مینی پروژه سوم: جمع آوری و تحلیل داده های بسکتبال

منتور: شادلین

گروه: G4

مقدمه:

در این پروژه با استفاده از وب اسکرپینگ داده‌های مربوط به تیم‌ها، بازیکنان، جوایز و فصل‌های مختلف بسکتبال دنیا از وب‌سایت هدف استخراج شده و در یک پایگاه داده به‌صورت ساخت‌یافته ذخیره می‌شود. پس از گردآوری و سامان‌دهی داده‌ها، با بهره‌گیری از کوئری‌ها، آزمون‌های آماری و ترسیم نمودارهای تحلیلی، داده‌ها مورد بررسی و تحلیل قرار می‌گیرند تا الگوها و بینش‌های معناداری از آن‌ها به دست آید.

مرحله اول: جمع آوری دیتا از سایت

سایت *basketball-reference.com* یکی از منابع جامع در حوزه بسکتبال است که اطلاعات تاریخی و آماری مربوط به لیگ‌های مختلف از جمله NBA، WNBA، ABA و مسابقات اروپایی را در اختیار قرار می‌دهد. این سایت شامل داده‌هایی درباره بازیکنان، تیم‌ها، اخبار، امتیازات، فصل‌ها (از جمله لیگ ABA از فصل 1968-69)، جدول زمان‌بندی بازی‌ها، رکوردهای ملی در جام جهانی بسکتبال و بسیاری شاخص‌های دیگر است.

برای دستیابی به داده‌های مورد نیاز در این پروژه، چهار Crawler مختلف طراحی و پیاده‌سازی شد. در این مرحله:

با استفاده از کتابخانه Selenium، صفحات وب هدف باز و لینک‌های مرتبط برای داده‌های مورد نظر پیمایش شدند. در این فرآیند، مرورگر Firefox با پیکربندی اختصاصی، فهرستی از *User-Agent*ها و تنظیمات غیرفعال‌سازی قابلیت شناسایی اتوماسیون مرورگر مورد استفاده قرار گرفت.

داده‌های HTML استخراج‌شده با کمک کتابخانه BeautifulSoup پردازش و به‌صورت ساخت‌یافته در قالب لیست مدیریت شدند (همراه با مدیریت خطا). در نهایت، داده‌ها با استفاده از کتابخانه Pandas به DataFrame تبدیل و برای مراحل بعدی ذخیره شدند. در این پروژه، داده‌های استخراج‌شده توسط چهار Crawler به شرح زیر هستند:

1. بازیکنان برتر و جوایز
   - جستجوی 50 بازیکن برتر مرتبط با جایزه ***Trophy Jordan Michael*** برای سال‌های 2019 تا 2025
   - ذخیره اطلاعات شامل: نام جایزه، نام بازیکن، سن، تیم بازیکن در فصل مربوطه، لینک صفحه بازیکن و سایر جزئیات موجود
   - خروجی در قالب فایل CSV

2. Crawler تیم‌های NBA
   - استخراج داده‌ها از صفحه ***teams*** شامل: شناسه تیم، نام کامل، موقعیت جغرافیایی، تاریخ حضور در لیگ‌ها، تعداد حضور در ***Playoffs***، تعداد قهرمانی‌ها و رکوردهای آماری
   - ذخیره داده‌ها در فایل CSV

3. Crawler اطلاعات بازیکنان
   - استخراج داده‌ها از صفحه ***totals_stats*** شامل: نام کامل بازیکن، تیم، تاریخ و محل تولد، سال‌های تجربه، پوزیشن، قد و وزن
   - ذخیره داده‌ها در فایل CSV

4. Crawler رتبه‌بندی بازیکنان جام جهانی
   - جمع‌آوری اطلاعات سالانه بازیکنان تیم‌های شرکت‌کننده در جام جهانی
   - داده‌های شامل: تاریخ تولد، پوزیشن، کالج بازیکن و لینک صفحه اطلاعات بازیکن
   - ذخیره خروجی در فایل CSV

مرحله دوم: پاک‌سازی و ذخیره‌سازی داده‌ها در پایگاه داده

قبل از وارد کردن داده‌ها به پایگاه داده، لازم بود به دلیل ذخیره شدن تمامی اطلاعات به صورت رشته‌ای (str)، فرمت هر ستون بررسی و پاک‌سازی شود. در این مرحله با استفاده از کتابخانه mysql.connector و با خواندن فایل

JSON تنظیمات پایگاه داده، اتصال به دیتابیس برقرار شد. همچنین از Pandas و NumPy برای پردازش و آماده‌سازی داده‌ها بهره گرفته شد.

