# BAN502-803: Course Project - Phase 1

## Gregory Moore

### Data Exploration, Preparation, and Visualization

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

##   
## Attaching package: 'GGally'

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

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)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

##   
## 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(MASS)

##   
## Attaching package: 'MASS'

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

library(caret)

##   
## Attaching package: 'caret'

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

library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

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.

## Initial Look at the data, look for fields with large amounts of NA values, and look for whitespaces

#str(rain)  
summary(rain)

## Date MinTemp MaxTemp Rainfall   
## Length:28003 Min. :-8.50 Min. :-3.00 Min. : 0.000   
## Class :character 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 0.000   
## Mode :character Median :12.00 Median :22.60 Median : 0.000   
## Mean :12.16 Mean :23.18 Mean : 2.265   
## 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 0.650   
## Max. :30.50 Max. :47.00 Max. :268.600   
## NA's :132 NA's :64 NA's :295   
## WindGustDir WindGustSpeed WindDir9am WindDir3pm   
## Length:28003 Min. : 7.00 Length:28003 Length:28003   
## 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 :1840   
## 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 Length:28003 Length:28003   
## 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 :196 NA's :532

rain\_blank = rain %>% filter\_all(any\_vars(. == ""))  
rain\_blank

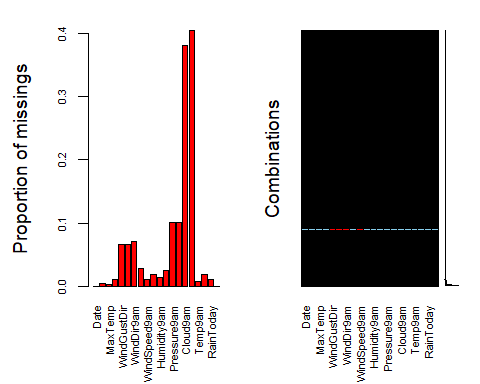
## # A tibble: 0 x 20  
## # ... with 20 variables: Date <chr>, MinTemp <dbl>, MaxTemp <dbl>,  
## # Rainfall <dbl>, WindGustDir <chr>, WindGustSpeed <dbl>,  
## # WindDir9am <chr>, WindDir3pm <chr>, WindSpeed9am <dbl>,  
## # WindSpeed3pm <dbl>, Humidity9am <dbl>, Humidity3pm <dbl>,  
## # Pressure9am <dbl>, Pressure3pm <dbl>, Cloud9am <dbl>, Cloud3pm <dbl>,  
## # Temp9am <dbl>, Temp3pm <dbl>, RainToday <chr>, RainTomorrow <chr>

rain\_1blank = rain %>% filter\_all(any\_vars(. == " "))  
rain\_1blank

## # A tibble: 0 x 20  
## # ... with 20 variables: Date <chr>, MinTemp <dbl>, MaxTemp <dbl>,  
## # Rainfall <dbl>, WindGustDir <chr>, WindGustSpeed <dbl>,  
## # WindDir9am <chr>, WindDir3pm <chr>, WindSpeed9am <dbl>,  
## # WindSpeed3pm <dbl>, Humidity9am <dbl>, Humidity3pm <dbl>,  
## # Pressure9am <dbl>, Pressure3pm <dbl>, Cloud9am <dbl>, Cloud3pm <dbl>,  
## # Temp9am <dbl>, Temp3pm <dbl>, RainToday <chr>, RainTomorrow <chr>

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)



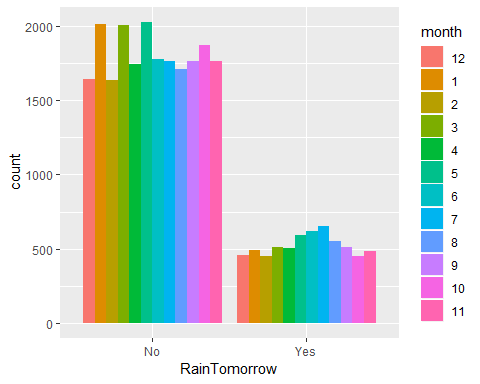
## Split “Date” into three separate fields and convert to factors

rain\_test = separate(rain, Date, sep="/", into = c("month", "day", "year"))  
rain\_test = rain\_test %>%   
 mutate(month = as\_factor(as.character(month))) %>%  
 mutate(day = as\_factor(as.character(day))) %>%  
 mutate(year = as\_factor(as.character(year)))   
#str(rain\_test)

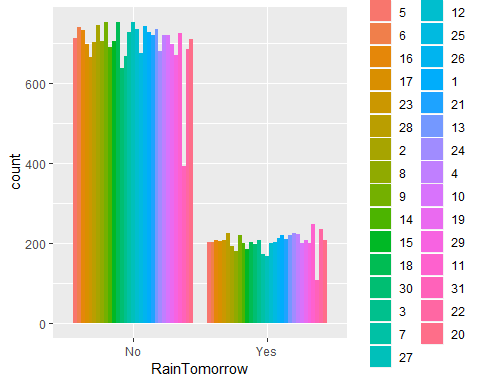
## 

## Visualize new date fields to see if any relationships exist with RainTomorrow

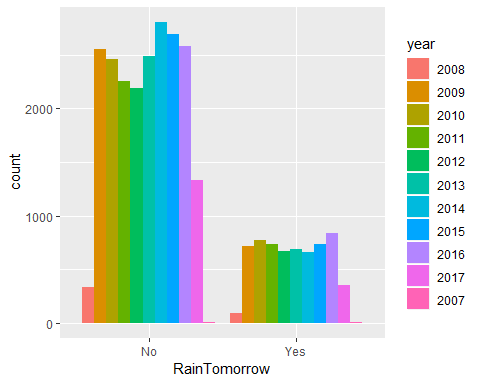
ggplot(rain\_test, aes(x=RainTomorrow, fill = month)) + geom\_bar(position = "dodge")



ggplot(rain\_test, aes(x=RainTomorrow, fill = day)) + geom\_bar(position = "dodge")



ggplot(rain\_test, aes(x=RainTomorrow, fill = year)) + geom\_bar(position = "dodge")



**Surprising results as no real pattern emerged. Consider dropping Date all together**

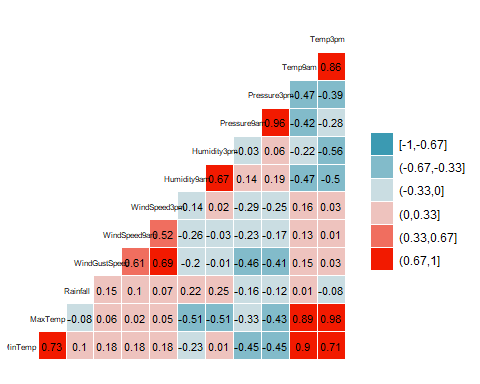
## 

## Convert all CHAR datatypes into Factors and visualize realatoinships between all variables

rain = rain %>%   
 mutate(Cloud9am = as\_factor(as.character(Cloud9am))) %>%  
 mutate(Cloud3pm = as\_factor(as.character(Cloud3pm))) %>%  
 mutate(WindGustDir = as\_factor(as.character(WindGustDir))) %>%  
 mutate(WindDir9am = as\_factor(as.character(WindDir9am))) %>%  
 mutate(WindDir3pm = as\_factor(as.character(WindDir3pm))) %>%  
 mutate(RainToday = as\_factor(as.character(RainToday))) %>%  
 mutate(RainTomorrow = as\_factor(as.character(RainTomorrow)))   
#str(rain)  
#summary(rain)

ggcorr(rain, label = TRUE, label\_round = 2, label\_size = 3, hjust = .6, size = 2, nbreaks = 6)

## Warning in ggcorr(rain, label = TRUE, label\_round = 2, label\_size = 3,  
## hjust = 0.6, : data in column(s) 'Date', 'WindGustDir', 'WindDir9am',  
## 'WindDir3pm', 'Cloud9am', 'Cloud3pm', 'RainToday', 'RainTomorrow' are not  
## numeric and were ignored



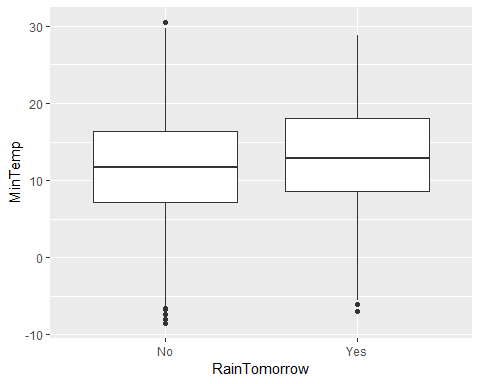
#ggsave("correlation.png")

### Things above to note:

**Rainfall has very little correlation with the other quantitative data. This field could possibly be dropped. Highest correlation (besides MIN/MAX temps with 9am and 3pm measurements) is Temperature and Humidity.**

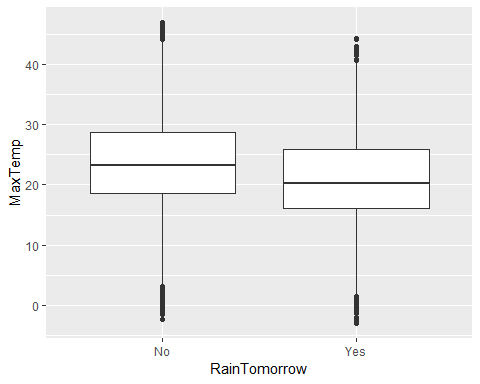
ggplot(rain,aes(x=RainTomorrow, y=MinTemp)) + geom\_boxplot()

## Warning: Removed 132 rows containing non-finite values (stat\_boxplot).



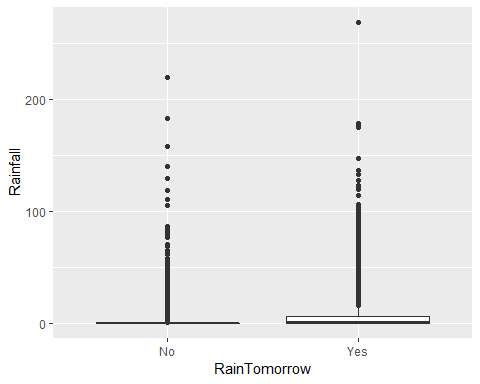
ggplot(rain,aes(x=RainTomorrow,y=MaxTemp)) + geom\_boxplot()

## Warning: Removed 64 rows containing non-finite values (stat\_boxplot).

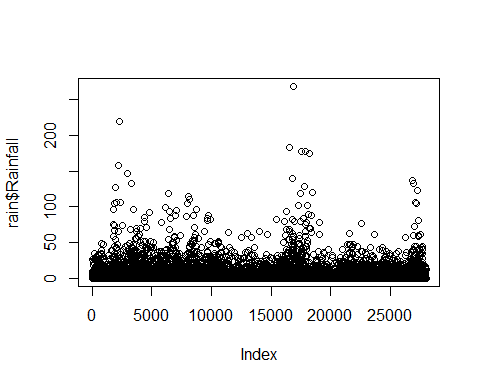


ggplot(rain,aes(x=RainTomorrow,y=Rainfall)) + geom\_boxplot()

## Warning: Removed 295 rows containing non-finite values (stat\_boxplot).



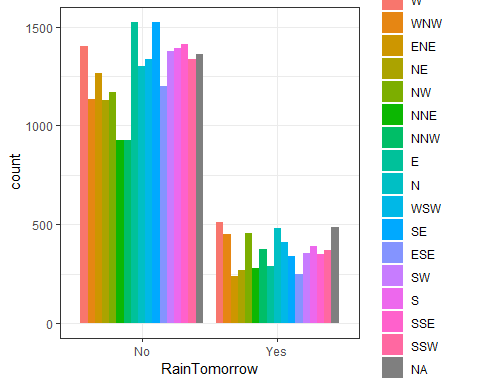
plot(rain$Rainfall)



