CSCI-585 Spring 2023 Midterm Exam Rubrics

Q1 [3+2 = 5 points].
a. In an EER diagram, we could have overlapping subtypes, as you know. Assuming there are 3 subtypes A, B, C, what are **three** different ways of modeling such a situation? Illustrate by drawing tables.

Assume the following attributes for each entity:
i. Supertype S: S_ID (Primary key), S1, S2
ii. Subtype A: S_ID (Primary key, Foreign key), A1, A2
iii. Subtype B: S_ID (Primary key, Foreign key), B1, B2
iv. Subtype C: S_ID (Primary key, Foreign key), C1, C2

Solutions:
1. Universal table with mapping table

S_ID	S1	S2	Subtype_ID	A1	A2	B1	B2	C1	C2

Subtype_ID	Subtype_type
000	Unused
001	A only
010	B only
011	C only
100	A and B
101	B and C
110	A and C
111	A, B and C

There is a single table containing all attributes of all entities. A subtype column contains a binary string/integer referenced from a mapping table. Each tuple contains values only for the attributes that are common (S1, S2) and for the attributes that are specific to its subtype(s).
Pros: simple; easy to maintain if the number of subtypes is fixed and known beforehand.

Cons: highly inflexible, wastes storage due to null values.

2. Universal table only

S_ID	S1	S2	Subtype_A	Subtype_B	Subtype_C	A1	A2	B1	B2	C1	C2

Instead of a mapping table and the Subtype_ID column, there is a boolean attribute for each subtype. A value of 1 indicates that the tuple belongs to that subtype.
Pros: same as #1.
Cons: even more inflexible that #1.

3. Supertype table with separate subtype tables

S_ID	S1	S2	Subtype_ID

S_ID	A1	A2

S_ID	B1	B2

S_ID	C1	C2

Pros: highly flexible, modular, loosely coupled.
Cons: to view all the data for a given S_ID, you have to first retrieve its Subtype_ID and then query the concerned tables.

4. Subtype tables only

S_ID	S1	S2	A1	A2

S_ID	S1	S2	B1	B2

S_ID	S1	S2	C1	C2

Pros: one fewer table to maintain.
Cons: since there is no way to know the subtype, you have to query each table for a given S_ID. Redundant storage of common attributes.

Grading:
+ 1 pt for each unique method with tables
- 0.5 pt per method if only explanation provided without table
- 0.5 pt overall if any notation other than tables used
Students may have other solutions.
Explanation provided here is for understanding only.

b. When we talk about entity supertypes and subtypes in EER, we are making an analogy with a class hierarchy e.g. a C++ or Java one. But the analogy between a table and a typical class isn't quite accurate. Why not? And what would make them equivalent, conceptually speaking?

The analogy isn't accurate because:
i. There is no inheritance between supertypes and subtypes. It is simply a visualization of the human understanding of relationships between entities. Eventually, all entities are implemented as tables with perhaps a primary key-foreign key relationship between them.
ii. As a result of i, subtypes-supertypes are more tightly coupled than superclasses-subclasses. A subclass can be instantiated independent of the superclass. On the other hand, if a primary key-foreign key relationship exists between a supertype and a subtype, you cannot add a row in the latter without first adding one in the former.
iii. Class inheritance is a mechanism for code reuse and polymorphism, whereas supertype-subtype is a mechanism for modeling complex entities with shared characteristics. Inheritance allows a class to inherit the behavior of its parent class and also to define its own behavior. In contrast, supertype-subtype is used to model complex entities with shared attributes and relationships, where the subtype inherits the attributes and relationships of the supertype and may also have its own unique attributes and relationships.

iv. **A table can hold only data whereas a class may have methods/functions. Hence, you cannot implement function overloading and overriding in tables.**

Just mentioning iv. is sufficient :) - Saty

How to make them equivalent:
i. Let the subtype implementation inherit the supertype's attributes and thus allow the subtype to be an actual extension of the supertype.
ii. Add behavior to entities using methods/functions and allow these to be inherited as well.

