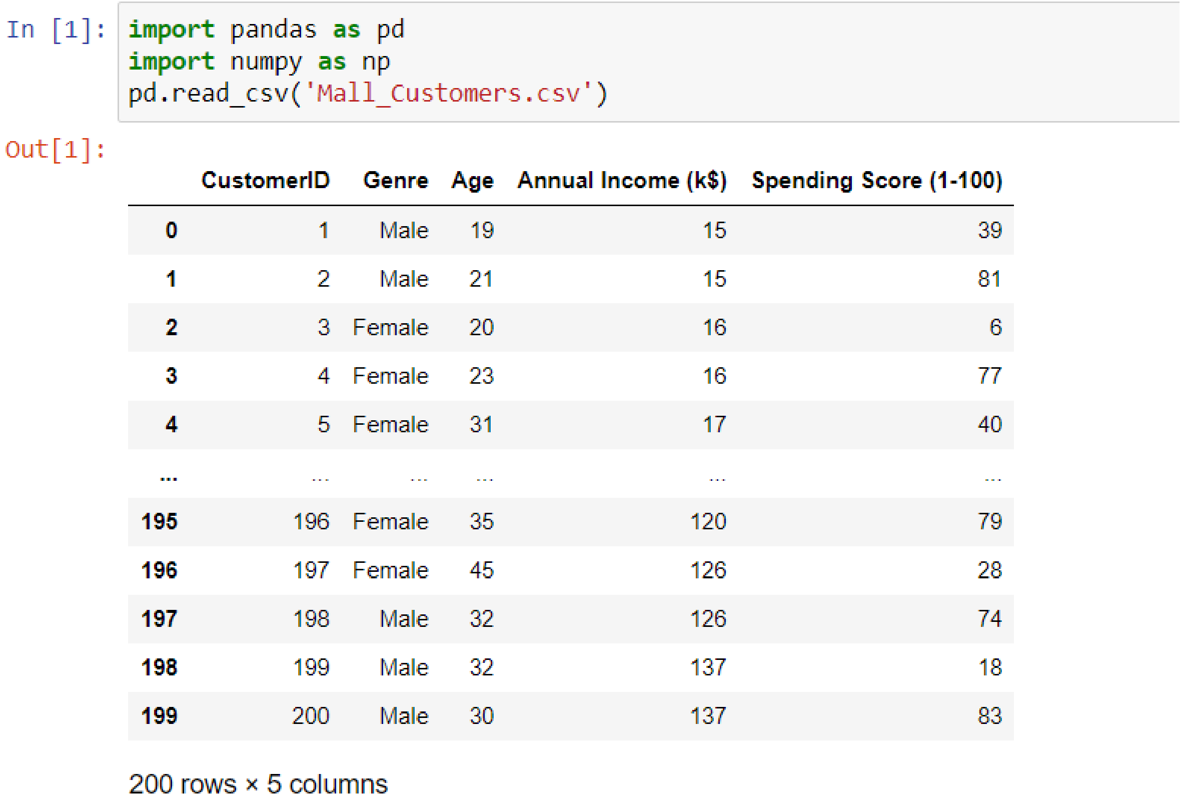
Machine Learning

Project plan

Name: Md Touhidul Islam

Source of dataset: https://www.kaggle.com/shwetabh123/mall-customers

In this project I am going to work with mall Customer's data. It has 200 rows and 5 columns. The columns in this dataset are CustomerID, Gender, Age, annual income Spending Score. Most of the calculation will do using age and Annual income of the customer.



```
In [1]: import pandas as pd
        import numpy as np
        pd.read_csv('Mall_Customers.csv')
```

Out[1]:

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

This the view of our data Mall_Customers.csv.

What am I going to do with this data?

1. Read in the data
2. Explore and visualize the data
3. Preprocess the data for machine learning
4. Apply machine learning methods
5. Explore and visualize the results

- **Read in the data:** First, I will read data using pandas.
- **Explore and visualize the data:** Exploration allows for deeper understanding of a dataset, making it easier to navigate and use the data later. The better an analyst knows the data they're working with, the better their analysis will be. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. We can see clean data.
- **Preprocess the data for machine learning:** Above I have selected dataset. Then must import all the crucial libraries like NumPy, pandas, Matplotlib.

  NumPy: Used to calculation. Also add multidimensional arrays and matrices in code.

  Pandas: This library use for data manipulation and analysis.

  Matplotlib: This library is a 2D plotting library that use to plot any types of charts.

  This stage we will import our dataset and must fix our directory. Here we are using Jupyter IDE. After that we will extract the independent and dependent variables. Identifying and handling the missing values and deleting the row, we will calculate mean, median or mode of a particular column.

  Encoding the categorical data: Categorical data refers to the information that has specific categories within the dataset. We have above one categorical variable- Gender. Here we will encode this Gender variable.

Splitting the dataset: Splitting the dataset is the next step in data preprocessing in machine learning. Every dataset for Machine Learning model must be split into two separate sets – training set and test set. Usually, the dataset is split into 70:30 ratio or 80:20 ratio. This means that you either take 70% or 80% of the data for training the model while leaving out the rest 30% or 20%. The splitting process varies according to the shape and size of the dataset in question.

Here we will train and test the data.

Example for split the data set:

To split the dataset, you have to write the following line of code –

```
 from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)
```

Here, the first line splits the arrays of the dataset into random train and test subsets. The second line of code includes four variables:

x_train – features for the training data

x_test – features for the test data

y_train – dependent variables for training data

y_test – independent variable for testing data

Feature scaling: here we will work with feature scaling. Used data: Age and Annual Income.

Feature scaling marks the end of the data preprocessing in Machine Learning. It is a method to standardize the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds.

- **Apply machine learning methods**: Here, we will work with method of machine leaning.

  Method:
  Regression and classification.

- **Explore and visualize the results:** Here, we will visualize result of the method.