Analytics Foundations: Breakout Session 6

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:spring, 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. Values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. Values are divided to 50 (max)
hum	Normalized humidity. Values are divided to 100 (max)
windspeed	Normalized wind speed. 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#

Data:

You can obtain the dataset by running the following code:

```
bike <- read.csv('https://raw.githubusercontent.com/IAA-
Faculty/statistical_foundations/master/bike.csv')
```

Questions:

1. You want to build a couple of different models and see which one is better. We will learn in later lectures how to do this with test datasets, but for right now we will only do this with training data. First, we need to split the data into training and test. Run the following code to get the training and test split:

```
set.seed(123)

bike <- bike %>% mutate(id = row_number())

train <- bike %>% sample_frac(0.7)

test <- anti_join(bike, train, by = 'id')
```

   What is the observation count in each dataset (train and test)?

2. You can't decide which variable is better to predict number of users (**cnt**), actual temperature (**temp**) or what the temperature feels like (**atemp**). Since you know they are highly correlated with each other, you shouldn't probably include both in your model. One strategy is to build two models – one with each of the variables and see which predicts better. Build one model with the following variables: actual temperature (**temp**), humidity (**hum**), and wind speed (**windspeed**). What is the Adjusted R-squared for this model? Which variables are significant at the 0.01 level?

3. Build a second model with the following variables: feeling temperature (**atemp**), humidity (**hum**) and wind speed (**windspeed**). What is the Adjusted R-squared for this model? Which variables are significant at the 0.01 level?

4. Ideally, we would compare these models using test data. We will learn how to do this in future labs. For now, which model has the higher adjusted R-squared in the training data?