Data Mining
Assignment Solution 2

Vivek Patani

February 23, 2016

Please read readme.md for code execution directions, I have tried to maintain a generic structure for each question in coding terms.

1. Question 1 Solution

Download data set Iris and answer the following questions:

1.1 Section 1

Calculate the average value and standard deviation for each of the four features.

Average for Petal Width: 1.2
Variance for Petal Width: 0.58
Standard Deviation for Petal Width: 0.762

Average for Petal Length: 3.76
Variance for Petal Length: 3.09
Standard Deviation for Petal Length: 1.758

Average for Sepal Length: 5.84
Variance for Sepal Length: 0.68
Standard Deviation for Sepal Length: 0.82

Average for Sepal Width: 3.05
Variance for Sepal Width: 0.19
Standard Deviation for Sepal Width: 0.436

Code Located in q1/q1.py

1.2 Section 2

Repeat the previous step but separately for each type of flower.

Average for Sepal Width & Iris Setosa: 3.42
Variance for Sepal Width & Iris Setosa: 0.14
Standard Deviation for Sepal Width & Iris Setosa: 0.37416573867739417

Average for Sepal Length & Iris Setosa: 5.01

Variance for Sepal Length & Iris Setosa: 0.12
Standard Deviation for Sepal Length & Iris Setosa: 0.34641016151377546

Average for Petal Width & Iris Setosa: 0.24
Variance for Petal Width & Iris Setosa: 0.01
Standard Deviation for Petal Width & Iris Setosa: 0.1

Average for Petal Length & Iris Setosa: 1.46
Variance for Petal Length & Iris Setosa: 0.03
Standard Deviation for Petal Length & Iris Setosa: 0.17320508075688773

Average for Sepal Width & Iris Versicolor: 2.77
Variance for Sepal Width & Iris Versicolor: 0.1
Standard Deviation for Sepal Width & Iris Versicolor: 0.31622776601683794

Average for Sepal Length & Iris Versicolor: 5.94
Variance for Sepal Length & Iris Versicolor: 0.26
Standard Deviation for Sepal Length & Iris Versicolor: 0.5099019513592785

Average for Petal Width & Iris Versicolor: 1.33
Variance for Petal Width & Iris Versicolor: 0.04
Standard Deviation for Petal Width & Iris Versicolor: 0.2

Average for Petal Length & Iris Versicolor: 4.26
Variance for Petal Length & Iris Versicolor: 0.22
Standard Deviation for Petal Length & Iris Versicolor: 0.469041575982343

Average for Sepal Width & Iris Virginica: 2.97
Variance for Sepal Width & Iris Virginica: 0.1
Standard Deviation for Sepal Width & Iris Virginica: 0.31622776601683794

Average for Sepal Length & Iris Virginica: 6.59
Variance for Sepal Length & Iris Virginica: 0.4
Standard Deviation for Sepal Length & Iris Virginica: 0.6324555320336759

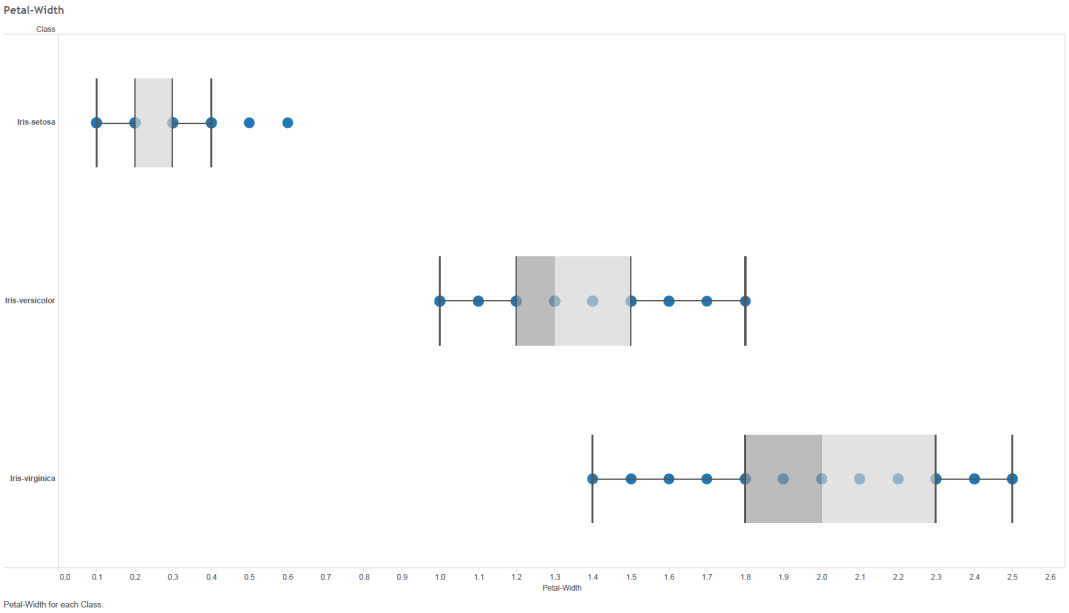
Average for Petal Width & Iris Virginica: 2.03
Variance for Petal Width & Iris Virginica: 0.07
Standard Deviation for Petal Width & Iris Virginica: 0.2645751311064591

