МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ "ЛЬВІВСЬКА ПОЛІТЕХНІКА"

Інститут **КНІТ**
Кафедра **ПЗ**

ЗВІТ

До лабораторної роботи № 2
З дисципліни: *"Видобування та опрацювання даних"*
На тему: "Етап розуміння даних за CRISP-DM"

Лектор:
асист. каф. ПЗ
Угриновський Б.В.

Виконав:
ст. гр. ПЗ-45
Хруставчук М.Л.

Прийняв:
асист. каф. ПЗ
Симець І.І.

« ____ » _____ 2026 р.

∑= _____ .....

Львів – 2026

Тема роботи: Етап розуміння даних за CRISP-DM.

Мета роботи: Навчитися аналізувати дані та їх характеристики з допомогою статистичних та візуальних інструментів бібліотек pandas та matplotlib.

ТЕОРЕТИЧНІ ВІДОМОСТІ

Етап розуміння даних (Data Understanding) у CRISP-DM включає первинний аналіз набору даних з метою визначення його структури, якості та характеристик. На цьому етапі досліджують склад датасету, перевіряють коректність значень, наявність пропусків, а також визначають основні властивості змінних.

У Data Mining виділяють основні типи даних: числові, категоріальні, текстові та часові. Числові дані бувають дискретними або неперервними. Категоріальні поділяються на номінальні (без порядку) та порядкові (з порядком). Текстові дані є неструктурованими, а часові ряди відображають значення показників у хронологічному порядку.

Важливою частиною аналізу є дослідження розподілу даних, який показує частоту появи значень у вибірці. Розподіли бувають дискретними та неперервними. Найпоширеніші приклади: нормальний, логнормальний, Бернуллі, біноміальний і Пуассонівський. Аналіз розподілів допомагає виявляти аномальні значення та оцінювати адекватність даних для статистичних методів.

Для аналізу залежностей між числовими ознаками використовується кореляційна матриця, що показує лінійний зв’язок у межах від -1 до 1. Значення близькі до 1 означають сильну пряму залежність, до -1 – сильну обернену, а близькі до 0 – відсутність лінійного зв’язку. Кореляція не є доказом причинності.

ЗАВДАННЯ

1. Завантажити датасет у DataFrame, показати перші 10 рядків.
2. Описати вибрані колонки для аналізу.
3. Визначити типи даних, показати категорії.

4. Знайти min/max числових значень, оцінити адекватність.
5. Перевірити пропуски/null/некоректні значення.
6. Побудувати 3+ розподіли, визначити тип розподілу.
7. Побудувати 3+ графіки, коротко описати.
8. Побудувати кореляційну матрицю, навести 3 приклади зв’язків.

ХІД ВИКОНАННЯ

1. Завантажити датасет у DataFrame, показати перші 10 рядків.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

from google.colab import drive
drive.mount('/content/drive')

df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/players_fifa23.csv')
df.head(10)
```

	short_name	player_positions	overall	potential	value_eur	wage_eur	age	height_cm	weight_kg	preferred_foot	...	mentality_composure	defending_marking_awareness	defending_standing_tackle
0	L. Messi	RW	91	91	54000000.0	195000.0	35	169	67	Left	...	96.0	20	35
1	K. Benzema	CF, ST	91	91	64000000.0	450000.0	34	185	81	Right	...	90.0	43	24
2	R. Lewandowski	ST	91	91	84000000.0	420000.0	33	185	81	Right	...	88.0	35	42
3	K. De Bruyne	CM, CAM	91	91	107500000.0	350000.0	31	181	70	Right	...	89.0	68	65
4	K. Mbappé	ST, LW	91	95	190500000.0	230000.0	23	182	73	Right	...	88.0	26	34
5	Cristiano Ronaldo	ST	90	90	41000000.0	220000.0	37	187	83	Right	...	95.0	24	32
6	M. Neuer	GK	90	90	13500000.0	72000.0	36	193	93	Right	...	70.0	17	10
7	T. Courtois	GK	90	91	90000000.0	250000.0	30	199	96	Left	...	66.0	20	18
8	V. van Dijk	CB	90	90	98000000.0	230000.0	30	193	92	Right	...	90.0	92	92
9	M. Salah	RW	90	90	115500000.0	270000.0	30	175	71	Left	...	92.0	38	43

Рис. 1. Результат відображення перших 10 записів

2. Описати вибрані колонки для аналізу.

Серед специфічних колонок та тих, що потребують пояснення предметної області, можна виділити такі:

- value_eur – ринкова (трансферна) вартість гравця в євро;
- player_positions – позиція(ї) гравця на полі (категоріальна змінна, може містити кілька позицій);
- age – вік гравця;
- height_cm – зріст гравця;

- weight_kg – вага гравця;
- overall – загальний рейтинг гравця у FIFA (інтегральна оцінка рівня);
- pace – швидкість;
- shooting – удар/завершення атак;
- passing – точність пасу.
- dribbling – дриблінг.
- defending – захисні навички.
- physic – фізика/боротьба/витривалість.

