Relational Schema Design Exercise

Felicia Liu (liu318@illinois.edu)

CS598 Foundations of Data Curation

I. Background

An auto dealer company consists of Inventory Department, Sales Department and Customer Relations Department, and at present, each department manages their data separately. The three departments are working on integrating their data into a shared database which contains all the information from the three different datasets. The solution is to design and develop a relational database which can be adapted effectively and efficiently for all departments in this company.

II. Evidence of in-depth examination of data

The company has the below three files and the format of the three files are different.

File Name	Department	File Format
File A	Inventory	Text
File B	Sales	CSV
File C	Customer relations	Word

A. File A is a text file from the Inventory department. From file A, the data of one particular inventory is organized in one single row. Overall, the data is understandable. Moreover, the different values are separated from each other by a tab space. However, this file still contains the following issues.

   a. There are no titles/headings for this file. The readers are unable to identify what the column represents. For instance, in row 1, the number “$35,240.00” can represent trade-in value, can represent retail price or represent cost price.

   b. For certain columns, the mandatory information is missing.

c. For certain columns, like the “Price” column, it is better to populate float type rather than string type for further calculation.

d. The current information is not sufficient to differentiate the inventory. It is better to add more information including colors etc.,

e. MSRP (manufacturer's suggested retail price) should be an integer, but is a String. That makes this field useless for calculations.

f. Some values for Engine are separated by slashes, making the data unreadable.

g. Number of doors is a String, which should be an integer.

h. There can be subcategories in the color of the vehicle, which is mentioned in brackets, making the data unclear.

B. File B is a csv file from the Sales department. Overall, this file benefits in finding the titles of the data present in the inventory and customer files. Furthermore, this file contains the details of the Customer who purchased the car; it also includes some information about the purchased vehicle and the sales information of the car. Nevertheless, this file still contains the following data quality issues.

  a. Data inconsistency issue identified. For instance, in customer details, there are some missing values for city, state, and country details.

  b. The definition of the columns are unclear. For instance, the column “Year” does not match the actual sales year, and it does not match the manufacturing year in the “Inventory” file.

  c. The “TradeInValue” and “PurchasedPrice” columns contain the dollar sign, which may make the fields difficult to calculate.

  d. For certain columns that should be mandatory, the values are missing. For example, the column “PurchasedPrice” must be populated with.

  e. Model attribute here is one single entity, which makes subclassification of the vehicle between submodels difficult.

  f. Repeat Customer Field looks redundant, as it seems evident that if a customer is repeating he will get listed in the discount field.

  g. Subcategories of Color are mentioned in one single attribute.

h. Few data values are missing from the purchase price and the MSRP attribute, which should be mandatory.

i. For the Discount attribute, It is given that the discount is offered to a few customers, but an essential field of the Discount amount is missing, which should be the part of the sales file.

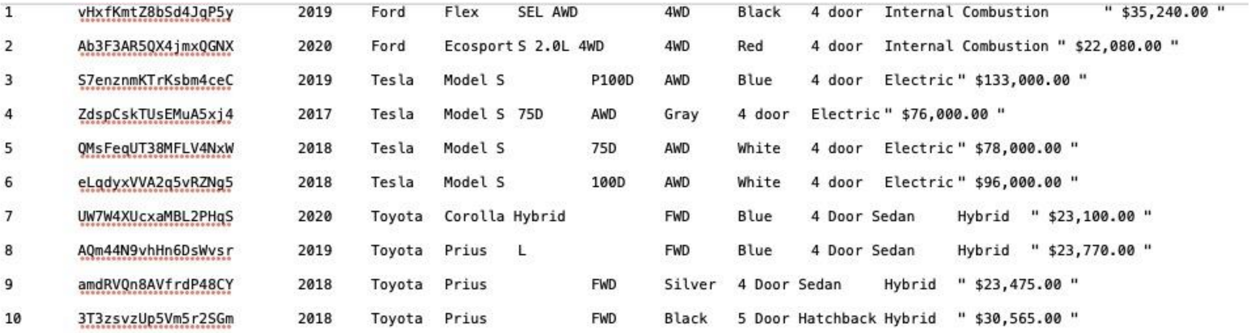
C. File C is a word document from the Customer Relations department. This file contains the personal details of the customers of the auto dealer, including their complete addresses, profession and their inquiries about purchasing the vehicle. This file mainly contains the following issues.

   a. There are no titles/headings for this file, which makes the file hard to comprehend.

   b. A single tuple is ended by 2 consecutive character returns, which is an empty line in the word document.

