# Data management and visualization

| | |
|---|---|
| **Iniziato** | mercoledì, 9 ottobre 2024, 18:14 |
| **Stato** | Completato |
| **Terminato** | mercoledì, 9 ottobre 2024, 18:14 |
| **Tempo impiegato** | 9 secondi |
| **Valutazione** | **0,00** su un massimo di 31,00 (**0%**) |

**Domanda 1**

Risposta non data

Punteggio max.: 1,50

In MongoDB, what is the best modeling pattern to handle changes in the contract conditions of an insurance company?

- ○ (a) Schema versioning
- ○ (b) Bucket pattern
- ○ (c) Document versioning
- ○ (d) none of the other answers is correct
- ○ (e) Attribute pattern

Risposta errata.

La risposta corretta è: Document versioning

**Domanda 2**

Risposta non data

Punteggio max.: 1,50

Consider the following query in MongoDB on a collection containing reviews.

```
db.review.find({date: {$lt: new Date('2023-01-01')}, user.city: 'Turin'}, {text:1, rating:1})
```

Which of the following statements is correct?

- ○ (a) none of the other answers is correct
- ○ (b) Display the text, the rating, and the _id of each review made before the year 2023 by users from Turin
- ○ (c) Display the documents of the reviews made before 2023 by users from Turin.
- ○ (d) Display the average rating of the reviews made by users from Turin before 2023
- ○ (e) Display the rating of the reviews made by users from Turin before 2023.

Risposta errata.

La risposta corretta è: Display the text, the rating, and the _id of each review made before the year 2023 by users from Turin

**Domanda 3**

What type of data visualization is commonly associated with the representation of categorical data using standardized pictorial symbols?

_____

○ (a) Bar chart

○ (b) Line graph

○ (c) Heatmap

○ (d) Scatter plot

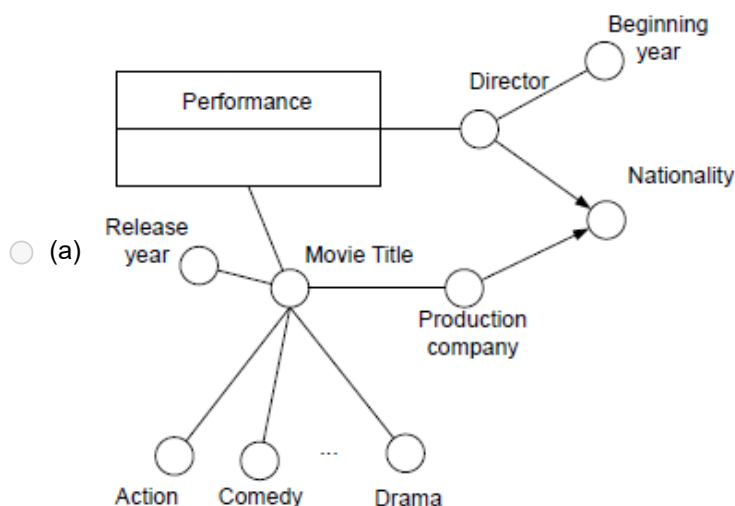○ (e) Isotype graph

---

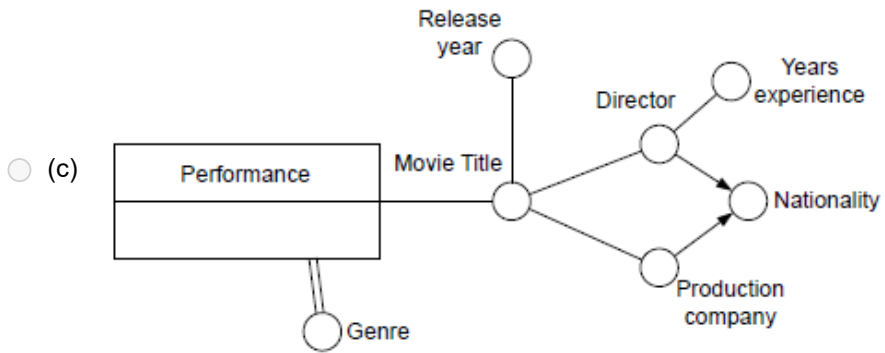Risposta errata.

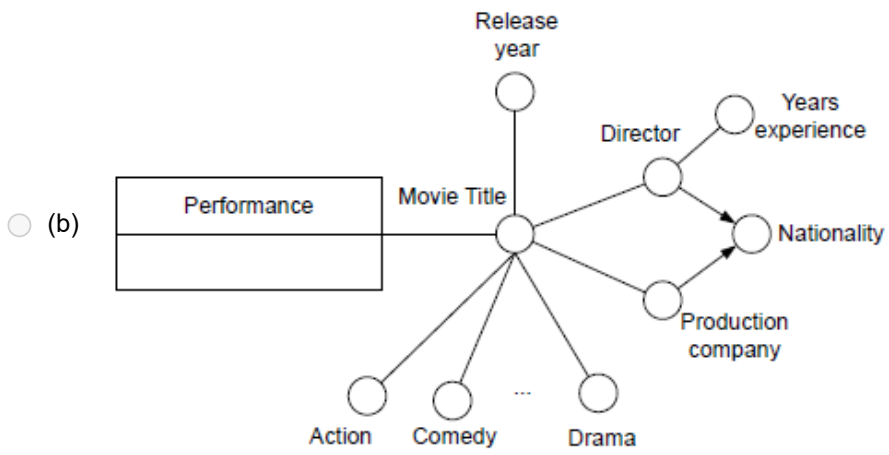La risposta corretta è: Isotype graph

Data analysts at a prominent movie production company are pursuing a mission to gain deeper insights into their film production and distribution operations. Their primary focus is on understanding the performance of movies, revenue generation, and audience engagement in terms of tickets sold at each movie projection in the cinemas. To accomplish this, they want to consider the following aspects of each ticket sold:

