# Unit -2
# ( topic : Student's t-test ROC Curves & EDA )


National Forensic Sciences University

Knowledge | Wisdom | Fulfilment

An Institution of National Importance
(Ministry of Home Affairs, Government of India)



By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

# Test of Significance

For applying the tests of significance, we first set up a hypothesis which is a definite statement about the population parameter called null hypothesis denoted by $H_0$.

Any hypothesis which is complementary to the null hypothesis ($H_0$) is called an alternative hypothesis denoted by $H_1$.

For example, if we want to test the null hypothesis that the population has a specified mean $\mu_0$, then we have $H_0 : \mu = \mu_0$

Alternative hypothesis will be

- $H_1 : \mu \neq \mu_0$ ( $\mu > \mu_0$ or $\mu < \mu_0$) (Two tailed alternative hypothesis).
- $H_1 : \mu > \mu_0$ (Right tailed alternative hypothesis or single tailed).
- $H_1 : \mu < \mu_0$ (Left tailed alternative hypothesis or single tailed).

Hence alternative hypothesis helps to know whether the test is two tailed or one tailed test.

The main aim of the sampling theory is to draw a valid conclusion about the population parameters. On the basis of the same results. In doing this we may commit the following two type of errors.

- Type I error: When $H_0$ is true, we may reject it.

$$P(\text{Reject } H_0 \text{ when it is true}) = P\left(\text{Reject}\frac{H_0}{H_0}\right) = \alpha$$

Where $\alpha$ is called the size of the type I error

- Type II error: When $H_0$ is wrong we may accept it.

$$P(\text{Accept } H_0 \text{ when it is wrong}) = P\left(\text{Accept}\frac{H_0}{H_1}\right) = \beta$$

Where $\beta$ is called the size of the type II error

# Student's t-test

The one-sample t-test is the statistical test used to determine whether an unknown population mean is different from a specific value.
For example, comparing the mean height of the students with respect to the national average height of an adult.

To test whether the mean of a sample drawn from a normal population deviates significantly from a stated value when variance of the population is unknown.

$H_0$ : There is no significant difference between the sample mean $\bar{X}$ and the population mean $\mu$ i. e., we use the static

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}, \quad \text{where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This test static is known as one sample t-test.

# Question : 1

When operating normally, a manufacturing process produces tablets for which the Mean weight is intended to be 5 grams. Nothing is known about the overall Standard Deviation. A random sample of 12 tablets produces the following data:

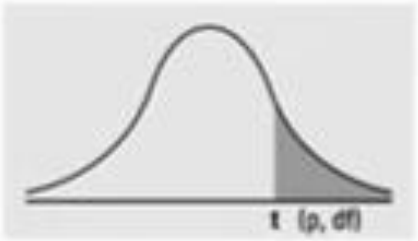
Sample Mean = 5.05g
Sample SD = 0.1

i) Conduct a suitable hypothesis test at the 5% level to test the null hypothesis that the mean weight of tablets being produced is not above 5g.

$$t_c = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

In Exam H0 is given

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).




| df/p | 0.40 | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.324920 | 1.000000 | 3.077684 | 6.313752 | 12.70620 | 31.82052 | 63.65674 | 636.6192 |
| 2 | 0.288675 | 0.816497 | 1.885618 | 2.919986 | 4.30265 | 6.96456 | 9.92484 | 31.5991 |
| 3 | 0.276671 | 0.764892 | 1.637744 | 2.353363 | 3.18245 | 4.54070 | 5.84091 | 12.9240 |
| 4 | 0.270722 | 0.740697 | 1.533206 | 2.131847 | 2.77645 | 3.74695 | 4.60409 | 8.6103 |
| 5 | 0.267181 | 0.726687 | 1.475884 | 2.015048 | 2.57058 | 3.36493 | 4.03214 | 6.8688 |
| 6 | 0.264835 | 0.717558 | 1.439756 | 1.943180 | 2.44691 | 3.14267 | 3.70743 | 5.9588 |
| 7 | 0.263167 | 0.711142 | 1.414924 | 1.894579 | 2.36462 | 2.99795 | 3.49948 | 5.4079 |
| 8 | 0.261921 | 0.706387 | 1.396815 | 1.859548 | 2.30600 | 2.89646 | 3.35539 | 5.0413 |
| 9 | 0.260955 | 0.702722 | 1.383029 | 1.833113 | 2.26216 | 2.82144 | 3.24984 | 4.7809 |
| 10 | 0.260185 | 0.699812 | 1.372184 | 1.812461 | 2.22814 | 2.76377 | 3.16927 | 4.5869 |
| 11 | 0.259556 | 0.697445 | 1.363430 | 1.795885 | 2.20099 | 2.71808 | 3.10581 | 4.4370 |
| 12 | 0.259033 | 0.695483 | 1.356217 | 1.782288 | 2.17881 | 2.68100 | 3.05454 | 43178 |

When operating normally, a manufacturing process produces tablets for which the Mean weight is intended to be 5 grams. Nothing is known about the overall Standard Deviation. A random sample of 12 tablets produces the following data:

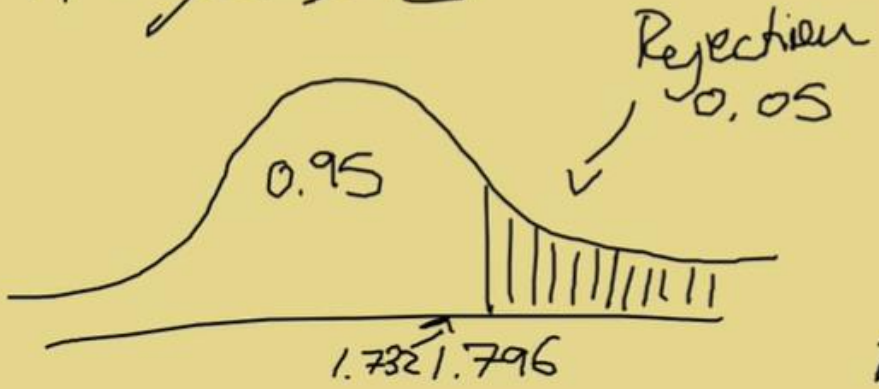
Sample Mean = 5.05g
Sample SD = 0.1

i) Conduct a suitable hypothesis test at the 5% level to test the null hypothesis that the mean weight of tablets being produced is not above 5g.

