Lecture -9

PYTHON INSTITUTE
Open Education & Development Group

Programing in Python

Instructor : AALWAHAB DHULFIQAR

Advisor : Dr. Tejfel Mate

CISCO

What you will learn:

Data Analysis
Correlations
Analysis Using Descriptive Statistics

Chapter 3: Data Analysis

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 3 - Sections & Objectives

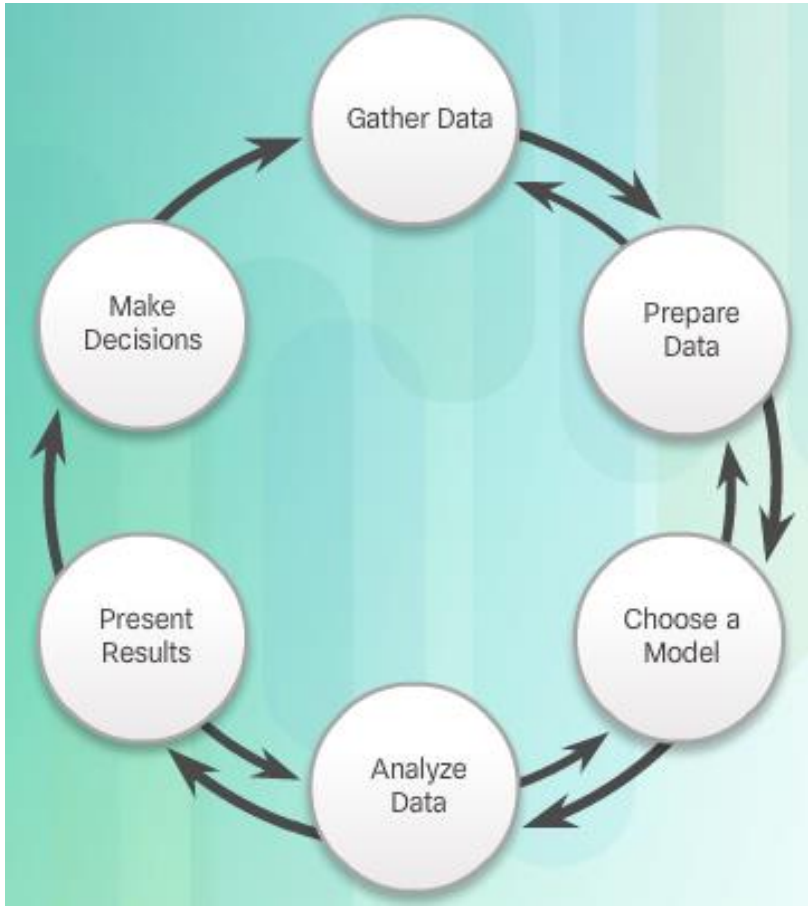
- 3.1 Analyzing Data
  - Analyze data using basic statistics.
- 3.2 Preparation for Chapter 3 Internet Meter Lab
  - Configure data for analysis.
- 3.3 Summary
  - Summarize the concepts presented in this chapter.

