

BIFS-613

Assigned Exercise (AE) #2

110 pts

Due: Sunday 09/22/19 by 11:59 PM EST (end of Week 2)

Purpose: Practice data visualization using ggplot using expression (“exprs”) data for *TP53* tumor suppressor gene in NCI-60 cancer cell lines.

Submit three files:

1. Part I as separate file named: “LastName_FirstName_AE2.Rmd”
 2. Part II as separate file named: “LastName_FirstName_AE2.html”
 3. Part III as “LastName_FirstName_AE2.pdf”
 - a. Insert your answers into **AE2_template.docx** (in Week 2 Course Content) and then save/export as pdf
-

PART I: Create RMarkdown file (75 pts total)

Download the **BIFS613_AE2.Rmd** file (in Week 2 Course Content). Then open the Rmd file in RStudio. Download the **TP53_exprs.csv** file (in Week 2 Course Content), move to the same director/folder as you have Rmd saved. The data will be loaded in code chunk 2 (code already provided in Rmd).

A. Compare expression of TP53 across cancer types/tissues (50 pts total)

Code chunk #3 shows a “bad” barplot. Note that it can be plotted either using `geom_col()` or `geom_bar(position = “identity”)`

1. Code chunk #4 - Instead of barplot, use a boxplot. Hint: use `geom_boxplot()` (5 pts)
2. Code chunk #5 - Add code to plot the individual sample points on top of the boxplot. Hint: add `geom_point()` to code used in chunk #4 (you want both the boxplot and the individual sample points) (5 pts)
3. Code chunk #6 - Instead of `geom_point()`, use `geom_jitter()` (5 pts)
4. Code chunk #7 - Add `outlier.shape = NA` to `geom_boxplot()` (5 pts)
5. Code chunk #8 - Add code to indicate the sex of the patient. Hint: add `color = sex` in the aesthetics of `geom_point()` (5 pts)
6. Code chunk #9 - Use custom colors for sex: red for females and blue for males. Hint: use `scale_color_manual()` (5 pts)
7. Code chunk #10 - Apply a theme of your choice (list here: <https://ggplot2.tidyverse.org/reference/ggtheme.html>) (5 pts)
8. Code chunk #11 - Improve readability of the x-axis text by angling the names. Hint: in `theme()` use `angle`, `hjust`, and/or `vjust` (5 pts) - *google for help!*
9. Code chunk #12 - Add a title to the plot using `ggtitle()` (5 pts)

10. Code chunk #13 - Save the plot as “TP53_exprs_tissue.png” Hint: use `print()` and `ggsave()` (5 pts) - *google for help!*
11. **Extra credit:** Add figure legend to final figure in code chunk #13. Hint: use the `fig.cap` chunk option (5 pts)

B. Compare TP53 expression to patient age (25 pts total)

1. Code chunk #14 - Plot age (x) vs. exprs (y) scatterplot using `geom_point()` (5 pts)
2. Code chunk #15 - Fit a line to the data using linear regression model (`lm`) and `geom_smooth()` (5 pts) - *we'll discuss linear regression later this semester*
3. Code chunk #16 - Fit a separate line for each sex. Hint: use aesthetic `color = sex` (5 pts)
4. Code chunk #17 - Modify this plot to have similar style as the “TP53_exprs_tissue.png” plot above: add the same theme and the custom colors for each sex (female = red; male = blue). Also add a title. (5 pts)
5. Code chunk #18 - Save the plot as “TP53_exprs_age.png” (5 pts)
6. **Extra credit:** Add figure legend to final figure in code chunk #18. (5 pts)

➤ Submit your final Rmd code as “LastName_FirstName_AE2.Rmd”

PART II: Generate report from RMarkdown file (10 pts total)

Click “knit” button at the top of Rmd file, and select “knit to html.” Make sure the code runs correctly and you generate the html report. You will receive full credit if you have made the edits to the Rmd from above, have added your name and today’s date to the YAML, and the Rmd successfully generates your html report (even if incorrect code/plots). Note: The html will be in the directory (folder) where you have the Rmd saved.

➤ Submit the html report from your final Rmd code as “LastName_FirstName_AE2.html”

PART III: Interpretation of results in report (25 pts total)

➤ Answer the following questions. Remember, it’s OK to google!

- A. What information does the boxplot tell you that the barplot doesn’t? Hint: What value is the barplot plotting? What values are represented by the boxplot’s box, lines (“whiskers”), and center line? (5 pts)
- B. Compare non-jitter (`geom_point`) vs jitter (`geom_jitter`). Why did we use `outlier.shape = NA` for boxplot when plotting individual sample points on top of the boxplot and using jitter? Hint: look closely at the individual CNS sample data. (5 pts)

- C. Why are the tissues plotted in the order that they are on the x-axis of the “p53_exprs_tissue.png” plot? **(5 pts)**
- D. What question does the “p53_exprs_tissue.png” plot answer? **(5 pts)**
- E. What question does the “p53_exprs_age.png” plot answer? **(5 pts)**