# Applied Regression - Problem H1

Harshita Agarwala

October 31, 2017

Importing important libraries to use

```
library("tidyr", lib.loc="~/R/win-library/3.4")
library("ggplot2", lib.loc="~/R/win-library/3.4")

library("readr", lib.loc="~/R/win-library/3.4")
library("tidyverse", lib.loc="~/R/win-library/3.4")
```

```
## Loading tidyverse: tibble
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
LAozone <- read_csv("C:/Users/MAX/TUM/Applied Regression/LAozone.csv",col_names=TRUE)
```

```
## Parsed with column specification:
## cols(
##   ozone = col_integer(),
##   vh = col_integer(),
##   wind = col_integer(),
##   humidity = col_integer(),
##   temp = col_integer(),
##   ibh = col_integer(),
##   dpg = col_integer(),
##   ibt = col_integer(),
##   vis = col_integer(),
##   doy = col_integer(),
##   id = col_integer()
## )
```

a) Summarizing

```
summary(LAozone[,2:10])
```

```
##        vh            wind           humidity          temp      
##  Min.   :5320   Min.   : 0.000   Min.   :19.00   Min.   :25.00  
##  1st Qu.:5690   1st Qu.: 3.000   1st Qu.:47.00   1st Qu.:51.00  
##  Median :5760   Median : 5.000   Median :64.00   Median :62.00  
##  Mean   :5750   Mean   : 4.891   Mean   :58.13   Mean   :61.75  
##  3rd Qu.:5830   3rd Qu.: 6.000   3rd Qu.:73.00   3rd Qu.:72.00  
##  Max.   :5950   Max.   :21.000   Max.   :93.00   Max.   :93.00  
##       ibh              dpg              ibt             vis       
##  Min.   : 111.0   Min.   :-69.00   Min.   :-25.0   Min.   :  0.0  
##  1st Qu.: 877.5   1st Qu.: -9.00   1st Qu.:107.0   1st Qu.: 70.0  
##  Median :2112.5   Median : 24.00   Median :167.5   Median :120.0  
##  Mean   :2572.9   Mean   : 17.37   Mean   :161.2   Mean   :124.5
```

```
##  3rd Qu.:5000.0   3rd Qu.: 44.75   3rd Qu.:214.0   3rd Qu.:150.0
##  Max.   :5000.0   Max.   :107.00   Max.   :332.0   Max.   :350.0
##       doy
##  Min.   :  3.00
##  1st Qu.: 90.25
##  Median :177.50
##  Mean   :181.73
##  3rd Qu.:275.75
##  Max.   :365.00
```

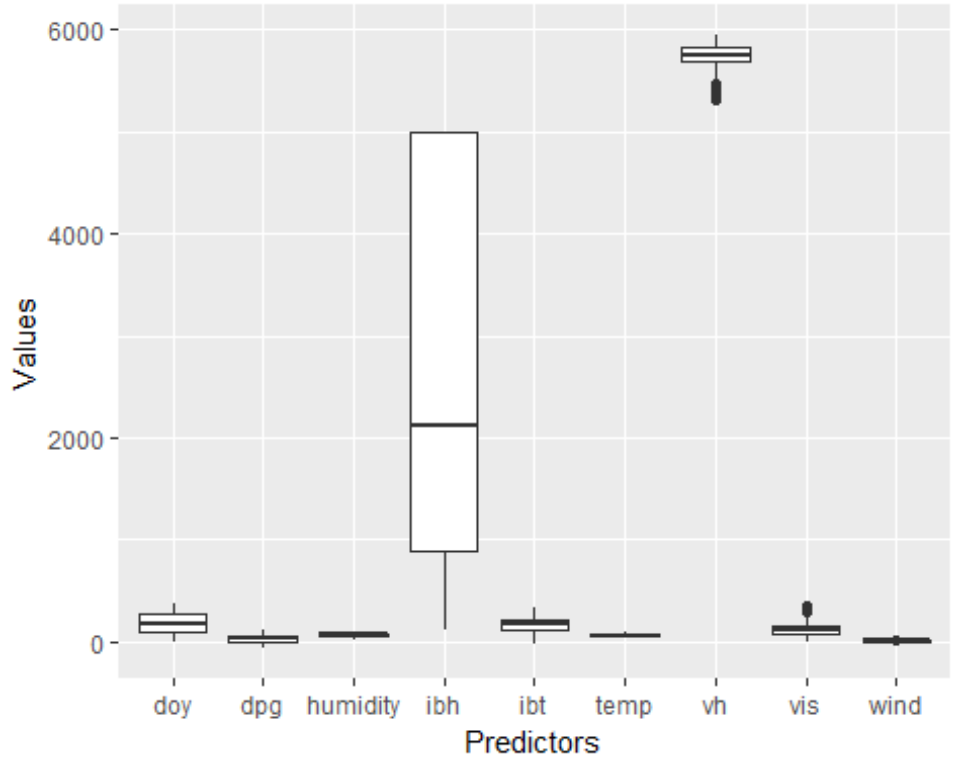
b)Converting to Long format

```
LAozone_long<-gather(LAozone,variable,value,vh:doy,factor_key=FALSE)
LAozone_long[2:4]
```

