# Water Solubility

Paul J. Kowalczyk

2019-10-30

# The ODSC Logo

. . . and a link to ODSC West

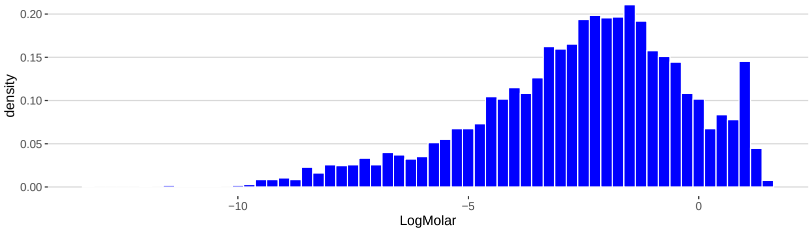
# Read Data

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

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

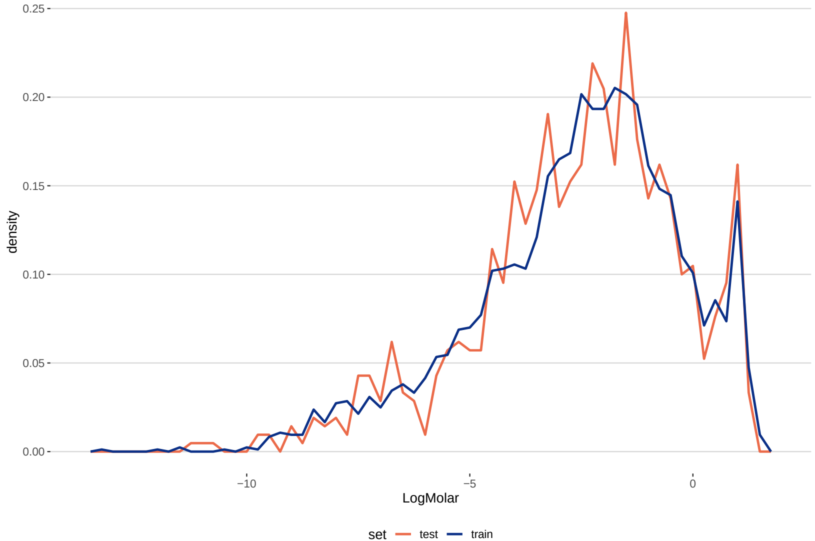
# Distribution of Endpoint Values

```
LogMolar <-
  ggplot(df, aes(LogMolar, stat(density))) +
  geom_histogram(binwidth = 0.25, color = 'white', fill =
  theme(legend.position = "none") +
  ggthemes::theme_hc()
LogMolar
```



# Build training and test sets

- stratified data partition: LogMolar
- 80% train / 20% test



# Data Curation: near-zero variance

Initial number of variables in the dataset: 115.

The variables with near-zero variance are:

```
nzv <- caret::nearZeroVar(X_train, freqCut = 100/0)
names(df[ , nzv])
```

```
## [1] "NumHDonors"   "SMR_VSA6"     "SlogP_VSA7"   "SlogP_
## [5] "VSA_EState10" "VSA_EState3"  "VSA_EState4"  "VSA_ES
```

Remove the near-zero variance variables

```
X_train <- X_train[ , -nzv]
X_test <- X_test[ , -nzv]
```

Number of variables in the dataset, following removal of those with near zero variance: 107

# Data Curation: highly correlated variables

For all pairs of variables whose pairwise correlation exceeds 0.85, remove that variable whose mean correlation to all other variables is the greater.

Identify highly correlated variables

```
allCorrelations <- cor(X_train)
highCorr <- findCorrelation(allCorrelations, cutoff = 0.85)
```

Remove highly correlated variables

```
X_train <- X_train[ , -highCorr]
X_test <- X_test[ , -highCorr]
```

Having removed the highly correlated variables, there are 72 variables remaining.

# Data Curation: names of removed variables (due to high correlation)

```
##  [1] "Chi0"                          "Chi1"
##  [3] "ExactMolWt"                    "HeavyAtomCount"
##  [5] "HeavyAtomMolWt"                "Kappa1"
##  [7] "LabuteASA"                     "MinAbsPartialCharge"
##  [9] "MolMR"                         "MolWt"
## [11] "NumAromaticRings"              "NumHAcceptors"
## [13] "NumHDonors"                    "NumValenceElectrons"
## [15] "SMR_VSA7"                      "VSA_EState10"
## [17] "Chi0n"                         "Chi0v"
## [19] "Chi1n"                         "Chi1v"
## [21] "Chi2n"                         "Chi2v"
## [23] "Chi3n"                         "Chi3v"
## [25] "FpDensityMorgan1"              "FpDensityMorgan2"
## [27] "MaxAbsEStateIndex"             "MaxAbsPartialCharge"
## [29] "Kappa2"                        "NumAliphaticCarbocycles
## [31] "NumAliphaticHeterocycles" "NumAliphaticRings"
```

# Data Curation: Linear combinations

Identify variables that are a linear combination

```
comboInfo <- findLinearCombos(X_train)
names(X_train[ , comboInfo$remove])
```

```
## [1] "NumSaturatedRings" "PEOE_VSA9"          "SlogP_VSA8'
```

Remove those variables that are a linear combination

```
X_train <- X_train[ , -comboInfo$remove]
X_test <- X_test[ , -comboInfo$remove]
```

Having removed variables that are a linear combination, there are 69 variables in the dataset.

# Principal Components Analysis