YOUR SCIENCE

CREATING OPPORTUNITY

Applied Analytics
AI & ML for BI & Automation

4. Exploratory Data Analysis
Including ML Techniques

Your Science

Mathematical Consulting

Prof. Norbert Poncin

2025

Your Science Center: 31, Boulevard Prince Henri, L-1724 Luxembourg

Phone: +352 621 674 917, **Email**: info@yourscience.eu, **Website**: yourscience.eu

Business Permit: 10154927, **LBR**: A44409, **VAT**: LU35024328

YOUR SCIENCE

CREATING OPPORTUNITY

Applied Analytics
AI & ML for BI & Automation

4. Exploratory Data Analysis
Including ML Techniques

Your Science

Mathematical Consulting

Prof. Norbert Poncin

2025

Your Science Center: 31, Boulevard Prince Henri, L-1724 Luxembourg

Phone: +352 621 674 917, **Email**: info@yourscience.eu, **Website**: yourscience.eu

Business Permit: 10154927, **LBR**: A44409, **VAT**: LU35024328

Contents

Complements to Exploratory Data Analysis, Including Machine Learning Techniques in Industrial Case Studies

Exploratory Data Analysis (EDA) is a comprehensive process that combines various techniques to uncover the *structure and characteristics* of data. It involves the following key components:

- **Data Cleaning and Preprocessing**: Identifying and handling missing values, outliers, and data transformations.
- **Descriptive Statistics**: Summarizing data with measures of central tendency, dispersion, and the shape of distributions.
- **Exploratory Visualizations**: Using charts, plots, and tables to identify patterns and trends.
- **Univariate, Bivariate, and Multivariate Analysis**:
  - **Univariate Analysis**: Examines individual variables, focusing on distributions and frequencies.
  - **Bivariate Analysis**: Investigates *relationships* between two variables using methods such as correlations and scatter plots.
  - **Multivariate Analysis**: Explores *relationships* among multiple variables using techniques like principal component analysis (PCA).

While earlier chapters addressed most of these aspects, this chapter focuses on core methods of Bivariate Analysis, with Chapter VI including a discussion of Multivariate Analysis.

As the present course emphasizes AI and machine learning (ML), we integrate the ML models Decision Trees (DTs) and Random Forests (RFs) into the EDA process. These models reveal significant *relationships* within the data, thereby making EDA more complete.

1 Correlation and Association

1.1 Automated Correlation Analysis with SMS Notification and Email Report

The following script **automatically** computes and visualizes correlation coefficients and correlation matrices for various customer features. Below, we provide the necessary background on correlation coefficients. Additionally, the code sends an **SMS notification** to stakeholders via the third-party service Twilio, which enables businesses to send and receive SMS, voice messages, and more through web-based APIs. The SMS notifies stakeholders about a **personal email** with the subject 'New CRM Insights Report Available'. This email includes the **CRM Insights Report**, which contains the computed correlation coefficients and their visualization.

Python Code

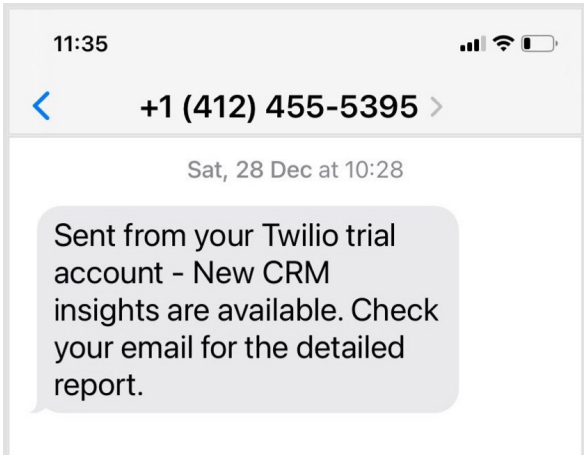
In the following code, please replace:

- `XXXXXXXXX` with your Twilio 'account_sid',
- `YYYYYYYYY` with your Twilio 'auth_token', and
- `ZZZZZZZZZ` with your email account password.

