Assignment 2 Report

Dataset: https://data.world/uci/abalone

Dataset obtained from data.world, an online free repository of datasets.

Dataset format: Tabular

The original dataset is a static, flat table which contains data on various characteristics of abalones, including sex, height, length, etc. The dataset is multivariate, as there are multiple attributes for each key.

Data Type: Attribute-based

The events are laid out in a tabular dataset and each record is an attribute describing an abalone. The abalones are solely identified by their attributes rather than a pre-assigned key value.

Attribute Types & Semantics

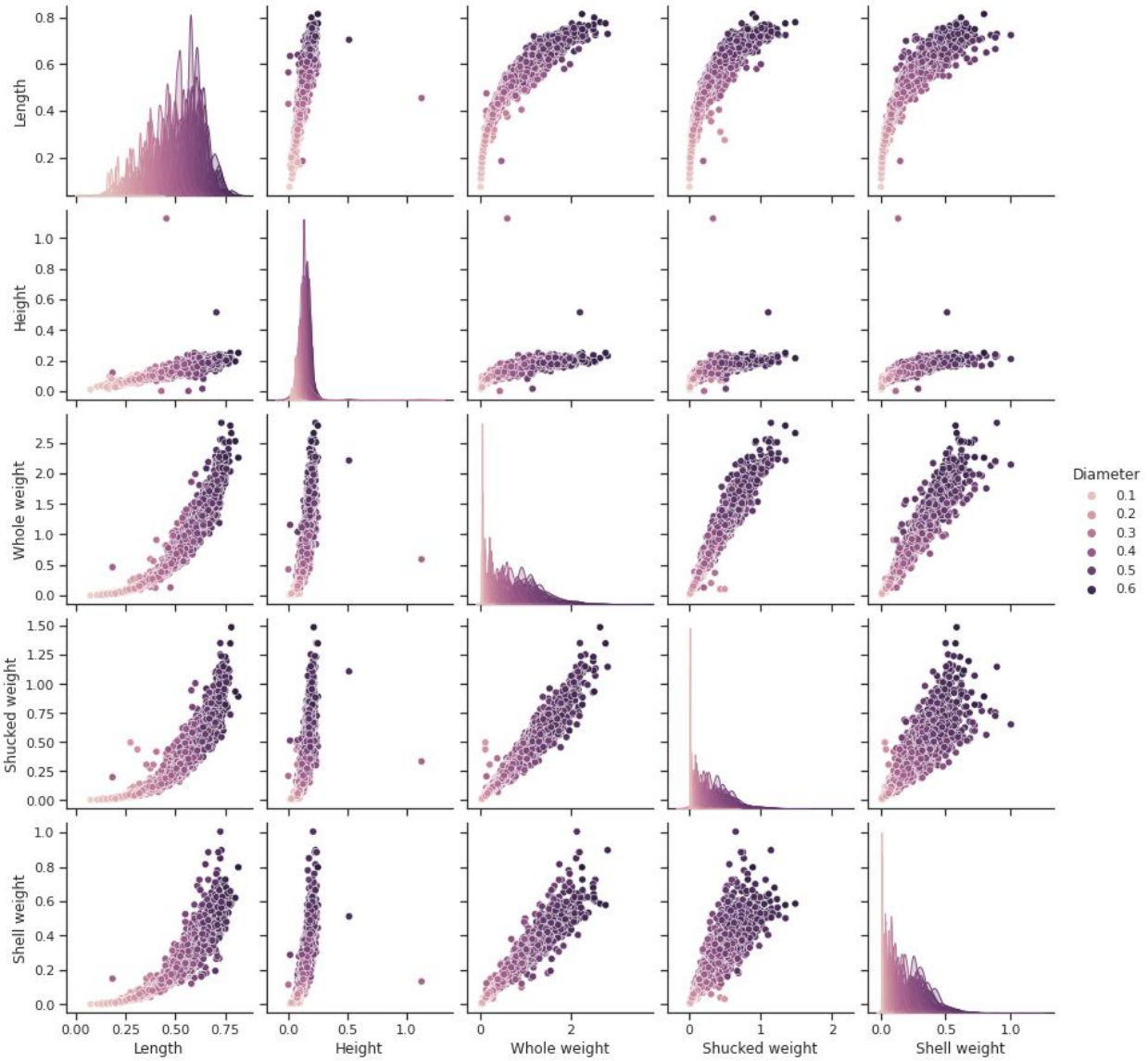
There are nine attributes that describe an abalone in the table: sex (nominal: male, female, and infant), length (a continuous value in millimeters), diameter (a continuous value in millimeters), height (a continuous value in millimeters), whole weight (a continuous value in grams), shuck weight (a continuous value in grams), viscera weight (a continuous value in grams), shell weight (a continuous value in grams), and rings (integer value). I will only be considering 6 attributes: length, diameter, height, whole weight, shuck weight, and shell weight.

