Data management and visualization

Started on	Tuesday, 9 February 2021, 12:57 PM
State	Finished
Completed on	Tuesday, 9 February 2021, 12:57 PM
Time taken	9 secs
Grade	**0.00** out of 31.00 (**0**%)

Question 1

Not answered

Marked out of 1.50

In Datawarehouse analysis, the aggregation window defined at the physical level in extended SQL:

- (a) can be specified only on a single sort key
- (b) is based on a physical structure (e.g. an index)
- (c) is specified with the range clause
- (d) is defined by counting the rows
- (e) is appropriate for sequence data with gaps and sparse data

Risposta errata.

The correct answer is: is defined by counting the rows

Question 2

Not answered

Marked out of 1.00

In NoSQL design, the extended reference pattern has the advantage of:

- (a) reducing the overall number of documents in a collection
- (b) reducing data denormalization
- (c) reducing the CPU workload for frequent computations
- (d) reducing the join operations
- (e) reducing the reference to document extensions
- (f) reducing future technical debt
- (g) reducing document complexity

Risposta errata.

The correct answer is: reducing the join operations

Question 3

Not answered

Marked out of 1.00

Select the right configuration of a MongoDB replica set:

- (a) 2 secondary nodes, 1 arbiter node
- (b) 1 secondary node, 1 arbiter node
- (c) 1 primary node, 2 secondary nodes , 2 arbiter nodes
- (d) 2 primary nodes, 2 secondary nodes, 1 arbiter node
- (e) 2 primary nodes, 1 secondary node
- (f) 2 primary nodes, 2 secondary nodes, 2 arbiter nodes

Risposta errata.

The correct answer is: 1 primary node, 2 secondary nodes , 2 arbiter nodes

Question 4

Not answered

Marked out of 1.50

Which one of the following examples is **NOT** related to a Gestalt principle?

- (a) the bars representing smaller values are shorter
- (b) the points of a data series are connected
- (c) the color of the legend is similar to the color of the elements of the graph
- (d) the direct labeling technique improves the readability of the visualization
- (e) the points of a group are enclosed by a fine line

Risposta errata.

The correct answer is: the bars representing smaller values are shorter

Question 5

Not answered

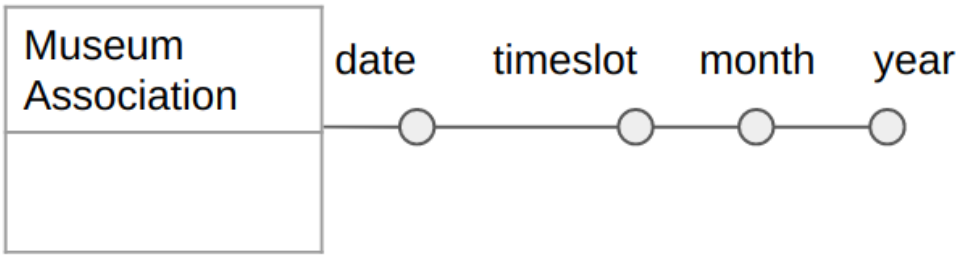
Marked out of 0.50

Data analysts of the National Association of Italian Museums are interested in analyzing the average revenue per ticket.
In particular, they would like the analyses to address the following features.

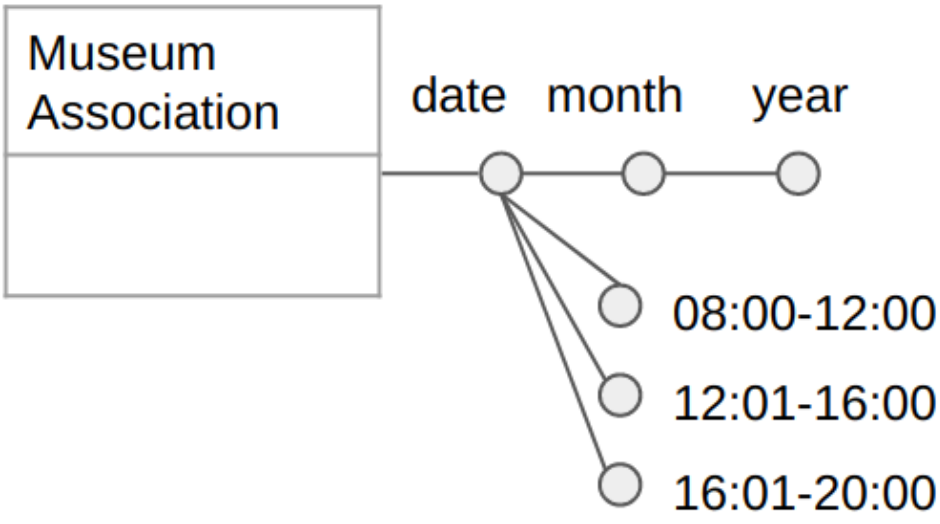
- Museums are analyzed according to their city and region. A museum has a unique name, and it is located in a specific city. The same city can host different museums.
- A museum may have some additional services available for its public. The systems records which services are available for each museum.
  Examples of additional services are “guided tours”, “audioguides”, “wardrobe”, “cafè”, “Wi-Fi”. The number of possible additional services is large and growing, hence the full list is not known a priori.
- The tickets sold by each museum are recorded. There are 4 different types of tickets: “Full price”, “Reduced-student” (for students from 18 to 24 years old), “Reduced-junior” (for young people less than 18 years old), and “Reduced-senior” (for people over 70 years old).
- The analyses must be carried out considering the date, month and year, and the time slot of the ticket emission. The time slot is stored in 3 ranges of 4-hour blocks (08:00-12:00, 12:01-16:00, 16:01-20:00).

Choose the correct conceptual schema from the proposed ones to properly define the time dimension according to the given specifications (at most one answer is correct).

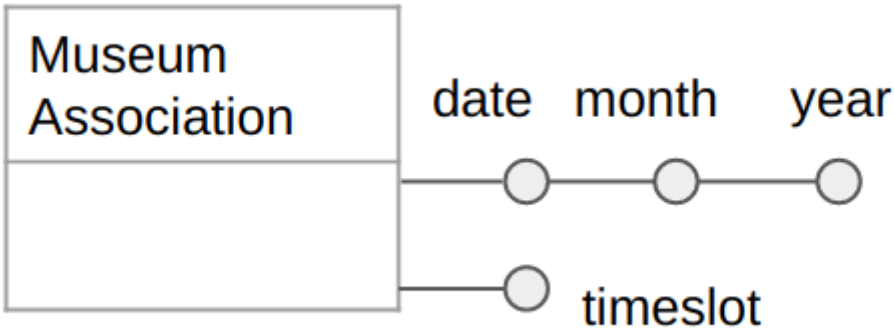
(a)



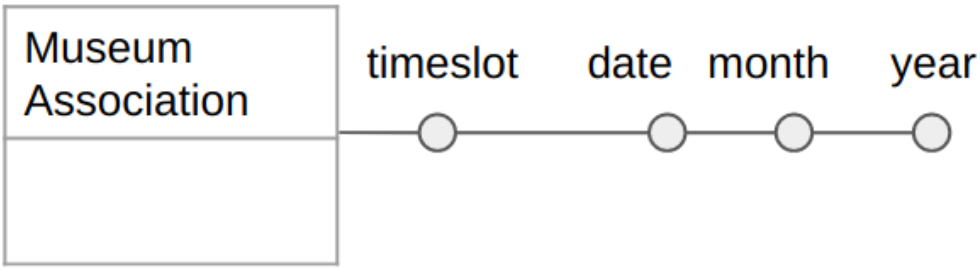
(b)



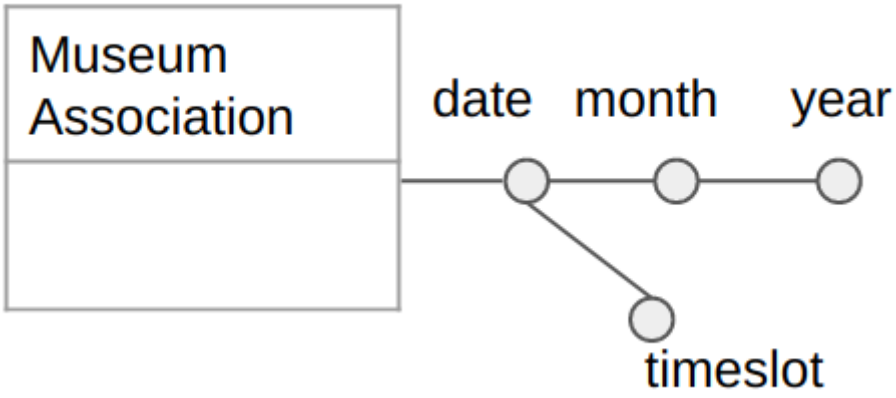
(c)



(d)

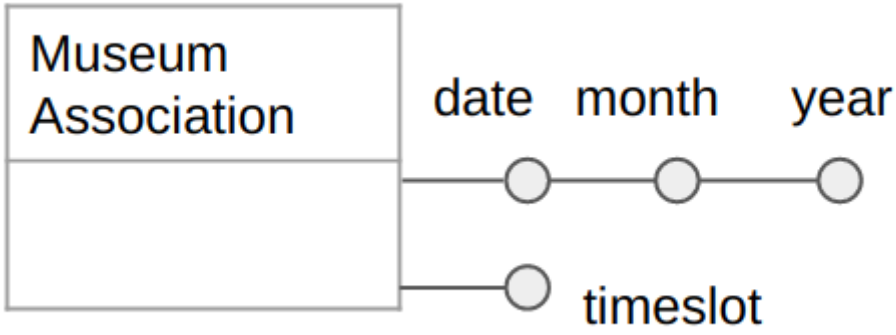


(e)



Risposta errata.

The correct answer is:



Question 6

Not answered

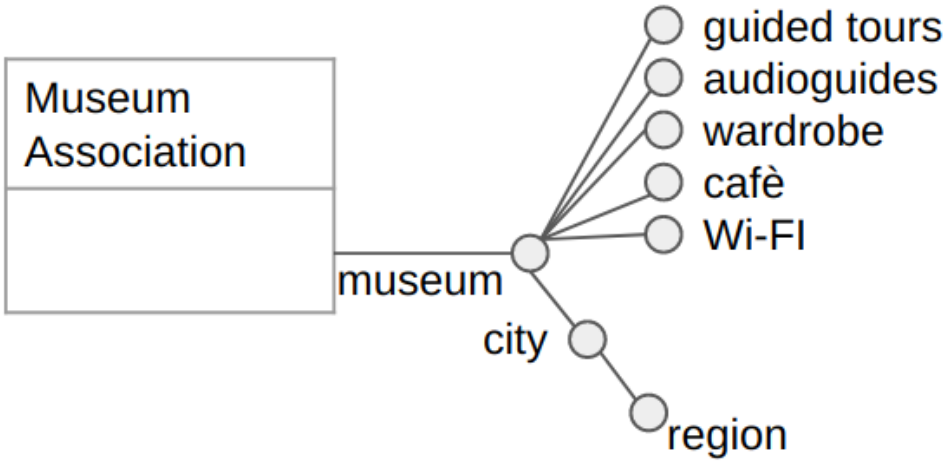
Marked out of 0.50

Data analysts of the National Association of Italian Museums are interested in analyzing the average revenue per ticket.
In particular, they would like the analyses to address the following features.

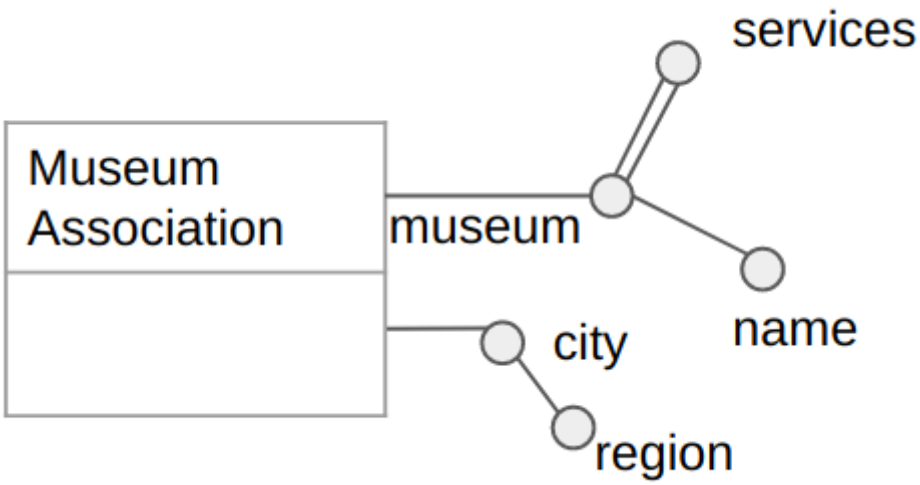
- Museums are analyzed according to their city and region. A museum has a unique name, and it is located in a specific city. The same city can host different museums.
- A museum may have some additional services available for its public. The systems records which services are available for each museum.
  Examples of additional services are “guided tours”, “audioguides”, “wardrobe”, “cafè”, “Wi-Fi”. The number of possible additional services is large and growing, hence the full list is not known a priori.
- The tickets sold by each museum are recorded. There are 4 different types of tickets: “Full price”, “Reduced-student” (for students from 18 to 24 years old), “Reduced-junior” (for young people less than 18 years old), and “Reduced-senior” (for people over 70 years old).
- The analyses must be carried out considering the date, month and year, and the time slot of the ticket emission. The time slot is stored in 3 ranges of 4-hour blocks (08:00-12:00, 12:01-16:00, 16:01-20:00).

Choose the correct conceptual schema from the proposed ones to properly define the characteristics of museum analytics according to the given specifications (at most one answer is correct).

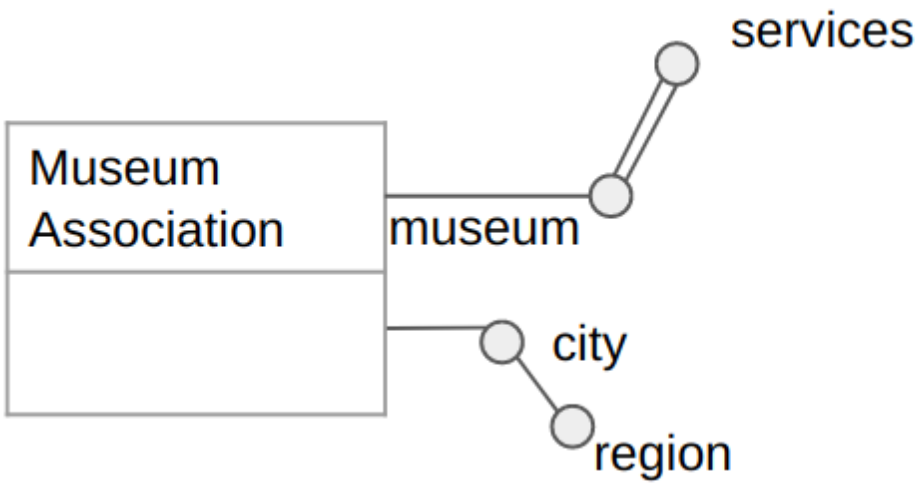
(a)



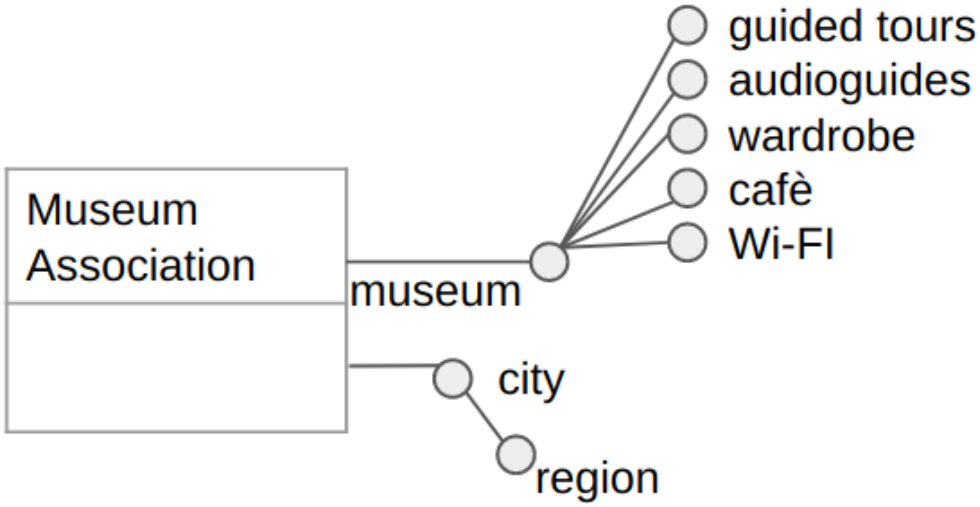
(b)



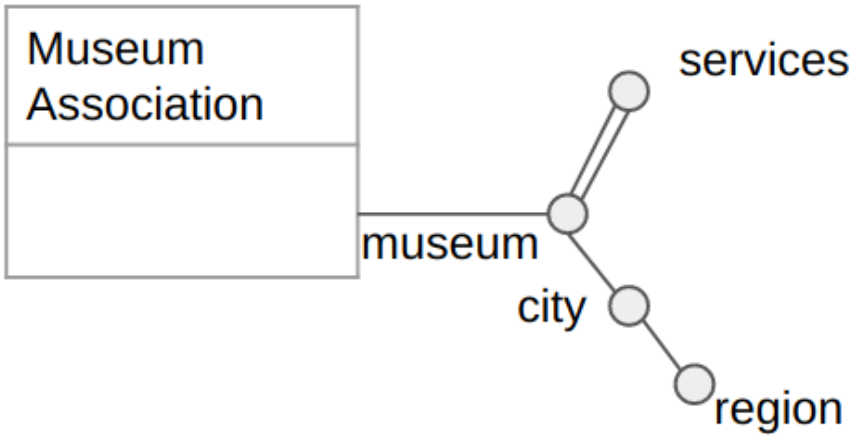
(c)



(d)

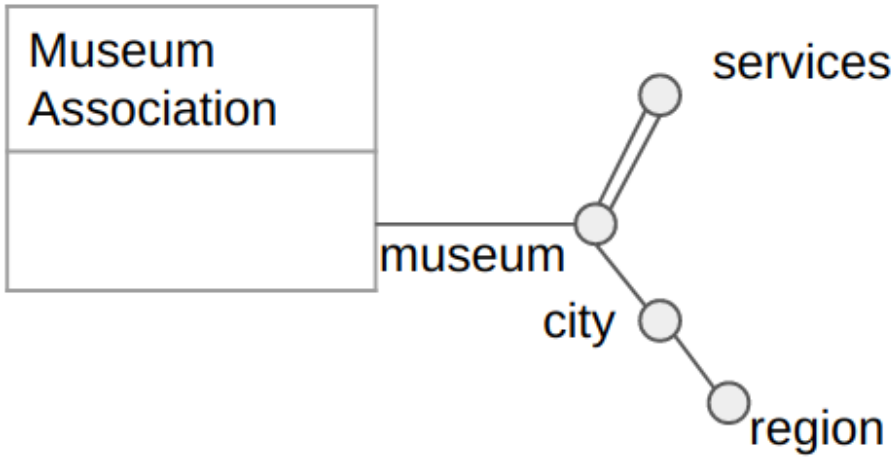


(e)



Risposta errata.

The correct answer is:



Question 7

Not answered

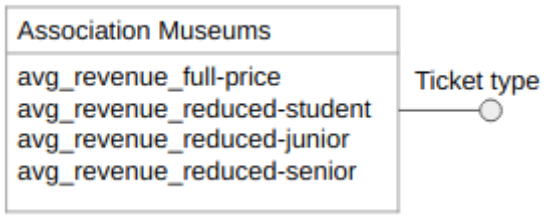
Marked out of 1.00

Data analysts of the National Association of Italian Museums are interested in analyzing the average revenue per ticket.
In particular, they would like the analyses to address the following features.

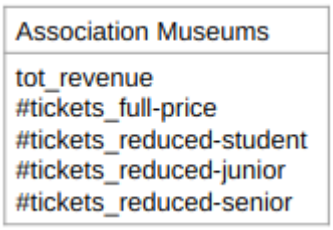
- Museums are analyzed according to their city and region. A museum has a unique name, and it is located in a specific city. The same city can host different museums.
- A museum may have some additional services available for its public. The systems records which services are available for each museum.
  Examples of additional services are “guided tours”, “audioguides”, “wardrobe”, “cafè”, “Wi-Fi”. The number of possible additional services is large and growing, hence the full list is not known a priori.
- The tickets sold by each museum are recorded. There are 4 different types of tickets: “Full price”, “Reduced-student” (for students from 18 to 24 years old), “Reduced-junior” (for young people less than 18 years old), and “Reduced-senior” (for people over 70 years old).
- The analyses must be carried out considering the date, month and year, and the time slot of the ticket emission. The time slot is stored in 3 ranges of 4-hour blocks (08:00-12:00, 12:01-16:00, 16:01-20:00).

Choose the best solution for the ticket information and measures in the conceptual schema design among those proposed (at most one answer is correct).

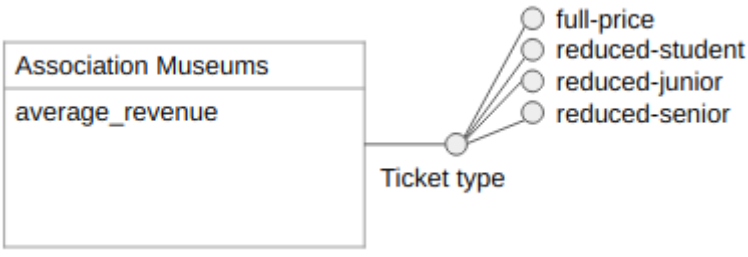
(a)



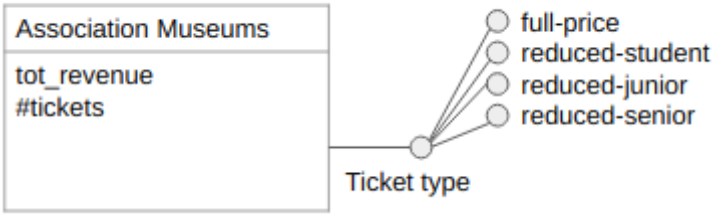
(b)



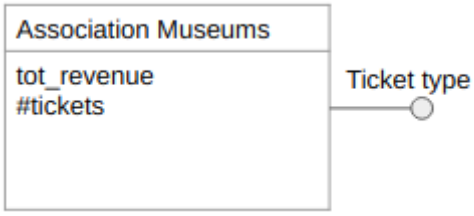
(c)



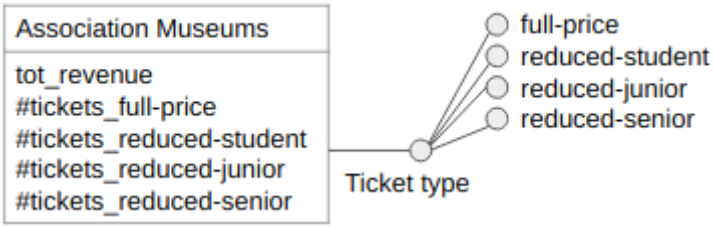
(d)



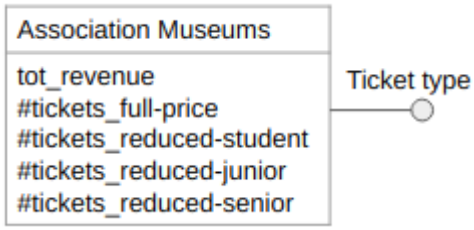
(e)



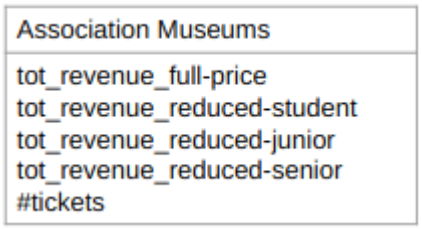
(f)



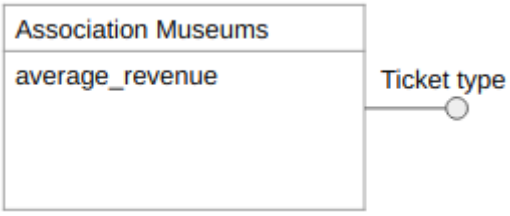
(g)



(h)

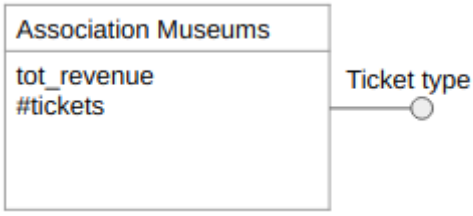


(i)



Risposta errata.

The correct answer is:

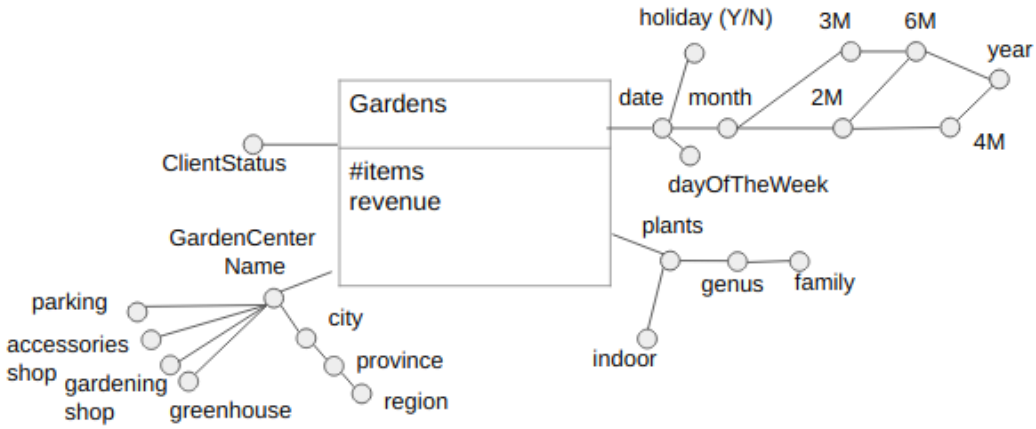


Question 8

Not answered

Marked out of 2.00

Given the following conceptual schema:



- Each garden center has a unique name. A garden center can have 0 or more services. There are 4 available services: “parking”, “accessories shop”, “gardening shop” and “greenhouse”.
- The cardinality of “ClientStatus” is 3, and it can be “1” for “Silver”, “2” for “Gold” and “3” for “Platinum”.
- A plant can be either an indoor or an outdoor plant. The genus and family of the plant are stored.

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.

Gardens(TimeId, GardenCenterId, PlantId, ClientStatus, #items, revenue)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek, holiday)
GardenCenter(GardenCenterId, GardenCenter, city, province, region, greenhouse, accessoriesShop, gardenShop, parking)
Plant(PlantId, plant, genus, family, indoor)

Question 9

Not answered

Marked out of 4.00

```
MusicStreaming(TimeId, SongId, PlatformId,
NumberOfStreamings, NumberOfLikes)
Time(TimeId, date, month, 2M, 3M, 6M, year, dayOfTheWeek)
Song(SongId, Song, album, classic, indie, pop, .., rock)
UserLocation(UserLocationId, province, region, country)
```

For each song and month, compute the following metrics:

- the total number of streamings
- the cumulative total number of streamings since the beginning of the year
- assign a rank to each song, separately for each album, based on the monthly number of streamings (rank 1st the most streamed song of the album for each month)

Write the requested SQL query.

```
SELECT month, song,
    SUM(NumberOfStreamings),
      SUM(SUM(NumberOfStreamings)) OVER (PARTITION BY songId, year
ORDER BY month ROWS UNBOUNDED PRECEDING)
      RANK() OVER (PARTITION BY album, month ORDER BY
SUM(NumberOfStreamings) DESC),

FROM Song S, Time T, MusicStreaming MS

WHERE S.SongId=MS.SongId AND T.Timeid=MS.Timeid

GROUP BY song, songId, month, year, album
```

Question 10

Not answered

Marked out of 4.00

```
MusicStreaming(TimeId, SongId, PlatformId,
NumberOfStreamings, NumberOfLikes)
Time(TimeId, date, month, 2M, 3M, 6M, year, dayOfTheWeek)
Song(SongId, song, album, classic, indie, pop, .., rock)
UserLocation(UserLocationId, province, region, country)
```

Separately for each song and province of the user, compute the following metrics:

- the average number of monthly likes
- the percentage of the number of likes with respect to the total number of likes received by users in the same country
- the number of likes of the album in the user province

Write the requested SQL query.

```
SELECT province, song,
    SUM(NumberOfLikes)/ COUNT(DISTINCT month),
      100*SUM(NumberOfLikes)/SUM(SUM(NumberOfLikes)) OVER (PARTITION
BY songId, country),
      SUM(SUM(NumberOfLikes)) OVER (PARTITION BY album, province)

FROM Song S, MusicStreaming MS, UserLocation Lm, Time T

WHERE S.SongId=MS.SongId AND L.UserLocationId=MS.PlatformId AND
T.TimeId=MS.TimeId

GROUP BY song, songId, province, country, album
```

Question 11

Not answered

Marked out of 2.00

Given the following document structure:

```
{
"address":
    {"building":"768",
    "coord":[-73.9685872,40.7679509],
    "street":"Madison Avenue",
    "zipcode":"10065",
    "borough":"Manhattan",
    "city": "New York"},
"sold_items": ["Smartphones", "PC", "TV"],
"reviews":[
    {"date": {date:{"2019-11-05"}}, "score":10, "description": "Lorem ipsum"},
    {"date": {date("2020-02-21"}}, "score":8, "description": "Lorem ipsum"}
    ],
"name":"Elettronic-store"
}
```

Select all the shops located in Rome that sell smartphones or TV and received at least one review with a score greater than 8. Show only the name, the street and the building.

```
db.shops.find({sold_items:{$in: ["Electronics", "Home"]},
            "address.city":"Rome",
            "reviews.score":{$gt:8} },
            {_id:0, name: 1, "address.street":1,"address.building":1 })
```

Question 12

Not answered

Marked out of 3.00

Given the following document structure:

```
{
"name":"Electrostore",
"address":
    {"building":"A1",
     "street":"via Torino",
     "zipcode":"12345",
     "borough":"Campidoglio",
     "city": "Rome"},
"sold_items": ["Smartphone", "PC", "TV"],
"reviews":[
    {"date": "2019-11-05", "score":10, "description": "Lorem ipsum"},
    {"date": "2020-02-21", "score":7, "description": "Lorem ipsum"}
    ]
}
```

For each city, compute the average and the maximum review score.

