Designing Schema for an Auto Dealer

SHRASHTI SINGHAL
Data Curation- IS 531
Department of Information Science
The University of Illinois at Urbana Champaign

Abstract

The setting is an auto dealer. In this company, there are three departments, including Inventory, Sales and Customer relations. Right now, each department manages their information differently.

These departments would like to integrate their data into a shared database, to be able to answer questions like, “What engine is in Customer Smith’s car?.” It’s a challenge to answer right now because the Sales department has information about which car Customer Smith bought; the Inventory department tracks which engines are in which cars; and the Customer Relations has entirely separate, slightly redundant details on Customer Smith.

The solution is to design and populate a database that is effective and efficient for all departments, and that will also support combining data from different departments.

I. Legacy Dataset

The Auto Dealer has three files to save his Inventory, Sales and Customer Related Data. The formats of all three files are different from each other.

I. I File A (Inventory): It is a text file. Though it is a text file, the data of one particular inventory is organized in one single row. The data is readable. The different values are separated by each other by a tab space.

Problem Areas:

*This document doesn’t contain headings of the attributes. Finding which data values belong to which attribute is a challenge.

*Most of the attribute fields looks mandatory while few are optional. The optional attributes include Sub model and Type of Vehicle.

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

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

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

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

I.II File B (Sales): This is a CSV file. It is the most organized file with headings of the attribute. This file can be helpful in finding the titles of the data present in the inventory and customer files. 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.

Problem Areas:

*In Customer details, City, State, and Country details have some missing fields, which are otherwise present in the Customer file. This makes the data inconsistent. Moreover, Complete Customer details in the sales files are redundant.

*Sale Date is the date of sale of the vehicle, with inconsistent date formats.

*Year attribute is unclear. It is not the year of Sale, as the Sale Date attribute is different from the Year attribute here. Moreover, it doesn’t even match with the year of Manufacturing in the inventory file. If it is the year of manufacturing, then the data is inconsistent with the Inventory file.

*MSPR, Purchase Price and Trade-In Value, all the values are in dollars. Adding a dollar to the data makes the field as text and leaves it useless for calculations.

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

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

*Subcategories of Colour are mentioned in one single attribute.

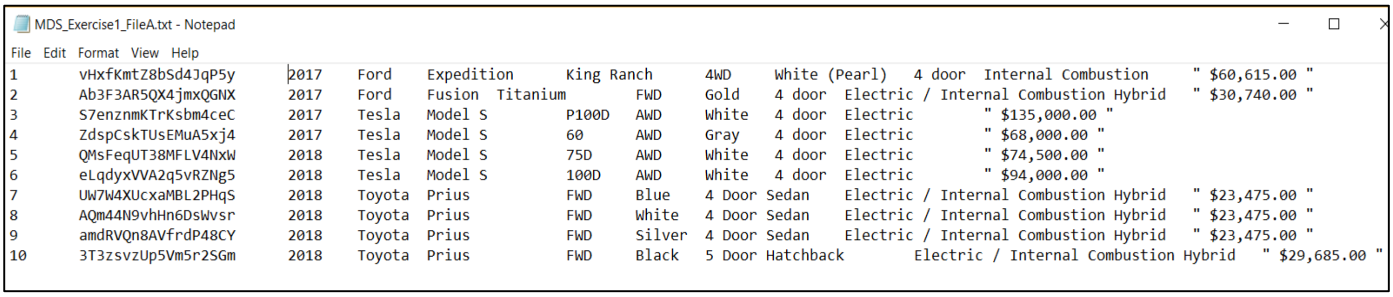
* Few data values are missing from the purchase price and the MSRP attribute, which should be mandatory.

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

I.III File C (Customer): This is a word document file, which is the most unreadable format when needed for data search. 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. A single tuple is ended by 2 consecutive character returns, which is an empty line in the word document. Attributes are separated by tabs in the document.

Problem Areas:

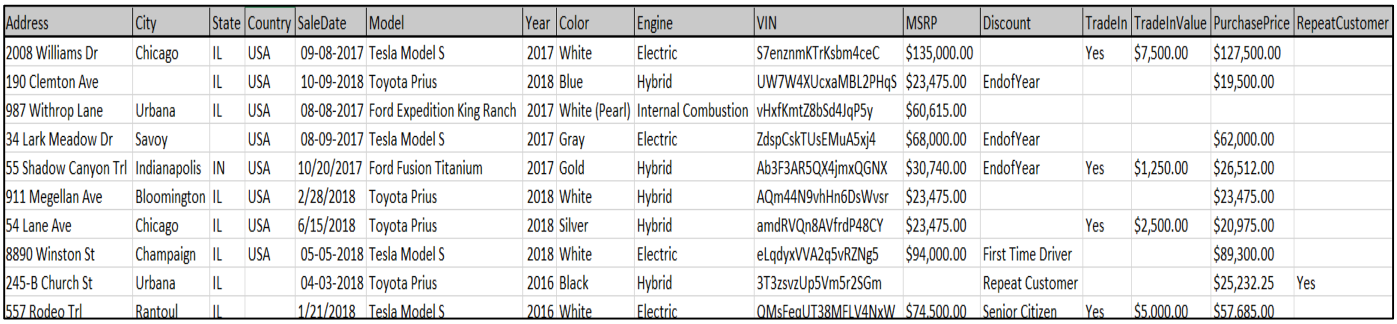
*This file contains no headings of the attributes listed. To check the attribute headings, we will need to compare the data to that of the Sales file.