3.1 Analyzing Data

Cisco | Networking Academy®
Mind Wide Open™

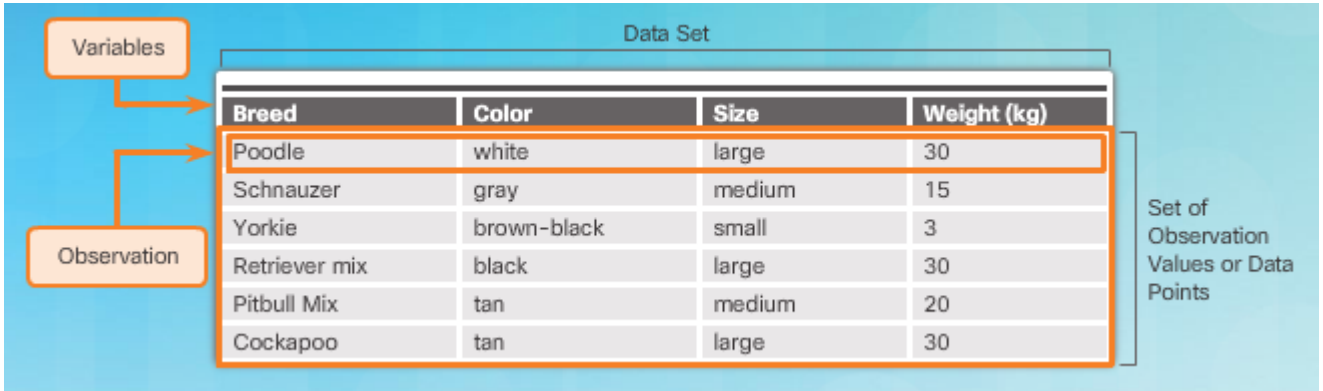
Preliminaries

- Data is changed from its raw format into information after it has been gathered, prepared, analyzed, and presented in a usable format.
- Exploratory data analysis is a set of procedures designed to produce descriptive and graphical summaries of data with the notion that the results *may* reveal interesting patterns



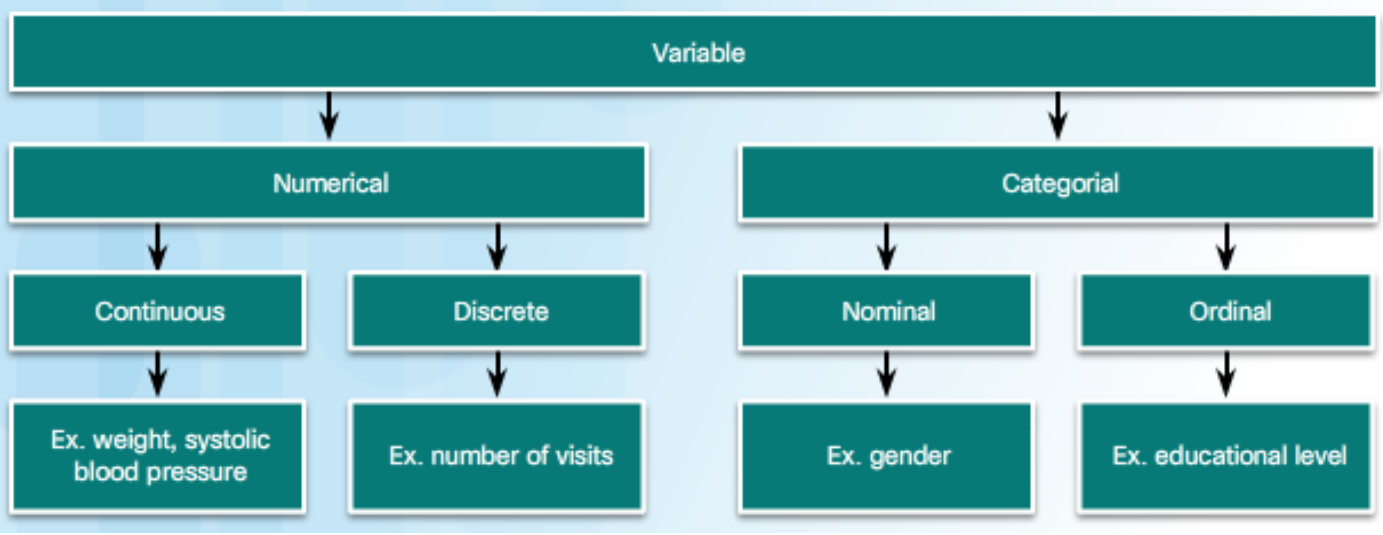
Preliminaries cont…

- **IoT Concerns**
  - IoT data may come in large volume and in different forms.
  - IoT data may require more advanced analytic tools for structured and unstructured data
  - IoT data is frequently streaming in real time or nearly real time.
- **Observations, Variables, and Values**
  - A variable is anything that varies from one instance to another and is something that can be measured, manipulated or controlled.
  - The recordings of the values, patterns and occurrences for a set of variables is an observation.
  - The set of values for a specific observation is called a data point.



