

پروژه سه‌فازی تحلیل داده‌های بسکتبال NBA

از استخراج داده تا طراحی پایگاه داده و تحلیل آماری

اطلاعات گروه Group 3

منتور :

شادلین سلطان پور

اعضا گروه :

- ثنا خالقی
- آرش نادری
- محمد امین دارا
- حسام صارمی زاده
- نرگس نورآذر


محتوا :

فاز اول: استخراج داده‌های خام از وبسایت Basketball Reference

فاز دوم: طراحی و نرمال‌سازی پایگاه داده رابطه‌ای

فاز سوم: تحلیل‌های آماری توصیفی و تست فرضیه‌های کارشناسی

فاز اول: استخراج داده‌های بسکتبال

ابزارهای استفاده‌شده 

کتابخانه	کاربرد
requests, BeautifulSoup	دریافت و پارس صفحات HTML
Selenium	استخراج داده از صفحات داینامیک
pandas	ساخت و ذخیره DataFrame
re, datetime	پردازش دقیق رشته‌ها و تاریخ‌ها

هدف 

جمع‌آوری اطلاعات بازیکنان، تیم‌ها، قهرمانان فصل و لیست جوایز برای ساخت دیتاست قابل تحلیل

داده‌های جمع‌آوری‌شده در فاز اول

- لیست ۶۰ بازیکن برتر هر فصل (2019-2025)
- بازیکنان تیم‌های قهرمان فصل
- اطلاعات بیوگرافی و آماری بازیکنان (سن، قد، وزن، ملیت، تیم، پوزیشن، تجربه، PTS)
- لیست بازیکنان جایزه Michael Jordan Trophy
- اطلاعات پایه تیم‌های فعال لیگ (نام، موقعیت، تأسیس، قهرمانی، پلی‌آف)

فاز دوم: طراحی پایگاه داده بسکتبال

جداول نهایی پایگاه داده

هدف 

Players, Positions, Player_Positions

Teams, Player_Teams

Seasons, Champions, Champion_Players

Player_Ranks, Michael_Jordan_Trophy

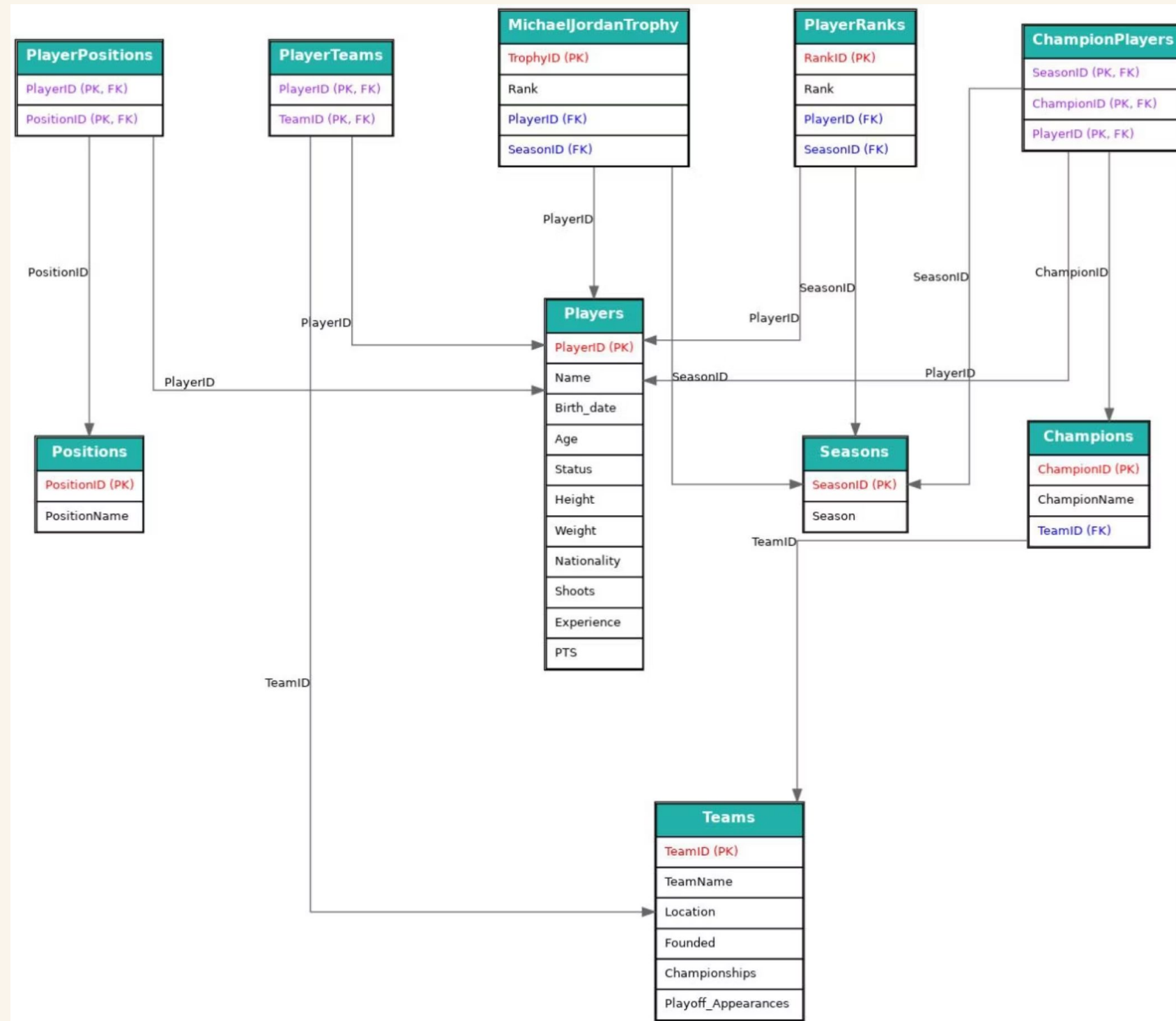
ساخت پایگاه داده رابطه‌ای نرمال برای مدیریت اطلاعات بازیکنان، تیم‌ها، فصل‌ها و جوایز

