Homework 1

Group 3
Nathalia Negri
Troy Yang

305-318-3446
781-812-4931

negri.n@husky.neu.edu
yang.tr@husky.neu.edu

Percentage of Effort Contributed by Student 1:________50%________

Percentage of Effort Contributed by Student 2:________50%________

Signature of Student 1:____________ Negri ________________

Signature of Student 2:____________ ___________________

Submission Date:________May 22, 2017__________

Problem 2.3 10 points (see 2.3_GenrmanCredit.xlsx for full data set)
The data does not look like it was randomly sampled as you can see that every 8th case is part of the sample, therefore, this data might not be useful. This is a systemic sample, which could technically provide a random sample, however, after reviewing the full data set, we find that the 15 records only come from the first 11% of the full data set.

Problem 2.5 10 points
When a model is fit to training data, zero error with those data is not necessarily good because it insinuates overfitting of the model. This means that it is such a perfect fit to the training data that the model is unlikely to be accurate, or even useful, when performing with future data.

Problem 2.7 15 points
1000 records
50 variables
5% values missing

Probability of having a value in a cell = 0.95
(0.95)^50 = 0.077
1 – 0.077 = 0.923
0.923 x 1000 = 923 records are removed

Problem 2.8 15 points
Mean Age = 45 (rounded to nearest integer)
Mean Income = $98,667
Stdev Age = 15
Stdev Income = $62,867

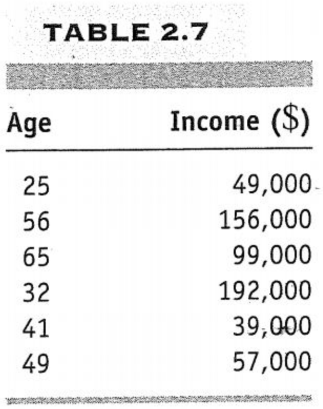


TABLE 2.7

Age	Income ($)
25	49,000
56	156,000
65	99,000
32	192,000
41	39,000
49	57,000

Normalized Age values:
(25 – 45)/15 = -1.33
(56 – 45)/15 = 0.73
(65 – 45)/15 = 1.33
(32 – 45)/15 = -0.87
(41 – 45)/15 = -0.27
(49 – 45)/15 = 0.27

Normalized Income values:
(49000 – 98667)/62867 = -0.79
(156000 – 98667)/ 62867 = 0.91
(99000 – 98667)/ 62867 = 0.005
(192000 – 98667)/ 62867 = 1.48
(39000 – 98667)/ 62867 = -0.95
(57000 – 98667)/ 62867 = -0.66

