ProjectPhase1\_vilain

Kelly Vilain

June 19, 2019

Libraries

options(tidyverse.quiet=TRUE)  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.3

library(caret)

## Warning: package 'caret' was built under R version 3.5.3

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(cluster) #algorithms for clustering

## Warning: package 'cluster' was built under R version 3.5.3

library(factoextra) #visualization

## Warning: package 'factoextra' was built under R version 3.5.3

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(dendextend) #viewing clustering dendograms

## Warning: package 'dendextend' was built under R version 3.5.3

##   
## ---------------------  
## Welcome to dendextend version 1.12.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------

##   
## Attaching package: 'dendextend'

## The following object is masked from 'package:stats':  
##   
## cutree

library(GGally)

## Warning: package 'GGally' was built under R version 3.5.3

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(mice)

## Warning: package 'mice' was built under R version 3.5.3

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM)

## Warning: package 'VIM' was built under R version 3.5.3

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.3

##   
## Attaching package: 'data.table'

## The following object is masked from 'package:dendextend':  
##   
## set

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

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

rain = read\_csv("rain.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## Date = col\_character(),  
## WindGustDir = col\_character(),  
## WindDir9am = col\_character(),  
## WindDir3pm = col\_character(),  
## RainToday = col\_character(),  
## RainTomorrow = col\_character()  
## )

## See spec(...) for full column specifications.

str(rain)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 20 variables:  
## $ Date : chr "12/5/2008" "12/6/2008" "12/16/2008" "12/17/2008" ...  
## $ MinTemp : num 17.5 14.6 9.8 14.1 20.5 20.1 9.6 14 12.5 17.4 ...  
## $ MaxTemp : num 32.3 29.7 27.7 20.9 31.8 32.7 23.9 28.3 28.4 43 ...  
## $ Rainfall : num 1 0.2 NA 0 0 0 0 0 0 0 ...  
## $ WindGustDir : chr "W" "WNW" "WNW" "ENE" ...  
## $ WindGustSpeed: num 41 56 50 22 41 48 41 48 37 39 ...  
## $ WindDir9am : chr "ENE" "W" NA "SSW" ...  
## $ WindDir3pm : chr "NW" "W" "WNW" "E" ...  
## $ WindSpeed9am : num 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm : num 20 24 22 9 20 30 11 24 9 17 ...  
## $ Humidity9am : num 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : num 33 23 28 82 24 15 22 15 16 8 ...  
## $ Pressure9am : num 1011 1009 1013 1012 1008 ...  
## $ Pressure3pm : num 1006 1005 1010 1010 1006 ...  
## $ Cloud9am : num 7 NA 0 8 NA NA NA NA NA NA ...  
## $ Cloud3pm : num 8 NA NA 1 NA NA NA NA NA NA ...  
## $ Temp9am : num 17.8 20.6 17.3 17.2 23.8 24.6 14.9 17.9 17.2 25.6 ...  
## $ Temp3pm : num 29.7 28.9 26.2 18.1 30.8 32.1 22.1 27.6 26.6 41.5 ...  
## $ RainToday : chr "No" "No" NA "No" ...  
## $ RainTomorrow : chr "No" "No" "No" "Yes" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Date = col\_character(),  
## .. MinTemp = col\_double(),  
## .. MaxTemp = col\_double(),  
## .. Rainfall = col\_double(),  
## .. WindGustDir = col\_character(),  
## .. WindGustSpeed = col\_double(),  
## .. WindDir9am = col\_character(),  
## .. WindDir3pm = col\_character(),  
## .. WindSpeed9am = col\_double(),  
## .. WindSpeed3pm = col\_double(),  
## .. Humidity9am = col\_double(),  
## .. Humidity3pm = col\_double(),  
## .. Pressure9am = col\_double(),  
## .. Pressure3pm = col\_double(),  
## .. Cloud9am = col\_double(),  
## .. Cloud3pm = col\_double(),  
## .. Temp9am = col\_double(),  
## .. Temp3pm = col\_double(),  
## .. RainToday = col\_character(),  
## .. RainTomorrow = col\_character()  
## .. )

#summary(rain)

