ODSC Notebook

Load libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```
## Warning: package 'tibble' was built under R version 3.6.1
```

```
## Warning: package 'tidyr' was built under R version 3.6.1
```

```
## Warning: package 'dplyr' was built under R version 3.6.1
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
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(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

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

```
library(ggplot2)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.6.1
```

```
library(jtools)
```

```
## Warning: package 'jtools' was built under R version 3.6.1
```

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

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

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

read data

```
df <-
  read.csv('data/water_solubility.csv',
           header = TRUE,
           stringsAsFactors = FALSE) %>%
  na.omit()
```

review input data

```
head(df[sample(nrow(df), 10), ])
```

```
##              CAS LogMolar BalabanJ   BertzCT      Chi0    Chi0n     Chi0v
## 3051 110235-47-7  -4.8575 2.289333 565.01146 12.087576 9.858529  9.858529
## 4090  60145-22-4  -8.1199 2.733183 597.20008 13.447229 8.577188 13.112762
## 3497    148-56-1  -0.8186 2.890827 812.10423 15.344935 9.339998 10.972991
## 1031    147-85-3   0.1483 2.349937  96.60179  5.983128 4.554132  4.554132
## 474      95-50-1  -2.9742 3.134862 162.63834  5.983128 4.065330  5.577188
```

```
## 222      75-43-4  -0.7383 2.323790  10.75489  3.577350 1.711244  3.223102
##          Chi1     Chi1n    Chi1v     Chi2n    Chi2v     Chi3n    Chi3v
## 3051 8.275188 5.3824611 5.382461 3.5841984 3.584198 2.1372143 2.137214
## 4090 8.396755 4.6932945 6.961081 3.4905981 5.963035 2.2861831 4.135442
## 3497 8.927487 4.7617127 7.748095 3.7412079 6.871830 2.3831406 4.610484
## 1031 3.804530 2.7668829 2.766883 1.9871762 1.987176 1.3584089 1.358409
## 474  3.804530 2.2053147 2.961244 1.4141088 2.228509 0.8243726 1.580281
## 222  1.732051 0.6546537 1.527525 0.2474358 1.237179 0.0000000 0.000000
##          Chi4n     Chi4v EState_VSA1 EState_VSA10 EState_VSA11 EState_VSA2
## 3051 1.4671410 1.4671410    0.000000     0.000000            0    0.000000
## 4090 1.5136535 3.1755609    0.000000     0.000000            0    0.000000
## 3497 1.5663639 3.3718355   47.264686    30.006839            0    0.000000
## 1031 0.8953267 0.8953267    5.969305     4.794537            0    6.041841
## 474  0.4392946 0.7107613    0.000000     0.000000            0    0.000000
## 222  0.0000000 0.0000000    5.090039     4.390415            0    0.000000
##      EState_VSA3 EState_VSA4 EState_VSA5 EState_VSA6 EState_VSA7
## 3051    5.948339    17.07524           0    6.923737    43.32194
## 4090   41.262703     0.00000           0   24.265468     0.00000
## 3497   18.471269     0.00000           0    0.000000     0.00000
## 1031    0.000000    19.38640           0    0.000000     0.00000
## 474    10.045267     0.00000           0   12.132734    12.13273
## 222     0.000000     0.00000           0    0.000000     0.00000
##      EState_VSA8 EState_VSA9 ExactMolWt FpDensityMorgan1 FpDensityMorgan2
## 3051   27.125614    0.000000   223.1109         1.117647         1.882353
## 4090    0.000000   69.605639   357.8444         0.500000         1.000000
## 3497    9.714500    5.138974   328.9752         1.250000         1.850000
## 1031    5.316789    5.106527   115.0633         1.750000         2.500000
## 474     0.000000   23.201880   145.9690         0.875000         1.250000
## 222    23.201880    0.000000   101.9439         1.500000         1.500000
##      FpDensityMorgan3 FractionCSP3 HallKierAlpha HeavyAtomCount
## 3051         2.588235    0.1428571         -2.34             17
## 4090         1.555556    0.0000000          0.18             18
## 3497         2.350000    0.1250000         -1.66             20
## 1031         2.750000    0.8000000         -0.57              8
## 474          1.500000    0.0000000         -0.20              8
## 222          1.500000    1.0000000          0.51              4
##      HeavyAtomMolWt          Ipc    Kappa1   Kappa2     Kappa3 LabuteASA
## 3051        210.175  8697.098936 11.154545 5.129293  3.0539481 100.93665
## 4090        356.850  8811.616990 14.586228 5.663674  2.8482948 134.30809
## 3497        323.233 14085.917649 14.742956 4.165067  2.4844717 110.95869
## 1031        106.060    93.102593  5.564590 2.132000  0.9795772  47.69881
## 474         142.972    72.495266  5.928205 2.381841  1.1456790  58.03794
## 222         101.915     3.245112  4.510000 1.794900 22.0032718  33.51132
##      MaxAbsEStateIndex MaxAbsPartialCharge MaxEStateIndex MaxPartialCharge
## 3051          4.322769          0.32414323       4.322769       0.22810473
## 4090          6.129825          0.08423255       6.129825       0.06070626
## 3497         12.810456          0.41729960      12.810456       0.41729960
## 1031         10.139074          0.48008077      10.139074       0.32019188
## 474           5.576728          0.08271320       5.576728       0.05918034
## 222          10.530864          0.24671723      10.530864       0.24671723
##      MinAbsEStateIndex MinAbsPartialCharge MinEStateIndex MinPartialCharge
## 3051         0.5777315          0.22810473      0.5777315      -0.32414323
## 4090         0.3672455          0.06070626      0.3672455      -0.08423255
## 3497         0.3396065          0.34452829     -5.0170285      -0.34452829
```

