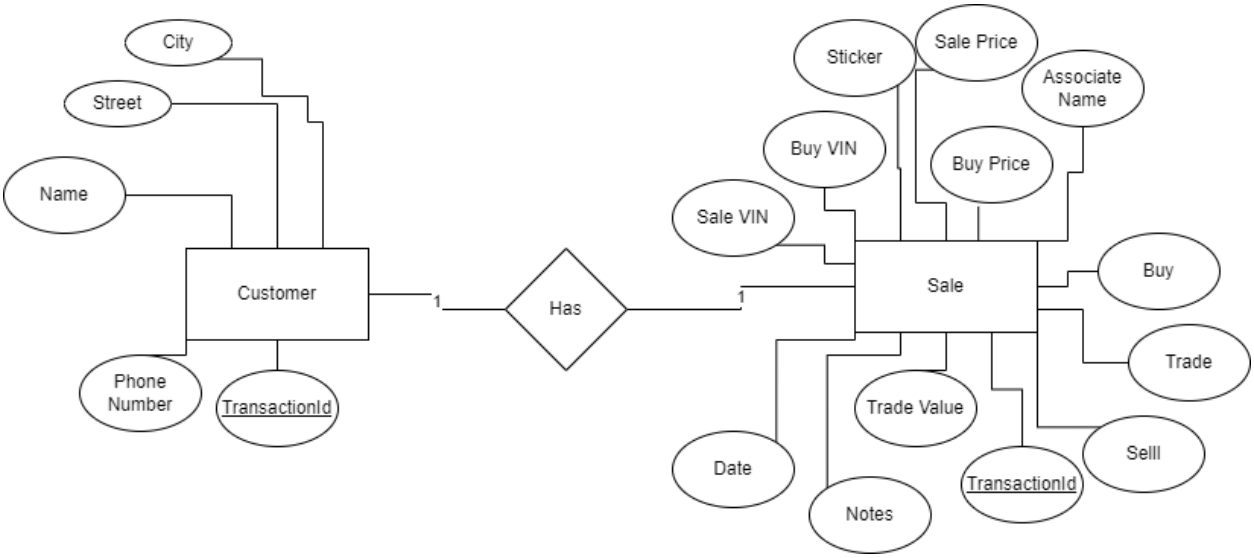
Saaif Ahmed

saaifza2

CS:598 Foundations of Data Curation

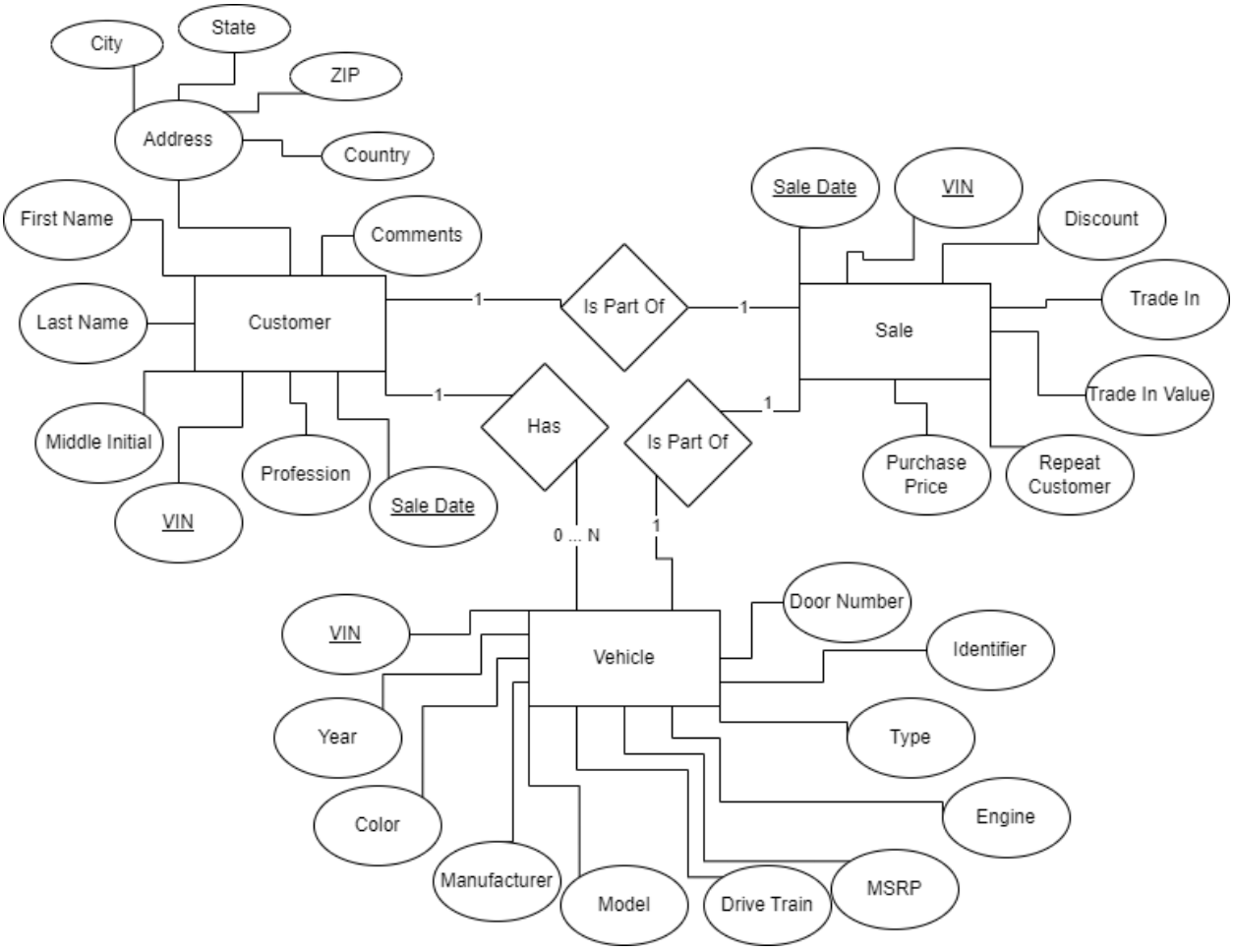
Assignment 1 Part 2