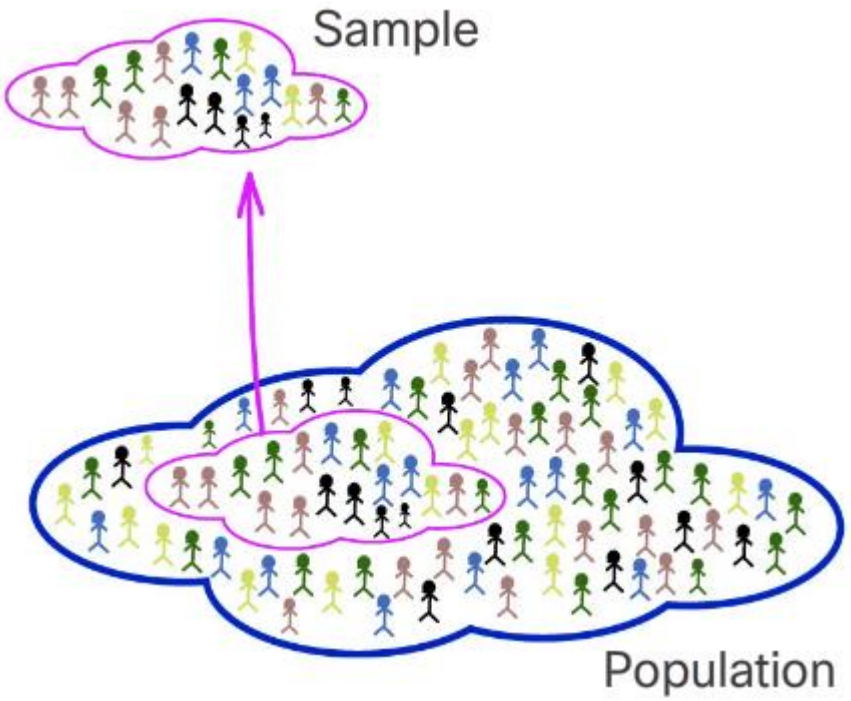
Preliminaries cont…

- Categorical variables include:
  - Nominal – Two or more categories or names that identify the object
  - Ordinal – Two or more categories in which order matter in the value
- Numerical variables include:
  - Continuous – quantitative along a continuum or range of values
  - Ratio - Interval variables where zero (0) means none
  - Discrete - Quantitative with a specific value from a finite set of values



Statistical Analysis

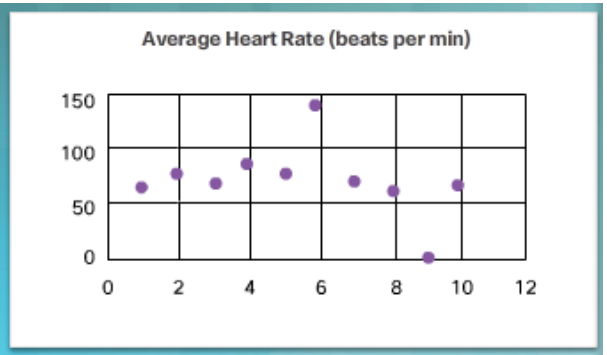
- Statistics is the collection and analysis of data using mathematical techniques.
- Sample and Population
  - A population is a group of similar entities such as people, objects, or events that share some common set of characteristics.
  - A sample is a representative group from the population.