MDS_Exercise1_FileA.txt - Notepad

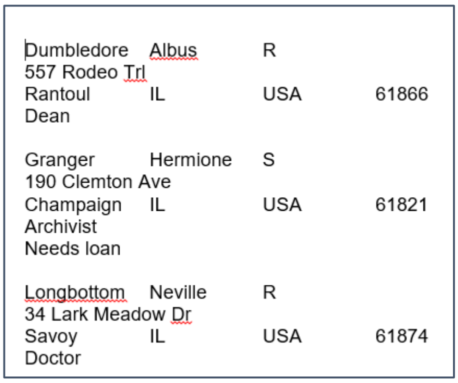
```
1       vHxfKmtZ8bSd4JqP5y      2017    Ford    Expedition      King Ranch      4WD     White (Pearl)   4 door  Internal Combustion     " $60,615.00 "
2       Ab3F3AR5QX4jmxQGNX      2017    Ford    Fusion  Titanium        FWD     Gold    4 door  Electric / Internal Combustion Hybrid   " $30,740.00 "
3       S7enznmKTrKsbm4ceC      2017    Tesla   Model S         P100D   AWD     White   4 door  Electric        " $135,000.00 "
4       ZdspCskTUsEMuA5xj4      2017    Tesla   Model S         60      AWD     Gray    4 door  Electric        " $68,000.00 "
5       QMsFeqUT38MFLV4NxW      2018    Tesla   Model S         75D     AWD     White   4 door  Electric        " $74,500.00 "
6       eLqdyxVVA2q5vRZNg5      2018    Tesla   Model S         100D    AWD     White   4 door  Electric        " $94,000.00 "
7       UW7W4XUcxaMBL2PHqS      2018    Toyota  Prius           FWD     Blue    4 Door Sedan    Electric / Internal Combustion Hybrid   " $23,475.00 "
8       AQm44N9vhHn6DsWvsr      2018    Toyota  Prius           FWD     White   4 Door Sedan    Electric / Internal Combustion Hybrid   " $23,475.00 "
9       amdRVQn8AVfrdP48CY      2018    Toyota  Prius           FWD     Silver  4 Door Sedan    Electric / Internal Combustion Hybrid   " $23,475.00 "
10      3T3zsvzUp5Vm5r2SGm      2018    Toyota  Prius           FWD     Black   5 Door Hatchback        Electric / Internal Combustion Hybrid   " $29,685.00 "
```

Image1 (Inventory): FileA.txt



Address	City	State	Country	SaleDate	Model	Year	Color	Engine	VIN	MSRP	Discount	TradeIn	TradeInValue	PurchasePrice	RepeatCustomer
2008 Williams Dr	Chicago	IL	USA	09-08-2017	Tesla Model S	2017	White	Electric	S7enznmKTrKsbm4ceC	$135,000.00		Yes	$7,500.00	$127,500.00	
190 Clemton Ave		IL	USA	10-09-2018	Toyota Prius	2018	Blue	Hybrid	UW7W4XUcxaMBL2PHqS	$23,475.00	EndofYear			$19,500.00	
987 Withrop Lane	Urbana	IL	USA	08-08-2017	Ford Expedition King Ranch	2017	White (Pearl)	Internal Combustion	vHxfKmtZ8bSd4JqP5y	$60,615.00					
34 Lark Meadow Dr	Savoy		USA	08-09-2017	Tesla Model S	2017	Gray	Electric	ZdspCskTUsEMuA5xj4	$68,000.00	EndofYear			$62,000.00	
55 Shadow Canyon Trl	Indianapolis	IN	USA	10/20/2017	Ford Fusion Titanium	2017	Gold	Hybrid	Ab3F3AR5QX4jmxQGNX	$30,740.00	EndofYear	Yes	$1,250.00	$26,512.00	
911 Megellan Ave	Bloomington	IL	USA	2/28/2018	Toyota Prius	2018	White	Hybrid	AQm44N9vhHn6DsWvsr	$23,475.00				$23,475.00	
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	
8890 Winston St	Champaign	IL	USA	05-05-2018	Tesla Model S	2018	White	Electric	eLqdyxVVA2q5vRZNg5	$94,000.00	First Time Driver			$89,300.00	
245-B Church St	Urbana	IL		04-03-2018	Toyota Prius	2016	Black	Hybrid	3T3zsvzUp5Vm5r2SGm		Repeat Customer			$25,232.25	Yes
557 Rodeo Trl	Rantoul	IL		1/21/2018	Tesla Model S	2016	White	Electric	QMsFeqUT38MFLV4NxW	$74,500.00	Senior Citizen	Yes	$5,000.00	$57,685.00	