### Things in the last two plots to note:

**Rainfall has very low averages and overall measurements and there is not a noticeable difference between YES and NO for the RainTomorrow factor levels. There are also many outliers that could skew the results. Drop Rainfall??**

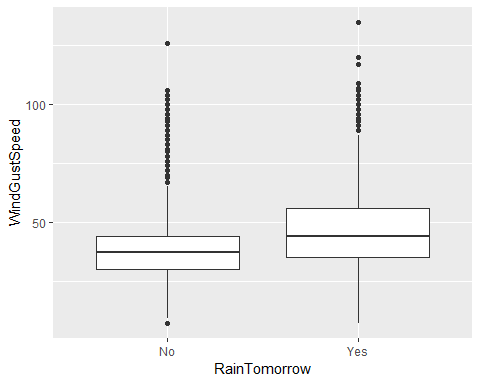
ggplot(rain, aes(x=RainTomorrow, fill = WindGustDir)) + geom\_bar(position = "dodge") + theme\_bw()



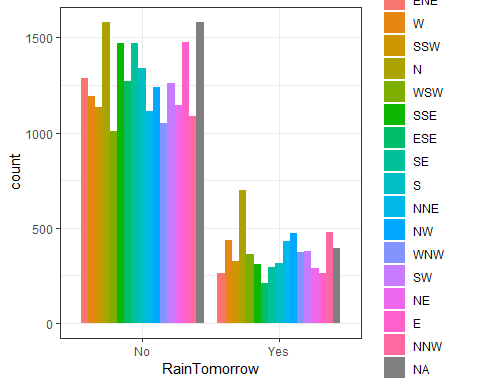
**Winds from the North?? Is this a pattern for YES values?? What about NO’s from the South?**

ggplot(rain,aes(x=RainTomorrow,y=WindGustSpeed)) + geom\_boxplot()

## Warning: Removed 1840 rows containing non-finite values (stat\_boxplot).

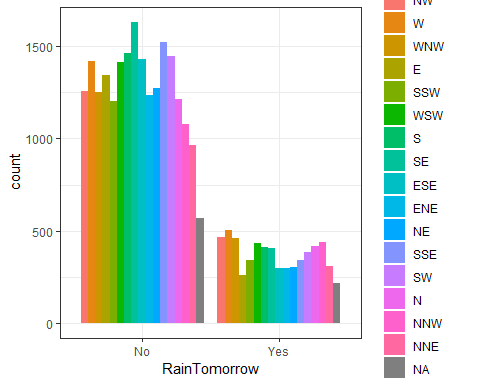


ggplot(rain, aes(x=RainTomorrow, fill = WindDir9am)) + geom\_bar(position = "dodge") + theme\_bw()



**Winds from the North?? Is THIS a pattern??**

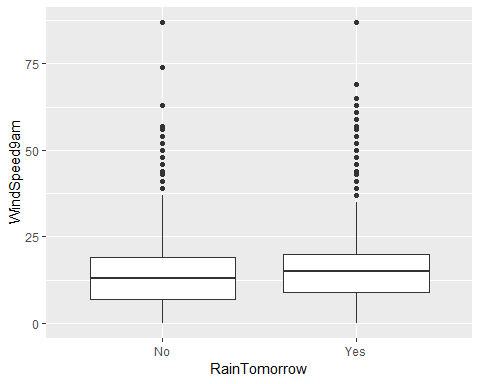
ggplot(rain, aes(x=RainTomorrow, fill = WindDir3pm)) + geom\_bar(position = "dodge") + theme\_bw()



**Spikes for NO with winds from the South and YES peaks with North again.**

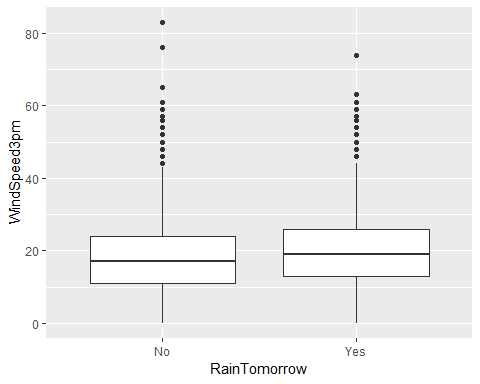
ggplot(rain,aes(x=RainTomorrow,y=WindSpeed9am)) + geom\_boxplot()

## Warning: Removed 308 rows containing non-finite values (stat\_boxplot).



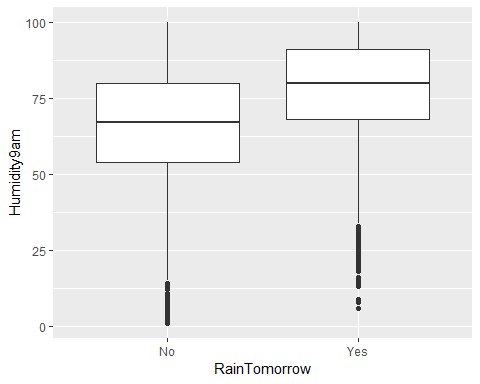
ggplot(rain,aes(x=RainTomorrow,y=WindSpeed3pm)) + geom\_boxplot()

## Warning: Removed 526 rows containing non-finite values (stat\_boxplot).



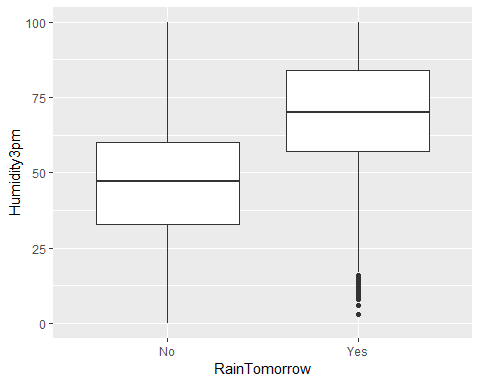
ggplot(rain,aes(x=RainTomorrow,y=Humidity9am)) + geom\_boxplot()

## Warning: Removed 366 rows containing non-finite values (stat\_boxplot).



ggplot(rain,aes(x=RainTomorrow,y=Humidity3pm)) + geom\_boxplot()

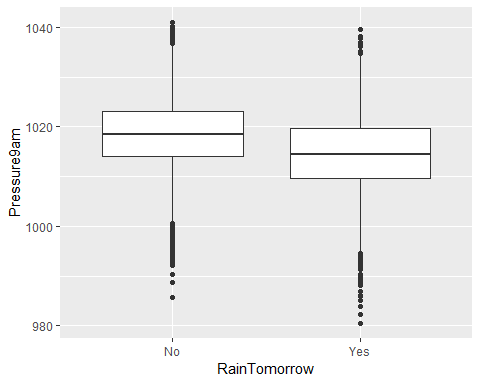
## Warning: Removed 694 rows containing non-finite values (stat\_boxplot).



**Humidity measurements are very interesting. They seemingly have some relationship with RainTomorrow**

ggplot(rain,aes(x=RainTomorrow,y=Pressure9am)) + geom\_boxplot()

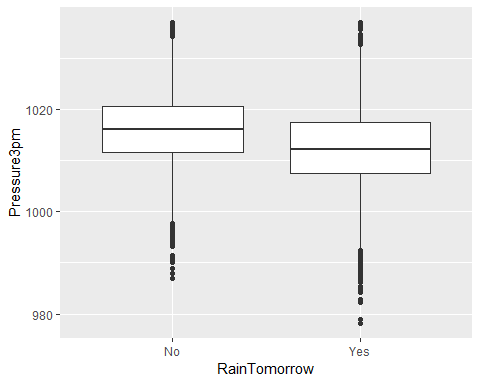
## Warning: Removed 2837 rows containing non-finite values (stat\_boxplot).



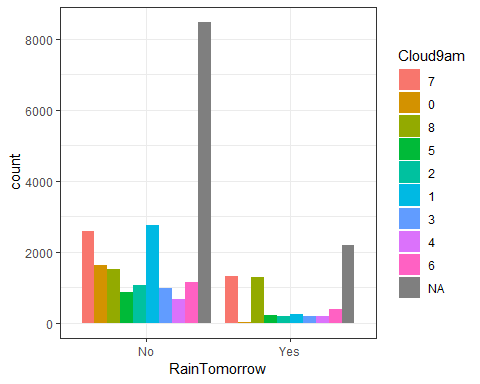
**Humidity measurements are also very interesting with its apparent relationship with RainTomorrow**

ggplot(rain,aes(x=RainTomorrow,y=Pressure3pm)) + geom\_boxplot()

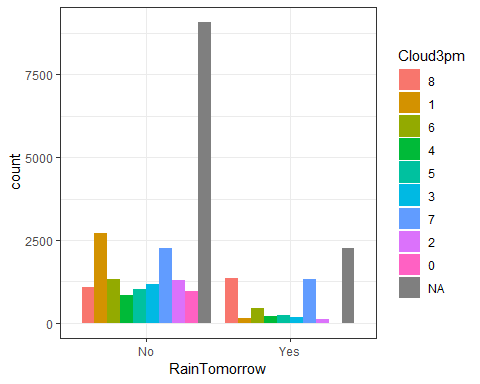
## Warning: Removed 2817 rows containing non-finite values (stat\_boxplot).



ggplot(rain, aes(x=RainTomorrow, fill = Cloud9am)) + geom\_bar(position = "dodge") + theme\_bw()



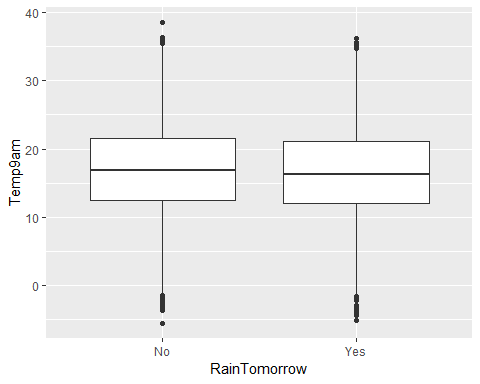
ggplot(rain, aes(x=RainTomorrow, fill = Cloud3pm)) + geom\_bar(position = "dodge") + theme\_bw()



**These Cloud measrements have the most noticable correlation of any of the Factors see up until this point.**

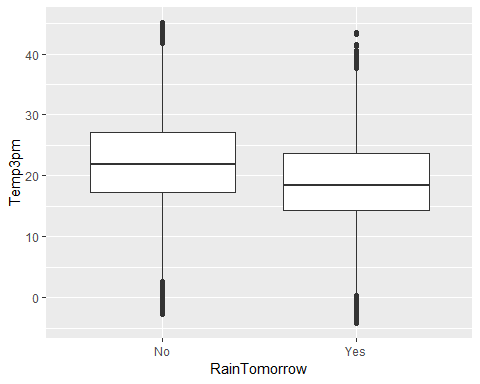
ggplot(rain,aes(x=RainTomorrow,y=Temp9am)) + geom\_boxplot()

## Warning: Removed 196 rows containing non-finite values (stat\_boxplot).



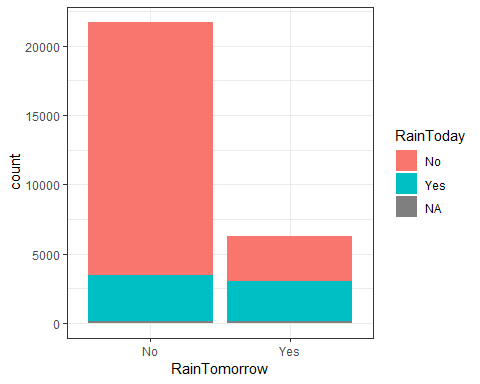
ggplot(rain,aes(x=RainTomorrow,y=Temp3pm)) + geom\_boxplot()

## Warning: Removed 532 rows containing non-finite values (stat\_boxplot).



**All four Temperature measurements were unremarkable. Maybe Drop from the dataset??**

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



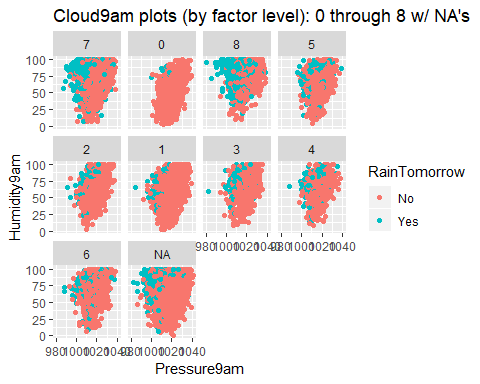
**RainToday seems to be a good indicator of RainTomorrow = YES.**

**\*\*But the NO ratio makes that trend almost unusable. Maybe it will shine more when combined with with some other fields/trends.**

#### Using the visualizations above, I now believe the Cloud cover, Pressure, and Humidity measurements (especially at 3pm) to have the strongest relationship with RainTomorrow (see plots below). However, it should be noted that while temperature shows no real correlation to our response variable, temperature does appear to have some significance to Pressure and Humidity and this fact may be helpful if/when we must impute our data.

