FEATURE ENGINEERING-2

ANOVA

> 2 categories

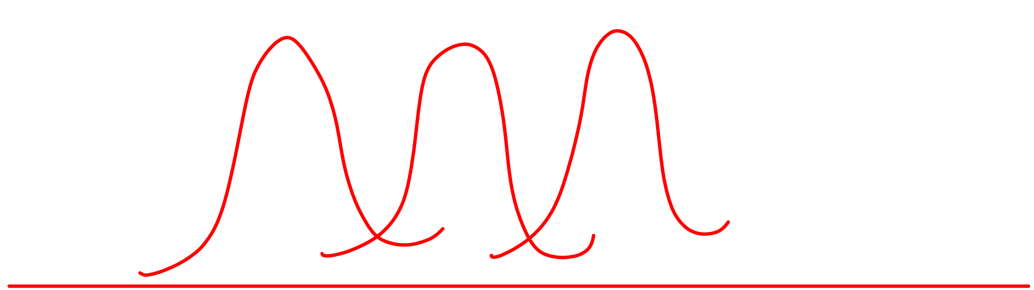
1. Gaussian → QQ Plot / Shapiro / KS Test

   Kolmogorov Smirnoff Test.

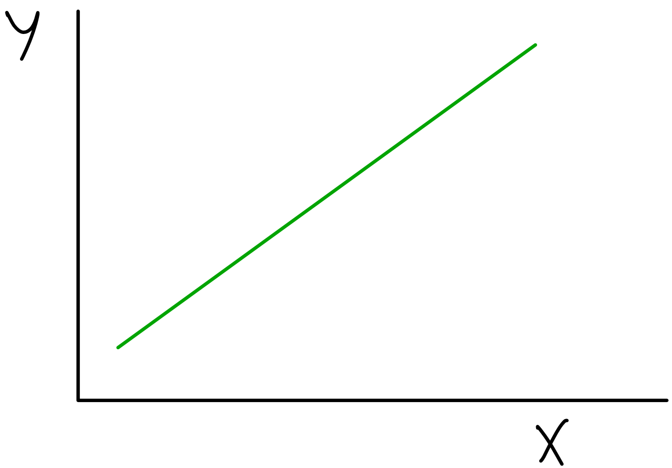
2. Independent

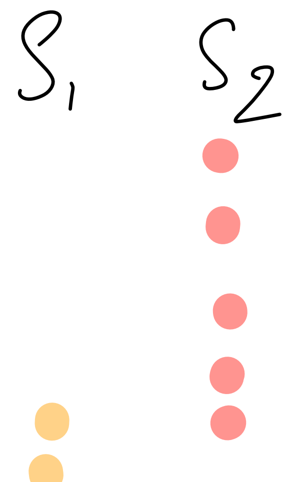
3. Equal Variances → Levene.

KRUSKAL'S TEST



$$f = \frac{MSB}{MSW}$$



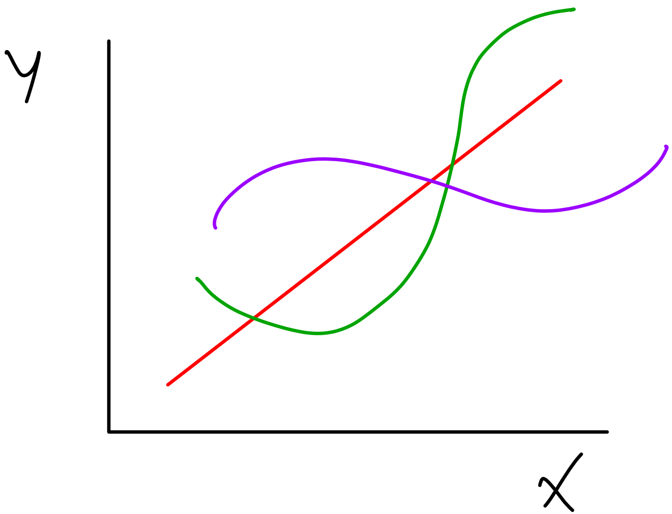


Any distribution $\longleftrightarrow$ Gaussian distribⁿ

% percentiles

% percentile

$1^{st}\ S_1$	$1^{st}\ S_2$
$2^{nd}\ S_1$	$2^{nd}\ S_2$
$3^{rd}\ S_1$	$3^{r}\ S_3$
$\vdots$	$\vdots$
$100^{th}\ S_1$	$100^{th}\ S_2$

