## Course Project

### Hackett, Evan

library(tidyverse)

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

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

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

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

library(VIM)

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

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

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

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

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

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

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

library(mice)

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

## 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(GGally)

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

##   
## Attaching package: 'GGally'

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

library(ggplot2)

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

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

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

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

str(rain)

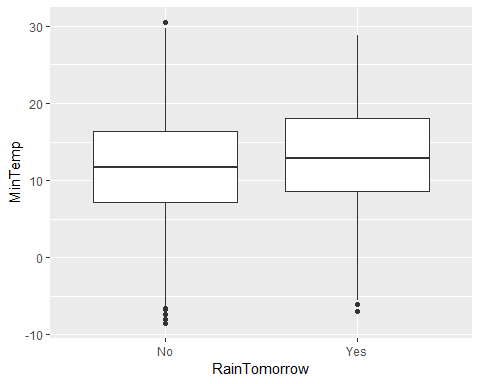
## Classes 'tbl\_df', 'tbl' and 'data.frame': 28003 obs. of 20 variables:  
## $ Date : chr "12/5/2008" "12/6/2008" "12/16/2008" "12/17/2008" ...  
## $ MinTemp : num 17.5 14.6 9.8 14.1 20.5 20.1 9.6 14 12.5 17.4 ...  
## $ MaxTemp : num 32.3 29.7 27.7 20.9 31.8 32.7 23.9 28.3 28.4 43 ...  
## $ Rainfall : num 1 0.2 NA 0 0 0 0 0 0 0 ...  
## $ WindGustDir : chr "W" "WNW" "WNW" "ENE" ...  
## $ WindGustSpeed: int 41 56 50 22 41 48 41 48 37 39 ...  
## $ WindDir9am : chr "ENE" "W" NA "SSW" ...  
## $ WindDir3pm : chr "NW" "W" "WNW" "E" ...  
## $ WindSpeed9am : int 7 19 NA 11 19 13 19 17 20 7 ...  
## $ WindSpeed3pm : int 20 24 22 9 20 30 11 24 9 17 ...  
## $ Humidity9am : int 82 55 50 69 54 56 44 43 38 40 ...  
## $ Humidity3pm : int 33 23 28 82 24 15 22 15 16 8 ...  
## $ Pressure9am : num 1011 1009 1013 1012 1008 ...  
## $ Pressure3pm : num 1006 1005 1010 1010 1006 ...  
## $ Cloud9am : int 7 NA 0 8 NA NA NA NA NA NA ...  
## $ Cloud3pm : int 8 NA NA 1 NA NA NA NA NA NA ...  
## $ Temp9am : num 17.8 20.6 17.3 17.2 23.8 24.6 14.9 17.9 17.2 25.6 ...  
## $ Temp3pm : num 29.7 28.9 26.2 18.1 30.8 32.1 22.1 27.6 26.6 41.5 ...  
## $ RainToday : Factor w/ 2 levels "0","1": 1 1 NA 1 1 1 1 1 1 1 ...  
## $ RainTomorrow : chr "No" "No" "No" "Yes" ...

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 0 :21525 Length:28003   
## 1st Qu.:12.30 1st Qu.:16.60 1 : 6183 Class :character   
## Median :16.70 Median :21.10 NA's: 295 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

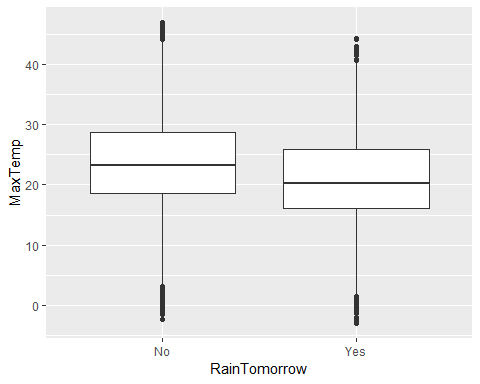
ggplot(rain,aes(x=RainTomorrow,y=MinTemp)) + geom\_boxplot()

