**IEDA 3302 IT for Ecommerce applications**

**HW 4. Fall 2023. *Max score: (2+2+2)+4+4+(4+2)= 20***

**Q1.** Write python functions to create the following visualizations. In all cases, please use the wines data set (with all data of red and white wines in a data frame).

**(a)** Parameters *residual sugar* and *alcohol* have a slightly negative correlation. Make a scatter plot to show the pairwise distribution of these two attributes, with each point illustrated by a circle that is yellow (for white wine) or red (for red wine).

**(b)** Suppose we wish to examine the extent of this relationship for low, medium or high quality wines separately. Make a visualization with separate scatterplots of these three categories plotted side by side in a single row.

**(c)** Suppose we wish to examine the distribution of the total sulphur dioxide in white wines. Plot the histogram of the total sulphur dioxide in all the white wines. Comment whether it looks close to a Gaussian distribution. Now use the distfit function (from the distfit library) to find the best fitting Normal distribution and plot the histogram together with the best fitting Normal distribution pdf (https://erdogant.github.io/distfit/pages/html/Performance.html).

**Q2.** From the wines data set, after normalizing all parameter values, delete the four parameters with the lowest variance. Run a PCA on this reduced data set and report the eigenvalues and eigenvectors. Can you use this analysis to further reduce the dimensionality of the data set to four parameters only? Explain your answer.

**Q3.** Recall that in a single parameter regularized regression, we find the optimal parameters that minimize the error function defined by . Confirm that each optimality condition, = 0, is a linear equation in terms of the *wi*.

**Q4.** Consider the price data (use the green line only) of the price of a smartphone over the last 90 days (please use the Market price data only, shown in green points/lines) from website price.com: <https://www.price.com.hk/product.php?p=574103&ct=history>.

**(a)** Use the scikit learn regression pipeline to fit polynomials of degree 1, degree 2 and degree 3 on the data for alternate days. Then use the remaining data to compute the R2 value in each case.

**(b)** Do you think that the model is a good predictor for this product over the next 90 days? Give reason.

[Hint: you can crawl the page for the price data, or alternatively just copy it from the source of the page and extract the useful values via a short python function].