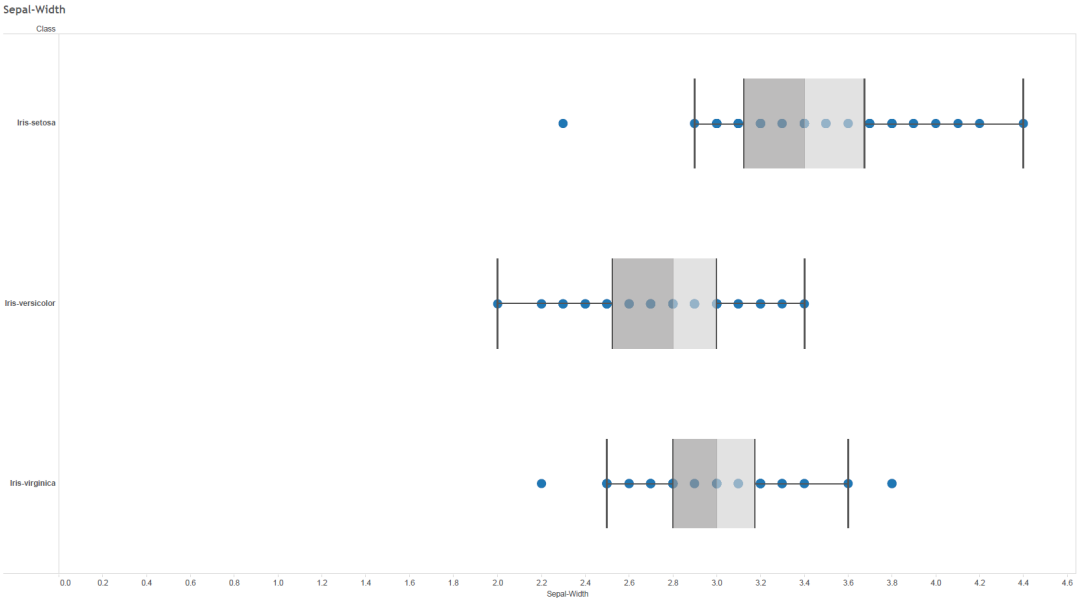
Average for Petal Length & Iris Virginica: 5.55

Variance for Petal Length & Iris Virginica: 0.3
Standard Deviation for Petal Length & Iris Virginica: 0.5477225575051661

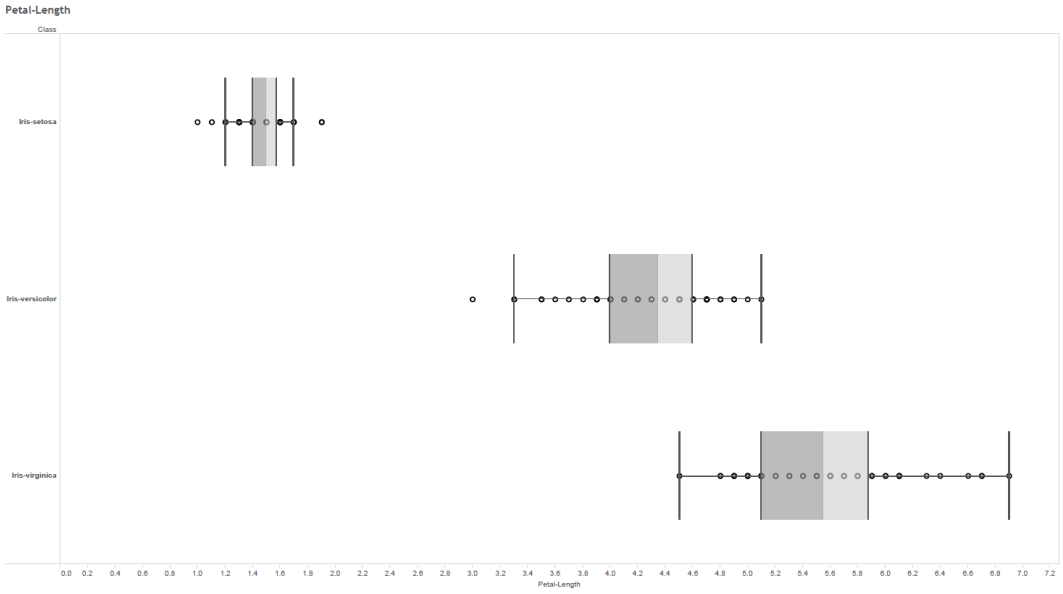
1.3 Section 3

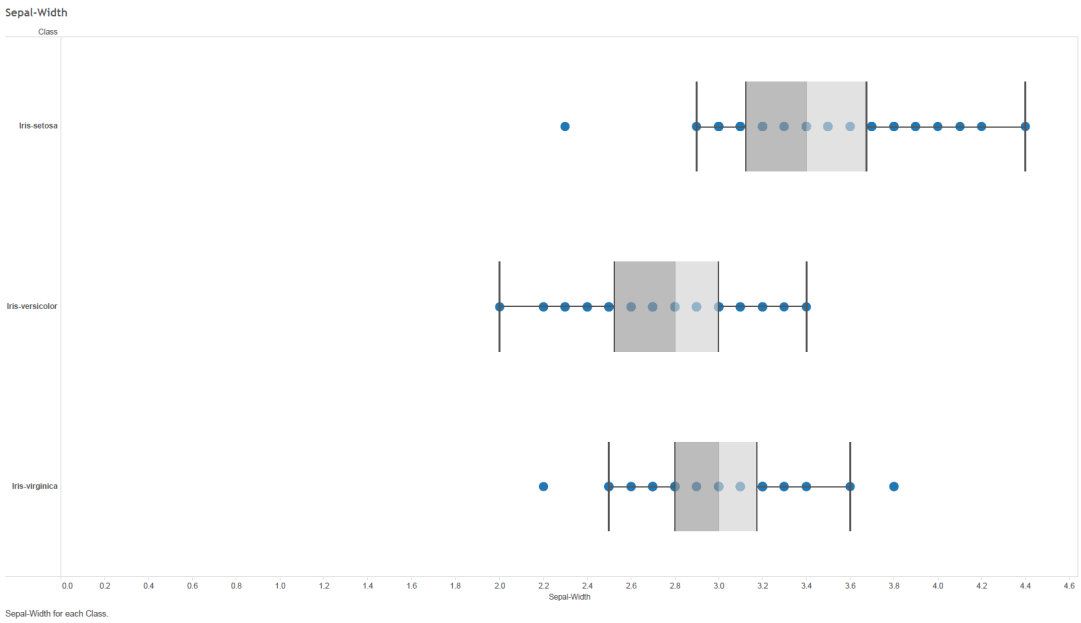
Draw four box plots, one for each feature, such that each figure shows three boxes, one for each type of flower. Properly label your figures and axes in all box plots. Make sure that the box plots look professional and appear in high resolution. Experiment with thickness of lines, font styles/sizes, etc. and describe what you tried and what looked the most professional.





Sepal-Width for each Class.





2. Question 2 Solution

Download data set Wine and answer the following questions

2.1 Section 1

Provide pairwise scatter plots for four most correlated and four least correlated pairs of features, using Pearson's correlation coefficient. Label all axes in all your plots and select fonts of appropriate style and size. Experiment with different ways to plot these scatter plots and choose the one most visually appealing and most professionally looking.

2.2 Section 2

Use Euclidean distance to find the closest example to every example available in the data set (exclude the class variable). Calculate the percentage of points whose closest neighbours have the same class label (for data set as a whole and also for each class).

2.3 Section 3

Repeat the previous step but after the data set is normalized using first 0-1 normalization and then z-score normalization. Investigate the reasons for discrepancy and provide evidence to support every one of your claims. Provide the code you used for normalizing and visualizing the data.

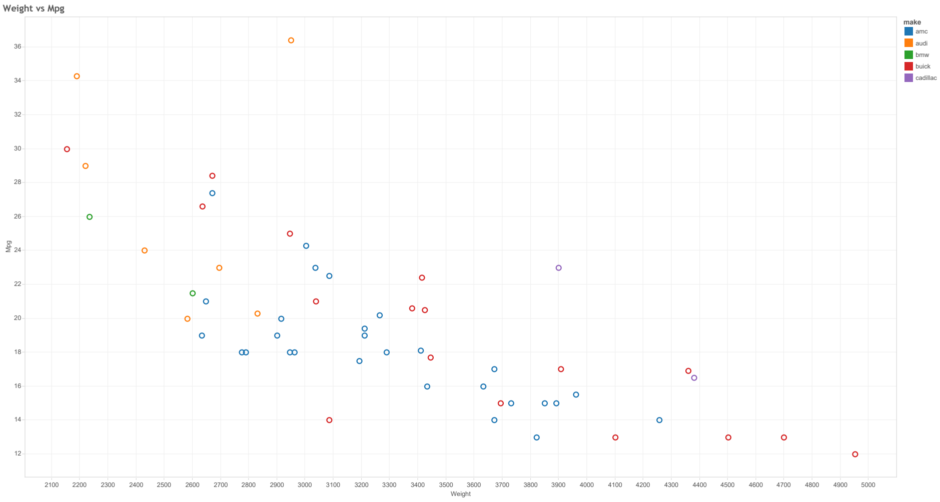
3. Question 3 Solution

Data exploration is often the first step in many data analysis tasks. Visualizing relationships between features as well as between features and the target variable(s), for example, can be exploited to design a good model or to understand why a particular model works. There are many software packages developed to make this step easier. In this question you will experiment with Tableau.

3.1 Section 1

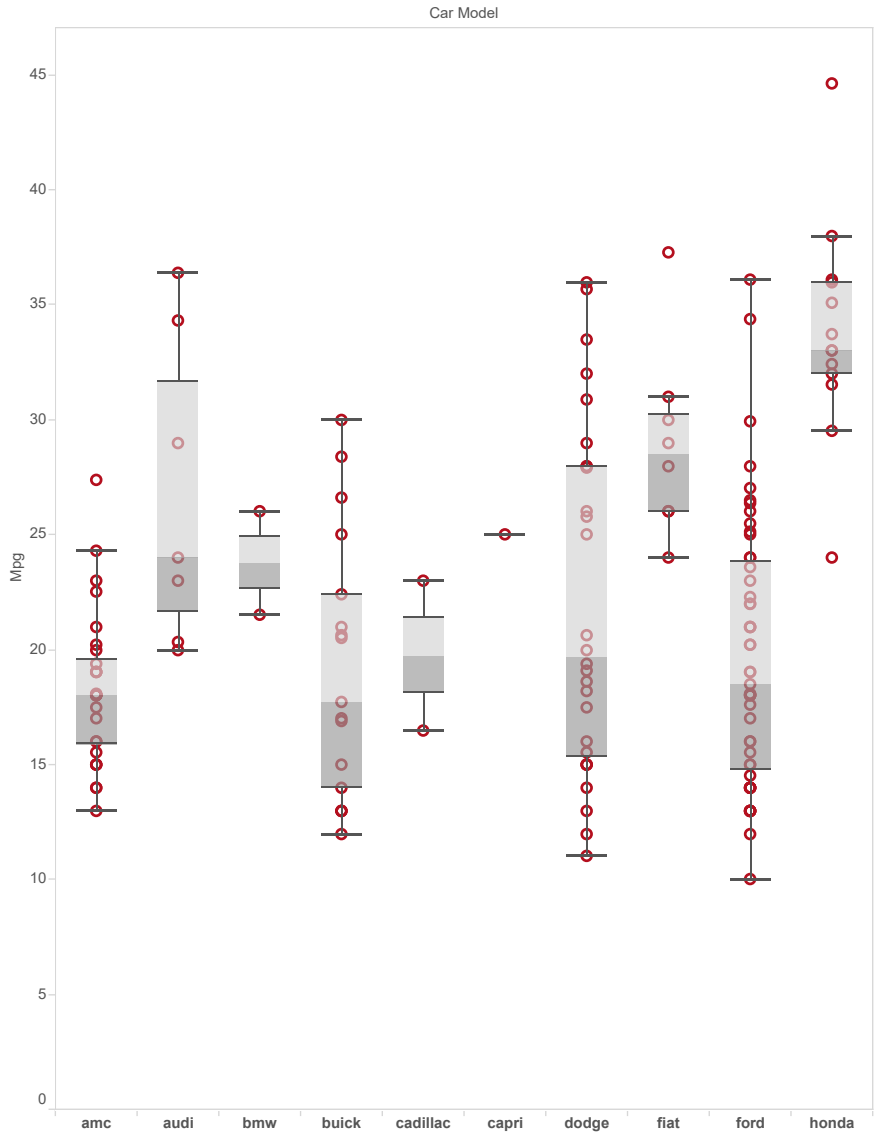
Download and study the Auto MPG data set from the UCI Machine Learning Repository. Import the data set into Tableau. Create a new feature make (Honda, Toyota, . . .) that contains the make of the automobile (extract this feature automatically from other features through Tableau) and in a single figure generate box plots of mpg for 10 makes of your choice. Then, for 5 makes of your choice create scatter plots of weight versus mpg. Include all figures in your submission and comment on what you observe.

