# BAN 502 Project: Part 1

## Data Exploration, Preparation, and Visualization

### Rachel Urban

library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.3.0  
## v tibble 2.0.1 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.3.0

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

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(caret)

##   
## Attaching package: 'caret'

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

library(ggcorrplot)  
library(leaps)  
library(GGally)

##   
## Attaching package: 'GGally'

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

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

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

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

str(rain)

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

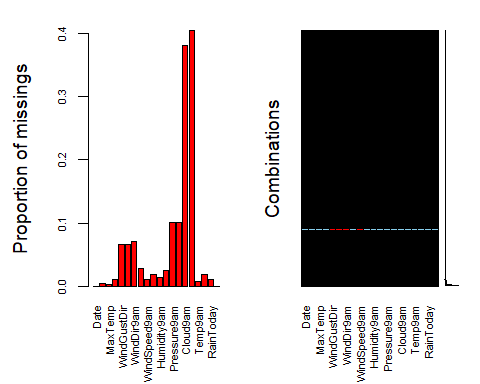
summary(rain)

## 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 = rain %>% mutate(RainToday = as.factor(RainToday)) %>%  
 mutate(RainTomorrow = as.factor(RainTomorrow))

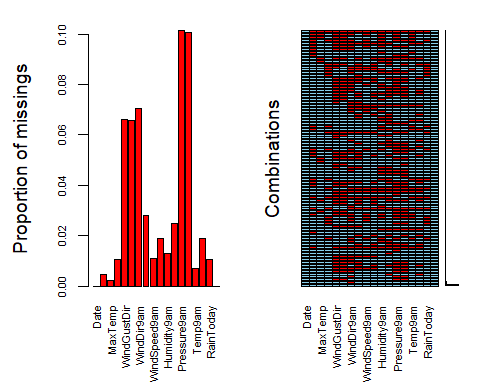
vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

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



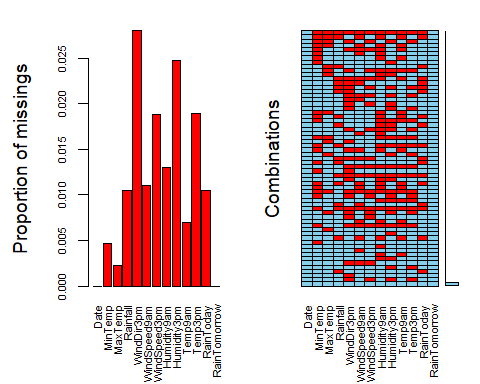
rain = rain %>% dplyr::select(-Cloud3pm) %>%   
 dplyr::select(-Cloud9am)  
   
vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

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



rain = rain %>% dplyr::select(-WindGustDir) %>%   
 dplyr::select(-WindGustSpeed) %>%  
 dplyr::select(-WindDir9am) %>%  
 dplyr::select(-Pressure9am) %>%  
 dplyr::select(-Pressure3pm)  
  
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 %>% dplyr::select(c("MinTemp","MaxTemp","Humidity9am","Humidity3pm","WindSpeed9am","WindSpeed3pm","RainToday","RainTomorrow"))  
  
str(rain)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 8 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 ...  
## $ Humidity9am : num 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : num 33 23 28 82 24 15 22 15 16 8 ...  
## $ WindSpeed9am: num 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm: num 20 24 22 9 20 30 11 24 9 17 ...  
## $ 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 ...

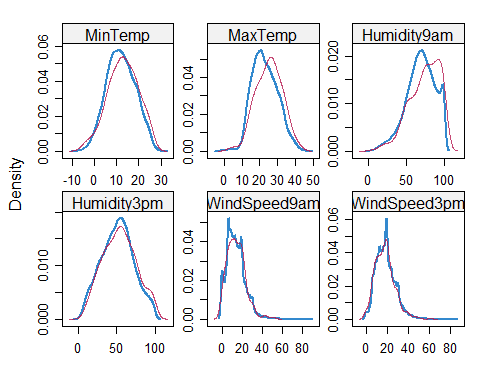
summary(rain)