Preprocessing

The preprocessing required was simple. I changed the column names to accurately reflect the values included within the column. I then removed the columns containing sex, viscera weight, and rings.

Visualization

I decided to represent this data using a scatterplot matrix in Python.



Analysis

From the visualization, we can observe a number of relationships between the different attributes of the dataset. The following attributes have a positive correlation with each other: length and height, shell weight and length, shucked weight and shell weight, whole weight and shell weight, and whole weight and shucked weight. Other relationships have more curious plots. The scatter plots created for the relationships between height and the three measured types of weight (whole weight, shucked weight, and shell weight) resemble a quadratic graph, indicating a quadratic relationship. We can observe based on the hue of the plot (which is based on diameter, as displayed to the right of the matrix and can be observed in the accompanying code) that the vast majority of diameter values in the data are greater than 0.3 mm. Diameter and length have a positive relationship, diameter and height have a positive relationship, and all of the relationships between diameter and the three types of weight are inverse.

Dataset: https://data.world/uci/forest-fires

Dataset obtained from data.world, an online free repository of datasets.

Dataset format: Tabular

The original dataset is a static, flat table which contains data about forest fires in the northeast region of Portugal including spatial data (x and y coordinates), temporal data (month, day) and meteorological data including temperature, wind speed, and rainfall.

Data Type: Item-based

The events are laid out in a tabular dataset and each record is identified by a combination of an x and y coordinate (within Portugal's Montesinho Park map) as well as the month and day, and there are several attributes describing various characteristics of the particular area relevant to the likelihood of a forest fire occurring there.

Attribute Types & Semantics

There are 13 attributes in the table: the x-axis spatial coordinate (integer ranging from 1 to 9), the y-axis spatial coordinate (integer ranging from 2 to 9) the month of the year (nominal, 'jan' to 'dec'), day of the week (nominal 'mon' to 'sun'), the FFMC index from the FWI system (decimal value ranging from 18.7 to 96.20), the DC index from the FWI system (decimal value ranging from 7.9 to 860.6), the ISI index from the FWI system (decimal value ranging from 0.0 to 56.10), temperature (in Celsius, a decimal value ranging from 2.2 to 33.30), relative humidity (a percent value ranging from 15.0 to 100), wind speed (kilometers per hour, decimal value ranging from 0.40 to 9.40), rainfall (mm/ m^2 , decimal value ranging from 0.0 to 6.4) and the burned area of the forest (measured in ha, decimal value ranging from 0.00 to 1090.84; the source of the dataset notes that this is an output variable). I will only be considering 5 attributes: day, the FFMC index, the ISI index, temperature, and wind.

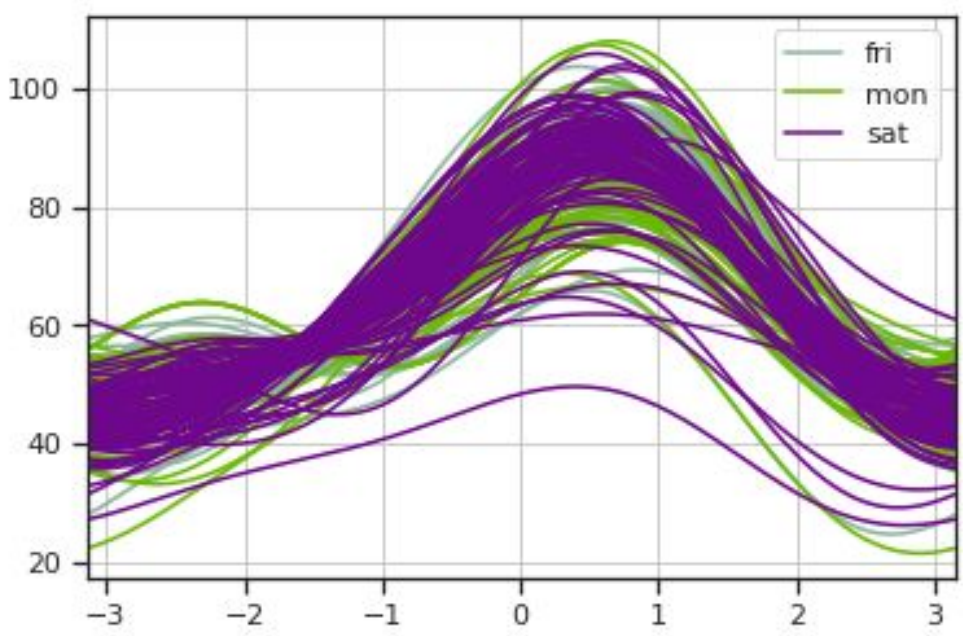
Preprocessing

To preprocess the data, I removed all of the columns that I did not decide to include in my visualization. Since I plan on visualizing the data with an Andrews curve, I'm limiting the

number of days to 3 (Friday, Saturday, and Monday) in order to not clutter the plot. To do so, I simply deleted the rows of the other days in Excel.

Visualization

I chose to visualize the data using an Andrews curve in Python.



Analysis

Andrews curves are useful for identifying any structure in data. We can observe based on the generated plot that the other attributes represented in our selected dataset (FFMC index, ISI index, temperature, and wind) can vary somewhat, but usually do not vary wildly and are fairly similar based on the day (for these three chosen days). Saturday contains the most outlying data.

Dataset: https://data.world/uci/wine

Dataset obtained from data.world, an online free repository of datasets.

Dataset format: Tabular

The original dataset is a static, flat table which contains data on various chemical characteristics of wine, including alcohol content, malic acid, ash, flavonoids, etc. The dataset is multivariate, as there are multiple attributes for each key.

Data Type: Attribute-based

The events are laid out in a tabular dataset and each record (a different type of wine) is not identified by a particular key. Each wine is solely identified by its attributes.

Attribute Types & Semantics

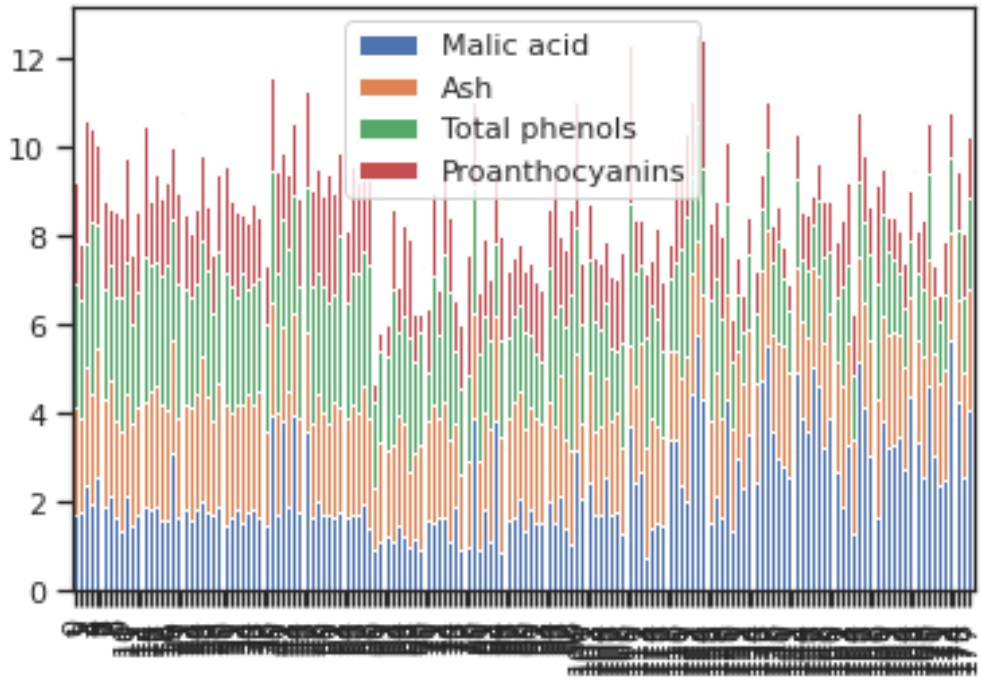
There are 14 attributes in the table: cultivar (which serves as the class identifier ranging from 1-3), alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, non-flavonoid phenols, proanthocyanins, color intensity, hue, 0280/0D315 of diluted wines, and proline. All values are continuous. I will only be considering four attributes in my visualization: malic acid, ash, total phenols, and proanthocyanins.

Preprocessing

To preprocess the data, I deleted all of the columns I did not previously choose for visualization in Excel. I also renamed the column header names to better describe the data they represent according to the dataset source.

Visualization

For this visualization, I decided to create a stacked bar graph in Python.



Analysis

Based on the created visualization, we can observe that the four attributes' values are not wildly different from each other. When we examine higher malic acid values indicated by a taller bar, we also notice a taller bar for ash, indicating a similarly high ash value. The height of the bars, in general, are fairly proportional in each stack. Naturally, there are height differences between different stacks of bars for the many different types of wine in this dataset due to the different chemical composition of each wine. It is not immediately obvious whether or not there are any outliers according to the visualization.