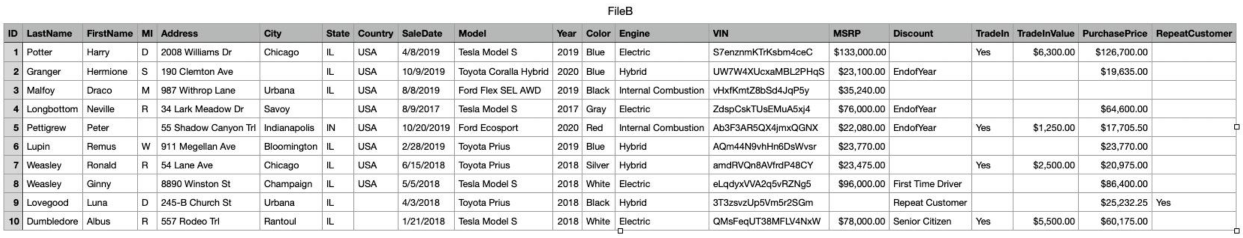
   c. Attributes are separated by tabs in the document.

File A



```
1     vHxfKmtZ8bSd4JqP5y    2019   Ford     Flex     SEL AWD        4WD    Black    4 door   Internal Combustion      " $35,240.00 "
2     Ab3F3AR5QX4jmxQGNX    2020   Ford     Ecosport S 2.0L 4WD     4WD    Red      4 door   Internal Combustion " $22,080.00 "
3     S7enznmKTrKsbm4ceC    2019   Tesla    Model S          P100D  AWD    Blue     4 door   Electric " $133,000.00 "
4     ZdspCskTUsEMuA5xj4    2017   Tesla    Model S  75D     AWD    Gray   4 door   Electric " $76,000.00 "
5     QMsFeqUT38MFLV4NxW    2018   Tesla    Model S          75D    AWD    White    4 door   Electric " $78,000.00 "
6     eLqdyxVVA2q5vRZNg5    2018   Tesla    Model S          100D   AWD    White    4 door   Electric " $96,000.00 "
7     UW7W4XUcxaMBL2PHqS    2020   Toyota   Corolla Hybrid          FWD    Blue     4 Door Sedan     Hybrid  " $23,100.00 "
8     AQm44N9vhHn6DsWvsr    2019   Toyota   Prius    L              FWD    Blue     4 Door Sedan     Hybrid  " $23,770.00 "
9     amdRVQn8AVfrdP48CY    2018   Toyota   Prius            FWD    Silver   4 Door Sedan     Hybrid  " $23,475.00 "
10    3T3zsvzUp5Vm5r2SGm    2018   Toyota   Prius            FWD    Black    5 Door Hatchback Hybrid  " $30,565.00 "
```

