Predicting investment attractiveness using machine learning: Egypt case study

By

Alaa Abdelkarim Ahmed Mohamed Hassen Seif

P-EM0199/23

Supervisor: Dr. Aslam Mia

Academic Year 2023/2024

Semester 1

Abstract

The study discusses the economic importance of foreign direct investment (FDI) in Egypt's investment industry and the country's difficulties in luring FDI. To forecast and predict the market, this paper emphasizes the importance of machine learning techniques and how they can aid investors in making more informed choices. Along with discussing FDI's influence in these processes, the paper also examines the factors determining openness and integration into the global economy. Analyzing the problems and obstacles Egypt's investment sector faces, figuring out what keeps investors away, and offering potential fixes are the goals of the study. Also, the goals of the study are to assess the state of the Egyptian investment market now, identify the factors influencing investor decisions, and assess the impact of those factors on investor choices. The study focuses on identifying and evaluating the issues in Egypt's investment industry that discourage investors or force them to choose to make their investments abroad or withdraw them. Thus, we found that The Neural Network is the best model for machine learning with an accuracy of 98% and we found that the GDP growth rate is the most influencing factor for foreign direct investment and ease of doing business is the most influencing factor for domestic investment.

Keywords: Foreign direct investment (FDI), Machine learning, The Neural Network, The GDP growth rate, Ease of doing business, and Domestic investment.

Table of Contents

List of figures

List of tables

Abbreviation List

Foreign direct investment	FDI
Gross domestic products	GDP
The Ownership, Location, and Internalization paradigm	OLI
the United Nations Conference on Trade and Development	UNCTAD
The Middle East and North Africa	MENA
Machine learning	ML
Logistic Regression	LR
Support Vector Classifier	SVC
support vector machines	SVM
Neural Networks	NNs
Multinational corporations	MNCs
General Authority for Investment and Free Zones	GAFI
Exploratory Data Analysis	EDA
Return on investment	ROI

Acknowledgment

First of all, I thank God Almighty for giving me the strength, knowledge, and understanding to complete this project after so much hard work, as his love was more than enough to motivate me throughout this journey. Then, I'd like to express my gratitude to my supervisor, Dr. Aslam Mia, for his guidance, support, and constructive criticism of this research. I also would like to thank Eng. Menna Selim for her support and impassioned encouragement. I also want to thank my family for their tremendous support, who have been a source of inspiration for my academic pursuits. God bless you. Finally, I am deeply grateful to all my DEBI friends for their support and encouragement, and I am truly thankful for their contributions to my success.

Chapter 1

1.1. Introduction

The goal of this chapter is to explain the study's background, state the goal and objectives, present the research questions, and finally provide the scope and a brief introduction to the remaining chapters of the report.

Egypt's investment sector is essential to the growth and success of its economy. Technology transfer, employment creation, and economic growth are all significantly influenced by foreign direct investment (FDI). FDI inflows to Egypt have decreased recently, at the same time, which raises questions about the nation's attraction to investors.

To maximize investing strategies, it is critical to effectively predict market trends and to understand macro-financial difficulties. Through the integration of machine learning methodologies for market forecasting and prediction, investors can utilize comprehensive analytical instruments in their making decision process (Lee et al., 2019).

Openness to the world economy brings benefits such as technological linkages, access to ideas, and larger markets. The determinants of openness and integration into the world economy are explored. While FDI plays a role, deeper policy changes such as institutional reforms and better governance may be necessary. FDI is seen as a stable source of capital flows and a means to transfer technological progress (Fakher15, 2014).

1.2. Background

Economic experts recommend that developing countries rely on foreign direct investment (FDI) as a source of external funding. When compared to other types of capital inflows, FDI is thought to be more advantageous for economic growth. However, it is still unclear if developing countries gain anything from investing a lot of resources in attracting FDI. Potential advantages could include beneficial consequences like improved supplier efficiency and technical training. The amount of investment made by both domestic and foreign investors is influenced by many factors, such as the GDP growth rate, exchange rates, economic output, liberalization, and inflation rate. To drive changes in FDI levels, it is imperative to evaluate these aspects according to (Salem, n.d.).

According to UNCTAD data, FDI flows to developing countries have been rising over time, averaging 30.6% in the 1990s and 31.6% in the 2000s. Despite the global financial crisis, developing countries received 45.6% of all foreign direct investment in 2010. While the share retreated to 44.5% in 2019, developing countries received 48% of total FDI in 2019 (Yassin et al., 2020).

One well-known paradigm for comprehending foreign investment decisions is the Ownership, Location, and Internalization (OLI) paradigm. This framework states that companies choose to produce abroad under three circumstances. Firstly, they have to have ownership-specific benefits over businesses in other countries, like exclusive property rights or intangibles like a competitive advantage in the market or innovative products. Second, there ought to be a benefit of location, like lower costs, easier access to resources, being close to a sizable market, or avoiding trade restrictions. Lastly, the company needs to want to hold onto these advantages and apply them to various markets (Blanton & Blanton, 2007).

The Middle East and North Africa (MENA) region has attracted much media attention due to economic, political, and security instability. The region is diverse, with wealthy countries rich in natural resources as well as resource-poor countries suffering from poverty and conflict. However, the MENA region faces challenges such as poor governance, corruption, a lack of infrastructure, and limited economic diversification, which have restricted investment growth and resulted in poor FDI performance. Despite these challenges, the region offers foreign investors market potential due to its strategic location, young population, and abundant oil and gas reserves. However, the MENA region has received little attention in terms of the factors that influence foreign direct investment. FDI inflows fluctuate and are concentrated in a few countries and sectors, primarily natural resources and non-tradable. Because of historical conflicts and political transitions triggered by events such as the Arab Spring, the region's business environment is unique (Dimitrova et al., 2020).

The globalization of the economy depends on the ongoing developments in computer science and data innovation, and although there are many approaches to calculating the cost of financial exchange, research has primarily focused on the latter. Due to the obvious shortcomings of traditional analytical methods in handling nonlinear problems, several machine learning algorithms are used in stock exchange queries. Financial backers might use a model that predicts the growth path of a stock's value and investment attractiveness to make smart decisions, improve productivity, and minimize potential losses (Padhi et al., 2022).

Egypt is one of the countries that have many resources and a high percentage of withdrawing investment. this study aims to fill the gap, determine the reasons for this, and give recommendations for the investment sector and Egypt’s economic growth.

1.3. Problem statement

The FDI has fallen by 12.5% in previous years, so it is important to analyze the issues and challenges facing Egypt's investment sector, resulting in investors withdrawing their investments and shifting their investments to other countries. It is important to identify the variables preventing investors and investigate possible solutions for improving Egypt's investment sector. so, it is important to predict what attracts investors and understand the factors influencing investors' decisions.

1.4. Research Objective

The Objectives of this study are:

RO1: To analyze the current situation of Egypt's investment market and understand the main challenges facing investors.

RO2: To determine the variables that affect investor choices and evaluate how they affect investors' decisions.

1.5. Scope of study

This study focuses on analyzing the problems in Egypt's investment sector that are turning away investors or making them choose to invest in other countries. It will include predicting what attracts investors most and identifying specific challenges that investors face, such as lack of infrastructure, political instability, corruption, and governmental hurdles, as well as economic diversification. In addition, the study will look into what draws investors to other countries and possible ways to develop Egypt's investment environment. The study will primarily cover the economic aspects related to investment decisions and their impact on Egypt's overall economy. However, it will not delve into broader political or social issues that may indirectly affect the investment sector.

1.6. Significance of the study

This paper provides a thorough investigation of determinants impacting investment attractiveness, which has important implications for global collaboration and economic policymaking. Using the information, policymakers may create focused plans that will increase the country's competitiveness, attract and retain investors, and promote sustainable economic growth. The study aids in risk reduction, economic diversification, and the development of investor trust by addressing concerns about business practices, legal frameworks, and political stability.

1.7. Research Questions

RQ1: What is the current situation of Egypt in investment market and what are the main challenges facing investors.

RQ2: What are the variables that affect investor choices.

1.8. Organization of the Paper

The organization of this report is as follows.

- Chapter 1: This chapter contains the introduction to the problem in this study and the aim, objectives, research questions, and scope of this study.
- Chapter 2: This chapter develops on the literature reviews conducted for this study.
- Chapter 3: This chapter explains the methodological structure of the study.
- Chapter 4: This chapter reports results and discussion
- Chapter 5: This chapter discusses the limitations of the study, recommendations regarding future work, and conclusion.

Chapter 2

2.0 Literature Review

This chapter aims to present the literature review performed for this study, Investment attractiveness, Supervised Machine Learning, and four machine learning algorithms employed in classification will be presented in this study.

2.1 Investment Attractiveness

The appraisal of investment performance may become more accurate with the use of machine learning algorithms. Traditional investment models may miss patterns, trends, and indicators that these models can spot by utilizing behavioral data, historical market data, and other relevant data. This makes it possible to make better decisions about risk management, portfolio optimization, and investment strategies.(Chia-Cheng et al., 2020)

Through data driven decision support, machine learning can assist investors in managing risk and uncertainty associated with venture capital investments. This can assist with a variety of tasks, including discovering possible acquirers, finding business possibilities, matching coinventors and deals, managing a portfolio more effectively, and more. It can also help in recognizing the risks involved with a company and projecting its future success, which could help investors learn more quickly (Arroyo et al., 2019).

2.2 Supervised Machine Learning

Machine learning (ML) has been around since the 1950s, but in the recent two to three decades, its use has grown significantly because of breakthroughs in database and data collection technologies, increased processing power, and the availability of open-source software. Computers can learn without explicit programming thanks to machine learning. Machine learning algorithms are classified into three primary groups: supervised learning, unsupervised learning, and reinforcement learning (Hu et al., 2021).

Using labeled data, supervised learning entails teaching an algorithm to provide predictions or classifications for new observations.

Supervised learning aims to acquire knowledge about mapping between input and output by utilizing sample input output pairs. This enables the model to classify or predict previously unobserved data. Supervised learning is frequently used in regression (e.g., forecasting sales, estimating housing prices) and classification (e.g., spam detection, and picture identification) (Berry et al., 2019).

2.3 Classification in machine learning

In machine learning, classification is a basic idea that entails grouping data into different classes or groups. It is a supervised learning method, which trains a model for future predictions using labeled data (Muñoz et al., 2018).

2.3.1 Logistic Regression

LR is a statistical method used to model the relationship between a dependent variable and one or more independent variables. Dichotomous outcomes can be modeled effectively, and it is able to assess the relationship between a categorical target variable and features that are either numerical, categorical, or both. LR can be used when the target variable is binary (either zero or one) or dichotomous (Nhu et al., 2020). However, a variation of this algorithm (multinomial logistic regression) is used to classify more than two categories of the target.

2.3.2 Support Vector Classifier (SVC)

The Support Vector Machine oversees determining the decision boundary to divide several classes and increase the margin which is the separation (perpendicular to the line) between the nearest dots and the line. Finding a hyperplane that splits a dataset into two groups most effectively is the foundation of SVMs. Data points that fall (exactly) on the margins' edges are known as support vectors. To calculate the margin, just these points are required. (Ke et al., 2023)

There are two types of data: separable and non-separable data. linear separable data, SVM tries to find the Hyperplane. Unfortunately, in real datasets, most of the data are non-linear separable so it is hard to find the hyperplane that's why SVM creates new two concepts: Soft Margin and Kernel Tricks. (Esteki & Naghsh-Nilchi, 2022)

The soft margin is a method used to balance the tradeoff between having a model that is more accurate on the training data and one that generalizes better to unseen data. This is done by allowing some misclassification of the training data, rather than trying to fit the model perfectly to the training data. The soft margin is often used in support vector machines (SVMs), which are a type of model used for classification tasks.

The kernel trick is a mathematical technique used to map data from a low-dimensional space into a higher-dimensional space, to perform a computation more efficiently or to improve the performance of a machine-learning algorithm. It is often used in conjunction with SVMs. In an SVM, the kernel trick is used to transform the data into a higher-dimensional space, where it is 7 easier to find a hyperplane that can separate the data points into different classes. The kernel function used to perform the transformation can be chosen based on the characteristics of the data. Sklearn.svm.

SVC () includes linear, poly, rbf, sigmoid, precomputed, or a callable as our kernel/transformation.

SVM has advantages like in the higher dimension, it is quite effective. Also, useful when there are more features than training samples. When classes can be separated, the best algorithm. Outliers have less effect. However, all these advantages have disadvantages, especially when working for a large amount of time requires time and does not work effectively with overlapping classes. (Gupta et al., 2019)

SVM can be applied in several fields: Document classification or document categorization, computer vision, and handwriting recognition.

2.3.3 Naïve Bayes

Naïve Bayes techniques are probabilistic classifiers based on Bayes theorem with strong individuality (naïve) among the attributes (Sarker, 2021). Naïve assumptions are applied in NB where each pair of attributes are independent conditionally given their corresponding class label. In other words, each feature is independent statistically of the other features for a given class. To put it simply, the value of an attribute is not associated with the value of an attribute Xj. NB techniques are computationally efficient. Despite their naïve assumptions, one example of successful implementation of NB is in the field of text classification. For a given observation, the probability of being a member of class y can be calculated as shown in the equation below:

Equation 1 *Naïve Bayes probability*

$$P(y|x_1, \ldots, x_n) = \frac{P(y) \times P(x_1|y) \times \ldots \times P(x_n|y)}{P(x_1, \ldots, x_n)}$$

For observation x, class y with the highest is predicted as the output of the model. The value of does not have any impact on choosing the class with the highest. Therefore, Equation 1 and Equation 2 predict the same class label,

Equation 2 *Naïve Bayes probability*

$$P(y|x_1, \dots, x_n) \approx P(y) \times P(x_1|y) \times \dots \times P(x_n|y)$$

where it approximates P(y) as the relative prevalence of class y using training set data points. NB techniques vary corresponding to their method to calculate the likelihood of the feature given class y; that is, for instance, multinomial Naïve Bayes utilizes a multinomial distribution for calculating while Gaussian Naïve Bayes technique utilizes a Gaussian distribution (Hastie et al., 2001).

2.3.4 The Neural Network

Neural Networks (NNs) are computer models that draw inspiration from the architecture and operations of the human brain. While they usually work on one task at a time, they can address multiple-state classification problems by employing several output units. They can conduct regression and classification tasks.

The input and activation functions of the unit, the network design, and the weight of each input link are the three main factors that affect ANN. The weights' current values which are originally set to random values determine the ANN's behavior. Instances of the training set are frequently exposed to the network during training. An instance's input values are put on the input units, and the network's output is compared to the intended output for that instance. Next, a little adjustment is made to the network's weights to move the output values of the network closer to the values for the intended output. An array of algorithms is available for neural network training (Osisanwo et al., 2017).

Studies utilizing ML approach integrated into investment analysis encompass a broad range of subfields. Examples include applying ML techniques to comprehend investment climates in particular economic expansion projects, leveraging neural networks (NN) to enhance discounted cash flow procedures, and evaluating investments utilizing a fuzzy logic and real choice combination (Aziz et al., 2022).

Overall, neural networks are powerful tools for performing complex tasks such as pattern recognition, classification, and regression, and they can be trained using various algorithms to improve their performance on specific tasks.

2.4 Foreign Direct Investment (FDI)

Foreign direct investment (FDI) refers to the acquisition by a resident or corporation of one country of a productive asset in another country, typically conducted by multinational corporations. FDI can be accomplished through the construction of new facilities or the acquisition of existing ones. It is also characterized by an investment that provides control over productive facilities overseas, typically defined as ownership of 10% or more of a company's equity (Kerner, 2014).

For many years, researchers have been curious to find out what makes a country attractive to foreign direct investment. In the past, economic variables like infrastructure, labor costs, market size, and exchange rates were thought to be important explanatory factors. However, after North's influential work in the 1990s, scholars started concentrating more on the power of institutions, which they defined as the "rules of the game in a society" or the artificial limitations that people impose on one another to shape social interactions. Positive or negative effects are commonly used to characterize the relationship between institutional factors and FDI attractiveness, with FDI being attracted by factors like democratic institutions, political stability, and rule of law and being discouraged by factors like corruption, tax policies, and cultural distance(Bailey, 2018a).

A change in focus among policymakers has resulted from developing country's increasing desire to bring in more foreign direct investment. Beginning in the early 1980s, many countries including developing countries have removed challenges to FDI flows. 2013 saw the adoption of 87 policy measures affecting foreign investment by 59 countries, according UNCTAD. From 25 to 27 percent of investment policies overall in 2012–2013, more restrictions or regulations were imposed. For the last few decades, global FDI inflows have increased dramatically, exceeding growth rates in global trade and GDP. Some countries, however, attract more FDI than others, and FDI inflows are not consistent across all countries. The desire to attract more FDI is driven by the possibility of reaping benefits like increased productivity, technology transfer, exposure to new procedures, and market access. In addition to being less destructive, FDI is thought to be less volatile than other short-term flows (Zghidi et al., 2016).

2.4.1 Economic indicators and their effect on FDI

In this part we will discuss the effect of each economic indicator on FDI and domestic investment.

GDP Growth rate

Economic growth per capita is primarily driven by improvements in productivity, which refers to producing more goods and services with the same inputs. In economics, economic growth refers to the growth of potential output at full employment, caused by growth in aggregate demand or observed output. It is commonly measured as the percentage rate of increase GDP (Almfraji & Almsafir, 2014).

The relationship between foreign direct investment FDI and GDP growth has been a topic of interest in recent times due to the liberalization of capital movement. According to World Bank data, FDI has increased significantly, indicating a positive relationship between FDI and economic growth. Though there is conflicting data from empirical studies regarding FDI's effect on economic growth, most of them point to the investment's beneficial effects. According to theoretical explanations, FDI can stimulate growth by introducing new technologies, raising labor productivity, and decreasing capital rental rates. FDI may, however, also have unfavorable outcomes, such as stifling competition and advancing its agenda. While some empirical studies have found no direct relationship, many have found a positive relationship between FDI and economic growth. The impact of economic growth on attracting FDI is also debated, with some studies suggesting that higher growth rates attract more FDI, while others argue against it (Zhang, 2001).

FDI can be greatly influenced by GDP growth, but the effectiveness of this investment is contingent upon the caliber of institutions in the country to which it is directed. To increase FDI and encourage economic growth, the government needs to enhance investor protection, government stability, and the business climate to improve the institutional environment (Chowdhury & Mavrotas, 2006).

Overall, the relationship between FDI and economic growth is complex and varies across sectors and specific manufacturing industries. Analyzing FDI data at a more detailed level is essential to understand its connection to economic growth accurately.

According to Salem, there is a positive relationship between FDI and GDP growth(Salem, n.d.).

H1: GDP growth rate has a positive effect on FDI and domestic investment.

Inflation rate

An indicator of macroeconomic instability, inflation is a factor in the economy that investors should be aware of. In a host country, a high rate of inflation can cause currency devaluation, which reduces real investment earnings and purchasing power. When government spending drives up the money supply, inflation results. High inflation raises borrowing costs and affects interest rates, which deters investors from making investments in a host country. Additionally, it lowers the returns from domestic investment and raises production costs in the host market. Consequently, a high inflation rate in a host country encourages outward FDI (Wako, 2021).

High inflation rates decrease the value and returns of investments, leading to financial losses for investors. High inflation also raises investment risk, uncertainties, and balance of payments deficits. Changes in these variables are significant to foreign capital owners because they impact how risk is perceived (Topal, 2016).

According to Adebayo et al, inflation can have significant effects on FDI inflows in emerging economies. The relationship between inflation and FDI inflows varies across countries, with some studies finding positive impacts, while others find negative or insignificant impacts (Adebayo et al., 2020).

H2: Inflation rate has a negative effect on FDI and domestic investment.

Exchange rate

FDI plays a crucial role in achieving sustainable economic development, and FDI flows are sensitive to exchange rate considerations. Exchange rate volatility can affect investment decisions and profits of multinational firms(Latief & Lefen, 2018).

Exchange rates can influence both the total amount of FDI and the allocation of investment spending across different countries. Exchange rate volatility provides opportunities for investors to invest in foreign currency to obtain higher yields, which can impact the price of exports and imports(Loungani & Razin, 2001).

The world economy has experienced economic crises and macroeconomic uncertainty in the last two decades, leading to price and exchange rate volatility. Exchange rate volatility has a negative impact on FDI inflows (Dal Bianco & Loan, 2017).

H3: Exchange rate has a negative effect on FDI and domestic investment.

2.4.2 Political Stability indicators and their effect on FDI

In this part we will discuss the effect of each political stability indicator on FDI and domestic investment.

Political Stability Index

Political events affect investors' perceptions of market risk and force adjustments to asset valuations and portfolio allocation decisions. A global exodus of capital and possible financial and economic crises has resulted from growing conflicts. Financial markets are impacted by a country's monetary and fiscal policies, which are shaped in large part by political developments. Investors, regulators, and academics should all be aware of how political developments impact asset returns and volatility (Ahmed, 2017).

Media coverage and public awareness of episodes of political violence are very important and have a potential impact on key economic indicators such as FDI (Kurecic & Kokotovic, 2017a).

Economists view political instability as a major obstacle to economic progress. Its widespread occurrence and negative impacts have sparked significant interest in the field. Policymakers with shortened horizons may implement suboptimal policies, while frequent policy changes create volatility and deter investment. Additionally, instability reduces FDI (Nazeer & Masih, 2017).

H4: Political stability has a positive effect on FDI and domestic investment.

Rule of Law

An effective legal system and the application of the rule of law are essential for attracting FDI and investment to a country. A clear and efficient legal framework that guarantees the protection of property rights is regarded as an essential requirement for taking FDI into account (Bailey, 2018b).

H5: Rule of law has a positive effect on FDI and domestic investment.

Corruption Perception

Egypt faces challenges in its investment policy despite some successes and an increase in FDI as a percentage of GDP. To maximize its potential as an investment destination, Egypt needs to address obstacles to business establishment and operation, improve transparency in the investment policy framework, and enhance investment promotion efforts. While corruption has been declining over time, Egypt still has a high level of corruption and a lack of transparency. One important political factor influencing FDI is corruption in the host country. FDI inflows can be hampered or facilitated by corruption, depending on how it affects governance and operating expenses. Corruption can result in ineffective practices, higher transaction costs, and economic distortion (Moustafa, 2020).

Good institutions are found to increase trade, while corruption in the host country has a negative relationship with FDI (Fakher15, 2014).

H6: Corruption has a negative effect on FDI and domestic investment.

2.4.3 Infrastructure indicator and their effect on FDI and domestic investment

The infrastructure of an economy, including roads, ports, railways, water supply, telecommunication systems, and internet access, is an essential factor that influences the decisions of foreign investors in a particular country (Alam Iqbal et al., 2019). Also, infrastructure is a crucial factor in attracting FDI as it directly impacts the productivity and profitability of MNCs (Fernandez, 2020).

while natural resources and market size play a significant role in attracting FDI to African countries, other factors such as infrastructure development, also influence investment decisions. The determinants of FDI in Africa are multifaceted and go beyond natural resources alone (Asiamah et al., 2019).

H7: Infrastructure has a positive effect on FDI and domestic investment.

2.4.4 Legal indicators and their effect on FDI

In this part we will discuss the effect of each legal indicator on FDI and domestic investment.

Government Investment Policies (Government Effectiveness)

For many years, economic geography has examined the decision made by MNEs regarding where to invest FDI. Businesses are diverse and have different preferences for locations and ways of evaluating different aspects of locations. For heterogeneous firms, the same location factors might have different effects. International companies' decisions about FDI and where to locate their investments are based on their understanding of the features of the host nation. The strategic intent and inherent qualities of a firm play a major role in determining these choices(Ye et al., 2019).

Changes in government regulations on FDI, such as privatization of state-owned enterprises and reforming tax policies, have influenced FDI inflows in specific regions. The theory of economic regulations suggests that variations in economic regulations can affect investment decisions. something should take under consideration that the multidimensionality of formal institutions and develop a typology of institutions to understand their effect on FDI decisions. In addition to contributing to the understanding of multinational enterprise strategy. It notes that there has been a shift in government policies from strict controls and restrictions to deregulation and liberalization, with many countries now welcoming FDI. However, the regulatory landscape remains uneven across countries, and institutional and other barriers still exist(Contractor et al., 2020).

H8: Government effectiveness has a positive effect on FDI and domestic investment.

Ease of doing business

According to World Bank the ease of doing business refers to the level of difficulty or simplicity involved in starting and operating a business in a particular country or region. It encompasses various factors such as the regulatory environment, legal framework, administrative procedures, access to finance, property rights, and contract enforcement.

H9: Ease of doing business has a positive effect on FDI and domestic investment.

Trade Openness

Trade openness refers to the expansion of world trade through the reduction or elimination of trade barriers, such as import tariffs. It offers consumers a wider variety of goods at lower prices.

Additionally, greater openness can stimulate foreign investment, leading to employment for the local workforce and the introduction of new technologies(Hao, 2023).

the openness of an economy plays a crucial role in attracting FDI and investment. More open economies tend to be more vulnerable to losing access to foreign financing, and a decrease in trade restrictions tends to increase FDI in host countries (Zenasni & Benhabib, 2013) .

According to Salem, there is a positive relationship between FDI and trade openness(Salem, n.d.).

H10: Trade openness has a positive effect on FDI and domestic investment.

Market size

Understanding the extent and possible effects of FDI on different economies requires an understanding of market size. Market size, which is determined by taking the company's sales over some time and dividing it by the total sales of the industry over the same period, refers to the total sales in an industry generated by a specific company. This measure gives a broad impression of a company's size for its industry and rivals. Businesses need to know the size of the market to determine the potential demand for their goods and services as well as to spot growth and expansion prospects or market entry (Kurecic & Kokotovic, 2017b).

In summary, market size is a fundamental factor in the context of FDI and investment, as it influences the potential demand for products or services, employment generation, and overall economic output.

H11: Market size has a positive effect on FDI and domestic investment.

2.5 Theoretical Underpinnings

In this part we will discuss the theory applied in this study

2.5.1 Eclectic Paradigm

The eclectic paradigm integrates elements of multiple theories of international business to comprehend FDI patterns, trends, and determinants. To invest globally, businesses require ownership (O), location (L), and internalization (I) advantages, according to the framework. Advantages of ownership include things like economies of scale, technology, and brands. Location advantages are host country attributes like resources, and infrastructure. Internalization advantages provide incentives for firms to internalize markets through FDI over alternatives like licensing(Stoian & Filippaios, 2008).

The paradigm of the eclectic (OLI) (Dunning, 1988). This theory states that a company's investments in its home or host country are determined by factors such as OLI, ownership (O), location (L), and internationalization (I). The ownership of intangible assets such as trademarks and patents increase brand awareness and trust. The benefits of location include lower transportation costs, easier access, and increased cross-border activity due to internationalization(Siddiqui & Aumeboonsuke, 2014).

According to the seminal eclectic or ownership, location, internalization (OLI) paradigm, MNE location decisions are determined by the interaction of ownership advantages (e.g., organizational know-how or proprietary technology), internalization advantages related to the particular transaction (e.g., licensing vs. FDI), and location advantages of the host country (e.g., political stability) Dunning 1980 (Nguyen & Do, 2020).

According to Dunning's eclectic theory (1993), cross-border FDI flows can be explained by ownership advantage (O), location factor (L), and internalization of transaction costs (I). Nonetheless, ownership advantages and the location factor are no longer sufficient to explain why some countries attract more foreign direct investment than others in considering the growth of internationalization, competition, and globalization(Ullah & Khan, 2017).

Main components in OLM paradigm:

Ownership Advantage

These are the advantages that multinational companies have over domestic companies in the marketplace because of their privileged ownership of tangible or intangible assets that are unique to their nationality. If the company has an internalizing advantage, it may decide to sell its ownership advantage assets or internalize them (Peter, 2010).

The importance of labor standards in a company's location choices is contingent upon how sensitive the company is to reputational and operating costs, which are impacted by ownership and industry factors (Maggioni et al., 2019).

The Ownership Advantage Dimension of the Ownership Location Internalization (OLI) Paradigm:

Political stability

The locational dimension of the OLI paradigm can be impacted by the political climate and stability of the host country. A supportive regulatory framework can lower political risk and lower internalization costs. States can also use financial incentives, human rights conditions, and preferential taxation policies to enhance the appeal of their location (Rashid et al., 2017).

Africa has an abundance of resources, but it has trouble attracting FDI. Just a small percentage of FDI that went into countries that were developing in 2009 went to Africa. Many times, institutional policies, political instability, and a lack of infrastructure are mentioned as barriers to FDI in Africa (Singh et al., 2011).

Corruption

According to Asbullah et al, corruption may have a negative impact on FDI, alter the composition of capital inflows and lower inward FDI stocks. Furthermore, there is proof that corruption in a country that imports capital can skew capital inflows away from FDI and towards foreign bank loans. On the other hand, some research suggests that corruption may even encourage FDI inflows, and there may even be a negative correlation between FDI and corruption, as well as differences in corruption between the host and source nations. A decrease in FDI inflows to the region coincided with the early 2011 Arab Spring eruptions in Egypt and Tunisia, suggesting that corruption may affect FDI. All things considered, the impact of corruption on FDI is complex and can change depending on certain circumstances. (Asbullah et al., 2022).

According to Mustafa, 2021, corruption in Egypt is positively associated with total FDI in both the short and long run (Moustafa, 2021).

A large market size can be attractive to firms with ownership advantages like superior technology, branding, or management skills. It offers a larger potential customer base and economies of scale for production and distribution. Government policies can create ownership and location advantages through subsidies, tax breaks, infrastructure investments, or skills development programs. These policies can make a market more attractive for FDI even during crises, especially if targeted towards sectors considered resilient or with high growth potential. High unemployment can be an ownership advantage for firms with internal efficiency advantages like superior production processes or lean management practices They can potentially hire skilled labor at lower costs during a crisis (Andreas & Carl, 2021).

Location

MNCs and FDI are the focus of the concept of location-specific advantage, which states that MNCs look for the best production location on a global scale that offers a high market potential, an ample labor force, and efficient raw material procurement. The incentives and variables that push multinational corporations to make investments in overseas markets are linked to this location-specific advantage (Park & Roh, 2019).

