Date: / /
Page:

Different ways to handle Errors -

- When data Scientist work with raw data, they can't safely assume that data is error-free.

- Mission should be, find & tackle errors in most efficient way.

① Accept the Errors:
- We enter "Navi Mumbai" instead of "Navi Mum.", if we accept the error, it will affect data science technique & algorithm.
- And whatever operations we perform, that may be differ from actual result, because value that we provide, not the same.

② Reject the error :
ⓐ Listwise -
- In this case, missing the value of the cells then we delete the entire row that containing missing variable/value.

ⓑ Pairwise -
- In this case, only missing values/variables are ignored.

③ Correct the Error :
- Spelling mistakes in customer names, Addresses & locations are a common errors, with are corrected systematically.

Date : __/__/____
Page: _____

(4) Create a default value -

- NaN (Not a Number) is a default missing value or dumming data, that can be easily detected and manipulated by using function in pandas.

Note on Missing value treatment -

⇨ Why data has missing value -

① Data field rename during upgradation,
② During data migration from old system to new system.
③ Data field not recorded due to data unavailability.
④ Incorrect data provided by the subject-matter expert.

⇨ Practical actions for missing data - (By using Pandas)

① Drop the columns, where all elements are missing -
df1 = df.dropna(axis=1, how="all")

② Drop the column, where any of elements is missing -
df1 = df.dropna(axis=1, how="any")

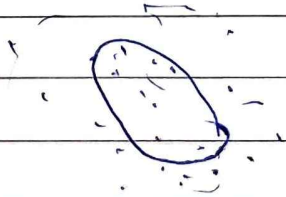
③ Keep only rows, that contains max. two missing values -
df1 = df.dropna(thresh=2)

axis="0" = row	how="all" = all missing values
axis='1' = Column	how='any' = any missing values

⇨ Actions for missing values - (By using Techniques).

Outlier Detection & treatment :

① Elliptic Envelope :
Scikit-learn package.



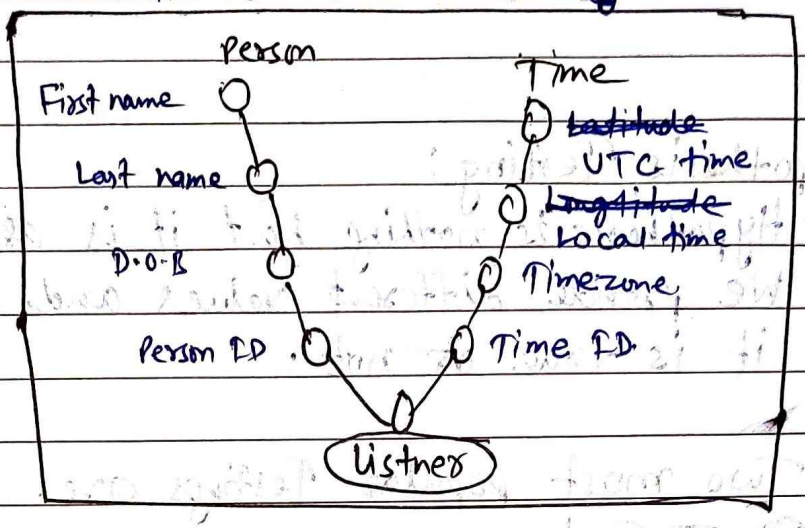
② Isolation forest -
⇒ max, min → average,
⇒ select the data nearer to the average.

③ Novelty Detection -
→ 'Sklearn.svm.OneClassSVM tool used
→ Observe the data and retrieve the original data by using this tool.

Various types of Sun Model - (4 types)

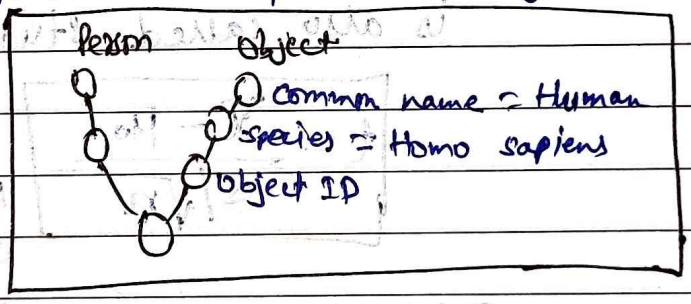
① Person-to-time :-

- It explains the relationship b/w Person & Time in the data vault.
- This model shows all characteristics of these two data category.
- 2-Dimentional data created.



