Assignment 2: Exploratory Data Analysis

Ron Cordell

W209 - Section 3

Florida 2000 Elections

Using county data from the 2000 Presidential elections in Florida, examine the raw data and formulate 3 hypotheses. For each of the 67 Florida counties, the data include the type of voting machine used, the number of columns in the presidential ballot, the undervote, the overvote, and the official certified votes for each of the twelve presidential candidates. Of particular interest are the Buchanan vote in Palm Beach county, and the overvote as a function of voting machine type and number of columns. Data is from the CMU Statistical Data Repository.

Hypotheses:

1. Votomatic voting machines have a higher overvote as a fraction of the total vote than other voting technologies.
2. The 2 column vote has a higher overvote as a fraction of the total vote than the 1 column vote.
3. The Buchanan vote in Palm Beach county is significantly larger than it should be.

Definitions:

- **Overvote**: occurs when a voter chooses more than the allowed number of candidates for an elected position. For example, if only one choice is allowed but the voter chose two. In the case of overvotes the vote is cancelled.
- **Undervote**: occurs when a voter chooses less than the required number of candidates for an elected position. It is within the voter's right to do so; for example, when choosing not to vote for any candidate for that position. The vote is still valid.

Hypothesis 1: Votomatic voting machines have a higher overvote as a fraction of the total vote than other technologies.

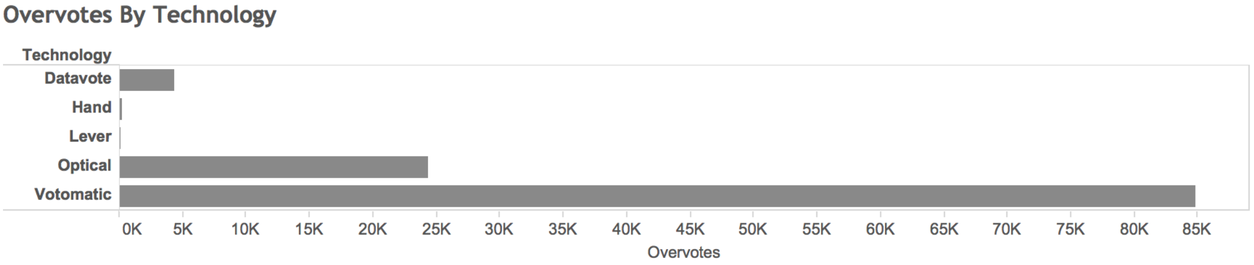

Figure 1: Overvotes By Technology

What's informative about this view: This view shows the number of overvotes by technology type for the Florida 2000 Presidential election. This view shows that the highest counts of overvotes occur with the Votomatic voting machines.

What could be improved about this view: This view uses only the raw counts of overvotes instead of the proportion of overvotes of the total vote count. Doing so would account for populations of counties and provide a better sense for *error rate* as opposed to *error count*.

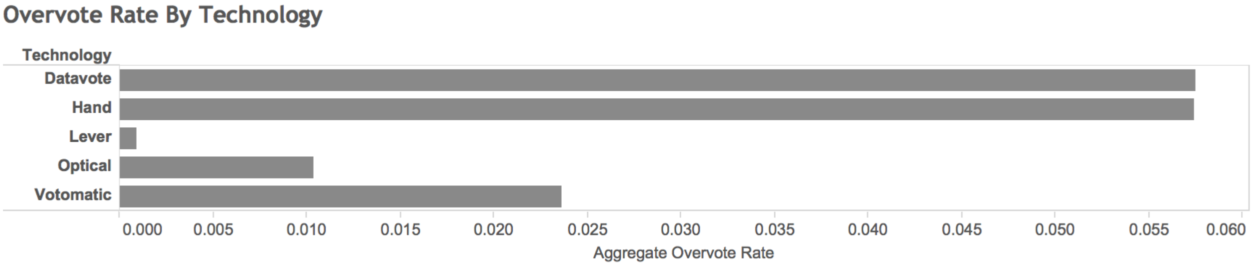

Figure 2: Overvote Rate By Technology and Column

What's informative about this view: This view shows the overvote rate by technology type for the Florida 2000 Presidential election. This view shows that the highest counts of overvotes occur with the Datavote voting machines and hand voting.

What could be improved about this view: We can get a better overall view of the column format with respect to overvote rate by combining this view with the format.

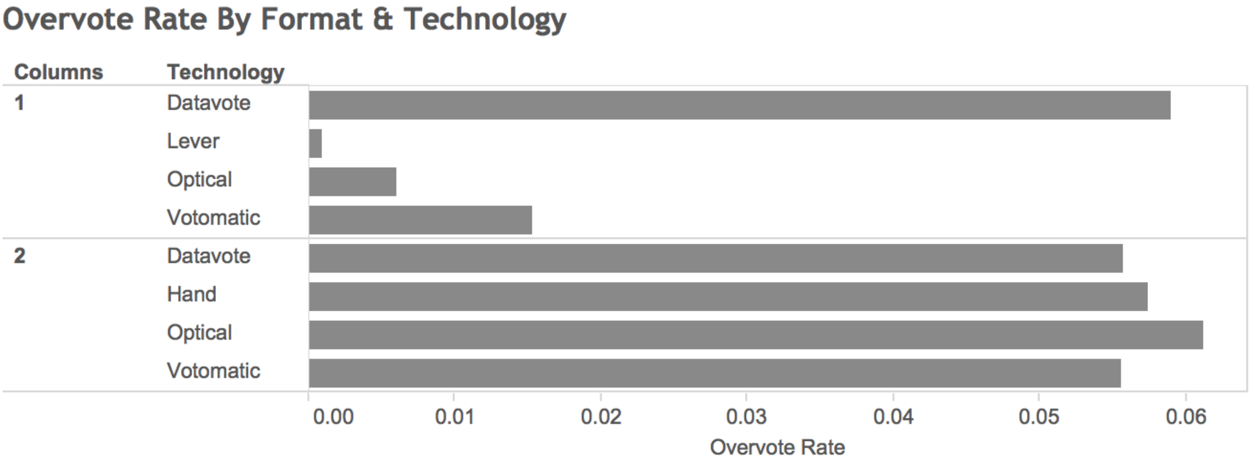

Figure 3: Overvote Rate By Technology and Column

What's informative about this view: This view shows the overvotes as a percentage of the total votes. Clearly the technology with the highest error rate regardless of number of columns is the Datavote. While all technologies show an increase in overvote rates from 1 column format to 2 column format, only the Datavote is high in both categories.

What could be improved about this view: There is more going on here than just voting technology. The 2 column format is more prone to overvotes than the single column format for most technology types except the Datavote. We can bring more into the picture here on what counties and what candidates were involved in where the Datavote technology was used.

Clearly the hypothesis that the Votomatic has a higher overvote is incorrect when comparing the overvotes as a function of the total vote but there is a likely correlation between the overvote and the technology as well.

Hypothesis 2: Two column format has a higher overvote rate than single column format

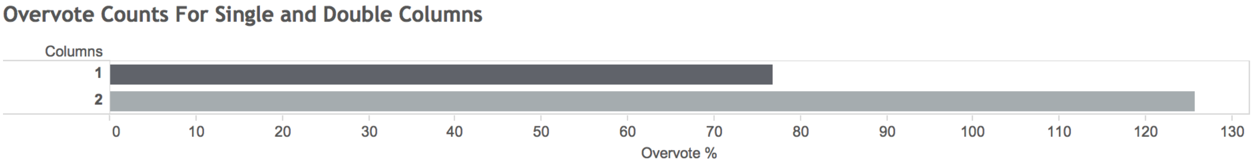

Figure 4: Overvote Rate for 1 and 2 Column Formats

This investigation follows from the findings that not only does the technology have a role to play in the overvotes but so does the format of the ballot. Cleary the overvote rate for a 2 column format is significantly higher than for 1 column format ballots.

We can get a little deeper to see how the overvote rate varies as a function of ballot columns, technology, and county.

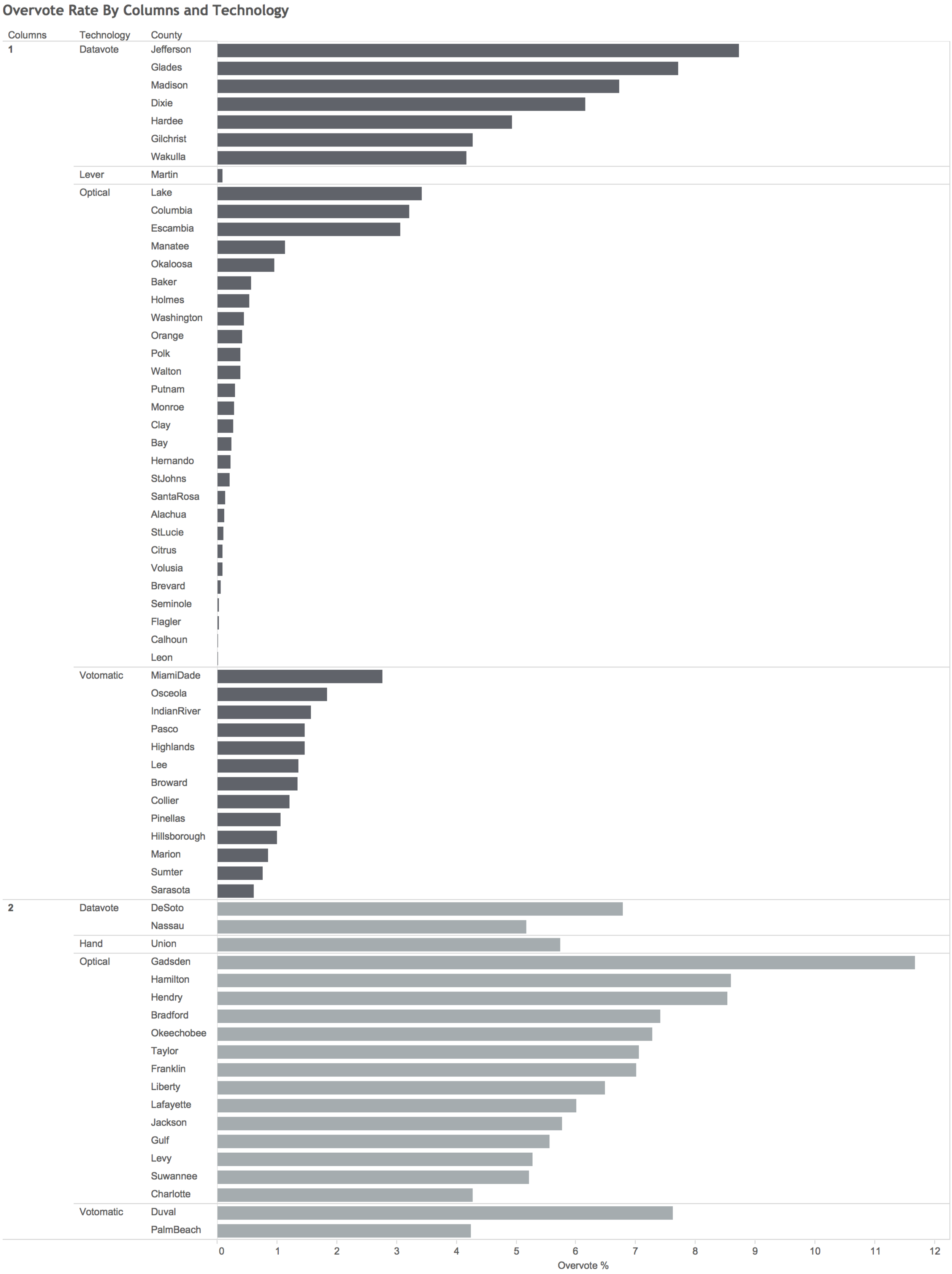

Figure 5: County Overvote Rates

Hypothesis 3: The Buchanan vote count in Palm Beach county is larger than it should be.

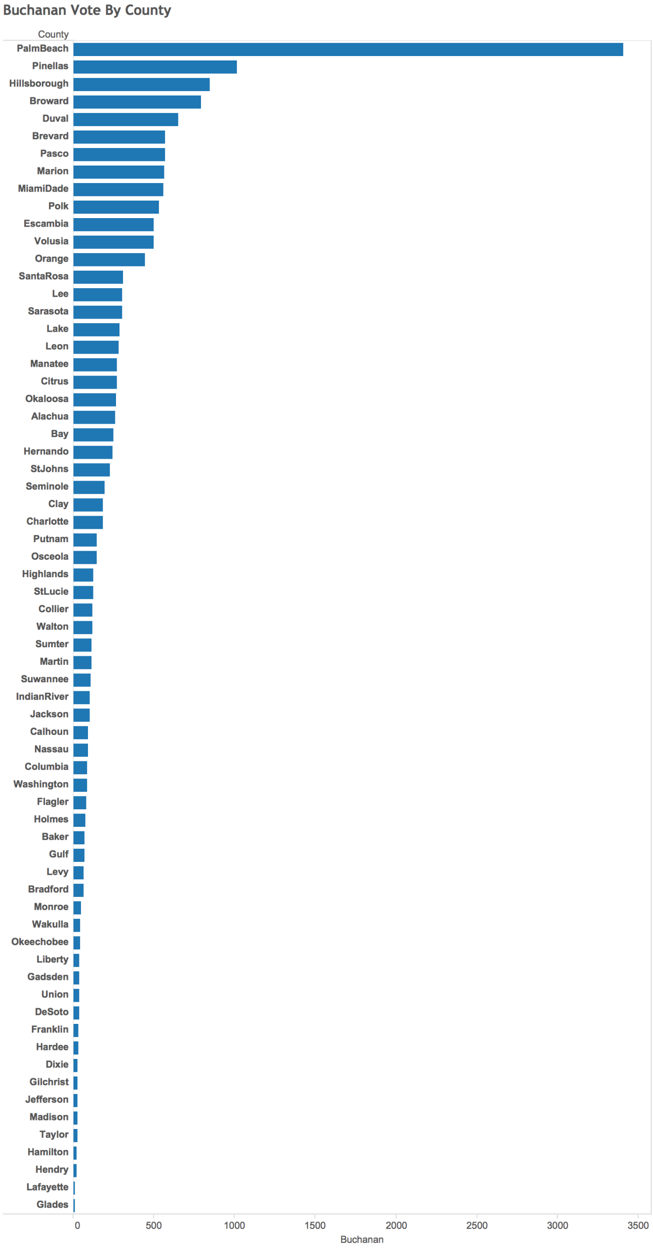

Figure 6: Buchanan Votes By County

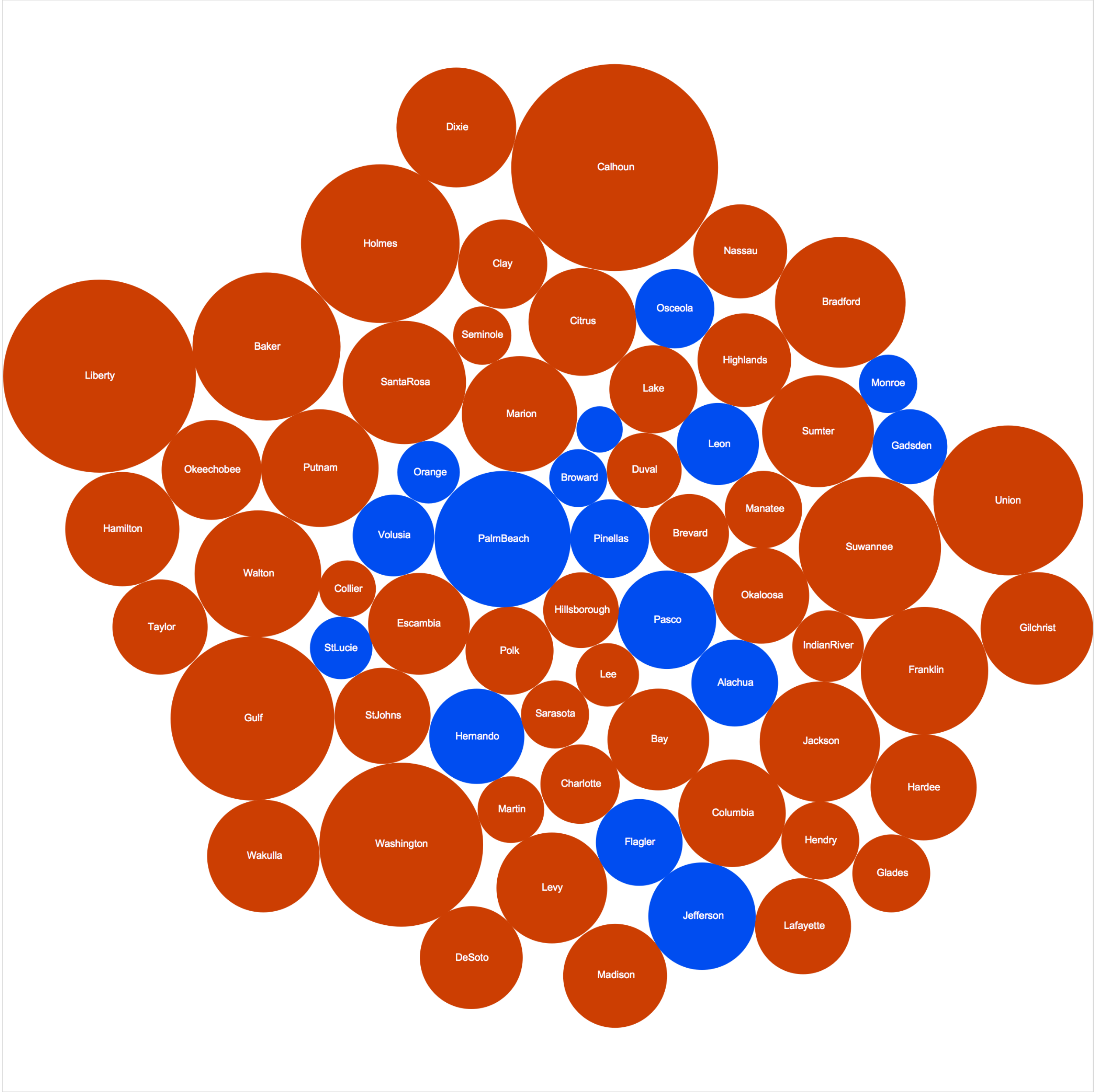

Figure 7: Buchanan Relative Votes

This graphic shows the relationships between the Republican, Democrat, and Buchanan votes by county. In this view the traditional red/blue colors indicated which party carried the county and the size of the circle is the proportion of Buchanan votes to total votes. Pat Buchanan was an extremely conservative candidate so if we also equate the Republican party with conservatism we would expect to see a corresponding increase in the Buchanan vote for those counties that voted Republican. We do see this trend in general in the graphic - the red circles tend to be larger than the blue ones - but this is not always the case. Palm Beach county is the highest Buchanan vote proportion and stands out from the other blue circles with the highest proportion of Buchanan votes. Palm Beach County is also the county at the center of the controversy of the "hanging chad" of the 2000 election. Palm Beach county used a 2 column format and Votomatic card punch machine. In the 2 column format Bush's name would be opposite Buchanan's name and have adjacent chads. One possible explanation is that the Buchanan votes were actually meant to be Bush votes.