**Portuguese banking institution Dataset Analysis Report**

Exploratory data analysis (EDA) is about getting an overall understanding of data. EDA includes exploring data to find its main characteristics and identifying patterns and visualizations. EDA provides meaningful insights into data that can be used in a variety of applications, e.g., machine learning. Python can be effectively used to do EDA as it has a rich set of easy-to-use libraries like Pandas, Seaborn, Numpy, and Matplotlib. In this workshop, we will cover the basics of EDA using a real-world data set, including, but not limited to, correlating, converting, completing, correcting, creating, and charting the data. In addition, we will be using Jupyter Notebooks (an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text).

**Problem statement:**

A bank wants to promote its deposit-term rate with the client. I will be analyzing this data as a data analyst to find insights that can be used to promote term deposits effectively. The company is interested in the insights that would help them in their marketing campaigns.

**Introduction of data:**

This is related to direct marketing campaigns for a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required to determine if the product (a bank term deposit) would be (or not be) subscribed.

**Bank client data:**

1. Age: client's age, and it is a numerical value

2. Job: The kind of Job the Clients do; it is a categorical variable, like Admin, Management, Housemaid, Entrepreneur, Students, Blue-collar, Self-employed, Retired, Technician, Services and Unknown

3. Marital: The marital status of the clients is a categorical variable. i.e. Married, Divorced, Single; Note Divorced means divorced or widowed)

4. Education: it is categorical i.e. Unknown, Secondary, Primary, Tertiary)

5. Default: The client has credit in default? (Binary: Yes, No)

6. Balance: average yearly balance, in euros (numeric)

7. Housing: has a housing loan? (Binary: Yes, No)

8. Loan: Has a personal loan? (Binary: Yes, No )

**Related to the last contact of the current campaign:**

9. Contact: contact communication type (categorical: unknown, telephone cellular)

10. Day: last contact day of the month (numeric)

11. Month: last contact month of the year (categorical: Jan, Feb, Mar, …, Nov, Dec)

12. Duration: last contact duration, in seconds (numeric)

**Other attributes of the current campaign:**

13. Campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact)

14. Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means the client was not previously contacted)

15. Previous: number of contacts performed before this campaign and for this client (numeric)

16. Poutcome: outcome of the previous marketing campaign (categorical: unknown, other, failure, success)

**Output variable (desired target):**

17. y: has the client subscribed to a term deposit? (binary: yes, no)

**Importing Python Libraries Package:**

I installed libraries to perform different operations on data. Let me explain these libraries a bit:

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with “relational” or “labelled” data both easy and intuitive.

Numpy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Matplotlib: Matplotlib is a Python 2D plotting library that produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

A screenshot of a computer program

AI-generated content may be incorrect.

Image 1. Screenshot of the python libraries imported. Image by Author

**Importing Data into Data Frames:**

To start working with data, first we need to import data from CSV files into a Pandas DataFrame. This will be done using Pandas’ read\_csv method. We will further learn how different delimiters are used by this function.

We need to make sense of our data in what it looks like. We will explore some more pandas’ functions here like

● See data in tabular form by using. head() to view the top five samples and. tail() to view the last 5 samples.

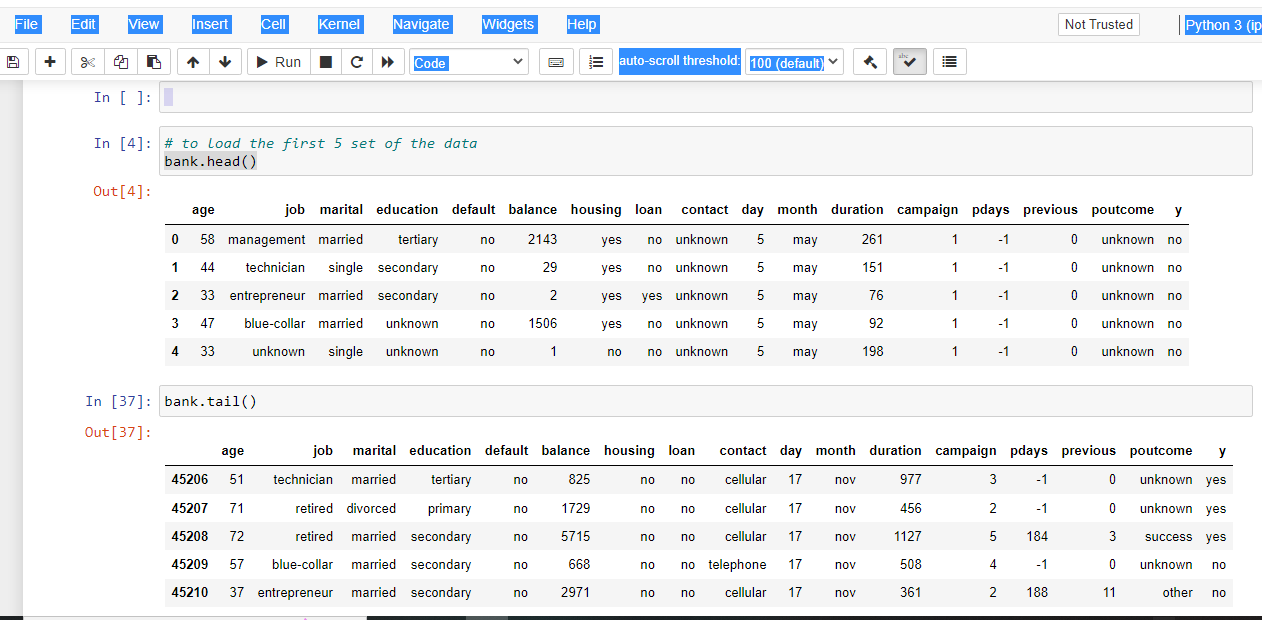


Image 2. Screenshot of the top 5 samples and last 5 samples. Image by author

● Overall summary of DataFrame using .info( )

A screenshot of a computer

AI-generated content may be incorrect.

