Assignment 2
Principal Component Analysis

To be completed in groups of UP TO THREE (3)
Due: Friday February 18, 2022 @11:59pm.
Grading: 5% of course grade (50 points available)

Submission Instructions

Please submit this assignment *electronically* before the due date. Late submissions will **not** be accepted. Submit via the A2L dropbox for the appropriate assignment. Be sure that you have the names and student numbers of all students on the front page of your submission. Submit your answers as a **single .pdf** file including all relevant figures, tables, and math. You may include relevant code embedded in the report, but you **must submit a .zip along with your report** that includes all your code for the assignment.

Please upload your files with the **naming convention** `A0X_macID1_madID2_macID3`, where the first McMaster ID is for the person uploading the submission.

Up to 10 points may be deducted from your submission for sloppy or otherwise unprofessional work. This is rare, but possible. The definition of unprofessional work may include:

- Low-resolution screenshots of figures and tables.
- Giving no context to an answer relating to the task (*i.e.* "See code" or "113.289" with no units, context, or discussion whatsoever).
- Clear changes in author denoted by format changes, blatant writing style changes, or other factors that may deduct from the cohesiveness of the report.
- Failing to provide references for work that is obviously not yours (particularly bad cases will be considered as academic dishonesty).

Problem 1 [10 Points]

Consider the dataset `distillation.csv` from the A2 data companion for this assignment. This dataset contains operational information for an industrial distillation tower splitting two components, known as the "heavy key" (comes out the bottom) and "light key" (comes out the top) over a period of 20 days. **Note that you do not need to know anything about distillation columns to answer these problems**. The first column in the data set is time (in minutes). The remaining columns and their tags are:

- Column 2 (`CONDL`) – The level measurement in the column condenser (inches)
- Column 3 (`CONDP`) – The condenser pressure (kPa)
- Column 4 (`REBL`) – The level measurement in the reboiler (inches)
- Column 5 (`TOPS`) – Production rate of the tops product (gal/hr)
- Column 6 (`FEED`) – Feed rate to the tower (gal/hr)
- Column 7 (`BOTTOMS`) – Production rate of the bottoms product (gal/hr)
- Column 8 (`MFTOPS`) – Mole fraction of heavy key in the tops product
- Column 9 (`MFFEED`) – Mole fraction of heavy key in feed
- Column 10 (`MFFEED2`) – Mole fraction of light key in the feed
- Column 11 (`TEMPFEED`) – Feed temperature (°C)
- Column 12 (`PRESS`) – Feed stream pressure (kPa)
- Column 13 (`PRESSB`) – Pressure measured near the bottom of the tower (kPa)
- Column 14 (`PRESST`) – Pressure measured near the top of the tower (kPa)
- Column 15 (`RR`) – Reflux ratio (a column operating parameter)

1. Analyze the raw data and report at least three pairs of correlated variables along with their scatter plots to prove correlation. If you find any data to have zero variance, indicate so and remove it from the data set. Plot a scatterplot matrix of the final variables to be analyzed. [**8**]

2. Report the mean and standard deviations of all the variables. Then, mean centre and scale the data to obtain mean-centered and scaled data matrix to use for PCA. Plot tags `MF_TOPS` versus `RR` on a scatterplot to show the data has been scaled. [**2**]

Problem 2 [20 Points]

Build a PCA model on the above `distillation.csv` dataset to answer the following questions. The PCA model should be built by running `nipalspca.m` in `MATLAB`. These files are user-defined functions of the form:

```
[t, p, R2] = nipalspca(x,A)
```

which fits a PCA model using an input data set `x` and number of desired components `A`, and returns a matrix of scores `t`, loadings `p`, and the cumulative R^2 of the model for each component `R2`. Be sure to use the data set after removing variables with no variance. Moreover, do not include the time column in your PCA.

1. Build a PCA model with THREE principal components and report the performance metric R^2 using different numbers of components. Make an argument that a fourth component would not substantially explain additional variance in the data. [**5**]

2. Make a score plot using the first two score vectors. Use different markers to show points with lower time tags versus higher time tags. What do you observe? Discuss key features of the plot briefly. [**5**]

3. Make a loadings bar plot for each of the first two principal components. Identify which variables contribute the most variance to each component, and any correlations. [**5**]

4. Make a score plot for the first two components as above and overlay the loadings as points or vectors to show a score/loadings plot. Then, use this plot and the bar plot above to explain what kinds of values in **X** you might expect to have a LOW t_1 (score for component 1) and HIGH t_2 (score for component 2)? *HINT: see information below* [**5**]

To make things a little easier, the following functions have been provided uploaded to A2L in the "`MATLAB` Content" section (free of charge!) should you decide you wish to use them:

- `loading_plot(p,number,Dataset)`: accepts a loading vector `p`, a `number` representing which component it is plotting, and an optional input `Dataset` that is a string array of variable names. If `Dataset` is not provided, it labels the columns in X as "variable 1, variable 2" etc.
- `score_loading_plot(t1, t2, p1, p2, Dataset)`: accepts any two scores `t1` and `t2` and corresponding component loadings `p1` and `p2` and creates a score plot with the loadings overlaid as vectors. Also plots the confidence intervals for Hotelling's T^2 (more on this later in the course). Also accepts an optional `Dataset` string array that labels the loading variables.
- `scoreplot(t1, t2)`: accepts any two score vectors `t1` and `t2` and plots them as a score plot. Also plot Hotelling's T^2 confidence intervals.

Problem 3 [10 Points]

In the previous problem, we used the so-called NIPALS algorithm in a function provided to you to fit a PCA model. We'll learn more about that later. However, note that there are other ways of fitting loadings and scores to a PCA model. One of the methods is through what is called the Eigenvalue Decomposition of the centered and scaled data matrix.

Build a function in `MATLAB` or `Python` that accepts a data matrix and requested number of components and returns score vectors **t**, loading vectors **p**, and R^2. Your function should use Eigenvalue Decomposition and NOT include the NIPALS algorithm in the function provided as a part of the assignment. Report the scores and loadings for a 2-component model of the `distillation.csv` dataset and confirm that they are the same as provided by the NIPALS algorithm. If they are different in some way, explain why.

Problem 4 [10 Points]

In class we have performed centering and scaling of data as a pre-processing step before building PCA models. However, there are many different strategies for data preprocessing before model building. In fact, PCA can be used as a data pre-processing method when building other models on a dataset.

Your task in this question is to do a literature/media search on data preprocessing methods and discuss one of these methods. You are expected to:

1. Give a descriptive overview of the method according to your source. [**5**]

2. Discuss what types of data this method could potentially work for, including any pros or cons of using this preprocessing step. Provide an example (from your source) of the preprocessing step in action. [**5**]

Be sure to include citations and references to your sources.