

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

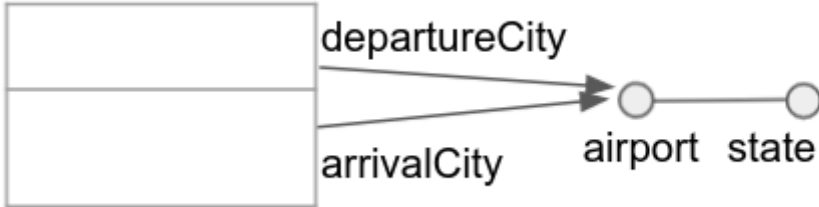
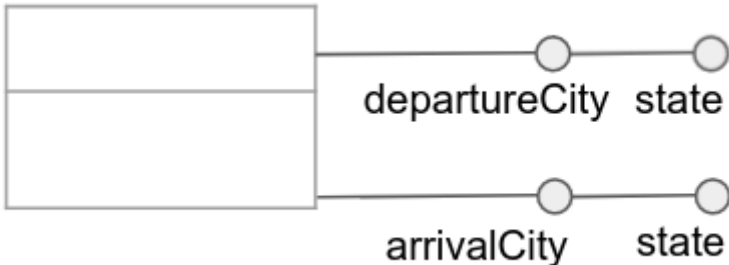

Data analysts of an international flight operator are interested in analyzing some metrics for different flights.

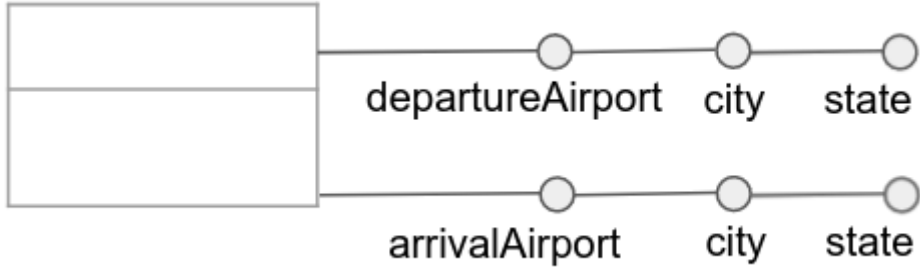
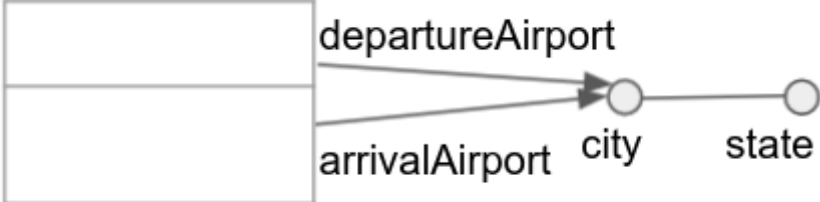
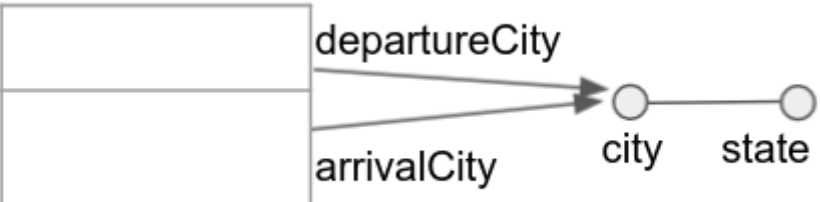

A flight is characterized by the departure and destination airport.

