Part 1 - Data Cleaning by Omission

UNI:sv2414

NOTE: The actual code can be viewed in the Rmd file “Part 1 - Data Cleaning by Omission.Rmd”

Check if packages are installed, install if required, and load

```
## Loading required package: plyr
## Loading required package: ggplot2
```

Import CSV file containing wind power generation data

Summary of “Windpower” data frame is shown below:

```
##         PCTimeStamp    WTG01_Grid.Production.PossiblePower.Avg...1.
##  1/1/2013 0:00:    1   Min.   : -3
##  1/1/2013 0:10:    1   1st Qu.:191
##  1/1/2013 0:20:    1   Median :518
##  1/1/2013 0:30:    1   Mean   :476
##  1/1/2013 0:40:    1   3rd Qu.:772
##  1/1/2013 0:50:    1   Max.   :850
##  (Other)      :52554   NA's   :725
##  WTG02_Grid.Production.PossiblePower.Avg...2.
##  Min.   : -3
##  1st Qu.:204
##  Median :530
##  Mean   :484
##  3rd Qu.:774
##  Max.   :850
##  NA's   :710
##  WTG03_Grid.Production.PossiblePower.Avg...3.
##  Min.   : -2
##  1st Qu.:214
##  Median :552
##  Mean   :497
##  3rd Qu.:792
##  Max.   :850
##  NA's   :927
```

```
##  WTG04_Grid.Production.PossiblePower.Avg...4.
##  Min.   : -3
##  1st Qu.:222
##  Median :598
##  Mean   :518
##  3rd Qu.:819
##  Max.   :850
##  NA's   :654
##  WTG05_Grid.Production.PossiblePower.Avg...5.
##  Min.   : -3
##  1st Qu.:192
##  Median :553
##  Mean   :495
##  3rd Qu.:805
##  Max.   :850
##  NA's   :685
##  WTG06_Grid.Production.PossiblePower.Avg...6.
##  Min.   : -2
##  1st Qu.:206
##  Median :537
##  Mean   :489
##  3rd Qu.:786
##  Max.   :850
##  NA's   :652
##  WTG07_Grid.Production.PossiblePower.Avg...7. WTG01_Total.Active.power..8.
##  Min.   : -6                                  Min.   :3109970
##  1st Qu.:180                                  1st Qu.:3895622
##  Median :497                                  Median :4744043
##  Mean   :472                                  Mean   :4608851
##  3rd Qu.:785                                  3rd Qu.:5262894
##  Max.   :850                                  Max.   :6045048
##  NA's   :710                                  NA's   :725
##  WTG02_Total.Active.power..9. WTG03_Total.Active.power..10.
##  Min.   : 609852              Min.   :3254759
##  1st Qu.:1391641              1st Qu.:4066341
##  Median :2341906              Median :5022455
##  Mean   :2189450              Mean   :4870726
##  3rd Qu.:2894952              3rd Qu.:5586471
##  Max.   :3690817              Max.   :6413840
##  NA's   :710                  NA's   :927
##  WTG04_Total.Active.power..11. WTG05_Total.Active.power..12.
##  Min.   :3341303               Min.   :3230186
##  1st Qu.:4168935               1st Qu.:4023712
##  Median :5159420               Median :4986563
##  Mean   :5003720               Mean   :4827587
##  3rd Qu.:5736883               3rd Qu.:5531891
##  Max.   :6575858               Max.   :6326853
```

```
##  NA's   :654                  NA's   :685
##  WTG06_Total.Active.power..13. WTG07_Total.Active.power..14.
##  Min.   :3264175              Min.   :3136754
##  1st Qu.:4085929              1st Qu.:3919523
##  Median :5057922              Median :4875312
##  Mean   :4903693              Mean   :4712335
##  3rd Qu.:5624685              3rd Qu.:5410488
##  Max.   :6451012              Max.   :6193951
##  NA's   :652                  NA's   :710
##  MET_Avg..Wind.speed.1..15. MET_Min..Wind.speed.1..16.
##  Min.   : 0.00              Min.   : 0.0
##  1st Qu.: 5.60              1st Qu.: 3.6
##  Median : 8.50              Median : 6.0
##  Mean   : 8.08              Mean   : 5.6
##  3rd Qu.:10.90              3rd Qu.: 7.8
##  Max.   :18.80              Max.   :15.8
##  NA's   :3                  NA's   :3
##  MET_Max..Wind.speed.1..17. GRID1_KWH_DEL
##  Min.   : 0.0               Min.   :     78
##  1st Qu.: 7.6               1st Qu.:2741038
##  Median :11.1               Median :5152314
##  Mean   :10.6               Mean   :5052061
##  3rd Qu.:13.9               3rd Qu.:7152592
##  Max.   :31.5               Max.   :9999699
##  NA's   :3                  NA's   :59
```

