

**Introduction to R Programming (112022-MHJ)**

**Assignment Title :** U.S Weather Report Data Management

**Name :** Afif Rifa’ie bin Mohd Gazali

**TP Number :** TP072295

**Intake** : APDMP2301DSBA(BI)(PR)

**Lecturer :** Ms. Minnu Helen Joseph

**Table of content**

**Introduction** 1

**Data import** 1

**Data exploration** 2

**Data cleaning** 5

**Data pre-processing** 5

**Question 1 The United States Rain Analysis**

Analysis 1.1 raining and non-raining days 6

Analysis 1.2 raining and not raining days for tomorrow 7

Analysis 1.3 Total rainfall for raining days 8

Analysis 1.4 Relation between cloud at 3pm and Rain Today 9

**Question 2 Level of dryness in the United States**

Analysis 2.1 Average temperature at 3pm 10

Analysis 2.2 Compare temperature and sunshine 11

Analysis 2.3 Compare temperature and humidity 12

Analysis 2.4 Compare temperature and evaporation 13

**Question 3 Common Breeze in the United States**

Analysis 3.1 Average wind gust speed 14

Analysis 3.2 Major wind gust speed occurrence 15

Analysis 3.3 highest wind gust speed for each wind gust direction 16

Analysis 3.4 average wind speed at 3pm when raining 17

Analysis 3.5 Wind gust direction when wind gust speed is over 50km/h while raining and identify the rain volume 18

**Question 4 Why no rain take majority in the United States?**

Analysis 4.1 - Relationship between sunshine and evaporation 19

Analysis 4.2 - Relationship between sunshine and humidity 20

**Conclusion** 21

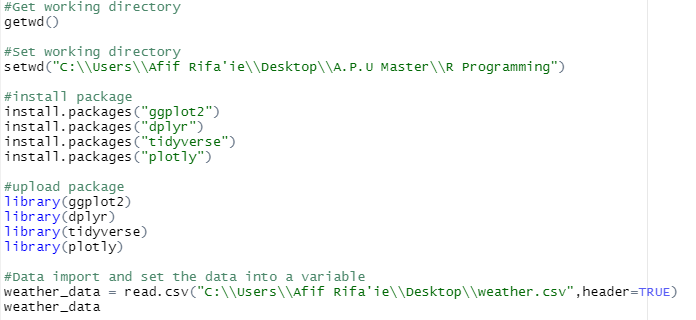
**Reference** 21

**Introduction**

The purpose of this report is to investigate and analyse weather data from the United States by utilising a variety of methods that are capable of retrieving the necessary information. The data contains 24 columns and 366 rows of information regarding rain factors and tabulation in the United States for a period of one year. The variables included in the data are temperature, humidity, and a number of others. The method begins with performing data importing, data exploration, data cleaning and data pre-processing, followed by investigation and visualisation into a visual graph using R Studio.

**Data Import**

The first step of data analysis is to import the given data set. The data was given in format called csv. Which the data in the file is separated by commas.

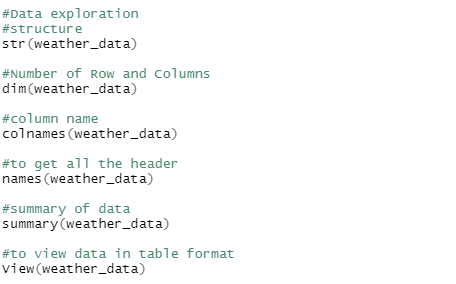


*Figure 1: setting working directory, installing packages and importing data set*

Based on the above snippet, the first thing I did was set the working directory and save it in my personal file. Then I download the necessary packages and install it in my R Studios. To import the data, I used the read.csv() function, which round bracket is the location of the data set on my computer, and I set the given data set as a weather\_data variable. I install packages with install.packages() function and the library() function to store the code in the R environment under a directory called library.

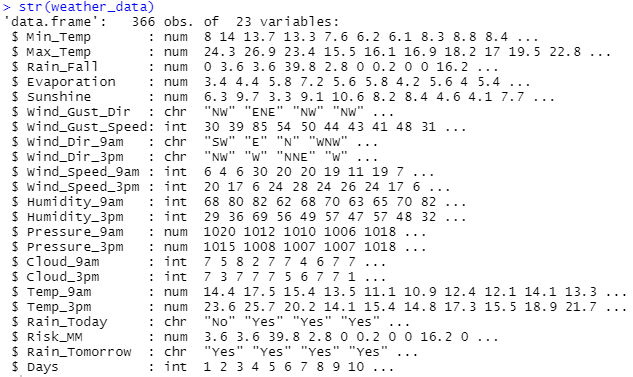
**Data Exploration**

After importing the Data Set into the R Studio. The next step is data exploration which to understand the data set. This is to do research to understand better of the rain attributes. I did research on each of the columns name in the data. After the research, I get basic information on what are the meaning of each column provided in the data set. After doing research, the next step is to know the basic information of the data such its structure, dimensions, columns names and summary.



*Figure 2: Data exploration*

To explore the data, I use str() function to find out the class of each column in data set. Then, I use colnames() function to find out the names of the data. I use the summary() function to get the data's mean, lowest, and highest values. View() function is used to view the data set in table format. The output of the function as below.



