Homework 4 (Neural Networks)

COSC 3337 Dr. Rizk

Problem statement and Datasets

- Your task for this homework will be to construct a neural network for both a regression and classification task. We'll be using the california dataset for the regression portion, and iris for classification. By the end of this hw you should be comfortable using sklearn's neural network on a dataset.
- The data we'll be using for this hw will come directly from sklearn. Sklearn has built in datasets that we can download by running a single command.
- Datasets: 2 dataset
  - California housing dataset
  - Iris dataset
    - sepal length (cm)
    - sepal width (cm)
    - petal length (cm)
    - petal width (cm)

California Housing Dataset

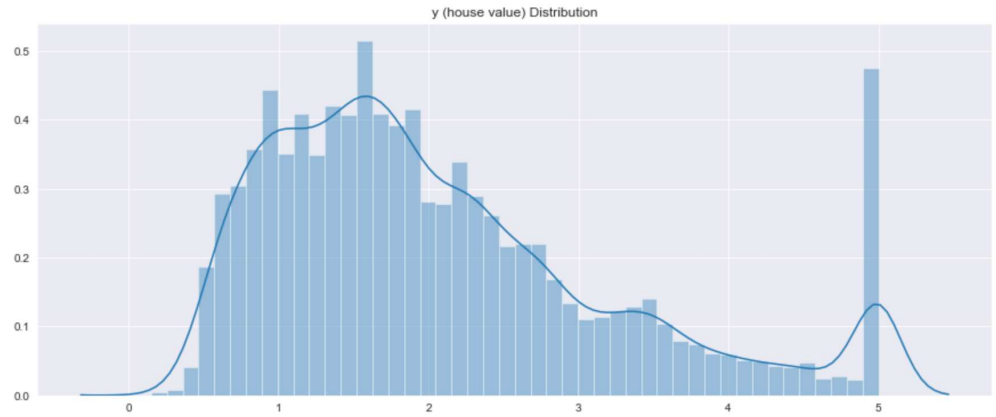
- Attributes
  - MedInc
  - HouseAge
  - AveRooms
  - AveBedrms
  - Population
  - AveOccup
  - Latitude
  - Longitude

Part1 Reading and Understanding the Data (Regression Part)

- Import ***fetch_california_housing*** from sklearn. We then proceed to saving this data into data frame (X), and our target/labels in a variable (y). Use these variables (X data frame and y) for the regression portion of this homework.
- TODO:
  - Print some basic statistics of your data.
  - Print some general information about your data using pandas.
  - By looking at our previous output, are there any missing values in this dataset? How do we know?

Part2 Visualization

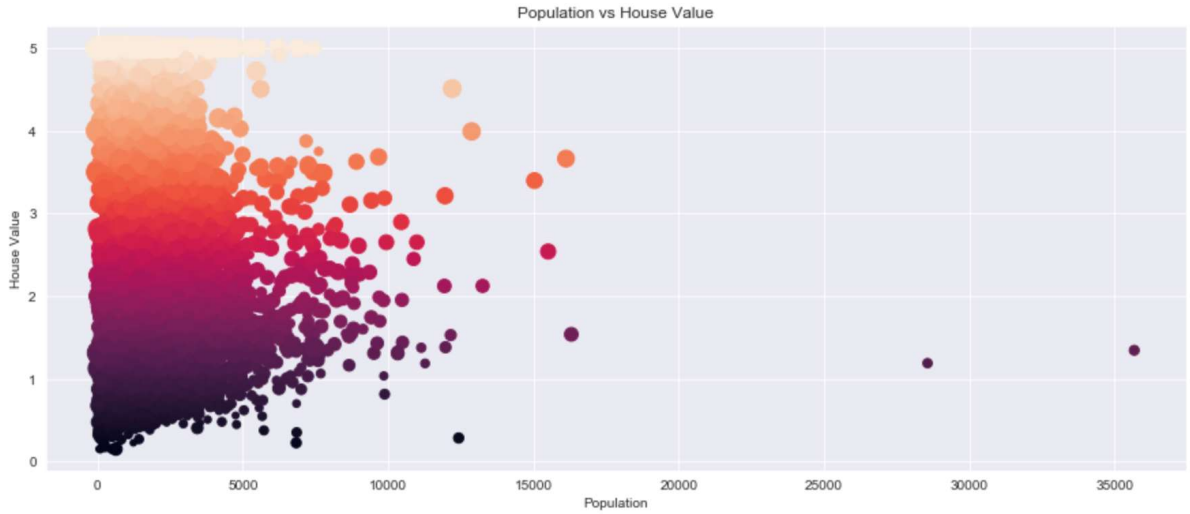
- Create a histogram of the median house value (y)