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

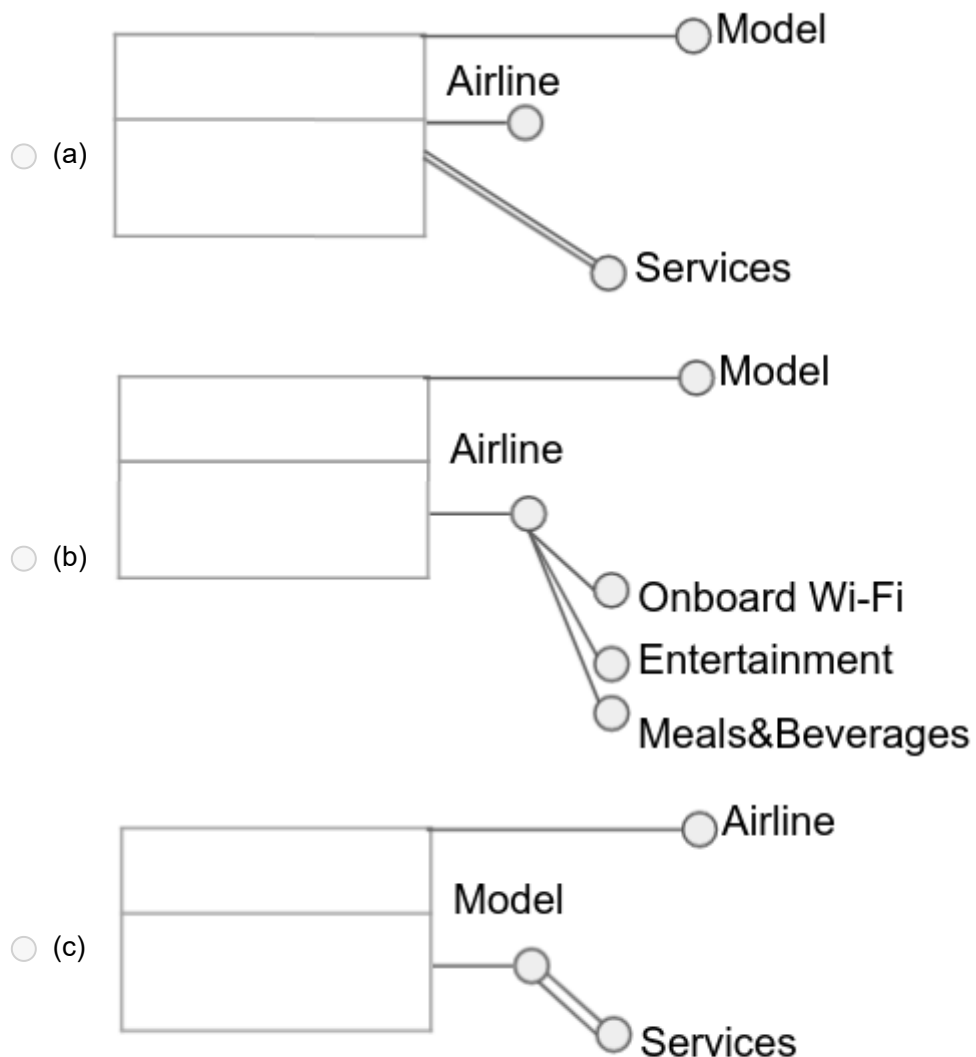
Data analysts of an international flight operator are interested in analyzing some metrics for different flights.

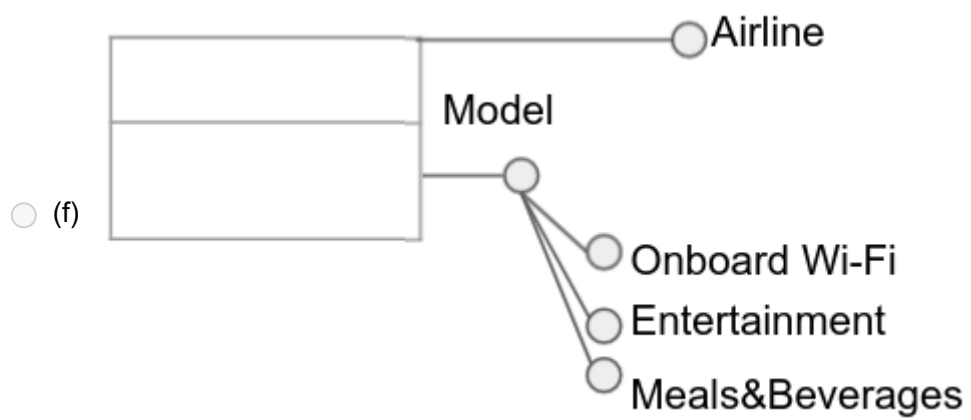
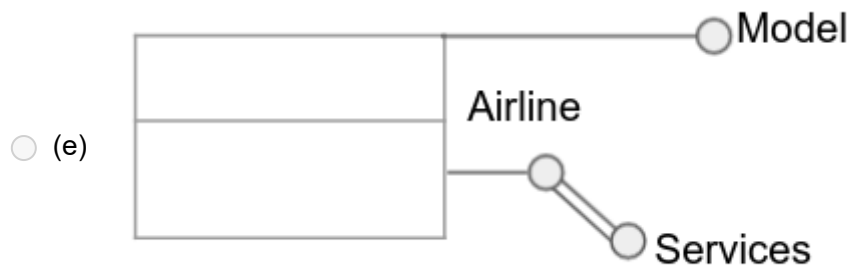
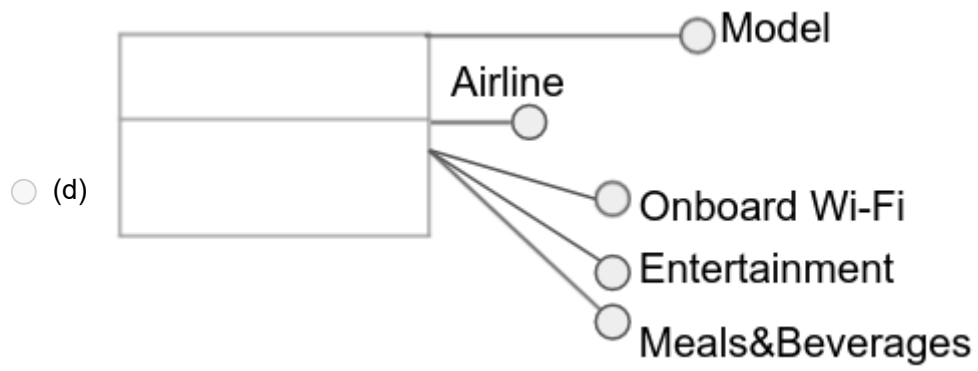
A flight is characterized by the departure and destination airport.

