Assignment 1: 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

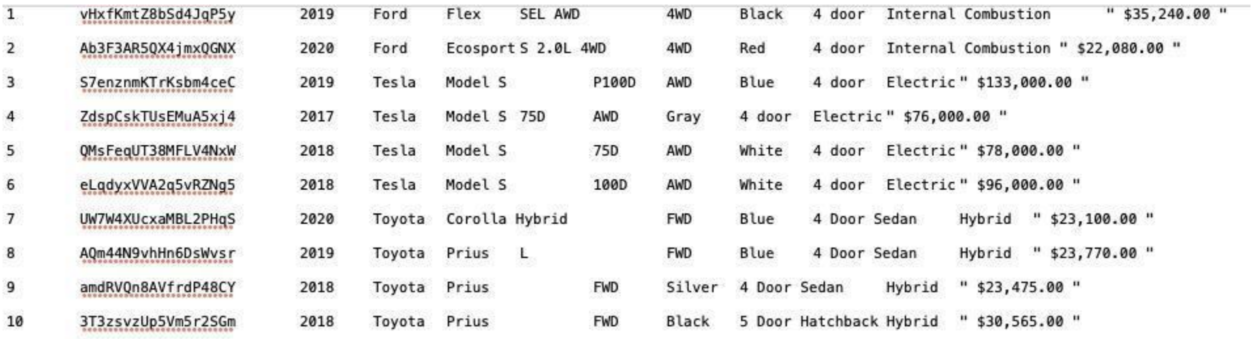
i. File A is a text format database from the Inventory department, which mostly contains information like VIN number, car model name, year, power, styles etc. Even though there is no column title for file A, each particular data point is stored in one row, which makes file A easier to understand and comprehend. In addition, the various data values are separated with some spaces in an organized way.

However, this file still contains the following issues.

- 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.

- For certain columns, the mandatory information is missing.
- For certain columns, like the "Price" column, it is better to populate float type rather than string type for further calculation. For instance, in the first row, the price column is shown as "$35,240.00", which makes it difficult to calculate.
- The first 4 columns are organized in a readable way, however, starting from the 5th column, since there are several missing values in certain columns, the whole file becomes messy and difficult to read.
- In some rows, for example the 2nd row, "4WD" shows up twice, the readers are unable to judge if it is duplicated or it should show up twice.



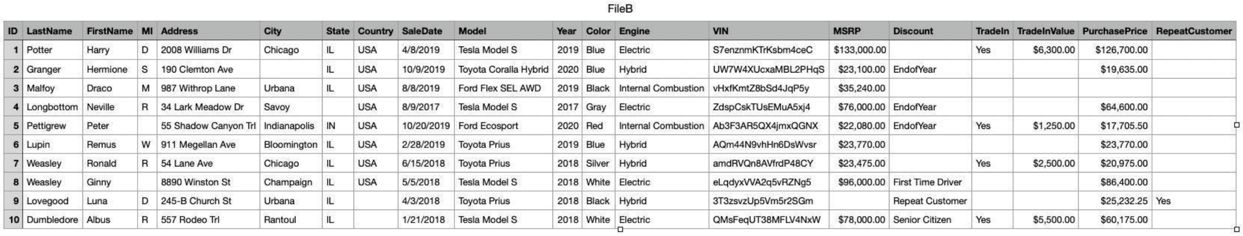
```
1       vHxfKmtZ8bSd4JqP5y      2019    Ford    Flex    SEL AWD         4WD     Black   4 door  Internal Combustion     " $35,240.00 "
2       Ab3F3AR5QX4jmxQGNX      2020    Ford    Ecosport S 2.0L 4WD     4WD     Red     4 door  Internal Combustion " $22,080.00 "
3       S7enznmKTrKsbm4ceC      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 "
```

Screenshot of File A

ii. File B is a CSV format table from the Sales department, which contains 15 variables and 10 entries to indicate the customer's information including name, address, sales date, model, year, color, engine, purchase price etc. Overall, this file is beneficial to identify the titles of the data presenting in all the 3 files. *Nevertheless, this file still contains the following data quality issues.*

- Data inconsistency issue identified. For instance, in customer details, there are some missing values for city, state, and country details.
- The definition of the columns are unclear. For instance, the column "Year" is not consistent with the column "SaleDate", and it does not match the manufacturing year in the "Inventory" file.