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



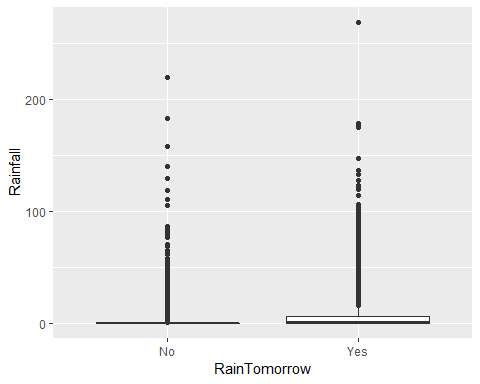
ggplot(rain,aes(x=RainTomorrow, y= MaxTemp)) + geom\_boxplot()

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



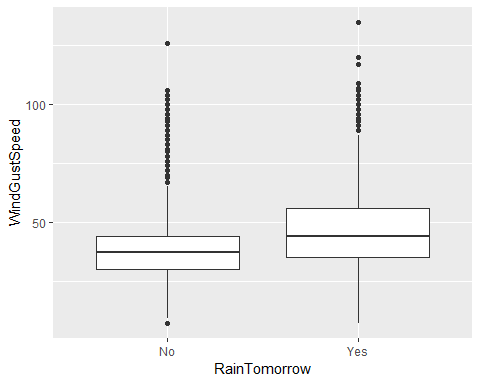
ggplot(rain,aes(x=RainTomorrow, y= Rainfall)) + geom\_boxplot()

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



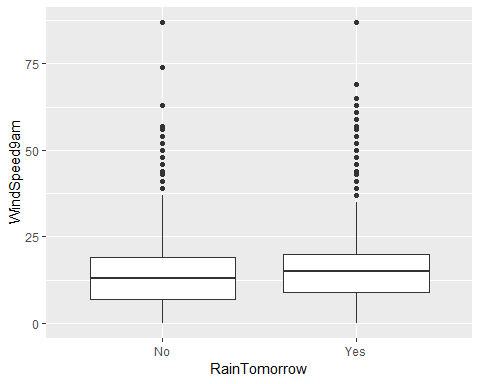
ggplot(rain,aes(x=RainTomorrow, y= WindGustSpeed)) + geom\_boxplot()

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



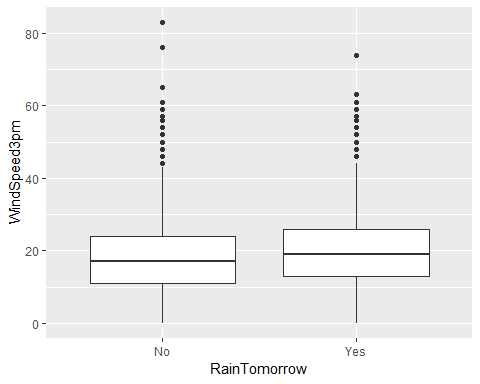
ggplot(rain,aes(x=RainTomorrow, y= WindSpeed9am)) + geom\_boxplot()

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



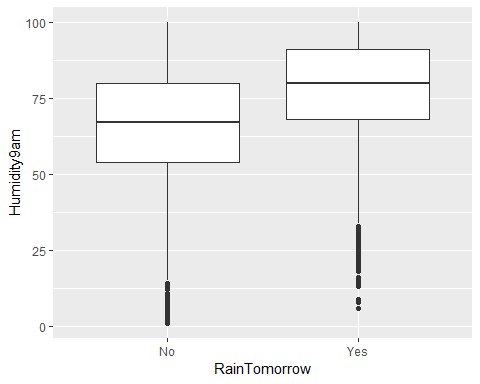
ggplot(rain,aes(x=RainTomorrow, y= WindSpeed3pm)) + geom\_boxplot()

