Statistics for Modelling

Data Science and Modeling

1. A process is characterized by a well defined set of inputs, well defined set of transformations and well defined output. The process may be physical such as an automobile or abstract such as El-Nino

2. Well defined inputs and outputs means the frequency of range of values inputs and outputs can take on various attributes seem to have a relationship. Some values occur more frequently than others

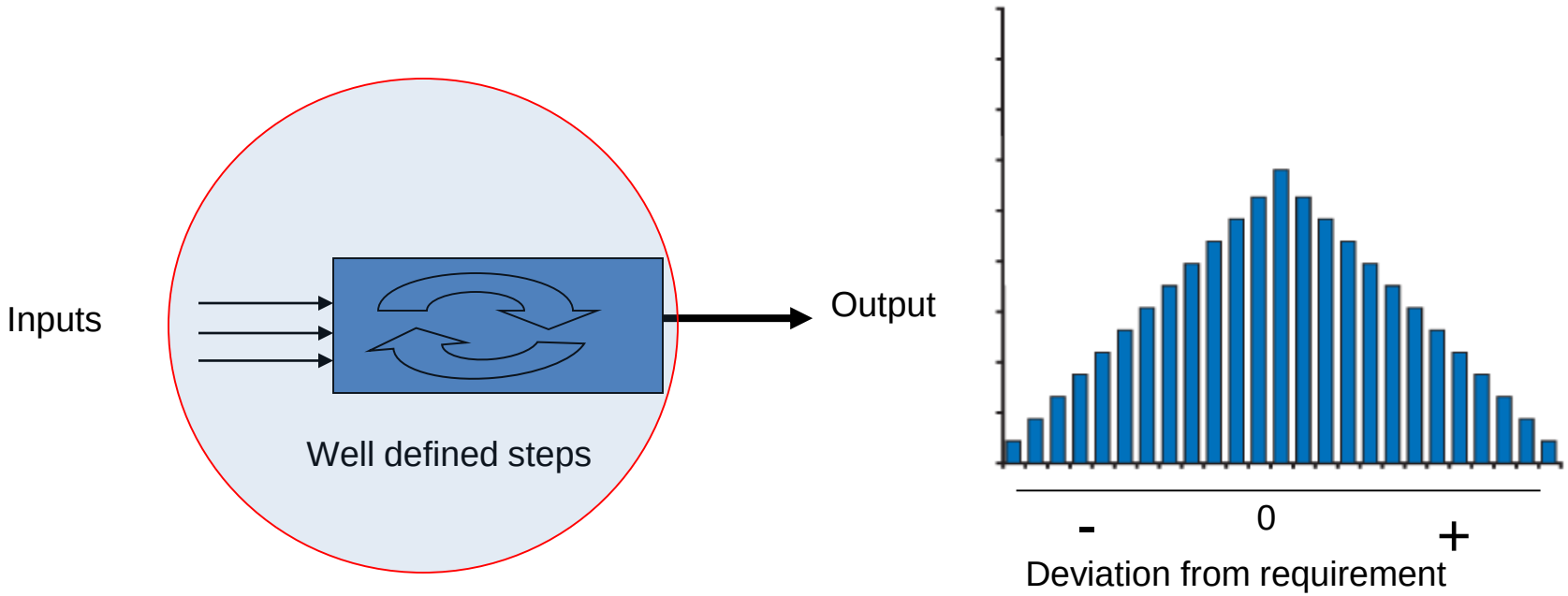
3. A process that meets the requirements in bullet 1 is a well defined stable process. We may not fully understand some processes such as El-Nino, as long as we observe some patterns, we can attempt to model it

4. Models are representations of physical process. The representation may be in form of equations, rules, clusters etc. For e.g. El-Nino $\frac{\partial}{\partial t}\iiint K\,dV = -\oint K\mathbf{u}\cdot\hat{n}\,d\sigma - \oint(\tilde{p}+p_s)\mathbf{u}\cdot\hat{n}\,d\sigma - g\iiint\tilde{\rho}w\,dV + \iint_{z=0}\mathbf{v}\cdot\boldsymbol{\tau}_o\,d\sigma\,\rho_o - \iiint\kappa_{MV}(\mathbf{u}_z\cdot\mathbf{u}_z)\,dV\,\rho_o - \iiint\kappa_{MH}[(\mathbf{u}_x\cdot\mathbf{u}_x)+(\mathbf{u}_y\cdot\mathbf{u}_y)]\,dV.$ https://journals.ametsoc.org/doi/full/10.1175/1520-0442%282000%29013%3C1496%3ATEOENO%3E2.0.CO%3B2