- Answer:
  - What can you conclude from the plot you created?

Part2 Visualization

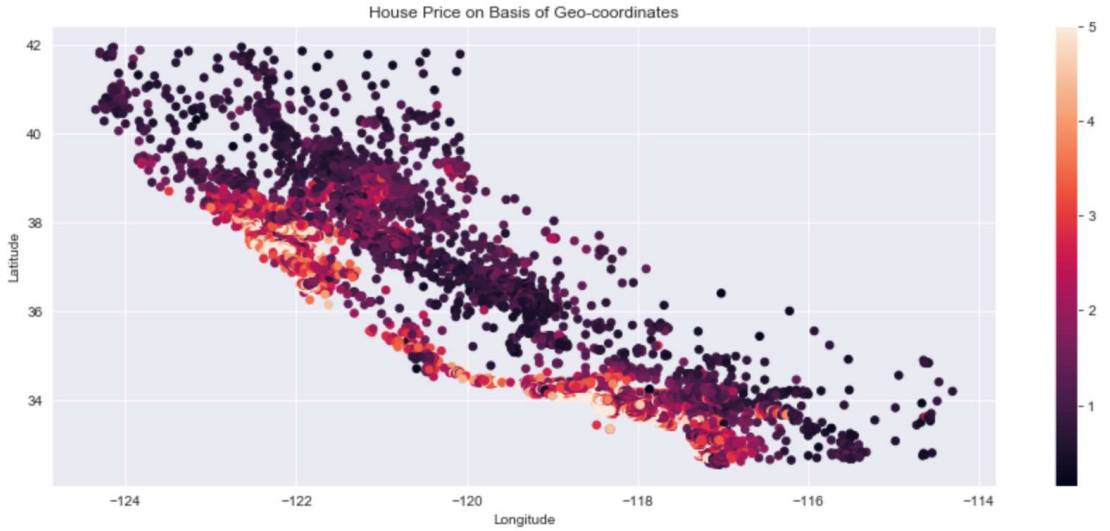
- Create a scatter plot of the Population vs. House Value (y).



- Answer:
  - What can you conclude from the plot you created?

Part2 Visualization

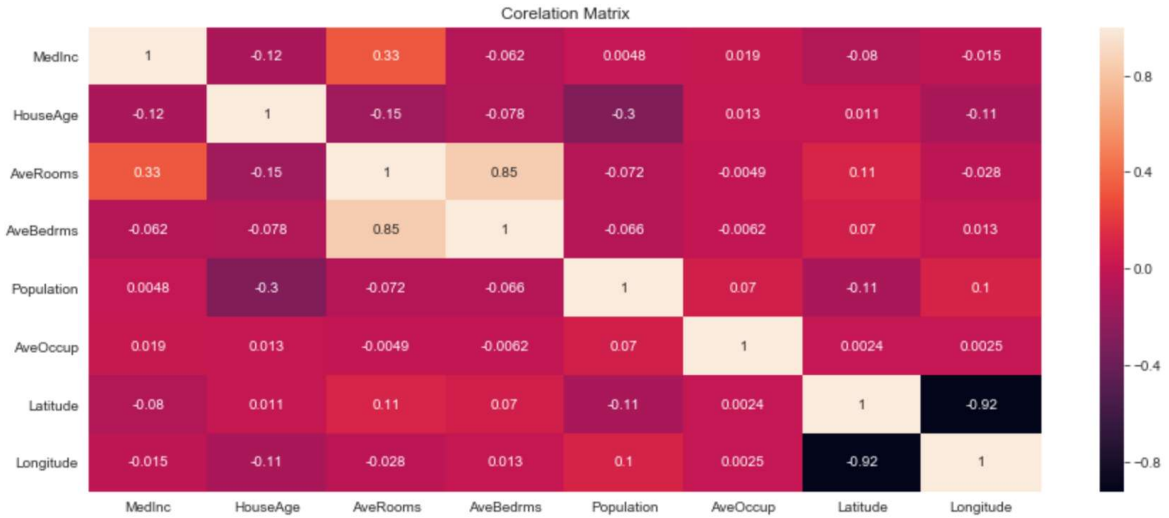
- Create a scatter plot of Longitude vs. Latitude, and color each point by its corresponding house value (y).



