EndpointValuesPlots.R

US16120

Mon Nov 26 10:55:53 2018

library(rcdk)

## Loading required package: rcdklibs

## Loading required package: rJava

library(tidyverse)

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.8  
## 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()  
## x dplyr::matches() masks rcdk::matches()

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(purrr)  
library(stringr)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

library(corrplot)

## corrplot 0.84 loaded

library(ggplot2)  
library(ggthemes)  
library(gridExtra)

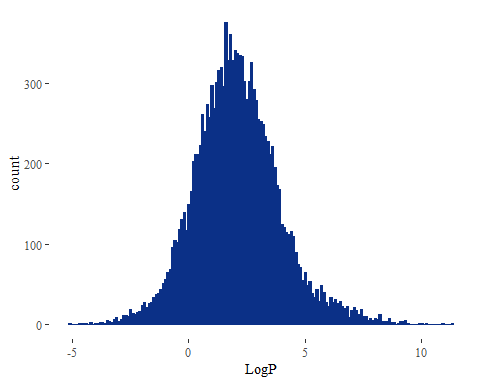
##   
## Attaching package: 'gridExtra'

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

# read data  
  
## LogP  
  
## training data  
train <-  
 read.csv('cache/TR\_LogP\_10537\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(LogP, everything()) %>%  
 na.omit()  
  
## test data  
test <-  
 read.csv('cache/TST\_LogP\_3513\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(LogP, everything()) %>%  
 na.omit()  
  
LogP <- rbind(train, test)  
  
summary(LogP$LogP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -5.08 0.89 2.00 2.08 3.16 11.29

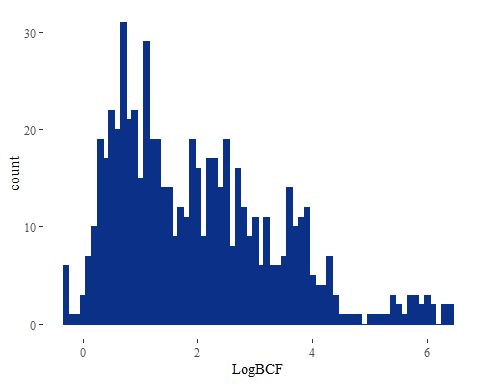
p1 <- ggplot(LogP, aes(LogP)) +  
 geom\_histogram(binwidth = 0.1, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p1



## BCF  
  
train <-  
 read.csv('cache/TR\_BCF\_469.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(ChemID, LogBCF)  
  
## test data  
test <-  
 read.csv('cache/TST\_BCF\_157.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(ChemID, LogBCF)  
  
BCF <- rbind(train, test)  
  
summary(BCF$LogBCF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.350 0.850 1.780 2.002 2.857 6.430

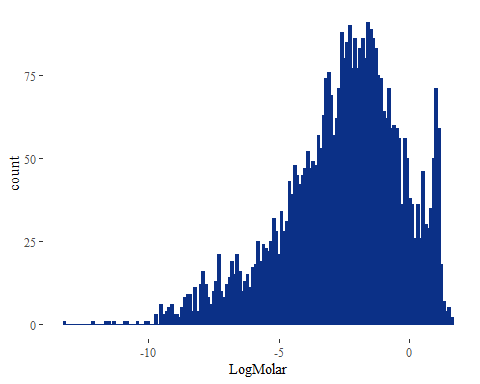
p6 <- ggplot(BCF, aes(LogBCF)) +  
 geom\_histogram(binwidth = 0.1, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p6



## LogS  
  
## training data  
train <-  
 read.csv('cache/TR\_WS\_3158\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(LogMolar, everything()) %>%  
 na.omit()  
  
## test data  
test <-  
 read.csv('cache/TST\_WS\_1066\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(LogMolar, everything()) %>%  
 na.omit()  
  
LogS <- rbind(train, test)  
  
summary(LogS$LogMolar)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -13.172 -3.877 -2.284 -2.573 -0.988 1.581

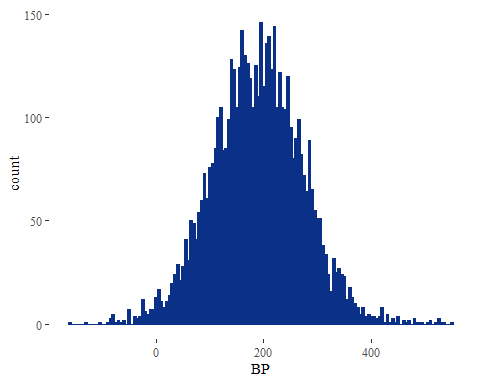
p2 <- ggplot(LogS, aes(LogMolar)) +  
 geom\_histogram(binwidth = 0.1, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p2



## Boiling Point  
  
train <-  
 read.csv('cache/TR\_BP\_4077\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(BP, everything()) %>%  
 na.omit()  
  
## test data  
test <-  
 read.csv('cache/TST\_BP\_1358\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(BP, everything()) %>%  
 na.omit()  
  
BP <- rbind(train, test)  
  
summary(BP$BP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -161.5 133.0 189.3 188.9 245.0 548.0

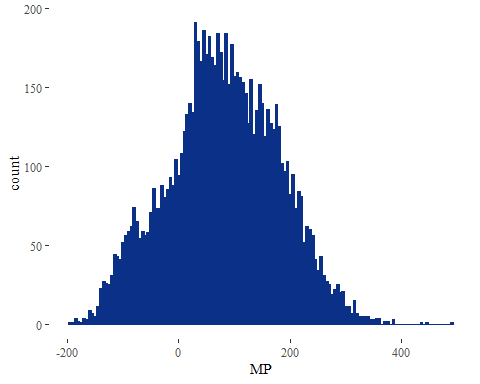
p3 <- ggplot(BP, aes(BP)) +  
 geom\_histogram(binwidth = 5, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p3



## Melting Point  
  
train <-  
 read.csv('cache/TR\_MP\_6486\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(MP, everything()) %>%  
 na.omit()  
  
## test data  
test <-  
 read.csv('cache/TST\_MP\_2167\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X,-CAS,-ROMol,-SMILES,-ID) %>%  
 select(MP, everything()) %>%  
 na.omit()  
  
MP <- rbind(train, test)  
  
summary(MP$MP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -196.00 16.00 80.00 80.45 151.20 492.50

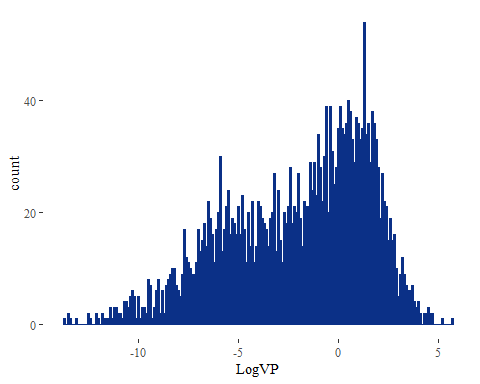
p4 <- ggplot(MP, aes(MP)) +  
 geom\_histogram(binwidth = 5, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p4



## Vapor Pressure  
  
train <-  
 read.csv('cache/TR\_VP\_2034\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X, -CAS, -ROMol, -SMILES, -ID) %>%  
 select(LogVP, everything()) %>%  
 na.omit()  
  
## test data  
test <-  
 read.csv('cache/TST\_VP\_679\_descrs.csv',  
 header = TRUE,  
 stringsAsFactors = FALSE) %>%  
 select(-X, -CAS, -ROMol, -SMILES, -ID) %>%  
 select(LogVP, everything()) %>%  
 na.omit()  
  
VP <- rbind(train, test)  
  
summary(VP$LogVP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -13.6778 -4.7696 -1.2573 -2.0395 0.8633 5.6682

p5 <- ggplot(VP, aes(LogVP)) +  
 geom\_histogram(binwidth = 0.1, color = '#0B3087', fill = '#0B3087') +  
 ggthemes::theme\_tufte()  
p5



gridExtra::grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2)

