Cody Whitt
pkz325
CPSC 4530
Spring 2023
Assignment 2

Github for Project: https://github.com/codydev901/4530_assignment2

Dataset 1: Life Expectancy & Schooling By Continent/Year (4 Dimensions)

Source and Inspiration

This dataset was found on Kaggle while browsing through the highly rated available datasets. There seemed to be questions on the quality of some of the attributes, but the attributes looked at for this analysis seemed ok. I wanted to see if the dataset reflected what one would consider “common preconceptions”, which will be discussed further in the analysis section. Note: a 2nd dataset was used to help in a transformation step.

Link:
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
https://www.kaggle.com/datasets/statchaitya/country-to-continent

Raw Data Filename: Life Expectancy Data.csv, countryContinent.csv

Dataset Type

These datasets are both in the form of a table, in particular flat tables. They exist as a single .csv file which was downloaded from the above link and used locally from then on. Both are completely static in the context of this analysis, as one of the attributes is temporal (years between 2000-2015 inclusive).

Data Type

Speaking in terms of the primary dataset (life expectancy), data exists in this dataset in the form of items and attributes. Items exist as individual rows in the table and the attributes are defined as the column labels (the first row of the .csv). The data contains 2938 items and 22 attributes in its raw form, where each item can best be described as holding various pieces of human-development type information for a country in a given year between 2000 and 2015. Country/Year combination would represent a primary key for each item.

Attribute Types and Semantics

As mentioned above, the raw data contained 22 attributes, however I used this as my 4-dimensional data set, so I will only cover the 4 I kept. These were: *Country, Year, Life expectancy* and *Schooling*. *Country* holds a string and represents the name of a country/geopolitical entity (for example, The United States or Germany). This is an unordered, categorical attribute. *Year* holds an integer, and is a temporal, ordered, discrete and quantitative attribute. *Life expectancy* holds what is assumed to be the average life expectancy for that item (country/year pair) and is stored as a float (quantitative continuous), and *Schooling* holds what is assumed to be the average number of years devoted to education for that item as a float (quantitative continuous).