- Answer:
  - What can you conclude from the plot you created?

Part2 Visualization

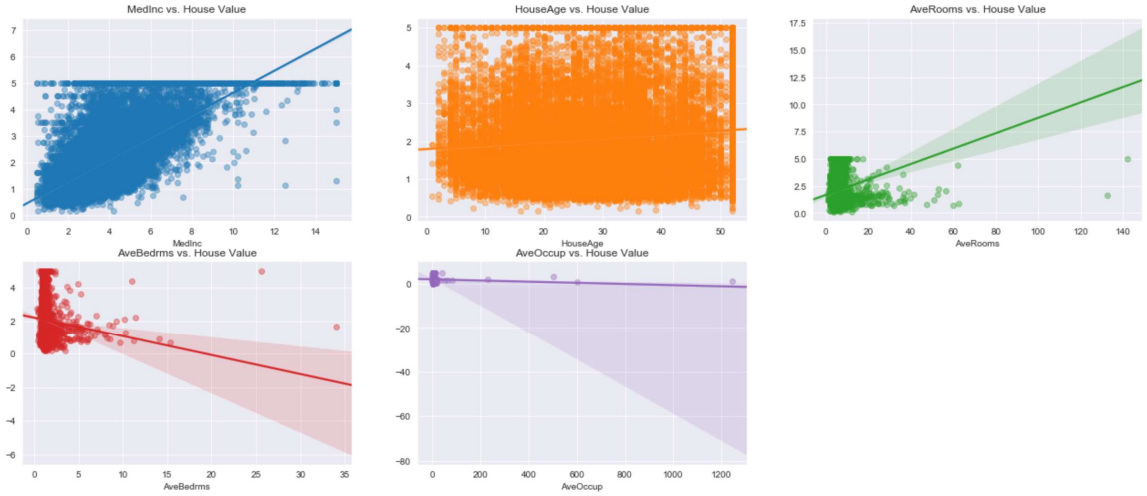
- Create a heatmap of the data.



- Answer:
  - What can you conclude from the heatmap you created?