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.

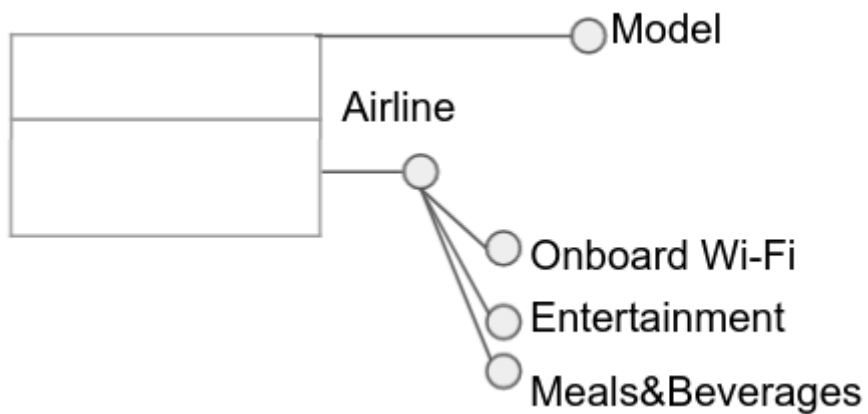
Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).





Risposta errata.

La risposta corretta è:



Domanda 7

Risposta non data

Punteggio max.:

0,50

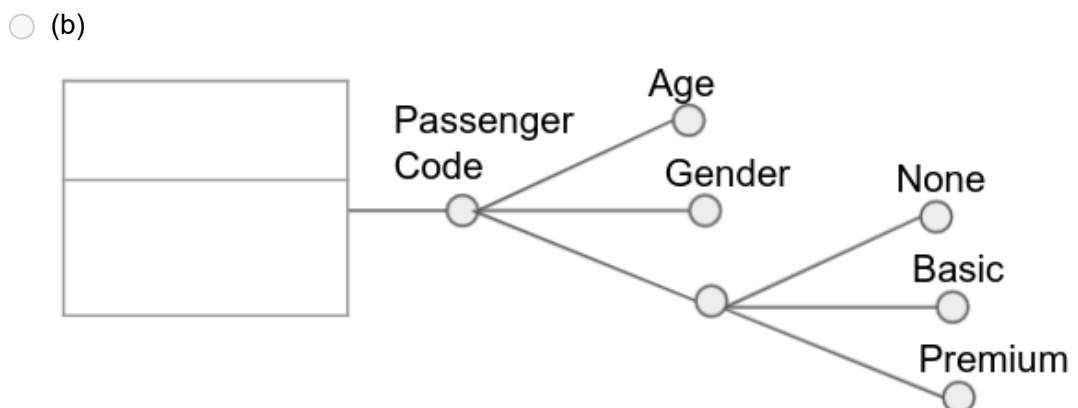
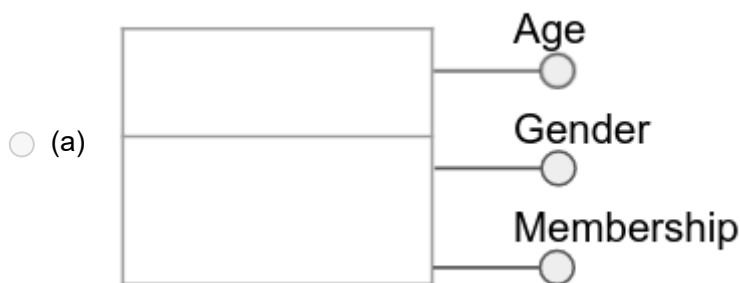
Data analysts of an international flight operator are interested in analyzing some metrics for different flights.

