Weather data from Barajas Airport, Madrid between 1997 and 2015

Team Radical

December 18, 2016

Introduction

Aviation is greatly affected by weather compared to other modes of transportation. From thunderstorms and snow storms, to wind and fog as well as temperature and pressure extremes, every phase of flight has the potential to be impacted by weather. Bad weather may result in delays of flights and damages to the aircrafts during flight which eventually causes loss to the commercial aviation companies. Although there are many available forecasting techniques with the Air Traffic Control of various airports, there are still cases of wrong forecasts or sudden climate changes that were not expected at times. Our goal is to study the reasons for the events(rain, fog, thunderstorm,..etc) caused and its relationship with other weather variables like Temperature, Visibility,..etc from weather reports of Barajas Airport, Madrid between 1997 and 2015 and see if we can forecast the events that can be occurred weather and its change to ensure a safe flight from Barajas Airport, Madrid.

Data

The data is orginally collected from https://www.wunderground.com/ and is made available to public by https://www.kaggle.com

(https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv)

Loading the dataset

```
weather <- read.csv(file = "weather.csv", header = TRUE)
```

Exploring the dataset

```
## [1] "Total rows in dataset:  6812  & total values in dataset:  156676"
```

```
## [1] "Total NA values in entire dataset: 7517"
```

There are total 23 variables in the dataset, each described as follows:-
(1) CET - Date range from 1997-01-01 to 2015-12-31
(2) Max.TemperatureC - Maximum Temperature in Celsius throughout the date range CET
(3) Mean.TemperatureC - Mean Temperature in Celsius throughout the date range CET
(4) Min.TemperatureC - Minimum Temperature in Celsius throughout the date range CET
Dew point is the highest temperature at which airborne water vapor will condense to form liquid dew.A higher dew point means there will be more moisture in the air. Dew point is sometimes called “frost point” when the temperature is below freezing. The measurement of dew point is related to humidity.
(5) Dew.PointC - Dew Point in Celsius throughout the date range CET
(6) MeanDew.PointC - Mean Dew Point in Celsius throughout the date range CET
(7) Min.DewpointC - Minimum Dew Point in Celsius throughout the date range CET
Humidity is the amount of water vapor in the air. Water vapor is the gaseous state of water and is invisible. Humidity indicates the likelihood of precipitation, dew, or fog. It is expressed in gram per cubic meter.

(8) Max.Humidity - Maximum Humidity throughout the date range CET
(9) Mean.Humidity - Mean Humidity throughout the date range CET
(10) Min.Humidity - Minimum Humidity throughout the date range CET
Sea Level pressure is the atmospheric pressure at sea level. In the dataset, it is expressed in hPa which means hectoPascal.
(11) Max.Sea.Level.PressurehPa - Maximum Sea Level Pressure throughout the date range CET
(12) Mean.Sea.Level.PressurehPa - Mean Sea Level Pressure throughout the date range CET
(13) Min.Sea.Level.PressurehPa - Minimum Sea Level Pressure throughout the date range CET
Visibility is a measure of the distance at which an object or light can be clearly discerned. In the dataset, Visibility is expressed in Kilometers(Km).
(14) Max.VisibilityKm - Maximum Visibility in Kilometers throughout the date range CET
(15) Mean.VisibilityKm - Mean Visibility in Kilometers throughout the date range CET
(16) Min.VisibilitykM - Minimum Visibility in Kilometers throughout the date range CET
Wind speed is caused by air moving from high pressure to low pressure, usually due to changes in temperature. In the dataset, Wind speed is expressed in kilometer per hour(km/h).
(17) Max.Wind.SpeedKm.h - Maximum Wind Speed in km/h throughout the date range CET
(18) Mean.Wind.SpeedKm.h - Mean Wind Speed in km/h throughout the date range CET
A gust is a sudden increase of the wind's speed that lasts no more than 20 seconds. This usually occurs when wind speeds reach a peak of at least 16 knots. A wind gust usually comes in 2-minute intervals. A wind gust comes quite suddenly and abruptly. There are a number of different reasons for wind gusts to occur. One of the causes for a wind gust is when there is a sudden shift from high pressure to low pressure. Another cause for a wind gust to occur is the terrain.
(19) Max.Gust.SpeedKm.h - Maximum Speed in km/h throughout the date range CET
(20) Precipitationmm - Precipitation is the standard way of measuring rainfall or snowfall is the standard rain gauge. It is measured in millimeters(mm).
(21) CloudCover - Cloud cover (also known as cloudiness, cloudage or cloud amount) refers to the fraction of the sky obscured by clouds when observed from a particular location. Okta is the usual unit of measurement of the cloud cover. It's value ranges from 0(completely clear sky) to 9(sky obstructed from view)
(22) Events - Natural weather changes like Rain, Snow, Fog, Thunderstorm etc caused one at a time or mixture of more than one.
(23) WindDirDegrees - Wind direction is reported by the direction from which it originates. For example, a northerly wind blows from the north to the south. Wind direction is usually reported in cardinal directions or in azimuth degrees. For example, a wind coming from the south is given as 180 degrees; one from the east is 90 degrees.

