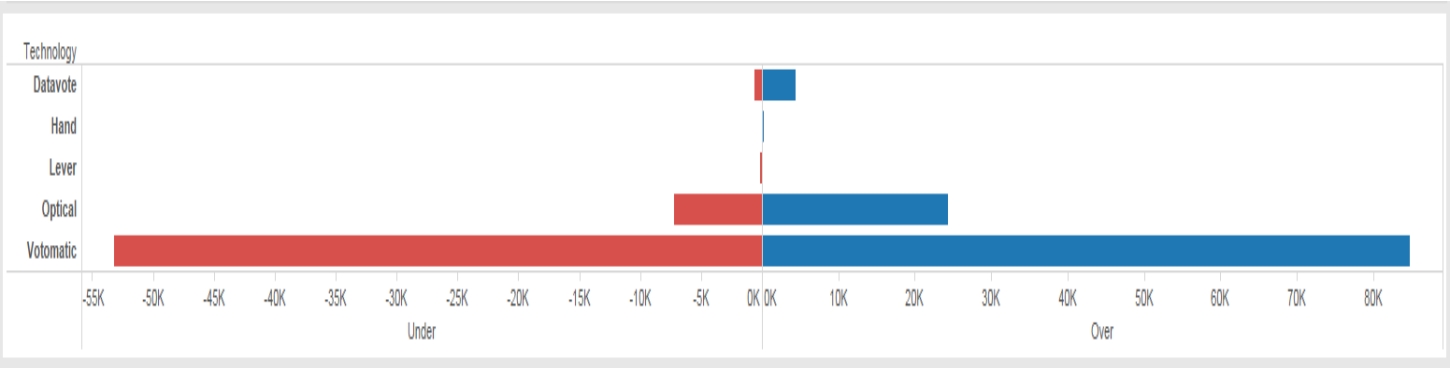
A2. Exploratory Data Analysis

Simon Macarthur

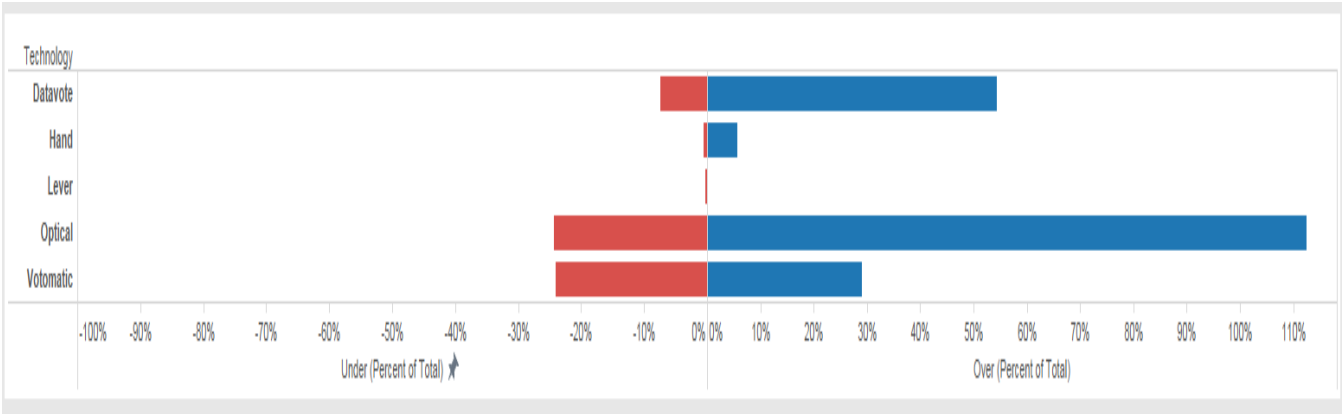
W209 - 2

Hypothesis 1: The technology used impacted upon the number of Undervotes and Overvotes.



