AE 02: Bike rentals in Washington, DC

Exploring and modeling relationships

Dav King

Sep 05, 2022

```r
library(tidyverse)
library(tidymodels)
library(ggridges)
library(viridis)
```

Data

Our dataset contains daily rentals from the Capital Bikeshare in Washington, DC in 2011 and 2012. It was obtained from the `dcbikeshare` data set in the dsbox R package.

We will focus on the following variables in the analysis:

- `count`: total bike rentals
- `temp_orig`: Temperature in degrees Celsius
- `season`: 1 - winter, 2 - spring, 3 - summer, 4 - fall

Click here for the full list of variables and definitions.

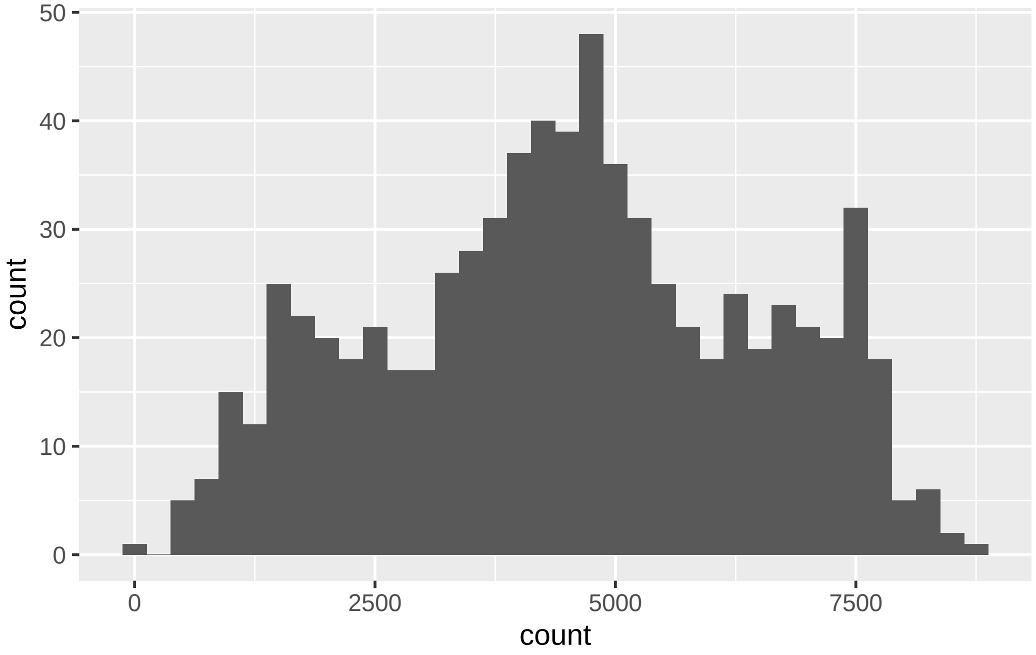
```r
bikeshare <- read_csv("data/dcbikeshare.csv")
```

Daily counts and temperature

Exercise 1

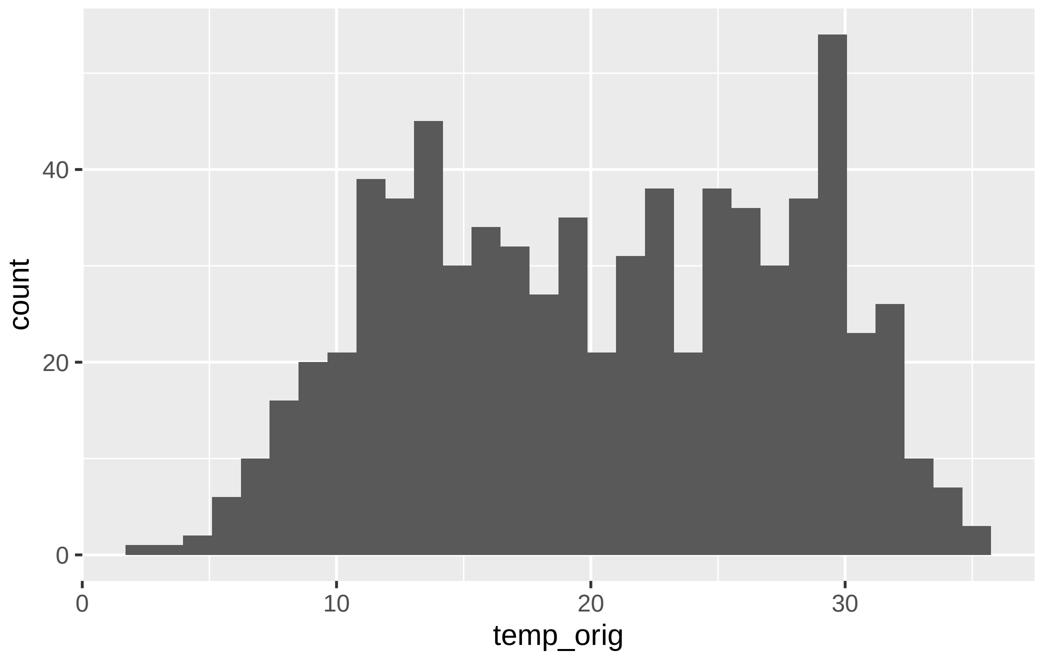
Visualize the distribution of daily bike rentals and temperature as well as the relationship between these two variables.

```r
ggplot(bikeshare, aes(x = count)) +
  geom_histogram(binwidth = 250)
```



