(Short) Problem Set 6

The sixth problem set requires you to solve two small

Grading will reflect your performance on both the coding as well as the documentation and interpretation tasks – however, there is no fixed grading scheme between the two categories.

This problem set is **due on July 19th by 12.00** through the wuecampus submission functionality.

Cooperation with other groups is not permitted and will lead to severe credit deductions.

1. Beer Analytics: Brewtoad.com offers a comprehensive list of craft beer recipes and statistics. Crawl the listing website (https://www.brewtoad.com/recipes?page=5&sort=rank&view_as_table=true) to extract Name, Style, OG, FG, IBU, ABV and Color for at least 1,000 brews each from four distinct brewing styles. [try to use not too similar styles]
    a. Run a Principal Components Analysis and explain the results.
    b. Use k-Means Clustering with k=number of brewing styles to cluster your data (omitting the brewing style). Assess the quality of your clustering graphically and by means of a table.
    c. Use hierarchical clustering to create dendrograms for different linkage behaviors. Assess the quality of your clusterings by means of tables.

2. With Obama's second term nearing an end we are all getting excited (or amused) by the raging campaigning between Hillary and Donald. Let's put the fight into historic perspective by analyzing historical voting patterns of states.

   The package *cluster* contains the data set *votes.repub*:

   This is a data frame with the percentage of votes given to the republican candidate in presidential elections from 1856 to 1976. Rows represent the 50 states, and columns the 31 elections.

   Source: S. Peterson (1973): A Statistical History of the American Presidential Elections. New York: Frederick Ungar Publishing Co. Data from 1964 to 1976 is from R. M. Scammon, American Votes 12, Congressional Quarterly.

   a. Visualize this spatial-temporal data using line-charts identifying states by color/facetting. What are obvious disadvantages of this representation?
   b. Use the US census regions (https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States) to group the states and replicate the analysis from a. – i.e., colored or facetted by region.
   c. Run hierarchical clusterings with complete and average linkage on this data set and compare the results by looking at the dendrograms.
      - Analyze to what extent the distance matrix is affected by NA values – e.g., by comparing the distance of a state pair with NAs (prior to filtering) and after filtering without NAs.
      - Discuss what the distance metric measures here. Assuming we want to use such data to look forward, what kind of post-processing may be warranted to avoid historical distortions carrying too far into the future?
   d. Create a heatmap as illustrated in https://cran.r-project.org/web/packages/dendextend/vignettes/introduction.html#gplots and discuss the additional insights from this visualization.
   e. Extract 4 clusters from your dendrograms and discuss if they are aligned with the census area – i.e., can voting behavior be approximated by spatial proximity?