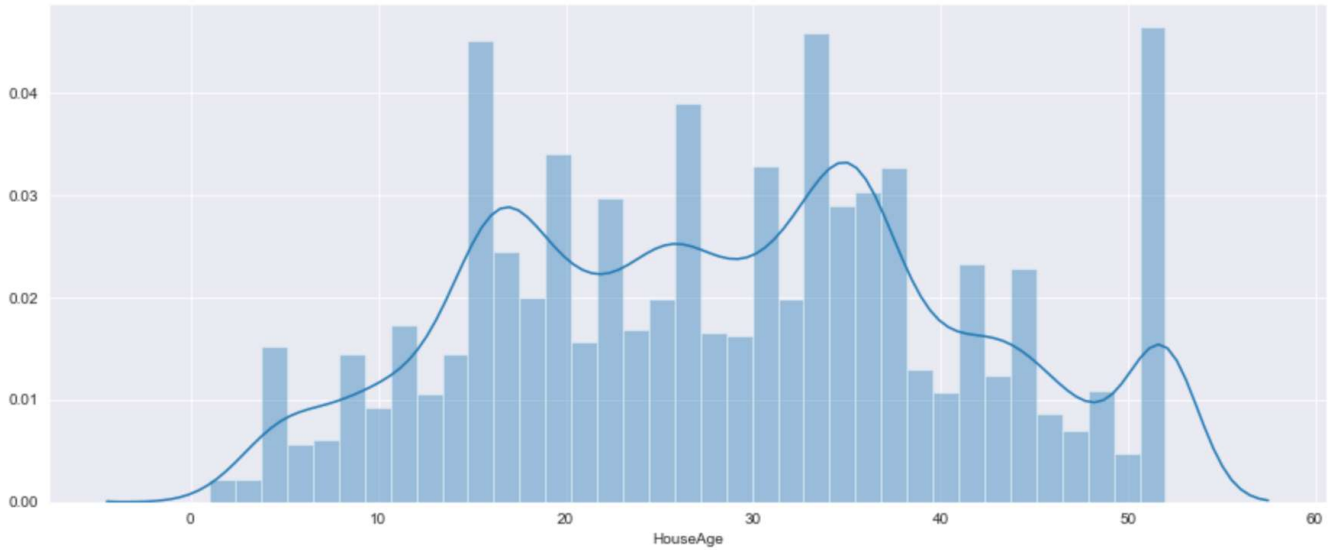
Part2 Visualization

- Create the following plots:
  1. A scatter plot of MedInc vs. house value (y)
  2. A scatter plot of HouseAge vs. house value (y)
  3. A scatter plot of AveRooms vs. house value (y)
  4. A scatter plot of AveBedrms vs. house value (y)
  5. A scatter plot of AveOccup vs. house value (y)
- **Answer:**
  - What can you conclude from the plots you created?



Part2 Visualization

- Create a histogram for HouseAge.



- Answer:
  - What can you conclude from the plot you created?

Part3 Model Creation and Evaluation

- Perform the following
  1. Scale the data
  2. Split your data into a training and testing set, with test size of 0.30
  3. Create a neural network for regression using sklearn. See sklearn.neural_network.MLPRegressor if you need help. Use a hidden layers of size 130 -> 64 -> 32 -> 16, and activation function 'relu'
  4. Run the test data through your model to obtain predictions. Save these predictions into a variable called 'predictions'.
  5. Print the R^2 of your model.

Part3 Model Creation and Evaluation

- Answer
  - How did your model perform?
  - Try experimenting with different hidden layer sizes.
  - Try using gridsearchCV to find the best set of parameters.
  - Write the best setting working.
  - Also, hopefully you noticed that our data contains outliers. You can also try removing them to see if that improves performance.
- Instruction(for submission)
  - Save any cells you use for experimenting so that we can see what you tried.

Part4 Reading and Understanding the Data (Classification part)

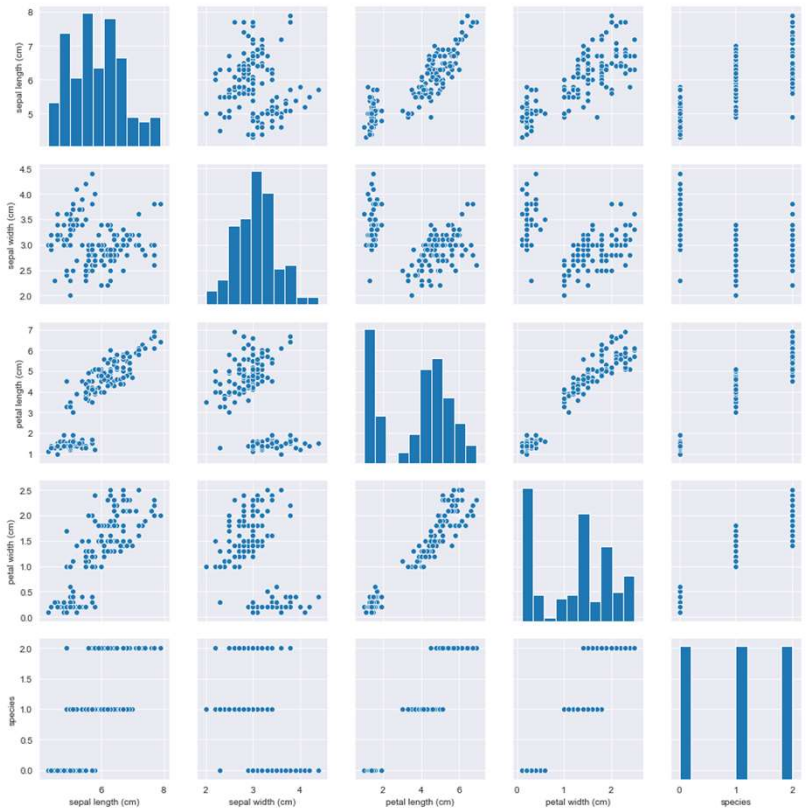
- Import Iris dataset
- TODO:
  - Print some basic statistics of your data.
  - Print some general information about your data using pandas.
  - By looking at our previous output, are there any missing values in this dataset? How do we know?

