Analytics Foundations: Breakout Session 5

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 are interested in the combined effect of season (**season**) as well as whether the day is a working day (**workingday**) on the impact to total users (**cnt**). You think there might be an interaction to these – weekends would be impacted differently for different seasons. Test to see if there is an interaction with season and working day to predict user count.

2. The problem sometimes that data scientists face is that we look at things through our own perspective without necessarily taking into account other ones. Do a quick poll in your breakout group. How many in there have used bicycles as their primary transportation method for all life's needs (work, school, groceries, etc.)? Maybe our analysis differs depending on if the bikes were used for primary transportation as compared to casual rentals. What if we separated casual from registered users…

3. With that context, let's test something further. Instead of total count of users (**cnt**), repeat (1) above for both registered users (**registered**) and casual users (**casual**) separately (build two different models). Do either of them have significant interactions between season and working day? If so, slice the data by season to see which one(s) have significant differences between working days and non-working days.