1- grafico #01Bilancino_Inizio.jpg per visualizzare la situazione dei missing

2- osservo i missing fino a #2003-12-31", osservo il dataset dopo il taglio
02Bilancino_cut_no_missing

3-#### guardo le variabili target ####
Le mie variabili TARGET sono: Lake_Level, Flow_Rate
-Grafico 03Bilancino_Flow_Rate ##Sembra che ci siano alcuni picchi Flow_Rate a volte a gennaio, a volte in primavera, non tutti gli anni. Il lago Bilancino e' un lago artificiale nel Muggello, che incontra il fiume Sieve (il Sieve e' sia un immissario che un emissario del lago). La media del suo flusso d'acqua e' molto bassa: 2,78 l/s. A volte ci sono dei picchi molto forti e improvvisi ma si torna rapidamente alla normalità'(sicuramente sono picchi pilotati dalla diga).
storia: Dopo l'alluvione di Firenze, nel 1966 si penso' di creare uno sbarramento sul fiume Sieve, in località Bilancino, onde limitare i rischi di alluvione nella piana dell'Arno e sopperire al bisogno idrico e idroelettrico di Firenze e Prato. Il lavori sono iniziati nel 1984, conclusi nel 1996.
La riva del lago, durante la stagione estiva, è balneabile.

-Grafico 04Bilancino_Lake_level
Il Lake_level ha una variabilità molto bassa,
inquanto il livello dell'acqua è sempre compreso tra 243 e 253 m
#nel corso di oltre 16 anni. Un picco verso il basso c'e' stato nel
#2012 / 2013
#ma e' risalito velocemente e si e' stabilizzato poco piu' in basso del livello precedente

- variabili target insieme, grafico 05Bilancino_target.jpg (stessi commenti di prima, da decidere se inserire due grafici, oppure uno solo, comparato)

4- Grafico 06Bilancino_target_season.
#Il livello del lago è più basso in autunno,
#mentre e' più alto in primavera.
#In estate e in inverno la mediana del livello del lago, è di circa 250 metri.
#Per il Flow_rate, le differenze non sono molte.
#Gli outlayers si trovano principalmente in primavera e in inverno.

Grafico 09box_target comparato, non intervengo sugli outliers perché essendo un lago artificiale, le alterazioni dei flussi d'acqua sono sicuramente volute.

5-Correlazione grafico 11Bilancino_correlazione
Commento:
#si può notare che le variabili riguardanti le precipitazioni
#sono fortemente correlate positivamente tra loro, tra 0,8 e 0,9 per queste variabili,
#quindi si potra' fare una riduzione. Le restanti relazioni sono molto
#basse
#quindi la rimozione delle variabili riguarderà solo le precipitazioni

6- Rainfall analysis grafico 10Bilancino_rain
PRECIPITAIZIONI comparazione tra localita'####
#Abbiamo precipitazioni da cinque diverse regioni e
#la loro quantità, varia da 0 a 125 mm. Gli aumenti rapidi si applicano
#a tutte le variabili e I FENOMENI SONO FORTEMENTE CORRELATI.
DECIDO di lasciare 3 variabili, per il processo di modellazione
#scelgo la prima coppia delle regioni "S Piero" e "Mangona", a Cavallina
#perché questi dati sono i meno correlata e quindi ci fornisce maggiori informazioni

7- grafico 12Bilancino_temp
Temperature: the temperature mean is 14.53 °C
#La temperatura alla localita' Le_Croci e' stagionale, tipica di una regione
#del centro Italia.
#La temperatura e' quindi una delle variabili, dipendente, continua.

#Nella maggioranza dei casi è compresa tra 0 e 30 gradi C,
#con stagionalità visibile e connessione con le stagioni.
#Essendo l'unica variabile sulla temperatura,
#la utilizzo sicuramente per creare il modello
#posso pensare di analizzare i casi sotto allo zero, per vedere se sono collegati
#a precipitazioni nevose

8- grafico 13Bilancino_LL_rainfall
notiamo che quando
#i livelli di pioggia sono alti il livello del lago è basso e viceversa.
#Il Lake level scende soprattutto nell'ultima parte dell'anno ma risale a
#dopo gennaio, dopo le piogge; possiamo considerare alcuni mesi di ritardo
#per ripristinare il livello dell'acqua

