Financial data
This dataset (.csv) collects 200+ financial indicators for all the stocks of the US stock market. The financial indicators are found in the 10-K filings that publicly traded companies release yearly.
The last column of the dataset represent the class of each stock, where:

- if the value of a stock increases during **2018**, then class=1;
- if the value of a stock decreases during **2018**, then class=0.

In other words, stocks that belong to class 1 are stocks that one should buy at the start of year 2018, and sell at the end of year 2018.

1. Some financial indicator values are missing (nan cells), so the user can select the best technique to clean each dataset (dropna, fillna, etc.).
2. There are outliers, meaning extreme values that are probably caused by mistypings. Also in this case, the user can choose how to clean each dataset (have a look at the 1% - 99% percentile values).
3. The third-to-last column, Sector, lists the sector of each stock. Indeed, in the US stock market each company is part of a sector that classifies it in a macro-area. Since all the sectors have been collected (Basic Materials, Communication Services, Consumer Cyclical, Consumer Defensive, Energy, Financial Services, Healthcare, Industrial, Real Estate, Technology and Utilities), the user has the option to perform per-sector analyses and comparisons.
4. The second-to-last column, PRICE VAR [%], lists the percent price variation of each stock for the year. For example, if we consider the dataset 2018_Financial_Data.csv, we will have **percent price variation for the year 2018** (meaning from the first trading day on Jan 2018 to the last trading day on Dec 2018).
5. The last column, class, lists a binary classification for each stock, where
    - for each stock, if the PRICE VAR [%] value is positive, class = 1. From a trading perspective, the 1 identifies those stocks that an hypothetical trader should **BUY** at the start of the year and sell at the end of the year for a profit.
    - for each stock, if the PRICE VAR [%] value is negative, class = 0. From a trading perspective, the 0 identifies those stocks that an hypothetical trader should **NOT BUY**, since their value will decrease, meaning a loss of capital.

This dataset has been developed in order to understand whether or not it is possible to classify the future performance of a stock by looking at the financial information released in the 10-K filings.
How can you achieve that?

1. Build a classification model for the stocks that would and would not increase their value in 2018

Big Mart Sales

This dataset contains information collected by *BigMart* (a supermarket chain in the US) about sales data for 1559 products across 10 stores in different cities.
The attributes recorded for each product and store have been defined as here below

- **Item_Identifier:** Unique product ID
- **Item_Weight:** Weight of product
- **Item_Fat_Content:** Whether the product is low fat or not
- **Item_Visibility:** The % of total display area of all products in a store allocated to the particular product
- **Item_Type:** The category to which the product belongs
- **Item_MRP:** Maximum Retail Price (list price) of the product
- **Outlet_Identifier:** Unique store ID
- **Outlet_Establishment_Year:** The year in which store was established
- **Outlet_Size:** The size of the store in terms of ground area covered
- **Outlet_Location_Type:** The type of city in which the store is located
- **Outlet_Type:** Whether the outlet is just a grocery store or some sort of supermarket
- **Item_Outlet_Sales:** Sales of the product in the particular store. This is the outcome variable to be predicted.

BigMart has collected such data in order to understand what kind of products sell more in what kind of stores. Furthermore, it would like to investigate how much the *item_visibility* impacts sales. We, as third users, may be interested in segmenting the products according to their specifics and/or sales at different stores.
How can you achieve that?

1. Build a predictive model and find out the sales of each product at a particular store (or at generic stores with different characteristics
2. Cluster items according to the available covariates, perhaps considering also the different sales in different stores (you should *spread()* the dataset for this last task.

Brazilian Houses

This dataset contains information about 10962 houses to rent in different Brazilian cities. The data have been gathered through a *web-crawler* (data have been automatically scraped from publicly available rent ads in the web), therefore be aware of possible errors or inconsistency in the data (outliers, duplicates, missing values, etc.).
The following 13 different features have been collected.

- **city:** City where the property is located
- **area:** Property area
- **rooms:** Number of rooms
- **bathroom:** Number of bathrooms
- **parking spaces:** Number of parking spaces
- **floor:** Floor number
- **animal:** Accept or does not accept animals
- **furniture:** Furnished or not furnished
- **hoa (R$):** Monthly *Homeowners Association Tax* (tassa condominiale), in *Real*
- **rent amount (R$):** Monthly requested rent amount, in *Real*
- **property tax (R$):** Yearly property taxes, in *Real*
- **fire insurance (R$):** Monthly fire insurance cost, in *Real*

These data have been collected in order to better understand the house-rent market in some of the most important cities in Brazil. A new company wants to enter the real-estate market, and wants to understand what kind of houses grant the larger (rent) revenue before investing its money: what are the driving forces leading to high rents?
Furthermore, we may want to segment the rent-houses market in different groups: does it check with the geographical positioning?

