Applied Data Science
Trainer: Prashant Nair

Dealing with categorical data

We need a technique that can represent the categorical data in numerical form considering,

ⓐ Mathematical nature of the data is not disturbed.

ⓑ The domain nature and pattern of the data remains unchanged.

OHE (One Hot Encoding) ⟶ Dummy Variables.

city
BOM
DEL
CHN
BOM
CHN
CHN

① Extract the unique values of the column

['BOM', 'CHN', 'DEL']

② Sort the above list in ascending order.

['BOM', 'CHN', 'DEL']

③ Create DUMMY VARIABLES based on above list and show which variable is active for the given record using binary '1'.

city	BOM	CHN	DEL
BOM	1	0	0
DEL	0	0	1
CHN	0	1	0
BOM	1	0	0
CHN	0	1	0
CHN	0	1	0

Hypothesis Testing

Pre-requisite

① Significance Level [SL] / α (alpha) values

→ % error value the project can tolerate

→ usually SL values is determined in two ways:

ⓐ Assumed values: 0.05 , 0.01, 0.1
5% 1% 10%

ⓑ Identify the nearby values of error using cross-validation method.
(practical SL)

Samples
↓
Statistics
(Approximation)
[There can be errors or inaccuracies in the result or prediction]

- How much % accuracy must be maintained by the process?
- How much % errors in prediction or in analysis report, the project can tolerate?

eg: SL = 0.2
out of 10 predictions done by the data product, the business shall allow 2 incorrect predictions

② Confidence Level (CL): (how much minimum accuracy expected from the model, report...).

$CL = 1 - SL.$

determines the minimum evaluation metric threshold to be achieved by the trained data product.
(trained model, analysis report, strategy report ...)

③ p-value: (probability value)

Calculated SL value derived from statistical testing formulae/tools

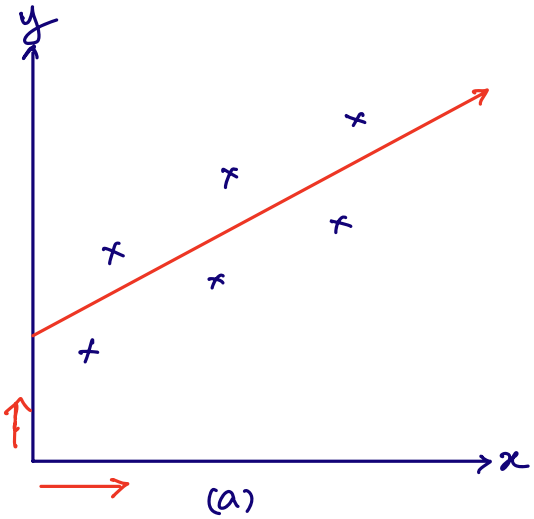
Statistical Testing

test to identify if the given statistical feature is present in my dataset / column.

- Test for Normalization
  - check if the given column has normal distribution
- Test for Correlation
  - check if correlation exists between two variables / columns.
- Test for Feature elimination
  - removing irrelevant / less significant independent variables.

Correlation

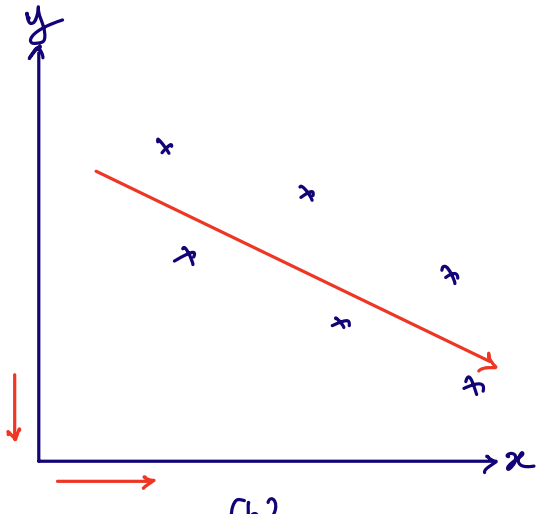
Correlation is all about identifying understanding whether the given two variables / columns in the dataset has a linear relationship.



(a)

Positive Correlation

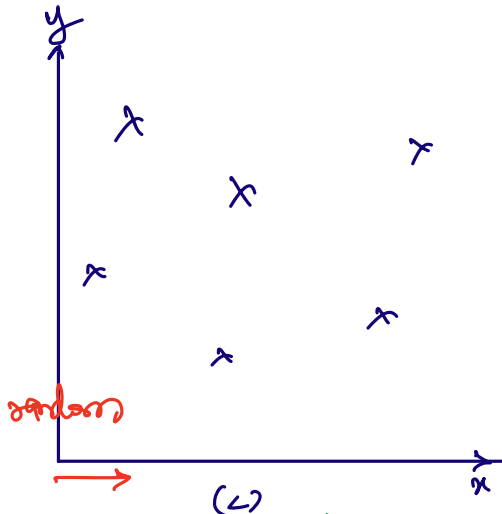
with increase in x, y increases.



(b)

Negative Correlation

with the increase in x, y decreases.



(c)

No correlation

with increase in x, y is random

Goal:
determine whether the evidence provided by the sample data is STRONG enough to support the said assumption.

(assumption) Hypothesis Testing

Goal: Test if col1 and col2 have linear relationship

H_0 → col1 and col2 has NO LINEAR RELATIONSHIP

H_a → col1 and col2 has LINEAR RELATIONSHIP.

Who win? via statistical testing

Null Hypothesis (H_0)

Alternate Hypothesis (H_1) (H_a)

Goal: Check if the given liquid is an ACID or BASE

Note: When it comes to formulating the Hypothesis, always remember:

ⓐ Hypothesis must be formed as a BINARY OUTCOME.

ⓑ Only ONE STATISTICAL feature must be tested at a time.

Hypothesis 1 → Check if the given liquid is ACID.

H_0 → given liquid is NOT an ACID.

H_a → given liquid is an ACID.

Hypothesis 2 → Check if the given liquid is BASE.

H_0 → given liquid is NOT a BASE.

H_a → given liquid is a BASE.

Statistical Testing

- Correlation test
- Normality Test
- Non-parametric test
- Parametric test
- Chi-square test

Applied on Quantitative Variables (Numerical)

Qualitative Data (categorical)

1. Pearson's Correlation Test
2. Spearman Test
3. Kendall Tau Test

1. Shapiro Test
2. Anderson-Darling Test
3. Normal Test

1. Wilcoxon Test
2. Mann-Whitney U test
3. Friedman test

1. student t-test
2. paired student t-test
3. ANOVA (Oneway)

chi-square test