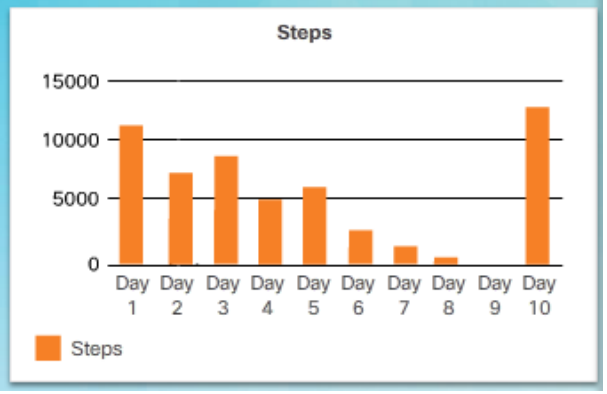


Statistical Analysis cont…

- Descriptive statistics
  - describe or summarize the values and observations of a data set.
- Inferential statistics
  - process of collecting, analyzing and interpreting data gathered from a sample to make generalizations or predictions about a population

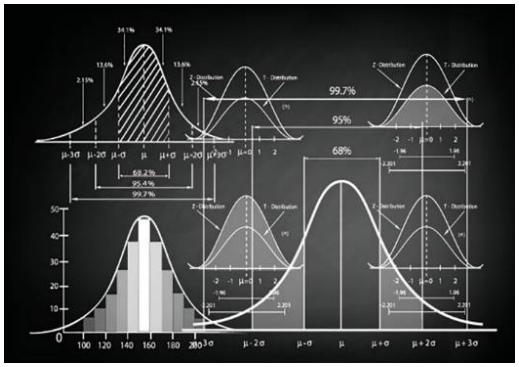
Day	Steps	Average Heart Rate (beats per min)
Day 1	10716	69
Day 2	8000	76
Day 3	9527	70
Day 4	5000	85
Day 5	6267	78
Day 6	2950	140
Day 7	1800	72
Day 8	60	64
Day 9	0	0
Day 10	12298	66

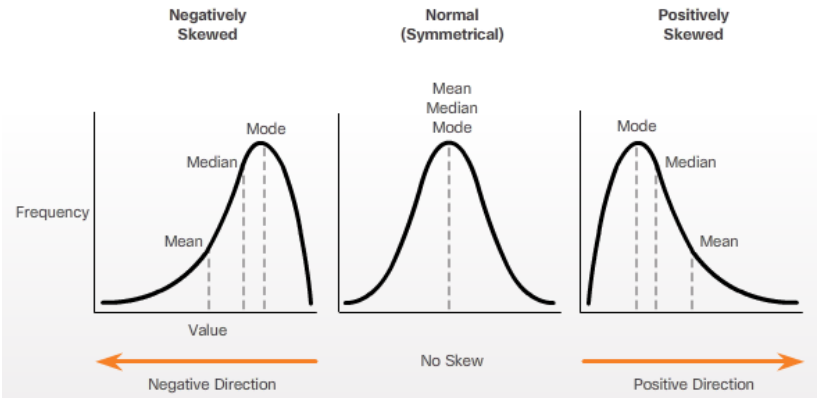


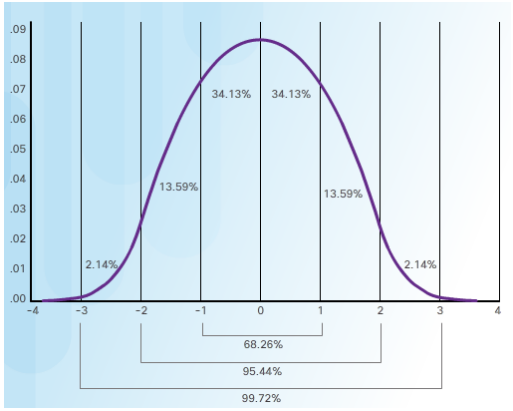


Characteristics of Samples

- Distribution
  - a variable and its frequency or probability
- Centrality
  - The mean, median, and mode
- Dispersion
  - the variability in the distribution







Analysis Using Descriptive Statistics

- Pandas
  - open source library for Python that adds high-performance data structures and tools for analysis of large data sets
  - Import data from files
  - Import data from web
  - Descriptive statistics in pandas

```
import pandas as pd
```

```
url =
'http://manage.hdx.rwlabs.org/hdx/api/exporter/indicator/csv/TT014/source/mdgs/fromYear/
1950/toYear/0/language/en/TT014_Baseline.csv'
```

```
some_cols = pd.read_table(url, sep=',', usecols = [1,2,7])
```

```
some_cols.head()
```

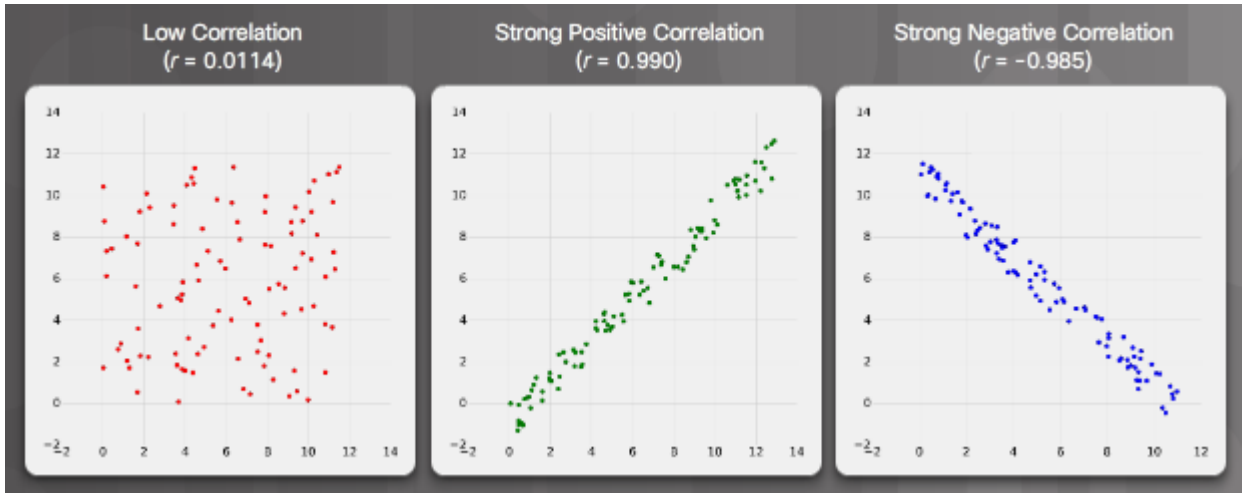
	Country name	2015	2010
0	AFGHANISTAN	27.7	27.3
1	ANGOLA	36.8	38.6
2	ALBANIA	20.7	16.4

```
some_cols.describe()
```

	2015	2010
count	189.000000	187.000000
mean	20.403704	17.381283
std	12.072191	11.140854

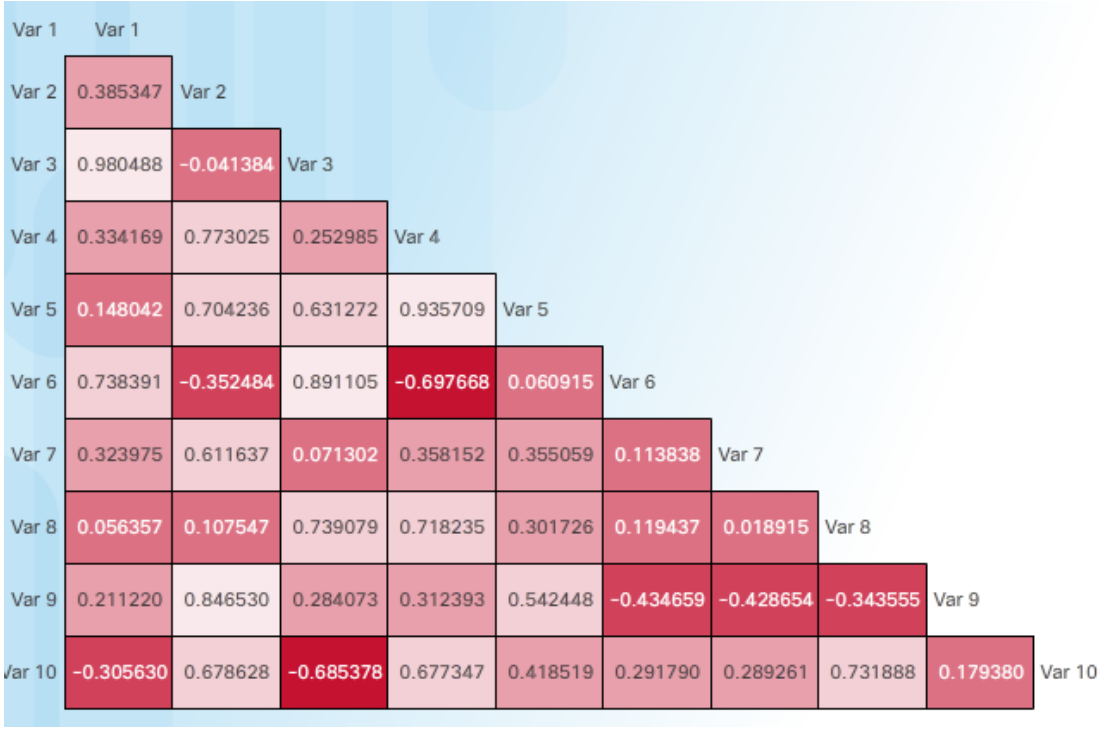
Analysis Using Correlation

- "Correlation does not imply causation"
  - Causation is a relationship in which one thing changes, or is created, directly because of something else.
  - Correlation is a relationship between phenomena in which two or more things change at a similar rate.
  - Correlations can be positive or negative.



Analysis Using Correlation cont…

- Correlations can be calculated for multiple variables simultaneously
- Heat map
  - values for correlation coefficients relate to one another



3.2 Preparation for Chapter 3 Internet Meter Lab

Cisco | Networking Academy®
Mind Wide Open™

Preparation for Chapter 3 Internet Meter Lab

Basic Analysis with pandas

- More often than not, the data sets that you work with will have incompatibilities
- Cleaning data can involve removing missing or unwanted values, or altering the format of the values to make them consistent

- **NaNs** (Not a Number) values are used to represent data that is undefined or cannot be represented. pandas refers to missing data as NaN values
  - NaTs are used for timestamps
- **Pandas** has many built-in functions for:
  - converting the datatypes
  - manipulating data frames
  - running statistical analysis on data sets.

3.3 Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- Exploratory data analysis produces descriptive and graphical summaries of data with the notion that the results *may* reveal interesting patterns.
- IoT data may be structured or unstructured and data must be organized in real time.
- Observations, variables, and values are critical to an analysis.
- Variables include Numerical (Continuous and Discrete) and Categorical (Nominal and Ordinal)
- Statistics is the collection and analysis of data using mathematical techniques.
  - The interpretation of data and the presentation of findings.
  - The discovery of patterns or relationships between variables.
- Statistics uses samples and populations.
- Statistical analysis includes descriptive and inferential statistics.

Summary cont…

- Distribution is a simple association between a value and the number or percentage of times it appears in a data sample.
- Centrality includes the mean, median, and mode.
  - These values that are closer to the center of the distribution occur with greater frequency.
- Dispersion is the variability in the distribution.
- Pandas is an open source library for Python with tools for analysis of large data sets
  - Importing data from files
  - Importing data from Web
  - Viewing descriptive statistics
- “Correlation does not imply causation”
- Data commonly needs cleaning, converting, and manipulating before data analysis.