$n = 12 \quad \bar{x} = 5.05 \quad s = 0.1$

$H_0 : \mu = 5$

$H_1 : \mu > 5$




$$t = \frac{5.05 - 5}{0.1/\sqrt{12}}$$

$$\approx 1.732$$

Fail to reject $H_0$

# Question : 2

Celia has had a regular car journey to work which has previously taken an average of 42 minutes. The local council has recently altered the road layout and Celia is convinced that this has increased the journey time. To investigate further, she records the time taken for 12 journeys. The 12 journeys have a mean time of 50 minutes and a standard deviation of 15 minutes.

Conduct a suitable hypothesis test at the 5% level to identify whether there is evidence for her contention that the average journey time has increased.

Celia has had a regular car journey to work which has previously taken an average of 42 minutes. The local council has recently altered the road layout and Celia is convinced that this has increased the journey time. To investigate further, she records the time taken for 12 journeys. The 12 journeys have a mean time of 50 minutes and a standard deviation of 15 minutes.

Conduct a suitable hypothesis test at the 5% level to identify whether there is evidence for her contention that the average journey time has increased.

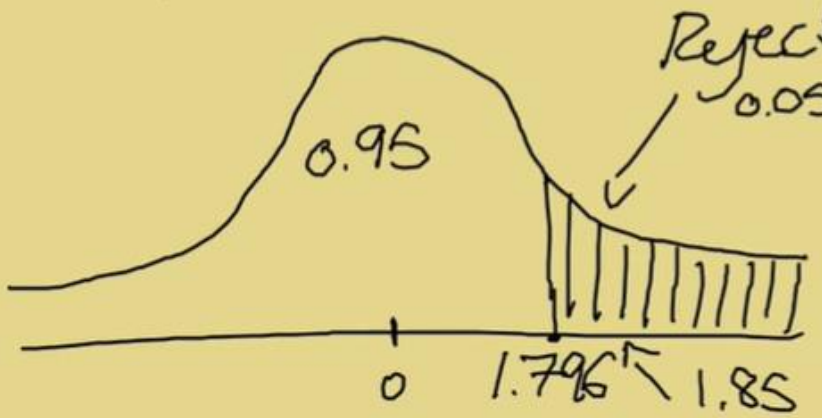
$n = 12 \quad \bar{x} = 50$

$s = 15$

$H_0 : \mu = 42$

$H_1 : \mu > 42$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



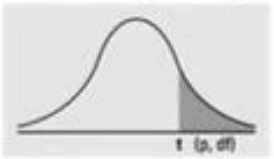

$$t = \frac{50 - 42}{15/\sqrt{12}}$$

Reject $H_0$. $t > 1.85$

# Question : 3

A researcher is interest in finding out whether the average regular gasoline price is higher than $2.45 in Mid-West region. The sample analysed consists of 25 observations, a sample mean of $2.65 and a sample standard deviation of $0.35. (a) State the null and alternative hypothesis. (b) At a 99% confidence level, is there enough evidence to discard the null hypothesis?

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).




| df/p | 0.40 | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.324920 | 1.000000 | 3.077684 | 6.313752 | 12.70620 | 31.82052 | 63.65674 | 636.6192 |
| 2 | 0.288675 | 0.816497 | 1.885618 | 2.919986 | 4.30265 | 6.96456 | 9.92484 | 31.5991 |
| 3 | 0.276671 | 0.764892 | 1.637744 | 2.353363 | 3.18245 | 4.54070 | 5.84091 | 12.9240 |
| 4 | 0.270722 | 0.740697 | 1.533206 | 2.131847 | 2.77645 | 3.74695 | 4.60409 | 8.6103 |
| 5 | 0.267181 | 0.726687 | 1.475884 | 2.015048 | 2.57058 | 3.36493 | 4.03214 | 6.8688 |
| 6 | 0.264835 | 0.717558 | 1.439756 | 1.943180 | 2.44691 | 3.14267 | 3.70743 | 5.9588 |
| 7 | 0.263167 | 0.711142 | 1.414924 | 1.894579 | 2.36462 | 2.99795 | 3.49948 | 5.4079 |
| 8 | 0.261921 | 0.706387 | 1.396815 | 1.859548 | 2.30600 | 2.89646 | 3.35539 | 5.0413 |
| 9 | 0.260955 | 0.702722 | 1.383029 | 1.833113 | 2.26216 | 2.82144 | 3.24984 | 4.7809 |
| 10 | 0.260185 | 0.699812 | 1.372184 | 1.812461 | 2.22814 | 2.76377 | 3.16927 | 4.5869 |
| 11 | 0.259556 | 0.697445 | 1.363430 | 1.795885 | 2.20099 | 2.71808 | 3.10581 | 4.4370 |
| 12 | 0.259033 | 0.695483 | 1.356217 | 1.782288 | 2.17881 | 2.68100 | 3.05454 | 43178 |
| 13 | 0.258591 | 0.693829 | 1.350171 | 1.770933 | 2.16037 | 2.65031 | 3.01228 | 4.2208 |
| 14 | 0.258213 | 0.692417 | 1.345030 | 1.761310 | 2.14479 | 2.62449 | 2.97684 | 4.1405 |
| 15 | 0.257885 | 0.691197 | 1.340606 | 1.753050 | 2.13145 | 2.60248 | 2.94671 | 4.0728 |
| 16 | 0.257599 | 0.690132 | 1.336757 | 1.745884 | 2.11991 | 2.58349 | 2.92078 | 4.0150 |
| 17 | 0.257347 | 0.689195 | 1.333379 | 1.739607 | 2.10982 | 2.56693 | 2.89823 | 3.9651 |
| 18 | 0.257123 | 0.688364 | 1.330391 | 1.734064 | 2.10092 | 2.55238 | 2.87844 | 3.9216 |
| 19 | 0.256923 | 0.687621 | 1.327728 | 1.729133 | 2.09302 | 2.53948 | 2.86093 | 3.8834 |
| 20 | 0.256743 | 0.686954 | 1.325341 | 1.724718 | 2.08596 | 2.52798 | 2.84534 | 3.8495 |
| 21 | 0.256580 | 0.686352 | 1.323188 | 1.720743 | 2.07961 | 2.51765 | 2.83136 | 3.8193 |
| 22 | 0.256432 | 0.685805 | 1.321237 | 1.717144 | 2.07387 | 2.50832 | 2.81876 | 3.7921 |
| 23 | 0.256297 | 0.685306 | 1.319460 | 1.713872 | 2.06866 | 2.49987 | 2.80734 | 3.7676 |
| 24 | 0.256173 | 0.684850 | 1.317836 | 1.710882 | 2.06390 | 2.49216 | 2.79694 | 3.7454 |
| 25 | 0.256060 | 0.684430 | 1.316345 | 1.708141 | 2.05954 | 2.48511 | 2.78744 | 3.7251 |
| 26 | 0.255955 | 0.684043 | 1.314972 | 1.705618 | 2.05553 | 2.47863 | 2.77871 | 3.7066 |
| 27 | 0.255858 | 0.683685 | 1.313703 | 1.703288 | 2.05183 | 2.47266 | 2.77068 | 3.6896 |

