Data Mining Tools and Techniques
Homework 1 – Analyzing Data and Posing Questions

Krishna Jayani
University of Houston – Clear Lake
2700 Bay Area Blvd.
Houston, Texas 77058
Jayanik1647@uhcl.edu

1. Data Set 1: CarDealershipSales

CarDealershipSales data describes the details of Automotive New Top 150 Dealership Groups (2020). The original data set has 150 rows. Each row corresponds to details of dealership group. This includes 2020 rank (Ordinal), Dealership group name, DealershipRating (Nominal), Address, Phone, Total new retail units (Ordinal), Total used units (Interval or higher), Total number of dealerships (Interval or higher), Group revenue all departments (Internal or higher), 2019 rank (Ordinal).

A) *The descriptive statistics are divided into two categories (Numeric and Non-numeric). According to the data set 1, Non-numeric descriptive statistics are Dealership group name, Dealership Rating, Address and numeric descriptive statistics are 2020 rank, Phone, Total new retail units, Total used units, Total number of dealerships, Group revenue all departments, 2019 rank. The table show the numeric descriptive statistics includes minimum, maximum and average.*

Name	*Minimum*	*Maximum*	*Average*
2020 rank	*1*	*150*	*75.5*
Total new retail units	*4,476*	*33,622*	*23,903*
Total used units	*1,910*	*225,713*	*17,651*
Group revenue all departments	*254,198,798*	*21,609,000,000*	*1,521,893,865*
2019 rank	*1*	*150*	*71.90977*

B) *To identify the attribute whether it is descriptive, discrete, continuous or discontinuous, it depends on the measures of each metric. Descriptive data may include Dealership group name and Address. Discrete data may include 2020 rank, Phone and 2019 rank. Continuous data may include Dealership rating, Total new retail units, Total used units, Total number of dealerships, Group revenue all departments. Discontinuous data may not include any attribute. The reason behind is that data has not values which have more than one gap at different intervals.*

C) *The data has supervised data set. Based on total new retail units, data showed the rank 2020. This means that one attribute serves as the output variable, dependent variable or class variable. The attribute 2020 rank is class variable.*

D) *The data has sequence data set. Due to 2020 rank, the data arranged in sequentially.*

E) *The dependent variable named total new retail units and measurement scale may be ordinal.*

2. Data Set 2: USA Traffic Accident

USA Traffic accidents datasets shows the accidents occurs in 50 states of the USA. There are many tens of millions of records. Each row represents the accidents that includes AccidentID, OfficerID, Date, Time, Accident Location, attributes of two vehicles as VIN, Make, Model, Year, Odometer.

A) *The datasets has the AccidentID and OfficerID. Then the total accidents for the given date and time would be the class variable and become supervised type of problem. After adding the accidents to a weekly(daily, monthly) basis it now considered as the time-series regression type of problem. The dependent variable would be the date, time and it is nominal. By aggregating the number of accidents for each day and time, it would be possible to raise regression type of question,*

Were there more accidents during the full-moon or during the day-light in 2018?

If we examined the dataset in terms of accident location, then there is no dependent variable and this would be an unsupervised type of problem. It becomes clustering type of problem, as we classified into 3 different location type. Clustering the accident based on the location would create the following question,

How many accidents occurs in the High Frequency Accident Location, Moderate Frequency Accident Location and Low Frequency Accident Location?

Did weather influenced the number of accident based on the Latitude and Longitude?

If the analysis groups the data based on the predefined categorial class, which is driver and vehicle characteristics. This would be the classification type of problem and class variable would be the vehicle characteristics, accident and driver's age.

We might ask, what would be the accident rate if the driver's range is between 40-50 years old and vehicle's Odometer is 10k?

What is the risk(low/high) of accident if vehicle's odometer is 20k and age of both drivers?

B) *Additional questions would be,*
Weather Data: What is the number of accidents based on snow-day versus non-snow days.
What is the number of accidents based on inclement weather (snow, or rain) versus sunny or cloudy.

Data Set 3: Volcano Eruption Data

Volcano Eruption data describes the data of National Centers for Environmental Information. The data set has 22 attributes and 838 rows. That includes Searching passing, Death Description, Missing Description, Damage Description. Houses Destroyed, House Destroyed Description, Total Death, Total Description.

A) *The datasets has Country and Location. If we adding the Elevation attribute to this, then class variable would be Country, Location and becomes supervised type of problem. The dependent variable would be elevation and it is ordinal. Classification type question raised, it would be*

If the elevation would be greater that 200, which type is considered?

If we aggregate Name, Location and Elevation, then Elevation attribute would be class variable and become supervised type problem. After adding latitude and longitude to the dataset, the data would *become spatial data. Dependent variable would be latitude and it is ordinal, so regression type question raised at this time,*

Is that possible that elevation related to the both latitude and longitude?

If dataset examined in terms of country, latitude and longitude, there would be no dependent variable. This would be an unsupervised type of problem and spatial data. If we classified different type of volcano, then clustering type of problem would be raised,

How many times volcano occurs in particular latitude and longitude?

Is the type of volcano depending on location of country?

B) *The following are the questions raised after adding more datasets.*

Temperature data. What is the highest and lowest temperature of country face the volcano eruption?

Season data: Which season has highest number of volcanos occurred?

Ozone Layer data: Is the gap of ozone affected and volcano eruption increases?

Data Set 4: AirBNB

The datasets contains 41 attributes(it has many rows). It is possible to create various type of problem from this dataset.

A) *We have most important features that influence price are host_name, host_since, maximum_nights, security_deposite, accommodation, host_identity. If these attributes are adding, then we can predict the total host listing which is numerical. It would be regression type of problem, and would raise the question as,*

How the low prices distinguish from those that have high prices of the houses?

If an analysis groups the data by listing_name, host_name, price, property_type, room_type and also we can rate on the basis on the location, cleanliness, location score and host of other parameter, then those attributes would be the class variables. It now becomes the clustering type of problem and would give the how the listing are distributed across neighborhood. Clustering the maximum/minimum listing with rating would leads to the question:

Which area is the best based on the maximum listing and rating in Texas?

If we examined the relationship between property type, neighborhood and zip code, we can classified different boroughs constitute of different rental types. Which has more than 20 types, become class variable. This would raise the classification type of problem as,

What type of listing has the highest and how much percentage of the property in that neighborhood?

B) *From the datasets, additional questions would be created as*
How do prices of listing vary by location? What localities in Texas are rated highly by guests?
Are does the demand and prices of the rentals correlated?
What are the different type of properties in Texas? Do they vary by neighborhood?

3. CONCLUSION

Based on the analysis of different datasets we did: We could see that it takes a very thorough consideration to classify the data, most cases there is no set classes. By analyzing incomplete data sets may give us faulty results and they are not useful. Based on the given data sets that we need to ask to help us analyze the data. Knowledge of the domain(requirement gathering) like where the data is coming from is important to analysis the data. Before training the model, we need to collect appropriate information from the different sources for analyzing the data sets. We had issue analyzing the USA Traffic data since we had *little domain knowledge. Although the assignment gave us very good experience to analyze different data sets and get to know what issues we may face.*