Climb Dataset - Data Analysis

Code

Rony Avivi - 207134347, Or Somech - 205984792

Introduction

Mount Rainier, “An Icon on the Horizon”,stands as an icon in the Washington landscape. Mount Rainier is an active volcano,ascending to 14,410 feet above sea level. It’s known to be the most glaciated peak in the contiguous U.S.A., spawning five major rivers.

In this study, we will analyze historical weather records and climbing records to examine the relationships between weather features (such as average temperature, wind speed, etc.) and months, and determine how weather affects climbing success rates.

In our project you will get familiar with the common weather conditions on this majestic mountain, the different routes to the summit, and the climbing statistics.

We will focus on:

1. Tidying our data.
2. Creating visualizations to understand relations between different features.
3. Creating tests and models to check a variety of hypotheses.

Our goals are to demonstrate and practice the different methods which we have learned about in the course by examining the relationship between the different variables from our data set.

Let`s start climbing!

Part One - Data Import And Tidying

Explaining our Dataset

the “Mount Rainier Weather and Climbing Data” was taken from kaggle.
The weather has been captured from https://www.nwac.us (https://www.nwac.us) and the climbing statistics from http://www.mountrainierclimbing.us/routes (http://www.mountrainierclimbing.us/routes). The data comes in a csv file.

Content:
climbing_statistics.csv - contains data for 4000+ groups who tried to climb the summit between 2014-2015.
Rainier_Weather.csv - contains data for 450+ daily weather conditions on the mountain.

Hide

```
# Reading the data
climbing_statistics <- read.csv('C:\\Users\\User\\Desktop\\data science\\D\\Data analysis\\Da
ta analysis Project\\climbing_statistics.csv')
weather_statistics <- read.csv('C:\\Users\\User\\Desktop\\data science\\D\\Data analysis\\Dat
a analysis Project\\Rainier_Weather.csv')
```

Let’s take a look at our data using glimpse function:

Code

```
## Rows: 4,077
## Columns: 5
## $ ï..Date            <chr> "11/27/2015", "11/21/2015", "10/15/2015", "10/13/20~
## $ Route              <chr> "Disappointment Cleaver", "Disappointment Cleaver",~
## $ Attempted          <int> 2, 3, 2, 8, 2, 10, 2, 2, 2, 2, 12, 12, 2, 12, 3, 12~
## $ Succeeded          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 6, 0, 3, 0, 9, 6, ~
## $ Success.Percentage <dbl> 0.00000000, 0.00000000, 0.00000000, 0.00000000, 0.0~
```

Code

```
## Rows: 464
## Columns: 7
## $ Date                  <chr> "12/31/2015", "12/30/2015", "12/29/2015", "12/28~
## $ Battery.Voltage.AVG   <dbl> 13.84500, 13.82292, 13.83458, 13.71042, 13.36250~
## $ Temperature.AVG       <dbl> 19.062917, 14.631208, 6.614292, 8.687042, 14.140~
## $ Relative.Humidity.AVG <dbl> 21.87083, 18.49383, 34.07292, 70.55792, 95.75417~
## $ Wind.Speed.Daily.AVG  <dbl> 21.977792, 3.540542, 0.000000, 0.000000, 0.00000~
## $ Wind.Direction.AVG    <dbl> 62.32583, 121.50542, 130.29167, 164.68375, 268.4~
## $ Solare.Radiation.AVG  <dbl> 84.915292, 86.192833, 85.100917, 86.241250, 31.0~
```

Explaining each feature:

climbing_statistics.csv

Date - a date between 2014-2015.
Route - route name.
Attempted - number of people attempted climbing to the summit on specific date.
Succeeded - number of people succeeded climbing to the summit on specific date.
Success Percentage - percentage of the people succeeded climbing to the summit on a specific date.

Rainier_Weather.csv

Date - a date between 2014-2015.
Battery Voltage AVG - average battery voltage.
Temperature AVG - day's average temperature.
Relative Humidity AVG - day's average humidity.
Wind Speed Daily AVG - day's average wind speed.
Wind Direction AVG - day's average wind direction.
Solare Radiation AVG - day's average solar radiation.

Tidying our data:

As you can see we have two separated data sets. Both are linked together by the "Date" column.
The problems we encountered to tidy our data:

1. The "Date" parameter in the climbing_statistics csv has punctuation characters in it.
2. The "Date" parameter in both data sets is char, we would like to convert it to numeric.
3. The "Succeeded" parameter was greater than the "Attempted" parameter in some of the rows because of wrong data.

Hide

```
# Fixing the "Date" parameter name in the climbing_statistics csv
colnames(climbing_statistics)[1] <- c("Date")
```

Hide

```
# Extract year and month - for climbing data
climbing_statistics$DATE<- as.Date(climbing_statistics$Date,format="%m/%d/%Y")
  climbing_statistics$YEAR<- as.numeric(format(climbing_statistics$DATE,"%Y"))
  climbing_statistics$MONTH<-as.numeric(format(climbing_statistics$DATE,'%m'))

# Sanity check
head(climbing_statistics, n=10)
```

```
##          Date                  Route Attempted Succeeded Success.Percentage
## 1  11/27/2015 Disappointment Cleaver         2         0                   0
## 2  11/21/2015 Disappointment Cleaver         3         0                   0
## 3  10/15/2015 Disappointment Cleaver         2         0                   0
## 4  10/13/2015          Little Tahoma         8         0                   0
## 5   10/9/2015 Disappointment Cleaver         2         0                   0
## 6   10/3/2015 Disappointment Cleaver        10         0                   0
## 7   10/3/2015 Disappointment Cleaver         2         0                   0
## 8   10/2/2015          Kautz Glacier         2         0                   0
## 9   10/2/2015 Disappointment Cleaver         2         0                   0
## 10  9/30/2015 Disappointment Cleaver         2         0                   0
##          DATE YEAR MONTH
## 1  2015-11-27 2015    11
## 2  2015-11-21 2015    11
## 3  2015-10-15 2015    10
## 4  2015-10-13 2015    10
## 5  2015-10-09 2015    10
## 6  2015-10-03 2015    10
## 7  2015-10-03 2015    10
## 8  2015-10-02 2015    10
## 9  2015-10-02 2015    10
## 10 2015-09-30 2015     9
```

Hide

```
# Extract year and month - for weather data
weather_statistics$DATE<-as.Date(weather_statistics$Date, format="%m/%d/%Y")
  weather_statistics$YEAR<-as.numeric(format(weather_statistics$DATE, "%Y"))
  weather_statistics$MONTH<-as.numeric(format(weather_statistics$DATE,"%m"))

# Sanity check
head(weather_statistics, n=10)
```

```
##          Date Battery.Voltage.AVG Temperature.AVG Relative.Humidity.AVG
## 1  12/31/2015            13.84500       19.062917              21.87083
## 2  12/30/2015            13.82292       14.631208              18.49383
## 3  12/29/2015            13.83458        6.614292              34.07292
## 4  12/28/2015            13.71042        8.687042              70.55792
## 5  12/27/2015            13.36250       14.140417              95.75417
## 6  12/26/2015            13.53167       17.512917              47.57458
## 7  12/25/2015            13.83708        3.215042              33.72250
## 8  12/24/2015            13.68167        2.815375              76.06583
## 9  12/23/2015            13.37167        2.005458              90.89167
## 10 12/22/2015            13.68125        4.028125              91.31667
##    Wind.Speed.Daily.AVG Wind.Direction.AVG Solare.Radiation.AVG       DATE YEAR
## 1             21.977792           62.32583             84.91529 2015-12-31 2015
## 2              3.540542          121.50542             86.19283 2015-12-30 2015
## 3              0.000000          130.29167             85.10092 2015-12-29 2015
## 4              0.000000          164.68375             86.24125 2015-12-28 2015
## 5              0.000000          268.47917             31.09071 2015-12-27 2015
## 6              0.000000          268.46667             43.40721 2015-12-26 2015
## 7              0.000000          268.47917             86.81050 2015-12-25 2015
## 8              0.000000          255.00000             76.49446 2015-12-24 2015
## 9              0.000000          244.96250             30.40321 2015-12-23 2015
## 10             0.000000          244.94583             70.34350 2015-12-22 2015
##    MONTH
## 1     12
## 2     12
## 3     12
## 4     12
## 5     12
## 6     12
## 7     12
## 8     12
## 9     12
## 10    12
```

Hide

```
# Extracting rows where the "Succeeded" value is greater than "Attempted" value

climbing_statistics <- climbing_statistics %>% filter(climbing_statistics$Succeeded <= climbi
ng_statistics$Attempted)
```

Part 2 - Visualization

We examined our data through different visualizations to analyze connections between different variables using the `ggplot2` package.

