Zach Rickman & Dalton Price
STAT 3010 Final Project
Auburn University
zlr0005@auburn.edu & dhp0015@auburn.edu

Efficiency With NBA Players' Contracts

27th April 2022

TABLE OF CONTENTS

SECTION 1: OVERVIEW

In our project we set out to identify a correlation between how much an NBA player is paid, along with their age, to their production on the court.

(Pictured: G. Antetokounmpo driving against J. Tatum)

The dataset used consists of two hundred randomly selected contracts of NBA players that were fulfilled sometime within the years 2010-2020. It contains information about the player name, time span of the contract, average salary per year, and all stats that the player accumulated during the NBA season before signing a contract.

Here is an example of a few of the data points and variables collected

(200 player contracts total, along with 28 columns of variables)

	NAME	CONTRACT_START	CONTRACT_END	AVG_SALARY	AGE	GP	W	L	MIN	PTS	...
0	Wesley Matthews	2019	2020	2564753.0	32.0	69.0	27.0	42.0	2091.0	840.0	...
1	Brook Lopez	2015	2017	21165675.0	27.0	72.0	34.0	38.0	2100.0	1236.0	...
2	DeAndre Jordan	2011	2014	10759763.5	22.0	80.0	31.0	49.0	2047.0	566.0	...
3	Markieff Morris	2015	2018	8143323.5	25.0	82.0	39.0	43.0	2581.0	1258.0	...
4	Dwight Howard	2018	2019	13410739.0	32.0	81.0	35.0	46.0	2463.0	1347.0	...

SECTION 2: DATA COMPARISONS AND VISUALIZATION

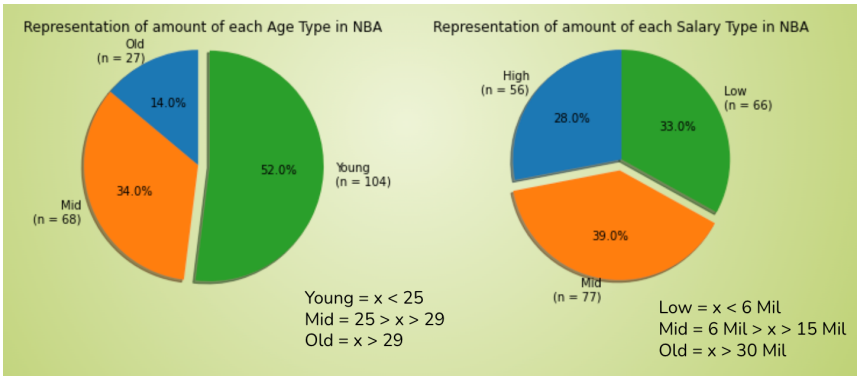
Null hypothesis: If a player is paid more and is more experienced, then their stats are more productive.
Alternative: Salary and/or age have no correlation to the productivity of a player.

This table contains certain variables we deemed the most important data to work with for our research (AVG_SALARY, AGE, W, PTS, GP, FG%). With each variable, this table describes different statistics formulas on the very left column. These calculations give us a grasp of the count, mean, quartiles, etc. (see below).

	AVG_SALARY	AGE	W	PTS	GP	FG%
count	1.990000e+02	199.000000	199.000000	199.000000	199.000000	199.000000
mean	1.107361e+07	25.934673	34.216080	813.447236	64.170854	46.743719
std	7.897820e+06	2.842810	14.485749	499.930031	19.573765	8.094826
min	8.232440e+05	20.000000	0.000000	0.000000	1.000000	0.000000
25%	4.767000e+06	24.000000	25.000000	443.500000	59.000000	42.850000
50%	9.500000e+06	25.000000	35.000000	734.000000	72.000000	45.500000
75%	1.638890e+07	28.000000	45.000000	1154.000000	78.000000	49.800000
max	3.359950e+07	36.000000	64.000000	2376.000000	82.000000	100.000000