Grading:
+ 0.5 pt for listing at least 1 difference
+ 0.5 pt for listing at least 1 way to overcome the difference
Let students discuss the difference and the solution in their own words as long as the answer is coherent and their understanding is correct.

Q2 [1+4 = 5 points]
a. What is the benefit of normalization, what is its drawback?
Normalization aims to improve the database structure in order to create an appropriate database design. The main goal of database normalization is to minimize data redundancies. By reducing redundancies in the database, data anomalies (e.g. insert/update anomalies) will be reduced as well.

Benefit & Drawbacks [Total 1 point]

The **benefits** of normalization include:

- Reducing data redundancy and inconsistency, which can improve data integrity and accuracy.
- Improving data consistency and maintainability, which can simplify database design and management.
- Reducing data storage requirements and improving query performance, which can save disk space and processing time.

The **drawbacks** of normalization include::

- Increasing the complexity of database design and management, which can make it harder to understand and maintain the database.
- **Reducing query performance for complex queries, which can slow down the response time of the database.**
- Limiting the flexibility of the database schema, which can make it harder to adapt to changing requirements.

Grading:
+ 0.5 pt for listing at least 1 benefit
+ 0.5 pt for listing at least 1 drawback

Please note it is an open ended question, so any of the benefits/drawbacks mentioned above and/or other are correct, if justified

*****

b. In a class of students, each student has an ID and a name. Each student is assigned (given) a book by a popular author to read; many students could be assigned the same book (e.g. many might be assigned to read 'The adventures of Tom Sawyer', by Mark Twain). The class teacher uses the spreadsheet to keep the track of the # of hours a student puts in, towards the reading her/his book.

Show using a table, how the teacher would store data incorrectly. Show how you would help fix the table. To save time when you answer, you can use 'simple' values A, B, C.. for your data [they don't need to be 'real']

How the teacher would store data incorrectly [Total 2 points]

Each row represents a student and their assigned book, along with the number of hours they have put in towards reading it. The columns represent the different attributes of the data, such as ID, Name, Book, Author, and Hours

Example 1 :

ID	Name	Book	Author	Hours
1	A	Book1	Author1	10
2	B	Book2	Author2	12
3	C	Book3	Author3	8
4	D	Book1	Author1	15
5	E	Book2	Author2	10

Example 2 :
This table is not normalized because it contains multiple attributes in repetition

Table : Books

ID	Book	Author
1	Book1	Author1
2	Book2	Author2
3	Book3	Author3
4	Book1	Author1
5	Book2	Author2

Table : Student and Books

ID	Name	Book
1	A	Book1
2	B	Book2
3	C	Book3
4	D	Book1
5	E	Book2

Table : Student and hours

ID	Name	Hours
1	A	10
2	B	12
3	C	8
4	D	15
5	E	10

Grading:

[Full points] +2 point for representing data in any of the above forms /denormalized form/ forms showing redundancy

[Partial points] +1 point for any valid explanation and justification of why they find data representation to be incorrect by the teacher.

-0.5 point for missing the concept of normalization/denormalization to represent data

-1 point for missing any fields or their explanation or incorrect justification

Show how you would help fix the table [2 points]

Example 1:

Table 1 represents the students in the class, with each row representing a different student and their ID (SID) and name. Table 2 represents the books assigned to the students, with each row representing a different book and its author. Table 3 represents the assignments of books to students, with each row representing a different student and book combination. Table 4 represents the number of hours each student has put in towards reading their assigned book, with each row representing a different student and the number of hours they have read.

