Поиск типичного игрока регулярного чемпионата НБА в 2020-2021 сезоне

Прикладные задачи анализа данных, задание №1

Сюй Минчуань, ММП ВМК МГУ

22 сентября 2021 г.

Общее описание датасета

- Собраны индивидуальные статистики игроков НБА в прошлом регулярном чемпионате (2020-2021 Regular Season)
- Показаны только данные игроков, которые имеют квалификацию при выполнении некоторых специфичных условий (присутствие мятчов, величины статистик...) см. `https://basketball.realgm.com/info/glossary`
- **ЗАДАЧА**: Представим, что генеральный директор НБА или какой-то команды хочет найти типичного игрока, чтобы узнавать о среднем уровне способности игроков для последующего анализа и принятия решений.

Предобработка данных и пример данных

In [2]:
```
player = pd.read_csv("active_player.csv")
player.drop(['FGM','FGA','3PM','3PA','FTM','FTA','ORB','DRB'],axis=1, inplace=True)
player.head()
```

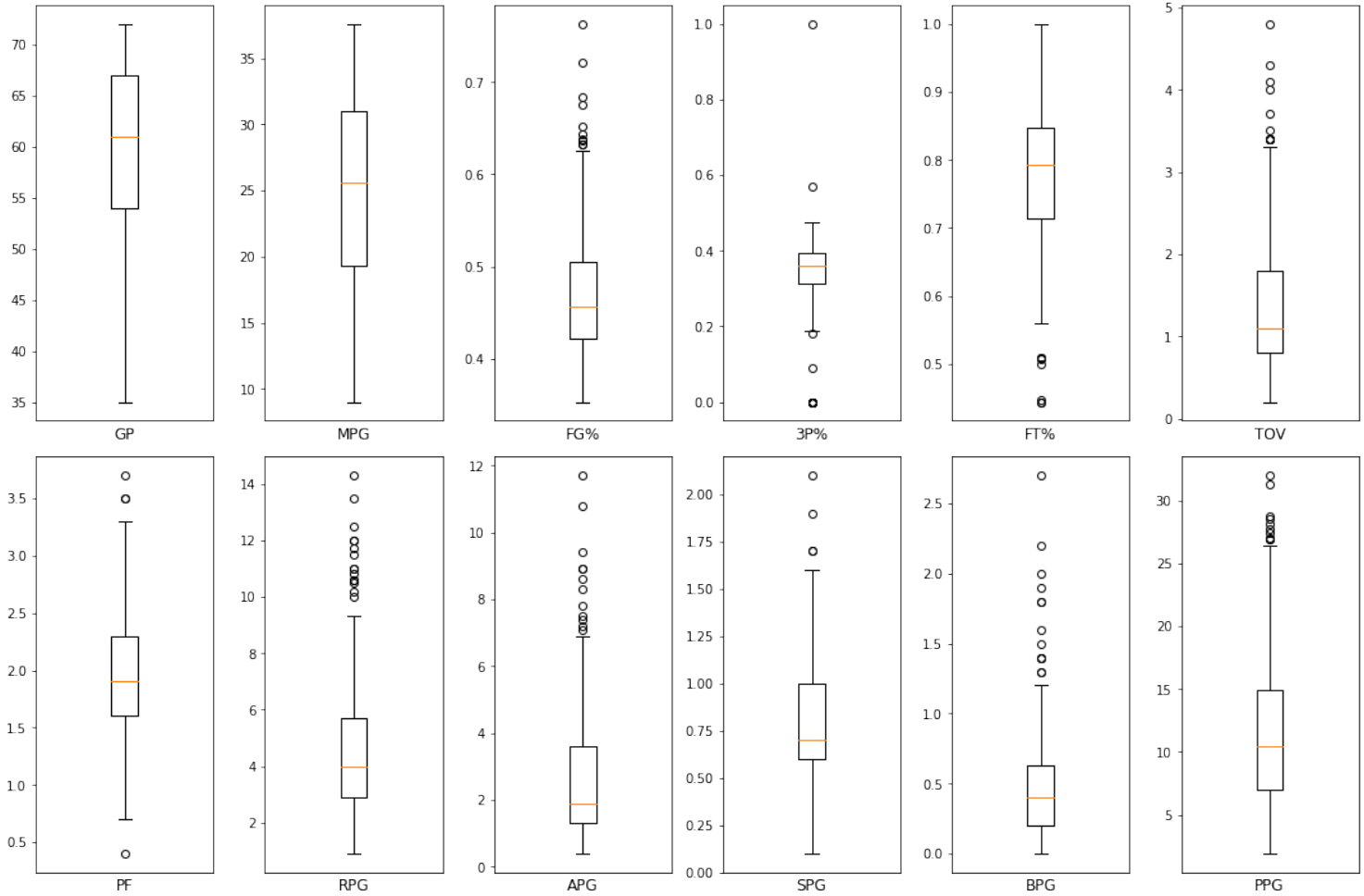
Out[2]:

