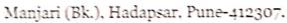
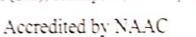


Pune District Education Association's

College Of Engineering







Title: Data Weangling, I Objectives: students should be able to per data wrangling operation using python Weangling: I perform the following Import all the required python libraries 2. Locate an open source data from the web perride a dear description of the data & its source 3 load the dataset into pandas data frame. Data Perpercelling : Check for missing values in it using pandas isnull () describe function to get some initial statistics. Provide variable descriptions, types of variables, etc. check the dimension of the data frame formatting & Data Normalisation: Summa the type of variables by checking





Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC

G. Turn categorical variables into quantitative variables in python. In addition to the codes of off, explain every operation that you do in the above steps of explain everything that you do to import/read/scrape the data set.

Requirement: 1 Basic of python programming.
2 Concept of data preprocessing, Data formatting,
Data Normalization of Data cleaning.

Theory:
Data wrangling in python

Data wrangling is the process of gathering,
collecting of transforming row data into another

Data weangling is the process of gathering, collecting of teansforming row data into another format for better understanding, decision - making, accessing of analysis in less time. Data weangling is also known as Data mugging.

Topoetance of Data Weargling
Data weargling is a very important step.

The below example will explain its importance as:

Books selling website want to show top-selling books of different domains, according to user preference. For example a new user search for motivational books which sell the most or having a





Manjari (Bk.), Hadapsar, Pune-412307.

-	
	high eating, etc. But on their website, there are
-	plenty of Eaw data from different users. Here the
-	concept of data mugging or data wrangling is used
-	As we know Data is not wrangled by system. This
-	process is done by Data scientists
+	Data weangling in python is a crucial topic
-	The Data Science & Data Analysis Pandas Francios
	of pylnon is used for Data wrangling. Pandas in
-	a specifically developed
	The Derroll
-	like data sorting, Data filtration, Data grouping,
	etc J
-	Data Weangling in python deals with the below funtionalities.
	funtionalities.
-	1. <u>Data exploration</u> : In this process, the data is studied, analyzed of understood by visualizing representation of data.
-	studied, analyzed of undetitood by visualizing
_	representation of data.
	2. Dealing with missing values:
-	Most of the dataset having a vast amount
_	COLLUID MILATING VALUE - KNI- XX
	ture of his kind all a
	TECQUENT Value
ï	a sumply by dispoina the issue
	having a NaN value.
_	





- P.T.O.

Manjari (Bk.), Hadapsar, Pune-412307.

	3. Rephasing Data: In this process, data is manipulated
	according to the requirements, where new data can
	be added or pre-existing data can be modified.
	4. Filtering Data: Sometimes dataset are comprised.
	of unwanted rows of columns which are required
	to be removed or filtered.
	5. Other; After dealing with the row dataset as
	per our requirements of then it can be used for
	required purpose like data analyzing, machine
	learning, data visualization, model tearning, etc.
	J'
	Below is an example which simplements the
	above funtionalities on a raw dataset:
ł	· Data exploration, here we assign the data of
	then we visualize the data in a tabular format.
	Pune District Education Association
L	CSV file / Dataset - Eitanic train CSV
	· Required python libraries
	import numpy as np.
	import pandas as pd
	imposet matplotlib. pyplot as plt
_	import seaborn as ans.



Pune District Education Association's





Manjari (Bk.), Hadapsar, Pune-412307.

	: Required syntax
	# Load the Dataset into pandas data frame
	df = pd. read_csv ('titanic_teain.csv') df
	# For showing top result
	df. head ()
	वहजन हिताय. बहुजन सुखाय।
_	# For showing bottom result df. tail ()
	df. tail ()
	# Calculation No. 11 Maria
/	# calculating Null Values. df. isnull(). sum()
	of annual (). Sun ()
	# Calculating Null Values in Age of Cabin' column of ['Age']. isnull (). sum() df ['Cabin']. isnull (). sum()
_	of ['Age']. inul (). sum()
	df ['cabin']. isnull (). sum()
	
	# (ret some initial statistics
	df. describe ()
	# Getting some information about dataset
	df. info ()
	D.T.A
-	P.T.O.





Pune District Education Association's

College Of Engineering



Manjari (Bk.), Hadapsar, Pune-412307.

```
# Finding Data types
of dtypes
# Finding Dimensions of data frame
   of shape
# Making impute function for filling Null Values
   def impute - age (cols):
       Age = cols[0]
       PClass = cots[1]
      if pd. isnull (Age):
         if Pclass ==1:
             return 37
         elif Pclass == 2:
             return 24
         return Age
# Displaying two-dimensional data in geid format where the color intensity represents value.
 sns. heatmap (
sns. heatmap (df. inull (), yticklabels = False, char = False,
      (map = 'vividis')
```





Manjari (Bk.), Hadapsar, Pune-412307.

```
# Applying impute function of droping column
(abin respectively df ['Age', 'Pclass']]. apply (impute_age, axu=1) df.diop ('Cabio', axis = 1, inplace = Teuc)
# df. drop_duplicates ()
# Data type conversion

df ['Age'] = df ['Age'] astype ('int')

df ['Age'] = df ['Age']. round (0). astype ('int')
# converting categorial variables to Quantitative
 variables
  cat = pd. get_dumnies (df, columns = ['sex'])
 # Female = 0 & male = 1
   cat [ 'Sex_female n] District Education Association
    (at [ 'sex_male']
  # df. columns
 Conclusion: Hence we have throughly studied how
  to perform the following operations on python on
  any open source dataset. (eg. data, csv)
 1. Impost the all required libraries
                                                           P.T.O.
```





Manjari (Bk.), Hadapsar, Pune-412307.

