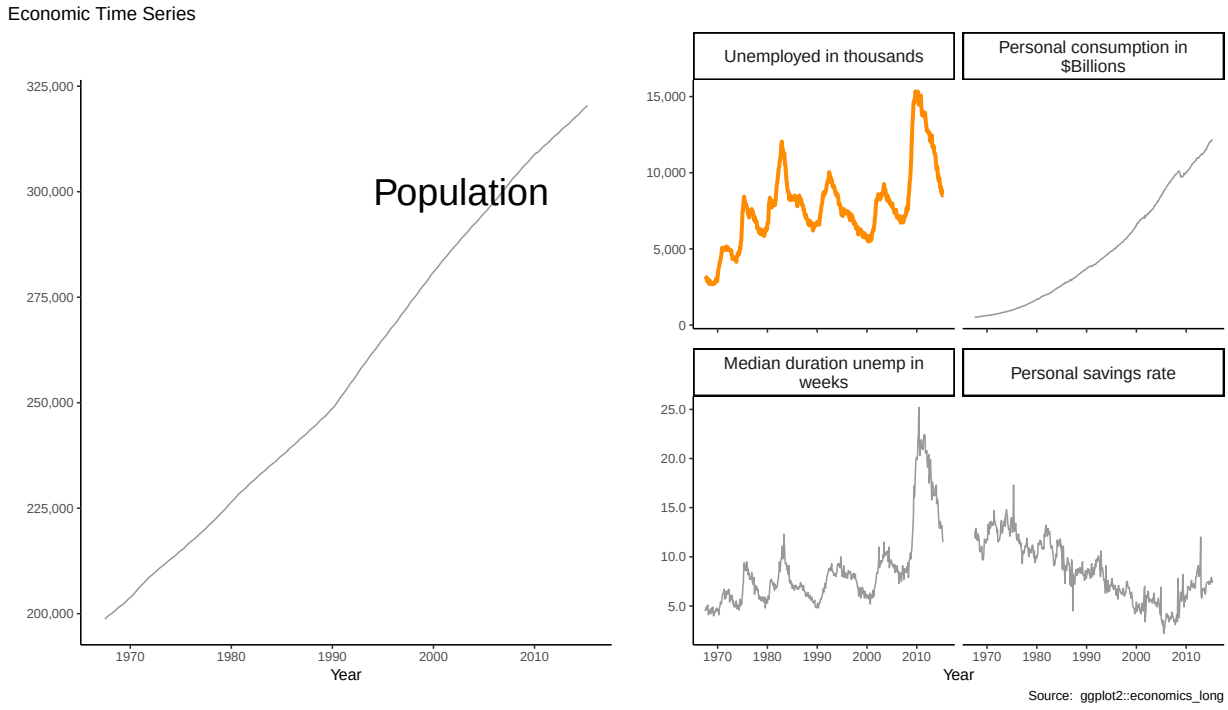
Unemployment over Time

An DIIG introduction to reproducilbe computational analysis and report writing

John Little

2021-02-05

R, RStudio, and the Tidyverse are three components of a modern computational environment that leverages important characteristics of Data Science. This tool suite is extensible but available to researchers working at various level of complexity. This example report demonstrates how computational analysis can be supported by a reproducible workflow. You will see how RStudio and the Tidyverse are geared towards reproducibility; can render reports in various formats; aids in the important and time consuming task of data wrangling (80% of any data project); brings sophisticated visualization functions; and leverages sophisticated data management techniques.



Explanation

This report was generated by leveraging the various Tidyverse packages listed, and briefly introduced, below.

dplyr

https://dplyr.tidyverse.org

Subset and sort

- `filter()` rows [observations]
- `select()` columns [variables]
- `arrange()` arrange rows by a variable

New variables and column totals

- `mutate()` create a new variable
- `summarize()` column totals
- `group_by()` when used with summarize, column subtotals

Add data from other data frames

- `left_join()`
- `anti_join()`

tidyr

https://tidyr.tidyverse.org

- `pivot_longer()`
- `pivot_wider()`

Visuzlization [ggplot2] & EDA

https://ggplot2.tidyverse.org https://docs.ropensci.org/skimr

- `skimr::skim()` - a good place to start for Exploratory Data Analysis
- `geom_histogram()`
- `geom_bar()` (and `geom_col()`)
- `geom_boxplot()`
- `geom_line()` (time series)
- `geom_point()`