1. **Introduction / Narrative**

The problem statement here involves building a suitable database across all (three) given departments for an auto company.

The following three departments have their own data and in ‘different formats’.

Inventory department: The text file provided has names of cars and their features such as color, model, year, Drive Train, doors and price. The data is formatted and organized in terms of variable-length fields, delimited with spaces. Thus, it takes quite some time to interpret the columns and arrange them in place.

Sales department: The csv file provided has details of customers and the type of car they have purchased. It showcases sales information such as date on which car was sold, its price, discounts availed, whether the car was bought for a trade-in of their old car etc. In addition, it also contains certain details of purchased car, such as model, color, engine. Personal information of the customers such as address and whether he/she is a repeat customer is also provided. The information is formatted in the form of a csv file and is much easier to read compared to the inventory data.

Customer Relations department: This data contains personal information of the customers. Again, this is also formatted and organized in terms of variable-length fields, delimited with spaces making it very hard to read and write standard subroutines.

What kinds of information is shared among the files?

-Among the inventory and sales department, information such as VIN, model, color, engine, year, price is shared.

-In the sales and customer relations department, common fields are: First name, last name, MI, Address, City, state and country.

- There is no relation between inventory and customer relations data unless they are connected through the sales data.

1. **Database schema design**

**Chosen data model: Relational model**

Relational database schema: Consists of a set of 3 tables.

{

**InventoryTable** (VIN, Year\_manufacted, Model, DriveTrain, Color, Door, Engine, MSRP)

**SalesTable** (ID, *cust\_id, VIN*, Year\_bought , SaleDate, Discount, TradeIn, TradeInValue, PurchasePrice, RepeatCustomer)

**CustomerRelationsTable** (cust\_id, Last Name, FirstName, MI, Address, City, State, Country, PostalCode, Profession, FinancialState)

}

**Data types of the attributes:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| String | Numeric | List of strings | Boolean (Also contains NULLs) | Date format |
| VIN, LastName, FirstName, DriveTrain, Color, Door, Discount, TradeInValue, MI, PurchasePrice, City, State, Country | Year\_manufacted, Year\_bought, ID, MSRP (floating number), PostalCode, cust\_id | Model, Engine, Address, Profession, FinancialState | TradeIn  RepeatCustomer | SaleDate |

**Primary and foreign keys:**

Primary key for Inventory table: VIN

Primary key for Sales table: ID (This ID is unique identifier for the sales table)

Foreign key for Sales table: cust\_id, VIN (cust\_id is a unique identifier given to each customer in the database)

Primary key for Customer Relations table: cust\_id

1. **Example table- Please refer to the workbook**
2. **Process for creating the database scheme and tables.**

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

The data model chosen here is a relational model. The database scheme includes a set of 3 tables, one for each department.

1. In an attempt to build an efficient database, the primary and foreign keys were first identified.
   * Primary key for the inventory dataset has been identified as VIN for our convenience (ID can also be chosen)

**Note: Since it is always required to use the combination of first and last names to perform merging of tables, this has been avoided by the introducing a unique identifier called cust\_id to the sales and customer relations table . Here every customer is given a unique cust\_id.** This should not be confused with the ID column in sales table as it will result in duplication (Same customer can have 2 IDs in the sales table if he has bought 2 cars)

* + Hence, the primary key for the sales dataset is ID. Foreign keys for this dataset are VIN and cust\_id
  + Primary key for customer relations is cust\_id.

1. Secondly, all columns that are repetitive are removed and retained in only one of the relational tables. For example, since the columns - model, color were present in both datasets, it has been retained only in the inventory table and has been removed from the sales tables. This is the process of **normalization**, to remove functional dependencies. By doing this we will **reduce the complexity** of updating/adding the same information across multiple tables and also reduce the space taken up by the database.. Further, if we need to obtain any information across different tables, a simple merge with unique identifiers will be sufficient.
2. Further, for some cases the year of manufacturing a car and year of buying a car are different. This is distinguished by assigning different names for the columns in their respective tables. To make the sales table more compact, we could also eliminate the year\_bought column, because, a part of the sale\_date column already contains information about the year in which a particular car was bought.

* **Did you leave out any information? If so, why?**
  + Repeating columns across tables were removed and retained only in one of the tables.
  + In the inventory dataset, there are sub-models provided (such as P100D, 60) for only the “Tesla S models”. This was not considered in building the database, as it is not provided for other models. Since the data is very small, providing this detail for just one car model ‘might’ not be useful for our analysis. In case the dataset grows, and sub-models for more cars is given, it might be useful.
  + Following the same logic as above, Sedan/Hatchback information is provided only for the “Toyota Prius” cars and this has not been considered in our table.
* **Why did you choose certain things as attributes? As keys?**
  + Attributes have been added in their respective tables keeping in mind the information is relevant to the table name. For example, it makes more sense to add customer details such as address, name etc in the customer relations tables than adding it to the sales table.
  + The primary keys are unique identifiers for tables i.e they display a unique combination of a set of values inside the table. The VIN is one such key that can be used in the inventory table. ID which is also already present in the sales table is also a unique identifier. The sales table also contains VIN through which the two tables can be easily merged. Further, for each customer, we create our own cust\_id which serves as a primary key, which when added to both the sales and customer relations table can be used to combine the tables.
* **What were the hardest decisions you had to make in this design process?**

One hard decision that I had to make was whether or not to include the values for car type-Sedan/Hatchback and Model type- P100D,100D in the inventory dataset. Since it is not present for all model types, it was hard to make a call due to the ambiguity that it could be helpful for a future analysis. Finally, I decided to leave it out as the tuples that have the column are very small in number and do not mean much to any analysis with small observations.

* **How does your schema design support data independence?**

The above relational representation supports abstraction by only presenting the intrinsic features of the data and thus causing the interaction with stored data through indirect mapping. Due to the mapping between the physical schema and logical schema, 1) any change in the storage method will not affect the data directly and the subroutines run without an error. 2)new additions to the data will not be interpreted wrongly.

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

The database schema discussed above helps achieve data analysis in a much effective manner due to its standard format throughout all departments. Also, the format can be reused over time with more incoming data.

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

Activities that can enhance or sustain the database for future discovery and use for new purposes: Preservation, Reformatting,

Additional activity that I would recommend would be Integration: For effective analysis, it would be ideal to merge the three available datasets (or more when available), in order to learn more about a customer-car relation.