Assignment No.5

Title: Perform the following operations using Python on the Air quality and Heart Diseases data sets.

1. Data cleaning
2. Data integration
3. Data transformation
4. Error correcting
5. Data model building

Objectives:

1. To understand and apply the Analytical concept of big data using Python.
2. To study detailed concept Python.

SOFTWARE REQUIREMENTS:

1. Windows 8 or above
2. Python SDK
3. 3. IDE (p0079charm, anaconda, cloud notebook)

THEORY:

Data cleaning or data preparation is an essential part of statistical analysis. In fact, in practice it is often more time-consuming than the statistical analysis itself.

1) 1) Data cleaning

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import random as rd

ds = pd.read_excel("AirQuality.xlsx")

ds_heart = pd.read_csv("heart.csv")

ds.head()

ds.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 9357 entries, 0 to 9356 Data columns (total 15 columns): # Column Non-Null Count Dtype----------------------------0 Date 9357 non-null object 1 Time 9357 non-null object 2 CO(GT) 9357 non-null object 3 PT08.S1(CO) 9357 non-null int64 4 NMHC(GT) 9357 non-null int64

ds.isnull().sum()
ds.dropna()
```

2) Data integration

```
ds1 = ds.loc[111:999, ['Date', 'Time', 'C6H6(GT)', 'RH']]
ds2 = ds.iloc[[1,3,5,2,4,22,43,54,67,7,8,9,50,10,11]]
```

```
ds_integration = pd.concat([ds1,ds2])
```

```
ds_integration
```

3) Data transformation

```
ds_integration.transpose()

ds.drop(columns = "NOx(GT)")
ds2.drop(1)
ds.melt()


ds_merged = pd.concat([ds,ds_heart])
ds_merged
```

4)Error correcting:

#Error Correction

##Check for the data characters mistakes ###feature 'ca' ranges from 0–3, however, df.nunique() listed 0–4. So lets find '4' and change them to NaN.

```
df['ca'].unique()
```

```
array([0, 2, 1, 3, 4], dtype=object)
```

```
#to count the number in of each category descending order
df.ca.value_counts()
```

```
0    175
1     65
2     38
3     20
4      5
Name: ca, dtype: int64
```

5) Data model building

 # Step1: Divide the dataset into taining and Testing

#step 2: design a model

Step 3: perform the accuracy measures

CONCLUSION: Thus, we have learnt how to Perform the different Data Cleaning and Data modeling operations using Python**.**