1. Just how thirsty *is* Thursday?

We have previously explored some breathalyzer test data from Ames, Iowa, USA, here. Now, we focus on working with the date and time information in the data. As a college town (Go Cyclones!), Ames has had its fair share of alcohol-related incidents. (For example, Google "VEISHEA riots 2014".) In this project, we examine breath alcohol test data in Ames from January 2013 to December 2017 that is published by the State of Iowa, specifically focusing on *when* these tests are administered.

The data file "breath*alcohol*datetimes.csv" contains 1,556 observations from breath alcohol tests collected by the Ames and Iowa State University Police Departments. The columns in this dataset are:

1. **DateTime** - date & time of test (datetime, "America/Chicago")
2. **Location** - who administered the test, Ames PD or ISU PD? (char.)
3. **Gender** - gender (M,F) of person being tested (char.)
4. **Res1** - first breath alcohol reading (num.)
5. **Res2** - second breath alcohol reading (num.)

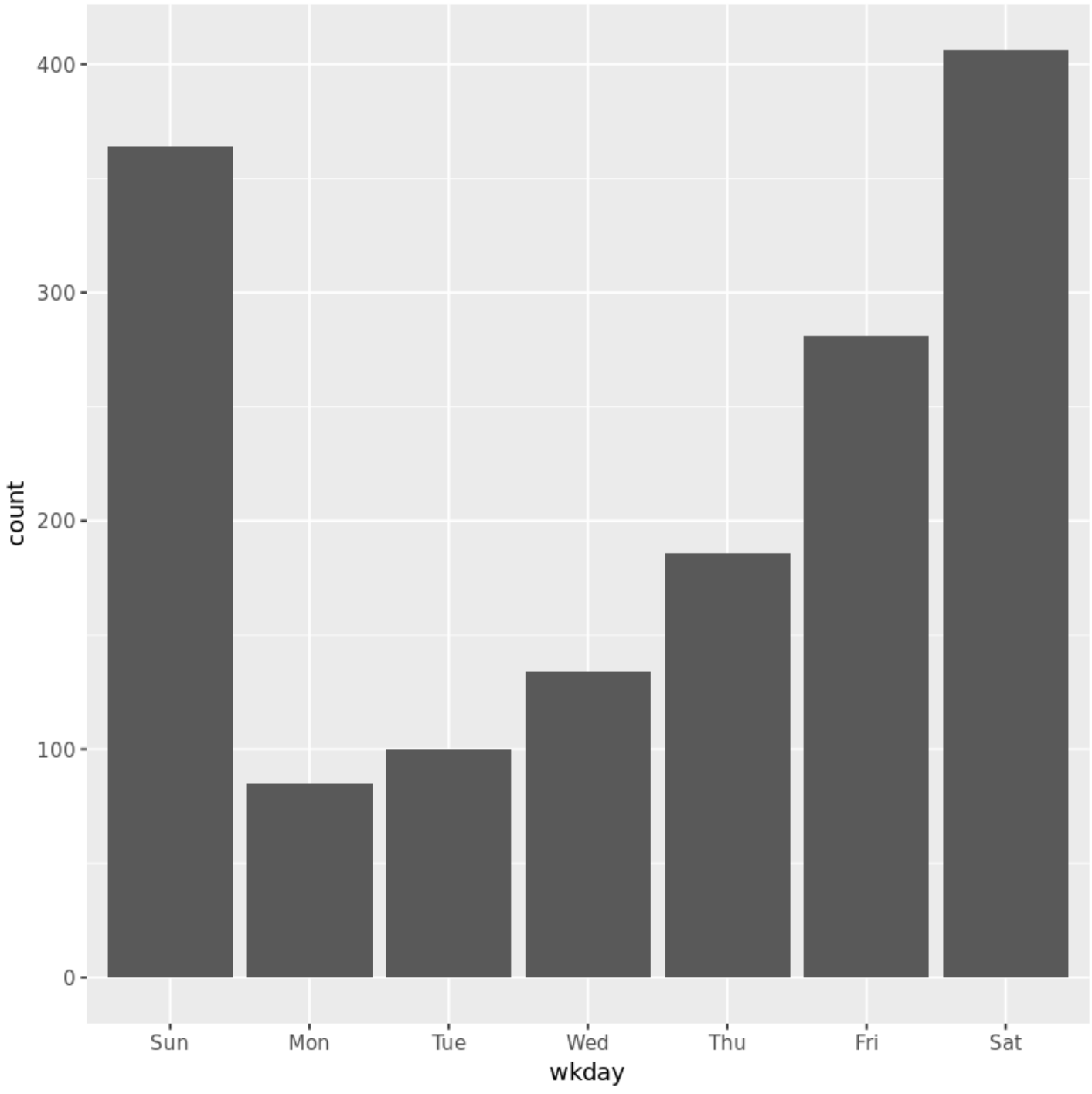
First, we create a bar chart showing number of tests by day of the week to see when the most tests were done.

