Graded Project

HIVE

Week 8

Table of Content

Healthcare

Context

This data is about the cause of death from 2001 to 2008 in France, this will help the France government to identify the reasons which are the reasons for deaths across different genders and take necessary steps.

Data

We have a predefined dataset (CauseofDeath.csv) having 4 columns. The dataset has different attributes like.

1. Time(year)

2. SEX (male or female)

3.Cause

4.Value (count of deaths)

Objective

Show case your hive skills.

Learning Outcomes

Once you successfully complete this exercise, you will be able to work on dynamic

partitioned tables, windows functions in Hive, using SERDE for building the hive tables.

Load Data Into HDFS

The first step is to create a folder and upload data into HDFS

On the CloudX Lab web console:

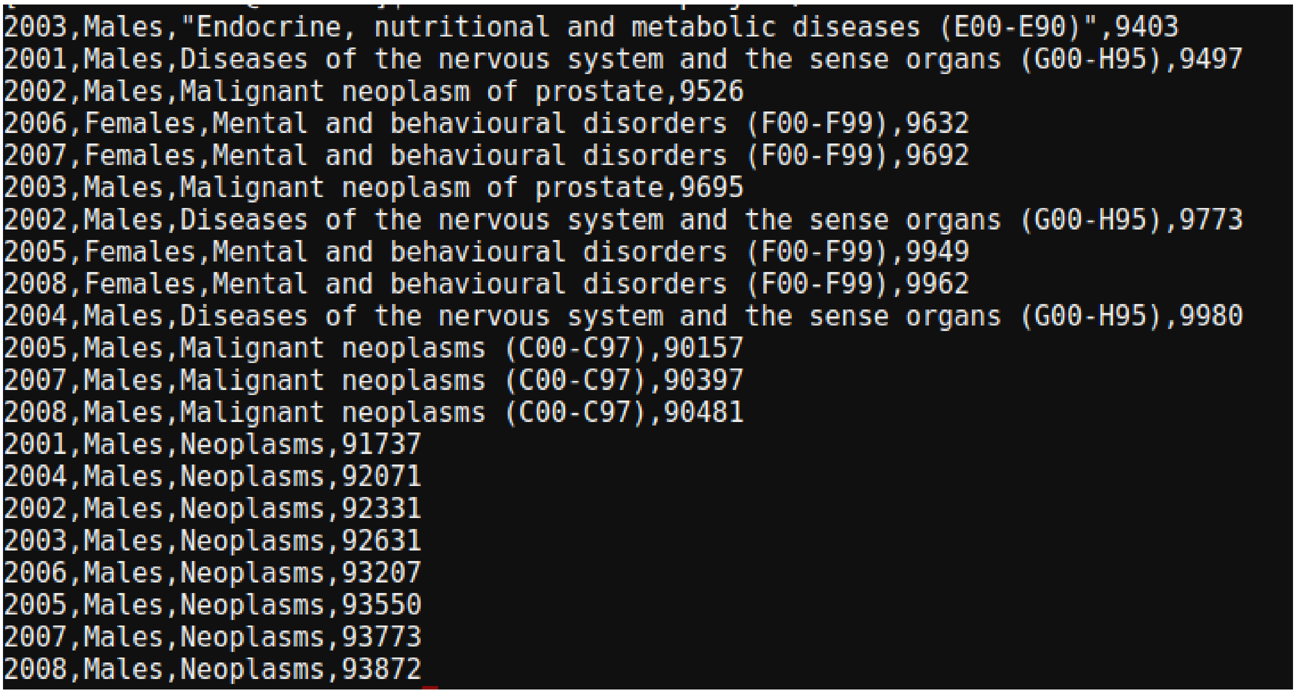
```
ls

hdfs dfs -ls

hdfs dfs -mkdir project

hdfs dfs -put CauseofDeath.csv project

hdfs dfs -tail project/CauseofDeath.csv
```



```
2003,Males,"Endocrine, nutritional and metabolic diseases (E00-E90)",9403
2001,Males,Diseases of the nervous system and the sense organs (G00-H95),9497
2002,Males,Malignant neoplasm of prostate,9526
2006,Females,Mental and behavioural disorders (F00-F99),9632
2007,Females,Mental and behavioural disorders (F00-F99),9692
2003,Males,Malignant neoplasm of prostate,9695
2002,Males,Diseases of the nervous system and the sense organs (G00-H95),9773
2005,Females,Mental and behavioural disorders (F00-F99),9949
2008,Females,Mental and behavioural disorders (F00-F99),9962
2004,Males,Diseases of the nervous system and the sense organs (G00-H95),9980
2005,Males,Malignant neoplasms (C00-C97),90157
2007,Males,Malignant neoplasms (C00-C97),90397
2008,Males,Malignant neoplasms (C00-C97),90481
2001,Males,Neoplasms,91737
2004,Males,Neoplasms,92071
2002,Males,Neoplasms,92331
2003,Males,Neoplasms,92631
2006,Males,Neoplasms,93207
2005,Males,Neoplasms,93550
2007,Males,Neoplasms,93773
2008,Males,Neoplasms,93872_
```

