W209-3 Assignment 2 – Sanjay Dorairaj

In this assignment you will be selecting one of the provided datasets and exploring it with Tableau Desktop. Begin by inspecting the available data without visualizing the data, and write down three hypotheses. Next, investigate each of your hypotheses by visualizing relevant variables (including derived variables, if that helps) in multiple ways. Look for correlations, clusters, outliers, or any other patterns. See if you can find evidence for or against each hypothesis. As you explore, take screen captures of plots that show the development of your analysis. Try to find something unexpected in the data. For only one of the hypotheses, describe your exploration process using the template below, noting changes and refinements you made to the visualizations as you went along, as well as what worked or didn’t work during your exploration process. For the remaining two hypotheses, simply provide screen captures of the first and last visualizations you made.

Note: this assignment will be graded primarily on how thoroughly you explored the data set with various visualizations. Your first and last visualizations should be quite different. In the template, you should show a variety of plot types as well as refinement within individual plots. Use of Tableau is required. Your writing should also be clear.

# Dataset Selection

The dataset selected for the purpose of this analysis is the County data from the 2000 Presidential Election in Florida. The data is briefly described below:

[County data from the 2000 Presidential Election in Florida](https://www.dropbox.com/s/1arzn0nup8jua6t/fl2000clean.csv?dl=0)  (also see the [text file with column details](https://www.dropbox.com/s/5e51qg7ub8vu3po/fl2000.txt?dl=0))

“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. (Note: be sure you save the text file properly. Some browsers may add extra characters.) [Background on the Florida butterfly ballot](http://en.wikipedia.org/wiki/Ballot#Design)”

# Hypothesis Selection

Based on a preliminary examination of the data, the below three hypotheses were chosen for further exploration using data visualization methods.

1. The sum of undervoting and overvoting combined is insignificant compared to the total votes cast
2. Overvoting and undervoting is more likely to occur in the highly criticized Votomatic system
3. The Green Party led by Nader steered liberal voters away from the Democratic base.

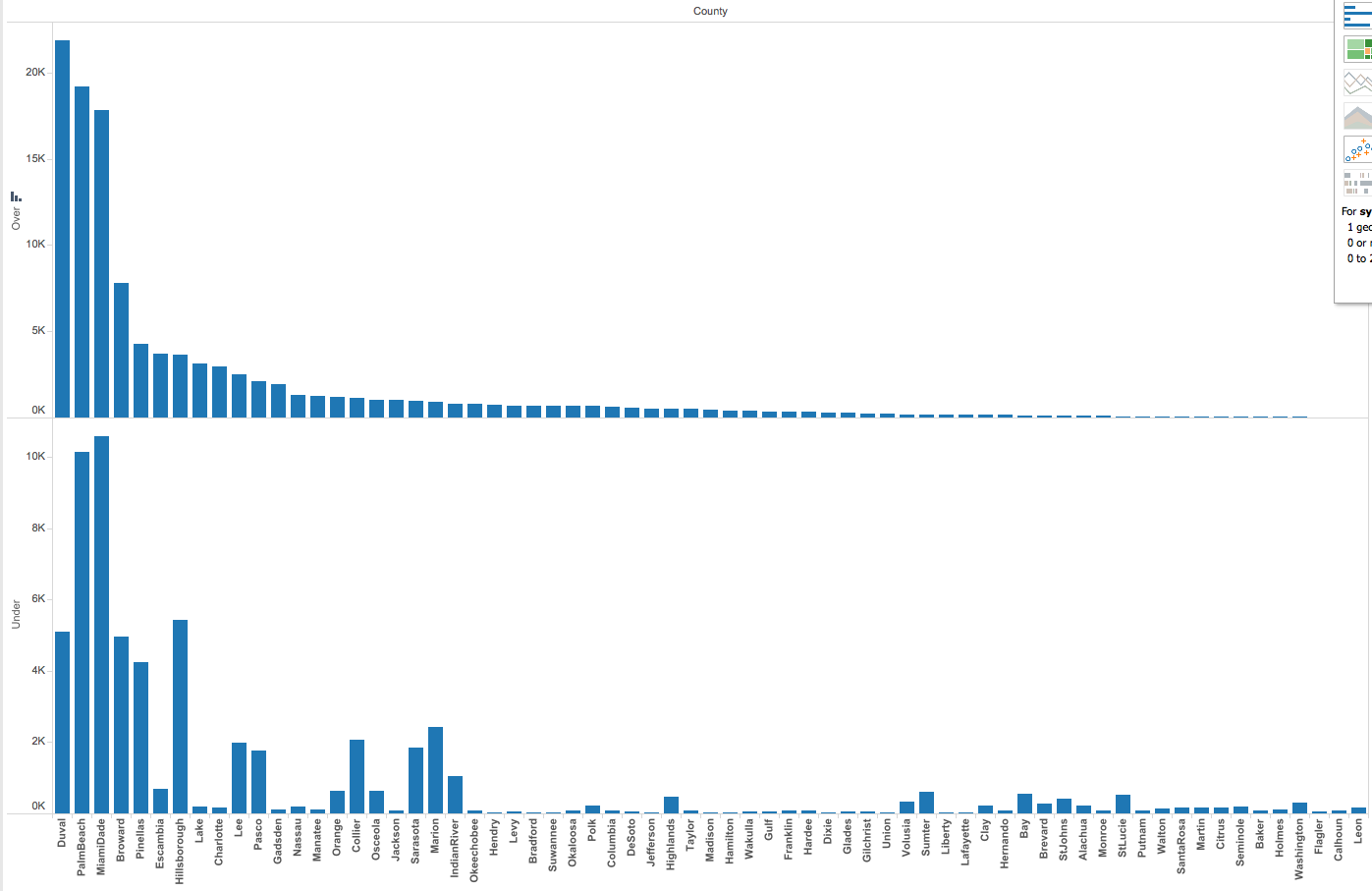
The following sections attempt to explore these hypotheses in further. **The third hypothesis is explored in greater detail per the suggested template.**

# Hypothesis 1 – The sum of undervoting and overvoting combined is insignificant compared to the total votes cast

The reason for choosing this hypothesis is because of an assumption that voting anomalies are few and that given the large number of people who cast votes, these anomalies are unlikely to be significant to the overall election results.

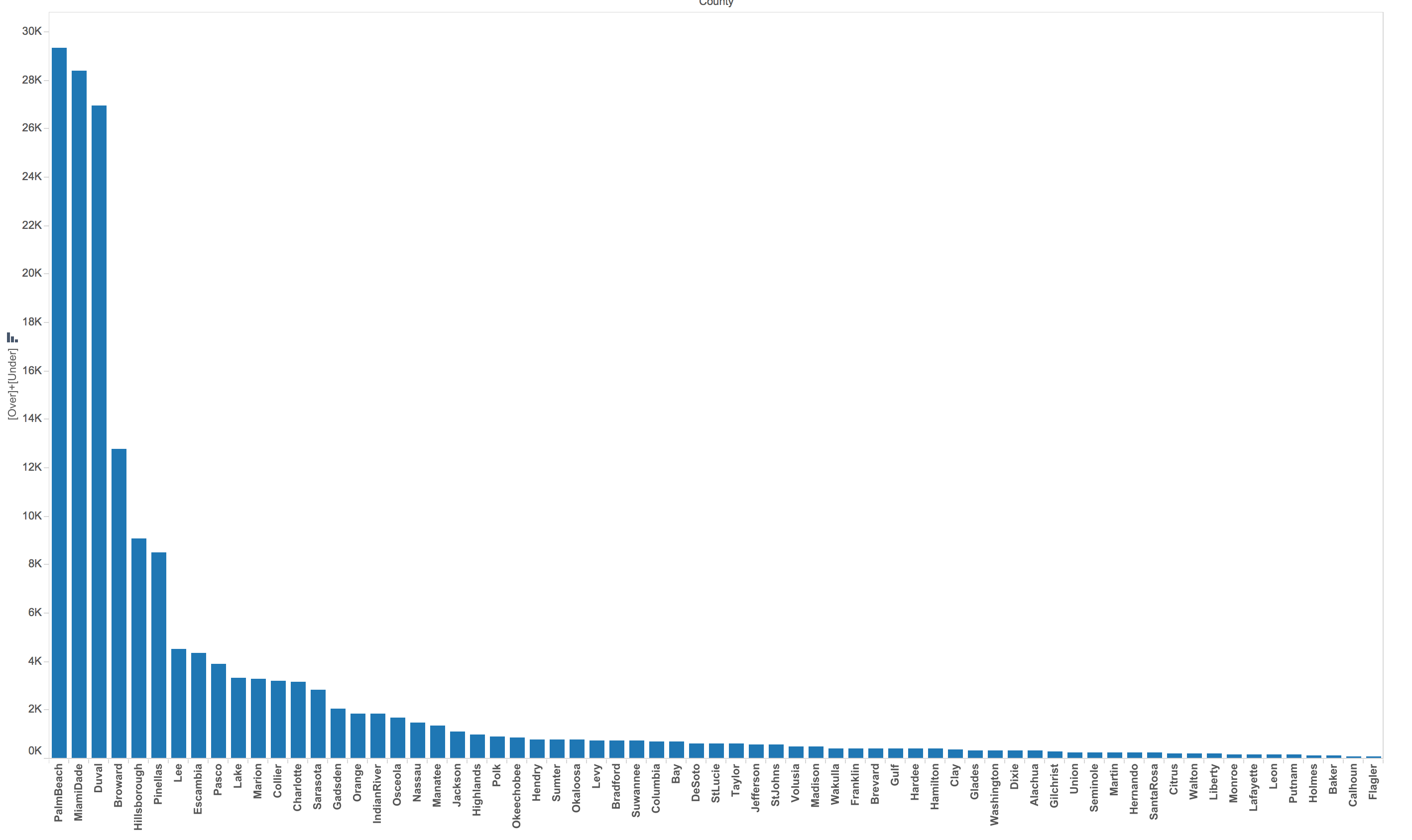
**Undervoting** refers to a case where fewer votes than available were cast. This may be a result of the user skipping some votes due to ambiguity or as a personal choice. **Overvoting** refers to cases where more votes were cast than allowable and this typically results from either a voting device malfunction or fraud.

The initial chart lines up the number of under and over votes in descending order.



An interesting observation here is the large number of under and over votes especially in counties such as Palm Beach and Miami Dade that generated the most amount of controversy in the 2000 elections. Another interesting observation is that counties with a large number of overvotes also appear to have a high number of undervotes, indicating a correlation between the overvoting and undervoting, which may be a result of similar voting conditions in each county.

The final slide depicts the sum of overvoting and undervoting together across counties.

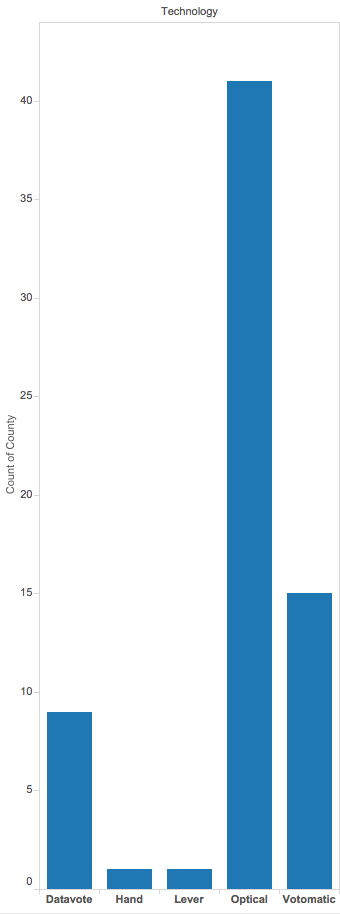


The high number of anomalous votes makes clear that the hypothesis that these sorts of anomalous votes were insignificant is incorrect. Certain counties especially Miami Dade show total anomalous votes of around 28,000 and equally close margins of victory.

# Hypothesis 2 – Overvoting and undervoting is more likely to occur in the highly criticized Votomatic system

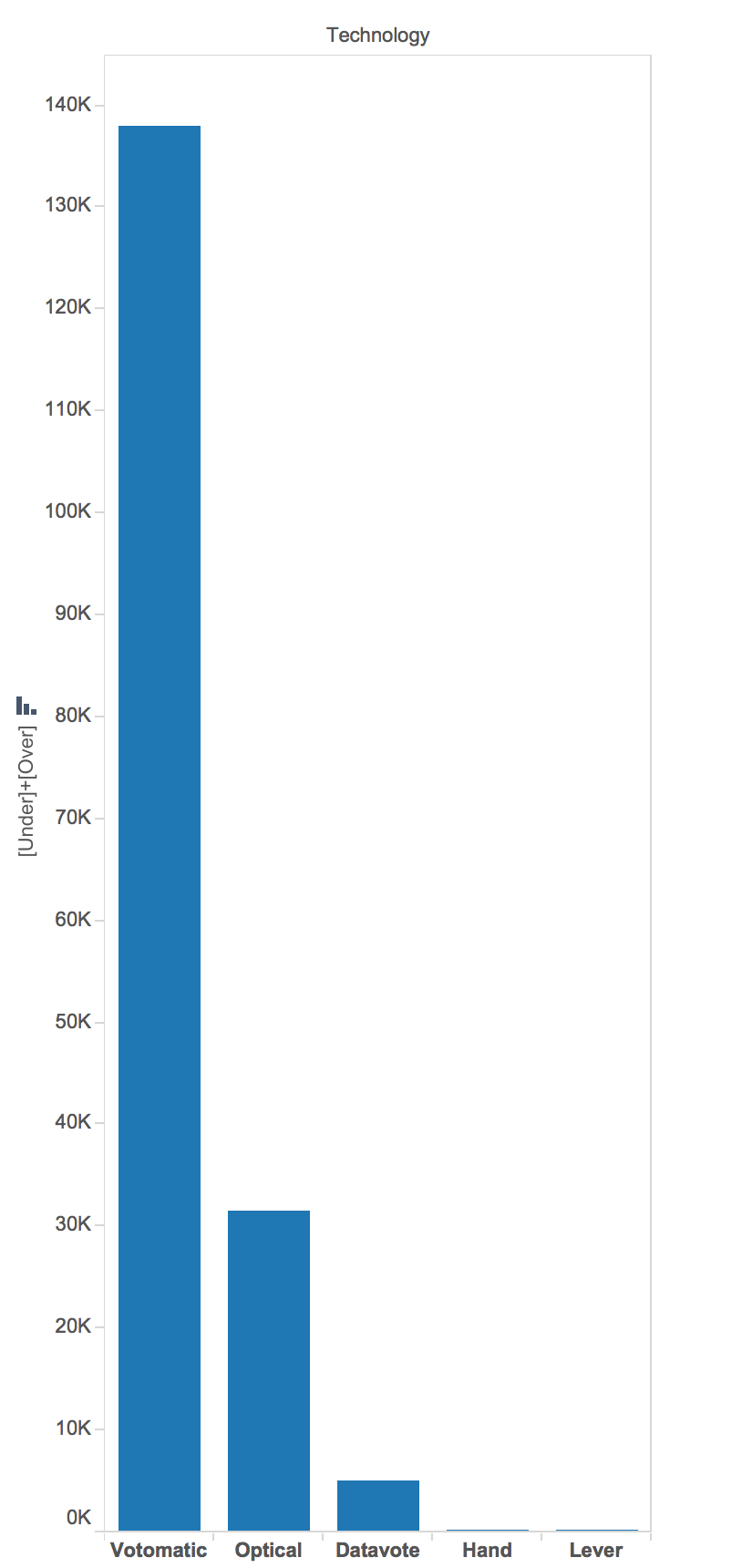
The Votomatic system was recently banned from elections after it was found that it leaded to a lot of confusion and consequently errors in the voting process. This was especially an issue in the 2000 elections. This hypothesis therefore claims that a majority of anomalies, represented as a sum of over and under votes, are most likely to take place with the Votomatic machine.

The initial chart looks at the distribution of the use of various voting systems in Florida.



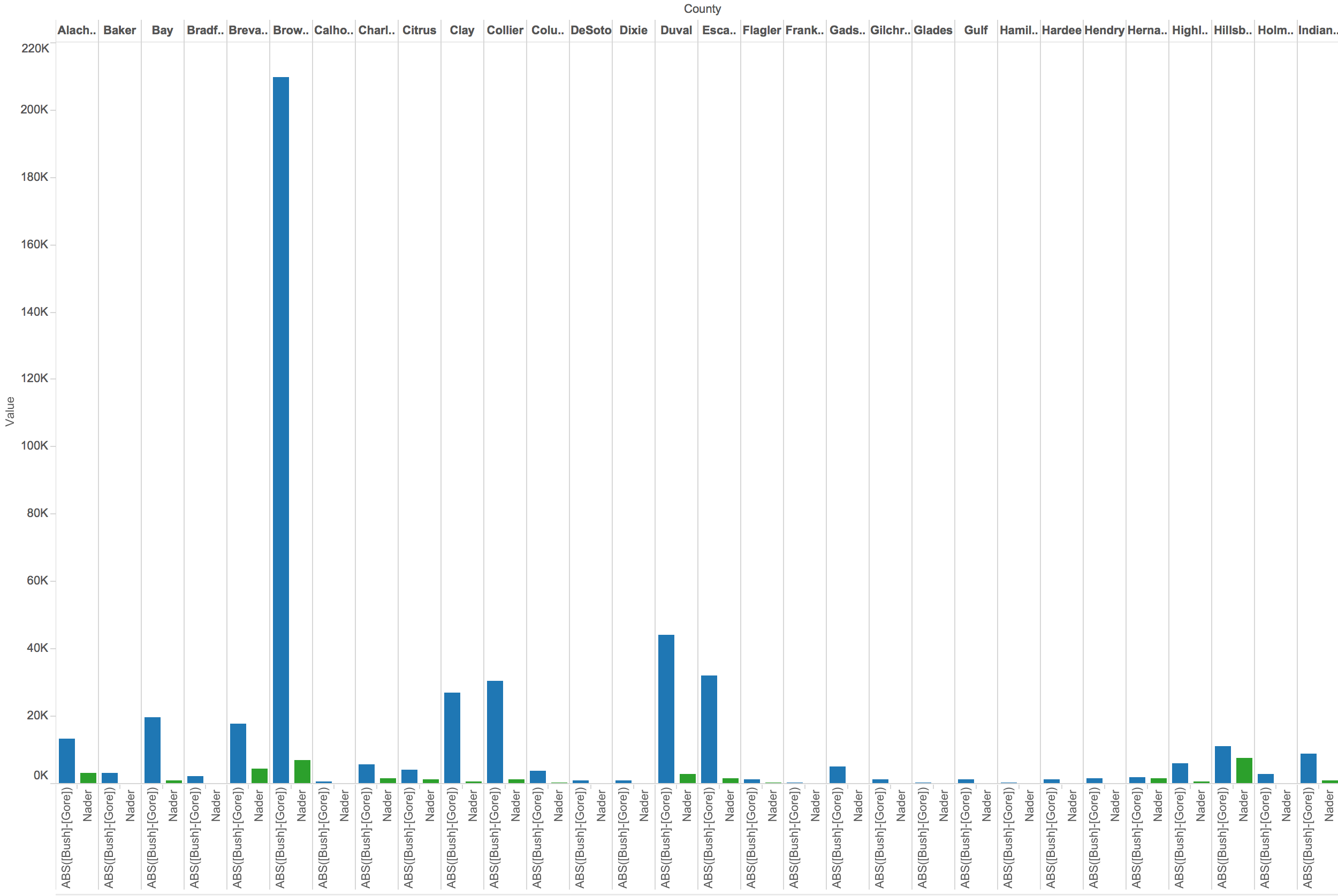
From the initial chart it is clear that the Votomatic system is the second most popular voting system although a distant second.

The final chart shows the distribution of total under and over votes across various voting machine types. It is clear from this chart that a large majority of anomalous votes took place when the Votomatic machine was used.



# Hypothesis 3 – The Green Party led by Nader steered liberal voters away from the Democratic base

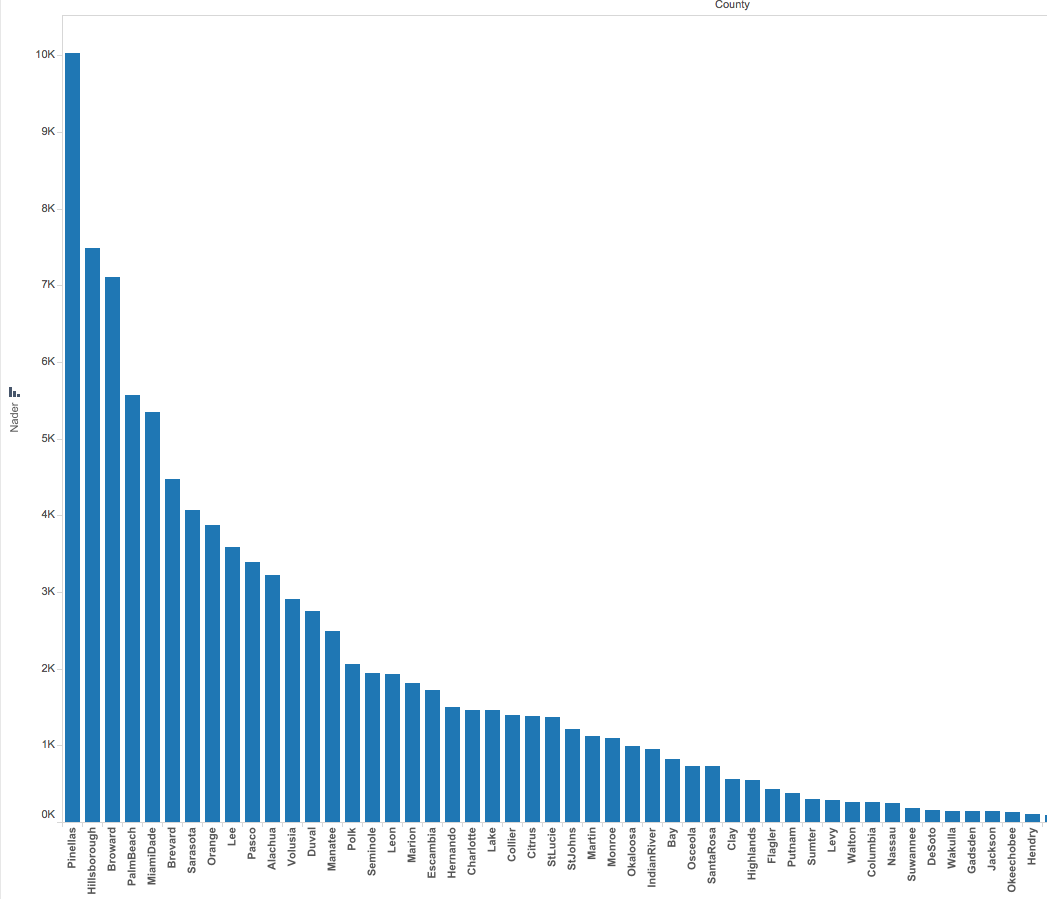
**Hypothesis**: The Green Party led by Ralph Nader steered liberal voters from the democratic base. Furthermore, the Green Party did sufficient damage to the Democratic candidate Al Gore as to cause the candidate to lose the elections. This is based off the assumption that were it not for the Green Party candidate, those votes would have gone to the democratic candidate.



**What’s informative about this view:** This view shows the Green Party votes (in Green) and the difference between Republican and Democratic votes (in Blue). The view indicates that in several cases, Green Party votes were negligible to the outcome of the elections.

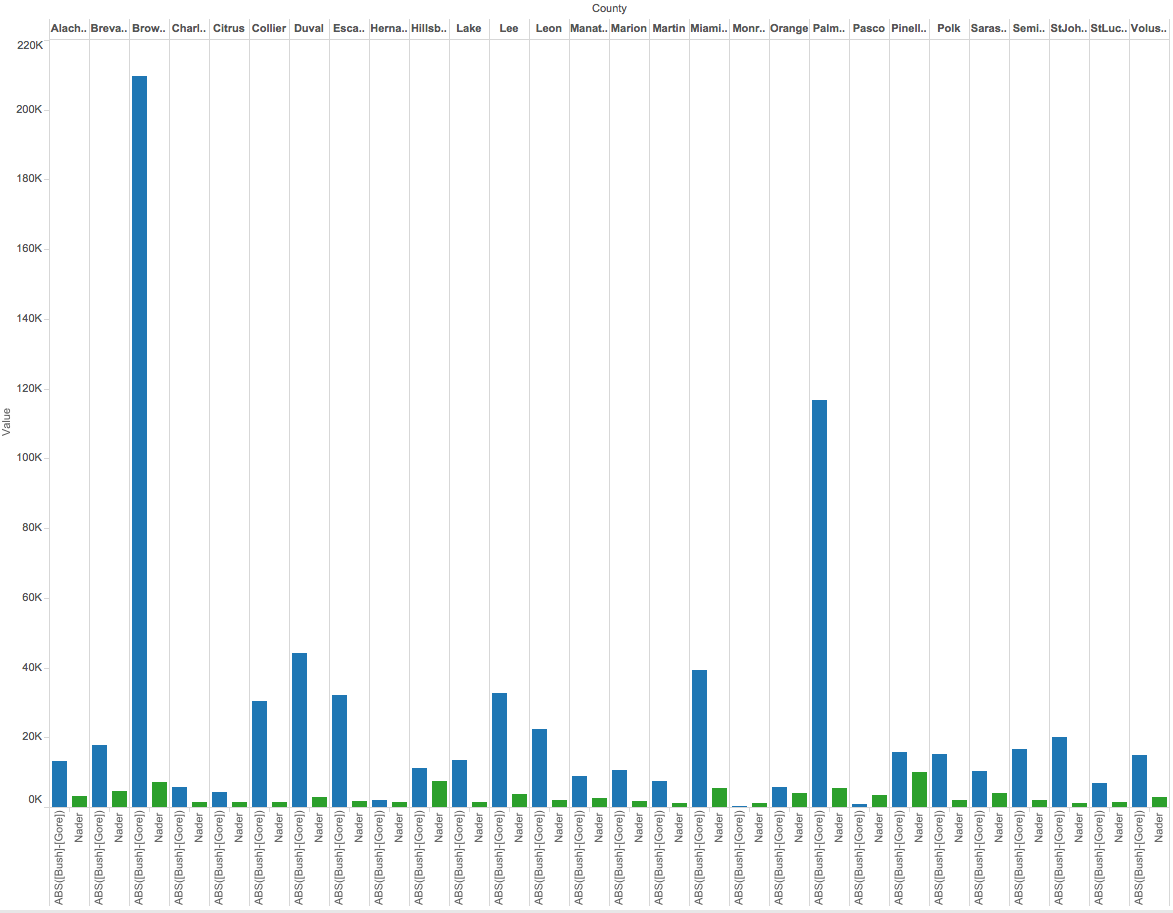
**What could be improved about this view:** The above view shows information on only a small set of counties and their results due to the restricted size of the visible viewing area. It would be useful to focus on those counties with smaller margins of victory and examine the influence of Green Party votes in those counties

**Distribution of Green Party votes**



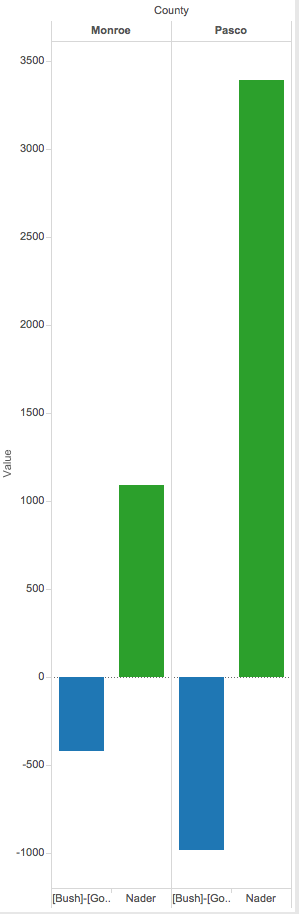
The above view shows that the largest number of votes secured in a county by the Green Party was 10000. We will therefore only examine counties that had 1000 and more Green Party votes and examine the margins of victory in those counties. This should provide us with a better picture of the influence of Green Party votes.

The final chart shows margins of victory and Green Party votes for just the selected counties with a higher number of **Green Party** votes.



A careful examination of the above view reveals that in most counties where the Green Party secured 1000 or more votes, the margins of victory were sufficiently high so as to ensure a Republican victory. Two counties however, **Monroe and Pascoe**, have sufficiently low margins of victory that it is possible that the Green Party could have made a difference to the outcome of the election.

In the case of the two counties Monroe and Pascoe identified above, both counties were already won by the Democratic party, indicating no difference in the outcome of the election.



**Conclusion:** The data appears to suggest that the Green Party did not have a significant affect on the outcome of the 2000 elections in Florida.