STA160 Midterm Project

Haolin Li | 913838107 | May 3, 2020

Introduction

In this report, I will conduct data analysis on two separate datasets, seeds and automobile from UCI Machine Learning Repository, to gain more knowledge about how data are structured and more, as well as to practice data visualization, preprocessing, clustering and more.

Seeds Data Set

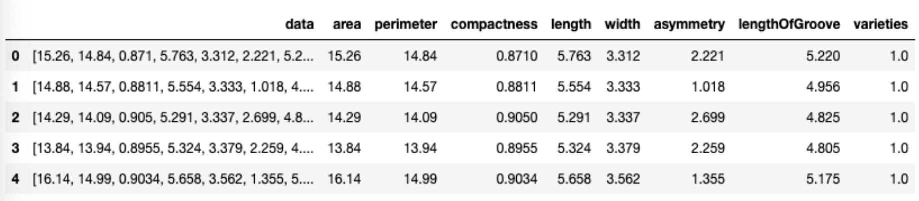
This dataset is downloaded from the UCI Machine Learning Repository. There are total of 210 elements in the dataset, belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each. Elements are randomly selected for the experiment. Then high-quality visualization of the internal kernel structure was detected using a soft X-ray technique, to obtain the seven geometric parameters of wheat kernels: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove, all of these parameters are real-valued continuous.

Automobile Data Set

This dataset is also downloaded from the UCI Machine Learning Repository. The dataset was collected from three sources: vehicle specification, personal auto manuals and insurance collision report. There are total of 205 elements in the dataset with 26 attributes that are about the vehicles, they are: symbolling, normalized losses, make, fuel-type, aspiration, number of doors, body style, drive wheels, engine location, wheel base, length, width, height, weight, engine type, fuel system, bore, stroke, compression ratio, horsepower, pear rpm, city mpg, highway mpg and the price. All of these attributes are a collection of categorical and numerical data.

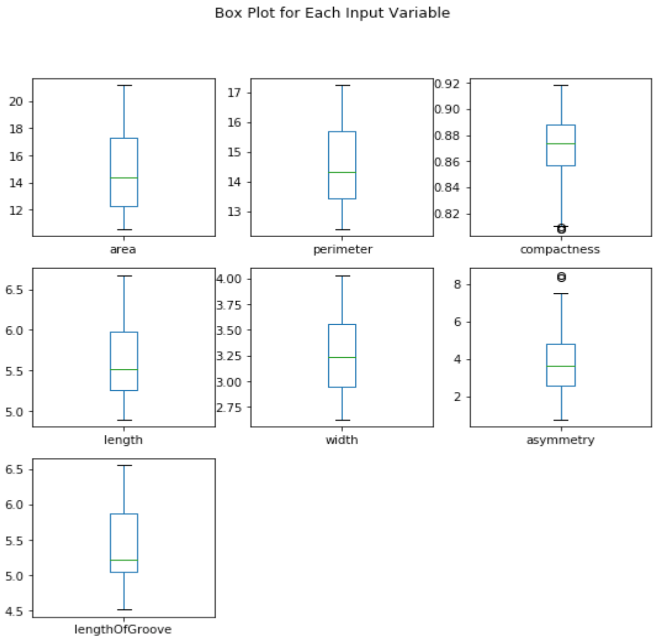
Data Exploratory

The datasets from the website has some cleaning to do. Each data stores the information as one single string or list. For the seed dataset, each wheat information is store as one string and within the string, features are separated by one or more tabs. For the automobile dataset, each vehicle information is store as a list and within the list, attributes are separated by a comma. Therefore using the split() function, I was able to clean up the data and form into a data frame. Below is an example of the data frame for seed data set.