مراحل کلیدی: 

- اتصال به MySQL با mysql.connector و SQLAlchemy
- پردازش داده‌ها با pandas
- نرمال‌سازی اسامی با unicodedata, rapidfuzz
- ساخت جداول میانی و نهایی
- بارگذاری داده‌ها با to_sql در MySQL

مزایا: 

- حذف تکرار
- یکپارچگی داده‌ها
- امکان گزارش‌گیری و تحلیل پیشرفته



استخراج داده با کوئری SQL

پس از طراحی پایگاه داده، نوبت به استخراج اطلاعات مورد نیاز برای تحلیل‌های آماری می‌رسد. در این مرحله، با استفاده از کوئری‌های SQL داده‌ها را به شکل مطلوب آماده می‌کنیم.



اتصال به پایگاه داده

برقراری ارتباط پایدار با MySQL با بهره‌گیری از ابزار قدرتمند SQLAlchemy.



اجرای کوئری‌های پیچیده

استفاده از تابع `pd.read_sql` برای اجرای کوئری‌های پیشرفته شامل `JOIN`, `CASE`, `DATEDIFF` و `ROUND`.



خروجی مستقیم به DataFrame

تبدیل یکپارچه و مستقیم نتایج کوئری به DataFrame پاندا برای شروع فوری تحلیل‌های آماری در محیط پایتون.



محاسبه متغیر

های تحلیلی

تعریف و محاسبه متغیرهای کلیدی جدید مانند «چابکی» و «توانایی ذاتی» بازیکنان،

مستقیماً در کوئری SQL.

