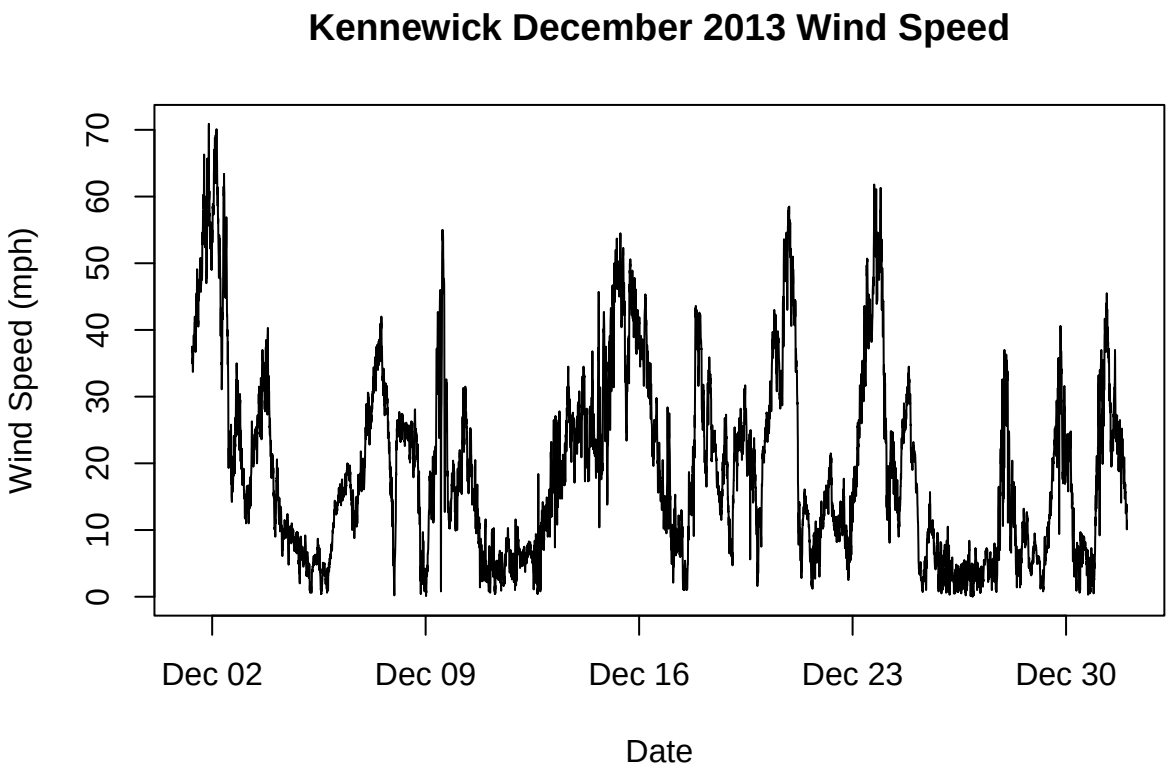
CSI 2300: Intro to Data Science

In-Class Exercise 25: Model Validation

1. Load in the met tower data used in the lecture notes using the code below, and plot the wind speed over time. Describe what you see in the wind speed time series.

```
suppressMessages(library(lubridate))
load("kenn_dec2013.Rda")

kenn_dec2013$Date.Time.UTC <- mdy_hms(kenn_dec2013$Date.Time.UTC)
plot(kenn_dec2013$Date.Time.UTC, kenn_dec2013$Wind.Speed.MPH, type = "l",
     xlab = "Date", ylab = "Wind Speed (mph)",
     main = "Kennewick December 2013 Wind Speed")
```



2. In the lecture, the difference between independent and dependent observations was described. Are the wind speed observations more likely to be independent or dependent? Explain your answer.

The wind speed observations are likely to be dependent. This is because wind speed can be influenced by various factors such as weather patterns, time of day, and seasonal changes.

As a result, the wind speed at one time point may be correlated with the wind speed at another time point, leading to a lack of independence in the observations.

3. If the observations are dependent, then one way to create the train/test sets are to split the data frame in such a way that the order can be preserved.
   - Do this here by using the first 50% of the data frame as the training set and the second half of the data frame as the testing set. - The goal is to model wind speed, and we can't build a model using the dates, so remove the two columns containing dates from both sets.

