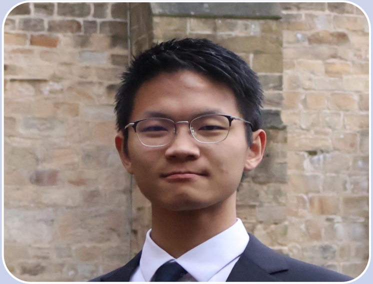
BANKBYTES

Mark Kimutai Kitur
kiturmark@gmail.com
Kenya
Dedan Kimathi university
Of technology
Data Science



Yue hu
hy1550278246@gmail.com
The United Kingdom
Durham University
Data Science

Rakshith Nagaraju
rakshith.nagaraj6@gmail.com
The United Kingdom
University of Liverpool
Data Science

Data Understanding

Pillars of the data

- The Bank dataset columns were divided into four main pillars .
- These pillars made it easier to understand factors that might greatly influence the outcome of the marketing campaign.
- The pillars are:
  - Personal Information
  - Contact Information
  - Campaign Information
  - Economic Indicators

1. Personal Information

- These attributes provide demographic and financial details about the clients which are essential for understanding their profiles and predicting their profiles.

a. Age: Different age groups show varying levels of interest in the product.

b. Job: The type of job indicates client’s financial stability thus influencing their decision.

c. Marital status: This affects financial responsibilities and priorities.

d. Education: Level of education often correlates with income and financial literacy.

e. Loan: Personal loans reflect financial behavior and risk.

f. Default: This attribute shows if a client has credit. This illustrates creditworthiness and financial reliability.

g. Housing: This attribute shows if a client has housing loan which indicates financial commitments and ability to invest.

2. Contact Information

- This characteristics provide details about the method and timing of the last contact with the client, which can influence the client’s response.

a) Contact: This describes means of communication, which their effectiveness greatly vary.

b) Month: The time of the year can affect client responses due to seasonal factors and financial cycles.

c) Day of the week: client availability and mood might differ on different days of the week.

d) Duration: This is duration of call in seconds. Longer call might indicate more interest from the client, but this is known only after the call.

3. Campaign Information

- These attributes track the number of contacts made and the outcomes of previous interactions, providing insights into the client's engagements history.

a) Campaign: Number of contacts during this campaign. Frequent contacts might indicate persistence but could also lead to fatigue.

b) Pdays: Days since the client was last contacted (999 means not contacted before).

c) Previous: This column shows number of contacts before this campaign. Past engagement level could affect the current campaign's success.

d) Poutcome: Outcome of the previous campaign. Previous campaign success can be a strong predictor of the current campaign success.

4. Economic Indicators

- These attributes provide contextual economic data which can affect the overall success of the marketing campaign.

a) Emp.var.rate: This column shows employment variation rate (quarterly indicator). This reflects economic stability and job market conditions.

b) Cons.price.idx: Consumer price index (monthly indicator). This illustrates inflation and cost of living, affecting disposable income.

c) Cons.conf.idx: Consumer confidence index is really important as it reflects consumer optimism or pessimism about the economy, influencing spending behavior.

d) Euribor3m: This column provides information on Euribor 3-month rate (daily indicator). This affects loan interest rates, influencing financial decisions.

e) Nr.employed: Number of employees (quarterly indicator). This reflects labour market size and the economic health.

Lastly we had y (output variable) indicator which show if

Problem Description

Objective

The main objective of this project is to develop a predictive model for the bank to identify potential customers who are likely to buy their term deposit product. The model aims to save resources and time by focusing marketing efforts on customers with a higher likelihood of purchasing the product, thus increasing the efficiency and success rate of telemarketing campaigns.

Specific Problems

(1) Predictive Factors Identification: Identify the key factors that influence a customer's likelihood of purchasing a term deposit. These factors could include demographic information, financial status, previous interactions with the bank, and past purchase behavior.

(2) Model Development: Develop and validate a machine learning model that can accurately predict the probability of a customer buying a term deposit. The model should be able to handle the complexity and variety of data available, ensuring high predictive performance.

(3) Resource Optimization: Utilize the predictive model to prioritize and focus marketing efforts on high-potential customers. This will help the bank save time and resources while improving the overall success rate of their telemarketing campaigns.

Data Resources

To develop a predictive model for identifying potential customers likely to buy the bank's term deposit product, a variety of data types have been utilized. The data is collected by the bank.

The data is expected to be in structured formats, which is in CSV files.