Data Preprocessing

Though we have 23 variables(columns) in the dataset, our main focus is on only few variables that is used in further analysis of the project. Our variables of interest will be explained further.

```
## [1] "Data loss if na.omit() is performed on dataset: 3966 rows"
```

In the Events column, as seen below there no NA values, but from the observations in the dataset we have noticed some values to be empty strings

```
##  [1] ""                       "Fog"
##  [3] "Fog-Rain"               "Fog-Rain-Snow"
##  [5] "Fog-Rain-Thunderstorm"  "Fog-Snow"
##  [7] "Fog-Thunderstorm"       "Rain"
##  [9] "Rain-Hail"              "Rain-Hail-Thunderstorm"
## [11] "Rain-Snow"              "Rain-Snow-Thunderstorm"
## [13] "Rain-Thunderstorm"      "Snow"
## [15] "Thunderstorm"           "Tornado"
```

As, we have gone through the dataset we have observed that only events are recorded when there is only a

weather change from Normal. So, we are replacing the empty values into string value “Normal”

```
## [1] "Total missing values in CloudCover column 1372"
```

For the missing values (NA) in CloudCover columns, we are replacing the NA values into -1 (“instead of Missing Values - NA”) Also, cloud cover range only from 0 to 8 in general.
Refer - https://en.wikipedia.org/wiki/Okta

In order to not loose data from all rows by omitting NA by rows, we are transferring our interested variables for analysis into new R object

```
## [1] "Variables(columns) of interest for data analysis: "
```

```
##  [1] "CET"               "CET_Year"          "CET_Month"
##  [4] "CET_Date"          "Mean.TemperatureC" "Mean.Humidity"
##  [7] "MeanDew.PointC"    "Mean.VisibilityKm" "Precipitationmm"
## [10] "CloudCover"        "Events"
```

```
## [1] "Column-wise NA totals in the final dataset that we are using for analysis: "
```

```
##                   TotalsNAs  typeof   class
## CET                       0 integer  factor
## CET_Year                  0  double numeric
## CET_Month                 0 integer  factor
## CET_Date                  0  double numeric
## Mean.TemperatureC         3 integer integer
## Mean.Humidity             2 integer integer
## MeanDew.PointC            2 integer integer
## Mean.VisibilityKm       940 integer integer
## Precipitationmm           0  double numeric
## CloudCover                0 integer  factor
## Events                    0 integer  factor
```

```
## [1] "Omitting rows containing NAs from final dataset: 940 rows omitting"
```

