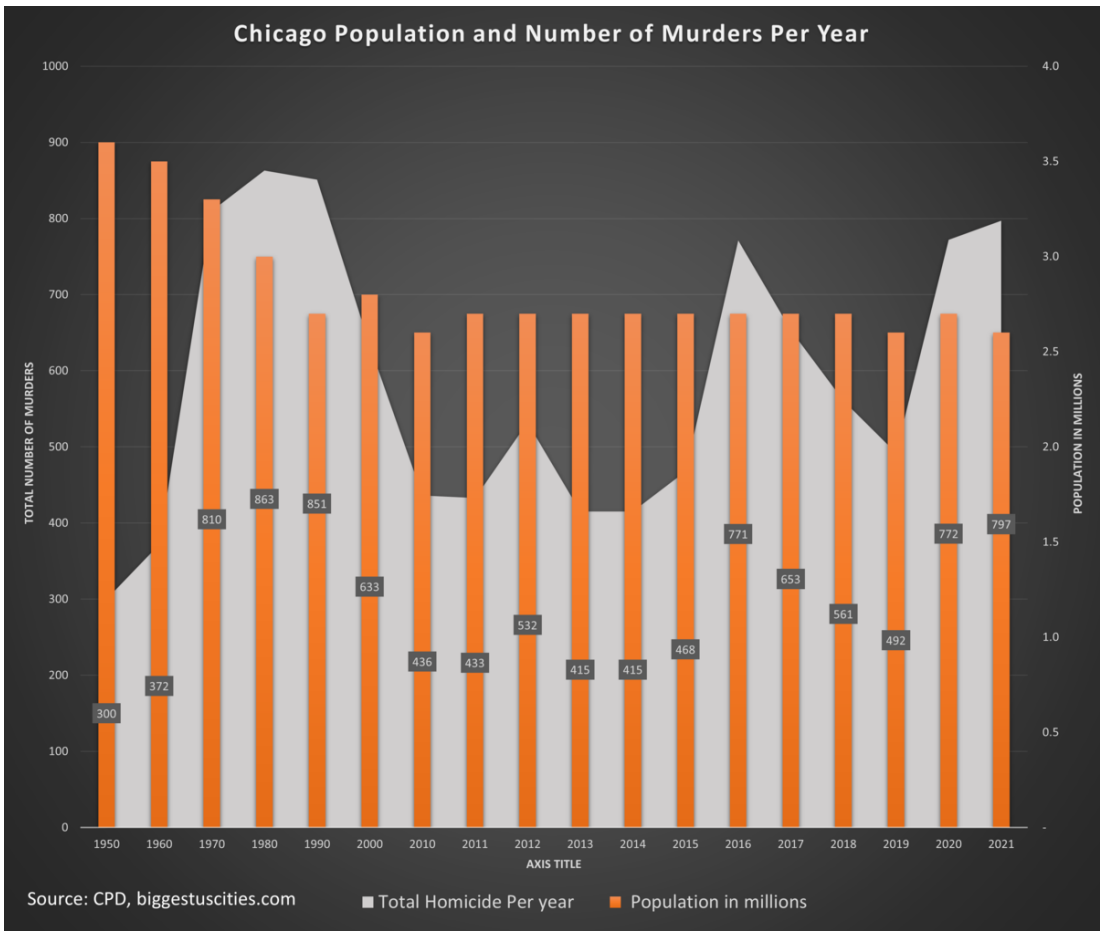
Data management and visualization

Iniziato	lunedì, 20 febbraio 2023, 16:24
Stato	Completato
Terminato	lunedì, 20 febbraio 2023, 16:24
Tempo impiegato	8 secondi
Valutazione	**0,00** su un massimo di 31,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 0,75



Visual Proportionality

Are the values encoded in a uniformly proportional way?

Domanda 2

Risposta non data

Punteggio max.: 4,00

```
Administrations(VaxID, PatientID, TimeID, ClinicID,
NumVaccinations, TotalCost
WelfareCoveredCost)
Vaccine(VaxID, BatchNumber,VaccineName, AssociatedPathology)
Patient(PatientID, AgeRange, AutoimmuneDisease, HeartDisease,
NeurologicalPathologies)
Clinic (ClinicID, ClinicName, Province, Region)
Time(TimeID, Month, 3m, 4m, 6m, year)
```

The data warehouse collects the total number and cost of vaccinations administered to patients over time. The clinic name is supposed to be unique for each clinic.

Separately for each **clinic** in the Piedmont region and for each **month**, select:

A. the average cost per vaccination
B. the ratio of total monthly cost with respect to the total cost of the year
C. the moving average of the monthly number of vaccinations over the last 5 months

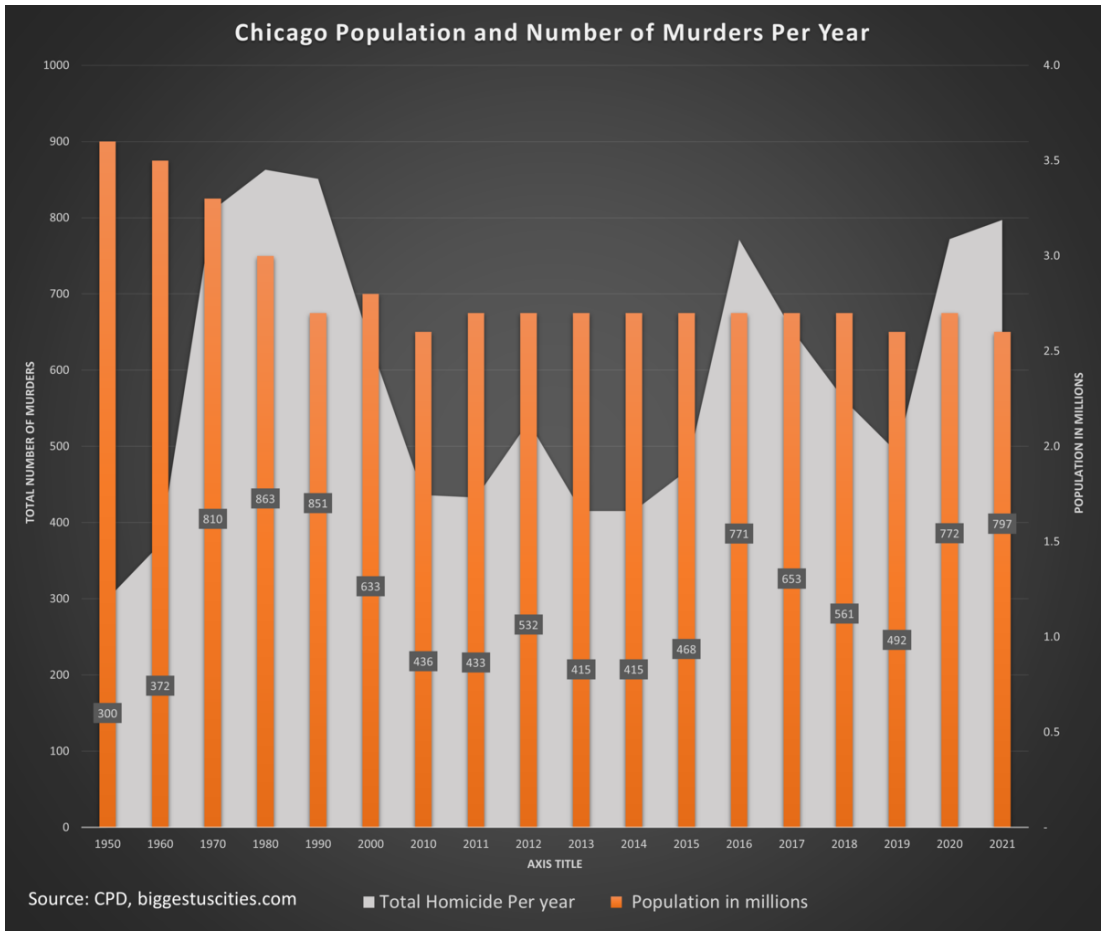
Write the requested SQL query.

```
SELECT ClinicName, month,
       SUM(TotalCost) / SUM(NumVaccinations) as A,
       SUM(TotalCost) / SUM(SUM(TotalCost)) OVER (PARTITION BY
ClinicName, year) as B
       AVG(SUM(NumVaccinations)) OVER (PARTITION BY ClinicName ORDER
BY month ROWS 4 PRECEDING) as C
//same as RANGE BETWEEN INTERVAL '4' ROWS PRECEDING and CURRENT ROW
FROM Administrations A, Clinic C, Time T
WHERE A.OperatorID=C.ClinicID and T.TimeId=A.TimeId and
Region="Piedmont"
GROUP BY ClinicName, ClinicID (optional), month, year
```

Domanda 3

Risposta non data

Punteggio max.: 0,25



Question

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

- (a) What is the trend of the Chicago population over several years?
- (b) What is the average number of murders in Chicago in the last decade?
- (c) What is the year with the highest number of murders in Chicago?
- (d) What is the number of murders per year in Chicago?
- (e) What is the correlation between the number of murders and the population of Chicago over several years?

Risposta errata.

La risposta corretta è: What is the correlation between the number of murders and the population of Chicago over several years?

Domanda 4

Risposta non data

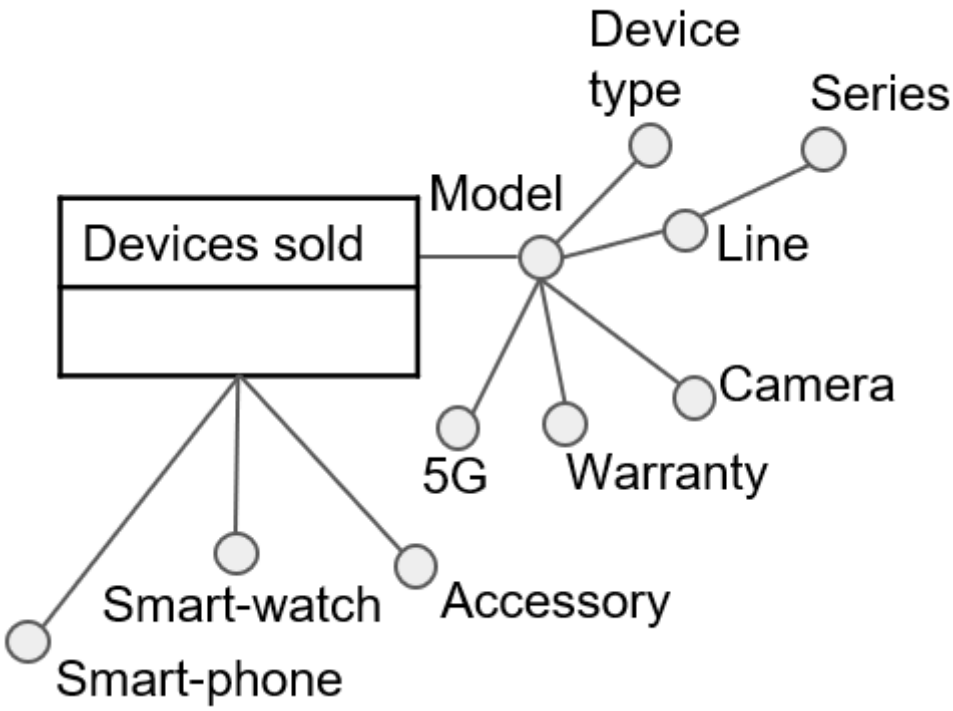
Punteggio max.: 0,50

Design a data warehouse to analyse the purchasing behaviour of consumer electronic devices manufactured by a tech company, in terms of average price per product, and the total number of products sold.