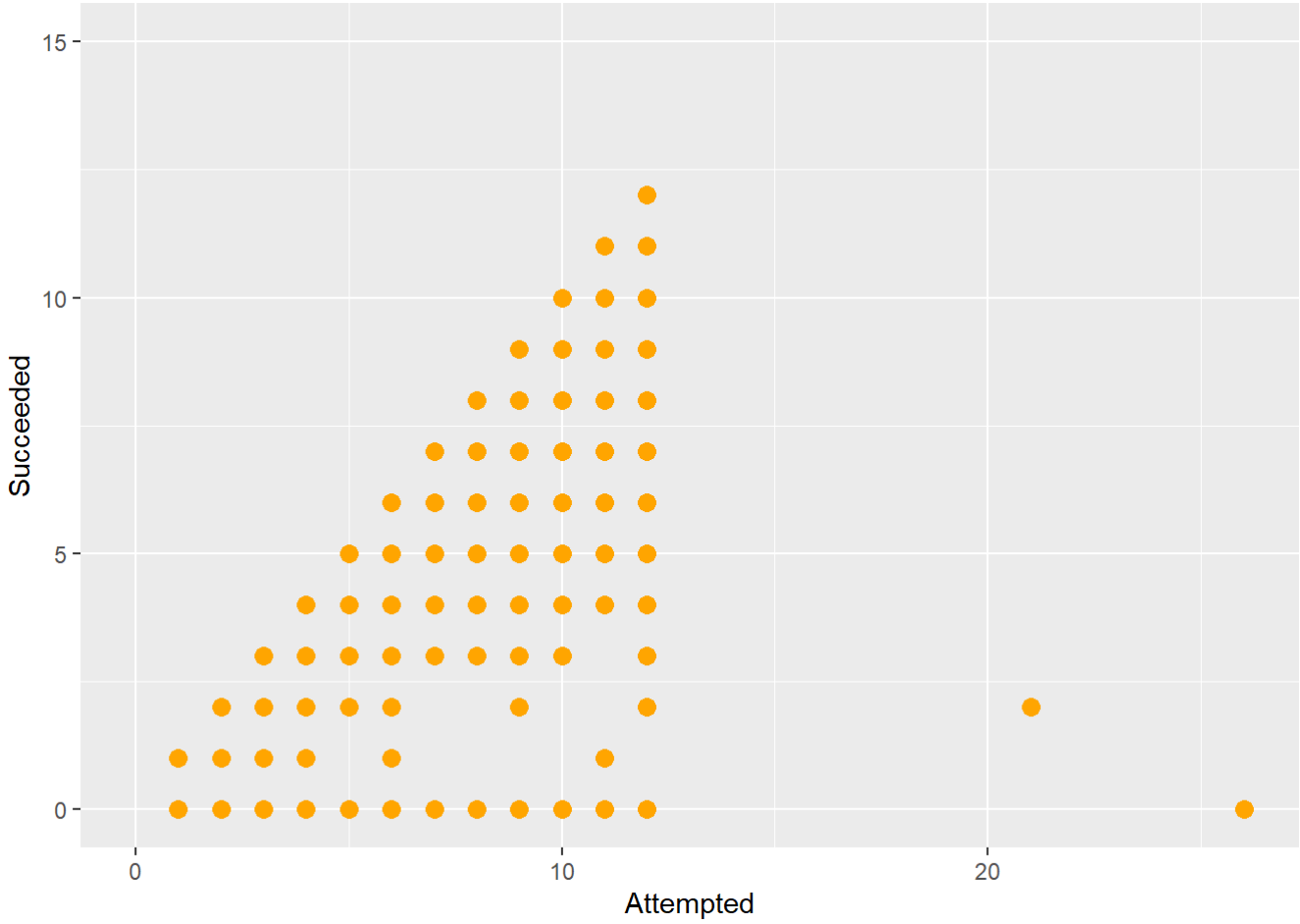
First, we want to create some general plots to help us understand our features better.

Attempts and successes

We created a plot that counts the number of people who attempted climbing the mountain and counts the number of those who succeeded reaching the summit.

Hide

```
ggplot(climbing_statistics, aes(x=Attempted, y=Succeeded),xlab("Attempt counts"), ylab("Succe
ss counts"))+ geom_point(color='orange',fill='orange',size=3)+ ylim(0,15) +xlim(0,26)
```

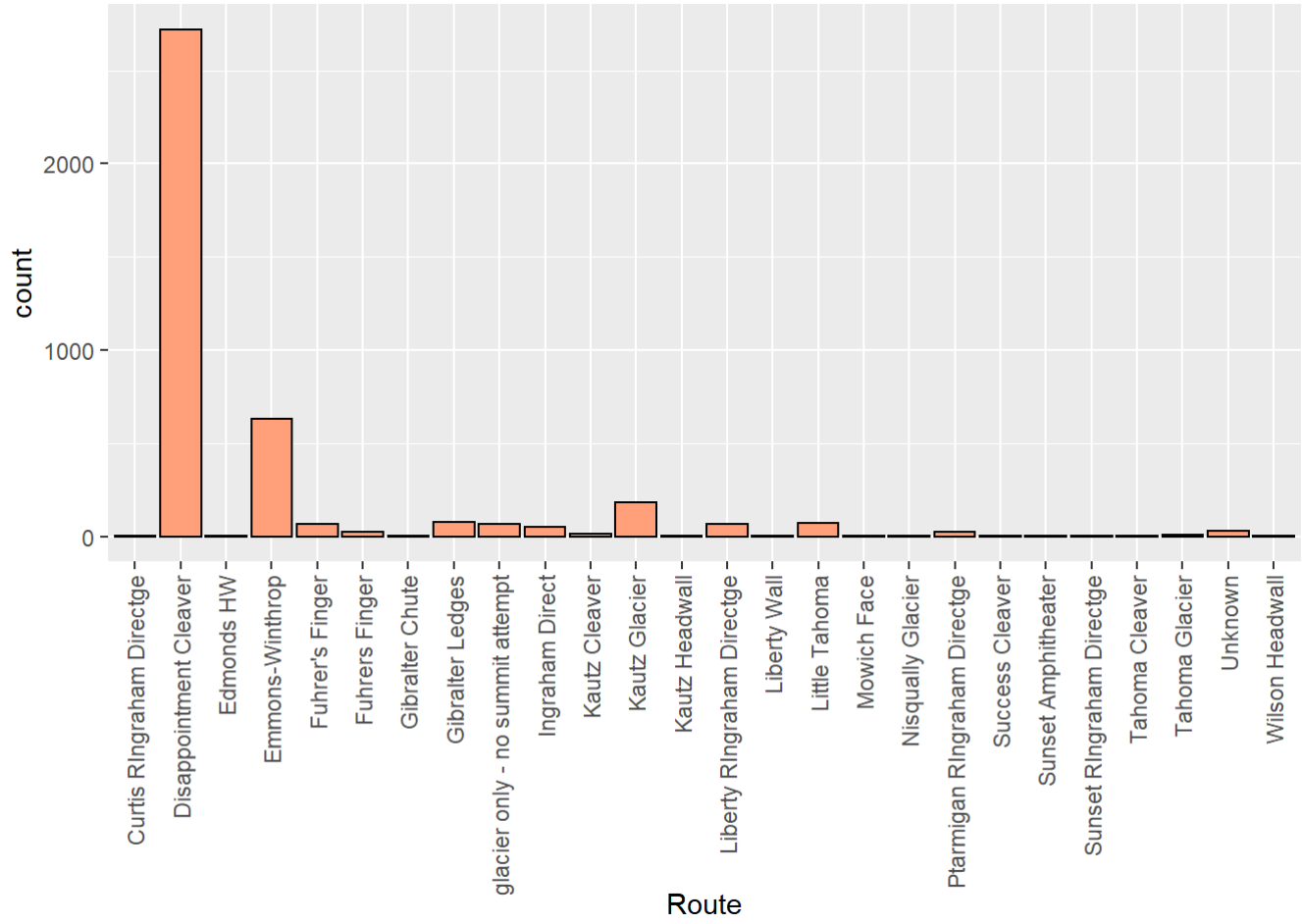


Plotting routes popularity

The routes parameter had 26 unique categories, let's check which is the most popular route!

Hide

```
ggplot(climbing_statistics,aes(x=Route)) + geom_bar(color = "black",fill="lightsalmon") + the
me(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



As we can see from the histogram, “Disapointment Cleaver” is the most popular route.

Attempts and successes over months

Let’s get down to business! we would like to check the connection between attempts and successes over month.

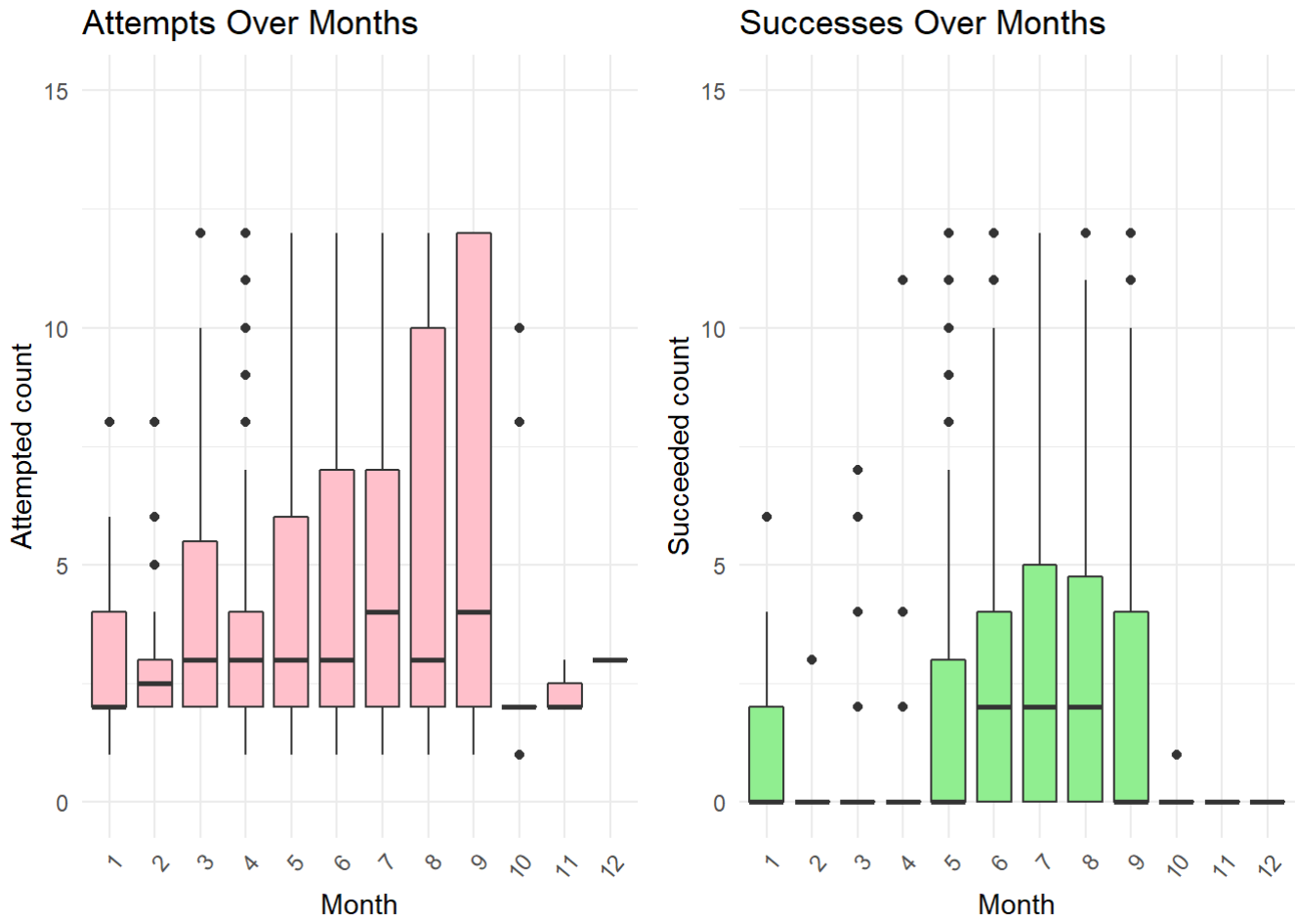
Hide