**Note:** All diagrams will be created with Chen Notation.

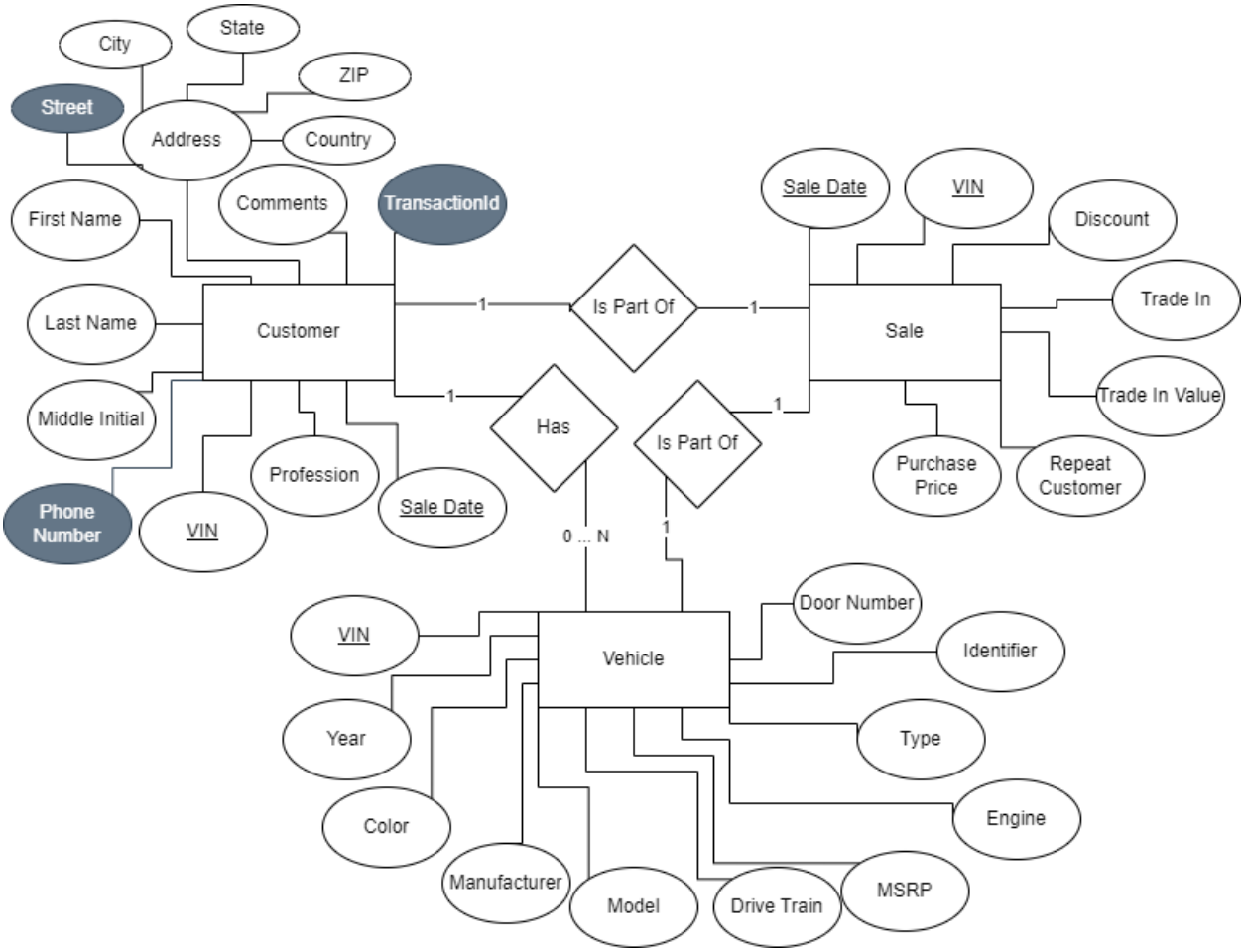
Question 1: ER Diagram for Pre-owned Dealer Database.

