Chapter 1

General Context

1.1 Subject

In recent years, with the emergence of new techniques and different inventions, many solutions have been developed to ease the lifestyle of the human beings. Nowadays generations have access to buy their needs from home and that with different ecommerce websites developed in the last years. E-commerce websites often feature a variety of features such as online catalogs, shopping carts, payment gateways, and customer service tools, making it easy for consumers to find and purchase the products they need. As a result, e-commerce has become a crucial aspect of modern business, and many companies are investing in e-commerce projects to improve their online presence and increase their sales revenue.

1.2 Problem Identification

Mathematics is nothing more but a language and set of tools used to model complex systems with certain efficiency and precision. In many fields, including science, engineering, finance and technology, mathematical reasoning is fundamental to understand the problems tackled; however, its utility can be more important in the use of the results rather than understanding the mathematical approach. As much as it is important to understand the mathematical stand of a problem, some cases require only the understanding of the context and the process to move on with more important steps depending on the goal and need.

1.3 Objectives

With e-commerce, consumers can shop from the comfort of their own homes, at any time of the day or night, while businesses can reach a wider audience and increase their sales potential. Our platform offers a variety of possibilities where customers can ask for certain products and check for their availability while making sure to specify the price and details for their demand.

1.4 Suggested solution

Artificial Intelligence has the potential to revolutionize the E-Commerce by implementing some models that will help us generate better results and predict the behavior of some elements.

1.5 DS Objectives

Out of the business objectives, we would extract 3 main data science objectives as follows:

- Selling Price prediction: When a customer is submitting an offer for a certain product, a price range will be suggested in function of the different attributes' product.
  - ⇨ Regression Model to predict the product price.
- Forecasting sales peak: Suppliers will be able to visualize customers' behaviors through the time range specified and find predictions for upcoming periods.
  - ⇨ Prediction Model to predict the sales through time.
- Recognize fraud cases: When a customer is paying for a certain product, some frauds can happen from time to time.
  - ⇨ Classification Model to classify transactions into frauds and non-frauds.

Chapter 2

Methodology of Work

3.1 CRISP-DM

3.1.1 Definition

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It is a widely-used data mining process model that outlines a set of steps to guide data analysts and data scientists through the stages of a data mining project.

3.1.2 Steps of CRISP-DM plan

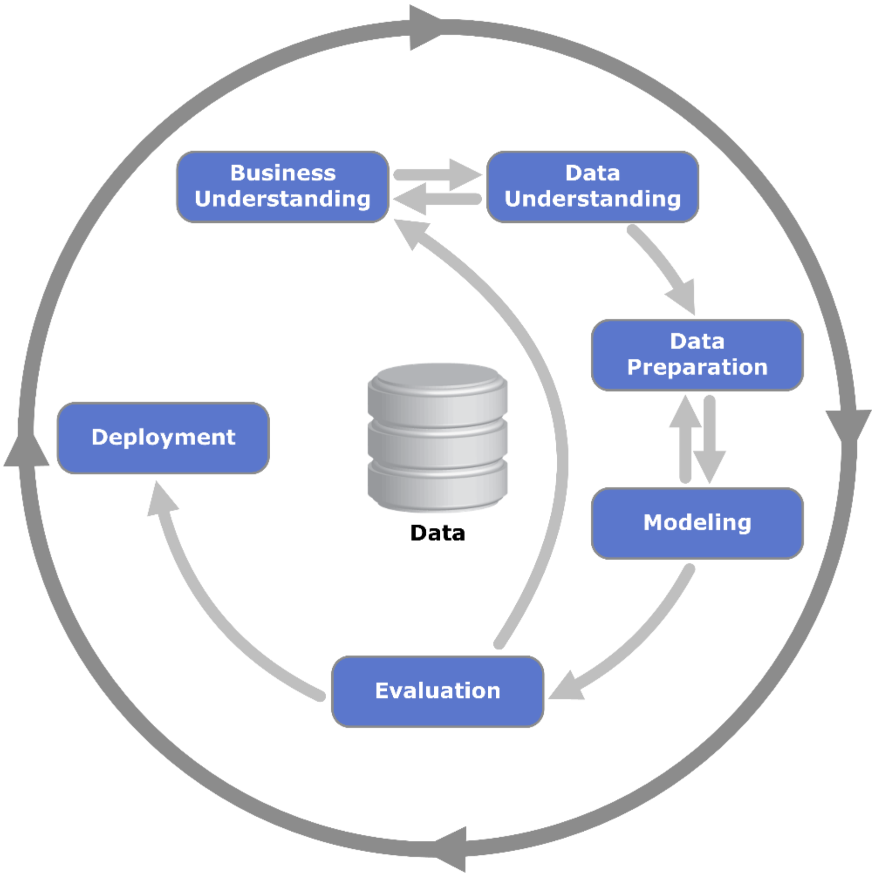


Figure 3.1: CRISP-DM methodology

The CRISP-DM model is comprised of six main phases:

- Business Understanding:
  In this phase, we will define the problem, the solution, and extract data mining objectives from the business objectives.
- Data Understanding:
  In this phase, the data requirements for the project are identified, and the available data is assessed for suitability and quality.
- Data Preparation:
  In this phase, the data is cleaned, transformed, and formatted to prepare it for analysis.
- Modeling:
  In this phase, statistical and machine learning models are developed and tested using the prepared data.
- Evaluation:
  In this phase, the model's performance is evaluated based on predefined criteria, and modifications are made as necessary.
- Deployment:
  In this phase, the final model is deployed into production, and the project's results are communicated to stakeholders.