5. Models help us understand the current behavior of a process and predict their future behavior. Behavior refers to how the process will transform given inputs to output and it’s impact

Descriptive Statistics and Modelling

Well defined stable process

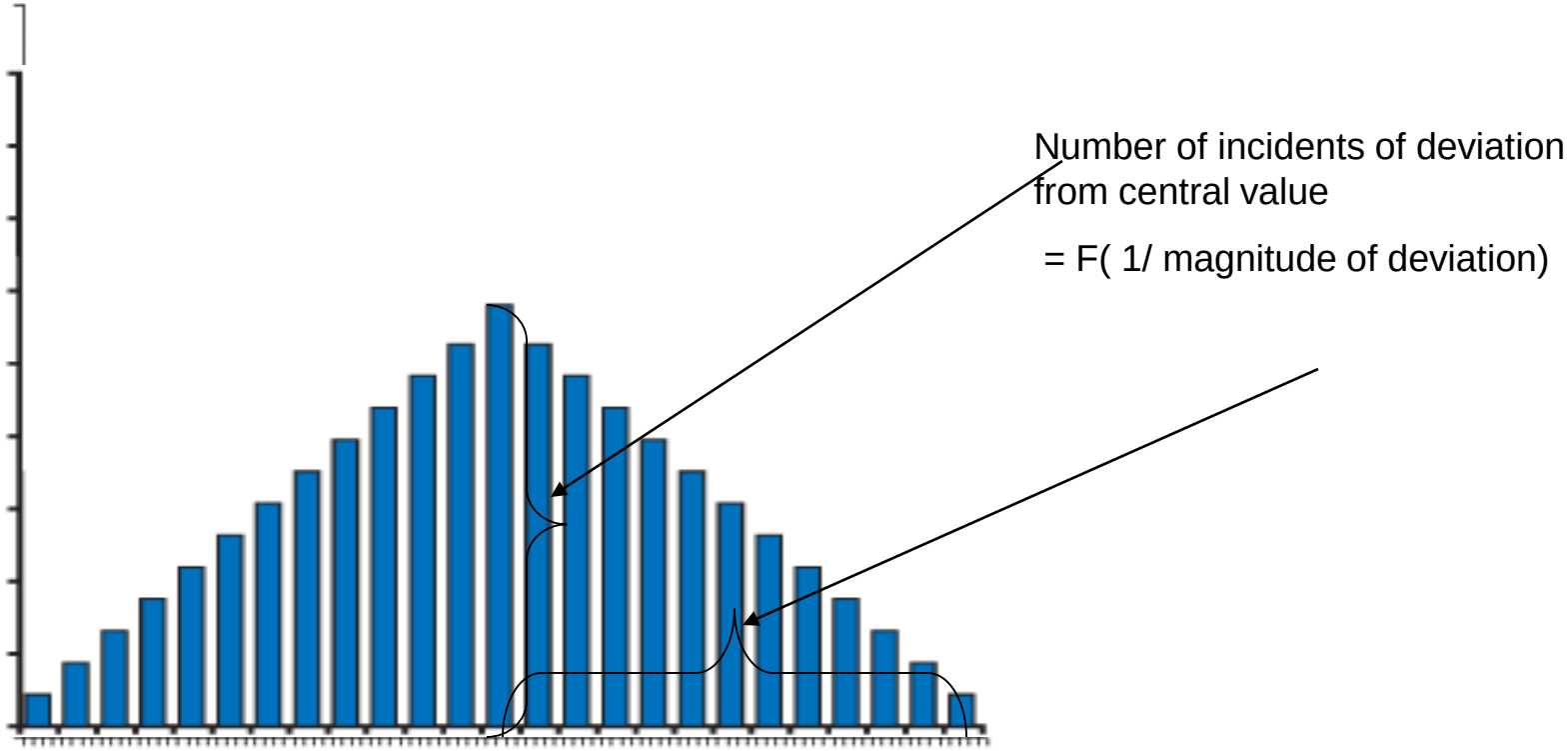


1. Output values vary due to random factors that come into play during the processing

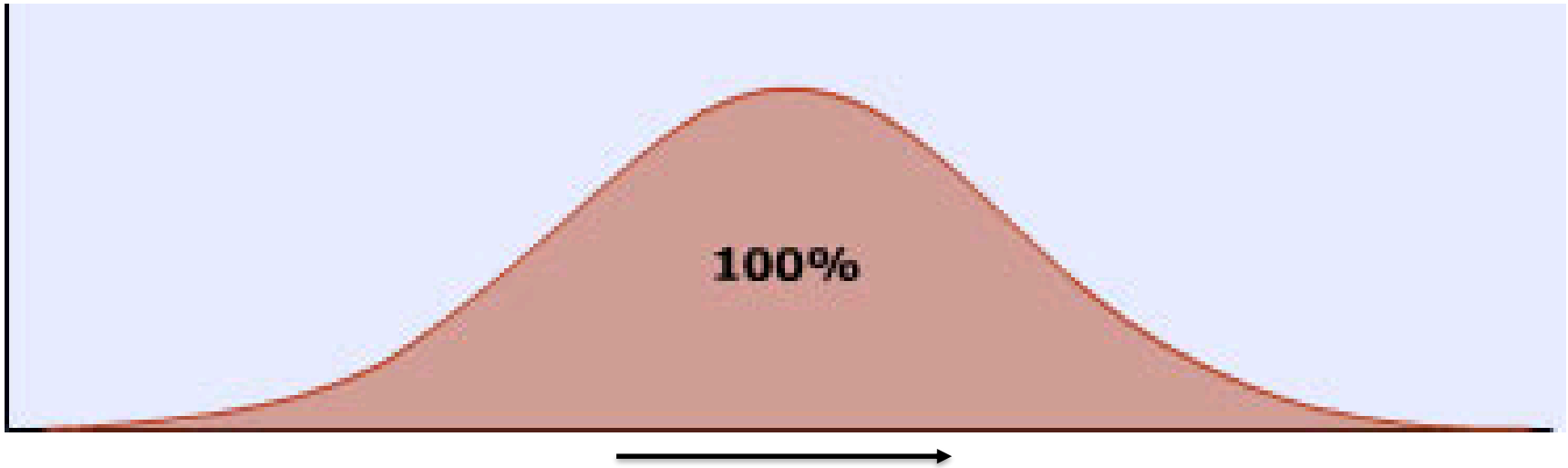
2. Output values very different from expected are generated less frequently. Most of the output values are clustered close to the expected value

Distribution characteristics of a well defined stable process

1. The output metric will show a frequency distribution as shown below. Generate outputs with value close to the expected value (central values)
2. The output metric value strays more and more from the expected value with lesser frequency
3. The frequency distribution on a continuous scale can be transformed to a continuous distribution

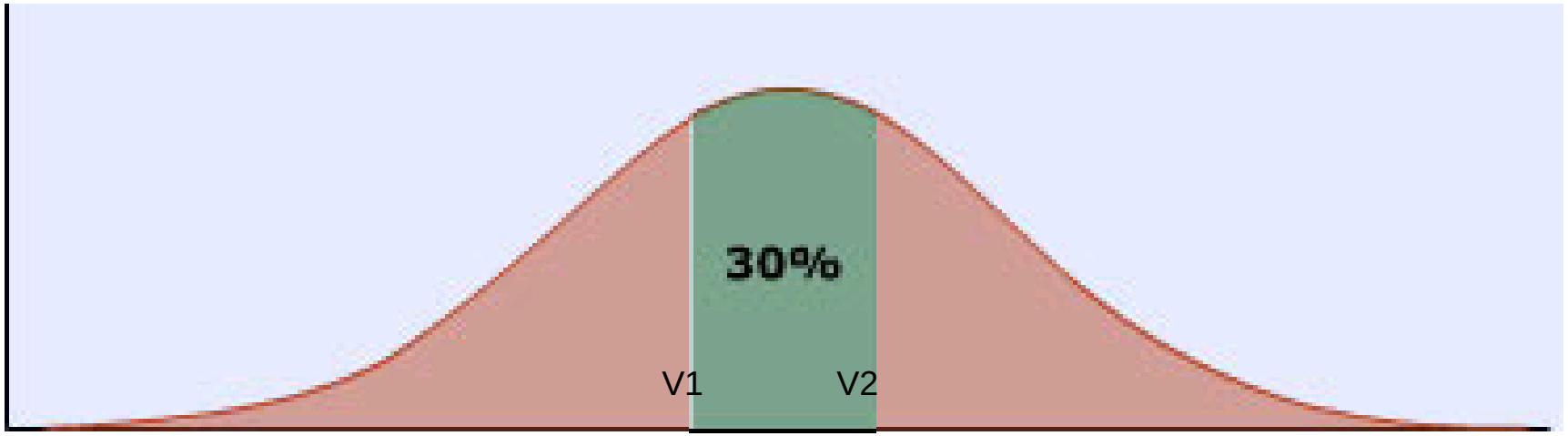


Area under the curve and probability



1. Area under the bell curve represents all possible (100%) values a parameter can take
2. Probability of output with value in the given range = Ratio of area of curve representing a particular range of values to overall area

Area under the curve and probability



Possible values of a given parameter e.g. diameter of a bolt up to infinity

1. Probability of producing output with values between V1 and V2 is area of the curve under between the two values / total area of the bell curve
2. What is the probability of producing output with
   - values less than V1
   - Values >v1
   - Values < v2
   - Value = v1 or v2

Note: normal distribution is one of the many possible distribution that may be used to represent a process. This is only for explanation

Basic Statistics (a.k.a descriptive statistics)

1. Basic statistical analysis include
    a. Measure of central values
    b. Measure the spread around central values
    c. Overall distribution shape

2. Descriptive analysis of the input and output attribute values helps understand the current state and behaviour of the process

Descriptive statistics – central value and spread

Time to prepare the pizza

ABC Pizzeria	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
XYZ Pizza To Go	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

	ABCPizzeria	XYZ Pizza To Go
Mean	7.15	7.15
Median	7.20	7.20
Mode	7.7	7.7

Image source: unknown

1. ABC pizzeria is much more consistent than XYZ pizza in terms of time to prepare
   a. If you want to be sure about the amount of time you will need to get your lunch you will go to ABC Pizza
   b. Going to XYZ Pizza will make one less sure about the time required to gather the lunch
   c. However, if you are a risk taker, you may go to XYZ Pizza even when under time pressure as there is a chance of getting the lunch within 4.2 minutes. Saving a full 2+ minutes compared to ABC!

- In short, central tendency alone is not important, one should know the spread also to take any decision

Descriptive statistics – Spread / Variance

1. Variance is measured as the average of sum of squared difference between each data point (represented by xi) and the mean represented by μ

2. N represents the number of data points (when number of data points is small (lesser than 30), we replace N with N-1 (degrees of freedom)

3. Sum of difference between each data point and mean will always be zero. Hence square the difference and sum up the squared term as shown below

$$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Descriptive statistics - Standard Deviation

1. Interpreting variance (a squared term) is not intuitive. Instead we under root it to get Standard Deviation which has the same units as the variable

2. Standard deviation, is a measure of average spread i.e. on an average what is the difference between any data point and the central value of the variable

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

3. Continuing with the previous example, standard deviation is square root of 106.28 = 10.30

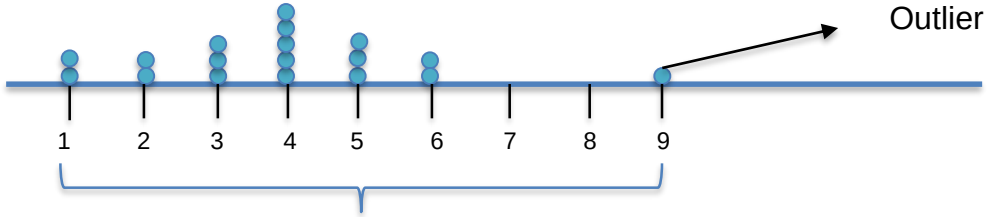
4. When the number of data points is small (less than 30), we replace N with N-1(degrees of freedom) in the denominator

Descriptive statistics - Range

1. Range is a measure of the difference between the max and the min values of any attribute

2. Range gives a first hand and quick view of the spread in the data

3. However, range will easily get distorted by outliers. Hence, it is not a reliable metric



Range = 9 – 1 = 8

Excluding outlier

Range = 6 – 1 = 5

Mean = 67/18 = 3.7
Median = 5

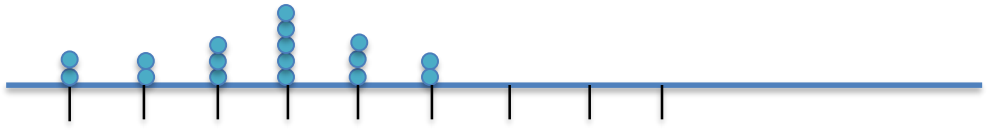
Excluding outlier

Mean = 58/17 = 3.4
Median = 3.5

4. Outliers will distort the central values (mean and median). Central values are expected values, an important characteristic of the process generating the data

Descriptive statistics - Quartiles

1. Quartile is a measure of dispersion of data around the median as central value
2. Median, which divides the data points into two equal half is also known as Quartile 2 or Q2
3. Quartile 1 or Q1 divides the values between minimum and Q2 into two equal half. In other words Q1 is that value which has 25% values below it and rest above
4. Quartile 3 or Q3 divides the data points between Q2 and max into two halves i.e. it has 75% of the values below it and 25% above



