Exploratory Data Analysis

Brooke Anderson

February 22, 2016

```
knitr::opts_knit$set(root.dir = "..") # Reset root directory for analysis
library(lubridate) # To help handle dates
library(dplyr) # Data wrangling
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2) # Plotting
```

Read in the data:

```
train <- read.csv("data/train.csv", as.is = TRUE) # `as.is` so `datetime` comes in as
                                                  # character, not factor
test <- read.csv("data/test.csv", as.is = TRUE)
```

How much data?

```
dim(train)
```

```
## [1] 10886    12
```

```
dim(test)
```

```
## [1] 6493    9
```

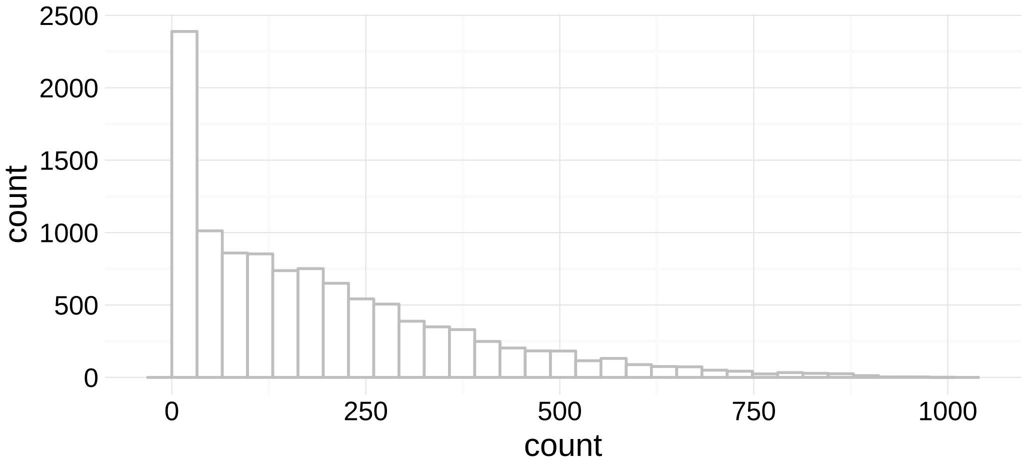
Type of data:

```
str(train)
```

```
## 'data.frame':    10886 obs. of  12 variables:
##  $ datetime  : chr  "2011-01-01 00:00:00" "2011-01-01 01:00:00" "2011-01-01 02:00:00" "2011-01-01 03
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
##  $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
##  $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
##  $ windspeed : num  0 0 0 0 0 ...
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
```

Distribution of response variable:

```
ggplot(train, aes(x = count)) +
  geom_histogram(color = "gray", fill = "white") +
  theme_minimal()
```

