11 July 2024

Machine learning for audit

Dr Stamatis KALOGIROU
Senior Data Scientist
HoT for Data Science and Analytics

EUROPEAN COURT OF AUDITORS

Luxembourg EU Summer School | 9-11 July 2024

Outline

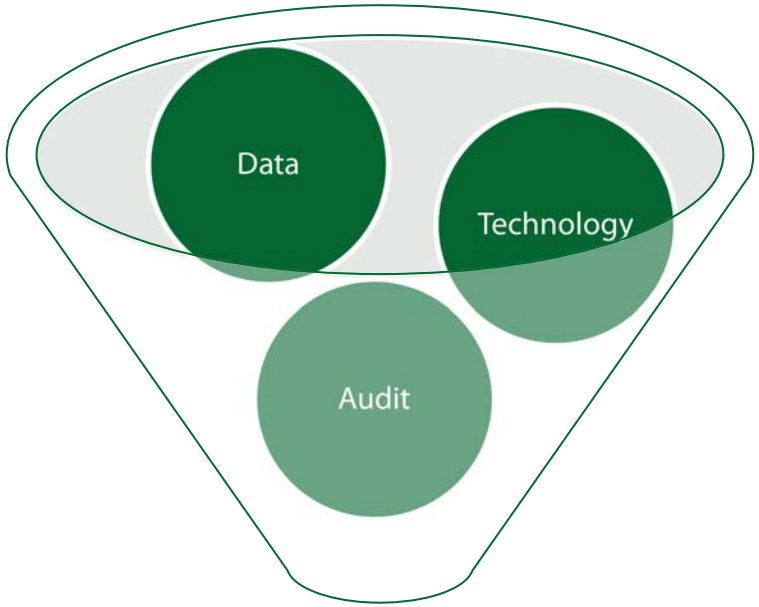
Who: the DATA Team

What: Data Science and Analytics in public audit

How:

- Machine Learning
- Artificial Intelligence

DATA = **DA**ta and **T**echnology for **A**udit



DATA Team

Created in 2021, the team is part of a 5-year Development Plan aiming to meet the demand of an increased use of data and technology in audit.

The DATA team aims to push the digitalisation of the audit work of the ECA through (among others):

- Developing Data Science/Analytics and IT audit internal expertise and services
- Scaling existing data activities
- Coordinating data related actions throughout the organisation
- Internally upskilling auditors to these technologies
- Observing the technological advancement and relative standards, evaluating, and embracing them as well as raising awareness about them.

Audit training offered by DATA

Data science path

- D1 - Introduction to Data Science (8h)
- D2 - Advanced Data Science and Analytics (8h)
- D3 - Exploring Machine Learning (8h)
- D4 - Data Science project (self-paced)

IT audit path

- I1 – Introduction to IT Audit (4h)
- I2 – IT Audit: from theory to practice (12h – 3 mornings)
- I3 – Data reliability assessments (2h)

AI path

- AI1 – Introduction to Generative AI (4h)
- AI2 - Natural Language Processing (8h)
- AI3 - The Regulatory Framework governing the use of AI (2h)
- AI4 - Auditing an AI/ML system (2h)
  AI5 - Geospatial Artificial Intelligence (4h)
- AI6 - Risks of AI in cybersecurity and privacy (2h)

Data collection, data automation, and data visualization training

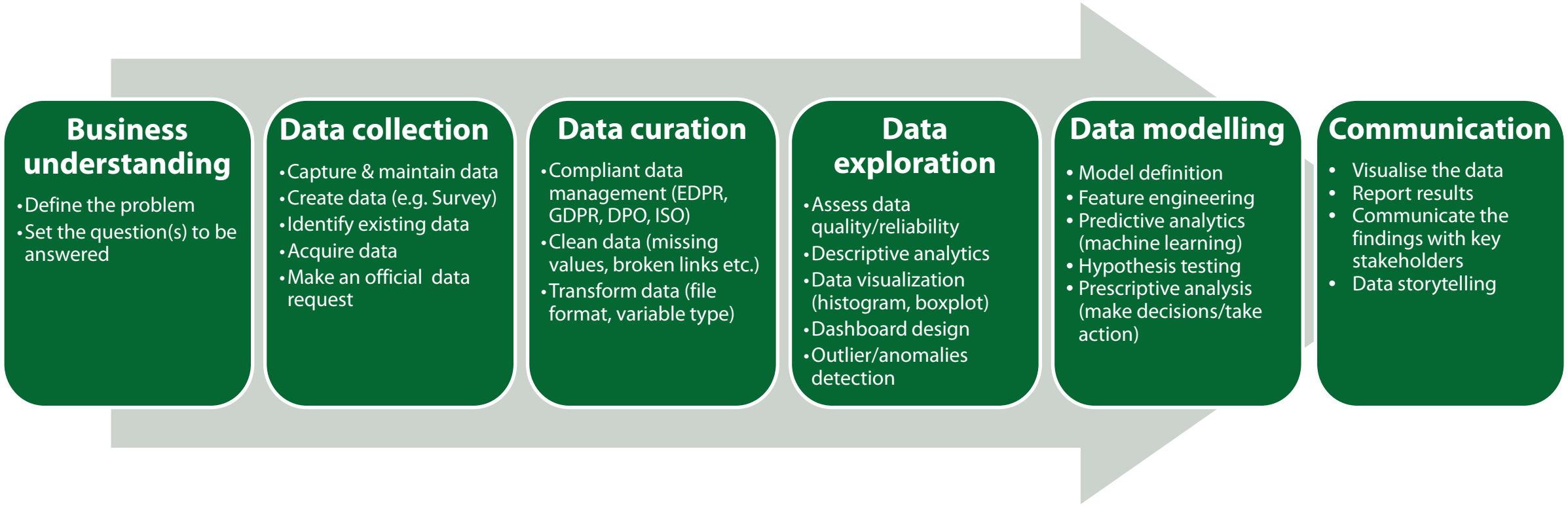
- S1 – Using survey for ECA (performance) audits – Training (6h)
- A1 – Automation 101 – Presentation (2h)
- D5 Data - visualization with PowerBI - Workshop (4h)

Data Science and Analytics in public audit

Why Data Science is important for auditors?

