Air Quality Analysis

Group 4

8/4/2021

Introduction

For our assignment, we have selected the "NewYork Air Quality" dataset from https://www.kaggle.com/mfaisalqureshi/newyork-air-quality. This data set has daily air quality measurements from May to September (5 months). The variables in our data set are Ozone, Solar.R, Wind, Temp, Month, and Day. The total number of rows in the dataset is 153.

Load libraries & import the data

```
library(tidyverse)
Air_Quality<-read.csv(file="airquality.csv")
```

```
str(Air_Quality) #Print the structure
```

```
## 'data.frame':    153 obs. of  7 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
names(Air_Quality) #List the variables
```

```
## [1] "X"       "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

```
head(Air_Quality, 15) #Print the top 15 rows
```

```
##     X Ozone Solar.R Wind Temp Month Day
## 1   1    41     190  7.4   67     5   1
## 2   2    36     118  8.0   72     5   2
## 3   3    12     149 12.6   74     5   3
## 4   4    18     313 11.5   62     5   4
## 5   5    NA      NA 14.3   56     5   5
## 6   6    28      NA 14.9   66     5   6
## 7   7    23     299  8.6   65     5   7
```

```
## 8   8    19       99 13.8   59     5   8
## 9   9     8       19 20.1   61     5   9
## 10 10    NA      194  8.6   69     5  10
## 11 11     7       NA  6.9   74     5  11
## 12 12    16      256  9.7   69     5  12
## 13 13    11      290  9.2   66     5  13
## 14 14    14      274 10.9   68     5  14
## 15 15    18       65 13.2   58     5  15
```

User-Defined Function

```r
square_of_solar<-function(){
  Air_Quality2<-Air_Quality
  (Air_Quality2$Solar.R)^2
}

square_of_solar()
```

```
##   [1]  36100  13924  22201  97969     NA     NA  89401   9801    361  37636
##  [11]     NA  65536  84100  75076   4225 111556  94249   6084 103684   1936
##  [21]     64 102400    625   8464   4356  70756     NA    169  63504  49729
##  [31]  77841  81796  82369  58564  34596  48400  69696  16129  74529  84681
##  [41] 104329  67081  62500  21904 110224 103684  36481  80656   1369  14400
##  [51]  18769  22500   3481   8281  62500  18225  16129   2209   9604    961
##  [61]  19044  72361  61504  55696  10201  30625  98596  76176  71289  73984
##  [71]  30625  19321  69696  30625  84681   2304  67600  75076  81225  34969
##  [81]  48400     49  66564  87025  86436  49729   6561   6724  45369  75625
##  [91]  64009  64516   6889    576   5929     NA     NA     NA  65025  52441
## [101]  42849  49284  18769  36864  74529  24649   4096   5041   2601  13225
## [111]  59536  36100  67081   1296  65025  44944  56644  46225  23409  41209
## [121]  50625  56169  35344  27889  38809  33489  35721   9025   8464  63504
## [131]  48400  52900  67081  55696  67081  56644    576  12544  56169  50176
## [141]    729  56644  40401  56644    196  19321   2401    400  37249  21025
## [151]  36481  17161  49729
```

Filter rows

```r
Air_Quality<-filter(Air_Quality, Air_Quality$Wind<10)
```

Independent Variables: Ozone, Solar.R, Wind, Temp

Dependent Variable: Day

Unused Variable: X

```
Air_Quality <- cbind(Air_Quality$Ozone,
                     Air_Quality$Solar.R, Air_Quality$Wind,
                     Air_Quality$Temp)
Air_Quality = as.data.frame(Air_Quality)
```

Remove missing values & duplicate rows

```
Air_Quality<-na.omit(Air_Quality)
Air_Quality %>% distinct()
```

```
head(Air_Quality, 15)
```

```
##      V1  V2  V3 V4
## 1    41 190 7.4 67
## 2    36 118 8.0 72
## 3    23 299 8.6 65
## 6    16 256 9.7 69
## 7    11 290 9.2 66
## 8    11  44 9.7 62
## 9     1   8 9.7 59
## 10    4  25 9.7 61
## 12  115 223 5.7 79
## 13   37 279 7.4 76
## 18   29 127 9.7 82
## 21   23 148 8.0 82
## 22   20  37 9.2 65
## 30  135 269 4.1 84
## 31   49 248 9.2 85
```

Rename columns