Q1 = 3
Median (Q2) = 4
Q3 = 5

a. The gap between Q3 and Q2 = 1 units while distance from Q2 to Q1 = 1 unit. The body of the distribution is intact
b. Distance between min and Q1 is 2 units and distance from Q3 to max is 1 unit. Left side tail is longer the right side tail. Indicates a skew on left side
c. Overall a slight asymmetry is noticed in the distribution

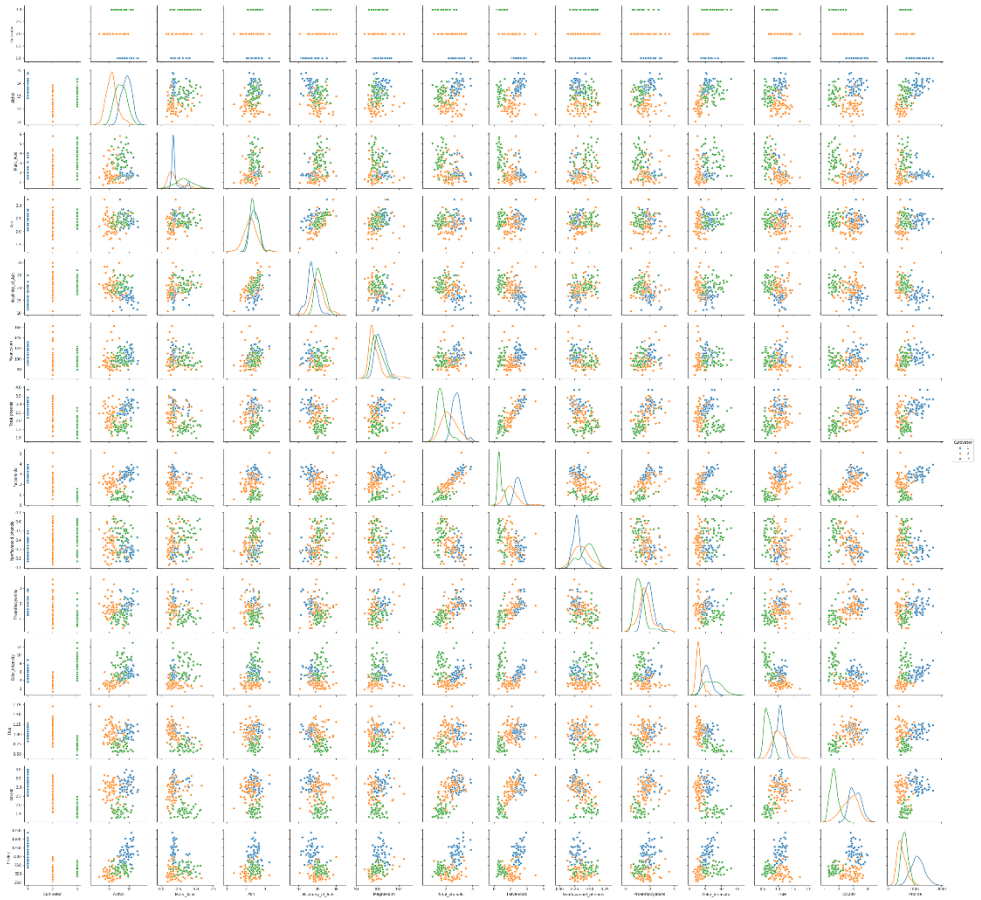
Descriptive statistics – Lab 1

1. Run descriptive statistics on wines dataset and discuss your findings –
    1. import pandas as pd
    2. wine_df = pd.read_csv('d:\\ML_Data\\wine.csv', names = ["Cultivator", "Alchol", "Malic_Acid", "Ash", "Alcalinity_of_Ash", "Magnesium", "Total_phenols", "Falvanoids", "Nonflavanoid_phenols", "Proanthocyanins", "Color_intensity", "Hue", "OD280", "Proline"])
    3. wine_df.describe().transpose()