- Movies and Production
    - Each movie is associated with various production details such as its title (it is considered as unique), genre, release year, and director.
    - Movies can belong to one or more genres (e.g., action, comedy, drama). The list is assumed to be fixed over time and in a limited number.
    - Directors have their own unique details, nationality and years of experience in the industry. Movies have exactly one Director, while a Director can be associated with more than one movie.
    - Production companies involved in each movie, including their name and location country, must be tracked.
- Audience Engagement
    - Audience demographics should include spectators' gender and the type of the purchased ticket (children, young, regular, senior).
    - Spectators can assign a rating after each projection. This can be either a number between 0 and 5 or "None" if unavailable.
- Distribution and Screening
    - Each ticket provides access to watching a movie in a cinema.
    - The distribution of movies involves details about cinemas and screening schedules.
    - Cinemas details to be tracked are the name (unique), the number of halls, the year of construction, and their locations (city, region, and country).
    - The screening schedule consists of the day of the week, the time slots (morning, afternoon or night), and whether it falls on a holiday.
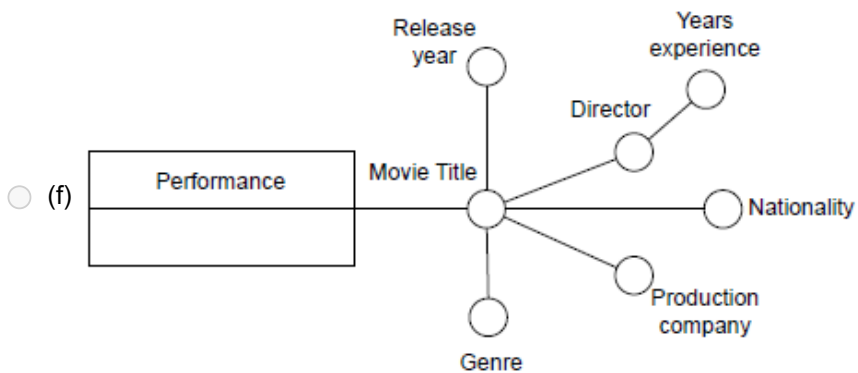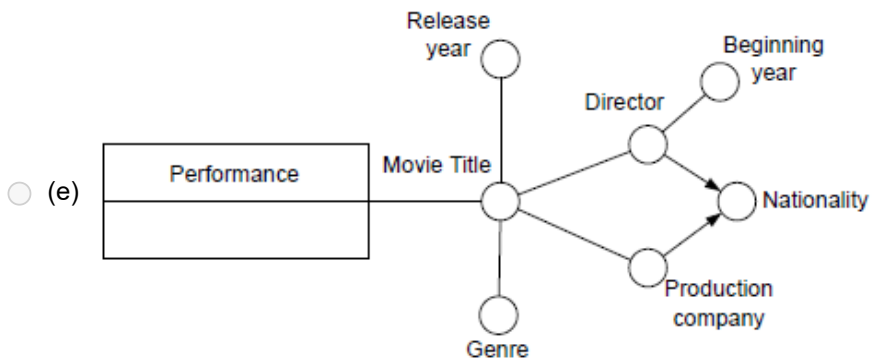
Box office performance is analyzed for various time frames: daily, weekly, monthly, quarterly, and annually.
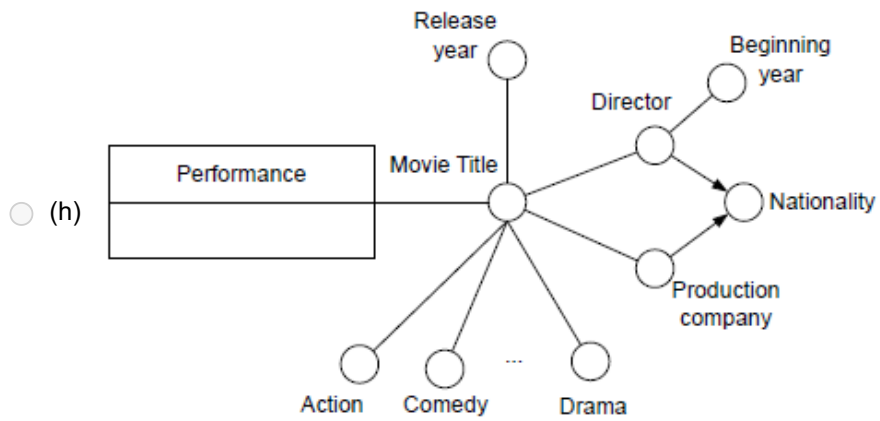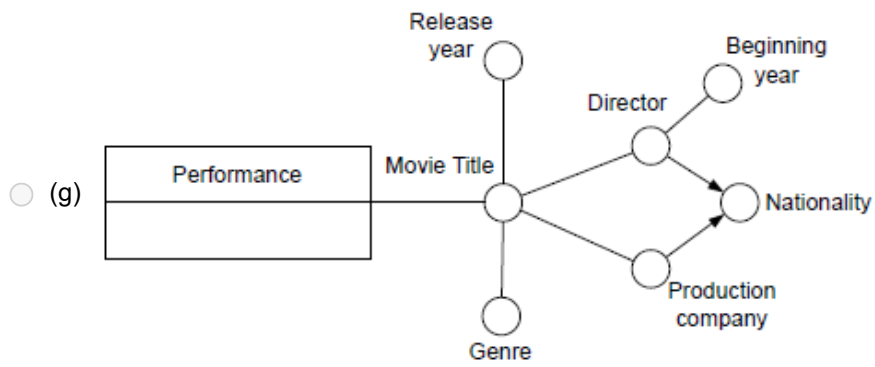
Choose the correct conceptual schema from the proposed ones to properly analyze the ticket metrics according to the given specifications (at most one answer is allowed).

(a)

(b)



(c)

(d) None of the others is correct.



(e)



(f)

○ (g)

Release year

Beginning year

Director

Performance | Movie Title

Nationality

Production company

Genre

○ (h)

Release year

Beginning year

Director

Performance | Movie Title

Nationality

Production company

Action    Comedy    ...    Drama

Risposta errata.

La risposta corretta è:

Release year

Beginning year

Director

Performance | Movie Title

Nationality

Production company

Action    Comedy    ...    Drama

Data analysts at a prominent movie production company are pursuing a mission to gain deeper insights into their film production and distribution operations. Their primary focus is on understanding the performance of movies, revenue generation, and audience engagement in terms of tickets sold at each movie projection in the cinemas. To accomplish this, they want to consider the following aspects of each ticket sold:

- Movies and Production
  - Each movie is associated with various production details such as its title (it is considered as unique), genre, release year, and director.
  - Movies can belong to one or more genres (e.g., action, comedy, drama). The list is assumed to be fixed over time and in a limited number.
  - Directors have their own unique details, nationality and years of experience in the industry. Movies have exactly one Director, while a Director can be associated with more than one movie.
  - Production companies involved in each movie, including their name and location country, must be tracked.
- Audience Engagement
  - Audience demographics should include spectators' gender and the type of the purchased ticket (children, young, regular, senior).
  - Spectators can assign a rating after each projection. This can be either a number between 0 and 5 or "None" if unavailable.
- Distribution and Screening
  - Each ticket provides access to watching a movie in a cinema.
  - The distribution of movies involves details about cinemas and screening schedules.
  - Cinemas details to be tracked are the name (unique), the number of halls, the year of construction, and their locations (city, region, and country).
  - The screening schedule consists of the day of the week, the time slots (morning, afternoon or night), and whether it falls on a holiday.