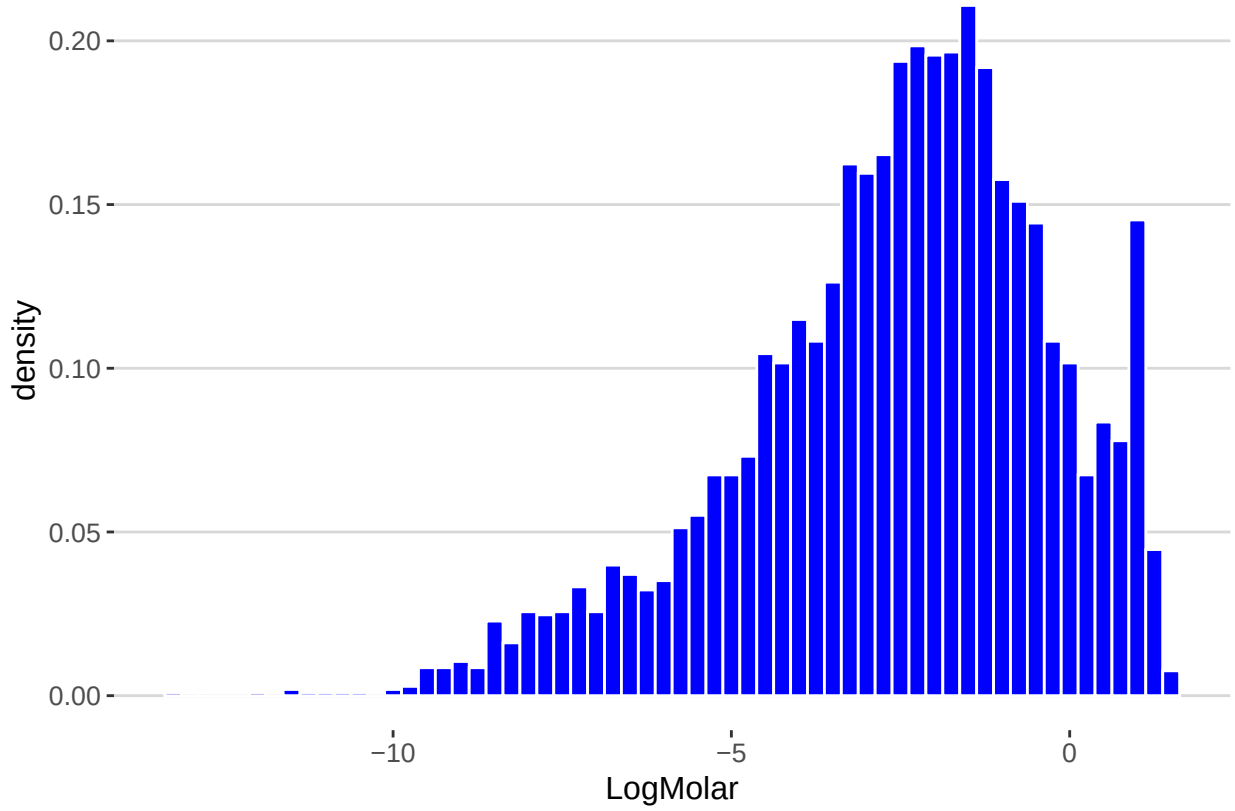
```
## 1031         0.2685185          0.32019188     -0.7199074      -0.48008077
## 474          0.6057099          0.05918034      0.6057099      -0.08271320
## 222          1.7222222          0.21187755     -1.7222222      -0.21187755
##       MolLogP   MolMR   MolWt NHOHCount NOCount NumAliphaticCarbocycles
## 3051  2.90002 69.0457 223.279         1       3                       0
## 4090  7.27400 81.9380 360.882         0       0                       0
## 3497  0.49530 62.1177 329.281         3       7                       0
## 1031 -0.17700 28.6605 115.132         2       3                       0
## 474   2.99340 36.4620 147.004         0       0                       0
## 222   1.71710 16.6020 102.923         0       0                       0
##      NumAliphaticHeterocycles NumAliphaticRings NumAromaticCarbocycles
## 3051                        0                 0                      1
## 4090                        0                 0                      2
## 3497                        1                 1                      1
## 1031                        1                 1                      0
## 474                         0                 0                      1
## 222                         0                 0                      0
##      NumAromaticHeterocycles NumAromaticRings NumHAcceptors NumHDonors
## 3051                       1                2             3          1
## 4090                       0                2             0          0
## 3497                       0                1             5          2
## 1031                       0                0             2          2
## 474                        0                1             0          0
## 222                        0                0             0          0
##      NumHeteroatoms NumRadicalElectrons NumRotatableBonds
## 3051              3                   0                2
## 4090              6                   0                1
## 3497             12                   0                1
## 1031              3                   0                1
## 474               2                   0                0
## 222               3                   0                0
##      NumSaturatedCarbocycles NumSaturatedHeterocycles NumSaturatedRings
## 3051                       