Data Exploration

This version of the exercises is for students who already feel comfortable with python, pandas, and data concepts. If not, follow the more guided version.

1 Fundamentals

1. Loading the data
   (a) Download the dataset `cars_synthtic.csv`found on the class website.
   (b) Using pandas, load the file and check its content
2. Data cleaning
   (a) One column is in the wrong format. Find it and fix it.
   (b) Some columns have a few missing values. Remove the corresponding lines. Be careful, one column has many missing values, do not remove those values
   (c) A column has aberrant values. Find them and remove them
3. Data Exploration
   (a) Explore some of the variables using relevant plots. You can use a library such as `AutoViz` or `pandas_profiling`
   (b) To really understand your data, you will however often have to spend time designing your own plots. In this example, use plotly's `px.scatter` function to design a plot in which: x is the `year`, y is the `price`, the symbol shape depends on the `type`, the symbol color corresponds to car's `color` and the symbol size corresponds to the car's `weight`. Try to check if you see some patterns in it. For instance, does it seem that the color or the type has an influence on the price?
4. Distributions
   (a) Explore the distribution of the length variable, and the length of SUV and normal cars. Using a `shapiro` test, find which one do not follow a normal distribution.
5. Dispersion, Correlation
   (a) For the following questions, we will focus on the numerical variables only (length,weight,width,price,year)
   (b) Recompute manually the correlation coefficient between those variables from the covariance matrix.
   (c) For each pair of variables, check if they are independent, linearly correlated, or non-linearly correlated.

2 Advanced

6. (a) On the class page, you can find a dataset corresponding to real data about used cars, for one brand. Download it (you can also find the reference to the original dataset, containing other brands, if you prefer).
   (b) Apply a similar analysis on this real data.