Box office performance is analyzed for various time frames: daily, weekly, monthly, quarterly, and annually.

Choose the correct conceptual schema from the proposed ones to properly analyze the ticket metrics according to the given specifications (at most one answer is allowed).
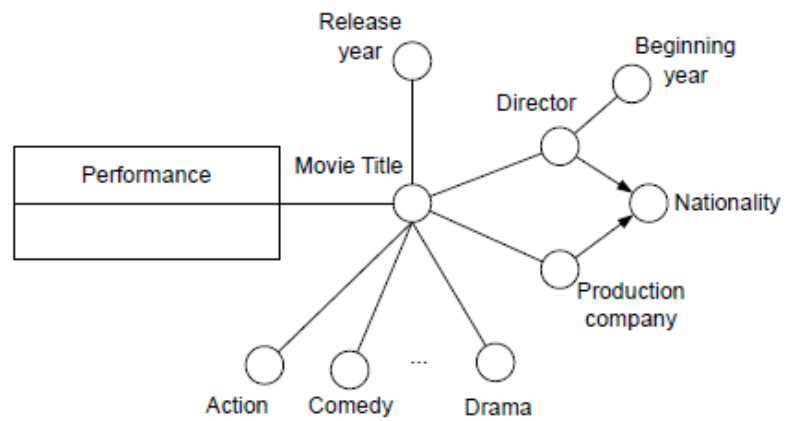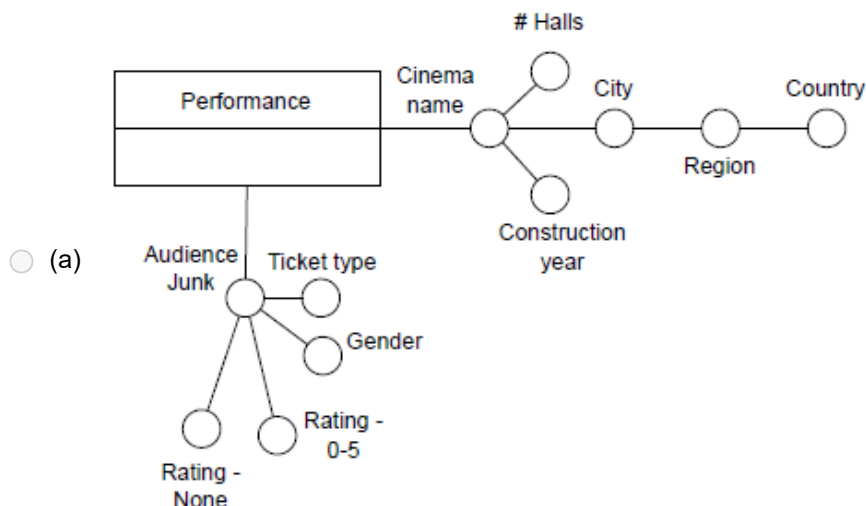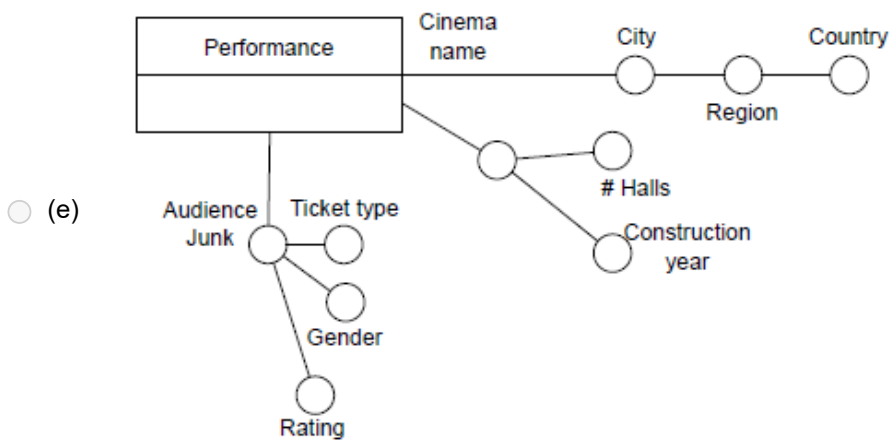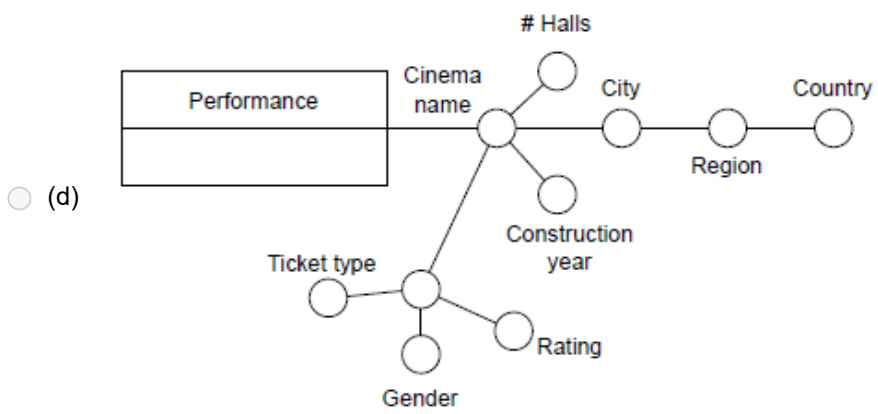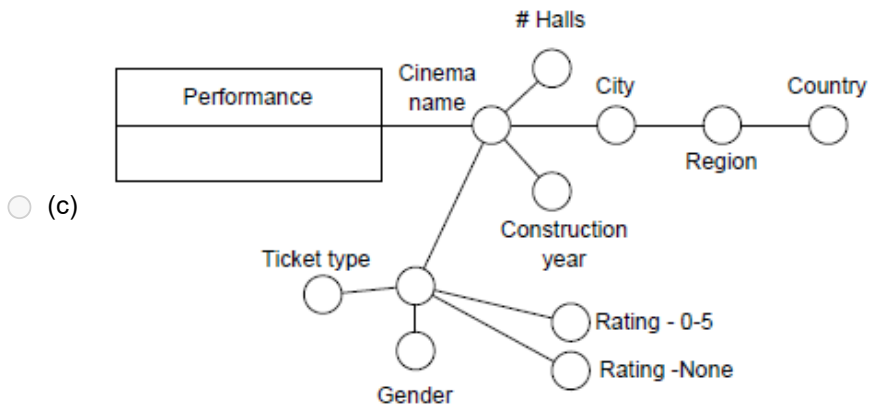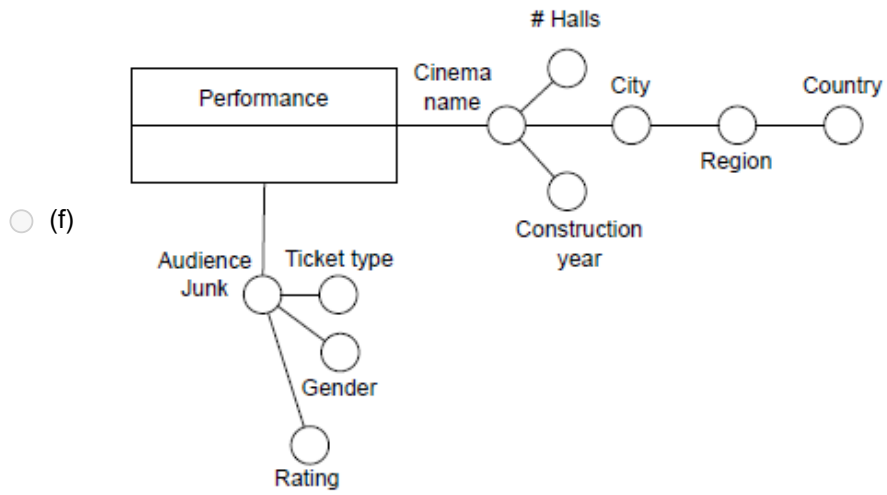


