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 – Xuyuan Zhang**
- **Student #2 – Yunqiu 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)**

If a customer purchases multiple products in one store, the store ID and customer ID will appear multiple times for different products. In other words, when a customer buys different products in a store, all the information regarding store ID and customer ID is repeated, which causes data redundancy. There are six columns in the "flat-file" format spreadsheet: units, dollars, weekNum, StoreID, CustomerID, ProductID.

To convert the flat-file form spreadsheet to relational databases, we create two entities named Shopping Trip and Transaction. StoreID, weekNum, CustomerID are listed as attributes under Shopping Trip because those information indicates when, where and who make the shopping trip. ProductID, units and dollars are listed as attributes under Transaction, which indicates further information regarding a specific transaction. There is still an attribute missing to

connect the two entities and uniquely identify each record in the Shopping Trip entity, so we add attribute VisitNum, which records a specific customer visit.

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	Chr	4	None	None	PK
Brand	Product Brand	Chr	15	None	None	
Price	Unit Price	Num	3.2	None	-9	
ProDate	Date of Production	Date	11	None	None	
ExpDate	Date of Expiration	Date	11	None	None	
CategoryID	Product Category	Chr	6	None	None	FK
SupplierID	Product Supplier	Chr	6	None	None	FK

Note:

- The width for price is 3.2. 3 stands for number of digits before decimal place, and 2 stands for number of digits after decimal place.
- The widths for ProDate and ExpDate are 11. They are in the format of mm/dd/yyyy.

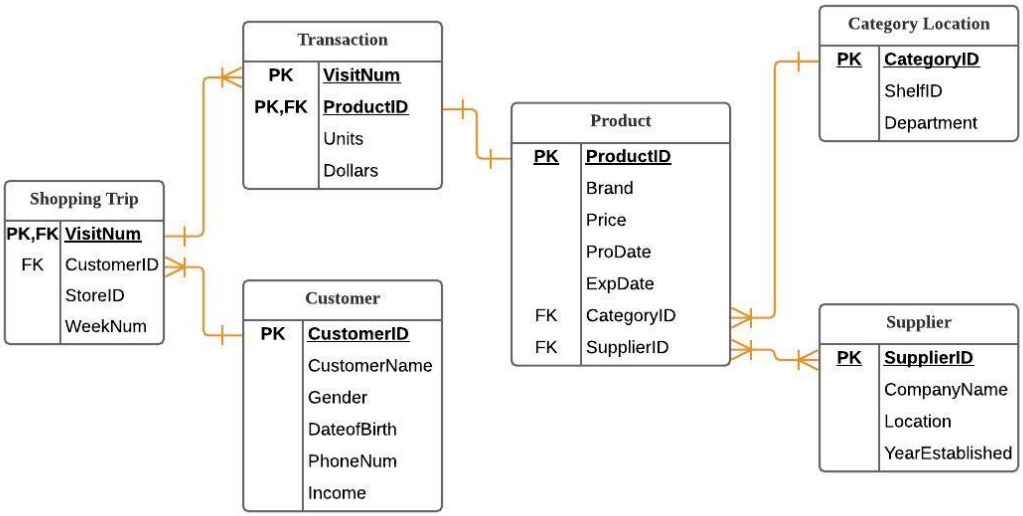
CUSTOMERS table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
CustomerID	Customer number	Chr	10	None	None	PK
Name	Customer name	Chr	15	None	None	
Gender	Customer gender	Num	1.0	1 = Female 2 = Male	-9	
DateofBirth	Customer date of birth	Date	11	None	None	
PhoneNumber	Customer phone number	Num	10.0	None	-9	
Income	Customer Income	Num	10.0	None	-9	

Note:

- The width for gender is 1.0. 1 stands for number of digits before decimal place, and 0 stands for number of digits after decimal place. The same rational applies to PhoneNumber and Income.
- DateofBirth is in the format of mm/dd/yyyy.

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)**



* Note: In the Transaction entity, VisitNum and ProductID both serve as primary keys, VisitNum uniquely identifies a specific customer trip but the customer may buy multiple products in a shopping trip, so only VisitNum and ProductID together can uniquely identify a specific transaction. In other words, the VisitNum is just like a receipt for a shopping trip, while VisitNum and ProductID together locates a specific line on the receipt.

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)**

The two sets of information to track are which promotion has been offered and whether a customer accepts the promotion offer. Therefore, in the flat-file excel spreadsheet, we'd like to add two variables: PromotionCode and OfferAccept. The PromotionCode variable records the specific promotion code for a given product-store-week combination. Each promotion code has 6 digits made

of the English alphabet or numbers. The OfferAccept records if a customer accepts or rejects to use the promotion code. 1 stands for acceptance and 0 stands for rejection. Also, to record further information regarding a promotion, we'd like to add three more variables: StartDate, ExpDate and DiscountRate, which records when the promotion code starts to be valid, when it expires and what percentage of discount it applies to the original price.

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)**

In the Product entity, a new attribute called PromotionCode is required to be added. This attribute can uniquely identify every product-store-week combination because it is linked with a certain product and you can get information about stores and week numbers by tracing back to the Shopping Trip entity.