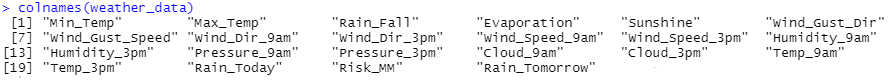
*Figure 3: str() function output*

str() function gives an overview on how many observations and variables are in the data set. This function also provides the information about the type of data, which useful when determining the graph for visualization.



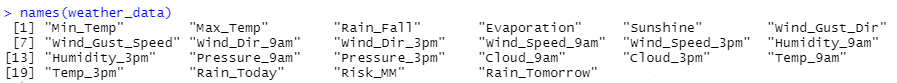
*Figure 4: dim() function output*

dim() function provides information about how many rows and columns are in the data set. There are 366 rows and 23 columns in the weather\_data.



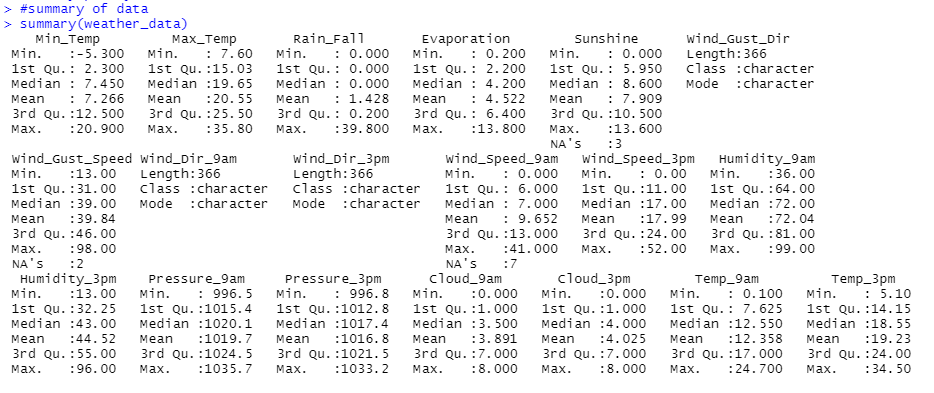
*Figure 5: colnames() function output*

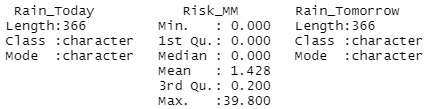
colnames() function provides the names of each column in the data frame. It is also function to change the names of the columns.



*Figure 6 : names() function output*

Similarly, names() function gives the name of the object's header in a data frame.

****

****

*Figure 7 : summary() function output*

summary() function provides an overview of the variables, including their minimum, first quartile, median, mean, third quantile, and maximum values.

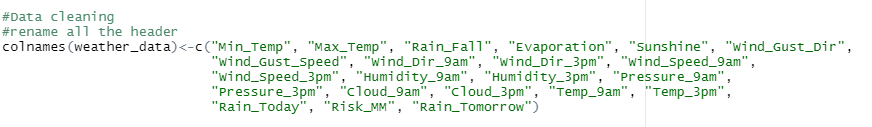
****

*Figure 8 : Views() function*

View() function is used to produce table style data viewer.

**Data Cleaning**

Once I explored the data, I proceed with data cleaning. Basically means to make sure that all the columns are set up in a way that is easy for the user to understand and to make sure the columns are set without spacing and inconsistent use of capital and small letter. As the snipped below, I reset the column names with proper spacing using underscore and start the name with capital letter. I state the new columns name into the vector c(“ ”) and set it back to the weather\_data.



*Figure 9: Data cleaning and rename the header*

**Data Pre-processing**

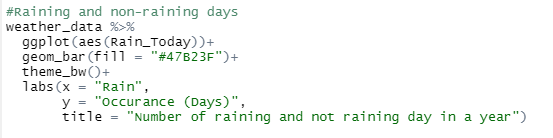
After cleaning the data set, the next step is to do the data pre-processing. From the data set, there are no dates given for when each of data attributes happened. As below snipped, I add new Days columns to the data set. With Days column, it would be useful to measure when each of the attribute happened. First, I set new vector of 1:366 because there are 366 rows of data. The assumption I make is each of the row represent a day. Then I used cbind() function to add new Days column in the data set.



*Figure 10: Add new column*

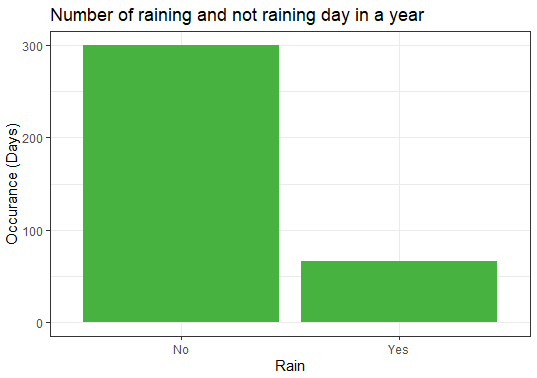
**Question 1 The United States Rain Analysis**

Analysis 1.1: Days with and without rain over the course of a year.



