Data 101

Alipio Jorge

10/11/2020

What is Data?

- Look it up!
  - Philosophy: *things known or assumed as facts, making the basis of reasoning or calculation*
  - Science and engineering: *the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.*

How is data represented?

- Raw data
  - what the world gives
- Table or set of tables
  - sooner or later
- Mathematically
  - vectors, matrices, tensors
- Only?
  - NO, it's just what **successful** algorithms prefer
  - and that is what DS end up with 99.9% of the time

Tables

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

- Data Object (entity)
  - Row, example, case, individual, observation ...
- Attribute
  - Column, variable, feature, dimension, ...
  - statistically: random variables

Variables

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

- Independent Variable
  - Predictor, descriptor, input variable
- Dependent Variable
  - **Class**, target attribute, output variable

Variable types

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

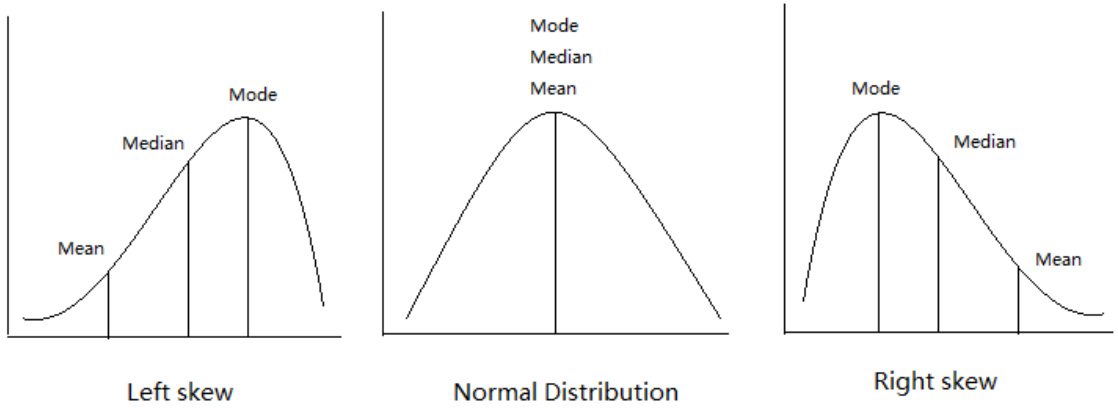
- **Discrete**
  - Nominal, Binary (or Boolean), Ordinal
- **Continuous**
  - Numeric
- Note: numeric and continuous are not synonimous, but are often treated as such

Basic Statistics

- Data understanding
  - Variables can be **summarized** using statistical measures
- Sample
  - the set of cases you get in your data set
  - each variable can be seen as a sample

Basic statistics - central

- **Central** tendency
  - μ Mean or average (numeric vars)
    - weighted mean
  - Median (ordinal vars, including numeric)
  - Mode (any var)



Basic statistics - dispersion

Ordinal, including numeric

- Range
  - Max − Min
- Quartiles, Q_1, Q_2 or median, Q_3
  - divide your data in four equal sized sets
- *Five-number summary*
  - Min, Q_1, median, Q_3, Max
- *IQR* - Inter quartile range
  - $IQR = Q_3 - Q_1$