In [2]:

```
# load necessary packages
library(tidyverse)
library(lubridate)

# read in the data from breath_alcohol_datetimes.csv
ba_dates <- read_csv('datasets/breath_alcohol_datetimes.csv')

# change DateTime column to America/Chicago with force_tz
ba_dates <- ba_dates %>% mutate(DateTime = force_tz(DateTime,
                                                    tzone =
                                                    'America/Chicago'))

# create a wkday column in the ba_dates
ba_dates <- ba_dates %>% mutate(wkday = wday(DateTime, label = TRUE,
                                             abbr = TRUE))

# create a bar chart of # tests by day of week
ggplot(data = ba_dates, aes(x = wkday)) +
  geom_bar()
```

```
-- Attaching packages --------------------------------------- tidyverse 1.2.1 --
<U+221A> ggplot2 3.0.0     <U+221A> purrr   0.2.5
<U+221A> tibble  1.4.2     <U+221A> dplyr   0.7.6
<U+221A> tidyr   0.8.1     <U+221A> stringr 1.3.1
<U+221A> readr   1.1.1     <U+221A> forcats 0.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

Attaching package: 'lubridate'

The following object is masked from 'package:base':

    date
```

```
Parsed with column specification:
cols(
  DateTime = col_datetime(format = ""),
  Location = col_character(),
  Gender = col_character(),
  Res1 = col_double(),
  Res2 = col_double()
)
```



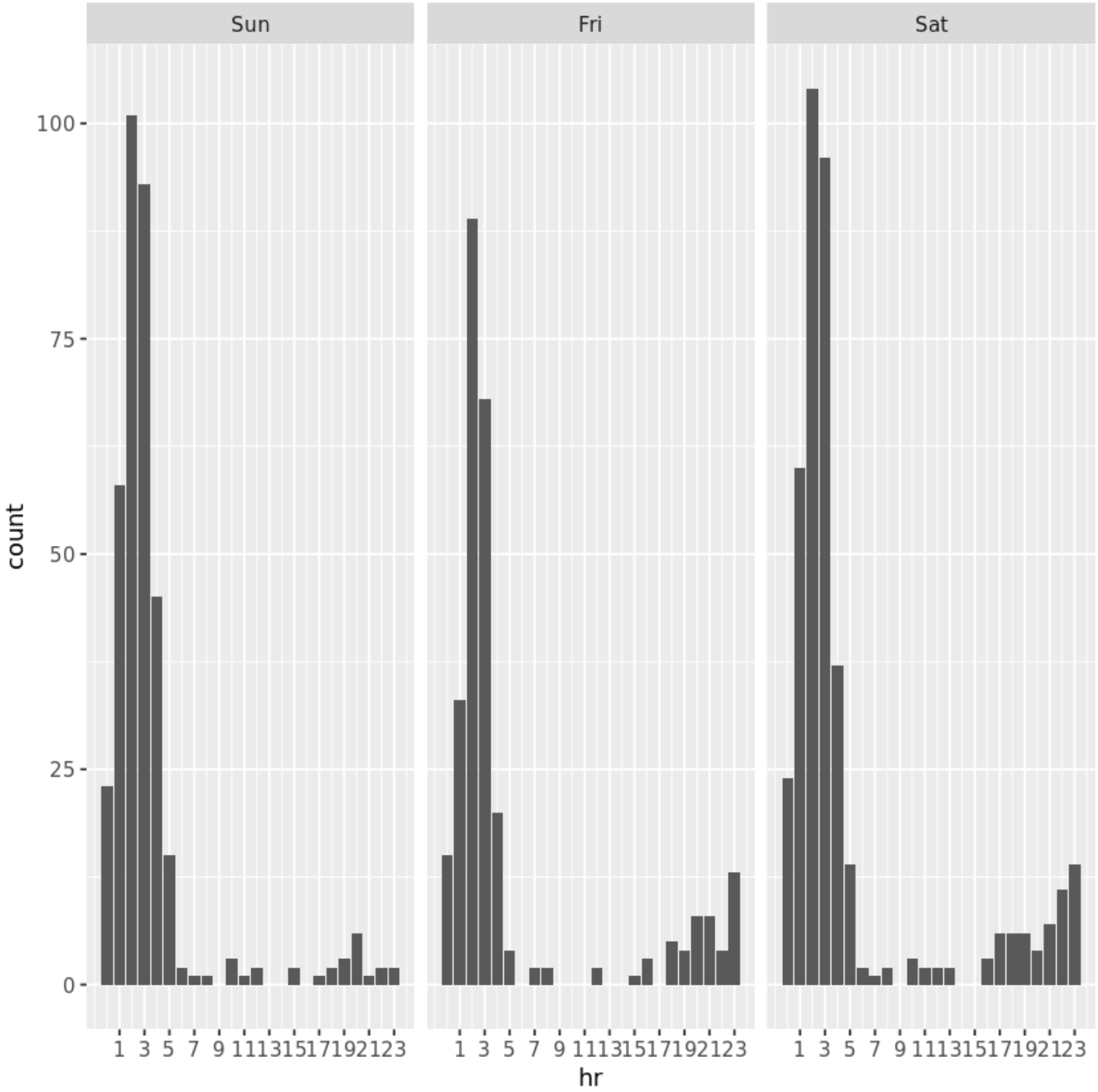
2. What makes Sunday so fun-day?

Well, that wasn't terribly surprising: Friday and Saturday are two of the most common days of the week for breathalyzer tests in a college town. But what might be *somewhat* surprising is that more tests occur on Sunday than on Friday. But *when* on Sunday are these tests being administered? To investigate, we look at the hour of the test and compare this data for Friday, Saturday, and Sunday.

