Assignment #1

Name: Saad Majidu Kulumba

The company is divided into sales, inventory, and customer departments. Each department is independent in terms of data management. However, it turns out that they would like to integrate their data into a central DB which will enable the analyzers to answer questions such as, "what is Ecrin's car make?." As an example. This is a challenge at the moment since data is across departments.

Three files are saved in different file formats: .txt, .csv, and Docx. That is for inventory, Sales, and Customer data, respectively.

About the Files

FileA (Inventory)

Inventory is a quick, easy-to-read text file since it is tab-spaced. A single row represents one record in this file.

File A Problems

- At a glance, this file is missing attribute headers. One has to figure it out by looking at the values, which is challenging.
- The price values are strings instead of actual integers. As a result, there is a need for extra work to make this valuable data for analysis and computations.
- There are some missing values, column (SEL, S2.OL)in particular. perhaps it was something optional. This h to be included to make this file uniform
- The door_field is a mix of integers and strings, which should be just an integer.

FileB (Sales)

Sales is a comma-separated value file with headers making it easy to understand by just scanning through

File B Problems

- Inconsistency. i.e., data is missing in City, State, and Country fields. Similar data is present in the customer file.
- Some pieces of data are not in the expected order. As a result, it leads to misunderstanding of the record. For instance, the discount and trade-in values are scattered.
- The $ sign MSRP field, Trade-In, and purchase price make these values unusable for computations.
- The discount field should be some integer value to be used for calculations.
- RepeatCustomer field is redundant, in my opinion, since it appears under discount.
- The model field should be expanded into Make and model for easy sub-classification of cars.

File C (Customer)

File C is a word format file, relatively easy to read at a glance. The data in this file is most probably the customer name, surname, address, profession, MI, and zip code.

File C Problems

- This file does not have headers to tell the different fields. So I made comparisons between the records to figure it out.

