

Univariate Graphics

STAT 133

Gaston Sanchez

Department of Statistics, UC–Berkeley

`gastonsanchez.com`

`github.com/gastonstat/stat133`

Course web: `gastonsanchez.com/teaching/stat133`

Looking at one single variable

Univariate Statistical Graphics

Getting started with graphics for exploration requires understanding charts and plots for single variables

Univariate graphics by type of variable

Qualitative Variable

- ▶ Bar chart
- ▶ Dot chart
- ▶ Pie chart
- ▶ Pareto chart

Quantitative Variable

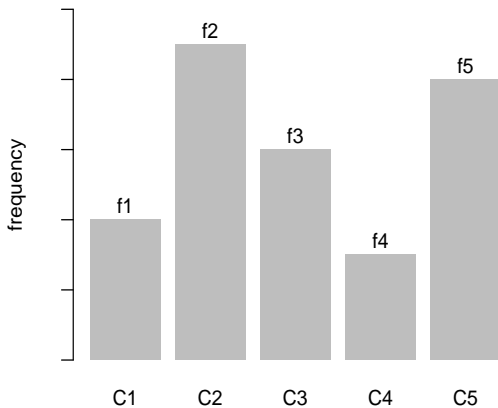
- ▶ All of qualitative
- ▶ Histogram
- ▶ Density curve
- ▶ Boxplot
- ▶ Ogive

Bar Charts

From Frequency Tables ...

Category	Absolute Frequency	Relative Frequency
C_1	f_1	f_1/n
C_2	f_2	f_2/n
C_3	f_3	f_3/n
\dots	\dots	\dots
C_k	f_k	f_k/n
$total$	n	1

to Bar-charts



Bar-charts

Elements of vertical bar-charts

- ▶ categories on horizontal axis
- ▶ frequencies on vertical axis
- ▶ length of bar equal to frequency

(Note that you can also make a horizontal bar-chart, in which case the axes play inverse roles)

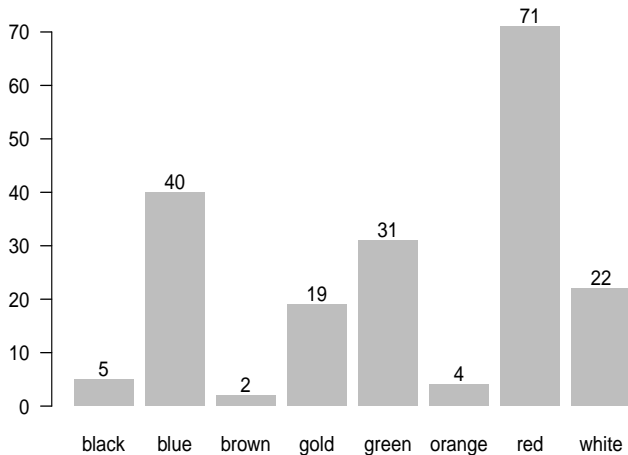
Bar-chart: predominant color in flags



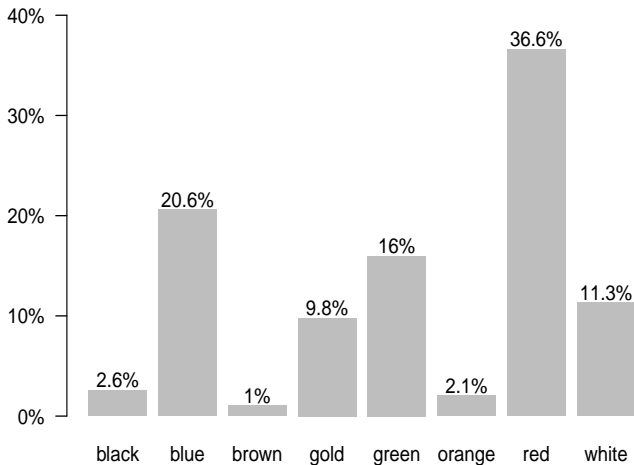
Predominant Color in Flags

##	color	count	percent
## 1	black	5	2.58
## 2	blue	40	20.62
## 3	brown	2	1.03
## 4	gold	19	9.79
## 5	green	31	15.98
## 6	orange	4	2.06
## 7	red	71	36.60
## 8	white	22	11.34

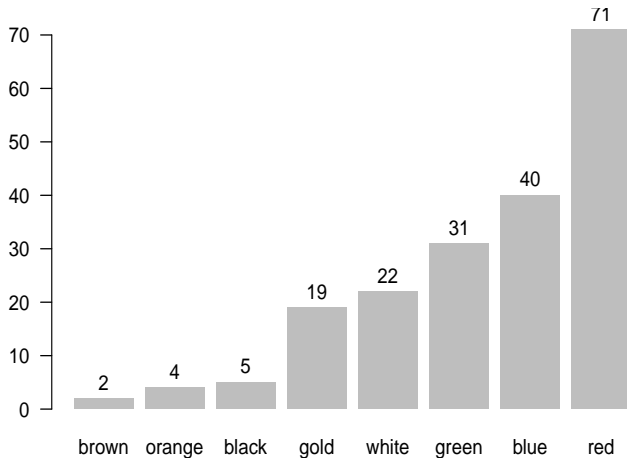
Bar-chart example



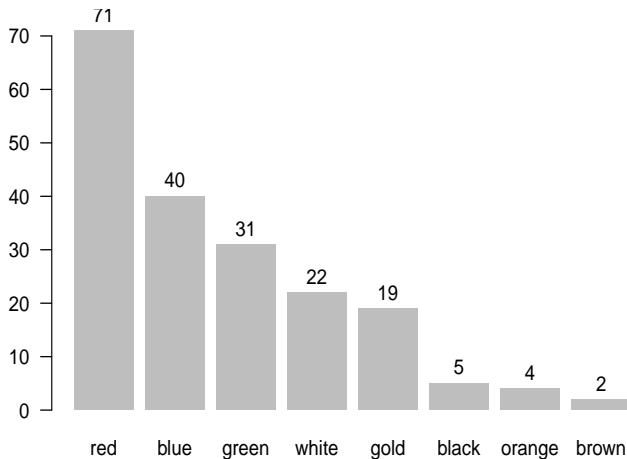
Bar-chart: predominant color in flags



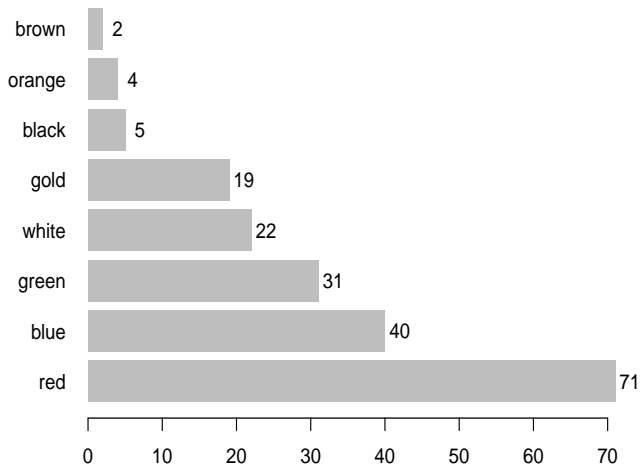
Bar-chart: predominant color in flags



Bar-chart: predominant color in flags



Bar-chart: predominant color in flags

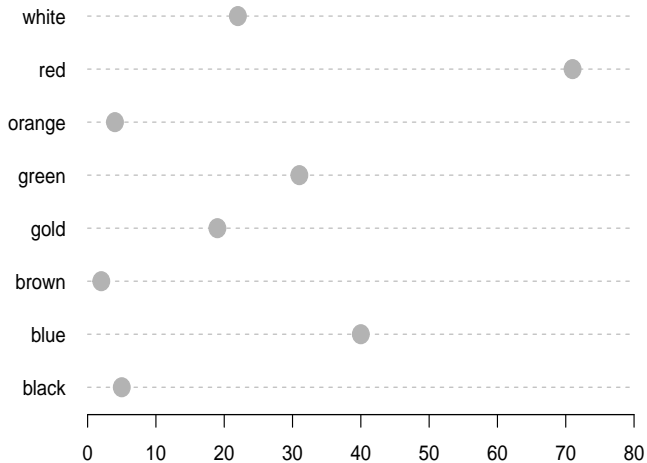


Dot charts

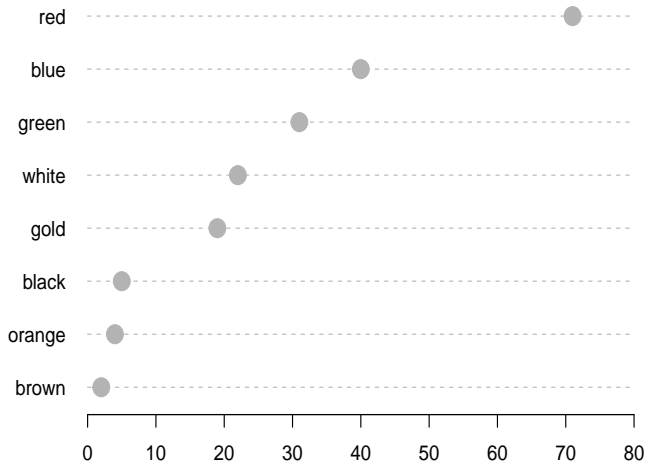
Dot charts

- ▶ Dot-charts are very similar to bar charts.
- ▶ Instead of using bars, dot-charts display frequencies with dots.
- ▶ They are simpler and cleaner than bar charts
- ▶ They are also less used than bar charts

Dot-chart: predominant color in flags



Ranked Dot-charts



Ranked dot-chart patterns



all values roughly the same



differences decrease by
roughly the same amount



differences from one value to
the next vary significantly



differences from one value
to the next increase

Ranked dot-chart patterns



differences from one value
to the next decrease



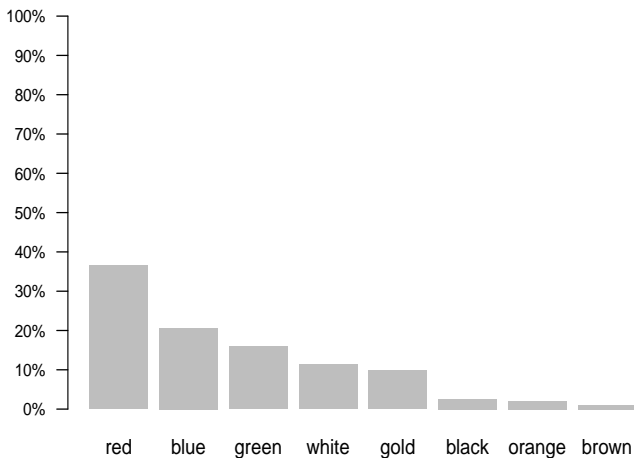
shifting differences from
one value to the next



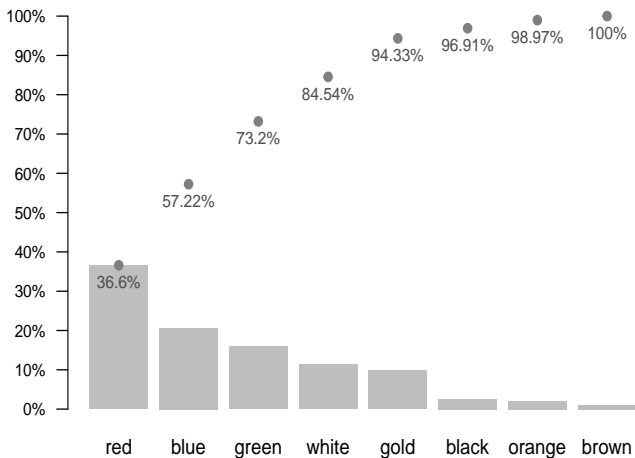
one or more values are extraordinarily
different from the rest

Pareto charts

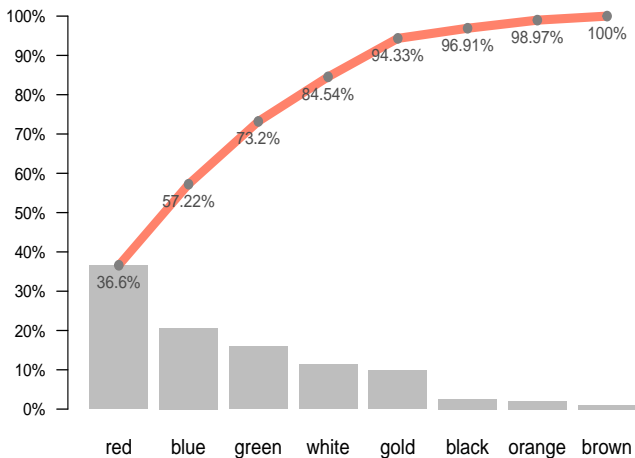
Bar-chart with Pareto Line



Bar-chart with Pareto Line



Bar-chart with Pareto Line

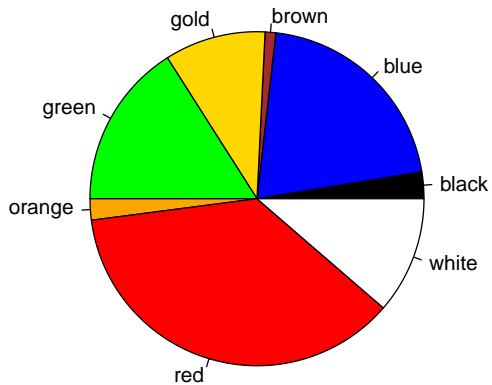


Pareto charts

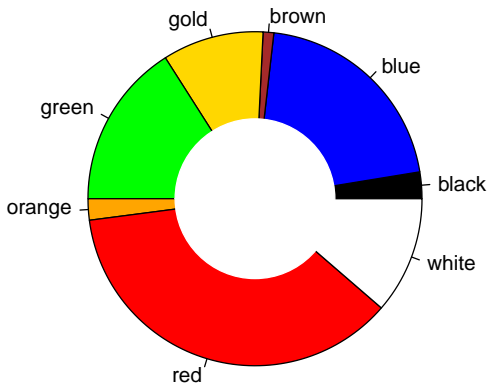
- ▶ Pareto charts contains both bars and a line graph
- ▶ Individual values are representing in descending order
- ▶ Cumulative frequencies are represented by the line
- ▶ The left vertical axis is the frequency of occurrence

Pie charts

Pie Chart



Donut Chart



Pie charts disadvantages

- ▶ Pie charts force us to compare either 2-D areas formed by each slice or the angles formed
- ▶ Visual perception handles neither of these comparisons easily or accurately

Univariate Quantitative Charts

NFL Ticket prices (2013)

##	teams	tickets	teams	tickets
## 1	cowboys	110.20	falcons	83.71
## 2	patriots	117.84	vikings	78.69
## 3	giants	111.69	rams	74.49
## 4	bears	103.60	seahawks	71.21
## 5	jets	110.28	cardinals	79.56
## 6	redskins	94.80	dolphins	71.14
## 7	ravens	100.19	raiders	64.80
## 8	eagles	93.01	titans	65.28
## 9	texans	88.98	lions	67.60
## 10	chargers	84.55	bengals	68.96
## 11	steelers	81.13	jaguars	68.44
## 12	packers	82.61	chiefs	64.92
## 13	49ers	83.54	buccaneers	63.59
## 14	saints	74.99	bills	57.75
## 15	broncos	84.27	panthers	66.84
## 16	colts	86.32	browns	54.20

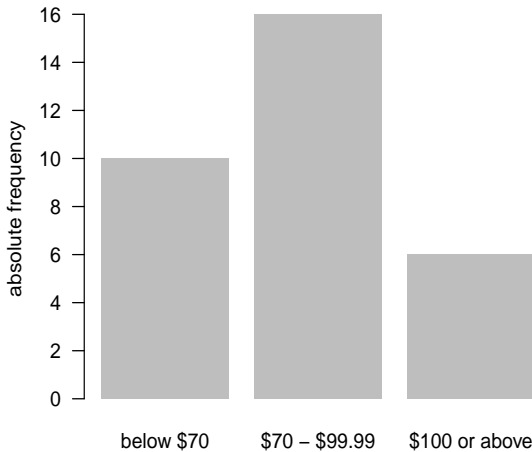
Bar charts for quantitative variables

- ▶ We can use bar charts with quantitative variables
- ▶ In this case we need to first categorize the variable, and then get a frequency table

Frequency Table of Ticket Prices

Category Name	Absolute Frequency	Relative Frequency
Below \$70	10	0.3125
\$70 - \$99.99	16	0.5000
\$100 or above	6	0.1875
Total	32	1.00

NFL Ticket prices (2013)

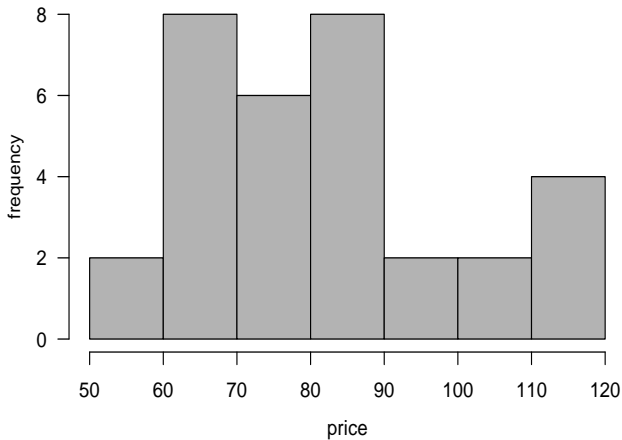


Histograms

Histograms

Histograms provide a way of viewing the general distribution of values in a quantitative variable

NFL Ticket prices (2013)



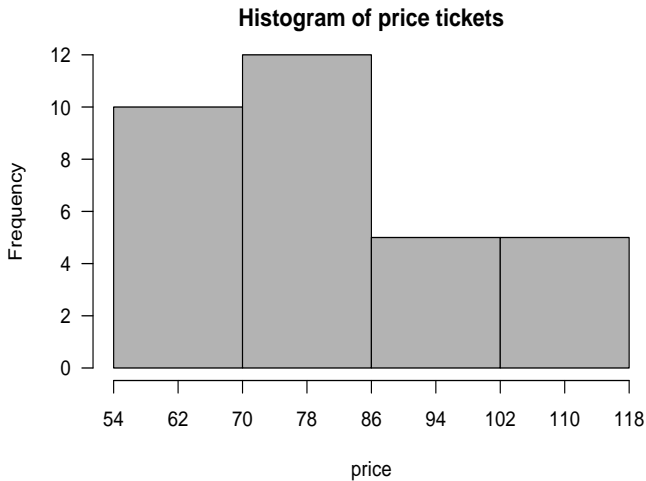
Building a Histogram

1. **Partition of values:** The range of the data values is partitioned into a number of non-overlapping “cells” or bins.
2. **Counting frequencies:** The number of data values falling into each cell is counted (either absolute or relative freqs)
3. **Drawing Bars:** The observations falling into a cell are represented as a “bar” drawn over the cell

About Histograms

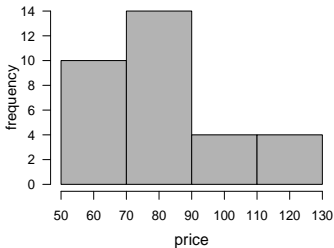
- ▶ The bins represent ranges of values
- ▶ The bins (intervals) must be adjacent, and usually of equal size
- ▶ The bars are adjacent (not discontinuous)
- ▶ The areas of the bars are meaningful
- ▶ Height of bars equal to the frequency
- ▶ Width equal to the bin size
- ▶ The area of a bar gives the proportion of data values which fall in the bin

Histogram with 4 bins

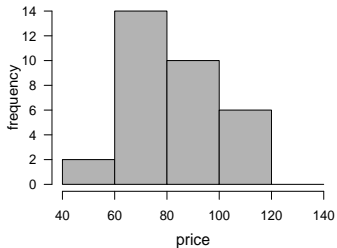


Histograms with different bins

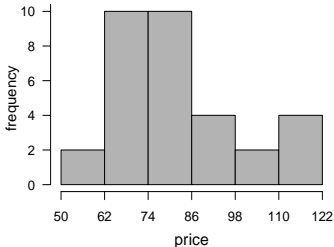
Histogram of price tickets (4 bins)



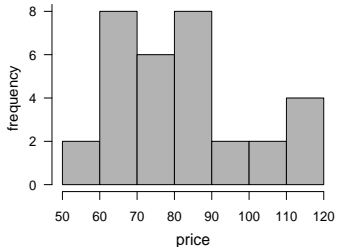
Histogram of price tickets (5 bins)



Histogram of price tickets (6 bins)

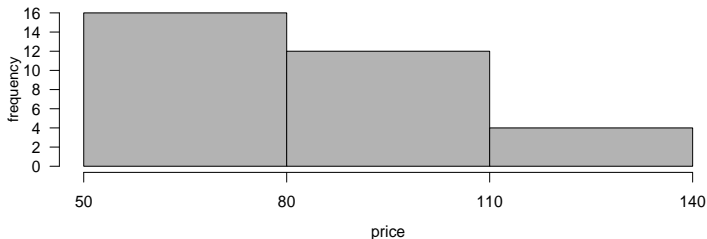


Histogram of price tickets (7 bins)

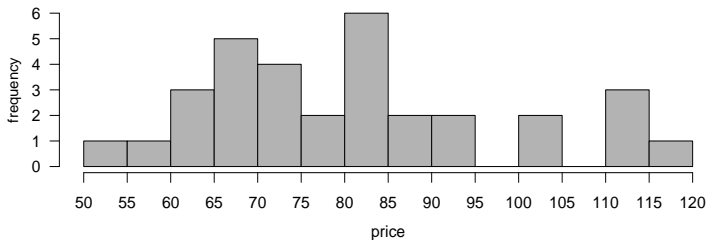


Avoid too few and too many bins

Histogram of price tickets (3 bins)



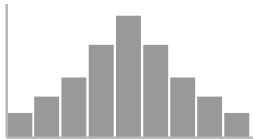
Histogram of price tickets (14 bins)



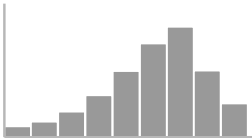
About Histograms

- ▶ The shape of a histogram depends on the chosen bins
- ▶ This suggests that there is a fundamental instability at the heart of its construction
- ▶ The bars are adjacent (not discontinuous)
- ▶ The areas of the bars are meaningful

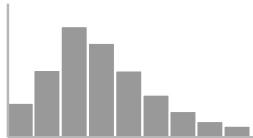
Histogram patterns



Symmetrical

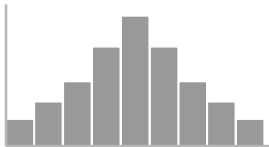


Skewed to the left

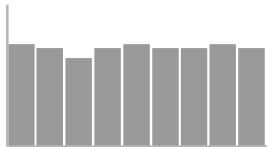


Skewed to the right

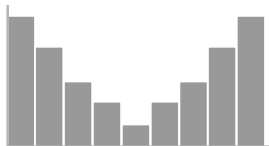
Histogram patterns



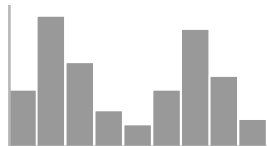
Curved



Flat or Uniform

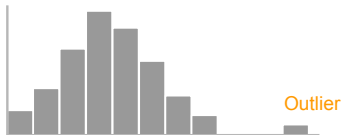
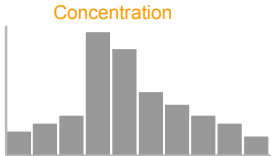


Curved Downward

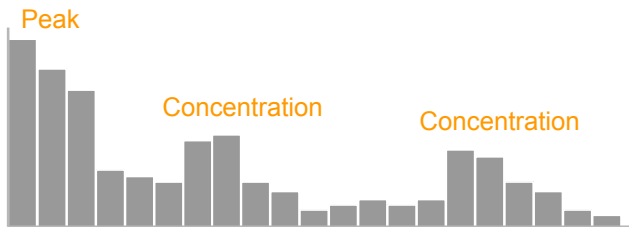


Multiple peaks
(e.g. bimodal, trimodal. etc)

Histogram patterns



Histogram patterns



Box plots

Building a Histogram

1. **Box-and-whisker plots**, most commonly known as “box plots”
2. created by John Tukey
3. simple and effective way to display the distribution of values
4. relies on the so-called **5-summary indicators**

Box plots based on 5-number summary

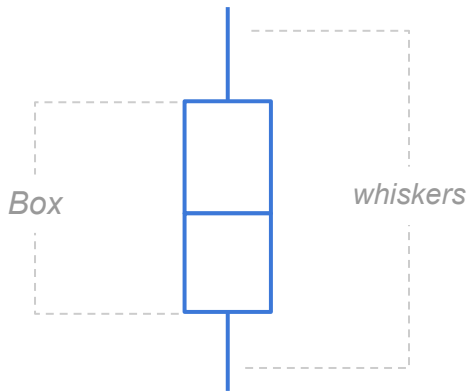
5 summary indicators

Box plots based on 5-number summary

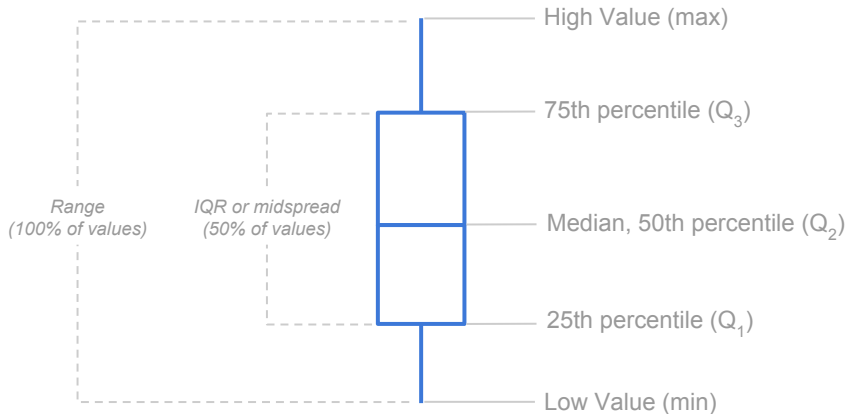
5 summary indicators

1. minimum
2. 25th percentile (1st quartile)
3. 50th percentile (2nd quartile, or median)
4. 75th percentile (3rd quartile)
5. maximum

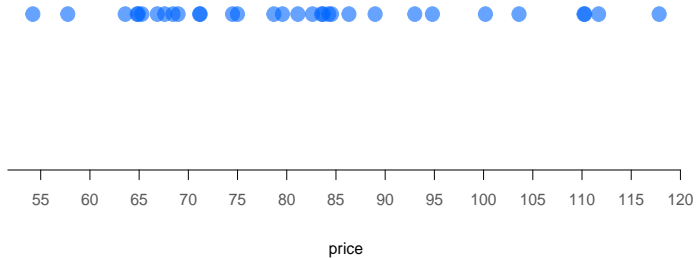
Box plot basics



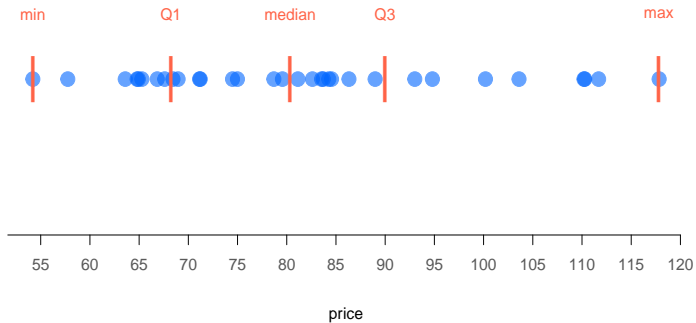
Box plot basics



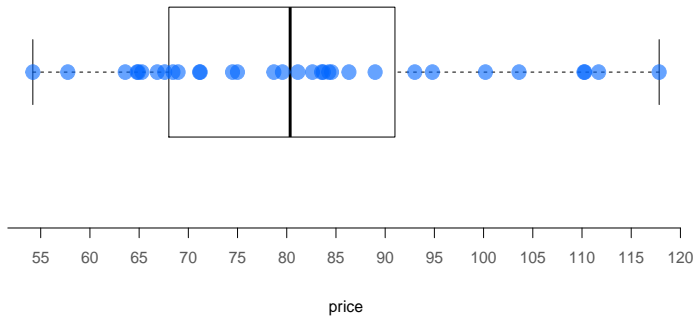
NFL Ticket Prices



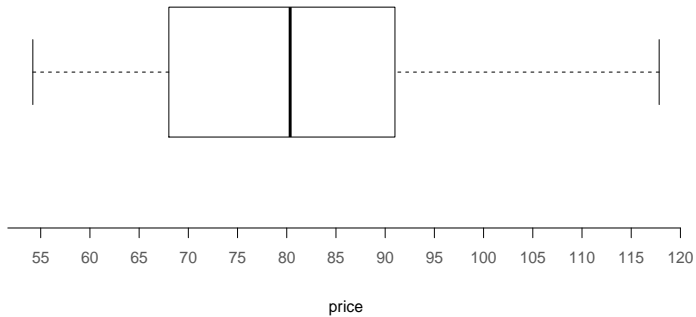
5 number summary



Box plot



Box plot

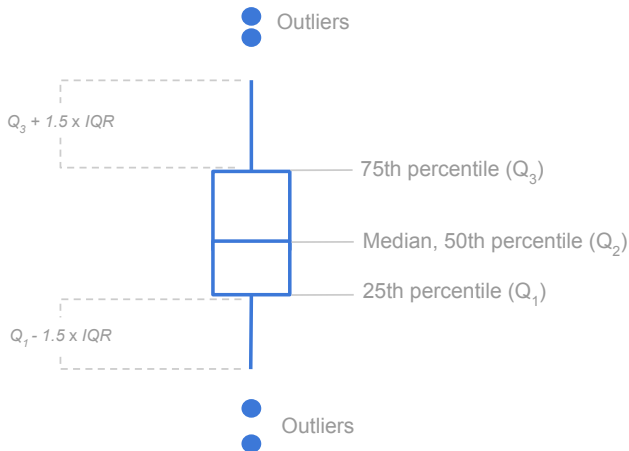


Box plot and outliers

The $1.5 \times \text{IQR}$ rule for outliers

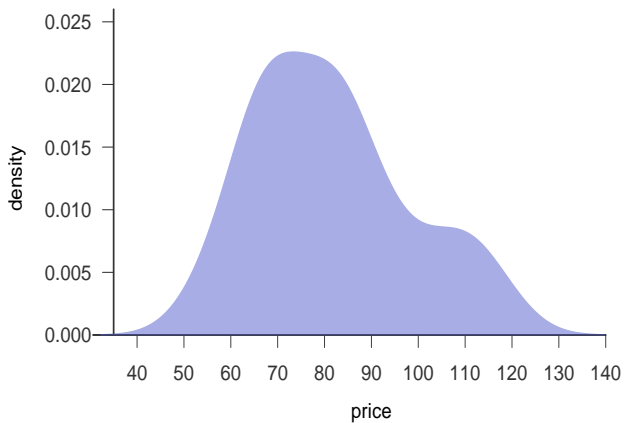
Call an observation a suspected outlier if it falls more than $1.5 \times \text{IQR}$ above the third quartile or below the first quartile

Modified Box plot



Density Curves

Density Curve



Density Curve

A Density Curve

- ▶ Describes the distribution of values by a smooth curve
- ▶ Is always on or above the horizontal axis
- ▶ Has area equal to 1 underneath it
- ▶ Is an idealized distribution

Density Curve

About Density Curve

- ▶ The **mode** is the peak point of the curve (could be more than one or none)
- ▶ The **median** is the equal-areas point
- ▶ The **mean** is the balance point
- ▶ The median and the mean are always equal on a symmetric density curve

Ogives

Ogives

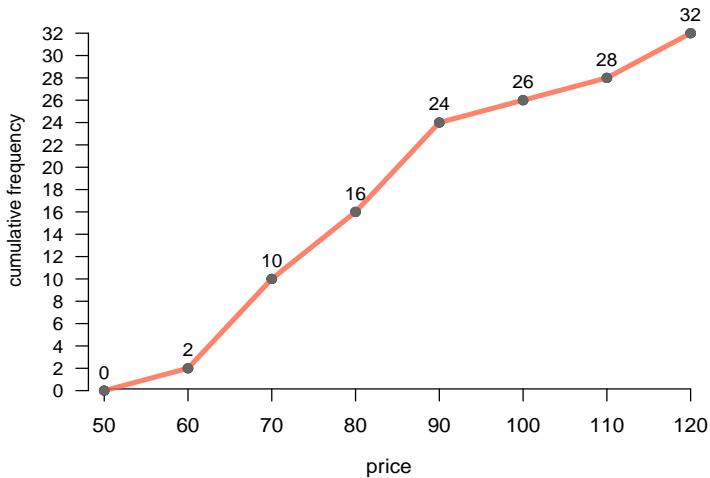
About Ogives

- ▶ Ogives help us examine the cumulative distribution of values in a quantitative variable
- ▶ An ogive tells us how many data are less than the indicated value on the horizontal axis
- ▶ An ogive shows how slowly or rapidly the data values accumulate over the range of the data

Frequency Table NFL Price Tickets

Bin	Interval	Mid-point	Frequency	Cum Freq
1	[50-60)	55	2	2
2	[60-70)	65	8	10
3	[70-80)	75	6	16
4	[80-90)	85	8	24
5	[90-100)	95	2	26
6	[100-110)	105	2	28
7	[110-120)	115	4	32

Ogive



Ogives

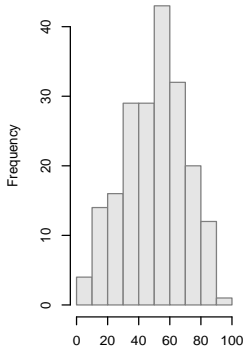
Building an Ogive

- ▶ Make a frequency table showing bin intervals and cumulative frequencies.
- ▶ An ogive begins on the horizontal axis at the lower boundary of the first bin.
- ▶ For each bin, make a dot over the upper interval limit at the height of the cumulative frequency.
- ▶ Connect the dots with line segments.

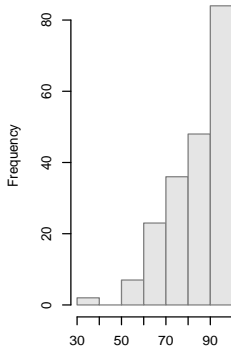
Distributions and Ogives

Three histograms

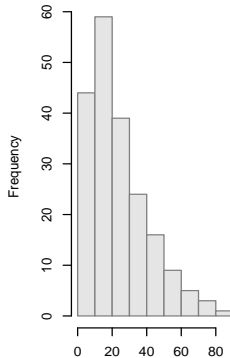
Histogram 1



Histogram 2

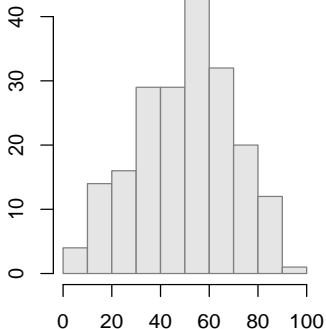


Histogram 3

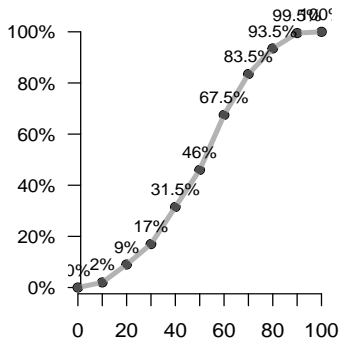


Symmetric Distribution

Histogram 1

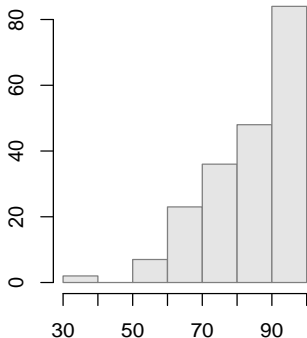


Ogive 1

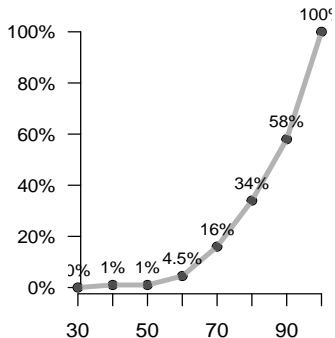


Skewed to the left Distribution

Histogram 2



Ogive 2



Skewed to the right Distribution

