SUPERVISED LEARNING CAPSTONE

NBA SALARIES FOR THE 2016-2017 SEASON

Introduction

- Dataset is available on Kaggle under Social Power NBA
- Contains data from Basketball Reference, ESPN, NBA.com, FiveThirtyEight, Twitter, and Wikipedia
- The data is from the 2016-2017 NBA season
- Essentially divided into two parts;
  - *Player data*
  - *Team data*

Question

What factors, on-court and off-court, contribute to an NBA player’s salary

Player DataFrame

- RK
- PLAYER
- POSITION
- AGE
- MP
- FG
- FGA
- FG%
- 3P
- 3PA
- 3P%
- 2P
- 2PA
- 2P%
- eFG%
- FT
- FTA
- FT%
- ORB
- DRB
- TRB
- AST
- STL
- BLK
- TOV
- PF
- POINTS
- TEAM
- GP
- ORPM
- DRPM
- RPM
- WINS_RPM
- PIE
- PACE
- W
- SALARY_MILLIONS
- PAGEVIEWS
- TWITTER_FAV
- TWITTER_RETWEET

Dealing with Null Values

```
players_df.isna().sum()
```

```
PLAYER                    0
POSITION                  0
AGE                       0
MP                        0
FG                        0
FGA                       0
FG%                       0
3P                        0
3PA                       0
3P%                       7
2P                        0
2PA                       0
2P%                       0
eFG%                      0
FT                        0
FTA                       0
FT%                       2
ORB                       0
DRB                       0
TRB                       0
AST                       0
STL                       0
BLK                       0
TOV                       0
PF                        0
POINTS                    0
TEAM                      0
GP                        0
ORPM                      0
DRPM                      0
RPM                       0
WINS_RPM                  0
PIE                       0
PACE                      0
W                         0
SALARY_MILLIONS           0
PAGEVIEWS                 0
TWITTER_FAVORITE_COUNT    3
TWITTER_RETWEET_COUNT     3
```

```
players_df[players_df.isna().any(axis=1)]
```