Shema diagram

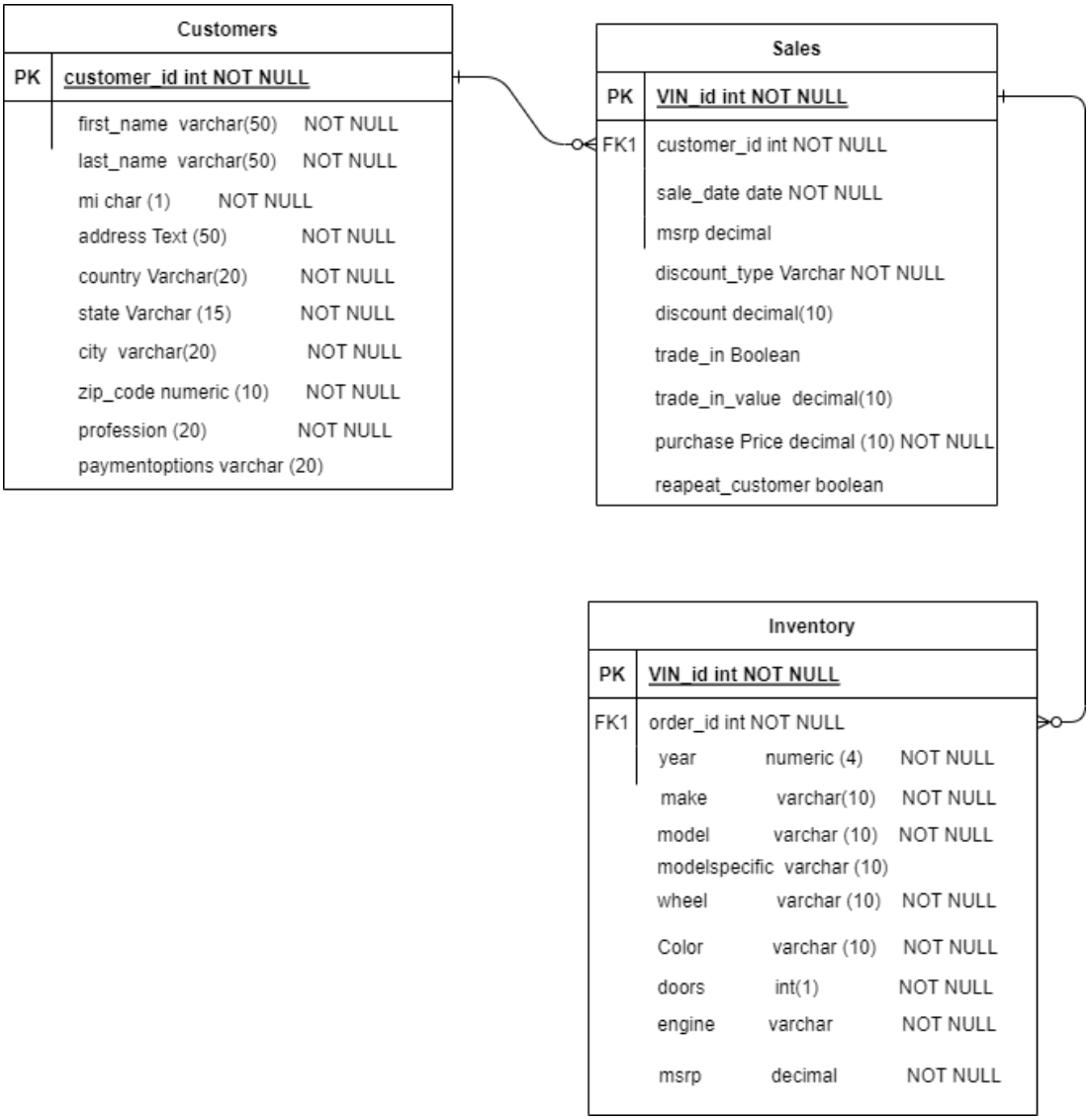


Figure 1 logical database schema

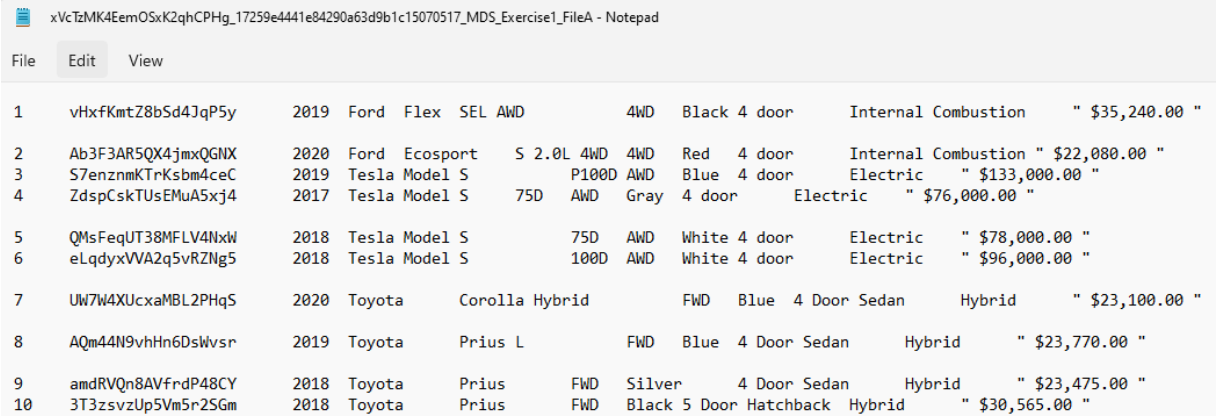
PK -Primary Key | FK – Foreign Key

VIN is the vehicle identification number that is unique to each vehicle. This is primarily in the Sales table

CustomerID is created to be unique in the customer table

An example of each table populated with data from the file

Inventory- File A



xVcTzMK4EemOSxK2qhCPHg_17259e4441e84290a63d9b1c15070517_MDS_Exercise1_FileA - Notepad

