Analytics Foundations: Problem Set 12

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. 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')
```

2. There are abnormal times where the number of casual users is greater than or equal to the number of registered users. You can use the following code to create a variable **casual_high** that captures this:

```
train$casual_high <- train$casual >= train$registered
```

3. You want to know if the occurrence of these times are related to the season of the year. Even though season is ordinal, you want to just test a general association and **not** a linear one so use the Pearson Chi-square test. What do you find at a significance level of 0.001? If you were to perform a Mantel-Haenszel Chi-square test would you reach the same conclusion?

4. You also want to know if the occurrence of these times are related to the whether the day is a holiday or not. Perform the appropriate chi-squared test to test this association as well as an odds ratio for a measure of strength. What do you find at a significance level of 0.001? Interpret the odds ratio.