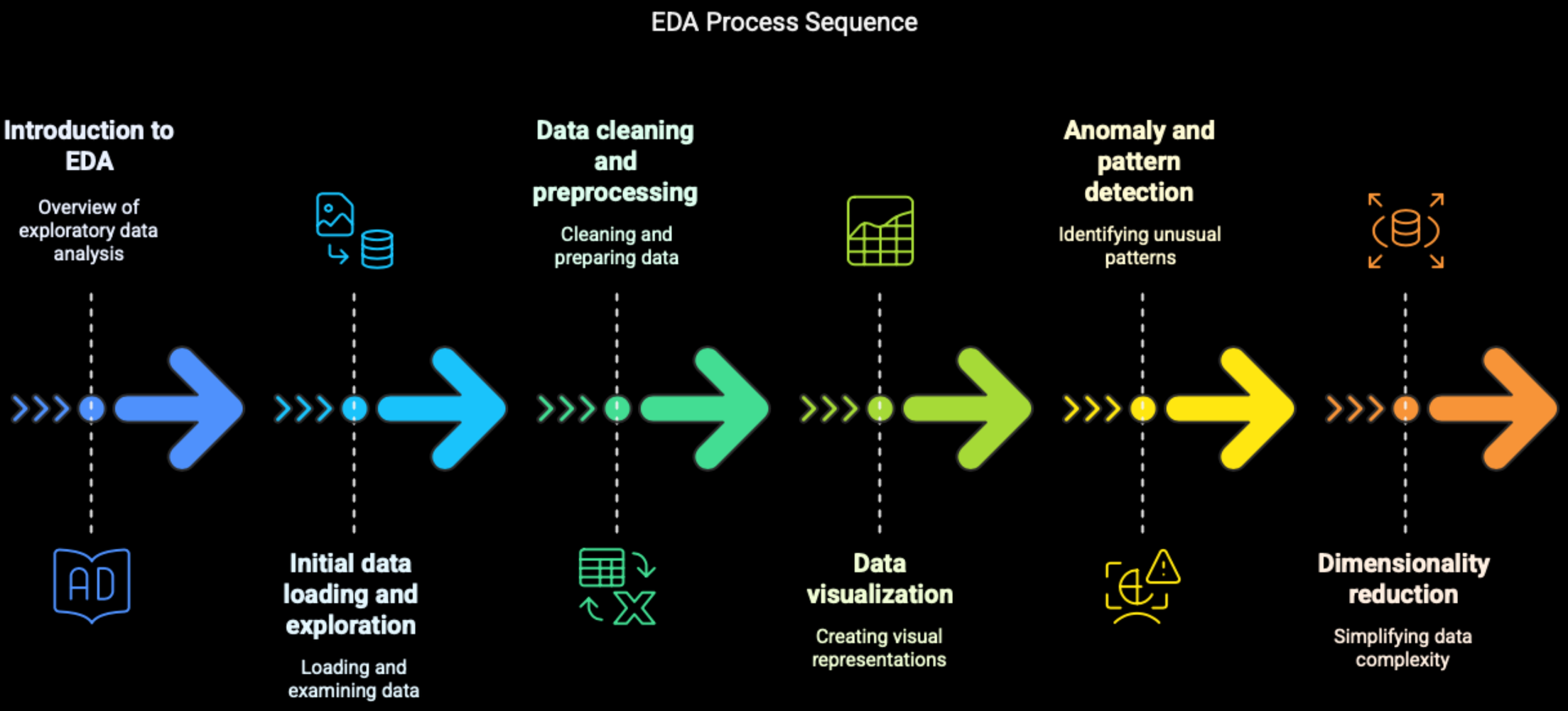
Data Analysis

Exploratory Data Analysis (EDA) With Python

Edgar Rios Linares

Roadmap


EDA Process Sequence
Introduction to EDA
Overview of exploratory data analysis
Initial data loading and exploration
Loading and examining data
Data cleaning and preprocessing
Cleaning and preparing data
Data visualization
Creating visual representations
Anomaly and pattern detection
Identifying unusual patterns
Dimensionality reduction
Simplifying data complexity

Introduction

Definition: Process that allows us to understand the structure and characteristics of a data set.

