<u>Data types</u>: categorical: nominal(no order), ordinal(order); numerical/quantitative: continuous(Interval: no fixed 0, ratio: fixed 0), discrete(only certain values possible);

<u>Tables</u>: **Reference table**(all data), **Demonstration table**(illustrate a point);

<u>Location statistics</u>: mean, median, mode(may be multiple), percentiles($L_P = 1 + \frac{P}{100}(n-1)$, let l and h be the observations at $\lfloor L_P \rfloor$ and $\lceil L_P \rceil$,then $p^{th}\ percentile\ =\ l\ +\ (L\_P\ -\ \lfloor L_P \rfloor(h-l))$;

<u>Scale statistics</u>: **Range**(max – min), **IQR**(Q3– Q1), **Variance**($\frac{1}{n-1}\sum_i(x_i - \overline{x})^2$), **Std**($\sqrt{varience}$), **MAD**(median of |x_i-median|);

z-score($\frac{x-\overline{x}}{\sigma}$, number of std away from mean, above 2.5 = outlier);

<u>Association statistics</u>: **Covariance**($\frac{1}{n-1}\sum_i(x_i - \overline{x})(y_i - \overline{y})$),**Correlation**($\frac{covariance}{\sigma_x \sigma_y}$ or $\frac{1}{n-1}\sum_i z_{x_i} z_{y_i}$);

Strip plot(data: 1 numerical, task: (cluster, outlier), small datasets); **KDE**(create: choose kernel function + bandwidth => kernel for every datapoint => sum kernels, area: per points 1/n, cannot see maximal and minimal); **Skew**: symmetric, right-skewed(positive) = long right tail, left-skewed(negative) = long left tail;

Empirical Cumulative Distribution Function(returns fraction of observations ≤l x)

<u>Search</u>(location, target): lookup(1, 1), browse(1, 0), locate(0, 1), explore(0, 0);

<u>Targets</u>(all data: (trends/patterns, outliers, features), attributes only: (single: (distribution, extremes), many: (dependency, correlation, similarity));

<u>Marks and Channels</u>(marks: geometric primitive(points, lines, areas, ...), channels: appearance of marks(position, colour(colour-blindness, use blue & orange), size, ...));

<u>Design Strategies</u>(position before colour, natural order: (position, length thickness, area, brightness, saturation), no order: (shape, line style, hue));

<u>Idioms</u>: **Scatterplot**(data: 2 numerical, mark: point, channels: horizontal & vertical position, task: (relations, outliers, trends, clusters)), **Bar chart**(data:(1 categorical, 1 numerical), mark: thick line, channels: (length for value, region per key), task: lookup and compare), **Stacked bar chart**(data: (2 categorical, 1 numerical), mark: vertical stack of lines, channels: (length and hue, region per key), task: (compare lowest category and totals, lookup, part-to-whole, trends)), **Normalized stacked bar chart**(same as previous but every bar = 100%), **Line chart**(data: (2 numerical or 1 numerical and 1 ordinal-categorical), mark: points and connecting lines, channels: (lengths for quantitative values, ordered by key into horizontal regions), task: (trends, relationship)), **Histogram**(data: 1 numerical, task: (distribution, counts), recommended bins = $\sqrt{n}$), **Cumulative Histogram**(data: 1 numerical, task: (distribution, thresholds), height of a bin is sum observations less), **Heatmap**(data: (2 categorical-usually ordinal, 1 numerical) or (3 numerical : 2 binned), mark: (2d matrix, indexed by 2 key), channels: colour by quantity, task: (clusters, outliers, trends)), **Box plot**(data: (1 numerical, ?1 categorical if multiple plots),box: Q1+Q3 + median, outliers: dots for above Q3+1.5IQR & below Q1-1.5IQR, endpoints of whiskers: max and min non-outlier), **Violin plot**(box + kde), **Pie chart**(data: (1 categorical + derived $= angle = \frac{\#category}{\#total}360°$) or (1 categorical, 1 numerical), mark: area, channel: angle + colour, layout: radial, scale: at most 12 categories, task: lookup and compare part-to-whole), **Scatterplot matrix**(all possible pairs of axes, by category, scatterplots), **Parallel coordinates**(parallel axes: line representing each item, rectilinear axes: item as point);

<u>Effectiveness</u>(clearly indicates relationship of values, represents quantities accurately, easy comparison, easy observation of ranked order of values, obvious how information should be used);

<u>DMM categories(</u>supervised = predefined target / labelled, global = all the data, local = some data);

<u>Linear regression</u>(kind: (supervised, global) relate y to x by y = ax + b, SSD = Sum of Squared Deviations = sum residuals^2, $R^2 = 1 - \frac{SSD}{(n-1)\sigma_y^2}$);

<u>clustering</u>(kind: (unsupervised, global), partition into groups, depends on distance => equal treatment of attributes(same units for similar attributes, relevant distance units, standardize units for dissimilar attributes e.g. z-score));

<u>k-means clustering(</u>kind of cluster, pick k centroids => assign points to nearest centroid => recompute mean of clusters as centroid => repeat till stable);

<u>Distances</u>: **Manhattan**(add up horizontal and vertical distance),**Network**(sum of lengths of edges, good if sparse network and we know all possible movements), **Euclidean**($\sqrt{x^2+y^2}$);

<u>Decision tree(</u>kind: (supervised, global), separate into 'positive' and 'negative' cases);

<u>Evaluation</u>: **Confusion matrix**, **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$, **Precision** = $\frac{TP}{TP+FP}$, **Recall/Sensitivity** =$\frac{TP}{TP+FN}$, **Specificity** =$\frac{TN}{TN+FP}$

<u>Association rule</u>(kind: (unsupervised, local), find high-confidence associations between subsets / frequently occurring associations, $support(X) = \frac{|X|}{total}$, $confidence(X \Rightarrow Y) = \frac{|X \cap Y|}{|X|}$, creation: find frequent itemsets $length$ 1: A, $length$ 2: $A\ \cap B$, $length$ 3: $A\ \cap B\ \cap B$ until no sufficient support left then replace ∩ with ⇒, e.g. $A\ \cap B\ \Rightarrow C$);

<u>Data Modelling</u>(redundancy: same data too many times, leads to inconsistency, **primary key**: minimal set of attributes in table identifying each row), **Logical Schema**(data modal, logical structure), **Instance**(actual content of database, relation instance = table);

<u>Database Design</u>: **E-R Model**(entity = unique object, entities have attributes, entity set = set of entities that share properties, relationship = association of entities, relationship set = collection of relationship on entity sets, relationship sets can have attributes), **E-R Diagrams syntax**(rectangle = entity set, diamond = relationship set, lines = entity set so relationship set, dotted/dashed line =

rectangle contains attribute set of relationship set, attributes inside rectangle, underline = primary key, simple line = many, arrow = one, double line = all ex: (one to one: ← rel →, many to one: ←rel -, many to many: - rel -, all to one: - rel =);

<u>SQL</u>: **Order**(SELECT => FROM => WHERE => GROUP BY => HAVING), **LIKE**(string comparison, %: 0 or more characters), **BOOL**(AND, OR, NOT), **AS**(renames: old-name AS new-name, optional: borrower b = borrower AS b), **Set operators**(UNION, EXCEPT, INTERSET; used to join queries, eliminates duplicates append ALL to retain everything), **Aggregate functions**(COUNT: number of values, MIN: minimum, MAX: maximum, AVG: average, SUM: sum)

<u>Kinds of data</u>: **Primary data**(collected by me), **Secondary data**(collected by others), **Features**(measurable properties or characteristics);

<u>Scientific method</u>: **Deductive reasoning**(If the premises are true, the conclusion is valid), **Inductive reasoning**(inference from particular to general case, from finite sample to whole population), **Occam's Razor/Parsimony**(prefer the simpler explanation), **Validity**(accurate to the real world, inner: conclusions in the study valid, external: can the conclusions be applied beyond the context of the study), **Reliability**(similar conditions => similar results), **Reproducibility**(ability to replicate findings), **Precision**(errors by the measuring instrument i.e. +-0.5mm), **Accuracy**(difference from reality), **Errors**(sources: (measurement process, definition, inadequacy of technology), random: no pattern, systematic: consistent e.g. offset);

<u>Data sampling</u>: **Population**(complete set), **Sample**(part of the set, biased if some part is overrepresented), **Convenience sampling**(data that is easier to collect, advantages: (time, effort, money, ...), disadvantages: possible bias threat to external validity), **Random sampling**(everyone equally likely to be included, ignoring knowledge about the population), **Stratified random sampling**(strata: disjoint parts forming the population, proportionate stratified random sampling: random sample from every stratum in equal proportion to the proportion of this stratum in the population, disproportionate stratified random sampling: want to overrepresent a stratum), **Voluntary sampling**(individuals select themselves, self-selection bias is difficult to measure);