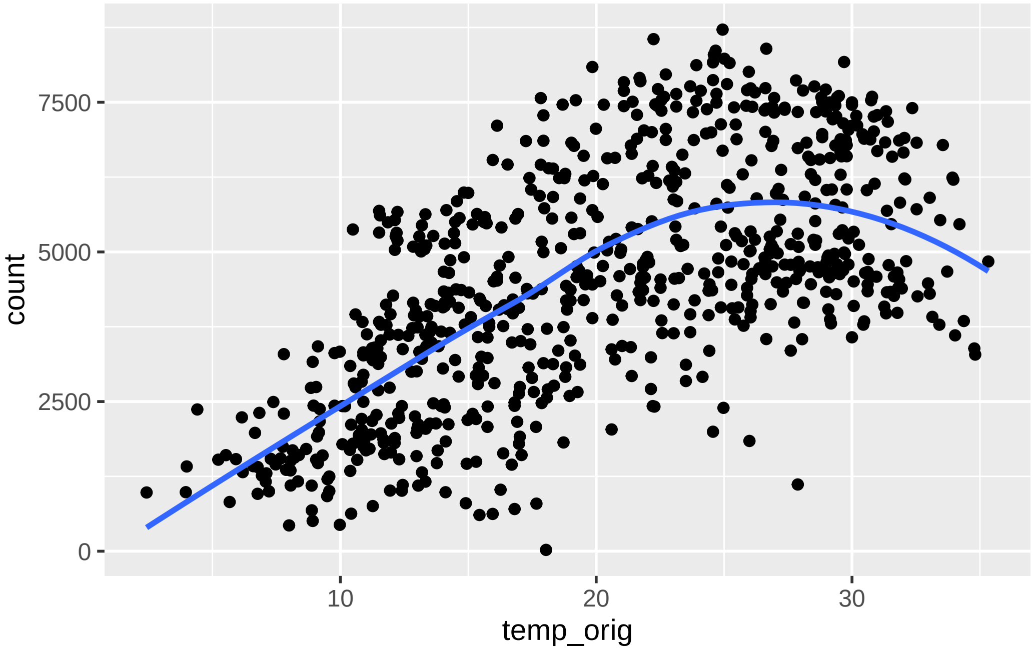
```r
ggplot(bikeshare, aes(x = temp_orig)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(bikeshare, aes(y = count, x = temp_orig)) +
  geom_point() +
  geom_smooth(se = F)
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Exercise 2

Describe the distribution of daily bike rentals and the distribution of temperature based on the visualizations created in Exercise 1. Include the shape, center, spread, and presence of any potential outliers.

The distribution of daily bike rentals seems to be roughly trimodal, with a peak at around 2000, a much larger peak around 4500, and then a peak around 7000 of roughly comparable size to the first one. The distribution of daily bike rentals is roughly centered at 4500, and ranges from 0 to ~9000. The only notable outlier is days with almost 0 rentals, but given a standard definition of outliers no days would fall into that category.

The distribution of temperatures is roughly uniform between 10 and 30, with much less frequent temperatures outside of those extremes. The data are roughly centered around 20 degrees celsius, with a range from 0 to around 35 degrees. Nothing would likely be statistically identifiable as an outlier.

Exercise 3

There appears to be one day with a very small number of bike rentals. What was the day? Why were the number of bike rentals so low on that day? *Hint: You can Google the date to figure out what was going on that day.*

