## Course Project - Part 1

### Nicholas Barrett, 02/24/19

options(tidyverse\_quiet = TRUE)  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------ tidyverse 1.2.1 --

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

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

library(GGally)

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

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

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 3.5.2

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

library(rpart)

## Warning: package 'rpart' was built under R version 3.5.2

library(caret)

## Warning: package 'caret' was built under R version 3.5.2

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.5.2

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.2

##   
## Attaching package: 'gplots'

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

library(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 3.5.2

library(rattle)

## Warning: package 'rattle' was built under R version 3.5.2

## Rattle: A free graphical interface for data science with R.  
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

rain <- read.csv("rain.csv")

glimpse(rain)

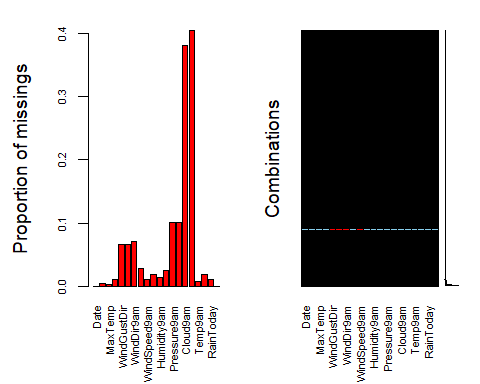
## Observations: 28,003  
## Variables: 20  
## $ Date <fct> 12/5/2008, 12/6/2008, 12/16/2008, 12/17/2008, 12...  
## $ MinTemp <dbl> 17.5, 14.6, 9.8, 14.1, 20.5, 20.1, 9.6, 14.0, 12...  
## $ MaxTemp <dbl> 32.3, 29.7, 27.7, 20.9, 31.8, 32.7, 23.9, 28.3, ...  
## $ Rainfall <dbl> 1.0, 0.2, NA, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,...  
## $ WindGustDir <fct> W, WNW, WNW, ENE, WNW, WNW, W, W, NE, NW, WNW, W...  
## $ WindGustSpeed <int> 41, 56, 50, 22, 41, 48, 41, 48, 37, 39, 44, 28, ...  
## $ WindDir9am <fct> ENE, W, NA, SSW, W, N, WSW, W, SSE, SSE, W, ENE,...  
## $ WindDir3pm <fct> NW, W, WNW, E, W, WNW, SSW, WSW, S, SSW, W, SSW,...  
## $ WindSpeed9am <int> 7, 19, NA, 11, 19, 13, 19, 17, 20, 7, 20, 17, 9,...  
## $ WindSpeed3pm <int> 20, 24, 22, 9, 20, 30, 11, 24, 9, 17, 28, 15, 11...  
## $ Humidity9am <int> 82, 55, 50, 69, 54, 56, 44, 43, 38, 40, 34, 34, ...  
## $ Humidity3pm <int> 33, 23, 28, 82, 24, 15, 22, 15, 16, 8, 28, 17, 1...  
## $ Pressure9am <dbl> 1010.8, 1009.2, 1013.4, 1012.2, 1007.8, 1005.2, ...  
## $ Pressure3pm <dbl> 1006.0, 1005.4, 1010.3, 1010.4, 1005.7, 1001.7, ...  
## $ Cloud9am <int> 7, NA, 0, 8, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ Cloud3pm <int> 8, NA, NA, 1, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ Temp9am <dbl> 17.8, 20.6, 17.3, 17.2, 23.8, 24.6, 14.9, 17.9, ...  
## $ Temp3pm <dbl> 29.7, 28.9, 26.2, 18.1, 30.8, 32.1, 22.1, 27.6, ...  
## $ RainToday <fct> No, No, NA, No, No, No, No, No, No, No, No, No, ...  
## $ RainTomorrow <fct> No, No, No, Yes, No, No, No, No, No, No, No, No,...

summary(rain)

