On The Plague Trail

1. Business Problem

1.1 Problem Description

Predict the total number of people infected by the 7 different pathogens.

Plague is an epidemic event caused by Bacteria. A group of senior scientists misplaced a package containing fatal plague bacteria during one of their trips. With no means of tracking where the package is, scientists are now trying to come up with a solution to stop the plague. This plague has 7 different strains that are unique for each continent. This strain is expanding rapidly in each continent.

1.2 Problem Statement

The dataset contains escalations of the plague for all the seven strains. The dataset is a time series in which the training set contains the number of individuals that are infected by the plague over a defined period of time. Your mission, should you choose to accept it, is to defend the world against this plague by building an algorithm that can minimize the damage.

1.3 Data Description

You can find the dataset : https://www.kaggle.com/shivammittal99/hackerearth-on-the-plague-trail#train.csv

1.4 Real world/Business Objectives and constraints

Objectives: 1. Predict the columns of PA, PB, PC, PD, PE, PF, PG. 2. Minimize the difference between predicted and actual values (RMSE and MAPE)

1.5 Column Description

ID:

A calculated unique ID for each research.

DateTime:

Represents the data and time on which the event is recorded

TempOut:

Outside Temperature

HiTemp:

Highest Temperature

LowTemp:

Lowest Temperature

OutHum:

Outside Humidity

DewPt:

Dew Point

WindSpeed:

Wind Speed

WindDir:

Wind Direction

WindRun:

Wind Run Flow

HiSpeed:

Highest Speed of the wind

HiDir:

Direction of the wind which has highest speed

WindChill:

Chillness of the wind

HeatIndex:

Heat Index

THWIndex:

THW Index

Bar:

Barometer Reading

Rain:

Rain

RainRate:

Frequency of Rain

HeatDD:

Heat DD

CoolDD:

Cool DD

InTemp:

Temperature Inside

InHum:

Humidity Inside

InDew:

Dew Inside

InHeat:

Heat Inside

InEMC:

EMC Inside

InAirDensity:

Air Density

WindSamp:

Wind - Attribute 1

WindTx:

Wind - Attribute 2

ISSRecpt:

Reception

ArcInt:

Attribute

PA:

Total No of People infected by Pathogen A

PB:

Total No of People infected by Pathogen B

PC:

Total No of People infected by Pathogen C

PD:

Total No of People infected by Pathogen D

PE:

Total No of People infected by Pathogen E

PF:

Total No of People infected by Pathogen F

PG:

Total No of People infected by Pathogen G

1.6 Mapping the real world problem to a Machine Learning Problem

```
We need to predict the number of people affected by pathogens (PA,PB,PC,PD,PE,PF,PG).
It is a Regression problem
```

In [2]:

```
#import all the necessary packages.

from PIL import Image
import requests
from io import BytesIO
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import warnings
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
import math
import time
import re
import os
import seaborn as sns
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics import pairwise_distances
from matplotlib import gridspec
from scipy.sparse import hstack
import plotly
import plotly.figure_factory as ff
from plotly.graph_objs import Scatter, Layout

plotly.offline.init_notebook_mode(connected=True)
warnings.filterwarnings("ignore")
```

In [3]:

```
import os
os.chdir('C:/Users/kingsubham27091995/Desktop/AppliedAiCouse/CASE STUDIES/On the Plague trail')
```

In [4]:

```
train_data=pd.read_csv("train.csv")
```

In [5]:

```
print("Number of data points:{0} and Number of features:{1}".format(train_data.shape[0],train_data
.shape[1]))
```

```
Number of data points:40000 and Number of features:37
```

In [5]:

```
train_data.head(5)
```

Out[5]:

	ID	DateTime	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	WindRun	...	WindTx	ISSRecp
0	PR00001	07-12-2040 0:15	53.5	53.6	53.5	85	49.1	2	SSE	0.5	...	1	100.0
1	PR00002	07-12-2040 0:30	53.5	53.5	53.4	85	49.1	2	SSE	0.5	...	1	100.0
2	PR00003	07-12-2040 0:45	53.3	53.5	53.2	85	48.9	2	SSE	0.5	...	1	100.0
3	PR00004	07-12-2040 1:00	53.1	53.3	53.0	86	49.0	2	S	0.5	...	1	100.0
4	PR00005	07-12-2040 1:15	52.9	53.1	52.9	86	48.8	2	S	0.5	...	1	100.0

5 rows × 37 columns

In [6]:

```
train_data.tail(5)
```

Out[6]:

	ID	DateTime	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	WindRun	...	WindTx	ISS
39995	PR39996	04-01-2042 0:00	55.0	55.1	55.0	88	51.5	1	SSE	0.25	...	1	100
39996	PR39997	04-04-2042 12:00	60.1	60.5	59.1	72	51.0	3	SSE	0.75	...	1	100
39997	PR39998	08-04-2041 10:30	79.6	79.6	75.6	40	53.1	1	S	0.25	...	1	100
39998	PR39999	08-04-2041 11:00	81.2	82.0	80.6	38	53.2	3	SSE	0.75	...	1	100
39999	PR40000	08-04-2041 11:15	82.9	83.0	80.9	37	54.0	3	SSE	0.75	...	1	100

5 rows × 37 columns

In [7]:

```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 37 columns):
ID              40000 non-null object
DateTime        40000 non-null object
TempOut         40000 non-null float64
HiTemp          40000 non-null float64
LowTemp         40000 non-null float64
OutHum          40000 non-null int64
DewPt           40000 non-null float64
WindSpeed       40000 non-null int64
WindDir         40000 non-null object
WindRun         40000 non-null float64
HiSpeed         40000 non-null int64
HiDir           40000 non-null object
WindChill       40000 non-null float64
HeatIndex       40000 non-null float64
THWIndex        40000 non-null float64
Bar             40000 non-null float64
Rain            40000 non-null float64
RainRate        40000 non-null float64
HeatDD          40000 non-null float64
CoolDD          40000 non-null float64
InTemp          40000 non-null float64
InHum           40000 non-null int64
InDew           40000 non-null float64
InHeat          40000 non-null float64
InEMC           40000 non-null float64
InAirDensity    40000 non-null float64
WindSamp        40000 non-null int64
WindTx          40000 non-null int64
ISSRecpt        40000 non-null float64
ArcInt          40000 non-null int64
PA              40000 non-null int64
PB              40000 non-null int64
PC              40000 non-null int64
PD              40000 non-null int64
PE              40000 non-null int64
PF              40000 non-null int64
PG              40000 non-null int64
dtypes: float64(19), int64(14), object(4)
memory usage: 11.3+ MB
```

Basic statistics for each features

In [8]:

```
for i in train_data.columns:
    print("Basic statistics for feature : {0}".format(i))
    print(train_data[i].describe())
    print("---------------------------------------------")
```

```
Basic statistics for feature : ID
count        40000
unique       40000
top        PR35754
freq             1
Name: ID, dtype: object
---------------------------------------------
Basic statistics for feature : DateTime
count                40000
unique               40000
top        02-08-2041 3:30
freq                     1
Name: DateTime, dtype: object
---------------------------------------------
Basic statistics for feature : TempOut
count     40000.000000
mean         58.508625
```

```
std          12.119640
min          29.300000
25%          51.100000
50%          56.400000
75%          65.300000
max         110.300000
Name: TempOut, dtype: float64
---------------------------------------------
Basic statistics for feature : HiTemp
count    40000.000000
mean        58.975230
std         12.323427
min         29.500000
25%         51.300000
50%         56.800000
75%         66.000000
max        111.000000
Name: HiTemp, dtype: float64
---------------------------------------------
Basic statistics for feature : LowTemp
count    40000.000000
mean        58.056785
std         11.916335
min         29.300000
25%         50.800000
50%         56.100000
75%         64.700000
max        108.600000
Name: LowTemp, dtype: float64
---------------------------------------------
Basic statistics for feature : OutHum
count    40000.000000
mean        72.915750
std         20.873482
min          4.000000
25%         58.000000
50%         79.000000
75%         91.000000
max         98.000000
Name: OutHum, dtype: float64
---------------------------------------------
Basic statistics for feature : DewPt
count    40000.000000
mean        48.156873
std          7.895771
min          1.200000
25%         43.600000
50%         49.700000
75%         53.900000
max         66.900000
Name: DewPt, dtype: float64
---------------------------------------------
Basic statistics for feature : WindSpeed
count    40000.000000
mean         2.348650
std          2.346365
min          0.000000
25%          0.000000
50%          2.000000
75%          4.000000
max         16.000000
Name: WindSpeed, dtype: float64
---------------------------------------------
Basic statistics for feature : WindDir
count      40000
unique        17
top          SSE
freq        9870
Name: WindDir, dtype: object
---------------------------------------------
Basic statistics for feature : WindRun
count    40000.000000
mean         0.587163
std          0.586591
min          0.000000
25%          0.000000
50%          0.500000
```

```
75%          1.000000
max          4.000000
Name: WindRun, dtype: float64
---------------------------------------------
Basic statistics for feature : HiSpeed
count    40000.000000
mean         6.028675
std          4.808251
min          0.000000
25%          2.000000
50%          5.000000
75%          9.000000
max         33.000000
Name: HiSpeed, dtype: float64
---------------------------------------------
Basic statistics for feature : HiDir
count     40000
unique       17
top         SSE
freq       8470
Name: HiDir, dtype: object
---------------------------------------------
Basic statistics for feature : WindChill
count    40000.000000
mean        58.373335
std         12.167000
min         29.000000
25%         50.800000
50%         56.300000
75%         65.200000
max        110.300000
Name: WindChill, dtype: float64
---------------------------------------------
Basic statistics for feature : HeatIndex
count    40000.000000
mean        58.139203
std         11.858623
min         29.100000
25%         51.000000
50%         56.100000
75%         64.600000
max        107.100000
Name: HeatIndex, dtype: float64
---------------------------------------------
Basic statistics for feature : THWIndex
count    40000.000000
mean        58.003950
std         11.912303
min         28.800000
25%         50.800000
50%         55.900000
75%         64.500000
max        107.100000
Name: THWIndex, dtype: float64
---------------------------------------------
Basic statistics for feature : Bar
count    40000.000000
mean        30.071947
std          0.145422
min         29.619000
25%         29.961000
50%         30.055000
75%         30.168000
max         30.534000
Name: Bar, dtype: float64
---------------------------------------------
Basic statistics for feature : Rain
count    40000.000000
mean         0.000505
std          0.004234
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max          0.190000
Name: Rain, dtype: float64
---------------------------------------------
```

```
Basic statistics for feature : RainRate
count    40000.000000
mean         0.003950
std          0.058002
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max          5.940000
Name: RainRate, dtype: float64
---------------------------------------------
Basic statistics for feature : HeatDD
count    40000.000000
mean         0.094455
std          0.084451
min          0.000000
25%          0.000000
50%          0.090000
75%          0.145000
max          0.372000
Name: HeatDD, dtype: float64
---------------------------------------------
Basic statistics for feature : CoolDD
count    40000.000000
mean         0.026837
std          0.061124
min          0.000000
25%          0.000000
50%          0.000000
75%          0.003000
max          0.472000
Name: CoolDD, dtype: float64
---------------------------------------------
Basic statistics for feature : InTemp
count    40000.000000
mean        69.171345
std          2.036967
min         58.900000
25%         68.300000
50%         69.200000
75%         70.100000
max         82.000000
Name: InTemp, dtype: float64
---------------------------------------------
Basic statistics for feature : InHum
count    40000.000000
mean        47.259250
std         13.889228
min         16.000000
25%         36.000000
50%         46.000000
75%         58.000000
max         88.000000
Name: InHum, dtype: float64
---------------------------------------------
Basic statistics for feature : InDew
count    40000.000000
mean        47.181495
std          8.363692
min         21.100000
25%         40.600000
50%         48.300000
75%         53.800000
max         66.600000
Name: InDew, dtype: float64
---------------------------------------------
Basic statistics for feature : InHeat
count    40000.000000
mean        67.406550
std          2.685041
min         55.900000
25%         66.100000
50%         67.700000
75%         68.800000
max         81.100000
Name: InHeat, dtype: float64
---------------------------------------------
```

```
Basic statistics for feature : InEMC
count    40000.000000
mean         9.043872
std          2.415366
min          3.940000
25%          7.220000
50%          8.640000
75%         10.750000
max         19.360000
Name: InEMC, dtype: float64
---------------------------------------------
Basic statistics for feature : InAirDensity
count    40000.000000
mean         0.074569
std          0.000644
min          0.072900
25%          0.074100
50%          0.074500
75%          0.074900
max          0.077400
Name: InAirDensity, dtype: float64
---------------------------------------------
Basic statistics for feature : WindSamp
count    40000.000000
mean       351.205575
std          0.697801
min        323.000000
25%        351.000000
50%        351.000000
75%        351.000000
max        353.000000
Name: WindSamp, dtype: float64
---------------------------------------------
Basic statistics for feature : WindTx
count    40000.0
mean         1.0
std          0.0
min          1.0
25%          1.0
50%          1.0
75%          1.0
max          1.0
Name: WindTx, dtype: float64
---------------------------------------------
Basic statistics for feature : ISSRecpt
count    40000.000000
mean        99.997938
std          0.106524
min         94.400000
25%        100.000000
50%        100.000000
75%        100.000000
max        100.000000
Name: ISSRecpt, dtype: float64
---------------------------------------------
Basic statistics for feature : ArcInt
count    40000.0
mean        15.0
std          0.0
min         15.0
25%         15.0
50%         15.0
75%         15.0
max         15.0
Name: ArcInt, dtype: float64
---------------------------------------------
Basic statistics for feature : PA
count    40000.000000
mean       372.452375
std        645.413994
min          1.000000
25%          7.000000
50%         55.000000
75%        403.000000
max       2980.000000
Name: PA, dtype: float64
---------------------------------------------
```

```
---------------------------------------------
Basic statistics for feature : PB
count    40000.000000
mean       197.904025
std        321.658543
min          1.000000
25%          6.000000
50%         38.000000
75%        234.000000
max       1440.000000
Name: PB, dtype: float64
---------------------------------------------
Basic statistics for feature : PC
count    40000.000000
mean       117.700025
std        180.131998
min          1.000000
25%          5.000000
50%         28.000000
75%        148.000000
max        786.000000
Name: PC, dtype: float64
---------------------------------------------
Basic statistics for feature : PD
count    40000.0000
mean        76.2855
std        110.3007
min          1.0000
25%          5.0000
50%         22.0000
75%        101.0000
max        470.0000
Name: PD, dtype: float64
---------------------------------------------
Basic statistics for feature : PE
count    40000.000000
mean        52.868375
std         72.429328
min          1.000000
25%          4.000000
50%         17.000000
75%         73.000000
max        303.000000
Name: PE, dtype: float64
---------------------------------------------
Basic statistics for feature : PF
count    40000.000000
mean        38.638975
std         50.285082
min          1.000000
25%          4.000000
50%         14.000000
75%         55.000000
max        207.000000
Name: PF, dtype: float64
---------------------------------------------
Basic statistics for feature : PG
count    40000.000000
mean        29.472725
std         36.520023
min          1.000000
25%          3.000000
50%         12.000000
75%         43.000000
max        148.000000
Name: PG, dtype: float64
---------------------------------------------
```

