1. **Write a short narrative description of each file we have given you (about 150~300 words). Be sure to remark on the overall quality of the data (are there errors, missing values, etc.), the contents of each data file, and any data which overlaps between files.**

File A contains data about vehicles. Every vehicle is identified using a VIN which is a unique id associated with every vehicle. The other field includes the details of the vehicle like model number, manufacturer ,color of the car, year , whether the car is sedan or hatchback, whether the car is electric or not and the price of the car. The data is not properly formatted and one can’t identify the attributes as they are not present in the file. Thus, there could be a margin of error in inferring the data fields. The data seems to be complete with no missing values. The data is redundant as some information is also present in file B/sales department. Ideally the details should be present with the inventory department only and the sales department should be having a vehicle id with which the details can be fetched from the inventory department.

File B contains the data from sales department. The data contains customer information and basic details like name ,address , sale data like purchase price, tradein value ,discount. It also contains details about vehicle like vehicle identification number vin ,model , color,engine. The data is redundant as some information present in file a and file c like vehicle data and customer details. There are some missing values in the data for various attributes. This is acceptable/tolerable for some of the attributes but not for some attributes like purchase price which is a crucial piece of information. Ideally the sales department should have only sales data and a reference to customer details and vehicle details which can be then used to fetch information from sales department data by using shared database. Some of the missing values can be fetched from other files like address details like city and state which are present in File C. The discrepancy/missing data in one file/department being present in other department/file shows the importance of normalization and removing data redundancy by creating a reference to data.

FileC contains data about the basic details of the customer the name , address , profession and whether he needs a loan, financing or inquiring into financing option. The name and address is present for all the records however for one of the records profession is not present. Also loans, financing option is missing for some of the records. The address details and name are also present in other files and this makes data redundant. Ideally the customer details should have been present in one place with a customer id and then other files/departments could have a reference to the customer id attribute of the customer relations department.

* How did you decide to represent the data in the way that you did?

The basic idea was to create a different table for minute details and a parent/central table which contains high level data about every sale i.e. sale id, vin and customer id.

* Did you leave out any information? If so, why?

No, we didn’t leave any information that is present in file. The field repeat customer could be removed because it is most likely used to provide repeat customer discount. However it’s good to have data rather than omitting it.

* Why did you choose certain things as attributes? As keys?

The id is central table is primary key and set to auto increment. Sale ID , Vin and Customer ID were chosen as foreign keys for the central table so that they can be used to get minute details from respective tables. The attributes and keys for other tables are chosen according to the data given. Some new fields like customer id has been created so that it avoids duplication of data if an existing customer has bought a new vehicle. Instead of adding a new tuple with customer details we can just use a customer id in the central table for which we only have data in the customer table. Similary for VehicleData the attributes are according to the data provided and Vin is the primary key. Same goes for the sales department.

* What were the hardest decisions you had to make in this design process?

The hardest decisions were to keep which attributes in which table. Some of the data can be present in any of the three tables.

* How does your schema design support data independence?

The schema is not impacted by how the data is stored. The same schema can be stored in multiple types of physical storage that is take care by data base management software. Also adding new data to this schema will not cause any discrepancy in the existing data this is because of the fact that we are following a relation model to represent our data.

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

Collection : The data can be collected from various departments to populate the database.

Organization : We are using relational data model to organize the data

Storage The relational model/schema can be stored using any relational data management software

Preservation The attributes will help in understanding the data. We have a schema and that will help us in identifying data present in tuples.

Discoverability: The data can be searched using the keys that are present in central table. To fetch the key a search can be performed on the tuples of tables of various departments to get the key.

Access The access can be taken care by the database management software where we will translate this data model into the real world.

Workflow Support the ability to systematize data workflows. This will also be take care by database management software. A utility can be written on the top of the database apis to automate various workflows and that can be documented. Eg we can document how data will be populated in various tables.

Identification We have various constraints and attributes and keys which will help in authenticating and validating the data

Integration This is also a portion of how the data will be populated In various relations/table. The same data coming from a source can be populated in various tables according to the attributes.

Compliance This will be ensured by the legal team of the organization. Any data that we need to populate in the database should be first approved by the legal team.

Security: This is also managed by database management software but we need to ensure that the permissions and ACLs are managed accordingly

* What are the pros and cons of your schema design?

Our schema provides data authentication, referential integrity, data independence and also tries to eliminate data redundancy using normalization.

Since we have designed a relational schema for storing our data it comes with few of disadvantages. For eg searching for a customer id(primary key) using a name will fetch results slower and then searching the vehicle that customer has will take a search to central table using customer id and then the final step would be to query the vehicle table using vin we got from central table.

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

Supporting reproducibility of data to reproduce meaningful insights can be helpful. The data can be used to get various insights on the cars that various customers are trying to purchase by profiling them on the basis of profession, geography etc. This will help the sales department to target new customers. Sharing this data with an internal R&D team will help in analysis of data which is essentially the goal of data curation.