Type of Data

There are two type of data:

a. Categorical data:

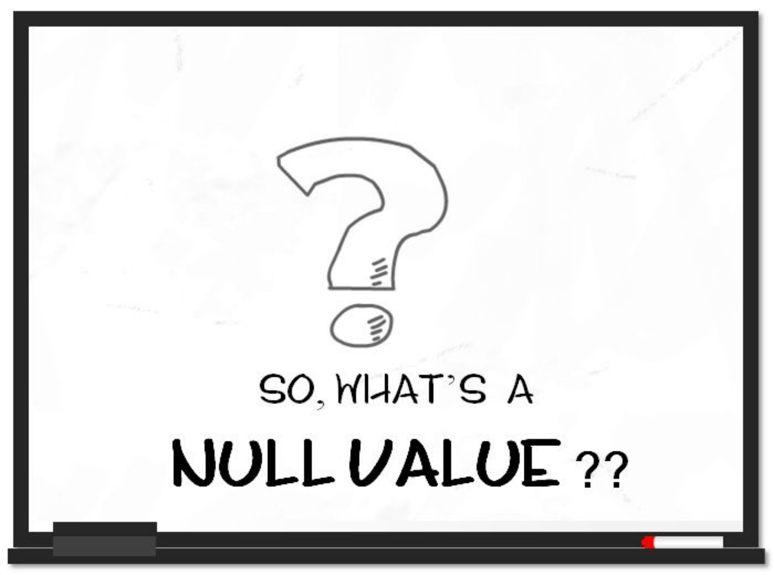
   job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome

b. Numeric data:

   Age, duration, campaign, pdays, previous, emp_var_rate, cons_price_idx, cons_conf_idx, euribor3m, nr_employed

Problems in the data & Proposed Solutions.

Null (NA) values.



Null value is when there is no record of data in a particular instance in the database.

It represents the absence of a value.

The key to good data analysis is having complete data.

NA values in the data?

The dataset being used for this project is the “Bank Marketing” dataset that can be found at this link.

More specifically the “bank-additional” dataset for training and testing.

This data contains **0** NA values in every property (column). This means that the data is complete.

Solution to null values.

Since there are no null values in the entire dataset as mentioned, there is no need to employ methods like mean values or dropping data, etc. in order to achieve complete data.

Therefore there is nothing that needs to be done about null values.

Standardized data.

Standardized data refers to whether the entries in the dataset are consistent in every instance of occurrence or whether it needs to be cleaned.

All categorical data falls within the mentioned categories without any outliers like spelling mistakes, punctuations or capitalizations.

Therefore, on initial observation, this data is standardized and does not require cleaning.

The categories for the data upon initial inspection are as follows:

job: housemaid, services, admin., blue-collar, technician, retired, management, unemployed, self-employed, unknown, entrepreneur, student

marital: married, single, divorced, unknown

education: basic.4y, high.school, basic.6y, basic.9y, professional.course, unknown, university.degree, illiterate

default: no, unknown, yes

housing: no, unknown, yes

loan: no, unknown, yes

contact: telephone, cellular

month: may, jun, jul, aug, oct, nov, dec, mar, apr, sep

day_of_week: mon, tue, wed, thu, fri

poutcome: nonexistent, failure, success

y: no, yes

Need for standardization?

All categorical data in the dataset are confirmed to be within the specified categories as mentioned above.

There are no spelling errors, capitalizations etc that could throw off the model during training.

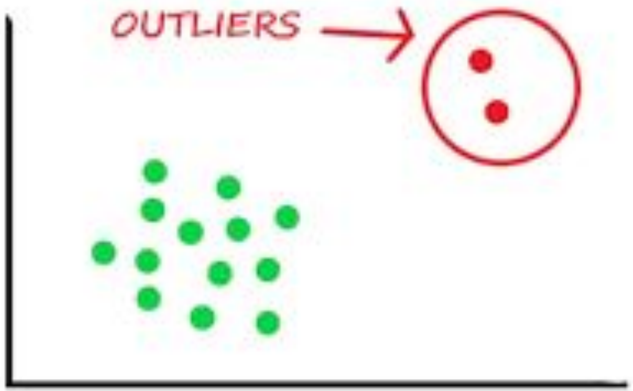
Therefore there is nothing that needs to be done about the standardization of categorical data.

Outliers.

Outlier data refers to data points that significantly deviate from the majority of observations in a dataset.