Image2 (Sales): FileB.csv



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

Image3 (Customer): FileC.docx

II. Curatorial Activities:

II. I Collection, Sharing, Communication, Modification and Reformatting: Data is collected from all the three departments. The data is obtained in three different formats. The inventory data is in text format, Sales data is in CSV format, and Customer relations data is a word document.

As data is converted to CSV, which is the simplest way of storing the data. The data can be easily converted to other formats. CSV supports reformatting for use by different tools or to match new format standards.

CSV, as the most widely used tool for data storage, sharing becomes more comfortable. Moreover, excel has features to display the data in various forms, example; line, bar and pie charts, tables, pivot tables, etc. Therefore it supports communication. CSV also promotes easy management of corrections and updates.

II.II Processing: Data is processed in the below steps:

Inventory File:

1. Convert Inventory file text file into excel. Copy data from the text file into an excel sheet.
2. Arrange data values for 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.
3. Provide Attributes to inventory file. This can be done by checking the data from Sales file. Give the same headings as in Sales file. For remaining, data fields give the appropriate titles.
4. The data value in the engine column has multiple values separated by slashes. Remove the redundant text.
5. Make a subcategory of color, to accommodate different color shades.
6. Make the MSRP field as an integer by removing the dollar symbol from there.
7. Make the number of doors attribute as an integer by removing the redundant string.

Sales File:

8. 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.
9. 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.
10. Make all sale date data values to one format which is MM/DD/YYYY.
11. Year attribute here is neither sale date year nor manufacturing year. It has mismatches with both

the years' data. For keeping consistency, we delete the year attribute.

12. MSPR, Purchase Price and Trade-In Value attributes are strings. These attribute wouldn't be available for analytics in this format. We remove the dollar symbol from these data to convert into an integer.
13. The model attribute is a single entity here, while in 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.
14. Repeat customer could be deleted. But we don't have enough proofs if all repeat customer gets the discount. To have the complete record, we will retain this field.
15. Colour attribute looks redundant, as it is available in inventory records.
16. 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 Sales record from Inventory record.
17. An important field is missing which is the discount amount. It is useful for analytics purpose and a mandate in Sales file.
18. Purchase Price attribute data is missing for few rows. We will complete it by deducting the discount price from MSPR.

Customer File:

19. Copy the data from the word file into an excel sheet.
20. Convert data into rows and columns.
21. Provide heading/attribute names to the data by comparing it to the Sales file and appropriate headings to the data specific to this file.

II.III Organization & Workflow: Employ an appropriate data model and use appropriate standards. Document schema attributes (including specifying data types and constraints).

Relational Model will be implemented because Data from different sources can be integrated more easily. Data will be more accessible to check for validity and quality. Data independence is supported.

The processing of data has been carried out a well-designed modular system of transformations.

Keys:

Inventory File: VIN (Vehicle Identification Number is unique among vehicles around the world. No two different cars would have the same VIN. Therefore VIN is the primary key of the inventory tuple. VIN is the foreign key to communicate with the Sales file.

Sales File: VIN is the primary key for the Sales file as well. Lastname, Firstname and Address would be the compound foreign, unique key, to communicate with the customer records. VIN is the foreign key for the Inventory file.

Customer File: There is no unique attribute which satisfies to be a primary key. Therefore, we have created another id, which would be a customer id, unique for all customers. We can derive unique customer id for existing data by combining attributes the first name, last name, and address. Customer Id can be the primary key for customer records and serve as a foreign key for the Sales file.

