Assignment 2 Report Robert Hall

Dataset 1: Air Pollution Data – LSTM Multivariate 4-Variables

Description

The original data-set is from Kaggle and consists of information about air pollution including various variables that relate to the weather. This data is a subset of weather data and does not encompass the full weather data trends overall. The data is a time series data-set where it starts out on January 1st 2010 and goes to the date December 31st 2014. Since this is not a time-series analysis we are not using the time-series variable column. The data can be accessed from Kaggle via the website:

https://www.kaggle.com/rupakroy/lstm-datasets-multivariate-univariate

Dataset Format, Type, Attribute Types and Semantics

The data-set from Kaggle is a flat table which consists of the time-series data for air pollution which includes the pollution(calculated from PM2.5 concentration), the dew point, the temperature, the pressure, wind speed, wind direction, snow and rain. The data used in this analysis was the columns consisting of: Pollution , Dew, Temperature, and Pressure. The columns used are float values where specifically the dew and temperature can take on negative values. We are not including the following columns in this analysis: Wind Speed, Snow, Rain, and the Date. The reason for not including these column variables is that they are not going to provide us much information about the analysis and might throw off any observations we can have due to the data-points being on a different scale compared to the others. PCA could be used for these points but is not applicable for this analysis, however. The "pollution" column is the PM2.5 concentration of pollution in the air. The "Dew" column is the dew point value for the air. The "temp" column is the temperature of the air reported (in celsius). The "press" column is the pressure value reported for the air. Lastly the "wnd_dir" column is the direction the wind was blowing at the time of data recording. The most important information in this analysis is the mentioned columns above and that the time-series column is not used in this analysis due to this not being a time analysis. The columns not used in this analysis are: "wnd_spd", "snow" , "rain" and "date".

Preprocessing

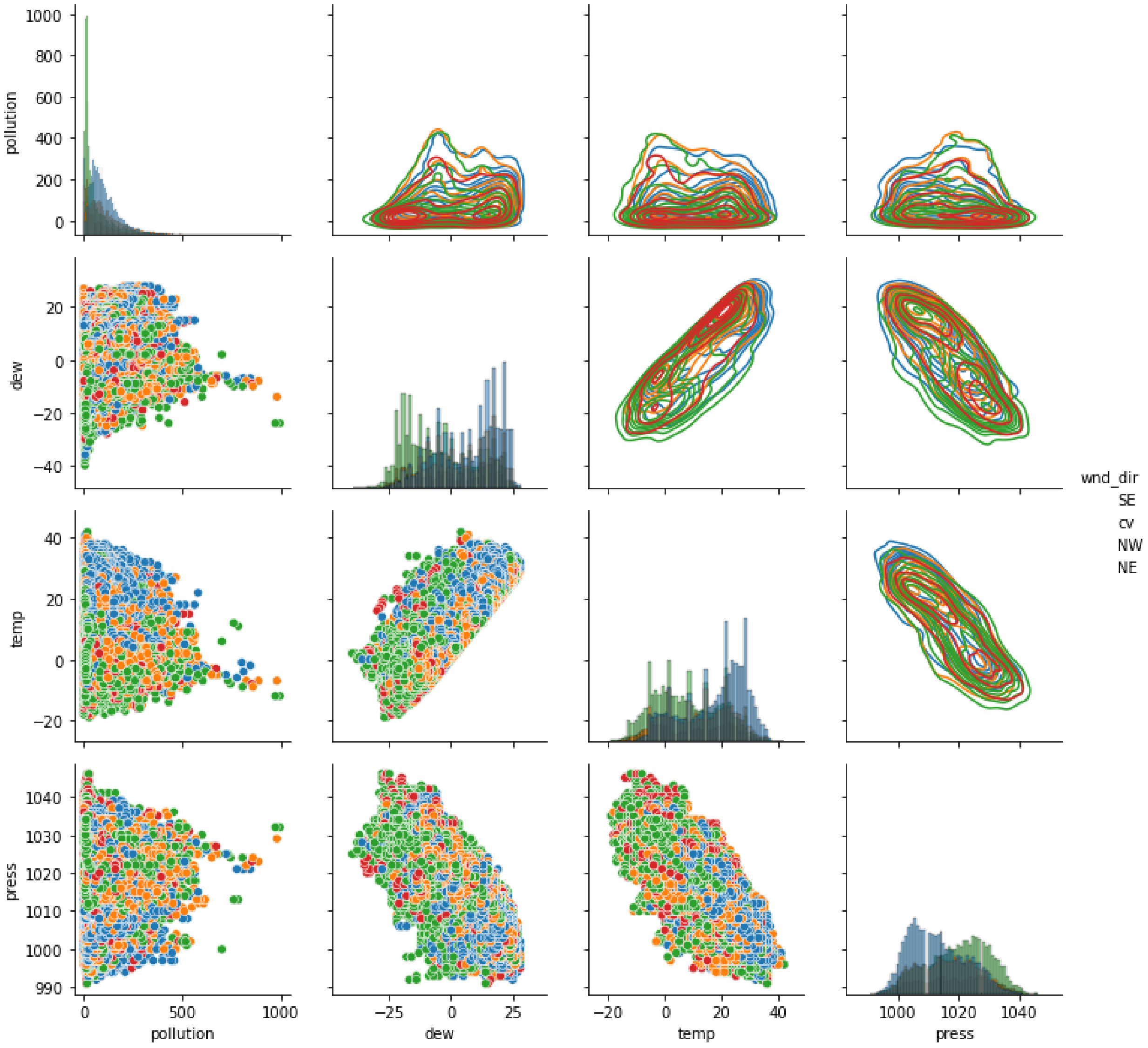
The data whenever loaded into the system pulls all variables and columns, the data was split into the columns that we are using for the analysis which were: *pollution , dew , temp , press , and wnd_dir .* For this reason we subset the data frame created from reading in the .csv file to the proper format we need, below is a sample snippet of how this was achieved:

```
air_pollution_important =
df[['pollution','dew','temp','press','wnd_dir']]
```

