Accelerated Certificate Programs (ACPs)

Data Science & Predictive Analytics For Business Professionals

Strategic Business Analysis Using Predictive Analytics

I&C SCI\_X425.57

Spring 2018

Homework#5

Date Given: April 13, 2018 Due Date: April 16, 2018

=============================================================================

**Problem#1: Data Sampling**

Analyze the data source in ‘kc-house-data.csv’ file. This data source is a part of databases available in the public domain. This file contains 21,613 observations of real-estate properties of King county in Washington state. The data for the following 21 variables are provided.

1. id
2. date
3. price
4. bedrooms
5. bathrooms
6. sqft\_living
7. sqft\_lot
8. floors
9. waterfront
10. view
11. condition
12. grade
13. sqft\_above
14. sqft\_basement
15. yr\_built
16. yr\_renovated
17. zipcode
18. latitude
19. longitude
20. sqft\_living15
21. sqft\_lot15

Write R code with the following functionalities. Read the raw data source file ‘kc-house-data.csv’. Split the data source into 2 parts - (1) Training Data (2) Testing Data. The Training data should contain 70% of the observations and the Testing data should contain the remaining 30%. The selection of the 70% of the Training Data should be done randomly. To make sure that every student gets the same split, use zero (0) as the seed value of your random number generator (set.seed(0)).

Compute the following.

* Average house price of the training data.
* Average house price of the testing data.

**Answer**:

Average house price of the training data = $540,683.10

Average house price of the testing data = $538,700.10

**Problem#2: Data Discretization**

Download the Arrythmia data set from the UCI Machine Learning Repository <http://archive.ics.uci.edu/ml> (arrythmia.csv). Retrieve the first 5 columns of the first 100 observations.

1. Normalize all records to a mean of 0 and a standard deviation of 1.
2. Discretize each numerical attribute into 10 equi-width ranges. Compute the count of elements in each bin for all the columns.

Answer for the first column data only:

1. Normalized data for the first 5 rows of the first column.

1.8537956

0.6003395

0.4683968

0.5343681

1.8537956

-2.2364296

1. First column: 10 equi-width bins should have the following count.

1 1 2 11 23 18 17 12 6 9

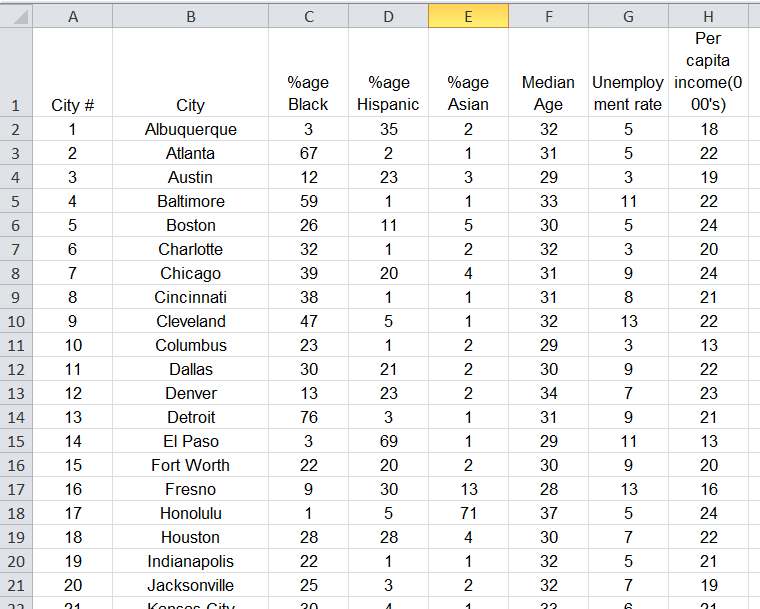
**Problem#3: Data Normalization**

The following data is given for the 49 of America’s largest cities.

Filename “RawDataUSCities.csv

* Percentage Black
* Percentage Hispanic
* Percentage Asian
* Media Age
* Unemployment rate
* Per Capita income

For example, Atlanta’s demographic information is as follows: 67% black, 2% Hispanic, 1% Asian, has a median age of 31, a 5% unemployment rate, and a per-capita income of $22,000.



1. To compare these numbers with each other, we have to standardize them. Standardize each demographics attribute data by computing the z-values of data value. (subtract the attribute’s mean and divide by the attribute’s standard deviation).
2. Scale all the demographics attribute data to the range [0,1].

Answer for the first 6 cities.

Normalized Data

PercentageBlack PercentageHispanic PercentageAsian MedianAge

[1,] -1.17872113 1.2389537 -0.36257405 0.06134197

[2,] 2.35518849 -0.7644344 -0.45230197 -0.43961742

[3,] -0.68176509 0.5104489 -0.27284613 -1.44153619

[4,] 1.91344978 -0.8251431 -0.45230197 0.56230135

[5,] 0.09127764 -0.2180558 -0.09339029 -0.94057681

[6,] 0.42258167 -0.8251431 -0.36257405 0.06134197

UnemploymentRate PerCapitaIncomeThousand

[1,] -0.7514633 -0.8752312

[2,] -0.7514633 0.3243864

[3,] -1.4953360 -0.5753268

[4,] 1.4801550 0.3243864

[5,] -0.7514633 0.9241951

[6,] -1.4953360 -0.2754224

Scaled data from 0 – 1.

PercentageBlack PercentageHispanic PercentageAsian MedianAge

[1,] 0.02666667 0.50000000 0.01428571 0.4444444

[2,] 0.88000000 0.01470588 0.00000000 0.3333333

[3,] 0.14666667 0.32352941 0.02857143 0.1111111

[4,] 0.77333333 0.00000000 0.00000000 0.5555556

[5,] 0.33333333 0.14705882 0.05714286 0.2222222

[6,] 0.41333333 0.00000000 0.01428571 0.4444444

UnemploymentRate PerCapitaIncomeThousand

[1,] 0.2 0.2777778

[2,] 0.2 0.5000000

[3,] 0.0 0.3333333

[4,] 0.8 0.5000000

[5,] 0.2 0.6111111

[6,] 0.0 0.3888889