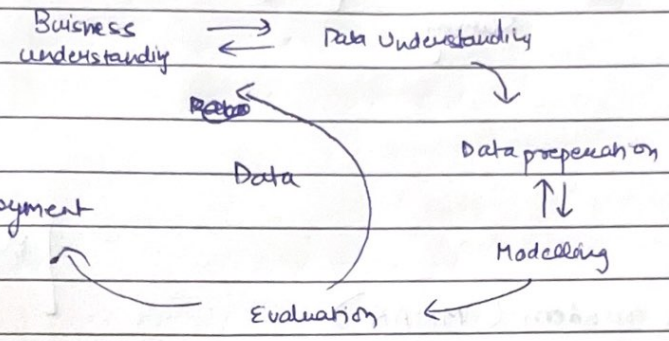
CRISP DM Framework

↳ Cross industry standard for data mining.

DATE: __/__/____
PAGE: _____



Data preparation

ML
↳ Learn the pattern
↓
Null values

- ↳ Missing value treatment
- ↳ Duplicates values (Remove)
- ↳ Outliers (Boxplot → Used to detect outlier, Detection & treatment)
- ↳ Feature engineering
  - Select relevant features
  - Create new features from existing one
  - Modify the existing features
- ↳ f1 f2 ↳ (we'll check the correlation, the most correlated data will be traced as relevant feature) ② Distance | time | speed
- ↳ Data Encoding

* Null | Missing value treatment

① Missing completely at random (MCAR) — Independent of the data that we have or the data that is missing

Students	Marks
A	39
B	39
C	-
D	30
e	31
f	-

→ No dependence of C on the people who appeared for exams & who didn't appeared.

Good Write

② Missing at random (MAR)

↳ some dependency on observed data.

DATE: __/__/____
PAGE: _____

Name	Age
V	21
A	25
S	

Eg: Girls not giving contact no. in the open survey.

[df.isnull().sum()]

③ Missing not at random (MNAR)

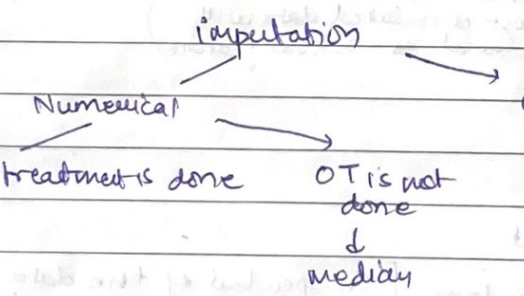
↳ Some dependency on observed data

* How to treat null values

↳ Never do missing value treatment without connecting to business team.

* if missing value < 1% and the data is large, you can drop the missing value.

* if data is small and you dont want to drop the missing value (be it small or large) you will do imputation.



* impute missing value with a constant (Replace with zero 0) or the number that is not possible for that particular column)

Student marks

50
60
70
80
—
90

Here you cannot use mean & median because if student appears for the exam

* Create a new column with a flag if the column is missing and input some const. value in the missing column.

Name	Marks	Name of observed
A	28	0
B	-	1
C	30	0
D	-	0

Good Write