## MinTemp MaxTemp Humidity9am Humidity3pm   
## Min. :-8.50 Min. :-3.00 Min. : 1.00 Min. : 0.00   
## 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 57.00 1st Qu.: 36.00   
## Median :12.00 Median :22.60 Median : 70.00 Median : 52.00   
## Mean :12.16 Mean :23.18 Mean : 68.86 Mean : 51.54   
## 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 83.00 3rd Qu.: 66.00   
## Max. :30.50 Max. :47.00 Max. :100.00 Max. :100.00   
## NA's :132 NA's :64 NA's :366 NA's :694   
## WindSpeed9am WindSpeed3pm RainToday RainTomorrow  
## Min. : 0.00 Min. : 0.00 No :21525 No :21713   
## 1st Qu.: 7.00 1st Qu.:13.00 Yes : 6183 Yes: 6290   
## Median :13.00 Median :19.00 NA's: 295   
## Mean :13.97 Mean :18.59   
## 3rd Qu.:19.00 3rd Qu.:24.00   
## Max. :87.00 Max. :83.00   
## NA's :308 NA's :526

imp\_rain = mice(rain, m=1, method='pmm', printFlag=FALSE)  
summary(imp\_rain)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## MinTemp MaxTemp Humidity9am Humidity3pm WindSpeed9am   
## "pmm" "pmm" "pmm" "pmm" "pmm"   
## WindSpeed3pm RainToday RainTomorrow   
## "pmm" "pmm" ""   
## PredictorMatrix:  
## MinTemp MaxTemp Humidity9am Humidity3pm WindSpeed9am  
## MinTemp 0 1 1 1 1  
## MaxTemp 1 0 1 1 1  
## Humidity9am 1 1 0 1 1  
## Humidity3pm 1 1 1 0 1  
## WindSpeed9am 1 1 1 1 0  
## WindSpeed3pm 1 1 1 1 1  
## WindSpeed3pm RainToday RainTomorrow  
## MinTemp 1 1 1  
## MaxTemp 1 1 1  
## Humidity9am 1 1 1  
## Humidity3pm 1 1 1  
## WindSpeed9am 1 1 1  
## WindSpeed3pm 0 1 1

densityplot(imp\_rain)



rain\_complete = complete(imp\_rain)   
str(rain\_complete)

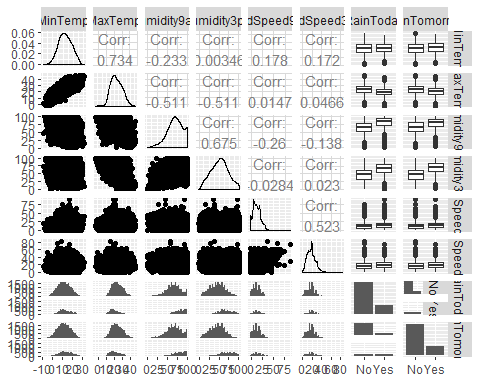
## 'data.frame': 28003 obs. of 8 variables:  
## $ MinTemp : num 17.5 14.6 9.8 14.1 20.5 20.1 9.6 14 12.5 17.4 ...  
## $ MaxTemp : num 32.3 29.7 27.7 20.9 31.8 32.7 23.9 28.3 28.4 43 ...  
## $ Humidity9am : num 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : num 33 23 28 82 24 15 22 15 16 8 ...  
## $ WindSpeed9am: num 7 19 19 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm: num 20 24 22 9 20 30 11 24 9 17 ...  
## $ 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 1 2 1 1 1 1 1 1 ...

summary(rain\_complete)

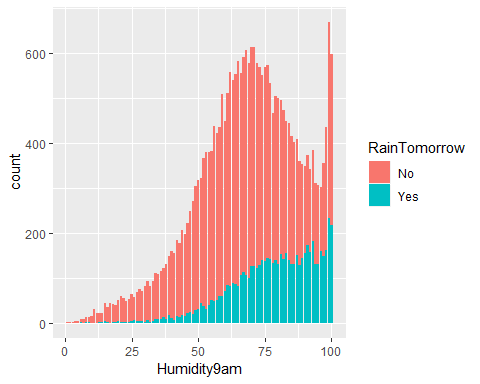
## MinTemp MaxTemp Humidity9am Humidity3pm   
## Min. :-8.50 Min. :-3.00 Min. : 1.00 Min. : 0.0   
## 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 57.00 1st Qu.: 36.5   
## Median :12.00 Median :22.60 Median : 70.00 Median : 52.0   
## Mean :12.16 Mean :23.18 Mean : 68.91 Mean : 51.6   
## 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 83.00 3rd Qu.: 66.0   
## Max. :30.50 Max. :47.00 Max. :100.00 Max. :100.0   
## WindSpeed9am WindSpeed3pm RainToday RainTomorrow  
## Min. : 0.00 Min. : 0.00 No :21719 No :21713   
## 1st Qu.: 7.00 1st Qu.:13.00 Yes: 6284 Yes: 6290   
## Median :13.00 Median :19.00   
## Mean :13.97 Mean :18.57   
## 3rd Qu.:19.00 3rd Qu.:24.00   
## Max. :87.00 Max. :83.00

ggpairs(rain\_complete)