Part4 Reading and Understanding the Data (Classification part)

- Create a pairplot of the iris data. We'll temporarily add y to our X data frame (name it species) so we can easily access the species type. Just remember to drop the column once you're finished.
- TODO:
  - Create a pairplot of the iris data

- Answer:
  - What can you conclude from the plot you created?

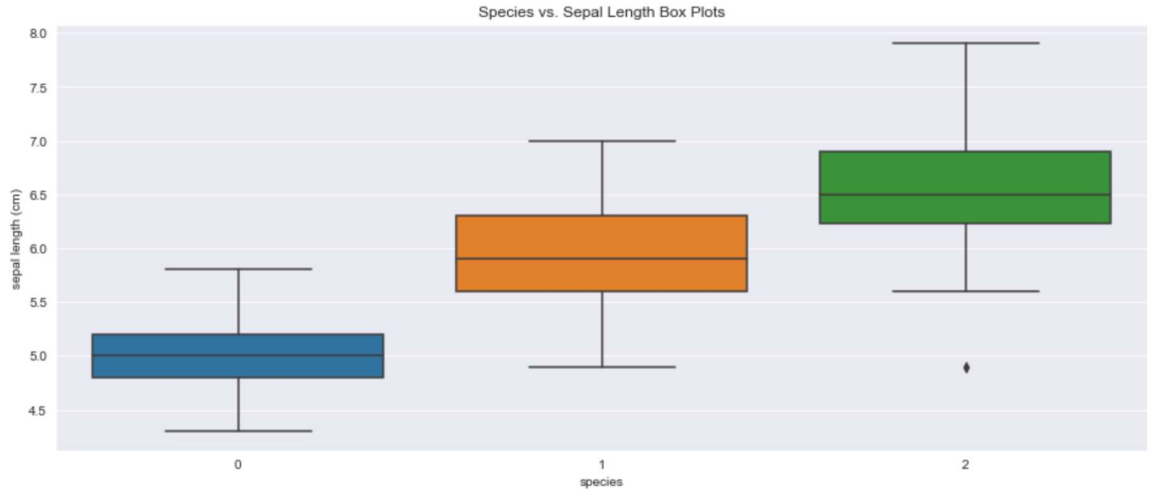
Part4 Reading and Understanding the Data (Classification part)



Part4 Reading and Understanding the Data (Classification part)