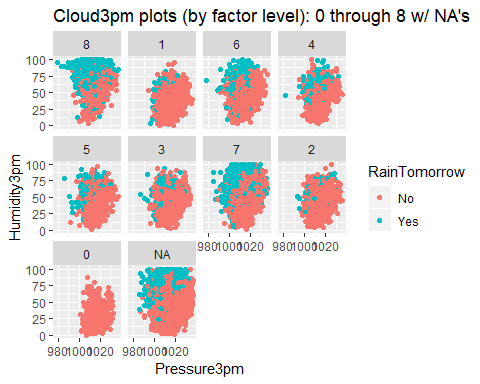
ggplot(rain, aes(x=Pressure9am, y=Humidity9am, color=RainTomorrow)) + geom\_point() + facet\_wrap(~ Cloud9am) +   
 labs(title="Cloud9am plots (by factor level): 0 through 8 w/ NA's")

## Warning: Removed 3036 rows containing missing values (geom\_point).



ggplot(rain, aes(x=Pressure3pm, y=Humidity3pm, color=RainTomorrow)) + geom\_point() + facet\_wrap(~ Cloud3pm) +   
 labs(title="Cloud3pm plots (by factor level): 0 through 8 w/ NA's")

## Warning: Removed 3241 rows containing missing values (geom\_point).



## End of Phase 1

## Make a few decisions based on the visualized data and build a few forward stepwise regression models to test routes

### Model where Date is split and kept and Cloud coverage is cleaned

rain\_test = separate(rain, Date, sep="/", into = c("month", "day", "year"))  
rain\_test = rain\_test %>%   
 mutate(month = as\_factor(as.character(month))) %>%  
 mutate(day = as\_factor(as.character(day))) %>%  
 mutate(year = as\_factor(as.character(year)))   
#str(rain\_test)

rain\_test = rain\_test %>%   
 mutate(Cloud9am = as\_factor(as.character(Cloud9am))) %>%  
 mutate(Cloud3pm = as\_factor(as.character(Cloud3pm))) %>%  
 mutate(WindGustDir = as\_factor(as.character(WindGustDir))) %>%  
 mutate(WindDir9am = as\_factor(as.character(WindDir9am))) %>%  
 mutate(WindDir3pm = as\_factor(as.character(WindDir3pm))) %>%  
 mutate(RainToday = as\_factor(as.character(RainToday))) %>%  
 mutate(RainTomorrow = as\_factor(as.character(RainTomorrow)))   
#summary(rain\_test)

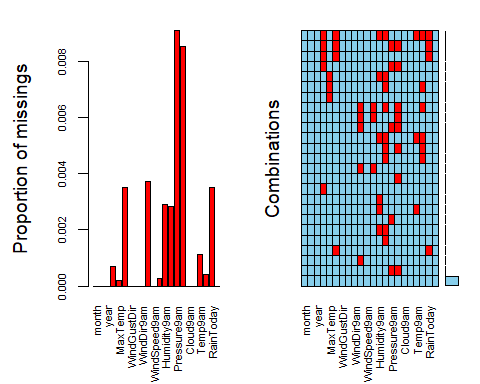
rain\_test\_1 = rain\_test %>% drop\_na(Cloud9am) %>%  
 drop\_na(Cloud3pm)   
#summary(rain\_test\_1)

**The next three WIND variables had a lot of NA’s, so they will also be scrubbed from the dataset.**

rain\_test\_1 = rain\_test\_1 %>% drop\_na(WindGustSpeed) %>%  
 drop\_na(WindGustDir) %>%  
 drop\_na(WindDir9am)

vim\_plot = aggr(rain\_test\_1, numbers = TRUE, prop = c(TRUE, FALSE), cex.axis=.7)

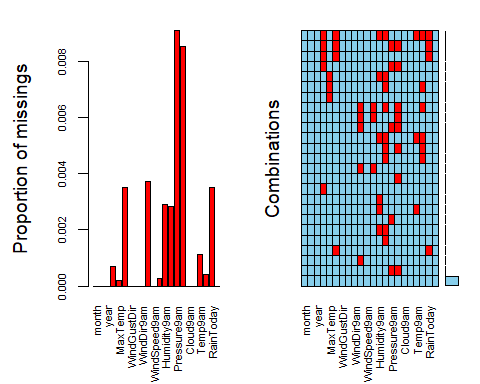
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



rain\_test\_1 = rain\_test\_1 %>% drop\_na(WindGustDir)  
#summary(rain\_test\_1)

vim\_plot = aggr(rain\_test\_1, numbers = TRUE, prop = c(TRUE, FALSE), cex.axis=.7)

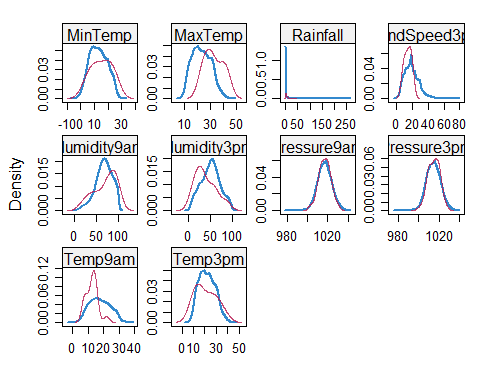
## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



rain\_imp = mice(rain\_test\_1, m=1, method = "pmm", seed = 12345)

##   
## iter imp variable  
## 1 1 MinTemp MaxTemp Rainfall WindDir3pm WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainToday  
## 2 1 MinTemp MaxTemp Rainfall WindDir3pm WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainToday  
## 3 1 MinTemp MaxTemp Rainfall WindDir3pm WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainToday  
## 4 1 MinTemp MaxTemp Rainfall WindDir3pm WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainToday  
## 5 1 MinTemp MaxTemp Rainfall WindDir3pm WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainToday

densityplot(rain\_imp)



rain\_complete = complete(rain\_imp)

set.seed(1234)  
train.rows = createDataPartition(y = rain\_complete$RainTomorrow, p=0.7, list = FALSE)   
train\_cloud = rain\_complete[train.rows,]   
test\_cloud = rain\_complete[-train.rows,]

allmod = glm(RainTomorrow ~., train\_cloud, family = "binomial")   
summary(allmod)