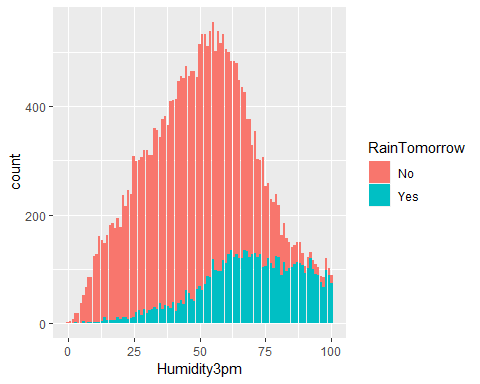
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()`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



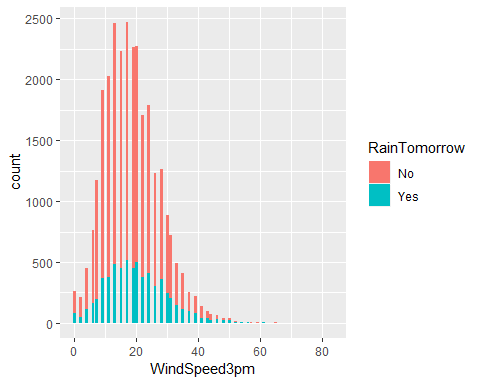
ggplot(rain\_complete,aes(x=Humidity9am,fill=RainTomorrow)) + geom\_bar()



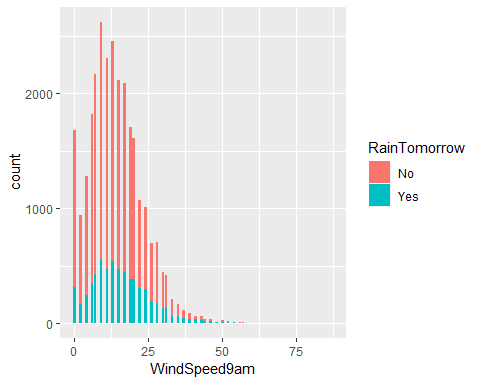
ggplot(rain\_complete,aes(x=Humidity3pm,fill=RainTomorrow)) + geom\_bar()



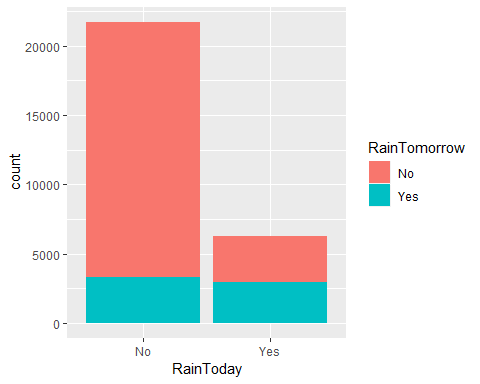
ggplot(rain\_complete,aes(x=WindSpeed3pm,fill=RainTomorrow)) + geom\_bar()



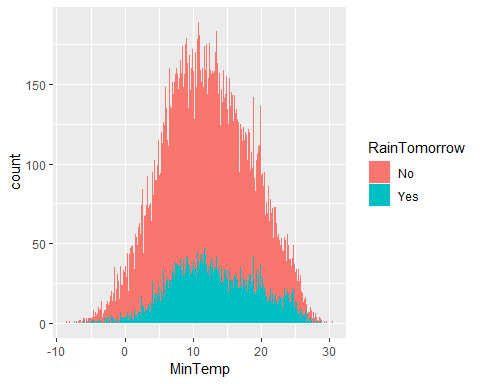
ggplot(rain\_complete,aes(x=WindSpeed9am,fill=RainTomorrow)) + geom\_bar()



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



ggplot(rain\_complete,aes(x=MinTemp,fill=RainTomorrow)) + geom\_bar()



ggplot(rain\_complete,aes(x=MaxTemp,fill=RainTomorrow)) + geom\_bar()