	count	mean	std	min	25%	50%	75%	max
Cultivator	178	1.938202	0.775035	1	1	2	3	3
Alchol	178	13.00062	0.811827	11.03	12.3625	13.05	13.6775	14.83
Malic_Acid	178	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.8
Ash	178	2.366517	0.274344	1.36	2.21	2.36	2.5575	3.23
AlcalinityAsh	178	19.49494	3.339564	10.6	17.2	19.5	21.5	30
Magnesium	178	99.74157	14.28248	70	88	98	107	162
Total_phenols	178	2.295112	0.625851	0.98	1.7425	2.355	2.8	3.88
Falvanoids	178	2.02927	0.998859	0.34	1.205	2.135	2.875	5.08
Nonflavanoid	178	0.361854	0.124453	0.13	0.27	0.34	0.4375	0.66
Proanthocyanins	178	1.590899	0.572359	0.41	1.25	1.555	1.95	3.58
Color_intensity	178	5.05809	2.318286	1.28	3.22	4.69	6.2	13
Hue	178	0.957449	0.228572	0.48	0.7825	0.965	1.12	1.71
OD280	178	2.611685	0.70999	1.27	1.9375	2.78	3.17	4
Proline	178	746.8933	314.9075	278	500.5	673.5	985	1680

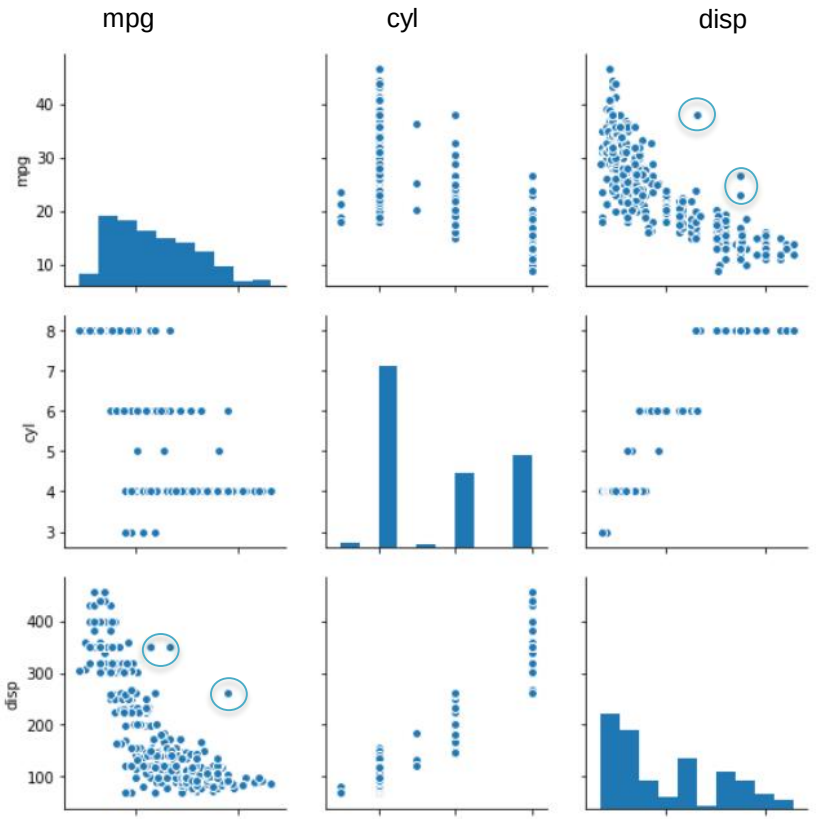
Descriptive statistics (univariate / bi-variate analysis

1. Pair plot –
    1. import seaborn as sns
    2. sns.pairplot(wine_df, hue='Cultivator' , diag_kind = 'kde')



Descriptive statistics – Pairplot analysis

1. Pairplot is a square matrix. Each row and each column in the matrix is one attribute in the dataframe.
2. The relation between the attributes (how they influence one another) is reflected in the scatter plot
3. When the row number and column number in the matrix are same, we are comparing an attribute with itself. Scatter plot will not make sense. Instead, we get histogram by default
4. Histogram helps visualize the distribution characteristics on each attribute.
5. Scatter plot reflects how one attribute interacts with other. Is it a strong or weak relationship, is it positive or negative relationship or, is there no relationship
6. Can visualize outlier / leverage points (shown in red circle) which would not stand out in univariate analysis