```
selected_columns = [
    'value_eur',
    'player_positions',
    'age',
    'height_cm',
    'weight_kg',
    'overall',
    'pace',
    'shooting',
    'passing',
    'dribbling',
    'defending',
    'physic'
]

df_selected = df[selected_columns].copy()
df_selected.head(10)

is_gk = df_selected['player_positions'].str.contains('GK', na=False)
df_selected = df_selected[~is_gk].copy()
```

3. Визначити типи даних, показати категорії.

У моєму наборі даних представлені такі типи:

- Числові (numeric): value_eur, age, height_cm, weight_kg, overall, pace, shooting, passing, dribbling, defending, physic.
- Категоріальні (categorical / text): player_positions.

Позиції гравців записуються скорочено. Типові скорочення у FIFA:

- CB (Centre Back) – центральний захисник;
- LB / RB (Left Back / Right Back) – лівий / правий захисник;
- LWB / RWB (Left Wing Back / Right Wing Back) – лівий / правий вінгбек;
- CDM (Central Defensive Midfielder) – опорний півзахисник;

- CM (Central Midfielder) – центральний півзахисник;
- CAM (Central Attacking Midfielder) – атакувальний півзахисник;
- LM / RM (Left Midfielder / Right Midfielder) – лівий / правий півзахисник;
- LW / RW (Left Winger / Right Winger) – лівий / правий вінгер;
- CF (Centre Forward) – відтягнутий форвард;
- ST (Striker) – нападник.

```
df_selected.dtypes
df_selected['player_positions'].nunique()

positions = (
    df_selected['player_positions']
    .dropna()
    .str.split(', ')
    .explode()
)

positions.unique()
positions.value_counts()
```

4. Знайти min/max числових значень, оцінити адекватність.

```
df_selected.select_dtypes(include=['int64', 'float64']).agg(['min', 'max'])
```

	value_eur	age	height_cm	weight_kg	overall	pace	shooting	passing	dribbling	defending	physic
min	15000.0	16	155	49	46	28.0	16.0	25.0	28.0	15.0	30.0
max	190500000.0	41	206	102	91	97.0	92.0	93.0	94.0	91.0	90.0

Рис. 2. Мінімальні та максимальні значення для числових колонок

Загалом значення виглядають адекватними та реалістичними для футбольного датасету. Мінімальні та максимальні межі не містять очевидних аномалій або фізично неможливих значень. Діапазони відповідають реальним характеристикам професійних гравців, тому дані можна вважати коректними для подальшого аналізу.

5. Перевірити пропуски/null/некоректні значення.

```
df_selected.isnull().sum()
(df_selected.isnull().sum() / len(df_selected)) * 100
```

	0
value_eur	0.431165
player_positions	0.000000
age	0.000000
height_cm	0.000000
weight_kg	0.000000
overall	0.000000
pace	0.000000
shooting	0.000000
passing	0.000000
dribbling	0.000000
defending	0.000000
physic	0.000000

Рис. 3. Інформація про кількість пустих значень для конкретних колонок

6. Побудувати 3+ розподіли, визначити тип розподілу.