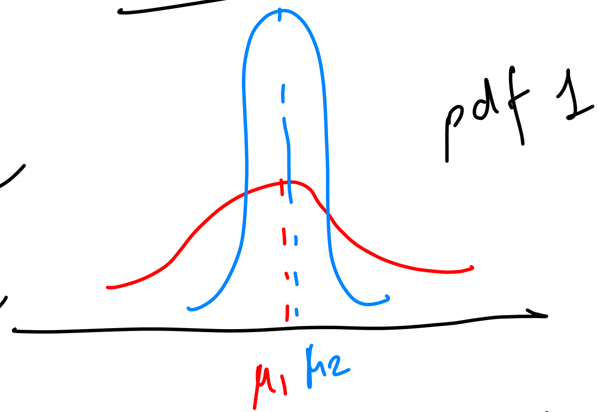


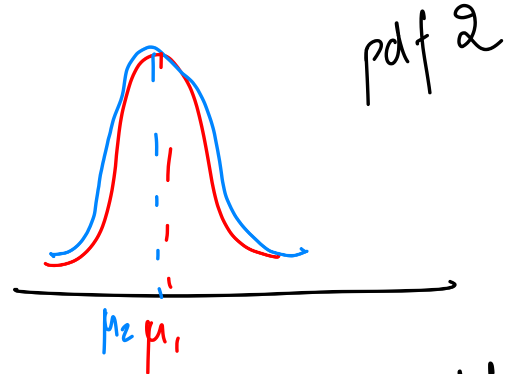
KS Test

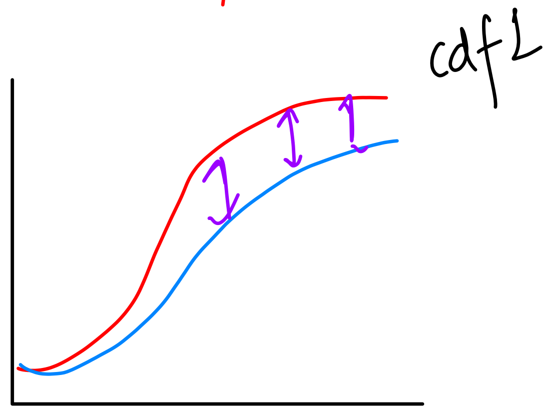
ttest $\rightarrow$ μ_1 μ_2

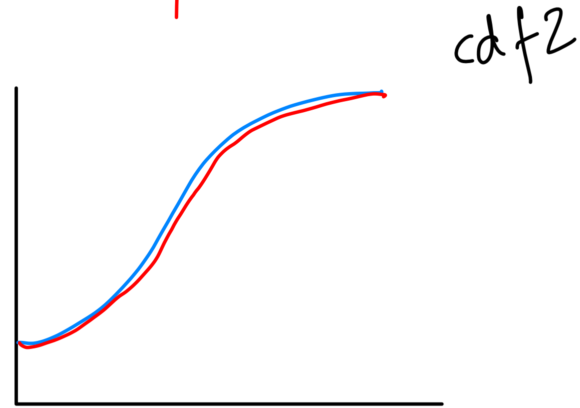
