

# P556 HOMEWORK 1

Fall 2022

Due on Sep 21st, 11:59pm

The goal of this homework is to expose you to various python packages and functions. You can use whatever packages and functions you want to complete the following questions.

All homework assignments need to be submitted to IU github as a jupyter notebook (.ipynb). To make things easier, please name your submission with prefix "hw1\_groupx\_xx\_xx.xx.ipynb". (Use the latest version of group assignment csv file. The group id is the index for your group assignment, see figure below. xx indicates each of your group member's last name). All tasks can be done in a single jupyter notebook.

1	Groups of 3	
2	Yash Rajabhau Vibhute	Nik
3	Muhammad Siddiq Khan	Mo
4	Mandar Mhatre	Chi
5	Aadhya Kaul	Jigr

Figure 1: Use the group index indicated in the group csv file. The group id starts from 2 as shown in this picture.

## Question 1: Density Plot (20 points)

The exercise will work on the Australia Fire dataset, which can be located here: <https://www.kaggle.com/datasets/carlosparadis/fires-from-space-australia-and-new-zeland>

We will work with fire\_nrt\_V1\_96617.csv which is described here: <https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/viirs-i-band-375-m-active-fire-data>

Remember that during the lecture, we emphasized several times that plotting your data is very important. Here plot the longitude vs latitude several ways within a single figure (each in its own axes):

- Using the matplotlib defaults. (matplotlib is a python package we used in our sample code).
- Adjusting alpha and marker size to compensate for overplotting.
- Using a hexbin plot.
- Subsampling the dataset.

For each but the first one, ensure that all the plotting area is used in a reasonable way and that as much information as possible is conveyed; this is somewhat subjective and there is no one right answer.

Answer this question based on what you find: in what areas are most of the anomalies (measurements) located?

## Question 2: Visualizing class membership (20 points)

Visualize the distribution of Brightness temperature I-4 as a histogram (with appropriate settings). Let's assume we are certain of a fire if the value of temperature I-4 is saturated as visible from the histogram.

- Do a small multiples plot of whether the brightness is saturated, i.e. do one plot of lat vs long for those points with brightness saturated and a separate for those who are not (within the same figure on separate axes). You can pick any of the methods from the question above that you find most suitable. Can you spot differences in the distributions?

- Plot both groups in the same axes with different colors. Try changing the order of plotting the two classes (i.e. draw the saturated first then the non-saturated or the other way around). Make sure to include a legend. How does that impact the result?
- Can you find a better way to compare the two distributions?

### Question 3: Regression Sydney Dataset (30 points)

You can load the Sydney dataset from <https://www.kaggle.com/shree1992/housedata> where you can also find a description. The goal is to predict the 'price' column. For this task, you can ignore the date.

- Determine which features are continuous vs. categorical. Drop rows without a valid sales price.
- Visualize the univariate distribution of each continuous variable, and the distribution of the target. Do you notice anything? Is there something that might require special treatment?
- Visualize the dependency of the target on each continuous feature (2d scatter plot).
- Split the data in training and testing set. Use ColumnTransformer to encode categorical variables. Impute missing values using SimpleImputer. Evaluate Linear Regression (OLS), Ridge, Lasso and Elasticnet (although we haven't talked about these methods yet, but you can easily find references online and you can use provided functions by Scikit-learn or other packages directly) using cross-validation with the default parameters. Does scaling the data with StandardScaler help? Use the preprocessing that works best going forward.

### Question 4: Classification on the 'credit-g' dataset (30 points)

You can download the dataset with 'fetch\_openml('credit-g')' and see its description at <https://www.openml.org/d/31>

- Determine which features are continuous and which ones are categorical.
- Visualize the univariate distribution of each continuous variable, and the distribution of the target.
- Split the data in training and testing set. Preprocess the data (such as treatment of categorical variables) and evaluate an initial Logistic Regression model (directly use the provided function) with a training/validation split.
- Use ColumnTransformer to encode categorical variables. Evaluate Logistic Regression, Linear Support Vector Machines and nearest neighbors (You can directly call these functions). How different are the results? How does scaling the continuous features with StandardScaler influence the results?