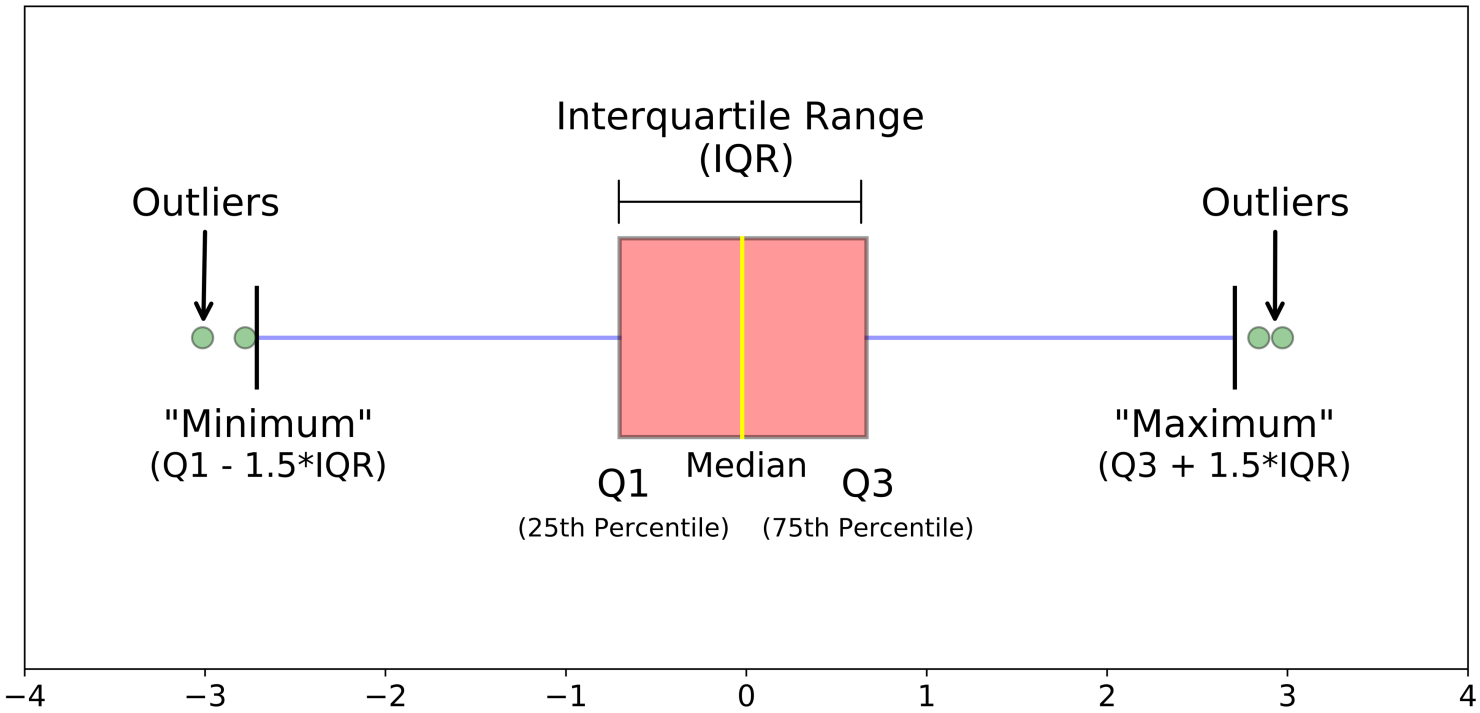
Basic statistics - dispersion

Numeric

- σ^2 - Variance
  - the mean of the squared differences of the observations to their mean
  - also: the mean of the squared observations minus the the square of their mean
- σ Standard deviation
  - the squared root of the variance

Basic statistics - Outliers

- **Outliers**
  - values well out of the expected range of observations
  - various ways to define outliers
  - Commonly:
    - $x \geq Q_3 + 1.5 \times IQR$ or
    - $x \leq Q_1 - 1.5 \times IQR$
  - Outliers can be visualized with a **box plot**
    - whisker chart, gráfico de bigodes (PT)



