Assessment 3

A1844790 - Dylan Rohan

## Executive summary

## Methods

## Results

## Discussion

## Conclusion

## References

## Appendix

library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("dplyr")  
library("lubridate")

## Warning: package 'lubridate' was built under R version 4.1.2

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library("stringr")  
library("inspectdf")  
library("olsrr")

## Warning: package 'olsrr' was built under R version 4.1.2

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

library("e1071")   
mwc <- read\_csv("melbourne.csv")

## Rows: 365 Columns: 21

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): Date, Direction of maximum wind gust, 9am wind direction, 9am win...  
## dbl (15): Minimum temperature (Deg C), Maximum Temperature (Deg C), Rainfal...  
## time (1): Time of maximum wind gust

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

knitr::kable(head(mwc), caption = "The data prior to cleaning")

The data prior to cleaning

| Date | Minimum temperature (Deg C) | Maximum Temperature (Deg C) | Rainfall (mm) | Evaporation (mm) | Sunshine (hours) | Direction of maximum wind gust | Speed of maximum wind gust (km/h) | Time of maximum wind gust | 9am Temperature (Deg C) | 9am relative humidity (%) | 9am cloud amount (oktas) | 9am wind direction | 9am wind speed (km/h) | 9am MSL pressure (hPa) | 3pm Temperature (Deg C) | 3pm relative humidity (%) | 3pm cloud amount (oktas) | 3pm wind direction | 3pm wind speed (km/h) | 3pm MSL pressure (hPa) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2019-01-1 | 15.5 | 26.2 | 0.0 | 7.0 | 11.0 | S | 35 | 17:44:00 | 19.8 | 74 | 7 | S | 6 | 1013.0 | 24.4 | 45 | 1 | SSW | 11 | 1011.5 |
| 2019-01-2 | 18.4 | 22.2 | 0.0 | 7.0 | 7.5 | SSW | 39 | 15:23:00 | 19.5 | 64 | 8 | SSE | 7 | 1013.9 | 21.4 | 62 | 1 | SSW | 19 | 1012.9 |
| 2019-01-3 | 15.9 | 29.5 | 0.0 | 6.6 | 9.3 | SSW | 26 | 14:53:00 | 18.1 | 75 | 8 | S | 2 | 1012.6 | 24.6 | 60 | 0 | SSW | 13 | 1009.9 |
| 2019-01-4 | 18.0 | 42.6 | 0.0 | 7.8 | 12.2 | NW | 54 | 12:03:00 | 29.5 | 31 | 0 | NNE | 9 | 1005.5 | 42.0 | 16 | 1 | NW | 15 | 1001.0 |
| 2019-01-5 | 17.4 | 21.2 | 0.4 | 15.4 | 5.8 | SSW | 39 | 08:24:00 | 18.0 | 63 | 7 | S | 13 | 1013.5 | 19.1 | 58 | 7 | S | 11 | 1013.4 |
| 2019-01-6 | 14.6 | 22.1 | 1.4 | 6.4 | 13.3 | SSW | 33 | 11:12:00 | 17.7 | 55 | 1 | SW | 9 | 1020.4 | 20.6 | 48 | 1 | SSW | 13 | 1019.5 |

colnames(mwc)

## [1] "Date" "Minimum temperature (Deg C)"   
## [3] "Maximum Temperature (Deg C)" "Rainfall (mm)"   
## [5] "Evaporation (mm)" "Sunshine (hours)"   
## [7] "Direction of maximum wind gust" "Speed of maximum wind gust (km/h)"  
## [9] "Time of maximum wind gust" "9am Temperature (Deg C)"   
## [11] "9am relative humidity (%)" "9am cloud amount (oktas)"   
## [13] "9am wind direction" "9am wind speed (km/h)"   
## [15] "9am MSL pressure (hPa)" "3pm Temperature (Deg C)"   
## [17] "3pm relative humidity (%)" "3pm cloud amount (oktas)"   
## [19] "3pm wind direction" "3pm wind speed (km/h)"   
## [21] "3pm MSL pressure (hPa)"

#simplify column names:  
mwc <- rename(mwc,   
 date = "Date",  
 min\_t\_C = "Minimum temperature (Deg C)",  
 max\_t\_C = "Maximum Temperature (Deg C)",  
 rainfall\_mm = "Rainfall (mm)",  
 evap\_mm = "Evaporation (mm)",  
 sun\_hrs = "Sunshine (hours)",   
 gust\_dir = "Direction of maximum wind gust",  
 gust\_kmh = "Speed of maximum wind gust (km/h)",  
 gust\_time = "Time of maximum wind gust",  
 temp\_9am\_C = "9am Temperature (Deg C)",  
 humid\_9am = "9am relative humidity (%)",  
 cloud\_9am = "9am cloud amount (oktas)",  
 wind\_dir\_9am = "9am wind direction",  
 wind\_speed\_9am = "9am wind speed (km/h)",  
 pres\_9am = "9am MSL pressure (hPa)",  
 temp\_3pm\_C = "3pm Temperature (Deg C)",  
 humid\_3pm = "3pm relative humidity (%)",  
 cloud\_3pm = "3pm cloud amount (oktas)",  
 wind\_dir\_3pm = "3pm wind direction",  
 wind\_speed\_3pm = "3pm wind speed (km/h)",  
 pres\_3pm = "3pm MSL pressure (hPa)"  
 )  
#create day, month, year columns  
mwc <- mwc %>%  
 mutate(date = as.Date(date),   
 day = day(date), month = month(date), year = year(date))  
#evaporation brought next to date to represent the subject of evaporation in a meaningful way.  
mwc <- mwc[,c(1, 5, 24, 23, 22, 2, 3, 4, 6:21)]  
#make day = day of week  
mwc$day <- weekdays(mwc$date)

#Set value types  
#low level categorical = factor, year could be considered a character as in the scheme of things time is infinite but we only have from 2019 so in this context it'll be a factor  
mwc$year <- factor(mwc$year)  
month\_levels <- c("6", "7", "8", "9", "10", "11", "12", "1", "2", "3", "4", "5")  
mwc$month <- factor(mwc$month, levels = month\_levels) #setting levels from lowest average to highest  
day\_levels <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")  
mwc$day <- factor(mwc$day, levels = day\_levels)  
mwc$gust\_dir <- factor(mwc$gust\_dir)  
mwc$wind\_dir\_9am <- factor(mwc$wind\_dir\_9am)  
mwc$wind\_dir\_3pm <- factor(mwc$wind\_dir\_3pm)  
#All measurements appear to have a set amount of significant figures suggesting all other variables are discrete =integers  
mwc$evap\_mm <- as.integer(mwc$evap\_mm)  
mwc$min\_t\_C <- as.integer(mwc$min\_t\_C)  
mwc$max\_t\_C <- as.integer(mwc$max\_t\_C)  
mwc$rainfall\_mm <- as.integer(mwc$rainfall\_mm)  
mwc$sun\_hrs <- as.integer(mwc$sun\_hrs)  
mwc$gust\_kmh <- as.integer(mwc$gust\_kmh)  
mwc$temp\_9am\_C <- as.integer(mwc$temp\_9am\_C)  
mwc$humid\_9am <- as.integer(mwc$humid\_9am)  
mwc$cloud\_9am <- as.integer(mwc$cloud\_9am)  
mwc$wind\_speed\_9am <- as.integer(mwc$wind\_speed\_9am)

## Warning: NAs introduced by coercion

mwc$pres\_9am <- as.integer(mwc$pres\_9am)  
mwc$temp\_3pm\_C <- as.integer(mwc$temp\_3pm\_C)  
mwc$wind\_speed\_9am <- as.integer(mwc$wind\_speed\_9am)  
mwc$humid\_3pm <- as.integer(mwc$humid\_3pm)  
mwc$cloud\_3pm <- as.integer(mwc$cloud\_3pm)  
mwc$wind\_speed\_3pm <- as.integer(mwc$wind\_speed\_3pm)  
mwc$pres\_3pm <- as.integer(mwc$pres\_3pm)  
inspect\_types(mwc) #Looking good, time variables already sorted for me.