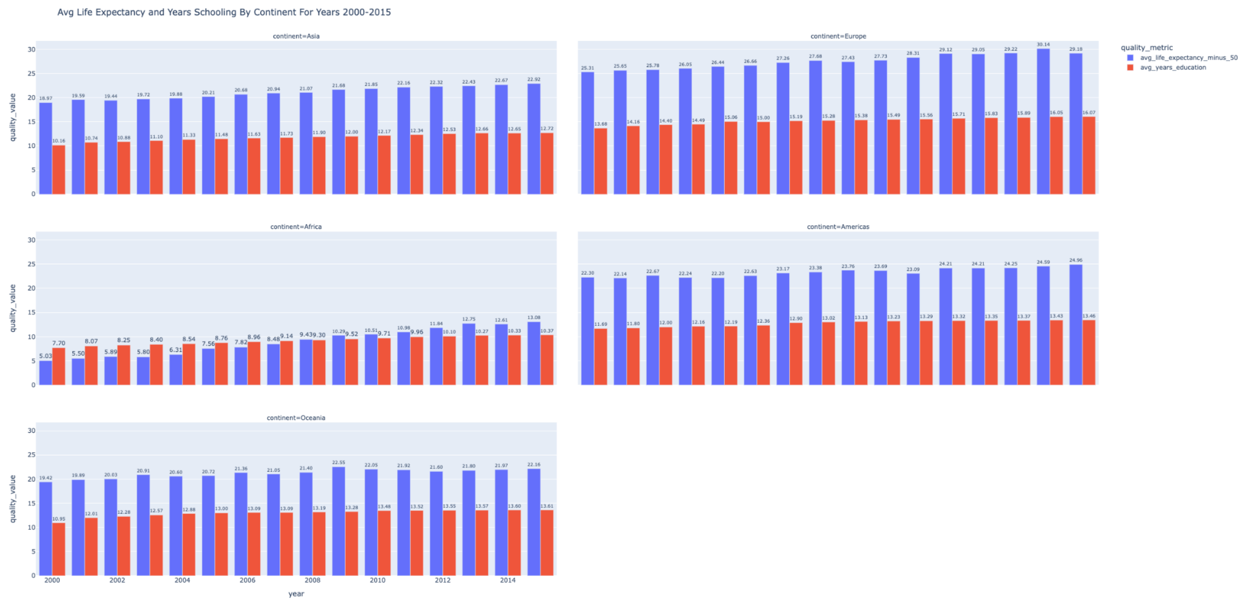
Pre-Processing and Plotting

I used Python to process and plot the data, in particular Pandas and Plotly. The data was first read into a DataFrame and then immediately filtered to just include the attributes above (in order that the null-check did not discard rows based on attributes we were not interested in). I then checked for missing values, where a little under 200 were found. This reduced the number of rows to 2768. After that I renamed the columns to clean things up a little (for example, the "Life expectancy " column had an extra space at the end. This resulted in the following attributes: *year, country, life_expectancy, years_education*. In order to reduce the number of unique values in the categorical attribute (country), I then decided to perform a transformation step where I converted the *country* to *continent*. This was done by using the 2nd dataset mentioned above to make a map where could be indexed by country to produce its associated continent. This transformed 173 unique countries into 5 continents. I then grouped the data by (continent/year) and took the average of the associated life expectancy and years education and stored those calculations in a new data frame. This then resulted in the parsed data being in a position to easily compare change in life expectancy and years education between the continents over the years. I performed one more transformation in regards to how I handled the resulting *avg_life_expectancy*, which will be discussed below.

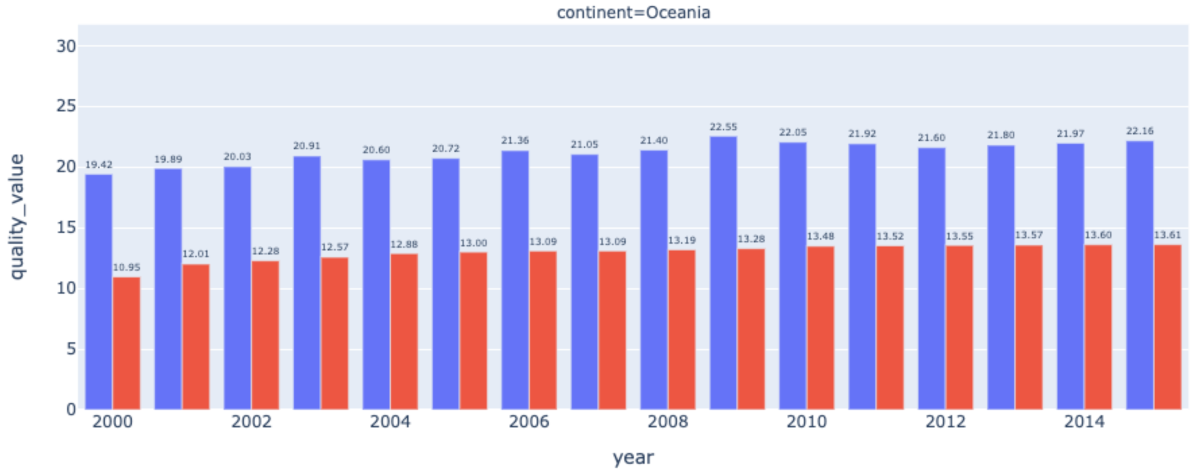
To plot, I chose to use a group of bar charts. In particular, the *continent* attribute was handled by the use of multiple displays and the two "quality metrics" (*life expectancy* and *years education)* were shown as separate clusters in the bar charts, with the color channel providing an additional point of differentiation. The x-axis holds the *year*, and the y-axis the values for the quality metrics. Since *life_expectancy* was overall much larger in magnitude than *years_education*, I decided to reduce its magnitude by a constant (I subtracted 50 and labeled it appropriately). This allowed the two groups of clustered bars to be more easily compared/the change over time seen. This adds a slight cognitive load for the user, but I felt it was preferable to some other form of normalization since "years" is a familiar concept and I wanted the actual values to be easily inferred from the figure. This is a region-based technique.

Parsed Filename: school_life_parsed.csv

Figure



Full Figure



Zoomed Sub-plot.

