COMPUTER SCIENCE 4373/5473
Assignment #2

Points: 100 **Weight**: 2%
Due: Friday, Sept. 11, 2020 at 11:59pm on Blackboard
Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: In this assignment, you will work individually to develop a program for the preprocessing data. You will use Python in Anaconda distribution to write a Jupyter notebook. Your program will read data from the provided CSV file into a DataFrame. The data file has 6 columns: A, B, C, D, E, and F, where A and B are categorical and the rest are numeric. Your program will than solve the following problems and write output to a text file. All values in the output should be formatted to have up to four (4) digits after the decimal point. Unless explicitly stated otherwise, you should use Python packages such as pandas and numpy that come installed with Anaconda 3. (You may also use UTSA VDI at www.utsa.edu/vdi, which has Anaconda 3 pre-installed.) Submit your solution via BlackBoard Learn in a zipped file yourName-hwk02.zip, which should contain your notebook, additional library if any, and the input/output files.
Notice: *Given the nature of on-line course, we will require you to practice using Words, Markdown, or HTML to write and format your homework solutions (**no scanned smeared image please**). It will prepare you for taking the on-line exams, where only a Words style editor (with HTML support) is available.*

1. [**30**] (Smoothing) Write functions that use the cut() and qcut() functions provided by DataFrame to smooth data in a given column using the following methods. Apply these function on column F.

   (a) Equal-depth binning with bin means for depth k, for example $k = 100$.

   (b) Equal-depth binning with bin boundaries for depth k, for example $k = 100$.

   (c) Equal-width binning with bin median for 10 bins.

2. [**30**] (Data Reduction) Write a function that takes a DataFrame, a set of column names (of numeric columns), and an integer p (less than the total number of columns in the table), and use PCA method in Scikit-Learn (specifically, sklearn.decomposition.PCA) to reduce the set of columns into p new columns. Apply this function to reduce the columns {C, D, E, F} into two columns p1, and p2.

3. [**40**] (Correlation) For this question, you will need to use packages scipy.stats

   (a) Compute the covariance and the correlation coefficient for each pair of the numeric columns.

   (b) Use the crosstab() function to construct the contingency table for columns A and B, similar to the following sample, where the distinct values in attribute A are $\{a_1, a_2\}$ and in attribute B are $\{b_1, b_2, b_3\}$ (this is just a sample and may not be the same as the data in your data file). Write a sequence of Python code to perform the Pearson's chi-square (χ^2) test of independence with a confidence level of 0.001 to determine if the two attributes are correlated. You should use stats.chi2() to get the χ^2 distribution. Print sufficient information to report the result of the test.

		A		
		a_1	a_2	all
B	b_1	??	??	??
	b_2	??	??	??
	b_3	??	??	??
	all	??	??	??