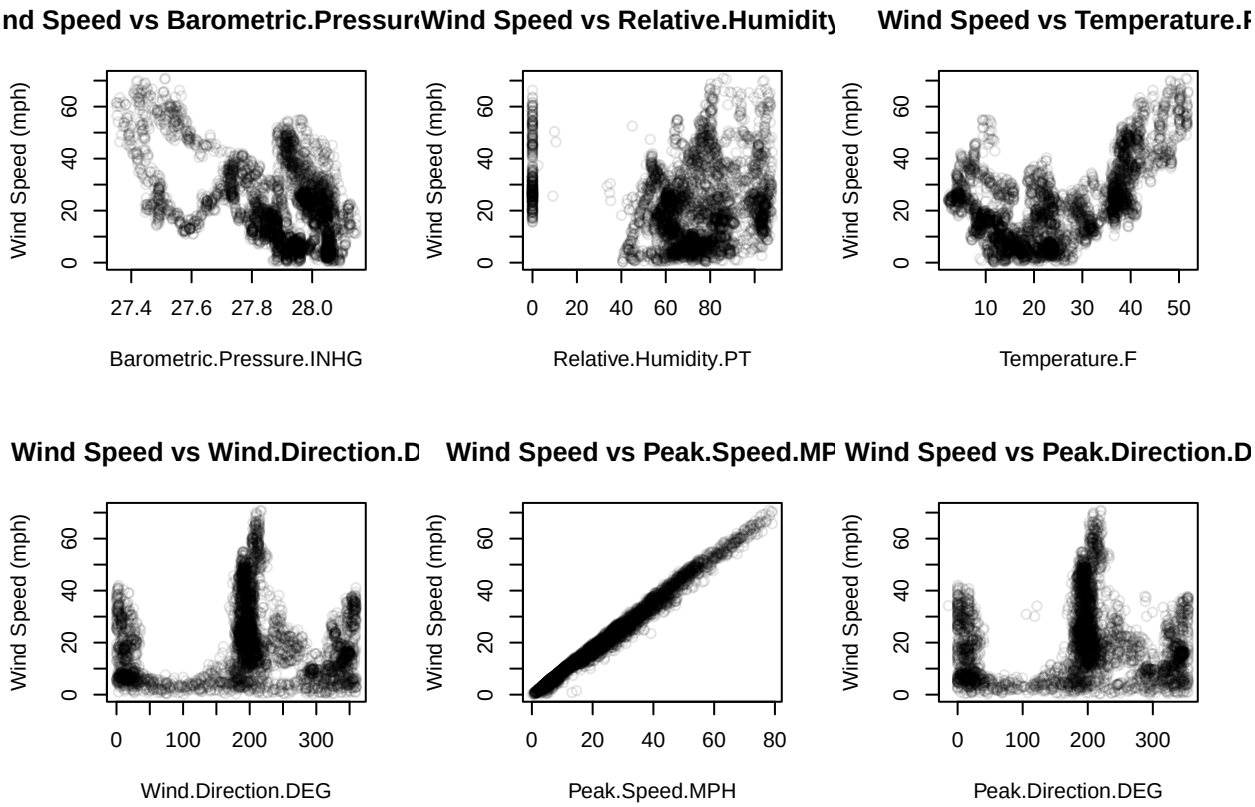
```
train_index <- 1:4416
test_index <- 4417:nrow(kenn_dec2013)

train_frame <- kenn_dec2013[train_index,-c(1,7)]
test_frame <- kenn_dec2013[test_index, -c(1,7)]
```

4. All decisions about modeling wind speed should be made using only the information contained in the training set. Construct pairwise scatterplots of wind speed against each of the other numeric variables in the training data frame. You should have 6 figures, and they can be organized into one panel using the command `par(mfrow=c(2,3))`. Which variables appear to be most strongly related to wind speed?

```
par(mfrow=c(2,3))

for (col in colnames(train_frame)[-5]) {
  plot(train_frame[, col], train_frame$Wind.Speed.MPH, col=rgb(0, 0, 0, .1),
       xlab = col, ylab = "Wind Speed (mph)",
       main = paste("Wind Speed vs", col))
}
```



The barometric pressure and windspeed seem to have a inverse correlation.

