Python
WEEK 10
Statistics & Visualization

AI Academy

COMPUTER PROGRAMMING WITH PYTHON

Instructor - James E. Robinson, III

Teaching Assistant - Travis Martin

LIGHTNING REVIEW

- Variables
- Input / Output
- Expressions
- Functions
- Conditional Control
- Looping
- Data Types
- Logging
- Functions
- Scope
- Decorator
- Recursion
- Dynamic Prg
- Exceptions
- Classes
- Objects
- Encapsulation
- Public v/s Private
- Dunder Methods
- Instances
- Inheritance
- Types of Inheritance
- Polymorphism
- Method Overriding
- Queue
- Stacks
- Graphs
- Trees
- Binary Trees
- Traversal Methods
- Searching
- Files
- Opening / Closing
- Reading / Writing
- Context Managers
- File Exceptions
- Handling Exceptions
- File Formats
- Numpy
- Ndarray
- Pandas
- Series
- Dataframes

TOPICS COVERED

- Statistics
  - Mean, Median, Mode
  - Variance
  - Standard Deviation
  - Outliers
  - Normal Distribution
  - Feature Scaling
  - Quartiles

- Visualizations
  - Box Plot
  - Pie Chart
  - Scatter Plot
  - Line Plot
  - Histogram

MEAN , MEDIAN , MODE

MEASURES OF CENTRAL TENDENCY

Mean
Adding all numbers in the data set and then dividing by the number of values in the set

Median
Middle value when a data set is ordered from least to greatest

Mode
Value that occurs most often (high frequency) in a data set

Example
1,3,4,3,5,6,4,2,0,7,8,9,2,3,6,9,3,8,1

mean
84 / 19 = 4.421

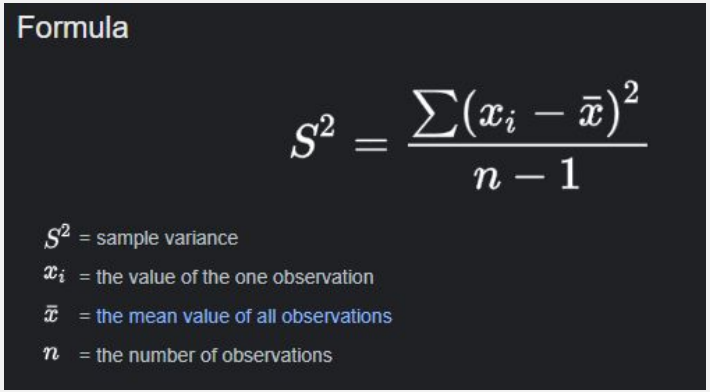
median
0,1,1,2,2,3,3,3,3,**4**,4,5,6,6,7,8,8,9,9

mode
0,1,1,2,2,**3,3,3,3**,4,4,5,6,6,7,8,8,9,9

VARIANCE

MEASURES OF CENTRAL TENDENCY

Variance
The average degree to which each point differs from the mean; average of all data points within a group



Example
1,3,4,3,5,6,4,2,0,7,8,9,2,3,6,9,3,8,1

mean(x) = 4.421
n = 19

By formula

variance = 7.506

Note:
var() function exists in numpy and pandas to calculate variance

STANDARD DEVIATION

MEASURES OF CENTRAL TENDENCY

Standard Deviation
Describes the spread of a group of numbers from the mean; also the square root of the variance

Formula

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

s = sample standard deviation
N = the number of observations
x_i = the observed values of a sample item
$\overline{x}$ = the mean value of the observations

Example
1,3,4,3,5,6,4,2,0,7,8,9,2,3,6,9,3,8,1

variance = 7.506

std_deviation = √7.506 = 2.739

Note:
std() function exists in numpy and pandas to calculate standard deviation

"Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out."

OUTLIERS

INEFFICIENCIES IN DATASETS

Outliers are values within a dataset that vary greatly from the others meaning that, they're either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty.

Outliers may have a negative effect on the result of an analysis and can skew the central tendencies of a data set.

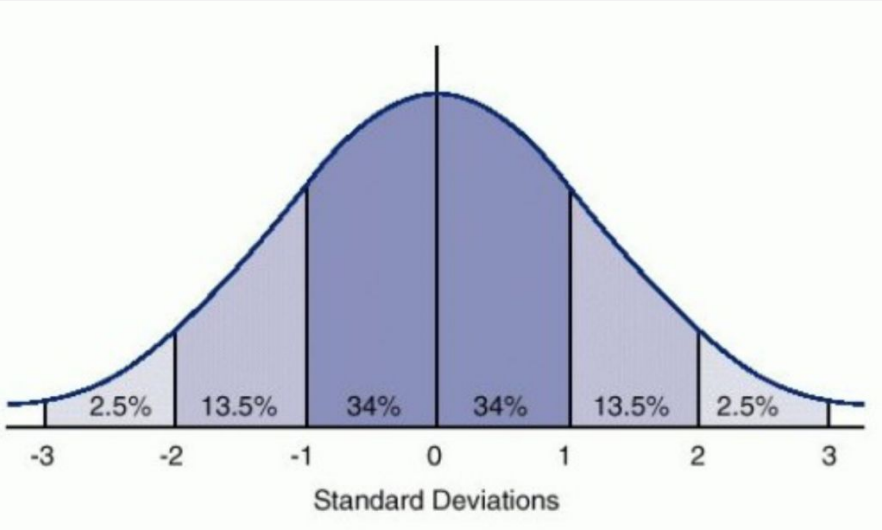
Outliers can be dealt with by -

- Removing them from the dataset
- Using a standard value for the records
- Replacing value with mean, median or mode

NORMAL DISTRIBUTION

GAUSSIAN DISTRIBUTION

Probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve



Note:

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of data falls within 1 SD of the mean.
- 95% of data falls within 2 SDs of the mean.
- 99.7% of data falls within 3 SDs of the mean.

"A clear normal distribution is indication of an accurate dataset"

FEATURE SCALING

STANDARDIZATION AND NORMALIZATION

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

Normalization

Scale values such that data has values between 0 and 1

X_new = (X - X_min)/(X_max - X_min)

Normalization ensures that each variable in a data set contributes equally to the analysis

Standardization

Scale values such that data has a mean of 0 and a standard deviation of 1

X_new = (X - mean)/Std

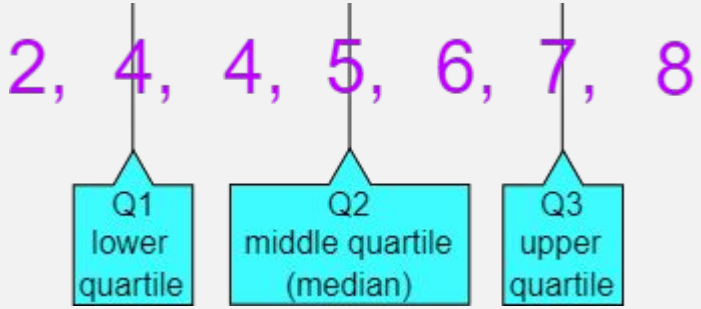
Standardization ensures that the shape of distribution remains the same

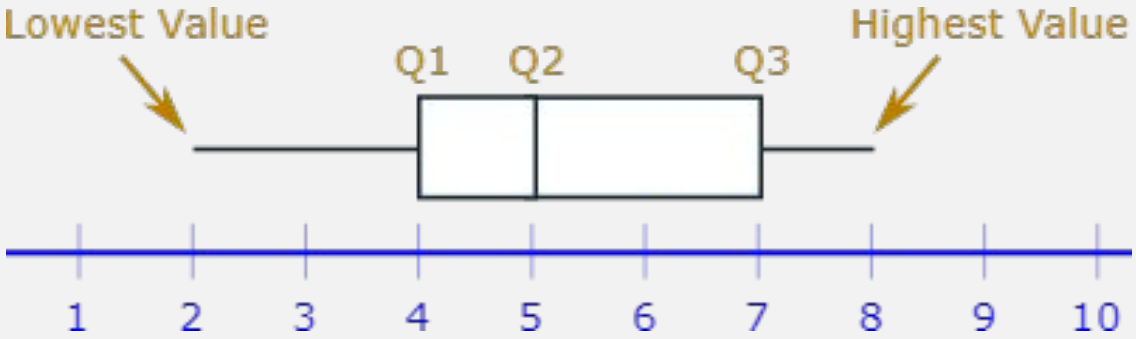
QUARTILES

DIVISION OF DATA BY THREE POINTS

Quartile - quartile divides data by three points

- Lower quartile or Q1
- Median or Q2
- Upper quartile or Q3





VISUALIZATION

REPRESENTATION OF DATA IN A DIAGRAMMATIC FASHION

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.

The main goal of data visualization is to make it easier to **identify patterns, trends and outliers** in large data sets.

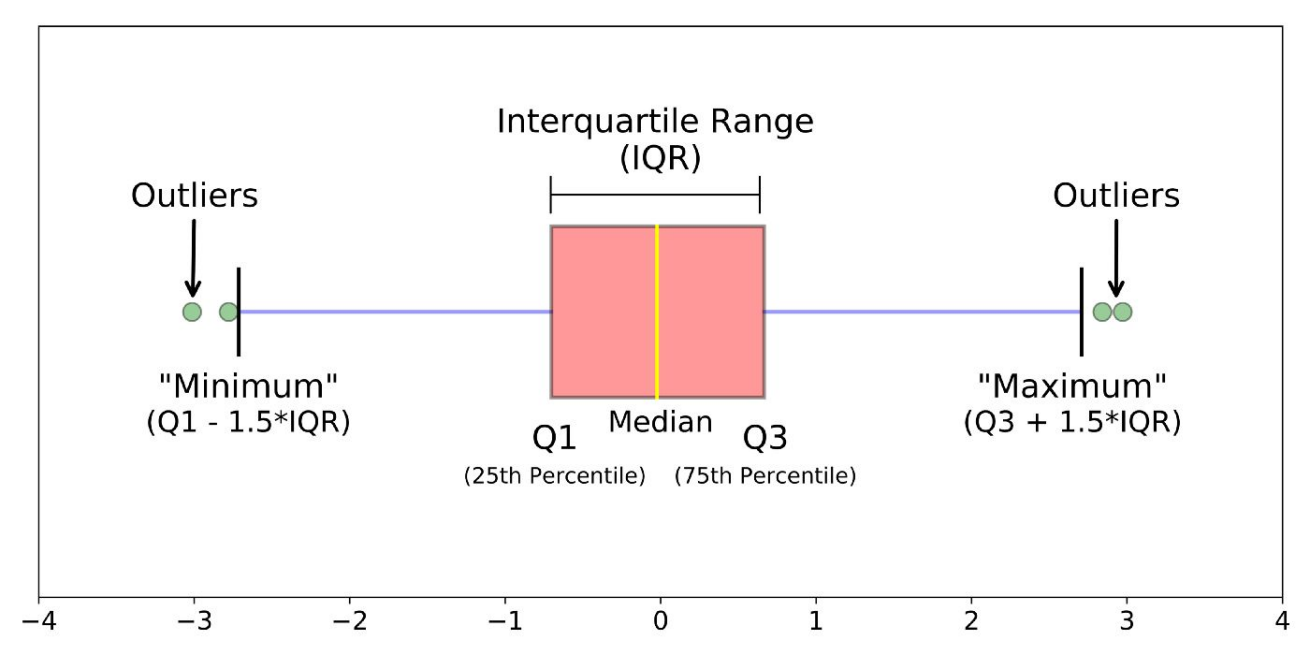
Common examples of visualization are:

- Box Plot
- Scatter Plot
- Histogram
- Bar Plot
- Pie Chart

Note:
Libraries like matplotlib, seaborn, plotly in python are used to help visualize data.

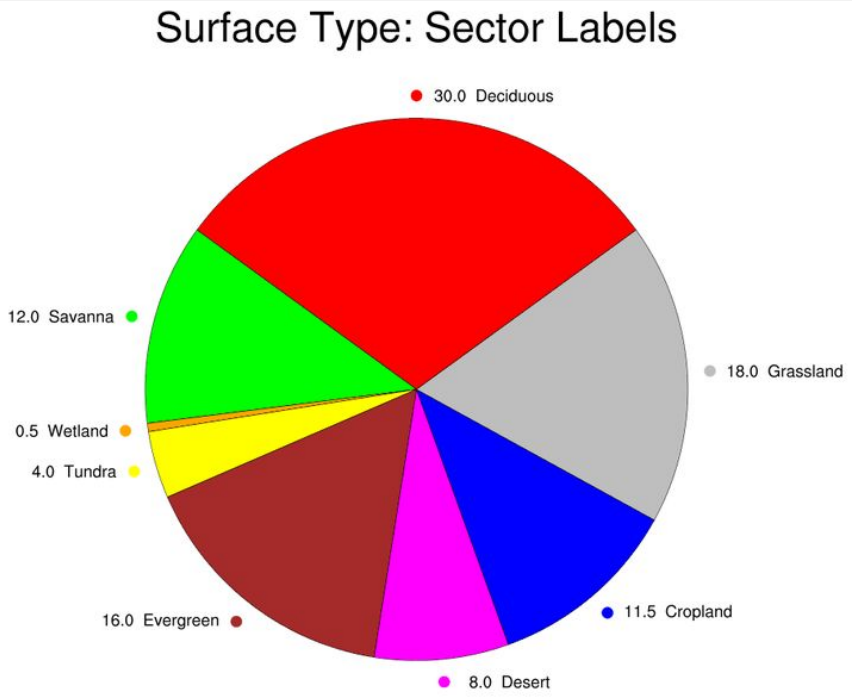
BOX PLOT

SUMMARY OF THE SET OF DATA VALUES HAVING PROPERTIES LIKE MINIMUM, FIRST QUARTILE, MEDIAN, THIRD QUARTILE AND MAXIMUM



PIE CHART

CIRCULAR STATISTICAL PLOT THAT CAN DISPLAY ONLY ONE SERIES OF DATA WHERE THE AREA OF THE CHART IS THE TOTAL PERCENTAGE OF THE GIVEN DATA

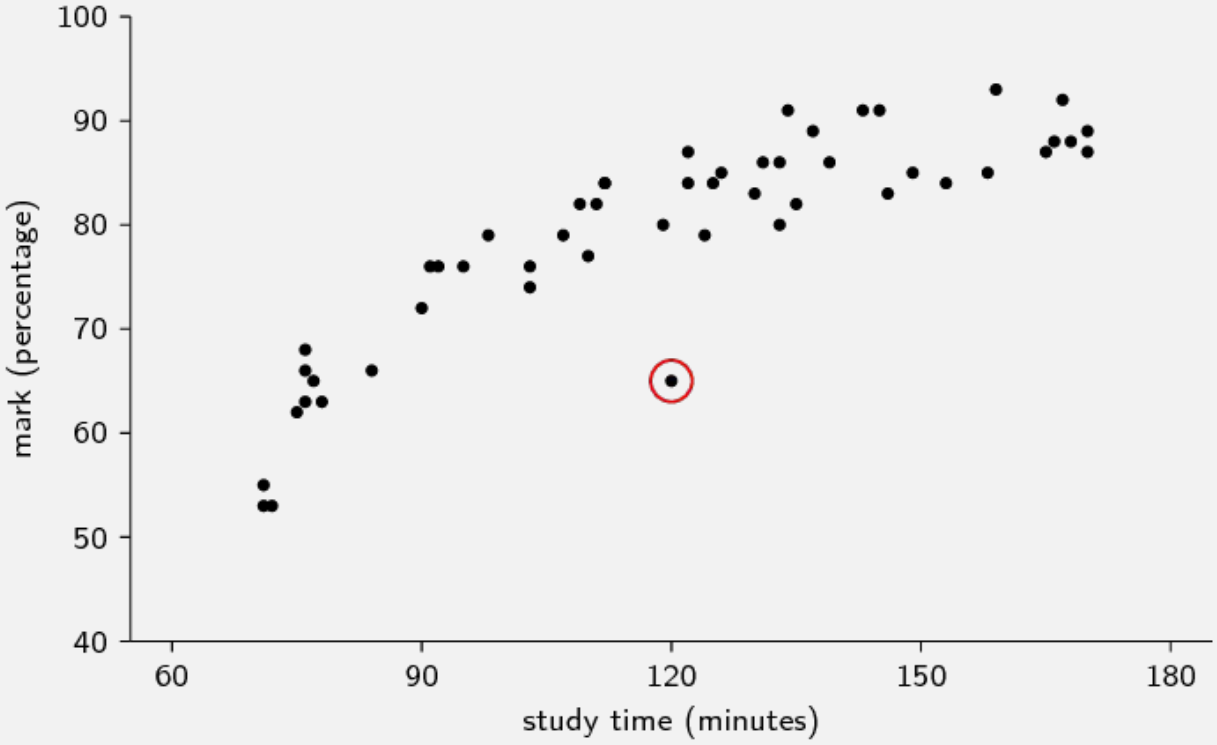


AI Academy

NC STATE

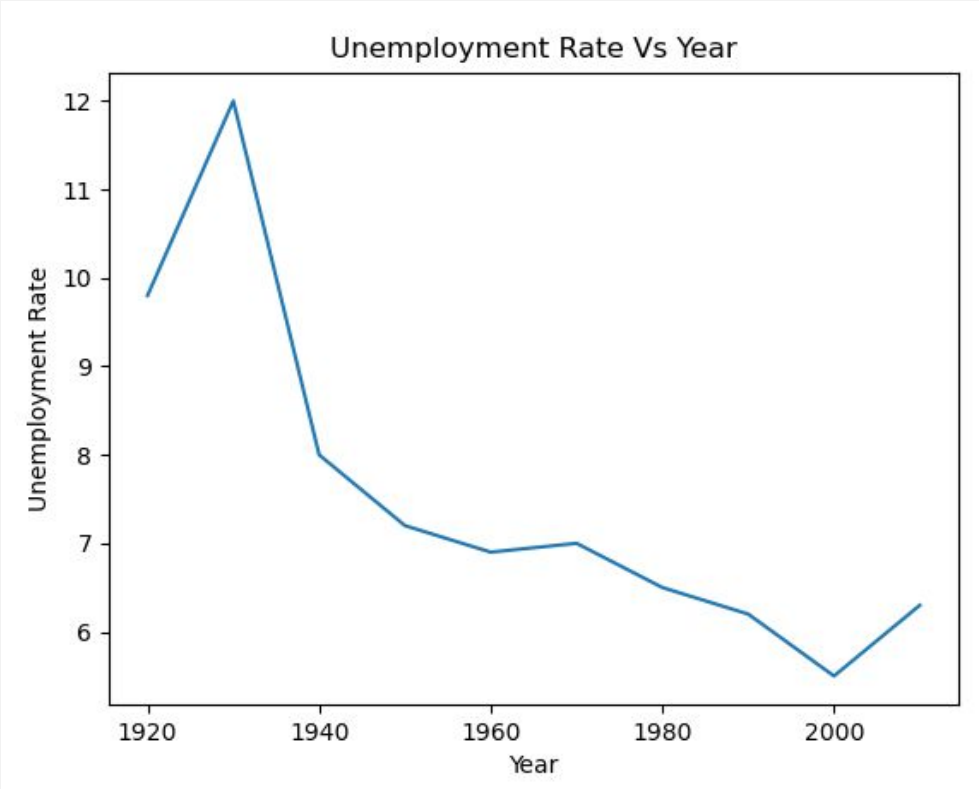
SCATTER PLOT

SUMMARY OF DATA WHERE EACH VALUE IN THE DATA IS REPRESENTED BY A DOT



LINE PLOT

REPRESENT THE RELATION BETWEEN TWO DATA X AND Y ON A DIFFERENT AXIS

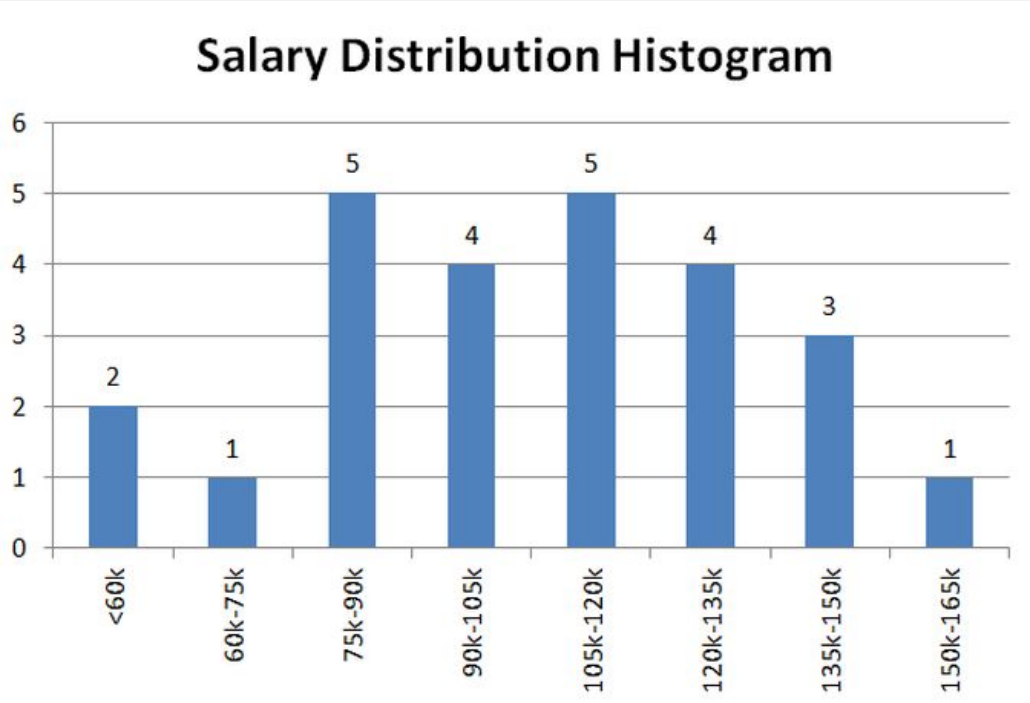


AI Academy

NC STATE

HISTOGRAM

TAKING MANY DATA POINTS AND GROUPING THEM INTO LOGICAL RANGES OR BINS



WEEK SUMMARY

- Learned the basic terms in statistics
- Learned concept of Distribution
- Learned to scale data for data mining
- Understood scaling techniques
- Learned about visualization
- Understood different kinds of visualizations

THANK YOU

FOR ADDITIONAL QUERIES OR DOUBTS
CONTACT:
jerobins@ncsu.edu

AI Academy