistricts
Nanded	16
Nashik	16
Yavatmal	16
Ahmednagar	15
Chandrapur	15
Jalgaon	15
Nagpur	15
Pune	15

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

~~2. Count the total number of districts~~

~~3. Select the top 10 districts with the highest number of sub districts~~

Select the top 10 villages with the highest number of households

Query - 4

Select the top 10 villages with the highest number of households

Query - 4

Select the top 10 villages with the highest number of households

```
SELECT district, subdistrict,village_town, Households FROM census11 WHERE rural_urban = 'Rural' ORDER BY
Households DESC LIMIT 10
```

Query - 4

Select the top 10 villages with the highest number of households

SELECT district, subdistrict,village_town, Households FROM census11 WHERE rural_urban = 'Rural' ORDER BY Households DESC LIMIT 10

District	SubDistrict	Village_Town	Households
Pune	Haveli	Fursungi	15595
Thane	Kalyan	Nandiwali Tarf Pachanand (N.V.)	9087
Pune	Haveli	Manjari Bk	8401
Nashik	Niphad	Pimpalgaon Baswant	8187
Ahmednagar	Shevgaon	Shevgaon	8013
Pune	Haveli	Keshavnagar-Mundwa	7537
Pune	Haveli	Lahagaon	7526
Sangli	Jat	Jat	7411

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

~~2. Count the total number of districts~~

~~3. Select the top 10 districts with the highest number of sub districts~~

~~4. Select the top 10 villages with the highest number of households~~

Find the top 10 villages (**having at-least 50 households**) with highest percentage of:

Female population

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

~~2. Count the total number of districts~~

~~3. Select the top 10 districts with the highest number of sub districts~~

~~4. Select the top 10 villages with the highest number of households~~

Find the top 10 villages (**having at-least 50 households**) with highest percentage of:

Female population **Female literate population**

Analyse data using SQL

To-Do

~~1. Locate the table in the database and print the first 10 rows~~

~~2. Count the total number of districts~~

~~3. Select the top 10 districts with the highest number of sub districts~~

~~4. Select the top 10 villages with the highest number of households~~

Find the top 10 villages (**having at-least 50 households**) with highest percentage of:

Female population **Female literate population** **Female working population**

Query - 5 (Calculated fields)

Find the top 10 villages, having at-least 50 households, with highest percentage of Female population

Query - 5 (Calculated fields)

Find the top 10 villages, having at-least 50 households, with highest percentage of Female population

```
SELECT district, subdistrict, village_town, (cast(femalepopulation as decimal)/population)*100 as
percent_female_pop FROM census11 WHERE rural_urban = 'Rural' AND households >= 50 ORDER BY
percent_female_pop DESC LIMIT 10
```

Query - 5 (Calculated fields)

Find the top 10 villages, having at-least 50 households, with highest percentage of Female population

```
SELECT district, subdistrict, village_town, (cast(femalepopulation as decimal)/population)*100 as
percent_female_pop FROM census11 WHERE rural_urban = 'Rural' AND households >= 50 ORDER BY
percent_female_pop DESC LIMIT 10
```

District	SubDistrict	Village_Town	percent_female_pop
Gondia	Deori	Charbhata	77.05287
Gadchiroli	Chamorshi	Tumdi	75.07599
Gadchiroli	Dhanora	Sode	71.55050
Ratnagiri	Dapoli	Borivali	71.32616
Raigad	Mangaon	Nhave	70.82803
Raigad	Mhasla	Dehen	69.96337
Nandurbar	Talode	Lobhani	69.77863

Case Study - Tracking field visits

Source: SNEHA

To-Do:

Filter out all cases which are closed

1. For each **cluster, center and CO (community organiser)** :
   1. Count the total number of pregnant women
   2. Count the number of high risk pregnancies
   3. Find the distribution of pregnant women by month of pregnancy
2. Find the cluster, center and CO with the highest number of pregnancies in the sixth and seventh month
3. For all women in this group, find the total number of field visits
4. For all women in the above group, count the total number of visits per month

Case Study - Tracking field visits

To-Do:

1. For each **cluster, center and CO (community organiser)** :
    1. Count the total number of pregnant women
    2. Count the number of high risk pregnancies
    3. Find the distribution of pregnant women by month of pregnancy
2. Find the cluster, center and CO with the highest number of pregnancies in the sixth and seventh month
3. For all women in this group, find the total number of field visits
4. For all women in the above group, count the total number of visits per month

SQL file

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

We don't have all the information in one table so we have to get information from multiple tables.

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

We don't have all the information in one table so we have to get information from multiple tables.

Break the query

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

We don't have all the information in one table so we have to get information from multiple tables.

Break the query

Part 1 Part 2 Final Query

Find the women with high risk pregnancy (current)

```
SELECT id FROM case_anc_visit_reduced WHERE closed=FALSE AND high_risk_preg='Yes'
```

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

We don't have all the information in one table so we have to get information from multiple tables.

