Citi Bike Analysis

Saurabh Sankhe

February 8, 2019

```
#Loading the dataset
mydata <- read.csv("C:/Users/Saurabh/Desktop/Sem-2 Course
Documents/Multivariate Analysis/station_72.csv")

#Printing the names of the variables
names(mydata)
```

```
## [1] "datetime"    "demand"      "temperature" "humidity"    "windspeed"
## [6] "visibility"  "condition"
```

```
#Printing the head of the dataframe
head(mydata)
```

```
##           datetime demand temperature humidity windspeed visibility
## 1 08-01-2016 00:00      0        71.1     91.5  4.000000   7.500000
## 2 08-01-2016 01:00      0        71.1     90.0  8.000000   6.900000
## 3 08-01-2016 02:00      0        70.0     93.0  7.000000   4.600000
## 4 08-01-2016 03:00      0        70.0     90.0  9.000000   8.100000
## 5 08-01-2016 04:00      0        70.0     90.0  9.666667   8.466667
## 6 08-01-2016 05:00      0        70.0     89.0  9.666667   9.800000
##   condition
## 1         2
## 2         1
## 3         1
## 4         2
## 5         3
## 6         3
```

```
#Printing the Summary of the dataset
summary(mydata)
```

```
##              datetime        demand        temperature       humidity
##  01-01-2017 00:00:   1   Min.   : 0.000   Min.   :14.00   Min.   : 13.00
##  01-01-2017 01:00:   1   1st Qu.: 0.000   1st Qu.:43.00   1st Qu.: 47.00
##  01-01-2017 02:00:   1   Median : 2.000   Median :57.00   Median : 60.00
##  01-01-2017 03:00:   1   Mean   : 4.153   Mean   :56.92   Mean   : 62.29
##  01-01-2017 04:00:   1   3rd Qu.: 6.000   3rd Qu.:71.10   3rd Qu.: 78.00
##  01-01-2017 05:00:   1   Max.   :35.000   Max.   :96.10   Max.   :100.00
##  (Other)         :8610                    NA's   :4       NA's   :13
##    windspeed        visibility       condition
##  Min.   : 0.200   Min.   : 3.500   Min.   :1.000
##  1st Qu.: 9.000   1st Qu.: 4.050   1st Qu.:1.000
##  Median :10.000   Median : 5.800   Median :1.000
```

```
##  Mean   : 8.999   Mean   : 6.183   Mean   :1.193
##  3rd Qu.:10.000   3rd Qu.: 7.500   3rd Qu.:1.000
##  Max.   :10.000   Max.   :26.500   Max.   :8.000
##  NA's   :14       NA's   :13
```

```
#Printing the total number of rows with na values
sum(rowSums(is.na(mydata)) > 0)
```

```
## [1] 26
```

We can say from the above results that there are 26 rows with na values.

```
#Printing the head of rows with NA values
head(mydata[rowSums(is.na(mydata)) > 0, ])
```

```
##             datetime demand temperature humidity windspeed visibility
## 24  08-01-2016 23:00      2        72.0       71        NA         3.5
## 89  08-04-2016 16:00      1        78.1       52         9          NA
## 163 08-07-2016 18:00     13        84.0       NA        NA          NA
## 310  8/13/2016 21:00      0        90.0       NA        NA         3.5
## 327  8/14/2016 14:00      4        93.9       NA         8         6.9
## 535   8/23/2016 6:00      5        64.0       58        NA          NA
##     condition
## 24          1
## 89          1
## 163         1
## 310         1
## 327         1
## 535         1
```

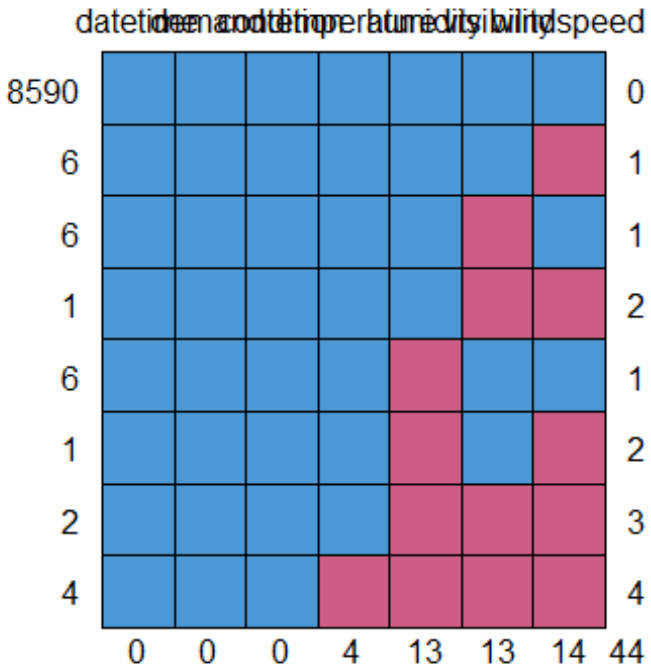
```
#Loading the required libraries
library(mice)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(ggplot2)

#Checking the pattern for null values
md.pattern(mydata)
```