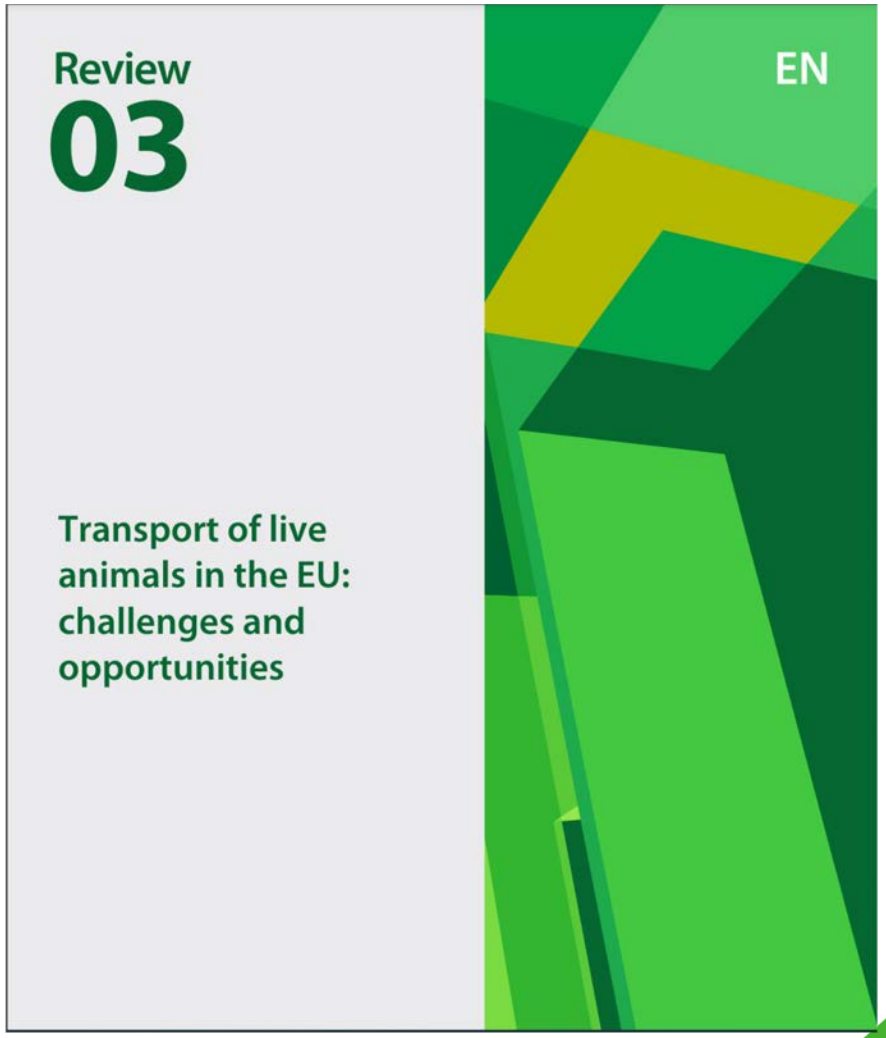
- Main driver of using data analytics: **improve audit quality**
- More **effectively audit** by analysing large amounts of data
- **Better understanding of** the auditee's information and **better identification of the risks**.
- Data analytics tools can **transform the data into insights** that are **understandable** to both auditors and auditees
- Possibility to tailor analysis to auditee-specific risks or to provide data directly into computerised audit procedures (automation)
- **Auditors can arrive at the result more efficiently**.
- *Source: ACCA Global

Data science service offering at the ECA



Animal Transport (Review)

- The audit team requested the support of the DATA team to
  - a) Identify data sources/variables based on their relevance to the audit task.
  - b) Provide support to understand:
    - the nature of the data, and
    - assess if they are suitable to answer the audit questions (fit for purpose).
  - c) Provide support in the communication with the data providers during the request of data and act as a point of contract for technical matters.



d) Download publicly available data and metadata and securely store to use them as audit evidence (Comext | Eurostat).

e) Acquire privately owned / confidential data and inform the DPO for EDPR/GDPR compliance (TRACES | DG Sante).

f) Actively participate in the meetings with the data provides to discuss the nature of data and know quality issues.

g) Clean, curate and harmonise the data; identify and treat outliers; transform as necessary to allow analytics.

h) Assess the quality of data and check for known data issues with data providers

i) Provide clean, processed data to the audit team for their own data analysis.

j) Create geospatial data and analytics based on addresses.

- This required a creation of an internal copy of the *open street map* that is then reused of other audit tasks.

k) Provide Descriptive Data Analytics based on a set of questions with several dimensions (more efficient).

l) Provide Data Visualisation for Data Exploration.

m) Discuss and provide data visualisation for publication (interactive dashboards). – *in production*

Lesson Learned: a clear business understanding and a well-defined request for data science services led to a more efficient and effective use of resources.

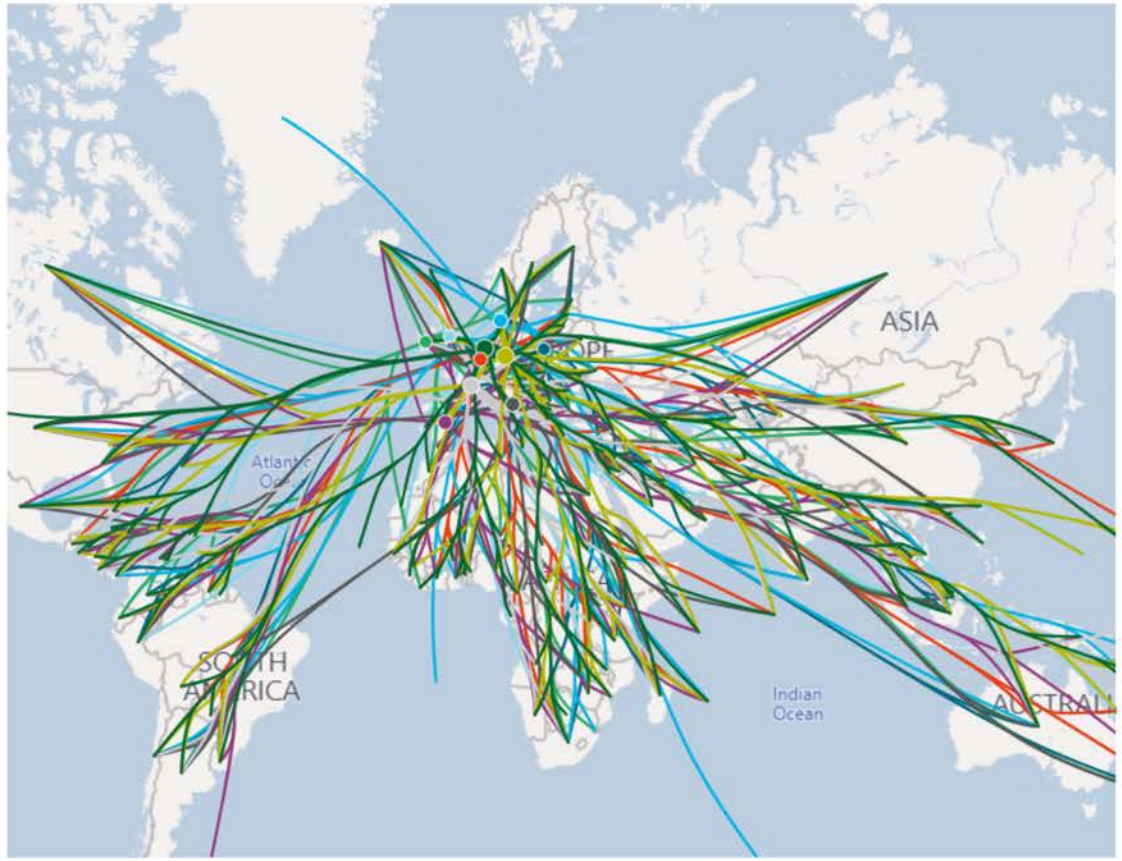
Published dashboard

Declarant Country	Partner Country	Year	Trade Type	Flow	Label EN - CN4	Label EN - CN8
All	All	All	All	All	All	All