Factor conversion and recoding

rain = rain %>% mutate(RainToday = as.factor(RainToday)) %>%   
 mutate(RainToday = fct\_recode(RainToday, "No" = "0", "Yes" = "1" )) %>%  
 mutate(RainTomorrow = fct\_recode(RainTomorrow, "No" = "0", "Yes" = "1" )) %>%  
 mutate(WindGustDir = as.factor(WindGustDir)) %>%   
 mutate(WindDir9am = as.factor(WindDir9am)) %>%   
 mutate(WindDir3pm = as.factor(WindDir3pm))

## Warning: Unknown levels in `f`: 0, 1  
  
## Warning: Unknown levels in `f`: 0, 1

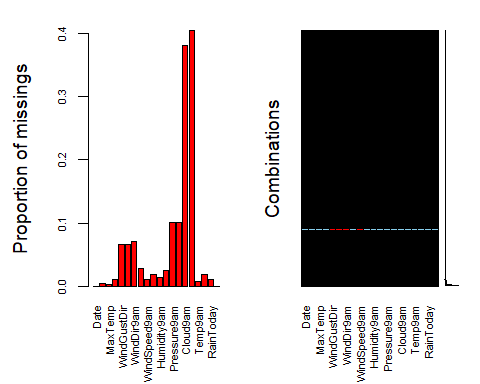
str(rain)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 20 variables:  
## $ Date : chr "12/5/2008" "12/6/2008" "12/16/2008" "12/17/2008" ...  
## $ MinTemp : num 17.5 14.6 9.8 14.1 20.5 20.1 9.6 14 12.5 17.4 ...  
## $ MaxTemp : num 32.3 29.7 27.7 20.9 31.8 32.7 23.9 28.3 28.4 43 ...  
## $ Rainfall : num 1 0.2 NA 0 0 0 0 0 0 0 ...  
## $ WindGustDir : Factor w/ 16 levels "E","ENE","ESE",..: 14 15 15 2 15 15 14 14 5 8 ...  
## $ WindGustSpeed: num 41 56 50 22 41 48 41 48 37 39 ...  
## $ WindDir9am : Factor w/ 16 levels "E","ENE","ESE",..: 2 14 NA 12 14 4 16 14 11 11 ...  
## $ WindDir3pm : Factor w/ 16 levels "E","ENE","ESE",..: 8 14 15 1 14 15 12 16 9 12 ...  
## $ WindSpeed9am : num 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm : num 20 24 22 9 20 30 11 24 9 17 ...  
## $ Humidity9am : num 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : num 33 23 28 82 24 15 22 15 16 8 ...  
## $ Pressure9am : num 1011 1009 1013 1012 1008 ...  
## $ Pressure3pm : num 1006 1005 1010 1010 1006 ...  
## $ Cloud9am : num 7 NA 0 8 NA NA NA NA NA NA ...  
## $ Cloud3pm : num 8 NA NA 1 NA NA NA NA NA NA ...  
## $ Temp9am : num 17.8 20.6 17.3 17.2 23.8 24.6 14.9 17.9 17.2 25.6 ...  
## $ Temp3pm : num 29.7 28.9 26.2 18.1 30.8 32.1 22.1 27.6 26.6 41.5 ...  
## $ RainToday : Factor w/ 2 levels "No","Yes": 1 1 NA 1 1 1 1 1 1 1 ...  
## $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...

View missingness

vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



#the cex.axis reduces size of text on x-axis so labels fit better

## I don’t know why the above output looks different in the knitted file than it does in my RStudio project:

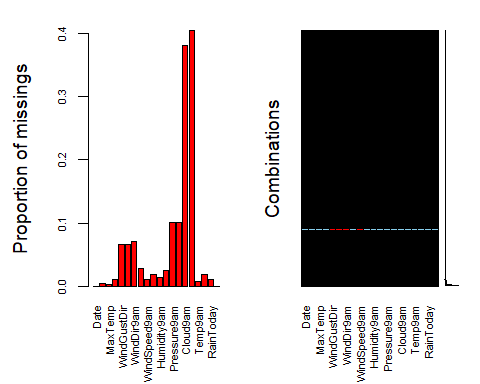
## 

## Create a dataset called rain1 and perform row-wise deletion of NA’s since our dataset is so large.

Column-wise deletion of the Date, Cloud9am and Cloud3pm variables.

rain1 = rain %>% select(-Date, -Cloud9am, -Cloud3pm)   
vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

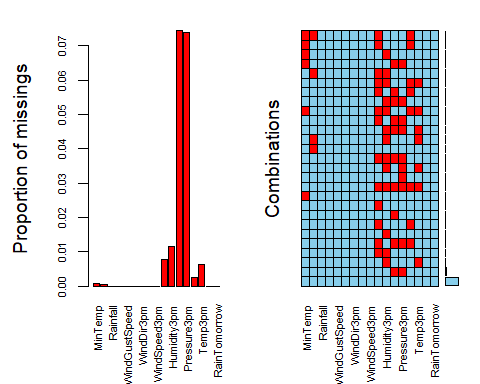
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



Row-wise deletion of any categorical rows with NAs is:

rain1 = rain1 %>% drop\_na(WindGustDir, WindDir9am, WindDir3pm, RainToday)   
vim\_plot = aggr(rain1, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

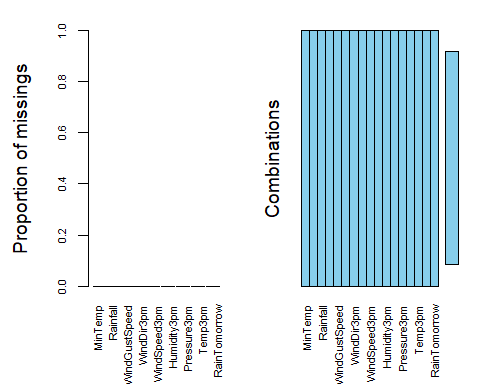
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



Row-wise deletion of the rest of the rows with NA’s.

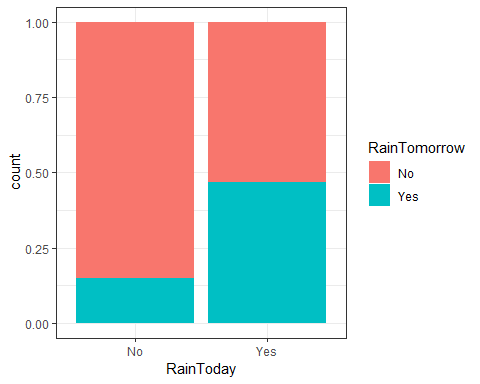
rain1 = rain1 %>% drop\_na()   
vim\_plot = aggr(rain1, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

## Warning in plot.aggr(res, ...): not enough horizontal space to display  
## frequencies



# Visualizations of rain1 data

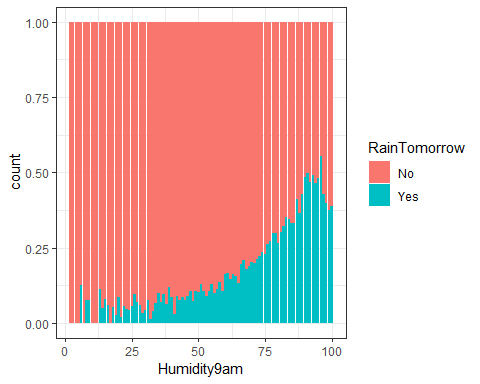
ggplot(rain1, aes(x = RainToday, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain1\_RainToday.png")

## Saving 5 x 4 in image

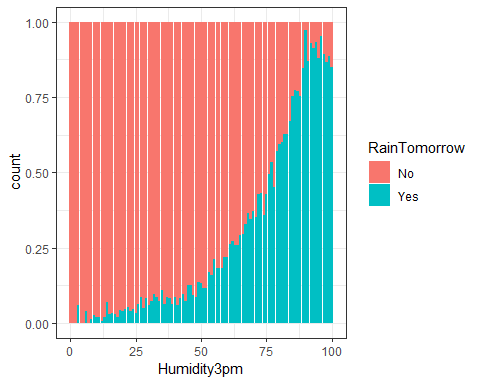
ggplot(rain1, aes(x = Humidity9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain1\_Humidity9.png")