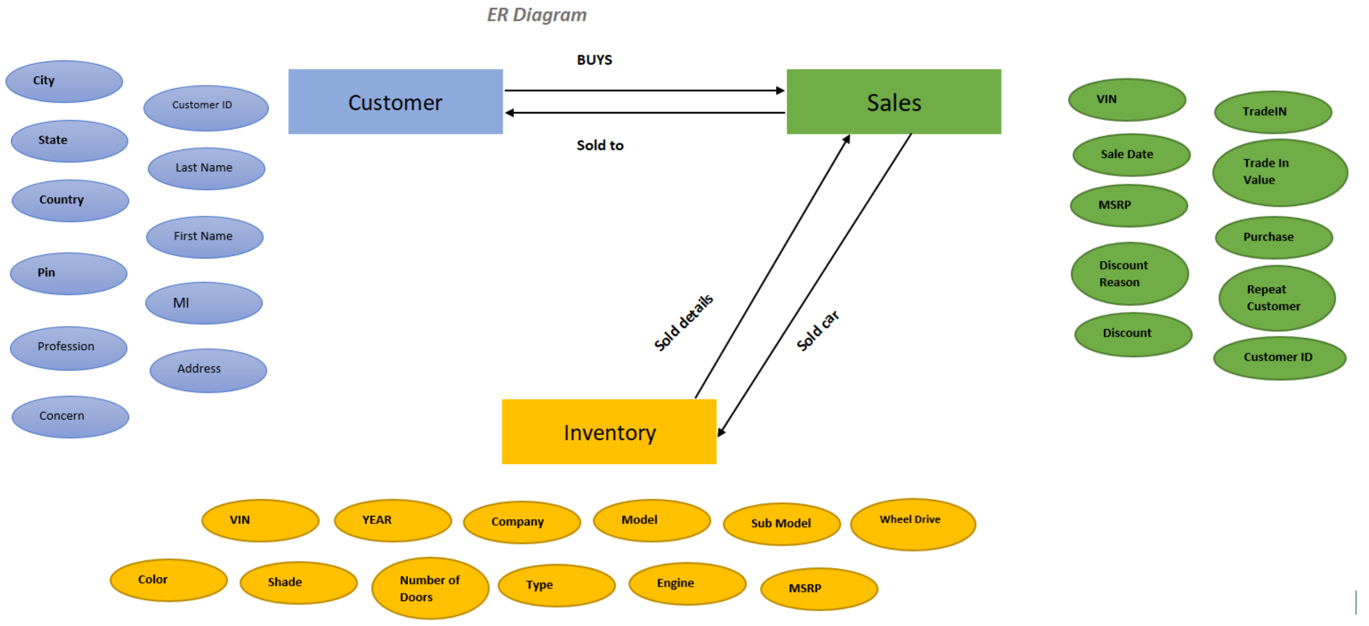
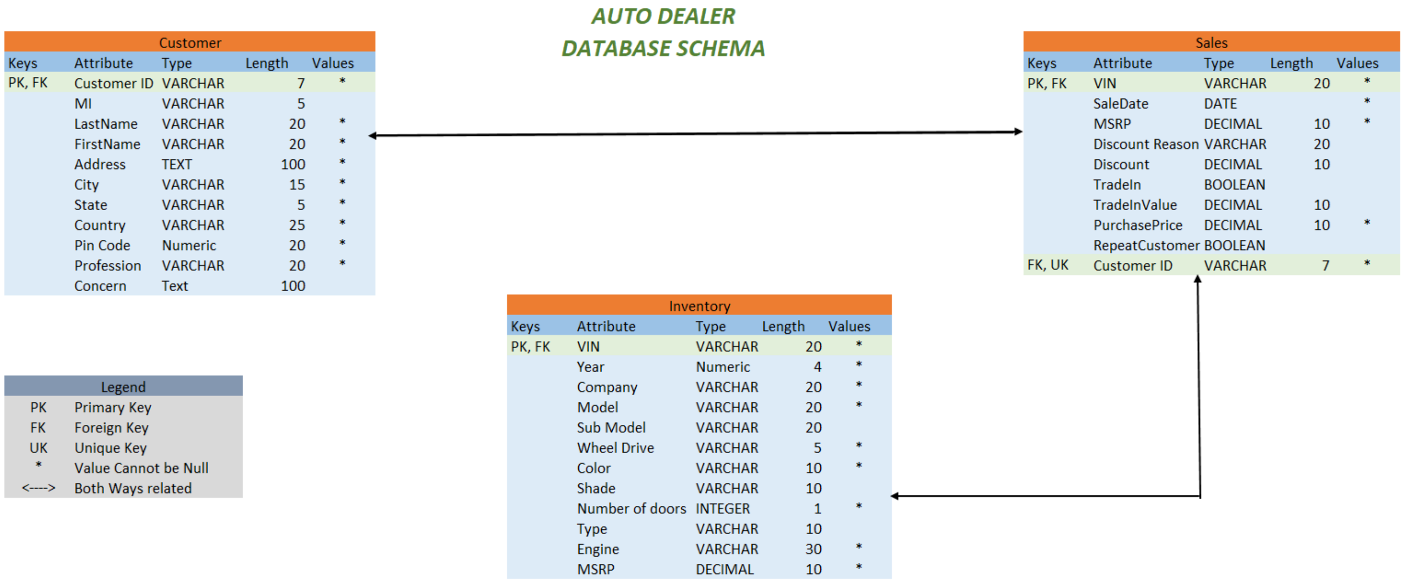