Find initial number of rows

```
## [1] "There are 52560 rows."
```

Remove columns I don't need

Updated summary:

```
##         PCTimeStamp    MET_Avg..Wind.speed.1..15. GRID1_KWH_DEL
##  1/1/2013 0:00:    1   Min.   : 0.00              Min.   :     78
##  1/1/2013 0:10:    1   1st Qu.: 5.60              1st Qu.:2741038
##  1/1/2013 0:20:    1   Median : 8.50              Median :5152314
##  1/1/2013 0:30:    1   Mean   : 8.08              Mean   :5052061
##  1/1/2013 0:40:    1   3rd Qu.:10.90              3rd Qu.:7152592
##  1/1/2013 0:50:    1   Max.   :18.80              Max.   :9999699
##  (Other)      :52554   NA's   :3                  NA's   :59
```

Rename columns

Updated summary:

```
##           DateTime      AvgWindSpeed    MeterReading
##  1/1/2013 0:00:    1   Min.   : 0.00   Min.   :     78
##  1/1/2013 0:10:    1   1st Qu.: 5.60   1st Qu.:2741038
##  1/1/2013 0:20:    1   Median : 8.50   Median :5152314
##  1/1/2013 0:30:    1   Mean   : 8.08   Mean   :5052061
##  1/1/2013 0:40:    1   3rd Qu.:10.90   3rd Qu.:7152592
##  1/1/2013 0:50:    1   Max.   :18.80   Max.   :9999699
##  (Other)      :52554   NA's   :3       NA's   :59
```

Convert DateTime to Date-Time values (useful for Part 2)

Check for negative values

```
## [1] "No negative values found in Wind Speed values."
```

```
## [1] "No negative values found in Meter Reading values."
```

