STATS 415, Homework 1

Due at 11:59pm on Thursday, Sep 10, 2020

Note: Turn in a pdf scan of your homework on Canvas. Please limit your answer to Q3 to 12 pages, organized into a coherently typed data analysis report. Answers to Q1 and Q2 may be either typed or handwritten. Please clearly write your name, your UMID, and your GSI/lab number on the homework.

1. Consider the students of STATS 415 in F'20 as your sample (of convenience). [10 points per question]

   (a) Name one variable **related to academics** you could collect or measure on this sample in each of the following categories: categorical, ordinal and continuous.

   (b) Name a population about which we could plausibly make inferences on the variables you listed, based on the data collected from this sample.

   (c) Name a population about which we could not make inferences on the variables based on the data collected from this sample.

2. Consider a document-term matrix, where f_{ij} is the frequency of the jth word (term) in the ith document and n is the number of documents. Consider the variable transformation that is defined by

   $$f_{ij}^* = f_{ij} \ln \frac{n}{g_j},$$

   where g_j is the number of documents in which the jth term appears, known as the document frequency of the term. This transformation is called the inverse document frequency transformation.

   (a) What might be the purpose of this transformation? Illustrate your answer by considering a rare term and a common term, and giving a specific example comparing f_{ij} and f_{ij}^*. [10 points]

   (b) Based on this document-term matrix, give one potential statistical task of supervised learning and unsupervised learning respectively. [10 points]

3. This exercise relates to the College data set, which can be found in the file College.csv on the following webpage:

   http://faculty.marshall.usc.edu/gareth-james/ISL/data.html.

   It contains the following variables for different universities and colleges in the US:

   - Private : Public/private indicator
   - Apps : Number of applications received
   - Accept : Number of applicants accepted
   - Enroll : Number of new students enrolled
   - Top10perc : New students from top 10% of high school class
   - Top25perc : New students from top 25% of high school class
   - F.Undergrad : Number of full-time undergraduates
   - P.Undergrad : Number of part-time undergraduates
   - Outstate : Out-of-state tuition
   - Room.Board : Room and board costs
   - Books : Estimated book costs
   - Personal : Estimated personal spending
   - PhD : Percent of faculty with Ph.D.'s
   - Terminal : Percent of faculty with terminal degree
   - S.F.Ratio : Student/faculty ratio
   - perc.alumni : Percent of alumni who donate
   - Expend : Instructional expenditure per student
   - Grad.Rate : Graduation rate

   Perform exploratory data analysis of this dataset and write up your findings in a report. Comment on any interesting or significant features. Inlcude the variables names in whatever tables or plots you report. You can do any exploration you like as long as you include

   - Some numerical summaries for each variable
   - Some multivariate numerical summaries (e.g., pairwise correlations)
   - Some graphical summaries for each variable. Include at least one boxplot and at least one histogram.

- Some multivariate graphical summaries (at least one with pairwise scatter plots, and at least one with side-by-side boxplots).

Pay special attention to ensure that you are using appropriate summaries for different types of variables; for example, do not compute the mean of a categorical variable, even if its values are coded with numbers. [50 points]