In [4]:

```
# create hour variable
ba_dates <- ba_dates %>% mutate(hr = hour(DateTime))
```

```
# create weekend data frame
weekend <- ba_dates %>% filter(wkday %in% c('Fri', 'Sat', 'Sun'))

# plot side-by side bar charts counting hour of the day of tests for each weekend day
ggplot(data = weekend) +
  geom_bar(aes(x = hr)) +
  facet_grid(.~wkday) +
  scale_x_continuous(breaks = 1:12*2-1) # for ease of readability
```



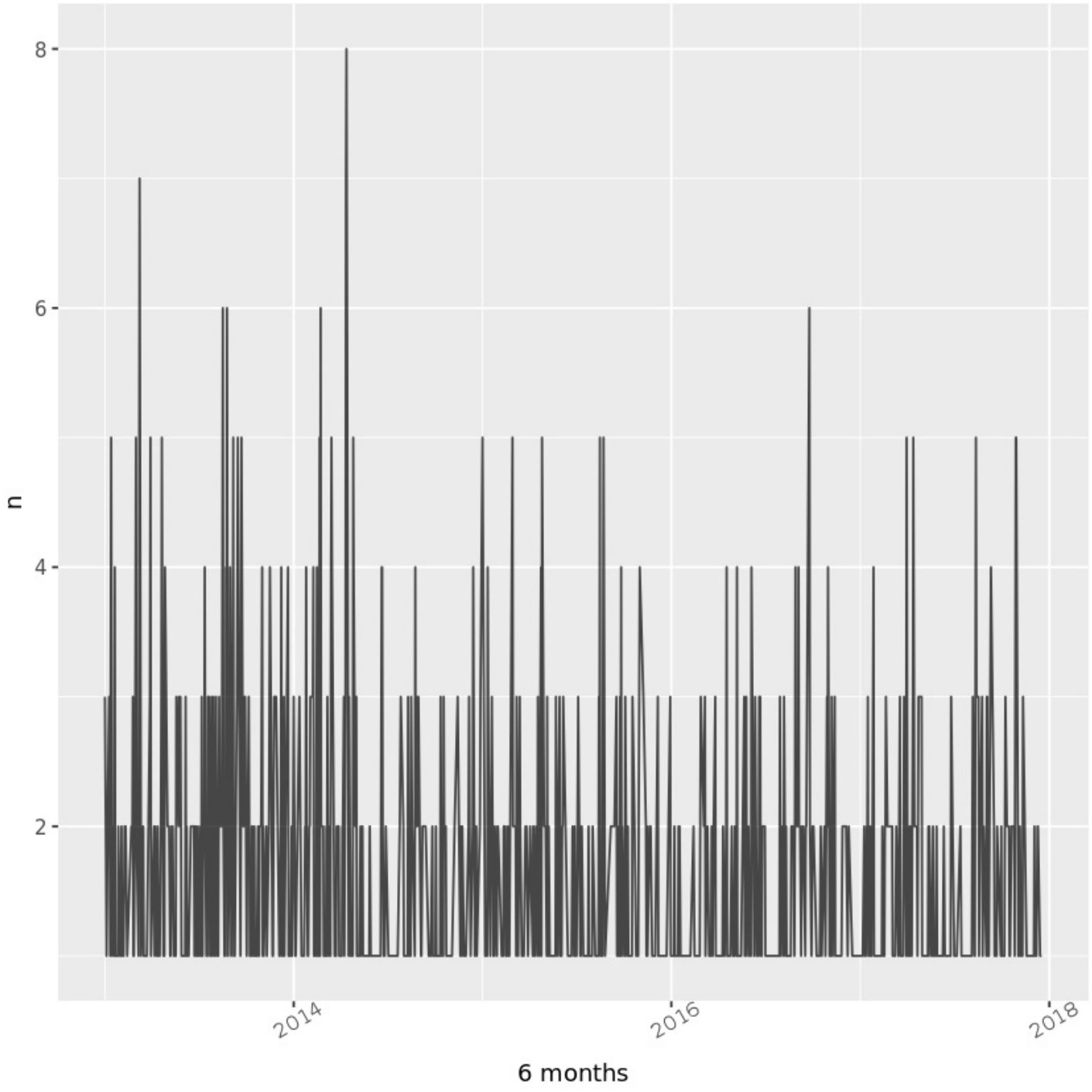
3. Trends in testing over time

We learned that most of the tests administered on Sundays are during early morning, from midnight to 5am. Strangely, the same pattern also exists on Friday mornings. (This is likely because Thursdays are "Mug Nights" in Ames, where you can get discounted drinks if you bring in the designated reusable mug.) Returning to the full dataset, we now explore the pattern of alcohol tests over the years. To look at the "bigger picture," let's count up the number of tests per day, and visualize the resulting time series using a line plot.

In [6]:

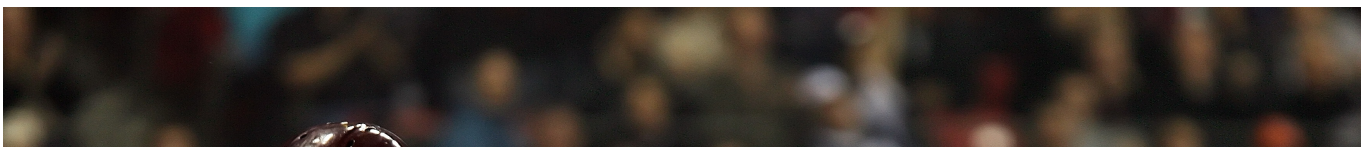
```
# create a date column rounded to the nearest day. as.Date() is for the plot later
ba_dates <- ba_dates %>% mutate(date = as.Date(round_date(DateTime,
                                                          unit = 'day')))
```

```
# count number of tests per date
ba_summary <- ba_dates %>% count(date)

# pipe the result from above into ggplot() using geom_line to create a time series plot.
ba_summary %>%
  ggplot() +
  geom_line(aes(x = date, y = n), alpha = .7) + # change alpha for readability
  scale_x_date('6 months') +
  theme(axis.text.x = element_text(angle = 30)) # make x-axis more readable
```



4. College football

In the time series, we see many days that have zero breathalyzer tests administered. In the entire five year period, there were at most eight tests in a day. There are many days with three or more tests in a day, and we wonder if the Iowa State football schedule may match up with some of those high test days. We next explore the Iowa State football schedule for 2013-2017. Data were downloaded from sports-reference.com.