```
bikeshare %>%
  arrange(count) %>%
  slice(1)
```

```
# A tibble: 1 x 17
  instant dteday     season    yr  mnth holiday weekday workingday weathersit
    <dbl> <date>      <dbl> <dbl> <dbl>   <dbl>   <dbl>      <dbl>      <dbl>
1     668 2012-10-29      4     1    10       0       1          1          3
# ... with 8 more variables: temp <dbl>, atemp <dbl>, hum <dbl>,
#   windspeed <dbl>, casual <dbl>, registered <dbl>, count <dbl>,
#   temp_orig <dbl>
```

The day was 15642, when Hurricane Sandy was hitting DC (and about half of the rest of the country).

Exercise 4

Describe the relationship between daily bike rentals and temperature based on the visualization created in Exercise 1. Comment on how we expect the number of bike rentals to change as the temperature increases.

Generally, the number of daily bike rentals increases with increasing temperatures, so we would expect higher temperatures to correlate with higher numbers of bike rentals. However, when we fit a loess (moving) regression line to the data, we can see that this relationship only holds true until a certain point around 25 degrees celsius - after that, bike rentals sharply drop off with increasing temperatures, and we would expect hotter temperatures to see fewer of these rentals. This makes sense - as the weather gets warmer, more and more people will go ride bikes, but after a certain temperature it gets too hot and this begins to drop off again.

Exercise 5

Suppose you want to fit a model so you can use the temperature to predict the number of bike rentals. Would a model of the form

$$\text{count} = \beta_0 + \beta_1 \ \text{temp\_orig} + \epsilon$$

be the best fit for the data? Why or why not?

If we wanted to run a linear regression, this would be the most fitting model for doing so. However, these data are not best modeled in a linear fashion. Because the number of bike rentals tends to increase until a certain temperature and then fall back off somewhat, we would want to construct a non-linear model of regression to understand our data.

```
allData <- linear_reg() %>%
  set_engine("lm") %>%
  fit(count ~ temp_orig, bikeshare)
tidy(allData)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    1215.     161.       7.54 1.43e-13
2 temp_orig       162.       7.44     21.8  2.81e-81
```

```
newTemp <- tibble(temp_orig = 25)

predict(allData, newTemp)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1 5264.
```

$$\hat{count} = 1215 + 161 \times temp\_orig$$

Daily counts, temperature, and season

Exercise 6

In the raw data, seasons are coded as 1, 2, 3, 4 as numerical values, corresponding to winter, spring, summer, and fall respectively. Recode the `season` variable to make it a categorical variable (a factor) with levels corresponding to season names, making sure that the levels appear in a reasonable order in the variable (i.e., not alphabetical).

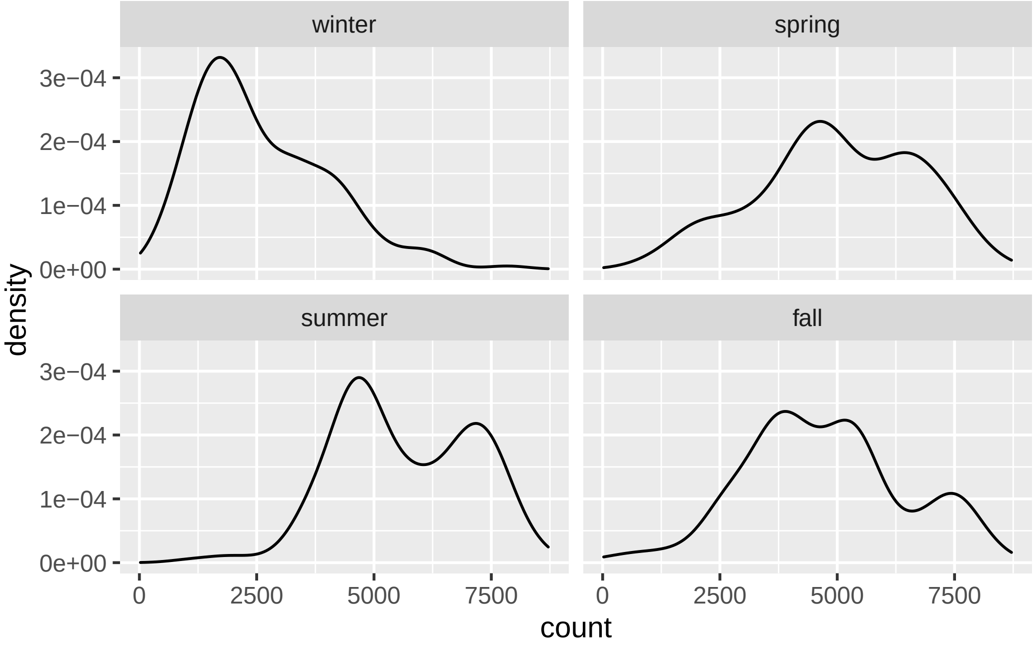
```
bikeshare <- bikeshare %>%
  mutate(season = factor(season, labels = c("winter", "spring", "summer", "fall")))

