**Assignment 1. Relational Schema Design**

**1. Input Data Review and Preprocessing**

The information in these three files was recorded by each department in very different formats which cannot be used effectively in one database; format conversion is needed.

Information is duplicated among the files, e.g. File B contains all the information from File A except for drive and body type; Make and Model are separate columns in File A and one column in File B. File B also contains all the information from File C except for the Zip Code, Occupation, and Needs Financing columns from File C. Overall, File A and C have the same number of columns and rows while their number is significantly greater in File B. The number of rows is the same in all three files.

For ease of presentation, the original data converted to a tabular format is shown on the Original Data tab of the Workbook. More detailed information about the data quality, completeness, and steps needed to modify the information before it can be imported into a database is presented below. Preprocessed input data ready for the final analysis before the import into a database is provided on the Changes Documented tab of the Workbook. This tab is used to document all the changes the data underwent before it was imported into the database and further optimized there.

**File A Inventory**

a) Quality of Data

* The data is about car inventory; it is provided in the txt format which is not portable. “Columns” have no headings (you have to guess what they are). There is a different number of “words” (sets of characters divided by spaces) in similar elements of different records (e.g. “Prius”, “Model S”, and “Expedition King Ranch” for model, or “Gold” and “White (Pearl)” for color) which makes it hard for automatic parsing and conversion into a table format. I had to copy the content to MS Word, insert commas manually, and then use the *Convert Text to Table (with commas)* feature of MS Word to get a table.
* The Body column contains incomplete or incorrect information about 6 cars and more complete about 4 cars (sedan, hatchback). This can be easily corrected as Tesla Model S is a 5-door liftback (not 4-door as mentioned in File A) and the Fords from File A are SUVs. The only unclear detail is that sometimes SUVs are classified as 3-door SUV or 5-dore SUV. This issue needs clarification from the Inventory Department. So far, I left it as 4-door SUV pending clarification.
* The Drive column needs clarification from the Inventory Department in regard to the possibility of showing Tesla (now AWD) and Ford (4WD) as one value of 4x4 instead of two (AWD and 4WD). It doesn’t look that this dealership sells trucks with more than two axles.
* The VIN column seems to be a problem: with a few exceptions, VINs are not 17 characters long (as required according to Carfax.com and Wikipedia), and several different lengths are used (to say nothing about the fact that all the VINs are fake, but that was done for the purpose of the exercise, I guess).
* The Inventory does not indicate if a certain car is still part of the inventory or is sold. In fact, all of the cars in this list are sold according to File B except, maybe, for File A record ID#1 whose status is somewhat unknown as there is no Purchase Price in File B.
* Otherwise, the records seem to be complete without missing data.

b) Documenting Changes Before Import to Database

* Added meaningful column headings based on File B.
* Added the Sold column.
* Completed the Body column

**File B Sales**

a) Quality of Data

The data is about car sales; it is provided in the portable csv format (comma-delimited data) which can be opened in MS Excel in a tabular format. These are the same cars mentioned as inventory in File A. Some data is incomplete, namely:

* Incomplete **record ID#3**: price missing. Needs clarification (deal is not closed or the data was not properly recorded?)
* Incomplete **record ID#5:** this is Ford Explorer Eddie Bauer Edition according to the File A; can be easily corrected.
* Incomplete **record ID#6**: a) State missing; both Bloomington IL and Bloomington IN are close enough to be possible cities of residence; this can be easily corrected as according to File C the state should be IL; b) Information on the discount OR trade in is missing. Seems to be a sloppy record. Needs clarification.
* Potentially wrong **record ID#7**: a July 2015 sales date for a 2016 model (too early for next year’s models?). May need clarification
* Incomplete **records ID#9 and ID#10**: Country missing; can be easily corrected as both places seem to be and actually are (verified) in Illinois, USA
* Zip codes are missing in addresses (can be found in File C)
* Same problem with the VIN column as in File A.

General observations:

* it is not clear what happens with trade-in cars: they don’t seem to become part of the inventory (needs to be clarified with the department);
* white is the most preferred car color;
* Harry Potter lives in Chicago IL; I didn’t know that.

b) Documenting Changes Before Import to Database

* Records ID#5, 9, and 10 were corrected.
* For simplicity purposes, the MI column was combined with the First Name column. I can’t think of a situation when someone would need a search based on middle names.
* The excessive Repeat Customer column is removed: data is already shown in the Discount column (you can run a search by the Discount column in order to find all Repeat Customers)
* The excessive Trade In column is removed: all data is shown in the Trade In Value column; NULL means no trade in, and a special code, e.g. 1111, can mean there was a trade in, but the exact value is unknown. This will work to get meaningful search results.
* Having four columns instead of one for address may be excessive from the standpoint of reserved space on disk, but I left them because one may need to run a specific search by City, State, or Country.
* The Model column was split into two column: Make and Model (see File A Inventory).

**File C Customer Relations**

a) Quality of Data

* The data is about customers and is placed in a .docx file formatted in such a way that it needs manual manipulations to convert it into a tabular format for import into a database. I had to insert commas manually and then use the *Convert Text to Table (with commas)* feature of MS Word to get a table.
* These are exactly the same customers who purchased vehicles according to File B.
* There are no “column” heading; I came up with them based on the other two files.
* There is no ID column.
* Harry Potter teaches at UIC; I didn’t know that.

b) Documenting Changes Before Import to Database

* Added the ID column.
* Added meaningful column headings.
* Tried to unify the values in the Needs Financing column.

