# Set A1.01: Workflow from geoprocessing to producing R graph

A1.01\_US State by number of counties.csv

Based on this data set, we will draw a scattered plot.

A1.01\_df\_CountyN.R

We transform the State into the categorical variable. This step is required in producing a base R plot.

A1.01\_Change plot symbols for categorical variable.R

This R code addresses the challenge of plotting categorical data (US states) on the x-axis using the plot() function. The standard plot() function treats categorical data like numeric data, leading to incorrectly spaced "scattered bars" instead of distinct points for each state.

The code resolves this by:

1. **Creating an empty plot:** It first creates an empty plot with no data points, but with properly defined axes, labels, and limits.
2. **Adding points with symbols():** It then uses the symbols() function to place circular points at the correct coordinates for each state and its corresponding number of counties.
3. **Customizing x-axis labels:** Finally, it adds the state names as x-axis labels at their appropriate positions using the axis() function.

This approach effectively visualizes the relationship between US states and their number of counties with correctly spaced points on the x-axis, providing a clear and accurate representation of the categorical data.

A1.01\_Complete Code.R

This R code demonstrates the integration of ArcGIS and R for spatial data analysis and visualization. It begins by importing a shapefile of US counties from ArcGIS Online into R. The code then cleans and preprocesses the data, including renaming columns for easier handling and analyzing duplicate county names.

Next, it aggregates the data to count the number of counties within each US state. Finally, it visualizes this aggregated data using a customized plotting function that incorporates variable labels for clear and informative axis labels. This workflow showcases how R can be used to effectively analyze and visualize spatial data obtained from ArcGIS Online.

# Set B1.01: De-clustering the text entities for normalization

B1.01\_CACountyRegion\_dta.xlsx

This dataset provides a mapping of California regions to their constituent counties.

It will be transformed into a normalized county-level database, where each row represents a single county with an associated region. This normalized structure will facilitate easier analysis and data manipulation.

B1.01\_Counties and Regions.csv

This is the resulting normalized data set.

B1.01\_Create CA data set for nine counties.R

This R script de-clusters and normalizes a dataset of California regions and their associated counties. It reads an Excel file where each region is listed alongside a string of multiple counties. The script then separates these county strings to create a tidy dataset with one county per row, linked to its corresponding region. This normalized and de-clustered data is saved as a CSV file named "Counties and Regions.csv".

# Set B2.01: Data Simulation and Anonymization Function

B2.01\_s1.csv

The dataset is a simulated version of a real-world dataset. To protect privacy, identifiable information has been replaced with fake IDs. The remaining values have been slightly modified (within +/- 5% of their original values) to maintain the overall data distribution and patterns while ensuring anonymity.

B2.01\_combined function to simulate data and produce coy IDs.R

This R function (`sim\_data`) simulates a dataset by perturbing numeric values within a +/- 5% range and replacing identifying information with anonymized IDs. It takes a data frame and an ID prefix as input, applies the simulation and anonymization, and returns the modified data frame.