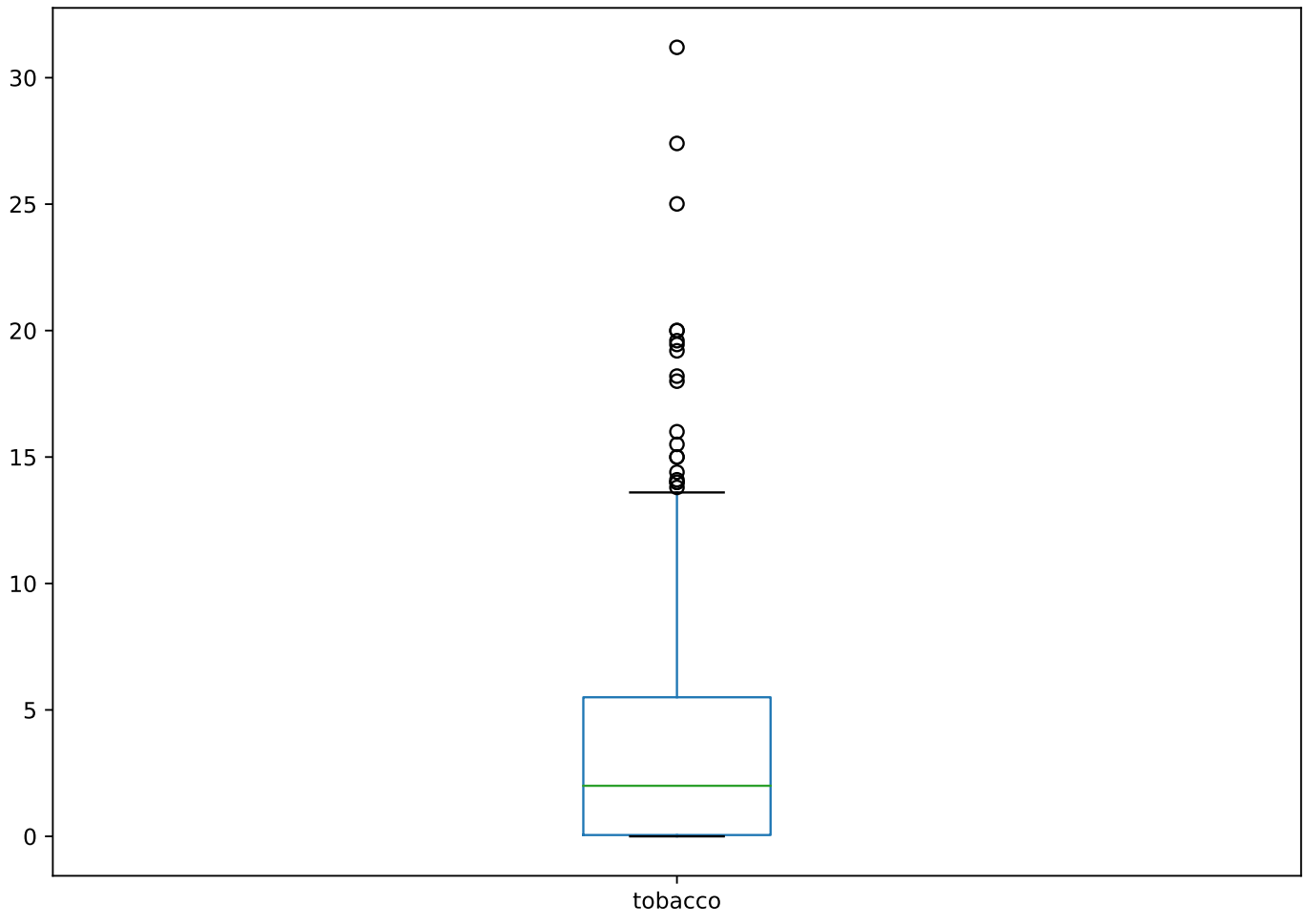
A Statistical summary in pandas

```
import pandas as pd
d = pd.read_csv('../Dados/SAheart.csv')
d[['sbp','age','chd']].describe(include='all')
```

```
##                 sbp         age  chd
## count   462.000000  462.000000  462
## unique         NaN         NaN    2
## top            NaN         NaN   No
## freq           NaN         NaN  302
## mean    138.326840   42.816017  NaN
## std      20.496317   14.608956  NaN
## min     101.000000   15.000000  NaN
## 25%     124.000000   31.000000  NaN
## 50%     134.000000   45.000000  NaN
## 75%     148.000000   55.000000  NaN
## max     218.000000   64.000000  NaN
```