## Date MinTemp MaxTemp Rainfall   
## 5/4/2015 : 21 Min. :-8.50 Min. :-3.00 Min. : 0.000   
## 1/5/2009 : 19 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 0.000   
## 5/11/2014: 18 Median :12.00 Median :22.60 Median : 0.000   
## 1/14/2014: 17 Mean :12.16 Mean :23.18 Mean : 2.265   
## 1/23/2016: 17 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 0.650   
## 1/26/2016: 17 Max. :30.50 Max. :47.00 Max. :268.600   
## (Other) :27894 NA's :132 NA's :64 NA's :295   
## WindGustDir WindGustSpeed WindDir9am WindDir3pm   
## W : 1913 Min. : 7.00 N : 2275 SE : 2036   
## SE : 1863 1st Qu.: 31.00 SSE : 1781 W : 1924   
## E : 1810 Median : 39.00 SE : 1764 S : 1877   
## S : 1779 Mean : 40.02 E : 1739 SSE : 1862   
## N : 1778 3rd Qu.: 48.00 NW : 1711 WSW : 1850   
## (Other):17009 Max. :135.00 (Other):16756 (Other):17669   
## NA's : 1851 NA's :1840 NA's : 1977 NA's : 785   
## 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 :308 NA's :526 NA's :366 NA's :694   
## Pressure9am Pressure3pm Cloud9am Cloud3pm   
## Min. : 980.5 Min. : 978.2 Min. :0.000 Min. :0.000   
## 1st Qu.:1013.0 1st Qu.:1010.5 1st Qu.:1.000 1st Qu.:2.000   
## Median :1017.7 Median :1015.3 Median :5.000 Median :5.000   
## Mean :1017.7 Mean :1015.3 Mean :4.459 Mean :4.513   
## 3rd Qu.:1022.4 3rd Qu.:1020.0 3rd Qu.:7.000 3rd Qu.:7.000   
## Max. :1041.0 Max. :1037.0 Max. :8.000 Max. :8.000   
## NA's :2837 NA's :2817 NA's :10673 NA's :11341   
## Temp9am Temp3pm RainToday RainTomorrow  
## Min. :-5.60 Min. :-4.20 No :21525 No :21713   
## 1st Qu.:12.30 1st Qu.:16.60 Yes : 6183 Yes: 6290   
## Median :16.70 Median :21.10 NA's: 295   
## Mean :16.96 Mean :21.63   
## 3rd Qu.:21.50 3rd Qu.:26.40   
## Max. :38.60 Max. :45.20   
## NA's :196 NA's :532

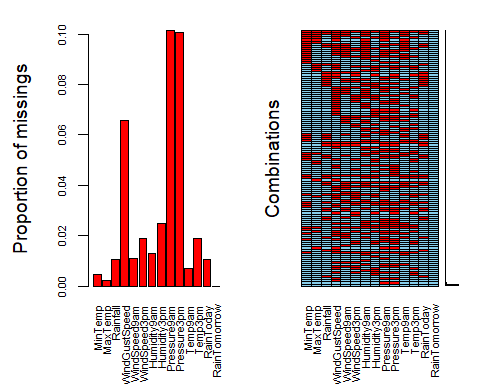
vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

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



rain = rain %>% select(-Date, -Cloud3pm, -Cloud9am, -WindGustDir, -WindDir3pm, -WindDir9am)  
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)



rain = rain %>% drop\_na()  
str(rain)

## 'data.frame': 23494 obs. of 14 variables:  
## $ MinTemp : num 17.5 14.6 14.1 20.5 20.1 9.6 14 12.5 17.4 19.8 ...  
## $ MaxTemp : num 32.3 29.7 20.9 31.8 32.7 23.9 28.3 28.4 43 32.7 ...  
## $ Rainfall : num 1 0.2 0 0 0 0 0 0 0 0 ...  
## $ WindGustSpeed: int 41 56 22 41 48 41 48 37 39 44 ...  
## $ WindSpeed9am : int 7 19 11 19 13 19 17 20 7 20 ...  
## $ WindSpeed3pm : int 20 24 9 20 30 11 24 9 17 28 ...  
## $ Humidity9am : int 82 55 69 54 56 44 43 38 40 34 ...  
## $ Humidity3pm : int 33 23 82 24 15 22 15 16 8 28 ...  
## $ Pressure9am : num 1011 1009 1012 1008 1005 ...  
## $ Pressure3pm : num 1006 1005 1010 1006 1002 ...  
## $ Temp9am : num 17.8 20.6 17.2 23.8 24.6 14.9 17.9 17.2 25.6 27.6 ...  
## $ Temp3pm : num 29.7 28.9 18.1 30.8 32.1 22.1 27.6 26.6 41.5 27.1 ...  
## $ RainToday : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...

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

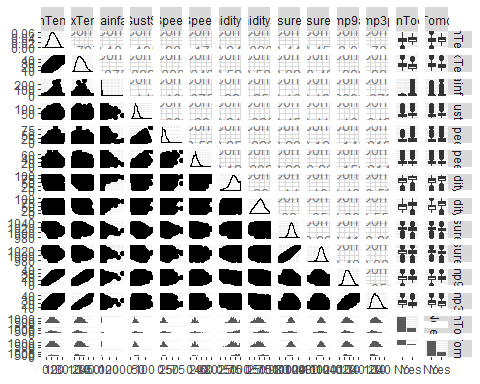
## Warning: Unknown levels in `f`: 1, 0  
  
## Warning: Unknown levels in `f`: 1, 0

str(rain)