- Create the following plot

  A boxplot of species on the x-axis and sepal length (cm) on the y-axis

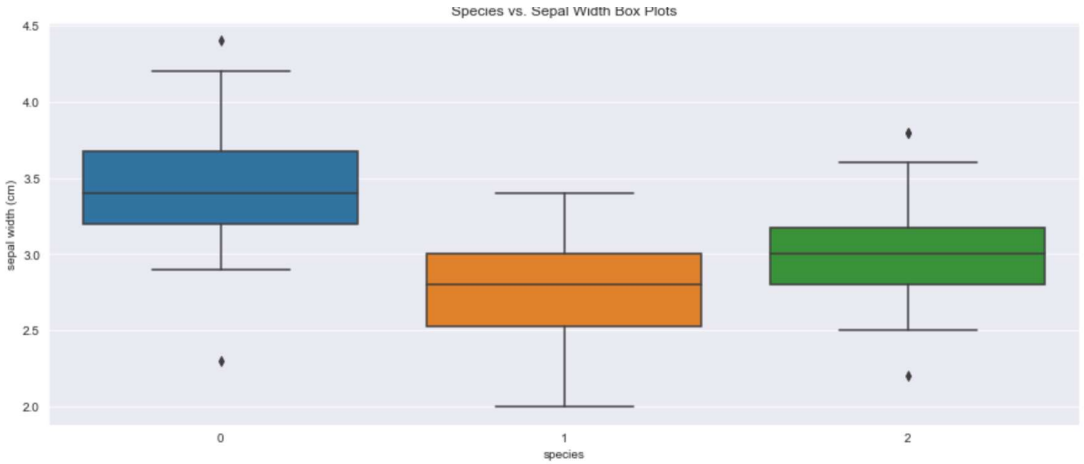


- Answer:
  - What can you conclude from the plots you created?

Part4 Reading and Understanding the Data (Classification part)

- Create the following plot

  A boxplot of species on the x-axis and sepal width (cm) on the y-axis

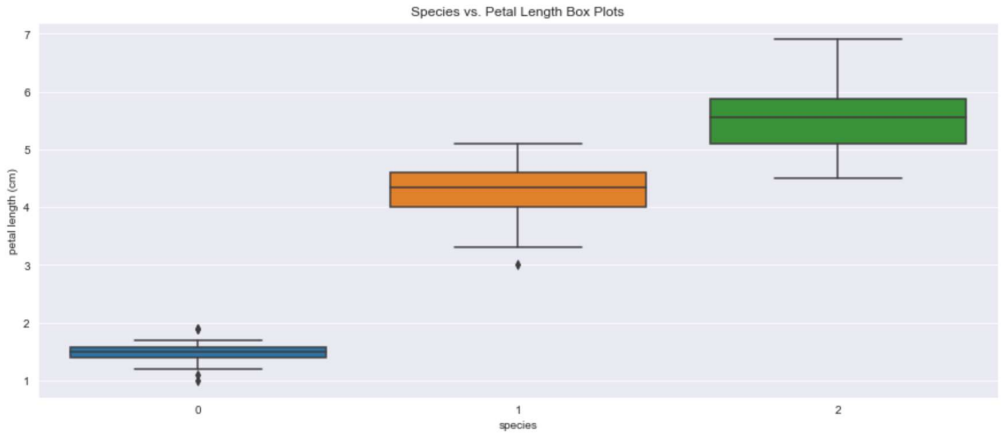


- Answer:
  - What can you conclude from the plots you created?

Part4 Reading and Understanding the Data (Classification part)