## Saving 5 x 4 in image

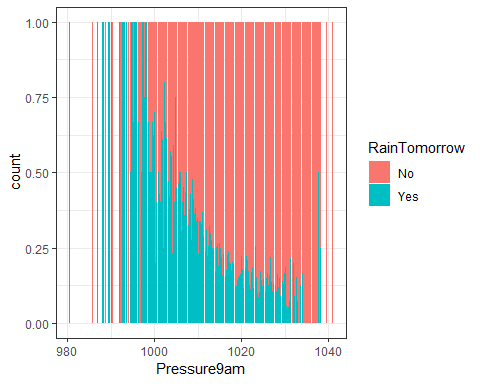
ggplot(rain1, aes(x = Humidity3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain1\_Humidity3.png")

## Saving 5 x 4 in image

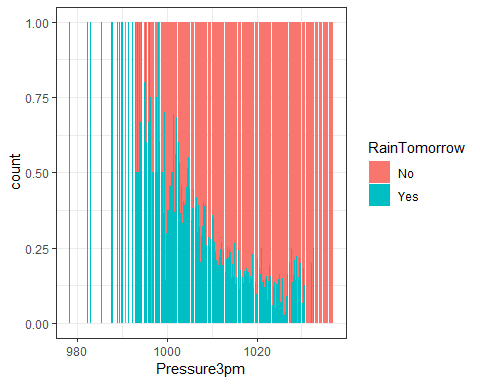
ggplot(rain1, aes(x = Pressure9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain1\_Pressure9.png")

## Saving 5 x 4 in image

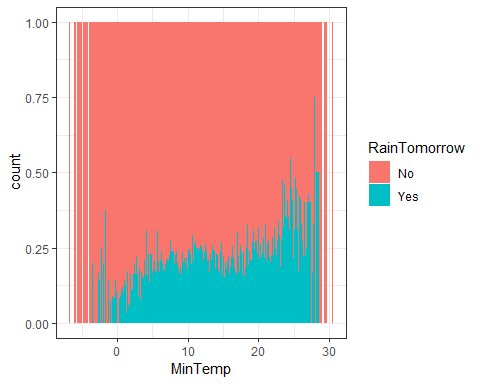
ggplot(rain1, aes(x = Pressure3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



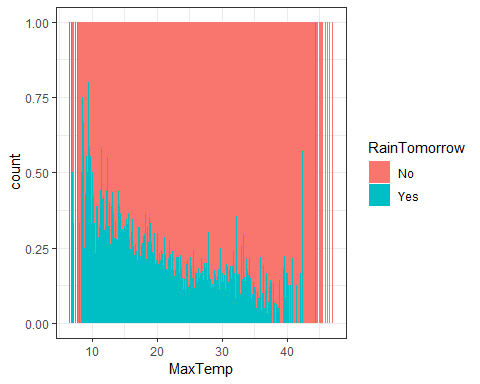
ggsave("rain1\_Pressure3.png")

## Saving 5 x 4 in image

ggplot(rain1, aes(x = MinTemp, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

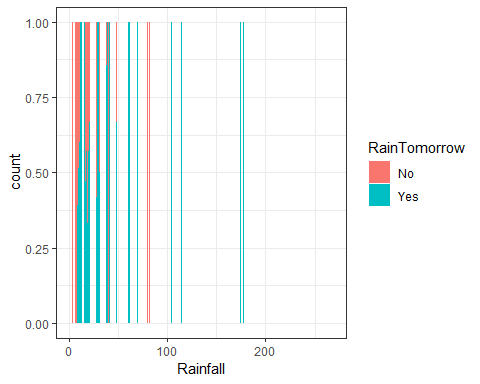


ggplot(rain1, aes(x = MaxTemp, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

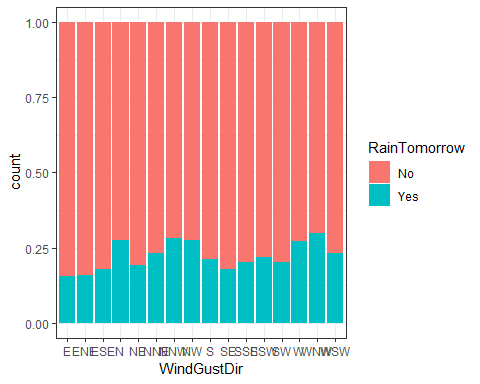


ggplot(rain1, aes(x = Rainfall, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