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



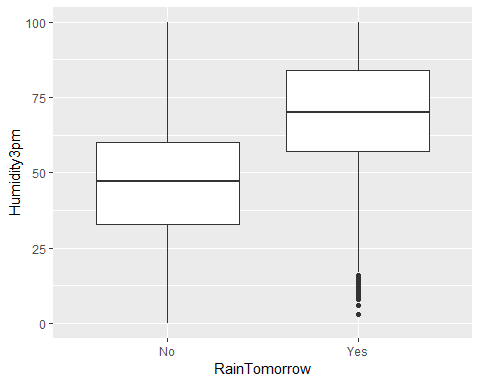
ggplot(rain,aes(x=RainTomorrow, y= Humidity9am)) + geom\_boxplot()

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



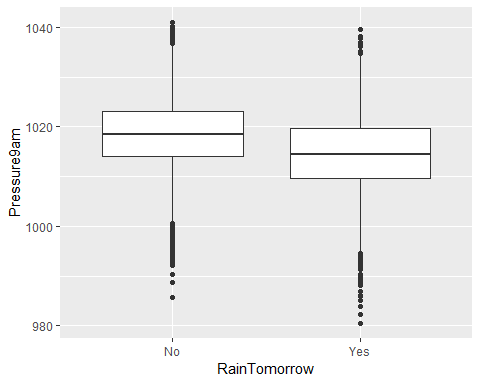
ggplot(rain,aes(x=RainTomorrow, y= Humidity3pm)) + geom\_boxplot()

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



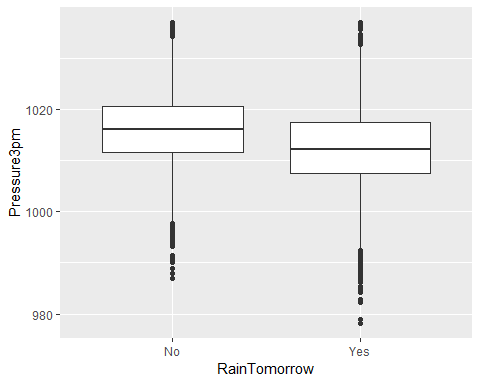
ggplot(rain,aes(x=RainTomorrow, y= Pressure9am)) + geom\_boxplot()

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



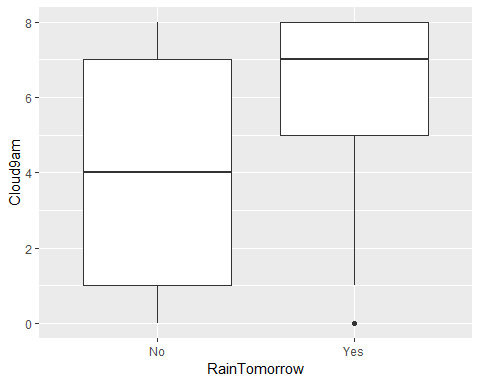
ggplot(rain,aes(x=RainTomorrow, y= Pressure3pm)) + geom\_boxplot()

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



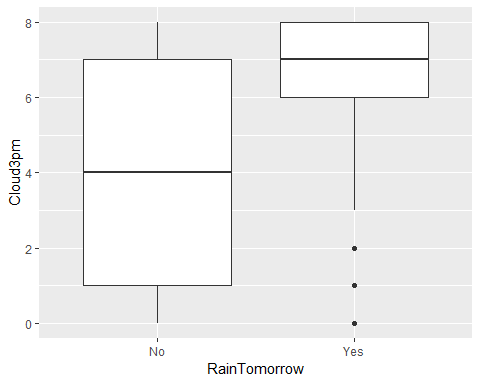
ggplot(rain,aes(x=RainTomorrow, y= Cloud9am)) + geom\_boxplot()