Show only the first 10 cities with the highest number of reviews.

```
db.collection.aggregate([
{$unwind: "$reviews"},
{$group: {
        _id: "$address.city",
        'countReviews':   {$sum: 1},
        'maxReviewScore': {$max: '$reviews.score'},
        'avgReviewScore': {$avg: '$reviews.score'}
}},
{$sort:
        {countReviews: -1}
},
{$limit: 10}
])
```

Question 13

Not answered

Marked out of 4.00

Design a MongoDB database to manage a warehouse for parcel delivery according to the following requirements.

Customers of the parcel delivery service are citizens identified by their social security number. They can be senders or recipients of delivered parcels. They are characterized by their name, surname, email address, a telephone number, and by different addresses, one for each type, e.g., one billing address, one home address, one work address, etc. Each address consists of street name, street number, postal code, city, province, and country.

Parcels are characterized by a unique barcode and their physical dimensions (specifically: width, height, depth, and weight). All widths, heights, and depths are always expressed in meters. All weights are always expressed in kilograms.

The recipient and the sender information required to deliver each parcel must be always available when accessing the data of a parcel. Recipient and sender information required to deliver a parcel consists of: full name, street name, street number, postal code, city, province, and country. For instance, a recipient information can be: Mario Rossi, corso Duca degli Abruzzi, 24, 10129, Torino, Torino, Italy.

The parcel warehouse is divided into different areas. Each area is identified by a unique code, e.g., 'area_51' and consists of different lines. Each line is identified by unique code, e.g., 'line_12', and hosts several racks. Each rack is identified by unique code, e.g., 'rack_33', and is made up of shelves. Each parcel is placed on a specific shelf of the warehouse, identified by a unique code, e.g., 'shelf_99'. The database is required to track the location of each parcel within the warehouse.

Given a parcel, the database must be designed to efficiently provide its full location, from the shelf, up to the area, through the rack and line.

Given a customer, the database must be designed to efficiently provide all her/his parcels as a sender, and all his/her parcels as a recipient.

Write a sample document for each collection of the database. Explicitly indicate the design patterns used.

Parcel

```
{
    _id: <string>, // barcode
    dimensions: { // also 1st-level attribute is fine
        width: <number>,
        height: <number>,
        depth: <number>,
        weight: <number>
    }
    recipient: {
        _id: <string>, // SSN, _id of the customer
        street: <string>,
        civic_number: <string>,
        zip_code: <string>,
        city: <string>,
        province: <string>
    },
    sender:{
        _id: <string>,// SSN, _id of the customer
        street: <string>,
        civic_number: <string>,
        zip_code: <string>,
        city: <string>,
        province: <string>
    },
    father_pos: <string>, // code of area/lane/rack/shelf
    locations: [
        <string> // code of area/lane/rack/shelf
    ]
}
```

Tree pattern for the position. The full list of tree-pattern ancestors is required. The parent ancestor of the tree-pattern is optional.

No collection for the areas, since no data are tracked except their code.

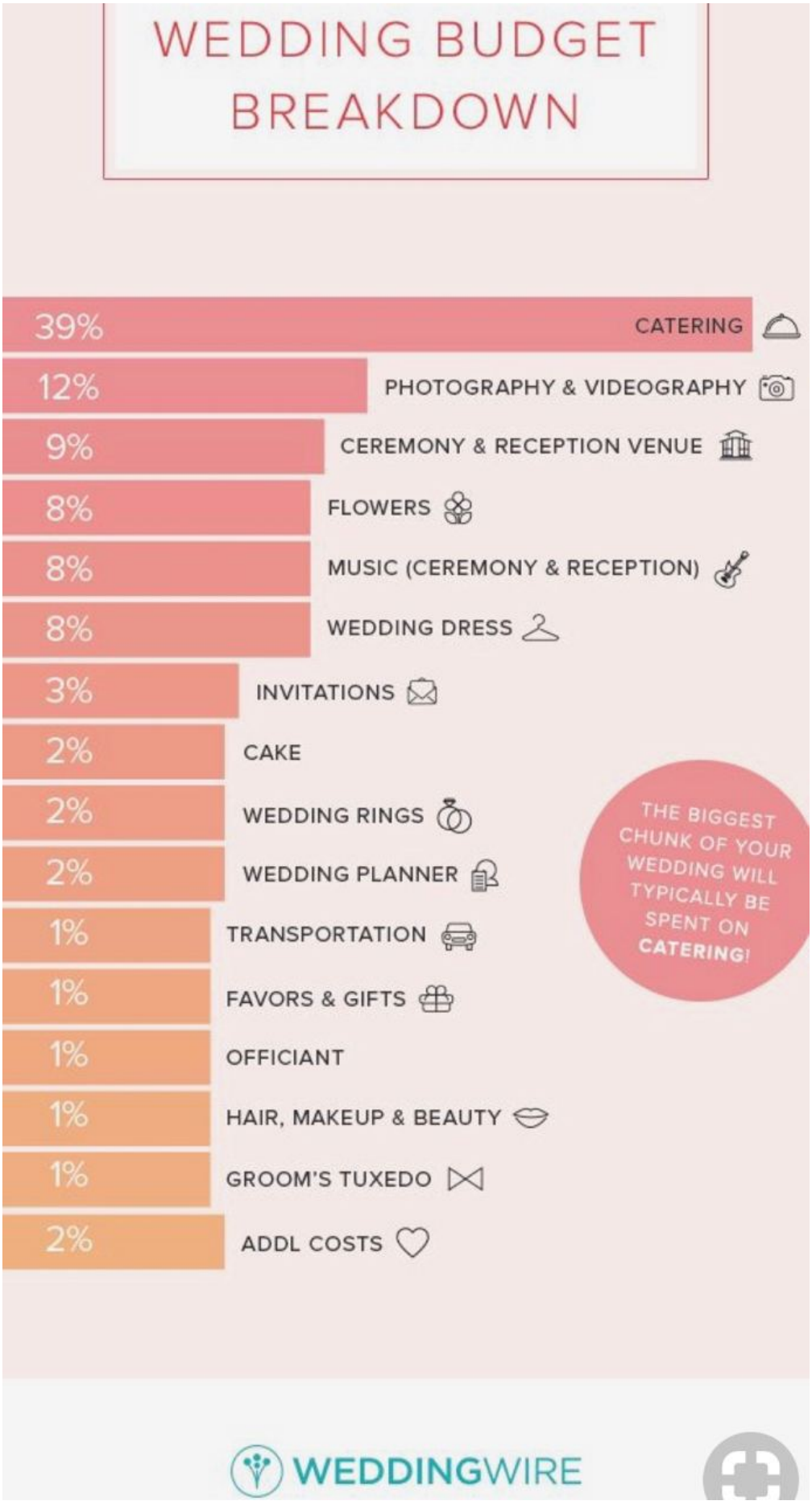
Extended reference pattern for recipient and sender address information. The recipient and sender _ids are required to look up all parcels of a given customer.

Customers

```
{
    _id: <string>, // fiscal code
    name: <string>,
    surname: <string>,
    email: <string>,
    tel: <string>,
    addresses:
        {
            home:
                {
                 street_name: <string>,
                 street_num: <string>,
                 postal_code: <int>,
                 city: <string>,
                 province: <string>,
                 country: <string>
                },
            'billing':
                {
                 street_name: <string>,
                 street_num: <string>,
                 postal_code: <int>,
                 city: <string>,
                 province: <string>,
                 country: <string>
                },
            work: {...}
        },
}
```

Attribute pattern (optional) for the addresses attribute.

Information



Analyze the above graph reporting the average breakdown of wedding costs. According to their website, WeddingWire is "the largest and most trusted global marketplace connecting engaged couples with local wedding professionals". WeddingWire published these data on a blog post dated December 2020: "We surveyed thousands of couples around the country in our WeddingWire Newlywed Report to share their wedding budgets with us".

Question 14

Not answered

Marked out of 0.25

Question

Is there a clearly defined question addressed by the visualization? Write it down.

Question 15

Not answered

Marked out of 1.25

Data

Is the data quality appropriate? Identify the inadequate characteristics and explain.

Question 16

Not answered

Marked out of 0.75

Visual Proportionality

Are the values encoded in a uniformly proportional way?

Question 17

Not answered

Marked out of 0.75

Visual Utility

All the elements in the graph convey useful information?

Question 18

Not answered

Marked out of 0.50

Visual Clarity

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

Question 19

Not answered

Marked out of 0.25

Design data

Design the visualization based on the following data structure (to be completed).

Question 20

Not answered

Marked out of 1.25

Design schema & Sketch

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

Question 21

Not answered

Not graded

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

Anything written here will NOT be evaluated.

DM & Visualization - Exam 2021-02-01 - Solution

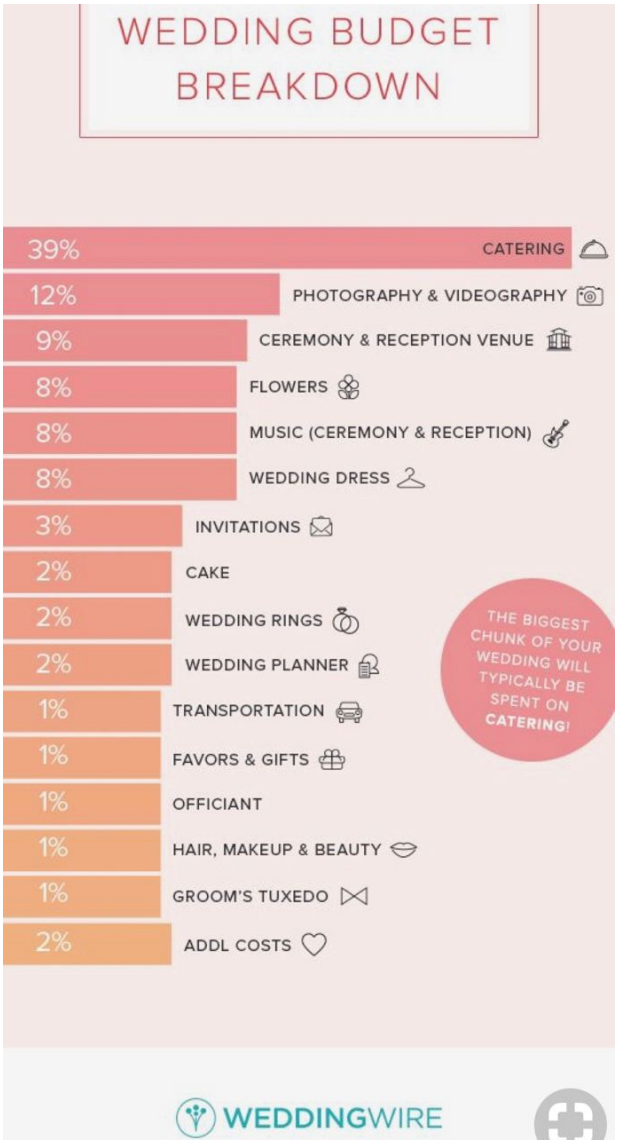


Figure 1: Breakdown of wedding costs according to WeddingWire

Analysis

Analyze the above graph reporting the average breakdown of wedding costs. According to their website, WeddingWire is "the largest and most trusted global marketplace connecting engaged couples with local wedding professionals". WeddingWire published these data on a blog post dated December 2020: "We surveyed thousands of couples around the country in our WeddingWire Newlywed Report to share their wedding budgets with us".

Question: Is there one (or more) question addressed by the visualization?

The question is very clear: what is the budget breakdown per category of expenses for an average wedding?

Data: Is the data quality appropriate?

Accuracy: data are comparable and the values are reasonable according to common judgment.

Completeness: data are complete, several categories are reported and we can assume that the list is exhaustive.

Consistency: the percentages of the categories correctly sum to 100%. The meaning of additional costs is unclear.

Currency: data are referred to the year 2020, so it is updated.

Credibility: the source is mentioned at the bottom and they are domain experts. We don't know how many couples answered the survey.

Understandability: data are understandable, but it is better to report absolute numbers instead of percentages.

Precision: a higher precision, maybe to the first decimal digit, would be more appropriate as many values are equal.

Visual Proportionality: Are the values encoded in a uniformly proportional way?

Not at all, as the lengths of the bars representing 1% and 2% (or 2% and 3%) are almost equal.

Visual Utility: All the elements in the graph convey useful information?

Several elements are useless: the colored background, the icons of the categories, the icon at the bottom-right, the textual comment, the rectangle around the title.

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

The usage of direct labeling is appropriate and very clear. However, the meaning of the different colors associated with the bars is not clear.

Design

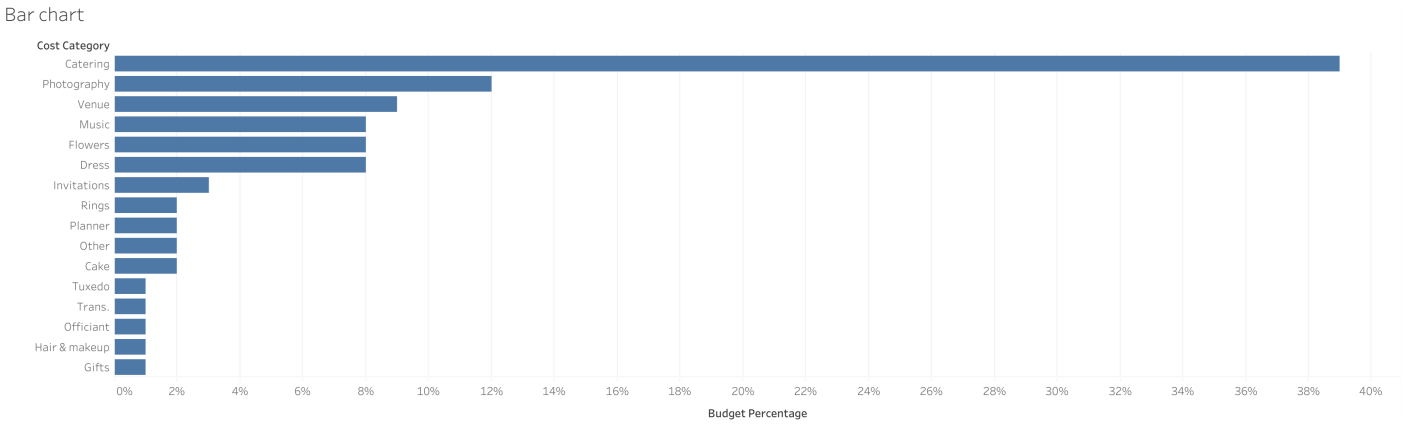
Design the visualization based on the following data structure

Field	Dim./Measure	Description
BUDGET_PERCENTAGE	Measure	Percentage of the total budget
COST_CATEGORY	Dimension	The categories of wedding expenses

Design schema

Schema	Details
Columns:	SUM(BUDGET_PERCENTAGE)
Rows:	COST_CATEGORY
Graph type:	Bar
Color:	Default
Size:	Default
Label:	Default

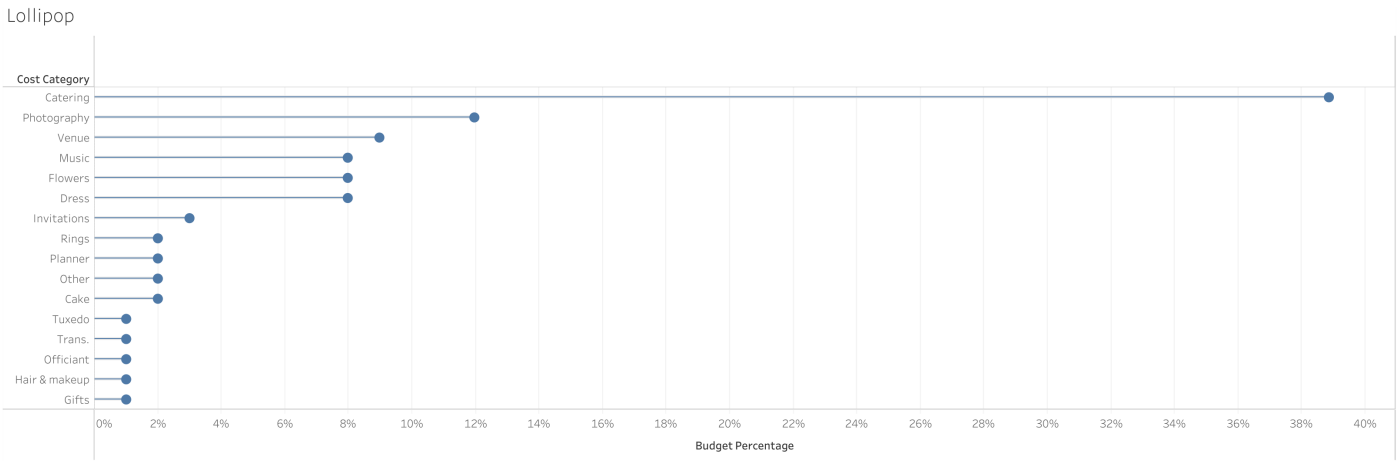
Sketch of the resulting graph



Design schema

Schema	Details
Columns:	SUM(BUDGET_PERCENTAGE), SUM(BUDGET_PERCENTAGE)
Rows:	COST_CATEGORY
Graph type:	Bar, Circle
Color:	Default
Size:	Smaller, Default
Label:	Default

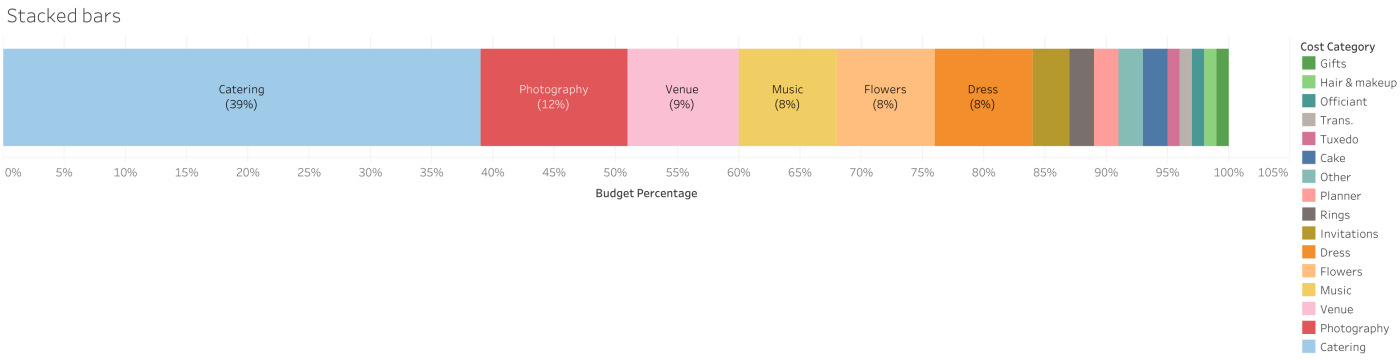
Sketch of the resulting graph



Design schema

Schema	Details
Columns:	SUM(BUDGET_PERCENTAGE)
Rows:	-
Graph type:	Bar
Color:	COST_CATEGORY
Size:	Default
Label:	COST_CATEGORY, SUM(BUDGET_PERCENTAGE)

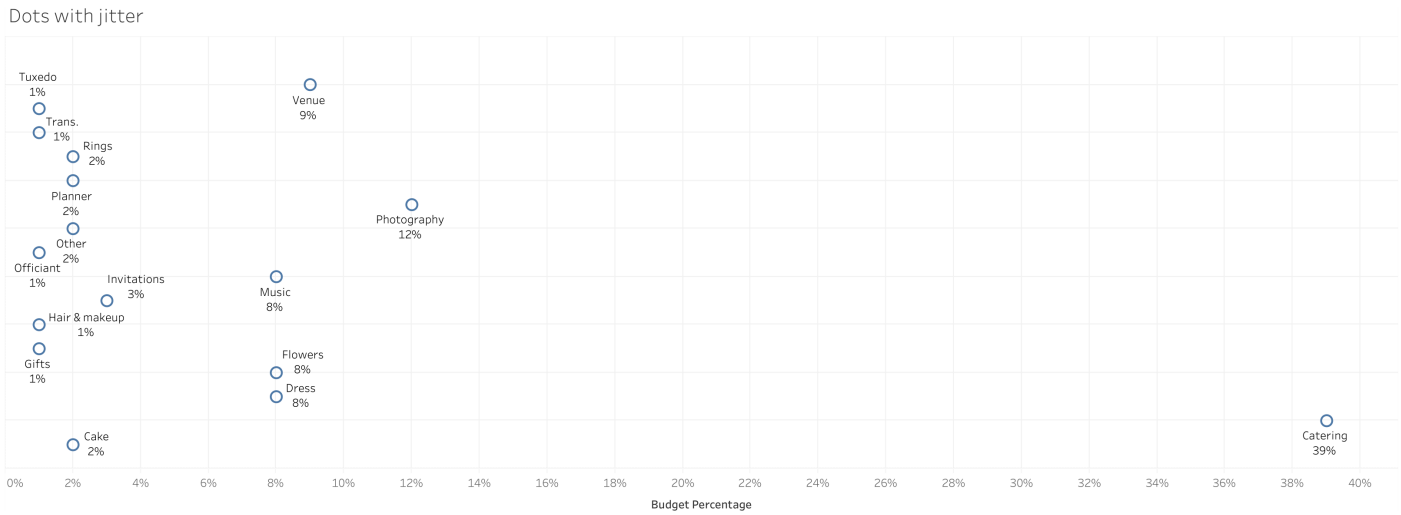
Sketch of the resulting graph



Design schema

Schema	Details
Columns:	SUM(BUDGET_PERCENTAGE)
Rows:	INDEX(COST_CATEGORY)
Graph type:	Shape
Color:	Default
Size:	Bigger
Label:	COST_CATEGORY, SUM(BUDGET_PERCENTAGE)

Sketch of the resulting graph



Theory

Which one of the following examples is **NOT** related to a Gestalt principle?

- the points of a group are enclosed by a fine line
- the color of the legend is similar to the color of the elements of the graph
- the direct labeling technique improves the readability of the visualization
- *the bars representing smaller values are shorter*
- the points of a data series are connected

Data management and visualization

Iniziato	giovedì, 25 febbraio 2021, 07:07
Stato	Completato
Terminato	giovedì, 25 febbraio 2021, 07:08
Tempo impiegato	18 secondi
Valutazione	**0,00** su un massimo di 31,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 1,50

During the ETL process, the correct order of data warehouse loading is:

- (a) update indices, then dimensions, tables, and finally materialized views
- (b) update materialized views, then indices, tables, and finally dimensions
- (c) update dimensions, then tables, finally materialized views and indices
- (d) update tables, then dimensions, and finally materialized views and indices
- (e) update materialized views, then indices, dimensions, and finally tables

Risposta errata.

La risposta corretta è: update dimensions, then tables, finally materialized views and indices

Domanda 2

Risposta non data

Punteggio max.: 1,00

The approximation pattern has the advantage of:

- (a) reduction in the overall number of documents in a collection
- (b) improvement of performance when there are a lot of join operations
- (c) fewer writes to the database
- (d) reduction in the overall size of the working set

Risposta errata.

La risposta corretta è: fewer writes to the database

Domanda 3

Risposta non data

Punteggio max.: 1,00

In the aggregation pipeline, which stage operator is used to execute a recursive search on a collection:

- (a) $group
- (b) $graphLookup
- (c) $project
- (d) $match

Risposta errata.

La risposta corretta è: $graphLookup

Domanda 4

Risposta non data

Punteggio max.: 1,50

Which one of the following visualizations is the most appropriate one for representing a measure as a statistical distribution, a dimension with a high cardinality and another dimension with a low cardinality? For example, think about a visualization representing incomes (measure), level of education (dimension with high cardinality), and gender (dimension with low cardinality).

- (a) Heatmaps
- (b) Multiple box plots
- (c) Gauges
- (d) Stacked bars
- (e) Pie charts

Risposta errata.

La risposta corretta è: Multiple box plots

Domanda 5

Risposta non data

Punteggio max.: 0,50

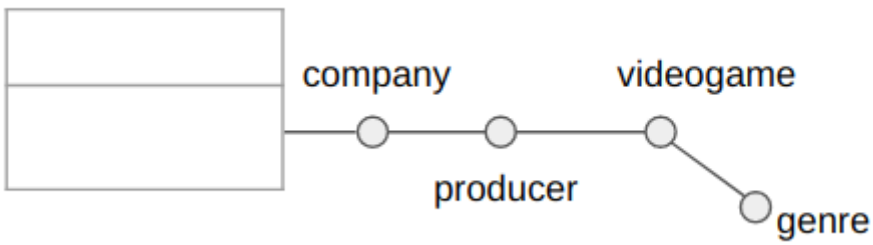
Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.

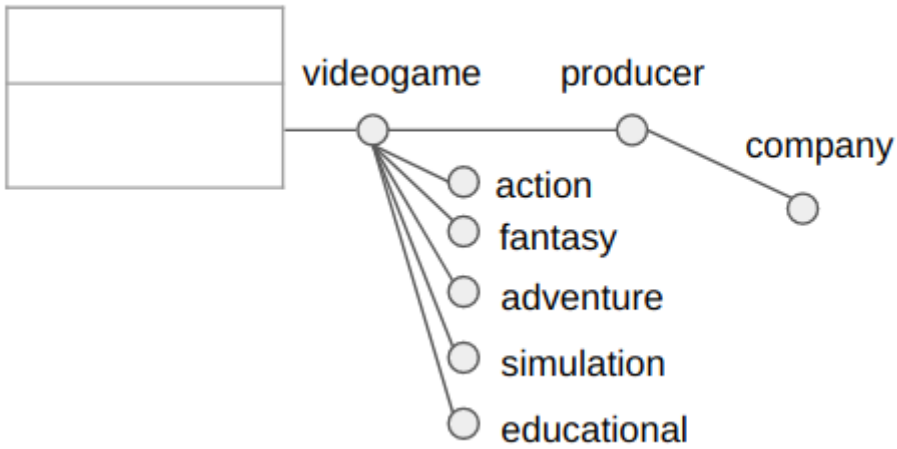
The customer **age group** (18-30, 31-50, >50 years old) is also required.