1. **Create the table using the above data set, also create the partition based on the time? After creating the partitioned tables, you need to display the partitions.**

Create database in hive:

```
create database hive_project;
use hive_project;
show tables;
```

Create table causeOfDeath:

```
create table if not exists causeOfDeath (time int, sex STRING, cause
STRING, value int)
row format serde  'org.apache.hadoop.hive.serde2.OpenCSVSerde'
with serdeproperties  (
'separatorChar'=',',
'quoteChar'='"',
'escapeChar'='\\'
)
stored as textfile
tblproperties('skip.header.line.count'= '1');
```

Load the data into the table:

```
load data inpath 'project/CauseofDeath.csv' overwrite into table causeOfDeath;
```

Create partitioned table:

```
create table if not exists cause_partition(sex STRING, cause
STRING, value int)
```

```
partitioned by (time int)
row format serde  'org.apache.hadoop.hive.serde2.OpenCSVSerde'
with serdeproperties  (
'separatorChar'=',',
'quoteChar'='"',
'escapeChar'='\\'
)
stored as textfile ;
```

Configure Hive to allow partitions:

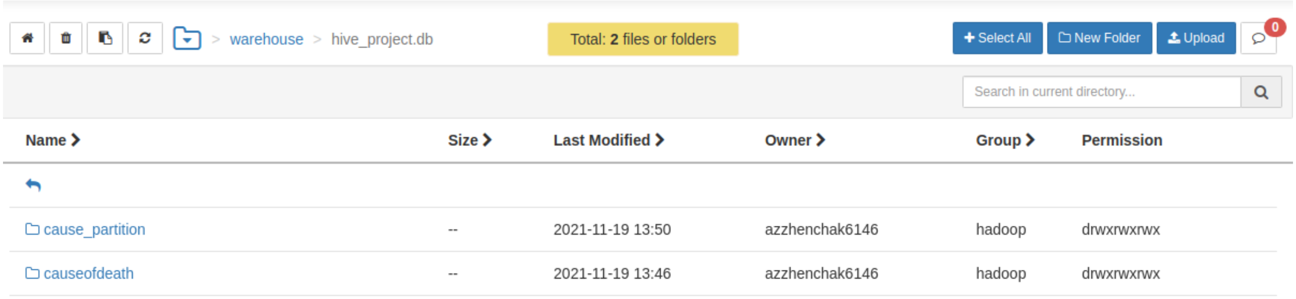
```
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.dynamic.partition=true;
```

Load data into partition table:

```
from causeofdeath c INSERT OVERWRITE TABLE   cause_partition
PARTITION(time)
SELECT c.sex,c.cause,c.value,c.time;
```

