Pair Assignment 3: Gathering, Cleaning, Merging and Exploring our Data

Course: Introduction to Collaborative Social Science Data Analysis

Malte Berneaud-Kötz & Jonas Markgraf

Hertie School of Governance

14 April 2016

1. Introduction[1]

2. Gathering the Data

2.1 Bank Board Data

We hand-collect a unique panel dataset on the composition of Boards of Directors in Bavaria's *Sparkassen*. This dataset includes detailed information on board member profiles which enables us to identify mayors on bank boards:

- name of board members;
- occupation of board members (identifier for mayors on board);
- position within board: normal board member, chairman, or vice chairman.

Annual information on Board of Directors is hand-collected from savings banks' annual reports available in PDF format on *Bundesanzeiger* for the years from 2006 to 2015; access to data prior 2006 is proprietary (Bureau van Djik's *Bankscope* database), which restricts our observation period. The dataset on boardroom composition constitutes the first comprehensive and systematic investigation of Bavarian savings banks' corporate governance as information on German public banks' boards has not been systematically collected yet.

[1] This paper is based on and a part of a research project by Guillermo Rosas (Washington University in St. Louis; grosas@wustl.edu) and Jonas Markgraf (Hertie School of Governance; markgraf@hertie-school.org).

2.2 Municipal Election Data

A database on mayoral elections in Bavaria is available from the state statistical office upon request. It contains data on direct municipal elections between 1948 and 2014. With this database we are covering 79 of the 416 German *Sparkassen* (19%) and 2,099 municipalities (19% of all municipalities in Germany). The data for mayoral elections contains the following variables:

- election date;
- name of election winner and (at least) the first loser;
- party affiliation of candidates;
- vote shares of candidates;
- dummy for competitive elections (at least two candidates);
- dummy for 'first-time mayor';
- number of eligible voters in voting district (size of municipality).

After obtaining the raw data, we cleaned the data set and subsequently created additional variables needed in our analysis. The individual steps taken are outline in the

3. Cleaning the Data

3.1 *Sparkassen* Data

The data we obtained on the *Sparkassen* had really long and unwieldy names, which we changed to make them more manageable. Moreover, we standardized them to follow use underscores to seperate words and use lower case.

The strings containing the municipality names and the names of mayor candidates contained unnecessary whitespace which we trimmed using the str_trim function from the stringr package.

Furthermore, we created an additional variable for top positions in *Sparkassen* boards. The raw data includes information on the position of the respective person (normal board member, vice chairman or chairman); since chairmen and vice chairmen alternative over years in Bavarian savings banks, there is no difference between chairmen and vice chairmen and the can be amalgamated into one category (top position).

Finally, we created three sub-dataframes by subsetting the initial dataframe in order to analyze different aspects of *Sparkassen* boards in greater detail. Hence, we created a subset containing unique board member profiles, another one with unique profiles of mayors on the board, and a dataframe containing only profiles of persons in top positions on the board.

3.2 Creating Additional Variables

In addition to the variables available in the data set already, we created variables (1) distinguishing 'first time mayors', (2) identifying competitive elections where there were more than one candidate, (3) the number of times a single mayor was elected, which allowed us to identify mayor's first re-election. This last point is important because our research is interested in the first re-election of mayors specifically.

3.3 Municipal Election Data

The municipal election data as provided by the Bavarian Statistical Service was provided as an Excel worksheet, which also meant that the columns where named in a way which was not computer-readable. As a result, we had to clean the names of the data set almost entirely. Aside from containing spaces, they also frequently contained line breaks and carriage returns.

Aside from the variables included in the data set, we extracted information on Phd titles from the names of the candidates, to include it as a control variable. Addtionally, we created a variable which indicates whether or not the election of the mayor is is . We want to use this variable in subsetting our data set later, as it identifies mayors which could have possibly leveraged their position in a *Sparkassen* board to secure re-election.