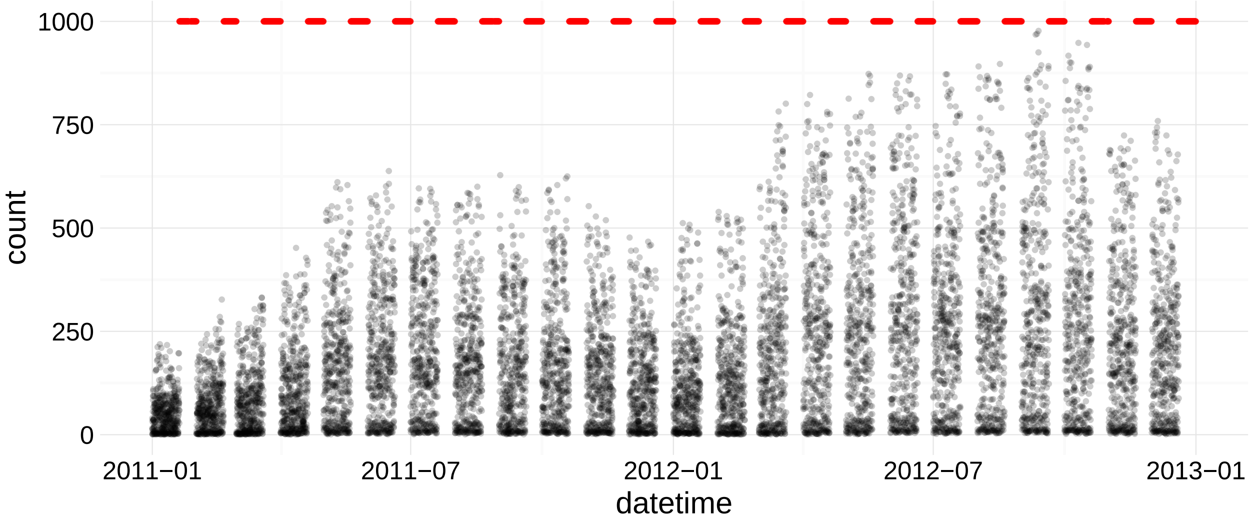


Exploring patterns in bike use by time

Convert `datetime` to the right kind of R object and create columns for months, hours, and day of year of each observation:

```
train <- mutate(train,
                datetime = ymd_hms(datetime),
                hour = hour(datetime),
                month = month(datetime),
                yday = yday(datetime),
                season = factor(season, levels = c(1, 2, 3, 4),
                                labels = c("Spring", "Summer", "Fall", "Winter")),
                workingday = factor(workingday, levels = c(0, 1),
                                    labels = c("Holiday / weekend",
                                               "Working day")))
test  <- mutate(test,
                datetime = ymd_hms(datetime),
                hour = hour(datetime),
                month = month(datetime),
                yday = yday(datetime),
                season = factor(season, levels = c(1, 2, 3, 4),
                                labels = c("Spring", "Summer", "Fall", "Winter")),
                workingday = factor(workingday, levels = c(0, 1),
                                    labels = c("Holiday / weekend",
                                               "Working day")))
```

```
ggplot(train, aes(x = datetime, y = count)) +
  geom_point(alpha = 0.2, size = 0.5) +
  geom_point(aes(y = 1000), data = test, # Plot times of testing data
             color = "red", alpha = 0.2, size = 0.5) +
  theme_minimal()
```



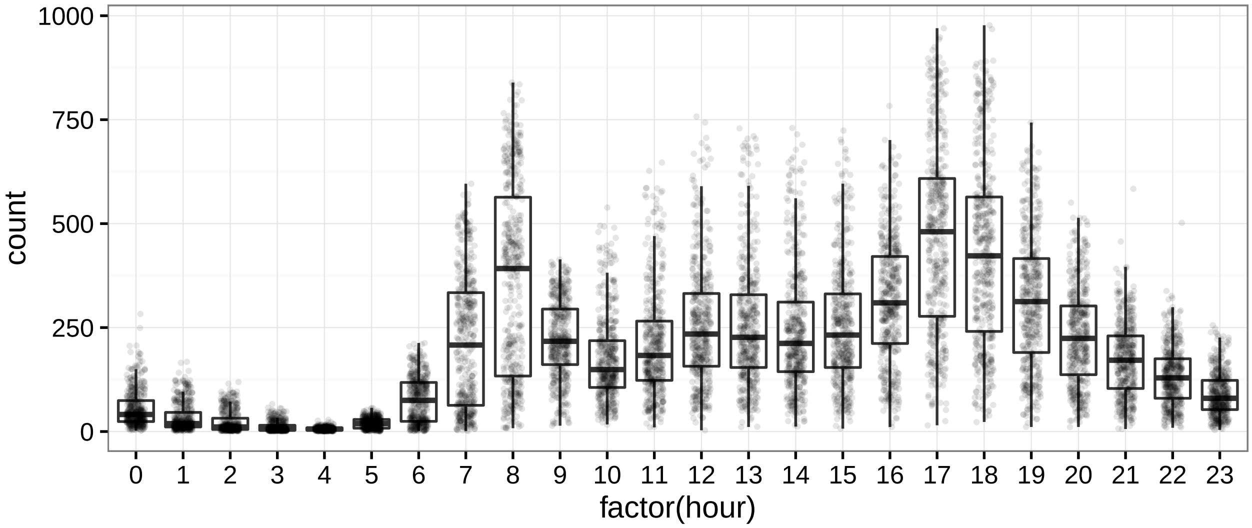
The training observations go from 2011-01-01 to 2012-12-19 23:00:00. The testing observations go from 2011-01-20 to 2012-12-31 23:00:00. The `test` data times are interspersed with the `train` times.

A few things stand out:

- There's a clear seasonal trend, with more bike use in the summer than winter
- There's more variation in bike use in the warmer seasons
- There's an increase trend over the time period in bike use (maybe they made more bikes available or opened more locations between the start and end of the period?)
- There are always observations when few or no bikes are being used. Perhaps this is observations taken during the middle of the night?

