**Plan**

* **deskreptive Statistik**
* graphical representations/data exploration (Form von daten verstehen) Fraktion verteilung, Proteingesamtmenge
* Normalization (shift\_02) of data and cleanup:

**Data cleanup**

You will be analyzing multiple proteomic datasets together (2 samples, 3 replicates each). It is essential that you explore each dataset and clean it. Cleaning can refer to many things:

* Removing missing values
* Imputing missing values
* Removing low variance columns/rows
* Removing batch effects
* Removing outlier samples (only if it is due to technical issues !!)
* Making sure that data is in the correct format, for example, numbers should be encoded as numeric and not as characters. Categorical variables should be factors etc.
* Re-ordering rows/columns in meaningful and useful ways
* Statistic tests
* Handle the difference in total protein for the reps -> div by total n
* define what outliers are
* clean data of outliers in the reps
* average out the cleaned reps for a 3680 x 25 df
* Could remove proteins with low diff between ctrl RNAse eg 5-10% to reduce data dim
* do stat test by prot on RNAse data in same distribution as ctrl H0 = RNAse not significantly different from ctrl (note maybe only compare the tail end eg frac 19-25)
* Set threshold value to identify the proposed RNAdps Cluster by r^2 value from stat test
* handle the difference in total protein for the reps -> div by total n Global and local minima
* finding per protein: Vector of data from data frac 1-25, maxima where element n-1<n>n+1, global max

**Data exploration**

* Look at the distribution of the overall data, specific samples or features.
* **Visualize the data distribution**
* Visualize the inter-dependencies among specific samples/features of interest
* Check some of your hypothesis like - is something high/low between two conditions etc
* **Maxima and minima**
* **Clustering**
* **Shift bestimmen**
* **Interpretieren und dann:**
* **Definition of selektionskriterien**
* **Data modelling**
* For example, by using **complementary information** from databases on RNA-binding proteins, gene ontology or protein-protein interactions, try to *predict* whether a protein is a **TRUE** RNA-dependent protein. Test how well this could work.