Data

- Measurement
  - What is being measured?
  - How is it being measured?
  - Why is it being measured?
- Mean and Variation
  - What is the central tendency of the data?
  - Why are all the observations in the dataset not the same?
- Relationships with other variables

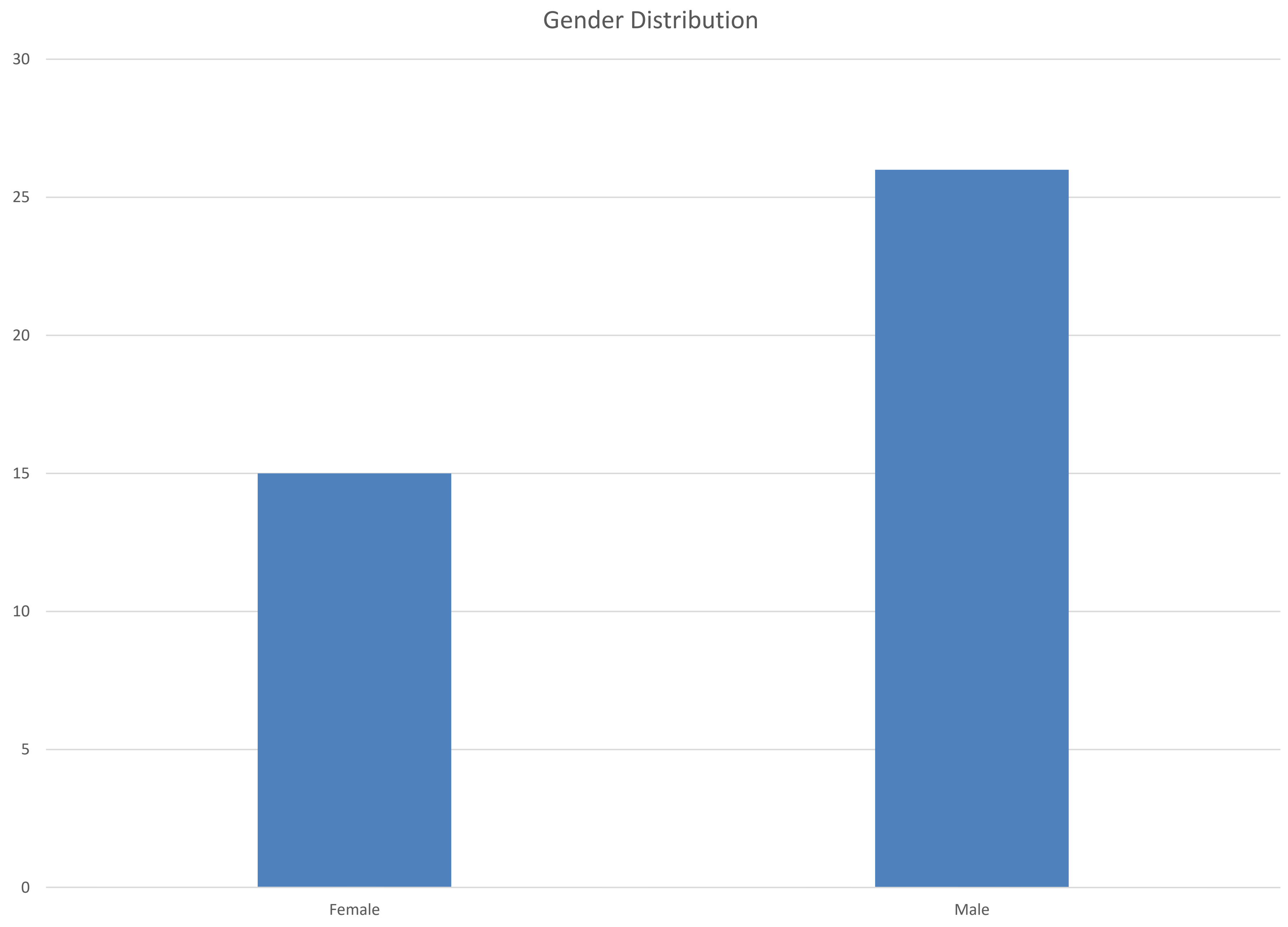
Examining Data

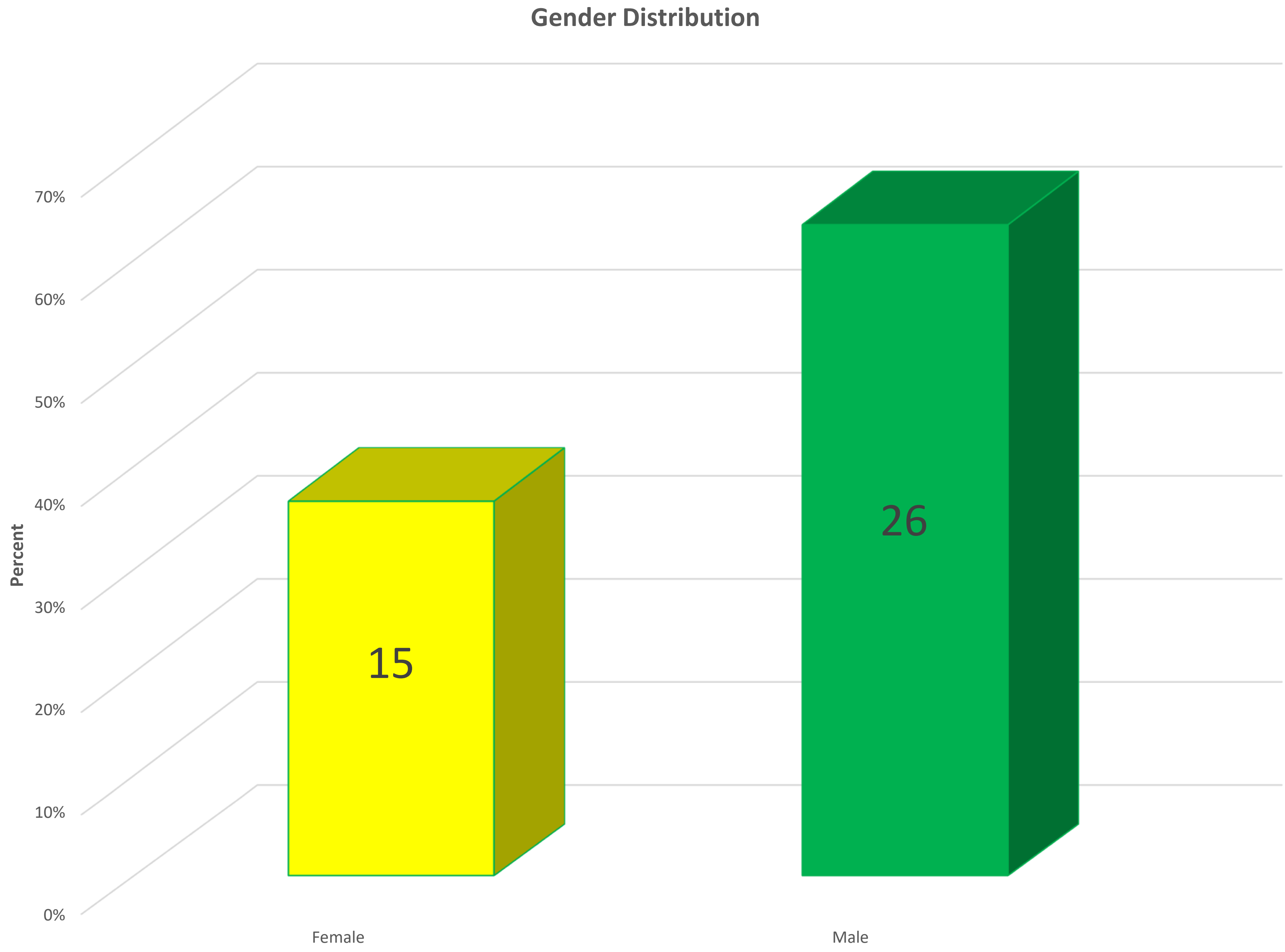
- Begin with graphs
  - Then go on to numerical summaries

- First look at each variable by itself
  - Then look at relationships between variables

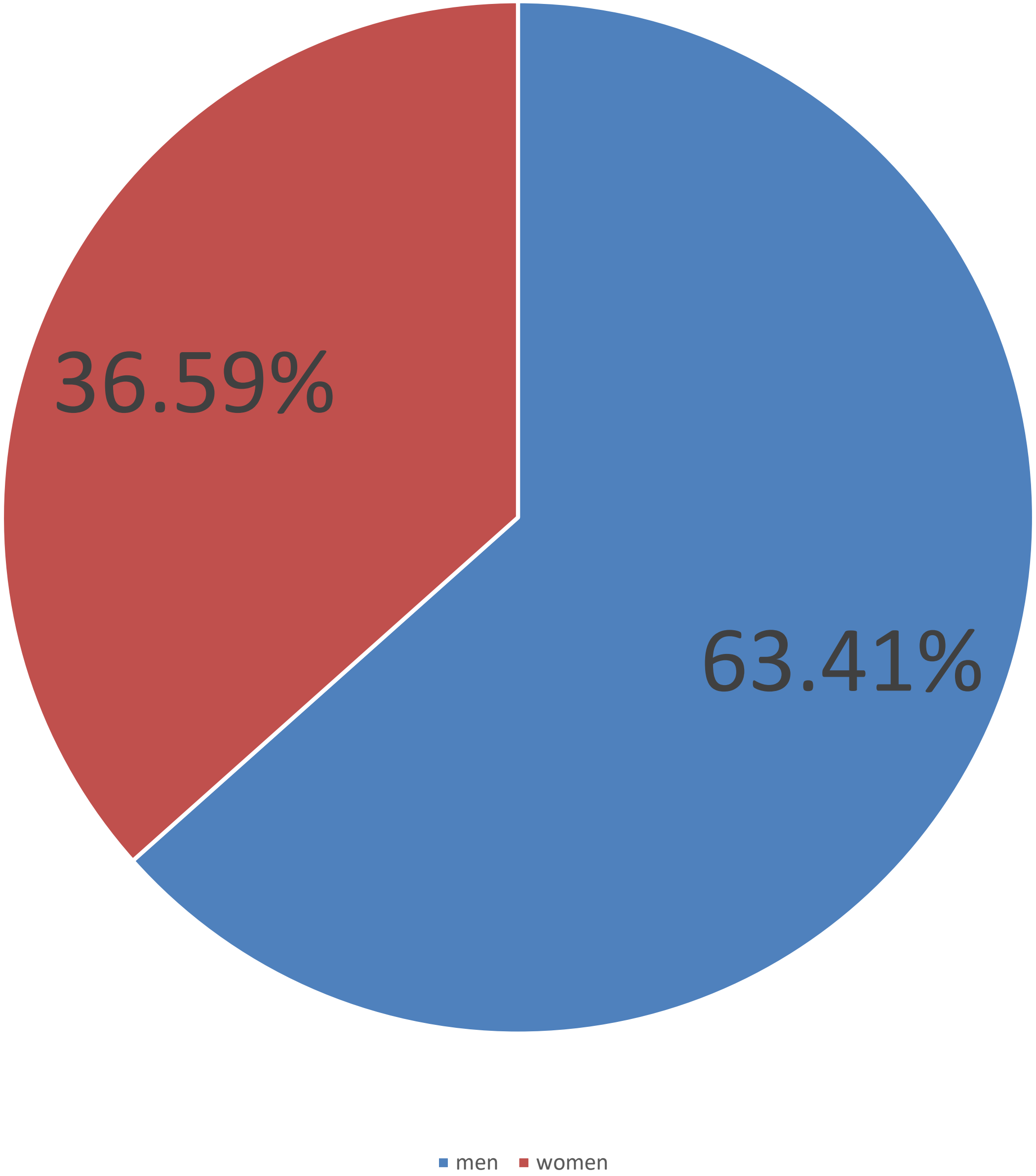
Graphing Data

- Graphs for qualitative data
  - Bar graphs (count or percent)
  - Pie charts (percents)
  - * Be careful with 3D figures
- Graphs for quantitative data
  - Stem plots (stem and leaf plots)
  - Histogram
  - Time series graphs
  - * Pay attention to the scale

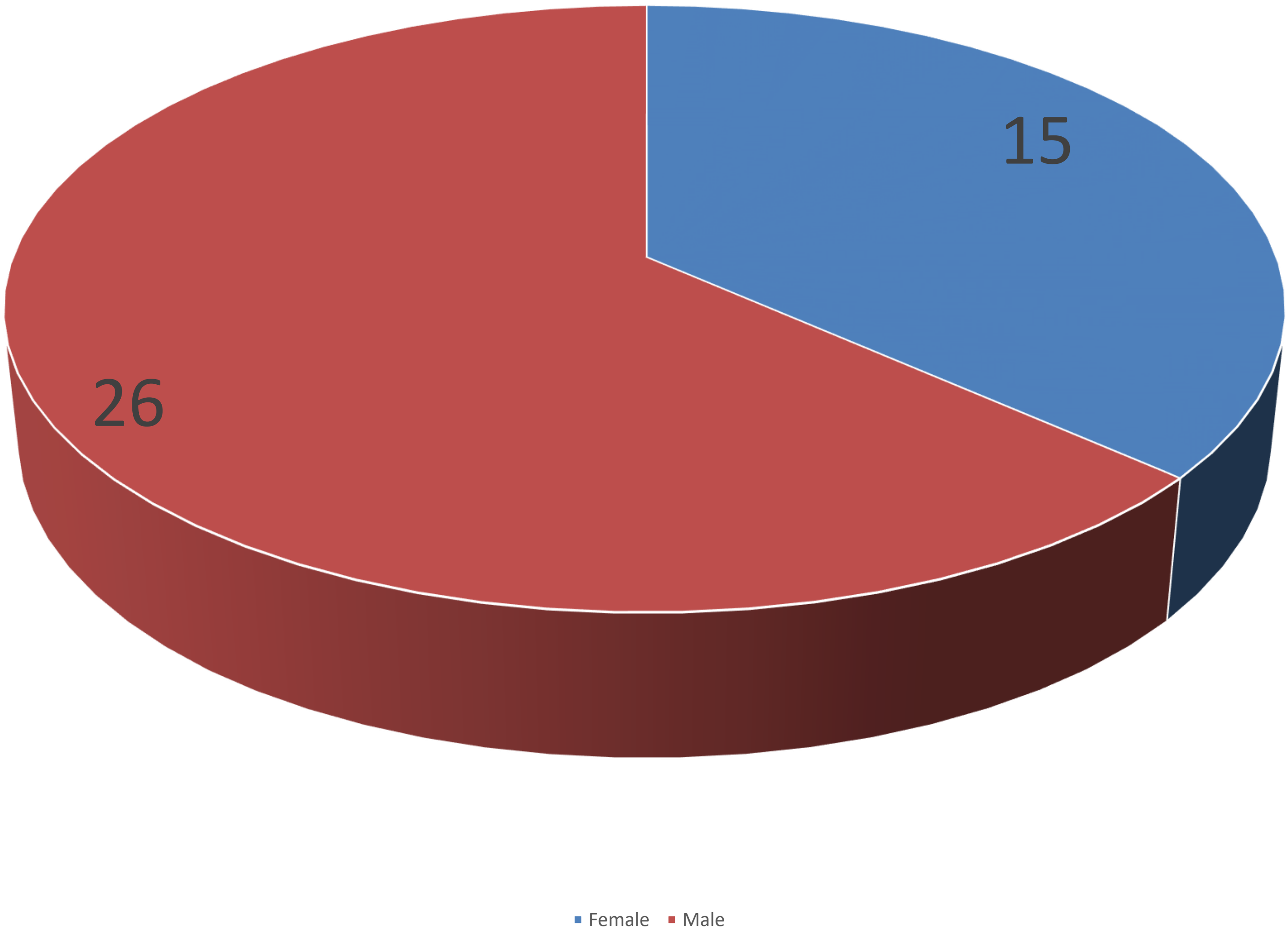

Gender Distribution
30
25
20
15
10
5
0
Female
Male

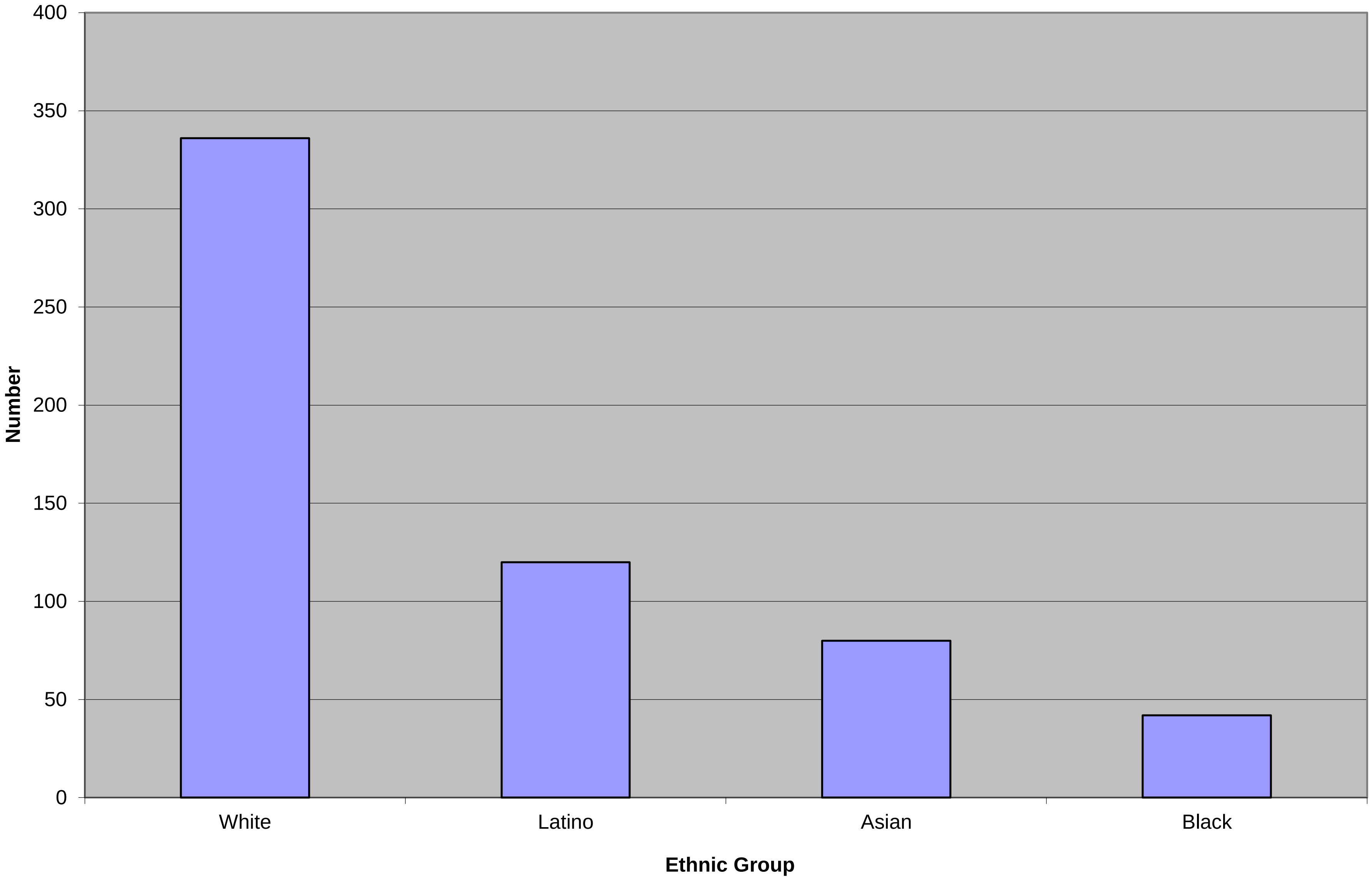

Gender Distribution
Percent
70%
60%
50%
40%
30%
20%
10%
0%
15
26
Female
Male

Gender Distribution



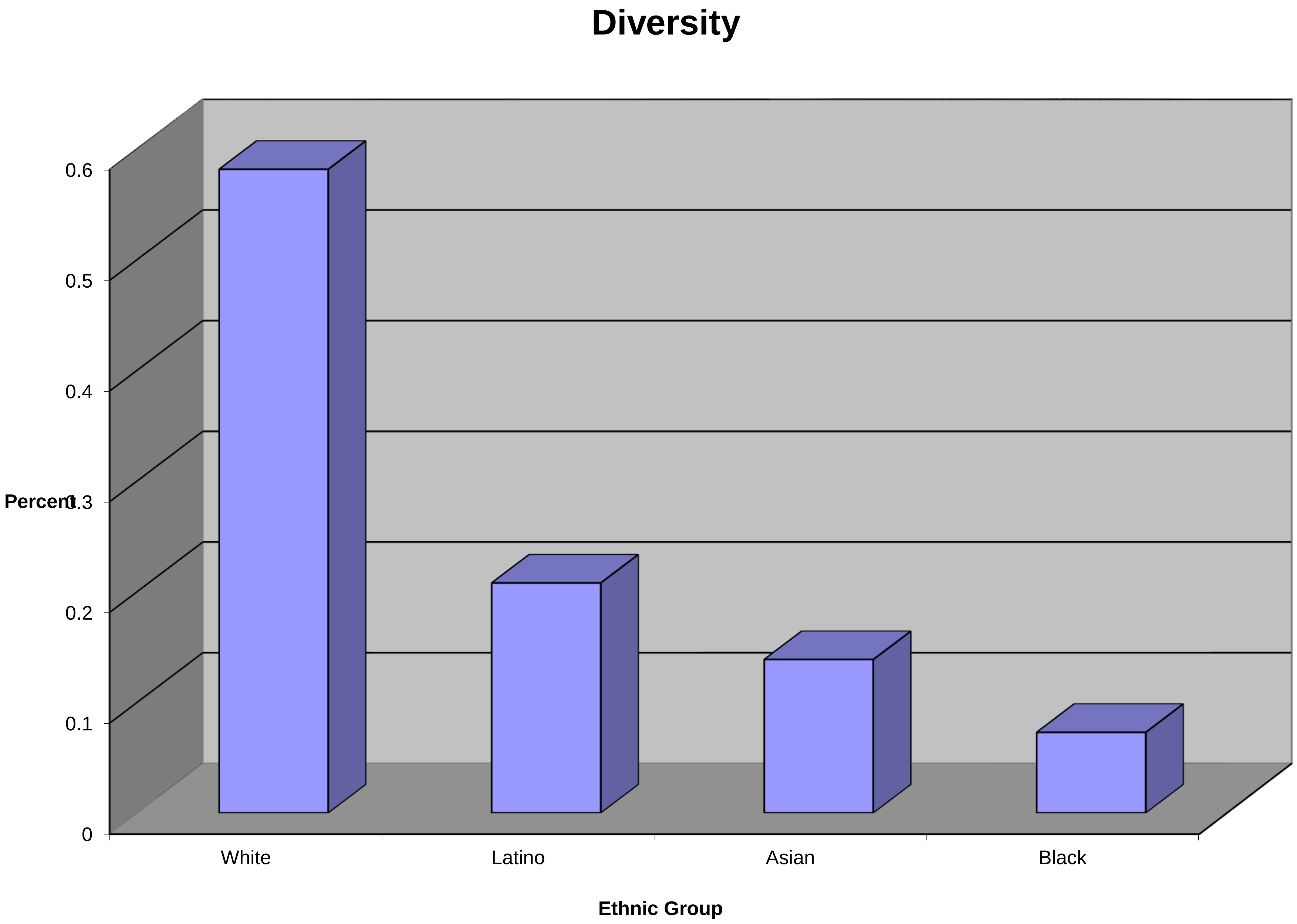
Gender Distribution



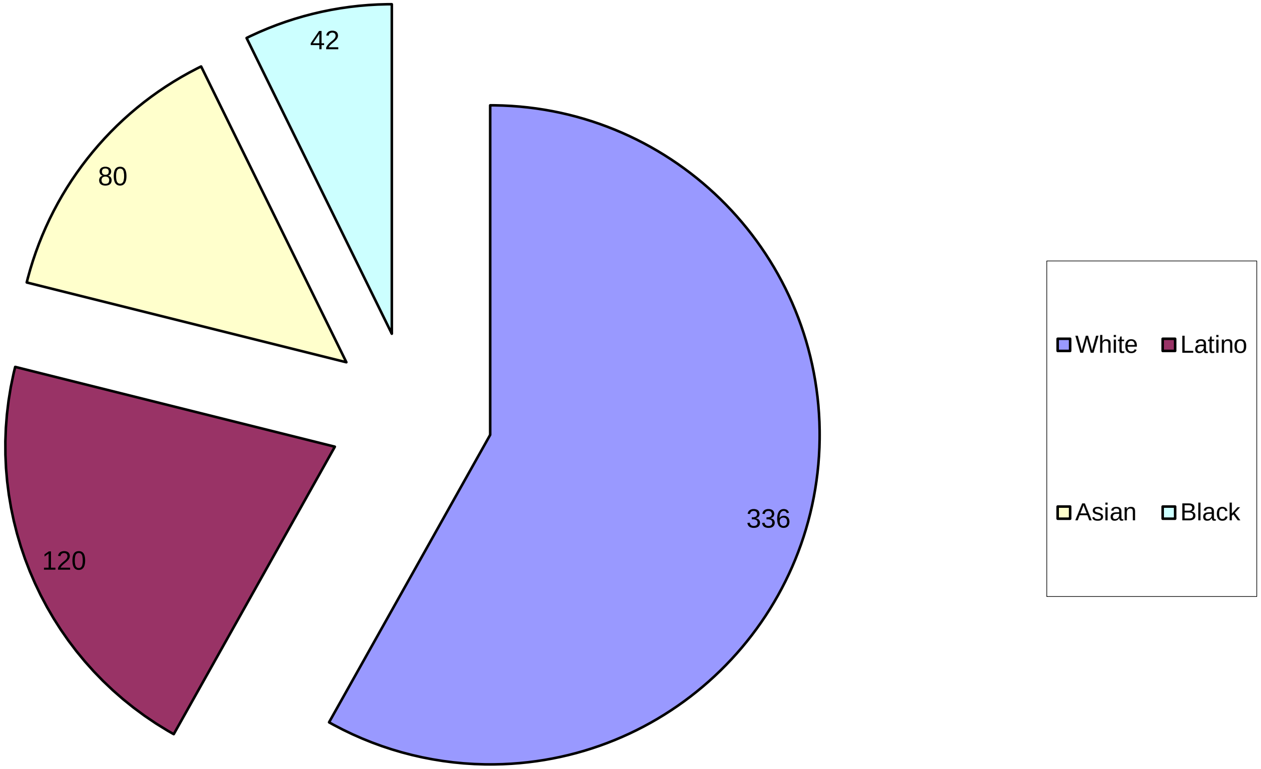


Diversity

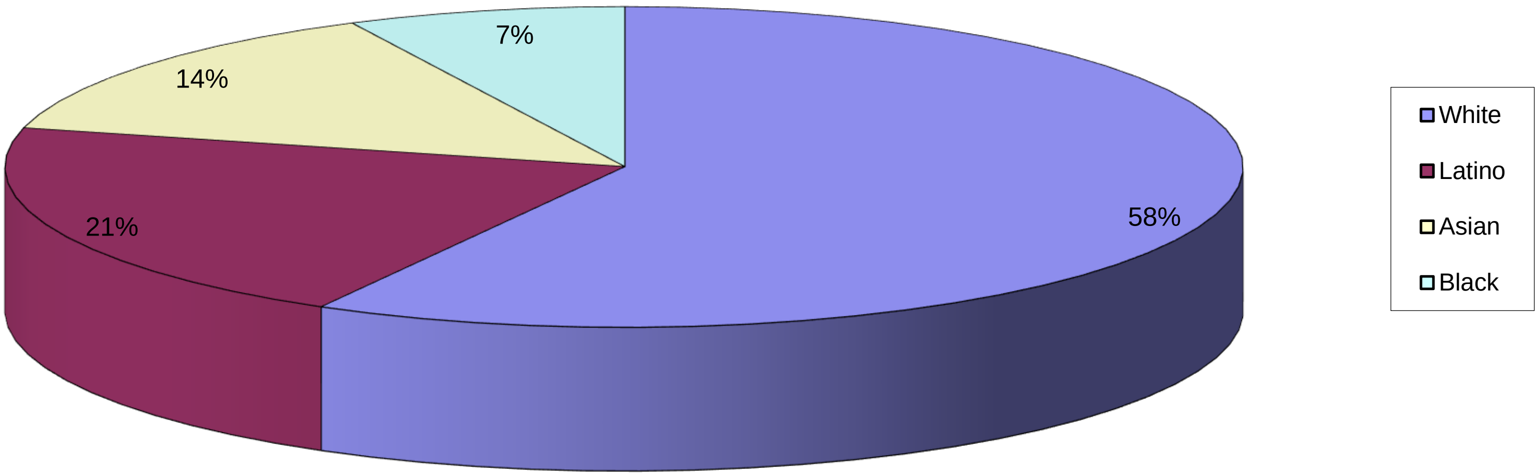
400
350
300
250
200
150
100
50
0
Number
White
Latino
Asian
Black
Ethnic Group


Diversity
Percent
0.6
0.5
0.4
0.3
0.2
0.1
0
White
Latino
Asian
Black
Ethnic Group

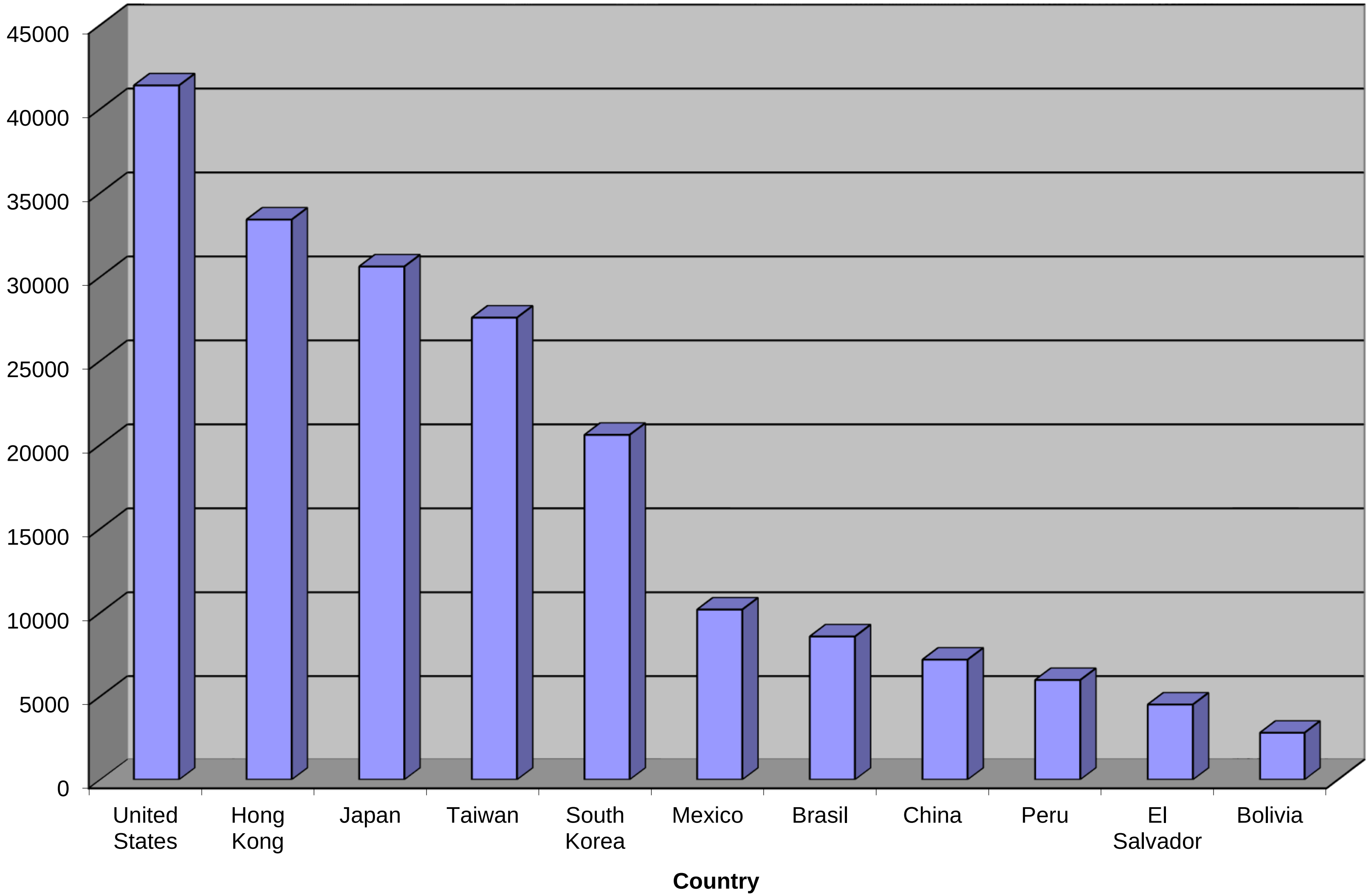
Diversity



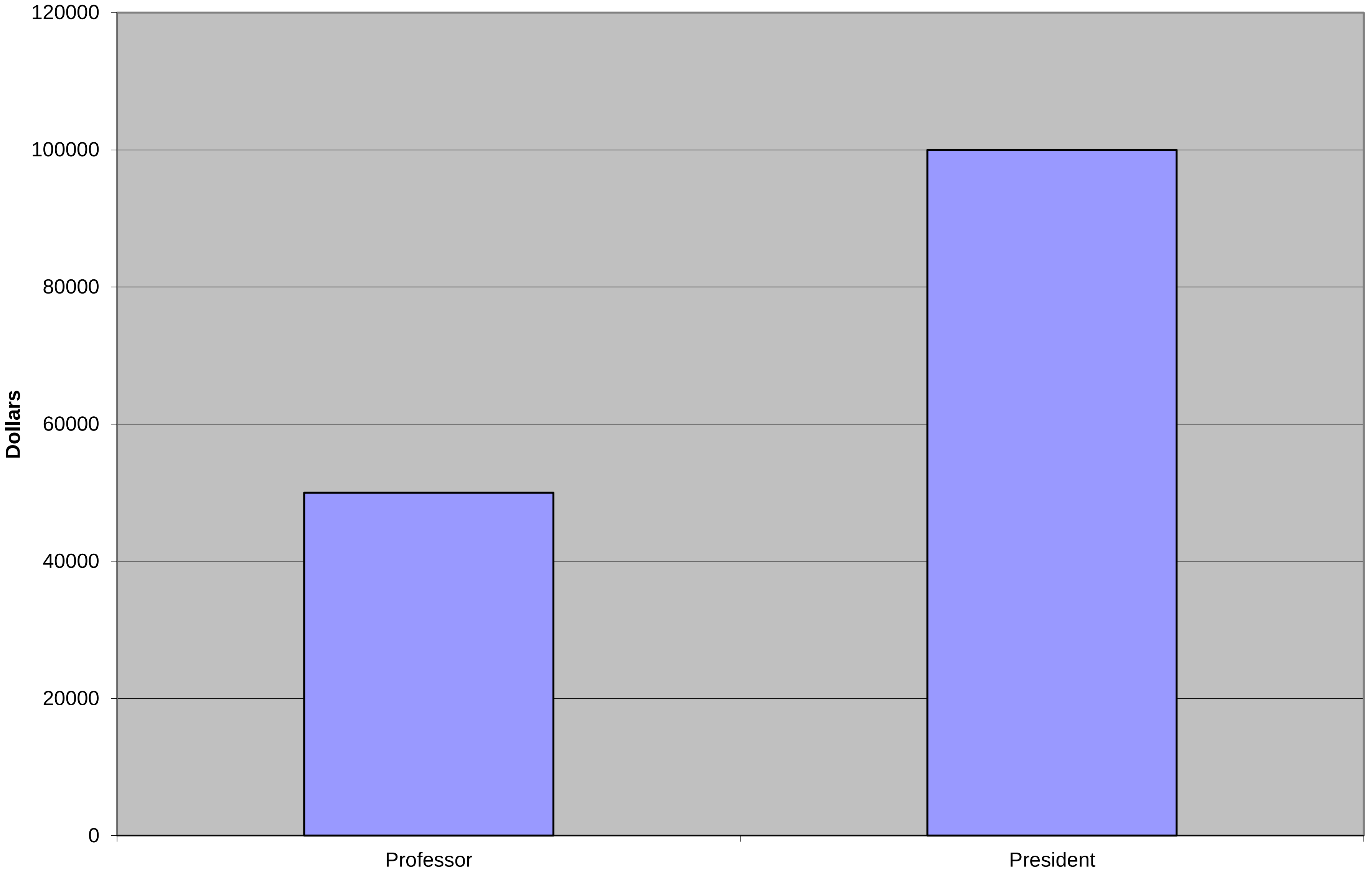
Diversity

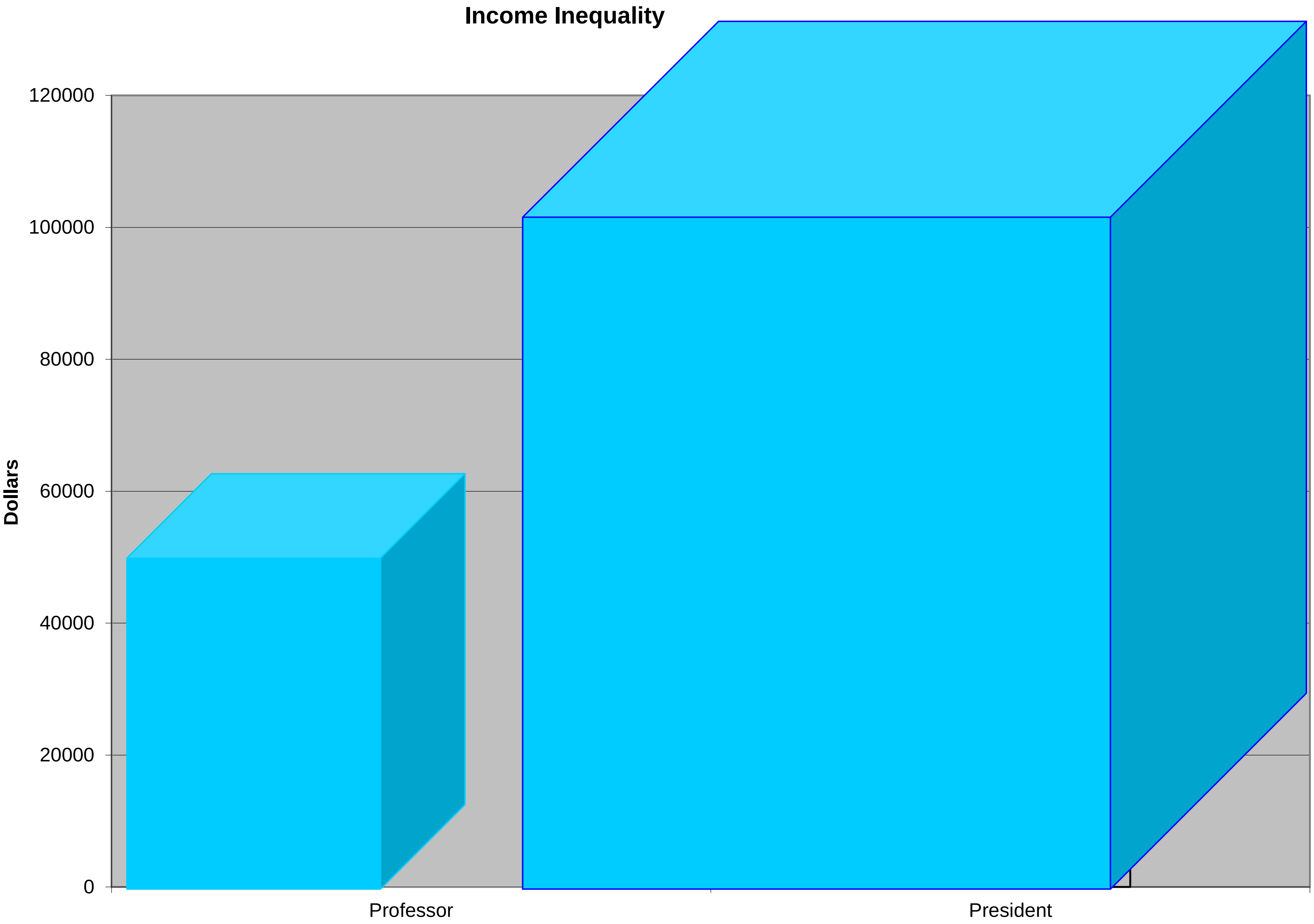


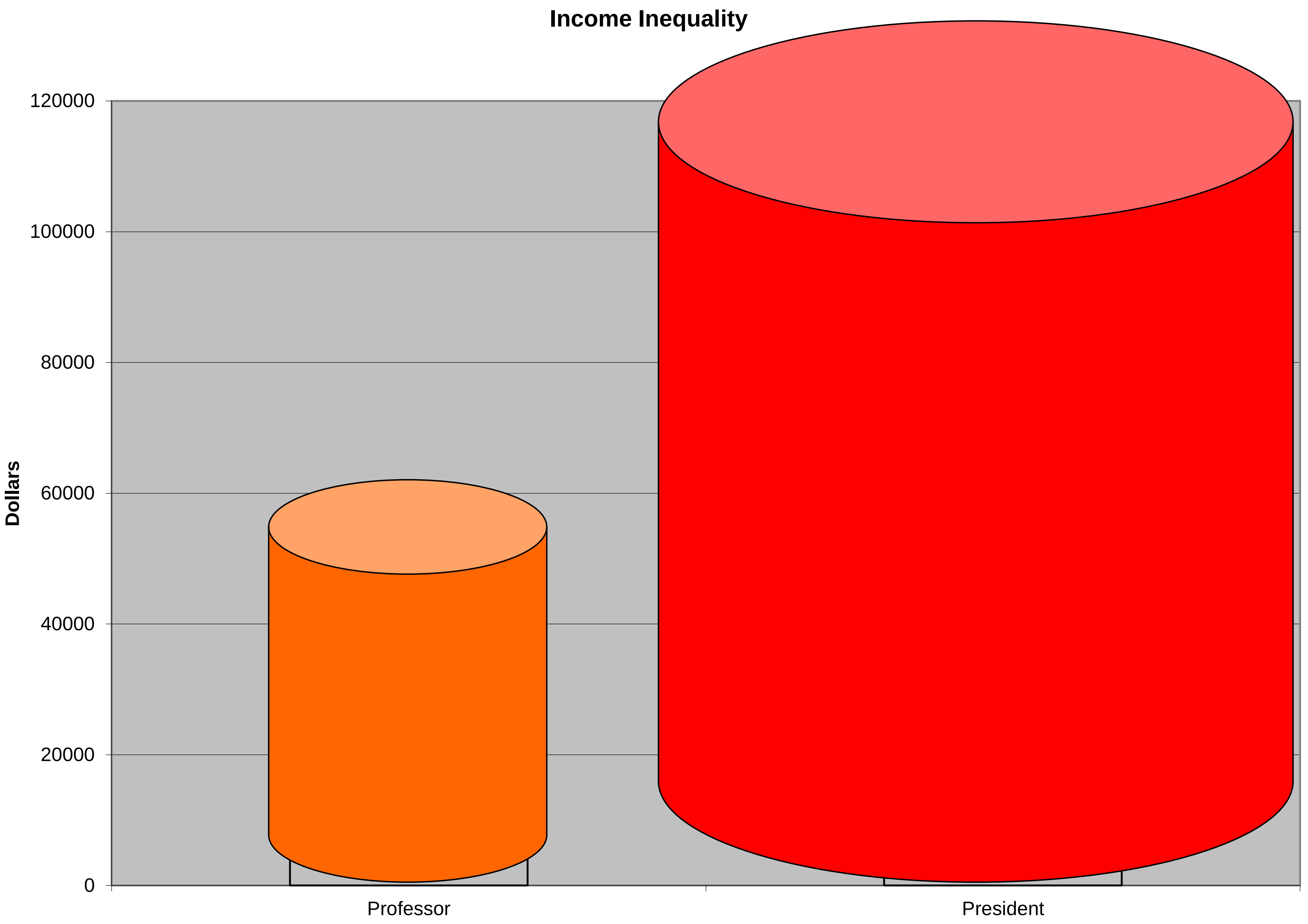
GDP Per Capita



Income Inequality




Income Inequality
Dollars
120000
100000
80000
60000
40000
20000
0
Professor
President



Income Inequality

Graphing Data

- Graphs for qualitative data
  - Bar graphs (count or percent)
  - Pie charts (percents)
  - * Be careful with 3D figures
- Graphs for quantitative data
  - Stem plots (stem and leaf plots)
  - Histogram
  - Time series graphs
  - * Pay attention to the scale

Stem Plots

1. Stem – all but the right most digit
2. Leaf – final digit
3. Write stems in a vertical column, largest to smallest
4. Write each leaf in a row next to the stem

Stem Plots

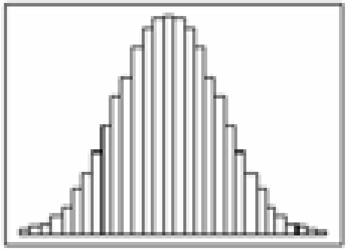
Examine the overall distribution of the data

- overall pattern and striking deviations
- shape, center, and spread
- outliers: fall outside the overall pattern

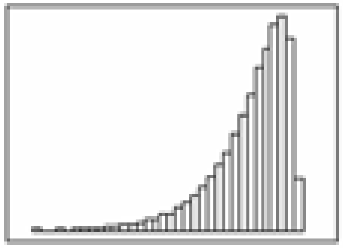
Does the distribution have one peak (unimodal) or several peaks?

Is the distribution symmetric, skewed to the right, or skewed to the left?

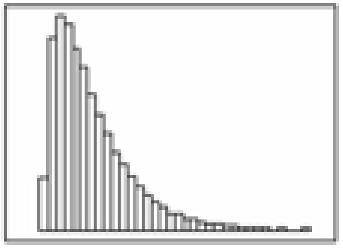
Symmetric or Skewed



Symmetric
Bell shaped



Skewed to
the Left



Skewed to
the Right

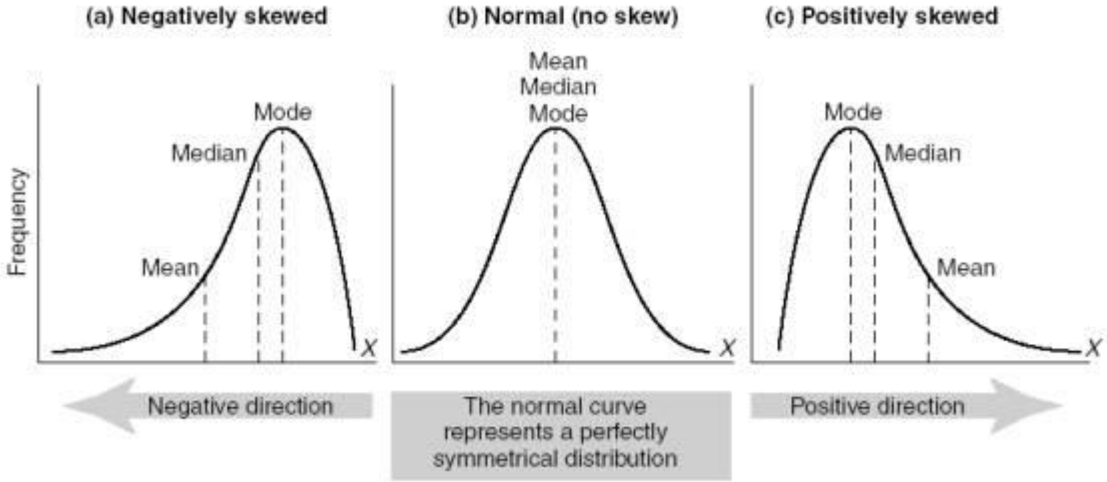


FIGURE 15.6 Examples of normal and skewed distributions

Histograms

- Stem plots are of limited usefulness
  - Hard with large datasets
  - Can’t choose your own interval sizes
- Histograms work better
- You can choose the size of your intervals
- You can display counts or percentages

Time Series Plots

- When data are collected over time
- Plot observations in time order
- Observe any seasonal variation
  - Repetition at regular known intervals
- Observe trends over time
  - Persistent long term rise or fall
- * Notice the scale of the axes!

Numerical summaries

- Use numbers to describe the center and spread of any dataset
- Measures of Center
  - Mean: average value
  - Median: middle value
  - Mode: most common value
- Measures of spread
  - Range
  - Interquartile Range
  - Five number summary (box plots)
  - Variance and Standard Deviation

Choosing measures of center and spread

- Use the sample mean and sample standard deviation if you have a symmetric distribution
- Use the five number summary if you have a skewed distribution
- A plot or graph gives you the best overall picture of a distribution

Changing units of measure

- When you change the units of measure
  - Feet to inches
  - Pounds to kilograms
- The mean, variance and standard deviation will change
- Example: Let Y be a linear transformation of X

  Y = *a*X + *b* where *a* and *b* are constants

$$\bar{Y} = a\bar{X} + b$$

$$s_Y = a(s_X)$$

$$s_y^2 = a^2(s_X^2)$$

How to explore your data

- Plot your data with a stem plot or histogram
- Look at the overall pattern and any striking deviations
- Calculate some numerical summaries to describe the center and the spread of the distribution

Relationships between variables

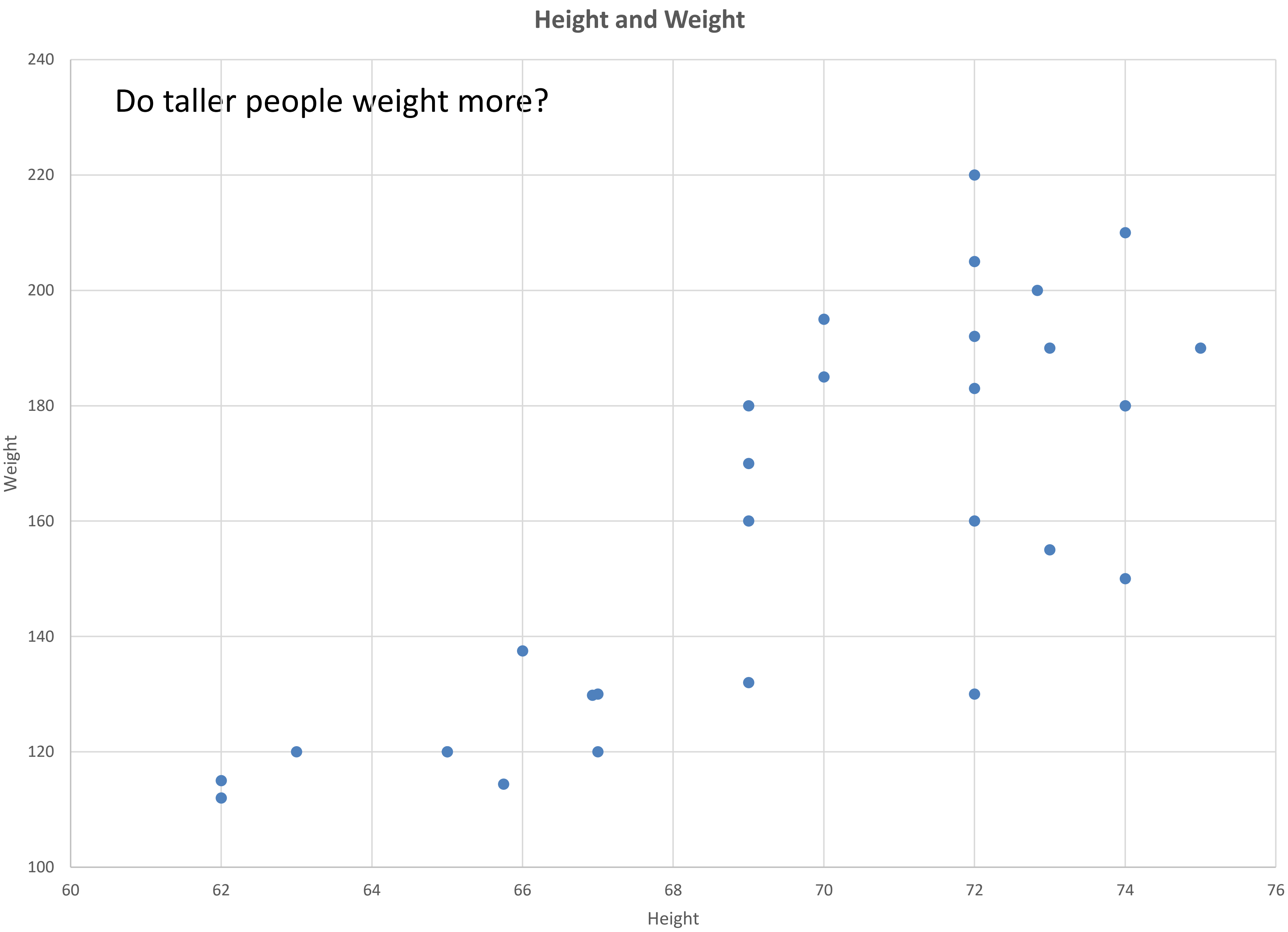
- Is there a relationship between two variables?
- Is it a positive relationship or negative relationship?
- How strong is this relationship?
- Is it an explanatory relationship?
  - Dependent variable: response variable measuring the outcome of a study
  - Independent variable: explanatory variable which explains the change in the dependent variable

Quantitative Data

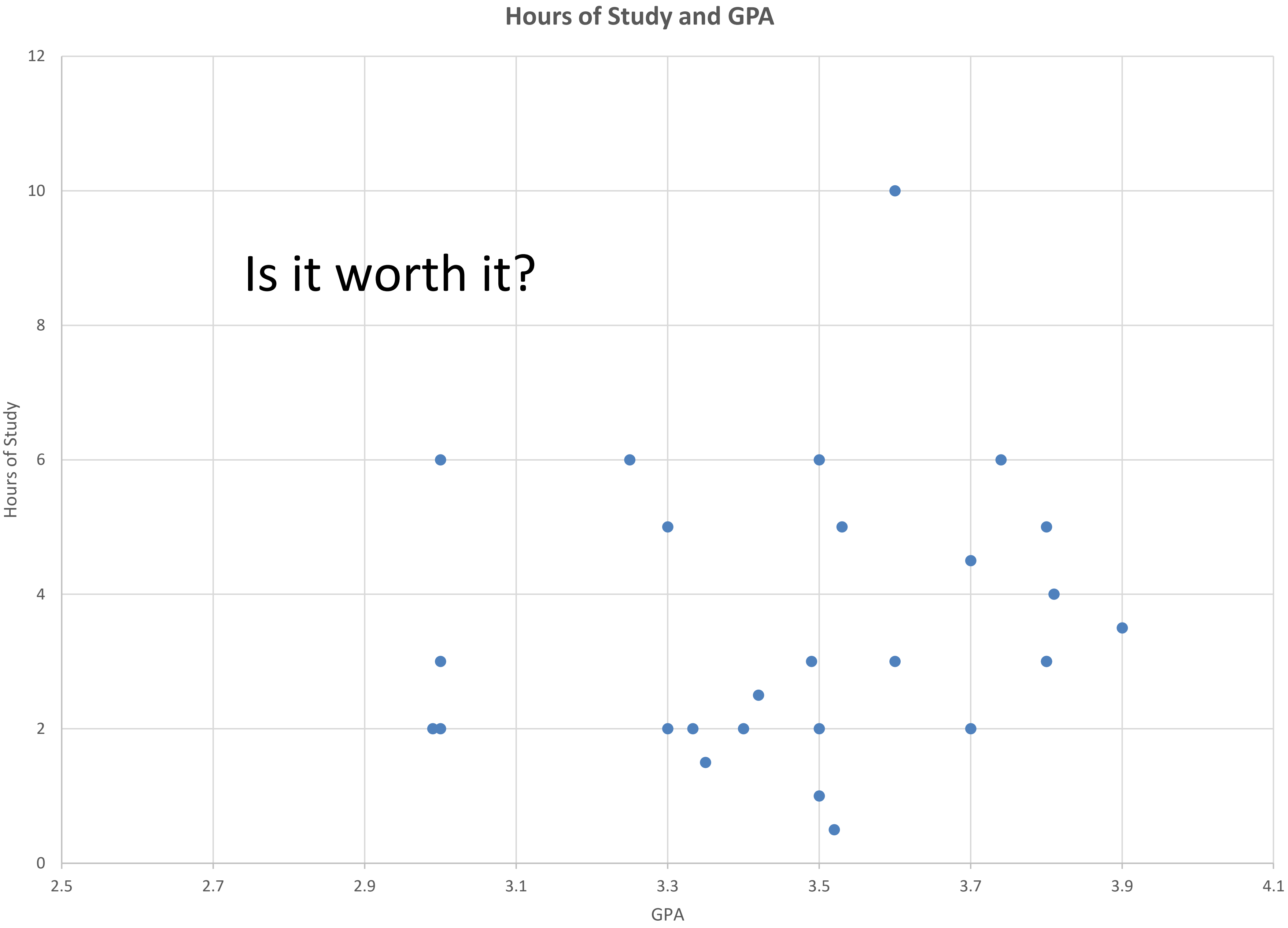
- Start with a graphical display, then add numerical summaries
- Look for overall patterns and deviations from those patterns
- If the pattern is regular, we can try to model the relationship and use regression analysis

Scatterplot

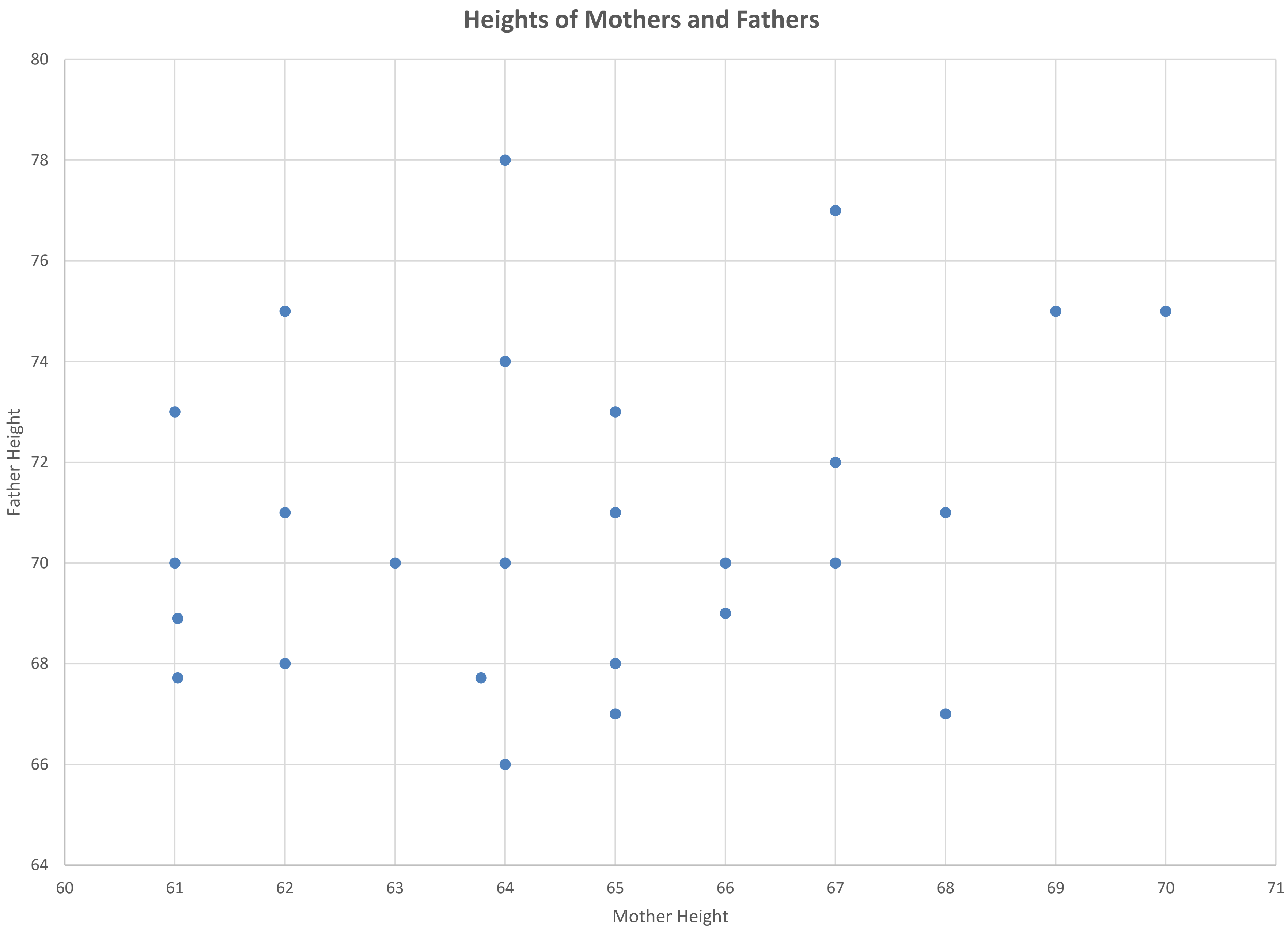
- Displays the relationship between to quantitative variables on the same entity
- Put the dependent (response) Y variable on the vertical axis
- Put the independent (explanatory) X variable on the horizontal axis

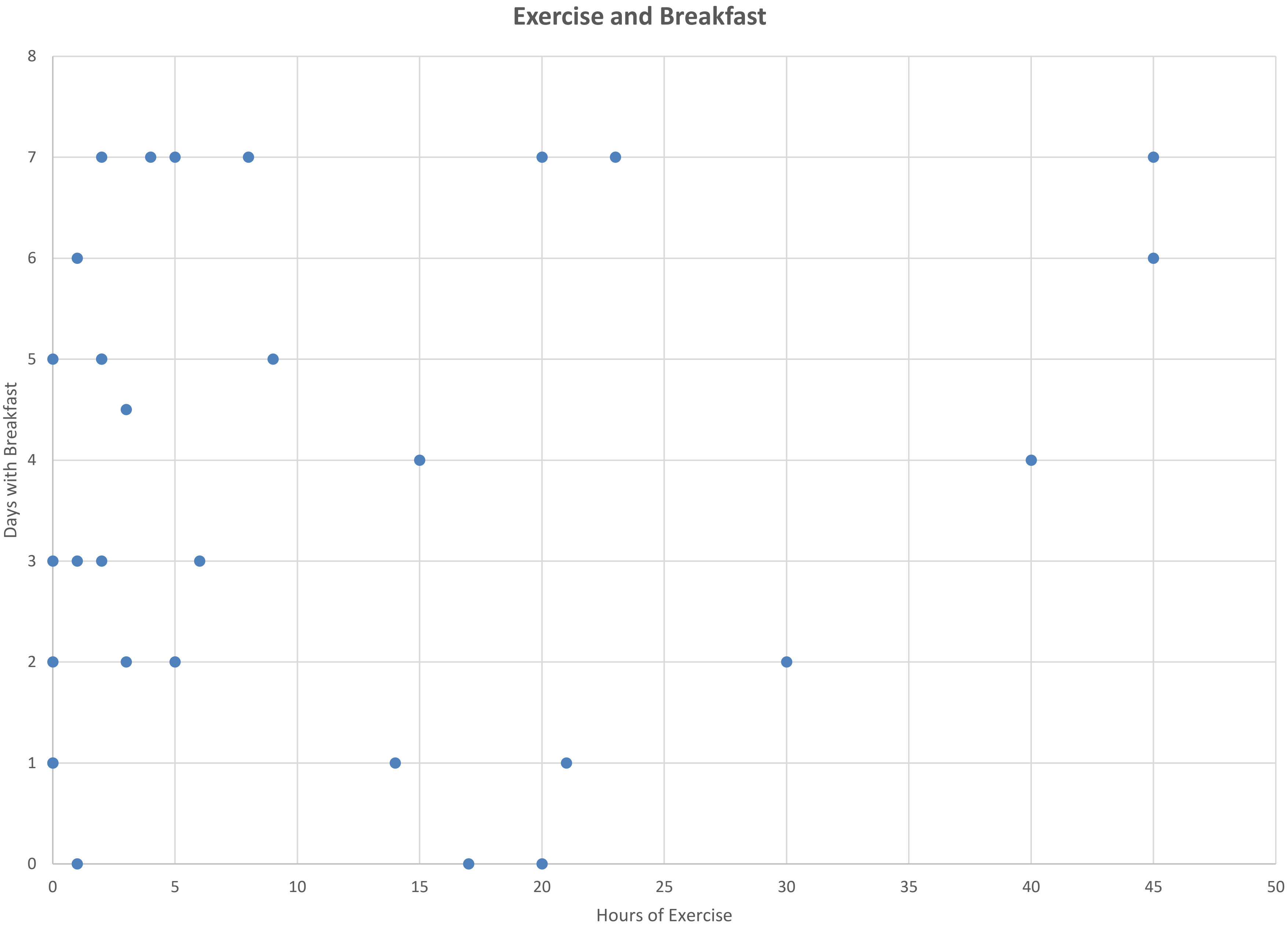


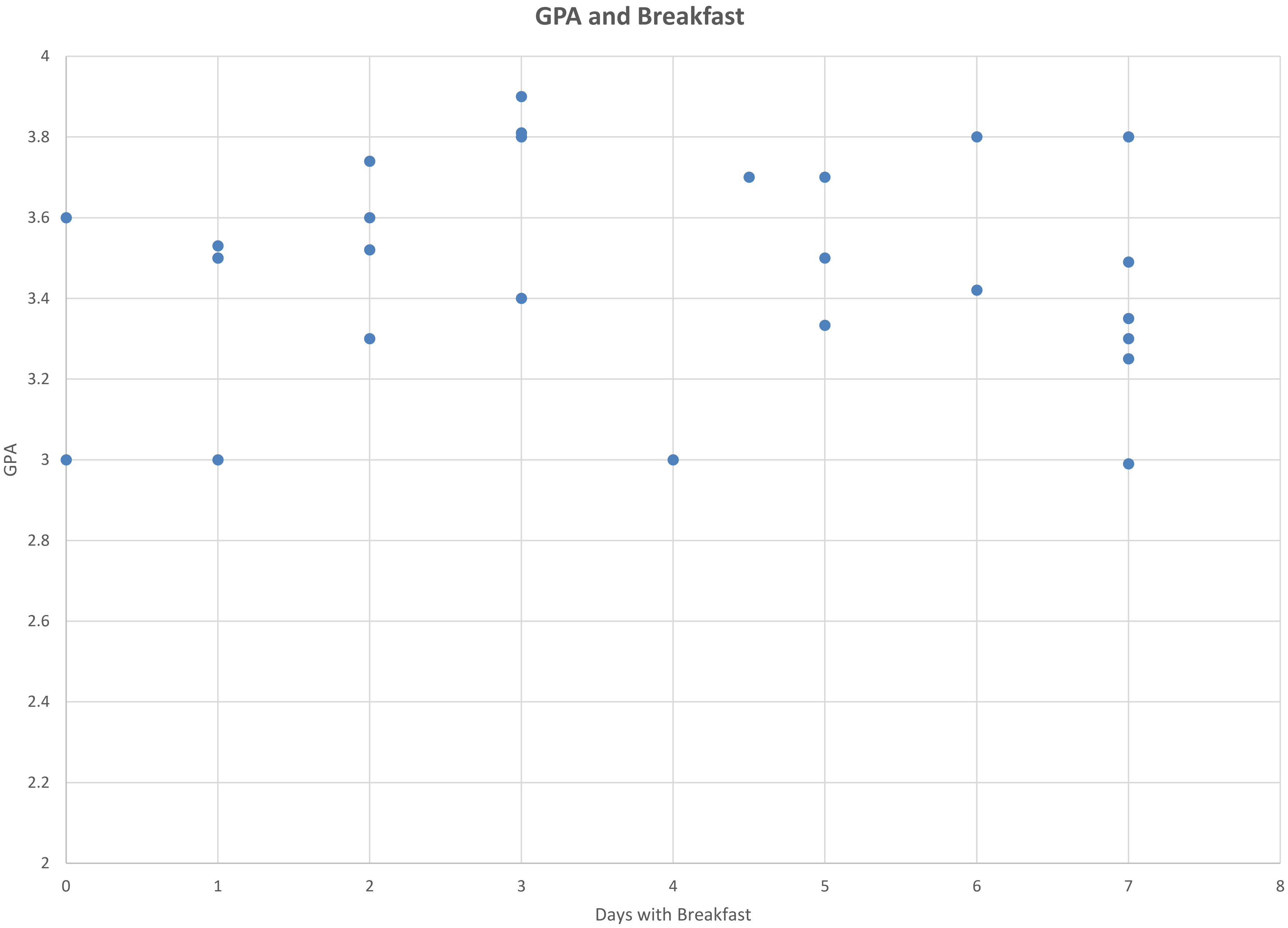
Do taller people weight more?

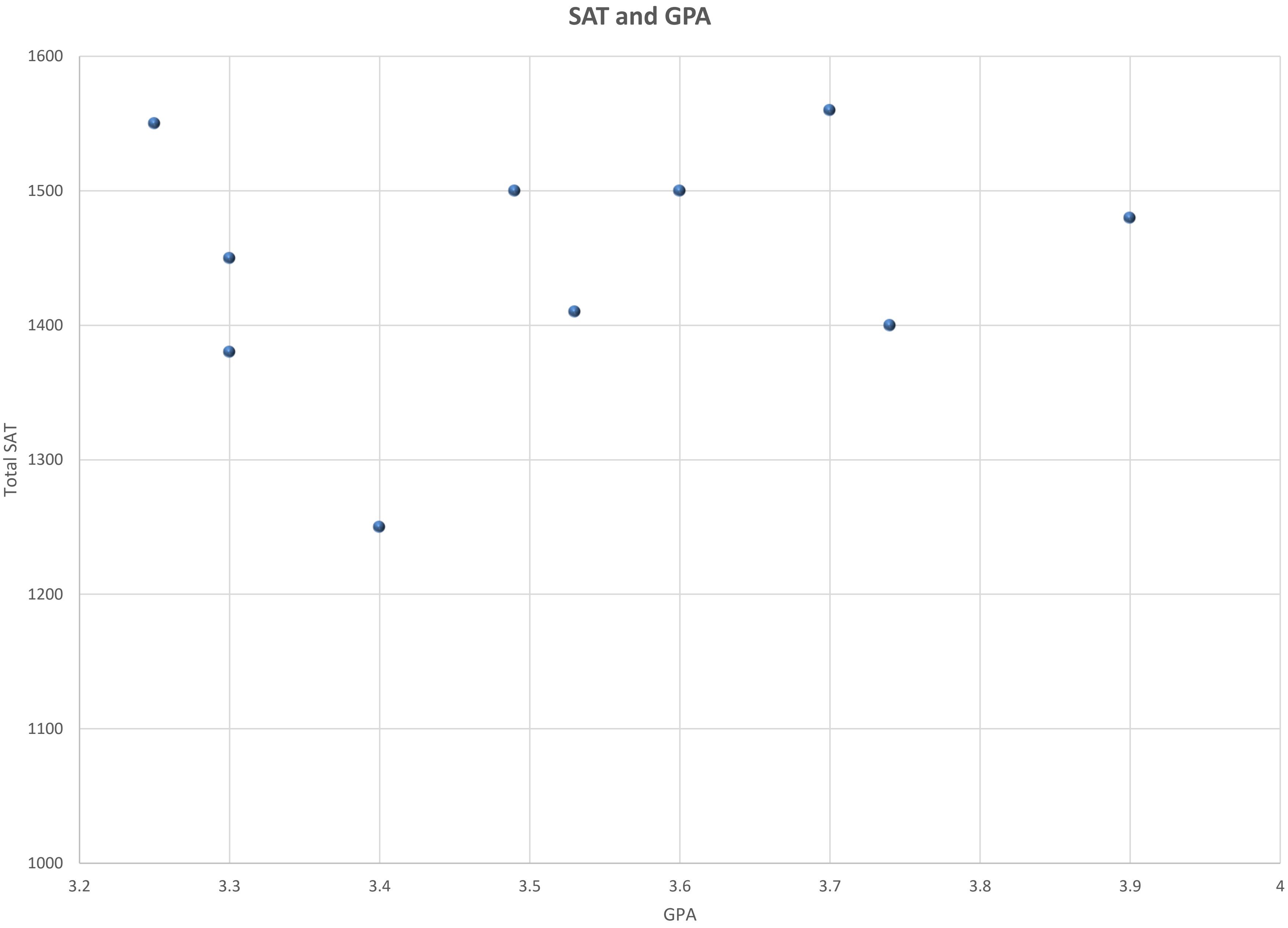


Is it worth it?


Heights of Mothers and Fathers
Father Height
80
78
76
74
72
70
68
66
64
60
61
62
63
64
65
66
67
68
69
70
71
Mother Height


Exercise and Breakfast
Days with Breakfast
0
1
2
3
4
5
6
7
8
Hours of Exercise
0
5
10
15
20
25
30
35
40
45
50


GPA and Breakfast
4
3.8
3.6
3.4
3.2
3
2.8
2.6
2.4
2.2
2
GPA
0
1
2
3
4
5
6
7
8
Days with Breakfast


SAT and GPA
Total SAT
1600
1500
1400
1300
1200
1100
1000
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4
GPA

Measuring Relationships

- Covariance (sample)

$$cov_{xy} = \frac{\sum_N (X_i - \bar{X})\,(Y_i - \bar{Y})}{N - 1}$$

- Correlation Coefficient

$$r_{xy} = \frac{cov_{xy}}{SD_x SD_y}$$

$$-1 \leq r_{xy} \leq 1$$

Qualitative Data

- Examine relationships by looking at tables
- You can present counts or percent
- You can have percent of row totals or column totals or both
- * Be careful of inappropriate aggregation (Simpson's paradox)

Region and Major

	US	Europe	Asia	Total
Engineering	61,941	158,931	280,772	501,644
Natural Science	111,158	140,126	242,879	494,163
Social Science	182,166	116,353	236,018	534,537
Total	355,265	415,410	759,669	1,530,344

Marginal Distribution

	US	Europe	Asia	
Engineering	4.05%	10.39%	18.35%	32.78%
Natural Science	7.26%	9.16%	15.87%	32.29%
Social Science	11.90%	7.60%	15.42%	34.93%
	23.21%	27.14%	49.64%	100.00%

Conditional on Region

	US	Europe	Asia
Engineering	17.44%	38.26%	36.96%
Natural Science	31.29%	33.73%	31.97%
Social Science	51.28%	28.01%	31.07%
	100.00%	100.00%	100.00%

Conditional on Major

	US	Europe	Asia	
Engineering	12.35%	31.68%	55.97%	100.00%
Natural Science	22.49%	28.36%	49.15%	100.00%
Social Science	34.08%	21.77%	44.15%	100.00%

Death Penalty and Race

- Examination of 326 death penalty cases
- About half involve a white defendant
- About half involve a black defendant
- Every defendant was convicted of killing someone.

	Death Penalty		
Defendant	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

	Death Penalty		
Defendant	Yes	No	Total
White	11.87	88.13	100%
Black	10.24	89.76	100%
Total	11.04	88.95	100%

Death Penalty

- The probability of getting the death penalty appears to be about the same for whites and blacks.
- But what if we look at the data more carefully?
- Is killing a black person the same as killing a white person?

White Defendant	Death Penalty		
	Yes	No	Total
White Victim	19	132	151
Black Victim	0	9	9
Total	19	141	160

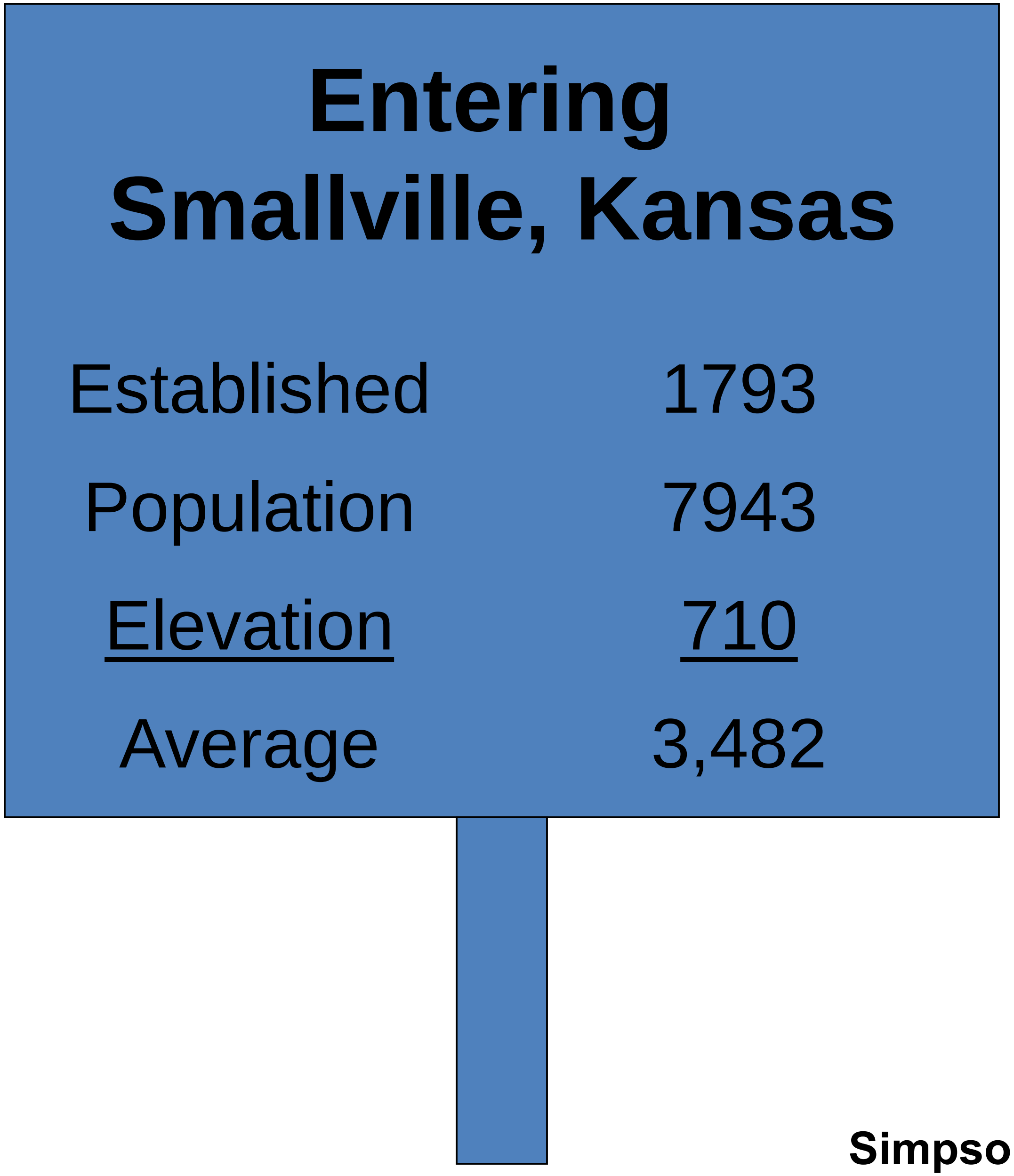
Black Defendant	Death Penalty		
	Yes	No	Total
White Victim	11	52	63
Black Victim	6	97	103
Total	17	149	166

White Defendant	Death Penalty		
	Yes	No	Total
White Victim	12.5	87.5	100%
Black Victim	0	100	100%

Black Defendant	Death Penalty		
	Yes	No	Total
White Victim	17.4	82.5	100%
Black Victim	5.8	94.2	100%

Simpson’s Paradox

- Blacks are more likely to get the death penalty
- But the overall probability of getting the death penalty looked to be the same
- You are more likely to get the death penalty for killing a white person
- White are more likely to kill other whites
- Blacks are more likely to kill other blacks



Entering
Smallville, Kansas

Established	1793
Population	7943
Elevation	710
Average	3,482

Batting Average

MLB Batting Averages

	1995		**1996**		**Combined**	
Derek Jeter	12/48	0.25	183/582	0.314	195/630	**0.31**
David Justice	104/411	**0.253**	45/140	**0.321**	149/551	0.27

Batting Averages

	1995		**1996**		**1997**		**Combined**	
Derek Jeter	12/48	0.25	183/582	0.314	190/654	0.291	385/1284	**0.3**
David Justice	104/411	**0.253**	45/140	**0.321**	163/495	**0.329**	312/1046	0.298

College Admissions

Admission to UC Berkeley

	Applicants	**% admitted**
Men	8442	**44%**
Women	4321	35%

College Admissions

Major	**Men**		**Women**	
	Applicants	% admitted	Applicants	% admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Kidney Stones

Kidney Stone Treatment

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Kidney Stones

	Treatment A	**Treatment B**
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Statistics

- Statistics
  - The collection, organization, presentation, analysis and interpretation of numerical facts and data
- Descriptive Statistics
  - The collection, organization and presentation of data (summarizing and describing a given data set)
- Inferential Statistics
  - The way we draw general conclusions about the phenomena under consideration, beyond the facts of the observed data
    - Deriving rational decisions from incomplete data
    - Wise decision making in the face of uncertainty

Inferential Statistics

- Population
  - Total set of observations on measurements or outcomes.
  - Size of the population can be finite or infinite.
- Sample
  - Set of measurements or outcomes selected from a population
  - Some samples are created by the experimenter, but most samples in economics are created by “nature.”

Statistical Inference

- How we generalize the sample characteristics to the entire population?
  - How certain are we that the implications are true?
  - How certain are we that a given theory is true?
- Since samples are drawn randomly, we need to understand some things about randomness and thus probability.

Probability

- **Gerolamo Cardano**
  - **1501-1576**
- *Book of Games of Chance*
  - *1526 (published 1663)*

- Gambler, effective cheating methods