1. Build a predictive model and find out the rent amount according to the house specifics
2. Cluster the houses for rental according to their characteristics.

Garment Workers

This dataset includes important attributes of the garment manufacturing process and the productivity of the employees which had been collected manually and also been validated by the industry experts. Data have been collected across different days along the year, and each row (instance) contains different characteristics of a specific *worker team* devoted at performing a specific task.
The following 14 different features have been collected.

- **date:** Date of the recording
- **quarter:** Portion of the month (month was divided into 4 quarters)
- **department:** Department to which the team belong to
- **day:** Day of the week
- **team:** Associated team number
- **targeted_productivity:** Targeted productivity set by the Authority for the team at that specific day.
- **smv:** Standard Minute Value, it is the allocated time for a task
- **wip:** Work in progress. Includes the number of unfinished items for products
- **over_time:** amount of overtime by each team in minutes
- **idle_time:** The amount of time when the production was interrupted due to several reasons
- **idle_men:** The number of workers who were idle due to production interruption
- **no_of_workers:** *Number of workers in each team*
- **actual_productivity:** The actual % of productivity that was delivered by the workers. It ranges from 0-1.

The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories.
In particular, they would like to understand what really impacts the productivity of a team. From a practical point of view, they also know that a productivity score larger than 0.8 is good enough, while a productivity score lower than 0.8 is not.

1. Build a predictive model for the actual productivity of different teams in different days. The response in is (0, 1). What can we do to make the regression setting feasible?
2. Build a classification model to understand what teams and in what days generally have a good enough performance.

Telecom Churn

This dataset contains information about US Telecom customers. Each row represents a customer. The dataset has the following variables on the columns:

- state: code of the US state of customer residence
- account length: number of months the customer has been with the current telco provider
- area code
- international plan: the customer has an international plan (Yes=1)
- voice mail plan: the customer has a voice-mail plan (Yes=1)
- number vmail messages: number of voice-mail messages
- total day minutes: total minutes of calls during the day
- total day calls: total number of calls during the day
- total day charge: total charge of day charge
- total eve minutes: total minutes of calls during the evening
- total eve calls: total number of calls during the evening
- total eve charge: total charge of evening charge
- total night minutes: total minutes of calls during the night
- total night calls: total number of calls during the night
- total night charge: total charge of night charge
- total intl minutes: total minutes of international calls
- total intl calls: total number of international calls
- total intl charge: total charge of international charge
- number customer service calls: number of calls to customer service
- churn: customer changed telco provider (Yes=1)

Telecom wants to understand what customers are more likely to churn (change provider), according to the available covariates so that it can target these customers with an ad-hoc promotional campaign.

1. Build a classification model to help Telecom in targeting the *willing-to-churn* customers before it is too late
2. Cluster customers according to their behavior

Wine Quality

This dataset contains information about the white variant of the Portuguese *Vinho Verde* wine. Due to privacy and logistic issues, only physicochemical variables and a sensory variable (i.e. quality) are available (e.g. there is no data about grape types, wine brand, wine selling price,etc.). Wine samples are recorded on the rows. The dataset has the following features:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality: score between 0 and 10

What makes a wine taste good? We own a wine-producing company and we want to select our vineyards in order to get the best wine ever. If we get a wine with some specifics, what is going to be its quality?

1. Build a regression model to predict the quality of the wine according to its specifics. Treat the outcome as continuous (even if it is not), but keep in mind that the score is between 0 and 10…
2. Bin the quality in *bad wines (score<6)* and *good wines (score>=6).* What makes a wine taste bad (or good?)