In [9]:

```
train_data.isnull().sum()
```

Out[9]:

```
ID               0
DateTime         0
```

```
TempOut          0
HiTemp           0
LowTemp          0
OutHum           0
DewPt            0
WindSpeed        0
WindDir          0
WindRun          0
HiSpeed          0
HiDir            0
WindChill        0
HeatIndex        0
THWIndex         0
Bar              0
Rain             0
RainRate         0
HeatDD           0
CoolDD           0
InTemp           0
InHum            0
InDew            0
InHeat           0
InEMC            0
InAirDensity     0
WindSamp         0
WindTx           0
ISSRecpt         0
ArcInt           0
PA               0
PB               0
PC               0
PD               0
PE               0
PF               0
PG               0
dtype: int64
```

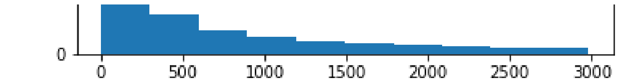
Checking for Skewness and Log Transformations:

In [6]:

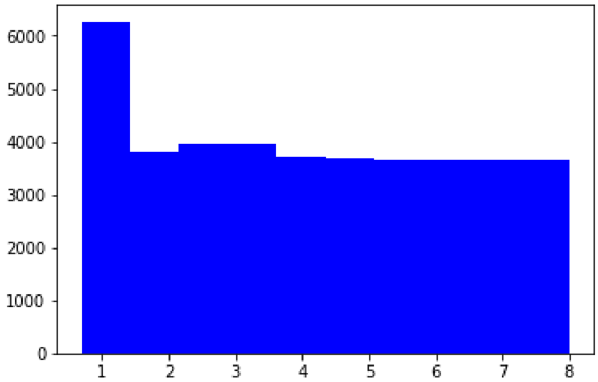
```
# Determining the Skewness of data
outputs= ["PA","PB","PC","PD","PE","PF","PG"]
for result in outputs:
    print("For "+ result)
    print("="*50)
    print ("Skew is:", train_data[result].skew())
    plt.hist(train_data.PA)
    plt.show()

    print("For "+result+" after log transformation")
    print("="*50)
    # After log transformation of the data it looks much more center aligned
    train_data['Skewed_PA'] = np.log(train_data['PA']+1)
    print ("Skew is:", train_data['Skewed_PA'].skew())
    plt.hist(train_data['Skewed_PA'], color='blue')
    plt.show()
```

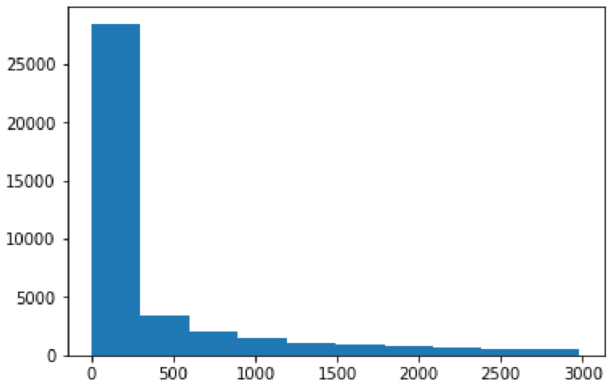
```
For PA
==================================================
Skew is: 2.1808251642863983
```



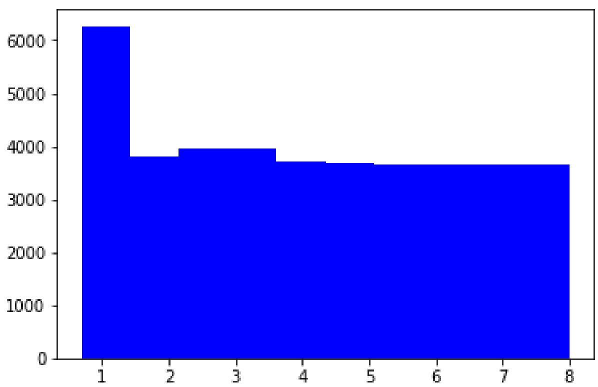

```
For PA after log transformation
==================================================
Skew is: 0.09780414828042107
```



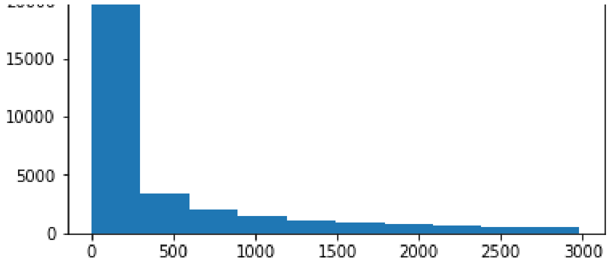

```
For PB
==================================================
Skew is: 2.0367205108020845
```



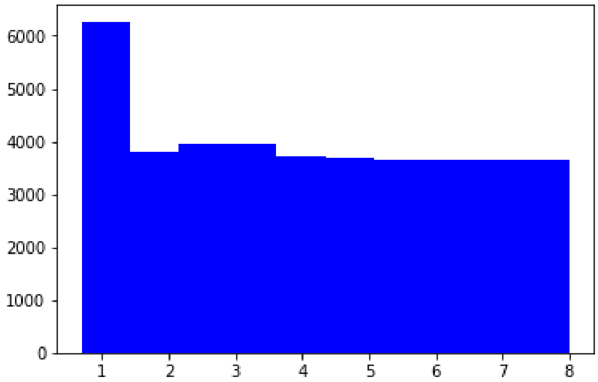

```
For PB after log transformation
==================================================
Skew is: 0.09780414828042107
```



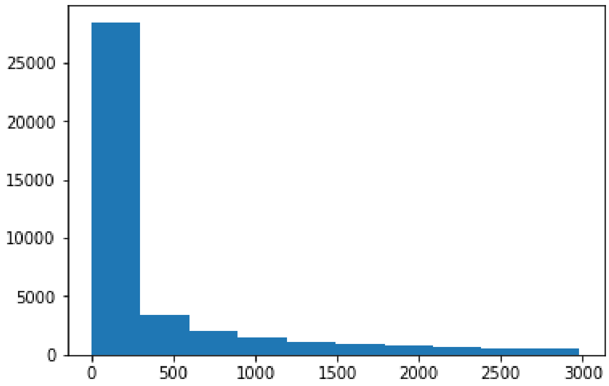

```
For PC
==================================================
Skew is: 1.9103130220711935
```



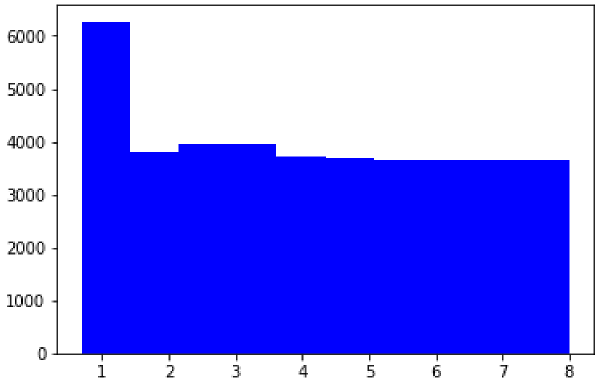
For PC after log transformation
==
Skew is: 0.09780414828042107



