Brandeis University | INTERNATIONAL BUSINESS SCHOOL

BUS211f(1) Analyzing Big Data I **Fall 2018**
Brandeis International Business School

Analysis 3 – ERD Exercise for Customer Purchase Data

Student Name(s)

For verification, please list Team Members below:

- **Student #1 –** Ran Dou
- **Student #2 –** Qimo Li

Notes:

- *Point values of each part are shown below; 10 points will be allocated for the quality of your submission (organization, clarity, grammar, on-time submission etc.).*
- *All team members will receive the same grade. It is up to the team to ensure that all members deserve the same grade.*
- *Type or paste your responses into the boxes below.*

Deliverables: Your upload will consist of **ONE FILE**:

□ The Template file with responses to questions, saved as a **pdf**. This should include the pictures of the ERD prepared on LucidCharts or a similar tool. Points will be deducted for non-pdf submissions.

Overview: In this exercise, your team will develop an Entity Relationship Diagram from a dataset of customer purchases (consumerDataFrame.csv on LATTE), as well as a data model defining fields in part of the database.

1. The data contains consumer purchases for a product category (think canned soup or yogurt) from various stores of a supermarket. This data is in a "flat-file" form, which contains data redundancies. Name one redundancy in this dataset and mention how you would store the data more effectively in a relational database. Be specific in listing relevant attributes of each entity and indicate relationships to other entities. ***(10 points)***

   We found that one redundancy in this dataset is that the same weekNum-StoreID-CustomerID combination was recorded for multiple times. This redundancy exists because the numbers in these three attributes repeat several times in each purchase of a product that makes the data shown inefficiently.

   We would move these three attributes above to a new entity called "Shopping Trip" and create a new attribute named "ShoppingTripID" in this entity. We would also create an attribute named "TransactionID" that shows every transaction related to the combination of "weekNum", "StoreID" and "CustomerID," and we would keep this new attribute in our original modified entity called "Transaction".

   The relationships among each entity: each customer has multiple transactions and shopping trips, and each shopping trip contains various transactions. Moreover, each transaction is corresponding to each product. We created these relations to help us store the data more logically and effectively. The modified entities (Transaction and Shopping Trip) are shown below. Furthermore, each shopping trip is corresponding to multiple transactions.

Transaction table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
TransactionID	Identification of the purchase for each item	Character	5	None	None	PK
ShoppingTripID	Identification of the shopping trip the transaction belongs to	Character	5	None	None	FK
CustomerID	Identification of customer who made this transaction	Character	5	None	None	FK
ProductID	Product number	Character	4	None	None	FK
dollars	Total price of the item in the transaction	Numeric	3	None	None	
units	Quantity of item	Numeric	2	None	None	

Shopping Trip table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
ShoppingTripID	Identification of the shopping trip	Character	5	None	None	PK
CustomerID	Identification of customer who went to this shopping trip	Character	5	None	None	FK
StoreID	Identification of where this shopping happened	Character	1	None	None	FK
weekNum	The number of week this shopping trip happened	Numeric	4	None	None	

2. The retail chain also has information on products and customers that can be related to the above purchases data. Here are two blank data dictionary tables for the **Product** and **Customer** data tables. Complete them with *at least* 3 attributes each, listing all fields (attributes) and place the abbreviations "PK" or "FK" in the "Key?" column to identify Primary and Foreign keys. Add rows to the Tables as necessary. ***(10 points)***