**4. Schema and Tables**

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

All possible data modification / corrections mentioned above in Section 1 were introduced in order to make the data as consistent and complete as possible. Different database tables were updated based on these modifications, based on each other and common sense information for all the issues mentioned above in Section 1.

The currency format of Excel cells showing prices/discounts was changed to number with two decimal places (float) in the database for simplicity. In Table 1, attribute *Year* is of integer type and *Sold* is of Boolean type (it seemed logical to me). All attributes with a limited range of values were assigned the type *Selection from a list of strings*. All other attributes are of string type which is logical, including *Zip Code* because zip codes can contain letters in foreign countries (and we have attribute *Country*).

All the primary keys for the three tables are cross-referenced in Table 2; however, depending on the specific usage requirements for the database, it is possible to introduce foreign keys in Table 3 and Table 1 to relate them to each other and to Table 2.

Open issues mentioned above in Section 1 still need clarifications from respective departments.

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

Below is a more detailed description of all the changes made in each table (both deletions and additions):

* **Table 1** is modified File A: all attribute names were added at the preprocessing stage above; attribute *Car* *ID* was modified to show a unique car ID code (data type = string) – this is the primary key; attribute *Sold* (type = Boolean) is added to facilitate search for cars that are sold or that are still part of the inventory; attributes are rearranged for a more logical representation of data.
* **Table 2** is based on File B: attribute names were provided as part of File B or modified based on common sense; attribute *Transaction* *ID* was modified to show a unique transaction ID code (data type = string) – this is the primary key. The following data optimization was implemented: File B columns *Repeat Customer* and *Trade In* were removed as containing duplicate information from File B columns *Discount* and *Trade-In Value*, respectively (see the preprocessing stage description); File B columns *Model*, *Year*, *Color*, *Engine*, *VIN*, *MSRP* were all replaced with one attribute *Car* (type = string) which is a foreign key associated with Table 1 because this information is already provided in Table 1 and belongs there logically; File B columns *LastName*, *FirstName*, *MI*, *Address*, *City*, *State*, *Country* were all replaced with one attribute *Customer* (type = string) which is a foreign key associated with Table 3 because this information is already provided in Table 3 and belongs there logically.
* **Table 3** is modified File C: all attribute names were added at the preprocessing stage above; attribute *Customer* *ID* was added to show a unique customer ID code (data type = string) – this is the primary key; data entries for attribute *Needs Financing* were unified at the preprocessing stage above;

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

To optimize the data in order to avoid duplication of data (and related mistakes: see Wikipedia article again! + RM section in book).

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

The hardest step for me during the design process was to determine how to distribute attributes among the three relations and optimize them to avoid data redundance and related issues – Wiki article), and also to clean File B as it is big and contains a lot of inconsistencies and duplicate information.

*How does your schema design support data independence?*

My schema supports data independence by the fact that now it is possible to add new data constructs without having an impact on programs. Also, it is possible to change storage methods without such impact. Operations on data can now be defined formally. All interactions with the data are conducted in terms of tables (containing attributes and values); subsequently, this can be translated into data storage instructions.

The original input data was taken from different sources and was provided in very different formats. Now it is integrated and can be used more easily by all data sources (departments). This definitely makes it easier to check data for validity and quality and improves the overall efficiency of the company’s operations though a unified standard approach to handling data.

*How may your schema design support the overarching goals of data curation*

Since data curation is concerned with all aspects of the management of data, my schema design directly supports data curation by making this management easier and more effective on the bases of the qualities described in the previous two paragraphs. As far as I understand the areas of curatorial activities, my schema supports

1) Collection as once you have an appropriate data model, you have a better idea about what data should be collected and how it should be presented;

2) Organization through an appropriate data model and standards;

3) Storage because the data has physical independence making storage more reliable;

4) Preservation as my schema ensures the understanding of the semantics of data elements and makes it easier to audit it;

5) Ease of Access and retrieval through appropriate data organization;

6) Identification because the schema contains identifiers and facilitates in developing methods of identification;

7) Integration through schema alignment by accommodating data from multiple sources;

8) Reproducibility though documenting data processing and analysis;

9) Sharing though a uniform format;

10) Communication because well-organized and effectively integrated data serves as a good basis for analysis and insights;

11) Provenance through the identification of inputs, modifications, and actions responsible for data values;

12) Modification, Compliance, and Workflow indirectly through a simpler structure of the database based on an appropriate schema which makes it easier to understand and analyze data and take appropriate actions.

*Which curation activities could enhance or sustain the database for future discovery and use for new purposes?*

The curation activities that could enhance or sustain the database for future discovery and use for new purposes may include:

* Preservation;
* Discoverability;
* Access;
* Workflow;
* Identification;
* further Integration (in case of new data and data types);
* Reproducibility;
* Provenance (in case of new data and data types);
* Modification and management of change;
* Security.

*What additional activities would you recommend?*

In the future, I would recommend to implement:

1. Data validation - clarify all the open issues mentioned in Section 1 hereof in order to make the data more understandable and reliable;
2. Access/revision control to ensure that data can be changed only by authorized employees and that the history of such changes is recorded in the database.