##   
## Call:  
## glm(formula = RainTomorrow ~ ., family = "binomial", data = train\_cloud)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2505 -0.5488 -0.2637 -0.0237 3.1818   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 50.437078 6.039500 8.351 < 2e-16 \*\*\*  
## month1 -0.204976 0.161982 -1.265 0.205718   
## month2 0.007237 0.167227 0.043 0.965481   
## month3 0.124491 0.157428 0.791 0.429072   
## month4 0.158417 0.162827 0.973 0.330596   
## month5 0.365179 0.164413 2.221 0.026343 \*   
## month6 0.029496 0.171544 0.172 0.863483   
## month7 0.131258 0.173887 0.755 0.450340   
## month8 0.147515 0.175257 0.842 0.399955   
## month9 0.108912 0.171450 0.635 0.525270   
## month10 0.153206 0.163233 0.939 0.347950   
## month11 0.178794 0.155850 1.147 0.251291   
## day6 0.222331 0.238085 0.934 0.350392   
## day16 0.143839 0.245299 0.586 0.557619   
## day17 0.216315 0.244450 0.885 0.376207   
## day23 0.315971 0.245606 1.286 0.198271   
## day28 0.520705 0.239592 2.173 0.029758 \*   
## day2 0.070575 0.247437 0.285 0.775473   
## day8 0.274101 0.246776 1.111 0.266685   
## day9 0.451715 0.239157 1.889 0.058922 .   
## day14 0.017880 0.249240 0.072 0.942811   
## day15 0.297792 0.249018 1.196 0.231749   
## day18 -0.043163 0.253127 -0.171 0.864603   
## day30 0.326931 0.251673 1.299 0.193933   
## day3 0.234575 0.247311 0.949 0.342875   
## day7 -0.230432 0.254514 -0.905 0.365264   
## day27 -0.107290 0.249992 -0.429 0.667798   
## day12 -0.161730 0.242892 -0.666 0.505505   
## day25 0.289604 0.250174 1.158 0.247023   
## day26 0.086859 0.242274 0.359 0.719958   
## day1 0.195088 0.240709 0.810 0.417669   
## day21 -0.320378 0.252062 -1.271 0.203718   
## day13 0.237433 0.245913 0.966 0.334286   
## day24 0.406806 0.243218 1.673 0.094407 .   
## day4 -0.003483 0.245559 -0.014 0.988683   
## day10 -0.082904 0.250066 -0.332 0.740247   
## day19 0.317176 0.256820 1.235 0.216826   
## day29 0.053040 0.250366 0.212 0.832225   
## day11 0.363106 0.238924 1.520 0.128572   
## day31 0.414934 0.299742 1.384 0.166265   
## day22 0.120550 0.240815 0.501 0.616657   
## day20 0.143277 0.248159 0.577 0.563697   
## year2009 0.223081 0.260052 0.858 0.390986   
## year2010 0.418678 0.258473 1.620 0.105273   
## year2011 0.118675 0.260907 0.455 0.649214   
## year2012 0.352289 0.263546 1.337 0.181311   
## year2013 0.314141 0.261258 1.202 0.229202   
## year2014 0.222145 0.260149 0.854 0.393154   
## year2015 0.451868 0.259873 1.739 0.082069 .   
## year2016 0.146672 0.260893 0.562 0.573985   
## year2017 0.218860 0.287792 0.760 0.446967   
## year2007 1.544139 0.894137 1.727 0.084175 .   
## MinTemp 0.003456 0.017459 0.198 0.843105   
## MaxTemp -0.023644 0.028836 -0.820 0.412248   
## Rainfall 0.004293 0.003912 1.097 0.272526   
## WindGustDirWNW -0.073746 0.163359 -0.451 0.651675   
## WindGustDirENE -0.436745 0.205741 -2.123 0.033771 \*   
## WindGustDirNE -0.233502 0.205300 -1.137 0.255385   
## WindGustDirNW -0.353882 0.183557 -1.928 0.053866 .   
## WindGustDirNNE -0.240205 0.211398 -1.136 0.255844   
## WindGustDirNNW -0.167302 0.192960 -0.867 0.385926   
## WindGustDirE -0.284193 0.198412 -1.432 0.152047   
## WindGustDirN -0.425407 0.179744 -2.367 0.017946 \*   
## WindGustDirWSW -0.060607 0.158993 -0.381 0.703059   
## WindGustDirSE -0.073605 0.195998 -0.376 0.707259   
## WindGustDirESE -0.139209 0.205154 -0.679 0.497418   
## WindGustDirSW -0.046330 0.169242 -0.274 0.784274   
## WindGustDirS -0.163895 0.188262 -0.871 0.383992   
## WindGustDirSSE 0.167254 0.193102 0.866 0.386411   
## WindGustDirSSW -0.117109 0.183931 -0.637 0.524320   
## WindGustSpeed 0.055265 0.003734 14.802 < 2e-16 \*\*\*  
## WindDir9amW -0.231177 0.192846 -1.199 0.230620   
## WindDir9amSSW -0.418191 0.207850 -2.012 0.044222 \*   
## WindDir9amN 0.088257 0.180950 0.488 0.625732   
## WindDir9amWSW -0.304495 0.198708 -1.532 0.125431   
## WindDir9amSSE -0.537868 0.194577 -2.764 0.005705 \*\*   
## WindDir9amESE -0.277911 0.204362 -1.360 0.173864   
## WindDir9amSE -0.322854 0.198125 -1.630 0.103196   
## WindDir9amS -0.522305 0.201894 -2.587 0.009681 \*\*   
## WindDir9amNNE 0.640223 0.192321 3.329 0.000872 \*\*\*  
## WindDir9amNW -0.213759 0.197824 -1.081 0.279898   
## WindDir9amWNW -0.307007 0.198264 -1.548 0.121508   
## WindDir9amSW -0.204414 0.194418 -1.051 0.293067   
## WindDir9amNE 0.176491 0.195116 0.905 0.365707   
## WindDir9amE -0.188654 0.193997 -0.972 0.330822   
## WindDir9amNNW -0.325368 0.199315 -1.632 0.102589   
## WindDir3pmW -0.142809 0.184242 -0.775 0.438270   
## WindDir3pmWNW -0.071319 0.183131 -0.389 0.696949   
## WindDir3pmE -0.380466 0.211901 -1.795 0.072575 .   
## WindDir3pmSSW -0.133716 0.209743 -0.638 0.523785   
## WindDir3pmWSW -0.217800 0.191955 -1.135 0.256526   
## WindDir3pmS -0.308922 0.206330 -1.497 0.134334   
## WindDir3pmSE -0.102870 0.207548 -0.496 0.620147   
## WindDir3pmESE -0.314821 0.207356 -1.518 0.128948   
## WindDir3pmENE -0.359807 0.209168 -1.720 0.085399 .   
## WindDir3pmNE -0.287806 0.206752 -1.392 0.163911   
## WindDir3pmSSE -0.368693 0.210558 -1.751 0.079941 .   
## WindDir3pmSW -0.232500 0.200371 -1.160 0.245906   
## WindDir3pmN -0.228472 0.192155 -1.189 0.234440   
## WindDir3pmNNW 0.087225 0.186166 0.469 0.639403   
## WindDir3pmNNE -0.296067 0.205900 -1.438 0.150457   
## WindSpeed9am -0.013508 0.005005 -2.699 0.006959 \*\*   
## WindSpeed3pm -0.028280 0.005062 -5.587 2.31e-08 \*\*\*  
## Humidity9am 0.002427 0.003647 0.665 0.505874   
## Humidity3pm 0.059541 0.003864 15.409 < 2e-16 \*\*\*  
## Pressure9am 0.157988 0.021487 7.353 1.94e-13 \*\*\*  
## Pressure3pm -0.213970 0.021621 -9.896 < 2e-16 \*\*\*  
## Cloud9am0 -0.592418 0.301311 -1.966 0.049283 \*   
## Cloud9am8 0.133742 0.094163 1.420 0.155511   
## Cloud9am5 -0.145131 0.129356 -1.122 0.261885   
## Cloud9am2 -0.027712 0.137275 -0.202 0.840016   
## Cloud9am1 -0.436539 0.127842 -3.415 0.000639 \*\*\*  
## Cloud9am3 -0.109471 0.133745 -0.819 0.413069   
## Cloud9am4 -0.024075 0.147787 -0.163 0.870595   
## Cloud9am6 -0.149125 0.110445 -1.350 0.176946   
## Cloud3pm1 -1.425588 0.157993 -9.023 < 2e-16 \*\*\*  
## Cloud3pm6 -0.474960 0.117836 -4.031 5.56e-05 \*\*\*  
## Cloud3pm4 -0.523498 0.147633 -3.546 0.000391 \*\*\*  
## Cloud3pm5 -0.600530 0.138265 -4.343 1.40e-05 \*\*\*  
## Cloud3pm3 -0.808534 0.144162 -5.609 2.04e-08 \*\*\*  
## Cloud3pm7 -0.189207 0.095137 -1.989 0.046726 \*   
## Cloud3pm2 -1.067580 0.159817 -6.680 2.39e-11 \*\*\*  
## Cloud3pm0 -3.858065 1.023553 -3.769 0.000164 \*\*\*  
## Temp9am 0.043682 0.026809 1.629 0.103233   
## Temp3pm -0.005120 0.031792 -0.161 0.872053   
## RainTodayYes 0.527834 0.081223 6.499 8.11e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11020.6 on 9931 degrees of freedom  
## Residual deviance: 6869.4 on 9806 degrees of freedom  
## AIC: 7121.4  
##   
## Number of Fisher Scoring iterations: 8

emptymod = glm(RainTomorrow ~1, train\_cloud, family = "binomial")   
summary(emptymod)

##   
## Call:  
## glm(formula = RainTomorrow ~ 1, family = "binomial", data = train\_cloud)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7466 -0.7466 -0.7466 -0.7466 1.6815   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.13492 0.02339 -48.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11021 on 9931 degrees of freedom  
## Residual deviance: 11021 on 9931 degrees of freedom  
## AIC: 11023  
##   
## Number of Fisher Scoring iterations: 4

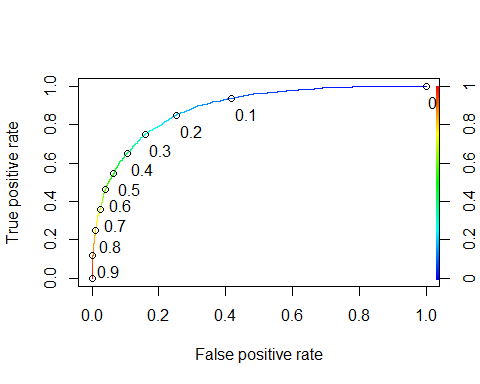
forwardmod = stepAIC(emptymod, direction = "forward", scope=list(upper=allmod,lower=emptymod), trace = FALSE)  
summary(forwardmod)

##   
## Call:  
## glm(formula = RainTomorrow ~ Humidity3pm + Pressure3pm + Cloud3pm +   
## WindGustSpeed + Pressure9am + RainToday + WindSpeed3pm +   
## WindDir9am + Cloud9am + WindSpeed9am + Temp9am + MaxTemp,   
## family = "binomial", data = train\_cloud)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0821 -0.5588 -0.2726 -0.0249 3.2807   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 48.846971 5.563104 8.781 < 2e-16 \*\*\*  
## Humidity3pm 0.058359 0.002592 22.518 < 2e-16 \*\*\*  
## Pressure3pm -0.207965 0.018176 -11.442 < 2e-16 \*\*\*  
## Cloud3pm1 -1.444595 0.154042 -9.378 < 2e-16 \*\*\*  
## Cloud3pm6 -0.480903 0.114097 -4.215 2.50e-05 \*\*\*  
## Cloud3pm4 -0.542358 0.143429 -3.781 0.000156 \*\*\*  
## Cloud3pm5 -0.609599 0.134407 -4.535 5.75e-06 \*\*\*  
## Cloud3pm3 -0.835107 0.140799 -5.931 3.01e-09 \*\*\*  
## Cloud3pm7 -0.203793 0.092659 -2.199 0.027850 \*   
## Cloud3pm2 -1.087368 0.156118 -6.965 3.28e-12 \*\*\*  
## Cloud3pm0 -3.923252 1.023783 -3.832 0.000127 \*\*\*  
## WindGustSpeed 0.055304 0.003546 15.598 < 2e-16 \*\*\*  
## Pressure9am 0.154069 0.017704 8.702 < 2e-16 \*\*\*  
## RainTodayYes 0.604183 0.068063 8.877 < 2e-16 \*\*\*  
## WindSpeed3pm -0.027812 0.004850 -5.734 9.79e-09 \*\*\*  
## WindDir9amW -0.037094 0.181897 -0.204 0.838409   
## WindDir9amSSW -0.291083 0.196106 -1.484 0.137726   
## WindDir9amN 0.133887 0.169351 0.791 0.429184   
## WindDir9amWSW -0.155879 0.188597 -0.827 0.408509   
## WindDir9amSSE -0.344242 0.180708 -1.905 0.056785 .   
## WindDir9amESE -0.271664 0.197452 -1.376 0.168870   
## WindDir9amSE -0.210223 0.188706 -1.114 0.265269   
## WindDir9amS -0.380814 0.188443 -2.021 0.043296 \*   
## WindDir9amNNE 0.689093 0.183735 3.750 0.000177 \*\*\*  
## WindDir9amNW -0.024703 0.187197 -0.132 0.895013   
## WindDir9amWNW -0.114601 0.186553 -0.614 0.539014   
## WindDir9amSW -0.103803 0.184304 -0.563 0.573289   
## WindDir9amNE 0.193294 0.190442 1.015 0.310117   
## WindDir9amE -0.222479 0.189543 -1.174 0.240488   
## WindDir9amNNW -0.196556 0.186532 -1.054 0.292001   
## Cloud9am0 -0.608403 0.293690 -2.072 0.038304 \*   
## Cloud9am8 0.128451 0.091041 1.411 0.158273   
## Cloud9am5 -0.138861 0.126165 -1.101 0.271055   
## Cloud9am2 -0.022383 0.133030 -0.168 0.866386   
## Cloud9am1 -0.441460 0.123043 -3.588 0.000333 \*\*\*  
## Cloud9am3 -0.099239 0.129748 -0.765 0.444359   
## Cloud9am4 -0.047417 0.144550 -0.328 0.742888   
## Cloud9am6 -0.159792 0.107762 -1.483 0.138121   
## WindSpeed9am -0.014970 0.004721 -3.171 0.001520 \*\*   
## Temp9am 0.035081 0.015925 2.203 0.027605 \*   
## MaxTemp -0.027989 0.016635 -1.683 0.092459 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11020.6 on 9931 degrees of freedom  
## Residual deviance: 6983.4 on 9891 degrees of freedom  
## AIC: 7065.4  
##   
## Number of Fisher Scoring iterations: 8

## Decent AIC score of 7065.4, only 1 Temp measurement kept but it is not significant to RainTomorrow, Wind from the NNE and Pressure at 9am have a possitive influance to RainTomorrow, and Humidity at 3pm also has a possitive influence.

predictions = predict(forwardmod, type="response")  
ROCRpred = prediction(predictions, train\_cloud$RainTomorrow)

ROCRpred = prediction(predictions, train\_cloud$RainTomorrow)   
  
###You shouldn't need to ever change the next two lines:  
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



opt.cut = function(perf, pred){   
 cut.ind = mapply(FUN=function(x, y, p){   
 d = (x - 0)^2 + (y-1)^2   
 ind = which(d == min(d))   
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])   
 }, perf@x.values, perf@y.values, pred@cutoffs)   
}   
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.8120861  
## specificity 0.7860564  
## cutoff 0.2343838

Confusion matrix

table(train\_cloud$RainTomorrow, predictions > 0.2343838)

##   
## FALSE TRUE  
## No 5908 1608  
## Yes 454 1962

Accuracy

(5908+1962)/nrow(train\_cloud)

## [1] 0.7923882

Sensitivity

(1962)/(1962 + 454)

## [1] 0.8120861

Speciﬁcity

(5908)/(5908+1608)

## [1] 0.7860564

t1 = table(train\_cloud$RainTomorrow,predictions > 0.3)   
(t1[1,1]+t1[2,2])/nrow(train\_cloud)

## [1] 0.818969

t1 = table(train\_cloud$RainTomorrow,predictions > 0.4)   
(t1[1,1]+t1[2,2])/nrow(train\_cloud)