## Warning: Removed 10673 rows containing non-finite values (stat\_boxplot).



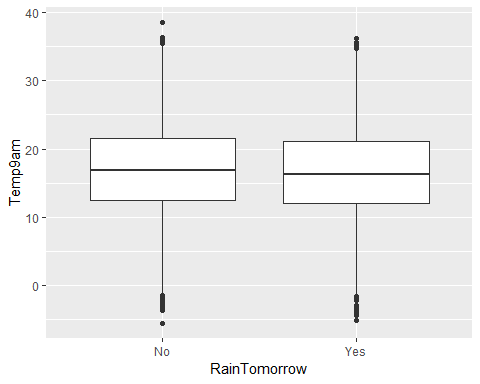
ggplot(rain,aes(x=RainTomorrow, y= Cloud3pm)) + geom\_boxplot()

## Warning: Removed 11341 rows containing non-finite values (stat\_boxplot).



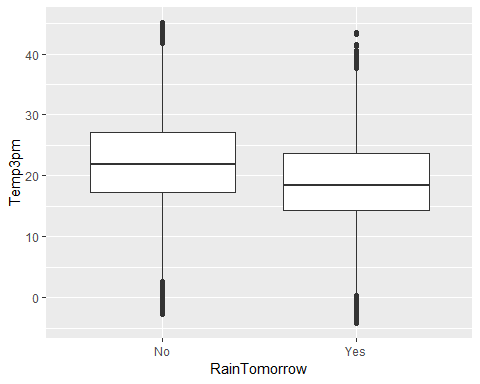
ggplot(rain,aes(x=RainTomorrow, y= Temp9am)) + geom\_boxplot()

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



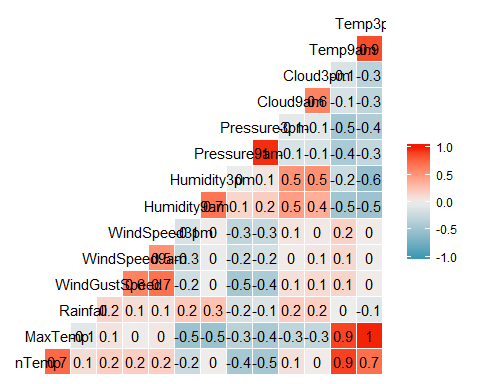
ggplot(rain,aes(x=RainTomorrow, y= Temp3pm)) + geom\_boxplot()

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



ggcorr(rain, label = TRUE)

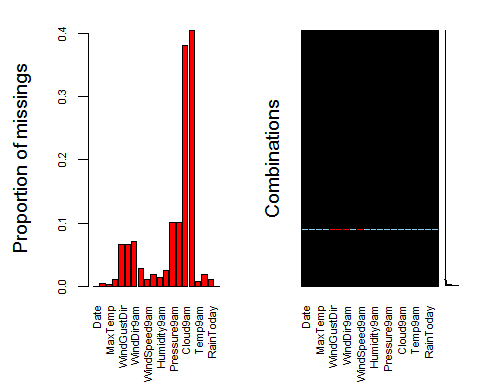
## Warning in ggcorr(rain, label = TRUE): data in column(s) 'Date',  
## 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', 'RainTomorrow' are  
## not numeric and were ignored



View Missingness

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

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

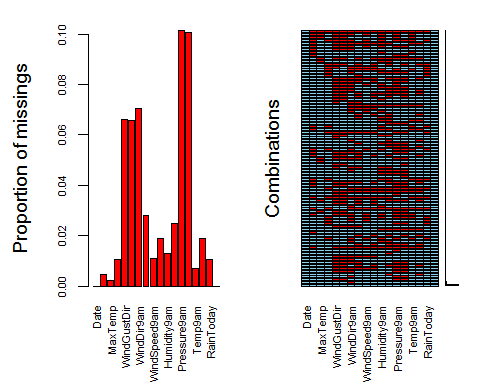


#the cex.axis reduces size of text on x-axis so labels fit better

Deleting out Cloud Coverage variables, deleting due to large number of missing observations.