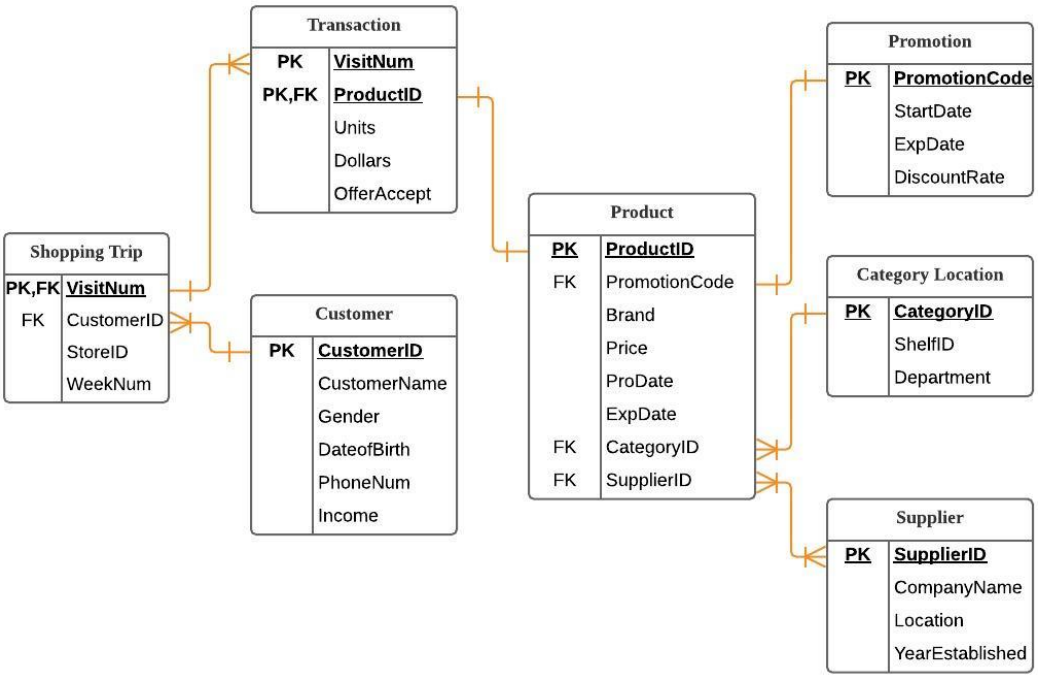
In addition, a new entity named Promotion is also added to the diagram. Four attributes are included in this entity:

- **PromotionCode: the code that can uniquely identify every product-store-week combination**
- **StartDate: the start date of the promotion**
- **ExpDate: the expiration date of the promotion**
- **DiscountRate: the discount rate of the promotion**

PromotionCode in Promotion is the Primary Key (PK) since it can uniquely identify each item in that table. PromotionCode in Product is the Foreign Key since it is used to link with another entity.

The relationship between Product and Promotion is one-to-one because each product can have only one promotion and each promotion can only be applied to one product.

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 paragraph maximum) message explaining how your recommendations address the business needs of the retail chain. **(10 points)**

By converting different variables in the fiat-file excel spreadsheet into different entities with multiple attibutes, the retail chain can increase their data efficiency. It also shows the hierarchy of variables by referring to attributes under a given entity and the relationships between entities. Also, it would be easier for the retail chain to add more information to different entities without creating data redundancy.

In addition, this ERM structure can help the retail chain manage the company better. The company can evaluate the effect of promotion on sales for each product by tracking its PromotionCode and OfferAccept Record, and then come up with better sales strategies through further analysis. For example, for promotion codes that were used with higher frequency, the retail chain can implement it more frequently to boost sales. Furthermore, it is easy to know where this product is located by tracking its CategoryID and who the supplier is by tracking its SupplierID. By tracking the record under Category Location and Suppliers entities, the retail chain can better understand products under which department/category generate most sales and products from which supplier are more popular.

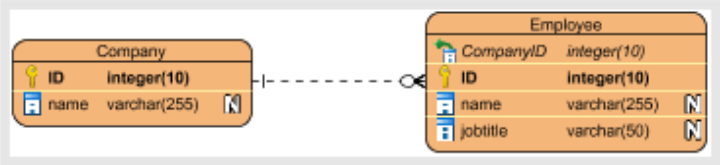
A Note on Data Dictionaries:

A data dictionary lists the fields (variables) in a data table, typically including information like that shown in this figure:

Variable Name	Label	Type (Width)	Value Codes	Missing Code
ID	Identification Number	String (4)	none	none
Age	Age on Jan 1, 2010	Numeric (3.0)	none	-9
Gender		Numeric (1.0)	1=Female 2=Male	9
TDATE	Test Date	Date (11) (mm/dd/yyyy)	none	None
SCORE	Test Score	Numeric (6.2)	None	-9

Note that each attribute or element in the data table is given a short Variable name, followed by a more descriptive label. For our assignment, indicate the type of data (character, numeric, date) and make a reasonable estimate of the width of the data column. For categorical variables, think about possible Value Codes, and about how a missing value would be represented.

As shown below, the dictionary must be consistent with the ERD, and should also indicate which variables are Primary or Foreign keys. Ignore the "Nullable" and "Unique" columns.



Entity Relationship Diagram1

Data Dictionary

Entity Name / Column Name	Column Description	Data Type	Length	Primary Key	Nullable	Unique
Company	A company is a business unit that provides good or service.					
ID	For the unique identification of company records.	integer	10	true	false	false
name	Name of the company.	varchar	255	false	true	false
Employee	An employee is someone who work in a company.					
CompanyID		integer	10	false	false	false
ID	For the unique identification of employee records.	integer	10	true	false	false
jobtitle	The position of the employee in a company.	varchar	50	false	true	false
name	Name of the employee.	varchar	255	false	true	false