Computational Finance
Practice set 1 - The Tidyverse

27th of February 2025

Pierfrancesco Alaimo Di Loro

For this exercise, we will use the data from the following publication:

"Kuo, Pei-Lun and Jager, Leah and Taub, Margaret and Hicks, Stephanie. (2019, February 14). opencasestudies/ocs-healthexpenditure: Exploring Health Expenditure using State-level data in the United States (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.2565307"

The healthcare policy in the United States is extremely complex. There are various forms of health insurance ranging from federal government healthcare programs to those of private insurance companies. Our interest is to understand, in broad terms, how insurance coverage and healthcare spending are distributed across various federal states. More specifically, the key questions are the following.

- Is there any relationship between the number of insured individuals (under various forms of insurance) and the total healthcare expenditure of a state?
- How does spending vary across different regions of the United States?
- Were there any variations between 2013 and 2014?

We will consider two datasets extracted from the Henry J Kaiser Family Foundation (KFF), which you can find on the Elearning platform.

1. Create a folder and name it *Practice2*. Inside it, create another folder and name it *Data*. Create two subfolders named *raw_data* and *tidy_data* and place the CSV files "healthcare-coverage.csv," "healthcare-spending.csv," and "stateNamesAndAbbs.csv" inside the *raw_data* folder.
   (a) Open *RStudio* and create a new project within the *Practice2* folder.
   (b) Create a new `Rscript` and require the `Tidyverse` package.
2. Read the data "healthcare-coverage.csv" and "healthcare-spending.csv" as `tibbles` into two objects named `datc` and `dats`. The first one contains the *number of individuals under various forms of insurance for each state and year*. The second one contains *healthcare spending in millions of dollars for each year and state*. There is something wrong with the reading...
   (a) Use the `read_lines()` function to read the various lines of the file in a *non-structured* way. Check the **first** few lines.

(b) The `read_csv()` function has an argument that allows you to *skip* a certain number of rows during reading.

(c) Take a full look at the read data again. There's something wrong with the last rows! Remove them or reread the data without reading them.

(d) Repeat the process for "healthcare-spending.csv".

(e) Save **both datasets** into a **single file** in *.rda* format using the `save()` command. You can name it "Practice2_data.rda".

(f) Save the script as *00data_import.R* and close it.

3. Create a new `Rscript` and load the usual packages. Let's focus on the *coverage*.

(a) Take a quick look at the data. Some columns that should be `double` are instead `character`. This happens because the *missing* values are recorded as "N/A".

- Use the `na_if()` function from `dplyr` to replace all "N/A" with `NA`.
- Convert all variables except `location` to numeric.

(b) The dataset has many columns referring to the same `type` of insurance but in different `years`. This **does not match the *tidy* format!**

- Consolidate the information related to these various columns into a single column. *long format* where the key is the pair `type_year` and the value is the `coverage`.
- Split `type` and `year` from the `type_year` column using `separate()`.

(c) For each year, recalculate the total for each state from the individual items. Verify if it matches the total already recorded in the data.

(d) They are not equal because our total is the **official population**, which may not coincide with the *insurable* population. Delete the rows with this information and add it as another column named `tot_pop`.

(e) Repeat the same steps for *spending*. Column for the type unnecessary here!

(f) Merge *coverage* and *spending* into a single `tibble` named `datAll`.

(g) Read the file "stateNamesAndAbbs.csv". Merge them with the `tibble` `datAll`.

(h) Since we want to perform the analysis at the individual state level, remove the row referring to the *United States*.

(i) Calculate the following quantities and add them as columns:

- The *percentage of individuals covered by each type of insurance* over the total population `coverage_perctot`
- The *percentage of individuals covered by any insurance* over the total population `covered_perctot` (consider that one type of coverage is *uninsured*)
- The *per-capita healthcare spending* of each state `spending_capita` (remember it's in millions)

(j) Count how many missing values there are in the dataset and on which variables.

(k) All NA values are in the same column. They all belong to the same `type` of coverage: eliminate all rows corresponding to that `type`.

(l) Save **the final dataset** into a **single file** in *.rda* format in the *tidy_data* folder using the `save()` command. You can name it "Practice2_data_tidy.rda". Save the script as *01data_clean.R* and close it.

4. Create a new `Rscript` and load the usual packages.

(a) Import the dataset in *tidy* and *cleaned* format from the previous step.

(b) Represent the distribution of various types of insurance across all United States (combining all states) in a bar chart.

(c) Represent, in a single chart, the same distribution varying by `region` of the state.

(d) Represent on the same panel but on different graphs the same chart for different years. Try using the `+ facet_wrap(~ year)` command!

(e) Represent the distribution of the *per-capita* public healthcare spending aggregating all the states. Choose whether to use *boxplot*, *histogram*, or *density plot*.

(f) Not really *Normal*. See what happens if you apply the *logarithm* to spending.

(g) Represent the same distribution varying by region and then compare it between 2013 and 2014 (on different charts)

(h) Check the relationship between *per-capita spending* (x-axis) and *percentage of covered population* (y-axis) through a *scatter plot*.

(i) Add a trend line that better captures the relationship (use `geom_smooth(method="lm")`) and color each point by *region*.

(j) Do the same for the two years separately (maybe using `facet_wrap()`).

(k) Represent the relationship between *the percentage of each type of insurance* (y-axis) and *healthcare spending* in 2014.

(l) Add, alongside each point, the state abbreviation using `geom_text`.

(m) Copy the commands from the graph in (i). Use `geom_path()` to connect the points referring to the same state in different years, coloring by region.

(n) Run the command `pdf("AllPlots.pdf")` in the console, run all the graphs in the script, and then run `dev.off()` in the console. Check what appeared in your project folder.7Save the script as *02data_viz.R* and close it.