## # A tibble: 4 x 4  
## type cnt pcnt col\_name   
## <chr> <int> <dbl> <named list>  
## 1 integer 16 66.7 <chr [16]>   
## 2 factor 6 25 <chr [6]>   
## 3 Date 1 4.17 <chr [1]>   
## 4 hms difftime 1 4.17 <chr [1]>

#Tibble ready for analysis  
knitr::kable(head(mwc), caption = "The tidy data")

The tidy data

| date | evap\_mm | year | month | day | min\_t\_C | max\_t\_C | rainfall\_mm | sun\_hrs | gust\_dir | gust\_kmh | gust\_time | temp\_9am\_C | humid\_9am | cloud\_9am | wind\_dir\_9am | wind\_speed\_9am | pres\_9am | temp\_3pm\_C | humid\_3pm | cloud\_3pm | wind\_dir\_3pm | wind\_speed\_3pm | pres\_3pm |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2019-01-01 | 7 | 2019 | 1 | Tuesday | 15 | 26 | 0 | 11 | S | 35 | 17:44:00 | 19 | 74 | 7 | S | 6 | 1013 | 24 | 45 | 1 | SSW | 11 | 1011 |
| 2019-01-02 | 7 | 2019 | 1 | Wednesday | 18 | 22 | 0 | 7 | SSW | 39 | 15:23:00 | 19 | 64 | 8 | SSE | 7 | 1013 | 21 | 62 | 1 | SSW | 19 | 1012 |
| 2019-01-03 | 6 | 2019 | 1 | Thursday | 15 | 29 | 0 | 9 | SSW | 26 | 14:53:00 | 18 | 75 | 8 | S | 2 | 1012 | 24 | 60 | 0 | SSW | 13 | 1009 |
| 2019-01-04 | 7 | 2019 | 1 | Friday | 18 | 42 | 0 | 12 | NW | 54 | 12:03:00 | 29 | 31 | 0 | NNE | 9 | 1005 | 42 | 16 | 1 | NW | 15 | 1001 |
| 2019-01-05 | 15 | 2019 | 1 | Saturday | 17 | 21 | 0 | 5 | SSW | 39 | 08:24:00 | 18 | 63 | 7 | S | 13 | 1013 | 19 | 58 | 7 | S | 11 | 1013 |
| 2019-01-06 | 6 | 2019 | 1 | Sunday | 14 | 22 | 1 | 13 | SSW | 33 | 11:12:00 | 17 | 55 | 1 | SW | 9 | 1020 | 20 | 48 | 1 | SSW | 13 | 1019 |

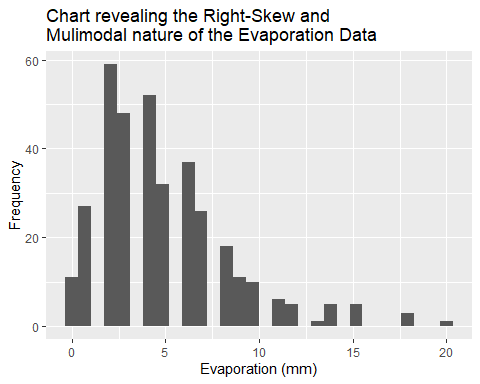
# EDA and Bivariate summaries

## Evaporation

ggplot(mwc, aes(x = evap\_mm)) +   
 geom\_histogram() +  
 labs( title = "Chart revealing the Right-Skew and \nMulimodal nature of the Evaporation Data",  
 x = "Evaporation (mm)",  
 y = "Frequency")

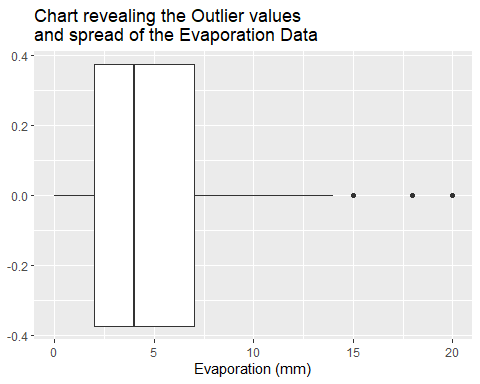
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 8 rows containing non-finite values (stat\_bin).



ggplot(mwc, aes(x = evap\_mm)) +  
 geom\_boxplot() +  
 labs( title = "Chart revealing the Outlier values \nand spread of the Evaporation Data",  
 x = "Evaporation (mm)")

## Warning: Removed 8 rows containing non-finite values (stat\_boxplot).



summary(mwc$evap\_mm)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 2.000 4.000 4.936 7.000 20.000 8

sd(mwc$evap\_mm, na.rm = TRUE)

## [1] 3.504218

filter(mwc, evap\_mm >=15)

## # A tibble: 9 x 24  
## date evap\_mm year month day min\_t\_C max\_t\_C rainfall\_mm sun\_hrs  
## <date> <int> <fct> <fct> <fct> <int> <int> <int> <int>  
## 1 2019-01-05 15 2019 1 Saturday 17 21 0 5  
## 2 2019-01-25 18 2019 1 Friday 21 42 0 11  
## 3 2019-01-26 15 2019 1 Saturday 19 26 0 10  
## 4 2019-02-25 15 2019 2 Monday 18 34 0 11  
## 5 2019-03-01 20 2019 3 Friday 22 38 0 8  
## 6 2019-03-02 18 2019 3 Saturday 22 36 0 11  
## 7 2019-03-03 18 2019 3 Sunday 25 36 0 7  
## 8 2019-04-17 15 2019 4 Wednesday 22 30 0 6  
## 9 2018-12-07 15 2018 12 Friday 25 38 0 8  
## # ... with 15 more variables: gust\_dir <fct>, gust\_kmh <int>, gust\_time <time>,  
## # temp\_9am\_C <int>, humid\_9am <int>, cloud\_9am <int>, wind\_dir\_9am <fct>,  
## # wind\_speed\_9am <int>, pres\_9am <int>, temp\_3pm\_C <int>, humid\_3pm <int>,  
## # cloud\_3pm <int>, wind\_dir\_3pm <fct>, wind\_speed\_3pm <int>, pres\_3pm <int>

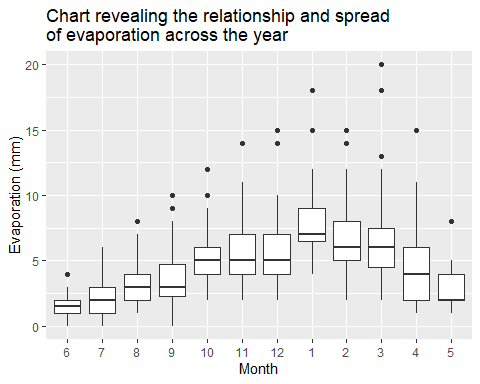
# shape = Right skewed,  
# outliers (9)= 5 at 15mm, 3 at 18mm, 1 at 20mm  
# Spread [0,20] = mean = 4.936mm, SD = 3.504218 mm, IQR = 5mm  
# Location: median = 4mm  
skewness(mwc$max\_t\_C, na.rm =TRUE)

## [1] 0.9333363

## month

ggplot(mwc, aes(x = month, evap\_mm)) +  
 geom\_boxplot()+  
 labs( title = "Chart revealing the relationship and spread \nof evaporation across the year",  
 y = "Evaporation (mm)",  
 x = "Month")

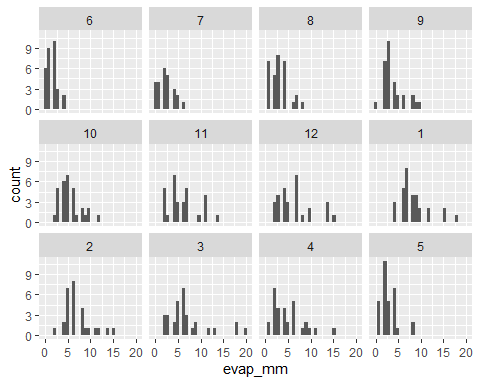
## Warning: Removed 8 rows containing non-finite values (stat\_boxplot).



#evap (Quant) vs month (cat) (side by side box-plot)  
ggplot(mwc, aes(evap\_mm)) + geom\_histogram() + facet\_wrap(~month)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 8 rows containing non-finite values (stat\_bin).

