Module 3: Assignment 3

Ellen Bledsoe

2023-10-31

Assignment Details

Purpose

The goal of this assignment is to assess your ability to use the appropriate statistical test for comparing means between groups and make decisions based on the interpretation of the results.

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 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

November 7 at midnight MST

Assignment Questions

For this assignment, we will be using some data about solar panels.

In addition to our wind turbines that are creating energy, we are also testing out solar panels from many different companies. We want to choose the company that makes solar panels that produce the most watts per hour.

We've set up solar panels on two different sides of the station: the north and the south. The plan is to first compare the solar panels companies on the north side to see which company's panels are creating the most energy. From there, we will do the same with the solar panel companies from the southern side.

Once the company that produces the most energy has been determined for each side, we will run another test with the top two companies in the same location and choose the best one.

(Note: This is pretty clearly not the best way to actually test solar panels, but it works best for the assignment. . .)

Set-Up

1. As always, let's read in our first data set, `solar_panels.csv`. Remember to also load the `tidyverse` and to give the data frame a name!

```
library(tidyverse)
solar_panels <- read_csv("../data/solar_panels.csv")
solar_panels
```

```
## # A tibble: 250 x 4
##    panel_id company    direction watts_per_hour
##       <dbl> <chr>      <chr>              <dbl>
##  1        1 Solar Gain North               311.
##  2        2 Solar Gain North               309.
##  3        3 Solar Gain North               304.
##  4        4 Solar Gain North               309.
##  5        5 Solar Gain North               304.
##  6        6 Solar Gain North               304.
##  7        7 Solar Gain North               300.
##  8        8 Solar Gain North               308.
##  9        9 Solar Gain North               314.
## 10       10 Solar Gain North               304.
## # i 240 more rows
```

Once you bring in the data, take a look at it so you understand what each row means and what all the columns mean.

The North Side

We are going to start with data comparing the watts per hours from solar panels on the north side of the station.

2. As you might have noticed in the data you read in, there is a column that tells us whether the solar panels are located on the north side or the south side. Create a data frame that includes only solar panels that are on the north side (remember `filter()`?). Save the output as a new data frame called `north`.

```
north <- solar_panels %>%
  filter(direction == "North")
```

3. Based on our task at hand, identify our dependent and independent variables:

   - **independent:** maker
   - **dependent:** watts_per_hour

4. Let's first summarize our data so we know what we are working with. Calculate the mean number of watts per hour produced by each solar panel company on the north side.

```
north %>%
  group_by(company) %>%
  summarise(mean = mean(watts_per_hour))
```

```
## # A tibble: 2 x 2
##   company        mean
##   <chr>         <dbl>
## 1 Solar Gain     308.
## 2 Sunny Side Up  326.
```

5. Write out your two statistical hypotheses for the test:

   - **Null Hypothesis:** there is no difference in the watts per hour produced by the two companies
   - **Alternative Hypothesis:** there is a difference in the watts per hour produced by the two companies

6. Write code to run the appropriate test. If this test involves a post-hoc comparison, add that line of code, as well.

   Hint: remember to use the correct data frame!

```
# students should be using a t-test here, not an ANOVA
t.test(watts_per_hour ~ company, data = north)
```

```
##
##  Welch Two Sample t-test
##
## data:  watts_per_hour by company
## t = -5.3614, df = 52.216, p-value = 1.903e-06
## alternative hypothesis: true difference in means between group Solar Gain and group Sunny Side U
## 95 percent confidence interval:
##  -24.94567 -11.35903
## sample estimates:
##    mean in group Solar Gain mean in group Sunny Side Up
##                    308.1425                    326.2948
```

7. Interpret the results of the statistical test you ran in question 6 above. In your answer, be sure to include whether the p-value is greater or smaller than our significance level (0.05), what that means in terms of statistical significance, and whether we should reject or fail to reject the null hypothesis. If your test involved a post hoc test, interpret those results as well (which companies are significantly different from each other, if any). (2 points)

   Answer: $p < 0.05$ so there is a significant difference so we reject the null—there is a difference in means between the companies; no post-hoc

8. Using your results from this entire section, do the solar panels from one company obviously outperform the others? If so, which company? Explain your rationale.