	Player	Team	GP	MPG	FG%	3P%	FT%	TOV	PF	RPG	APG	SPG	BPG	PPG
0	Stephen Curry	GSW	63	34.2	0.482	0.421	0.916	3.4	1.9	5.5	5.8	1.2	0.1	32.0
1	Bradley Beal	WAS	60	35.8	0.485	0.349	0.889	3.1	2.3	4.7	4.4	1.1	0.4	31.3
2	Damian Lillard	POR	67	35.8	0.451	0.391	0.928	3.0	1.5	4.2	7.5	0.9	0.3	28.8
3	Joel Embiid	PHI	51	31.1	0.513	0.377	0.859	3.1	2.4	10.6	2.8	1.0	1.4	28.5
4	Giannis Antetokounmpo	MIL	61	33.0	0.569	0.303	0.685	3.4	2.8	11.0	5.9	1.2	1.2	28.1

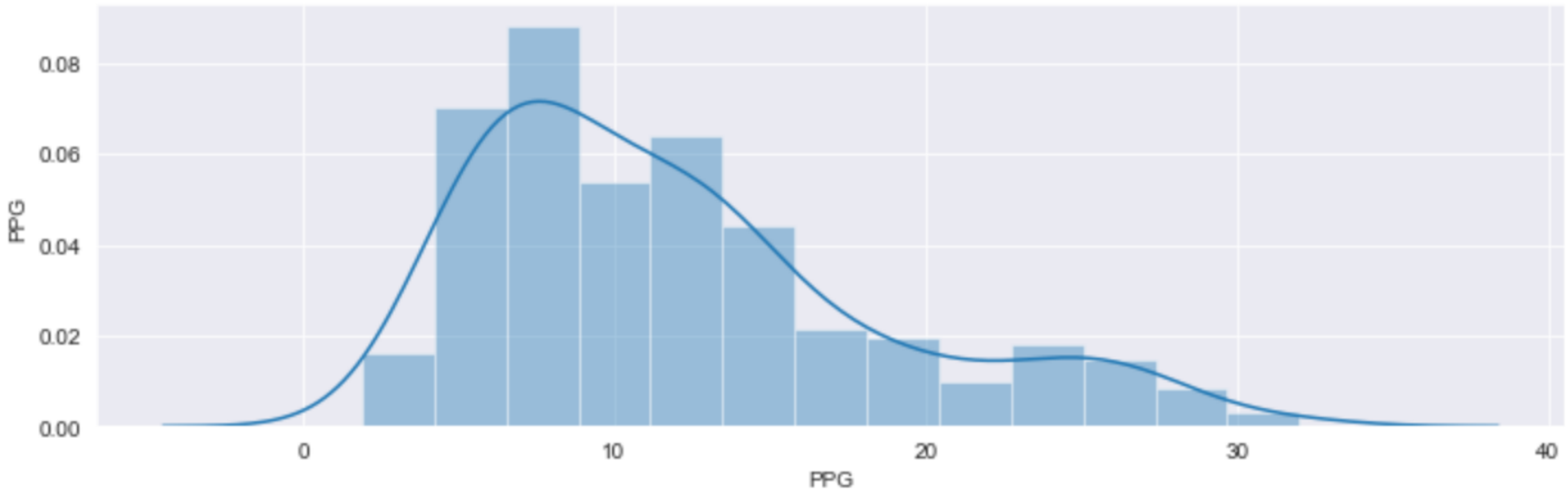
GP - Games Played MPG - Minutes Per Game FG% - Field Goals Percentage 3P% - 3-Points Percentage FT% - Free Throws Percentage TOV - Turnovers PF - Personal Fouls RPG - Rebounds Per Game APG - Assists Per Game

