Kyle Ford

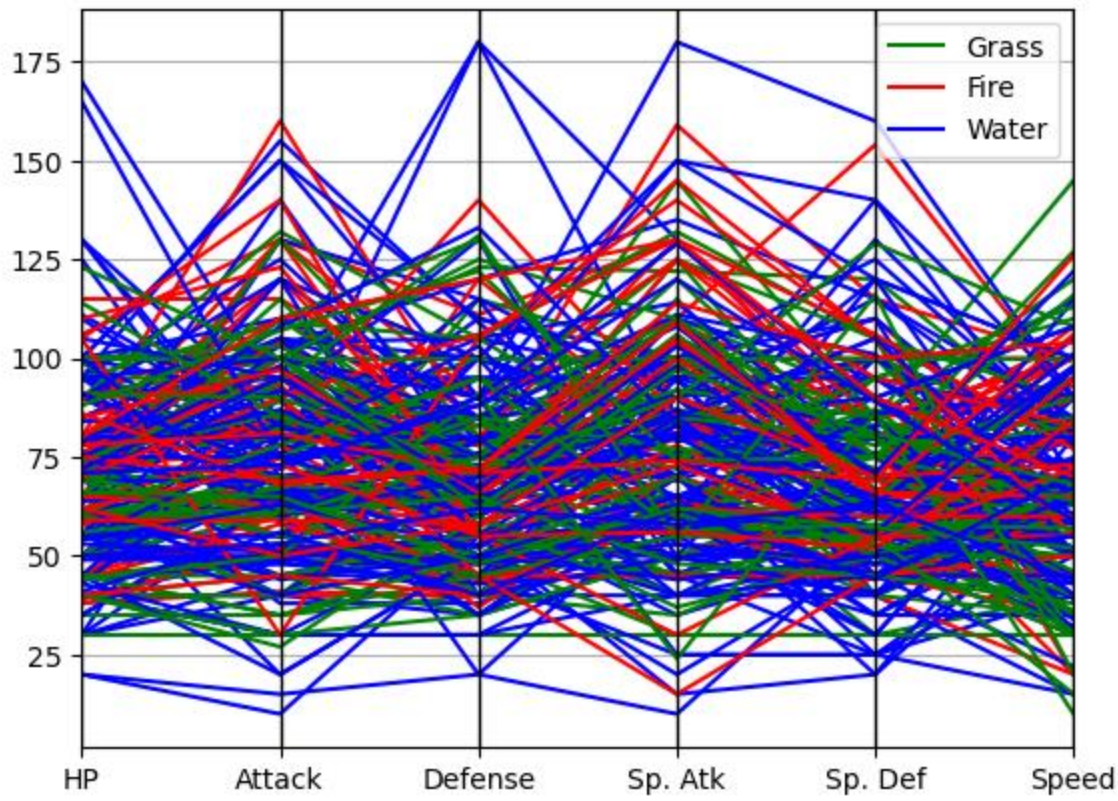
CPSC 5530

Due Date: 2/26/2023

Assignment 2:

