**Adam Jones HW 3 Part B+C**

BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #3

**Part B: Principal Components Analysis (PCA) on University Ranking**

The data set **Universities.csv** contain information on 1,302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school). You should remove all categorical variables and missing numeric measurements from the data set prior to doing the tasks below.

1. Conduct a principal component analysis (PCA) of the given data set and comment on the results.
   1. **95.59% of the total variance across all 18 components found in the first two principal components (pre-normalization).**
   2. **After normalization, 57.68% of the total variance across all 18 components can be found in the first two principal components.** 
      1. **PC1 (31.07%) is dominated by the student faculty ratio variable.**
      2. **PC2 (26.61%) is dominated by the # of applications received variable.**
      3. **Components 3-18 are all on relatively similar scales with many components having similar % of variance as other components.**
      4. **The % of variance across PC3, 4, and 5 is very similar**
      5. **The % of variance across PC6 and 7 is very similar**
      6. **The % of variance across PC8 and 9 is very similar**
      7. **The % of variance across PC10, 11, and 12 is very similar**
      8. **The % of variance across PC13 and 14 is very similar**
      9. **The % of variance across PC17 and 18 is very similar**
2. Should the data be normalized? Discuss what characterizes the components you consider key.
   1. **The data should be normalized. 95.59% of the total variance across all 18 components is found in the first two components.**

**Part C: Final Project Proposal**

Please write one paragraph describing your final project proposal. Please include these items in your project proposal:

* Business question(s) you want to answer or solve
  + **Which raw material specifications (independent variables) have the biggest impact on finished product quality (dependent variables)?**
  + **Which raw material specifications tend to be clustered together?**
  + **Independent Variables (predictors)**
    - **W (Alveo)**
    - **P over L (Alveo)**
    - **Farinograph Stability**
    - **Farinograph Absorption**
    - **Protein-12% Moisture Basis**
    - **Wet Gluten-Wheat**
    - **Falling Number-Wheat**
    - **Protein-Dry Basis**
    - **Gluten Index**
  + **Dependent Variables (outcomes)**
    - **Total Bread Score**
    - **Loaf Volume**
    - **Voume du pain**
    - **Note totale**
    - **Volume Chronotec-1H45**
    - **Volume Chronotec-2H45**
    - **Volume 40min Soft Bread**
    - **Volume 65min Soft Bread**
    - **Total Score Soft Bread**
* Data source(s) you plan to use
  + **Internal SharePoint Application for Quality Control results**
  + **123,000 records dating back to 2001**
  + **ALL lab results recorded in a SINGLE field! Major clean-up required as I must scrape the column using an if statement to pull out only the results I want. I would like to discuss this dataset with you as I am not sure if this is going to cause an issue.**
* Supervised and/or unsupervised learning methods you plan to use to mine the data source(s). A minimum of two methods is required.
  + **Logistical Regression because I would like to see how strongly certain raw material specifications correlate to a finished good analysis result (outcome variable)**
  + **K-Means because I would like to see if there are common groups of quality control variables (I will include all the independent AND dependent variables).**

**Extra Credit: Introduction to GitHub**

The purpose of the extra credit assignment is to introduce you to GitHub, a platform to manage and share projects. Programmers, data scientists, and analysts use GitHub to share codes and data. Please work through the tutorial in the Tutorial\_R\_Project\_Git\_No\_Command\_Line.docx file. To earn the extra credit point, you will need to send me the link to a repository with the **crash.csv** data set and **.Rmd** file for Part A. You can include the link in the file for Parts B & C of the homework assignment.