SPG - Steals Per Game BPG - Blocks Per Game PPG - Points Per Game

Диаграммы размаха относительно статистик

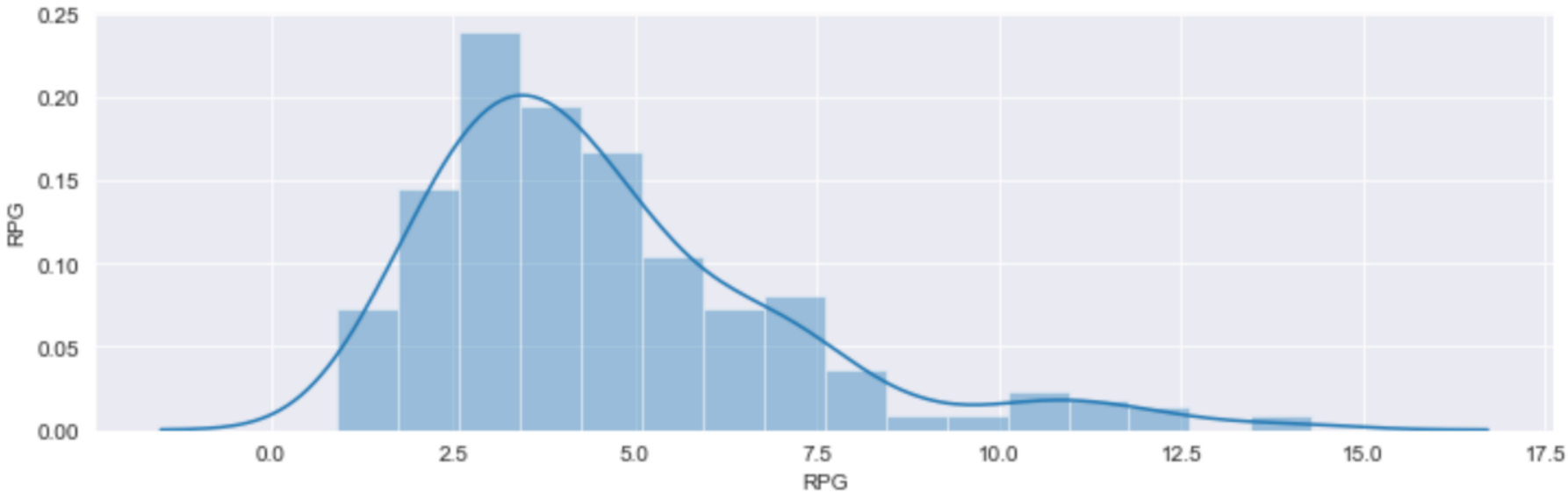


Гистограммы и распределения



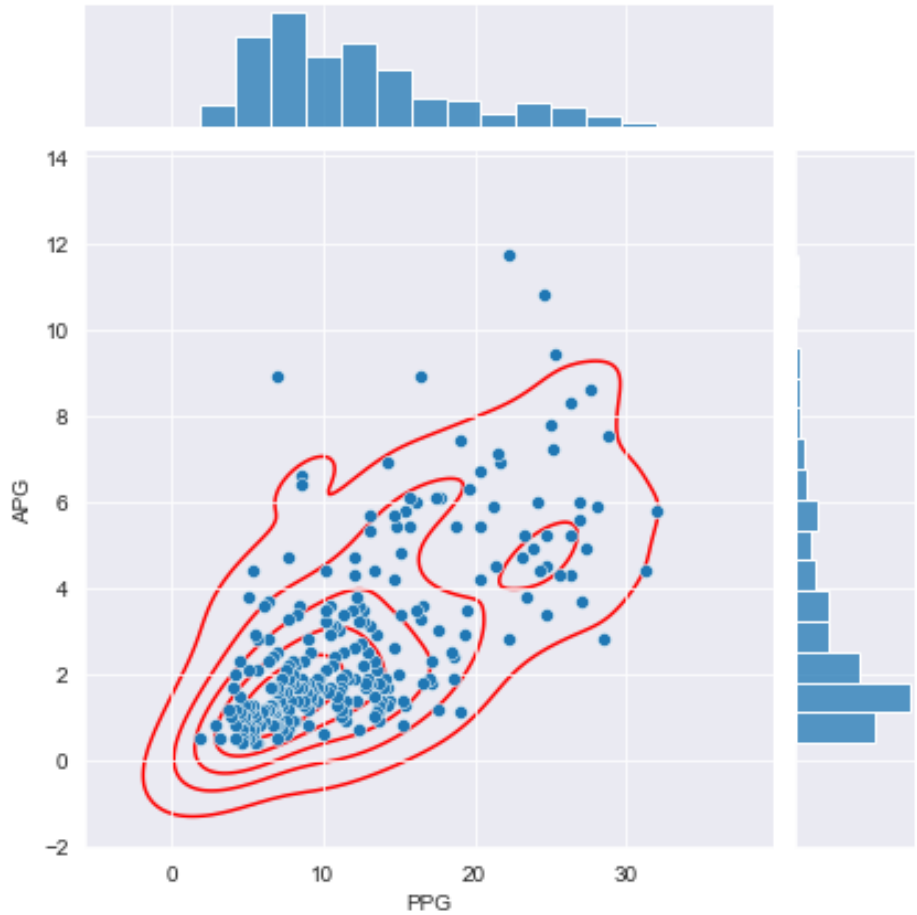
по "PPG"