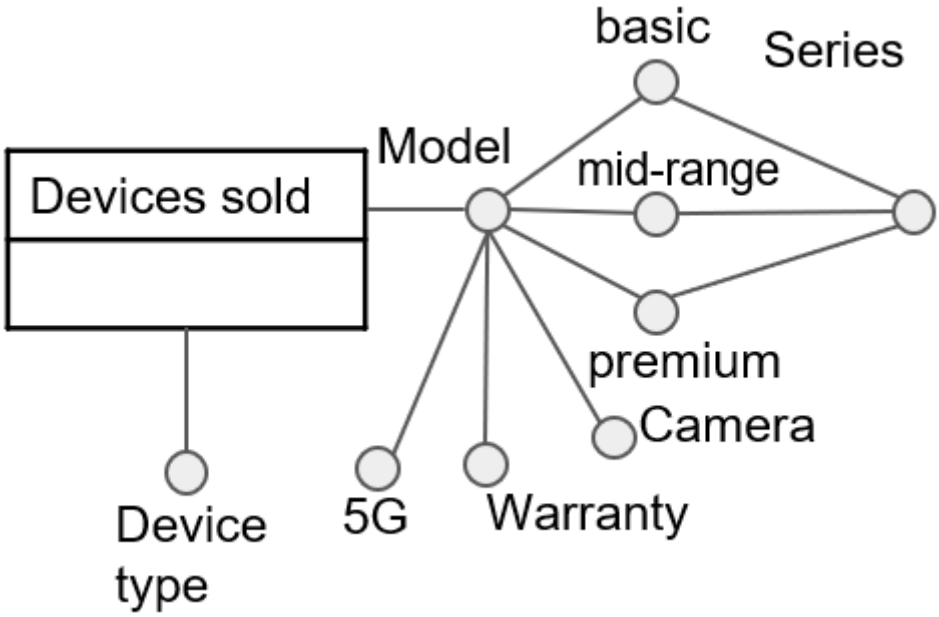
- The tech company manufactures and sells different electronic **devices**
  - All devices have a model and a series. Each model belongs to one series only. Each series collects different models. Each series belongs to one of the following three lines: “basic”, “mid-range”, and “premium”.
  - Device models can be divided into the following types: “smart-phone”, “smart-watch”, or “accessory”.
  - For each model, the following features must be recorded:
    - camera (present/absent),
    - 5G (present/absent),
    - warranty period (1, 2, or 3 years)
- The **products** sold by the tech company are manufactured in and shipped to different places around the world.
  - Both manufacturing places and shipping destinations must be analysed in terms of commercial area, continent and country.
  - A commercial area includes a group of countries, possibly belonging to different continents. Each country belongs to one commercial area only.
- The **consumer** buying the products is characterized by
  - Her/his age, in one of the following ranges: 18-30, 31-45, 45-65, or 65+ years
  - Loyalty program enrolment: yes/no
- The analysis must be performed for each **date**, day of the week, holiday (yes/no), month, 4-month period, semester, and year.

Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

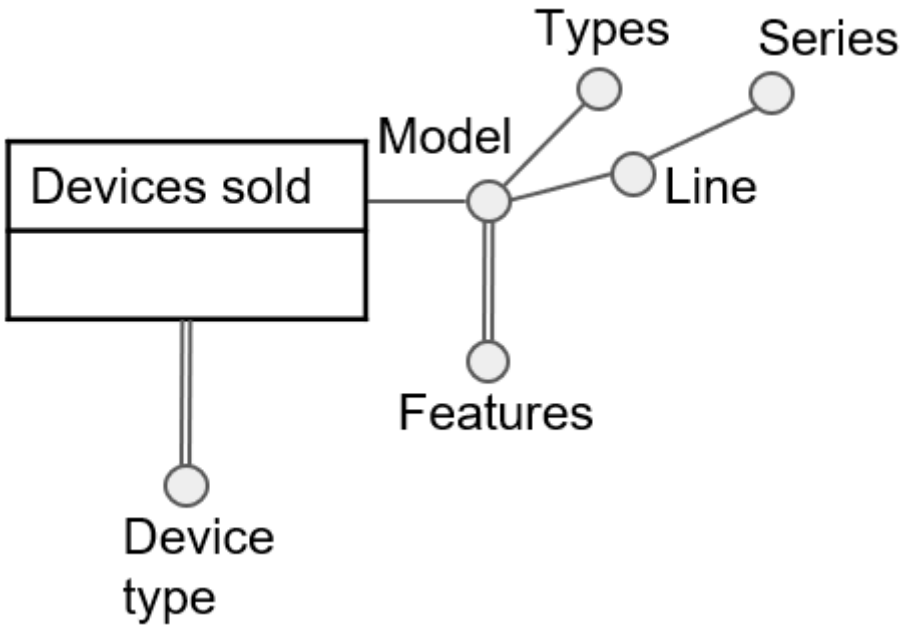
(a)



(b)

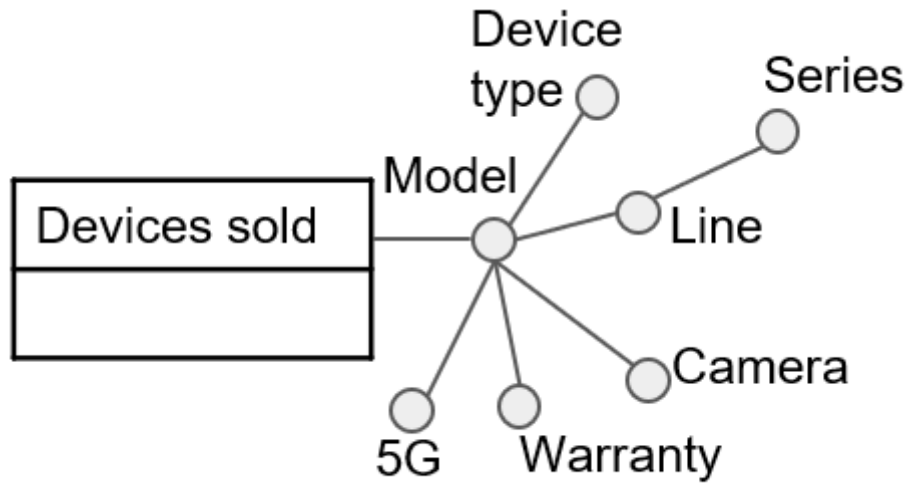


(c)

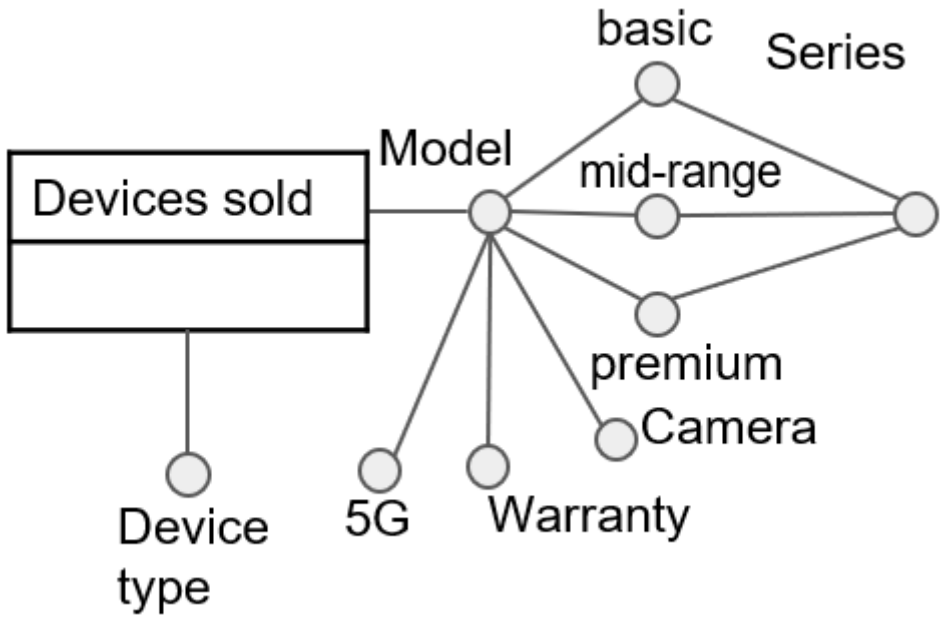


(d) None of the others is correct.

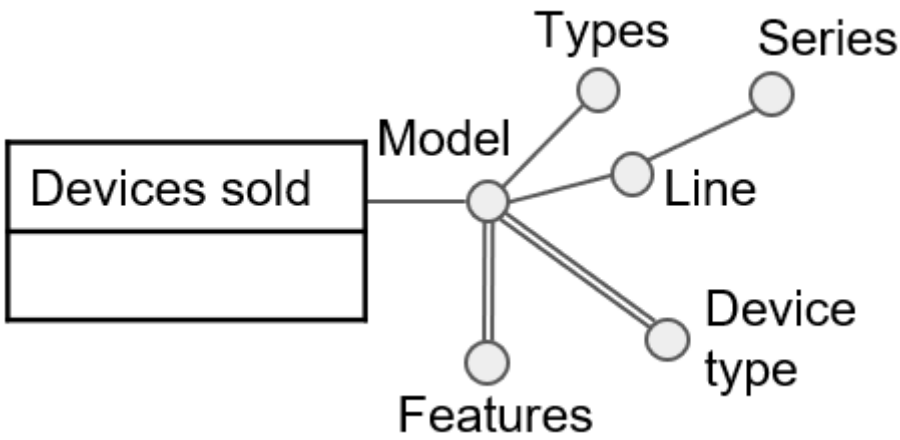
(e)


Device
type
Series
Model
Devices sold
Line
Camera
5G
Warranty

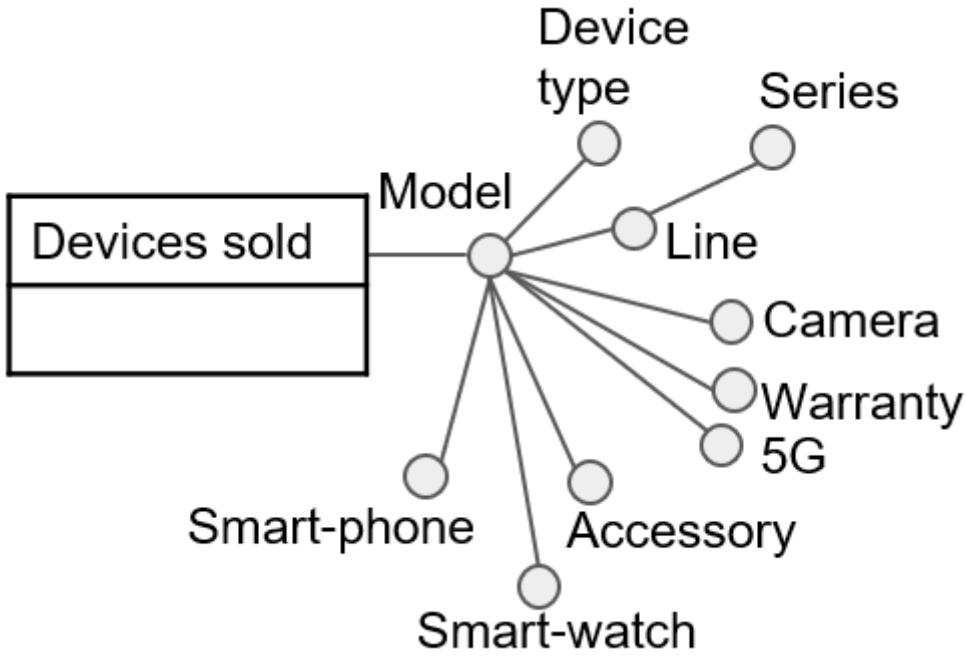
(f)