   Answer: Sunny Side Up is the winner (higher mean, statistically significant so difference is real)

The South Side

Now, we are going to test the differences in the number of watts per hour for the solar panel companies we placed on the south side of the station.

9. Create a data frame for the solar panels that were placed on the south side, and name the data frame `south`.

```
south <- solar_panels %>%
  filter(direction == "South")
```

10. Calculate the average number of watts per hour generated for each company.

```
south %>%
  group_by(company) %>%
  summarise(mean = mean(watts_per_hour))
```

```
## # A tibble: 3 x 2
##   company              mean
##   <chr>               <dbl>
## 1 Bright Future        309.
## 2 Function in the Sun  318.
## 3 Panel du Soliel      307.
```

11. There is one company that has a higher average, but is that higher average a real one? Run the correct statistical test to find out. If that test involves a post-hoc test, run that as well.

```
# half point for running ANOVA, half point for running TukeyHSD
mod <- aov(watts_per_hour ~ company, data = south)
summary(mod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## company       2   3095  1547.6   7.238  0.001 **
## Residuals   147  31432   213.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = watts_per_hour ~ company, data = south)
##
## $company
##                                           diff        lwr       upr     p adj
## Function in the Sun-Bright Future     8.347326   1.422886 15.271767 0.0136101
## Panel du Soliel-Bright Future        -2.198102  -9.122542  4.726338 0.7331384
## Panel du Soliel-Function in the Sun -10.545428 -17.469868 -3.620988 0.0012312
```

12. Interpret the results of the statistical test you ran in question 11 above. In your answer, be sure to include whether the p-value is greater or smaller than our significance level (0.05), what that means in terms of statistical significance, and whether we should reject or fail to reject the null hypothesis. If your test involved a post hoc test, interpret those results as well (which companies are significantly different from each other, if any). (2 points)

    Answer: overall test is significant ($p < 0.05$ so reject null). Performed pairwise comparisons, and Function in the Sun is significantly different from the other two companies (which are not significantly different from each other).

13. Using your results from this entire section, is one company clearly above the others or is that not very clear? If so, which company? Explain your rationale.

    Answer: Function in the Sun is significantly different from the other two companies and has the higher wattage, so it is the winner.

The Final Test

We are finally ready to pit our two top companies against each other! We've moved both of their solar panels over to the north side and have re-run the experiment. We need to read in the second data set for the results of this second experiment.

```
final <- read_csv("../data/final_solar.csv")
```

```
## Rows: 100 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): company, direction
## dbl (2): panel_id, watts_per_hour
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

14. Let's summarize! Calculate the means for the companies that are in the final test of which solar panels are best.

    ```
    final %>%
      group_by(company) %>%
      summarise(mean = mean(watts_per_hour))
    ```

    ```
    ## # A tibble: 2 x 2
    ##   company              mean
    ##   <chr>               <dbl>
    ## 1 Function in the Sun  321.
    ## 2 Sunny Side Up        330.
    ```

15. Run the correct statistical test to determine if there is a real difference between the companies. If the test involves a post-hoc test, include that code, as well.

    ```
    t.test(watts_per_hour ~ company, data = final)
    ```

    ```
    ##
    ##  Welch Two Sample t-test
    ```

```
##
## data:  watts_per_hour by company
## t = -1.9632, df = 62.558, p-value = 0.05407
## alternative hypothesis: true difference in means between group Function in the Sun and group Sun
## 95 percent confidence interval:
##  -17.5627392   0.1568707
## sample estimates:
## mean in group Function in the Sun       mean in group Sunny Side Up
##                           321.0573                           329.7602
```

16. Interpret the results of the statistical test you ran in question 12 above. In your answer, be sure to include whether the p-value is greater or smaller than our significance level (0.05), what that means in terms of statistical significance, and whether we should reject or fail to reject the null hypothesis. If your test involved a post hoc test, interpret those results as well (which companies are significantly different from each other, if any). (2 points)

    Answer: $p > 0.05$, so not significantly different and we do not reject the null.

17. Using your results from this entire section, is one company a clear winner? If so, which company? Explain your rationale. What if the cost of one company's solar panels were dramatically cheaper? Would that change our decision?

    Answer: No significant difference in wattage, so go with the cheap ones!