Outliers can skew the results of statistical analyses, leading to misleading conclusions. For example, a single outlier can drastically affect the mean, making it unrepresentative of the data's central tendency.

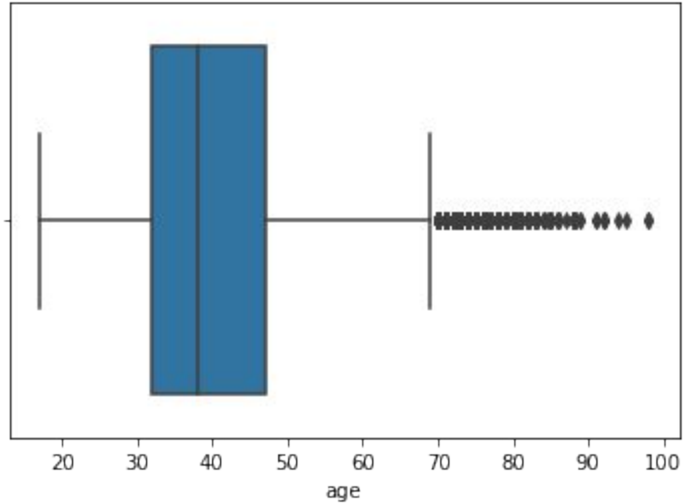
Detecting and managing outliers is crucial, as they can provide valuable insights or indicate problems with data collection.



Outliers in the data

The plots show the how the outliers are scattered in the data:

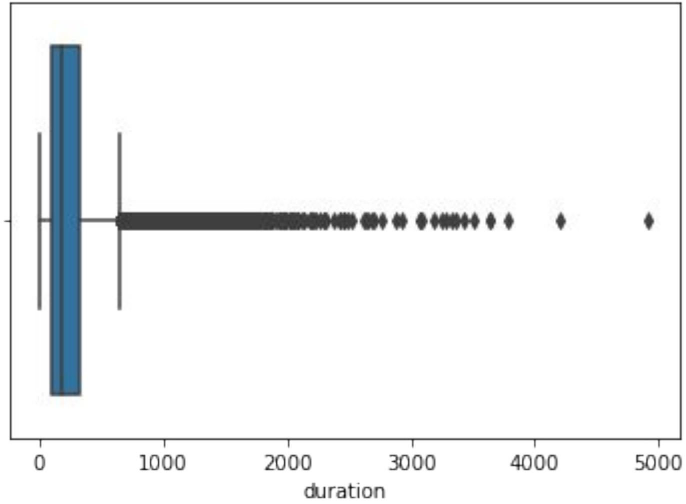
- age
-



Outliers in the data

The plots show the how the outliers are scattered in the data:

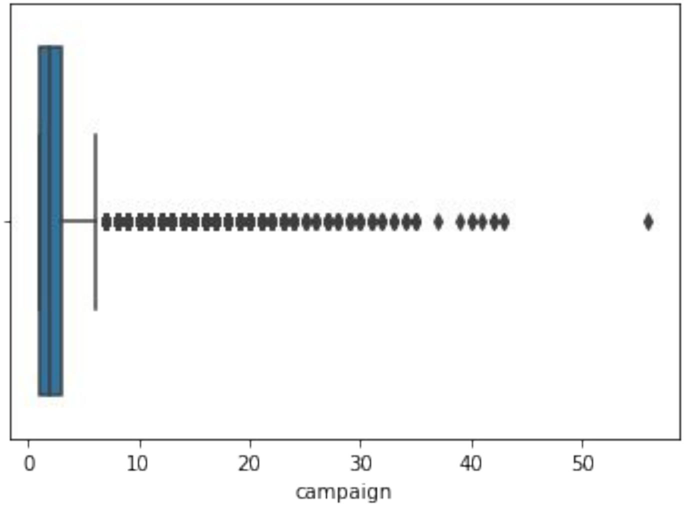
- age
- duration
-



Outliers in the data

The plots show the how the outliers are scattered in the data:

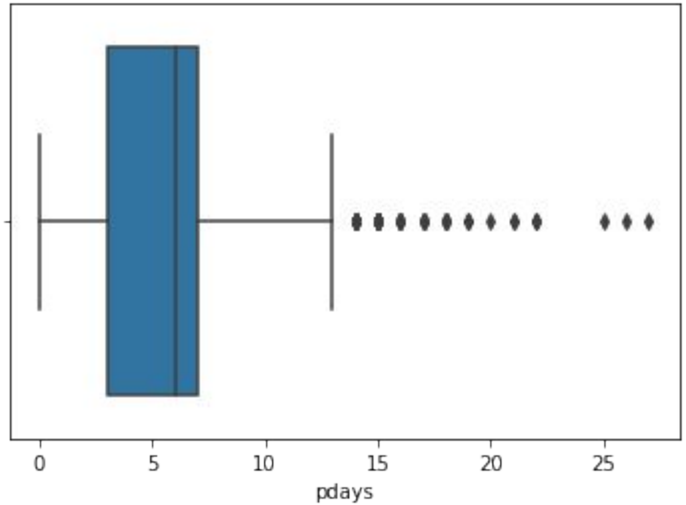
- age
- duration
- campaign
-



Outliers in the data

The plots show the how the outliers are scattered in the data:

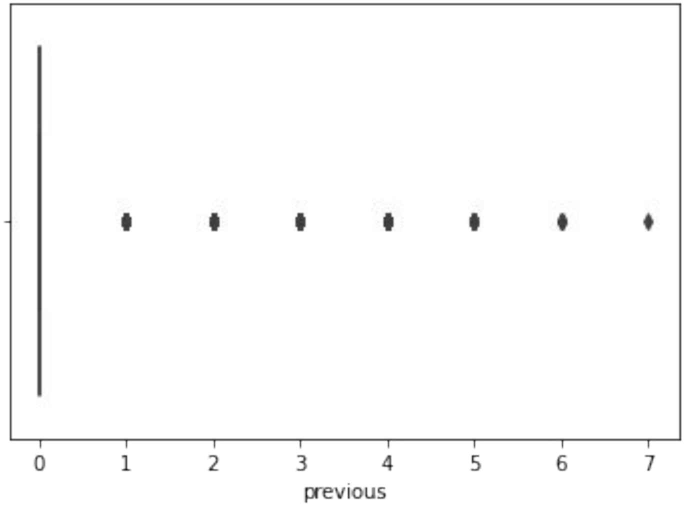
- age
- duration
- campaign
- pdays
-



Outliers in the data

The plots show the how the outliers are scattered in the data:

- age
- duration
- campaign
- pdays
- previous



Handling outliers.

Handling outlier data is essential because outliers can significantly distort statistical analyses and lead to incorrect conclusions. They can skew measures of central tendency like the mean, inflate variance, and affect the results of regression models, making them unreliable.

Additionally, outliers can provide valuable insights, such as identifying errors in data collection or revealing important, atypical phenomena. By addressing outliers, researchers can improve the quality and reliability of their findings.

Without removing any data, we can employ techniques such as:

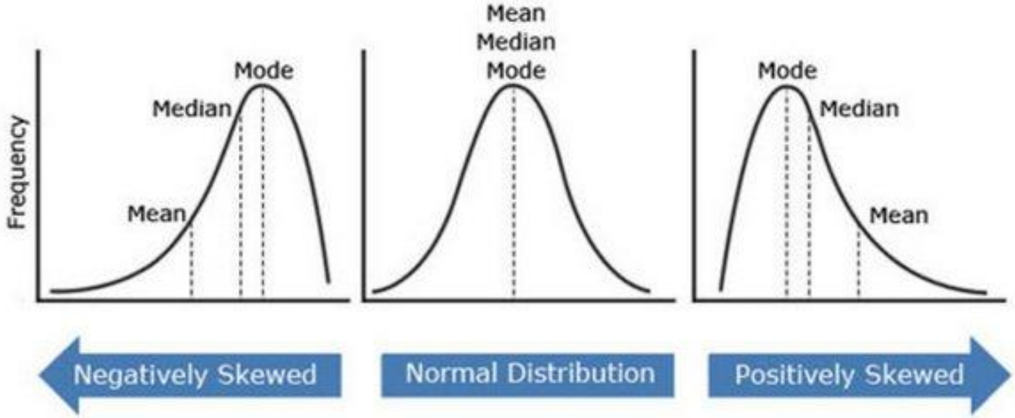
- **Trimming the data.**
- **Capping the data.**

Skewness.

Skewed data occurs when a dataset is not symmetrically distributed around the mean, resulting in one tail being longer or fatter than the other.

In a positively skewed distribution, the right tail is longer, and the mean is greater than the median. In a negatively skewed distribution, the left tail is longer, and the mean is less than the median.

