CHANDIGARH UNIVERSITY
UNIVERSITY INSTITUTE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Submitted by:	**Submitted To:** Ajay Kumar (E13141)
Subject Name:	Machine Learning Lab
Subject Code:	20CSP-317
Branch:	CSE
Semester:	5th

LAB INDEX

Sr. No	Program	Date	Evaluation				Sign
			LW (12)	VV (8)	FW (10)	Total (30)	
1.							
2.							
3.							
4.							
5.							

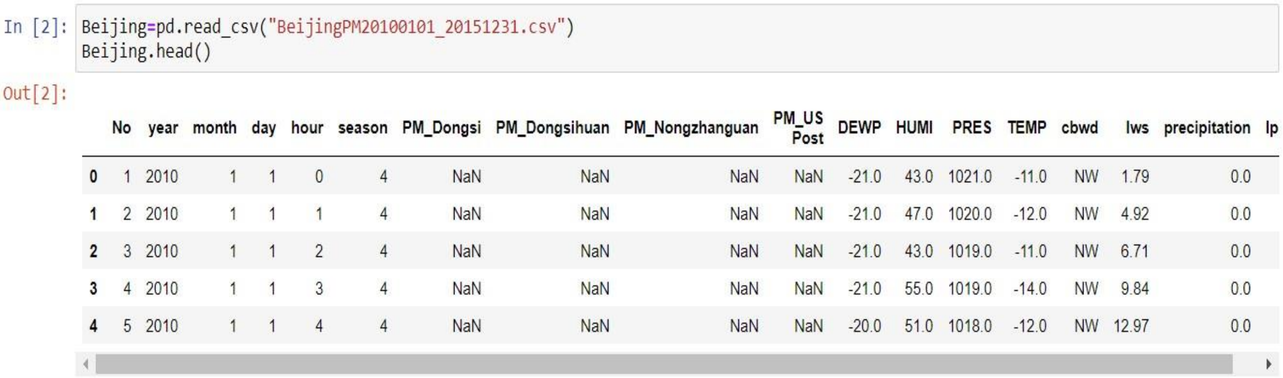
Experiment-2

Aim/Overview of the practical: To perform Data Visualization

Code and output:

```
# Importing Libraries
import pandas as pd import
numpy as np import
matplotlib.pyplot as plt import
seaborn as sns

# Reading the data
```

Beijing=pd.read_csv("BeijingPM20100101_20151231.csv")
Beijing.head()

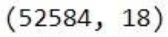


```
In [2]: Beijing=pd.read_csv("BeijingPM20100101_20151231.csv")
        Beijing.head()
```

Out[2]:

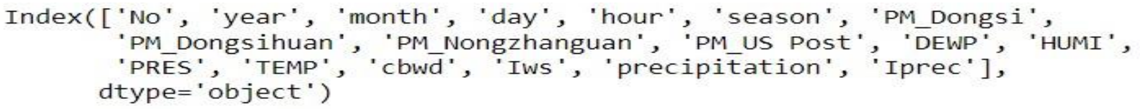
	No	year	month	day	hour	season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhanguan	PM_US Post	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Ip
0	1	2010	1	1	0	4	NaN	NaN	NaN	NaN	-21.0	43.0	1021.0	-11.0	NW	1.79	0.0	
1	2	2010	1	1	1	4	NaN	NaN	NaN	NaN	-21.0	47.0	1020.0	-12.0	NW	4.92	0.0	
2	3	2010	1	1	2	4	NaN	NaN	NaN	NaN	-21.0	43.0	1019.0	-11.0	NW	6.71	0.0	
3	4	2010	1	1	3	4	NaN	NaN	NaN	NaN	-21.0	55.0	1019.0	-14.0	NW	9.84	0.0	
4	5	2010	1	1	4	4	NaN	NaN	NaN	NaN	-20.0	51.0	1018.0	-12.0	NW	12.97	0.0	