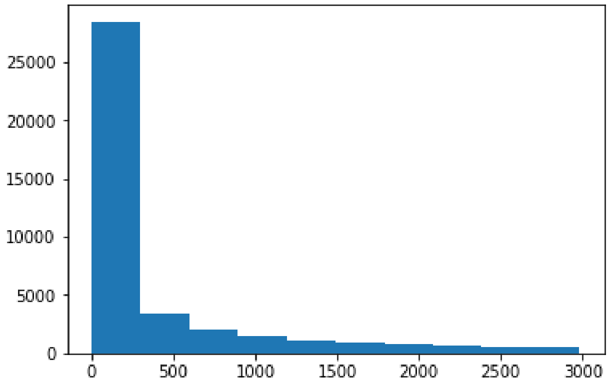
For PD
==
Skew is: 1.7983386554551613



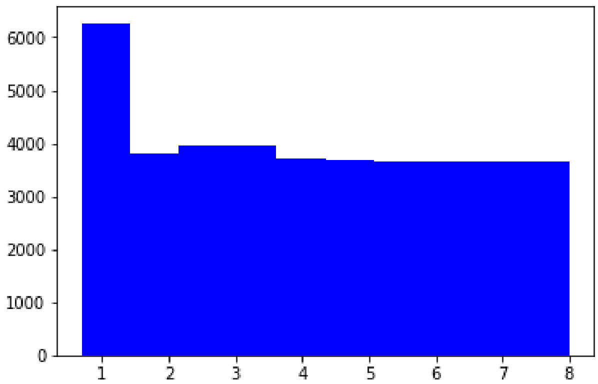
For PD after log transformation
==
Skew is: 0.09780414828042107



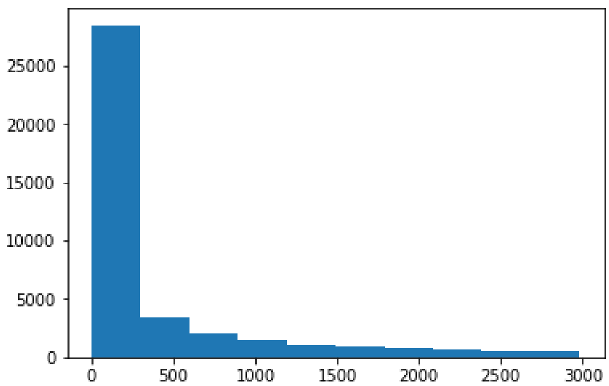
For PE
==
Skew is: 1.6983667147272488



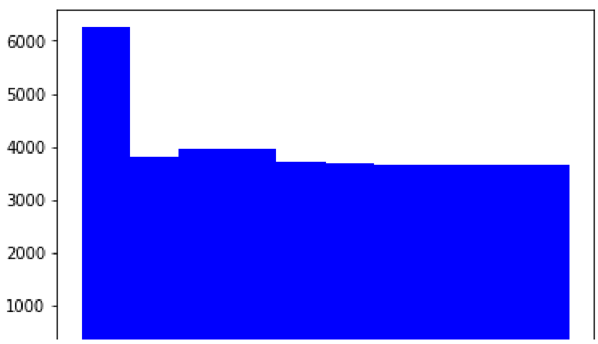

```
For PE after log transformation
==================================================
Skew is: 0.09780414828042107
```

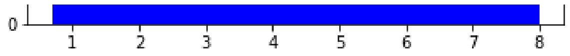



```
For PF
==================================================
Skew is: 1.60844986096591
```

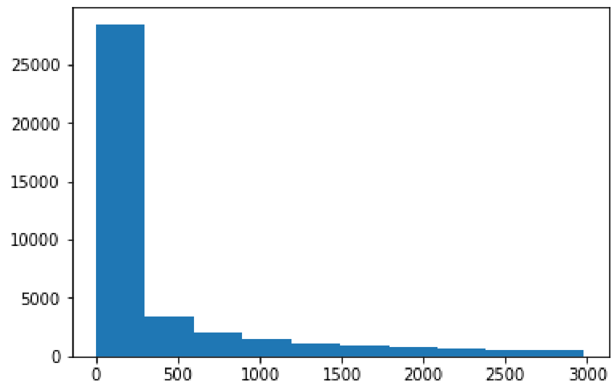



```
For PF after log transformation
==================================================
Skew is: 0.09780414828042107
```

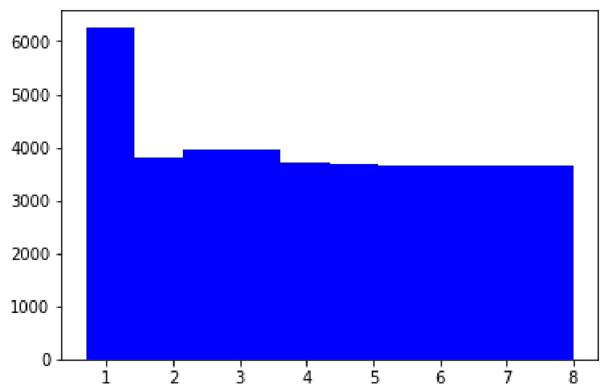





```
For PG
==================================================
Skew is: 1.5271096920926972
```




```
For PG after log transformation
==================================================
Skew is: 0.09780414828042107
```



Converting to Date Format

In [7]:

```
# how to replace elements in list python: https://stackoverflow.com/a/2582163/4084039
cols = ['Date' if x=='DateTime' else x for x in list(train_data.columns)]


#sort dataframe based on time pandas python: https://stackoverflow.com/a/49702492/4084039

train_data['Date'] = pd.to_datetime(train_data['DateTime'])
train_data.drop('DateTime', axis=1, inplace=True)
#train_data.sort_values(by=['DateTime'], inplace=True)


# how to reorder columns pandas python: https://stackoverflow.com/a/13148611/4084039
train_data = train_data[cols]


train_data.head(2)
```

Out[7]:

	ID	Date	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	WindRun	...	ISSRecpt	ArcInt	I
0	PR00001	2040-07-12 00:15:00	53.5	53.6	53.5	85	49.1	2	SSE	0.5	...	100.0	15	

	ID	2040Date	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	WindRun	...	ISSRecpt	ArcInt	I
1	PR00002	07-12 00:30:00	53.5	53.5	53.4	85	49.1	2	SSE	0.5	...	100.0	15	1

2 rows × 38 columns

In [8]:

```
train_data['Year'] = train_data['Date'].dt.year
```

In [9]:

```
train_data['Month'] = train_data['Date'].dt.month
train_data['Day'] = train_data['Date'].dt.day
```

In [10]:

```
train_data.head(2)
```

Out[10]:

	ID	Date	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	WindRun	...	PB	PC	PD	PE	P
0	PR00001	2040-07-12 00:15:00	53.5	53.6	53.5	85	49.1	2	SSE	0.5	...	1	1	1	1	1
1	PR00002	2040-07-12 00:30:00	53.5	53.5	53.4	85	49.1	2	SSE	0.5	...	1	1	1	1	1

2 rows × 41 columns

In [14]:

```
import pandas_profiling as pp
pp.ProfileReport(train_data)
```

Out[14]:

Overview

Dataset info

Number of variables	40
Number of observations	40000
Total Missing (%)	0.0%
Total size in memory	12.2 MiB
Average record size in memory	320.0 B

Variables types

Numeric	19
Categorical	2
Boolean	0
Date	1
Text (Unique)	1
Rejected	17
Unsupported	0

Warnings

- ArcInt has constant value 15 Rejected

- `CoolDD` has 29824 / 74.6% zeros Zeros
- `HeatDD` has 10258 / 25.6% zeros Zeros
- `HeatIndex` is highly correlated with `WindChill` ($\rho = 0.99688$) Rejected
- `HiSpeed` is highly correlated with `WindRun` ($\rho = 0.94861$) Rejected
- `HiTemp` is highly correlated with `TempOut` ($\rho = 0.99902$) Rejected
- `ISSRecpt` is highly skewed ($\gamma_1 = -51.663$) Skewed
- `InDew` is highly correlated with `InHum` ($\rho = 0.96322$) Rejected
- `InEMC` is highly correlated with `InDew` ($\rho = 0.93992$) Rejected
- `LowTemp` is highly correlated with `HiTemp` ($\rho = 0.9978$) Rejected
- `PB` is highly correlated with `PA` ($\rho = 0.999$) Rejected
- `PC` is highly correlated with `PB` ($\rho = 0.99919$) Rejected
- `PD` is highly correlated with `PC` ($\rho = 0.99934$) Rejected
- `PE` is highly correlated with `PD` ($\rho = 0.99945$) Rejected
- `PF` is highly correlated with `PE` ($\rho = 0.99953$) Rejected
- `PG` is highly correlated with `PF` ($\rho = 0.99958$) Rejected
- `Rain` has 39022 / 97.6% zeros Zeros
- `RainRate` is highly skewed ($\gamma_1 = 47.628$) Skewed
- `RainRate` has 39295 / 98.2% zeros Zeros
- `THWIndex` is highly correlated with `HeatIndex` ($\rho = 0.99897$) Rejected
- `WindChill` is highly correlated with `LowTemp` ($\rho = 0.99687$) Rejected
- `WindRun` is highly correlated with `WindSpeed` ($\rho = 1$) Rejected
- `WindSamp` is highly skewed ($\gamma_1 = -23.693$) Skewed
- `WindSpeed` has 10818 / 27.0% zeros Zeros
- `WindTx` has constant value 1 Rejected

Variables

~~ArcInt~~

Constant

This variable is constant and should be ignored for analysis

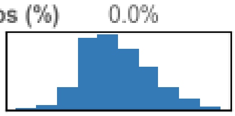
Constant value	15

Bar

Numeric

Distinct count	846
Unique (%)	2.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	30.072
Minimum	29.619
Maximum	30.534
Zeros (%)	0.0%



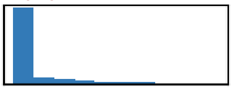
Toggle details

CoolDD

Numeric

Distinct count	391
Unique (%)	1.0%

Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0
Mean 0.026837
Minimum 0
Maximum 0.472
Zeros (%) 74.6%



Toggle details

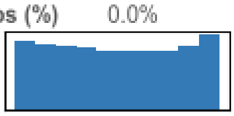
Date
Date

Distinct count 40000
Unique (%) 100.0%
Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0
Minimum 2040-07-01 00:15:00
Maximum 2042-04-12 13:45:00

Toggle details

Day
Numeric

Distinct count 31
Unique (%) 0.1%
Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0
Mean 15.586
Minimum 1
Maximum 31
Zeros (%) 0.0%

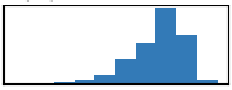


Toggle details

DewPt
Numeric

Distinct count 531
Unique (%) 1.3%
Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0
Mean 48.157

Minimum 1.2
Maximum 66.9
Zeros (%) 0.0%

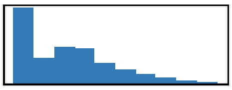


Toggle details

HeatDD

Numeric

Distinct count 354
Unique (%) 0.9%
Missing (%) 0.0%
Missing (n) 0
Infinite (%) 0.0%
Infinite (n) 0
Mean 0.094455
Minimum 0
Maximum 0.372
Zeros (%) 25.6%



Toggle details

~~HeatIndex~~

Highly correlated

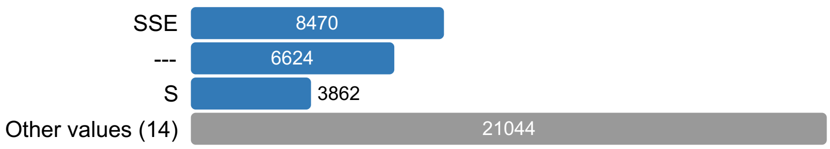
This variable is highly correlated with `WindChill` *and should be ignored for analysis*

Correlation 0.99688

HiDir

Categorical

Distinct count 17
Unique (%) 0.0%
Missing (%) 0.0%
Missing (n) 0



Toggle details

~~HiSpeed~~

Highly correlated

This variable is highly correlated with `WindRun` *and should be ignored for analysis*

Correlation 0.94861

~~HiTemp~~

Highly correlated

This variable is highly correlated with `TempOut` *and should be ignored for analysis*

Correlation 0.99902

Correlation [illegible]

ID

Categorical, Unique

First 3 values

Last 3 values

Toggle details

ISSRecpt

Numeric

Distinct count	3
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	99.998
Minimum	94.4
Maximum	100
Zeros (%)	0.0%

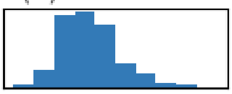
Toggle details

InAirDensity

Numeric

Distinct count	46
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	0.074569
Minimum	0.0729
Maximum	0.0774
Zeros (%)	0.0%



Toggle details

~~InDew~~

Highly correlated

This variable is highly correlated with `InHum` *and should be ignored for analysis*

Correlation 0.96322

~~InEMC~~

Highly correlated

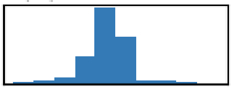
This variable is highly correlated with `InDew` *and should be ignored for analysis*

Correlation 0.93992

InHeat

Numeric

Distinct count	248
Unique (%)	0.6%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	67.407
Minimum	55.9
Maximum	81.1
Zeros (%)	0.0%

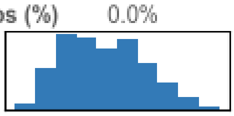


Toggle details

InHum

Numeric

Distinct count	73
Unique (%)	0.2%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	47.259
Minimum	16
Maximum	88
Zeros (%)	0.0%

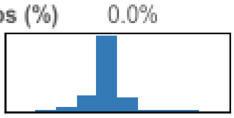


Toggle details

InTemp

Numeric

Distinct count	231
Unique (%)	0.6%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	69.171
Minimum	58.9
Maximum	82
Zeros (%)	0.0%



Toggle details

~~LowTemp~~

Highly correlated

This variable is highly correlated with `HiTemp` and should be ignored for analysis

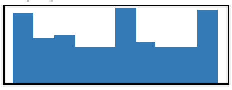
This variable is highly correlated with H1Temp and should be ignored for analysis

Correlation 0.9978

Month
Numeric

Distinct count	12
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	6.4581
Minimum	1
Maximum	12
Zeros (%)	0.0%

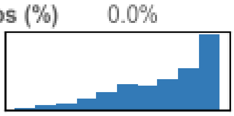


Toggle details

OutHum
Numeric

Distinct count	95
Unique (%)	0.2%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	72.916
Minimum	4
Maximum	98
Zeros (%)	0.0%

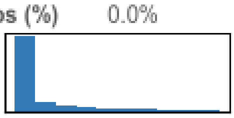


Toggle details

PA
Numeric

Distinct count	2980
Unique (%)	7.4%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	372.45
Minimum	1
Maximum	2980
Zeros (%)	0.0%



Toggle details

~~PB~~
Highly correlated

This variable is highly correlated with PA and should be ignored for analysis

Correlation 0.999

~~PC~~
Highly correlated

This variable is highly correlated with PB and should be ignored for analysis

Correlation 0.99919

~~PD~~
Highly correlated

This variable is highly correlated with PC and should be ignored for analysis

Correlation 0.99934

~~PE~~
Highly correlated

This variable is highly correlated with PD and should be ignored for analysis

Correlation 0.99945

~~PF~~
Highly correlated

This variable is highly correlated with PE and should be ignored for analysis

Correlation 0.99953

~~PG~~
Highly correlated

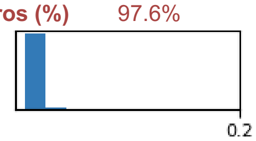
This variable is highly correlated with PF and should be ignored for analysis

Correlation 0.99958

Rain
Numeric

Distinct count	15
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	0.0005055
Minimum	0
Maximum	0.19
Zeros (%)	97.6%



Toggle details

RainRate
Numeric

Distinct count	89
Unique (%)	0.2%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	0.0039495
Minimum	0
Maximum	5.94
Zeros (%)	98.2%

Toggle details

~~THWIndex~~

Highly correlated

This variable is highly correlated with `HeatIndex` *and should be ignored for analysis*

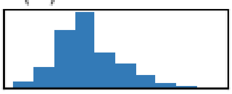
Correlation 0.99897

TempOut

Numeric

Distinct count	744
Unique (%)	1.9%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	58.509
Minimum	29.3
Maximum	110.3
Zeros (%)	0.0%



Toggle details

~~WindChill~~

Highly correlated

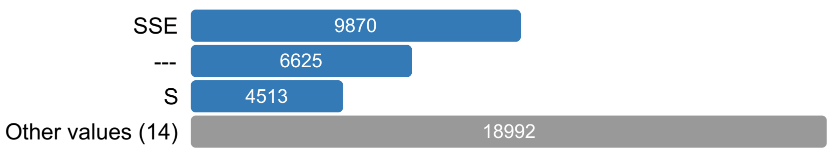
This variable is highly correlated with `LowTemp` *and should be ignored for analysis*

Correlation 0.99687

WindDir

Categorical

Distinct count	17
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0



Toggle details

~~WindRun~~

Highly correlated

This variable is highly correlated with `WindSpeed` and should be ignored for analysis

Correlation	1

WindSamp

Numeric

Distinct count	8
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	351.21
Minimum	323
Maximum	353
Zeros (%)	0.0%

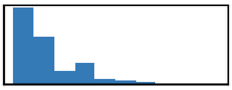
Toggle details

WindSpeed

Numeric

Distinct count	17
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	2.3487
Minimum	0
Maximum	16
Zeros (%)	27.0%



Toggle details

~~WindTx~~

Constant

This variable is constant and should be ignored for analysis

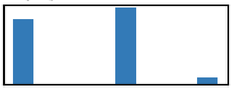
Constant value	1

Year

Numeric

Distinct count	3
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	2040.6
Minimum	2040
Maximum	2042
Zeros (%)	0.0%



Toggle details

Correlations

Sample

	ID	Date	TempOut	HiTemp	LowTemp	OutHum	DewPt	WindSpeed	WindDir	Wir
0	PR00001	2040-07-12 00:15:00	53.5	53.6	53.5	85	49.1	2	SSE	0.5
1	PR00002	2040-07-12 00:30:00	53.5	53.5	53.4	85	49.1	2	SSE	0.5
2	PR00003	2040-07-12 00:45:00	53.3	53.5	53.2	85	48.9	2	SSE	0.5
3	PR00004	2040-07-12 01:00:00	53.1	53.3	53.0	86	49.0	2	S	0.5
4	PR00005	2040-07-12 01:15:00	52.9	53.1	52.9	86	48.8	2	S	0.5

Inferences:

1. 'THWIndex' is highly correlated with HeatIndex (ρ = 0.99897) , so we can reject any 1 row
2. 'WindChill' is highly correlated with LowTemp (ρ = 0.99687)
3. 'HiTemp' is highly correlated with TempOut (ρ = 0.99902)
4. 'AcrInt' has constant value 15. So, we can reject it
5. WindTx has constant value 1
6. Rain has 39022 / 97.6% zeros
7. RainRate has 39295 / 98.2% zeros

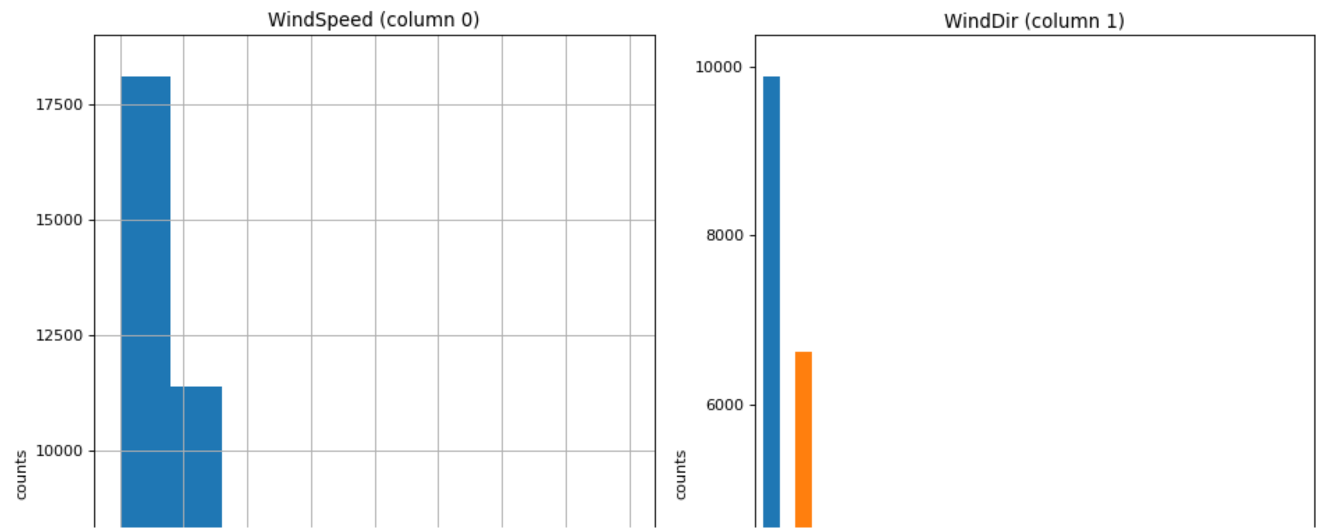
Checking Distribution Graphs

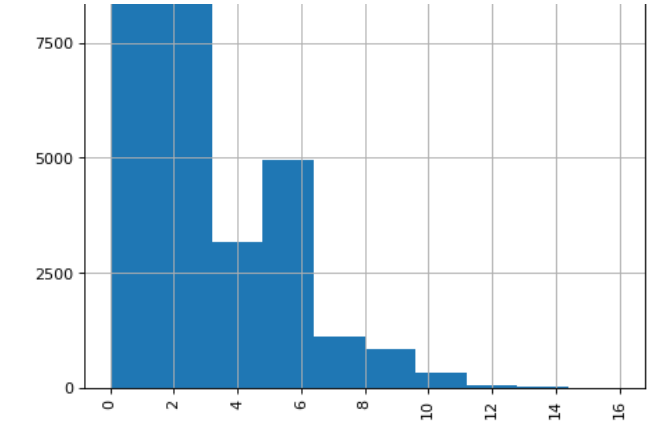
In [13]:

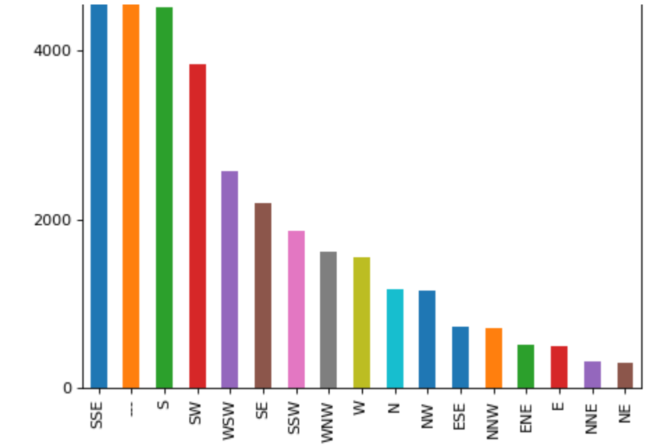
```
# Distribution graphs (histogram/bar graph) of column data
def plotPerColumnDistribution(df, nGraphShown, nGraphPerRow):
    nunique = df.nunique()
    df = df[[col for col in df if nunique[col] > 1 and nunique[col] < 50]] # For displaying purposes, pick columns that have between 1 and 50 unique values
    nRow, nCol = df.shape
    columnNames = list(df)
    nGraphRow = (nCol + nGraphPerRow - 1) / nGraphPerRow
    plt.figure(num = None, figsize = (6 * nGraphPerRow, 8 * nGraphRow), dpi = 80, facecolor = 'w', edgecolor = 'k')
    for i in range(min(nCol, nGraphShown)):
        plt.subplot(nGraphRow, nGraphPerRow, i + 1)
        columnDf = df.iloc[:, i]
        if (not np.issubdtype(type(columnDf.iloc[0]), np.number)):
            valueCounts = columnDf.value_counts()
            valueCounts.plot.bar()
        else:
            columnDf.hist()
        plt.ylabel('counts')
        plt.xticks(rotation = 90)
        plt.title(f'{columnNames[i]} (column {i})')
    plt.tight_layout(pad = 1.0, w_pad = 1.0, h_pad = 1.0)
    plt.show()
```

In [14]:

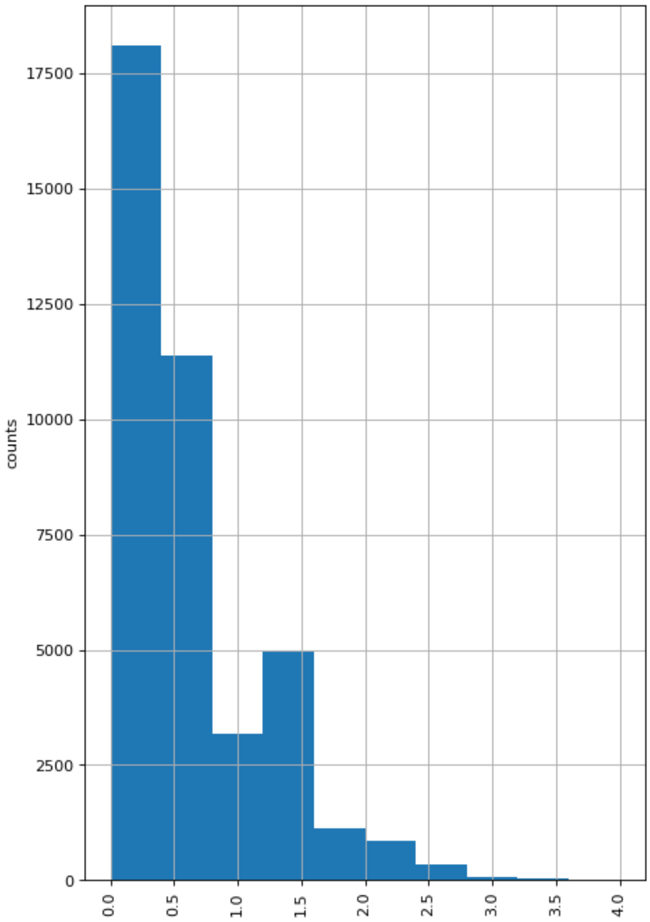
```
plotPerColumnDistribution(train_data, 10, 2)
```



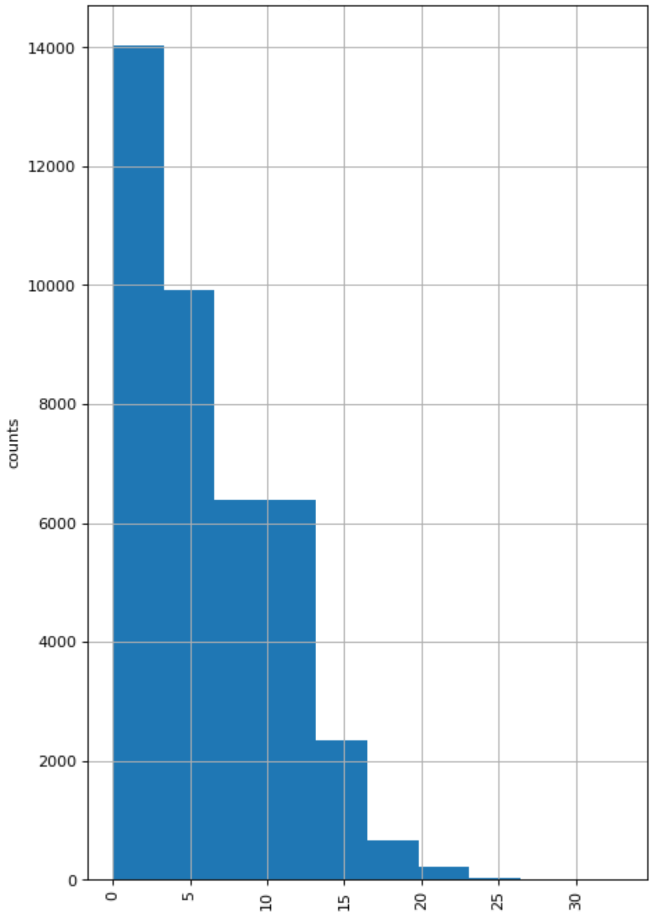




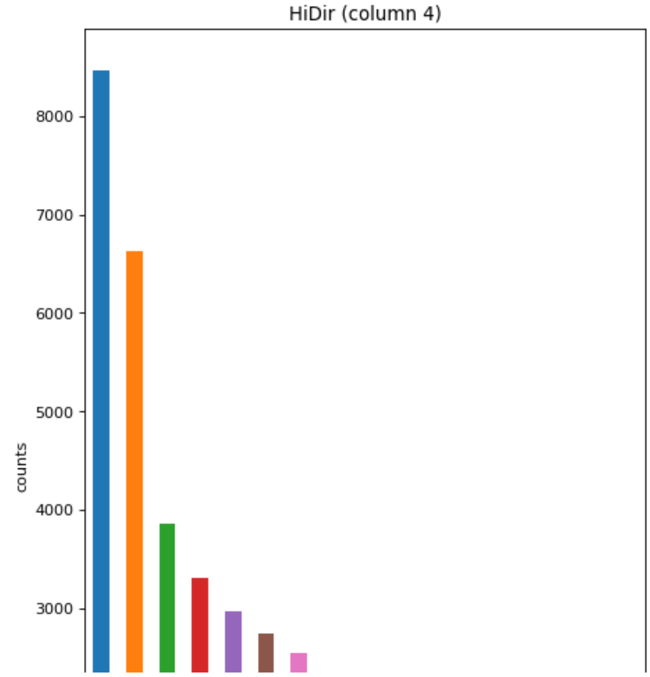
WindRun (column 2)



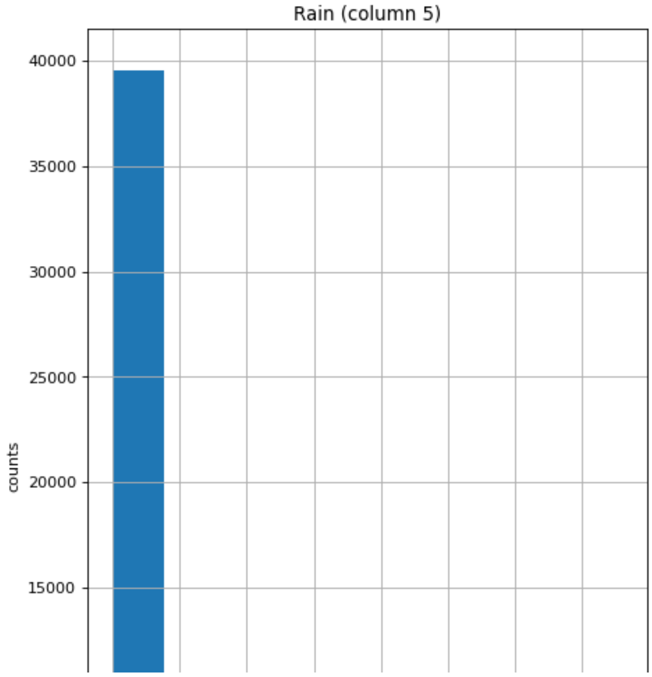
HiSpeed (column 3)

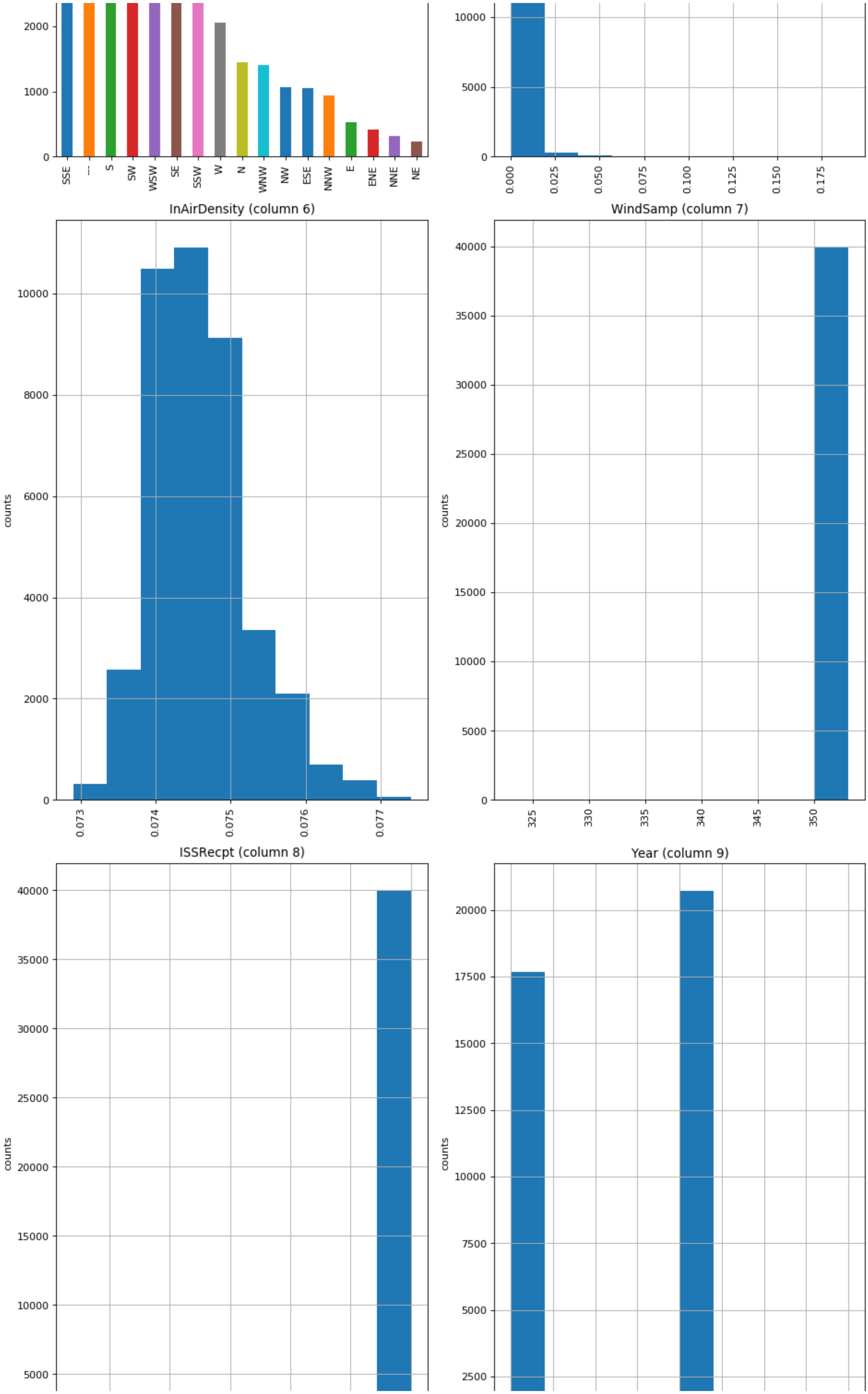


HiDir (column 4)

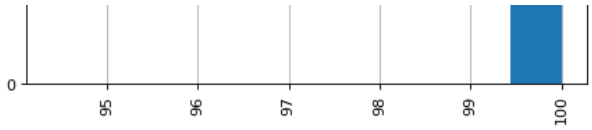


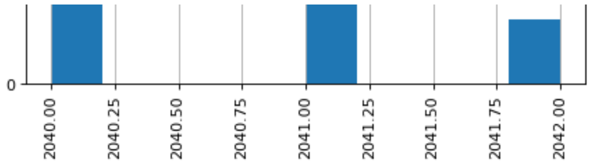
Rain (column 5)




InAirDensity (column 6)
WindSamp (column 7)
ISSRecpt (column 8)
Year (column 9)
counts
SSE

S
SW
WSW
SE
SSW
W
N
WNW
NW
ESE
NNW
E
ENE
NNE
NE





In [11]:

```
train_data.columns
```

Out[11]:

```
Index(['ID', 'Date', 'TempOut', 'HiTemp', 'LowTemp', 'OutHum', 'DewPt',
       'WindSpeed', 'WindDir', 'WindRun', 'HiSpeed', 'HiDir', 'WindChill',
       'HeatIndex', 'THWIndex', 'Bar', 'Rain', 'RainRate', 'HeatDD', 'CoolDD',
       'InTemp', 'InHum', 'InDew', 'InHeat', 'InEMC', 'InAirDensity',
       'WindSamp', 'WindTx', 'ISSRecpt', 'ArcInt', 'PA', 'PB', 'PC', 'PD',
       'PE', 'PF', 'PG', 'Skewed_PA', 'Year', 'Month', 'Day'],
      dtype='object')
```

Categorical Features

In [12]:

```
train_data.select_dtypes(include=['O']).columns.values
```

Out[12]:

```
array(['ID', 'WindDir', 'HiDir'], dtype=object)
```

Numerical Features

In [20]:

```
numerical_features = train_data.select_dtypes(include=[np.number])
numerical_features.dtypes
```

Out[20]:

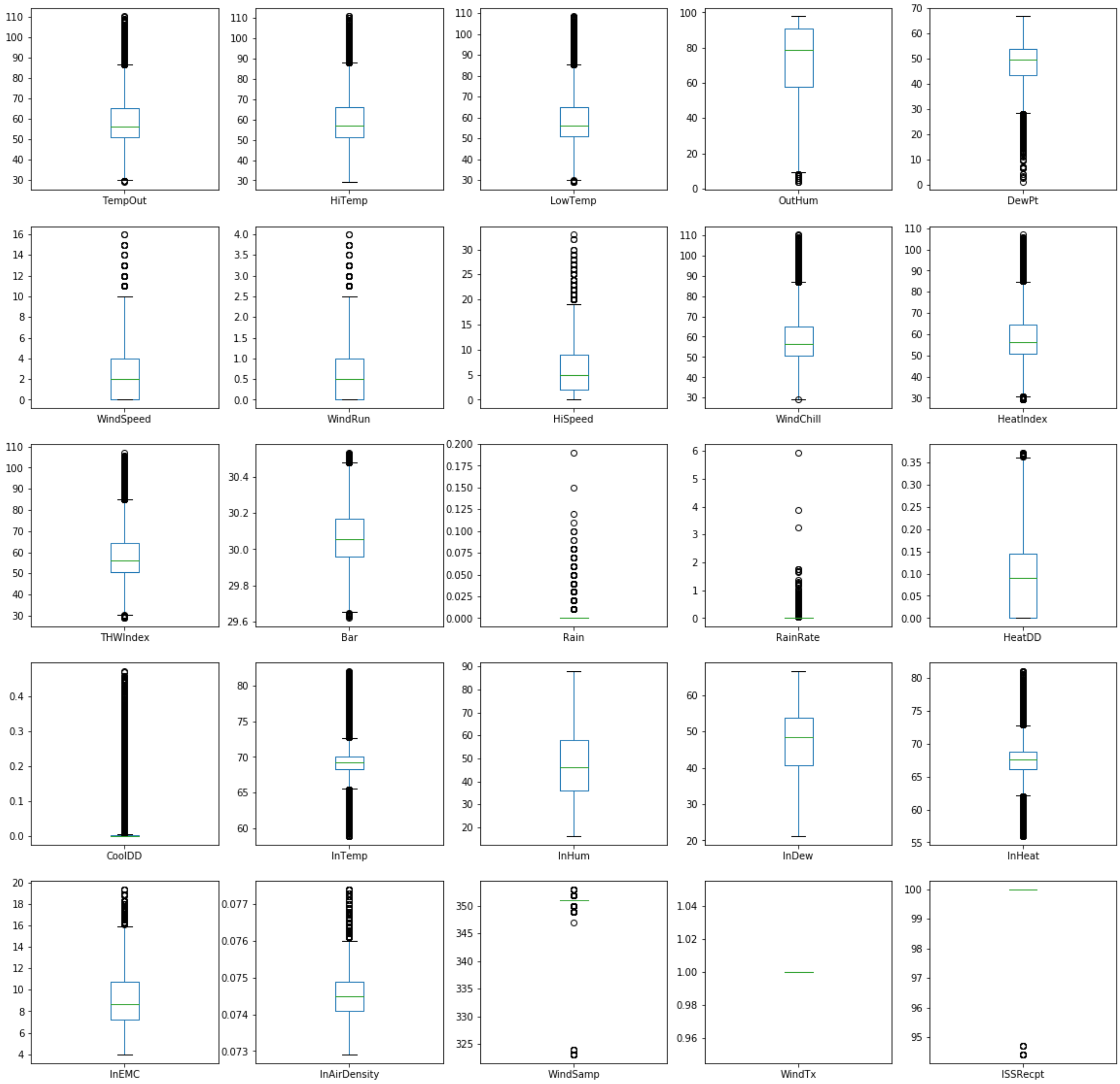
```
TempOut         float64
HiTemp          float64
LowTemp         float64
OutHum            int64
DewPt           float64
WindSpeed         int64
WindRun         float64
HiSpeed           int64
WindChill       float64
HeatIndex       float64
THWIndex        float64
Bar             float64
Rain            float64
RainRate        float64
HeatDD          float64
CoolDD          float64
InTemp          float64
InHum             int64
InDew           float64
InHeat          float64
InEMC           float64
InAirDensity    float64
WindSamp          int64
WindTx            int64
ISSRecpt        float64
ArcInt            int64
PA                int64
PB                int64
PC                int64
PD                int64
PE                int64
```

```
PF                 int64
PG                 int64
Skewed_PA        float64
Year               int64
Month              int64
Day                int64
dtype: object
```

BarPlots

In [6]:

```python
fig, axarr = plt.subplots(5,5, figsize=(20, 20))
cols = ['TempOut', 'HiTemp', 'LowTemp', 'OutHum', 'DewPt',
       'WindSpeed', 'WindRun', 'HiSpeed', 'WindChill',
       'HeatIndex', 'THWIndex', 'Bar', 'Rain', 'RainRate', 'HeatDD', 'CoolDD',
       'InTemp', 'InHum', 'InDew', 'InHeat', 'InEMC', 'InAirDensity',
       'WindSamp', 'WindTx', 'ISSRecpt']
k = 0
for i in range(0,5):
    for j in range(0,5):
        train_data[cols[k]].plot.box(ax=axarr[i][j])
        k = k + 1
```



Let's view some features

In [20]:

```
train_data['WindDir'].value_counts()
```

Out[20]:

```
SSE    9870
---    6625
S      4513
SW     3842
WSW    2567
SE     2188
SSW    1860
WNW    1609
W      1549
N      1172
NW     1148
ESE     724
NNW     714
ENE     508
E       494
NNE     320
NE      297
Name: WindDir, dtype: int64
```

> 6625 values are missing in this feature

In [30]:

```
train_data['Rain'].value_counts()
```

Out[30]:

```
0.00    39022
0.01      545
0.02      188
0.03       95
0.04       65
0.05       37
0.06       18
0.08       11
0.07       10
0.10        3
0.09        2
0.12        1
0.15        1
0.19        1
0.11        1
Name: Rain, dtype: int64
```

> Maximum rows are zeros

In [32]:

```
train_data['RainRate'].value_counts().head(5)
```

Out[32]:

```
0.00    39295
0.04       82
0.06       51
0.07       41
0.05       41
Name: RainRate, dtype: int64
```

> Maximum rows are zeros

Correlation with Output Features:

In [57]:

```
zero_feat=[ 'TempOut', 'HiTemp', 'LowTemp', 'OutHum', 'DewPt',
       'WindSpeed', 'WindRun', 'HiSpeed', 'WindChill',
       'HeatIndex', 'THWIndex', 'Bar', 'Rain', 'RainRate', 'HeatDD', 'CoolDD',
       'InTemp', 'InHum', 'InDew', 'InHeat', 'InEMC', 'InAirDensity',
       'WindSamp', 'WindTx', 'ISSRecpt', 'ArcInt', 'Year', 'Month', 'Day']
out_feat=['PA','PB','PC','PD','PE','PF','PG']
```