What's informative about this view: This view shows number of undervotes and overvotes by Technology used. This view is helpful in gaining an overview of whether there is a difference in undervoting and overvoting by the type of technology used. It shows that there is a significant difference, with the 'votomatic' system having the majority of undervotes and overvotes.

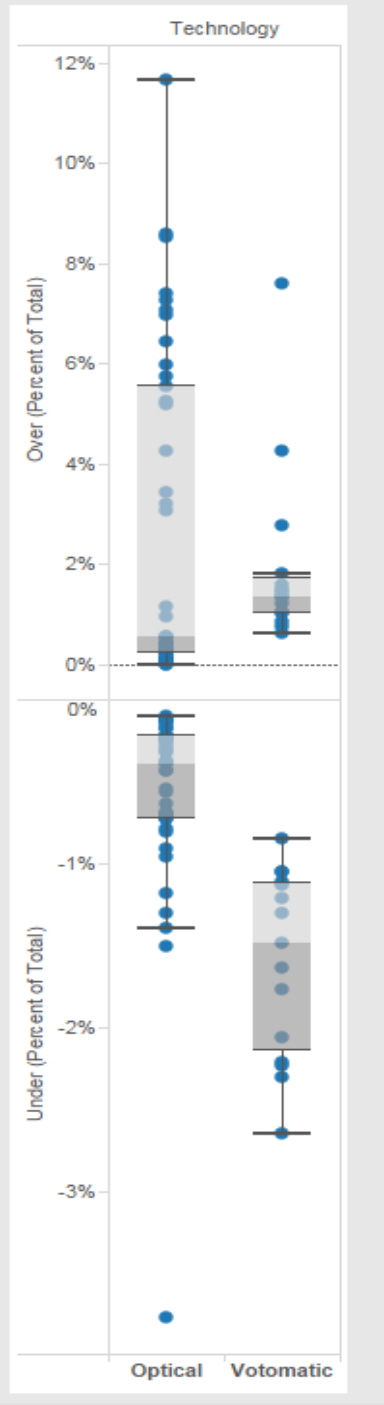
What could be improved about this view: Whilst this view tells you how many undervotes and overvotes occur, it does not provide insight into how many votes are cast using these technologies. Just because Votomatic is the highest, does not mean it had the highest number of issues per voter.



What's informative about this view: This view shows number of undervotes and overvotes by Technology used. However, this view shows the data as a percentage of total accepted votes.

This view is helpful in understanding whether the technology had a small or large effect on the votes discarded.

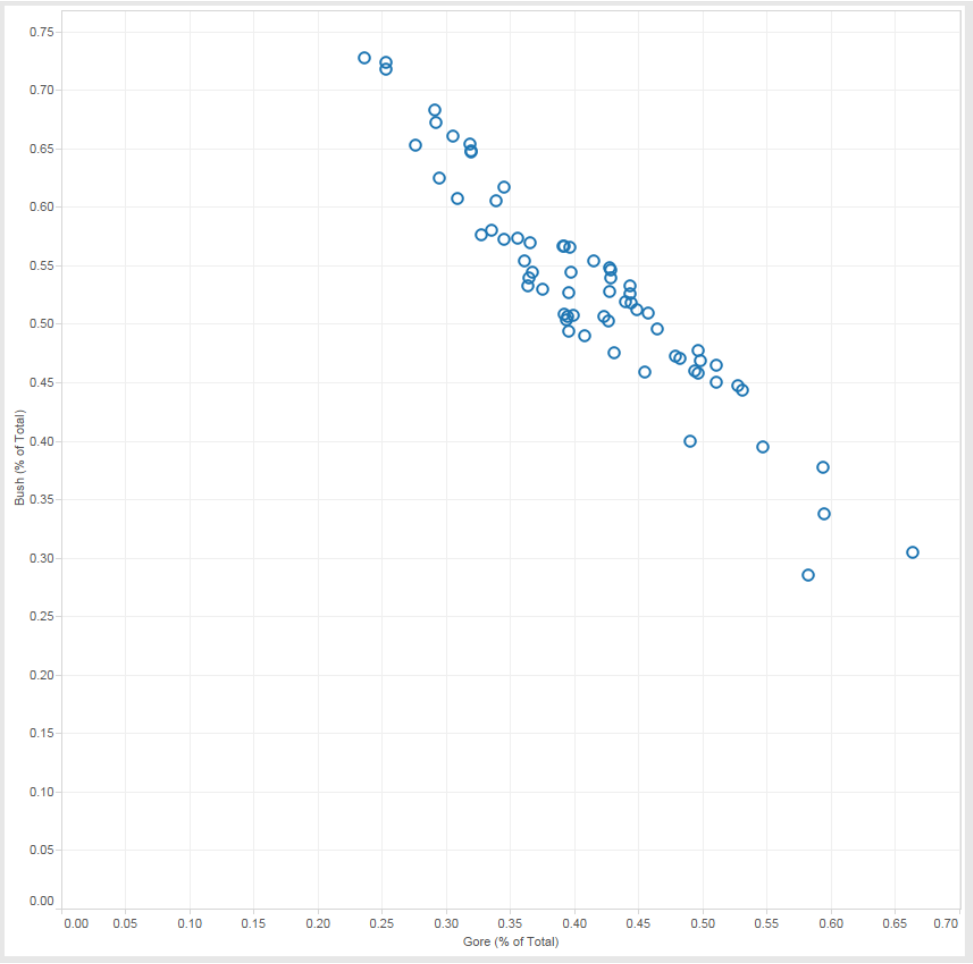
What could be improved about this view: It is difficult to understand the driving force behind the overvoting. Was it restricted to a certain county? Was there any trend in technologies between counties?



What's informative about this view: This view shows a box and whisker plot of only two technology types – Optical and Votomatic, by County. This view is helpful in gaining an understanding of if there were any specific technology issues in particular counties. It can be seen that Washington was an significant outlier in the Optical technology – being 4% under, double that of the nearest. With overvoting, it can be seen that the votomatic had three outliers, being Duval at 8% of the total and Palm Beach and MiamiDade being 4% and 3% respectively.

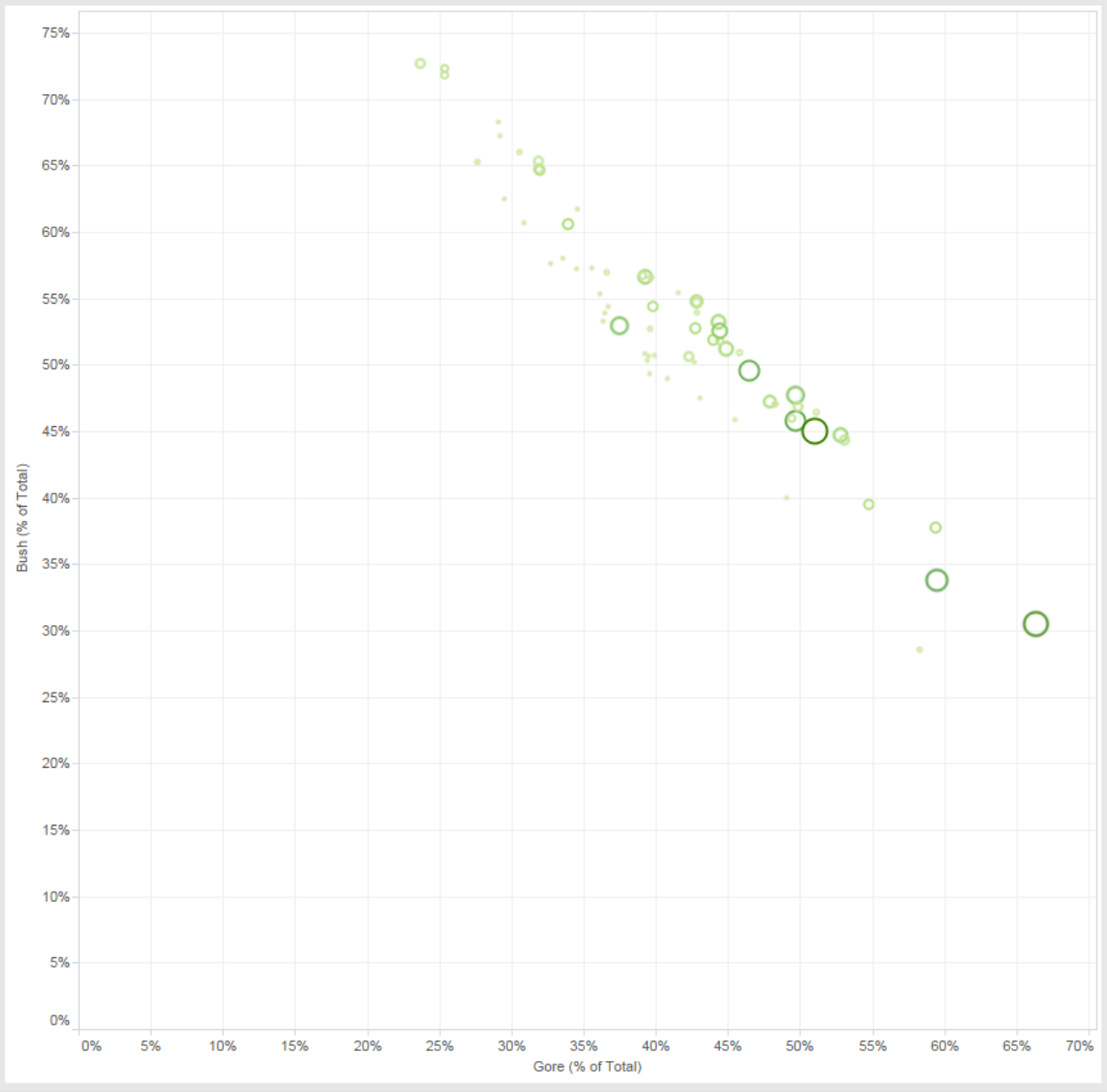
Conclusion These data views suggest that there is a linkage between the technology used and amounts of undervoting and overvoting. In particular, when looking at this data by county, it becomes apparent that the issue was not state wide, but there were some definite outliers within each of the technologies used.

Hypothesis 2: There is a relationship in voting between Bush and Gore. That is, for the majority if they did not vote for Bush, they voted Gore.



What's informative about this view: This view shows there is a definite relationship between Gore and Bush votes, showing that as one has a high amount, the other tends to have a low vote (as opposed to voting for another candidate). It highlights this really was a two-horse race.

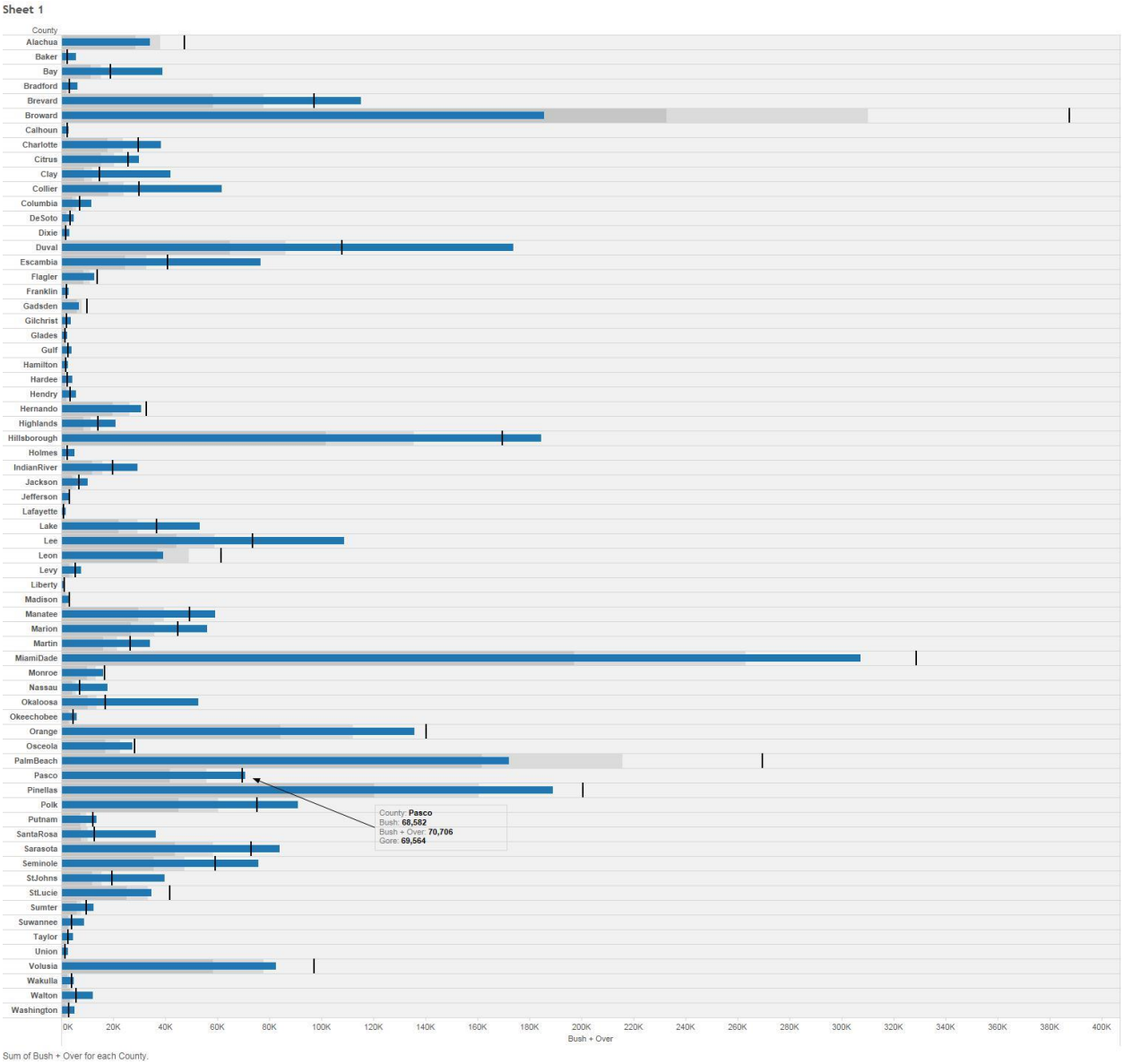
What could be improved about this view: Whilst this view shows the relationship between the two, it is hard to see how many votes were included in total for this county. This is important as 70% of 2,000 votes is very different from 70% of 200,000 votes.



What's informative about this view: This view now provides the color highlighting and respective size for the number of total votes in this county. This provides much more insight into the valuable high and low percentages for Gore and Bush respectively.

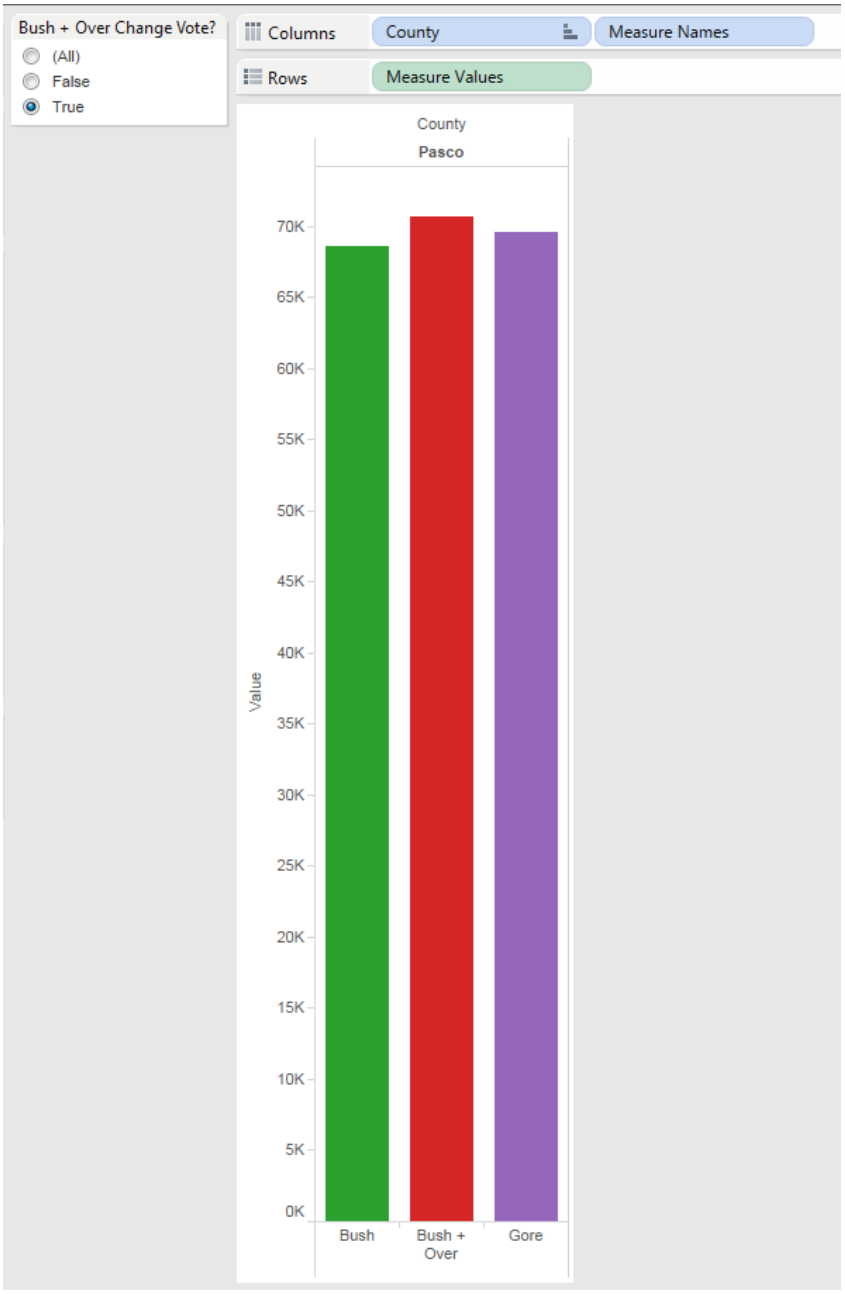
Conclusion The data highlights the fact that this was a two-horse race, and also provides insight into which are the more impactful counties included in this information.

Hypothesis 3: The number of over-votes would have changed the final decision in voting preferences for counties, if they had been lodged correctly.



What's informative about this view: This view shows the difference between the Bush vote + assumes all the over-votes were lodged correctly and voted for Bush compared to the Gore vote per county. It can be seen that in the county of Pasco this may have made a difference, as Gore won the vote, but if you include the over-votes to Bush it shows that it would have swung to Bush.

What could be improved about this view: Whilst this information can be obtained from this view, it is also hard to visually see this information in the view chosen. It, therefore makes it hard to determine which county, if any, would be impacted by this.



What's informative about this view: This view is much cleaner, it provides insight immediately that there is only one county in which the number of over-votes would change the result. It highlights what the value of the Bush vote was initially, what the value is including the over-votes, and what the value of the Gore vote was. It also makes use of filters to easily identify those counties which could be impacted versus which were not.

Conclusion Whilst one county was identified in this process, the overall data does not support the hypothesis. The change in votes in this one county would not have been enough to change the overall vote for the state.