Homework 1 Quarto Test

Code

AUTHOR
Seán Cassidy 21212291

PUBLISHED
June 6, 2024

Analysis of Air Quality in New York City

Identification of Interesting Relationship between Air Quality Variables

For this short report we want to investigate the airquality dataset, which looks at a number of air quality measures taken in ***New York City*** in **1973**. The dataset is useful to see if air quality is Bad or Good but here we want to explore a significant relationship between 2 of our variables.

To begin we want to examine the dataset for an understanding of its structure.

```
'data.frame':   153 obs. of  6 variables:
 $ 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 ...
```

We can see that our dataset is made up of 6 variables out of 153 observations, with all the variables being of integer type except Wind which is a continuous numeric variable. More information on this dataset can be found at link. With an understanding of our structure we can summarise the key statistics for each variable.

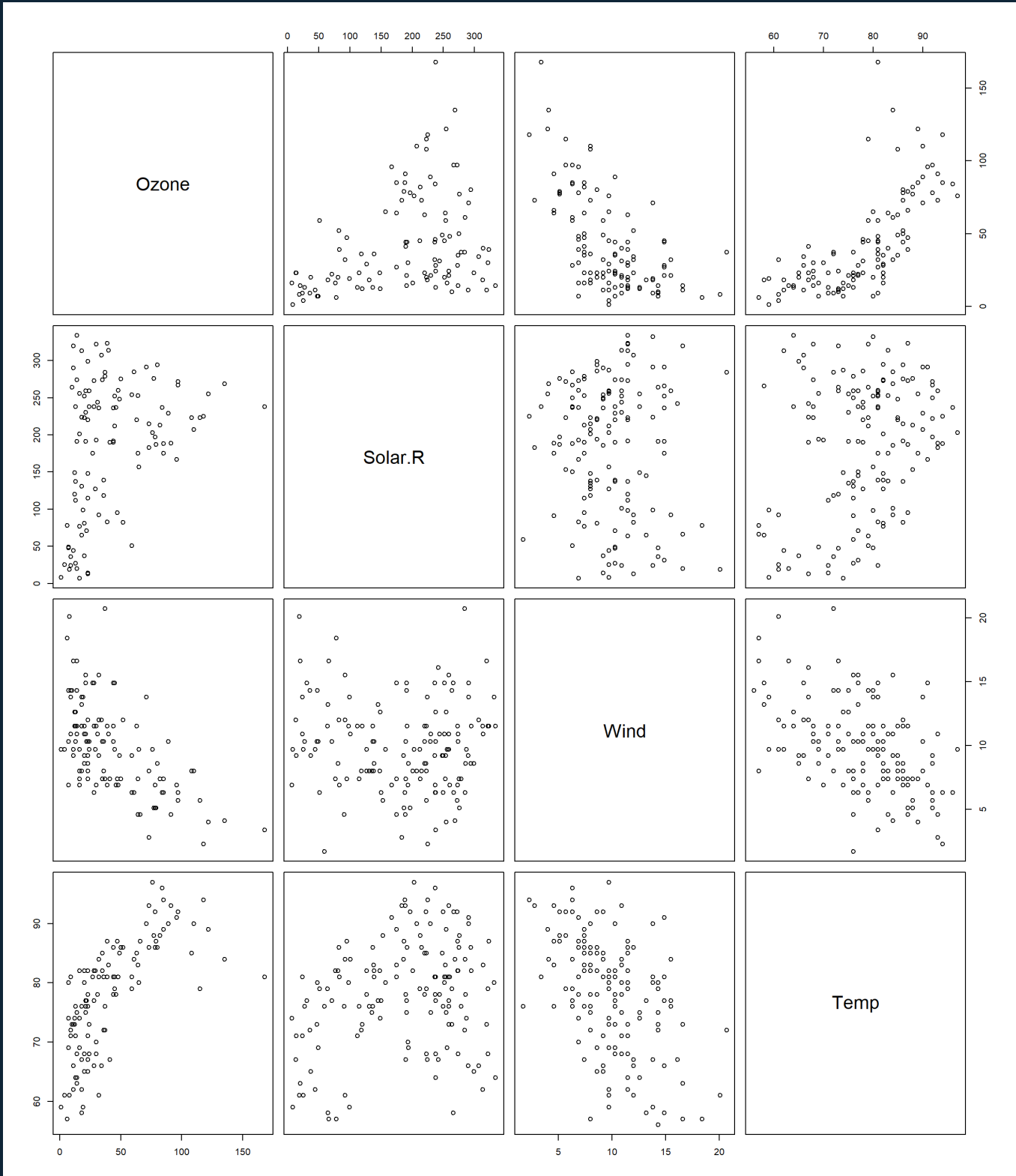
```
     Ozone           Solar.R          Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7                                       
     Month            Day      
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  
```

From the summary we can see that the different measures for the variables have a high range of difference in terms of their values, indicating that the variables are across different scales. Day and Month are both

indicator variables so the summary statistics offer little value to us for these variables. We will therefore focus on the other 4 variables as part of our relationship analysis.