*Figure 11: Analysis 1.1*

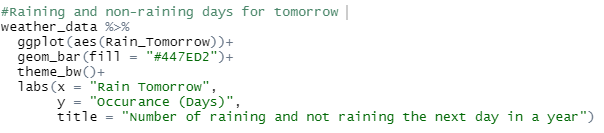
In this assignment, pipe function %>% is used to start the code as it takes output of one function, and passes it into the another function as an argument. The stored data variable is weather\_data. geom\_bar() function under ggplot() package is used to build a bar chart graph. The x-axis shows days it rained or didn't rain, and the y-axis shows how many days it happened. The bar’s colour is set by the fill function in geom\_bar(). The theme\_bw function is to set theme as grey theme with grey lines on white background. While labs() function is used to change the names of the x axis, y axis and to give title of the graph.



*Figure 12 : output of analysis 1.1*

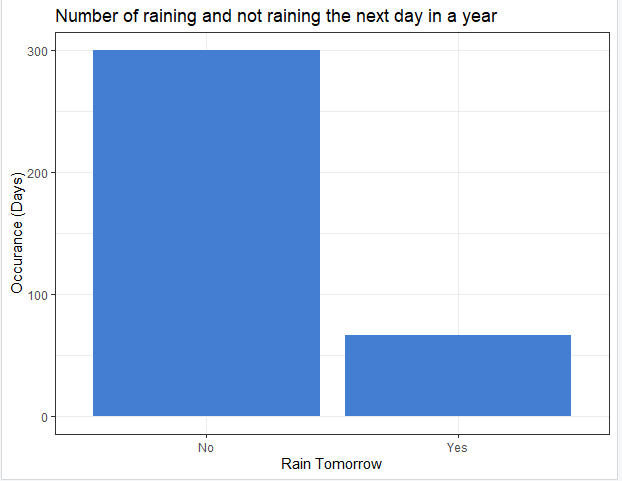
The figure above shows that the number of days when it does not rain is more than the number of days when it does rain. Meaning that it does not rain very often in the U.S.

Analysis 1.2 Number of rainy and not rainy days for tomorrow



*Figure 13: Analysis 1.2*

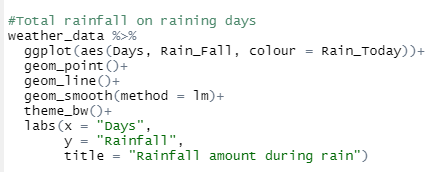
The code is used to make a bar chart graph, where x representing will it rain tomorrow and y axis representing how often it will rain. The bar's colour is set by the fill function in geom\_bar().



*Figure 14: output of analysis 1.2*

Similar to rain today bar chart, rain tomorrow has similar amount of rainy and not rainy days where no rain takes the majority of the days in the United States throughout a year.

Analysis 1.3 Total rainfall for raining days



*Figure 15: Analysis 1.3*

This code is used to build a scatter plot with a tabulation line that shows how much rain falls throughout a year, with X axis representing the number of days and Y axis representing the amount of rain. In addition to that, colour = Rain\_Today in ggplot() function to differentiate categorical variable for Rain\_Today in the plot. Green represent yes for raining days while orange represent not raining days. geom\_smooth(method = lm) function to add a straight regression line into the scatter plot.

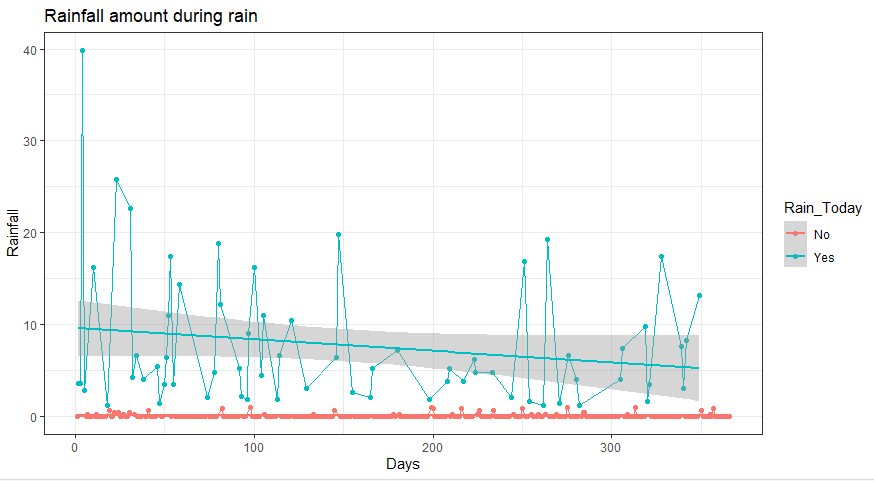
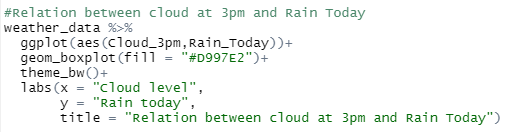