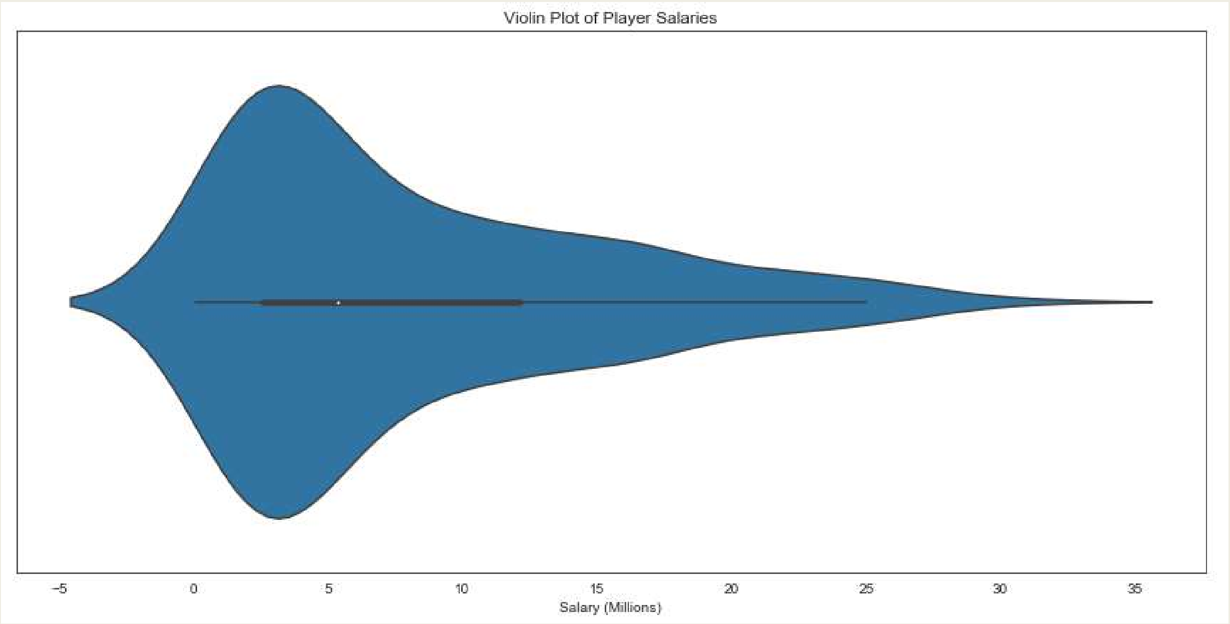
	PLAYER	POSITION	AGE	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TO
128	Tyson Chandler	C	34	27.6	3.3	4.9	0.671	0.0	0.0	NaN	3.3	4.9	0.671	0.671	1.9	2.6	0.734	3.3	8.2	11.5	0.6	0.7	0.5	1
146	David Lee	PF	33	18.7	3.1	5.3	0.590	0.0	0.0	NaN	3.1	5.3	0.590	0.590	1.0	1.4	0.708	1.9	3.7	5.6	1.6	0.4	0.5	1
175	Ian Mahinmi	C	30	17.9	2.1	3.6	0.586	0.0	0.0	NaN	2.1	3.6	0.586	0.586	1.4	2.4	0.573	1.5	3.3	4.8	0.6	1.1	0.8	1
204	Anthony Brown	SF	24	14.5	1.6	4.5	0.360	0.6	2.5	0.259	1.0	2.1	0.478	0.430	0.0	0.0	NaN	0.7	2.3	3.0	0.7	0.5	0.1	0
213	Dragan Bender	PF	19	13.3	1.3	3.7	0.354	0.7	2.3	0.277	0.7	1.4	0.483	0.441	0.1	0.3	0.364	0.5	1.9	2.4	0.5	0.2	0.5	0
222	Omer Asik	C	30	15.5	1.0	2.1	0.477	0.0	0.0	NaN	1.0	2.1	0.477	0.477	0.7	1.3	0.590	1.5	3.7	5.3	0.5	0.2	0.3	0
226	Miles Plumlee	C	28	10.8	1.0	2.0	0.478	0.0	0.0	NaN	1.0	2.0	0.478	0.478	0.6	0.9	0.641	0.8	1.3	2.1	0.5	0.4	0.3	0
230	Rakeem Christmas	PF	25	7.6	0.7	1.5	0.442	0.0	0.0	NaN	0.7	1.5	0.442	0.442	0.7	1.0	0.724	0.9	1.0	1.9	0.1	0.1	0.2	0
233	Cole Aldrich	C	28	8.6	0.7	1.4	0.523	0.0	0.0	NaN	0.7	1.4	0.523	0.523	0.2	0.4	0.682	0.8	1.7	2.5	0.4	0.4	0.4	0
235	Bruno Caboclo	SF	21	4.4	0.7	1.8	0.375	0.2	0.7	0.333	0.4	1.1	0.400	0.438	0.0	0.0	NaN	0.6	0.6	1.1	0.4	0.2	0.1	0
238	Alonzo Gee	SF	29	6.8	0.2	1.1	0.214	0.0	0.2	0.000	0.2	0.8	0.273	0.214	0.4	0.7	0.556	0.3	0.8	1.2	0.5	0.4	0.1	0