PRODUCT table

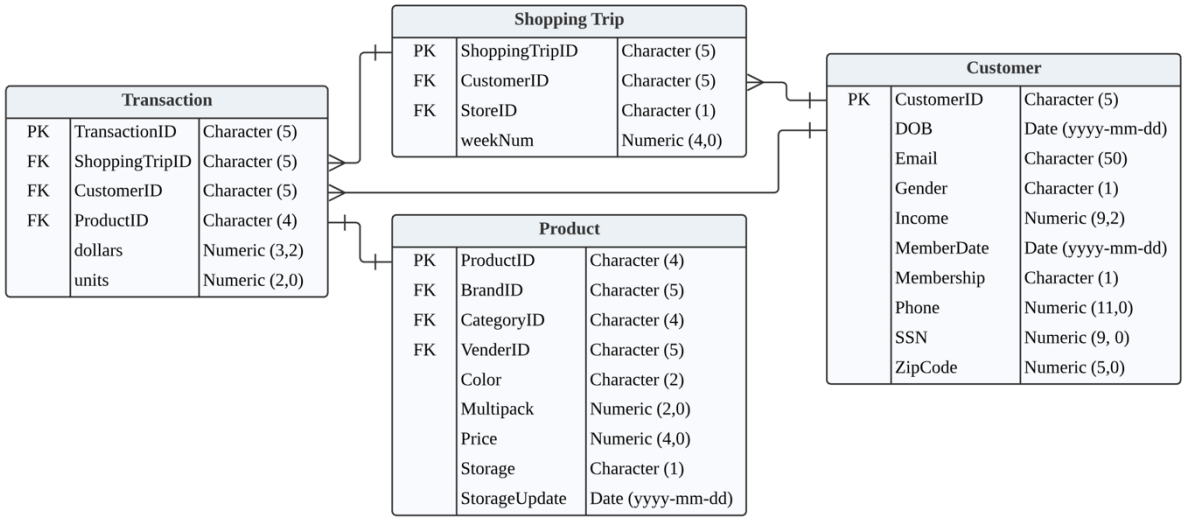
VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
ProductID	Product Number	Character	4	None	None	PK
BrandID	Identification of the brand associated with the product	Character	5	None	None	FK
CategoryID	Identification of the category associated with the product	Character	4	None	None	FK
Color	Color description of the product	Character	2	None	0	

Multipack	The number of products based on the product unit	Numeric	2	None	None	
Price	The overall price of the product	Numeric	4	None	None	
Storage	Status of a product is stock available or out of stock	Character	1	1=Out of Stock 2=StockAvailable	None	
StorageUpdate	The most recent date a product update its storage information	Date (yyyy-mm-dd)	10	None	None	
VenderID	Identification of the vender that supplies the product	Character	5	None	None	FK

CUSTOMER table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key ?
CustomerID	Customer number	Character	5	None	None	PK
DOB	Date of birth of customer	Date (yyyy-mm-dd)	10	None	0	
Email	Email address	Character	50	None	0	
Gender	Sexual classification	Character	1	1=Female 2=Male	0	
Income	Monthly purchasing power of customer	Numeric	9	None	0	
MemberDate	Date of joining the membership	Date (yyyy-mm-dd)	10	None	0	
Membership	Whether the customer has the membership	Character	1	1=No 2=Yes	None	
Phone	Phone number	Numeric	11	None	0	
SSN	Social Security Number	Numeric	9	None	0	
ZipCode	Code in the postal address of a customer	Numeric	5	None	0	

3. Prepare a simple ERD (use LucidChart or equivalent tool) that shows how to convert the customer purchases flat-file into a relational database. Show the links between the purchases data and the product and customer tables. Make sure the ERD is complete and ***includes cardinalities***. ***(25 points)***



The retail chain frequently offers its customers price promotions (discounts, coupons, etc). These discounts are availed at the checkout counter and entered into the system. Therefore, the retail manager knows the promotions offered for every product-store-week combination, and whether the customer decided to accept the offer.

4. We want to augment and expand the current dataset so that it can track both sets of information. Which variable(s) would you add to the flat-file Excel spreadsheet to record this information? Be specific in listing these variables, and the type of data they contain. ***(5 points)***

   We would like to add two variables, "PromotionID" and "PromotionAccept", to the flat-file Excel to record the promotions offered for every product-store-week combination and whether the customer decided to accept the offer. The two new variables are both displayed under the "Transaction" entity, and "PromotionID" can be a foreign key that uniquely identifies the observations in the entity that describes the promotions offered by each product-store-week combination.

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key ?
PromotionID	Identification of promotions offered for every product-store-week combination	Character	4	None	None	FK
PromotionAccept	Whether the customer decided to accept the offer	Character	1	1=Accepted 2=Not Accepted	None	

5. What additional entity/ies and attributes would be required in the relational database to support the special price promotion application? Identify primary and foreign keys. Be specific in listing relevant attributes of each entity, and indicate relationships to other entities. ***(10 points)***

