1. **Problem definition and description**

The main purpose of the project is to evaluate the contribution(correlation) of each ingredient material on the concrete strength using given data. Firstly, it was approached with the Pearson correlation which was followed by those assumptions: continuous variables, related pairs, absence of outliers, bivariate normality, linearity, and homoscedasticity. Outliers were deleted based on the 6-sigma rule. Since normality, linearity, and homoscedasticity are difficult to test, it would be acceptable to conduct:

* Normality test (univariate normality)
* Plotting scatterplot (check for linearity)

For those who do not satisfy the assumptions, the Spearman correlation was calculated instead.

1. **Core code**

def exclude\_outliers\_arr(arr: np.array):

# exclude outliers based on the 6 sigma rule

std\_temp = arr.std()

mean\_temp = arr.mean()

lower\_limit = arr > mean\_temp - 3 \* std\_temp

upper\_limit = arr < mean\_temp + 3 \* std\_temp

return arr[lower\_limit & upper\_limit]

def correlation\_test(p: int):

if p < 0.05:

print("The H0 rejected. (correlated)")

else:

print("The H0 can’t be rejected. (uncorr)")

"""

H0: The two sets of data are uncorrelated. α = 0.05

"""

def calculate\_correlation(x: np.array, y: np.array, column\_name: str):

if column\_name == 'Water (component 4)(kg in a m^3 mixture)':

r, p = stats.pearsonr(x, y)

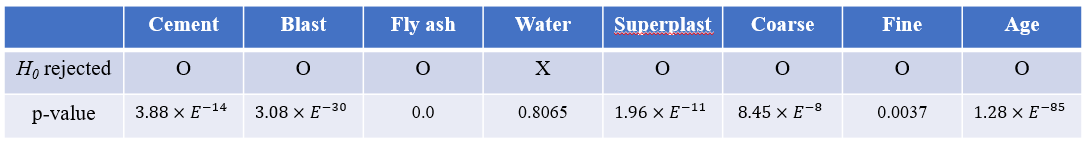
else:

r, p = stats.spearmanr(x, y)

print("corr:", r, "p-value:", p)

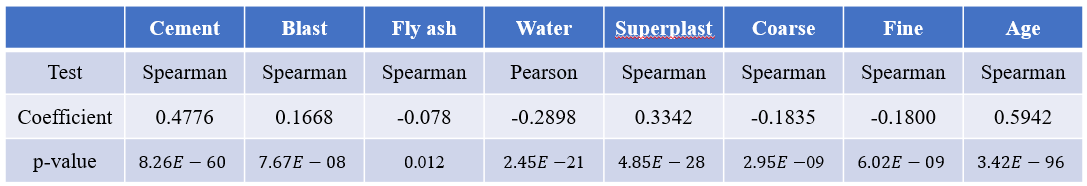
correlation\_test(p)

1. **Results and plots**



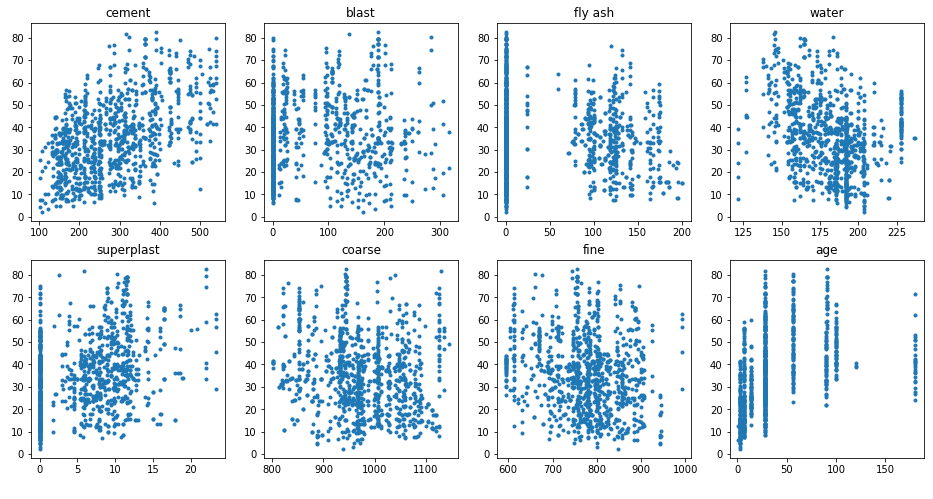
**Table 1. Result of the normality test**

(H0: The dataset follows normal distribution, α=0.05)

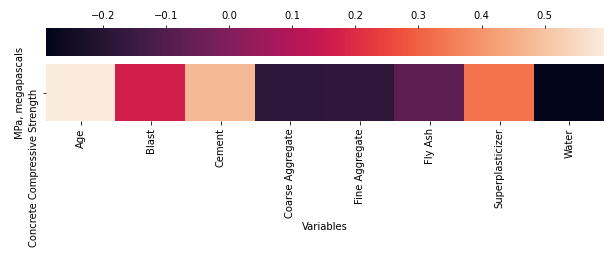


**Table 2. Result of the correlation coefficient**

(H0: The two sets of data are correlated., α=0.05)



**Figure 1. Scatter plots of each variables  
with concrete strength**



**Figure 2. Heatmap based on the Table 2 (Corr.)**

1. **Discussion**

The project could be divided into two parts: i) to test if the variables’ distribution matches the assumption, ii) to calculate the correlation and analyze the results.

In part i), the normality test and plotting scatterplot were conducted (Table 1, Figure 1). ‘Water’ was the only one that satisfies the normality. Plus, the scatter plots barely show any linearity, but slightly represent a positive correlation for cement and a negative correlation for water.

Correlation Coefficients were estimated based on the results obtained in part i) (Table 2, Figure 2). ‘Water’, which matches the assumption, was the only variable that calculated by the Pearson method, and the other variables were by Spearman correlation since they do not satisfy normality.

Components did not show strong correlations in general except age and cement mixture shows a high positive correlation with concrete strength. Although correlation coefficient could be a way to figure out the relationships between variables, it could not explain the causal relationships. To make sure the real contribution of each component, descriptive methods such as regression analysis should be done.

1. **Refernces**

Lund Research Ltd, 2018. *Pearson Product-Moment Correlation*. Available online: https://statistics.laerd.com/  
statistical-guides/pearson-correlation-coefficient-statisti  
cal-guide.php (Accessed on March 31, 2021)