3. fejezet: Adatelemzés

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 3 - Sections & Objectives

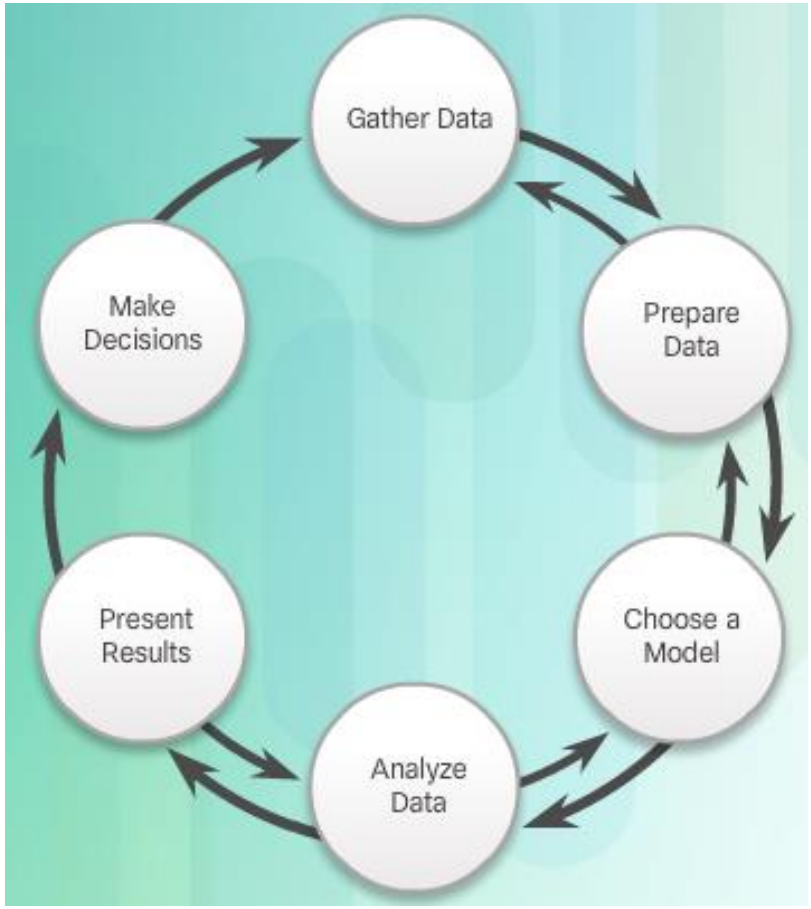
- 3.1 Az adatok elemzése
  - Az adatok elemzése alapvető statisztikák segítségével.
- 3.2 Felkészülés a 3. fejezet Internet Meter laborjára
  - Az adatok konfigurálása elemzéshez.
- 3.3 Összefoglaló
  - A fejezetben bemutatott fogalmak összefoglalása.