Image 4: ER Diagram- Conceptual Schema



AUTO DEALER
DATABASE SCHEMA

Customer

Keys	Attribute	Type	Length	Values
PK, FK	Customer ID	VARCHAR	7	*
	MI	VARCHAR	5	
	LastName	VARCHAR	20	*
	FirstName	VARCHAR	20	*
	Address	TEXT	100	*
	City	VARCHAR	15	*
	State	VARCHAR	5	*
	Country	VARCHAR	25	*
	Pin Code	Numeric	20	*
	Profession	VARCHAR	20	*
	Concern	Text	100	

Sales

Keys	Attribute	Type	Length	Values
PK, FK	VIN	VARCHAR	20	*
	SaleDate	DATE		*
	MSRP	DECIMAL	10	*
	Discount Reason	VARCHAR	20	
	Discount	DECIMAL	10	
	TradeIn	BOOLEAN		
	TradeInValue	DECIMAL	10	
	PurchasePrice	DECIMAL	10	*
	RepeatCustomer	BOOLEAN		
FK, UK	Customer ID	VARCHAR	7	*

Inventory

Keys	Attribute	Type	Length	Values
PK, FK	VIN	VARCHAR	20	*
	Year	Numeric	4	*
	Company	VARCHAR	20	*
	Model	VARCHAR	20	*
	Sub Model	VARCHAR	20	
	Wheel Drive	VARCHAR	5	*
	Color	VARCHAR	10	*
	Shade	VARCHAR	10	
	Number of doors	INTEGER	1	*
	Type	VARCHAR	10	
	Engine	VARCHAR	30	*
	MSRP	DECIMAL	10	*

Legend

PK	Primary Key
FK	Foreign Key
UK	Unique Key
*	Value Cannot be Null
<---->	Both Ways related

Image 5: Relational Database Schema- Logical Schema

II.IV. Storage, Identification, Security, and Access: All the three records are stored in an excel workbook. Each worksheet is for a particular department and contains the data relevant to them. These sheets can be password protected by the individual department. The data model will support the ability to identify, authenticate, and validate data.

II.V. Preservation: To make the data understandable and usable in the future, the database schema is developed, and the data in these tables are modified according to the database schema. This document will serve as a documented preservation strategy.

II.VI. Discoverability & Integration: Data is discoverable as the data is converted in desired data types to make it quantifiable. Moreover, keys are provided to each dataset so that data tuples can communicate amongst each other.

