I executed SQL queries against the 3 tables below.

Table 'dataset' has ids of films and directors from tables 'films' and 'directors'

Table of Contents:


```
import pandas as pd
dataset = pd.read_csv('dataset.csv')
films = pd.read_csv('films.csv')
directors = pd.read_csv('directors.csv')
```

```
dataset_columns = dataset.columns.to_list()
films_columns = films.columns.to_list()
directors_columns = directors.columns.to_list()

print('The table "films" has columns:', films_columns)
print('The table "directors" has columns:', directors_columns)
print('The table "dataset" has columns:', dataset_columns)
```

```
The table "films" has columns: ['film_id', 'film_title']
The table "directors" has columns: ['director_id', 'director']
The table "dataset" has columns: ['index', 'film_id', 'year', 'runtime', 'genre',
'rating', 'director_id']
```

SQL queries (various)

- Top 5 directors with the highest average rating

```
select director, round(avg(rating),2) as average_rating from dataset
left join directors using(director_id)
group by director
order by average_rating desc
limit 5;
```

index	director	average_rating
0	Christopher Nolan	8.46
1	Peter Jackson	8.4
2	Stanley Kubrick	8.23
3	Akira Kurosawa	8.22
4	Quentin Tarantino	8.18

5 rows x 3 columns

- Films directed by Akira Kurosawa or Martin Scorsese and released between 1980 and 1990

```
select film_title, director, year from dataset
left join films using(film_id)
left join directors using(director_id)
where year between 1980 and 1990 and director in ('Akira Kurosawa','Martin
Scorsese');
```

index	film_title	director	year
0	Ran	Akira Kurosawa	1985
1	Kagemusha	Akira Kurosawa	1980
2	Goodfellas	Martin Scorsese	1990
3	Raging Bull	Martin Scorsese	1980
4	The King of Comedy	Martin Scorsese	1982
5	After Hours	Martin Scorsese	1985

6 rows x 4 columns

- Unique directors

```
select distinct director from directors;
```

index	director
0	Akira Kurosawa
1	Alexander Mackendrick
2	Alexander Payne
3	Alfonso Cuarón
4	Alfred Hitchcock
5	Andrei Tarkovsky
6	Bradley Cooper
7	Brian De Palma
8	Bryan Singer
9	Christopher Miller

Page **1 of 5** | Go to page: 1 Show 10 To return to sample view Collapse rows

- Minimum, maximum and average rating of all films in the dataset

```
select min(rating) as minimum_rating, max(rating) as maximum_rating,
round(avg(rating),1) as average_rating from dataset;
```

index	minimum_rating	maximum_rating	average_rating
0	7.6	9	8

1 rows x 4 columns

- Minimum and maximum rating for each director

```
select director, min(rating) as minimum_rating, max(rating) as maximum_rating from
dataset left join directors using(director_id)
group by director
order by minimum_rating desc;
```

index	director	minimum_rating	maximum_rating

0	Andrei Tarkovsky	8.1	8.2
1	Ingmar Bergman	8.1	8.2
2	Stuart Rosenberg	8.1	8.1
3	Akira Kurosawa	8	8.6
4	Orson Welles	8	8.3
31 rows hidden, showing first and last five			
36	Steven Spielberg	7.6	8.9
37	Terrence Malick	7.6	7.8
38	Wes Anderson	7.6	8.1
39	Woody Allen	7.6	8
40	Guy Ritchie	7.6	8.3

41 rows x 4 columns

- 3 random films

```
select film_title from films
order by random()
limit 3;
```

index	film_title
0	The Lady Vanishes
1	Höstsonaten
2	The Lego Movie

3 rows x 2 columns

- Film description including film title, year and director

```
select film_title, concat('"',film_title,'"',' ','was released in',' ', year,'
','and',' ','directed by',
               ' ', director) as film_description
from dataset
left join films using(film_id)
left join directors using(director_id);
```

index	film_title	film_description
0	Shichinin no samurai	"Shichinin no samurai" was released in 1954 and directed by Akira Kurosawa
1	Tengoku to jigoku	"Tengoku to jigoku" was released in 1963 and directed by Akira Kurosawa
2	Ikiru	"Ikiru" was released in 1952 and directed by Akira Kurosawa
3	Ran	"Ran" was released in 1985 and directed by Akira Kurosawa
4	Yôjinbô	"Yôjinbô" was released in 1961 and directed by Akira Kurosawa
178 rows hidden, showing first and last five		
183	Catch Me If You Can	"Catch Me If You Can" was released in 2002 and directed by Steven Spielberg
184	The Curious Case of Benjamin Button	"The Curious Case of Benjamin Button" was released in 2008 and directed by David Fincher
185	The Social Network	"The Social Network" was released in 2010 and directed by David Fincher
186	The Gentlemen	"The Gentlemen" was released in 2019 and directed by Guy Ritchie

187	Casino	"Casino" was released in 1995 and directed by Martin Scorsese

188 rows x 3 columns

- Films that have a rating which is higher than the average rating of all the films in the dataset