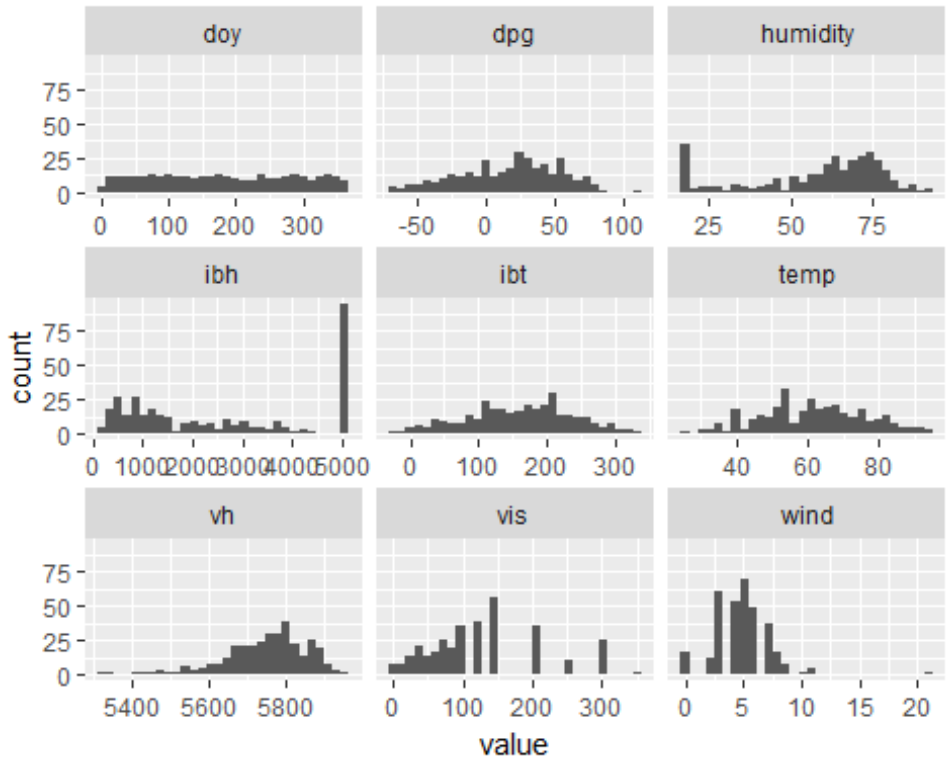
```
## # A tibble: 2,970 x 3
##       id variable value
##    <int>    <chr> <int>
##  1     1       vh  5710
##  2     2       vh  5700
##  3     3       vh  5760
##  4     4       vh  5720
##  5     5       vh  5790
##  6     6       vh  5790
##  7     7       vh  5700
##  8     8       vh  5700
##  9     9       vh  5770
## 10    10       vh  5720
## # ... with 2,960 more rows
```

c)Boxplots and Histograms

```
ggplot(LAozone_long,aes(variable,value))+geom_boxplot()+labs(x="Predictors",y="Values")
```

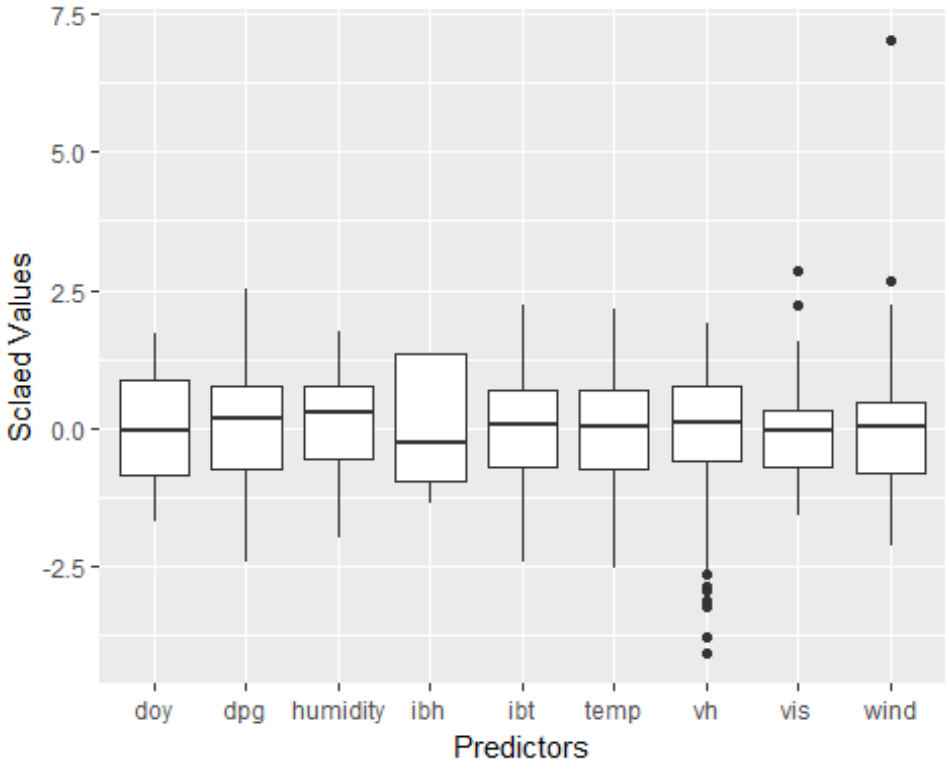


```
ggplot(LAozone_long,aes(value))+facet_wrap(~variable,scales='free_x')+geom_histogram()
```