```
AttemptsChart<-ggplot(data=climbing_statistics, aes(x=as.factor(MONTH), y=Attempted)) + geom_
boxplot(fill="pink") +
  ggtitle("Attempts Over Months") + xlab("Month") + ylab("Attempted count") + theme_minimal()
+
  ylim(0,15) + theme(axis.text.x = element_text(angle = 50,  hjust = 1))

SuccessesChart<-ggplot(data=climbing_statistics, aes(x=as.factor(MONTH), y=Succeeded)) + geom
_boxplot(fill="lightgreen") +
  ggtitle("Successes Over Months") + xlab("Month") + ylab("Succeeded count") + theme_minimal
() +
  ylim(0,15) + theme(axis.text.x = element_text(angle = 50, hjust = 1))

  grid.arrange(AttemptsChart, SuccessesChart, nrow= 1)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



As we can see from the box plots, although most of the people attempt climbing on September, the number of people who succeeded reaching the summit is greater in July.

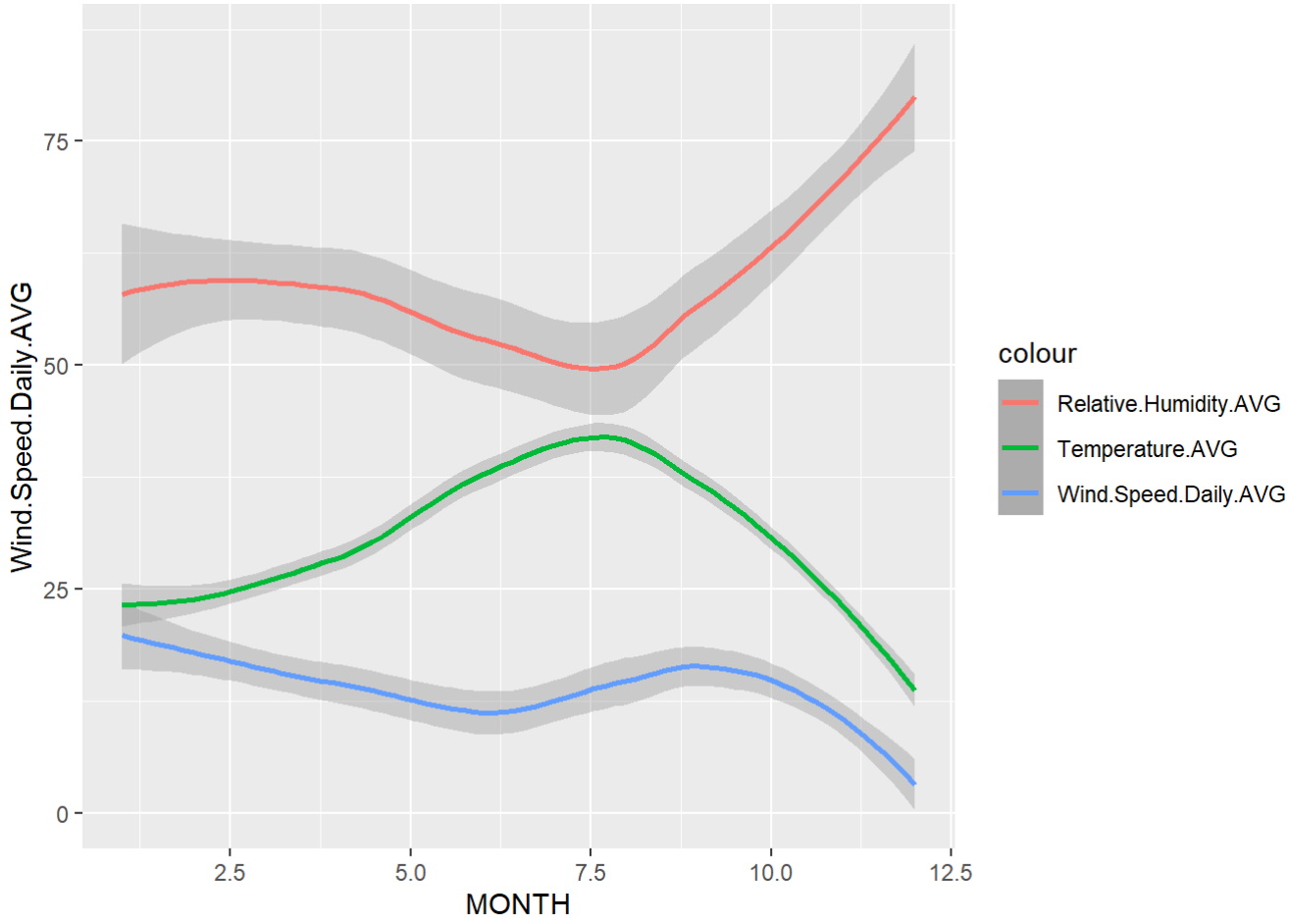
Because we know most successes happen in the summer, we would like to check the relations between some weather conditions (that we are assuming have a great affect on the success rate) across months and to see if there are any differences between June, July, and August, and the rest of the year.

AVG Wind Speed, Relative Humidity and Temperature over Months

Hide

```
ggplot()+
geom_smooth(weather_statistics,mapping =  aes(x=MONTH,y=Wind.Speed.Daily.AVG,color="Wind.Spee
d.Daily.AVG"))+
geom_smooth(weather_statistics,mapping = aes(x=MONTH,y=Relative.Humidity.AVG,color="Relative.
Humidity.AVG"))+
geom_smooth(weather_statistics,mapping = aes(x=MONTH,y=Temperature.AVG,color="Temperature.AV
G"))
```

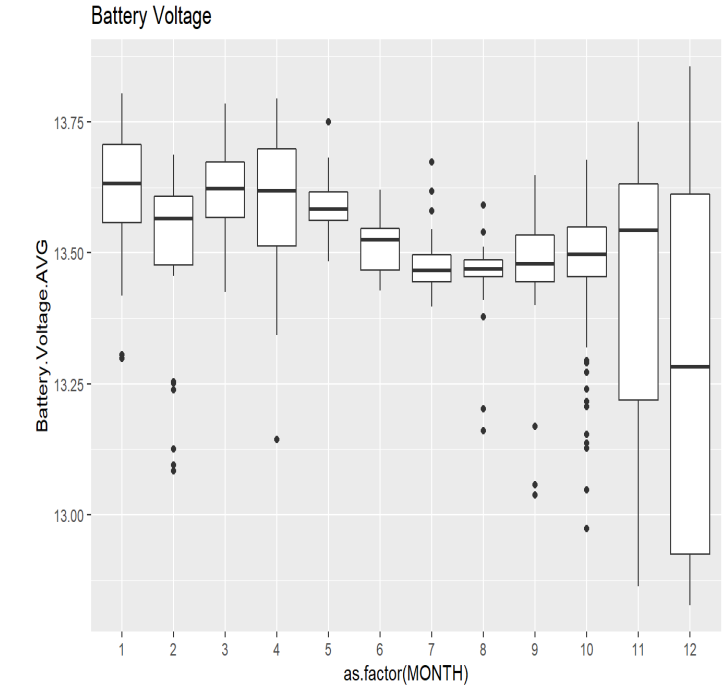
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

