Step 1:
File Description:

File A:

First file doesn’t have column headers. This file is tab delimited file. This file contains car inventory details:
1.year make
2. color of car
3.Max retail price of car
4 Electric or non-electric
5.Type of car: Sedan, Hatchback
6.VIN number of car

First two records in inventory has limited edition cars like Eddie Bauer edition Ford. File A shares information with the file B but not with file C.
Some redundant information is present in this file like tesla model price is same for white and gray model. This file can be divided into two multiple tables to avoid data redundancy like mentioned.

File B:

File B has column headers clearly defined. This file is comma delimited file format.This file has sales information from the auto dealers for the car. File has information like customer first name and last name,VIN,car type, Address of customers, type of Discount given on car ,trade in flag ,tradein value, Selling Price.
This file as gives information if customer is a repeat customer for the auto dealer.

File A and File B share information like car VIN,car type,MSRP,year of car,engine.

File C:

File C has information about the customer like details about the profession of the customer, if customer needs loan or financing, address of customer. File is in doc format.The information is not organized into columns so column metadata is missing in this file. File C and file B share information like customer last name first name, address. File C has zip code also which is missing in the file B. So file C customer address can be considered as master information. File C has not of null values which we have discovered while doing data profiling on the file. In the data model we have marked them as NA as default value in case there is null value.

Step 4:
Process for creating the database scheme and tables

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

Varun:

I used normalization process to come up with 4 different tables.
1.CarDetails : master car table with car details.
2.Inventory : Inventory details for the auto dealer
3.Customer : master table for customer information.
4.SalesTransaction : Fact table with transactions sales.

I created one extra table for cardetail to avoid repeating the information again and again for the cardetails.
This car detail table is master table with all types of cars while inventory table will have entry if dealer has some inventory about the table.
Sales transaction table will have the sales information with surrogate key added to identify each record.
With inventory table natural key of VIN is used as primary key.
.I have added customer active status to ensure that if the address of the customer is changed we are still tracking the old address of the customer and marking it as inactive record.
We will create new customer record in this case with new address.

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

I did not leave out information. Information given here is valuable for discovery purpose.
I renamed one column: Purchase price for car to selling price in sales table to make it more appropriate for table column

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

Varun:
I used attributes for tables which can define the table.
I have used keys for example VIN in inventory table as this key can uniquely identity the tuple.

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

Varun:
Hardest decision was to create two tables for:
CarDetails and Inventory.

I was thinking to create one initially and then 2.
Finally decided to go for 2 to have more granular based on normalization.

How does your schema design support data independence?

Varun:
Yes schema supports logical data independence as new field can be added to the schema tables.
Physical tables mapping can be adjusted to achieve data independence.

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

Varun:
Yes we can join these tables and extract useful information which is the main goal of data curation.
Information like which user bought which car on which date can be queries easily using the join on the surrogate keys and natural key combinations.

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

Varun: Storing more historical data to more analysis.
Also security of the sensitive information I want to do on this data set.

Data profiling to confirm the data types for the fields and precision can be done.
Data profiling can help to find for example exact precision required for the fields which can help in saving storage.
Compliance to legal and local policy requirement is one requirement for that source of data need to be known like who entered this data and if security is there in entry system
Also address validation can be done on the address field to improve data quality of the customer master records.