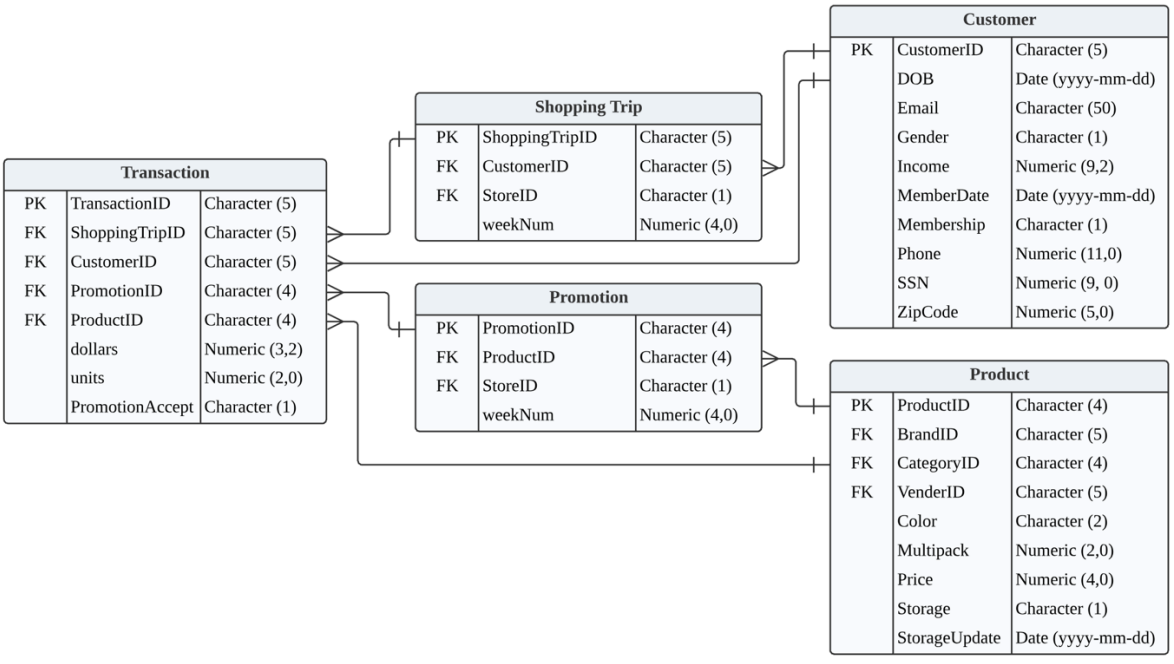
   First, an additional entity called "Promotion" would be required, which records the promotion provided for each product-store-week combination. This new entity includes four attributes, "PromotionID", "ProductID", "StoreID" and "weekNum". The data dictionary for this entity is shown below.

   In the "Promotion" entity, the primary key is "PromotionID" because it can uniquely identify all table records. Moreover, "ProductID" and "StoreID" are both foreign keys. "ProductID" is the foreign key connect to "Product" entity, each product can have distinct promotions in different weeks and stores. "StoreID" can be the foreign key connect to the entity that records the information of stores, similarly, each store can have distinct promotions for different products and in different weeks.

Promotion table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key ?
PromotionID	Identification of promotions offered for every product-store-week combination	Character	4	None	None	PK
ProductID	Product Number	Character	4	None	None	FK
StoreID	Identification of the store the promotion offered	Character	1	None	None	FK
weekNum	The number of the store the promotion offered	Numeric	4	None	None	

6. Modify the ERD in part (2) above to convey the promotion application information. Make sure the ERD is complete and ***includes cardinalities***. ***(20 points)***



7. Finally prepare a very brief (two paragraphs maximum) message explaining how your recommendations address the business needs of the retail chain. ***(10 points)***

Generally, we added two new attributes "TransactionID", "ShoppingTripID", "PromotionID" and "PromotionAccept" to the original flat-file Excel spreadsheet and converted the spreadsheet to an entity in the ERD called "Transaction". Moreover, we created four more entities "Shopping Trip", "Promotion", "Customer" and "Product" to further explain the foreign keys in the "Transaction" entity.

Above all, these recommendations are essential for businesses to create more efficient customer/supplier collaborations of the retail chain. First and foremost, implementing promotions on products can stimulate customers to purchase and thus help the retail businesses generate profits and keep the potential customers. Next, it's important to organize the product storage tracking system (checking the status of a product is stock available or out of stock periodically) because this helps the retail businesses operate the distribution system more smoothly and efficiently for more than one store. Further and more importantly, successful retail businesses need to sell the products that meet the needs of most customers. By connecting each shopping trip, promotion, customer, and product to each transaction, businesses can predict that the most popular and the most out-of-date products among the customers. Selecting the potential customers and products is vital to business operations.