This ER Diagram illustrates the two main concepts present in the pre-owned database. There is a clear focus on Customers and Sales within the database. Thus, the ER diagram creates them as distinct entities. Within this diagram, and the database itself, a sale must have a customer and a customer must have a corresponding sale. This is shown by the linkage between the 2 entities with the key TransactionId. We show a 1 to 1 relationship with customers and transactions with the presumption that each customer makes 1 transaction at a time. In the scenario where a customer would make 2 transactions in one day, that would be 2 different TransactionId values to identify the unique parts of that. The ER diagram preserves all the data from the database, applying every attribute in each table to a corresponding entity. Lastly as a comment, the data had to be cleaned to create the diagram. Duplicate values, null values, and empty values made it unclear briefly what the relationships were. The benefit of this diagram is that all the data's intricacies are preserved and maintained with the same naming convention of the database itself. The largest consequence of this design is the lack of abstraction within the data itself to separate the pure transaction information from other factors such as discounts found in the Notes.




Question 2: ER Diagram for Part 1 Schema.

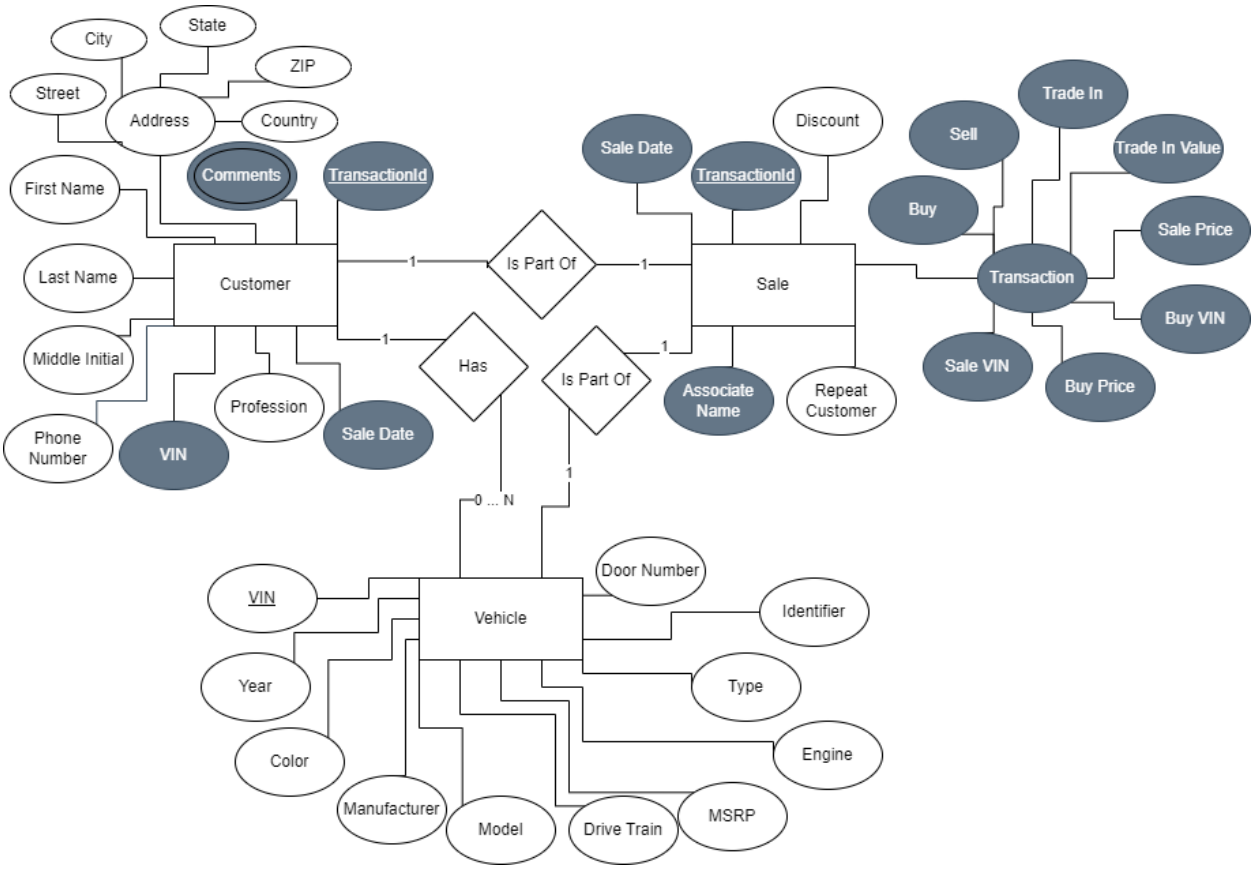
This ER diagram illustrates the 3 main concepts within the previous part's datasets. These concepts are customer, sale, and product. The fundamental idea between the schema is that any sale requires a product and a customer, and in this scenario the vehicle is the product. While the schema from Part 1 has more information present alongside the sales, for the sake of an ER diagram these inclusions are non-essential. No data is lost through this ER diagram however, as all the information that is not immediately present within an entity can be found using the key VIN. For Customer and Sale information the keys are a combination of Sale