0                        0                 0
## 4090                       0                        0                 0
## 3497                       0                        0                 0
## 1031                       0                        1                 1
## 474                        0                        0                 0
## 222                        0                        0                 0
##      NumValenceElectrons PEOE_VSA1 PEOE_VSA10 PEOE_VSA11 PEOE_VSA12
## 3051                  84  5.316789   5.693928          0   5.948339
## 4090                  94  0.000000   0.000000          0   0.000000
## 3497                 110  5.316789  11.234019          0  10.023291
## 1031                  46 10.423316   6.041841          0   0.000000
## 474                   42  0.000000   0.000000          0   0.000000
## 222                   26  0.000000   0.000000          0   5.090039
##      PEOE_VSA13 PEOE_VSA14 PEOE_VSA2 PEOE_VSA3 PEOE_VSA4 PEOE_VSA5
## 3051    0.00000   0.000000  0.000000  9.967957   0.00000  0.000000
## 4090    0.00000   0.000000  0.000000  0.000000   0.00000  0.000000
## 3497   10.02329   6.176299  0.000000 13.556771  21.58904  4.397711
## 1031    0.00000   5.969305  4.794537  0.000000   0.00000  0.000000
## 474     0.00000   0.000000  0.000000  0.000000   0.00000  0.000000
## 222     0.00000   0.000000  0.000000  4.390415   0.00000  0.000000
##      PEOE_VSA6 PEOE_VSA7 PEOE_VSA8 PEOE_VSA9 RingCount  SMR_VSA1 SMR_VSA10
## 3051  24.11954  37.96701  11.38131   0.00000         2  0.000000 11.635726
```

```
## 4090  69.60564  24.26547  16.14954  25.11317         2  0.000000 69.605639
## 3497   0.00000  12.13273   0.00000  16.14632         2 30.006839 32.072504
## 1031   0.00000  19.38640   0.00000   0.00000         1  9.901065  5.969305
## 474   35.33461  12.13273   0.00000  10.04527         1  0.000000 23.201880
## 222   23.20188   0.00000   0.00000   0.00000         0  4.390415 23.201880
##      SMR_VSA2 SMR_VSA3 SMR_VSA4  SMR_VSA5 SMR_VSA6 SMR_VSA7 SMR_VSA8
## 3051        0 9.967957 0.000000 13.847474 5.316789 47.78606        0
## 4090        0 0.000000 0.000000  0.000000 0.000000 54.40127        0
## 3497        0 0.000000 9.536685 15.967265 5.316789 17.69619        0
## 1031        0 5.316789 0.000000 18.883484 6.544756  0.00000        0
## 474         0 0.000000 0.000000  