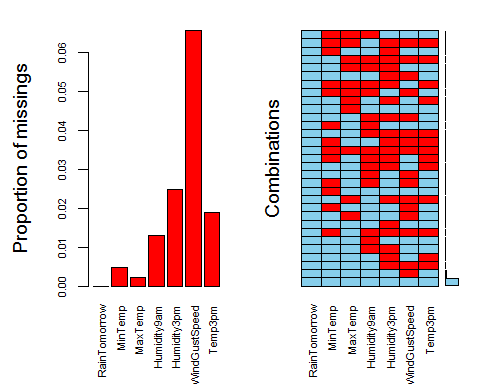
rain = rain %>% select(-Cloud9am, -Cloud3pm)   
vim\_plot = aggr(rain, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

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



#select only variables relevant to our analysis  
rain = rain %>% select(c("RainTomorrow","MinTemp","MaxTemp","Humidity9am","Humidity3pm","WindGustSpeed","Temp3pm"))  
  
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)



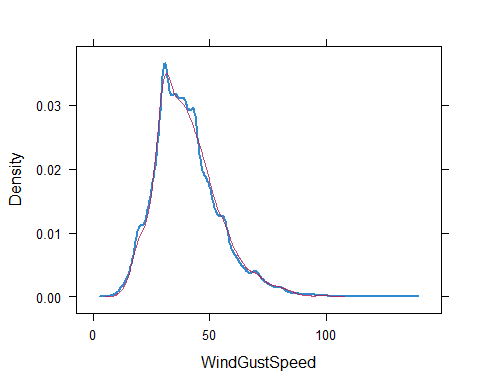
imp\_WindGustSpeed = mice(rain, m=1, method='pmm', printFlag=FALSE)

## Warning: Number of logged events: 1

summary(imp\_WindGustSpeed)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## RainTomorrow MinTemp MaxTemp Humidity9am Humidity3pm   
## "" "pmm" "pmm" "pmm" "pmm"   
## WindGustSpeed Temp3pm   
## "pmm" "pmm"   
## PredictorMatrix:  
## RainTomorrow MinTemp MaxTemp Humidity9am Humidity3pm  
## RainTomorrow 0 1 1 1 1  
## MinTemp 0 0 1 1 1  
## MaxTemp 0 1 0 1 1  
## Humidity9am 0 1 1 0 1  
## Humidity3pm 0 1 1 1 0  
## WindGustSpeed 0 1 1 1 1  
## WindGustSpeed Temp3pm  
## RainTomorrow 1 1  
## MinTemp 1 1  
## MaxTemp 1 1  
## Humidity9am 1 1  
## Humidity3pm 1 1  
## WindGustSpeed 0 1  
## Number of logged events: 1   
## it im dep meth out  
## 1 0 0 constant RainTomorrow

densityplot(imp\_WindGustSpeed, ~WindGustSpeed)



Merge the imputed values into our titanic data frame

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

## RainTomorrow MinTemp MaxTemp Humidity9am   
## Length:28003 Min. :-8.50 Min. :-3.00 Min. : 1.00   
## Class :character 1st Qu.: 7.60 1st Qu.:17.90 1st Qu.: 57.00   
## Mode :character Median :12.00 Median :22.60 Median : 70.00   
## Mean :12.16 Mean :23.18 Mean : 68.92   
## 3rd Qu.:16.80 3rd Qu.:28.20 3rd Qu.: 83.00   
## Max. :30.50 Max. :47.00 Max. :100.00   
## Humidity3pm WindGustSpeed Temp3pm   
## Min. : 0.00 Min. : 7.00 Min. :-4.20   
## 1st Qu.: 36.00 1st Qu.: 31.00 1st Qu.:16.60   
## Median : 52.00 Median : 39.00 Median :21.10   
## Mean : 51.58 Mean : 40.04 Mean :21.67   
## 3rd Qu.: 66.00 3rd Qu.: 48.00 3rd Qu.:26.40   
## Max. :100.00 Max. :135.00 Max. :45.20