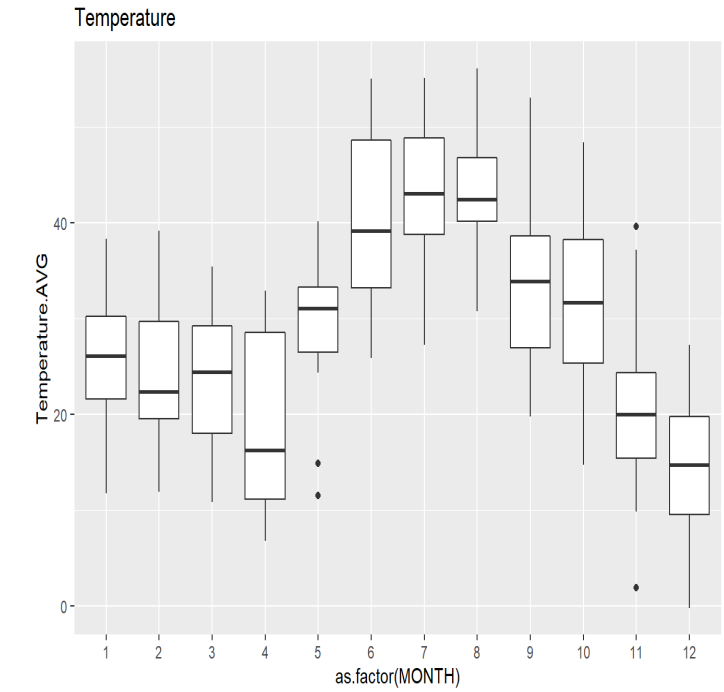


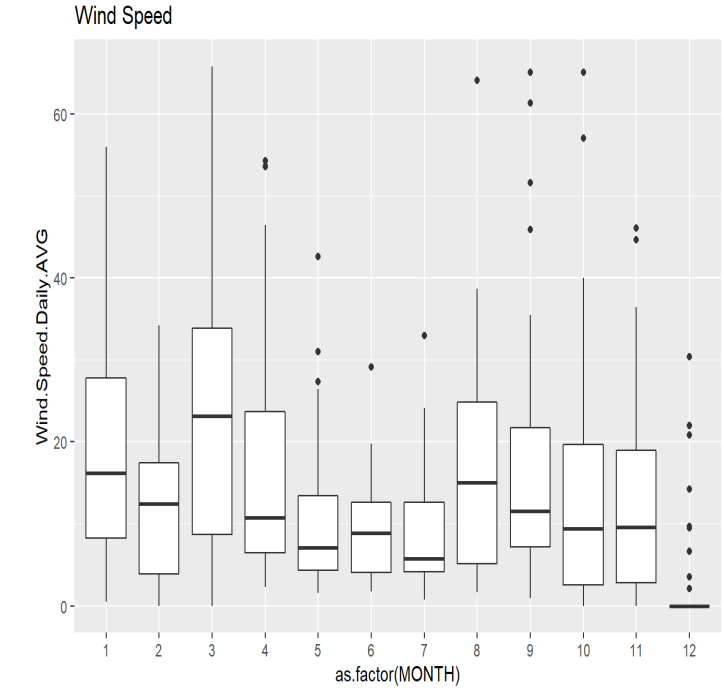
Judging from the graph, the average relative humidity and the average daily wind speed starts to decrease while the average temperature starts to increase when near month of June.
Perhaps, lower average relative humidity, lower average daily wind speed and higher average temperature will increase the chance of success.

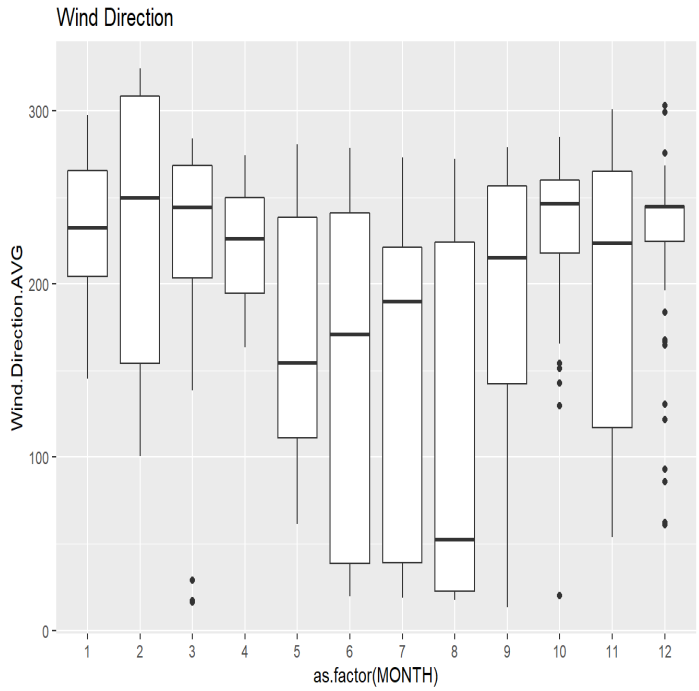
Now, let’s expend our check and examine all of the relations between the weather features across months.

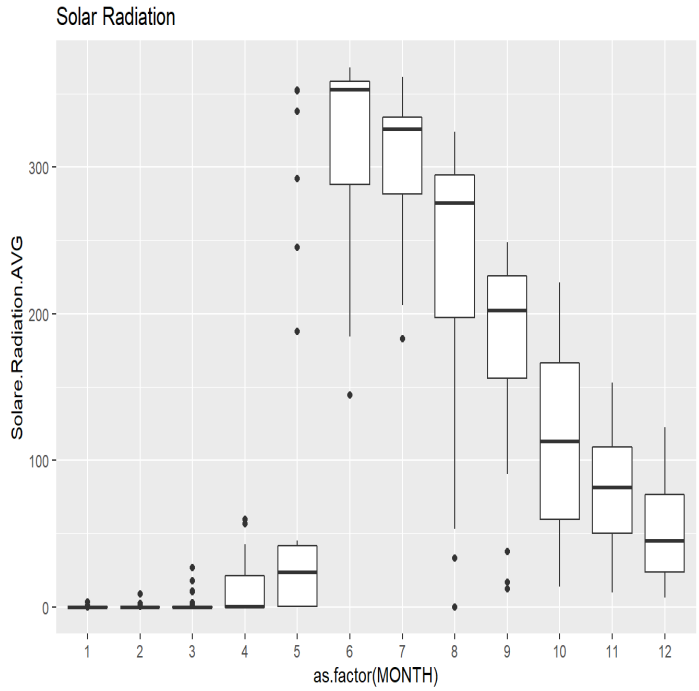
Weather parameters over months


Battery Voltage
Battery.Voltage.AVG
13.75
13.50
13.25
13.00
1
2
3
4
5
6
7
8
9
10
11
12
as.factor(MONTH)


Temperature
Temperature.AVG
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
as.factor(MONTH)


Wind Speed
Wind.Speed.Daily.AVG
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
as.factor(MONTH)





From the box plots we can see the weather features distribution in each month.
Generally, we can see that compared to the rest of the year, in the summer months the battery voltage is lower,the temperature is higher, the wind speed is lower and the solar radiation is higher.

Part 3 - Modeling

ANOVA test:

Analysis of Variance (ANOVA) is a statistical technique, commonly used to studying differences between two or more group means. ANOVA in R primarily provides evidence of the existence of the mean equality between the groups. This statistical method is an extension of the t-test.

We want to test our weather data variables across months using ANOVA test.

- H_0 : the means of the different groups are the same.
- H_1 : At least one sample mean is not equal to the others.

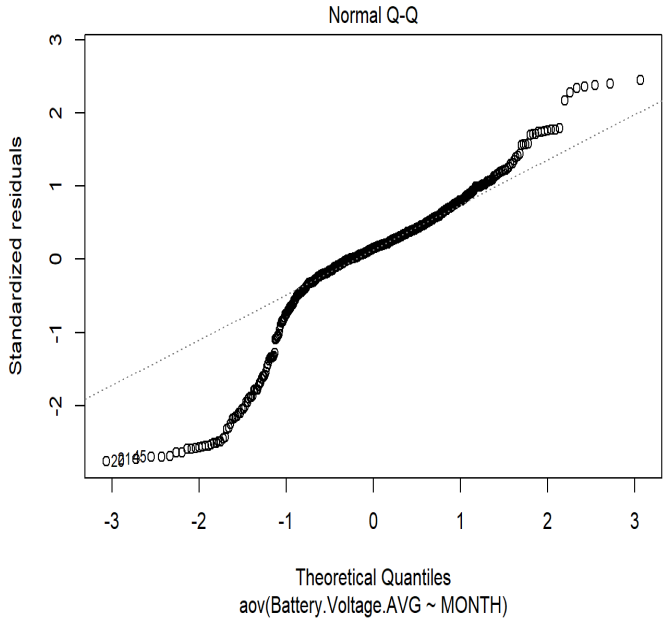
Hide

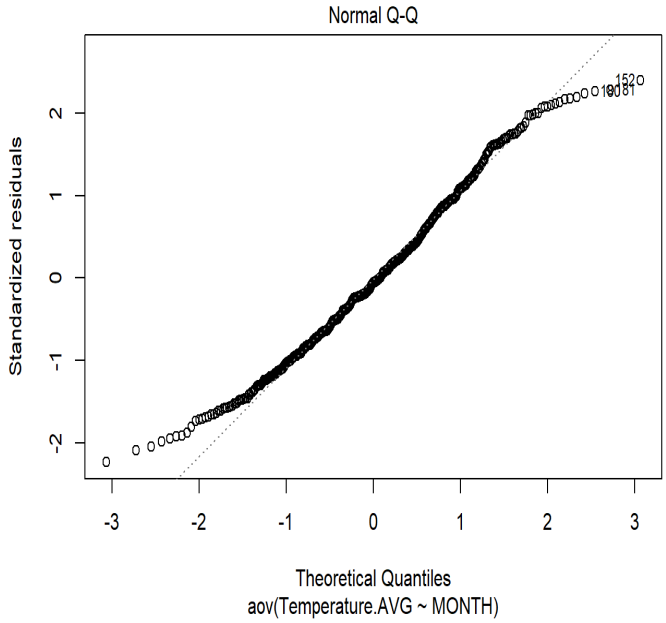
```
BatteryOverMonth <- aov(Battery.Voltage.AVG ~ MONTH, data=weather_statistics)
TempOverMonth <- aov(Temperature.AVG ~ MONTH, data=weather_statistics)
HumidityOverMonth <- aov(Relative.Humidity.AVG ~ MONTH, data=weather_statistics)
WindSpeedOverMonth <- aov(Wind.Speed.Daily.AVG ~ MONTH, data=weather_statistics)
WindDirectionOverMonth <- aov(Wind.Direction.AVG ~ MONTH, data=weather_statistics)
SolarRadiationOverMonth <- aov(Solare.Radiation.AVG ~ MONTH, data=weather_statistics)
```