3.1 Analyzing Data

Cisco | Networking Academy®
Mind Wide Open™

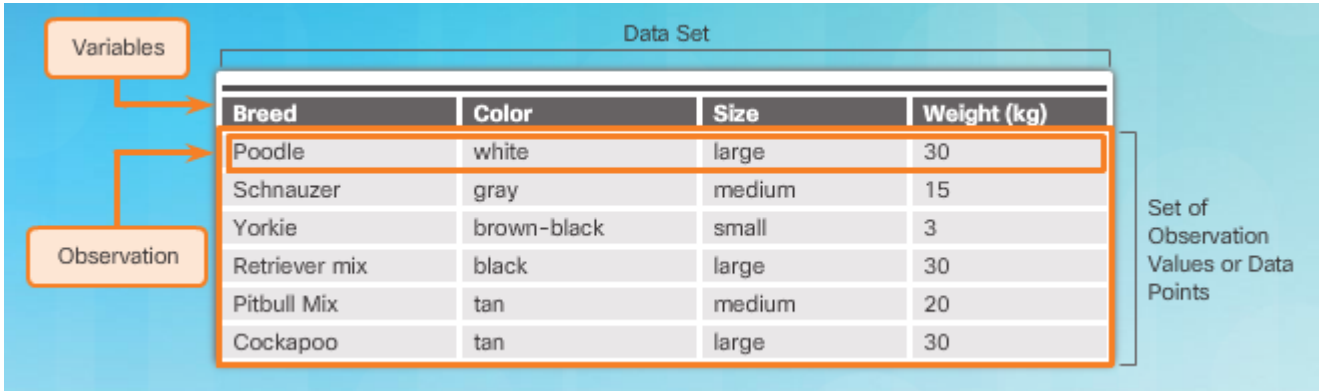
Előzetesek

- Az adatok nyers formájukból információvá alakulnak át, miután összegyűjtötték, előkészítették, elemezték és használható formában bemutatták őket.
- A feltáró adatelemzés olyan eljárások összessége, amelyek célja az adatok leíró és grafikus összefoglalóinak készítése azzal a gondolattal, hogy az eredmények érdekes mintákat tárhatnak fel.



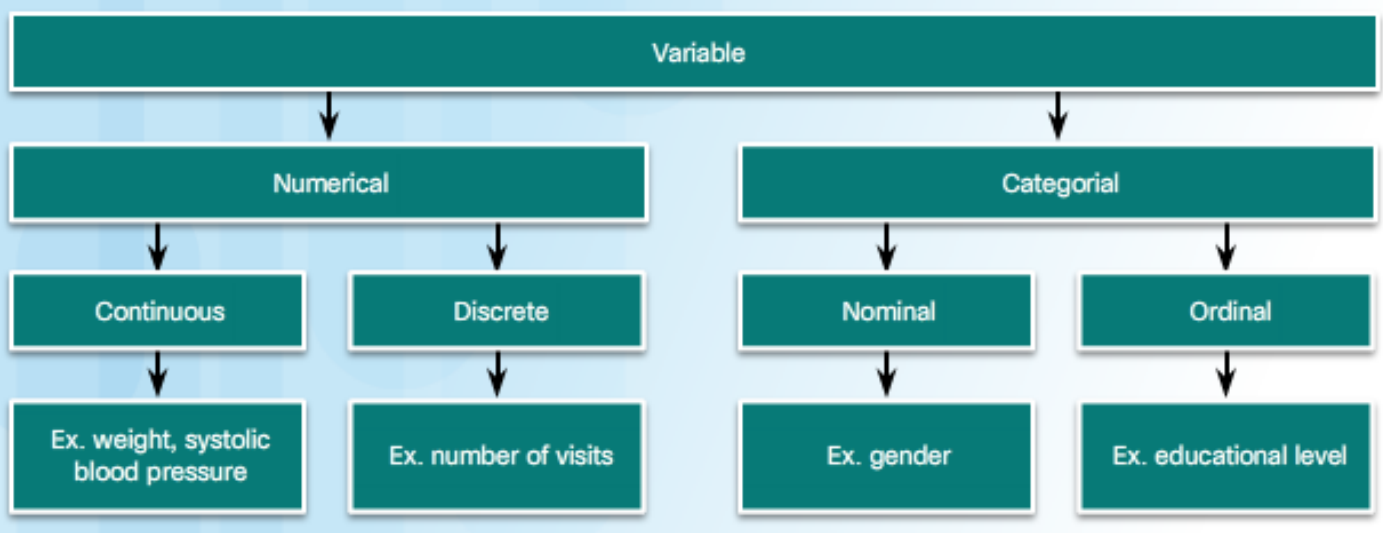
Előzetesek cont…

- **IoT aggályok**
  - Az IoT-adatok nagy mennyiségben és különböző formában érkezhetnek.
  - Az IoT-adatokhoz fejlettebb elemzőeszközökre lehet szükség a strukturált és strukturálatlan adatokhoz.
  - Az IoT-adatok gyakran valós időben vagy közel valós időben áramlanak.
- **Observations, Variables, and Values**
  - A variable minden, ami egyik esetről a másikra változik, és ami mérhető, manipulálható vagy ellenőrizhető.
  - A változók egy halmazára vonatkozó értékek, mintázatok és előfordulások rögzítése egy észrevétel (observation).
  - Egy adott megfigyeléshez tartozó értékek halmazát Data Pointnak nevezzük.



