Outlier Detection

Problem 1)

Suppose you want to buy an antique car, because you're a famous collector. You have a list with many characteristics of each car. A car that **stands out** would be a good idea, but a car that "stands out" can be very good or very bad. So which car to buy?

Input Data: mtcars

Description:

The dataset that we are going to use in this case study, called mtcars, was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

A data frame with 32 observations on 11 variables.

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

As a collector, you are only interested in three characteristics: mpg, qseq and hp. So a filter need to be done in the dataset. Could you identify the “outstanding” cars.

```
> cars
                   mpg  qsec  hp
Mazda RX4         21.0 16.46 110
Mazda RX4 Wag     21.0 17.02 110
Datsun 710        22.8 18.61  93
Hornet 4 Drive    21.4 19.44 110
Hornet Sportabout 18.7 17.02 175
Valiant           18.1 20.22 105
Duster 360        14.3 15.84 245
Merc 240D         24.4 20.00  62
Merc 230          22.8 22.90  95
Merc 280          19.2 18.30 123
Merc 280C         17.8 18.90 123
Merc 450SE        16.4 17.40 180
Merc 450SL        17.3 17.60 180
Merc 450SLC       15.2 18.00 180
Cadillac Fleetwood 10.4 17.98 205
```

Figure 1: The dataset after removing the irrelevant attributes

Task: Find out which cars that **stand out** (outliers) and would be interesting to collectors.

Problem 2)

In the dataset “bankloan.csv”, data scientists want to use the following features to predict whether the company obligator is default or not.

Data Attributes

x1 age of company (in years)

x2 ownership (1: public, 2: cooperative, 3: government)

x3 type (1: company, 2: institute)

x4 stock type (1: public joint stock, 2: private joint stock, 3: limited LTD, . . .)

x5 number of managers

x6 managers’ average age (in years)

x7 managers’ total stock

x8 the ratio of asset to capital

x9 the ratio of collateral to loan

x10 activity background in this segment (0: without any background, 1: with background)

x11 the remaining months for borrowers to withdraw the obligations

x12 collateral code

x13 duration between default so far (in years)

x14 duration between the last payment so far (in years)

x15 payment after due date (0: no, 1: yes)

x16 default in past years (0: nondefault, 1: default)

x17 the ratio of past debt to previous credit value

x18 the ratio of default to previous credit value

x19 class label (0: nondefault, 1: default)

Tasks:

1. Use Boxplots to visualize the univariate outliers for variable x1, x5, x6, x7, x11, x13, and x14 respectively. (**7 boxplots**)
2. Find out and **list** the top 10 multivariate outliers. (using the Gower distance is an option).