Lab 1: Working. with data and statistics

In this lab we are going to start working with data to calculate basic statistics and then we will visualize the end results.

We will begin by creating an array of grades which will be used to calculate some interesting values.

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

This array contains data with grades from an exam, take a minute to analyze the data and get basic insights and get used to it.

In this lab you will implement and use the following functions from scratch: (do not use built in functions). These functions take the input list grades and return the specified value.

Min: Return the minimum value from the input.

Max: Return the maximum value from the input.

Mean: Return the average from the values in the input.

Variance: Return the spread of the values on the input (how far each number in the set is from the mean).

Standard deviation: Return the dispersion of the input relative to its mean.

Median: Return the middle number on a sorted input.

Median absolute deviation: Return the average distance of the input points from the median.

For more information regarding these functions and how to implement some in python refer to lectures 4 and 5.

Task 1: Get some insight:

- Calculate the min, max, spread, and mean of the array provided above. What do these four values mean? Explain.

Task 2: Some more insight:

- Following the same logic calculate the variance and the standard deviation. Explain once again what these values mean.

Task 3: Median and Median absolute deviation

- Finally calculate the median and the Median absolute deviation of the data

Task 4: Let´s visualize the data!

- By using the following code, we will be able to see how the data looks in a histogram:

```
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()
```

- What is the most frequent value?
- Is there a value that does not match with the rest? Or appears too far from the rest in the graph? Those are called outliers

Task 5: Apply to real dataset

- 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

- Provide some insights from this data:

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 a dataframe.

a. Count the number of districts loaded in this exercise
b. Calculate the mean of house values among all the districts
c. Create a histogram for **ammount_of_households, median_income , housing_median_age and median_house_value**.
d. What do you notice about the graphs? Specifically focus on end of housing_median_age and median_house_value graphs.
e. 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.
f. For additional exercise repeat b & c but now divided by each categorical value of "ocean_proximity" (I.e. return the mean house value for districts NEAR BAY).