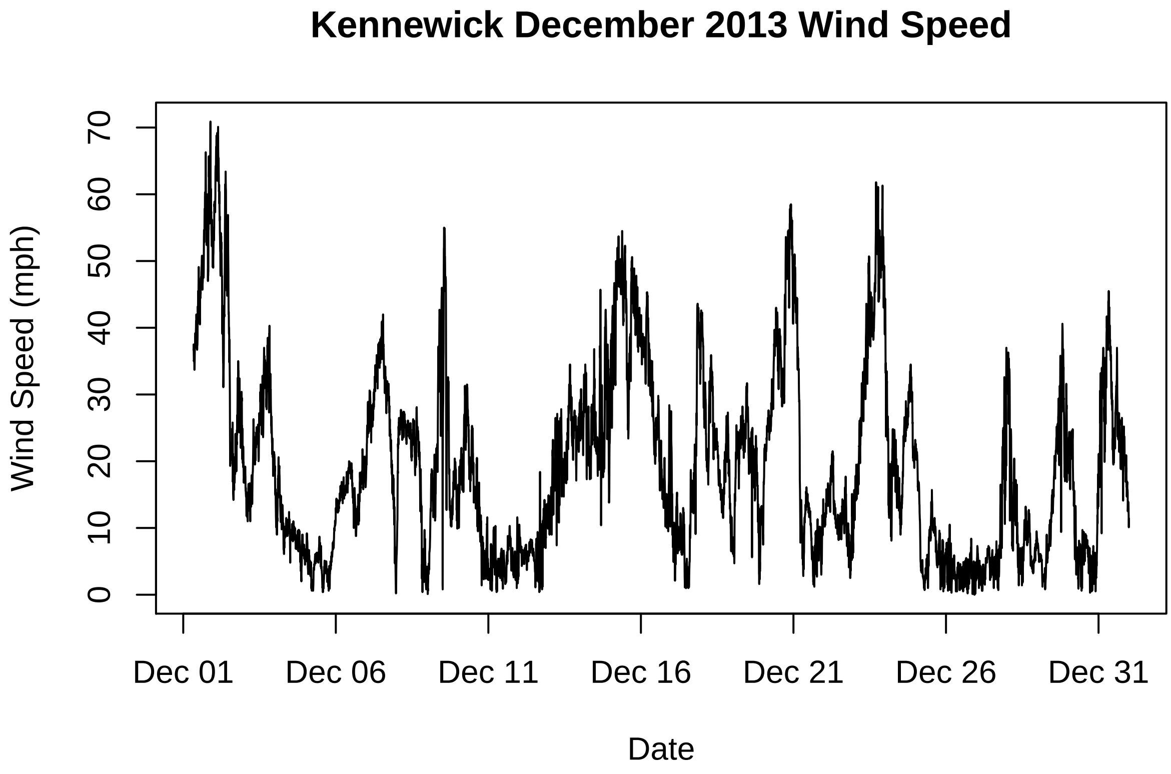
CSI 2300: Intro to Data Science

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.

```r
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.

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.

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?

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.)

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	xx	xx	xx
BIC	xx	xx	xx

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