## [1] 0.8361861

t1 = table(train\_cloud$RainTomorrow,predictions > 0.5)   
(t1[1,1]+t1[2,2])/nrow(train\_cloud)

## [1] 0.8408176

t1 = table(train\_cloud$RainTomorrow,predictions > 0.6)   
(t1[1,1]+t1[2,2])/nrow(train\_cloud)

## [1] 0.8390052

predictions\_test = predict(forwardmod, newdata = test\_cloud, type="response")   
table(test\_cloud$RainTomorrow,predictions\_test > 0.5)

##   
## FALSE TRUE  
## No 3001 220  
## Yes 459 576

Accuracy for Test

(3001+576)/nrow(test\_cloud)

## [1] 0.8404605

Sensitivity for Test

(576)/(576+459)

## [1] 0.5565217

Speciﬁcity for Test

(3001)/(3001+220)

## [1] 0.9316982

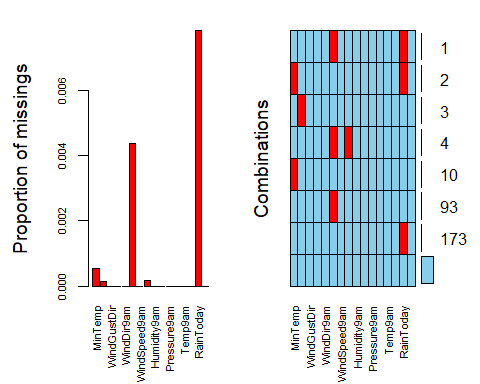
**This model produced some very good numbers with the test\_cloud dataset with .556 Sensitivity and an accuracy score of .8404.It is also very important to note that none of the new date values or Rainfall were kept or showed any influence or significance to RainTomorrow. I will drop the Date and Rainfall values from this point going forward.**

## Model for removing Cloud9am and Cloud3pm

rain\_no\_cloud = rain %>% dplyr::select(-Cloud9am,-Cloud3pm, -Rainfall, -Date)   
#summary(rain\_no\_cloud)  
#str(rain\_no\_cloud)

rain\_no\_cloud = rain\_no\_cloud %>% drop\_na(Pressure3pm) %>%  
 drop\_na(Pressure9am) %>%  
 #drop\_na(WindDir3pm) %>%  
 drop\_na(WindDir9am) %>%  
 drop\_na(Humidity3pm) %>%  
 drop\_na(Humidity9am) %>%  
 drop\_na(WindGustDir)   
   
#str(rain\_no\_cloud)  
#summary(rain\_no\_cloud)

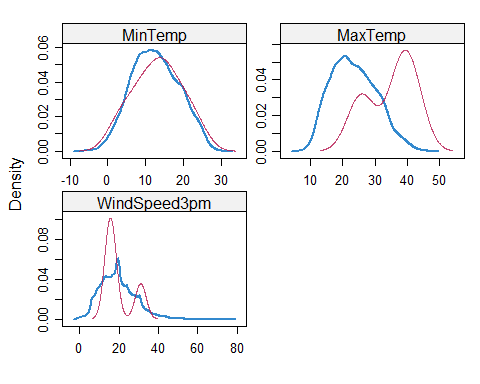
vim\_plot = aggr(rain\_no\_cloud, numbers = TRUE, prop = c(TRUE, FALSE), cex.axis=.7)



rain\_imp2 = mice(rain\_no\_cloud, m=1, method = "pmm", seed = 12345)

##   
## iter imp variable  
## 1 1 MinTemp MaxTemp WindDir3pm WindSpeed3pm RainToday  
## 2 1 MinTemp MaxTemp WindDir3pm WindSpeed3pm RainToday  
## 3 1 MinTemp MaxTemp WindDir3pm WindSpeed3pm RainToday  
## 4 1 MinTemp MaxTemp WindDir3pm WindSpeed3pm RainToday  
## 5 1 MinTemp MaxTemp WindDir3pm WindSpeed3pm RainToday

densityplot(rain\_imp2)



rain\_complete\_nr = complete(rain\_imp2)

set.seed(1234)   
train.rows = createDataPartition(y = rain\_complete\_nr$RainTomorrow, p=0.7, list = FALSE)   
train2 = rain\_complete\_nr[train.rows,]   
test2 = rain\_complete\_nr[-train.rows,]

allmod2 = glm(RainTomorrow ~., train2, family = "binomial")   
summary(allmod2)

##   
## Call:  
## glm(formula = RainTomorrow ~ ., family = "binomial", data = train2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8114 -0.5495 -0.3052 -0.1223 3.0437   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 50.956247 4.544121 11.214 < 2e-16 \*\*\*  
## MinTemp 0.043206 0.012703 3.401 0.000671 \*\*\*  
## MaxTemp -0.014210 0.021275 -0.668 0.504191   
## WindGustDirWNW -0.047707 0.128966 -0.370 0.711444   
## WindGustDirENE -0.093676 0.160704 -0.583 0.559955   
## WindGustDirNE -0.336884 0.162253 -2.076 0.037867 \*   
## WindGustDirNW -0.012080 0.136703 -0.088 0.929584   
## WindGustDirNNE -0.315690 0.163134 -1.935 0.052971 .   
## WindGustDirNNW -0.092835 0.151275 -0.614 0.539426   
## WindGustDirE -0.046146 0.158940 -0.290 0.771560   
## WindGustDirN -0.118304 0.144099 -0.821 0.411653   
## WindGustDirWSW -0.020876 0.127790 -0.163 0.870236   
## WindGustDirSE 0.056064 0.152422 0.368 0.713005   
## WindGustDirESE 0.151181 0.161024 0.939 0.347795   
## WindGustDirSW -0.042593 0.137715 -0.309 0.757104   
## WindGustDirS -0.054066 0.145271 -0.372 0.709761   
## WindGustDirSSE 0.080576 0.151723 0.531 0.595368   
## WindGustDirSSW -0.200172 0.143512 -1.395 0.163072   
## WindGustSpeed 0.060006 0.002881 20.827 < 2e-16 \*\*\*  
## WindDir9amW -0.074397 0.150748 -0.494 0.621646   
## WindDir9amSSW -0.183399 0.154137 -1.190 0.234109   
## WindDir9amN 0.139635 0.139512 1.001 0.316886   
## WindDir9amWSW 0.012372 0.153279 0.081 0.935671   
## WindDir9amSSE -0.374116 0.150103 -2.492 0.012688 \*   
## WindDir9amESE -0.235396 0.161068 -1.461 0.143885   
## WindDir9amSE -0.245288 0.153877 -1.594 0.110925   
## WindDir9amS -0.312767 0.150836 -2.074 0.038121 \*   
## WindDir9amNNE 0.436203 0.146714 2.973 0.002948 \*\*   
## WindDir9amNW -0.190205 0.150154 -1.267 0.205250   
## WindDir9amWNW -0.270034 0.154710 -1.745 0.080912 .   
## WindDir9amSW -0.049433 0.148845 -0.332 0.739806   
## WindDir9amNE 0.148263 0.155656 0.953 0.340843   
## WindDir9amE -0.205075 0.155017 -1.323 0.185863   
## WindDir9amNNW -0.256632 0.149368 -1.718 0.085774 .   
## WindDir3pmW -0.011192 0.141157 -0.079 0.936805   
## WindDir3pmWNW -0.188520 0.138795 -1.358 0.174381   
## WindDir3pmE -0.360266 0.164511 -2.190 0.028530 \*   
## WindDir3pmSSW -0.290295 0.161435 -1.798 0.072142 .   
## WindDir3pmWSW -0.203325 0.149218 -1.363 0.173009   
## WindDir3pmS -0.173236 0.154430 -1.122 0.261957   
## WindDir3pmSE -0.372466 0.158264 -2.353 0.018600 \*   
## WindDir3pmESE -0.357268 0.160006 -2.233 0.025559 \*   
## WindDir3pmENE -0.317924 0.160068 -1.986 0.047014 \*   
## WindDir3pmNE -0.574321 0.159634 -3.598 0.000321 \*\*\*  
## WindDir3pmSSE -0.474678 0.157904 -3.006 0.002646 \*\*   
## WindDir3pmSW -0.196320 0.154377 -1.272 0.203483   
## WindDir3pmN -0.193251 0.148587 -1.301 0.193396   
## WindDir3pmNNW 0.001320 0.141996 0.009 0.992581   
## WindDir3pmNNE -0.514290 0.160779 -3.199 0.001380 \*\*   
## WindSpeed9am -0.008077 0.003882 -2.081 0.037465 \*   
## WindSpeed3pm -0.042050 0.003923 -10.718 < 2e-16 \*\*\*  
## Humidity9am 0.008110 0.002794 2.903 0.003696 \*\*   
## Humidity3pm 0.062542 0.002962 21.118 < 2e-16 \*\*\*  
## Pressure9am 0.147613 0.015715 9.393 < 2e-16 \*\*\*  
## Pressure3pm -0.204432 0.015983 -12.790 < 2e-16 \*\*\*  
## Temp9am 0.017988 0.019749 0.911 0.362374   
## Temp3pm -0.043335 0.023944 -1.810 0.070315 .   
## RainTodayYes 0.585321 0.057940 10.102 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16798 on 15750 degrees of freedom  
## Residual deviance: 11109 on 15693 degrees of freedom  
## AIC: 11225  
##   
## Number of Fisher Scoring iterations: 6

emptymod2 = glm(RainTomorrow ~1, train2, family = "binomial")   
summary(emptymod2)

##   
## Call:  
## glm(formula = RainTomorrow ~ 1, family = "binomial", data = train2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7141 -0.7141 -0.7141 -0.7141 1.7271   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.23639 0.01908 -64.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16798 on 15750 degrees of freedom  
## Residual deviance: 16798 on 15750 degrees of freedom  
## AIC: 16800  
##   
## Number of Fisher Scoring iterations: 4

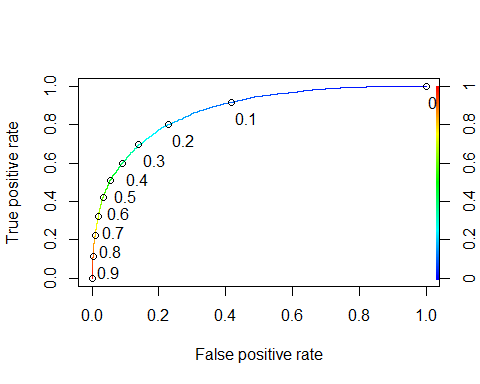
forwardmod2 = stepAIC(emptymod2, direction = "forward", scope=list(upper=allmod2,lower=emptymod2), trace = FALSE)  
summary(forwardmod2)