basic
Series
Model
mid-range
Devices sold
premium
Camera
Device
type
5G
Warranty

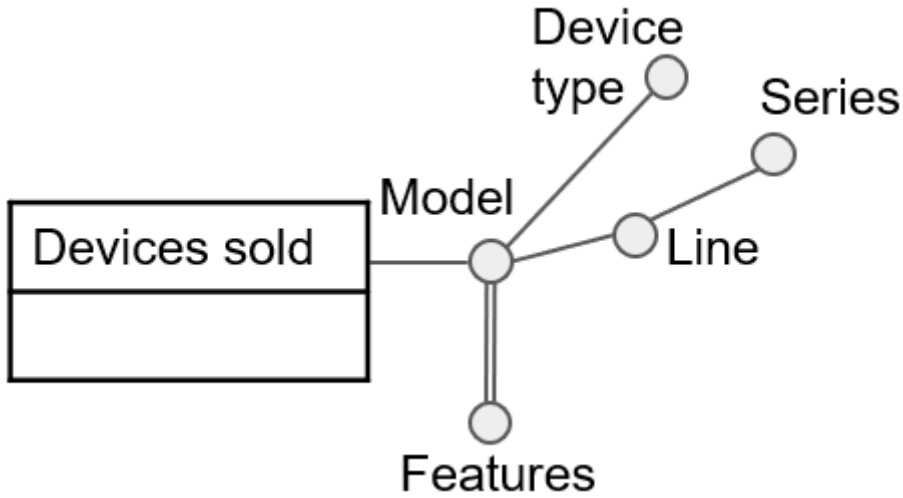
(g)


Types
Series
Model
Devices sold
Line
Device
type
Features

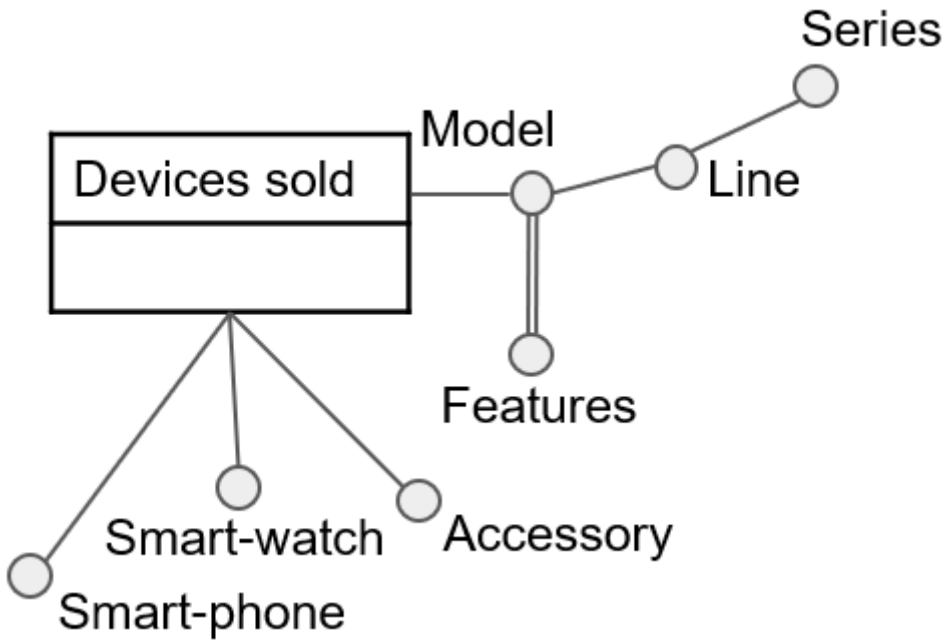
(h)


Device type
Series
Model
Devices sold
Line
Camera
Warranty
5G
Smart-phone
Accessory
Smart-watch

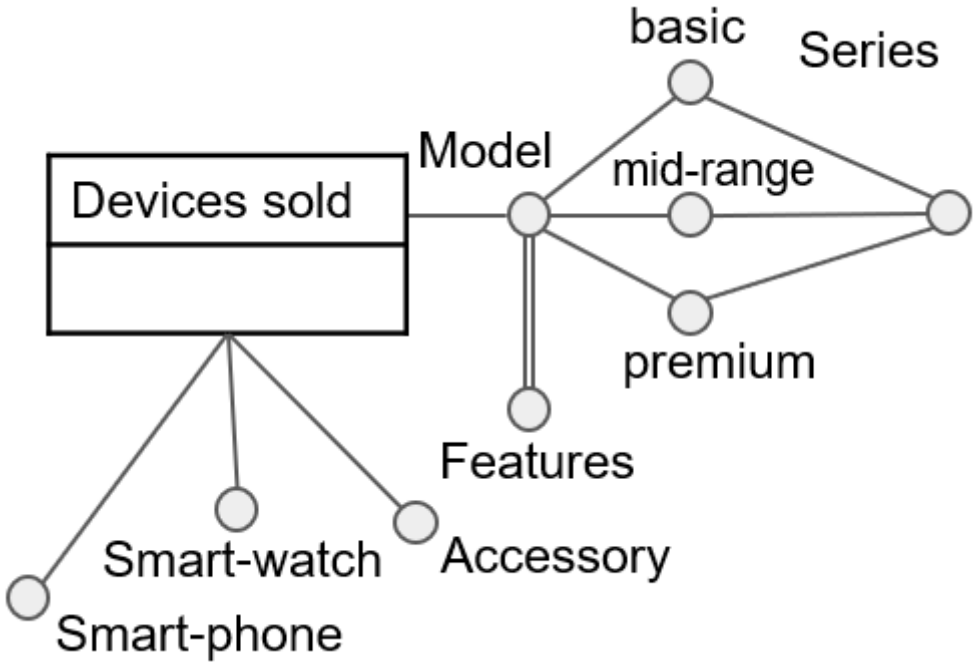
(i)


Device type
Series
Model
Devices sold
Line
Features

(j)


Series
Model
Devices sold
Line
Features
Smart-watch
Accessory
Smart-phone

(k)



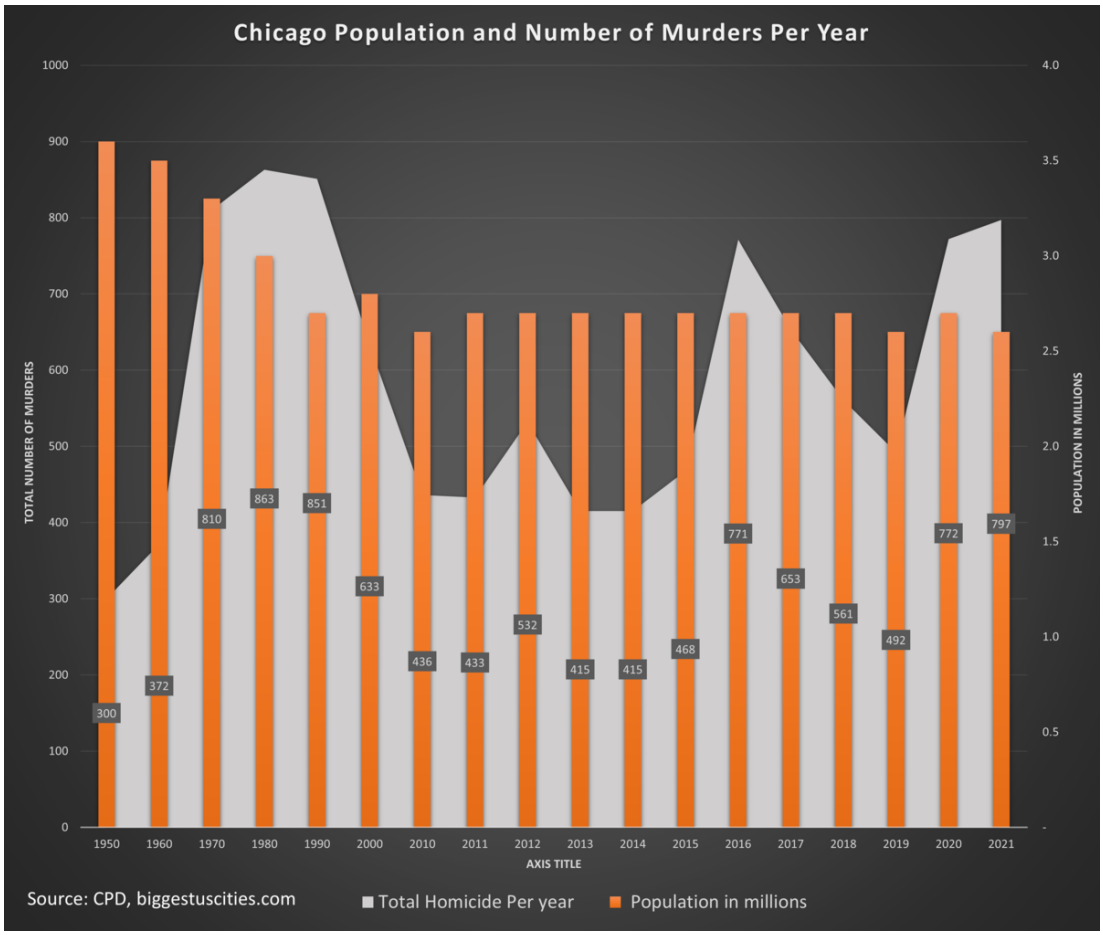
Risposta errata.

La risposta corretta è: None of the others is correct.

Domanda 5

Risposta non data

Punteggio max.: 0,50



Visual Clarity

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

Domanda 6

Risposta non data

Punteggio max.: 3,00

The following document structure represents a document of a movie. Each document collects some information about the movie directors and the movie release in each country.

```
{
_id: ObjectId(),
title: <string>,
abstract:<string>,
keywords: [ <string>, <string>, … ],
category:  [ <string>, <string>, … ],
directors: [{
        _id: ObjectId(),
          name: <string>,
          surname: <string>,
          nationality: <string>}
…
],
releases: [
          {date: <date>,
           country: <string>
          },
           …
           ],
review_score: <float>
}
```

Considering only the movies released in Italy and belonging to the category “comedy”, separately for each nationality of the director, compute the average review score and the total number of movies.

Sort the results in descending order according to the average review score.

```
db.collection_name.aggregate([
{$match: {
  “category”: “comedy”,
  “releases.country”: "Italy"
}
},
{ $unwind: ‘$directors},
{ $group: {
  '_id': ‘$directors.nationality,
  ‘avg_score’: {'$avg’: '$review_score’},
  'tot':  { '$sum: 1}
}
},
{ $sort: { avg_score: -1} }
])
```

Domanda 7

Risposta non data

Punteggio max.: 2,00

The following document structure represents a document of a movie. Each document collects some information about the movie directors and the movie release in each country.

```
{
_id: ObjectId(),
title: <string>,
abstract:<string>,
keywords: [ <string>, <string>, … ],
category:  [ <string>, <string>, … ],
directors: [{
        _id: ObjectId(),
          name: <string>,
          surname: <string>,
          nationality: <string>}
…
],
releases: [
          {date: <date>,
           country: <string>
          },
           …
           ],
review_score: <float>
}
```

Write a MongoDB query to find all the movies in the category “Fantasy” having George Lucas among the directors and with a review score higher than 4. Show only the title and the keywords.

```
db.shows.find(
  { categories: “Fantasy”,
    directors: { $elemMatch: { name: "George", surname: “Lucas” } },
    review_score: {$gt: 4}  },
  {title: 1, keywords: 1, _id:0}
)
```

Domanda 8

Risposta non data

Punteggio max.: 1,50

Which one of the following elements should **not** be removed from a visualization according to the guidelines for visual utility?

