# Project 1-Data Exploration, Preparation, and Visualization

## BAN502 Predictive Analytics

### Susan Wiggins

Libraries needed for Project 1 Data Exploration:

options(tidyverse.quiet = TRUE)  
library(tidyverse)  
library(mice) #package for imputation

## 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) #visualizing missingness

## 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(leaps) #best subset selection  
library(MASS) #access to forward and backward selection algorithms

##   
## Attaching package: 'MASS'

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

library(weathermetrics)  
library(caret) #for splitting functions

##   
## Attaching package: 'caret'

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

Reading in the data set:

library(readr)  
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.

View(rain)

### Structure and summary and Visualizations for understanding of variables in the data set:

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

Converting variables to factors:

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

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

glimpse(rain)

## Observations: 28,003  
## Variables: 20  
## $ Date <chr> "12/5/2008", "12/6/2008", "12/16/2008", "12/17/2...  
## $ 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 <chr> "W", "WNW", "WNW", "ENE", "WNW", "WNW", "W", "W"...  
## $ WindGustSpeed <dbl> 41, 56, 50, 22, 41, 48, 41, 48, 37, 39, 44, 28, ...  
## $ WindDir9am <chr> "ENE", "W", NA, "SSW", "W", "N", "WSW", "W", "SS...  
## $ WindDir3pm <chr> "NW", "W", "WNW", "E", "W", "WNW", "SSW", "WSW",...  
## $ WindSpeed9am <dbl> 7, 19, NA, 11, 19, 13, 19, 17, 20, 7, 20, 17, 9,...  
## $ WindSpeed3pm <dbl> 20, 24, 22, 9, 20, 30, 11, 24, 9, 17, 28, 15, 11...  
## $ Humidity9am <dbl> 82, 55, 50, 69, 54, 56, 44, 43, 38, 40, 34, 34, ...  
## $ Humidity3pm <dbl> 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 <dbl> 7, NA, 0, 8, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ Cloud3pm <dbl> 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 <chr> "No", "No", NA, "No", "No", "No", "No", "No", "N...  
## $ RainTomorrow <fct> No, No, No, Yes, No, No, No, No, No, No, No, No,...

rain = rain %>% mutate(Cloud9am = as\_factor(as.character(Cloud9am)))%>%   
mutate(Cloud9am = fct\_recode(Cloud9am, "clear sky" = "0", "few clouds" = "1", "2", "scattered clouds" = "3", "4","broken" = "5", "6","7", "completely overcast" = "8"))  
  
  
glimpse(rain)

## Observations: 28,003  
## Variables: 20  
## $ Date <chr> "12/5/2008", "12/6/2008", "12/16/2008", "12/17/2...  
## $ 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 <chr> "W", "WNW", "WNW", "ENE", "WNW", "WNW", "W", "W"...  
## $ WindGustSpeed <dbl> 41, 56, 50, 22, 41, 48, 41, 48, 37, 39, 44, 28, ...  
## $ WindDir9am <chr> "ENE", "W", NA, "SSW", "W", "N", "WSW", "W", "SS...  
## $ WindDir3pm <chr> "NW", "W", "WNW", "E", "W", "WNW", "SSW", "WSW",...  
## $ WindSpeed9am <dbl> 7, 19, NA, 11, 19, 13, 19, 17, 20, 7, 20, 17, 9,...  
## $ WindSpeed3pm <dbl> 20, 24, 22, 9, 20, 30, 11, 24, 9, 17, 28, 15, 11...  
## $ Humidity9am <dbl> 82, 55, 50, 69, 54, 56, 44, 43, 38, 40, 34, 34, ...  
## $ Humidity3pm <dbl> 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 <fct> , NA, clear sky, completely overcast, NA, NA, NA...  
## $ Cloud3pm <dbl> 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 <chr> "No", "No", NA, "No", "No", "No", "No", "No", "N...  
## $ RainTomorrow <fct> No, No, No, Yes, No, No, No, No, No, No, No, No,...

rain = rain %>% mutate(Cloud3pm = as\_factor(as.character(Cloud3pm)))%>%   
mutate(Cloud3pm = fct\_recode(Cloud3pm, "clear sky" = "0", "few clouds" = "1", "2", "scattered clouds" = "3", "4","broken" = "5", "6","7", "completely overcast" = "8"))  
  
  
glimpse(rain)

## Observations: 28,003  
## Variables: 20  
## $ Date <chr> "12/5/2008", "12/6/2008", "12/16/2008", "12/17/2...  
## $ 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 <chr> "W", "WNW", "WNW", "ENE", "WNW", "WNW", "W", "W"...  
## $ WindGustSpeed <dbl> 41, 56, 50, 22, 41, 48, 41, 48, 37, 39, 44, 28, ...  
## $ WindDir9am <chr> "ENE", "W", NA, "SSW", "W", "N", "WSW", "W", "SS...  
## $ WindDir3pm <chr> "NW", "W", "WNW", "E", "W", "WNW", "SSW", "WSW",...  
## $ WindSpeed9am <dbl> 7, 19, NA, 11, 19, 13, 19, 17, 20, 7, 20, 17, 9,...  
## $ WindSpeed3pm <dbl> 20, 24, 22, 9, 20, 30, 11, 24, 9, 17, 28, 15, 11...  
## $ Humidity9am <dbl> 82, 55, 50, 69, 54, 56, 44, 43, 38, 40, 34, 34, ...  
## $ Humidity3pm <dbl> 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 <fct> , NA, clear sky, completely overcast, NA, NA, NA...  
## $ Cloud3pm <fct> completely overcast, NA, NA, few clouds, 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 <chr> "No", "No", NA, "No", "No", "No", "No", "No", "N...  
## $ RainTomorrow <fct> No, No, No, Yes, No, No, No, No, No, No, No, No,...