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

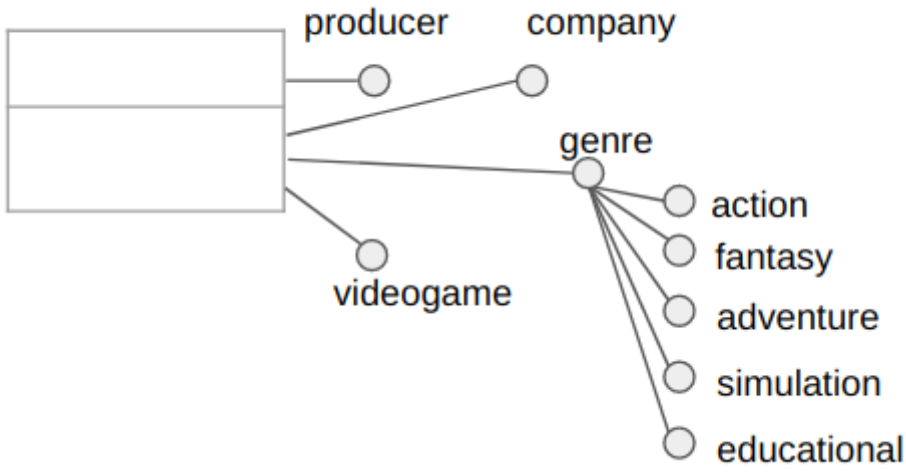
(a)



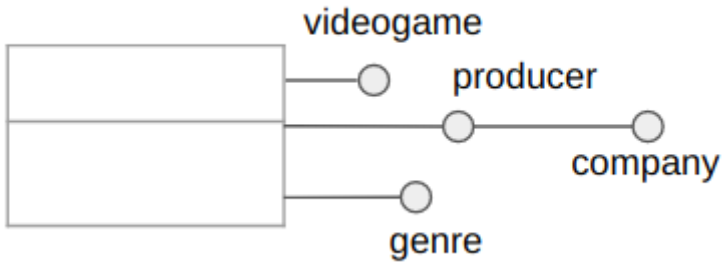
(b)



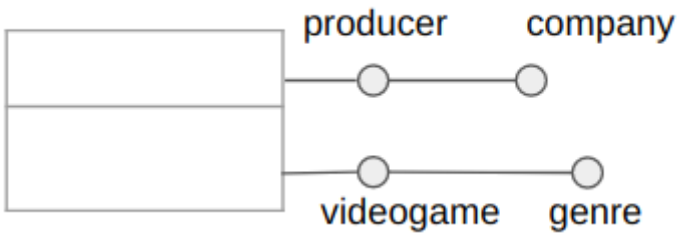
(c)


producer
company
genre
action
fantasy
adventure
simulation
educational
videogame

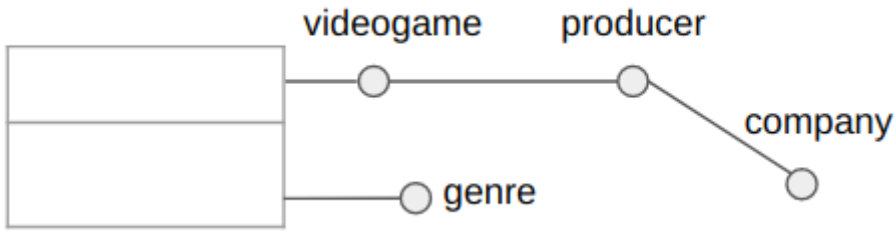
(d)


videogame
producer
company
genre

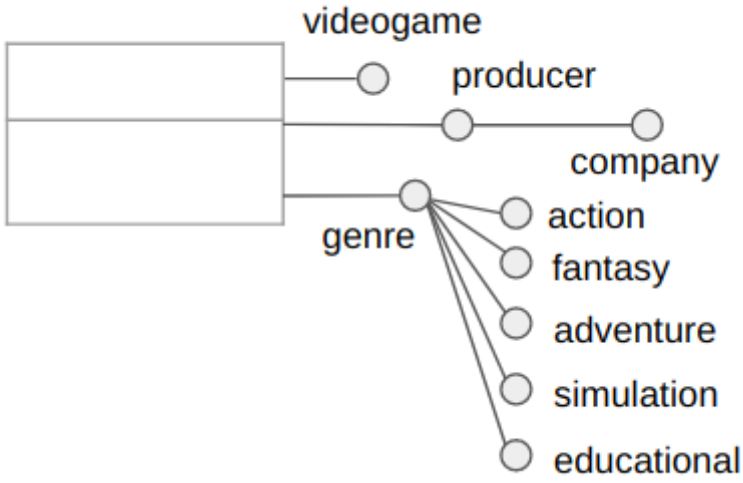
(e)


producer
company
videogame
genre

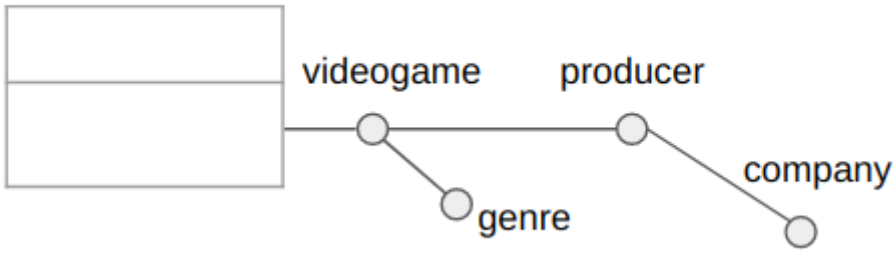
(f)


videogame
producer
company
genre

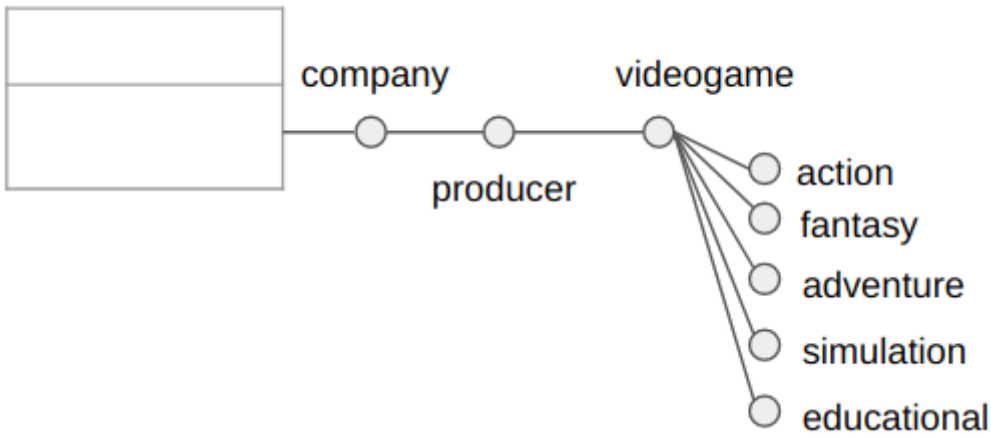
(g)


videogame
producer
company
genre
action
fantasy
adventure
simulation
educational

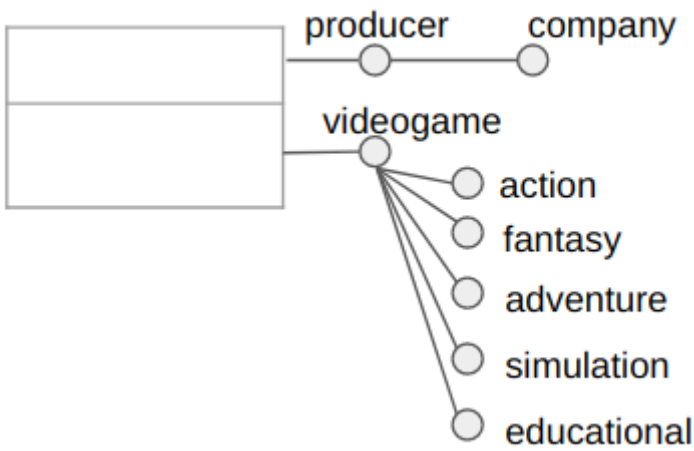
(h)


videogame
producer
company
genre

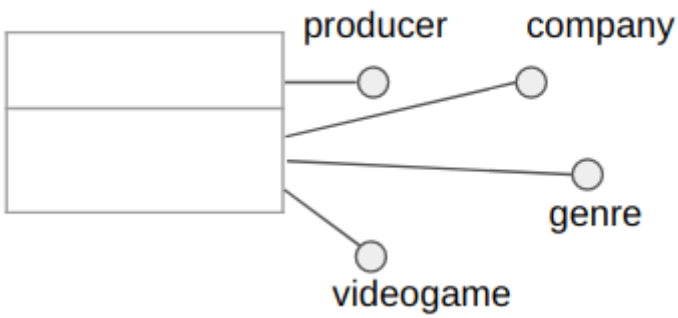
(i)


company
videogame
producer
action
fantasy
adventure
simulation
educational

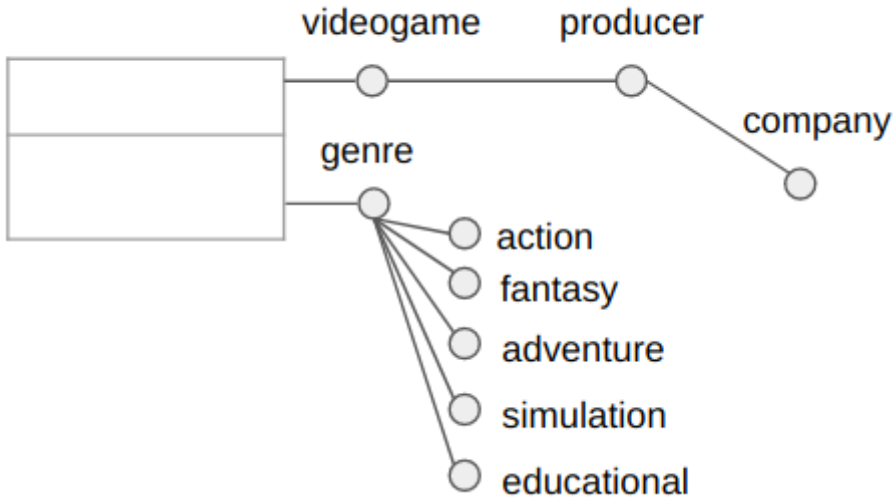
(j)


producer
company
videogame
action
fantasy
adventure
simulation
educational

(k)

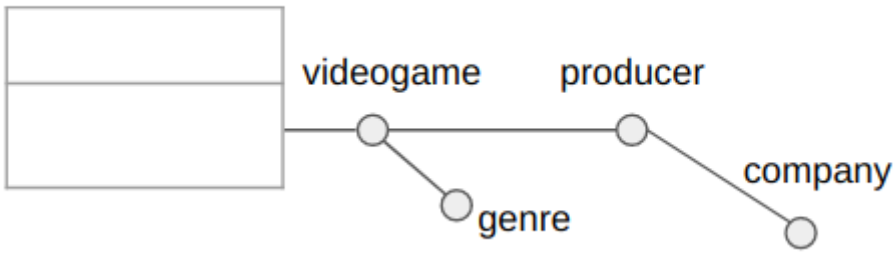

producer
company
genre
videogame

(I)



Risposta errata.

La risposta corretta è:



Domanda 6

Risposta non data

Punteggio max.: 0,50

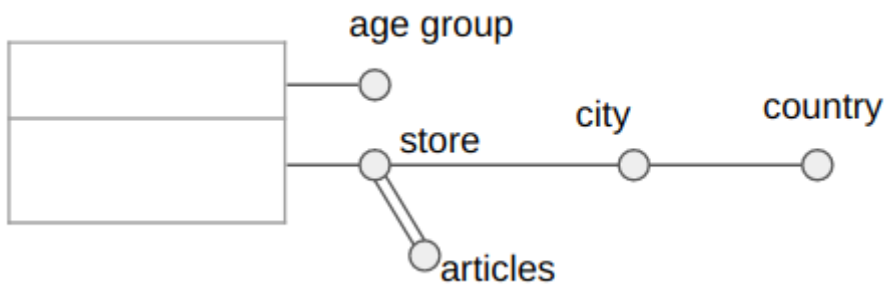
Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.

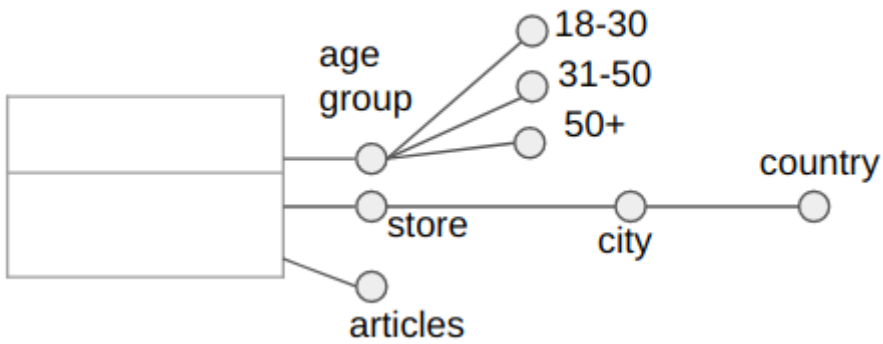
The customer **age group** (18-30, 31-50, >50 years old) is also required.

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

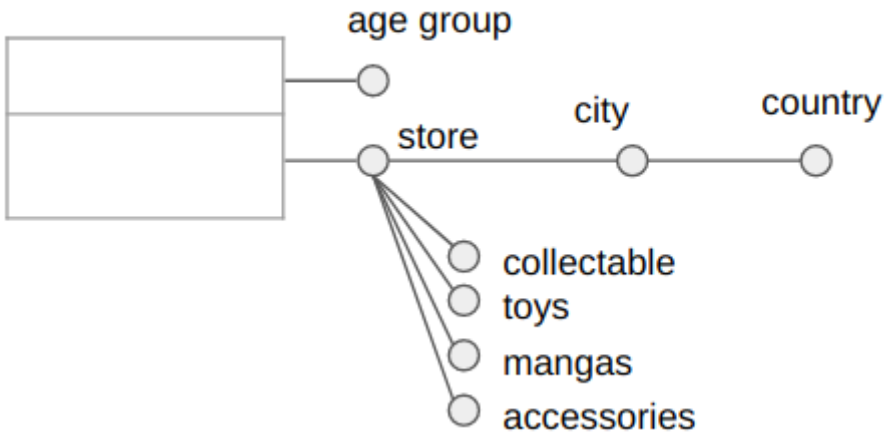
(a)



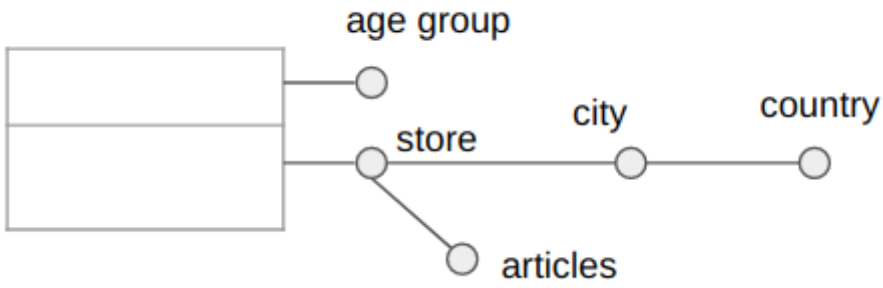
(b)



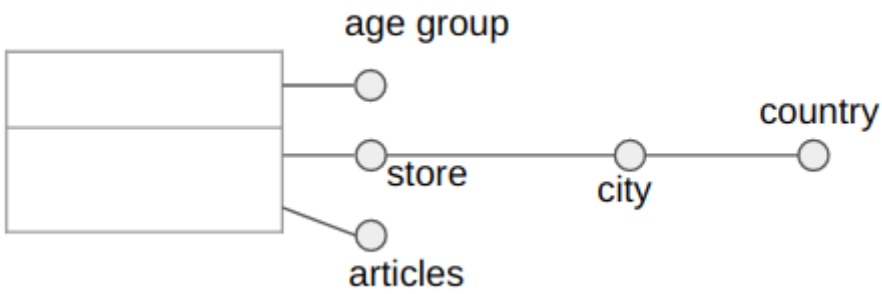
(c)


age group
store
city
country
collectable
toys
mangas
accessories

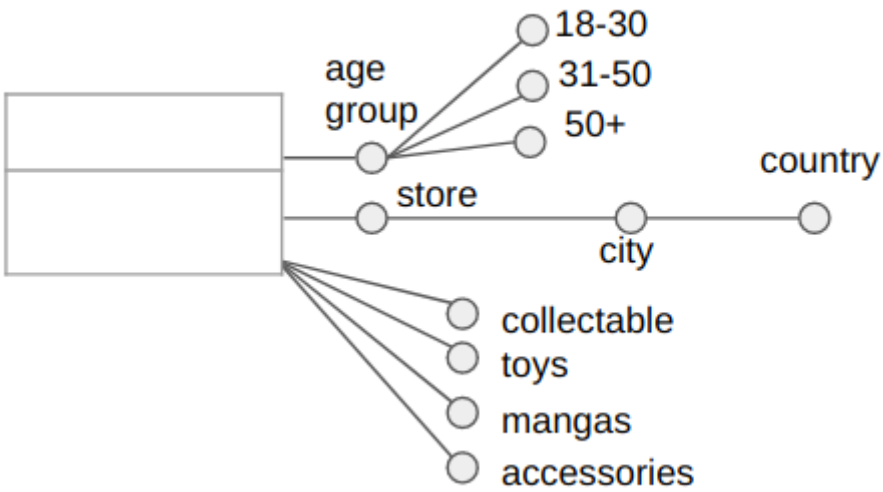
(d)


age group
store
city
country
articles

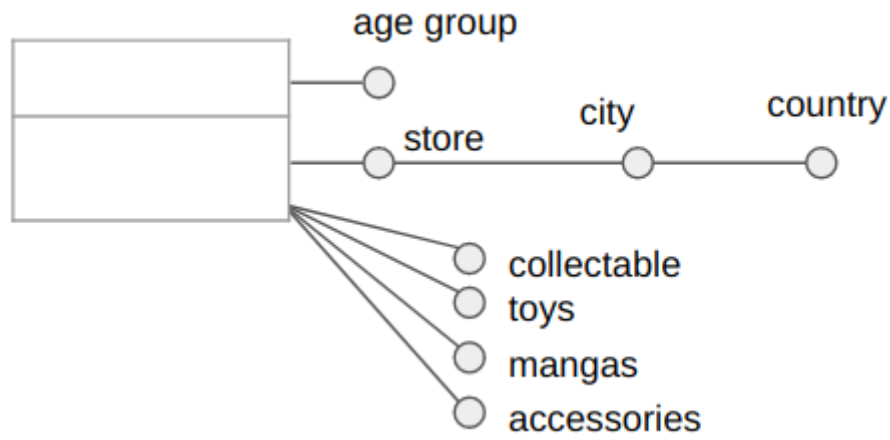
(e)


age group
store
city
country
articles

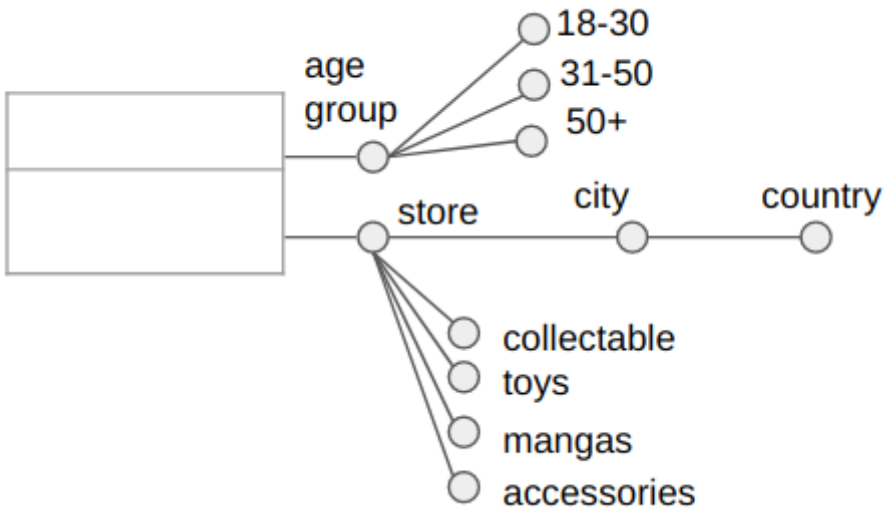
(f)


18-30
31-50
50+
age
group
store
city
country
collectable
toys
mangas
accessories

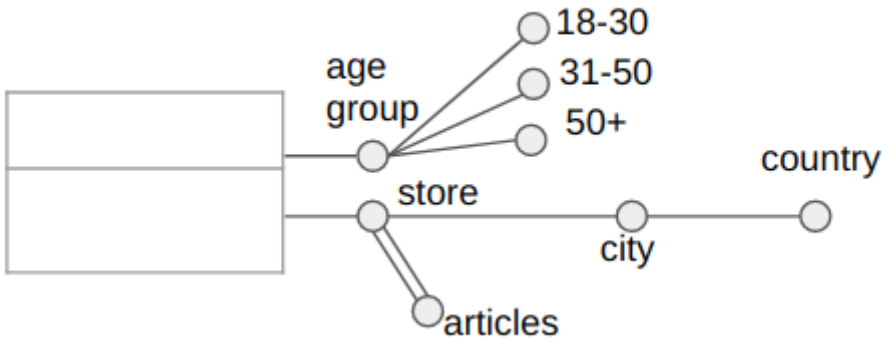
(g)



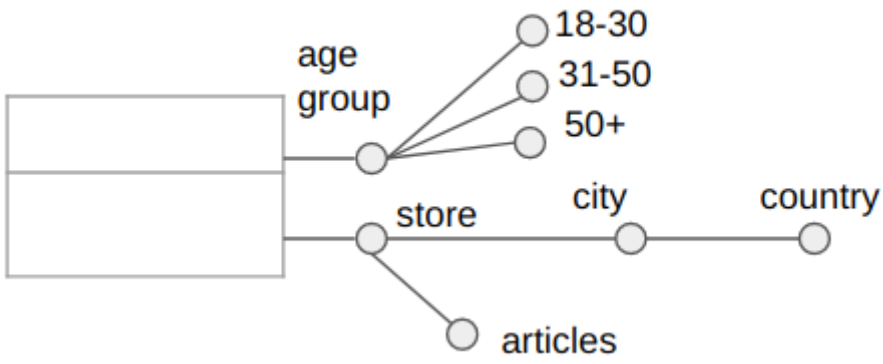
(h)



(i)

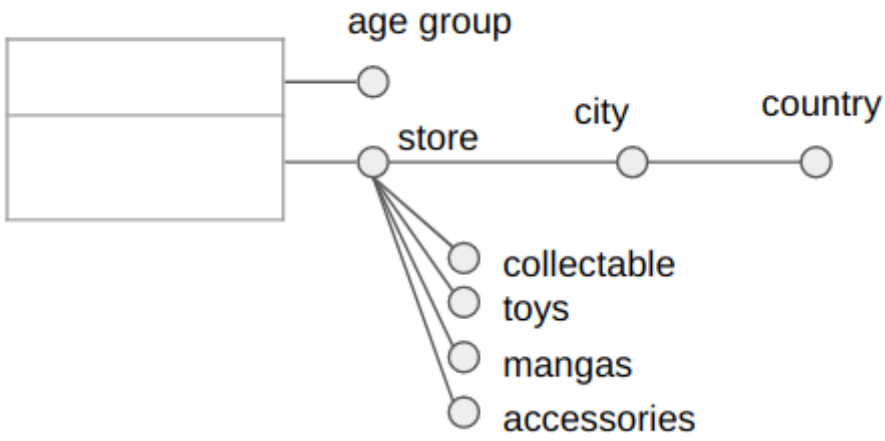


(j)



Risposta errata.

La risposta corretta è:



Domanda 7

Risposta non data

Punteggio max.: 1,00

Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.
- The customer **age group** (18-30, 31-50, >50 years old) is also required.

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) Average revenues per video game
- [ ] (b) Average number of video games
- [ ] (c) Total number of customers
- [ ] (d) Total number of stores
- [ ] (e) Average revenue for each video game purchase
- [ ] (f) Total number of video game purchases
- [ ] (g) Total revenues of the producer
- [ ] (h) Average price of the video game
- [ ] (i) Total number of producers
- [ ] (j) Total age of the customers
- [ ] (k) Total number of different video games
- [ ] (l) Total revenues
- [ ] (m) Total revenues of the store

Risposta errata.

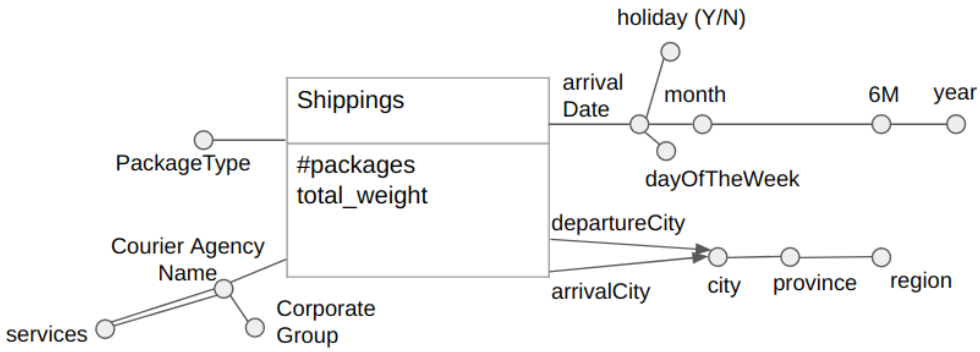
La risposta corretta è: Total revenues, Total number of video game purchases

Domanda 8

Risposta non data

Punteggio max.: 2,00

Given the following conceptual schema:



- For each shipping, the departure and arrival cities, provinces and regions are recorded.
- The cardinality of “PackageType” is 3, and it can be “1” for “Small”, “2” for “Medium” and “3” for “Large”.
- For the shipping, the courier agency is recorded. The courier agency has a unique name. Each agency belongs to a Corporate group.
- An agency may have some additional services available. Examples of additional services are “package tracking”, “package insurance” and “notification by SMS”. The systems records which services are available for each agency. The number of possible additional services is large and growing, hence the full list is not known a prior.
- The system stores the arrival date, the day of the week and if the day was an holiday or not. It also stores the month, year and semester.

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.

```
CourierAgency(CourierAgencyId, CourierAgencyName,
CorporateGroup)
Services(ServiceId, ServiceName)
Agency-HAS-Services(CourierAgencyId, ServiceId)
Location(LocationId, city, province, region)
Time(TimeId, arrivalDate, dayOfTheWeek, holiday, month,
6M, year)
Shippings(CourierAgencyId, TimeId, ArrivalLocationId,
DepartureLocationId, PackageType, #packages,
total_weight)
```

Domanda 9

Risposta non data

Punteggio max.: 4,00

Gardens(TimeId, GardenCenterId, PlantId, numberOfPlants, revenue)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek, holiday)
GardenCenter(GardenCenterId, GardenCenter, city, province, region, greenhouse, accessoryShop, gardenShop, parking)
Plant(PlantId, plantSpecies, genus, family, indoor)

A plant species can be either an indoor or an outdoor plant. A plant species belongs to only one genus and one genus belong to only one family.
A garden center can have 0 or more services. There are 4 available services: "parking", "accessories shop", "gardening shop" and "greenhouse".
`Indoor, greenhouse, accessoryShop, gardenShop,` and `parking` attributes can be `"True"` or `"False"`.

Separately for each plant species and month, compute the following metrics:
A. the daily average number of plants
B. the monthly percentage of the number of plants of the species with respect to the number of plants of the genus
C. the cumulative total number of plants since the beginning of the year

Write the requested SQL query.

```
SELECT plantSpecies, month,
    SUM(numberOfPlants)/COUNT(DISTINCT date) as A,
    100*SUM(numberOfPlants)/SUM(SUM(numberOfPlants)) OVER
(PARTITION BY genus, month) as B,
      SUM(SUM(numberOfPlants)) OVER (
            PARTITION BY plantId, plantSpecies, year
            ORDER BY month
            ROWS UNBOUNDED PRECEDING) as C
FROM Plant P, Time T, Gardens G
WHERE P.PlantId=G.PlantId AND T.Timeid=G.Timeid
GROUP BY plantId, plantSpecies, month, year, genus
```

Domanda 10

Risposta non data

Punteggio max.: 4,00

```
Gardens(TimeId, GardenCenterId, PlantId, numberOfPlants,
revenue)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek,
holiday)
GardenCenter(GardenCenterId, GardenCenter, city, province,
region, greenhouse, accessoryShop, gardenShop, parking)
Plant(PlantId, plantSpecies, genus, family, indoor)
```

A plant species can be either an indoor or an outdoor plant. A plant species belongs to only one genus and one genus belong to only one family.
A garden center can have 0 or more services. There are 4 available services: “parking”, “accessories shop”, “gardening shop” and “greenhouse”.
`Indoor, greenhouse, accessoryShop, gardenShop,` and `parking` attributes can be `“True”` or `“False”`.