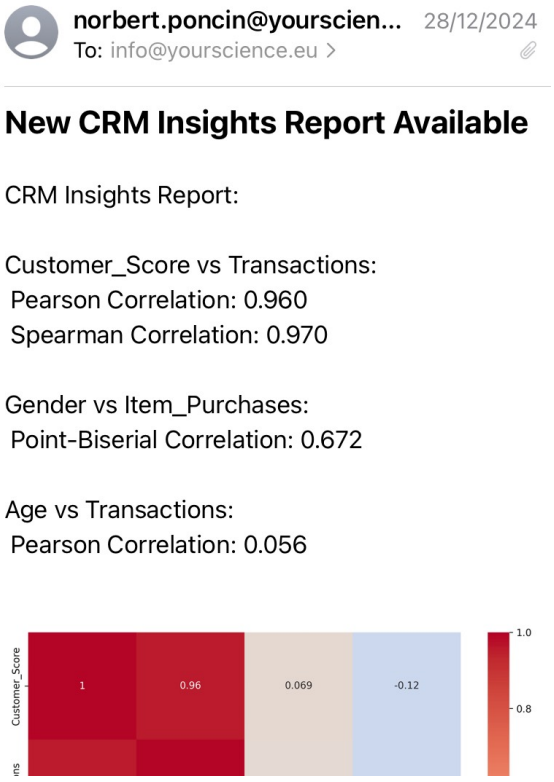
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import pearsonr, spearmanr, pointbiserialr
import os
import smtplib
from email.message import EmailMessage
from twilio.rest import Client

# Define path to Downloads folder
downloads_folder = os.path.expanduser("~/Downloads/")
file_path = os.path.join(downloads_folder, "CRMDataCorr.png")
```

SMS and Email

You receive a text message containing the text specified in the code:



Additionally, you receive an email with the content specified in the code:



Exercise 1. *Work through the following questions:*

- **Correlation Encoding**: *Customer_Score (or Loyalty_Score) is strongly correlated with Transactions amounts. Examine the code and analyze why this (logical) correlation appears in the simulated data. Perform a similar analysis for the relationship between Gender and Item_Purchases.*

- **Poisson Distribution Overview**: *To model the number of Item_Purchases, the code uses a Poisson distribution. Consider asking your AI assistant what characterizes data that follow a Poisson distribution and request examples to deepen your understanding. Describe in your own words how the shape of the Poisson distribution evolves as its parameter increases (focusing on shifts in the peak, tail behavior, and overall symmetry).*

- **SMS and Email Notification Code**: *Are you curious about how the automatic SMS warning and the email containing the report are implemented in Python? Explain in detail, using your own words, what happens in the code.*

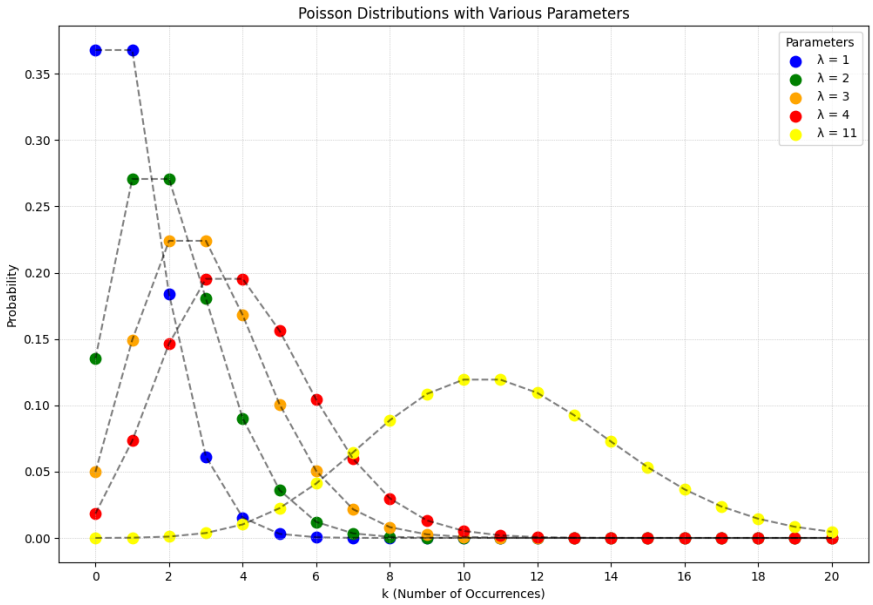


Figure 2: Poisson Distribution

```
    Gender  Requested_Demo Source                 Timestamp
99    Male            True  Email 2024-12-28 15:51:50+01:00

