**PROGRAMMING FOR DATA ANALYSIS**

**U.S Weather Report Data Management**

APU2F2102CS(IS)

Abiyyu Taj Mahasin Bagindo

TP058652

Date Assigned: March 17th 2021

Date Completed: May 28th 2021

**Introduction**

Humans must make countless decisions every day. These decisions can be something as minor as what to eat for breakfast, or something as major as getting a large loan from the bank. Before these decisions are made, there are various factors that are considered to make sure the decision that will be made gives the best outcome. Logically, the bigger and more complex the decision, there will be more complex factors to consider, and the number of factors may increase as well.

Just like humans have always had, to help them with this problem of decision making, humans develop tools. For example, explorers use compass to help them make the decision of where to go next. Scientists have multiple meters and displays to help them measure and make the decision of what to do next. And the most recent and most advance example is the invention of computers to allow humans to do calculations at a massive scale and fast rate.

To ease the process of writing programs for computers, humans made programming languages. There is a wide variety of programming languages, each with its own purpose, strengths, and weaknesses. Among those languages, is the R programming language. R is a GNU-based programming language, made for graphical and statistical purposes at Bell Laboratories (R: What is R?, 2021).

The objective of this project is to analyze the annual weather data report from the United States using the R programming language to help decision making. This expansive weather data contains multiple nation-wide variables such as temperature, rainfall, air pressure, sunshine, and more. The data records the aforementioned variables of every day in United States during a span of a whole year. This means a wide analysis can be made, and multiple conclusions can be reached. This helps tremendously in decision making where weather is an important factor.

**First Question – When is the best time to dry your clothes?**

**Analysis 1-1: Which days have zero rainfall?**

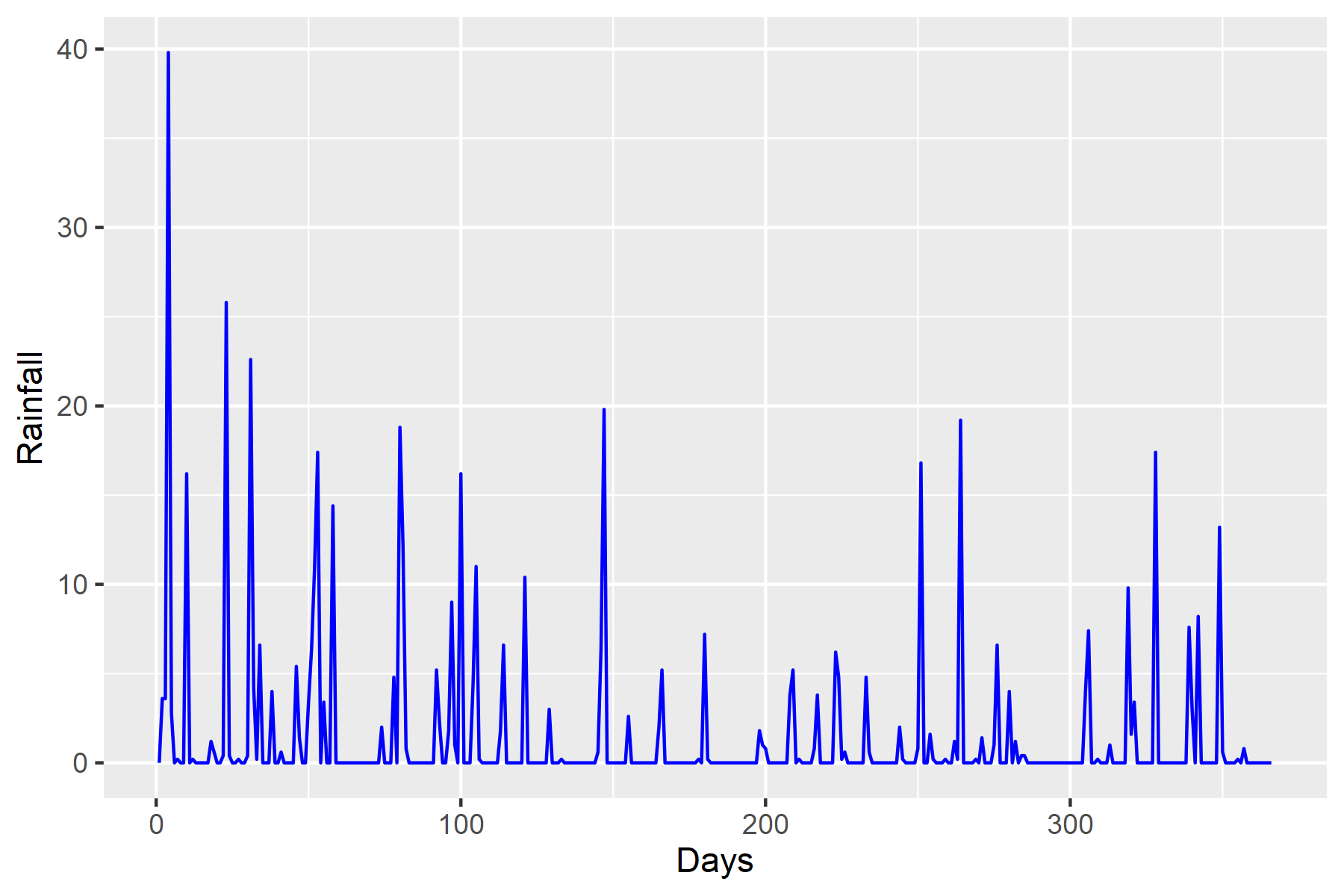


Figure 1 Rainfall graph

The first analysis is to filter out days with zero rainfall, because the last thing you want when drying your clothes is for them to be wet again. The code below starts with the initialization of a NULL vector with the name *rainyday* which will be used to store the days with zero rainfall. A variable called *rainydayindex* is then assigned a value of 1. A for loop that loops 366 times for each day of the year that contains an if statement then starts. The if statement checks whether the contents of the second column of the *rainfallframe* data frame, which is used to store the rainfall data is below 0.2 or not because 0.2 is the lowest value of rainfall. If it is lower than 0.2, that day has zero rainfall and the value of the counter variable *i* is then assigned to the index *rainydayindex* of the vector *rainyday*. The *rainydayindex* variable is then incremented by 1 and the loop repeats. This algorithm is used multiple times throughout the program and we will refer to it as *The Data Frame Crawler Algorithm* in the next parts of the document to reduce repetition. The algorithm shows that there are 264 days with zero rainfall.

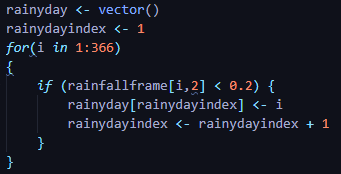


Figure 2 Zero rainfall filter code

**Analysis 1-2: Which days have high evaporation rate?**

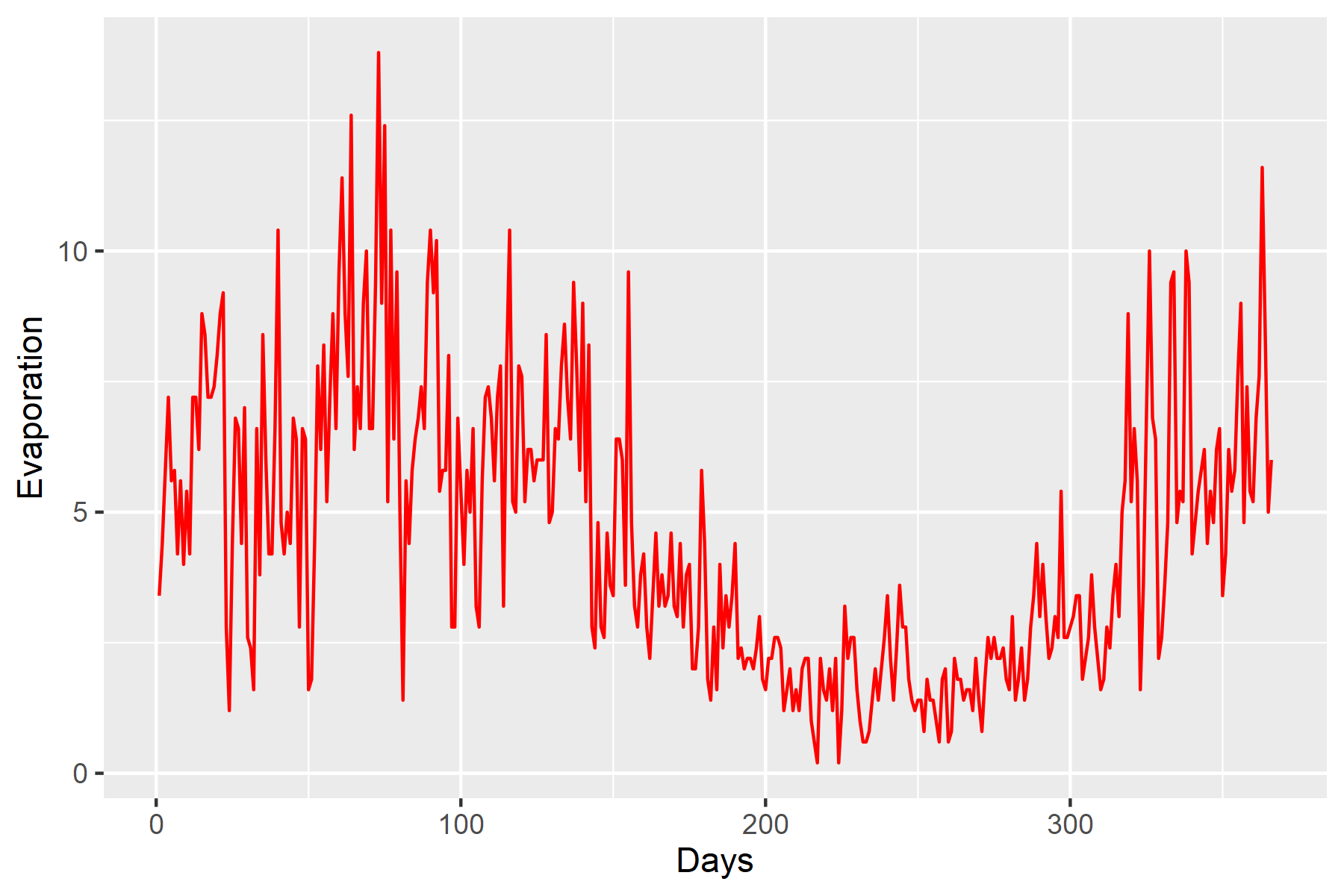


Figure 3 Evaporation graph

The second analysis is to filter out the days that have above average evaporation rate. The higher the evaporation rate, the faster the water on your clothes will evaporate, therefore the faster it will dry. The algorithm used in the code below is the data frame crawler algorithm which is the same one used in the previous analysis. With the difference being the data that is fed to the algorithm and the value that is used to filter out days. In this analysis, we need to find the days with high evaporation, as a bottom line for the filter, the average evaporation rate has been used. The average evaporation rate is obtained by using the *mean()* function which returns the average of the argument used in the form of a float. To use it as an integer, the *round()* function is then used which returns a rounded version of the argument. This rounded value is then assigned to the *evameanr* variable as an integer using the *as.integer()* function. The *evadayindex* is then printed at the end to display the days where the evaporation rate is above average. The algorithm shows that there are 147 days with above average evaporation.

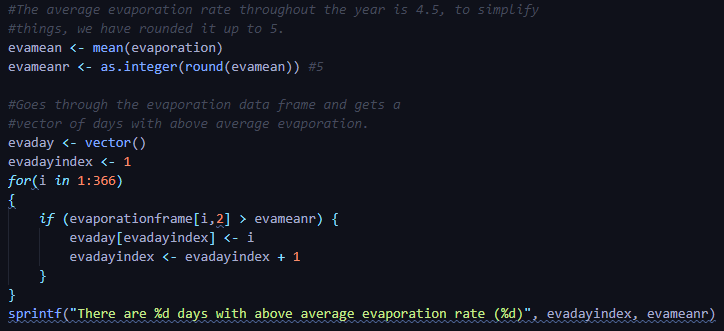


Figure Evaporation mean and filter code.

**Analysis 1-3: Which days have above average sunshine?**

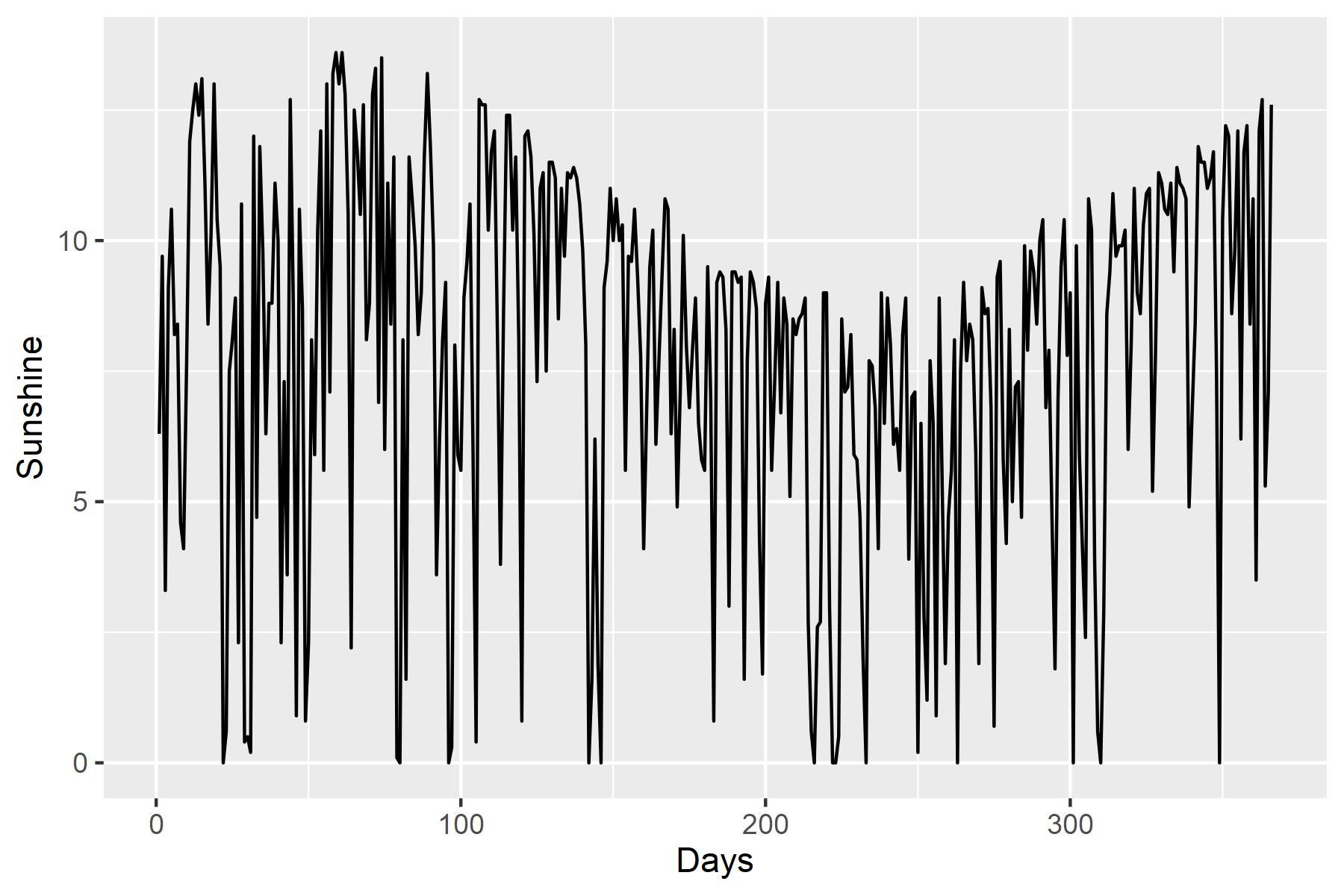


Figure 5 Sunshine graph

The whole activity of drying clothes naturally, is only possible due to the energy carried by sunlight. Therefore, the third analysis is to filter out days with above average sunshine. Once again, we used the same data frame crawler algorithm to find the days with above average sunshine. The code that finds the days with above average sunshine is the same as the one used in Analysis 1-2 with only the variable changes. The algorithm shows that there are 214 days with above average sunshine.

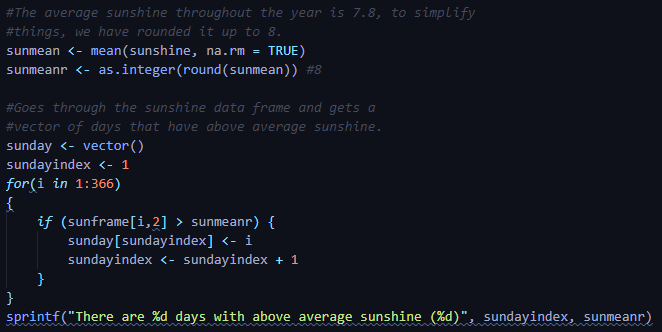


Figure 6 Sunshine mean and filter code

**First Question: Results**

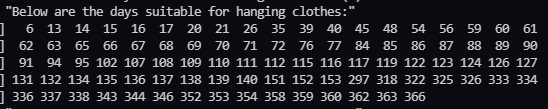


Figure 7 First question results

By using the three aforementioned analysis methods, it has been found that there are 264 days that do not rain, 147 days with above average evaporation rate (5), and 214 days with above average sunshine (8). The vectors containing the days will then be compared with each other to find days that exist in all three vectors. Those days will have all three properties for a good day to hang clothes. First, a null vector that will contain all days with zero rainfall and above average evaporation rate called *rainevaday* will be initialized. A variable called rainevaindex will also be initialized and assigned the value of 1. The following process works similarly to the data frame crawler algorithm but with an added layer and it compares values of two vectors instead. When the compared values of the two vectors matched, it will then be assigned as a value to the *rainevaday* vector. The process is then repeated with another vector which contains the days with above average sunshine and the days where all three properties are met will be printed. We will call this algorithm the *Vector Comparison Algorithm* for future references.

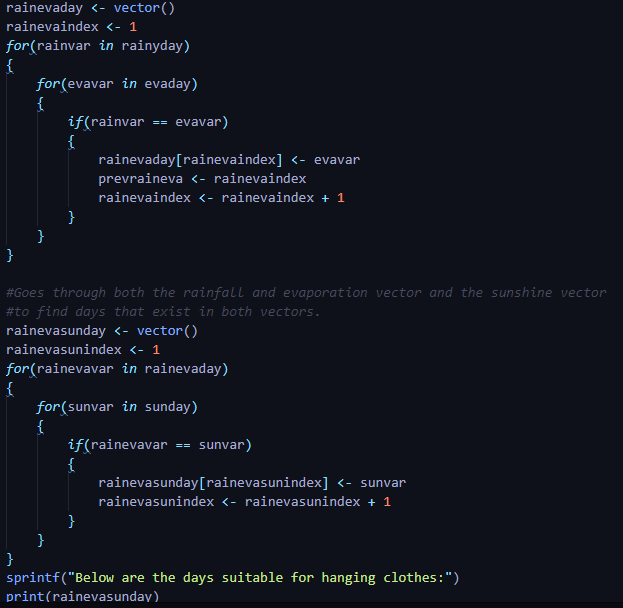


Figure Vector comparison code

**Second Question – When is the best time to exercise in the evening?**

**Analysis 2-1: Which days do not rain?**

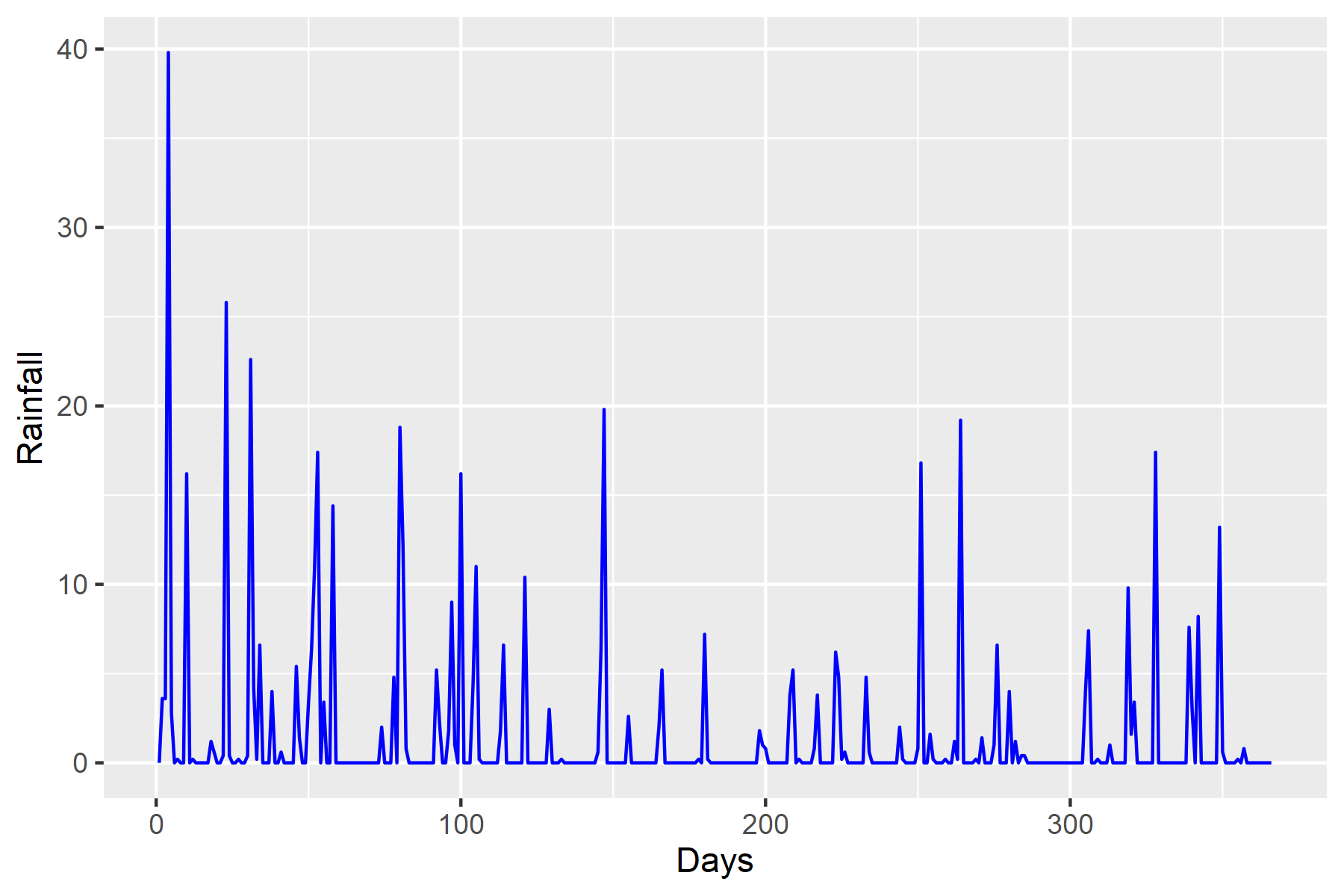


Figure Rainfall graph

Exercising is healthy, however doing it during rain will result in the opposite of healthy. Therefore, a good day for exercising will not rain. According to the weather.csv file, a minimum of 1 rainfall will be required for a day to be considered having a rain. Therefore, for this analysis days with less than 1 rainfall will be filtered. The analysis once again uses the data frame crawler algorithm to find days with less than 1 rainfall.

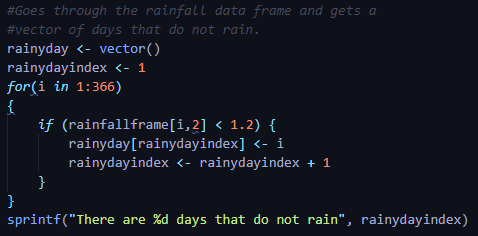


Figure No rain filter code

**Analysis 2-2: Which days have comfortable levels of humidity in the evening?**

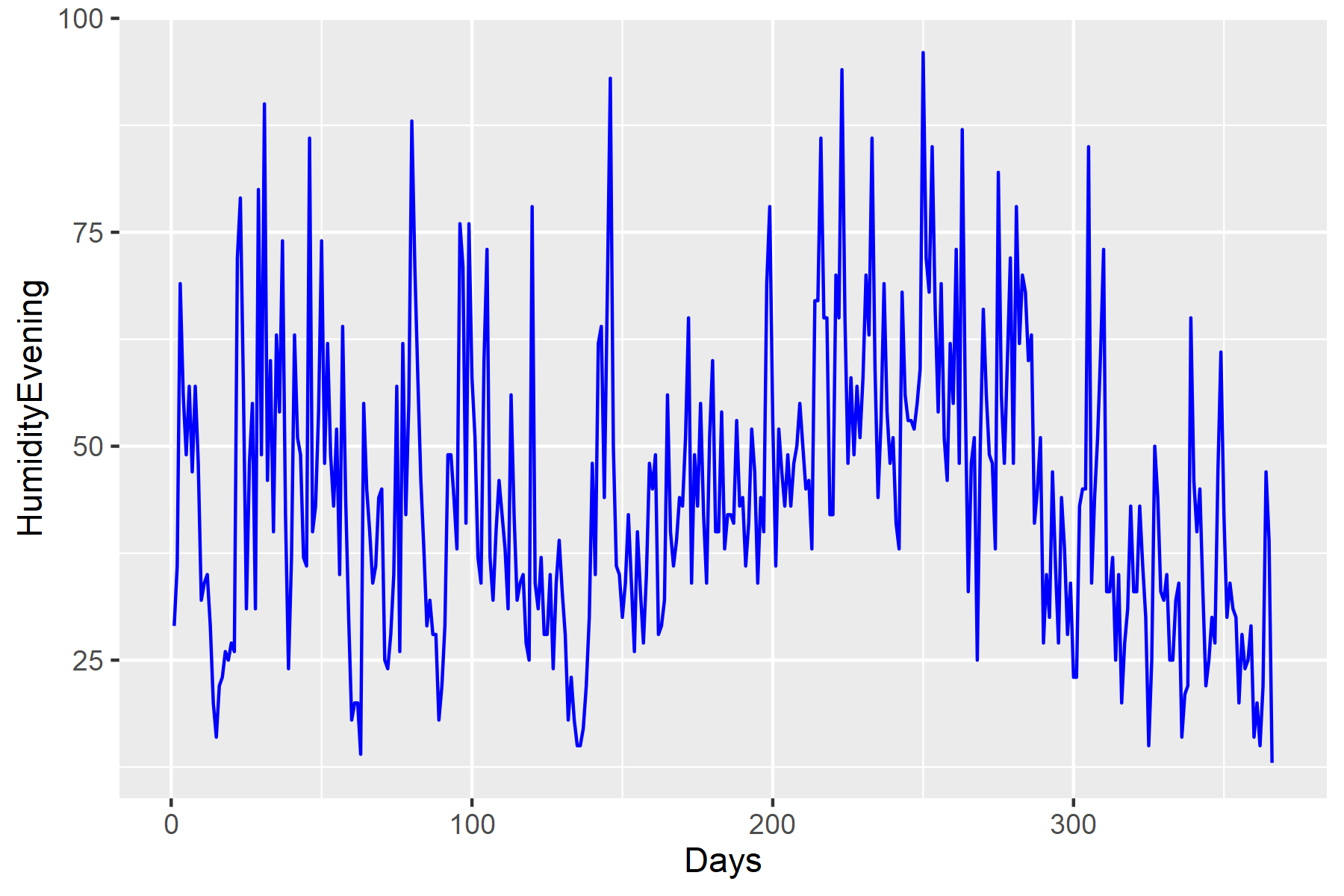


Figure Evening Humidity Graph

The air contains various gasses, among those gasses are water vapor. The level of water vapor in the air is called air humidity. The air humidity represents what percentage of the air is water vapor. The level of humidity that is comfortable and healthy for the human body is around 30% to 50% (Three Signs Your Home Has Poor Indoor Humidity, 2021). The data frame crawler algorithm will be used to find days with humidity in between 30% and 50%.

**Analysis 2-3: Which days have non-windy evenings?**

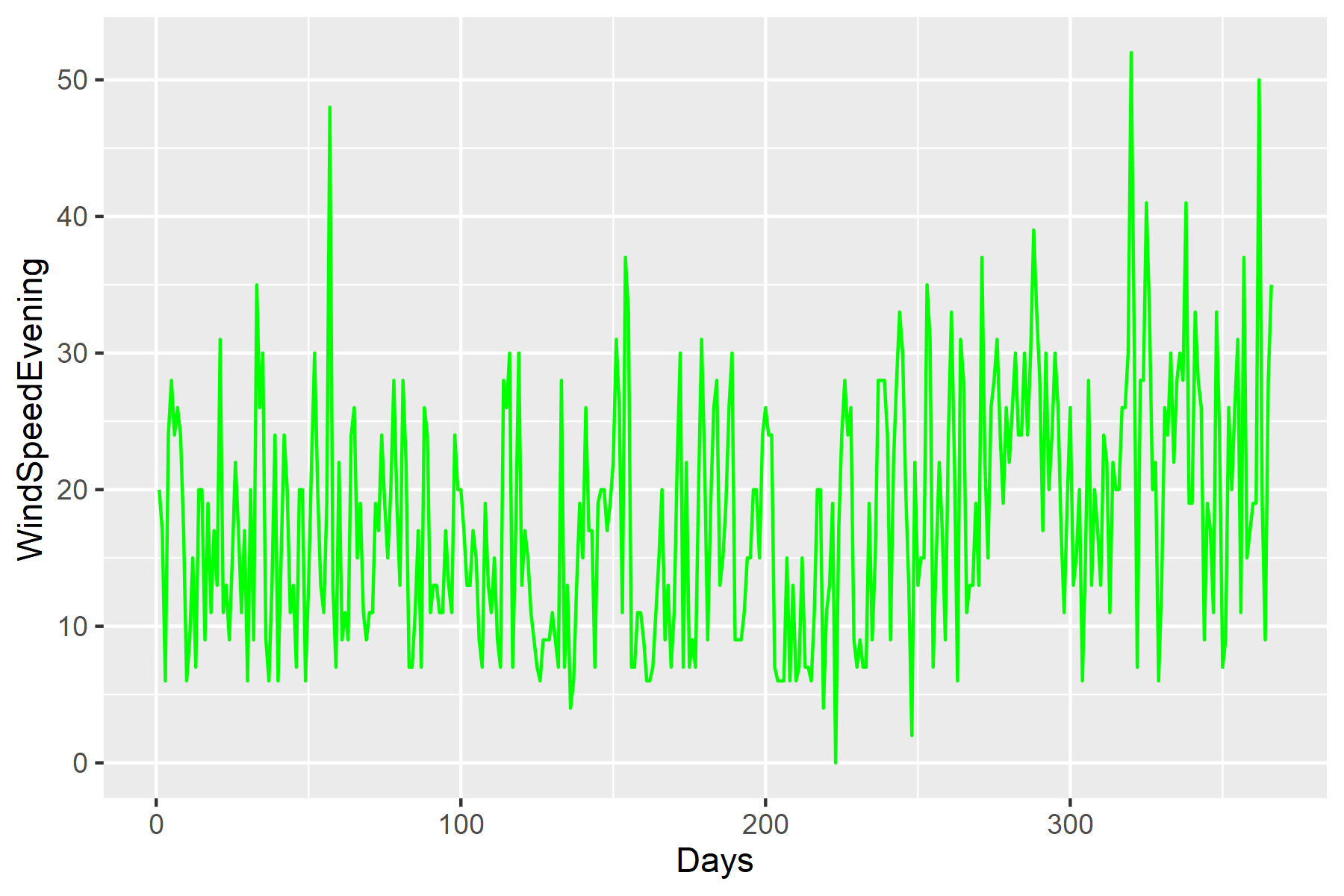


Figure Evening wind speed graph

We will be looking into days with lower-than-average wind speed during the evening to reduce wind resistance when doing exercises that involves moving from point A to point B. The algorithm used will be the same as the one used in Analysis 1-2 and 1-3 to determine which values of wind speed is below or above average.

**Second Question: Results**

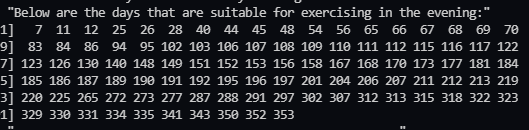


Figure 13 Second question results

The three analysis shows that there are 301 days with no rain, 166 days with comfortable humidity levels, and 213 days with non-windy evenings. The vector compare algorithm is used in to compare the three analysis and find the list of days that fit the three conditions.

**Third Question – When is the best time to fly from the west to the east?**

**Analysis 3-1: Which days have winds travelling to the east?**

Wind is a major part of aviation; this concept applies to everything that flies which ranges from animals to planes. Aviation companies, inform their pilots on the daily jet stream, which boosts the plane towards whichever direction the jet stream is flowing. This means the planes will consume less energy and therefore be more fuel efficient (Behind the Forecast: How the jet stream affects weather and flight times, 2021).

|  |  |
| --- | --- |
| Compass | Number |
| N | 1 |
| NNE | 2 |
| NE | 3 |
| ENE | 4 |
| E | 5 |
| ESE | 6 |
| SE | 7 |
| SSE | 8 |
| S | 9 |
| SSW | 10 |
| SW | 11 |
| WSW | 12 |
| W | 13 |
| WNW | 14 |
| NW | 15 |
| NNW | 16 |

For this reason, the first analysis of this question will be to find days where the wind blows eastbound to help flights that fly to the east. The algorithm that is used to find the days where the wind blows to the east is the data frame crawler algorithm, however the data that is fed to the algorithm must be processed beforehand unlike the other data. This is because, the other weather data are numeric while the wind direction data are strings that represent the respective compass direction. To process this data, an if-else statement has been written to convert the strings to numbers. The wind direction is fed to the if-else statement, if the if-else statement is true, the data will be converted to its corresponding numeric equivalent shown in the table above. The strings are checked using the *grepl* function which checks whether a string can be found in another string or not. Due to this limitation, the if-else must be in a certain order to prevent the compass data from being converted to the wrong numeric equivalent.

