# Module 2 - MATH 241

## Contents



*Posted: February 15, 2022*

**Due: February 25, 2022**

**Instructions:**

- **Please knit your R Markdown file as PDF file.** Don't put your name in any part of the document. Your document upload will correspond to your name automatically in Gradescope.
- **Please provide complete solutions for each problem.** If it involves mathematical computations, explanations, or analysis, please provide your reasoning or detailed solutions.
- **It is recommended that you work on the module incrementally.** The modules are designed for you to work on it at anytime until the designated deadline. This method will let you transfer what you have learned from lectures and mini-assignments in small parts rather than stressing on completing the entire assignment at the last minute.
- **Note that some problems have multiple solutions or ways to solve it.** Make sure that your solutions are clear enough to showcase your work and understanding of the material.
- **Creativity and collaborations are encouraged.** Use all of the resources you have and what you need to complete the module. Each student must take personal responsibility and submit their work individually. Please abide by the Reed College Honor Principle.
- If you can't figure out why a code chunk is preventing you from knitting the document, replace "r" at the top of the code chunk with "r eval = FALSE, echo = TRUE." The code will not be executed, but it will be printed in your pdf, earning you some partial credit.
- Please follow the general module guidelines written in the course website. If you have any questions, please send them to the instructor as a direct message on Slack or through email.

**R Packages:**

- Below are pre-loaded general packages required for this module assignment. You can load more packages here or throughout the module if necessary.
- Note that you need to install R packages before you can use them. You can use the `install.packages()` in the R console, or go to the "Tools" tab and click "Install Packages..." in R Studio.
- Be careful on loading R packages because sometimes any two packages can have conflicting functions when calling them.

```
# pre-load packages here
library(tidyverse)
```

# I. The Art of Data Visualizations (Homework)

## Materials

According to the book *Visualize this : the FlowingData guide to design, visualization, and statistics* by Nathan Yau Yau (2011), data visualization is both an art and a science. Recent advances in technology and computer science had made possible for us to present data in more efficient and compelling ways. Sometimes a good visualization speaks for itself without any further explanations but a more complicated data set or statistical analysis requires more than a visualization.

In this section of the module, you will use these embedded figures below to apply the basic foundations of data visualizations.

- ***US Life vs Death 2003-2021 by Nathan Yau.*** The file `us-life-vs-death-2003-2021-nyau.png` is the PNG file used for the for the embedded image below. This figure was taken from the online post, When there were more deaths than births in the U.S. by Nathan Yau. This figure compares the number of births vs the number of deaths in each year from 2003 to 2021 in the US. It shows the time-series evolution of births and deaths while highlighting their differences.