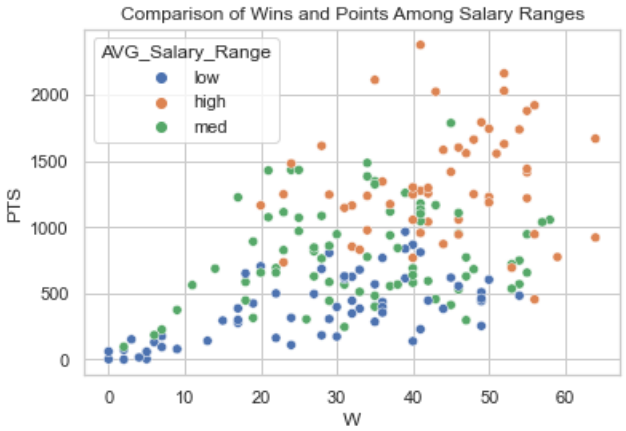
The next table adds two categorical variables (AGE_Range & AVG_Salary_Range) containing three ranges of data each from their respective data points (AVG_SALARY & AGE).

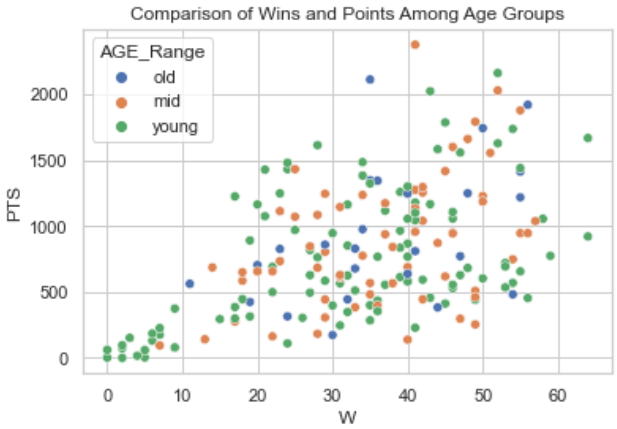
	AVG_SALARY	AGE	W	PTS	GP	FG%	AGE_Range	AVG_Salary_Range
0	2564753.0	32.0	27.0	840.0	69.0	40.0	old	low
1	21165675.0	27.0	34.0	1236.0	72.0	51.3	mid	high
2	10759763.5	22.0	31.0	566.0	80.0	68.6	young	med
3	8143323.5	25.0	39.0	1258.0	82.0	46.5	young	med
4	13410739.0	32.0	35.0	1347.0	81.0	55.5	old	med



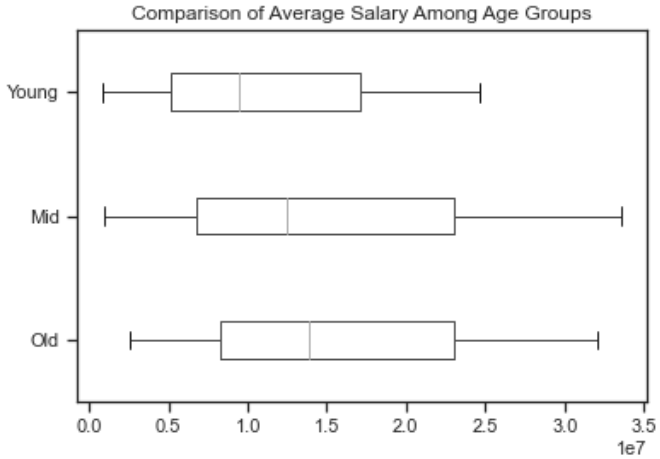
SECTION 2: DATA COMPARISONS AND VISUALIZATION

From these two scatterplots we can see how a player's wins correlate with the amount of points they score in a season. While the trend is not absolute, this comparison shows that players who on average score more points are more likely to accumulate more wins. It also depicts that age does not necessarily impact success, while high salary players tend to perform on average.