Partner Country	Value (€)	Quantity (KG)	Quantity (#)
Germany	14,140,716,232	8,816,867,439	3,168,338,345
Netherlands	12,167,980,297	7,639,044,589	2,676,624,638
France	10,094,430,457	3,733,426,424	817,058,305
Italy	8,479,915,861	2,887,005,654	156,688,965
Denmark	5,936,298,682	2,710,880,276	447,771,054
Belgium	5,766,508,805	3,225,592,415	1,127,904,945
United Kingdom	5,012,481,142	425,872,156	293,398,129
Total	103,647,360,992	43,373,560,917	14,443,663,668

Label CN4	Value (€)	Quantity (KG)	Quantity (#)
Live bovine animals	31,529,595,527	12,022,799,589	47,695,762
Live swine	26,653,237,308	15,047,284,214	336,947,490
Live poultry, fowls of the species Gallus domesticus, ducks, geese, turkeys and guinea fowls	18,470,713,390	12,778,638,400	13,965,739,647
Live horses, asses, mules and hinnies	11,776,389,321	769,306,634	3,952,072
Total	103,647,360,992	43,373,560,917	14,443,663,668

Declarant Country	Value (€)	Quantity (KG)	Quantity (#)
Netherlands	16,206,102,231	8,897,386,276	3,836,817,963
Germany	15,232,398,145	8,163,306,798	2,523,196,436
France	12,643,345,191	3,933,830,110	779,345,455
Italy	7,877,063,393	2,709,778,460	145,983,470
Spain	7,635,621,084	2,524,735,020	444,427,928
Denmark	6,726,872,226	2,667,698,479	424,675,699
Total	103,647,360,992	43,373,560,917	14,443,663,668



BACK TO START

Machine Learning

What is machine learning and where is it used?

- **Machine learning** (**ML**) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions (Wikipedia),

  or simply is the

- **"field of study that gives computers the ability to learn without being explicitly programmed"** (Arthur Samuel, 1959)

- At the ECA, we use machine learning methods during the data exploration (anomaly detection) and data modelling (regression).

Wikipedia contributors. (2023, November 30). Machine learning. In *Wikipedia, The Free Encyclopedia*. Retrieved 14:03, November 30, 2023, from https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1187635450

Main ML algorithm categories

Supervised Learning

- models are trained on "labelled data", i.e. the output to a specific input is already known (classification & regression).
- The algorithm is trained on multiple correct responses (y_i) on inputs (x_i).
- Then the algorithm is expected to make a good prediction of an unknown y' based on an input (x)

- *Example*: Estimation of CO_2 emissions based on car mass and engine size.

Unsupervised learning

- models are trained on data without labelled responses with the goal of discovering hidden patterns or intrinsic structures within the data.

- *Example:* a clustering algorithm that groups beneficiaries of EU funds based on the amount received, the legal entity size and other factors.

Supervised Learning

– Regression

Regression

- Regression refers to the process of studying the cause-and-effect relationship between a dependent variable and a set of independent variables.
- There is a part of machine learning that focuses on this task, which refers to explanatory machine learning models.
- In the supervised learning discussed here, we can consider regression as a mapping of one or more input variables x, to an output variable y.
- To better train a machine learning algorithm, we wish to use as big data as possible.
- This is because in predictive machine learning, the objective is to increase the accuracy in the prediction of the output in each input.
- For example, a real estate agent would appreciate a hedonic price model that could predict the market value of a property with a good accuracy.
- Thus, at the business environment, the focus is on predictive analytics.

- Assuming that y is the output (dependent variable), and x_i is each input (independent variable), a model can be expressed as a function of a vector X of x_i s

$$y = f(X)$$

- At a simple form, this could be a multiple linear regression

$$y = a_0 + \Sigma\beta x_i + \varepsilon$$

- However, most of the time the algorithms are more complex, such as random forests, a tree-based algorithm with bootstrapping.

Unsupervised Learning

- Clustering
- Anomaly Detection

Unsupervised learning

- Only the input data x is available (not the output y)
- The algorithm must find structure in the data

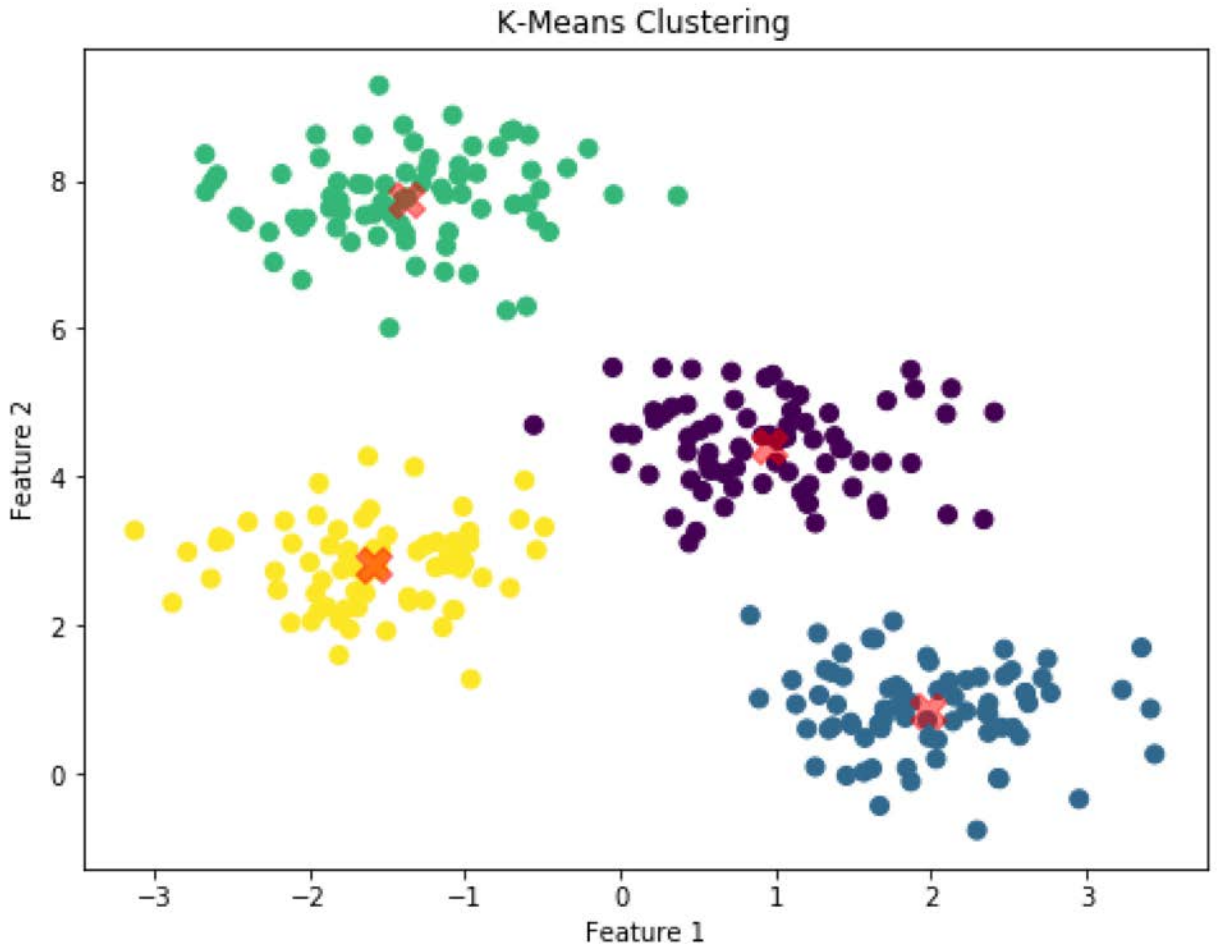
- We focus on two methods
  - Clustering
    - Grouping of entities with similar characteristics into “homogeneous” classes
  - Anomaly detection
    - Identifying unusual data points