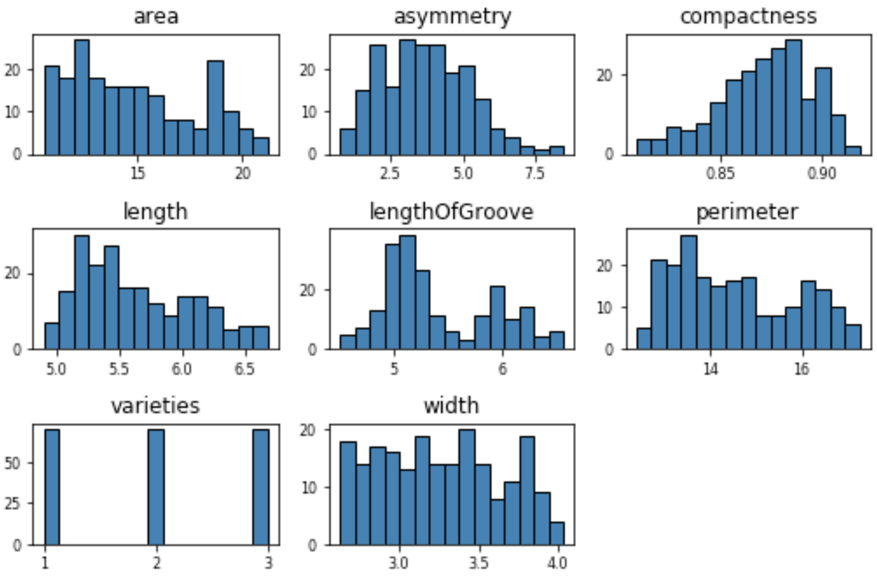


	data	area	perimeter	compactness	length	width	asymmetry	lengthOfGroove	varieties
0	[15.26, 14.84, 0.871, 5.763, 3.312, 2.221, 5.2...	15.26	14.84	0.8710	5.763	3.312	2.221	5.220	1.0
1	[14.88, 14.57, 0.8811, 5.554, 3.333, 1.018, 4....	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1.0
2	[14.29, 14.09, 0.905, 5.291, 3.337, 2.699, 4.8...	14.29	14.09	0.9050	5.291	3.337	2.699	4.825	1.0
3	[13.84, 13.94, 0.8955, 5.324, 3.379, 2.259, 4....	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1.0
4	[16.14, 14.99, 0.9034, 5.658, 3.562, 1.355, 5....	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1.0

To begin with exploring the variables, we often use box plot to explore for the numeric variables because it gives a clear idea of the distribution of the input variables. The boxplots of the features of wheat below show us the distribution of each feature. From that, we can observe that asymmetry has generally a normal distribution.

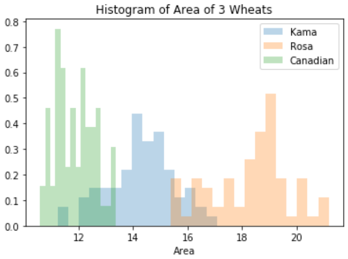


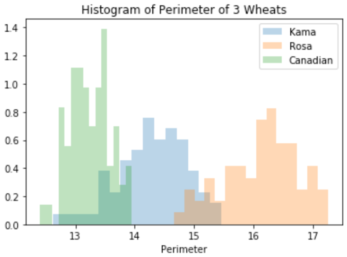
One of the quickest and most effective ways to visualize all numeric data and their distribution is to use histograms. From the histograms, we have a basic idea of the data distribution of any of the features, not regarding the types of wheats.

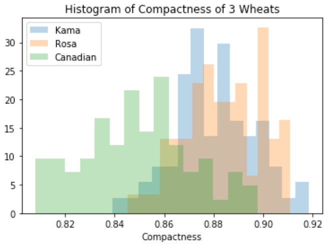


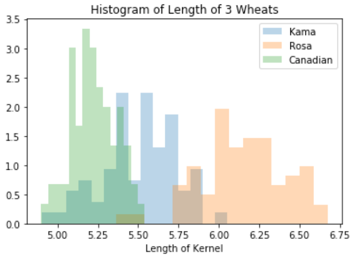
Variable Comparison

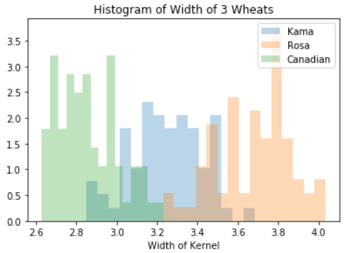
In this section, I want to compare the labels and features of the seed dataset in different dimensions. I generated histograms of each feature regarding all three types of wheat, to see the distributions of each feature of each wheat, and how they are different, as well as similar in structure, basically I am seeking overlap and similarity in distribution.