Purpose: Identify patterns, outliers, relationships between variables, and problems in the data before applying machine learning models.

Importance: Allows us to make informed decisions in the modeling phase and avoid biases or errors in the results.

Loading and exploration

Data types:

Numeric (discrete and continuous).

Categorical (nominal and ordinal).

Mixed (combination of numeric and categorical).

Data loading methods:

Using pandas (pd.read_csv(), pd.read_excel(), pd.read_sql()).

Structure verification with .info(), .head(), .describe().

Data cleaning and preprocessing

Handling missing values:

Elimination (dropna()).

Imputation with mean, etc.

Handling duplicates (drop_duplicates()).

Normalization and standardization:

MinMaxScaler to scale values between 0 and 1.

StandardScaler to normalize data with mean 0 and standard deviation 1.

Data visualization

Histograms:

Distribution of values of a variable.

Scatter diagrams:

Relationship between two variables.

Boxplots:

Detection of outliers.

Heat maps:

Correlation matrices

Anomaly and pattern detection

Identifying outliers:

Using standard deviation.

Using percentiles and interquartile range (IQR).

Analyzing distributions to find biases or unusual patterns.

Dimensionality reduction

Principal Component Analysis (PCA): Allows the number of variables to be reduced without losing relevant information.

Variable selection using filtering or grouping techniques.

Practical Example

Credit Risk – German Credit

1.Preprocessing

Read dataset

Load data

```
[1] # import libraries
    import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
#load dataset
dataset = pd.read_csv('german_credit_data.csv')
dataset.head(5)
```

	Unnamed: 0	age	sex	job	housing	saving_accounts	checking account	credit amount	duration	purpose	risk
0	0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	2	49	male	1	own	little	NaN	2096	12	education	good
3	3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	4	53	male	2	free	little	little	4870	24	car	bad

Practical Example

Credit Risk – German Credit

1.Preprocessing

Structure verification

```
[13] # Dimension

     dataset.shape
```

```
(1000, 11)
```

```
[4]  # info
     dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0        1000 non-null   int64
 1   age               1000 non-null   int64
 2   sex               1000 non-null   object
 3   job               1000 non-null   int64
 4   housing           1000 non-null   object
 5   saving_accounts   817 non-null    object
 6   checking account  606 non-null    object
 7   credit amount     1000 non-null   int64
 8   duration          1000 non-null   int64
 9   purpose           1000 non-null   object
 10  risk              1000 non-null   object
dtypes: int64(5), object(6)
memory usage: 86.1+ KB
```

Practical Example

Credit Risk – German Credit

1.Preprocessing

Structure verification

```
# statistical description

dataset.describe()
```

	Unnamed: 0	age	job	credit amount	duration
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	499.500000	35.546000	1.904000	3271.258000	20.903000
std	288.819436	11.375469	0.653614	2822.736876	12.058814
min	0.000000	19.000000	0.000000	250.000000	4.000000
25%	249.750000	27.000000	2.000000	1365.500000	12.000000
50%	499.500000	33.000000	2.000000	2319.500000	18.000000
75%	749.250000	42.000000	2.000000	3972.250000	24.000000
max	999.000000	75.000000	3.000000	18424.000000	72.000000

Practical Example

Credit Risk – German Credit

1.Preprocessing

Structure verification

```
# unique values

dataset.nunique()
```

	0
Unnamed: 0	1000
age	53
sex	2
job	4
housing	3
saving_accounts	4
checking account	3
credit amount	921
duration	33
purpose	8
risk	2

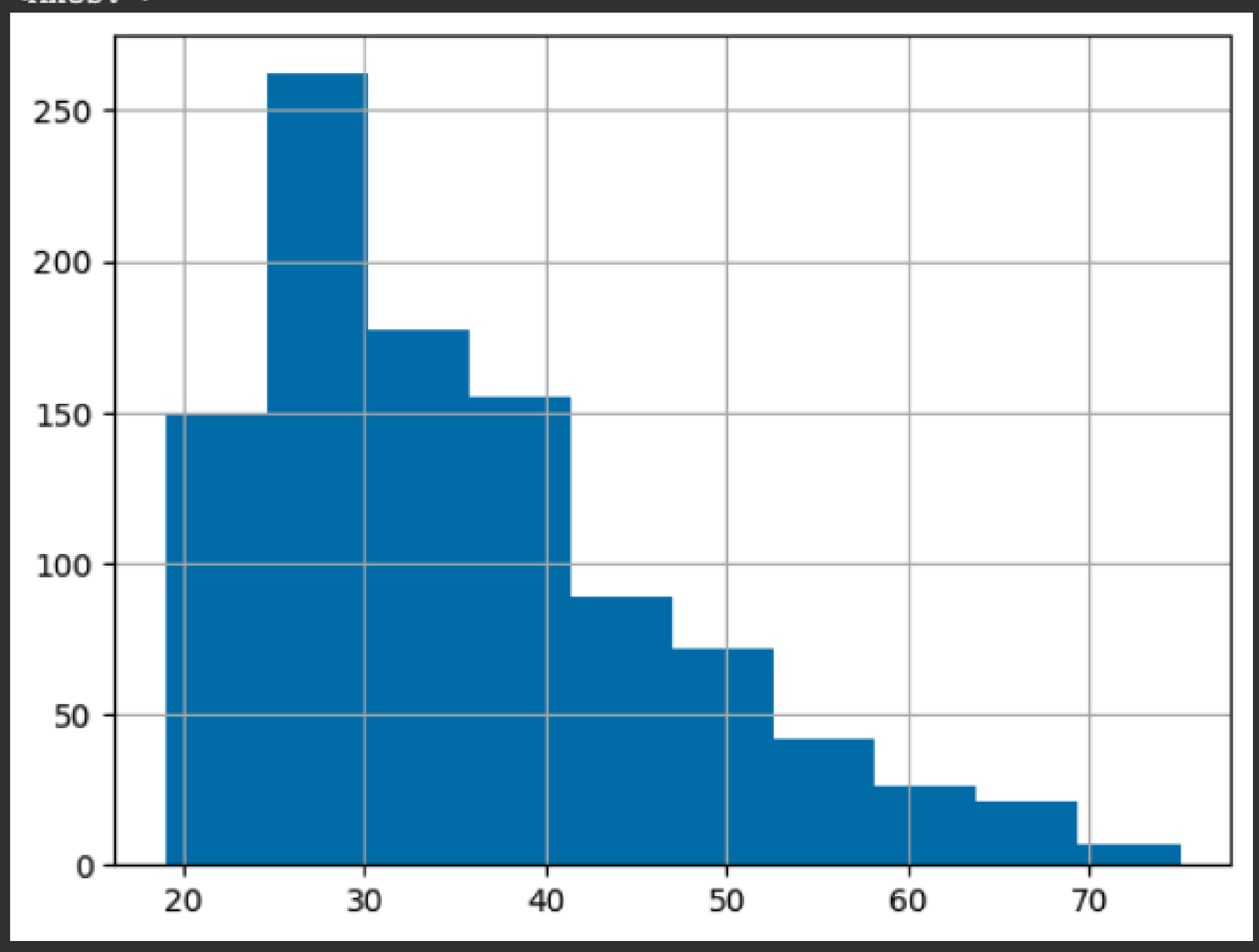
Practical Example

Credit Risk – German Credit

Data Visualization

```
#age histogram
dataset.age.hist()
```

```
<Axes: >
```



Practical Example

Credit Risk – German Credit

Data Visualization

```
#correlation matrix
corrmat = dataset.corr(numeric_only=True)
corrmat
```

	Unnamed: 0	age	job	credit amount	duration
Unnamed: 0	1.000000	-0.010096	-0.027345	0.013488	0.030788
age	-0.010096	1.000000	0.015673	0.032716	-0.036136
job	-0.027345	0.015673	1.000000	0.285385	0.210910
credit amount	0.013488	0.032716	0.285385	1.000000	0.624984
duration	0.030788	-0.036136	0.210910	0.624984	1.000000

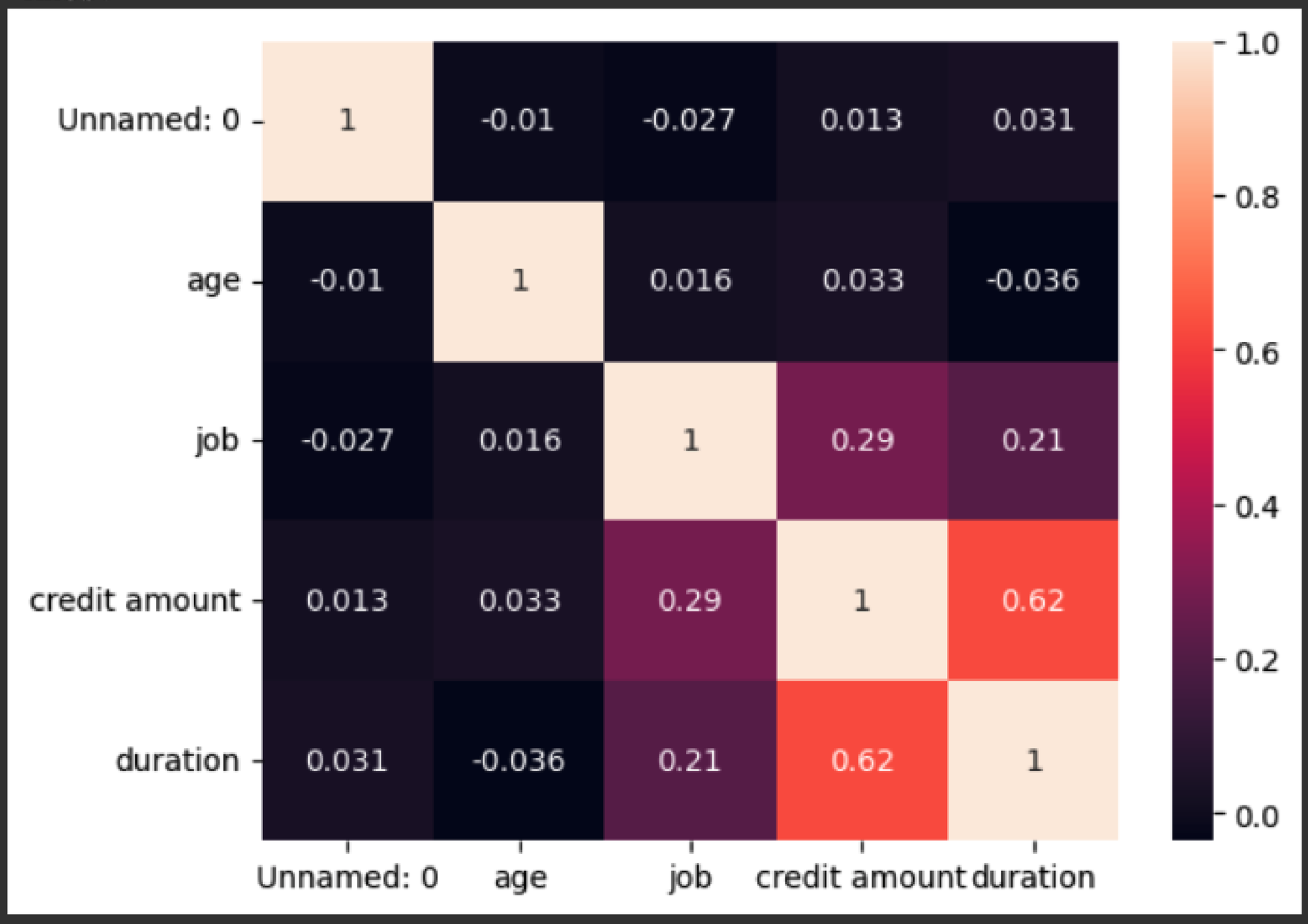
Practical Example

Credit Risk – German Credit

Data Visualization

```
import seaborn as sns
sns.heatmap(corrmat, annot=True)
```

```
<Axes: >
```



Educator in AI

Artificial Intelligence

Data Engineering

Machine Learning

Data Science

📍Linkedin —> https://www.linkedin.com/in/erlinares/

👋 Follow us on 𝕏: https://x.com/erlinares

💻 GitHub: https://github.com/erlinares/365_AI_Journey/

💬 Discord: https://discord.gg/5fFM2zh8

Edgar Rios Linares