	2. Locate an open source data from the web, provide
	clear description of data of its source.
	3. Load the dataset into pandas dataframe
	1. Data properties i check for missing values in
	4. Data perpeocessing: check for missing values in
	description. Type of Variables, etc., check dimension
	of data feame.
	S. Data formatting P data normalization: Summarize
	the type of variables by checking the data type
	of variables in the data set. If variables aren't
	in the correct data type, apply peoper type conversion
	6. Turn categorial variables into Quantitative
	variables in python. In addition to codes of ofp,
	explain every operation that you do in above steps
	of explain everything that you do to import / read/
	scrape the dataset.
	The second secon
01/	Explain Dataframes with suitable examples.
-	Data fearnes are data displayed in a format as a
	table. Data feamer can have different types of data
	inside it.
	Example :- lets have - Data of teaining pulse duration
	Fraining Pulse duration
	1. Stringth 100 60 2. Stamina 150 30
	3 other 120 45
	P.T.O.





Manjari (Bk.), Hadapsar, Pune-412307.

	•
Q2.	What is the limitations of label encoding method?
	limitation of label encoding method:
	label encoding converts the data in machine
	readable from but it arrians a unique number to
	readable form, but it assigns a unique number to
	each class of priority issues in training of data
	sets. A label with a high value may be considered
	to have high priority than the label having lower
	value.
() з.	what is need of data normalization?
	The main objective of database normalization
	is to eliminate redundant data, minimize data
	modification errors of simplify the query process.
•	process process decreased A
Q4.	What are different techniques for handling the
	marrial made and a data /
	Common Technique District Education Association
	1. Mean of Median Imputation
	2. Multivariate Imputation by chained equations
	3. Random forest
(05.	What is meant by Data Dreprocessing?
7	what is meant by Data preprocessing? 1. Checking of missing values using pandas isnull()
	function of massing remains the control of the cont
	2. By using describe () function to get some initial
	statistics.
	P.T.O.





Manjari (Bk.), Hadapsar, Pune-412307.

	· 1000 · 1000
	3. Pervide variable description. 4. Type of variables 5. Checking the dimension of data frame.
Q a	What is meant by Data Wrangling? 1. Data exploration: The data is studied, analyzed of understood by visualizing representation of data. 2. Dealing with missing values. 3. Data reshaping: Data is manipulated according to the requirement, where new data can be added of pre-existing data can be modified. 4. Filtering data: Some data set exists with unwanted rows of columns which are required to be removed or filtered.
07./	Improve data wability Converts data into compatible format.
Ø 8.	What is meant by data weangling process? Cleaning, organising of enriching raw data so that it can be used for decision making process
.е 🔾	Uses of Pandas library. Panda is open bibrary.
	— P.T.0





Manjari (Bk.), Hadapsar, Pune-412307.

	uses:
	1. Data cleaning 2. Data fill
	3 Data normalization 4 merges & joins
	5. Data visualization 6. Statistical analysis
	7. Data inspection 8. Loading & saving data
Q10.	The of number library.
	Numpy is a numerical python library.
	Uses:
	1. Working with arrays, and Hollal 2. Working in domain of linear algebra.
	3. Fourier teansformation
	4. Working with mateices. It is also open source
	Library Tomas Comments
	s. Working is vector - vector multiplication.
() 1/	Uses of matplotlib Distribution Association
7-	Matplotlib is used for data visualization of
	geaphical plotting library (histogram, scatter plots,
	bar chart set)
(12.	Uses of seaborn library.
9	Seaboen library is used for making statistical
	geaphics in python.
	— - P.T. 0.





Manjari (Bk.), Hadapsar, Pune-412307.

()13.	Difference between matplotlib & seaborn
	·Matplotlib:
	It is a python library used to plot various geaphs with the help of additional libraries like numpy of
	with the help of additional dibraries like numpy of
	Pandas.
	Matplotlib creates simple graphs, including histograms,
	bargeaphs, pie charts, scatter plots, lines & other
	visual representation of data.
	Mainly used to plot 20 graphs of arrays.
	Mainly used to plot 2D graphs of arrays. It uses syntax that is relatively complicated f extensive.
	eg: matplotlib.pyplot.bar (x-axis, y-axis) is syntax for bar
	geaph.
	· Seaboen:
	It is also a python library that utilizes matphotlib,
	pardas & numpy to plot graphs.
	It is a superset of matprotlib dibrary.
	It has relatively simple syntax.
	eg: seaboen barplot (x-axis, y-axis) syntax for bar geaph!
<u> </u>	Seabourn is more comfortable with panda data
	feames.
	It prevents overlapping with the help of default
A	themes.
1411.	
()14.	How to install any library in python program
	pip install package_name.
	eg: pip install scaboen.