EXAM 1

Part 1: Python Code

Download "BusinessPlanData.dta" and "Exam1_Part1.ipynb" files from Data Mining Google Drive catalog and follow the instructions in the Jupyter Notebook.

Part 2: Exercise

As the price of oil rises, there is increased worldwide interest in alternate sources of energy. A Financial Times/Harris Poll surveyed people in six countries to assess attitudes toward a variety of alternate forms of energy (Harris Interactive website, February 27, 2008). The data in the following table are a portion of the poll's findings concerning whether people favor or oppose the building of new nuclear power plants. The level of significance is 5%.

	Country					
Response	Great Britain	France	Italy	Spain	Germany	United States
Strongly favor	141	161	298	133	128	204
Favor more than oppose	348	366	309	222	272	326
Oppose more than favor	381	334	219	311	322	316
Strongly oppose	217	215	219	443	389	174

- State Null and alternative hypotheses to test whether people's attitude toward building new nuclear power plants is independent of country.

- Fill in empty parts of the Contingency Table with Expected Frequencies.

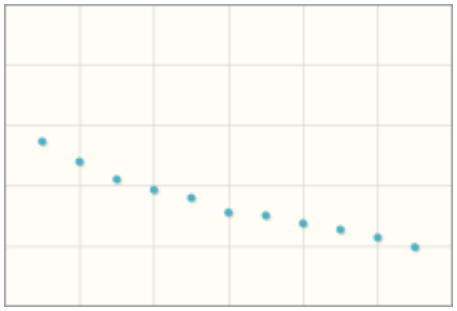
	Great Britain	France	Italy	Spain	Germany	United States	Total
Strongly favor	179.5	177.7	172.6				
Favor more than oppose	310.7	307.5	298.7	317.0			1843
Oppose more than favor	317.4	314.2	305.2	323.9	324.4	297.9	1883
Strongly oppose	279.3	276.5	268.5	285.0	285.5	262.1	1657
Total	1087	1076	1045	1109	1111	1020	6448

- Show the steps you have taken to test the hypothesis: State your decision and decision rule.

- Provide your findings within the context of the problem.

Part 3: Multiple Choice Questions

1. What we can conclude about Pearson correlation vs Spearman correlation from the plot below.



   a. Pearson correlation = Spearman correlation
   b. Pearson correlation > Spearman correlation
   c. Pearson correlation < Spearman correlation
   d. We have not enough info to conclude

2. Which one is true about P-value.
   a. P-value is a conditional probability
   b. P-value is an area
   c. In hypothesis testing the P-value approach is equivalent to the Test Statistics approach
   d. All of the above

3. Consider an experiment with four groups, with seven values in each. How many degrees of freedom are there in determining the total variation for the ANOVA summary table.
   a. 21
   b. 24
   c. 27
   d. 28

4. Which one is a Feature Selection technique.
   a. Univariate Selection
   b. Feature Importance
   c. VIF
   d. All of the above

5. In KDD and data mining noise is referred to as
   a. Meta data
   b. Complex data
   c. Random errors in database
   d. Data integration