Гистограммы и распределения

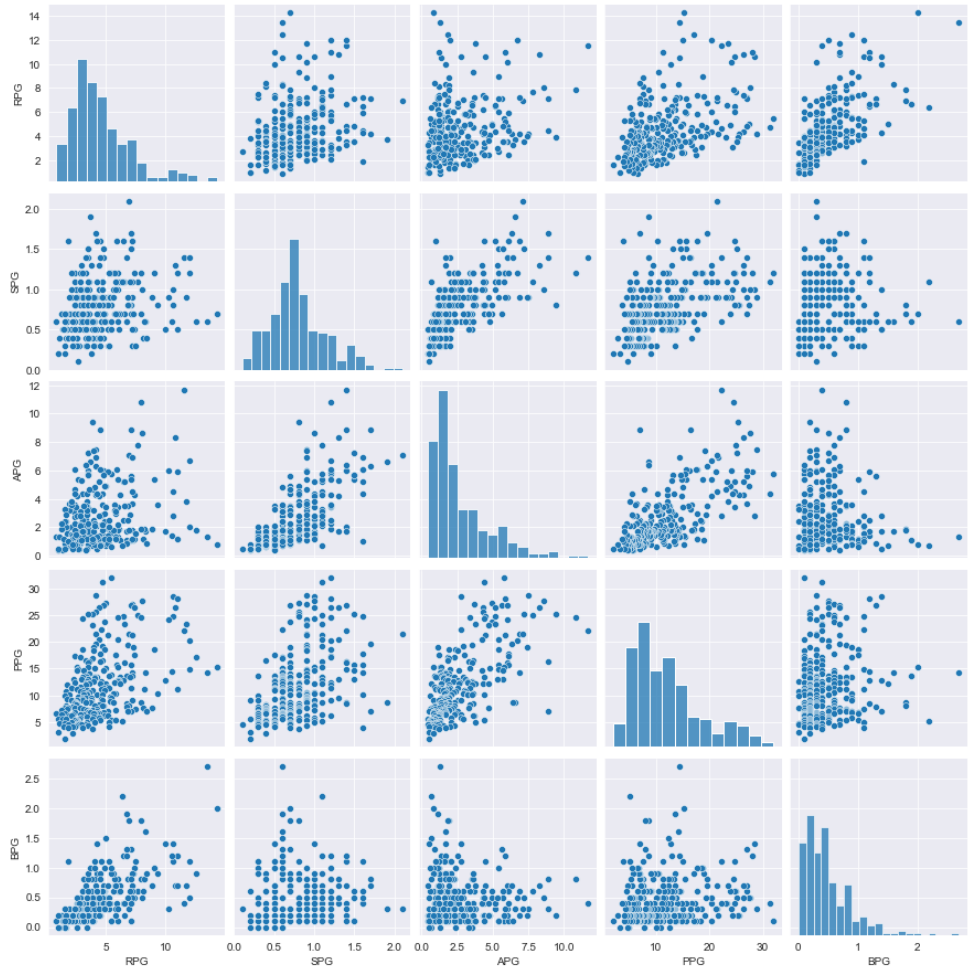


по "RPG"

график отношения jointplot



pairplot по пяти статистикам



Среднее как решение оптимизационной задачи

$$L(a) = \sum_{i=1}^{m} (x_i - a)^2 \to \min_a$$

Дифференцируя, приравнивая производную функции $L(a)$ к нулю, получаем

$$\frac{dL}{da} = \sum_{i=1}^{m} 2(x_i - a) = 0 \quad \Rightarrow \quad a = \frac{\sum_{i=1}^{m} x_i}{m}$$

