**Database Problem Set**

**Nicholas Wilkinson**

The database I used is based on the attributes of a car evaluation data set designed for machine learning. Because the data just contains the different combinations for the data, I decided to create a simple Java program, which I titled ‘DatabaseCreation’, to create a database with the same attributes but slightly modified.

Each row is a car. The first column, ‘Buy Cost’, is the cost of the car, with 1 meaning low cost, 2 meaning medium cost, 3 meaning high cost, and 4 meaning heavy cost. The maintenance cost is based on the purchase cost, with low price cars being unable to be low maintenance, and high price cars being unable to be high maintenance. This is because higher cost cars would have better components and be better put together. ‘Num Doors’ is the number of doors the vehicle has, which is either 2 or 4. Hatch doors are not considered with this count. ‘Num Passengers’ is the number of passengers, ranging from 2 to 8. The number of passengers is based on the number of doors, with 2 door cars having 2 to 3 passengers, and 4 door cars having 4 to 8 passengers. ‘Luggage Capacity’ is the size of the luggage, with 0 meaning no luggage, 1 meaning small luggage, 2 meaning medium luggage, and 3 meaning large luggage. ‘Safety’ is how safe the car is, with 1 meaning low safety, 2 meaning medium safety, and 3 meaning high safety. Safety is based on the cost of the car, with high and very high cost cars having medium or high safety, and low and medium cars having low or medium safety. This is because more expensive cars have better safety features.

Since the data values are randomly generated, outside of conditional properties, all values are equally likely when the car is created. For instance, when a car is made, it is just as likely for it to have a low buy cost as it is to have a very high buy cost. However, if the car has a high buy cost, it is impossible for the car to have a very high maintenance cost, but is equally likely to have a low, medium, or high maintenance cost.

Despite the fact that I can calculate the actual probabilities, having created the backend for the database creation, I will use the data generated to find the probabilities. However, I will use the actual probabilities to double check my work. I will also be using the functions of Excel to assist with calculations like the mean and conditional probabilities.

* **1.3 Question:**

Find the expect, variance, and standard deviation of all attributes.

B = Buy Cost, M = Maintenance Cost, D = Num Doors, P = Num Passengers, L = Luggage Capacity, S = Safety

* **2.3 Question:**

What is the count of cars that are very high buy and very high maintenance? Low buy or low maintenance?

Very High Buy Very High Maintenance = **1265**

Low Buy Low Maintenance = **4892**

* **2.4 Question:**

If a car is randomly selected, what is the probability that a car has 4 doors?

**0.505**

* **2.5 Question:**

If 2 cars selected randomly, what is the probability both have high safety?

**0.063**

* **2.6 Question:**

How many combinations of attributes are there for a car, according to this data?

**1728**

* **2.7 Question:**

Given a car has a high buy cost, what is the probability it has a high maintenance cost? Are the two attributes independent?

**0.086**

**Not Independent**

* **2.8 Question:**

What is the probability that a car has a large luggage and has high safety?

**0.063**

* **2.10 Question:**

Using the Theorem of Total Probability and Bayes Theorem, find the probability for each value of maintenance, given that the car has a very high buy cost.

* **3.2 Question:**

Find the probability distribution for all values for the number of passengers.

* **3.3 Question:**

If 496 cars from the list are chosen, what are the expected values for all of the attributes?

Buy Cost: Low = 122, Medium = 121, High = 127, Very High = 126

Maintenance Cost: Low = 80, Medium = 166, High = 165, Very High = 85

Num Doors: 2 = 246, 4 = 250

Num Passengers: 2 = 121, 3 = 124, 4 = 50, 5 = 50, 6 = 50, 7 = 50, 8 = 51

Luggage Capacity: None = 127, Small = 123, Medium = 121, Large = 125

Safety: Low = 123, Medium = 249, High = 124

* **3.4 Question:**

If 10 cars are randomly picked, what is the probability 5 have large luggage boots? Use Binomial Probability.

**0.059**

* **3.5 Question:**

If cars are randomly picked, what is the probability the first car with a small luggage boot is the fourth car? Use Geometric Probability.

**0.105**

* **3.7 Question:**

If 100 cars are randomly picked, what is the probability 30 of them can seat 4 people? Use Hypergeometric Probability.

*\*My calculator cannot handle the large values. Below is the equation to answer the problem.\**

* **3.8 Question:**

From the data, it can be seen that for every 100 cars, roughly 17 have very high maintenance costs. If 100 cars are picked at random, what is the chance that 10 have a very high maintenance cost? 25? Use Poisson Probability.

* **3.11 Question:**

Using the mean and standard deviation calculated before, what is the probability that a car can house a number of passengers within 1.2 standard deviations? Use Tchebysheff’s Theorem.

**0.306**

* **4.2 Question:**

Create a rough distribution and density function for buy cost, number of doors, and luggage capacity.

**Buy Cost:**

**Num Doors:**

**Luggage Capacity:**

* **4.3 Question:**

Find mean for the above distribution functions

**Buy Cost:**

**Num Doors:**

**Luggage Capacity:**

* **4.4 Question:**

Assume that the buy cost of 30 cars at a dealership is uniformly distributed. If a customer comes in not knowing what car they want, what is the probability that the dealer picks a car that is between medium and high price?

**1/4**

* **5.2 Question:**

Make a joint probability function for buy and maintenance cost.

B = Buy Cost, M = Maintenance Cost

* **5.3 Question:**

Find the marginal probability functions for buy and maintenance cost.

* **5.4 Question:**

Are buy and maintenance costs dependent on each other?

**Dependent**