Clustering

- Clustering refers to the process of classifying data points in groups (categories) based on one or more variables with the attempt to make a "homogeneous" groupings as possible.
- For example, an area typology of EU Regions on risk of poverty based on several characteristics can be the result of a clustering exercise.
- In this case, there is no known output for input data, i.e. we do not have information to which group each data point belongs to. If we did, then it would have been a classification exercise (e.g. logistic regression).
- Example applications:
  - DNA Analysis
  - News / Videos grouping
  - Market Segmentation
  - Social network analysis
  - Anomaly detection

K-means clustering algorithm

- It is a popular unsupervised learning algorithm, which groups data into k distinct, non-overlapping subsets (clusters) in which the distance of each data point to the cluster centroid is minimised.
- k is normally a small integer number.
- The data points can have one or more variables.
- How it works:
  - Step 0. The algorithm starts with random values of k centroids (mean values of variables for each subset)
  - Step 1. The algorithm assigns each data point to the nearest cluster centroid.
  - Step 2. It recalculates each cluster's centroid
  - Step 3. Repeat steps 1 and 2 until there are no more changes.



Anomaly Detection

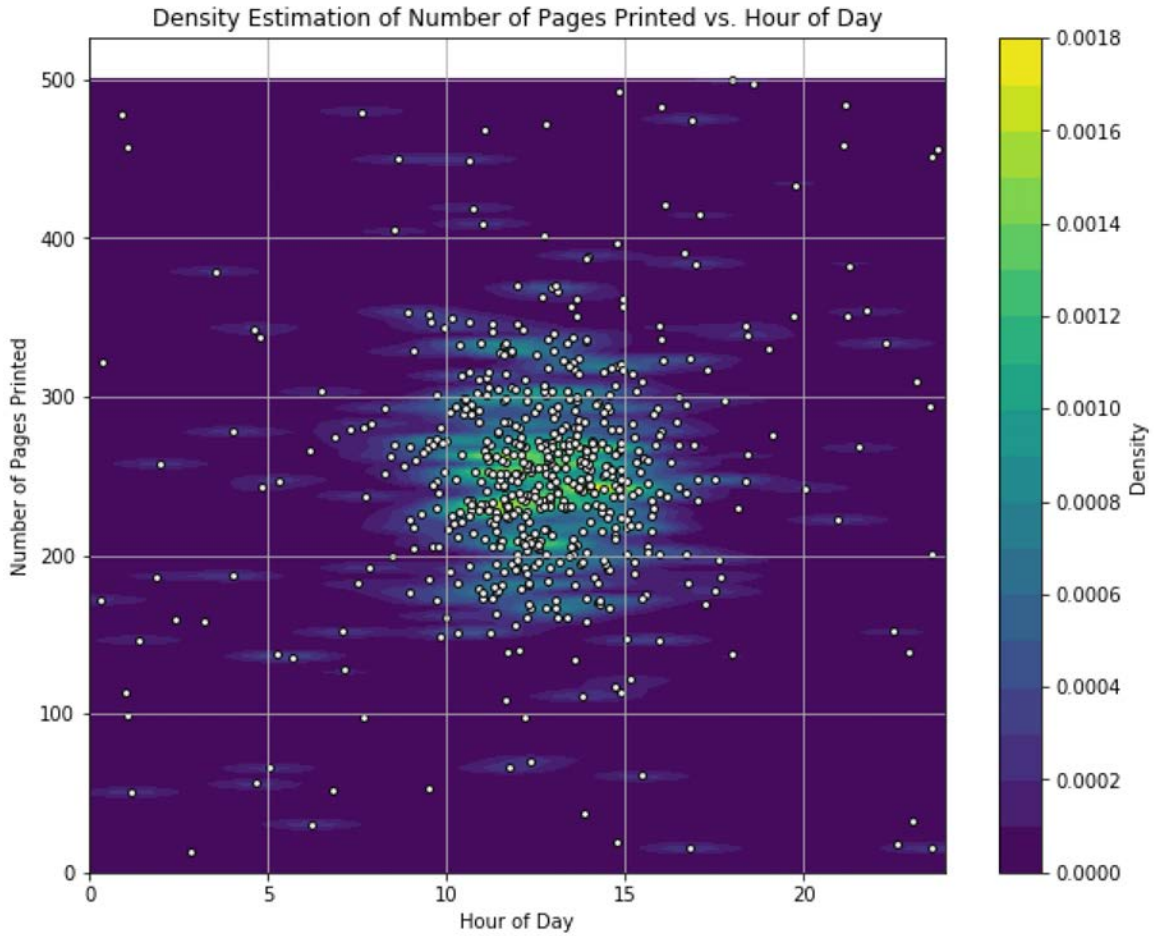
- **Definition**: Anomaly detection is the process of identifying unexpected items or events in data sets, which differ significantly from the norm.
- **Importance**: Anomaly detection is crucial in various fields like **fraud detection, irregularity suspect cases,** network security, fault detection, system health monitoring, and outlier detection.
- Use cases:
  - Fraud Detection in Finance: Identifying unusual transactions that could indicate fraud.
  - Intrusion Detection in Cybersecurity: Detecting unusual patterns that could signify a security breach.
  - Health Monitoring: Identifying abnormal patterns in health data for early disease detection.
  - Industrial Fault Detection: Monitoring equipment and environment to detect early signs of malfunction.
- Anomaly Detection Techniques
  - Statistical Methods: such as Z-score, Grubbs' test.
  - **Machine Learning-Based Methods: supervised, unsupervised, and semi-supervised techniques.**
  - Deep Learning Approaches: neural networks and autoencoders for complex anomaly detection tasks.

Density estimation

1. Density estimation involves estimating the **probability distribution** (or density function) for a dataset. The goal is to understand how the data is distributed across the space it occupies.

2. **Purpose**: By understanding the density distribution, one can identify areas of low probability, which are potential regions for anomalies.

3. A new data point with a probability below a threshold ε is a potentially anomaly.



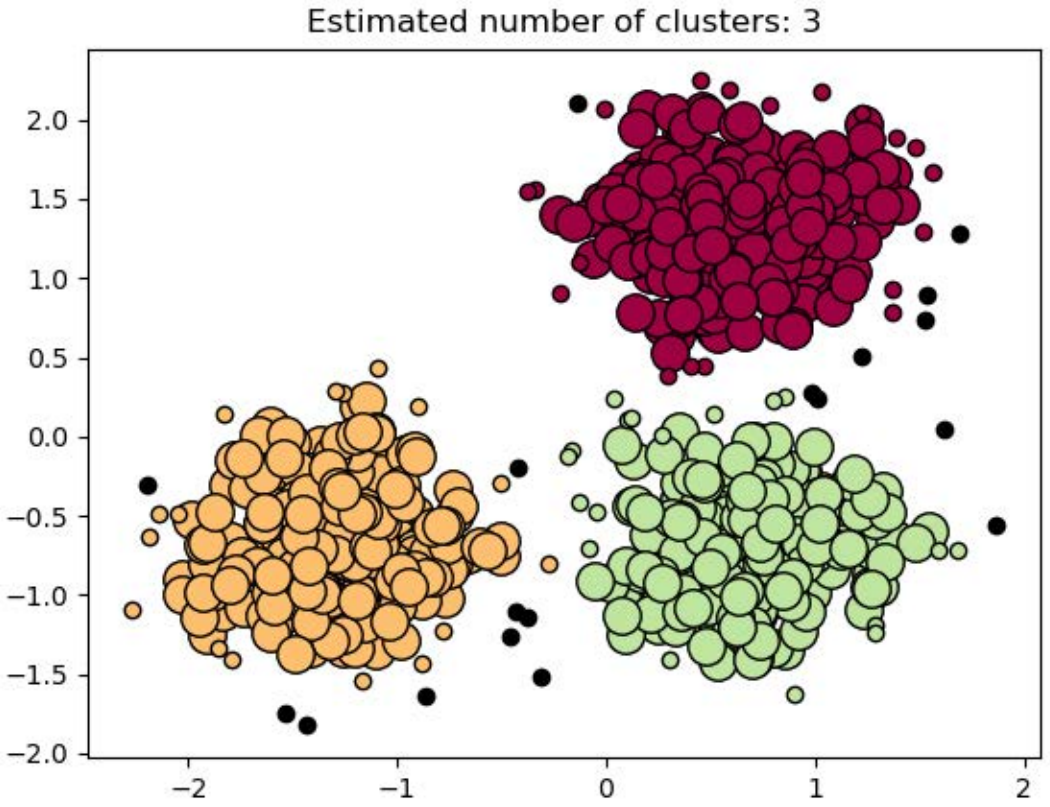
Example using Kernel Density Estimation
Graph is generated using AI (custom ChatGPT).

Density-Based Spatial Clustering Applications with Noise

- *DBSCAN* is a clustering algorithm that defines clusters as continuous regions of high density.
- There are two parameters that must be determined beforehand: the cluster distance denoted by ε and the minimum number of samples per cluster denoted by m.
- The algorithm works in the following way:

1. For each element (data point) in the dataset, the algorithm counts how many other elements are located within a small distance ε from it. This region is often referred to as the element's ε-neighborhood.
2. If an element has at least m minimum samples in its ε-neighborhood (including itself), the it is considered a *core element*, namely it is an element located in a dense region.
3. All elements in the neighborhood of a core element are grouped into the same cluster. This neighborhood may include other core elements, in which case a long sequence of neighboring core elements forms a single cluster.
4. Any element that is neither a core element, nor does it have one in its neighborhood is considered an anomaly.

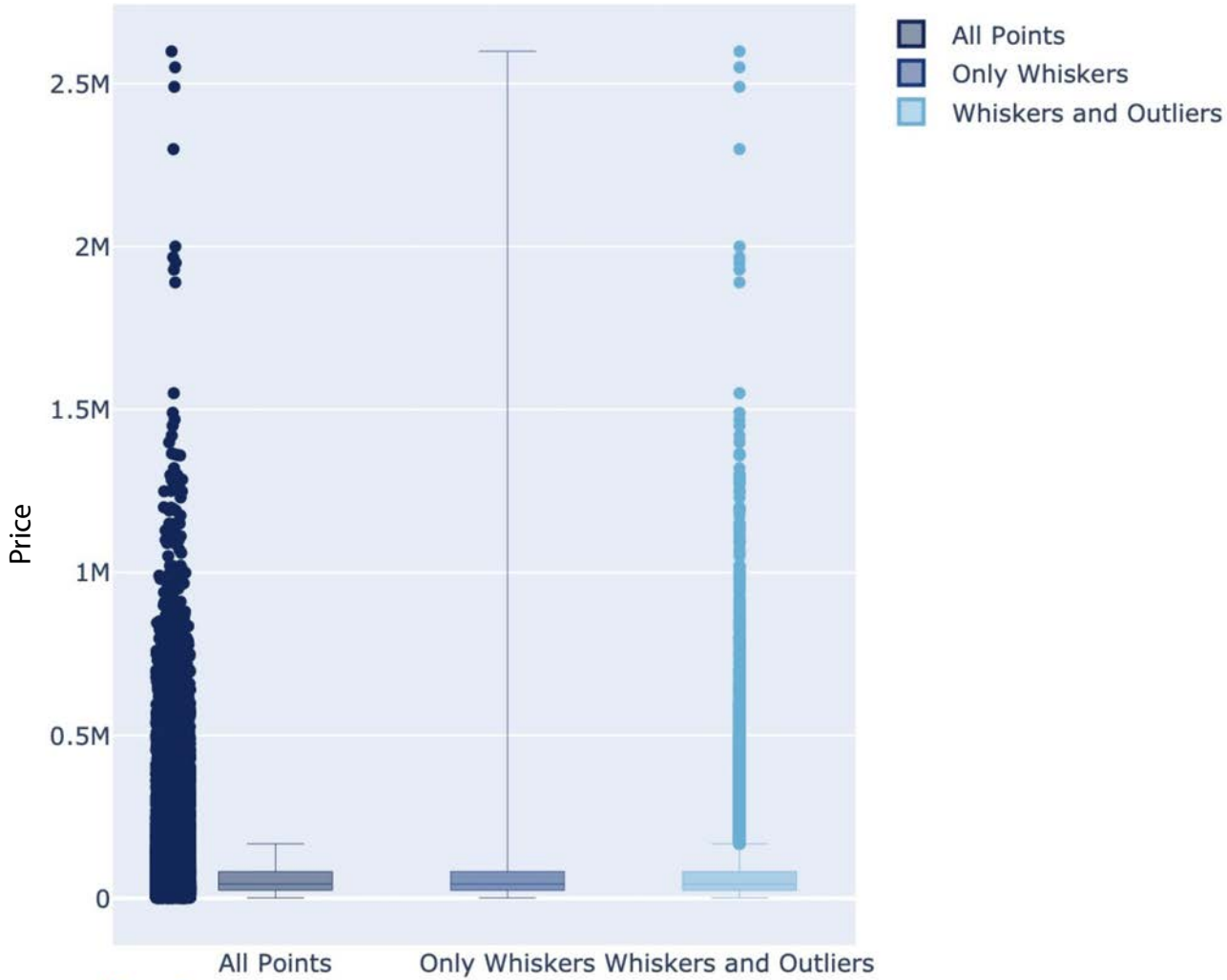
Illustration

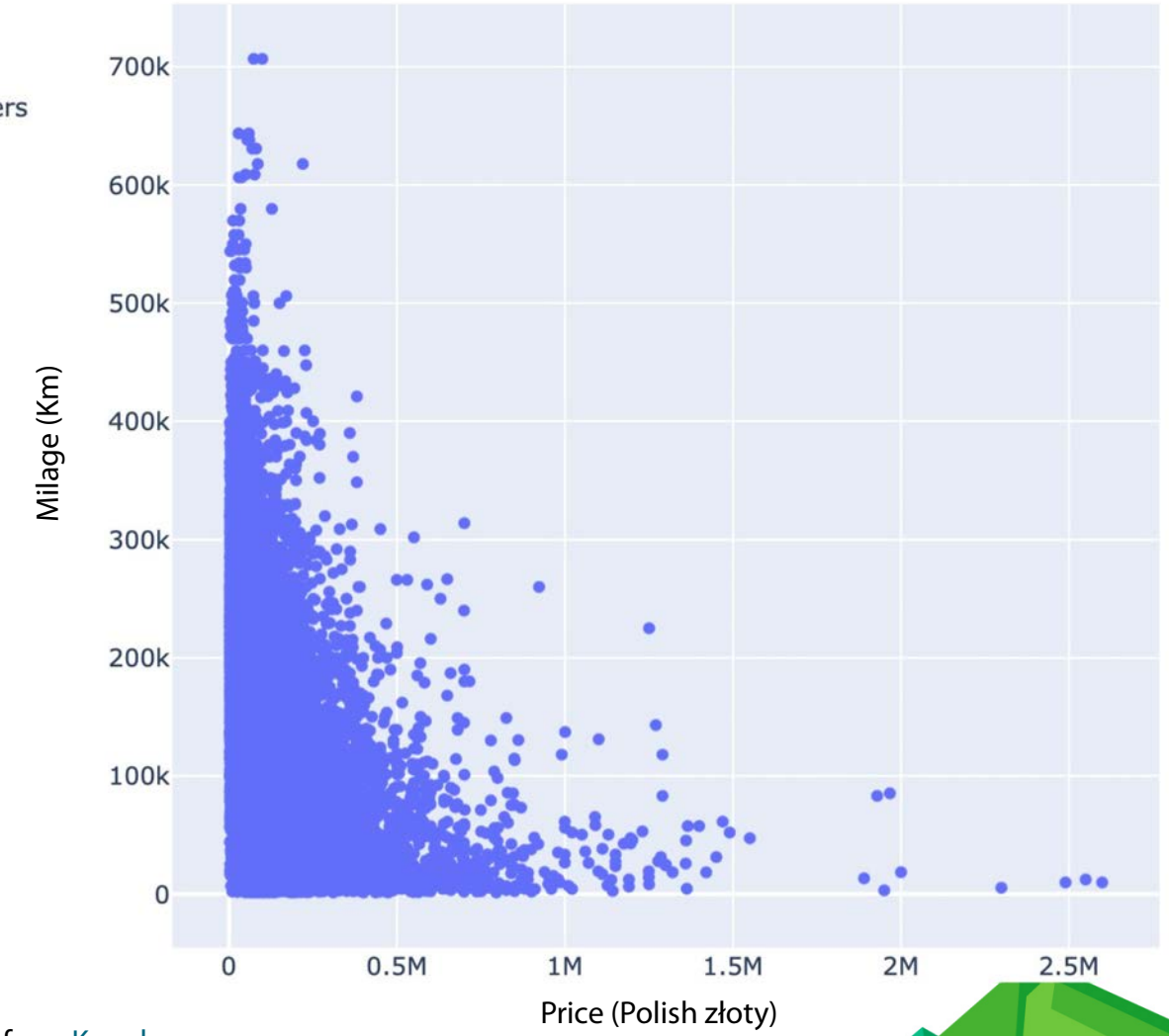


- Core elements (large dots) and non-core elements (small dots) are color-coded according to the assigned cluster.
- Elements tagged as noise (anomalies) are represented in black.

Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py

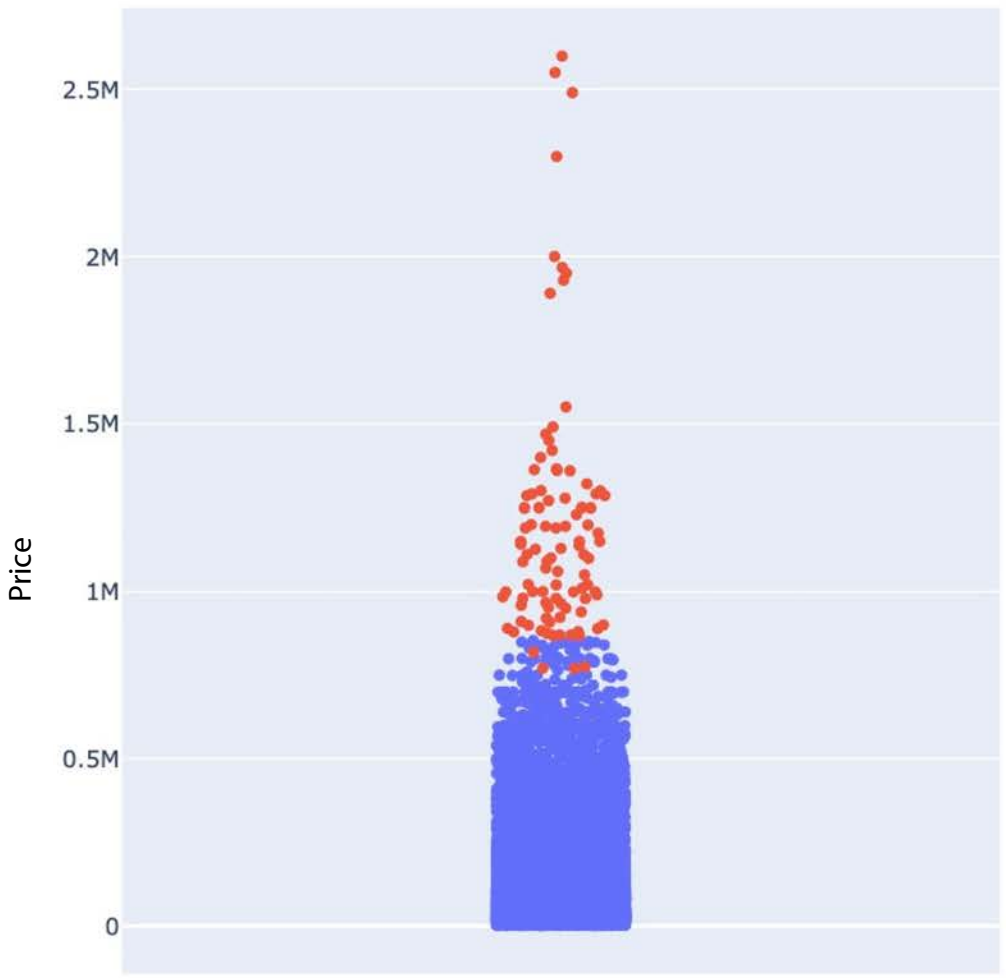
Sample data: Poland Used Cars

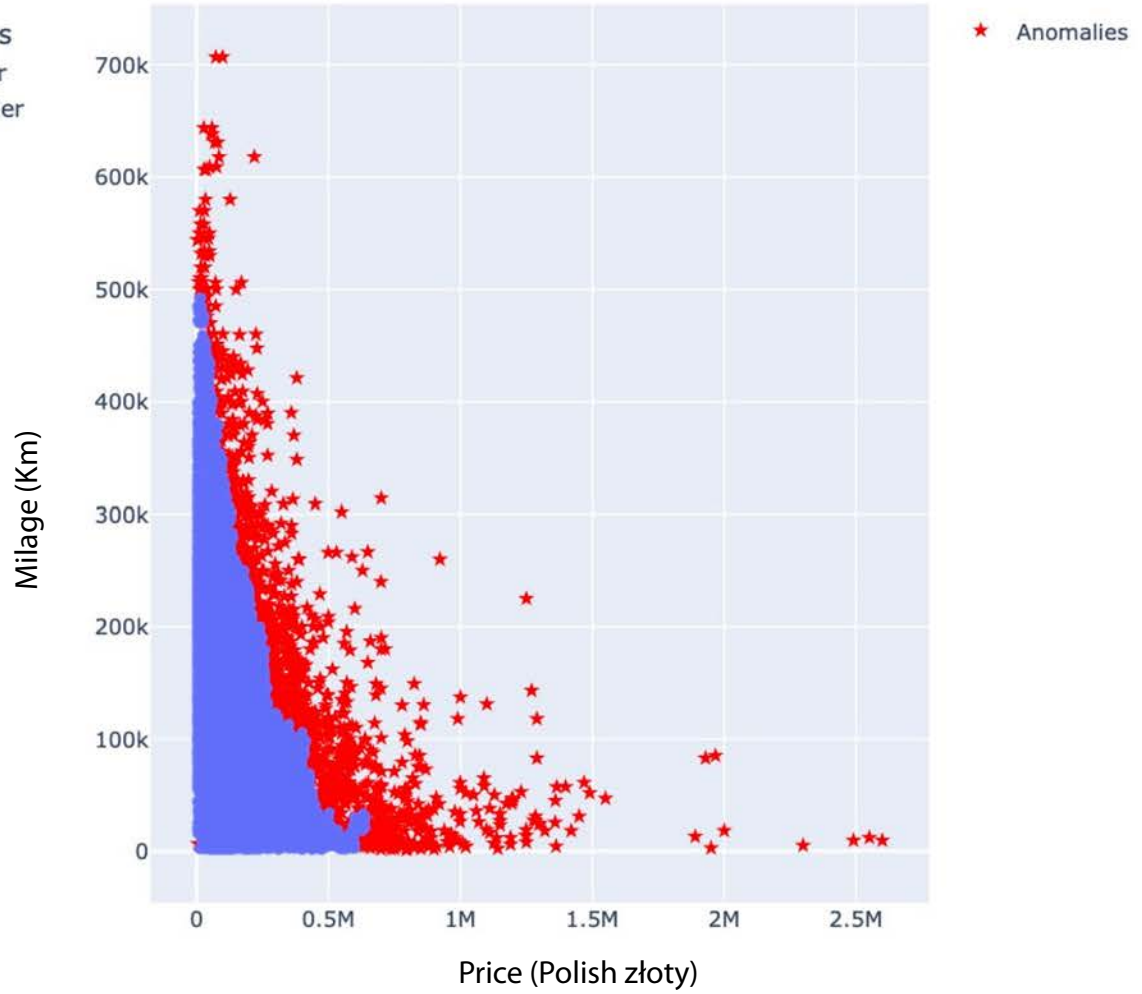




Sample dataset can be downloaded from Kaggle

Density-Based Spatial Clustering Applications with Noise





$\varepsilon = 0.1$

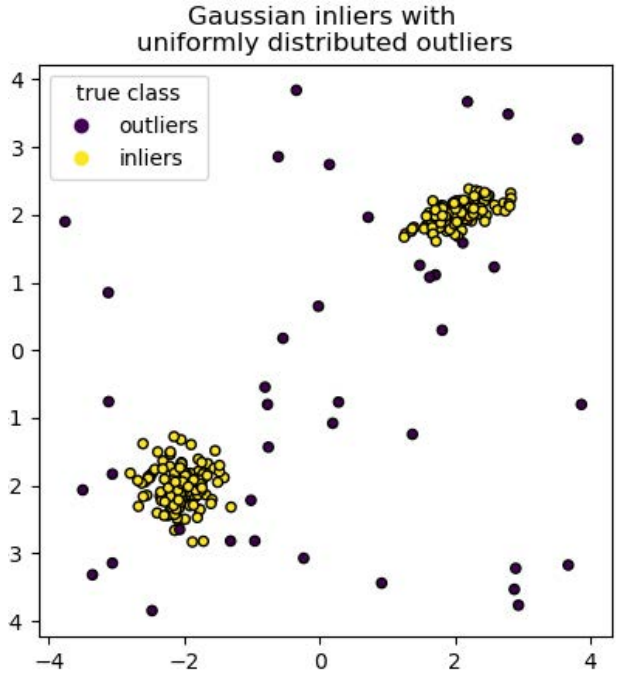
$m = 10$

Isolation Forest

- The *Isolation Forest* Algorithm is an efficient algorithm for outlier detection, especially in high-dimensional datasets. The algorithm builds a *Random Forest* in which each *Isolation Tree* (or *iTree*) is grown randomly:

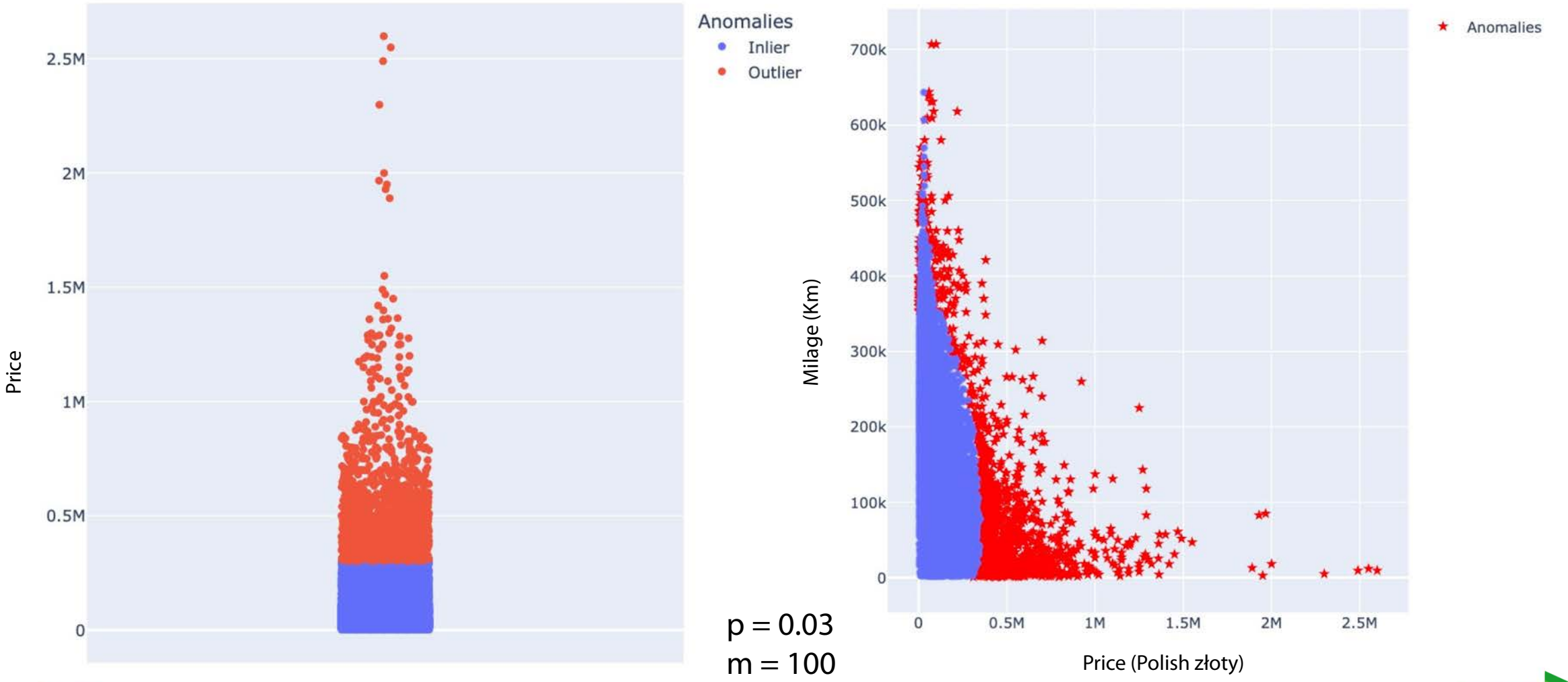
1. At each node, it picks a feature randomly, then it picks a random threshold value (between the min and max values) to split the dataset in two.
2. The dataset gradually gets chopped into pieces this way, until all elements end up isolated from the other elements.

- Anomalies are usually far from other elements, so on average (across all the iTrees) they tend to get isolated in fewer steps than normal elements.
- Two parameters we can adjust to fine tune the model is the percentage of contamination denoted by p (the threshold for classifying an element as an anomaly), as well as the maximum number of samples denoted by m (number of samples to be drawn from the dataset to build each tree).



Source: https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html

Isolation Forest example



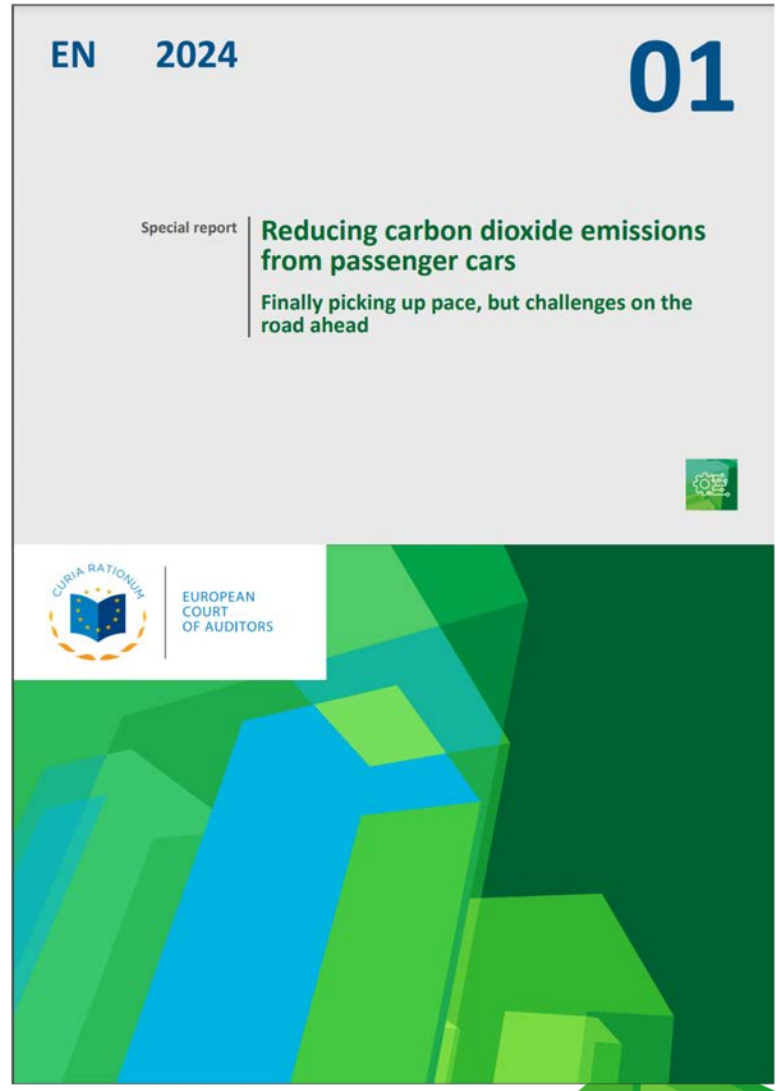
p = 0.03
m = 100

Unsupervised learning in practice

- We have applied clustering and anomaly detection in ongoing audits with interesting findings.
- Applicable in full population data.
- Business rules (e.g. choice of variables) are made by the audit team.
- Technical choices (e.g. size of cluster and distance) are made by the DATA team.

Supervised learning (regression) in practice

- We fit some regression models in the audit looking at CO_2 emissions from passenger cars (Special report 01/2024)
- We examined the relationship between mass, engine size and CO_2 emission based on real fuel consumption data (collected in cars since 2021). We also looked at the CO_2 emission gaps (lab to real) by fuel type, engine size and mass.
- We tested if heavy cars with small engines emit more CO_2 (or have a bigger real-lab gap) than heavy cars with larger engines.
- The analysis showed that there was no strong evidence to make a conclusion.
- More and better-quality real fuel consumption (and thus CO_2 emissions) data is needed to redo this analysis.



EUROPEAN COURT OF AUDITORS

Special report 01/2024: Reducing carbon dioxide emissions from passenger cars
Finally picking up pace, but challenges on the road ahead

Thank you for your attention!

Questions and discussion time.

Contact details

Stamatis Kalogirou

Senior Data Scientist

stamatis.kalogirou [at] eca.europa.eu

linkedin.com/in/stamatis

Want to know more about ECAs work:

Website: **https://www.eca.europa.eu/**

Twitter **@EUAuditors**

LinkedIn: linkedin.com/company/euauditors

Videos on YouTube: **EUAuditorsECA**

EUROPEAN COURT OF AUDITORS