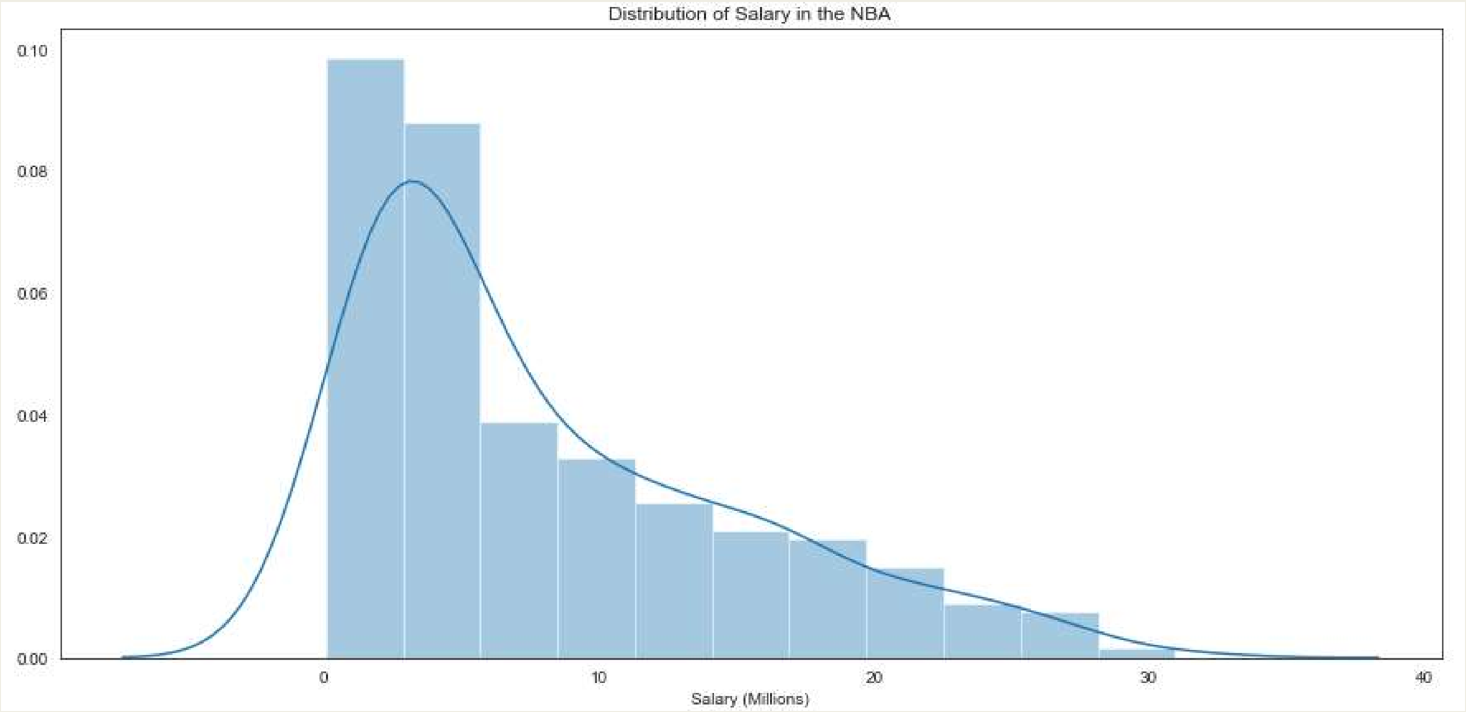
Player Salary

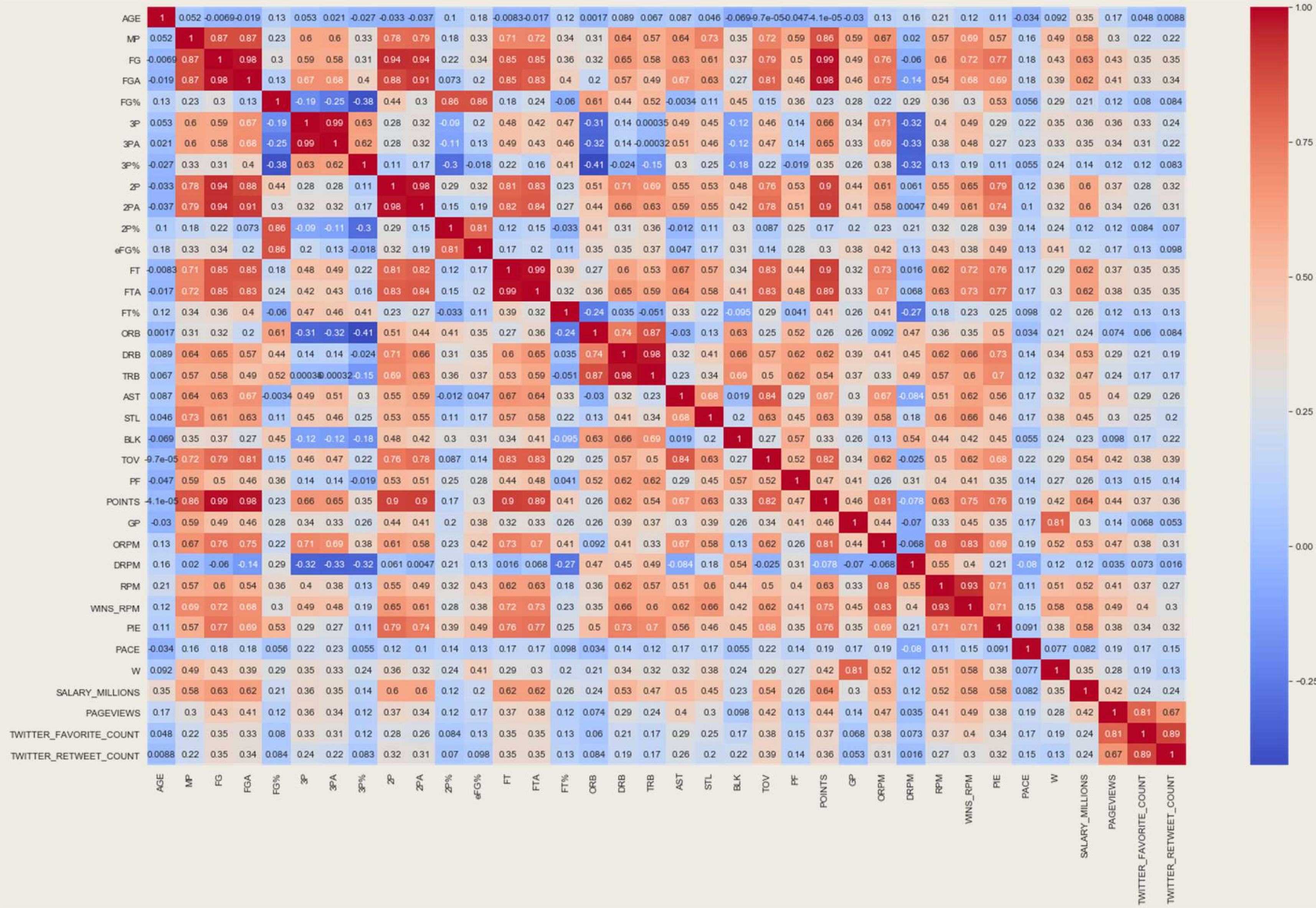
```
players_df['SALARY_MILLIONS'].describe()
```

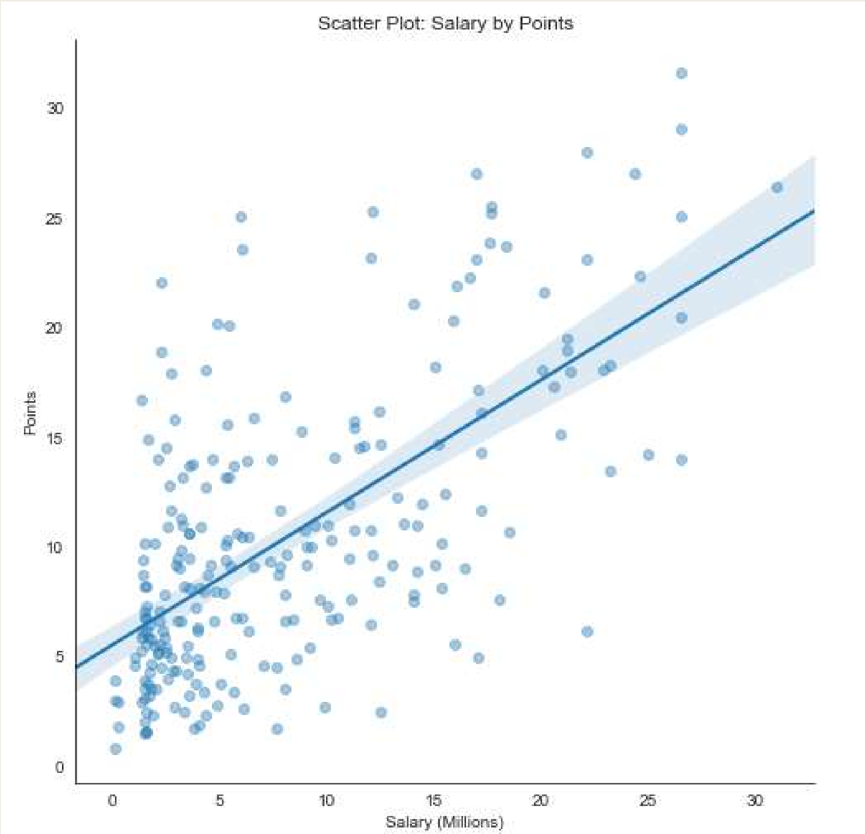
```
count    239.00000
mean       8.09184
std        6.95558
min        0.06000
25%        2.58000
50%        5.37000
75%       12.09500
max       30.96000
Name: SALARY_MILLIONS, dtype: float64
```