Analysis

The purpose of this dataset/figure that I was just to take a look at the common preconceptions that 1. certain continents have higher quality of life in terms of life expectancy and education, and 2. That life expectancy and education rates have been trending upwards. I feel that both the figure and the underlying data support these assumptions. For example Europe can be seen to have a higher life expectancy than Africa per the height of the blue bars, exactly how much higher in a given year can be calculated by taking the difference of the absolute values displayed above the bars. The trend that both life expectancy and education has increased can be seen by comparing the height of the bars from left to right and seeing an upwards trend, with the biggest gains seen in Africa, where 8 years were gained between 2000 and 2015. An overall observation is that Europe seems to lead the world by around ~4 years in both average life expectancy and average years of education.

Dataset 2: Stroke Events (5 Dimensions)

Source and Inspiration

This dataset was found on Kaggle while browsing through the highly rated available datasets. I wasn't sure exactly what I was going to try to do with it, but I figured once I started parsing it that I could attempt to use it to answer some questions in terms of relating some of them to the occurrence of having a stroke.

Link:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Raw Data Filename: healthcare-dataset-stroke-data.csv

Dataset Type

This dataset is in the form of a table, in particular a flat table. It exists as a single .csv file which was downloaded from the above link and used locally from then on.

Data Type

Data exists in this dataset in the form of items and attributes. Items exist as individual rows in the table and the attributes are defined as the column labels (the first row of the .csv). The data contains 5110 items and 12 attributes in its raw form, where each item could be described as holding medical information for a person and a binary classification on whether or not they experienced a stroke. This then allows the examination of the supporting medical information in relation to the potential presence of a stroke event.

Attribute Types and Semantics

I used this dataset as the 5-attribute set, so I will only discuss the 5 attributes I kept. After examining the possible attributes, I chose to look at the following: *gender, age, avg_glucose_level, bmi*, and *stroke*. I felt these in particular would be interesting in the context of relationship with a stroke event as they were both a mix of attribute types and easily relatable to a potential audience. *Gender* is stored as a string which could be described as a binary category (male/female), *age* is stored as an integer and is an ordered, quantitative, discrete attribute. *avg_glucose_level* and *bmi* both are stored as floats and are continuous quantitative attributes, and finally *stroke* is stored as an integer, but represents a binary classification where 1 indicates a stroke, and 0 doesn't.

Pre-Processing and Plotting

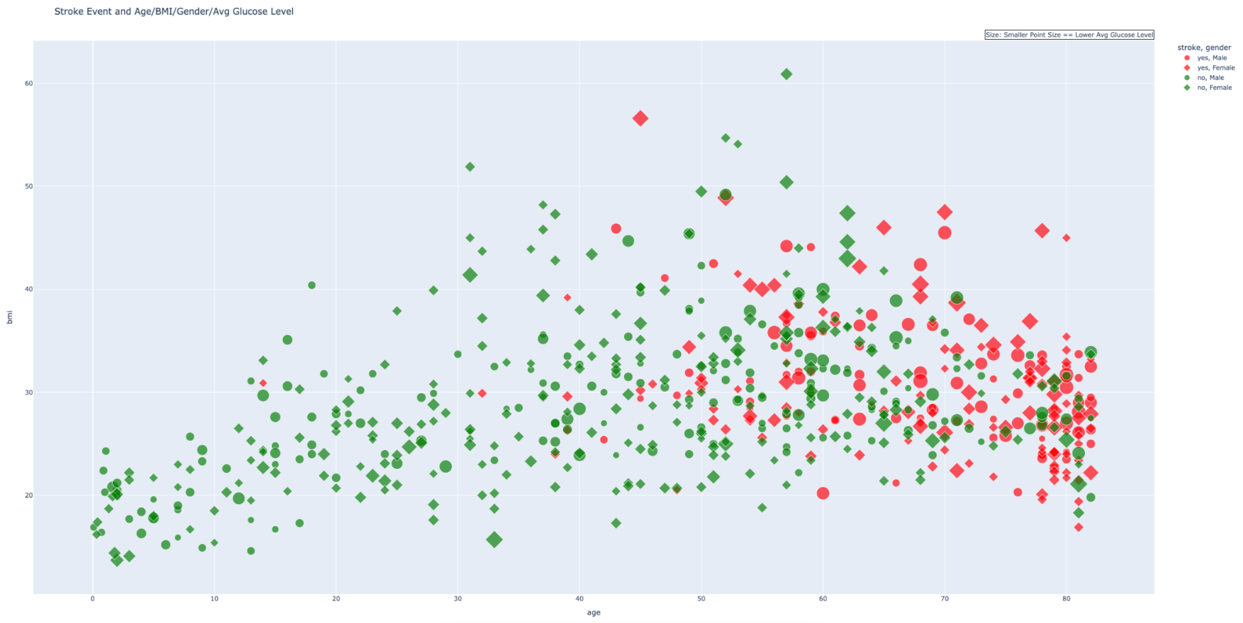
I used Python to parse and plot the data, in particular Pandas and Plotly. This dataset didn't require much pre-processing, but there is one important step I took that I will discuss.