In [8]:

```
# read in the football data
isu_fb <- read_csv('datasets/isu_football.csv')

# make Date a date variable
isu_fb <- isu_fb %>% mutate(Date = parse_date(Date, format = "%b %d, %Y"))

# filter ba_summary
ba_fb  <- ba_summary %>% filter(date %in% isu_fb$Date)

# arrange ba_fb by number of tests from high to low and print first six rows
ba_fb %>% arrange(desc(n)) %>% head()
```

```
Parsed with column specification:
cols(
  G = col_integer(),
  Date = col_character(),
  Time = col_time(format = ""),
  Day = col_character(),
  School = col_character(),
  Home = col_character(),
  Opponent = col_character(),
  Conf = col_character(),
  Res = col_character(),
  Pts = col_integer(),
  Opp = col_integer(),
  W = col_integer(),
  L = col_integer(),
  Streak = col_character(),
  Notes = col_character(),
  TV = col_character()
)
```

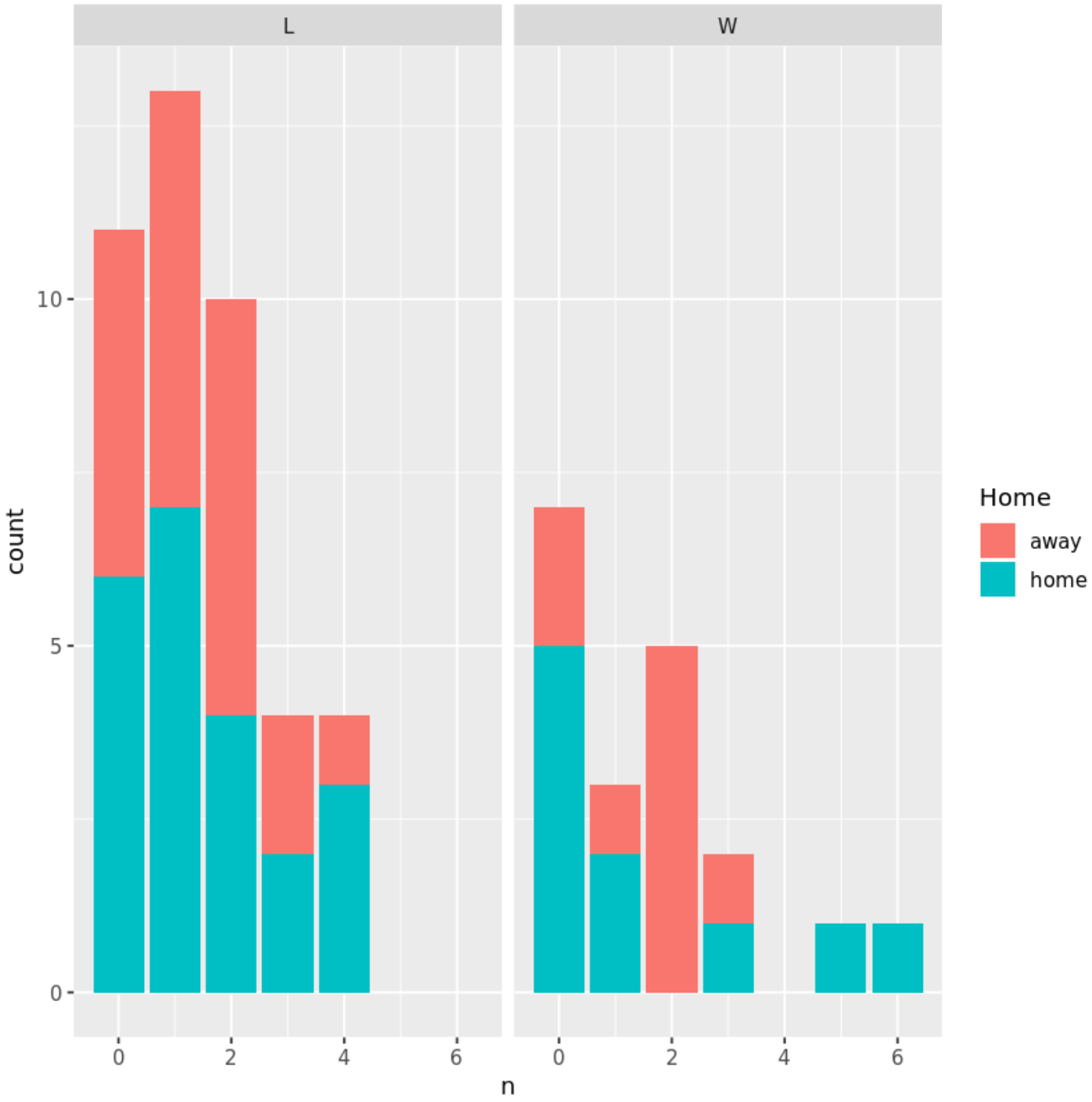
date	n
2016-09-24	6
2017-10-28	5
2013-08-31	4
2013-09-14	4
2013-11-16	4
2016-09-03	4

5. Home vs. away? Win vs. lose?

The most breathalyzer tests given on a football game day was on Sept. 24, 2016. This was a home game against San Jose State that Iowa State won 44-10. The win/loss information is in the `Res` column in `isu_fb`. Could the home game win have led to some excessive celebrations that resulted in more breathalyzer tests than an away win or a home loss?