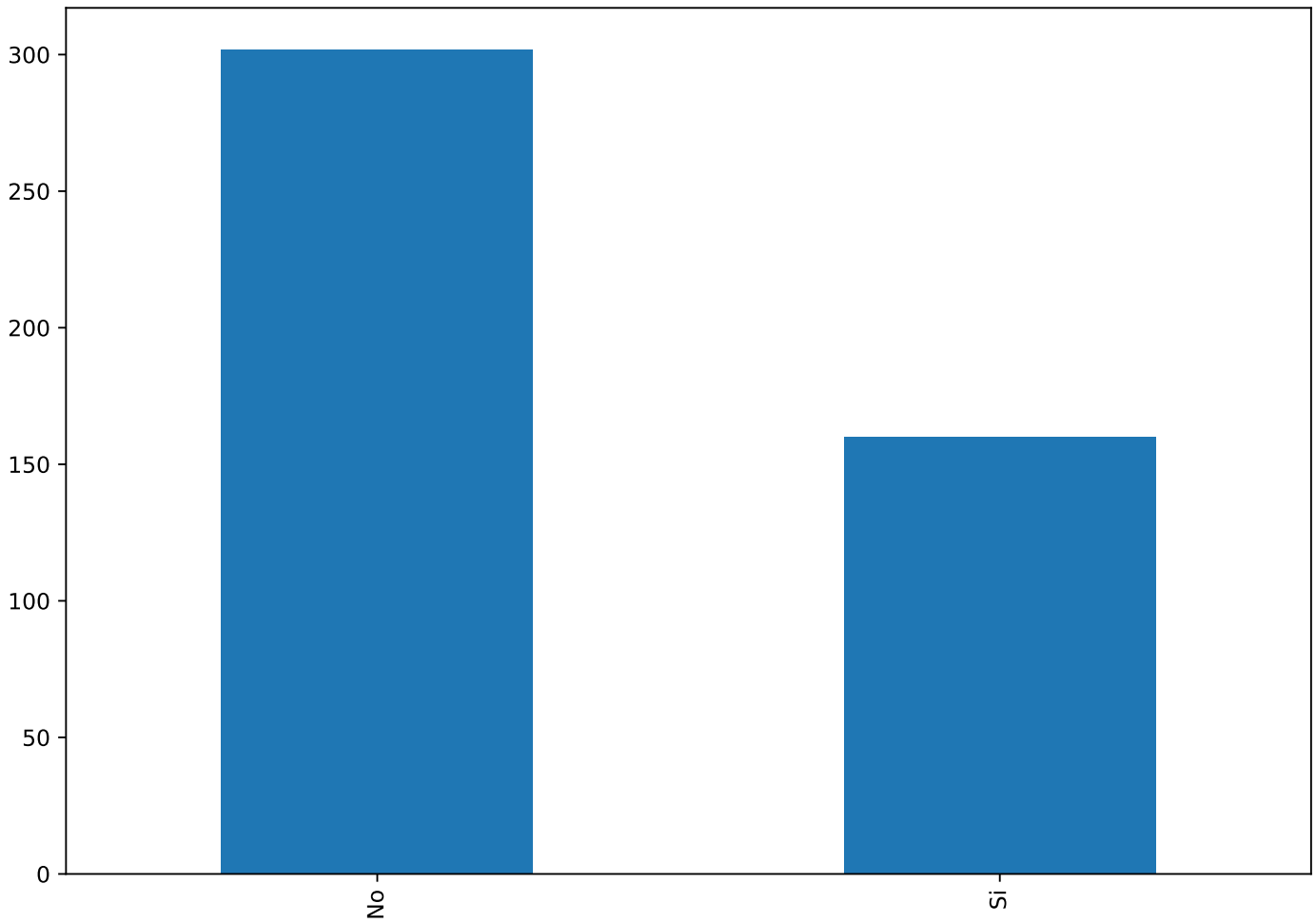
Graphical Displays - Box plot

```
d['tobacco'].plot.box()
```



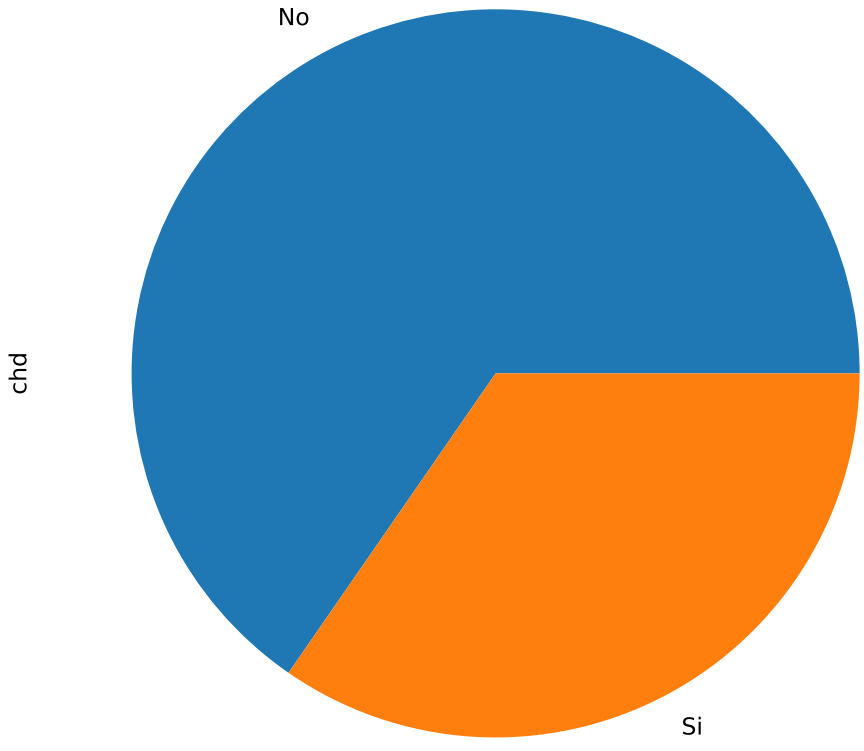
Graphical Displays - Bar plot

```
d['chd'].value_counts().plot.bar()
```