File Edit View

```
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 "
```

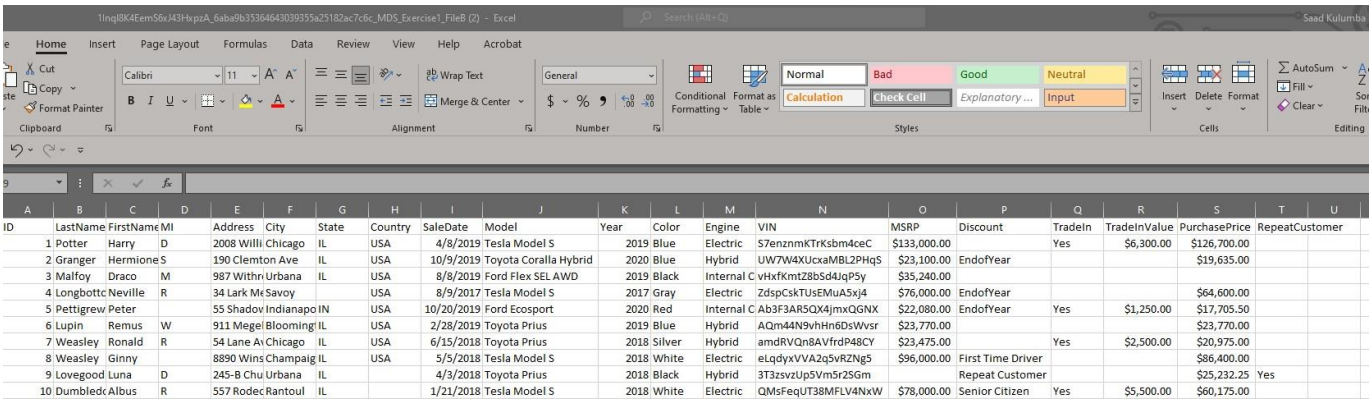
Tasks

- Moved file content to MS Excel
- Reformated MSRP to remove the dollar sign
- Organized values under correct fields
- Included headers
- Added a new field ModelSpecific to classify model
- Changed door field to be an integer instead of a varchar

Result Example table

					Inventory table- File A					
VIN ID	Year	Make	Model	SubModel	Wheel	color	Type	doors	Engine	MSRP
vHxfKmtZ8bSd4JqP5y	2019	Ford	Flex	SEL	AWD	Black		4	Internal Combustion	35,240.00
Ab3F3AR5QX4jmxQGNX	2020	Ford	Ecosport	S 2.0L	4WD	Red		4	Internal Combustion	22,080.00
S7enznmKTrKsbm4ceC	2019	Tesla	Model S	P100D	AWD	Blue		4	Electric	133,000.00
ZdspCskTUsEMuA5xj4	2017	Tesla	Model S	75D	AWD	Gray		4	Electric	76,000.00
QMsFeqUT38MFLV4NxW	2018	Tesla	Model S	75D	AWD	White		4	Electric	78,000.00
eLqdyxVVA2q5vRZNg5	2018	Tesla	Model S	100D	AWD	White		4	Electric	96,000.00
UW7W4XUcxaMBL2PHqS	2020	Toyota	Corolla		FWD	Blue	Sedan	4	Hybrid	23,100.00
AQm44N9vhHn6DsWvsr	2019	Toyota	Prius L		FWD	Blue	Sedan	4	Hybrid	23,770.00
amdRVQn8AVfrdP48CY	2018	Toyota	Prius		FWD	Silver	Sedan	4	Hybrid	23,475.00
3T3zsvzUp5Vm5r2SGm	2018	Toyota	Prius		FWD	Black	Hatchbak	5	Hybrid	30,565.00

Sales- File B



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 Willi	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 Withn	Urbana	IL	USA	8/8/2019	Ford Flex SEL AWD	2019	Black	Internal C	vHxfKmtZ8bSd4JqP5y	$35,240.00					
4	Longbotto	Neville	R	34 Lark Me	Savoy		USA	8/9/2017	Tesla Model S	2017	Gray	Electric	ZdspCskTUsEMuA5xj4	$76,000.00	EndofYear			$64,600.00	
5	Pettigrew	Peter		55 Shadov	Indianapo	IN	USA	10/20/2019	Ford Ecosport	2020	Red	Internal C	Ab3F3AR5QX4jmxQGNX	$22,080.00	EndofYear	Yes	$1,250.00	$17,705.50	
6	Lupin	Remus	W	911 Megel	Blooming	IL	USA	2/28/2019	Toyota Prius	2019	Blue	Hybrid	AQm44N9vhHn6DsWvsr	$23,770.00				$23,770.00	
7	Weasley	Ronald	R	54 Lane A	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 Wins	Champaig	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 Chu	Urbana	IL		4/3/2018	Toyota Prius	2018	Black	Hybrid	3T3zsvzUp5Vm5r2SGm		Repeat Customer			$25,232.25	Yes
10	Dumbledo	Albus	R	557 Rodec	Rantoul	IL		1/21/2018	Tesla Model S	2018	White	Electric	QMsFeqUT38MFLV4NxW	$78,000.00	Senior Citizen	Yes	$5,500.00	$60,175.00	

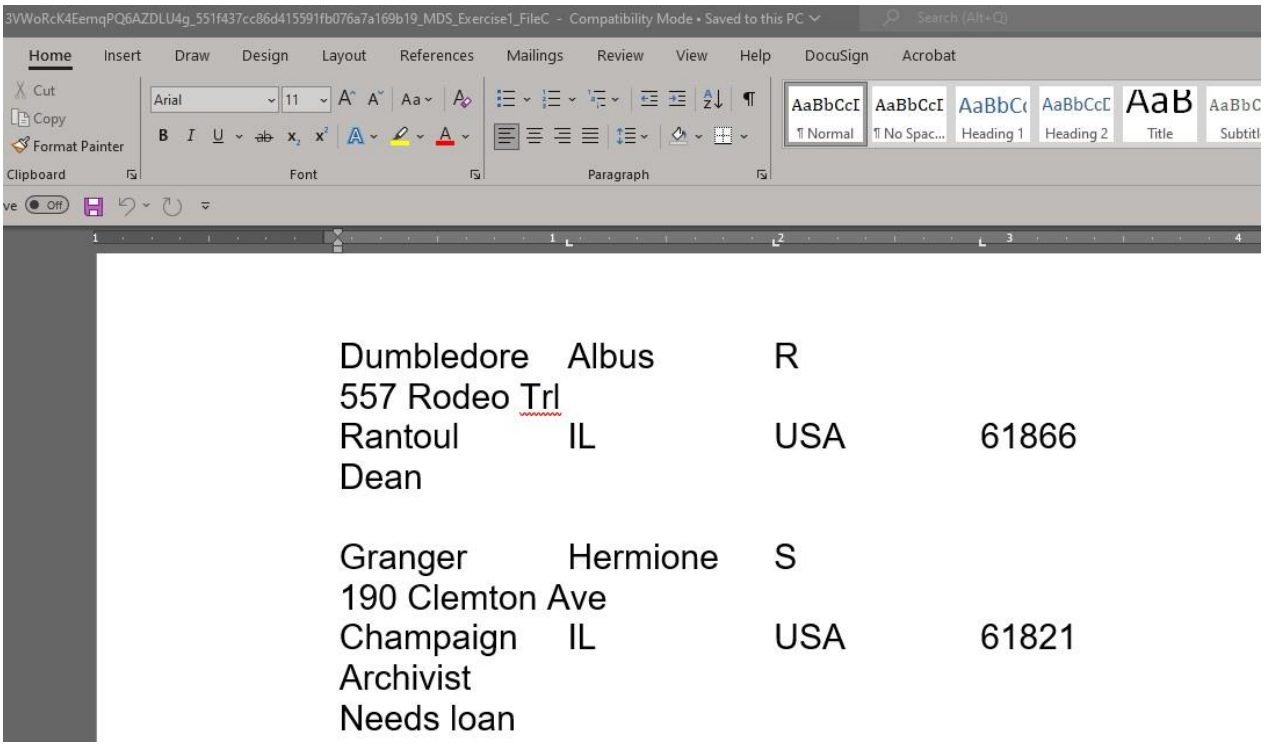
Tasks

- Eliminates fields that occur in the inventory table. This eliminates redudancy
- Personal information is already appearing in the customer table. So it's eliminated here
- Reformat MSRP, purchasePrice, and Tradein value by removing the dollar sign
- Excel did automatically convert the strings for me to numbers
- Eliminate color since it appears in inventory
- Add a discount field by calculating MSRP – purchase price

File B- Sales Table

CustomerID	VIN ID	SaleDate	MSRP	DiscountType	Discount	TradeIn	TradeInValue	PurchasePrice	RepeatCustomer
100001	S7enznmKTrKsbm4ceC	4/8/2019	133,000.00		6,300.00	Yes	6,300.00	126,700.00	
100002	UW7W4XUcxaMBL2PHqS	10/9/2019	23,100.00	EndofYear	3,465.00			19,635.00	
100003	vHxfKmtZ8bSd4JqP5y	8/8/2019	35,240.00		35,240.00				
100004	ZdspCskTUsEMuA5xj4	8/9/2017	76,000.00	EndofYear	11,400.00			64,600.00	
100005	Ab3F3AR5QX4jmxQGNX	10/20/2019	22,080.00	EndofYear	4,374.50	Yes	1,250.00	17,705.50	
100006	AQm44N9vhHn6DsWvsr	2/28/2019	23,770.00		0.00			23,770.00	
100007	amdRVQn8AVfrdP48CY	6/15/2018	23,475.00		2,500.00	Yes	2,500.00	20,975.00	
100008	eLqdyxVVA2q5vRZNg5	5/5/2018	96,000.00	First Time Driver	9,600.00			86,400.00	
100009	3T3zsvzUp5Vm5r2SGm	4/3/2018	25,232.25	Repeat Customer	0.00			25,232.25	Yes
100010	QMsFeqUT38MFLV4NxW	1/21/2018	78,000.00	Senior Citizen	17,825.00	Yes	5,500.00	60,175.00	

Customer – 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

Figure 2 Assignment file C

Tasks

- Transfer data from the word document to excel
- Transpose the columns into rows
- Name appropriate attribute names

Customer table -File C										
CustomerI	FirstName	LastNam	MI	Address	City	State	Country	ZipCode	Profession	PaymentOptions
100001	Dumbledore	Albus	R	557 Rodeo Trl	Rantoul	IL	USA	61866	Dean	
100002	Granger	Hermione	S	190 Clemton Ave	Champaign	IL	USA	61821	Archivist	Needs loan
100003	Longbottom	Neville	R	34 Lark Meadow Dr	Savoy	IL	USA	61874	Doctor	
100004	Lovegood	Luna	D	245-B Church St	Urbana	IL	USA	61802	Student	Needs loan
100005	Lupin	Remus	W	911 Megellan Ave	Bloomington	IL	USA	61701	Doctor - pediatrician	
100006	Malfoy	Draco	M	987 Withrop Lane	Urbana	IL	USA	61801	Unknown profession	
100007	Pettigrew	Peter		55 Shadow Canyon Trl	Indianapolis	IN	USA	46077	Librarian	Needs financing
100008	Potter	Harry	D	2008 Williams Dr	Chicago	IL	USA	60007	Professor, UIC	
100009	Weasley	Ginny		8890 Winston St	Champaign	IL	USA	61820	Stay at home mother	Inquiry into financing options
100010	Weasley	Ronald	R	54 Lane Ave	Chicago	IL	USA	60018	Research scientist	

Questions

1. **Why my representation?**

   The goal was to combine data across departments of this auto company. So by creating relationships between departments with distinct rows and columns eliminates redundancy. More storage space was being used by storing the same records in different tables.

   The way I've presented it, is more efficient and faster to use for querying

2. **Any info left out?**

   No, all information has been included. Except that I've reduced columns in the sales table because we already have them in the Inventory table. It's up to the user to combine which data they need.

3. **Why I chose CustomerID, VIN and order as keys**

   By looking at the data itself, I'm able to tell what kind of values are in the table. So I derive the attributes by examining the record. For instance, in file C. (customers), I was able to extract attributes of FirstName, LastName, MI, Address, City, State, and Zip.

   As for the key, I chose the most unique field to the table for the key. I did generate a customer ID which wasn't included. The VIN was unique enough to be the key to Inventory. In order to create a relation to these tables, I had to connect them using a Foreign Key. VIN appears in two tables, one in which it acts as a foreign key and the other as a primary key attribute.

4. **Difficult decisions of the process**

   I noticed that we had MSRP and Purchase price, which were not equal. So I came up with a discount field that was present. I included a calculation of MSRP – Purchase price to show how much discount the customer was given. Otherwise, it wasn't obvious.

5. **Data independence in the schema**

   This design supports data independence in a way that we are able to make the logic abstract to the user. Inventory data will be coming from a separate table as well as other tables. In case of changes in the user interface, this same date will still be available. Data can be added and deleted in separate tables and later combined when needed.

6. **Support for data curation goals**

   The overall goal of data curation is to incorporate all data management aspects. From collection to design, schema formatting, organizing, modifying, integrating, reformatting, workflow, communication, and discoverability. I have enhanced the schema By obtaining the given files and creating a design, defining appropriate attributes, assigning

primary and foreign keys, setting some constraints, and eliminating repeated data. In the end, integrate all tables into a relation. This contributes to the goals of data curation.

7. **Pros and cons of my design**

Pros

- Schema documentation leads to better organization and flow of information
- Its easily transferable and may be shared with other users
- Manages integrity by ensuring data validity. For example, it can help avoid data duplication

Cons

- Designing appropriate fields is rather tedious

8. **Additional activities that I'd recommend**

Another curation activity would be data security since we are handling some sensitive information. There should be a hierarchy of data access. A data breach could happen, and all data may be lost or stolen since we are living in the information age.