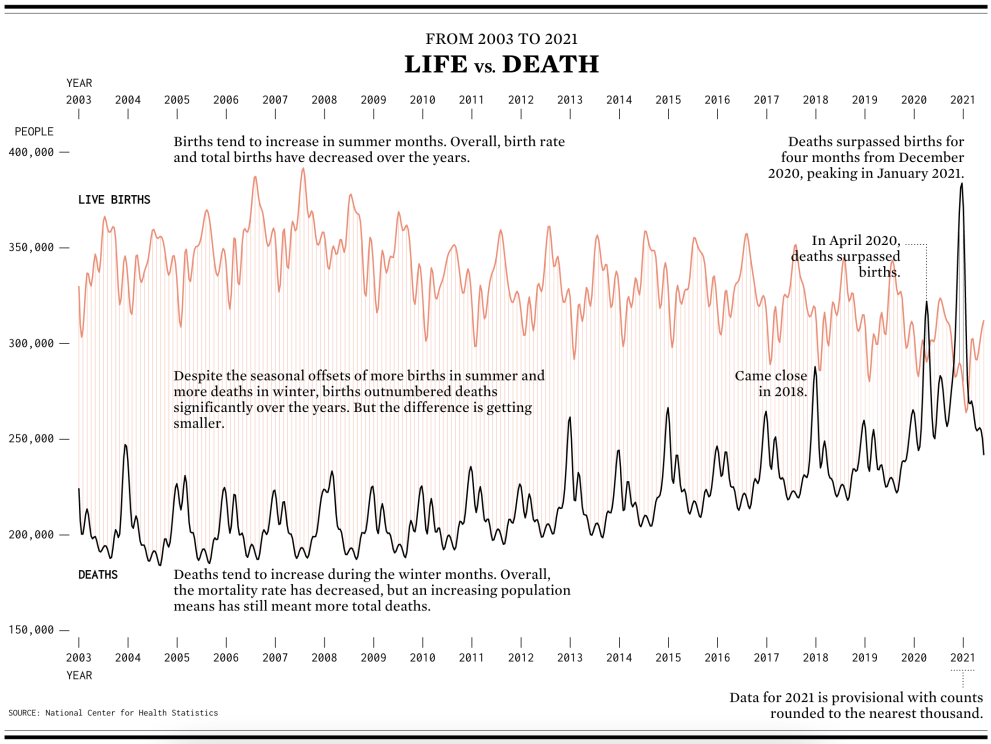


Figure 1: Life vs Death 2003-2021 by Nathan Yau.

- ***Life Expectancy at Birth for U.S. States and Census Tracts, 2010-2015.*** The National Center for Health Statistics (NCHS) at the Center for Disease Control and Prevention (CDC) provides a static map of the US colored according to life expectancy rates by state and by county (Tejada-Vera et al., 2020). The file `us-life-expectancy-2010-2015-cdc.png` is the PNG file used for the embedded image below.

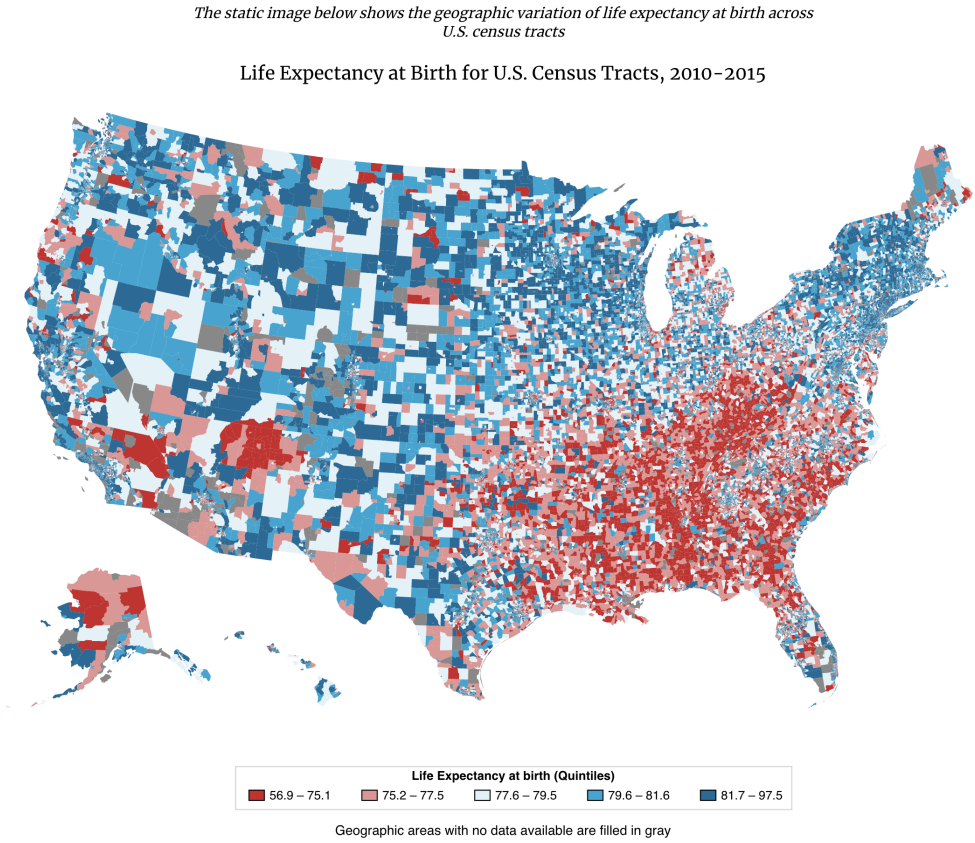


Figure 2: Life Expectancy at Birth for U.S. States and Census Tracts, 2010-2015.

**Exercises**

1. Using Fig. 1, apply the foundations of data visualizations to evaluate the plot, reflect on it, and provide a paragraph for each item below. Please look back from the lecture slides and mini-activities to get a sense of what these words mean in the context of data visualizations, and to help you write something about each item.

    a. Verification
    b. Dimensions
    c. Aesthetics
    d. Interpretations and Intentions

2. Using Fig. 2, answer the following questions.

    a. What are the variables presented in this Figure? Please provide details on what types of variables involved in this Figure.
    b. What are your observations regarding patterns and distributions of life expectancies across the US?
    c. How does the contrasting colors direct you to your observations? What are the weaknesses of using colors on visualizing data? Explain why.
    d. Can geography solely explain on why some regions in the US have low or high life expectancy? Explain why and provide any confounding variables that might explain such phenomenon. *Hint: You can add comments about poverty, education, unemployment, and/or healthcare.*

## II. Data Wrangling Utilizing Visualizations (Homework)

### Materials

This section of the module uses the following two independent data sets.

- The `county_complete` data set from the `usdata` package provides county level information. This data frame has 3142 observations - which corresponds to the counties in the US - with 188 variables. In this section, we are only interested in variables in 2015 because it corresponds to our next data set of interest. This subset of the `county_complete` data has the following variables:
  - `state` - The state of the county.
  - `name` - The name of the county.
  - `pop2015` - The county's 2015 population.
  - `civilian_labor_force_2015` - The county's number of people available for work in 2015.
  - `employed_2015` - The estimated employed individuals in 2015.
  - `unemployed_2015` - The estimated unemployed individuals in 2015.
  - `unemployment_rate_2015` - The estimated unemployment rate in 2015.

```
# load usdata package
library(usdata)

# Get 2015 subset of the county_complete data
cc_sub <- county_complete %>%
            select("state"|"name"|ends_with("2015")) %>%
            drop_na() %>%
            tibble()
```

- The file `us-life-expectancy-2010-2015-cdc-data.csv` is a dataset in a CSV file containing life expectancy at birth in 2010-2015 by state and by county from NCHS (Tejada-Vera et al., 2020). This dataframe has 6 variables and 73121 rows - which corresponds to a geographic area within each county in the US. This data has the following variables:
  - `state` - A nominal categorical variable with levels as the names of US states.
  - `county` - A nominal categorical variable with levels as the names of US counties.
  - `census_tract_number` - A tract number is a geographic area designated for census purposes.
  - `life_expectancy` - A numerical variable of the life expectancy estimate of an area.
  - `life_expectancy_range` - A pseudo-numerical variable where it contains confidence intervals of life the expectancy of an area.
  - `life_expectancy_standard_error` - A numerical variable of the life expectancy standard errors of an area.

```
# Load the CSV data
life_expt <- read_csv("us-life-expectancy-2010-2015-cdc-data.csv") %>%
              drop_na()
```

**Exercises**

This section serves as a continuation of practicing data wrangling. The exercises below asks you to create plots. In each plot, make sure you add descriptive titles and the labels for each variable is readable, and the R code within the code blocks are properly commented and wrapped around the margins.

1. Using the `cc_sub` data set, create a horizontal bar graph where the x-axis is the unemployment rate, y-axis is the states ordered by unemployment rate, and each bar is colored according to the log10-transformed total population by state. You can use the `geom_col()` to create a horizontal bar graph and use the `log10()` function for the transformation. Describe your observations of the resulting plot. Hint: `fill = log10(population)` with label "log10(population") or `scale_fill_gradient(trans = "log10")` with label "log10-scaled population."