Convert negative values, if any, to NA

Convert missing and negative Wind Speed values to 0

Print total number of faulty values found in preliminary cleaning

```
## [1] "Total number of missing or negative values is 65"
```

```
## [1] "These are contained in the following rows:"
```

```
##                  DateTime AvgWindSpeed MeterReading
## 10376 2013-03-14 01:10:00          3.2           NA
## 29720 2013-07-26 09:10:00          5.9           NA
## 30150 2013-07-29 08:50:00          6.1           NA
## 30151 2013-07-29 09:00:00          6.7           NA
## 30152 2013-07-29 09:10:00          6.0           NA
## 30153 2013-07-29 09:20:00          5.7           NA
## 30154 2013-07-29 09:30:00          5.6           NA
## 37776 2013-09-20 07:50:00          7.6           NA
## 37777 2013-09-20 08:00:00          7.2           NA
```

```
## 37778 2013-09-20 08:10:00          7.2           NA
## 37779 2013-09-20 08:20:00          7.3           NA
## 37780 2013-09-20 08:30:00          8.0           NA
## 37781 2013-09-20 08:40:00          8.5           NA
## 37782 2013-09-20 08:50:00          8.0           NA
## 37783 2013-09-20 09:00:00          8.2           NA
## 37784 2013-09-20 09:10:00          8.0           NA
## 37785 2013-09-20 09:20:00          7.6           NA
## 37786 2013-09-20 09:30:00          7.2           NA
## 37787 2013-09-20 09:40:00          7.4           NA
## 37788 2013-09-20 09:50:00          7.5           NA
## 37789 2013-09-20 10:00:00          7.5           NA
## 37790 2013-09-20 10:10:00          7.6           NA
## 37791 2013-09-20 10:20:00          7.5           NA
## 37792 2013-09-20 10:30:00          8.5           NA
## 37793 2013-09-20 10:40:00          7.7           NA
## 37794 2013-09-20 10:50:00          7.0           NA
## 37795 2013-09-20 11:00:00          6.5           NA
## 37796 2013-09-20 11:10:00          8.1           NA
## 37797 2013-09-20 11:20:00          7.3           NA
## 37798 2013-09-20 11:30:00          6.9           NA
## 37799 2013-09-20 11:40:00          6.2           NA
## 37800 2013-09-20 11:50:00          7.0           NA
## 37801 2013-09-20 12:00:00          6.8           NA
## 37802 2013-09-20 12:10:00          7.2           NA
## 37803 2013-09-20 12:20:00          6.5           NA
## 37804 2013-09-20 12:30:00          6.6           NA
## 37805 2013-09-20 12:40:00          6.5           NA
## 37806 2013-09-20 12:50:00          6.8           NA
## 37807 2013-09-20 13:00:00          6.9           NA
## 37808 2013-09-20 13:10:00          7.5           NA
## 37809 2013-09-20 13:20:00          7.3           NA
## 37810 2013-09-20 13:30:00          8.2           NA
## 37811 2013-09-20 13:40:00          7.4           NA
## 37812 2013-09-20 13:50:00          7.4           NA
## 37813 2013-09-20 14:00:00          7.3           NA
## 37814 2013-09-20 14:10:00          7.7           NA
## 37815 2013-09-20 14:20:00          8.0           NA
## 37816 2013-09-20 14:30:00          8.2           NA
## 37817 2013-09-20 14:40:00          7.9           NA
## 37818 2013-09-20 14:50:00          8.0           NA
## 37819 2013-09-20 15:00:00          7.8           NA
## 37820 2013-09-20 15:10:00          7.6           NA
## 37821 2013-09-20 15:20:00          7.4           NA
## 37822 2013-09-20 15:30:00          6.3           NA
## 37823 2013-09-20 15:40:00          7.0           NA
## 37824 2013-09-20 15:50:00          7.1           NA
```

```
## 37825 2013-09-20 16:00:00          7.1           NA
## 37826 2013-09-20 16:10:00          7.9           NA
## 37827 2013-09-20 16:20:00          8.3           NA
```

```
## [1] "Omitting these..."
```

```
## [1] "After omission of missing values, there are now 52495 observations remaining."
```

Add new column for kWh delivered in 10 minutes

Summary of "tenminkwh" column is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     217     417     404     584    2740
```

Remove power generated values greater than rated capacity

```
## [1] "Omitting these..."
```

```
## [1] "After omission of missing values, there are now 52492 observations remaining."
```

Import CSV file containing Manufacturer's PowerCurve

Summary of "mpc" data frame is shown below:

```
##  windspeed_mps     power_kW
##  Min.   : 0.0   Min.   :  0.0
##  1st Qu.: 7.5   1st Qu.:  8.9
##  Median :15.0   Median :626.6
##  Mean   :15.0   Mean   :475.7
##  3rd Qu.:22.5   3rd Qu.:846.9
##  Max.   :30.0   Max.   :850.0
```

Rename columns

Updated summary:

```
##    WindSpeed        Power       
##  Min.   : 0.0   Min.   :  0.0  
##  1st Qu.: 7.5   1st Qu.:  8.9  
##  Median :15.0   Median :626.6  
##  Mean   :15.0   Mean   :475.7  
##  3rd Qu.:22.5   3rd Qu.:846.9  
##  Max.   :30.0   Max.   :850.0
```

WE ARE ASSUMING METER READINGS ARE CORRECT AND WIND VALUES *MAY* BE FAULTY

Add new column for 10min generation according to MPC, Betz Limit, Kinetic Energy of Wind

This is just to examine inconsistencies in a plot

Summary of "tenminmpcurve", "tenminbetz", "tenminKEwind" columns are shown below:

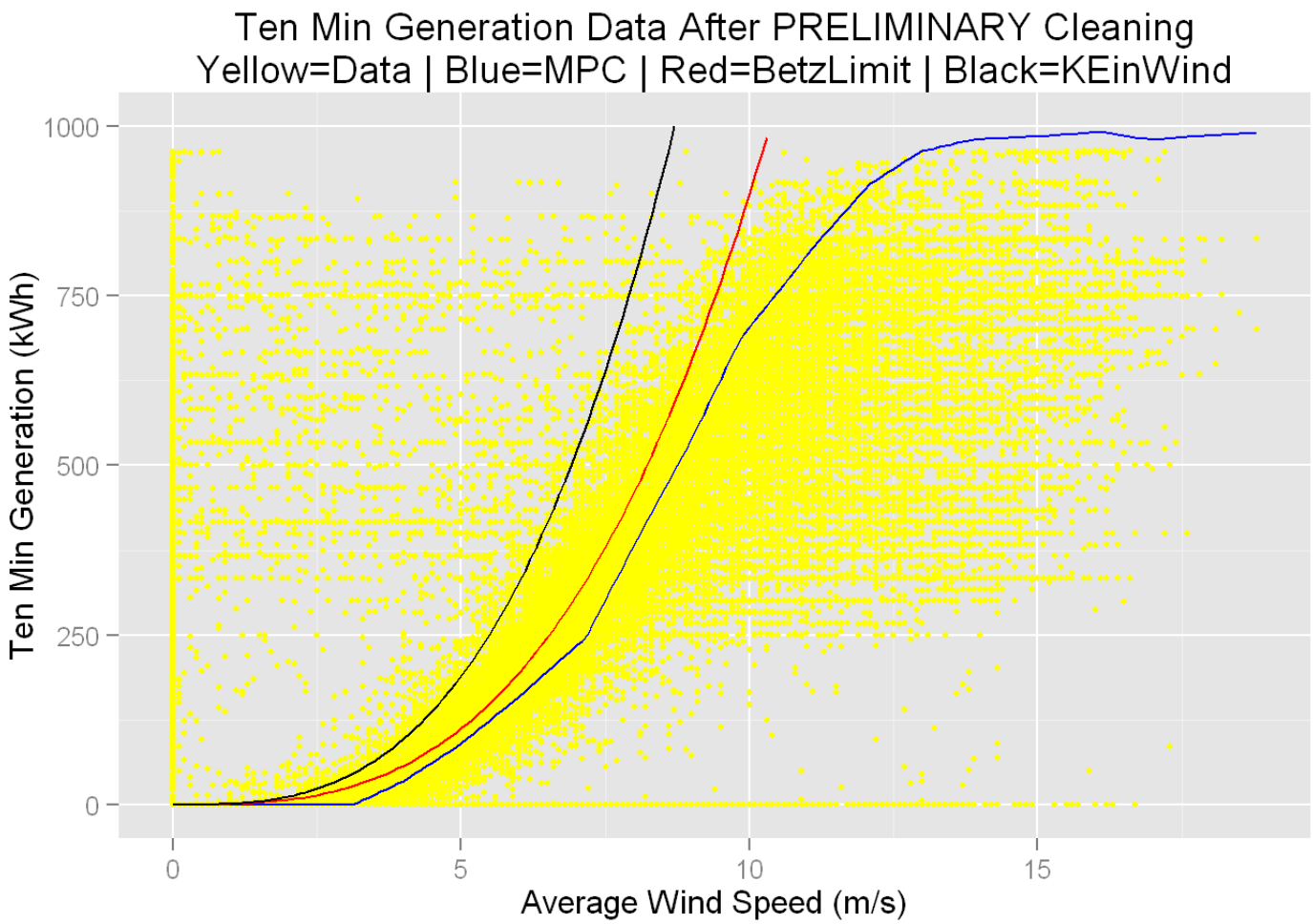
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     131     461     467     799     992
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     158     553     776    1170    5980
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     267     932    1310    1970   10100
```