$H_a = \mu_1 = \mu_2$

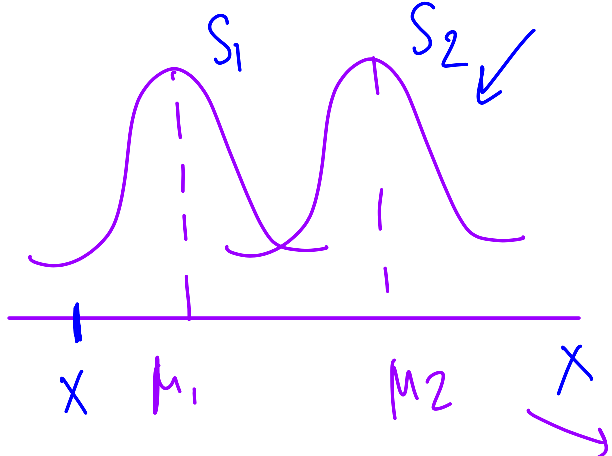
$H_a = \mu_1 \neq \mu_2$





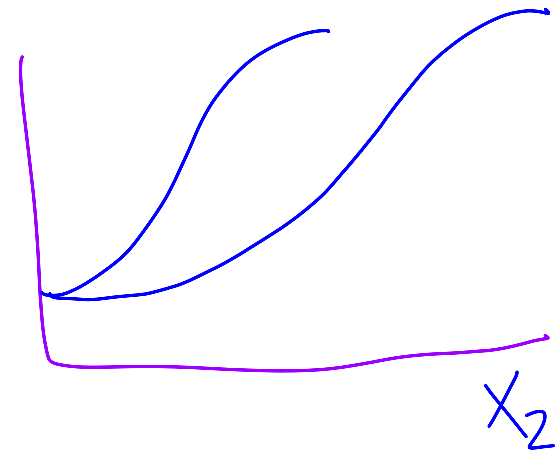






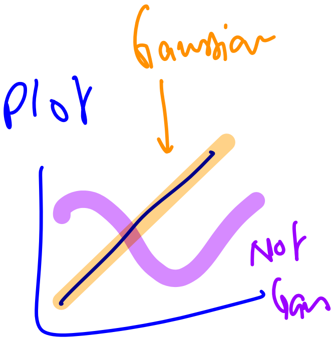
Hyp

H_a

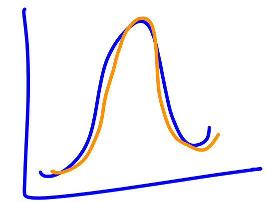