Exploratory Data Analysis

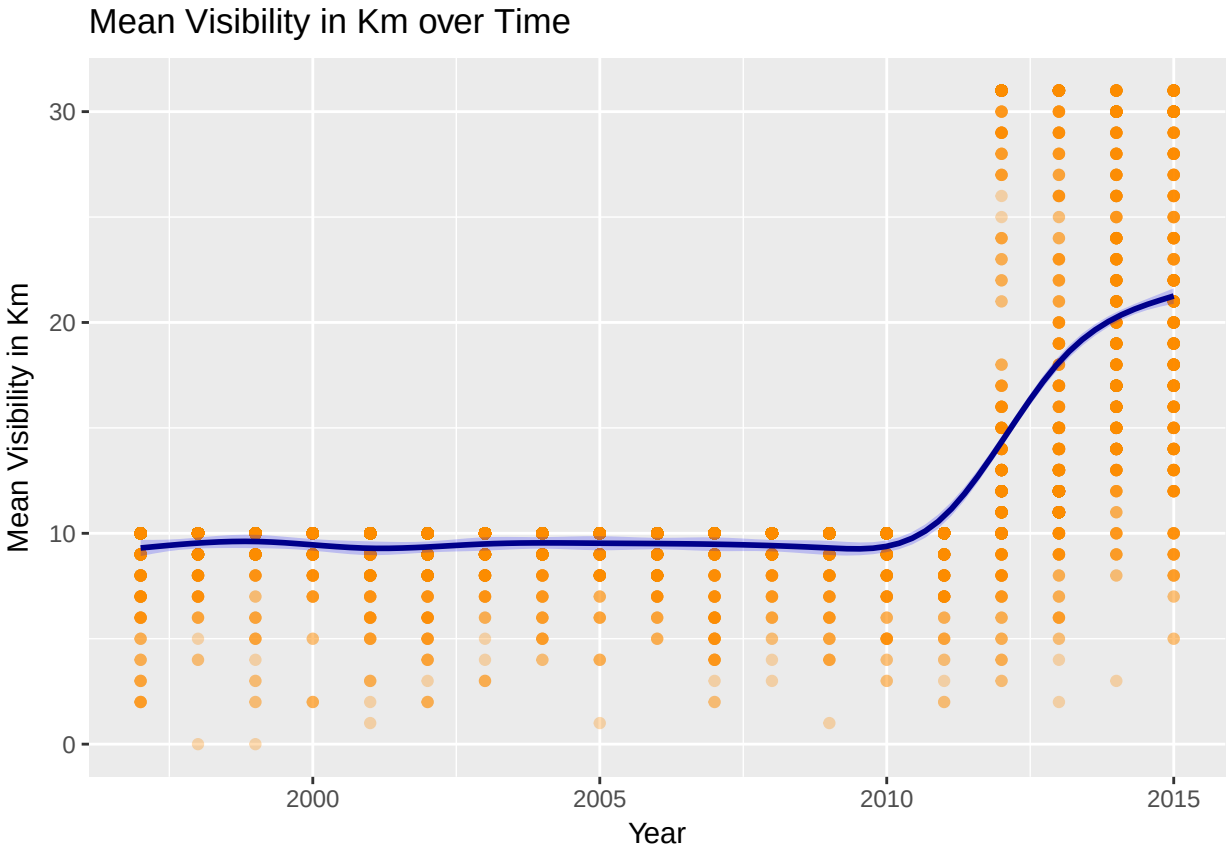
Summary statistics

```
summary(weather2ForAnalysis)
```

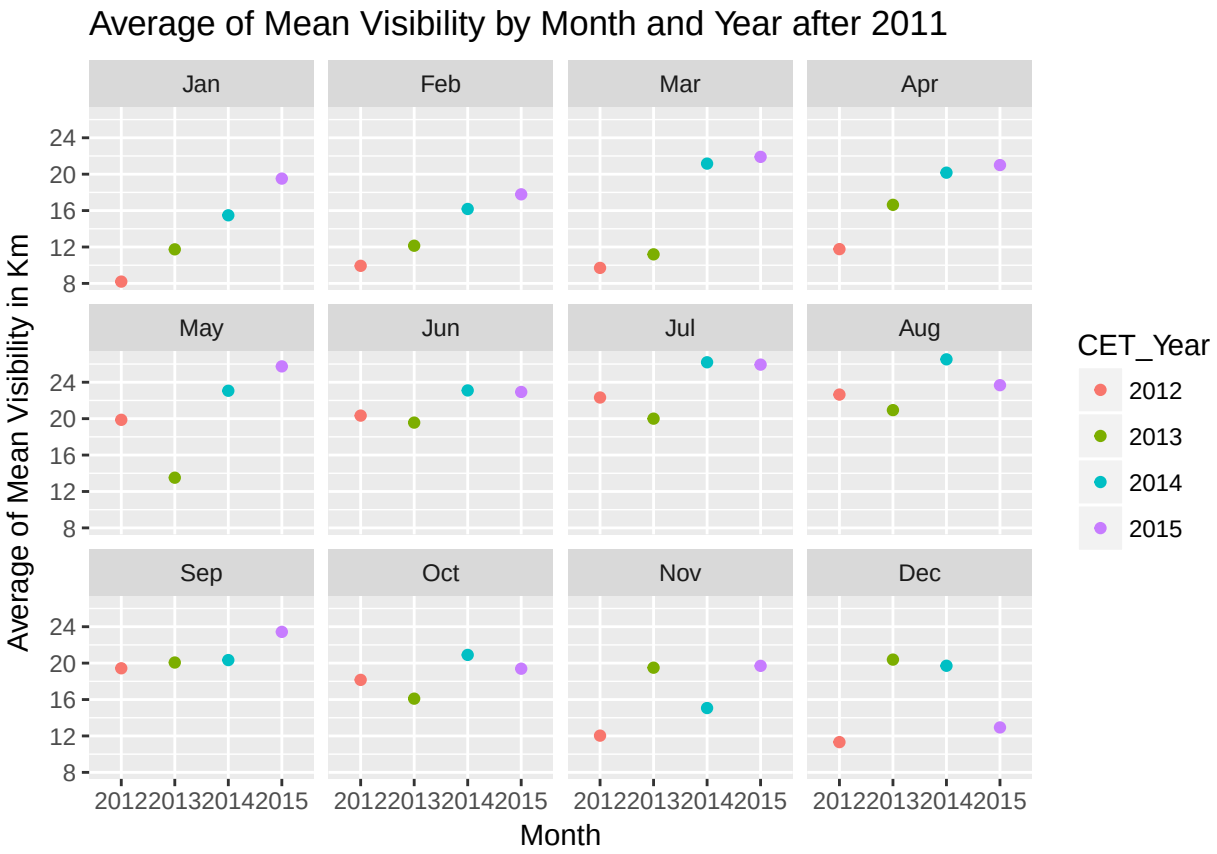
```
##       CET                CET_Year      CET_Month       CET_Date
##  Min.   :1997-01-01   Min.   :1997   May    : 545   Min.   : 1.00
##  1st Qu.:2001-11-27   1st Qu.:2001   Dec    : 520   1st Qu.: 8.00
##  Median :2006-10-02   Median :2006   Oct    : 503   Median :16.00
##  Mean   :2006-09-15   Mean   :2006   Jun    : 499   Mean   :15.65
##  3rd Qu.:2011-10-27   3rd Qu.:2011   Jan    : 497   3rd Qu.:23.00
##  Max.   :2015-12-31   Max.   :2015   Nov    : 497   Max.   :31.00
##                                      (Other):2811
##  Mean.TemperatureC Mean.Humidity   MeanDew.PointC    Mean.VisibilityKm
##  Min.   :-3.00     Min.   : 16.0   Min.   :-15.000   Min.   : 0.00
##  1st Qu.: 8.00     1st Qu.: 44.0   1st Qu.:  3.000   1st Qu.:10.00
##  Median :14.00     Median : 61.0   Median :  6.000   Median :10.00
##  Mean   :14.56     Mean   : 59.8   Mean   :  5.491   Mean   :11.72
##  3rd Qu.:21.00     3rd Qu.: 76.0   3rd Qu.:  9.000   3rd Qu.:10.00
##  Max.   :32.00     Max.   :100.0   Max.   : 16.000   Max.   :31.00
##
```

```
##  Precipitationmm    CloudCover                 Events
##  Min.   : 0.000   1      :1231   Normal           :4074
##  1st Qu.: 0.000   4      : 891   Rain             :1140
##  Median : 0.000   2      : 873   Rain-Thunderstorm: 247
##  Mean   : 0.129   3      : 870   Fog              : 233
##  3rd Qu.: 0.000   5      : 807   Fog-Rain         :  69
##  Max.   :32.000   6      : 556   Thunderstorm     :  45
##                   (Other): 644   (Other)          :  64
```