**Analysis 3-2: Which days have non-cloudy mornings?**

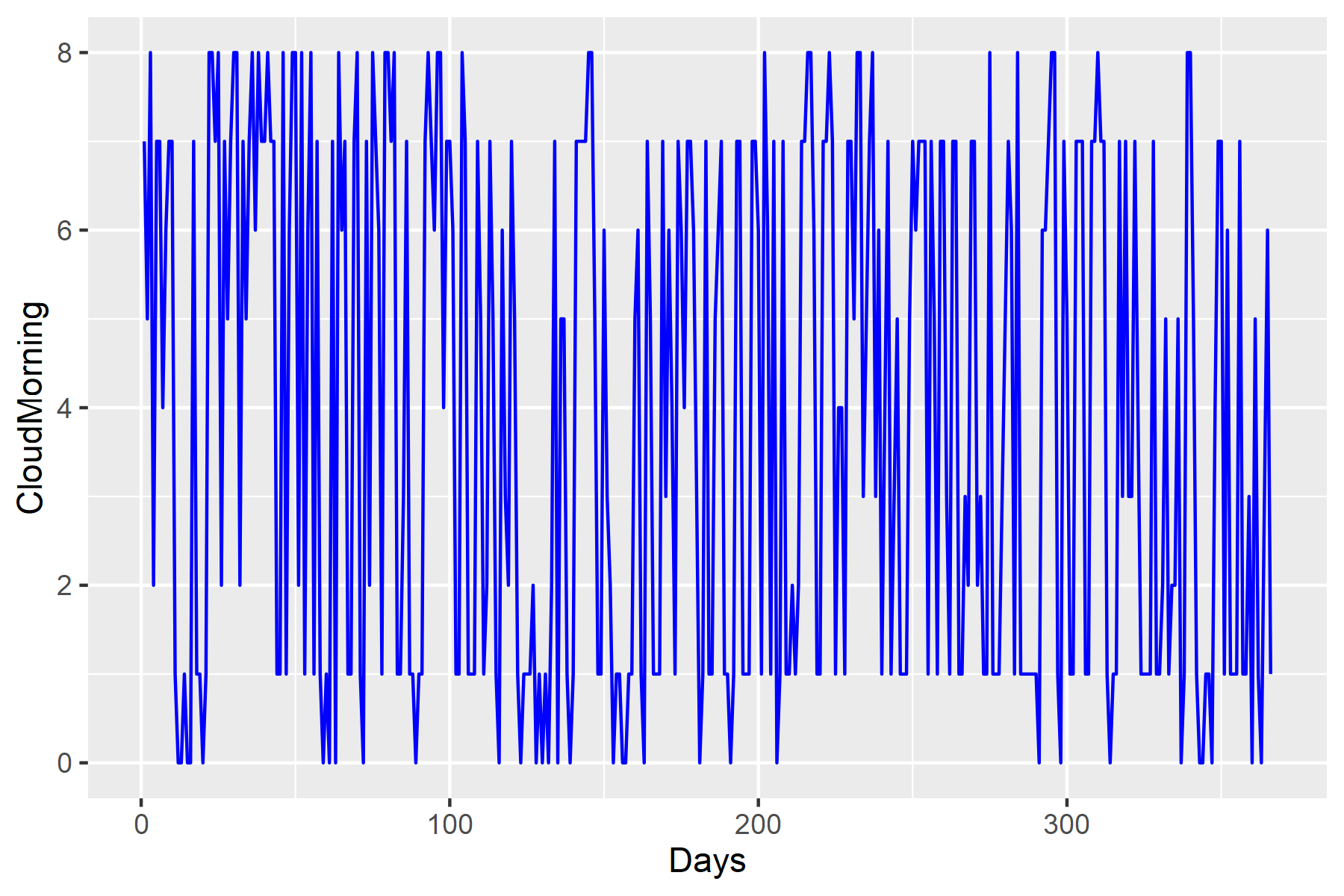


Figure Morning cloud graph

Clouds can render a normal and comfortable flight into the exact opposite. This is because, when a plane flies through a cloud, it goes through tiny water droplets which causes turbulence (Flying Through Clouds – Why Are They So Dangerous? | Southern Wings, 2021). For this reason, the second analysis of the third question will be to find days with minimal amounts of cloud in the morning. The algorithm used is the data frame crawler algorithm with the daily cloud average being used as the ceiling to find the below average values.

**Analysis 3-3: Which days have low rainfall?**

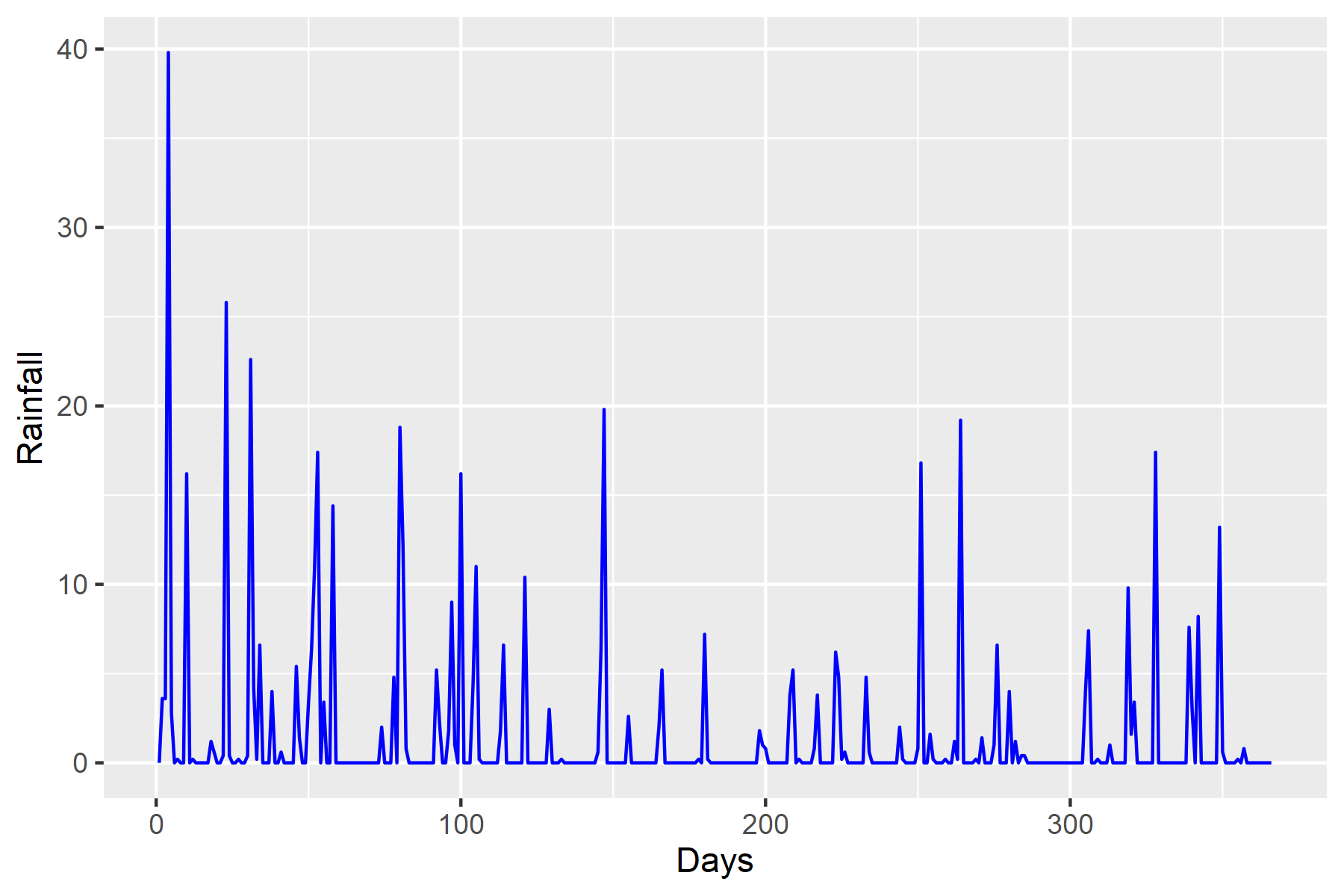


Figure Rainfall graph

A rainy weather is not the best weather to fly in, however low rainfall is still tolerable. For this analysis, days with less than 10 units of rainfall will be filtered out with the same data frame crawler algorithm.

**Third Question: Results**

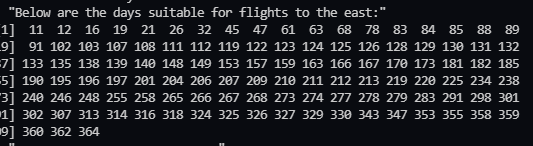


Figure 16 Third question results

The three analysis shows that there are 221 days with winds travelling to the east in the morning, 184 days with below average cloudiness, and 350 days with less than 10 rainfall. By using the vector compare algorithm, it has been discovered that there are 112 days with good weather conditions to fly from the west to the east.