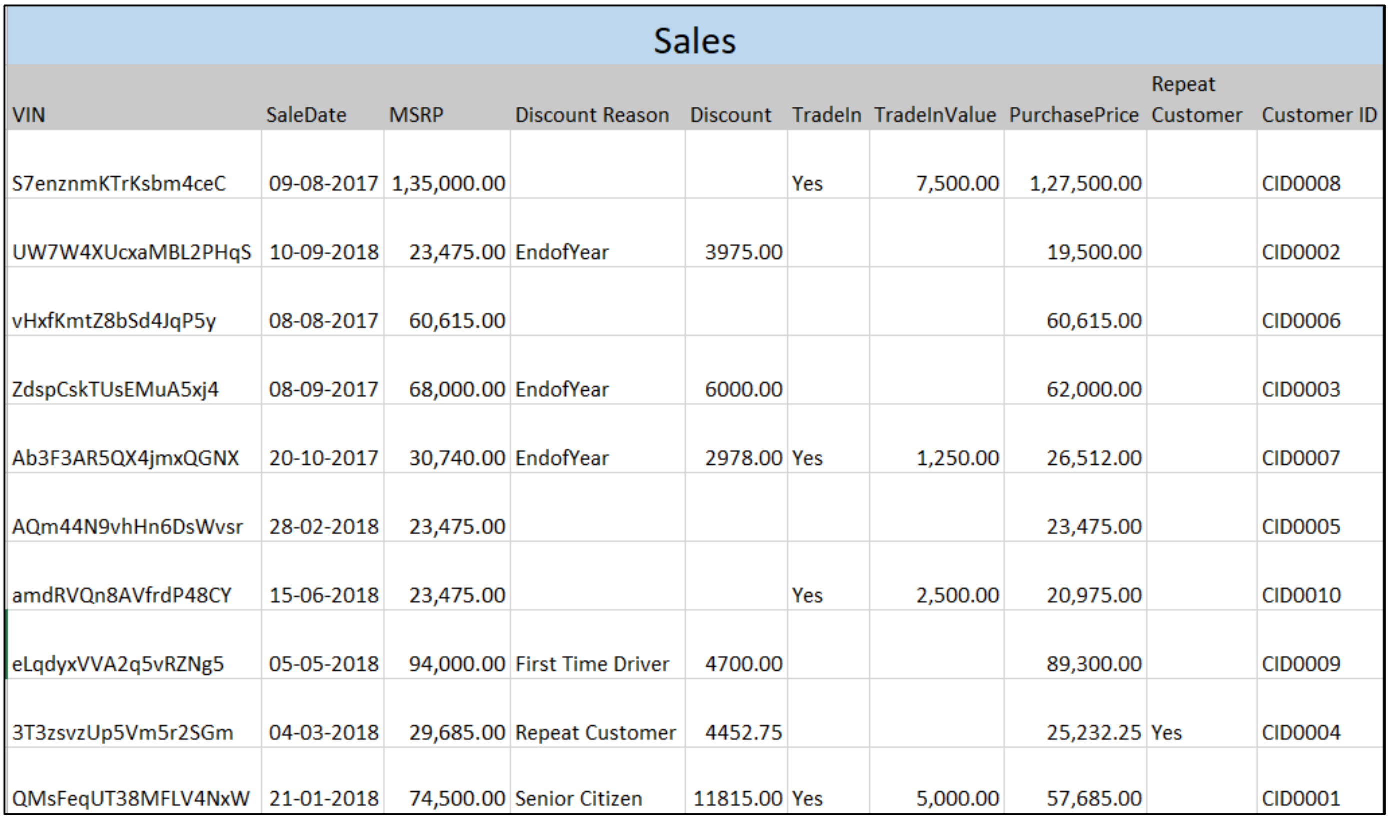
III. Example of Auto Dealer Database

The data has been populated in the tables using the relational Schema. Below are the examples of the data filled through the Schema, satisfying the constraints of the designed schema.

Inventory

VIN	Year	Company	Model	Sub Model	Wheel Drive	Color	Shade	Number of doors	Type	Engine	MSRP
vHxfKmtZ8bSd4JqP5y	2017	Ford	Expedition	King Ranch	4WD	White	Pearl	4		Internal Combustion	60,615.00
Ab3F3AR5QX4jmxQGNX	2017	Ford	Fusion	Titanium	FWD	Gold		4		Hybrid	30,740.00
S7enznmKTrKsbm4ceC	2017	Tesla	Model S	P100D	AWD	White		4		Electric	1,35,000.00
ZdspCskTUsEMuA5xj4	2017	Tesla	Model S	60	AWD	Gray		4		Electric	68,000.00
QMsFeqUT38MFLV4NxW	2018	Tesla	Model S	75D	AWD	White		4		Electric	74,500.00
eLqdyxVVA2q5vRZNg5	2018	Tesla	Model S	100D	AWD	White		4		Electric	94,000.00
UW7W4XUcxaMBL2PHqS	2018	Toyota	Prius		FWD	Blue		4	Sedan	Hybrid	23,475.00
AQm44N9vhHn6DsWvsr	2018	Toyota	Prius		FWD	White		4	Sedan	Hybrid	23,475.00
amdRVQn8AVfrdP48CY	2018	Toyota	Prius		FWD	Silver		4	Sedan	Hybrid	23,475.00
3T3zsvzUp5Vm5r2SGm	2018	Toyota	Prius		FWD	Black		5	Hatchback	Hybrid	29,685.00