In [10]:

```
# join ba_summary to isu_fb
isu_fb2 <- isu_fb %>% left_join(ba_summary, by = c("Date" = "date"))

# change nas to 0s
isu_fb2 <- isu_fb2 %>% mutate(n = ifelse(is.na(n), 0, n))

# plot
isu_fb2 %>%
  ggplot() +
  geom_bar(aes(x = n, fill = Home)) +
  facet_grid(.~Res)
```

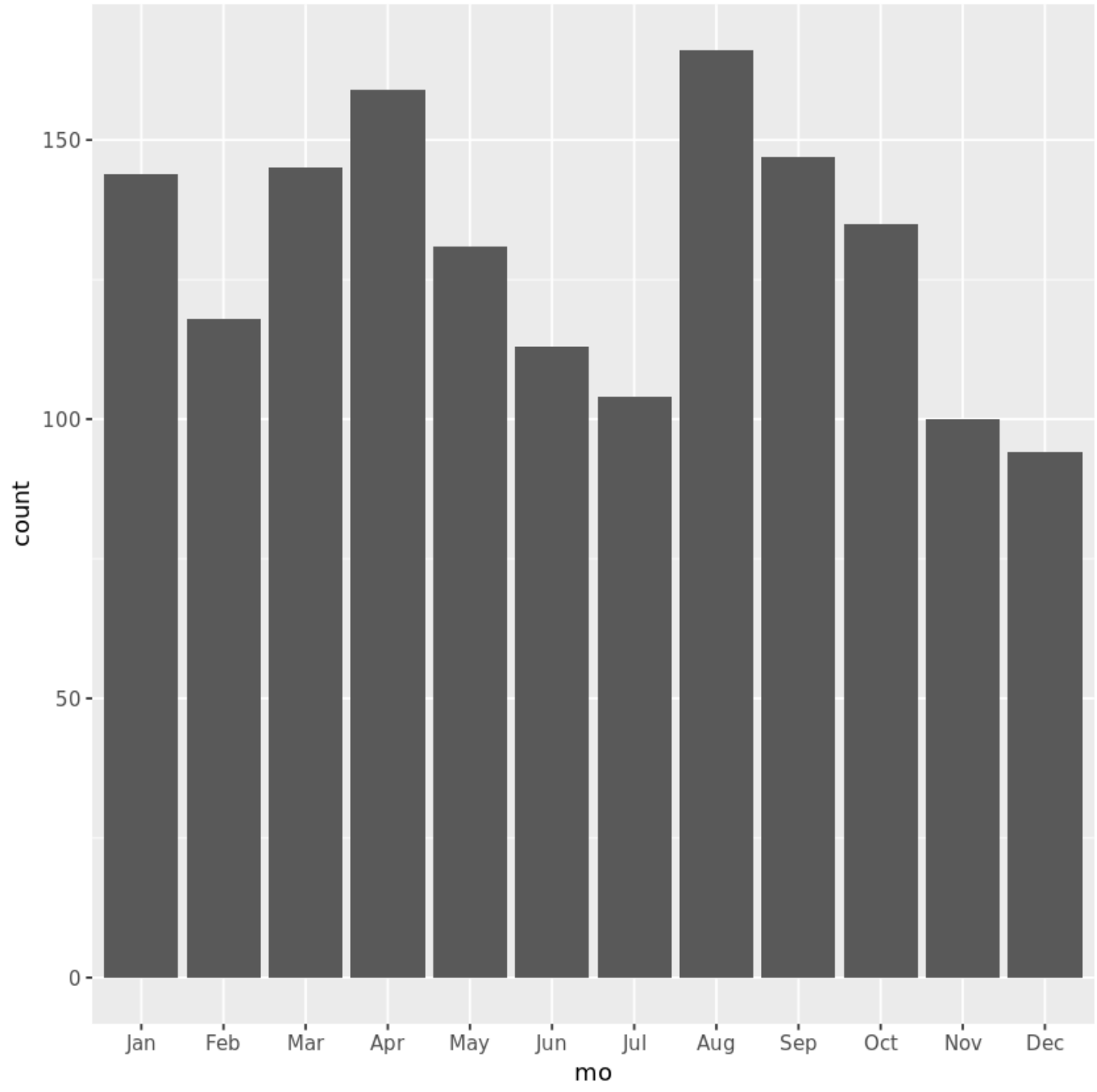


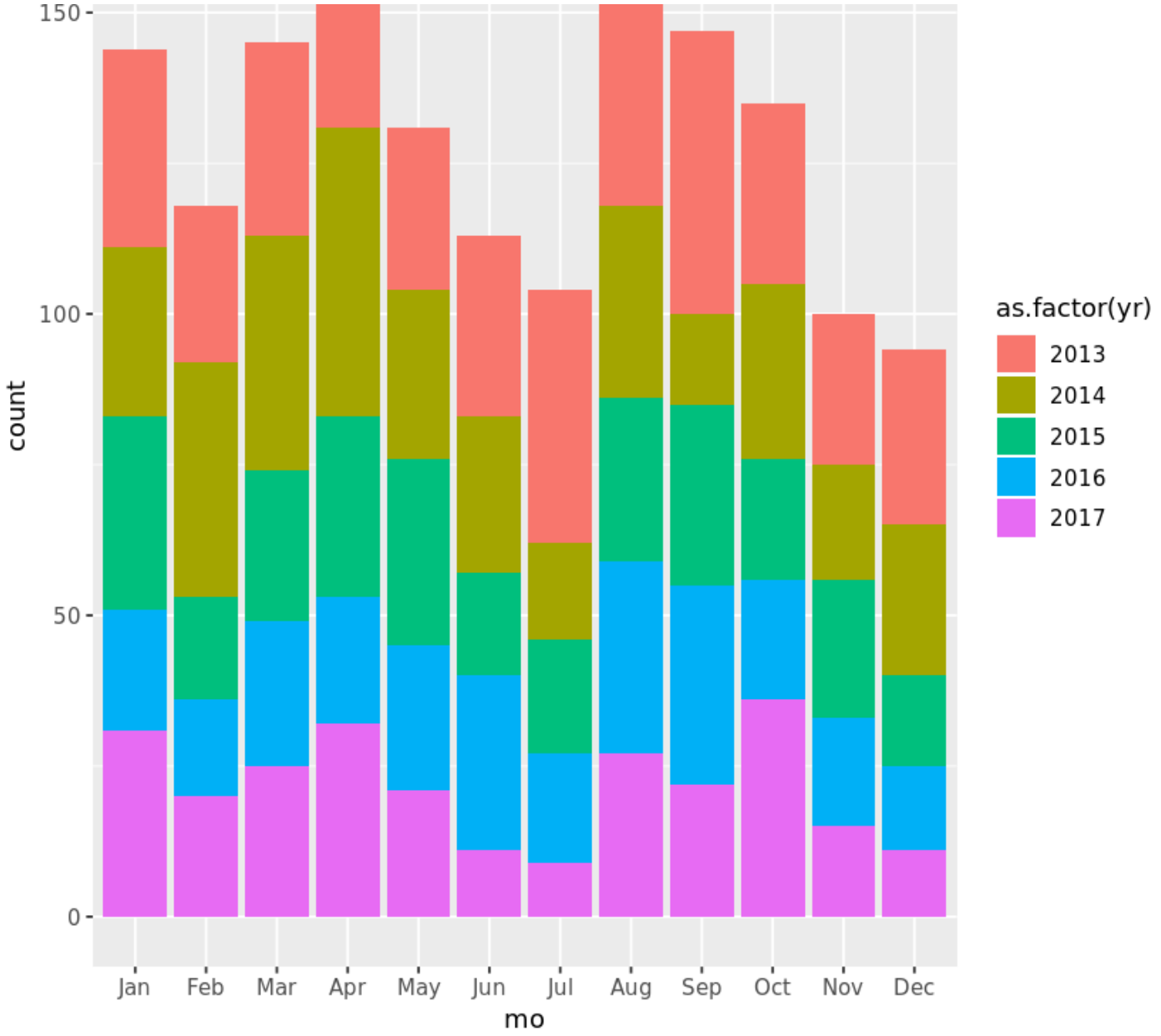
6. Monthly counts

The football season typically lasts from September through November. As we just saw, Iowa State football has more losses than wins in the last few years. The men's basketball team, however, has traditionally been very successful. The basketball season usually lasts from November through March. We now investigate the number of breathalyzer tests by month to see if the basketball months have more tests than the football months.

In [12]:

```
# create a mo and a yr column in ba_dates
ba_dates <- ba_dates %>% mutate(mo = month(date, label = T), yr = year(date))

# make bar chart by mo.
ba_dates %>% ggplot() + geom_bar(aes(x=mo))

# color by year
ba_dates %>% ggplot() + geom_bar(aes(x=mo, fill = as.factor(yr)))
```





7. VEISHEA: an old tradition

The monthly bar charts show that the months with the most test per day are August and April. April is a surprise because there are no major college sports in April, and students are busy studying for finals and finishing semester projects. Well, at Iowa State, there was a historical weeklong festival known as VEISHEA held in April every year. It was cancelled in 2014 due to the many drinking-related arrests, violence, and vandalism that occurred yearly. Looking at the VEISHEA weeks and subsequent non-VEISHEA weeks, can we see the effect of the cancellation in the breathalyzer data?