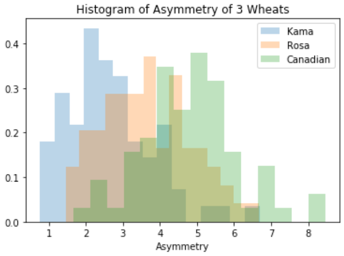


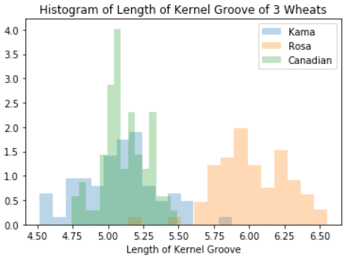






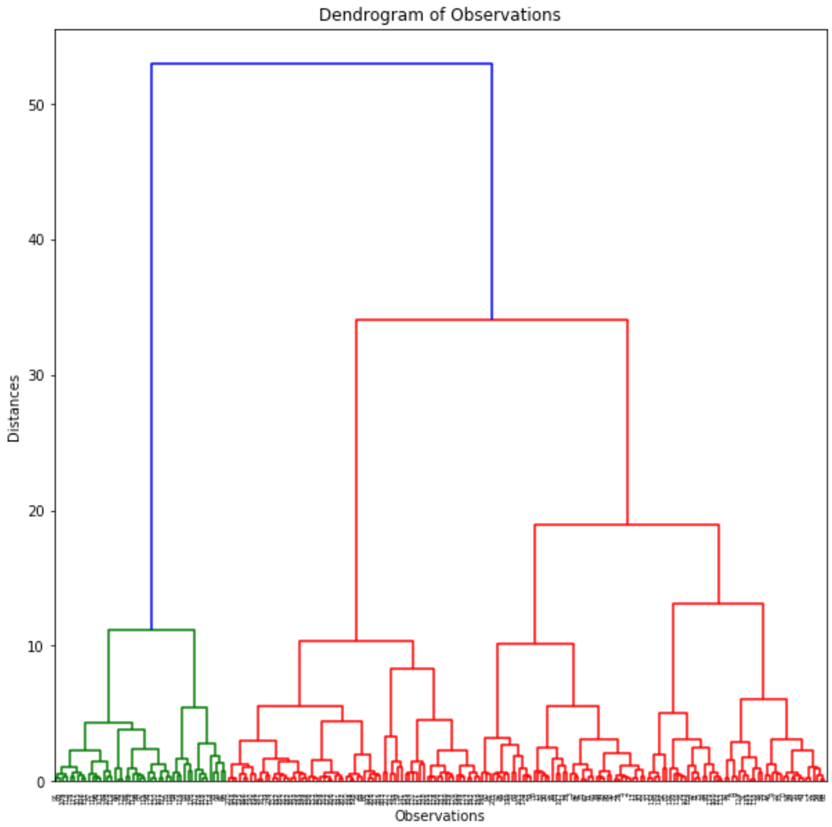






From the group of histograms above, we can see that there are three colors within each histogram, the green represents Kama, the orange represents Rosa and the blue represents Canadian. We can observe that there are overlays in each feature of the three wheats. For example, in the length of the kernel groove, Kama and Canadian have a similar data input and distribution. Therefore, these histograms show us how to compare multiple labels with respect to one single feature.