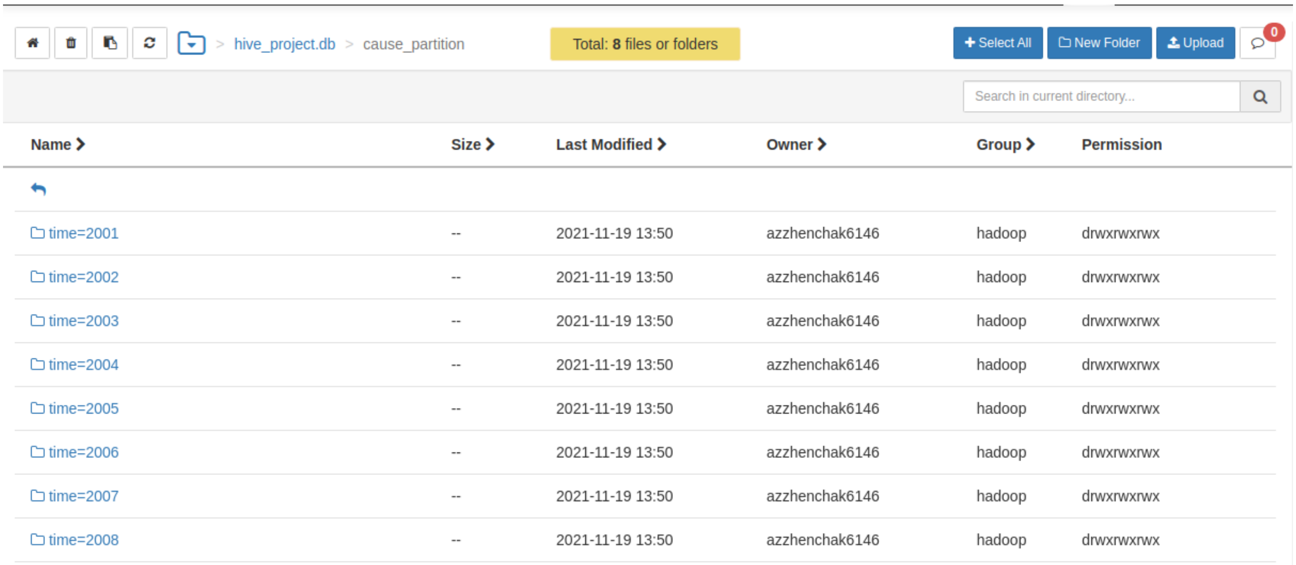
Check data in Ambari cloudxlab:

After partition, there were created 8 new files with data corresponding to the year



warehouse > hive_project.db
Total: 2 files or folders
Select All
New Folder
Upload
0
Search in current directory...

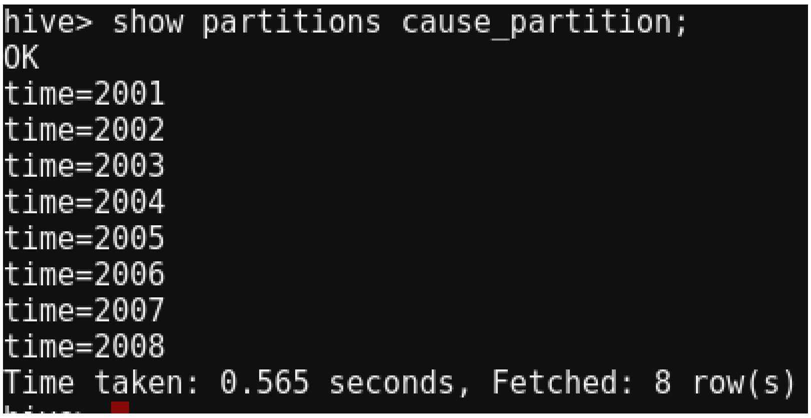
Name	Size	Last Modified	Owner	Group	Permission
cause_partition	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
causeofdeath	--	2021-11-19 13:46	azzhenchak6146	hadoop	drwxrwxrwx



hive_project.db > cause_partition
Total: 8 files or folders
Select All
New Folder
Upload
0
Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission
time=2001	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2002	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2003	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2004	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2005	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2006	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2007	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx
time=2008	--	2021-11-19 13:50	azzhenchak6146	hadoop	drwxrwxrwx