#bikeshare <- bikeshare %>%
  #mutate(season = case_when(
    #season == 1 ~ "winter",
    #season == 2 ~ "spring",
    #season == 3 ~ "summer",
    #season == 4 ~ "fall")) %>%
  #mutate(season = factor(season, levels =
                            #c("winter", "spring", "summer", "fall")))
```

Exercise 7

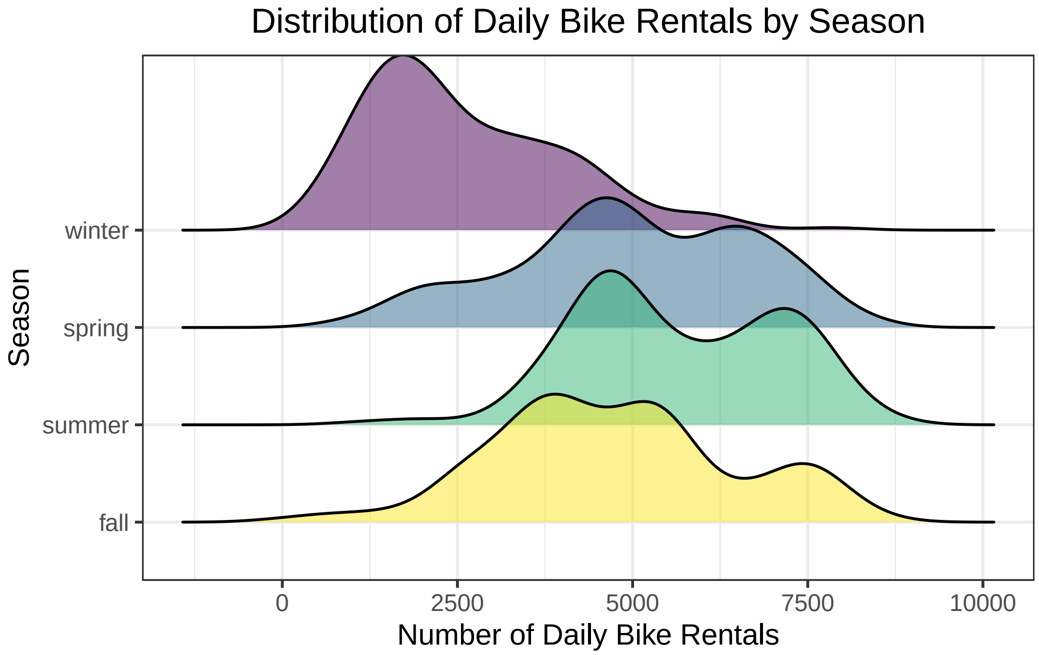
Next, let's look at how the daily bike rentals differ by season. Let's visualize the distribution of bike rentals by season using density plots. You can think of a density plot as a "smoothed out histogram". Compare and contrast the distributions. Is this what you expected? Why or why not?

```
bikeshare %>%
  ggplot(aes(x = count)) +
  geom_density() +
  facet_wrap(~ season)
```



```r
bikeshare %>%
  ggplot(aes(x = count, y = season, fill = season)) +
  geom_density_ridges(alpha = 0.5) +
  scale_fill_viridis(discrete = T) +
  theme_bw() +
  labs(x = "Number of Daily Bike Rentals", y = "Season",
       title = "Distribution of Daily Bike Rentals by Season") +
  guides(fill = "none") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_discrete(limits = rev)
```