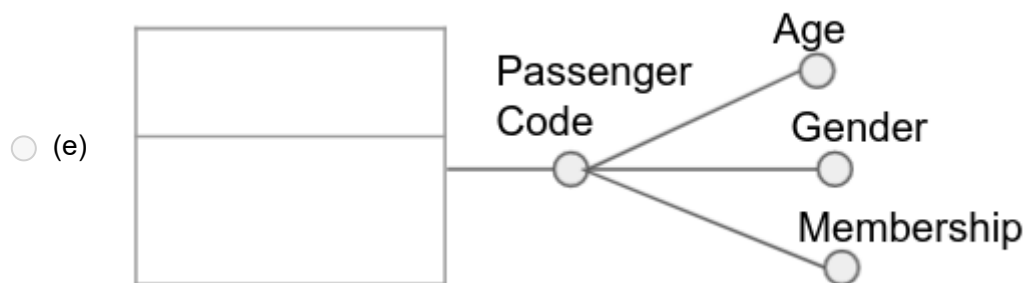
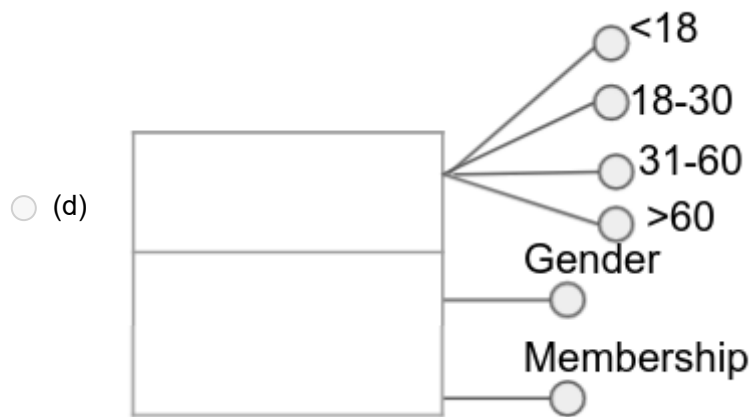
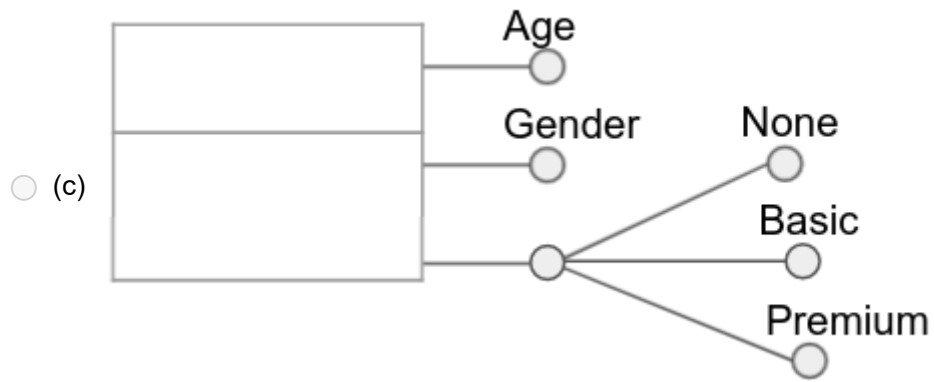
A flight is characterized by the departure and destination airport.

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.

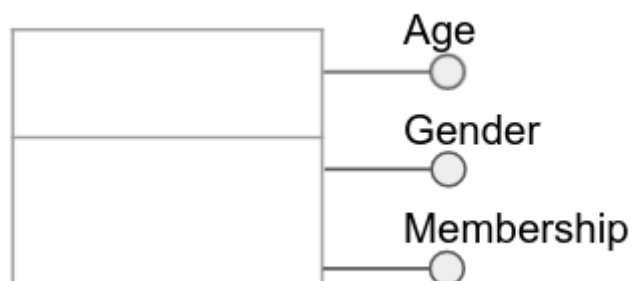
Select, among the following dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).





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)

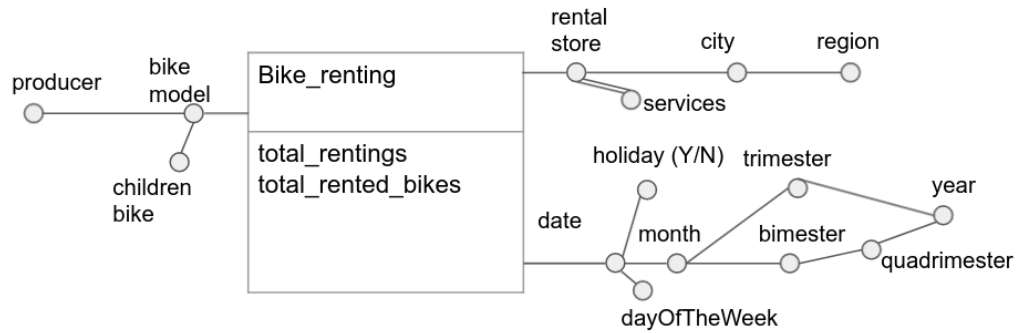
Domanda 9

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.

