STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Group No.: 11

Name	Email
Dylan Lasrado	dylan.lasrado@tamu.edu
Pragati Naikare	pragatinaikare311@tamu.edu
Romit Madhukar Bonkar	romit.bonkar@tamu.edu
Tanmay	tanmay@tamu.edu

STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

---

1. **The Transactions table violates the first, second, and third normal forms. For each level of normalization, explain why the Transactions table is in violation. Be specific.**

First Normal Form:

The Transactions table violates the First Normal Form as the table has repeating groups and the "Item1", "Item2" and "Item3" columns are not atomic. The Item column is repeated 3 times in the table as "Item1", "Item2" and "Item3" respectively. Each of these Item columns is not atomic as they contain two different pieces of information namely the type of the item and its price and hence can be further split into two columns. These columns would each reflect the 'Type of the item' and the 'Price of the item'.

Second Normal Form:

The Transactions table as well as the newly created OrderLine table violate the Second Normal form. In the Transactions table, CustomerName and CustomerStatus have a partial dependency on the composite key and can be uniquely recognized by using only CustomerID. Similarly, in the OrderLine table, the Cost can be identified by referring only to the Item column.

Third Normal Form:

The Transactions table has a transitive dependency, wherein the Promotions column is dependent on order type, customer status, and items which in turn is dependent on the customer ID and transaction date as a composite key.

2. **Create a collection of tables representing all of the data in Transactions that satisfies first, second, and third normal forms and does not contain redundant or duplicated information across the collection of tables. All data in the Transactions table should be represented somewhere in your collection of tables, but you can remove any redundant information. Provide complete tables with all of the rows and columns filled in, not just the names of the columns.**

The final database contains 5 tables:

STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Transactions Table			
TransactionDate	CustomerID	OrderType	StoreCredit
2023-09-01 10:00:00	1001	Online	$ 50
2023-09-01 10:00:00	1002	In-store	$ 30
2023-09-05 12:45:00	1003	Online	$ 10
2023-09-06 09:15:00	1001	In-store	$ 45
2023-09-10 16:45:00	1001	Online	$ 40
2023-09-12 11:30:00	1002	In-store	$ 25

Customer Table		
CustomerID	CustomerName	CustomerStatus
1001	Juan Johnson	Gold
1002	Ahmed Al-Masri	Silver
1003	David Lee	Bronze

Item Table	
Name	Cost
Laptop	$1200
Gaming Console	$500
Headphones	$100
Tablet	$600
Running Shoes	$100
Home Theater System	$700
Smartphone	$800
Coffee Maker	$50
Fitness Tracker	$80
Scooter	$80

STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Order Line Table			
TransactionDate	CustomerID	Name	Quantity
2023-09-01 10:00:00	1001	Laptop	1
2023-09-01 10:00:00	1002	Gaming Console	1
2023-09-05 12:45:00	1003	Headphones	1
2023-09-06 09:15:00	1001	Tablet	1
2023-09-10 16:45:00	1001	Running Shoes	1
2023-09-12 11:30:00	1002	Home Theater System	1
2023-09-01 10:00:00	1001	Smartphone	1
2023-09-05 12:45:00	1003	Coffee Maker	1
2023-09-10 16:45:00	1001	Fitness Tracker	1
2023-09-05 12:45:00	1003	Scooter	1

Promotions Table		
TransactionDate	CustomerID	Promotion
2023-09-01 10:00:00	1001	GoldStatusFreeShipping
2023-09-05 12:45:00	1003	BronzeStatusFreeShipping
2023-09-06 09:15:00	1001	GoldStatus10Off
2023-09-10 16:45:00	1001	GoldStatusFreeShipping

3. **For each level of normalization, explain the specific steps you have taken to bring your new collection of tables into compliance with the first, second, and third normal forms. Be specific. State any assumptions made regarding the data needed to explain why your new collection of tables satisfies the third normal form.**

First Normal Form:

To convert the original 'Transactions table' into a form of tables that satisfy the First Normal Form, we have created two tables namely a 'Transactions Table' and an 'Order Line Table'. The 'TransactionsDate' and 'CustomerID' are the keys to the 'Transaction Table' while the 'TransactionsDate', 'CustomerID', and 'Name' are the keys to the 'Order Line Table'.

