Module 3 Assignment 1

Ellen Bledsoe

2023-03-23

Assignment Details

Purpose

The goal of this assignment is to assess your ability to compare means numerically, visually, and statistically

Task

Write R code which produces the correct answers and correctly interpret the results of visualizations and statistical tests.

Criteria for Success

- Code is within the provided code chunks
- Code is commented with brief descriptions of what the code does
- Code chunks run without errors
- Code produces the correct result
  - Code that produces the correct answer will receive full credit
  - Code attempts with logical direction will receive partial credit
- Written answers address the questions in sufficient detail

Due Date

April 4 at midnight MST

Assignment Questions

In this assignment, we're going to explore another data set on wind turbines that generate a significant portion of the energy for us down here in Antarctica.

Set-Up

Let's load the `tidyverse` and read in the data set. Call the data `turbines`.

```r
library(tidyverse)
turbines <- read_csv("../data/wind_turbines.csv")
turbines
```

```
## # A tibble: 67 x 4
##    turbine_id manufacturer wind_speed power_output
##         <dbl> <chr>             <dbl>        <dbl>
##  1          1 Windmill Inc      14.8         65.4
##  2          2 Windmill Inc       9.08         3.84
##  3          3 Windmill Inc       6.51        32.0
##  4          4 Windmill Inc      13.9         37.7
##  5          5 Windmill Inc      17.1         49.9
##  6          6 Windmill Inc      15.0         46.6
##  7          7 Windmill Inc       8.97        48.5
##  8          8 Windmill Inc       7.79         2.97
##  9          9 Windmill Inc       3.49         1.94
## 10         10 Windmill Inc       9.89        65.3
## # ... with 57 more rows
```

1. Explore the data set, either through the environment or through code. Answer the following questions (2 point):

    a. How many turbine makers are there? 2
    b. What does each row of data represent? one turbine

    ```r
    # optional; only if you want space for coding
    ```

Numeric

2. Generate a summary of the data set that calculates the mean wind speed and mean power output for each wind turbine company. (2 point)

```r
turbine_summary <- turbines %>%
  group_by(manufacturer) %>%
  summarise(mean_wind_speed = mean(wind_speed),
            mean_power = mean(power_output))
turbine_summary
```

```
## # A tibble: 2 x 3
##   manufacturer   mean_wind_speed mean_power
##   <chr>                    <dbl>      <dbl>
## 1 Turbo Turbines           10.2        15.8
## 2 Windmill Inc              9.66       37.1
```

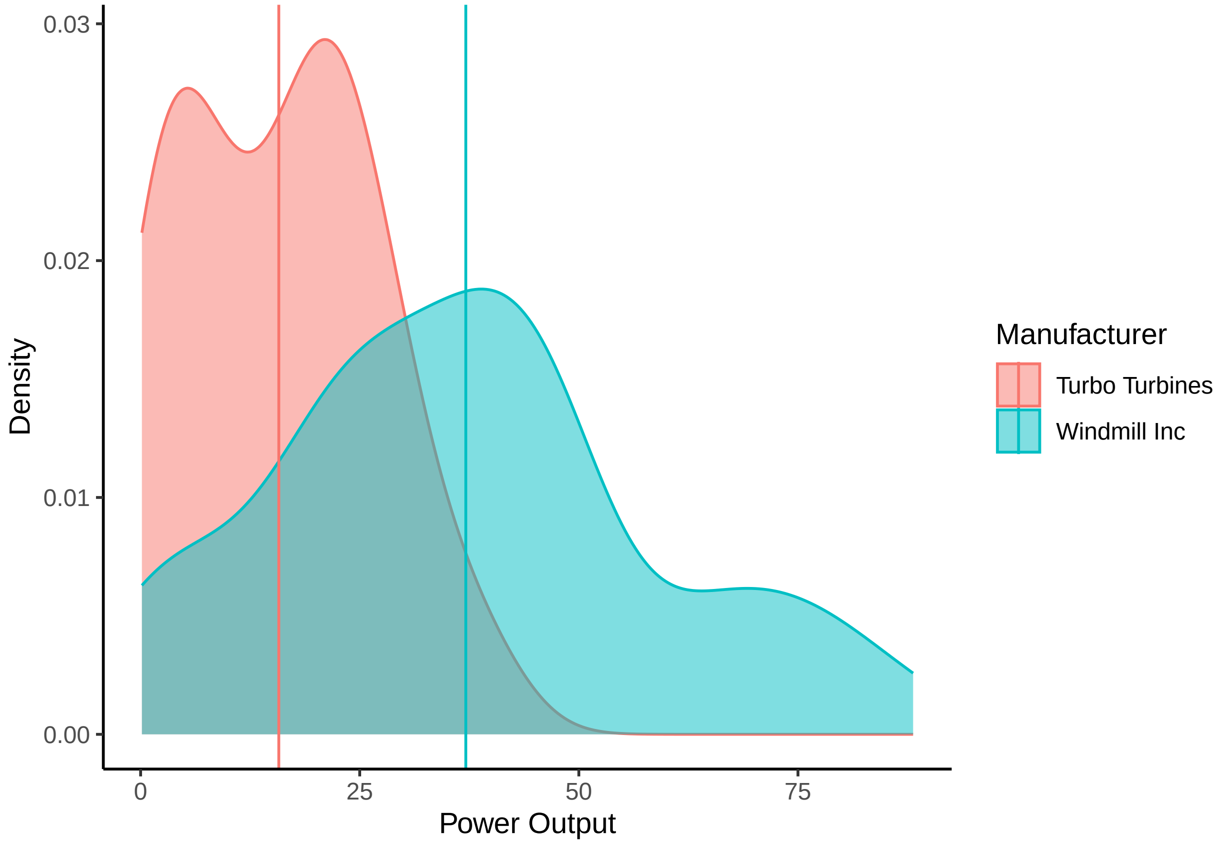
Visual

3. Create a density plot for the power output variable. (3 points)

    - be sure to have a density plot for each turbine producer; the color and the fill should be determined by the maker of the turbine
    - add in vertical lines for the mean values in the same color as the turbine makers

- make sure the x-axis, y-axis, and legend labels are capitalized and easier to understand (power output in measured in kilowatts, or kWh)
- use the `theme_classic()` function

```
ggplot(turbines, aes(power_output, color = manufacturer, fill = manufacturer)) +
  geom_density(alpha = 0.5) +
  geom_vline(data = turbine_summary, aes(xintercept = mean_power, color = manufacturer)) +
  labs(x = "Power Output", # any way of adding cleaner labels is fine
       y = "Density",
       color = "Manufacturer",
       fill = "Manufacturer") +
  theme_classic()
```



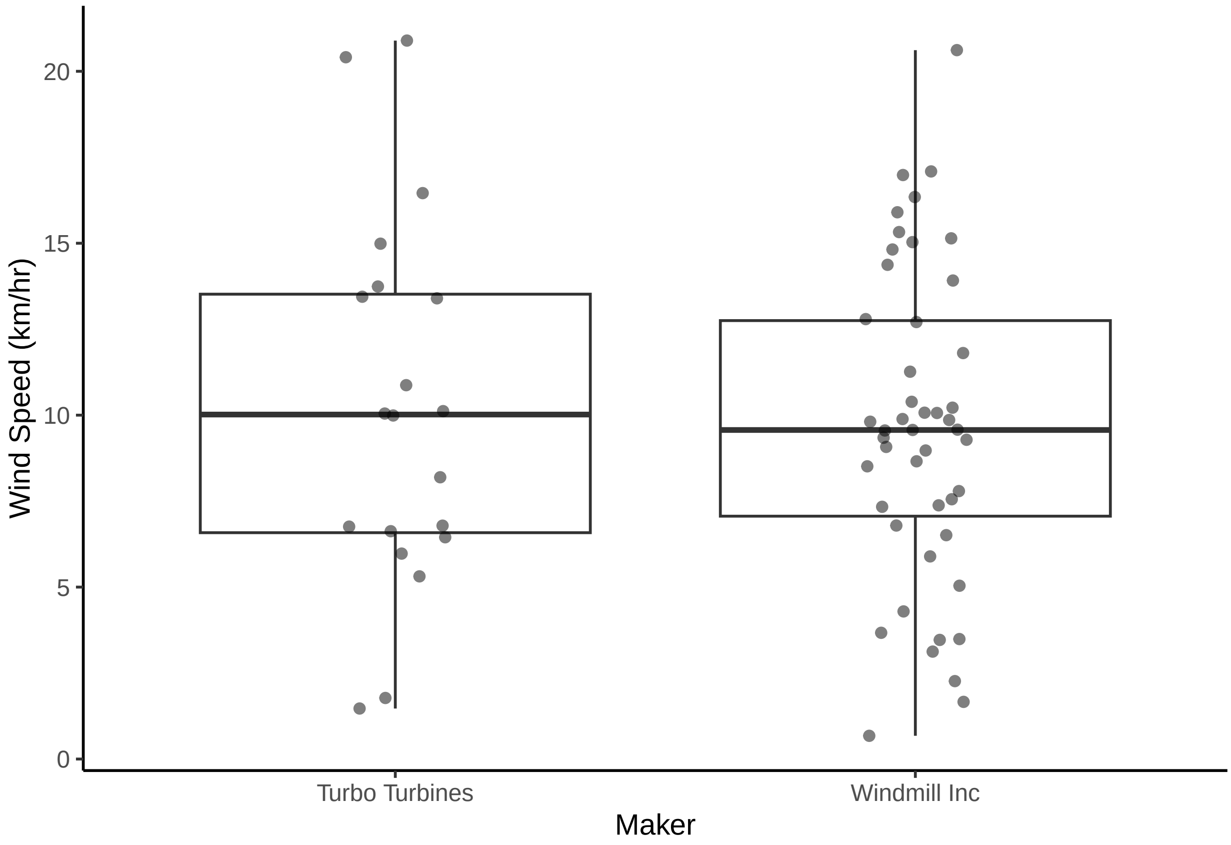
```
### the answer key that I gave them accidentally had wind speed as the x-axis label
### because of that, they can use wind speed OR power output in the question
```