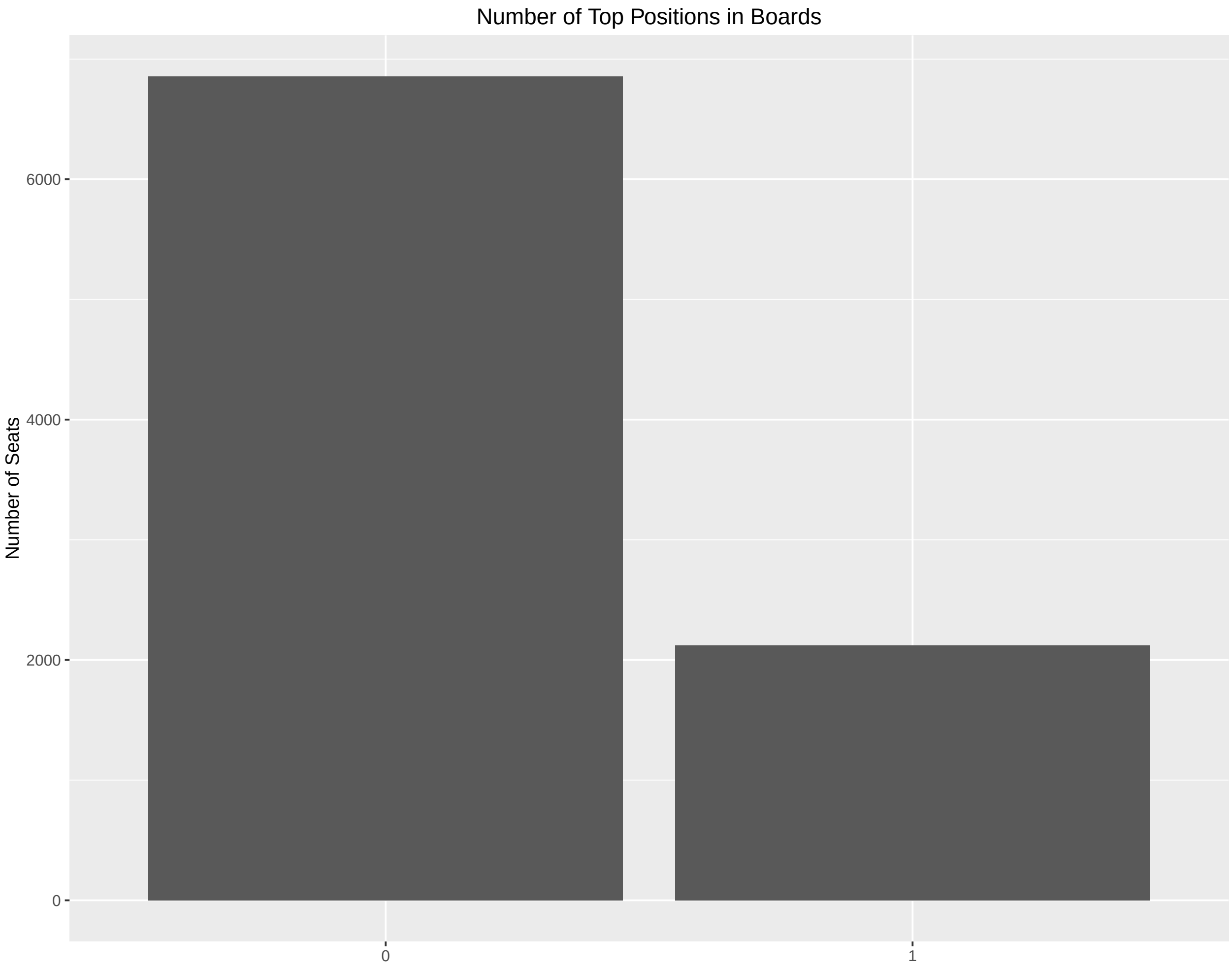
4. Merging the Data

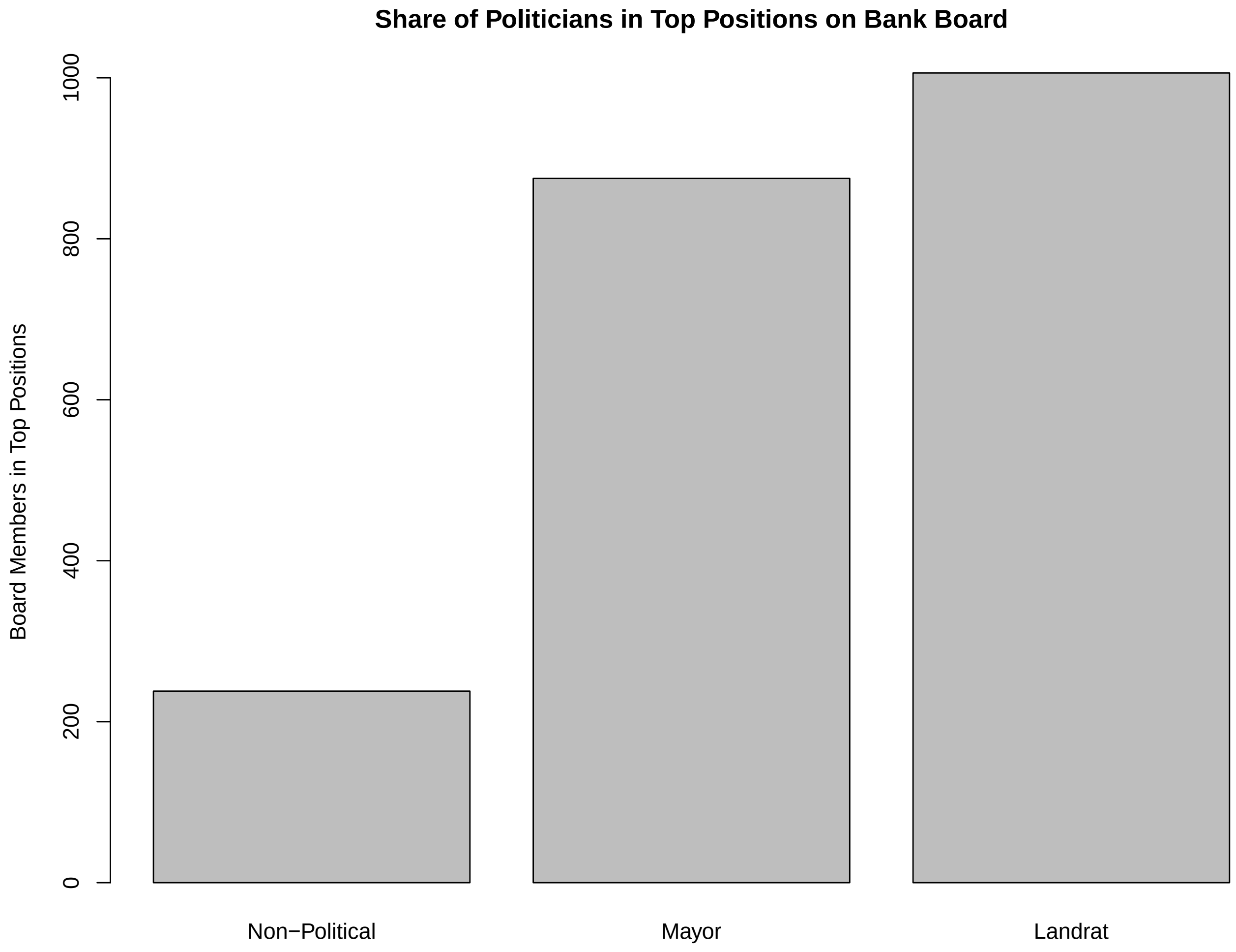
Both datasets were merged on the

5. Describing th

e Data

5.1. Sparkassen Dataset





5.2. Election Dataset

5.3. Merged Dataset

6. First Inferences