- The “TradeInValue” and “PurchasedPrice” columns contain the dollar sign, which may make the fields difficult to calculate.
- For certain columns that should be mandatory, the values are missing. For example, the column “PurchasedPrice” must be populated, however, in the 3rd row, the purchase price is missing.
- For columns like City, State, Country, there are some missing values.
- This file should be focused on the sales data, however, it contains several redundant pieces of information, making the file more difficult to understand.



FileB

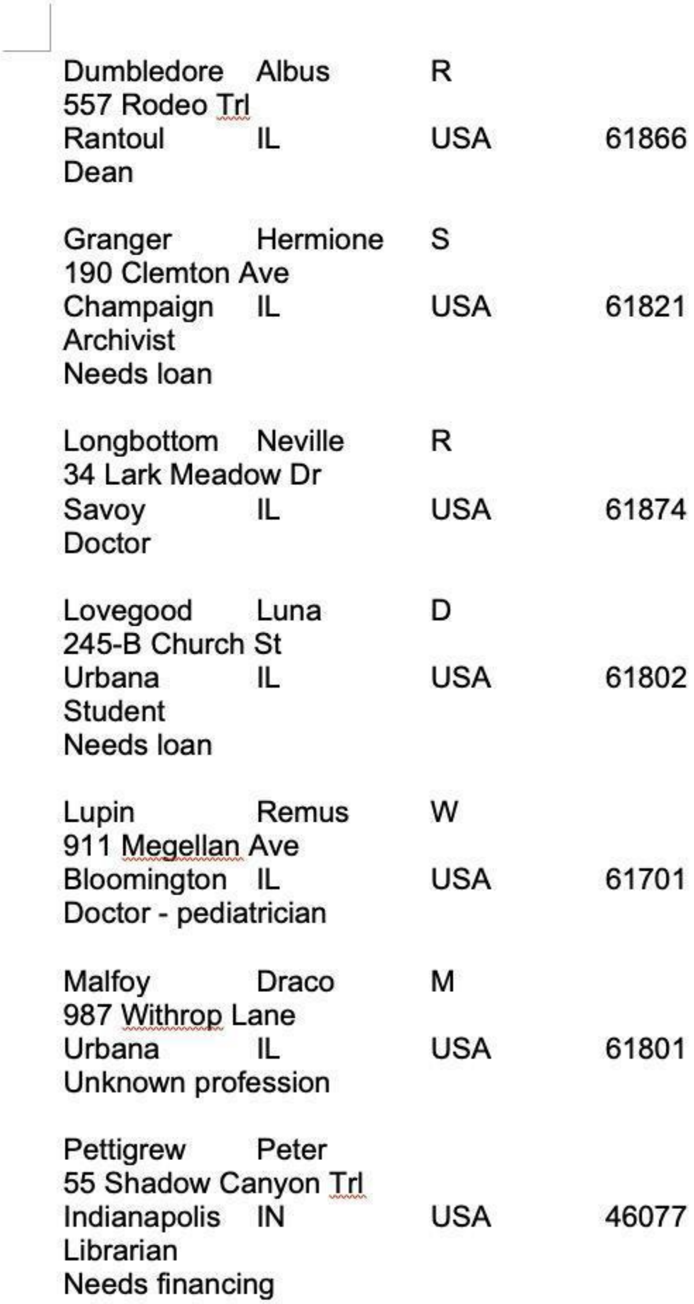
ID	LastName	FirstName	MI	Address	City	State	Country	SaleDate	Model	Year	Color	Engine	VIN	MSRP	Discount	TradeIn	TradeInValue	PurchasePrice	RepeatCustomer
1	Potter	Harry	D	2008 Williams Dr	Chicago	IL	USA	4/8/2019	Tesla Model S	2019	Blue	Electric	S7enznmKTrKsbm4ceC	$133,000.00		Yes	$6,300.00	$126,700.00	
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	

Screenshot of File B

iii. File C is a word format database from the Customer Relations department, which includes customer’s personal information like the address, city, country, occupation, inquiries regarding the services and warranties.

<u>This file mainly contains the following issues.</u>

- The word document is not a great choice to store multiple data and may result in difficulties in managing the data in the future.
- There are no titles/headings for this file, which makes the file hard to comprehend.
- Some information is not accurate or concise, for example, in the 1st row, the occupation is denoted as “Dean”, which should not be listed as the employment title.



```
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
```

Screenshot of File C

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.