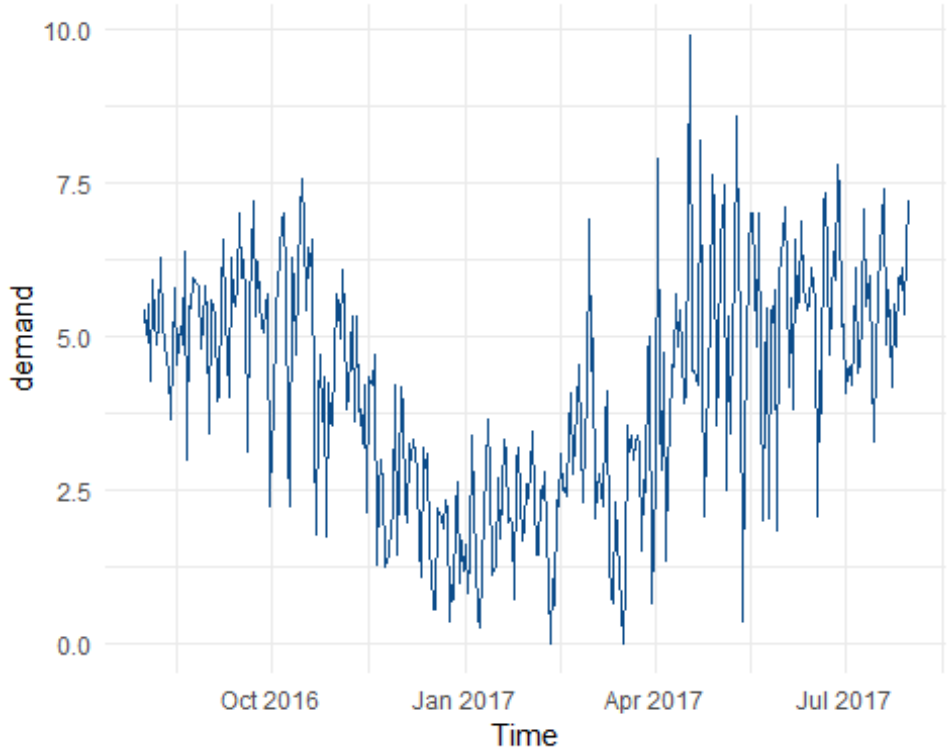



```
##      datetime demand condition temperature humidity visibility windspeed
## 8590        1      1         1           1        1          1         1
## 6           1      1         1           1        1          1         0
## 6           1      1         1           1        1          0         1
## 1           1      1         1           1        1          0         0
## 6           1      1         1           1        0          1         1
## 1           1      1         1           1        0          1         0
## 2           1      1         1           1        0          0         0
## 4           1      1         1           0        0          0         0
##             0      0         0           4       13         13        14
##
## 8590  0
## 6     1
## 6     1
## 1     2
## 6     1
## 1     2
## 2     3
## 4     4
##      44
```

From the above plot we can say that we have 44 NA values in total and 26 rows has those NA values.