The Location Dimension of the Ownership Location Internalization (OLI) Paradigm:

GDP and trade openness

The OLI paradigm highlights the significance of trade openness and GDP in influencing FDI. These factors play a crucial role in attracting FDI to economy, contributing to economic growth, modernization, and income growth (Alikhanov & Khudiyev, 2020).

Infrastructure

The location dimension encompasses the role of the government, which can improve a location's appeal to MNCs by providing direct investment support or by enhancing essential advantages like infrastructure (Batschauer da Cruz et al., 2022).

Internalization

MNCs can use the internalization from the OLI paradigm as a framework to assess alternative approaches for organizing the creation and exploitation of their unique advantages in conjunction with the locational advantages of host countries. The commercial benefits that exist in an intra-firm relationship between the foreign investor and the investment recipient are the main focus of this sub-paradigm. It emphasizes that a firm is more likely to choose to engage in foreign production itself rather than license the right to do so, the greater the net benefits of internalizing cross-border intermediate product markets. According to this framework, there are several reasons to use internalization to take advantage of ownership-specific and locational advantages. These include incomplete markets for production inputs or assets, high transaction costs or time lags, and the ability to retain exclusive rights to assets for a sizable competitive advantage (Amal et al., 2010).

The Internalization Dimension of the Ownership Location Internalization (OLI) Paradigm:

Exchange rate and ease of doing business.

The internationalization dimension in the context of FDI inflows primarily arises from inadequacies in the factor market. These deficiencies include exchange rate volatility, ease of doing business, and other related factors that influence FDI inflows. Regarding FDI inflows, the exchange rate and ease of doing business are key internationalization factors that influence the investment environment and influence the decisions of multinational corporations and other investors. It is a reflection of the larger political and economic landscape that influences a nation's desirability as a location for foreign investment (Nagpal & Jain, 2019).

Inflation

Fiscal governance and the management of the macroeconomic environment are significantly impacted by inflation. High levels of inflation can cause economic hardship and discourage foreign investors, which lowers FDI. Unpredictable inflation rates can also breed uncertainty among investors, who may then demand higher prices to offset inflationary risks. This could result in a decrease in the amount of money invested (Samal, 2018).

2.6 Summary of hypothesis

Table 1 *Table of hypothesis*

No	Hypothesis
H1	**GDP growth rate has a positive effect on FDI and domestic investment**
H2	**Inflation rate has a negative effect on FDI and domestic investment**
H3	**Exchange rate has a negative effect on FDI and domestic investment**
H4	**Political stability has a positive effect on FDI and domestic investment**
H5	**Rule of law has a positive effect on FDI and domestic investment**
H6	**Corruption has a negative effect on FDI and domestic investment**
H7	**Infrastructure has a positive effect on FDI and domestic investment**
H8	**Government effectiveness has a positive effect on FDI and domestic investment**
H9	**Ease of doing business has a positive effect on FDI and domestic investment**
H10	**Trade openness has a positive effect on FDI and domestic investment**
H11	**Market size has a positive effect on FDI and domestic investment**

Chapter 3

3 Methodology

This chapter illustrates the steps to conduct the study. The primary software where the dataset is being used is Jupiter Notebook (6.4.5) and Google Collab on a PC with a 2.50GHz Intel(R) Core (TM) i5-4200M CPU processor and 8.00 GB RAM. Using Windows 10 operating system. The Python programming language was employed throughout the project, IBM SPSS Statistics was used to analyze relations, and Power BI Desktop was used for visualization. The chapter consists of 2 stages, one for predicting investment attractiveness and the other for identifying what affects investment decisions. The full step-by-step taken throughout this study can be found in the Appendix.

3.1 Stage 1 for predicting investment attractiveness

3.1.1 Data Collection

The data was obtained from Egypt's General Authority for Investment and Free Zones (GAFI) with 234605 observations, 12 features, and 1 target column. The target is determining if the firm is attractive to investors or not.

The data was available in Excel file format in the Arabic language, so the first step was to translate the data into English using Google Spreadsheets translating function.

The first step in the coding process is to install the necessary libraries. The library will be saved in the Jupiter Notebook and maybe it will be used later. The dataset was then imported into Python using the read.excel() function to convert the Excel into a dataset with labeled axes. The descriptions of each feature are in the below table.

Table 2: *Data Table for domestic investment*

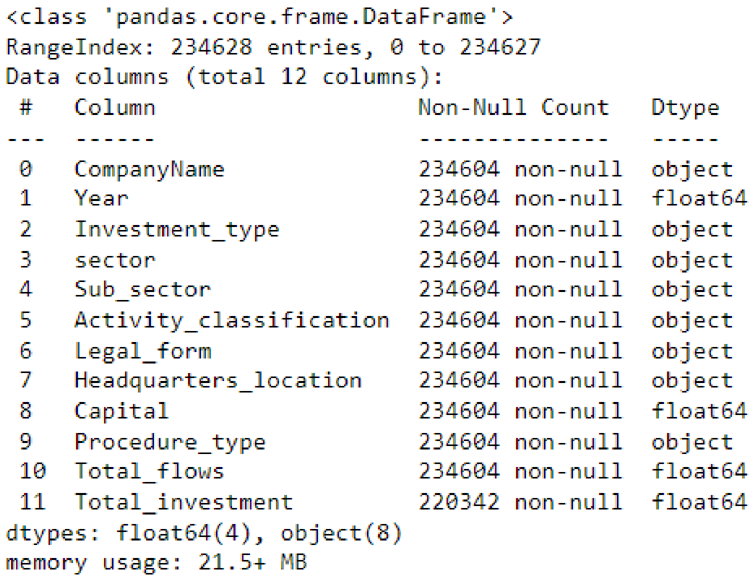
Name of the features	Data Type	Description
CompanyName	Text	Company name
Year	Date/Time	Establish year
Investment_type	Text	Type of investment
sector	Text	The sector of investment

Sub_sector	Text	Sub sector for the company
Activity_classification	Text	The activity that the firm practices
Legal_form	Text	The legal form
Headquarters_location	Text	The location of the firm
Capital	Number	The paid capital
Procedure_type	Text	The procedure type at the time of establishment.
Total_flows	Number	Total flow for the firm in a year
Total_investment	Number	Total investment from to sector

3.1.2 Exploratory Data Analysis (EDA)

Data exploration is a very important step that helps to get more information about data and show more understanding of it. the df.info() was used to know more information about the dataset and the number of values in each column and their type.

Figure 1 *Data information before preprocessing*

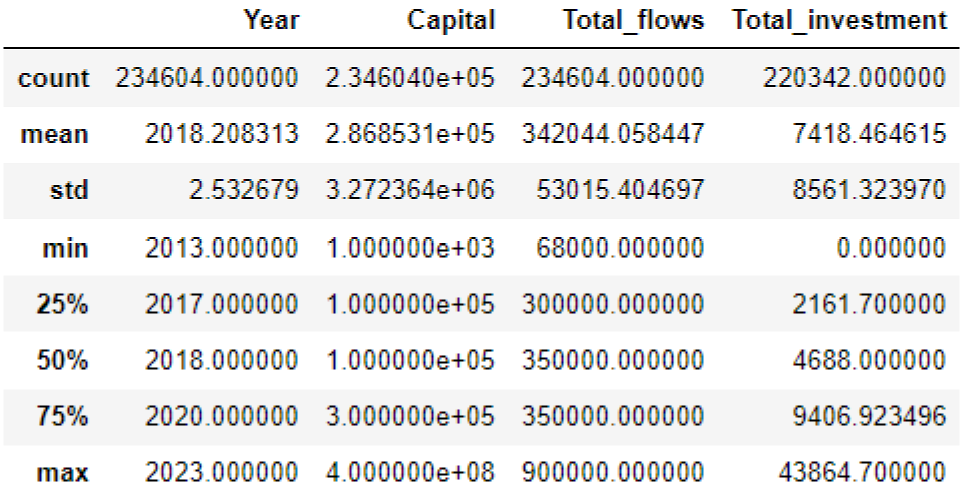


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234628 entries, 0 to 234627
Data columns (total 12 columns):
 #   Column                   Non-Null Count   Dtype  
---  ------                   --------------   -----  
 0   CompanyName              234604 non-null  object 
 1   Year                     234604 non-null  float64
 2   Investment_type          234604 non-null  object 
 3   sector                   234604 non-null  object 
 4   Sub_sector               234604 non-null  object 
 5   Activity_classification  234604 non-null  object 
 6   Legal_form               234604 non-null  object 
 7   Headquarters_location    234604 non-null  object 
 8   Capital                  234604 non-null  float64
 9   Procedure_type           234604 non-null  object 
 10  Total_flows              234604 non-null  float64
 11  Total_investment         220342 non-null  float64
dtypes: float64(4), object(8)
memory usage: 21.5+ MB
```

Statistical Summary

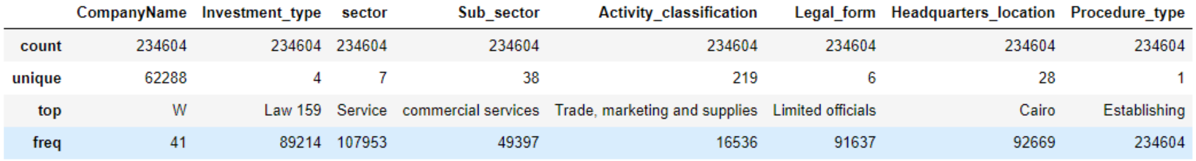
It is very important to get statistical information about data to discover patterns, the shape of the distribution, and validate assumptions. The function df.describe() was used to get the information for numeric columns and df.describe(include=['object']) function for categorical columns.

Table 3 Statistical Summary of Numerical Variables before preprocessing



	Year	Capital	Total_flows	Total_investment
count	234604.000000	2.346040e+05	234604.000000	220342.000000
mean	2018.208313	2.868531e+05	342044.058447	7418.464615
std	2.532679	3.272364e+06	53015.404697	8561.323970
min	2013.000000	1.000000e+03	68000.000000	0.000000
25%	2017.000000	1.000000e+05	300000.000000	2161.700000
50%	2018.000000	1.000000e+05	350000.000000	4688.000000
75%	2020.000000	3.000000e+05	350000.000000	9406.923496
max	2023.000000	4.000000e+08	900000.000000	43864.700000

Table 4 Statistical Summary of Categorical Variables before preprocessing

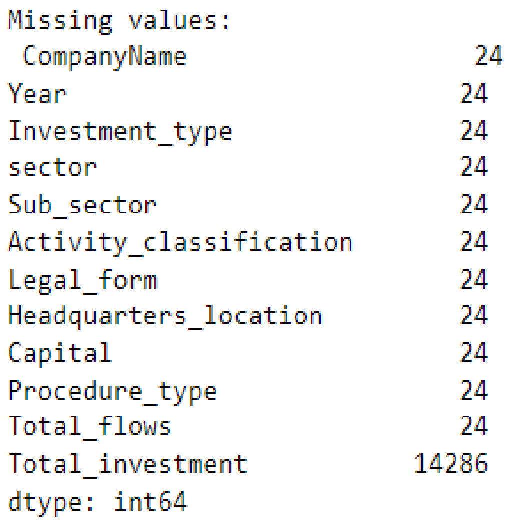


	CompanyName	Investment_type	sector	Sub_sector	Activity_classification	Legal_form	Headquarters_location	Procedure_type
count	234604	234604	234604	234604	234604	234604	234604	234604
unique	62288	4	7	38	219	6	28	1
top	W	Law 159	Service	commercial services	Trade, marketing and supplies	Limited officials	Cairo	Establishing
freq	41	89214	107953	49397	16536	91637	92669	234604

Checking missing values

Checking missing values using df.**isnull().sum() function** from **Pandas** library to check if the data contains any missing values. As a result, there are missing values in the data as shown in Figure 2.

Figure 2 *Checking missing values.*

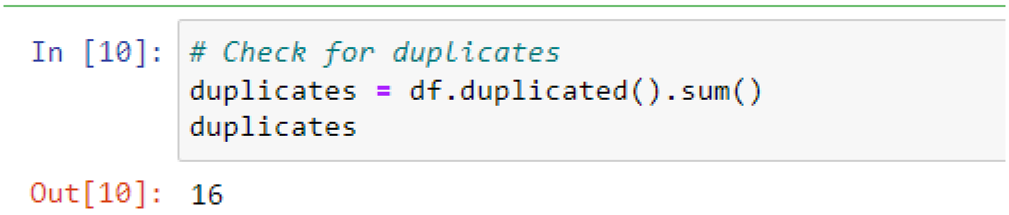


```
Missing values:
 CompanyName                    24
Year                           24
Investment_type                24
sector                         24
Sub_sector                     24
Activity_classification        24
Legal_form                     24
Headquarters_location          24
Capital                        24
Procedure_type                 24
Total_flows                    24
Total_investment            14286
dtype: int64
```

Checking duplicates

The function df.duplicated().sum() was used to check if the data contained any duplicates or not and get the number of them. As a result, 16 duplicates needed to be removed.

Figure 3 *Check duplicates.*

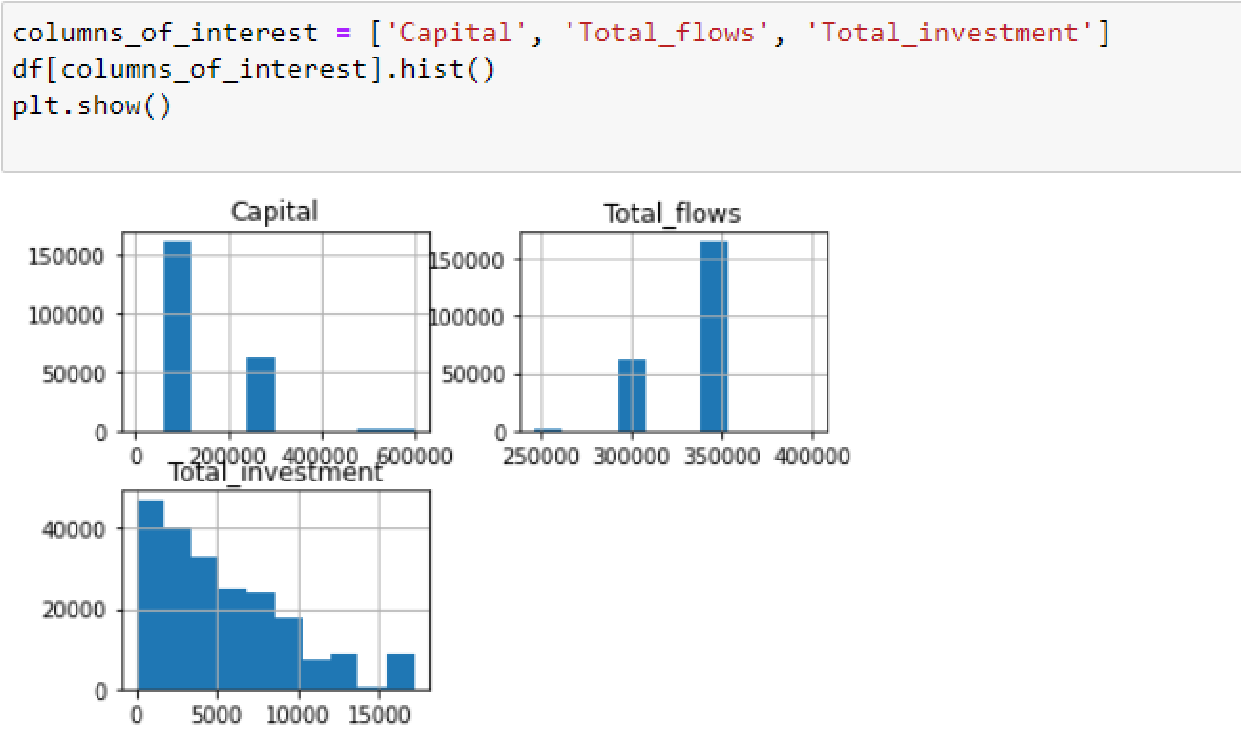


```
In [10]: # Check for duplicates
         duplicates = df.duplicated().sum()
         duplicates

Out[10]: 16
```

Data distribution

Knowing the distribution of the data is very important to discover the patterns of data. The df.hist() function was used to visualize the distribution of numerical columns by histogram. As shown in Figure 8 the data skewed in Total_flows and capital

Figure 4 *Distribution for numeric variables*



```
columns_of_interest = ['Capital', 'Total_flows', 'Total_investment']
df[columns_of_interest].hist()
plt.show()
```

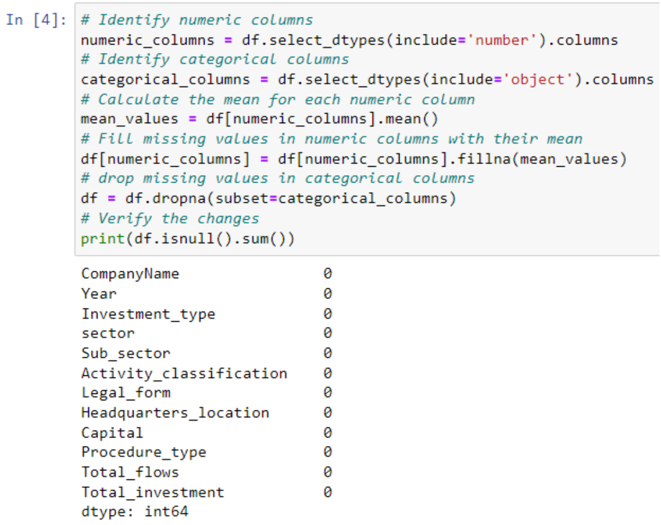
3.1.3 Data Pre-processing

Data pre-processing is the process of preparing raw data for machine learning algorithms. Data pre-processing steps for the project include handling missing values and outliers, encoding categorical variables, and feature selection.

Handling Missing Values

The function df.droupna() used to handle missing values in the categorical columns and df.fillna() to fill numeric columns with the mean as shown in Figure 5.

Figure 5 *Handle missing values*

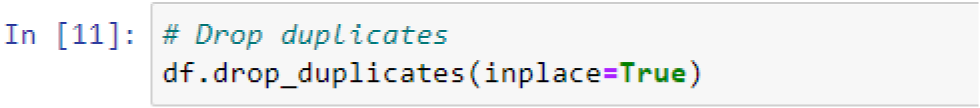


```
In [4]: # Identify numeric columns
        numeric_columns = df.select_dtypes(include='number').columns
        # Identify categorical columns
        categorical_columns = df.select_dtypes(include='object').columns
        # Calculate the mean for each numeric column
        mean_values = df[numeric_columns].mean()
        # Fill missing values in numeric columns with their mean
        df[numeric_columns] = df[numeric_columns].fillna(mean_values)
        # drop missing values in categorical columns
        df = df.dropna(subset=categorical_columns)
        # Verify the changes
        print(df.isnull().sum())

        CompanyName                0
        Year                       0
        Investment_type            0
        sector                     0
        Sub_sector                 0
        Activity_classification    0
        Legal_form                 0
        Headquarters_location      0
        Capital                    0
        Procedure_type             0
        Total_flows                0
        Total_investment           0
        dtype: int64
```

Duplicates values

The df.drop_duplicates() function was used to drop duplicates as shown in Figure 6.

Figure 6 *Drop duplicates.*

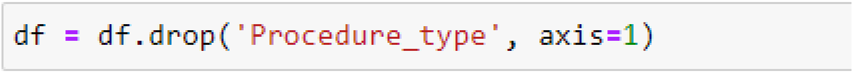


```
In [11]: # Drop duplicates
         df.drop_duplicates(inplace=True)
```

Drop unnecessary variables

The exploration step shows that there in categorical variables Procedure_type has only 1 value so it is more appropriate to drop it using df.drop() function as shown in figure 5

Figure 7 *Drop Procedure_type*



```
df = df.drop('Procedure_type', axis=1)
```

Outliers Values

To check the outliers values the function:

plt.figure(figsize=(15, 8))

for i, column in enumerate(numeric_columns, 1):

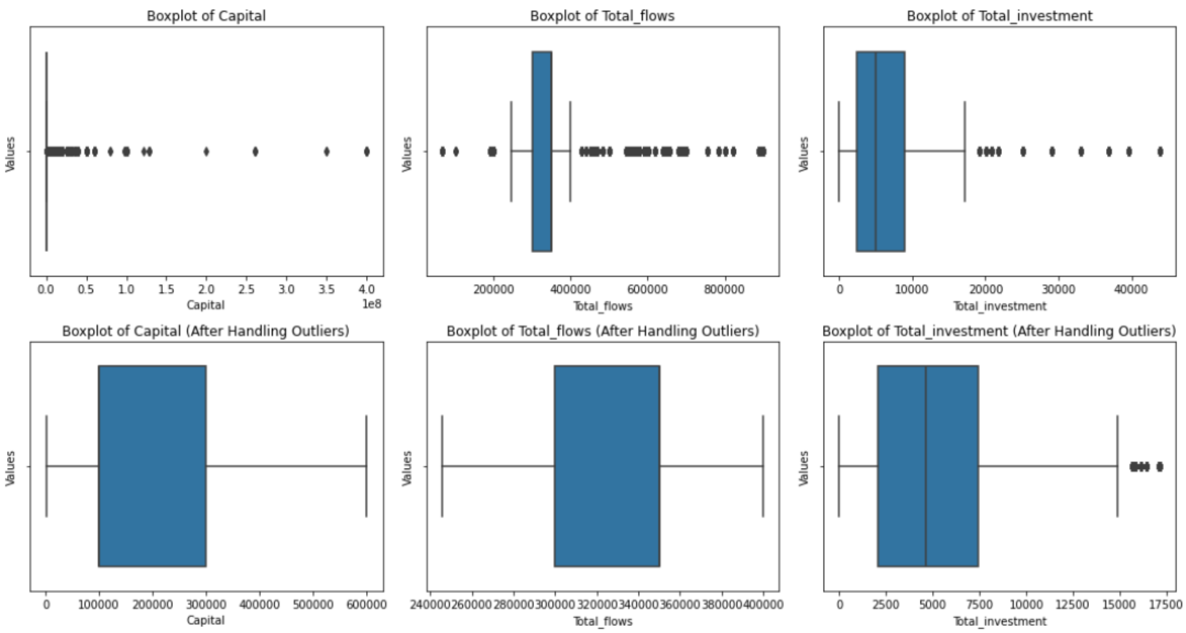
```
    plt.subplot(2, 3, i)
    sns.boxplot(x=df[column])
    plt.title(f'Boxplot of {column}')
    plt.xlabel(column)
    plt.ylabel('Values')
Q1 = df[numeric_columns].quantile(0.25)
Q3 = df[numeric_columns].quantile(0.75)
IQR = Q3 - Q1
outliers = ((df[numeric_columns] < (Q1 - 1.5 * IQR)) | (df[numeric_columns] > (Q3 + 1.5 * IQR)))
```

was used to check if there are outliers or not and identify them. As a result, there were outliers and the function

```
df[numeric_columns] = df[numeric_columns].where(~outliers, df[numeric_columns].mean(), axis=0)
for i, column in enumerate(numeric_columns, 1):
    plt.subplot(2, 3, i + 3)
    sns.boxplot(x=df[column])
    plt.title(f'Boxplot of {column} (After Handling Outliers)')
    plt.xlabel(column)
    plt.ylabel('Values')
plt.tight_layout()
plt.show()
```

used to replace them with the median.

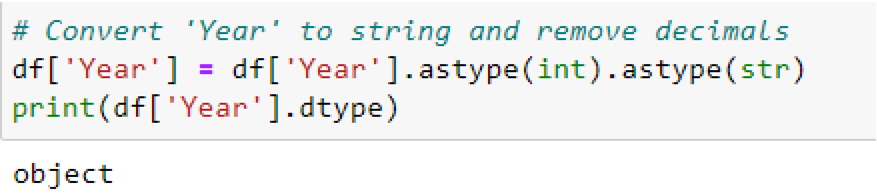
Figure 8 *Handling outliers*



Type Conversion

Specifying the correct data type for each variable is very important to get correct and accurate results. From information collected in the data exploration step, it was found that all variables are in the appropriate data type except Year variable.

Figure 9 *Year Type Conversion*



```
# Convert 'Year' to string and remove decimals
df['Year'] = df['Year'].astype(int).astype(str)
print(df['Year'].dtype)
```

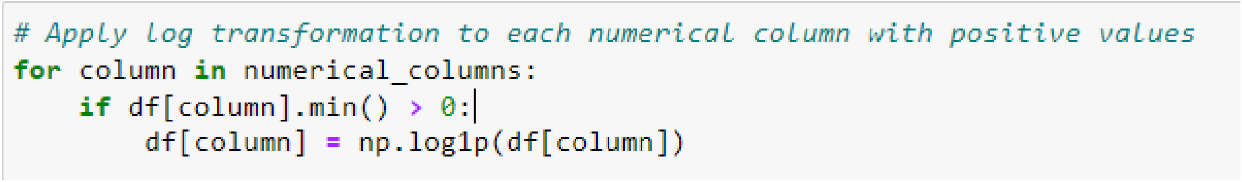
```
object
```

Scaling data

To fix the problem of skewed data and make data more balanced some trials with different techniques have been done to get the best data distribution. Those techniques are log transformation, square root transformation, cube root transformation, and Quantile Transformation

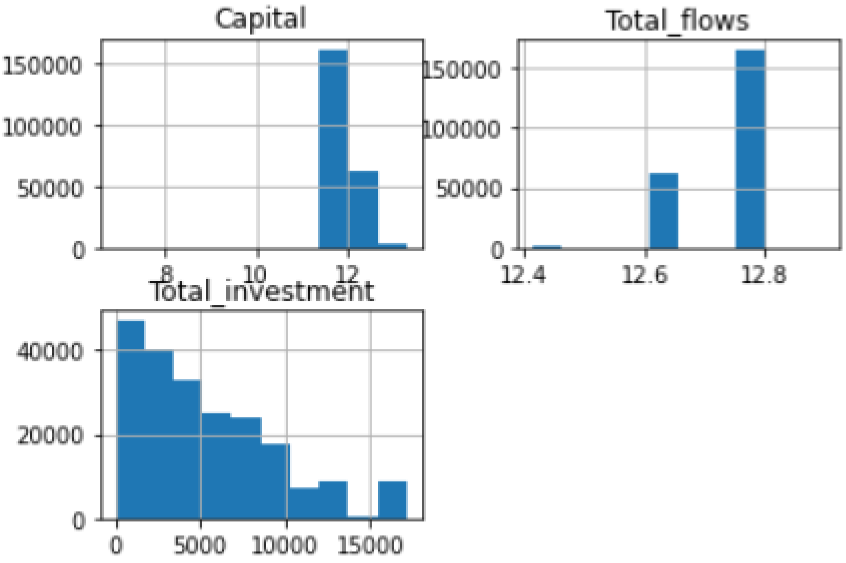
Log transformation

Figure 10 *Log transformation*



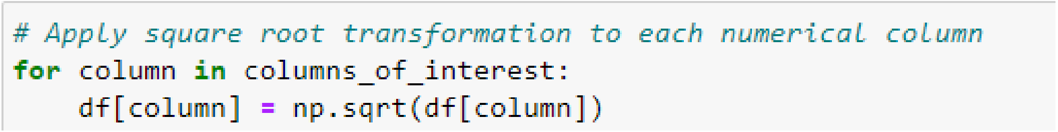
```
# Apply log transformation to each numerical column with positive values
for column in numerical_columns:
    if df[column].min() > 0:
        df[column] = np.log1p(df[column])
```

Figure 11 *Disruption after Log transformation*



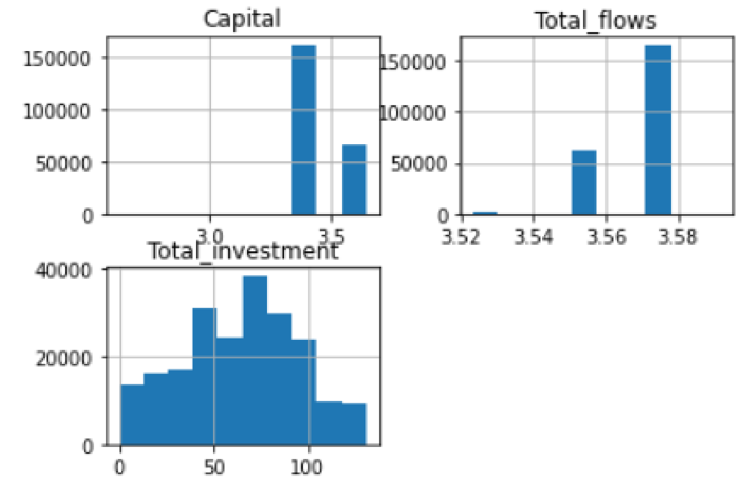
Square root transformation

Figure 12 *Square root transformation*



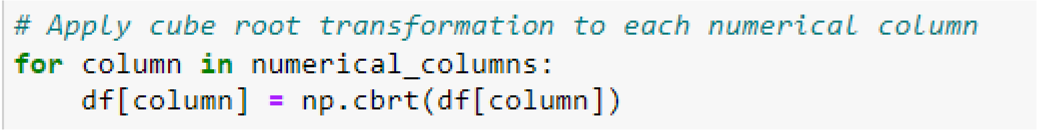
```
# Apply square root transformation to each numerical column
for column in columns_of_interest:
    df[column] = np.sqrt(df[column])
```

Figure 13 *Distribution after square root transformation*



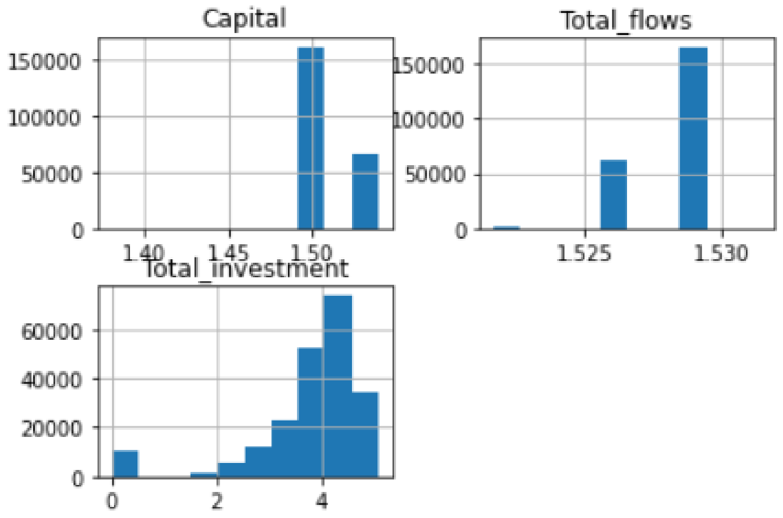
Cube root transformation

Figure 14 *Cube root transformation*



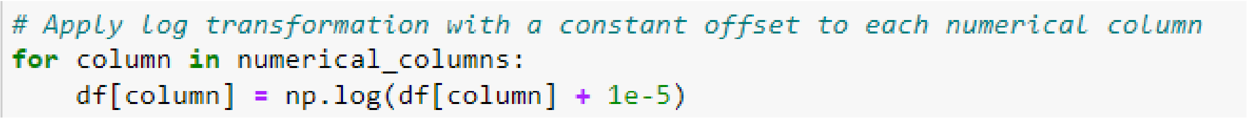
```
# Apply cube root transformation to each numerical column
for column in numerical_columns:
    df[column] = np.cbrt(df[column])
```

Figure 15 *Distribution after cube root transformation*



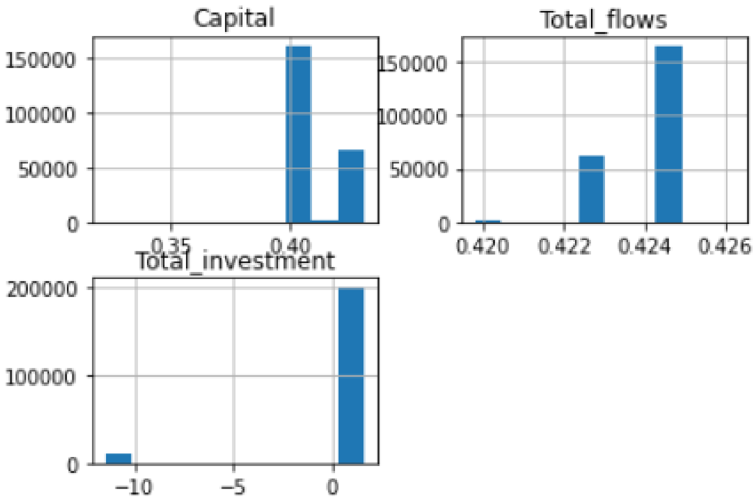
Log transformation with a constant offset

Figure 16 *Log transformation with a constant offset*



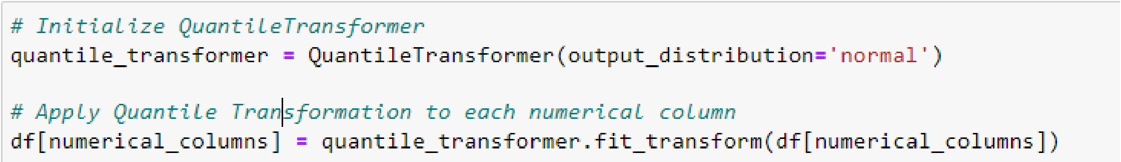
```
# Apply log transformation with a constant offset to each numerical column
for column in numerical_columns:
    df[column] = np.log(df[column] + 1e-5)
```

Figure 17*Distribution after Log transformation with a constant offset*



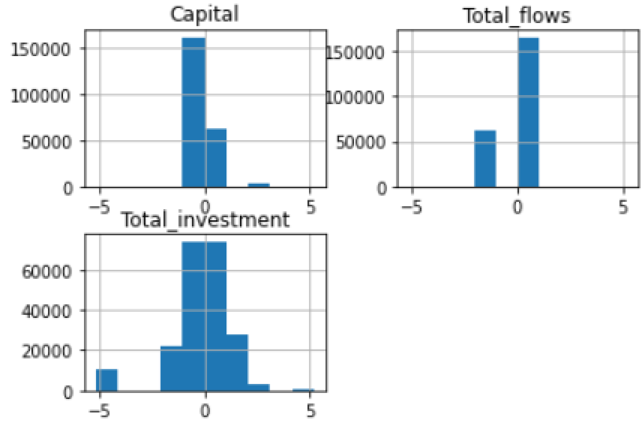
Quantile Transformation

Figure 18 *Quantile Transformation*



```
# Initialize QuantileTransformer
quantile_transformer = QuantileTransformer(output_distribution='normal')

# Apply Quantile Transformation to each numerical column
df[numerical_columns] = quantile_transformer.fit_transform(df[numerical_columns])
```

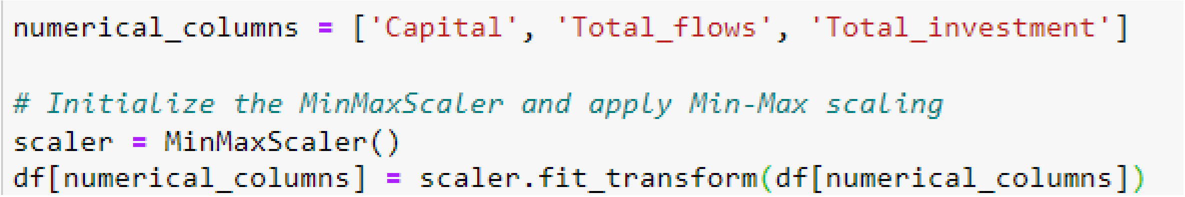
Figure 19 *Distribution after Quantile Transformation*



Scaling data

Scaling data is an important step for ensuring that the model will do well also Scaling helps to balance the effect of regularization across all features. One of the popular scaling techniques is MinMaxScaler. This scaling method is useful when you want to normalize the range of numerical features, ensuring that all features contribute equally to the model training process. MinMaxScaler() function was used to scale the data

Figure 20 *Scaling data using MinMaxScaler*

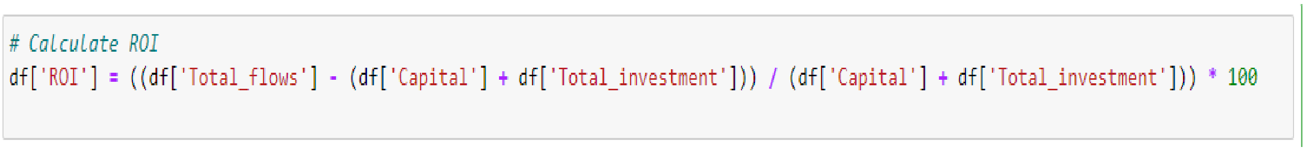


```
numerical_columns = ['Capital', 'Total_flows', 'Total_investment']

# Initialize the MinMaxScaler and apply Min-Max scaling
scaler = MinMaxScaler()
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

Feature engineering

Feature engineering can help the model better understand the underlying patterns in the data. In this project, the generated variable is a return on investment (ROI) created from Total_flows, Capital, and Total_investment.

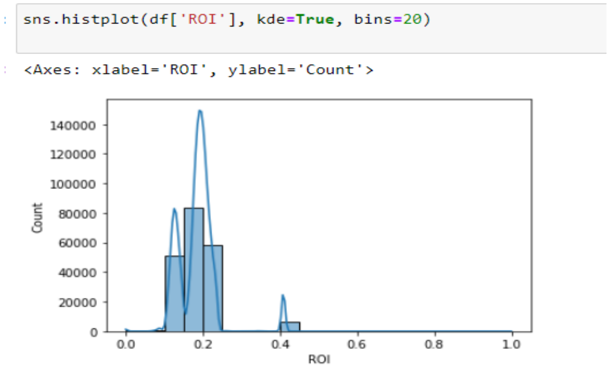
Figure 21 *ROI Calculation*



```
# Calculate ROI
df['ROI'] = ((df['Total_flows'] - (df['Capital'] + df['Total_investment'])) / (df['Capital'] + df['Total_investment'])) * 100
```

ROI distribution

Figure 22 *ROI distribution*

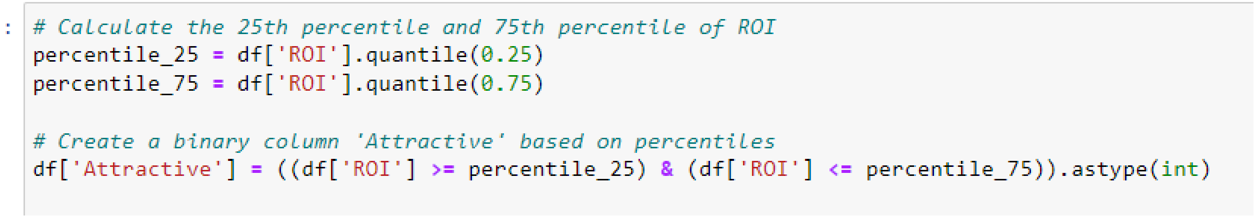


```
sns.histplot(df['ROI'], kde=True, bins=20)
```

```
<Axes: xlabel='ROI', ylabel='Count'>
```

Create a binary variable from ROI as a target variable 1 if the investment attracts investors and 0 if not.

Figure 23 *Create target variable*



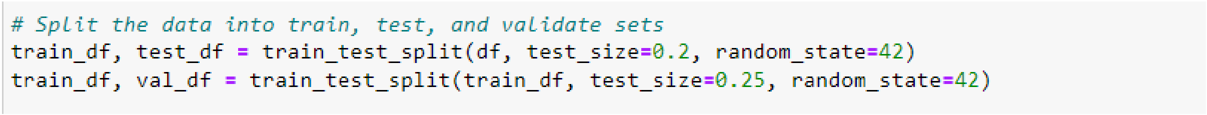
```
# Calculate the 25th percentile and 75th percentile of ROI
percentile_25 = df['ROI'].quantile(0.25)
percentile_75 = df['ROI'].quantile(0.75)

# Create a binary column 'Attractive' based on percentiles
df['Attractive'] = ((df['ROI'] >= percentile_25) & (df['ROI'] <= percentile_75)).astype(int)
```

3.1.4 Data Modeling

Data splitting

In this project, the data is split into train, validation, and test sets using train_test_split() function to ensure that the model learns well.

Figure 24 *Data splitting*

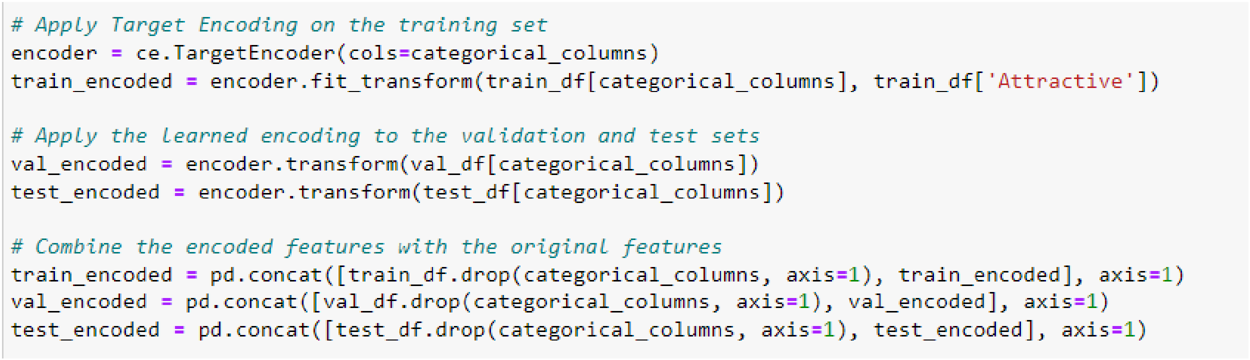


```
# Split the data into train, test, and validate sets
train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)
train_df, val_df = train_test_split(train_df, test_size=0.25, random_state=42)
```

Target encoding for categorical variables

Target encoding is a powerful tool for capturing valuable information from categorical variables in a way that aligns with the target variable's characteristics.

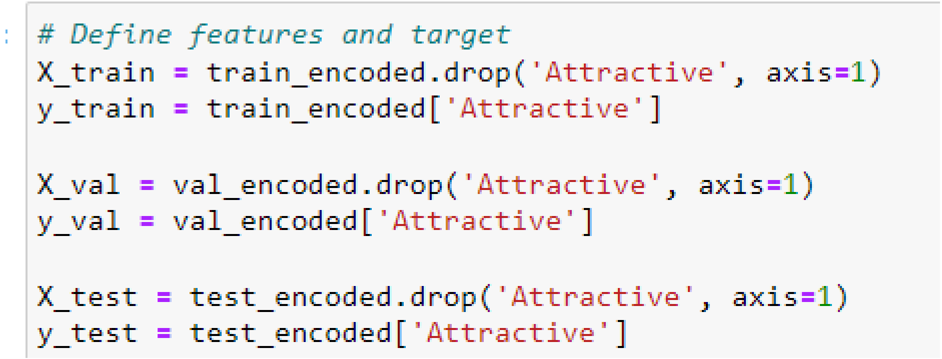
Figure 25 *Target encoding*



```
# Apply Target Encoding on the training set
encoder = ce.TargetEncoder(cols=categorical_columns)
train_encoded = encoder.fit_transform(train_df[categorical_columns], train_df['Attractive'])

# Apply the learned encoding to the validation and test sets
val_encoded = encoder.transform(val_df[categorical_columns])
test_encoded = encoder.transform(test_df[categorical_columns])

# Combine the encoded features with the original features
train_encoded = pd.concat([train_df.drop(categorical_columns, axis=1), train_encoded], axis=1)
val_encoded = pd.concat([val_df.drop(categorical_columns, axis=1), val_encoded], axis=1)
test_encoded = pd.concat([test_df.drop(categorical_columns, axis=1), test_encoded], axis=1)
```

Define Target and features variables

Figure 26 *Define Target and features variables*

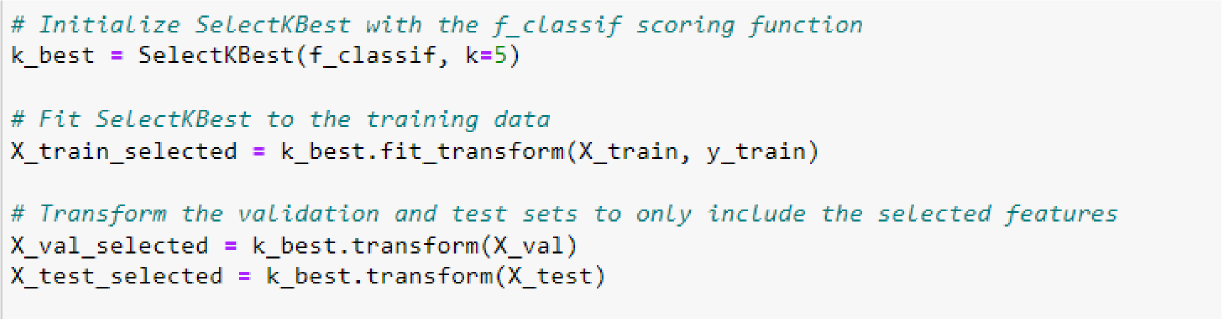


```
# Define features and target
X_train = train_encoded.drop('Attractive', axis=1)
y_train = train_encoded['Attractive']

X_val = val_encoded.drop('Attractive', axis=1)
y_val = val_encoded['Attractive']

X_test = test_encoded.drop('Attractive', axis=1)
y_test = test_encoded['Attractive']
```

Feature selection

Feature selection is important before the analysis to make sure that the model doing well. the function SelectKBest() was used to select the features.

Figure 27 *Feature selection*



```
# Initialize SelectKBest with the f_classif scoring function
k_best = SelectKBest(f_classif, k=5)

# Fit SelectKBest to the training data
X_train_selected = k_best.fit_transform(X_train, y_train)

# Transform the validation and test sets to only include the selected features
X_val_selected = k_best.transform(X_val)
X_test_selected = k_best.transform(X_test)
```

After finishing data preprocessing data now is ready for Modeling. In this project There are four models performed in this section using Logistic Regression, Support vector machine classifier (SVC) method, the Neural Network (Multi-Layer Perceptron) model, and Naive Bayes model.

The machine learning algorithms pseudocode that is applied at each algorithm is as follows:

Install the Library

Import Data

Manage Null Value

Drop duplicates

Handel outliers

Scale the data

Train &Test and validate the Data

Target Encode the Data

Feature Selection

Indicate the performance of the confusion matrix.

3.2 Stage 2: What affects investment decisions.

3.2.1 Data collection

The data was collected from World Bank, OECD, IMF, and Egypt Central Bank with 23 records with 17 independent variables and 2 dependent variables which are FDI, and domestic_investment. The data is collected manually Using Microsoft Excel. The descriptions of each variable are in the below table.

Table 5 *Data Table for what affect investment*

Name of the variable	**Description**
Foreign_direct_investment_net_inflows (% of GDP)	Total foreign direct investment enters Egypt
domestic_investment	Total national investment
GDP_growth (annual %)	Gross domestic products
Inflation_Rate	Inflation rate over the years
exchange_rate	Currency exchange rate
Political_Stability_Percentile_Rank	Rank for political stability and zero terrorists
Control_of_Corruption_Percentile_Rank	The rank of Egypt for controlling corruption
LPI_Rank	The logistics index represents the infrastructure index
Government_Effectiveness_Percentile_Rank	The rank of Egypt if the government doing well and issued effective laws
Ease_of_doing_business_score	How is it easy to do business in Egypt
FDI_restrictiveness	The restrictiveness by the government in FDI
Population	Egypt's population over the years
Rule_of_law	The score of how low is applicable in Egypt
Export_volume_index	Export volume out of Egypt
Import_volume_index	Import volume enters Egypt
Tariff_rate_weighted_mean	The rate decided by the central bank in Egypt for import or export to and from Egypt
PPP	Purchasing power parity for consumers

Time_required_to_start_a_business(days)	Number of days need to finish papers and contracts in Egypt
Year	Number of years from 2000 to 2023

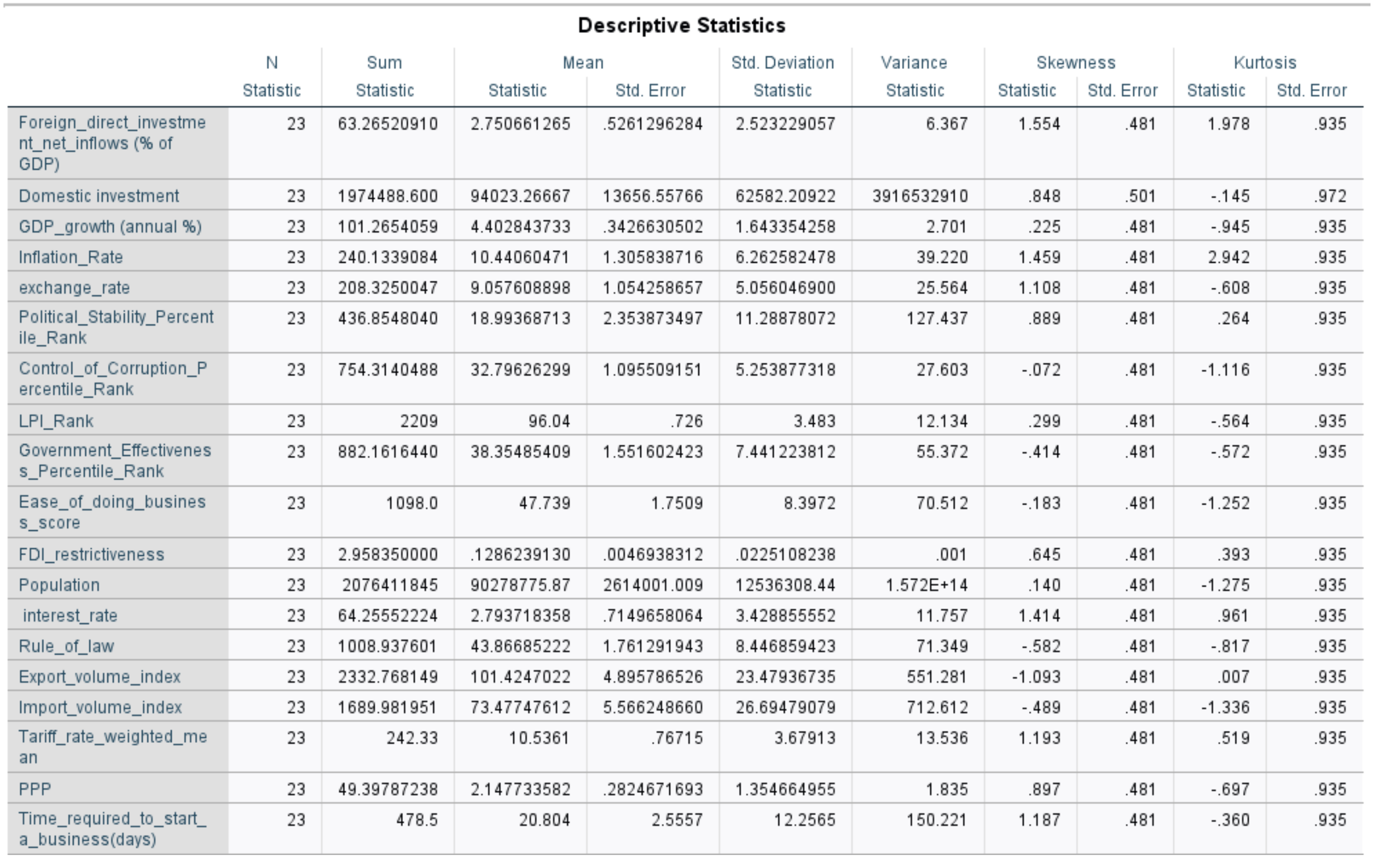
3.2.2 Data exploration and preprocessing

As the data is collected manually, the data is cleaned with no missing or duplicate values, and the data is ready to measure how independent variables affect dependent variables.

Descriptive Analysis

Descriptive analysis is a crucial step in exploring and summarizing the main features of a dataset. It provides a concise and meaningful overview of the key characteristics, tendencies, and patterns within the data. The descriptive statistics table is shown below.

Table 6 *Descriptive statistics table*



Descriptive Statistics

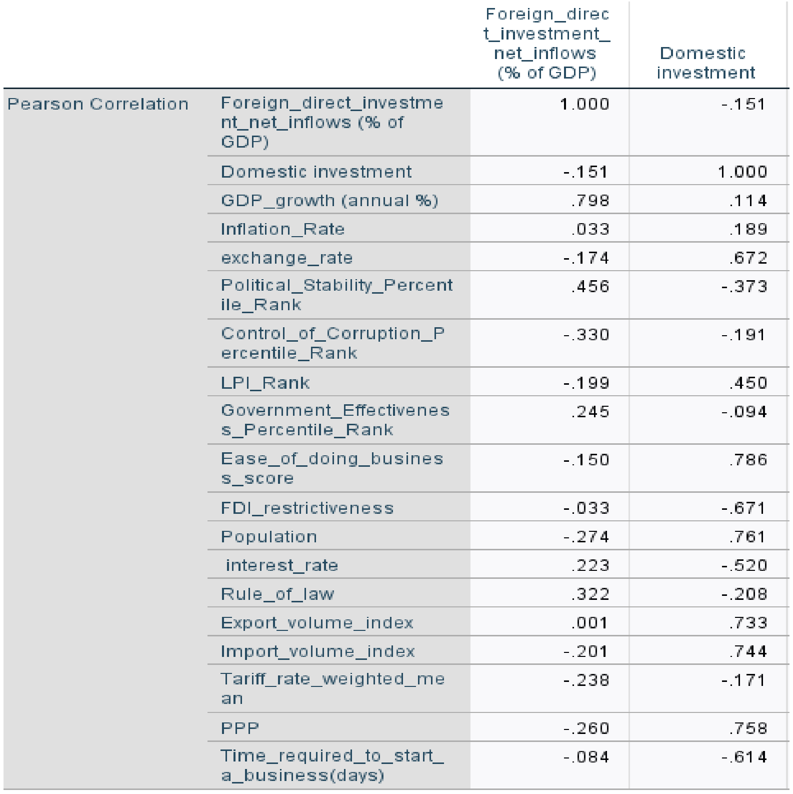
	N	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Foreign_direct_investment_net_inflows (% of GDP)	23	63.26520910	2.750661265	.5261296284	2.523229057	6.367	1.554	.481	1.978	.935
Domestic investment	23	1974488.600	94023.26667	13656.55766	62582.20922	3916532910	.848	.501	-.145	.972
GDP_growth (annual %)	23	101.2654059	4.402843733	.3426630502	1.643354258	2.701	.225	.481	-.945	.935
Inflation_Rate	23	240.1339084	10.44060471	1.305838716	6.262582478	39.220	1.459	.481	2.942	.935
exchange_rate	23	208.3250047	9.057608898	1.054258657	5.056046900	25.564	1.108	.481	-.608	.935
Political_Stability_Percentile_Rank	23	436.8548040	18.99368713	2.353873497	11.28878072	127.437	.889	.481	.264	.935
Control_of_Corruption_Percentile_Rank	23	754.3140488	32.79626299	1.095509151	5.253877318	27.603	-.072	.481	-1.116	.935
LPI_Rank	23	2209	96.04	.726	3.483	12.134	.299	.481	-.564	.935
Government_Effectiveness_Percentile_Rank	23	882.1616440	38.35485409	1.551602423	7.441223812	55.372	-.414	.481	-.572	.935
Ease_of_doing_business_score	23	1098.0	47.739	1.7509	8.3972	70.512	-.183	.481	-1.252	.935
FDI_restrictiveness	23	2.958350000	.1286239130	.0046938312	.0225108238	.001	.645	.481	.393	.935
Population	23	2076411845	90278775.87	2614001.009	12536308.44	1.572E+14	.140	.481	-1.275	.935
interest_rate	23	64.25552224	2.793718358	.7149658064	3.428855552	11.757	1.414	.481	.961	.935
Rule_of_law	23	1008.937601	43.86685222	1.761291943	8.446859423	71.349	-.582	.481	-.817	.935
Export_volume_index	23	2332.768149	101.4247022	4.895786526	23.47936735	551.281	-1.093	.481	.007	.935
Import_volume_index	23	1689.981951	73.47747612	5.566248660	26.69479079	712.612	-.489	.481	-1.336	.935
Tariff_rate_weighted_mean	23	242.33	10.5361	.76715	3.67913	13.536	1.193	.481	.519	.935
PPP	23	49.39787238	2.147733582	.2824671693	1.354664955	1.835	.897	.481	-.697	.935
Time_required_to_start_a_business(days)	23	478.5	20.804	2.5557	12.2565	150.221	1.187	.481	-.360	.935

3.2.3 Correlation analysis

The importance of correlation analysis lies in its ability to detect the relations and provide valuable insights into the patterns and associations within a dataset.

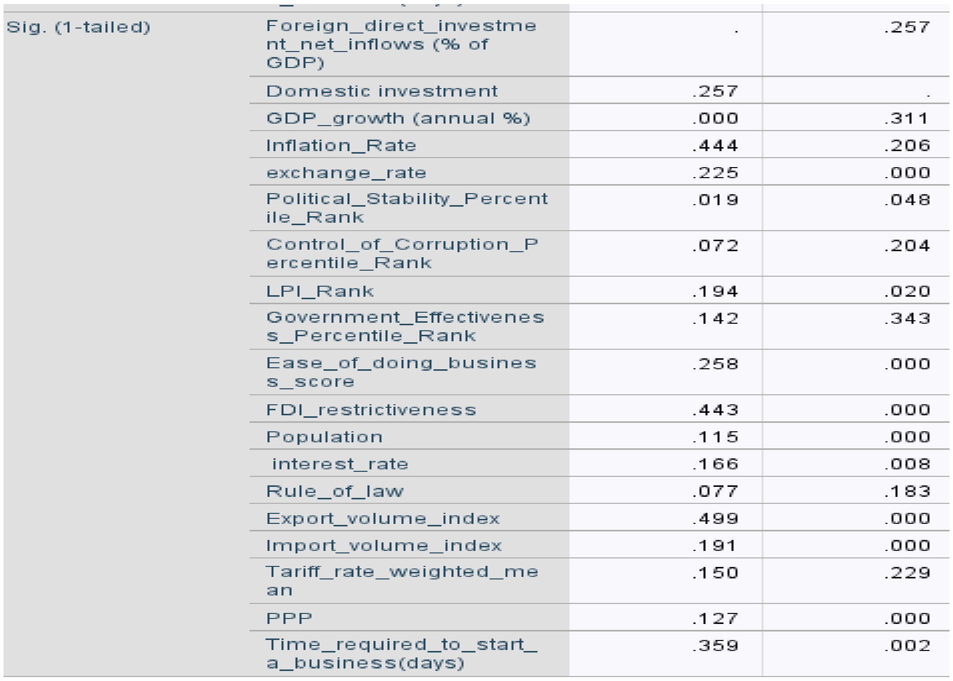
The Pearson correlation is a parametric measure for evaluating the statistical evidence for a linear relationship. The range of values is -1.0 to 1.0. A perfect negative correlation is represented by a correlation of -1.0, whereas a perfect positive correlation is represented by a correlation of 1.0.

Table 7 *The Pearson correlation*



		Foreign_direct_investment_net_inflows (% of GDP)	Domestic investment
Pearson Correlation	Foreign_direct_investment_net_inflows (% of GDP)	1.000	-.151
	Domestic investment	-.151	1.000
	GDP_growth (annual %)	.798	.114
	Inflation_Rate	.033	.189
	exchange_rate	-.174	.672
	Political_Stability_Percentile_Rank	.456	-.373
	Control_of_Corruption_Percentile_Rank	-.330	-.191
	LPI_Rank	-.199	.450
	Government_Effectiveness_Percentile_Rank	.245	-.094
	Ease_of_doing_business_score	-.150	.786
	FDI_restrictiveness	-.033	-.671
	Population	-.274	.761
	interest_rate	.223	-.520
	Rule_of_law	.322	-.208
	Export_volume_index	.001	.733
	Import_volume_index	-.201	.744
	Tariff_rate_weighted_mean	-.238	-.171
	PPP	-.260	.758
	Time_required_to_start_a_business(days)	-.084	-.614

Table 8 *Sig correlation*



Sig. (1-tailed)	Foreign_direct_investment_net_inflows (% of GDP)	.	.257
	Domestic investment	.257	.
	GDP_growth (annual %)	.000	.311
	Inflation_Rate	.444	.206
	exchange_rate	.225	.000
	Political_Stability_Percentile_Rank	.019	.048
	Control_of_Corruption_Percentile_Rank	.072	.204
	LPI_Rank	.194	.020
	Government_Effectiveness_Percentile_Rank	.142	.343
	Ease_of_doing_business_score	.258	.000
	FDI_restrictiveness	.443	.000
	Population	.115	.000
	interest_rate	.166	.008
	Rule_of_law	.077	.183
	Export_volume_index	.499	.000
	Import_volume_index	.191	.000
	Tariff_rate_weighted_mean	.150	.229
	PPP	.127	.000
	Time_required_to_start_a_business(days)	.359	.002

Chapter 4

4.1 Results

This chapter explains, discusses, evaluates, and visualizes the results in the previous chapter.

4.1.1 Model evaluating

To measure the performance of the classifiers, four performance metrics were used: Accuracy, and F1 score. The confusion matrix is deployed as shown below.

Table 9 *Confusion Matrix*

	Attractive (1)	**Not attractive (0)**
Attractive (1)	True Positive (TP)	False Positive (FP)
Not attractive (0)	False Negative (FN)	True Negative (TN)

Where:

True Positive (TP): The Observation is Positive, and the model classified it as Positive.

False Negative (FN): The Observation is Positive, but the model classified it as Negative.

True Negative (TN): The Observation is Negative, and the model classified it as Negative.

False Positive (FP): The Observation is Negative, but the model classified it as Positive.

Accuracy:

Accuracy is the ratio of correct prediction the machine learning achieves in the dataset.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \qquad (1)$$

F1 Score

Precision and recall are the two main components of the F1 score. The F1 score aims to mix the precision and recall measures into a single rating.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

$F1 = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$ (4)

4.1.2 Accuracy score

Table 10 *Accuracy score*

Model	Accuracy on validation	Accuracy on test
the Neural Network (Multi-Layer Perceptron)	**0.98**	**0.979**
Logistic Regression	0.822	0.824
Naïve Bayes	0.77	0.78
Support Vector Classifier (SVC)	0.9739	0.9733

4.1.3 F1 Score

Table 11 *Models results*

Model		**precision**	**recall**	**F1 Score**
The Neural Network (Multi-Layer Perceptron)	**0**	**0.98**	**0.98**	**0.98**
	1	**0.98**	**0.98**	**0.98**
Logistic Regression	0	0.83	0.81	0.82
	1	0.82	0.84	0.83
Naive Bayes	0	0.84	0.68	0.76
	1	0.74	0.88	0.80
Support Vector Classifier (SVC)	0	0.98	0.97	0.97
	1	0.97	0.98	0.97

4.1.4 Confusion matrix

Table 12 *Confusion matrix*

	Attractive (1)	**Not attractive (0)**
Attractive (1)	19225	421
Not attractive (0)	318	20012

4.1.5 Hyperparameter tuning results in the champion model

Table 13 *Hyperparameter tuning results*

Champion model		**precision**	**recall**	**F1 Score**	**Accuracy on validation**	**Accuracy on test**
the Neural Network (Multi-Layer Perceptron)	0	0.98	0.98	0.98	0.9815	0.9813
	1	0.98	0.98	0.98		

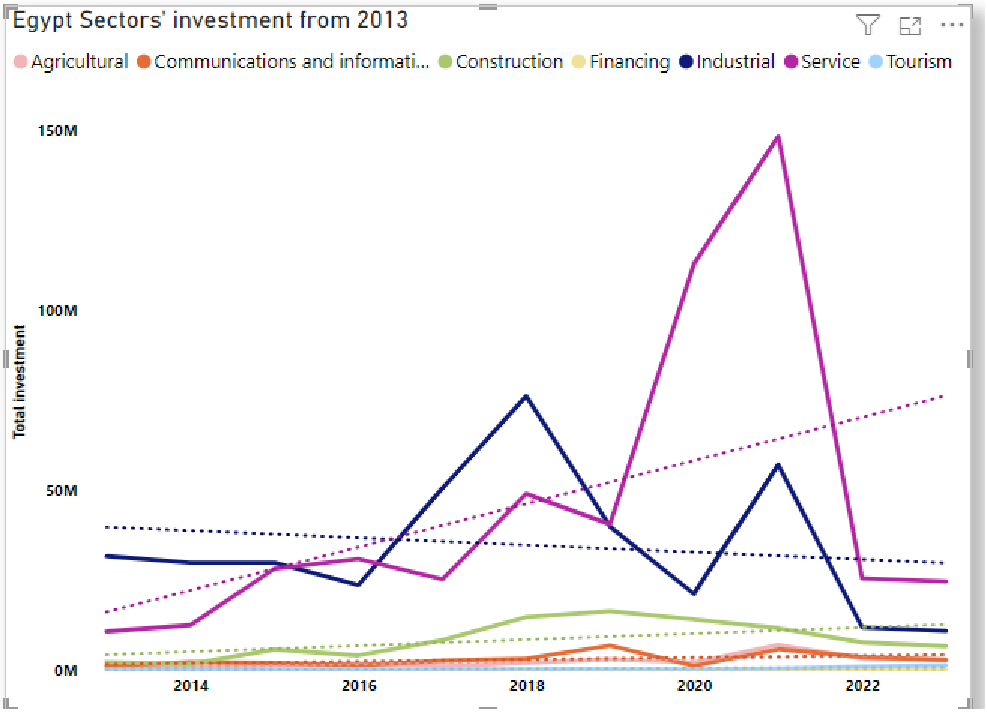
4.1.6 Results from correlation analysis

- There is a strong positive relation between foreign direct investment and GDP growth which represents economic growth.
- There is a strong positive relationship between domestic investment and ease of doing business score, population, export volume index, import volume index, PPP, and exchange rate.
- There is a strong negative relationship between domestic investment time required to start a business and FDI_restrictiveness.
- The political stability index is a significant variable in both domestic investment and foreign direct investment.

4.2 Visualizations and Discussion

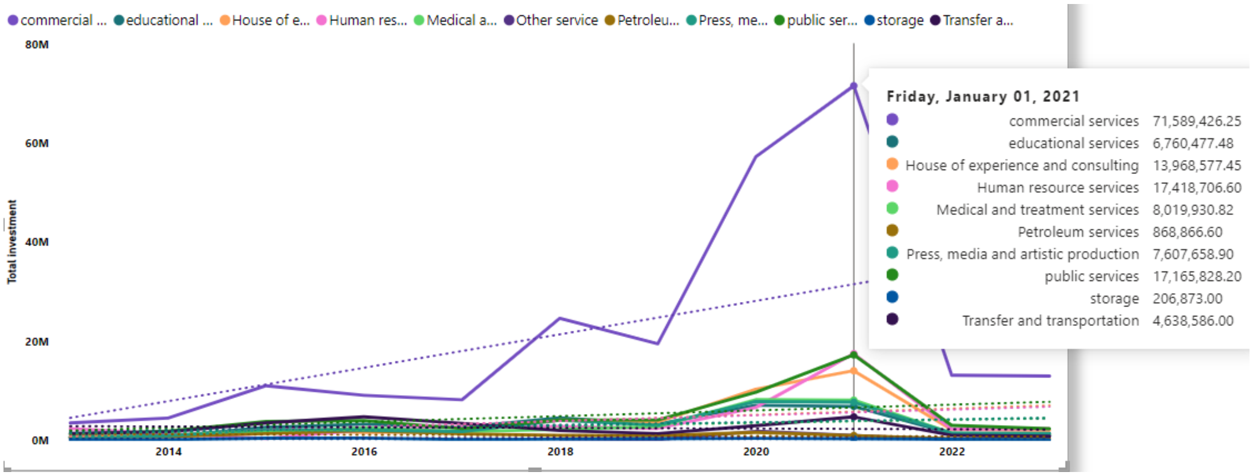
Total investment in Egypt over the years

Figure 28 *Total investment for each sector*



The graph shows the sector most invested in is the service sector. The investment in service sector started to increase in 2019 reaching the highest value in 2021. The reason for this is during the COVID-19 crisis and quarantine restrictions people stayed home and online trading and services increased dramatically. As shown in Figure 29 the highest value of total investments in subsectors in the service sector is commercial services.

Figure 29 *Total investments in the service sector*



Inflation and political stability affect Domestic investment and FDI

Figure 30 *High Inflation rate and low political stability index affect Domestic investment and FDI*

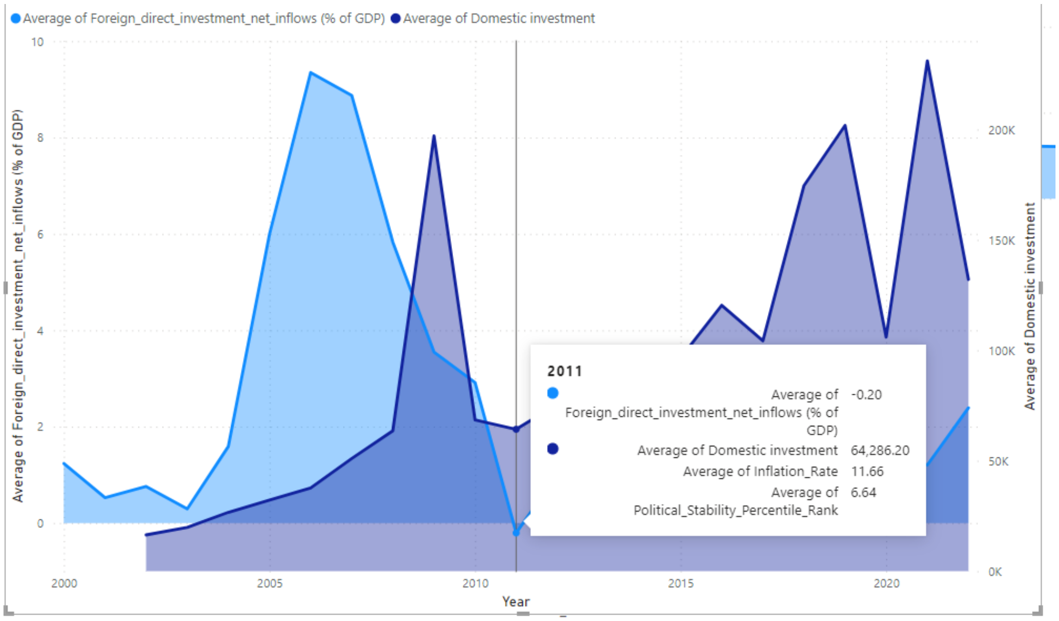
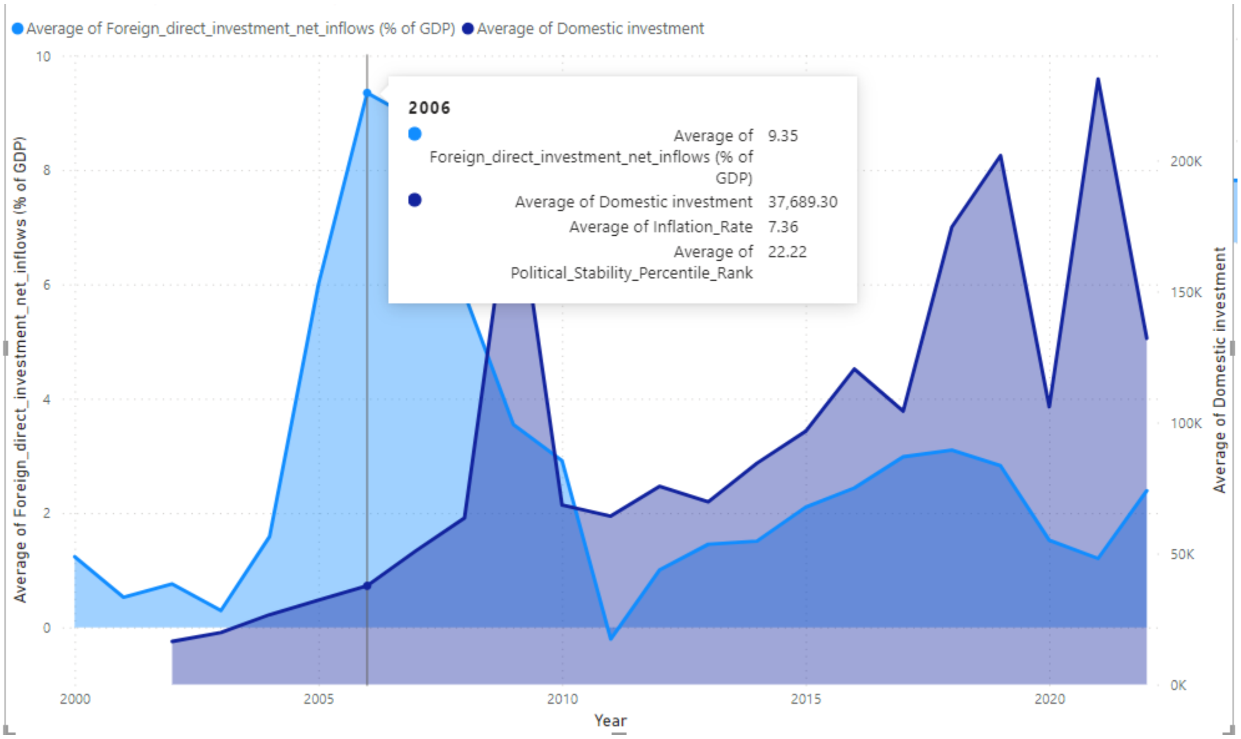


Figure 31 *Low Inflation rate and high political stability index affect Domestic investment and FDI*



Figures 30 and 31 show the effect of political stability and inflation rate on both domestic and foreign investments. In 2011 because of the revolution that happened at this time political stability decreased and the inflation rate increased so both investments decreased sharply even foreign investors withdrew their investments causing negative FDI.

Chapter 5

This chapter aims to illustrate the limitation, give recommendation to GAFI and future researchers, and summarize what discussed in the paper.

5.1 Limitations of the study

- Data Limitation: The study makes use of information from other sources as well as Egypt's General Authority for Investment and Free Zones (GAFI). The dependability of these sources affects the data's completeness and correctness.
- Translation Problems: Because the source dataset was written in Arabic, there may be translation mistakes made during the Google Spreadsheets translation process. The analysis's accuracy may be impacted by these mistakes.
- Assumption of Linearity: A linear relationship between features and the target variable is assumed by the models, which include support vector machines and logistic regression. Not every situation in the real world will fit this assumption.
- Generalizability: The models and findings drawn from this research are particular to the Egyptian dataset and might not be readily transferable to other geographical areas or economic situations.
- External Factors: Outside forces can have an impact on political and economic environments. Investment decisions may be impacted by unanticipated events or policy changes that were not taken into account in the study.

5.2 Recommendations

1. Increase Data Transparency: Increase the data's transparency about investments. Investor confidence can be increased by timely and reliable information on economic indicators, trends particular to a given industry, and investment prospects.
2. Invest in Political Stability: Put laws and other measures into place to improve political stability. A stable political climate is likely to attract more investors to a place. An environment that is favorable to investment can be created by exhibiting a commitment to political stability.
3. Encourage economic diversification to lessen reliance on certain industries. The study focuses on relationships that exist between investment attractiveness and specific economic indices. A more varied economy can draw in a wider pool of investors and reduce risks.

4. Simplify corporate processes: examine and simplify corporate processes, particularly in industries where complex processes could turn away investors. Streamlining the startup and operation processes of a company can draw in both local and foreign investors.
5. Handle Inflationary Fears: Put policies in place to handle worries about inflation. Elevated rates of inflation have been recognized as a factor that adversely impacts investments, both foreign and domestic. A more hospitable environment for investment can be achieved through targeted inflation control strategies.
6. An environment favorable to investors: Make the law more investor friendly. For investor confidence, a fair, open, and effective judicial system must be maintained. An investment is perceived more favorably when legal procedures are transparent and predictable.
7. Participate in partnerships between the public and private sectors (PPPs): Encourage and assist in PPPs. Working together, the public and private sectors can provide profitable investment opportunities that promote economic expansion.
8. Constant Monitoring and Adjustment: Set up systems for keeping an eye on market developments and economic data at all times. Adaptability and responsiveness to investor demands will be demonstrated by revising policies regularly in response to changing economic conditions.
9. Financial Policies That Encourage Investment: Put in place financial policies that encourage investment. Offering financial support systems, tax benefits, and incentives can draw in investors and promote capital preservation.
10. International Cooperation: Promote international cooperation and demonstrate your dedication to a world economy that is integrated. Egypt can become a more desirable location for investments if it participates in international organizations and follows best practices from around the world.
11. Investment Promotion Initiatives: Start focusing campaigns to promote investments. Make marketing efforts to draw attention to Egypt's advantages as an investment destination, including its economic strengths, policy improvements, and investment prospects.
12. Public Relations Approach: Create an efficient and transparent communication plan. Perceptions can be positively impacted by informing the public and investors about economic strategy, government policies, and investment prospects.

Recommendations for Future work

- Improved Data Collection: To guarantee more precise and thorough datasets, future research should concentrate on enhancing data quality, maybe in cooperation with pertinent agencies.
- Cross-validation: To evaluate the models' resilience and generalizability in various economic situations and validate them using data from various nations or regions.
- Constant Monitoring: Constant monitoring and model upgrades would improve the forecasting accuracy of the models over time, as investment landscapes and economic situations are dynamic.
- Include Qualitative Data: To give a more thorough insight into the elements impacting investment decisions, take into consideration expert viewpoints and qualitative data.
- Conduct scenario analysis to evaluate the models' performance in various political or economic scenarios. This will give stakeholders a better understanding of the opportunities and hazards that may arise.

5.3 Conclusion

In summary, the purpose of this study was to evaluate and forecast Egypt's investment attractiveness. Using thorough data analysis, preprocessing, and the utilization of machine learning models, significant understandings regarding the variables impacting investment choices were obtained. The study emphasizes the significance of sector-specific developments, political stability, and economic indicators.

It is imperative to recognize the limits of the study, including the possibility of data flaws and the dependence on past patterns. The results should be interpreted cautiously because economic situations are inherently changing.

Notwithstanding these drawbacks, the models that have been constructed exhibit encouraging accuracy and serve as a basis for further study and decision-making. Models will need to be continuously improved upon and adjusted as economic environments change to remain relevant.

References

Adebayo, T., Akinsola, G., Olanrewaju, V., & Abolaji, A. (2020). Does Inflation Asymmetrically Affect Foreign Direct Investment in an Emerging Market? An Application of the Non-Linear Autoregressive Distributed Lag (Nardl) Model. *EuroEconomica*, *39*(3).

Ahmed, W. M. (2017). The impact of political regime changes on stock prices: The case of Egypt. *International Journal of Emerging Markets*, *12*(3), 508–531.

Alam Iqbal, B., Sami, S., & Turay, A. (2019). Determinants of China's outward foreign direct investment in Asia: A panel data analysis. *Economic and Political Studies*, *7*(1), 66–86.

Alikhanov, A., & Khudiyev, N. (2020). TESTING OF FDI AND NON-OIL FDI INFLOWS IN AZERBAIJAN USING DUNNING'S ECLECTIC MODEL. *Economic and Social Development: Book of Proceedings*, *3*, 620–626.

Almfraji, M. A., & Almsafir, M. K. (2014). Foreign direct investment and economic growth literature review from 1994 to 2012. *Procedia-Social and Behavioral Sciences*, *129*, 206–213.

Amal, M., Tomio, B. T., & Raboch, H. (2010). Determinants of foreign direct investment in Latin America. *GCG: Revista de Globalización, Competitividad y Gobernabilidad*, *4*(3), 116–133.

Andreas, R., & Carl, P. (2021). *Capital flows during times of crises: A study of 21st century economic crises and their impact on FDI-flows*.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, *7*, 124233–124243.

Asbullah, M. H., Shaari, M. S., Zainol, N., & Abidin, S. (2022). Determinants of foreign direct investment (FDI). *Sciences*, *11*(3), 213–232.

Asiamah, M., Ofori, D., & Afful, J. (2019). Analysis of the determinants of foreign direct investment in Ghana. *Journal of Asian Business and Economic Studies*, *26*(1), 56–75.

Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *European Financial Management*, *28*(3), 744–770.

Bailey, N. (2018a). Exploring the relationship between institutional factors and FDI attractiveness: A meta-analytic review. *International Business Review*, *27*(1), 139–148.

Bailey, N. (2018b). Exploring the relationship between institutional factors and FDI attractiveness: A meta-analytic review. *International Business Review*, *27*(1), 139–148.

Batschauer da Cruz, C. B., Eliete Floriani, D., & Amal, M. (2022). The OLI Paradigm as a comprehensive model of FDI determinants: A sub-national approach. *International Journal of Emerging Markets*, *17*(1), 145–176.

Berry, M. W., Mohamed, A., & Yap, B. W. (2019). *Supervised and unsupervised learning for data science*. Springer.

Blanton, S. L., & Blanton, R. G. (2007). What attracts foreign investors? An examination of human rights and foreign direct investment. *The Journal of Politics*, *69*(1), 143–155.

Chia-Cheng, C., Chun-Hung, C., & Ting-Yin, L. (2020). Investment performance of machine learning: Analysis of S&P 500 index. *International Journal of Economics and Financial Issues*, *10*(1), 59.

Chowdhury, A., & Mavrotas, G. (2006). FDI and growth: What causes what? *World Economy*, *29*(1), 9–19.

Contractor, F. J., Dangol, R., Nuruzzaman, N., & Raghunath, S. (2020). How do country regulations and business environment impact foreign direct investment (FDI) inflows? *International Business Review*, *29*(2), 101640.

Dal Bianco, S., & Loan, N. C. T. (2017). FDI inflows, price and exchange rate volatility: New empirical evidence from Latin America. *International Journal of Financial Studies*, *5*(1), 6.

Dimitrova, A., Rogmans, T., & Triki, D. (2020). Country-specific determinants of FDI inflows to the MENA region: A systematic review and future research directions. *Multinational Business Review*, *28*(1), 1–38.

Esteki, S., & Naghsh-Nilchi, A. R. (2022). Frequency component Kernel for SVM. *Neural Computing and Applications*, *34*(24), 22449–22464.

Fakher15, A. (2014). Quality of institutions and integration in the world economy: Applied study on Egypt. *Journal of Economics and Business*, *17*(2).

Fernandez, M. (2020). Analysis of foreign direct investment in Egypt. *International Journal of Economics and Management Studies*, *7*(11), 8–17.

Gupta, A., Mumtaz, S., Li, C.-H., Hussain, I., & Rotello, V. M. (2019). Combatting antibiotic-resistant bacteria using nanomaterials. *Chemical Society Reviews*, *48*(2), 415–427.

Hao, Y. (2023). The dynamic relationship between trade openness, foreign direct investment, capital formation, and industrial economic growth in China: New evidence from ARDL bounds testing approach. *Humanities and Social Sciences Communications*, *10*(1), 1–11.

Hastie, T., Tibshirani, R., Botstein, D., & Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology*, *2*(1), 1–12.

Hu, L., Chen, J., Vaughan, J., Aramideh, S., Yang, H., Wang, K., Sudjianto, A., & Nair, V. N. (2021). Supervised machine learning techniques: An overview with applications to banking. *International Statistical Review*, *89*(3), 573–604.

Ke, K., Chen, Y., Zhou, X., Yam, M. C., & Hu, S. (2023). Experimental and numerical study of a brace-type hybrid damper with steel slit plates enhanced by friction mechanism. *Thin-Walled Structures*, *182*, 110249.

Kerner, A. (2014). What we talk about when we talk about foreign direct investment. *International Studies Quarterly*, *58*(4), 804–815.

Kurecic, P., & Kokotovic, F. (2017a). The relevance of political stability on FDI: A VAR analysis and ARDL models for selected small, developed, and instability threatened economies. *Economies*, *5*(3), 22.

Kurecic, P., & Kokotovic, F. (2017b). The relevance of political stability on FDI: A VAR analysis and ARDL models for selected small, developed, and instability threatened economies. *Economies*, *5*(3), 22.

Latief, R., & Lefen, L. (2018). The effect of exchange rate volatility on international trade and foreign direct investment (FDI) in developing countries along "one belt and one road." *International Journal of Financial Studies*, *6*(4), 86.

Lee, T. K., Cho, J. H., Kwon, D. S., & Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Systems with Applications*, *117*, 228–242.

Loungani, P., & Razin, A. (2001). How beneficial is foreign direct investment for developing countries? *Finance and Development*, *38*(2), 6–9.

Maggioni, D., Santangelo, G. D., & Koymen-Ozer, S. (2019). MNEs' location strategies and labor standards: The role of operating and reputational considerations across industries. *Journal of International Business Studies*, *50*, 948–972.

Moustafa, E. (2021). The relationship between perceived corruption and FDI: a longitudinal study in the context of Egypt. *Transnational Corporations Journal*, *28*(2).

Moustafa, E. (2020). *Foreign Direct Investment and Corruption in Egypt: A Cointegration Analysis*.

Muñoz, M. A., Villanova, L., Baatar, D., & Smith-Miles, K. (2018). Instance spaces for machine learning classification. *Machine Learning*, *107*, 109–147.

Nagpal, A., & Jain, M. (2019). The Dubious Relationship Between Make in India and Foreign Direct Investment: The Story So Far and the Road Ahead. *Paradigm*, *23*(1), 98–115.

Nazeer, A. M., & Masih, M. (2017). *Impact of political instability on foreign direct investment and Economic Growth: Evidence from Malaysia*.

Nguyen, V. C., & Do, T. T. (2020). Impact of exchange rate shocks, inward FDI and import on export performance: A cointegration analysis. *The Journal of Asian Finance, Economics and Business*, *7*(4), 163–171.

Nhu, V.-H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., Jaafari, A., Chen, W., Miraki, S., & Dou, J. (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. *International Journal of Environmental Research and Public Health*, *17*(8), 2749.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Padhi, D. K., Padhy, N., Bhoi, A. K., Shafi, J., & Yesuf, S. H. (2022). An intelligent fusion model with portfolio selection and machine learning for stock market prediction. *Computational Intelligence and Neuroscience*, *2022*.

Park, B. I., & Roh, T. (2019). Chinese multinationals' FDI motivations: Suggestion for a new theory. *International Journal of Emerging Markets*, *14*(1), 70–90.

Peter, S. (2010). *OWNERSHIP, LOCATION SPECIFIC AND INTERNALIZATION DETERMINANTS AND FOREIGN DIRECT INVESTMENT INFLOWS*.

Rashid, M., Looi, X. H., & Wong, S. J. (2017). Political stability and FDI in the most competitive Asia Pacific countries. *Journal of Financial Economic Policy*, *9*(02), 140–155.

Salem, M. I. (n.d.). *DETERMINANTS OF FOREIGN DIRECT INVESTMENT IN EGYPT: AN EMPIRICAL ANALYSIS*.

Samal, J. E. (2018). *Effect of Inflation on Foreign Direct Investments in Kenya*.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Siddiqui, H. A. A., & Aumeboonsuke, V. (2014). Role of interest rate in attracting the FDI: Study on ASEAN 5 economy. *International Journal of Technical Research and Applications*, *2*(3), 59–70.

Singh, S., Dhillon, K., Kaulich, F., & Chen, W. (2011). Location Determinants of FDI in Sub-Saharan Africa: An Empirical Analysis. In *Dynamics of Globalization: Location-Specific Advantages or Liabilities of Foreignness?* (pp. 327–356). Emerald Group Publishing Limited.

Stoian, C., & Filippaios, F. (2008). Dunning's eclectic paradigm: A holistic, yet context specific framework for analysing the determinants of outward FDI: Evidence from international Greek investments. *International Business Review*, *17*(3), 349–367.

Topal, M. H. (2016). The effect of country risk on foreign direct investment: A dynamic panel data analysis for developing countries. *Journal of Economics Library*, *3*(1), 141–155.

Ullah, I., & Khan, M. A. (2017). Institutional quality and foreign direct investment inflows: Evidence from Asian countries. *Journal of Economic Studies*, *44*(6), 1030–1050.

Wako, H. A. (2021). Foreign direct investment in sub-Saharan Africa: Beyond its growth effect. *Research in Globalization*, *3*, 100054.

Yassin, B. M., Elfiky, F., & El Nimer, N. (2020). The Role of Institutional Determinants in Attracting Foreign Direct Investment to Egypt: Empirical Study. *FWU Journal of Social Sciences*, *14*(3).

Ye, Y., Wu, K., Xie, Y., Huang, G., Wang, C., & Chen, J. (2019). How firm heterogeneity affects foreign direct investment location choice: Micro-evidence from new foreign manufacturing firms in the Pearl River Delta. *Applied Geography*, *106*, 11–21.

Zenasni, S., & Benhabib, A. (2013). The determinants of foreign direct investment and their impact on growth: Panel data analysis for AMU countries. *International Journal of Innovation and Applied Studies*, *2*(3), 300–313.

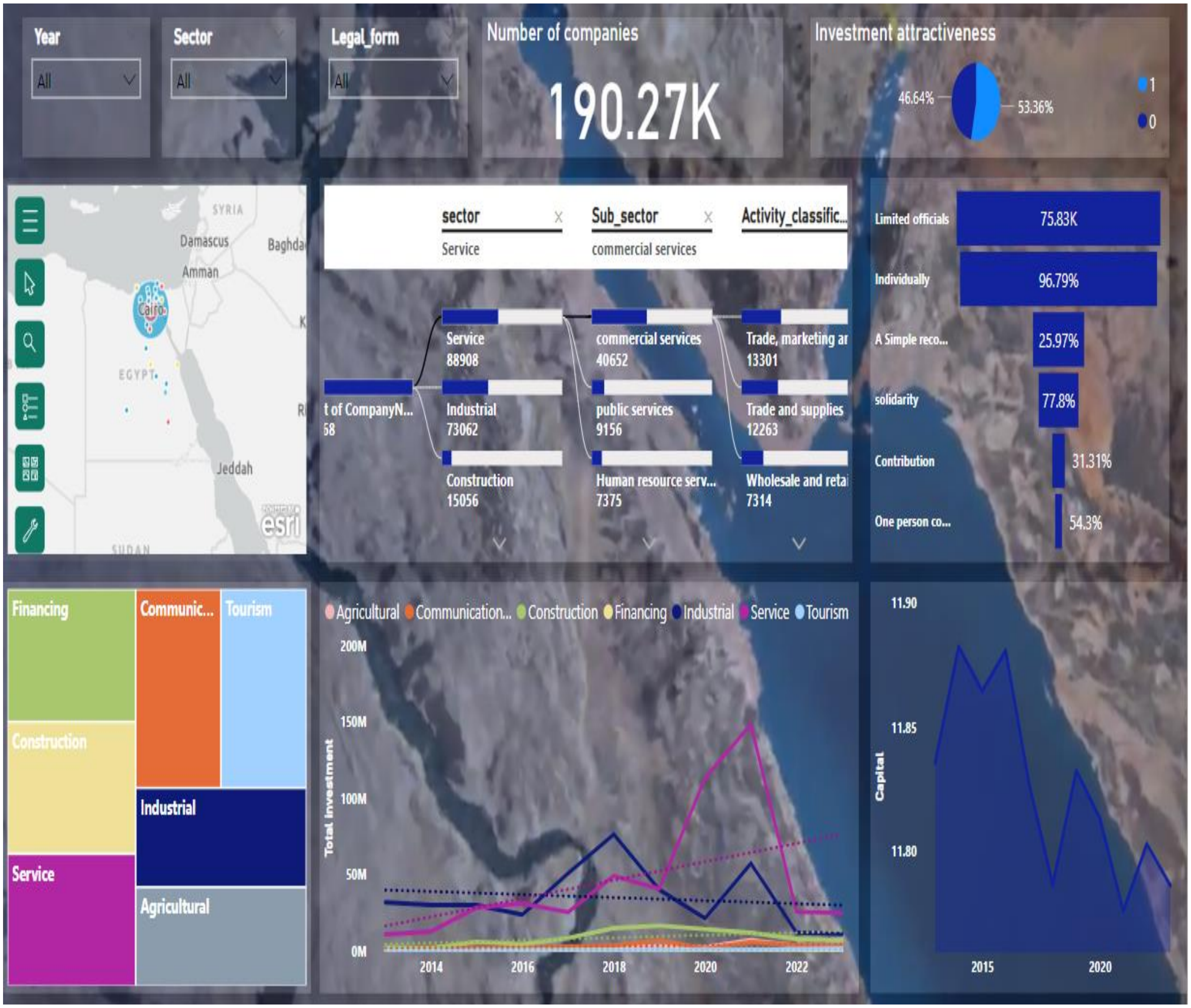
Zghidi, N., Mohamed Sghaier, I., & Abida, Z. (2016). Does economic freedom enhance the impact of foreign direct investment on economic growth in North African countries? A panel data analysis. *African Development Review*, *28*(1), 64–74.

Zhang, K. H. (2001). How does foreign direct investment affect economic growth in China? *Economics of Transition*, *9*(3), 679–693.

Appendixes

Appendix A

Dashboard



Year: All

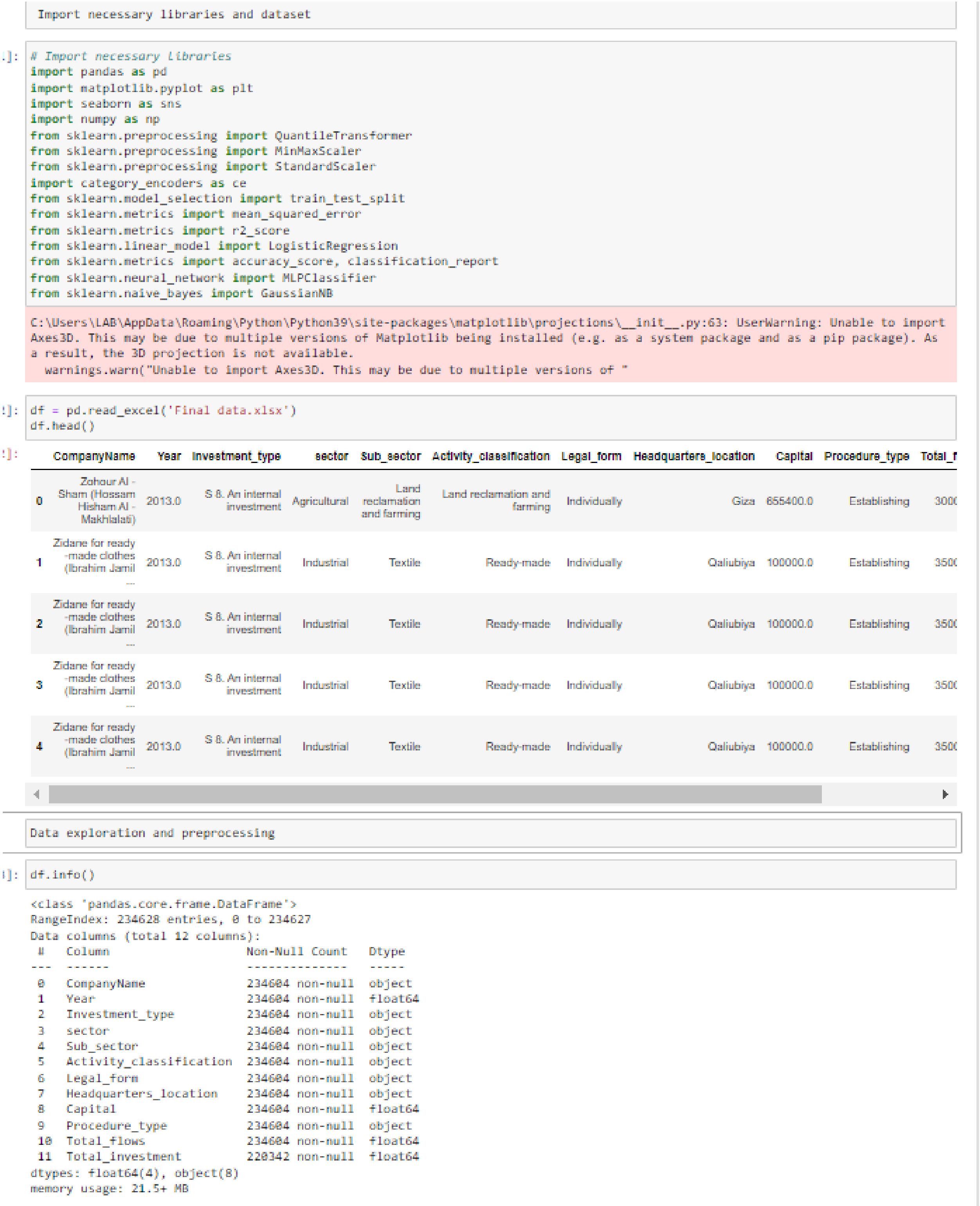
Sector: All

Legal_form: All

Number of companies: 190.27K

Appendix B

Python code



```
Import necessary libraries and dataset
```

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.preprocessing import QuantileTransformer
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
import category_encoders as ce
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.neural_network import MLPClassifier
from sklearn.naive_bayes import GaussianNB
```

```
C:\Users\LAB\AppData\Roaming\Python\Python39\site-packages\matplotlib\projections\__init__.py:63: UserWarning: Unable to import
Axes3D. This may be due to multiple versions of Matplotlib being installed (e.g. as a system package and as a pip package). As
a result, the 3D projection is not available.
  warnings.warn("Unable to import Axes3D. This may be due to multiple versions of "
```

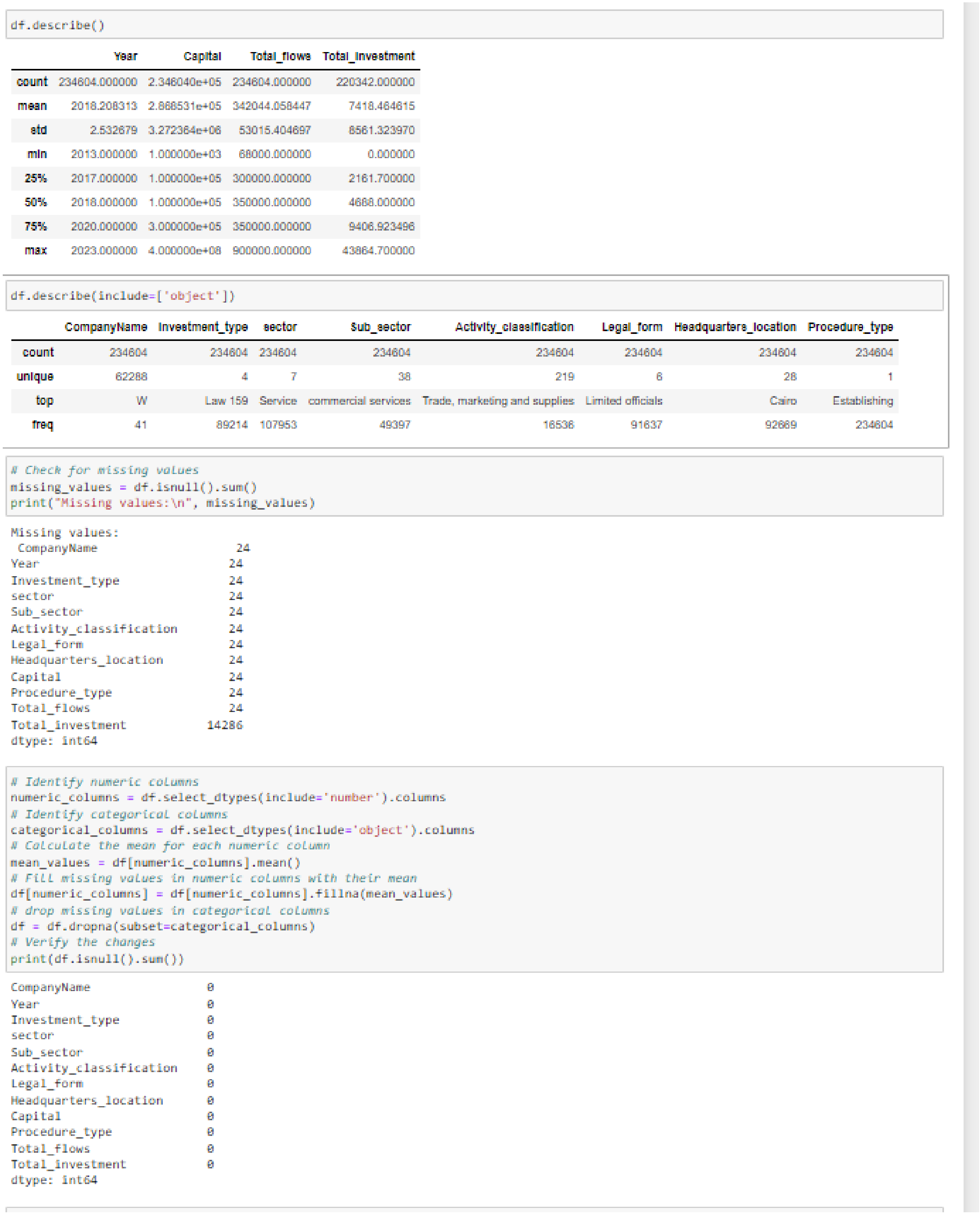
```
df = pd.read_excel('Final data.xlsx')
df.head()
```

	CompanyName	Year	Investment_type	sector	Sub_sector	Activity_classification	Legal_form	Headquarters_location	Capital	Procedure_type	Total_f
0	Zohour Al - Sham (Hossam Hisham Al - Makhlalati)	2013.0	S 8. An internal investment	Agricultural	Land reclamation and farming	Land reclamation and farming	Individually	Giza	655400.0	Establishing	300(
1	Zidane for ready -made clothes (Ibrahim Jamil ...	2013.0	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	Establishing	350(
2	Zidane for ready -made clothes (Ibrahim Jamil ...	2013.0	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	Establishing	350(
3	Zidane for ready -made clothes (Ibrahim Jamil ...	2013.0	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	Establishing	350(
4	Zidane for ready -made clothes (Ibrahim Jamil ...	2013.0	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	Establishing	350(

```
Data exploration and preprocessing
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234628 entries, 0 to 234627
Data columns (total 12 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   CompanyName              234604 non-null  object
 1   Year                     234604 non-null  float64
 2   Investment_type          234604 non-null  object
 3   sector                   234604 non-null  object
 4   Sub_sector               234604 non-null  object
 5   Activity_classification  234604 non-null  object
 6   Legal_form               234604 non-null  object
 7   Headquarters_location    234604 non-null  object
 8   Capital                  234604 non-null  float64
 9   Procedure_type           234604 non-null  object
 10  Total_flows              234604 non-null  float64
 11  Total_investment         220342 non-null  float64
dtypes: float64(4), object(8)
memory usage: 21.5+ MB
```



```
df.describe()
```

	Year	Capital	Total_flows	Total_investment
count	234604.000000	2.346040e+05	234604.000000	220342.000000
mean	2018.208313	2.868531e+05	342044.058447	7418.464615
std	2.532679	3.272364e+06	53015.404697	8561.323970
min	2013.000000	1.000000e+03	68000.000000	0.000000
25%	2017.000000	1.000000e+05	300000.000000	2161.700000
50%	2018.000000	1.000000e+05	350000.000000	4668.000000
75%	2020.000000	3.000000e+05	350000.000000	9406.923496
max	2023.000000	4.000000e+08	900000.000000	43864.700000

```
df.describe(include=['object'])
```

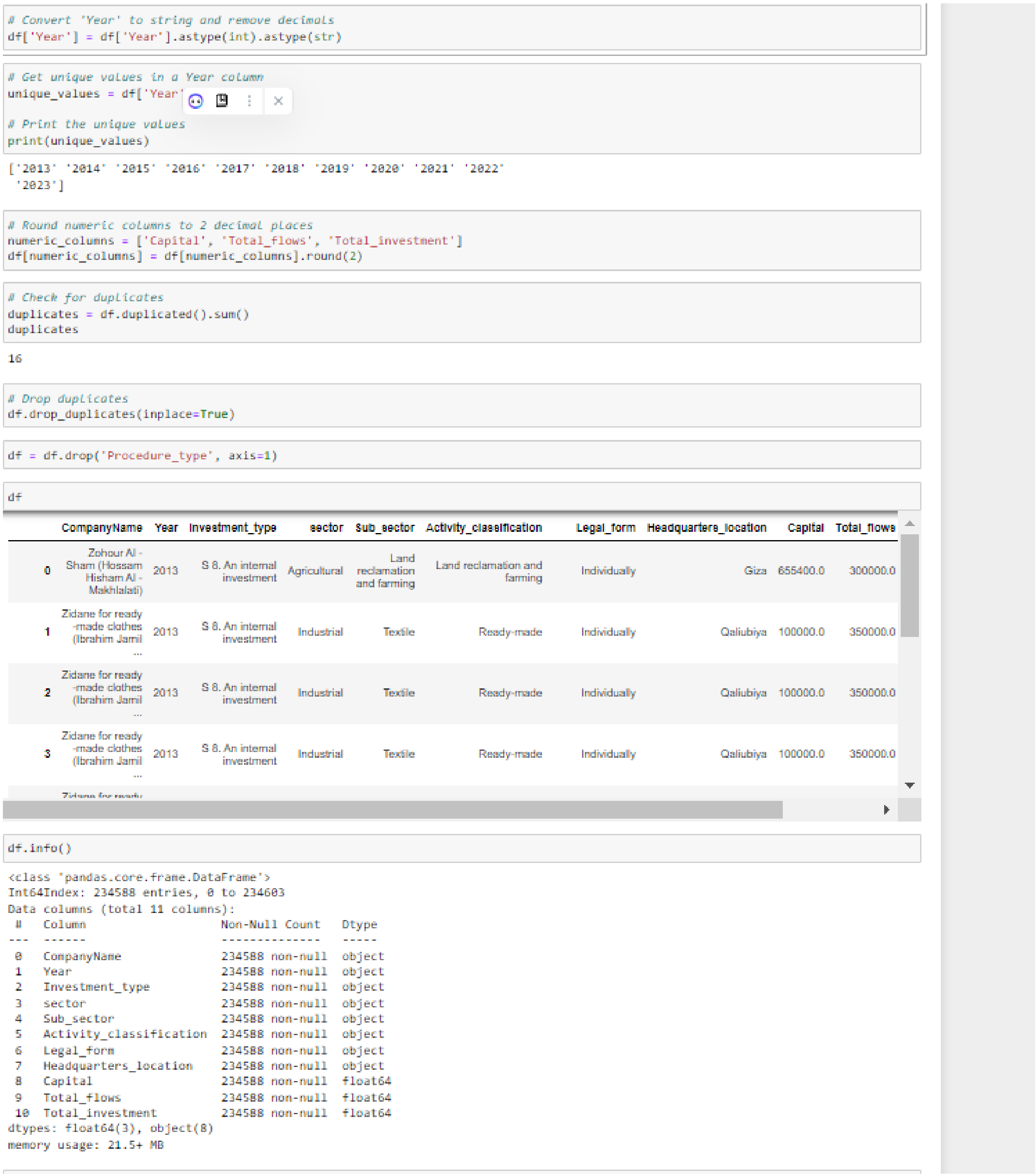
	CompanyName	Investment_type	sector	Sub_sector	Activity_classification	Legal_form	Headquarters_location	Procedure_type
count	234604	234604	234604	234604	234604	234604	234604	234604
unique	62288	4	7	38	219	6	28	1
top	W	Law 159	Service	commercial services	Trade, marketing and supplies	Limited officials	Cairo	Establishing
freq	41	89214	107953	49397	16536	91637	92689	234604

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing values:\n", missing_values)
```

```
Missing values:
 CompanyName                 24
Year                        24
Investment_type             24
sector                      24
Sub_sector                  24
Activity_classification     24
Legal_form                  24
Headquarters_location       24
Capital                     24
Procedure_type              24
Total_flows                 24
Total_investment         14286
dtype: int64
```

```
# Identify numeric columns
numeric_columns = df.select_dtypes(include='number').columns
# Identify categorical columns
categorical_columns = df.select_dtypes(include='object').columns
# Calculate the mean for each numeric column
mean_values = df[numeric_columns].mean()
# Fill missing values in numeric columns with their mean
df[numeric_columns] = df[numeric_columns].fillna(mean_values)
# drop missing values in categorical columns
df = df.dropna(subset=categorical_columns)
# Verify the changes
print(df.isnull().sum())
```

```
CompanyName                0
Year                       0
Investment_type            0
sector                     0
Sub_sector                 0
Activity_classification    0
Legal_form                 0
Headquarters_location      0
Capital                    0
Procedure_type             0
Total_flows                0
Total_investment           0
dtype: int64
```



```
# Convert 'Year' to string and remove decimals
df['Year'] = df['Year'].astype(int).astype(str)
```

```
# Get unique values in a Year column
unique_values = df['Year'

# Print the unique values
print(unique_values)
```

```
['2013' '2014' '2015' '2016' '2017' '2018' '2019' '2020' '2021' '2022'
 '2023']
```

```
# Round numeric columns to 2 decimal places
numeric_columns = ['Capital', 'Total_flows', 'Total_investment']
df[numeric_columns] = df[numeric_columns].round(2)
```

```
# Check for duplicates
duplicates = df.duplicated().sum()
duplicates
```

```
16
```

```
# Drop duplicates
df.drop_duplicates(inplace=True)
```

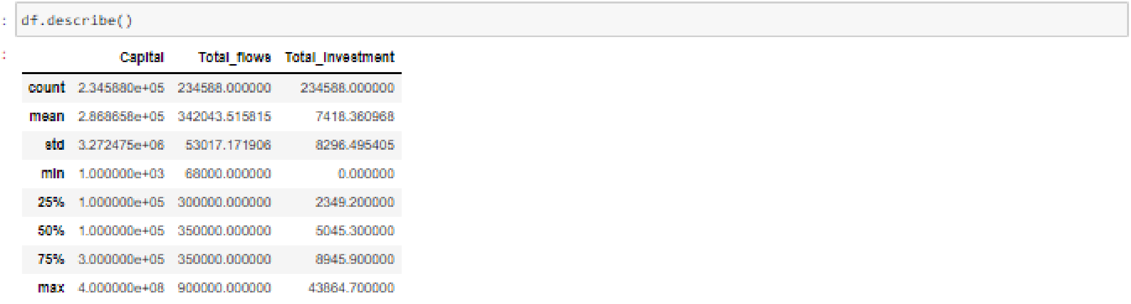
```
df = df.drop('Procedure_type', axis=1)
```

```
df
```

	CompanyName	Year	Investment_type	sector	Sub_sector	Activity_classification	Legal_form	Headquarters_location	Capital	Total_flows
0	Zohour Al - Sham (Hossam Hisham Al - Makhlalati)	2013	S 8. An internal investment	Agricultural	Land reclamation and farming	Land reclamation and farming	Individually	Giza	655400.0	300000.0
1	Zidane for ready -made clothes (Ibrahim Jamil ...	2013	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	350000.0
2	Zidane for ready -made clothes (Ibrahim Jamil ...	2013	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	350000.0
3	Zidane for ready -made clothes (Ibrahim Jamil ...	2013	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	100000.0	350000.0

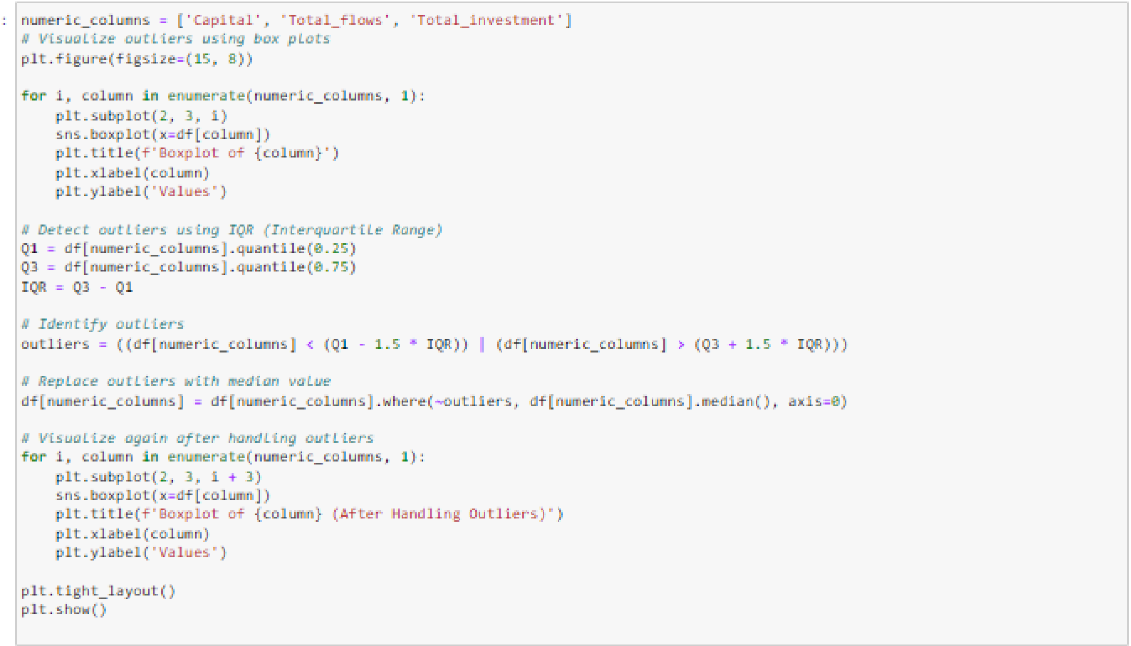
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 234588 entries, 0 to 234603
Data columns (total 11 columns):
 #   Column                   Non-Null Count   Dtype  
---  ------                   --------------   -----  
 0   CompanyName              234588 non-null  object 
 1   Year                     234588 non-null  object 
 2   Investment_type          234588 non-null  object 
 3   sector                   234588 non-null  object 
 4   Sub_sector               234588 non-null  object 
 5   Activity_classification  234588 non-null  object 
 6   Legal_form               234588 non-null  object 
 7   Headquarters_location    234588 non-null  object 
 8   Capital                  234588 non-null  float64
 9   Total_flows              234588 non-null  float64
 10  Total_investment         234588 non-null  float64
dtypes: float64(3), object(8)
memory usage: 21.5+ MB
```

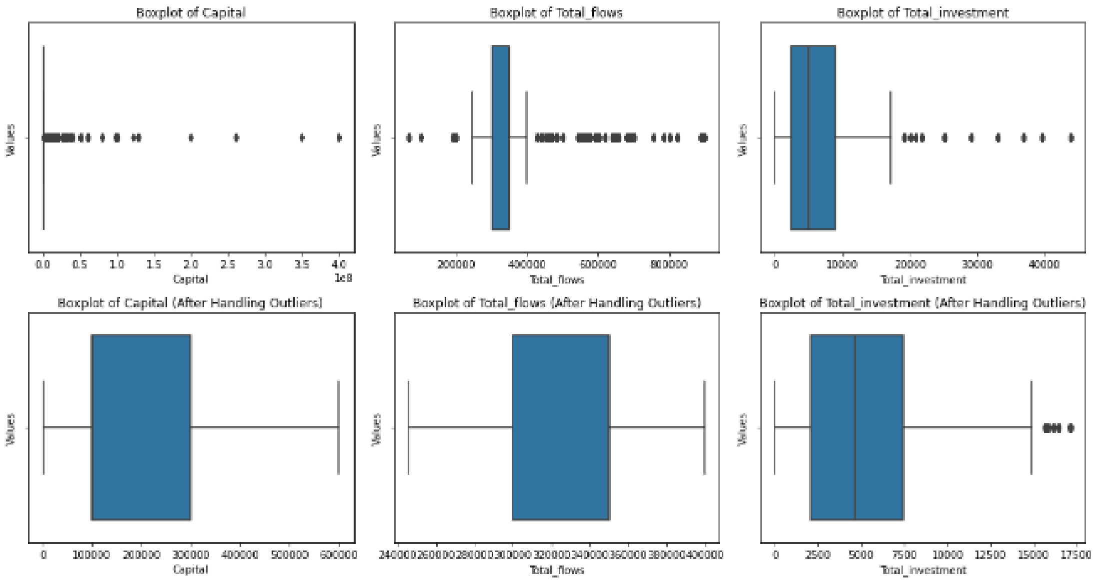


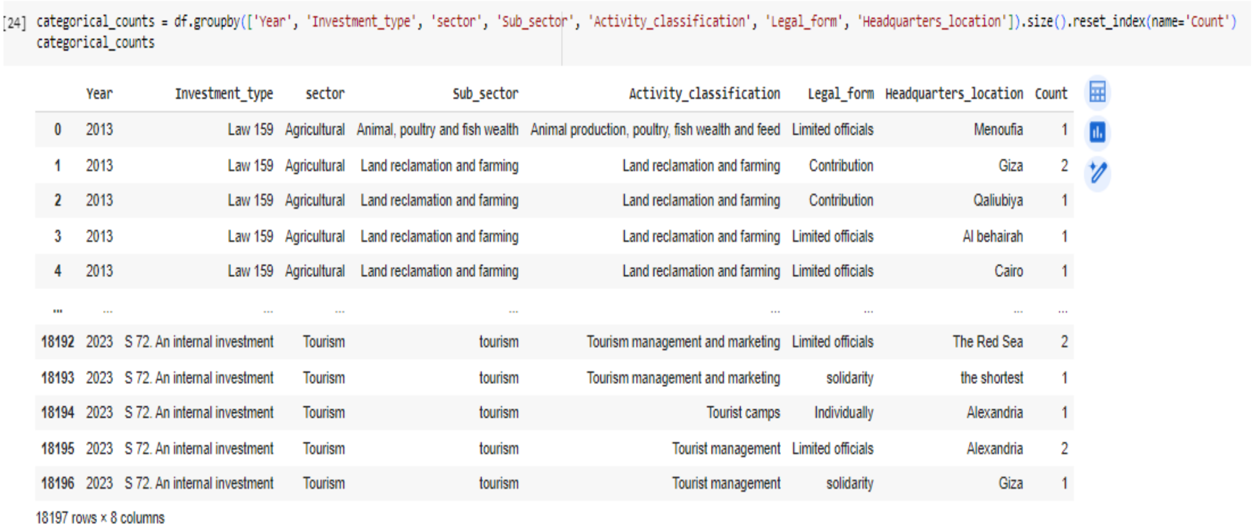
```
df.describe()
```

	Capital	Total_flows	Total_investment
count	2.345880e+05	234588.000000	234588.000000
mean	2.868658e+05	342043.515815	7418.360968
std	3.272475e+06	53017.171906	8296.495405
min	1.000000e+03	68000.000000	0.000000
25%	1.000000e+05	300000.000000	2349.200000
50%	1.000000e+05	350000.000000	5045.300000
75%	3.000000e+05	350000.000000	8945.900000
max	4.000000e+08	900000.000000	43864.700000



```
numeric_columns = ['Capital', 'Total_flows', 'Total_investment']
# Visualize outliers using box plots
plt.figure(figsize=(15, 8))

for i, column in enumerate(numeric_columns, 1):
    plt.subplot(2, 3, i)
    sns.boxplot(x=df[column])
    plt.title(f'Boxplot of {column}')
    plt.xlabel(column)
    plt.ylabel('Values')

# Detect outliers using IQR (Interquartile Range)
Q1 = df[numeric_columns].quantile(0.25)
Q3 = df[numeric_columns].quantile(0.75)
IQR = Q3 - Q1

# Identify outliers
outliers = ((df[numeric_columns] < (Q1 - 1.5 * IQR)) | (df[numeric_columns] > (Q3 + 1.5 * IQR)))

# Replace outliers with median value
df[numeric_columns] = df[numeric_columns].where(~outliers, df[numeric_columns].median(), axis=0)

# Visualize again after handling outliers
for i, column in enumerate(numeric_columns, 1):
    plt.subplot(2, 3, i + 3)
    sns.boxplot(x=df[column])
    plt.title(f'Boxplot of {column} (After Handling Outliers)')
    plt.xlabel(column)
    plt.ylabel('Values')

plt.tight_layout()
plt.show()
```

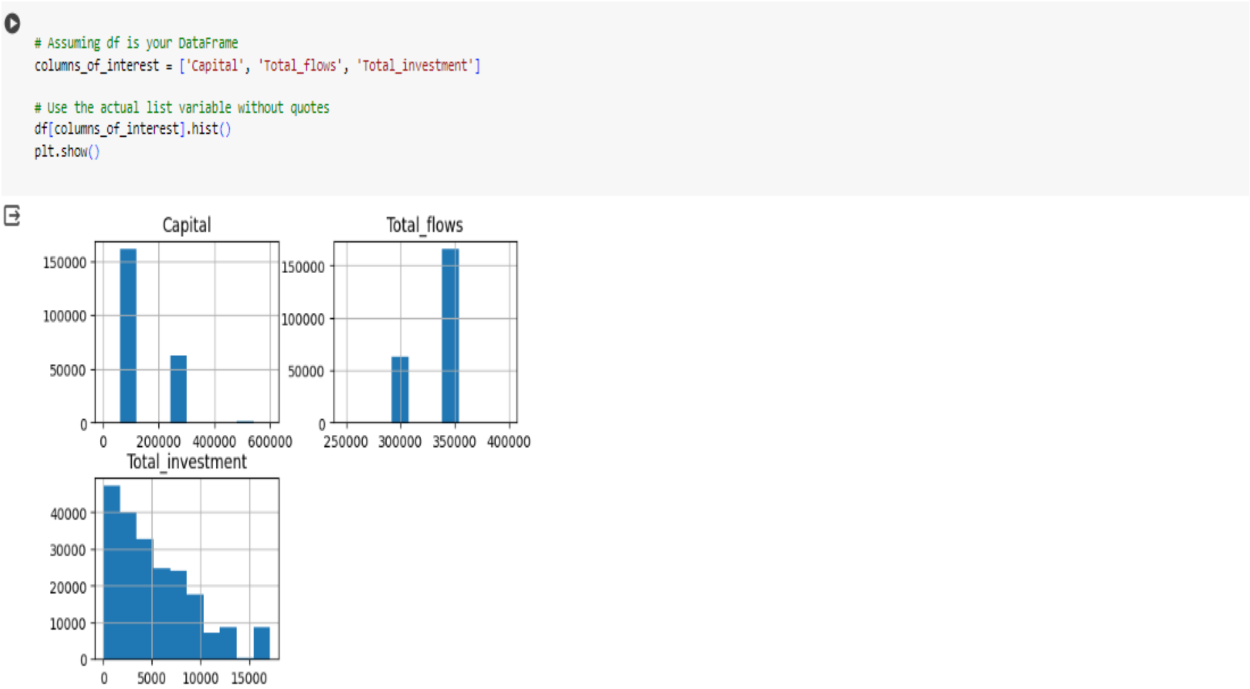




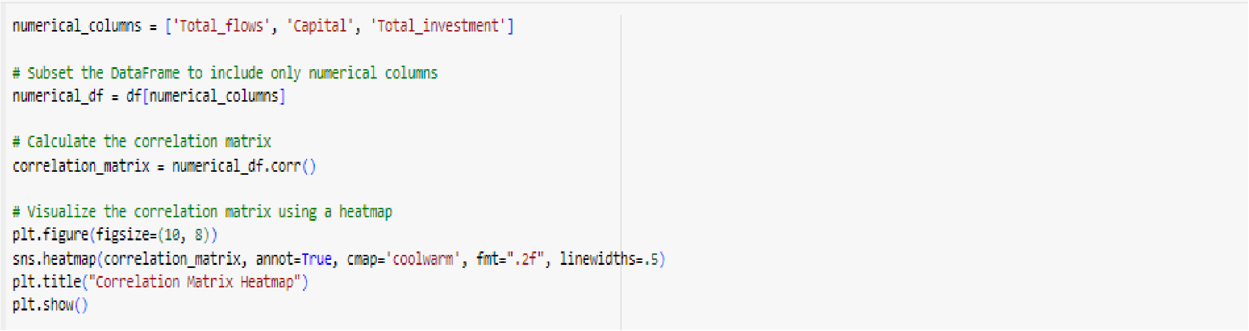
```
[24] categorical_counts = df.groupby(['Year', 'Investment_type', 'sector', 'Sub_sector', 'Activity_classification', 'Legal_form', 'Headquarters_location']).size().reset_index(name='Count')
     categorical_counts
```

	Year	Investment_type	sector	Sub_sector	Activity_classification	Legal_form	Headquarters_location	Count
0	2013	Law 159	Agricultural	Animal, poultry and fish wealth	Animal production, poultry, fish wealth and feed	Limited officials	Menoufia	1
1	2013	Law 159	Agricultural	Land reclamation and farming	Land reclamation and farming	Contribution	Giza	2
2	2013	Law 159	Agricultural	Land reclamation and farming	Land reclamation and farming	Contribution	Qaliubiya	1
3	2013	Law 159	Agricultural	Land reclamation and farming	Land reclamation and farming	Limited officials	Al behairah	1
4	2013	Law 159	Agricultural	Land reclamation and farming	Land reclamation and farming	Limited officials	Cairo	1
...	...	...	...	...	...	...	...	...
18192	2023	S 72. An internal investment	Tourism	tourism	Tourism management and marketing	Limited officials	The Red Sea	2
18193	2023	S 72. An internal investment	Tourism	tourism	Tourism management and marketing	solidarity	the shortest	1
18194	2023	S 72. An internal investment	Tourism	tourism	Tourist camps	Individually	Alexandria	1
18195	2023	S 72. An internal investment	Tourism	tourism	Tourist management	Limited officials	Alexandria	2
18196	2023	S 72. An internal investment	Tourism	tourism	Tourist management	solidarity	Giza	1

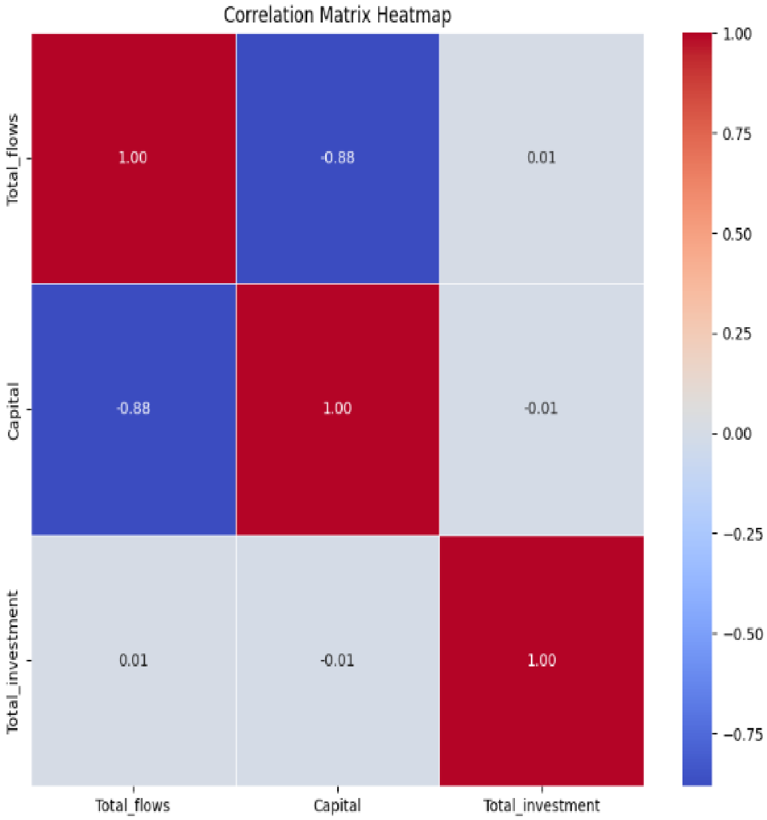
18197 rows × 8 columns

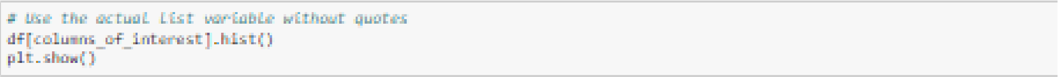


```
# Assuming df is your DataFrame
columns_of_interest = ['Capital', 'Total_flows', 'Total_investment']

# Use the actual list variable without quotes
df[columns_of_interest].hist()
plt.show()
```

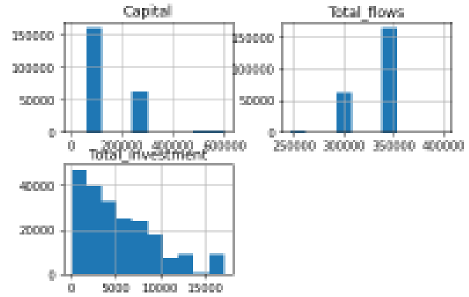


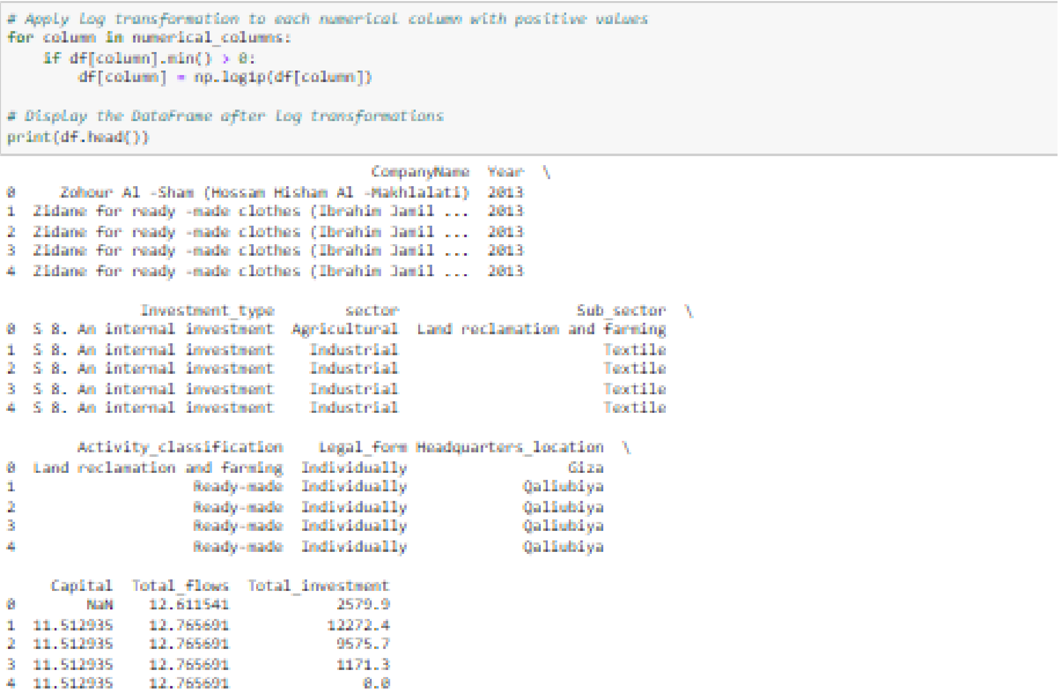
```
numerical_columns = ['Total_flows', 'Capital', 'Total_investment']

# Subset the DataFrame to include only numerical columns
numerical_df = df[numerical_columns]

# Calculate the correlation matrix
correlation_matrix = numerical_df.corr()

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title("Correlation Matrix Heatmap")
plt.show()
```





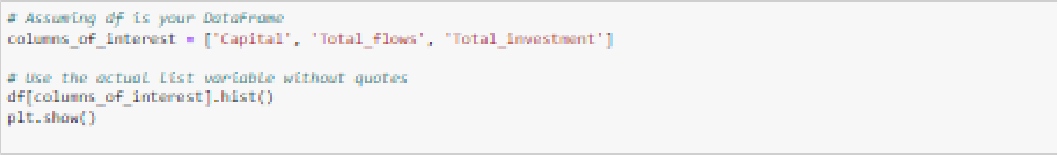
```
# Use the actual list variable without quotes
df[columns_of_interest].hist()
plt.show()
```



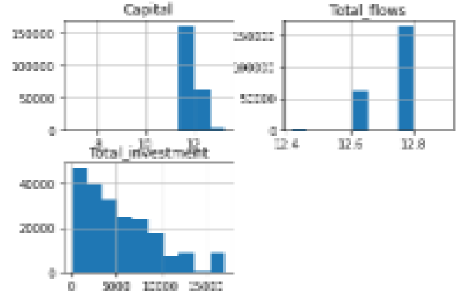


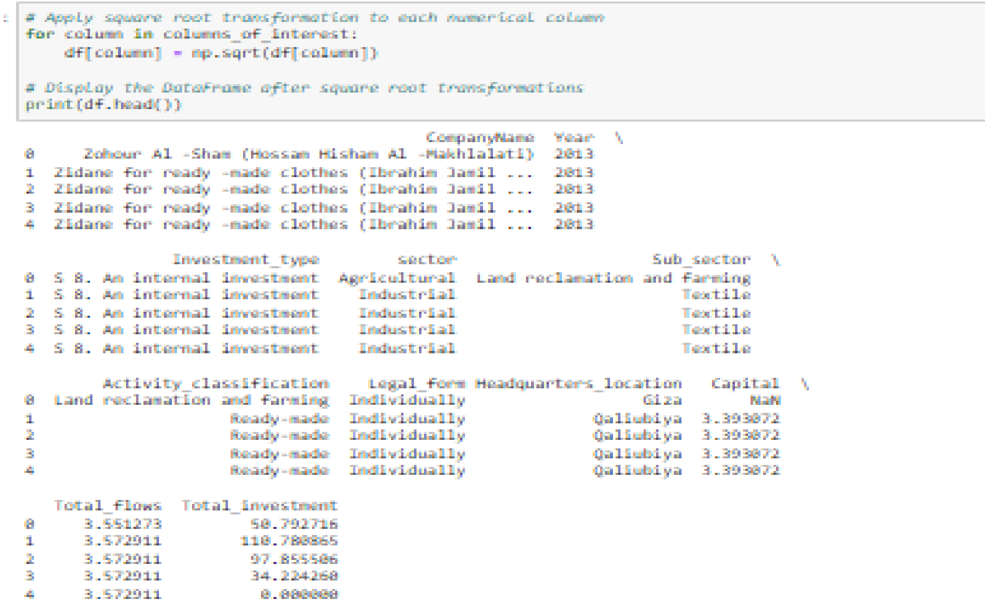
```
# Apply log transformation to each numerical column with positive values
for column in numerical_columns:
    if df[column].min() > 0:
        df[column] = np.log1p(df[column])

# Display the DataFrame after log transformations
print(df.head())
```

```
                                         CompanyName  Year  \
0     Zohour Al -Sham (Hossam Hisham Al -Makhlalati)  2013
1  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
2  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
3  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
4  Zidane for ready -made clothes (Ibrahim Jamil ...  2013

                Investment_type        sector                    Sub_sector  \
0  S 8. An internal investment  Agricultural  Land reclamation and farming
1  S 8. An internal investment    Industrial                       Textile
2  S 8. An internal investment    Industrial                       Textile
3  S 8. An internal investment    Industrial                       Textile
4  S 8. An internal investment    Industrial                       Textile

          Activity_classification    Legal_form Headquarters_location  \
0  Land reclamation and farming  Individually                  Giza
1                    Ready-made  Individually             Qaliubiya
2                    Ready-made  Individually             Qaliubiya
3                    Ready-made  Individually             Qaliubiya
4                    Ready-made  Individually             Qaliubiya

     Capital  Total_flows  Total_investment
0        NaN    12.611541            2579.9
1  11.512935    12.765691           12272.4
2  11.512935    12.765691            9575.7
3  11.512935    12.765691            1171.3
4  11.512935    12.765691               0.0
```



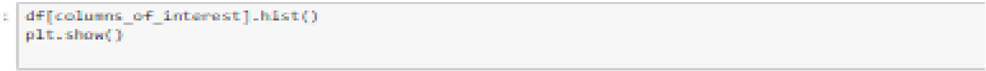
```
# Assuming df is your DataFrame
columns_of_interest = ['Capital', 'Total_flows', 'Total_investment']

# Use the actual list variable without quotes
df[columns_of_interest].hist()
plt.show()
```



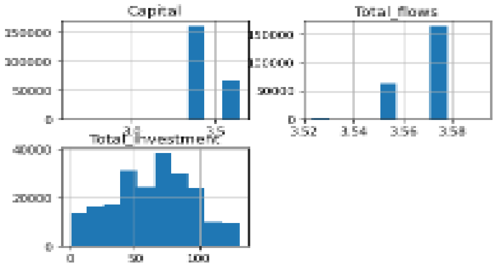


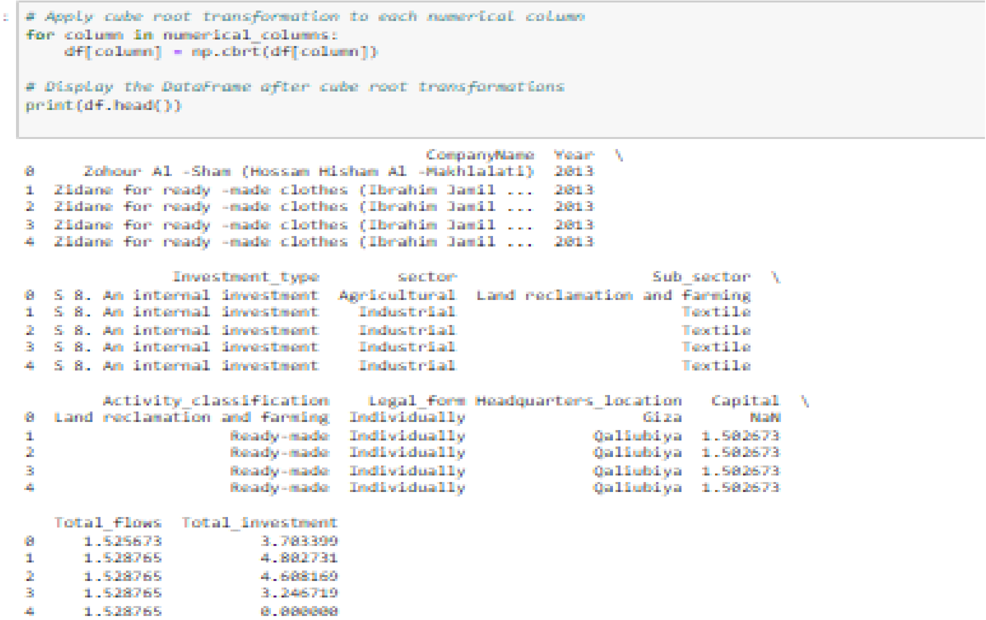
```
# Apply square root transformation to each numerical column
for column in columns_of_interest:
    df[column] = np.sqrt(df[column])

# Display the DataFrame after square root transformations
print(df.head())
```

```
                                      CompanyName  Year  \
0     Zohour Al -Sham (Hossam Hisham Al -Makhlalati)  2013
1  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
2  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
3  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
4  Zidane for ready -made clothes (Ibrahim Jamil ...  2013

             Investment_type        sector                    Sub_sector  \
0  S 8. An internal investment  Agricultural  Land reclamation and farming
1  S 8. An internal investment    Industrial                      Textile
2  S 8. An internal investment    Industrial                      Textile
3  S 8. An internal investment    Industrial                      Textile
4  S 8. An internal investment    Industrial                      Textile

         Activity_classification    Legal_form Headquarters_location   Capital  \
0  Land reclamation and farming  Individually                  Giza       NaN
1                     Ready-made  Individually             Qaliubiya  3.393072
2                     Ready-made  Individually             Qaliubiya  3.393072
3                     Ready-made  Individually             Qaliubiya  3.393072
4                     Ready-made  Individually             Qaliubiya  3.393072

   Total_flows  Total_investment
0     3.551273         50.792716
1     3.572911        110.780865
2     3.572911         97.855506
3     3.572911         34.224260
4     3.572911          0.000000
```



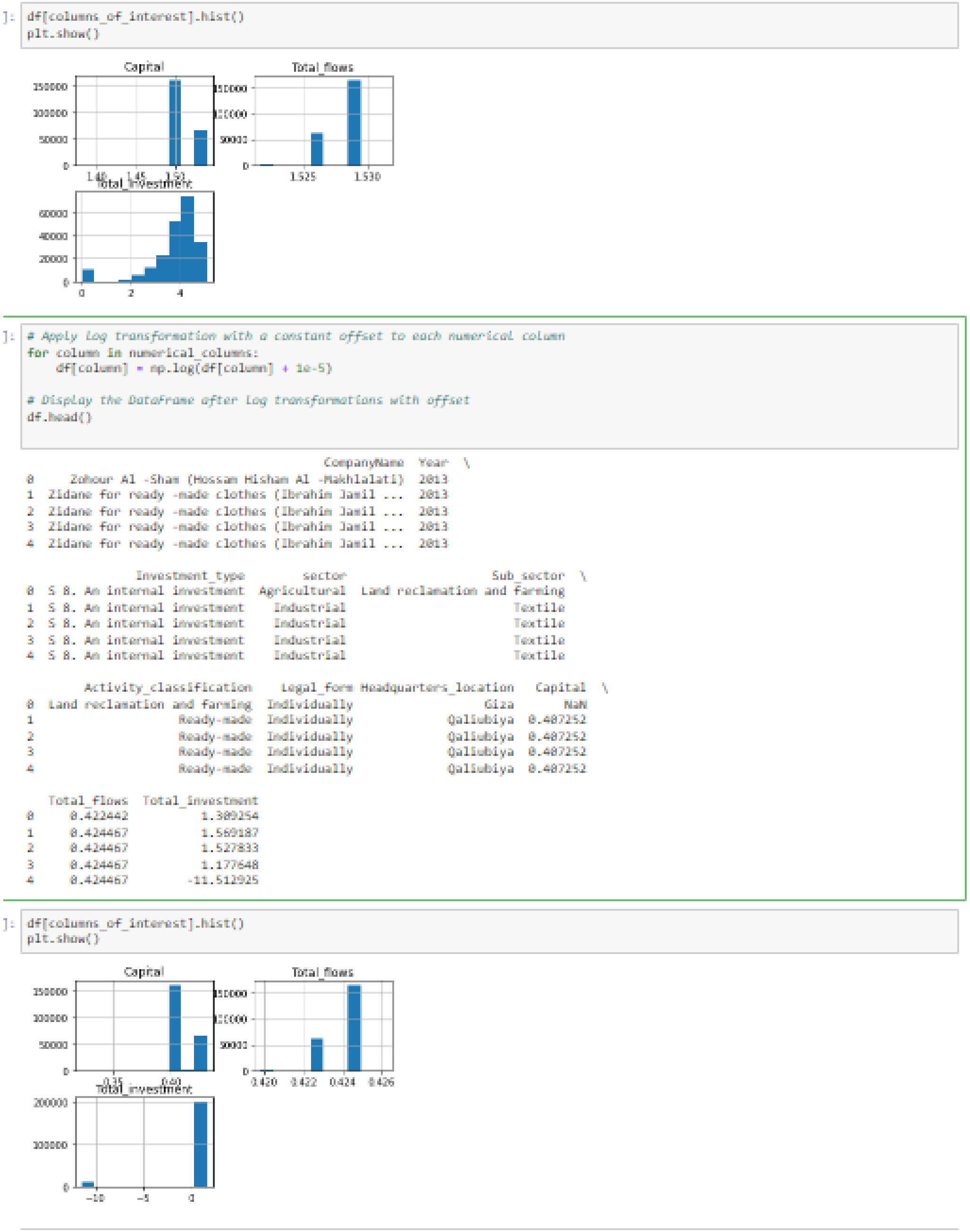
```
df[columns_of_interest].hist()
plt.show()
```





```
# Apply cube root transformation to each numerical column
for column in numerical_columns:
    df[column] = np.cbrt(df[column])

# Display the DataFrame after cube root transformations
print(df.head())
```

```
                                      CompanyName  Year  \
0     Zohour Al -Sham (Hossam Hisham Al -Makhlalati)  2013
1  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
2  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
3  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
4  Zidane for ready -made clothes (Ibrahim Jamil ...  2013

             Investment_type        sector                    Sub_sector  \
0  S 8. An internal investment  Agricultural  Land reclamation and farming
1  S 8. An internal investment    Industrial                      Textile
2  S 8. An internal investment    Industrial                      Textile
3  S 8. An internal investment    Industrial                      Textile
4  S 8. An internal investment    Industrial                      Textile

         Activity_classification    Legal_form Headquarters_location   Capital  \
0  Land reclamation and farming  Individually                  Giza       NaN
1                     Ready-made  Individually             Qaliubiya  1.502673
2                     Ready-made  Individually             Qaliubiya  1.502673
3                     Ready-made  Individually             Qaliubiya  1.502673
4                     Ready-made  Individually             Qaliubiya  1.502673

   Total_flows  Total_investment
0     1.525673          3.703399
1     1.528765          4.802731
2     1.528765          4.608169
3     1.528765          3.246719
4     1.528765          0.000000
```

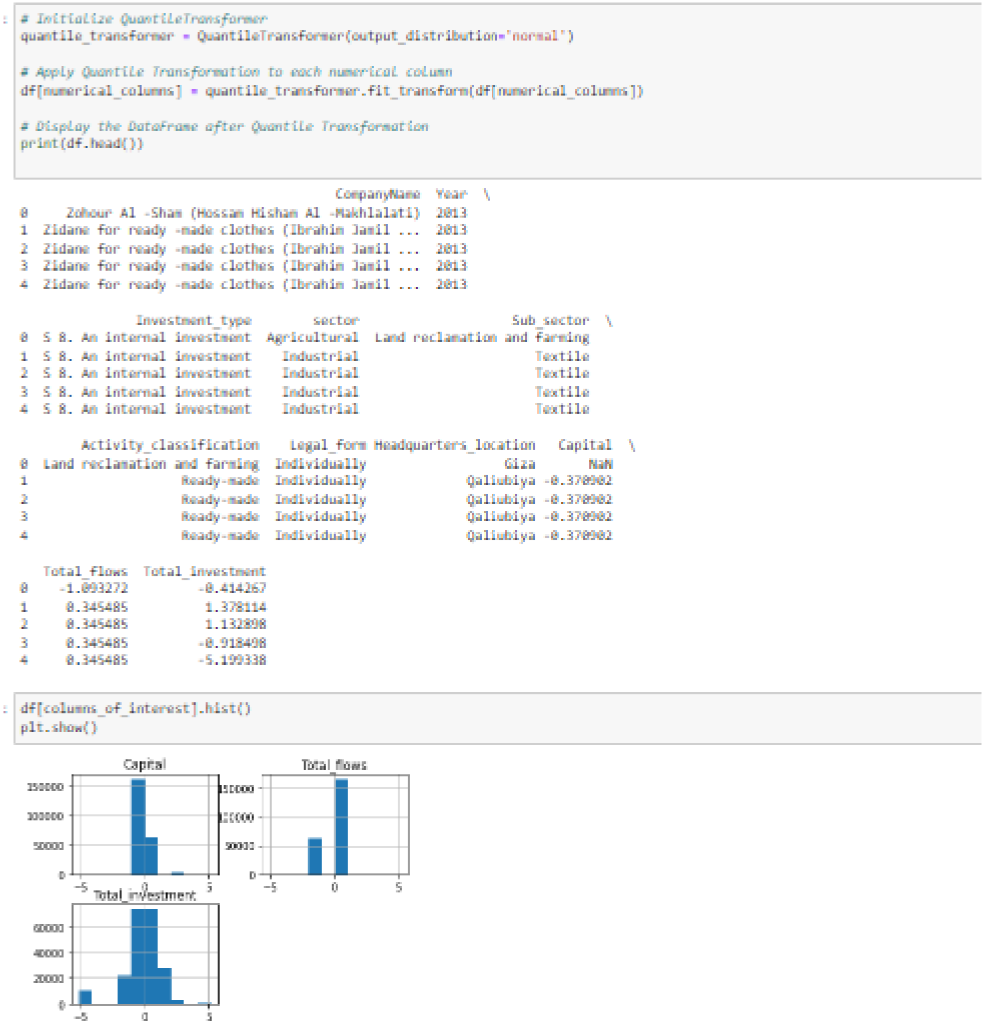


```
]: df[columns_of_interest].hist()
   plt.show()
```

```
]: # Apply log transformation with a constant offset to each numerical column
   for column in numerical_columns:
       df[column] = np.log(df[column] + 1e-5)

   # Display the DataFrame after log transformations with offset
   df.head()
```

```
                                         CompanyName  Year  \
0     Zohour Al -Sham (Hossam Hisham Al -Makhlalati)  2013
1  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
2  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
3  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
4  Zidane for ready -made clothes (Ibrahim Jamil ...  2013

              Investment_type        sector                     Sub_sector  \
0  S 8. An internal investment  Agricultural  Land reclamation and farming
1  S 8. An internal investment    Industrial                       Textile
2  S 8. An internal investment    Industrial                       Textile
3  S 8. An internal investment    Industrial                       Textile
4  S 8. An internal investment    Industrial                       Textile

        Activity_classification    Legal_form Headquarters_location   Capital  \
0  Land reclamation and farming  Individually                  Giza       NaN
1                    Ready-made  Individually             Qaliubiya  0.407252
2                    Ready-made  Individually             Qaliubiya  0.407252
3                    Ready-made  Individually             Qaliubiya  0.407252
4                    Ready-made  Individually             Qaliubiya  0.407252

   Total_flows  Total_investment
0     0.422442          1.389254
1     0.424467          1.569187
2     0.424467          1.527833
3     0.424467          1.177648
4     0.424467        -11.512925
```

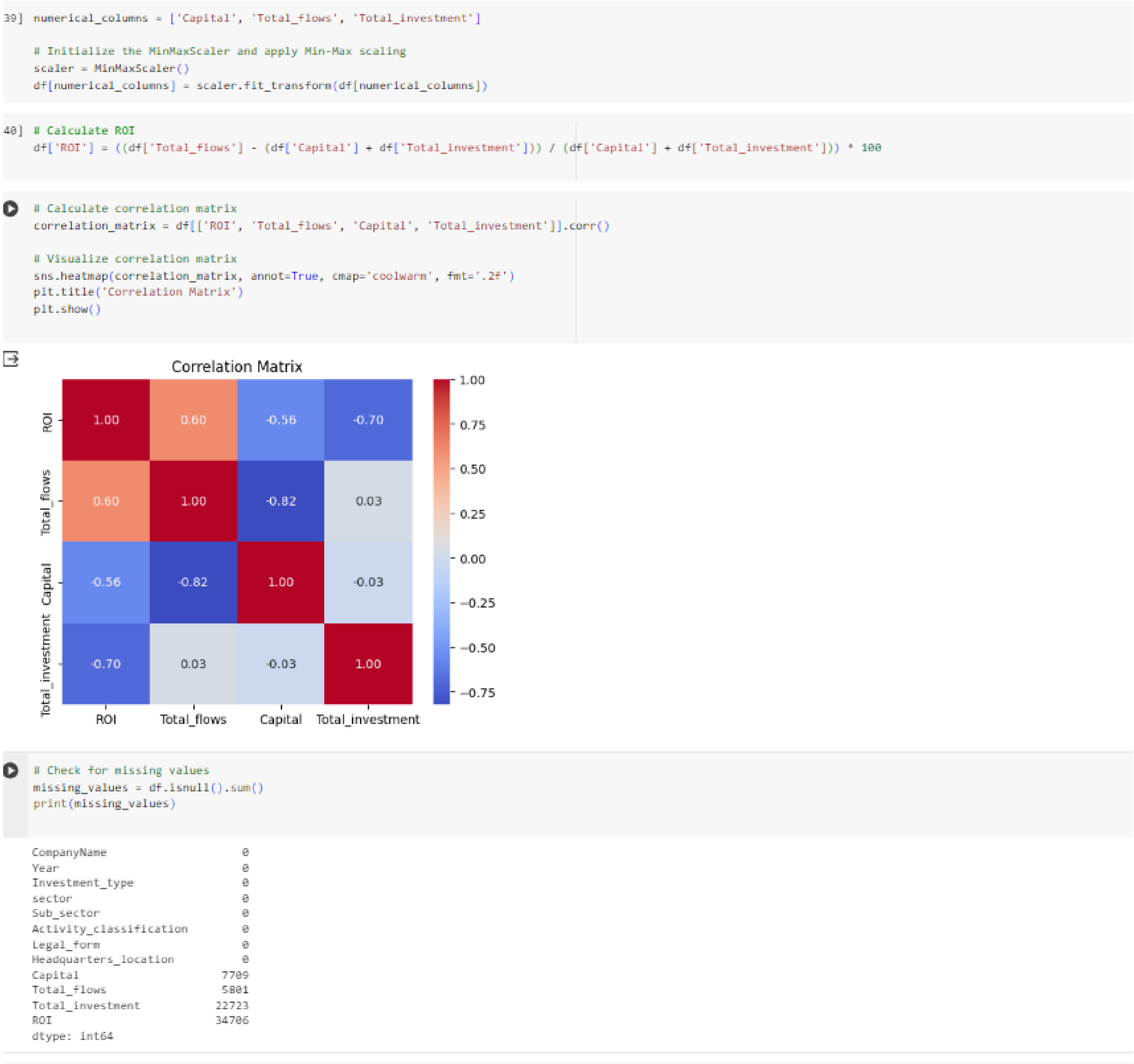
```
]: df[columns_of_interest].hist()
   plt.show()
```



```
# Initialize QuantileTransformer
quantile_transformer = QuantileTransformer(output_distribution='normal')

# Apply Quantile Transformation to each numerical column
df[numerical_columns] = quantile_transformer.fit_transform(df[numerical_columns])

# Display the DataFrame after Quantile Transformation
print(df.head())
```

```
                                      CompanyName  Year  \
0     Zohour Al -Sham (Hossam Hisham Al -Makhlalati)  2013
1  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
2  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
3  Zidane for ready -made clothes (Ibrahim Jamil ...  2013
4  Zidane for ready -made clothes (Ibrahim Jamil ...  2013

              Investment_type        sector                    Sub_sector  \
0  S 8. An internal investment  Agricultural  Land reclamation and farming
1  S 8. An internal investment    Industrial                      Textile
2  S 8. An internal investment    Industrial                      Textile
3  S 8. An internal investment    Industrial                      Textile
4  S 8. An internal investment    Industrial                      Textile

        Activity_classification    Legal_form Headquarters_location   Capital  \
0  Land reclamation and farming  Individually                  Giza       NaN
1                    Ready-made  Individually             Qaliubiya -0.370902
2                    Ready-made  Individually             Qaliubiya -0.370902
3                    Ready-made  Individually             Qaliubiya -0.370902
4                    Ready-made  Individually             Qaliubiya -0.370902

   Total_flows  Total_investment
0    -1.093272         -0.414267
1     0.345485          1.378114
2     0.345485          1.132898
3     0.345485         -0.918498
4     0.345485         -5.199338
```

```
df[columns_of_interest].hist()
plt.show()
```



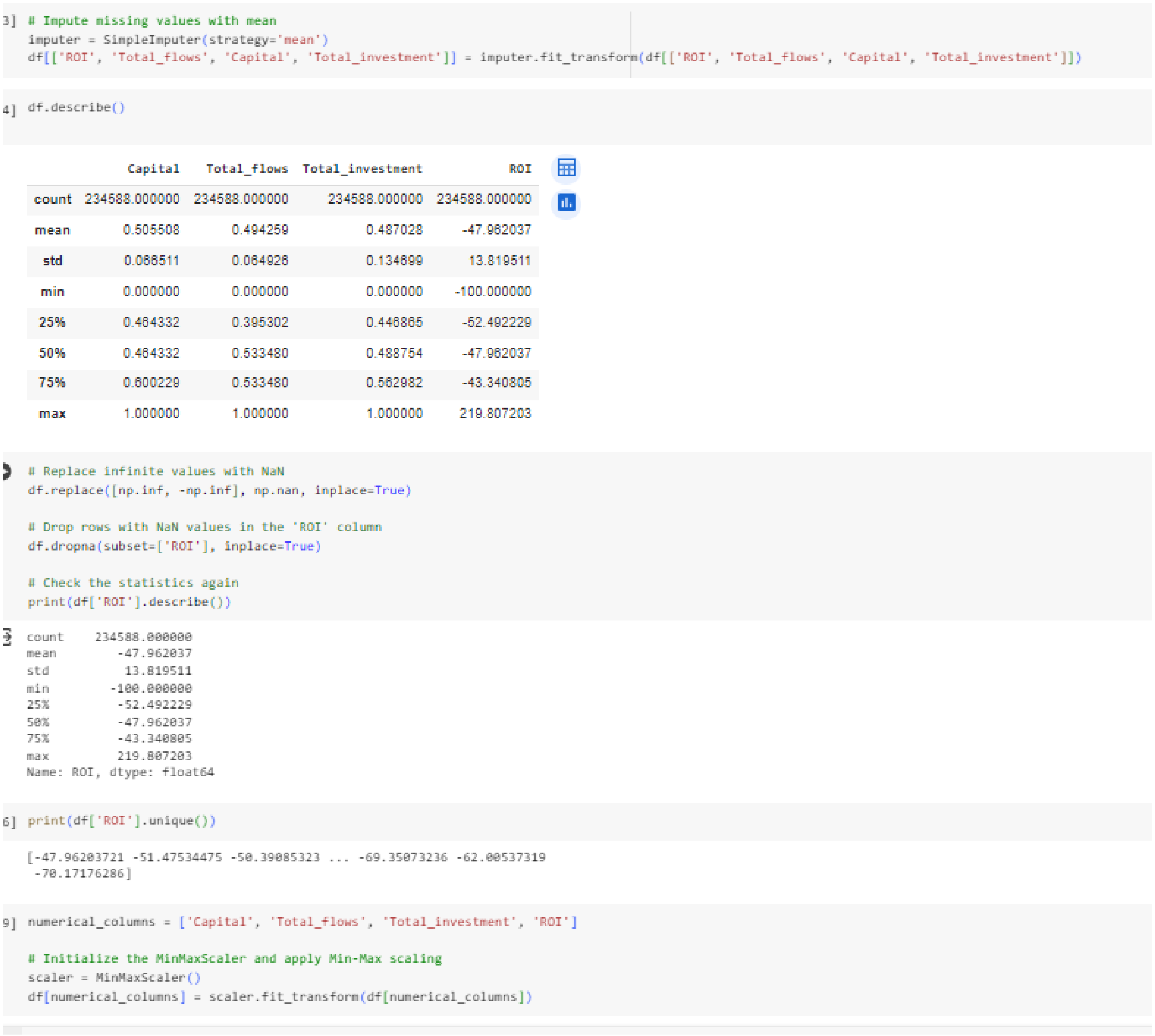
```
39] numerical_columns = ['Capital', 'Total_flows', 'Total_investment']

    # Initialize the MinMaxScaler and apply Min-Max scaling
    scaler = MinMaxScaler()
    df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

```
40] # Calculate ROI
    df['ROI'] = ((df['Total_flows'] - (df['Capital'] + df['Total_investment'])) / (df['Capital'] + df['Total_investment'])) * 100
```

```
# Calculate correlation matrix
correlation_matrix = df[['ROI', 'Total_flows', 'Capital', 'Total_investment']].corr()

# Visualize correlation matrix
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```

```
# Check for missing values
missing_values = df.isnull().sum()
print(missing_values)
```

```
CompanyName                   0
Year                          0
Investment_type               0
sector                        0
Sub_sector                    0
Activity_classification       0
Legal_form                    0
Headquarters_location         0
Capital                    7709
Total_flows                5801
Total_investment          22723
ROI                       34706
dtype: int64
```



```
3] # Impute missing values with mean
   imputer = SimpleImputer(strategy='mean')
   df[['ROI', 'Total_flows', 'Capital', 'Total_investment']] = imputer.fit_transform(df[['ROI', 'Total_flows', 'Capital', 'Total_investment']])
```

```
4] df.describe()
```

	Capital	Total_flows	Total_investment	ROI
count	234588.000000	234588.000000	234588.000000	234588.000000
mean	0.505508	0.494259	0.487028	-47.962037
std	0.066511	0.064926	0.134699	13.819511
min	0.000000	0.000000	0.000000	-100.000000
25%	0.464332	0.395302	0.446865	-52.492229
50%	0.464332	0.533480	0.488754	-47.962037
75%	0.600229	0.533480	0.562982	-43.340805
max	1.000000	1.000000	1.000000	219.807203

```
# Replace infinite values with NaN
df.replace([np.inf, -np.inf], np.nan, inplace=True)

# Drop rows with NaN values in the 'ROI' column
df.dropna(subset=['ROI'], inplace=True)

# Check the statistics again
print(df['ROI'].describe())
```

```
count    234588.000000
mean        -47.962037
std          13.819511
min        -100.000000
25%         -52.492229
50%         -47.962037
75%         -43.340805
max         219.807203
Name: ROI, dtype: float64
```

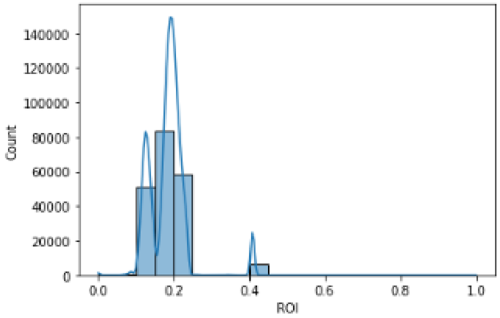
```
6] print(df['ROI'].unique())
```

```
[-47.96203721 -51.47534475 -50.39085323 ... -69.35073236 -62.00537319
 -70.17176286]
```

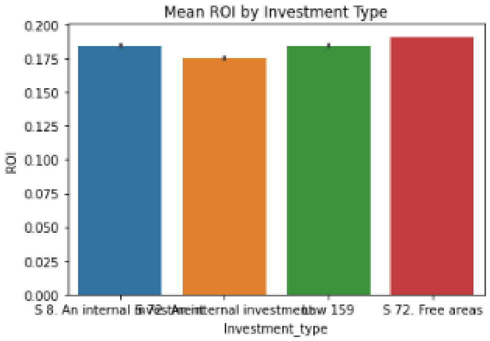
```
9] numerical_columns = ['Capital', 'Total_flows', 'Total_investment', 'ROI']

   # Initialize the MinMaxScaler and apply Min-Max scaling
   scaler = MinMaxScaler()
   df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

```
sns.histplot(df['ROI'], kde=True, bins=20)
```

```
<Axes: xlabel='ROI', ylabel='Count'>
```



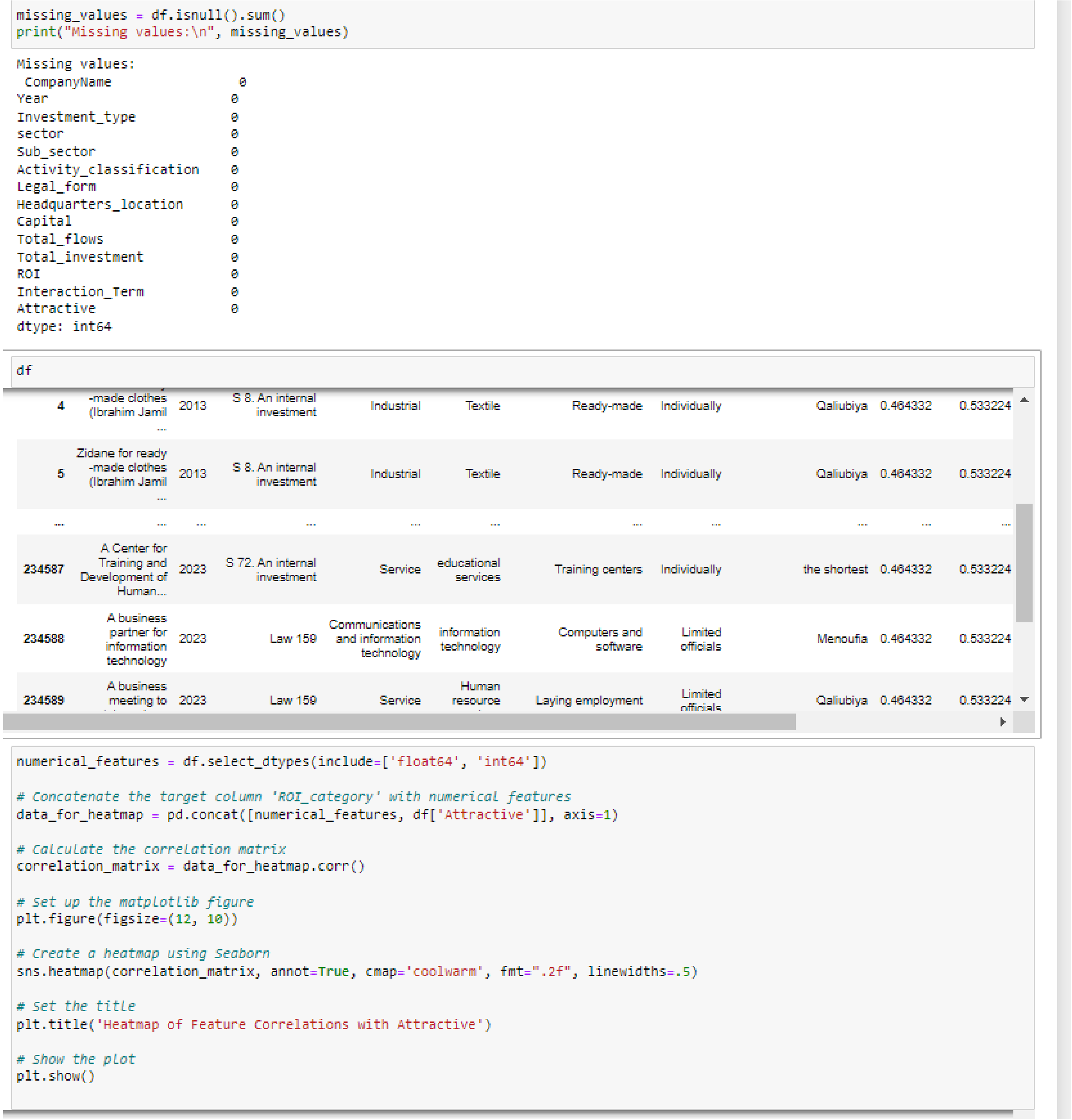
```
categorical_columns = ['CompanyName','Investment_type', 'sector', 'Sub_sector', 'Activity_classification', 'Legal_form', 'Headqua

# Group data by categorical columns and visualize the mean ROI for each group
mean_roi_by_category = df.groupby(categorical_columns)['ROI'].mean().reset_index()

# Visualize the mean ROI using a barplot
sns.barplot(x='Investment_type', y='ROI', data=mean_roi_by_category)
plt.title('Mean ROI by Investment Type')
plt.show()
```



```
# Calculate the 25th percentile and 75th percentile of ROI
percentile_25 = df['ROI'].quantile(0.25)
percentile_75 = df['ROI'].quantile(0.75)

# Create a binary column 'Attractive' based on percentiles
df['Attractive'] = ((df['ROI'] >= percentile_25) & (df['ROI'] <= percentile_75)).astype(int)
```

```
print(df['Attractive'].unique())
```

```
[0 1]
```



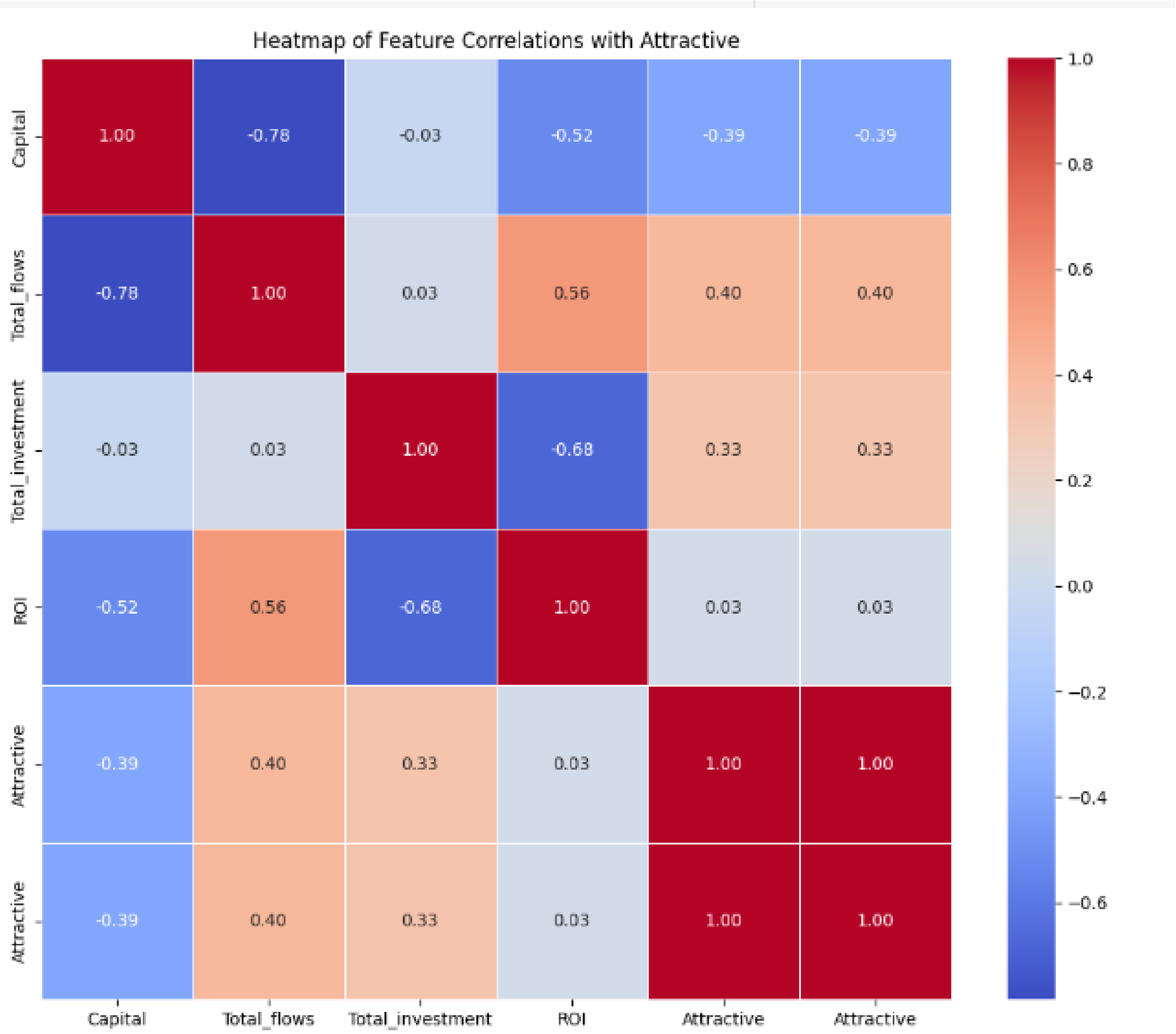
```
missing_values = df.isnull().sum()
print("Missing values:\n", missing_values)
```

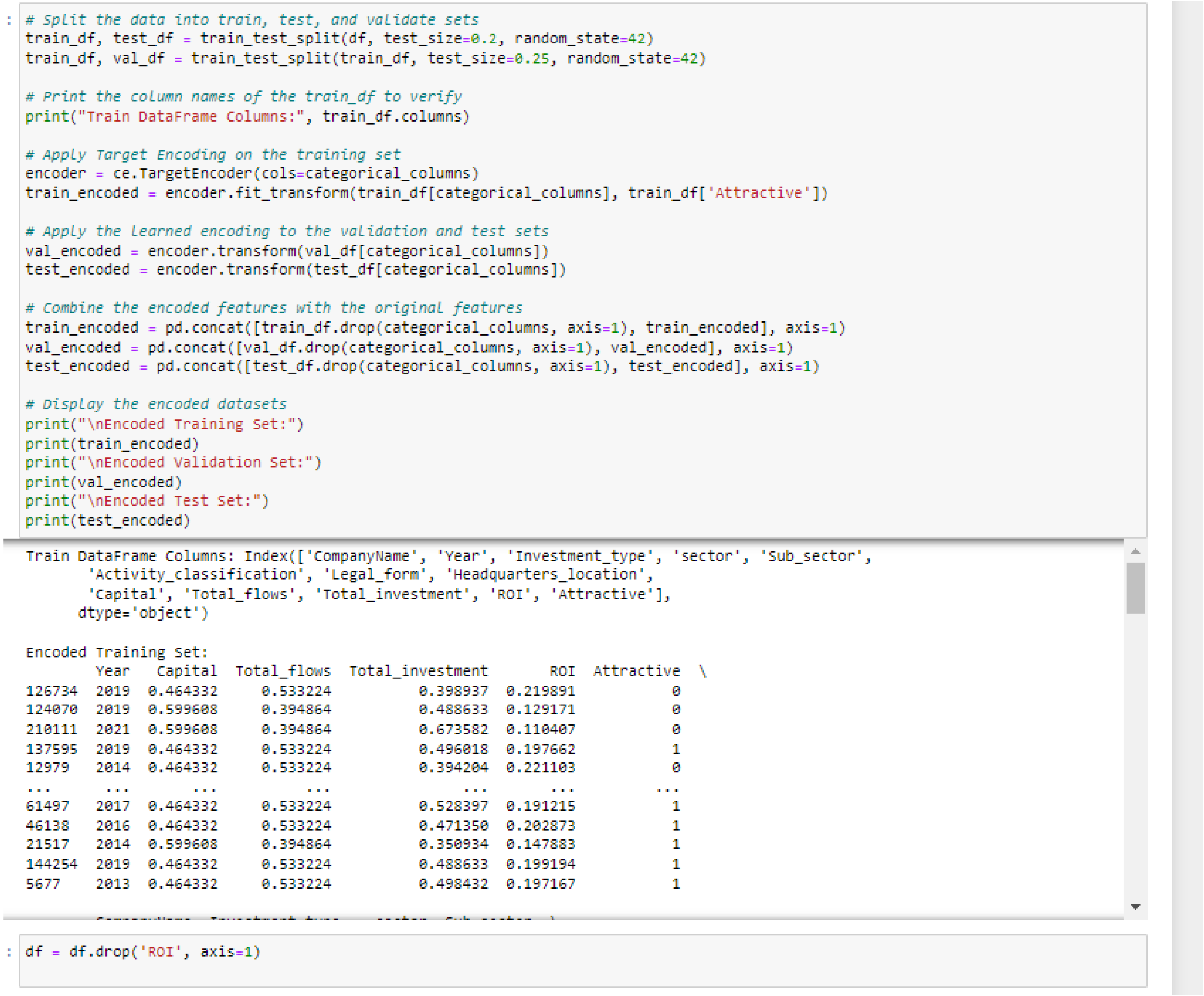
```
Missing values:
 CompanyName                0
Year                      0
Investment_type           0
sector                    0
Sub_sector                0
Activity_classification   0
Legal_form                0
Headquarters_location     0
Capital                   0
Total_flows               0
Total_investment          0
ROI                       0
Interaction_Term          0
Attractive                0
dtype: int64
```

```
df
```

4	-made clothes (Ibrahim Jamil ...	2013	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	0.464332	0.533224
5	Zidane for ready -made clothes (Ibrahim Jamil ...	2013	S 8. An internal investment	Industrial	Textile	Ready-made	Individually	Qaliubiya	0.464332	0.533224
...	...	...	...	...	...	...	...	...	...	...
234587	A Center for Training and Development of Human...	2023	S 72. An internal investment	Service	educational services	Training centers	Individually	the shortest	0.464332	0.533224
234588	A business partner for information technology	2023	Law 159	Communications and information technology	information technology	Computers and software	Limited officials	Menoufia	0.464332	0.533224
234589	A business meeting to	2023	Law 159	Service	Human resource	Laying employment	Limited officials	Qaliubiya	0.464332	0.533224

```
numerical_features = df.select_dtypes(include=['float64', 'int64'])

# Concatenate the target column 'ROI_category' with numerical features
data_for_heatmap = pd.concat([numerical_features, df['Attractive']], axis=1)

# Calculate the correlation matrix
correlation_matrix = data_for_heatmap.corr()

# Set up the matplotlib figure
plt.figure(figsize=(12, 10))

# Create a heatmap using Seaborn
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)

# Set the title
plt.title('Heatmap of Feature Correlations with Attractive')

# Show the plot
plt.show()
```

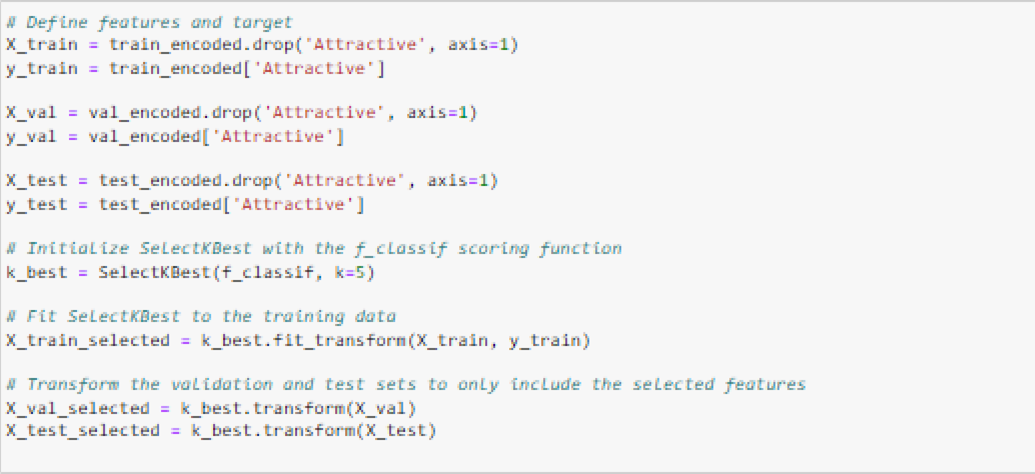

Heatmap of Feature Correlations with Attractive
Capital
Total_flows
Total_investment
ROI
Attractive
Attractive
1.00
-0.78
-0.03
-0.52
-0.39
-0.39
-0.78
1.00
0.03
0.56
0.40
0.40
-0.03
0.03
1.00
-0.68
0.33
0.33
-0.52
0.56
-0.68
1.00
0.03
0.03
-0.39
0.40
0.33
0.03
1.00
1.00
-0.39
0.40
0.33
0.03
1.00
1.00
Capital
Total_flows
Total_investment
ROI
Attractive
Attractive
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6



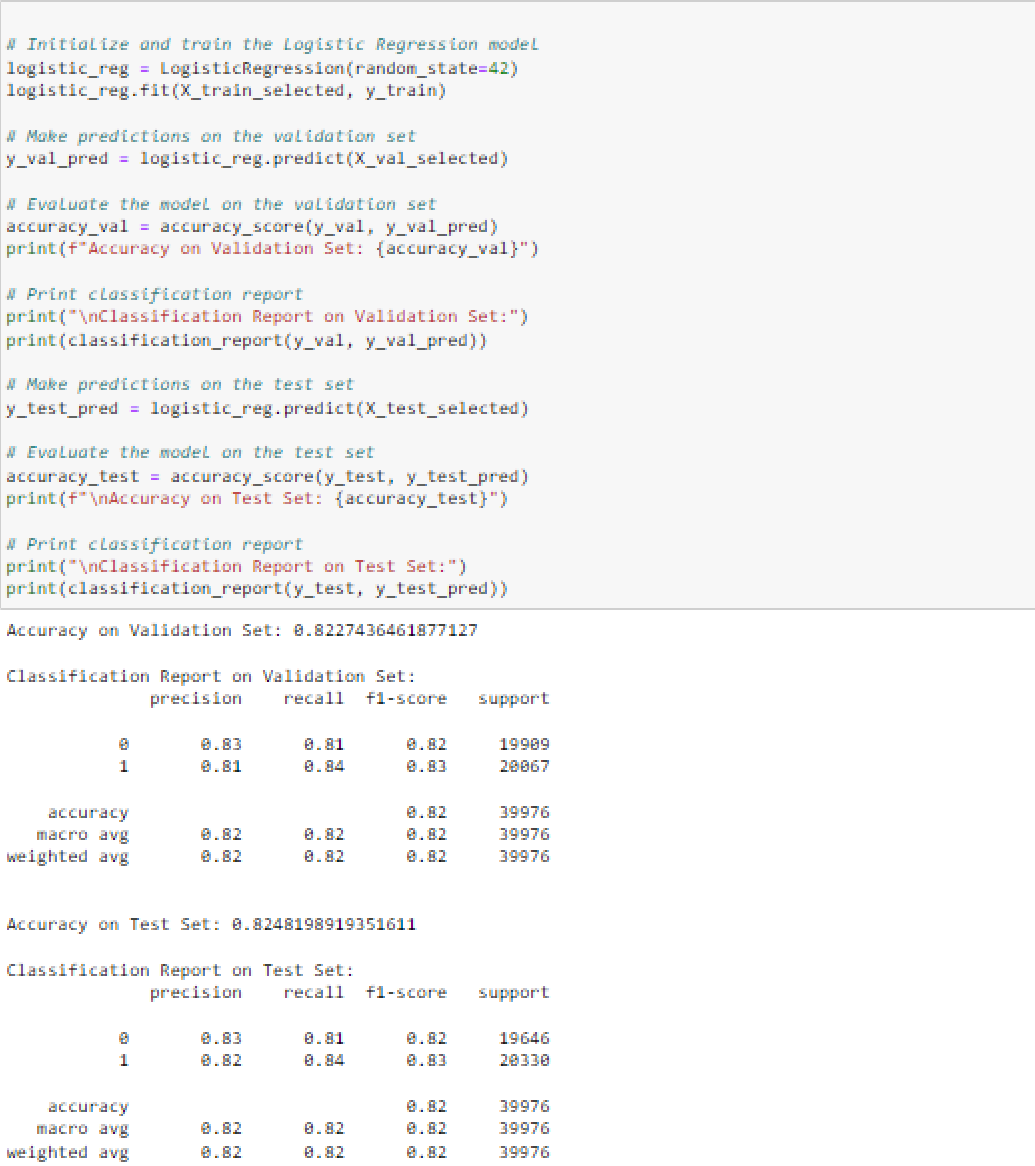
```
# Split the data into train, test, and validate sets
train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)
train_df, val_df = train_test_split(train_df, test_size=0.25, random_state=42)

# Print the column names of the train_df to verify
print("Train DataFrame Columns:", train_df.columns)

# Apply Target Encoding on the training set
encoder = ce.TargetEncoder(cols=categorical_columns)
train_encoded = encoder.fit_transform(train_df[categorical_columns], train_df['Attractive'])

# Apply the learned encoding to the validation and test sets
val_encoded = encoder.transform(val_df[categorical_columns])
test_encoded = encoder.transform(test_df[categorical_columns])

# Combine the encoded features with the original features
train_encoded = pd.concat([train_df.drop(categorical_columns, axis=1), train_encoded], axis=1)
val_encoded = pd.concat([val_df.drop(categorical_columns, axis=1), val_encoded], axis=1)
test_encoded = pd.concat([test_df.drop(categorical_columns, axis=1), test_encoded], axis=1)

# Display the encoded datasets
print("\nEncoded Training Set:")
print(train_encoded)
print("\nEncoded Validation Set:")
print(val_encoded)
print("\nEncoded Test Set:")
print(test_encoded)
```

```
Train DataFrame Columns: Index(['CompanyName', 'Year', 'Investment_type', 'sector', 'Sub_sector',
       'Activity_classification', 'Legal_form', 'Headquarters_location',
       'Capital', 'Total_flows', 'Total_investment', 'ROI', 'Attractive'],
      dtype='object')

Encoded Training Set:
        Year   Capital  Total_flows  Total_investment       ROI  Attractive  \
126734  2019  0.464332     0.533224          0.398937  0.219891           0
124070  2019  0.599608     0.394864          0.488633  0.129171           0
210111  2021  0.599608     0.394864          0.673582  0.110407           0
137595  2019  0.464332     0.533224          0.496018  0.197662           1
12979   2014  0.464332     0.533224          0.394204  0.221103           0
...      ...       ...          ...               ...       ...         ...
61497   2017  0.464332     0.533224          0.528397  0.191215           1
46138   2016  0.464332     0.533224          0.471350  0.202873           1
21517   2014  0.599608     0.394864          0.350934  0.147883           1
144254  2019  0.464332     0.533224          0.488633  0.199194           1
5677    2013  0.464332     0.533224          0.498432  0.197167           1

        CompanyName  Investment_type    sector  Sub_sector  \
```

```
df = df.drop('ROI', axis=1)
```

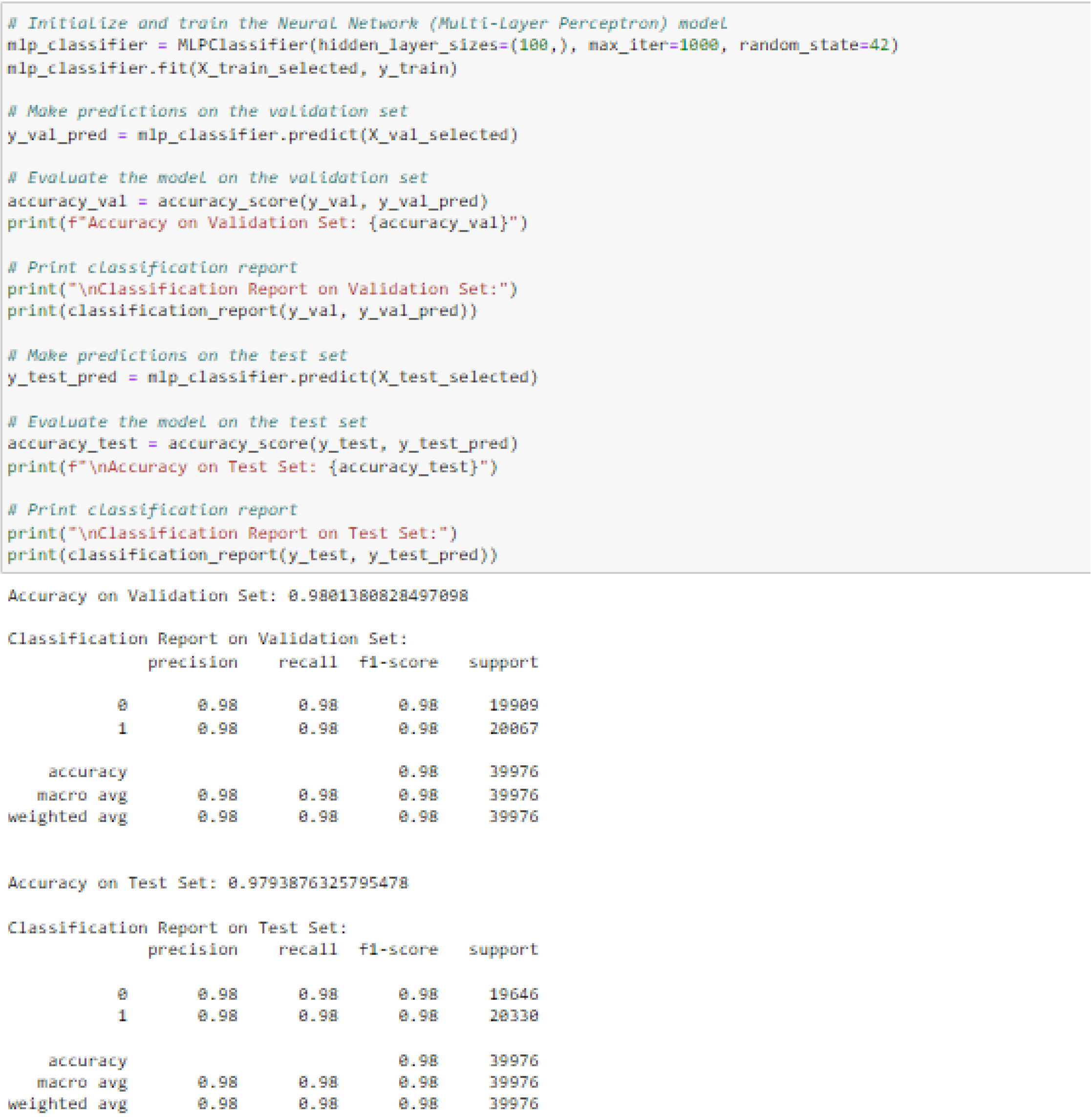


```python
# Define features and target
X_train = train_encoded.drop('Attractive', axis=1)
y_train = train_encoded['Attractive']

X_val = val_encoded.drop('Attractive', axis=1)
y_val = val_encoded['Attractive']

X_test = test_encoded.drop('Attractive', axis=1)
y_test = test_encoded['Attractive']

# Initialize SelectKBest with the f_classif scoring function
k_best = SelectKBest(f_classif, k=5)

# Fit SelectKBest to the training data
X_train_selected = k_best.fit_transform(X_train, y_train)

# Transform the validation and test sets to only include the selected features
X_val_selected = k_best.transform(X_val)
X_test_selected = k_best.transform(X_test)
```



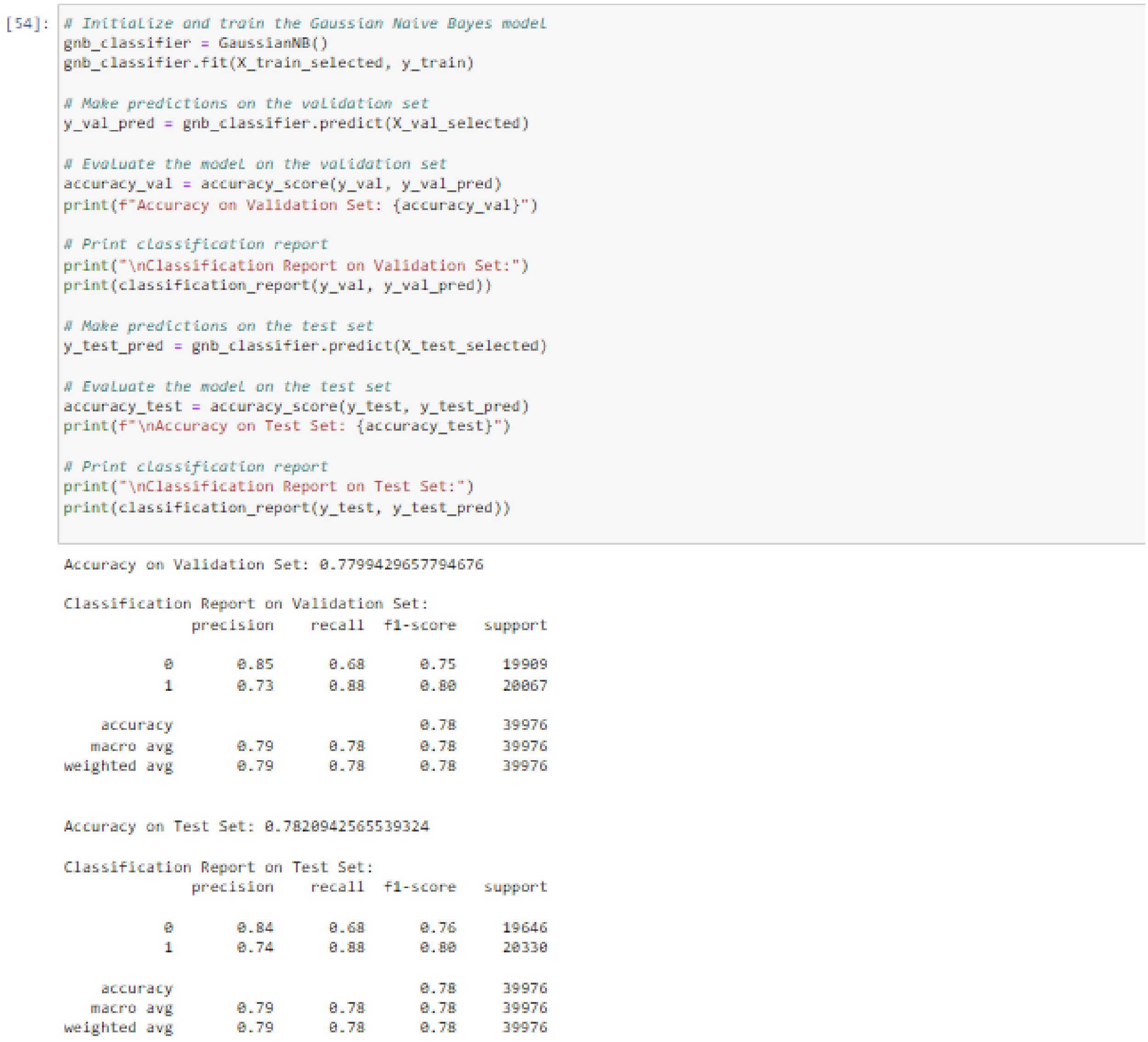
```python
# Initialize and train the Logistic Regression model
logistic_reg = LogisticRegression(random_state=42)
logistic_reg.fit(X_train_selected, y_train)

# Make predictions on the validation set
y_val_pred = logistic_reg.predict(X_val_selected)

# Evaluate the model on the validation set
accuracy_val = accuracy_score(y_val, y_val_pred)
print(f"Accuracy on Validation Set: {accuracy_val}")

# Print classification report
print("\nClassification Report on Validation Set:")
print(classification_report(y_val, y_val_pred))

# Make predictions on the test set
y_test_pred = logistic_reg.predict(X_test_selected)

# Evaluate the model on the test set
accuracy_test = accuracy_score(y_test, y_test_pred)
print(f"\nAccuracy on Test Set: {accuracy_test}")

# Print classification report
print("\nClassification Report on Test Set:")
print(classification_report(y_test, y_test_pred))
```

```
Accuracy on Validation Set: 0.8227436461877127

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.83      0.81      0.82     19909
           1       0.81      0.84      0.83     20067

    accuracy                           0.82     39976
   macro avg       0.82      0.82      0.82     39976
weighted avg       0.82      0.82      0.82     39976


Accuracy on Test Set: 0.8248198919351611

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.83      0.81      0.82     19646
           1       0.82      0.84      0.83     20330

    accuracy                           0.82     39976
   macro avg       0.82      0.82      0.82     39976
weighted avg       0.82      0.82      0.82     39976
```



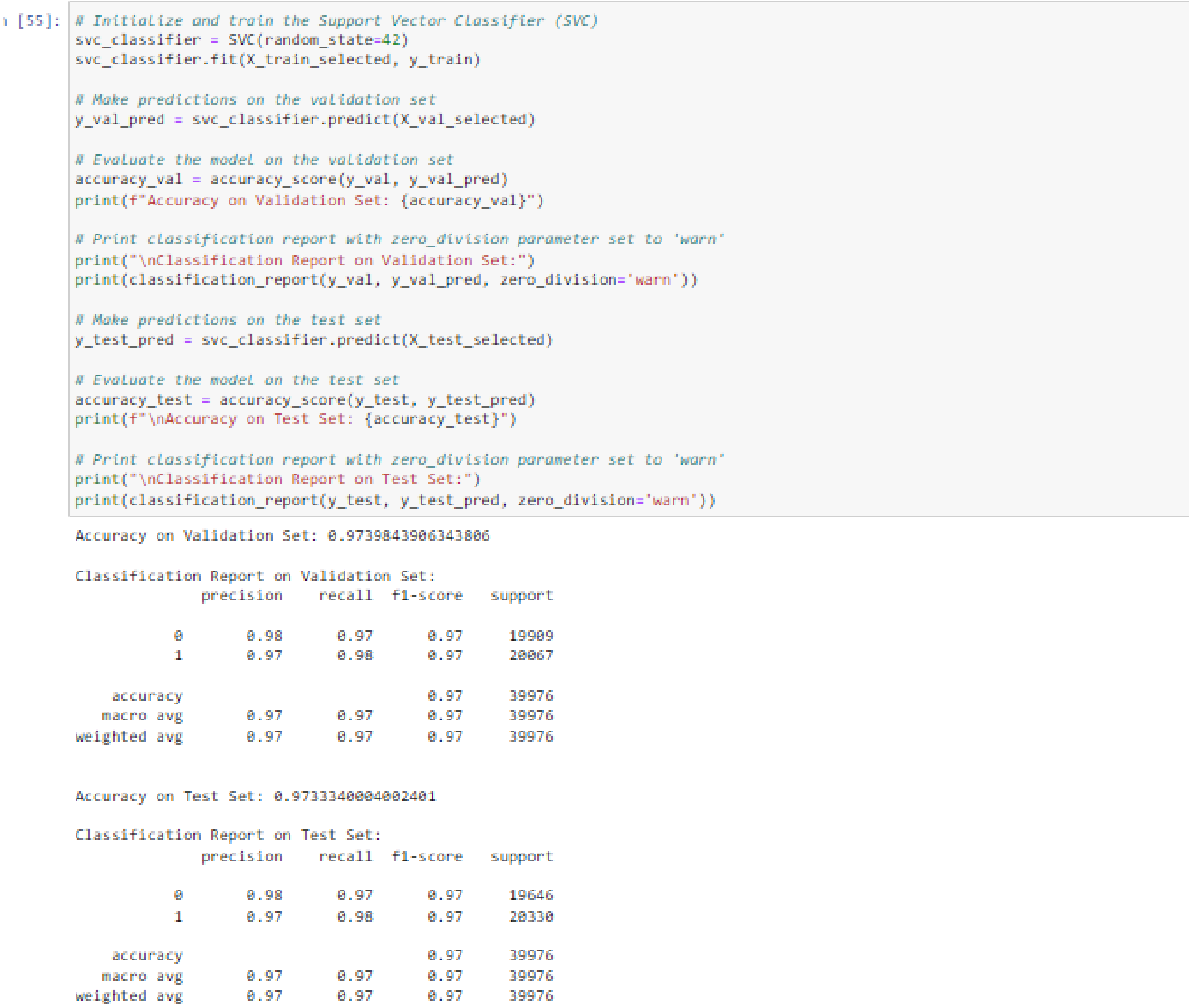
```python
# Initialize and train the Neural Network (Multi-Layer Perceptron) model
mlp_classifier = MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
mlp_classifier.fit(X_train_selected, y_train)

# Make predictions on the validation set
y_val_pred = mlp_classifier.predict(X_val_selected)

# Evaluate the model on the validation set
accuracy_val = accuracy_score(y_val, y_val_pred)
print(f"Accuracy on Validation Set: {accuracy_val}")

# Print classification report
print("\nClassification Report on Validation Set:")
print(classification_report(y_val, y_val_pred))

# Make predictions on the test set
y_test_pred = mlp_classifier.predict(X_test_selected)

# Evaluate the model on the test set
accuracy_test = accuracy_score(y_test, y_test_pred)
print(f"\nAccuracy on Test Set: {accuracy_test}")

# Print classification report
print("\nClassification Report on Test Set:")
print(classification_report(y_test, y_test_pred))
```

```
Accuracy on Validation Set: 0.9801380828497098

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     19909
           1       0.98      0.98      0.98     20067

    accuracy                           0.98     39976
   macro avg       0.98      0.98      0.98     39976
weighted avg       0.98      0.98      0.98     39976


Accuracy on Test Set: 0.9793876325795478

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     19646
           1       0.98      0.98      0.98     20330

    accuracy                           0.98     39976
   macro avg       0.98      0.98      0.98     39976
weighted avg       0.98      0.98      0.98     39976
```



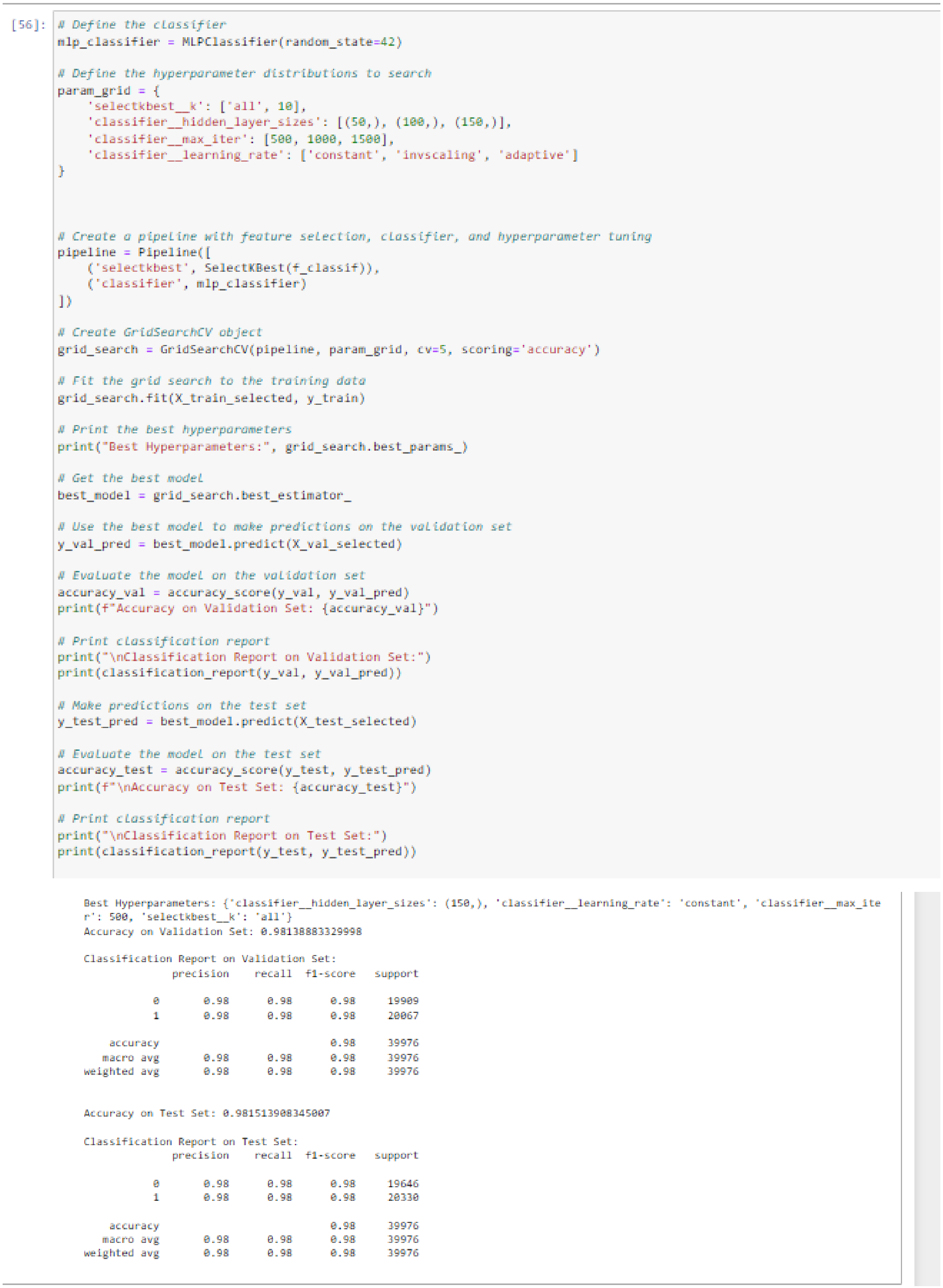
```
[54]: # Initialize and train the Gaussian Naive Bayes model
      gnb_classifier = GaussianNB()
      gnb_classifier.fit(X_train_selected, y_train)

      # Make predictions on the validation set
      y_val_pred = gnb_classifier.predict(X_val_selected)

      # Evaluate the model on the validation set
      accuracy_val = accuracy_score(y_val, y_val_pred)
      print(f"Accuracy on Validation Set: {accuracy_val}")

      # Print classification report
      print("\nClassification Report on Validation Set:")
      print(classification_report(y_val, y_val_pred))

      # Make predictions on the test set
      y_test_pred = gnb_classifier.predict(X_test_selected)

      # Evaluate the model on the test set
      accuracy_test = accuracy_score(y_test, y_test_pred)
      print(f"\nAccuracy on Test Set: {accuracy_test}")

      # Print classification report
      print("\nClassification Report on Test Set:")
      print(classification_report(y_test, y_test_pred))
```

```
Accuracy on Validation Set: 0.7799429657794676

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.85      0.68      0.75     19909
           1       0.73      0.88      0.80     20067

    accuracy                           0.78     39976
   macro avg       0.79      0.78      0.78     39976
weighted avg       0.79      0.78      0.78     39976


Accuracy on Test Set: 0.7820942565539324

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.84      0.68      0.76     19646
           1       0.74      0.88      0.80     20330

    accuracy                           0.78     39976
   macro avg       0.79      0.78      0.78     39976
weighted avg       0.79      0.78      0.78     39976
```



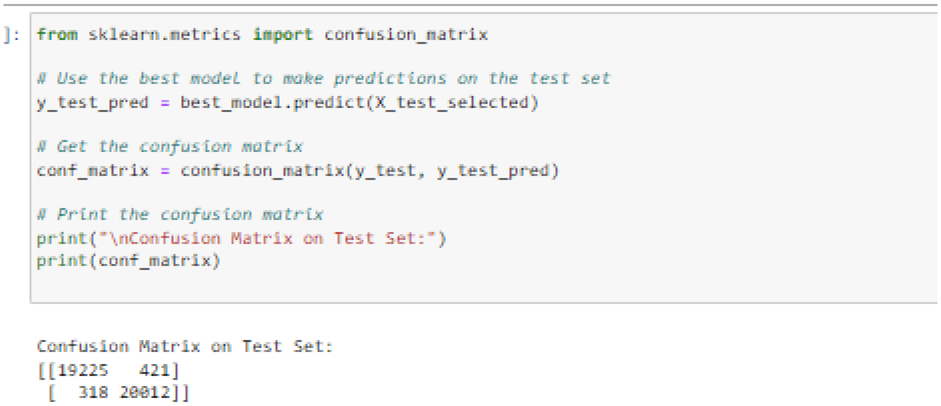
```
[55]: # Initialize and train the Support Vector Classifier (SVC)
      svc_classifier = SVC(random_state=42)
      svc_classifier.fit(X_train_selected, y_train)

      # Make predictions on the validation set
      y_val_pred = svc_classifier.predict(X_val_selected)

      # Evaluate the model on the validation set
      accuracy_val = accuracy_score(y_val, y_val_pred)
      print(f"Accuracy on Validation Set: {accuracy_val}")

      # Print classification report with zero_division parameter set to 'warn'
      print("\nClassification Report on Validation Set:")
      print(classification_report(y_val, y_val_pred, zero_division='warn'))

      # Make predictions on the test set
      y_test_pred = svc_classifier.predict(X_test_selected)

      # Evaluate the model on the test set
      accuracy_test = accuracy_score(y_test, y_test_pred)
      print(f"\nAccuracy on Test Set: {accuracy_test}")

      # Print classification report with zero_division parameter set to 'warn'
      print("\nClassification Report on Test Set:")
      print(classification_report(y_test, y_test_pred, zero_division='warn'))
```

```
Accuracy on Validation Set: 0.9739843906343806

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.98      0.97      0.97     19909
           1       0.97      0.98      0.97     20067

    accuracy                           0.97     39976
   macro avg       0.97      0.97      0.97     39976
weighted avg       0.97      0.97      0.97     39976


Accuracy on Test Set: 0.9733340004002401

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.98      0.97      0.97     19646
           1       0.97      0.98      0.97     20330

    accuracy                           0.97     39976
   macro avg       0.97      0.97      0.97     39976
weighted avg       0.97      0.97      0.97     39976
```



```
[56]: # Define the classifier
mlp_classifier = MLPClassifier(random_state=42)

# Define the hyperparameter distributions to search
param_grid = {
    'selectkbest__k': ['all', 10],
    'classifier__hidden_layer_sizes': [(50,), (100,), (150,)],
    'classifier__max_iter': [500, 1000, 1500],
    'classifier__learning_rate': ['constant', 'invscaling', 'adaptive']
}



# Create a pipeline with feature selection, classifier, and hyperparameter tuning
pipeline = Pipeline([
    ('selectkbest', SelectKBest(f_classif)),
    ('classifier', mlp_classifier)
])

# Create GridSearchCV object
grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='accuracy')

# Fit the grid search to the training data
grid_search.fit(X_train_selected, y_train)

# Print the best hyperparameters
print("Best Hyperparameters:", grid_search.best_params_)

# Get the best model
best_model = grid_search.best_estimator_

# Use the best model to make predictions on the validation set
y_val_pred = best_model.predict(X_val_selected)

# Evaluate the model on the validation set
accuracy_val = accuracy_score(y_val, y_val_pred)
print(f"Accuracy on Validation Set: {accuracy_val}")

# Print classification report
print("\nClassification Report on Validation Set:")
print(classification_report(y_val, y_val_pred))

# Make predictions on the test set
y_test_pred = best_model.predict(X_test_selected)

# Evaluate the model on the test set
accuracy_test = accuracy_score(y_test, y_test_pred)
print(f"\nAccuracy on Test Set: {accuracy_test}")

# Print classification report
print("\nClassification Report on Test Set:")
print(classification_report(y_test, y_test_pred))
```

```
Best Hyperparameters: {'classifier__hidden_layer_sizes': (150,), 'classifier__learning_rate': 'constant', 'classifier__max_ite
r': 500, 'selectkbest__k': 'all'}
Accuracy on Validation Set: 0.98138883329998

Classification Report on Validation Set:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     19909
           1       0.98      0.98      0.98     20067

    accuracy                           0.98     39976
   macro avg       0.98      0.98      0.98     39976
weighted avg       0.98      0.98      0.98     39976


Accuracy on Test Set: 0.981513908345007

Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     19646
           1       0.98      0.98      0.98     20330

    accuracy                           0.98     39976
   macro avg       0.98      0.98      0.98     39976
weighted avg       0.98      0.98      0.98     39976
```



```
from sklearn.metrics import confusion_matrix

# Use the best model to make predictions on the test set
y_test_pred = best_model.predict(X_test_selected)

# Get the confusion matrix
conf_matrix = confusion_matrix(y_test, y_test_pred)

# Print the confusion matrix
print("\nConfusion Matrix on Test Set:")
print(conf_matrix)
```

```
Confusion Matrix on Test Set:
[[19225   421]
 [  318 20012]]
```