Descriptive statistics – PairPlot analysis

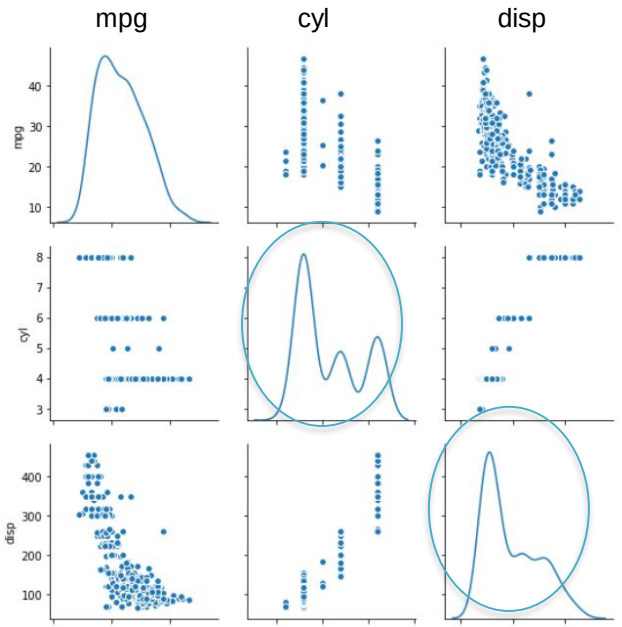
7. Using KDE in pairplot, we can identify potential clusters in the attributes

8. Presence of multiple peaks (red circle) on an attribute that is not the target (predicted column) may be due to mix up of gaussians

9. Mix up of gaussians will result in long tails and less reliable central values.

10. Less reliable the central values, less reliable the models that hinge on the central values

11. One can analyze the attributes and their relations to the target variables by each gaussian for richer insights



Cars-mpg dataset
Predict mpg based on other attributes

Descriptive statistics – PairPlot analysis

12. One can analyze the attributes and their relations to the target variables by each gaussian for richer insights
13. In the figures, each gaussian is marked by different color
14. Orange color represents small cars (low horsepower, less displacement and weight) while blue represents large cars
15. Looks like this data consists of three types of cars (small, mid-size and large cars)
16. The strength of relation between mpg (target column) and the various attributes is different for the different types of cars
17. Displacement seems to be strong predictor of mpg for mid size and large cars but not so for small cars!
18. Does it make sense to build one single model representing all the cars when their characteristics are different?

Fig A

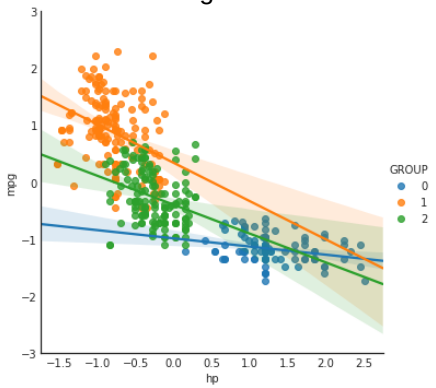
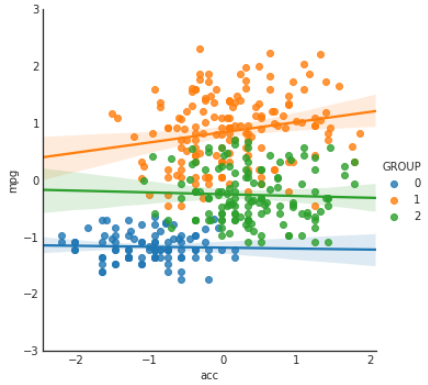