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 1: Inventory 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 2: Customer Relations 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

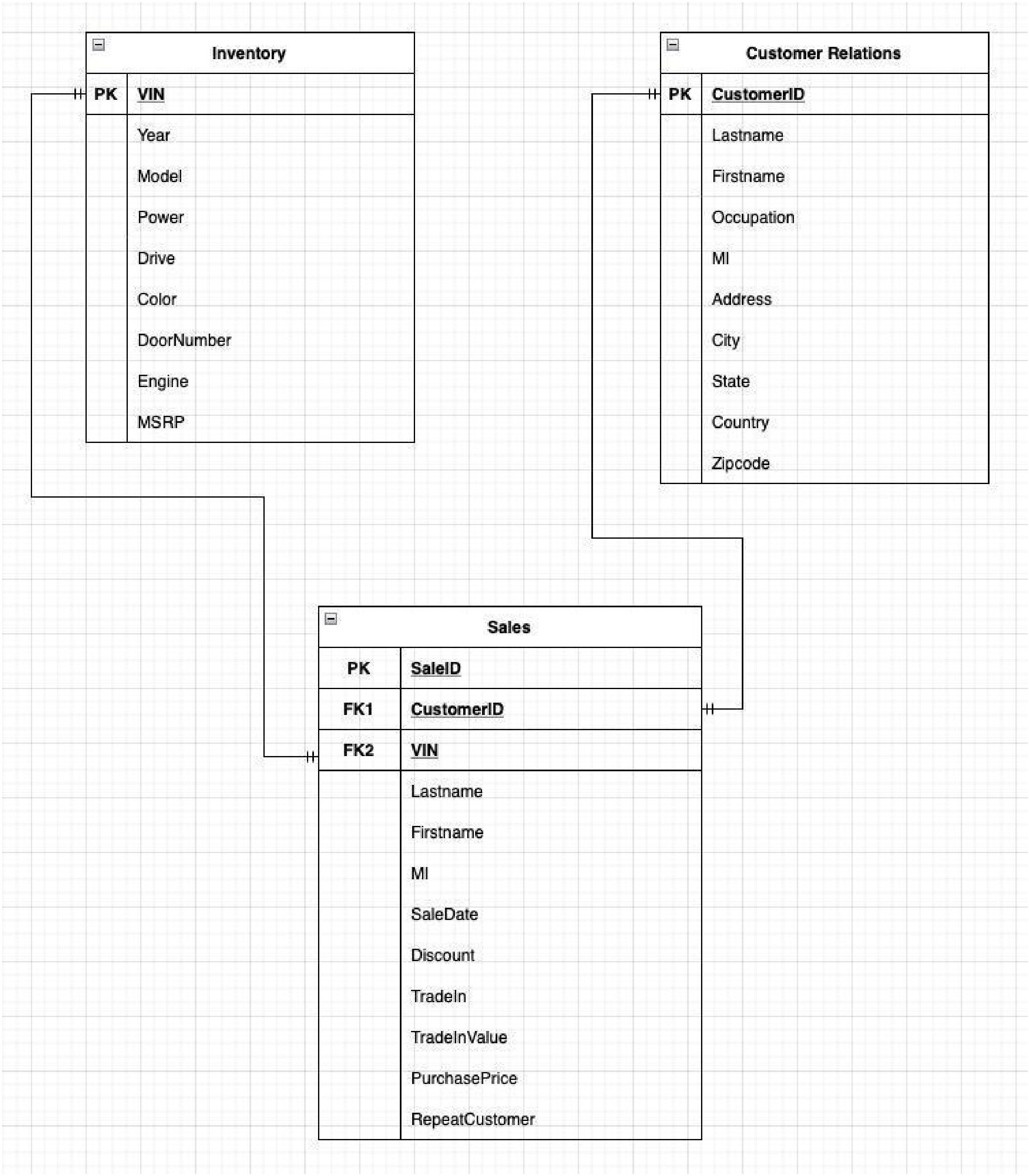
Table 3: Sales Table

The relationship for the three tables is presented as below.

- In the “Inventory” table, column “VIN” acts as the primary key since it is the unique number for each vehicle. In the “Sales” table, column “VIN” serves as the foreign key to link the “Inventory” table and “Sales” table.

- In the “Customer Relations” table, column “CustomerID” is the primary key because this table aims at managing the customer information, and each customer should have a unique ID. In the “Sales” table, column “CustomerID” works as the foreign key to link the “Customer Relations” table and “Sales” table.

- In the “Sales” table, the column “SaleID” is the primary key since it stands for each unique order.