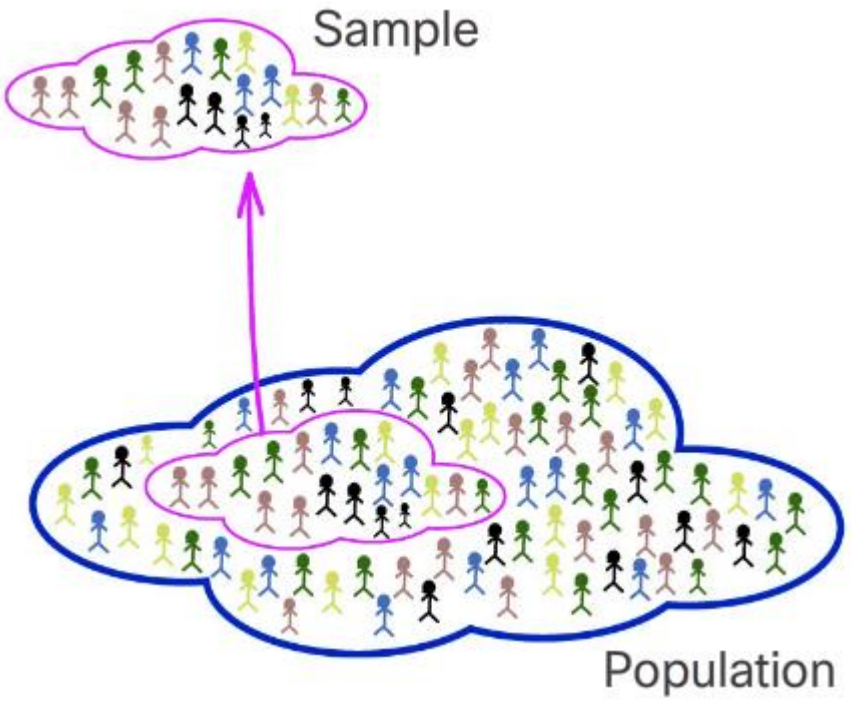
Előzetesek cont…

- A kategorikus változók (Categorical variables):
  - Nominal – Két vagy több kategória vagy név, amelyek azonosítják az objektumot.
  - Ordinal – Két vagy több kategória, amelyekben a sorrend számít az értékben
- A numerikus változók (Numerical variables):
  - Continuous – mennyiségi értékek egy kontinuum vagy értéktartomány mentén
  - Ratio - Intervallum változók, ahol a nulla (0) azt jelenti, hogy nincs.
  - Discrete - Kvantitatív, meghatározott értékkel egy véges értékkészletből



Statistical Analysis

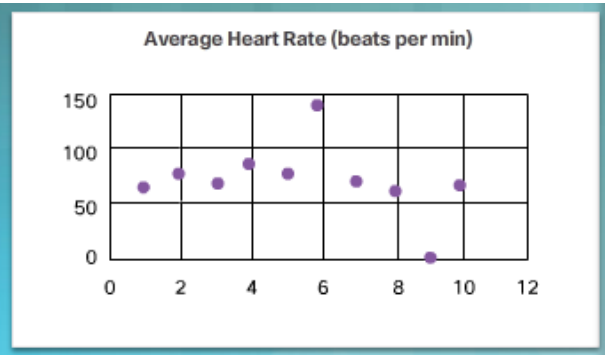
- Statistics is the collection and analysis of data using mathematical techniques.
- Sample and Population
  - A population is a group of similar entities such as people, objects, or events that share some common set of characteristics.
  - A sample is a representative group from the population.