Figure 16: Output of analysis 1.3

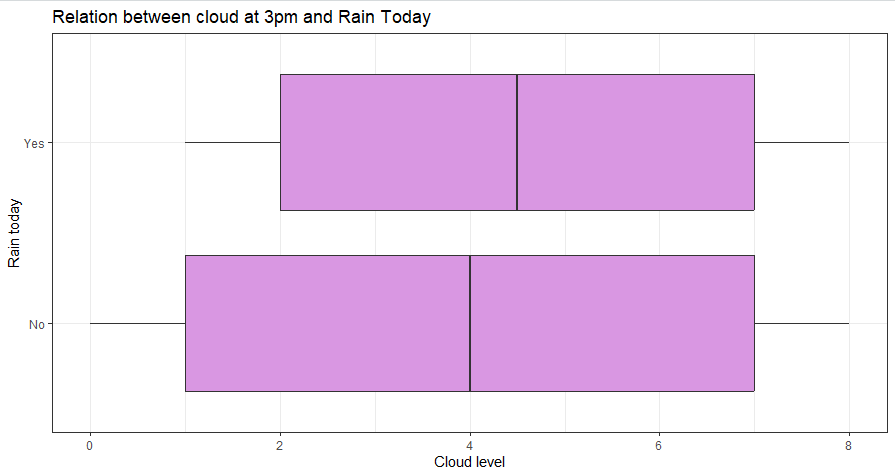
The figure above shows that most of the rain falls between 0 and 20 throughout the year, and only one day has high amount of rain fall.

Analysis 1.4 Relation between cloud at 3pm and rain today



*Figure 17 : Analysis 1.4*

The code is used to build a boxplot graph. The x axis representing cloud level, the y axis representing rain today. The colour of the boxplot is set by the fill function in geom\_boxplot.

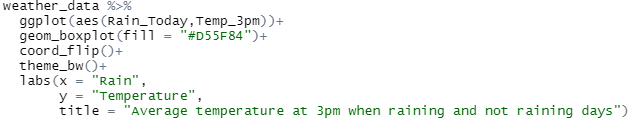


*Figure 18: Output of analysis 1.4*

The line in the middle of the box shows the average value, while the left and right walls show the lowest and highest values. The boxplot above shows that on days when it rains, the clouds are more likely to be in the 2–7 cloud level range than on days when it doesn't rain, when the cloud level can go as low as 1.

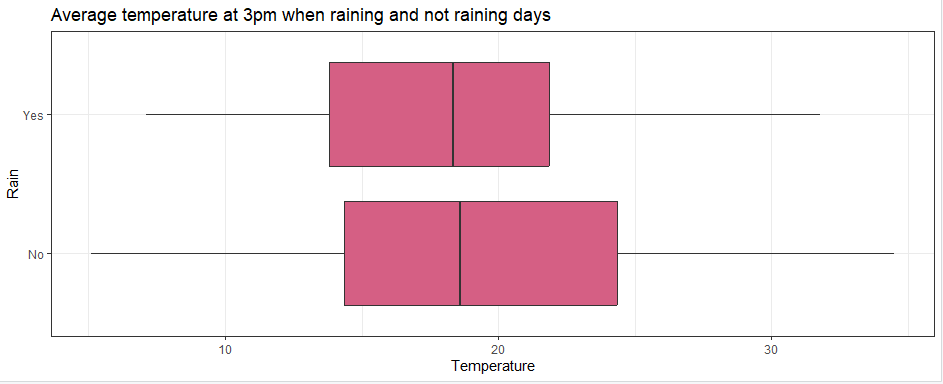
**Question 2 Level of dryness in the United States**

Analysis 2.1 –Average Temperature at 3pm



*Figure 19: Analysis 2.1*

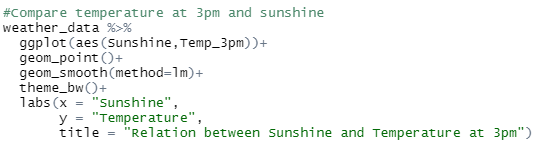
The code is used to build a boxplot graph. The x axis representing rain today, the y axis representing temperature at 3pm. The colour of the boxplot is set by the fill function in geom\_boxplot. The coord\_flip() function is used to flip the graph so that it can have better view.



*Figure 20: Output of analysis 2.1*

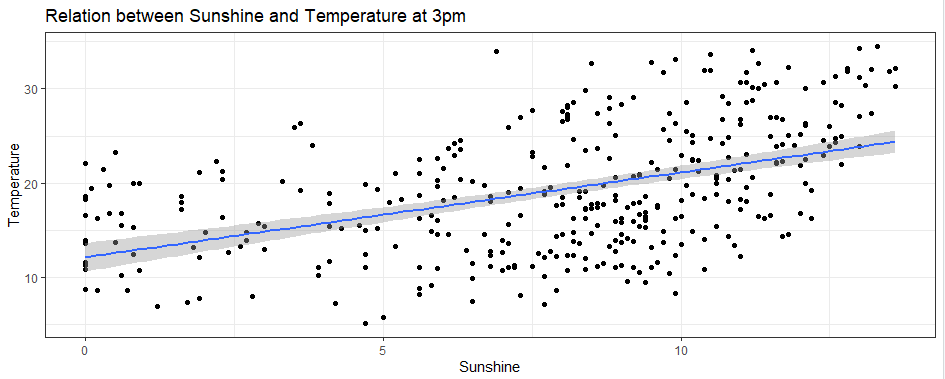
The line in the middle of the box shows the average value, while the left and right walls show the lowest and highest values. The boxplot above shows that when it rains, the average temperature is a little lower than when it doesn't. The lowest temperature is 5.10 when raining. On days when it doesn't rain, the temperature can get as high as 34.50.

