1. **Monitoring of CO2 emissions from passenger cars – Regulation 443/2009**

European Environment Agency

The Regulation (EC) No 443/2009 requires Member States to record information for each new passenger car registered in its territory. Every year, each Member State shall submit to the Commission all the information related to their new registrations. In particular, the following details are required for each new passenger car registered: manufacturer name, type approval number, type, variant, version, make and commercial name, specific emissions of CO2, mass of the vehicle, wheel base, track width, engine capacity, fuel type and fuel mode. Additional information, such as engine power, were also submitted.

http://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-8

TRANSPORT – CAR – ENVIRONMENT – REGULATION

2. Speed Dating

Department of Statistics – Columbia University

Speed dating data with over 8,000 observations of matches and non-matches, with answers to survey questions about how people rate themselves and how they rate others on several dimensions. This is a large and rich dataset which might take you some time to fully understand. It should be fun to play with.

http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

FUN – DATING – HUMAN

3. Bike Sharing

Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data. The dataset contains 17389 instances and 16 attributes.

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

BIKE – TRANSPORT – LOGISTIC

4. Loans

Lending Club Corporation

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

Another dataset contains the list and details of all loan applications that did not meet Lending Club's credit underwriting policy. You have to Sign in to download the full version of the files.

https://www.lendingclub.com/info/download-data.action

FINANCE – BANK – BUSINESS

5. OpenFlights – Airport, Airline & Route Database

OpenFlights

The OpenFlights Airports Database contains 6977 airports spanning the globe.

The OpenFlights Airlines Database contains 5888 airlines.

The OpenFlights/Airline Route Mapper Route Database contains 59036 routes between 3209 airports on 531 airlines spanning the globe.

http://openflights.org/data.html

TRANSPORT – AIRLINE – AIRPORT – ROUTE

6. The Insurance Company Benchmark – KDD Cup

Information and Computer Science, University of California, Irvine

This data set contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was collected to answer the following question: Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

http://kdd.ics.uci.edu/databases/tic/tic.html

INSURANCE – BUSINESS – MARKETING

7. Mailing campaign for NPO – KDD Cup 1998

Information and Computer Science, University of California, Irvine

The dataset consists in a regression problem where the goal is to estimate the return from a direct mailing in order to maximize donation profits.

http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html

NON-PROFIT – CAMPAIGN – MARKETING

8. Customer relationship prediction – KDD Cup 2009

KDD Cup / Orange

This dataset offers the opportunity to work on large marketing databases from the French Telecom company Orange to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). Both training and test sets contain 50,000 examples. For the large dataset, the first 14,740 variables are numerical and the last 260 are categorical. For the small dataset, the first 190 variables are numerical and the last 40 are categorical.

http://kdd.org/kdd-cup/view/kdd-cup-2009/Data

CHURN – CUSTOMER RELATIONSHIP – MARKETING

9. Fuel prices

ETALAB, data.gouv.fr

The dataset consists in daily prices for gas stations in France from 2007 to 2014. It contains information such as the address, geographical information, working hours, prices, services provided and permanent or temporary closure if it is the case. It also contains historical information to allow comparisons.

https://www.data.gouv.fr/en/datasets/prix-des-carburants-en-france/

FUEL – GEOGRAPHICAL – ENERGY

10. Medical expense refunds (Medicam)

ETALAB, data.gouv.fr

The Medic'AM dataset reports the medical expenses refunds by the French health insurance. For each medicament, the dataset provides its name, its category, the refunded basis, the number of refunded medicaments, the refunded amount and the prescribers basis. The dataset contains data from 2008 to 2013.

https://www.data.gouv.fr/en/datasets/medicaments-rembourses-par-lassurance-maladie/

HEALTH – INSURANCE – FRAUD

11. Establishment Specific Injury & Illness Data (OSHA Data Initiative)

United State Departement of Labor

The Occupational Safety and Health Administration (OSHA) collected work-related injury and illness data from employers within specific industry and employment size specifications from 1996 through 2011. The data provided is used by OSHA to calculate establishment specific injury and illness incidence rates. This searchable database contains a table with the name, address, industry, and associated Total Case Rate (TCR), Days Away, Restricted & Transfer (DART) case rate, and the Days Away From Work (DAFWII) case rate for the establishments that provided OSHA with valid data.

https://www.osha.gov/pls/odi/establishment_search.html

INJURY – MEDICAL – LABOUR

Disclaimer: those datasets are provided by third parties. They can be made unavailable by those parties without notice. ESSEC is not liable in any case for any problem occurred because of any type of usage of those datasets.