In [7]:
```
mean_data = player_data.mean()
np.round(mean_data, 2)
```

Оценка минимального контраста

Рассматривается задача нахождения медианы в общем случае:

$$mid(X) = argmin \sum_{i=1}^{m} f(x_i, a)$$

Мешалкин Л.Д. в 1977-ом году предлагал

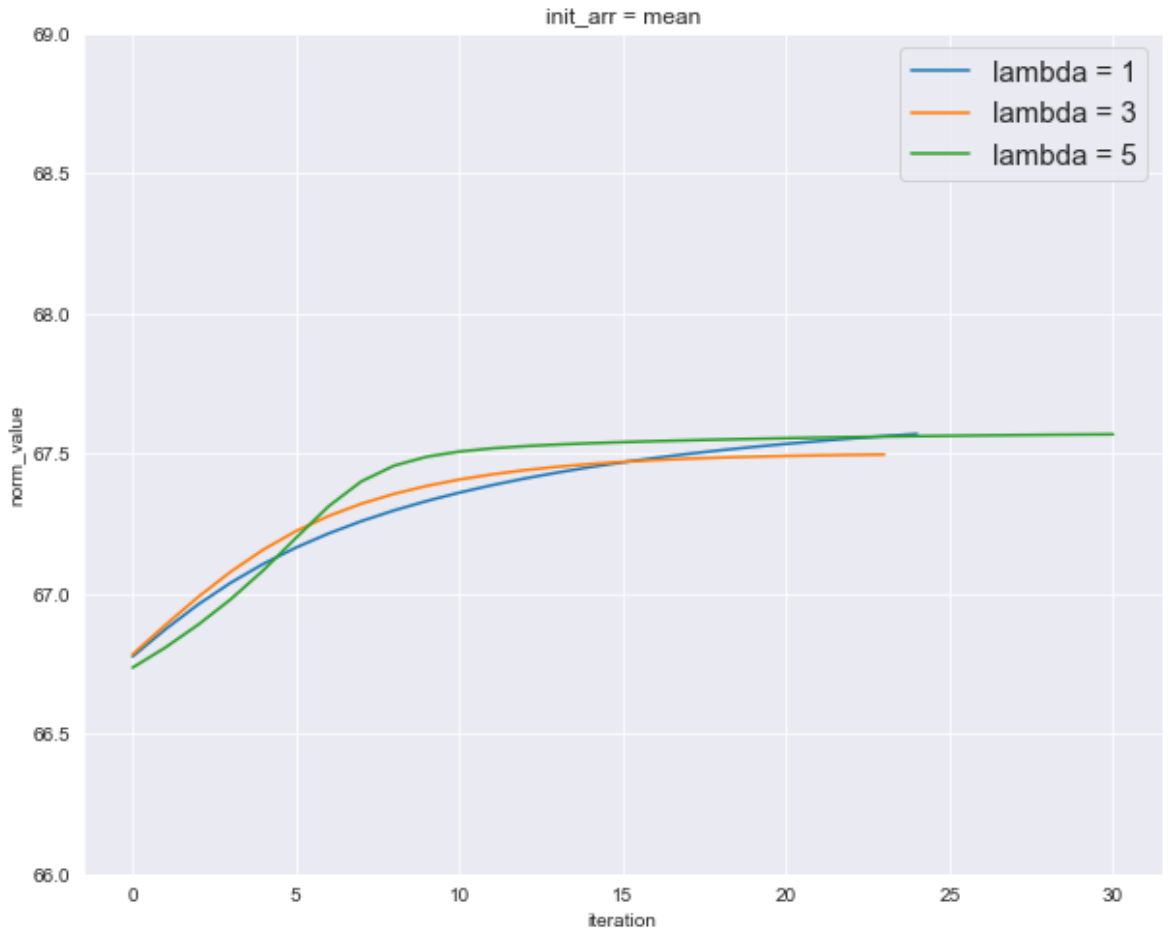
$$f(x, a) = -\frac{1}{\lambda} exp\left(-\frac{\lambda(x-a)^2}{2}\right)$$