Consider only the garden centers having the “parking” service.
Separately for each garden center and plant genus, compute the following metrics:

A. the average revenue per plant
B. the total revenues of the plant family, for each garden center
C. assign a rank to each garden center within its province, based on its total revenues (rank 1st the garden center with the highest revenue in its province for each plant genus)

Write the requested SQL query.

```
SELECT gardenCenter, genus,
    SUM(revenue)/SUM(numberOfPlants) as A,
      SUM(SUM(revenue)) OVER (PARTITION BY family,
gardenCenter) as B,
      RANK() OVER (PARTITION BY province, genus
                   ORDER BY SUM(revenue) DESC) as C
FROM Plant P, Gardens G, GardenCenter GC
WHERE P.PlantId=G.PlantId AND
G.GardenCenterId=GC.GardenCenterId AND parking=True
GROUP BY gardenCenter, genus, family, province
```

Domanda 11

Risposta non data

Punteggio max.: 2,00

Given the following document structure, representing the measurements received by sensors, where each document collects the measures received in one day:

```
{"_id":ObjectId("5553a998e4b02cf7151190b8"),
"start": Date("2021-02-01T00:00:00.000Z"),
"end": Date("2021-02-01T23:00:00.000Z"),
“sensor”:{
        “_id”: 1000,
        "position":{"type":"Point","coordinates":[-47.9,47.6]},
        "elevation":200,
        “city”: “Turin”,
        “country”: “Italy”
},
"temperature":[
        {ts: Date("2021-02-01T00:00:00.000Z"), value: 12},
        {ts: Date("2021-02-01T01:00:00.000Z"), value: 11},
        …
        {ts: Date("2021-02-01T23:00:00.000Z"), value: 9}
],
nTemp: 24,    // total number of elements in the temperature list
sumTemp: 372    // sum of the values of all elements in the temperature list
}
```

Update the document of the sensor with "_id" equal to 1000 by adding a new "temperature" measurement with "value" 16 received at the timestamp "ts" 2021-02-02T01:10:00.000Z.

Also concurrently update the corresponding statistics (i.e., "nTemp" and "sumTemp").

Suppose that the document with "start" attribute equal to "2021-02-02" exists.

N.B. Use the syntax new Date (string) to manage date attributes, e.g., "start": new Date("2021-02-02")

```
db.measures.updateOne(
{
    'sensor._id': 1000,
    'start': new Date("2021-02-02"),
},
{
    $inc:{ nTemp: 1, sumTemp: 16},
    $push: { temperature: { ts: new Date(“2021-02-20 10:00”), value: 16}}
})
```

Domanda 12

Risposta non data

Punteggio max.: 3,00

Given the following document structure, representing the measurements received by sensors, where each document collects the measures received in one day:

```
{"_id":ObjectId("5553a998e4b02cf7151190b8"),
"start": Date("2021-02-01T00:00:00.000Z"),
"end": Date("2021-02-01T23:00:00.000Z"),
“sensor”:{
        “_id”: 1000,
        "position":{"type":"Point","coordinates":[-47.9,47.6]},
        "elevation":200,
        “city”: “Turin”,
        “country”: “Italy”
},
"temperature":[
        {ts: Date("2021-02-01T00:00:00.000Z"), value: 12},
        {ts: Date("2021-02-01T01:00:00.000Z"), value: 11},
        …
        {ts: Date("2021-02-01T23:00:00.000Z"), value: 9}
],
nMeasure: 24,    // total number of elements in the temperature list
totMeasure: 372    // sum of the values of each element in the temperature list
}
```

Considering the sensor located in Italy and the measures received in the month of January 2021, show the sensor id, sensor city and the date in which the average measure of the sensor was greater than or equal to 15.

N.B. Use the syntax new Date (string) to manage date attributes, e.g., new Date("2021-02-01")

```
db.measures.aggregate([
   {$match: {
        “sensor.country”: "Italy",
        “start”: {$gte: new Date (“2021-01-01”)},
        “start”: {$lte: new Date (“2021-01-31”)},
   {$addFields: {
        avg: { $divide: [“$tot”, “$n” ]}
   }},
   { $match : {avg: {$gte: 15}}},
   {$project: { “sensor._id”: 1, “sensor.city”: 1, start: 1}}
])
```

Domanda 13

Risposta non data

Punteggio max.: 4,00

Design a MongoDB database to manage museum exhibitions according to the following requirements.

Museums are characterized by their name, address, a telephone number, and a website (if available). The address consists of geographical coordinates, street name and number, postal code, and city.

The items exhibited in the museums are identified by a progressive number and characterized by a title, a description and the list of author names. The items are categorized as either archaeological finds, or paintings, or sculptures. The database must record all the main features of each item, such as its dimensions (i.e., width, height, weight, etc.). Each feature has at least a name and a value, and possibly a unit of measure. For instance, the main material is a feature of an archaeological find, the geometrical sizes are features of a painting. For each item, the museum to which it belongs must be recorded, with the museum name frequently accessed together with the item itself.

Several exhibitions are hosted in each museum. The exhibition is characterized by a title, a description, the list of curator names. You must record all the items associated with each exhibition, they can be in the order of hundreds. An item can be part of different exhibitions. Moreover, each exhibition can be hosted by several museums in different periods. You must record the start and end dates of each exhibition in each museum.

Given an item, the database must be designed to efficiently provide the name of the museum that owns it.

Given an exhibition, the database must be designed to efficiently provide the name of the museum and the geographical coordinates where it has been hosted.

Furthermore, given an exhibition, the list of items included in the exhibition and their number must be efficiently returned.

Write a sample document for each collection of the database.

Explicitly indicate the design patterns used.

Museum

```
{
_id: ObjectId(),
name: <string>,
address: {
        street: <string>,
        number: <string|number>,
        postal_code: <number>,
        city: <string>,
        province: <string>,
        geo_ref: {type: <string>, coordinates: [ <number> ] }
}
tel: <string>,
website: <string> // optional
}
```

Items

```
{
_id: <number>,
title: <string>,
description:<string>,
authors: [ <string> ],
category: <string>,
features: [ {k: <string>, v: <string>, u: <string>} ],
museum: {
      _id: itemId(),
      name: <string>
      }
}
```

Exhibition

```
{
_id: ObjectId(),
title: <string>,
description:<string>,
curators: [ <string> ],
events: [
        {start: <date>,
        end: <date>,
        museum:{
                _id: itemId(),
                name: <string>,
                geo_ref: {type: <string>, coordinates: [ <number> ] }
        }
],
items: [<number> ] // _id of items
}
```

Pattern used:

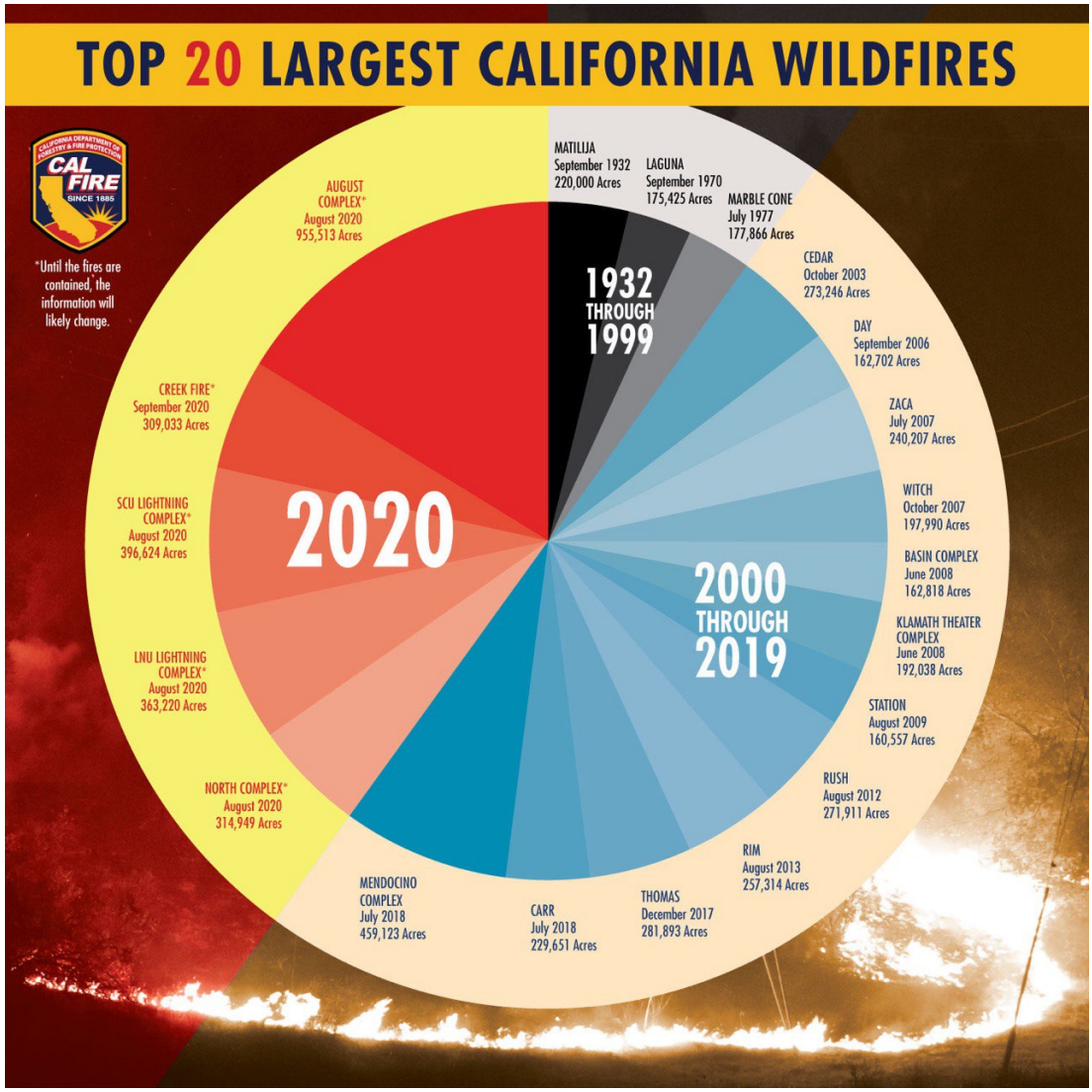
Polymorphic pattern to track the website information in the museum collection.

Attribute pattern (with polymorphic pattern) to track the features of each item.

Extended reference pattern to track the museum information associated with each item.

Bucket pattern with extended reference pattern to track when an exhibition is hosted in a museum.

Informazione



Analyze the above graph illustrating the top 20 largest California wildfires. The visualization was created by CAL FIRE, that is the California Department of Forestry and Fire Protection.

Domanda 14

Risposta non data

Punteggio max.: 0,25

Question

Is there a clearly defined question addressed by the visualization? Write it down.

Domanda 15

Risposta non data

Punteggio max.: 1,25

Data

Is the data quality appropriate? Identify the inadequate characteristics and explain.

Domanda 16

Risposta non data

Punteggio max.: 0,75

Visual Proportionality

Are the values encoded in a uniformly proportional way?

Domanda 17

Risposta non data

Punteggio max.: 0,75

Visual Utility

All the elements in the graph convey useful information?

Domanda 18

Risposta non data

Punteggio max.: 0,50

Visual Clarity

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

Domanda 19

Risposta non data

Punteggio max.: 0,25

Design data

Design the visualization based on the following data structure (to be completed).

Domanda 20

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 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.

/

DM & Visualization - Exam 2021-02-15 - Solution

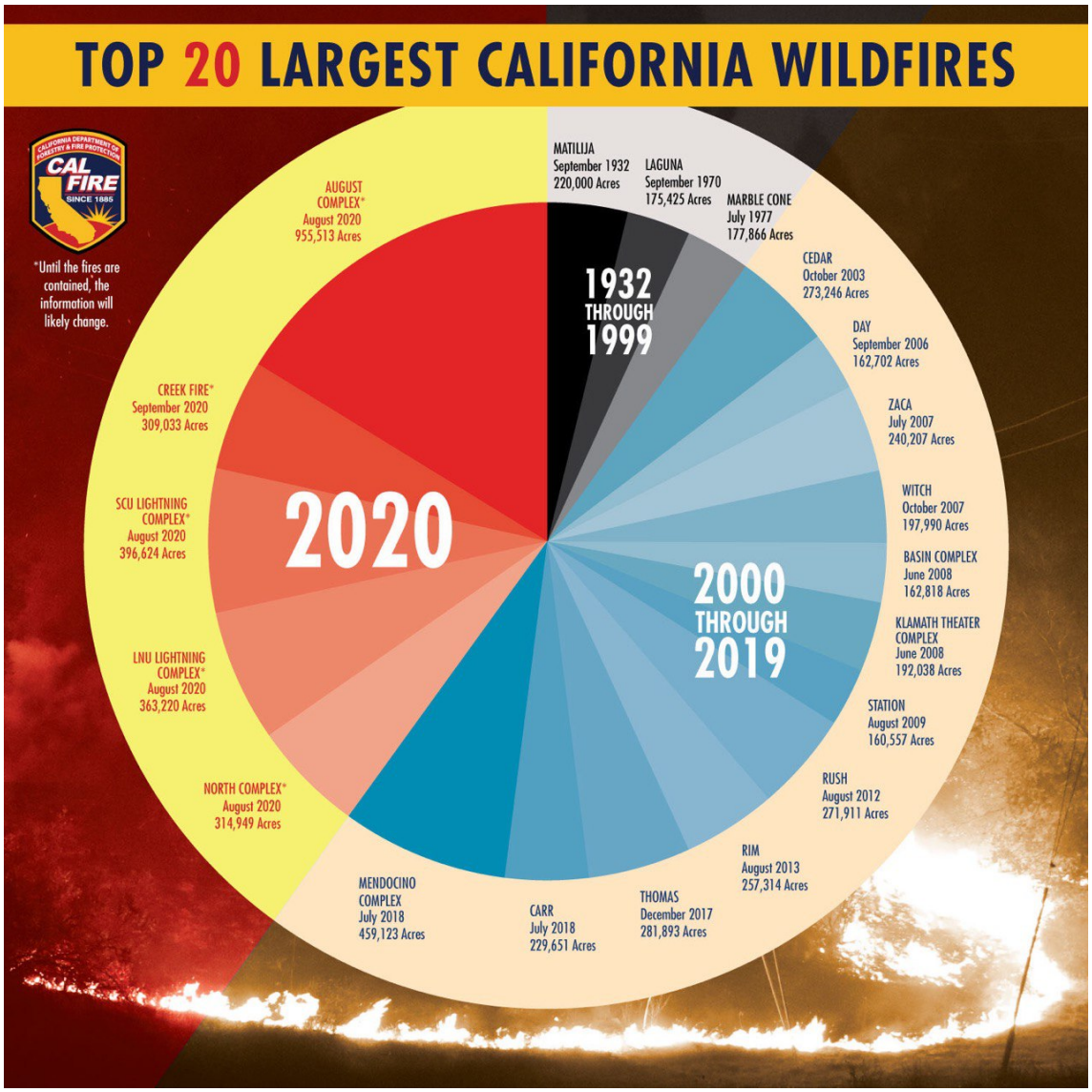


Figure 1: Top 20 largest California wildfires

Analysis

Analyze the above graph illustrating the top 20 largest California wildfires. The visualization was created by CAL FIRE, that is the California Department of Forestry and Fire Protection.

Question: Is there one (or more) question addressed by the visualization?

The question is quite clear: what is the size of wildfires that happened in 2020 in California and how does it compare with the size of wildfires from 2000-2019 and 1932-1999?

Data: Is the data quality appropriate?

Accuracy: the values reported are reasonable, the largest wildfire is about 0.9% the area of California.

Completeness: data are not complete, as only the top 20 wildfires are considered.

Consistency: the three different timeframes are of different lengths. The values of 2020 are estimated.

Currency: data were last updated in September 2020, but it was current when the visualization was created.

Credibility: the source is mentioned in the logo and it is trusted because it is a government agency.

Understandability: data are understandable in the USA, but in general it is better to use square kilometers.

Precision: precision is too detailed, apart the size of the first wildfire.

Visual Proportionality: Are the values encoded in a uniformly proportional way?

The slices of the piechart and the colors of the timeframes are proportional to the size of the wildfires. However, this visualization has serious perceptual problems because it is very difficult to compare areas and shades.

Visual Utility: All the elements in the graph convey useful information?

Several elements are useless: the image in the background, the CAL FIRE logo, the different shades for each timeframe, the donut around the piechart.

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

The usage of direct labeling is appropriate. The choice of piechart is wrong because the values are not part of a whole. The shades for each timeframe are difficult to interpret. It is not clear that only 20 wildfires are considered.

Design

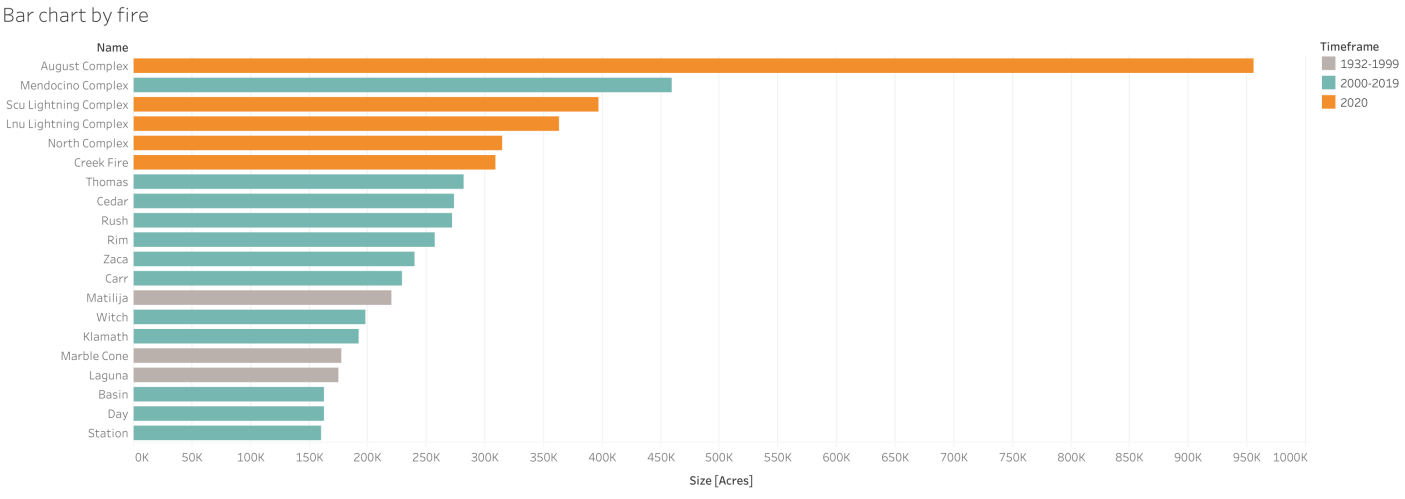
Design the visualization based on the following data structure

Field	Dim./Measure	Description
FIRE_NAME	Dimension	The name of the wildfire
FIRE_SIZE	Measure	The area affected by the wildfire
FIRE_DATE	Dimension	The month and year when wildfire happened
FIRE_TIMEFRAME	Dimension	The three timeframes: 1932-1999, 2000-2019, 2020

Design schema

Schema	Details
Columns:	SUM(FIRE_SIZE)
Rows:	FIRE_NAME
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	Default

Sketch of the resulting graph

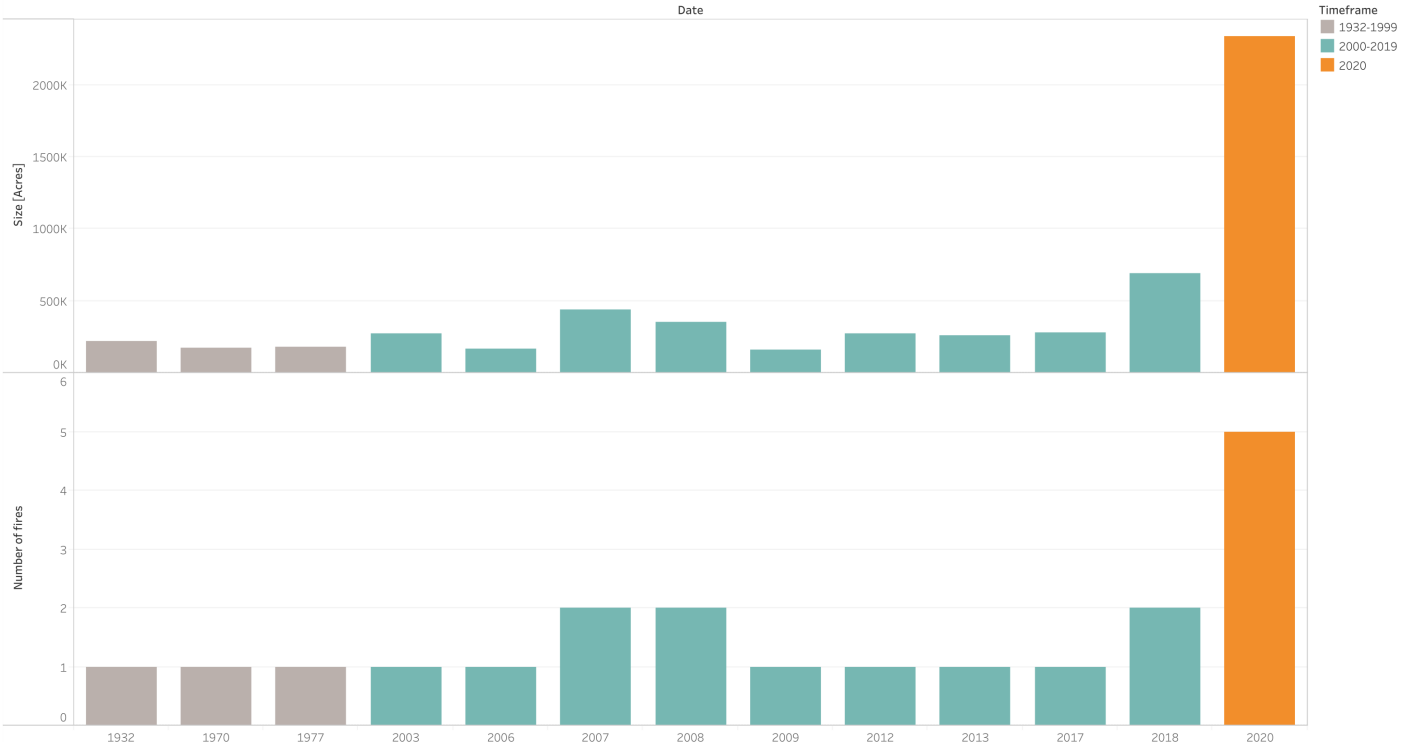


Design schema

Schema	Details
Columns:	YEAR(FIRE_DATE)
Rows:	SUM(FIRE_SIZE), CNT(FIRE_SIZE)
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	Default

Sketch of the resulting graph

Bar chart by year

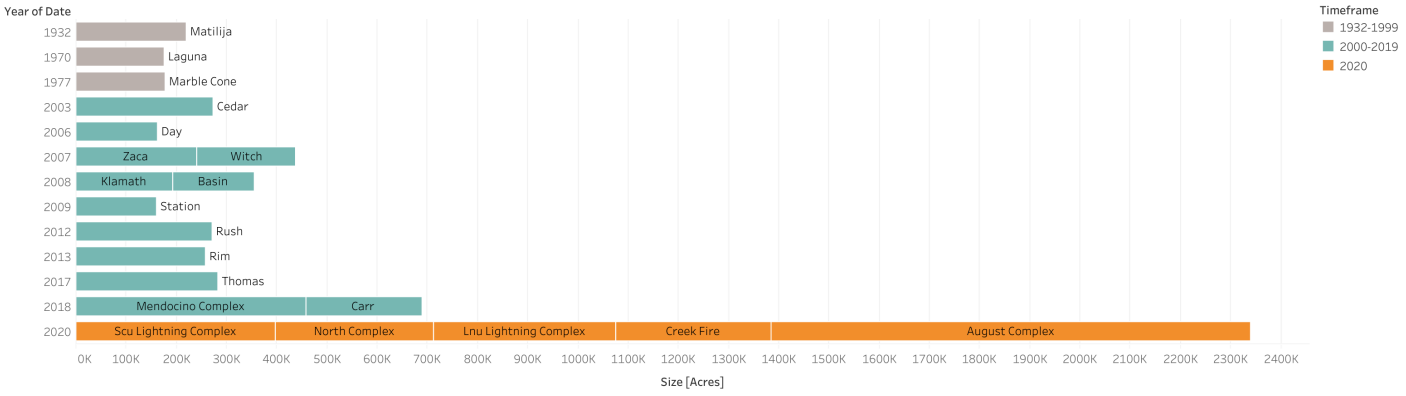


Design schema

Schema	Details
Columns:	SUM(FIRE_SIZE)
Rows:	YEAR(FIRE_DATE)
Graph type:	Bar
Color:	FIRE_TIMEFRAME
Size:	Default
Label:	FIRE_NAME

Sketch of the resulting graph

Stacked bar chart by year

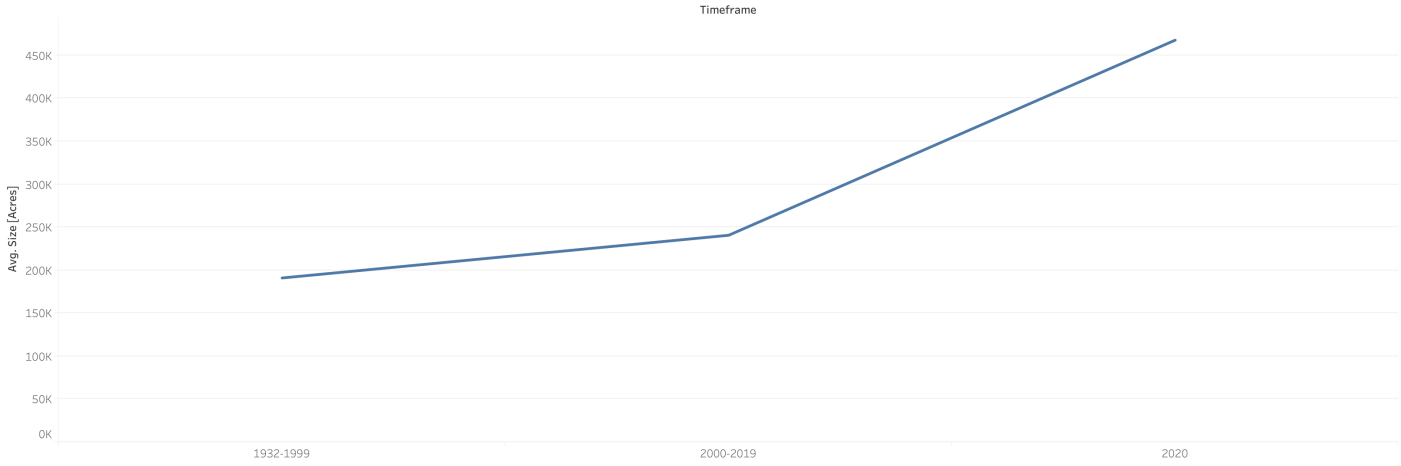


Design schema

Schema	Details
Columns:	FIRE_TIMEFRAME
Rows:	AVG(FIRE_SIZE)
Graph type:	Line
Color:	Default
Size:	Default
Label:	Default

Sketch of the resulting graph

Line chart per avg. size



Theory

Which one of the following visualizations is the most appropriate one for representing a measure as a statistical distribution, a dimension with a high cardinality and another dimension with a low cardinality? For example, think about a visualization representing incomes (measure), level of education (dimension with high cardinality), and gender (dimension with low cardinality).

- Gauges
- *Multiple box plots*
- Stacked bars
- Pie charts
- Heatmaps

Data management and visualization

Iniziato	sabato, 26 giugno 2021, 07:53
Stato	Completato
Terminato	sabato, 26 giugno 2021, 07:53
Tempo impiegato	18 secondi
Valutazione	**0,00** su un massimo di 31,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 1,00

Given a collection of documents, each describing a photo, the statement

```
db.photos.updateMany( {user: "john", tag: "seaside" }, { $addToSet: {tag: "Riccione"}  } ):
```

- (a) removes the tag “Riccione” to one photo belonging to the user “john” and having the value “seaside” in the tag list
- (b) adds the tag “Riccione” to one photo belonging to the user “john” and having the tag field equal to “seaside”
- (c) adds the tag “Riccione” to all the photos belonging to the user “john” and having the value “seaside” in the tag list
- (d) sets the tag field to be equal to “Riccione” to all the photos belonging to the user “john” and having the tag field equal to “seaside”

Risposta errata.

La risposta corretta è: adds the tag “Riccione” to all the photos belonging to the user “john” and having the value “seaside” in the tag list

Domanda 2

Risposta non data

Punteggio max.: 1,00

In the MongoDB aggregation pipeline, which stage operator is used to output a new document for each element of an array:

- (a) $unwind
- (b) $match
- (c) $group
- (d) $foreach
- (e) $project

Risposta errata.

La risposta corretta è: $unwind

Domanda 3

Risposta non data

Punteggio max.: 1,50

In a master-slave distributed database setting, when the replication is asynchronous:

- (a) a failure of the master always causes the data to be lost
- (b) data can be lost only if the majority of the slaves fail
- (c) data can be lost even if the master has already committed
- (d) data cannot be lost if the slaves do not fail

Risposta errata.

La risposta corretta è: data can be lost even if the master has already committed

Domanda 4

Risposta non data

Punteggio max.: 1,50

Which one of the following answers is a direct consequence of Steven's law?

- (a) Ordinal measure should be mapped to increasing saturation and intensity
- (b) It is important to avoid comparisons between areas
- (c) For every single attribute no more than four distinct levels are discernible
- (d) There is no common magnitude assessment for the curvature
- (e) The length of non-aligned objects is harder to compare

Risposta errata.