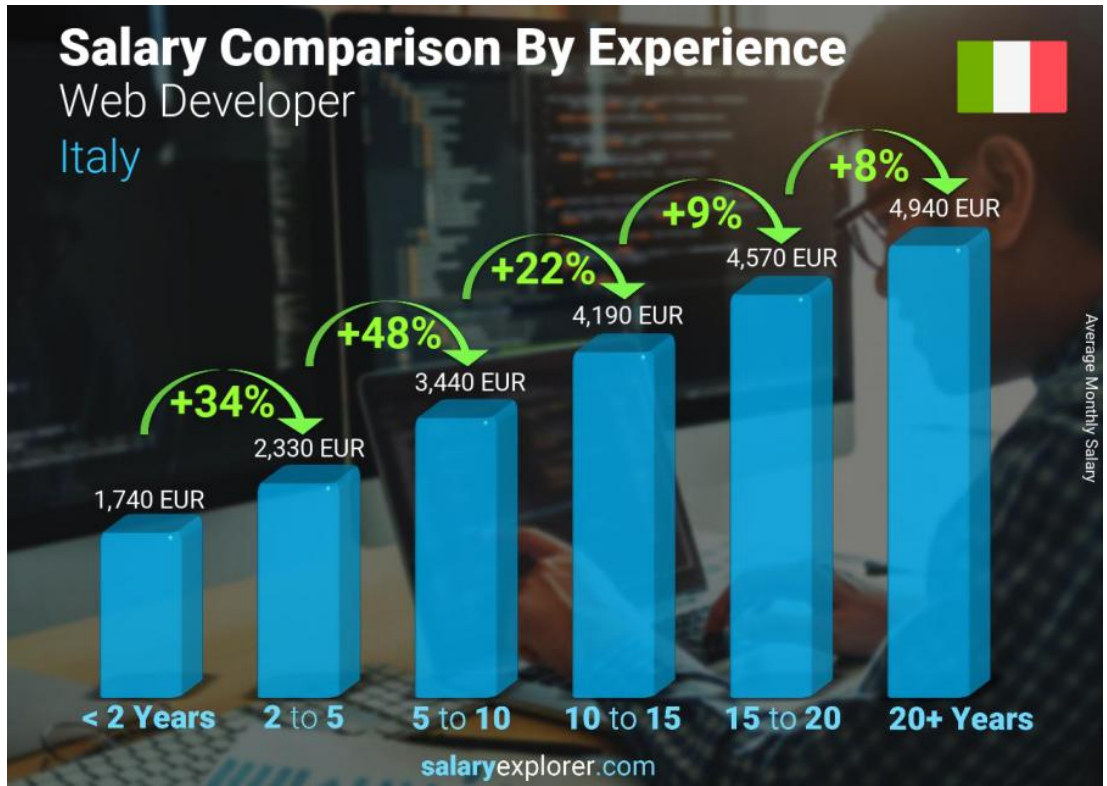
Extended reference for the review collection to show the hotel info.

Domanda 15

Risposta non data

Punteggio max.:

0,25



Question

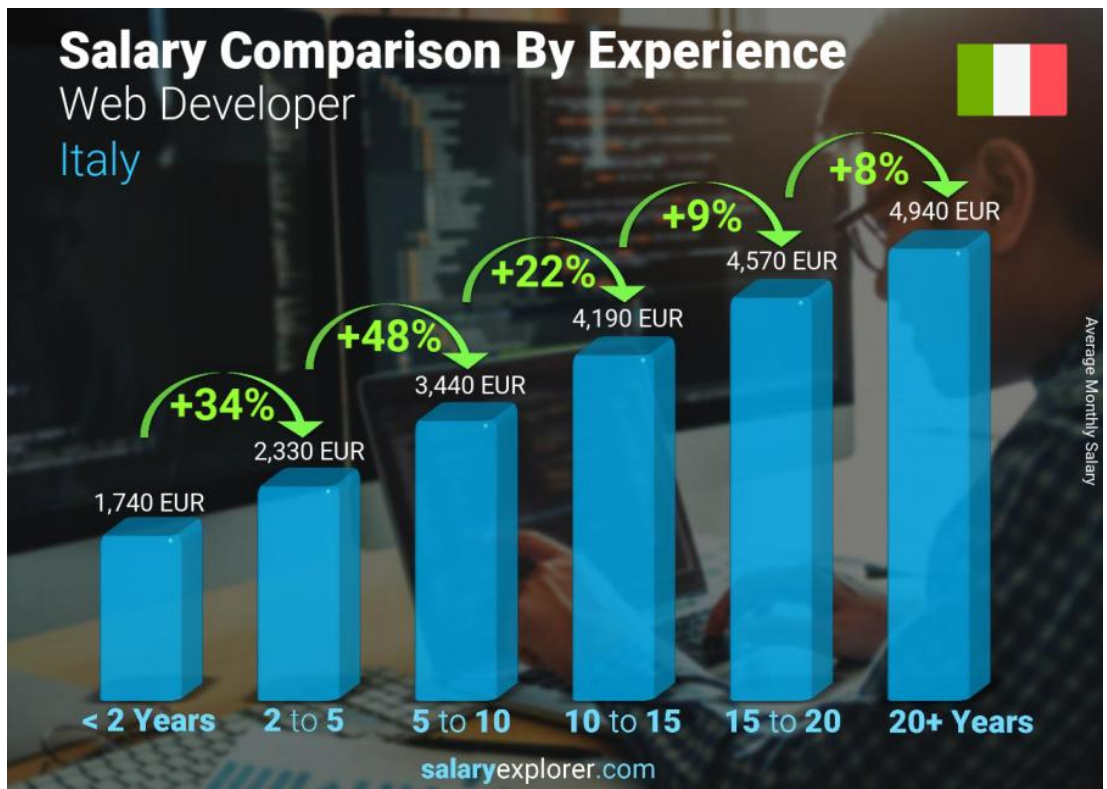
Is there a clearly defined question addressed by the visualization? Write it down.