A researcher is interested in finding out whether the average regular gasoline price is higher than \$2.45 in Mid-West region. The sample analysed consists of 25 observations, a sample mean of \$2.65 and a sample standard deviation of \$0.35. (a) State the null and alternative hypothesis. (b) At a 99% confidence level, is there enough evidence to discard the null hypothesis?

$n = 25 \quad \bar{x} = 2.65$

$S = 0.35$

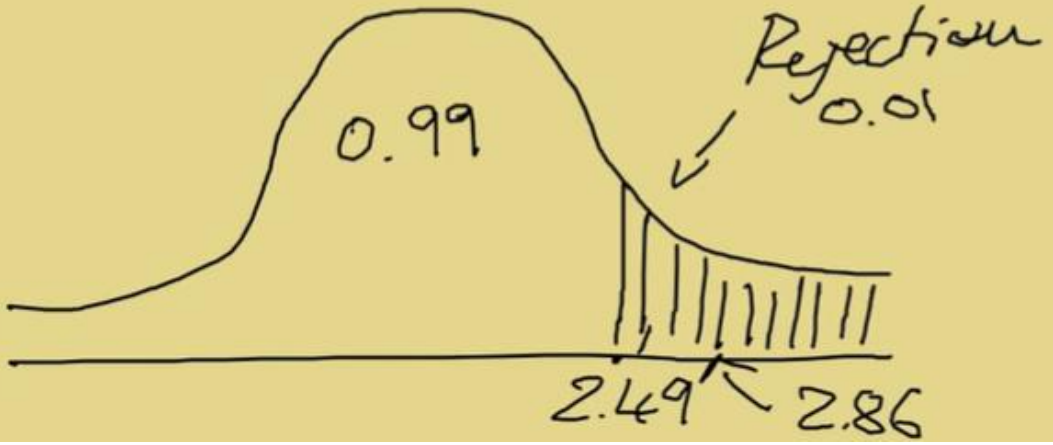
$H_0: \mu = 2.45$

$H_1: \mu > 2.45$

$$t_c = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

$$t_c = \frac{2.65 - 2.45}{0.35/\sqrt{25}}$$

$\approx 2.86$




Reject

# Receiver Operating Characteristics

**The Receiver Operating Characteristics(ROC) curve**

**The ROC curve is a evaluation measure**

**that is based on two basic evaluation measures**

**- specificity and sensitivity**

- Specificity = True Negative Rate
- Sensitivity = Recall = True Positive Rate

| Instance | Yes |
|---|---|
| 1 | 0.008 |
| 2 | 0.011 |
| 3 | 0.021 |
| 4 | 0.009 |
| 5 | 0.014 |
| 6 | 0.015 |
| 7 | 0.012 |
| 8 | 0.015 |

| | Cut-off = 0.020 | | | Cut-off = 0.015 | | | Cut-off = 0.010 | | |
|---|---|---|---|---|---|---|---|---|---|
| Actual | Instance | Predict | Type | Instance | Predict | Type | Instance | Predict | Type |
| N | 1 | N | TN | 1 | N | TN | 1 | N | TN |
| N | 2 | N | TN | 2 | N | TN | 2 | Y | FP |
| Y | 3 | Y | TP | 3 | Y | TP | 3 | Y | TP |
| N | 4 | N | TN | 4 | N | TN | 4 | N | TN |
| N | 5 | N | TN | 5 | N | TN | 5 | Y | FP |
| N | 6 | N | TN | 6 | Y | FP | 6 | Y | FP |
| N | 7 | N | TN | 7 | N | TN | 7 | Y | FP |
| Y | 8 | N | FN | 8 | Y | TP | 8 | Y | TP |