Plot to see inconsistencies

```
## Warning: Removed 16026 rows containing missing values (geom_path).
## Warning: Removed 24997 rows containing missing values (geom_path).
```



Everything above the MPC should be moved to the MPC

Add a column to calculate equivalent power in KW for ten minute intervals

Summary of "eqPower" column is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     186     357     347     501     826
```

Import CSV file containing Manufacturer's PowerCurve sorted by Power and cleaned

(This is because we are interpolating for wind speed based on power)

Summary of “mpc2” data frame is shown below:

```
##     power_kW   windspeed_mps
##  Min.   :  0   Min.   : 3.13
##  1st Qu.:409   1st Qu.: 8.61
##  Median :840   Median :14.08
##  Mean   :635   Mean   :14.08
##  3rd Qu.:850   3rd Qu.:19.56
##  Max.   :850   Max.   :25.03
```

Rename columns

Updated summary:

```
##      Power       WindSpeed
##  Min.   :  0   Min.   : 3.13
##  1st Qu.:409   1st Qu.: 8.61
##  Median :840   Median :14.08
##  Mean   :635   Mean   :14.08
##  3rd Qu.:850   3rd Qu.:19.56
##  Max.   :850   Max.   :25.03
```

Add a column to interpolate for wind values based on MPC

Summary of “mpcWind” column is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.13    6.77    8.22    7.79    9.25   13.00
```

Add a new column for final cleaned wind speed values

Compare measured and interpolated wind values, and assign NA where measured value is less than interpolated value

Summary of “finalWSvalue” column before omission of NAs is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       3       9      10      10      12      19   23231
```

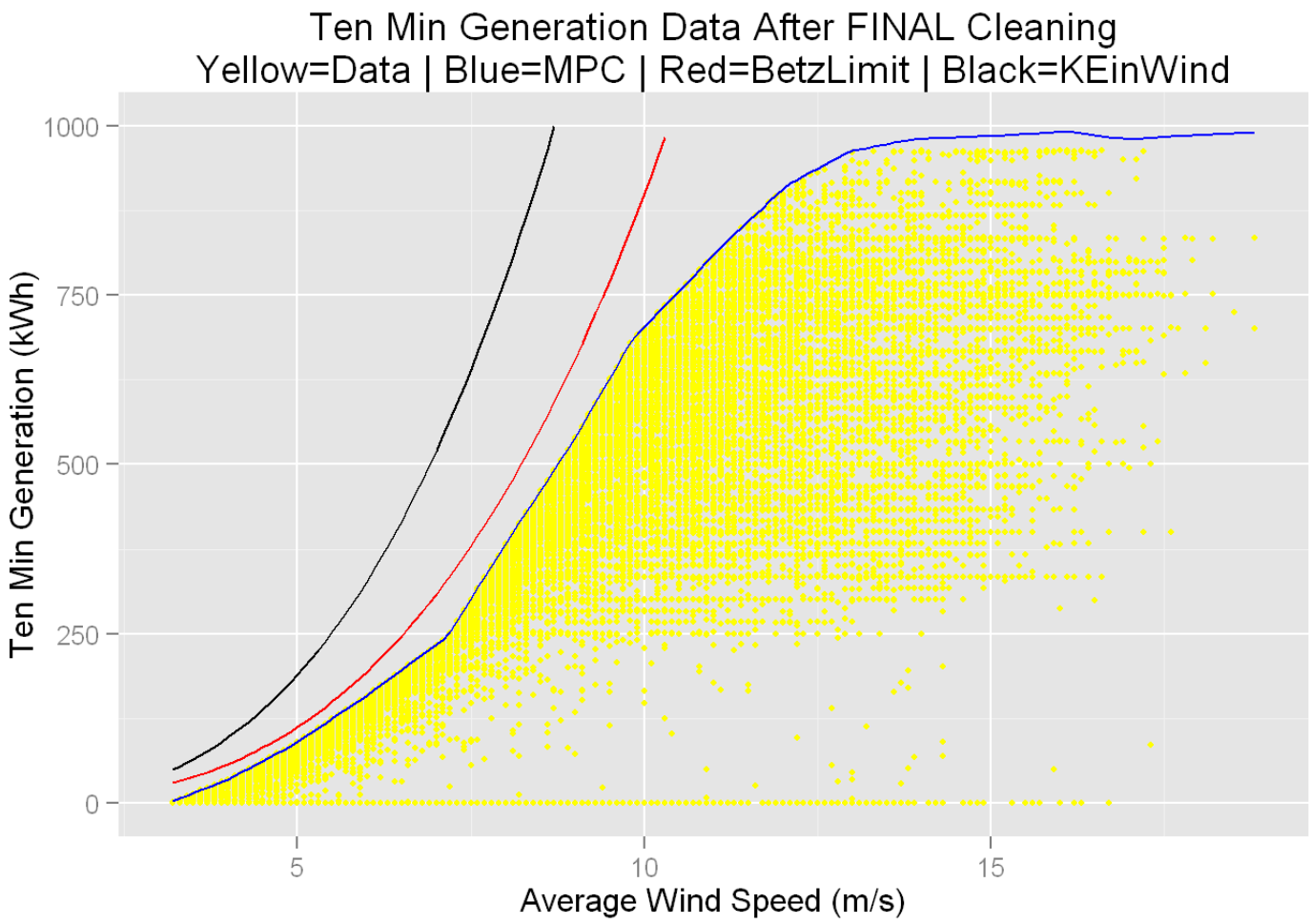
Rows containing NAs are omitted

Summary of "finalWSvalue" column after omission of NAs is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.2     9.0    10.5    10.4    12.1    18.8
```