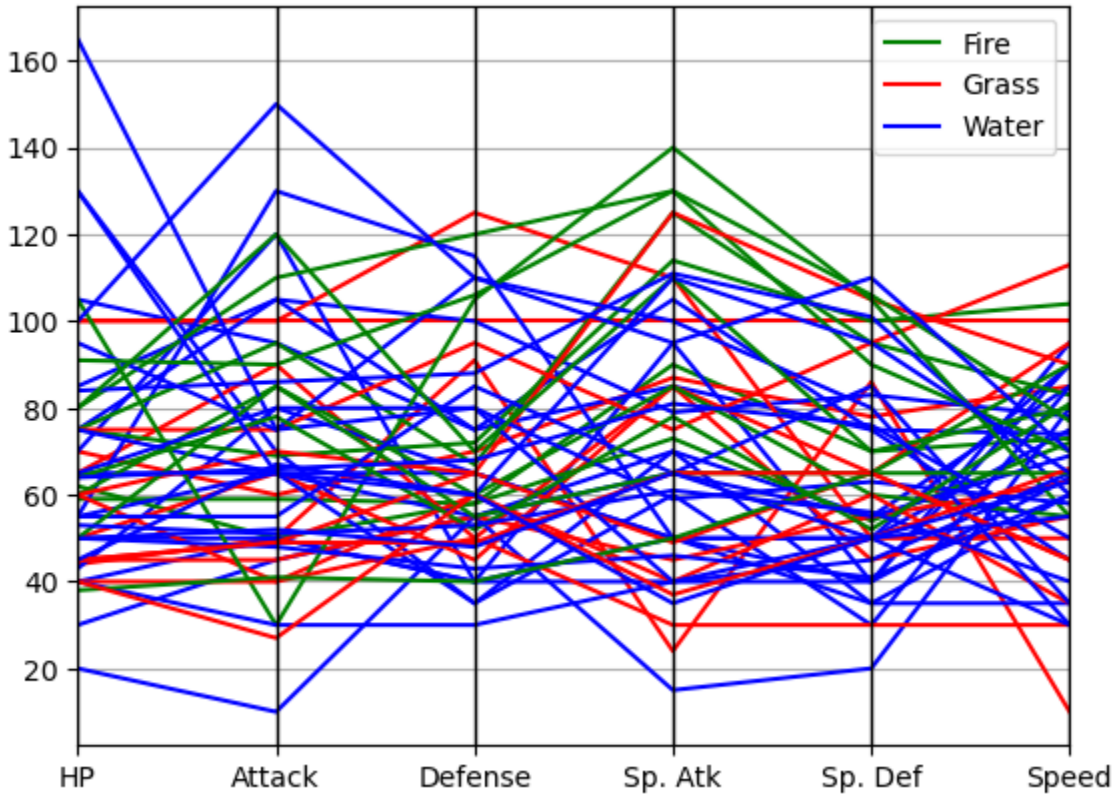
Dataset 1: Pokémon dataset: 6D

1. Data set downloaded from the URL https://www.kaggle.com/code/residentmario/multivariate-plotting/data
2. This data set is organized in the form of a table. Specifically, it is a simple table. Each row in the table represents a different entry corresponding to a different Pokémon. Each column contains a different attribute used to describe the characteristics of the Pokémon.
3. This dataset contains both items and attributes. As described in part (2), each column represents a different attribute, or characteristic, of the Pokémon, such as Name, Type, and various numerical statistics describing the Pokémon's abilities. Each row is an item which corresponds to an individual Pokémon.
4. This dataset has 12 attributes, namely Name, Type 1, Type 2, Total, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed, Generation, and Legendary. Each of these could be theoretically plotted in order to observe trends among differing Pokémon. The Name field refers to the Pokémon name, the Types 1 and 2 attributes refer to the physical type of the Pokémon, the HP, Attack, Defense, Sp. Atk, Sp. Def, and Speed attributes refer to numerical statistics describing a Pokémon's abilities, the Generation attribute is a categorical attribute describing which generation the Pokémon belongs to, and the Legendary attribute is a binary categorical attribute denoting whether the Pokémon is legendary or not.
5. First, the dataset is read into Python as a pandas data frame. We begin by checking the types to make sure they are compatible for comparison. Specifically, we are interested in comparing the numerical statistics for different Pokemon types. This will answer the question of which type of Pokemon is better suited in which type of fighting attribute. We find that our data types are acceptable, and we can begin visualization. Since we are interested in comparing differing numerical quantities, we choose a parallel coordinates plot to show the relationship between attributes. Doing so for the whole data frame yields the following visualization:



Note that this visualization is incredibly cluttered, with different lines occluding and overlapping others such that it is difficult to discern any trends. Thus, as part of the data preprocessing process, we print the total number of Pokemon represented in our data frame, and take a random sample of 200 of them. Doing so will still give us a representative sample, but make the visualization easier to read.

6. Visualization:



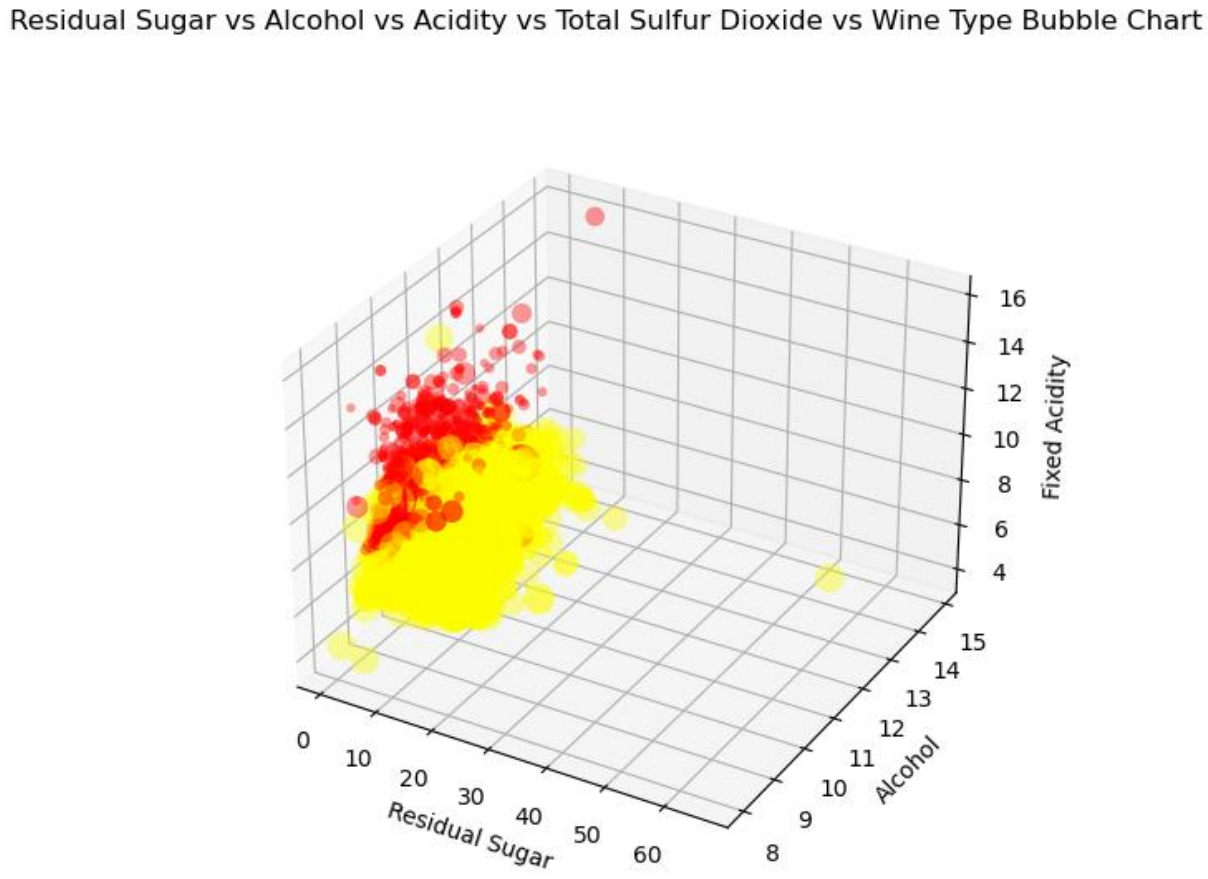
Note that at the extreme high and low ends of our y-axis values are blue lines, representing water type Pokemon. We conclude that they are likely to experience extreme values. Also note that red lines representing fire type Pokemon are concentrated in the center of the plot, indicating average statistical values for fire Pokémon. Finally, note that green lines representing grass Pokemon tend to trend higher than fire Pokémon, but lower than the extreme values represented by water Pokémon.

Dataset 2: Red and White Wine Dataset: 5D

1. Data set downloaded from URL https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57
2. This data comes in the form of a table. Specifically, this dataset is the combination of two tables, one representing the qualities of red wine, and the other representing the qualities of white wine.
3. This dataset makes use of items and attributes to describe the data. Each column in the dataset represents an attribute, which is a characteristic of the wine. Alternatively, each row in the dataset represents a different type of wine, either red or white.
4. This dataset contains a total of 12 dimensions/attributes. Specifically, it contains the attributes fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. The attributes fixed acidity, volatile acidity, and citric acid refer to the numerical quantity of amounts of

different types of acid present in each wine. The attributes residual sugar, chlorides, free sulfur dioxide, and total sulfur dioxide represent the total amount of other chemical substances present in the wine. The pH attribute refers to the pH scale of the wine. The density attribute describes the density of the wine. Finally, the quality attribute is an ordinal data type ranking the wine's quality on a scale of 1-10.

5. There are two main parts of preprocessing that must be done for this data set. First, when importing this dataset as a pandas data frame, we must add the additional argument sep = ";". This is because the original csv file is presented as a single column, with all of the attribute values separated by a semicolon. Second, we must merge two csv files to create a single data frame containing the data for both red and white wines. This is because the original dataset is comprised of two tables, one for red wine and the other for white wine. During this process, we also introduce a function to add an additional attribute – wine quality. This is a categorical attribute that distinguishes between wines of low, medium, and high quality.
6. Visualization:

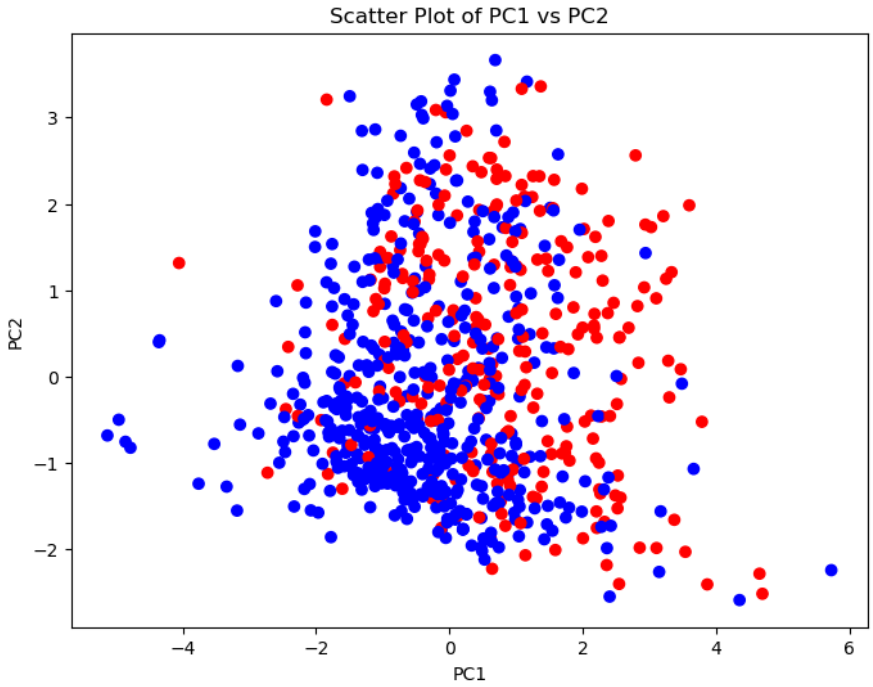


The five dimensions we plot are residual sugar, alcohol, acidity, total sulfur dioxide, and wine type. We do this via a 3-dimensional bubble chart. Three of our dimensions, residual sugar, alcohol, and fixed acidity, are represented on the x-, y-, and z- axes. The wine type attribute is represented via color, with red wines represented by red bubbles, and white wines represented by yellow ones. Finally, we employ the depth feature to

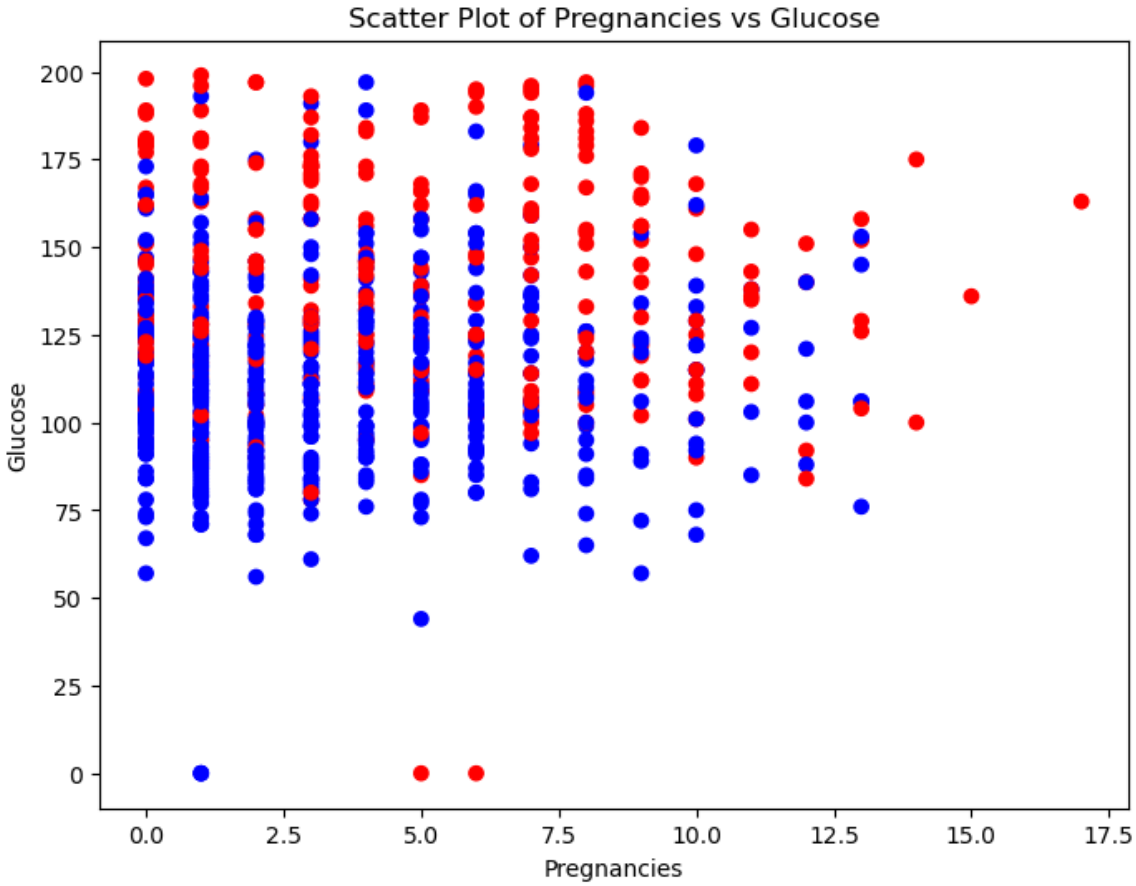
encode the total sulfur dioxide dimension. This plot illustrates that red wines tend to have higher levels of fixed acidity, but that white wines tend to have higher levels of residual sugar. The alcohol level appears to be fairly evenly distributed.

Dataset 3: Diabetes Dataset: PCA

1. Dataset downloaded from URL https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
2. This dataset is in the form of a simple table where the columns represent attributes, and the rows represent specific entries.
3. The data is in the form of items and attributes. The attributes are the columns in the table, corresponding to pregnancies, glucose, blood pressure, skin thickness, BMI, diabetes pedigree function, age, and outcome. All of these attributes except for outcome are numerical, quantitative data types. The outcomes data type is a binary categorical data type.
4. This dataset features 8 dimensions/attributes. Pregnancies corresponds to the number of times a person had become pregnant. The glucose attribute corresponds to the amount of glucose in a person's blood at the time of testing. The blood pressure attribute refers to a person's recorded blood pressure at the time of testing. The skin thickness attribute refers to the thickness of a person's skin. BMI refers to a person's measure body mass index at the time of testing. The diabetes pedigree function refers to a numeric value calculated based on the likelihood a person will get diabetes based on genetic and ancestral information. The age attribute refers to how old a person is. The outcome attribute refers to whether or not a person has been diagnosed with diabetes or not.
5. Since we are performing PCA on this dataset, the first step is to determine our feature variables and our target variables. We will fit the PCA across the feature variables to attempt to better represent the target variable. For this dataset, we choose pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age as the feature variables, and outcome as the target variable. Thus, in order to perform the PCA, we create a new data frame where the outcome column is dropped. This creates a data frame where only the feature variables are present. The next step in PCA is to scale, or normalize the data. After that, we run the PCA, specifying the number of principal components to be two, on our scaled data.
6. PC1 vs PC2 scatter plot:



Pregnancies vs Glucose scatter plot:



These two plots illustrate the effect of PCA. The goal is to better be able to predict the outcome of an event, in this case whether or not a person will be diagnosed with diabetes.

In terms of visualization, we can see this effect in terms of clustering, where similar results cluster together. Note that the plot of PC1 vs PC2 demonstrates much higher levels of clustering than the plot of two of the features. It is difficult to discern any observable pattern from the second graph, while the first one demonstrates much higher similarity via clustering.