Domanda 16

Risposta non data

Punteggio max.:

1,25



Data

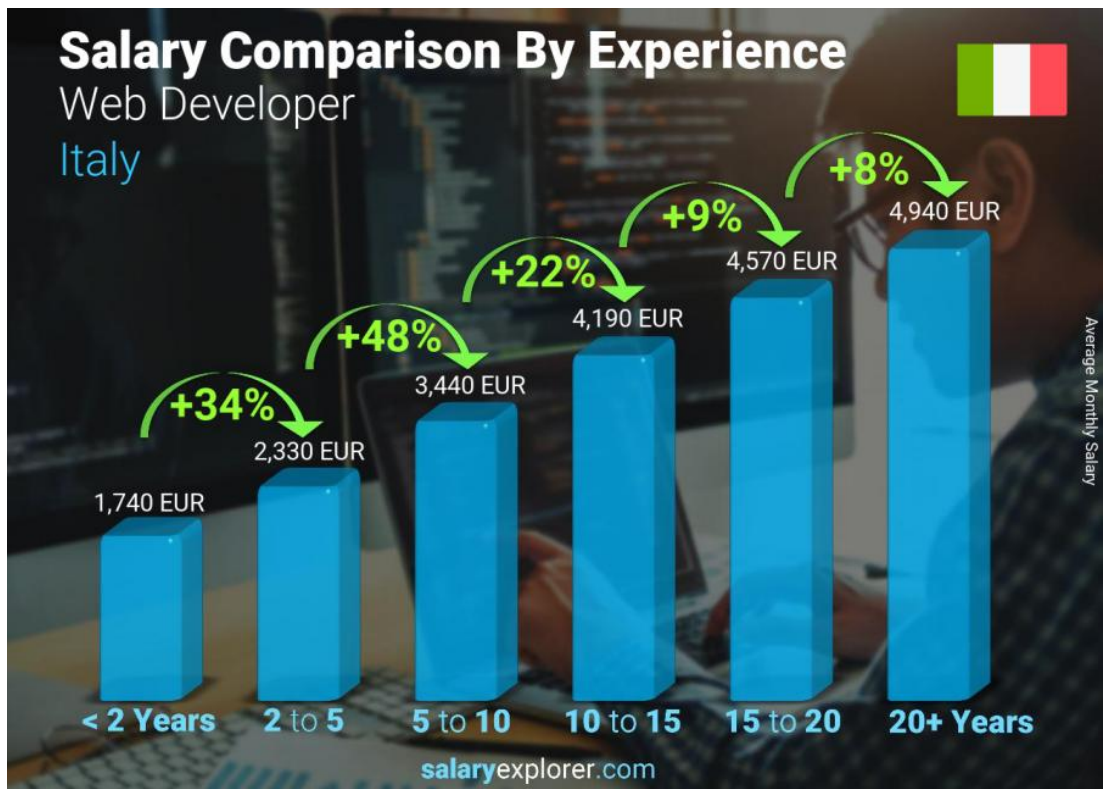
Is the data quality appropriate? Identify the inadequate characteristics and explain.

Domanda 17

Risposta non data

Punteggio max.:

0,75



Visual Proportionality

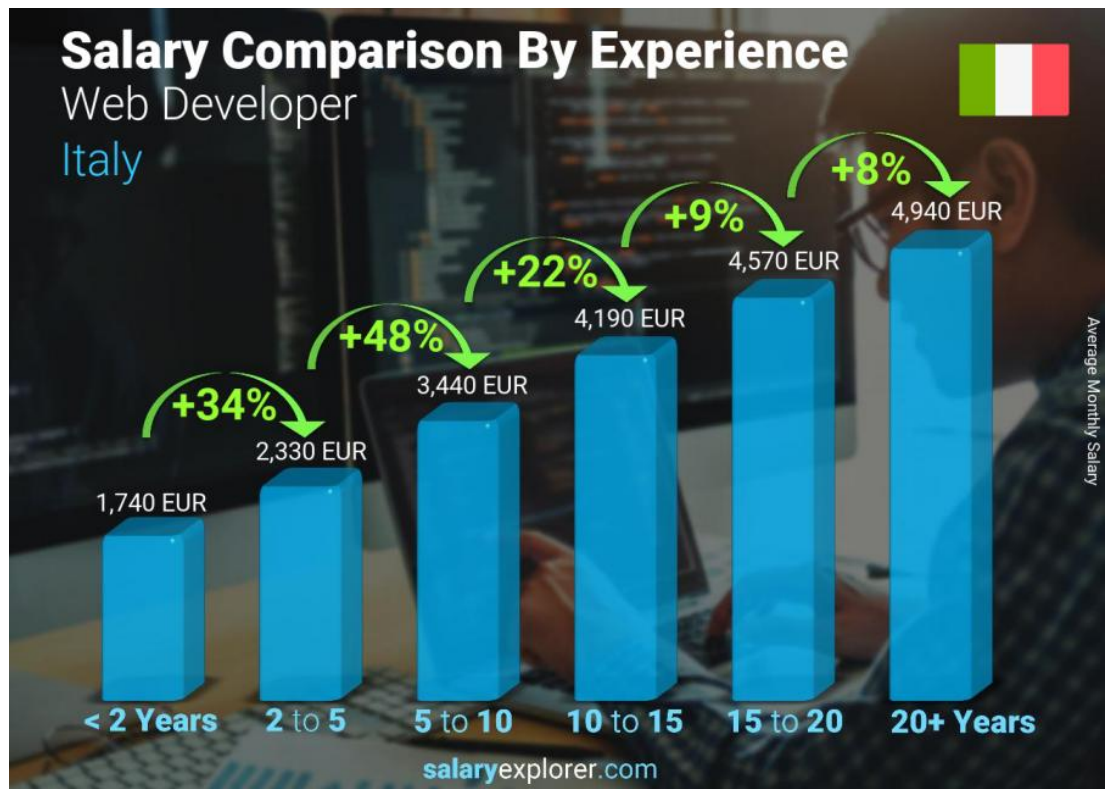
Are the values encoded in a uniformly proportional way?

Domanda 18

Risposta non data

Punteggio max.:

0,75



Visual Utility

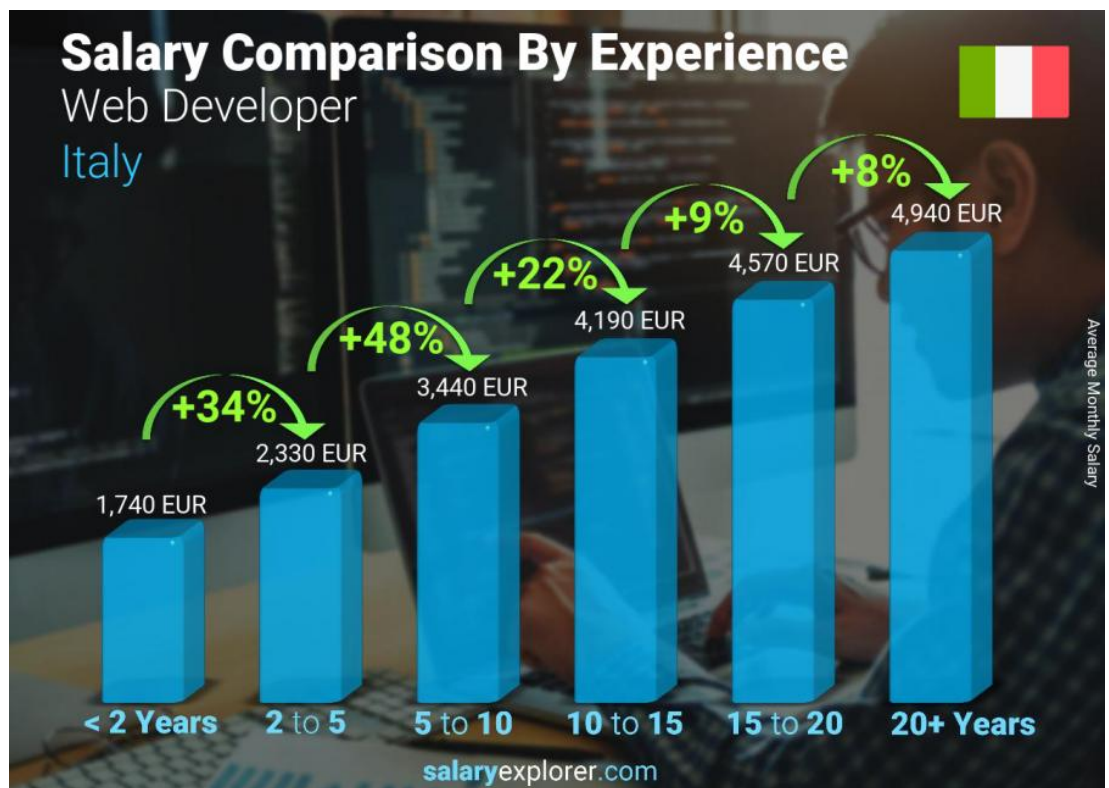
All the elements in the graph convey useful information?

Domanda 19

Risposta non data

Punteggio max.:

0,50



Visual Clarity

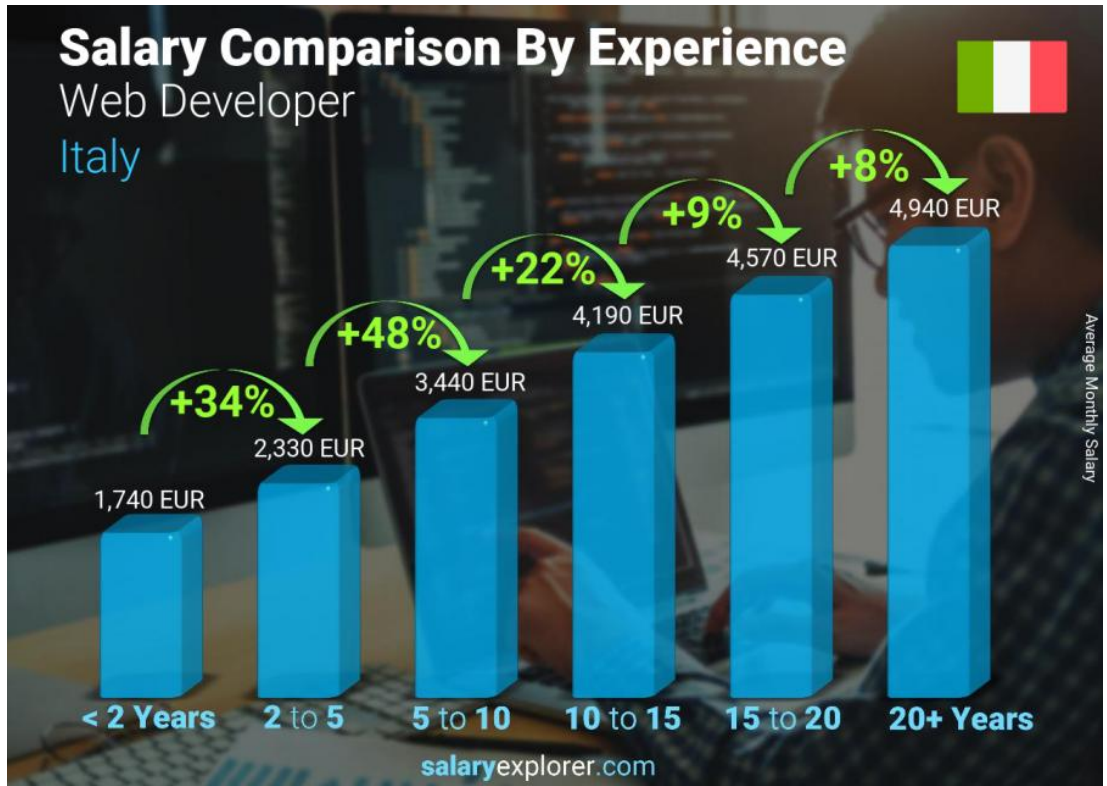
Are the data in the graph clearly identifiable and understandable (properly described)?

Domanda 20

Risposta non data

Punteggio max.:

0,25



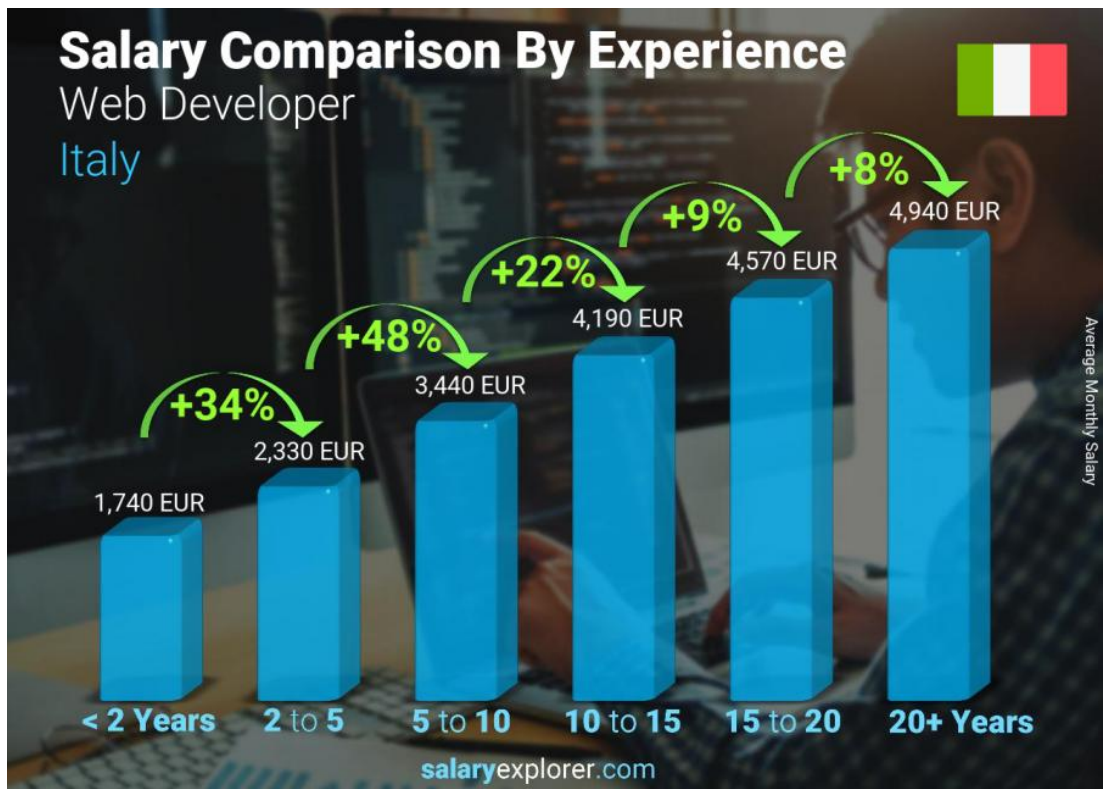
Design data

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.
