Lab 1: Data and statistics

In this lab, we will begin by working with data to calculate fundamental statistics. Next, you will develop a data loader to import a real dataset. Utilizing the functions you have created, you will gain valuable insights from the dataset and generate data plots for visualization purposes.

The data to be used is available in Canvas and shows information about different housing districts. The name of the file is **housing.csv**.

Note that in addition to the Tasks described below the following exercises in the course book are mandatory for completing this lab (Chapter-Exercise): **3-2**, **5-3**, **8-4**, **10-6**, **10-7**.

Task 1: Create basic functions:

Your task is to implement several functions from scratch without using built-in functions. These functions will operate on the input list of grades and return specific values. Here are the functions you need to implement:

Min: This function should find and return the minimum value from the input list.

Max: Implement this function to find and return the maximum value from the input list.

Mean: You need to calculate and return the average value across values in the input list.

Variance: This function should compute and return the spread of the values in the input list, which measures how far each number in the set is from the mean.

Standard deviation: Implement this function to calculate and return the dispersion of the input list relative to its mean.

Median: You should sort the input list and return the middle number from the sorted list as the median value.

Median absolute deviation: This function should compute the average distance of the input points from the median and return the result.

For more information regarding these functions and how to implement some in python refer to lectures 4 and 5 and to the following online resource which defines most of the quantiles mentioned previously https://en.wikipedia.org/wiki/Median_absolute_deviation.

Task 2: Get some insight:

Now you will continue by defining a list. Imagine that you are a teacher and you just received the grades that a group of students got in an exam. This list will be used to calculate some interesting values.

```
grades = [8, 6, 1, 7, 8, 9, 8, 7, 10, 7, 6, 9, 7]
```

2.1 For the provided array, obtain: Min, max, mean, standard deviation, median and median absolute deviation.

2.2 Explain what each of these values mean.

Task 3: Load and inspect the housing data:

- Load the housing data contained as a resource provided for this lab: **housing.csv**
  - **Use built-in functions**

```
import pandas as pd

pd.read_csv ("housing.csv")
```

You can find in book (page 46-47) the way of download the data from the internet

- The output of the load function should look like this:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	NEAR BAY
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	3.6591	299200.0	NEAR BAY
7	-122.25	37.84	52.0	3104.0	687.0	1157.0	647.0	3.12	241400.0	NEAR BAY
8	-122.26	37.84	42.0	2555.0	665.0	1206.0	595.0	2.0804	226700.0	NEAR BAY
9	-122.25	37.84	52.0	3549.0	707.0	1551.0	714.0	3.6912	261100.0	NEAR BAY
10	-122.26	37.85	52.0	2202.0	434.0	910.0	402.0	3.2031	281500.0	NEAR BAY

This dataset provides a set of housing districts and their information. Take a moment to analyze what information it shows.

Task 4: Apply functions on real dataset:

Use the functions created in the previous tasks to find information related to this dataset, you may need to modify these functions to receive values inside the table.

4.1 Count the number of districts loaded in this exercise

4.2 Calculate the mean of house values among all the districts

4.3 Create a histogram for **ammount_of_households, median_income , housing_median_age and median_house_value**.

- An example of how to create a histogram using matplotlib is shown here:

```
from matplotlib import pyplot as plt

grades = [8, 6, 1, 7, 8, 9, 8, 7, 10, 7, 6, 9, 7]

plt.hist(grades)

plt.title("histogram")

plt.show()
```

4.4 What do you notice about the graphs? Specifically focus on end of housing_median_age and median_house_value graphs.

4.5 What do you think about the magnitude of the values in median_house_value? What may have happened to them in the processing, think about the units.

Additional Task: For each ocean proximity category in the dataset calculate the mean house value.

Hint: To perform this task you will need to group and calculate the mean of all districts per "ocean_proximity" category, an example in how to execute this task can be found in the following picture:

Median_house_value	Ocean_proximity
450,000	NEAR BAY
390,000	NEAR BAY
400,000	NEAR BAY
500,000	ISLAND
480,000	ISLAND

Calculate mean of these values only if they share category

Mean Value	Ocean_proximity
413,333	NEAR BAY
490,000	ISLAND

4.6 Think about the following two cases:

1. Let's think about the first task of this lab. Imagine you are a teacher, and you want to analyze the performance of your students on a recent exam. You have the exam scores of all the students in your class. Which metric is more adequate for this analysis?

2. Consider a study on the income distribution of a country. The dataset includes the incomes of individuals, and it is known to have a few extremely high-income earners, such as billionaires. Which metric would me better in this case?