Image 3. Screenshot of statistical information of the data. source by me

The information helps us know the type of dataset and general information. In this bank data, there are 7 numerical variables and 9 categorical variables.

**Data Cleaning**

I want to find out if there are any missing (null) values in columns. Check using Pandas’ isnull. and alsoduplicate values.

**ANALYSIS**

Our Analysis will be in three stages;

1. Univariate

2. Bi-variate

3. Multivariate

**Univariate Analysis:**

The univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own. Examples of our Univariate Analysis are count plots, boxplots, histograms, etc.

From this dataset

A graph of a number of people

AI-generated content may be incorrect.

Image by Author: Visualize the histogram of the client’s Age

In this graph, the dashed circle shows the range of most customers’ ages, which explains that most people in the bank are in the age bracket of 35–50.

A graph of sales

AI-generated content may be incorrect.

Image by author

This graph shows that the most common job categories are blue-collar and management jobs. A blue-collar worker is a person from the working class who performs manual labour. Examples are farmers, mechanics, power plant operators, etc. They are based on hourly wages.

A graph showing the number of customers

AI-generated content may be incorrect.

Image by author

It explains that most customers in the marketing campaign are secondary holders.

A graph of a number of customers

AI-generated content may be incorrect.

Image by author

Married customers are involved in the marketing campaign.

A blue and orange pie chart

AI-generated content may be incorrect.

Image by author

This graph illustrates that only 11.70 per cent of the customers subscribed to the deposit term offered by the marketing campaign.

In bivariate analysis, we will know the reason why and what the marketing campaign can do to promote the rate of subscription.

**Bi-variate Analysis:**

According to the definition, the bivariate analysis examines any concurrent relationship between two variables or qualities. This study explores the relationship between two variables as well as the depth of this relationship to figure out if there are any discrepancies between the two variables and any causes of these differences.

A graph of different colored columns

AI-generated content may be incorrect.

Image by author

This graph shows the relationship between deposit (the target variable) and the job category.

A graph showing the number of housing loan

AI-generated content may be incorrect.

Image by author

A graph with blue and orange squares

AI-generated content may be incorrect.

Image by author

This graph shows that clients with loans, either personal or housing, tend not to show interest in deposit subscriptions. Both charts show that customers require a housing loan mostly, and their loan is small.

A graph of a person and person

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

A graph showing a number of people

AI-generated content may be incorrect.

It shows that no matter the age or balance in euros of the client, it has the same patterns concerning the deposit subscription.

**Multivariate Analysis**

Multivariate analysis is based on the observation and analysis of more than one statistical outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

A graph of red and blue dots

AI-generated content may be incorrect.

Image by author

**Correlation**

Correlation is a statistical term describing the degree to which two variables move in coordination with one another.

There are two types of correlation

1. Spearman correlation

2. Pearson correlation

Spearman’s rank Correlation: Spearman’s rank correlation measures the strength and direction of association between two ranked variables. It gives the measure of the monotonicity of the relation between two variables.

Pearson Correlation: Pearson Correlation measures the strength of the linear relationship between two variables.

In this case, I will be using spearman’s rank correlation

A chart with different colored squares

AI-generated content may be incorrect.

Image by author

**Duration, previous, balance**, and **pdays**are positively correlated to deposit, while **credit default, personal loan**, and **housing loan** are negatively correlated to deposit.

**DASHBOARD**

The main use of a dashboard is **to show a comprehensive overview of data from different sources**. Dashboards are useful for monitoring, measuring, and analyzing relevant data in key areas. I will be using two Python libraries. Dash and plotly

**DASH**

Dash is **an open-source Python framework used for building analytical web applications**. It is a powerful library that simplifies the development of data-driven applications. It’s especially useful for Python data scientists who aren’t very familiar with web development.

**PLOTLY**

The Plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use cases.

A screenshot of a computer program

AI-generated content may be incorrect.

Image by author

A screenshot of a computer program

AI-generated content may be incorrect.

Image by author



Image by author

**INSIGHT and CONCLUSION.**

From the various chart above, the following can be derived;

1. Most of the clients that are involved in this marketing campaign are in the age range of 35–50.

2. Most of the client's job categories are blue-collar, managerial, and technician.

3. Since it is mostly blue-collar and technician job categories, this is the reason why their educational level is mostly secondary. A blue-collar worker is a manual worker, like a farmer. Plant operator etc

4. Clients or customers with loans, either housing or personal, tend not to subscribe to this deposit due to paying for their existing loan.

5. Both married and single people subscribe to the deposit terms.

6. Only 11.70 % of the bank customers subscribed to the deposit terms.

7. With or without a deposit term, the age follows the same pattern.

8. The client who subscribes to the deposit term is more likely to be married. This can be a result of the responsibility they have at hand.

9. Clients with low balances tend to subscribe more than those with high –balances.

The rate of deposit-term subscriptions decreases as people get older.

**RECOMMENDATION**

The company should focus more on youths aged 35–50.

The company should focus more on clients with high-income balances.

The company should shift its focus to the client without any form of a loan.