The CRISP-DM model is an agile process, and the results of each phase inform the other phases. The model is intended to be flexible and iterative, allowing for adjustments and refinements as the project progresses.

Chapter 3

Phase 1: Data Analysis

5.1 Data Collection

Data collection is the process of gathering data from various sources, such as databases, files, surveys, and external sources. It involves designing and executing data collection methods and storing the collected data for analysis.

5.1.1 Data Collection Methods

There are mainly three methods to collect data:

- Data gathering through forms and questionnaires.
- Web scrapping.
- Direct download.

5.1.2 Data

- E-commerce Products dataset: 2451 lines in function of 8 features (Products and their categories and subcategories with the date and price).
- Transactions dataset: 623 lines in function of 12 features (Customers, payments methods and transactions amounts).
- Sales dataset: 1048575 lines in function of 26 features (item, status, amount, method and price).

5.2 Data Understanding

Data understanding is the process of exploring and familiarizing oneself with the collected data. It involves examining the data to identify patterns, trends, and relationships and gain a better understanding of the quality and completeness of the data.

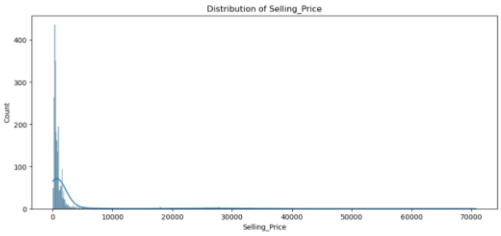
Chapter 4

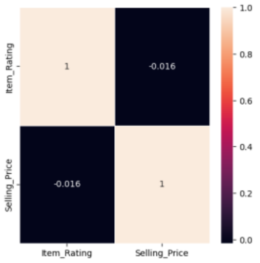
Phase 2: Data Preparation

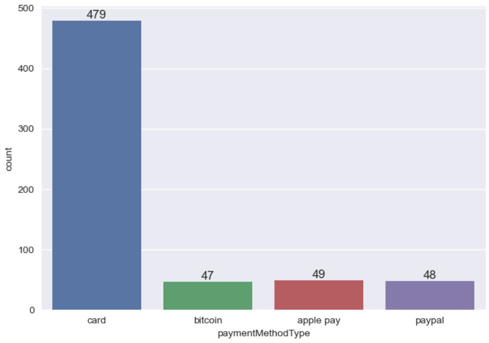
The data preparation phase involved cleaning, transforming, and organizing the raw data to ensure it was accurate, complete, and ready for analysis. As we use two different types of data, both images and problems, the preparation phase differs from one type to another.

6.1 Visualizations:

Data visualization is an important aspect of data science because it allows us to communicate complex information in a visual and intuitive way. A well-designed visualization can help to identify patterns, trends, and outliers in data that may not be apparent from raw data or summary statistics.

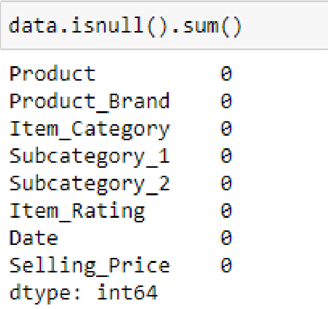






6.2 NaN and duplicate values:

NaN (Not a Number) values and duplicate values are common issues that data scientists encounter when working with datasets. Those values may be caused because of the missing data or undefined one.

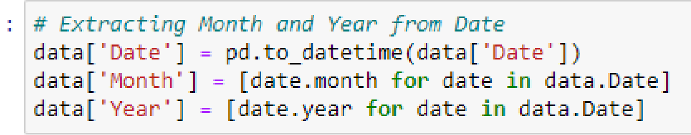


```
data.isnull().sum()
```

```
Product          0
Product_Brand    0
Item_Category    0
Subcategory_1    0
Subcategory_2    0
Item_Rating      0
Date             0
Selling_Price    0
dtype: int64
```

6.3 Feature Engineering:

Feature engineering is the process of transforming raw data into features that can be used to train machine learning models. The goal of feature engineering is to extract relevant information from the raw data and create informative, discriminating, and non-redundant features that capture the underlying patterns and relationships in the data.

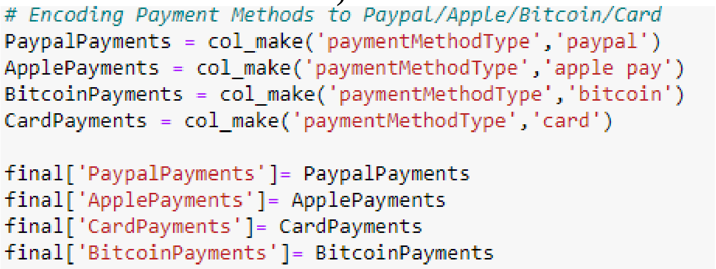


```
# Extracting Month and Year from Date
data['Date'] = pd.to_datetime(data['Date'])
data['Month'] = [date.month for date in data.Date]
data['Year'] = [date.year for date in data.Date]
```

6.4 Encoding:

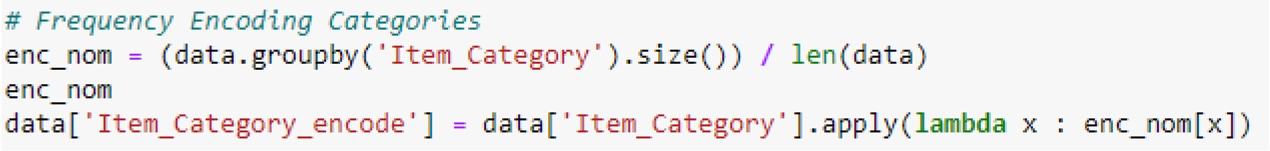
Encoding is the process of converting categorical data into a numerical representation that can be used as input for machine learning algorithms. Categorical data refers to data that consists of non-numerical values, such as color, gender, or country of origin.

Some methods of encoding include One Hot Encoding (creates binary column for each category), Binary Encoding (creates a binary representation for each category) and Frequency Encoding (replaces each category with its numerical values).



```
# Encoding Payment Methods to Paypal/Apple/Bitcoin/Card
PaypalPayments = col_make('paymentMethodType','paypal')
ApplePayments = col_make('paymentMethodType','apple pay')
BitcoinPayments = col_make('paymentMethodType','bitcoin')
CardPayments = col_make('paymentMethodType','card')

final['PaypalPayments']= PaypalPayments
final['ApplePayments']= ApplePayments
final['CardPayments']= CardPayments
final['BitcoinPayments']= BitcoinPayments
```

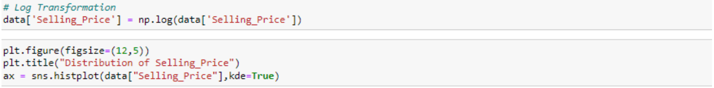


```
# Frequency Encoding Categories
enc_nom = (data.groupby('Item_Category').size()) / len(data)
enc_nom
data['Item_Category_encode'] = data['Item_Category'].apply(lambda x : enc_nom[x])
```

6.5 Scaling:

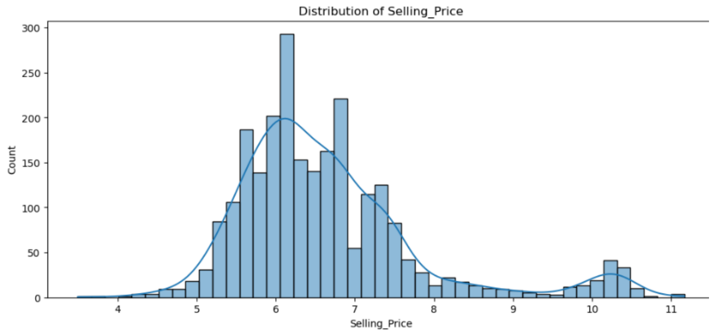
Scaling is a crucial step in the data preprocessing phase, and it is particularly important when dealing with numerical features that have different scales or units. The goal of scaling is to transform the features so that they have similar scales and ranges, which can improve the performance of machine learning algorithms.

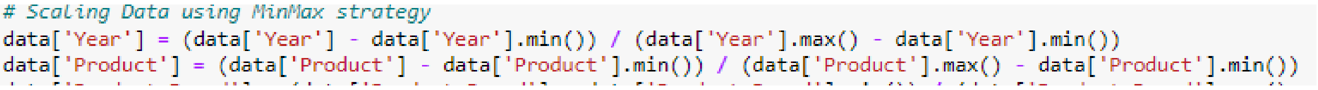
Some examples of scaling are the minmax scaler, log transformation and standardization.



```
# Log Transformation
data['Selling_Price'] = np.log(data['Selling_Price'])
```

```
plt.figure(figsize=(12,5))
plt.title("Distribution of Selling_Price")
ax = sns.histplot(data["Selling_Price"],kde=True)
```





```
# Scaling Data using MinMax strategy
data['Year'] = (data['Year'] - data['Year'].min()) / (data['Year'].max() - data['Year'].min())
data['Product'] = (data['Product'] - data['Product'].min()) / (data['Product'].max() - data['Product'].min())
```

Chapter 5

Phase 3: Modeling

The modeling phase in data science involves selecting and training a machine learning model on the preprocessed data. The goal of this phase is to develop a model that can make accurate predictions on new, unseen data.

7.1 Prediction models: