Data Science Project Number 1

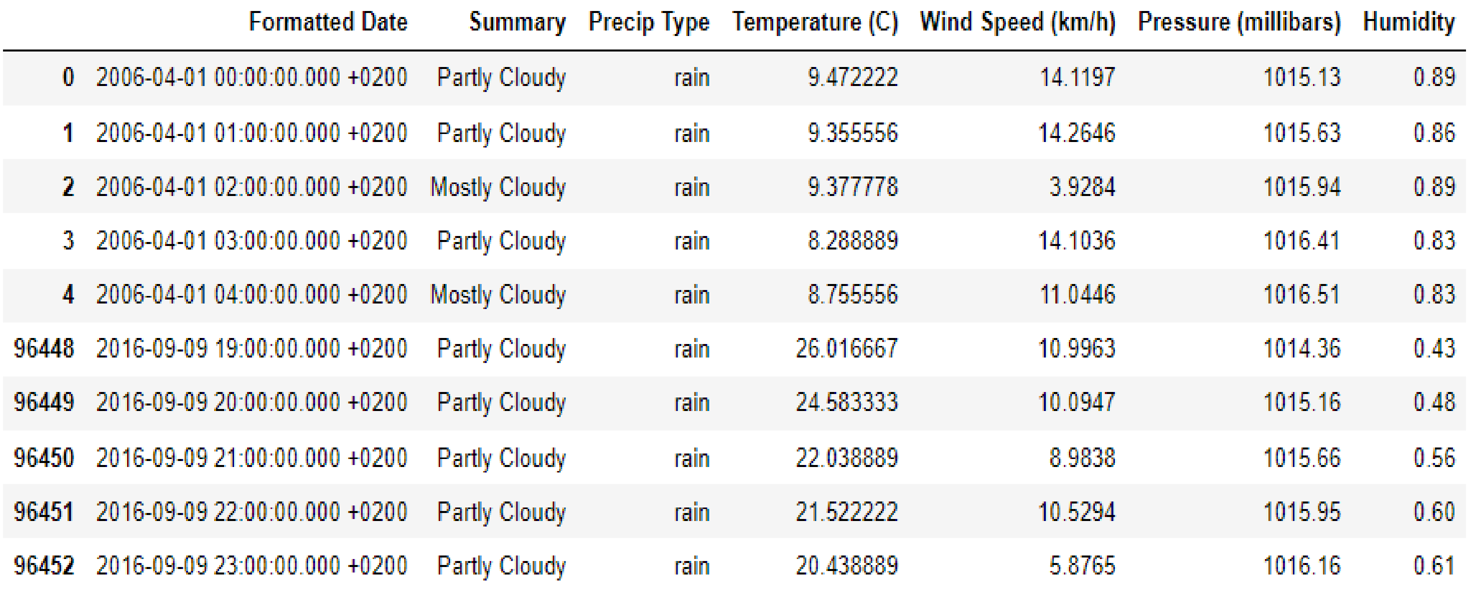
Mohammad Bagher Soroush

Project Master:
Mohammad Reza Momeni

Summer 2022

Introduction: the dataset we have is about England Weather and our goal is to study air humidity and factors that affecting on it.

Here's a quick look of our dataset (the first five and the last five lines):

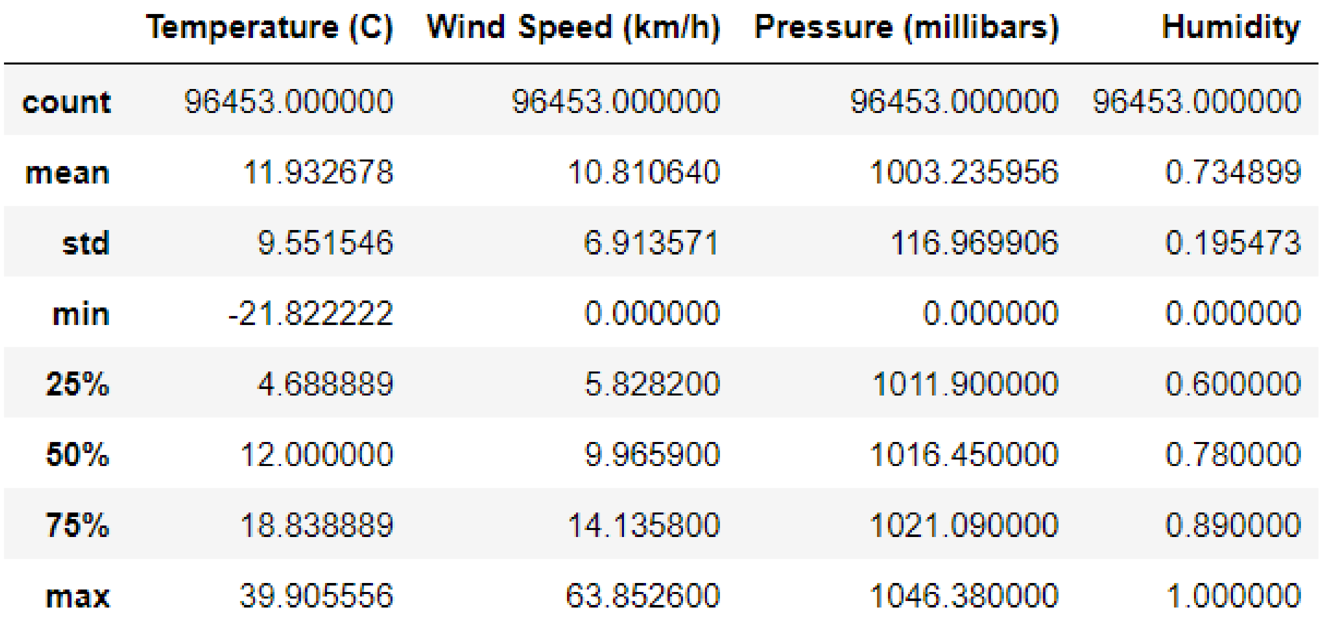


	Formatted Date	Summary	Precip Type	Temperature (C)	Wind Speed (km/h)	Pressure (millibars)	Humidity
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	14.1197	1015.13	0.89
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	14.2646	1015.63	0.86
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	3.9284	1015.94	0.89
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	14.1036	1016.41	0.83
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	11.0446	1016.51	0.83
96448	2016-09-09 19:00:00.000 +0200	Partly Cloudy	rain	26.016667	10.9963	1014.36	0.43
96449	2016-09-09 20:00:00.000 +0200	Partly Cloudy	rain	24.583333	10.0947	1015.16	0.48
96450	2016-09-09 21:00:00.000 +0200	Partly Cloudy	rain	22.038889	8.9838	1015.66	0.56
96451	2016-09-09 22:00:00.000 +0200	Partly Cloudy	rain	21.522222	10.5294	1015.95	0.60
96452	2016-09-09 23:00:00.000 +0200	Partly Cloudy	rain	20.438889	5.8765	1016.16	0.61

As you can see our dataset has 96453 rows and 7 columns of information of England weather over 10 years.

The columns include: date (with hours), summery of weather, type of precipitation, temperatures (in Celsius), wind speed (in kilometers per hour), pressure (in millibars) and our target in this project, air humidity (in percent)

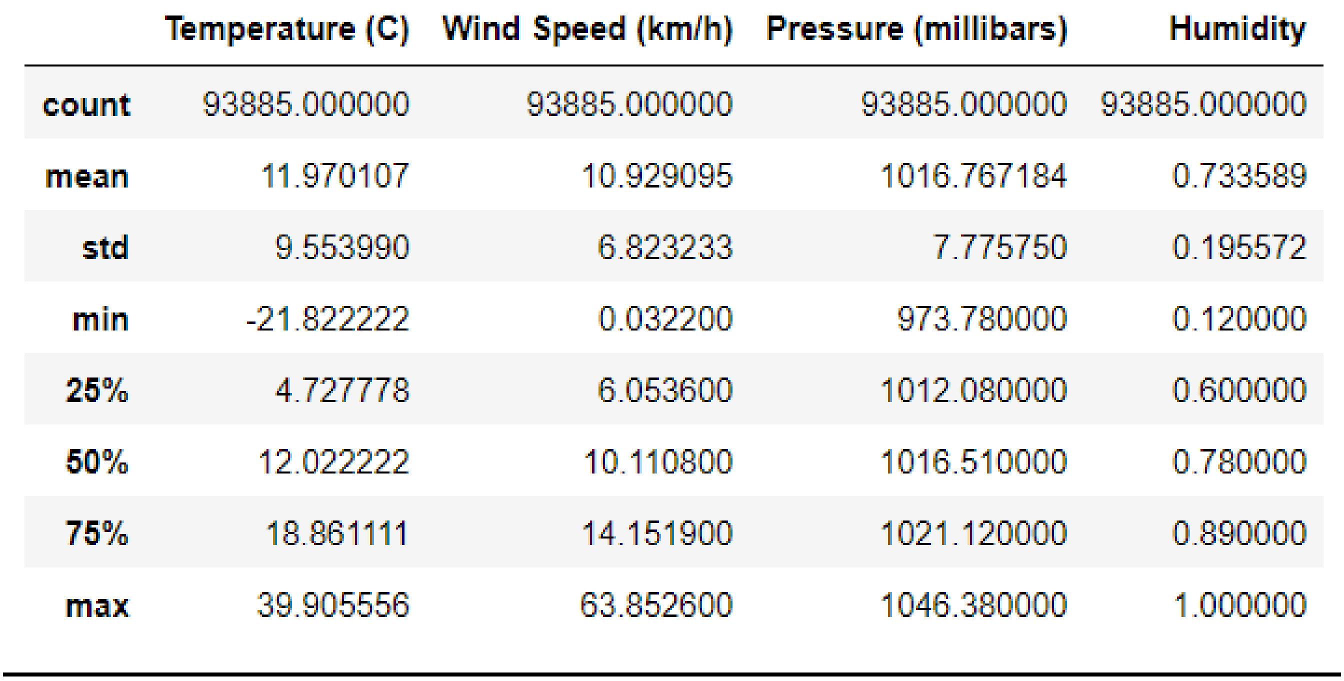
First of all, we create a data frame and observing the description of it:



	Temperature (C)	Wind Speed (km/h)	Pressure (millibars)	Humidity
count	96453.000000	96453.000000	96453.000000	96453.000000
mean	11.932678	10.810640	1003.235956	0.734899
std	9.551546	6.913571	116.969906	0.195473
min	-21.822222	0.000000	0.000000	0.000000
25%	4.688889	5.828200	1011.900000	0.600000
50%	12.000000	9.965900	1016.450000	0.780000
75%	18.838889	14.135800	1021.090000	0.890000
max	39.905556	63.852600	1046.380000	1.000000

The first thing that got our attention is the min of wind speed, pressure and humidity columns. because it's impossible for them to have a zero value and we have to remove these illogical samples.

Now let's observing the new data frame:

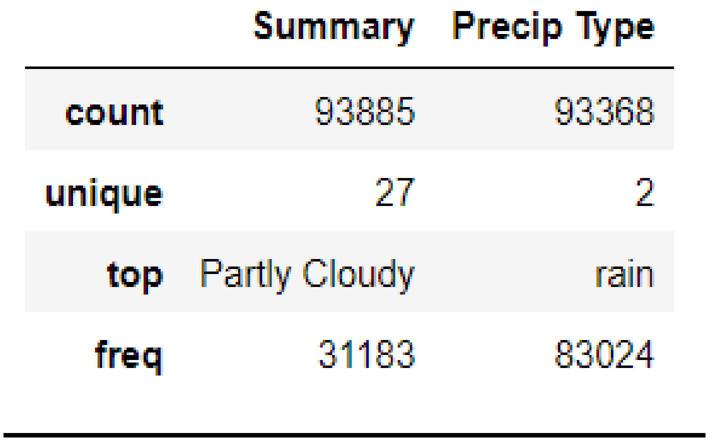


	Temperature (C)	Wind Speed (km/h)	Pressure (millibars)	Humidity
count	93885.000000	93885.000000	93885.000000	93885.000000
mean	11.970107	10.929095	1016.767184	0.733589
std	9.553990	6.823233	7.775750	0.195572
min	-21.822222	0.032200	973.780000	0.120000
25%	4.727778	6.053600	1012.080000	0.600000
50%	12.022222	10.110800	1016.510000	0.780000
75%	18.861111	14.151900	1021.120000	0.890000
max	39.905556	63.852600	1046.380000	1.000000

The number of rows in our data frame has been reduced from 96453 to 93885 (about 2568 rows) and closer to reality.

Regarding the describe table, we can see England is a cold, rainy country and also has a high air humidity.

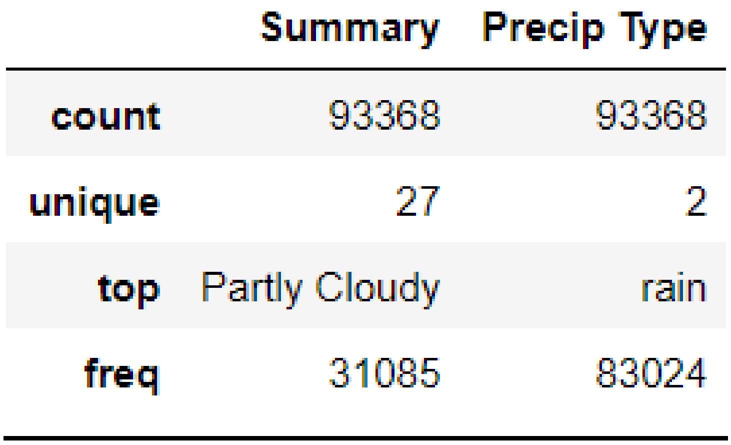
Let's observing the other two column of our data frame that is "Summery" and "Precipe Type":



	Summary	Precip Type
count	93885	93368
unique	27	2
top	Partly Cloudy	rain
freq	31183	83024

As you can see, the count of "Summery" and "Precip Type" is unequal which means we have an error in our data frame that calls "missing value" and we have to remove it.

The description of the new data frame:

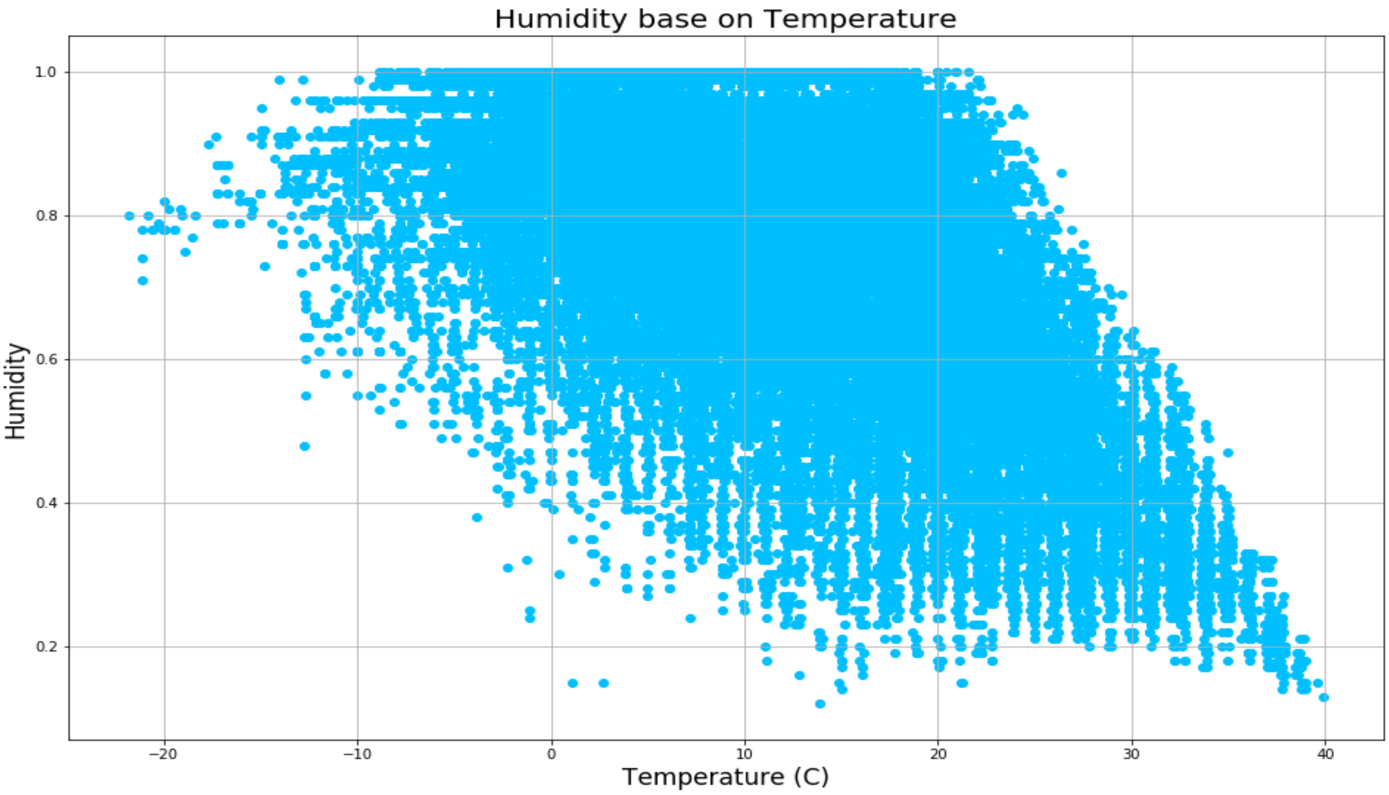


	Summary	Precip Type
count	93368	93368
unique	27	2
top	Partly Cloudy	rain
freq	31085	83024

The number of rows has been reduced again to 93368 (about 517 rows) and now we have a clean data frame.

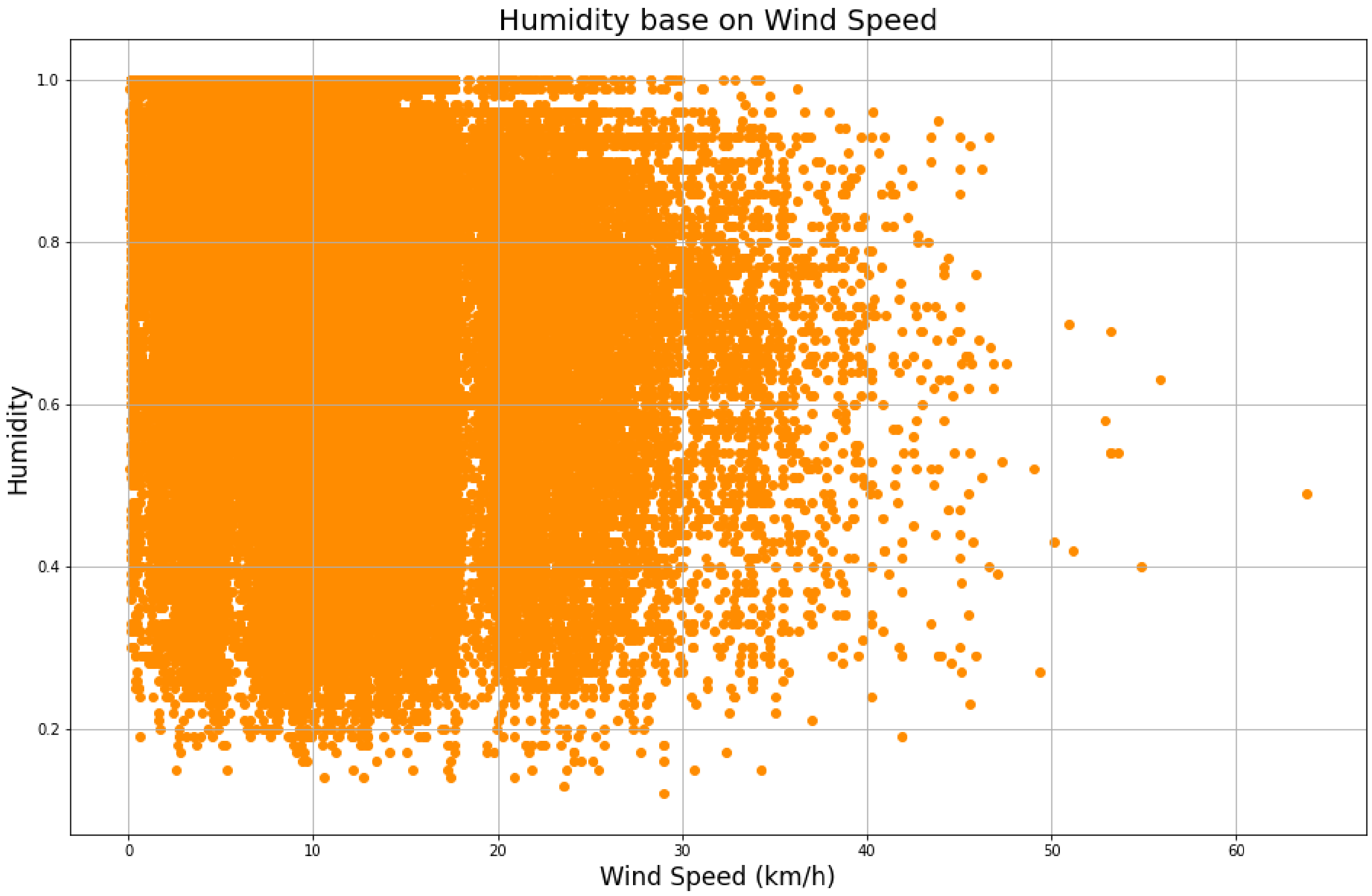
Regarding to the description table, England has about 27 different type of weather and a partly cloudy sky in many days of year. Although England is cold country, but not cold enough to be a snow country in general and it's rainy most of the time.

For the better study the air humidity, we have to checking the plots base on different features:



Air humidity base on temperature:

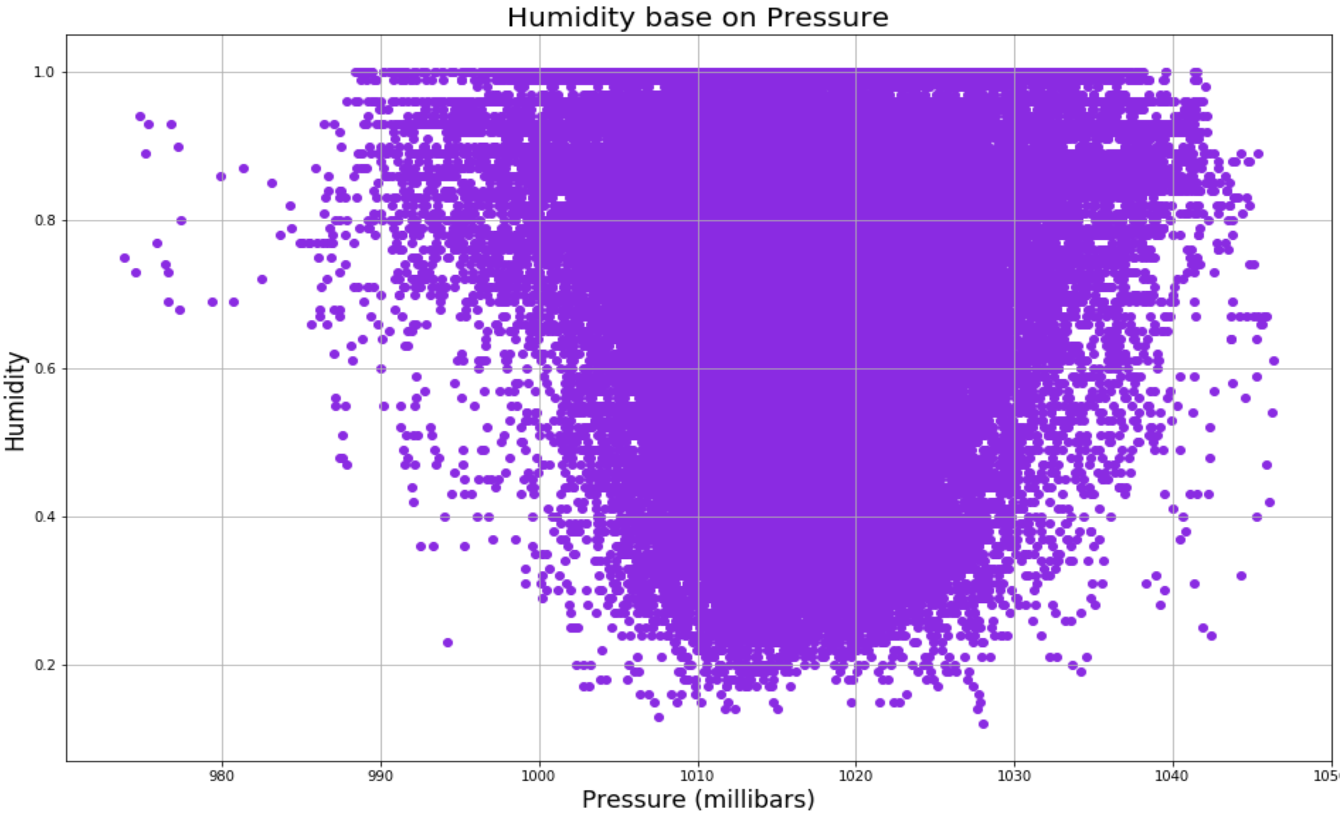
Between -20C and 20C, we have the maximum of humidity (90%). after that by increasing the temperature, the max humidity will decrease till 40C. the min of humidity at the min of temperature is about 80% while the min of humidity is under 20% after 35C.