In [14]:

```
# In 2013, VEISHEA was held from April 15-21. In 2014, it was held from April 7-13.
v13 <- interval(make_date(2013, 4, 15) , make_date(2013, 4, 21),
                tzone = 'America/Chicago')
v14 <- interval(make_date(2014, 4, 7) , make_date(2014, 4, 13),
                tzone = 'America/Chicago')
# Other comparable VEISHEA weeks in 2015-2017
v15 <- interval(make_date(2015, 4, 13) , make_date(2015, 4, 19), tzone = "America/Chicag
o")
v16 <- interval(make_date(2016, 4, 11) , make_date(2016, 4, 17), tzone = "America/Chicag
o")
v17 <- interval(make_date(2017, 4, 10) , make_date(2017, 4, 16), tzone = "America/Chicag
o")

# filter ba_dates for only the 5 veishea intervals
veishea <- ba_dates %>% filter(date %within% v13 |
                               date %within% v14 |
```

```
                                date %within% v15 |
                                date %within% v16 |
                                date %within% v17)

# count up years
veishea %>% count(date)
```

date	n
2013-04-16	1
2013-04-17	1
2013-04-19	2
2013-04-20	3
2013-04-21	5
2014-04-08	3
2014-04-09	1
2014-04-10	2
2014-04-11	3
2014-04-12	7
2014-04-13	8
2015-04-16	2
2015-04-17	1
2015-04-18	3
2015-04-19	2
2016-04-12	1
2016-04-13	1
2016-04-16	2
2016-04-17	4
2017-04-10	1
2017-04-12	3
2017-04-13	5
2017-04-15	2