Table 1: Students

SID	Name
1	A
2	B
3	C
4	D
5	E

Table 2: Books

Book	Author
Book1	Author1
Book2	Author2
Book3	Author3

Table 3:Assignments

SID	Book
1	Book1
2	Book2
3	Book3
4	Book1
5	Book2

Table4: Hours

SID	Hours
1	10
2	12
3	8
4	15
5	10

Example 2:

Table 1: Students

SID	Name
1	A
2	B
3	C
4	D
5	E

Table 2: Books and student

Book	Author	Name
Book1	Author1	A,D
Book2	Author2	B, E
Book3	Author3	C

Table4: Hours

Name	Hours
A	10
B	12
C	8
D	15
E	10

TO FIX IT, ALL WE NEED TO DO, IS STORE (SID, Book, Hours) - this is similar to the (EmployeeID, ProjectID, Hours) in the lecture slides. - Saty

Grading:

[Full points] +2 point for representing data in the above form/ normalized form

[Partial points]

+1.5 points for representing data in any other normalized form partially

+1 point for any valid explanation and justification of why they find data representation to be correct by the teacher.

+0.5 points for representing at least 1 table correctly

-1 point for missing any fields or their explanation or incorrect justification
-1 point for redundant data representation as it misses the entire point of normalization

Note : all other normalized forms can be given partial credits based on how close they are to the correct representations

*****

Q3: [1+2+2] = 5 points

> Q3 [1+2+2 = 5 points].
>
> a. What is an example of data that is suitable for a single-user DB? What is another example, for a special-purpose DB?
>
> b. Why is structural dependence a bad thing, when it comes to storing data? Illustrate structural dependence using a small example of your own [with some sample data].
>
> c. On the flip side (of structural dependence), we have layered data abstraction - what benefit does layering provide? Explain in two or three sentences (NOT more!).

a. Example of data for single-user DBs (0.5 pts):
Data related to confidential information; personal information and so on. **(also, contact list on our phones)**

Example of data for special-purpose DBs (0.5 pts):
Any data that have specific use cases: students' midterm grades, stock price, weather record of a local area, **and, molecule data, building info systems ("BIM"), CAD drawings' 'bill of materials' (BOM), etc.**

Grading:

1. One type of data for single-user DBs +0.5pts
2. One type of data for special purpose DBs +0.5pts
3. If only answer the examples of DBs, one can only get 0.5pts for this question.

*** pitfall: what is an example of **DATA** that is suitable for single-user/special-purpose DB.

b. Why is structural dependence a bad thing when it comes to storing data (1pt)?
When data is stored with structural dependence, the drawbacks are (at least) twofold: i) it may cause trouble for applications that want to add/delete/modify/query the data but do not know the structure of the data. ii) when the data structure changed, all the applications that used to customize to the old structures need to be modified.

Illustrate structural dependence using a small example of your own(1pt):
A list of data containing the name and the name and DOB of students. When we want to add another attribute of students to the list, e.g. address, all the applications that are

used to query the data list using the old order should be changed or they are not able to query the data correctly.

Alex
05/02
Bill
06/13
Chris
01/31

Grading:

1. Only one drawback +0.5pt; at least two drawbacks +1pt
2. Examples related to data structures +0.5 pt; explanations +0.5 pt

c. What benefit does layering provide (one point: 1pts; >= three points: 2pts)?

Development efficiency, flexibility; Data reusability; Database design flexibility.

Grading:

1. The first mentioned benefit +1pts
2. Each one more benefit +0.5pts.

Q4 (5 points)

Q4 [1+2+2 = 5 points].

a. In the 'story' of connectivity, about how it all started, leading to where we are today, what were key stops along the way, ie. what were milestones? You can simply list them, no need to elaborate.

b. How does data connectivity occur today, ie what is the dominant architecture? Explain in your own words, using your own diagram.

c. Briefly discuss two ways via which the web server was (is) augmented to serve data to the client.

a. Story of connectivity (1point)

Grading:

1. Provides a clear and concise list of key milestones in the history of connectivity, ranging from the invention of the telegraph to the impact of the COVID-19 pandemic