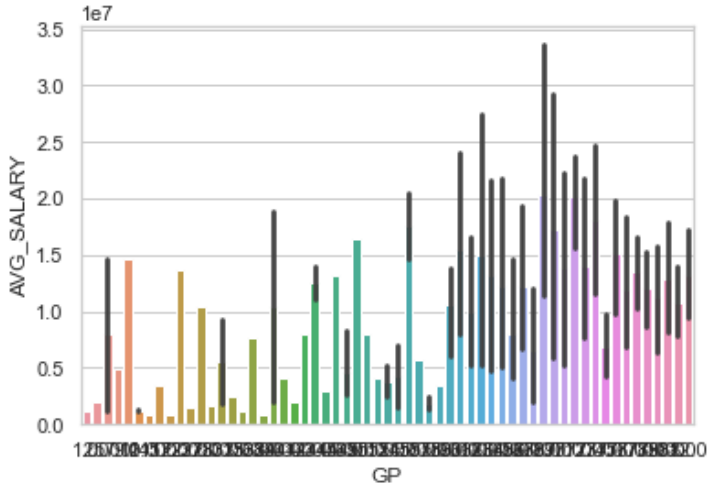


We find from this graph a comparison between the variables depicted within our categorical variables. Young players are not proven and do not make the same salary as "Mid" or "Old" players. There is also a great difference in outliers with the "Young" category. (X-axis: $0-$35 Million).

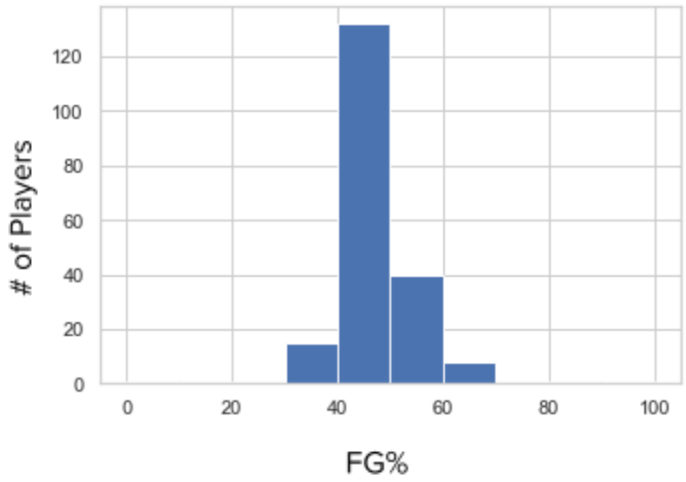


SECTION 2: DATA COMPARISONS AND VISUALIZATION

This graph shows that on average the more games you play, the salary is more consistently high. The peak is at the 67 game mark, which can indicate that stars are rested to not risk injury.



The histogram below shows the distribution of players by FG%. This data is seen to be possibly normal, so we can dissect this with further testing.



SECTION 3: TESTS & ANOVA RESEARCH

In this section we are testing how FG% correlates with our two categorical variables.

Number of Unique Values:

High 56
Med 77
Low 66
Name: AVG_Salary_Range, dtype: int64

One Way ANOVA Test

F_onewayResult(statistic=4.120524424061244, p value=0.01766354429199324e-)

Normality Tests

These tests show that the High and Mid Salary Ranges consist of values with normal distribution regarding FG%

```
normaltest(H)
```
NormaltestResult(statistic=26.385633462656486, p value=1.86394e-06)
```
normaltest(M)
```
NormaltestResult(statistic=18.782768128629115, p value=8.343989e-09)
```
normaltest(L)
```
NormaltestResult(statistic=39.9474826386561, p value=2.115993e-05)

Equal Variance Test

LeveneResult(statistic=1.302824534530697, p value=0.2741059494835371)

Two Way ANOVA Test

This test shows us that there are significant enough differences in the groups tested.

	sum_sq	df	F	PR(>F)
C(AVG_Salary_Range)	519.098175	2.0	4.021843	0.019467
C(AGE_Range)	50.599748	2.0	0.392034	0.676226
C(AVG_Salary_Range):C(AGE_Range)	138.463386	4.0	0.536390	0.709161
Residual	12261.622883	190.0	NaN	NaN

SECTION 4: CONCLUSION REGARDING DATA, ANOVA, AND TESTS

When exploring our data, we found that players who on average score more points are more likely to accumulate more wins. It also shows that age does not necessarily impact success, while high salary players tend to perform better on average. Young players are not proven and do not make the same salary as “Mid” or “Old” players.

These tests show that the High and Low Salary Ranges consist of values with normal distribution regarding FG%. FG% was the only variable that we could find that was normally distributed. This concludes us to believe that stars and medium paid players score at a higher efficiency rating, which proves our original hypothesis.

On average it is worth it to pay your most productive players

SECTION 5: LIMITATIONS OF ANALYSIS

There were a few different limits to our analysis that arose some questions to ask. FG% can be a broad stat since it accounts for 2-pointers and 3-pointers, so a player can have a lower percentage with more ppg. Lastly, our histogram would not plot our findings by category, so we had to rely on our testing findings afterward.

What other statistics in the set are normally distributed that we did not get to explore? Is FG% the best indicator of skill? What leads to players becoming overpaid?

SECTION 6: PYTHON CODE AND SOURCES

```
import pandas as pd;
from scipy.stats import shapiro
df = pd.read_csv('nba_contracts_history 2.csv')
```

```
df.head()
```

```
subset = df[['AVG_SALARY', 'AGE', 'W', 'PTS', 'GP', 'FG%']]
fg = df[['FG%']]
```

```
subset.head()
```

In [336…

```
def age(x):
    if x > 29:
        return 'old'
    elif x > 25:
        return 'mid'
    else:
        return 'young'
subset['AGE_Range'] = subset['AGE'].map(age)

def salaries(x):
    if x > 15000000:
        return 'high'
    elif x > 6000000:
        return 'med'
    else:
        return 'low'
...
subset['AVG_Salary_Range'] = subset['AVG_SALARY'].map(salaries)
```

```
subset.head()
```

```
subset.groupby(['AVG_Salary_Range']).PTS.describe()
```

```
import seaborn as sn
import matplotlib.pyplot as plt
sn.scatterplot('W','PTS', hue = 'AGE_Range', data = subset)
plt.title('Comparison of Wins and Points Among Age Groups')
plt.show()
```

```
sn.scatterplot('W','PTS', hue = 'AVG_Salary_Range', data = subset)
plt.title('Comparison of Wins and Points Among Salary Ranges')
plt.show()
```

```
labels = 'Old', 'Mid', 'Young'
sizes = [14, 34, 52]
explode = (0, 0, 0.1)

fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')

plt.show()

labels = 'High', 'Mid', 'Low'
sizes = [28, 39, 33]
explode = (0, 0.1, 0)

fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')

plt.show()
```

```
import seaborn as sn
import matplotlib.pyplot as plt
sn.set_theme(style="whitegrid")
#tips = sn.load_dataset("subset")
ax = sn.barplot(x="GP", y="AVG_SALARY", data=subset)
```

```
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline
plt
plt.hist(df['FG%'])
plt.show()
```

```
import scipy.stats as stats
H = subset.loc[subset['AVG_Salary_Range']=='high', 'FG%']
M = subset.loc[subset['AVG_Salary_Range']=='med', 'FG%']
L = subset.loc[subset['AVG_Salary_Range']=='low', 'FG%']

stats.levene(H, M, L, center='mean')
```

```
subset['AVG_Salary_Range'].value_counts()
```

```
med     77
low     66
high    56
Name: AVG_Salary_Range, dtype: int64
```

```
from scipy.stats import f_oneway

f_oneway(H,M,L)
```

```
from scipy.stats import normaltest
normaltest(H)
```

```
NormaltestResult(statistic=26.385633462656486, pvalue=1.86
```

```
from scipy.stats import normaltest
normaltest(M)
```

```
NormaltestResult(statistic=18.782768128629115, pvalue=8.34
```

```
from scipy.stats import normaltest
normaltest(L)
```

```
NormaltestResult(statistic=39.9474826386561, pvalue=2.11
```

```
from scipy.stats import levene
stats.levene(H, M, L, center='median')
```

```
LeveneResult(statistic=1.302824534530697, pvalue=0.27410
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

#perform two-way ANOVA
model = ols('FG ~ C(AVG_Salary_Range) + C(AGE_Range) + C(AVG_Salary_Range):C(AGE_Range)', data = subset).fit()
sm.stats.anova_lm(model, typ=2)
```

Pictures for Project:

https://www.sportingnews.com/us/nba/news/heat-hawks-predictions-odds-schedule-live-streams-playoffs/twr0f0f93bxwpcsso3i0ujuz

https://www.silverscreenandroll.com/2020/3/27/21197055/giannis-antetokounmpo-jayson-tatum-list-top-players-almost-all-lakers-kobe-lebron-kareem-shaq-magic

Dataset:

https://www.kaggle.com/datasets/jarosawjaworski/current-nba-players-contracts-history?resource=download&select=nba_contracts_history.csv

Further help:

GeekforGeeks

Canvas STAT 3010 Lab3