Air humidity base on wind speed:

In the range of 0 km/h and 30 km/h we have the max of humidity (90%). After that by increasing the wind speed, the max and the min of humidity by sort will decrease and increase and we have smaller rang between max and min of humidity.

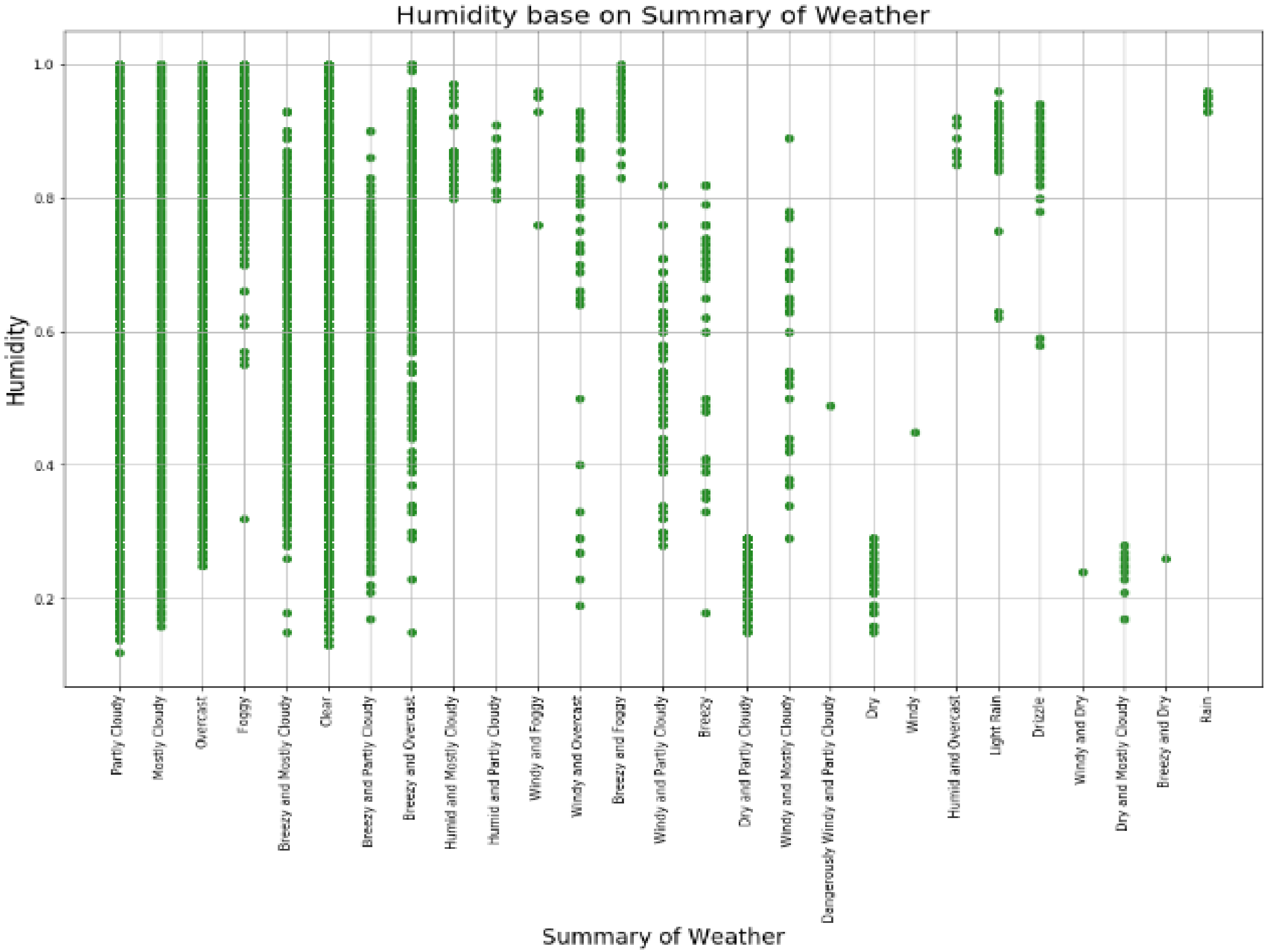
We also have a 20% min of humidity in the range of 0 km/h and 30 km/h and this number increased to almost 60%. But the recorded data between 30 km/h and 45 km/h is less than the firs range we were talking and after 45 km/h, it's verry low records over 10 years and 93368 samples.