Date and VIN to handle the edge case scenario where a car is sold, traded in, and then purchased again by the same customer. That distinction allows the tables to be unique always. The relationships are extremely intuitive. Again, a sale must have a customer and a product. Thus, in the ER diagram we see that a customer "is a part of" a sale, and a vehicle "is a part of" sale. These are modeled as a 1 to 1 relationship because each sale is a unique transaction. Additionally, the customer can have 0-to-many cars and a car can only have 1 owner. This 1-to-many relationship is due to the fact that a car can be traded in, therefore a customer can no longer be the owner of the car but the transaction is marked as a trade in, and still able to be stored for records. The benefit of this diagram is the zero emission of any data from the datasets. The consequence is the lack of abstraction to the types of transactions such as sales, purchases, trade-ins, and their respective crucial information such as VIN numbers and prices.




Question 3: Intermediate Step #1.

**Note:** Changes in each intermediate step will be marked in grey.

To begin integrating the schemas together we can break it down into 3 steps, integrating into the customers entity, the sale entity, and final the vehicle entity. The pre-owned database holds more information surrounding customers that can be easily implemented as attributes to the existing customers entity. By making the Address attribute a composite attribute we can identify the different parts of the address we may want to use for data processing. For example, if the dealership wants to analyze their customers from Chicago, they are now capable. We also include the TransactionId for the customer. This is a unique identifier that can be easily added to entries that do not have this value, by means of a hashing function, and will serve a greater purpose in the final design. Furthermore, it preserves the data being integrated from the pre-owned dataset. This integration step is the combining schema step.