Analysis 2.2 – Compare temperature and sunshine



*Figure 21: Analysis 2.1*

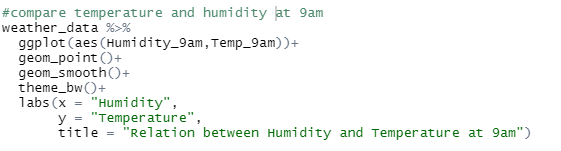
The code is used to build a scatter plot. The X-axis shows the amount of sunshine, and the y-axis shows the temperature at 3 p.m. For the straight linear line on the graph, the geom\_smooth() function with method = lm is used.



*Figure 22: Output of analysis 2.1*

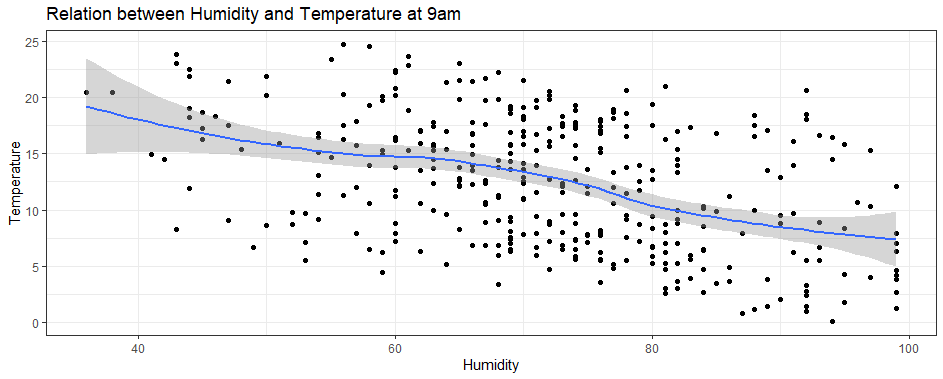
The figure above shows that there is a positive relationship between the amount of sunshine and the temperature. When there is more sunshine, the temperature goes up.

Analysis 2.3 – Compare temperature and humidity



*Figure 23: Analysis 2.3*

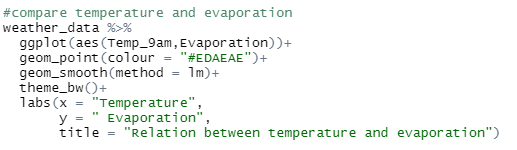
The code above is used to build a scatter plot. The x axis representing humidity, the y axis representing temperature at 9am. The geom\_smooth() function is used to get the linear line in a graph.



*Figure 24: Output of analysis 2.4*

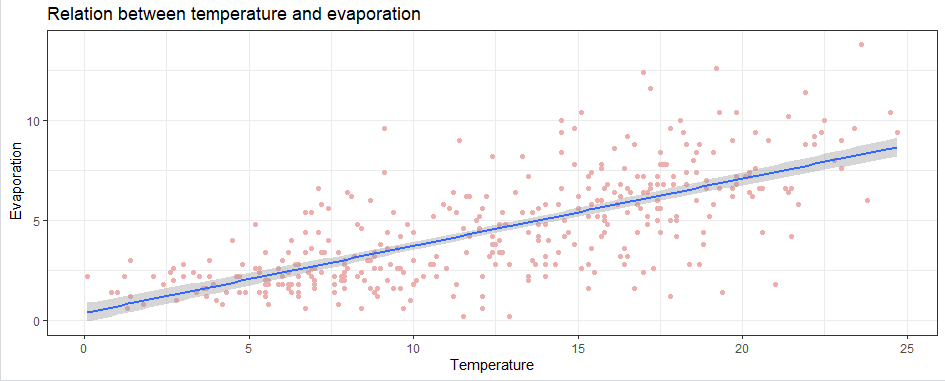
The above figure shows that the humidity has negative relationship against temperature. When humidity is high, the temperature goes down.

Analysis 2.4 – Compare temperature and evaporation



*Figure 25: Analysis 2.4*

The above code is used to build a scatter plot. The x axis representing temperature at 9am, the y axis representing evaporation. The fill function in geom\_boxplot is to set the colour of the point. The geom\_smooth(lm) function is to get the linear line in the graph.

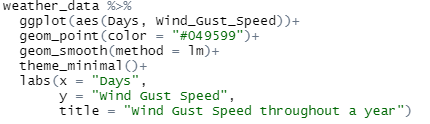


*Figure 26: Output of analysis 2.4*

The above figure shows that the evaporation has positive relations with the temperature. The amount of water that evaporates depends on how high the temperature is.

