Assignment 1

**Part 1**

MDS\_Exercise1\_FileA.txt

File A is a text file which is tab-separated and contains the inventory record of each individual vehicle. Each car has a unique identity (VIN) and characterized with details and feature which includes year, make model, trim, drive trans, color, doors, engine, and MSRP. File A overlaps with File B on the vehicle information The overall quality of File A is poor compared with other two files.

There are several issues in File A:

* Missing column headings which make the table hard to identify its attributes.
* Spacing format inconsistency for each line which raises a lot of confusion. Some fields use tab to separate, some fields use two tabs to separate, and some fields just use one space to delimit.
* Some fields of the same column call different names but meaning the same thing, for example, 4WD and AWD
* MSRP column is in quoted string instead of number

.

MDS\_Exercise1\_FileB.csv

File B is a csv file which contains customer information, car information of corresponding customer, and purchase information. For customer information, it includes customer’s name and address. For car information, it includes the Model, Year, Color, Engine, VIN, and MSRP. For purchase information, it includes customer ID, Sale Date, Discount, Trade-In, Trade-In Value, Purchase Price, and Repeat Customer. The overall quality is the best compare to other two files. Formatting is good, but it contains some missing data.

The following is the summary of missing information:

* Hermione Granger’s address lacks city information
* Luna Lovegood and Albus Dumbledore’s address lacks country information
* Luna Lovegood’s car lacks MSRP
* Draco Malfoy’s car lacks Purchase Price

For those missing values, some of those can be found in other files. Address-related information can be found in File C while MSRP can be found in File A. Purchase price cannot be found.

Vehicle information can be found in both File B and File A. Personal information is overlapped in both File B and File C.

MDS\_Exercise1\_FileC.docx

File C is a Microsoft Word file which contains personal and financial information such as name, address, occupation, and loan status. Personal information of File B and File C are overlapped. The overall quality of File C is better than File A but worse than File B.

File C has some issue:

* Format issue: Instead of table format, it is like business card format scrolling down for each person
* Attributes names and tuple names are also missing. The information of each customer is placed vertically.
* Some customers contains loan status, but some do not.

**Part 4**

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

Since some information is overlapping among File A, File B and File C, I use the information of File B to create a Transaction Table and Inventory Table. Then, I use customer information to create Customer Table. I create customer ID in both Transaction Table and Customer Table. Therefore, sales information can be found in Transaction Table while customers information can be found in Customer Table. Transaction Table and Inventory Table contains VIN. Hence, VIN is used for identifying car information in Inventory Table and sales information in Transaction Table.

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

No, I did not leave out any information provided except duplicate data. All information is included for future usage.

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

Vehicle’s VIN is unique for each vehicle, so I used it as the primary key for the inventory table and transaction table. I create customer ID as primary key for searching customer information in customer table and searching sales information in transaction table.

1. What is the hardest decisions you had to make in this design process?

The hardest decision is to determine how to group the data in one table and how many tables I should create in well organized sense. For example, repeated customers can only be found once in Customer Table but can be found many times in Transaction Table. In order to simplify searching, I create customer ID to represent each customer. With customer ID, customer information and transaction information are separated in two tables. With VIN, car information and transaction information are separated. Therefore, I decide to use three tables to represent all data.

1. How does your schema design support data independence?

Using my schema design to store data in the database, the input format of customer, vehicle, and transaction will be consistent and spacing issue mentioned in part 1 will be resolved. Therefore, physical data independence is achieved.

Adding additional information in customer table such as email address or adding used car information in inventory table will not affect information in the transaction table. Therefore, logical data independence is achieved.

1. How may your schema design support the overarching goals of data curation?

Management of data would be easier which provides more storage space, effective searching, easy modification and better platform for data analysis.

1. What are the pros and cons of your schema design?

Pro:

* Clear relation and representation
* No duplicate data
* Support data independence
* Data locality. Data is separated for customers, vehicle and transaction records.

Con:

* If the scale goes up, tables may be gone too long. It can be solved by creating smaller table inside each major table.

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

For future discovery, modification and provenance would enhance and sustain the database. Records can be modified and identified for better use. I would like to recommend adding sharing activities. Most of these data are used in auto dealer, sharing with marketing will make better promotions.