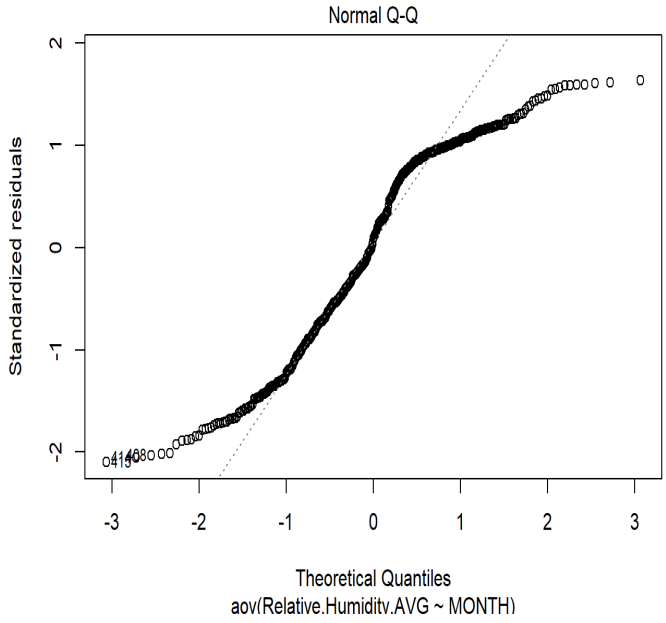
Before we use this test we need to assume:

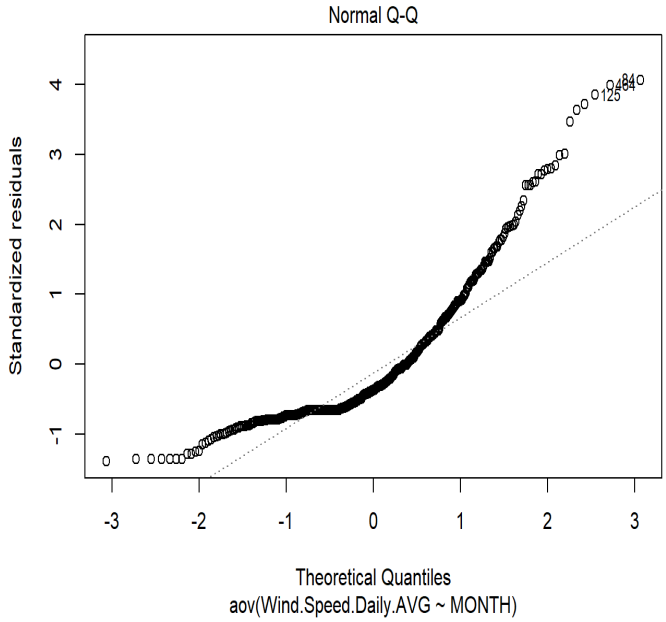
1. The data of each factor level are normally distributed.
2. These normal populations have a common variance.

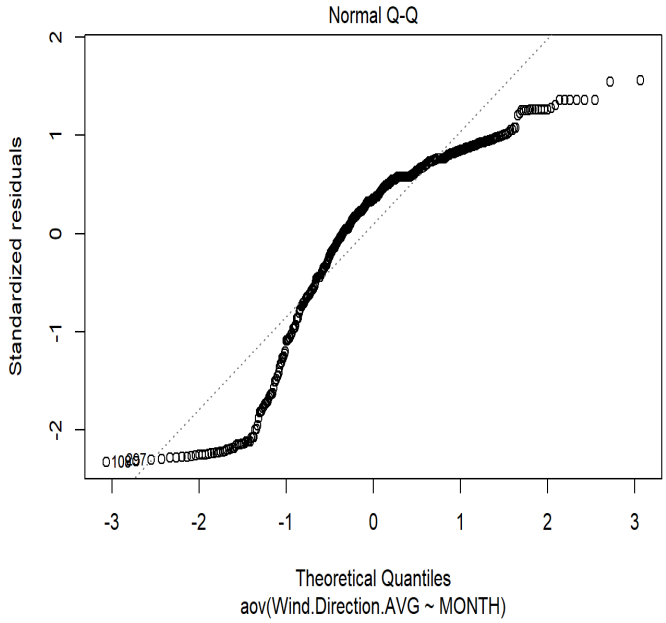
1. Checking the normality assumption

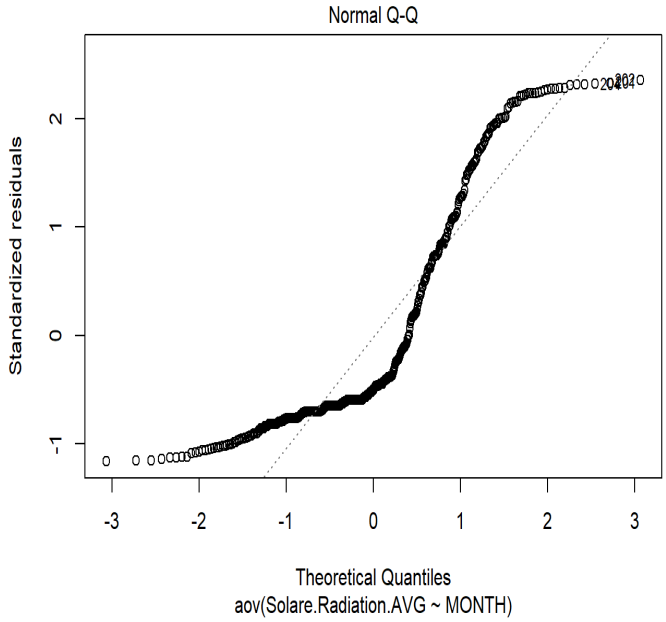

Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Battery.Voltage.AVG ~ MONTH)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Temperature.AVG ~ MONTH)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Relative.Humiditv.AVG ~ MONTH)

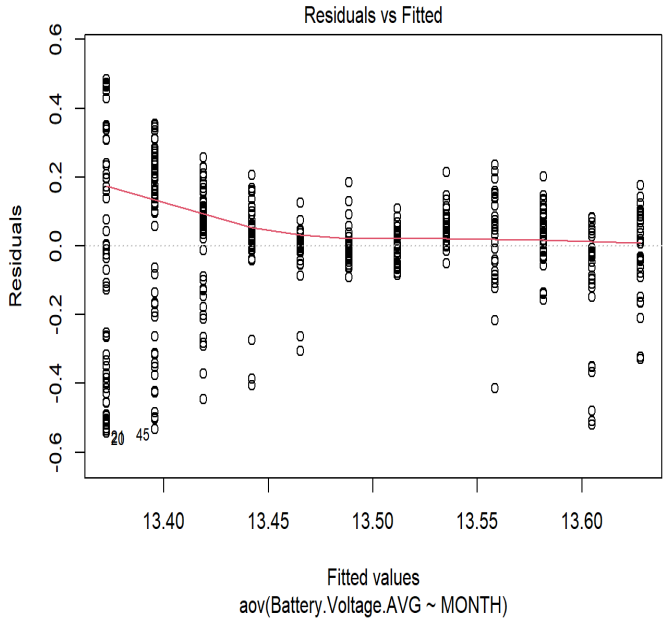

Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Wind.Speed.Daily.AVG ~ MONTH)

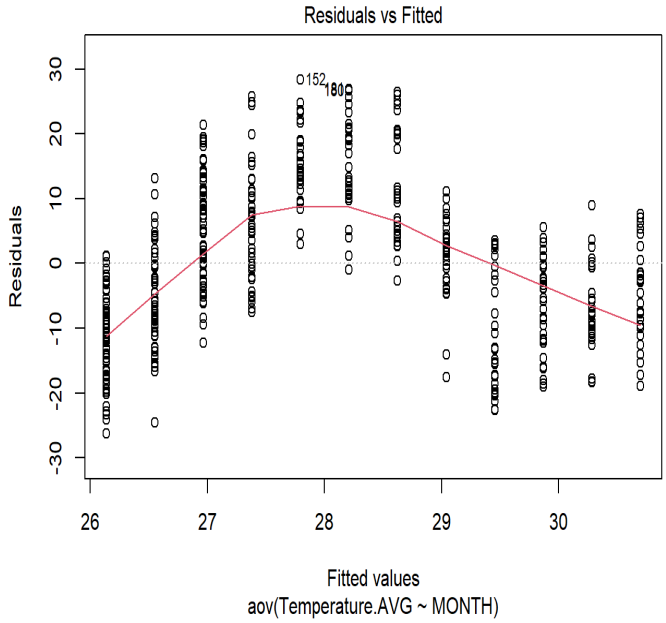

Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Wind.Direction.AVG ~ MONTH)

