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 – Mengyang Yao**
- **Student #2 -**

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)***

   One redundancy is that when a customer purchases another product in the same transition, all the same information is repeatedly stored again, which is wasting space and redundant.
   I would prefer to store in an ERD, which includes two entities (four in the end), CUSTOMERS and PRODUCT. The relationship will be: one customer owns multiple products.
   The PRODUCT table contains PRODUCT _ID, ITEM_QUANTITY, TOT_UNIT_COST, VISIT_NBR and STORE_NBR.
   CUSTOMERS table includes CUSTOMER_ID, ZIP_CODE, SIC, BUS_CR_TYP_STAT_CD, RENEWAL_DATE.

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)***

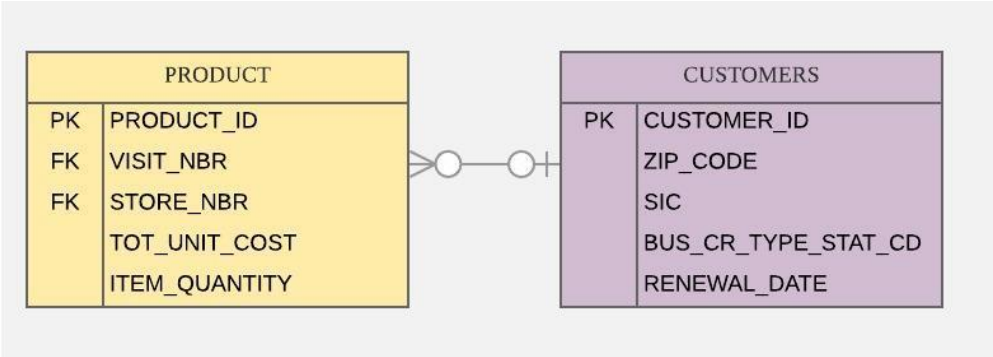
PRODUCT table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
ProductID	Product Number	Chr	10	NONE	NONE	PK
ITEM_QUANTITY	Number of same item	num	10	NONE	0	
TOT_UNIT_COST	The cost of the item	num	50	NONE	0	
VISIT_NBR	Created every time a card scanned	num	255	NONE	NONE	FK
STORE_NBR	Store identification number	num	50	NONE	NONE	FK

CUSTOMERS table

VARName	Label	DataType (chr, num, date)	Width	Value Codes	Missing Code	Key?
CustomerID	Customer number	Chr	10	NONE	NONE	PK
ZIP_CODE	The zip-code of the member	num	50	NONE	NONE	
SIC	Standard industry	num	50	NONE	NONE	
BUS_CR_TYP_STAT_CD	Business Credit Type Status Code	num	10	NONE	NONE	
RENEWAL_DATE	Date a membership should be renewed	date	25	NONE	NONE	

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.

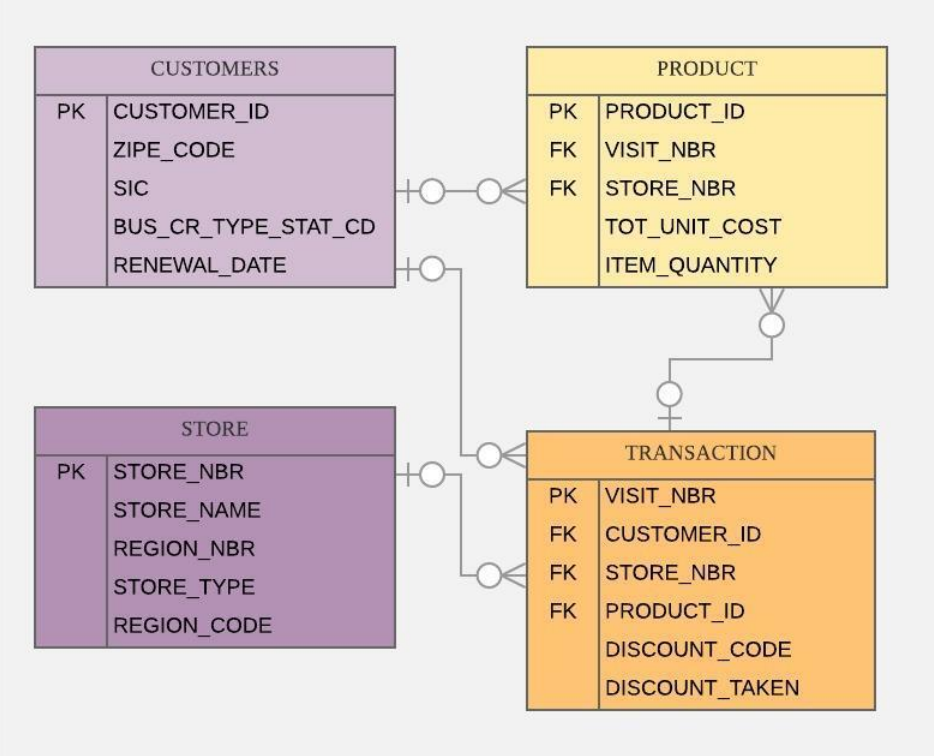
4. We want to augment and expand the current dataset so that it can track both sets of information. Which variable(s) would you add to the flat-file Excel spreadsheet to record this information? Be specific in listing these variables, and the type of data they contain. ***(5 points)***

   I want to add two variables: DISCOUNT_CODE and DISCOUNT_TAKEN. DISCOUNT_CODE is the unique code number to each product-store-week combination, it is stored as character(nominal); the DISCOUNT_TAKEN is whether a customer chooses to take the discount chance in one transaction for one product, it is stored as integer(nominal), 1 represents taken, 0 otherwise.

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)***

   Two entities are required.
   One named TRANSACTION, the primary key is VISIT_NBR, the foreign keys are PRODUCT_ID, CUSTOMER_ID and STORE_ID, other attributes includes DISCOUNT_CODE and DISCOUNT_TAKEN.
   The other one is STORE, the primary key is STORE_NBR, other attributes are STORE_NAME, REGION_NBR, STORE_TYPE, and REGION_CODE.
   One transaction owns multiple products, while one customer owns multiple transactions. One store owns multiple transactions.

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)***

   The TRANSACTION entity contains information for every visit per customer per store. Each row has a discount code for discount combination when a customer checks out, also a DISCOUNT_TAKEN notes whether or not they take the discount chance. In this way, the retail manager knows the promotions offered for every product-store-week combination, and whether the customer decided to accept the offer.

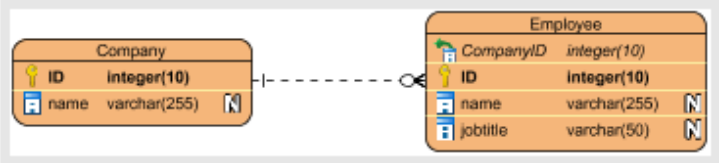
A Note on Data Dictionaries:

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	Entity Description					
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