Data Analysis Project

# Problem Statement

You can choose between two data-sets:

* **Dataset A: Credit Risk**

The dataset contains 1000 entries. In this dataset, each entry represents a person who takes a credit by a bank. Each person was classified as good or bad credit risks according to the set of attributes and the amount of credit was given.

* **Dataset B: Real State Data**

This comprehensive real estate dataset contains over 5,000 property listings from South Carolina, collected in 2025 from Realtor.com. The dataset captures diverse property types including single-family homes, condominiums, land parcels, townhomes, and other residential properties. This dataset provides a rich snapshot of South Carolina's real estate market.

The goal is to analyze the data and generate a report that summarizes some info about the datasets.

# Motivation

It is a sneak peek into what kind of work is waiting for you in the next semester without giving too many spoilers. In the next course, you will see how to solve these kinds of problems on a bigger scale with better tools but it will also show the basics of what those tools are easing.

# Good to Know Python Concepts

Concepts such as data types, functions, loops, data structures (list, dictionary, etc), *file reading and writing (new concept)* and *data visualization (new concept)* are good to know/explore.

# Choose your dataset

* **Dataset A: Credit Risk**

The dataset contains 1000 entries. In this dataset, each entry represents a person who takes a credit by a bank. Each person was classified as good or bad credit risks according to the set of attributes and the amount of credit was given.

**Attributes:**

* 1. Age (numeric)
  2. Sex (text: male, female)
  3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
  4. Housing (text: own, rent, or free)
  5. Saving account (text - little, moderate, quite rich, rich)
  6. Checking account (text - little, moderate, quite rich, rich)
  7. Credit amount (numeric, in Deutsch Marks)
  8. Duration (numeric, in month)
  9. Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

note**:** A checking account is for daily spending, providing easy access to funds via debit cards and checks for bills and transactions, while a savings account is for long-term saving, typically offering interest on deposits to help money grow.

* **Dataset B: Real State Data**

This comprehensive real estate dataset contains over 5,000 property listings from South Carolina, collected in 2025 from Realtor.com. The dataset captures diverse property types including single-family homes, condominiums, land parcels, townhomes, and other residential properties. This dataset provides a rich snapshot of South Carolina's real estate market.

**Attributes:**

1. type: (text: Primary property category (single\_family, condos, land, townhomes, multi\_family, farm))
2. sub\_type: (text: Detailed property classification (condo, townhouse, co\_op))
3. sqft: (numeric: Property size in square feet)
4. baths: (numeric: Number of bathrooms (decimal values indicate half baths))
5. beds: (numeric: Number of bedrooms)
6. stories: (numeric: Number of floors/stories in the property)
7. year\_built: (numeric: Construction year of the property)
8. listPrice: (numeric: Property listing price in USD)

# Steps dataset A

1. Look at the data and familiarize yourselves with it, understanding the structure and column names.
2. Read the data, use the **csv** module to read the data from a CSV file
3. Explore data: try to print out its column names, print out how many rows it has, check for missing data.

**Don’t forget to pay attention to the description of the dataset.**

1. Transform your data. Think about which attribute can be modified, converted, or filtered.
2. Compute basic statistics:
   1. What is the average Credit amount given?
   2. What is the average age of all clients?
   3. Is there a difference between the credit amount given for Male and Female?
   4. What was the most common purpose overall for credit given?
3. Print out your results
4. Using the **matplotlib** module to visualize results, try to visualize the quantity of each Purpose in a bar chart.
5. Try the same chart again but separating by Sex. One for Female and other for male. Is there a difference?
6. Give your insights of your findings of the data. How the missing data impact your analysis?

bonus: explore more options of data description and plots visualization!

# Steps dataset B

1. Look at the data and familiarize yourselves with it, understanding the structure and column names.
2. Read the data, use the **csv** module to read the data from a CSV file
3. Explore data: try to print out its column names, print out how many rows it has, check for missing data.

**Don’t forget to pay attention to the description of the dataset.**

1. Transform your data. Think about which attribute can be modified, converted, or filtered.
2. Compute basic statistics:
   1. What is the average of Prices?
   2. What is the average size of the Houses?
   3. Is there a difference in the price for “single\_family”,” condos” and “land”?
   4. What is the average of Prices for the different quantities of bedrooms? Use 1,2,3 and 4 bedrooms.
3. Print out your results
4. Using the **matplotlib** module to visualize results, try to visualize the Pricing by the size of the property in a Scatterplot.
5. Try a bar chart to check the quantity of types of properties.
6. Give your insights of your findings of the data. How the missing data impact your analysis?

bonus: explore more options of data description and plots visualization!