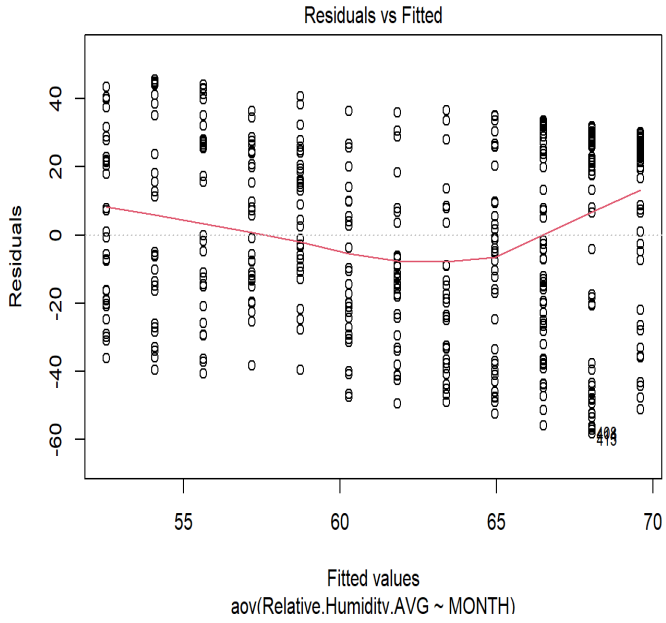


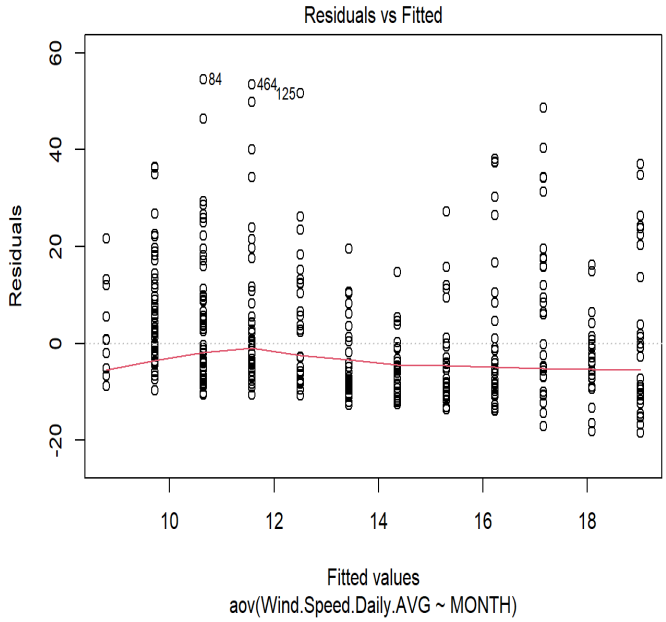
In all of the features, all the points fall approximately along the reference line, so we can assume normality.

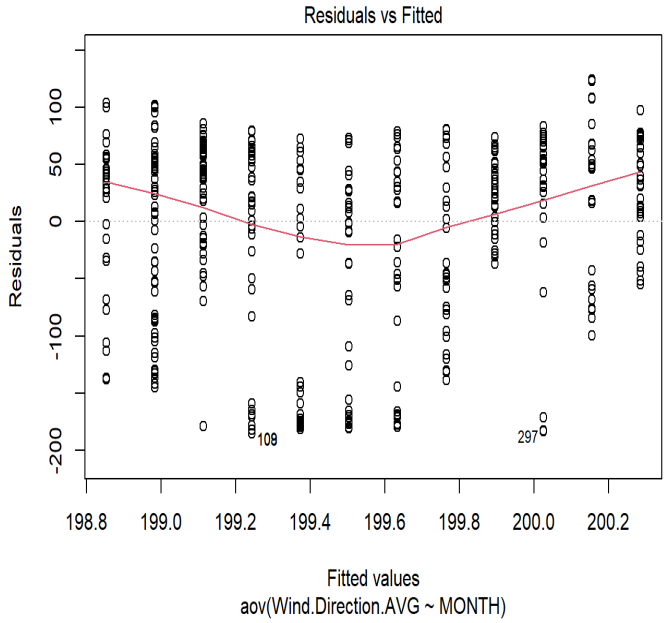
2. Checking homogeneity of variances

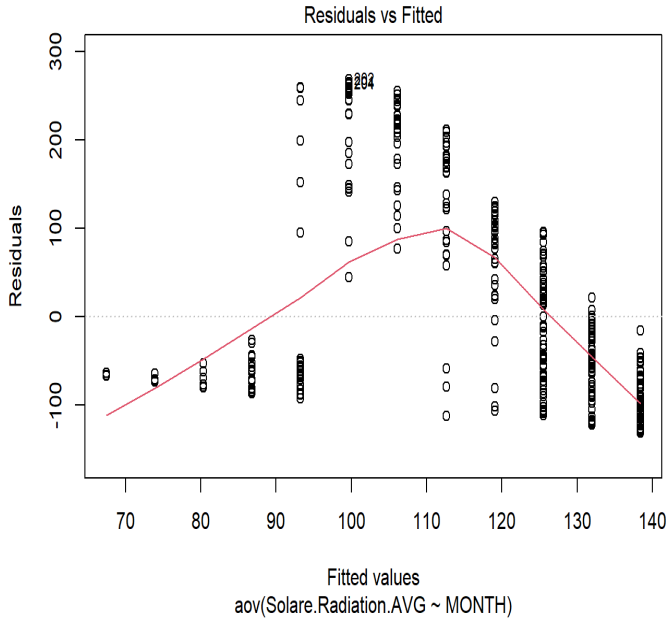

Residuals vs Fitted
Residuals
Fitted values
aov(Battery.Voltage.AVG ~ MONTH)


Residuals vs Fitted
Residuals
Fitted values
aov(Temperature.AVG ~ MONTH)


Residuals vs Fitted
Residuals
Fitted values
aov(Relative.Humiditv.AVG ~ MONTH)


Residuals vs Fitted
Residuals
Fitted values
aov(Wind.Speed.Daily.AVG ~ MONTH)


Residuals vs Fitted
Residuals
Fitted values
aov(Wind.Direction.AVG ~ MONTH)



We can see that: 'TempOverMonth', 'WindDirectionOverMonth' and 'SolarRadiationOverMonth' failed the test for homogeneity of variance.
We chose to **not** conduct the ANOVA test on does parameters because this result gives us a strong evidence that the groups are not selected from identical populations. We haven't yet tested whether the means are distinct, but we already know that the variances are different. This is why that may be a good stopping point, because we have strong evidence that the populations the data are sampled from are not identical.

Hide

```
# Performing the ANOVA test
# Summary of the analysis for: BatteryOverMonth, HumidityOverMonth and WindSpeedOverMonth.

summary(BatteryOverMonth)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## MONTH         1   3.18   3.180   81.36 <2e-16 ***
## Residuals   462  18.06   0.039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
summary(HumidityOverMonth)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## MONTH         1  14178   14178   18.23 2.38e-05 ***
## Residuals   462 359311     778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
summary(WindSpeedOverMonth)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## MONTH         1   5093    5093   28.27 1.65e-07 ***
## Residuals   462  83235     180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpret the ANOVA Results:

Df program: The degrees of freedom for the variable program. This is calculated as #groups -1.
In our case, there were 2 different workout programs in all the test, so this value is: 2-1 = 1.

Df Residuals: The degrees of freedom for the residuals. This is calculated as #total observations – # groups.

Sum Sq program: The sum of squares associated with the variable program.

Sum Sq Residuals: The sum of squares associated with the residuals or “errors”.

Mean Sq. Program: The mean sum of squares associated with program. This is calculated as Sum Sq. program / Df program.

Mean Sq. Residuals: The mean sum of squares associated with the residuals. This is calculated as Sum Sq. residuals / Df residuals.

F Value: The overall F-statistic of the ANOVA model. This is calculated as Mean Sq. program / Mean sq.

Pr(>F): The p-value associated with the F-statistic with numerator df and denominator df.
The most *important* value in the entire output is the p-value because this tells us whether there is a significant difference in the mean values between the three groups.

Conclusions

Since the p-values in all of our ANOVA tables are extremely tiny numbers and less than 0.05, we have sufficient evidence to **reject** all of the null hypothesis.

Linear reggression

Background

Due to the fact that there is a connection between temperature and the number of people who succeeded to reach the mountain summit, we assume that as the temperature rises so is the number of people’s successes.

Transformation

Before we started analyzing the data, we wanted to make sure that it is reliable.
We can see that there is a gap between the dates in the “weather_statistics” table and the “climbing_statistics” table:

1. In “weather_statistics” the earliest and the latest date are not the same as in “climbing_statistics”.
2. In “climbing_statistics” there are several appearances of the same date. In each appearance there is different value of success according to the route that was chosen.

Here is an example for the mismatches:

Hide

```
# climbing_statistics.csv
head(climbing_statistics,7)
```

```
##         Date                  Route Attempted Succeeded Success.Percentage
## 1 11/27/2015 Disappointment Cleaver         2         0                  0
## 2 11/21/2015 Disappointment Cleaver         3         0                  0
## 3 10/15/2015 Disappointment Cleaver         2         0                  0
## 4 10/13/2015          Little Tahoma         8         0                  0
## 5  10/9/2015 Disappointment Cleaver         2         0                  0
## 6  10/3/2015 Disappointment Cleaver        10         0                  0
## 7  10/3/2015 Disappointment Cleaver         2         0                  0
##         DATE YEAR MONTH
## 1 2015-11-27 2015    11
## 2 2015-11-21 2015    11
## 3 2015-10-15 2015    10
## 4 2015-10-13 2015    10
## 5 2015-10-09 2015    10
## 6 2015-10-03 2015    10
## 7 2015-10-03 2015    10
```

Hide