9- neve
snow
#Temperature sotto lo zero / eventuale pioggia/neve, da analizzare
(bilancino_featured)
```
bilancino_featured <- add.seasons(Lake_Bilancino_cut) %>%
  mutate(snow.yes = as.factor(ifelse(Temperature_Le_Croci <=0 & Rainfall_Le_Croci > 0, 1,0)),
         snow.no = as.factor(ifelse(Temperature_Le_Croci > 0 & Rainfall_Le_Croci <= 0,1,0)))
str(bilancino_featured)
```

10- un esempio su come che i lag di pioggia (se si ripete con altri dataset, non scrivere)
```
bilancino_months <- bilancino_featured %>%
  mutate(lag1 = lag(Rainfall_Le_Croci, +1),
         lag3 = lag(Rainfall_Le_Croci,+3),
         lag5 = lag(Rainfall_Le_Croci,+5),
         lag7 = lag(Rainfall_Le_Croci,+7),
         )
str(bilancino_months)
```

10- RANDOM FOREST
Primo test immagine 16Bilancino_RF
RF con Rainfall_S_Piero, Rainfall_Mangona Rainfall Cavallina - Target Lake Level####
RMSE Test: 1.78
#L'autunno ha maggiore influenza sul modello.
#Ha un impatto sul modello oltre 3 volte maggiore rispetto alla
#seconda variabile più importante, che è la stagione della primavera.
#
#Le piogge dalle regioni incluse,si sono rivelate avere
#il minor impatto sul modello.
```
Sample size of test (predict) data: 2008
               Number of grow trees: 200
  Average no. of grow terminal nodes: 745.68
         Total no. of grow variables: 8
       Resampling used to grow trees: swr
    Resample size used to grow trees: 2008
                            Analysis: RF-R
                              Family: regr
               % variance explained: 33.27
                Test set error rate: 3.18
```
errore RMSE Test: 1.78 e
la varianza spiega circa il 33% dei casi

11- valori reali/predetti grafico 17Bilancino_RF_LakeLevel
(Da migliorare dopo il grafico)

12- grafico 18Bilancino_RF_flowrate

RF con Rainfall_S_Piero, Rainfall_Mangona Rainfall Cavallina - Target Flow_rate####

Sample size of test (predict) data: 2008
Number of grow trees: 200
Average no. of grow terminal nodes: 736.845
Total no. of grow variables: 8
Resampling used to grow trees: swr
Resample size used to grow trees: 2008
Analysis: RF-R
Family: regr
% variance explained: 9.9
Test set error rate: 17.81
#####RMSE Test: 4.22

- 13 - valori reali/predetti grafico grafico 19Bilancino_RF_FL (da migliorare)

14- GBM. - Ho escluso solo le variabili rainfall S.Agaa e rainfall le croci, guardando la matrice di correlazione
####Flow_rate target con GBM dataset originale
Grafico 20Bilancino_Sqerror01

RMSE 3.4211 Flow_RAte GB #### migliora il modello random forest

Controllo l'incidenza delle variabili
A tibble: 6 x 2
var rel.inf
<chr> <dbl>
1 Rainfall_Cavallina 38.6
2 Rainfall_Mangona 21.2
3 Temperature_Le_Croci 19.3
4 Rainfall_S_Piero 13.6
5 Season.Winter 2.56
6 Season.Spring 2.33

Grafico 21flowrate_features.jpg al primo posto trovo la pioggiaCavallina
Grafico 22flow_rate_pred pred/actual
+ controlla questo grafico
23Bilancino_flowrate_pred_act

———-
Lake_level_df GBM
Grafico 24lake_level_GB vedo dove tagliare i rami
RMSE 1.5908 Lake_Level GB

Controllo l'incidenza delle variabili:
A tibble: 6 x 2
var rel.inf
<chr> <dbl>
1 Season.Autumn 50.2
2 Temperature_Le_Croci 23.7
3 Season.Winter 8.17
4 Rainfall_Mangona 5.71
5 Rainfall_S_Piero 5.00
6 Rainfall_Cavallina 4.99
Grafico 25lake_level_features per vedere l'incidenza delle top 6 variabili

plot predicted vs actual
Grafico 26lake_level_pred oppure
27Bilancino_lakelevel_pred_act. (Ricontrollo I grafici)