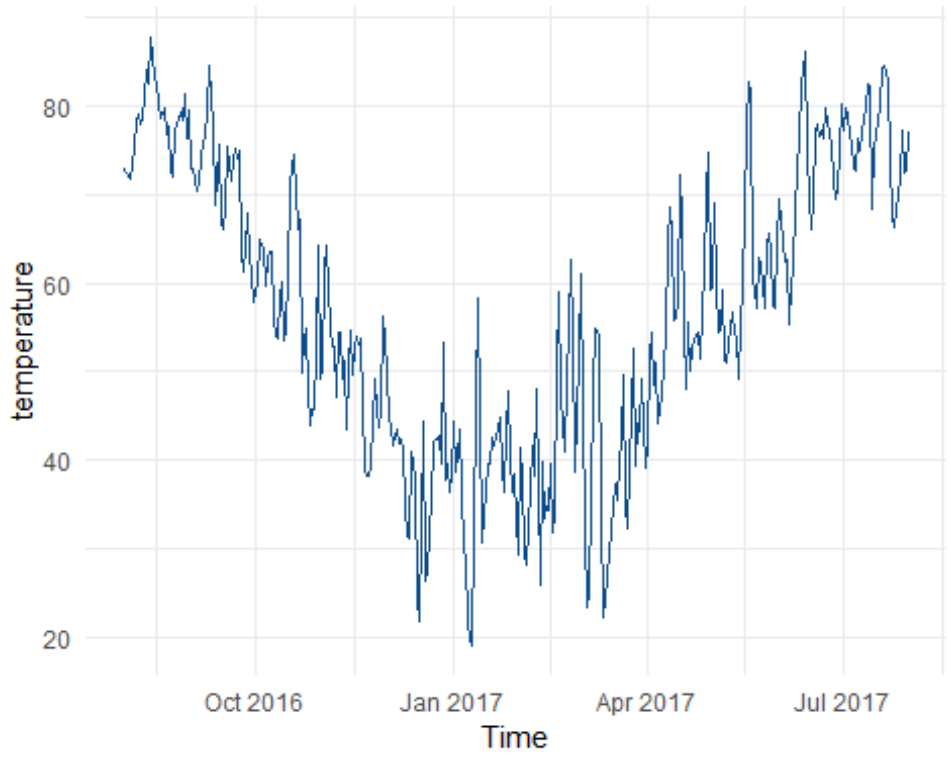
```
# Extracting Date from datetime
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date

library(plyr)

##
## Attaching package: 'plyr'

## The following object is masked from 'package:lubridate':
##
##     here

mydata$datetime <- mdy_hm(mydata$datetime)
mydata$newdate = as.POSIXct(strptime(mydata$datetime, format="%Y-%m-%d"))
mydata$newdate = as.Date(mydata$newdate, "%m%d%Y")

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%m%d%Y'

#Plotting demand against time
agg_demand=aggregate(demand~newdate,data=mydata,mean)
ggplot(data = agg_demand) +
  aes(x = newdate, y = demand) +
  geom_line(color = "#0c4c8a") +
  theme_minimal()+
  xlab("Time")
```

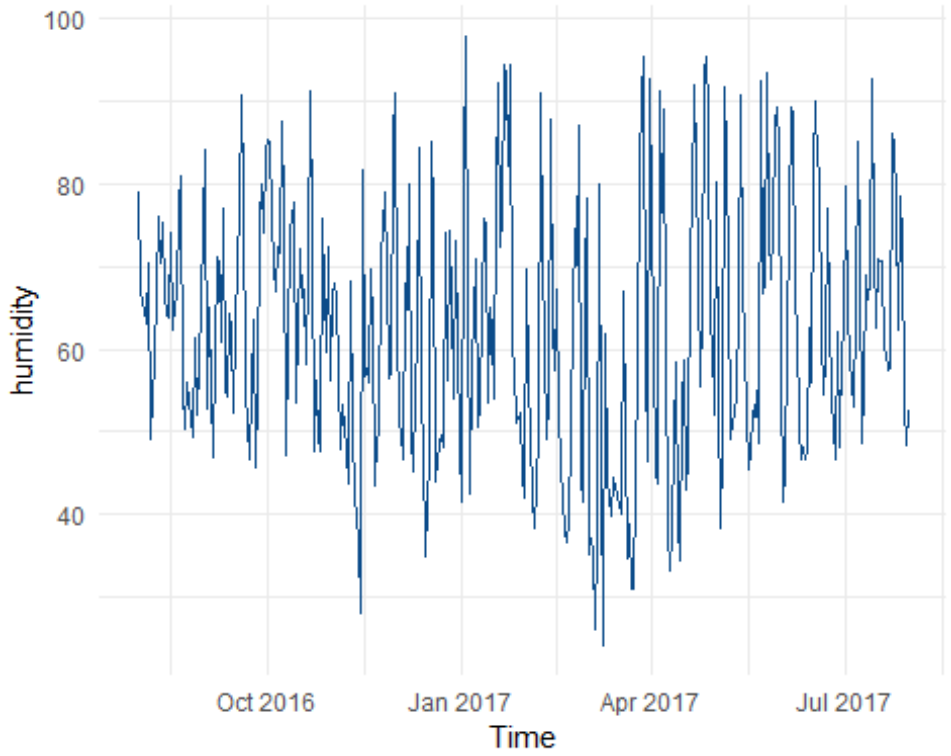


It can be inferred from the above graph that demand fell in winter i.e from November till March and is high from April till october