**Fourth Question – When is the best time to ski?**

**Analysis 4-1: When is winter?**

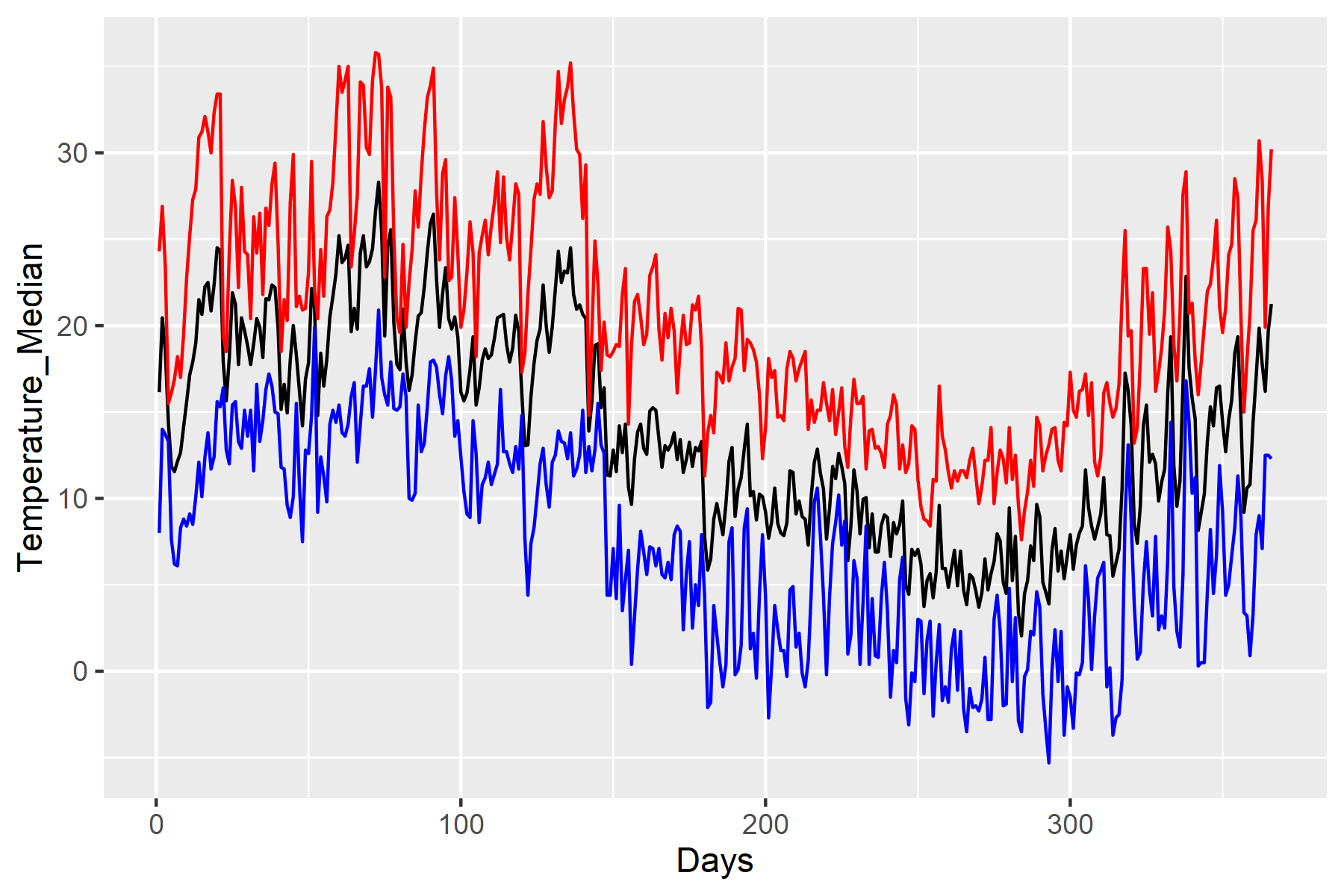
****

Figure Temperature graph showing daily Min/Max temperatures and the median temperature

Skiing is an activity that requires snow, therefore the best and probably the only time skiing can be conducted is during the winter. Since winter is a season, and seasons last for around 90 days, the temperature median graph shows that winter starts at around the 124th day and ends at around the 214th day.

**Analysis 4-2: Which days have non-windy mornings?**

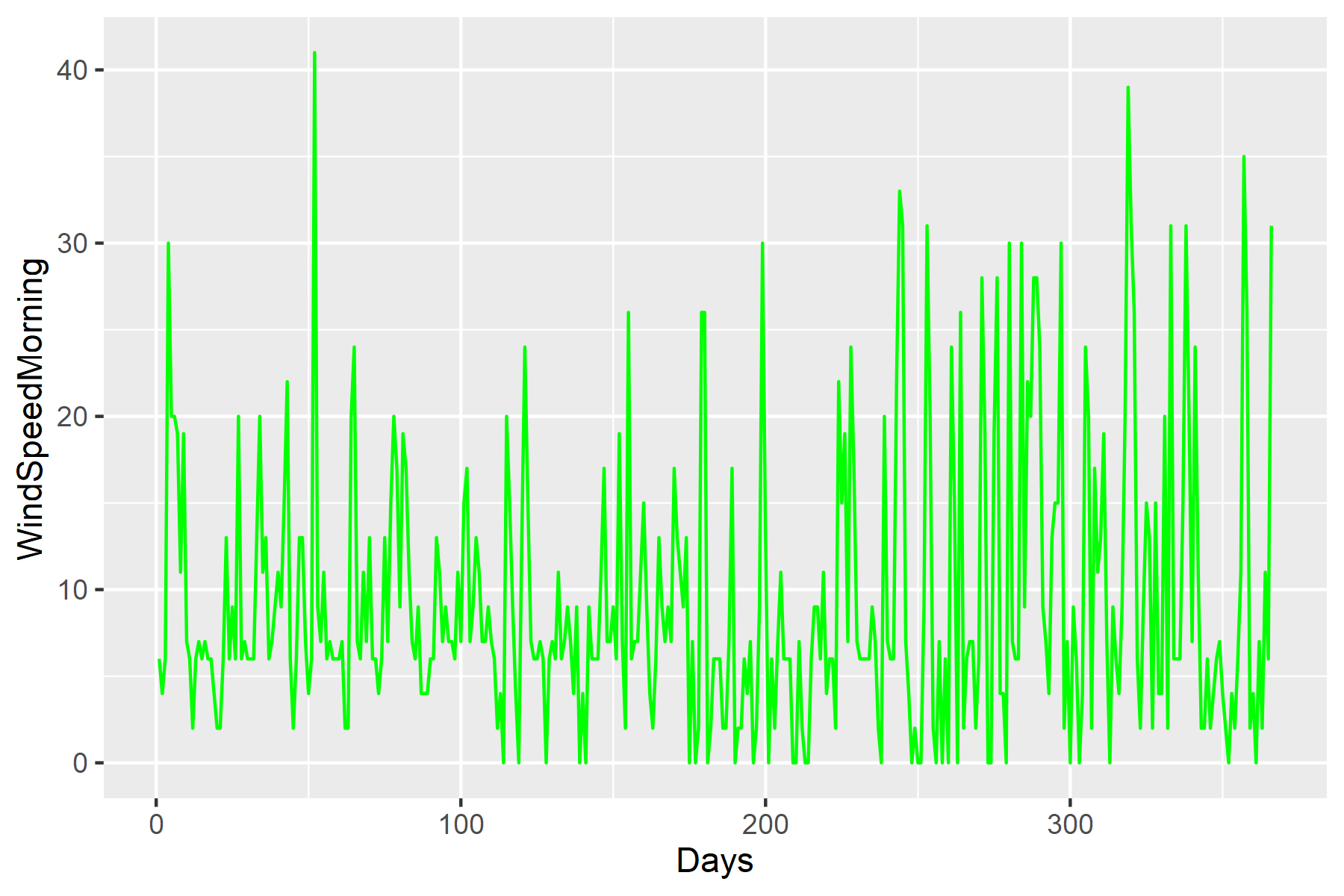


Figure Morning wind speed graph

Mornings are the best time to ski due to the condition of the snow being fresher due to the lack of people in the morning. And since there are less people, the area will be less crowded, and more movements will be possible. However, morning winds can be chilly and cold. Therefore, the next analysis is to find the days with low morning wind speed. The data frame crawler algorithm is used again.

**Analysis 4-3: Which days lack of gust?**

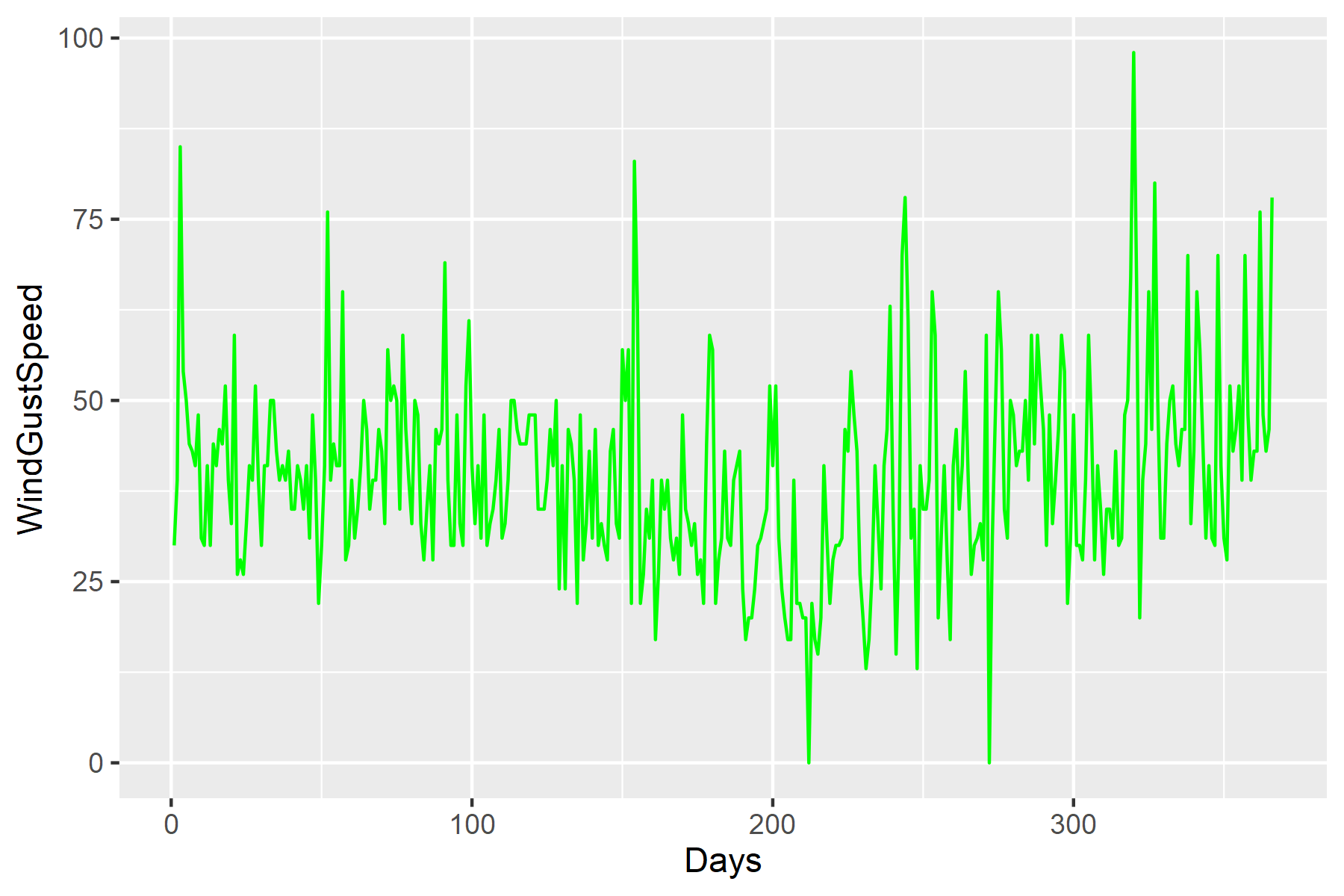


Figure Wind gust speed graph

Strong gusts of wind can only either benefit or does not benefit. For the case of skiing, it is the latter. Strong gusts of wind may be chilly and even intervene with movement which may lead to unwanted circumstances. The third analysis will be to find the days with low gust speed. By using the data frame crawler algorithm, the list of days with low gust speed can be found.