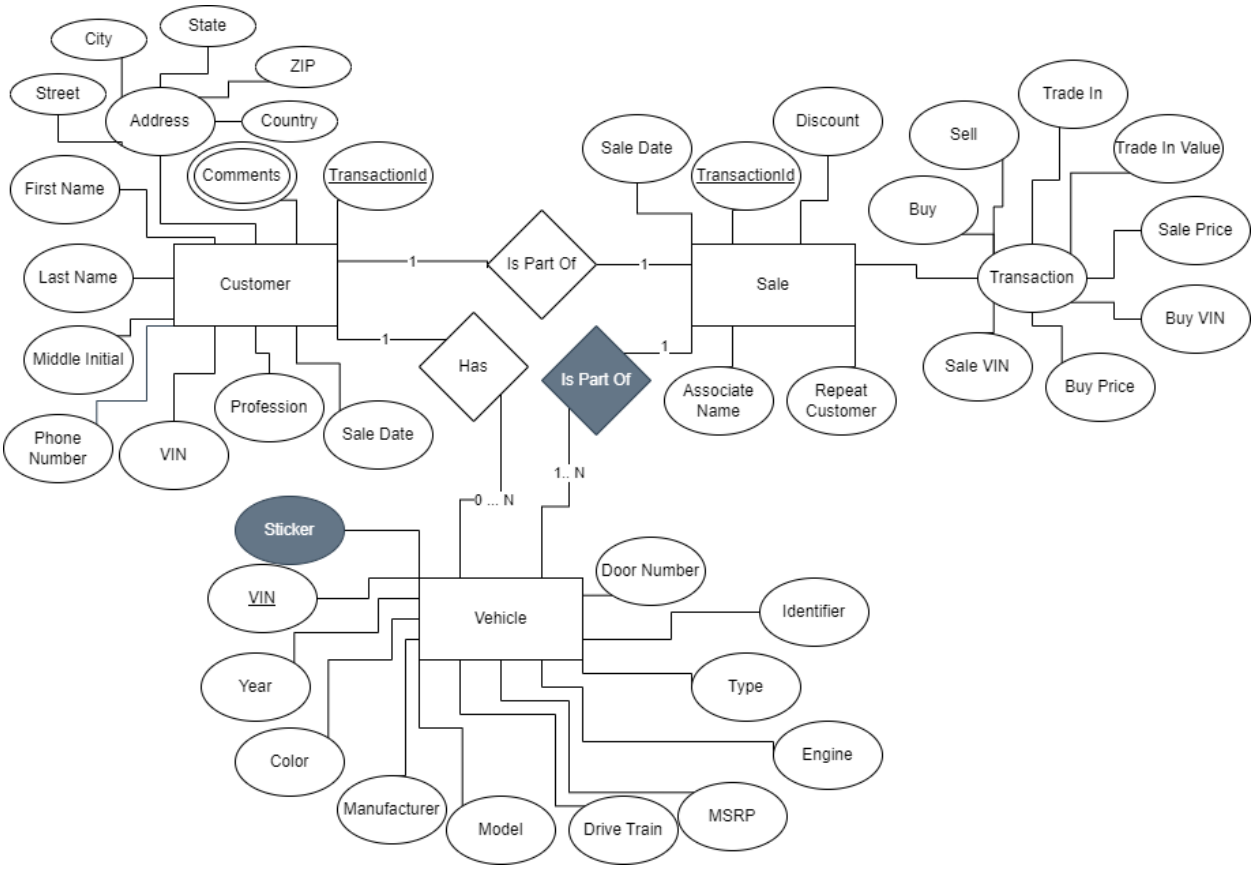



Question 3: Intermediate Step #2.