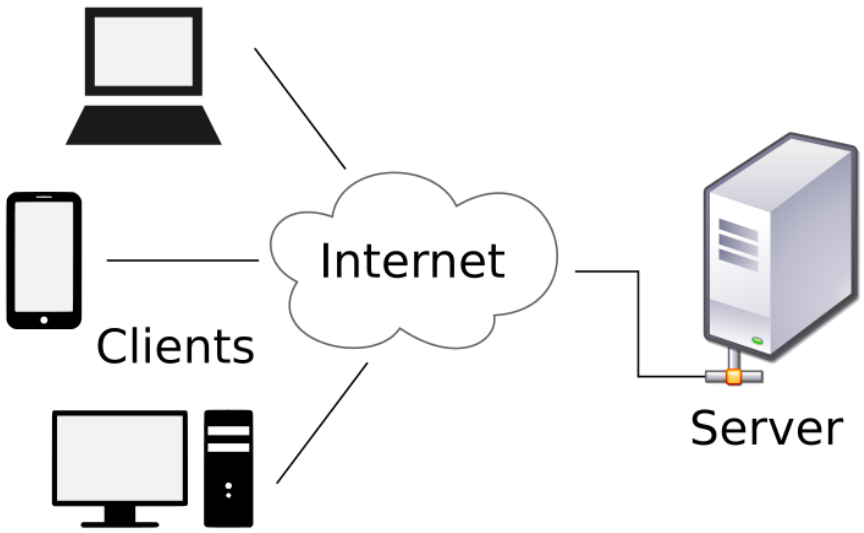
on remote communication technologies in 2020 (This is just an example, just the story line should match the history). **[+1point]**
Partial grading points: 0.5 pts

The story can start with MS and Sun providing ODBC and JDBC respectively, then continuing with server-side scripting (eg. 'cgi', CFML etc), with microservices+container+cloud being today's tech.

b. Architecture (2 point)

Grading:

1. Provides a clear and accurate description of the dominant architecture for data connectivity today, explaining how client devices communicate with servers over the internet to request and receive data. **[+1point]**
2. Includes a diagram that effectively illustrates the client-server architecture, demonstrating an understanding of how data connectivity occurs in modern applications and services. **[+1point]** (Attaching a sample architecture for reference, it might differ from the student's diagram)



If no diagram, give partial points for description [+0.5 points]
If no description, give partial points for diagram [+0.5 points]

The dominant architecture would be REST or GraphQL APIs, or even 'MCC' - microservices, container, cloud (where REST/GraphQL runs in containers on a cloud server).

c. Briefly discuss the two ways via which server serves data to client

The server can either send raw data (XML or JSON or plaintext etc) to the client, or, add markup to data and send HTML.

Grading:

1. Provides a clear and accurate description of two common methods for augmenting web servers to serve data to clients, (Example: Caching, and Compression). **[+1point]**
2. A Brief explanation of 1st method **[+0.5 points]**
3. A brief explanation of 2nd method **[+0.5 points]**
4. **Partial Points [+1 points]**: if no methods are specified but has explained the working

Q5 [4+1 = 5 points]
a. Pick any two apps/sites on your phone/tablet/laptop using which you access data, describe how your UI actions (eg. searching, or doing data filtering) might result in SQL, using one example for each app/site (so two examples total).

Example 1: Doordash
When a user searches for restaurants in a specific area, Doordash's UI would trigger an SQL query that filters all restaurants based on location from the database. For example, if a user searches for "Mexican food" in "San Francisco", the UI would trigger an SQL query such as:

SELECT * FROM restaurants WHERE cuisine = "Mexican" AND city = "San Francisco"

Example 2: Amazon
When a user filters a search result by a specific category or price range, Amazon's UI would trigger an SQL query that retrieves data from the database based on the filter criteria. For example, if a user searches for "smartphones" and filters by "under $500", the UI would trigger an SQL query such as:

SELECT * FROM products WHERE category = "smartphones" AND price < 500

Grading:
Example 1: (0.5 mark for stating the name of the app + 1.5 mark for description)
Example 2: (0.5 mark for stating the name of the app + 1.5 mark for description)

Please note that writing a SQL query is optional. Since this is an open-ended question, there could be many valid different apps that students may write (for example. Lyft, Uber, Instagram, Spotify, Airbnb, etc*).* ***Please consider all valid answers.***

b. Assuming (like in 'a' above) that your app-driven-querying does turn into SQL, where would such conversion (ie. transformation from UI-based query to SQL) occurs?