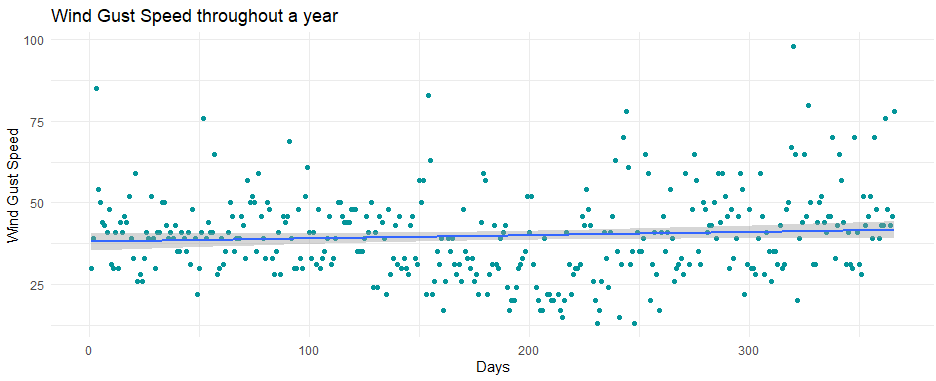
**Question 3 Common Breeze in the United States**

Analysis 3.1 –Average Wind Gust Speed



*Figure 27: Analysis 3.1*

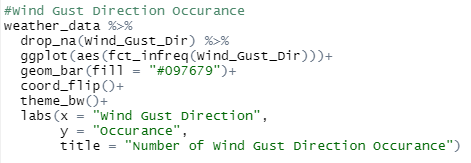
The code is used to build a scatter point graph, where the x axis shows the number of days and the y axis shows the speed of the wind gusts. The geom\_smooth(lm) function is get the linear line in the graph.



*Figure 28: Output of analysis 3.1*

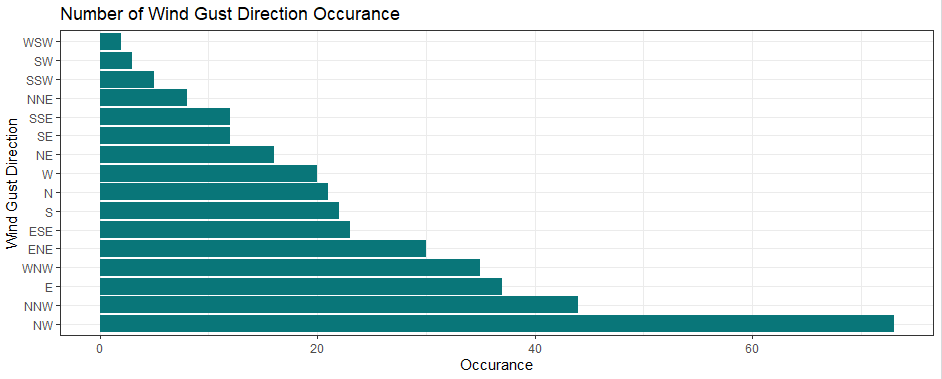
The figure above shows that the average gust speed in the United States is 39.00 miles per hour. The highest possible value is 98, and the lowest possible value is 13.

Analysis 3.2 –Major Wind Gust Speed occurrence



*Figure 29: Analysis 3.2*

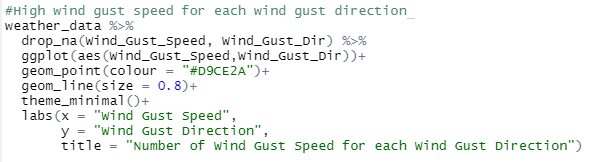
The code is used to build a bar chart graph, with the x axis showing the direction of wind gusts and the y axis showing how often they happen. The fct\_infreq function is used to make the number of occurrence of wind gust direction in ascending order. The coord\_flip() function is used to turn the graph around so it can be seen better.



*Figure 30: Output of analysis 3.2*

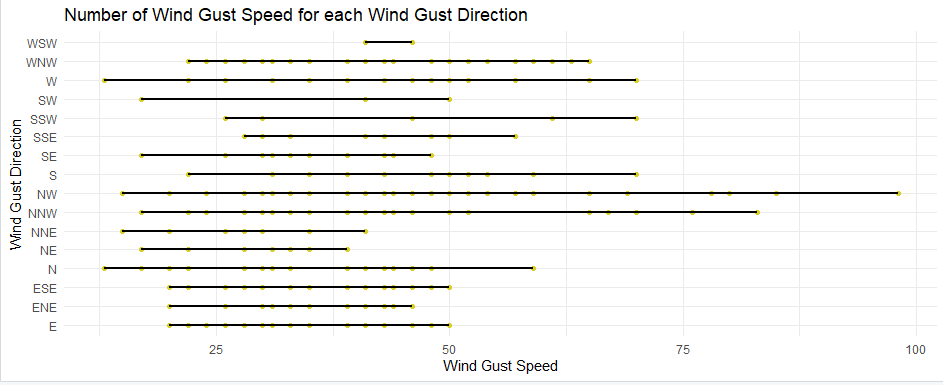
The figure above shows that wind gusts are most likely to come from the NW and least likely to come from the WSW.

