31437

10th Jan 2022

# Group A: Data Science

## Assignment 1

Title: Data Wrangling, I

Problem Statement:

Perform the following operations using python on any open source dataset (eg data.csv)

1. Import all the required Python Libraries
2. Locate an open source data from the web (eg. https://www.kaggle.com). Provide a clear description of the data and its source (i.c URL of the website.
3. Load the dataset into pandas data frame
4. Data Preprocessing: check for missing values in the data using isnull(), describe() function to get some intial statistics. Provide variable descriptions. Types of variables etc Check the dimensions of the data frame.
5. Data formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e. character, numeric, integer, factor and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantative variables in python

31437

Objectives:

→ Should be able to make raw data useful by applying scientific data processing libraries

→ Understand and apply the python libraries such as numpy, pandas, matplotlib.etc

Outcomes:

→ Understand the importance of data wrangling

→ Able to ~~make~~ take out some useful info from raw data.

Theory:

Data Wrangling:

Data wrangling referred to as data munging, is the process of transforming and mapping data from one raw data into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analytics typically spend the majority of their time in process of data wrangling compared to the actual analysis of the data.

The process of data wrangling may include further munging, data visualization, data aggregation, training a satistical model, as well as many other uses.

Data wrangling typically follows a set of general steps ~~with~~ which begin with extracting the data in a raw form from the data source, parsing the data into predefined data structures. and finding depositing the result content into a data sink for storage and future use

Methods and Functions used:

1) import pandas as pd

i) pd.read_csv()

This method is used to read a comma separated values file into data frame.

Also supports optionally iterating or breaking of the file into chunks

2) data_frame.head(limit)

Returns the first 5 rows of the data frame. To override the default, you may insert a value between the parenthesis to change the number of rows returned.

3) data_frame.shape

Returns a tuple representing the dimensions. for ex- an output (48,14) represents 48 rows and 14 columns.

4) data_frame.describe()

provides descriptive satistics that summarises the central tendency, dispersion and shape

31437

5) • value_counts()
Method returns counts for each unique values. in the column you selected.

6) • is null ()
Returns dataframe with value 'True' where it finds null values and false where it encounters anytype of value.

7) • isnull ().sum()
Returns count of null values in column

8) df ['col'] • unique ()
Returns the unique values in column 'col'.

9) dataFrame • dtypes
This attribute returns the dtypes in the dataframe. It returns a series with the data type of each column.

10) pandas • map()
It is used to map values from two series having one column same, for mapping two series, the last column of the first series should be same as index column of the second series, also the values should be unique

11) data_frame.astype()
This method is used to cast a pandas object to a specified dtype. astype() function also provides the capability to convert any suitable existing column to categorical type. Data Frame. astype() function comes very handy when we want to convert a particular column data type to another data type.

Packages/Libraries used

1) Pandas
Pandas is a software library written for the python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and times series, it is a free software release under the three clause BSD license.
It allows importing data from various file formats such as csv, JSON, SQL databases / queires.

2) Numpy:
Numpy is a library for the python programming language, adding support for large, multi-dimensional arrays and matrices along with a large collection of high-level mathematical functions to operate on these array

Numpy is open-source software and has many contributors.

Analysis:

Pandas is an very efficient scientific data processing library by using which a person can scrap out useful data from raw data for processing and using the data in many applications.

Conclusion:

Data wrangling is one of the most important Technique to turn raw data into useful assest where pandas perfoms an very Important role of converting the so called raw data into productive data sets that can be utilized later for different purposes.