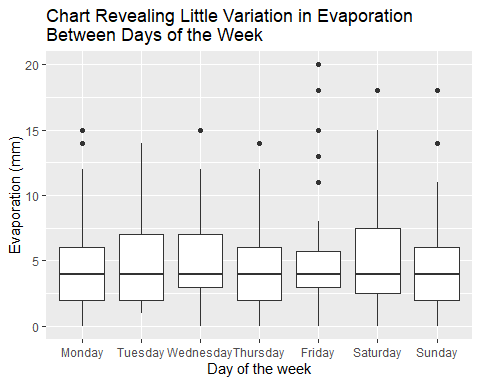


#each month appears to have a fairly right scewed shape many appearing to be multimodal, the month alters the domain,   
#Jan to December, clear seasonal effects. This will have an interaction term with temp variables  
#January(with the summer solstice) marking the maximum and June(with the winter solctice) marking the minimum.  
# We travel an elliptical path around the sun which creates a regular weather pattern that we feel as seasons resulting from the angle of incidence on the atmosphere and consequently the amount of energy we receive from the sun.  
# quite a few outliers throughout the year, the least variation is in June.  
#all domains are within (-1, 21)

## Day of week

# not expecting to find anything informative here  
ggplot(mwc, aes(x = day, evap\_mm)) +  
 geom\_boxplot() +  
 labs( title = "Chart Revealing Little Variation in Evaporation \nBetween Days of the Week",  
 y = "Evaporation (mm)",  
 x = "Day of the week")

## Warning: Removed 8 rows containing non-finite values (stat\_boxplot).



#evap (Quant) vs day (cat) (side by side box-plot)  
unique(mwc$day)

## [1] Tuesday Wednesday Thursday Friday Saturday Sunday Monday   
## Levels: Monday Tuesday Wednesday Thursday Friday Saturday Sunday

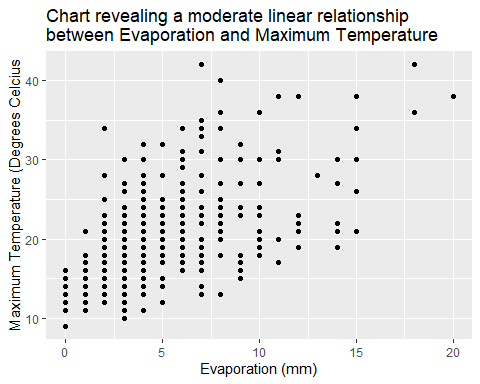
#Not seeing too much going on from day to day. Makes sense, this cycles doesn't influence weather patterns and is an arbitrary measure created for humans not nature.   
day\_signif<- lm(evap\_mm~day, data =mwc)  
anova(day\_signif)

## Analysis of Variance Table  
##   
## Response: evap\_mm  
## Df Sum Sq Mean Sq F value Pr(>F)  
## day 6 63.9 10.654 0.8657 0.5203  
## Residuals 350 4307.6 12.307

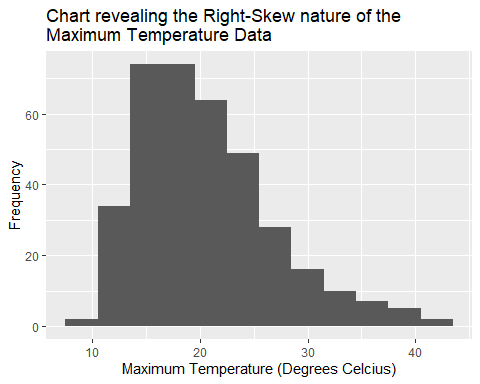
## Maximum Temp in deg C

ggplot(mwc, aes(x = evap\_mm, max\_t\_C)) +   
 geom\_point()+  
 labs( title = "Chart revealing a moderate linear relationship \nbetween Evaporation and Maximum Temperature",  
 x = "Evaporation (mm)",  
 y = "Maximum Temperature (Degrees Celcius")

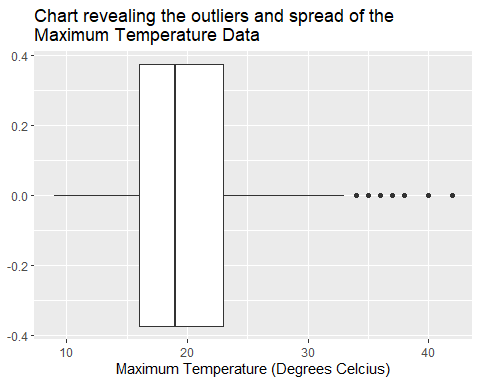
## Warning: Removed 8 rows containing missing values (geom\_point).



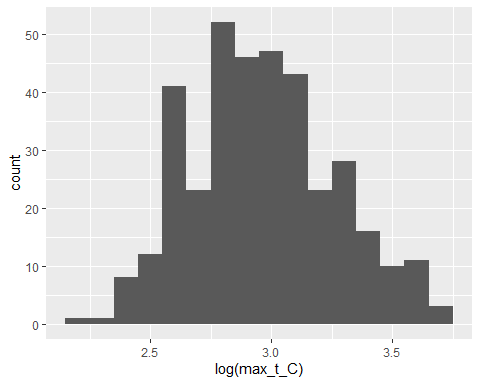
#evap (Quant) vs Max T (quant) (scatterplot)  
#Moderate, positive linear trend. Variation likely due to wind speed. Evaporation depends on the energy of water molecules. higher energy (hotter) molecules will break hydrogen bonds and become gaseous. Wind speed reduces the energy need to leave the surface were the hottest particles are highly likley to be.  
ggplot(mwc, aes(max\_t\_C)) +   
 geom\_histogram(binwidth=3) +  
 labs( title = "Chart revealing the Right-Skew nature of the \nMaximum Temperature Data",  
 x = "Maximum Temperature (Degrees Celcius)",  
 y = "Frequency")



ggplot(mwc, aes(max\_t\_C)) +   
 geom\_boxplot() +  
 labs( title = "Chart revealing the outliers and spread of the \nMaximum Temperature Data",  
 x = "Maximum Temperature (Degrees Celcius)")



#right skewed  
ggplot(mwc, aes(log(max\_t\_C)))+ geom\_histogram(binwidth=0.1)



#normalizes well with log(x)  
summary(mwc$max\_t\_C)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.0 16.0 19.0 20.4 23.0 42.0

sd(mwc$max\_t\_C, na.rm = TRUE)

## [1] 6.321035

filter(mwc, max\_t\_C > 33.5)

## # A tibble: 20 x 24  
## date evap\_mm year month day min\_t\_C max\_t\_C rainfall\_mm sun\_hrs  
## <date> <int> <fct> <fct> <fct> <int> <int> <int> <int>  
## 1 2019-01-04 7 2019 1 Friday 18 42 0 12  
## 2 2019-01-24 8 2019 1 Thursday 17 40 0 12  
## 3 2019-01-25 18 2019 1 Friday 21 42 0 11  
## 4 2019-01-28 6 2019 1 Monday 15 34 0 12  
## 5 2019-01-29 10 2019 1 Tuesday 17 36 0 12  
## 6 2019-01-30 12 2019 1 Wednesday 18 38 0 9  
## 7 2019-02-02 8 2019 2 Saturday 15 34 0 12  
## 8 2019-02-03 11 2019 2 Sunday 19 38 0 11  
## 9 2019-02-06 2 2019 2 Wednesday 18 34 0 8  
## 10 2019-02-25 15 2019 2 Monday 18 34 0 11  
## 11 2019-02-28 8 2019 2 Thursday 16 36 0 10  
## 12 2019-03-01 20 2019 3 Friday 22 38 0 8  
## 13 2019-03-02 18 2019 3 Saturday 22 36 0 11  
## 14 2019-03-03 18 2019 3 Sunday 25 36 0 7  
## 15 2018-11-01 2 2018 11 Thursday 12 34 0 3  
## 16 2018-12-06 7 2018 12 Thursday 14 35 0 12  
## 17 2018-12-07 15 2018 12 Friday 25 38 0 8  
## 18 2018-12-24 7 2018 12 Monday 14 34 0 13  
## 19 2018-12-27 NA 2018 12 Thursday 16 37 0 12  
## 20 2018-12-28 NA 2018 12 Friday 21 36 0 6  
## # ... with 15 more variables: gust\_dir <fct>, gust\_kmh <int>, gust\_time <time>,  
## # temp\_9am\_C <int>, humid\_9am <int>, cloud\_9am <int>, wind\_dir\_9am <fct>,  
## # wind\_speed\_9am <int>, pres\_9am <int>, temp\_3pm\_C <int>, humid\_3pm <int>,  
## # cloud\_3pm <int>, wind\_dir\_3pm <fct>, wind\_speed\_3pm <int>, pres\_3pm <int>

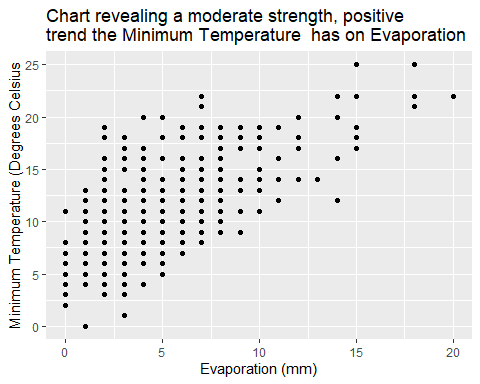
cor(mwc$evap\_mm, mwc$max\_t\_C, use = "complete.obs")

## [1] 0.5841801

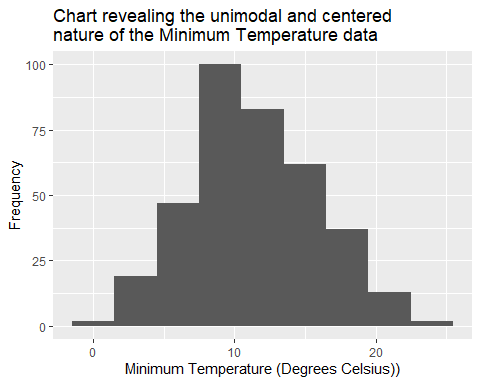
## Minimum Temp in deg C

# As above but less of a slope is what I'm predicting. Colder days might have larger wind speeds though...   
ggplot(mwc, aes(x = evap\_mm, min\_t\_C)) + geom\_point() +  
 labs( title = "Chart revealing a moderate strength, positive \ntrend the Minimum Temperature has on Evaporation ",  
 x = "Evaporation (mm)",  
 y = "Minimum Temperature (Degrees Celsius")

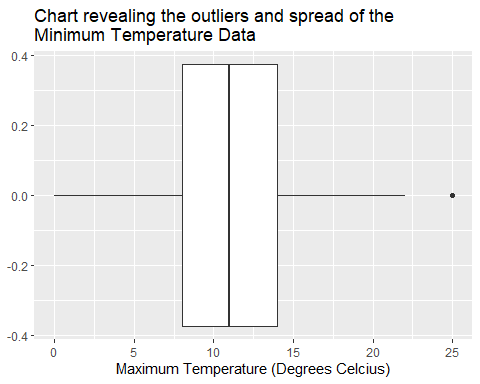
## Warning: Removed 8 rows containing missing values (geom\_point).