```
Air_Quality<-rename(Air_Quality, Ozone=V1, Solar_Rad=V2, Wind=V3, Temperature=V4)
```

Reorder rows in descending order

```
Air_Quality %>% arrange(desc(Air_Quality$Ozone))
```

Add new variables

```
Air_Quality$Double_Wind = (Air_Quality$Wind)*2
Air_Quality$Half_Ozone = (Air_Quality$Ozone)/2
head(Air_Quality, 8)
```

```
##    Ozone Solar_Rad Wind Temperature Double_Wind Half_Ozone
## 1     41       190  7.4          67        14.8       20.5
## 2     36       118  8.0          72        16.0       18.0
## 3     23       299  8.6          65        17.2       11.5
## 6     16       256  9.7          69        19.4        8.0
## 7     11       290  9.2          66        18.4        5.5
## 8     11        44  9.7          62        19.4        5.5
## 9      1         8  9.7          59        19.4        0.5
## 10     4        25  9.7          61        19.4        2.0
```

Create a training set using random number generator engine

```
set.seed(1234)
Air_Quality %>% sample_frac(0.80, replace = FALSE)
```

```
##    Ozone Solar_Rad Wind Temperature Double_Wind Half_Ozone
## 1     20        81  8.6          82        17.2       10.0
## 2     32       236  9.2          81        18.4       16.0
## 3     48       260  6.9          81        13.8       24.0
## 4     65       157  9.7          80        19.4       32.5
## 5    118       225  2.3          94         4.6       59.0
## 6     96       167  6.9          91        13.8       48.0
## 7    115       223  5.7          79        11.4       57.5
## 8     11       290  9.2          66        18.4        5.5
## 9     59        51  6.3          79        12.6       29.5
## 10    30       193  6.9          70        13.8       15.0
## 11    16       256  9.7          69        19.4        8.0
## 12    16        77  7.4          82        14.8        8.0
## 13    23       115  7.4          76        14.8       11.5
## 14    23        14  9.2          71        18.4       11.5
## 15    80       294  8.6          86        17.2       40.0
## 16    11        44  9.7          62        19.4        5.5
## 17    49       248  9.2          85        18.4       24.5
## 18   135       269  4.1          84         8.2       67.5
## 19   168       238  3.4          81         6.8       84.0
## 20    78       197  5.1          92        10.2       39.0
## 21    85       188  6.3          94        12.6       42.5
## 22    85       175  7.4          89        14.8       42.5
## 23    97       272  5.7          92        11.4       48.5
## 24    23       299  8.6          65        17.2       11.5
## 25    64       253  7.4          83        14.8       32.0
## 26    82       213  7.4          88        14.8       41.0
## 27    46       237  6.9          78        13.8       23.0
## 28    47        95  7.4          87        14.8       23.5
```

```
## 29    36       118  8.0          72         16.0       18.0
## 30    73       215  8.0          86         16.0       36.5
## 31     4        25  9.7          61         19.4        2.0
## 32   108       223  8.0          85         16.0       54.0
## 33   110       207  8.0          90         16.0       55.0
## 34    73       183  2.8          93          5.6       36.5
## 35    23       148  8.0          82         16.0       11.5
## 36   122       255  4.0          89          8.0       61.0
## 37    24       259  9.7          73         19.4       12.0
## 38    97       267  6.3          92         12.6       48.5
## 39    91       189  4.6          93          9.2       45.5
## 40    18       131  8.0          76         16.0        9.0
## 41    37       279  7.4          76         14.8       18.5
## 42    29       127  9.7          82         19.4       14.5
## 43    50       275  7.4          86         14.8       25.0
## 44    28       238  6.3          77         12.6       14.0
## 45    76       203  9.7          97         19.4       38.0
## 46    41       190  7.4          67         14.8       20.5
```

Calculate descriptive statistics

```
summary(Air_Quality)
```

```
##      Ozone          Solar_Rad          Wind        Temperature
##  Min.   :  1.00   Min.   :  7.0   Min.   :2.300   Min.   :59.00
##  1st Qu.: 23.25   1st Qu.:135.2   1st Qu.:6.300   1st Qu.:76.00
##  Median : 49.50   Median :205.0   Median :7.400   Median :82.00
##  Mean   : 57.29   Mean   :187.3   Mean   :7.286   Mean   :81.09
##  3rd Qu.: 81.50   3rd Qu.:253.8   3rd Qu.:9.050   3rd Qu.:87.75
##  Max.   :168.00   Max.   :299.0   Max.   :9.700   Max.   :97.00
##   Double_Wind      Half_Ozone
##  Min.   : 4.60   Min.   : 0.50
##  1st Qu.:12.60   1st Qu.:11.62
##  Median :14.80   Median :24.75
##  Mean   :14.57   Mean   :28.65
##  3rd Qu.:18.10   3rd Qu.:40.75
##  Max.   :19.40   Max.   :84.00
```

```
mean(Air_Quality$Ozone)
```

```
## [1] 57.2931
```

```
median(Air_Quality$Ozone)
```

```
## [1] 49.5
```

```
range(Air_Quality$Ozone)
```

```
## [1]   1 168
```

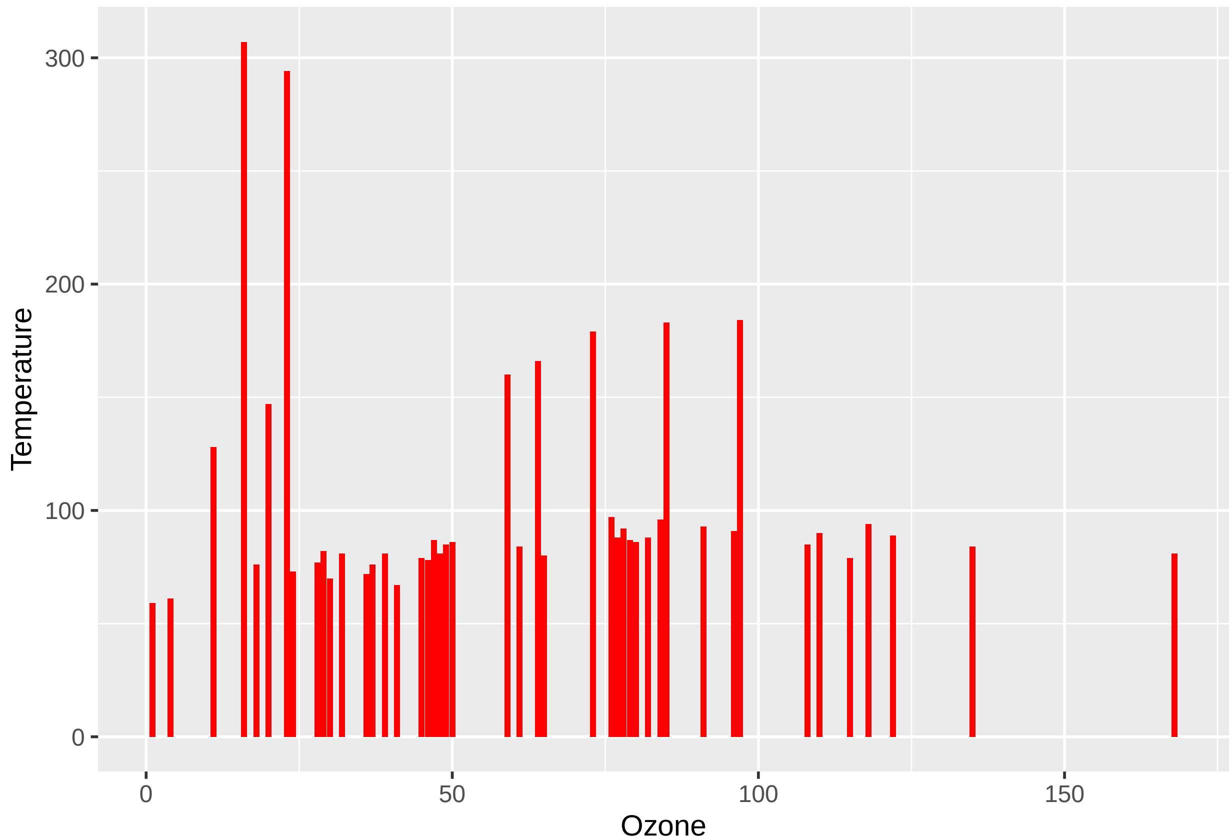
User-defined mode function

```
user_mode<-function(x){
  modeVal<-unique(x)

