BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #2

General Instruction: The homework assignment is due on the assigned date at 5:45 PM. Assignment turned in after the due date and time will lose 2 points for every day late. No assignment will be accepted one week after it is assigned. Here is a breakdown of the point distribution.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Preferred File Name** |
| Part A: 3 questions  Part B: 3 questions | 3 x 3/each = 9 points  3 X 4/each = 12 | LastName\_text.docx |
| R markdown documentation for Parts A & B | 2 parts x 2 points/each part = 4 points | LastName.html |
| ***Total*** | ***25 points*** |  |

To help facilitate the grading process, please use the file naming convention listed in the table above. Please upload the required files onto Blackboard for grading.

The html outputs of your markdown files will not be graded, but they will be checked if necessary to verify your findings and recommendations. Point deductions may occur if there are major discrepancies between your written answers and memorandums and the knitted markdown files.

Please upload your Word documents individually onto Blackboard. Please put your html files in a zipped folder and upload it. Blackboard’s upload feature does not accept html files.

**Underwriting at General Casualty Kansas City Insurance**

You are a newly hired data scientist at General Casualty Kansas City (GCKC) insurance. The CEO of GCKC would like to try to improve their pricing and product offerings for automobile insurance. They have traditionally only operated in the Midwest (IL, IN, MI, OH, WI, IA, KS, MN, MO, NE, ND, SD) but would like to go nationwide. She gives you data on fatal automobile accidents from 2011 and asks you to “see what the data tell you.” The file **crash.csv** contains nearly 5000 observations on the following variables:

* State
* Atmospheric condition - weather
* Crash date (mm/dd/yy)
* Fatalities in crash
* Roadway – type of location where crash occurred
* Age of driver or occupant
* Alcohol results – blood alcohol level for the driver
* Person type – who was killed
* Drug involved
* Race of driver or occupant
* Gender of driver or occupant
* Severity of the injury
* Year of the crash (2011)
* Month of the crash (integer)
* Day of the month of the crash (integer)
* Day of the week of the crash (integer)

GCKC is still seeking approval for their expansion from the regulators and may only be allowed to add one additional Census region at a time.

The other regions of U.S. are defined below:

**Northeast**: Connecticut (CT), Maine (ME), Massachusetts (MA), New Hampshire (NH), Rhode Island (RI), Vermont (VT), New Jersey (NJ), New York (NY), and Pennsylvania (PA).

**South**: Delaware (DE), Florida (FL), Georgia (GA), Maryland (MD), North Carolina (NC), South Carolina (SC), Virginia (VA), District of Columbia (DC), West Virginia (WV), Alabama (AL), Kentucky (KY), Mississippi (MS), Tennessee (TN), Arkansas (AR), Louisiana (LA), Oklahoma (OK), and Texas (TX).

**West**: Arizona (AZ), Colorado (CO), Idaho (ID), Montana (MT), Nevada (NV), New Mexico (NM), Utah (UT), Wyoming (WY), Alaska (AK), California (CA), Hawaii (HI), Oregon (OR), and Washington (WA).

Because the CEO hasn’t given you a target variable, you decide that an unsupervised learning method is appropriate to start. You choose to perform cluster analyses on the provided data set.

**Part A:** Define the question(s) and prepare the data

Because these data are from law enforcement files, they are not in the best shape for insurance analysis. A fair amount of cleaning and preprocessing is necessary. [This is a non-trivial exercise. Make sure you understand the data and what each entry represents. Then decide how YOU want to proceed.] You will be working with other data scientists in the department so your code should clearly show what you have done. Documentation is key here. You will be asked to report to the senior data scientist what you have found and how you intend to proceed. In no more than 3 or 4 paragraphs (not including tables and /or charts if you choose to use them), answer these questions:

1. Perform some preliminary EDA on the data and describe the sample. What do you learn about fatal automobile accidents from this exercise? How can you use that information?
2. What steps will you take to prepare the data? Why?
3. What questions will you be able to answer with the cleaned data? What limitations will you face?

(Note: because the audience here are other data scientists, technical terminology is fine.)

**Part B:** Unsupervised learning analysis

The senior data scientist agrees with your approach and tells you to continue the analysis and brief the CEO. Perform a cluster analysis using **2** different approaches. Choose one of the approaches and in no more than 3 or 4 paragraphs, answer the following questions for the CEO about the results.

1. How many types or groups of crashes are prevalent in the provided data set after you have conducted the analysis? How did you determine that? Describe the characteristics of each group.
2. Are there geographic differences in the crash types or fatalities? If yes, what were they? How might that affect the CEO’s decision?
3. Can you make at least two or three recommendations for policy or pricing options to the CEO based on your analysis? These recommendations can range from suggestions on whether or not to offer automobile insurance to certain areas or drivers as well to suggesting pricing differentials for different customers. Explain your reasoning.

Because you are briefing the CEO, technical language is NOT appropriate. She doesn’t understand terms like “k-means” and “centroid” so do not use them in your response. Your paragraphs should contain clear, concise, and correct language with no technical jargon.

NOTE: There are no wrong answers to these questions. There are badly formed questions or badly explained answers. The key to this assignment (and all real-world work) is to clearly explain your approach (what you did and why you did it that way) and the results you got.

Bonus: (2 points) In no more than two paragraphs, explain the difference between the two clustering exercises. How did the clusters differ? Why do you think that is? How did you choose which to present to the CEO? Technical language is fine.