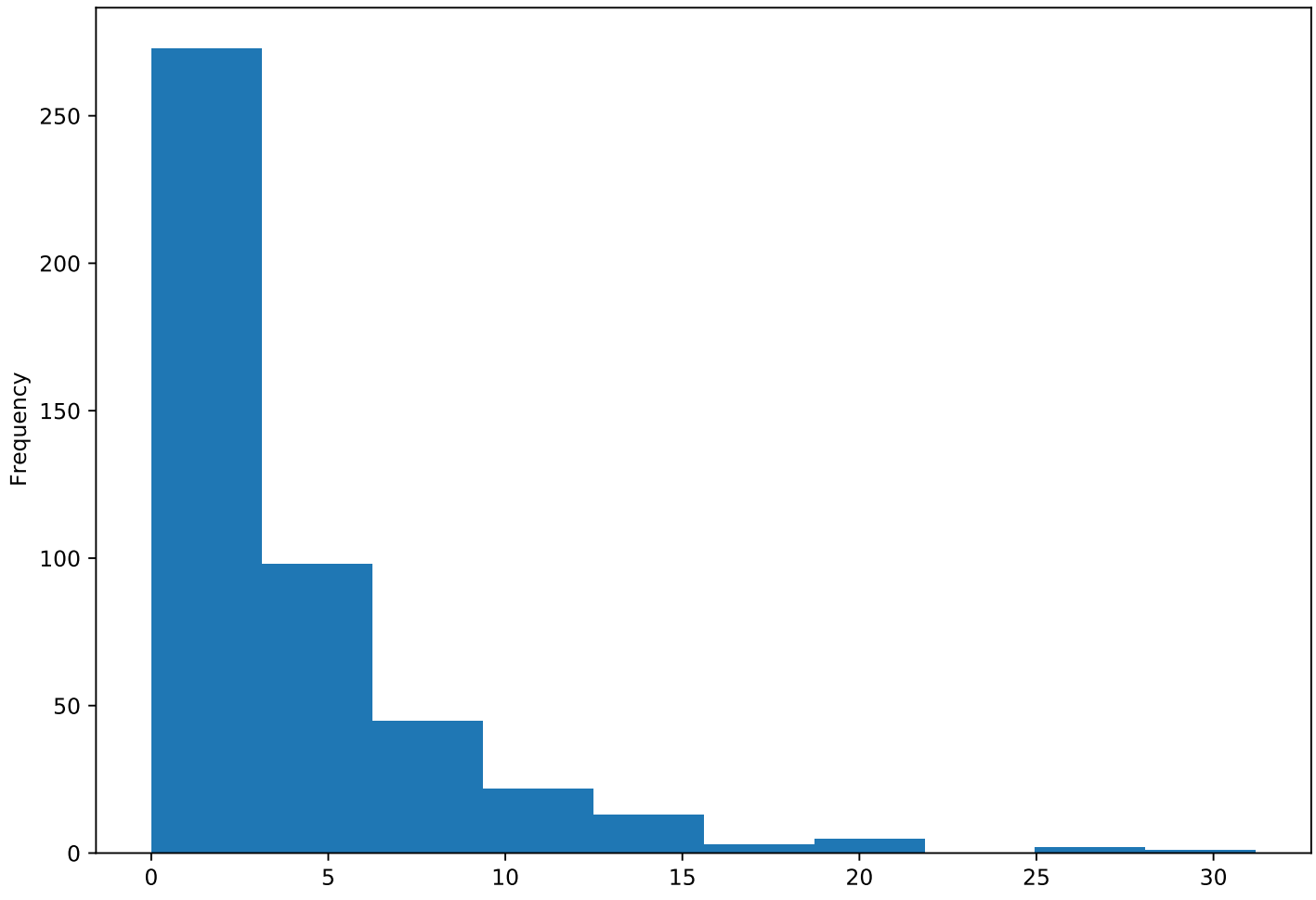
Graphical Displays - Pie plot

```
d['chd'].value_counts().plot.pie()
```



