


Statistics Assignment 1:

Introduction to R and using R to calculate various statistics of the dataset "Wind.csv".

Creating a variable

After loading up and importing the given dataset "wind.csv" I input the following command **speed<-c(26, 41, 31, 20, 56, 59)** to create a variable "speed" to store values of the maximum windspeed for the first 6 days in the dataset. The resultant output is as below.

Data	
▶ wind	32 obs. of 4 variables 
Values	
speed	num [1:6] 26 41 31 20 56 59

In this photo, the dataset "wind" has been included for clarity. Below this, there's a values tab which indicates my variable speed has been inputted successfully and points to the values of the maximum windspeed for the first 6 days of the dataset.

Finding mean value of variable "speed"

Now that a variable has been added, R is capable of finding statistical values such as the mean. This is important in determining the average windspeed in our sample of 6 days. Inputting the command **mean(speed)** gave me a value of **38.83333** as shown

```
> mean(speed)
[1] 38.83333
```

Viewing the wind dataset

Being able to see my dataset is crucial when it comes to creating statistical inferences, altering the dataset, and picking specific parts of the dataset to work on. The command **View(wind)** provided me with this representation of data to the right.

The groups on the top are my columns which consist of "Day", "Direction", "Speed", and "Temperature". The observed values in this sample dataset will help me draw conclusions and find relationships between one and the other variables.

Finding the mean windspeed of all observed days

Now that I can see my data, I'm curious as to what the average windspeed is for not just 6 values; but for all. To do this, I inputted the following-
mean(wind\$Speed) and found that the average windspeed for every given day is **40.53125**.

```
> mean(wind$Speed)
[1] 40.53125
```

While coming to this conclusion, I learned that R is also capital sensitive and was given an error for inputting the command with the value "speed" instead of specifying it as "Speed".

```
> mean(wind$speed)
[1] NA
Warning message:
In mean.default(wind$speed) :
  argument is not numeric or logical: returning NA
```

	Day	Direction	Speed	Temperature
1	1	E	26	17.4
2	2	E	41	16.2
3	3	S	31	17.6
4	4	E	20	17.7
5	5	NW	56	22.3
6	6	SW	59	18.6
7	7	SW	46	13.5
8	8	E	24	18.9
9	9	SW	48	19.1
10	10	SW	48	12.0
11	11	E	31	14.1
12	12	S	52	11.3
13	13	S	31	14.4
14	14	NE	41	13.9
15	15	SW	28	17.9
16	16	E	37	17.4
17	17	NE	30	20.8
18	18	NE	31	23.0
19	19	NE	31	19.5
20	20	W	52	28.4
21	21	N	50	23.8
22	22	E	61	24.4
23	23	W	48	10.8
24	24	SW	33	15.2
25	25	E	37	16.8
26	26	E	41	15.3
27	27	NE	33	16.4
28	28	W	35	20.6
29	29	NW	50	23.2
30	30	S	72	20.5
31	31	SW	43	13.1
32	32	NE	31	18.5

Gathering the dimensions of the dataset

Wanting to view how large my dataset is and the number of days data has been collected is also important. To view the dimensions of the “wind” dataset, I typed the line **dim(wind)** and the resultant was this.

```
> dim(wind)
[1] 32  4
```

This shows me that my dataset contains 32 days of collected data and the 4 variable types which data sits under, which is simply enough represented as 32 rows and 4 columns.

Finding the highest daily windspeed

I now can calculate the largest recorded value of windspeed over any of the 32 days. **max(wind\$Speed)** provided me with the answer **72** which if measured in Km/h is a very impressive force.

```
> max(wind$Speed)
[1] 72
```

Showing all the different values of wind-direction

Knowing all the values of a specific variable could come in useful especially if we’re expecting certain measured values. For the wind-direction, using **levels(wind\$Direction)** allowed me to see that all measured wind-directions in this dataset are “North”, “North East”, “North West”, “East”, “South”, “South West”, and “West”. Drawing a conclusion from this output that the direction South East is not included in my data.

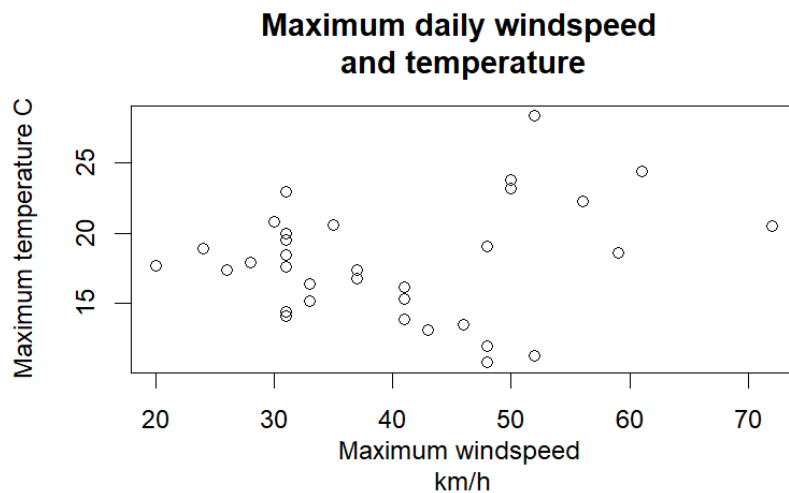
```
> levels(wind$Direction)
[1] "E"  "N"  "NE" "NW" "S"  "SW" "W"
```

Which wind-direction is usually the highest windspeed?

Interestingly enough, finding the average windspeed for every direction isn’t that difficult. Instead of having to calculate the mean windspeed for every direction individually, I can do it all at once. This proves to be an extremely efficient and powerful tool to use even more so with larger datasets. My input **tapply(wind\$Speed, wind\$Direction, mean)** made it clear that from this sample, on average North-West blows at the highest speed of 53.

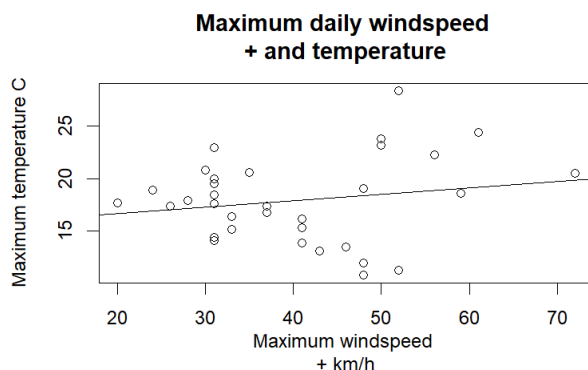
```
> tapply(wind$Speed, wind$Direction, mean)
      E      N      NE      NW      S      SW      W
35.33333 50.00000 32.83333 53.00000 46.50000 43.57143 45.00000
```

Plotting the data



Here is the data plotted using the **plot** command in R. Adding a title and labeling the x and y axes helps improve clarity. From the observed data, most maximum wind speed values appear to cluster around the 30–35 km/h range. However, this alone is not sufficient evidence to conclude that the average maximum wind speed falls within this range. Similarly, the majority of temperature observations lie between 15–20°C. Notably as wind speed increases the temperature values become more scattered and unpredictable suggesting greater variability rather than a consistent trend.

For further analysis, I created a linear model to identify any potential patterns I may have overlooked. While the data itself doesn't reveal much to me, visualizing it in graph form helps to give a clearer perspective. I plotted a best-fit line onto the original graph using the coefficient information from the linear model and output the following graph.

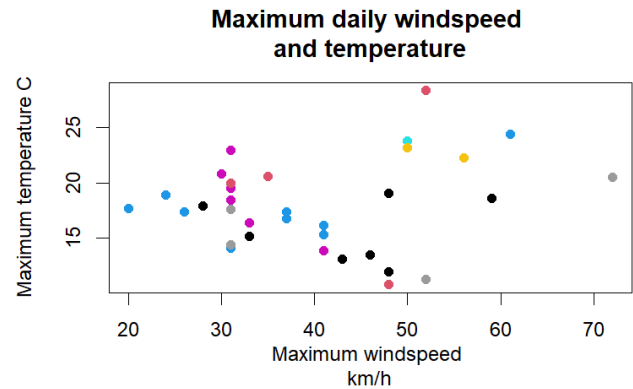


The trend line shows a weak positive slope, suggesting a possible relationship between temperature and maximum wind speed, where higher wind speeds may correspond to higher temperatures. However, the data

points are widely scattered around the trend line, making it difficult to identify a clear or strong trend with confidence.

More data analysis on plots

After plotting the graph in R with color coding to represent different wind directions, I still do not see a clear trend, though some clustering of certain wind conditions appears to be present. The use of color helps in identifying potential patterns by introducing an additional variable, providing more context for analysis. One noticeable observation is the clustering of pink and blue points. When the wind is blowing in the pink direction, it often corresponds to wind speeds around 30 km/h. Similarly, when the wind is blowing in the blue direction, temperatures are frequently in the range of 15–20°C. While these patterns suggest a possible hidden trend, more data is needed to determine a definitive relationship between the variables.



Conclusion

After exploring R and familiarizing myself with its commands and functions, I found working with the given dataset to be both engaging and insightful. Through this assignment I learned how to plot data effectively, identify potential trends, calculate key values such as the mean, determine the dataset's size, learn from my mistakes, and most importantly I learned to enjoy the process of data analysis.