## 'data.frame': 23494 obs. of 14 variables:  
## $ MinTemp : num 17.5 14.6 14.1 20.5 20.1 9.6 14 12.5 17.4 19.8 ...  
## $ MaxTemp : num 32.3 29.7 20.9 31.8 32.7 23.9 28.3 28.4 43 32.7 ...  
## $ Rainfall : num 1 0.2 0 0 0 0 0 0 0 0 ...  
## $ WindGustSpeed: int 41 56 22 41 48 41 48 37 39 44 ...  
## $ WindSpeed9am : int 7 19 11 19 13 19 17 20 7 20 ...  
## $ WindSpeed3pm : int 20 24 9 20 30 11 24 9 17 28 ...  
## $ Humidity9am : int 82 55 69 54 56 44 43 38 40 34 ...  
## $ Humidity3pm : int 33 23 82 24 15 22 15 16 8 28 ...  
## $ Pressure9am : num 1011 1009 1012 1008 1005 ...  
## $ Pressure3pm : num 1006 1005 1010 1006 1002 ...  
## $ Temp9am : num 17.8 20.6 17.2 23.8 24.6 14.9 17.9 17.2 25.6 27.6 ...  
## $ Temp3pm : num 29.7 28.9 18.1 30.8 32.1 22.1 27.6 26.6 41.5 27.1 ...  
## $ RainToday : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...

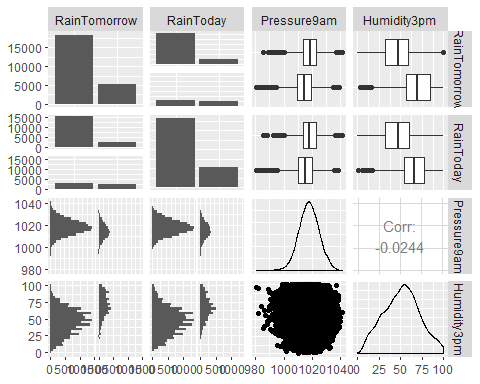
ggpairs(rain)

## `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`.  
## `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`.  
## `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`.



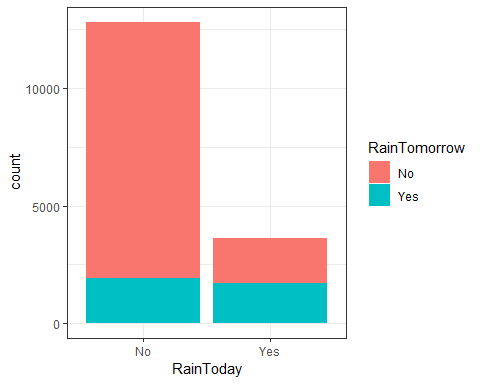
ggpairs(rain, columns = c("RainTomorrow", "RainToday", "Pressure9am", "Humidity3pm"))

## `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`.

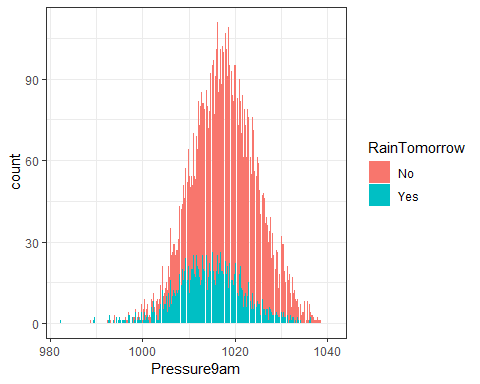


set.seed(12345)  
  
train.rows = createDataPartition(y = rain$RainTomorrow, p=0.7, list = FALSE)   
train = rain[train.rows,]   
test = rain[-train.rows,]

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



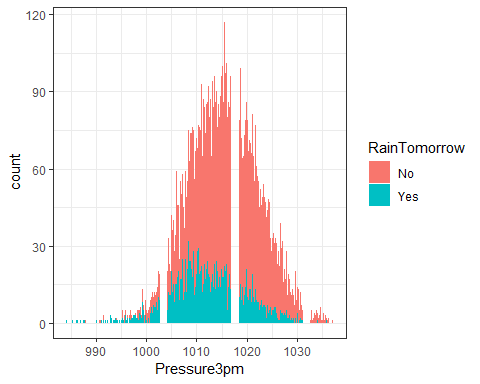
ggplot(train, aes(x = Pressure9am, fill = RainTomorrow)) + geom\_bar() + theme\_bw()



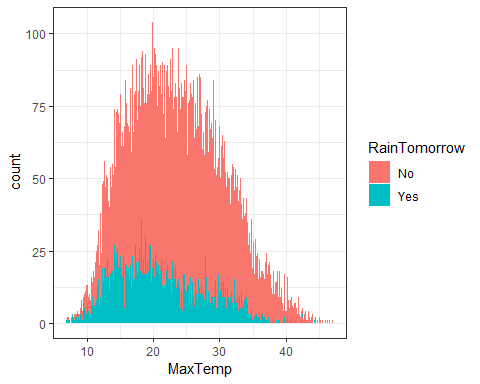
ggsave("9amPressure.png")

