Phase 1

# Course Project Phase 1

## BAN 502

### Danielle Harris

# Load the libraries

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(GGally)

##   
## Attaching package: 'GGally'

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

library(ggcorrplot)  
library(mice)

## Loading required package: lattice

##   
## 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 3.5.2

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.2

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

## 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

# Read in the data

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

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## Date = col\_character(),  
## WindGustDir = col\_character(),  
## WindGustSpeed = col\_integer(),  
## WindDir9am = col\_character(),  
## WindDir3pm = col\_character(),  
## WindSpeed9am = col\_integer(),  
## WindSpeed3pm = col\_integer(),  
## Humidity9am = col\_integer(),  
## Humidity3pm = col\_integer(),  
## Cloud9am = col\_integer(),  
## Cloud3pm = col\_integer(),  
## RainToday = col\_character(),  
## RainTomorrow = col\_character()  
## )

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

# View the structure of the data

str(rain)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 65499 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: int 41 56 50 22 41 48 41 48 37 39 ...  
## $ WindDir9am : chr "ENE" "W" NA "SSW" ...  
## $ WindDir3pm : chr "NW" "W" "WNW" "E" ...  
## $ WindSpeed9am : int 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm : int 20 24 22 9 20 30 11 24 9 17 ...  
## $ Humidity9am : int 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : int 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 : int 7 NA 0 8 NA NA NA NA NA NA ...  
## $ Cloud3pm : int 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")=List of 2  
## ..$ cols :List of 20  
## .. ..$ Date : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ MinTemp : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ MaxTemp : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Rainfall : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ WindGustDir : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ WindGustSpeed: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ WindDir9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ WindDir3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ WindSpeed9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ WindSpeed3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Humidity9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Humidity3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Pressure9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Pressure3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Cloud9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Cloud3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Temp9am : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ Temp3pm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ RainToday : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ RainTomorrow : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

summary(rain)

## Date MinTemp MaxTemp Rainfall   
## Length:65499 Min. :-8.50 Min. :-3.00 Min. : 0.00   
## Class :character 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 0.00   
## Mode :character Median :12.00 Median :22.60 Median : 0.00   
## Mean :12.16 Mean :23.18 Mean : 2.26   
## 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 0.65   
## Max. :30.50 Max. :47.00 Max. :268.60   
## NA's :37628 NA's :37560 NA's :37791   
## WindGustDir WindGustSpeed WindDir9am WindDir3pm   
## Length:65499 Min. : 7.00 Length:65499 Length:65499   
## Class :character 1st Qu.: 31.00 Class :character Class :character   
## Mode :character Median : 39.00 Mode :character Mode :character   
## Mean : 40.02   
## 3rd Qu.: 48.00   
## Max. :135.00   
## NA's :39336   
## WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm   
## Min. : 0.00 Min. : 0.00 Min. : 1.00 Min. : 0.00   
## 1st Qu.: 7.00 1st Qu.:13.00 1st Qu.: 57.00 1st Qu.: 36.00   
## Median :13.00 Median :19.00 Median : 70.00 Median : 52.00   
## Mean :13.97 Mean :18.59 Mean : 68.86 Mean : 51.54   
## 3rd Qu.:19.00 3rd Qu.:24.00 3rd Qu.: 83.00 3rd Qu.: 66.00   
## Max. :87.00 Max. :83.00 Max. :100.00 Max. :100.00   
## NA's :37804 NA's :38022 NA's :37862 NA's :38190   
## Pressure9am Pressure3pm Cloud9am Cloud3pm   
## Min. : 980.5 Min. : 978.2 Min. :0.00 Min. :0.00   
## 1st Qu.:1013.0 1st Qu.:1010.5 1st Qu.:1.00 1st Qu.:2.00   
## Median :1017.7 Median :1015.3 Median :5.00 Median :5.00   
## Mean :1017.7 Mean :1015.3 Mean :4.46 Mean :4.51   
## 3rd Qu.:1022.4 3rd Qu.:1020.0 3rd Qu.:7.00 3rd Qu.:7.00   
## Max. :1041.0 Max. :1037.0 Max. :8.00 Max. :8.00   
## NA's :40333 NA's :40313 NA's :48169 NA's :48837   
## Temp9am Temp3pm RainToday RainTomorrow   
## Min. :-5.60 Min. :-4.20 Length:65499 Length:65499   
## 1st Qu.:12.30 1st Qu.:16.60 Class :character Class :character   
## Median :16.70 Median :21.10 Mode :character Mode :character   
## Mean :16.96 Mean :21.63   
## 3rd Qu.:21.50 3rd Qu.:26.40   
## Max. :38.60 Max. :45.20   
## NA's :37692 NA's :38028

