DATA SCIENCE WITH PYTHON
CS 498 (SPECIAL TOPICS IN COMPUTER SCIENCE)

BASIC STATISTICS

Dr. Malak Abdullah
Computer Science Department
Jordan University of Science and Technology

WHY STUDY STATISTICS?

1. ***Data are everywhere***
2. ***Statistical techniques are used to make many decisions that affect our lives***
3. ***No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively***

WHAT IS STATISTICS?

1. Collecting Data
   e.g., Survey
2. Presenting Data
   e.g., Charts & Tables
3. Characterizing Data
   e.g., Average

Data Analysis

Why?

Decision-Making

TWO PROCESSES

- Describing sets of data

and

- Drawing conclusions (making estimates, decisions, predictions, etc. about sets of data based on sampling)

APPLICATION AREAS

- Economics
  - Forecasting
  - Demographics
- Engineering
  - Construction
  - Materials
- Sports
  - Individual & Team Performance
- Business
  - Consumer Preferences
  - Financial Trends

STATISTICS

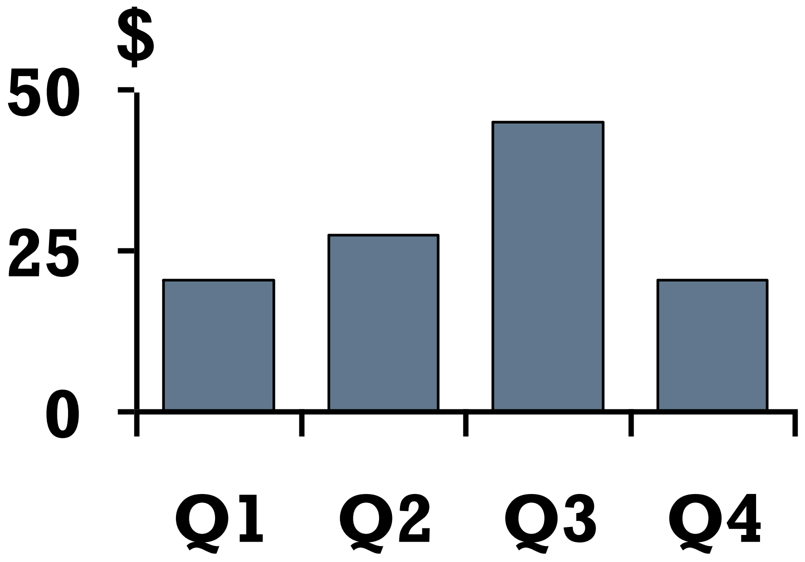
- The science of collectiong, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions
- Statistical analysis – used to manipulate summarize, and investigate data, so that useful decision-making information results.

TWO AREAS OF STATISTICS:

- **Descriptive Statistics**: collection, presentation, and description of sample data.
- **Inferential Statistics**: making decisions and drawing conclusions about populations.

DESCRIPTIVE STATISTICS

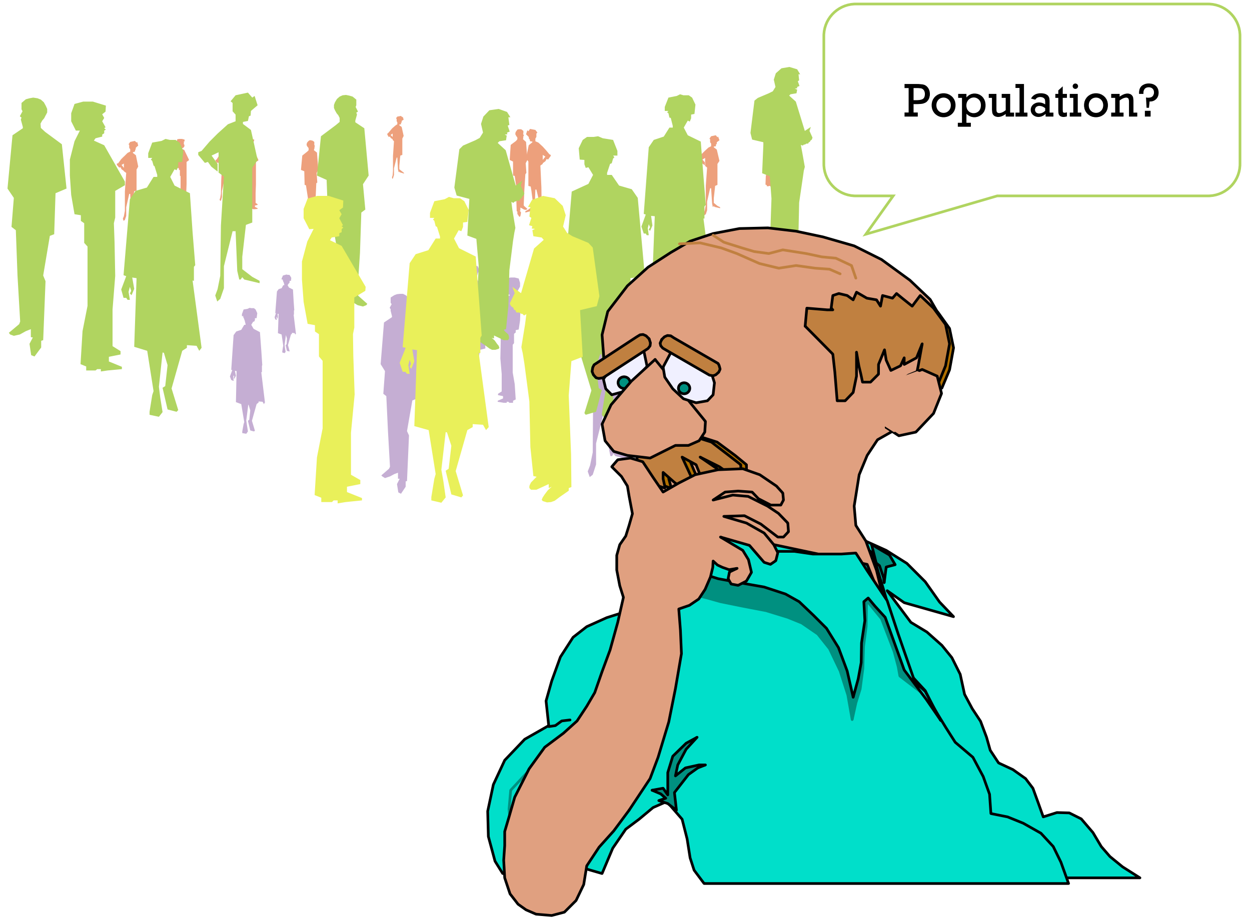
1. Involves
   - Collecting Data
   - Presenting Data
   - Characterizing Data
2. Purpose
   - Describe Data



$\bar{X} = 30.5 \quad S^2 = 113$

INFERENTIAL STATISTICS

1. Involves
   - Estimation
   - Hypothesis Testing
2. Purpose
   - Make decisions about population characteristics



TYPES OF STATISTICS/ANALYSES

Descriptive Statistics

- Frequencies
- Basic measurements

Describing a phenomena

How many? How much?

BP, HR, BMI, IQ, etc.

Inferential Statistics

- Hypothesis Testing
- Correlation
- Confidence Intervals
- Significance Testing
- Prediction

Inferences about a phenomena

Proving or disproving theories

Associations between phenomena

If sample relates to the larger population

E.g., Diet and health

Example: A recent study examined the math and verbal SAT scores of high school seniors across the country. Which of the following statements are descriptive in nature and which are inferential.

- The mean math SAT score was 492.
- The mean verbal SAT score was 475.
- Students in the Northeast scored higher in math but lower in verbal.
- 80% of all students taking the exam were headed for college.
- 32% of the students scored above 610 on the verbal SAT.
- The math SAT scores are higher than they were 10 years ago.

INFERENTIAL STATISTICS

Inferential statistics can be used to prove or disprove theories, determine associations between variables, and determine if findings are significant and whether or not we can generalize from our sample to the entire population

The types of inferential statistics we will go over:

- Correlation
- T-tests/ANOVA
- Chi-square
- Logistic Regression

INFERENTIAL STATISTICS

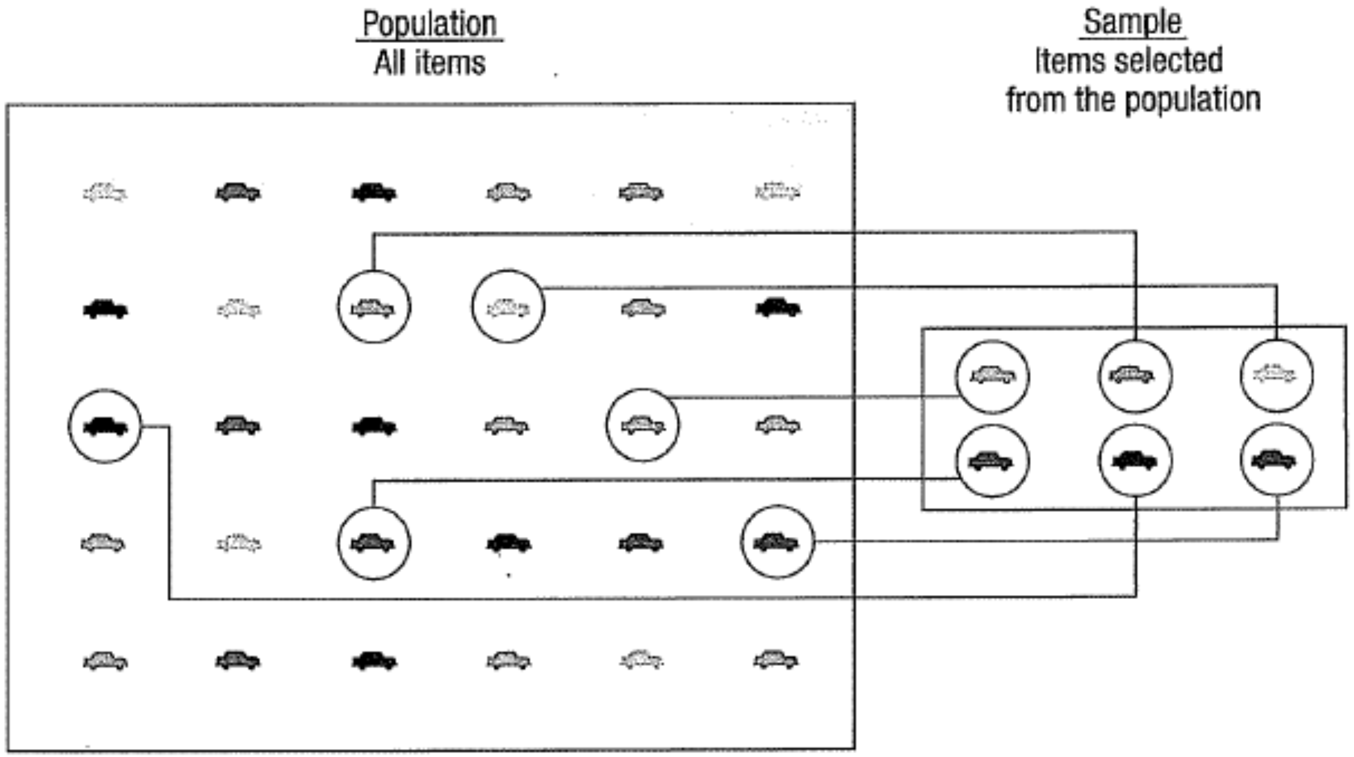
- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results

INTRODUCTION TO BASIC TERMS

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Sample: A subset of the population.



Statistical inference is the process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population

PROCESS OF STATISTICAL INFERENCE: EXAMPLE

1. Population consists of heights of all **JUST students.**

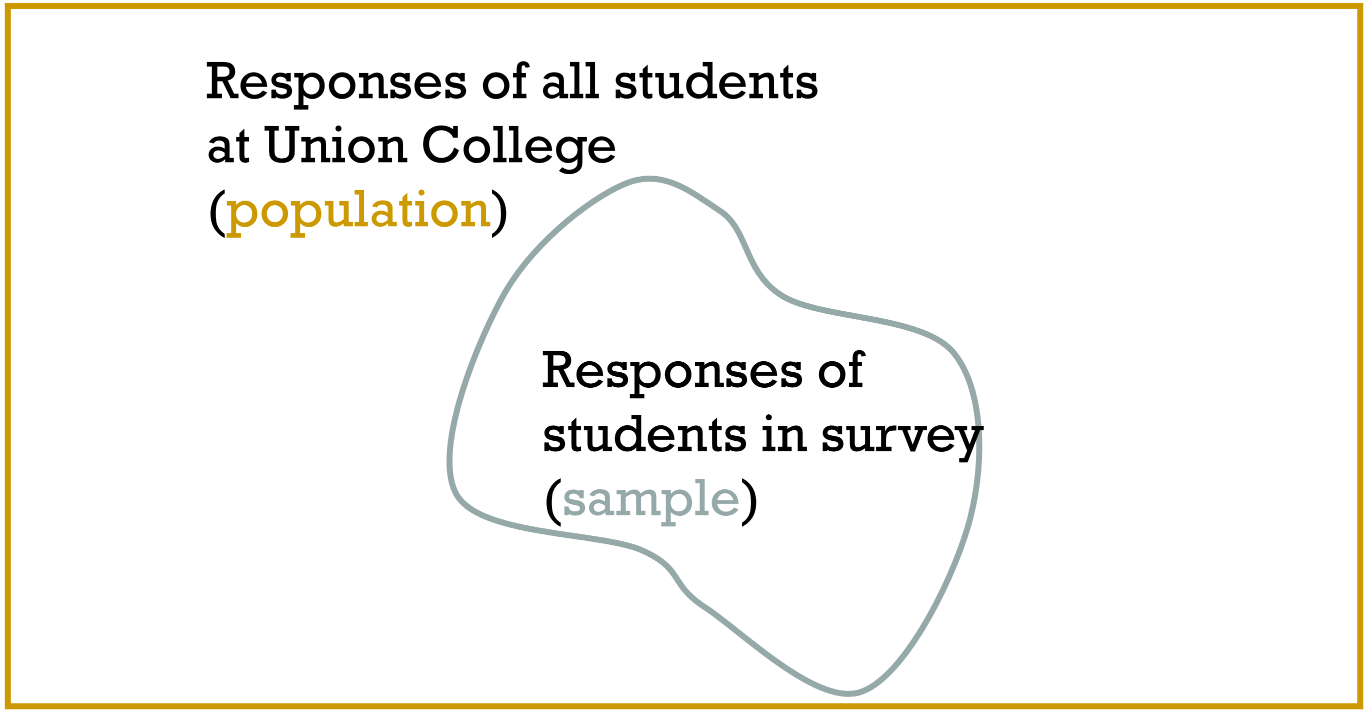
2. A sample of 25 students are randomly selected and measured.

3. The sample data provide a sample average height of 5′ 5″.

4. The sample average is used to estimate the <u>population average</u>.

POPULATIONS & SAMPLES

- **Example**:
- In a recent survey, 250 college students at Union College were asked if they smoked cigarettes regularly. 35 of the students said yes. Identify the population and the sample.



SAMPLING METHODS

Sampling methods can be:

- **random** (each member of the population has an equal chance of being selected)
- **nonrandom**

The actual process of sampling causes **sampling errors.** For example, the sample may not be large enough or representative of the population.

Back to Basic Terms

Variable: A characteristic about each individual element of a population or sample.

Data (singular): The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

Data (plural): The set of values collected for the variable from each of the elements belonging to the sample.

Statistic: A numerical value summarizing the sample data.

Example: A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the "age" of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *statistic* is the "average" age for all faculty in the sample.

Two kinds of variables:

Qualitative, or Attribute, or Categorical, Variable: A variable that categorizes or describes an element of a population (hair color, Department).

Quantitative, or Numerical, Variable: A variable that quantifies an element of a population (height and weight, number of children in a family).

QUALITATIVE DATA

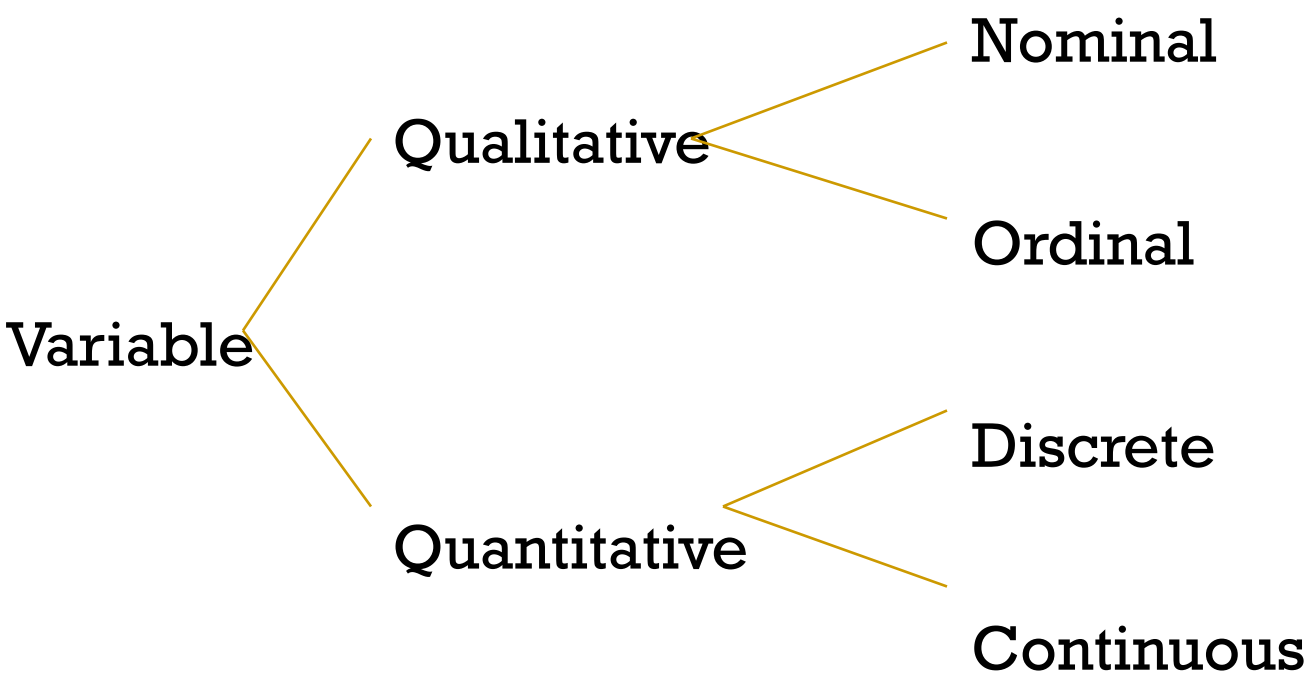
Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

QUANTITATIVE DATA

Quantitative data are always numbers and are the **result of counting or measuring** attributes of a population. Quantitative data can be separated into two subgroups:

- **discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf, ...)
- **continuous** (if it is the result of *measuring* (distance traveled, weight of luggage, …)

Qualitative and quantitative variables may be further subdivided:



QUALITATIVE AND QUANTITATIVE DATA

- **Example**:
- The grade point averages of five students are listed in the table. Which data are qualitative data and which are quantitative data?

Student	GPA
Sally	3.22
Bob	3.98
Cindy	2.75
Mark	2.24
Kathy	3.84

Qualitative data

Quantitative data

Nominal Variable: A qualitative variable that categorizes (or describes, or names) an element of a population (hair color).

Ordinal Variable: A qualitative variable that incorporates an ordered position, or ranking (level of happiness).

Discrete Variable: A quantitative variable that can assume a countable number of values (number of students in each class).

Continuous Variable: A quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values (salary of an employee).

Note:

1. In many cases, a discrete and continuous variable may be distinguished by determining whether the variables are related to a count or a measurement.
2. Discrete variables are usually associated with counting.
3. Continuous variables are usually associated with measurements. The values of discrete variables are only limited by your ability to measure them.

DESCRIPTIVE STATISTICS

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

DESCRIPTIVE STATISTICS

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

MEAN

Most commonly called the "average."

Add up the values for each case and divide by the total number of cases.

$$\text{Y-bar} = \frac{(Y1 + Y2 + \ldots + Yn)}{n}$$

$$\text{Y-bar} = \frac{\Sigma\, Y\mathit{i}}{n}$$

MEAN

Class A--IQs of 13 Students	
102	115
128	109
131	89
98	106
140	119
93	97
110	

$\Sigma\, Y_i = 1437$

$\bar{Y}_A = \frac{\Sigma\, Y_i}{n} = \frac{1437}{13} = 110.54$

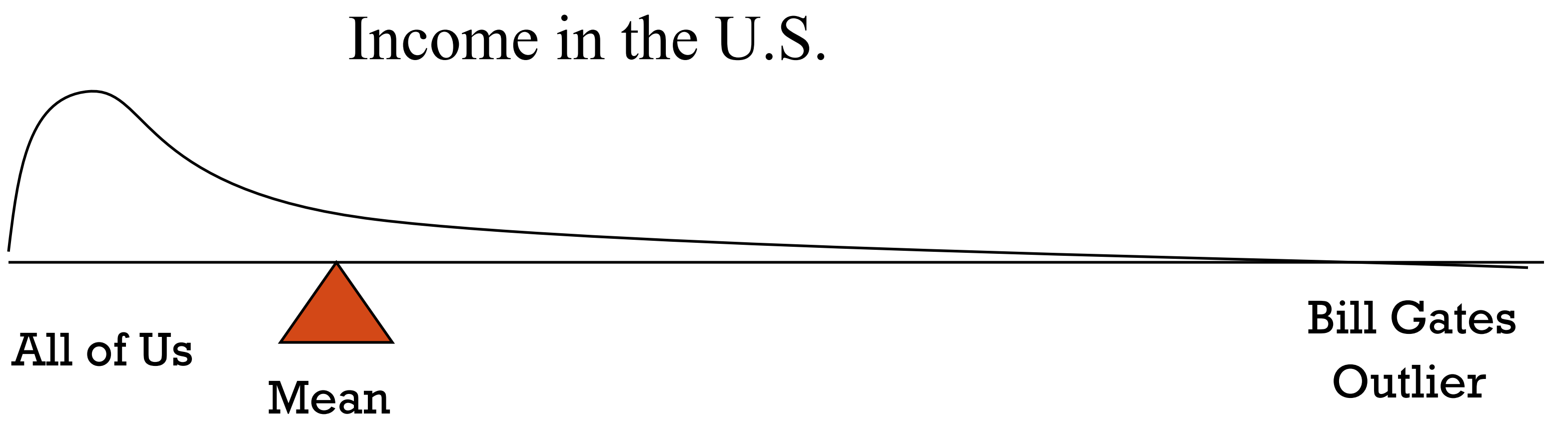
Class B--IQs of 13 Students	
127	162
131	103
96	111
80	109
93	87
120	105
109	

$\Sigma\, Y_i = 1433$

$\bar{Y}_B = \frac{\Sigma\, Y_i}{n} = \frac{1433}{13} = 110.23$

MEAN

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience



MEDIAN

The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

The 50^{th} percentile.

MEDIAN

Class A--IQs of 13 Students

89

93

97

98

102

106

109

110

115

119

128

131

140

Median = 109

(six cases above, six below)

MEDIAN

If the first student were to drop out of Class A, there would be a new median:

~~89~~
93
97
98
102
106
109
110
115
119
128
131
140

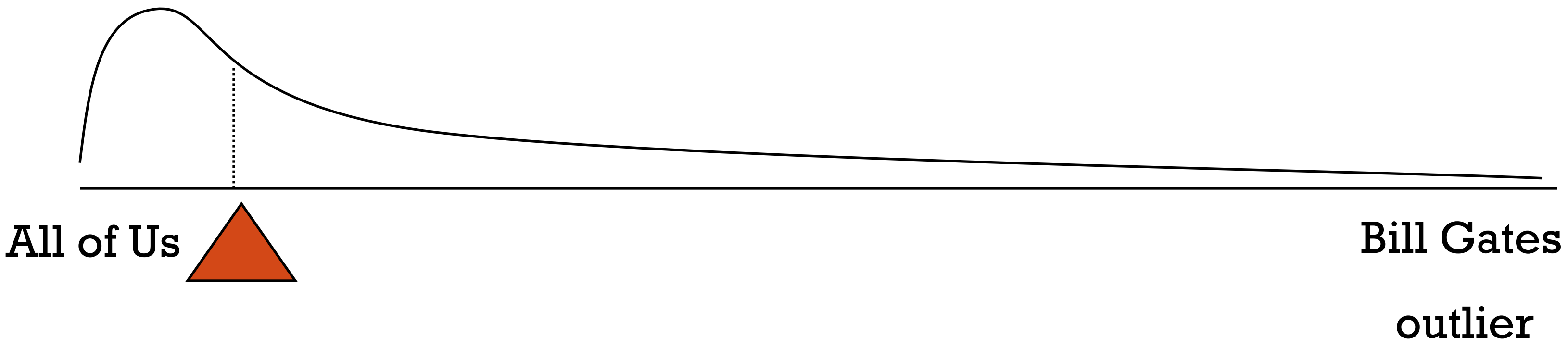
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

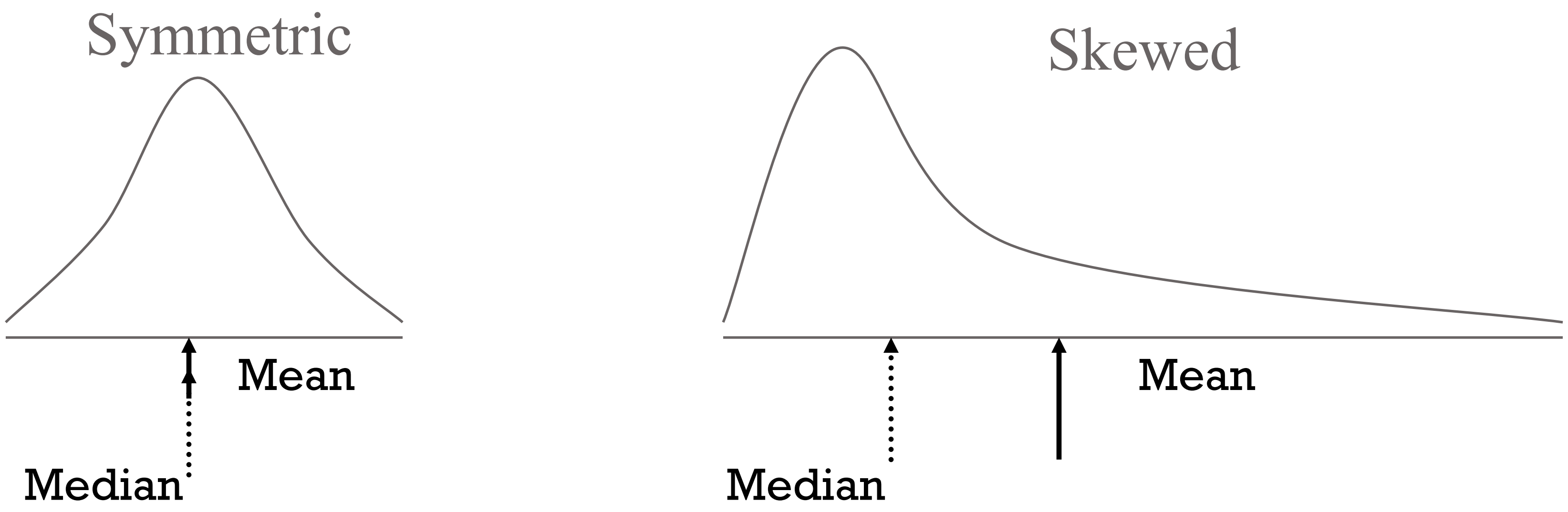
MEDIAN

1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.



MEDIAN

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.



MODE

The most common data point is called the mode.

The combined IQ scores for Classes A & B:

80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120

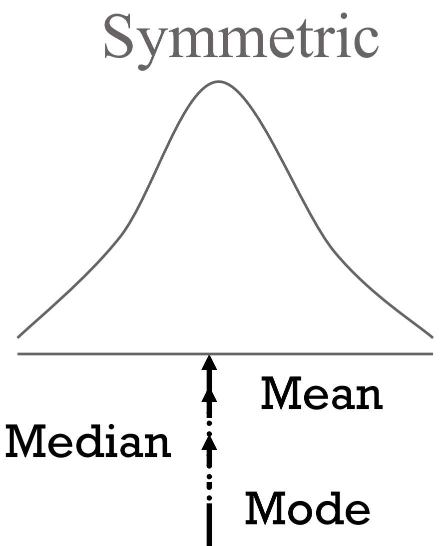
127 128 131 131 140 162

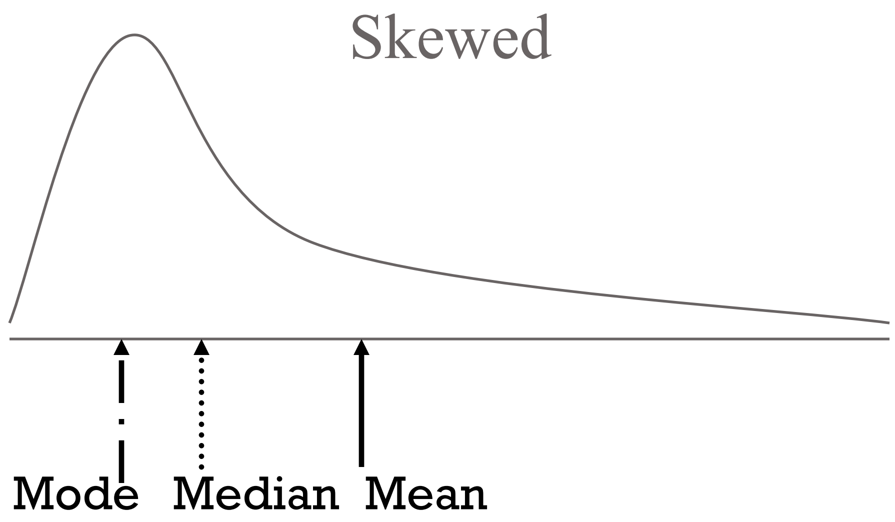
A la mode!!

BTW, It is possible to have more than one mode!

MODE

1. It may give you the most likely experience rather than the "typical" or "central" experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.





DESCRIPTIVE STATISTICS

Summarizing Data:

- ✓ Central Tendency (or Groups' "Middle Values")
  - ✓ Mean
  - ✓ Median
  - ✓ Mode

  https://repl.it/@DrMalak/CentralTendencies

- ❖ Variation (or Summary of Differences Within Groups)
  - ❖ Range
  - ❖ Interquartile Range
  - ❖ Variance
  - ❖ Standard Deviation

RANGE

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	
Class A Range = 140 - 89 = 51		**Class B Range = 162 - 80 = 82**	

INTERQUARTILE RANGE

A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?

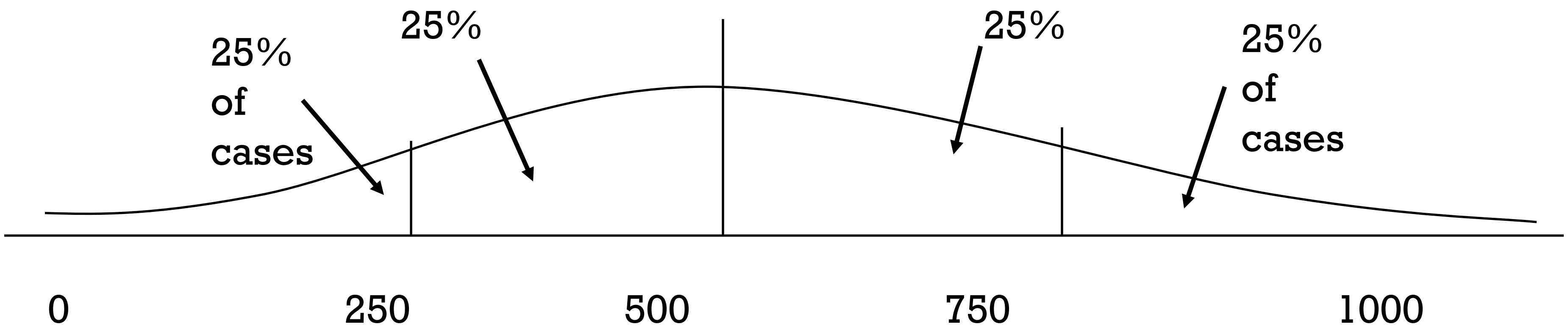
0 quartile = 0 quantile = 0 percentile

1 quartile = 0.25 quantile = 25 percentile

2 quartile = .5 quantile = 50 percentile (median)

3 quartile = .75 quantile = 75 percentile

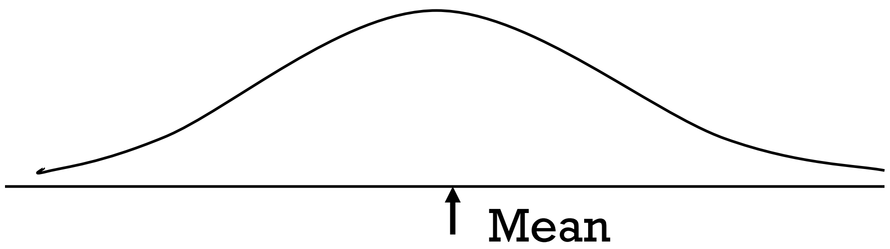
4 quartile = 1 quantile = 100 percentile



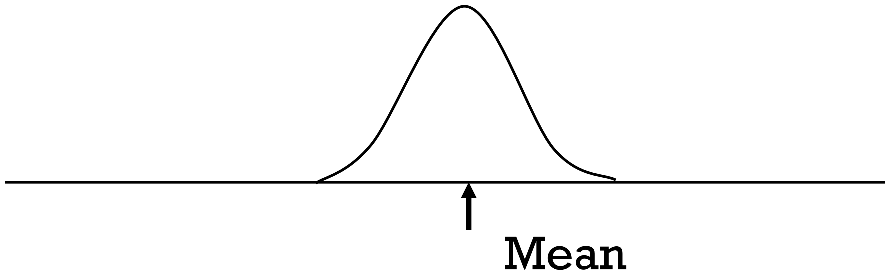
VARIANCE

A measure of the spread of the recorded values on a variable. A measure of dispersion.

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



VARIANCE

Variance is a number that at first seems complex to calculate.

Calculating variance starts with a "deviation."

A deviation is the distance away from the mean of a case's score.

Y*i* – Y-bar

If the average person's car costs $20,000, my deviation from the mean is - $14,000!

6K - 20K = -14K

VARIANCE

The deviation of 102 from 110.54 is? Deviation of 115?

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar$_A$ *= 110.54*

VARIANCE

The deviation of 102 from 110.54 is?

102 - 110.54 = -8.54

Deviation of 115?

115 - 110.54 = 4.46

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

$\text{Y-bar}_A = 110.54$

VARIANCE

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.

Squaring the deviations will eliminate negative signs...

A Deviation Squared: $(Y_i - \text{Y-bar})^2$

Back to the IQ example,

A deviation squared for 102 is: of 115:

$(102 - 110.54)^2 = (-8.54)^2 = 72.93$ $(115 - 110.54)^2 = (4.46)^2 = 19.89$

VARIANCE

If you were to add all the squared deviations together, you'd get what we call the "Sum of Squares."

Sum of Squares (SS) = $\Sigma\,(Y\mathit{i} - \text{Y-bar})^2$

$SS = (Y1 - \text{Y-bar})^2 + (Y2 - \text{Y-bar})^2 + \ldots + (Yn - \text{Y-bar})^2$

VARIANCE

Class A, sum of squares:

$(102 – 110.54)^2 + (115 – 110.54)^2 +$
$(126 – 110.54)^2 + (109 – 110.54)^2 +$
$(131 – 110.54)^2 + (89 – 110.54)^2 +$
$(98 – 110.54)^2 + (106 – 110.54)^2 +$
$(140 – 110.54)^2 + (119 – 110.54)^2 +$
$(93 – 110.54)^2 + (97 – 110.54)^2 +$
$(110 – 110.54) = SS = 2825.39$

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar = 110.54

VARIANCE

The last step…

The approximate average sum of squares is the variance.

SS/N = Variance for a population.

SS/n-1 = Variance for a sample.

Variance = $\Sigma(Y_i - \text{Y-bar})^2 / n - 1$

VARIANCE

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

How helpful is that???

STANDARD DEVIATION

To convert variance into something of meaning, let's create standard deviation.

The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\frac{\Sigma(\mathrm{Y}i - \text{Y-bar})^2}{n - 1}}$$

STANDARD DEVIATION

For Class A, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

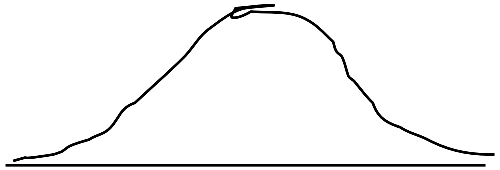
Review:

1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

STANDARD DEVIATION

1. Larger s.d. = greater amounts of variation around the mean.

   For example:



19 25 31

Y = 25

s.d. = 3

13 25 37

Y = 25

s.d. = 6

2. s.d. = 0 only when all values are the same (only when you have a constant and not a "variable")
3. If you were to "rescale" a variable, the s.d. would change by the same magnitude—if we changed units above so the mean equaled 250, the s.d. on the left would be 30, and on the right, 60
4. Like the mean, the s.d. will be inflated by an outlier case value.

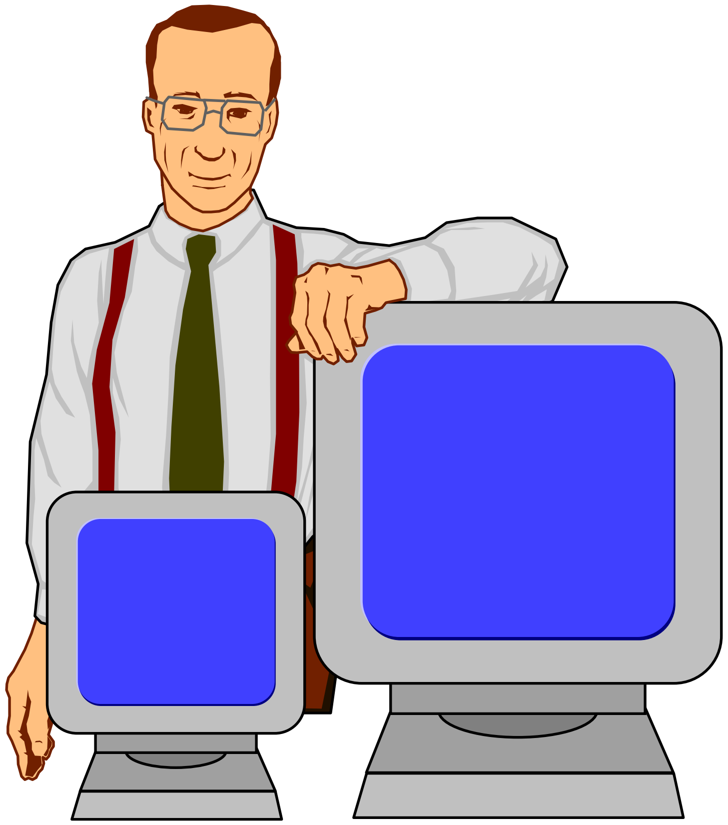
STATISTICAL COMPUTER PACKAGES

1. Typical Software
   - SPSS
   - MINITAB
   - Excel

2. Need Statistical Understanding
   - Assumptions
   - Limitations

```
import statistics

lst=[90,90,70,77,80,88,76,56,34,99,23,30]
print (lst)

print (statistics.mean(lst))
print (statistics.median(lst))
print (statistics.mode(lst))
print (statistics.stdev(lst))
print (statistics.variance(lst))
```

```
[90, 90, 70, 77, 80, 88, 76, 56, 34, 99, 23, 30]
67.75
76.5
90
25.919894079469326
671.8409090909091
```



BOX-PLOTS

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

A box-plot shows: Upper and lower quartiles

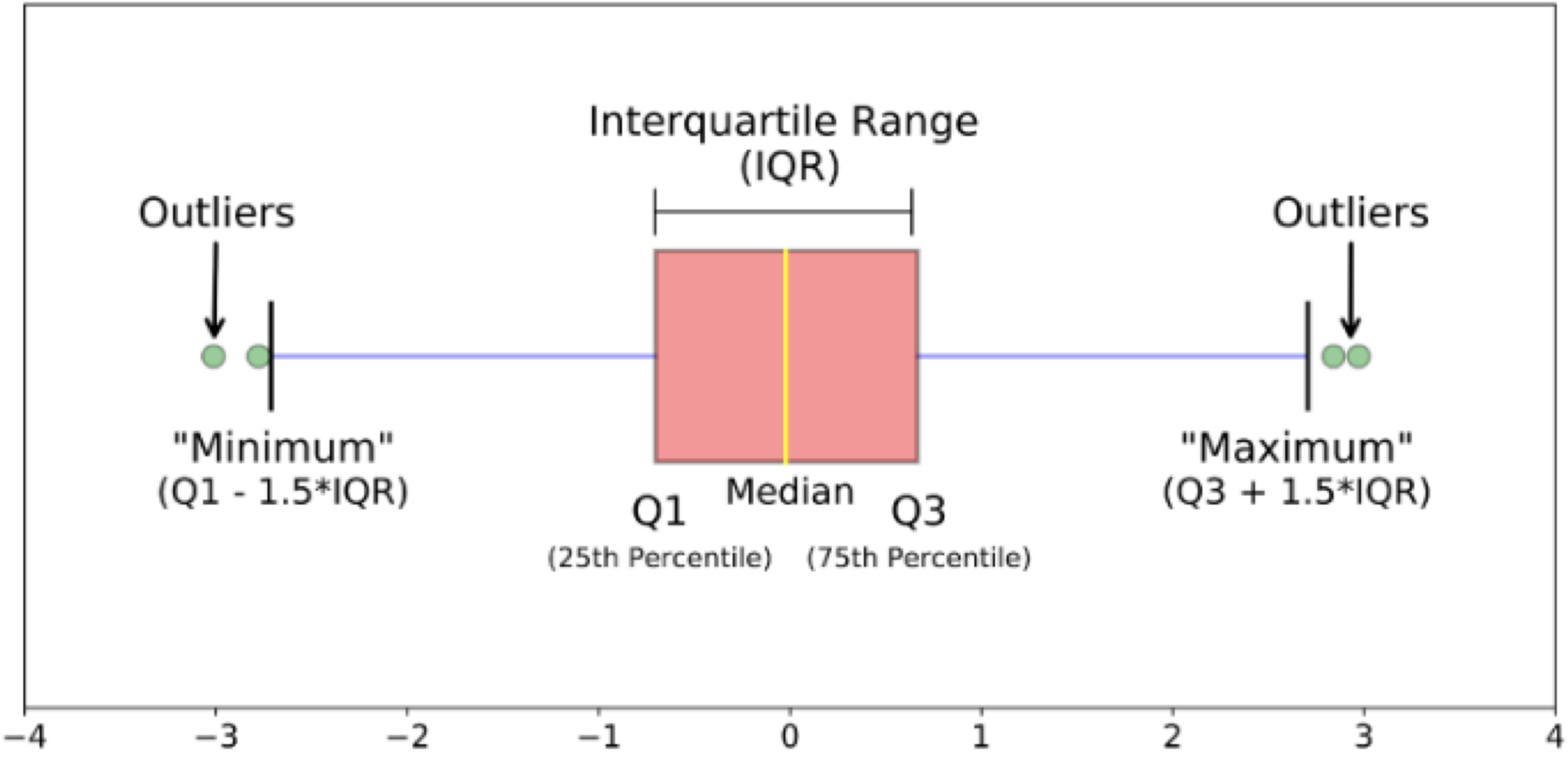
Mean

Median

Range

Outliers (1.5 IQR)

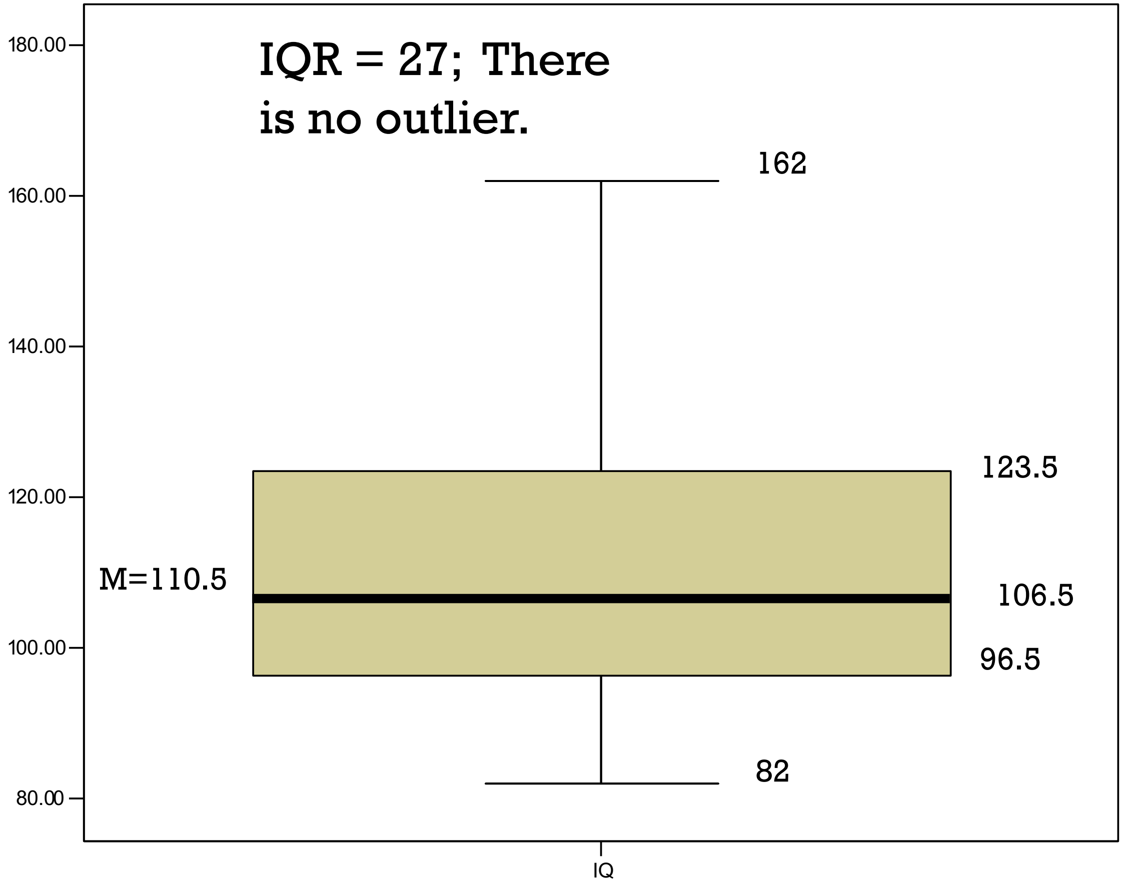
Understanding Boxplots

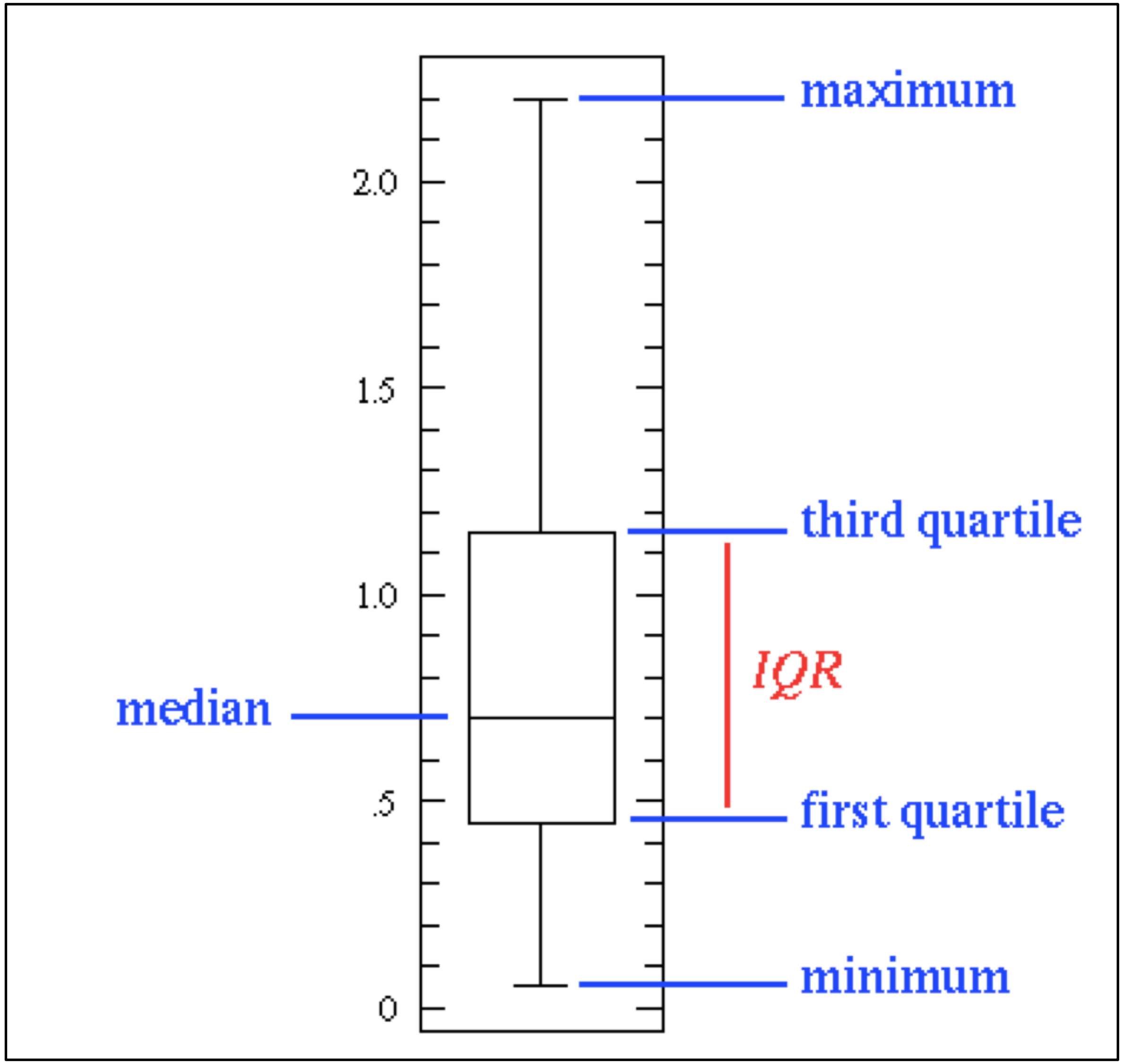


Different parts of a boxplot

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

BOX-PLOTS

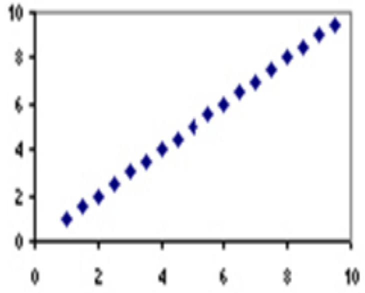

IQR = 27; There
is no outlier.
180.00
160.00
140.00
120.00
100.00
80.00
162
123.5
M=110.5
106.5
96.5
82
IQ


maximum
2.0
1.5
third quartile
1.0
IQR
median
.5
first quartile
minimum
0

CORRELATION

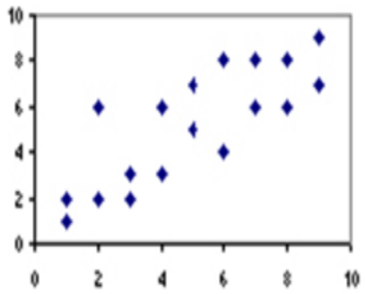
- A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.
- When to use it?
  - When you want to know about the association or relationship between two continuous variables
    - Ex) Taller people tend to be heavier.
    - If an **increase** in one **variable** tends to be associated with a **decrease** in the other then this is known as a **negative correlation**. An example would be height above **sea level** and temperature
- What does it tell you?
  - If a linear relationship exists between two variables, and how strong that relationship is
- What do the results look like?
  - The correlation coefficient = Pearson's *r*
  - Ranges from -1 to +1

CORRELATION



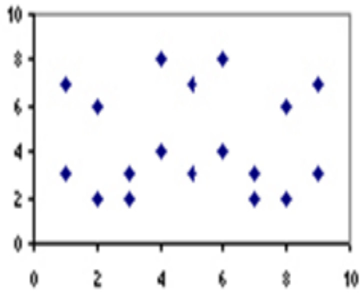
Maximum positive correlation

(r = 1.0)



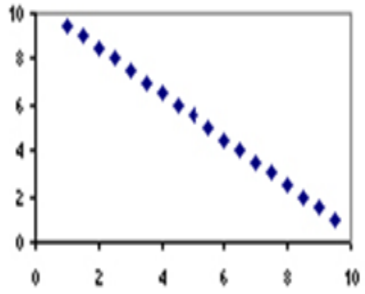
Strong positive correlation

(r = 0.80)



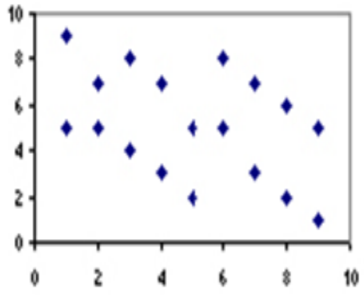
Zero correlation

(r = 0)



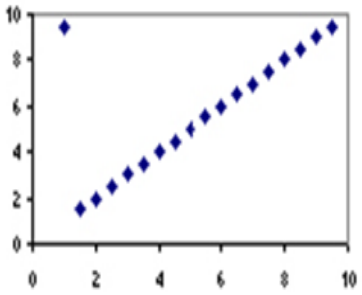
Minimum negative correlation

(r = -1.0)



Moderate negative correlation

(r = -0.43)



Strong correlation & outlier

(r = 0.71)

<u>Guide for interpreting strength of correlations:</u>

- 0 – 0.25 = Little or no relationship
- 0.25 – 0.50 = Fair degree of relationship
- 0.50 - 0.75 = Moderate degree of relationship
- 0.75 – 1.0 = Strong relationship
- 1.0 = perfect correlation

CORRELATION

- How do you interpret it?
  - If r is **positive**, high values of one variable are associated with high values of the other variable (both go in SAME direction - ↑↑ OR ↓↓)
    - Ex) Diastolic blood pressure tends to rise with age, thus the two variables are positively correlated
  - If r is **negative**, low values of one variable are associated with high values of the other variable (opposite direction - ↑↓ OR ↓ ↑)
    - Ex) Heart rate tends to be lower in persons who exercise frequently, the two variables correlate negatively
  - Correlation of 0 indicates NO linear relationship
- How do you report it?
  - "Diastolic blood pressure was positively correlated with age ($r = .75, p < .05$)."

Tip: Correlation does NOT equal causation!!! Just because two variables are highly correlated, this does NOT mean that one CAUSES the other!!!

EXAMPLES

- Positive
- The more time you spend running on a treadmill, the more calories you will burn.
- As the temperature goes up, ice cream sales also go up.
- As a student's study time increases, so does his test average.

- Negative
- A student who has many absences has a decrease in grades.
- As one exercises more, his body weight becomes less.

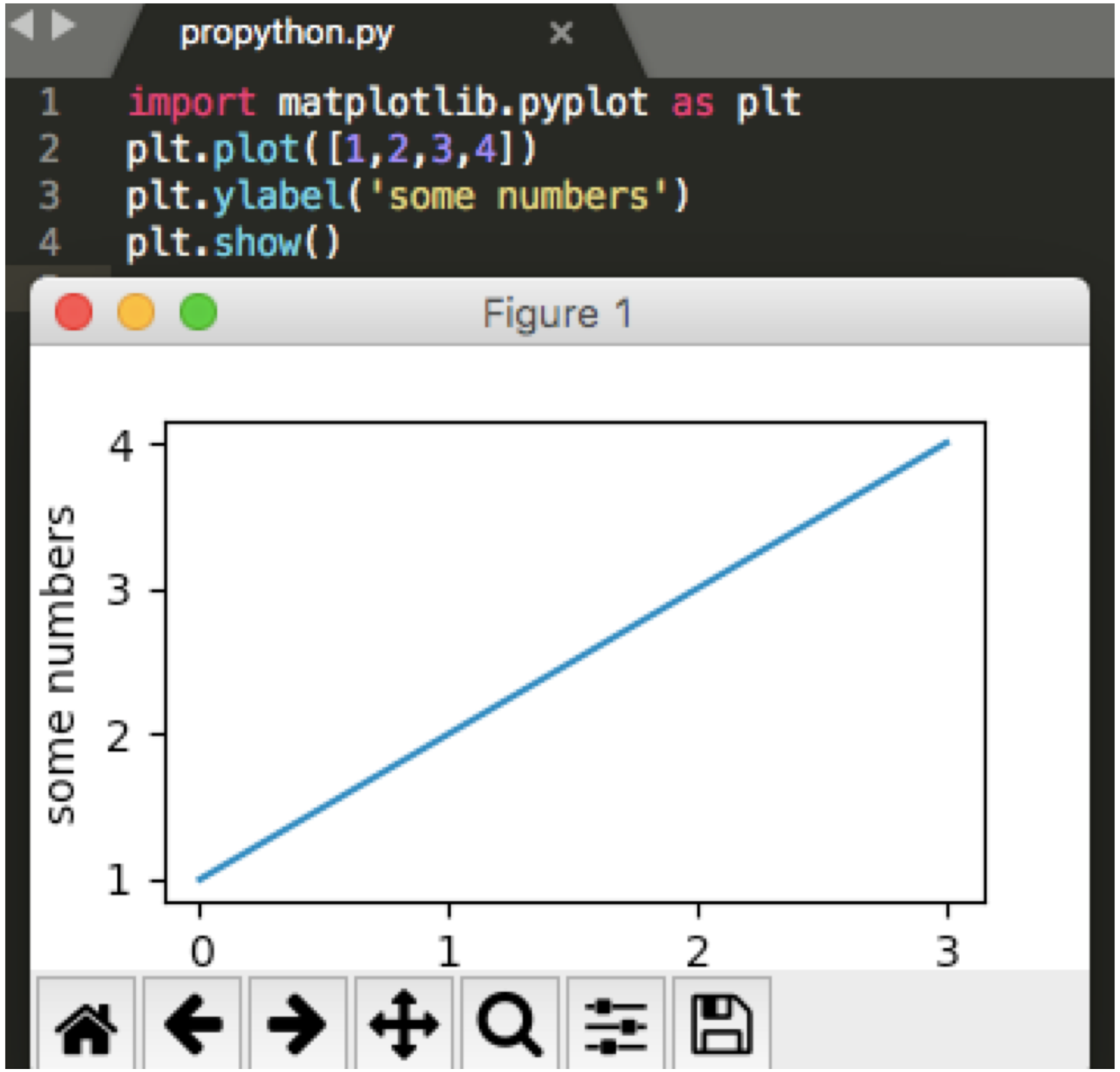
VISUALIZATION

New Chapter !!! ☺

IMPORT MATPLOTLIB.PYPLOT AS PLT

- matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot unction makes some change to a figure: eg, create a figure, create a plotting area in a figure, plot some lines in a plotting area, decorate the plot with labels, etc….

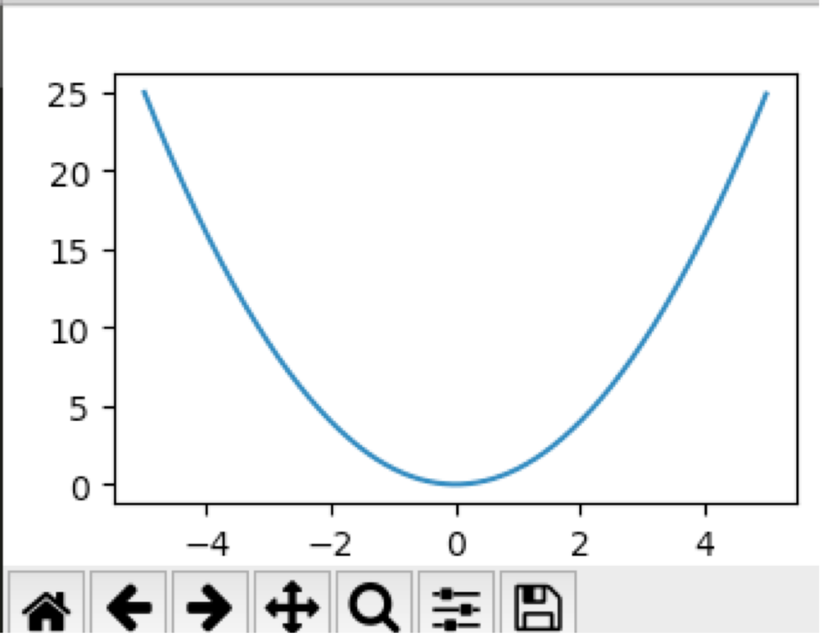
https://matplotlib.org/users/pyplot_tutorial.html

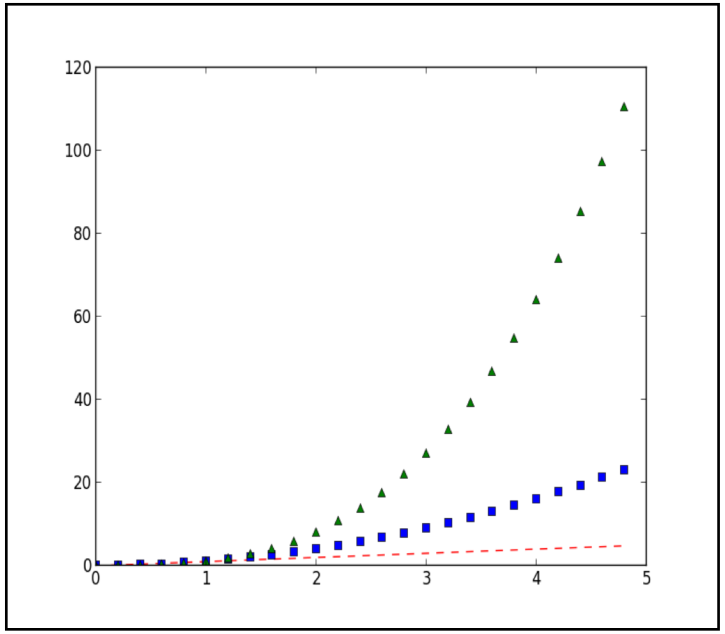


The required libraries are NumPy, SciPy and MatPlotLib. Installing PyLab is an easy way of setting up all the three libraries correctly.

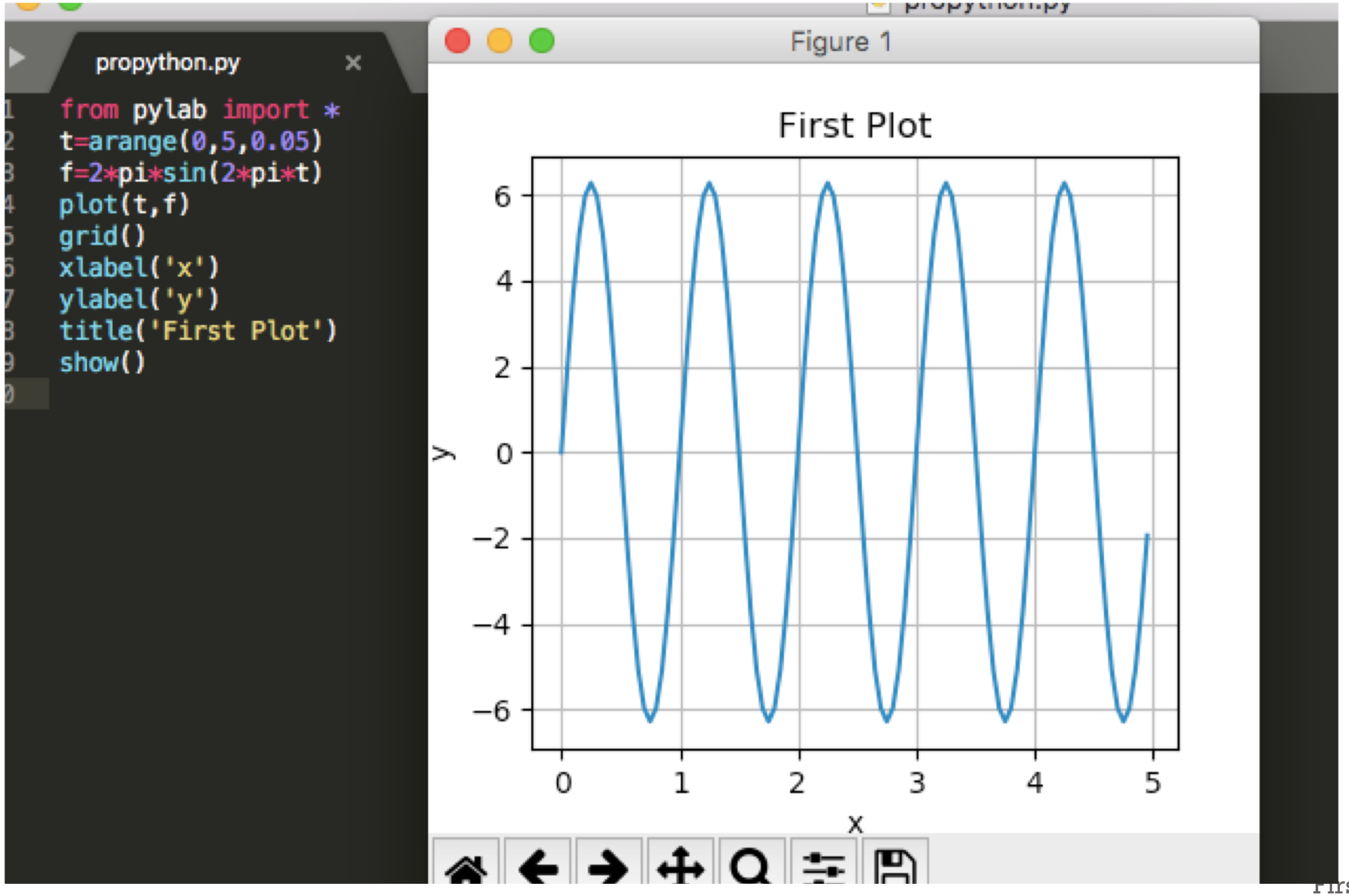
propython.py

```
from pylab import *
x = arange(-5.0, 5.0, 0.01)
y = x**2
plot(x, y)
show()


# evenly sampled time at 200ms intervals
t = np.arange(0., 5., 0.2)

# red dashes, blue squares and green triangles
plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^')
show()

```

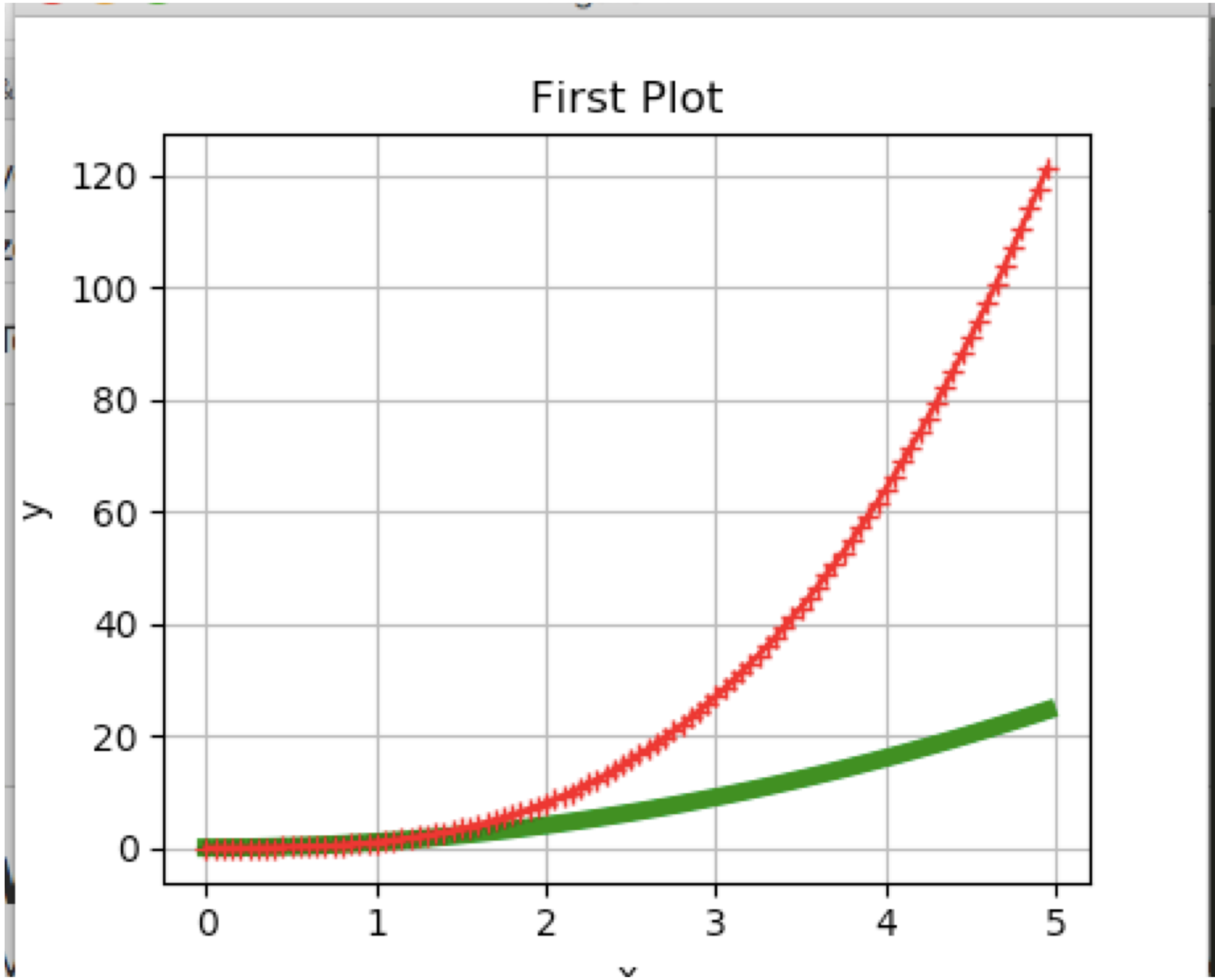




ANOTHER EXAMPLE



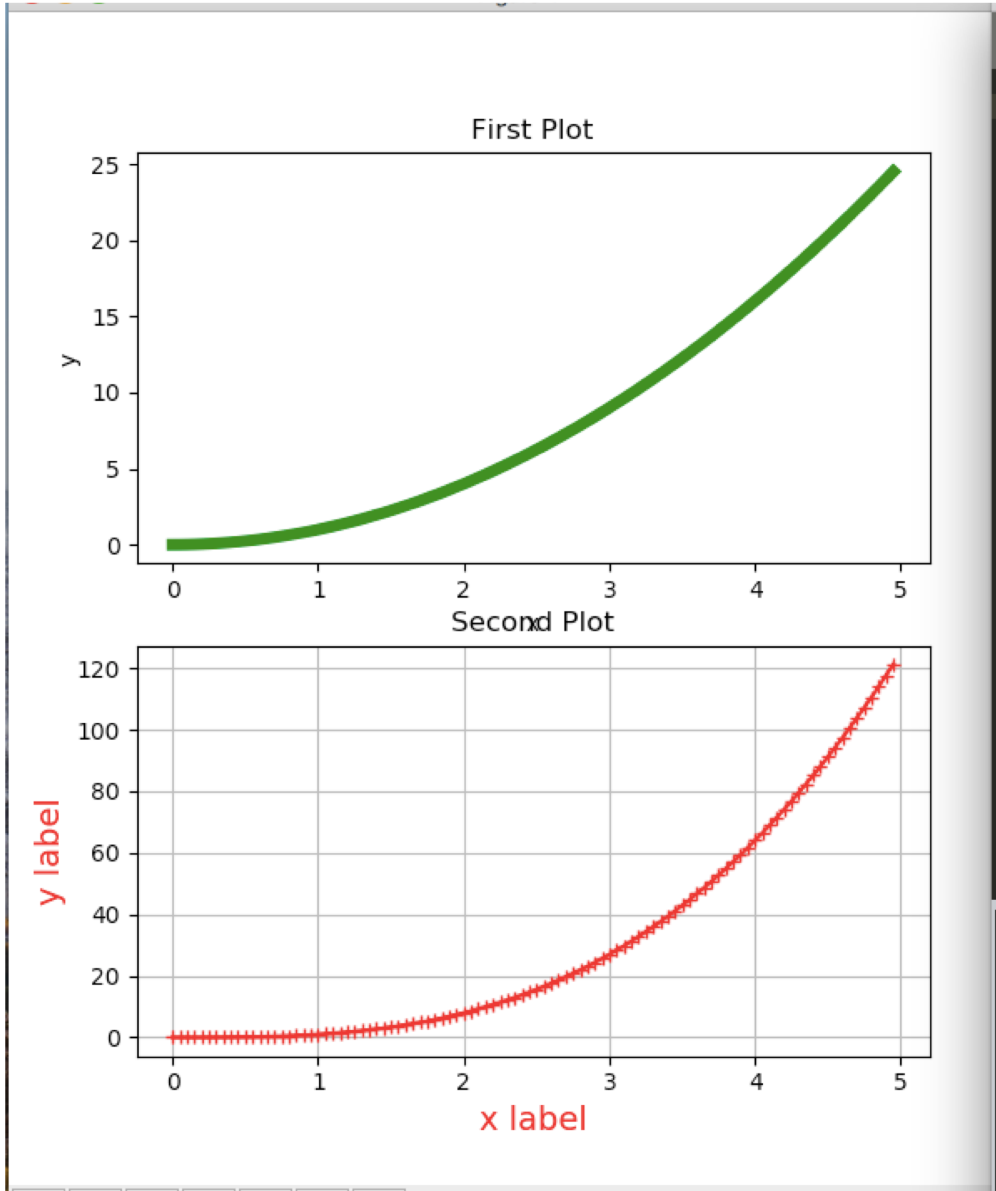
PLOTTING



propython.py

```
from pylab import *
t=arange(0,5,0.05)
f=t**2
z=t**3
plot(t,f, color='g' , linewidth=5.0)
plot(t,z, color='r', marker='+')
grid()
xlabel('x')
ylabel('y')
title('First Plot')
show()
```

WORKING WITH MULTIPLE FIGURES



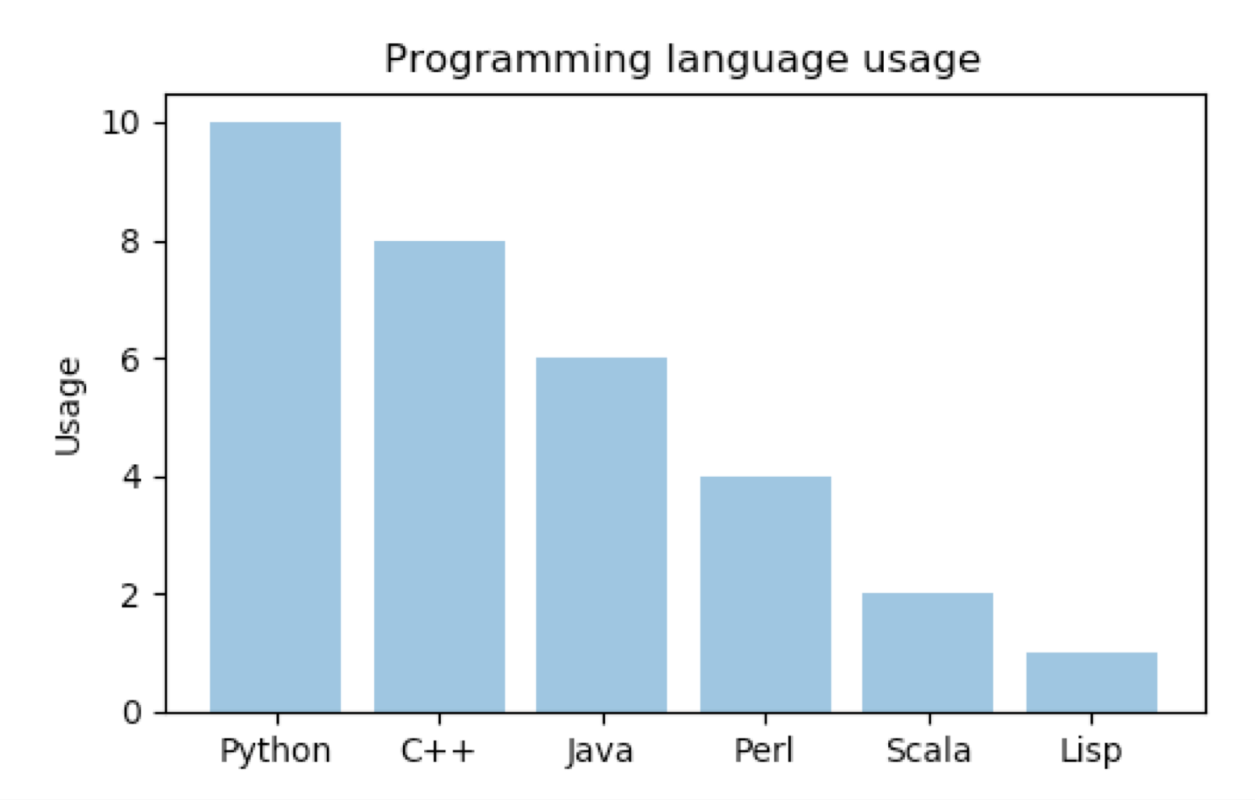
propython.py

```python
from pylab import *
t=arange(0,5,0.05)
f=t**2
z=t**3
subplot(211)
plot(t,f, color='g' , linewidth=5.0)
xlabel('x')
ylabel('y')
title('First Plot')
subplot(212)
plot(t,z, color='r', marker='+')
grid()
xlabel('x label', fontsize=14, color='red')
ylabel('y label', fontsize=14, color='red')
title('Second Plot')
show()
```

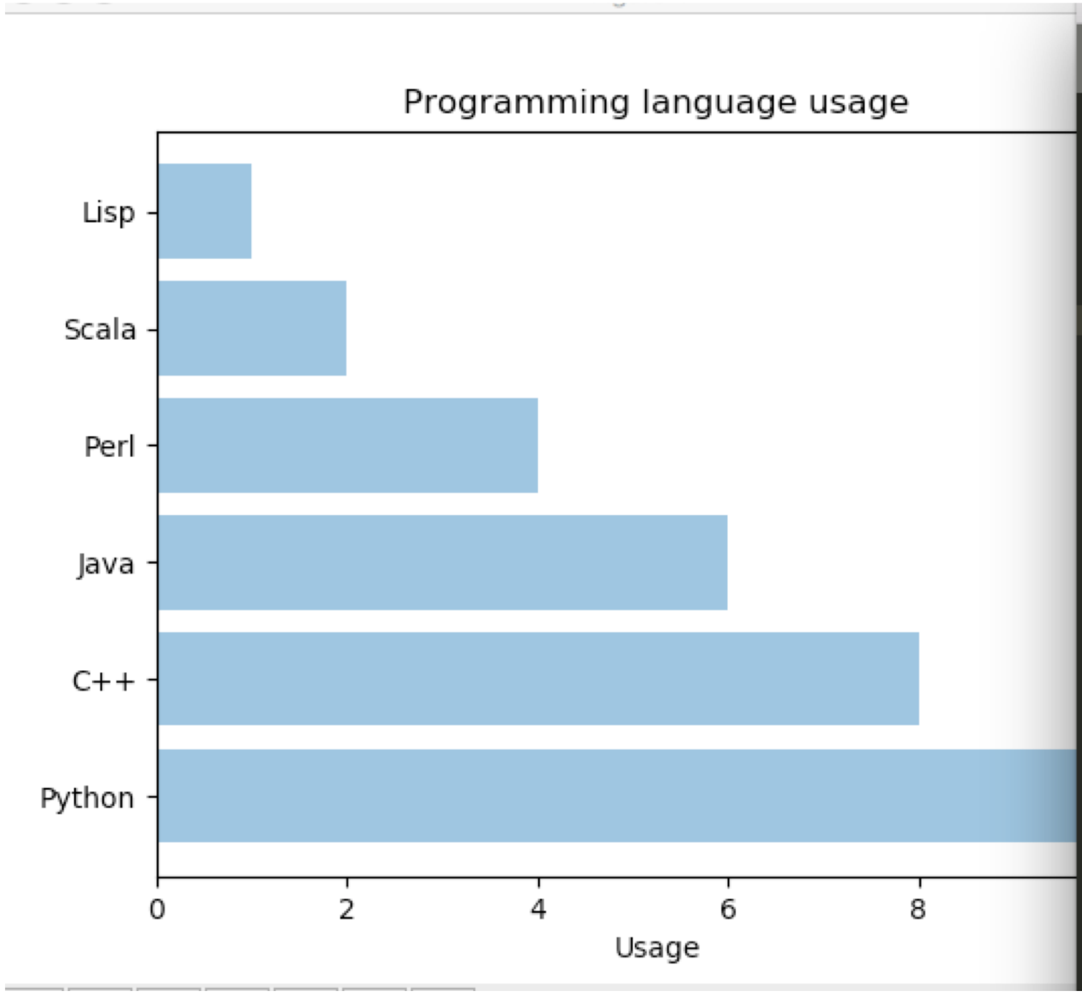
BARCHART

propython.py

```
from pylab import *
objects = ('Python', 'C++', 'Java', 'Perl', 'Scala', 'Lisp')
y_pos = arange(len(objects))
performance = [10,8,6,4,2,1]

bar(y_pos, performance, align='center', alpha=0.5)
xticks(y_pos, objects)
ylabel('Usage')
title('Programming language usage')

show()
```



BARCHART HORIZONTALLY



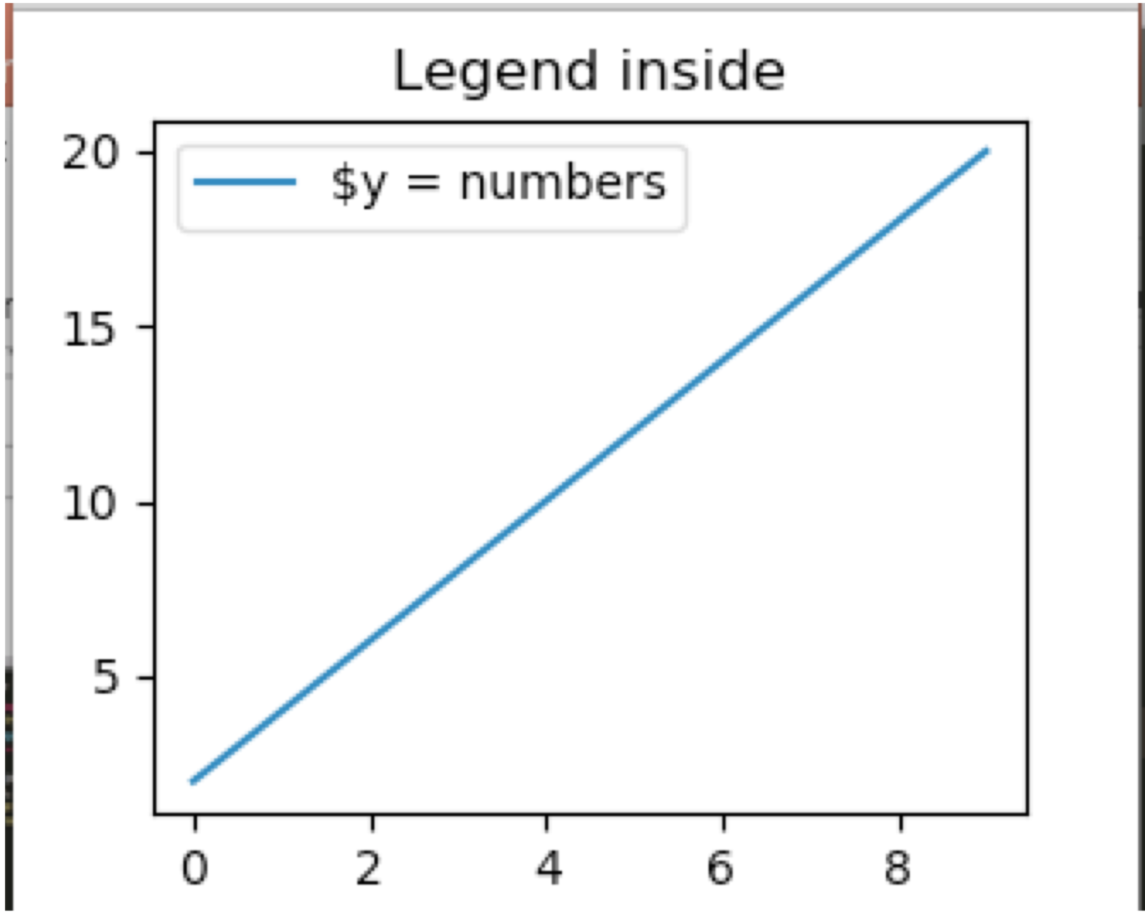
propython.py

```python
from pylab import *
objects = ('Python', 'C++', 'Java', 'Perl', 'Scala', 'Lisp')
y_pos = arange(len(objects))
performance = [10,8,6,4,2,1]

barh(y_pos, performance, align='center', alpha=0.5)
yticks(y_pos, objects)
xlabel('Usage')
title('Programming language usage')

show()
```

propython.py

```
from pylab import *
# data to plot
n_groups = 4
means_sami = (90, 55, 40, 65)
means_fadi = (85, 62, 54, 20)

# create plot
fig, ax = subplots()
index = arange(n_groups)
print (index)
bar_width = 0.35     # the width of the bars
opacity = 0.8  #light or dark color

rects1 = bar(index, means_sami, bar_width,
                 alpha=opacity,
                 color='b',
                 label='Sami')

rects2 = bar(index + bar_width, means_fadi, bar_width,
                 alpha=opacity,
                 color='g',
                 label='Fadi')

xlabel('Person')
ylabel('Scores')
title('Scores by person')
xticks(index + bar_width, ('A', 'B', 'C', 'D'))
legend()

tight_layout()
show()
```

SAVE THE FIGURE



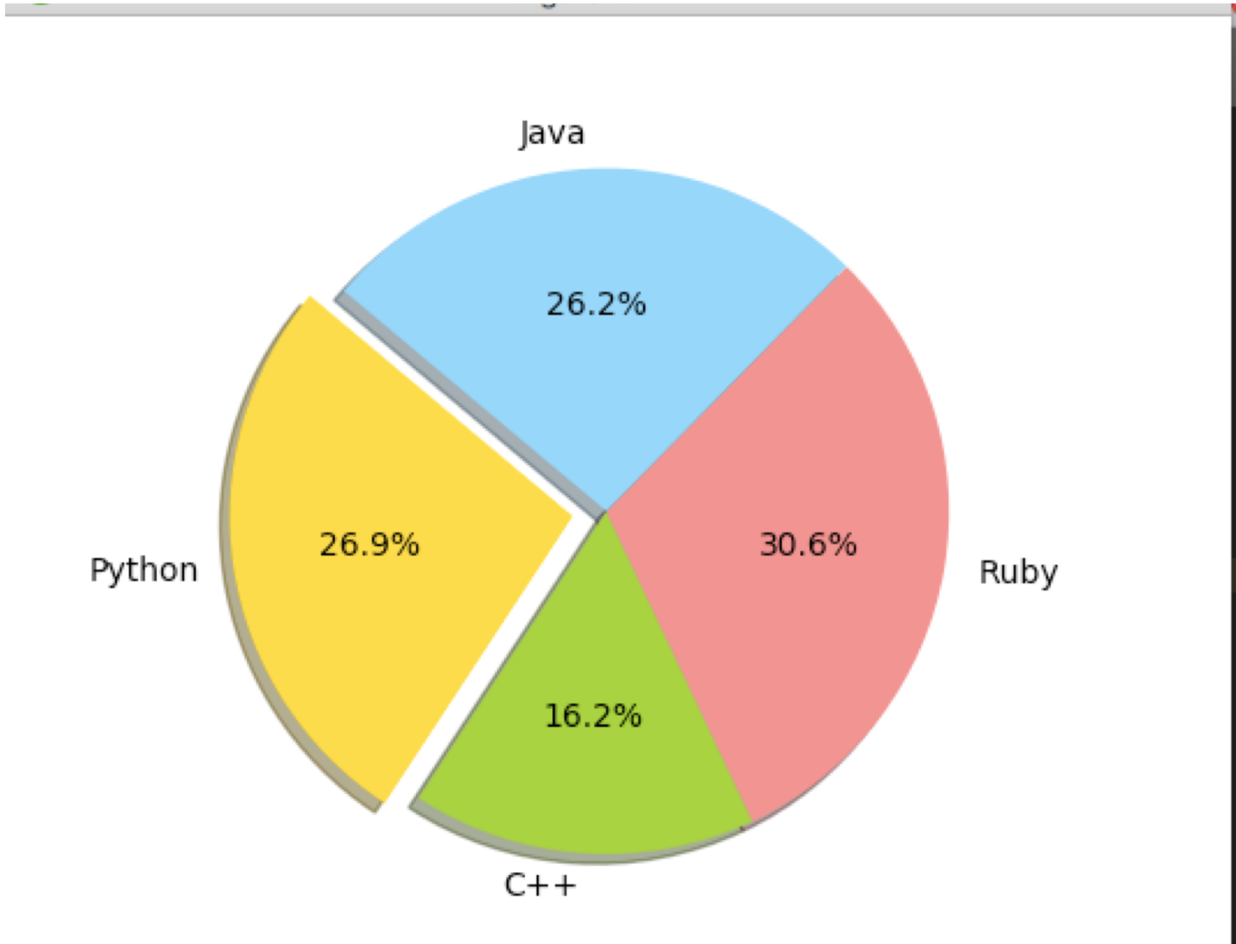
propython.py

```
from pylab import *

y = [2,4,6,8,10,12,14,16,18,20]
x = arange(10)
fig = figure()
ax = subplot(111)
ax.plot(x, y, label='$y = numbers')
title('Legend inside')
ax.legend()
plt.show()

fig.savefig('plot.png')
```

fig.savefig('plot.pdf')

display plot.png

PIE CHART



propython.py

```python
from pylab import *

# Data to plot
labels = 'Python', 'C++', 'Ruby', 'Java'
sizes = [215, 130, 245, 210]
colors = ['gold', 'yellowgreen', 'lightcoral', 'lightskyblue']
explode = (0.1, 0, 0, 0)  # explode 1st slice

# Plot
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
        autopct='%1.1f%%', shadow=True, startangle=140)

plt.axis('equal')
plt.show()
```