Fig B



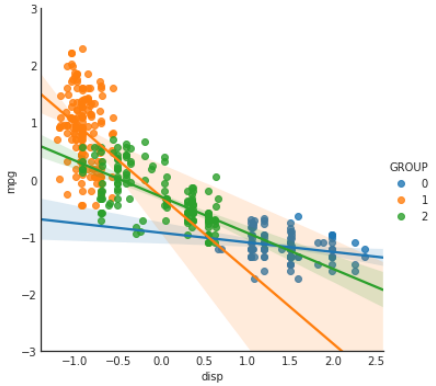


Fig C

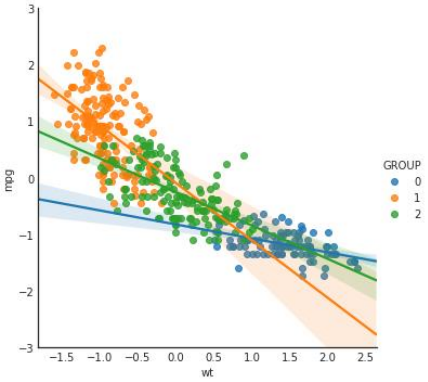


Fig D

Cars-mpg dataset
Predict mpg based on other attributes

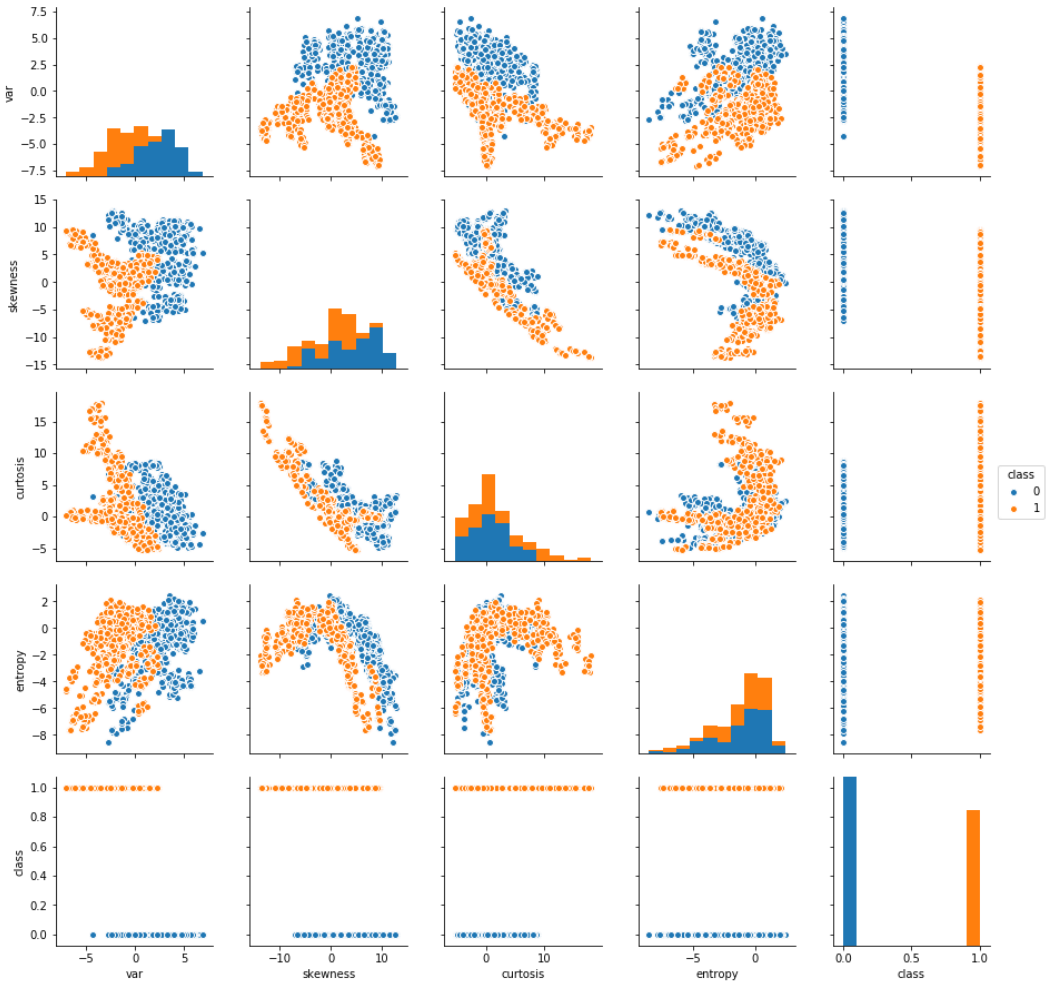
Descriptive statistics – PairPlot analysis

19. In case of classification problems, we will naturally find as many peaks as the classes in the target column. This is not the same as mix of multi-gaussians on attributes.

20. In the figure, we see in the last column, last row is the target variable “class” which is binary (blue – real notes, orange – fake notes)

21. Pairplot helps us to identify attributes that can help us separate these two classes.
    a. In the last row, class Vs attributes – Higher value of curtosis is only for fake
    b. In the last row, first column – higher values of ‘var’ is only for the real notes
    c. For rest of the data points, the probability of belonging to one class or other changes gradually in “skewness”
    d. All these distribution characteristics together help build good models with high score in classification on this dataset

22. The distribution of the two classes and the shape of the boundaries may also help guess which algorithms may do well and which not
    a. Decision tree being axis parallel algorithm may not do as well as a logistic or KNN in this case



data_banknote_authentication
Classify notes into real or fake

Descriptive Statistics Findings

1. Distribution characteristics of the attributes (univariate analysis)
    1. Sample statistics including central values (mean, median)
    2. Spread inform of inter quartile range, distance of central values from quartiles
    3. Skews in the distributions, long tails, potential outliers, missing values
    4. Mix of gaussians

2. Bi-variate analysis helps understand-
    1. Degree of interactions between the variables
    2. Distribution patterns of classes (are they linearly separable)
    3. Presence of hidden outliers , leverage points
    4. Identify potential good attributes
    5. May help in choice of algorithms to build models on

3. Together, the univariate and bi-variate analysis give an insight into the process that generated the