- (a) Background shades
- (b) Separated legend
- (c) Heavy grid lines
- (d) Decorative colors
- (e) Logos and other pictures

Risposta errata.

La risposta corretta è: Separated legend

Domanda 9

Risposta non data

Punteggio max.: 4,00

```
Administrations(VaxID, PatientID, TimeID, ClinicID,
NumVaccinations, TotalCost
WelfareCoveredCost)
Vaccine(VaxID, BatchNumber,VaccineName, AssociatedPathology)
Patient(PatientID, AgeRange, AutoimmuneDisease, HeartDisease,
NeurologicalPathologies)
Clinic (ClinicID, ClinicName, Province, Region)
Time(TimeID, Month, 3m, 4m, 6m, year)
```

The data warehouse collects the total number and cost of vaccinations administered to patients over time. The clinic name is supposed to be unique for each clinic.

Considering the year 2022, separately for each **vaccine name**, **associated pathology**, and **month**:

A. the total number of vaccinations
B. the percentage of monthly vaccinations with respect to all the monthly vaccines with the same associated pathology
C. assign an increasing rank to each vaccine for each month, according to its total number of monthly vaccinations

Write the requested SQL query.

```
SELECT VaccineName, month,
       SUM(NumVaccinations) as A,
       100 * SUM(NumVax) / SUM(SUM(NumVax)) OVER (PARTITION BY
month, AssociatedPathology) as B
       RANK() OVER (PARTITION BY month, AssociatedPathology ORDER BY
SUM(NumVaccinations) DESC) as C
FROM Administrations A, Vaccine V, Time T
WHERE A.VaxID=V.VaxID and T.TimeId=A.TimeId and year=2022
GROUP BY VaccineName, AssociatedPathology, month
```

Domanda 10

Risposta non data

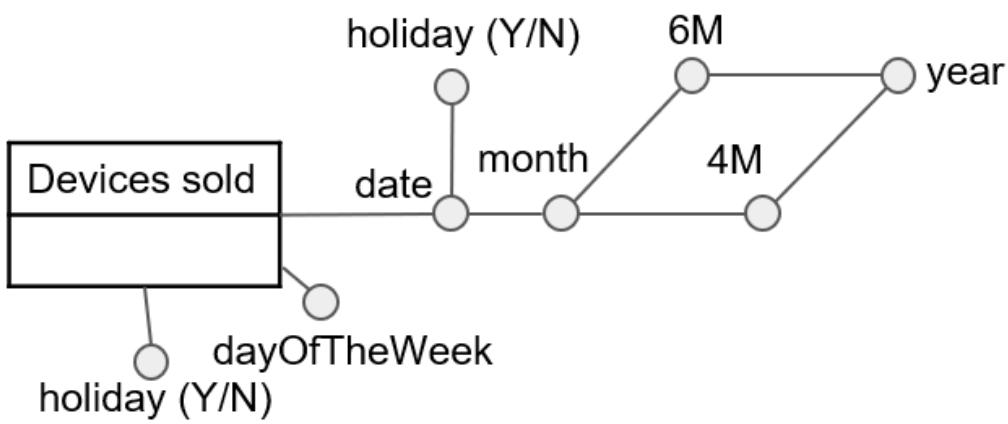
Punteggio max.: 0,50

Design a data warehouse to analyse the purchasing behaviour of consumer electronic devices manufactured by a tech company, in terms of average price per product, and the total number of products sold.

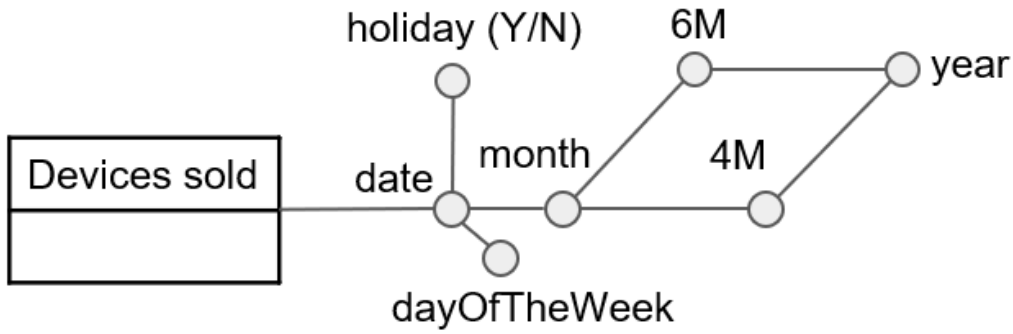
- The tech company manufactures and sells different electronic **devices**
  - All devices have a model and a series. Each model belongs to one series only. Each series collects different models. Each series belongs to one of the following three lines: “basic”, “mid-range”, and “premium”.
  - Device models can be divided into the following types: “smart-phone”, “smart-watch”, or “accessory”.
  - For each model, the following features must be recorded:
    - camera (present/absent),
    - 5G (present/absent),
    - warranty period (1, 2, or 3 years)
- The **products** sold by the tech company are manufactured in and shipped to different places around the world.
  - Both manufacturing places and shipping destinations must be analysed in terms of commercial area, continent and country.
  - A commercial area includes a group of countries, possibly belonging to different continents. Each country belongs to one commercial area only.
- The **consumer** buying the products is characterized by
  - Her/his age, in one of the following ranges: 18-30, 31-45, 45-65, or 65+ years
  - Loyalty program enrolment: yes/no
- The analysis must be performed for each **date**, day of the week, holiday (yes/no), month, 4-month period, semester, and year.

Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

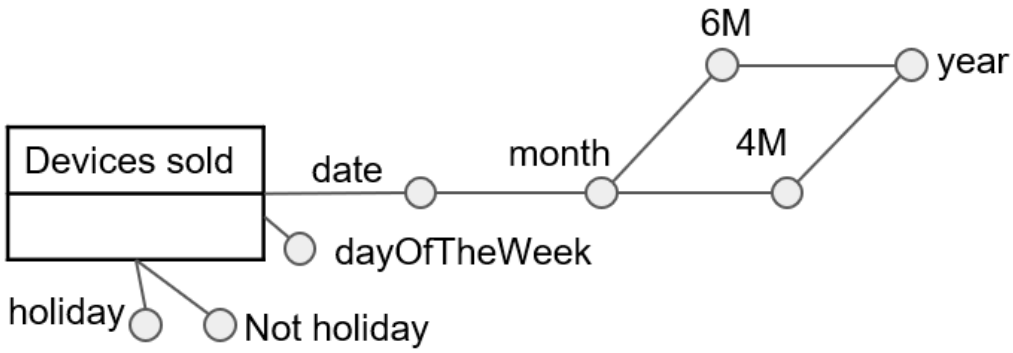
(a)



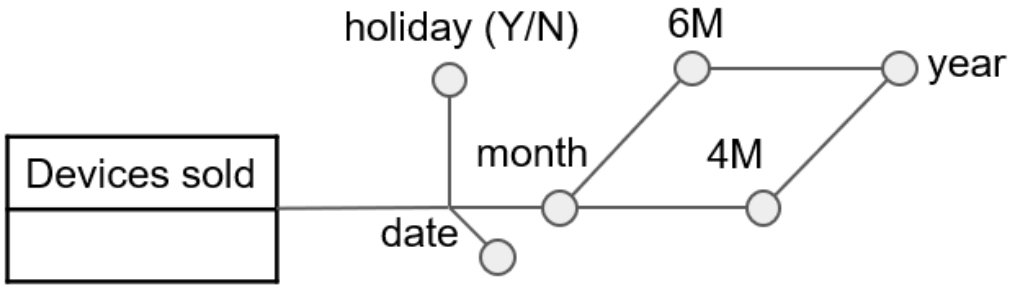
(b)



(c)

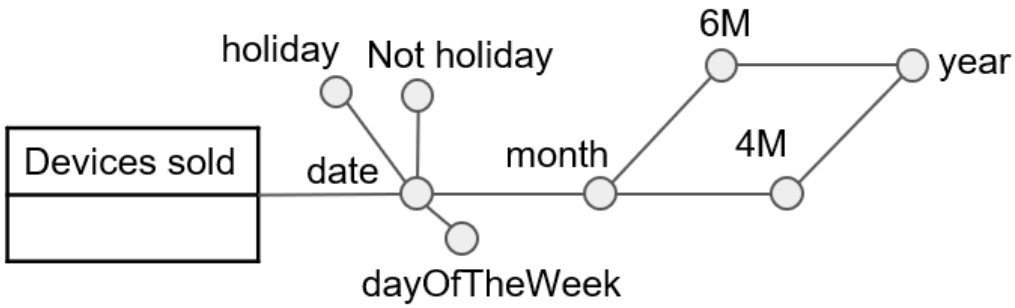


(d)



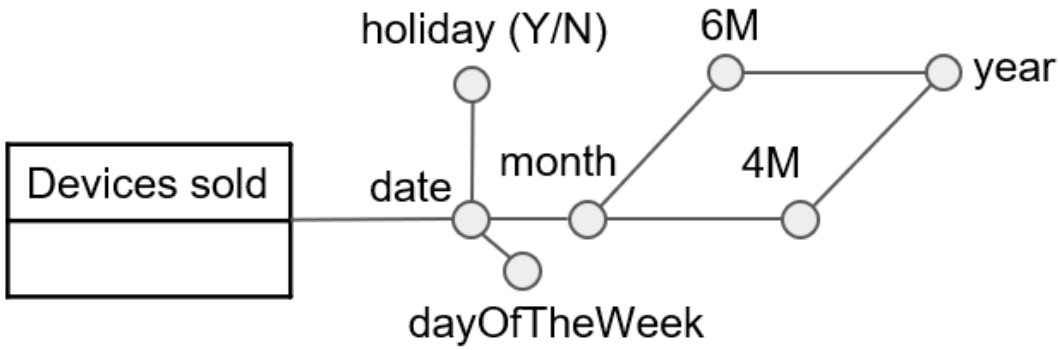
(e) None of the others is correct.

(f)



Risposta errata.

La risposta corretta è:



Domanda 11

Risposta non data

Punteggio max.: 1,50

Design a data warehouse to analyse the purchasing behaviour of consumer electronic devices manufactured by a tech company, in terms of average price per product, and the total number of products sold.

- The tech company manufactures and sells different electronic **devices**
  - All devices have a model and a series. Each model belongs to one series only. Each series collects different models. Each series belongs to one of the following three lines: “basic”, “mid-range”, and “premium”.
  - Device models can be divided into the following types: “smart-phone”, “smart-watch”, or “accessory”.
  - For each model, the following features must be recorded:
    - camera (present/absent),
    - 5G (present/absent),
    - warranty period (1, 2, or 3 years)
- The **products** sold by the tech company are manufactured in and shipped to different places around the world.
  - Both manufacturing places and shipping destinations must be analysed in terms of commercial area, continent and country.
  - A commercial area includes a group of countries, possibly belonging to different continents. Each country belongs to one commercial area only.
- The **consumer** buying the products is characterized by
  - Her/his age, in one of the following ranges: 18-30, 31-45, 45-65, or 65+ years
  - Loyalty program enrolment: yes/no
- The analysis must be performed for each **date**, day of the week, holiday (yes/no), month, 4-month period, semester, and year.

