NBA Player Salaries (2022-23 Season)

Projekt PAD

Daniel Fekete

Úvod

V analýze sme použili dataset z adresy: https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season/data Tento dataset poskytuje informácie o hráčoch NBA v sezóne 2022 - 2023 a budeme z neho používať tietot premenné:

- Player Name: Meno hráča
- Salary: Plat hráča
- Position: Pozícia na ihrisku (PG, PF, atď.)
- Age: Vek hráča
- Team: Tím, za ktorý hrá
- GP: Počet odohraných zápasov
- GS: Počet zápasov, v ktorých hráč začal v základnej zostave
- PTS: Priemerný počet bodov na zápas

Na začiatok naimportujeme dôležité a potrebné knižnice:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
from scipy import stats
from scipy.stats import ttest_ind, levene
from scipy.stats import chi2_contingency

import seaborn as sns
```

Import datasetu:

```
df = pd.read_csv("nba_2022-23_all_stats_with_salary.csv")
```

Vykreslenie časti dataframu:

```
df.head()
```

```
   Unnamed: 0        Player Name    Salary Position  Age     Team  GP
GS  \
```

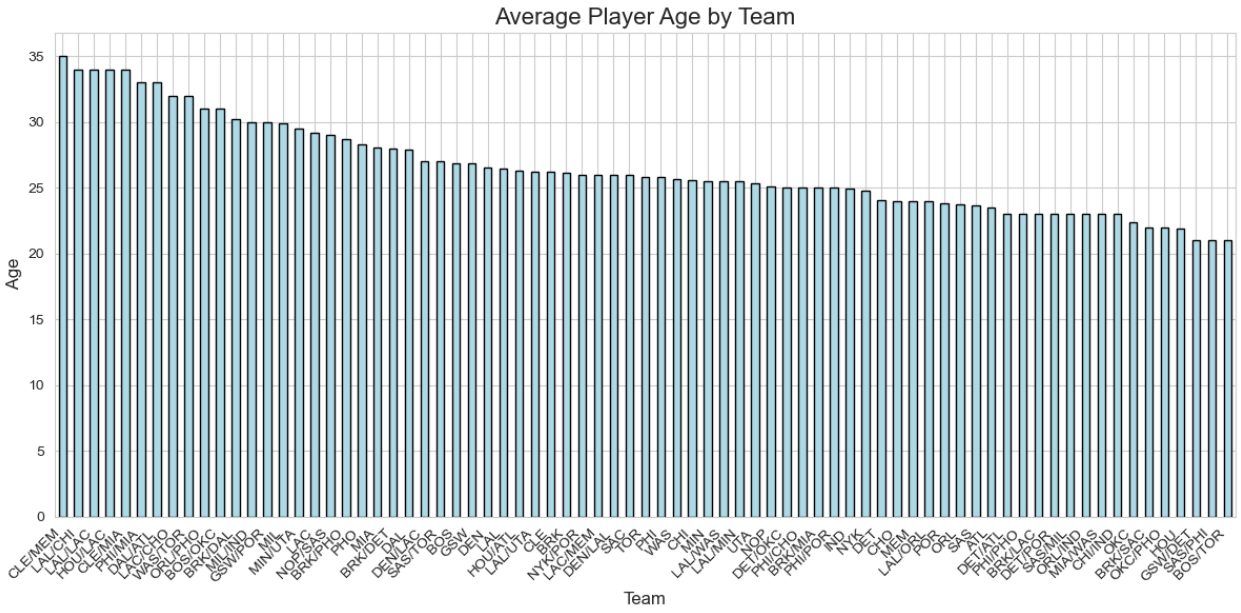
```
0           0      Stephen Curry  48070014       PG   34      GSW  56
56
1           1          John Wall  47345760       PG   32      LAC  34
3
2           2  Russell Westbrook  47080179       PG   34  LAL/LAC  73
24
3           3       LeBron James  44474988       PF   38      LAL  55
54
4           4       Kevin Durant  44119845       PF   34  BRK/PHO  47
47

     MP    FG  ...  TOV%  USG%  OWS  DWS   WS  WS/48  OBPM  DBPM  BPM
VORP
0  34.7  10.0  ...  12.5  31.0  5.8  2.0  7.8  0.192   7.5   0.1  7.5
4.7
1  22.2   4.1  ...  17.1  27.0 -0.4  0.7  0.3  0.020  -0.8  -0.4 -1.2
0.1
2  29.1   5.9  ...  18.4  27.7 -0.6  2.6  1.9  0.044   0.3  -0.1  0.2
1.2
3  35.5  11.1  ...  11.6  33.3  3.2  2.4  5.6  0.138   5.5   0.6  6.1
4.0
4  35.6  10.3  ...  13.4  30.7  4.7  2.1  6.8  0.194   6.0   1.2  7.1
3.9

[5 rows x 52 columns]
```

```
data_cleaned = df.drop(columns=['Unnamed: 0'])
```

Exploračná analýza - prehľad

Vytvoríme základné vizualizácie

```
age_bins = pd.cut(data_cleaned['Age'], bins=[18, 23, 28, 33, 38, 43],
labels=['18-23', '24-28', '29-33', '34-38', '39-43'])
age_distribution = age_bins.value_counts().sort_index()

average_age_by_team = df.groupby('Team')
['Age'].mean().sort_values(ascending=False)
plt.figure(figsize=(12, 6))
average_age_by_team.plot(kind='bar', color='lightblue',
edgecolor='black')
plt.title('Average Player Age by Team', fontsize=16)
plt.ylabel(' Age', fontsize=12)
plt.xlabel('Team', fontsize=12)
plt.xticks(rotation=45, ha='right', fontsize=10)
plt.tight_layout()
plt.show()
```

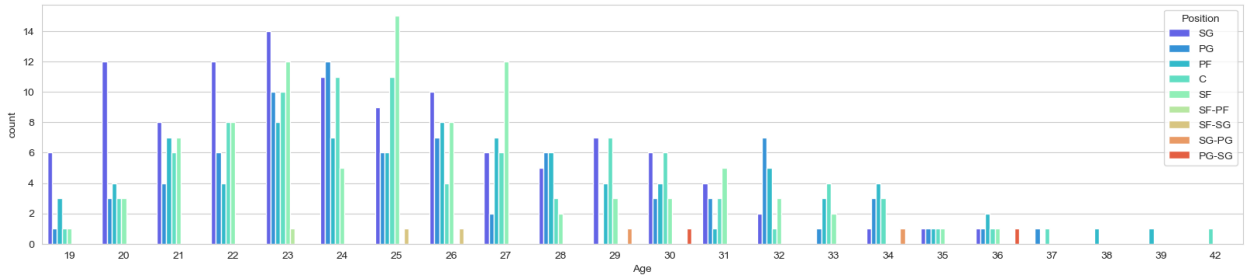


Záver:

Graf ukázal priemerný vek hráčov v tímoch. Z výsledkov môžeme vyvodiť, že medzi jednotlivými tímami existujú rozdiely v priemernom veku hráčov.

```
plt.figure(figsize=(20,4))
sns.set_style('whitegrid')
sns.countplot(x='Age',hue='Position', data=df,palette='rainbow')
```

```
<Axes: xlabel='Age', ylabel='count'>
```

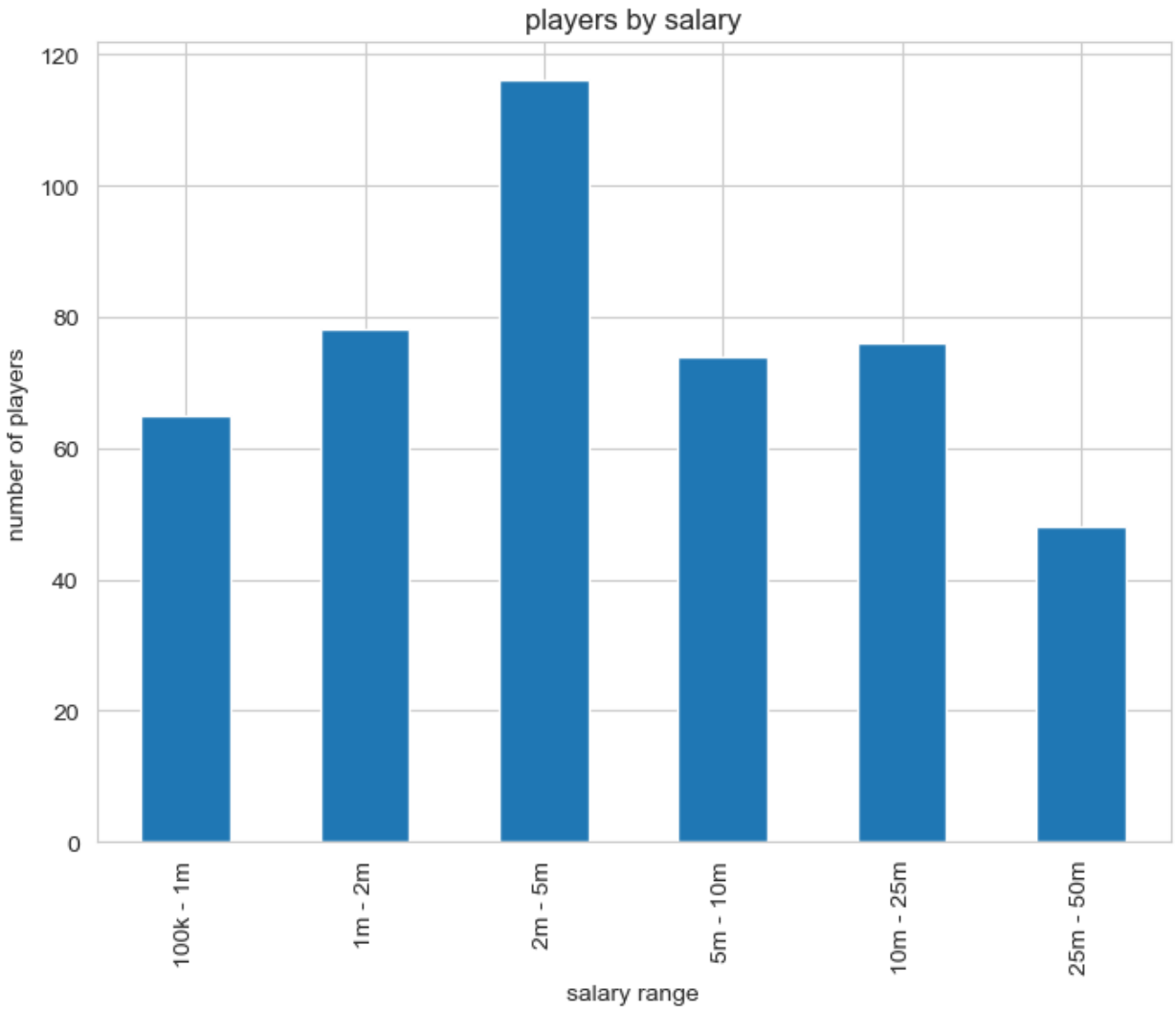


Záver:

Vekové rozloženie hráčov NBA vykazuje určitú symetriu, podobnú normálnemu rozdeleniu. Najviac hráčov patrí do vekového rozpätia medzi 22 a 27 rokmi. V datasetoch nie sú zaznamenaní hráči starší ako 43 rokov, čo naznačuje vekový limit pre profesionálnych hráčov NBA.