Show partitions:



```
hive> show partitions cause_partition;
OK
time=2001
time=2002
time=2003
time=2004
time=2005
time=2006
time=2007
time=2008
Time taken: 0.565 seconds, Fetched: 8 row(s)
```

2. Total data points collected for each year?

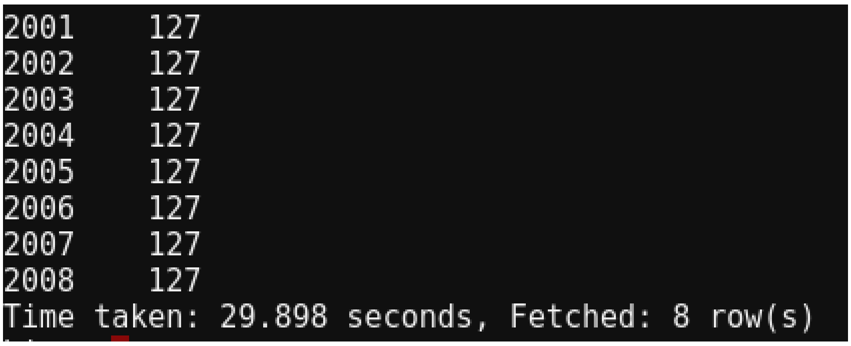
Counting number data points for each year

select time, count(*)

from causeofdeath

group by time;

Output:



```
2001    127
2002    127
2003    127
2004    127
2005    127
2006    127
2007    127
2008    127
Time taken: 29.898 seconds, Fetched: 8 row(s)
```

Data is equally distributed among each year

3. What are different types of cause of death?

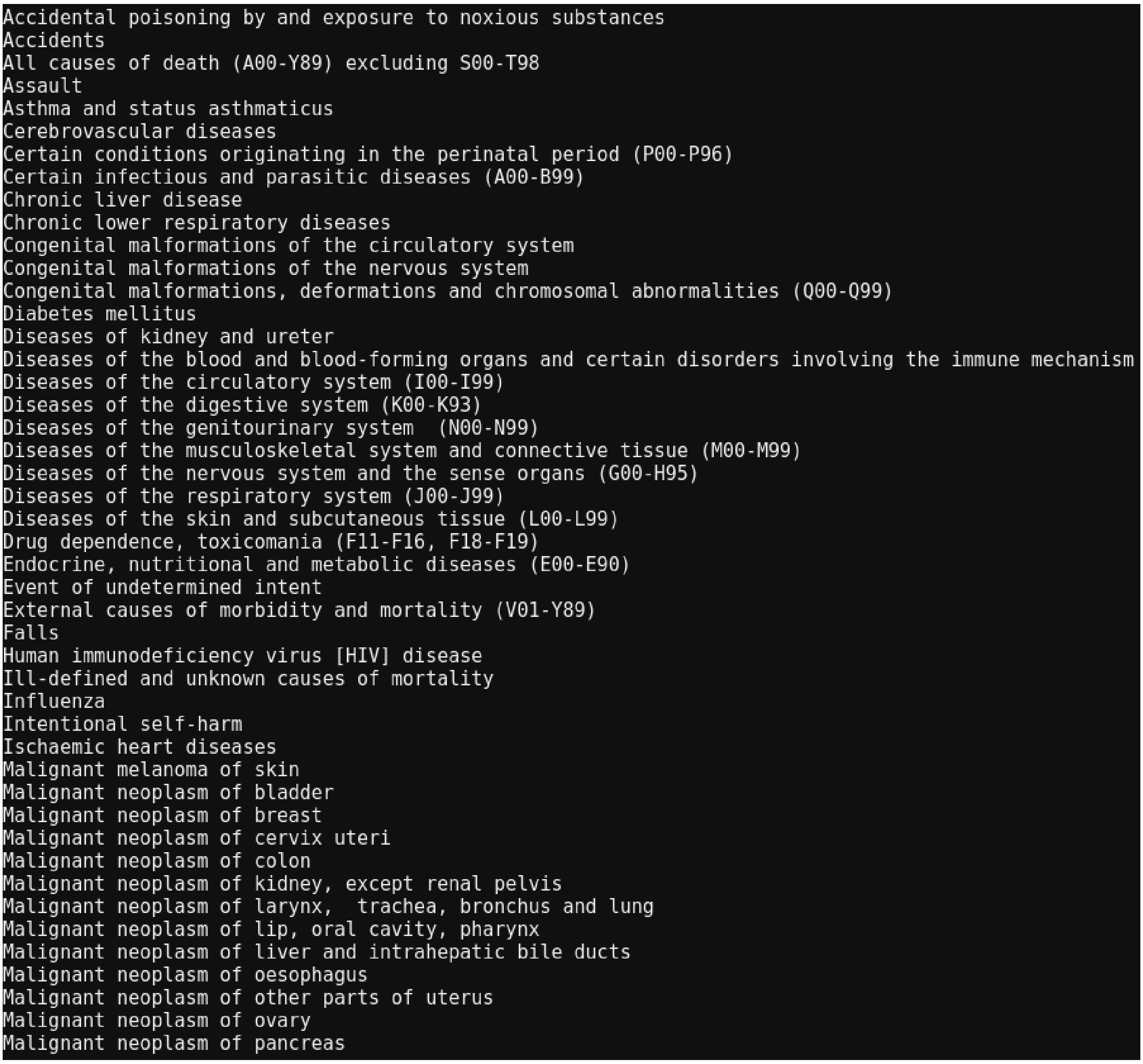
Types of causes of deaths:

select cause

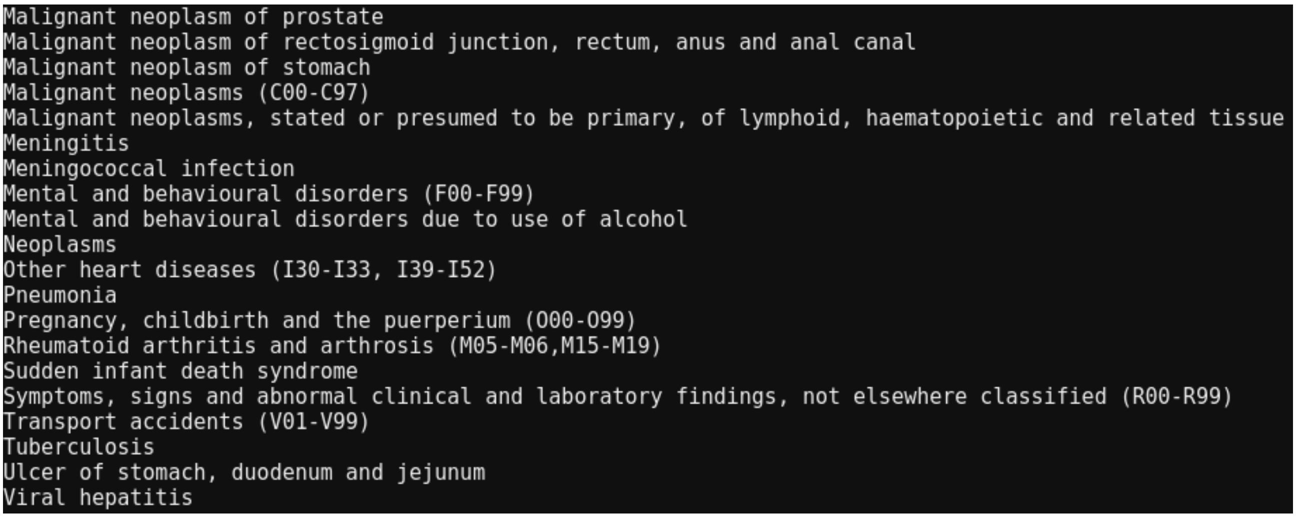
from causeofdeath

group by cause;