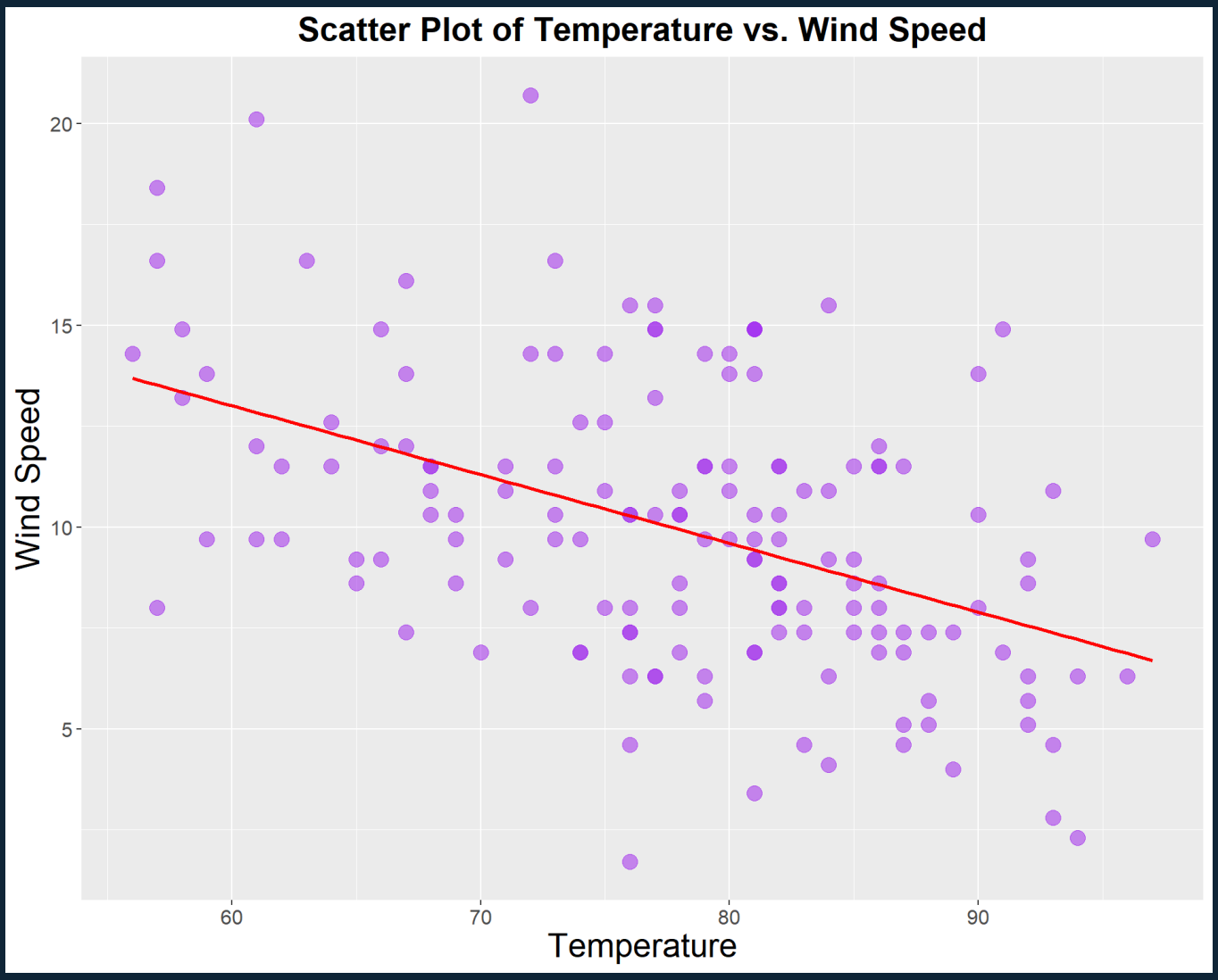
Air Quality Plots

To first establish which our four variables have a relationship that would be most interesting to examine in more detail, we can plot the four variables together in scatter plot to examine each relationship between all variables.



Air Quality data

There are a number of interesting possible relationships to select but we will look further into the relationship between wind and temperature. Logically we would assume that as weather is hotter that the level of wind will be lower, and this does appear to be the case as a strong negative relationship exists between our variables, but lets examine this in more detail.



Wind and Temperature Relationship Analysis

From the scatter plot and linear regression line it is clear there is a negative relationship between wind and temperature, which we can interpret as wind speed is likely to decrease as temperature increases. Interestingly from our plot we can note a number of significant outlier values from the regression line which highlights that there are occasions present in the data where there are high incidents of wind with high temperature and vice versa.