# Factor conversion. Convert the response variable RainTomorrow

rain = rain %>% mutate(RainTomorrow = as.factor(RainTomorrow)) %>%   
 mutate(RainTomorrow = fct\_recode(RainTomorrow, "No" = "1", "Yes" = "2" )) %>%  
 mutate(WindGustDir= as.factor(WindGustDir)) %>%   
 mutate(WindDir9am= as.factor(WindDir9am)) %>%   
 mutate(WindDir3pm= as.factor(WindDir3pm)) %>%  
 mutate(RainToday= as.factor(RainToday)) %>%  
 mutate(RainToday = fct\_recode(RainToday, "No"= "1", "Yes" = "2"))

## Warning: Unknown levels in `f`: 1, 2  
  
## Warning: Unknown levels in `f`: 1, 2

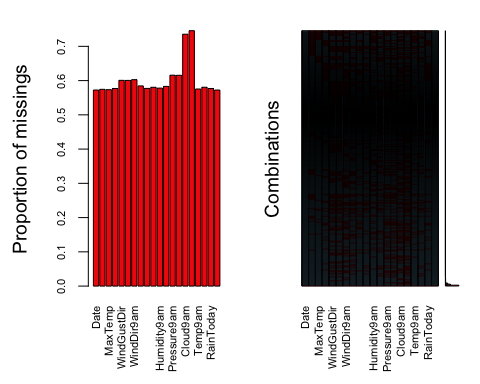
str(rain)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 65499 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: int 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 : int 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm : int 20 24 22 9 20 30 11 24 9 17 ...  
## $ Humidity9am : int 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : int 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 : int 7 NA 0 8 NA NA NA NA NA NA ...  
## $ Cloud3pm : int 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)

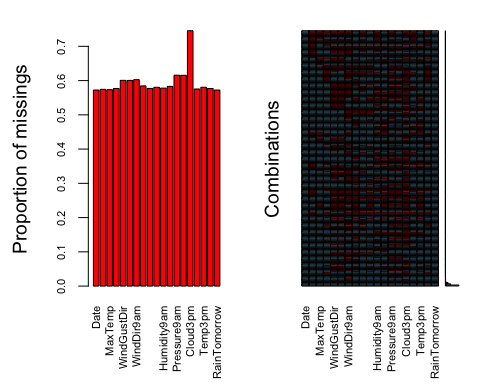
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



# Column-wise deletion of the “Cloud9am” variable

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

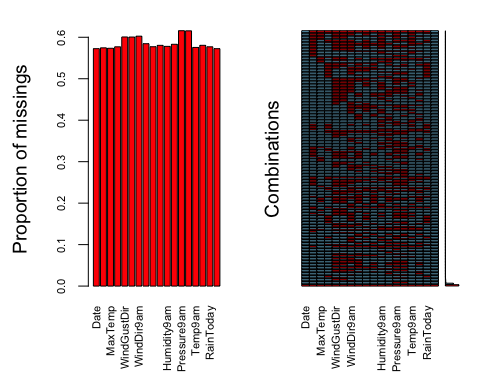
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



# Column-wise deletion of the “Cloud3pm” variable

rain = rain %>% select(-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)

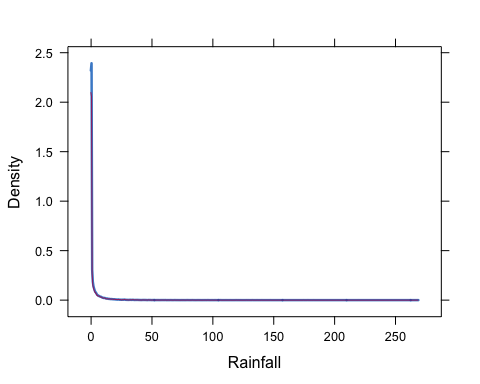


# Use the “mice” package to do imputation

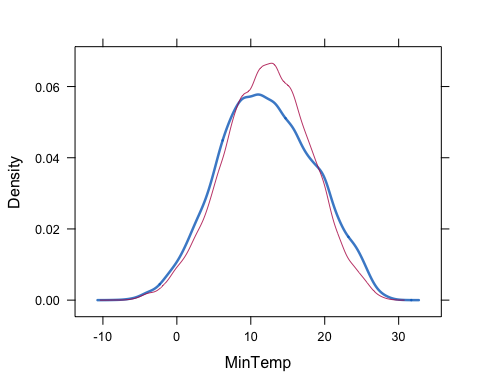
#select only variables relevant to the analysis  
rain2 = rain %>% select(c("Rainfall","MinTemp", "MaxTemp", "Humidity9am", "Humidity3pm", "Pressure9am", "Pressure3pm", "Temp9am", "Temp3pm", "RainToday"))  
  