#evap (Quant) vs Min T (quant) (scatterplot)  
#don't be fooled by the y-axis, the slop is much lower. Moderate , positive linear trend.  
ggplot(mwc, aes(min\_t\_C)) + geom\_histogram(binwidth=3) +   
 labs( title = "Chart revealing the unimodal and centered \nnature of the Minimum Temperature data",  
 x = "Minimum Temperature (Degrees Celsius))",  
 y = "Frequency")



#fairly normal distribution  
ggplot(mwc, aes(min\_t\_C)) +   
 geom\_boxplot() +  
 labs( title = "Chart revealing the outliers and spread of the \nMinimum Temperature Data",  
 x = "Maximum Temperature (Degrees Celcius)")



summary(mwc$min\_t\_C)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 8.00 11.00 11.38 14.00 25.00

sd(mwc$min\_t\_C, na.rm = TRUE)

## [1] 4.568038

filter(mwc, min\_t\_C > 23)

## # A tibble: 2 x 24  
## date evap\_mm year month day min\_t\_C max\_t\_C rainfall\_mm sun\_hrs  
## <date> <int> <fct> <fct> <fct> <int> <int> <int> <int>  
## 1 2019-03-03 18 2019 3 Sunday 25 36 0 7  
## 2 2018-12-07 15 2018 12 Friday 25 38 0 8  
## # ... with 15 more variables: gust\_dir <fct>, gust\_kmh <int>, gust\_time <time>,  
## # temp\_9am\_C <int>, humid\_9am <int>, cloud\_9am <int>, wind\_dir\_9am <fct>,  
## # wind\_speed\_9am <int>, pres\_9am <int>, temp\_3pm\_C <int>, humid\_3pm <int>,  
## # cloud\_3pm <int>, wind\_dir\_3pm <fct>, wind\_speed\_3pm <int>, pres\_3pm <int>

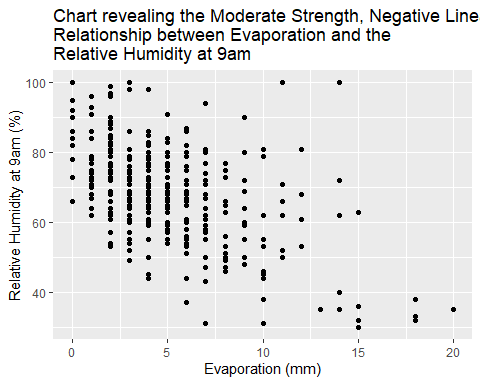
cor(mwc$evap\_mm, mwc$min\_t\_C, use = "complete.obs")

## [1] 0.6493404

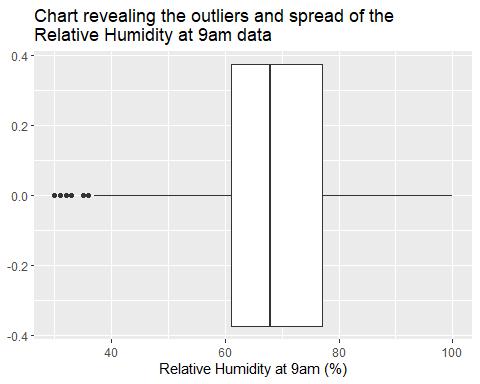
## Relative Humidity at 9am

# The amount of water you can have in the air is limited. Higher humidity makes it more difficult for water to become gaseous as crowding molecules become heavier and sink back into a liquid.  
# correlates with season, and therefore months (interaction term)  
ggplot(mwc, aes(x = evap\_mm, humid\_9am)) + geom\_point() +   
 labs( title = "Chart revealing the Moderate Strength, Negative Linear \nRelationship between Evaporation and the \nRelative Humidity at 9am",  
 x = "Evaporation (mm)",  
 y = "Relative Humidity at 9am (%)")

## Warning: Removed 8 rows containing missing values (geom\_point).

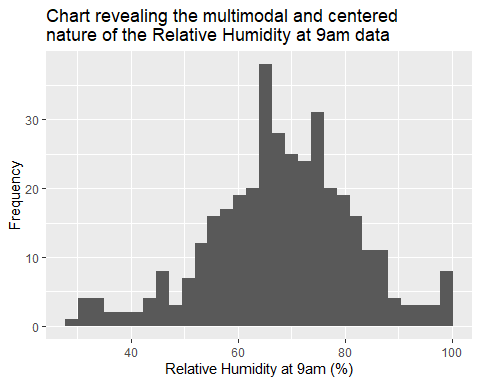


#evap (Quant) vs humid\_9am (quant) (scatter plot)  
# indeed, higher humidity results in less evaporation, moderate negative linear trend with a few outliers   
  
ggplot(mwc, aes(humid\_9am)) +   
 geom\_boxplot() +  
 labs( title = "Chart revealing the outliers and spread of the \nRelative Humidity at 9am data",  
 x = "Relative Humidity at 9am (%)")



ggplot(mwc, aes(humid\_9am)) +   
 geom\_histogram() +  
 labs( title = "Chart revealing the multimodal and centered \nnature of the Relative Humidity at 9am data",  
 x = "Relative Humidity at 9am (%)",  
 y = "Frequency")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



summary(mwc$humid\_9am)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 30.00 61.00 68.00 68.22 77.00 100.00

sd(mwc$humid\_9am, na.rm = TRUE)

## [1] 13.55518

filter(mwc, humid\_9am <45)

## # A tibble: 19 x 24  
## date evap\_mm year month day min\_t\_C max\_t\_C rainfall\_mm sun\_hrs  
## <date> <int> <fct> <fct> <fct> <int> <int> <int> <int>  
## 1 2019-01-04 7 2019 1 Friday 18 42 0 12  
## 2 2019-01-25 18 2019 1 Friday 21 42 0 11  
## 3 2019-02-25 15 2019 2 Monday 18 34 0 11  
## 4 2019-03-01 20 2019 3 Friday 22 38 0 8  
## 5 2019-03-02 18 2019 3 Saturday 22 36 0 11  
## 6 2019-03-03 18 2019 3 Sunday 25 36 0 7  
## 7 2019-03-08 7 2019 3 Friday 11 26 0 1  
## 8 2019-03-29 13 2019 3 Friday 14 28 0 2  
## 9 2019-04-16 10 2019 4 Tuesday 17 30 0 9  
## 10 2019-04-17 15 2019 4 Wednesday 22 30 0 6  
## 11 2018-08-03 6 2018 8 Friday 7 18 0 2  
## 12 2018-08-10 4 2018 8 Friday 9 20 0 9  
## 13 2018-09-18 10 2018 9 Tuesday 11 20 0 6  
## 14 2018-10-06 6 2018 10 Saturday 8 23 0 11  
## 15 2018-11-19 10 2018 11 Monday 15 30 0 11  
## 16 2018-11-20 14 2018 11 Tuesday 22 27 0 0  
## 17 2018-12-07 15 2018 12 Friday 25 38 0 8  
## 18 2018-12-08 14 2018 12 Saturday 22 30 0 1  
## 19 2018-12-28 NA 2018 12 Friday 21 36 0 6  
## # ... with 15 more variables: gust\_dir <fct>, gust\_kmh <int>, gust\_time <time>,  
## # temp\_9am\_C <int>, humid\_9am <int>, cloud\_9am <int>, wind\_dir\_9am <fct>,  
## # wind\_speed\_9am <int>, pres\_9am <int>, temp\_3pm\_C <int>, humid\_3pm <int>,  
## # cloud\_3pm <int>, wind\_dir\_3pm <fct>, wind\_speed\_3pm <int>, pres\_3pm <int>

cor(mwc$evap\_mm, mwc$humid\_9am, use = "complete.obs")

## [1] -0.5302321

skewness(mwc$humid\_9am, na.rm = TRUE)

## [1] -0.2663345

# Model selection

#Evaporation (in mm) on a given day in Melbourne (our evap\_mm is a daily measure from melbourne)  
evap\_model1 <- lm(evap\_mm ~ month + max\_t\_C + min\_t\_C + humid\_9am + day + month:humid\_9am, data = mwc)  
summary(evap\_model1)

##   
## Call:  
## lm(formula = evap\_mm ~ month + max\_t\_C + min\_t\_C + humid\_9am +   
## day + month:humid\_9am, data = mwc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9681 -1.2465 -0.0359 1.0387 11.4716   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.456032 3.526334 0.129 0.89718   
## month7 3.052492 4.398946 0.694 0.48823   
## month8 1.705301 4.155641 0.410 0.68181   
## month9 6.837756 4.086223 1.673 0.09522 .   
## month10 1.077821 4.102677 0.263 0.79294   
## month11 6.160587 3.826672 1.610 0.10839   
## month12 7.887509 3.855434 2.046 0.04158 \*   
## month1 7.263099 3.985851 1.822 0.06934 .   
## month2 8.178222 4.257766 1.921 0.05563 .   
## month3 12.842347 3.738992 3.435 0.00067 \*\*\*  
## month4 9.298658 4.057539 2.292 0.02256 \*   
## month5 3.640927 4.238419 0.859 0.39096   
## max\_t\_C 0.029263 0.030644 0.955 0.34033   
## min\_t\_C 0.337267 0.044107 7.646 2.35e-13 \*\*\*  
## humid\_9am -0.026913 0.041803 -0.644 0.52015   
## dayTuesday 0.195277 0.437011 0.447 0.65528   
## dayWednesday 0.204757 0.436260 0.469 0.63914   
## dayThursday -0.338278 0.439892 -0.769 0.44245   
## dayFriday -0.195668 0.448242 -0.437 0.66275   
## daySaturday 0.838104 0.435609 1.924 0.05523 .   
## daySunday 0.284373 0.433699 0.656 0.51249   
## month7:humid\_9am -0.029736 0.056914 -0.522 0.60170   
## month8:humid\_9am 0.001185 0.054099 0.022 0.98254   
## month9:humid\_9am -0.078830 0.055065 -1.432 0.15323   
## month10:humid\_9am 0.019985 0.054685 0.365 0.71500   
## month11:humid\_9am -0.054833 0.049121 -1.116 0.26512   
## month12:humid\_9am -0.087963 0.048954 -1.797 0.07329 .   
## month1:humid\_9am -0.068005 0.052823 -1.287 0.19886   
## month2:humid\_9am -0.092242 0.057664 -1.600 0.11065   
## month3:humid\_9am -0.153088 0.047474 -3.225 0.00139 \*\*   
## month4:humid\_9am -0.117383 0.053267 -2.204 0.02825 \*   
## month5:humid\_9am -0.044326 0.054068 -0.820 0.41292   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.182 on 325 degrees of freedom  
## (8 observations deleted due to missingness)  
## Multiple R-squared: 0.6461, Adjusted R-squared: 0.6124   
## F-statistic: 19.14 on 31 and 325 DF, p-value: < 2.2e-16

anova(evap\_model1)

## Analysis of Variance Table  
##   
## Response: evap\_mm  
## Df Sum Sq Mean Sq F value Pr(>F)   
## month 11 1485.41 135.04 28.3699 < 2.2e-16 \*\*\*  
## max\_t\_C 1 295.29 295.29 62.0372 5.102e-14 \*\*\*  
## min\_t\_C 1 389.64 389.64 81.8589 < 2.2e-16 \*\*\*  
## humid\_9am 1 453.35 453.35 95.2444 < 2.2e-16 \*\*\*  
## day 6 50.27 8.38 1.7602 0.106676   
## month:humid\_9am 11 150.59 13.69 2.8761 0.001273 \*\*   
## Residuals 325 1546.96 4.76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#remove humid (0.52015)  
  
evap\_model2 <- lm(evap\_mm ~ month + min\_t\_C + day + month:humid\_9am, data = mwc)  
summary(evap\_model2)

##   
## Call:  
## lm(formula = evap\_mm ~ month + min\_t\_C + day + month:humid\_9am,   
## data = mwc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0140 -1.2382 -0.0375 1.0249 11.5233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.96333 3.48562 0.276 0.782437   
## month7 2.81270 4.39118 0.641 0.522275   
## month8 1.64746 4.15464 0.397 0.691969   
## month9 6.93611 4.08437 1.698 0.090423 .   
## month10 1.32077 4.09423 0.323 0.747211   
## month11 6.26999 3.82444 1.639 0.102084   
## month12 8.28679 3.83218 2.162 0.031312 \*   
## month1 7.46172 3.97988 1.875 0.061706 .   
## month2 8.29478 4.25544 1.949 0.052127 .   
## month3 13.07541 3.73051 3.505 0.000521 \*\*\*  
## month4 9.53586 4.04938 2.355 0.019120 \*   
## month5 3.59782 4.23761 0.849 0.396492   
## min\_t\_C 0.35073 0.04179 8.393 1.47e-15 \*\*\*  
## dayTuesday 0.17891 0.43662 0.410 0.682249   
## dayWednesday 0.17425 0.43503 0.401 0.689008   
## dayThursday -0.33043 0.43976 -0.751 0.452964   
## dayFriday -0.16872 0.44729 -0.377 0.706264   
## daySaturday 0.82175 0.43521 1.888 0.059894 .   
## daySunday 0.29014 0.43360 0.669 0.503880   
## month6:humid\_9am -0.02902 0.04174 -0.695 0.487316   
## month7:humid\_9am -0.05584 0.03925 -1.423 0.155805   
## month8:humid\_9am -0.02729 0.03531 -0.773 0.440180   
## month9:humid\_9am -0.10913 0.03654 -2.987 0.003032 \*\*   
## month10:humid\_9am -0.01129 0.03458 -0.326 0.744300   
## month11:humid\_9am -0.08393 0.02593 -3.236 0.001335 \*\*   
## month12:humid\_9am -0.12055 0.02540 -4.745 3.12e-06 \*\*\*  
## month1:humid\_9am -0.09670 0.03261 -2.965 0.003245 \*\*   
## month2:humid\_9am -0.12038 0.03919 -3.072 0.002309 \*\*   
## month3:humid\_9am -0.18334 0.02320 -7.904 4.19e-14 \*\*\*  
## month4:humid\_9am -0.14827 0.03369 -4.401 1.46e-05 \*\*\*  
## month5:humid\_9am -0.07246 0.03450 -2.100 0.036473 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.181 on 326 degrees of freedom  
## (8 observations deleted due to missingness)  
## Multiple R-squared: 0.6451, Adjusted R-squared: 0.6125   
## F-statistic: 19.76 on 30 and 326 DF, p-value: < 2.2e-16

anova(evap\_model2)

## Analysis of Variance Table  
##   
## Response: evap\_mm  
## Df Sum Sq Mean Sq F value Pr(>F)   
## month 11 1485.41 135.04 28.3776 <2e-16 \*\*\*  
## min\_t\_C 1 600.75 600.75 126.2452 <2e-16 \*\*\*  
## day 6 35.89 5.98 1.2572 0.2769   
## month:humid\_9am 12 698.16 58.18 12.2262 <2e-16 \*\*\*  
## Residuals 326 1551.30 4.76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#remove maxT (0.34033)  
  
evap\_model3 <- lm(evap\_mm ~ month + min\_t\_C + day + month:humid\_9am, data = mwc)  
summary(evap\_model3)

##   
## Call:  
## lm(formula = evap\_mm ~ month + min\_t\_C + day + month:humid\_9am,   
## data = mwc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0140 -1.2382 -0.0375 1.0249 11.5233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.96333 3.48562 0.276 0.782437   
## month7 2.81270 4.39118 0.641 0.522275   
## month8 1.64746 4.15464 0.397 0.691969   
## month9 6.93611 4.08437 1.698 0.090423 .   
## month10 1.32077 4.09423 0.323 0.747211   
## month11 6.26999 3.82444 1.639 0.102084   
## month12 8.28679 3.83218 2.162 0.031312 \*   
## month1 7.46172 3.97988 1.875 0.061706 .   
## month2 8.29478 4.25544 1.949 0.052127 .   
## month3 13.07541 3.73051 3.505 0.000521 \*\*\*  
## month4 9.53586 4.04938 2.355 0.019120 \*   
## month5 3.59782 4.23761 0.849 0.396492   
## min\_t\_C 0.35073 0.04179 8.393 1.47e-15 \*\*\*  
## dayTuesday 0.17891 0.43662 0.410 0.682249   
## dayWednesday 0.17425 0.43503 0.401 0.689008   
## dayThursday -0.33043 0.43976 -0.751 0.452964   
## dayFriday -0.16872 0.44729 -0.377 0.706264   
## daySaturday 0.82175 0.43521 1.888 0.059894 .   
## daySunday 0.29014 0.43360 0.669 0.503880   
## month6:humid\_9am -0.02902 0.04174 -0.695 0.487316   
## month7:humid\_9am -0.05584 0.03925 -1.423 0.155805   
## month8:humid\_9am -0.02729 0.03531 -0.773 0.440180   
## month9:humid\_9am -0.10913 0.03654 -2.987 0.003032 \*\*   
## month10:humid\_9am -0.01129 0.03458 -0.326 0.744300   
## month11:humid\_9am -0.08393 0.02593 -3.236 0.001335 \*\*   
## month12:humid\_9am -0.12055 0.02540 -4.745 3.12e-06 \*\*\*  
## month1:humid\_9am -0.09670 0.03261 -2.965 0.003245 \*\*   
## month2:humid\_9am -0.12038 0.03919 -3.072 0.002309 \*\*   
## month3:humid\_9am -0.18334 0.02320 -7.904 4.19e-14 \*\*\*  
## month4:humid\_9am -0.14827 0.03369 -4.401 1.46e-05 \*\*\*  
## month5:humid\_9am -0.07246 0.03450 -2.100 0.036473 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.181 on 326 degrees of freedom  
## (8 observations deleted due to missingness)  
## Multiple R-squared: 0.6451, Adjusted R-squared: 0.6125   
## F-statistic: 19.76 on 30 and 326 DF, p-value: < 2.2e-16

anova(evap\_model3)

## Analysis of Variance Table  
##   
## Response: evap\_mm  
## Df Sum Sq Mean Sq F value Pr(>F)   
## month 11 1485.41 135.04 28.3776 <2e-16 \*\*\*  
## min\_t\_C 1 600.75 600.75 126.2452 <2e-16 \*\*\*  
## day 6 35.89 5.98 1.2572 0.2769   
## month:humid\_9am 12 698.16 58.18 12.2262 <2e-16 \*\*\*  
## Residuals 326 1551.30 4.76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#remove day (0.2769)  
  
evap\_model4 <- lm(evap\_mm ~ month + min\_t\_C + month:humid\_9am, data = mwc)  
summary(evap\_model4)

##   
## Call:  
## lm(formula = evap\_mm ~ month + min\_t\_C + month:humid\_9am, data = mwc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4907 -1.2440 -0.0317 1.1340 11.0554   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.63358 3.42334 0.185 0.853282   
## month7 3.05364 4.37854 0.697 0.486034   
## month8 1.34502 4.14767 0.324 0.745928   
## month9 6.63708 4.07134 1.630 0.104008   
## month10 1.90611 4.05197 0.470 0.638366   
## month11 6.79424 3.80625 1.785 0.075171 .   
## month12 8.85762 3.81583 2.321 0.020877 \*   
## month1 7.84832 3.95089 1.986 0.047803 \*   
## month2 8.43501 4.23047 1.994 0.046984 \*   
## month3 13.37318 3.71764 3.597 0.000371 \*\*\*  
## month4 10.19378 4.03536 2.526 0.011998 \*   
## month5 4.05103 4.20535 0.963 0.336096   
## min\_t\_C 0.35538 0.04149 8.565 4.10e-16 \*\*\*  
## month6:humid\_9am -0.02334 0.04150 -0.562 0.574235   
## month7:humid\_9am -0.05317 0.03920 -1.356 0.175889   
## month8:humid\_9am -0.01690 0.03499 -0.483 0.629368   
## month9:humid\_9am -0.09702 0.03634 -2.670 0.007967 \*\*   
## month10:humid\_9am -0.01393 0.03446 -0.404 0.686360   
## month11:humid\_9am -0.08598 0.02597 -3.311 0.001031 \*\*   
## month12:humid\_9am -0.12233 0.02544 -4.808 2.32e-06 \*\*\*  
## month1:humid\_9am -0.09687 0.03194 -3.033 0.002612 \*\*   
## month2:humid\_9am -0.11643 0.03923 -2.968 0.003214 \*\*   
## month3:humid\_9am -0.18159 0.02314 -7.846 5.93e-14 \*\*\*  
## month4:humid\_9am -0.15201 0.03366 -4.516 8.76e-06 \*\*\*  
## month5:humid\_9am -0.07323 0.03436 -2.132 0.033779 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.19 on 332 degrees of freedom  
## (8 observations deleted due to missingness)  
## Multiple R-squared: 0.6359, Adjusted R-squared: 0.6096   
## F-statistic: 24.16 on 24 and 332 DF, p-value: < 2.2e-16

anova(evap\_model4)

## Analysis of Variance Table  
##   
## Response: evap\_mm  
## Df Sum Sq Mean Sq F value Pr(>F)   
## month 11 1485.41 135.04 28.167 < 2.2e-16 \*\*\*  
## min\_t\_C 1 600.75 600.75 125.307 < 2.2e-16 \*\*\*  
## month:humid\_9am 12 693.67 57.81 12.057 < 2.2e-16 \*\*\*  
## Residuals 332 1591.68 4.79   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#All checks out.  
  
#day is an arbitrary unnatural variable, month is useful only for determining climate, but temperatures and humidities actually are the climate, and so it makes sense that only those two really matter.  
#humidity in relation to time of year might be more important  
#Max-t is related to humidity, the minimum t might also be more meaningful when it comes to the amount of evap as measured in mm.  
#Check it:  
  
step(evap\_model1, direction = "backward")

## Start: AIC=587.47  
## evap\_mm ~ month + max\_t\_C + min\_t\_C + humid\_9am + day + month:humid\_9am  
##   
## Df Sum of Sq RSS AIC  
## - day 6 42.454 1589.4 585.14  
## - max\_t\_C 1 4.340 1551.3 586.47  
## <none> 1547.0 587.47  
## - month:humid\_9am 11 150.589 1697.5 598.64  
## - min\_t\_C 1 278.304 1825.3 644.53  
##   
## Step: AIC=585.14  
## evap\_mm ~ month + max\_t\_C + min\_t\_C + humid\_9am + month:humid\_9am  
##   
## Df Sum of Sq RSS AIC  
## - max\_t\_C 1 2.267 1591.7 583.65  
## <none> 1589.4 585.14  
## - month:humid\_9am 11 158.405 1747.8 597.06  
## - min\_t\_C 1 301.110 1890.5 645.07  
##   
## Step: AIC=583.65  
## evap\_mm ~ month + min\_t\_C + humid\_9am + month:humid\_9am  
##   
## Df Sum of Sq RSS AIC  
## <none> 1591.7 583.65  
## - month:humid\_9am 11 161.51 1753.2 596.15  
## - min\_t\_C 1 351.74 1943.4 652.93

##   
## Call:  
## lm(formula = evap\_mm ~ month + min\_t\_C + humid\_9am + month:humid\_9am,   
## data = mwc)  
##   
## Coefficients:  
## (Intercept) month7 month8 month9   
## 0.633577 3.053641 1.345022 6.637083   
## month10 month11 month12 month1   
## 1.906112 6.794242 8.857621 7.848325   
## month2 month3 month4 month5   
## 8.435006 13.373182 10.193781 4.051027   
## min\_t\_C humid\_9am month7:humid\_9am month8:humid\_9am   
## 0.355377 -0.023336 -0.029834 0.006433   
## month9:humid\_9am month10:humid\_9am month11:humid\_9am month12:humid\_9am   
## -0.073687 0.009411 -0.062646 -0.098989   
## month1:humid\_9am month2:humid\_9am month3:humid\_9am month4:humid\_9am   
## -0.073534 -0.093094 -0.158251 -0.128669   
## month5:humid\_9am   
## -0.049897

ols\_step\_backward\_p(evap\_model1, prem= 0.05)

##   
##   
## Elimination Summary   
## ---------------------------------------------------------------------------  
## Variable Adj.   
## Step Removed R-Square R-Square C(p) AIC RMSE   
## ---------------------------------------------------------------------------  
## 1 max\_t\_C 0.6451 0.6125 -19.0881 1601.5963 2.1814   
## 2 day 0.6359 0.6096 -12.6046 1598.7702 2.1896   
## ---------------------------------------------------------------------------