Since we subset the data this way, we only grab the correct columns we need for the analysis. It is also important to note there were no erroneous values in the data set or missing pieces which is good since we will not have to filter those out for the analysis. The reason we included the wind direction column is for the analysis to be able to color-code the values the wind was blowing to see if the direction has somewhat of an effect on the outcome of the data. The data was then placed into a Seaborn PairGrid function to be able to create a diagonal visualization to plot the points. (Visualization in next section) . On the diagonal of the PairGrid plot is a barchart with the data-points aggregated based on the values (pollution -

dew , temp - dew , etc.) Whereas on the upper section of the diagonal a Kernel Density Estimate was drafted to show the boundaries of the data points. On the lower section of the diagonal, however, a scatterplot of the data points were created to show the overall trends. It is important to note the KDE is showing us the estimate of the boundaries of the data and the scatterplots show us the actual plotting of the data. The data was then color-coded using the wind-direction variable for easier viewing.

Visualization



Analysis

From viewing the figure from the PairGrid plot above, one can see a couple things for certain. The first thing is the clear relationship between the temperature and the dew. It shows in this analysis that whenever the dew point increases, the temperature also increases in a linear fashion. The same can be said for the pressure - dew point , and the temperature - press relationships except the inverse such that they have a negative effect on each other. This means whenever the dew point and temperature increase the pressure gradually starts to decrease in turn. We can tell these statements are true from both the lower section of the diagonals scatter plots graphs and the Kernel Density estimate on the upper section showing both relationships as equal. The Kernel Density Estimate was included for this analysis because I believe it shows the relationship between the dew, pressure, and the temperature in a better light rather than a mess of data points in the scatterplot (even though the scatter plot shows the same positive vs. negative relationship the KDE also can show the boundaries).