imp = mice(rain2, m=1, method='pmm', printFlag=FALSE)  
summary(imp)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## Rainfall MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am   
## "pmm" "pmm" "pmm" "pmm" "pmm" "pmm"   
## Pressure3pm Temp9am Temp3pm RainToday   
## "pmm" "pmm" "pmm" "pmm"   
## PredictorMatrix:  
## Rainfall MinTemp MaxTemp Humidity9am Humidity3pm Pressure9am  
## Rainfall 0 1 1 1 1 1  
## MinTemp 1 0 1 1 1 1  
## MaxTemp 1 1 0 1 1 1  
## Humidity9am 1 1 1 0 1 1  
## Humidity3pm 1 1 1 1 0 1  
## Pressure9am 1 1 1 1 1 0  
## Pressure3pm Temp9am Temp3pm RainToday  
## Rainfall 1 1 1 1  
## MinTemp 1 1 1 1  
## MaxTemp 1 1 1 1  
## Humidity9am 1 1 1 1  
## Humidity3pm 1 1 1 1  
## Pressure9am 1 1 1 1

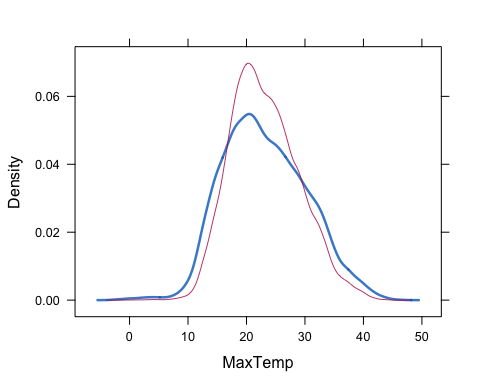
densityplot(imp, ~Rainfall) #red imputed, blue original



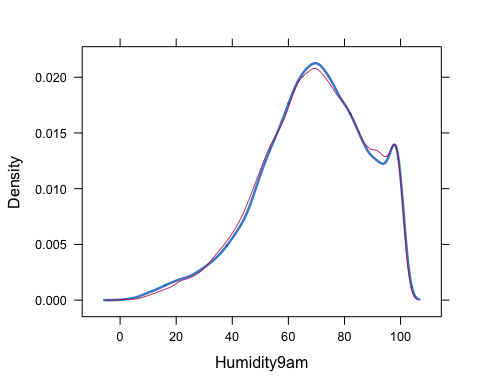
densityplot(imp, ~MinTemp) #red imputed, blue original



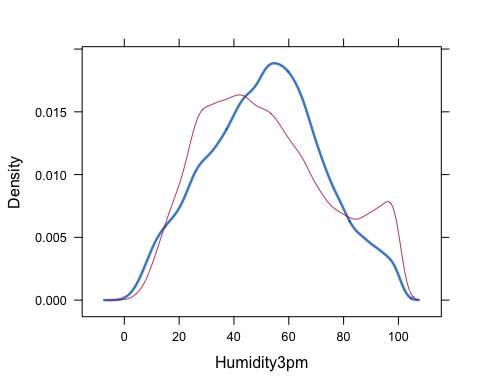
densityplot(imp, ~MaxTemp) #red imputed, blue original



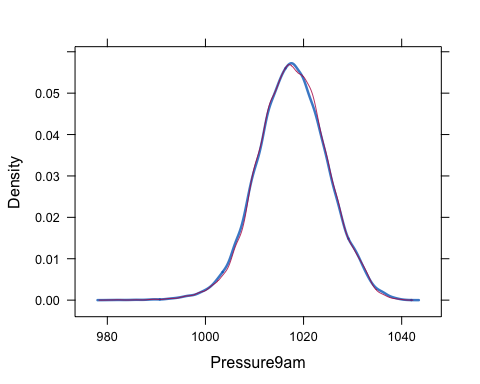
densityplot(imp, ~Humidity9am) #red imputed, blue original



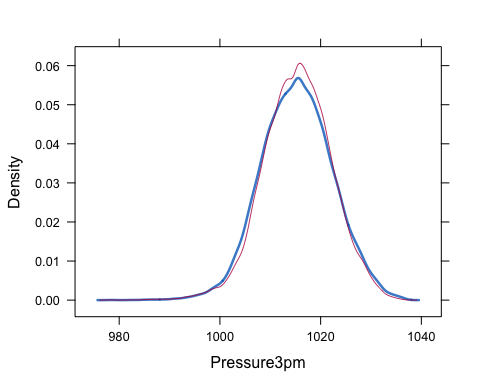
densityplot(imp, ~Humidity3pm) #red imputed, blue original