- Create the following plot

  A boxplot of species on the x-axis and petal length (cm) on the y-axis

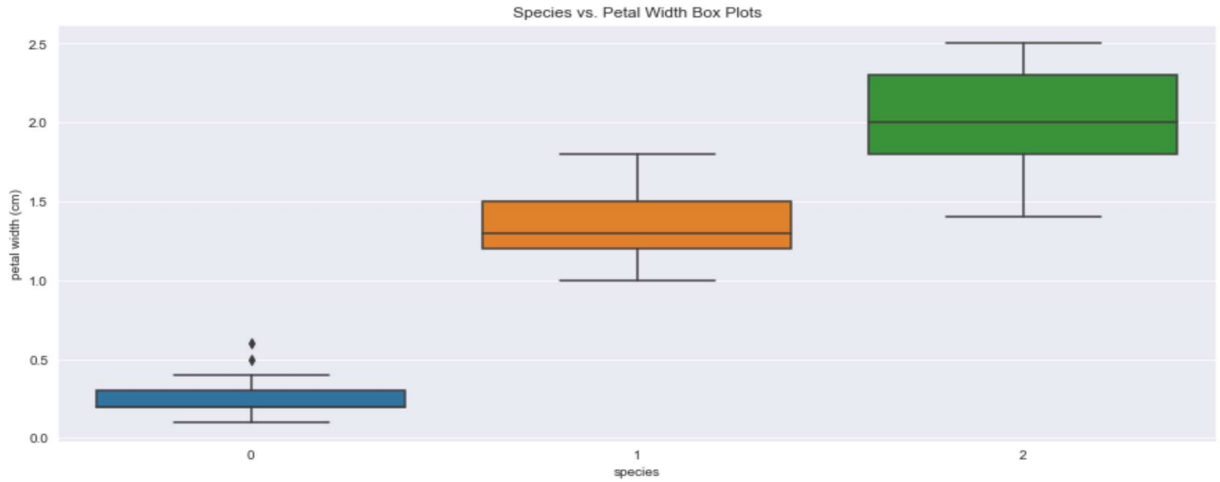


- Answer:
  - What can you conclude from the plots you created?

Part4 Reading and Understanding the Data (Classification part)

- Create the following plot

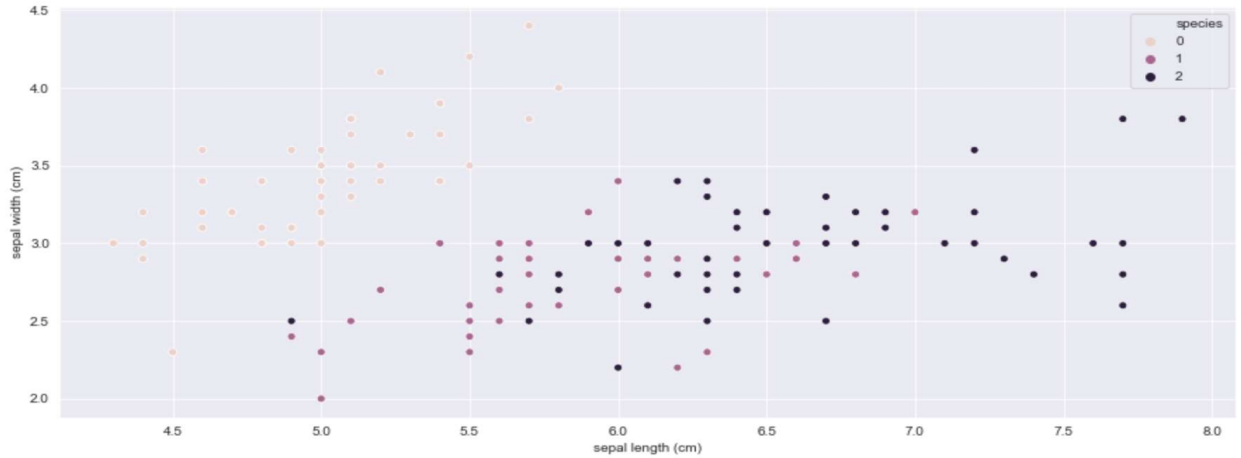
  A boxplot of species on the x-axis and petal width (cm) on the y-axis



- Answer:
  - What can you conclude from the plots you created?

Part4 Reading and Understanding the Data (Classification part)

- Create a scatterplot of sepal length (cm) vs. sepal width (cm) and color each point with a unique color according to their species.



- Answer
  - What can you conclude from the plots you created?

Part5: Model creation and evaluation

- Remove species column from the dataset
- Perform the following
  1. Scale the data
  2. Split your data into a training and testing set, with test size of 0.30
  3. Create a neural network for classification using sklearn. See sklearn.neural_network.MLPClassifier if you need help. Use a hidden layers of size 256->128->64->32, and activation function 'relu'.
  4. Run the test data through your model to obtain predictions. Save these predictions into a variable called 'predictions'.
  5. Print the mean accuracy of your model on the given test data and labels.

Part5 Model Creation and Evaluation

- Answer :
  - How did your model perform?
  - Try experimenting with different hidden layer sizes.
  - Try using gridsearchCV to find the best set of parameters.
  - Save any cells you use for experimenting so that we can see what you tried.
  - What works best for this model?

- Instruction(for submission)
  - Save any cells you use for experimenting so that we can see what you tried.

Part5 Model creation and evaluation

- Create a confusion matrix of the results.