Another important piece of information that can be extracted from this visualization is on the first column of plots, specifically the pollution variables compared to the dew, temperature, and pressure. We can see it sort of follows a semi-normal distribution curve but has such values that increase into the upper echelon of above 600 PM2.5 concentration to where it creates a visible tail where some data points follow it as well. This “tail” increase is not true for all the values plotted against it, meaning that for temperature the tail is created when the temperature is between 0 and - 20 degrees celsius , the tail is created when the dew point is between 0 and -20 degrees celsius (This is most likely true since dew point and temperature clearly are related as shown), but the pressure’s tail is created between the values of 1020 - 1030 psi. Further analysis of why this is the case specifically for pollution will require more data and further analysis of locations.

Data-Set 2 UCI’s Auto MPG data 5-Variable (PCA)

Description

The data-set used for this analysis is the data of various cars including information about the MPG, cylinders and the origin of the car. The data used for this originally comes from a survey done at the Statlib library but is now used for UCI’s machine learning repository. It is important to note that the data used in this analysis is just a sample of car data and does not encompass the full car market, rather just a small subset of the overall data. The data can be accessed from the website via:

https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Data-Set Format, Type, Attribute Types and Semantics

The data from UCI consists of a flat table consisting of information about cars which includes: MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year, Origin, and Car Name. Each row in the data-set corresponds to a single car instance where each row has the respective information regarding that car. The only columns used in this analysis were: MPG, Cylinders, displacement, weight, and acceleration. The “MPG” column is a floating point variable that is the miles per gallon the car gets. The “Cylinders” column is a float that is representative of the number of cylinders the car has. The “displacement” column is a float and the combined volume of the cylinders. The “weight” is a float where the column corresponds to the weight of the car itself. Lastly, the “acceleration” column is the speed in which the car can go. There are a couple of variables that were not used including: car name, horsepower, model year. The “origin” column is the location the car is from. The only information the data-set and the source gives about the numbers the origin corresponds to is that it is a discrete number. So it is expected that this origin column is the target and our analysis filters and shows these areas.

Preprocessing

Data set was initially read from the corresponding .csv file and placed into a data frame. Some work had to be done to create a format where we could standardize it and create the principal components from them. An array was created to hold the features we needed, then using pandas we subset the data into the features data-frame and the target data-frame (origin column). After we created the respective data-frames we used the StandardScaler function from sklearn to standardize the data in the features section to prepare it for the PCA. The sample snippet below shows how this was achieved:

```
features = ['mpg','cylinders','displacement','weight','acceleration']
x_features = auto_mpg.loc[:,features].values
y_target = auto_mpg.loc[:,['origin']].values
x_standardized = StandardScaler().fit_transform(x_features)
```

After standardizing the features section we created a PCA class from sklearn with the number of PCA components requested to two. We then transform the previously standardized feature set to create our array of principal components. Whenever this array is created it is then used to create a new dataframe that consists of the: 1st principal component, the 2nd principal component, and the origin column that was originally present in the data-set. (Important to note that the origin column was not standardized since it is our target variable and is used for separation on the visualizations).

With this creation we have our data-frame that has both principal components and the origin column ready for use in our visualizations and analysis. No other work needed to be done on the initial data-set to filter out any erroneous values or improper formatted sections. We just needed to transform that original data into a usable standardized format.