Random distribution

↳ Gaussian → { QQ Plot



{ KS Test

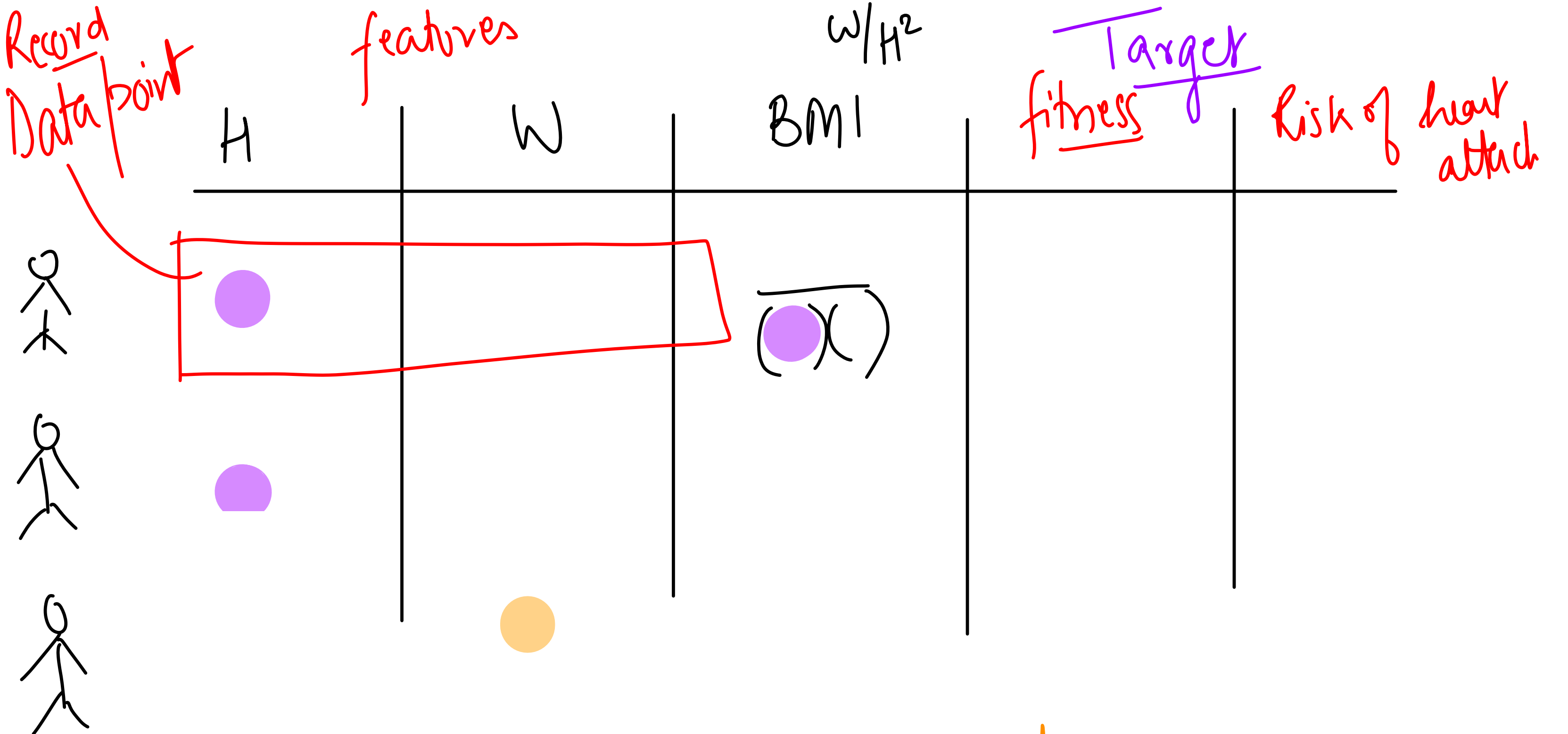


{ Shapiro pvalue

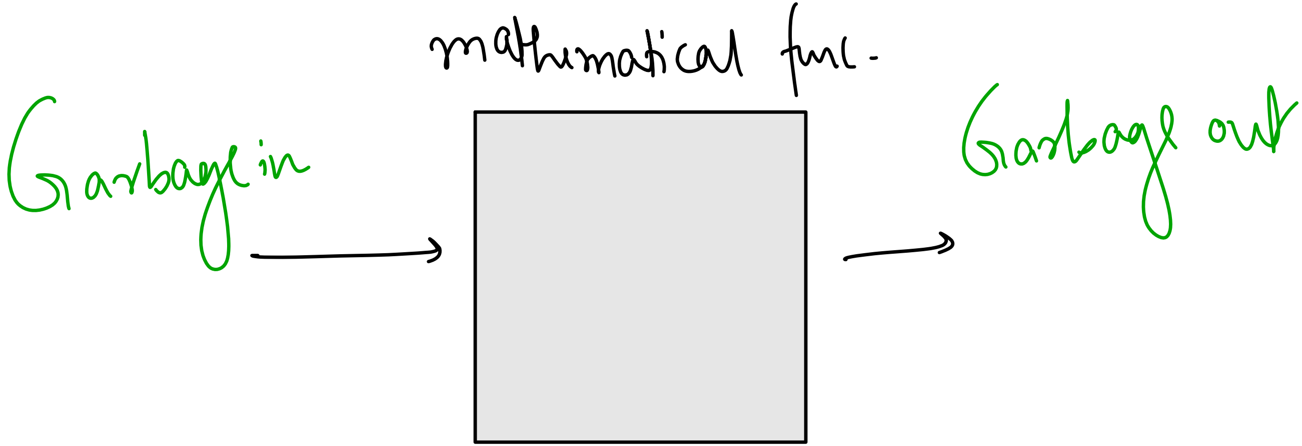
H_0: Gaussian

H_a: Not Gaussian.