0.000000 0.000000 34.31073        0
## 222         0 0.000000 0.000000  5.090039 0.000000  0.00000        0
##      SMR_VSA9 SlogP_VSA1 SlogP_VSA10 SlogP_VSA11 SlogP_VSA12 SlogP_VSA2
## 3051 11.84087   5.316789   11.635726           0     0.00000   9.967957
## 4090 11.12690   0.000000    0.000000           0    69.60564   0.000000
## 3497  0.00000  10.455762   18.858631           0     0.00000  23.174129
## 1031  0.00000   5.316789    0.000000           0     0.00000  23.662430
## 474   0.00000   0.000000    0.000000           0    23.20188   0.000000
## 222   0.00000   0.000000    4.390415           0    23.20188   5.090039
##      SlogP_VSA3 SlogP_VSA4 SlogP_VSA5 SlogP_VSA6 SlogP_VSA7 SlogP_VSA8
## 3051   0.000000   18.76461  18.311593   36.39820    0.00000     0.0000
## 4090   0.000000    0.00000   0.000000   24.26547   30.13580    11.1269
## 3497  26.222881    0.00000   5.563451   26.32141    0.00000     0.0000
## 1031   4.794537    0.00000  12.841643    0.00000    0.00000     0.0000
## 474    0.000000    0.00000   0.000000   24.26547   10.04527     0.0000
## 222    0.000000    0.00000   0.000000    0.00000    0.00000     0.0000
##      SlogP_VSA9   TPSA VSA_EState1 VSA_EState10 VSA_EState2 VSA_EState3
## 3051          0  37.81           0     0.000000           0           0
## 4090          0   0.00           0    36.086528           0           0
## 3497          0 118.69           0    -9.031870           0           0
## 1031          0  49.33           0     0.000000           0           0
## 474           0   0.00           0    11.153457           0           0
## 222           0   0.00           0     8.746914           0           0
##      VSA_EState4 VSA_EState5 VSA_EState6 VSA_EState7 VSA_EState8
## 3051           0           0           0           0     0.00000
## 4090           0           0           0           0     0.00000
## 3497           0           0           0           0    87.02812
## 1031           0           0           0           0     0.00000
## 474            0           0           0           0     0.00000
## 222            0           0           0           0    10.53086
##      VSA_EState9       qed
## 3051  36.1666667 0.7953660
## 4090   9.9134716 0.4687210
## 3497  -0.7462448 0.7772640
## 1031  23.0000000 0.4982089
## 474    8.4020988 0.5285863
## 222   -1.7222222 0.4091325
```

