Lorena Vasquez
CIS 3120
Project 2: Part B
Data Visualization Page Report

Project Two - Part B - Part 1

To begin the project, there are necessary libraries to bring into the program. One of them will be *'Import pandas as pd',* this helps the program to be able to read the csv file into a dataframe. Next to *import matplotlib.pyplot as plt* for any necessary basic graph plotting. Next is to *import seaborn as sns* to form some of the graphs at the end of the program. Then we remove the warnings from the output by importing *import warnings, warnings.filterwarnings("ignore")*.

The dataset that I chose came from Kaggle.com. The dataset is on different species of fish with their relevant weight, height, weight and length. To start, I will call the dataframe fish by the following command of using pandas to read the csv file. I check the data frame by *fish.head().* Next *fish.shape* to see the fish shape, which shows the number of columns and rows of the dataset. In the first scatter plot, it is to show the correlation of Weight and Width. There seems to be a positive correlation, where as height increases so does the width. The next scatterplot shows the correlation between the weight and height, again there is a positive correlation as weight increases height increases as well. After a certain point in the graph it does, level off and doesn't seem to increase much more. The last scatter plot then shows the correlation between the weight and length3, where there is a positive correlation. The joint plot shows the correlation between height and width. In the center it shows the correlation between the dependent and independent variable. The top graph shows the distribution of the independent and the one to the right, shows the distribution of the dependent variable. A facetgrid of the scatter plot made it easier. The facetgrid easily shows the differences between the species. It shows the correlation of the species and how they are correlated. The first box plot shows the distribution of the two variables of the weight and the height by species. In this plot it shows that on average the *Bream* species is bigger in height. We can also see that the *Perch* species has a long range of values. The last observation that I saw, is that the *Smelt* species, on average, was smallest compared to all the species. The next boxplot is to show the distribution of the species and width. In this case on average the Whitefish was on average the wider fish. Like the previous box plot the Perch species also has the largest range in width. Like the previous graph as well the smelt species also on average had the smallest width compared to the other species. The next box plot shows the distributions of the species, comparing the lengths across the species. On the contrary, for this graph Pike species on average were significantly greater than the rest of the species. And like the two previous graphs, the smelt species again on average were smaller in length.

Violin plots are like box plots. Like box plots you can see the distribution in ranges, and you can also see the mean and the interquartile ranges in the violins. The difference between the violin plots and the box plots is that they represent the variable distribution comparison across the species of fishes. So for each variable, they will run it across all of the species. That is why some of the violins are curved in and out or sparse. The next plot is the violin plot with the swarmplot. On its own the swamplot does not rely on the violinplot. It is complementary to the violin so it enhances the data shown from the violinplot. The swarmplot just shows the distribution of the ranges. Next plot is the sns.pairplot that uses the seaborn kdeplot. The KDE plot is a way to visualize the distribution of the observations in a dataset, almost like a histogram. The plot shows the data, using the continuous probability density curve, that is why we can see a curve bell shape within the graph. Smelt being the one to have a high density with a low height and the remaining species to range through the remaining of the ranges. Smelt reaches to a close 1.2 density and the rest of the species do not pass a .4 of density. The next graph is to show a seaborn pairplot. It generates a grid of axes where each numeric variable in the data is shared across the y axis, It builds two basic figures, histograms and scatter plots. The histogram on the diagonal line allows us to see a distribution of a single variable and the upper and lower triangles show the relationship between two varaibles. Going across all the variables, they all have positive correlations, when comparing both variables.

The next graph shows the boxplots grouped by species. So for each variable there will be a boxplot. The last graph is a parallel coordinate. Parallel coordinates help to understand datasets with numerous variables. In the plot, the different variables are presented with vertical parallel lines. Because parallel coordinates are specifically used with numerical values, I needed to remove the variables that were of string, and to keep one of the lengths instead of all of the lengths columns.

Sources:
Website used for data visualizations:
Kaggle:
https://www.kaggle.com/datasets/aungpyaeap/fish-market