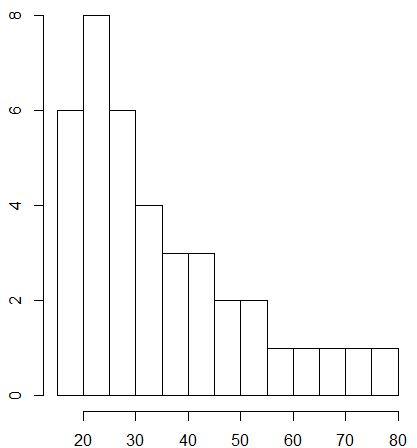
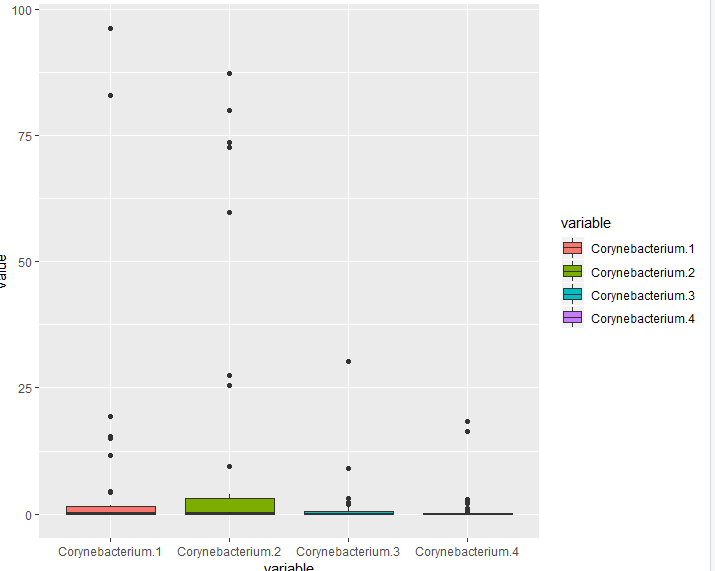
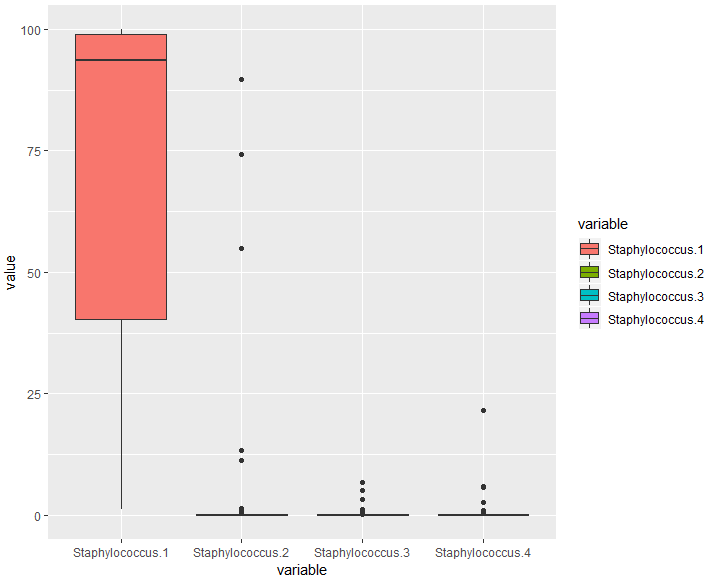
**Data cleaning**

* Remove rows with missing data
* Check if all relative abundances sum up to 100
* factorize BMI
* Remove leading and trailing whitespaces from Gender and factorize
* for strategy 1:
  + Split Age in different groups
  + 
  + per 10 years or equal group sizes?

**Strategy 1**

* Make discrete age groups
* univariate analysis
  + Genus composition, age, bmi, gender
  + Rapport normality of variables and subgroups (in appendix, will need for testing)
  + 
  + 
* bivariate analysis
  + Age and gender
  + BMI and gender
  + Bmi and Genus
  + Genus composition per age (take into account gender?)
    - Take a look at composition of abundance per species
    - Also possible to take a look at composition of abundance within each speies
    - Calculate the mean of each composition per age group and look at differences (H0 states no difference)
  + Correlogram, look at correlation between compositions and age

**Strategy 2**

* hard to know what we need to look for
* thinking about using permutation test?
* Use critical values?
* limit ourselves to pearson, kendall or spearman (R cor function)
  + - <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216287>
      * something with bootstrapping (do the same as in HW3 with the monte carlo stuff?)
    - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/>
    - <https://link.springer.com/article/10.1007/BF02291550>
    - <https://link.springer.com/article/10.1007/BF02589106>
    - Last two links are not open
    - <https://journals.sagepub.com/doi/full/10.1177/1948550617697177>
      * equivalence tests
    - <https://courses.lumenlearning.com/boundless-statistics/chapter/hypothesis-testing-correlations/>
      * 95% critical values for pearson correlation
      * <https://www.statisticssolutions.com/table-of-critical-values-pearson-correlation/>
  + limit ourselves to pearson, kendall and spearman (R cor)
    - pearson looks for strictly linear relation
    - spearman and kendall for monotonic relation (rank based)
    - <https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient/>
    - Kendal
      * based on concordant and discordant pairs, insensitive to error, p-value more accurate with smaller sample sizes
        + concordant: if both members of one obs are larger than their respective members of the other observations
        + discordant: if the two numbers in one observation differ in opposite directions
    - spearman
      * larger values, based on deviations, more sensitive to error and discrepancies in the data
  + Compare cor.test with known tests

**How are the relative abundances of the 8 species related to one another?**

* Maybe if one genus from a species is highly abundant, the other genera are less abundant because they occupy the same environment
* If one bacteria has a higher abundance, the values for the others go down (because working with relative abundances and not absolute values)
* Maybe one genus lives closely with a different one (symbiotic?)

**Schedule**

- 10/12 analyses done and reviewed

-1/12, test developed and tested, prepping & descriptives done (2,2)  
-8/12, hyp testing done, rapported and concluded (2)  
-9&10/12 review the result, agree on, amend where needed (2)

-13/12 presentation done (2)

-27/12 rapport done (4)

**Division of tasks**

-Stijn and Robin do the test development for correlation, look for literature etc

-Lin and meran do the descriptive part

meet on 4 december! to decide on how to proceed