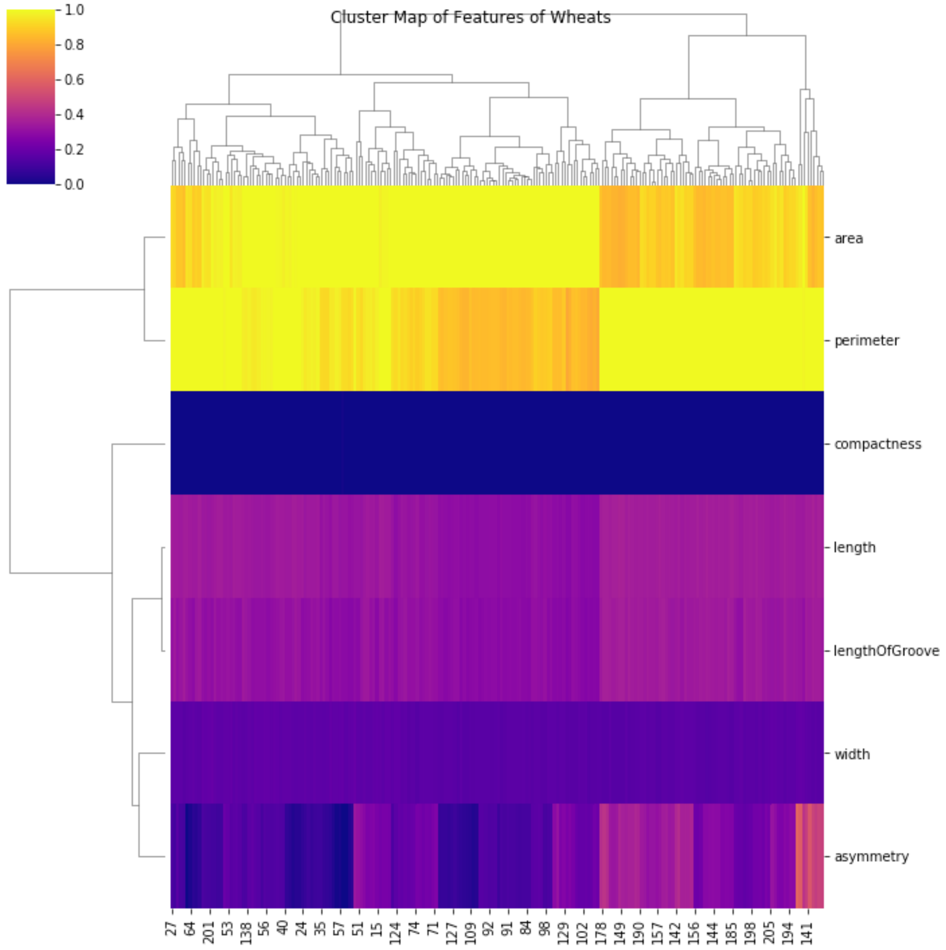
The reason I use simple histograms to compare multiple labels of each feature is because they are one-dimensional. We do not use multidimensional clustering algorithms for one-dimensional problem because one-dimensional data can be sorted. Therefore, when it comes to compare multiple labels of multiple features, we use cluster analysis because data is larger and complex, we have too many variables and too complex processes to model them. In this case, I obtained a dendrogram of the features of the seed dataset.



A dendrogram is a diagram that shows the hierarchical relationship between objects. It is commonly created from hierarchical clustering. The dendrogram above shows us that there are three classes in our dataset. The way we can read the dendrogram is to focus on the height at which any two objects are joined together. In the dendrogram above, each data point is a cluster and data points with similarity joined together and formed another cluster. The color in the dendrogram indicates that there are three classes in out dataset. Therefore, to compare multiple labels with respect to multiple features, we can use dendrogram.

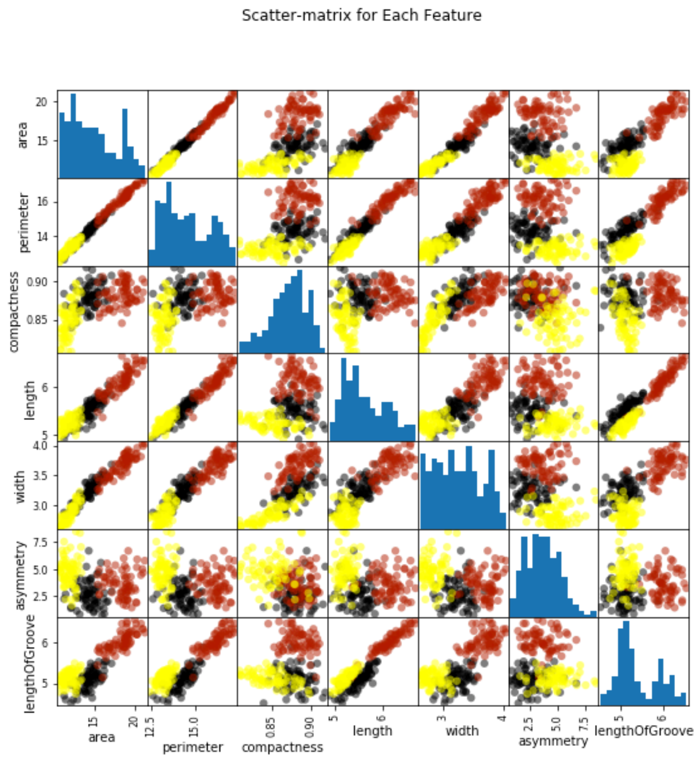
Variable Relations

In this section, I want to obtain the relationship between variables of each dataset. To do so, I obtained three different plots to visualize the relationship between the variables. The goal of this section here is to show the correlation between some variables from the data and visualize it with plots. Below I obtained a cluster map of features of the seed dataset.



