**Lab 1 – Overview of Data**

**CIS9660 – Data Mining for Business Analytics**

**Datasets required:** *ToyotaCorolla.csv, RidingMowers.csv, BostonHousing.csv, ApplianceShipments.csv*

**Deliverable:** Submit python notebook with answers to Blackboard (Submissions). Format: last-name\_first-name\_lab1.ipynb

**Problem 1. [1 point]** Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

**1.a.** Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).

**1.b.** In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.

**Problem 2. (Book pg. 58) [2 points] [Dataset: ToyotaCorolla.csv]**

The dataset *ToyotaCorolla.csv* contains data on used cars on sale during the late summer of 2004 in the Netherlands. Variables relate to characteristics of the used vehicles including *Price, Age, Kilometers, HP,* and other specifications.

2.1 [1 point]

**2.1.a. data dimension**. How many instances and features (attributes) does it have? (use shape on the data frame)

**2.1.b. review first few records.** Use the head method to retrieve the first 5 records.

**2.1.c. output attribute names.** Output the attribute names using the columns property.

**2.1.d. The most kilometers.** What car model is the one with the highest mileage?

**2.1.e. Oldies.** How many cars have more than 200,000 kilometers?

**2.2. Box plots [0.5 points].** Create a box plot with the mileage by fuel type. Is there any interesting insight/pattern?

**2.3 Histogram [0.5 points].** Plot a histogram of the variableprice, what is the distribution like?

**Problem 3. [1 point] Sales of Riding Mowers: Scatter Plots.[Dataset: RidingMowers.csv]**

A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in $ 1000s) and Lot Size (in 1000 ft2). The marketing expert looked at a random sample of 24 households, given in the file \_RidingMowers.csv.

**3.a.** Using Python, create a scatter plot of Lot Size vs. Income

**3.b.** Color-code by the outcome variable owner/nonowner. Make sure to obtain a well-formatted plot (create legible labels and a legend, etc.).

**Problem 4. [1 point] Shipments of Household Appliances: Line Graphs.**

The file \_ApplianceShipments.csv\_ contains the series of quarterly shipments (in millions of dollars) of US household appliances between 1985 and 1989.

**3.1.a.** Create a well-formatted time plot of the data using Python. [1 point]

**3.1.b.** Does there appear to be a quarterly pattern? For a closer view of the patterns, zoom in to the range of 3500–5000 on the y-axis.