Analytics Foundations: Problem Set 4

Today's dataset comes from a bike sharing company (Capital Bike Share). Each *hour,* the number of riders (**cnt**) is given, along with various other attributes as shown in the table below:

cnt	count of total rental bikes including both casual and registered
dteday	date
instant	record index (ID)
season	season (1:springer, 2:summer, 3:fall, 4:winter)
yr	year (0: 2011, 1:2012)
mnth	month (1 to 12)
hr	hour (0 to 23)
holiday	whether day is holiday or not
weekday	day of the week
workingday	if day is neither weekend nor holiday is 1, otherwise is 0.
weathersit	- 1: Clear, Few clouds, Partly cloudy, Partly cloudy - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Normalized temperature in Celsius. The values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (max)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users
registered	count of registered users

Source: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

1. Explore the correlation between the numeric variables **temp, atemp, hum,** and **windspeed** with your target **cnt**.

    a. Calculate the correlation coefficient for each pair of variables.

    b. Examine the scatter plot for each pair of variables.

    c. Are there any strong linear relationships between the input variables?

There is a high association between temp and atemp.

2. Do any of the input variables seem like they might be useful predictors the number of riders? Pick one variable that you think might work the best.

    a. Build a linear regression using that variable to predict the total number of hourly riders (**cnt**).

Using the variable temp for a simple linear regression, we have

$$\hat{Y} = -0.04 + 381.29x$$

    b. What is the R^2 for the model you built? Interpret this number in a sentence.

The R^2 value is 0.1638. Approximately 16.38% of the variation in bike rentals can be explained by the normalized temperature in Celsius.

    c. What is the value of the slope for the model you built? Interpret this number in a sentence.

The slope is 381.29. For every 1 degree increase in the normalized temperature in Celsius, we expect the average number of bikes rentals to increase by 381.29.