La risposta corretta è: It is important to avoid comparisons between areas

Domanda 5

Risposta non data

Punteggio max.: 0,50

Data analysts of an italian high-speed train operator are interested in designing a new datawarehouse to analyze some key performance indicators of their train trips.
A trip consists of a specific train travelling from a departure to a destination station, stopping by in different intermediate stations.
In the original database, the start and stop timestamps of each trip are recorded together with the scheduled times.
The new data warehouse must be designed to efficiently analyze

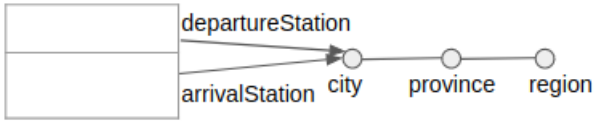
A. the average **duration** of the trips,
B. the average **length** (in km) of the trips,
C. the average number of minutes of **delay** of the trips (measured at the destination station),
D. and the average number of intermediate **stations** of the trips,

according to the following dimensions.

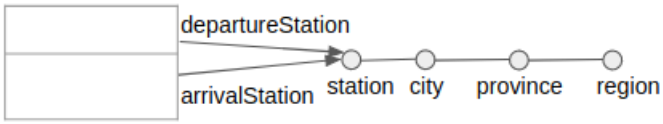
- Departure and destination **station**, the **city** of the station, its **province**, its **region**.
- The **model** of the train.
  - Each model is built by a specific **manufacturer**.
  - A manufacturer can be associated with many train models.
- Each train model offers several **services**. The systems stores which services are available for each train model.
  - Examples of additional services are “bar”, “restaurant”, “wi-fi”, “air conditioning”.
  - The number of additional services is large and growing, hence the full list is not known a priori.
- Each trips is characterized by an **interruption class,** defined as follows.
  - High class with 5 or more stops in intermediate stations
  - Medium class with 2, 3 or 4 stops in intermediate stations
  - Low class with less than 2 stops in intermediate stations

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

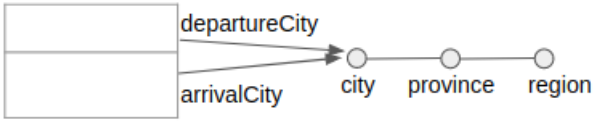
(a)



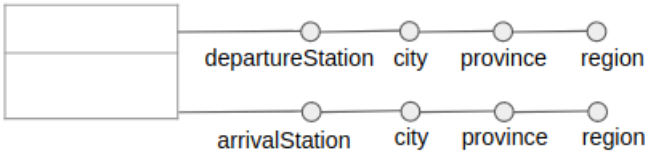
(b)



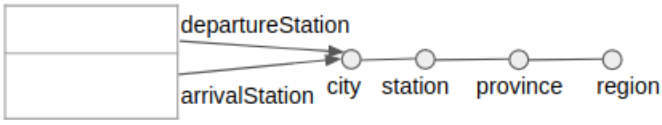
(c)



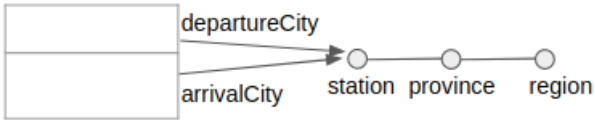
(d)



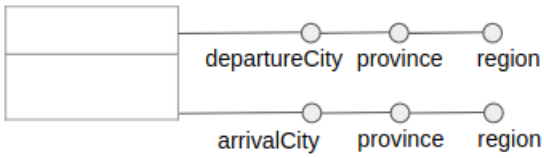
(e)



(f)

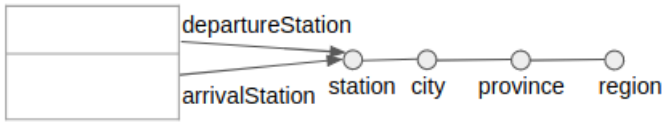


(g)



Risposta errata.

La risposta corretta è:



Domanda 6

Risposta non data

Punteggio max.: 0,50

Data analysts of an italian high-speed train operator are interested in designing a new datawarehouse to analyze some key performance indicators of their train trips.
A trip consists of a specific train travelling from a departure to a destination station, stopping by in different intermediate stations.
In the original database, the start and stop timestamps of each trip are recorded together with the scheduled times.
The new data warehouse must be designed to efficiently analyze

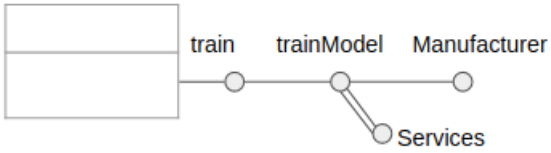
A. the average **duration** of the trips,
B. the average **length** (in km) of the trips,
C. the average number of minutes of **delay** of the trips (measured at the destination station),
D. and the average number of intermediate **stations** of the trips,

according to the following dimensions.

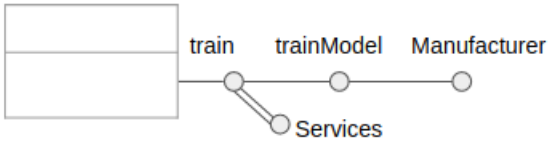
- Departure and destination **station**, the **city** of the station, its **province**, its **region**.
- The **model** of the train.
  - Each model is built by a specific **manufacturer**.
  - A manufacturer can be associated with many train models.
- Each train model offers several **services**. The systems stores which services are available for each train model.
  - Examples of additional services are “bar”, “restaurant”, “wi-fi”, “air conditioning”.
  - The number of additional services is large and growing, hence the full list is not known a priori.
- Each trips is characterized by an **interruption class,** defined as follows.
  - High class with 5 or more stops in intermediate stations
  - Medium class with 2, 3 or 4 stops in intermediate stations
  - Low class with less than 2 stops in intermediate stations

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

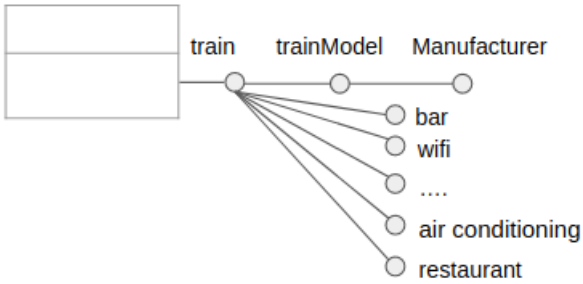
(a)



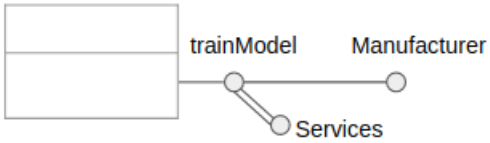
(b)



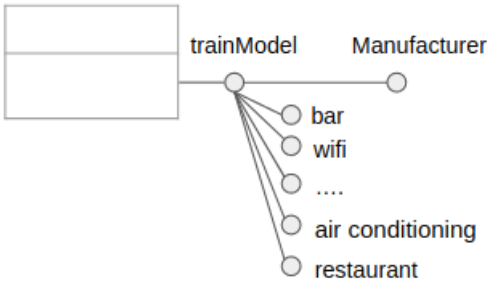
(c)



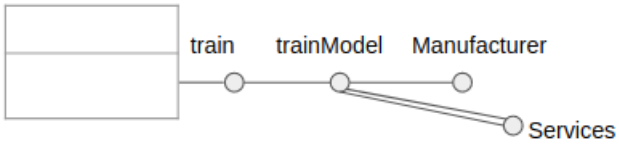
(d)



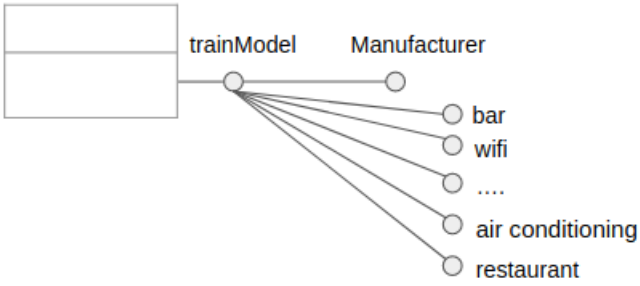
(e)



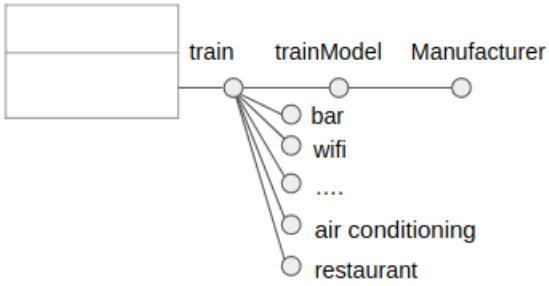
(f)



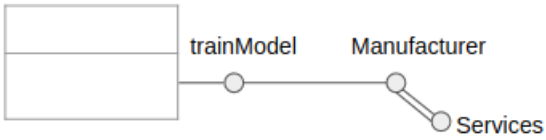
(g)



(h)

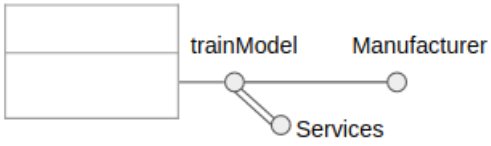


(i)



Risposta errata.

La risposta corretta è:



Domanda 7

Risposta non data

Punteggio max.: 0,50

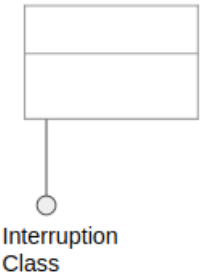
Data analysts of an italian high-speed train operator are interested in designing a new datawarehouse to analyze some key performance indicators of their train trips.
A trip consists of a specific train travelling from a departure to a destination station, stopping by in different intermediate stations.
In the original database, the start and stop timestamps of each trip are recorded together with the scheduled times.
The new data warehouse must be designed to efficiently analyze

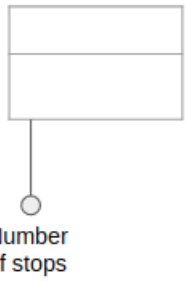
A. the average **duration** of the trips,
B. the average **length** (in km) of the trips,
C. the average number of minutes of **delay** of the trips (measured at the destination station),
D. and the average number of intermediate **stations** of the trips,

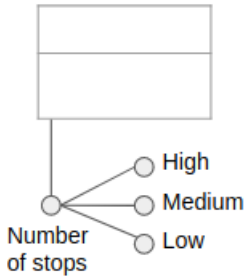
according to the following dimensions.

- Departure and destination **station**, the **city** of the station, its **province**, its **region**.
- The **model** of the train.
  - Each model is built by a specific **manufacturer**.
  - A manufacturer can be associated with many train models.
- Each train model offers several **services**. The systems stores which services are available for each train model.
  - Examples of additional services are “bar”, “restaurant”, “wi-fi”, “air conditioning”.
  - The number of additional services is large and growing, hence the full list is not known a priori.
- Each trips is characterized by an **interruption class,** defined as follows.
  - High class with 5 or more stops in intermediate stations
  - Medium class with 2, 3 or 4 stops in intermediate stations
  - Low class with less than 2 stops in intermediate stations

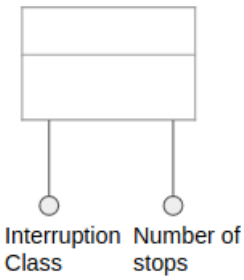
Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

(a) 

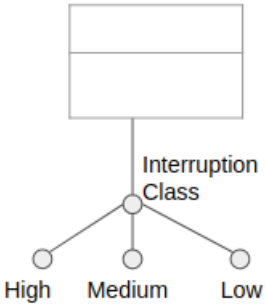
(b) 

(c) 

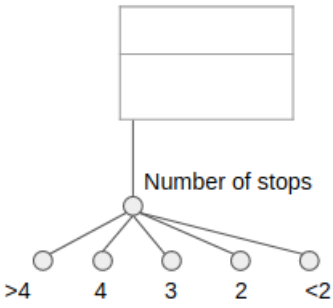
(d)



(e)

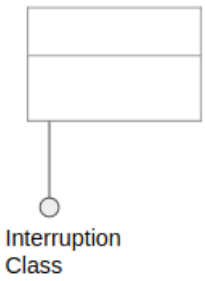


(f)



Risposta errata.

La risposta corretta è:



Domanda 8

Risposta non data

Punteggio max.: 1,00

Data analysts of an italian high-speed train operator are interested in designing a new datawarehouse to analyze some key performance indicators of their train trips.
A trip consists of a specific train travelling from a departure to a destination station, stopping by in different intermediate stations.
In the original database, the start and stop timestamps of each trip are recorded together with the scheduled times.
The new data warehouse must be designed to efficiently analyze

A. the average **duration** of the trips,
B. the average **length** (in km) of the trips,
C. the average number of minutes of **delay** of the trips (measured at the destination station),
D. and the average number of intermediate **stations** of the trips,

according to the following dimensions.

- Departure and destination **station**, the **city** of the station, its **province**, its **region**.
- The **model** of the train.
  - Each model is built by a specific **manufacturer**.
  - A manufacturer can be associated with many train models.
- Each train model offers several **services**. The systems stores which services are available for each train model.
  - Examples of additional services are “bar”, “restaurant”, “wi-fi”, “air conditioning”.
  - The number of additional services is large and growing, hence the full list is not known a priori.
- Each trips is characterized by an **interruption class,** defined as follows.
  - High class with 5 or more stops in intermediate stations
  - Medium class with 2, 3 or 4 stops in intermediate stations
  - Low class with less than 2 stops in intermediate stations

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) Average delay per destination station (minutes)
- [ ] (b) Total duration of trips (minutes)
- [ ] (c) Total delay of the trips (minutes)
- [ ] (d) Total number of trips (count)
- [ ] (e) Total number of intermediate stations of the trips (count)
- [ ] (f) Average length per trip (km)
- [ ] (g) Total number of train models (count)
- [ ] (h) Average number of intermediate stations per trip (count)
- [ ] (i) Average duration per trip (minutes)
- [ ] (j) Total number of departure stations per trip (count)
- [ ] (k) Total number of destination stations per trip (count)
- [ ] (l) Total length of the trips (km)
- [ ] (m) Average delay per trip (minutes)
- [ ] (n) Number of services (count)
- [ ] (o) Average number of trips (count)
- [ ] (p) Total number of trains (count)

Risposta errata.

La risposta corretta è: Total number of trips (count),
Total duration of trips (minutes)
,

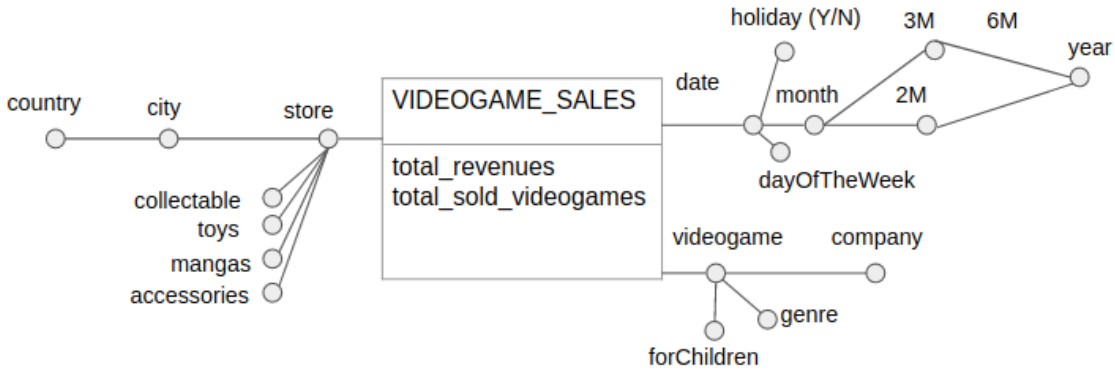
Total length of the trips (km)
, Total delay of the trips (minutes),
Total number of intermediate stations of the trips (count)

Domanda 9

Risposta non data

Punteggio max.: 1,50

Given the following conceptual schema:



- A video game has a unique name and it is distributed by a video game company.
- Each video game has a specific genre.
- A videogame can be appropriate for children or not.
  - The value of this field can be “0” for not appropriate and “1” for appropriate.
- A store is identified by a unique name. Stores are analyzed according to their city and country.
- Each store may sell some additional articles. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”.
- The system records the sales with their date, the day of the week and if the day was an holiday or not. It also records the month, year, bimester, trimester and semester of the sales.

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.

VideoGame(**CodV**, VideoGameName, forChildren, Genre, Company)
Store(**CodS,** Store, City, Country, Collectable, Toys, Mangas, Accessories)
Time(**CodT**, date, dayOfTheWeek, holiday, month, 2M, 3M, 6M, year)
Fact(**CodV**, **CodS**, **CodT**, total_revenues, total_sold_videogames)

Domanda 10

Risposta non data

Punteggio max.: 4,00

```
CourierAgency(CourierAgencyId, CourierAgencyName, CorporateGroup)
Location(LocationId, city, province, region)
Time(TimeId, arrivalDate, dayOfTheWeek, holiday, month, 6M, year)
Shippings(CourierAgencyId, TimeId, ArrivalLocationId, DepartureLocationId,
#packages, total_weight)
```

- For each shipping, the departure and arrival cities, provinces and regions are recorded.
- For the shipping, the courier agency is recorded. The courier agency has a unique name. Each agency belongs to a Corporate group.
- The system stores the arrival date, the day of the week and if the day was an holiday or not. It also stores the month, year and semester.

Separately for each courier **agency** and departure **city**, compute the following metrics:

A. the percentage of packages with respect to the total number of packages of the agency for the departure region
B. the average weight per package
C. assign a rank to each courier agency within its corporate group, based on its total number of packages (rank 1st the courier agency with the highest number of shipped packages in its corporate group for each departure city)

```
SELECT CourierAgencyName,  L.city
    100*SUM(#packages)/SUM(SUM(#packages))
                  OVER (PARTITION BY L.region, CourierAgencyId) as B,
    SUM(total_weight)/SUM(#packages) as A,
      RANK() OVER (PARTITION BY L.city, CorporateGroup
                   ORDER BY SUM(#packages) DESC) as C
FROM CourierAgency CA, Location L, Shippings S
WHERE CA.CourierAgencyId=s.CourierAgencyId and S.DepartureLocationId=L.LocationId
GROUP BY CourierAgencyId, CourierAgencyName, L.city, L.region, CorporateGroup
```

Domanda 11

Risposta non data

Punteggio max.: 4,00

```
CourierAgency(CourierAgencyId, CourierAgencyName, CorporateGroup)
Location(LocationId, city, province, region)
Time(TimeId, arrivalDate, dayOfTheWeek, holiday, month, 6M, year)
Shippings(CourierAgencyId, TimeId, ArrivalLocationId, DepartureLocationId,
#packages, total_weight)
```

- For each shipping, the departure and arrival cities, provinces and regions are recorded.
- For the shipping, the courier agency is recorded. The courier agency has a unique name. Each agency belongs to a Corporate group.
- The system stores the arrival date, the day of the week and if the day was an holiday or not. It also stores the month, year and semester.

Separately for each **month**, departure **province** and arrival **province**, compute the following metrics:

A. the daily average number of shipped packages
B. the cumulative total weight of delivered packets since the beginning of the semester

---

```
SELECT month, L1.province, L2.province, (6M),
      SUM(#packages)/COUNT(distinct date) as A,
    SUM(SUM(total_weight)) OVER (
                PARTITION BY L1.province, L2.province, 6M
                ORDER BY month
                ROWS UNBOUNDED PRECEDING) as B,
FROM Location L1, Location L2, Shippings S, Time T
WHERE S.DepartureLocationId=L1.LocationId and S.ArrivalLocationId=L2.LocationId and
T.TimeId=S.TimeId
GROUP BY month, L1.province, L2.province, 6M
```

Domanda 12

Risposta non data

Punteggio max.: 2,00

The following document structure represents cameras sold by an e-commerce.

Each document collects the aggregated metrics of one day.

```
{"_id": "nikon_d3500",
"model": "D3500",
"brand":{
        "name": "Nikon",
        "url": "https://www.nikon.it/"
}
"releaseDate": Date("2018-08-28T00:00:00.000Z"),
"category": "DSRL",
"price": 435,
"specs":{
        "resolution": 24,
"technology": "APS-C CMOS",
"min_ISO": 100,
"max_ISO": 25600,
"weight": 365,
"viewfinder": "optical",
"video_resolution": "1920 x 1080"
},
"scores": {
        "overall": 57,
        "image_quality": 48,
"versatility":  62,
"comfort": 85,
"speed": 41
}
}
```

Write a MongoDB query to display only the model, the price, and the brand name of cameras released in 2021, belonging to the “laser” category, and whose overall score is in the 70-90 range.

N.B. Use the syntax new Date (string) to manage date attributes, e.g., "start": new Date("2021-06-01")

```
db.cameras.find(
{
        category: ‘laser',
        releaseDate: {
                $gte: new Date('2021-01-01'),
                $lt: new Date('2022-01-01')
                },
        ‘scores.overall’: {
                $gte: 70,
                $lte: 90
                }
},
{model:1, "brand.name":1, price:1, _id:0}
)
```

Domanda 13

Risposta non data

Punteggio max.: 3,00

The following document structure represents cameras sold by an e-commerce.

Each document collects the aggregated metrics of one day.

```
{"_id": "nikon_d3500",
"model": "D3500",
"brand":{
        "name": "Nikon",
        "url": "https://www.nikon.it/"
}
"releaseDate": Date("2018-08-28T00:00:00.000Z"),
"category": "DSRL",
"price": 435,
"specs":{
        "resolution": 24,
"technology": "APS-C CMOS",
"min_ISO": 100,
"max_ISO": 25600,
"weight": 365,
"viewfinder": "optical",
"video_resolution": "1920 x 1080"
},
"scores": {
        "overall": 57,
        "image_quality": 48,
"versatility":  62,
"comfort": 85,
"speed": 41
}
}
```

Considering only cameras released since 2015, for each release year and for each category, select the median overall score.

N.B. Use the operator $year to extract the year from the date, e.g., $year: “$releaseDate”

Use the syntax new Date (string) to manage date attributes, e.g., "start": new Date("2021-06-01")

```
db.measures.aggregate([
{$match: {releaseDate: {$gte: new Date('2015-01-01')} } },
{$sort:
        {'$scores.overall': 1}
},
{$group:
        {
        '_id':      {
                'cat': '$category',
                'y': { $year: "$releaseDate" }
                },
        'value':
                {'$push': '$scores.overall'}
        }
},
{$project:
        {
        _id: 1,
        "median": {
                $arrayElemAt: ["$value", {
                        $floor: {
                                $multiply:
                                [0.50, {$size: "$value"}]
                                }
                        }]
         }
])
```

Domanda 14

Risposta non data

Punteggio max.: 4,00

Design a MongoDB database to manage online courses according to the following requirements.

Teachers are characterized by their name, surname, email, and list of subjects they can teach (e.g., maths, electronics, etc.). Each teacher can have one or more online profiles on different platforms (e.g., Facebook, LinkedIn, Wikipedia, etc.). For each online profile, if available, the database tracks the corresponding URL of the profile (e.g., https://en.wikipedia.org/wiki/Ranjitsinh_Disale). Note that for each teacher and each platform, at most one profile can exist. A teacher can teach different courses.

The courses are characterized by a name, a syllabus, a list of keywords, and the teacher. Each course has several editions. For each edition, the start date, the end date, and the number of enrolled students are known.

Given a course, the database must be designed to efficiently provide the name, the surname and the email of its teacher.

Furthermore, given a course, the number of editions and the average number of enrolled students in each edition must be efficiently returned.

Teachers are typically retrieved by subject (e.g., all those teaching maths), and by online profile platform (e.g., all those having a wikipedia page).

Write a sample document for each collection of the database.

Explicitly indicate the design patterns used.

Teacher

```
{
_id: ObjectId(),
name: <string>,
surname: <string>,
email: <string>,
profiles: {
        facebook: <url>,
        linkedin: <url>,
        ….
}
subjects: [ <string>]
}
```

Course

```
{
_id: ObjectId(),
name: <string>,
syllabus:<string>,
keywords: [ <string> ],
teacher: {
        _id: ObjectId(),
name: <string>,
surname: <string>,
email: <string>,
}
editions: [
{start: <date>,
end: <date>,
n_students: <number>
]
n_editions: <number>,
tot_students: <number>
}
```

Pattern used:

Polymorphic pattern to track the online profile information in the Teacher collection.

Extended reference pattern to track the teacher information associated with each course.

Bucket pattern to track when a course is provided.

Computed pattern for average students on each edition.

Domanda 15

Risposta non data

Punteggio max.: 0,25

Question

Is there a clearly defined question addressed by the visualization? Write it down.

Domanda 16

Risposta non data

Punteggio max.: 1,25

Data

Is the data quality appropriate? Identify the inadequate characteristics and explain.

Domanda 17

Risposta non data

Punteggio max.: 0,75

Visual Proportionality

Are the values encoded in a uniformly proportional way?

Domanda 18

Risposta non data

Punteggio max.: 0,75

Visual Utility

All the elements in the graph convey useful information?

Domanda 19

Risposta non data

Punteggio max.: 0,50

Visual Clarity

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

Domanda 20

Risposta non data

Punteggio max.: 0,25

Design data

Design the visualization based on the following data structure (to be completed).

Domanda 21

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 22

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.

DM & Visualization - Exam 2021-06-17 - Solution

Figure 1: The top 10 football clubs by market value

Analysis

Analyze the above graph illustrating the top 10 football clubs by market value.

Question: Is there one (or more) question addressed by the visualization?

The question is quite clear: what are the top 10 football clubs by market value and how does it compare with their annual revenue and with their agreement to join the European Super League?

Data: Is the data quality appropriate?

Accuracy: the values reported are reasonable, it is not clear how *current value* is computed.

Completeness: data are not complete, as only the top 10 football clubs are considered.

Consistency: data should be consistent if *current value* is computed with the same methodology for all clubs.

Currency: data were probably updated in 2020, but this is not reported in the visualization.

Credibility: the source is mentioned in the logo and it is a well-known and trusted magazine about finance.

Understandability: data are quite understandable, it is not clear that the revenue is annual.

Precision: precision is appropriate, but it varies among different clubs (from 0 to 2 decimal digits).

Visual Proportionality: Are the values encoded in a uniformly proportional way?

The football balls look proportional to the corresponding values. However, this visualization has serious perceptual problems because it is very difficult to compare areas represented by football balls. The balls are not aligned and they are also divided to represent two values.

Visual Utility: All the elements in the graph convey useful information?

Several elements are useless: the football field in the background, the balls, the stars, the logos.

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

Data are understandable because the numerical values are reported. It is very difficult to compare these values because they are not properly aligned. The number of stars has no real meaning. The rank is very clear because each football club is associated with a number.