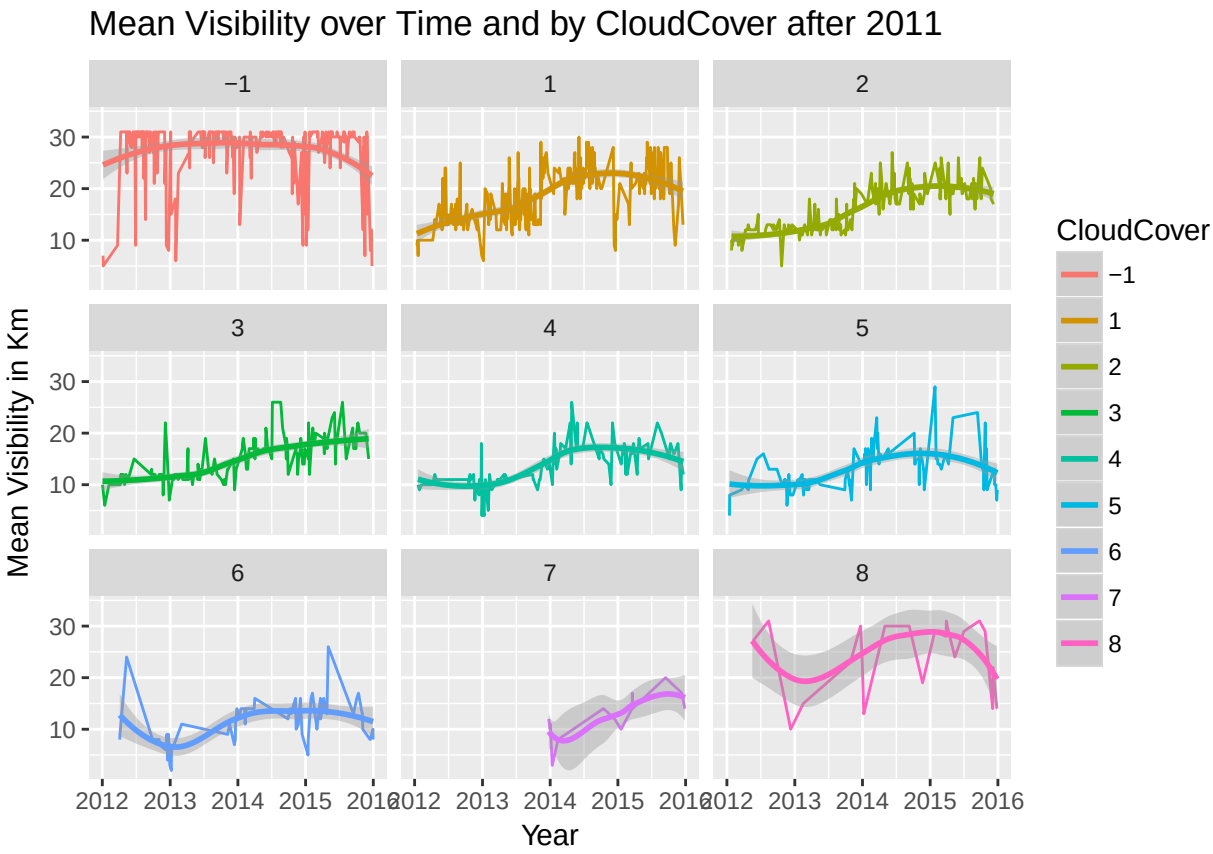
Visualization through various plots



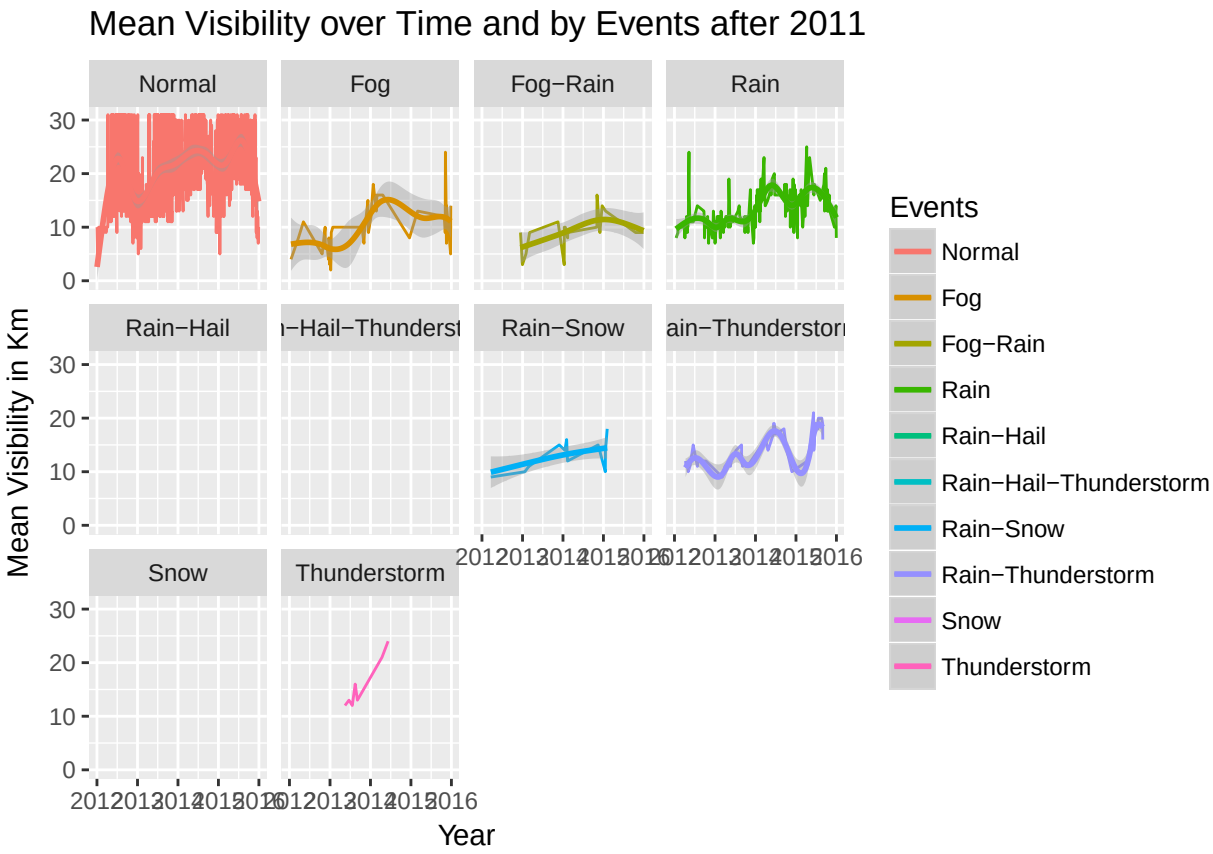
In this plot, we can see that over time the change in mean visibility is steady till the year 2011 and starting 2011 there is a sudden rise which indicates some error or change of criteria. In order to be more meaningful we shifted our focus from the observations recorded starting the year 2012



This plot gives the information of average mean visibility by month and year. It is clear that during the month of January after 2011, the average mean visibility kept rising from year to year, which indicates that the visibility range is becoming clearer. In December after 2011, the average mean visibility showed variation and eventually went down at the end of year 2015. This gives us a sign that the weather cycle is shifting during the timeframe over time.



