

QBIO490: Multi-omic Data Analysis

Intro to Clinical Data: Part II

Due: Friday 2/24 @ 11:59PM

Deliverables:

1. Submit your code by creating an R script (not a Notebook!) called `Intro_to_Clinical_Data_II_yourname.R` and adding it to your GitHub.
 - a. Because this is a script, you will need to change your working directory via `setwd("/PATH/TO/analysis_data")`

Since this is an optional partner activity, it is okay if your answers are the same as your partner's. Each person must individually push their code to Github. At the top of your R Script, write the name of you and your partner.

Before you start:

1. Read in the clinical data csv you saved in Part I.
2. We will also be working with drug and radiation data. The data were already queried and downloaded in your `analysis_data` folder, just prepare them with the following two commands:

```
clinical_drug <- GDCprepare_clinic(query = clin_query,  
                                  clinical.info = "drug")
```

```
clinical_rad <- GDCprepare_clinic(query = clin_query,  
                                  clinical.info = "radiation")
```

Answer the following questions as comments in your R script:

1. Look at the different column names of the `clinical` dataframe. Choose one that is interesting to you and your partner. Ensure that there are not too many NAs in this column by using `is.na(clinical$COLUMN_NAME)`. Remember that in coding, TRUE is equal to 1 and FALSE is equal to 0. You can then use the `sum()` function to find how many TRUES exist. Which variable have you chosen?
2. Is your variable categorical, discrete, or continuous?
3. Look at the different column names of the `clinical.drug` and `clinical.rad` data frames. Choose a variable from one of these data frames. Ensure there are not too

many NAs (there will likely be more NAs in the drug and radiation dfs than in the patient data, don't worry about it too much). Which variable have you chosen? Provide a brief description of the variable.

4. Is your variable categorical, discrete, or continuous?
5. Scientists generate hypotheses before experimenting or exploring data. Generate three hypotheses: (1) Relate your variables to each other, (2) Relate your first variable to survival in breast cancer, (3) Relate your second variable to survival in breast cancer.

Coding Activity:

1. Create either a boxplot or scatter plot comparing your two selected variables. What information can you glean from your plot? Why did you choose one plot over the other?
 - Helpful functions/packages: `plot()`, `hist()`, `boxplot()`, `pairs()`, `ggplot2` package + associated functions
2. For your first variable, perform a survival analysis and create a Kaplan-Meier plot.
 - The survival analysis needs a categorical variable. If you have a continuous variable, use an `ifelse()` statement to create a new column with a categorical version of the variable.
3. For your second variable, perform another survival analysis and create a Kaplan-Meier Plot.
4. Analyze your two KM plots. What do the KM plots suggest about the impact of your variables on breast cancer survival? What are the p-values? Do the differences in survival appear to be significant?
5. Save your plots as jpegs or pngs.

Check before submitting:

You **must** include informative comments throughout your code.

```
str(clinical) # view structure of clinical data frame  
head(clinical) # view first few rows of clinical data frame
```

You **must** install and load all necessary packages at the top of your coding fall.

```
if (!require(package)){  
  install.packages("package")  
}  
  
library(package)
```

You **must** change your working directory at the top of your coding file.

```
setwd("/Users/nicoleblack/Desktop/QBIO/qbio_nicole/analysis_data")
```

You **must** be able to run your script from top to bottom (with a clean environment) without any issues.

- Before turning it in, hit the broom in the top right corner of Environment to clear all values and data. Then run the entire script by hitting the run button in the top right of your source panel. Your code should run all the way through with no errors.