Subqueries

```
select film_title, rating, (select round(avg(rating),2) from dataset) as
average_rating from dataset inner join films using (film_id)
where rating > average_rating;
```

index	film_title	rating	average_rating
0	Shichinin no samurai	8.6	8.03
1	Tengoku to jigoku	8.4	8.03
2	Ikiru	8.3	8.03
3	Ran	8.2	8.03
4	Yôjinbô	8.2	8.03
72 rows hidden, showing first and last five			
77	Persona	8.1	8.03
78	Trois couleurs: Rouge	8.1	8.03
79	Casino	8.2	8.03
80	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	8.4	8.03
81	Catch Me If You Can	8.1	8.03

82 rows x 4 columns

- Number of unique directors
- Number of films
- Average number of films per director
- Maximum number of films per director
- Minimum number of films per director

```
select
count(distinct(director_id)) as num_of_dir,
count(film_id) as num_of_films,
count(film_id)/count(distinct(director_id)) as avg_films_per_dir,
(select max(counts) from (select count(film_id) as counts from dataset
 group by director_id)) as max_film_per_dir,

 (select min(counts) from (select count(film_id) as counts from dataset
 group by director_id)) as min_film_per_dir

from dataset;
```

index	num_of_dir	num_of_films	avg_films_per_dir	max_film_per_dir	min_film_per_dir
0	41	188	4	14	1

1 rows x 6 columns

- The titles of the drama films with the longest runtime

```
select film_title, runtime, genre from dataset
left join films using(film_id)
where genre like '%Drama%' and runtime in (select max(runtime) from dataset where
genre like '%Drama%');
```

index	film_title	runtime	genre
0	The Irishman	209	Biography, Crime, Drama

1 rows x 4 columns

CTE, IF, CASE WHEN

- Split films into groups using the length. Find the numberf films in each group.

```
with t1 (film_title, runtime,length) as
(select film_title, runtime, if(runtime > 120, 'long', if(runtime<60, 'short',
'average')) as length from dataset
left join films using (film_id))

select count(film_title) as amount_of_films, length from t1
group by length

;
```

index	amount_of_films	length
0	100	long
1	88	average

2 rows x 3 columns

- Create a column defining a century in which a film was released

```
select film_title, year, case
    when year>1999 THEN '21 century'
    else '20 century'
end as century
from dataset left join films using(film_id);
```

index	film_title	year	century
0	Shichinin no samurai	1954	20 century
1	Tengoku to jigoku	1963	20 century
2	Ikiru	1952	20 century
3	Ran	1985	20 century
4	Yôjinbô	1961	20 century
178 rows hidden, showing first and last five			
183	Repulsion	1965	20 century
184	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	1964	20 century
185	Catch Me If You Can	2002	21 century
186	The Curious Case of Benjamin Button	2008	21 century
187	The Social Network	2010	21 century

188 rows x 4 columns

- Directors who directed 5 and more films released 21st century

```
with t1 (film_title, year, runtime, genre, rating, director) as
(select film_title, year, runtime, genre, rating, director from dataset
 left join films using(film_id)
 left join directors using(director_id)
 where year >= 2000)

 select director, count(film_title) as num_of_films from t1
 group by director
 having count(film_title) >=5
 order by num_of_films desc
 ;
```

index	director	num_of_films
0	Christopher Nolan	8
1	Quentin Tarantino	6
2	Alfonso Cuarón	5
3	Clint Eastwood	5
4	David Fincher	5
5	Denis Villeneuve	5
6	Peter Jackson	5
7	Wes Anderson	5

8 rows x 3 columns

Window Functions

- Average films runtime per director

```
select distinct director,
       round(avg(runtime) OVER (partition by director)) AS avg_time_per_dir
from dataset left join directors using(director_id)
order by avg_time_per_dir ;
```

index	director	avg_time_per_dir
0	Alexander Mackendrick	94
1	Krzysztof Kieslowski	96
2	Woody Allen	96
3	Wes Anderson	97
4	Christopher Miller	100
31 rows hidden, showing first and last five		
36	Quentin Tarantino	144
37	Martin Scorsese	145
38	Andrei Tarkovsky	147
39	Lars von Trier	159
40	Peter Jackson	178

41 rows x 3 columns

- Number the films for every director by year

```
select director, film_title,
row_number() over (partition by director order by year) as rank,
year
from dataset left join directors using(director_id) left join films using(film_id);
```

index	director	film_title	rank	year
0	Akira Kurosawa	Rashômon	1	1950
1	Akira Kurosawa	Ikiru	2	1952
2	Akira Kurosawa	Shichinin no samurai	3	1954
3	Akira Kurosawa	Kumonosu-jô	4	1957
4	Akira Kurosawa	Kakushi-toride no san-akunin	5	1958
5	Akira Kurosawa	Yôjinbô	6	1961
6	Akira Kurosawa	Sanjuro	7	1962
7	Akira Kurosawa	Tengoku to jigoku	8	1963
8	Akira Kurosawa	Kagemusha	9	1980
9	Akira Kurosawa	Ran	10	1985