Chi-square test results:
Chi-square statistic: 8.050933441558438
p-value: 0.004548021477274153
Degrees of freedom: 1
```

Gender	No Demo (False)	Demo (True)
Female	15	41
Male	25	19

Table 2: Contingency Table – Demo Requests by Gender

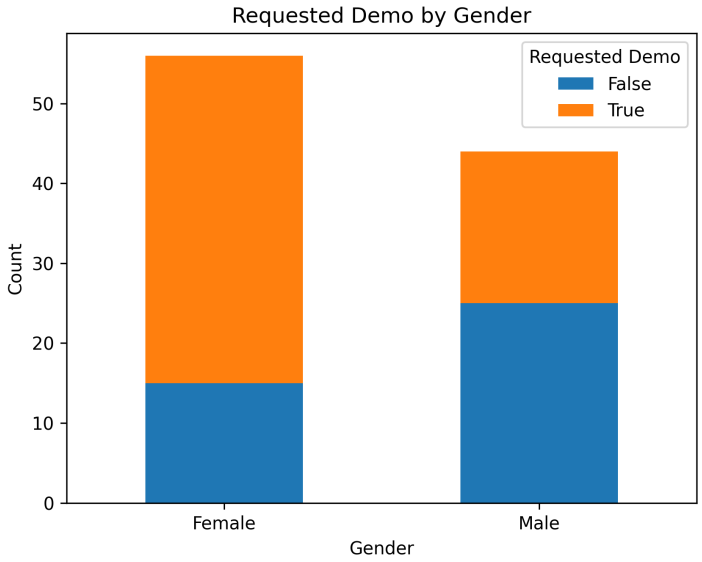


Figure 3: Bar Chart True|False vs. Female|Male

```
There is a statistically significant relationship between gender
and requesting a demo (p < 0.05).
```

Association Measures for Categorical Variables

Association measures are specifically designed to assess relationships between nominal variables, i.e., categorical variables without inherent order.

Cross-Tabulation

Categorical data analysis often uses cross-tabulation to explore relationships between *categorical variables*, such as gender (female, male) and demo request (yes, no). Through cross-tabulation, we generate contingency tables (frequency tables) and stacked bar charts (where 'demo' and 'no-demo' bars are stacked within each gender) to reveal patterns and potential correlations (see Table 2 and Figure 3). This approach typically includes statistical tests (e.g., chi-square test) to assess the significance of observed relationships.

Additional Information about the Test Run

We now explain the content of Code Line 106:

```
chi2, p, dof, expected = chi2_contingency(contingency_table)
```

and the output:

```
Chi-square test results:
Chi-square statistic: 8.050933441558438
p-value: 0.004548021477274153
Degrees of freedom: 1
```

Chi-square: `chi2` (better: χ^2) is the test statistic for Pearson's chi-square test. It measures the discrepancy between the *observed frequencies* O_i *in the contingency table* and the *expected frequencies* E_i *under the assumption of no correlation* between the variables:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \, .$$

Generalizing this formula to n classes $C_1, \ldots, C_n$, we get

$$\text{gini} = 1 - \sum_{i=1}^{n} p_i^2 \,,$$

where p_i is the proportion of C_i .

Other impurity metrics include the **Entropy** and the **Mean Squared Error** (MSE).

- **Samples**: The variable 'samples' in the nodes of the tree indicates the number of emails represented by that node. For instance, the root node contains 700 samples because 70% of the 1000 emails in the dataset are used for training, which the tree represents.
- **Value**: The variable 'value' is a 1D array where the first entry represents the number of `Not Spam` emails and the second entry represents the number of `Spam` emails in the node.
- **Class**: The variable 'class' is determined by the larger of the two entries in the 'value' array, representing the majority class in that node.

Precision, Recall, Accuracy, F1-Score, and AVGs

Precision, Recall, Accuracy, F1-Score, and Averages (AVGs) are performance metrics for evaluating classification models.

You may think about them in the following way:

If you predict whether an email is `Spam` or `Not Spam`, and you classify it as `Spam`, it is referred to as a **Positive**. If the prediction is correct, it's a **True Positive**, if incorrect, it's a **False Positive**. Similarly, if you predict an email is `Not Spam`, it is a **Negative**. If correct, it's a **True Negative**, and if wrong, it's a **False Negative**.

When evaluating the quality of the model, three key questions arise:

- **Precision**: Among all emails predicted as `Spam` (i.e., Positives = True Positives + False Positives), *what proportion is correctly predicted?* This

leads to the formula:

$$\text{Precision } \texttt{Spam} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} .$$

The case `Not Spam` is similar.

- **Recall**: Among all emails that are truly `Spam` (i.e., True Positives + False Negatives), *what proportion is correctly predicted*, i.e., what proportion does the model correctly recall? Formula:

$$\text{Recall } \texttt{Spam} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} .$$

- **Accuracy**: Among all emails (i.e., True Positives + False Positives + True Negatives + False Negatives), *what proportion is correctly predicted*?

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}} .$$

- **Support**: The number of samples in the dataset or subset.

- **Macro Average**: Since high Precision for `Spam` might be offset by low Precision for `Not Spam`, computing an average metric (Average Precision, Average Recall, Average F1-Score) is essential for overall performance assessment:

$$\text{Macro Average Precision} = \frac{\text{Precision for Class 1} + \text{Precision for Class 2}}{2} .$$

- **Weighted Average**: The average metric weighted by the number of samples in each class accounts for class imbalance. If n_1 is the number of samples in Class 1, n_2 the number of samples in Class 2, and $n =$

Cropped first two Trees of the Forest

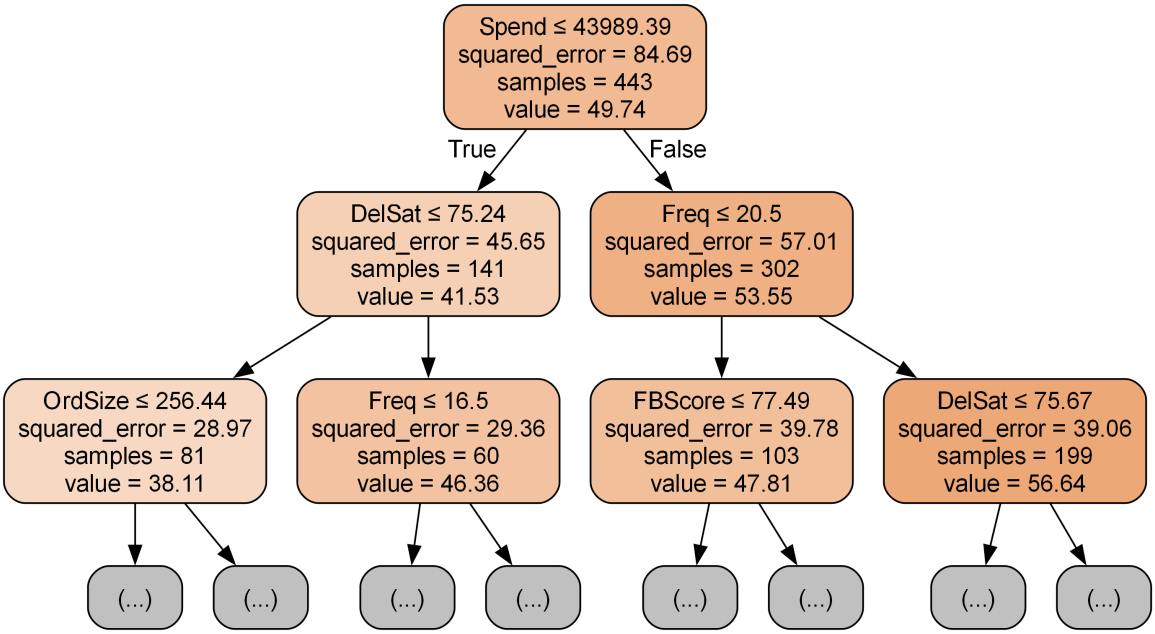


Figure 7: Cropped Tree 1

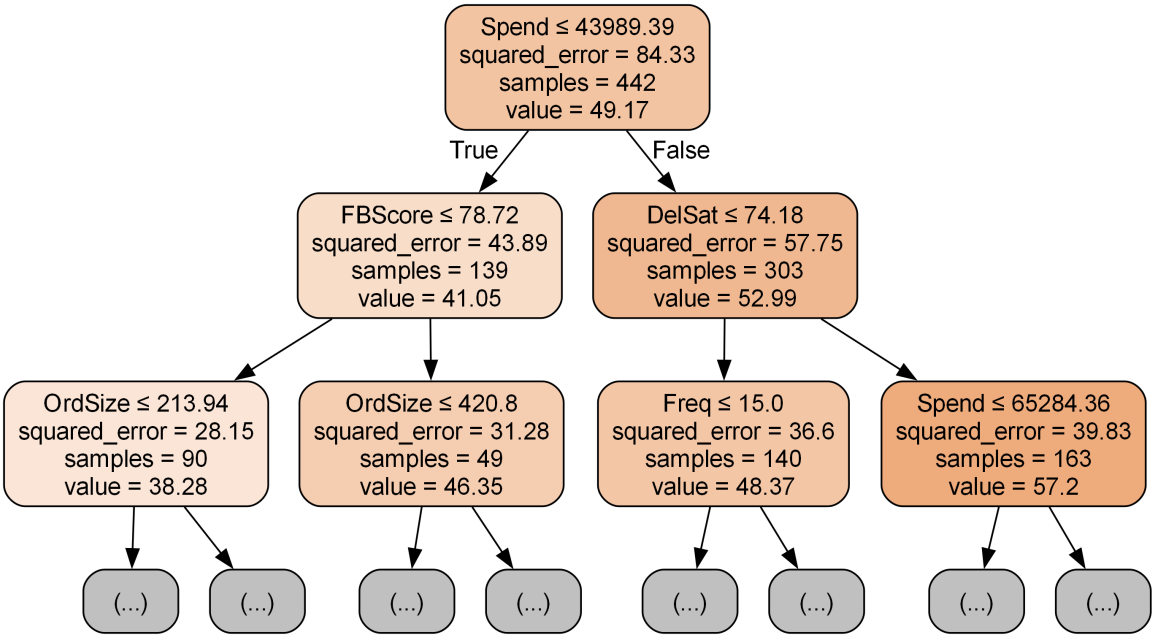


Figure 8: Cropped Tree 2

3 Learning Outcomes

After working through this chapter, the reader should be able to:

- Summarize the key steps involved in Exploratory Data Analysis (EDA).
- Describe the main types of variables, including numeric (continuous and discrete), categorical (nominal and ordinal), ranked, and binary variables.
- Explain the Chi-Squared Test and the concept of p-value.
- Outline the principles of Supervised Machine Learning models, with a focus on the modus operandi of Decision Trees and Random Forests for both Classification and Regression tasks.
- Describe the content of nodes in Decision Trees for Classification (particularly Gini impurity) and in Random Forests for Regression (including bootstrap sampling, random feature selection, multiple threshold candidates, and local MSE). Explain how these models are used to make predictions.
- Describe performance metrics such as Precision, Recall, F1-Score, and Accuracy.
- Explain the performance metrics used for regression, including global MSE and the coefficient of determination R^2.

YOUR SCIENCE

CREATING OPPORTUNITY

About the Author

Norbert Poncin is a Luxembourgish mathematician, who was originally educated as a mathematical analyst and has worked extensively in partial differential equations (PDEs) at the University of Liège. His Master's thesis focused on the propagation of singularities in boundary value problems (BVPs) for dynamic hyperbolic systems. Applying the finite element method (FEM), his subsequent dissertation addressed BVPs for complex elliptic systems of PDEs. For his doctoral thesis, he explored mathematical quantization, while his post-doctoral education at the Polish Academy of Sciences strongly emphasized theoretical physics and its models.

Norbert has served as a Full Professor of Mathematics at the University of Luxembourg for more than 25 years and collaborated with more than 25 foreign professors and post-doctoral scholars. He has organized numerous academic events, notably approximately 10 international research meetings and over 20 research seminars focusing on theories, frameworks, concepts and models in Physics and Engineering. Beyond a substantial publication record in Differential Geometry, Algebraic Topology, and related disciplines, he has contributed roughly 25 papers to the fields of Mathematical Physics and Quantum Theory.

He was the leading instructor for over 20 university courses. Spanning a diverse spectrum of subjects, including mathematical analysis, probability theory, inferential statistics, point and solid dynamics, Lagrangian and Hamiltonian mechanics, mechanics of deformable solids, fluid dynamics, special relativity, quantum physics, geometric methods in mathematical physics, and supersymmetric models, his teaching portfolio underscores his extensive experience in applied aspects of mathematics.

In 2023, Norbert Poncin founded the mathematical consulting agency Your Science, where he currently serves as director. His primary interests include data science and artificial intelligence, along with mathematical modeling and computational science.