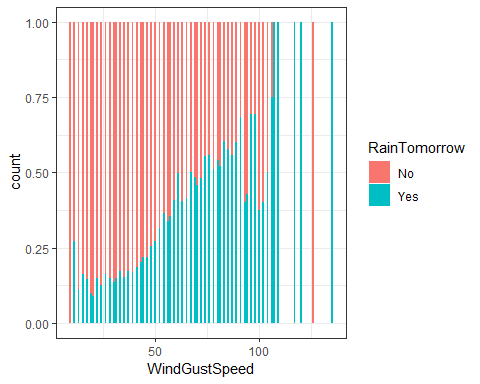
## Warning: position\_stack requires non-overlapping x intervals



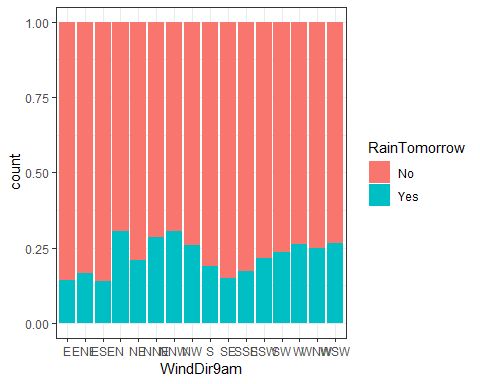
ggplot(rain1, aes(x = WindGustDir, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



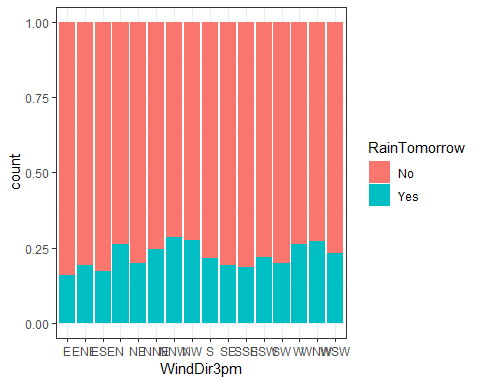
ggplot(rain1, aes(x = WindGustSpeed, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



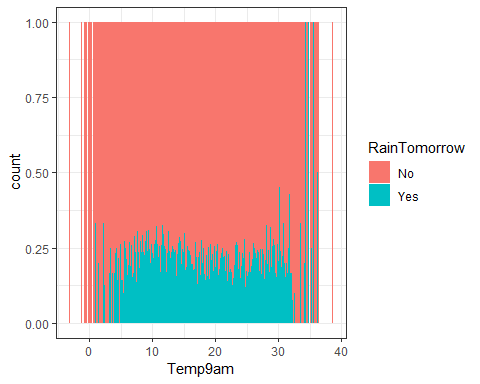
ggplot(rain1, aes(x = WindDir9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



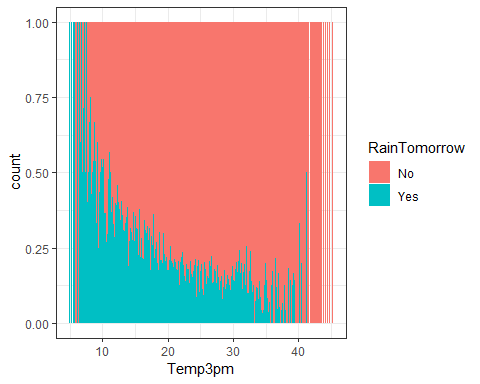
ggplot(rain1, aes(x = WindDir3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain1, aes(x = Temp9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain1, aes(x = Temp3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

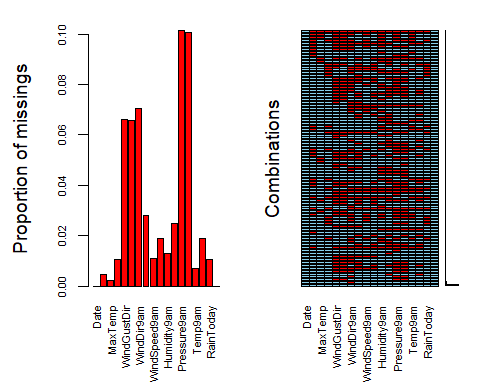


## Do it all again and make a rain2 dataset that performs imputations rather than row-wise deletion to see how risky it might be to impute this data rather than delete it.