```
# Rainier_Weather.csv
head(weather_statistics,7)
```

```
##         Date Battery.Voltage.AVG Temperature.AVG Relative.Humidity.AVG
## 1 12/31/2015            13.84500       19.062917              21.87083
## 2 12/30/2015            13.82292       14.631208              18.49383
## 3 12/29/2015            13.83458        6.614292              34.07292
## 4 12/28/2015            13.71042        8.687042              70.55792
## 5 12/27/2015            13.36250       14.140417              95.75417
## 6 12/26/2015            13.53167       17.512917              47.57458
## 7 12/25/2015            13.83708        3.215042              33.72250
##   Wind.Speed.Daily.AVG Wind.Direction.AVG Solare.Radiation.AVG       DATE YEAR
## 1            21.977792           62.32583             84.91529 2015-12-31 2015
## 2             3.540542          121.50542             86.19283 2015-12-30 2015
## 3             0.000000          130.29167             85.10092 2015-12-29 2015
## 4             0.000000          164.68375             86.24125 2015-12-28 2015
## 5             0.000000          268.47917             31.09071 2015-12-27 2015
## 6             0.000000          268.46667             43.40721 2015-12-26 2015
## 7             0.000000          268.47917             86.81050 2015-12-25 2015
##   MONTH
## 1    12
## 2    12
## 3    12
## 4    12
## 5    12
## 6    12
## 7    12
```

As we can see, in the file “climbing_statistics.csv” the latest date is 27/11/2015 while in the file “Rainier_Weather.csv” it’s 31/12/2015. Also we see that in “climbing_statistics.csv” the date 3/10/2015 appears twice.

We will transform the data so we could examine the influence of the temperature on the number of success climbs in each date.

Hide

```
# Transforming the data
new_climbing_statistics <- climbing_statistics %>% filter(climbing_statistics$DATE >= as.Date
("2014-09-23"))
new_weather_statistics <- weather_statistics %>% filter(weather_statistics$DATE <= as.Date("2
015-11-27"))

temperature_success_list <- data.frame(new_weather_statistics$DATE,new_weather_statistics$Tem
perature.AVG)
temperature_success_list$success_num <- 0

temp_success_list <- data.frame(new_climbing_statistics$DATE,new_climbing_statistics$Succeede
d)


for(i in 1:length(temperature_success_list$new_weather_statistics.DATE))
{
  day <- temperature_success_list$new_weather_statistics.DATE[i]
  success_sum <- 0
  for(j in 1:length(temp_success_list$new_climbing_statistics.DATE))
  {
    if(temp_success_list$new_climbing_statistics.DATE[j] == day)
    {
      success_sum <- success_sum + temp_success_list$new_climbing_statistics.Succeeded[j]
    }
  }
  temperature_success_list$success_num[i] <- success_sum
}
```

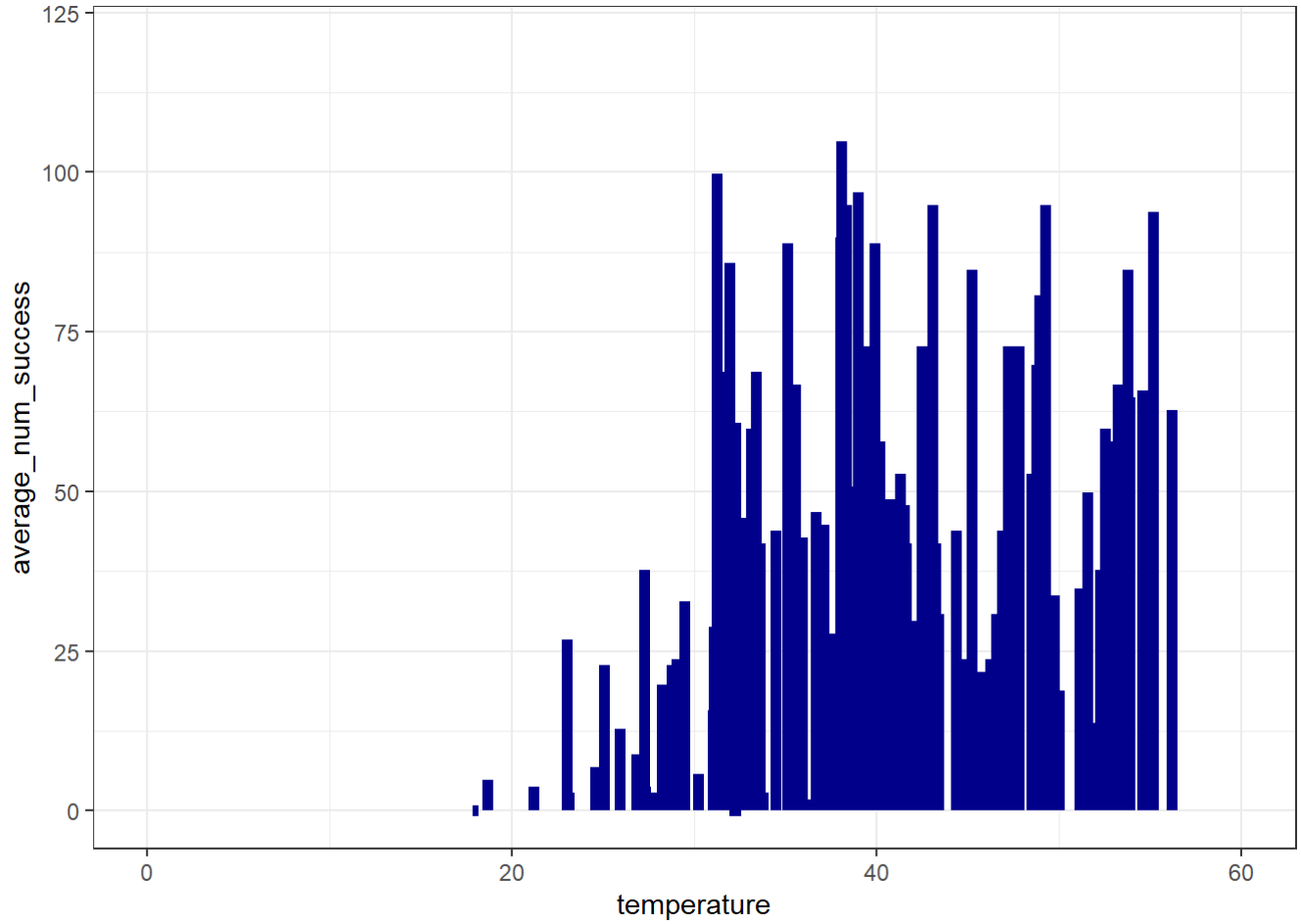
Test our theory

In order to test our theory, we calculated an average success for each of the temperatures that was measured in the sample. The results can be seen in the following graph:

Hide

```
mean <-aggregate(temperature_success_list[,3], list(temperature_success_list$new_weather_stat
istics.Temperature.AVG), mean)
mean<-rename(mean,temperature=Group.1)
mean<-rename(mean,num_of_success=x)

