	# Basic # Data Ucaning # Data Preparation. Date / //
Q	Some important concepts to Rect in mind
	while doing Data Cleaning and
e ^t	Preparation man has made in the formation of the second
	an derived the second to the second
1	Handling Missing Data
1-	and are burger house to more house, and is is
f3	why is who entrous how a fet grights is
	. To maintain data integrity
	. To audid Bias and distortion 1 11111
4.34	· To maintain data completences
	the publications.
	How?
paige	· completely filter out rows or walles using
	: abandaparati odmila sou materiolo
ruta S	Imputation i.e., filling missing values
	either by mean median or by anyother
AND THE STATE OF T	value.
	either i domain specific imputation
The state of the s	2. Time sovies imputation or
The property of the contract o	3. otatistical imputation.
2	Data Transformation.
	why?
	· grregular scales of data and different
	range makes analysis difficult.
	· Duplicacy of data which will be af no use
	. To handle data skewness
Sec.	. For Further processing and analysis DT is

0 0 20	# Basic # Data cleaning
	# Data huchanation DOMS Page No.
	Date / /
	Viscontinuo han intertheritation of
•	For make data in torm which is can be used
	for twithou computation URE handling categorical
i.	data. Consularios de sons torrestores
•	Bupsa La compress rated pure con the porting of
H 2.0	To make data in form wwitable for other
	libraries ure mumby.
	How? Then head the entrapation
y god	1. Remowing Duplicates:
	Pandas has inbuilt tous to remoue duplicates
	COST INDICATE TO DELLE TO
	2. Transforming Data using Function mapping
	or didion any mapping, and bard
The latest	ntim dom of anothrut so prenoitive priau
	data by using map () tundion and
	moni pulation ocioting columns or creating
1 1.0	Enouve one la distribution de mom
A TEN	exception problem statement and productive
	3. Replacing values: tos chou and
	using replace fundion to replace
	existing halves by null or any other
3	Malue as vie auined.
	TOPICA NOTCOLETA LA
	4. using 2 we can also rename indocand
	manipulate them as per requirement
	whout vicating new data structure.
	we can also use Henome() feul in
	bandas.

Basic # Data aconing Doms Date /

Date / /

Б.	Discretization and Binning:
60000	sometimes it is hourd to a natyse
boring of the	continuous data uso it is discretifed into
	"bins" for analysis.
HOURS TO	you can consult in the bridge of
stotto	continuous ada -> coregorical data
	by weating bins: (use pandas)
6.	Detecting and Filturing outliers:
-	· you can use matplot lib to detet outliers
aster ald all	h o d oo o
~	to filter out the outliers
enia	· we can use vovous statistical tests
	and methods for detecting outliers
CHUS C	. use imputation or addition for filtering
	Lacor flag themen contained alob the
raling	For abedion there are navious
	motods available but used acrording
	to problem. Statement and ucquirement
	1. Vasualization, state on tooley 18
3	oldor 201 Zaliscore moleur prion
	uno 30 UIQR mothodo patrane
	4. Mahalan bis distance
	5. Isolation forest
- John	6, Local outive factor (LOF)
	SINGHALON PAGES OF CROSSES STORY DOOR
	white we called the position of the transmit

	# Basic # Data Purparation. Date / /
7.	Random sampling & ucorduling for audicing
	bias and noise learning.
₩ 8.	concerting categorical variables at data
	into column form. where each column
	represents a categorical voveiable.
	This is uncalled for modelling and consider
	as a important which.
NOTE:	Fot string manifulation we caus use regix
	by importing re Unrowy. i.r. regular ochrenion.
3.	More on Categorical Data. Categorical data
1 200	into numurical
	why? representation.
•	Algorithm compatiblity:
	many Me algo, are compatible into
	numurical inputs.
	Preventing misintul Nuctation of data and handling
	multiple categories.
all second	
	How?
	There are various techniques to do so
	-using dummy variables or indoving
	-one hot encoding
	-label encoding
	- target encoding and various other.
Section 1	