d)Scaling and Plotting

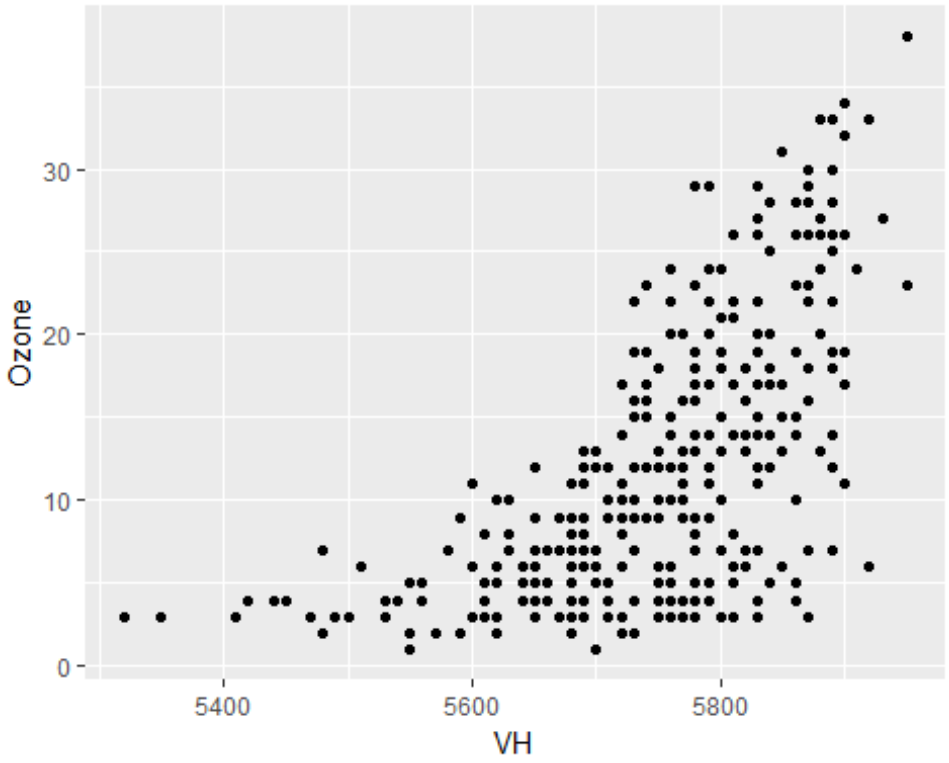
```
LAozone_scale<-scale(LAozone,center=TRUE,scale=TRUE)
LAozone_new<-as.data.frame(LAozone_scale)
LAozone_long2<-gather(LAozone_new,variable,value,vh:doy,factor_key=FALSE)
ggplot(LAozone_long2,aes(variable,value))+geom_boxplot()+labs(x="Predictors",y="Sclaed Values")
```



The most skewed variable is IBH. It is a Positive skew which means that most number of values of IBH are below the mean. As in a positive skew, the Mean >= Median > Mode

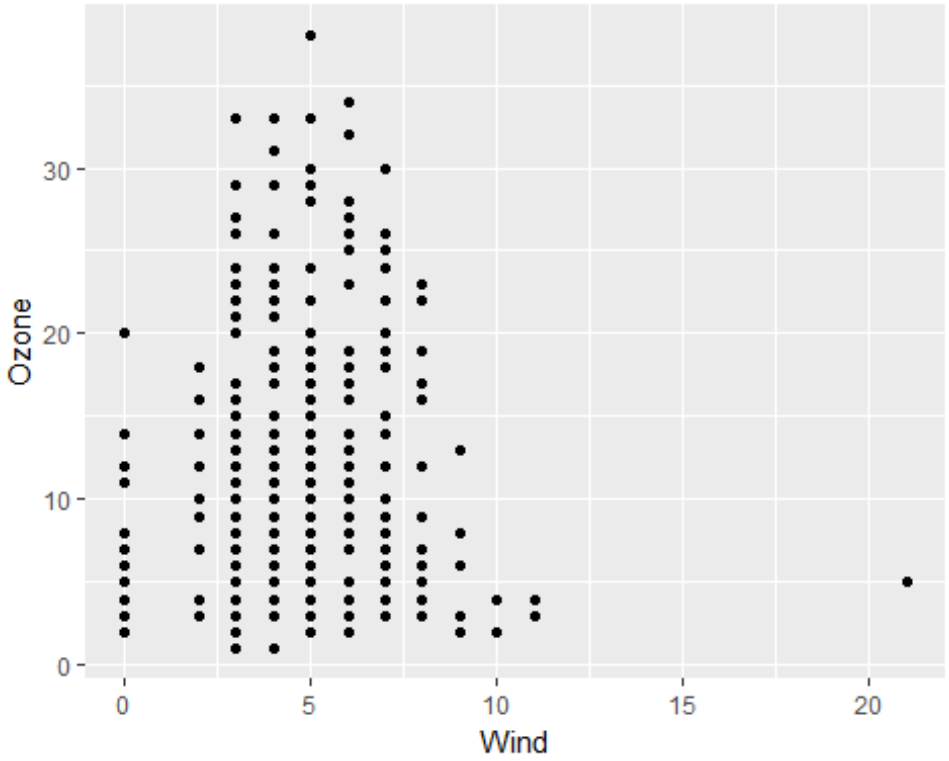
e)Scatter Plots of each Predictor variable with response

```
ggplot(LAozone,aes(vh,ozone))+geom_point()+labs(x="VH",y="Ozone")
```



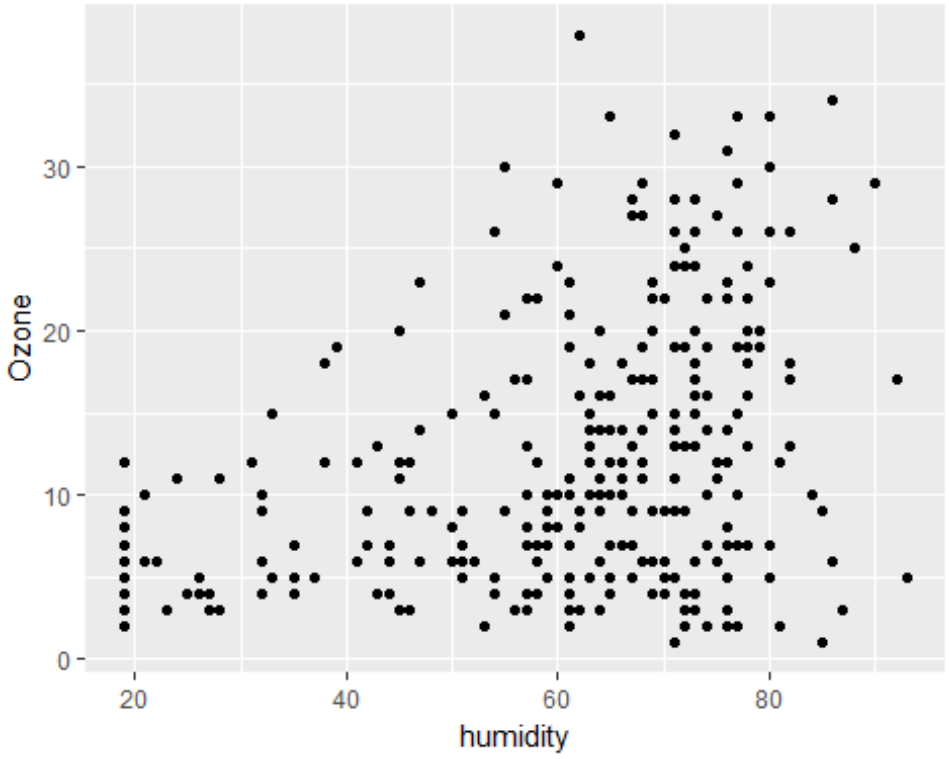

It is clear from the scatter plot that the value of ozone slowly increases with increase in VH till VH is about 5650. Beyond this the value of ozone increases rapidly with increase in VH.

```
ggplot(LAozone,aes(wind,ozone))+geom_point()+labs(x="Wind",y="Ozone")
```



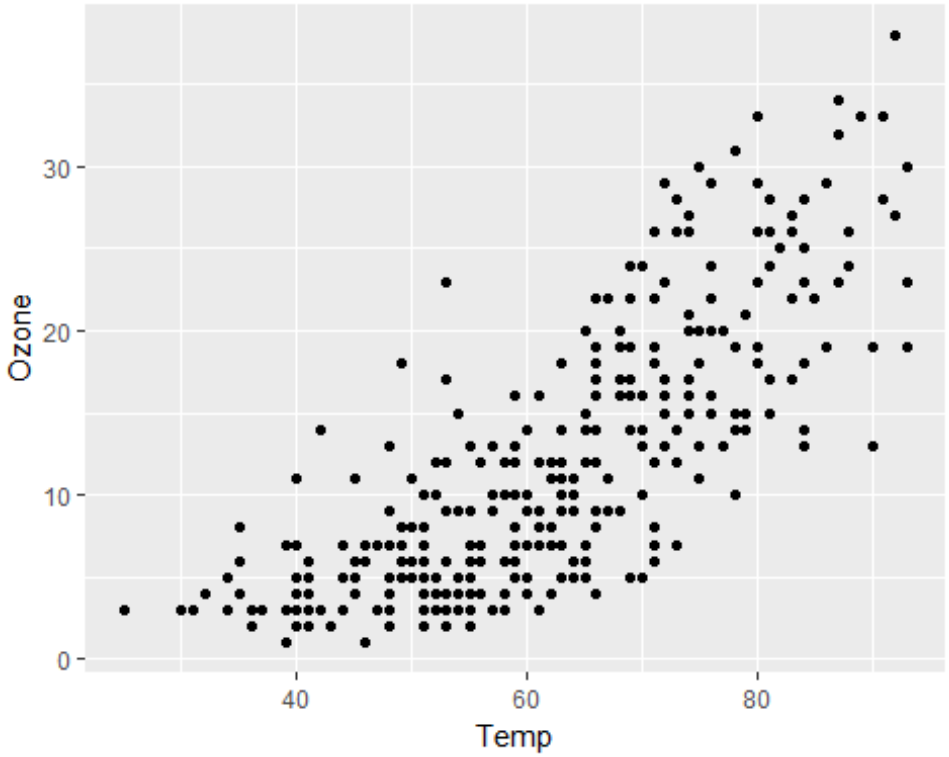

There is no clear relation between wind and ozone as can be seen from the graph. However, the value of ozone varies between [0,40] units when the wind is between [0,11] units and for a higher value of wind the ozone level is zero.

```
ggplot(LAozone,aes(humidity,ozone))+geom_point()+labs(x="humidity",y="Ozone")
```



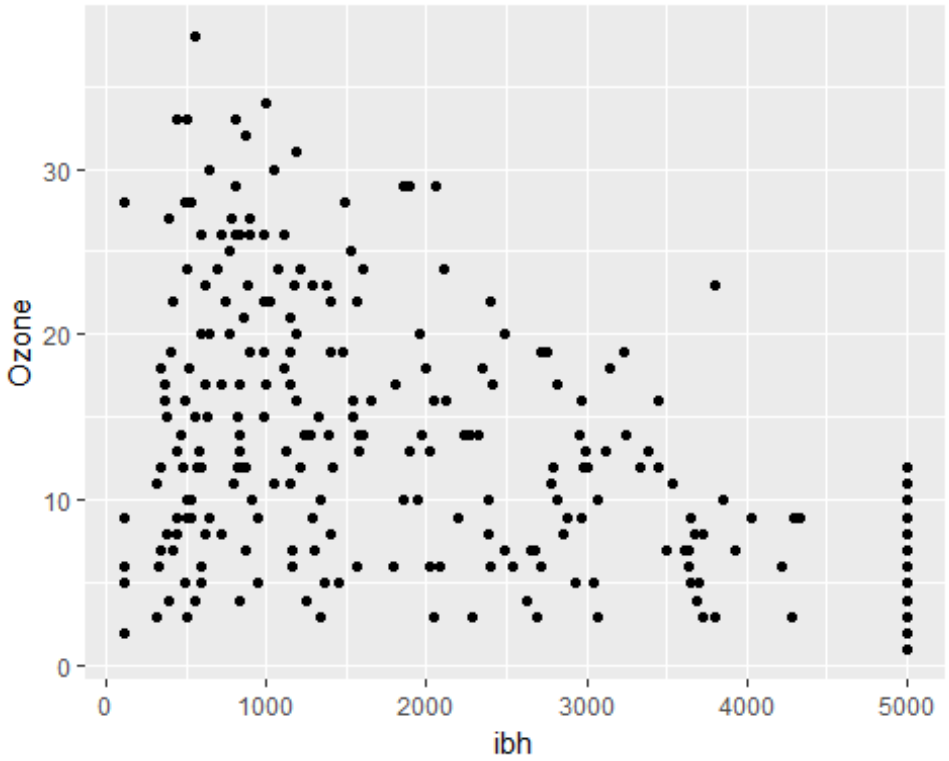

It can been observed from the scatter plot that the value of ozone slowly increases with increase in humidity till humidity is about 50 units. Beyond this the value of ozone increases rapidly with increase in humidity. However there many exceptions and outliers in this case.

```
ggplot(LAozone,aes(temp,ozone))+geom_point()+labs(x="Temp",y="Ozone")
```




The value of ozone is almost stable when temperature is below 50 units. Beyond this the value of ozone increases with increase in Temp.