Column-wise deletion of the Cloud9am and Cloud3pm variables.

rain2 = rain %>% select(-Cloud9am, -Cloud3pm)   
vim\_plot = aggr(rain2, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

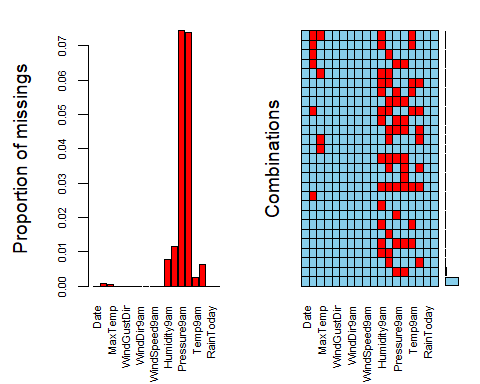
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



Row-wise deletion of any categorical rows with NAs is:

rain2 = rain2 %>% drop\_na(WindGustDir, WindDir9am, WindDir3pm, RainToday)   
vim\_plot = aggr(rain2, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



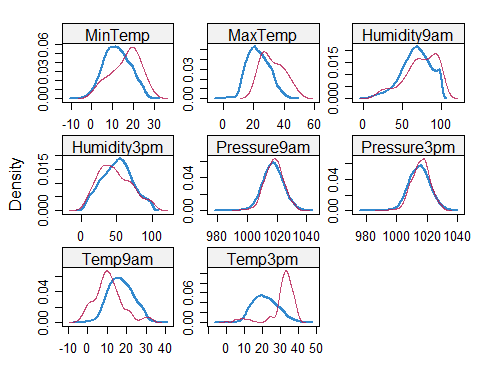
Use the “mice” package to do imputute data.

#select only variables relevant to our analysis  
#rain2 = rain2 %>% select(c("Date", "Rainfall", "WindGustDir", "WindGustSpeed", "WindDir9am", "WindSpeed9am", "Humidity9am", "Pressure9am", "Temp9am", "RainToday", "RainTomorrow"))  
rain2\_imp = mice(rain2, m=1, method = "pmm", seed = 12345)

##   
## iter imp variable  
## 1 1 MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm  
## 2 1 MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm  
## 3 1 MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm  
## 4 1 MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm  
## 5 1 MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm

## Warning: Number of logged events: 1

densityplot(rain2\_imp) #red imputed, blue original

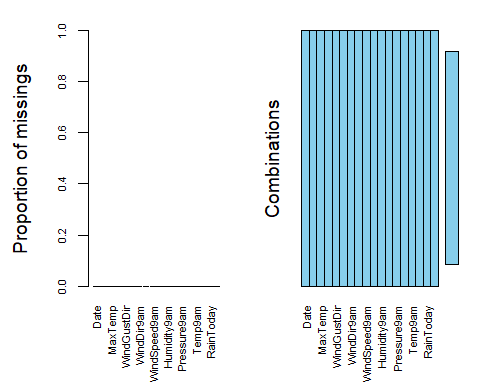


Merge the imputed values into our rain data frame

rain2\_complete = complete(rain2\_imp)   
#summary(rain2\_complete)

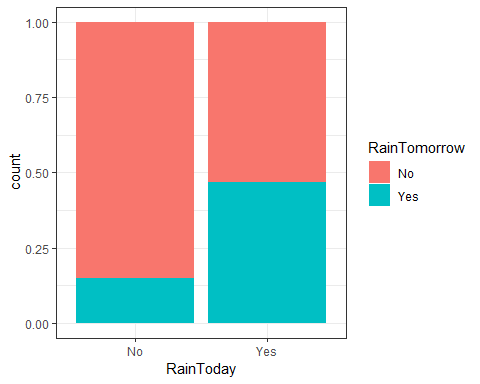
vim\_plot = aggr(rain2\_complete, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

## Warning in plot.aggr(res, ...): not enough horizontal space to display  
## frequencies



# Visualizations of rain2 data

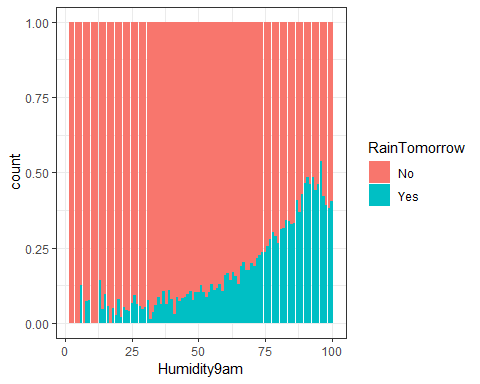
ggplot(rain2\_complete, aes(x = RainToday, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain2\_RainToday.png")