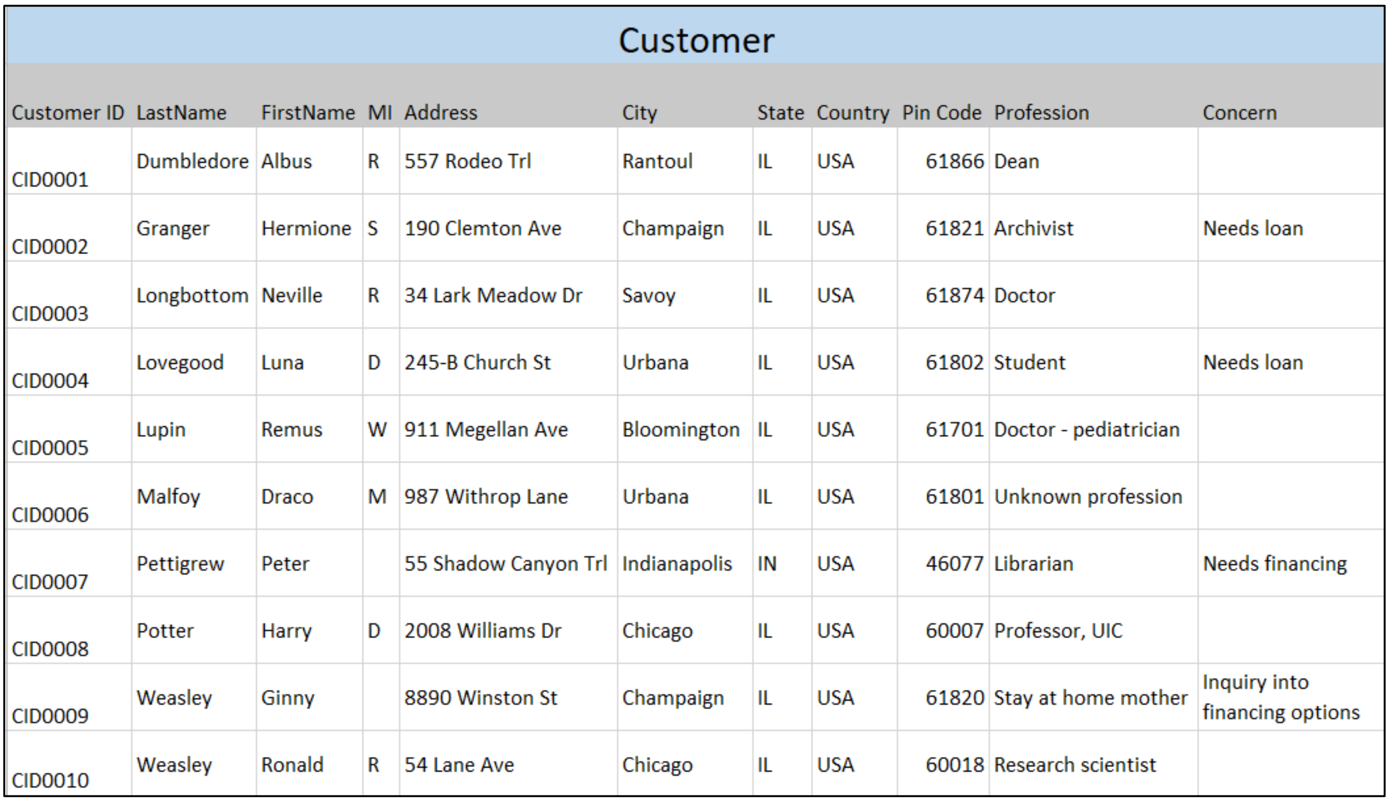
Table 1: Inventory Table



Sales

VIN	SaleDate	MSRP	Discount Reason	Discount	TradeIn	TradeInValue	PurchasePrice	Repeat Customer	Customer ID
S7enznmKTrKsbm4ceC	09-08-2017	1,35,000.00			Yes	7,500.00	1,27,500.00		CID0008
UW7W4XUcxaMBL2PHqS	10-09-2018	23,475.00	EndofYear	3975.00			19,500.00		CID0002
vHxfKmtZ8bSd4JqP5y	08-08-2017	60,615.00					60,615.00		CID0006
ZdspCskTUsEMuA5xj4	08-09-2017	68,000.00	EndofYear	6000.00			62,000.00		CID0003
Ab3F3AR5QX4jmxQGNX	20-10-2017	30,740.00	EndofYear	2978.00	Yes	1,250.00	26,512.00		CID0007
AQm44N9vhHn6DsWvsr	28-02-2018	23,475.00					23,475.00		CID0005
amdRVQn8AVfrdP48CY	15-06-2018	23,475.00			Yes	2,500.00	20,975.00		CID0010
eLqdyxVVA2q5vRZNg5	05-05-2018	94,000.00	First Time Driver	4700.00			89,300.00		CID0009
3T3zsvzUp5Vm5r2SGm	04-03-2018	29,685.00	Repeat Customer	4452.75			25,232.25	Yes	CID0004
QMsFeqUT38MFLV4NxW	21-01-2018	74,500.00	Senior Citizen	11815.00	Yes	5,000.00	57,685.00		CID0001

Table 2: Sales Table



Customer

Customer ID	LastName	FirstName	MI	Address	City	State	Country	Pin Code	Profession	Concern
CID0001	Dumbledore	Albus	R	557 Rodeo Trl	Rantoul	IL	USA	61866	Dean	
CID0002	Granger	Hermione	S	190 Clemton Ave	Champaign	IL	USA	61821	Archivist	Needs loan
CID0003	Longbottom	Neville	R	34 Lark Meadow Dr	Savoy	IL	USA	61874	Doctor	
CID0004	Lovegood	Luna	D	245-B Church St	Urbana	IL	USA	61802	Student	Needs loan
CID0005	Lupin	Remus	W	911 Megellan Ave	Bloomington	IL	USA	61701	Doctor - pediatrician	
CID0006	Malfoy	Draco	M	987 Withrop Lane	Urbana	IL	USA	61801	Unknown profession	
CID0007	Pettigrew	Peter		55 Shadow Canyon Trl	Indianapolis	IN	USA	46077	Librarian	Needs financing
CID0008	Potter	Harry	D	2008 Williams Dr	Chicago	IL	USA	60007	Professor, UIC	
CID0009	Weasley	Ginny		8890 Winston St	Champaign	IL	USA	61820	Stay at home mother	Inquiry into financing options
CID0010	Weasley	Ronald	R	54 Lane Ave	Chicago	IL	USA	60018	Research scientist	

Table 3: Customer Table

IV. Questions:

1. How did you decide to represent the data in the way that you did?
   The data is presented in a Relational model that is to conceptualize all information in terms of relations (rows, columns, values).