Inventory	
PK	**VIN**
	Year
	Model
	Power
	Drive
	Color
	DoorNumber
	Engine
	MSRP

Customer Relations	
PK	**CustomerID**
	Lastname
	Firstname
	Occupation
	MI
	Address
	City
	State
	Country
	Zipcode

Sales	
PK	**SaleID**
FK1	**CustomerID**
FK2	**VIN**
	Lastname
	Firstname
	MI
	SaleDate
	Discount
	TradeIn
	TradeInValue
	PurchasePrice
	RepeatCustomer

ER Diagram

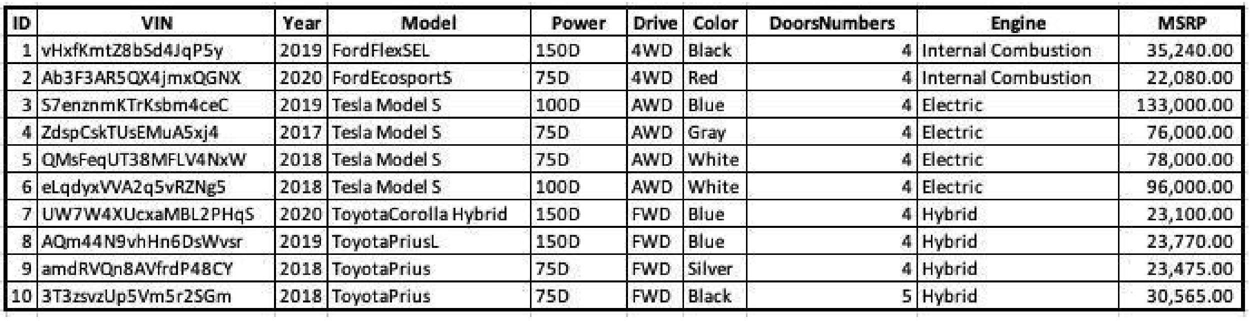
For further detail, please refer to the Assignment1_Relational_Schema_Design_Exercise.xlxs file

IV. Create an example of each table, populated with data from the files.

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:

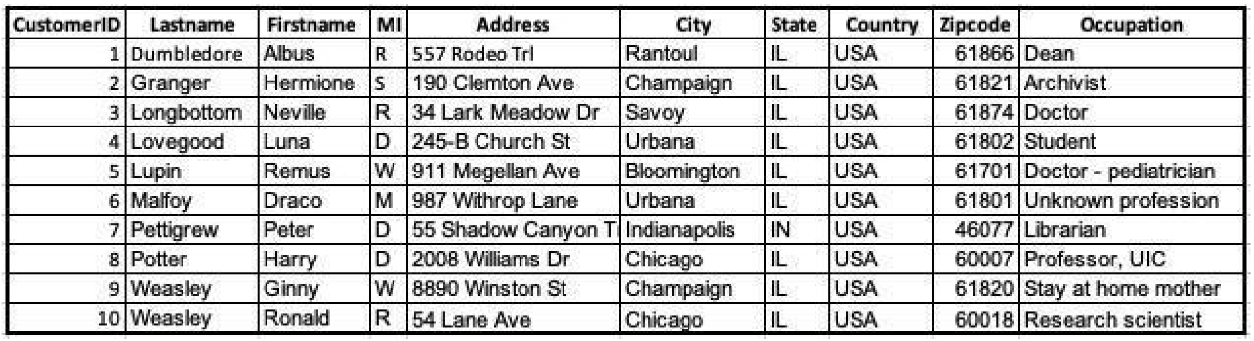
- Dependent on custom tools and application
- Dependent on memory and workplace practices
- Difficult to preserve fDifficult to documentor future use
- Difficult to repurpose and reuse
- Data Inconsistency between different files

Therefore, the preliminary goal is to leverage an adaptable technologic tool to manage and analyze the dataset, making it more organized and readable. As a consequence, the three tables are converted to the .xlsx format shown below.