Korea Real Estate

This dataset contains information about apartment transaction data generated from August 2007 to August 2017 in Daebong district, Daegu city, South Korea. Each row refers to an apartment. The dataset has the following variables:

- SalePrice: price in US dollar
- YearBuilt
- YearSold
- MonthSold
- Size(sqf): size of apartment in square feet
- Floor: the floor where the apartment is located
- HallwayType
- HeatingType
- AptManageType: how the apartment was managed
- N_Parkinglot: count number of parking spaces on the ground
- TimeToBusStop: measures time takes from apartment to bus stop
- TimeToSubway: measures time takes from apartment to subway station
- N_APT: number of apartments building in an apartment complex
- N_manager: number of people manage apartment facilities (e.g. security,cleaner,etc.)
- N_elevators: total number of elevators in an apartment complex
- SubwayStation: name of subway station nearby apartment
- N_FacilitiesNearBy(PublicOffices): number of public offices nearby apartment
- N_FacilitiesNearBy(Hospitals): number of hospitals nearby apartment
- N_FacilitiesNearBy(DepStores): number of department stores nearby apartment
- N_FacilitiesNearBy(Malls): number of malls nearby apartment
- N_FacilitiesNearBy(ETC): number of buildings like hotels or special schools nearby apartment
- N_FacilitiesNearBy(Park): number of parks nearby apartment
- N_SchoolNearBy(Elementary): number of elementary schools nearby apartment
- N_SchoolNearBy(Middle): number of middle schools nearby apartment
- N_SchoolNearBy(High): number of high schools nearby apartment
- N_SchoolNearBy(University): number of universities nearby apartment
- N_FacilitiesInApt: number of facilities for residents like swimming pool, gym,play ground
- N_FacilitiesNearBy(Total): total number of facilities nearby apartment
- N_SchoolNearBy(Total): total number of schools nearby apartment

A new company wants to enter the real-estate market in Korea, and wants to understand what kind of houses, and in what areas (facilities, etc.) grant the larger sale prices.

1. Build a predictive model for the sale prices of houses in Korea according to the available covariates

Car Prices

This dataset contains information about automobiles. Each row represents a different car model. For each observation we have information about the price and the car features, as well as an insurance risk score.

- Car_ID: Unique id of each observation (Interger)
- Symboling: Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
- carCompany: Name of car company (Categorical)
- fueltype: Car fuel type i.e gas or diesel (Categorical)
- aspiration: Aspiration used in a car (Categorical)
- doornumber: Number of doors in a car (Categorical)
- carbody: body of car (Categorical)
- drivewheel: type of drive wheel (Categorical)
- enginelocation: Location of car engine (Categorical)
- wheelbase: Weelbase of car (Numeric)
- carlength: Length of car (Numeric)
- carwidth: Width of car (Numeric)
- carheight: height of car (Numeric)
- curbweight: The weight of a car without occupants or baggage. (Numeric)
- enginetype: Type of engine. (Categorical)
- cylindernumber: cylinder placed in the car (Categorical)
- enginesize: Size of car (Numeric)
- fuelsystem: Fuel system of car (Categorical)
- boreratio: Boreratio of car (Numeric)
- stroke: Stroke or volume inside the engine (Numeric)
- compressionratio: compression ratio of car (Numeric)
- horsepower: Horsepower (Numeric)
- peakrpm: car peak rpm (Numeric)
- citympg: Mileage in city (Numeric)
- highwaympg: Mileage on highway (Numeric)
- price: Price of car (Numeric)

I car producing company has to come up with a new model of car. It wants to target a specific segment of the market, so it needs a model able to predict the price of the new-designed car according to its specifics. Furthermore, it is well known that car market is strongly segmented in different types (van, suv, coupet, etc.). What cluster will the new car belong?

1. Build a predictive model for the sale prices of cars according to their characteristics
2. Can you get a clusterization of car-type just looking at the available covariates (excluding model and carbody)?

Affairs

This dataset contains information about extramarital affairs. For each subject the following features were recorded:

- rate_marriage: How rate marriage, 1 = very poor, 2 = poor, 3 = fair, 4 = good, 5 = very good
- age: Age
- yrs_married: No. years married. Interval approximations.
- children: No. children
- religious: How religious, 1 = not, 2 = mildly, 3 = fairly, 4 = strongly
- educ: Level of education, 9 = grade school, 12 = high school, 14 = some college, 16= college graduate, 17 = some graduate school, 20 = advanced degree
- occupation: 1 = student, 2 = farming, agriculture; semi-skilled, or unskilled worker; 3 = white-colloar; 4 = teacher, counselor social worker, nurse; artist, writers; technician, skilled worker, 5 = managerial, administrative, business, 6 = professional with advanced degree
- occupation_husb: Husband's occupation. Same as occupation.
- affairs: measure of time spent in extramarital affairs

Some possible questions of interest (not exhaustive: these are just some ideas): Is the amount of time spent in extramarital affairs affected by the other covariates? Does the presence of children (or other features) tend to affect in some way the probability of being in a successful marriage?