Diego Garcia-Olano
DAKD Fall, 2013
Assignment 2

Data Visualization with Slopegraphs

In selecting an example of what amounts to an expressive visualization in my opinion, I like many other students in the course went through countless examples and even implemented an interactive web-based datamap of the United States 2012 election inspired by the Gastner project referenced in class [1] though mine was targeted at the discovery of counties with similar situational and voting outcomes of my former home of Austin, Texas; a democrat dominated "blue" island within a largely republican "red" state[2]. Although the project has many merits, primarily that of allowing users to view counties with similar properties in the United States, I feel it has a ways to go before really telling the story I was hoping to express which is why these counties vote as they do; a concern I hope to alleviate in the near future by incorporating other data sets reflecting other social factors such as education attainment, population density, etc.

In the meantime, I decided to select an elegantly simple and expressive technique called the slopegraph which is a type of visualization created in 1983 by noted pioneer of information design and professor of statistics, political science, and computer science at Yale University, Edward Tufte, in his book "The visual display of quantitative information". [3][4] Slopegraphs compare changes over time for a list of ranked items. The leading example found on page 178 of the book can be seen in Figure 1.

This graphic ranks 15 countries by government tax collections in 1970 and in 1979, with the names spaced in proportion to the percentages. It is uncluttered and shows both the exact values a table would hold, but also clearly visually represents how the values have changed over time for each country and how they've changed in relation to one another. Because proportion is held via vertical spacing, the slopes of the lines can be easily read to show the overall general upward trend over the years for all countries, and easily compared to show countries whose percentages have grown faster or slower relative to others via their steepness or horizontal aspect respectively. This chart does this in a remarkably minimalist way, discarding the flashy, nonessential elements Tufte is outwardly critical of and having each element of the visualization show pertinent data.

From a data mining perspective, slopegraphs are of minimal use for data exploration purposes except perhaps for datasets with low dimensionality, but on the other hand, they do quite an effective job of telling a story of comparison between points over time. To further demonstrate this point in a later book, Beautiful Evidence, Tufte describes a scenario involving the rates of survival for patients with different sorts of cancer over a span of time. The table of data in Figure 2 is a subset of what he provides [5]. The table by itself contains an abundance of information, both the percentages per type over 5,10,15 and 20 years, but aside from initial ordering of survival, it is difficult to tell how the different types fare over time as a whole and individually. Putting the information into a slopegraph visualization renders Figure 3, again a subset, of the total data.

This depiction clearly shows the rate of change over time for each type of cancer, and contains the same specific data the simple table provides. The one drawback to this visualization in fact is the lack of maintenance of the relationship of order for the types as whole. Looking to the final column and the relation between “Prostate” and “Thyroid” order shows this drawback, but in this case, the point of the visualization is to show how each type behaves over time in comparison with others and not necessarily for all points. Thus, we can observe the fall in percentage of survival for the “Prostate” type over time versus the stability of “Thyroid” percentage while still maintaining its final general ordering amongst the other types through this.

The slopegraph form has been adapted and modified slightly for many other uses, including a quite informative visualization by MIT professor Ben Fry in his book Visualizing Data[6] amongst other interesting ones[7]. Fry's case involving the league standings of professional baseball teams in relation to their team's salary spending is interesting because as a standalone graphic, seen in Figure 4, it adds a different aspect of comparison, that of the win-loss ratio to that of spending as opposed to the usual measurement over time typical of slopegraphs. In the case of the online version [8], it adds back the common dimension of time, showing the relationship of how standings change over time in addition to how well a team’s spending is working overtime via the relationship of the connecting lines. Fry's version of the slopegraph adds thickness to lines, thicker lines for higher ordered elements in the right hand side, and color to lines, red for those with a positive change and blue for negative change. These additions intensify the visibility of negative relations of change for teams spending more than their relative standings, shown by thick red lines with steep increasing slopes, and on the opposite side, the positive relationships of teams with higher standing than relative cost shown by thin blue descending lines. This use of slopes as a quick visual determiner of relation between data points highlights one of the underlying benefits of these table graphics which is that slopes are easy to read.

In practice, slopegraphs are actually quite young, having been only officially given a name by Tufte in 2011 after almost thirty years of existence. Because of that, technical considerations needing to be accounted for and due in part to it's main use being for mainly univariate data, there don't exist too many pre-built libraries to implement them though I was able to find some in R[9] and javascript [10]. That said, they are a powerful and intuitive form of graphic which could be of great educational value when explaining datasets and thus I expect to see them more so in the future.

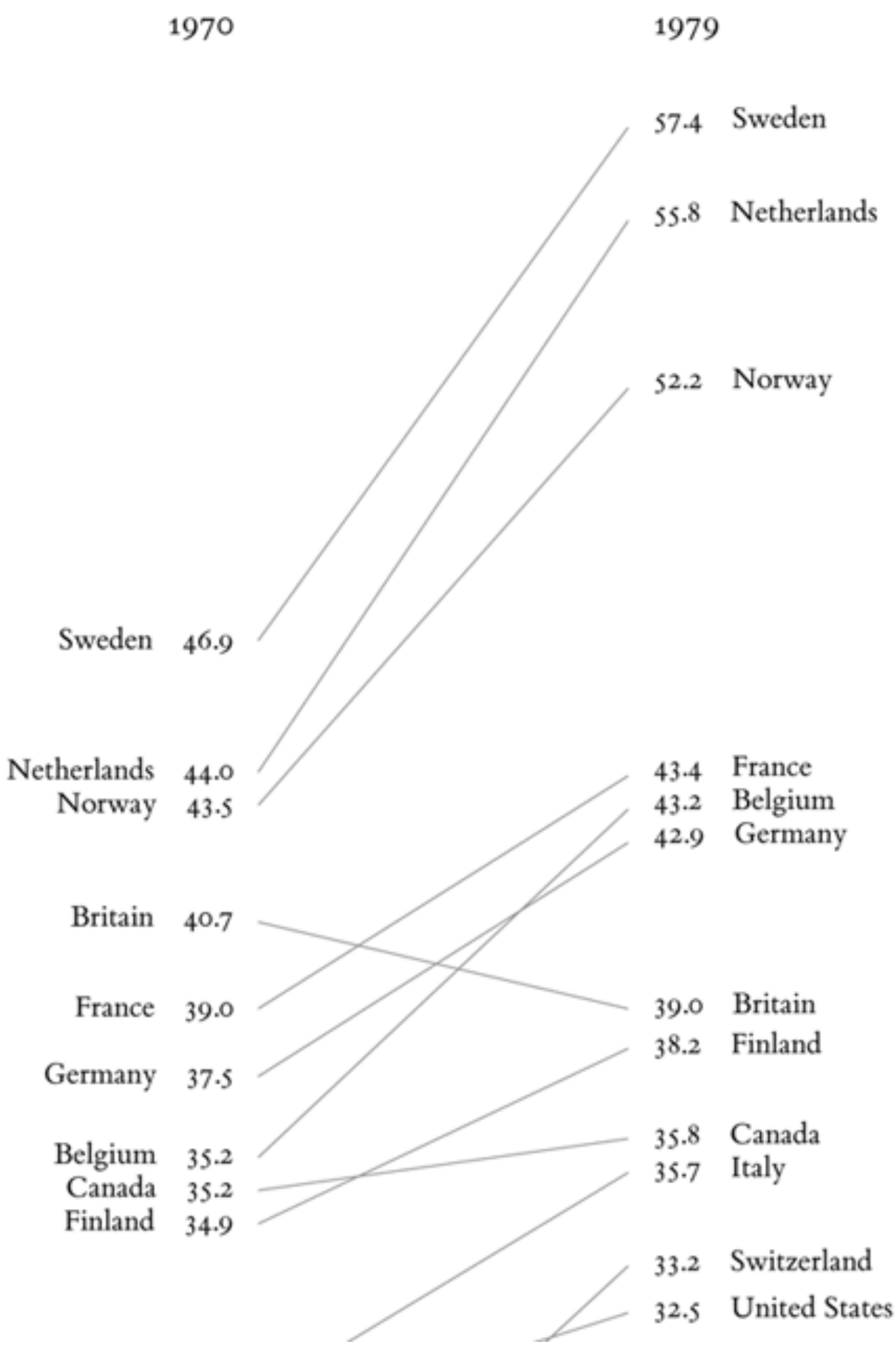
References:

[1] http://www-personal.umich.edu/~mejn/election/2012/   Michael Gastner, Cosma Shalizi, and Mark Newman (University of Michigan)
[2] http://www.diegoolano.com/electionmap/
[3] Tufte, Edward. The visual display of quantitative information. Cheshire, Connecticut: Graphics Press, 1983. p158,
[4] http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk
[5] Tufte, Edward. Beautiful Evidence. Cheshire, CT: Graphics Press, 2006. p174-6,
[6] Fry, Ben. Visualizing Data. Oreily Media, 2008.
[7] http://www.netindex.com/download/allcountries/
[8] http://fathom.info/salaryper/
[9] http://www.jameskeirstead.ca/blog/slopegraphs-in-r/
[10] http://skedasis.com/d3/slopegraph/

Figures:

Figure 1.

Current Receipts of Government as a Percentage of Gross Domestic Product, 1970 and 1979



1970
1979
57.4 Sweden
55.8 Netherlands
52.2 Norway
Sweden 46.9
Netherlands 44.0
Norway 43.5
43.4 France
43.2 Belgium
42.9 Germany
Britain 40.7
France 39.0
39.0 Britain
38.2 Finland
Germany 37.5
35.8 Canada
35.7 Italy
Belgium 35.2
Canada 35.2
Finland 34.9
33.2 Switzerland
32.5 United States

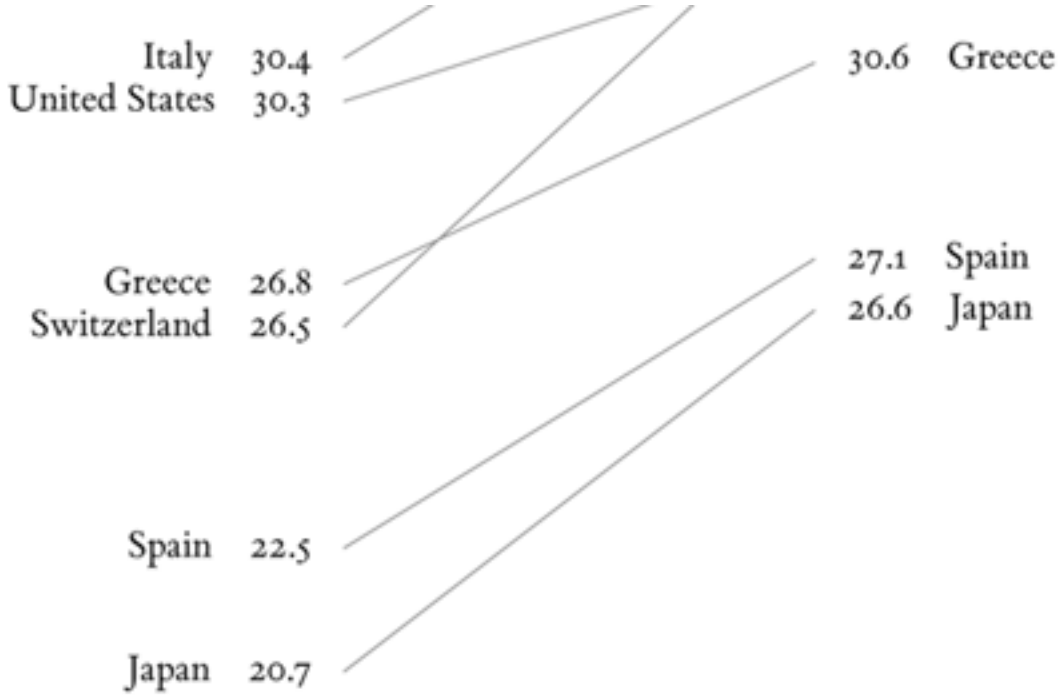
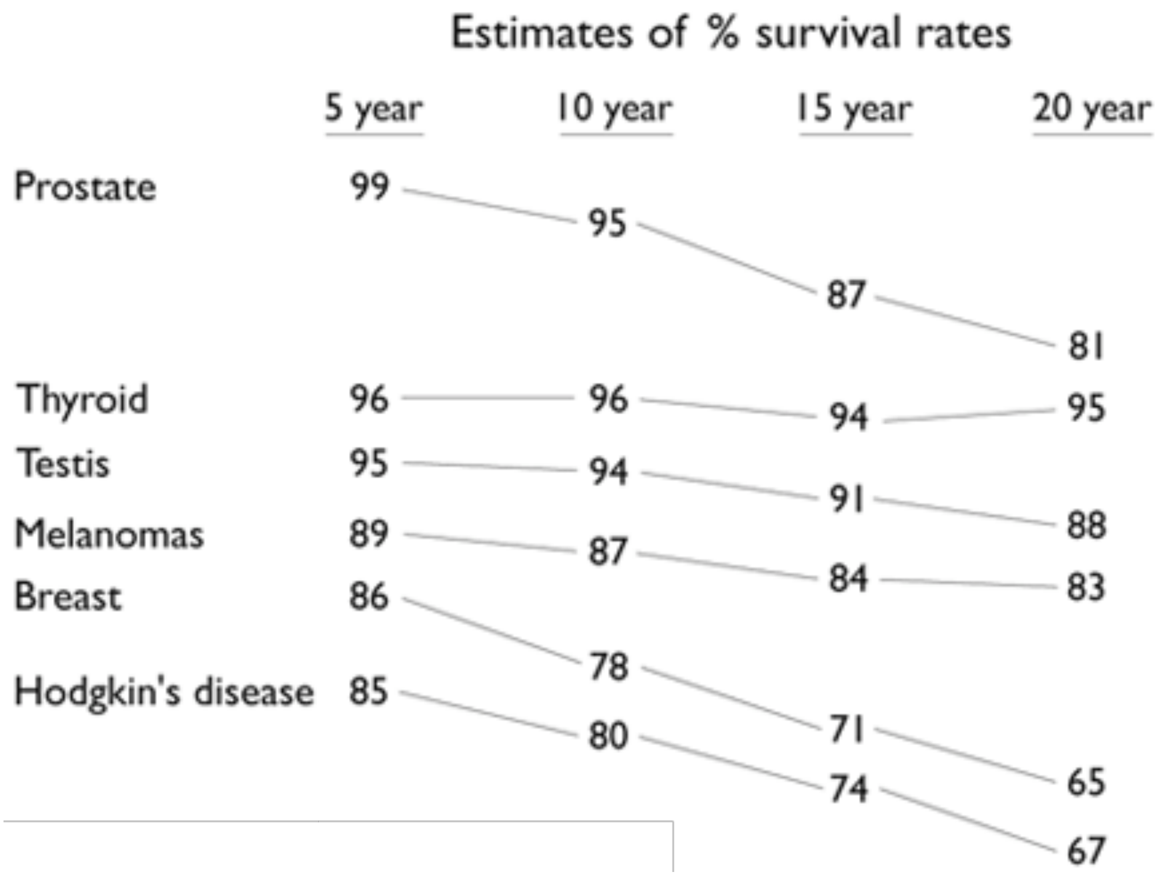


Figure 2.

Estimates of relative survival rates, by cancer site

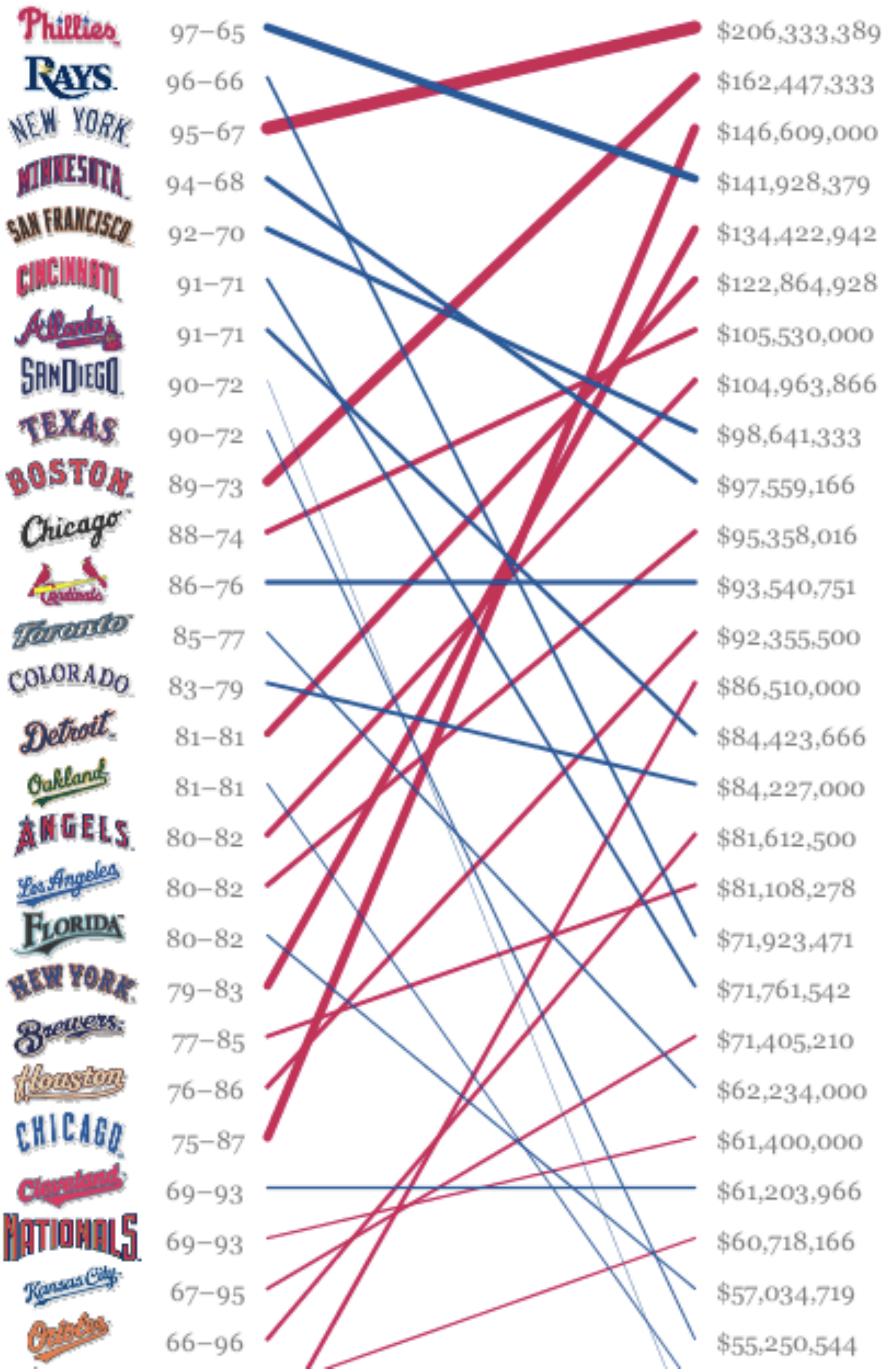
	% survival rates and their standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8

Figure 3.



Estimates of % survival rates
5 year
10 year
15 year
20 year
Prostate
99
95
87
81
Thyroid
96
96
94
95
Testis
95
94
91
88
Melanomas
89
87
84
83
Breast
86
78
71
65
Hodgkin's disease
85
80
74
67

Figure 4


Phillies
97–65
$206,333,389
Rays
96–66
$162,447,333
New York
95–67
$146,609,000
Minnesota
94–68
$141,928,379
San Francisco
92–70
$134,422,942
Cincinnati
91–71
$122,864,928
Atlanta
91–71
$105,530,000
San Diego
90–72
$104,963,866
Texas
90–72
$98,641,333
Boston
89–73
$97,559,166
Chicago
88–74
$95,358,016
Cardinals
86–76
$93,540,751
Toronto
85–77
$92,355,500
Colorado
83–79
$86,510,000
Detroit
81–81
$84,423,666
Oakland
81–81
$84,227,000
Angels
80–82
$81,612,500
Los Angeles
80–82
$81,108,278
Florida
80–82
$71,923,471
New York
79–83
$71,761,542
Brewers
77–85
$71,405,210
Houston
76–86
$62,234,000
Chicago
75–87
$61,400,000
Cleveland
69–93
$61,203,966
Nationals
69–93
$60,718,166
Kansas City
67–95
$57,034,719
Orioles
66–96
$55,250,544