After reading in the raw data, I first checked for null values. Roughly 200 were found, which brought the number of items from 5909 to 4909. After that I removed the attributes I wasn’t interested in by subsetting on the columns of the DataFrame. Since the *stroke* attribute was what I was most interested in terms of overall context, I then took a quick look to see how that attribute was represented throughout the data and saw that the dataset was class imbalanced in that 209 records contained a stroke, while 4700 did not. In both an attempt to balance that and reduce the number of items for plotting – I wanted to use a scatterplot for this, which are generally best used up to comparisons of hundreds of points, I decided to take a random sample/subset of the non-stroke data where I decided to only use 418 non-stroke data points (chosen by num_stroke * 2). All the had-stroke data points were included. This was done using the built-in .sample() function on Panda’s DataFrames, with a random_state set for reproducibility and replace set to False (so that the same sampled point couldn’t be taken twice). This then left a dataset with 627 items, where ⅓ of them were items which had a stroke, and ⅔ did not.

To plot I used a single scatter plot where I placed *age* on the x-axis, and *bmi* on the y-axis, since these were the two quantitative attributes. I then used color to encode *stroke* where red indicated a stroke, and green indicated no-stroke. *Gender* was encoded by the symbol of the points, and *avg_glucose_level* by the size of the points. The color/marker/size encoding represents use of dimension embedding. This is a point-based technique.

Parsed Data Filename: stroke_parsed.csv

Figure



Analysis

So a few trends that I found interesting, first is that BMI increased by age, but then began to decrease at age 70+. I suppose this makes sense because a higher-BMI is generally regarded as unhealthy and would lead to shorter life spans (less higher BMI points at higher age). Event of a stroke itself seems to be highly correlated with age as the “red” points are concentrated at age 50+ which is perhaps expected. However what I did find most interesting is perhaps women are more likely to experience a stroke before the age of 40, as the 6 data points indicated a stroke before that age are all women (seen by the diamond point). Overall BMI and Glucose Level don’t appear to have that much of a correlation with having a stroke in terms of what is able to be seen with this visualization – there is not a clear differentiation between stroke status at age 60+ and BMI (y-axis) or glucose level (point size) for example.

Dataset 3: Red Wine Quality (6+1 Dimensions)

Source and Inspiration

This dataset was found on Kaggle while browsing through the highly rated available datasets. I am using the white wine version in another class/in a different way, so thought it would be interesting to look at the red wine version in context of this assignment.

Link:
https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Raw Data Filename: winequality-red.csv

Dataset Type

This dataset is in the form of a table, in particular a flat table. It exists as a single .csv file which was downloaded from the above link and used locally from then on.

Data Type

Data exists in this dataset in the form of items and attributes. Items exist as individual rows in the table and the attributes are defined as the column labels (the first row of the .csv). The data contains 1599 items and 12 attributes in its raw form, where each item represents a unique wine and associated physical/chemical data with a quality score. This allows for the examination of the physicochemical attributes in relation to quality (which is derived from a human ranking).