# True positive rate vs False positive rate

**True positive rate (TPR) = $TP/(TP+FN)$**

**and False positive rate (FPR) = $FP/(FP+TN)$**

**Use different cut-off thresholds (0.00, 0.01, 0.02,..., 1.00),**

**calculate the TPR and FPR, and plot them into graph.**

**That is receiver operating characteristic (ROC) curve.**

**Example**

| TP=1 | FN=1 |
|---|---|
| FP=0 | TN=6 |

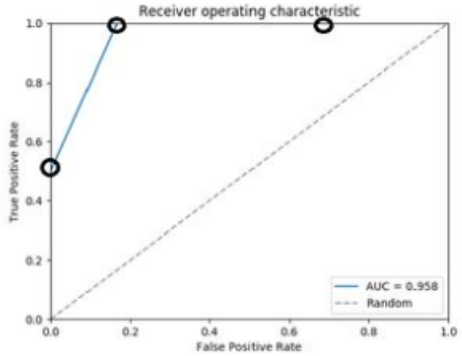
TPR = 0.5
FPR = 0

| TP=2 | FN=0 |
|---|---|
| FP=1 | TN=5 |

TPR = 1
FPR = 0.167

| TP=2 | FN=0 |
|---|---|
| FP=4 | TN=2 |

TPR = 1
FPR = 0.667

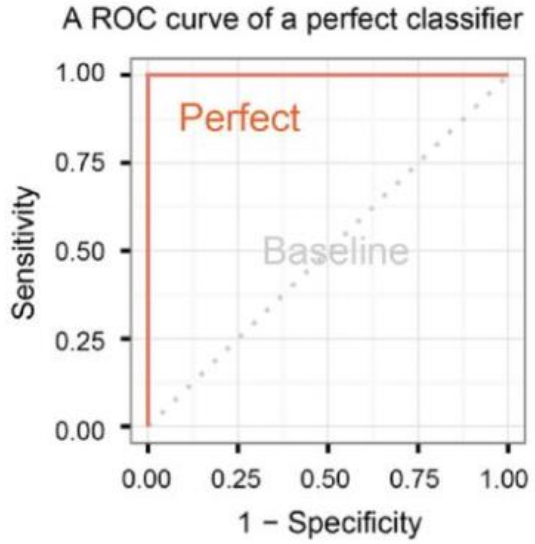


# Example of Changing Thresholds

| Score | Prediction for (1) | Prediction for (0.7) | Prediction for (0.6) | Prediction for (0.5) | Prediction for (0) | Y |
|---|---|---|---|---|---|---|
| .25 | | | | | | 0 |
| .45 | | | | | | 0 |
| .55 | | | | | | 1 |
| .67 | | | | | | 0 |
| .82 | | | | | | 1 |
| .95 | | | | | | 1 |

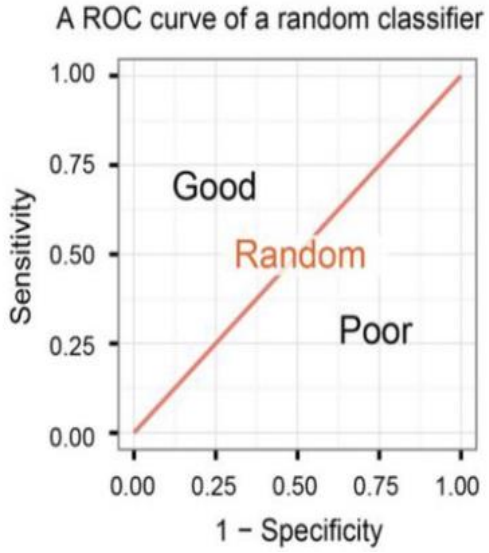
# Perfect Classifier

**A classifier with the perfect performance level shows**
**a combination of two straight lines**
**It is important to notice that classifiers with meaningful performance levels**
**usually lie in the area between the random ROC curve and the perfect ROC curve**




**A classifier with the random performance level always shows a straight line**
**Two areas separated by this ROC curve**

- ROC curves in the area with the top left corner indicate good performance levels
- ROC curves in the other area with the bottom right corner indicate poor performance levels

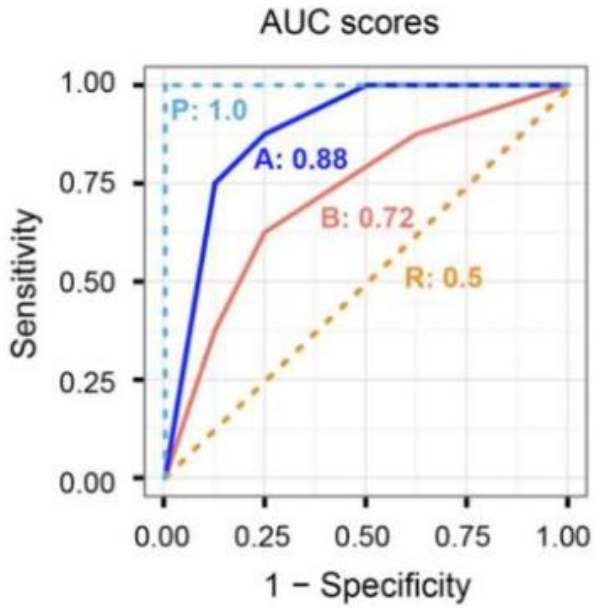


# Area Under Curve (AUC)

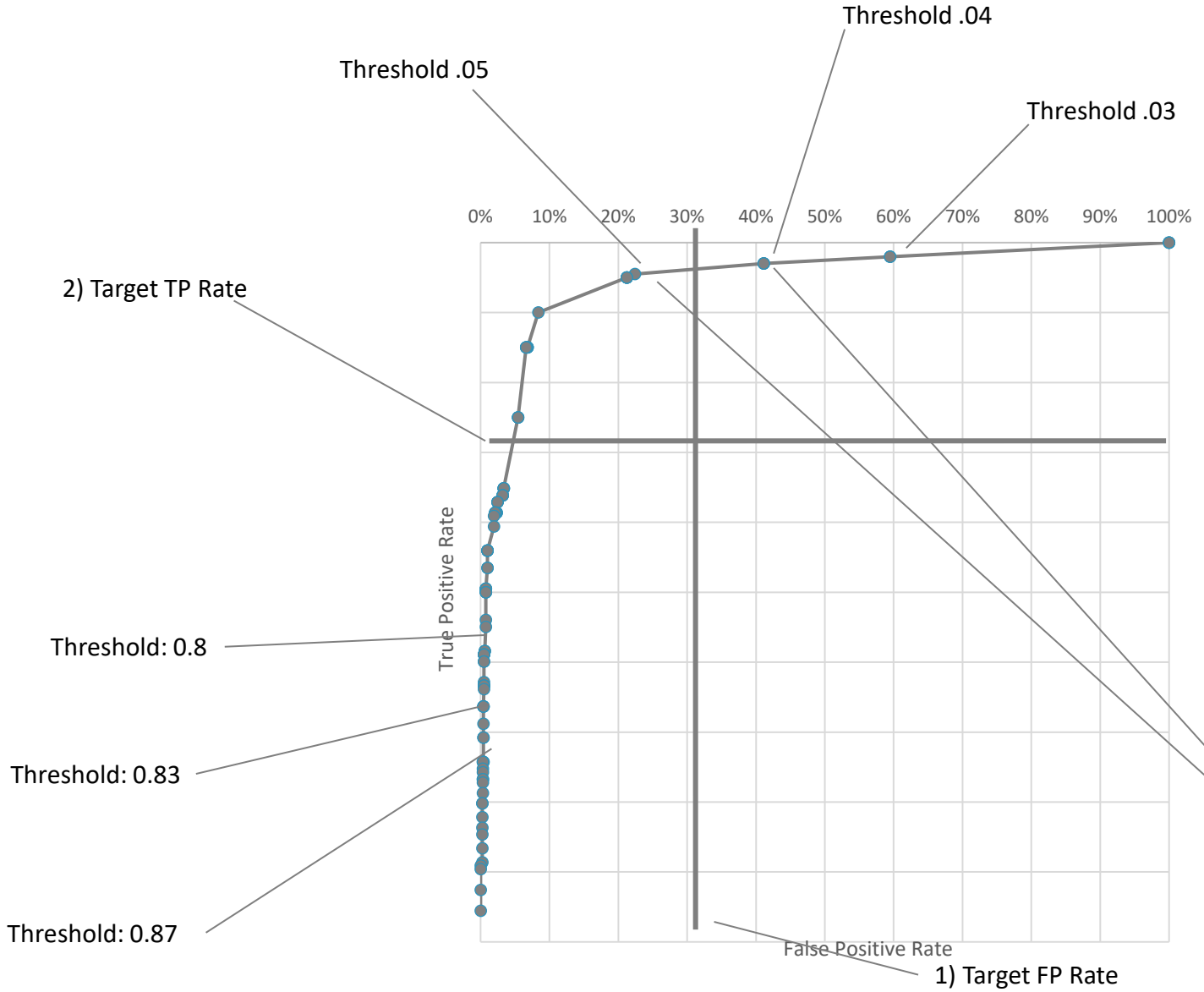
**AUC(Area under the ROC curve) score**

- An advantage of using ROC curve is a single measure called AUC score
- As the name indicates, it is an area under the curve calculated in the ROC space
- Although the theoretical range of AUC score is between 0 and 1,
  the actual scores of meaningful classifiers are greater than 0.5,
  which is the AUC score of a random classifier

- ROC curves clearly shows classifiers A outperforms classifier B



# Threshold value




Threshold value decrease :

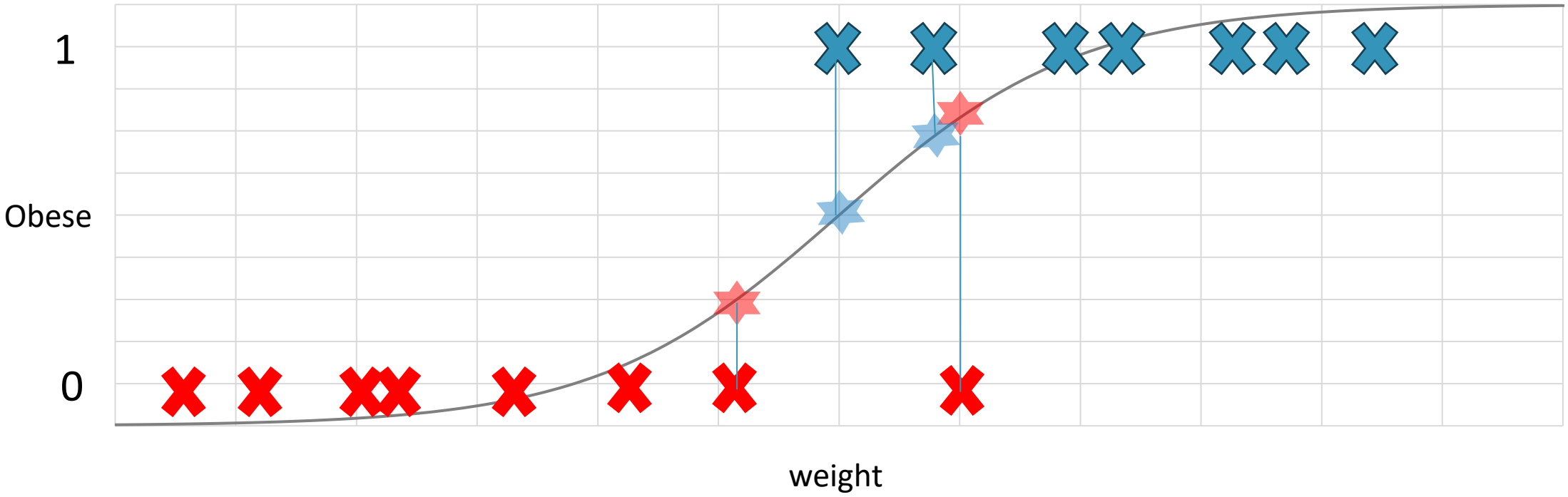
Positive prediction increase.
True/False Positive increase.
True/False Negative decrease.

Threshold value increase :

Negative prediction increase.
True/False Negative increase.
True/False Positive decrease.

# Classifications and Probability Estimates

- What happens if you vary the threshold?