##   
## Call:  
## glm(formula = RainTomorrow ~ Humidity3pm + WindGustSpeed + Pressure3pm +   
## WindSpeed3pm + RainToday + Pressure9am + WindDir3pm + WindDir9am +   
## Humidity9am + MinTemp + Temp3pm + WindSpeed9am, family = "binomial",   
## data = train2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8058 -0.5493 -0.3061 -0.1242 2.9821   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 50.659784 4.466591 11.342 < 2e-16 \*\*\*  
## Humidity3pm 0.063204 0.002444 25.865 < 2e-16 \*\*\*  
## WindGustSpeed 0.060076 0.002818 21.318 < 2e-16 \*\*\*  
## Pressure3pm -0.201363 0.015779 -12.762 < 2e-16 \*\*\*  
## WindSpeed3pm -0.041954 0.003823 -10.973 < 2e-16 \*\*\*  
## RainTodayYes 0.586405 0.057614 10.178 < 2e-16 \*\*\*  
## Pressure9am 0.144761 0.015463 9.362 < 2e-16 \*\*\*  
## WindDir3pmW 0.011077 0.134076 0.083 0.934157   
## WindDir3pmWNW -0.168683 0.136358 -1.237 0.216063   
## WindDir3pmE -0.343975 0.152695 -2.253 0.024278 \*   
## WindDir3pmSSW -0.338617 0.149450 -2.266 0.023466 \*   
## WindDir3pmWSW -0.186931 0.140896 -1.327 0.184598   
## WindDir3pmS -0.181316 0.140939 -1.286 0.198274   
## WindDir3pmSE -0.304881 0.145295 -2.098 0.035873 \*   
## WindDir3pmESE -0.298146 0.148273 -2.011 0.044348 \*   
## WindDir3pmENE -0.372268 0.147729 -2.520 0.011738 \*   
## WindDir3pmNE -0.704726 0.146724 -4.803 1.56e-06 \*\*\*  
## WindDir3pmSSE -0.436267 0.145417 -3.000 0.002699 \*\*   
## WindDir3pmSW -0.207324 0.143994 -1.440 0.149922   
## WindDir3pmN -0.252907 0.140240 -1.803 0.071327 .   
## WindDir3pmNNW -0.019109 0.139208 -0.137 0.890819   
## WindDir3pmNNE -0.628982 0.151159 -4.161 3.17e-05 \*\*\*  
## WindDir9amW -0.043151 0.147904 -0.292 0.770478   
## WindDir9amSSW -0.171754 0.151312 -1.135 0.256334   
## WindDir9amN 0.141311 0.136507 1.035 0.300580   
## WindDir9amWSW 0.029055 0.150690 0.193 0.847103   
## WindDir9amSSE -0.311421 0.147413 -2.113 0.034637 \*   
## WindDir9amESE -0.177688 0.159576 -1.113 0.265494   
## WindDir9amSE -0.174541 0.151315 -1.153 0.248706   
## WindDir9amS -0.269116 0.147693 -1.822 0.068435 .   
## WindDir9amNNE 0.412436 0.145233 2.840 0.004514 \*\*   
## WindDir9amNW -0.167906 0.147574 -1.138 0.255214   
## WindDir9amWNW -0.247237 0.151992 -1.627 0.103813   
## WindDir9amSW -0.028842 0.146448 -0.197 0.843873   
## WindDir9amNE 0.129287 0.155005 0.834 0.404237   
## WindDir9amE -0.179501 0.154413 -1.162 0.245044   
## WindDir9amNNW -0.238742 0.145863 -1.637 0.101683   
## Humidity9am 0.007244 0.002097 3.455 0.000551 \*\*\*  
## MinTemp 0.049484 0.009724 5.089 3.60e-07 \*\*\*  
## Temp3pm -0.046467 0.010551 -4.404 1.06e-05 \*\*\*  
## WindSpeed9am -0.008083 0.003788 -2.134 0.032852 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16798 on 15750 degrees of freedom  
## Residual deviance: 11126 on 15710 degrees of freedom  
## AIC: 11208  
##   
## Number of Fisher Scoring iterations: 6

**Really bad 11208 AIC score and, once again, only one Tempurature Factor used, and also two NNE wind directions show significance to RainTomorrow.**

predictions = predict(forwardmod2, type="response")  
ROCRpred = prediction(predictions, train2$RainTomorrow)

###You shouldn't need to ever change the next two lines:  
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



opt.cut = function(perf, pred){   
 cut.ind = mapply(FUN=function(x, y, p){   
 d = (x - 0)^2 + (y-1)^2   
 ind = which(d == min(d))   
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])   
 }, perf@x.values, perf@y.values, pred@cutoffs)   
}   
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.7858956  
## specificity 0.7912502  
## cutoff 0.2182195

Confusion matrix

table(train2$RainTomorrow, predictions > 0.2182195)

##   
## FALSE TRUE  
## No 9658 2548  
## Yes 760 2785

Accuracy

(9658+2785)/nrow(train2)

## [1] 0.7899816

Sensitivity

(2785)/(2785+760)

## [1] 0.7856135

Speciﬁcity

(9658)/(9658+2548)

## [1] 0.7912502

t1 = table(train2$RainTomorrow,predictions > 0.3)   
(t1[1,1]+t1[2,2])/nrow(train2)

## [1] 0.8229954

t1 = table(train2$RainTomorrow,predictions > 0.4)   
(t1[1,1]+t1[2,2])/nrow(train2)

## [1] 0.8393753

t1 = table(train2$RainTomorrow,predictions > 0.5)   
(t1[1,1]+t1[2,2])/nrow(train2)

## [1] 0.8481366

predictions\_test = predict(forwardmod2, newdata = test2, type="response")   
table(test2$RainTomorrow,predictions\_test > 0.5)

##   
## FALSE TRUE  
## No 4932 299  
## Yes 768 750

Accuracy

(4932+750)/nrow(test2)

## [1] 0.8419025

Sensitivity for Test

(750)/(750+768)

## [1] 0.4940711

Speciﬁcity for Test

(4932)/(4932+299)

## [1] 0.9428408

**While having a similar accuracy score as the last model with .8419, there was close to a 5% drop in its sensitivity score.**

### Best Guess while cleaning Cloud Coverage, Humidity, and Pressure’s NA’s while dropping Date and Rainfall

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.

rain3 = rain %>% dplyr::select(-Date, -Rainfall)

rain3 = rain3 %>%   
 mutate(Cloud9am = as\_factor(as.character(Cloud9am))) %>%  
 mutate(Cloud3pm = as\_factor(as.character(Cloud3pm))) %>%  
 mutate(WindGustDir = as\_factor(as.character(WindGustDir))) %>%  
 mutate(WindDir9am = as\_factor(as.character(WindDir9am))) %>%  
 mutate(WindDir3pm = as\_factor(as.character(WindDir3pm))) %>%  
 mutate(RainToday = as\_factor(as.character(RainToday))) %>%  
 mutate(RainTomorrow = as\_factor(as.character(RainTomorrow)))   
str(rain3)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 18 variables:  
## $ 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 ...  
## $ WindGustDir : Factor w/ 16 levels "W","WNW","ENE",..: 1 2 2 3 2 2 1 1 4 5 ...  
## $ WindGustSpeed: num 41 56 50 22 41 48 41 48 37 39 ...  
## $ WindDir9am : Factor w/ 16 levels "ENE","W","SSW",..: 1 2 NA 3 2 4 5 2 6 6 ...  
## $ WindDir3pm : Factor w/ 16 levels "NW","W","WNW",..: 1 2 3 4 2 3 5 6 7 5 ...  
## $ 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 : Factor w/ 9 levels "7","0","8","5",..: 1 NA 2 3 NA NA NA NA NA NA ...  
## $ Cloud3pm : Factor w/ 9 levels "8","1","6","4",..: 1 NA NA 2 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 ...

summary(rain3)

## MinTemp MaxTemp WindGustDir WindGustSpeed   
## Min. :-8.50 Min. :-3.00 W : 1913 Min. : 7.00   
## 1st Qu.: 7.60 1st Qu.:17.90 SE : 1863 1st Qu.: 31.00   
## Median :12.00 Median :22.60 E : 1810 Median : 39.00   
## Mean :12.16 Mean :23.18 S : 1779 Mean : 40.02   
## 3rd Qu.:16.80 3rd Qu.:28.20 N : 1778 3rd Qu.: 48.00   
## Max. :30.50 Max. :47.00 (Other):17009 Max. :135.00   
## NA's :132 NA's :64 NA's : 1851 NA's :1840   
## WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm   
## N : 2275 SE : 2036 Min. : 0.00 Min. : 0.00   
## SSE : 1781 W : 1924 1st Qu.: 7.00 1st Qu.:13.00   
## SE : 1764 S : 1877 Median :13.00 Median :19.00   
## E : 1739 SSE : 1862 Mean :13.97 Mean :18.59   
## NW : 1711 WSW : 1850 3rd Qu.:19.00 3rd Qu.:24.00   
## (Other):16756 (Other):17669 Max. :87.00 Max. :83.00   
## NA's : 1977 NA's : 785 NA's :308 NA's :526   
## Humidity9am Humidity3pm Pressure9am Pressure3pm   
## Min. : 1.00 Min. : 0.00 Min. : 980.5 Min. : 978.2   
## 1st Qu.: 57.00 1st Qu.: 36.00 1st Qu.:1013.0 1st Qu.:1010.5   
## Median : 70.00 Median : 52.00 Median :1017.7 Median :1015.3   
## Mean : 68.86 Mean : 51.54 Mean :1017.7 Mean :1015.3   
## 3rd Qu.: 83.00 3rd Qu.: 66.00 3rd Qu.:1022.4 3rd Qu.:1020.0   
## Max. :100.00 Max. :100.00 Max. :1041.0 Max. :1037.0   
## NA's :366 NA's :694 NA's :2837 NA's :2817   
## Cloud9am Cloud3pm Temp9am Temp3pm   
## 7 : 3921 7 : 3564 Min. :-5.60 Min. :-4.20   
## 1 : 2986 1 : 2868 1st Qu.:12.30 1st Qu.:16.60   
## 8 : 2828 8 : 2426 Median :16.70 Median :21.10   
## 0 : 1672 6 : 1777 Mean :16.96 Mean :21.63   
## 6 : 1543 2 : 1410 3rd Qu.:21.50 3rd Qu.:26.40   
## (Other): 4380 (Other): 4617 Max. :38.60 Max. :45.20   
## NA's :10673 NA's :11341 NA's :196 NA's :532   
## RainToday RainTomorrow  
## No :21525 No :21713   
## Yes : 6183 Yes: 6290   
## NA's: 295   
##   
##   
##   
##

rain\_Cloud = rain3 %>% drop\_na(Cloud9am) %>%  
 drop\_na(Cloud3pm) %>%  
 drop\_na(Humidity9am) %>%  
 drop\_na(Humidity3pm) %>%  
 drop\_na(Pressure9am) %>%  
 drop\_na(Pressure3pm)   
summary(rain\_Cloud)

## MinTemp MaxTemp WindGustDir WindGustSpeed   
## Min. :-5.8 Min. : 6.60 E :1132 Min. : 11.00   
## 1st Qu.: 8.2 1st Qu.:18.00 SW :1060 1st Qu.: 31.00   
## Median :12.9 Median :23.10 W :1059 Median : 39.00   
## Mean :13.1 Mean :23.59 WSW :1015 Mean : 40.74   
## 3rd Qu.:18.0 3rd Qu.:29.00 N : 992 3rd Qu.: 48.00   
## Max. :30.5 Max. :46.30 (Other):9128 Max. :120.00   
## NA's :14 NA's :3 NA's : 818 NA's :815   
## WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm   
## N :1278 SW :1125 Min. : 0.00 Min. : 0.0   
## E :1088 SE :1092 1st Qu.: 9.00 1st Qu.:13.0   
## SSE :1056 WSW :1090 Median :15.00 Median :19.0   
## W :1002 S :1088 Mean :15.19 Mean :19.5   
## SE : 983 W :1056 3rd Qu.:20.00 3rd Qu.:24.0   
## (Other):9342 (Other):9659 Max. :69.00 Max. :76.0   
## NA's : 455 NA's : 94 NA's :32 NA's :23   
## Humidity9am Humidity3pm Pressure9am Pressure3pm   
## Min. : 2.00 Min. : 0.00 Min. : 980.5 Min. : 978.2   
## 1st Qu.: 56.00 1st Qu.: 37.00 1st Qu.:1012.7 1st Qu.:1010.2   
## Median : 69.00 Median : 52.00 Median :1017.4 Median :1015.0   
## Mean : 67.71 Mean : 51.49 Mean :1017.4 Mean :1015.0   
## 3rd Qu.: 82.00 3rd Qu.: 66.00 3rd Qu.:1022.1 3rd Qu.:1019.8   
## Max. :100.00 Max. :100.00 Max. :1039.6 Max. :1037.0   
##   
## Cloud9am Cloud3pm Temp9am Temp3pm   
## 7 :3609 7 :3381 Min. :-0.90 Min. : 5.00   
## 1 :2694 1 :2560 1st Qu.:12.60 1st Qu.:16.70   
## 8 :2305 8 :2082 Median :17.30 Median :21.60   
## 0 :1389 6 :1642 Mean :17.66 Mean :22.08   
## 6 :1365 2 :1282 3rd Qu.:22.60 3rd Qu.:27.20   
## 2 :1083 3 :1259 Max. :36.40 Max. :44.80   
## (Other):2759 (Other):2998   
## RainToday RainTomorrow  
## No :11485 No :11502   
## Yes : 3666 Yes: 3702   
## NA's: 53   
##   
##   
##   
##