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

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

glimpse(rain)

## Observations: 28,003  
## Variables: 20  
## $ Date <chr> "12/5/2008", "12/6/2008", "12/16/2008", "12/17/2...  
## $ 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 <chr> "W", "WNW", "WNW", "ENE", "WNW", "WNW", "W", "W"...  
## $ WindGustSpeed <dbl> 41, 56, 50, 22, 41, 48, 41, 48, 37, 39, 44, 28, ...  
## $ WindDir9am <chr> "ENE", "W", NA, "SSW", "W", "N", "WSW", "W", "SS...  
## $ WindDir3pm <chr> "NW", "W", "WNW", "E", "W", "WNW", "SSW", "WSW",...  
## $ WindSpeed9am <dbl> 7, 19, NA, 11, 19, 13, 19, 17, 20, 7, 20, 17, 9,...  
## $ WindSpeed3pm <dbl> 20, 24, 22, 9, 20, 30, 11, 24, 9, 17, 28, 15, 11...  
## $ Humidity9am <dbl> 82, 55, 50, 69, 54, 56, 44, 43, 38, 40, 34, 34, ...  
## $ Humidity3pm <dbl> 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 <fct> , NA, clear sky, completely overcast, NA, NA, NA...  
## $ Cloud3pm <fct> completely overcast, NA, NA, few clouds, 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,...

Converting Celsius to Farhenheit since United States looks at this scale for Temperature readings:

data(rain)

## Warning in data(rain): data set 'rain' not found

rain$MinTempF <- convert\_temperature(rain$MinTemp,  
 old\_metric = "celsius", new\_metric = "fahrenheit")  
  
rain$MaxTempF <- convert\_temperature(rain$MaxTemp,  
 old\_metric = "celsius", new\_metric = "fahrenheit")  
  
rain$Temp9amF <- convert\_temperature(rain$Temp9am,  
 old\_metric = "celsius", new\_metric = "fahrenheit")  
  
rain$Temp3pmF <- convert\_temperature(rain$Temp3pm,  
 old\_metric = "celsius", new\_metric = "fahrenheit")  
  
  
  
  
str(rain)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 24 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 : Factor w/ 6 levels "","clear sky",..: 1 NA 2 3 NA NA NA NA NA NA ...  
## $ Cloud3pm : Factor w/ 6 levels "completely overcast",..: 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 ...  
## $ MinTempF : num 63.5 58.3 49.6 57.4 68.9 ...  
## $ MaxTempF : num 90.1 85.5 81.9 69.6 89.2 ...  
## $ Temp9amF : num 64 69.1 63.1 63 74.8 ...  
## $ Temp3pmF : num 85.5 84 79.2 64.6 87.4 ...

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   
## Min. : 980.5 Min. : 978.2 : 7566   
## 1st Qu.:1013.0 1st Qu.:1010.5 clear sky : 1672   
## Median :1017.7 Median :1015.3 completely overcast: 2828   
## Mean :1017.7 Mean :1015.3 broken : 1095   
## 3rd Qu.:1022.4 3rd Qu.:1020.0 few clouds : 2986   
## Max. :1041.0 Max. :1037.0 scattered clouds : 1183   
## NA's :2837 NA's :2817 NA's :10673   
## Cloud3pm Temp9am Temp3pm RainToday   
## completely overcast: 2426 Min. :-5.60 Min. :-4.20 No :21525   
## few clouds : 2868 1st Qu.:12.30 1st Qu.:16.60 Yes : 6183   
## : 7783 Median :16.70 Median :21.10 NA's: 295   
## broken : 1258 Mean :16.96 Mean :21.63   
## scattered clouds : 1359 3rd Qu.:21.50 3rd Qu.:26.40   
## clear sky : 968 Max. :38.60 Max. :45.20   
## NA's :11341 NA's :196 NA's :532   
## RainTomorrow MinTempF MaxTempF Temp9amF   
## No :21713 Min. :16.70 Min. : 26.60 Min. : 21.92   
## Yes: 6290 1st Qu.:45.68 1st Qu.: 64.22 1st Qu.: 54.14   
## Median :53.60 Median : 72.68 Median : 62.06   
## Mean :53.89 Mean : 73.72 Mean : 62.52   
## 3rd Qu.:62.24 3rd Qu.: 82.76 3rd Qu.: 70.70   
## Max. :86.90 Max. :116.60 Max. :101.48   
## NA's :132 NA's :64 NA's :196   
## Temp3pmF   
## Min. : 24.44   
## 1st Qu.: 61.88   
## Median : 69.98   
## Mean : 70.94   
## 3rd Qu.: 79.52   
## Max. :113.36   
## NA's :532

Difference in the MaxTemperture and the Minimum Temperature in the data set along with a conversion on the temperature scale:

rain <- mutate(rain, Temperture\_Change = MaxTemp - MinTemp)  
  
  
rain$Temperture\_ChangeF <- convert\_temperature(rain$Temperture\_Change,  
 old\_metric = "celsius", new\_metric = "fahrenheit")  
  
  
rain$Cloud3pm[rain$Cloud3pm==""] = NA #convert blanks in cabin to NA  
rain$Cloud9am[rain$Cloud9am==""] = NA #convert blanks in cabin to NA  
  
str(rain)

