Assignment No 1

Name - Kudharon Sumit Dattatroya
Class - TE
Div - 4
Subject - DSBDAL

Problem Statement -

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

1. Import all the required python libraries.
2. Locate an open source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data & its source (i.e., URL of the web site).
3. Load the Dataset into Pandas data frame.
4. Data Preprocessing : Check for missing values in the data using Pandas insull(), describe() function to get some initial statistics. Provide variable descriptions. Type of variables etc. check the dimensions of the data frame.
5. Data Formatting & Data Normalization : Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor & logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversations.
6. Turn categorical variables into quantitative variables in Python.

In addition to the codes & outputs, explain every operation that you do in the above steps and explain everything that you do import/read/ scrape the data set.

Theory.

1) Explain Dataframe with Suitable example.

→ A Dataframe is a two dimensional data structure, i.e., data is aligned in a tabular fashion in rows & columns.

Features of Dataframe -

- Potentially columns are of different types.
- Size-mutable
- Labeled axes (rows & columns)
- Can perform Arithmetic operations on rows & columns.

Structure -

Let us assume that we are creating a dataframe with student's data.

Regd.No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

Pandas.DataFrame -

Pandas.DataFrame(data, index, columns, dtype, copy)

Now, Create an Dataframe -

```
import pandas as pd
data = [['Alex',10],['Bob',12],['clarke',13]]
df = pd.Dataframe(data, columns=['Name','Age'])
Print(df)
```

Output -

	Name	Age
0	Alex	10
1	Bob	12
2	Clarke	13

2) Explain the steps of Data wrangling ?

→ Data wrangling is the pratice of converting and then plotting data from one "raw" form into another.

There are 6 steps of data wrangling -

① Data Discovery -

This is an all-encompassing term that describes understanding what your data is all about. In this first step you get familiar with your data

② Data Structuring -

When you collect raw data it initially is in all shapes & sizes and has no definite structure. Such data needs to be restructured to suit the analytical model that your enterprise plans to deploy.

③ Data Cleaning -

Raw data comes with some errors that need to be fixed before data is passed on to the next stage. Cleaning involves the tracking of outliers making corrections or deleting bad data completely.

④ Data Enriching -

By this stage you have kind of become familiar with the data in hand. Now is the time to ask yourself this question - do you need to embellish the raw data? Do you want to augment it with other data.

⑤ Data Validating -

The activity surfaces data quality issues, and they have to be addressed with the necessary transformations. The rules of validation rules require repitative programming steps to check the authencity and the quality of your data.

⑥ Data publishing -

Once all the above steps are completed, the final output of your data wrangling efforts are pushed downstream for your analytics needs

3) What is the need of data normalization?

→ Data normalization is a method in which data attributes are structured to improve the cohesion of the types of entities contwin a data model

Why data normalization required?

① Duplicate Data Reduce -

Reducing the number of duplicates in your database is one of the biggest impacts of normalizing your results. until matching and combining duplicates, normalizing the data will be make it easier to find the duplicates if you don't use a deduplication tool that automatically does it like Ring Lead Cleanse.

② Segmentation for marketing -

Another advantage of normalizing the information is that it will assist the leaders of the marketing team section, especially with job titles. Job titles differ widely between business and sectors, making it almost difficult to equate a given job title with something actionable for segmentation or lead scoring. So it can be very useful to standardize this value and a variety of approaches are possible

③ Metrics & performance -

When it comes to analyzing data, databases that arenot structured and poorly managed can cause significant headaches. Your data coowd be considerably easier to coork through by standardizing your data by using a single organizational approach of appropriate capitialization. No to mention since they won't have to spend time sorting the info, the Sales and marketing departments will save Precious time. Translating insane details into a structured list allows you the freedom to take steps that coill be hard or difficult to do properly otherwise.

4) What are the different Techniques for handling the missing data?

→ There are two primary ways to handle missing data -

1) Deleting the Missing Values -

Generally this approach is not recommended. It is one of the quick & dirty techniques one can use to deal with missing values. If the missing value is the type of MNAR then it should not be delated. If the missing value is of type of MAR or McAR then it can be deleted.

There are two ways to defete the missing values -

① Deleting entire row.

② Deleting the entire column.

2) Imputing the Missing Value -

There are different ways of replacing the missing values.

① Replacing with arbitrary values.

② Replacing with mean.

③ Replacing with mode,

④ Replacing with median

5) What is type conversion, how to do it in Python?

→ There are two types of data conversion-

① Implicit type conversion.

② Explicit type conversion.

① Implicit type conversion -

In this type Conversion of data types in python, the Python interpreter automatically converts one data type to another without any user involvement.

Example -

```
x=10
Print("x is of type:", type(x))
Y=10.6
Print("Y is of type:", type(Y))
x=x+y
print(x)
Print("x is of type:", type(x))
```

output-

```
x is of type: <class 'int'>
Y is of type: <class 'float'>
20.6
x is of type: <class 'float'>
```

② Explicit type conversion -

In this type conversion the data type is manually changed by user as per their requirement.

There are some explicit type conversions listed below -

1) int(a, base)	5) oct()	9) dict()
2) float()	6) tuple()	10) str()
3) ord()	7) set()	11) complex(real, imag)
4) hex()	8) list()	12) chr(number)

Example -

```
s = 'Sumit'      # string type
t = tuple(s)     # converting string into tuple.
print(t)
```

output -

```
('S','U','m','i','t')
```

6) What is .csv ?

→ csv stands for Comma Separated Values. A csv file is a plain text file that stores tables & spreadsheet information