Beijing.shape



```
(52584, 18)
```

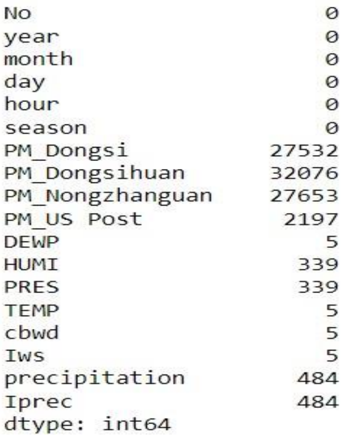
Beijing.columns



```
Index(['No', 'year', 'month', 'day', 'hour', 'season', 'PM_Dongsi',
       'PM_Dongsihuan', 'PM_Nongzhanguan', 'PM_US Post', 'DEWP', 'HUMI',
       'PRES', 'TEMP', 'cbwd', 'Iws', 'precipitation', 'Iprec'],
      dtype='object')
```

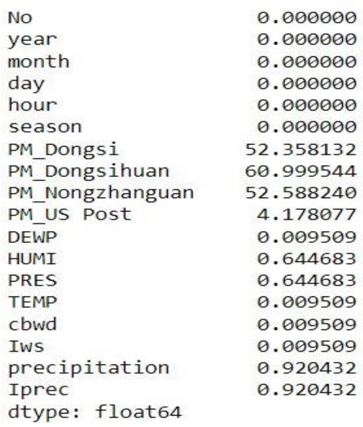
# Calculating the percentage of NaN values in the Data set

Beijing.isnull().sum()



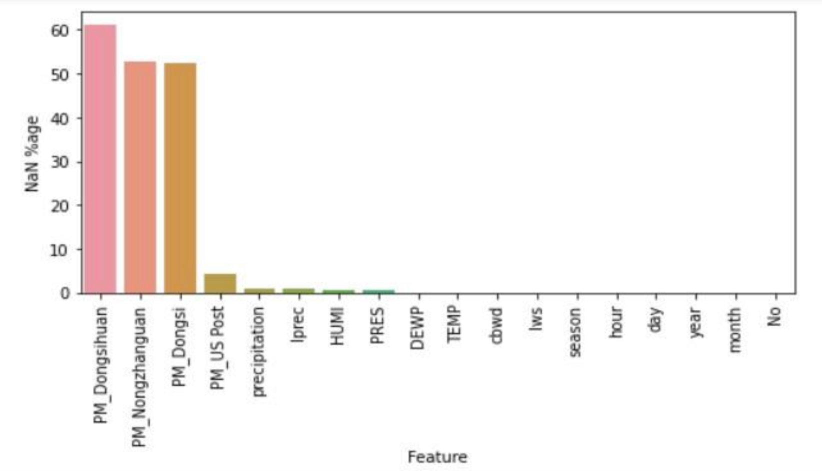
```
No                     0
year                   0
month                  0
day                    0
hour                   0
season                 0
PM_Dongsi          27532
PM_Dongsihuan      32076
PM_Nongzhanguan    27653
PM_US Post          2197
DEWP                   5
HUMI                 339
PRES                 339
TEMP                   5
cbwd                   5
Iws                    5
precipitation        484
Iprec                484
dtype: int64
```

Beijing.isnull().mean()*100

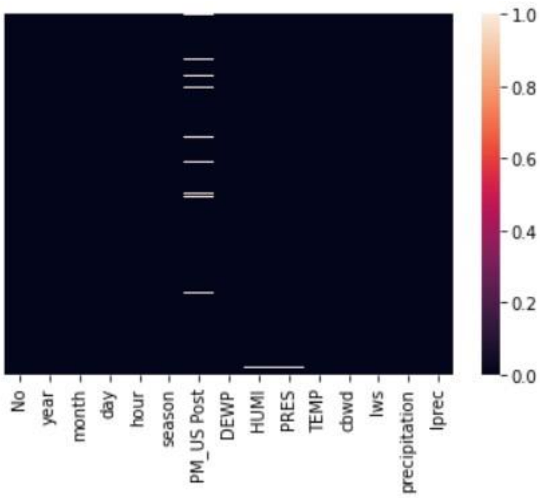


```
No                    0.000000
year                  0.000000
month                 0.000000
day                   0.000000
hour                  0.000000
season                0.000000
PM_Dongsi            52.358132
PM_Dongsihuan        60.999544
PM_Nongzhanguan      52.588240
PM_US Post            4.178077
DEWP                  0.009509
HUMI                  0.644683
PRES                  0.644683
TEMP                  0.009509
cbwd                  0.009509
Iws                   0.009509
precipitation         0.920432
Iprec                 0.920432
dtype: float64
```