**Fourth Question: Results**

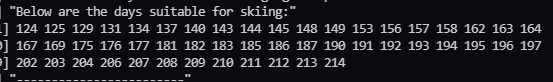


Figure 20 Fourth question results

The three analysis methods show that there are 90 days during winter, 223 days with non-windy mornings, and 192 days with below average gust speed. With the vector compare algorithm, the three vectors can be compared, and it has been found that there are 51 days with the three conditions for skiing.

**Fifth Question – When will a storm occur?**

**Analysis 5-1: Which days have low air pressure?**

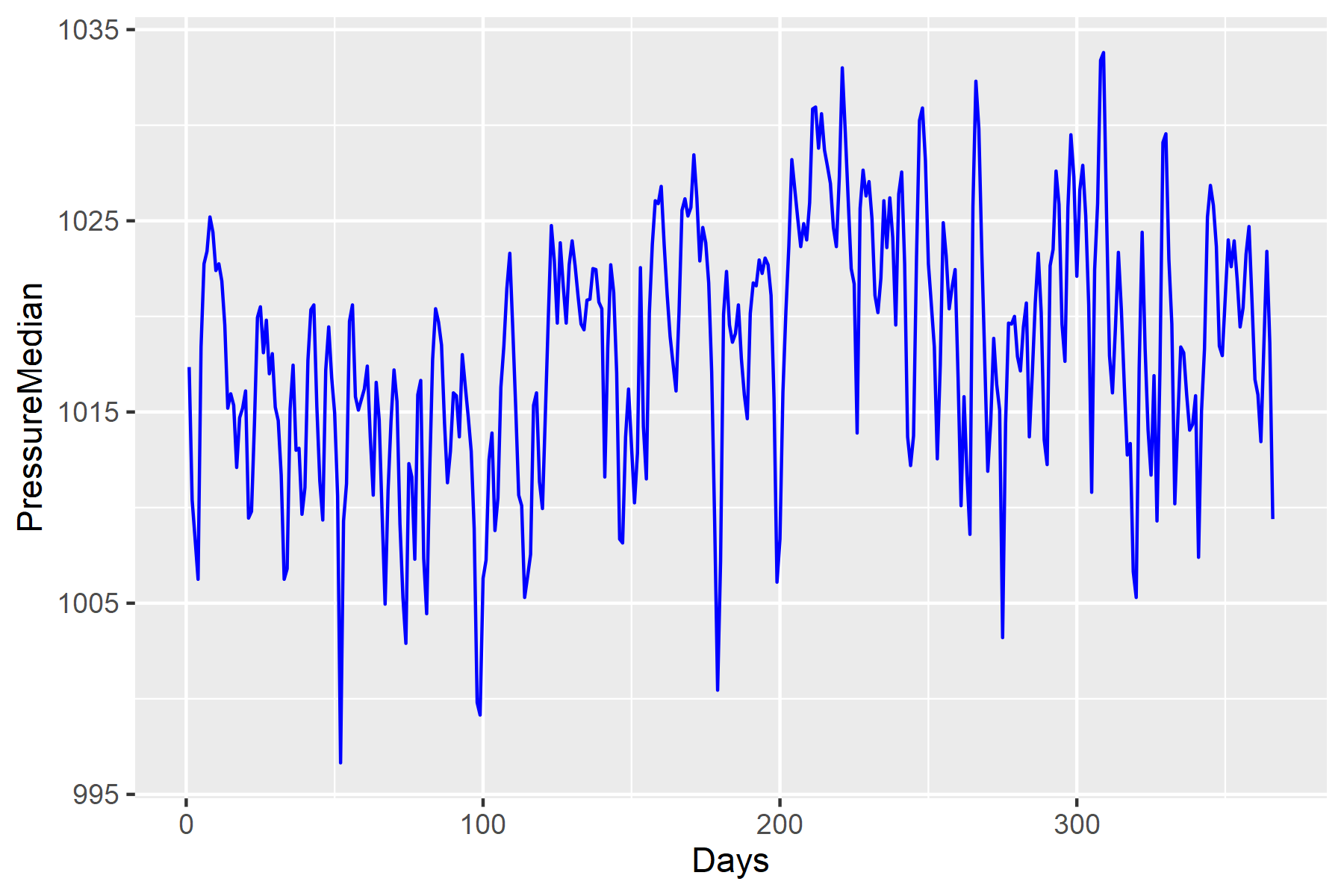


Figure Median air pressure graph

Signs of a storm coming is when there is a drop in air pressure. This is because low air pressure cells bring storm clouds (Collecting Weather Data | Physical Geography, 2021). This means finding days with air pressure drops is crucial to finding days where storms occur. To do this, two vectors are made from the morning air pressure column and the evening air pressure column. The average of the two which is the median is then calculated and kept in another vector. This vector is then used to form a data frame. The data frame crawler algorithm is then used to find the list of days with lower-than-average air pressure.

**Analysis 5-2: Which days have high rainfall?**

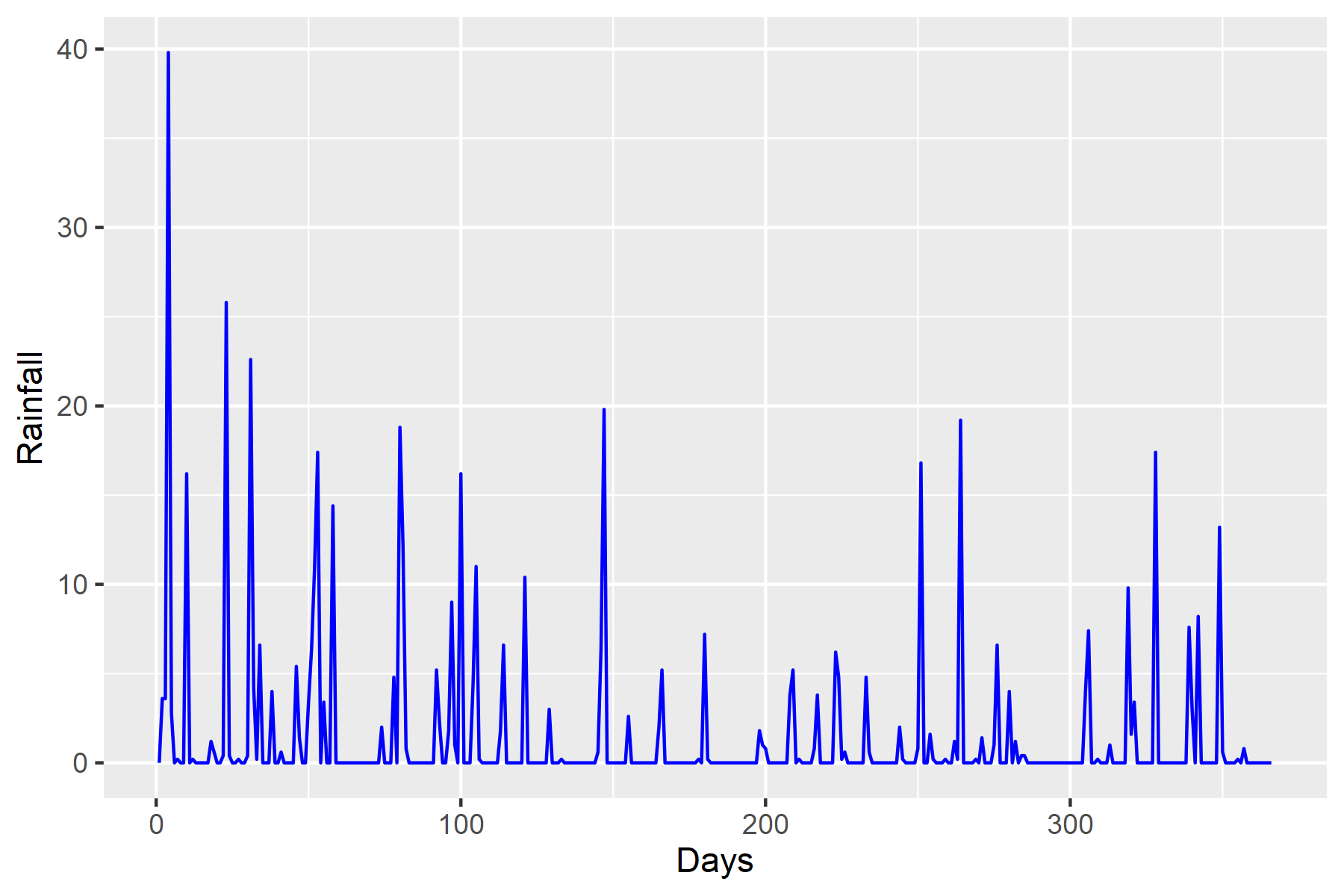


Figure Rainfall graph

Storms will always bring heavy rain. Therefore, the obvious analysis to find days where storms occur is to find days with heavy rainfall. In this analysis, the data frame crawler algorithm is utilized to find days with heavy rainfall.