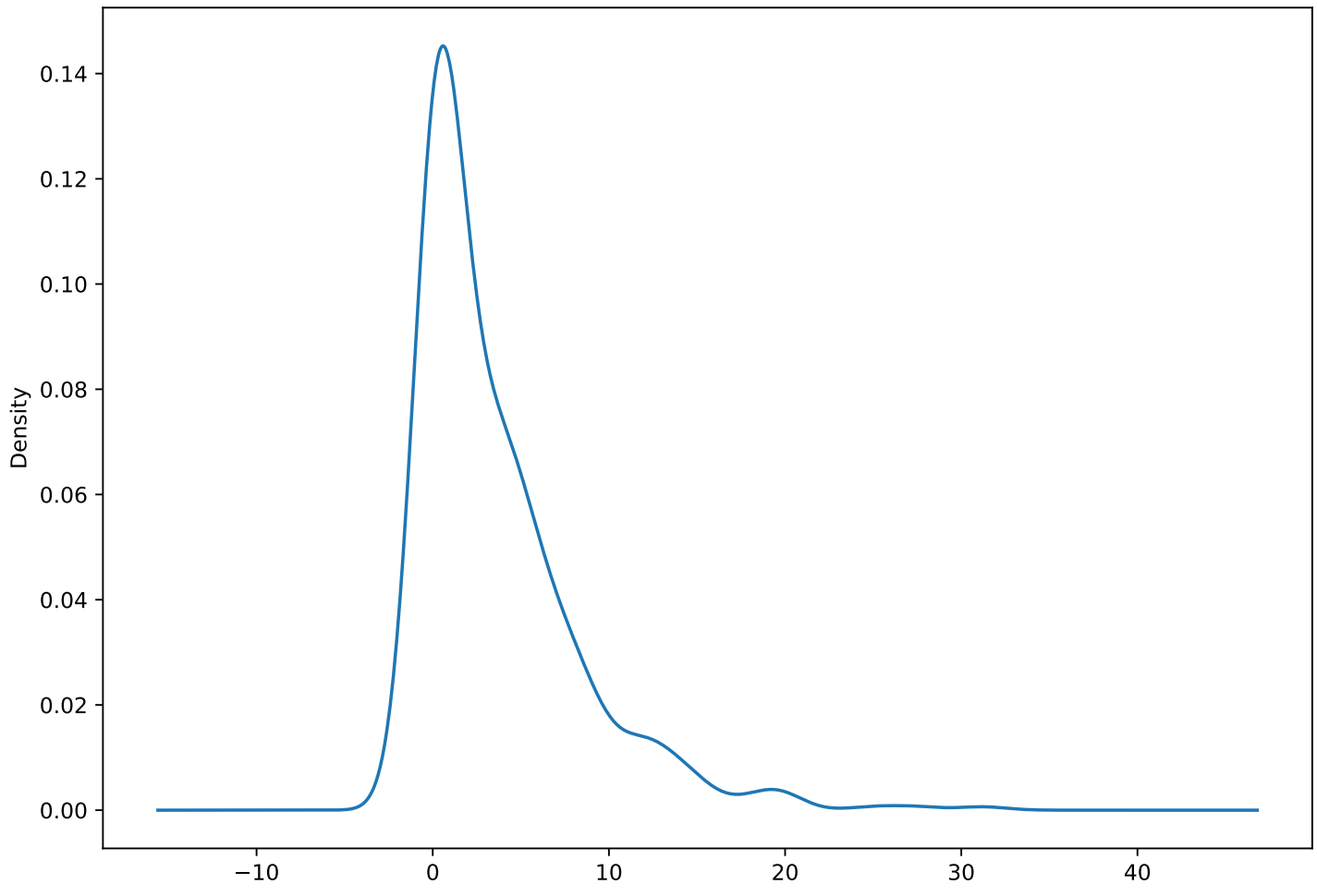
Graphical Displays - Histogram

```
d['tobacco'].plot.hist()
```



Graphical Displays - Density

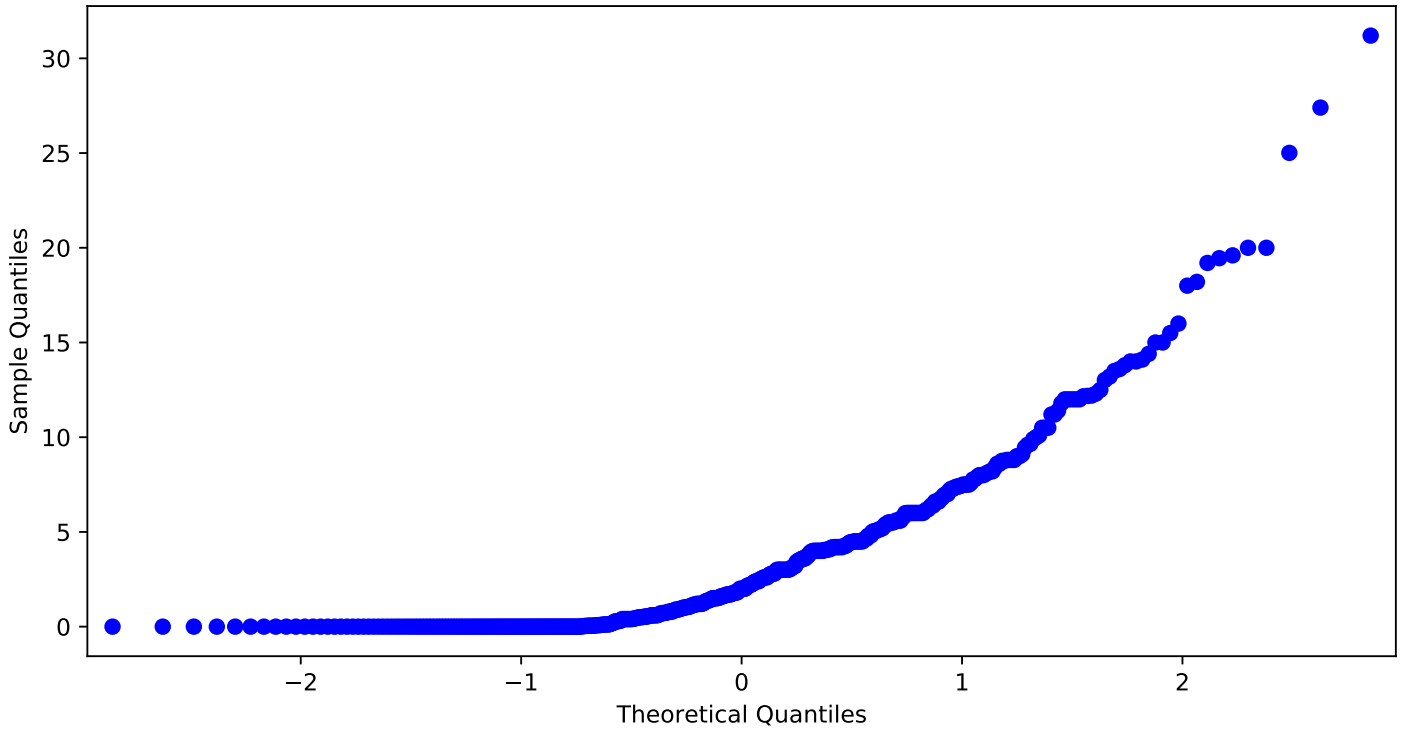
```
d['tobacco'].plot.density()
```



Graphical Displays - Quantile plot

- Visual comparison with the standard normal
  - or between two variables

```
import statsmodels.api as sm
sm.qqplot(d['tobacco'])
```



Multivariate statistics

- When we work with more than one variable at a time
- How much two variables jointly change?
  - **Covariance**
- How much does a variable "agree" with another?
  - **Correlation** (Pearson, Kendall, Spearman, . . .)