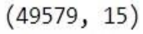
```
NaN_percentage = pd.DataFrame(Beijing.isnull().mean()*100,columns=["NaN %age"]).reset_index().sort_values(by='NaN %age',ascending=False)
NaN_percentage.rename(columns={"index":"Feature"},inplace=True)

# Visualization for dropping NaN values

plt.figure(figsize=(8,3.5))
sns.barplot(x="Feature",y="NaN %age",data=NaN_percentage)
plt.xticks(rotation=90)
```



```
for f in Beijing.columns:
    if(Beijing[f].isnull().mean()*100>30):
Beijing.drop(f,inplace=True,axis=1)
```

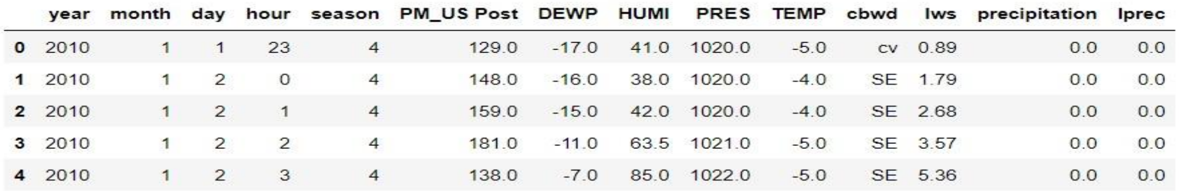


```
Beijing.dropna(inplace=True)

Beijing.shape
```



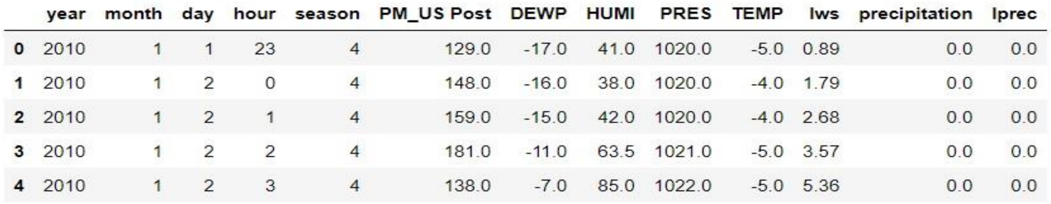
```
(49579, 15)
```

```
Beijing.reset_index(inplace=True)

#Dropping unecessary features

Beijing.drop(["index","No"],axis=1,inplace=True)
Beijing.head()
```