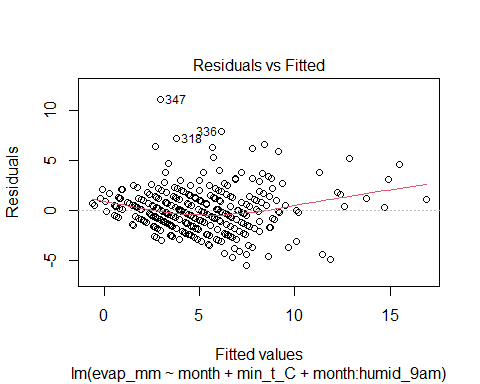
ols\_check <- lm(evap\_mm ~ month + min\_t\_C + humid\_9am + month:humid\_9am, data = mwc)  
step\_check <- lm(evap\_mm ~ month + min\_t\_C + humid\_9am + month:humid\_9am, data = mwc)  
#the above two functions found the same answer. so is ours statistically differnt?  
evap\_model4 <- lm(evap\_mm ~ month + min\_t\_C + month:humid\_9am, data = mwc)  
anova(ols\_check, evap\_model4)

## Analysis of Variance Table  
##   
## Model 1: evap\_mm ~ month + min\_t\_C + humid\_9am + month:humid\_9am  
## Model 2: evap\_mm ~ month + min\_t\_C + month:humid\_9am  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 332 1591.7   
## 2 332 1591.7 0 0

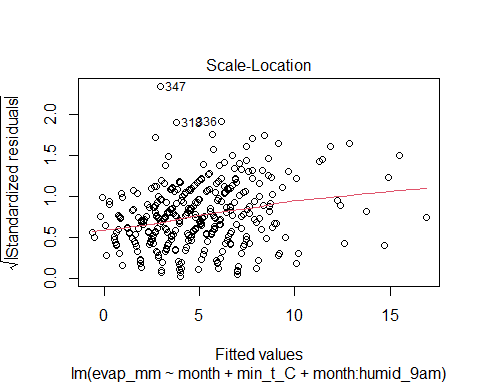
#ANOVA finds no statistically significant difference between these models.

# Model Diagnostics

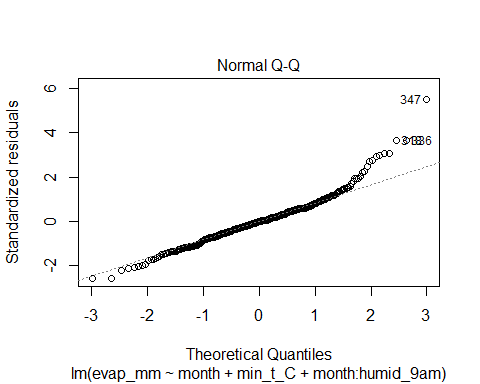
plot(evap\_model4, which = 1) #linearity



plot(evap\_model4, which = 3) #homoscedasticity



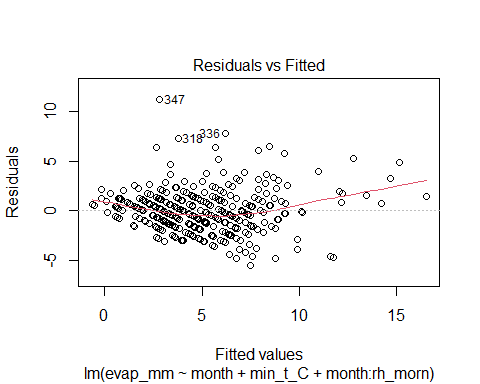
plot(evap\_model4, which = 2) #noise is normally distributed



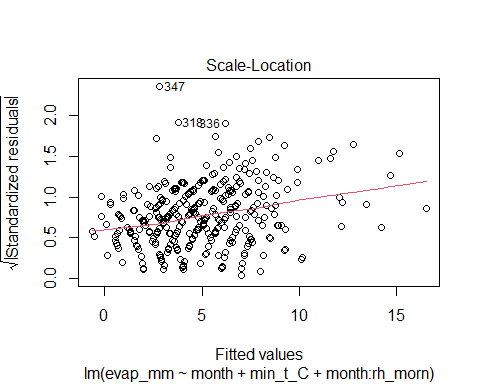
#Assumptions about linearity is fairly good for the bulk of the data,   
#homoscedasticity shows a slight linear trend but nothing exponential.  
#Noise is fairly normally distributed.  
vars<- mwc[,c( "evap\_mm", "month", "day", "min\_t\_C", "max\_t\_C", "humid\_9am")]  
vars <- vars %>%  
 mutate(maxtemp = (max\_t\_C)^-0.3, rh\_morn = (humid\_9am)^1.4)  
vars <- tibble(vars)  
vars

## # A tibble: 365 x 8  
## evap\_mm month day min\_t\_C max\_t\_C humid\_9am maxtemp rh\_morn  
## <int> <fct> <fct> <int> <int> <int> <dbl> <dbl>  
## 1 7 1 Tuesday 15 26 74 0.376 414.  
## 2 7 1 Wednesday 18 22 64 0.396 338.  
## 3 6 1 Thursday 15 29 75 0.364 422.  
## 4 7 1 Friday 18 42 31 0.326 122.  
## 5 15 1 Saturday 17 21 63 0.401 330.  
## 6 6 1 Sunday 14 22 55 0.396 273.  
## 7 9 1 Monday 17 23 55 0.390 273.  
## 8 7 1 Tuesday 16 24 72 0.385 398.  
## 9 7 1 Wednesday 16 20 62 0.407 323.  
## 10 8 1 Thursday 13 21 53 0.401 259.  
## # ... with 355 more rows

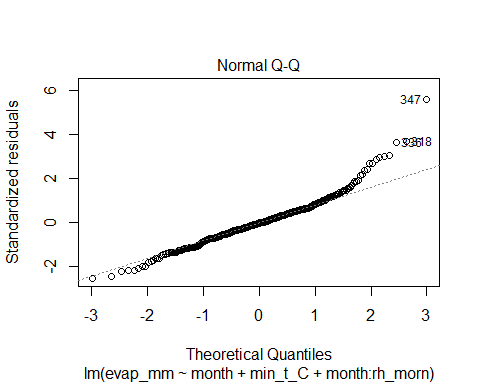
chech <- lm(evap\_mm~month+min\_t\_C+month:rh\_morn, data = vars)  
plot(chech, which = 1)



plot(chech, which = 3)



plot(chech, which = 2)



# Prediction

#February 29, 2020, if this day has a minimum temperature of 13.8 degrees and reaches a   
#maximum of 23.2 degrees, and has 74% humidity at 9am.   
predict(evap\_model4, newdata = tibble(month = "2", min\_t\_C = 13.8, humid\_9am = 74), interval = "prediction")

## fit lwr upr  
## 1 5.356955 0.912332 9.801579

#5.356955mm  
#9.801579mm at most

December 25, 2020, if this day has a minimum temperature of 16.4 degrees and reaches a maximum of 31.9 degrees, and has 57% humidity at 9am.

predict(evap\_model4, newdata = tibble(month = "12", min\_t\_C = 16.4, humid\_9am = 57), interval = "prediction")

## fit lwr upr  
## 1 8.346807 3.920063 12.77355

8.346807 mm 12.77355 mm at most

January 13, 2020, if this day has a minimum temperature of 26.5 degrees and reaches a maximum of 44.3 degrees, and has 35% humidity at 9am

predict(evap\_model4, newdata = tibble(month = "1", min\_t\_C = 26.5, humid\_9am = 35), interval = "prediction")

## fit lwr upr  
## 1 14.50894 9.708499 19.30938

14.50894mm 19.30938mm at most

July 6, 2020, if this day has a minimum temperature of 6.8 degrees and reaches a maximum of 10.6 degrees, and has 76% humidity at 9am.

predict(evap\_model4, newdata = tibble(month = "7", min\_t\_C = 6.8, humid\_9am = 76), interval = "prediction")

## fit lwr upr  
## 1 2.062808 -2.343371 6.468986

2.062808 mm 6.468986 mm at most

Confidence interval = 95% chance of finding the mean evap\_mm measurement prediction interval = 95% chance of finding a specific evap\_mm measurement We want particular days values so we use prediction intervals

citation("tidyverse")

##   
## Wickham et al., (2019). Welcome to the tidyverse. Journal of Open  
## Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {Welcome to the {tidyverse}},  
## author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},  
## year = {2019},  
## journal = {Journal of Open Source Software},  
## volume = {4},  
## number = {43},  
## pages = {1686},  
## doi = {10.21105/joss.01686},  
## }

citation("dplyr")

##   
## To cite package 'dplyr' in publications use:  
##   
## Hadley Wickham, Romain François, Lionel Henry and Kirill Müller  
## (2021). dplyr: A Grammar of Data Manipulation. R package version  
## 1.0.7. https://CRAN.R-project.org/package=dplyr  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {dplyr: A Grammar of Data Manipulation},  
## author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller},  
## year = {2021},  
## note = {R package version 1.0.7},  
## url = {https://CRAN.R-project.org/package=dplyr},  
## }

citation("lubridate")