```
# 1 value_eur - Lognormal distribution
plt.figure(figsize=(5,3))
sns.histplot(
    np.log10(df_selected[df_selected['value_eur'] > 0]['value_eur']),
    bins=50
)
plt.title('Normal distribution after log transform (value_eur)')
plt.xlabel('log10(value_eur)')
plt.show()

# 2 overall - Normal distribution
plt.figure(figsize=(5,3))
sns.histplot(df_selected['overall'], bins=30)
plt.title('Normal distribution (overall)')
plt.show()

# 3 age - Not normal distribution
plt.figure(figsize=(5,3))
sns.histplot(df_selected['age'], bins=range(15, 46))
plt.title('Not normal distribution (age)')
plt.show()
```

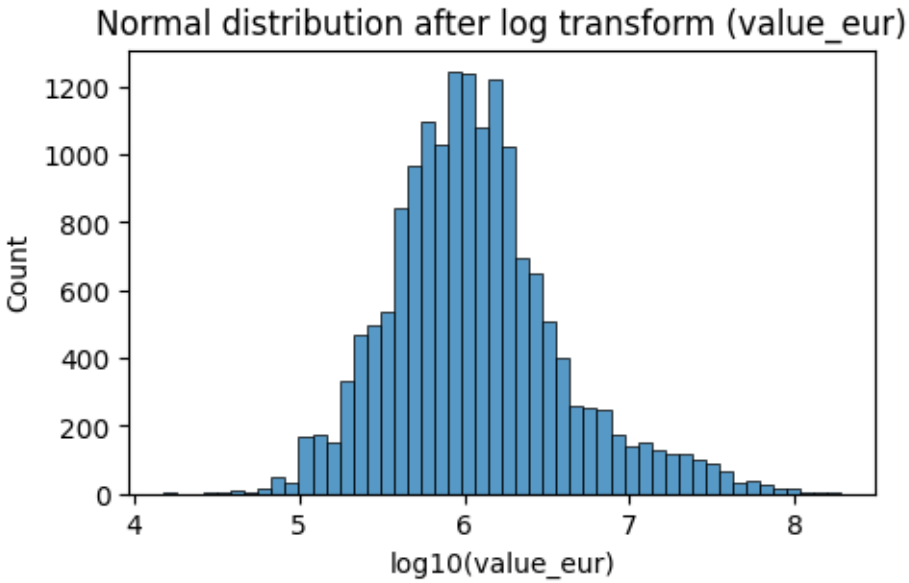

Рис. 4. Розподіл ринкової вартості гравців (log10 value_eur)