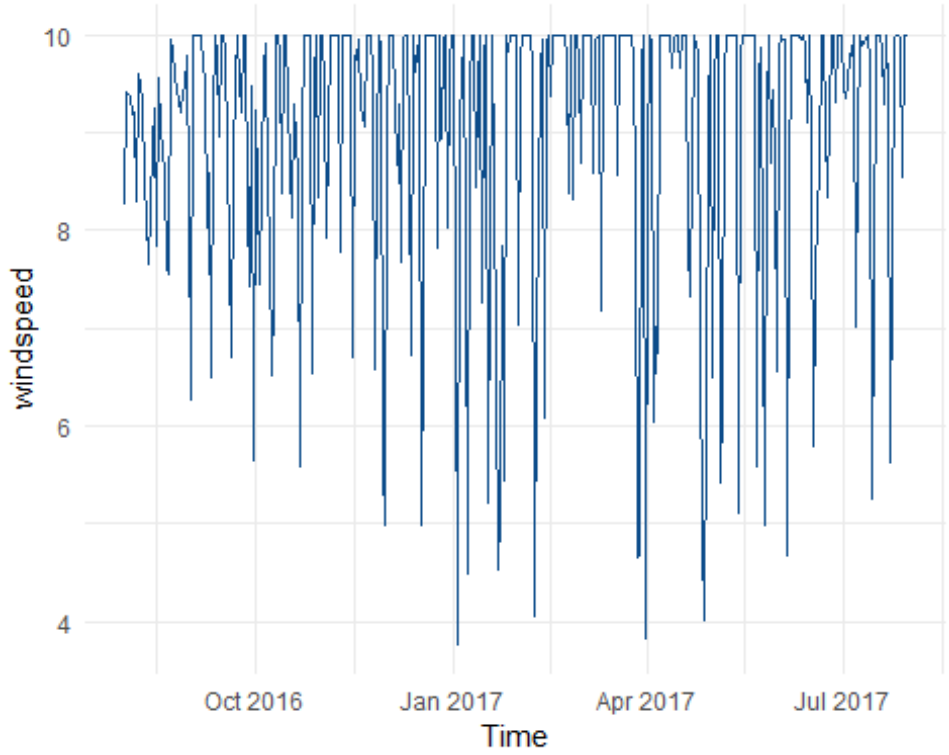
```
#Creating new matrix for temperature,humidity,windspeed,visibility
agg_temp = aggregate(temperature~newdate, data=mydata,mean)
agg_humidity=aggregate(humidity~newdate,data=mydata,mean,na.rm=TRUE)
agg_windspeed=aggregate(windspeed~newdate,data=mydata,mean,na.rm=TRUE)
agg_visibility=aggregate(visibility~newdate,data=mydata,mean, na.rm=TRUE)

#Plotting temperature,humidity,windspeed,visibility against time
ggplot(data = agg_temp) + aes(x = newdate, y = temperature) + geom_line(color
= "#0c4c8a") + theme_minimal() + xlab("Time")
```



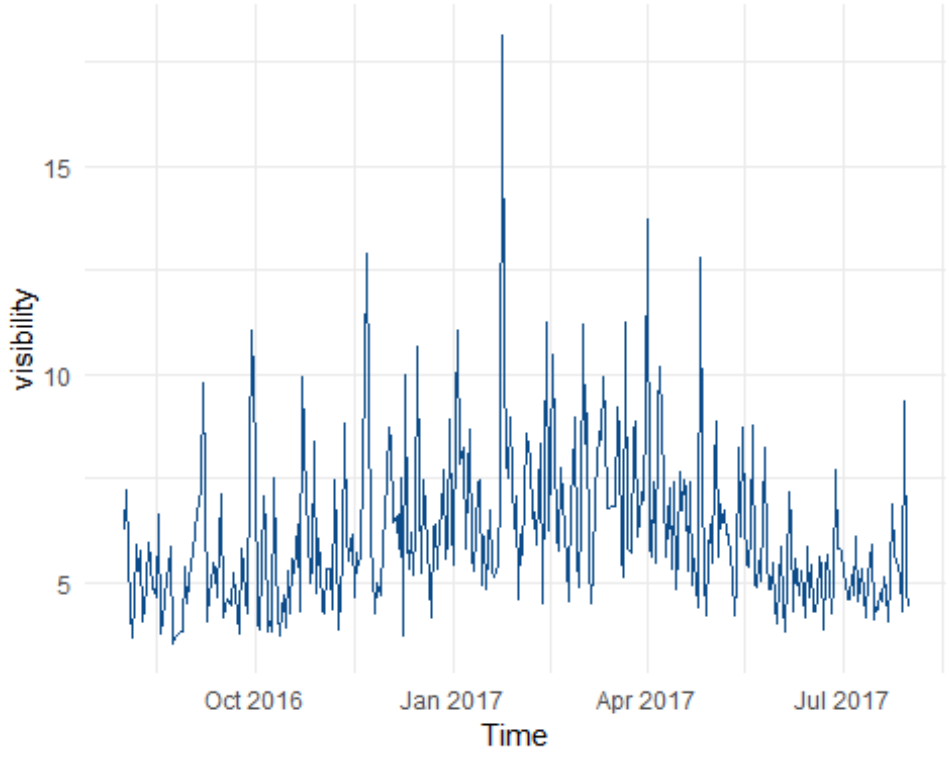

```
ggplot(data = agg_humidity, aes(newdate,humidity)) + geom_line(color =
"#0c4c8a") + theme_minimal() + xlab("Time")
```



```
ggplot(data = agg_windspeed, aes(newdate,windspeed)) + geom_line(color =
"#0c4c8a") + theme_minimal() + xlab("Time")
```