8. Looking at BAC

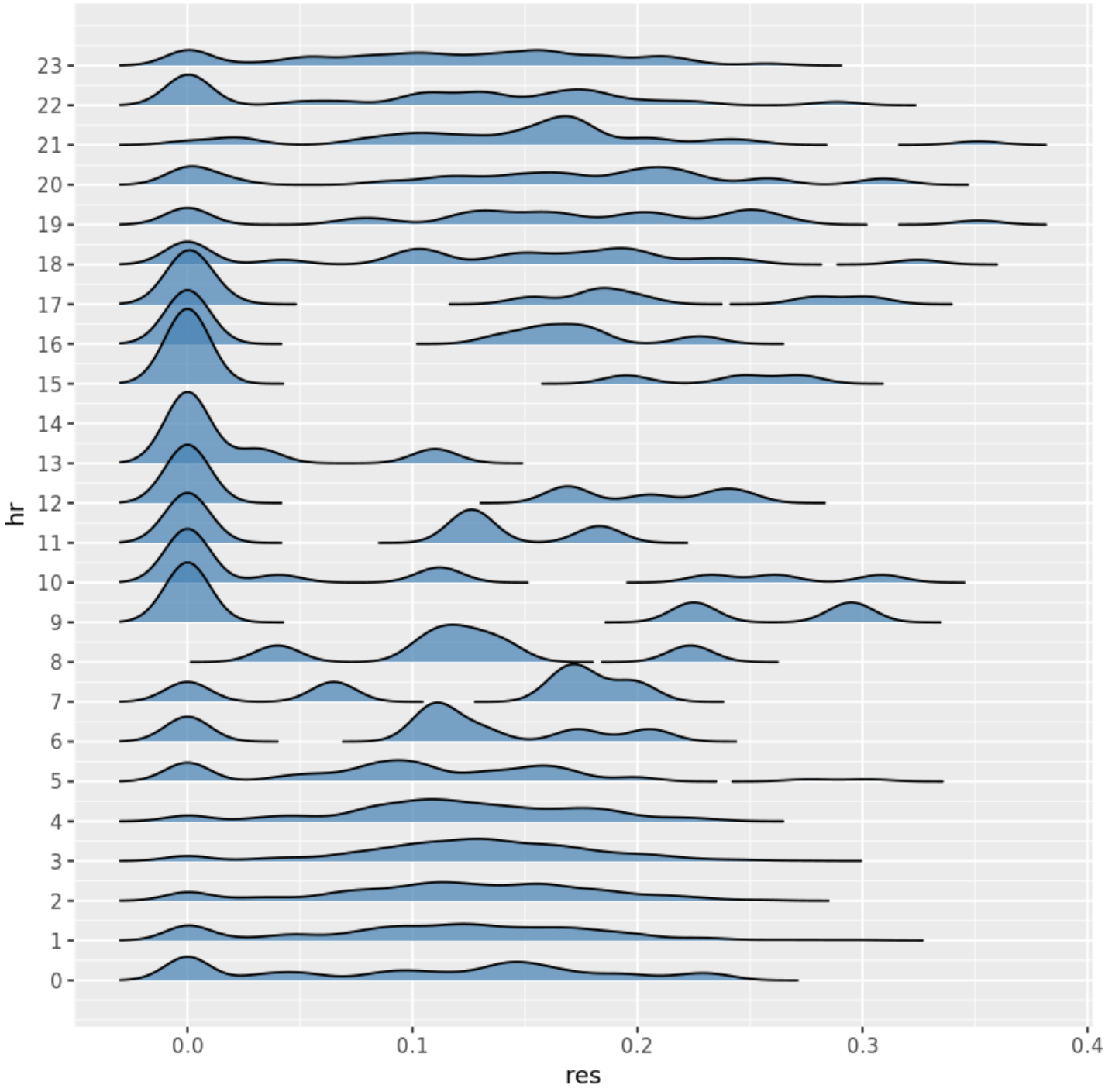
Finally, let's look at the actual results of the breathalyzer tests. Based on our knowledge from Section 2, we suspect that the highest BAC results occur late and night and in the early morning, since those times are most common for tests on the weekends.

In [16]:

```
# take a mean of res1, res2
ba_dates   <- ba_dates %>% mutate(res = (Res1 + Res2) / 2)

# library the ggridges package
library(ggridges)

# make ridgeline plot
ggplot(data = ba_dates, aes(x = res, y = hr, group = hr)) +
  # some style choices made already
  geom_density_ridges(alpha = 0.7, fill = "steelblue", bandwidth = .01, rel_min_height =
0.0001) +
  scale_y_continuous(breaks = 0:23)
```

```
Attaching package: 'ggridges'

The following object is masked from 'package:ggplot2':
```

```
scale_discrete_manual
```