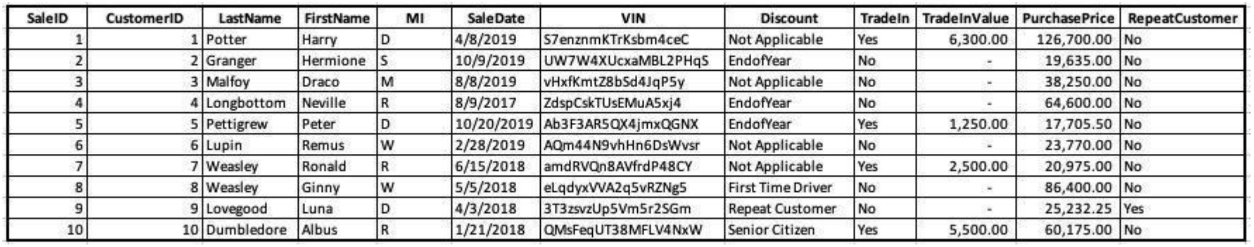
ID	VIN	Year	Model	Power	Drive	Color	DoorsNumbers	Engine	MSRP
1	vHxfKmtZ8bSd4JqP5y	2019	FordFlexSEL	150D	4WD	Black	4	Internal Combustion	35,240.00
2	Ab3F3AR5QX4jmxQGNX	2020	FordEcosportS	75D	4WD	Red	4	Internal Combustion	22,080.00
3	S7enznmKTrKsbm4ceC	2019	Tesla Model S	100D	AWD	Blue	4	Electric	133,000.00
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	ToyotaPriusL	150D	FWD	Blue	4	Hybrid	23,770.00
9	amdRVQn8AVfrdP48CY	2018	ToyotaPrius	75D	FWD	Silver	4	Hybrid	23,475.00
10	3T3zsvzUp5Vm5r2SGm	2018	ToyotaPrius	75D	FWD	Black	5	Hybrid	30,565.00

File A table example



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 T	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 B table example



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

File C table example

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

V. Discussion of curation objectives, decisions, and activities

- Q: How did you decide to represent the data in the way that you did?

  A:

  The company has three departments, and preliminarily provided three databases, therefore, I determined to follow the structure and created three tables, which are inventory table, customer relations table, sales table respectively.

  In the inventory table, I set up the VIN number as a unique identifier to record the inventories. For each vehicle, the inventory department should provide relevant information like year, model, number of doors etc, as a consequence, I included these variables in the inventory database.

  In the customer relations table, the customer management department will need to assign a unique code to each customer to manage the customer relationship. As a result, I assigned the customer ID to be the primary key in this database. What is more, for each customer, the customer management department will need to capture the customer's information like name, address, city, state, etc. Hence, in the customer relations table, I included these columns in the database.

In the Sales table, first of all, the sales department will need a unique number to record each order. In this case, I generated a “SaleID” column to serve as the primary key. In addition, for each order, we will have a customer to make the order, and will have a commodity to be sold. Therefore, we include the “CustomerID” and “VIN” as foreign keys in this table. Additionally, we also hope to understand the customer’s information, as well as the commodity’s relevant information. Consequently, in the Sales table, I also included the customer’s name, vehicle’s purchase price, trade-in price etc.

- Q: Did you leave out any information? If so, why?

  A: In comparison with the original Sales table, I left out a couple of columns like “Address”, “City”, “State”, “Country”, “Engine”, “Color”, “Year” etc due to the reason that in the Sales table, we do not need to know the address of the customer, and we do not need to understand detailed information for each vehicle that is sold. This information will make the database more difficult to understand and read.

- Q: Why did you choose certain things as attributes? As keys?

  A: For instance, in the Sales table, I created a “SaleID” attribute and it acts as the primary key in this database since it represents a unique order number for each order. In addition, for each order, the sales department will have a customer to make the order, and will have a commodity to be sold. Therefore, I generated the “CustomerID” and “VIN” as foreign keys in this table. Additionally, some basic information like the customer’s first name, last name, the VIN number of the sold car, and the purchase price of the sold commodity are critical for each sale. As a result, I included these pieces of information in the Sales table.

- Q: What were the hardest decisions you had to make in this design process?

  A: The hardest decision in this design process for me was to determine that in each table, what attributes should be included. For the original inventory database and customer database, since there is no title for each attribute, I need to spend time understanding what the attributes stand for. Furthermore, in the original sales table, there are several redundant attributes, and I need to decide which variables to be dropped from the updated sales table.

- Q: How does your schema design support data independence?

  A: To support data independence, I generated a relational model and applied these two principals when generating the relational schema: Abstraction and Indirection.

  - ***Abstraction:*** Before implementing the concise data points in each table, I designed three relational tables. In the relational tables, I just developed the attributes in each table, and the attribute data types like string, integer etc.
  - ***Indirection:*** Besides including the primary key in each table, I also designed the foreign keys as the lineage between the three tables. For instance, in the Sales table, I included the "VIN" and "CustomerID" as the foreign keys to link the other two tables.

- Q: How may your schema design support the overarching goals of data curation (revisit objectives and activities of Week 1)?