```
salary_bins = pd.cut(data_cleaned['Salary'], bins=[100000,
1000000,2000000, 5000000,10000000, 25000000, 50000000], labels=['100k
- 1m', '1m - 2m','2m - 5m',  '5m - 10m', '10m - 25m', '25m - 50m'])
salary_distribution = salary_bins.value_counts().sort_index()
plt.figure(figsize=(8, 6))
```

```
salary_distribution.plot(kind='bar')
plt.title('players by salary')
plt.xlabel('salary range')
plt.ylabel('number of players')
plt.show()
```

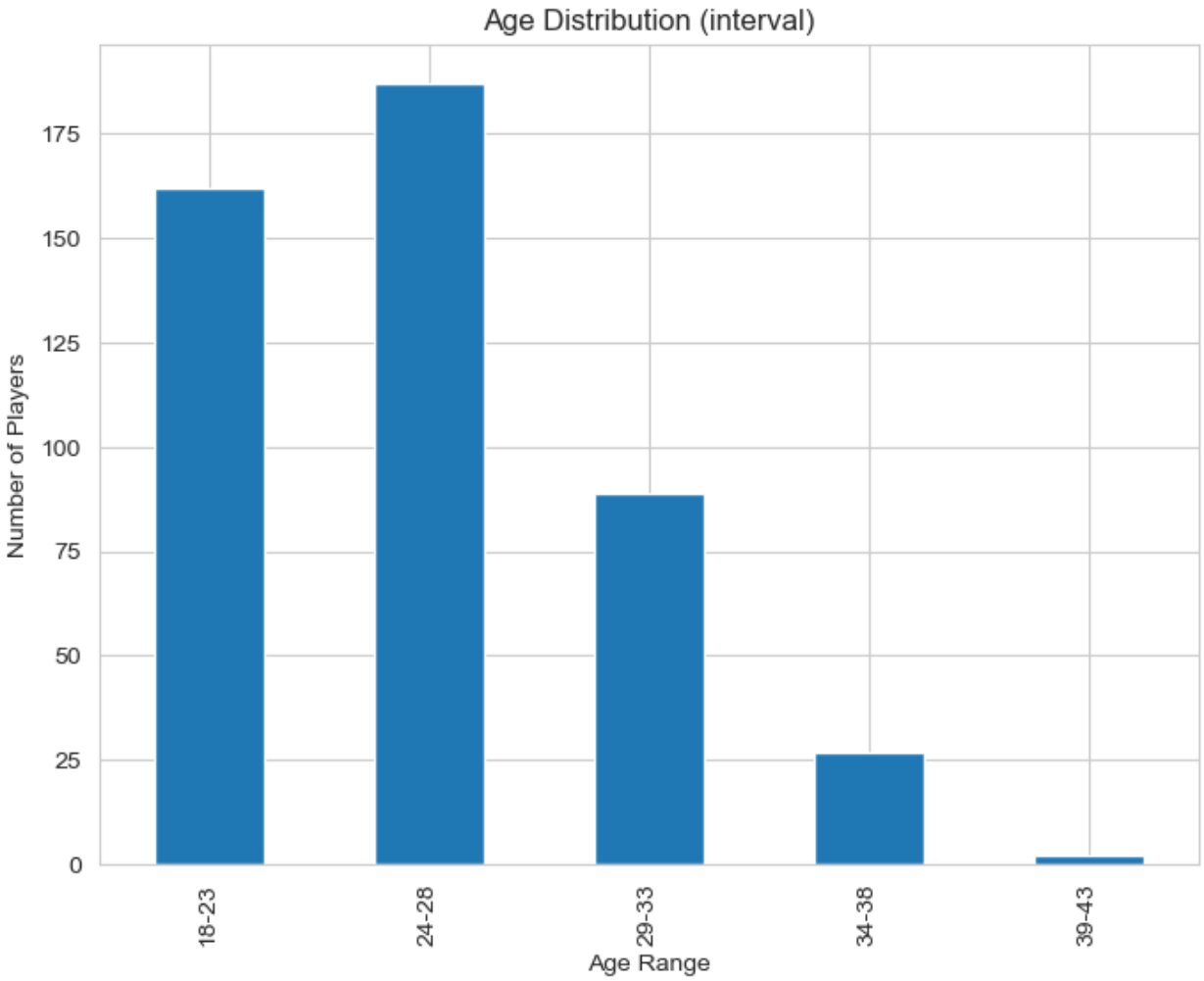


Záver:

Graf ukazuje, že väčšina hráčov má platy v intervaloch 100 tisíc až 2 milióny dolárov, pričom počet hráčov postupne klesá s vyššími platovými kategóriami. Najmenej hráčov zarába v rozmedzí 25 až 50 miliónov dolárov, čo odráža, že len elitní hráči dosahujú najvyššie zárobky.