Air humidity base on air pressure:

Between 988 millibars and 1038 millibars we have the maximum of humidity (90%) and no changes.

By increasing the pressure to 1010 millibars, the min of humidity is reduced to 20% and it's been the same till pressure of 1020 millibars. After that this process is reversed and the min of humidity increased to 80% at 1040 pressure.



Air humidity base on type of precipitation:

We have the maximum humidity (90%) in the 'Partly Cloudy', 'Mostly Cloudy', 'Overcast', 'Clear' and 'Breezy and Foggy' weathers and the most of samples are recorded in the four first of them.

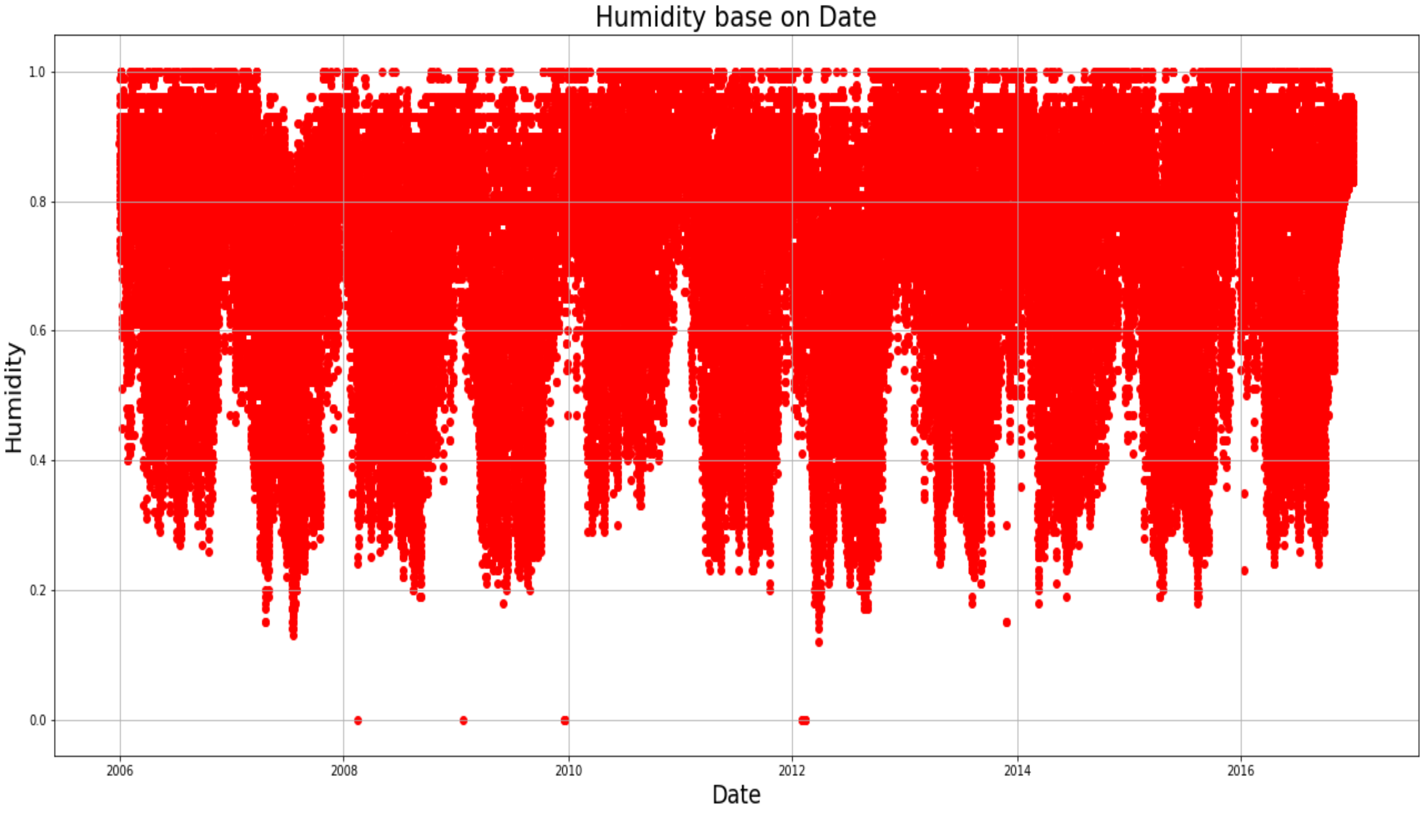
In 'Dangerously Windy', 'Windy', 'Breezy and Dry', 'Windy and Foggy' and 'Windy and Dry' weathers we have verry few records that tell us the humidity have been so less than other weathers between 2006 and 2016 in those weathers.

There is an interesting tip. the England is mostly rainy but the max of humidity at the 'rain' weather is not the maximum of humidity and it's between 85% and 88%.