Page **1 of 19** | Go to page: 1 Show 10 To return to sample view Collapse rows

- Rank films by rating

```
select film_title, rating, dense_rank() over (order by rating desc) as rank
from dataset left join films using(film_id);
```

index	film_title	rating	rank
0	The Dark Knight	9	1
1	The Lord of the Rings: The Return of the King	8.9	2
2	Pulp Fiction	8.9	2
3	Schindler's List	8.9	2
4	Inception	8.8	3
178 rows hidden, showing first and last five			
183	Minority Report	7.6	15
184	Close Encounters of the Third Kind	7.6	15
185	The Thin Red Line	7.6	15
186	The Royal Tenenbaums	7.6	15
187	Match Point	7.6	15

188 rows x 4 columns

- Calculate the difference between the last and the first film of each director in the dataset. Order by descending difference.

```
with t1 (director, year, rank) as
(select director, year,
row_number() over (partition by director_id order by year) as rank
from dataset left join directors using(director_id))
select distinct director, (last_value(year) OVER (partition by director)) -
(first_value(year) OVER (partition by director)) as difference
from t1
order by difference desc
;
```

index	director	difference
0	Martin Scorsese	43
1	Steven Spielberg	40

2	Roman Polanski	37
3	Woody Allen	36
4	Akira Kurosawa	35
5	Clint Eastwood	32
6	Stanley Kubrick	31
7	Quentin Tarantino	27
8	Terrence Malick	25
9	Alfonso Cuarón	23

Page **1 of 5** | Go to page: 1 Show 10 To return to sample view Collapse rows

UNION, INTERSECT, EXCEPT

- Select information about the films which were released in 2005 and in 2011

```
select * from dataset
where year = 2011
union all
select * from dataset
where year = 2005
;
```

index	index	film_id	year	runtime	genre	rating
68	68	73	2011	158	Crime, Drama, Mystery	7.8
183	183	188	2011	96	Comedy, Fantasy, Romance	7.7
52	52	57	2005	140	Action, Adventure	8.2
96	96	101	2005	129	Drama, Romance	7.8
187	187	192	2005	124	Drama, Romance, Thriller	7.6

- Select information about the films which were released after 2011 and after 2008 (without duplicates)

```
select * from dataset
where year > 2011
union
select * from dataset
where year > 2008
order by film_id;
```

index	index	m_id	year	runtime	genre	rating	director_id
12	12	13	2013	115	Adventure, Comedy, Drama	7.7	356
15	15	16	2018	135	Drama	7.7	357
16	16	17	2013	91	Drama, Sci-Fi, Thriller	7.7	357
38	38	43	2018	136	Drama, Music, Romance	7.6	360
43	43	48	2018	134	Biography, Drama, Music	8	362
31		dden, showing first and last five					
173	173	178	2014	99	Adventure, Comedy, Crime	8.1	393
174	174	179	2018	101	Animation, Adventure, Comedy	7.9	393
175	175	180	2009	87	Animation, Adventure, Comedy	7.9	393
176	176	181	2012	94	Comedy, Drama, Romance	7.8	393
183	183	188	2011	96	Comedy, Fantasy, Romance	7.7	394

41 rows x 8 columns

- Select information about the films which were released before 2011 but after 2008

```
select * from dataset
where year < 2011
intersect
select * from dataset
where year > 2008
```

index	index	film_id	year	runtime	genre	rating
47	47	52	2010	148	Action, Adventure, Sci-Fi	8.8
71	71	76	2010	120	Biography, Drama	7.7
73	73	78	2010	131	Drama, Mystery, War	8.3
87	87	92	2009	128	Action, Adventure, Mystery	7.6
108	108	113	2010	138	Mystery, Thriller	8.2
127	127	132	2009	153	Adventure, Drama, War	8.3
175	175	180	2009	87	Animation, Adventure, Comedy	7.9

7 rows x 8 columns

- Select information about the films which were released after 2006 except films released in 2008

```
select * from dataset
where year > 2006
except
select * from dataset
where year = 2008
order by year;
```

index	index	film_id	year	runtime	genre	rating
72	72	77	2007	157	Crime, Drama, Mystery	7.7
95	95	100	2007	123	Drama, Mystery, Romance	7.8
87	87	92	2009	128	Action, Adventure, Mystery	7.6
127	127	132	2009	153	Adventure, Drama, War	8.3
175	175	180	2009	87	Animation, Adventure, Comedy	7.9
33 rows hidden,		last five				
86	86	91	2019	113	Action, Comedy, Crime	7.8
112	112	117	2019	209	Biography, Crime, Drama	7.9
117	117	122	2019	137	Comedy, Drama, Romance	7.9
132	132	137	2019	161	Comedy, Drama	7.6
165	165	170	2019	108	Comedy, Drama, War	7.9

43 rows x 8 columns