فاز سوم: تحلیل‌های آماری و تست فرضیه‌ها

ساختار فاز سوم 

بخش اول: تحلیل‌های آماری توصیفی

بخش دوم: تست فرضیه‌ها با آزمون‌های آماری

هدف 

تحلیل آماری داده‌های ساخت‌یافته لیگ NBA برای بررسی ویژگی‌های

بازیکنان، تیم‌های قهرمان و اعتبارسنجی فرضیه‌های کارشناسی

مقایسه قد بازیکنان MVP با ۵۰ بازیکن برتر فصل

1.82

اختلاف میانگین

سانتی متر

توزیع MVP ها

بازه 201-211

198.48

میانگین قد بازیکنان برتر

سانتی متر

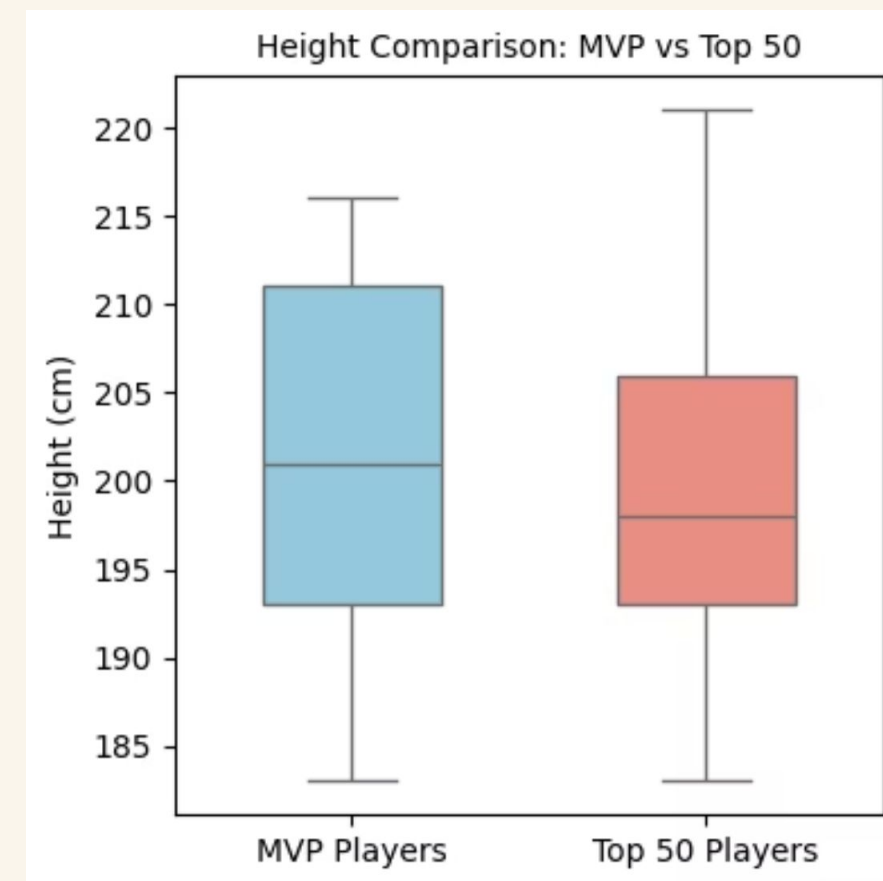
توزیع برترها

بازه 198-206

200.30

میانگین قد MVP ها

سانتی متر



تحلیل قد و تجربه بازیکنان در دو فصل آخر

♦ قد

قهرمان‌ها: میانگین 200.43، میانه 201

برترها: میانگین 198.63، میانه 198

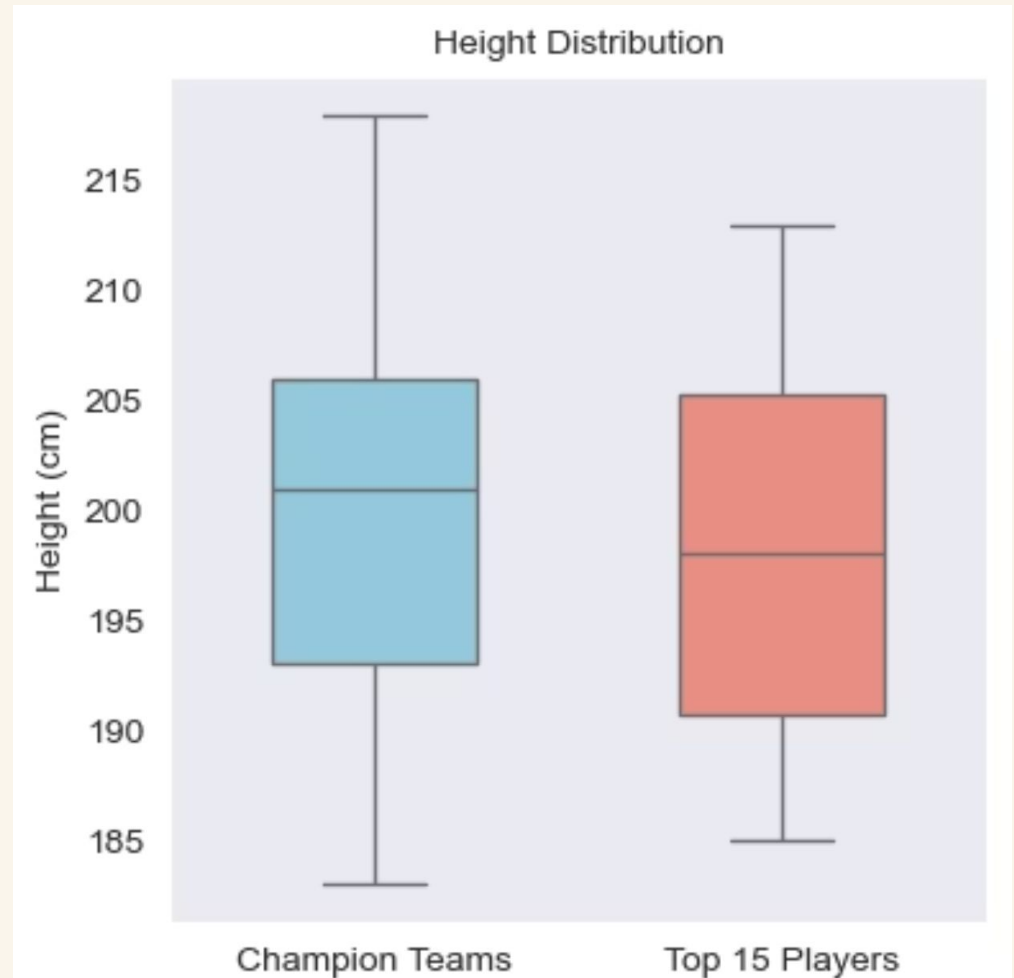
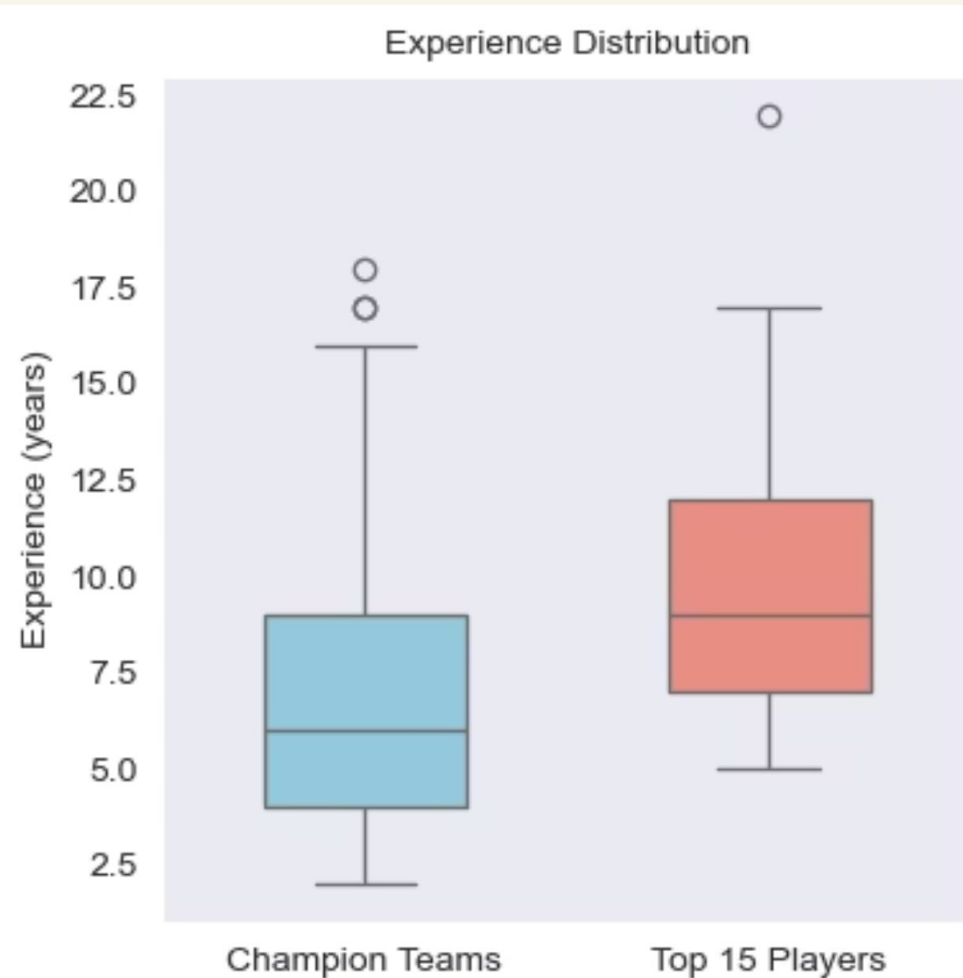
📌 نتیجه: قهرمان‌ها قد بلندتر و یکنواخت‌تر دارند

♦ تجربه

قهرمان‌ها: میانگین 7.48، میانه 6

برترها: میانگین 10.10، میانه 9

📌 نتیجه: قهرمان‌ها جوان‌تر و برترها باتجربه‌ترند



انتخاب بازیکن مناسب برای باشگاه

📌 معیارها:

- تعداد Trophy
- سن
- میانگین امتیاز (PTS)
- تجربه

📌 گزینه‌ها:

تجربه	PTS	سن	Trophy	بازیکن
۷ فصل	۲۸.۶	۲۶	۵	Luka Dončić
۱۲ فصل	۲۳.۹	۳۰	۵	Giannis
۱۶ فصل	۲۴.۷	۳۷	۳	Curry

بلندمدت

Luka

فیزیک و تجربه

Giannis

رهبری و ثبات

Curry

آیا چابکی بازیکنان برتر افزایش یافته؟

1 تعریف چابکی

نسبت قد به وزن

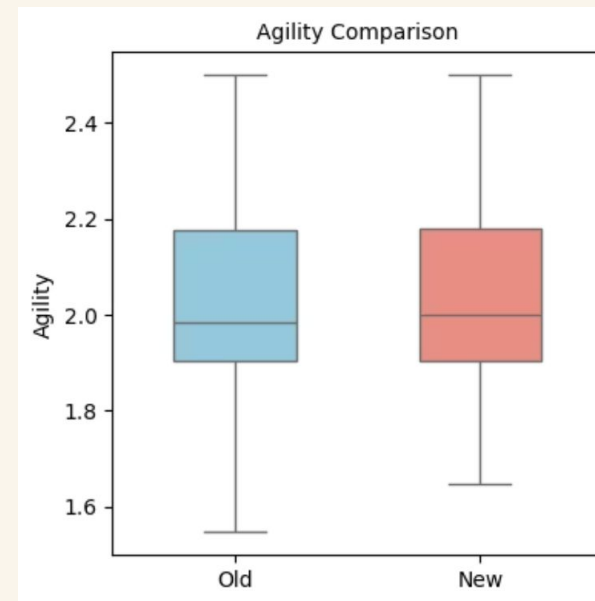
2 مقایسه دو دوره

قدیم (22-2020) vs جدید (24-2023)

3 آزمون‌ها

Shapiro-Wilk → هر دو نرمال

t-test → p-value = 0.487



نتایج:

- میانگین جدید: 2.0394
- میانگین قدیم: 2.0378
- اختلاف: 0.0016

نتیجه: فرضیه رد شد

آیا توانایی ذاتی بازیکنان قهرمان افزایش یافته؟

1 تعریف Skill

تجربه / سن

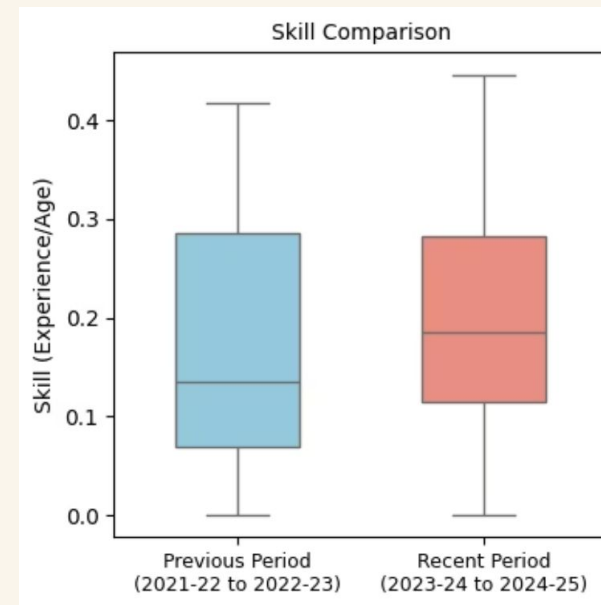
2 مقایسه دو دوره

قدیم (22-2020) vs جدید (24-2023)

3 آزمون‌ها

Shapiro-Wilk → جدید نرمال، قدیم غیرنرمال

Mann-Whitney U → p-value = 0.2070



نتایج:

- میانگین جدید: 0.203
- میانگین قدیم: 0.181
- اختلاف: 0.022

نتیجه: فرضیه رد شد

جمع‌بندی سه فاز پروژه

فاز اول

استخراج داده‌های دقیق و متنوع از منابع آنلاین

فاز دوم

طراحی پایگاه داده نرمال با ساختار رابطه‌ای و قابل تحلیل

فاز سوم

تحلیل‌های آماری نشان دادند:

- قد بلندتر در موفقیت تیمی و MVP مؤثر است

- تجربه بالا در عملکرد انفرادی مهم است

- تیم‌های قهرمان ترکیب جوان‌تر و فیزیکی‌تری

دارند

تست فرضیه‌ها نشان دادند:

- چابکی و توانایی ذاتی در دوره جدید تفاوت

معنادار ندارند

- تحلیل آماری مانع تصمیم‌گیری‌های شهودی و

نادقیق می‌شود