Weight vs Mpg



Weight vs. Mpg. Color shows details about make.

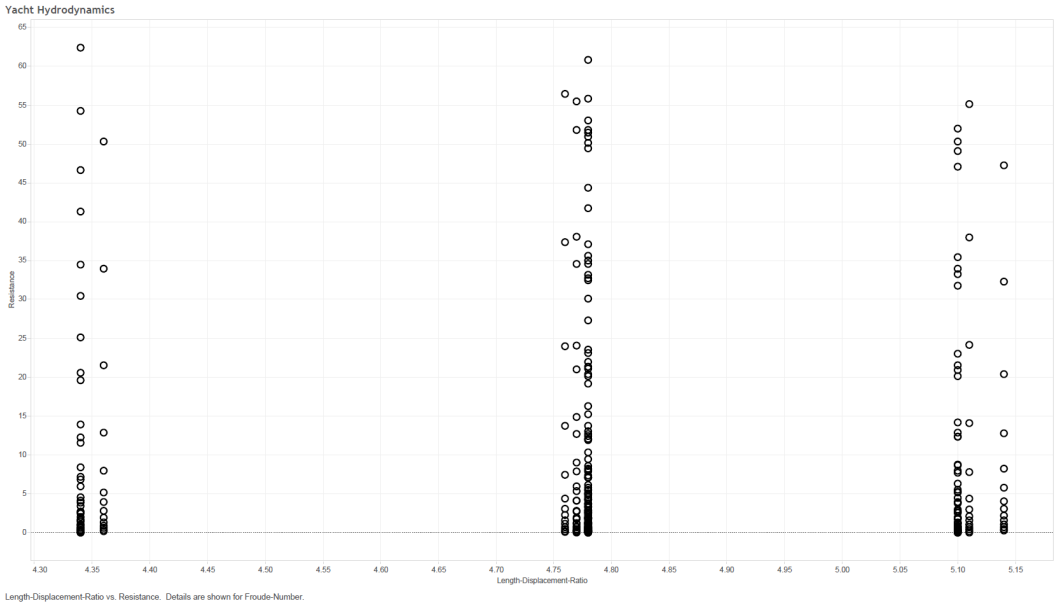
Car Model - Mpg Plot



Mpg for each Car Model.

3.2 Section 2

Pick 3 data sets of your choice from UCI Machine Learning Repository. Visualize each the data set in meaningful ways that show hidden patterns. Experiment with colors, size, shapes, filters,groups and sets. Feel free to experiment with other advanced functionalities of Tableau.



3.3 Section 3

Tell us about your experience with Tableau. What did you learn? What did you like/dislike about Tableau?