exploratory data analysis

types of variables:	**Categorical** (nominal no order ex color of eyes or ordinal order ex.lvl of education variables)/ **Numerical** variables: discrete and continous variables
numerical summaries	**quantile**: value that proportion p the data is smaller than Q(p) and 1-p bigger first quantile Q1: p=0.25, **median** Q2: p=0.5 and third quantile: p=0.75 Q3, **IQR** is the interquartile range = Q3-Q1 contains 50% of the data Formula for the rank is p(n-1)+1 if not integer extrapolate with 2 values between with weight
measures of center	**MODE**: most frequent value **MEDIAN**: Q(0.50)/ **MEAN**: average, tot/n if **unimodal** and **symetric** distribution mean=median, right skewed mode<median<mean
variance and sd	
Graphics	**pies**, **barplots** (frequency or rf, any order, specific categories ex faculties), **contingency tables** (2 or + categorical variables), **mosaic plot** (translation of CT, if aligned, independant), **frequency table** (numerical variable, f, rf=proportion, cumulative f, cummulative rf, densities rf/amplitude, order), **hitograms** (translatio of FT, area proportional to class frequency = density, numerical variables, order needed, size can be an interval no precise value as bp), **BOXPLOT** (IQR and 1.5*IQR, put median, LB, UB), **QQ-plot** (compare two distribution theorical and empitrical, if 45° same distribution)

Statistical inference

simpson paradox	heterogenous sources: divide to more homogenous subgroups: ex by major because could bias the proportion : **controlling for the confounding factor** men chose the easiest program whereas women chose the more difficult to enter: the solution is to use a weighted average of the admission rates
sampling the population	**population**: what we want to analyse, want to find the population's parameters, these are true and ifxed values but usually unknown **sample**: what we have, piece of the population chosen randomly, parameters are random variables, should be as large as possible to limit bias, sample have incomplete information, if finite population without replacement of sample can affect results
point estimation	**estimators** an estimator is a parametor calculated with the simple. it tries to estimate the true parameter of the population it is a random variables and parameter are fixed but unknown within a certain certitude: **confidence intervals**
Estimator	to estimate a parameter and its uncertainty: ex: μ, the more sampling, the more precise because variance decreases with N large concentrated distribution around true value

Statistical inference (cont)	
central limit thm	when we sum random variables from the same distribution: sum/n= new variable that follow a normal distribution when n is large special case for proportion (binomial)
estimating variance s^2 and s ~2,	if x follow a normal distrbution, follos khi 2 distribution with n-1 degrees of freedom similar to variance estimation
confidence intervals	from central limit thm: C is a certain value for with prob of (1-a) that the estimator is in the interval, small alfa, bigger interval, not exactly 95/100 but around value, prob, if normal distribution use student distribution so modify CI to be more precise,
for proportions :	^p estimate mean
for median	
for variance	
for the difference of means	when 0 is not in the interval: significante différence
theory of estimation	depends on situation, can evaluate the quality of estimator, good one has nu bias,