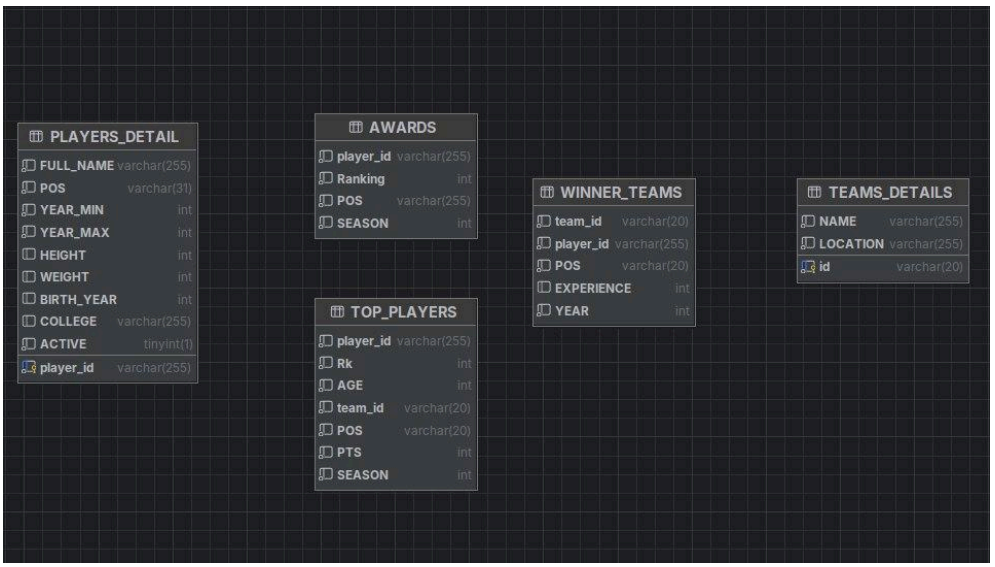
در پردازش داده‌های بازیکنان برتر، ابتدا فایل‌های هر سال با یکدیگر ترکیب شدند. سپس از لینک هر بازیکن شناسه (ID) استخراج شد و ستون لینک حذف گردید. رتبه‌بندی بازیکنان MVP نیز از ستون جوایز (Awards) استخراج شد و تمامی داده‌های مربوط به جدول MVP در یک DataFrame نهایی جمع‌آوری شد تا جدول TOP_PLAYERS شامل ستون‌های شناسه بازیکن، رده، سن، تیم، پوزیشن، امتیاز و فصل ایجاد شود.

اطلاعات تکمیلی بازیکنان نیز پردازش شد؛ به‌عنوان مثال، علامت * از نام بازیکنان حذف شد، قد به سانتی‌متر و وزن به کیلوگرم تبدیل گردید و سال تولد استخراج شد تا در جدول PLAYER_LIST ذخیره شود. اطلاعات تیم‌ها و تیم‌های قهرمان نیز خوانده و در لیست‌های مربوطه نگهداری شد.

سپس با ایجاد یک Cursor برای اجرای دستورات SQL، به پایگاه داده متصل شدیم. در صورتی که پایگاه داده قبلی وجود داشت، حذف شد و پایگاه داده جدید با نام NBA_DB ساخته شد. در این پایگاه داده، جداول مربوطه ایجاد و داده‌ها ذخیره شدند. جدول AWARDS شامل رتبه بازیکنان در جوایز و ستون‌های شناسه بازیکن، رنک، پوزیشن و فصل بود. جدول TOP_PLAYERS اطلاعات بازیکنان برتر شامل شناسه، رده، سن، تیم، پوزیشن، امتیاز و فصل را در خود داشت. جدول PLAYERS_DETAIL حاوی جزئیات کامل بازیکنان مانند قد، وزن، سال تولد، کالج، سال شروع و پایان فعالیت و وضعیت فعال بودن بود. جدول WINNER_TEAMS اطلاعات بازیکنان تیم قهرمان هر فصل، پوزیشن، تجربه و سال را شامل می‌شد و جدول TEAMS_DETAILS شناسه، نام و مکان تیم‌ها را ذخیره می‌کرد.

