Investment Adviser Public Disclosure Data Visualization Report

This project analyzes data from the Investment Adviser Public Disclosure (IAPD) system, which contains records on Investment Adviser Representatives (IARs) registered with firms across the United States. The data includes a range of information, such as employment history, personal and registration information, exam and certification results, business affiliations, and regulatory disclosures.

The raw data was provided as structured XML files. A Python script was used to extract and organize this information into a set of normalized tables, each covering different aspects of the IARs—including employment, registration status, prior employment and registration history, exams, and disclosures. The output of this step was a collection of clean datasets, which served as input to the R-based data wrangling and analysis.

The R workflow picked up from the Python output and focused on the challenge of cleaning and standardizing employer organization names. Raw names were inconsistent and often included legal suffixes (e.g., “LLC,” “Inc”), duplicate spacing, punctuation, and brand variants. First, a custom function lowercased each name and remove punctuation. Second, common filler words were removed from tokenized names, helping isolate core identifiers like “morgan stanley” or “edward jones.”

These tokenized names were then grouped by common roots—primarily using the first token and word frequency across the dataset. Then search through company name tokens and group them if they share similar core structures. After grouping, a group_label was created for each group using the two most frequent tokens, and a lookup table was used to map each raw organization name to its cleaned and grouped counterpart. Finally, a set of case_when() rules was applied to assign a **standardized brand label** to a subset of 50 organizations (e.g., “Morgan Stanley”), ensuring consistency across variations.

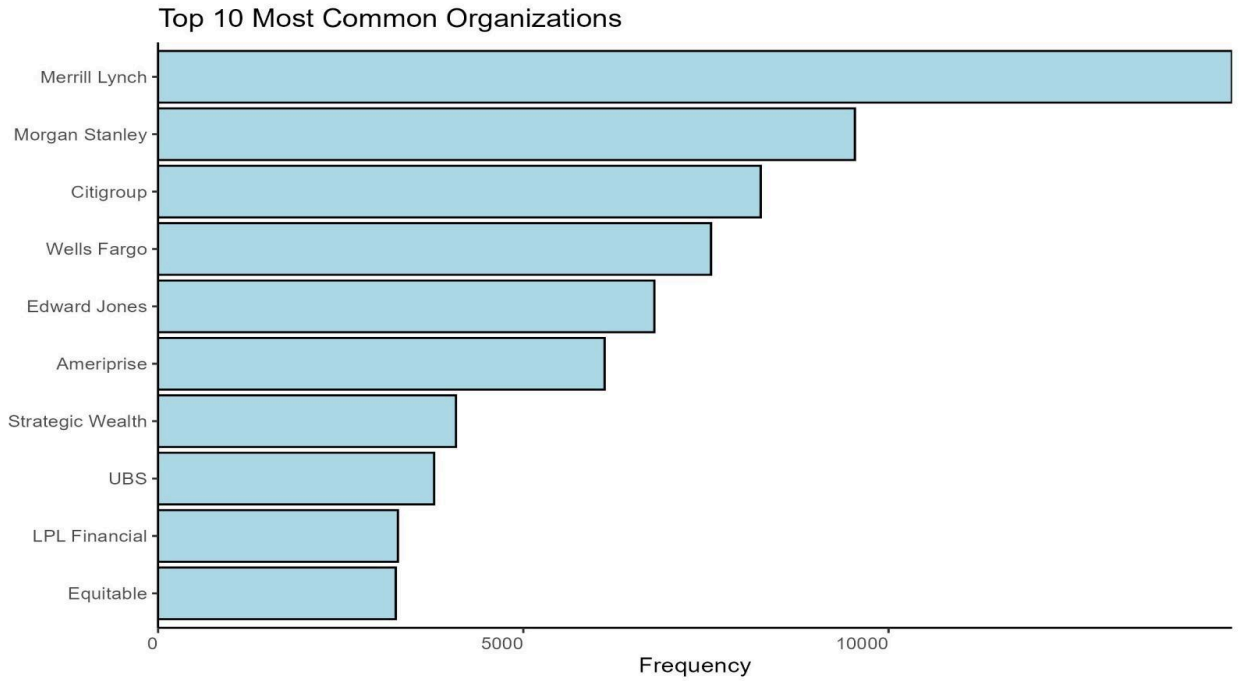
Data Enrichment and Visual Insights

Following standardization, the dataset was enriched with several derived fields. These included the average registration duration per individual, converted into years, as well as total registration counts and city coverage for each firm. Employment locations were used to extract U.S. state information, which was then mapped to census regions.

A variety of visualizations were developed to explore trends in the data. A bar chart showing the top 10 most common firms highlighted the most prominent players in the industry. Registration duration was analyzed using a histogram and boxplot to show the typical range and outliers in adviser registration tenure.