Salary Visualizations




Distribution of Salary in the NBA
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
Salary (Millions)


AGE
MP
FG
FGA
FG%
3P
3PA
3P%
2P
2PA
2P%
eFG%
FT
FTA
FT%
ORB
DRB
TRB
AST
STL
BLK
TOV
PF
POINTS
GP
ORPM
DRPM
RPM
WINS_RPM
PIE
PACE
W
SALARY_MILLIONS
PAGEVIEWS
TWITTER_FAVORITE_COUNT
TWITTER_RETWEET_COUNT
1.00
0.75
0.50
0.25
0.00
−0.25



Scatter Plot: Salary by Points
Points
Salary (Millions)
0
5
10
15
20
25
30

Models

- Linear Regression
- Ridge Regression
- Lasso Regression
- Support Vector Regression
- KNN Regressor

First Attempt on DataFrame

Model	Test Data	Training Data
Linear Regression	.900	.901
Ridge Regression	.898	.897
Lasso Regression	.892	.889
Support Vector Regression	.029	.027
KNN Regressor	1	1

Choosing Features

- Running the model on the whole DataFrame is not necessary because of many duplicates
- FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, FT, FTA, FT%, ORB, DRB, TRB
- Select the count of made shots (3P, 2P, and FT), percentages (3P, 2P, and FT), ORB, and DRB

Second Attempt on Selected Features

Model	Test Data	Training Data
Linear Regression	.895	.894
Ridge Regression	.895	.893
Lasso Regression	.891	.889
Support Vector Regression	.029	.027
KNN Regressor	1	1

Third Attempt after PCA

Model	Test Data	Training Data
Linear Regression	.887	.882
Ridge Regression	.887	.882
Lasso Regression	.887	.882
Support Vector Regression	.029	.027
KNN Regressor	1	1

Looking at Linear Regression

	Coefficients	P-Values
AGE	5.01364	1.20088e-11
MP	-1.38886	3.19260e-01
3P%	-32.48583	2.99796e-01
2P%	-56.09916	4.17129e-01
eFG%	84.34400	4.36069e-01
FT%	21.78574	4.12845e-01
ORB	8.59516	2.76485e-01
DRB	9.39552	3.47463e-02
AST	10.97233	3.61890e-02
STL	20.53206	9.59184e-02
BLK	-6.14907	5.81260e-01
TOV	-26.72319	1.99961e-02
PF	-10.06663	1.99729e-01
POINTS	9.75343	1.12024e-04
GP	0.21085	4.78199e-01
ORPM	-4.48459	6.85177e-02
DRPM	4.71470	8.00497e-02
RPM	0.23011	9.29526e-01
WINS_RPM	-3.40960	2.63745e-01
PIE	-5.36153	8.35685e-02
PACE	-1.15407	3.41886e-03
W	0.37486	3.49045e-01
PAGEVIEWS	0.00234	4.85225e-01
TWITTER_FAVORITE_COUNT	-0.00893	2.88996e-01
TWITTER_RETWEET_COUNT	0.02378	2.37218e-01

Looking at Linear Regression (cont.)

- The significant variables in this model are
  - *AGE (player age)*
  - *DRB (defensive rebounds)*
  - *AST (assists)*
  - *TOV (turnovers)*
  - *POINTS (points)*
  - *PACE (pace of play)*

Conclusion

- Regression models were used to predict NBA salary
- First attempt was on the whole DataFrame
- Second attempt was on selected features
- Third attempt went further and utilized PCA
- The scores for linear, lasso, and ridge regressions stayed relatively the same but decreased after each iteration
- Support Vector Regression was not successful at all
- KNN Regressor was always overfit
- There are limitations that arise due to this dataset