(a)

(b)



(c)



(d)



(e)

Cinema
name

City

Country

Performance

Region

(f)

Construction
year

Audience   Ticket type
Junk

Gender

Rating

(g) None of the others is correct.

# Halls

Cinema
name

City

Country

Performance

Region

(h)

Construction
year

Audience   Ticket type
Junk

Rating -
None

Gender

Rating - 0

Rating - 5

# Halls

Cinema
name

City

Country

Performance

Region

(i)

Construction
year

Rating

Gender   Ticket type

Risposta errata.

La risposta corretta è:

Data analysts at a prominent movie production company are pursuing a mission to gain deeper insights into their film production and distribution operations. Their primary focus is on understanding the performance of movies, revenue generation, and audience engagement in terms of tickets sold at each movie projection in the cinemas. To accomplish this, they want to consider the following aspects of each ticket sold:

- Movies and Production
  - Each movie is associated with various production details such as its title (it is considered as unique), genre, release year, and director.
  - Movies can belong to one or more genres (e.g., action, comedy, drama). The list is assumed to be fixed over time and in a limited number.
  - Directors have their own unique details, nationality and years of experience in the industry. Movies have exactly one Director, while a Director can be associated with more than one movie.
  - Production companies involved in each movie, including their name and location country, must be tracked.
- Audience Engagement
  - Audience demographics should include spectators' gender and the type of the purchased ticket (children, young, regular, senior).
  - Spectators can assign a rating after each projection. This can be either a number between 0 and 5 or "None" if unavailable.
- Distribution and Screening
  - Each ticket provides access to watching a movie in a cinema.
  - The distribution of movies involves details about cinemas and screening schedules.
  - Cinemas details to be tracked are the name (unique), the number of halls, the year of construction, and their locations (city, region, and country).
  - The screening schedule consists of the day of the week, the time slots (morning, afternoon or night), and whether it falls on a holiday.

Box office performance is analyzed for various time frames: daily, weekly, monthly, quarterly, and annually.
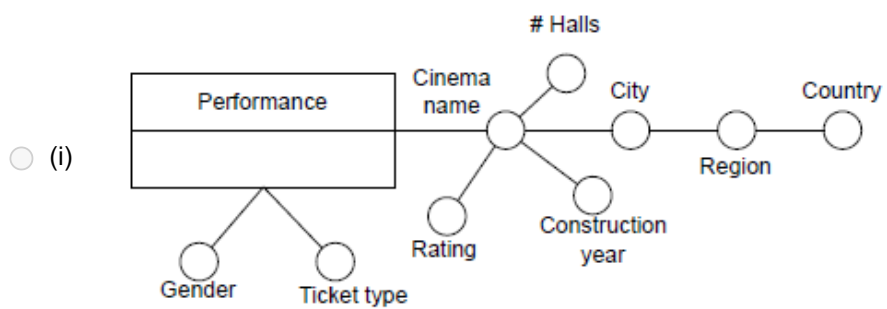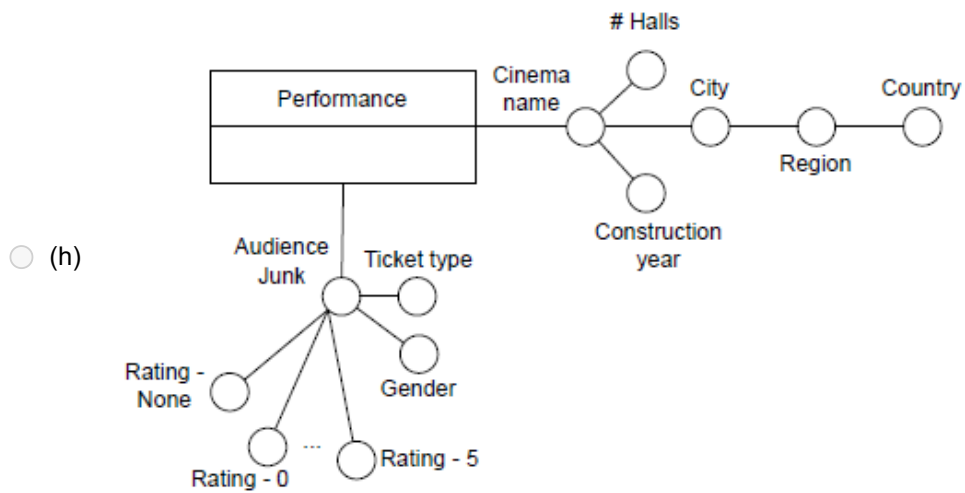
Choose the correct conceptual schema from the proposed ones to properly analyze the ticket metrics according to the given specifications (at most one answer is allowed).
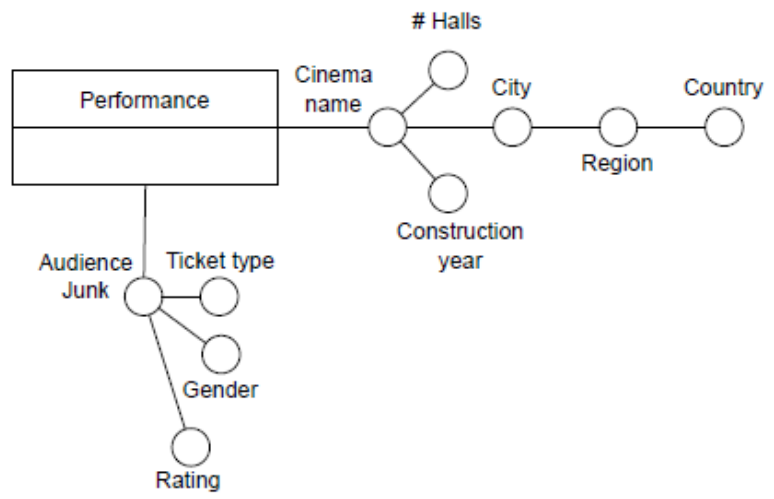


