Vehicle inventory

it appears that this file attempts to store details about vehicles, but it is missing columns heading which makes it harder to mentally scan the data and classify it. Also, since the file is in text format, any null values affect how visually data looks and drastically decreases its readability and quality. I strongly recommend against storing and sharing data in text files unless it is coming directly from a database extract. For column 2, because of the missing column headers I can’t confirm the nature of its data. From what I can tell, it can either be the VIN for the vehicle, or it could be another identification value for the tuple (row). In either case it makes the seed row id in column 1 redundant. In addition to that, it looks like the vehicle’s price is stored as a string where it makes more sense to store it as decimal

Sales

This dataset is more readable because of the columns heading. There are a few missing values, but they easily can be filled out by observing the remaining data. Since FirstName and LastName are spelled in full, there is something to be said about the use of “MI” in column 3 heading instead of the full meaning (i.e.: middle initial). The currency columns (MSRP, TradeIn, and PurchasePrice) better off to be decimal instead of currency value for ease of operations and data manipulation. The discount amount can be calculated but would make it easier if it is spelled out in the dataset. There is one sale without PurchasePrice so not sure what make out of this, did someone get a vehicle without paying for it? There is a ‘yes’ value in the “RepeateCustomer” column but no other required for the same customer? So not sure what “RepeateCustomer” means here. Also all customer data appear to be redundant with what’s in File3 and can be extracted into its own table.

People to Personal

This dataset is hard to read and comprehend which makes the quality of this document poor. The Microsoft word is not the right tool to share database content. Content is not formatted and no columns heading. It appears that each database record is stretched over 4-5 lines in the document. The nature of the data is understandable with little effort, and I can infer the purpose of each column. The data itself is almost all redundant in comparison with the second dataset. The People information (customers data) is duplicated between the two sets and can be easily modeled through Primary key/ Foreign key. The only datum that worth keeping is the service/request.

The process of creating the database schema:

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

I thought about a relational way to represent the datasets I have in the three files. After thinking about for some time, I found that the core two datasets can be ‘vehicles’ and ‘customers’ and then all other data can be linked from them (e.g.: sales, customer service.)

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

I left out the “CustomerRepeated” column as I don’t think is accurate (see page one for a refresher). I also left out a lot of columns from the excel file dataset because it is just redundant with the addition of customer table. I also left out all data from the doc file except for the “service/request” string

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

I have chosen id:integer as a key for all tables except “vehicles” because it represents a unique identification number for each row and each row can be uniquely identified by it.

I have chosen “vin” as the key for “vehicles” table because it’s known that vins are unique and a vehicle can uniquely be identified by a its vin number

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

Deciding what do some values mean in the provided files. Also, it took me some time to analyze the files and decide what tables I want to use to re-represent the data

* How does your schema design support data independence?

My schema allows for views to be created and hide the tables from directly getting accessed by queries. This will protect the applications from breaking when a change to the schema happens.

As for physical independence: It really depends on the db engine but for this schema it shouldn’t matter what indexing algorithm to use or how to re-organize a heap file because these changes should not affect the results of queries but rather affect the performance of them

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

It supports the collection and acquisition of data. It employs an appropriate data model and uses appropriate standards. Support the ability to search for and locate relevant data. Support the ability to identify, authenticate, and validate data

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

Pros: data independent, relational, employs appropriate standards, scalable, searchable

Cons: # of tables

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

I’d recommend the following curation activities:

* Storage
* Compliance
* Sharing
* Reformatting