The conversion of a UI-based query to SQL occurs on the server-side of the app/site with the help of a web-to-database middleware. Once the user initiates a query through the UI, the request is sent to the middleware from the server, which processes the request and translates it into SQL. The middleware then sends the SQL query to the database, retrieves the relevant data, and then formats the data and sends it back to the server, which finally sends it to the UI for display. The web-to-database middleware acts as an intermediary between the server and the database, handling the translation of requests and queries between the two systems.

Grading:
+1 for any valid explanation. (Student must write either “backend server” or “web-to-database middleware” in the answer, -0.5 if not written)

Q6. [2+2+1 = 5 points]

a. In 2PL for data access during transactions, what is the most important phase? Explain.

   Sample Answer 1 based on Conservative/Static 2PL:
   Growing phase. A transaction acquires all the locks it needs to read and modify data items in the growing phase. Once a lock is acquired, it cannot be released until the transaction commits or aborts. This makes sure that the same data cannot be updated by another transaction while it is being read or modified by the current transaction.

   Sample Answer 2 based on 2PL:Locking phase.
   A transaction acquires all the locks on the data items it needs to access before making any modifications or performing any other operations on them in the locking phase. This makes sure that the transaction has exclusive access to the data items and prohibits concurrent modification of them by other transactions, which can result in data inconsistency and other problems.

   Grading:
   +1: For mentioning Locking phase/Growing phase/Phase 1
   +1: Valid explanation for the phase mentioned

*****

Alternately, a student might consider the EXECUTION phase to be the most important (the "pyramid plateau" I discussed in class, where we can think of 2PL as 3PL in fact: lock acquisition, EXECUTION, lock release) - this is ok, too.

b. In 2PL, when we release locks, if we release locks prematurely, what issue might that cause? How would we fix it?

Sample Answer:
Issues: Data Inconsistency, Incorrect Reads, Lost Update, Rollbacks, Inconsistent retrievals

Prevention: Release locks only after transaction fully finishes.

Grading:
+1: Identifying at least one correct issue as mentioned
+1: Valid technique for prevention

*****

c. What issue might arise, when we do TM without locks? How would we fix it?

Sample Answer:
Issue: Dirty Read or Data Inconsistencies

Prevention:

- Use Concurrency Control Techniques.
- Use locking mechanism: Shared and Exclusive Locks
- Use Recovery Managers

A different answer: there might be data corruption on account of different transactions overwriting cell values; we would fix it by analyzing the log files that record transactions, identify errors and manually rolling back the offending transactions. (Background, for the graders: T.M without locks is only rarely practiced, when we are sure that such collisions will be extremely rare).

Grading:
+0.5: Identifying at least one correct issue as mentioned
+0.5: Valid technique for prevention

*****

Q7(1+4 = 5 points)

a.Codd's relational operators for data processing, lead to 'closure'. Why is this advantageous? Illustrate.

Sample Answer:

The advantage of closure is that it allows for efficient and effective data processing.
In addition, closure ensures that the data stored in the database is consistent and free of redundancy. Since the relational operators are based on set theory, they ensure that each relation contains only unique and relevant data. This eliminates the need for duplicate data storage and reduces the risk of inconsistencies and errors in the data.
Overall, the use of relational algebra and closure leads to more efficient and effective data processing, improved query performance, and better data consistency and integrity. It is therefore a fundamental concept in the design and implementation of relational databases.