Levene:



Feature Importance → Feature Selection



prob

Credit History

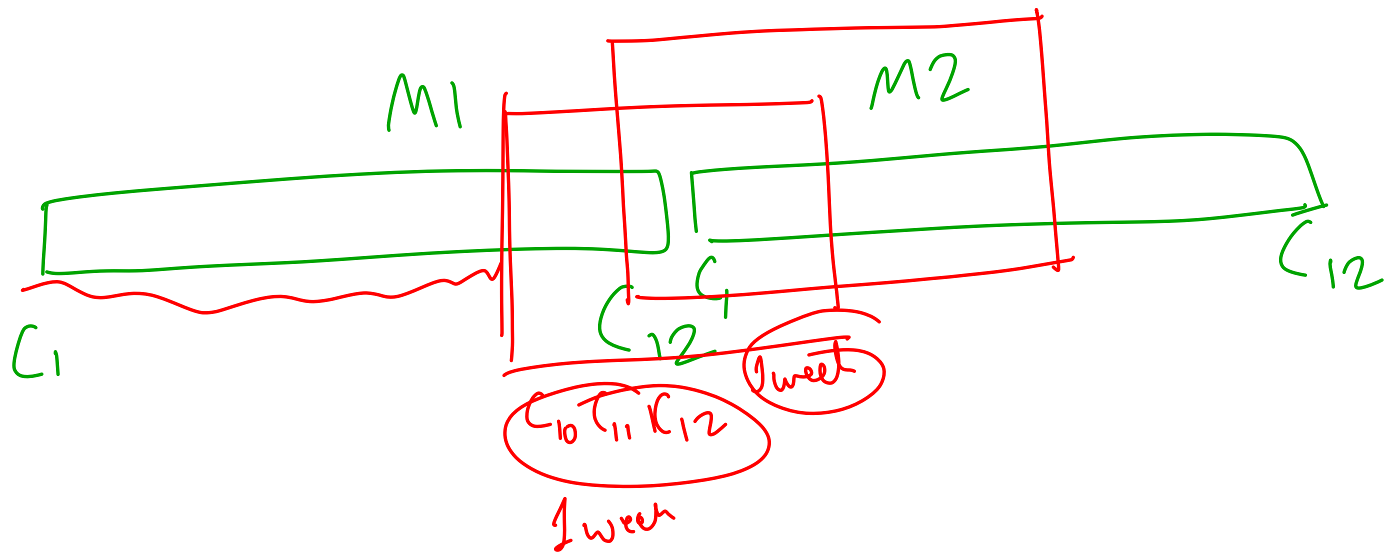
Yes	1
No	0
Yes	1
No	0
Yes	1
Yes	1
Yes	1

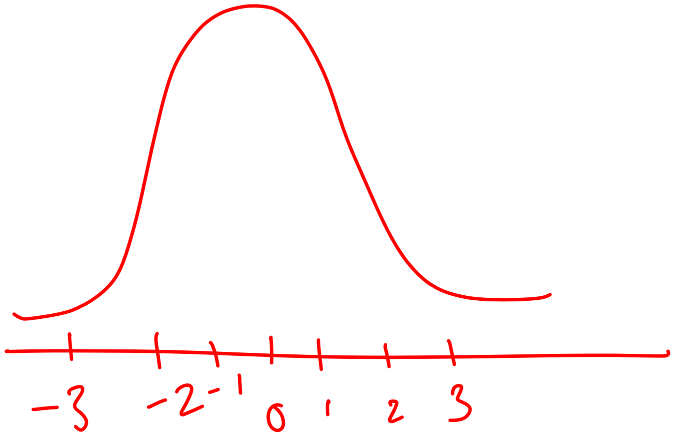
$$\frac{1+0+1+0+1+1+1}{7}$$

$\frac{5}{7}$ % $\times 100$

0.84 % → Dataset

Credit History


M1
M2
C1
C12
C1
C12
C10 C11 C12
1 week
1 week



-3.1, -3, -2.9, -2.8 . . . 0 . . .