```
Picking joint bandwidth of 481
```



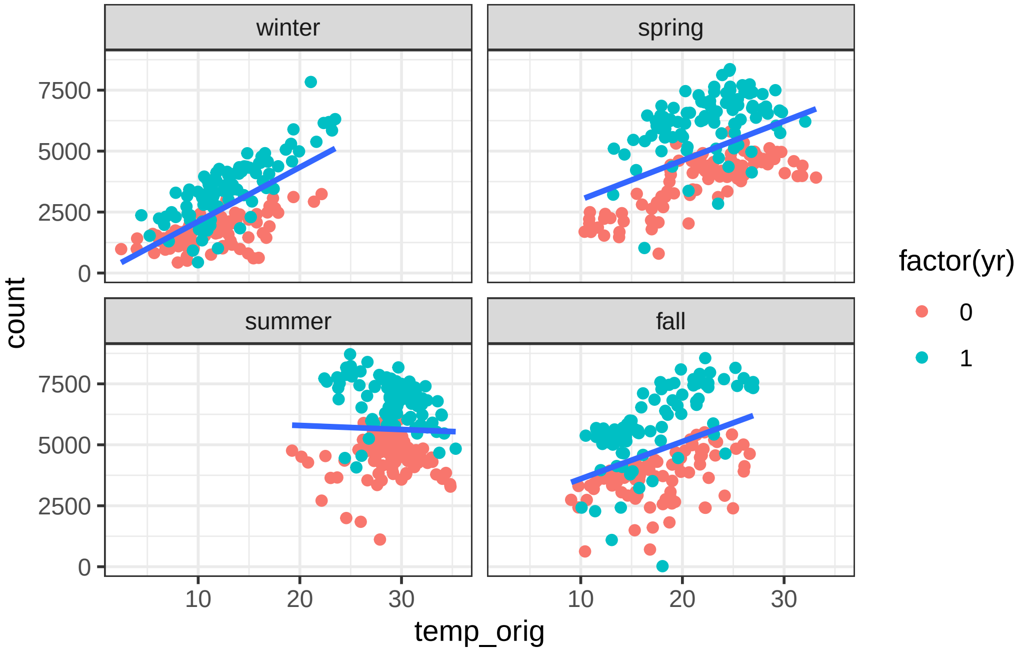
There is a much higher frequency of days with few bike rentals in winter than there are in any other season - which makes sense, because winter is the coldest month and Washingtonians are no fans of the cold. The other seasons are all roughly comparable in distribution, which makes some sense - summer has the most days with really high counts of daily bike rentals, and also the fewest days with very few rentals, but spring and fall have their own highlights as well.

Exercise 8

We want to evaluate whether the relationship between temperature and daily bike rentals is the same for each season. To answer this question, first create a scatter plot of daily bike rentals vs. temperature faceted by season.

```
ggplot(bikeshare, aes(x = temp_orig, y = count)) +
  geom_point(aes(color = factor(yr))) +
  geom_smooth(method = "lm", se = F) +
  facet_wrap(~ season) +
  theme_bw()
```

```
`geom_smooth()` using formula 'y ~ x'
```



Exercise 9

- Which season appears to have the **strongest** relationship between temperature and daily bike rentals? Why do you think the relationship is strongest in this season?

The strongest relationship is definitely winter, which makes a whole world of sense - when people are used to it being cold outside, even a hint of warm weather is enough to get them excited.

- Which season appears to have the **weakest** relationship between temperature and daily bike rentals? Why do you think the relationship is weakest in this season?

The weakest relationship is the summer - when it's already hot outside, there's no reason why even hotter temperatures would make MORE people want to be outside.

Modeling

Exercise 10

Filter your data for the season with the strongest apparent relationship between temperature and daily bike rentals.

```
winter = bikeshare %>%
  filter(season == "winter")
```

Exercise 11

Using the data you filtered in Exercise 10, fit a linear model for predicting daily bike rentals from temperature for this season.

```
winterModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(count ~ temp_orig, data = winter)
tidy(winterModel)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -111.      238.    -0.466 6.42e- 1
2 temp_orig       222.       18.5   12.0   7.28e-25
```

Exercise 12

Use the output to write out the estimated regression equation.

$$\widehat{count} = -111 + 238 \times temp\_orig$$

Exercise 13

Interpret the slope in the context of the data.

For every 1 degree Celsius increase in temperature, we would expect to see an increase of 238 bike rentals on any given day.

Exercise 14

Interpret the intercept in the context of the data.

On a day when the temperature is 0 degrees Celsius, we expect -111 bikes to be rented in DC.

Synthesis

Exercise 15

Suppose you work for a bike share company in Durham, NC, and they want to predict daily bike rentals in 2022. What is one reason you might recommend they use your analysis for this task? What is one reason you would recommend they not use your analysis for this task?

I might recommend they use my analysis because it shows that temperature (a universal) is a strong predictor of bike rentals. However, I very much might not recommend they use my analysis because this is an analysis of Washington, D.C., and it cannot be appropriately extrapolated to Durham.

The following exercises will be completed only if time permits.

Exercise 16

Pick another season. Based on the visualization in Exercise 8, would you expect the slope of the relationship between temperature and daily bike rentals to be smaller or larger than the slope of the model you've been working with so far? Explain your reasoning.

[Add your answer here]

Exercise 17

For this season you picked in Exercise 16, fit a linear model for predicting daily bike rentals from temperature. Note, you will need to filter your data for this season first. Use the output to write out the estimated regression equation and interpret the slope and the intercept of this model.

```
# add your code here
```

[Add your answer here]