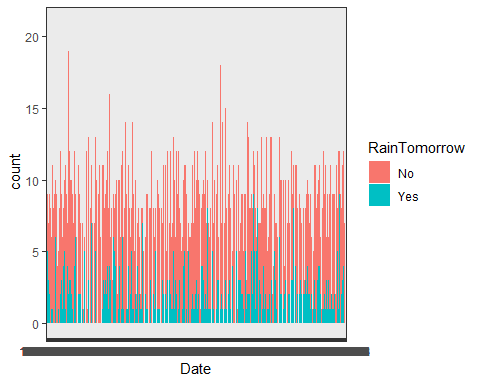
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 26 variables:  
## $ Date : chr "12/5/2008" "12/6/2008" "12/16/2008" "12/17/2008" ...  
## $ MinTemp : num 17.5 14.6 9.8 14.1 20.5 20.1 9.6 14 12.5 17.4 ...  
## $ MaxTemp : num 32.3 29.7 27.7 20.9 31.8 32.7 23.9 28.3 28.4 43 ...  
## $ 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 : Factor w/ 6 levels "","clear sky",..: NA NA 2 3 NA NA NA NA NA NA ...  
## $ Cloud3pm : Factor w/ 6 levels "completely overcast",..: 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 ...  
## $ MinTempF : num 63.5 58.3 49.6 57.4 68.9 ...  
## $ MaxTempF : num 90.1 85.5 81.9 69.6 89.2 ...  
## $ Temp9amF : num 64 69.1 63.1 63 74.8 ...  
## $ Temp3pmF : num 85.5 84 79.2 64.6 87.4 ...  
## $ Temperture\_Change : num 14.8 15.1 17.9 6.8 11.3 12.6 14.3 14.3 15.9 25.6 ...  
## $ Temperture\_ChangeF: num 58.6 59.2 64.2 44.2 52.3 ...

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   
## Min. : 980.5 Min. : 978.2 : 0   
## 1st Qu.:1013.0 1st Qu.:1010.5 clear sky : 1672   
## Median :1017.7 Median :1015.3 completely overcast: 2828   
## Mean :1017.7 Mean :1015.3 broken : 1095   
## 3rd Qu.:1022.4 3rd Qu.:1020.0 few clouds : 2986   
## Max. :1041.0 Max. :1037.0 scattered clouds : 1183   
## NA's :2837 NA's :2817 NA's :18239   
## Cloud3pm Temp9am Temp3pm RainToday   
## completely overcast: 2426 Min. :-5.60 Min. :-4.20 No :21525   
## few clouds : 2868 1st Qu.:12.30 1st Qu.:16.60 Yes : 6183   
## : 0 Median :16.70 Median :21.10 NA's: 295   
## broken : 1258 Mean :16.96 Mean :21.63   
## scattered clouds : 1359 3rd Qu.:21.50 3rd Qu.:26.40   
## clear sky : 968 Max. :38.60 Max. :45.20   
## NA's :19124 NA's :196 NA's :532   
## RainTomorrow MinTempF MaxTempF Temp9amF   
## No :21713 Min. :16.70 Min. : 26.60 Min. : 21.92   
## Yes: 6290 1st Qu.:45.68 1st Qu.: 64.22 1st Qu.: 54.14   
## Median :53.60 Median : 72.68 Median : 62.06   
## Mean :53.89 Mean : 73.72 Mean : 62.52   
## 3rd Qu.:62.24 3rd Qu.: 82.76 3rd Qu.: 70.70   
## Max. :86.90 Max. :116.60 Max. :101.48   
## NA's :132 NA's :64 NA's :196   
## Temp3pmF Temperture\_Change Temperture\_ChangeF  
## Min. : 24.44 Min. : 0.00 Min. :32.00   
## 1st Qu.: 61.88 1st Qu.: 7.20 1st Qu.:44.96   
## Median : 69.98 Median :10.40 Median :50.72   
## Mean : 70.94 Mean :11.02 Mean :51.84   
## 3rd Qu.: 79.52 3rd Qu.:14.50 3rd Qu.:58.10   
## Max. :113.36 Max. :30.00 Max. :86.00   
## NA's :532 NA's :177 NA's :177

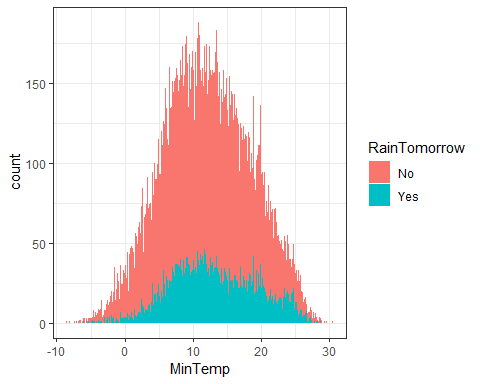
Visualization of Variables in Rain dataset:

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



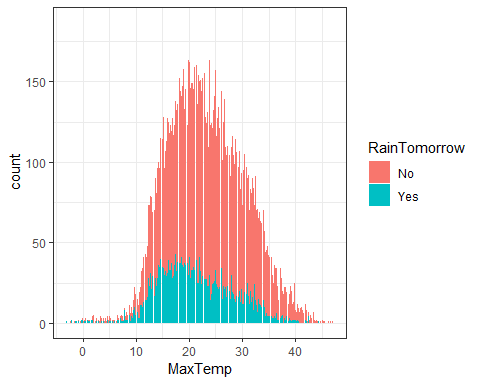
ggplot(rain, aes(x= MinTemp, fill = RainTomorrow)) + geom\_bar() + theme\_bw()

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



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

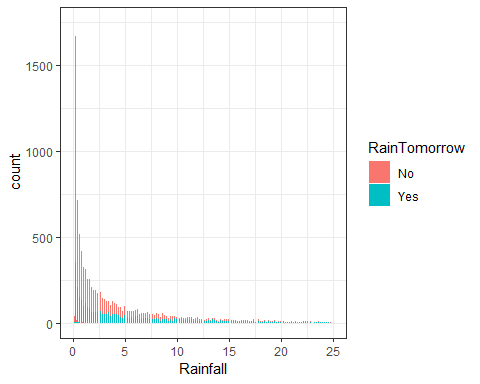
## Warning: Removed 64 rows containing non-finite values (stat\_count).



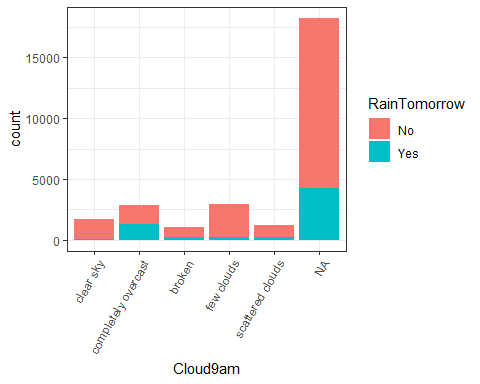
ggplot(rain, aes(x= Rainfall, fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + scale\_x\_continuous(limit = c(0, 25)) +  
 scale\_y\_continuous(limit = c(0, 1750))

## Warning: Removed 831 rows containing non-finite values (stat\_count).

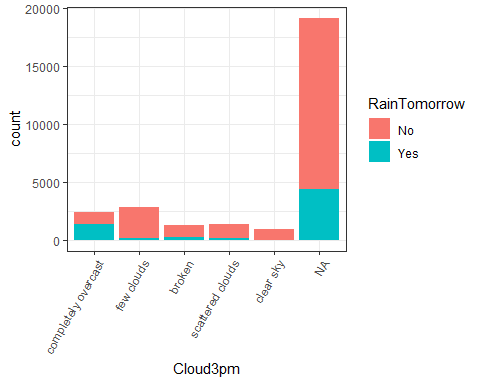
## Warning: Removed 4 rows containing missing values (geom\_bar).



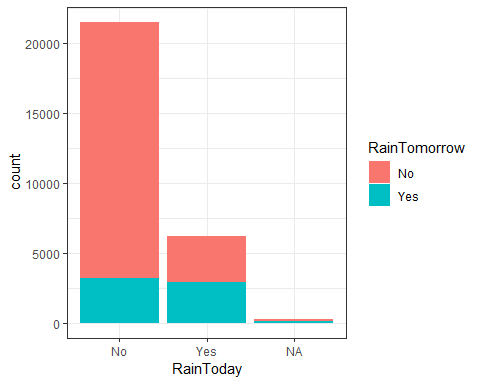
ggplot(rain, aes(x= Cloud9am, fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))



ggplot(rain, aes(x= Cloud3pm, fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))



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

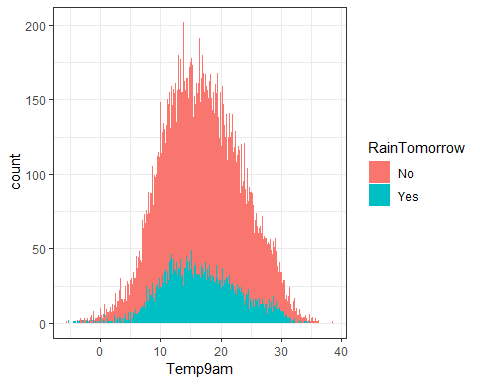


t\_rain = table(rain$RainTomorrow, rain$RainToday) #create a table object  
prop.table(t\_rain, margin = 2 ) #crosstab with proportions

##   
## No Yes  
## No 0.8492451 0.5312955  
## Yes 0.1507549 0.4687045

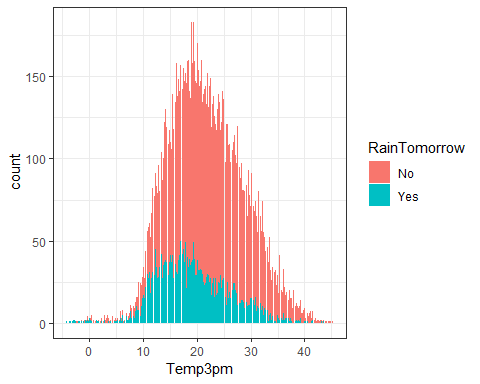
ggplot(rain, aes(x= Temp9am , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw()