5. Perform backward selection and obtain the model based on the BIC criteria (setting `k=log(n)`) and another model based a stonger penalty criteria (setting `k=n`), call it MIC for "my information criteria". Which variables are selected for each model? (See Lecture 21 for reference on backward selection.)

```
m <- lm(Wind.Speed.MPH ~ ., data = train_frame)

bic <- step(m, direction = "backward", k = log(nrow(train_frame)), trace = 0)

summary(bic)
#
# Call:
# lm(formula = Wind.Speed.MPH ~ Barometric.Pressure.INHG + Relative.Humidity.PT +
#     Temperature.F + Wind.Direction.DEG + Peak.Speed.MPH + Peak.Direction.DEG,
#     data = train_frame)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -11.4551  -0.7857   0.1973   0.9636   5.2569
```

```
#
# Coefficients:
#                              Estimate Std. Error t value Pr(>|t|)
# (Intercept)                -8.847e+01  4.449e+00 -19.884  < 2e-16 ***
# Barometric.Pressure.INHG    3.135e+00  1.585e-01  19.778  < 2e-16 ***
# Relative.Humidity.PT        4.382e-03  8.874e-04   4.939 8.16e-07 ***
# Temperature.F               8.661e-03  2.162e-03   4.006 6.27e-05 ***
# Wind.Direction.DEG         -3.551e-03  3.967e-04  -8.952  < 2e-16 ***
# Peak.Speed.MPH              8.982e-01  1.768e-03 508.069  < 2e-16 ***
# Peak.Direction.DEG          2.440e-03  3.956e-04   6.170 7.45e-10 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.501 on 4409 degrees of freedom
# Multiple R-squared:  0.9895,  Adjusted R-squared:  0.9895
# F-statistic: 6.938e+04 on 6 and 4409 DF,  p-value: < 2.2e-16

mic <- step(m, direction = "backward", k = nrow(train_frame), trace = 0)

summary(mic)
#
# Call:
# lm(formula = Wind.Speed.MPH ~ Peak.Speed.MPH, data = train_frame)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -11.0901  -0.7925   0.2353   0.9803   4.8454
#
# Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
# (Intercept)    -0.403330   0.043590  -9.253   <2e-16 ***
# Peak.Speed.MPH  0.883908   0.001458 606.402   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.596 on 4414 degrees of freedom
# Multiple R-squared:  0.9881,  Adjusted R-squared:  0.9881
# F-statistic: 3.677e+05 on 1 and 4414 DF,  p-value: < 2.2e-16
```

The BIC model used the following variables: Barometric.Pressure.INHG Relative.Humidity.PT Temperature.F Wind.Direction.DEG Peak.Speed.MPH Peak.Direction.DEG

The MIC model only used one variable: Peak.Speed.MPH

6. Use the two models fitted on the training dataset to make predictions for the testing

set. Replace the `xx`'s in the table below with the R^2, RMSE, and MAE computed on the testing data for each of the two models from #5. RMSE and MAE can be computed from the residuals as `sqrt(mean(residuals^2))` and `mean(abs(residuals))`, respectively. (See in-class exercises for Lecture 20 if you need to reference making predictions.)

Model	R^2	RMSE	MAE
MIC	.986	1.508	1.083
BIC	.985	1.571	1.123

```
test_mic <- predict(mic, newdata = test_frame)
test_bic <- predict(bic, newdata = test_frame)

cor(test_mic, test_frame$Wind.Speed.MPH)^2
# [1] 0.9868012
cor(test_bic, test_frame$Wind.Speed.MPH)^2
# [1] 0.9857786

errors_mic <- test_frame$Wind.Speed.MPH - test_mic
errors_bic <- test_frame$Wind.Speed.MPH - test_bic

sqrt(mean(errors_mic^2))
# [1] 1.508162
sqrt(mean(errors_bic^2))
# [1] 1.571305

mean(abs(errors_mic))
# [1] 1.083121
mean(abs(errors_bic))
# [1] 1.123443
```

7. Which model makes better predictions in the testing set?

The model with the lower RMSE and MAE values is considered to make better predictions. In this case, we would compare the RMSE and MAE values from both models (MIC and BIC) to determine which one performs better on the testing set. The model with the lower values for both metrics would be preferred.