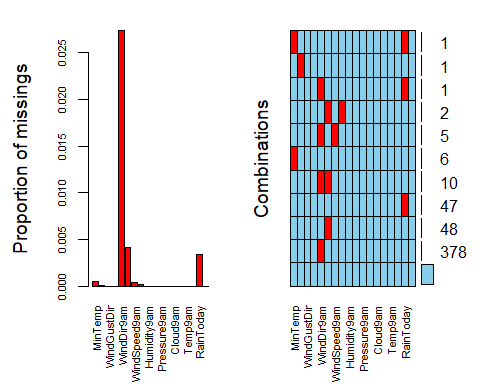
str(rain\_Cloud)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 15204 obs. of 18 variables:  
## $ MinTemp : num 17.5 14.1 15.4 6.6 7.2 9.3 8.8 7.6 5.7 -1 ...  
## $ MaxTemp : num 32.3 20.9 22.3 21.6 19.2 20.3 11.6 12 13.2 12 ...  
## $ WindGustDir : Factor w/ 16 levels "W","WNW","ENE",..: 1 3 3 2 13 4 2 1 2 9 ...  
## $ WindGustSpeed: num 41 22 30 26 22 31 54 61 39 20 ...  
## $ WindDir9am : Factor w/ 16 levels "ENE","W","SSW",..: 1 3 6 1 1 8 11 2 12 NA ...  
## $ WindDir3pm : Factor w/ 16 levels "NW","W","WNW",..: 1 4 12 2 2 11 3 6 3 14 ...  
## $ WindSpeed9am : num 7 11 11 6 7 9 19 17 15 0 ...  
## $ WindSpeed3pm : num 20 9 9 11 6 22 19 22 24 11 ...  
## $ Humidity9am : num 82 69 69 59 81 77 88 82 91 98 ...  
## $ Humidity3pm : num 33 82 78 39 49 50 74 62 76 64 ...  
## $ Pressure9am : num 1011 1012 1017 1018 1027 ...  
## $ Pressure3pm : num 1006 1010 1019 1017 1024 ...  
## $ Cloud9am : Factor w/ 9 levels "7","0","8","5",..: 1 3 3 4 3 5 3 3 3 1 ...  
## $ Cloud3pm : Factor w/ 9 levels "8","1","6","4",..: 1 2 1 2 4 1 4 2 1 5 ...  
## $ Temp9am : num 17.8 17.2 18.3 12.6 10.1 11.6 8.8 7.8 7.9 3.2 ...  
## $ Temp3pm : num 29.7 18.1 19.6 21.6 19.1 19.5 11.4 11.7 12.6 11.2 ...  
## $ RainToday : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 2 1 1 ...  
## $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 2 2 1 1 1 ...

rain\_Cloud = rain\_Cloud %>% drop\_na(WindGustDir) %>%  
 drop\_na(WindGustSpeed)# %>%  
 #drop\_na(WindDir9am)  
   
summary(rain\_Cloud)

## MinTemp MaxTemp WindGustDir WindGustSpeed   
## Min. :-5.8 Min. : 6.60 E :1132 Min. : 11.00   
## 1st Qu.: 8.1 1st Qu.:18.00 SW :1060 1st Qu.: 31.00   
## Median :12.8 Median :23.30 W :1059 Median : 39.00   
## Mean :13.1 Mean :23.72 WSW :1015 Mean : 40.74   
## 3rd Qu.:18.1 3rd Qu.:29.20 N : 992 3rd Qu.: 48.00   
## Max. :30.5 Max. :46.30 SSW : 991 Max. :120.00   
## NA's :7 NA's :1 (Other):8137   
## WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm   
## N :1225 SW :1061 Min. : 0.00 Min. : 0.0   
## E :1035 S :1035 1st Qu.: 9.00 1st Qu.:13.0   
## SSE :1031 SE :1024 Median :15.00 Median :19.0   
## SE : 946 WSW :1015 Mean :15.28 Mean :19.5   
## ENE : 927 W :1007 3rd Qu.:20.00 3rd Qu.:24.0   
## (Other):8828 (Other):9184 Max. :69.00 Max. :76.0   
## NA's : 394 NA's : 60 NA's :5 NA's :2   
## Humidity9am Humidity3pm Pressure9am Pressure3pm   
## Min. : 2.00 Min. : 0.00 Min. : 980.5 Min. : 978.2   
## 1st Qu.: 56.00 1st Qu.: 36.00 1st Qu.:1012.7 1st Qu.:1010.2   
## Median : 69.00 Median : 52.00 Median :1017.3 Median :1014.9   
## Mean : 67.51 Mean : 50.91 Mean :1017.3 Mean :1014.9   
## 3rd Qu.: 82.00 3rd Qu.: 65.00 3rd Qu.:1022.0 3rd Qu.:1019.7   
## Max. :100.00 Max. :100.00 Max. :1039.6 Max. :1037.0   
##   
## Cloud9am Cloud3pm Temp9am Temp3pm   
## 7 :3414 7 :3219 Min. :-0.90 Min. : 5.00   
## 1 :2562 1 :2433 1st Qu.:12.60 1st Qu.:16.70   
## 8 :2166 8 :1956 Median :17.30 Median :21.80   
## 0 :1340 6 :1528 Mean :17.71 Mean :22.22   
## 6 :1268 2 :1230 3rd Qu.:22.70 3rd Qu.:27.50   
## 2 :1038 3 :1192 Max. :36.40 Max. :44.80   
## (Other):2598 (Other):2828   
## RainToday RainTomorrow  
## No :10914 No :10918   
## Yes : 3423 Yes: 3468   
## NA's: 49   
##   
##   
##   
##

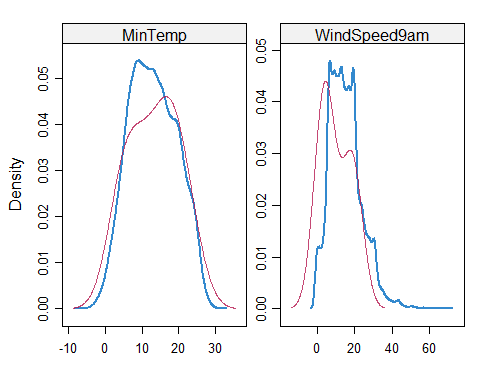
vim\_plot = aggr(rain\_Cloud, numbers = TRUE, prop = c(TRUE, FALSE), cex.axis=.7)



rain\_imp3 = mice(rain\_Cloud, m=1, method = "pmm", seed = 12345)

##   
## iter imp variable  
## 1 1 MinTemp MaxTemp WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm RainToday  
## 2 1 MinTemp MaxTemp WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm RainToday  
## 3 1 MinTemp MaxTemp WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm RainToday  
## 4 1 MinTemp MaxTemp WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm RainToday  
## 5 1 MinTemp MaxTemp WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm RainToday

densityplot(rain\_imp3)



rain\_complete3 = complete(rain\_imp3)

str(rain\_complete3)

## 'data.frame': 14386 obs. of 18 variables:  
## $ MinTemp : num 17.5 14.1 15.4 6.6 7.2 9.3 8.8 7.6 5.7 -1 ...  
## $ MaxTemp : num 32.3 20.9 22.3 21.6 19.2 20.3 11.6 12 13.2 12 ...  
## $ WindGustDir : Factor w/ 16 levels "W","WNW","ENE",..: 1 3 3 2 13 4 2 1 2 9 ...  
## $ WindGustSpeed: num 41 22 30 26 22 31 54 61 39 20 ...  
## $ WindDir9am : Factor w/ 16 levels "ENE","W","SSW",..: 1 3 6 1 1 8 11 2 12 7 ...  
## $ WindDir3pm : Factor w/ 16 levels "NW","W","WNW",..: 1 4 12 2 2 11 3 6 3 14 ...  
## $ WindSpeed9am : num 7 11 11 6 7 9 19 17 15 0 ...  
## $ WindSpeed3pm : num 20 9 9 11 6 22 19 22 24 11 ...  
## $ Humidity9am : num 82 69 69 59 81 77 88 82 91 98 ...  
## $ Humidity3pm : num 33 82 78 39 49 50 74 62 76 64 ...  
## $ Pressure9am : num 1011 1012 1017 1018 1027 ...  
## $ Pressure3pm : num 1006 1010 1019 1017 1024 ...  
## $ Cloud9am : Factor w/ 9 levels "7","0","8","5",..: 1 3 3 4 3 5 3 3 3 1 ...  
## $ Cloud3pm : Factor w/ 9 levels "8","1","6","4",..: 1 2 1 2 4 1 4 2 1 5 ...  
## $ Temp9am : num 17.8 17.2 18.3 12.6 10.1 11.6 8.8 7.8 7.9 3.2 ...  
## $ Temp3pm : num 29.7 18.1 19.6 21.6 19.1 19.5 11.4 11.7 12.6 11.2 ...  
## $ RainToday : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 2 1 1 ...  
## $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 2 2 1 1 1 ...

set.seed(1234)   
train.rows = createDataPartition(y = rain\_complete3$RainTomorrow, p=0.7, list = FALSE)   
train3 = rain\_complete3[train.rows,]   
test3 = rain\_complete3[-train.rows,]

allmod3 = glm(RainTomorrow ~., train3, family = "binomial")   
summary(allmod3)