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

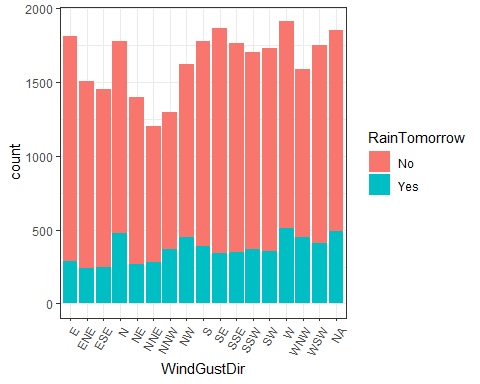


ggplot(rain, aes(x= Temp3pm , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw()

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

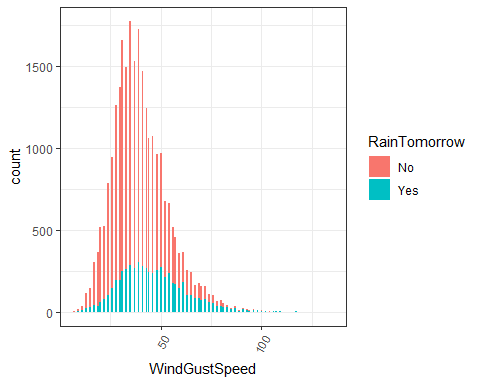


ggplot(rain, aes(x= WindGustDir , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

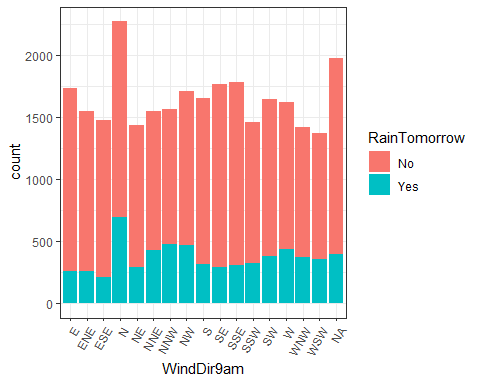


ggplot(rain, aes(x= WindGustSpeed , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

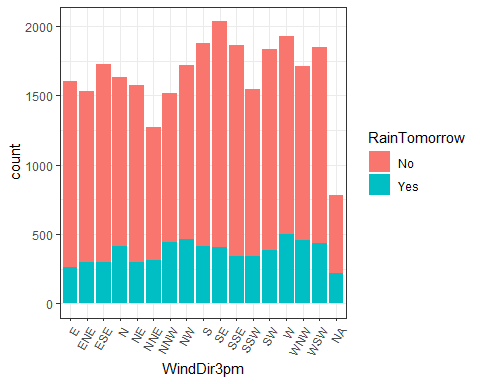
## Warning: Removed 1840 rows containing non-finite values (stat\_count).



ggplot(rain, aes(x= WindDir9am , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

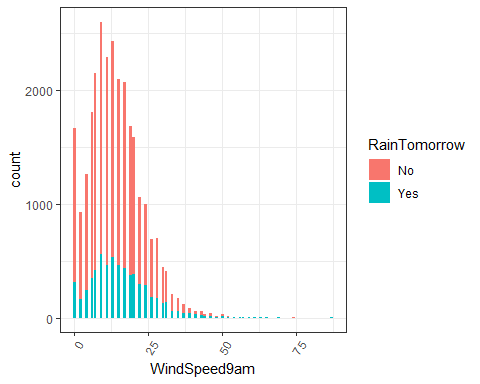


ggplot(rain, aes(x= WindDir3pm, fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))



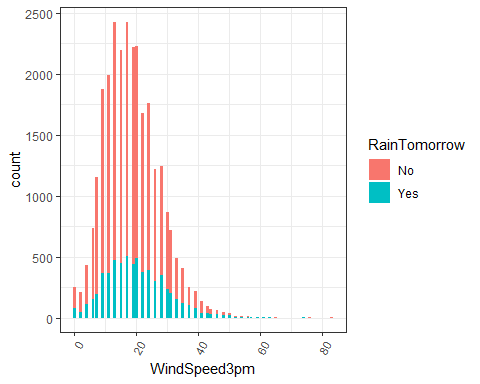
ggplot(rain, aes(x= WindSpeed9am , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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



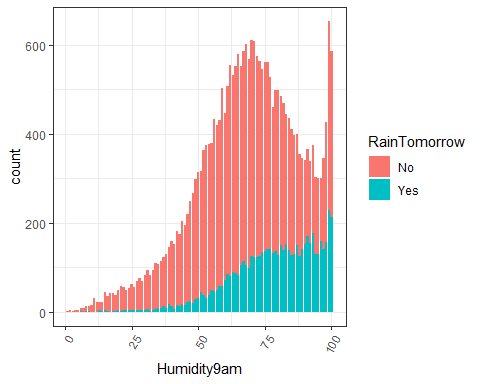
ggplot(rain, aes(x= WindSpeed3pm , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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



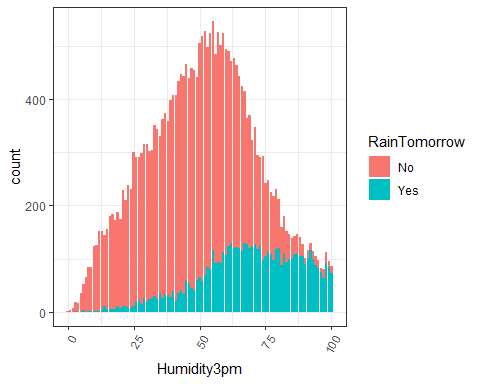
ggplot(rain, aes(x= Humidity9am , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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



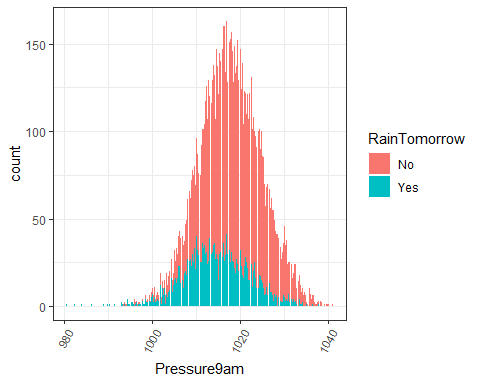
ggplot(rain, aes(x= Humidity3pm , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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



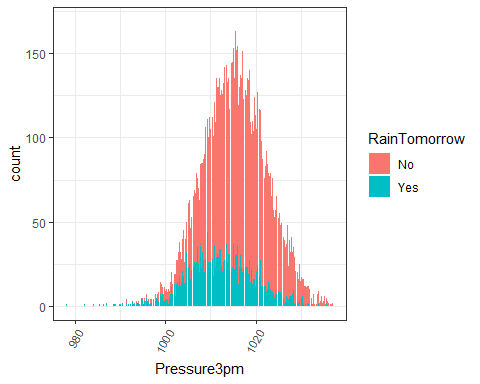
ggplot(rain, aes(x= Pressure9am , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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

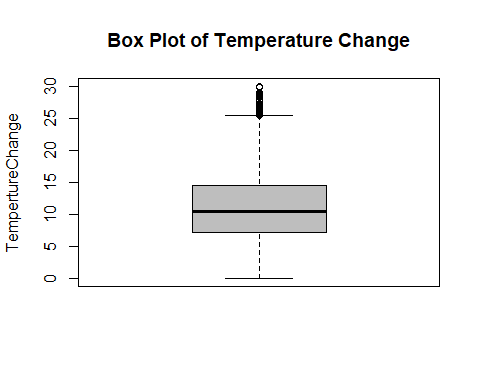


ggplot(rain, aes(x= Pressure3pm , fill = RainTomorrow)) + geom\_bar() +   
 theme\_bw() + theme(axis.text.x = element\_text(angle=60, hjust=1))

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

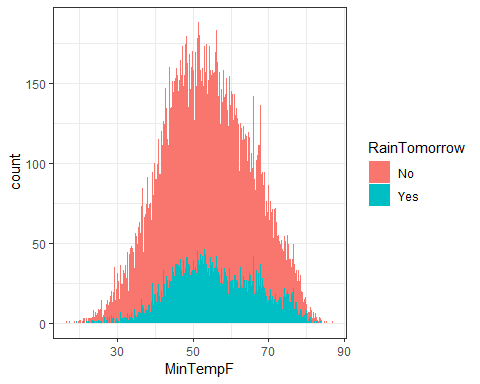


rain$Temperture\_Change %>% boxplot(main="Box Plot of Temperature Change", ylab="TempertureChange", col = "grey")



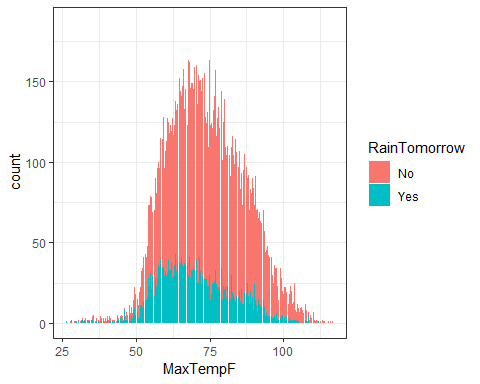
ggplot(rain, aes(x= MinTempF, fill = RainTomorrow)) + geom\_bar() + theme\_bw()

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



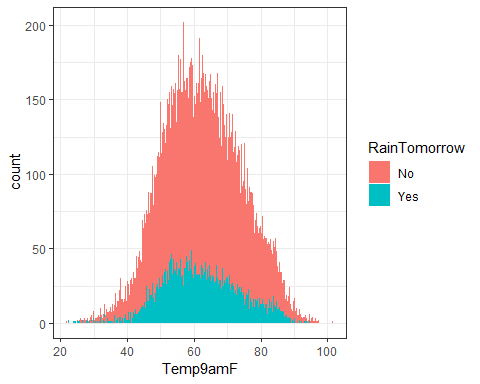
ggplot(rain, aes(x= MaxTempF, fill = RainTomorrow)) + geom\_bar() + theme\_bw()

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



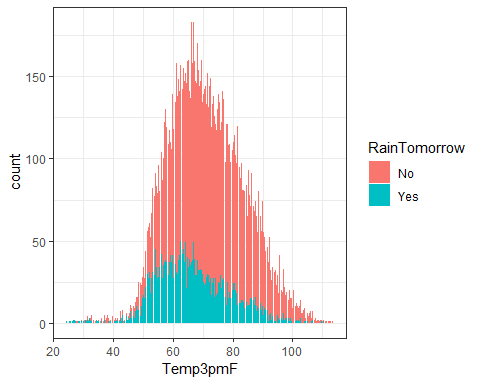
ggplot(rain, aes(x= Temp9amF, fill = RainTomorrow)) + geom\_bar() + theme\_bw()

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



ggplot(rain, aes(x= Temp3pmF, fill = RainTomorrow)) + geom\_bar() + theme\_bw()

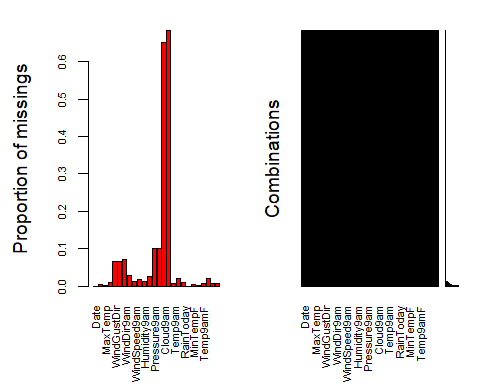
## Warning: Removed 532 rows containing non-finite values (stat\_count).



### Identify Missingness of data and appropriately deal with missingness:

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

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



Removal of Cloud9am and Cloud3pm due to such a large degree of data missing around over 60% missing data points.

Column-wise deletion of the “Cloud9am” and “Cloud3pm” variable, not sure why this was not removed. Next, Letting use the “mice” package to do imputation instead.

#rain = rain %>% select(-Cloud9am)   
#vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)  
  
#rain = rain %>% select(-Cloud3pm)   
#vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)  
  
rain\_subset = rain %>% dplyr:: select(Date, MinTempF, MaxTempF, Rainfall, RainTomorrow, RainToday, Temperture\_ChangeF, Humidity9am, Humidity3pm)  
head(rain)

## # A tibble: 6 x 26  
## Date MinTemp MaxTemp Rainfall WindGustDir WindGustSpeed WindDir9am  
## <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 12/5~ 17.5 32.3 1 W 41 ENE   
## 2 12/6~ 14.6 29.7 0.2 WNW 56 W   
## 3 12/1~ 9.8 27.7 NA WNW 50 <NA>   
## 4 12/1~ 14.1 20.9 0 ENE 22 SSW   
## 5 12/2~ 20.5 31.8 0 WNW 41 W   
## 6 12/2~ 20.1 32.7 0 WNW 48 N   
## # ... with 19 more variables: WindDir3pm <chr>, WindSpeed9am <dbl>,  
## # WindSpeed3pm <dbl>, Humidity9am <dbl>, Humidity3pm <dbl>,  
## # Pressure9am <dbl>, Pressure3pm <dbl>, Cloud9am <fct>, Cloud3pm <fct>,  
## # Temp9am <dbl>, Temp3pm <dbl>, RainToday <fct>, RainTomorrow <fct>,  
## # MinTempF <dbl>, MaxTempF <dbl>, Temp9amF <dbl>, Temp3pmF <dbl>,  
## # Temperture\_Change <dbl>, Temperture\_ChangeF <dbl>

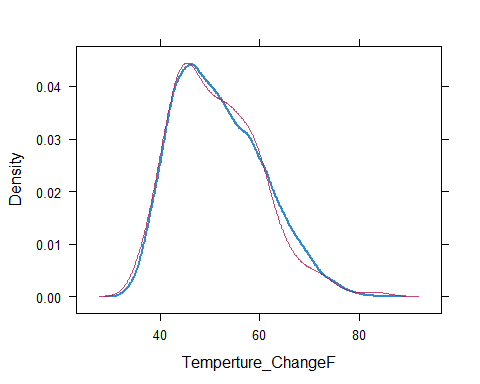
imp\_age = mice(rain\_subset, m=1, method='pmm', printFlag=FALSE)

## Warning: Number of logged events: 21

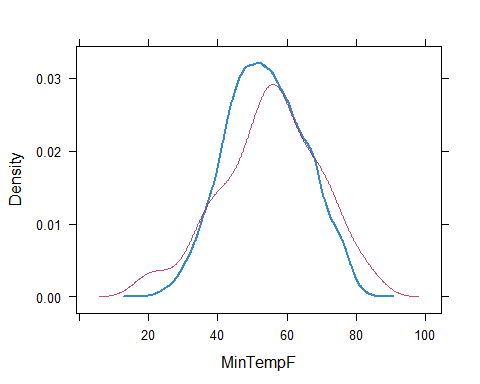
summary(imp\_age)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## Date MinTempF MaxTempF   
## "" "pmm" "pmm"   
## Rainfall RainTomorrow RainToday   
## "pmm" "" "pmm"   
## Temperture\_ChangeF Humidity9am Humidity3pm   
## "pmm" "pmm" "pmm"   
## PredictorMatrix:  
## Date MinTempF MaxTempF Rainfall RainTomorrow RainToday  
## Date 0 1 1 1 1 1  
## MinTempF 0 0 1 1 1 1  
## MaxTempF 0 1 0 1 1 1  
## Rainfall 0 1 1 0 1 1  
## RainTomorrow 0 1 1 1 0 1  
## RainToday 0 1 1 1 1 0  
## Temperture\_ChangeF Humidity9am Humidity3pm  
## Date 1 1 1  
## MinTempF 1 1 1  
## MaxTempF 1 1 1  
## Rainfall 1 1 1  
## RainTomorrow 1 1 1  
## RainToday 1 1 1  
## Number of logged events: 21   
## it im dep meth out  
## 1 0 0 constant Date  
## 2 1 1 Rainfall pmm MaxTempF  
## 3 1 1 RainToday pmm MaxTempF  
## 4 1 1 Humidity9am pmm MaxTempF  
## 5 1 1 Humidity3pm pmm MaxTempF  
## 6 2 1 Rainfall pmm MaxTempF