<u>Data Cleaning and Preprocessing(diagnosing and editing faulty data)</u>: **Sources**(equipment or transmission, collection circumstances, manual entry, non-optimal collection protocol), **Handling**(discard: potentially(lose a lot of data, introduce bias), impute: (fill in constant, data mining or estimate), do nothing), **Noise reduction in time series data**(e.g. smoothing kernel);

<u>Filtering time series(choose window size)</u>: **Median**(replace middle by median of values in window), **Mean**(replace middle by mean of values in window, far and near values have same influence), **Gaussian**(choose width ω, gaussian kernel T for ω, choose pair (t_n, x_n), assign weights $w_i = T(t_n - t_i, \omega)$ to each value x_i, further away = lower weight, $\sum_i^n w_i w_{n-i}, \sum w = 1$), **Convolution**(example: (gaussian, mean), sum weights = 1), **Differentiation**($[w_1, w_0, w_1] \Rightarrow [-w_1, -w_0, 0, w_0, w_1]$);

<u>Variables(features)</u>: **Independent**(see how it's change impacts dependent value), **Confounding**(variable that's not considered), **Differentiation**($p'(t_i) = \frac{p_{i+1} - p_{i-1}}{t_{i+1} - t_{i-1}}$);

<u>Distributions</u>: **Discrete uniform**($P(X = k) = \frac{1}{n}$), **Binomial**($\binom{n}{k} = \frac{n!}{k!(n-k)!}, P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, P(X \le k) = \sum_k^l P(X = k)$), **Geometric**($P(X = k) = (1-p)^{k-1} p$), **Poisson**(rare events), **Expectation**($\sum_k kP(X = k)$, binomial: np), **Variance**($\sum_k (k - E(X))^2 P(X = k)$, binomial: np(1-p)), **Continuous probability**(density function, P(X=x)=0, $P(a \le X \le b) = \int_a^b f(x)dx$, (E, Var, STD replace sum with indefinity integral)), **Normal**(N(μ, σ^2), continuous probability, symmetric around mean μ, if X ~ N(μ, σ^2) then z-scores ~ N(0, 1), normal quantile z_a: value such that $Z \sim N(0,1), P(Z \le z_a) = a$);

<u>Estimations</u>: **Central limit theorem**(for large n the distribution of $\overline{X}$ can be approximated by $N\left(E(X), \frac{\sigma_X^2}{n}\right)$, for large n the mean is normally distributed), **Binomial**(estimate $\hat{p} = \frac{\#success}{n}, E(\hat{p}) = p, Var(\hat{p}) = \frac{p(1-p)}{n}$, central limit theorem applies);

<u>Confidence intervals(interval containing true unknown value)</u>: significance level = α, confidence level = 1-α, **Two sided**($\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$), **One sided**(right: $\mu \le \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$, left: $\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu$, if sigma unknown use sample std, and studentized t_{n-1} instead of z), **Proportion**(same but replace mean with estimate $\hat{p}$), **Difference**($\mu_1 - \mu_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, $\widehat{p_1} - \widehat{p_2} \pm z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$);

<u>Hypothesis testing</u>: **Procedure**(choose null hypothesis and alternative hypothesis, if enough evidence =>reject null hypothesis) **P-value**(probability of observing a more extreme test statistic in the direction of the alternate hypothesis than observed (P(X≥e k)), if p-value < a => reject), **Critical value**(choose test statistic e.g. mean, compute confidence interval based on a and hypothesis, if test statistic outside confidence interval => H_0 rejected);

<u>Error types</u>: **Type I**(erroneously rejecting H_0), **Type II**(erroneously not rejecting H_0 β), **Power of the test**(correctly rejecting H_0, 1-β);

<u>Assumptions-hypothesis tests</u>: independent observations, come from same distribution, means(normality or "large sample size"), proportions(np > 5, n(1-p) > 5), Normality testing(graphical, goodness-of-fit test e.g. Anderson-Darling(gives p-value, small => not normal), kde, QQplot);

<u>Assumptions – Linear regression(diagnostic)</u>: expectation is a linear function(scatter plot), additive error term epsilon(scatter plot, residual plot), normality of epsilon(QQplot), independence of observations(QQ plot of standardized residuals, kde plot of standardized residuals, Anderson-Darling test on standardized residuals), equal variance for all observations, no clear pattern in residual plot, studentized residuals within (-2.5, 2.5)