در نهایت، تمامی داده‌ها با استفاده از حلقه‌های for روی DataFrameها خوانده شده و به جداول پایگاه داده وارد شدند.

در نهایت جداول به این صورت شماتیک می شوند:



مرحله سوم: نوشتن کوئری‌ها برای استخراج داده‌های مورد نیاز تحلیل آماری

در این مرحله، برای پاسخ‌گویی به فرضیه‌ها و پرسش‌های پژوهش، داده‌های مورد نیاز با استفاده از کوئری‌های SQL استخراج شدند و نتایج در فایل‌های CSV ذخیره گردید. ابتدا قد بازیکنان در جدول AWARDS برای فصل‌های 2019 تا 2024 با یک کوئری JOIN بر اساس شناسه بازیکن‌ها از جدول PLAYERS DETAIL استخراج شد. مشابه آن، قد بازیکنان در جدول TOP PLAYERS نیز با استفاده از JOIN و گروه‌بندی بر اساس نام، قد و فصل به دست آمد.

برای بررسی تجربه و قد بازیکنان تیم‌های قهرمان و بازیکنان برتر در فصل‌های 2023 و 2024، داده‌ها از جداول WINNER_TEAMS و TOP_PLAYERS همراه با اطلاعات PLAYERS_DETAIL استخراج شد. در این مرحله تجربه بازیکنان با محاسبه تفاوت فصل جاری و سال شروع فعالیت به دست آمد.

شمارش بازیکنان نامزد در جایزه ها MVP نیز با استفاده از یک *CTE* (WITH clause) انجام شد. در این کوئری ابتدا بازیکنان با موقعیت *PG* و حضور در فصل‌های موردنظر فیلتر شده و تعداد نامزدی‌های هر بازیکن محاسبه و بر اساس تعداد و نام مرتب شد تا سه بازیکن برتر مشخص گردند.

نسبت قد به وزن برای 20 بازیکن برتر دو فصل اخیر (2023 و 2024) نیز با ایجاد یک CTE استخراج شد و شاخص *AGILITY* برای هر بازیکن محاسبه شد. همچنین توان بازیکنان با تقسیم تجربه بر سن محاسبه شد؛ برای این منظور ابتدا اطلاعات بازیکنان تیم‌های قهرمان با سال تولد و تجربه جمع‌آوری شد و سپس شاخص *POTENTIAL* محاسبه گردید تا بازیکنان با بیشترین پتانسیل در هر فصل شناسایی شوند.

تمام نتایج استخراج شده از این کوئری‌ها در فایل‌های CSV ذخیره شدند تا برای تحلیل‌های آماری و ترسیم نمودارها آماده باشند.