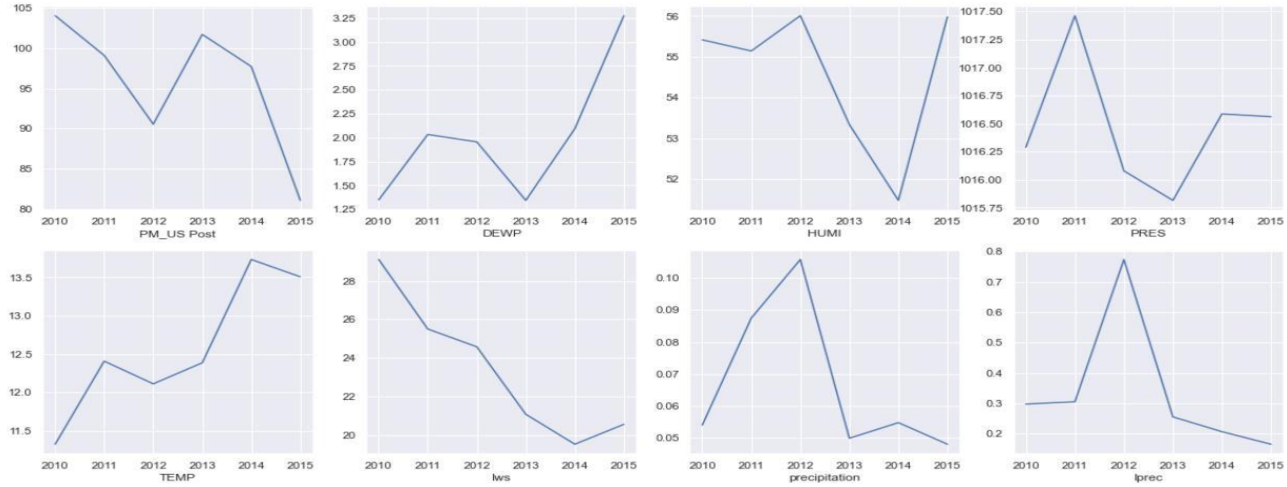
	year	month	day	hour	season	PM_US Post	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Iprec
0	2010	1	1	23	4	129.0	-17.0	41.0	1020.0	-5.0	cv	0.89	0.0	0.0
1	2010	1	2	0	4	148.0	-16.0	38.0	1020.0	-4.0	SE	1.79	0.0	0.0
2	2010	1	2	1	4	159.0	-15.0	42.0	1020.0	-4.0	SE	2.68	0.0	0.0
3	2010	1	2	2	4	181.0	-11.0	63.5	1021.0	-5.0	SE	3.57	0.0	0.0
4	2010	1	2	3	4	138.0	-7.0	85.0	1022.0	-5.0	SE	5.36	0.0	0.0

```
#Data processing on numerical features

Beijing_numerical=Beijing.select_dtypes(exclude="object").copy()

Beijing_numerical.head()
```

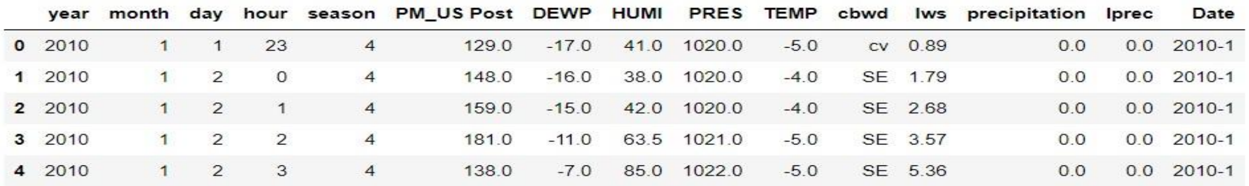


	year	month	day	hour	season	PM_US Post	DEWP	HUMI	PRES	TEMP	Iws	precipitation	Iprec
0	2010	1	1	23	4	129.0	-17.0	41.0	1020.0	-5.0	0.89	0.0	0.0
1	2010	1	2	0	4	148.0	-16.0	38.0	1020.0	-4.0	1.79	0.0	0.0
2	2010	1	2	1	4	159.0	-15.0	42.0	1020.0	-4.0	2.68	0.0	0.0
3	2010	1	2	2	4	181.0	-11.0	63.5	1021.0	-5.0	3.57	0.0	0.0
4	2010	1	2	3	4	138.0	-7.0	85.0	1022.0	-5.0	5.36	0.0	0.0

```
# Visualizing the Time series data for Yearly trends

f = ["year","hour","month","day","season"] sns.set()
plt.figure(figsize=(20,20)) for i,c in
enumerate(Beijing_numerical.drop(f,axis=1).columns):     if c
not in f:
        plt.subplot(4,4,i+1)
        plt.plot(Beijing_numerical.groupby("year").mean()[c])
plt.xlabel(c)
```