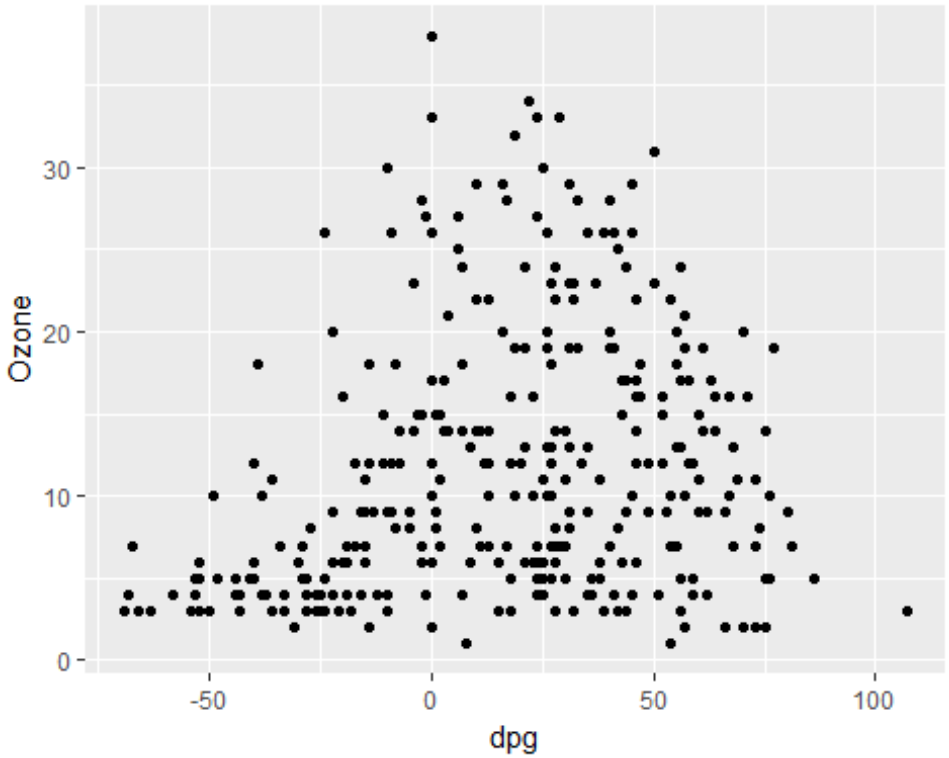
```
ggplot(LAozone,aes(ibh,ozone))+geom_point()+labs(x="ibh",y="Ozone")
```




We can observe from the scatter plot that the value of ozone ranges between [0,35] units when ibh is between [0,1000] units and the upper limit of this range slowly decrease with increase in the value of ibh.

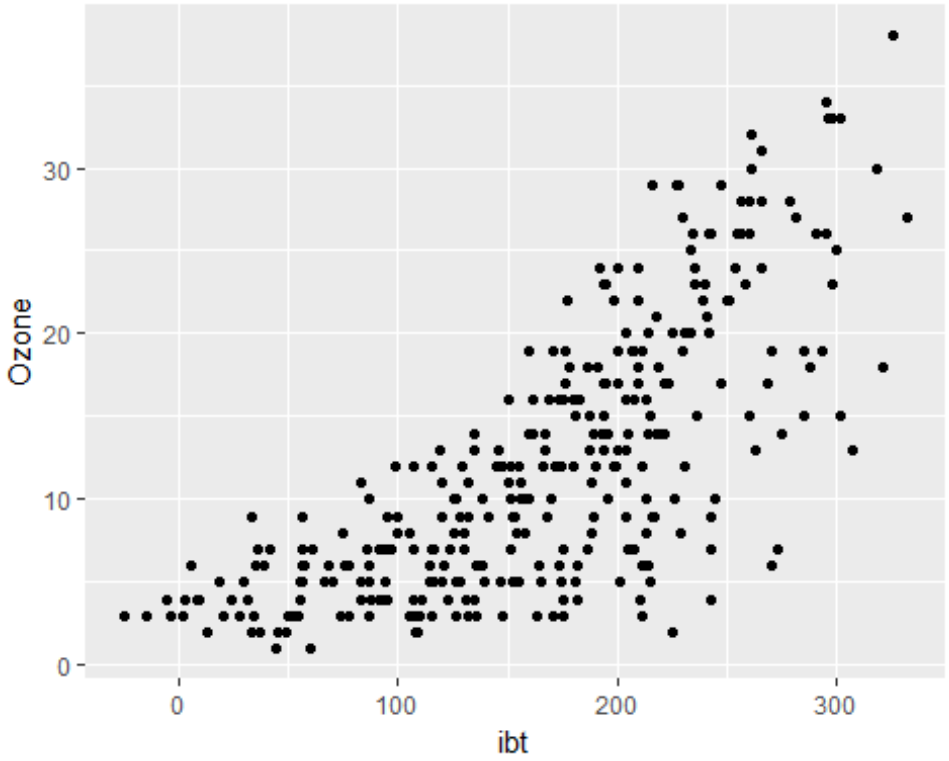
It becomes negligible when the ibh is around 4000 units but suddenly shoots up when ibh is 5000. Again, many outliers and exceptions.

```
ggplot(LAozone,aes(dpg,ozone))+geom_point()+labs(x="dpg",y="Ozone")
```



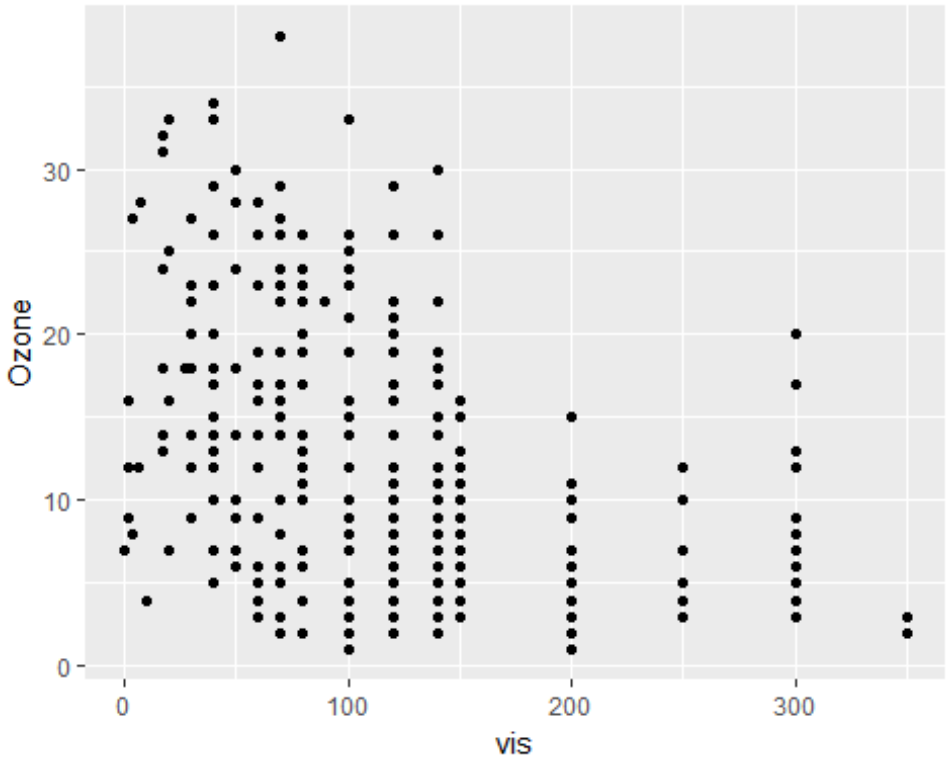

It is clear from the scatter plot that the value of ozone remains stable when the value of dpg is below -25 units. It increases rapidly when the dpg is between [-25,25] units and then starts to drop for higher values.

```
ggplot(LAozone,aes(ibt,ozone))+geom_point()+labs(x="ibt",y="Ozone")
```



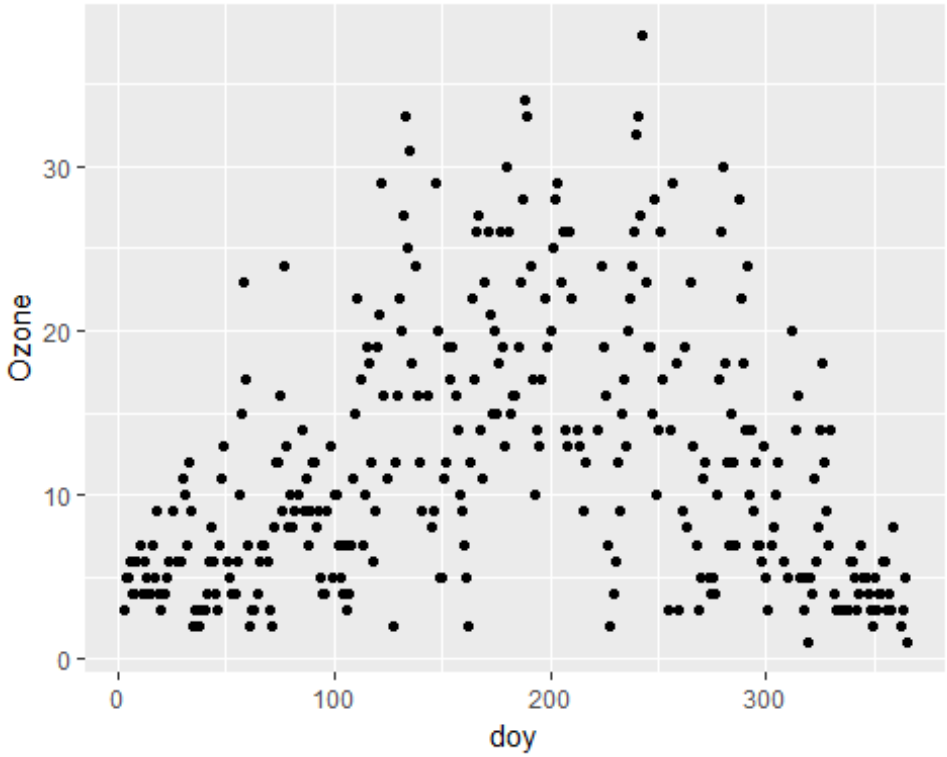

It is clear from the scatter plot that the value of ozone slowly increases with increase in ibt till ibt is about 125 units. Beyond this the value of ozone increases rapidly with increase in ibt.

```
ggplot(LAozone,aes(vis,ozone))+geom_point()+labs(x="vis",y="Ozone")
```




There is no clear relation between vis and ozone. Ozone ranges in [0,40] when vis is between [0,150] units and in [0,20] when vis more than 150.

```
ggplot(LAozone,aes(doy,ozone))+geom_point()+labs(x="doy",y="Ozone")
```




The value of ozone increases rapidly with increase in doy till doy reaches 130 units and then it stabalizes till doy is 220. Beyond this it decreases with icrease in doy. However, there are many outliers in this relation.

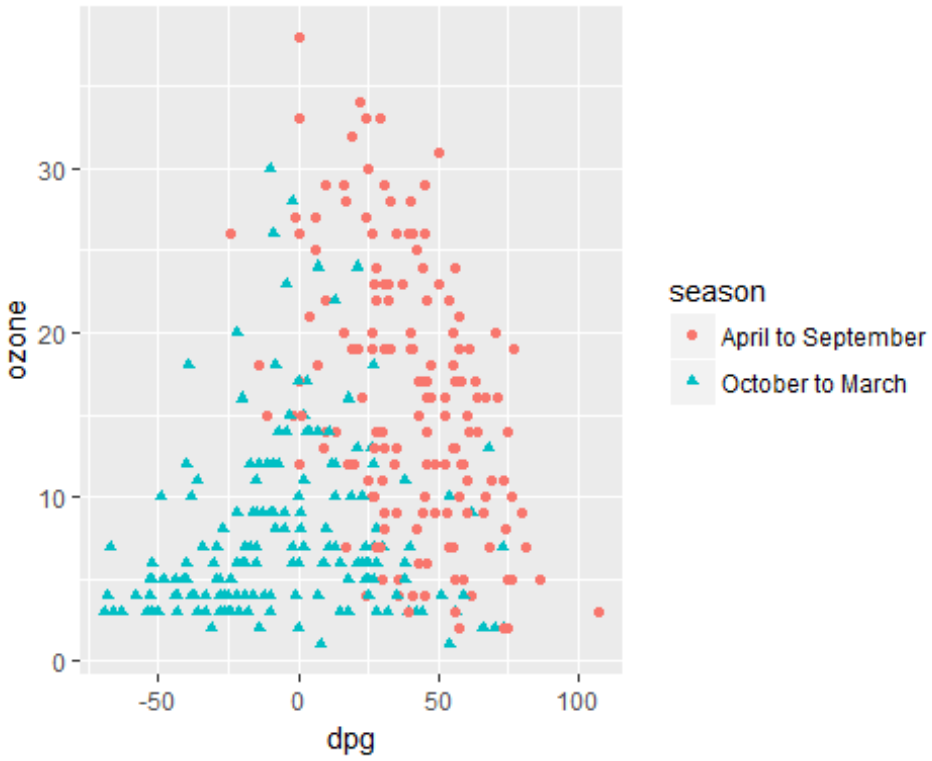
f) Introducing Season column and plotting Ozone vs dpg with season classification.

```
add_column(LAozone,'season'=NA)
```

```
## # A tibble: 330 x 12
##    ozone    vh  wind humidity  temp   ibh   dpg   ibt   vis   doy    id
##    <int> <int> <int>    <int> <int> <int> <int> <int> <int> <int> <int>
##  1     3  5710     4       28    40  2693   -25    87   250     3     1
##  2     5  5700     3       37    45   590   -24   128   100     4     2
##  3     5  5760     3       51    54  1450    25   139    60     5     3
##  4     6  5720     4       69    35  1568    15   121    60     6     4
##  5     4  5790     6       19    45  2631   -33   123   100     7     5
##  6     4  5790     3       25    55   554   -28   182   250     8     6
##  7     6  5700     3       73    41  2083    23   114   120     9     7
##  8     7  5700     3       59    44  2654    -2    91   120    10     8
##  9     4  5770     8       27    54  5000   -19    92   120    11     9
## 10     6  5720     3       44    51   111     9   173   150    12    10
## # ... with 320 more rows, and 1 more variables: season <lgl>
```

```
for (i in c(1:330)){
  if(LAozone[i,'doy']>=91 && LAozone[i,'doy']<=273){LAozone[i,'season']<-'April to Septem
ber'}}
for (i in c(1:330)){
  if(is.na(LAozone[i,'season'])){LAozone[i,'season']<-'October to March'}}

ggplot(LAozone, aes(x=dpg, y=ozone)) + geom_point(aes(shape=season,color=season))
```




The following observations can be made against the graph in section (e): 1. The value of ozone remains stable when the value of dpg is below -25 units. 2. It increases rapidly when the dpg is between [-25,25] units and then starts to drop for higher values. 3. Moreover with the inclusion of season classes, the value of dpg is is mostly in the range [0,100] for the season 'April to September' and [-75,50] for 'October to March' season. 4. Also, the value of ozone is varied and average is higher for the season 'April to September' and less varied with a lower average for the 'October to March' season.