Case Study 2: Bike Sharing

Abbie M Popa

Instructions

Answer the following questions regarding the bike sharing dataset. The case study will be worth 65 point, and you will lose 2.5 points for each hour it is late. The case study MUST be completed in groups of 2-3 (unless you have an SDS accomodation stating otherwise, of course).

Important 10 points will be awarded based on having a clean, well-organized markdown report.

- 5 pts, file is a knit pdf
- 1 pts, does not print the entire data set at any point
- 2 pts, all code is in code blocks with appropriate line breaks
- 2 pts, all questions are answered in complete sentences outside of code blocks, and which question they are (e.g., quesiton 1, 2, 3. . .) is clearly labelled

Question 1: Set-Up (10 pts)

(a) (1 pt) The bike sharing data set contains two documents, day, and hour, adjust the `R` code to read them into your current `R` session.

```
day <- read.csv("dir/day.csv")
hour <- read.csv("dir/hour.csv")
```

(b) (1 pt) Look at both datasets using `str()`.

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

```
- instant: record index
- dteday : date
- 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 (extracted from http://dchr.dc.gov/page/holiday-schedule)
- 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
- cnt: count of total rental bikes including both casual and registered
```

(c) (5 pts) Notice that some of the columns, e.g., "holiday" are represented as numbers, even though they represent categories (e.g., whether a day is a holidy or not). These columns will likely be easier to use if they are treated as factors. Convert any columns that should be treated categorically (rather than as continuous numbers) to factors.

(d) (3 pts) The dteday column is imported as a factor rather than a date, convert this column to a date.

Question 2 (15 pts)

(a) (5 pts) You want to know if (overall) ridership has increased since the company was founded. Make a graph where the x-axis shows the date, and the y axis shows the total ridership that day. Hint: Think about whether the "day" data or the "hour" data will be easier to use for this purpose. Give your plot a title.

(b) (3 pts) Describe the plot you see. Note when ridership goes up and when ridership goes down. Make at least one hypothesis about why this might be.

(c) (5 pts) You want to know if there are more total riders on a given day depending on the season. Make a plot showing the ridership for each season. Make sure to pick a plot type that works well for categorical data. Also, please make sure your plot is colored to make it easy to read.

(d) (2 points) Describe what you see in the season plot. Does it look like certain seasons are more or less popular. Which ones? Be sure to check the description of the data and state which season using words (i.e., winter, summer, spring, or fall) not just number.

Question 3 (14 pts)

(a) (5 points) You want to know if people are riding their bikes to work. First, make a plot with date on the x axis, total count on the y axis, with workingday coded by color and holiday colored by shape. Make your points large enough so you can see if there is a difference in shape.

(b) (2 points) Do you see a difference here? Why might that be?

(c) (3 points) You think perhaps people who ride their bike to work are more likely to subscribe to the bike share. Make the same plot you just made two more times, once for casual riders and once for registered riders. Make sure to give each plot a title describing which group it represents!

(d) (4 points) What do you notice looking at these two plots? Be sure to describe the pattern you see in the casual riders, and the pattern you see in the registered riders. Based on this information, what do you conclude about people who ride their bikes to work?

Question 4 (16 pts)

(a) (8 points) You think there is an effect of weather on ridership. Make at least 3 plots to investigate this relationship. You should be sure to use the x-axis, the y-axis, and color on at least one of your plots. Your plots should not all be the same type (e.g., don't just make three point plots). Also be sure to consider that weather might have a different effect on casual riders as opposed to registered riders. (Hint, since it's not the same temperature all day long, you may want to look at the hour data.) (4 points for clean, relevant plots, 4 points for written explanation of findings.)

(b) (8 points) You are interested in what time of day riders ride their bikes. You also think this might be affected by weather conditions or time of year. Make at least 3 plots to investigate time of day riders ride their bikes, and how this is affected by other variables. Be sure to use the x-axis, the y-axis, and color on at least one of your plots. Your plots should not all be the same type (e.g., don't just

make three point plots.) Also, be sure to consider how casual riders and registered riders may behave differently. (4 points for clean, relevant plots, 4 points for written explanation of findings.)