File B



FileB

ID	LastName	FirstName	MI	Address	City	State	Country	SaleDate	Model	Year	Color	Engine	VIN	MSRP	Discount	TradeIn	TradeInValue	PurchasePrice	RepeatCustomer
1	Potter	Harry	D	2008 Williams Dr	Chicago	IL	USA	4/8/2019	Tesla Model S	2019	Blue	Electric	S7enznmKTrKsbm4ceC	$133,000.00		Yes	$6,300.00	$126,700.00	
2	Granger	Hermione	S	190 Clemton Ave		IL	USA	10/9/2019	Toyota Coralla Hybrid	2020	Blue	Hybrid	UW7W4XUcxaMBL2PHqS	$23,100.00	EndofYear			$19,635.00	
3	Malfoy	Draco	M	987 Withrop Lane	Urbana	IL	USA	8/8/2019	Ford Flex SEL AWD	2019	Black	Internal Combustion	vHxfKmtZ8bSd4JqP5y	$35,240.00					
4	Longbottom	Neville	R	34 Lark Meadow Dr	Savoy		USA	8/9/2017	Tesla Model S	2017	Gray	Electric	ZdspCskTUsEMuA5xj4	$76,000.00	EndofYear			$64,600.00	
5	Pettigrew	Peter		55 Shadow Canyon Trl	Indianapolis	IN	USA	10/20/2019	Ford Ecosport	2020	Red	Internal Combustion	Ab3F3AR5QX4jmxQGNX	$22,080.00	EndofYear	Yes	$1,250.00	$17,705.50	
6	Lupin	Remus	W	911 Megellan Ave	Bloomington	IL	USA	2/28/2019	Toyota Prius	2019	Blue	Hybrid	AQm44N9vhHn6DsWvsr	$23,770.00				$23,770.00	
7	Weasley	Ronald	R	54 Lane Ave	Chicago	IL	USA	6/15/2018	Toyota Prius	2018	Silver	Hybrid	amdRVQn8AVfrdP48CY	$23,475.00		Yes	$2,500.00	$20,975.00	
8	Weasley	Ginny		8890 Winston St	Champaign	IL	USA	5/5/2018	Tesla Model S	2018	White	Electric	eLqdyxVVA2q5vRZNg5	$96,000.00	First Time Driver			$86,400.00	
9	Lovegood	Luna	D	245-B Church St	Urbana	IL		4/3/2018	Toyota Prius	2018	Black	Hybrid	3T3zsvzUp5Vm5r2SGm		Repeat Customer			$25,232.25	Yes
10	Dumbledore	Albus	R	557 Rodeo Trl	Rantoul	IL		1/21/2018	Tesla Model S	2018	White	Electric	QMsFeqUT38MFLV4NxW	$78,000.00	Senior Citizen	Yes	$5,500.00	$60,175.00	

File C

Dumbledore Albus R
557 Rodeo Trl
Rantoul IL USA 61866
Dean

Granger Hermione S
190 Clemton Ave
Champaign IL USA 61821
Archivist
Needs loan

Longbottom Neville R
34 Lark Meadow Dr
Savoy IL USA 61874
Doctor

Lovegood Luna D
245-B Church St
Urbana IL USA 61802
Student
Needs loan

Lupin Remus W
911 Megellan Ave
Bloomington IL USA 61701
Doctor - pediatrician

Malfoy Draco M
987 Withrop Lane
Urbana IL USA 61801
Unknown profession

Pettigrew Peter
55 Shadow Canyon Trl
Indianapolis IN USA 46077
Librarian
Needs financing

III. Evidence of understanding relations and schemas

After understanding the three original datasets, I sorted the columns and data types of the three tables like below.

Table 1: Inventory Table

Inventory		
	Column	**Type**
Primary Key	VIN	unique string
	Year	int
	Model	string
	Power	string
	Drive	string
	Color	string
	DoorNumber	int
	Engine	string selection
	MSRP	float

Table 2: Customer Relations Table

Customer Relations		
	Column	**Type**
Primary Key	CustomerID	unique int
	Lastname	string
	Firstname	string
	MI	string

	Address	string
	City	string selection
	State	string selection
	Country	string selection
	Zipcode	int
	Occupation	string

Table 3: Sales Table

Sales		
	Column	**Type**
Primary Key	SaleID	unique int
Foreign Key	CustomerID	int
	LastName	string
	FirstName	string
	MI	string
	SaleDate	datetime
Foreign Key	VIN	string
	Discount	string selection
	TradeIn	string selection
	TradeInValue	float
	PurchasePrice	float
	RepeatCustomer	string selection

For further detail, please refer to the Assignment1_Relational_Schema_Design_Exercise.xlxs file

IV. Discussion of curation objectives, decisions, and activities

The original datasets are managed in different technological tools, which leads to difficulties in managing and analyzing the data. It may have the below issues:

- Dependent on custom tools and application
- Dependent on memory and workplace practices
- Difficult to preserve fDifficult to documentor future use
- Difficult to repurpose and reuse
- Data Inconsistency between different files

Therefore, the preliminary goal is to leverage an adaptable technologic tool to manage and analyze the dataset, making it more organized and readable.