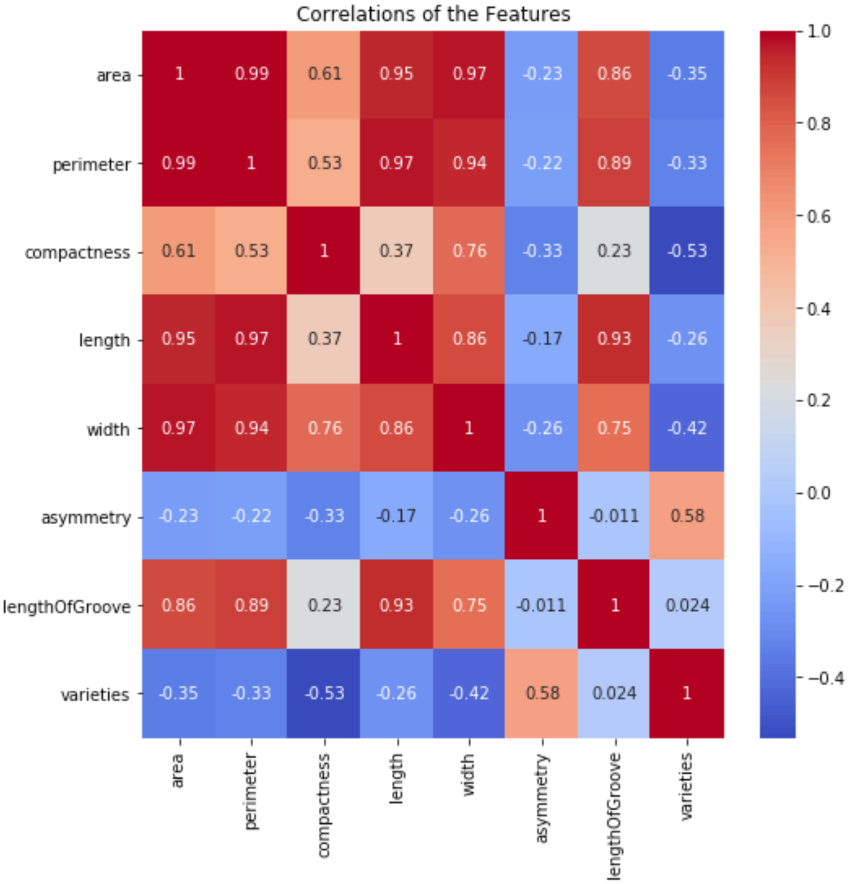
From this cluster map, we can see that there are cluster on the side to indicate the relationship between the features and cluster on the top to indicate the relationship between each data points. From center part of the plot, we can see that area and perimeter share the same color, and length and length of groove share the same color as well. This indicates that the two features are correlated, and it is also shown in the cluster map on the side where the height of the cluster is low, the lower the cluster the stronger they are correlation.

Perhaps the cluster map is not very easy to interpret since the color and clusters are very confusing. Below I obtained the scatter-matrix plot for each feature to display the correlation between features.



A scatter-matrix plot is an estimation of covariance matrix when covariance cannot be calculated or hardly to calculate. It is used in a lot of dimensionality reduction exercises. If there are k variables, scatter-matrix plot will be a k*k matrix, in this case, the scatter-matrix plot for seed dataset is a 7*7 matrix. The scatter-matrix plot shows us the correlation of the features. We observed from the cluster map area and perimeter are correlated, as well as length and length of groove. From the scatter-matrix, we can see the correlation between the variables very well where features with strong correlation has a linear correlation map.