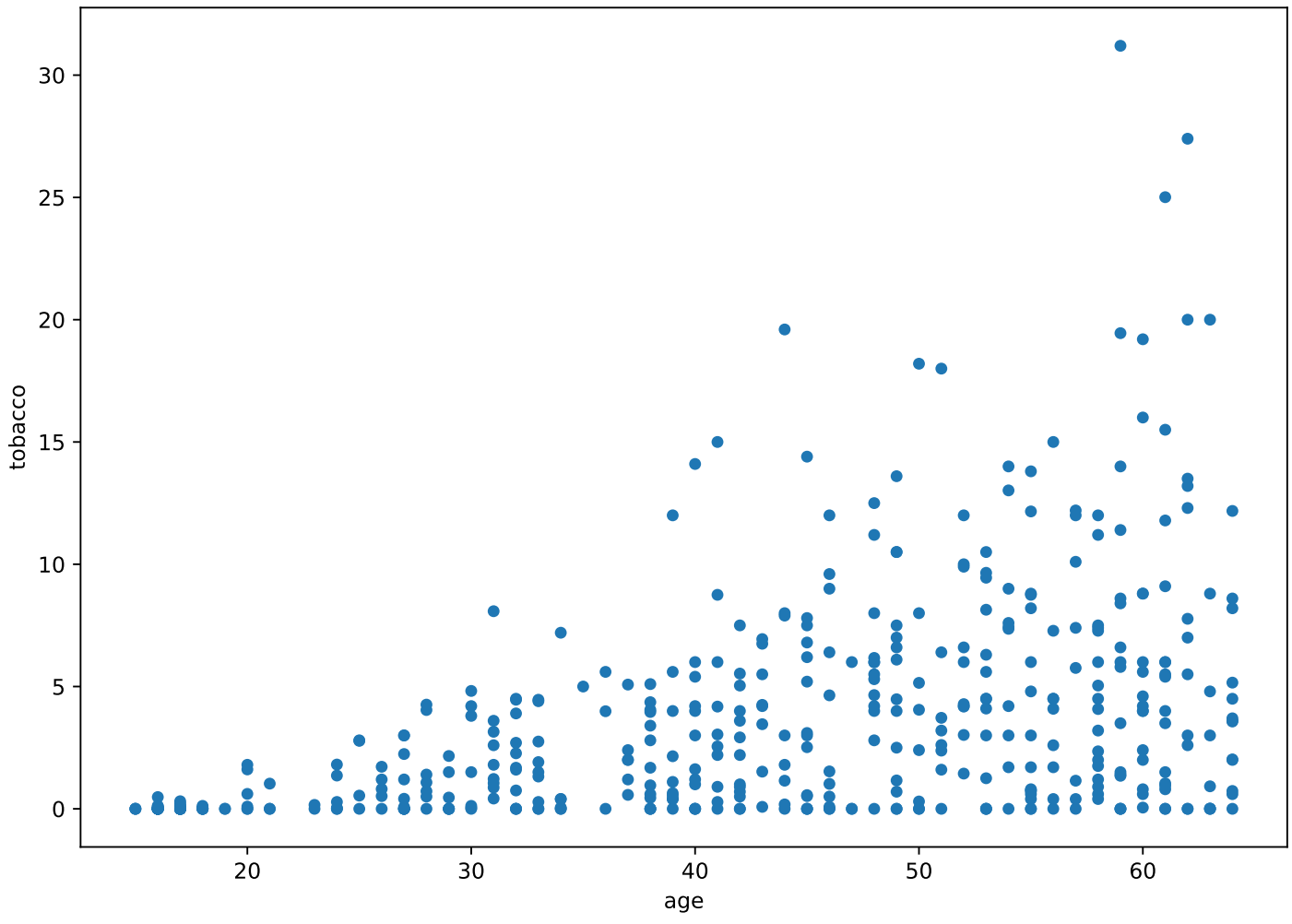
Get the correlation matrix of the variables in the dataset

```
d.corr()
```

```
##                  sbp   tobacco       ldl   ...   obesity
## sbp         1.000000  0.212247  0.158296   ...  0.238067
## tobacco     0.212247  1.000000  0.158905   ...  0.124529
## ldl         0.158296  0.158905  1.000000   ...  0.330506 -
## adiposity   0.356500  0.286640  0.440432   ...  0.716556
## typea      -0.057454 -0.014608  0.044048   ...  0.074006
## obesity     0.238067  0.124529  0.330506   ...  1.000000
## alcohol     0.140096  0.200813 -0.033403   ...  0.051620
## age         0.388771  0.450330  0.311799   ...  0.291777
##
```

Graphical Displays - Scatter plot

```
d.plot.scatter('age','tobacco')
```



Wrap up

- Homework to be submitted (counts as participation)
  - Produce a notebook with the theoretical definitions and examples of all the statistical measures. Include plots to show your examples.
  - Submit in moodle a **pdf** or a self standing **html** resulting from the notebook
  - **deadline** October 17th

References

- Data Mining Concepts and Techniques, Han, Kamber & Pei
- From the web
  - Towards Data Science, Data Science Made Simple, Medium, Wikipedia among others