Break the query

Part 1 | Part 2 | Final Query

Find the women who were visited at-least once in September, 2022

```
SELECT DISTINCT(caseid) FROM form_anc_visit_reduced WHERE TO_CHAR(anc_visit_date,'YYYY-MM') = '2022-09'
```

Learn more about working with Dates

SQL Subqueries - Working with more than 1 table

How many women with high risk pregnancy were visited at-least once in September, 2022

We don't have all the information in one table so we have to get information from multiple tables.

Break the query

Part 1 Part 2 **Final Query**

Find the womenID which are present in both Part 1 and 2

```
SELECT id FROM case_anc_visit_reduced WHERE closed=FALSE AND high_risk_preg='Yes' AND id IN (SELECT
DISTINCT(caseid) FROM form_anc_visit_reduced WHERE TO_CHAR(anc_visit_date,'YYYY-MM') = '2022-09')
```

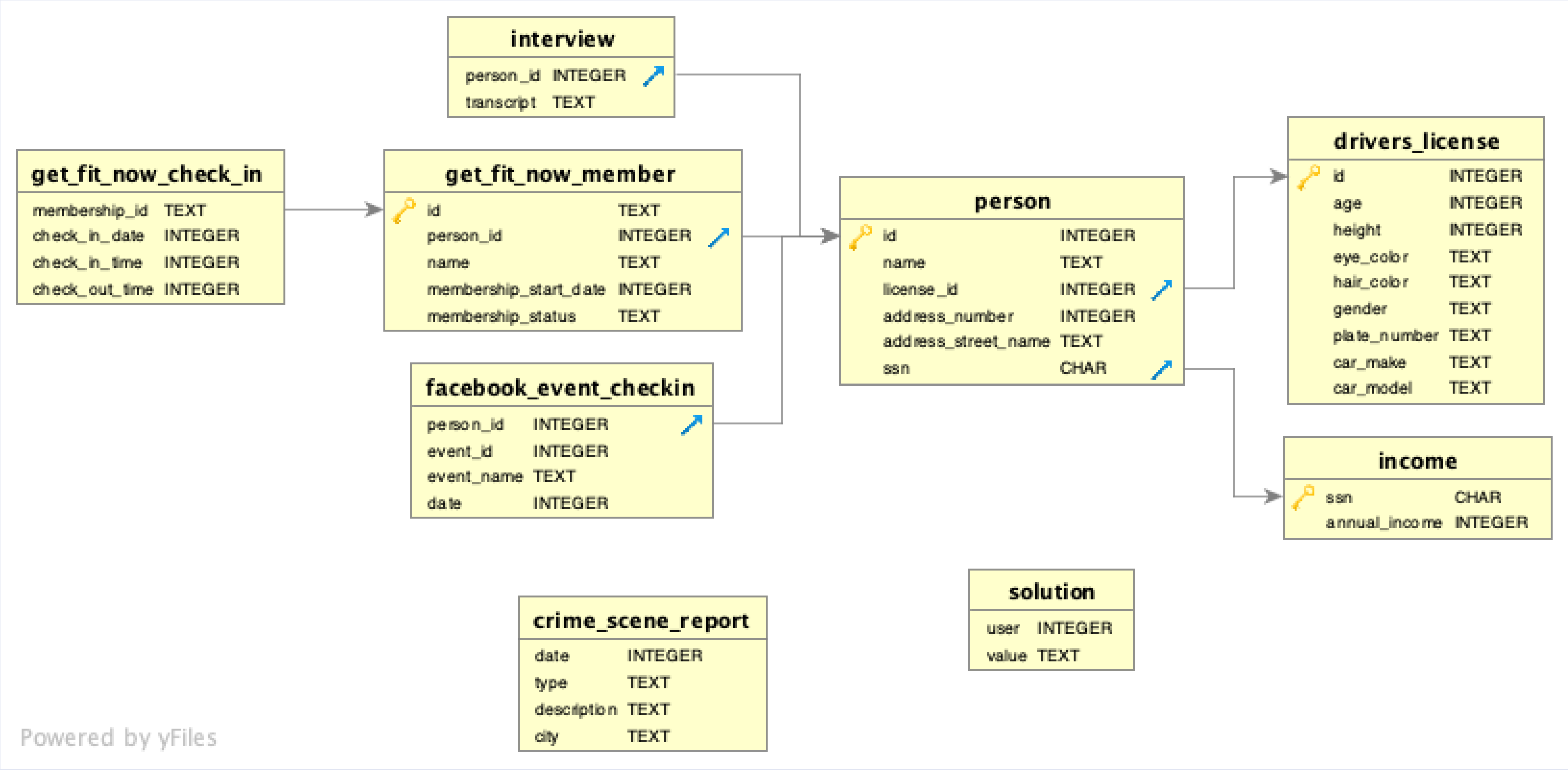
SQL Detective

Source: knightlab

SQL file

SQL Detective - Step by Step

Open the schema diagram



interview

person_id	INTEGER	↗
transcript	TEXT	

get_fit_now_check_in

membership_id	TEXT
check_in_date	INTEGER
check_in_time	INTEGER
check_out_time	INTEGER

get_fit_now_member

🔑	id	TEXT	
	person_id	INTEGER	↗
	name	TEXT	
	membership_start_date	INTEGER	
	membership_status	TEXT	

facebook_event_checkin

person_id	INTEGER	↗
event_id	INTEGER	
event_name	TEXT	
date	INTEGER	

person

🔑	id	INTEGER	
	name	TEXT	
	license_id	INTEGER	↗
	address_number	INTEGER	
	address_street_name	TEXT	
	ssn	CHAR	↗

drivers_license

🔑	id	INTEGER
	age	INTEGER
	height	INTEGER
	eye_color	TEXT
	hair_color	TEXT
	gender	TEXT
	plate_number	TEXT
	car_make	TEXT
	car_model	TEXT

income

🔑	ssn	CHAR
	annual_income	INTEGER

solution

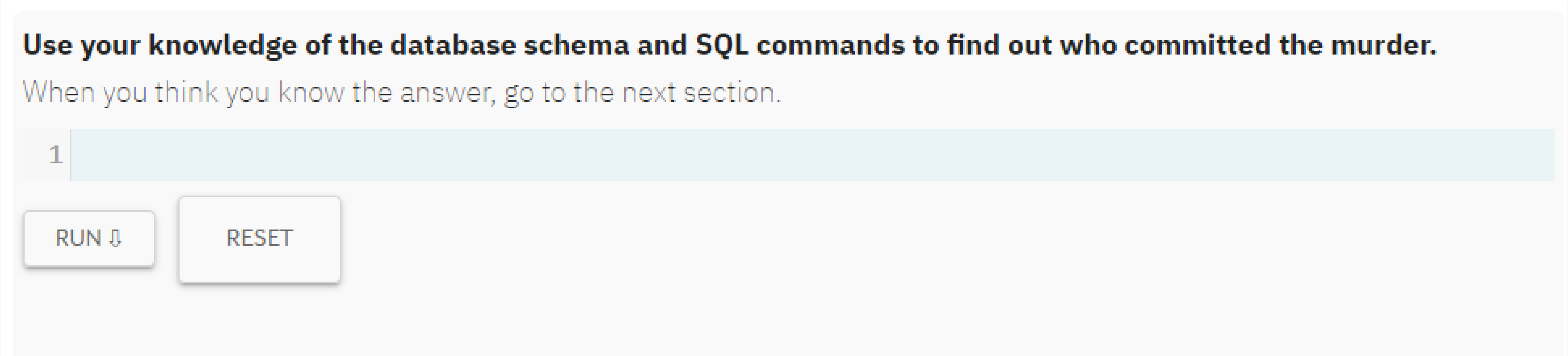
user	INTEGER
value	TEXT

crime_scene_report

date	INTEGER
type	TEXT
description	TEXT
city	TEXT

SQL Detective - Step by Step

Read the clues & Execute the commands inside the web page

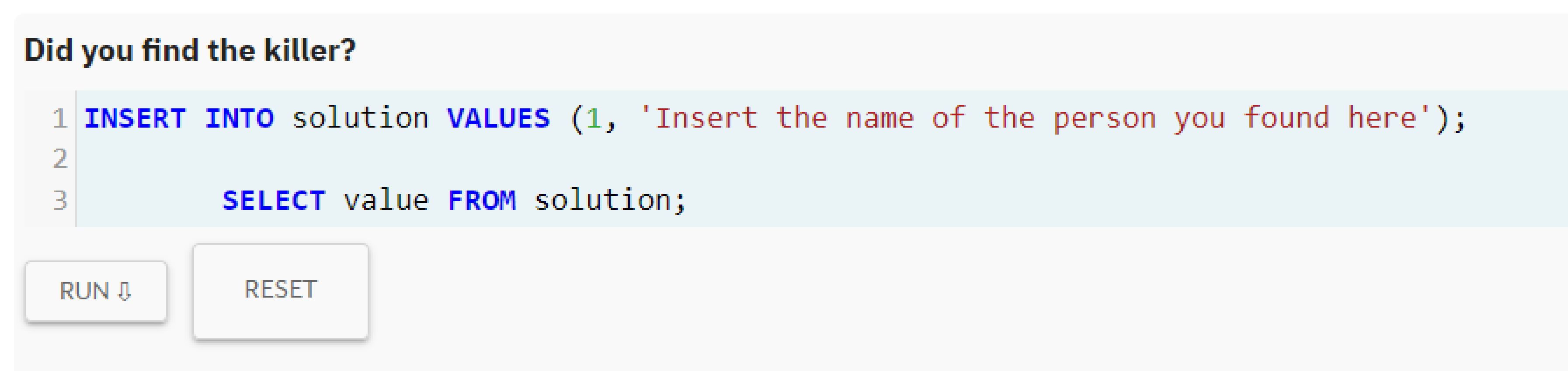


Use your knowledge of the database schema and SQL commands to find out who committed the murder.

When you think you know the answer, go to the next section.

SQL Detective - Step by Step

Check the answer!



Did you find the killer?

```sql
INSERT INTO solution VALUES (1, 'Insert the name of the person you found here');

        SELECT value FROM solution;
```

Resources to learn and practice SQL

Queries and Feedback

Share your feedback here -> https://forms.gle/nBwwbiTXCbAdv5Gz5