(a)



(b)

(c) Sales — Time slot, Day, Holiday (Y/N), Day of the Week, Week, Month, Quarter, Year

(d) Sales — Holiday (Y/N), Week, Day, Day of the Week, Time slot, Month, Quarter, Year

(e) Sales — Day, Holiday (Y/N), Week, Month, Day of the Week, Time slot, Year, Quarter

(f) Sales — Time slot, Day, Holiday (Y/N), Week, Day of the Week, Month, Quarter, Year

○ (g) None of the others is correct.
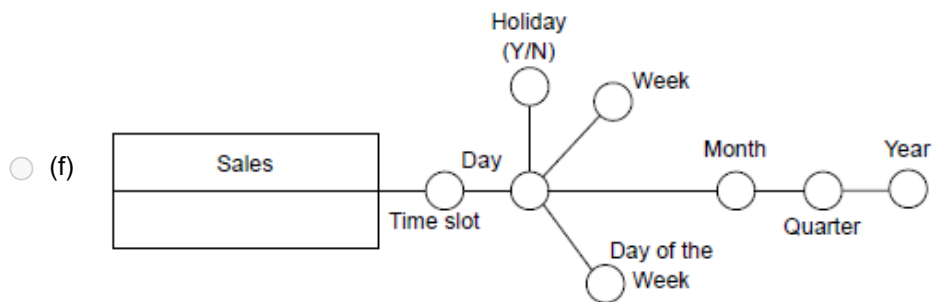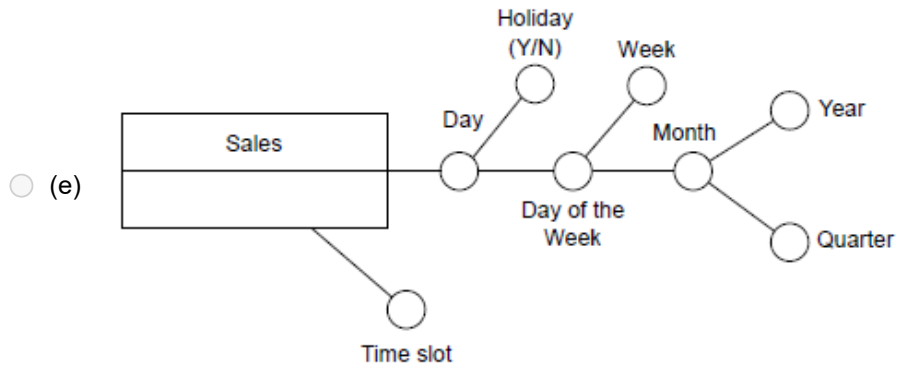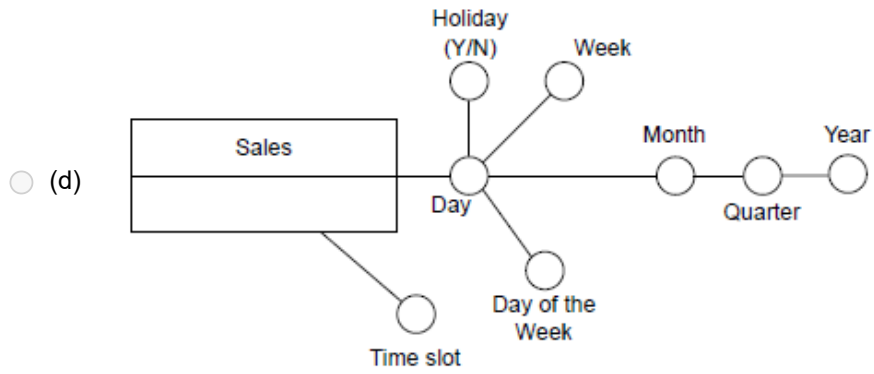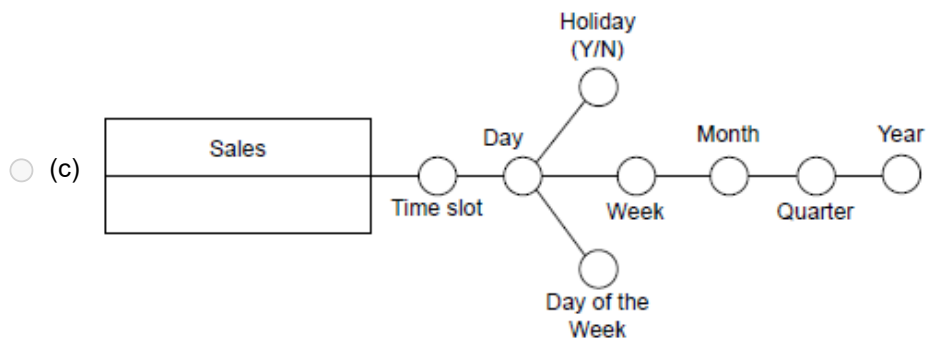
Risposta errata.

La risposta corretta è:

Holiday
(Y/N)        Week

Sales

Month        Year

Day

Quarter

Day of the
Week

Time slot

Data analysts at a prominent movie production company are pursuing a mission to gain deeper insights into their film production and distribution operations. Their primary focus is on understanding the performance of movies, revenue generation, and audience engagement in terms of tickets sold at each movie projection in the cinemas. To accomplish this, they want to consider the following aspects of each ticket sold:

- Movies and Production
  - Each movie is associated with various production details such as its title (it is considered as unique), genre, release year, and director.
  - Movies can belong to one or more genres (e.g., action, comedy, drama). The list is assumed to be fixed over time and in a limited number.
  - Directors have their own unique details, nationality and years of experience in the industry. Movies have exactly one Director, while a Director can be associated with more than one movie.
  - Production companies involved in each movie, including their name and location country, must be tracked.
- Audience Engagement
  - Audience demographics should include spectators' gender and the type of the purchased ticket (children, young, regular, senior).
  - Spectators can assign a rating after each projection. This can be either a number between 0 and 5 or "None" if unavailable.
- Distribution and Screening
  - Each ticket provides access to watching a movie in a cinema.
  - The distribution of movies involves details about cinemas and screening schedules.
  - Cinemas details to be tracked are the name (unique), the number of halls, the year of construction, and their locations (city, region, and country).
  - The screening schedule consists of the day of the week, the time slots (morning, afternoon or night), and whether it falls on a holiday.