Select all and only the required measures of the fact table in the conceptual schema design among the following (multiple-choice question). Hint: do consider the dimensions defined by the previous answers.

Scegli una o più alternative:

- [ ] (a) Total income (euros)
- [ ] (b) Average number of consumers per region (count)
- [ ] (c) Total number of products (count)
- [ ] (d) Average income per factory (euros)
- [ ] (e) Total number of orders (count)
- [ ] (f) Total warranty period (time)
- [ ] (g) Total number of consumers (count)
- [ ] (h) Average number of factories per country (count)
- [ ] (i) Average number of products per type (count)

- [ ] (j) Total number of available devices (count)
- [ ] (k) Maximum number of orders (count)
- [ ] (l) Average warranty period per product (time)
- [ ] (m) Total number of factories (count)

Risposta errata.

La risposta corretta è: Total income (euros), Total number of products (count)

Domanda 12

Risposta non data

Non valutata

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

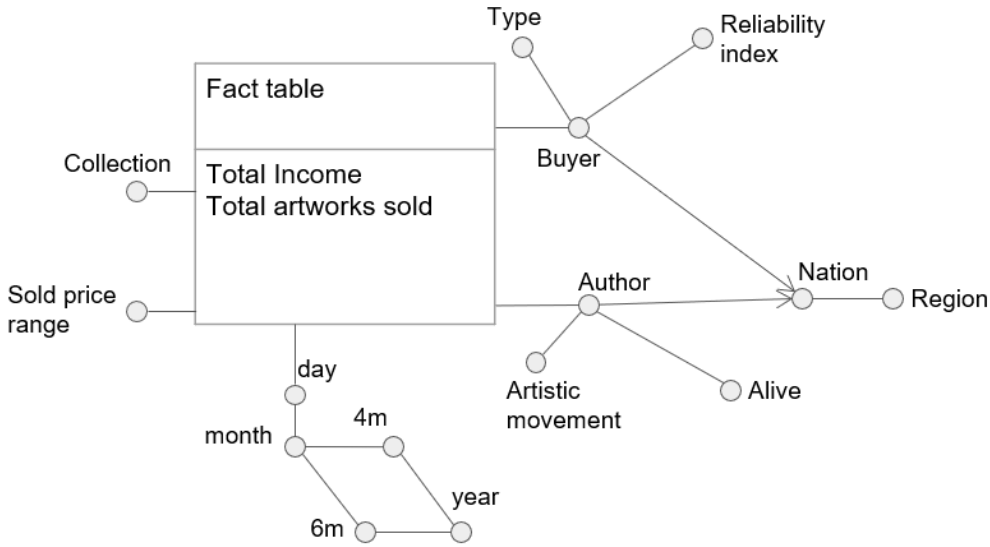
Anything written here will NOT be evaluated.

Domanda 13

Risposta non data

Punteggio max.: 1,00

Given the following conceptual schema:



- Buyers can be either private consumers, Art houses, or museum
- Reliability index is an integer number between 1 and 5
- Sold price range can be either **<10k$**, **10k$-100k$**, and **>100k$**

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 underline for identifying primary-key attributes.

```
Fact(AutID, BuyerID, TimeID, JID, TotalIncome,
TotalArtworksSold)
Junk(JID, PriceRange, Collection)
Author(AutID, ArtisticMovement, Alive, Nation, Region)
Buyer(PatientID, Type, ReliabilityIndex, Nation, Region)
Time(TimeID, Month, 4m, 6m, year)
```

`PriceRange` and `Collection` as 2 separate tables are also acceptable

Domanda 14

Risposta non data

Punteggio max.: 1,50

The schema versioning pattern has the advantage of:

- (a) none of the answers is correct
- (b) controlling the schema migration
- (c) avoiding join operations
- (d) reducing the required indexes during migration

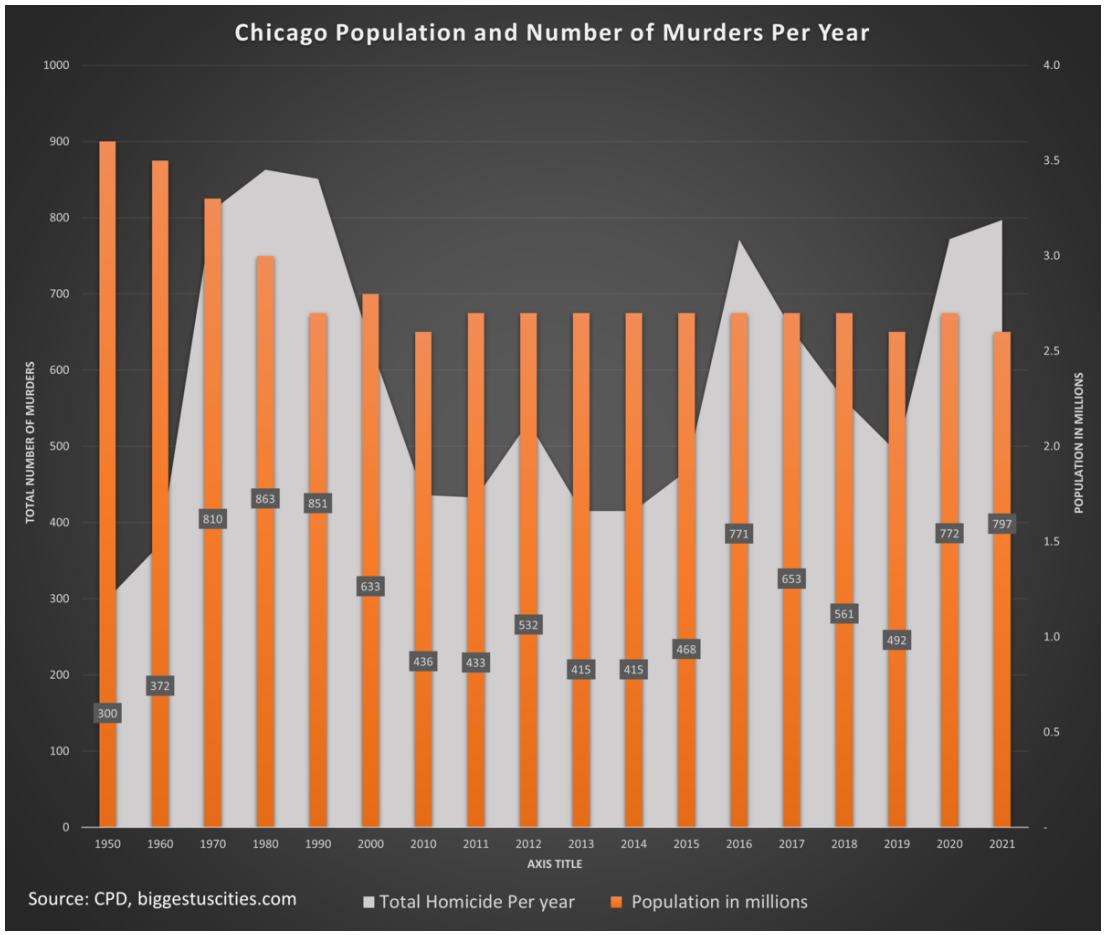
Risposta errata.

La risposta corretta è: controlling the schema migration

Domanda 15

Risposta non data

Punteggio max.: 0,25



Design data

Design the visualization based on the following data structure.

MURDERS Scegli...

POPULATION Scegli...

YEAR Scegli...

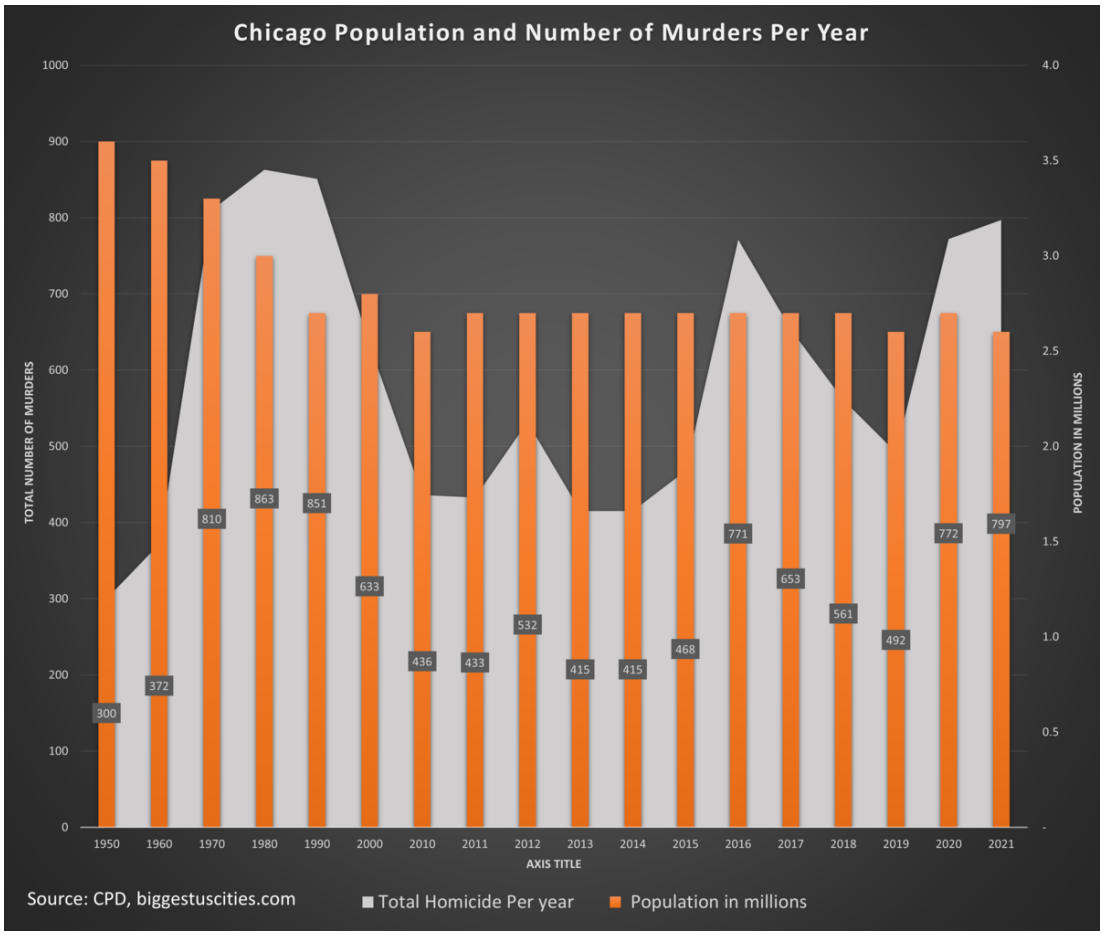
Risposta errata.

La risposta corretta è: MURDERS – Measure, POPULATION – Measure, YEAR – Dimension

Domanda 16

Risposta non data

Punteggio max.: 0,75



Visual Utility

All the elements in the graph convey useful information?

Domanda 17

Risposta non data

Punteggio max.: 2,00

The following document structure represents videos uploaded on a digital platform.

```
{"_id": ObjectID('xyz'),
"title": "D3500",
"author":{
        "name": "Alex",
        "page": "https://www.ytube.com/alex"
}
"uploadDate": Date("2022-12-14T14:03:00.000Z"),
"duration_sec": 321,
"resolutions": ["HD", "FullHD"],
"tags": ["holiday", "winter"]
}
```

You are required to extract the number of uploaded videos, separately for each tag, considering only the videos uploaded in 2022. We are only interested in showing the information for tags with at least 1000 videos. Which of the following queries satisfies the request?