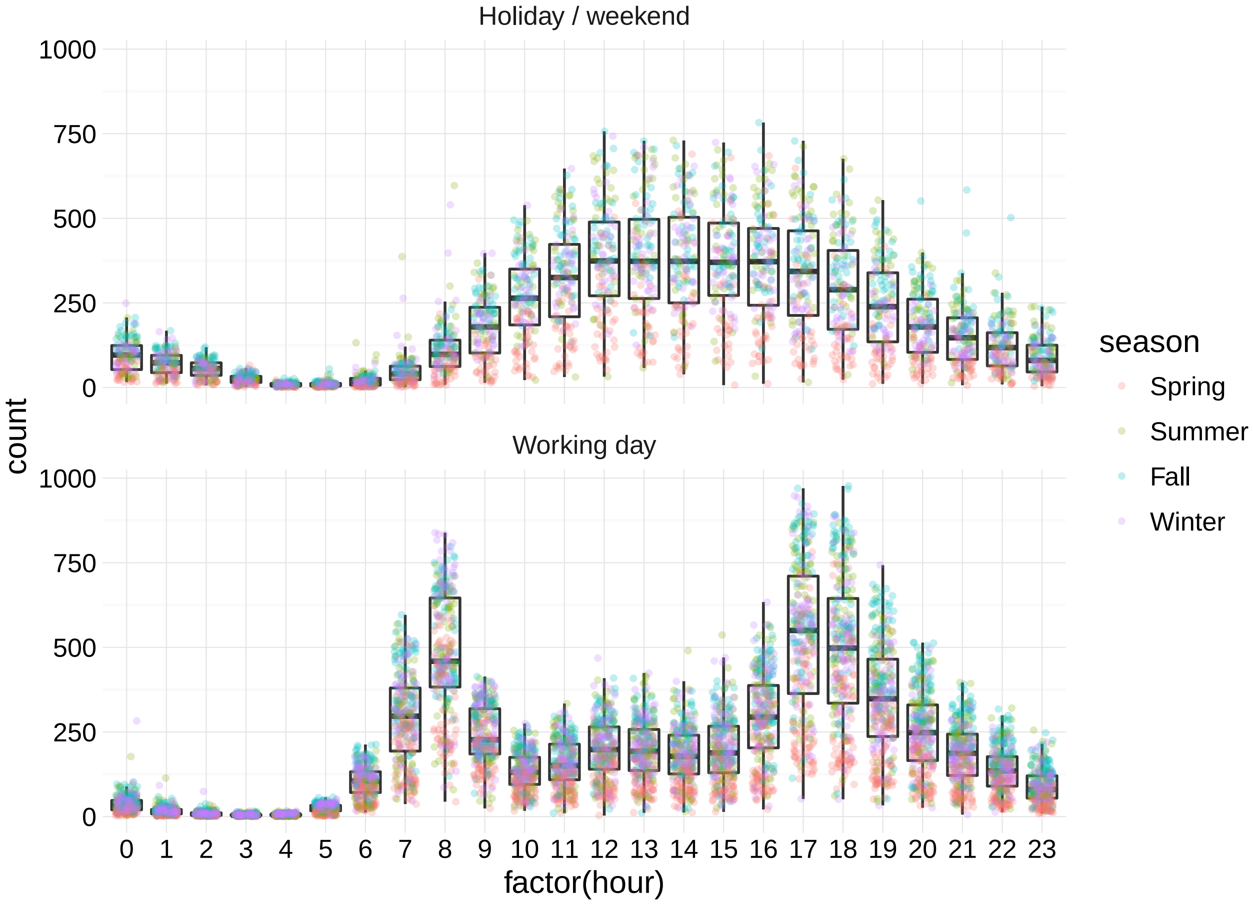
If you look at the counts by hour, it does look like most of the zero or near-zero counts occur between 10:00 pm and 5:00 am. There also seems to be a pretty big pick-up during times when people would commute (7:00 to 9:00 am and 5:00 to 7:00 pm):

```
ggplot(train, aes(x = factor(hour), y = count)) +
  geom_boxplot(outlier.shape = NA) + # Don't plot outliers since I'm overlaying points
  geom_jitter(alpha = 0.1, size = 0.5, width = 0.5) +
  theme_bw()
```



These patterns are pretty different for working days (`workingday` = 1) versus weekends or holidays (`workingday` = 0), which suggests that interactions between `hour` and `workingday` might be useful. An interaction with `season` might also be useful:

```r
ggplot(train, aes(x = factor(hour), y = count)) +
  geom_boxplot(outlier.shape = NA) + # Don't plot outliers since I'm overlaying points
  geom_jitter(aes(color = season),
              alpha = 0.25, size = 0.7, width = 0.7) +
  facet_wrap(~ workingday, ncol = 1) +
  theme_minimal()
```


Holiday / weekend
Working day
count
1000
750
500
250
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
factor(hour)
season
Spring
Summer
Fall
Winter