مرحله سوم: تحلیل های آماری

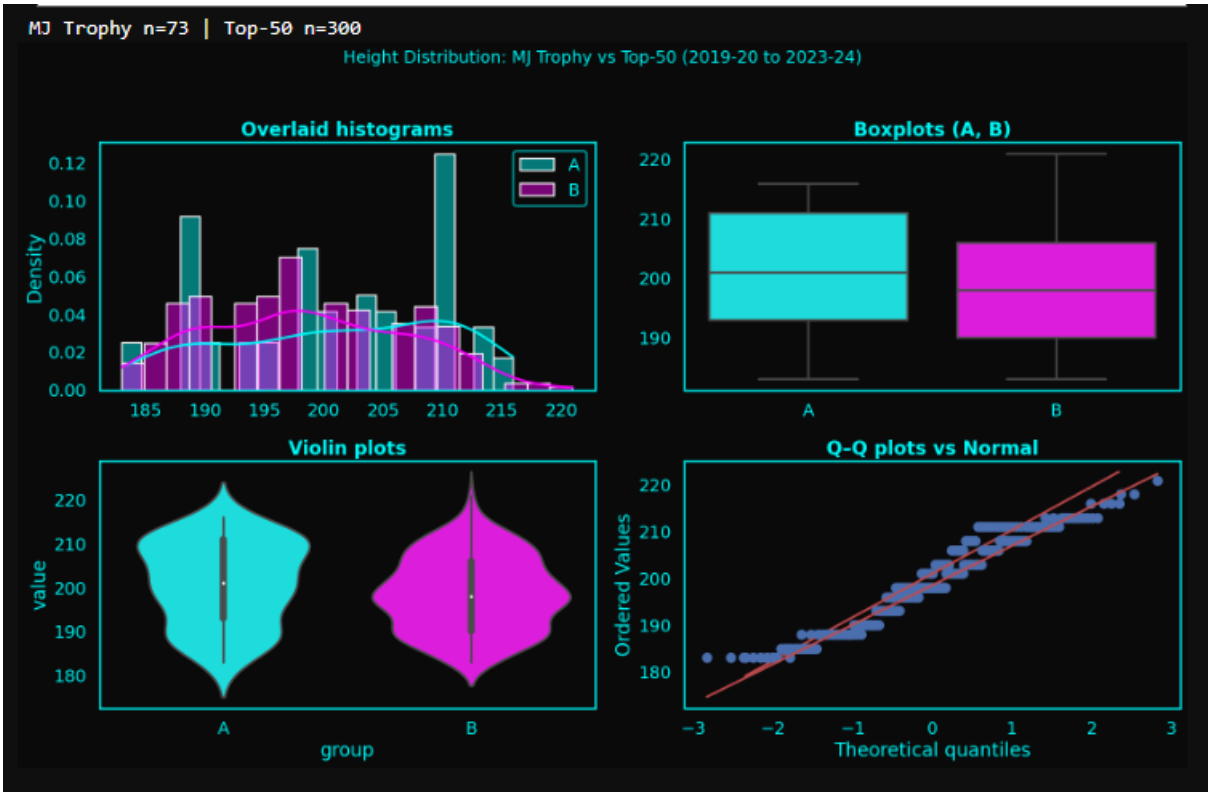
در این بخش با کمک کلاس Analyzer تعریف شده سوالات را تحلیل می کنیم.

سوال اول توزیع قد بازیکنانی که در لیست Trophy Jordan Michael حضور دارند را با ۵۰ بازیکن برتر فصل با توجه به آمار فصلها از فصل ۲۰۲۰-۲۰۱۹ تا پایان ۲۰۲۴-۲۰۲۳ بود. در نتیجه توزیع قد بازیکنان جایزه نسبت به بازیکنان *Top 50* کمی به سمت راست شیفت دارد، یعنی به‌طور متوسط قد بازیکنان MJT کمی بلندتر است.
پراکندگی پهنای جعبه‌ها و نمودارهای ویولین تقریباً مشابه است و واریانس کلی هر دو گروه نزدیک به هم است؛ اما گروه B کمی دم‌های سنگین‌تری دارد، به‌ویژه در سمت کوتاه‌تر.

شکل توزیع/نرمال بودن هر دو گروه تقریباً تک‌قله‌ای (unimodal) و نزدیک به توزیع نرمال هستند اما کاملاً نرمال نیستند:
در نمودار Q–Q، کوانتیل‌های میانی نزدیک خط قرار دارند اما دم‌ها از خط منحرف می‌شوند ⇒ نشان‌دهنده دم‌های سنگین‌تر از نرمال.

پس قد بازیکنان MJ Trophy کمی بیشتر از بازیکنان Top-50 است، اما این تفاوت بزرگ نیست؛ گروه Top-50 دامنه گسترده‌تری را پوشش می‌دهد. برای بررسی اینکه آیا این تفاوت کوچک از نظر آماری معنادار است یا خیر، می‌توان از

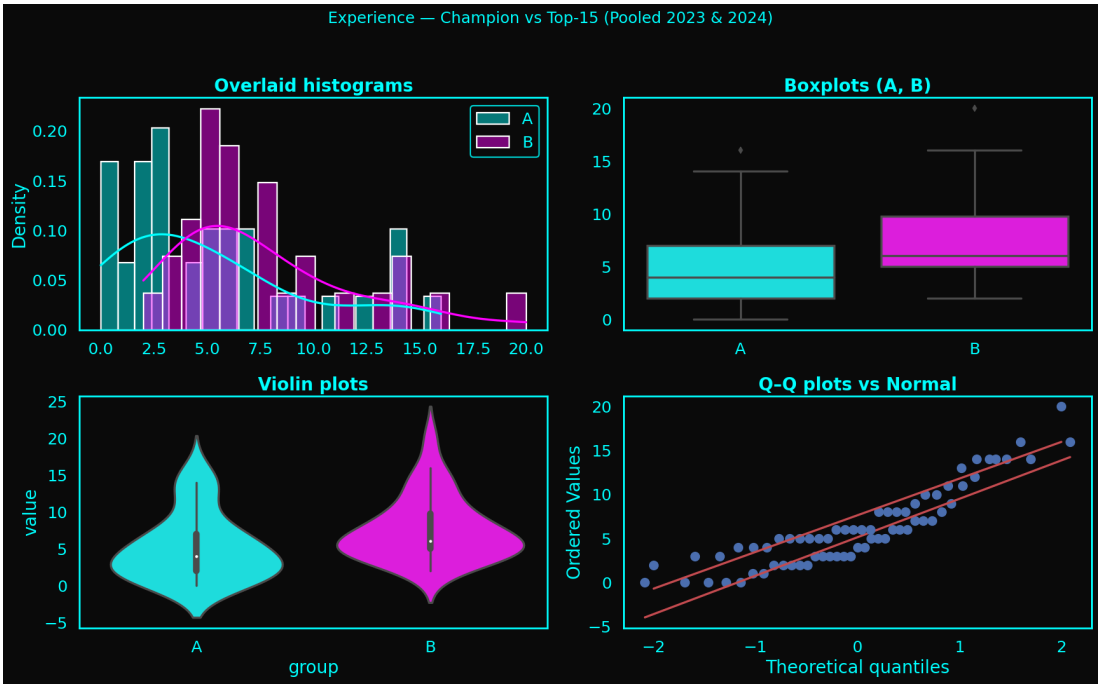
آزمون Welch t-test (اگر نرمال بودن قابل قبول باشد) یا آزمون Mann–Whitney استفاده کرد.

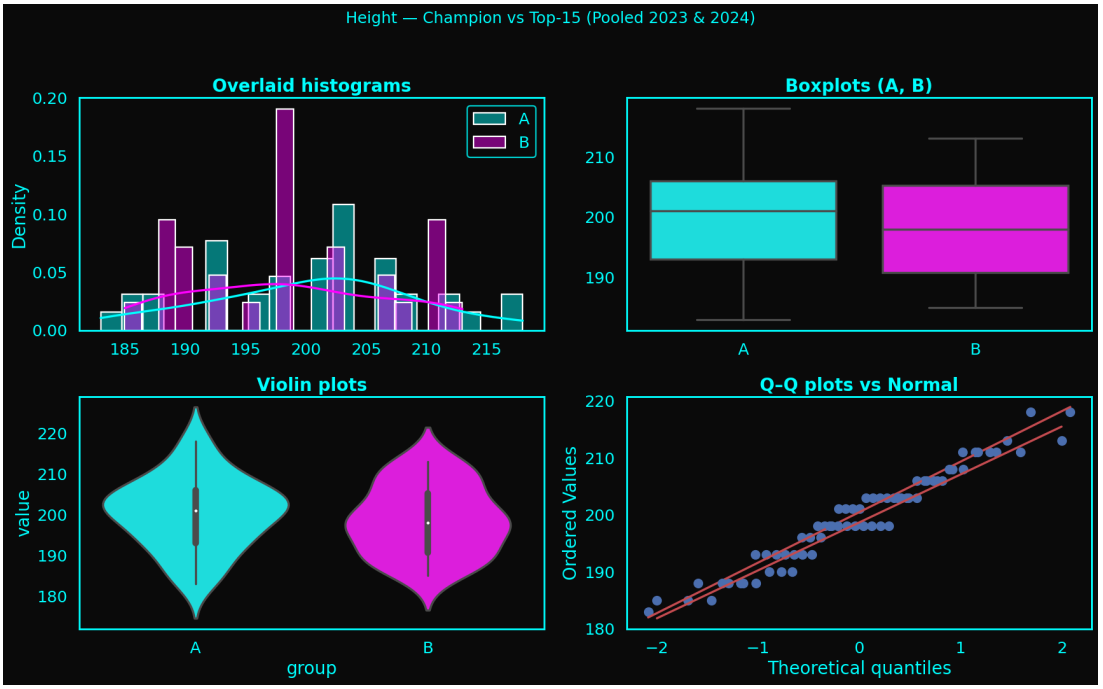


سوال دوم توزیع میزان تجربه افراد فعال، در تیم قهرمان و قد در دو فصل آخر را با توزیع میزان تجربه و قد ۱۵ بازیکن برتر این فصل مقایسه کردیم. در نتیجه توزیع تجربه بازیکنان قهرمان نسبت به بازیکنان *Top 15* کمی به سمت راست شیفت دارد و توزیع قد به صورت نرمال است، یعنی به‌طور متوسط قد بازیکنان قهرمان کمی بلندتر است اما تجربه ها تفاوت زیادی ندارند.
پراکندگی پهنای جعبه‌ها و نمودارهای ویولین تقریباً مشابه است و واریانس کلی هر دو گروه نزدیک به هم است؛ اما گروه B کمی دم‌های سنگین‌تری دارد، به‌ویژه در سمت کوتاه‌تر.

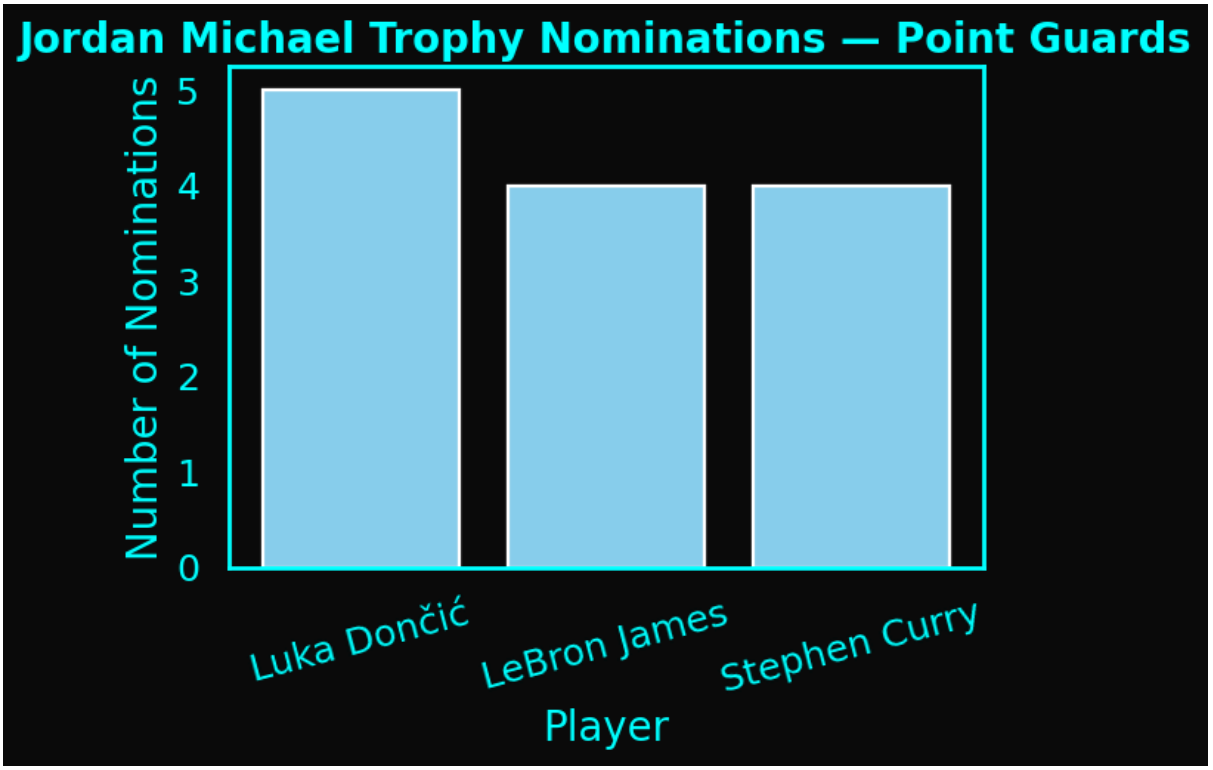
شکل توزیع/نرمال بودن هر دو گروه تقریباً تک‌قله‌ای (unimodal) و برای قد نزدیک به توزیع نرمال هستند.
در نمودار Q–Q، کوانتیل‌های میانی نزدیک خط قرار دارند و به طور کلی پراکندگی متعادل است.

پس تجربه بازیکنان تیم قهرمان فصل ها کمی بیشتر از بازیکنان Top-15 است، اما این تفاوت بزرگ نیست؛ گروه Top-15 دامنه گسترده‌تری را پوشش می‌دهد.
برای بررسی اینکه آیا این تفاوت کوچک از نظر آماری معنادار است یا خیر، می‌توان از آزمون Welch t-test (اگر نرمال بودن قابل قبول باشد) یا آزمون Mann–Whitney استفاده کرد.





در سوال سوم از ما بازیکنی در پوزیشن Guard Point خریداری کند که توانایی بالایی داشته باشد. معیار توانایی برای این باشگاه، وجود بازیکن در لیست Trophy Jordan Michael میباشد و بازیکنی که حضور بیشتری داشته، اولویت بالاتری دارد. لیستی از بازیکنان مناسب برای خرید با توجه به آمار فصلها از فصل ۲۰۲۰-۲۰۱۹ تا پایان ۲۰۲۴-۲۰۲۳ می خواهد. بر اساس داده‌ها، لوکا دونچیچ، لبرون جیمز و استفن کری بیشترین تعداد حضور را در فهرست جام مایکل جردن داشته‌اند. بنابراین، این سه بازیکن بهترین گزینه‌ها برای جذب در پست پوینت گارد هستند.



در سوال چهارم ادعا شد که میانگین چابکی افرادی که در ۲۰ نفر اول هر فصل حضور دارند، نسبت به گذشته افزایش یافته است. تعریف چابکی را میتوان نسبت قد به وزن افراد دانست. برای این سوال با کمک داده های جمع آوری شده و خروجی آزمون های آماری، با توجه به نسبت قد به وزن افراد که بدست آوردیم این ادعا رد شد. پس نمی توان این ادعا رو تایید کرد.

```
{'test_used': 'Independent t-test (equal var)',
 'alpha': 0.05,
 'n_a': 40,
 'n_b': 40,
 'statistic': 0.03278580009701344,
 'p_value': 0.9739291542122088,
 'significant': False,
 'effect_size': 0.007331127771364125}
```

در سوال پنجم هم کارشناسی مدعی شد امروزه به دلیل پیشرفت و بهبود شرایط، استفاده و شکوفا شدن توانایی ذاتی افراد نسبت به گذشته بهبود یافته است و برای مثال، میزان میانگین توانایی ذاتی بازیکنهای تیم قهرمان ۲ فصل اخیر، از ۲ فصل قبلی آن بیشتر بوده است. او تعریف توانایی ذاتی افراد را نسب میزان تجربه به سن فرد اعالم می کند. اما با توجه به نتیجه آزمون آماری داده ها نسبت به داده هایی که بدست آوردیم، این ادعا رد می شود.

```
{'test_used': 'Mann-Whitney U Test',
 'alpha': 0.05,
 'n_a': 37,
 'n_b': 39,
 'statistic': 755.0,
 'p_value': 0.7310674720727703,
 'significant': False,
 'effect_size': -0.04643104643104645}
```