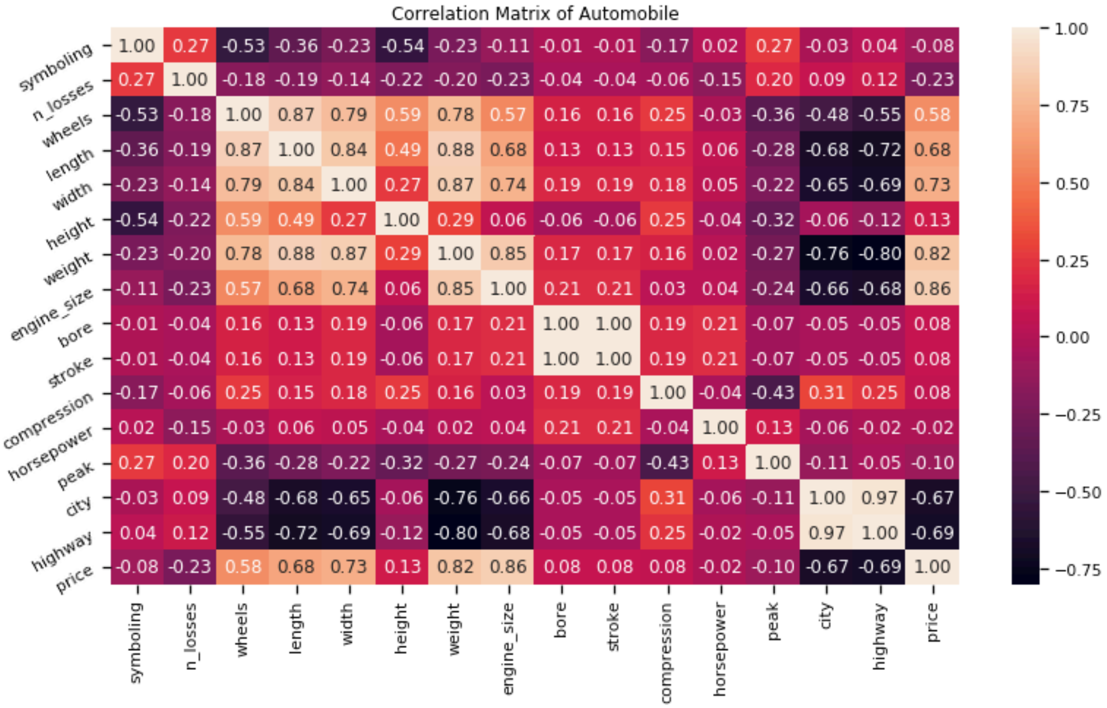
The two different plots from above to show the correlation of the variables are all imagery where there is only color, distribution and shape we can observe from. Below I obtained a correction matrix of the features for further proving the correlation between the variables.



	area	perimeter	compactness	length	width	asymmetry	lengthOfGroove	varieties
area	1	0.99	0.61	0.95	0.97	-0.23	0.86	-0.35
perimeter	0.99	1	0.53	0.97	0.94	-0.22	0.89	-0.33
compactness	0.61	0.53	1	0.37	0.76	-0.33	0.23	-0.53
length	0.95	0.97	0.37	1	0.86	-0.17	0.93	-0.26
width	0.97	0.94	0.76	0.86	1	-0.26	0.75	-0.42
asymmetry	-0.23	-0.22	-0.33	-0.17	-0.26	1	-0.011	0.58
lengthOfGroove	0.86	0.89	0.23	0.93	0.75	-0.011	1	0.024
varieties	-0.35	-0.33	-0.53	-0.26	-0.42	0.58	0.024	1

From the observation of the cluster map and scatter-matrix plot, we know that the correlation between area and perimeter are strong, which is proven from the correlation matrix above where the value of their covariance is 0.99. We also observed that the correlation between length and length of groove are strong, it is also proven from the correlation matrix above. The correlation can not only be observed from the value above, it can also be determined by the color where darker red means higher correlation and darker blue means lower correlation.

Below is the correlation matrix of the automobile dataset, similar to the correlation matrix for seed dataset above, we have number representing the value of covariance and we have color to indicate the strongness of the correlation. Because there are categorical and numerical variables in the automobile dataset, we only display the numerical variables here because adding categorical variables would make the matrix be more confusing.