В этом случае: $\psi(z) = zexp\left(\frac{-\lambda z^2}{2}\right), \xi(z) = exp\left(\frac{-\lambda z^2}{2}\right)$.

Формула для итеративного расчета:

$$a = \frac{\sum_{i=1}^{m} x_i \xi(x_i - a)}{\sum_{i=1}^{m} \xi(x_i - a)}$$

Берем разные λ



Другое способ оценивания...

1. Находим экстремумы по основным атрибутам (атриб. от 'GP' до 'PPG'),
   - По атрибутам 'TOV' и 'PF': min, max
   - По всем остальным атрибутам: max
2. Припишем для каждого игрока *скоринг* в зависимости от экстремумов.
   - ex. max по 'PPG' = 32.0, то скоринг игрока с 'PPG=20.0'по 'PPG' будет 20.0/32.0=0.625.
   - ex. min по 'TOV' = 0.2, максимум = 4.8, то скоринг игрока с 'TOV=3' по 'TOV' будет 1-(3-0.2)/(4.8-0.2) = 0.39
3. Просуммируем для каждого игрока скорингы всех статистик, получаем *итоговую оценку* игрока, и просортируем по итоговой оценке.
4. Находим медиану итоговых оценок, и тот игрок, у которого оценка ближайжая к медиане, будет типичный игрок.

Скоринги, итоговые оценки, и типичный игрок

Out[13]:

	Player	Team	GP	MPG	FG%	3P%	FT%	TOV	PF	RPG	APG	SPG	BPG	PPG	Total
0	Stephen Curry	GSW	0.875000	0.909574	0.631717	0.421	0.916	0.304348	0.545455	0.384615	0.495726	0.571429	0.037037	1.000000	7.091901
1	Bradley Beal	WAS	0.833333	0.952128	0.635649	0.349	0.889	0.369565	0.424242	0.328671	0.376068	0.523810	0.148148	0.978125	6.807740
2	Damian Lillard	POR	0.930556	0.952128	0.591088	0.391	0.928	0.391304	0.666667	0.293706	0.641026	0.428571	0.111111	0.900000	7.225157
3	Joel Embiid	PHI	0.708333	0.827128	0.672346	0.377	0.859	0.369565	0.393939	0.741259	0.239316	0.476190	0.518519	0.890625	7.073221
4	Giannis Antetokounmpo	MIL	0.847222	0.877660	0.745740	0.303	0.685	0.304348	0.272727	0.769231	0.504274	0.571429	0.444444	0.878125	7.203200

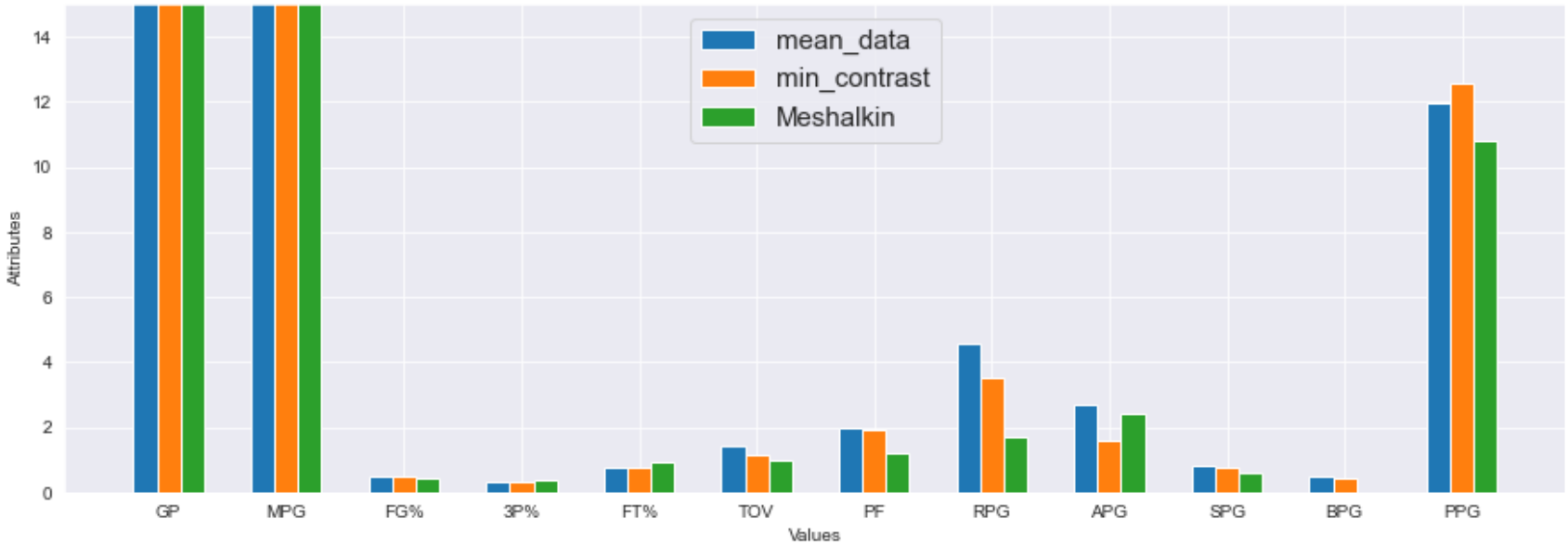
Out[15]:

	Player	Team	GP	MPG	FG%	3P%	FT%	TOV	PF	RPG	APG	SPG	BPG	PPG
128	Patty Mills	SAS	68	24.8	0.412	0.375	0.91	1.0	1.2	1.7	2.4	0.6	0.0	10.8

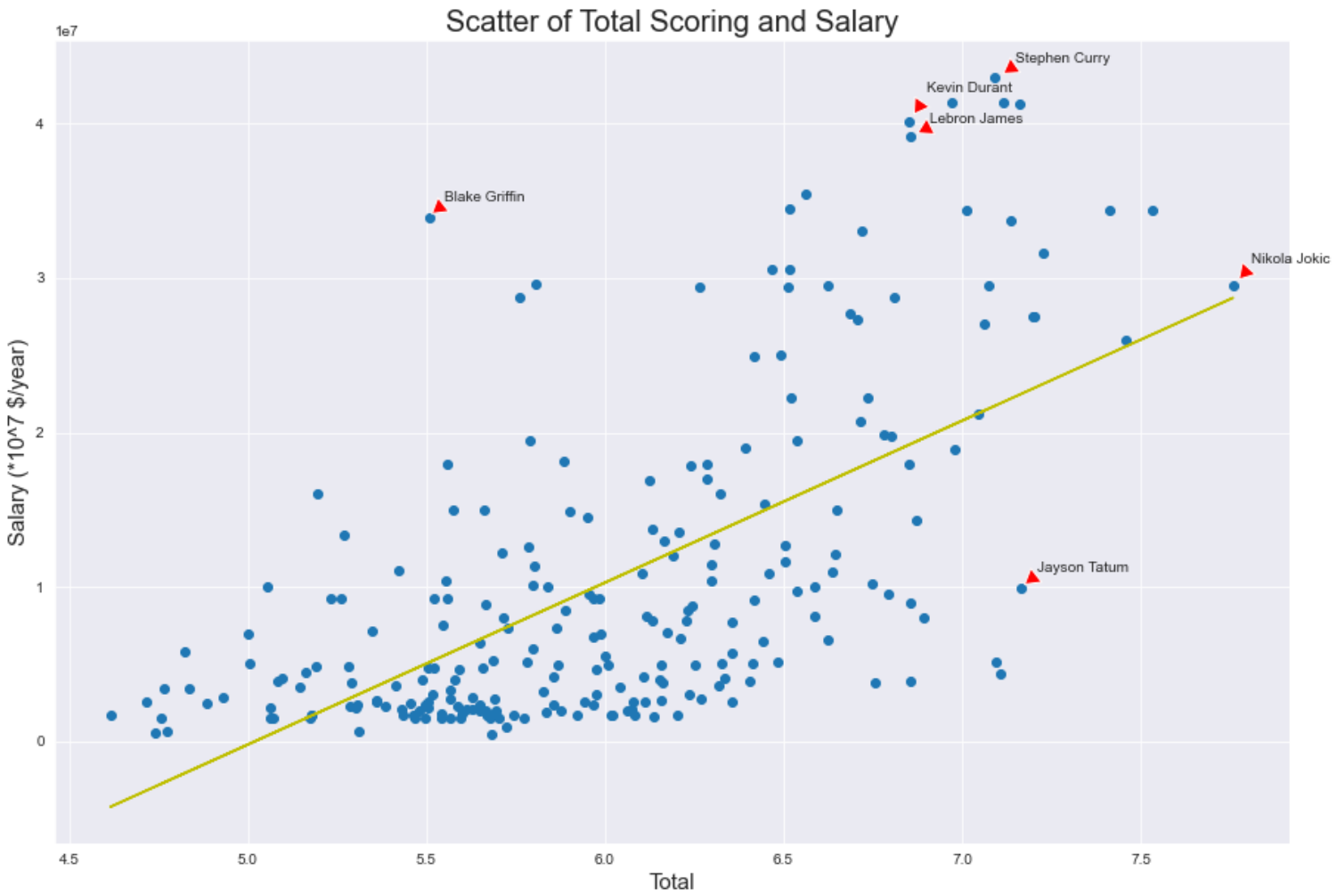
Изобразить типичных игроков, полученных ранее

	GP	MPG	FG%	3P%	FT%	TOV
mean_data	60.28	25.14	0.47	0.34	0.78	1.41
min_contrast	61.23	25.22	0.47	0.34	0.78	1.17
Meshalkin	68.00	24.80	0.41	0.38	0.91	1.00

	PF	RPG	APG	SPG	BPG	PPG
mean_data	1.97	4.58	2.72	0.80	0.49	11.99
min_contrast	1.95	3.50	1.61	0.78	0.44	12.58
Meshalkin	1.20	1.70	2.40	0.60	0.00	10.80



Сравним с настоящими зарплатами игроков...



Итоги и дальше

1. Что были?
   - датасет, предобработка, визуализация
   - три подхода получения среднего
   - оценки и зарплаты
2. Что дальше?
   - можно использовать весовую схему, приписав для каждых атрибутов веса, затем сопоставить зарплаты игроки и итоговые оценки, используя какую-то регрессионную модель (улучшить корреляцию этих двух величины)
   - ex. подробный анализ подбора λ и начального приближения, и прочее.

СПАСИБО!