In this plot, we have closely looked at three variables of interest namely Mean.VisibilityKm, CloudCover and Time in years starting from 2012 Time-series plot in the shows how the visibility is effected by cloud cover over years. As described previously, the less the cloud cover below the clear the sky is. -1 here is the replacement for missing values



This time series plot gives us visual information how the visibility got effected by the causation of an event starting the year 2012. When closely observed we can see that for events such as Rain-Hail, Rain-Hail-Thunderstorm and Snow were not much abundant.

Predictive Analysis

Our main forcus is on forecasting the events that can be occurred weather given various metrics of the weather from Barajas Airport, Madrid.

We have chosen to use the Decision tree analysis to build the prediction model since we are making use of various weather metrics measured in numerics in order to predict a categorical variable that is Events.

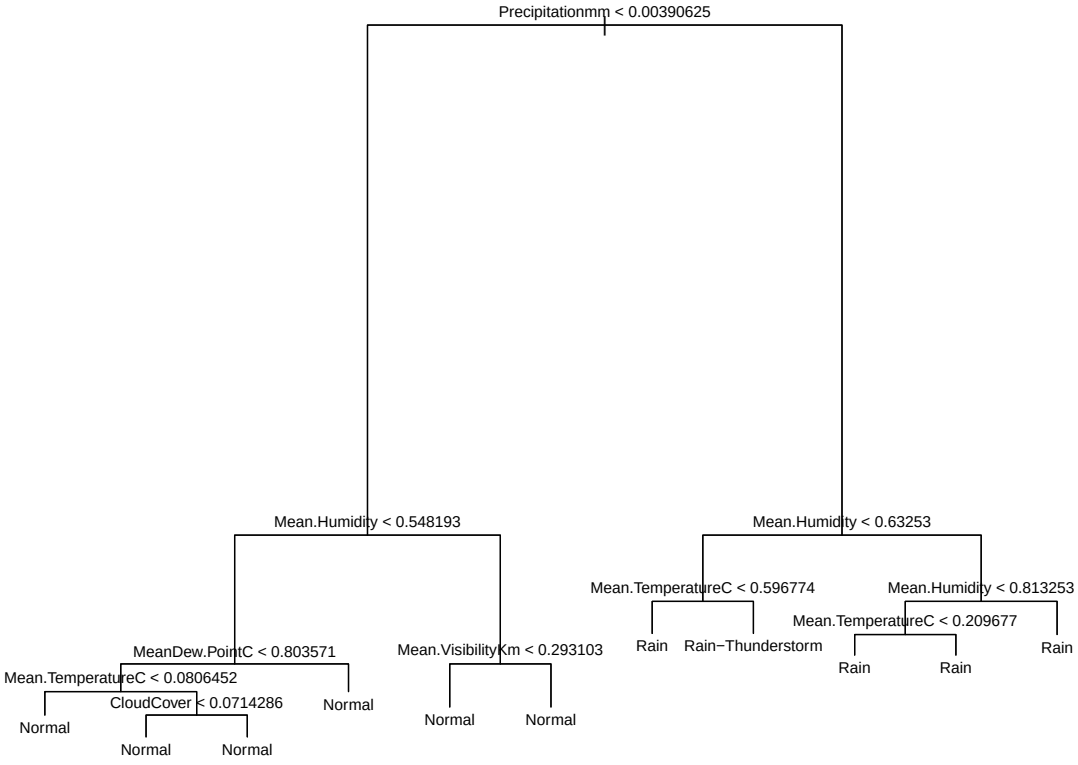
```
## [1] "The columns used for prediction model from original dataset are: "
```

```
##      Mean.TemperatureC Mean.Humidity MeanDew.PointC Mean.VisibilityKm
## 5353                 8            71              3                10
## 5364                 4            80              1                 9
## 5365                 3            87              0                 6
## 5366                 0            94              0                 4
## 5367                 3            90              3                 8
## 5368                 6            72              0                10
##      Precipitationmm CloudCover Events
## 5353               0          5 Normal
## 5364               0          3 Normal
## 5365               0          5 Normal
## 5366               0          7    Fog
## 5367               0          7   Rain
```