- Step 1: For file A, file B, and file C, I migrated the original data to csv format similar to below.

File A after migrating to csv format

ID	VIN	Year	Model	Power	Drive	Color	DoorsNumbers	Engine	MSRP
1	vHxfKmtZ8bSd4JqP5y	2019	Ford Flex SEL	150D	4WD	Black	4	Internal Combustion	35,240.00
2	Ab3F3AR5QX4jmxQGNX	2020	Ford Ecosport S	75D	4WD	Red	4	Internal Combustion	22,080.00
3	S7enznmKTrKsbm4ceC	2019	Tesla Model S	100D	AWD	Blue	4	Electric	133,000.00
4	ZdspCskTUsEMuA5xj4	2017	Tesla Model S	75D	AWD	Gray	4	Electric	76,000.00
5	QMsFeqUT38MFLV4NxW	2018	Tesla Model S	75D	AWD	White	4	Electric	78,000.00

6	eLqdyxVVA2q5vRZNg5	2018	Tesla Model S	100D	AWD	White	4	Electric	96,000.00
7	UW7W4XUcxaMBL2PHqS	2020	ToyotaCorolla Hybrid	150D	FWD	Blue	4	Hybrid	23,100.00
8	AQm44N9vhHn6DsWvsr	2019	ToyotaPrius L	150D	FWD	Blue	4	Hybrid	23,770.00
9	amdRVQn8AVfrdP48CY	2018	ToyotaPrius	75D	FWD	Silver	4	Hybrid	23,475.00
10	3T3zsvzUp5Vm5r2SGm	2018	Toyota Prius	75D	FWD	Black	5	Hybrid	30,565.00

File B after migrating to csv format

CustomerID	Lastname	Firstname	MI	Address	City	State	Country	Zipcode	Occupation
1	Dumbledore	Albus	R	557 Rodeo Trl	Rantoul	IL	USA	61866	Dean
2	Granger	Hermione	S	190 Clemton Ave	Champaign	IL	USA	61821	Archivist
3	Longbottom	Neville	R	34 Lark Meadow Dr	Savoy	IL	USA	61874	Doctor
4	Lovegood	Luna	D	245-B Church St	Urbana	IL	USA	61802	Student
5	Lupin	Remus	W	911 Megellan Ave	Bloomington	IL	USA	61701	Doctor - pediatrician
6	Malfoy	Draco	M	987 Withrop Lane	Urbana	IL	USA	61801	Unknown profession
7	Pettigrew	Peter	D	55 Shadow Canyon Trl	Indianapolis	IN	USA	46077	Librarian
8	Potter	Harry	D	2008 Williams Dr	Chicago	IL	USA	60007	Professor, UIC

9	Weasley	Ginny	W	8890 Winston St	Champaign	IL	USA	61820	Stay at home mother
10	Weasley	Ronald	R	54 Lane Ave	Chicago	IL	USA	60018	Research scientist

File C after migrating to csv format

SaleID	CustomerID	LastName	FirstName	MI	SaleDate	VIN	Discount	TradeIn	TradeInValue	PurchasePrice	RepeatCustomer
1	1	Potter	Harry	D	4/8/2019	S7enznmKTrKsbm4ceC	Not Applicable	Yes	6,300.00	126,700.00	No
2	2	Granger	Hermione	S	10/9/2019	UW7W4XUcxaMBL2PHqS	EndofYear	No	-	19,635.00	No
3	3	Malfoy	Draco	M	8/8/2019	vHxfKmtZ8bSd4JqP5y	Not Applicable	No	-	38,250.00	No
4	4	Longbottom	Neville	R	8/9/2017	ZdspCskTUsEMuA5xj4	EndofYear	No	-	64,600.00	No
5	5	Pettigrew	Peter	D	10/20/2019	Ab3F3AR5QX4jmxQGNX	EndofYear	Yes	1,250.00	17,705.50	No
6	6	Lupin	Remus	W	2/28/2019	AQm44N9vhHn6DsWvsr	Not Applicable	No	-	23,770.00	No
7	7	Weasley	Ronald	R	6/15/2018	amdRVQn8AVfrdP48CY	Not Applicable	Yes	2,500.00	20,975.00	No
8	8	Weasley	Ginny	W	5/5/2018	eLqdyxVVA2q5vRZNg5	First Time Driver	No	-	86,400.00	No
9	9	Lovegood	Luna	D	4/3/2018	3T3zsvzUp5Vm5r2SGm	Repeat Customer	No	-	25,232.25	Yes