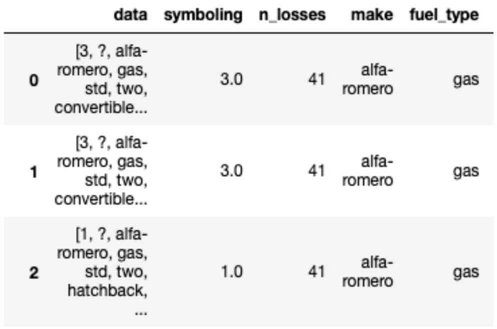


	symboling	n_losses	wheels	length	width	height	weight	engine_size	bore	stroke	compression	horsepower	peak	city	highway	price
symboling	1.00	0.27	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.01	-0.01	-0.17	0.02	0.27	-0.03	0.04	-0.08
n_losses	0.27	1.00	-0.18	-0.19	-0.14	-0.22	-0.20	-0.23	-0.04	-0.04	-0.06	-0.15	0.20	0.09	0.12	-0.23
wheels	-0.53	-0.18	1.00	0.87	0.79	0.59	0.78	0.57	0.16	0.16	0.25	-0.03	-0.36	-0.48	-0.55	0.58
length	-0.36	-0.19	0.87	1.00	0.84	0.49	0.88	0.68	0.13	0.13	0.15	0.06	-0.28	-0.68	-0.72	0.68
width	-0.23	-0.14	0.79	0.84	1.00	0.27	0.87	0.74	0.19	0.19	0.18	0.05	-0.22	-0.65	-0.69	0.73
height	-0.54	-0.22	0.59	0.49	0.27	1.00	0.29	0.06	-0.06	-0.06	0.25	-0.04	-0.32	-0.06	-0.12	0.13
weight	-0.23	-0.20	0.78	0.88	0.87	0.29	1.00	0.85	0.17	0.17	0.16	0.02	-0.27	-0.76	-0.80	0.82
engine_size	-0.11	-0.23	0.57	0.68	0.74	0.06	0.85	1.00	0.21	0.21	0.03	0.04	-0.24	-0.66	-0.68	0.86
bore	-0.01	-0.04	0.16	0.13	0.19	-0.06	0.17	0.21	1.00	1.00	0.19	0.21	-0.07	-0.05	-0.05	0.08
stroke	-0.01	-0.04	0.16	0.13	0.19	-0.06	0.17	0.21	1.00	1.00	0.19	0.21	-0.07	-0.05	-0.05	0.08
compression	-0.17	-0.06	0.25	0.15	0.18	0.25	0.16	0.03	0.19	0.19	1.00	-0.04	-0.43	0.31	0.25	0.08
horsepower	0.02	-0.15	-0.03	0.06	0.05	-0.04	0.02	0.04	0.21	0.21	-0.04	1.00	0.13	-0.06	-0.02	-0.02
peak	0.27	0.20	-0.36	-0.28	-0.22	-0.32	-0.27	-0.24	-0.07	-0.07	-0.43	0.13	1.00	-0.11	-0.05	-0.10
city	-0.03	0.09	-0.48	-0.68	-0.65	-0.06	-0.76	-0.66	-0.05	-0.05	0.31	-0.06	-0.11	1.00	0.97	-0.67
highway	0.04	0.12	-0.55	-0.72	-0.69	-0.12	-0.80	-0.68	-0.05	-0.05	0.25	-0.02	-0.05	0.97	1.00	-0.69
price	-0.08	-0.23	0.58	0.68	0.73	0.13	0.82	0.86	0.08	0.08	0.08	-0.02	-0.10	-0.67	-0.69	1.00

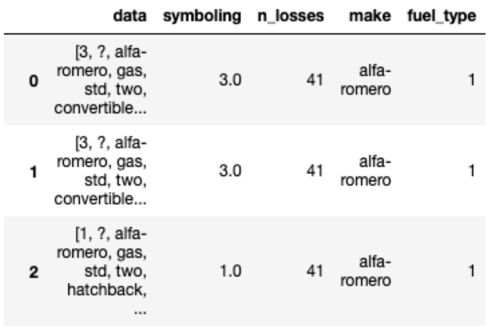
Categorical Variable

In this section, I will talk about how to deal with categorical variables in the seed dataset and automobile dataset. In the seed dataset, the only categorical variable is the label of the wheat, which is Kama, Rosa and Canadian. In the automobile dataset, we have 10 categorical variables and they are: make, fuel type, aspiration, number of doors, body style, drive wheels, engine location, engine type, number of cylinders and fuel system. Despite the units for number of doors are expressed as 2 door or 4 doors with number, they are actually categorical variable to distinguish what type of automobile is it, or in other word, the type of number of doors is fixed.

When we deal with categorical variable, we have two things to do. If we want to transform the data type to numeric, then we have to declare a variable where numbers represent different type of that variable. For example, in the dataset we have fuel type displayed as a string, “gas” or “diesel”. If we want to transform the variable into an integer, then we need to declare such that 0 = gas, 1 = diesel. Additionally, we can also do Boolean expression where true means gas, false means diesel, but this way only limits two types of choices. Below is the transformation of the fuel type variable.