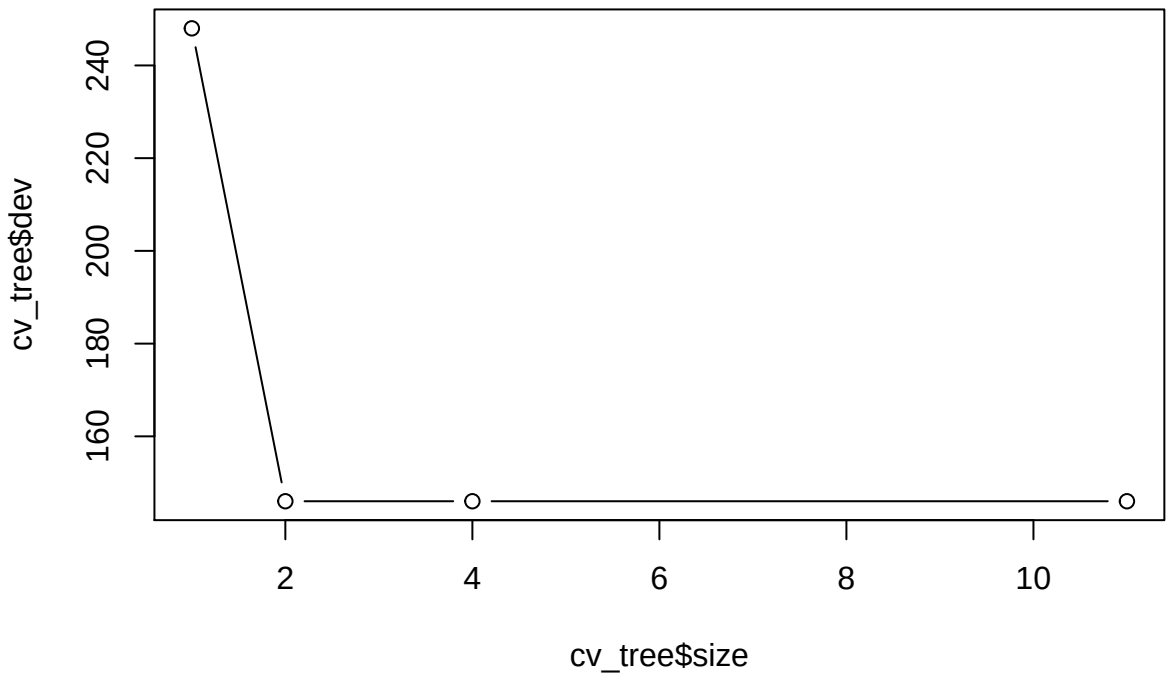
```
## 5368              0         3 Normal
```

```
## [1] "After normalization the data looks like this: "
```

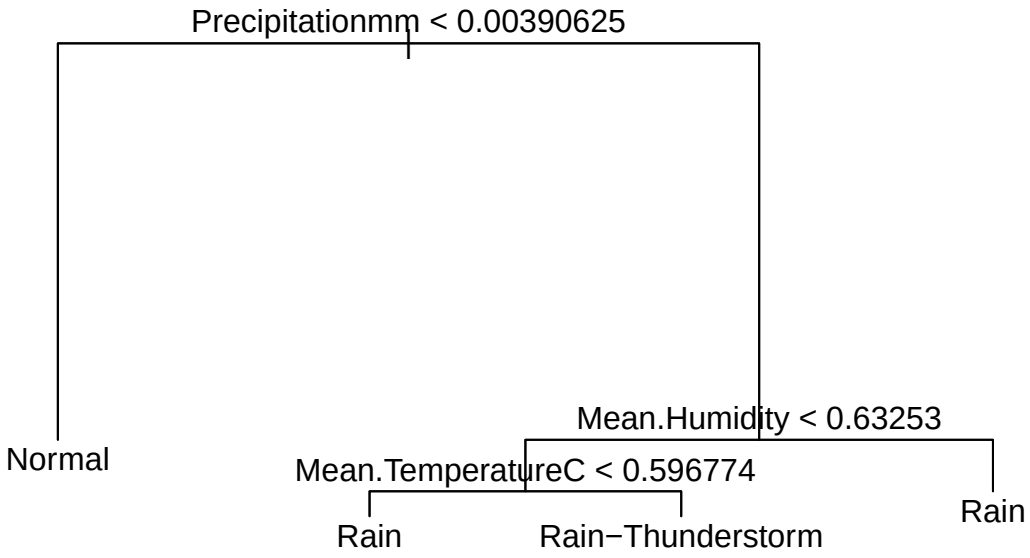
```
##   Mean.TemperatureC Mean.Humidity MeanDew.PointC Mean.VisibilityKm
## 1        0.25806452     0.6626506      0.5357143        0.27586207
## 2        0.12903226     0.7710843      0.4642857        0.24137931
## 3        0.09677419     0.8554217      0.4285714        0.13793103
## 4        0.00000000     0.9397590      0.4285714        0.06896552
## 5        0.09677419     0.8915663      0.5357143        0.20689655
## 6        0.19354839     0.6746988      0.4285714        0.27586207
##   Precipitationmm CloudCover Events
## 1               0  0.2857143 Normal
## 2               0  0.0000000 Normal
## 3               0  0.2857143 Normal
## 4               0  0.5714286    Fog
## 5               0  0.5714286   Rain
## 6               0  0.0000000 Normal
```




```
## [1] "size"   "dev"    "k"      "method"
```




```
## [1] "Our pruned model is built considering best value to be 4"
```




```
##  Accuracy
## 0.8074866
```