## Saving 5 x 4 in image

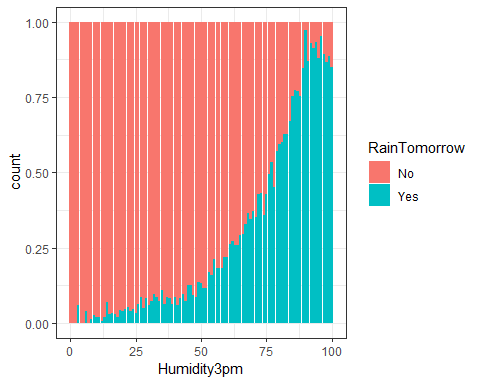
ggplot(rain2\_complete, aes(x = Humidity9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggsave("rain2\_Humidity9.png")

## Saving 5 x 4 in image

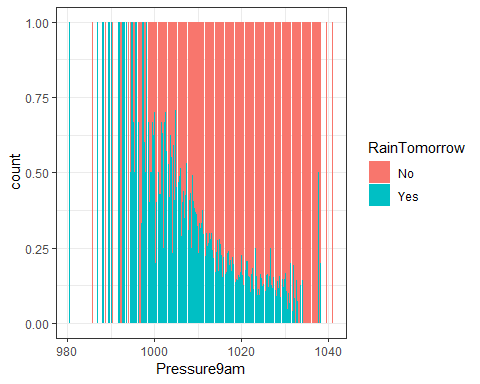
ggplot(rain1, aes(x = Humidity3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



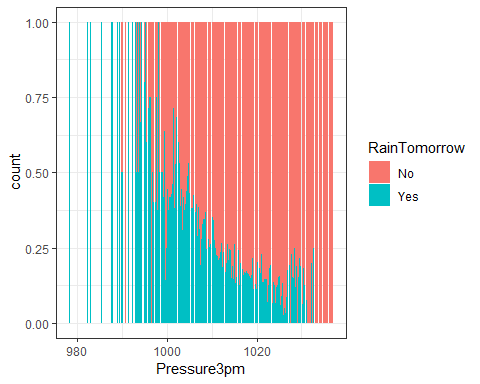
ggsave("rain2\_Humidity3.png")

## Saving 5 x 4 in image

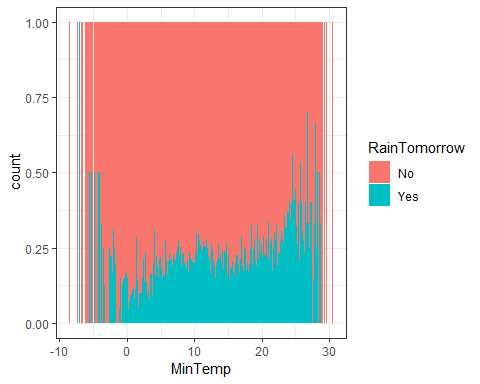
ggplot(rain2\_complete, aes(x = Pressure9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



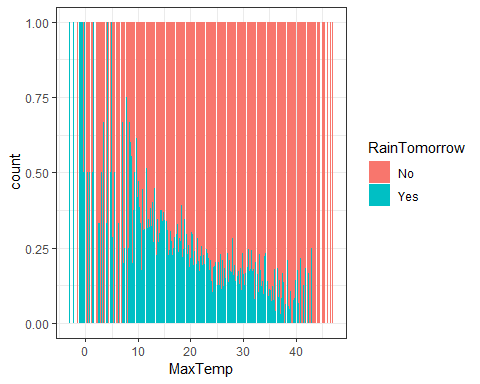
ggplot(rain2\_complete, aes(x = Pressure3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain2\_complete, aes(x = MinTemp, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