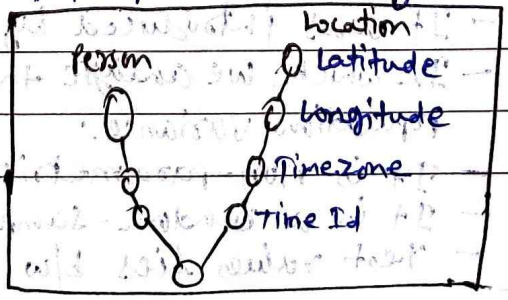
② Person-to-Object :-

- relationship b/w person & Object.
- Here object means person belongs to which category.



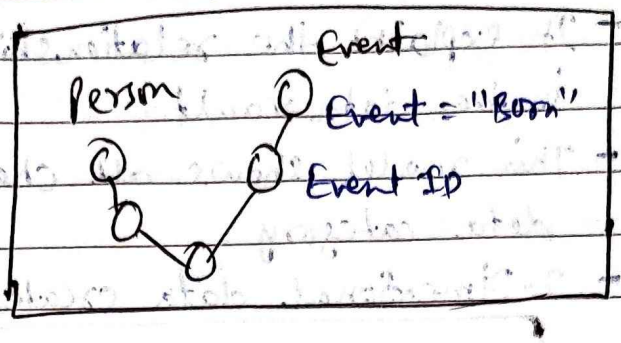
③ Person-to-Location :-

- It show ~~which~~ person belongs to which location.



Date : / /
Page:

④ Person-to-Event :
- Explain the relationship b/w person & Event.
- Here Event means what action occurse (Born, Death).



Hypothesis Testing :
- Hypothesis is nothing but it is assumption.
- We provide different values and, check either it is true or not.

⇨ Two most popular Testings are.

① T-Test -
- T-test is small sample test.
- It was developed by William Gosset in 1908.
- T-test was the name of a pen. So, it is also called student t-test.

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

$\bar{x}$ = mean of sample
μ_0 = mean of total population
S = Sample standard deviation
n = Total population

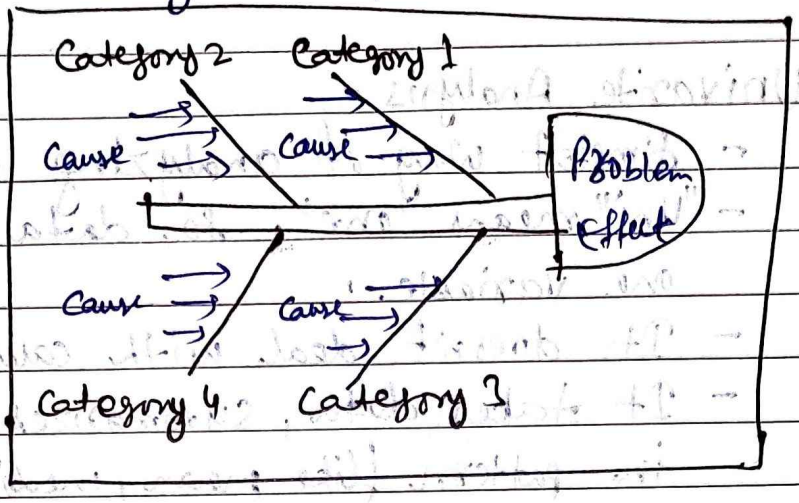
- T-test apply when $n < 30$, & st. dev = unknown

② Chi-Square Test -
- It was introduced by Karl-Pearson.
- In which, we analyse the sample & test the population variance.
- It is non-parametric test
- It is a random-sampling method
- Test values lies b/w 0 to 1.

Date : __/__/____
Page: _____

Fishbone Diagram :

- It is a useful tool that helps managers to find/track the reasion imperfactions, defects or failures.
- It showing the arrengemat of data simplest way in data vault.
- This diagram look like a fish's skeleton, that's why it is called fishbone diagram.



Hypothesis Testing -

③ Cross-Validation Test -
- It is used to acheive the goal by value prediction.
- It is generalized for all to put the value and got the output (statistical analysis).

④ Univarite Analysis -
- Simplest way of analyzing data.
- "Uni" means "one", so, data has only one variable.
- It doesn't deal with causes & relationships.
- It takes data, summarizes that and get in pattern (like mean, median, mode).