The relational model is a pure single high-level abstraction for conceptualizing knowledge; it is indifferent to the details of physical storage and processing. All interactions with the data are in relational (tabular) terms, such as attributes, values, tuples. Those interactions are translated into instructions expressed in terms of storage data instructions. Operations on data are based on

formally defined, well-understood activities from logic and set theory.
Moreover, the tree model is used to store hierarchical data, while ER Model is majorly used to show the relationship between different entities. We don't have either kind of the data set.

2. Did you leave out any information? If so, why?
No information has been left out. I leave it to analytics operation to filter the information needed by them. Here, I have captured all the details found out from the legacy dataset.
Though, I have rearranged the information. I have removed customer details such as first name, Last name, address, MI, city, state and country details from the Sales file. As those details were incomplete in Sales files, also it's better to maintain one kind of information in one place to avoid redundancy. The above values were replaced by Customer Id. I have also removed the year attribute from the Sales file for the same reason. Year attribute was present in the inventory file. Moreover, it was inconsistent with the data provided in the Inventory File.

3. Why did you choose certain things as attributes? As keys?
Selecting attributes was easy. The data at the auto dealer data set is stored as values and attributes. These attributes represent the category of data, example First name, Last name, Profession, etc. We just needed to find the appropriate name for the attributes, which suit the data category.
Keys are selected based on the uniqueness of the particular table as well as to support a way of interacting and exchanging information among the tables.
Example, VIN is selected as Primary key & foreign key for both Inventory and Sales files, as VIN is unique and supports interaction between the tables.
I wanted to remove the Customer details from the Sales file completely, but that would have hampered the process of interaction between other records to the customer file. Therefore Customer ID is used as the primary key for customer file and foreign key for Sales file.

4. What were the hardest decisions you had to make in this design process?
There were few hard decisions in designing the schema.
First was to add an extra attribute to the sales file, which is the discount amount. It was needed to calculate the missing purchase price in the sales file. Moreover, Discount information seems to be a must in the sales records.
Secondly, adding a customer Id as the primary key in the Customer file. This makes compulsory for the customer to remember their IDs, all the time in case of need to revisit to the store.
Thirdly, Retaining the Repeat customer attribute is dicey, assuming that all repeat customers might not get the discount.

5. How does your schema design support data independence?
Data independence is supported as CSV is used as a storage method which allows interoperability between most other softwares and can be changed without impacting the data. It provides data independence from storage. Therefore, if physical storage changes the end programs using the storage will continue to operate like before.
Also, new data constructs can be added without impacting the data, relationship and keys between the existing data.

6. How may your schema design support the overarching goals of data curation?
The overarching goal of Data Curation is concerned with all aspects of the data management, therefore, I have implemented most of the data Curation techniques/activities such as, Collection, designing Schema, Formatting, Organization, Modification, Integration, Reformatting, Workflow, Communication, and Discoverability including Data modelling in Conceptual and Logical Layer. This will allow efficient and reliable support to the analysis of data, and will also enable reuse over the time. How these activities have been incorporated and have enhanced the database design is mentioned above in this paper. We have used Data model, schema, primary key, and foreign key, not null, length and Unique key constraints, appropriate data types and removal of redundant and repeated data to maintain accurate and completeness of data. The above are all part of data curation activities.

7. Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional actions would you recommend?
Security and access control can be added by giving the users of the individual department of an auto dealer a password. So, no unwanted theft or tampering may occur.
Another curation activity is compliance, as we have personal details of customer's including their complete addresses and profession information, the auto dealer has to ensure compliance with legal, regulatory, and local policy requirements.
Provenance could enhance database for future purposes as a new derived/calculated attribute has

been added (discount amount) to Sales file taking account of Inventory file MSRP values. When one data set (or view) is derived from another, reliable use and understanding require that the inputs, calculations, and actions responsible for data values can be identified.

V. Conclusion

The dataset available at the auto dealer was unusable for analytics. It was

- Inefficient
- Error-prone
- Untrustworthy
- Difficult to document
- Difficult to repurpose and reuse
- Difficult to preserve for future use
- Dependent on memory and workplace practices
- Dependent on custom tools and application

After revamping the Database design by developing a database schema, merging different types of formatted data, formatting of data, removing redundant data and completing the incomplete data, answering the question like “What engine is in Customer Smith’s car?.” becomes easy.

Which would be, in Sales records, look for customer Smith’s ID, check his VIN, and search that VIN in inventory records. We will get all the data related to Mr. Smith’s car. His car engine, color, model, gates, type, etc.