(a)

```
db.videos.aggregate([
{$match: { releaseDate : {$gte: new Date ("2022"), $lte: new Date ("2023") } }
},
        {$group: {
                _id: {category: "$tags"},
                n: {$count: 1}
                }
          },
        {$match: { count: { $gte: 1000 } } }
])
```

(b)

```
db.videos.aggregate([
{$match: { releaseDate : {$gte: new Date ("2022-01-01"), $lte: new Date ("20
22-12-31") } } },
        {$group: {
                        _id: {
category: "$tags"
      },
                n: {$sum: 1}
                        }
                },
        {$match: { n: { $gte: 1000 } } }
])
```

(c)

```
db.videos.aggregate([
{$match: { releaseDate : {$gte: new Date ("2022-01-01"), $lte: new Date ("20
22-12-31") } } },
{$unwind: “$tags”},
      {$group: {
                  _id: {
category: “$tags”
    },
          n: {$sum: 1}
                }
          },
      {$match: { n: { $gte: 1000 } } }
])
```

(d) none of the other answers is correct

(e)

```
db.videos.aggregate([
{$match: { releaseDate : {$gte: new Date ("2022-01-01"), $lte: new Date ("20
22-12-31") } } },
{$unwind: “$tags”},
      {$group: {
                  _id: {
category: “$tags”
    },
          n: {$sum: 1}
                }
          }
      }
])
```

(f)

```
db.videos.aggregate([
      {$group: {
                  _id: {
category: “$tags”
    },
          n: {$sum: 1}
                }
          },
{$match: { releaseDate : {$gte: new Date ("2022-01-01"), $lte: new Date ("20
22-12-31")}, n: { $gte: 1000 }  } }
])
```

Risposta errata.

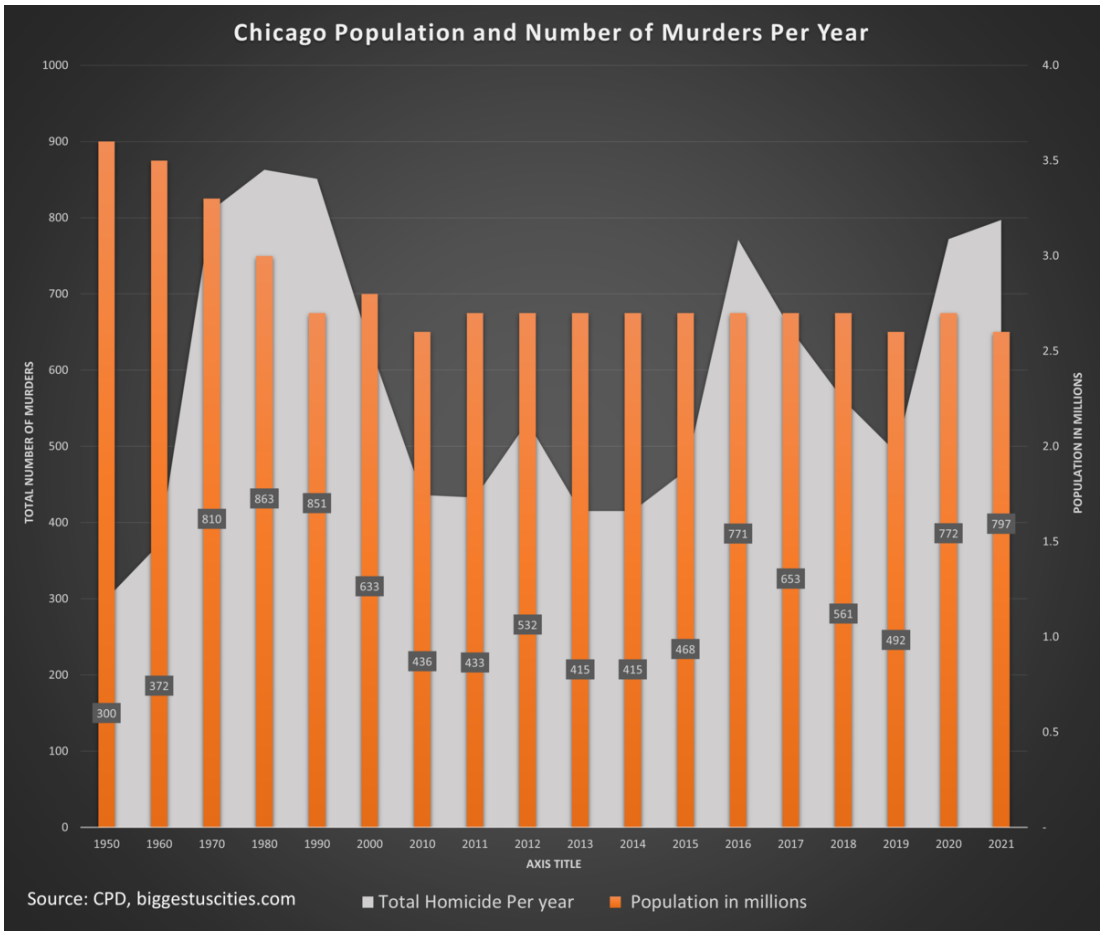
La risposta corretta è:

```
db.videos.aggregate([
{$match: { releaseDate : {$gte: new Date ("2022-01-01"), $lte: new Date ("2022-12-31") } } },
{$unwind: “$tags”},
        {$group: {
                        _id: {
category: “$tags”
      },
                n: {$sum: 1}
                        }
                },
        {$match: { n: { $gte: 1000 } } }
])
```

Domanda 18

Risposta non data

Punteggio max.: 1,25



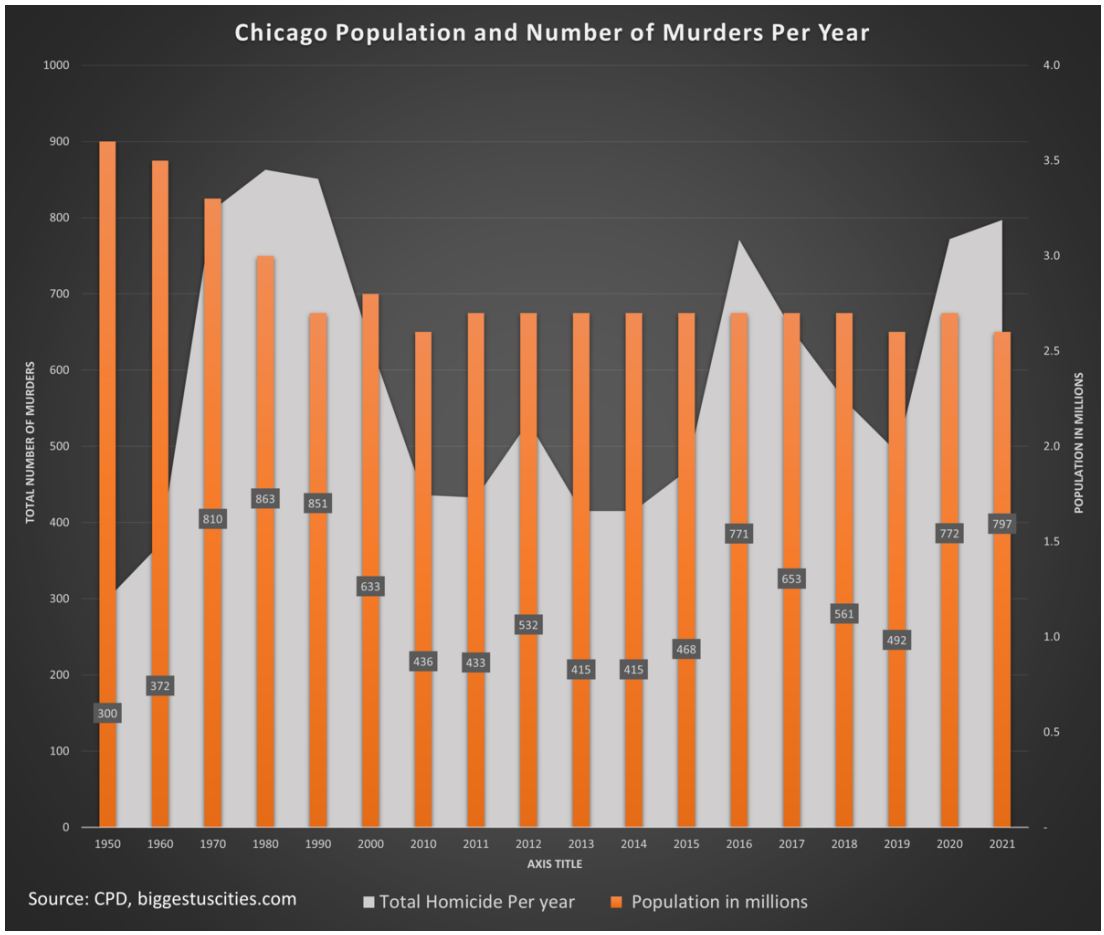
Design schema & Sketch

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

Domanda 19

Risposta non data

Punteggio max.: 1,25



Data

Is the data quality appropriate? Select true answers only.

Scegli una o più alternative:

- (a) Precision is not appropriate as murders shoud be represented with more decimal digits.
- (b) Precision is appropriate for the task as the population is reported in millions with a decimal digit.
- (c) The visualization is very clear because the labels are associated with the bars.
- (d) Data is not complete because several years are missing.
- (e) Data is not credible because no source is reported.
- (f) Data about murders is credible because the source is the Chicago Police Department (CPD).
- (g) Data is consistent because some decades are compared with single years.
- (h) The number of murders and the population are accurate data points because they are represented as percentages.
- (i) Data is reasonably updated for the task because the last data point is from 2021.
- (j) Data is complete because starting from 2010 all years are reported.

Risposta errata.

La risposta corretta è: Data is not complete because several years are missing., Data is reasonably updated for the task because the last data point is from 2021., Precision is appropriate for the task as the population is reported in millions with a decimal digit.,

Data about murders is credible because the source is the Chicago Police Department (CPD).

Domanda 20

Risposta non data

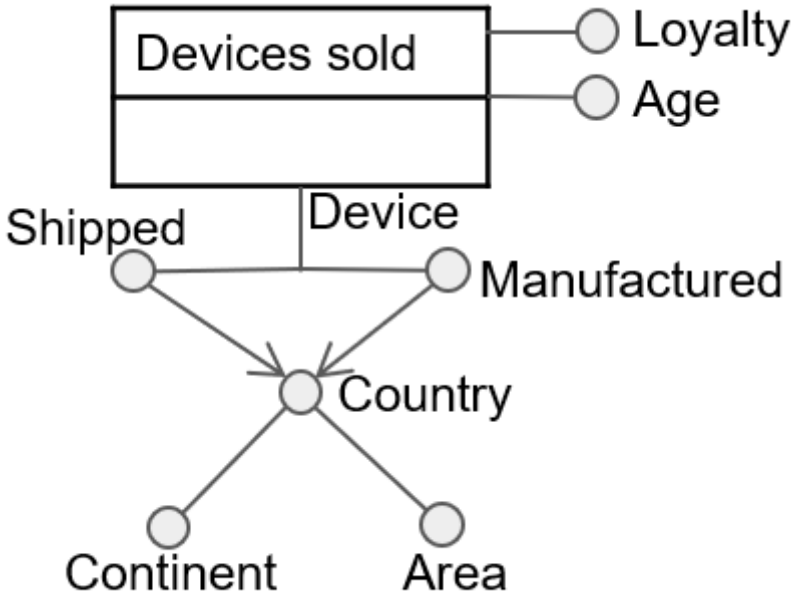
Punteggio max.: 0,50

Design a data warehouse to analyse the purchasing behaviour of consumer electronic devices manufactured by a tech company, in terms of average price per product, and the total number of products sold.

- The tech company manufactures and sells different electronic **devices**
  - All devices have a model and a series. Each model belongs to one series only. Each series collects different models. Each series belongs to one of the following three lines: “basic”, “mid-range”, and “premium”.
  - Device models can be divided into the following types: “smart-phone”, “smart-watch”, or “accessory”.
  - For each model, the following features must be recorded:
    - camera (present/absent),
    - 5G (present/absent),
    - warranty period (1, 2, or 3 years)
- The **products** sold by the tech company are manufactured in and shipped to different places around the world.
  - Both manufacturing places and shipping destinations must be analysed in terms of commercial area, continent and country.
  - A commercial area includes a group of countries, possibly belonging to different continents. Each country belongs to one commercial area only.
- The **consumer** buying the products is characterized by
  - Her/his age, in one of the following ranges: 18-30, 31-45, 45-65, or 65+ years
  - Loyalty program enrolment: yes/no
- The analysis must be performed for each **date**, day of the week, holiday (yes/no), month, 4-month period, semester, and year.

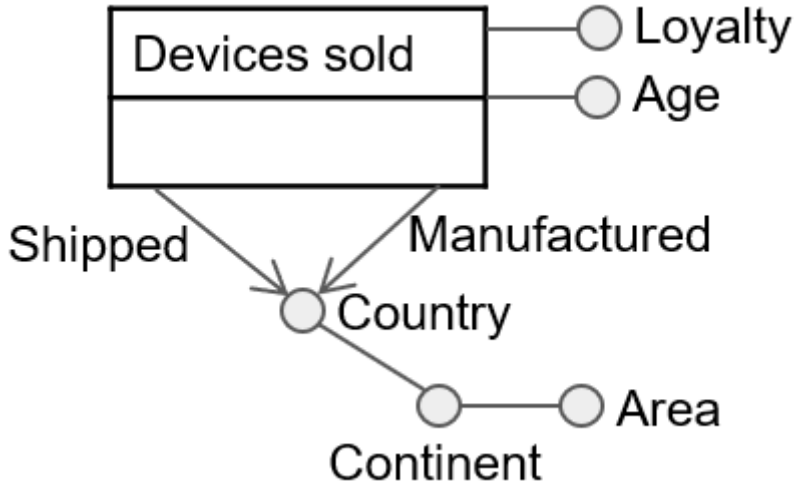
Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

(a)

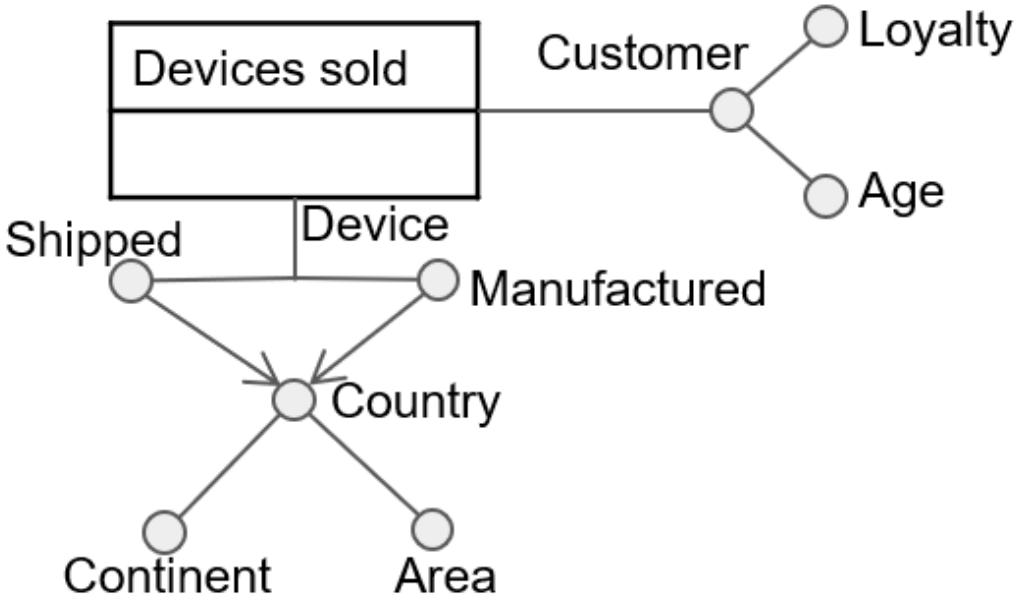


(b) None of the others is correct.

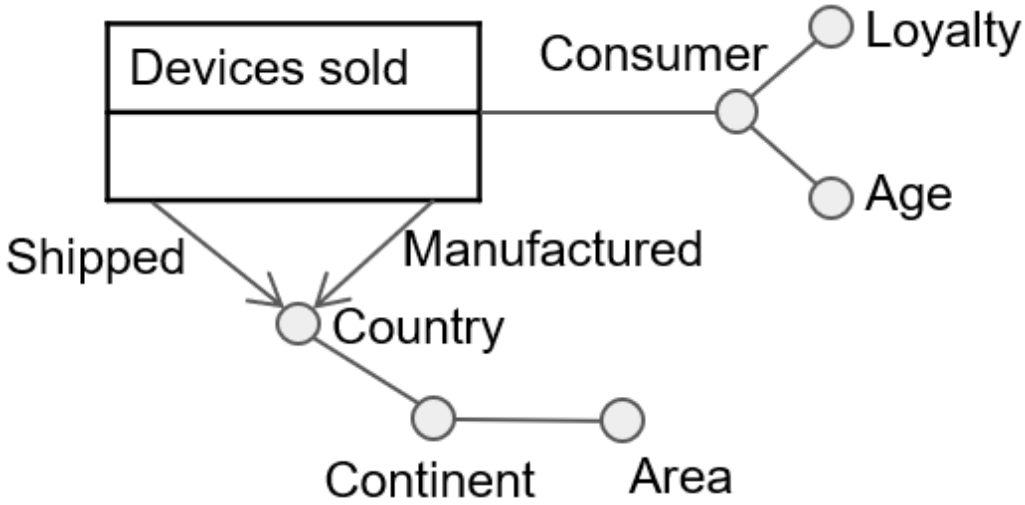
(c)



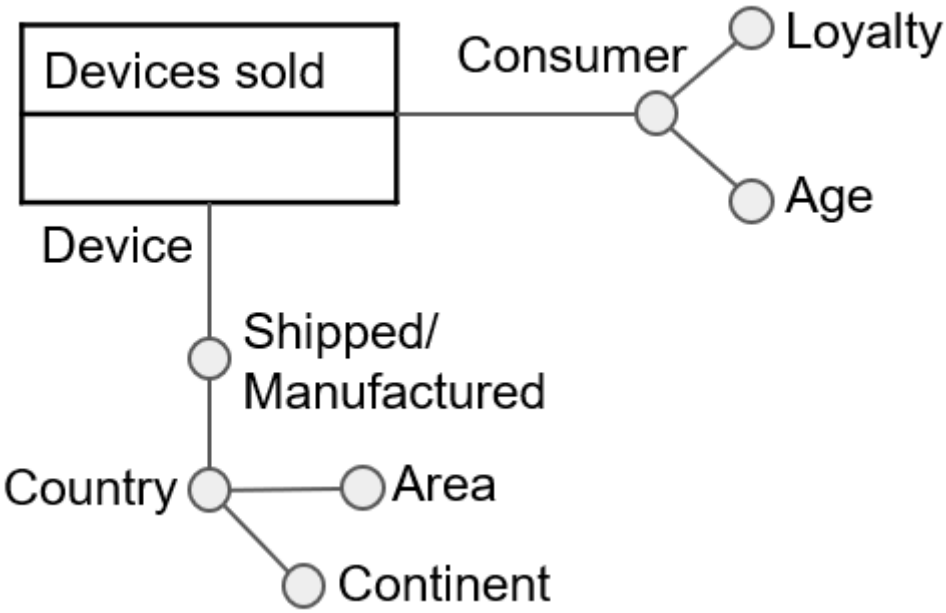
(d)



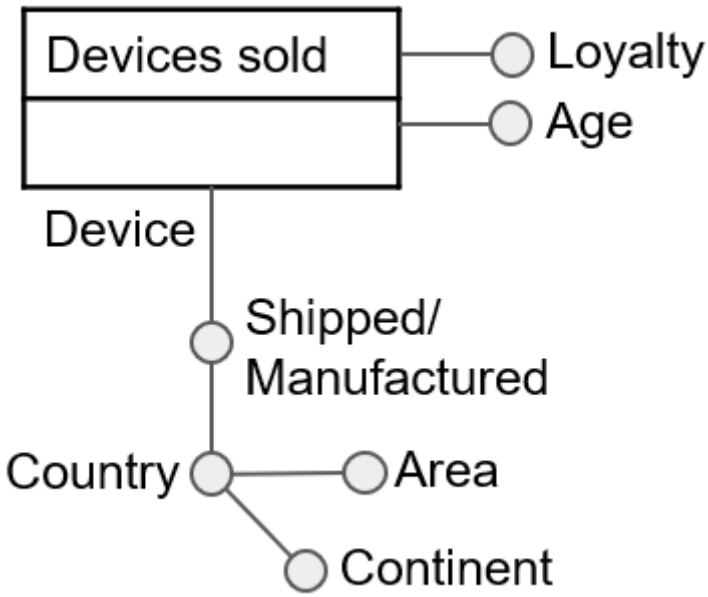
(e)


Devices sold
Consumer
Loyalty
Age
Shipped
Manufactured
Country
Continent
Area

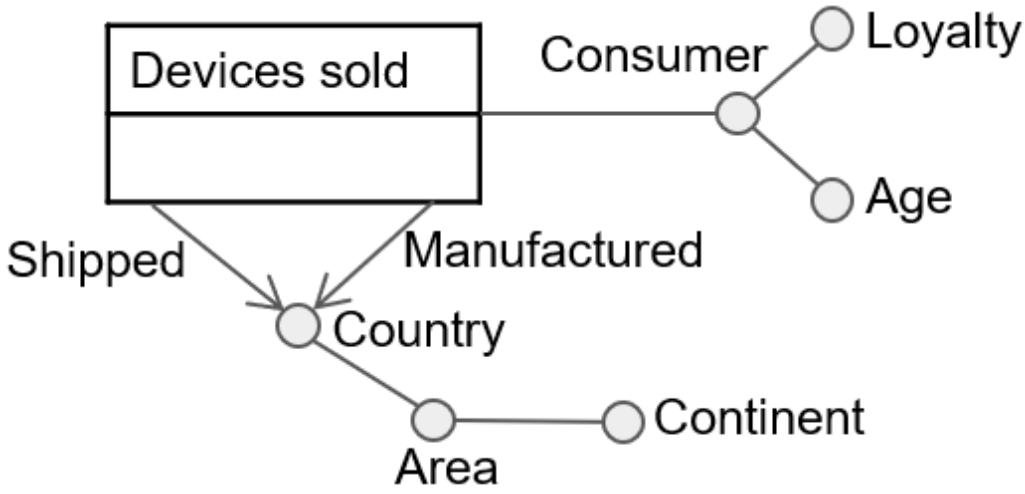
(f)