Design

Design the visualization based on the following data structure

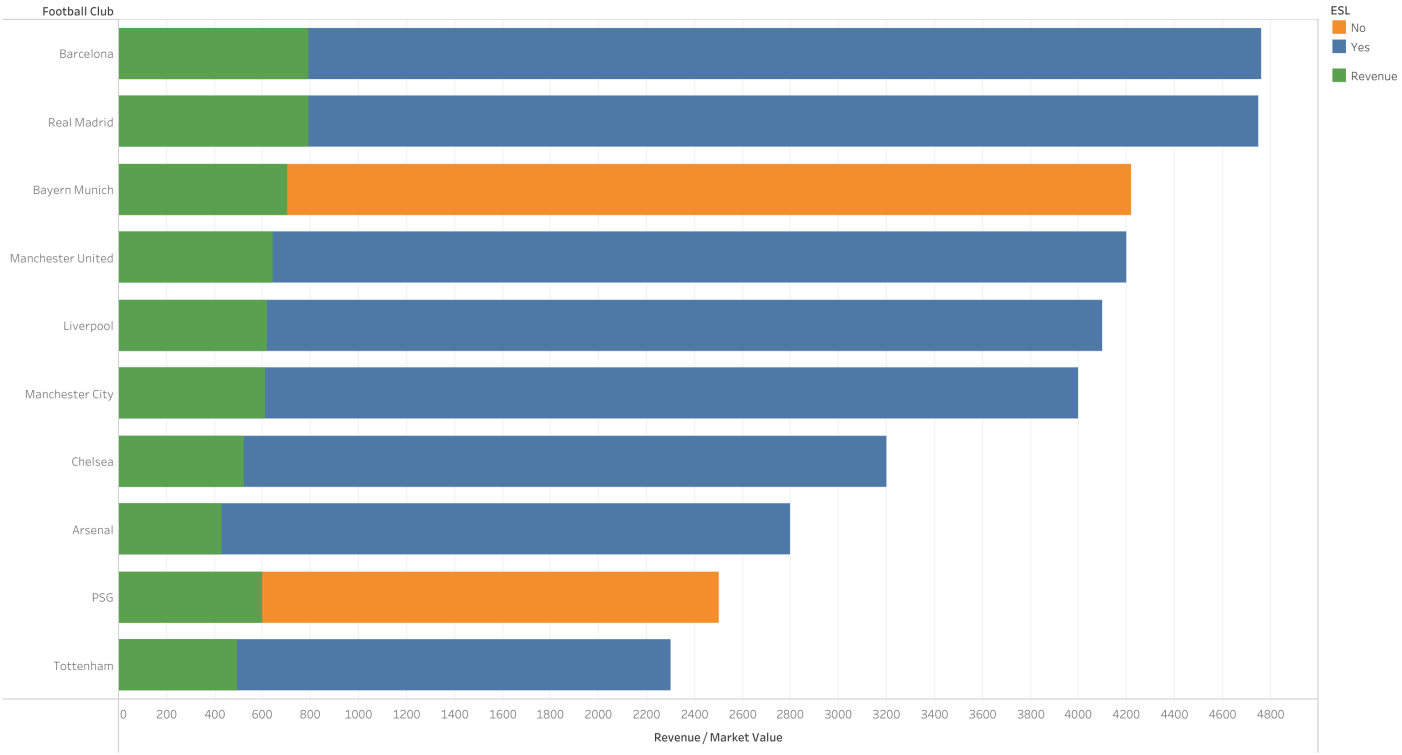
Field	Dim./Measure	Description
CLUB_NAME	Dimension	The name of the football clubs
CURRENT_VALUE	Measure	The current value of the club
REVENUE	Measure	The annual revenue of the club
SUPER_LEAGUE	Dimension	If the club agreed to join European Super League

Design schema

Schema	Details
Columns:	SUM(CURRENT_VALUE), SUM(REVENUE)
Rows:	CLUB_NAME
Graph type:	Bar with double axis
Color:	SUPER_LEAGUE
Size:	Default
Label:	Default

Sketch of the resulting graph

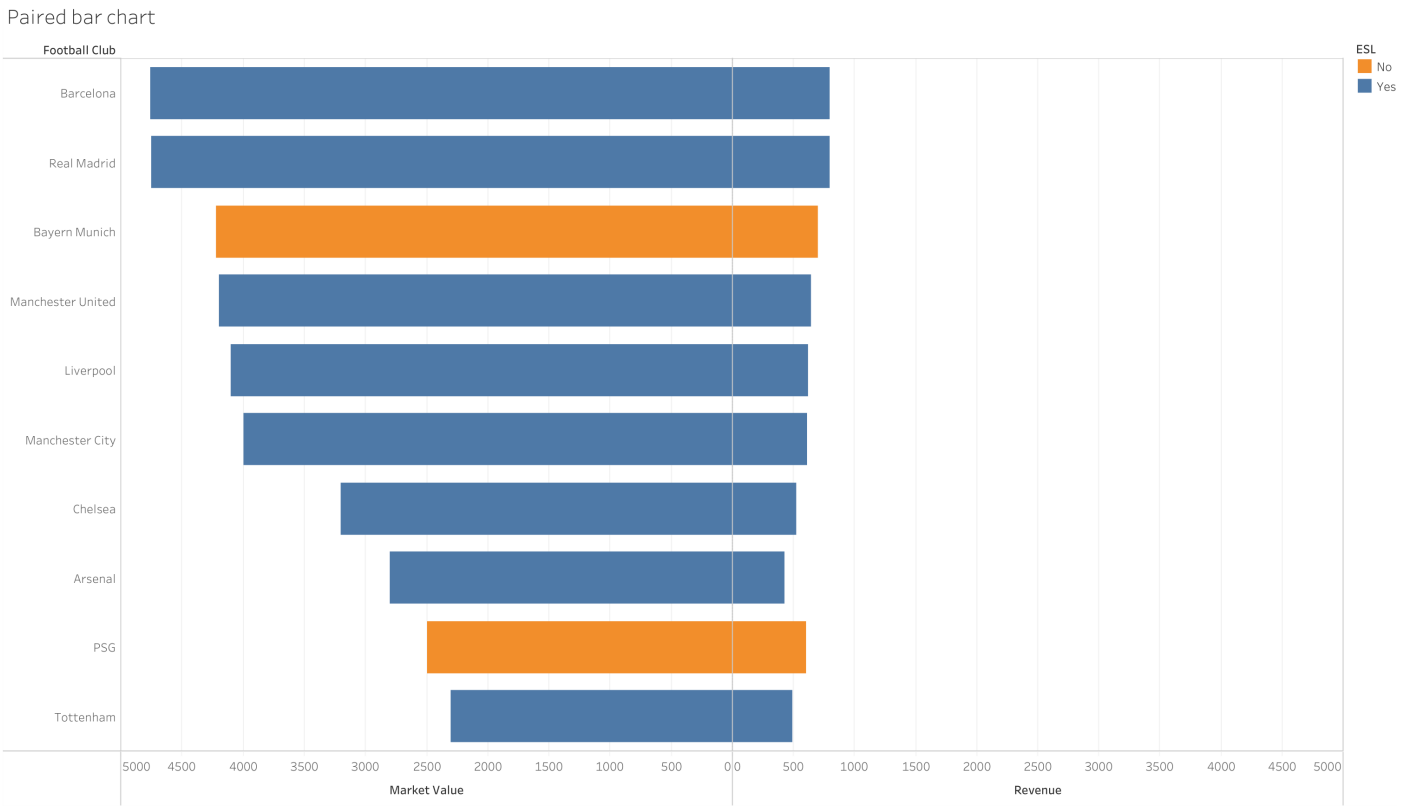
Stacked bar chart



Design schema

Schema	Details
Columns:	SUM(CURRENT_VALUE), SUM(REVENUE)
Rows:	CLUB_NAME
Graph type:	Bar with an inverted axis
Color:	SUPER_LEAGUE
Size:	Default
Label:	Default

Sketch of the resulting graph

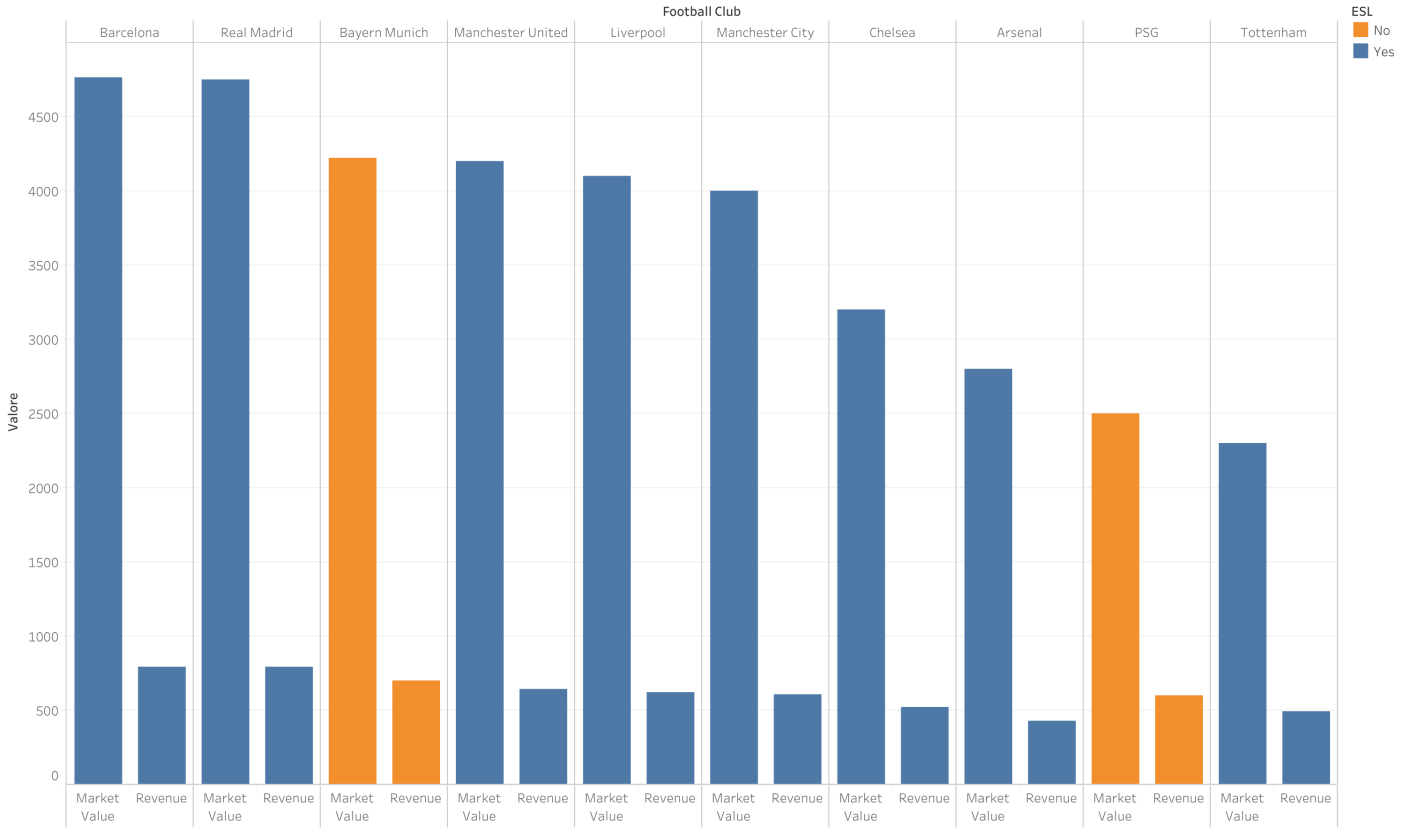


Design schema

Schema	Details
Columns:	CLUB_NAME, Names
Rows:	Measures
Graph type:	Bar
Color:	SUPER_LEAGUE
Size:	Default
Label:	Default

Sketch of the resulting graph

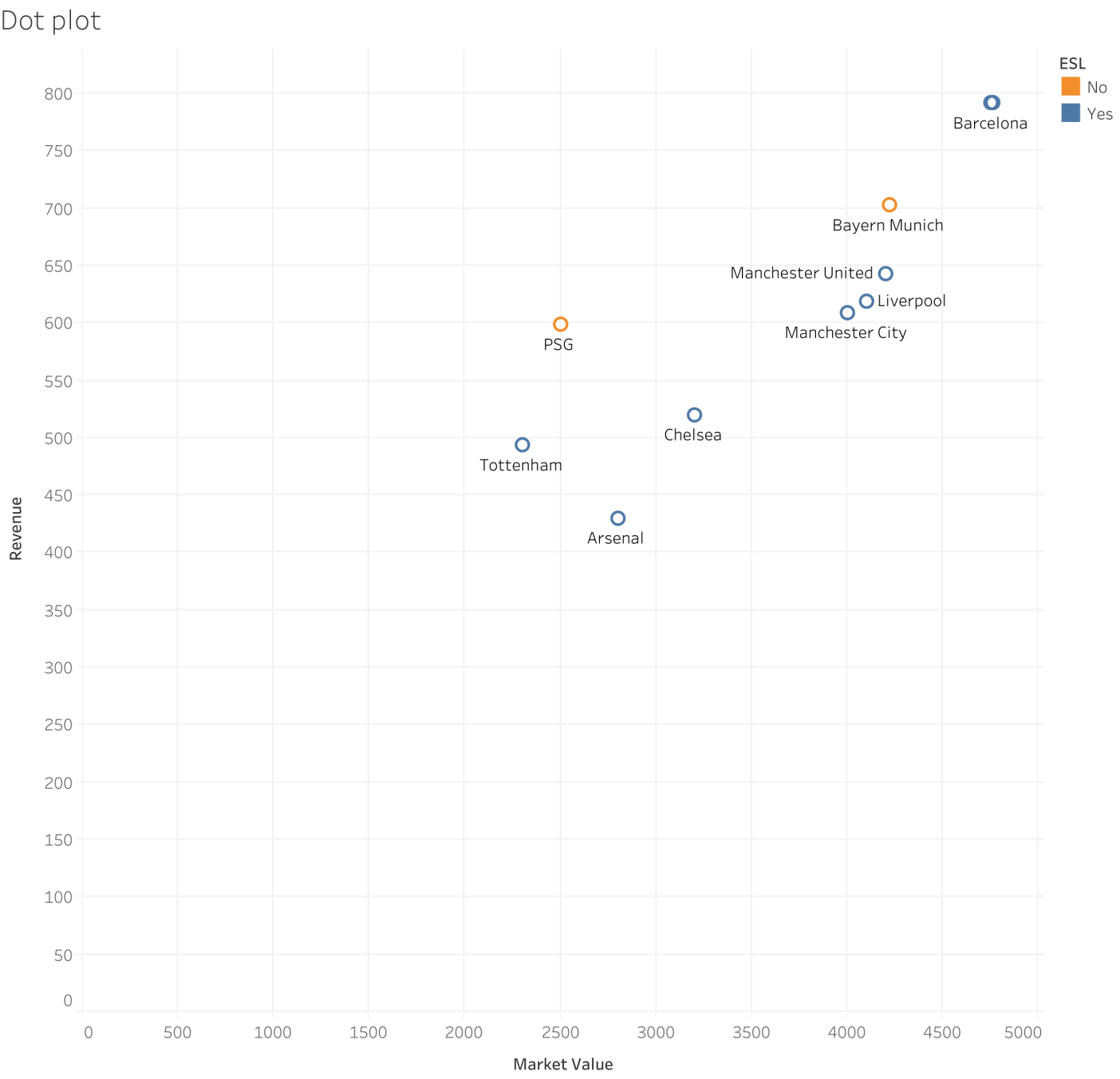
Vertical bar chart



Design schema

Schema	Details
Columns:	SUM(CURRENT_VALUE)
Rows:	SUM(REVENUE)
Graph type:	Shape
Color:	SUPER_LEAGUE
Size:	Default
Label:	CLUB_NAME

Sketch of the resulting graph



Theory

Which one of the following answers is a direct consequence of Steven's law?

- *It is important to avoid comparisons between areas*
- The length of non-aligned objects is harder to compare
- There is no common magnitude assessment for the curvature
- Ordinal measure should be mapped to increasing saturation and intensity
- For every single attribute no more than four distinct levels are discernible

Data management and visualization

Iniziato	mercoledì, 15 settembre 2021, 16:06
Stato	Completato
Terminato	mercoledì, 15 settembre 2021, 16:06
Tempo impiegato	11 secondi
Valutazione	**0,00** su un massimo di 31,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 1,00

An arbiter node of a MongoDB replica set:

- (a) does not hold data and can participate in elections
- (b) holds data and cannot participate in elections
- (c) holds data and can participate in elections
- (d) none of the other answers are correct

Risposta errata.

La risposta corretta è: does not hold data and can participate in elections

Domanda 2

Risposta non data

Punteggio max.: 1,00

The find() operator in MongoDB:

- (a) allows to retrieve specific fields of interest, returning always all documents from a collection
- (b) none of the other answers are correct
- (c) allows to specify the documents of interest and always returns all their fields from a collection
- (d) allows to specify both the documents of interest and the specific fields to be returned from a collection

Risposta errata.

La risposta corretta è: allows to specify both the documents of interest and the specific fields to be returned from a collection

Domanda 3

Risposta non data

Punteggio max.: 1,50

Considering the CAP theorem and its evolutions, when a distributed system provides Availability and Partition tolerance:

- (a) it is also immediately consistent
- (b) none of the other answers are correct
- (c) it will never be consistent
- (d) it can become eventually consistent

Risposta errata.

La risposta corretta è: it can become eventually consistent

Domanda 4

Risposta non data

Punteggio max.: 1,00

In a list of email addresses, you find a phone number. In the context of data quality, this is an issue of...

- (a) Precision
- (b) Understandability
- (c) Accuracy
- (d) Completeness
- (e) Credibility

Risposta errata.

La risposta corretta è: Accuracy

Domanda 5

Risposta non data

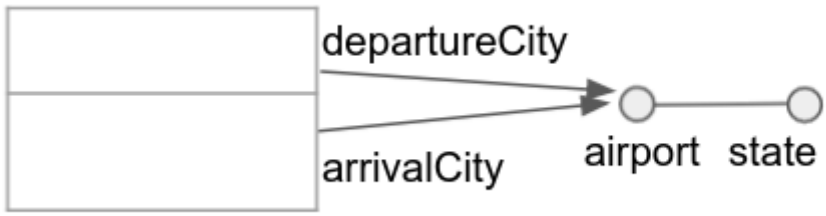
Punteggio max.: 0,50

Data analysts of an international flight operator are interested in analyzing some metrics for different flights.
A flight is characterized by the departure and destination airport.
The data warehouse must be designed to efficiently analyze the average **number of passengers**, the **average duration**, and the **average revenue** for **each flight,** according to the following dimensions.

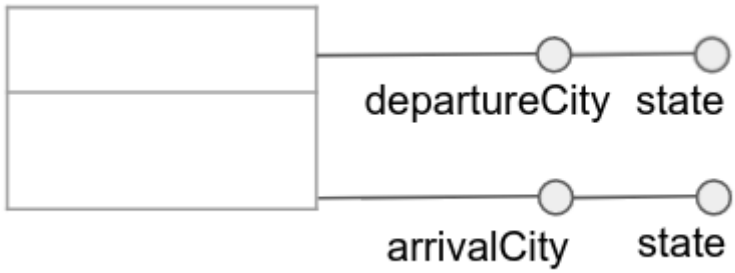
- A flight is characterized by the departure and destination **airport**. For an airport (e.g. Torino Caselle), the city and the state are also stored.
- For each flight, the system stores
  - the **airline** operating the flight (e.g. Delta airlines)
  - the **model** of the airplane
- Each airline offers three additional **services**: “OnBoard Wi-Fi”, “Entertainment” and “Meals&Beverages”. The system stores which services are available for each airline.
- For the passengers, the system records their **age group** (<18, 18-30, 31-60, >60 years old), their **gender**, and their **membership**. Specifically, the membership is “None” if the passenger is not registered to the airline fidelity program, “Basic” if the passenger is registered to the “Basic” fidelity program, and “Premium” otherwise.

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

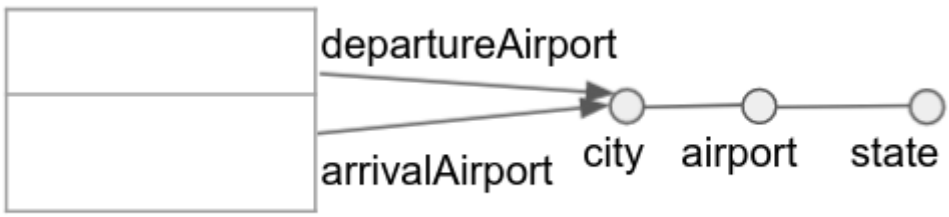
(a)



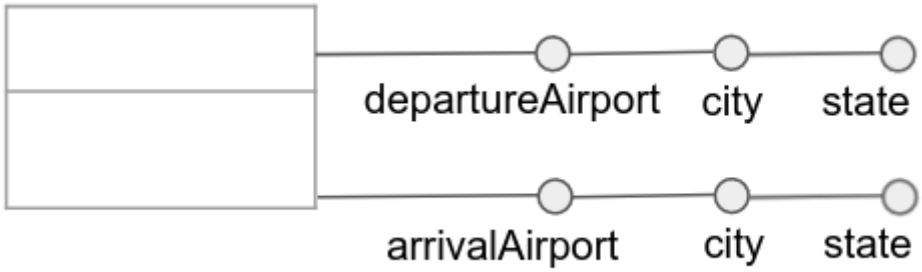
(b)



(c)

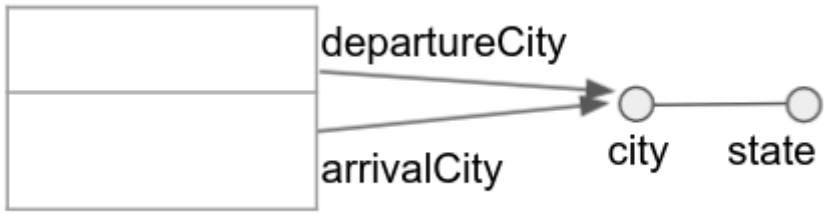


(d)

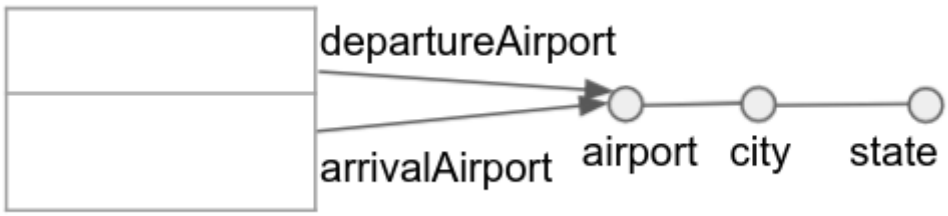


(e)

(f)

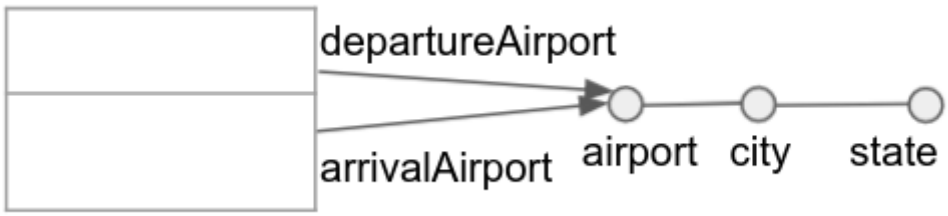


(g)



Risposta errata.

La risposta corretta è:



Domanda 6

Risposta non data

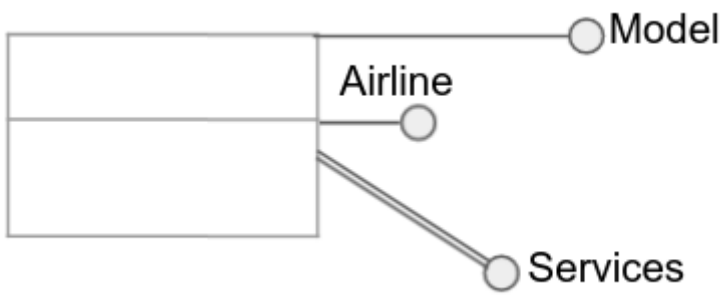
Punteggio max.: 0,50

Data analysts of an international flight operator are interested in analyzing some metrics for different flights.
A flight is characterized by the departure and destination airport.
The data warehouse must be designed to efficiently analyze the average **number of passengers**, the **average duration**, and the **average revenue** for **each flight,** according to the following dimensions.

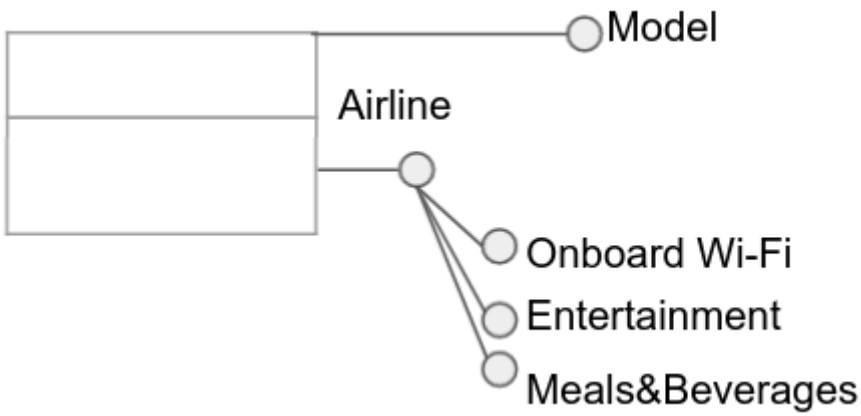
- A flight is characterized by the departure and destination **airport**. For an airport (e.g. Torino Caselle), the city and the state are also stored.
- For each flight, the system stores
  - the **airline** operating the flight (e.g. Delta airlines)
  - the **model** of the airplane
- Each airline offers three additional **services**: “OnBoard Wi-Fi”, “Entertainment” and “Meals&Beverages”. The system stores which services are available for each airline.
- For the passengers, the system records their **age group** (<18, 18-30, 31-60, >60 years old), their **gender**, and their **membership**. Specifically, the membership is “None” if the passenger is not registered to the airline fidelity program, “Basic” if the passenger is registered to the “Basic” fidelity program, and “Premium” otherwise.

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

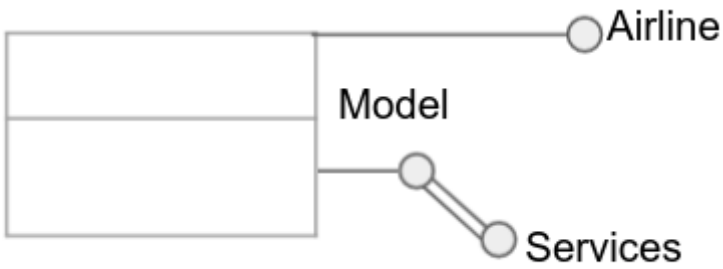
(a)



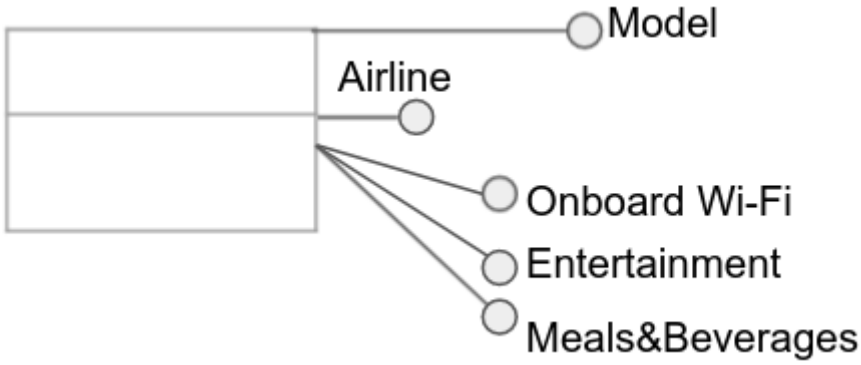
(b)



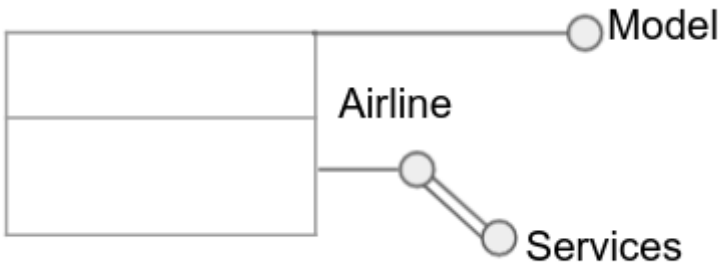
(c)



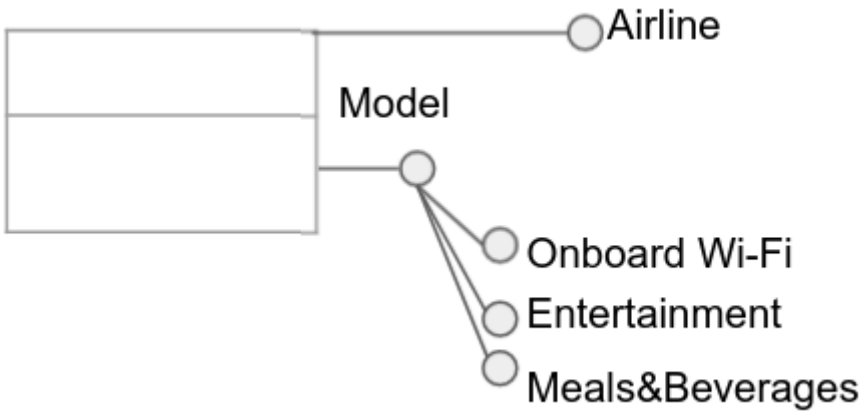
(d)



(e)

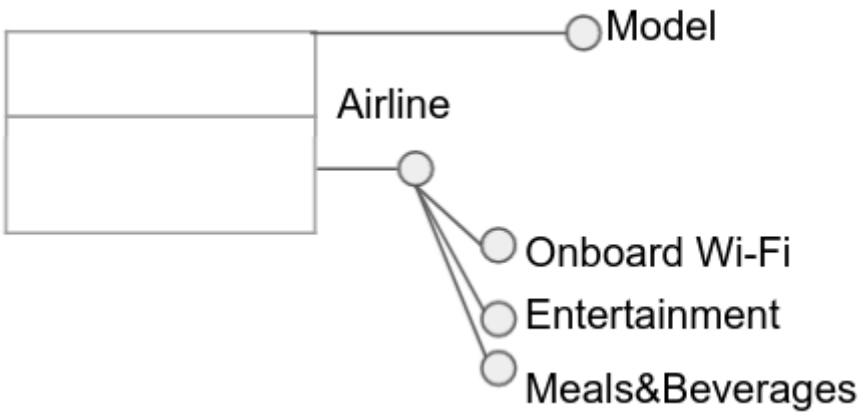


(f)



Risposta errata.