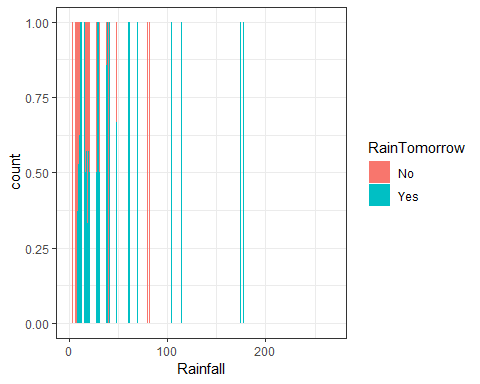


ggplot(rain2\_complete, aes(x = MaxTemp, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

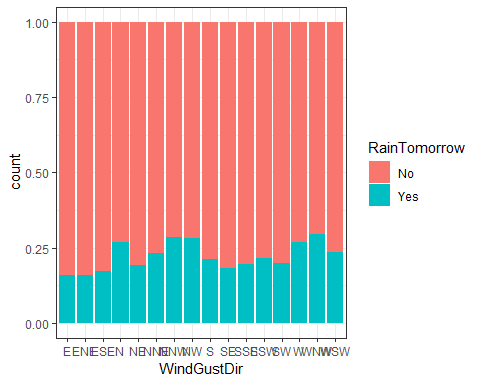


ggplot(rain2\_complete, aes(x = Rainfall, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

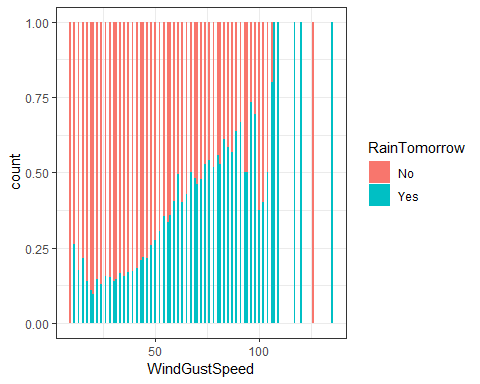
## Warning: position\_stack requires non-overlapping x intervals



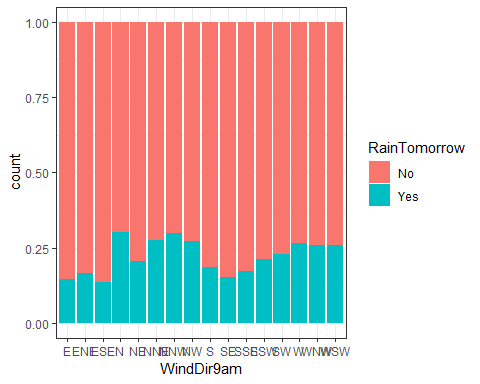
ggplot(rain2\_complete, aes(x = WindGustDir, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



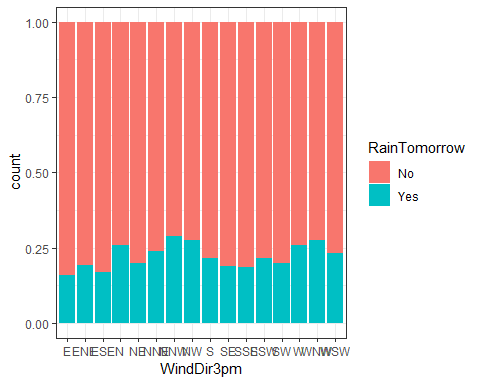
ggplot(rain2\_complete, aes(x = WindGustSpeed, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



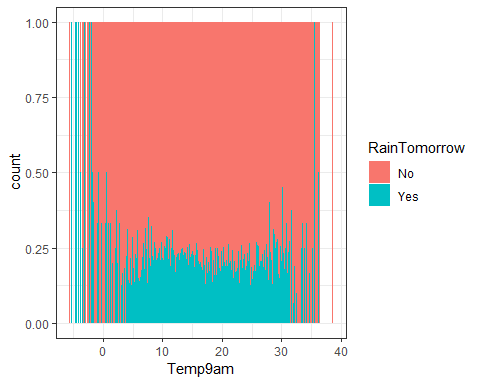
ggplot(rain2\_complete, aes(x = WindDir9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain2\_complete, aes(x = WindDir3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain2\_complete, aes(x = Temp9am, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()



ggplot(rain2\_complete, aes(x = Temp3pm, fill = RainTomorrow)) + geom\_bar(position="fill") + theme\_bw()