Devices sold
Consumer
Loyalty
Age
Device
Shipped/
Manufactured
Country
Area
Continent

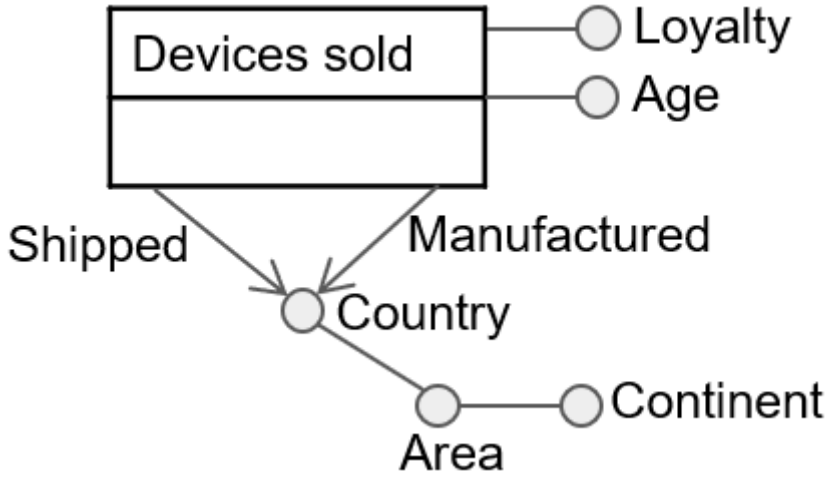
(g)


Devices sold
Loyalty
Age
Device
Shipped/
Manufactured
Country
Area
Continent

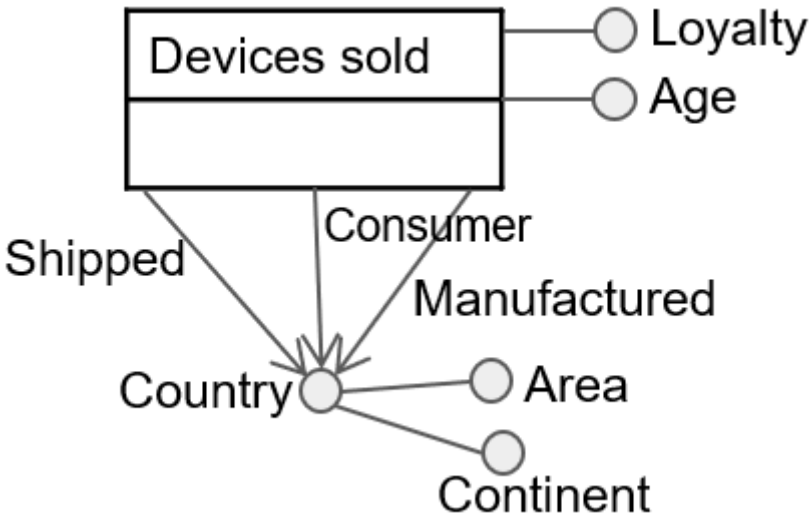
(h)


Devices sold
Consumer
Loyalty
Age
Shipped
Manufactured
Country
Continent
Area

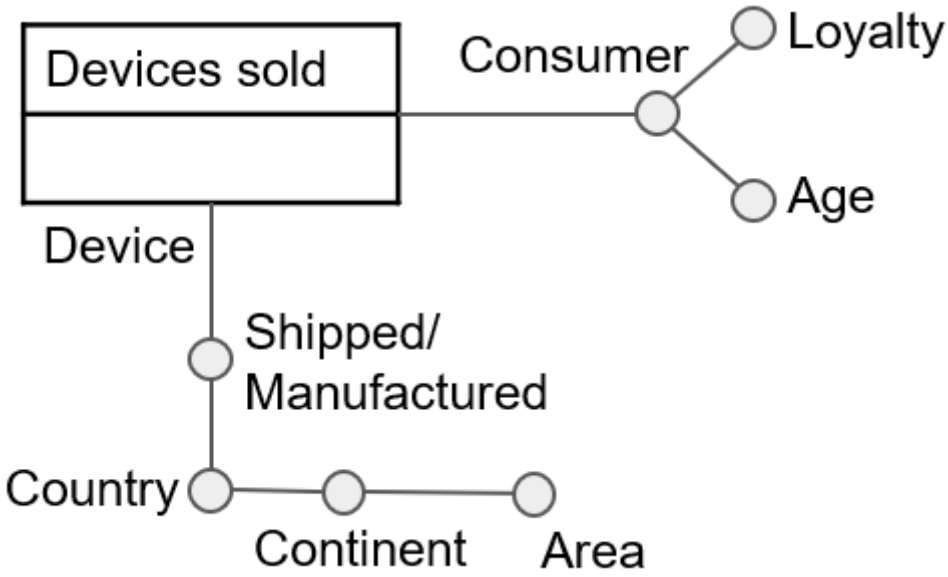
(i)


Devices sold
Loyalty
Age
Shipped
Manufactured
Country
Continent
Area

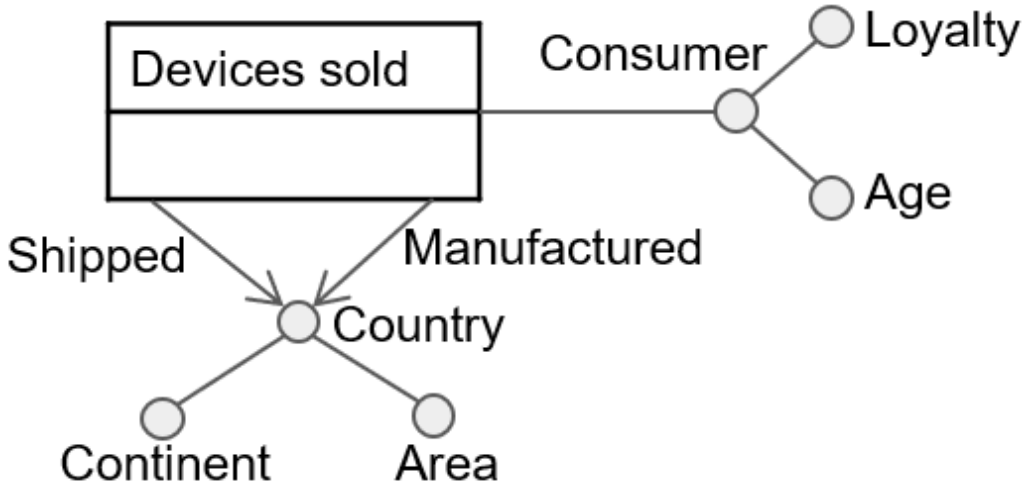
(j)


Devices sold
Loyalty
Age
Consumer
Shipped
Manufactured
Country
Area
Continent

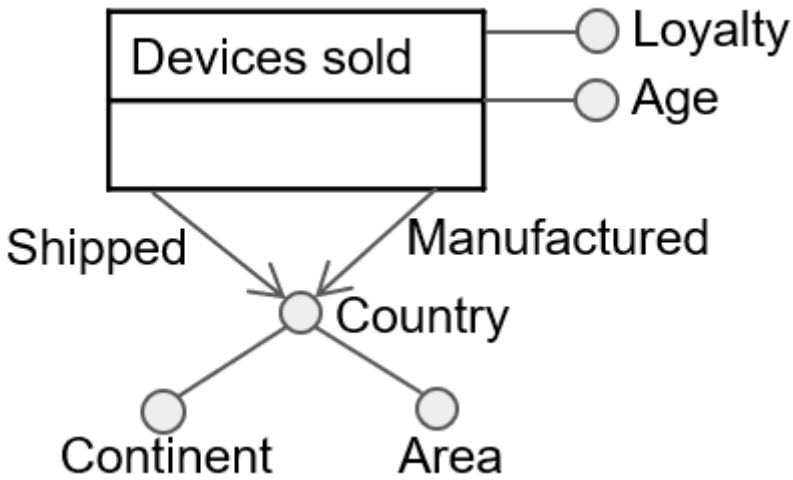
(k)



(l)

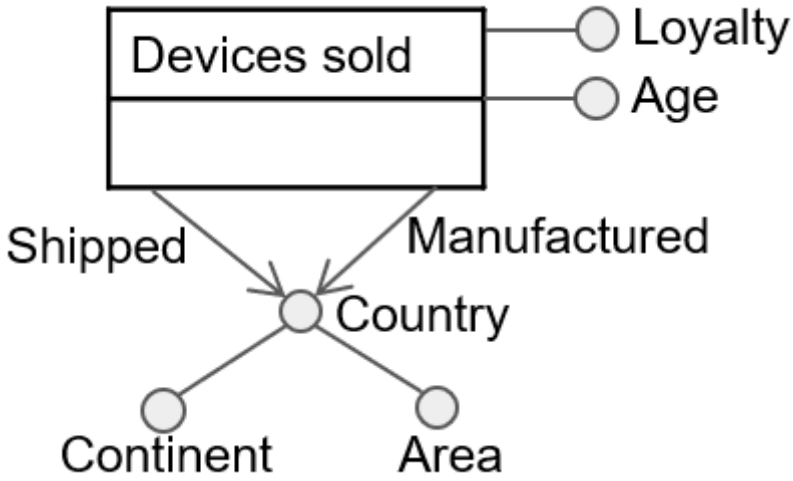


(m)



Risposta errata.

La risposta corretta è:



Domanda 21

Risposta non data

Punteggio max.: 4,00

Design a MongoDB database to collect restaurant menus and dish recipes according to the following requirements.

The dish recepits are characterized by a name, a description, the level of complexity (from 1 to 5), and the chef who proposed the dish. For the chef, the name, country of origin, website address (if available), and email address (if available) are known.

The ingredients used in each recipe must be recorded. Each ingredient consists of the name, the quantity, and (if necessary) the unit of measure. For example, a cake may use 500 ml of milk as an ingredient. In that case, the ingredient name is “milk”, the quantity is 500, and the unit of measure is “ml”. Each ingredient can be present only once in a recipe.

The menus proposed by different restaturants must be recorded. Each menu is characterized by the name of its restaurant, and the list of dishes.

For the restaurant, its location is known. The restaurant location is characterized by the street, the street number, the zip code, the city, the country and the geographical coordinates (latitude and longitude).

For each dish, its price must be recorded (i.e., the currency and the value).

Given a menu, the database must be designed to efficiently provide all the dish names and their chef names without additional lookups.

Write a sample document for each collection of the database.

Explicitly indicate the design patterns used.

Dish

```
{
_id: ObjectId(),
name: <string>,
description: <string>,
complexity_level: <int or float>,
chef: {
        name: <string>,
        country: <string>,
        website: <url>,
        email: <string>
  },
ingredients: { // otherwise a list with sub-dictionaries, i.e., attribute pattern
  “milk”: {value: 500, uom: “ml”},
  “eggs”: {value: 2},
  },
}
```

Menu

```
{
_id: ObjectId(),
name: <string>,
restaurant: {
        “street”: <string>,
        “street_number”: <string>,
        “zip_code”: <number>,
        “city”: <string>,
        “country”: <string>,
         // coordinates only, without Point, are acceptable [x,y] as legacy format
        “loc”: {“type”:”Point”,
                “coordinates”:[: <number>, <number>]
}
dishes: [
{
  _id: Objectid(),
  dish_name: <string>,
  chef_name: <string>,
  price: { val: <number>, currency: <string> } // otherwise price_val and price_currency
  }, ...
]
}
```

Pattern used:

- Extended reference pattern in menu collection for dishes name and chef
- Polymorphic pattern for chef information
- Polymorphic or Attribute pattern for ingredients