##   
## To cite lubridate in publications use:  
##   
## Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy  
## with lubridate. Journal of Statistical Software, 40(3), 1-25. URL  
## https://www.jstatsoft.org/v40/i03/.  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {Dates and Times Made Easy with {lubridate}},  
## author = {Garrett Grolemund and Hadley Wickham},  
## journal = {Journal of Statistical Software},  
## year = {2011},  
## volume = {40},  
## number = {3},  
## pages = {1--25},  
## url = {https://www.jstatsoft.org/v40/i03/},  
## }

citation("stringr")

##   
## To cite package 'stringr' in publications use:  
##   
## Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for  
## Common String Operations. R package version 1.4.0.  
## https://CRAN.R-project.org/package=stringr  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {stringr: Simple, Consistent Wrappers for Common String Operations},  
## author = {Hadley Wickham},  
## year = {2019},  
## note = {R package version 1.4.0},  
## url = {https://CRAN.R-project.org/package=stringr},  
## }

citation("inspectdf")

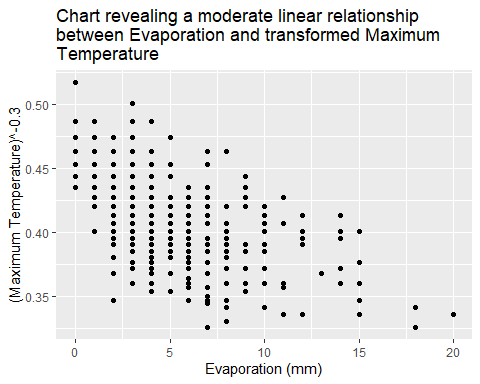
##   
## To cite package 'inspectdf' in publications use:  
##   
## Alastair Rushworth (2021). inspectdf: Inspection, Comparison and  
## Visualisation of Data Frames. R package version 0.0.11.  
## https://CRAN.R-project.org/package=inspectdf  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {inspectdf: Inspection, Comparison and Visualisation of Data Frames},  
## author = {Alastair Rushworth},  
## year = {2021},  
## note = {R package version 0.0.11},  
## url = {https://CRAN.R-project.org/package=inspectdf},  
## }

citation("olsrr")

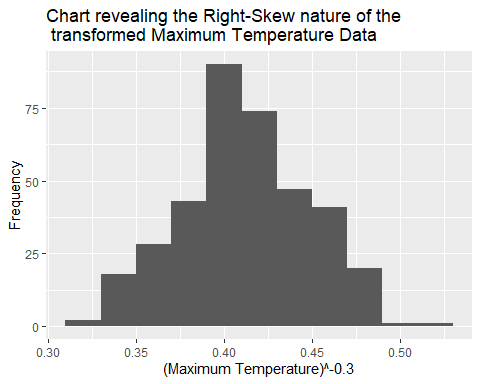
##   
## To cite package 'olsrr' in publications use:  
##   
## Aravind Hebbali (2020). olsrr: Tools for Building OLS Regression  
## Models. R package version 0.5.3.  
## https://CRAN.R-project.org/package=olsrr  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {olsrr: Tools for Building OLS Regression Models},  
## author = {Aravind Hebbali},  
## year = {2020},  
## note = {R package version 0.5.3},  
## url = {https://CRAN.R-project.org/package=olsrr},  
## }

vars<- mwc[,c( "evap\_mm", "month", "day", "min\_t\_C", "max\_t\_C", "humid\_9am")]  
vars <- vars %>%  
 mutate(maxtemp = (max\_t\_C)^-0.3, rh\_morn = (humid\_9am)^1.4)  
vars <- tibble(vars)  
  
ggplot(vars, aes(x = evap\_mm, maxtemp)) +   
 geom\_point()+  
 labs( title = "Chart revealing a moderate linear relationship \nbetween Evaporation and transformed Maximum \nTemperature",  
 x = "Evaporation (mm)",  
 y = "(Maximum Temperature)^-0.3")

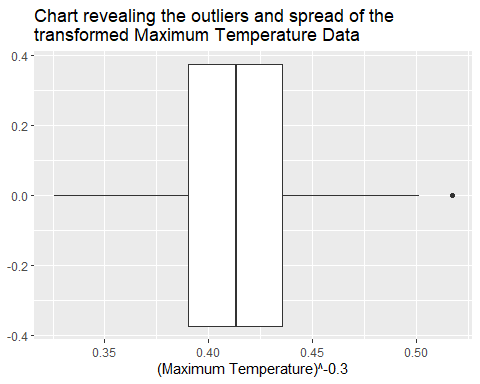
## Warning: Removed 8 rows containing missing values (geom\_point).



ggplot(vars, aes(maxtemp)) +   
 geom\_histogram(binwidth=0.02) +  
 labs( title = "Chart revealing the Right-Skew nature of the \n transformed Maximum Temperature Data",  
 x = "(Maximum Temperature)^-0.3",  
 y = "Frequency")

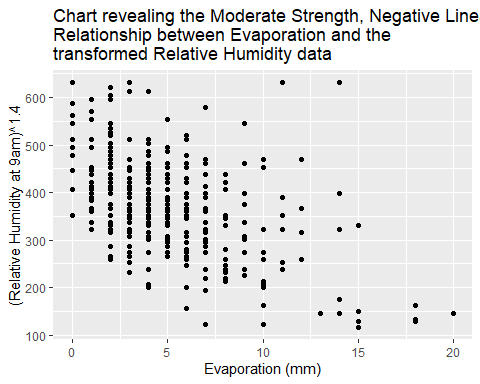


ggplot(vars, aes(maxtemp)) +   
 geom\_boxplot() +  
 labs( title = "Chart revealing the outliers and spread of the \ntransformed Maximum Temperature Data",  
 x = "(Maximum Temperature)^-0.3")

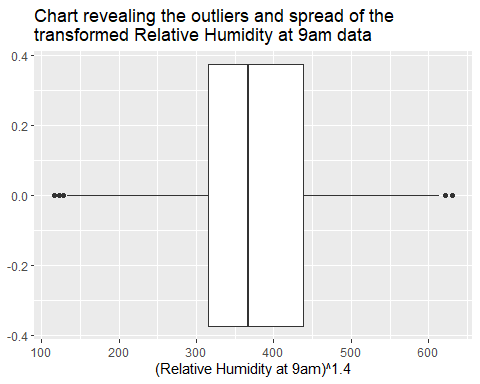


ggplot(vars, aes(x = evap\_mm, rh\_morn)) + geom\_point() +   
 labs( title = "Chart revealing the Moderate Strength, Negative Linear \nRelationship between Evaporation and the \ntransformed Relative Humidity data",  
 x = "Evaporation (mm)",  
 y = "(Relative Humidity at 9am)^1.4")

## Warning: Removed 8 rows containing missing values (geom\_point).

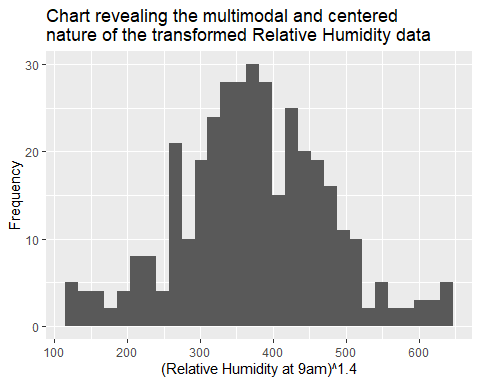


#evap (Quant) vs humid\_9am (quant) (scatter plot)  
# indeed, higher humidity results in less evaporation, moderate negative linear trend with a few outliers   
  
ggplot(vars, aes(rh\_morn)) +   
 geom\_boxplot() +  
 labs( title = "Chart revealing the outliers and spread of the \ntransformed Relative Humidity at 9am data",  
 x = "(Relative Humidity at 9am)^1.4")



ggplot(vars, aes(rh\_morn)) +   
 geom\_histogram() +  
 labs( title = "Chart revealing the multimodal and centered \nnature of the transformed Relative Humidity data",  
 x = "(Relative Humidity at 9am)^1.4",  
 y = "Frequency")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



citation("e1071")

##   
## To cite package 'e1071' in publications use:  
##   
## David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and  
## Friedrich Leisch (2021). e1071: Misc Functions of the Department of  
## Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R  
## package version 1.7-9. https://CRAN.R-project.org/package=e1071  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {e1071: Misc Functions of the Department of Statistics, Probability  
## Theory Group (Formerly: E1071), TU Wien},  
## author = {David Meyer and Evgenia Dimitriadou and Kurt Hornik and Andreas Weingessel and Friedrich Leisch},  
## year = {2021},  
## note = {R package version 1.7-9},  
## url = {https://CRAN.R-project.org/package=e1071},  
## }