Attribute Types and Semantics

As mentioned above, this dataset contained 12 attributes. I chose to use 7, though I make a slight distinction in calling it 6 + 1 above, in that I used 6 of the physicochemical attributes (quantitative, continuous floats) and quality (the +1) which is an ordinal/categorical attribute stored as an integer. Choosing which of the physicochemical attributes to use was an interesting part of using this dataset and will be discussed below, but I chose the following: *citric acid, total sulfur dioxide, free sulfur dioxide, residual sugar, chlorides,* and *volatile acidity*. *Quality* exists as an integer per above. It had a native range of 3 to 8, where higher value indicated a higher quality in terms of how a human rated the wine.

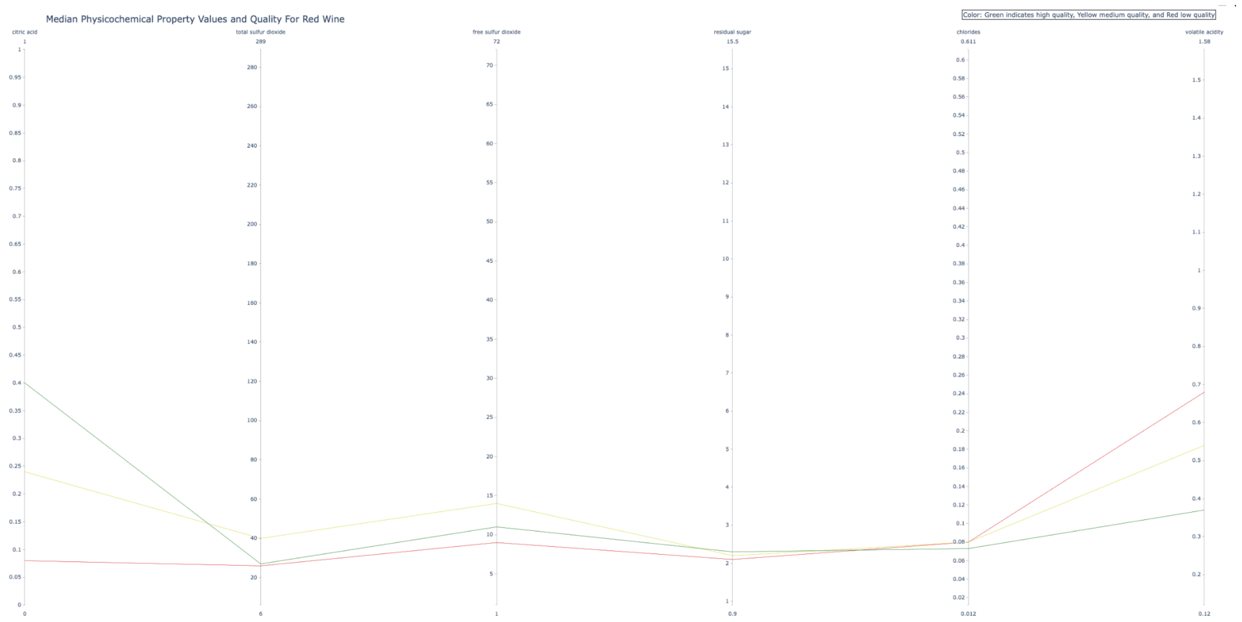
Pre-Processing and Plotting

I used Python to parse and plot the data, in particular Pandas and Plotly. The dataset was first read from .csv into a DataFrame and then checked for missing values, where in this case we were lucky and there were none. After that the most interesting part of this pre-processing occurred in which I needed to pick which attributes to use. I suppose this could be considered a simple dimension subsetting algorithm, but I ended up calculating the coefficient of variation that existed in the dataset for each of the attributes, and then picking the 6 with the most variation to keep. The thought was that I wanted to look at the relationship between potential differences in the physicochemical attributes in relation to quality, so the best ones to look at would be those with the most variation. The DataFrame was then sub-set to only include quality + the six attributes of interest. Next I did a further transformation step where I bucketed quality into 3 groups (from 6), where qualities of 3 and 4, 5 and 6, and 7 and 8 were grouped together. This represented low, medium and high quality wines. I did a quick polar coordinates plot with this and decided it was too messy (still 1599 data points), so I then decided to reduce the data further by taking the median value for each attribute at each of the three quality levels. I felt this would then allow for a quick comparison in quality vs. physicochemical attributes (though at the expense of seeing more fine-grained behavior). In order to preserve the range of the original dataset (non-median values), I captured the max/min those separately and used them in the figure to set the range of each attribute displayed. This then allowed the median values to be seen in context of the overall data. Two parsed datasets were then written to files to be used in plotting, one holding the median physicochemical attribute and quality data, the other holding the min/max ranges for each attribute to be used in the plot.