**Analysis 5-3: Which days have strong gusts of wind?**

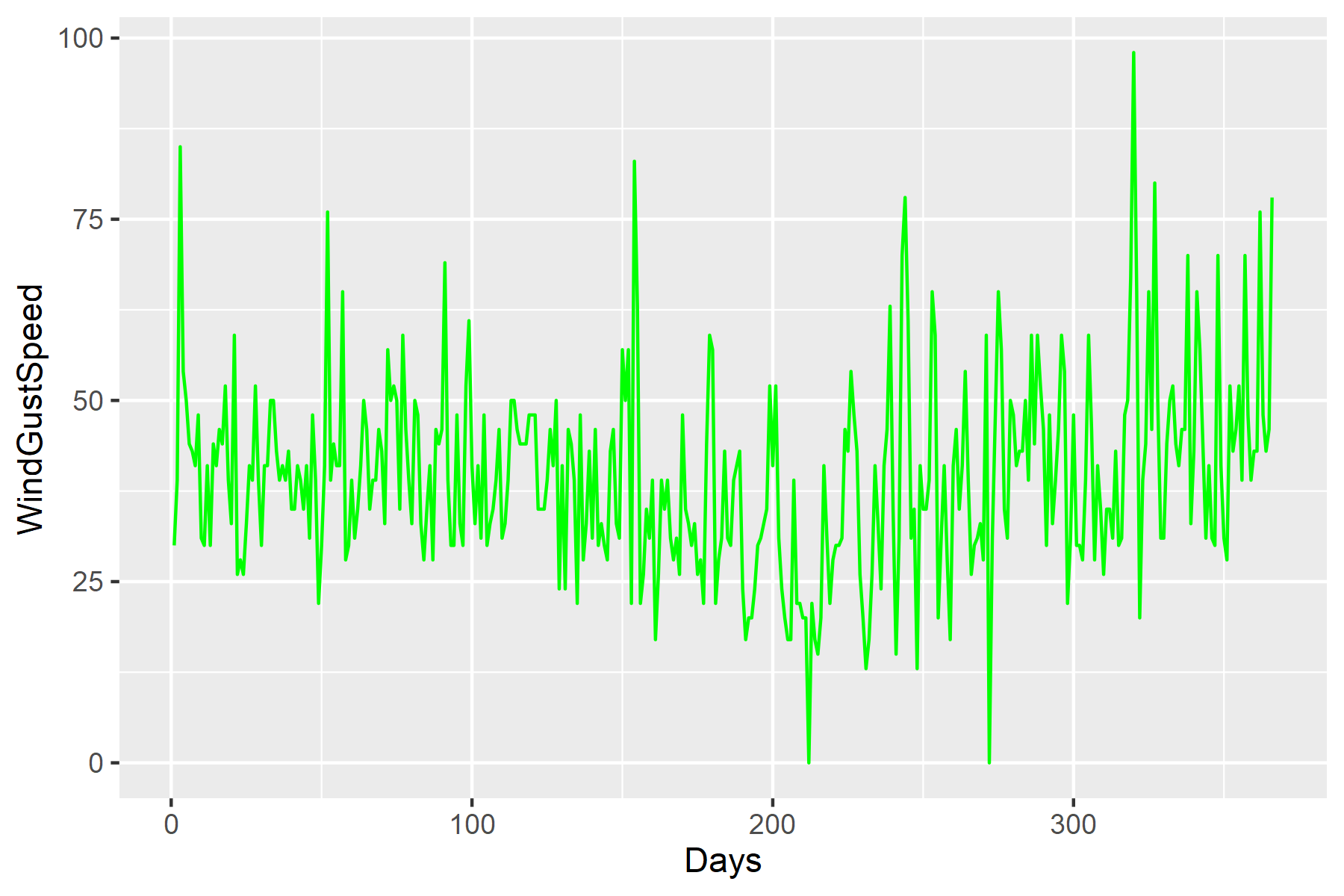


Figure Wind gust speed graph

The drop in pressure in storms produces strong gusts of wind (Knight, 2019). Therefore, It will be logical for this analysis to be to find days of the year where there are strong gusts of wind. The data frame crawler algorithm is used to filter out days with strong gusts of wind.

**Fifth Question: Results**



Figure 24 Fifth question results

The three analysis shows that there are 171 days with below average air pressure, 4 days with exceptionally high rainfall, and 176 days with above average gust speed. With the vector compare algorithm we can find days that belong to the tree vectors and there are 2 days that belong to those three vectors. This means, according to the analysis that has been made storms occurred twice throughout the year.

**Extra Feature**

When plotting graphs, R exports all the graphs in the program into a page in a file called Rplots.pdf. This fulfills the basic necessity of having the graph exported out of the program. However, by using a library called Cairo that is now a built-in library, plotted graphs can be plotted into various file formats with customizable specifications such as height, width, size unit, and more (Save a ggplot (or other grid object) with sensible defaults — ggsave, 2021).

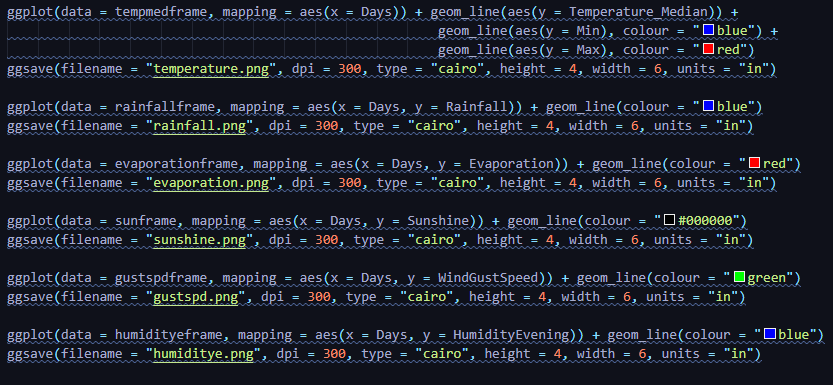


Figure Plotted graphs being exported to their respective .png files for this document

**Conclusion**

In conclusion, various methods can be used to create analysis in an effort to help decision making. Using the R programming language to process raw data is a great way to do so. Not only does R provide the tools needed, but it is also easy to use and learn. Without R, various graphs have to be plotted manually, and tables of thousands of data have to be painstakingly kept track of and then calculated. Therefore, the R programming language is a great and effective way to make analysis from raw data to help decision making.

**References**

Lennox. 2021. *Three Signs Your Home Has Poor Indoor Humidity*. [online] Available at: <https://www.lennox.com/lennox-life/comfort-matters/getting-comfortable/three-signs-your-home-has-poor-indoor-humidity> [Accessed 19 May 2021].

Wave3 News. 2021. *Behind the Forecast: How the jet stream affects weather and flight times*. [online] Available at: <https://www.wave3.com/2020/09/04/behind-forecast-how-jet-stream-affects-weather-flight-times/> [Accessed 21 May 2021].

Southern Wings. 2021. *Flying Through Clouds – Why Are They So Dangerous? | Southern Wings*. [online] Available at: <https://www.southernwings.co.nz/flying-through-clouds-why-are-they-so-dangerous/#:~:text=Turbulence%20When%20Flying%20Through%20Clouds&text=Clouds%20are%20made%20up%20of,to%20the%20surrounding%20external%20air.> [Accessed 24 May 2021].

CheckYeti Blog. 2018. *The best time to start to learn to ski*. [online] Available at: <https://www.checkyeti.com/blog/en/best-time-learn-ski#:~:text=The%20snow%20conditions%20are%20usually,you%20go%20skiing%20quite%20early.&text=Additionally%2C%20afternoon%20lessons%20are%20usually,bit%20cheaper%20than%20morning%20lessons.> [Accessed 24 May 2021].

Courses.lumenlearning.com. 2021. *Collecting Weather Data | Physical Geography*. [online] Available at: <https://courses.lumenlearning.com/geophysical/chapter/collecting-weather-data/> [Accessed 18 May 2021].

Knight, J., 2019. *Here's why the winds are so powerful long after the storms*. [online] WJLA. Available at: <https://wjla.com/weather/stormwatch7-weather-blog/heres-why-the-winds-are-so-powerful-long-after-the-storms> [Accessed 25 May 2021].

R-project.org. 2021. *R: What is R?*. [online] Available at: <https://www.r-project.org/about.html> [Accessed 28 May 2021].

Ggplot2.tidyverse.org. 2021. *Save a ggplot (or other grid object) with sensible defaults — ggsave*. [online] Available at: <https://ggplot2.tidyverse.org/reference/ggsave.html> [Accessed 28 May 2021].