Analysis 3.3 –Highest wind gust speed for each wind gust direction



*Figure 31: Analysis 3.3*

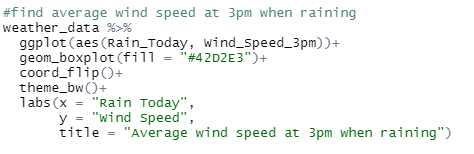
The code is used to build a scatter point graph with line. The x-axis shows the speed of wind gusts, and the y-axis shows the direction of wind gusts. drop\_na() fuction is to drop rows that contain missing values. The geom\_line() function is used to put a line in between of each point.



*Figure 32: Output of analysis 3.3*

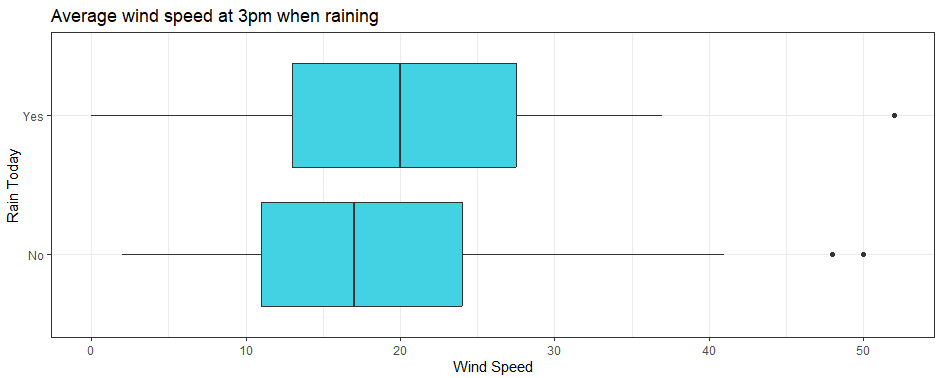
The figure above shows that the wind gust speed is highest at the NW, at 98.00, and lowest at the W and N, at 13.00.

Analysis 3.4 –Average wind speed at 3pm when raining



*Figure 31: Analysis 3.4*

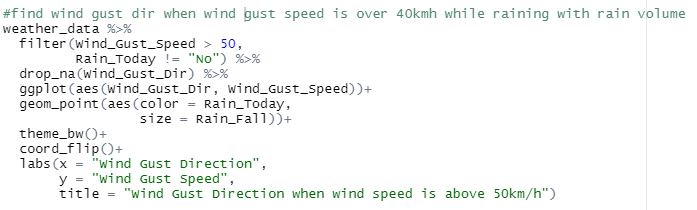
The code is used to build a boxplot graph. The x axis representing rainy and not rainy days and y representing how fast the wind speed is at 3pm. The coord\_flip() function is used to rotate the graph for better view.



*Figure 32: Output of analysis 3.4*

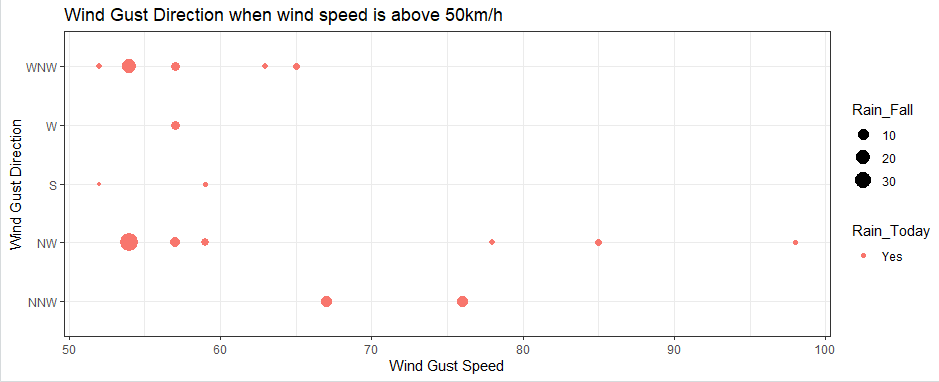
The line in the box shows the average value, and the walls on the left and right show the lowest and highest values. The figure above shows that when it rains, the average wind speed is 20, which is faster than when it doesn't rain. In the same way, both the slowest and the fastest wind speeds are higher when it rains.

Analysis 3.5 –Wind gust direction when wind gust speed is over 50km/h while raining and identify the rain volume



*Figure 33: Analysis 3.5*

This code is used to construct a scatter plot graph. The x-axis shows where wind gusts are coming from, and the y-axis shows how fast they are moving. The filter() function is used to make sure that only days when it rains are shown when the wind gust speed is over 50 km/h. The geom\_point() function is used with "aes" of "size" to show how much rain there is. The coord\_flip() function rotates the x and y axes to make the graph easier to read.

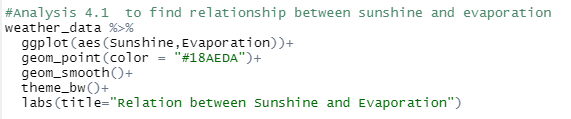


*Figure 34: Output of analysis 3.5*

The above scatter plot shows that when wind gusts are over 50 km/h, the strongest gusts are in the NW and the most rain falls are in the WMW and NW.

**Question 4 Why no rain take majority in the United States?**

Analysis 4.1 - Relationship between sunshine and evaporation



*Figure 35: Analysis 4.1*

This code is used to construct a scatter graph to visualise whether or not the amount of sunshine has an impact on the amount of evaporation. The x axis represents the amount of sunshine, and the y axis represents the amount of evaporation.

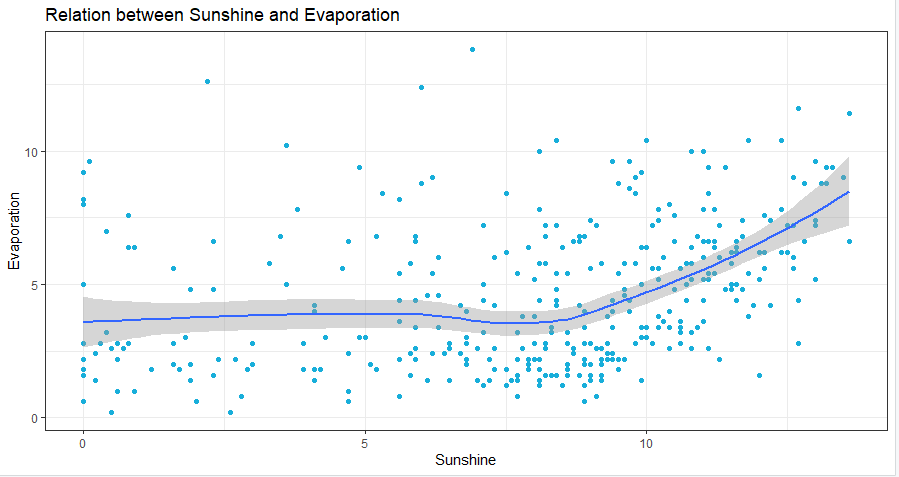
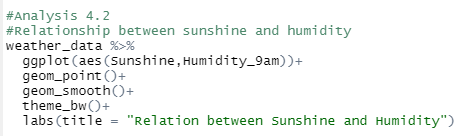


Figure 36: Output of analysis 4.1

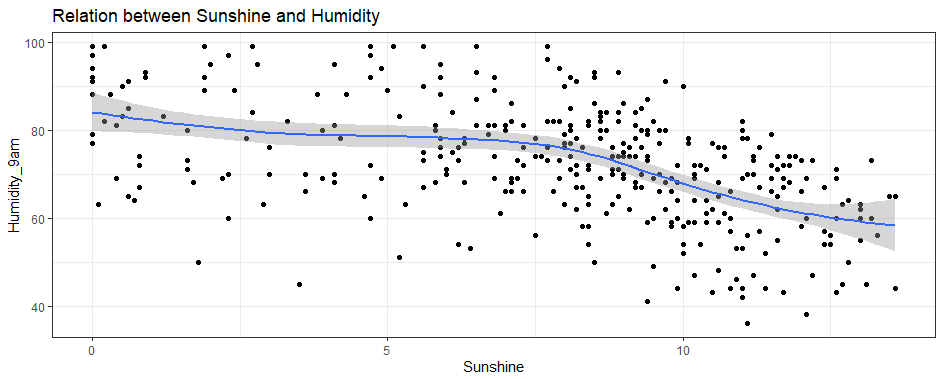
The figure above shows that the rate of evaporation tends to rise when the sunshine level is close to 10 or higher. After the sun's intensity reaches a higher level, which causes a rise in the rate of evaporation, it is likely that it will begin to rain in the United States.

Analysis 4.2 Relationship between sunshine and humidity



*Figure 37: Analysis 4.2*

This code is used to construct a point graph that compares whether or not the amount of sunshine throughout the year has an influence on the humidity level. The x-axis of the graph indicate the amount of sunshine, and the y-axis will represent the humidity level at 9 a.m.

****

*Figure 38: Output of analysis 4.2*

The figure above shows that when the sunshine level is close to 10 or higher, the rate of humidity tends to go down.

**Conclusion**

In conclusion, I have investigated and analysed the data on the weather in the United States using a variety of methods. I did this by employing the ggplot2 tools in order to generate a visual graph and draw a conclusion. Bar charts, box plots, and scatter graphs with linear lines are representations of graphs that are used in this report. According to the data given, there are a greater number of days in the United States that do not contain rain than there are days that do contain rain. This could be many reasons and one of it is the temperatures, which impact the degree of evaporation and, as a result, cause fewer days with rain.

**Reference**

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package ‘ggplot2’. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189.

Dinea, E. (2021, November 24). PREDICTION OF RAIN TOMORROW IN AUSTRALIA. Jovian. Retrieved December 17, 2022, from <https://jovian.ai/edaaydinea/prediction-of-raintomorrow>

R Programming 101. (2021, February 2). Ggplot for plots and graphs. An introduction to data visualization using R programming [Video]. YouTube. https://www.youtube.com/watch?v=HPJn1CMvtmI&list=LL&index=6