Visualizations

Figure 1. PCA 1 and PCA 2 Scatterplot

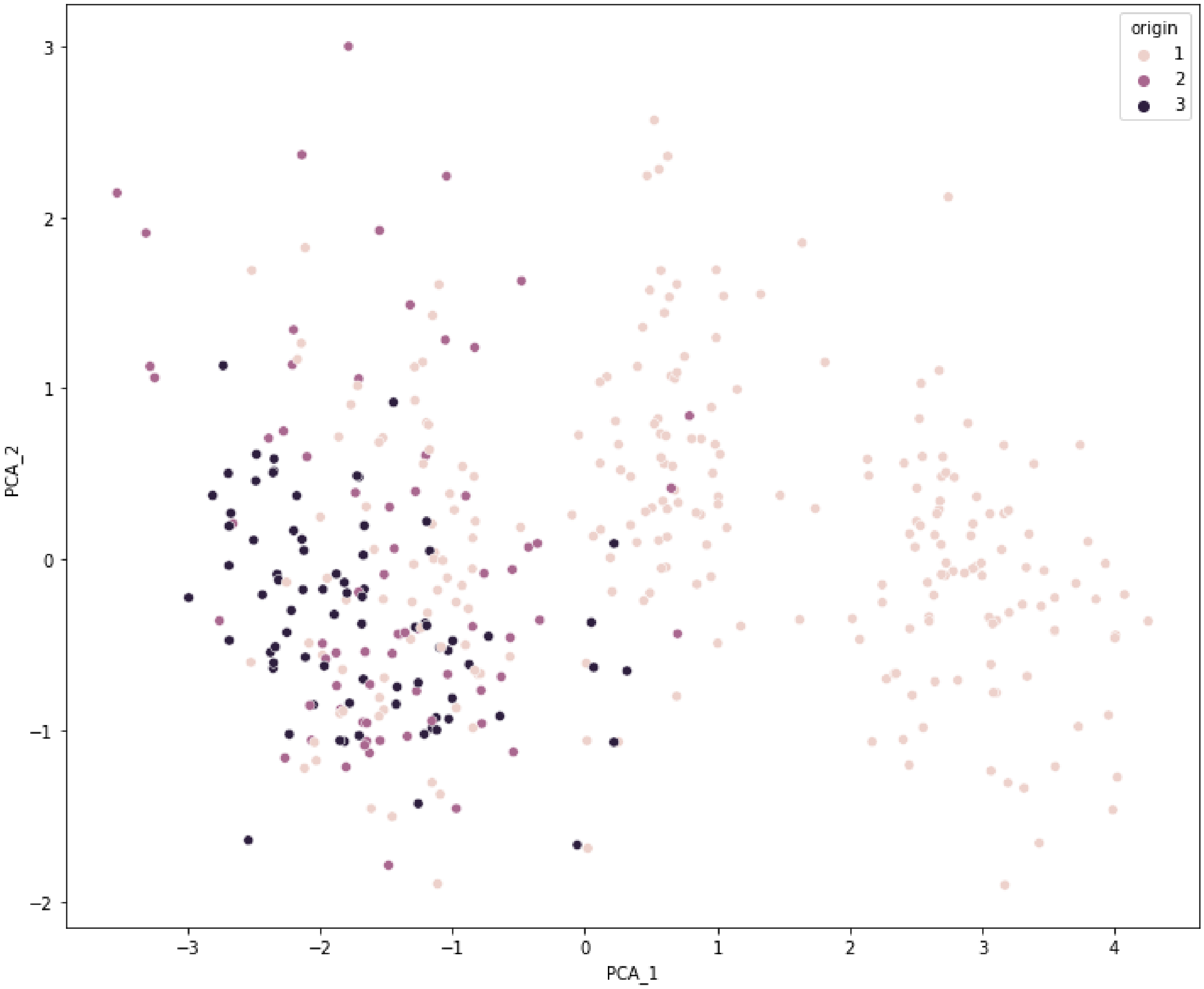
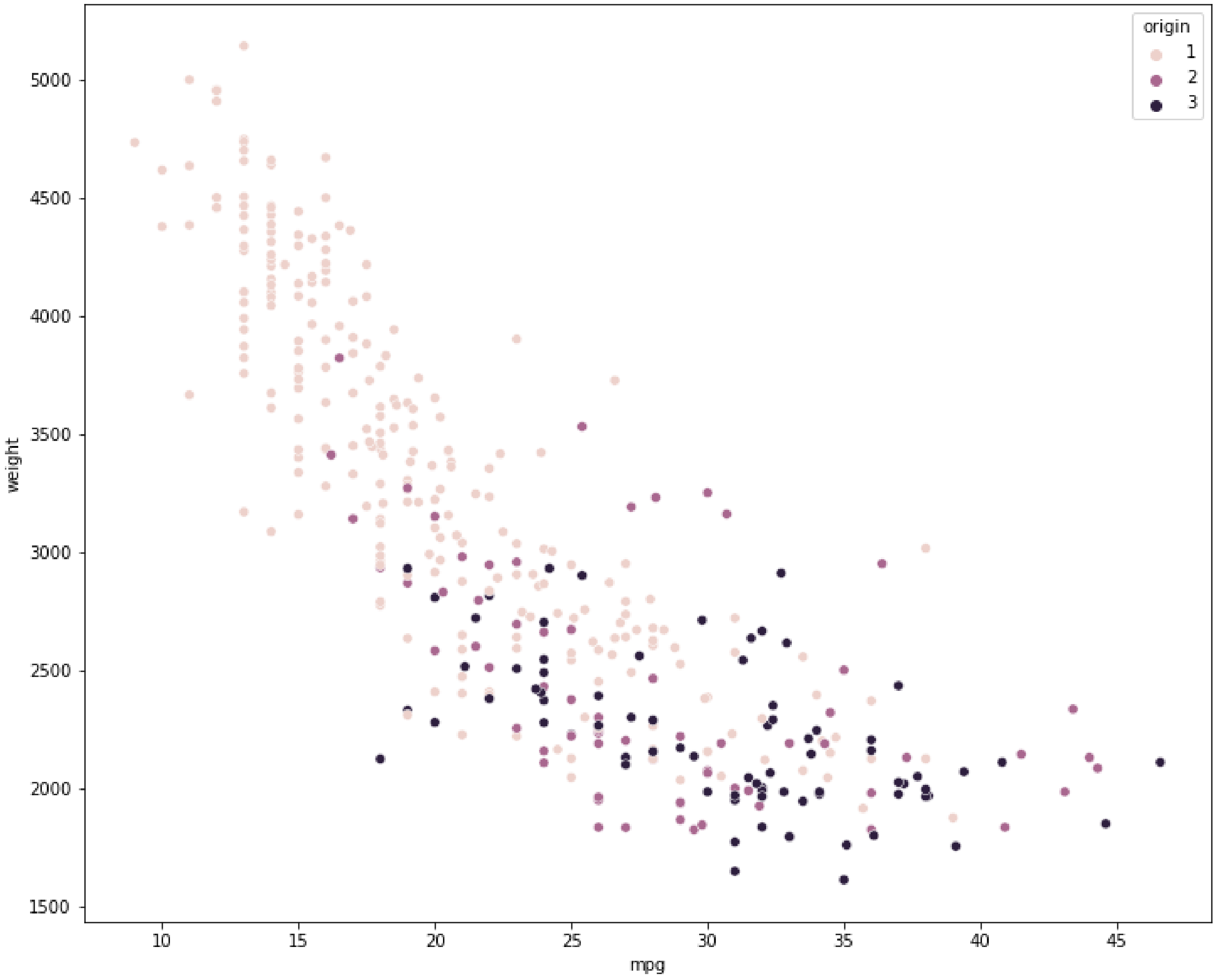


Figure 2: MPG and Weight Scatterplot



Analysis

From viewing the produced visualizations above one can see a couple things, from figure 1 one can see that origin 3 has quite a bit of data-points closely clustered together when reaching below 0 on the first principal component and grouped between 1 and -1 on the second principal component respectively. The same can also be said for origin 2 where they are closely related in about the same areas as origin 3. Origin 1 is where things start to vary quite a bit where one can see three distinct clusters on the PCA_1 axis : -2 - 0 , 0 - 2, and 2 - 4 respectively. Whenever we plot MPG and weight on the scatter plot on figure 2 we can see similar grouping for origins 2 and 3. An interesting piece to see is that for origin 1 has similar clustering when increasing the MPG and lowering the weight, but also has quite a bit of points on the plot on the higher-end weights above 2000lbs. From what I can see from both of these graphs I am assuming that origin 1 normally creates vehicles that are used for commercial purposes (whether it be transporting goods or the vehicles are used in developments) which would reflect in both the weight and MPG category. A similar statement can be made for Origin 2 and Origin 3 but the inverse where it appears they make smaller more fuel efficient vehicles. The statement of similarity for origin 2 and origin 3 is founded from the fact they are both clustered together on Figure 2 and the computed standardization has them clustered together when plotted against each other.

Data-Set 3 UCI's Abalone Data 6-Variables

Description

The data-set used for this analysis is measurements of Abalone (sea shell) consisting of various information about their characteristics including attributes such as the weight, length, diameter and others. It is important to note with this Abalone data that this does not encompass the whole data for Abalone rather this is from a study that looked at the species near them. The Abalone is a common name for marine gastropod molluscs where they are harvested and farmed for both food and decorative items. The data can be accessed from the website:

https://archive.ics.uci.edu/ml/datasets/abalone

Data-Set Format, Type, Attribute Types and Semantics

The data-set is a flat table consisting of multiple columns of Abalone data where each row in the table consists of one Abalone and it's characteristics. This meaning each of the data points relates to a single instance of Abalone characteristic defining. For this analysis we are looking at the following column characteristics: Length, Diameter, Height, Shucked Weight, Viscera Weight, and Shell Weight. The: Length, Diameter, and Height variables are floating point numbers that relate to the size and are measured in millimeters. The: Shucked Weight, Viscera Weight, and Shell Weight are also floating point numbers that relate to the weight and are measured in grams. The length variable is the longest shell measurement, the diameter is the perpendicular to the length, the height is measured with the meat in the shell. THe Whole weight attribute is the weight of the Abalone with all the internals. The Shucked Weight attribute is the weight of the meat of the Abalone after it has been harvested. The Viscera Weight is the gut weight after the Abalone has been dried and bled. There are other columns and attributes that are present in this data-set that where not used for this analysis and they are: Sex and Rings. The sex column relates to the gender of the Abalone and the Rings column relates to the age of the Abalone. These are not used in the analysis and are shown respectively in the visualizations in the later section.

Preprocessing

The data-set was read from the corresponding .csv file and ported into a dataframe using pandas. From there the data-frame was subsetted into the proper columns that we are using for the analysis which were : Length, Diameter, Height, Shucked Weight, Viscera Weight, and Shell Weight. This data-set did not contain any erroneous values or instances of missing data so no techniques were used to fix these issues. From the data-frame that was created the visualizations were created using a histogram plot to show the frequency of the continuous data. Initial attempts were tried with both line-plots and violin-plots but when graphed they appeared to be messy and did not give any important information for this analysis. Histogram plots were plotted using the Seaborn package which is a high-level API for mathplotlib. Added onto the histogram plots were the KDE which shows the density and curvature of the distribution of the data. Two plots were created where one uses a stacking technique to show individual distributions and a non-stacking technique to show them overlapping. Both were included for reference and can be found in the visualization section. No other preprocessing techniques were used since the data was fully intact and no changes were needed.

Visualizations

Figure 1. Non-Stacked Abalone Distribution.