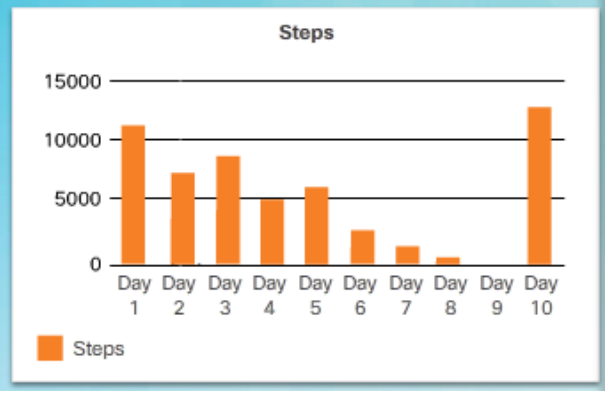


Statisztikai elemzés cont…

- Descriptive statistics
  - leírja vagy összefoglalja egy adathalmaz értékeit és megfigyeléseit(observations).
- Inferential statistics
  - a mintán gyűjtött adatok gyűjtésének, elemzésének és értelmezésének folyamata, hogy általánosításokat vagy előrejelzéseket lehessen tenni a populációra vonatkozóan.

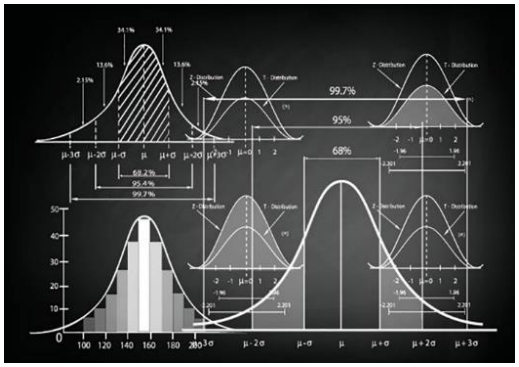
Day	Steps	Average Heart Rate (beats per min)
Day 1	10716	69
Day 2	8000	76
Day 3	9527	70
Day 4	5000	85
Day 5	6267	78
Day 6	2950	140
Day 7	1800	72
Day 8	60	64
Day 9	0	0
Day 10	12298	66

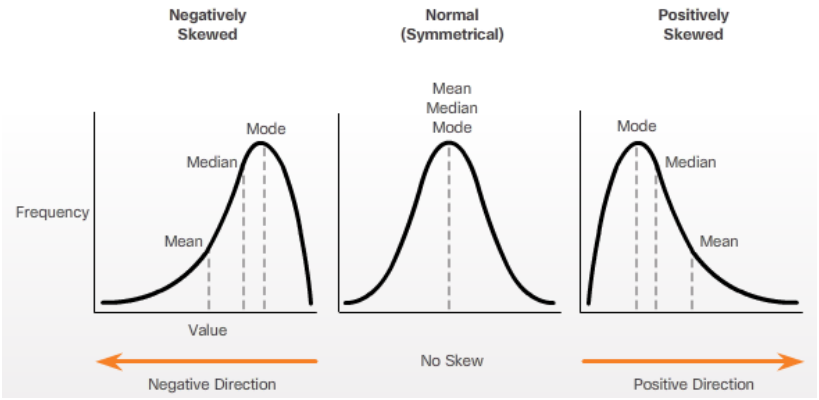


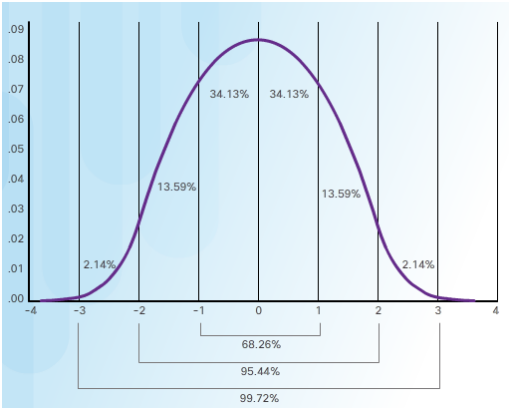


Characteristics of Samples

- Distribution
  - egy változó és annak gyakorisága vagy valószínűsége.
- Centrality
  - Az átlag, a medián és a módusz.
- Dispersion
  - az eloszlás változékonysága.







Elemzés Descriptive leíró statisztikák segítségével

- Pandas
  - nyílt forráskódú könyvtár Pythonhoz, amely nagy teljesítményű adatstruktúrákat és eszközöket kínál nagy adathalmazok elemzéséhez.
  - Adatok importálása fájlokból.
  - Adatok importálása az internetről.
  - Descriptive Leíró statisztika pandasban

```
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()
```

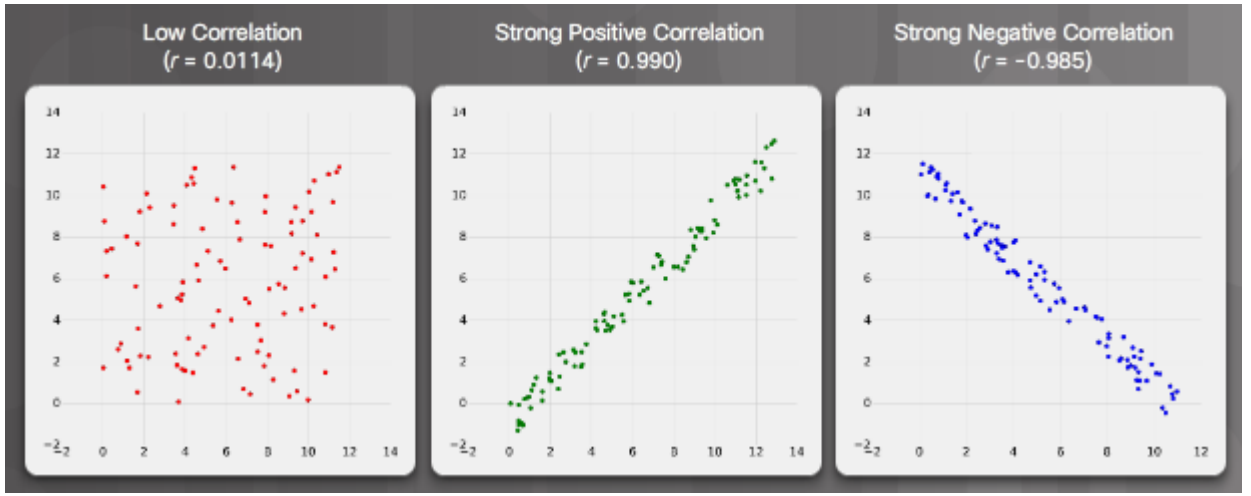
	Country name	2015	2010
0	AFGHANISTAN	27.7	27.3
1	ANGOLA	36.8	38.6
2	ALBANIA	20.7	16.4

```
some_cols.describe()
```

	2015	2010
count	189.000000	187.000000
mean	20.403704	17.381283
std	12.072191	11.140854

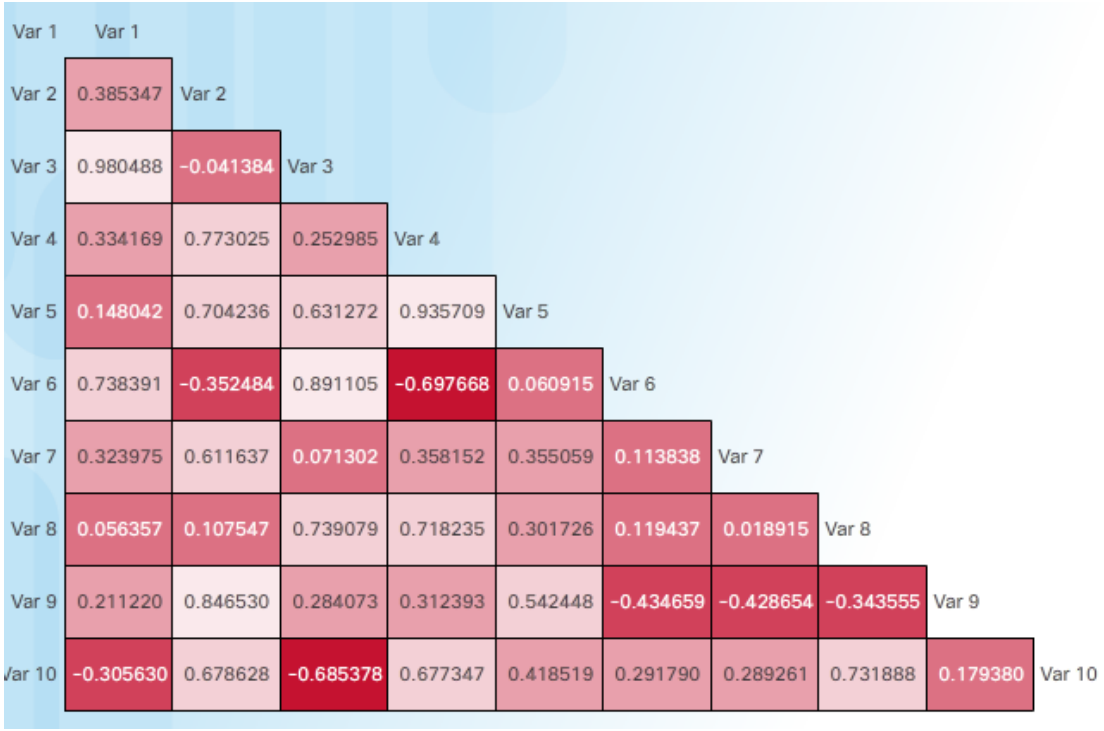
Analysis Using Correlation

- “Correlation does not imply causation”
  - Az ok-okozati viszony olyan kapcsolat, amelyben egy dolog közvetlenül valami más miatt változik, vagy jön létre.
  - A korreláció olyan jelenségek közötti kapcsolat, amelyben két vagy több dolog hasonló ütemben változik.
  - A korrelációk lehetnek pozitívak vagy negatívak.



Analysis Using Correlation cont…

- A korrelációk egyszerre több változóra is kiszámíthatók.
- Heat map
  - a korrelációs koefficiensek értékei egymáshoz viszonyítva



3.2 Felkészülés a 3. fejezet Internetmérő laboratóriumára

Cisco | Networking Academy®
Mind Wide Open™

Basic Analysis with pandas

- Gyakran előfordul, hogy az adatkészletek, amelyekkel dolgozol, nem kompatibilisek egymással.
- Az adatok tisztítása magában foglalhatja a hiányzó vagy nem kívánt értékek eltávolítását, vagy az értékek formátumának módosítását, hogy azok konzisztensek legyenek.

- **NaNs** (Not a Number) értékek a meghatározatlan vagy nem reprezentálható adatok jelölésére szolgálnak. A pandas a hiányzó adatokat NaN értékeknek nevezi..
  - NaTs az időbélyegzőket használják
- **Pandas** számos beépített funkcióval rendelkezik:
  - az adattípusok átalakítása
  - adatkeretek manipulálása
  - statisztikai elemzések futtatása adathalmazokon.

3.3 Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- A feltáró (Exploratory) adatelemzés leíró és grafikus összefoglalókat készít az adatokról azzal a gondolattal, hogy az eredmények érdekes mintákat tárhatnak fel..
- Az IoT-adatok lehetnek strukturáltak vagy strukturálatlanok, és az adatokat valós időben kell rendszerezni.
- Observations, variables, and values kritikusak az elemzés szempontjából.
- Variables include Numerical (Continuous and Discrete) and Categorical (Nominal and Ordinal)
- A statisztika az adatok matematikai módszerekkel történő gyűjtése és elemzése..
  - Az adatok értelmezése és az eredmények bemutatása.
  - A változók közötti minták vagy összefüggések felfedezése.
- A statisztika mintákat és populációkat használ.
- A statisztikai elemzés magában foglalja a leíró és következtetési statisztikát.

Summary cont…

- Az eloszlás egy egyszerű összefüggés egy érték és az adatmintában való megjelenésének száma vagy százalékos aránya között.
- A centralitás magában foglalja az átlagot, a mediánt és a móduszt.
  - These values that are closer to the center of the distribution occur with greater frequency.
- Dispersion is the variability in the distribution.
- Pandas egy nyílt forráskódú könyvtár Pythonhoz, nagy adathalmazok elemzéséhez szükséges eszközökkel.
  - Importing data from files
  - Importing data from Web
  - Viewing descriptive statistics
- “Az összefüggés nem jelent ok-okozati összefüggést”
- Az adatokat az adatelemzés előtt általában tisztítani, átalakítani és manipulálni kell.

köszönöm a figyelmet