Output:



```
Accidental poisoning by and exposure to noxious substances
Accidents
All causes of death (A00-Y89) excluding S00-T98
Assault
Asthma and status asthmaticus
Cerebrovascular diseases
Certain conditions originating in the perinatal period (P00-P96)
Certain infectious and parasitic diseases (A00-B99)
Chronic liver disease
Chronic lower respiratory diseases
Congenital malformations of the circulatory system
Congenital malformations of the nervous system
Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)
Diabetes mellitus
Diseases of kidney and ureter
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Diseases of the circulatory system (I00-I99)
Diseases of the digestive system (K00-K93)
Diseases of the genitourinary system  (N00-N99)
Diseases of the musculoskeletal system and connective tissue (M00-M99)
Diseases of the nervous system and the sense organs (G00-H95)
Diseases of the respiratory system (J00-J99)
Diseases of the skin and subcutaneous tissue (L00-L99)
Drug dependence, toxicomania (F11-F16, F18-F19)
Endocrine, nutritional and metabolic diseases (E00-E90)
Event of undetermined intent
External causes of morbidity and mortality (V01-Y89)
Falls
Human immunodeficiency virus [HIV] disease
Ill-defined and unknown causes of mortality
Influenza
Intentional self-harm
Ischaemic heart diseases
Malignant melanoma of skin
Malignant neoplasm of bladder
Malignant neoplasm of breast
Malignant neoplasm of cervix uteri
Malignant neoplasm of colon
Malignant neoplasm of kidney, except renal pelvis
Malignant neoplasm of larynx,  trachea, bronchus and lung
Malignant neoplasm of lip, oral cavity, pharynx
Malignant neoplasm of liver and intrahepatic bile ducts
Malignant neoplasm of oesophagus
Malignant neoplasm of other parts of uterus
Malignant neoplasm of ovary
Malignant neoplasm of pancreas
```



```
Malignant neoplasm of prostate
Malignant neoplasm of rectosigmoid junction, rectum, anus and anal canal
Malignant neoplasm of stomach
Malignant neoplasms (C00-C97)
Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue
Meningitis
Meningococcal infection
Mental and behavioural disorders (F00-F99)
Mental and behavioural disorders due to use of alcohol
Neoplasms
Other heart diseases (I30-I33, I39-I52)
Pneumonia
Pregnancy, childbirth and the puerperium (O00-O99)
Rheumatoid arthritis and arthrosis (M05-M06,M15-M19)
Sudden infant death syndrome
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)
Transport accidents (V01-V99)
Tuberculosis
Ulcer of stomach, duodenum and jejunum
Viral hepatitis
```

4. Total number of deaths, find out the year in which the deaths are more?

Total number of deaths:

```
select cast(sum(value) as int) from causeofdeath;
```

Output:

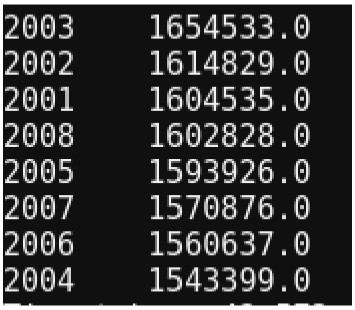
```
12745563
```

Total number of deaths which happend between 2001 and 2008 is 12745563

Total number of deaths in each year in descending order:

```
select time, sum(value) as total_deaths
from causeofdeath
group by time
order by total_deaths desc;
```

Output:



```
2003    1654533.0
2002    1614829.0
2001    1604535.0
2008    1602828.0
2005    1593926.0
2007    1570876.0
2006    1560637.0
2004    1543399.0
```

Max number of deaths in the year :

```
select time, sum(value) as total_deaths
from causeofdeath
group by time
order by total_deaths desc
limit 1;
```

Output:

```
2003 1654533.0
```

The maximum number of deaths happened during the year 2003 due to various causes.