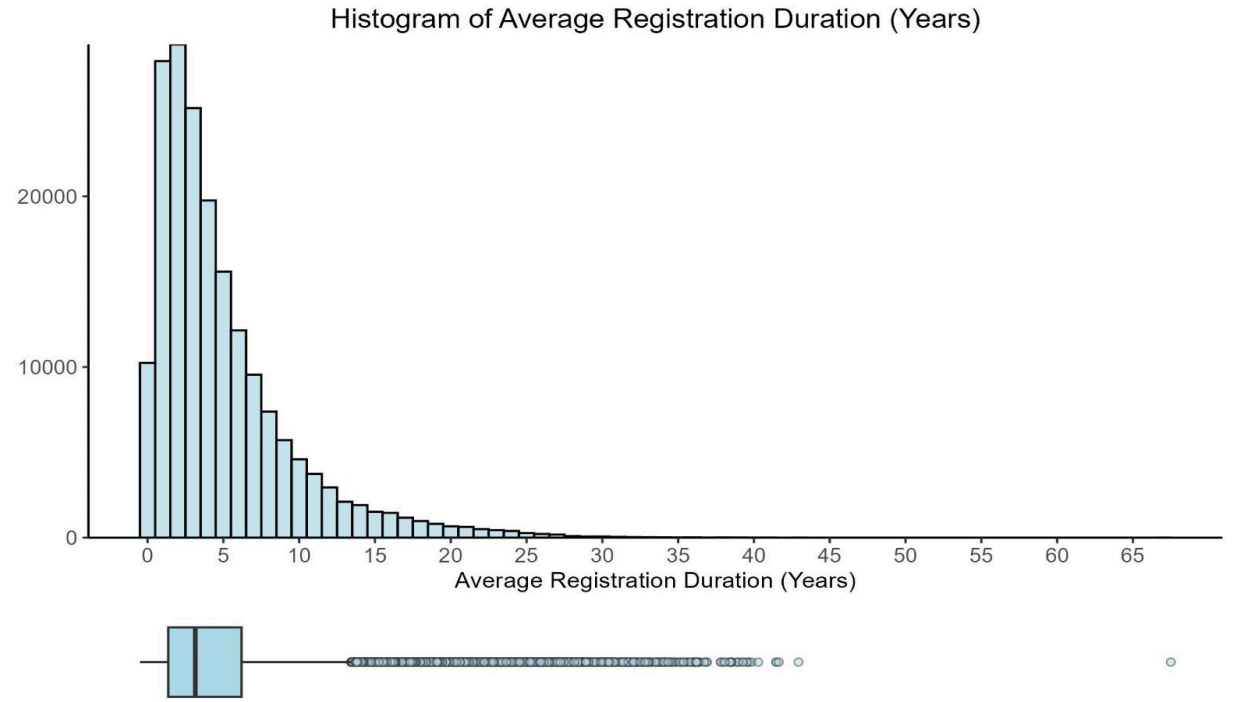
An interactive map created with Plotly shows the presence of the top five firms by IAR employment across U.S. states. Bubbles were placed based on state coordinates and adjusted with jitter to prevent overlap. Each bubble includes tooltips showing firm-level metrics such as number of individuals, and total registrations.

Graph 1:



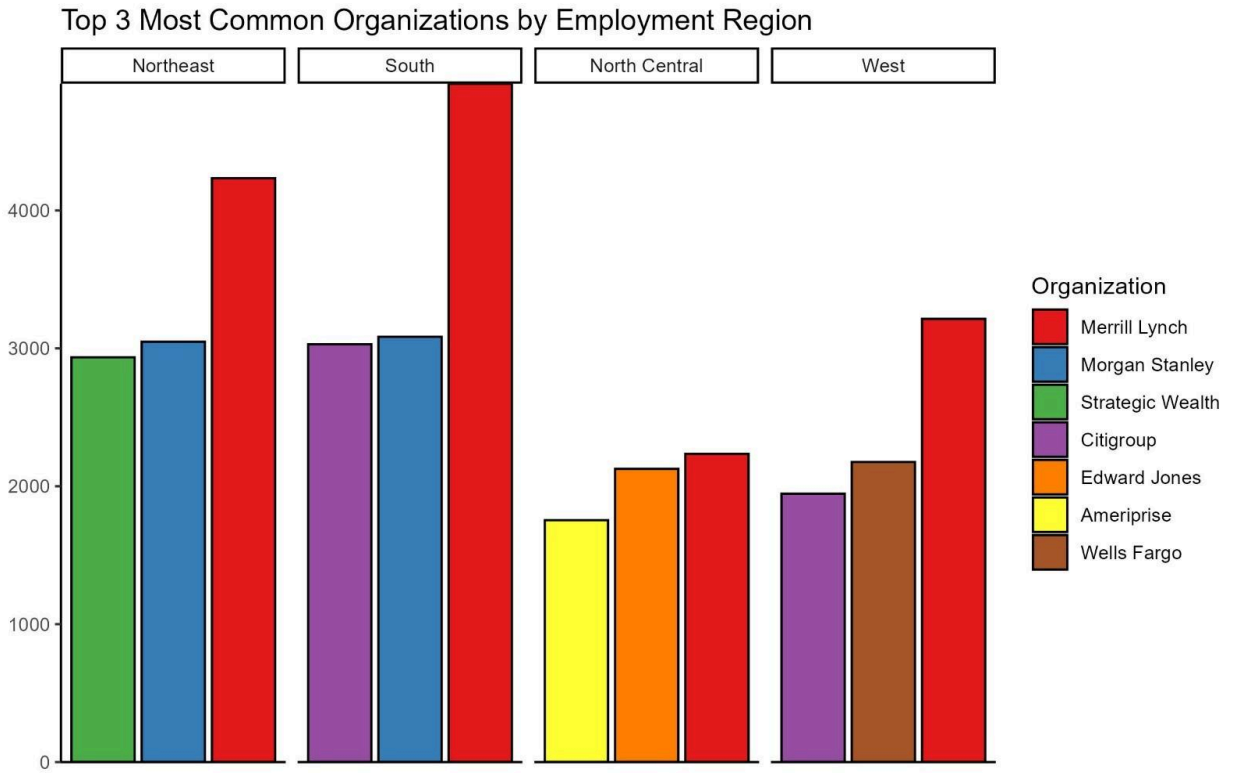
Description: Firms with the highest number of representatives.

Graph 2:



Description: Histogram and boxplot showing how long on average representatives remain registered.

Graph 3:



Description: Firms with the highest number of representatives by region

Graph 4:

Description: [Interactive bubble map] of firm activity across states.