In [58]:

```
from scipy.stats import pearsonr
for i in zero_feat:
    for j in out_feat:
        corr, _ = pearsonr(train_data[i], train_data[j])
        print('Pearsons correlation of {0} with {1} is {2}'.format(i,j,corr))
    print('-'*60)
```

```
Pearsons correlation of TempOut with PA is 0.1910002858025576
Pearsons correlation of TempOut with PB is 0.18772556227103043
Pearsons correlation of TempOut with PC is 0.1833779205680189
Pearsons correlation of TempOut with PD is 0.17826373141140262
Pearsons correlation of TempOut with PE is 0.17269184079347027
Pearsons correlation of TempOut with PF is 0.16701808535621332
Pearsons correlation of TempOut with PG is 0.16099174981231795
------------------------------------------------------------
Pearsons correlation of HiTemp with PA is 0.18918340806814024
Pearsons correlation of HiTemp with PB is 0.18594789469922177
Pearsons correlation of HiTemp with PC is 0.1816503753686059
Pearsons correlation of HiTemp with PD is 0.17659474022835622
Pearsons correlation of HiTemp with PE is 0.17108522044836355
Pearsons correlation of HiTemp with PF is 0.16547264128712647
Pearsons correlation of HiTemp with PG is 0.1595200676181685
------------------------------------------------------------
Pearsons correlation of LowTemp with PA is 0.1927218355997909
Pearsons correlation of LowTemp with PB is 0.18944478564716483
Pearsons correlation of LowTemp with PC is 0.18507912331743667
Pearsons correlation of LowTemp with PD is 0.17993485518558577
Pearsons correlation of LowTemp with PE is 0.17433018011105367
Pearsons correlation of LowTemp with PF is 0.16860856990606835
Pearsons correlation of LowTemp with PG is 0.1625346920347483
------------------------------------------------------------
Pearsons correlation of OutHum with PA is -0.030012234753616036
Pearsons correlation of OutHum with PB is -0.02877201248575463
Pearsons correlation of OutHum with PC is -0.02721669451236607
Pearsons correlation of OutHum with PD is -0.02556142471311104
Pearsons correlation of OutHum with PE is -0.023829752512048618
Pearsons correlation of OutHum with PF is -0.021922945867353983
Pearsons correlation of OutHum with PG is -0.02006048098244488
------------------------------------------------------------
Pearsons correlation of DewPt with PA is 0.27381544240158606
Pearsons correlation of DewPt with PB is 0.27207032443469126
Pearsons correlation of DewPt with PC is 0.26894305746494024
Pearsons correlation of DewPt with PD is 0.26462955923735054
Pearsons correlation of DewPt with PE is 0.2595842899281463
Pearsons correlation of DewPt with PF is 0.25439287336842775
Pearsons correlation of DewPt with PG is 0.24848267405885047
------------------------------------------------------------
Pearsons correlation of WindSpeed with PA is 0.16852172128185688
Pearsons correlation of WindSpeed with PB is 0.17472569269921417
Pearsons correlation of WindSpeed with PC is 0.17982218858402138
Pearsons correlation of WindSpeed with PD is 0.18392278894809455
Pearsons correlation of WindSpeed with PE is 0.1873219929494478
Pearsons correlation of WindSpeed with PF is 0.1897687629977747
Pearsons correlation of WindSpeed with PG is 0.1916203803971352
```

```
------------------------------------------------------------
Pearsons correlation of WindRun with PA is 0.16852172128185688
Pearsons correlation of WindRun with PB is 0.17472569269921417
Pearsons correlation of WindRun with PC is 0.17982218858402138
Pearsons correlation of WindRun with PD is 0.18392278894809455
Pearsons correlation of WindRun with PE is 0.1873219929494478
Pearsons correlation of WindRun with PF is 0.1897687629977747
Pearsons correlation of WindRun with PG is 0.1916203803971352
------------------------------------------------------------
Pearsons correlation of HiSpeed with PA is 0.18845232759624414
Pearsons correlation of HiSpeed with PB is 0.1944243541930638
Pearsons correlation of HiSpeed with PC is 0.199268297801943
Pearsons correlation of HiSpeed with PD is 0.20307339156236615
Pearsons correlation of HiSpeed with PE is 0.20615006597786067
Pearsons correlation of HiSpeed with PF is 0.2083205658435073
Pearsons correlation of HiSpeed with PG is 0.2098519457576899
------------------------------------------------------------
Pearsons correlation of WindChill with PA is 0.19252946750714536
Pearsons correlation of WindChill with PB is 0.18907425409310927
Pearsons correlation of WindChill with PC is 0.1845447537386401
Pearsons correlation of WindChill with PD is 0.17925016653529138
Pearsons correlation of WindChill with PE is 0.17350481918318794
Pearsons correlation of WindChill with PF is 0.1676704356816646
Pearsons correlation of WindChill with PG is 0.16148979889565493
------------------------------------------------------------
Pearsons correlation of HeatIndex with PA is 0.20292118361471348
Pearsons correlation of HeatIndex with PB is 0.1995355057195256
Pearsons correlation of HeatIndex with PC is 0.19503247362695195
Pearsons correlation of HeatIndex with PD is 0.18971809841047102
Pearsons correlation of HeatIndex with PE is 0.18391806243989695
Pearsons correlation of HeatIndex with PF is 0.178023427840309
Pearsons correlation of HeatIndex with PG is 0.1717535788906328
------------------------------------------------------------
Pearsons correlation of THWIndex with PA is 0.20432621072449905
Pearsons correlation of THWIndex with PB is 0.2007583078733473
Pearsons correlation of THWIndex with PC is 0.19607242887518211
Pearsons correlation of THWIndex with PD is 0.1905773156488461
Pearsons correlation of THWIndex with PE is 0.18460400756799886
Pearsons correlation of THWIndex with PF is 0.17854923307397585
Pearsons correlation of THWIndex with PG is 0.1721260037716632
------------------------------------------------------------
Pearsons correlation of Bar with PA is -0.17521848843177595
Pearsons correlation of Bar with PB is -0.1748313801265145
Pearsons correlation of Bar with PC is -0.1730044549429335
Pearsons correlation of Bar with PD is -0.17007231750163576
Pearsons correlation of Bar with PE is -0.16641313508650762
Pearsons correlation of Bar with PF is -0.16200681468941622
Pearsons correlation of Bar with PG is -0.157801432732328
------------------------------------------------------------
Pearsons correlation of Rain with PA is -0.04529380916583757
Pearsons correlation of Rain with PB is -0.04423212315786063
Pearsons correlation of Rain with PC is -0.04290220829446859
Pearsons correlation of Rain with PD is -0.04148394389640452
Pearsons correlation of Rain with PE is -0.03999027591774844
Pearsons correlation of Rain with PF is -0.03821913321079623
Pearsons correlation of Rain with PG is -0.0367590459219622
------------------------------------------------------------
Pearsons correlation of RainRate with PA is -0.02660439271941859
Pearsons correlation of RainRate with PB is -0.026097122566562846
Pearsons correlation of RainRate with PC is -0.025434354040211524
Pearsons correlation of RainRate with PD is -0.02474715726418719
Pearsons correlation of RainRate with PE is -0.023995432799517484
Pearsons correlation of RainRate with PF is -0.02309081258746043
Pearsons correlation of RainRate with PG is -0.0223535591826576
------------------------------------------------------------
Pearsons correlation of HeatDD with PA is -0.20682550678172687
Pearsons correlation of HeatDD with PB is -0.20516461054945415
Pearsons correlation of HeatDD with PC is -0.2022184369685048
Pearsons correlation of HeatDD with PD is -0.19826848295942517
Pearsons correlation of HeatDD with PE is -0.19369246041209423
Pearsons correlation of HeatDD with PF is -0.18888276856382277
Pearsons correlation of HeatDD with PG is -0.183397105373803
------------------------------------------------------------
Pearsons correlation of CoolDD with PA is 0.10877529701580482
Pearsons correlation of CoolDD with PB is 0.10430744349992314
Pearsons correlation of CoolDD with PC is 0.0993991810908948
Pearsons correlation of CoolDD with PD is 0.09429442632229229
```

```
Pearsons correlation of CoolDD with PE is 0.08910898357838196
Pearsons correlation of CoolDD with PF is 0.08403634146725539
Pearsons correlation of CoolDD with PG is 0.07916837360277684
------------------------------------------------------------
Pearsons correlation of InTemp with PA is 0.08673296488854398
Pearsons correlation of InTemp with PB is 0.08828129128081072
Pearsons correlation of InTemp with PC is 0.08933688043726996
Pearsons correlation of InTemp with PD is 0.08984366372304395
Pearsons correlation of InTemp with PE is 0.08996685423139901
Pearsons correlation of InTemp with PF is 0.08988941816013543
Pearsons correlation of InTemp with PG is 0.0893538142098953
------------------------------------------------------------
Pearsons correlation of InHum with PA is 0.1990532333263121
Pearsons correlation of InHum with PB is 0.19538819317256945
Pearsons correlation of InHum with PC is 0.1908711351104089
Pearsons correlation of InHum with PD is 0.18564682072351105
Pearsons correlation of InHum with PE is 0.18002540545783818
Pearsons correlation of InHum with PF is 0.17438936248460338
Pearsons correlation of InHum with PG is 0.16849545960718745
------------------------------------------------------------
Pearsons correlation of InDew with PA is 0.20784262251889057
Pearsons correlation of InDew with PB is 0.20588716994553413
Pearsons correlation of InDew with PC is 0.20295424787507924
Pearsons correlation of InDew with PD is 0.19916583981182134
Pearsons correlation of InDew with PE is 0.19484325154147278
Pearsons correlation of InDew with PF is 0.19038691408624242
Pearsons correlation of InDew with PG is 0.1855506771628572
------------------------------------------------------------
Pearsons correlation of InHeat with PA is 0.1734282888596046
Pearsons correlation of InHeat with PB is 0.1732871769414427
Pearsons correlation of InHeat with PC is 0.17229843546830725
Pearsons correlation of InHeat with PD is 0.17048405119396276
Pearsons correlation of InHeat with PE is 0.16813619680601616
Pearsons correlation of InHeat with PF is 0.1655950384512045
Pearsons correlation of InHeat with PG is 0.1625261338454233
------------------------------------------------------------
Pearsons correlation of InEMC with PA is 0.18589701458103222
Pearsons correlation of InEMC with PB is 0.1812595014596652
Pearsons correlation of InEMC with PC is 0.17586009953938897
Pearsons correlation of InEMC with PD is 0.16983294775209984
Pearsons correlation of InEMC with PE is 0.16349383814460505
Pearsons correlation of InEMC with PF is 0.15721139510787574
Pearsons correlation of InEMC with PG is 0.1507657093925662
------------------------------------------------------------
Pearsons correlation of InAirDensity with PA is -0.2207473713805517
Pearsons correlation of InAirDensity with PB is -0.2198994956745529
Pearsons correlation of InAirDensity with PC is -0.21765242160098808
Pearsons correlation of InAirDensity with PD is -0.21422269632731
Pearsons correlation of InAirDensity with PE is -0.21003078996415506
Pearsons correlation of InAirDensity with PF is -0.20530611591306755
Pearsons correlation of InAirDensity with PG is -0.2003625915456893
------------------------------------------------------------
Pearsons correlation of WindSamp with PA is -0.00015148350449021309
Pearsons correlation of WindSamp with PB is -8.462957406608333e-05
Pearsons correlation of WindSamp with PC is -1.4978061348912962e-05
Pearsons correlation of WindSamp with PD is 1.2126620369934841e-05
Pearsons correlation of WindSamp with PE is 5.508132430222083e-05
Pearsons correlation of WindSamp with PF is 0.00011379113048262728
Pearsons correlation of WindSamp with PG is 0.00020580560555023247
------------------------------------------------------------
Pearsons correlation of WindTx with PA is nan
Pearsons correlation of WindTx with PB is nan
Pearsons correlation of WindTx with PC is nan
Pearsons correlation of WindTx with PD is nan
Pearsons correlation of WindTx with PE is nan
Pearsons correlation of WindTx with PF is nan
Pearsons correlation of WindTx with PG is nan
------------------------------------------------------------
Pearsons correlation of ISSRecpt with PA is 0.00022444361608942105
Pearsons correlation of ISSRecpt with PB is 0.00026696296235558556
Pearsons correlation of ISSRecpt with PC is 0.0003246224902840238
Pearsons correlation of ISSRecpt with PD is 0.0003467297262089046
Pearsons correlation of ISSRecpt with PE is 0.00041327595311513424
Pearsons correlation of ISSRecpt with PF is 0.0004714694628948367
Pearsons correlation of ISSRecpt with PG is 0.0005770977449711354
------------------------------------------------------------
Pearsons correlation of ArcInt with PA is nan
```

```
Pearsons correlation of ArcInt with PB is nan
Pearsons correlation of ArcInt with PC is nan
Pearsons correlation of ArcInt with PD is nan
Pearsons correlation of ArcInt with PE is nan
Pearsons correlation of ArcInt with PF is nan
Pearsons correlation of ArcInt with PG is nan
------------------------------------------------------------
Pearsons correlation of Year with PA is 0.4448487494003768
Pearsons correlation of Year with PB is 0.46679586454791056
Pearsons correlation of Year with PC is 0.4869534918309897
Pearsons correlation of Year with PD is 0.5053908539368497
Pearsons correlation of Year with PE is 0.5221958635079577
Pearsons correlation of Year with PF is 0.5376610584736801
Pearsons correlation of Year with PG is 0.5515328019175089
------------------------------------------------------------
Pearsons correlation of Month with PA is -0.07025390230350499
Pearsons correlation of Month with PB is -0.08941605259271808
Pearsons correlation of Month with PC is -0.10791068156248412
Pearsons correlation of Month with PD is -0.12554666478199616
Pearsons correlation of Month with PE is -0.14214022811297822
Pearsons correlation of Month with PF is -0.1578749358923123
Pearsons correlation of Month with PG is -0.17218748373718223
------------------------------------------------------------
Pearsons correlation of Day with PA is 0.032350992133947425
Pearsons correlation of Day with PB is 0.02954100797740828
Pearsons correlation of Day with PC is 0.02688362999304832
Pearsons correlation of Day with PD is 0.02437313717582151
Pearsons correlation of Day with PE is 0.022183024708924314
Pearsons correlation of Day with PF is 0.020006056400236137
Pearsons correlation of Day with PG is 0.01770981303523274
------------------------------------------------------------
```

Inference:

1. We can see that columns which has mostly zeros has very low correlation(i.e. in negative or very small positive values) with Output Features
2. Correlation = **'nan'** , for Features having constant values like 'ArcInt' and 'WindTx'

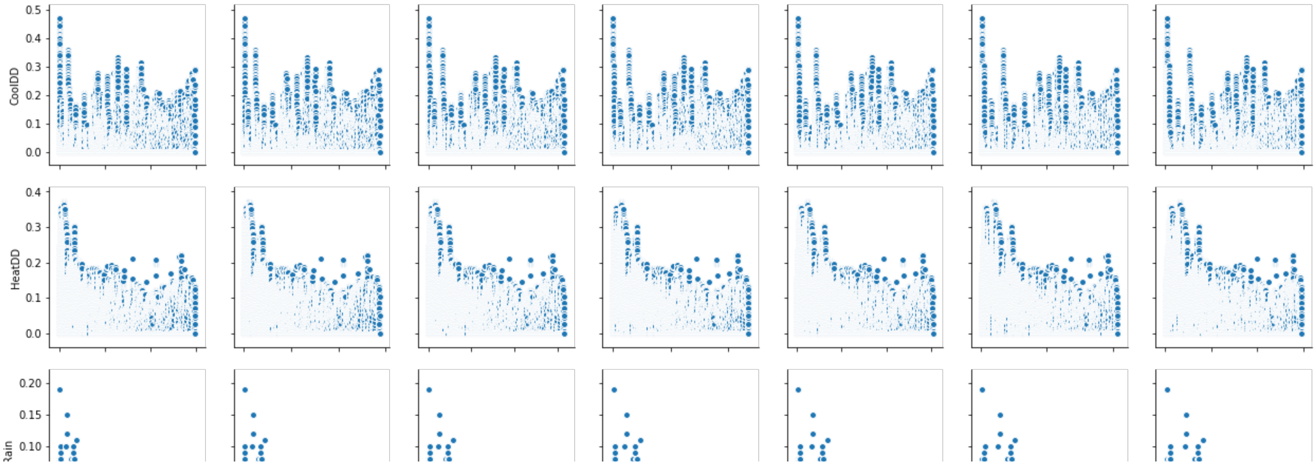
Summary:

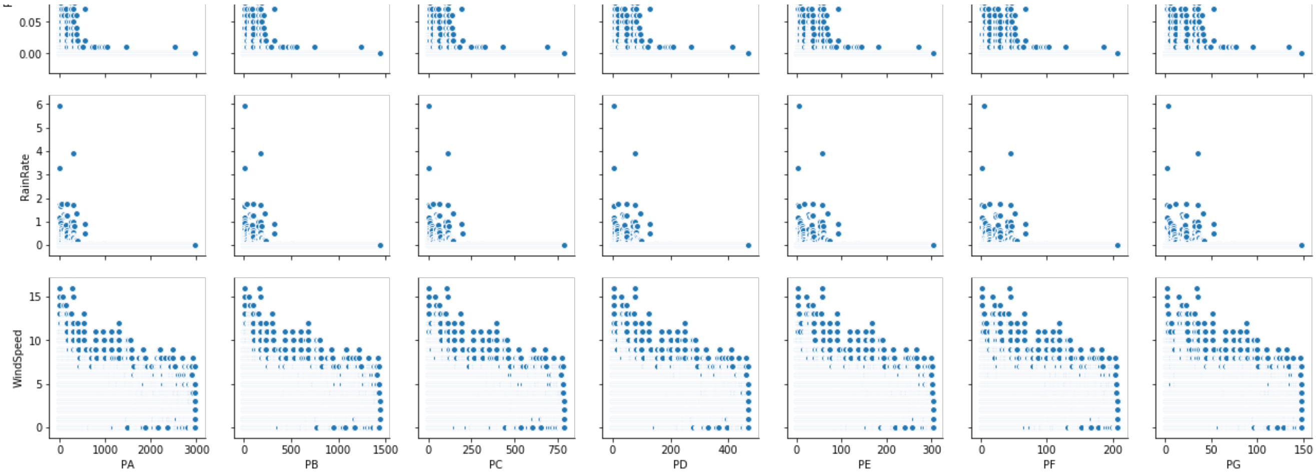
1. The Features should be less correlated with other features , but should be highly related with Output features. **So, the features which are less related or those who have very less correlation coefficients with respect to Output Features should be carefully handled.**

Pair Plots for features containing mostly zeros

In [14]:

```
sns.pairplot(train_data,
x_vars=['PA','PB','PC','PD','PE','PF','PG'],y_vars=['CoolDD','HeatDD','Rain','RainRate','WindSpeed'
])
plt.show()
```





Inference:

1. Most of the point lies at **x-axis=zero**

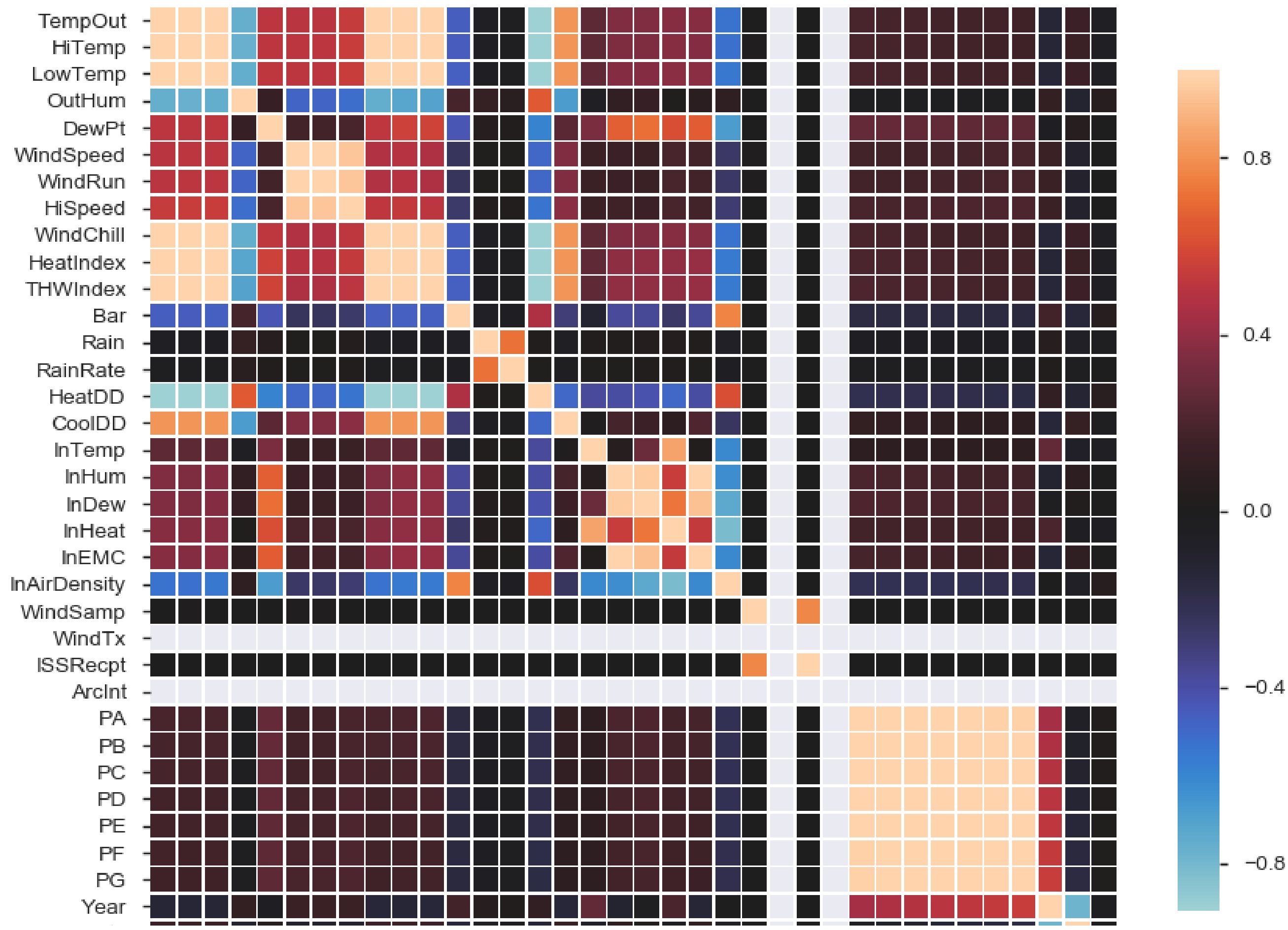
Let's check out the Correlation Matrix

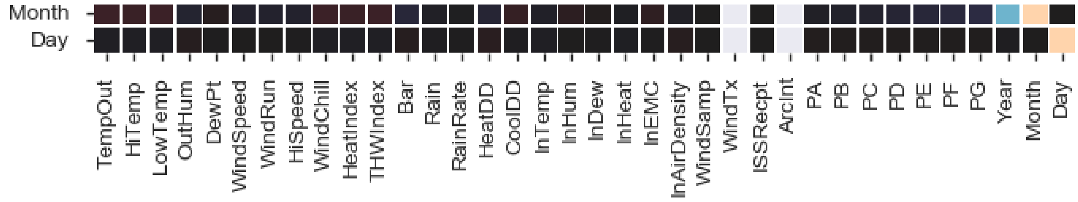
In [45]:

```python
# www.kaggle.com
def correlation_heatmap(train_data):
    correlations = train_data.corr()

    fig, ax = plt.subplots(figsize=(10,10))
    sns.heatmap(correlations, vmax=1.0, center=0, fmt='.2f',
                square=True, linewidths=.5,  cbar_kws={"shrink": .70})
    plt.show();

correlation_heatmap(train_data)
```





Inferences:

1. We can see high correlations between TWHIndex, WindChill and HeatIndex.
2. We can see high correlations between TempOut, HighTemp and LowTemp.
3. We can see high correlations between WindSpeed, WindRun and HiSpeed.

Conclusion:

> Correlations are very useful in many applications, especially when conducting regression analysis. However, it should not be mixed with causality and misinterpreted in any way. You should also always check the correlation between different variables in your dataset and gather some insights as part of your exploration and analysis.

How Can I Deal With This Problem?

> There are multiple ways to deal with this problem. The easiest way is to delete or eliminate one of the perfectly correlated features

Multicollinearity

> If your dataset has perfectly positive or negative attributes then there is a high chance that the performance of the model will be impacted by a problem called—"Multicollinearity". Multicollinearity happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy. This can lead to skewed or misleading results.