ECE 356 Project

Used-Car Sales Data

This document is a description of details relevant to those doing the course project using the **used-car-sales** dataset. It provides details about the dataset, as well as suggestions pertaining to the client application, the entity-relationship design, and the data-mining exercise.

Data Source

The main source for used-car sales data comes from a web-scraping of Car-Guru data, and is described in:

`https://www.kaggle.com/ananaymital/us-used-cars-dataset`

This is a single CSV with approximately three-million records of used-car sales listings, with 66 distinct attributes. The file "`used_cars_data.csv`" is available on marmoset04 in "`/var/lib/mysql-files/01-Cars/`". In addition, the Craig's List cars and trucks data comes from:

`https://www.kaggle.com/austinreese/craigslist-carstrucks-data`

This is a single CSV will a little less than half-a-million records of used-car sales listings with 26 distinct attributes. The file "`vehicles.csv`" is likewise available on marmoset04 in the same `01-Cars` directory.

You may also find that you want to incorporate address data, particularly if you are looking at data-mining issues pertaining to address correlations with car-sales data. The open-address datasets are available from

`https://www.kaggle.com/search?q=openaddresses+in%3Adatasets`

and on marmoset04 in "`/var/lib/mysql-files/USAddresses/`" with separate CSV files for each state, by two-character state code. For example, Texas addresses are contained within the `tx.csv` file, California ones in `ca.csv`, *etc.*

In all instances, you should look at any of the relevant datasets on Kaggle to determine what the different attributes are within the CSV.

Client Application

As noted in the main project document, there will be little additional to add to the generic client-application requirements listed there. If you want a sense of what a client application for the used-car sales data should do, you are advised to look at a website such as autotrader or kijiji autos. When looking at such sites, remember that you should should focus on determining functionality, not interface. Essentially, though, a used-car client application should allow a customer to search for used cars on some reasonable basis that a person

looking for a used car would want: by year, by make and/or model, by price, *etc.* Likewise, a person should be able to list a car for sale, modify the listing to change the price and/or add additional information, and remove the listing once the car is sold.

Entity-Relationship Design

Per the main project document, you will need to determine an appropriate ER design for your dataset. Prior projects done using this data have identified over a dozen appropriate entity sets, as well as associated relationship sets connecting various of those entity sets together. If you have difficulty in thinking about different relevant entity sets for this domain you should consult with your designated instruction-team member.

Data-Mining Investigation

For used-car sales data there are numerous possible data mining-exercises that are worth considering. For example, determining a classification tree for length of time listed would be extremely valuable to used-car dealerships. This can be best worded as the question, "what factors enable rapid sale of a vehicle *vs.* it taking a long time to sell?" Other questions for which a classification approach would be relevant include, "what factors determine the price of a used vehicle?" and, "what make and/or model sell most in a given geography?" That second question also points to a clustering approach that could be used.

If you have difficulty thinking about an appropriate data-mining exercise, you should consult with your designated instruction-team member. If you think you have a good idea for a data-mining exercise, it is probably worthwhile checking with your designated instruction-team member to confirm that it is of appropriate scope.