The repetition of columns is avoided by combining the 'Item1', 'Item2', and 'Item3' columns from the original 'Transactions table' into a single column 'Name' under the 'Order Line Table' which reflects the Type of Item. The Price of the item is shown under the 'Cost' column under the 'Order Line Table'. The information pertaining to order type, name and status of the customer, store credit available to the customer, promotions, and item count is part of the 'Transactions Table'.

Transactions Table

TransactionDate	CustomerID	OrderType	CustomerName	CustomerStatus	StoreCredit	Promotion
2023-09-01 10:00:00	1001	Online	Juan Johnson	Gold	$ 50	GoldStatusFreeShipping
2023-09-01 10:00:00	1002	In-store	Ahmed Al-Masri	Silver	$ 30	
2023-09-05 12:45:00	1003	Online	David Lee	Bronze	$ 10	BronzeStatusFreeShipping
2023-09-06 09:15:00	1001	In-store	Juan Johnson	Gold	$ 45	GoldStatus10Off
2023-09-10 16:45:00	1001	Online	Juan Johnson	Gold	$ 40	GoldStatusFreeShipping
2023-09-12 11:30:00	1002	In-store	Ahmed Al-Masri	Silver	$ 25	

Order Line Table

TransactionDate	CustomerID	Name	Cost
2023-09-01 10:00:00	1001	Laptop	$1200
2023-09-01 10:00:00	1002	Gaming Console	$500
2023-09-05 12:45:00	1003	Headphones	$100
2023-09-06 09:15:00	1001	Tablet	$600
2023-09-10 16:45:00	1001	Running Shoes	$100
2023-09-12 11:30:00	1002	Home Theater System	$700
2023-09-01 10:00:00	1001	Smartphone	$800
2023-09-05 12:45:00	1003	Coffee Maker	$50
2023-09-10 16:45:00	1001	Fitness Tracker	$80
2023-09-05 12:45:00	1003	Scooter	$80

Second Normal Form:

We need to separate the CustomerName and CustomerStatus columns from the Transactions table and create a new table, Customer, which will have CustomerID, CustomerName, and CustomerStatus. Similarly, we will separate the Cost column from the OrderLine table and create a new Item table, which will include ItemName and Cost columns.

STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Transactions Table				
TransactionDate	CustomerID	OrderType	StoreCredit	Promotion
2023-09-01 10:00:00	1001	Online	$ 50	GoldStatusFreeShipping
2023-09-01 10:00:00	1002	In-store	$ 30	
2023-09-05 12:45:00	1003	Online	$ 10	BronzeStatusFreeShipping
2023-09-06 09:15:00	1001	In-store	$ 45	GoldStatus10Off
2023-09-10 16:45:00	1001	Online	$ 40	GoldStatusFreeShipping
2023-09-12 11:30:00	1002	In-store	$ 25	

Customer Table		
CustomerID	CustomerName	CustomerStatus
1001	Juan Johnson	Gold
1002	Ahmed Al-Masri	Silver
1003	David Lee	Bronze

Item Table	
Name	Cost
Laptop	$1200
Gaming Console	$500
Headphones	$100
Tablet	$600
Running Shoes	$100
Home Theater System	$700
Smartphone	$800
Coffee Maker	$50
Fitness Tracker	$80
Scooter	$80

STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Order Line Table			
TransactionDate	CustomerID	Name	Quantity
2023-09-01 10:00:00	1001	Laptop	1
2023-09-01 10:00:00	1002	Gaming Console	1
2023-09-05 12:45:00	1003	Headphones	1
2023-09-06 09:15:00	1001	Tablet	1
2023-09-10 16:45:00	1001	Running Shoes	1
2023-09-12 11:30:00	1002	Home Theater System	1
2023-09-01 10:00:00	1001	Smartphone	1
2023-09-05 12:45:00	1003	Coffee Maker	1
2023-09-10 16:45:00	1001	Fitness Tracker	1
2023-09-05 12:45:00	1003	Scooter	1

Third Normal Form:

The "promotion" column in our database is currently dependent on multiple factors, namely "order type," "customer status," and "items purchased." This setup violates the principles of the Third Normal Form (3rd NF) in database design. To address this issue, we have two possible courses of action: either we can remove the "promotion" column entirely or create a new table to store all promotion-related data.

Option 1: Drop the "promotion" column completely

By doing this, we would simplify our database structure and adhere to 3NF principles.

However, if someone is new and not familiar with the rules for each promotion, they might face difficulties understanding how promotions work.

Option 2: Create a new table for all promotion-related data

This option allows us to retain all the information about promotions in a dedicated table. It provides a clear rule book for promotions, making it easier for new users or employees to understand how promotions work.

Our choice to create a new table for promotion-related data is motivated by the desire to retain information, improve data organization, and provide a user-friendly database that includes a comprehensive rule book for all promotions.

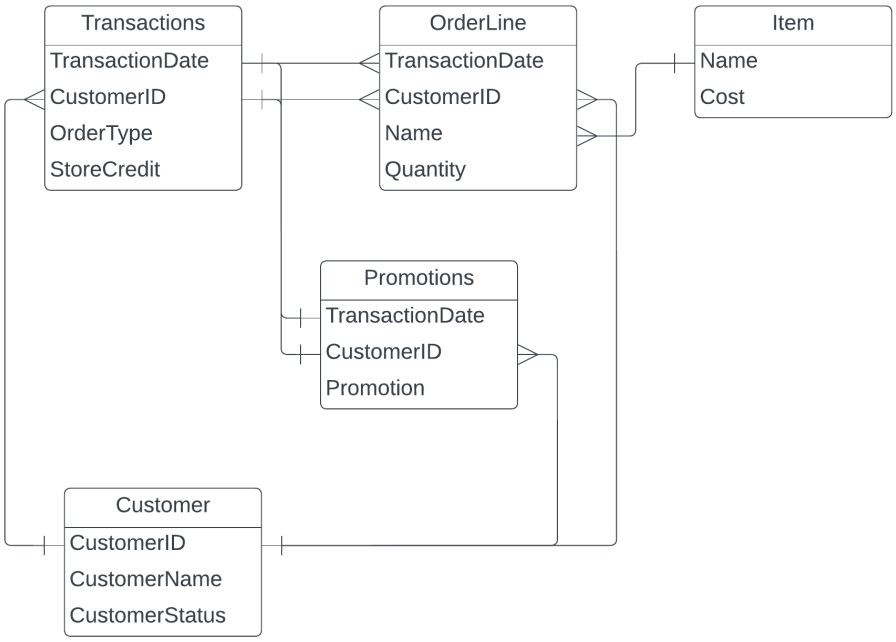
STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

---

4. Identify the primary keys for each table in your new collection of tables. Composite primary keys are acceptable.

Table	**Keys**
Transactions	Composite key (TransactionDate, CustomerID)
Item	Primary key (Name)
Order Line	Composite key (TransactionDate, CustomerID, Name)
Customer	Primary key (CustomerID)
Promotions	Composite key (TransactionDate, CustomerID)

5. Identify the types of relationships (one-to-one, one-to-many, many-to-many) among the fields in your tables. Be specific about which field corresponds to the 'one' vs. 'many'. Use the `primary key' and `foreign key' terminology when appropriate



STAT 624: COMPUTING TOOLS FOR DATA SCIENCE
NORMALIZE THIS - ASSIGNMENT

Contributions and Learning:

- **Dylan Lasrado**
  - ➢ **Contribution:**
    1. Participated in team discussions based on the interpretations of the 3 normal forms and possible solutions to the problem assigned.
    2. Converted the original Transactions table into the 1st Normal Form.
    3. Researched for examples through the slides provided and through examples on the internet to gain a deeper understanding of each normal form and how the conversion from one normal form to the next can be performed.
    4. Contributed towards the documentation of the solution we had arrived at.

  - ➢ **Learning:** Through this assignment, I have been able to gain a deeper understanding of database normalization. I was able to better understand each of the three types of normal forms namely 1st, 2nd, and 3rd normal forms, and the rules that must be kept in mind when converting these tables from one normal form to the other. I also gained a better understanding of the concepts of primary key and composite keys and how they are identified in tables. Lastly, the assignment also helped me learn about the different types of relationships (one-to-one, one-to-many, many-to-many) between the fields in the tables.

- **Pragati Naikare**
  - ➢ **Contribution:**
    1. Presented a solution with 3 tables namely Transaction, Customer, and Items.
    2. Converted Order Line table from 1NF to 2NF
    3. Discussed with my team members whether to drop the 'promotion' column or create a new table to preserve the information
    4. Discussed a few real-time scenarios in which our solution could potentially lead to data redundancy and made necessary modifications to address this issue.
    5. Identified the primary and composite keys for the tables.
    6. Question 3, provided a clear explanation of why we decided to retain the Promotion table instead of dropping it.
  - ➢ **Learning:** Gained a deeper understanding of the importance of database normalization and how to identify and resolve violations of the first, second, and third normal forms. Learned how to create a more efficient and organized data model by breaking down a complex table into smaller, related tables, thereby reducing redundancy and improving data integrity. Further, I gained a better understanding of primary keys and how they establish relationships between

tables, including the use of composite primary keys when needed. I learned how to structure a relational database from denormalized data, create normalized tables, and establish relationships between them. This project also reinforced the significance of collaboration and effective communication within a group to produce a well-organized and accurate database structure.

- **Romit**
  - ➢ **Contribution:**
    1. Researched for examples that can be used to understand how each Normal Formal can be handled.
    2. Converted Transactions table from 1NF to 2NF.
    3. Scheduled Zoom calls for discussions.
    4. Created ERD diagram to show relationships between fields of different tables
  - ➢ **Learning:** Understood different violations of the second normal form that may be caused by data in the first normal form. Learned how to handle multivalued and repeating columns. I was able to figure out the difference between different approaches used for Normalization and decide which approach resulted in a more appropriate result. Learned how to represent different types of relationships between related tables in an ERD diagram. Applied retail experience to figure out possible relationships between various entities at each normal form.
- **Tanmay**
  - ➢ **Contribution:**
    1. Converted tables from 2NF to 3NF by eliminating the transitive dependencies
    2. Highlighted the redundancies in data distribution and suggested solutions to simplify distribution.
    3. Collaborated on defining the relations between fields and entities.
    4. Prepare an entity relationship diagram for a visual representation of relations between entities.
    5. Contributed towards the documentation of all the results and learnings in the format described by the instructor.
  - ➢ **Learning:**
    Completing this task has offered substantial educational experience within the field of database structuring and normalization. It has enriched my comprehension of the three tiers of normalization (1NF, 2NF, and 3NF) and the precise rationale behind their application. A critical insight gained pertains to the necessity of minimizing data repetition to bolster data reliability and economize on storage resources, which constitutes a pivotal aspect of efficient data administration. Furthermore, I acquired an understanding of composite keys and their role in singularly identifying entries within tables, reflecting authentic relationships within the data. Additionally, this assignment enabled me to practice the art of

data modeling by dissecting a complex table into interconnected tables, a skill indispensable for orchestrating data logically and methodically. Furthermore, I learned how to embed business logic, such as promotional regulations determined by order type, customer status, and acquired items, into the design of the database. I also gained insight into identifying and defining relationships between entities in a way that can later help with developing efficient queries over the database.