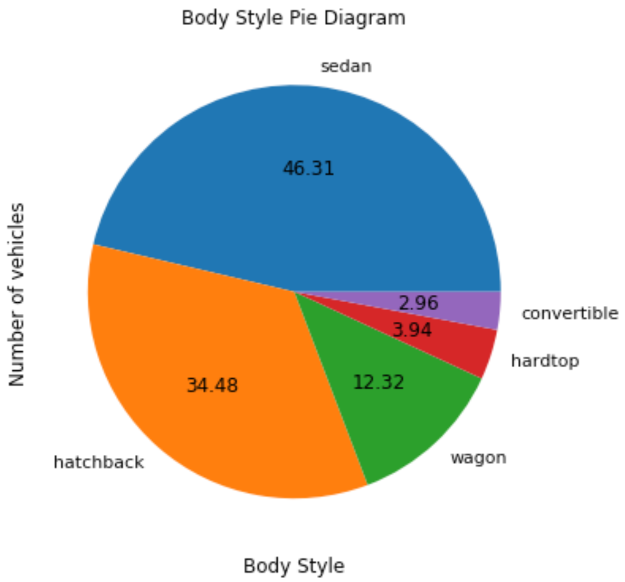


	data	symboling	n_losses	make	fuel_type
0	[3, ?, alfa-romero, gas, std, two, convertible...	3.0	41	alfa-romero	gas
1	[3, ?, alfa-romero, gas, std, two, convertible...	3.0	41	alfa-romero	gas
2	[1, ?, alfa-romero, gas, std, two, hatchback, ...	1.0	41	alfa-romero	gas

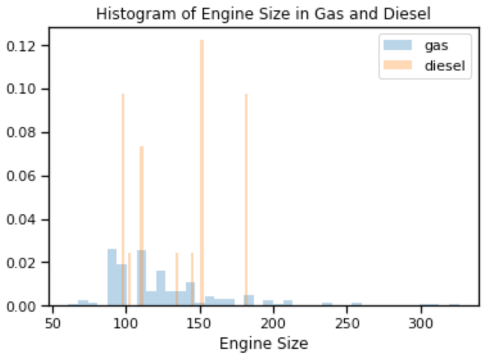


	data	symboling	n_losses	make	fuel_type
0	[3, ?, alfa-romero, gas, std, two, convertible...	3.0	41	alfa-romero	1
1	[3, ?, alfa-romero, gas, std, two, convertible...	3.0	41	alfa-romero	1
2	[1, ?, alfa-romero, gas, std, two, hatchback, ...	1.0	41	alfa-romero	1

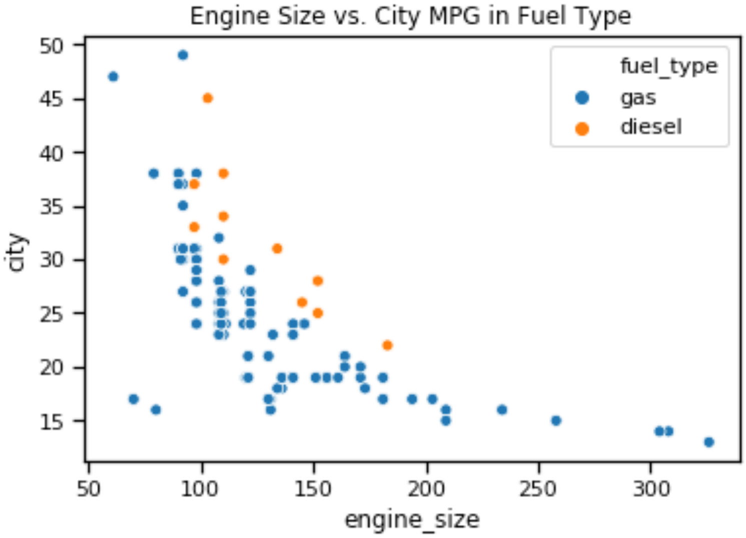
When we want to plot categorical variable, we often use pie chart or histogram, charts that where x is the categorical names and y is the number of values. Below is the pie chart of body style which shows the number of vehicles of each body style automobile.



When we want to compare categorical variable and numeric variable, we often use histogram to show the distribution of the two variables. Below is the histogram of engine size in different fuel type. We can see their distribution for each fuel type and seems like diesel fuel type automobile has a larger engine.



Another way to compare categorical data is to use it as the separator of two numerical variables. In this case below, we have the scatter plot showing the relationship between engine size and city MPG, but we also show that which car is using gas and which car is using diesel. Therefore, we can conclude that gas and diesel both has the same relationship that the larger the engine size the lower the city MPG.



Conclusion

In this report, we discussed and show how to compare multiple labels with respect to one single feature in one dimension. I also showed how to compare in multidimensional using dendrogram and clustering. I then displayed using plot to show how to measure correlations between variables, numerical vs. numerical and categorical vs. numerical.