Показує, як розподіляється ринкова вартість гравців у вибірці. Дає розуміння, чи більшість гравців дешеві або дорогі, і наскільки є нерівномірність між масовими гравцями та зірками.

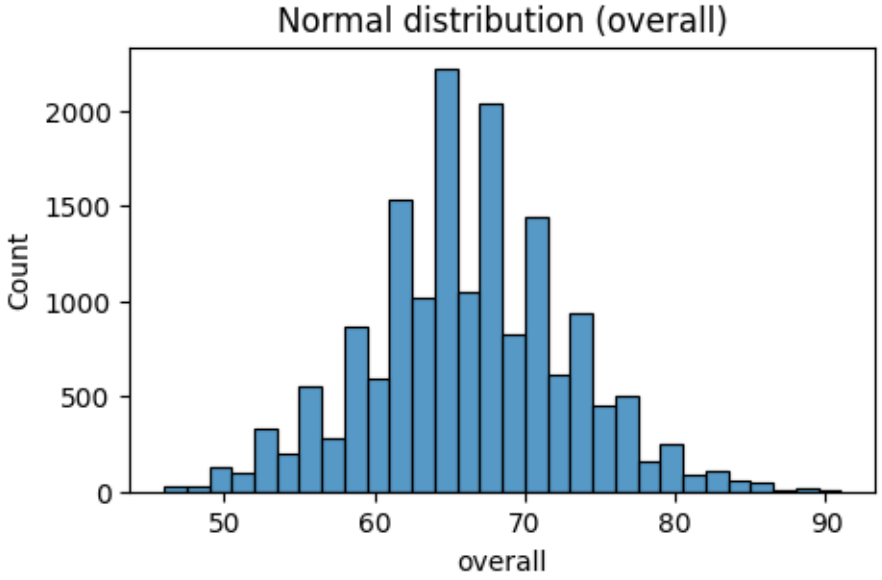

Рис. 5. Розподіл загального рейтингу гравців (overall)

Показує загальний рівень гравців у датасеті. Дозволяє оцінити, який рейтинг є типовим, та наскільки сильно гравці відхиляються від середнього рівня.

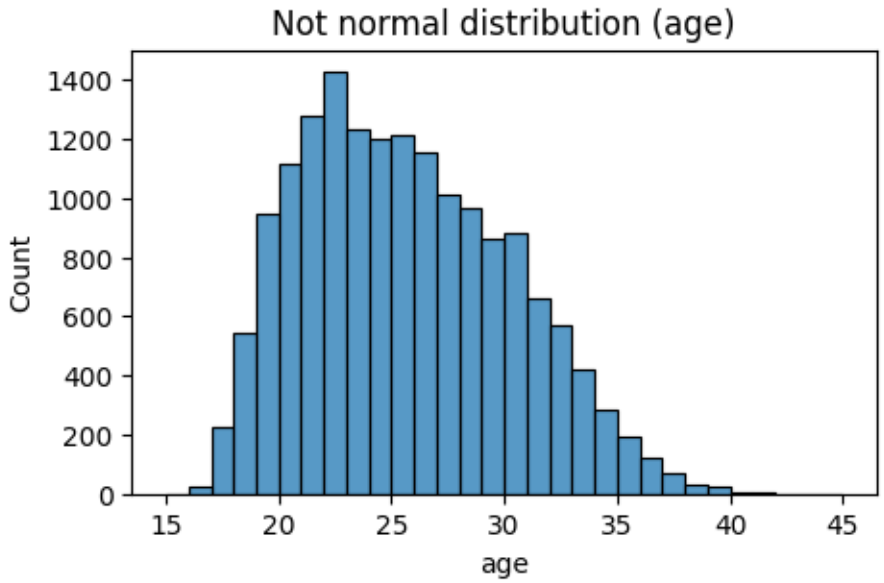

Рис. 6. Розподіл віку гравців (age)

Показує вікову структуру гравців. Дає уявлення про те, у якому віковому діапазоні зосереджена основна маса футболістів і наскільки представлена старша або молодша група.

7. Побудувати 3+ графіки, коротко описати.

```
# Графік 1: Top 15 Most Common Player Positions
positions = (
    df_selected['player_positions']
    .dropna()
    .str.split(', ')
    .explode()
)

top_positions = positions.value_counts().head(15)

plt.figure(figsize=(10,6))
sns.barplot(x=top_positions.values, y=top_positions.index)
plt.title('Top 15 Most Common Player Positions')
plt.xlabel('Number of Players')
plt.ylabel('Position')
plt.show()

# Графік 2: Market Value by Primary Position (Log Scale)
df_selected['primary_position']                                                    =
df_selected['player_positions'].str.split(', ').str[0]

top_primary = df_selected['primary_position'].value_counts().head(10).index
df_pos_top = df_selected[df_selected['primary_position'].isin(top_primary)]

plt.figure(figsize=(12,6))
sns.boxplot(data=df_pos_top, x='primary_position', y='value_eur')
plt.yscale('log')
plt.title('Market Value by Primary Position (Log Scale)')
plt.xlabel('Primary Position')
plt.ylabel('Market Value (log scale)')
```

```
plt.show()

# Графік 3: Average Overall Rating by Position
avg_overall = (
    df_pos_top
    .groupby('primary_position')['overall']
    .mean()
    .sort_values(ascending=False)
)

plt.figure(figsize=(10,6))
sns.barplot(x=avg_overall.index, y=avg_overall.values)
plt.title('Average Overall Rating by Position (Top 10)')
plt.xlabel('Position')
plt.ylabel('Average Overall Rating')
plt.show()
```

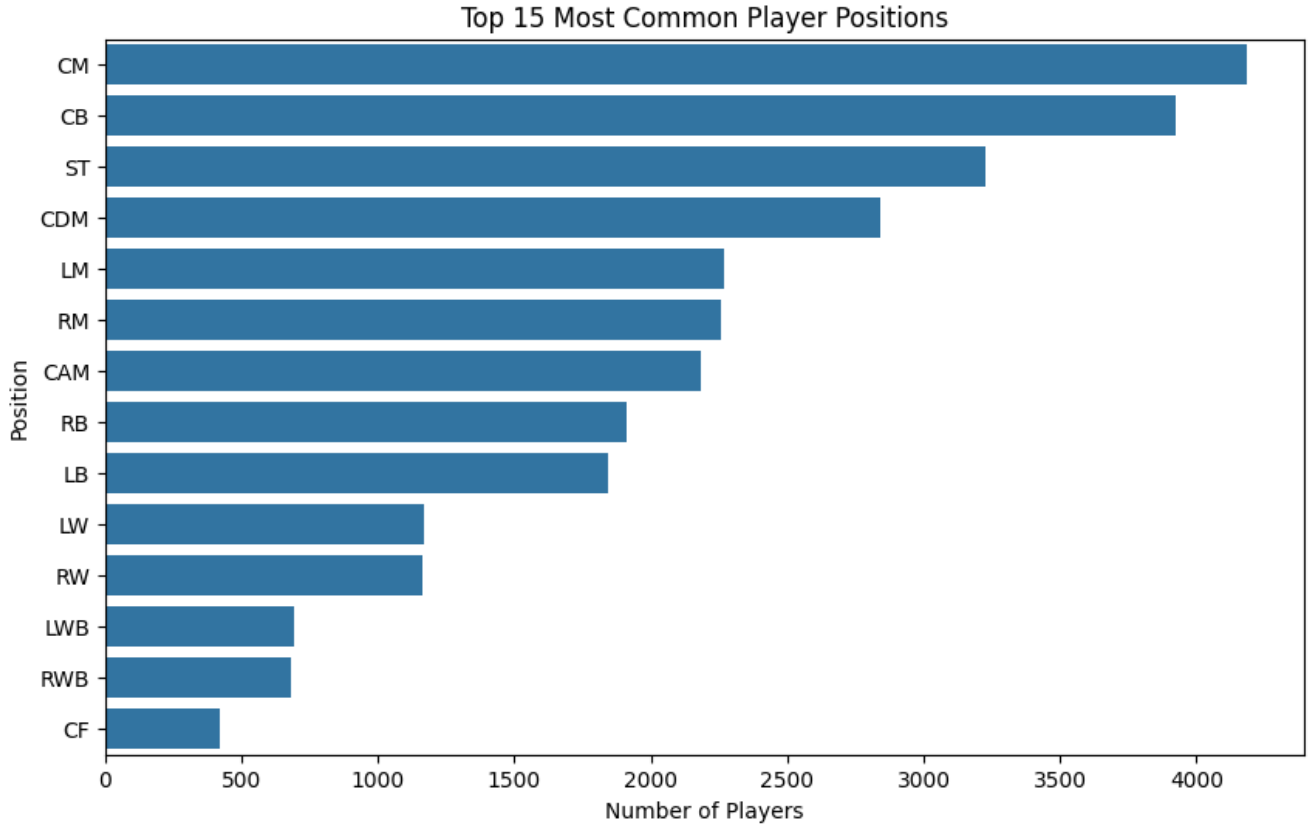

Рис. 7. Top 15 Most Common Player Positions

Графік демонструє кількість гравців на кожній позиції. Найбільш представленими є CM, CB та ST. Це свідчить про те, що вибірка містить переважно польових гравців центральної зони та нападників. Структура позицій відповідає типовому складу футбольних команд.

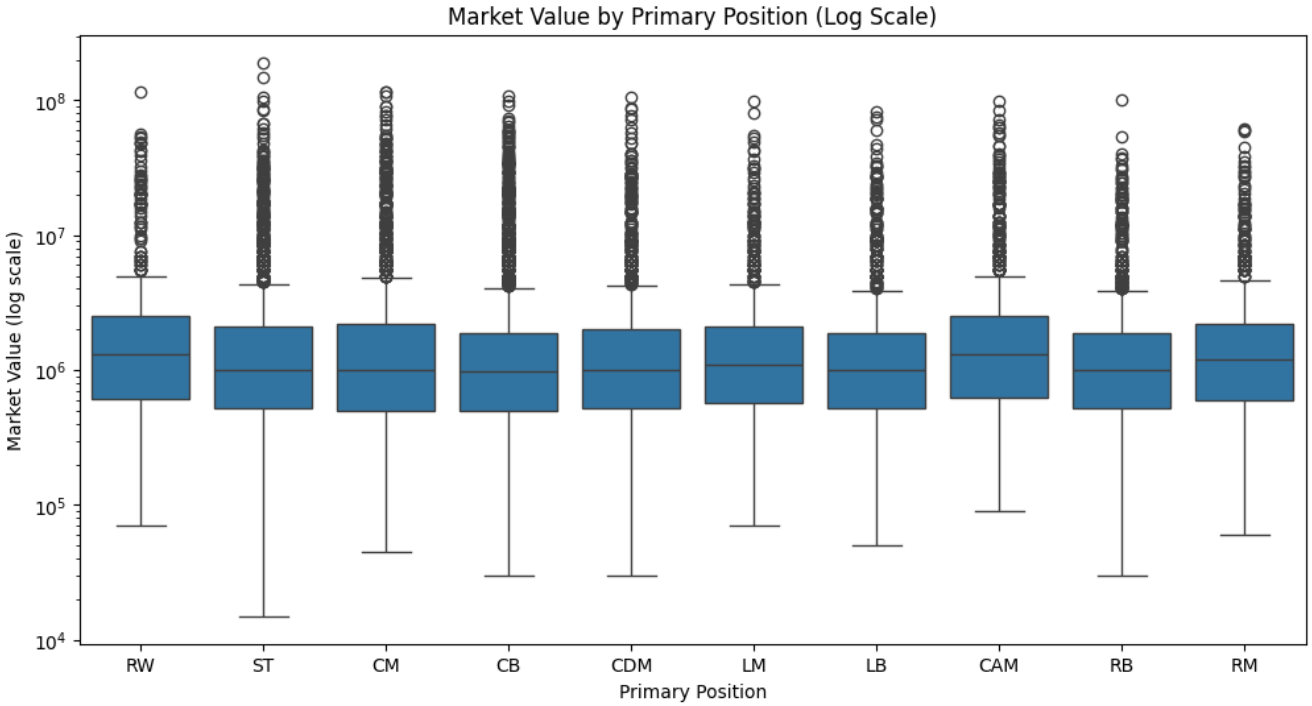

Рис. 8. Market Value by Primary Position (Log Scale)

Boxplot показує розподіл ринкової вартості за основною позицією гравця. Використання логарифмічної шкали дозволяє коректно відобразити великий діапазон вартості. Видно, що медіанна вартість різних позицій є подібною, але для атакувальних ролей (RW, ST, CAM) спостерігається більша кількість високовартісних гравців.

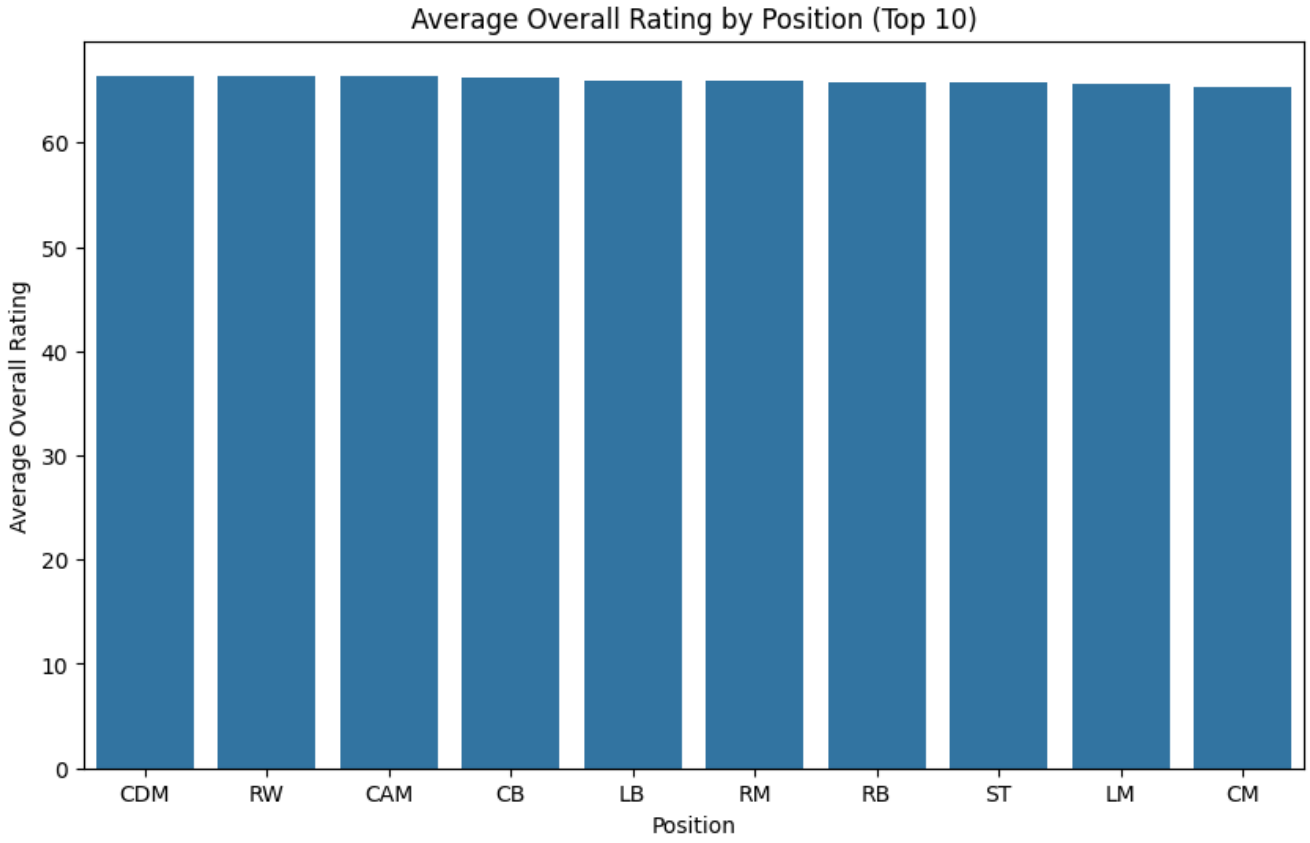

Рис. 9. Average Overall Rating by Position

Графік показує середній рейтинг гравців для кожної позиції. Середні значення дуже близькі між собою, що свідчить про відносно рівномірний рівень гравців у вибірці. Значних відмінностей між позиціями за середнім overall не спостерігається.

8. Побудувати кореляційну матрицю, навести 3 приклади зв’язків.

```
numeric_df = df_selected.select_dtypes(include=['int64', 'float64'])
corr = numeric_df.corr()

plt.figure(figsize=(10,8))
sns.heatmap(corr, annot=True)
plt.title('Correlation matrix (selected columns)')
plt.show()
```

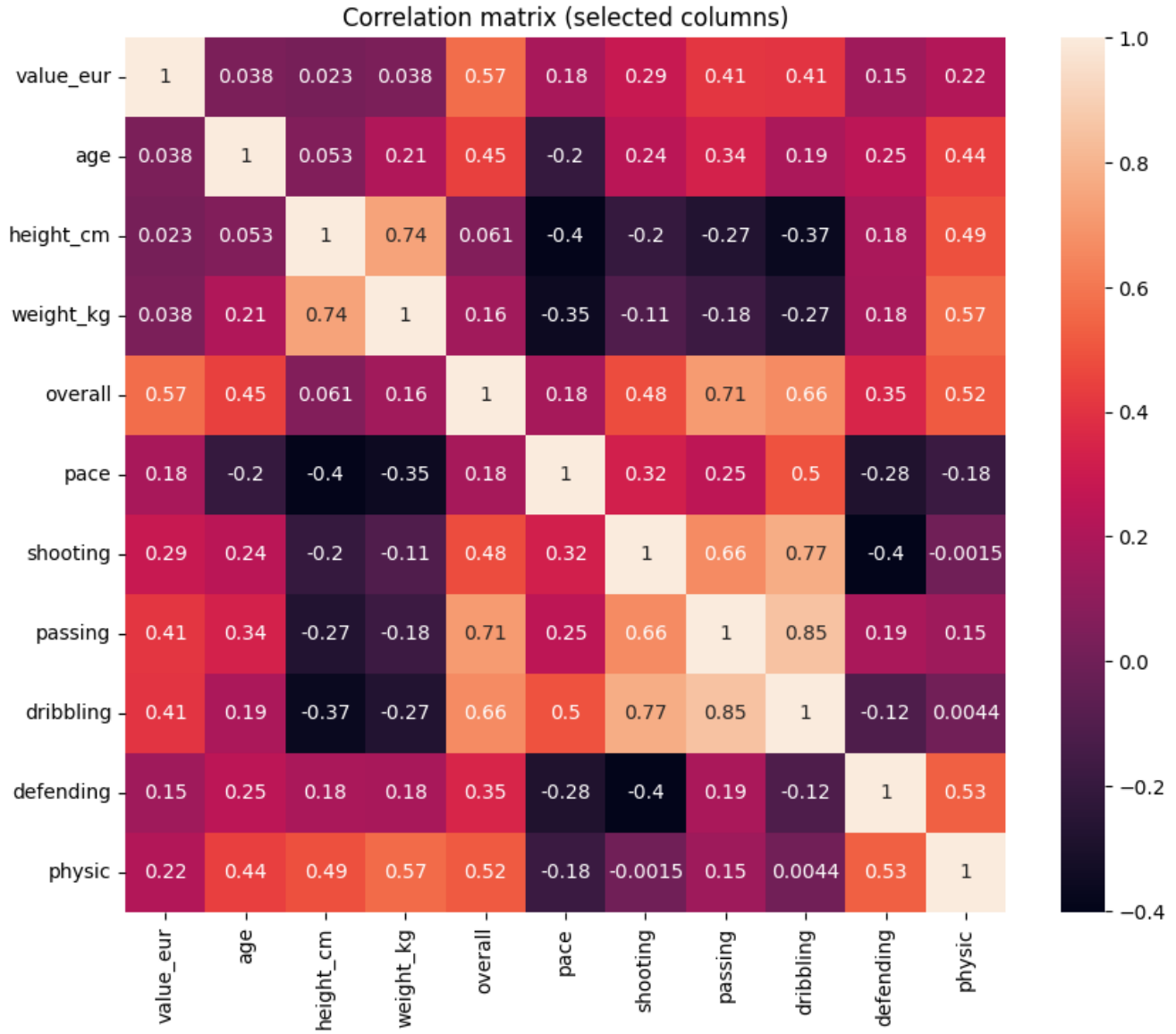

Рис. 10. Кореляційна матриця

З кореляційної матриці видно кілька помітних залежностей. Між value_eur і overall спостерігається достатньо сильна позитивна кореляція близько 0.57, що означає: чим вищий загальний рейтинг гравця, тим вища його ринкова вартість. Між passing і dribbling кореляція дуже висока, близько 0.85, що логічно, оскільки технічні навички зазвичай розвиваються комплексно. Також height_cm і weight_kg мають сильну позитивну залежність близько 0.74, адже вищі гравці зазвичай важчі. Крім цього, технічні показники, такі як shooting, dribbling і passing, мають помірну позитивну кореляцію з overall, що очікувано, оскільки overall є узагальненою оцінкою рівня гравця.

ВИСНОВКИ

У ході виконання лабораторної роботи було проведено етап розуміння даних відповідно до методології CRISP-DM. Було проаналізовано структуру датасету, типи змінних, розподіли показників та виявлено наявність пропусків. Побудовані візуалізації дозволили оцінити структуру позицій гравців та залежність ринкової вартості від характеристик. Кореляційний аналіз підтвердив логічні зв’язки між технічними показниками та загальним рейтингом гравців, що створює основу для подальшого моделювання.