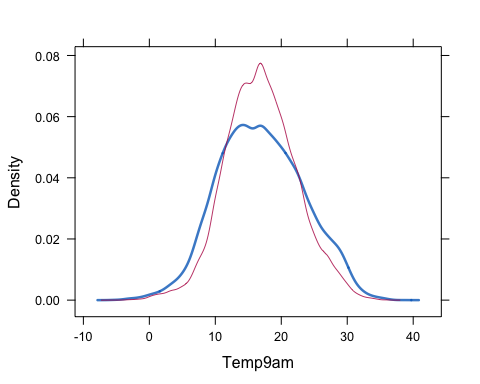
densityplot(imp, ~Pressure9am) #red imputed, blue original



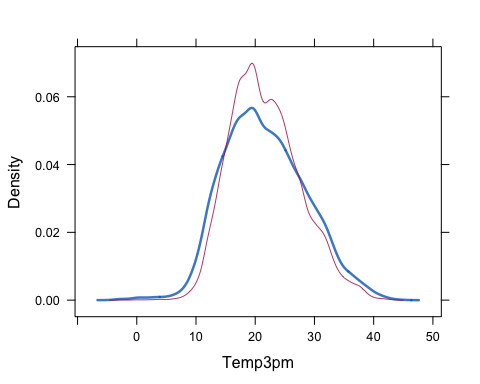
densityplot(imp, ~Pressure3pm) #red imputed, blue original



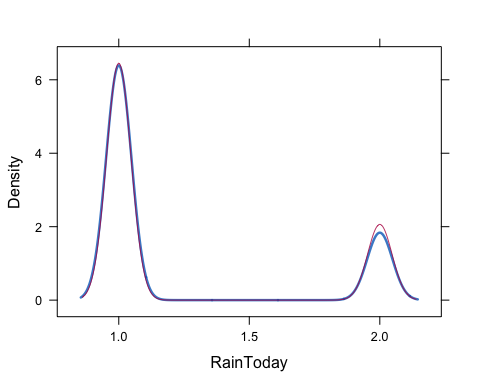
densityplot(imp, ~Temp9am) #red imputed, blue original



densityplot(imp, ~Temp3pm) #red imputed, blue original



densityplot(imp, ~RainToday) #red imputed, blue original



# Merge the imputed values into the rain data frame

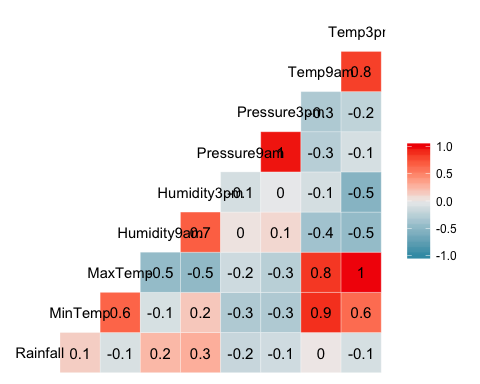
rain = complete(imp)   
summary(rain)

## Rainfall MinTemp MaxTemp Humidity9am   
## Min. : 0.000 Min. :-8.50 Min. :-3.00 Min. : 1.00   
## 1st Qu.: 0.000 1st Qu.: 7.90 1st Qu.:18.50 1st Qu.: 57.00   
## Median : 0.000 Median :12.10 Median :22.60 Median : 70.00   
## Mean : 2.436 Mean :12.18 Mean :23.16 Mean : 68.91   
## 3rd Qu.: 0.800 3rd Qu.:16.50 3rd Qu.:27.50 3rd Qu.: 83.00   
## Max. :268.600 Max. :30.50 Max. :47.00 Max. :100.00   
## Humidity3pm Pressure9am Pressure3pm Temp9am   
## Min. : 0.00 Min. : 980.5 Min. : 978.2 Min. :-5.60   
## 1st Qu.: 34.00 1st Qu.:1013.1 1st Qu.:1010.8 1st Qu.:12.90   
## Median : 50.00 Median :1017.7 Median :1015.4 Median :16.70   
## Mean : 51.64 Mean :1017.7 Mean :1015.4 Mean :16.93   
## 3rd Qu.: 67.00 3rd Qu.:1022.5 3rd Qu.:1020.0 3rd Qu.:20.80   
## Max. :100.00 Max. :1041.0 Max. :1037.0 Max. :38.60   
## Temp3pm RainToday   
## Min. :-4.20 No :50164   
## 1st Qu.:17.00 Yes:15335   
## Median :21.00   
## Mean :21.64   
## 3rd Qu.:25.80   
## Max. :45.20

# Correlation plots with ggcorr

#ggcorr  
ggcorr(rain, label = TRUE)

## Warning in ggcorr(rain, label = TRUE): data in column(s) 'RainToday' are  
## not numeric and were ignored



# Visualization

ggpairs(rain)

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