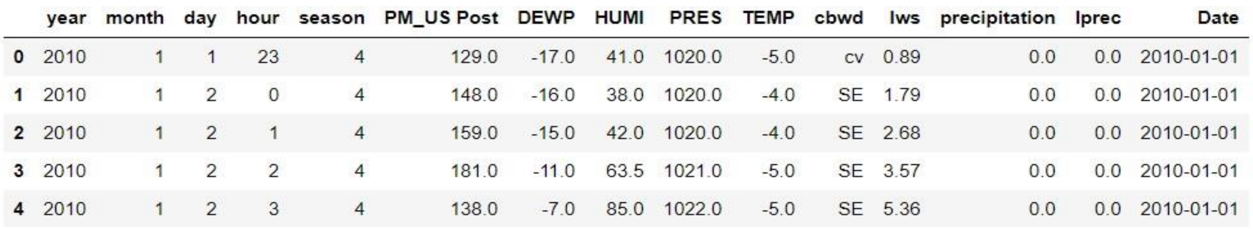


```
# Preparing time series data for visulizing Monthly trends dates=[]

for i in range(Beijing.shape[0]):
lst=[str(Beijing["year"][i]),str(Beijing["month"][i])]
st="-"     s=st.join(lst)
    dates.append(s)

Beijing["Date"]=dates
Beijing.head()
```



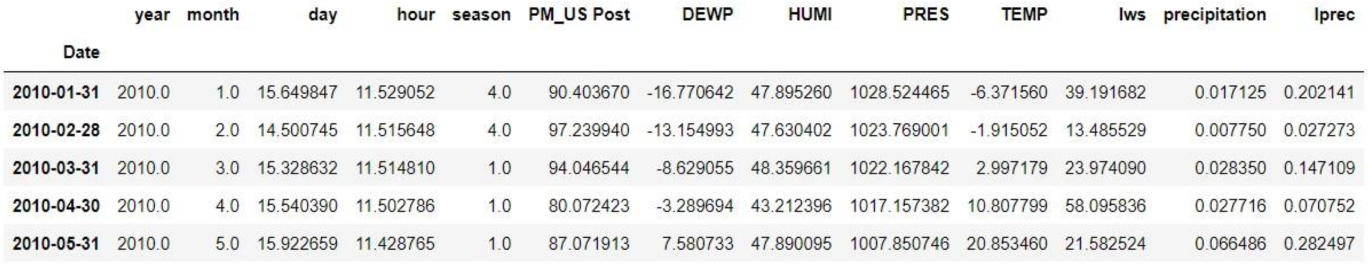
	year	month	day	hour	season	PM_US Post	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Iprec	Date
0	2010	1	1	23	4	129.0	-17.0	41.0	1020.0	-5.0	cv	0.89	0.0	0.0	2010-1
1	2010	1	2	0	4	148.0	-16.0	38.0	1020.0	-4.0	SE	1.79	0.0	0.0	2010-1
2	2010	1	2	1	4	159.0	-15.0	42.0	1020.0	-4.0	SE	2.68	0.0	0.0	2010-1
3	2010	1	2	2	4	181.0	-11.0	63.5	1021.0	-5.0	SE	3.57	0.0	0.0	2010-1
4	2010	1	2	3	4	138.0	-7.0	85.0	1022.0	-5.0	SE	5.36	0.0	0.0	2010-1

```
Beijing["Date"]=pd.to_datetime(Beijing["Date"])
Beijing.head()
```