Problem 2.9 15 points

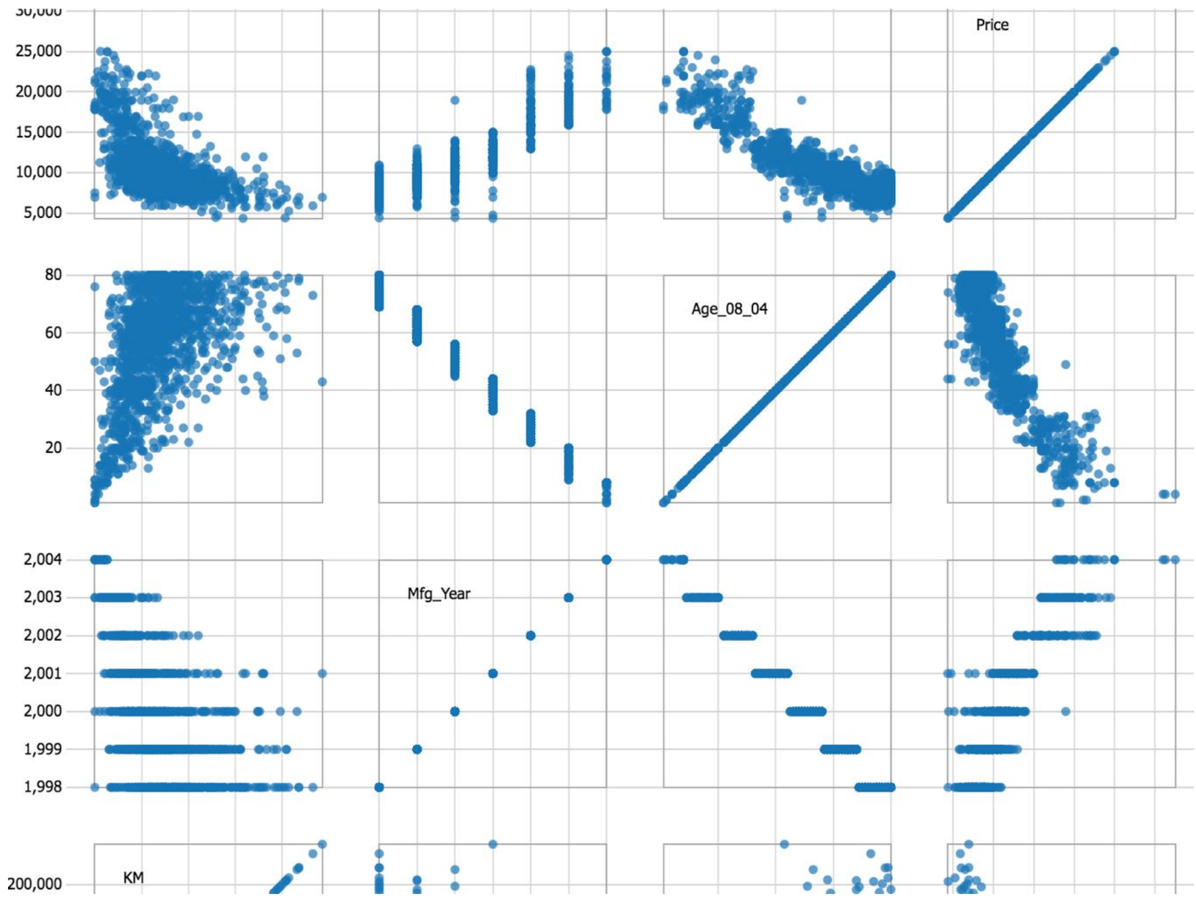
Yes, the largest Euclidian distance was 153,000 between age 56 and age 49. After normalizing the data, the Euclidian distance between these two cases becomes 2.51, which is not the largest out of the normalized data. The largest Euclidian distance from the normalized data is 2.65 between age 32 and age 65.

Problem 2.10 15 points

Model B. The training data is only used to build the model, the validation data is used to evaluate the model's performance. Therefore, it is more important for the model to be accurate using validation data.

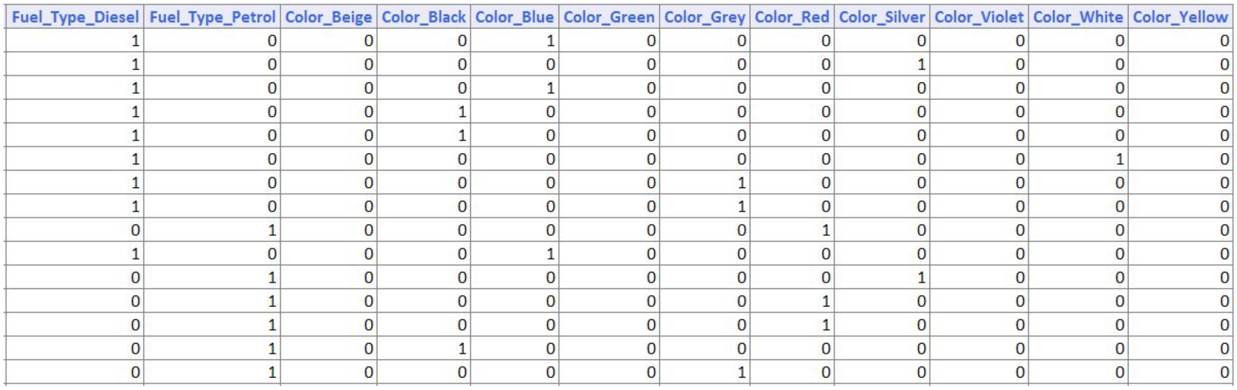
Problem 2.11 20 points (see 2.11_ToyotaCorolla.xlsx for full data set)

a. Multiple pairs among the variables seem to be correlated, here are a few (as there are a significant amount of relationships, I only listed a few):
   Price & Age, Price & Mfg_Yr, Mfg_Yr & Age, Price & KM