view distribution of endpoint values; save visualization

```
LogMolar <-
  ggplot(df, aes(LogMolar, stat(density))) +
```

```
  # geom_freqpoly(binwidth = 0.25, size = 1) +
  geom_histogram(binwidth = 0.25, color = 'white', fill = 'blue') +
  theme(legend.position = "none") +
  ggthemes::theme_hc()
LogMolar
```

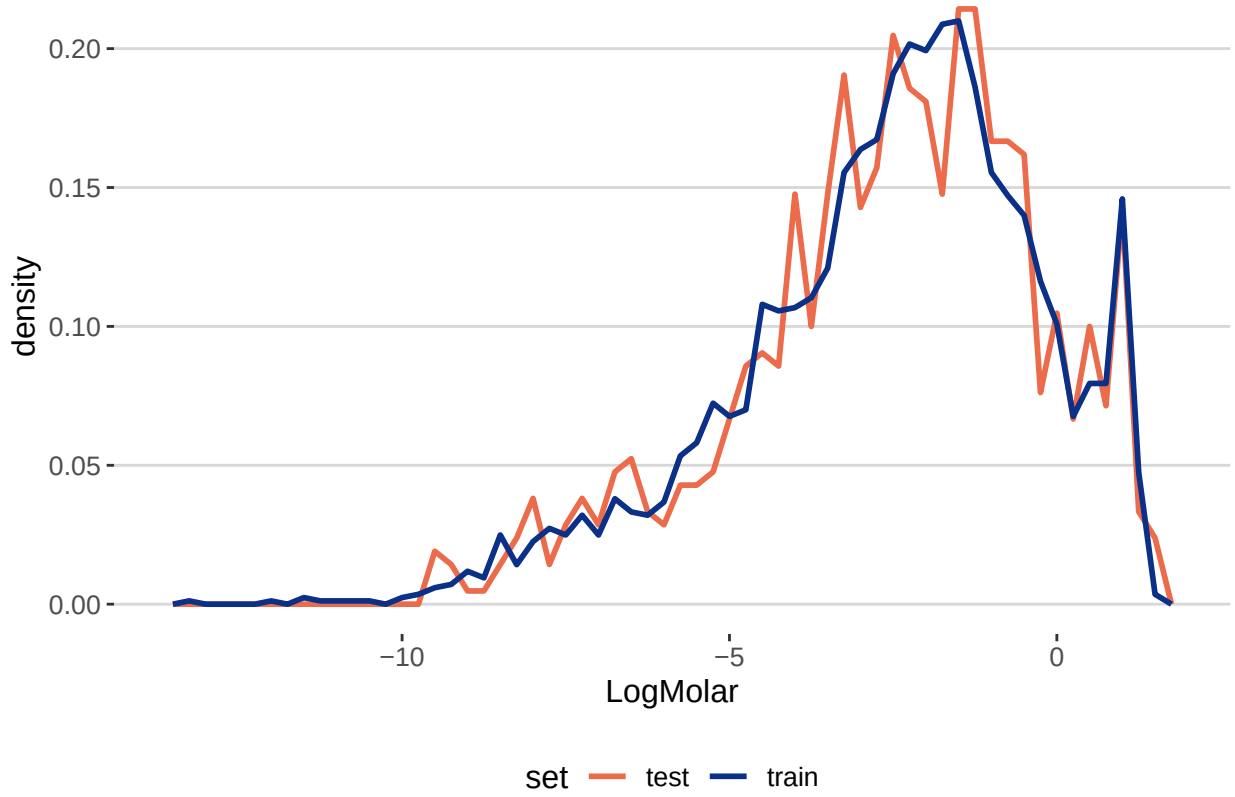



```
ggsave('graphics/WS_LogMolar_Histogram.jpg', plot = LogMolar)
```

```
## Saving 6.5 x 4.5 in image
```

```
inTrain <- caret::createDataPartition(df$LogMolar, p = 0.8, list = FALSE)
train <- df[inTrain, ]
test <- df[-inTrain, ]

X_train <- train[ , 3:ncol(train)]
y_train <- train[ , 2] %>% data.frame()
colnames(y_train) <- 'LogMolar'
X_test <- test[ , 3:ncol(test)]
y_test <- test[ , 2] %>% data.frame()
colnames(y_test) <- 'LogMolar'

TRAIN <- train %>%
  mutate(set = 'train')
TEST <- test %>%
```

```
  mutate(set = 'test')

LogMolar <- rbind(TRAIN, TEST)

LogMolar_train_test <-
  ggplot(LogMolar, aes(LogMolar, stat(density), colour = set)) +
  geom_freqpoly(binwidth = 0.25, size = 1) +
  scale_color_manual(values = c('#EB6B4A', '#0B3087')) +
  theme(legend.position = "none") +
  ggthemes::theme_hc()
LogMolar_train_test
```




```
ggsave('graphics/WS_LogMolar_TrainTest.jpg', plot = LogMolar_train_test)
```

```
## Saving 6.5 x 4.5 in image
```