3

1st

1st

1st, ← -3,

* Z test → μ, σ ✓ (Numerical v/s Categories)

* T test (ttest_1samp, ttest_rel, ttest_ind, ttest_ind_fromstats)
  σ ✗ → $n < 30$, (Numerical v/s Categories)

* KS Test → $\mu_1 \approx \mu_2$, distributions are diff.

* χ^2 → Goodness of fit (CATEGORICAL)
  → Test of Independance (CAT v/s CAT)

* ANOVA → More than 2 grps (Numerical)
  → Normality
  - QQ Plot
  - Shapiro
  - KS Test
  → Variance
  - Levene

* KRUSKAL → when assump^n of ANOVA FAIL

* CORRELATION → (NUMERICAL v/s NUMERICAL)

SCALER

HYPOTHESIS TESTING CHEAT SHEET

1 CONCEPTS

Central Limit Theorem (CLT)

The **distribution of sample means is Gaussian**, no matter what the shape of the original distribution is.

Assumptions: population mean and standard deviation should be finite and sample size >=30

Hypothesis Testing

- A method of statistical inference to decide whether the **data** at hand sufficiently **support** a **particular hypothesis**.
- A test statistic **directs** us to **either reject or not reject the null hypothesis.**

Null Hypothesis (H_0) represents the assumption that is made about the data sample whereas,
Alternative Hypothesis (H_a) represents a counterpoint.

P-value

Probability of observing the Test statistic as extreme or more than $T_{observed}$ considering the null hypothesis as true.

If **p-value < significance level; reject the null hypothesis**, else fail to reject the null hypothesis.

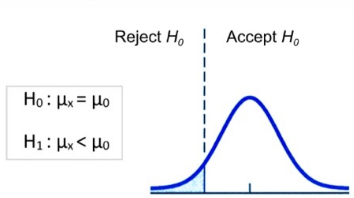
Critical Value

A **cut-off value** used to mark the **start of a region** where the **test statistic** is **unlikely** to fall in.

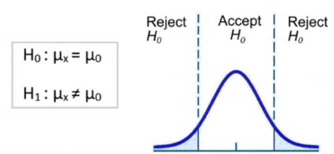
2 TYPES OF HYPOTHESIS TESTING

Type I error (α) - **Reject** a null hypothesis that is true.
Type II error (β) - **Not reject** a null hypothesis that is false.

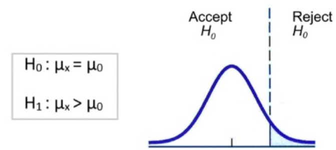
One Tailed - Left



Two Tailed



One Tailed - Right



Framework for Hypothesis Testing

1. Define the experiment and a sensible test statistic variable.
2. Define the null hypothesis and alternate hypothesis.
3. Decide a test statistic and a corresponding distribution.
4. Determine whether the test should be left-tailed, right-tailed, or two-tailed.
5. Determine the p-value.
6. Choose a significance level.
7. Accept or reject the null hypothesis by comparing the obtained p-value with the chosen significance level.

3 TESTS

One Tailed - Left

- Used to determine whether the population mean is significantly different from an assumed value.
- It uses Standard normal distribution as the baseline.

Assumptions:
- either the **standard deviation of the population** should be **known** or,
- we should estimate them well when the **sample size** is **not too small (n>30).**

$$Test\ statistic = Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Two sample Z-test

Used to **compare** the **means** of two populations.