A: The objective of data curation is to be concerned with all aspects of management of data in order to efficiently and reliably support the analysis of data, and enable reuse over time. To support this objective, the curatorial activities include collection, organization, storage, etc.

My schema design has fulfilled the objectives of data curation and also fulfilled majority of the curatorial activities, for example:

- ***Collection:*** Compared with the original database, the new schema design ensures that all the three tables are stored in the same format, which makes it easier to collect and acquisite the data.
- ***Organization:*** The original three tables do not have standard criteria. However, my new schema design defines each attribute in a table and the corresponding data types, which ensures deployment of an appropriate data model. To be specific, for each table, we standardize the attribute that can be populated and the data type that should be populated under the respective attribute.
- ***Discoverability:*** In comparison with the original database, the new databases are stored in an Excel format, which can be easily converted to a CSV format. By storing the data in an Excel format, the user can easily leverage the search function in Excel to identify the specific data values. In addition, the titles are assigned for the inventory table, customer table as well as the sales table, hence, it supports the ability to search for and locate relevant data. What is more, the Excel table can be easily imported into SQL databases or other technological tools like Python, Alteryx etc.

- Q: What are the pros and cons of your schema design?

  Pros:

  - It defines the attributes and the data types of each attribute to support the reliable and effective storage of the data points. Additionally, by standardizing the attributes in each table, it ensures that data will be understandable and usable in the future.
  - By designing the ER Diagram, the relationship between the schemas can be easily understood, leading to the support of sharing data between different departments and other organizations.
  - It improves the data quality issue. For instance, for the mandatory attributes, the schema design requires that it should not be blank, hence there will not have missing values.

  Cons:

  - The schema design is based on Excel and I do not have time to import the database to SQL server to generate some syntax to display the data lineage. To modify the data values in Excel is not as convenient as in the SQL server.

- Q: Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional activities would you recommend?

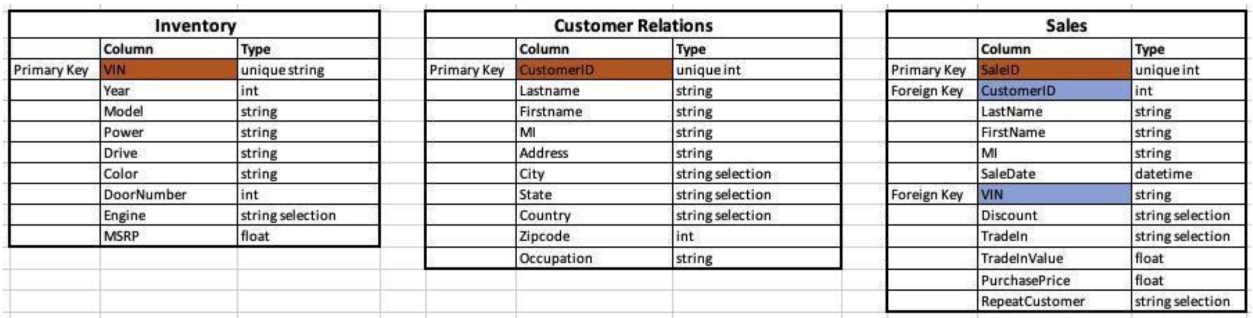
  A: Overall, the current schema design can fulfill majority of the curation activities, however, we can still make improvements in the following aspects:

  - ***Identification:*** The current schema design requires the user to populate the data points, however, it is difficult to identify if the user does not populate it accurately. For example, the schema design defines the attribute "VIN" to be a unique string, but if the user enters the duplicated strings, it is hard to find out the issues. The schema design needs to support the ability to identify, authenticate, and validate data accuracy.

- ***Compliance:*** At present, the schema design is not involved in the areas of compliance, which aims at ensuring the legal, regulatory, and local policy requirements.
- ***Security:*** The three databases are stored as an Excel format, which is easy to access and distribute. The schema can be designed to ensure that data is secure from tampering or inappropriate access and distribution.

VI. Overall quality analysis and completeness

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



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

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

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

Schema design for the inventory, customer and sales table

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”, columns “CustomerID” and “VIN” are added to answer the questions “Which customer makes this order”, “Which model is sold”, and these two columns serve as the foreign key to link the 3 various tables.

Generally, the schema design standardizes the attributes of the inventory table, customer relationship table, as well as the sales table, and also defines the data types respectively. By standardizing the attributes, it ensures that each database will contain the titles, leading to easier comprehension and distribution. In addition, it is easier to track and monitor if there are missing values in the mandatory attributes, enhancing the efficiency and effectiveness of the storage.