# Which Model should you use?

| | False Positive Rate | False Negative Rate |
|---|---|---|
| Model 1 | 41% | 3% |
| Model 2 | 5% | 25% |

Actually the same model
- different thresholds

Mistakes have different costs:

- Disease Screening – LOW FN Rate
- Spam filtering – LOW FP Rate

# Quiz Question

1. Imagine a phishing or malware classification model where phishing and malware websites are in the class labeled 1 (true) and harmless websites are in the class labeled 0 (false). This model mistakenly classifies a legitimate website as malware. What is this called?

- [ ] A false negative
- [ ] A true negative
- [ ] A false positive
- [ ] A true positive

1. Imagine a phishing or malware classification model where phishing and malware websites are in the class labeled 1 (true) and harmless websites are in the class labeled 0 (false). This model mistakenly classifies a legitimate website as malware. What is this called?

- [ ] A false negative
- [ ] A true negative
- [x] A false positive

  A negative example (legitimate site) has been wrongly classified as a positive example (malware site).

  Correct answer.
- [ ] A true positive

# Quiz Question

**2. In general, what happens to the number of false positives when the classification threshold increases? What about true positives? Experiment with the slider above.**

- ☐ True positives increase. False positives decrease.
- ☐ Both true and false positives increase.
- ☐ Both true and false positives decrease.

**3. In general, what happens to the number of false negatives when the classification threshold increases? What about true negatives? Experiment with the slider above.**

- ☐ True negatives increase. False negatives decrease.
- ☐ Both true and false negatives increase.
- ☐ Both true and false negatives decrease.

**2. In general, what happens to the number of false positives when the classification threshold increases? What about true positives? Experiment with the slider above.**

- True positives increase. False positives decrease. ☐
- Both true and false positives decrease. ✓

  As the threshold increases, the model will likely predict fewer positives overall, both true and false. A spam classifier with a threshold of .9999 will only label an email as spam if it considers the classification to be at least 99.99% likely, which means it is highly unlikely to mislabel a legitimate email, but also likely to miss actual spam email.

  Correct answer.
- Both true and false positives increase. ☐

**3. In general, what happens to the number of false negatives when the classification threshold increases? What about true negatives? Experiment with the slider above.**

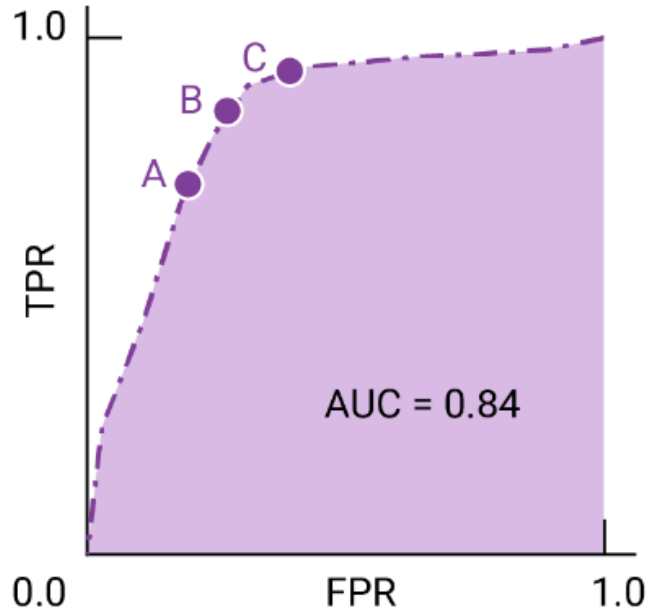
- True negatives increase. False negatives decrease. ☐
- Both true and false negatives decrease. ☐
- Both true and false negatives increase. ✓

  As the threshold increases, the model will likely predict more negatives overall, both true and false. At a very high threshold, almost all emails, both spam and not-spam, will be classified as not-spam.

  Correct answer.

# Quiz Question

Imagine a situation where it's better to allow some spam to reach the inbox than to send a business-critical email to the spam folder. You've trained a spam classifier for this situation where the positive class is spam and the negative class is not-spam. Which of the following points on the ROC curve for your classifier is preferable?




Answer is A

1. Increasing a binary classifier's threshold value is likely to produce which of the following effects?

- ◯ False positives increase
- ◯ False positives decrease
- ◯ False positives and false negatives both increase
- ◯ False positives and false negatives both decrease

2. The dataset that you split into train, test and evaluate sets has 9,998 negative examples and 2 positive examples. The resulting model has an accuracy rate of 99.9%. Can you trust this model based on that accuracy metric?

- ◯ Yes
- ◯ No

1. Increasing a binary classifier's threshold value is likely to produce which of the following effects?

- ○ False positives increase
- ◉ False positives decrease ✓ Correct!
- ○ False positives and false negatives both increase
- ○ False positives and false negatives both decrease

2. The dataset that you split into train, test and evaluate sets has 9,998 negative examples and 2 positive examples. The resulting model has an accuracy rate of 99.9%. Can you trust this model based on that accuracy metric?

- ○ Yes
- ◉ No ✓ Accuracy is not a good metric to use to evaluate models with class-imbalanced datasets, like in the scenario here. A model that always predicted the negative class would have an accuracy of 99.9% even though it would have no ability to identify positive examples. For class-imbalanced datasets, you should consider using other evaluation metrics, such as precision or recall.

3. In general, when precision increases, what happens to recall?

- ○ Recall is unaffected.
- ○ Recall increases exponentially.
- ◉ Recall decreases.

✓ Precision and recall tend to have an inverse relationship. When precision increases, recall tends to decrease.

4. True or False: The points on a binary classification model's ROC (receiver-operating characteristic) curve closest to (1,1) (the upper-right corner) generally represent the best-performing thresholds for the model

# EDA(Exploratory Data Analysis)

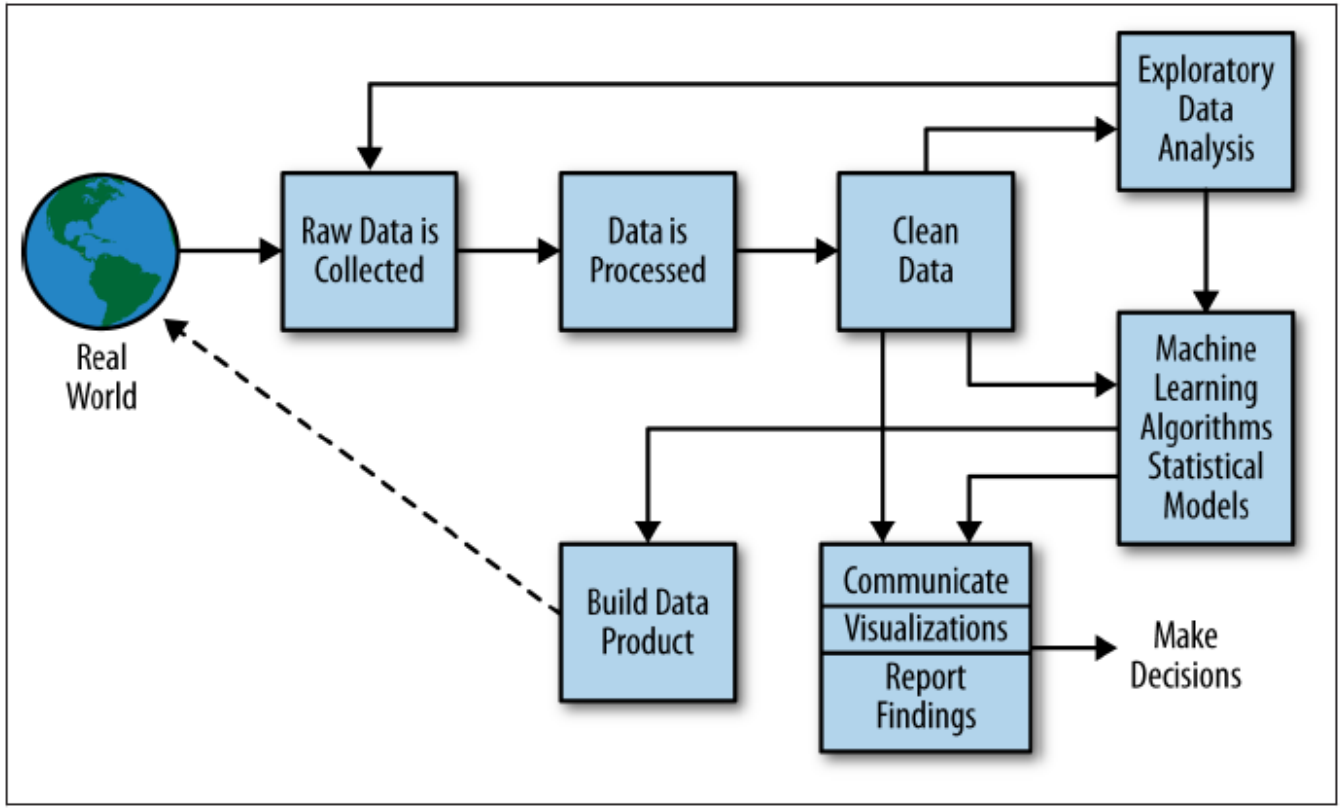
**Is EDA and data analysis the same?**

- Data analysis is a broad term involving different types of analysis like descriptive, diagnostic, predictive, and prescriptive. EDA is synonymous with descriptive analysis, where one explores the hidden relationships and patterns in the available data.

**How one should write an EDA report?**

- An EDA report must thoroughly explain the dataset's variables, their correlation, and any preprocessing performed on the dataset to make it suitable for applying a machine learning algorithm for further use in the organization. This report consists of multiple relevant visualizations that present a complete picture of the information in the available data and hence, can be used by senior management to make data-driven decisions.

# The Data Science Process



# Types of Exploratory Data Analysis

There are four main types of EDA:

1.Univariate non-graphical

Mean, Mode, Median, Variance, Standard deviation

1.Univariate graphical

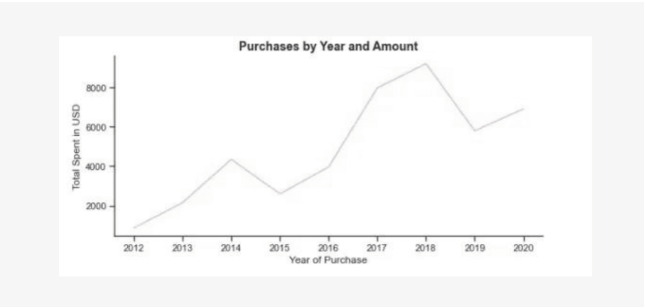
Stem-and-leaf Plots, Histogram, Group Chart, Percentage Chart, Box-plot

1.Multivariate nongraphical

Correlation, Covariance, Regression Analysis, ANOVA

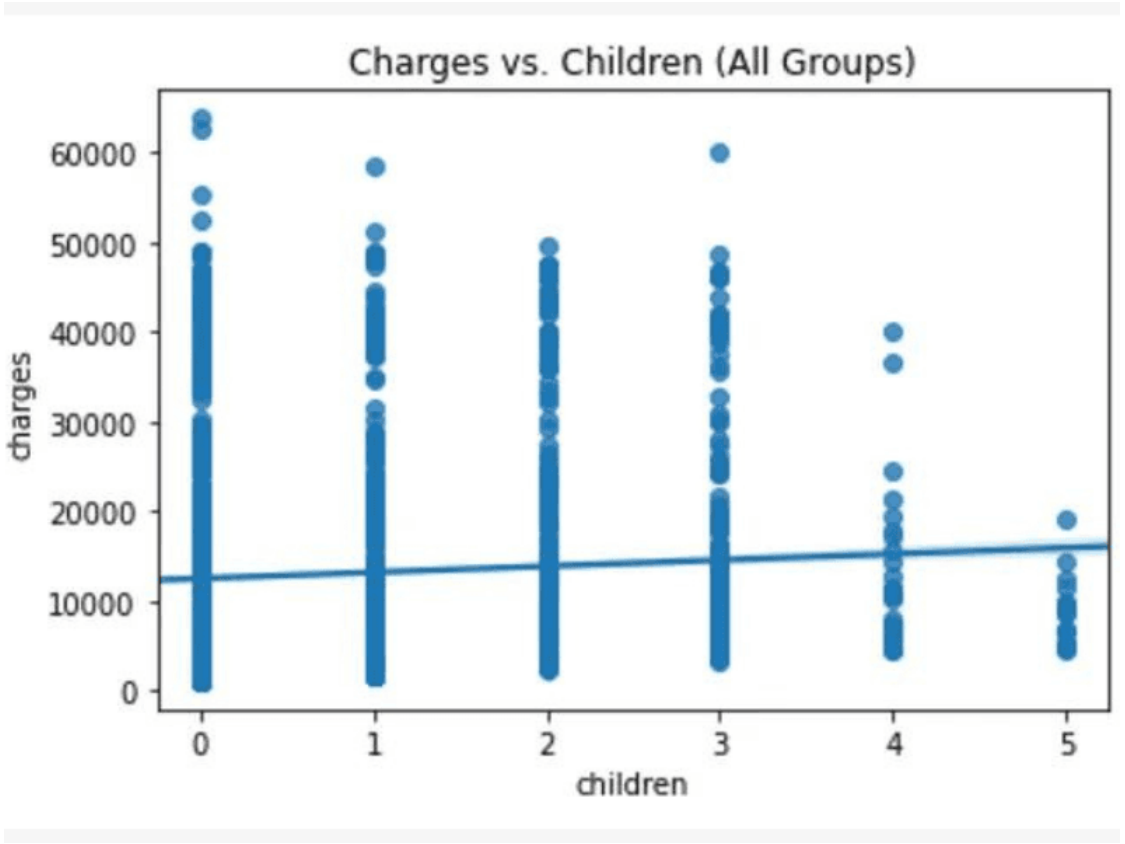
1.Multivariate graphical

Scatter plot, Run chart, Bubble chart, heat map, Box-plot



# Steam-Leaf Plot – Box Plot

| Stem | Leaf | # |
|---|---|---|
| 60 | 0 | 1 |
| 58 | | |
| 56 | 0 | 1 |
| 54 | 000000 | 6 |
| 52 | 00000 | 5 |
| 50 | 00000 | 5 |
| 48 | 00000000 | 8 |
| 46 | 000000000000 | 12 |
| 44 | 0000000000000000000 | 19 |
| 42 | 000000000000 | 12 |
| 40 | 0000000000000000000000 | 22 |
| 38 | 000000000000000 | 15 |
| 36 | 000000000000000000000000000 | 25 |
| 34 | 00000000000000000 | 17 |
| 32 | 000000000000000000000000 | 25 |
| 30 | 0000000000000000000000000000 | 28 |
| 28 | 000000000000000 | 15 |
| 26 | 0000000000000000000 | 18 |
| 24 | 0000000000000 | 13 |
| 22 | 00000000000 | 11 |
| 20 | 00000000 | 8 |
| 18 | 0000000 | 7 |
| 16 | 00000000 | 8 |
| 14 | 0000 | 4 |
| 12 | 000 | 3 |
| 10 | 0 | 1 |
| 8 | 0 | 1 |




Steam-Leaf Plot

# EDA Tools (for plot, Graph & summary statistic)

## 1. Python

EDA libraries like:

Matplotlib, Pandas, Seaborn, NumPy

You can find many open-source packages in Python, such as **D-Tale, AutoViz, PandasProfiling**, etc., that can automate the entire exploratory data analysis process and save time.

## 2. R

R programming language is a regularly used option to make statistical observations and analyze data, i.e., perform detailed EDA by data scientists and statisticians. Like Python, R is also an open-source programming language suitable for statistical computing and graphics.

Apart from the commonly used libraries like: ggplot, Leaflet, Lattice there are several powerful R libraries for automated EDA, such as **Data Explorer, SmartEDA, GGally**, etc.

## 3. MATLAB

MATLAB is a well-known commercial tool among engineers since it has a very strong mathematical calculation ability. Due to this, it is possible to use MATLAB for EDA, but it requires some basic knowledge of the MATLAB programming language.

# Philosophy of EDA

**John Tukey is considered as father of EDA.**

**Based on following principles:**

1. **Revelation : emphasizes using graph**
2. **Resistance : general pattern which cover majority of Data.**
3. **Reexpression :  to new scale to simplify the data analysis**

# Philosophy of EDA

**Exploratory Data Analysis (EDA)** is more than just a set of statistical techniques; it's a philosophical approach to understanding data.

**1. Curiosity and Open-Mindedness**

- **Questioning:**
- **Openness:**

**2. Visualization as Storytelling**

- **Visual Communication:** EDA recognizes the power of visualization to communicate complex ideas effectively. By creating informative and engaging visualizations, we can tell compelling stories about the data.
- **Insight Generation:**.

**3. Iterative Exploration**

- **Continuous Learning:** EDA is an iterative process. As we explore the data, we gain insights that can lead to new questions and avenues of exploration.
- **Refinement:** EDA involves refining our understanding of the data through repeated analysis and visualization.

**4. Data-Driven Decision Making**

- **Evidence-Based:** EDA is grounded in the belief that data-driven decisions are more informed and reliable than those based on intuition or assumptions.
- **Actionable Insights:** EDA aims to provide actionable insights that can be used to guide decision-making and problem-solving.

**5. Respect for the Data**

- **Data Integrity:** EDA emphasizes the importance of data quality and integrity. It involves cleaning and preparing the data to ensure that it is accurate and reliable.
- **Ethical Considerations:** EDA also considers ethical implications, such as data privacy and bias, when working with data.

**In essence, the philosophy of EDA is about using data to understand the world around us, to ask meaningful questions, and to make informed decisions.**



By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar