Rupesh Dharme 31124





Assignment az

DOP 10/01/2022

DOS 20/01/2022

Title: Data wrangling II

Problem statement:

create academic performance dataset of students and perform following operations using python

inconsistencies, if there are missing values / inconsistencies, if there are missing values / inconsistence use any of suitable techniques.

2. Scan all neumeric variables for outliers

1f there are, use any technique to deal with them.

3. Apply data transformation on at least

one of variables with any of following purpose:

to change scale for better understanding

to convert non-linear to linear relation or

to decrease skewness and convert to

normal distrubution.

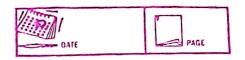
learning Objectives:

To learn and understand data wrangling.
To deal with missing values/inconsistencie/
To deal with outliers in dataset
To learn and perform data transformations

learning outcomes:

student will be able to

Rupesh	Dharme
31124	



- perform handling of outliers - perform data transformations for better understanding.

H/W S/W req:
Windows 10, 64 bits, 8 GB RAM, SR GB SSD
INTEL is, vscade, jupyter notebook.

Theory:

An outlier is an observation in given dataset that lies for from rest of the observations. It is It may occur due to variability in data / experimen

or human error. They may indicate heavy skewness - mean is accurate measure to present data when we do not have outliers.

- median is used when outlies are present. - mode is only measure of central tendency that is used with outliers when more than half of data is same.

some techniques to detect outliers

- 1. Boxplot
- 2. Z-score
- 3. Mer ayartle Range

Some techniques to trim outliers:

" Trimming 2. Quantile based flooring or capping 3 mean/median imputation.





As mean is highly influenced by outliers, advised to replace outliers with median value

Normalizati is a technique with the goal to change the values of numeric columns to a common scale without distorting differences in the ranges of values or losing information

Z-score is a variation of scaling that represents the numbers of standard deviations away from mean Ensures your feature distribution has mean =0 and std der = 1. Useful when there are few outliers butnot so extreme that you need clipping.

Another normalization method is the Min-Max scaling. All features are transformed into the range [0,1] meaning minimum corresponds to o and maximum to 1.

- i) The dataset has a shape of (1000,8)
 ii) There are null values in 'math score', 'reading score 'niriting score'
- iii) Moth score column is given in String data type so we type cast it into Int 64.
- iv) By plotting box plot, we come to know that there are outliers

IQR = Q3 -Q1

Upper bound = Q3 + 1.5 * TOR lower bound = Q1-1.5* IRR

Rupesh Dharme 31129





	31129	DATE	PAGE	
	vi) we apply one hot encoding columns to ensure there is I	on co	ategorical	P
	Conclusion: We have successfully implemedataset.	ented	data ana	sisyl
20 ,4		* :		
		4		
		1 1		,