##   
## Call:  
## glm(formula = RainTomorrow ~ ., family = "binomial", data = train3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9915 -0.5569 -0.2744 -0.0412 3.5492   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 41.801692 5.699328 7.334 2.23e-13 \*\*\*  
## MinTemp -0.001998 0.016842 -0.119 0.905578   
## MaxTemp -0.009885 0.027661 -0.357 0.720811   
## WindGustDirWNW 0.118719 0.157163 0.755 0.450016   
## WindGustDirENE -0.135566 0.195996 -0.692 0.489140   
## WindGustDirNE -0.154229 0.196844 -0.784 0.433326   
## WindGustDirNW -0.158793 0.174217 -0.911 0.362050   
## WindGustDirNNE -0.113313 0.202684 -0.559 0.576119   
## WindGustDirNNW -0.163614 0.189972 -0.861 0.389098   
## WindGustDirE -0.111075 0.192060 -0.578 0.563037   
## WindGustDirN -0.360481 0.176330 -2.044 0.040919 \*   
## WindGustDirWSW 0.003489 0.152794 0.023 0.981783   
## WindGustDirSE -0.043415 0.191647 -0.227 0.820785   
## WindGustDirESE -0.054671 0.199054 -0.275 0.783580   
## WindGustDirSW 0.026886 0.165234 0.163 0.870742   
## WindGustDirS -0.175121 0.180506 -0.970 0.331964   
## WindGustDirSSE 0.111297 0.185085 0.601 0.547621   
## WindGustDirSSW -0.196671 0.176414 -1.115 0.264925   
## WindGustSpeed 0.056606 0.003563 15.887 < 2e-16 \*\*\*  
## WindDir9amW -0.061150 0.182850 -0.334 0.738057   
## WindDir9amSSW -0.313910 0.199436 -1.574 0.115490   
## WindDir9amN 0.064667 0.172183 0.376 0.707234   
## WindDir9amWSW -0.276921 0.191008 -1.450 0.147119   
## WindDir9amSSE -0.471224 0.185416 -2.541 0.011040 \*   
## WindDir9amESE -0.460801 0.199953 -2.305 0.021192 \*   
## WindDir9amSE -0.299499 0.190028 -1.576 0.115008   
## WindDir9amS -0.443187 0.192553 -2.302 0.021356 \*   
## WindDir9amNNE 0.473846 0.183335 2.585 0.009750 \*\*   
## WindDir9amNW -0.309985 0.191750 -1.617 0.105962   
## WindDir9amWNW -0.318295 0.190306 -1.673 0.094418 .   
## WindDir9amSW -0.150870 0.186319 -0.810 0.418090   
## WindDir9amNE 0.075069 0.190679 0.394 0.693807   
## WindDir9amE -0.178513 0.187018 -0.955 0.339819   
## WindDir9amNNW -0.496235 0.189654 -2.617 0.008883 \*\*   
## WindDir3pmW -0.148052 0.177134 -0.836 0.403257   
## WindDir3pmWNW 0.034900 0.176088 0.198 0.842891   
## WindDir3pmE -0.357716 0.204791 -1.747 0.080683 .   
## WindDir3pmSSW -0.064725 0.200452 -0.323 0.746775   
## WindDir3pmWSW -0.245026 0.186150 -1.316 0.188080   
## WindDir3pmS -0.243000 0.198487 -1.224 0.220854   
## WindDir3pmSE 0.084792 0.198843 0.426 0.669798   
## WindDir3pmESE -0.271397 0.202190 -1.342 0.179502   
## WindDir3pmENE -0.374229 0.201713 -1.855 0.063560 .   
## WindDir3pmNE -0.125413 0.200022 -0.627 0.530661   
## WindDir3pmSSE -0.316221 0.204293 -1.548 0.121652   
## WindDir3pmSW -0.265041 0.193167 -1.372 0.170037   
## WindDir3pmN -0.201188 0.186637 -1.078 0.281049   
## WindDir3pmNNW 0.024953 0.185171 0.135 0.892806   
## WindDir3pmNNE -0.182981 0.202953 -0.902 0.367271   
## WindSpeed9am -0.009736 0.004699 -2.072 0.038291 \*   
## WindSpeed3pm -0.029895 0.004916 -6.081 1.20e-09 \*\*\*  
## Humidity9am 0.000825 0.003546 0.233 0.816044   
## Humidity3pm 0.056428 0.003732 15.119 < 2e-16 \*\*\*  
## Pressure9am 0.151039 0.019949 7.571 3.70e-14 \*\*\*  
## Pressure3pm -0.197649 0.020354 -9.711 < 2e-16 \*\*\*  
## Cloud9am0 -0.623646 0.276317 -2.257 0.024008 \*   
## Cloud9am8 0.075785 0.091257 0.830 0.406278   
## Cloud9am5 -0.139969 0.124748 -1.122 0.261855   
## Cloud9am2 -0.079033 0.134823 -0.586 0.557743   
## Cloud9am1 -0.524577 0.126961 -4.132 3.60e-05 \*\*\*  
## Cloud9am3 -0.249293 0.132959 -1.875 0.060798 .   
## Cloud9am4 -0.111837 0.146149 -0.765 0.444140   
## Cloud9am6 -0.201940 0.106756 -1.892 0.058544 .   
## Cloud3pm1 -1.403871 0.155526 -9.027 < 2e-16 \*\*\*  
## Cloud3pm6 -0.495105 0.114701 -4.316 1.59e-05 \*\*\*  
## Cloud3pm4 -0.555975 0.143912 -3.863 0.000112 \*\*\*  
## Cloud3pm5 -0.492430 0.132961 -3.704 0.000213 \*\*\*  
## Cloud3pm3 -0.824637 0.146401 -5.633 1.77e-08 \*\*\*  
## Cloud3pm7 -0.278705 0.092560 -3.011 0.002603 \*\*   
## Cloud3pm2 -1.205783 0.159614 -7.554 4.21e-14 \*\*\*  
## Cloud3pm0 -3.145140 0.737761 -4.263 2.02e-05 \*\*\*  
## Temp9am 0.040188 0.025720 1.563 0.118161   
## Temp3pm -0.019314 0.030620 -0.631 0.528185   
## RainTodayYes 0.513501 0.071544 7.177 7.10e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11125.1 on 10070 degrees of freedom  
## Residual deviance: 7128.3 on 9997 degrees of freedom  
## AIC: 7276.3  
##   
## Number of Fisher Scoring iterations: 8

emptymod3 = glm(RainTomorrow ~1, train3, family = "binomial")   
summary(emptymod3)

##   
## Call:  
## glm(formula = RainTomorrow ~ 1, family = "binomial", data = train3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7428 -0.7428 -0.7428 -0.7428 1.6868   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.1467 0.0233 -49.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11125 on 10070 degrees of freedom  
## Residual deviance: 11125 on 10070 degrees of freedom  
## AIC: 11127  
##   
## Number of Fisher Scoring iterations: 4

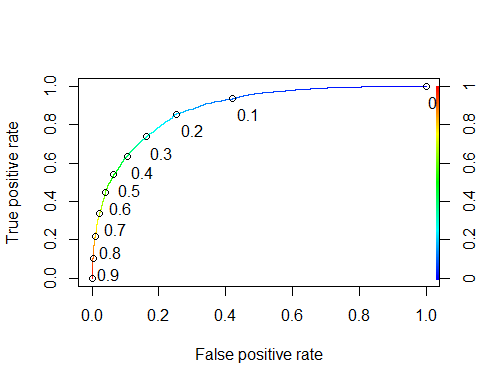
forwardmod3 = stepAIC(emptymod3, direction = "forward", scope=list(upper=allmod3,lower=emptymod3), trace = FALSE)  
summary(forwardmod3)

##   
## Call:  
## glm(formula = RainTomorrow ~ Humidity3pm + WindGustSpeed + Cloud3pm +   
## Pressure3pm + Pressure9am + RainToday + WindSpeed3pm + WindDir9am +   
## Cloud9am + WindSpeed9am + Temp9am + Temp3pm, family = "binomial",   
## data = train3)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0046 -0.5632 -0.2794 -0.0410 3.4876   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 44.600902 5.463255 8.164 3.25e-16 \*\*\*  
## Humidity3pm 0.055820 0.002757 20.248 < 2e-16 \*\*\*  
## WindGustSpeed 0.056153 0.003436 16.342 < 2e-16 \*\*\*  
## Cloud3pm1 -1.420091 0.153720 -9.238 < 2e-16 \*\*\*  
## Cloud3pm6 -0.491708 0.113535 -4.331 1.49e-05 \*\*\*  
## Cloud3pm4 -0.552472 0.142235 -3.884 0.000103 \*\*\*  
## Cloud3pm5 -0.498755 0.131436 -3.795 0.000148 \*\*\*  
## Cloud3pm3 -0.829083 0.144647 -5.732 9.94e-09 \*\*\*  
## Cloud3pm7 -0.278770 0.091761 -3.038 0.002381 \*\*   
## Cloud3pm2 -1.209569 0.157976 -7.657 1.91e-14 \*\*\*  
## Cloud3pm0 -3.163517 0.736491 -4.295 1.74e-05 \*\*\*  
## Pressure3pm -0.194891 0.018714 -10.414 < 2e-16 \*\*\*  
## Pressure9am 0.145349 0.018240 7.969 1.60e-15 \*\*\*  
## RainTodayYes 0.532220 0.067931 7.835 4.70e-15 \*\*\*  
## WindSpeed3pm -0.029035 0.004768 -6.089 1.13e-09 \*\*\*  
## WindDir9amW 0.056966 0.174574 0.326 0.744185   
## WindDir9amSSW -0.271703 0.190276 -1.428 0.153309   
## WindDir9amN 0.081368 0.163286 0.498 0.618262   
## WindDir9amWSW -0.207642 0.182921 -1.135 0.256313   
## WindDir9amSSE -0.372471 0.175426 -2.123 0.033734 \*   
## WindDir9amESE -0.414243 0.195083 -2.123 0.033719 \*   
## WindDir9amSE -0.211584 0.182670 -1.158 0.246748   
## WindDir9amS -0.379010 0.181218 -2.091 0.036487 \*   
## WindDir9amNNE 0.508923 0.177117 2.873 0.004061 \*\*   
## WindDir9amNW -0.172666 0.184226 -0.937 0.348629   
## WindDir9amWNW -0.167516 0.180745 -0.927 0.354026   
## WindDir9amSW -0.102955 0.177909 -0.579 0.562795   
## WindDir9amNE 0.093397 0.187881 0.497 0.619112   
## WindDir9amE -0.195337 0.184388 -1.059 0.289429   
## WindDir9amNNW -0.386071 0.181099 -2.132 0.033021 \*   
## Cloud9am0 -0.629485 0.271176 -2.321 0.020270 \*   
## Cloud9am8 0.093336 0.090319 1.033 0.301417   
## Cloud9am5 -0.114242 0.122708 -0.931 0.351853   
## Cloud9am2 -0.064957 0.132523 -0.490 0.624024   
## Cloud9am1 -0.505719 0.123578 -4.092 4.27e-05 \*\*\*  
## Cloud9am3 -0.217822 0.131161 -1.661 0.096768 .   
## Cloud9am4 -0.094951 0.144652 -0.656 0.511560   
## Cloud9am6 -0.192976 0.105558 -1.828 0.067529 .   
## WindSpeed9am -0.010942 0.004443 -2.463 0.013790 \*   
## Temp9am 0.030595 0.014871 2.057 0.039651 \*   
## Temp3pm -0.024542 0.016509 -1.487 0.137115   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11125.1 on 10070 degrees of freedom  
## Residual deviance: 7164.5 on 10030 degrees of freedom  
## AIC: 7246.5  
##   
## Number of Fisher Scoring iterations: 8

### This model produced a slightly worse AIC score than the first model (about +200 higher), but there does appear to be a higher number of variables that show some significance towards our RainTomorrow response variable.

predictions = predict(forwardmod3, type="response")  
ROCRpred = prediction(predictions, train3$RainTomorrow)

ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



opt.cut = function(perf, pred){   
 cut.ind = mapply(FUN=function(x, y, p){   
 d = (x - 0)^2 + (y-1)^2   
 ind = which(d == min(d))   
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])   
 }, perf@x.values, perf@y.values, pred@cutoffs)   
}   
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.8084843  
## specificity 0.7852937  
## cutoff 0.2371556

Confusion matrix

table(train3$RainTomorrow, predictions > 0.2371556)

##   
## FALSE TRUE  
## No 6002 1641  
## Yes 465 1963

Accuracy

(6002+1963)/nrow(train3)

## [1] 0.7908847

Sensitivity

(1963)/(1963+465)

## [1] 0.8084843

Speciﬁcity

(6002)/(6002+1641)

## [1] 0.7852937

t1 = table(train3$RainTomorrow,predictions > 0.3)   
(t1[1,1]+t1[2,2])/nrow(train3)

## [1] 0.8131268

t1 = table(train3$RainTomorrow,predictions > 0.4)   
(t1[1,1]+t1[2,2])/nrow(train3)

## [1] 0.8327872

t1 = table(train3$RainTomorrow,predictions > 0.5)   
(t1[1,1]+t1[2,2])/nrow(train3)

## [1] 0.8397379

t1 = table(train3$RainTomorrow,predictions > 0.6)   
(t1[1,1]+t1[2,2])/nrow(train3)

## [1] 0.8361632

predictions\_test3 = predict(forwardmod3, newdata = test3, type="response")   
table(test3$RainTomorrow,predictions\_test3 > 0.5)

##   
## FALSE TRUE  
## No 3097 178  
## Yes 461 579

Accuracy for Test

(3097+579)/nrow(test3)

## [1] 0.8519119

Sensitivity for Test

(579)/(579+461)

## [1] 0.5567308

Speciﬁcity for Test

(3097)/(3097+178)

## [1] 0.9456489

### Even with this model’s higer AIC score, it actually provided slightly better results for Accuracy, Sensitivity, and Specificity when compared to the first Forward Stepwise model: Accuracy(.8519 vs .8404), Sensitivity(.5567 vs .5562), & Specificity(.9456 vs .9316).