2000\_Election

Santonastaso

2022-07-31

# Who Won the 2000 Presidential Election?

The 2000 presidential election was decided on a few hundred votes. Ambiguous votes were recounted. Specifically undercounts (where the ballot was ambiguous and seemed no presidential vote was cast) and overcounts (where the ballot was ambiguous and seemed that more than one presidential vote was cast). There is an argument that Buchanan received more votes than expected in Palm Beach County (indeed, it’s an outlier) due to graphic designe layout of the printed ballot.

This uses regression analysis to predict whether this is true.

The data and story for them comes from : Winston, Wayne. *Danalytics Stories: Using Data to make Good Things Happen.*  Indianapolis, John Wiley & Sons, Inc., 2021.

#Set up our libraries  
library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ stringr 1.4.0  
## ✔ tidyr 1.2.0 ✔ forcats 0.5.1  
## ✔ readr 2.1.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

#Establish path and import the spreadsheet  
setwd('C:/Users/tsant/Documents/Data Science/Data\_Stories/Chapter 03')  
#I wrap it in 'as.data.frame' otherwise it comes out as a tibble which isn't very friendly to manipulation  
df <- as.data.frame(read\_excel("PalmBeachRegression.xlsx", sheet = "All Counties"))

## New names:  
## • `` -> `...4`  
## • `` -> `...5`  
## • `` -> `...6`  
## • `` -> `...7`  
## • `` -> `...8`  
## • `` -> `...9`

#Explore the data set. Make sure I got what I think I got:  
head(df)

## county Perot96 Buchanan00 ...4 ...5 ...6 ...7 ...8 ...9  
## 1 Alachua 8072 263 NA NA NA <NA> NA <NA>  
## 2 Baker 667 73 NA NA NA <NA> NA <NA>  
## 3 Bay 5922 248 NA NA NA <NA> NA <NA>  
## 4 Bradford 819 65 NA NA NA <NA> NA <NA>  
## 5 Brevard 25249 570 NA NA NA <NA> NA <NA>  
## 6 Broward 38964 788 NA NA NA <NA> NA <NA>

dim(df) #it looks like we have to clean the data up. There are notes in the Excel sheet past column 3

## [1] 67 9

str(df)

## 'data.frame': 67 obs. of 9 variables:  
## $ county : chr "Alachua" "Baker" "Bay" "Bradford" ...  
## $ Perot96 : num 8072 667 5922 819 25249 ...  
## $ Buchanan00: num 263 73 248 65 570 788 90 182 270 186 ...  
## $ ...4 : logi NA NA NA NA NA NA ...  
## $ ...5 : logi NA NA NA NA NA NA ...  
## $ ...6 : num NA NA NA NA NA NA NA NA NA NA ...  
## $ ...7 : chr NA NA NA NA ...  
## $ ...8 : logi NA NA NA NA NA NA ...  
## $ ...9 : chr NA NA NA NA ...

#This is a dplyr command  
df1 <- select(df, county, Perot96, Buchanan00)

#again, is this what I think it is?  
dim(df1)

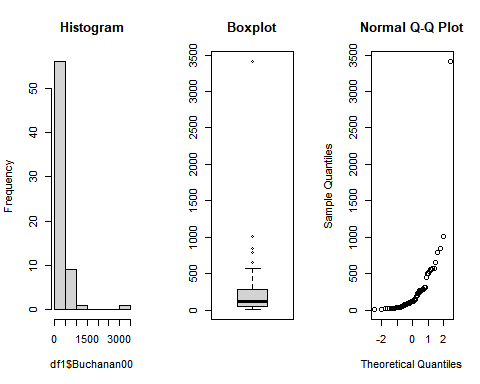
## [1] 67 3

str(df1)

## 'data.frame': 67 obs. of 3 variables:  
## $ county : chr "Alachua" "Baker" "Bay" "Bradford" ...  
## $ Perot96 : num 8072 667 5922 819 25249 ...  
## $ Buchanan00: num 263 73 248 65 570 788 90 182 270 186 ...

We’re going to do a regression analysis which uses x as a predictor of y. X is the independent variable and y is the dependent variabled. In this case, we’ll make a scatter plot of Buchanon on the y axis and Perot on the x axis. This suggests that the amount of votes Perot received in 1996 is predicting the amount of votes Buchanon received in 2000. *This relationship assumes that voters who voted for the third party candidate in 1996 will have proportionally voted for the third party candidate in 2000.*

#Let's look for outliers.  
#Constructing a figure of: Histogram, Boxplot and Q-Q plot  
par(mfrow = c(1, 3)) #we're telling R that we'll have 3 figures in a single row  
hist(df1$Buchanan00, main = "Histogram")  
boxplot(df1$Buchanan00, main = "Boxplot")  
qqnorm(df1$Buchanan00, main = "Normal Q-Q Plot")

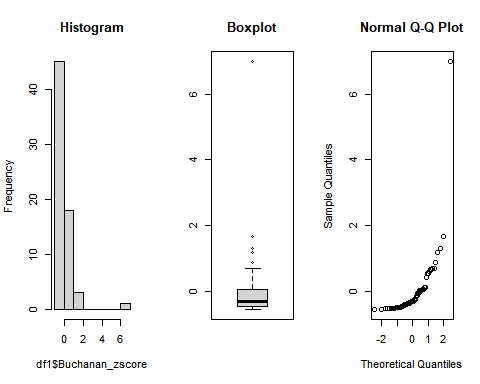
 Now to calculate the Z-scores. How far does each point lie from the mean?

meanB <- mean(df1$Buchanan00)  
meanP <- mean(df1$Perot96)  
sdB <- sd(df1$Buchanan00)  
sdP <- sd(df1$Perot96)  
df1 <- df1 %>% mutate(Buchanan\_zscore = (df1$Buchanan00 - meanB)/sdB)  
df1 <- df1 %>% mutate(Perot\_zscore = (df1$Perot96 - meanP)/sdP)

df1

## county Perot96 Buchanan00 Buchanan\_zscore Perot\_zscore  
## 1 Alachua 8072 263 0.005175001 0.094892049  
## 2 Baker 667 73 -0.417118340 -0.730365045  
## 3 Bay 5922 248 -0.028163947 -0.144716695  
## 4 Bradford 819 65 -0.434899112 -0.713425264  
## 5 Brevard 25249 570 0.687512136 2.009198748  
## 6 Broward 38964 788 1.172038179 3.537679645  
## 7 Calhoun 630 90 -0.379334199 -0.734488544  
## 8 Charlotte 7783 182 -0.174855318 0.062684176  
## 9 Citrus 7244 270 0.020733177 0.002614821  
## 10 Clay 3281 186 -0.165964932 -0.439045390  
## 11 Collier 6320 122 -0.308211110 -0.100361216  
## 12 Columbia 1970 89 -0.381556795 -0.585151001  
## 13 Desoto 965 36 -0.499354411 -0.697154159  
## 14 Dixie 652 29 -0.514912587 -0.732036734  
## 15 Duval 13844 652 0.869765051 0.738157943  
## 16 Escambia 8587 502 0.536375572 0.152286702  
## 17 Flagler 2185 83 -0.394892375 -0.561190127  
## 18 Franklin 878 33 -0.506022201 -0.706849954  
## 19 Gadsden 938 38 -0.494909218 -0.700163199  
## 20 Gilchrist 841 29 -0.514912587 -0.710973454  
## 21 Glades 521 9 -0.559364518 -0.746636151  
## 22 Gulf 1054 71 -0.421563533 -0.687235471  
## 23 Hamilton 406 23 -0.528248166 -0.759452432  
## 24 Hardee 851 30 -0.512689991 -0.709858994  
## 25 Hendry 1135 22 -0.530470763 -0.678208351  
## 26 Hernando 7272 242 -0.041499526 0.005735307  
## 27 Highlands 3739 127 -0.297098127 -0.388003155  
## 28 Hillsborough 25154 847 1.303171374 1.998611385  
## 29 Holmes 1208 76 -0.410450550 -0.670072798  
## 30 IndianRiver 4635 105 -0.345995251 -0.288147604  
## 31 Jackson 1602 102 -0.352663040 -0.626163103  
## 32 Jefferson 393 29 -0.514912587 -0.760901229  
## 33 Lafayette 316 10 -0.557141921 -0.769482566  
## 34 Lake 8813 289 0.062962511 0.177473481  
## 35 Lee 18389 305 0.098524055 1.244679684  
## 36 Leon 6672 282 0.047404335 -0.061132250  
## 37 Levy 1774 67 -0.430453919 -0.606994403  
## 38 Liberty 376 39 -0.492686622 -0.762795810  
## 39 Madison 578 29 -0.514912587 -0.740283733  
## 40 Manatee 10360 271 0.022955773 0.349880331  
## 41 Marion 11340 563 0.671953960 0.459097340  
## 42 Martin 5005 112 -0.330437075 -0.246912611  
## 43 Miami-Dade 24722 560 0.665286170 1.950466744  
## 44 Monroe 4817 47 -0.474905850 -0.267864445  
## 45 Nassau 1657 90 -0.379334199 -0.620033577  
## 46 Okaloosa 5432 267 0.014065387 -0.199325200  
## 47 Okeechobee 1666 43 -0.483796236 -0.619030563  
## 48 Orange 18191 446 0.411910166 1.222613391  
## 49 Osceola 6091 145 -0.257091390 -0.125882334  
## 50 PalmBeach 30739 3407 6.993018491 2.621036890  
## 51 Pasco 18011 570 0.687512136 1.202553124  
## 52 Pinellas 36990 1013 1.672122398 3.317685384  
## 53 Polk 14991 532 0.603053467 0.865986422  
## 54 Putnam 3272 148 -0.250423600 -0.440048404  
## 55 SantaRosa 4957 311 0.111859634 -0.252262016  
## 56 Sarasota 14939 305 0.098524055 0.860191234  
## 57 Seminole 9357 194 -0.148184160 0.238100066  
## 58 St.Johns 4205 229 -0.070393281 -0.336069353  
## 59 St.Lucie 8482 124 -0.303765917 0.140584879  
## 60 Sumter 2375 114 -0.325991882 -0.540015401  
## 61 Suwannee 1874 108 -0.339327461 -0.595849810  
## 62 Taylor 1140 27 -0.519357780 -0.677651121  
## 63 Union 425 37 -0.497131815 -0.757334960  
## 64 Volusia 17319 496 0.523039992 1.125432542  
## 65 Wakulla 1091 46 -0.477128446 -0.683111972  
## 66 Walton 2342 120 -0.312656303 -0.543693116  
## 67 Washington 1287 88 -0.383779392 -0.661268570

par(mfrow = c(1, 3)) #we're telling R that we'll have 3 figures in a single row  
hist(df1$Buchanan\_zscore, main = "Histogram")  
boxplot(df1$Buchanan\_zscore, main = "Boxplot")  
qqnorm(df1$Buchanan\_zscore, main = "Normal Q-Q Plot")



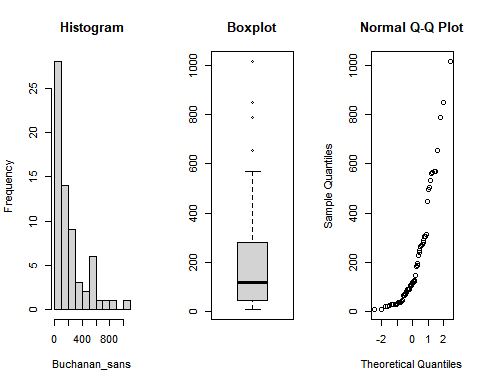
Ok, so the outlier is clear in all three figures.

Let’s now calculate the mean and standard deviation

mean = mean(df1$Buchanan00)  
std = sd(df1$Buchanan00)  
  
#find threshold values for outliers  
tmin = mean-(3\*std)  
tmax = mean+(3\*std)  
  
#find outlier  
df1$Buchanan00[which(df1$Buchanan00 < tmin | df1$Buchanan00 > tmax)]

## [1] 3407

#remove outlier  
Buchanan\_sans<-df1$Buchanan00[which(df1$Buchanan00 > tmin & df1$Buchanan00 < tmax)]  
  
par(mfrow = c(1, 3)) #we're telling R that we'll have 3 figures in a single row  
hist(Buchanan\_sans, main = "Histogram")  
boxplot(Buchanan\_sans, main = "Boxplot")  
qqnorm(Buchanan\_sans, main = "Normal Q-Q Plot")

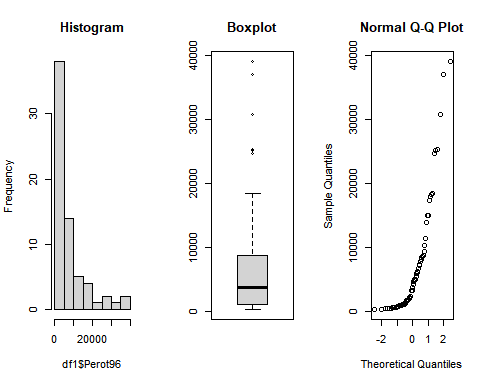


#hmm...looks like we might still have outliers. Do we?  
Buchanan\_sans[which(Buchanan\_sans < tmin | Buchanan\_sans > tmax)]

## numeric(0)

We don’t! great. How about Perot? par(mfrow = c(1, 3)) #we’re telling R that we’ll have 3 figures in a single row hist(Buchanan\_sans, main = “Histogram”) boxplot(Buchanan\_sans, main = “Boxplot”) qqnorm(Buchanan\_sans, main = “Normal Q-Q Plot”)

par(mfrow = c(1, 3))  
  
hist(df1$Perot96, main = "Histogram")  
boxplot(df1$Perot96, main = "Boxplot")  
qqnorm(df1$Perot96, main = "Normal Q-Q Plot")

 This looks like the 2nd Buchanan plots. But let’s just check:

mean = mean(df1$Perot96)  
std = sd(df1$Perot96)  
  
#find threshold values for outliers  
tmin = mean-(3\*std)  
tmax = mean+(3\*std)  
df1$Perot96[which(df1$Perot96 < tmin | df1$Perot96 > tmax)]

## [1] 38964 36990

I’m not going to touch these since we’re interested in the 2000 election, and we’re using the 1996 data as our baseline.

Great. On to Regression.

### Linear Regression with Palm Beach County *Included*

relation <- lm(df1$Buchanan00 ~ df1$Perot96)  
print(relation)

##   
## Call:  
## lm(formula = df1$Buchanan00 ~ df1$Perot96)  
##   
## Coefficients:  
## (Intercept) df1$Perot96   
## 1.34575 0.03592

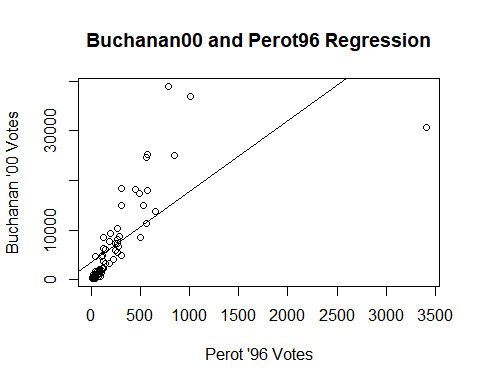
print(summary(relation))

##   
## Call:  
## lm(formula = df1$Buchanan00 ~ df1$Perot96)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -612.74 -65.96 1.94 32.88 2301.66   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.34575 49.75931 0.027 0.979   
## df1$Perot96 0.03592 0.00434 8.275 9.47e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 316.4 on 65 degrees of freedom  
## Multiple R-squared: 0.513, Adjusted R-squared: 0.5055   
## F-statistic: 68.48 on 1 and 65 DF, p-value: 9.474e-12

The regression line (R^2) suggests that about 51% of the variation in Buchanan00 can be explained by the Perot vote in ’96.

Let’s graph it:

plot(df1$Buchanan00, df1$Perot96, main = "Buchanan00 and Perot96 Regression",  
abline(lm(df1$Perot96~df1$Buchanan00)), xlab = "Perot '96 Votes", ylab = "Buchanan '00 Votes")



### Linear Regression with Palm County *Omitted*

#remove the outlying row:  
#identify the index of the outlier  
i <- which(df1$Buchanan00 == 3407)  
df2 <- df1[-c(i), ]

dim(df2)

## [1] 66 5

str(df2)

## 'data.frame': 66 obs. of 5 variables:  
## $ county : chr "Alachua" "Baker" "Bay" "Bradford" ...  
## $ Perot96 : num 8072 667 5922 819 25249 ...  
## $ Buchanan00 : num 263 73 248 65 570 788 90 182 270 186 ...  
## $ Buchanan\_zscore: num 0.00518 -0.41712 -0.02816 -0.4349 0.68751 ...  
## $ Perot\_zscore : num 0.0949 -0.7304 -0.1447 -0.7134 2.0092 ...

#check if PalmBeach is in there specifically  
"PalmBeach" %in% df2$county

## [1] FALSE

relation <- lm(df2$Buchanan00 ~ df2$Perot96)  
print(relation)

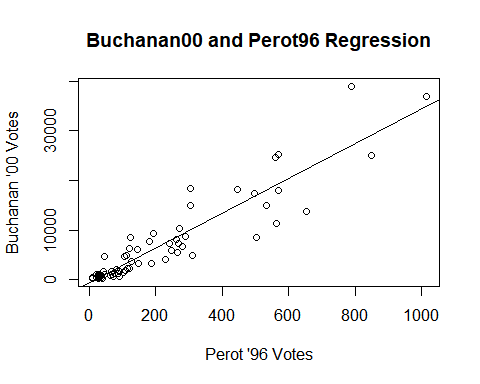
##   
## Call:  
## lm(formula = df2$Buchanan00 ~ df2$Perot96)  
##   
## Coefficients:  
## (Intercept) df2$Perot96   
## 45.84193 0.02435

print(summary(relation))

##   
## Call:  
## lm(formula = df2$Buchanan00 ~ df2$Perot96)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -206.70 -43.51 -16.02 26.92 269.03   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.841933 13.892746 3.30 0.00158 \*\*   
## df2$Perot96 0.024352 0.001273 19.13 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 87.75 on 64 degrees of freedom  
## Multiple R-squared: 0.8512, Adjusted R-squared: 0.8488   
## F-statistic: 366 on 1 and 64 DF, p-value: < 2.2e-16

The regression line fits much better (85%).

plot(df2$Buchanan00, df2$Perot96, main = "Buchanan00 and Perot96 Regression",  
abline(lm(df2$Perot96~df2$Buchanan00)), xlab = "Perot '96 Votes", ylab = "Buchanan '00 Votes")

 It seems there is a compelling argument to be made that the votes in Palm Beach County have been overestimated for Buchanan in 2000. And given the ambiguity of the Butterfly Ballot, it seems that many of the votes Buchanan received were intended for Gore. By ‘many’, we’re talking thousands…so not *that* many. But it’s an order of magnitude more than was required to push the hundreds that Bush won by over to Gore.