Plot to see cleaned results

```
## Warning: Removed 15597 rows containing missing values (geom_path).
## Warning: Removed 22792 rows containing missing values (geom_path).
```



This is satisfactory; everything is on or below the MPC!

Add new rows for actual and uncurtailed generation

Summary of "ActualGenerationkWh", "UncurtailedGenerationkWh" columns are shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     366     499     481     644     964
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.9   540.0   757.0   689.0   915.0   992.0
```

Calculate total annual actual and uncurtailed generation

```
## [1] "ANNUAL GENERATION IS 14077445.00kWh"
```

```
## [1] "ANNUAL UNCURTAILED GENERATION IS 20168418.70kWh"
```

Calculate total possible generation at nameplate capacity (850kW)

Calculate actual and uncurtailed Capacity Factors

```
## [1] "ACTUAL CAPACITY FACTOR IS 27.0pc"
```

```
## [1] "UNCURTAILED CAPACITY FACTOR IS 38.7pc"
```

Add a column for Kinetic Energy in wind at cleaned values of wind speeds

Summary of "KEinWind" column is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      50    1110    1760    2030    2690   10100
```

Add a column for Turbine Efficiency

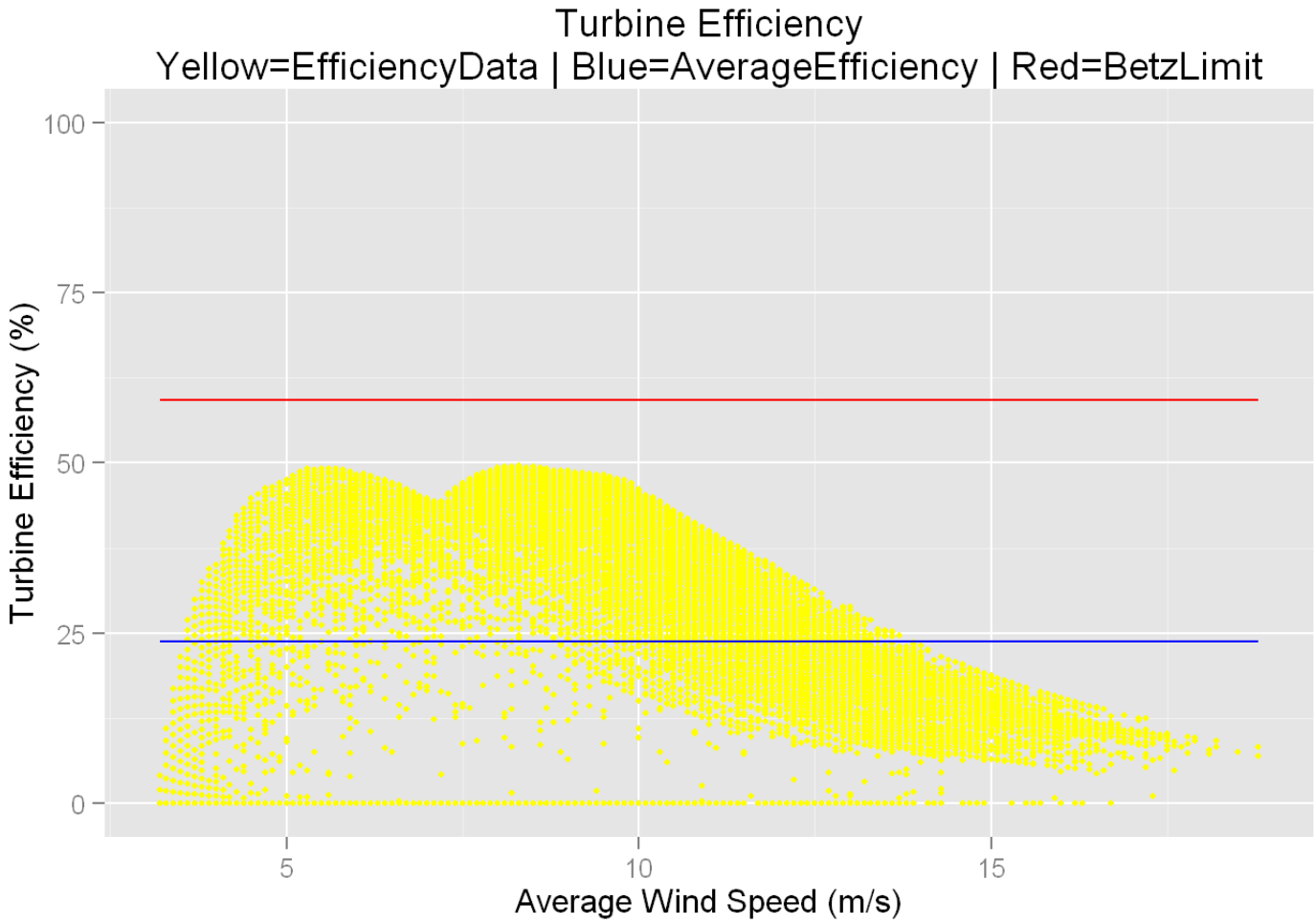
Summary of "TurbineEfficiency" column is shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    19.6    28.9    28.5    38.2    49.5
```

Calculate average Turbine Efficiency for the year

```
## [1] "Average Turbine Efficiency for the year is 23.72pc"
```

Plot turbine efficiency and compare to Betz Limit



It looks to be right!

Find final number of rows

```
## [1] "Final number of rows (after all cleaning) is 29261"
```

```
## [1] "Total number of rows omitted is 23299"
```

FINAL COMMENTS:

1.The actual, uncurtailed capacity factors and the turbine efficiency is found to be lesser when data is cleaned by omission (27.0%, 38.7%, 23%) than when it is cleaned by correction(40.7%, 52.4%, 28%).

2. Nearly half the rows were deleted on account of omission. This is not good and is expected to compromise the reliability of the solution.