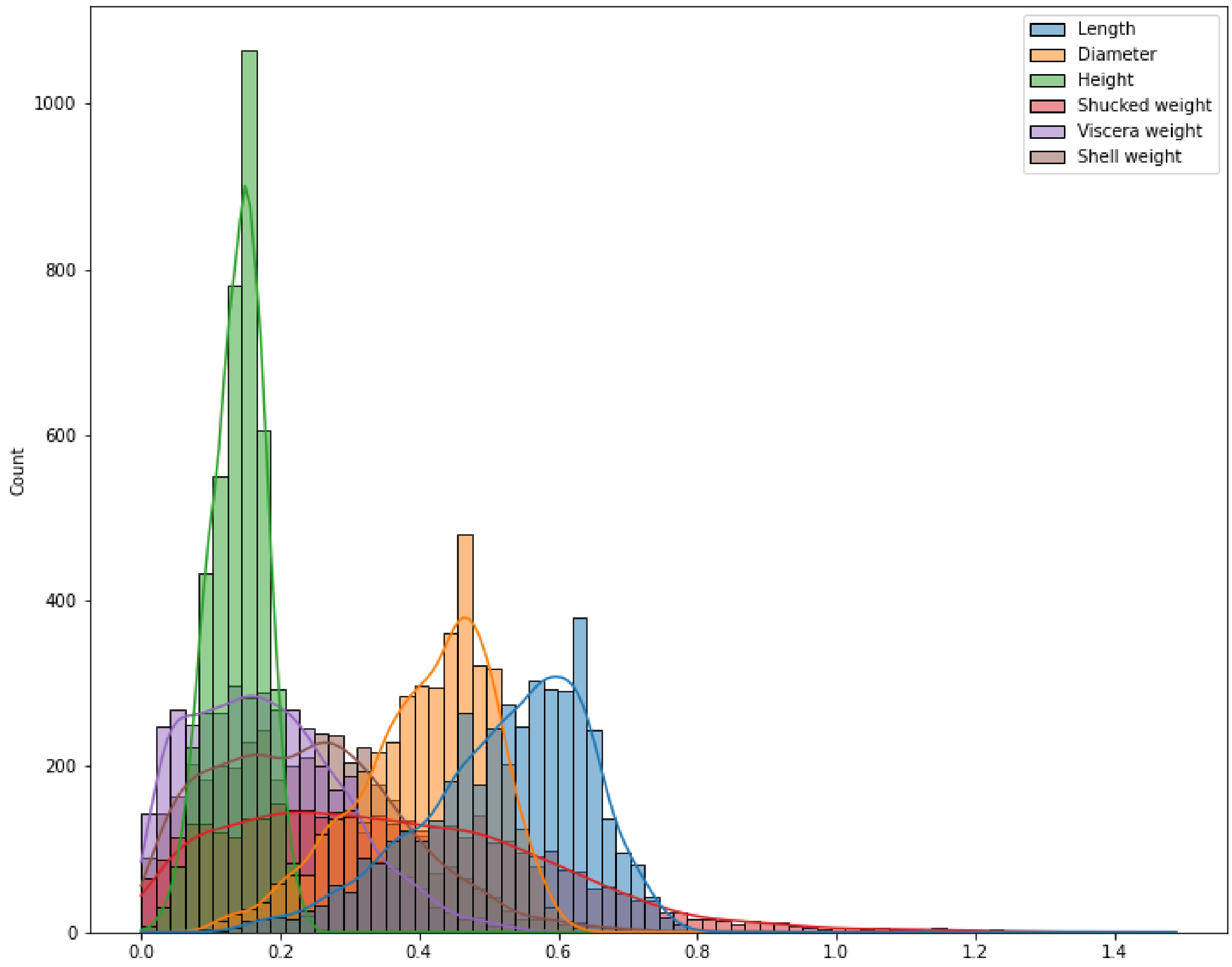
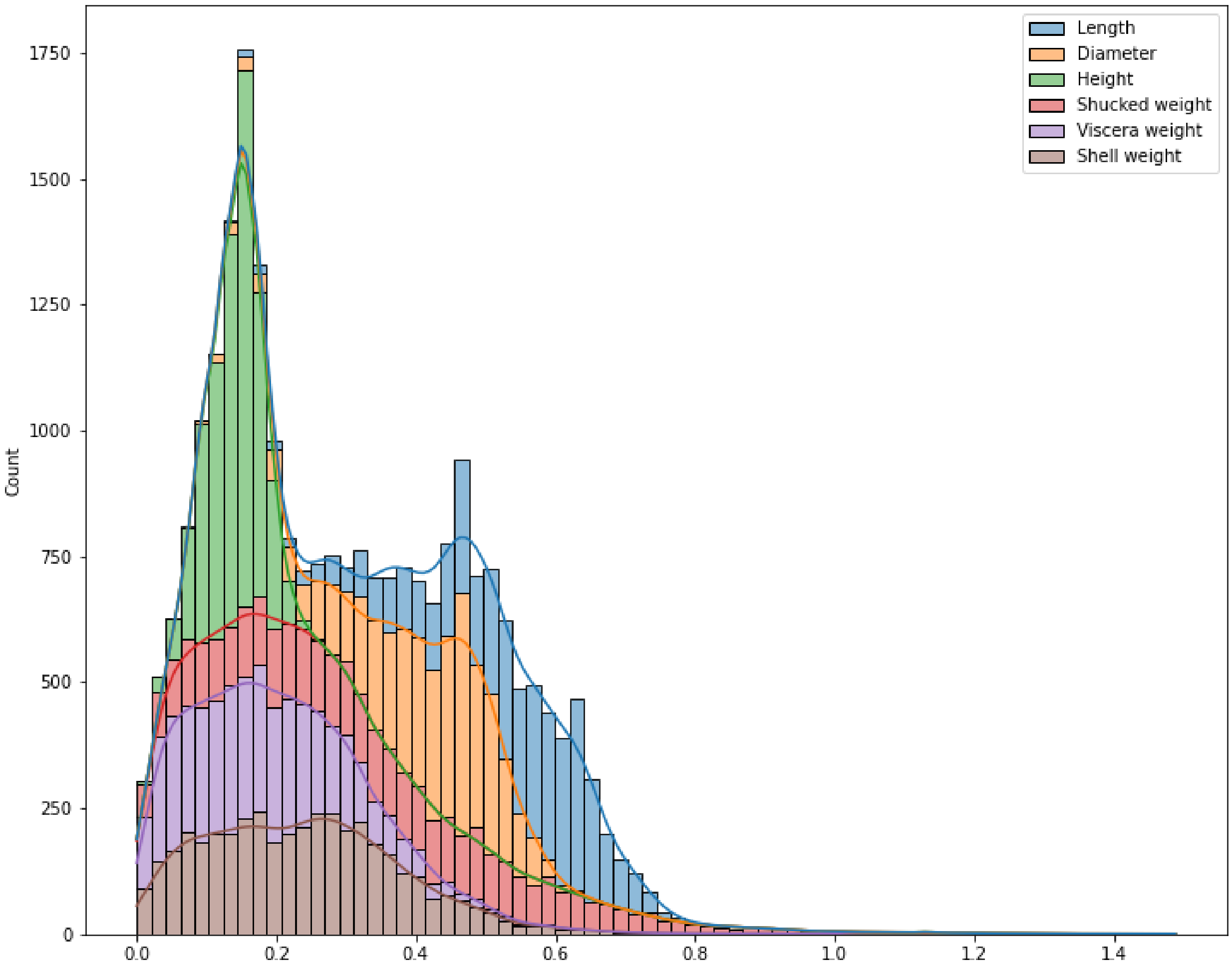


Figure 2. Stacked Abalone Distribution.


Length
Diameter
Height
Shucked weight
Viscera weight
Shell weight
Count
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Analysis

From viewing the visualizations above one can make a few assumptions. The first assumption that can be made is that most of the data-points follow a normal distribution and are spread out evenly. Further points can be made for the height variable of how many Abalone fall between the ranges of 0.0 - 0.2 mm and rarely go past 0.2 mm average in size. It is important to note that this height variable is the height included with the meat in the shell of the Abalone. A similar assumption can be made with the diameter variable how the diameter of the Abalone can be seen between the ranges of 0.1 - 0.6 mm frequently and follow the same pattern as the height in which the quantity of Abalone with diameters above 0.6 mm dropping frequently to where they are not plotted anymore (presumably due to no instances). The length variable is also important to look at on the visualizations due to it being distributed somewhat evenly across the ranges of 0.1 - 0.8 mm in length.

The same statements previously made about the height and diameter can also be made for the length in which the number of instances drops when approaching and going past the 0.8mm length. One of the more interesting points that can be found from the graph is the distribution of the Shucked weight variable (weight of the meat itself from the Abalone) , where it follows a left leaning distribution and has instances past the 0.8 grams measurement in numbers not seen in any other measurement. It is important to note with these visualizations that the weight variables are measured in grams , but since these are metric measurements they fit on the graph well and no extra work was needed to make them look presentable.