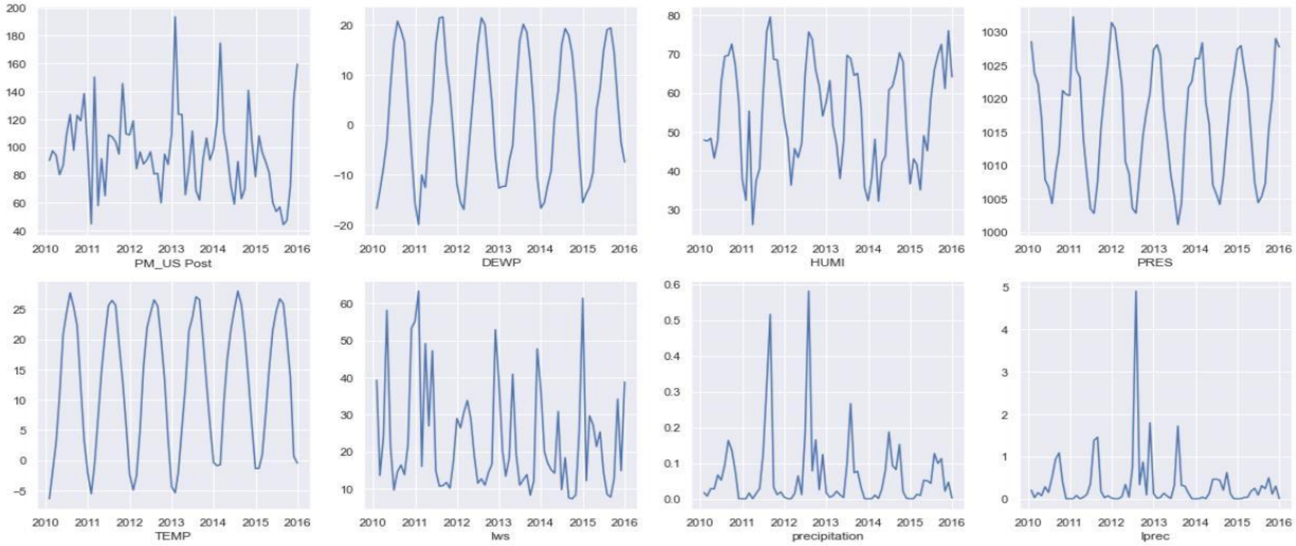
	year	month	day	hour	season	PM_US Post	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Iprec	Date
0	2010	1	1	23	4	129.0	-17.0	41.0	1020.0	-5.0	cv	0.89	0.0	0.0	2010-01-01
1	2010	1	2	0	4	148.0	-16.0	38.0	1020.0	-4.0	SE	1.79	0.0	0.0	2010-01-01
2	2010	1	2	1	4	159.0	-15.0	42.0	1020.0	-4.0	SE	2.68	0.0	0.0	2010-01-01
3	2010	1	2	2	4	181.0	-11.0	63.5	1021.0	-5.0	SE	3.57	0.0	0.0	2010-01-01
4	2010	1	2	3	4	138.0	-7.0	85.0	1022.0	-5.0	SE	5.36	0.0	0.0	2010-01-01

```
Beijing_dates=Beijing.groupby(pd.Grouper(key='Date', axis=0,
freq='M')).mean() Beijing_dates.head()
```



	year	month	day	hour	season	PM_US Post	DEWP	HUMI	PRES	TEMP	Iws	precipitation	Iprec
Date													
2010-01-31	2010.0	1.0	15.649847	11.529052	4.0	90.403670	-16.770642	47.895260	1028.524465	-6.371560	39.191682	0.017125	0.202141
2010-02-28	2010.0	2.0	14.500745	11.515648	4.0	97.239940	-13.154993	47.630402	1023.769001	-1.915052	13.485529	0.007750	0.027273
2010-03-31	2010.0	3.0	15.328632	11.514810	1.0	94.046544	-8.629055	48.359661	1022.167842	2.997179	23.974090	0.028350	0.147109
2010-04-30	2010.0	4.0	15.540390	11.502786	1.0	80.072423	-3.289694	43.212396	1017.157382	10.807799	58.095836	0.027716	0.070752
2010-05-31	2010.0	5.0	15.922659	11.428765	1.0	87.071913	7.580733	47.890095	1007.850746	20.853460	21.582524	0.066486	0.282497

```
f
= ["day","month","year","hour","season"]

plt.figure(figsize=(20,20))      for      i,c      in
enumerate(Beijing_dates.drop(f,axis=1).columns):   if c not
in f:
        plt.subplot(4,4,i+1)
plt.plot(Beijing_dates[c])      plt.xlabel(c)
```


PM_US Post
DEWP
HUMI
PRES
TEMP
Iws
precipitation
Iprec