La risposta corretta è:



Domanda 7

Risposta non data

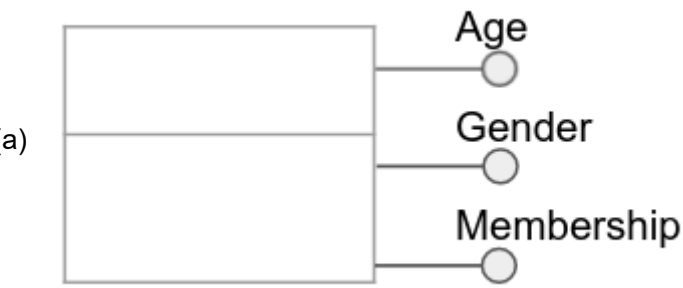
Punteggio max.: 0,50

Data analysts of an international flight operator are interested in analyzing some metrics for different flights.
A flight is characterized by the departure and destination airport.
The data warehouse must be designed to efficiently analyze the average **number of passengers**, the **average duration**, and the **average revenue** for **each flight,** according to the following dimensions.

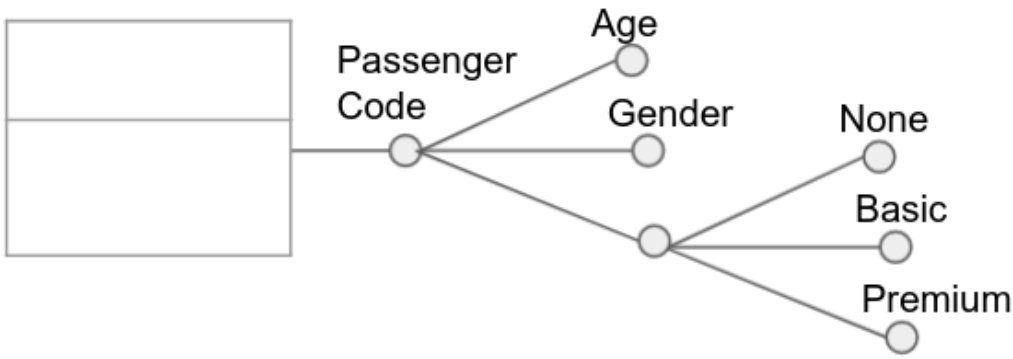
- A flight is characterized by the departure and destination **airport**. For an airport (e.g. Torino Caselle), the city and the state are also stored.
- For each flight, the system stores
  - the **airline** operating the flight (e.g. Delta airlines)
  - the **model** of the airplane
- Each airline offers three additional **services**: “OnBoard Wi-Fi”, “Entertainment” and “Meals&Beverages”. The system stores which services are available for each airline.
- For the passengers, the system records their **age group** (<18, 18-30, 31-60, >60 years old), their **gender**, and their **membership**. Specifically, the membership is “None” if the passenger is not registered to the airline fidelity program, “Basic” if the passenger is registered to the “Basic” fidelity program, and “Premium” otherwise.

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

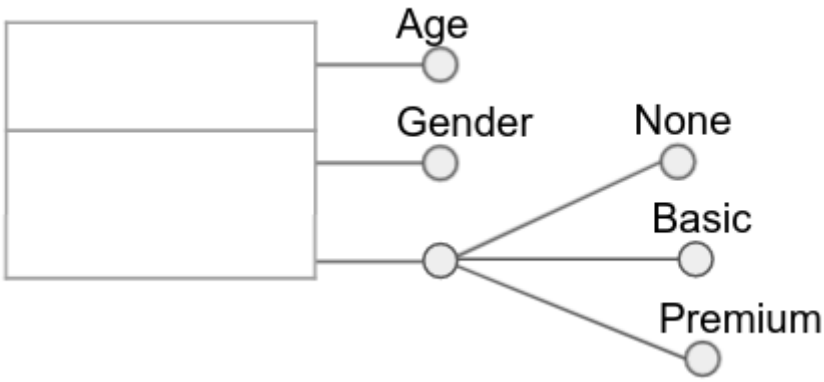
(a)



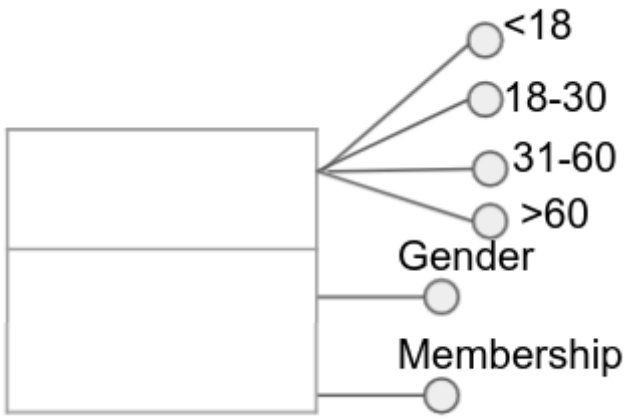
(b)



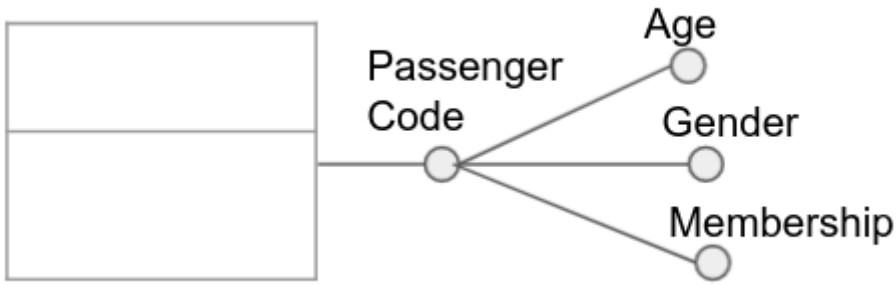
(c)



(d)

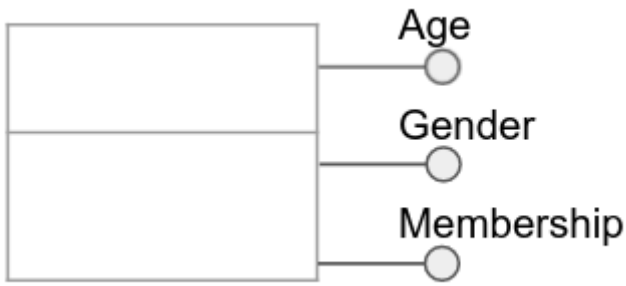


(e)



Risposta errata.

La risposta corretta è:



Domanda 8

Risposta non data

Punteggio max.: 1,50

Data analysts of an international flight operator are interested in analyzing some metrics for different flights.
A flight is characterized by the departure and destination airport.
The data warehouse must be designed to efficiently analyze the average **number of passengers**, the **average duration**, and the **average revenue** for **each flight,** according to the following dimensions.

- A flight is characterized by the departure and destination **airport**. For an airport (e.g. Torino Caselle), the city and the state are also stored.
- For each flight, the system stores
    - the **airline** operating the flight (e.g. Delta airlines)
    - the **model** of the airplane
- Each airline offers three additional **services**: “OnBoard Wi-Fi”, “Entertainment” and “Meals&Beverages”. The system stores which services are available for each airline.
- For the passengers, the system records their **age group** (<18, 18-30, 31-60, >60 years old), their **gender**, and their **membership**. Specifically, the membership is “None” if the passenger is not registered to the airline fidelity program, “Basic” if the passenger is registered to the “Basic” fidelity program, and “Premium” otherwise.

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) Average duration per flight (minutes)
- [ ] (b) Average number of passengers per flight (count)
- [ ] (c) Total number of departure airport per flight (count)
- [ ] (d) Total number of flights (count)
- [ ] (e) Total number of destination airport per flight (count)
- [ ] (f) Number of services (count)
- [ ] (g) Total number of airplane models (count)
- [ ] (h) Total duration of the flights (minutes)
- [ ] (i) Average revenue per flight (euros)
- [ ] (j) Total number of passengers of the flights (count)
- [ ] (k) Total number of airlines (count)
- [ ] (l) Average number of flights (count)
- [ ] (m) Total revenue of the flights (euros)

Risposta errata.

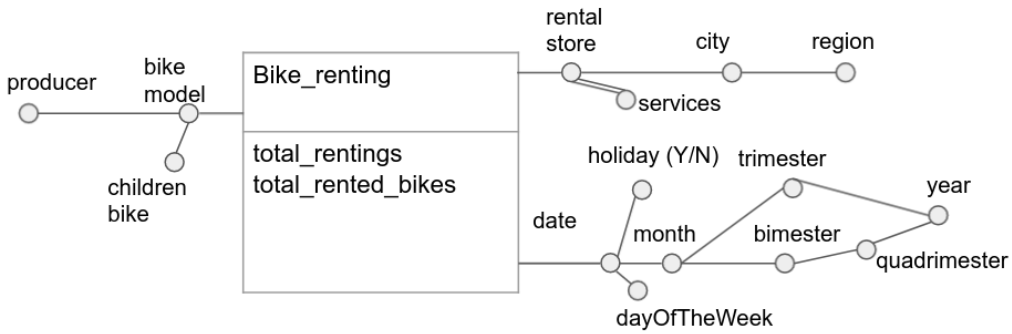
La risposta corretta è: Total number of flights (count), Total number of passengers of the flights (count), Total duration of the flights (minutes), Total revenue of the flights (euros)

Domanda 9

Risposta non data

Punteggio max.: 1,50

Given the following conceptual schema:



- The system stores the bike model and its producer. A bike model can be a model for children or not (field “children bike”)

The value of this field can be “1” if the bike is for children and “0” if not.

- A rental store is identified by a unique name. Stores are analyzed according to their city and region.
- Each rental store may offer many additional articles. Examples of additional services are “bar”, “restaurant”, “wi-fi”, “bike insurance”.
- The system records the renting with their date, the day of the week and if the day was an holiday or not. It also records the month, year, bimester, trimester and quadrimester of the rentings.

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.

VideoGame(**CodV**, VideoGameName, forChildren, Genre, Company)
Store(**CodS,** Store, City, Country, Collectable, Toys, Mangas, Accessories)
Time(**CodT**, date, dayOfTheWeek, holiday, month, 2M, 3M, 6M, year)
Fact(**CodV**, **CodS**, **CodT**, total_revenues, total_sold_videogames)

Domanda 10

Risposta non data

Punteggio max.: 4,00

```
VideoGame(CodV, VideoGameName, forChildren, Genre, Company)
Store(CodS, Store, City, Province, Country)
Time(CodT, date, dayOfTheWeek, holiday, month, bimester,
trimester, semester, year)
Fact(CodV, CodS, CodT, total_revenues, total_sold_videogames)
```

- A video game has a specific name, a specific genre, and it is distributed by a video game company.
- A videogame can be appropriate for children or not.
    - The value of this field can be “0” for not appropriate and “1” for appropriate.
- A store is identified by a unique name. Stores are analyzed according to their city and country.
- The system records the sales with their date, the day of the week and if the day was an holiday or not. It also records the month, year, bimester, trimester and semester of the sales.

Separately for each **videogame** and store **city**, compute the following metrics:

A. the percentage of copies of the videogame sold with respect to the total copies sold in the store province
B. assign a rank to each videogame separately for video game company and city, based on its sales (rank 1st the video game with the highest number of sold copies for each city)

```
SELECT VideoGameName, S.city, (S.province), (S.company)
    100*SUM(total_sold_videogames)/SUM(SUM(total_sold_videogames))
                  OVER (PARTITION BY S.province, V.CodV) as A,
      RANK() OVER (PARTITION BY S.city, V.company
                    ORDER BY SUM(#total_sold_videogames) DESC) as B
FROM VideoGame V, Fact F, Store S
WHERE V.CodV=F.CodV and S.StoreS=F.CodV
GROUP BY V.CodV, S.city, S.province, V.Company, VideoGameName
```

Note: Reading VideoGame is not strictly required (missing the Video Game name)

Domanda 11

Risposta non data

Punteggio max.: 4,00

```
VideoGame(CodV, VideoGameName, forChildren, Genre, Company)
Store(CodS, Store, City, Province, Country)
Time(CodT, date, dayOfTheWeek, holiday, month, bimester,
trimester, semester, year)
Fact(CodV, CodS, CodT, total_revenues, total_sold_videogames)
```

- A video game has a specific name, a specific genre, and it is distributed by a video game company.
- A videogame can be appropriate for children or not.
    - The value of this field can be “0” for not appropriate and “1” for appropriate.
- A store is identified by a unique name. Stores are analyzed according to their city and country.
- The system records the sales with their date, the day of the week and if the day was an holiday or not. It also records the month, year, bimester, trimester and semester of the sales.

Separately for each store **city** and **bimester**, compute the following metrics, only for the videogames appropriate for children:

A. the cumulative revenues since the beginning of the semester
B. the daily average revenues
C. the percentage of revenues in the bimester with respect to the revenues in the semester, for each city

---

```
SELECT city, bimester, (semester),
    SUM(SUM(total_revenues)) OVER (
                PARTITION BY city, semester
                ORDER BY bimester
                ROWS UNBOUNDED PRECEDING) as A,
      SUM(total_revenues)/COUNT(distinct date) as B,
      100*SUM(total_revenues)/SUM(SUM(total_revenues))
                OVER (PARTITION BY city, semester) as C,
FROM VideoGame V, Fact F, Store S
WHERE T.CodT=F.CodT and S.StoreS=F.CodV and V.CodV=F.CodV and
forChildren=1
GROUP BY city, bimester, semester
```

Domanda 12

Risposta non data

Punteggio max.: 2,00

The following document structure represents online courses.

```
{"_id": ObjectId(“xyz”),
"title": "Python 3.9",
"teacher":{
        "name": "John",
        “surname”: “Doe”,
        "webiste": "https://www.doe.com/",
        “nation”: “USA”
           },
"published": Date("2019-02-13T00:00:00.000Z"),
"category": "Computer Science",
“tags : [“Python”, “Coding”],
"price": 99,
“avg_score”: 4.8,
“number_reviews”: 47,
“enrolled_students”: 1234,
“details: {
            “hour_length”:  12,
            “number_of_lessons: 38,
            “final_test”: false
            }
}
```

Write a MongoDB query to display only the title, the category, and the price of courses containing the tag “Databases”, published in 2019, and whose length is less than 10 hours.

N.B. Use the syntax *new Date (string)* to manage date attributes, e.g., "attribute": new Date("2021-09-01")

```
db.courses.find(
{
            tag: “Databases”,
            published: {
                        $gte: new Date('2019-01-01'),
                        $lt: new Date('2020-01-01')
                        },
            ‘details.hour_length’: {
                        $lt: 10
                        }
},
{“title”:1, "category":1, “price”:1, “_id”:0}
)
```

Domanda 13

Risposta non data

Punteggio max.: 3,00

The following document structure represents online courses.

```
{"_id": ObjectId(“xyz”),
"title": "Python 3.9",
"teacher":{
        "name": "John",
        “surname”: “Doe”,
        "webiste": "https://www.doe.com/",
        “nation”: “USA”
           },
"published": Date("2019-02-13T00:00:00.000Z"),
"category": "Computer Science",
“tags : [“Python”, “Coding”],
"price": 99,
“avg_score”: 4.8,
“number_reviews”: 47,
“enrolled_students”: 1234,
“details: {
            “hour_length”:  12,
            “number_of_lessons: 38,
            “final_test”: false
            }
}
```

Considering only courses in the category Computer Science published in the year 2020, for each tag, select the average price and the maximum number of enrolled students.

N.B. Use the syntax *new Date (string)* to manage date attributes, e.g., "attribute": new Date("2021-09-01")

```
db.courses.aggregate([
{$match: {“published”: {$gte: new Date(‘2020-01-01’), $lt: new Date('2021-01-01')}
 }, “category”: “Computer Science” },
{ $unwind: ‘$tags’},
{$group:
            {
            '_id': ‘$tags’,
            'avg_price':  {'$avg': '$price'}
            'max_students':  {'$max': '$enrolled_students’}
            }
}
])
```

Domanda 14

Risposta non data

Punteggio max.: 4,00

Design a MongoDB database to store reviews of hotels from a website according to the following requirements.

The data to be displayed on the review website for each hotel include the hotel name, the number of stars, and the list of provided services, e.g., free wifi, baby parking, pet allowed, etc.

For each hotel, the venue information and its top 10 reviews must be always shown.

The venue information consists of the address, the city, and the country. Furthermore, the official website address might be included in the venue information.

Each review consists of a timestamp, a score (e.g., 4.5), the nickname of its author, the number of “likes”, and a textual description. Each review is related to one specific hotel.

Given a hotel, the database must be designed to efficiently provide all the data describing the hotel, its top 10 reviews (those having the highest numbers of “likes”), its total number of reviews, and their average score.

Instead, given a review, the database must efficiently provide the hotel name, its number of stars and its city.

Write a sample document for each collection of the database.

Important: besides the sample documents, explicitly indicate the design patterns used.

Hotel

```
{
_id: ObjectId(),
name: <string>,
stars: <number>,
services: [<string>],
venue: {
        address: <url>,
        city: <string>,
        country: <string>,
        website: <url>
},
top_reviews: [
{_id: ObjectId(),
timestamp: <date>,
score: <number>,
nickname: <string>,
likes: <number>,
description: <string>  }
],
tot_reviews: <number>,
avg_score: <number>
}
```

Review

```
{_id: ObjectId(),
timestamp: <date>,
score: <number>,
nickname: <string>,
likes: <number>,
description: <string>,
hotel: {
_id: ObjectId(),
name: <string>,
stars: <number>,
city: <string>
 }
}
```

Patterns used:

Polymorphic pattern to track the venue information in the hotel collection (due to the optional website info).

Subset pattern to track the top 10 reviews for each hotel.

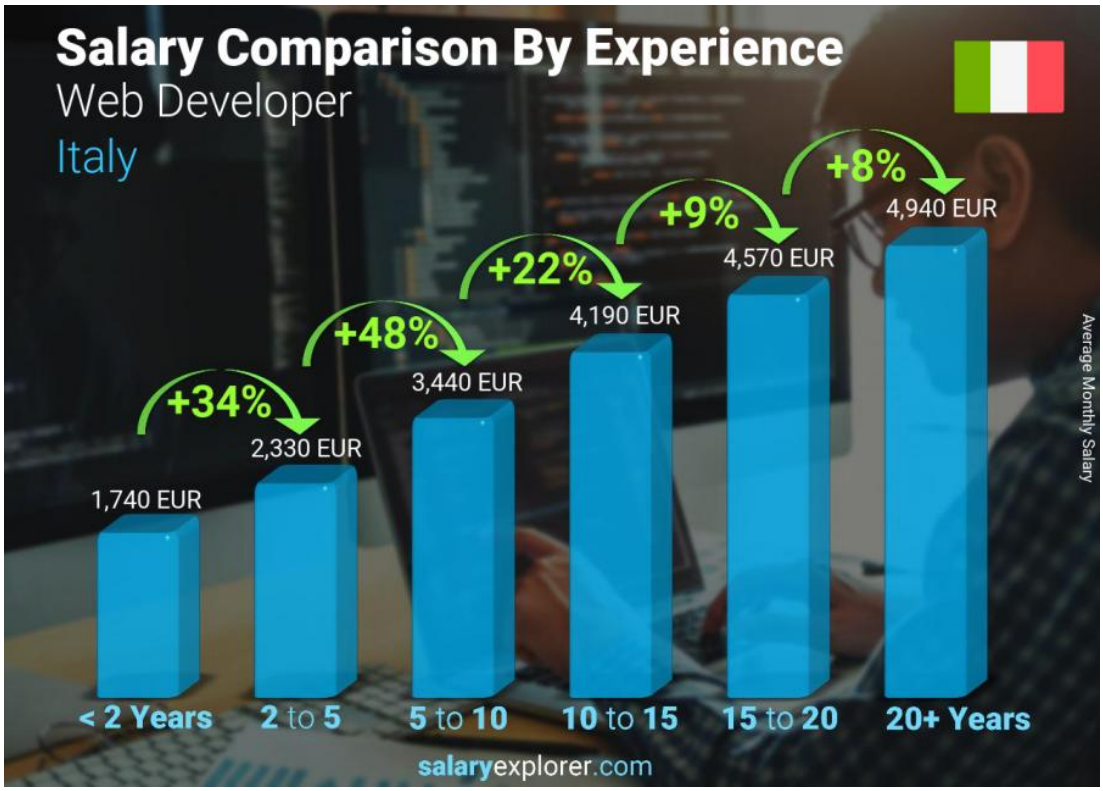
Computed pattern for the average score and total review count of each hotel.

Extended reference for the review collection to show the hotel info.

Domanda 15

Risposta non data

Punteggio max.: 0,25



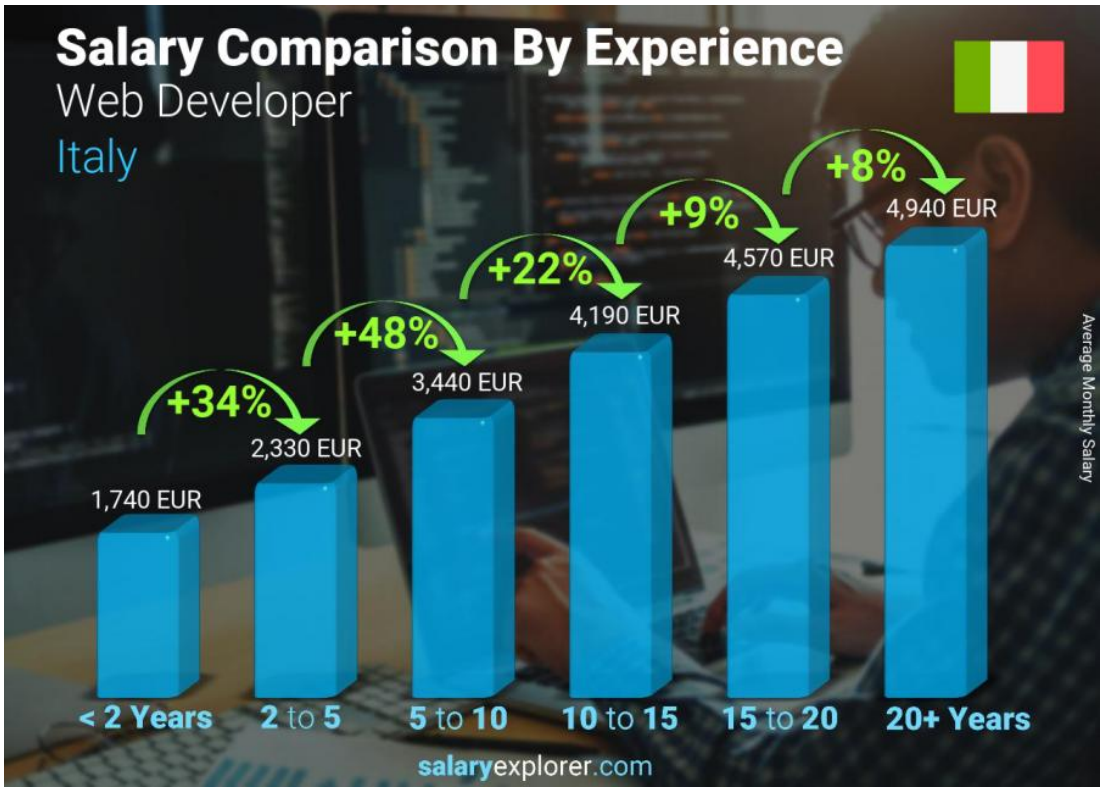
Question

Is there a clearly defined question addressed by the visualization? Write it down.

Domanda 16

Risposta non data

Punteggio max.: 1,25



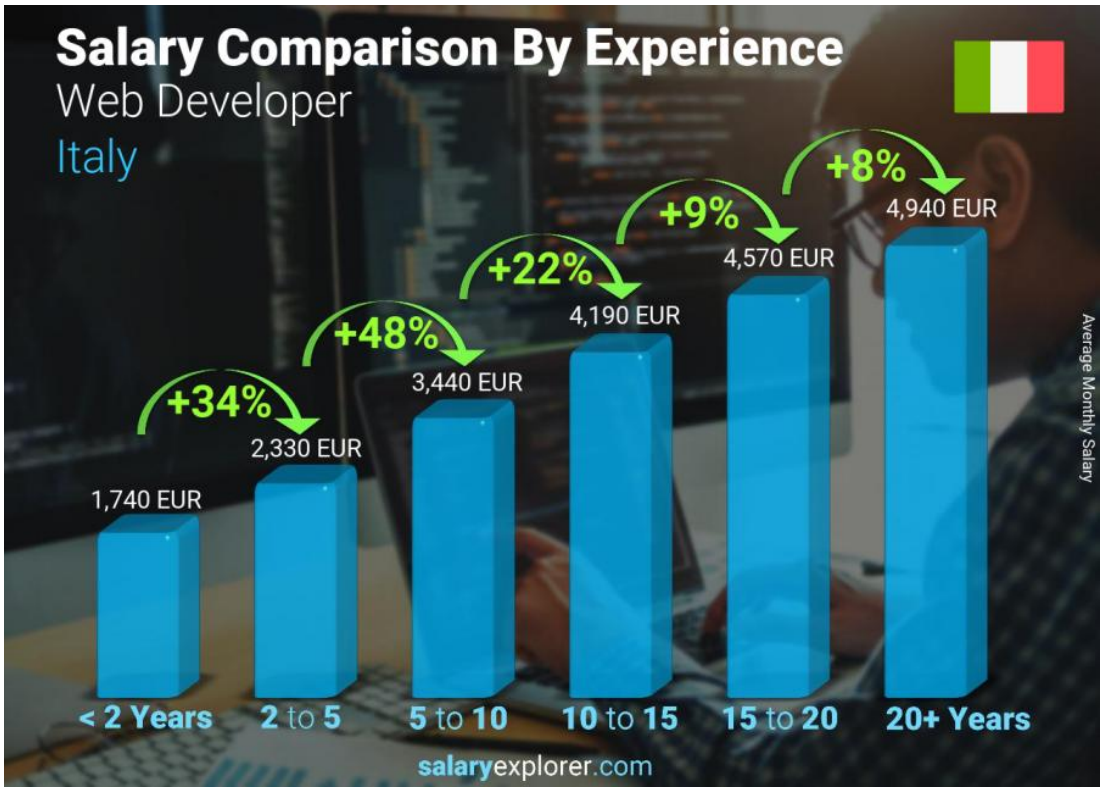
Data

Is the data quality appropriate? Identify the inadequate characteristics and explain.

Domanda 17

Risposta non data

Punteggio max.: 0,75



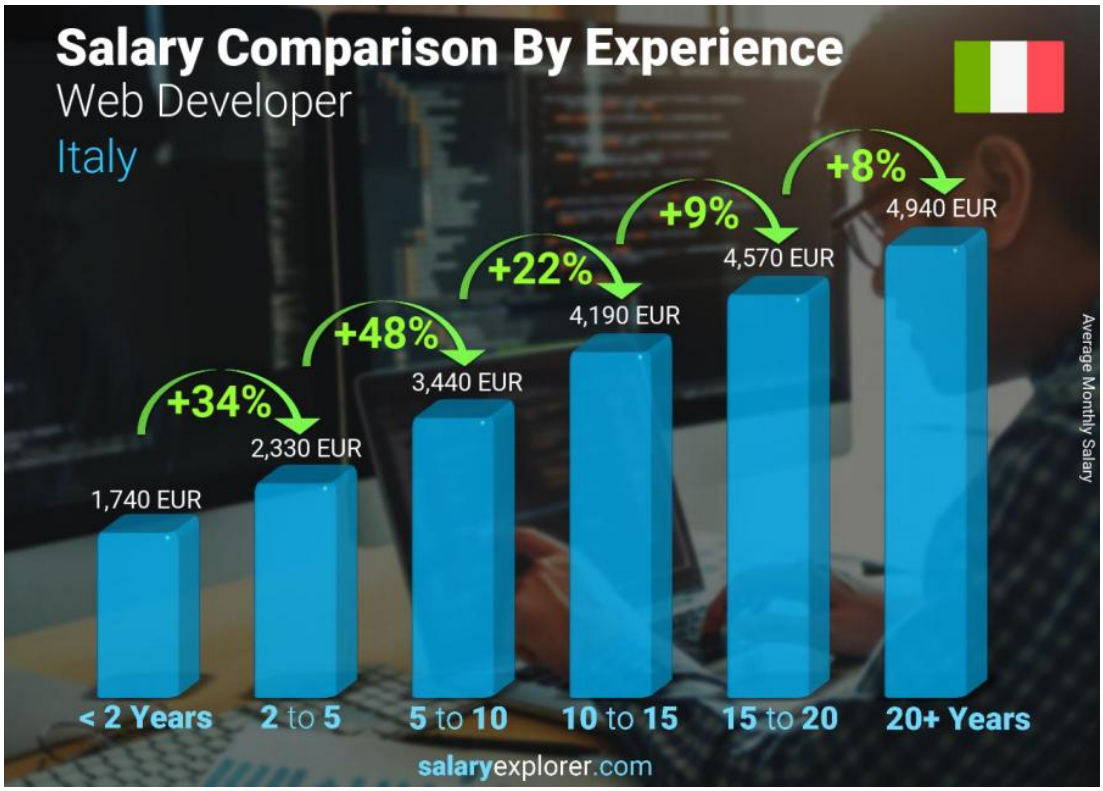
Visual Proportionality

Are the values encoded in a uniformly proportional way?

Domanda 18

Risposta non data

Punteggio max.: 0,75



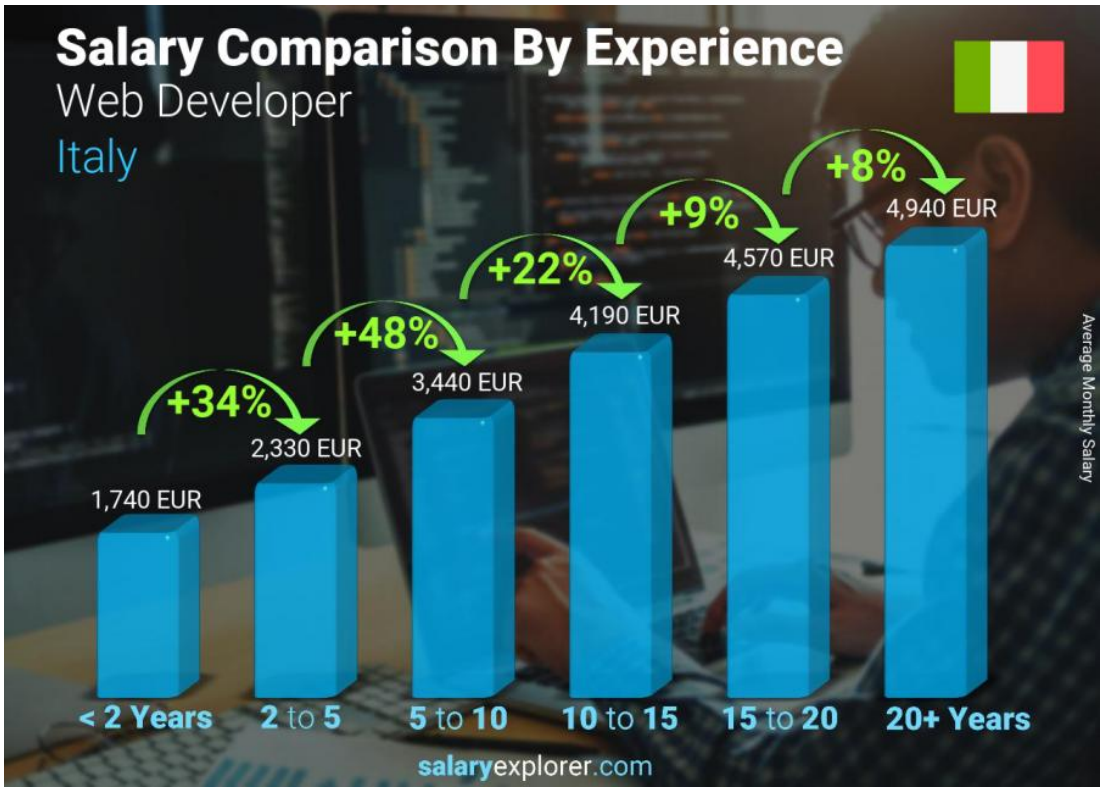
Visual Utility

All the elements in the graph convey useful information?

Domanda 19

Risposta non data

Punteggio max.: 0,50



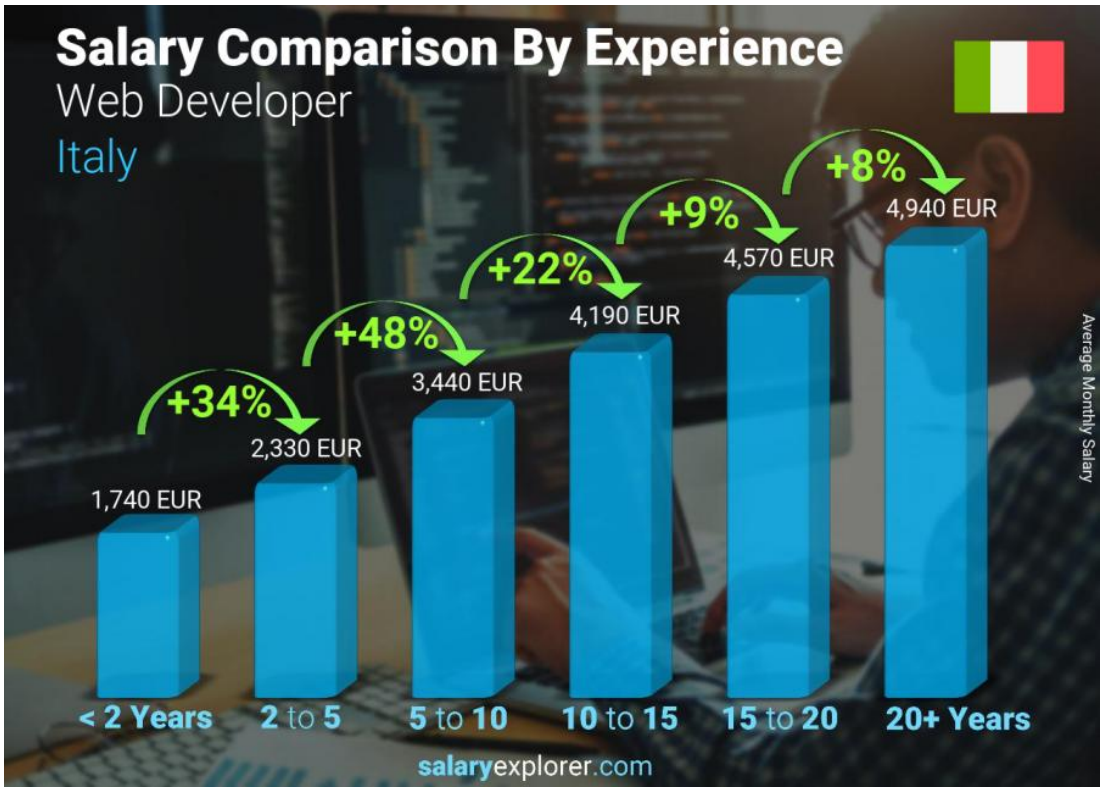
Visual Clarity

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

Domanda 20

Risposta non data

Punteggio max.: 0,25



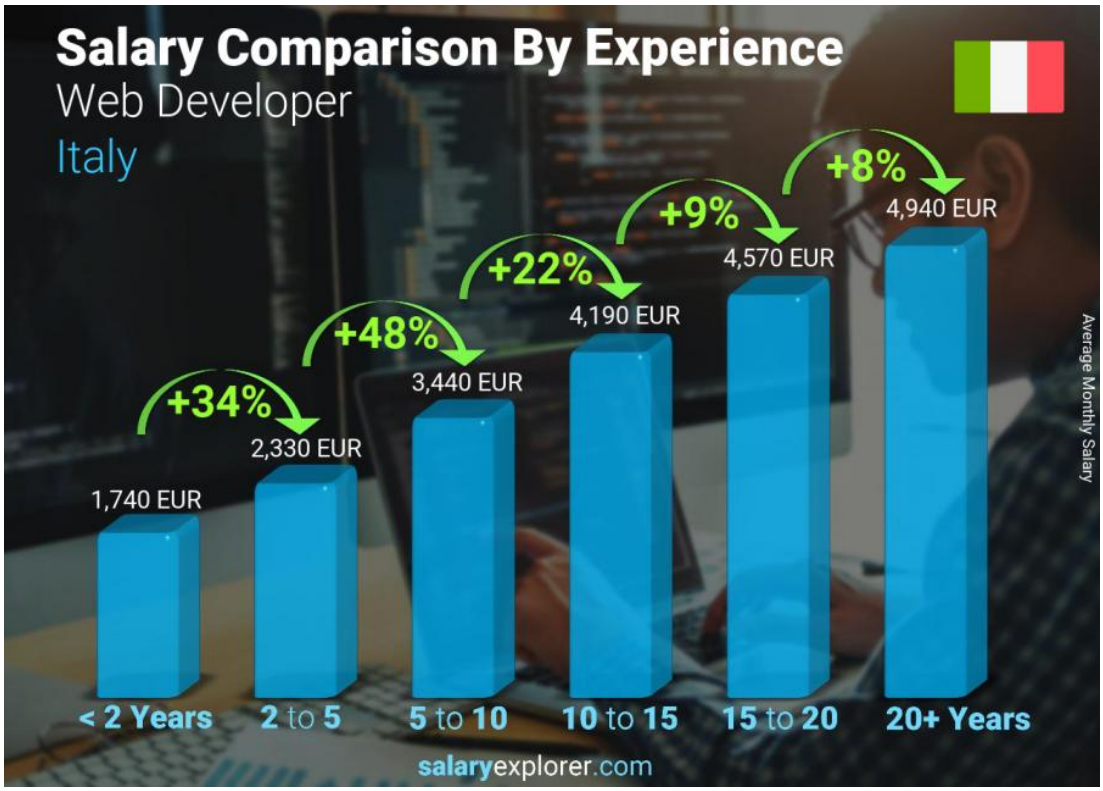
Design data

Design the visualization based on the following data structure (to be completed).

Domanda 21

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 22

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.

DM & Visualization - Exam 2021-09-01 - Solution

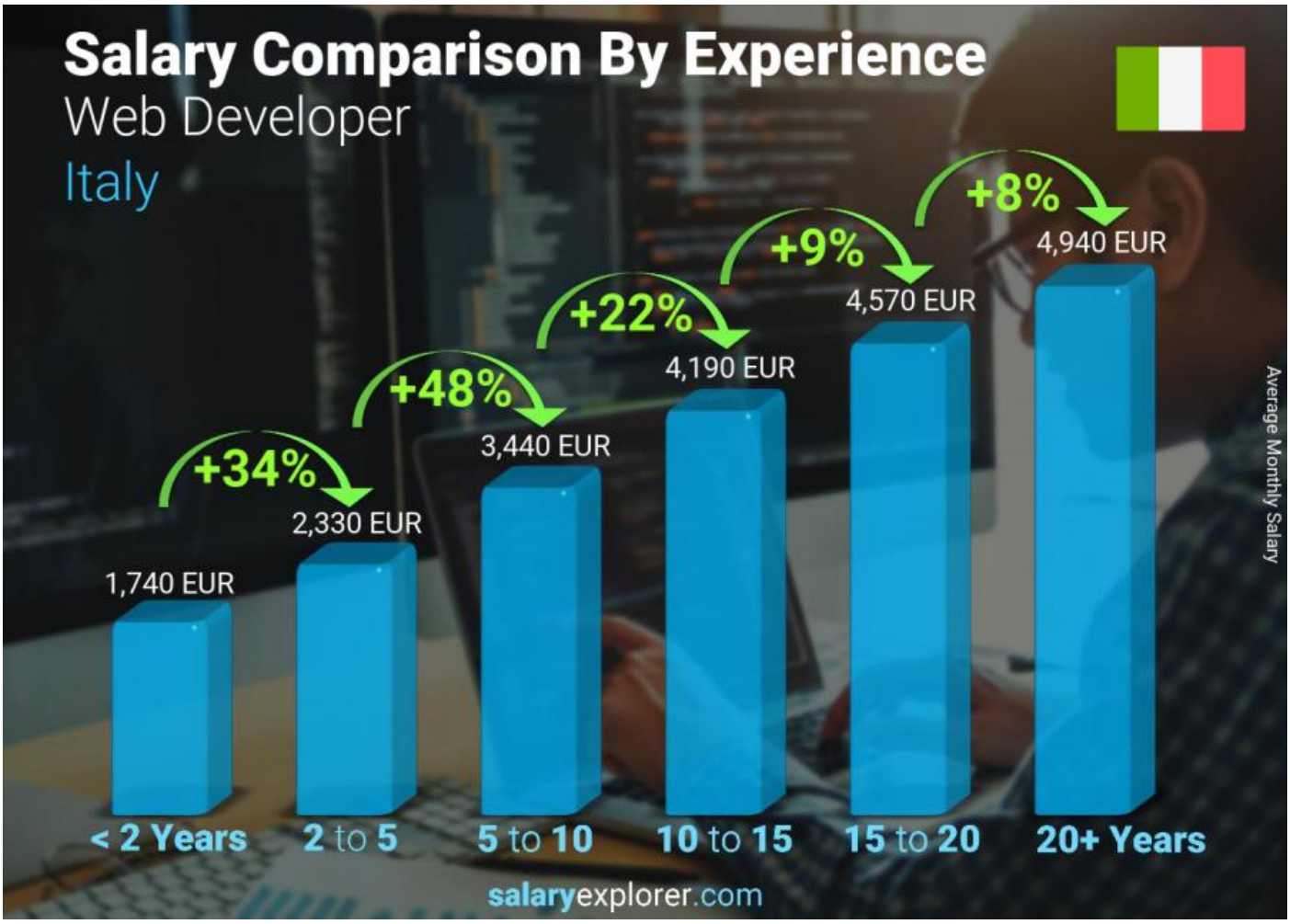


Figure 1: Salary Comparison By Experience

Analysis

Analyze the above graph illustrating a comparison of different salaries by years of experience.

Question: Is there one (or more) question addressed by the visualization?

The question is quite clear: what is the average monthly salary of a web developer in Italy by years of experience?

Data: Is the data quality appropriate?

Accuracy: the values are monetary amounts and they are expressed in Euros.

Completeness: data are complete, as all the ranges of years of experience are available.

Consistency: data are not consistent, as the ranges have different timespans.

Currency: unclear, because no date is available.

Credibility: the source is Salary Explorer, data were probably collected with a survey.

Understandability: the values probably represent the salaries before taxation, but this is not reported.

Precision: precision is appropriate for the task.

Visual Proportionality: Are the values encoded in a uniformly proportional way?

The lengths of the bars are not proportional to the corresponding salaries because the y axis does not start from zero. In any case, the usage of 3D and the translucent effect make the comparison quite difficult.

Visual Utility: All the elements in the graph convey useful information?

Not at all: the background, the translucent effect, the shades, the percentages, the green arrows, the flag of Italy can be removed.

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

The numerical values associated with each bar are clearly reported. It is difficult to evaluate proportions and understand the correct value because of the 3D effect and the missing axis. It is stated that the values represent the average monthly salaries. The meaning of the percentages is clear, even if they are useless. Each salary is associated with a range of years of experience.

Design

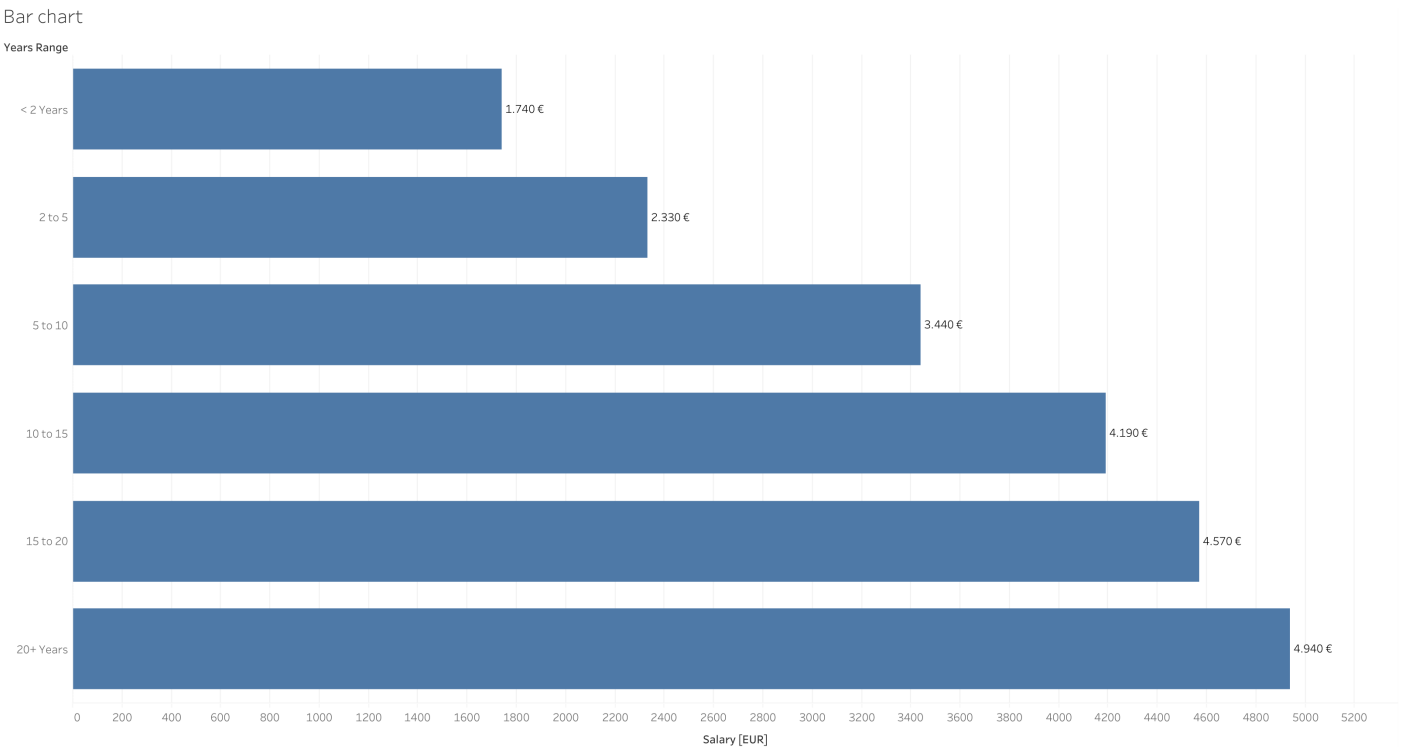
Design the visualization based on the following data structure

Field	Dim./Measure	Description
YEARS_RANGE	Dimension	The years of experience
SALARY	Measure	The average monthly salary

Design schema

Schema	Details
Columns:	SUM(SALARY)
Rows:	YEARS_RANGE
Graph type:	Bar
Color:	Default
Size:	Default
Label:	SUM(SALARY)

Sketch of the resulting graph

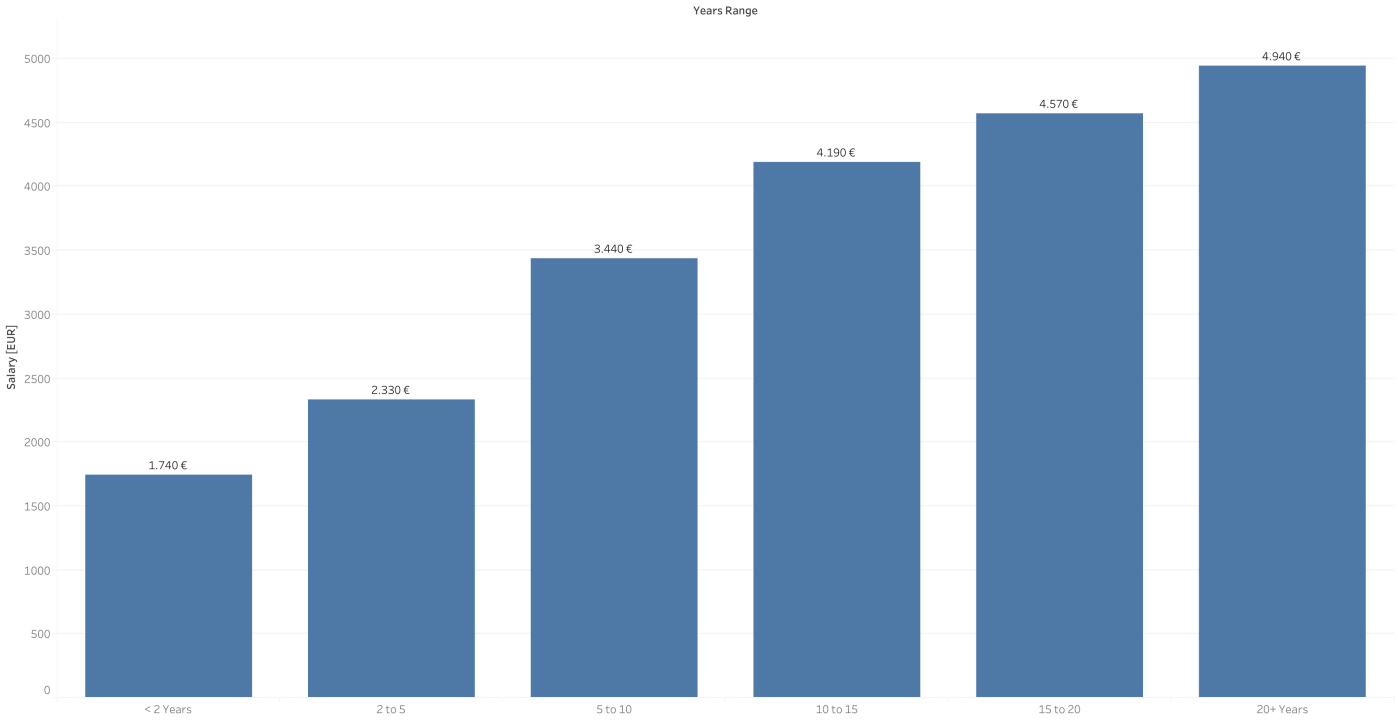


Design schema

Schema	Details
Columns:	YEARS_RANGE
Rows:	SUM(SALARY)
Graph type:	Bar
Color:	Default
Size:	Default
Label:	SUM(SALARY)

Sketch of the resulting graph

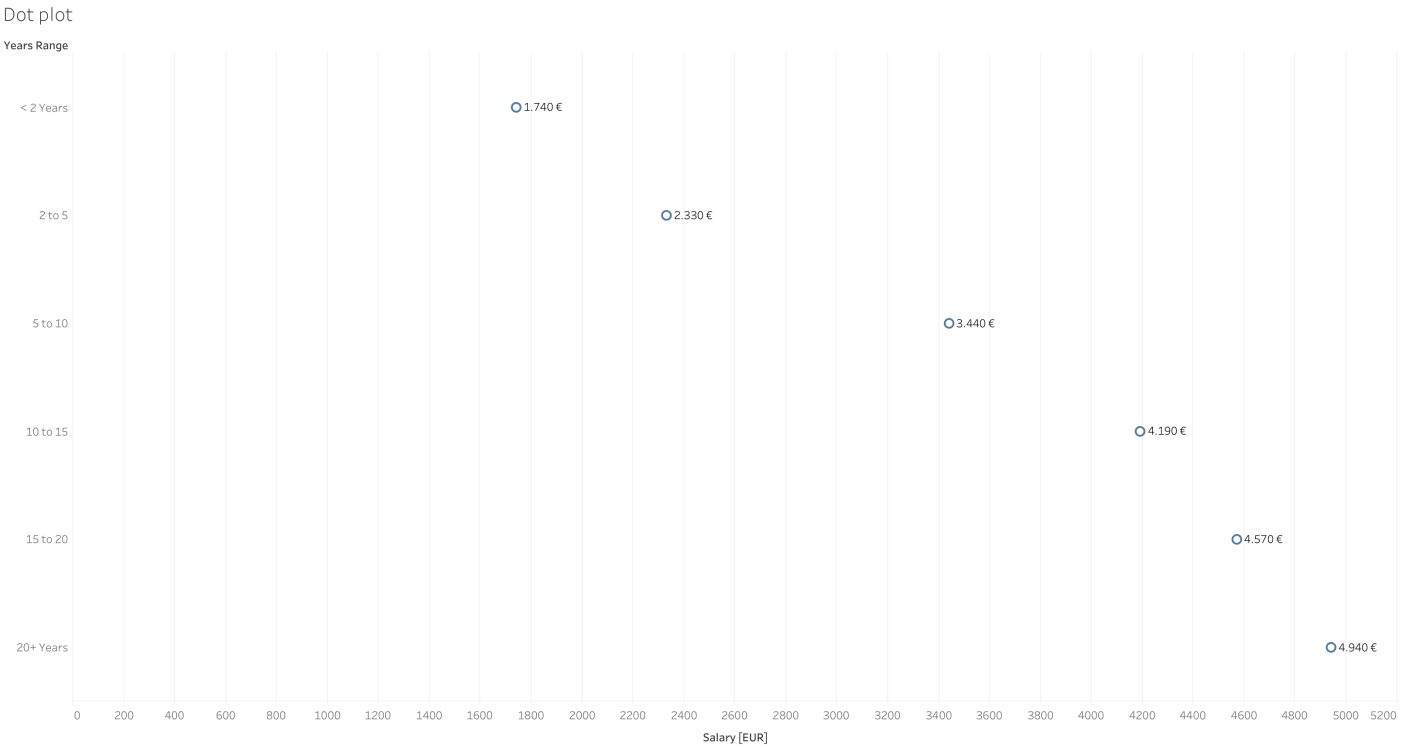
Vertical bar chart



Design schema

Schema	Details
Columns:	SUM(SALARY)
Rows:	YEARS_RANGE
Graph type:	Shape
Color:	Default
Size:	Default
Label:	SUM(SALARY)

Sketch of the resulting graph

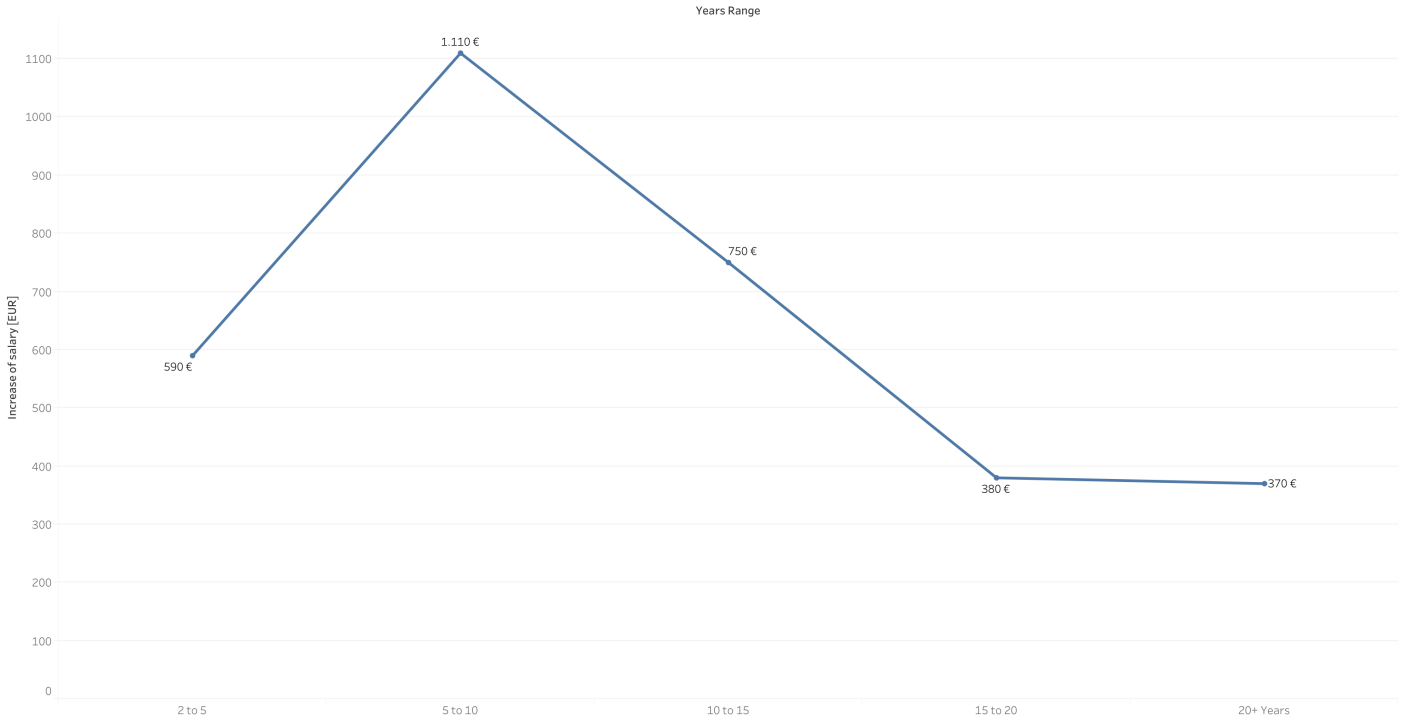


Design schema

Schema	Details
Columns:	YEARS_RANGE
Rows:	ZN(SUM(SALARY)) - LOOKUP(ZN(SUM(SALARY)), -1)
Graph type:	Line
Color:	Default
Size:	Default
Label:	ZN(SUM(SALARY)) - LOOKUP(ZN(SUM(SALARY)), -1)

Sketch of the resulting graph

Line plot



Theory

In a list of email addresses, you find a phone number. In the context of data quality, this is an issue of...

- *Accuracy*
- Completeness
- Credibility
- Understandability
- Precision