## Saving 5 x 4 in image

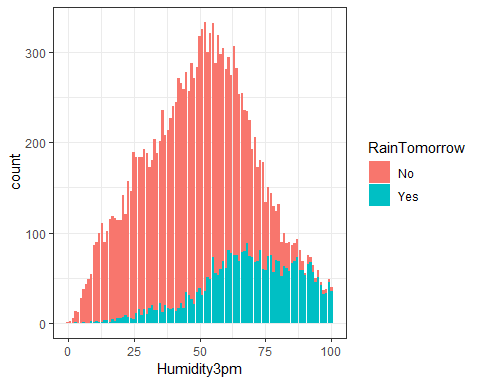
ggplot(train, aes(x = Pressure3pm, fill = RainTomorrow)) + geom\_bar() + theme\_bw()



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



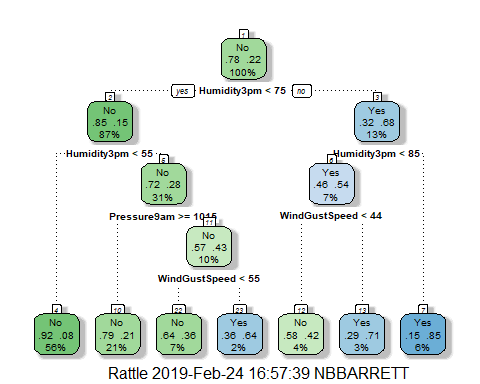
ggplot(train, aes(x = Humidity3pm, fill = RainTomorrow)) + geom\_bar() + theme\_bw()



ggsave("humidy.png")

## Saving 5 x 4 in image

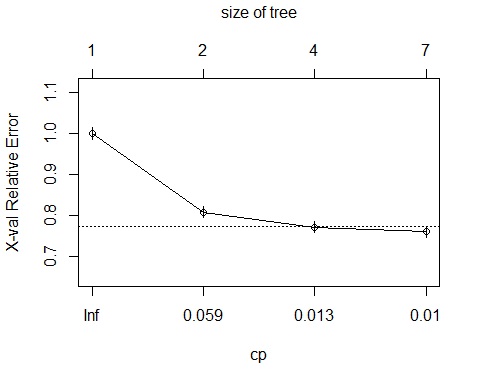
tree1 = rpart(RainTomorrow ~., train, method = "class")  
fancyRpartPlot(tree1)



printcp(tree1)

##   
## Classification tree:  
## rpart(formula = RainTomorrow ~ ., data = train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Humidity3pm Pressure9am WindGustSpeed  
##   
## Root node error: 3610/16447 = 0.21949  
##   
## n= 16447   
##   
## CP nsplit rel error xerror xstd  
## 1 0.218006 0 1.00000 1.00000 0.014704  
## 2 0.016205 1 0.78199 0.80803 0.013570  
## 3 0.010157 3 0.74958 0.77119 0.013322  
## 4 0.010000 6 0.71911 0.75956 0.013241

plotcp(tree1)



treepred = predict(tree1, train, type = "class")  
head(treepred)

## 1 3 6 7 8 9   
## No No No No No No   
## Levels: No Yes

confusionMatrix(treepred, train$RainTomorrow, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 12413 2172  
## Yes 424 1438  
##   
## Accuracy : 0.8422   
## 95% CI : (0.8365, 0.8477)  
## No Information Rate : 0.7805   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4423   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.39834   
## Specificity : 0.96697   
## Pos Pred Value : 0.77229   
## Neg Pred Value : 0.85108   
## Prevalence : 0.21949   
## Detection Rate : 0.08743   
## Detection Prevalence : 0.11321   
## Balanced Accuracy : 0.68265   
##   
## 'Positive' Class : Yes   
##

treepred\_test = predict(tree1, test, type = "class")  
head(treepred\_test)

## 2 4 5 11 12 14   
## No No No No No No   
## Levels: No Yes

confusionMatrix(treepred\_test, test$RainTomorrow, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 5304 962  
## Yes 197 584  
##   
## Accuracy : 0.8355   
## 95% CI : (0.8267, 0.8441)  
## No Information Rate : 0.7806   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4159   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.37775   
## Specificity : 0.96419   
## Pos Pred Value : 0.74776   
## Neg Pred Value : 0.84647   
## Prevalence : 0.21938   
## Detection Rate : 0.08287   
## Detection Prevalence : 0.11083   
## Balanced Accuracy : 0.67097   
##   
## 'Positive' Class : Yes   
##