Better answer: closure permits CHAINING of operations, ie. creating a tree (dataflow) of SQL operations.

Grading:

Keywords included + 1 (other similar keywords may also be acceptable)
Keywords: **Efficient, flexible**
Integrity and **Consistency** (offers a standard way).
Simplify the design and maintenance of the database.
Mentioning operation 'chaining' (or equivalent) is sufficient.

b. For these four data types, list an operation that does preserve closure, and one that does not: vector (e.g. with components x,y,z), matrix, complex number, color (with RGB components).

Sample Answer:

(1 point) 1. Vector:
(0.5point) Preserve: Vector addition, vector cross product keeps closure R^3 -> R^3
(0.5point) Not preserve: Dot product, doesn't keep closure R^3->R^1
(1 point) 2. Matrix:
(0.5point) Preserve: matrix-scalar multiplication, matrix addition etc.
(0.5point) Not preserve:
e1.matrix-matrix multiplication, it changes the matrix's dimension (only square matrix is an exception.)
e2. MaxPooling, or finding the determinant (which produces a scalar)
(1 point) 3. Complex Number:
(0.5point) Preserve: operations like conjugation, addition etc.
(0.5point) Not Preserve:
e1. Modulus
e2. Defined operation x->(x, x*2) from C-> C^2
(1 point) 4. Color:
(0.5point) Preserve:
brightness change, contrast change, color<->gray

Operations remaining in the same color channel domain(0,255) are also acceptable. E.g. color value/255, color normalization. From 0,255->0,1

(0.5point) Not Preserve:

Computing luminance [0.7*R+0.2*G+0.1*B] returns a float

Grading:

Other operations are acceptable. Please note operation is some form of computation; it can be an operator symbol or can be a function. 'Closure' means that the operation will output the same type as input.

Thus, any operation that makes results remain in the same domain as input could be regarded as a 'closure' operation and vice-versa.

Q8 [2+2+1 = 5 points]

1. During 2PC in distributed transactions, the transaction coordinator might fail. How would we fix that?
2. During 2PC, a non-coordinator site might fail partway (between phase 1 and phase 2) - how would we fix the problem (ie prevent bad transactions)?
3. During 2PC, a non-coordinator site might fail at the start (before phase 1) - how would we deal with that?

Sample Answer:

1. We can set up a **backup coordinator** or order a non-coordinator site (one that connects to every site) to work as a backup coordinator when the original one fails. The backup coordinator should have **access to the same transaction logs** and other relevant information as the original coordinator, so that it can continue the 2PC process from where the failed coordinator left off. If the coordinator falls between phase 1 and phase 2, none of the non-coordinator machines will get a COMMIT or ABORT message. After a **timeout period**, the non-coordinators will notify the designated backup, which will take over the reminder of the task
2. We can use a **timeout mechanism**: after a certain time not hearing back from a site, the timeout mechanism is called and the coordinator auto marks the transaction as failed and **reschedules** the tasks so this failed site will not be used in the next time.
3. Before the start of any transaction, we can set a protocol that for the coordinator to **send a message** to all the non-coordinator sites, simply ask a message back to the coordinator. If any of the sites is not responsive after a certain amount of time, the coordinator can mark that site as failed and schedule the tasks so this failed site will not be used in the next time.

Grading:

1. (2 points total) **+1:** if mension mechanism of "Backup coordinator" or "a non-coordinator site function as the coordinator"; **+1:** for the detail of the mechanism, including "back-up

coordinator should have the access to the log file", or "if fail down happened between phase 1 and phase 2, a timeout mechanism will be used", or other reasonable detail.
2. (2 points total) **+1:** if mension mechanism of "timeout"; **+1:** for the detail and explanation of how the mechanism works.
3. (1 point total) **+1** if mension sends a message to all sites before any transaction starts.

Other reasonable answers are okay, 1 and 2 require a brief explanation to get full points.

