

## 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.

#### 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.

#### Hypotheses:

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

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

First View - Image Here

**What's informative about this view:** This view shows the number of overvotes by columns and by technology type for the Florida 2000 Presidential election. This view shows that the most 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*.

Second View - Image Here

**What's informative about this view:** This view shows the overvotes as a percentage of the total votes. Clearly the highest error rates are the optical and Datavote technologies. In fact, the DataVote has the highest error rate across both 1 and 2 column formats.

**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.

Third View - Image Here

### **Hypothesis 3: What's up with the Buchanan vote in Palm Beach county?**

First View - Image Here

This shows how significantly different the vote is for Buchanan in Palm Beach county than for other counties in terms of raw vote counts.

Last View - Image Here

This shows that the Palm Beach votes for Buchanan are not really out of line as a percentage of total votes for the county as compared to other counties. However we can see when we bring in the binned overvote rate and the number of columns that