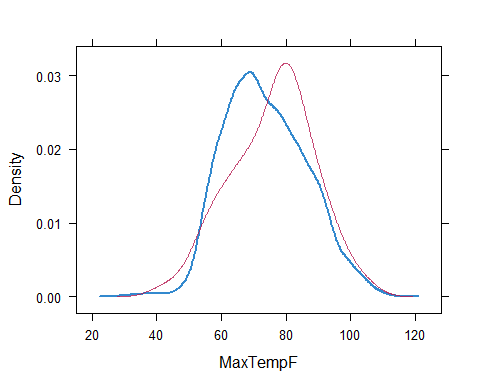
densityplot(imp\_age, ~Temperture\_ChangeF) #red imputed, blue original



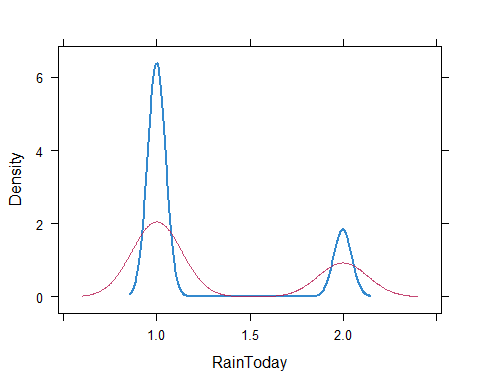
densityplot(imp\_age, ~MinTempF) #red imputed, blue original



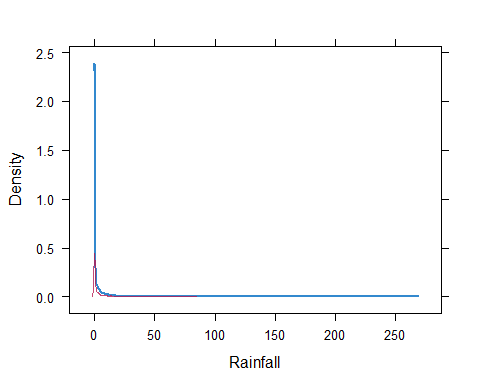
densityplot(imp\_age, ~MaxTempF) #red imputed, blue original



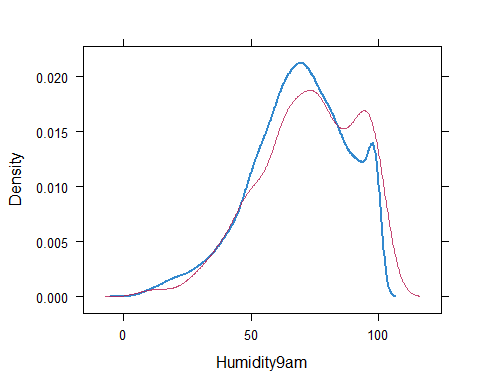
densityplot(imp\_age, ~RainToday) #red imputed, blue original



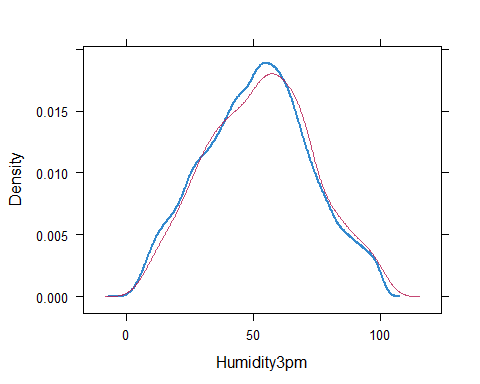
densityplot(imp\_age, ~Rainfall) #red imputed, blue original



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



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



Merge the imputed values into our titanic data frame

rain\_complete = complete(imp\_age)   
summary(rain\_complete)

## Date MinTempF MaxTempF Rainfall   
## Length:28003 Min. :16.70 Min. : 26.60 Min. : 0.000   
## Class :character 1st Qu.:45.68 1st Qu.: 64.22 1st Qu.: 0.000   
## Mode :character Median :53.60 Median : 72.68 Median : 0.000   
## Mean :53.90 Mean : 73.73 Mean : 2.283   
## 3rd Qu.:62.24 3rd Qu.: 82.76 3rd Qu.: 0.800   
## Max. :86.90 Max. :116.60 Max. :268.600   
## RainTomorrow RainToday Temperture\_ChangeF Humidity9am   
## No :21713 No :21728 Min. :32.00 Min. : 1.0   
## Yes: 6290 Yes: 6275 1st Qu.:44.96 1st Qu.: 57.0   
## Median :50.72 Median : 70.0   
## Mean :51.83 Mean : 68.9   
## 3rd Qu.:57.92 3rd Qu.: 83.0   
## Max. :86.00 Max. :100.0   
## Humidity3pm   
## Min. : 0.00   
## 1st Qu.: 37.00   
## Median : 52.00   
## Mean : 51.58   
## 3rd Qu.: 66.00   
## Max. :100.00

set.seed(12345)  
train.rows = createDataPartition(y = rain$RainTomorrow, p=0.7, list = FALSE) #70% in training  
train = rain[train.rows,]   
test = rain[-train.rows,]