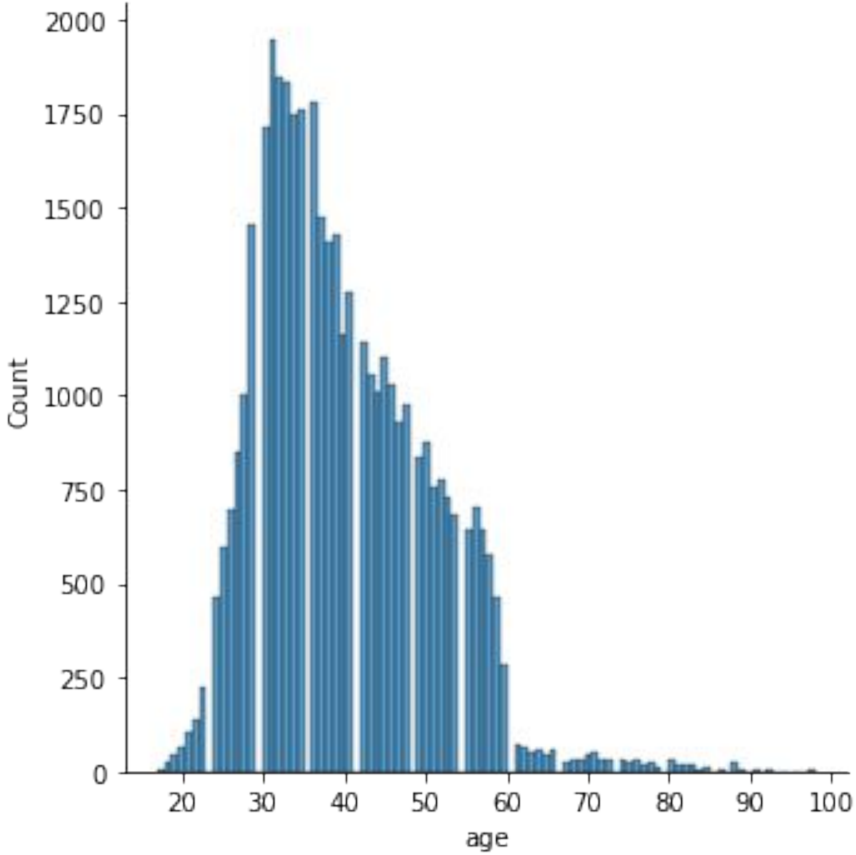
This asymmetry can affect statistical analyses, making measures like the mean less representative of the central tendency. Recognizing and adjusting for skewness is important for accurate data interpretation.



Skewness in the data

The plots show how the skewness exist for the numerical data:

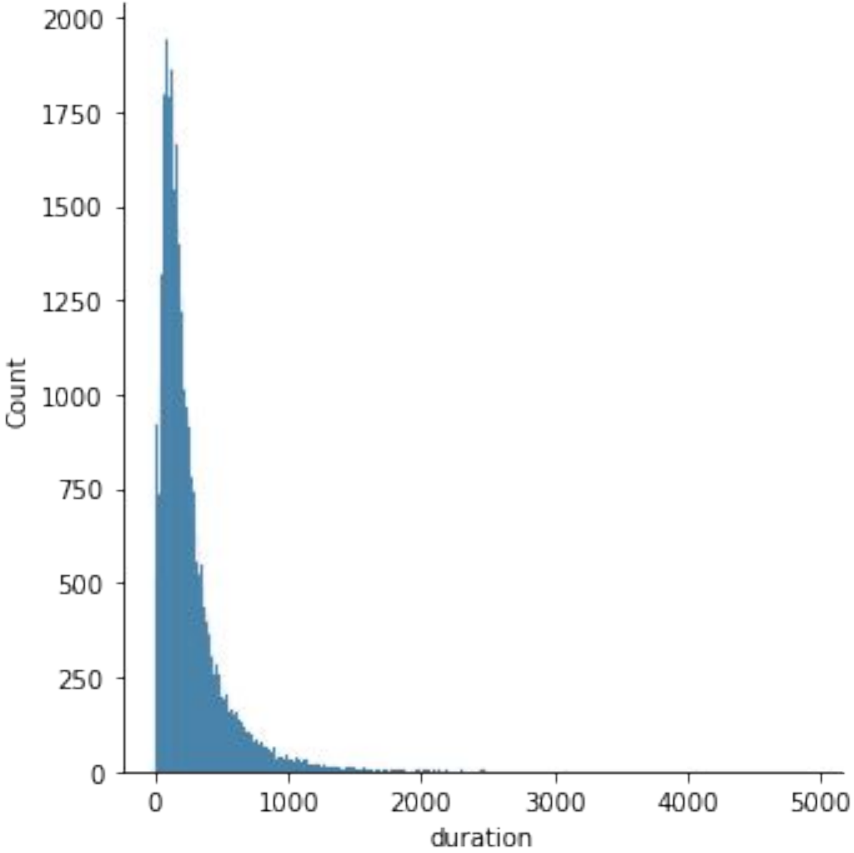
- age
-



Skewness in the data

The plots show how the skewness exist for the numerical data:

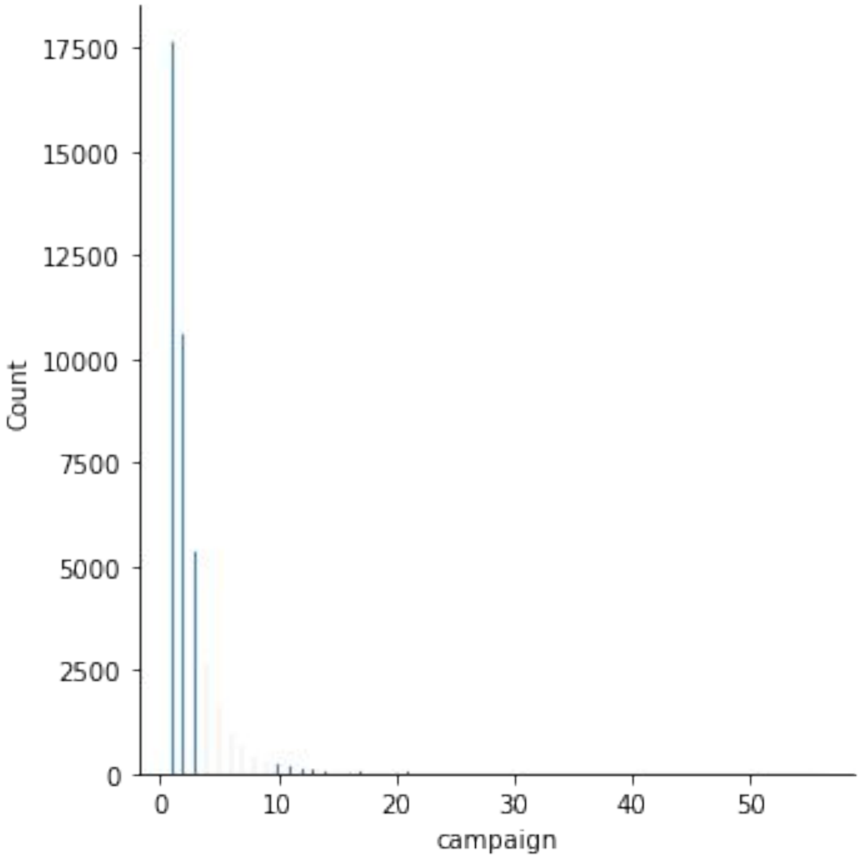
- age
- duration
-



Skewness in the data

The plots show how the skewness exist for the numerical data:

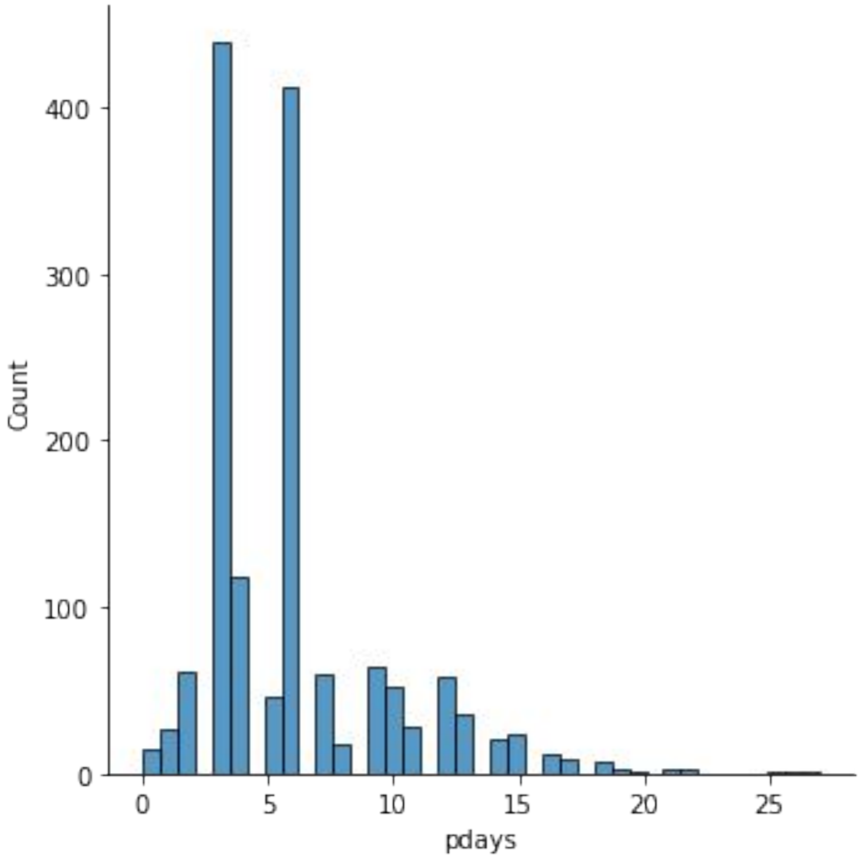
- age
- duration
- campaign
-



Skewness in the data

The plots show how the skewness exist for the numerical data:

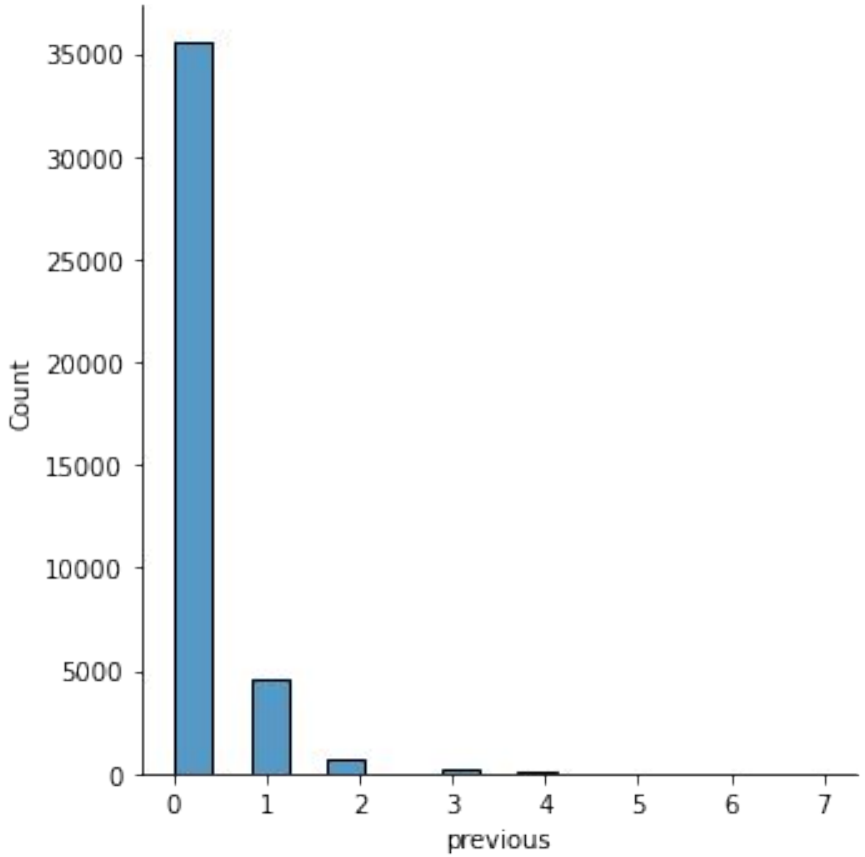
- age
- duration
- campaign
- pdays
-



Skewness in the data

The plots show how the skewness exist for the numerical data:

- age
- duration
- campaign
- pdays
- previous



Handling skewed data.

All the data appears to be positively skewed, and therefore we can apply techniques to transform it into normalized data.

In the skewed data, the tail region acts as an outlier for the statistical model and can impact the model’s performance. This is especially true for regression-based models. Too much skewness degrades the model’s ability to describe typical cases because the model has to deal with rare cases on extreme values.

We do not want to remove any data so we can employ techniques such as:

- **Log Transformation**
- **Square Root Transformation**