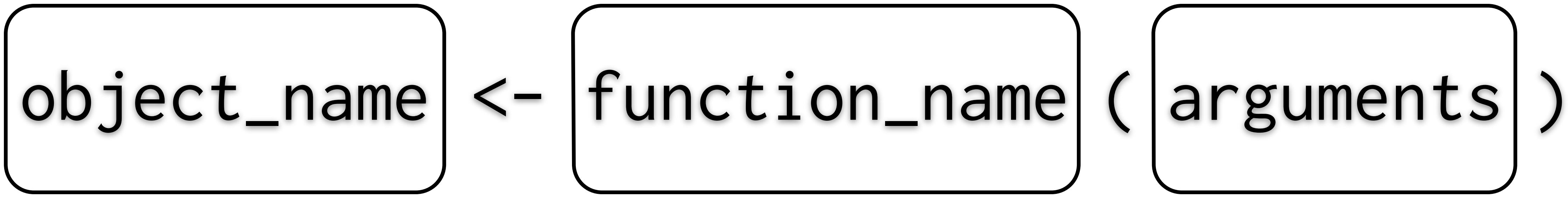
Math 107

Exploratory Data Analysis: One Variable
(Sections 2.1, 2.2, 2.3, 2.4)

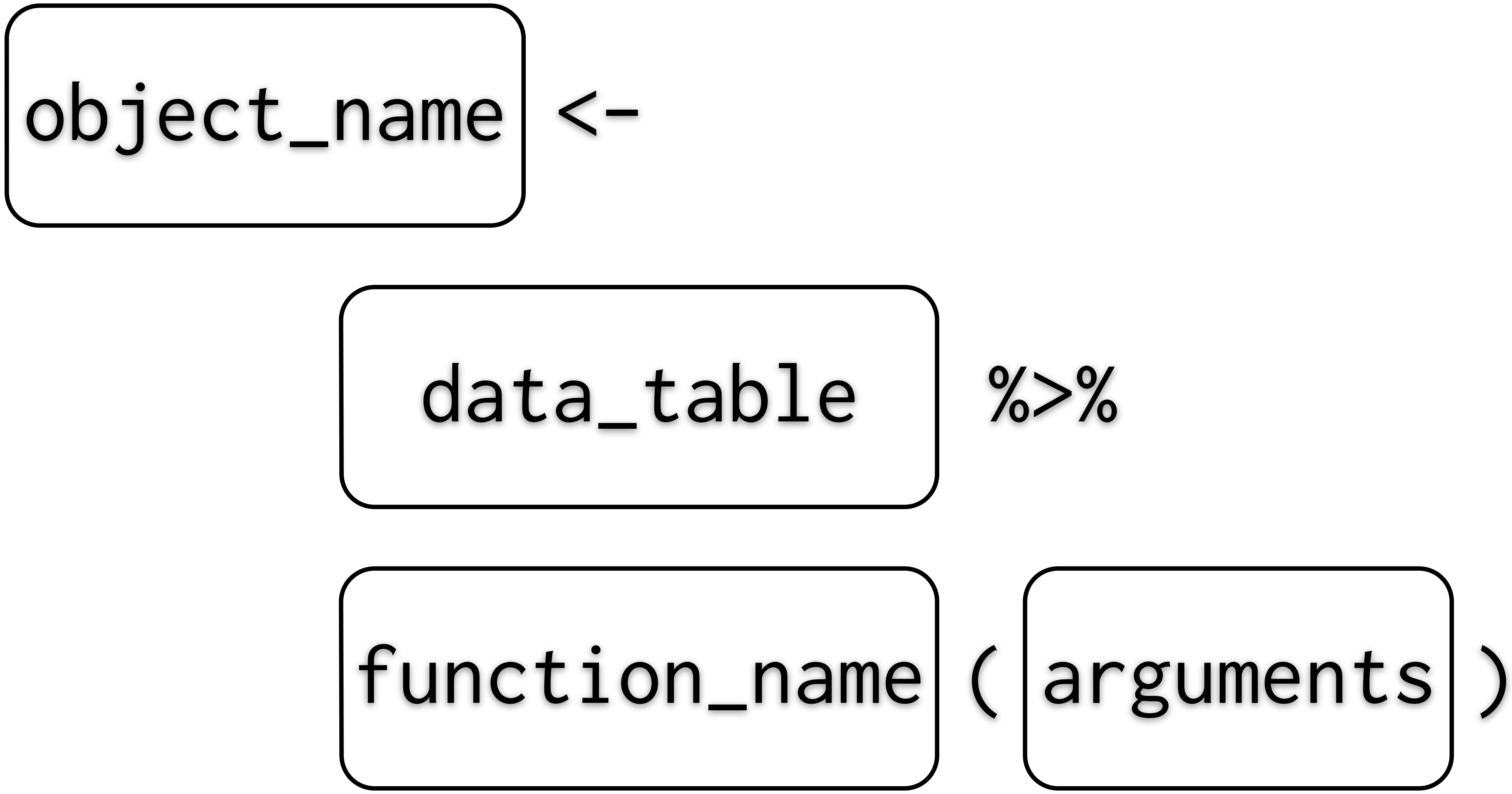
R Command Patterns

Function Application



```
object_name <- function_name ( arguments )
```

Chaining Syntax



```
object_name <-
        data_table %>%
        function_name ( arguments )
```

3 Major Objects

- Functions
- Data tables
- Variables

Basic rules

- The expression to the right of <- is the object you wish to refer to by name
- Function names are always followed immediately by an open parenthesis
- The spot to the left of the first %>% is occupied by a data table
- Arguments are in-between the pair of parentheses following a function name

Loading Tidy Data into R

Plain text

```
# Load in the data
HollywoodMovies2011 <- read.table(file.choose(),
      sep  = “,”, header = TRUE)

# Look at the first few rows (cases)
head(Hollywood2012)
```

Plain text

- `read.file` is our workhorse function
- Different file types require different separators:

Separator	Description
`sep = “ “`	white space separated
`sep = “\t”`	tab separated
`sep = “,”`	comma separated (.csv)

- Specify `header = TRUE` if there are column names

Excel

- If you have an excel file, then `read.table` won't work
- Instead use `read.xlsx`, which is a function in the `xlsx` package

```
# First, load the required packages
library(xlsx)


# Load in a file
HollywoodMovies2012 <- read.xlsx(file.choose(), 1)
```

EDA

Exploratory Data Analysis

In **Exploratory Data Analysis (EDA)** we strive to discover/summarize the main characteristics of a dataset.

We use **summary (descriptive) statistics** and **statistical graphics**.

The type of summary/graphic is determined by the type of variable(s) being analyzed (categorical/ quantitative).

Distribution

The **distribution** of a variable is a description of the values that a variable takes and how often it takes these values.

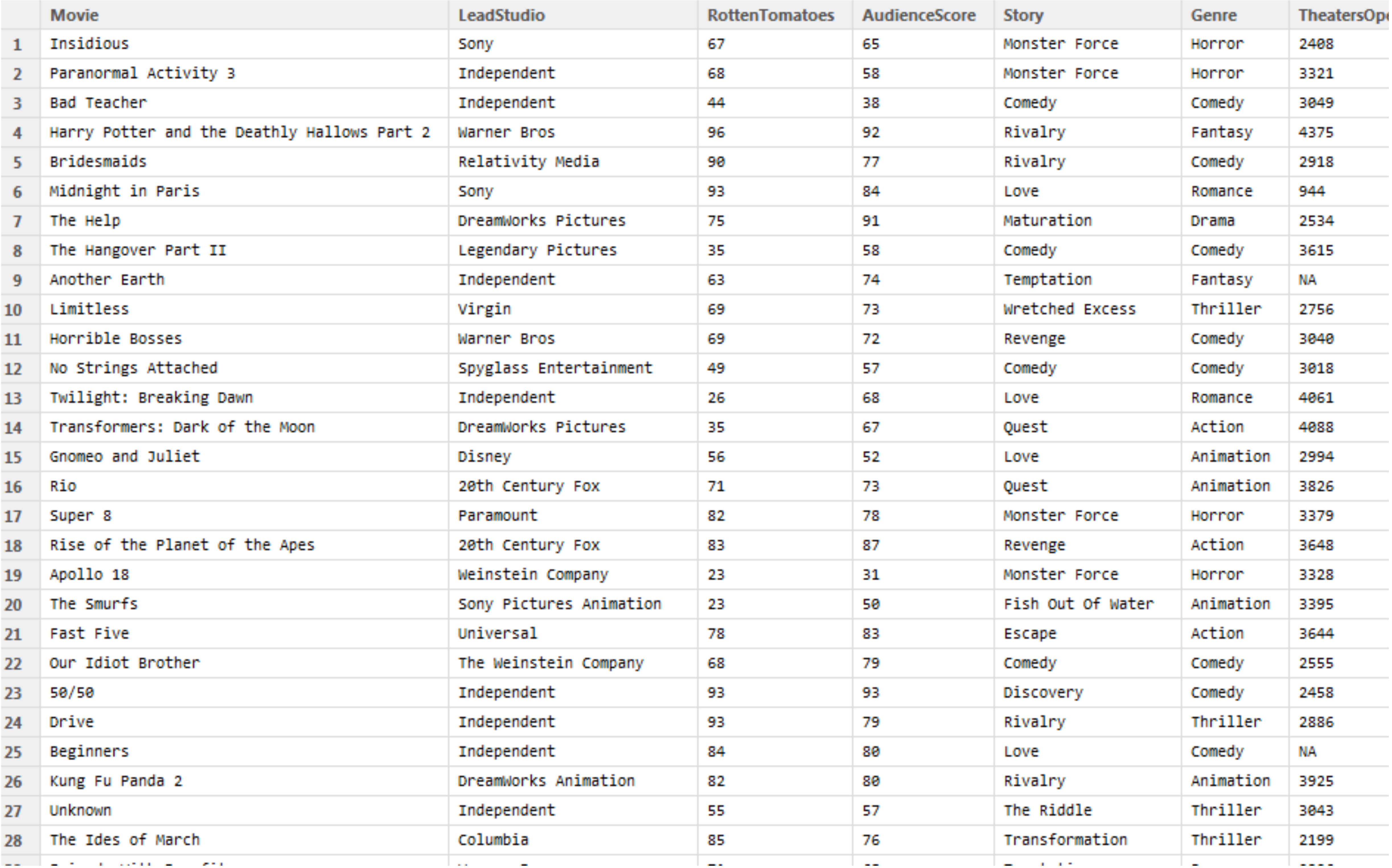
Tabular:

- Frequency table
- Relative frequency table

Graphical:

- Bar chart
- Dot plot
- Histogram
- Density plot
- Boxplot

Hollywood Movies in 2011



	Movie	LeadStudio	RottenTomatoes	AudienceScore	Story	Genre	TheatersOp
1	Insidious	Sony	67	65	Monster Force	Horror	2408
2	Paranormal Activity 3	Independent	68	58	Monster Force	Horror	3321
3	Bad Teacher	Independent	44	38	Comedy	Comedy	3049
4	Harry Potter and the Deathly Hallows Part 2	Warner Bros	96	92	Rivalry	Fantasy	4375
5	Bridesmaids	Relativity Media	90	77	Rivalry	Comedy	2918
6	Midnight in Paris	Sony	93	84	Love	Romance	944
7	The Help	Dreamworks Pictures	75	91	Maturation	Drama	2534
8	The Hangover Part II	Legendary Pictures	35	58	Comedy	Comedy	3615
9	Another Earth	Independent	63	74	Temptation	Fantasy	NA
10	Limitless	Virgin	69	73	Wretched Excess	Thriller	2756
11	Horrible Bosses	Warner Bros	69	72	Revenge	Comedy	3040
12	No Strings Attached	Spyglass Entertainment	49	57	Comedy	Comedy	3018
13	Twilight: Breaking Dawn	Independent	26	68	Love	Romance	4061
14	Transformers: Dark of the Moon	Dreamworks Pictures	35	67	Quest	Action	4088
15	Gnomeo and Juliet	Disney	56	52	Love	Animation	2994
16	Rio	20th Century Fox	71	73	Quest	Animation	3826
17	Super 8	Paramount	82	78	Monster Force	Horror	3379
18	Rise of the Planet of the Apes	20th Century Fox	83	87	Revenge	Action	3648
19	Apollo 18	Weinstein Company	23	31	Monster Force	Horror	3328
20	The Smurfs	Sony Pictures Animation	23	50	Fish Out Of Water	Animation	3395
21	Fast Five	Universal	78	83	Escape	Action	3644
22	Our Idiot Brother	The Weinstein Company	68	79	Comedy	Comedy	2555
23	50/50	Independent	93	93	Discovery	Comedy	2458
24	Drive	Independent	93	79	Rivalry	Thriller	2886
25	Beginners	Independent	84	80	Love	Comedy	NA
26	Kung Fu Panda 2	DreamWorks Animation	82	80	Rivalry	Animation	3925
27	Unknown	Independent	55	57	The Riddle	Thriller	3043
28	The Ides of March	Columbia	85	76	Transformation	Thriller	2199

One Categorical Variable

Frequency Table

A **frequency table** shows how many cases fall into each category.

Type	Frequency
Action	32
Adventure	1
Animation	12
Comedy	27
Drama	21
Fantasy	2
Horror	17
Romance	11
Thriller	13
Total	136

Proportion

The **proportion** in a category is found by

Notation:

for a sample: $\hat{p}$ ("p-hat")

for a population: **p**

Proportion

What proportion of Hollywood movies in 2011 were comedies?

Type	Frequency
Action	32
Adventure	1
Animation	12
Comedy	27
Drama	21
Fantasy	2
Horror	17
Romance	11
Thriller	13
Total	136

Relative Frequency Table

A **relative frequency table** shows the proportion of cases that fall into each category.

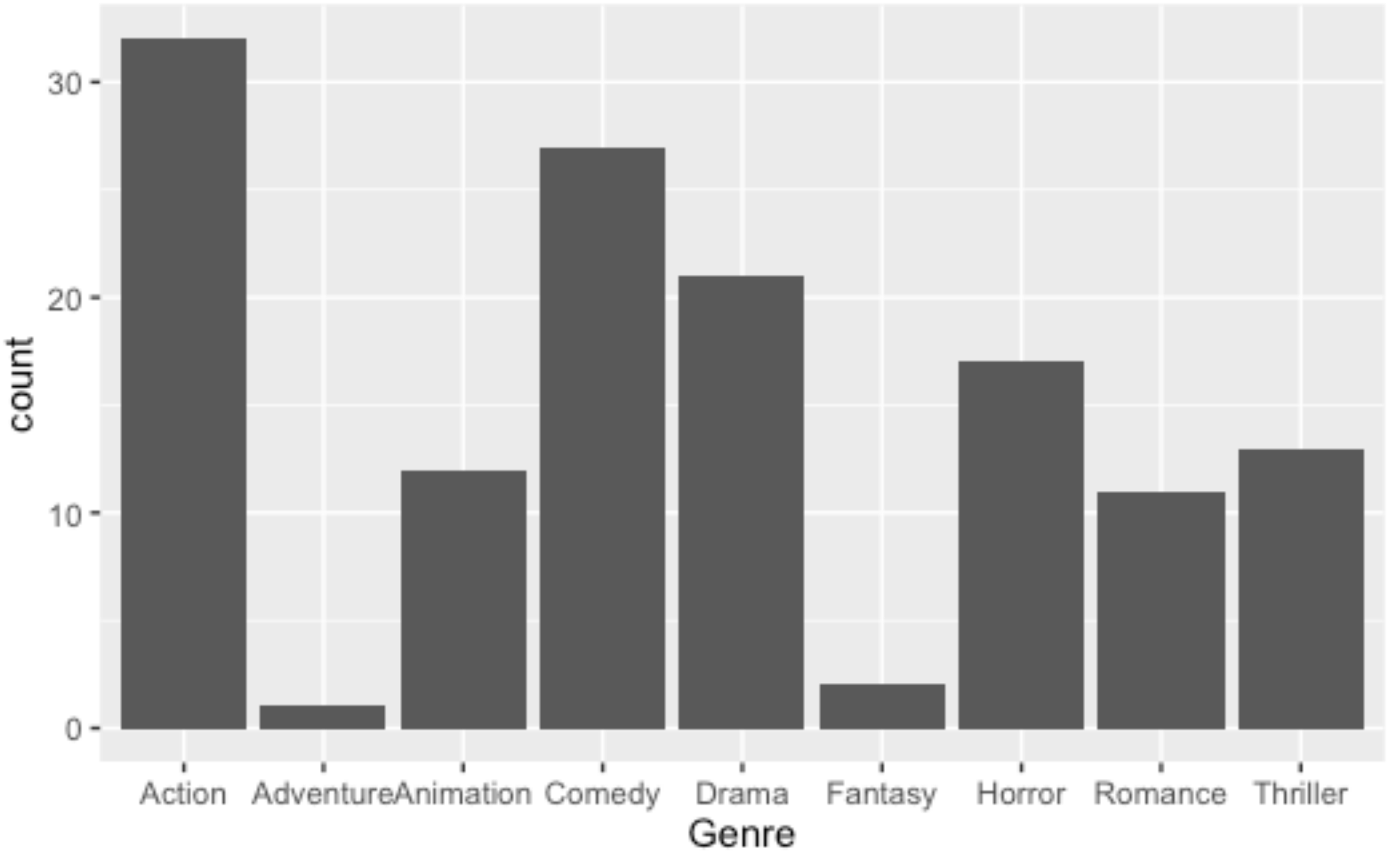
Type	Frequency
Action	0.235
Adventure	0.007
Animation	0.088
Comedy	0.199
Drama	0.154
Fantasy	0.015
Horror	0.125
Romance	0.081
Thriller	0.096
Total	1

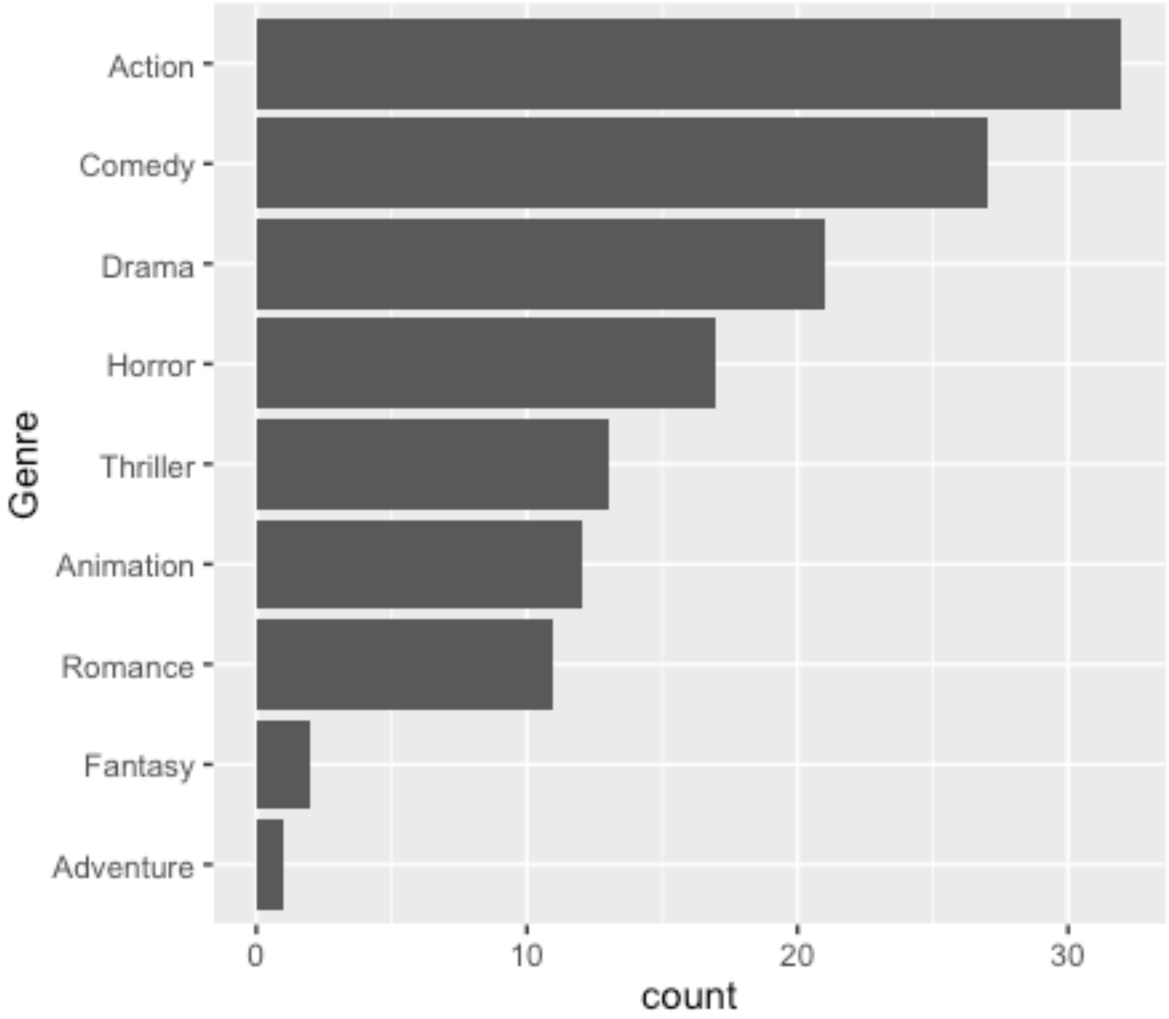
Tables in R

```
# First, load the required packages
library(dplyr)  # data wrangling and summarization extension

# calculate a frequency table
HollywoodMovies2011 %>%
  group_by(Genre) %>%
  summarise(count = n())

# calculate a relative frequency table
HollywoodMovies2011 %>%
  group_by(Genre) %>%
  summarise(count = n()) %>%
  mutate(freq = count / sum(count))
```

Bar Chart

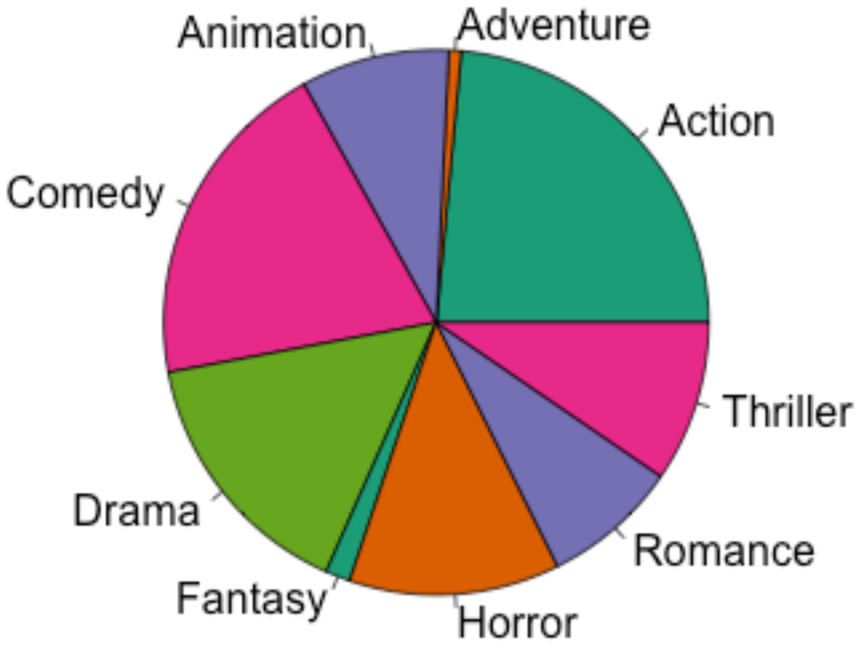
In a **bar chart**, the height/length of the bar is the number (or proportion) of cases falling into each category.





Pie Chart

In a **pie chart**, each category is displayed as a piece of a circle whose area is proportional to the proportion of cases in that category.



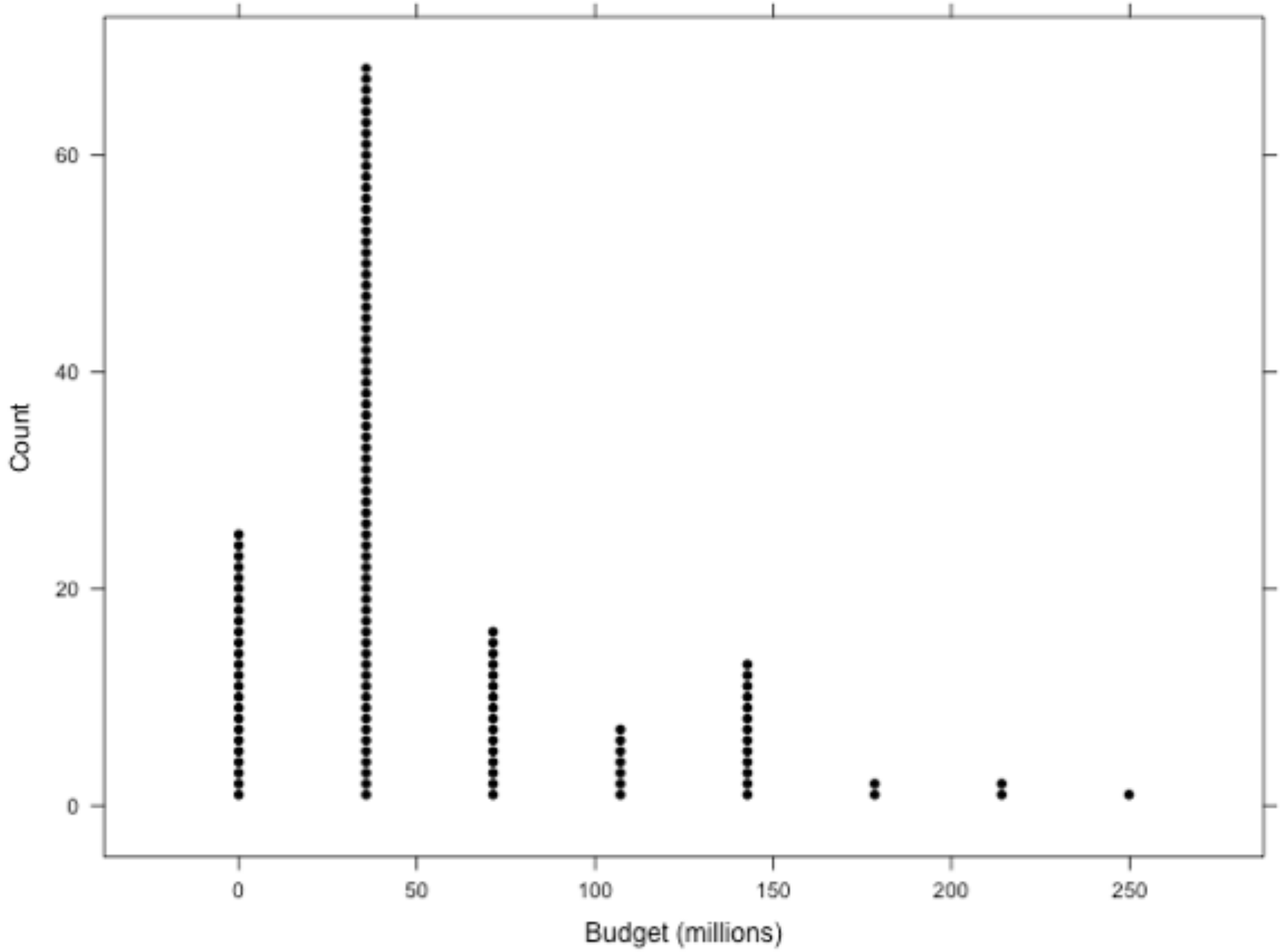
Plots in R

```
# First, load the required packages

library(ggplot2) # plotting extension


# Draw a bar chart

ggplot(data = HollywoodMovies2011) +

  geom_bar(mapping = aes(x = Genre))


# You can get fancier, but the commands become more

# complex

ggplot(data = HollywoodMovies2011) +
  geom_bar(mapping = aes(x = reorder(Genre, Genre, length))) +
  xlab("Genre") +
  coord_flip()
```

One Quantitative Variable

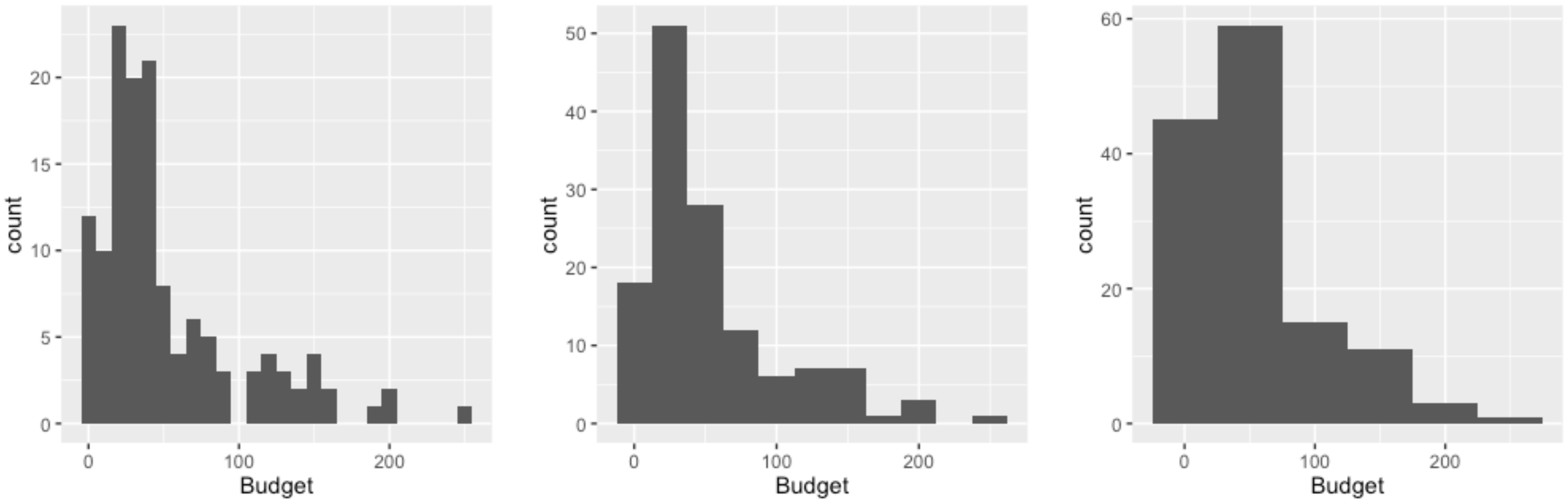
Dotplot

In a **dotplot**, each case is represented by a dot, and the dots are stacked above the corresponding values on the number line.



Histogram

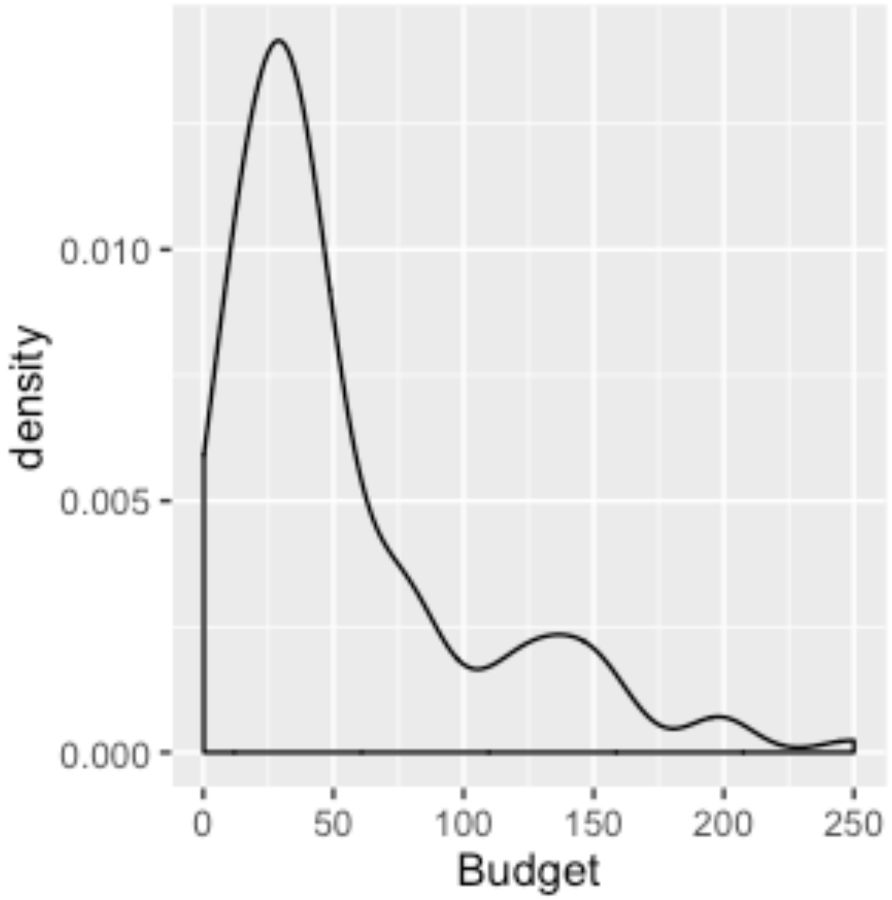
In a **histogram** the height of each bar is proportional to the number of cases within each **bin**.

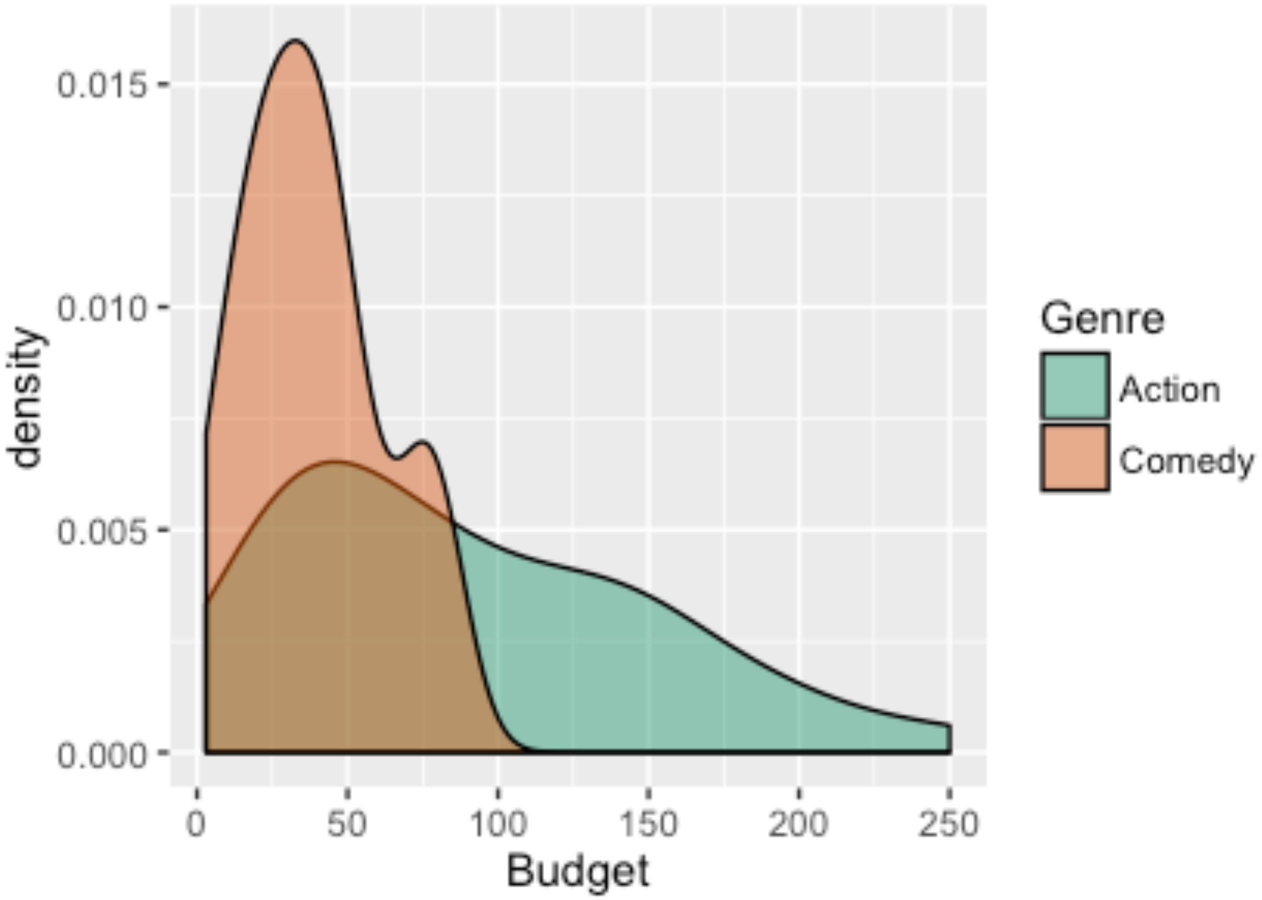


```r
# Make sure that the data are loaded and so is
# ggplot2

# drawing a histogram
ggplot(data = HollywoodMovies2011) +
  geom_histogram(mapping = aes(x = Budget), binwidth = 10)

ggplot(data = HollywoodMovies2011) +
  geom_histogram(mapping = aes(x = Budget), binwidth = 25)

ggplot(data = HollywoodMovies2011) +
  geom_histogram(mapping = aes(x = Budget), binwidth = 50)
```

Density Plots

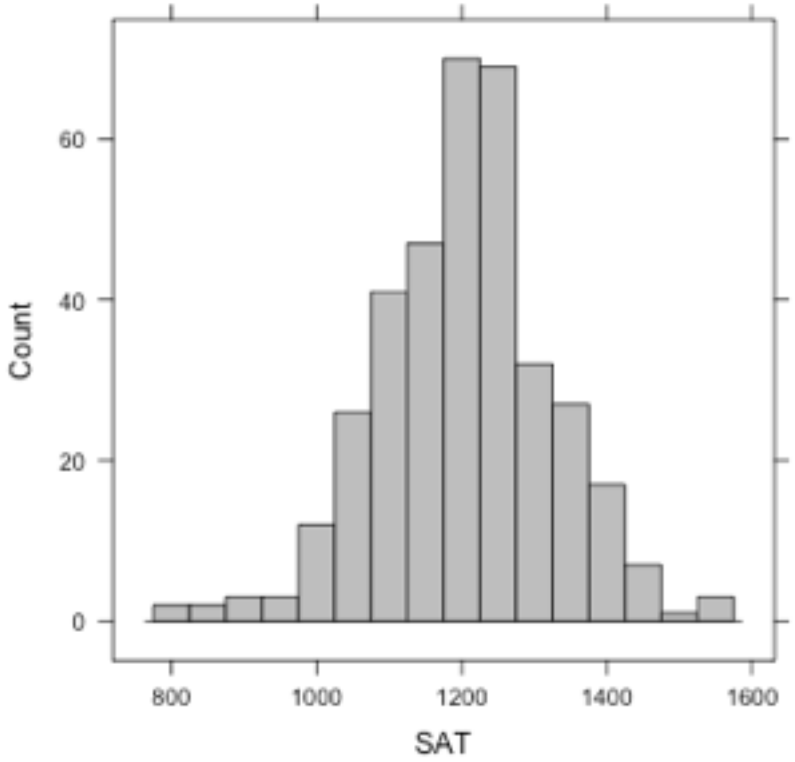
Density plots are similar to histograms, but use a smooth curve where the area is proportional to the frequency.

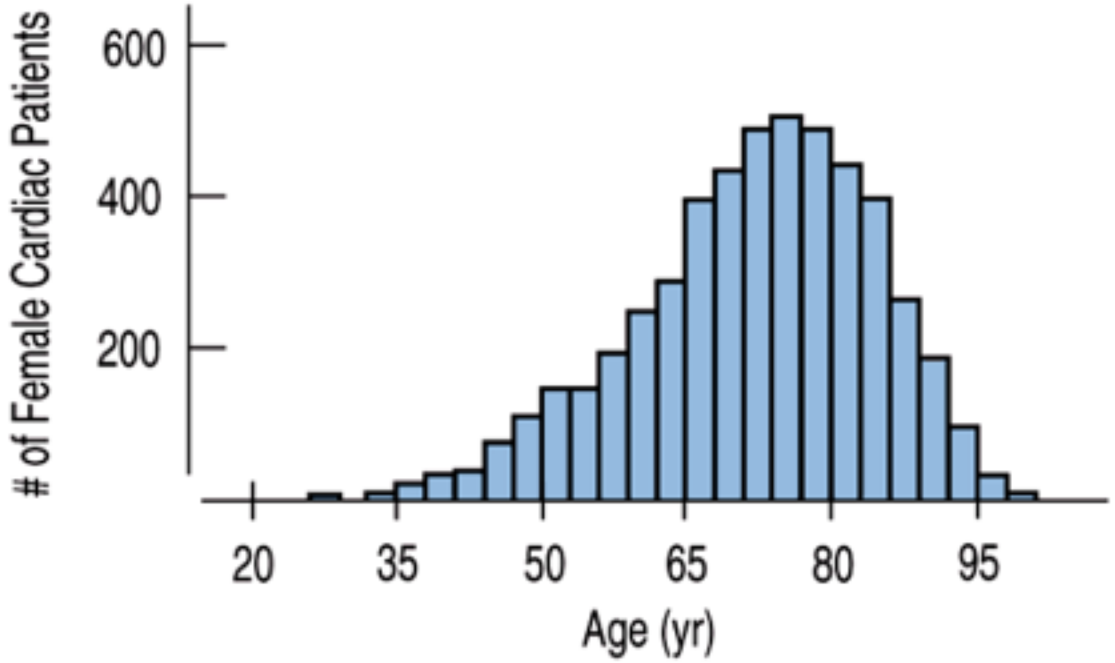


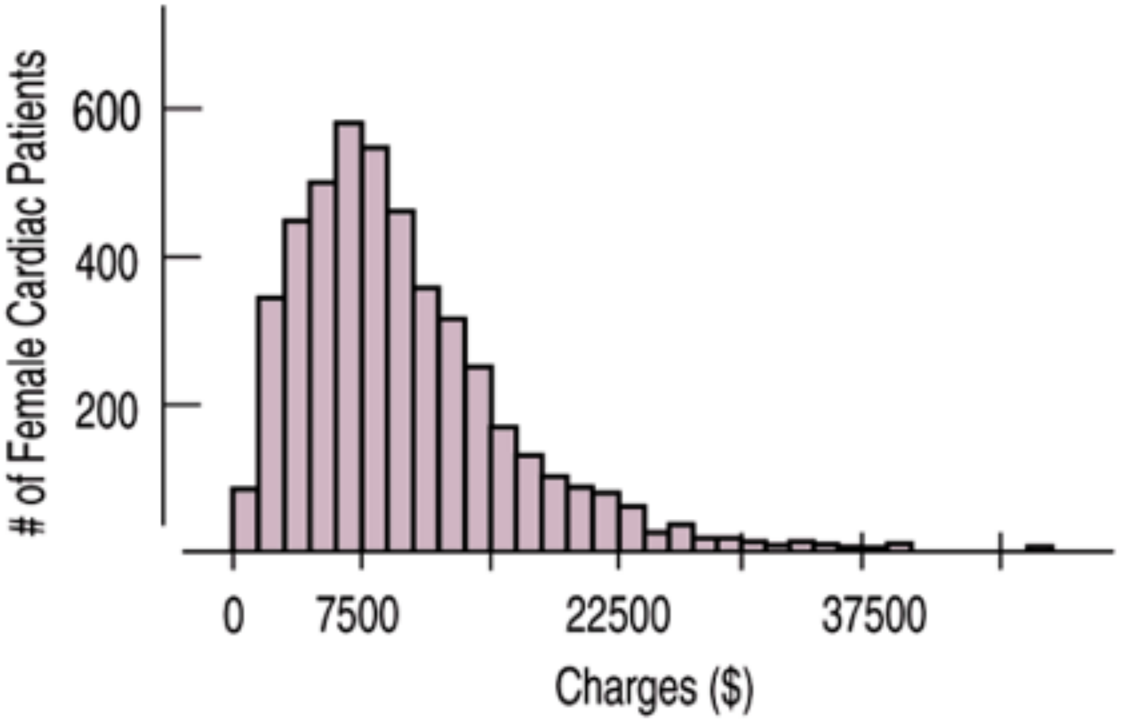


```
# Make sure that the data are loaded and so is ggplot2


# drawing a density plot

ggplot(data = HollywoodMovies2011) +

  geom_density(mapping = aes(x = Budget))


# Overlaying density plots

ggplot(data = HollywoodMovies2011) +

  geom_density(mapping = aes(x = Budget, fill = Genre))
```

Describing the Shape







Measures of Center

Notation

n = sample size

We often let **x** (or y) denote a variable, and x_1, x_2,..., x_3, represent the n values of the variable x.

Mean

The **mean** is the arithmetic average of all the data values.

Notation:

Sample value: $\overline{x}$

Population value: **μ (“mu”)**

In R: `mean(x)`

Median

The **median** is the middle value

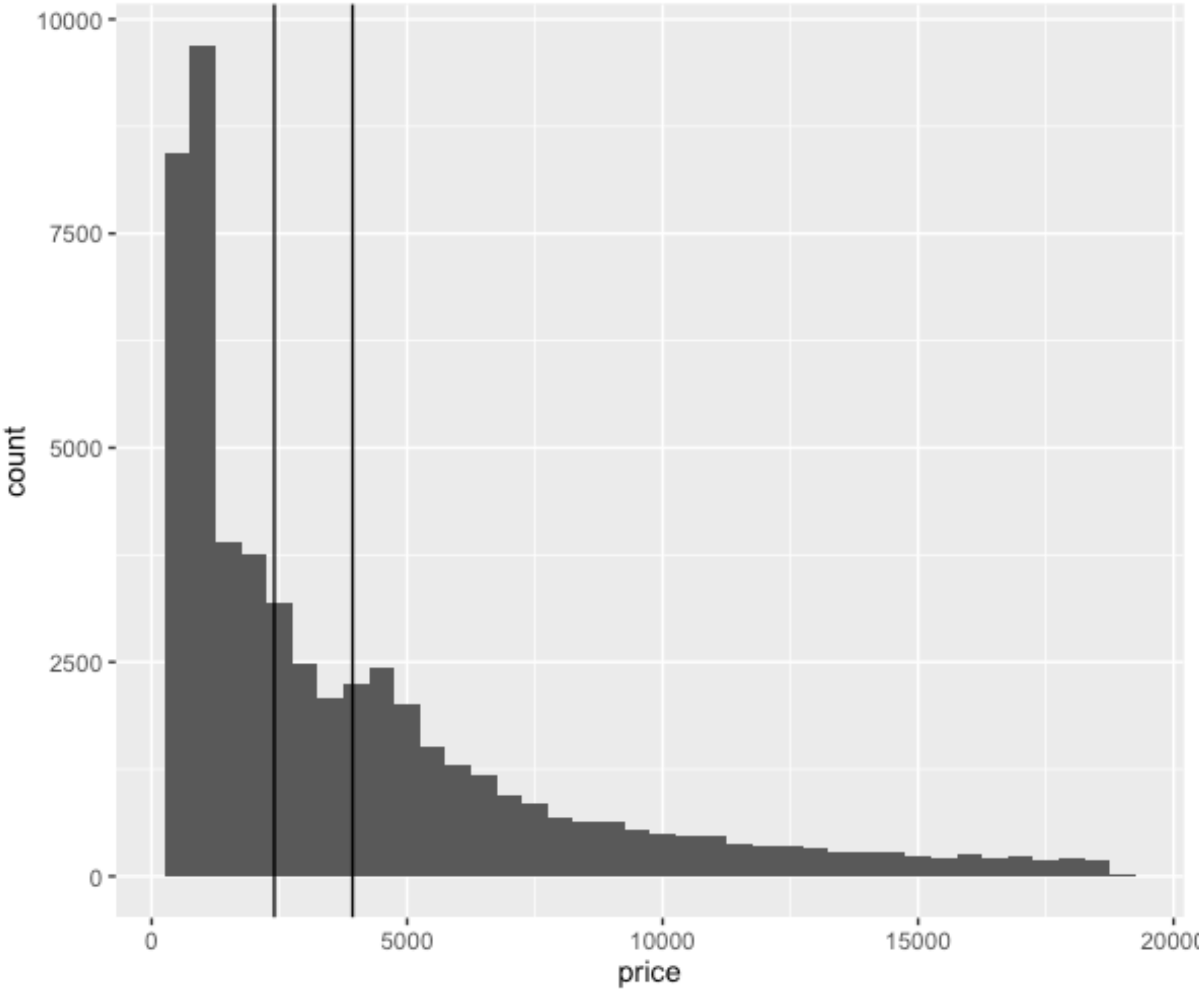
- If there are an odd number of data values, this will be the middle number.
- If there are an even number of data values, this will be the average of the middle two values.
- Denoted by **m**

In R: `median(x)`

```r
# mean and median of a column of a dataset

mean(HollywoodMovies2011$WorldGross, na.rm = TRUE)

median(HollywoodMovies2011$WorldGross, na.rm = TRUE)
```

Skewness

The mean is pulled in the direction of the skew.

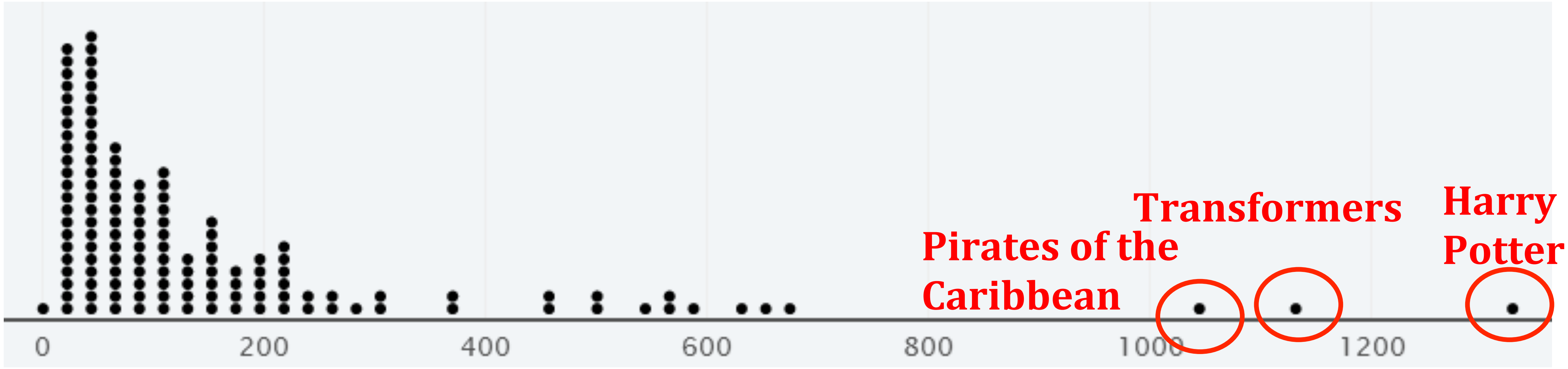


Your Turn

A distribution is skewed to the left. Which measure of center would you expect to be bigger?

Outliers

An **outlier** is an observation that is notably different from the other values in the dataset.



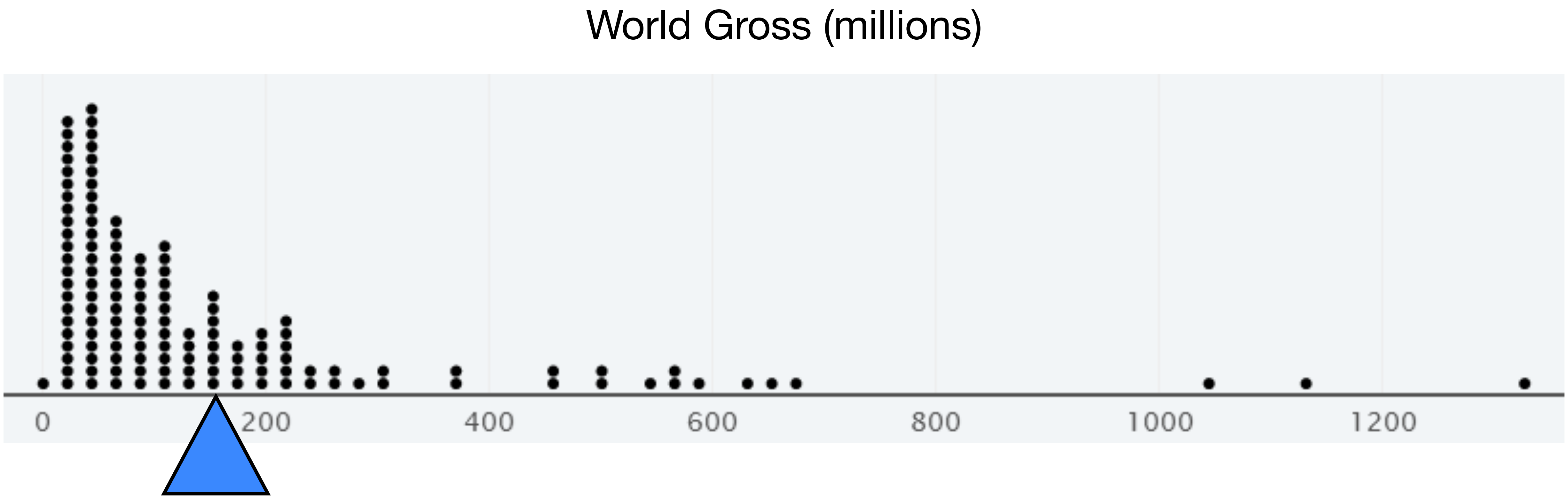
World Gross (millions)

Resistance

A statistic is **resistant/robust** if it is relatively unaffected by extreme values.

	Mean	Median
With Harry Potter	$150,742,300	$76,658,500
Without Harry Potter	$141,889,900	$75,009,000

Resistance



Outliers

When using statistics that are not resistant to outliers:

- Check whether outlier is an error
- If it’s not, is the outlier part of the target population?
- If so, run the analysis with and without the outlier. How much does the outlier influence the results? Report both analyses.