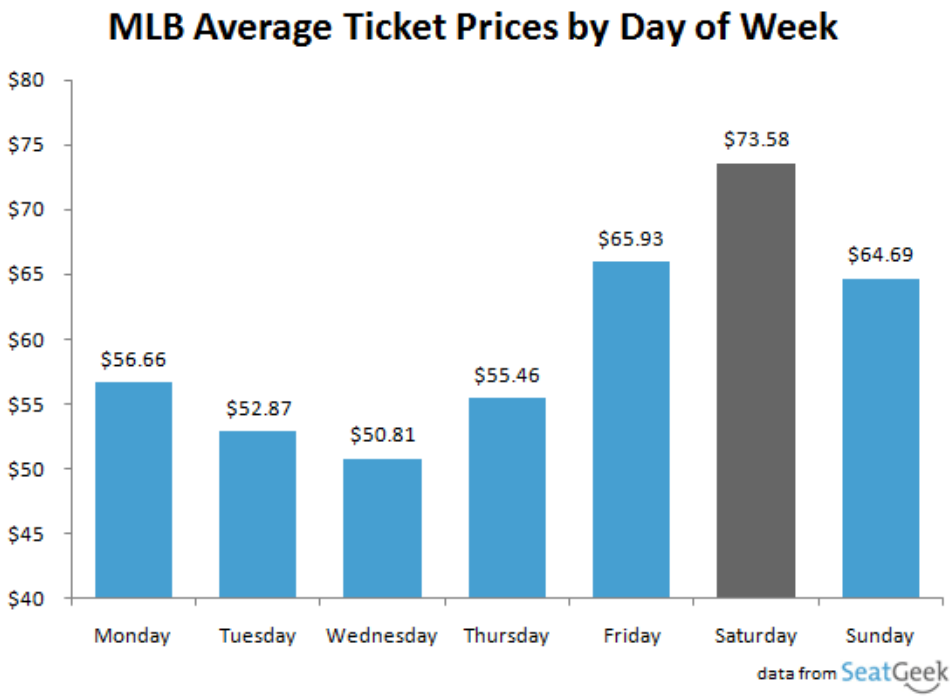
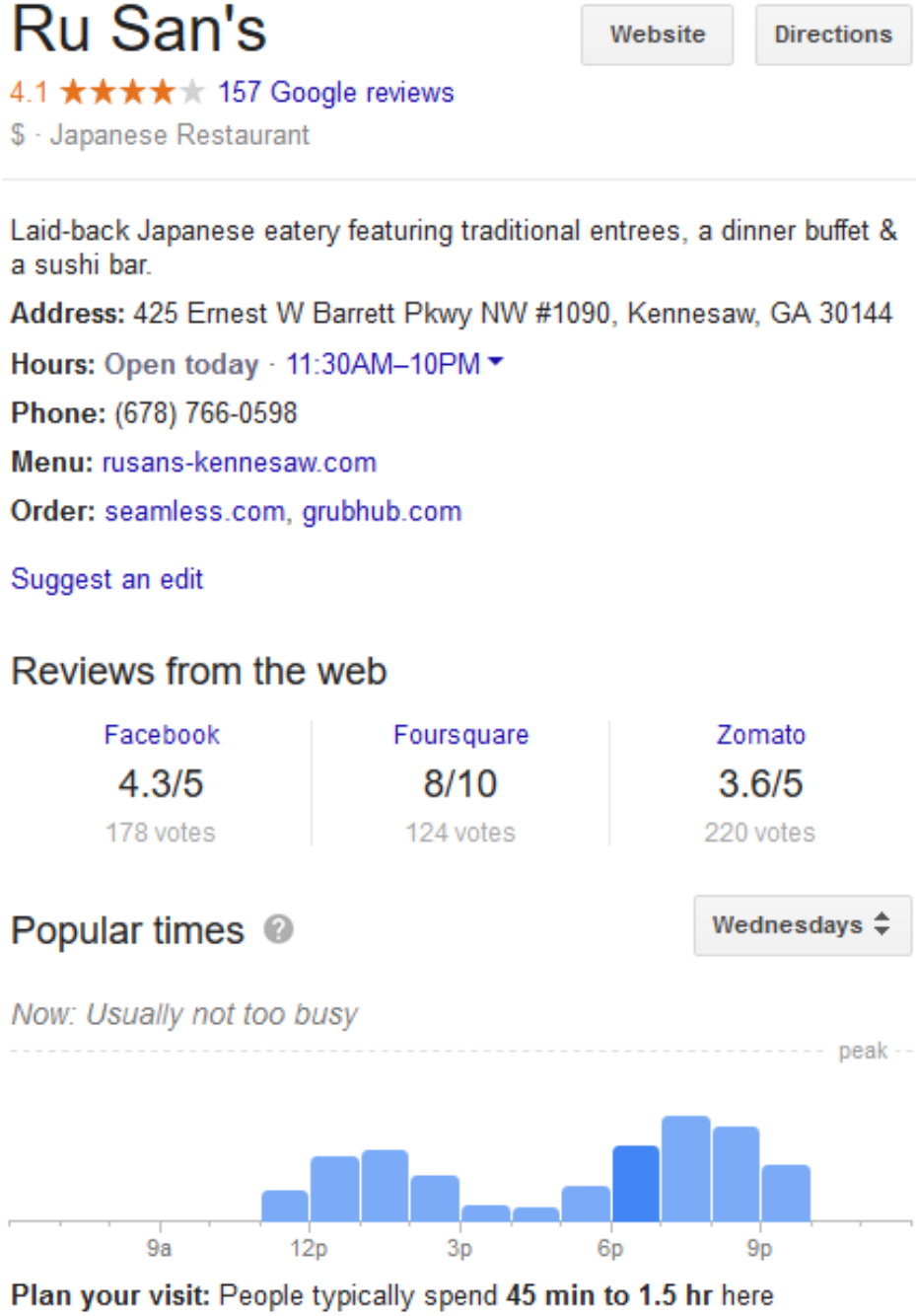
Section 2.1: Graphical Summaries of Data

1 Raw Data is Ugly

Graphical representations of data always look pretty in newspapers, magazines and books. What you haven't seen is the blood, sweat and tears that it sometimes takes to get those results.



http://seatgeek.com/blog/mlb/5-useful-charts-for-baseball-fans-mlb-ticket-prices-by-day-time



Google listing for Ru San's Kennesaw location

The previous examples are informative graphical displays of data. They all started life as bland data sets.

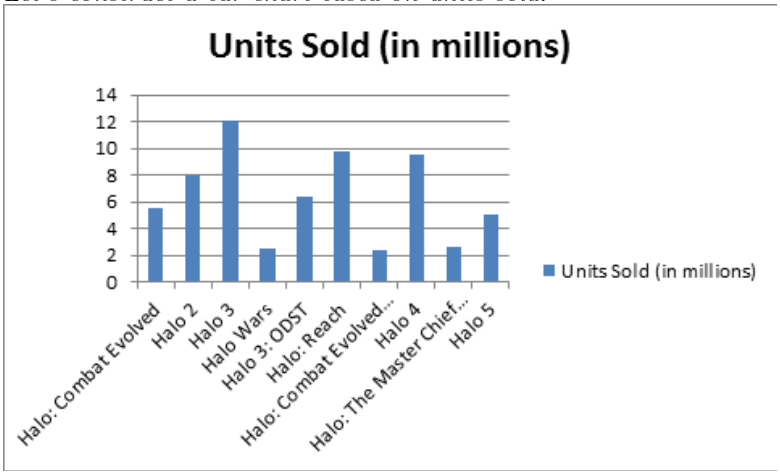
2 Bar Charts

Bar charts are a visual way to organize data. The height or length of a bar represents the number of points of data (**frequency distribution**) in a particular category. One can also let the bar represent the percentage of data (**relative frequency distribution**) in a category.

Example 1 *From (http://vgsales.wikia.com/wiki/Halo) consider the sales figures for Microsoft's videogame franchise, Halo. Here is the raw data.*

Year	*Game*	*Units Sold (in millions)*
2001	*Halo: Combat Evolved*	*5.5*
2004	*Halo 2*	*8*
2007	*Halo 3*	*12.06*
2009	*Halo Wars*	*2.54*
2009	*Halo 3: ODST*	*6.32*
2010	*Halo: Reach*	*9.76*
2011	*Halo: Combat Evolved Anniversary*	*2.37*
2012	*Halo 4*	*9.52*
2014	*Halo: The Master Chief Collection*	*2.61*
2015	*Halo 5*	*5*

Let's construct a bar chart based on units sold.

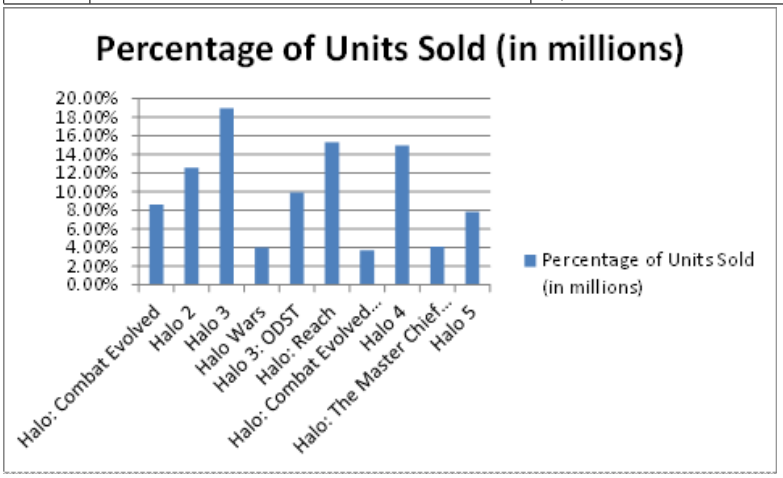


Translating frequencies into relative frequencies is an easy process. A relative frequency is the frequency divided by total number of data points.

Example 2 *Create a relative frequency chart for Halo games. First, determine how many data points exist. We do this by adding all the frequencies together. sum of all frequencies* $= 5.5+8+12.06+2.54+6.32+9.76+2.37+9.52+2.61+5 =$

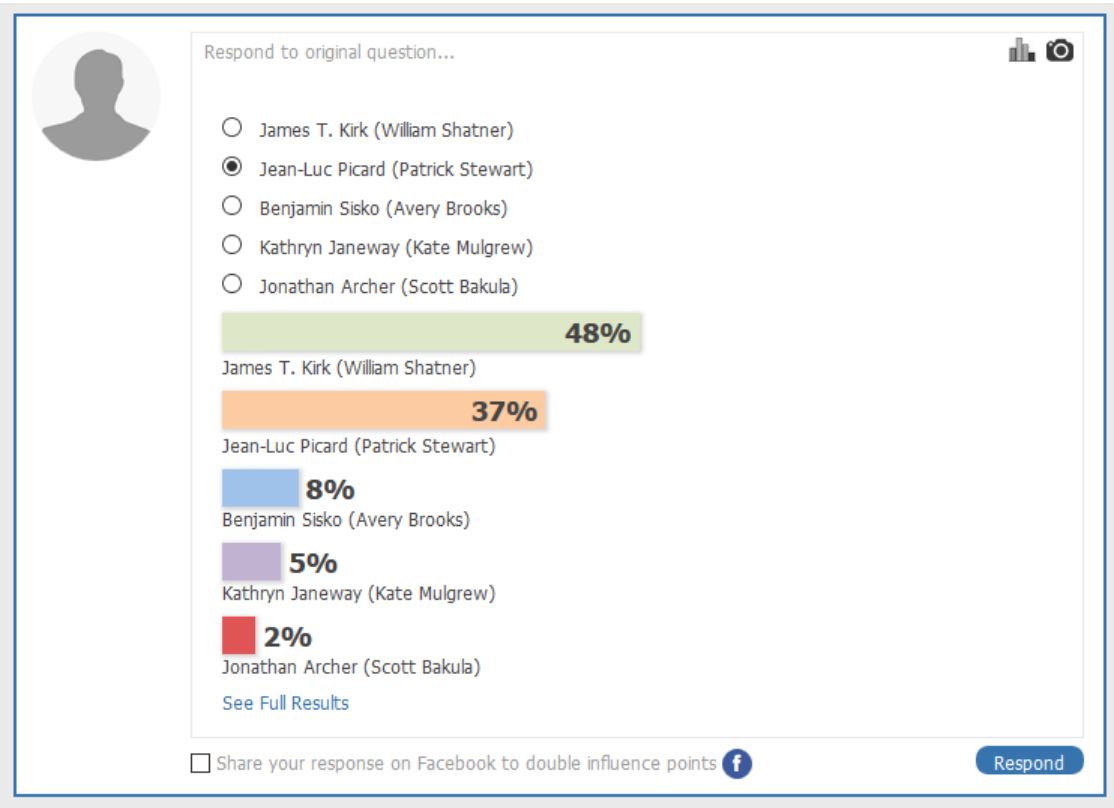
63.68. *So 63.68 million Halo games have been sold. Next we divide each frequency by 63.68 to turn it into a percentage of relative frequency.*

Year	*Game*	*Percentage of Units Sold (in millions)*
2001	*Halo: Combat Evolved*	$5.5/63.68 = 8.63\%$
2004	*Halo 2*	$8/63.68 = 12.56\%$
2007	*Halo 3*	$12.06/63.68 = 18.94\%$
2009	*Halo Wars*	$2.54/63.68 = 3.99\%$
2009	*Halo 3: ODST*	$6.32/63.68 = 9.92\%$
2010	*Halo: Reach*	$9.76/63.68 = 15.32\%$
2011	*Halo: Combat Evolved Anniversary*	$2.37/63.68 = 3.72\%$
2012	*Halo 4*	$9.52/63.68 = 14.95\%$
2014	*Halo: The Master Chief Collection*	$2.61/63.68 = 4.1\%$
2015	*Halo 5*	$5/63.68 = 7.85\%$

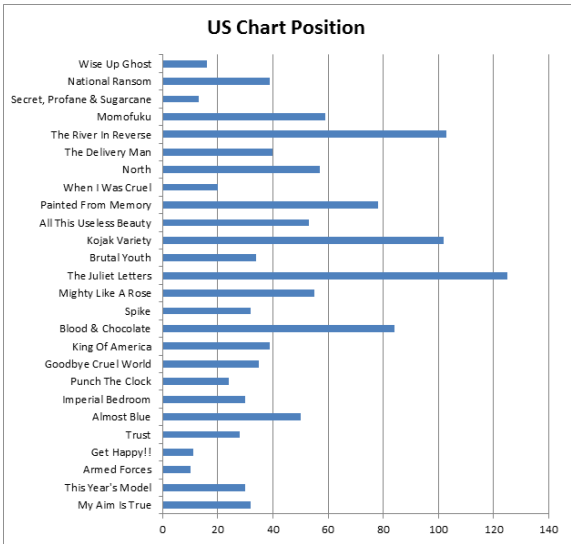


Note that despite changing from raw numbers to frequencies, the bars maintain the same height relative to each other. Why?

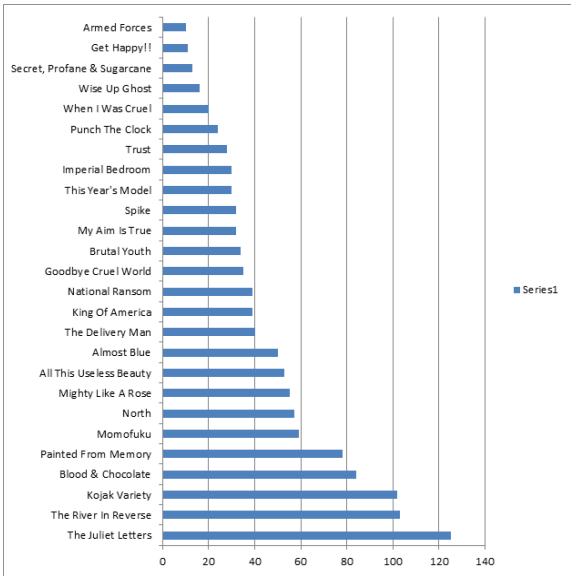
Example 3 *A website (https://www.rallypoint.com/answers/which-star-trek-captain-is-the-best#) asks which Star Trek Captain is the best. The resulting vote (as of June 2016) is easily viewed as a bar char*



Star Trek Captain Preference



Top position on US album chart for Elvis Costello in chronological order

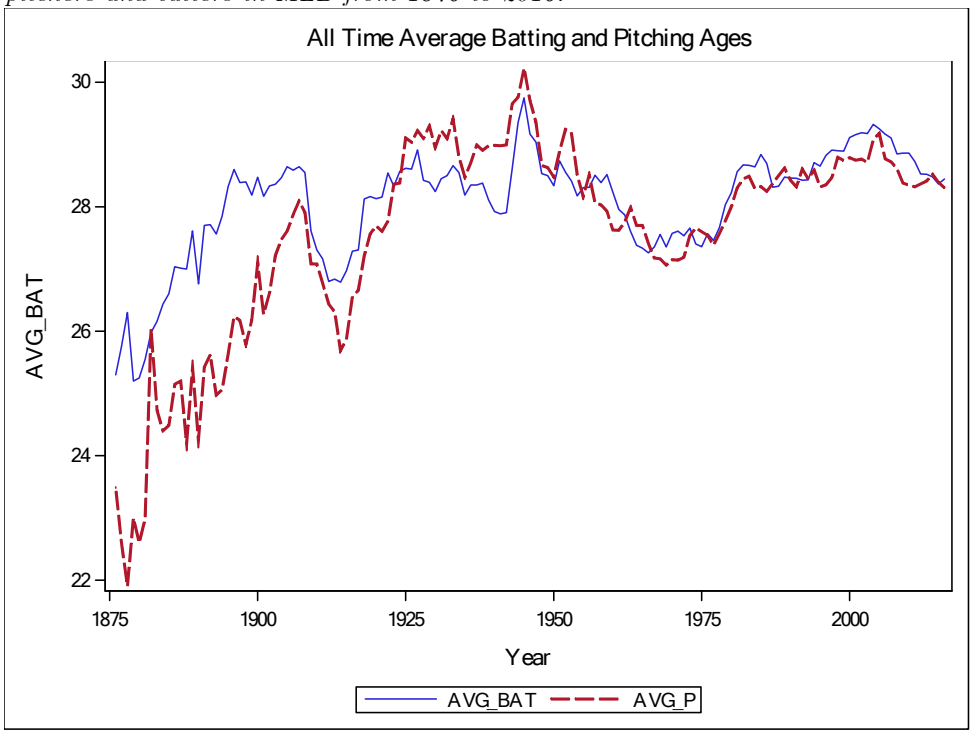


Position on US album chart for Elvis Costello

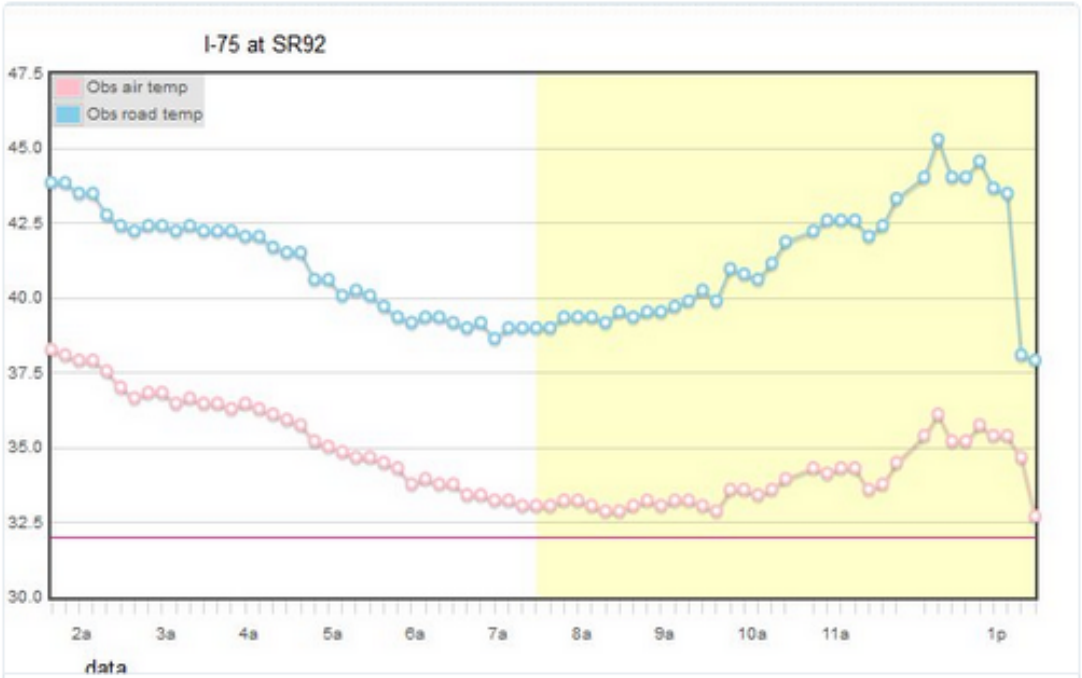
Definition 1 *A bar chart ordered by frequencies is called a* ***Pareto chart****.*

Definition 2 *A time series is a bar chart that represents data as it changes over time.*

Example 4 *Here we use a time series chart to represent the average age of pitchers and batters in MLB from 1876 to 2016.*



Example 5 *The following time series documents the drop in air temperature and road temperature at I-75 and GA-92 near Acworth.*



Brad Nitz
@BradNitzWSB

Follow

Road surface temp has fallen from 43° to 38° in the last 20 minutes on I-75 at SR 92 near Acworth.

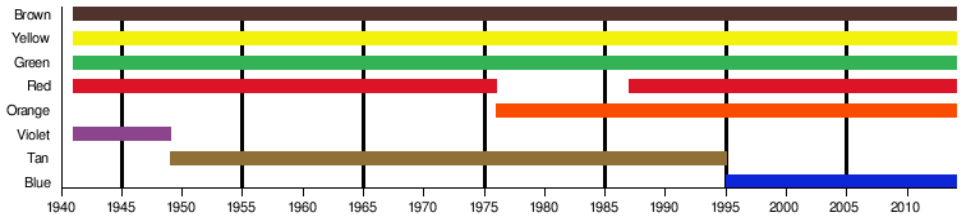
Time to head home.

1:44 PM - 6 Jan 2017

35 19

AJC.com on 1/6/2017

Example 6 *The following is a time series for colors of M&M's used from 1940 to 2010 (https://en.wikipedia.org/wiki/M%26M%27s).*



1940–2010

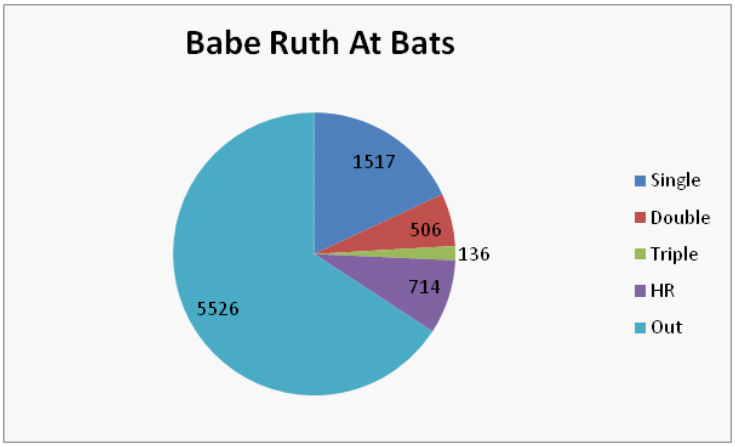
3 Pie Charts

Pie charts are a visual method for displaying the categories of a collection of data. The size of a slice of the pie chart is proportional to the percentage of data in that category.

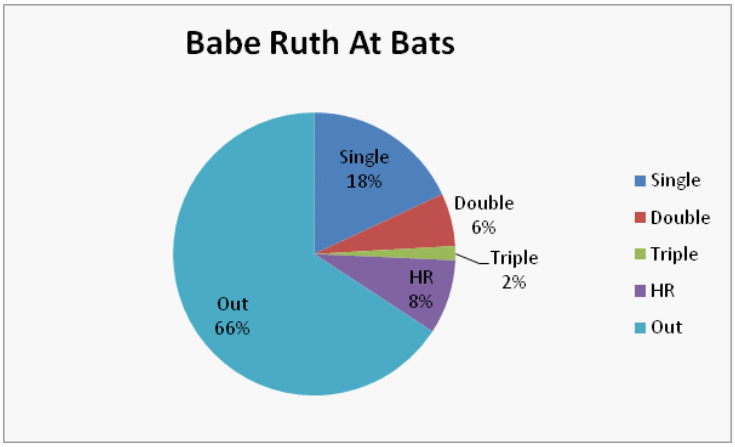
Example 7 *Consider the distribution of hits over the career of Babe Ruth. We visually represent this data in a pie chart. We can label the slices with either the raw numbers of percentages for each category*

Babe Ruth	At Bats
Single	1517
Double	506
Triple	136
HR	714
Out	5526

Babe Ruth Statistics



Pie chart with raw numbers

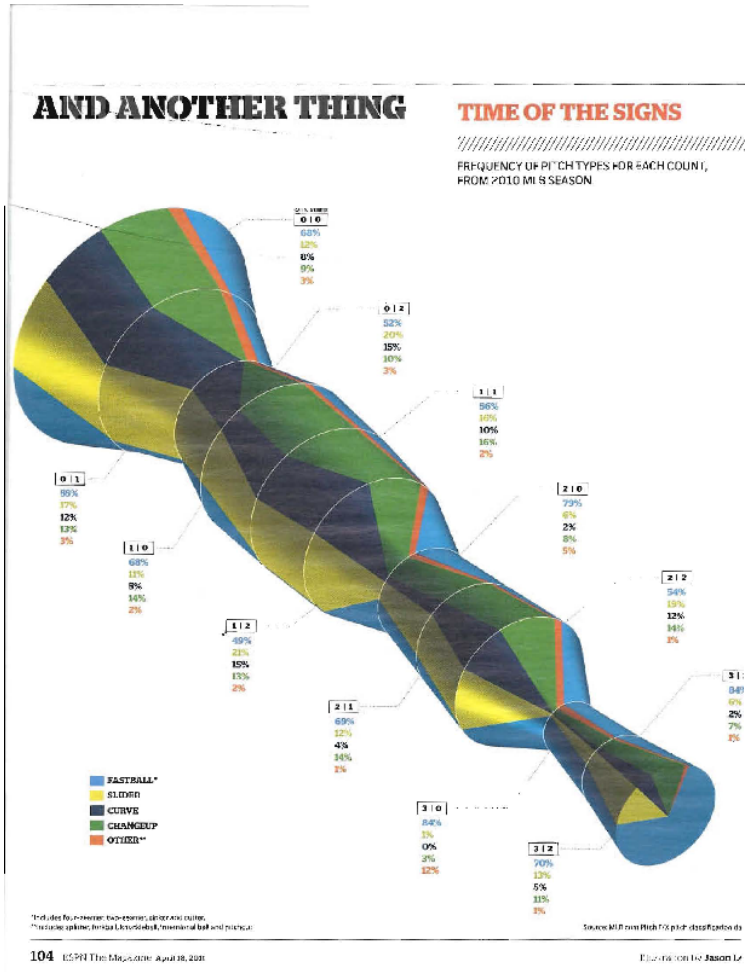


Pie chart with percentages and wedge labels

4 Pretty $\neq$ Good

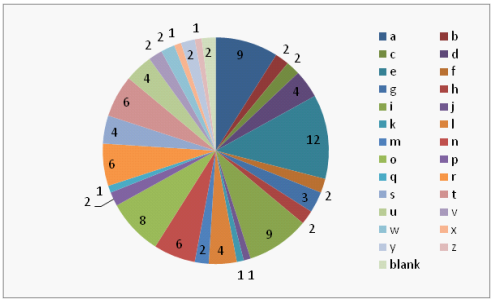
A graphical display of data must be informative. Avoid the temptation to make the graphic fancy at a cost of clarity of information.

Example 8 *Consider the following graph from ESPN the Magazine.*



Example 9 *You must be careful not to overload a pie chart with too much detail. The distribution of tiles at the start of a game of Scrabble is given below. This seems like a natural concept to represent with a pie chart. However, once you construct the pie chart for this distribution it is easy to see that too much information is contained in the picture.*

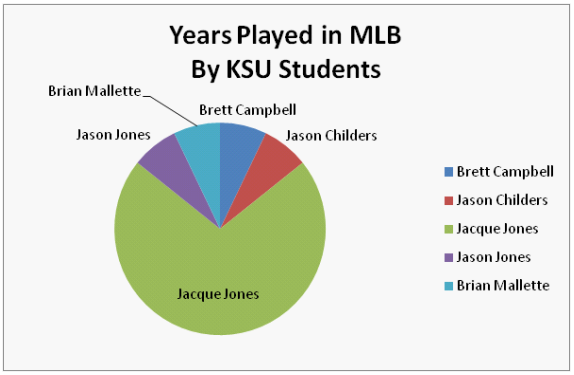
Letter	*Frequency*	*Letter*	*Frequency*	*Letter*	*Frequency*
a	*9*	*j*	*1*	*s*	*4*
b	*2*	*k*	*1*	*t*	*6*
c	*2*	*l*	*4*	*u*	*4*
d	*4*	*m*	*2*	*v*	*2*
e	*12*	*n*	*6*	*w*	*2*
f	*2*	*o*	*8*	*x*	*1*
g	*3*	*p*	*2*	*y*	*2*
h	*2*	*q*	*1*	*z*	*1*
i	*9*	*r*	*6*	*blank*	*2*



Frequencies of Scrabble tiles

5 Exercises

1. Based on the following pie chart, (as of 2011) which KSU student spent the most years playing professional major league baseball?

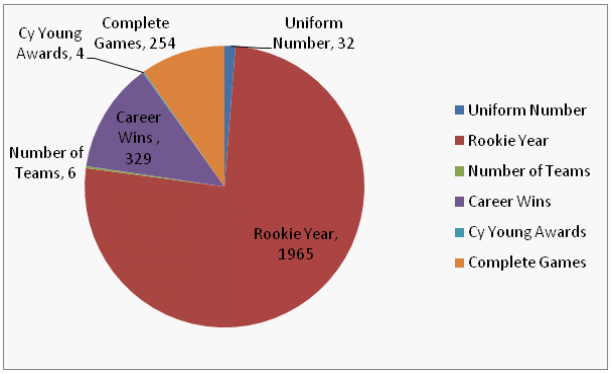


KSU Owls in MLB

2. Some statistics on Baseball Hall of Fame pitcher Steve Carlton are given below.

Steve Carlton	
Uniform Number	32
Rookie Year	1965
Number of Teams	6
Career Wins	329
Cy Young Awards	4
Complete Games	254

Does it make sense to use a pie chart to represent this data?



Quantitative data about Steve Carlton

3. Navidi/Monk Section 2.1 Exercises: 5-8, 13, 14, 17, 18,