5. Top 5 causes of death and Top 5 causes of death across different sex?

Top 5 causes of death:

select sum(value) as sum_val, cause

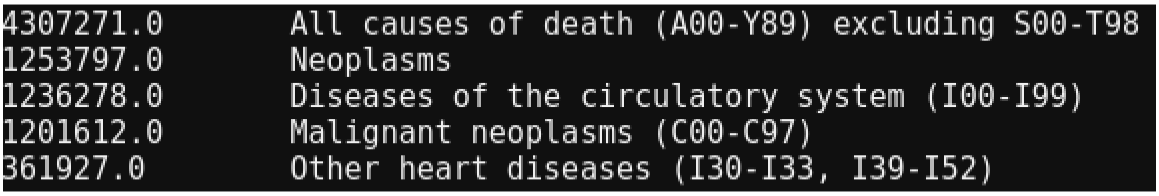
from causeofdeath

group by cause

order by sum_val desc

limit 5;

Output:



```
4307271.0       All causes of death (A00-Y89) excluding S00-T98
1253797.0       Neoplasms
1236278.0       Diseases of the circulatory system (I00-I99)
1201612.0       Malignant neoplasms (C00-C97)
361927.0        Other heart diseases (I30-I33, I39-I52)
```

“All causes of death (A00-Y89) excluding S00-T98” resulted in 4307271 number of deaths in both genders and stands at the top position for causing most of the deaths, followed by “Neoplasms” which caused 1253797 number of deaths.

Top 5 causes of death in “Males”:

select sum(value) as sum_val, cause

from causeofdeath

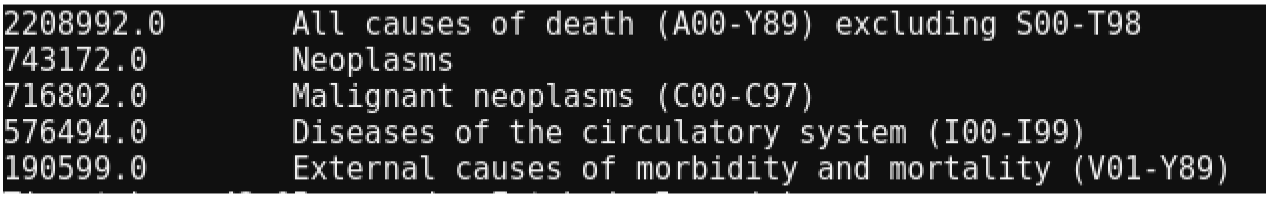
where sex='Males'

group by cause

order by sum_val desc

limit 5;

Output:



```
2208992.0       All causes of death (A00-Y89) excluding S00-T98
743172.0        Neoplasms
716802.0        Malignant neoplasms (C00-C97)
576494.0        Diseases of the circulatory system (I00-I99)
190599.0        External causes of morbidity and mortality (V01-Y89)
```

“All causes of death (A00-Y89) excluding S00-T98” resulted in 2208992 number of deaths for males and stands at the top position for causing most of the deaths, followed by “Neoplasms” which caused 743172 number of deaths.

Top 5 causes of death in “Females”:

select sum(value) as sum_val, cause

from causeofdeath

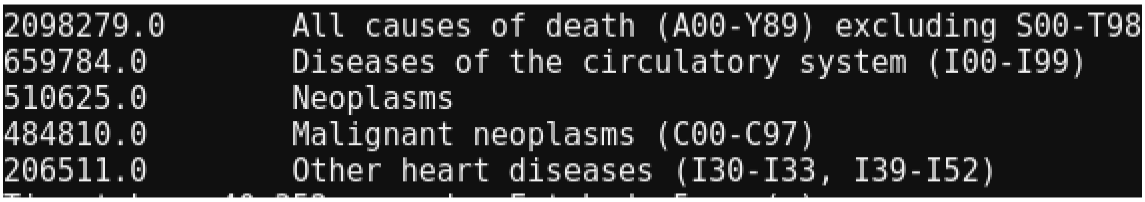
where sex='Females'

group by cause

order by sum_val desc

limit 5;

Output:



```
2098279.0       All causes of death (A00-Y89) excluding S00-T98
659784.0        Diseases of the circulatory system (I00-I99)
510625.0        Neoplasms
484810.0        Malignant neoplasms (C00-C97)
206511.0        Other heart diseases (I30-I33, I39-I52)
```

“All causes of death (A00-Y89) excluding S00-T98” resulted in 2098279 number of deaths for females and stands at the top position for causing most of the deaths, followed by “Diseases of the circulatory system (I00-I99)” which caused 659784 number of deaths.