10	10	Dumbledore	Albus	R	1/21/2018	QMsFeqUT38MFLV4NxW	Senior Citizen	Yes	5,500.00	60,175.00	No

- Step 2: For file A, I mainly changed the below parts to make the data more organized and understandable.
    - Arrange data values for the same attributes into one column. The copied data would not exactly fall in the same column, as there are some null values for few attributes.
    - Provide Attributes to the inventory file. This can be done by checking the data from the Sales file. Give the same headings as in the Sales file. For remaining, data fields give the appropriate titles.
    - Make a subcategory of color, to accommodate different color shades.
    - Make the MSRP field as an integer by removing the dollar symbol from there.

- Step 3: For file B, I mainly changed the below parts to make the data more organized and understandable.
    - City, State and Country attribute matches with the attributes in the word file. Though Data values for these attributes are incomplete as compared to Customer File. This causes data inconsistency. To maintain the completeness of data and avoid inconsistency, we will delete these attributes.
    - FirstName, LastName, and Address of the customers also seem irrelevant in this record. We would like to remove it from this record to maintain details at one place, in Customer records. But to make the communication possible between Sales and customer records, we would need customer information. So, replacing the customer details by customer id in the Sales file.
    - The model attribute is a single entity here, while in the inventory file we have subclassification of the model attributes. To maintain data

    consistency and completeness, we delete this attribute, as this attribute is more needed in inventory records.
  - Repeat customers could be deleted. But we don’t have enough proof if all repeat customers get the discount. To have the complete record, we will retain this field.
  - The Color attribute looks redundant, as it is available in inventory records.
  - Few MSRP data values are missing. MSRP should be an essential attribute for a Sales file. We used VLOOKUP in excel to complete this data in the Sales record from the Inventory record.

- Step 4: For file C, I mainly changed the below parts to make the data more organized and understandable.
  - Convert data into rows and columns.
  - Provide heading/attribute names to the data by comparing it to the Sales file and appropriate headings to the data specific to this file.

For further detail, please refer to the Assignment1_Relational_Schema_Design_Exercise.xlxs file

V. Overall quality analysis and completeness

Overall, the three files are converted to CSV/Excel as three different tables/datasets.

Table 1: Inventory Table

Inventory		
	Column	**Type**
Primary Key	VIN	unique string
	Year	int
	Model	string

	Power	string
	Drive	string
	Color	string
	DoorNumber	int
	Engine	string selection
	MSRP	float

Table 2: Customer Relations Table

Customer Relations		
	Column	**Type**
Primary Key	CustomerID	unique int
	Lastname	string
	Firstname	string
	MI	string
	Address	string
	City	string selection
	State	string selection
	Country	string selection
	Zipcode	int
	Occupation	string

Table 3: Sales Table

Sales		
	Column	**Type**

Primary Key	SaleID	unique int
Foreign Key	CustomerID	int
	LastName	string
	FirstName	string
	MI	string
	SaleDate	datetime
Foreign Key	VIN	string
	Discount	string selection
	TradeIn	string selection
	TradeInValue	float
	PurchasePrice	float
	RepeatCustomer	string selection

For table “Inventory”, the attribute “VIN”, which is constituted by unique characters to differentiate the stocks, serves as the primary key in this table. In addition, the table will contain columns including “Year”, “Model”, “Power”, “Color”, and all the missing values are populated.

For table “Customer Relations”, a column “CustomerID” is added to act as the primary key. For each customer, a unique ID will be assigned to that customer to better manage customer’s information.

For table “Sales”, a column “SaleID” is added to the table, and the data type is the unique integer to record each order. What is more, in the table “Sales”, column “CustomerID” and “VIN” are added to answer the questions “Which customer make this order”, “Which model is sold”, and these two columns serve as the foreign key to link the 3 various tables.