

PICT, PUNE	
	Assignment No.2
111	of the of the Manual Son Sugar and the
Made	Title Data Wrangling 2
-\	
Well	problem statement and and the statement
They times	create an 'Academic performance dataset of students
	and perform following operations using python.
	1) scan all variables for missing values inconsistiness
M. Carlot	If there are missing values and for inconsistencies
	use any of the suitable techniques to deal with it.
1	2) scan all numeric variables for outliers, If there are
	outliers use any of the suitable technique to deal with
	3) Apply the transformations on atleast one of the
- 4	variables. The purpose of this transformation should
7	be one of the following reason. to change the scale
	For better understanding of the variable, to convert
,	a nonlinear relation into linear one, or to decrease
	the skewness and convert the distribution into a
1.	normal distribution
1.	The state of the s
1	Learning Objectives
Jan party	1) To understand and implement various data processing
Length 9	and outliers handling techniques.
- hiri	2) Understand and implement various data transformation
	techniques to change scalar or convert distribution
	to a normal distribution.
	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
10 To	and the state of t
1	Will his medical the formation of the projection of the
	brief of the state



PICT, PUNE	
	Learning outcomes
:	
	By performing this assignment, we will be able to  1) Various data preprocessing and outliers handling
	techniques.
	3) various data transformation lechniques to change
	Scale or convert the distribution to normal distribution.
4	allow was a some a somewhere when it is the street was a some
7. A 1 at	SOFtware Requirements
	1) python 8.10 (with pandas, numpy, mat plotlib, seaborn
- 11-1	Scipy).
x - 1.1. 1.	2) jupyler Nolebook
P=	and the state of t
	Hardware Requirement
1	1) Inles is 10th generation
- 10	2) 8 GB RAM
	in horas colonial and and and
	Theory
-	Data Wrongling
	- Data Wrangling as the process of cleaning and unifying
	messy and complex data sets for easy access and
	andlysis.
	- This process hypically includes manually converting and
:	Trupping gord 5000 ODE tout form jobs markes formate
-11-74 -1-74	TO GITOW FUT MORE CONVENIENT CONCUMPTION and
	organisation of the data
	package Used
	1) pandas
	- It is a software library in python for data manipulation
	and analysis.



ı

PICT, PUNE	. 546-2 (4
D James	2) Numpy of the mount of the second of the
446 1	It is a library in python used to add support for large
	collection multidimensional arrays and matrices along
- 30	with large collection of high level mathematical
	Function to operate on the arrays
	A Comment & Comment of the Comment o
	3) scipy-stats
	It is a library used to perform advanced statistical
	Operations on dataset
the second	I don all mount for the distriction as a second
	4) Matphotlib
	- It is a library Used for plotting simple graphs and also
	for data visualization to detect outliers, data transformation
mant I don	Visualization etc.
- Surum	The course of the charge of above of the book of
10010	5) seaboin ve min forma and and
	- It is a library used to plot detailed and visually good
	looking graphs for more advanced data visualization to
	olefect outliers and check it data transformation happened
	correctly shire so read points (c)
· · · · · · · · · · · · · · · · · · ·	the property of the second sec
	pata togns formations
	skewness is defined as the measure of asymmetry of the
	probability distribution of the real valued random variable
	( realure attribules)
	the skewness outside the range of -1 to 1 are highly
	skewed and those in the range of -0.5 to 0.5 are
	Symmetrical or normal, remaining are moderally stement
Asset in the second	Symmetrical or normal, remaining are moderally stewed.



	- to convert the distribution to normal distribution means
	we need to change the skewness and take it in the
	range -0.5 to .0.5
1.	- some of the data bound formation techniques are.
1 II	n) logarithmic tounsformation
	2) Reciprocal transformation
	3) Square 100+ transformution
	9) Exponential boansformation
	5) Box-Cox towns formation.
	- The data transformation rephriques can only be applied
-	on pasitive values.
	in the second se
San Land	petect and eliminate outliers.
	- outliers are the data objects that deviate significantly from
	rest of the data objects and behaves in different manner.
	- They are delected on the basis of the kind of dataset
	attribules use are dealing with.
	- some of the ways to detect outliers are.
-	1) Using Boxplot or scatterplot ( Data Visualization)
	2) Using Ther- quartile range,
	3) Using a score
	4) Using standard deviation
<u> </u>	were now a superior and the first of the second will
	Functions used
	1) df. mean()
1.5	-religions the mean of each numeric cittibule
	and the second through the second
_ = (	2) df. max()
	- For calculating the maximum numeric value



PICT, PUNE	
	3) df. quantile
yA .	- returns the values at the given quantile over requested axi
	4) df.skew
	- returns the skewness of the values of numeric attributes data frome df.
. Inmo!	en de de la
AL THE	5) Seasom. Kdept of por)
	- plots a KDE plot for column 1001
1 (1	esternaturi Man and and and and as a part of
	e) nplog (df[001]) who have an all the second
	returns the data frame array object of logarithmic values.
	7) sapy. stats. boxcsx[df[001])
	- returns the towns formed values of each numeric record of
	column '001' of data forme 'df' so that the transformed
,	values from a normal distribution after box asx transforma
	is applied.
	observations
	1) out of the two plots to observe the outliers, boxplot
	is preferred for univariate while scatterplot is preferred
	bivariate / multivariate outliers.
	2) out of all the data transformation techniques, logarithmize
	and reciprocal tours formation gives most in accurate
	results.
,	resulks.
,	TESOT AS



PICT, PUNE	
	ATT COMPANY
we law and	3) the data transformations techniques can only be
	applied to positive data points.
	Commence of the second
	conclusion
	Henre, after performing this assignment, we learnt
	various processes and rechniques to handle outliers and
	apply data toansformations to convert distribution to
	normal distribution, we successfully implemented the
	techniques on provided dataset.
1	in the said of the said of the training of the said of
	Committee Transcont of the Committee of
•	Commerce describe in the contract of the contract of
- Line	the state of the second of the second of
	the and added the test in the live one on west - 2 miles
	And the state of t
	THE REPORT OF THE PARTY OF THE
	The state of the s
Andrew American	in the state of the service of the second of
	hilly all I made in the contract
The second 1st	and the first of the first of the state of the first of
The sale of the	the harm and the comment of the gray terms
A CONTRACT OF THE STATE OF THE	g Hand
w w	