State the region and the domain category that your data sets are about

Data is collected at location Nes, Ameland. The data is from water hight and water temperature.

Datasets were obtained here:

Water hight: <https://waterinfo.rws.nl/#!/details/publiek/golfhoogte/Nes(NES)-1/Significante___20golfhoogte___20in___20het___20spectrale___20domein___20Oppervlaktewater___20golffrequentie___20tussen___2030___20en___20500___20mHz___20in___20cm>

Water temperature: <https://waterinfo.rws.nl/#!/details/themakaarten/Waterbeheer/Nes(NES)-1/Temperatuur___20Oppervlaktewater___20oC>

The data was downloaded for 28 days into history. The data source automatically updates and removes data from before these 28 days. For this reason, the data used is supplied in the git repository, as newer data might give different results.

State the research question

The research question states: How does the independent variable water temperature influence the dependent variable of water hight?

H0: There is no effect on water hight caused by the water temperature

H1: There is a effect on water hight caused by the water temperature

Justify the chosen data storage and processing approach

The data is stored in their original file and are kept original, this way all processes that are done are shown.

The processing of the data is done by multiple functions.

The data is first cleaned, then translated, then unused data is dropped.

After this the used data is optimized for Furter processing. (optimized by combining some columns and renaming, converting to other dtypes)

Justify the chosen analysis approach

For analysis, the water hight is divided in 3 equal sized groups, grouped by temperature range/number of datapoints.

This was done to be able to use a one way anova. The groups are sorted by temperature and then divided in equal sized groups.

This might cause the groups to have the same temperature at the borders, but this should not be a problem since this is should not influence the anova testing.

A one way anova was performed because this test is used to compare the variance in the group means within a sample while considering only one independent variable. The independent variable being the temperature range of the groups and the dependent variable being the hight of the waves. If there is no effect on wave hight by the water temperature, then the mean hight of waves should be the same in each temperature group.

Justify the chosen data visualization approach

For the visualization, the data is first shown separate.

Then a scatter plot is shown with a regression line.

After the regrouping, a one way anova is done on the grouped data. This gives a pvalue to indicate if there is a significant effect.