****

Now that we checked the air humidity plots base on different features, it's time to study that base on date and combined some features together.

But for using the date, we have to change the format of "Formatted Date" to just year and after that we'll be able to draw our plot with annual step.

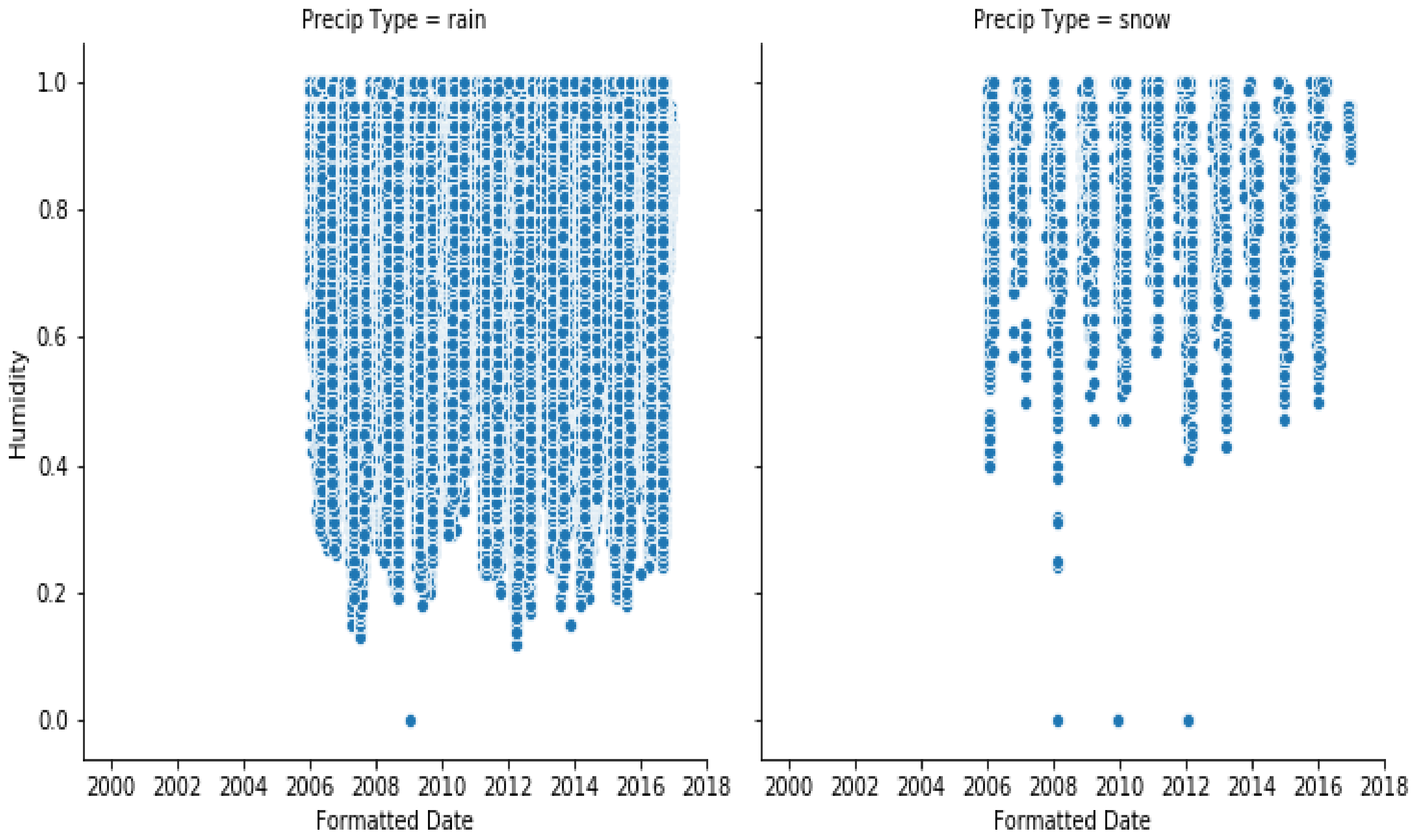


Air humidity base on year:

In the beginning and the ending of a year, the min of humidity is changing a lot and.

The min of humidity starts to decreasing almost in the middle of the year (summer) and when the spiring comes, min of humidity is increasing again.

If we compare the years, we see the maximum of between 2010 summer and 2012 summer is been recorded more than other years and we have the less records in 2007 (start at summer) and the minimum of humidity that recorded Is belong to 2007 and also 2012.



Air humidity base on date (separated by type of precipitation):

In the right plot, the density of samples is on the beginning of year and the ending of a year before which tells us England have snow precipitation just in the winter.

But regarding to the left plot, we have samples all over the year, even in the winter which confirm that fact England is a rainy country.

Although the density of samples in the rain plot is mush more than the snow plot, but the minimum of humidity in snow plot doesn't come lower than 30% while we have minimum of humidity less than 20% in every year at a rain precipitation except 2010 to 2012.

Result: all the features were effective on the humidity of England and 2011 and 2012 experienced the most humidity. Project code and description of its steps: