USC School of Business

Data Driven Decision Making: Descriptive Statistics

GSBA 545, Fall 2021

Professor Dawn Porter

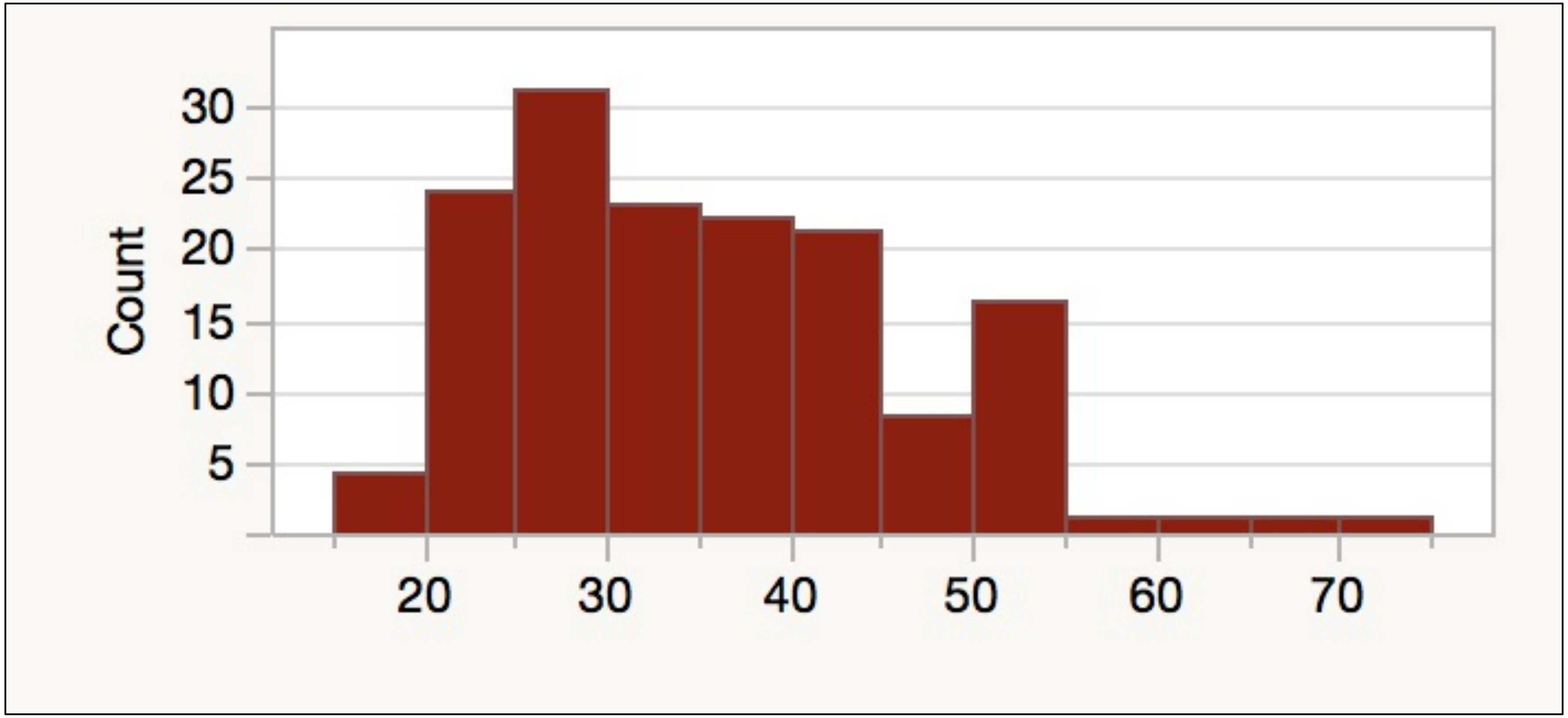
Descriptive Statistics

- Basic Terminology & Scales of Measurement

- Numerical Measures
  - Central Tendency
  - Dispersion

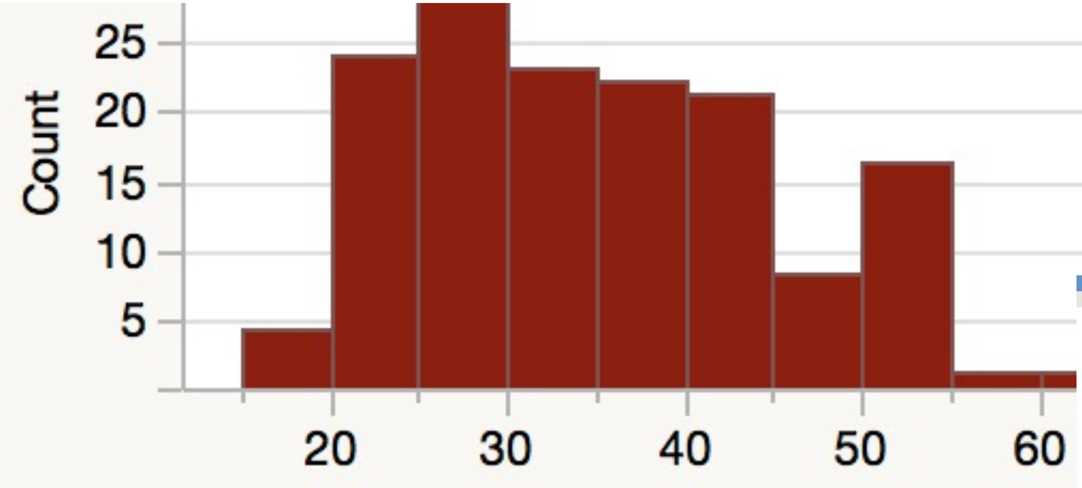
- Graphical Methods
  - Histograms
  - Box-and-Whisker plots
  - Bar Charts & Pie Charts
  - Scatterplots

MPG for 153 Hybrid Cars



Hybrid Car data: Population or Sample?

MPG for 153 Hybrid Cars



Population

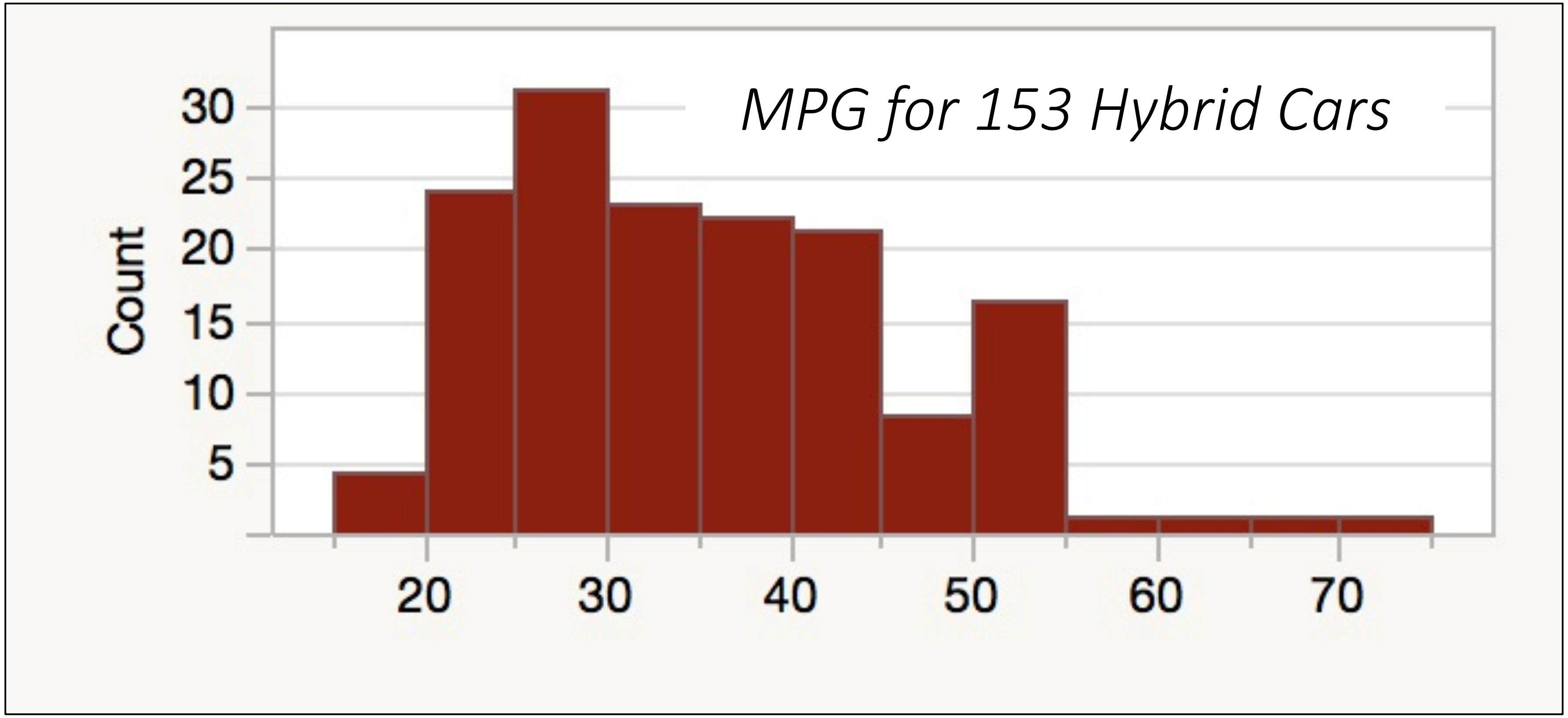
Sample

Total Results: 0

Powered by Poll Everywhere

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app

Population



Population

MPG for 153 Hybrid Cars

Population: Set of all items of interest in a statistical problem.

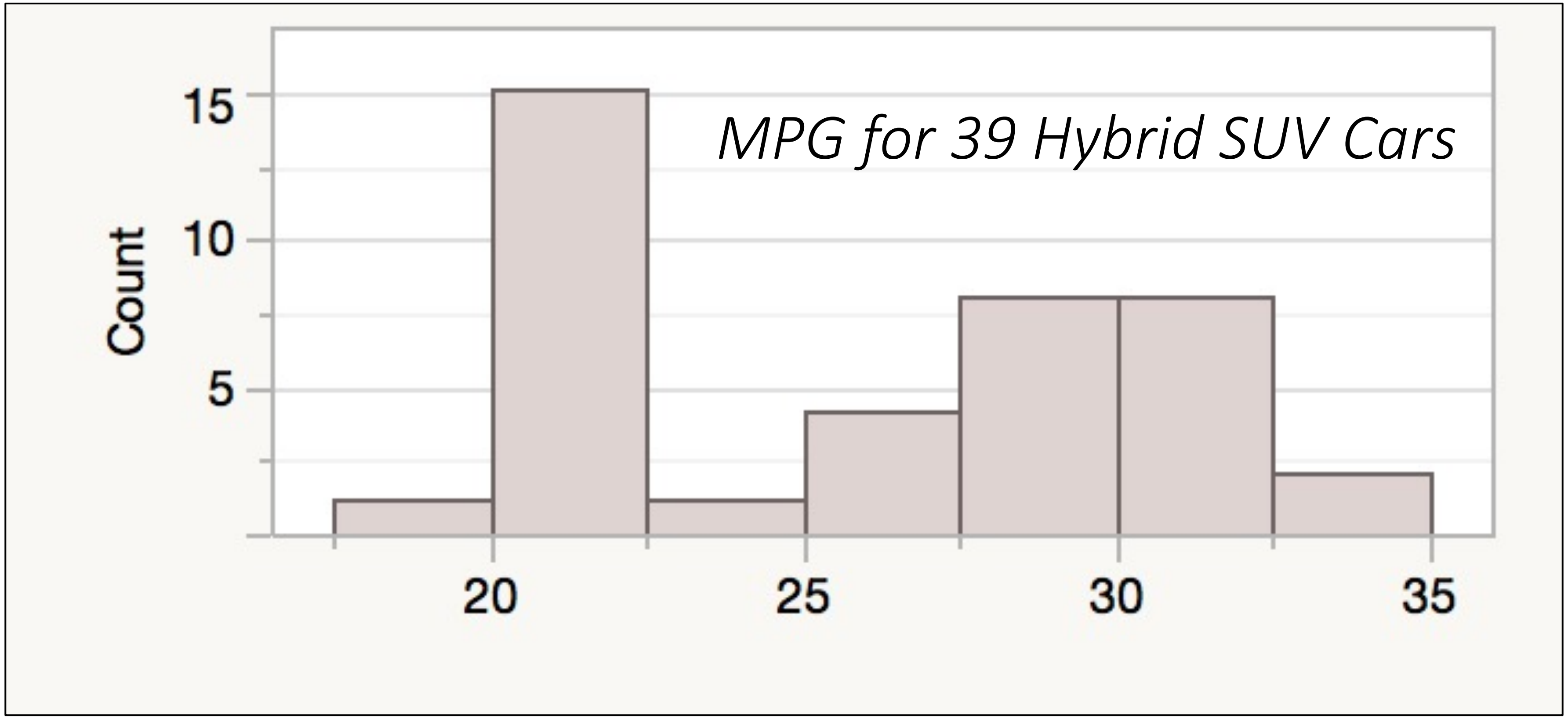
Parameter: Descriptive measure of population

- N = population size
- μ = population average
- σ = population standard deviation

Population: 153 Hybrid Cars

- N = 153
- μ = mean = average = 34.80 mpg
- σ = standard deviation = typical fluctuation = 10.97 mpg

Sample



MPG for 39 Hybrid SUV Cars

Sample: Set of data drawn from the population

Statistic: Descriptive measure of sample

- n = sample size
- $\bar{x}$ = sample average
- s = sample standard deviation

Sample: 39 Hybrid SUV Cars

- n = 39
- $\bar{x}$ = mean = average = 26.01 mpg
- s = standard deviation = typical fluctuation = 4.60 mpg

Data Types & Graphs

Numerical (quantitative)

- Natural measurement system
- Ratios and comparisons make sense

→

Histograms

Boxplots

Scatterplots

Categorical (qualitative)

- Nominal: no inherent ordering
- Ordinal: ordered, but distance between classes may vary

→

Bar Charts

Pie Charts

Side-by-side Boxplots

USC School of Business

Scales of Measurement

Discrete: Possible number of values is countable

- Number of Hybrid SUV Cars
- Number of Comedy films released in 2017
- Number of games in any given World Series

Continuous: Possible number of values is relatively infinite

- MPG of Hybrid Cars
- Height, weight, distance

Cross-sectional: Snapshot of data at a specific point in time

- Economic indicators for several countries in 2019

Time Series: Result of tracking one or more variables over time

- Economic indicators for only the US from 1900-2019

Distribution Shapes

Histograms and boxplots help uncover distribution shape:

Symmetrical (roughly equal tails)

- Bell-Shaped Distribution.

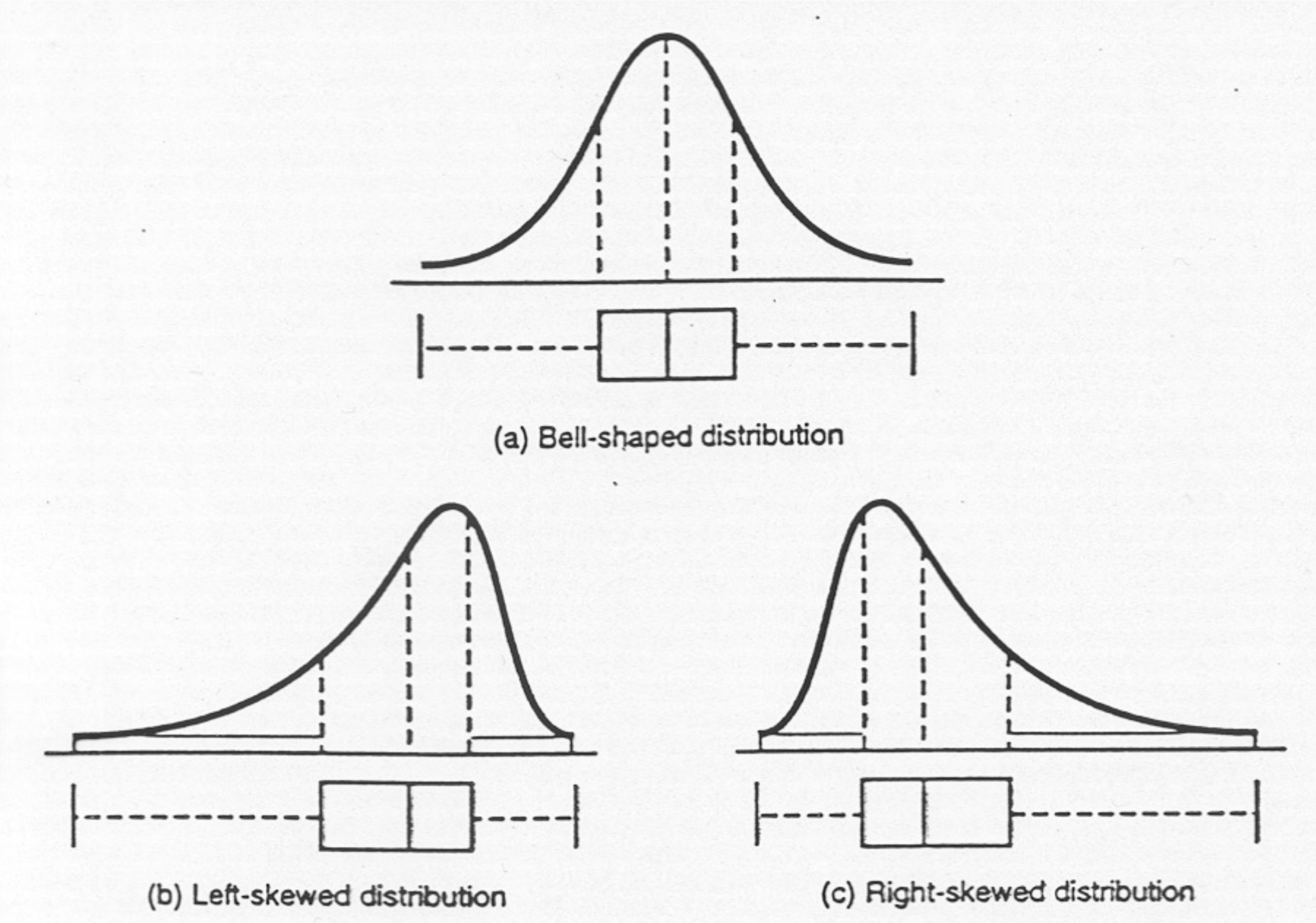
Positively Skewed – skewed right (long tail on right)

- Income Distributions.

Negatively Skewed – skewed left (long tail on left)

- Scores on an easy exam.

Distribution Shapes

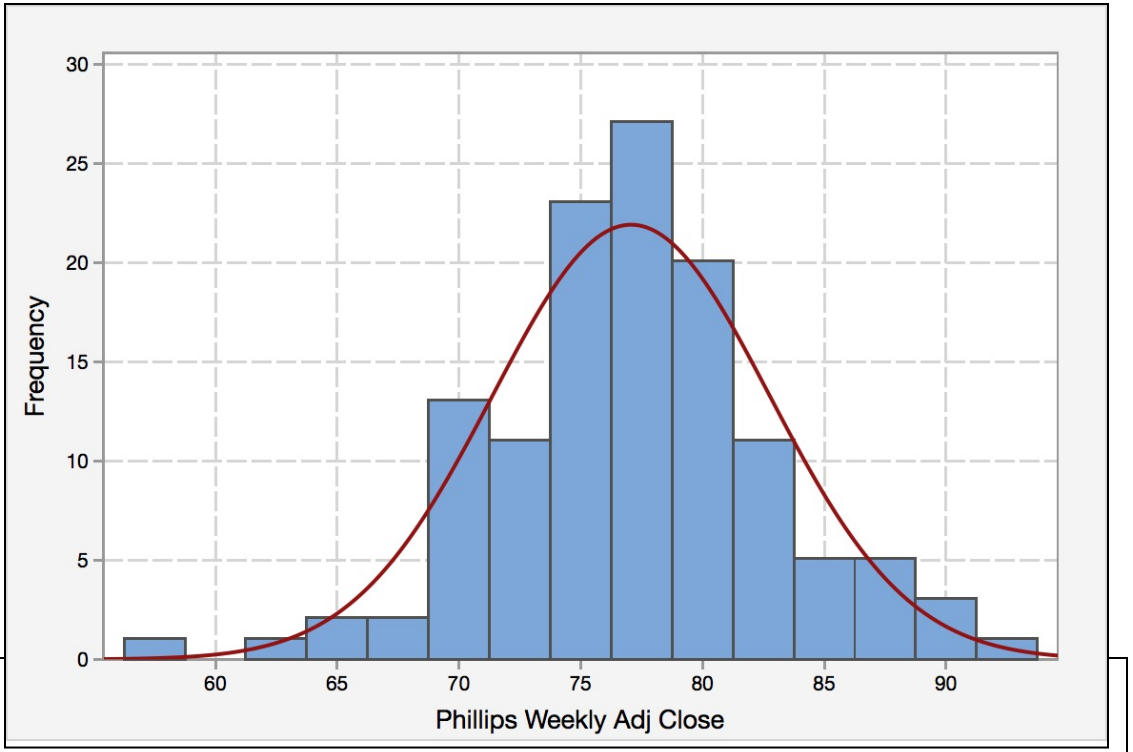


USC School of Business

Symmetric Distribution

Phillips Stock Prices:*

Mean ≈ Median and Median is *somewhat* close to being about halfway between 25th and 75th percentiles.



Statistics

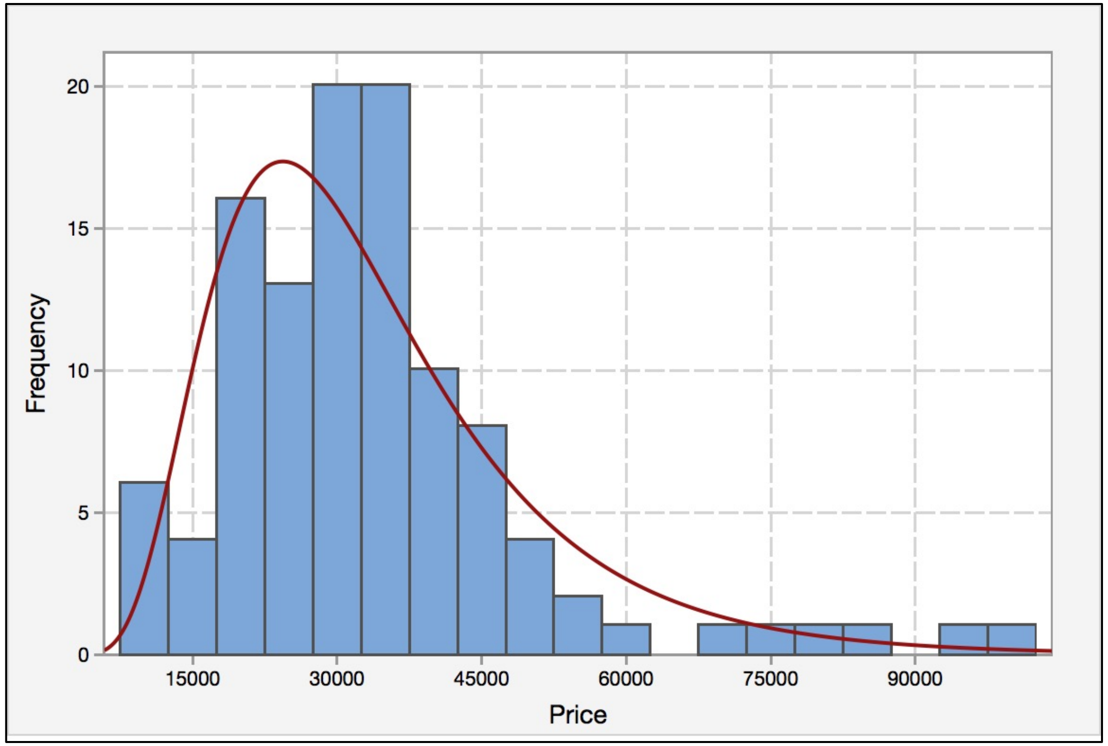
Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Adj Close	125	77.0677	5.6917	58.3264	73.9521	77.2600	79.8241	91.3296

*Weekly closing prices, 1/6/14 – 5/23/16

Right-Skewed Distribution

LA Used Car Prices:

Mean > Median and Mean is closer to the 75th percentile than to the 25th percentile.



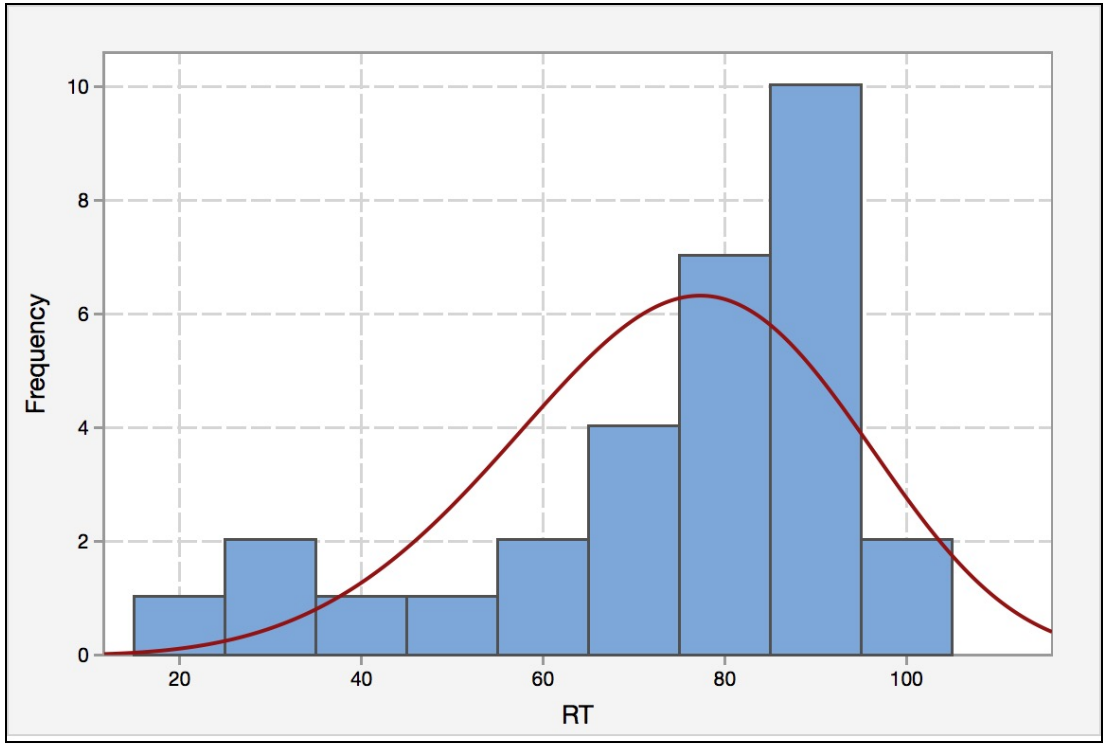
Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Price	110	33598	16314	9950	22991	31210	39220	99999

Left-Skewed Distribution

Top Movies in China, Rotten Tomatoes Score:

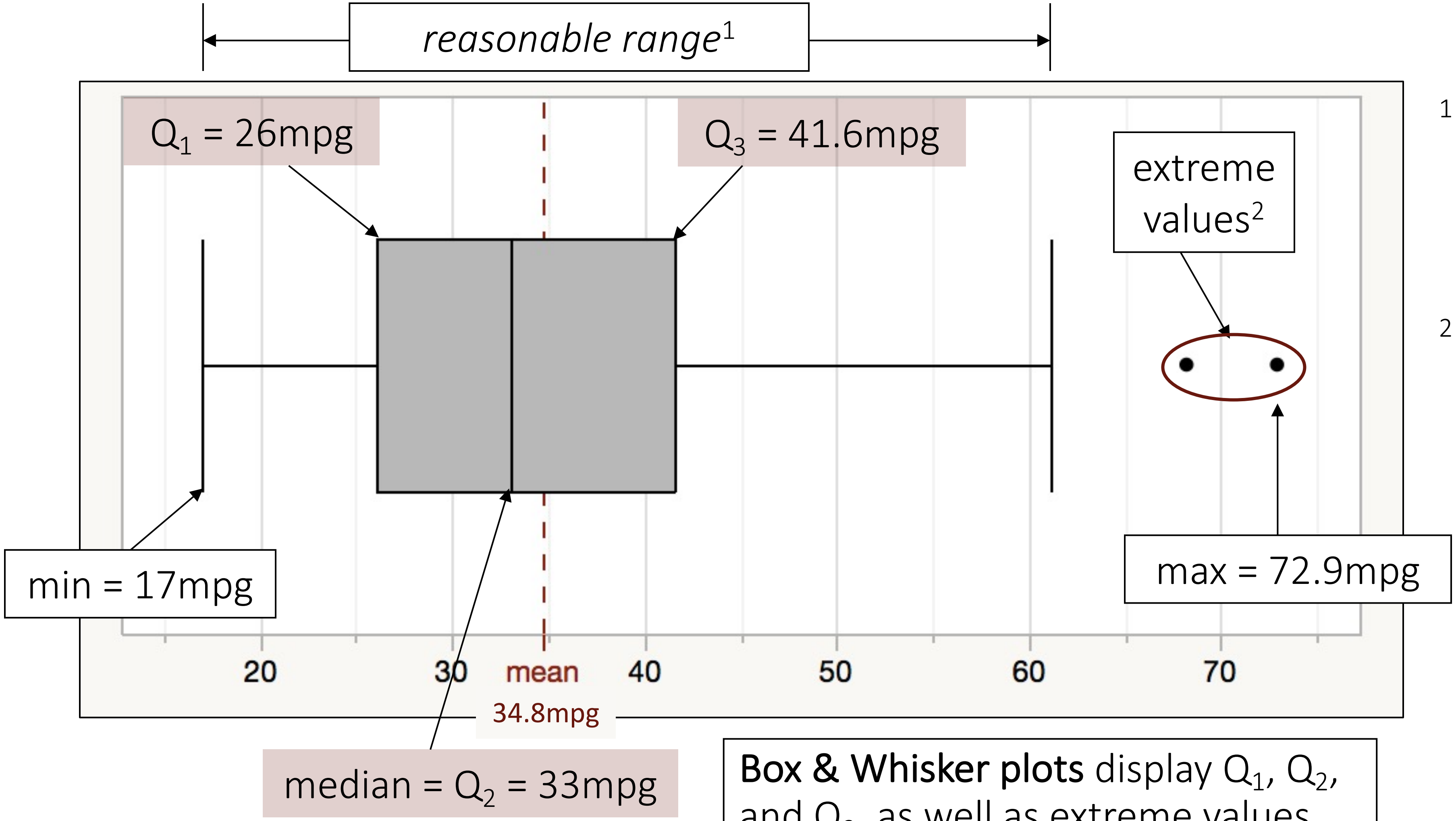
Mean < Median and Median is closer to 75^{th} than to 25^{th} percentile.



Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
RT	30	74.433	21.716	18.000	68.500	81.000	91.000	98.000

Box & Whisker Plot



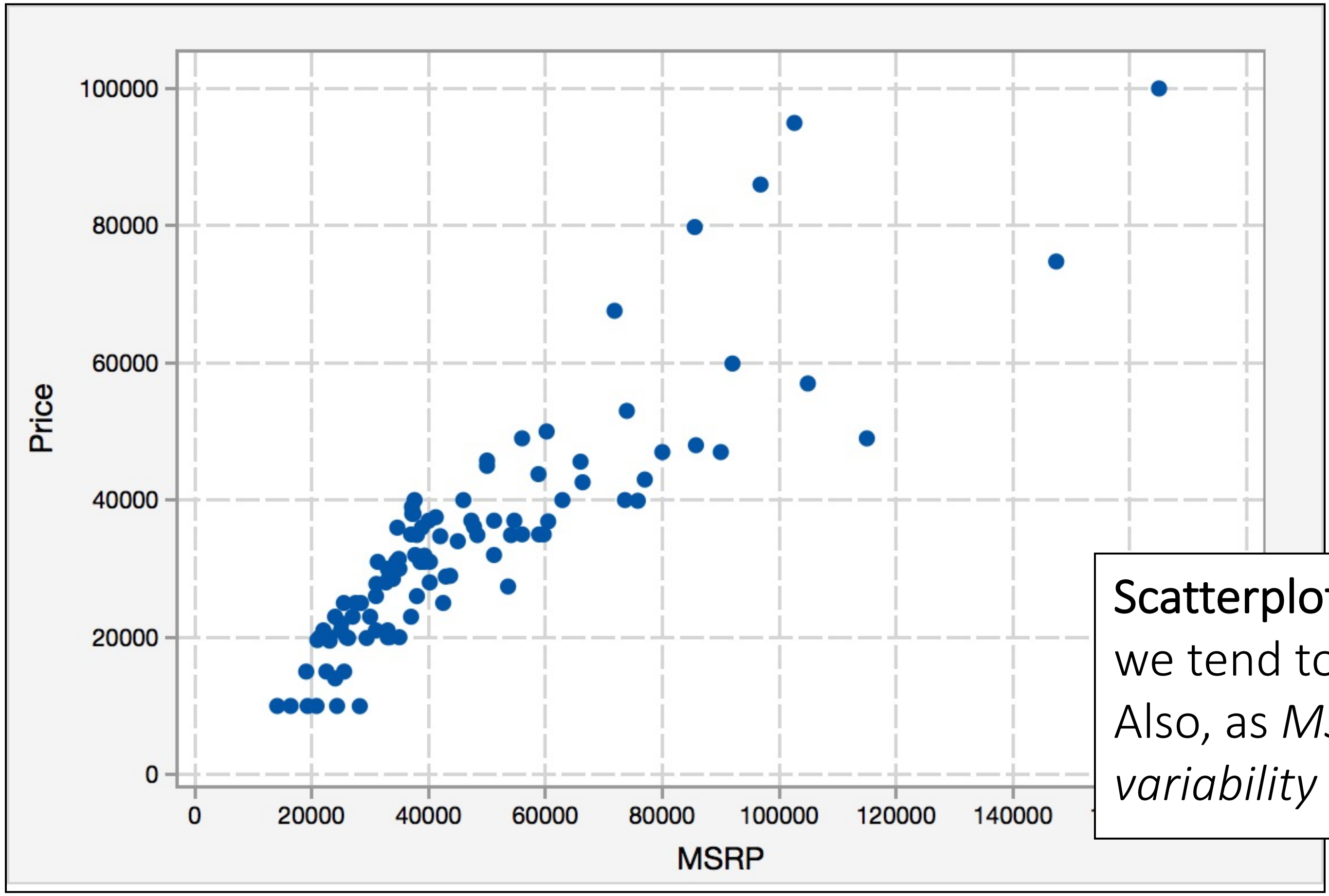
[1] Bounds of the *reasonable range* are:

Median $\pm$ 1.5 IQR

[2] *Extreme values* are defined as being at least 3 IQRs from the median.

Box & Whisker plots display Q_1, Q_2, and Q_3, as well as extreme values.

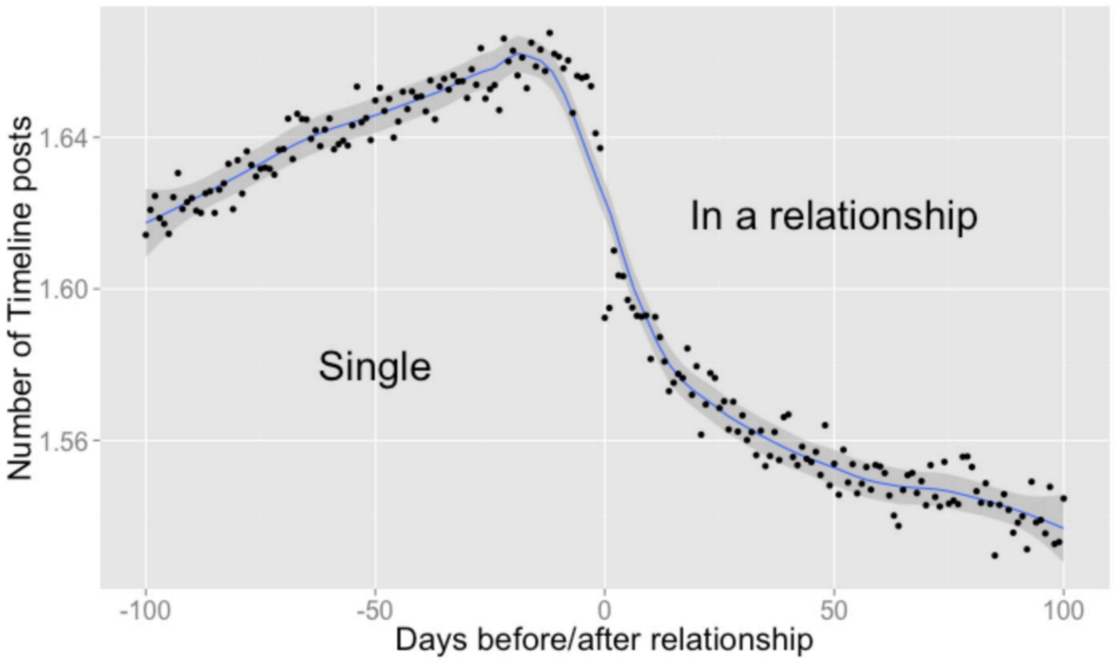
Scatterplot: Example



Scatterplot shows that, as *MSRP* increases, we tend to see higher prices for used cars. Also, as *MSRP* increases, there is more *variability* in the prices of used cars.

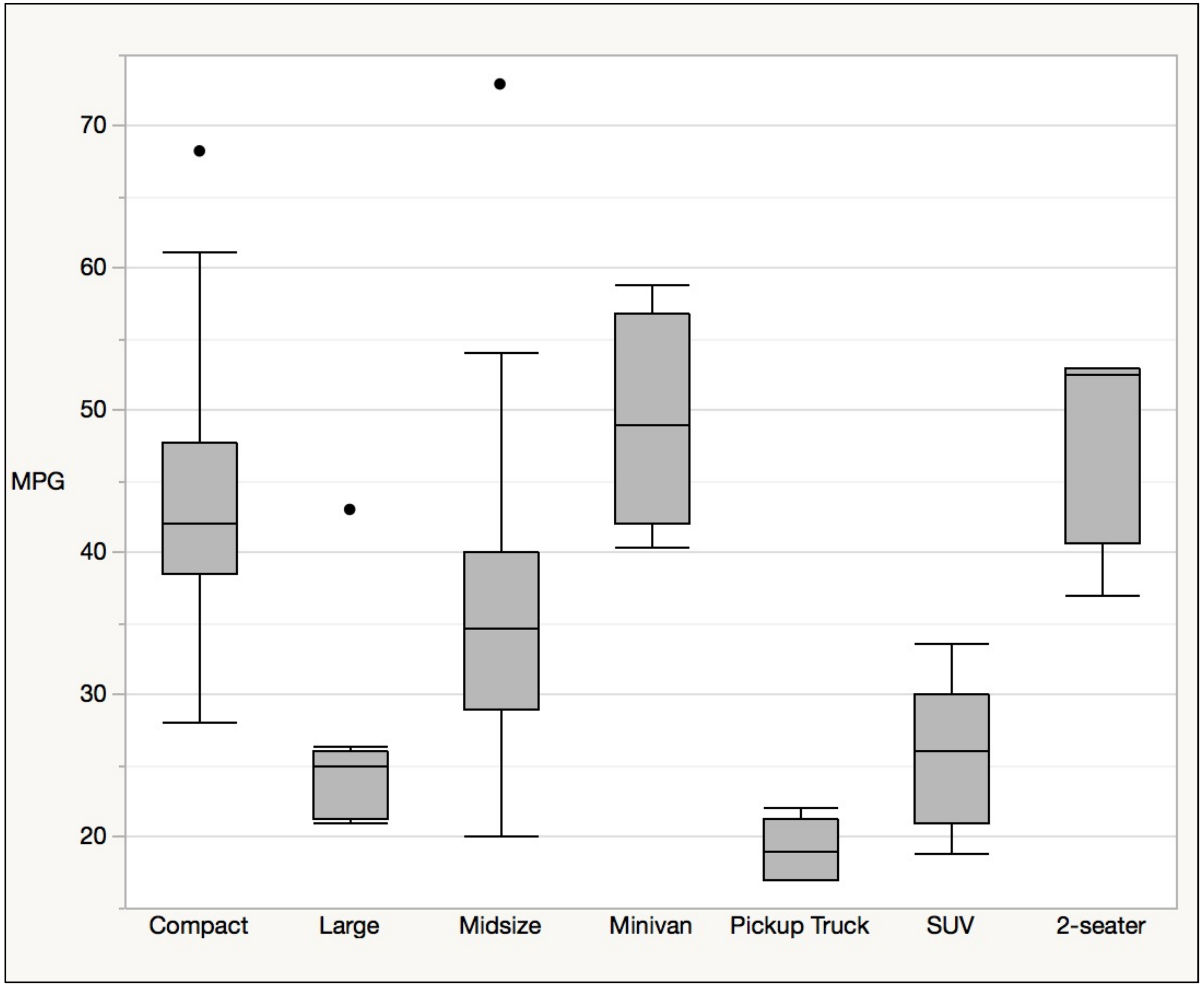
Scatterplot: Example

Facebook: In the 100 or so days before you're likely to start a relationship with someone, the number of interactions between users is expected to rise consistently. Then, right before the relationship begins, there's a free-fall in the number of timeline posts. After the relationship is established, the freefall is followed by a steady decline in the number of wall posts.



*Chartporn.org

Side-by-Side Boxplots

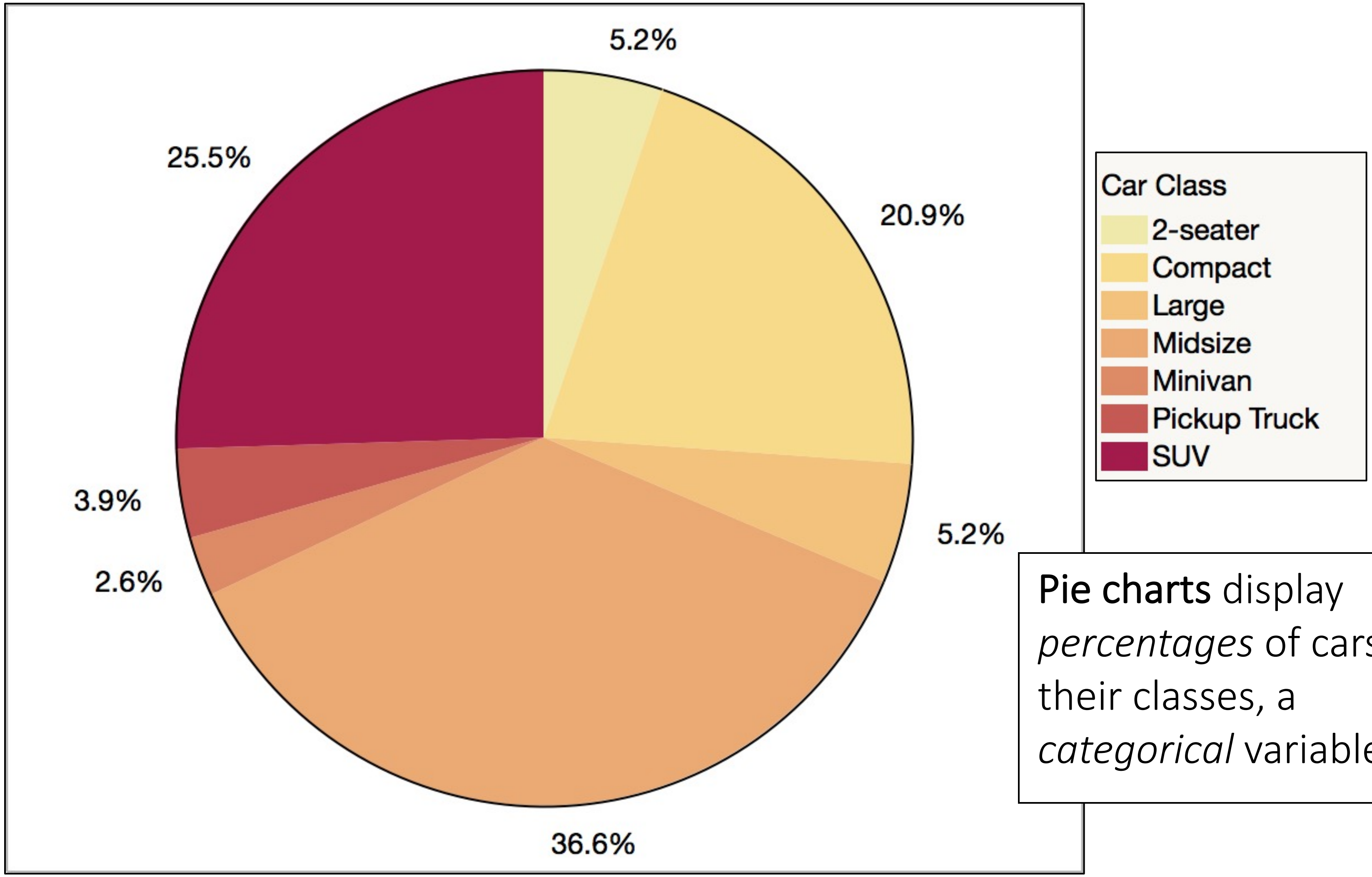


Hybrid Car MPG:

Use side by side boxplots to display relationships between categorical variables (car type) and a quantitative variable (MPG).

USC School of Business

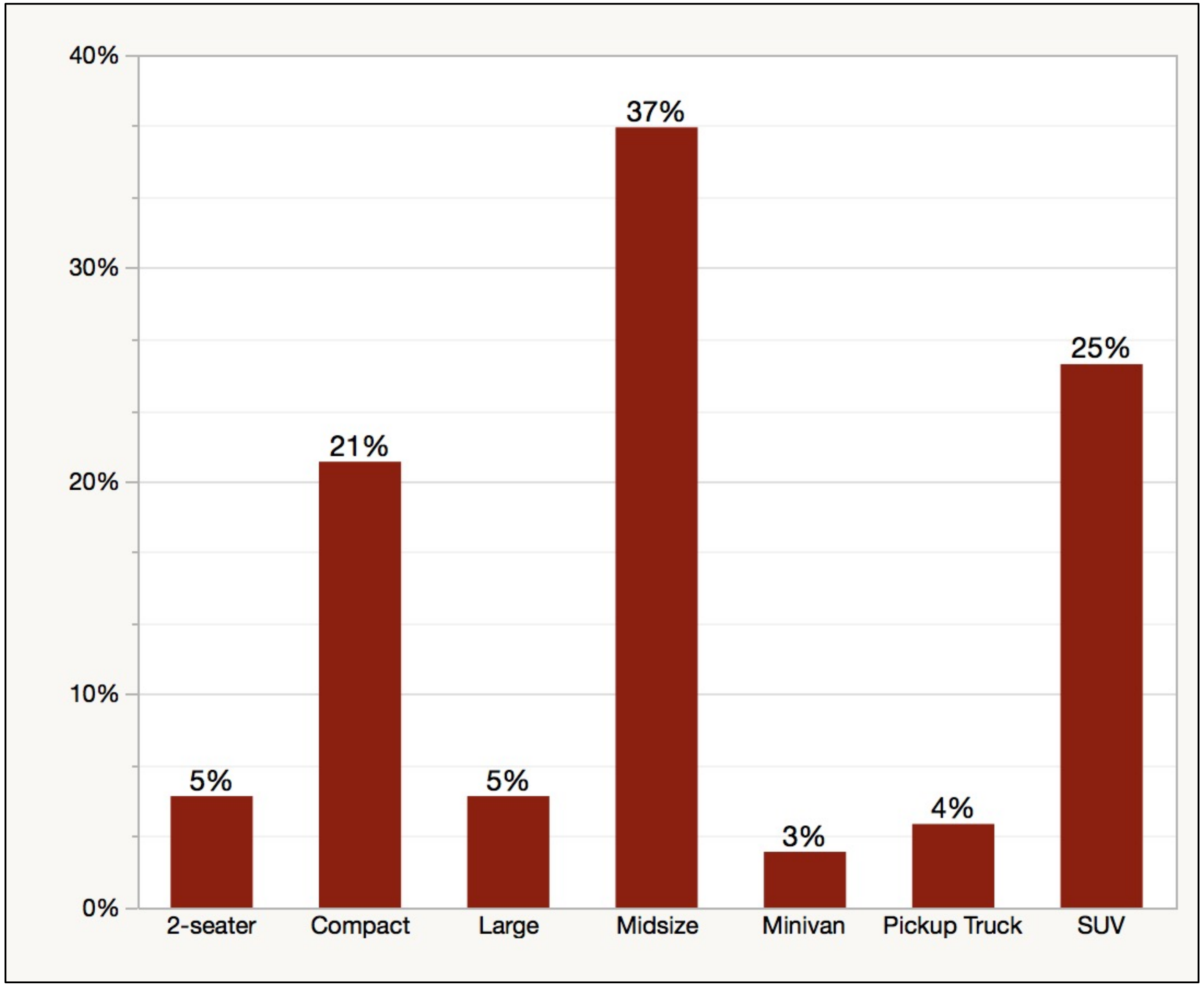
Pie Charts



Pie charts display *percentages* of cars by their classes, a *categorical* variable.

Bar Charts

Bar charts display *percentages* or *counts* of different cars by their classes, a *categorical* variable.



Central Tendency

How do we describe a dataset, especially if it is rather large, without having to present a table of meaningless numbers?

Generally, just two numbers will suffice:

1. Measure of central tendency (i.e. typical value, or location),
2. Measure of dispersion (fluctuation).

Common measures of **central tendency:**

Mean **(μ):**	Average or expected value
Median **(M_d):**	Middle point of ordered observations
Mode **(M_o):**	Most frequent value

Population Mean

The **mean** of a **population** of *N* measurements $x_1, \cdots, x_N$:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

Eg: Viewing our data set of the *Hybrid Cars' MPG* as a **population**, the population mean is

$$\mu = \frac{1}{153}\sum_{i=1}^{153} x_i = \frac{1}{153}(41.26 + 54.1 + \cdots + 37) = \boxed{34.7975 \text{ mpg}}$$

Sample Mean

The **mean** of a **sample** of *n* measurements $x_1, \cdots, x_n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Eg: Assessing only the SUV cars from the *Hybrid Car MPG* dataset, of which there are 39 rows, the **sample mean** is

$$\bar{x} = \frac{1}{39}\sum_{i=1}^{39} x_i = \frac{1}{39}(18.82 + 21 \ldots + 33.64) = \boxed{26.0077 \text{ mpg}}$$

Median

We can use $\bar{x}$ as an estimate of μ, but we then need to assess the *accuracy* of this and draw conclusions, or *make inferences*, about μ.

Problem: $\bar{x}$ is extremely sensitive to outliers.

- Outliers may be due to errors in recording data
- May be real (but exceptional) observations
- Usually set aside outliers before computing
- Can also use *median*

Whenever a dataset has extreme values, the **median** is the preferred measure of central location.

Median

Given n measurements arranged in order of magnitude,

Median = Middle value if n is odd, or
Average of two middle values if n is even.

Eg: CEO compensation for 5 food processing firms:

Pillsbury	698,000
Borden	1,200,000
Campbell Soup	646,000
Hershey Foods	573,000
Ralston Purina	750,000

Median

Converting to multiples of $1,000 and arranging in order:

573, 646, 698, 750, 1200

Median compensation is? $698,000

Mean compensation is? $773,400

- *Mean* > *median* because of outlier, Borden.

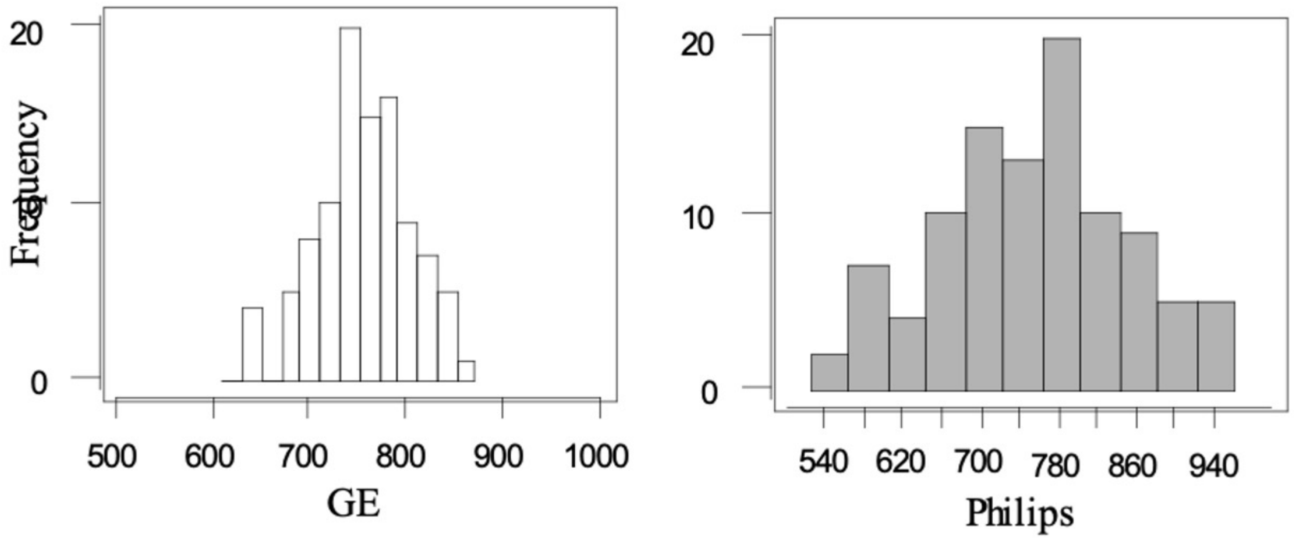
Removing Borden, *mean* = $666,750 < $672,000 = *median*

- Divides data set into two equal parts
- Half of data lies below median, half lies above it
- Resistant to outliers

Dispersion

Mean and **median** do not completely summarize a dataset… we also need to know how spread out the data is.

Lightbulb Lifetimes (hrs): GE vs Philips



GE has less volatility.

- GE exhibits better quality control: not much variation
- Philips has more fluctuation although average is same as GE

Range

Range: Largest minus smallest measurement

- Crude measure with little info about dispersion of values
- No resistance to outliers

Eg: **Range** of Hybrid Car MPG dataset

- Highest value: 72.92 mpg (Prius Alpha V)
- Lowest value: 17 mpg (Silverado 2WD)

Range = 72.92 mpg – 17 mpg = $\boxed{\text{55.92 mpg}}$

Interquartile Range (IQR)

Interquartile range (IQR): $Q_3 - Q_1 = 75^{th}$ %ile – 25^{th} %ile

- Width of "middle half" of dataset when ordered from smallest to largest
- Resistant to outliers (robust measure)

better indicators than range for spread.

1st quartile: $Q_1 = 25^{th}$ percentile, 25% of values lie below

Median of the lower half of the data.

2nd quartile: Q_2 or 50^{th} percentile = Median.

3rd quartile: Q_3 or = 75^{th} percentile, 75% of values lie below

Median of the upper half of the data.

Eg: IQR of *Hybrid Car MPG* dataset

- Q_3 = 41.565 mpg
- Q_1 = 26 mpg

IQR = 41.565 mpg – 26 mpg = $\boxed{\text{15.565 mpg}}$ ⇒ more reliable than previous slide

Interquartile Range (IQR)

20 customer satisfaction ratings:

1 3 5 5 7 8 8 8 8 8 8 9 9 9 9 9 10 10 10 10

Find the IQR for customer satisfaction ratings: IQR = 9 – 7.5 = 1.5

What is the 50th percentile? Average of 8 & 8 = 8

What is the 25th percentile? Average of 7 & 8 = 7.5

What is the 75th percentile? Average of 9 & 9 = 9

range would be 9 which is useless info

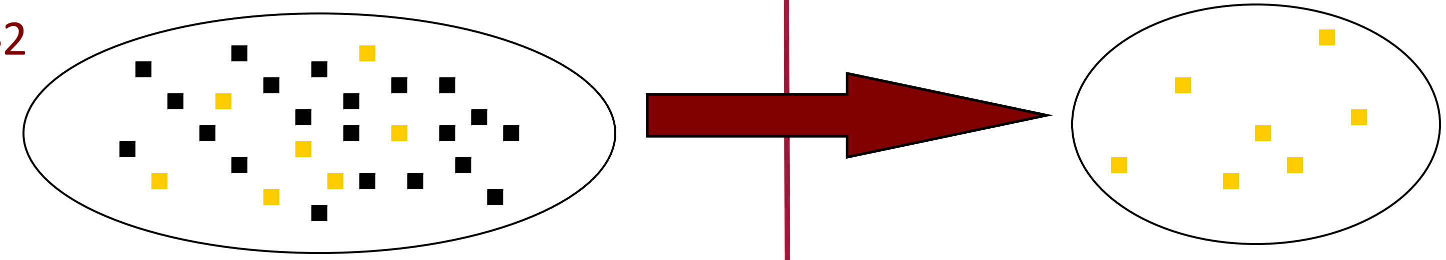
but its still good to know min & max

Variance & Standard Deviation

Population $X_1, X_2, \ldots, X_N$

Sample $x_1, x_2, \ldots, x_n$

σ^2



s^2

Population Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Population Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample Standard Deviation:

$$s = \sqrt{s^2}$$

-1 inflates the variance estimate slightly.

stdev.p & stdev.s in excel.

Variance & Standard Deviation

MPG of 153 Hybrid Cars:

Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{153}(41.26 + 54.1 + \cdots + 37) = \boxed{34.8 \text{ mpg}}$

Variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

$$= \frac{1}{153}[(41.26 - 34.8)^2 + \cdots + (37 - 34.8)^2]$$

$$= \boxed{120.3958 \text{ mpg}^2}$$

Standard Deviation: $\sigma = \sqrt{\sigma^2} = \boxed{10.9725 \text{ mpg}}$

Variance & Standard Deviation

MPG of 39 SUV Hybrid Cars:

Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{39}(18.82 + 21 + \cdots + 33.64) = \boxed{26 \text{ mpg}}$

Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$= \frac{1}{38}[(18.82 - 26)^2 + \cdots + (33.64 - 26)^2]$$

$$= \boxed{21.149 \text{ mpg}^2}$$

Standard Deviation: $s = \sqrt{s^2} = \boxed{4.599 \text{ mpg}}$

Coefficient of Variation

The coefficient of variation indicates how large the standard deviation is in relation to the mean and is useful for comparing levels of fluctuation between different variables.

The coefficient of variation for a population is computed as:

$$c_v = \left[\frac{\sigma}{\mu} \times 100\right] \%$$

And for a sample it is:

$$c_v = \left[\frac{s}{\bar{x}} \times 100\right] \%$$

way of comparing 2 set of records which do not have same middle.

Coefficient of Variation

MPG of the Hybrid Car dataset:

The coefficient of variation for the MPG of the population of 153 Hybrid Cars is:

$$c_v = \left[\frac{10.9725}{34.7975} \times 100\right]\% = \boxed{31.53\%}$$

And the coefficient of variation for the MPG of the sample of 39 SUV Hybrid Cars is:

$$c_v = \left[\frac{4.5988}{26.0077} \times 100\right]\% = \boxed{17.68\%}$$

→ The sample of SUV data is relatively less variable than the population.

Coefficient of Variation

Comparison of two stocks, Pfizer and Johnson & Johnson:

Monthly adjusted closing PFE and JNJ stock prices (4/1/08 – 3/1/18) had:

	PFE Adj Close	JNJ Adj Close
$\bar{x}$	22.18	76.67
s	8.36	29.33

The coefficient of variation for PFE is: $c_{v,PFE} = \left[\frac{8.36}{22.18} \times 100\right]\% = \boxed{37.68\%}$

And the coefficient of variation for the JNJ is: $c_{v,JNJ} = \left[\frac{29.33}{76.67} \times 100\right]\% = \boxed{38.25\%}$

→ The two stock prices seem to be relatively *equally* risky!

Empirical Rule

A normal population with mean μ and standard deviation σ has approximately

68.26% of the population measurements within one standard deviation of the mean:

$$[\mu - \sigma, \qquad \mu + \sigma]$$

95.44% of the population measurements within two standard deviations of the mean:

$$[\mu - 2\sigma, \qquad \mu + 2\sigma]$$

99.74% of the population measurements within three standard deviations of the mean:

$$[\mu - 3\sigma, \qquad \mu + 3\sigma]$$