ggplot(data=mean, aes(x=temperature, y =num_of_success))+
  geom_histogram(stat='identity',fill=I("lightblue"),
      col=I("darkblue"), size=2)+theme_bw() + labs(y = "average_num_success") + xlim(0,60) +
 ylim(0,120)
```



It can be seen in the graph that according to our hypothesis, there might be a connection between temperature and the number of climbers who succeeded to reach the summit. We will now examine whether it really exists and if so what is it.

Assumptions for linear reggression

Before we use this test we need to assume:

1. ϵ distributes normally with $(0, \sigma)$.
2. ϵ is homoscedastic.

We will check if our data is homoscedastic and if it distributes normally in the graphs below:

Hide

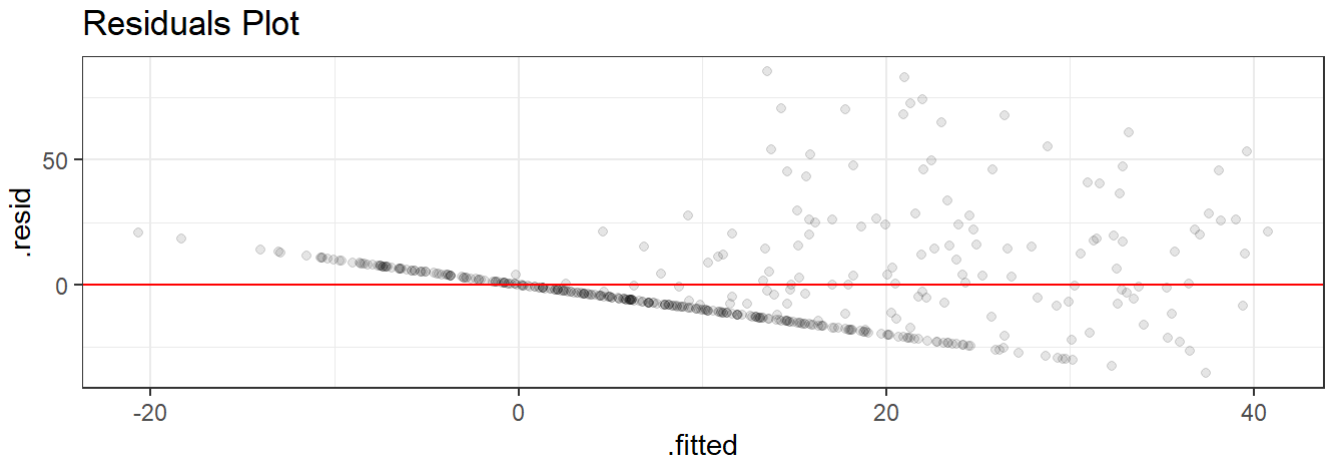
```
temperature_success_lm <- lm(success_num ~ new_weather_statistics.Temperature.AVG, data = tem
perature_success_list, conf.level = 0.95)
```

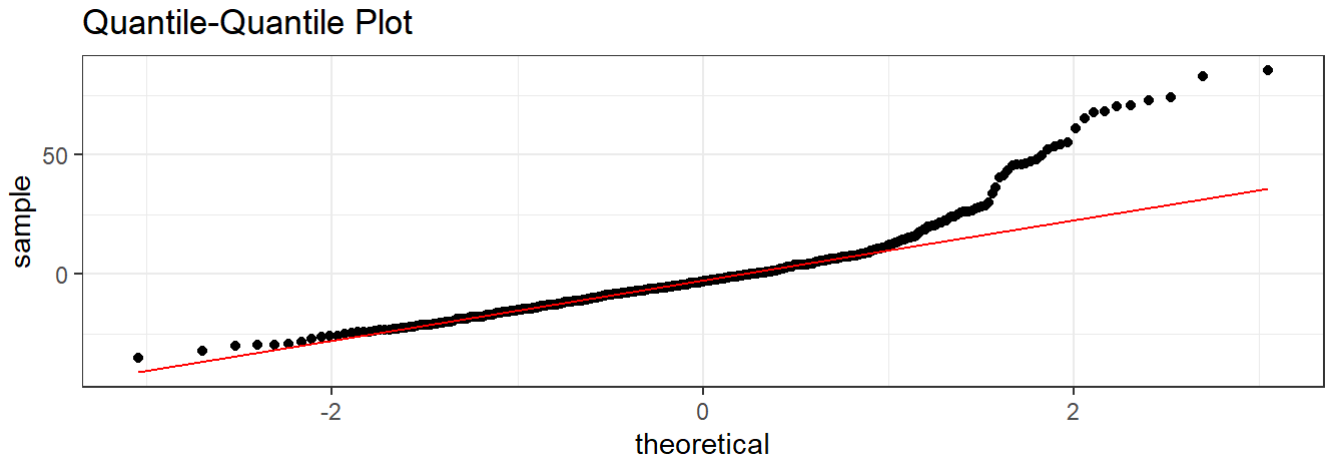
Hide

```
residuals_plot <- temperature_success_lm %>%
  ggplot(aes(x = .fitted, y = .resid)) + geom_point(alpha = 0.1) + geom_hline(yintercept = 0,
color = "red") + labs(title = "Residuals Plot")+theme_bw()

residuals_qq <- temperature_success_lm %>%
  ggplot(aes(sample = .resid)) + geom_qq() + geom_qq_line(col="red") + labs(title = "Quantile
-Quantile Plot")+theme_bw() + labs(x = "theoretical", y = "sample")

grid.arrange(residuals_plot, residuals_qq, nrow = 2)
```





The Residuals Plot shows that the data might be heteroscedastic and the QQ Plot shows that the data is distributed normally.
Nevertheless, we will assume that our data is homoscedastic in order to use the linear regression test.

Modeling

We would like to calculate the regression equation so that by using temperature we can predict the average number of people who succeeded climbing. That is, to find the connection between the weather and the success of climbing the summit.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Our hypothesis is that there is a connection between the temperature and the number of people who succeeded climbing.
Our hypothesis system:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Hide

```
temperature_success_lm
```

```
##
## Call:
## lm(formula = success_num ~ new_weather_statistics.Temperature.AVG,
##     data = temperature_success_list, conf.level = 0.95)
##
## Coefficients:
##                             (Intercept)  new_weather_statistics.Temperature.AVG
##                                  -20.49                                   1.09
```

$$Y = -20.49 + 1.09X$$

Hide

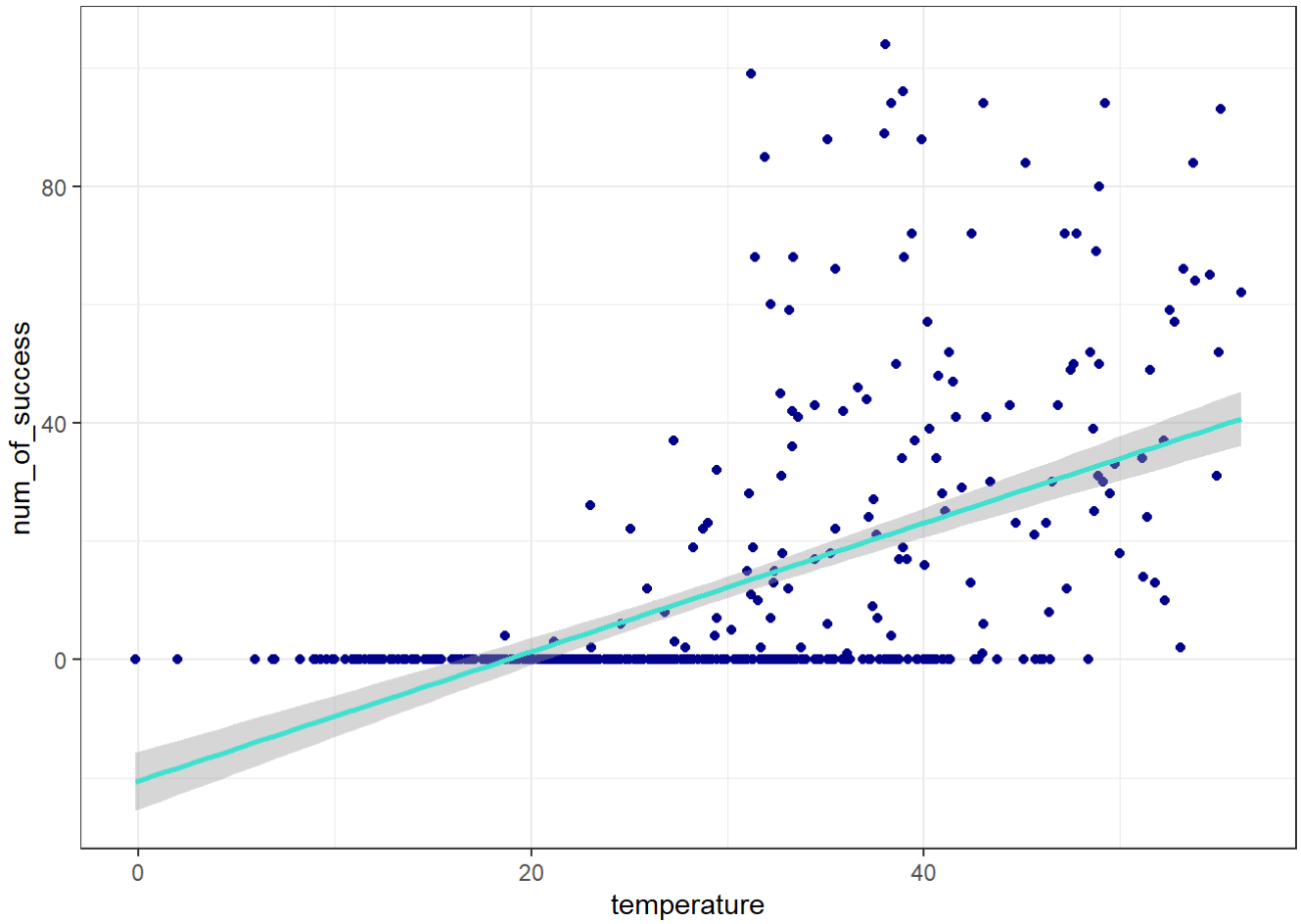
```
summary(temperature_success_lm)
```

```
##
## Call:
## lm(formula = success_num ~ new_weather_statistics.Temperature.AVG,
##     data = temperature_success_list, conf.level = 0.95)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.408 -11.290  -3.221   5.794  85.458
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -20.49261    2.49742  -8.206 2.72e-15
## new_weather_statistics.Temperature.AVG   1.09029    0.07972  13.676  < 2e-16
##
## (Intercept)                          ***
## new_weather_statistics.Temperature.AVG ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.9 on 428 degrees of freedom
## Multiple R-squared:  0.3041, Adjusted R-squared:  0.3025
## F-statistic:   187 on 1 and 428 DF,  p-value: < 2.2e-16
```

Hide

```
ggplot(mean, aes(x=temperature,y=num_of_success))+geom_point(color="darkblue")+
  stat_smooth(method = "lm", color = "turquoise")+theme_bw()
```

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



Conclusions

From the linear regression test, it can be seen that β_1 > 0 and P-value < α, therefore we reject the H_0 hypothesis at a confidence level of 95%. Meaning that we confirm our hypothesis that there is a correlation between temperature and the number of people who succeed in climbing the mountain.

Part 4 - Summary

In this research we used our experience and the techniques we learned during our course “Introduction to statistic and data analysis in R”.
First, we have arranged the data and subtracted unnecessary values to get more relevant results. After that we did some visualizations and checked connections and correlations between different parameters.
We did statistical tests and models to examine hypothesizes regarding the weather features and the months.
Finally we checked the connection between the average temperature and the success climbing rate to better understand connection between them.

Thank you for reading!