4. Generate a box-and-whisker plot using `ggplot2` that compares the wind speed between different turbine makers (3 points).

   The plot should:

   - have capitalized and more descriptive axis labels (hint: wind speed is measured in kilometers per hour—km/hr)
   - show raw data points in addition to the boxes. The points should be jittered.
   - use the `theme_classic()` function

```
ggplot(turbines, aes(manufacturer, wind_speed)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, alpha = 0.5) + #transparency & width are optional
  labs(x = "Maker", # any way of adding cleaner labels is fine
       y = "Wind Speed (km/hr)") +
  theme_classic()
```



Statistic

5. Write a null hypothesis and an alternative hypothesis for the question we are asking and that we will be using statistics to answer. (2 points)

   Null Hypothesis (H_0): there is no difference in power output and wind speed between the turbine makers **Alternative Hypothesis** (H_A): there is a difference in power output and wind speed between the turbine makers

6. Based on the mean values in the `turbine_summary` data frame and the plots you've created above, predict the outcome of each t-test (graded for completion, not accuracy). Explain your reasoning (1-2 sentences for each t-test is fine). (2 points)

   Answer: graded for completion only

   - power output (histogram) looks like probably yes, $p < 0.05$; wind speed (boxplot) looks like maybe no

7. Perform a t-test on the power output by turbine maker. (1 point)

```r
t.test(data = turbines, power_output ~ manufacturer)
```

```
##
##  Welch Two Sample t-test
##
## data:  power_output by manufacturer
## t = -5.2832, df = 62.905, p-value = 1.686e-06
## alternative hypothesis: true difference in means between group Turbo Turbines and group Windmill Inc
## 95 percent confidence interval:
##  -29.40840 -13.26639
## sample estimates:
## mean in group Turbo Turbines   mean in group Windmill Inc
##                     15.76615                     37.10355
```

8. In 2-3 sentences, interpret the output from question 7. Focus on what the p-value is in reference to the cutoff of 0.05, what that means, and whether that means we accept or reject the null hypothesis. (2 points)

   Answer: $p < 0.05$ so there is a significant difference and we reject the null

9. Perform another t-test, this time on the wind_speed variable by manufacturer. (1 point)

```r
t.test(data = turbines, wind_speed ~ manufacturer)
```

```
##
##  Welch Two Sample t-test
##
## data:  wind_speed by manufacturer
## t = 0.38194, df = 31.02, p-value = 0.7051
## alternative hypothesis: true difference in means between group Turbo Turbines and group Windmill Inc
## 95 percent confidence interval:
##  -2.290743  3.346450
## sample estimates:
## mean in group Turbo Turbines   mean in group Windmill Inc
##                    10.185139                     9.657286
```

10. In 2-3 sentences, interpret the output from question 9 (focus on the same ideas as question 8). (2 points)

    Answer: $p > 0.05$ so there is no difference and we fail to reject (it's ok if they say accept) the null