9. A more honest plot

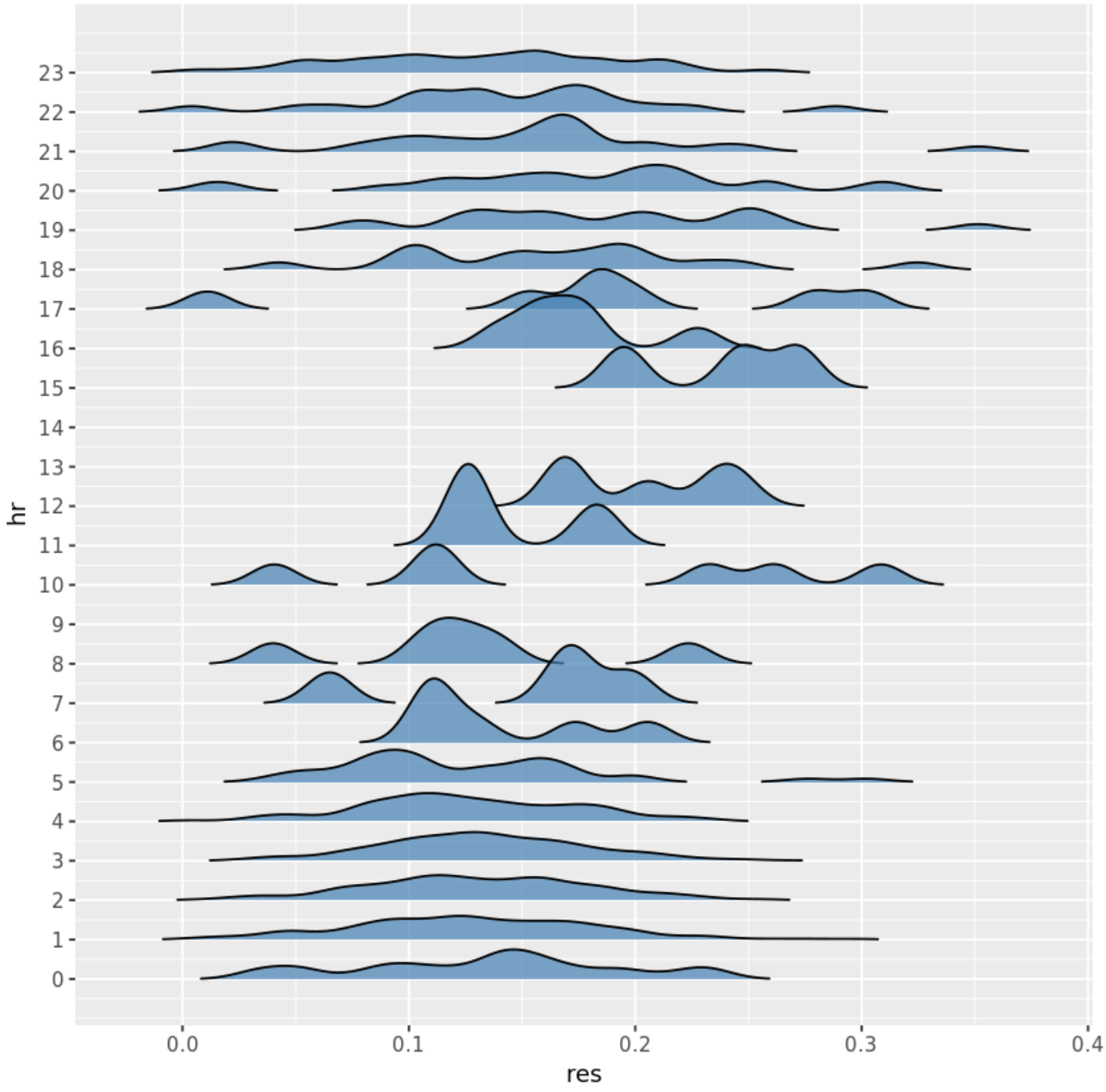
In the previous ridgeline plot, there are values below zero. This is impossible given the context: you cannot have negative alcohol concentration in your blood. We examine the zeroes below and make a more honest ridgeline plot.

In [18]:

```
# create a zero indicator variable
ba_dates <- ba_dates %>% mutate(zero = res == 0)

# tabulate the data by the zero column
ba_dates %>% count(zero)

# redo ridge with no 0s
ba_dates %>% filter(res != 0) %>%
ggplot(aes(x = res, y = hr, group = hr)) +
  geom_density_ridges(alpha = 0.7, fill = "steelblue", bandwidth = .01, rel_min_height =
```

```
0.005) +
  scale_y_continuous(breaks = 0:23)
```

zero	n
FALSE	1353
TRUE	203



10. The dangers of binge drinking

At a breath alcohol level of 0.16-0.30, a person will experience significant "speech, memory, coordination, attention, reaction time, [and] balance" impairment. Someone's "driving-related skills" and "judgement and decision making" are dangerously impaired, and they may experience "blackouts, vomiting [...] and loss of consciousness". BAC of 0.31 or above is life-threatening with "significant risk of death" (source). We conclude by looking at the time of day during which the most dangerous levels of alcohol consumption appear in the Ames data. Do the dates, times, and days of week match what one would expect?

While this report has taken a fun and playful tone, it is important to be aware of the seriousness of this issue. According to research, nearly 2,000 college students die each year from alcohol-related injuries. If you or a loved one are struggling with alcohol abuse, please seek help.

In [20]:

```
# filter the ba_dates data to contain only those with the most dangerous result
danger <- ba_dates %>% filter(res >= 0.31)

# print danger
print(danger)
```

```
# A tibble: 4 x 12
  DateTime            Location Gender  Res1  Res2 wkday    hr date       mo
  <dttm>              <chr>    <chr>  <dbl> <dbl> <ord> <int> <date>     <ord>
1 2015-01-05 20:49:00 Ames PD  M      0.313 0.312 Mon      20 2015-01-06 Jan
2 2014-11-13 19:23:00 Ames PD  F      0.352 0.351 Thu      19 2014-11-14 Nov
3 2014-10-21 18:03:00 Ames PD  F      0.325 0.324 Tue      18 2014-10-22 Oct
4 2013-05-28 21:01:00 ISU PD   F      0.352 0.351 Tue      21 2013-05-29 May
# ... with 3 more variables: yr <dbl>, res <dbl>, zero <lgl>
```