This is a large step in the integration process that can be broken down in to 3 substeps. First is the transition from the VIN + Sale Date keys to the TransactionId as a primary key. By including the TransactionId into the sale entity we are now able to uniquely identify each business exchange conducted by the dealership and who that business was conducted with. By doing this we eliminate all confusion in edge case scenarios.

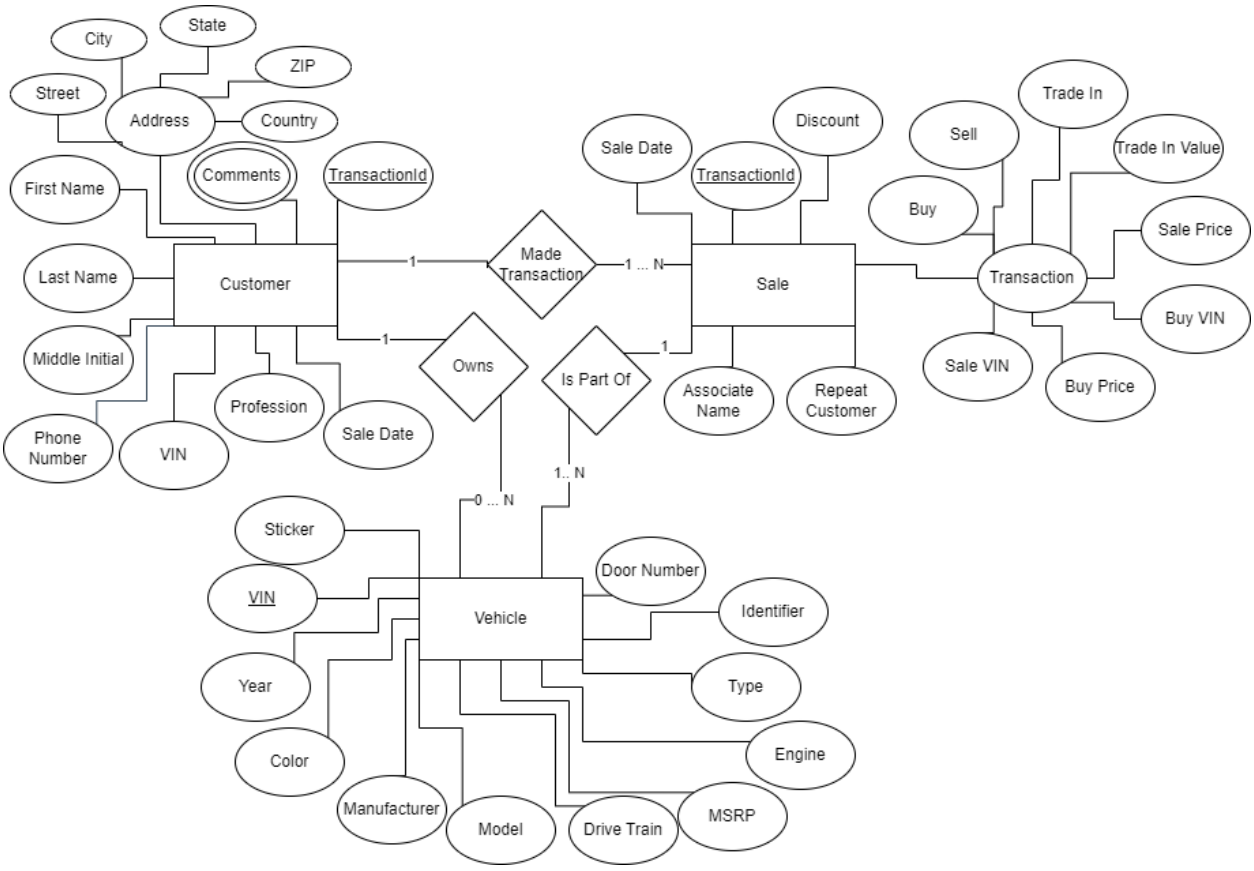
Secondly, the introduction of the Transaction composite attribute is very large. By merging with the pre-owned dealership, there is now a requirement to distinguish between Buys (the dealership buys the vehicle), Sells (the dealership sells a car to customer), and Trade Ins (a combination of a buy and a sell). The Transaction attribute has 3 attributes, buy, sell, and trade in, that can be used to identify what type of transaction a given TransactionId is. Because a Trade In is a combination of a buy and sell we need only include the 2 VINs, Buy VIN and Sale VIN, to identify the vehicles relating to that trade in. The Buy Price, Sale Price, Trade In Value are the main transactions record keepers within their specific form of transaction, and the Trade In Value in particular is to ensure no data is lost in the integration.

Lastly are the auxiliary changes made to increase the quality of the data and reduce any potential data loss. The Associate's Name is now added to the Sale entity to indicate who conducted the sale. The synonym problem is resolved between the different forms of VIN. The separation of Buy VIN and Sale VIN is needed to preserve the intent of the data, vehicles sold versus vehicles bought. The synonym problem is also resolved by Sale Date. We migrate "Data" from the pre-owned database to "Sale Date". Comments are now multivalued attributes as information about each customer is varied.




Question 3: Intermediate Step #3.

This stage of the integration is focused on integrating into the Vehicle entity and adjusting relation names and cardinality. Sticker is the last attribute not accounted for by previous iterations. For a dealership the sticker price being alongside the MSRP attribute makes the most sense for relevant business calculations. These include markups and markdowns in addition to market rate adjustments. In this iteration the only relationship changed is the “Is Part Of” connecting Sale and Vehicle. The others will be changed in the final diagram. For this relation change the cardinality of the Sale/Vehicle relationship changed. Now that sales can include multiple cars for a given TransactionId it is important to note the 1-to-many relationship present within the data.




Question 4: Final ER Diagram.

This is the final ER diagram. The only changes from the 3rd intermediate step and the final are relationship name changes, and cardinality. The relation between customer to sales is now 1-to-many. The intent of the data is to indicate that any customer may receive multiple vehicles from the dealership, especially if they fall under the Repeat Customer attribute. Thus 1 customer may be a part of many sales. The name has been changed as well now going by Made Transaction. This is to increase the readability and understanding of the diagram by following the logic that a customer makes transactions in a sale. The other relationship that changes is between customer and vehicle as it was name changed to “Owns”. Again, this was done to increase readability and discoverability of the data. Ownership of vehicles makes sense for a dealership and its records.

Question 5: Additional Narrative Prose

Much of the discussion surrounding intent, justifications, and explanation has been done in the comments under each ER diagram. This section will serve to defend the ER diagram and discuss potential changes. The ER diagram was made with the full intention of preserving as much data as possible through the integration. In addition, the ER diagram was made to support the curatorial activities that the previous schema had supported. Discoverability, organization, and problem solving were the main curatorial activities focused on in the previous design. This ER diagram features everything that the dealership would need to continue operations at high efficiency. Any realistic

question can be asked and answered through this ER diagram after the integration of the pre-owned data. For example, one can now determine how many vehicles that were traded in were bought by the dealership at 80% or lower than their MSRP. This can be checked by checking the transactions within the sales entity. The organization is very similar to the previous ER diagram and schema with the largest change concentrated on the transaction composite attribute. Furthermore, the discoverability is still at the same high level it was prior to the data integration.

There are changes that can be made to this diagram that may increase some aspects towards data curation and its objectives but fall short in other places. Primarily, a major difference can be in the implementation of the transaction composite attribute. An alternate implementation is to make 3 composite attributes of Buy, Sell, and Trade In, and generate sub-attributes for the transaction specific VINs, prices, and values. While this consolidates the information on the ER diagram quite neatly, it does lead to issues. Firstly, there is the matter of identifying the type of transaction within a query. These attributes would have to be populated in a similar fashion otherwise data would be lost in integration. Furthermore, the sub-attributes would overlap in purpose as a Trade In can be modeled as a Buy and Sell. Therefore, in the final ER diagram this method was not pursued. Another spot for improvement is in the comments attribute. Data is present in written information much more than it is in numerical form. While great analysis can be conducted via the numbers supplied by the ER diagram and subsequent schema, the comments for each transaction hold valuable information. One approach would be to analyze what exists within the comments and derive some standard naming conventions for the attributes found in them should any exist. This is a high-risk high-reward improvement that can be invested into to increase the quality of the integration.