```
plt.figure(figsize=(8, 6))

age_distribution.plot(kind='bar')
plt.title('Age Distribution (interval)')
plt.xlabel('Age Range')
```

```
plt.ylabel('Number of Players')
plt.show()
```

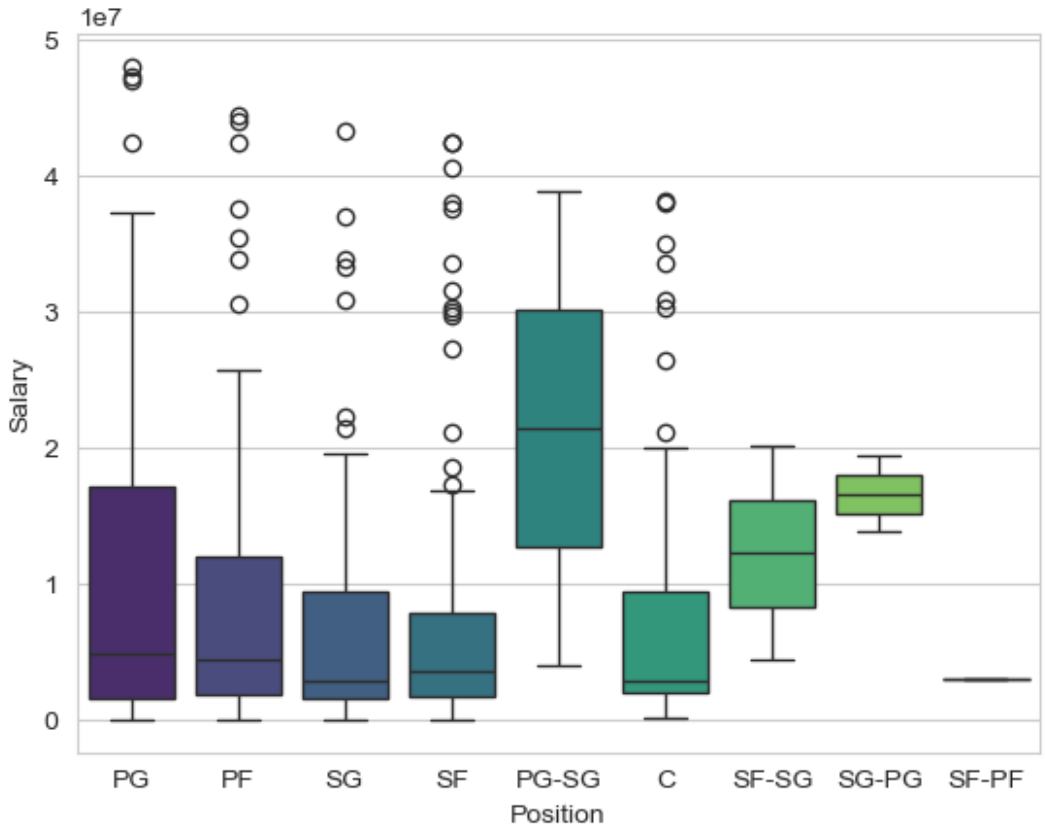


Graf znázorňuje vekové rozloženie hráčov NBA rozdelené do intervalov. Väčšina hráčov sa nachádza vo vekových kategóriách 24-28 rokov a 18-23 rokov, čo potvrdzuje, že väčšina profesionálnych hráčov NBA je vo svojom fyzickom vrchole počas tohto obdobia.

```
sns.boxplot(x='Position',y='Salary',data=df,palette='viridis')
```

```
C:\Users\danie\AppData\Local\Temp\ipykernel_59560\4222072432.py:1:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.boxplot(x='Position',y='Salary',data=df,palette='viridis')
```

```
<Axes: xlabel='Position', ylabel='Salary'>
```



Záver: Na základe krabicového grafu vidíme že niektoré pozície, ako napríklad PG, majú tendenciu vykazovať širší rozsah platov. Platy na pozicií PG-SG sa zdajú vacšie pretoze z nich máme menej dátových bodov

Čo ďalej: Použijeme metódy inferenčnej analýzy na overenie, či je medzi nimi štatisticky významný rozdiel.

Analýza 1

Inferenčná analýza - testy o stredných hodnotách

```
Chceme overiť, či je statisticky významný rozdiel medzi vekom Guards ( G ) a Forwards ( F )Naformulujeme preto nulovú hypotézu.
H0: Nie je Statisticky významný rozdiel medzi vekom Guards ( G ) a Forwards ( F )
```

Na overenie použijeme dvojvýberový t-test, keďže vek Guards a Forwards sú nezávislé. Pred aplikáciou t-testu však overíme rovnosť rozptylov pomocou Leveneovho testu. Táto analýza nám umožní zistiť, či veková štruktúra medzi týmito dvoma skupinami hráčov vykazuje štatisticky významné rozdiely.

```
guards = data_cleaned[data_cleaned['Position'].str.contains('G')]
['Age'].dropna()
forwards = data_cleaned[data_cleaned['Position'].str.contains('F')]
['Age'].dropna()

levene_stat, levene_p = levene(guards, forwards)
t_stat, t_p = ttest_ind(guards, forwards, equal_var=True)

{"Levene's Test": {"Statistic": levene_stat, "p-value": levene_p}, "t-
Test": {"Statistic": t_stat, "p-value": t_p}}
```

```
{"Levene's Test": {'Statistic': 0.01799705078116791,
  'p-value': 0.893353298456712},
 't-Test': {'Statistic': -1.4167319866640027, 'p-value':
0.1573892765075624}}

Na Základe výsledkov zamietame H0.

Záver: Vek Guards a Forwards sa podľa týchto analýz štatisticky
významne nelíši, a ich rozptyl je podobný.
```

Analýza 2

Inferenčná analýza - Korelačná analýza

Cieľ analýzy: Dataset o hráčoch NBA obsahuje (okrem iného) informácie o ročných platoch hráčov (Salary) a priemernom počte bodov na zápas (PTS) a počet odohraných zápasov (GP).

Cieľom nasledovnej analýzy bude zistiť, či existuje významná závislosť medzi týmito parametrami a určiť, do akej miery výkon hráča (reprezentovaný počtom bodov) a počet odohraných zápasov ovplyvňuje jeho finančné ohodnotenie.

Budeme teda zisťovať úroveň korelačného koeficientu medzi dvojicami:

- Salary X PTS
- Salary X GP

Formulácia nulových hypotéz:

```
H0 (A): Nie je štatisticky významná závislosť medzi premennými Salary
a PTS.
H0 (B): Nie je štatisticky významná závislosť medzi premennými Salary
a GP.
```

Keďže máme veľký počet údajov, použijeme parametrické metódy korelačnej analýzy na overenie tejto závislosti.

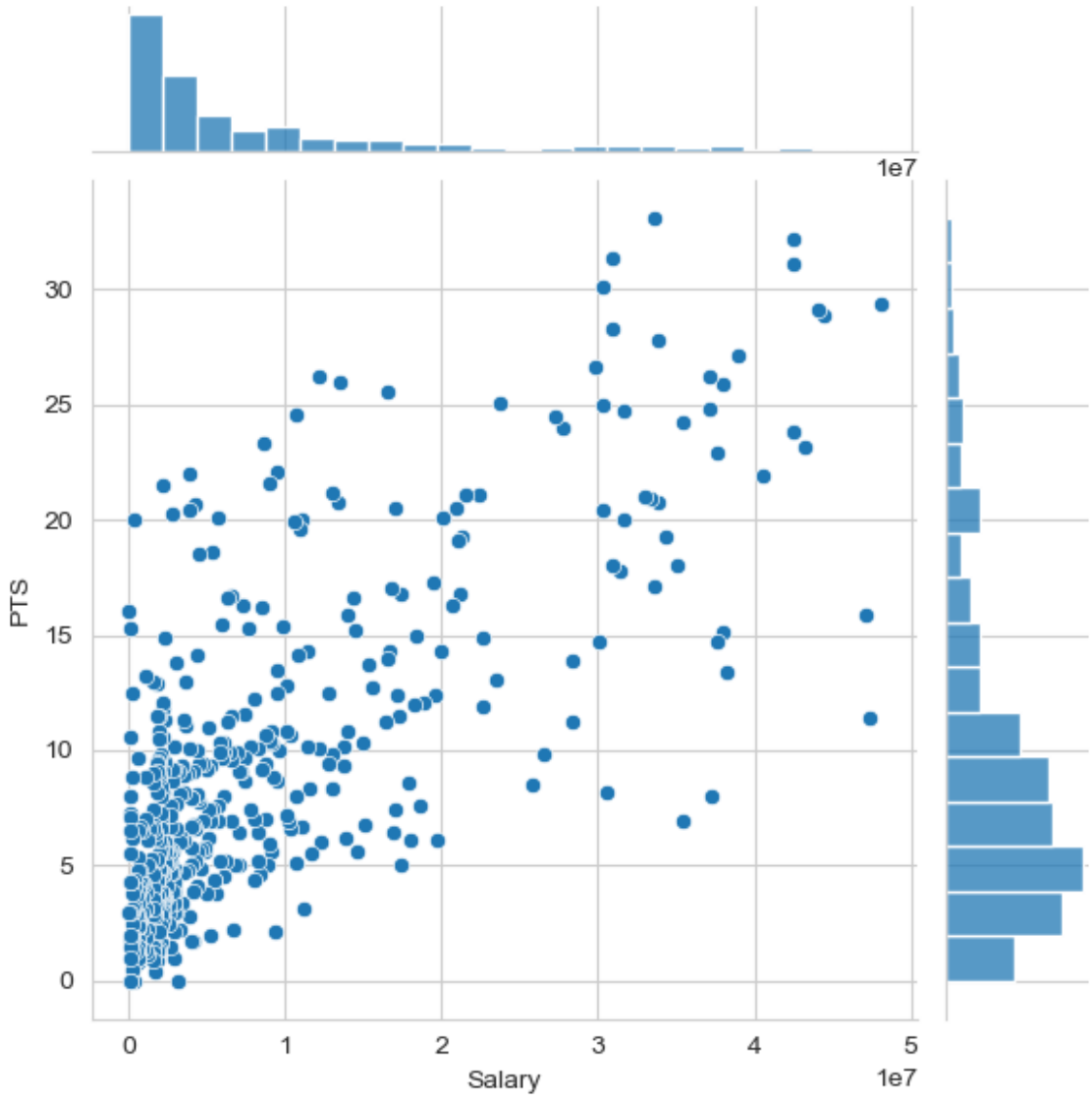
```
df[["Salary","PTS"]].corr()
```

```
          Salary       PTS
Salary  1.000000  0.727597
PTS     0.727597  1.000000

Vidíme, že medzi tým, že medzi počtom bodov a platom je vysoká
závislosť. Korelačný koeficient je až 0,73.

Overíme pre istotu aj p hodnotu.

stats.pearsonr(df["Salary"],
               df["PTS"])

PearsonRResult(statistic=0.7275966928493116,
pvalue=3.9581622553381053e-78)
```

Záver: Na základe výsledkov hypotézu H0 (A) zamietame. Znamená to, že je štatisticky významná závislosť medzi počtom bodov a platom.

Následne overujeme H0 (B)

```
sns.jointplot(data=df, x="Salary", y="PTS")

<seaborn.axisgrid.JointGrid at 0x1c264e8cd40>
```




```
df[["Salary","GP"]].corr()
```

```
            Salary        GP
Salary    1.000000  0.341707
GP        0.341707  1.000000
```

Na základe korelačnej matice je korelačný koeficient medzi platmi hráčov a počtom odohraných zápasov (GP) 0.34
čo naznačuje mierne pozitívnu závislosť. To znamená, že hráči s vyšším platom majú tendenciu odohrať viac zápasov, avšak vzťah nie je silný.

Overíme to aj pomocou p hodnoty.

```
stats.pearsonr(df["Salary"],
               df["GP"])
```

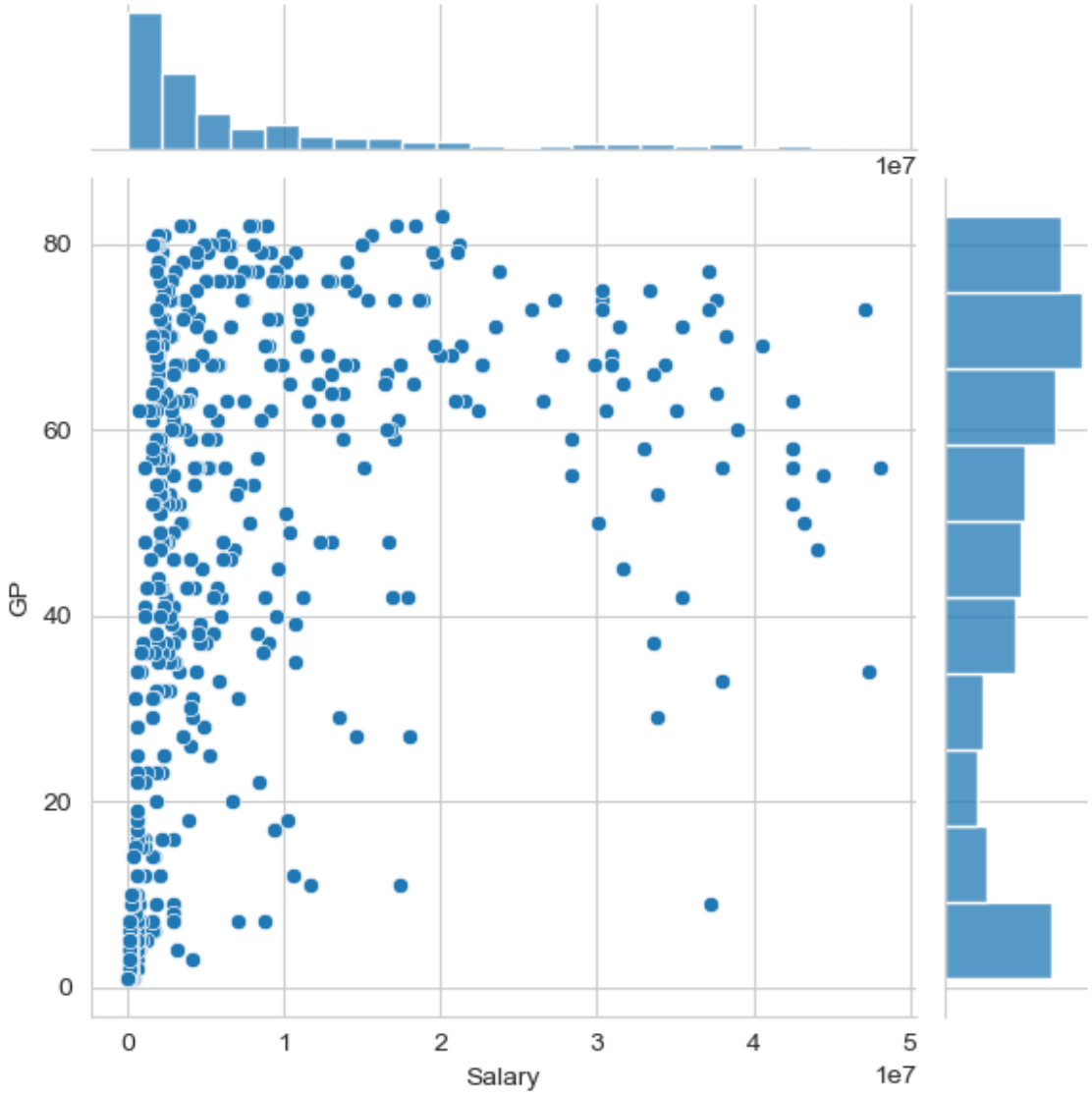
```
PearsonRResult(statistic=0.34170652737819396,
pvalue=3.0936906724153366e-14)
```

P-hodnota je veľmi nízka p < 0.001, čo znamená, že korelácia je štatisticky významná. Na základe týchto výsledkov zamietame nulovú hypotézu H0 (B)

Záver: Výsledky naznačujú, že výkon hráča, vyjadrený počtom bodov na zápas, má väčší vplyv na jeho plat ako počet odohraných zápasov. Teda kvalitní hráči, ktorí pravidelne skórujú, sú finančne odmeňovaní viac.

```
sns.jointplot(data=df, x="Salary", y="GP")
```

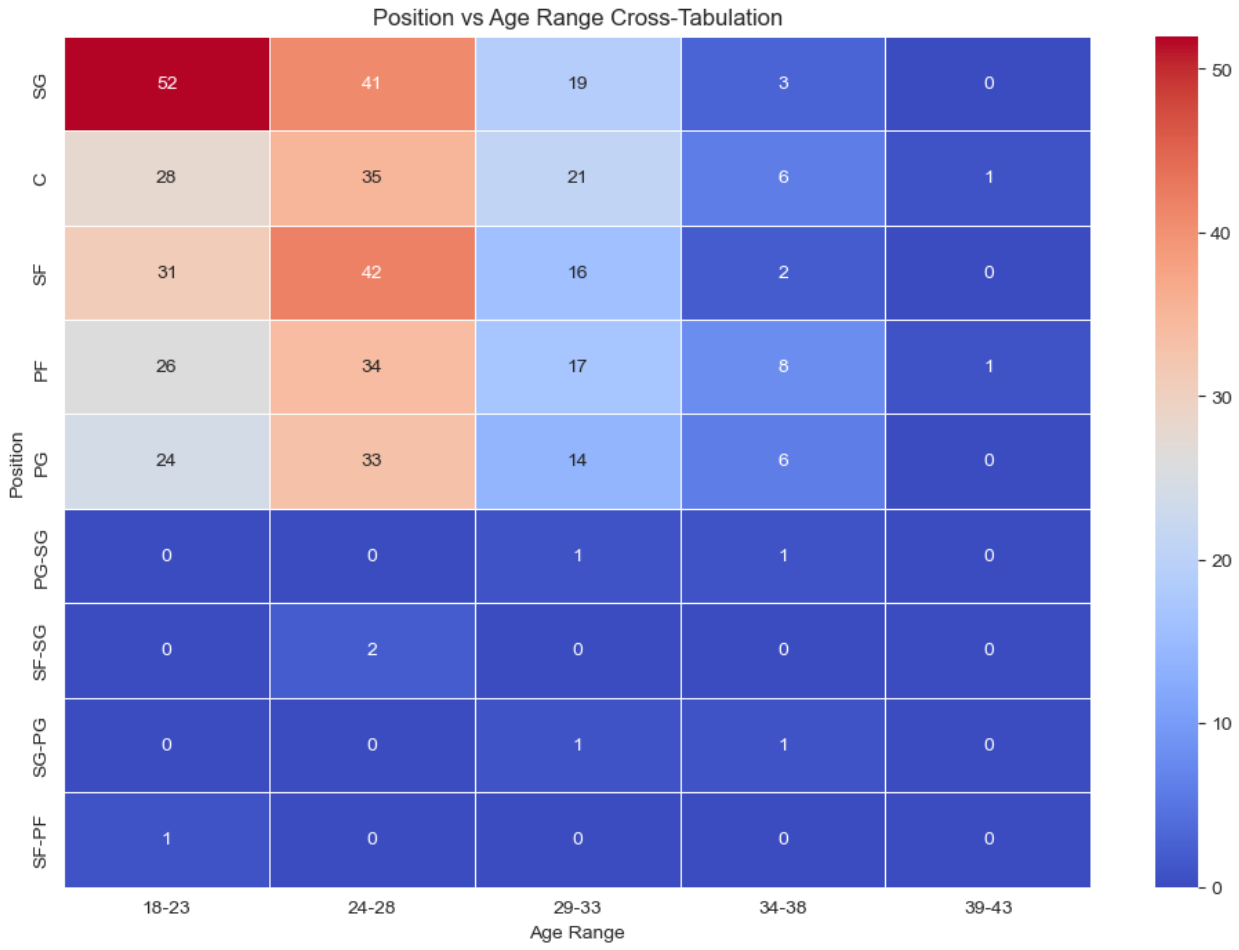
```
<seaborn.axisgrid.JointGrid at 0x1c263575430>
```



Analýza 4

Analýza kontingencie

```
Cielom tejto analyzy je preskúmať, či existuje vztah medzi pozíciou a
vekom
Zvolíme si nulovú hypotézu:
H0: Vek neovplyvňuje pozíciu hráča, premenné sú nezávislé

position_age_ct = pd.crosstab(df['Position'], age_bins)
position_age_ct['Total'] = position_age_ct.sum(axis=1)
position_age_ct_sorted = position_age_ct.sort_values('Total',
ascending=False).drop(columns=['Total'])
plt.figure(figsize=(12, 8))
sns.heatmap(position_age_ct_sorted, annot=True, fmt='d',
cmap='coolwarm', linewidths=0.5)
plt.title('Position vs Age Range Cross-Tabulation')
plt.xlabel('Age Range')
plt.ylabel('Position')
plt.show()
```



```
chi2, p, dof, ex = chi2_contingency(position_age_ct)

print(f"Chi2: {chi2}")
print(f"p-hodnota: {p}")
```

```
Chi2: 40.97124788737964
p-hodnota: 0.13292728895148495
```

Na základe p-hodnoty (0.133) nie je dôvod zamietnuť nulovú hypotézu. To naznačuje, že neexistuje štatisticky významná závislosť medzi pozíciou hráča a jeho vekovou kategóriou.

Záver

Analýza hráčov NBA v sezóne 2022-2023 zahŕňala rôzne štatistické prístupy na pochopenie vzťahov medzi platmi hráčov, ich výkonnosťou a demografickými ukazovateľmi. Kľúčové poznatky vyplývajúce z interferenčných a kontingenčných analýz sú nasledovné:

Silná závislosť medzi výkonom a platom: Existuje významná pozitívna korelácia medzi priemerným počtom bodov na zápas a výškou platu. Tento vzťah zdôrazňuje dôležitosť výkonu hráča pre jeho finančné ohodnotenie. Slabšia súvislosť medzi počtom zápasov a platom: Počet odohraných zápasov má slabšiu pozitívnu koreláciu s platom. To naznačuje, že pravidelná účasť v zápasoch je dôležitá, ale nie rozhodujúca pre finančné odmeny. Pozície hráčov: Vizualizácie ukázali, že niektoré pozície majú širšie rozdelenie platov, čo môže odrážať rozdielnu dôležitosť a špecializáciu v tíme. Tento výskum zdôrazňuje dôležitosť špičkovej výkonnosti pri finančnom ohodnocovaní hráčov a poskytuje pevný základ pre ďalšie analýzy v oblasti športovej ekonómie.