Q9. [1+2+2=5 points]

a) What are a couple of uses for 'computed columns' ? [1+2+2=5 points]

Computed columns are virtual columns in a database table that derive their values from expressions or functions that operate on other columns in the same table. Here are a few uses of computed columns:

1. Calculation of values: Computed columns can be used to calculate values that are based on other columns in the table. For example, you could use a computed column to calculate the total price of an order by multiplying the quantity ordered by the unit price.

2. Data transformation: Computed columns can also be used to transform data from one format to another. For example, you could use a computed column to format a date column in a specific way, or to concatenate columns into a single string.

3. Data validation: Computed columns can be used to enforce data validation rules. For example, you could use a computed column to ensure that the values in a column meet certain criteria, such as being within a certain range.

4. Indexing: Computed columns can be used as part of an index to improve query performance. For example, you could create an index on a computed column that combines the values of several columns, so that queries that use those columns will be more efficient.

Overall, computed columns can help simplify data management and improve data integrity by reducing the need for manual data manipulation and improving data consistency.

Grading:
+1: If at least one among the following are mentioned correctly.

b) Given a table with columns of sines and cosines (for 0 to 360 degrees, in increments of 1 degree), eg called COS and SIN, how would you verify the following formula/ identity?

To verify sin^2 Q+ cos^2Q=1 form the given table with attributes SIN and COS,

- Perform (SIN*SIN) +(COS*COS) and add the value to new column say result. If the value is 1 then return True else return False.
- Sample code: ALTER TABLE myTable ADD result AS (sin* sin + cos * cos)
- This will add a new computed column named `result` to the table `myTable`. The computed column will contain the result of the calculation sin^2 + cos^2 for each row in the table.
- You can then query the `result` column to retrieve the calculated values: SELECT result FROM myTable.

Grading:
+2: If similar steps are written. (Either theoretically or using SQL command).

c) Given a table with a pair of columns called X and Y, containing (x,y) values from a scatterplot for example, how would you calculate the (Pearson) correlation coefficient? Again, just describe the steps [no need for SQL].

(+1 point)

Here's a step-by-step explanation of how to calculate the Pearson correlation coefficient between two columns X and Y:

1. Calculate the mean of each column: Compute the mean (average) of each column by adding up the values in each column and dividing by the number of values (n).

Mean of X: $\mu_x = \Sigma(x_i) / n$ Mean of Y: $\mu_y = \Sigma(y_i) / n$

2. Compute the deviations from the mean: For each value in the columns X and Y, subtract the mean of the respective column to get the deviation.

Deviation of X: $x_i - \mu_x$ Deviation of Y: $y_i - \mu_y$

3. Calculate the product of the deviations: Multiply the deviations for each corresponding pair of values from columns X and Y.

Product of deviations: $(x_i - \mu_x)(y_i - \mu_y)$

4. Sum the products of the deviations: Add up the products of the deviations calculated in the previous step.

$\Sigma[(x_i - \mu_x)(y_i - \mu_y)]$

5. Calculate the sum of squared deviations for each column: Square the deviations for each value in columns X and Y, and then sum them up.

$\Sigma(x_i - \mu_x)^2$ and $\Sigma(y_i - \mu_y)^2$

6. Calculate the Pearson correlation coefficient: Divide the sum of the products of the deviations (step 4) by the square root of the product of the sum of squared deviations for X and Y (step 5).

(+1 point)

$\text{Pearson's } r = \Sigma[(x_i - \mu_x)(y_i - \mu_y)] / \sqrt{[\Sigma(x_i - \mu_x)^2 * \Sigma(y_i - \mu_y)^2]}$

That's it! The result, Pearson's r, will be a value between -1 and 1, indicating the strength and direction of the linear relationship between the variables in columns X and Y.

Grading:
+1: If the parameters required to measure correlation coefficient are similarly represented.
+1: If the Pearson's correlation coefficient is represented correctly.