2. Using the `life_expt` data, examine the `County` variable. You might notice that the levels of this variable contains the county name and the abbreviation of the state, as well as each county is repeated because each tract number is a geographic area within that county. Below are steps for creating a subset of this data so that we can use it together with the `cc_sub` data set.

    a. Create a subset where you `select()` the `state`, `county`, and the `life_expectancy` variables. Name this subset `life_expt_sub`. Create a new column named `le_mean_rep` that calculates the mean `life-expectancy` for each county. This will create the column with repeated mean values for each repeated counties. The resuling subset shoud have all four columns and the number of rows remains the same. Hint: `select(...) %>% group_by(...) %>% mutate(...)`
    b. Using the resulting subset from part a, group the dataset by county and use the `summarise()` function to create a new column named `mean_life_expectancy` and `state`. Here, you will use the `min()` function on the `le_mean_rep` and `state` variables to choose one value for each repeated group from the the `le_mean_rep` variable. The resulting data frame should have 3108 rows and 3 columns (county, state, and mean_life_expectancy). Below is a sample of what the resulting data subset `life_expt_sub` should be. Hint: `group_by(...) %>% summarise(...)`

3. Using the `life_expt_sub` data from problem 2, below are the steps for removing the state abbreviations of the county names.

    a. Apply the function `str_sub()` from the `stringr` package into the `county` variable to remove the last four chararcters of each string. Update the existing `county` variable when you apply the function. The county names should have no state abreviations and commas as a result. Hint: `str_sub(county,0,nchar(county)-4)`
    b. Create a horizontal boxplot showing the distributions of life expectancies by state. Make sure the boxplots are in descending order according to the median. Describe your observations of the resulting plot. Hint: `fct_reorder(state,mean_life_expectancy)` in combination with `ggplot2` will automatically sort the boxplots in descending order according to the median for each state.

## III. Statistical Thinking Utilizing Visualizations (Homework)

### Materials

This section of the module uses the following two independent data sets.

- The `cc_sub` data frame, which is the subset of the `county_complete` data frame from the previous section.
- The `life_expt_sub` data frame, which is the subset of the `life_expt` data frame from the previous section.

### Exercises

This section serves as a review on applying basic inferential methods learned from probability and statistics, as well as more practice on data wrangling and visualizations. The exercises below asks you to create plots. In each plot, make sure you add descriptive titles and the labels for each variable is readable, and the R code within the code blocks are properly commented and wrapped around the margins.

1. This problem uses both the `cc_sub` data frame and `life_expt_sub` data frame. Create a new data frame that joins the two data frames and name it `df`. Below are steps to accomplish this.
    a. Use the `full_join()` function to combine both datasets matching both county names and state. Drop any rows that has missing values. The resulting data frame should have 3102 rows and 8 variables.
    b. Use the county level data frame named `df` to create a scatter plot where the x-axis is the unemployment rate and the y-axis is the mean life expectancy. Describe your observations of the resulting plot, specifically by identifying any outliers and the association/relationship between the variables with and without the outliers.
2. This problem uses the `df` data frame from problem 1.
    a. Create a data frame where you group the counties by state. For each state, sum up the population, sum the total civilian labor work force, take the mean unemployment rate, and the mean life expectancy.
    b. Using the state level dataframe from part a, create a scatter plot where the x-axis as the unemployment rate, y-axis as the life expectancy, the size of each point is proportional to the population, and the color for each point is proportional to the ratio, civilian labor workforce over population. Have the color palette to be a gradient from blue to red. You can use `scale_colour_gradient(low="blue",high="red")`. Describe your observations, and specifically discuss about any associations between all variables.
3. This problem uses the `df` data frame from problem 1. Using the state level data, apply inference for simple linear regression between the unemployment rate (explanatory variable) and the life expectancy (response variable). Create a scatter plot that shows the linear fit with the 95% confidence interval. Discuss and interpret your results in context, as well as discuss any confounding variables that might also be worth to include as another explanatory variable to predict life expectancy.

## References

Tejada-Vera, B., Arias, E., Escobedo, L., & Salant, B. (2020). Life expectancy estimates by us. Census tract, 2010-2015. *National Center for Health Statistics*. https://www.cdc.gov/nchs/data-visualization/life-expectancy/

Yau, N. (2011). *Visualize this : The FlowingData guide to design, visualization, and statistics*. Wiley Pub.