Assumptions:
- either the **standard deviation (σ_1, σ_2)** of the populations should be **known**
- we should estimate them when the **sample sizes** are **not too small (n_1 , $n_2 \geq 30$).**

$$Test\ statistic = t = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

One sample t-test

The test statistic follows a **t - distribution** it is used when:
- the **sample size** is **too small (n < 30)** and/or,
- the **population standard deviation (σ)** is **unknown.**

$$Test\ statistic = z = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$$

Two sample t-test

Used when
- the **sample sizes** are **too small (n1 , n2 < 30)** and/or,
- the **population standard deviations (σ1, σ 2)** are **unknown**.

$$Test\ statistic = t = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

ANOVA (Analysis of Variance)

- Used to determine if there is a **statistically significant difference** between two or more **categorical groups** by testing for **differences of means** using variance.
- The test statistic **f** follows the **f** distribution represented by two parameters **(k-1) and (n-k). k = no. of groups, n = total sample size.**

$$Test\ statistic = f = \frac{MSB}{MSW}$$

where,
MSB = mean of the squared distances between the groups and **MSW** = the mean of the squared distances within the groups.

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2}{k-1} \qquad MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_ij-\bar{X}_i)^2}{n-k}$$

Assumptions:
- the variance of each group should be the same or close to each other.
- the total n observations should be independent of each other.

KS (Kolmogorov - Smirnov) test

- A **non - parametric test** used for determining whether the **distributions** of two samples are the **same or not**
- The test statistic T_{ks} follows a distribution called the kolmogorov distribution
- T_{KS} = the maximum absolute value of the difference in the CDFs of the two samples X and Y

4 CORRELATION

Degree of the mutual relationship between two variables

Pearson correlation coefficient(PCC)

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}$$

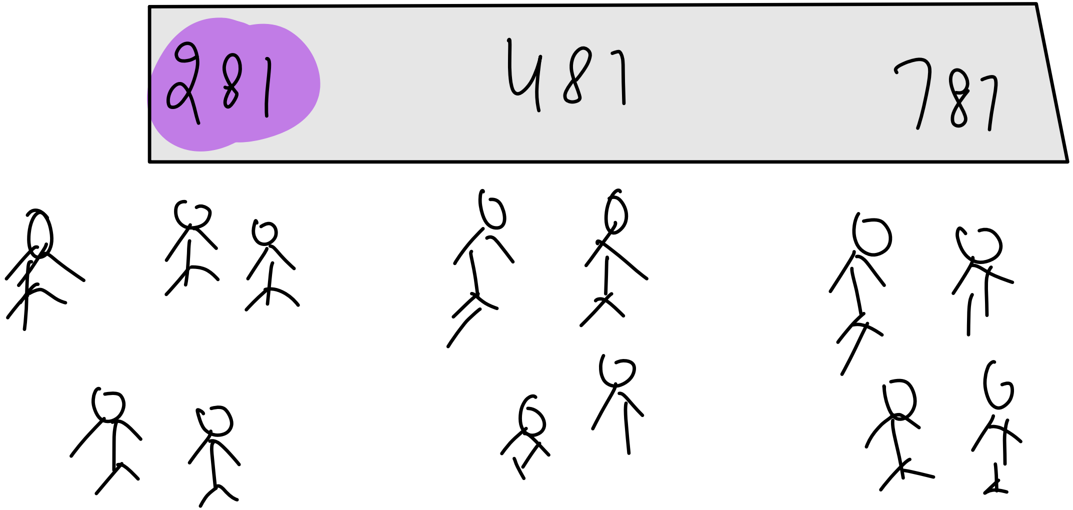
Limitation of PCC is that it only **captures** the **linear relationship** between the variables. It fails to capture the non-linear patterns.

Spearman Rank Correlation Coefficient

A statistical measure of the strength of a **monotonic relationship** between paired data. it captures the monotonicity of the variables rather than the linearity.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where,
d =difference between the two ranks of each observation and,
n = number of observations


281
487
787