Box office performance is analyzed for various time frames: daily, weekly, monthly, quarterly, and annually.

Select the following measures that meet the requirements described in the problem specification (more than one answer can be correct).

Scegli una o più alternative:

- (a) Total number of tickets (count)
- (b) Total umber of premium tickets (count)
- (c) Average rating (count)
- (d) Average number of theater halls per city (count)
- (e) Total number of theaters (count)
- (f) Average number of tickets (count)
- (g) Average revenue per ticket sold (euros)
- (h) Average revenue per theater region (euros)
- (i) Total number of spectators (count)
- (j) Total amount of euros per month (euros)
- (k) Average revenue for theater (euros)
- (l) Total revenue (euro)
- (m) Total number of ratings (count)
- (n) Average number of tickets per city (count)

Risposta errata.

La risposta corretta è: Total number of tickets (count), Total revenue (euro)

Given the following conceptual schema:
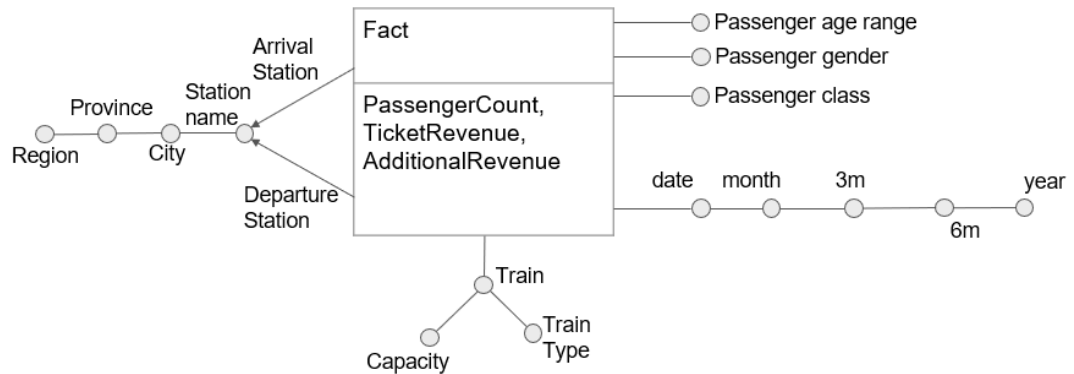


- Train capacity can be one value between "<100", "100-250", "250-500", ">500"
- Additional revenue is the revenue not associated with the ticket cost (e.g., food, services)

Write the logical design of the conceptual DW schema indicated in the picture.

Write each table on a new line.

Use the **bold** or the <u>underline</u> for identifying primary-key attributes.

---

```
Fact(DateKey, ArrivalStationKey, DepartureStationKey,
TrainKey, PassengerJunkKey, PassengerCount, TicketRevenue,
AdditionalRevenue)
Date(DateKey, Day, Month, Quarter, Semester, Year)
Station(StationKey, StationName, City, Province, Region)
Train(TrainKey, TrainName, TrainType, Capacity)
Passenger(PassengerJunkKey, AgeRange, Gender, Class)
```

```
Meetings(ServerId, OrganizerId, TimeId, PremiumCall, TotMinutes,
NumberParticipants)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek)
Organizer(OrganizerId, AgeRange, City, Region, Country)
Server(ServerId, Continent)
```

- Geographical attributes (city, region, and country) are associated with the organizer of each meeting.
- Premium calls can be either 0 or 1

Separately for each **server** and **bimester**, select:

A. the average number of participants per meeting

B. the percentage of the total call minutes of each server with respect to the call minutes of all the servers of the same continent

C. assign a rank to each server separately for continent and bimester, based on its total call minutes (rank 1st the server with the highest total call minutes for each continent)

---

```
SELECT ServerId, 2M,
  SUM(NumberParticipants) / COUNT(*) as A,
  100 * SUM(TotMinutes) / SUM(SUM(TotMinutes)) OVER (PARTITION BY
S.Continent, T.2M) as B,
  RANK() OVER (PARTITION BY S.Continent, T.2M,
                   ORDER BY SUM(TotMinutes) DESC) as C
FROM Meetings M, Server S, Time T
WHERE S.ServerId =M.ServerId and T.TimeId=M.TimeId
GROUP BY S.ServerId, T.2M, S.Continent
```

```
Meetings(ServerId, OrganizerId, TimeId, PremiumCall, TotMinutes,
NumberParticipants)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek)
Organizer(OrganizerId, AgeRange, City, Region, Country)
Server(ServerId, Continent)
```

- Geographical attributes (city, region, and country) are associated with the organizer of each meeting.
- Premium calls can be either 0 or 1

Separately for each **organizer region** and each **month**, select:

    A. the cumulative number of participants since the beginning of the year

    B. the daily average number of participants in the meetings

    C. The percentage of the monthly number of participants with respect to the monthly participants of all the regions of the same country

```
SELECT O.Region, month,
  SUM(SUM(NumberParticipants)) OVER (
      PARTITION BY O.Region, year
      ORDER BY month
      ROWS UNBOUNDED PRECEDING) as A,
         SUM(NumberParticipants) / COUNT(distinct date) as B
  100 * SUM(NumberParticipants) / SUM(SUM(NumberParticipants))
      OVER (PARTITION BY O.Country, T.month) as C,
FROM Meetings M, Organizer O, Time T
WHERE O.UserId=M.UserId and T.TimeId=M.TimeId
GROUP BY T.month, O.Region, O.Country, year
```

The following document structure represents a movie, with its title, categories, the release dates in the different countries around the world, the actors with their name, surname, and nationality, and the awards won, with the year, the award title, and the type of award. Finally, the IMDB score is also recorded.

```
{
    "_id" : ObjectId("61fa5b8f6f631bb5339dc4b7"),
    "title": "Star Wars",
    "categories": ["fantasy", "adventure"],
    "release": [
{"country": "USA",
"date": "1977-05-25"},
{"country": "Italy",
"date": "1977-10-20"}],
    "actors" : [{
    "id" : ObjectId("32af5b7a6f133ca5133dc4c8"),
    "name" : "Mark",
    "surname" : "Hamill",
"nationality": "USA"},
{
    "id" : ObjectId("32af5b7a6f133ca5133dc4c8"),
    "name" : "Harrison",
    "surname" : "Ford",
"nationality": "USA"}],
    "awards" : [
    {"year": "1978", "title": "Best Costumes", "type": "Oscar"},
  {"year": "1978", "title": "Best Movie", "type": "Oscar"}  ],
  "imdb_score": 4.8
}
```

Find all fantasy movies released in Italy after 1989 with an imdb rating greater than or equal to 4, in which at least one English actor plays, show the title and the full actor list only.

db.movie.find (
{"release": {$elemMatch: {'country: "Italy", date: {$gte: "1990-01-01"}}, "imdb.score": {$gte: 4}, actors.nationality: "UK", category: "fantasy" } , {_id:0, title:1, actors:1}
)

The following document structure represents a movie, with its title, categories, the release dates in the different countries around the world, the actors with their name, surname, and nationality, and the awards won, with the year, the award title, and the type of award. Finally, the IMDB score is also recorded.

```
{
    "_id" : ObjectId("61fa5b8f6f631bb5339dc4b7"),
    "title": "Star Wars",
    "categories": ["fantasy", "adventure"],
    "release": [
{"country": "USA",
"date": "1977-05-25"},
{"country": "Italy",
"date": "1977-10-20"}],
    "actors" : [{
    "id" : ObjectId("32af5b7a6f133ca5133dc4c8"),
    "name" : "Mark",
    "surname" : "Hamill",
"nationality": "USA"},
{
    "id" : ObjectId("32af5b7a6f133ca5133dc4c8"),
    "name" : "Harrison",
    "surname" : "Ford",
"nationality": "USA"}],
    "awards" : [
    {"year": "1978", "title": "Best Costumes", "type": "Oscar"},
  {"year": "1978", "title": "Best Movie", "type": "Oscar"}  ],
  "imdb_score": 4.8
}
```

Considering only movies released after 1999, separately for each category and type of award, calculate the total number of awards obtained and the average IMDB score. Sort the results in descending order according to the total number of awards.

```
db.movie.aggregate([
{$match: {
    "release.date":{$gte: "2000-01-01"}
},
{ $unwind: '$awards'},
{ $unwind: '$categories'},
{ $group: {
'_id': {
    prize_type: '$awards.type',
    category: '$categories'},
n: {
    $sum: 1
},
imdb: {
    $avg: '$imdb_score'}
}
{ $sort: { n: -1} }
])
```

```
db.movie.aggregate([
{$match: {
    "release.date":{$gte: "2000-01-01"}
},
{ $unwind: '$awards'},
{ $unwind: '$categories'},
{ $group: {
'_id': {
    prize_type: '$awards.type',
```

## Domanda 13

Design a MongoDB database to store the gaming sessions of video games.

Video games are characterised by a title, a description, a price in euros, and some categories (e.g., strategy, arcade, etc.).

Video game users are characterised by their first name, last name, nickname, email and some social profiles. For social profiles, we are required to track the social (e.g., Facebook, Instagram, etc.) and the user profile id (e.g., the URL).

Gaming sessions are characterised by the start timestamp, the end timestamp, the user, the video game and the level reached during the session. The level is an integer number.

To analyse the usage of video games, it is required to efficiently retrieve all the sessions of a video game for each day. Given a date and a video game, we want to display both the title of the video game and the nicknames of the users who played that video game on that day, besides all the gaming sessions. Finally, for each video game and on a given day, we want to efficiently retrieve the average duration of all the gaming sessions.

Write a sample document for each database collection.

**Important:** In addition to the example documents, explicitly state the design patterns used.

USER

```
{
_id: ObjectId(),
nickname: <string>,
name: <string>,
surname: <string>
email: <string>,
social: [{
type: <string>,
profile_id: <string>}]
}
```

GAME

```
{
_id: ObjectId(),
title: <string>,
description: <string>,
category: [<string>],
price: <number>
}
```

SESSIONS

```
{_id: ObjectId(),
game: {
_id: ObjectId(),
title: <string> },
day: <date>,
sessions: [
{starttime: <datetime>,
endtime: <datetime>,
level: <int>,
user: {
_id: ObjectId(),
nickname: <string>}
],
"tot_session_duration": <number>
"tot_number_of_sessions": <number>
}
```

**Pattern used:**

- Attribute pattern for social profiles
- Extended reference for sessions to show game information and user information.
- Bucket pattern for sessions
- Computed pattern for session stats

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |



| 50 people 7% of CO2 | 40 people 41% of CO2 | 9 people 37% of CO2 | 1 person 15% of CO2 |

https://www.oxfam.org/en/research/confronting-carbon-inequality

## Question

Which one of the following questions represents the purpose of this visualization?

○ (a) How does people's height correlate with their income and CO2 emissions?

○ (b) How does the distribution of CO2 emissions compare between different age groups?

○ (c) What percentage of people in each income class emit the most CO2?

○ (d) How do different socio-economic classes contribute to global CO2 emissions?

○ (e) What is the average CO2 emissions for each income class?

Risposta errata.

La risposta corretta è: How do different socio-economic classes contribute to global CO2 emissions?

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |
|---|---|---|---|



| 50 people 7% of CO2 | 40 people 41% of CO2 | 9 people 37% of CO2 | 1 person 15% of CO2 |
|---|---|---|---|

https://www.oxfam.org/en/research/confronting-carbon-inequality

## Data

Is the data quality appropriate? Select true answers only.

Scegli una o più alternative:

☐ (a) The lack of detailed methodology information could impact the overall credibility of the data.

☐ (b) Proper labeling explaining the relationship between income classes and CO2 emissions is missing.

☐ (c) The data might lack consistency if it doesn't account for potential variations in CO2 emissions based on factors other than income.

☐ (d) The lack of a specific data collection date raises concerns about the current relevance of the CO2 emissions values, possibly rendering the visualization obsolete.

☐ (e) The visualization does provide information about the total global CO2 emissions and the total population of each income class.

☐ (f) The percentages of CO2 emissions for each income class add up to 100%, ensuring consistent data representation.

☐ (g) The visualization might not be easy to understand for viewers without a background in CO2 emissions analysis.

☐ (h) The data includes information about the CO2 emissions for the four income classes, ensuring a complete representation of this aspect.

☐ (i) The CO2 emission percentages (7%, 41%, 37%, 15%) are presented with sufficient precision for the visualization's purpose.

☐ (j) Oxfam's reputation for producing reliable data ensures that even without a specific date, the information provided is likely up-to-date and relevant.

Risposta errata.

La risposta corretta è: The percentages of CO2 emissions for each income class add up to 100%, ensuring consistent data representation., The lack of a specific data collection date raises concerns about the current relevance of the CO2 emissions values, possibly rendering the visualization obsolete., The data includes information about the CO2 emissions for the four income classes, ensuring a complete representation of this aspect., The CO2 emission percentages (7%, 41%, 37%, 15%) are presented with sufficient precision for the visualization's purpose.

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |
|:---:|:---:|:---:|:---:|



| 50 people 7% of CO2 | 40 people 41% of CO2 | 9 people 37% of CO2 | 1 person 15% of CO2 |
|:---:|:---:|:---:|:---:|

https://www.oxfam.org/en/research/confronting-carbon-inequality

**Visual Proportionality**

Are the values encoded in a uniformly proportional way?

GLOBAL CO2 EMISSIONS IF
PEOPLE'S HEIGHT IS PROPORTIONAL
TO THEIR CO2 EMISSIONS

poorest 50%   middle 40%   middle 9%   richest 1%

50 people
7% of CO2

40 people
41% of CO2

9 people
37% of CO2

1 person
15% of CO2

https://www.oxfam.org/en/research/confronting-carbon-inequality

**Visual Utility**

All the elements in the graph convey useful information?

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |
|---|---|---|---|



| 50 people 7% of CO2 | 40 people 41% of CO2 | 9 people 37% of CO2 | 1 person 15% of CO2 |
|---|---|---|---|

https://www.oxfam.org/en/research/confronting-carbon-inequality

**Visual Clarity**

Are the data in the graph clearly identifiable and understandable (properly described)?

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |
|---|---|---|---|



| 50 people | 40 people | 9 people | 1 person |
|---|---|---|---|
| 7% of CO2 | 41% of CO2 | 37% of CO2 | 15% of CO2 |

https://www.oxfam.org/en/research/confronting-carbon-inequality
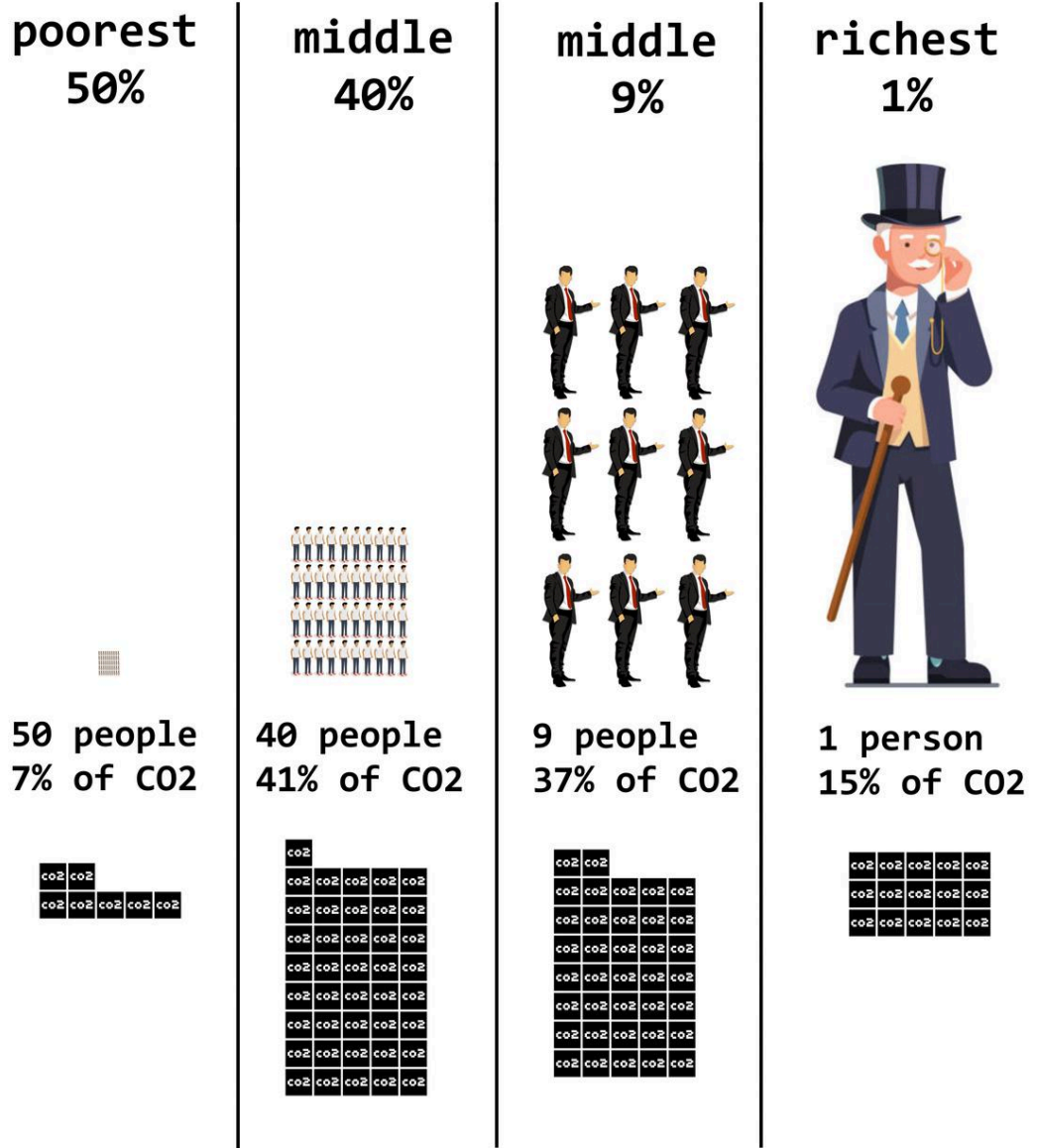
## Design data

Design the visualization based on the following data structure.

INCOME_CLASS   Scegli...

PEOPLE   Scegli...

CO2   Scegli...

Risposta errata.

# GLOBAL CO2 EMISSIONS IF PEOPLE'S HEIGHT IS PROPORTIONAL TO THEIR CO2 EMISSIONS

| poorest 50% | middle 40% | middle 9% | richest 1% |
|---|---|---|---|

50 people
7% of CO2

40 people
41% of CO2

9 people
37% of CO2

1 person
15% of CO2

https://www.oxfam.org/en/research/confronting-carbon-inequality

## Design schema & Sketch

Fill in the required schema elements; formulas can be used if required. Then describe in words the design proposal.

**Domanda 21**

Risposta non data

Non valutata

This is a blank question to be used as your personal notepad during the exam.

Anything written here will NOT be evaluated.