  #Match returns a vector of the positions of the first
  #matches of its arguments
  modeVal[which.max(tabulate(match(x, modeVal)))]
}

user_mode(Air_Quality$Ozone)
```

```
## [1] 23
```

Bar Plot

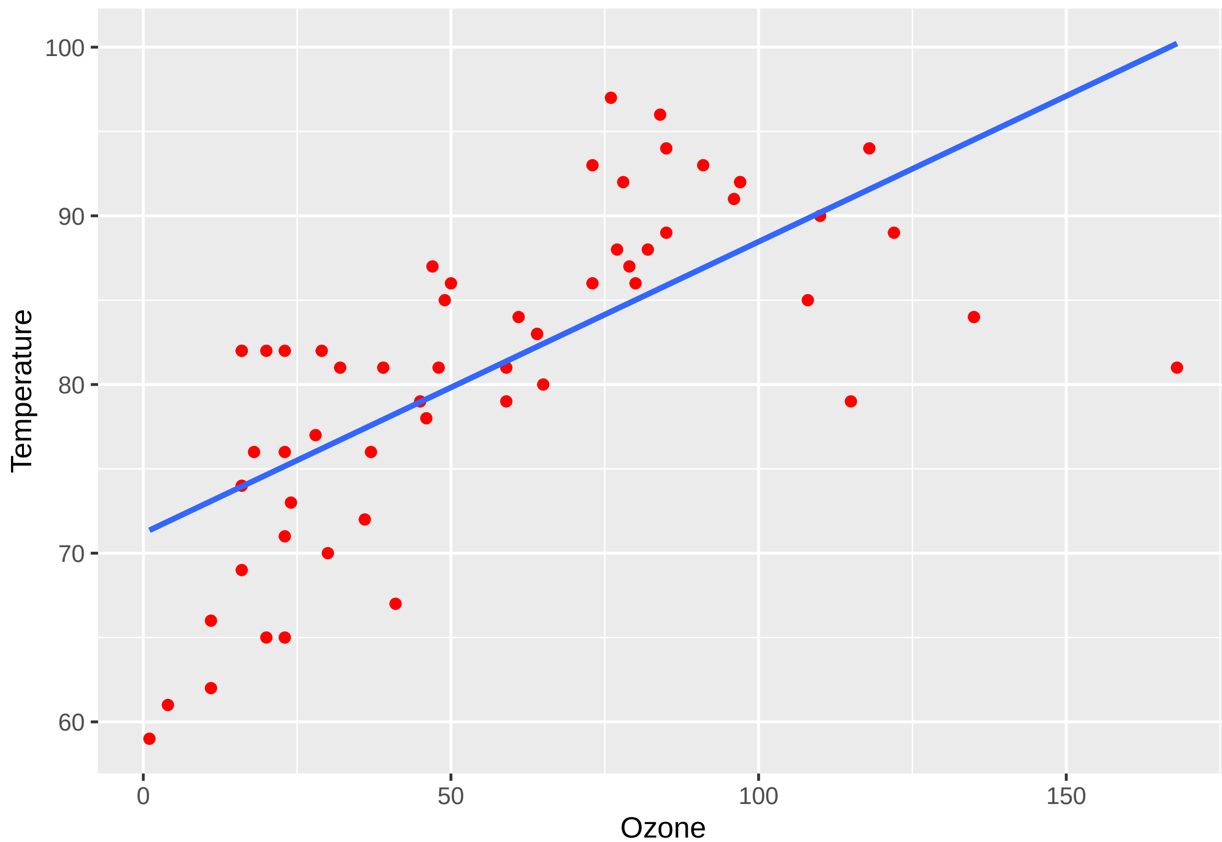
```
#Tell geom_bar that y-values will be provided
ggplot(data=Air_Quality, aes(x=Ozone, y=Temperature)) +
  geom_bar(stat="identity", fill="red") +
  labs(y="Temperature")
```



Scatter Plot

```
#Turn off confidence intervals
ggplot(data=Air_Quality, aes(x=Ozone, y=Temperature)) +
  geom_point(color="red") + labs(y="Temperature") +
  geom_smooth(method='lm', se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Calculate Pearson correlation

```
cor(Air_Quality$Ozone, Air_Quality$Temperature, method="pearson")
```

```
## [1] 0.6898136
```

Conclusion

Based on our analysis, there is a correlation between Ozone & Temperature. From the bar plot it can be seen that the temperature reaches its maximum around 25 for ozone. The scatter plot shows an exponential relationship between temperature & ozone.

Github Link: https://github.com/SkySpartan/BUS-4064-Assignment-1.git