Predictions – Test dataset

```
##    Mean.TemperatureC Mean.Humidity MeanDew.PointC Mean.VisibilityKm
## 1                 13            85              4                 3
## 2                 15            96              6                 2
## 3                 14            90              6                 7
## 4                 10            89              7                10
## 5                 12            83              9                 9
## 6                 14            84              2                 9
## 7                 20            58              4                10
## 8                 21            70             -2                 5
## 9                 21            43              2                 8
## 10                22            42              3                 9
## 11                20            36              2                 7
## 12                19            32              3                 7
## 13                22            58              2                 5
## 14                24            50              1                 3
## 15                22            58              2                10
## 16                14            44             -3                10
## 17                13            63              0                 8
## 18                16            67             -1                 7
## 19                16            55             -1                 2
```

```
## 20                18           38              4                1
## 21                20           50              3                6
## 22                18           62              5                3
## 23                19           72              7                3
## 24                18           47              6                6
## 25                14           93              4                7
## 26                26           78             12               10
## 27                28           98             14                9
## 28                28           97             13                9
## 29                27          100             12                8
## 30                28           89              4                7
## 31                28           85              9                5
## 32                27           96             10               10
## 33                28           90             10               10
## 34                29           89              9                9
## 35                27           83             12                9
## 36                18           62              5                9
## 37                14           72              7                4
## 38                26           47              9                7
## 39                20           88              4                5
## 40                25           59             13                9
##    Precipitationmm CloudCover           Events
## 1             0.00          5           Normal
## 2             0.00          3           Normal
## 3             0.51          1             Rain
## 4             0.76          4             Rain
## 5            32.00          4             Rain
## 6             3.05          6             Rain
## 7            13.97          3             Rain
## 8             2.03          2             Rain
## 9             2.03          4             Rain
## 10            0.00          2           Normal
## 11            0.00          4           Normal
## 12            4.06          2             Rain
## 13            0.00          3           Normal
## 14            8.89          3 Rain-Thunderstorm
## 15            0.00          6           Normal
## 16            0.00          2           Normal
## 17            0.00          3           Normal
## 18            1.02          4             Rain
## 19            1.02          1             Rain
## 20            1.02          2             Rain
## 21            2.03          6             Rain
## 22            0.00          2           Normal
## 23            0.00          1           Normal
## 24            0.51          5             Rain
## 25            2.03          4             Rain
## 26            0.51          5             Rain
## 27            2.03          2             Rain
## 28            0.00          6           Normal
## 29            0.25          4             Rain
## 30            0.00          4           Normal
## 31           11.94          3             Rain
## 32            0.51          4             Rain
```

```
## 33            0.00          3            Normal
## 34            1.04          2              Rain
## 35            3.67          4              Rain
## 36            2.03          5              Rain
## 37            0.00          7            Normal
## 38            0.00          5            Normal
## 39            0.51          5              Rain
## 40            2.03          6 Rain-Thunderstorm
```

Conclusion

The accuracy of our pruned decision tree classification model(at best = 4) used for prediction is 80.75%. This means the predicted values are expected to be 80.75% true when compared to the actual values. The final model predicts i.e., categorizies the test values into Events(leaf nodes) based on its nodes at higher levels. The prediction is carried out by categorizing the test value through a certain path along the nodes by checking the condition stated at each node. If a condition at a node is statisfied then it goes into next lower level and searches for the condition at the next child node and this process goes on till the leaf node is reached. Therefore, the leaf node where the process stops at is the category of the prediction variable(in our case, Events variable's category).

We have learnt from this project that for an event to be occurred in weather it completely depends on the metrics of the weather and their changes. Also, sometimes even global warming and pollution can be a reason for bad weather conditions. A clean and pollution free environment by the humans can atleast make the weather little better.