To plot I used a parallel coordinates plot since I felt it was appropriate in comparing the behavior of the 6 continuous quantitative attributes and the single categorical attribute. The categorical attribute was encoded by color (using the common red == bad, yellow == ok, green == good scheme) and an annotation added. This is a line-based technique.

Parsed Data FileNames: wine_quality_median_values.csv, wine_quality_original_range.csv

Figure



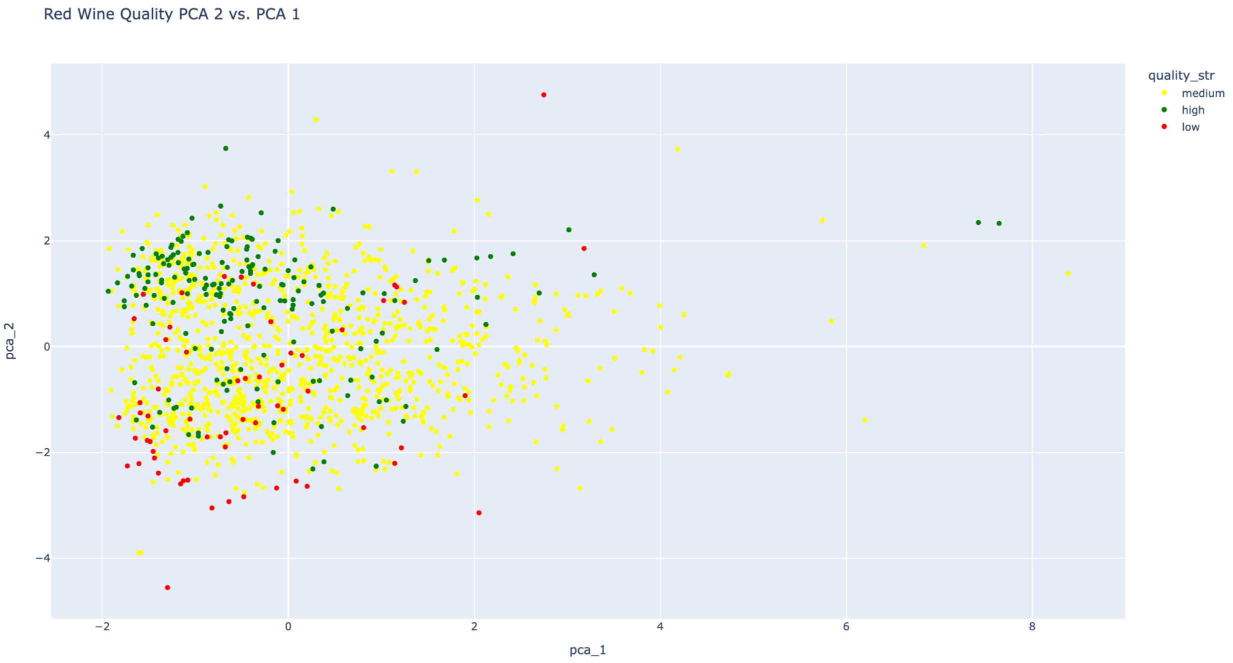
Analysis

I don’t really know anything about wine, but the figure does suggest a few trends that would perhaps be helpful in identifying what makes a “good wine” from its underlying physicochemical properties. First it appears that there is a positive correlation between Citric Acid and quality, with higher qualities having a larger magnitude there. Second, a negative correlation appears to exist between volatile acidity and quality, with lower quality wines having a higher volatile acidity. Residual sugar and chlorides don’t appear to have an effect on quality given their close median values, and the two sulfur attributes are inconclusive. Finally the presence of outliers/skewness in the original data can be seen by the location of the median values in relation to the min/max scales.

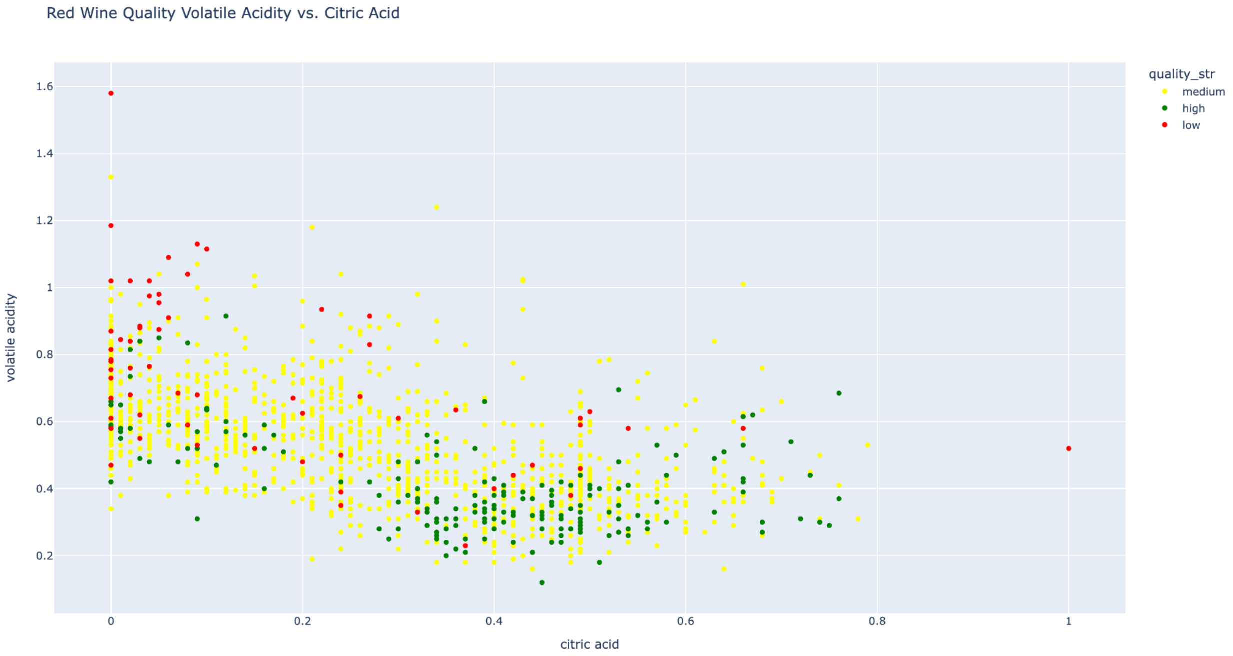
PCA

For the PCA portion of the assignment, I chose to use Dataset 3 which was the red wine quality one since it contained purely continuous quantitative attributes along with a categorical label, with the idea of attempting to relate the quantitative attributes to the category (wine quality). Since the parsed data set I used for the original wine analysis above only included median values, I wrote an intermediate version out for this that contained all data points. It will be interesting to see if the “recommendations” found in that earlier analysis hold true with this method. I followed the Medium post mentioned in the assignment in order to perform the PCA. It was performed in a separate .py file, pca_wine.py.

PCA 1 and PCA 2



Comparison of two selected Raw Attributes



PCA Analysis

So I chose Citric Acid and Volatile Acidity to use as the original data comparison points since the earlier analysis suggested they were most associated with change in wine quality. Both the PCA and the Raw Attribute plots show clustering between the 3 quality groups, though it is interesting to note that the “medium” quality group is present in both what could be considered the “low” and “high” clusters. There are also outliers in both areas. This is perhaps understandable as “quality” is subject to taste/interpretation (unlike the empirically collected physicochemical attributes). Seeing the different clusters in the PCA scatterplot does suggest some potential in relating quality to the underlying quantitative attributes, it doesn’t appear to be completely random etc.

A potential next step could be using only two (instead of three) quality buckets – the original data ranged from scores of 3 - 8 which was then grouped by 2. Perhaps a cleaner cluster split could be seen then. In terms of relating to the original DataSet 3 analysis, the trend of “higher citric acid/lower volatile acidity == higher quality wine” appears to have held up when re-examined here using the full dataset.