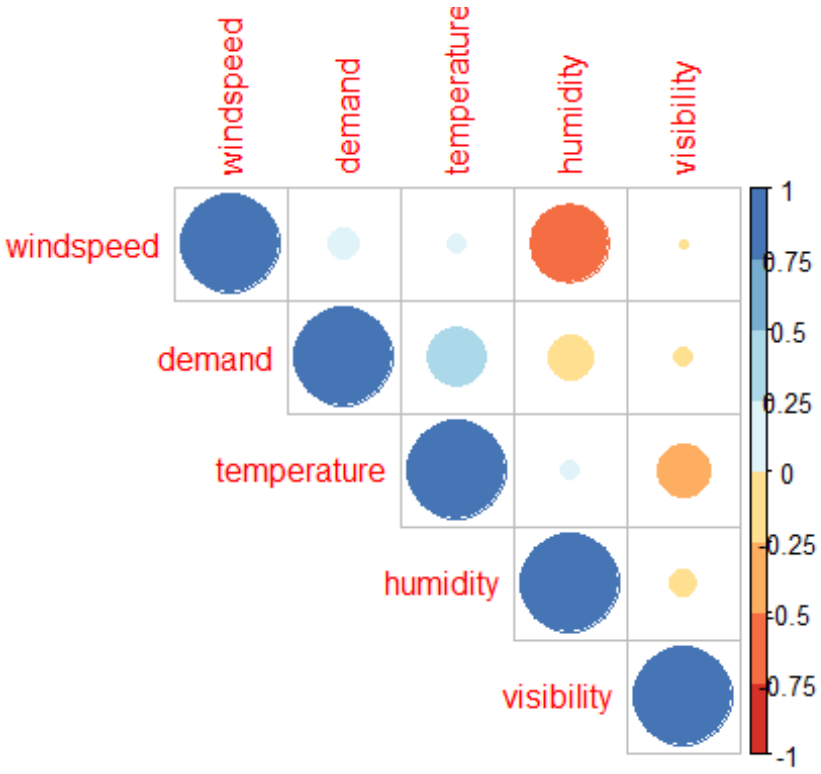



```
ggplot(data = agg_visibility, aes(newdate,visibility)) + geom_line(color =
"#0c4c8a") + theme_minimal() + xlab("Time")
```



The above charts say that windspeed and humidity are independent of time whereas temperature starts falling from October till March and starts rising from April so we can say that temperature follows same pattern as demand and it will have highest impact on demand.

```
#Calculating correlation Matrix
library(corrplot)

## corrplot 0.84 loaded

library(RColorBrewer)
corr_data <- mydata[,-c(1,8)]
corr_data <-
corr_data[!is.na(corr_data$demand)&!is.na(corr_data$temperature)&!is.na(corr_
data$humidity)&!is.na(corr_data$windspeed)&!is.na(corr_data$visibility)&!is.n
a(corr_data$condition),]
M <-cor(corr_data[,-6])
corrplot(M, type="upper", order="hclust", col=brewer.pal(n=8, name="RdYlBu"))
```



From the above plot we can say that windspeed and humidity are highly correlated whereas demand i.e the output variable is highly coorelated with temperature,humidity and weakly coorelated with visibility and windspeed.