# AE 02: Bike rentals in DC

Exploring multivariable relationships

INSERT YOUR NAME HERE

2021-08-24

## Bike rentals in DC

```
library(tidyverse)
```

## 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.

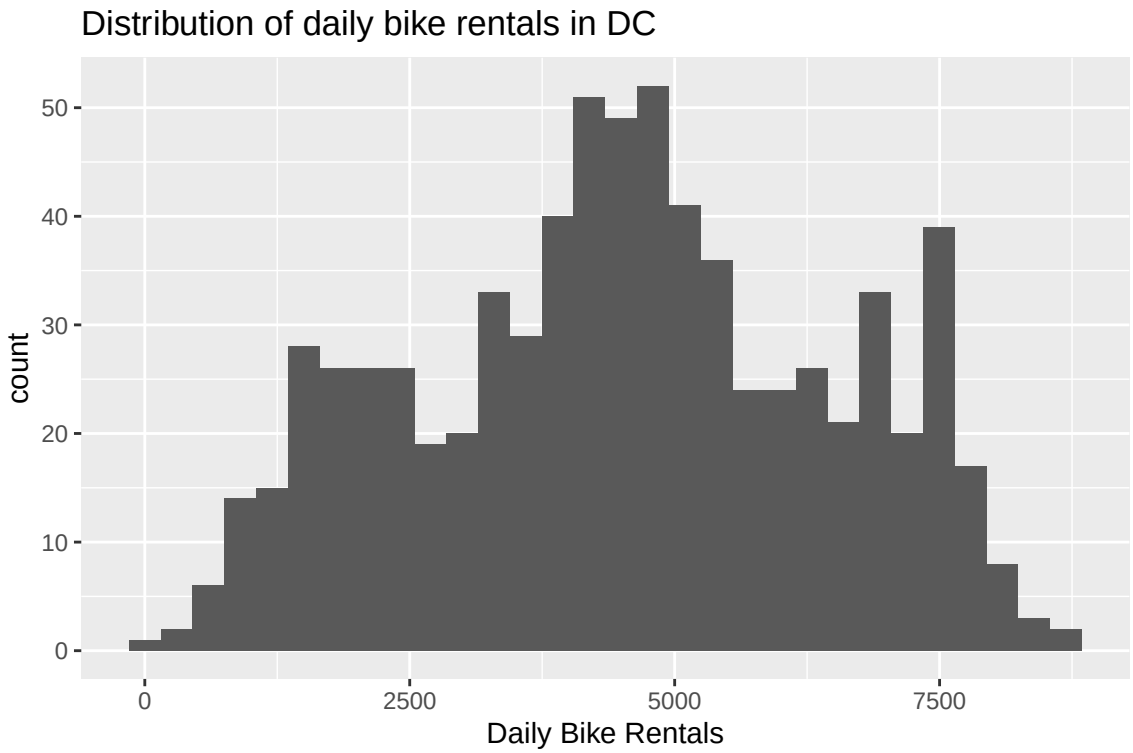
```
bikeshare <- read_csv("data/dcbikeshare.csv") %>%
  mutate(season = factor(season)) # treat season as categorical variable
```

## Exercise 1

Our analysis objective is to understand variability in the daily number of bike rentals. Let's start by look at the distribution of `count`, the total daily bike rentals.

```
ggplot(data = bikeshare, aes(x = count)) +
  geom_histogram() +
  labs(x = "Daily Bike Rentals",
       title = "Distribution of daily bike rentals in DC")
```

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




Describe the distribution of daily bike rentals. Include the shape, center, spread, and presence of any potential outliers.

## Exercise 2

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.*

## Exercise 3

Next, let's look at how the daily bike rentals differ by season. Visualize the distribution of daily bike rentals by season. Compare and contrast the distributions. Is this what you expected? Why or why not?

## Exercise 4

Let's look at the relationship between the temperature in degrees Celsius (`temp_orig`) and the number of bike rentals (`count`). Make a scatterplot to visualize the relationship between these two variables.

Use the scatterplot to describe how we expect the number of bike rentals to change as the temperature increases.

## 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

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

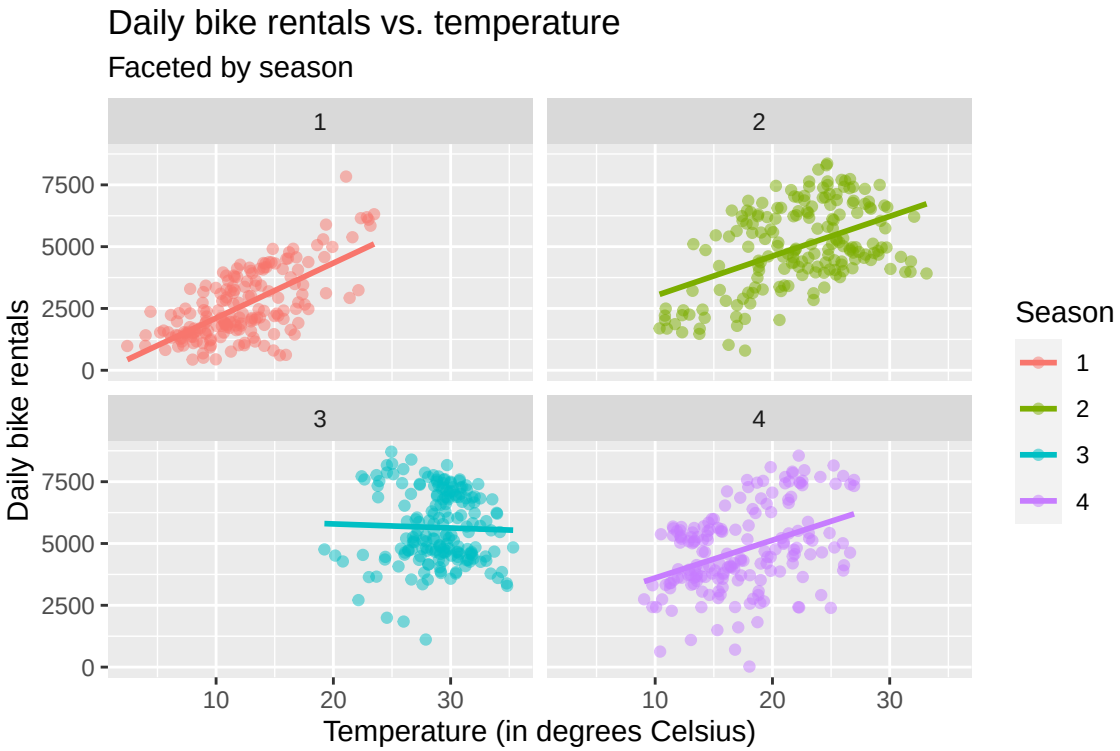
be the best fit for the data? Why or why not?

## Exercise 6

*Is the relationship between temperature and daily bike rentals the same for each season?* To answer this question, we'll examine a plot of daily bike rentals vs. temperature faceted by season.

```
ggplot(data = bikeshare, aes(x = temp_orig, y = count, color = season)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ season) +
  labs(x = "Temperature (in degrees Celsius)",
       y = "Daily bike rentals",
       color = "Season",
       title = "Daily bike rentals vs. temperature",
       subtitle = "Faceted by season")
```

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




- 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?
- 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?

## Exercise 7

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'd recommend they not use your analysis for this task?