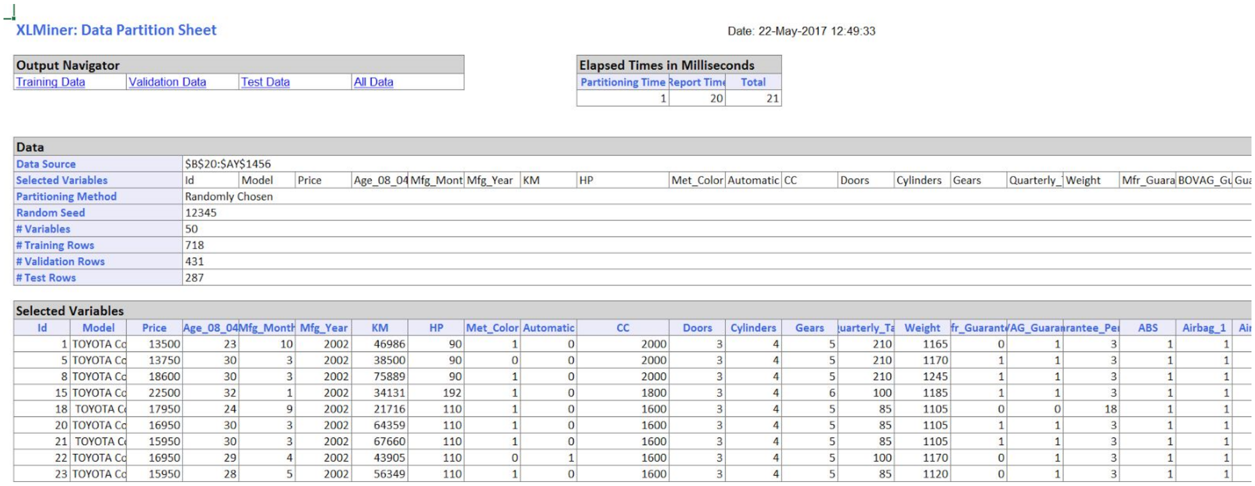
b. Preparation of data:

  i. There are 3 types of fuel in the Fuel Type column: Diesel, Petrol, and CNG. To prepare the data we would create a column for Diesel and Petrol, which would have binary values: 1 if true, 0 if false. If both Diesel and Petrol are 0, this means CNG is true. The same process would be done for the variable Color, where each color has its own column with binary value.



Fuel_Type_Diesel	Fuel_Type_Petrol	Color_Beige	Color_Black	Color_Blue	Color_Green	Color_Grey	Color_Red	Color_Silver	Color_Violet	Color_White	Color_Yellow
1	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0
0	1	0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0	0	0

  ii. The training data partition is what would be used to train the model(s). If there are multiple models in contention, the validation data is what would be used to determine which model to proceed with. Finally, once the model is selected, the test data is what would be used to determine the accuracy this model.



XLMiner: Data Partition Sheet

Date: 22-May-2017 12:49:33

Output Navigator			
Training Data	Validation Data	Test Data	All Data

Elapsed Times in Milliseconds		
Partitioning Time	Report Time	Total
1	20	21

Data	
Data Source	$B$20:$AY$1456
Selected Variables	Id Model Price Age_08_04 Mfg_Mont Mfg_Year KM HP Met_Color Automatic CC Doors Cylinders Gears Quarterly_ Weight Mfr_Guara BOVAG_Gu Gua
Partitioning Method	Randomly Chosen
Random Seed	12345
# Variables	50
# Training Rows	718
# Validation Rows	431
# Test Rows	287

Selected Variables

Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	HP	Met_Color	Automatic	CC	Doors	Cylinders	Gears	uarterly_Ta	Weight	fr_Guarant	AG_Guaran	rantee_Pe	ABS	Airbag_1	Ai
1	TOYOTA Co	13500	23	10	2002	46986	90	1	0	2000	3	4	5	210	1165	0	1	3	1	1	
5	TOYOTA Co	13750	30	3	2002	38500	90	0	0	2000	3	4	5	210	1170	1	1	3	1	1	
8	TOYOTA Co	18600	30	3	2002	75889	90	1	0	2000	3	4	5	210	1245	1	1	3	1	1	
15	TOYOTA Co	22500	32	1	2002	34131	192	1	0	1800	3	4	6	100	1185	1	1	3	1	1	
18	TOYOTA C	17950	24	9	2002	21716	110	1	0	1600	3	4	5	85	1105	0	0	18	1	1	
20	TOYOTA Co	16950	30	3	2002	64359	110	1	0	1600	3	4	5	85	1105	1	1	3	1	1	
21	TOYOTA C	15950	30	3	2002	67660	110	1	0	1600	3	4	5	85	1105	1	1	3	1	1	
22	TOYOTA Co	16950	29	4	2002	43905	110	0	1	1600	3	4	5	100	1170	0	1	3	1	1	
23	TOYOTA Co	15950	28	5	2002	56349	110	1	0	1600	3	4	5	85	1120	0	1	3	1	1	