

---

# 해외 축구선수 이적료 데이터 분석

---

파뿌리(파이썬 부시는 이십대들) 3기

2021 - 11 - 24

차승우

# 목차





























---

1. 데이터셋
2. 데이터 전처리
3. EDA(탐색적 자료분석)
4. 머신러닝 기법
5. 결론

# 축구선수 이적료

- 선수의 가치는 어떻게 측정될까?



tm NEWS TRANSFERS & RUMOURS MARKET VALUES COMPETITIONS FORUMS MY TM LIVE						
OVERVIEW TABLES TRANSFERS MARKET VALUES PLAYERS CLUBS INFORMATION & FACTS HIST						
6	 <b>Raheem Sterling</b> Left Winger	 	26		€90.00m	↓
7	 <b>Jadon Sancho</b> Left Winger	 	21		€90.00m	↓
8	 <b>Bruno Fernandes</b> Attacking Midfield		27		€90.00m	■
9	 <b>Marcus Rashford</b> Left Winger	 	24		€85.00m	■
10	 <b>Sadio Mané</b> Left Winger	 	29		€85.00m	↓
11	 <b>Heung-min Son</b> Left Winger		29		€85.00m	■
12	 <b>Phil Foden</b> Central Midfield		21		€80.00m	↑
13	 <b>Trent Alexander-Arnold</b> Right-Back		23		€75.00m	↓

- 계약 잔여기간, 나이, 기량 순으로 다양한 변수들을 고려하여 측정
- 기량은 해당 선수가 가진 (역량, 리더십, 부상우려 등) 다양한 측면을 통계적으로 집계하여 각각의 이적료를 산출하는데 반영함
- 가장 대표적인 축구 이적정보 사이트인 트랜스퍼 마켓에서 선수들의 객관적인 평가 이적료를 파악할 수 있음

---

# 1. 데이터셋

---

# 데이터셋

	id	name	age	continent	contract_until	position	prefer_foot	reputation	stat_overall	stat_potential	stat_skill_moves	value
0	0	L. Messi	31	south america	2021	ST	left	5.0	94	94	4.0	110500000.0
1	3	De Gea	27	europa	2020	GK	right	4.0	91	93	1.0	72000000.0
2	7	L. Suárez	31	south america	2021	ST	right	5.0	91	91	3.0	80000000.0
3	8	Sergio Ramos	32	europa	2020	DF	right	4.0	91	91	3.0	51000000.0
4	9	J. Oblak	25	europa	2021	GK	right	3.0	90	93	1.0	68000000.0

- Contract\_until : 선수의 계약기간 만료 년도 → 조금 남을 수록 이적료가 싸지지 않을까 예상
- Position : 포지션 → 이적료 최상위권은 보통 공격수 비율이 많지 않을까 예상
- Prefer\_foot : 선수의 선호 발 방향 → 같은 조건의 선수라면, 분포가 적은 왼발을 선호하는 선수가 더 가치 있을 것이라 예상
- Reputation : 선수의 평판(인기정도) → 이적료에 영향을 줄 것이라 예상
- Stat\_potential : 선수가 발전할 수 있는 정도 → 현재 능력치는 낮지만, 발전할 수 있는 정도가 높은 선수가 더 비쌀 것이라 예상
- Stat\_skill\_moves : 선수의 개인기 능력치 → 크게 영향을 줄 수도 있지만, 포지션별로 다를 것이라 예상 ex) 골키퍼의 개인기?
- Value : FIFA 선정 선수의 이적 시장 가격(유로) → Y값으로 설정

# 데이터셋

- 데이터 내용 확인하기

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8932 entries, 0 to 8931
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   id                    8932 non-null   int64
 1   name                  8932 non-null   object
 2   age                   8932 non-null   int64
 3   continent             8932 non-null   object
 4   contract_until        8932 non-null   object
 5   position              8932 non-null   object
 6   prefer_foot           8932 non-null   object
 7   reputation            8932 non-null   float64
 8   stat_overall          8932 non-null   int64
 9   stat_potential        8932 non-null   int64
10   stat_skill_moves      8932 non-null   float64
11   value                 8932 non-null   float64
dtypes: float64(3), int64(4), object(5)
memory usage: 837.5+ KB
```

〈데이터 요약정보〉

```
id          0
name        0
age         0
continent   0
contract_until 0
position    0
prefer_foot 0
reputation  0
stat_overall 0
stat_potential 0
stat_skill_moves 0
value       0
dtype: int64
```

〈결측치 존재유무 확인〉

---

## 2. 데이터 전처리

---

# 데이터 전처리

	age	continent	contract_until	position	prefer_foot	reputation	stat_overall	stat_potential	stat_skill_moves	value
0	31	south america	2021	ST	left	5.0	94	94	4.0	110500000.0
1	27	europa	2020	GK	right	4.0	91	93	1.0	72000000.0
2	31	south america	2021	ST	right	5.0	91	91	3.0	80000000.0
3	32	europa	2020	DF	right	4.0	91	91	3.0	51000000.0
4	25	europa	2021	GK	right	3.0	90	93	1.0	68000000.0

- ‘id’, ‘name’ 열은 이적료 예측에 있어서 불필요하다고 판단하여 삭제

	age	contract_until	reputation	stat_overall	stat_potential	stat_skill_moves	value	continent_africa	continent_asia	continent_europe	continent_oceania	continent_south america	position_DF	position_GK	position_MF	position_ST	prefer_foot_left	prefer_foot_right
0	31	2021	5.0	94	94	4.0	110500000.0	0	0	0	0	1	0	0	0	1	1	0
1	27	2020	4.0	91	93	1.0	72000000.0	0	0	1	0	0	0	1	0	0	0	1
2	31	2021	5.0	91	91	3.0	80000000.0	0	0	0	0	1	0	0	0	1	0	1
3	32	2020	4.0	91	91	3.0	51000000.0	0	0	1	0	0	1	0	0	0	0	1
4	25	2021	3.0	90	93	1.0	68000000.0	0	0	1	0	0	0	1	0	0	0	1

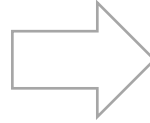
- 범주형 변수를 분석에 사용하기 위해 더미화 진행
- 더미변수 : 0,1로 표현되는 이진변수 (yes or no)



# 데이터 전처리

```
tr_data['contract_until'].unique()
```

```
array(['2021', '2020', '2019', '2023', '2022', '2024', 'Jun 30, 2019',  
      '2026', 'Dec 31, 2018', '2018', '2025', 'Jun 30, 2020',  
      'May 31, 2020', 'May 31, 2019', 'Jan 31, 2019', 'Jan 1, 2019',  
      'Jan 12, 2019'], dtype=object)
```



```
tr_data['contract_until'].unique()
```

```
array([2021.    , 2020.    , 2019.    , 2023.    , 2022.    , 2024.    ,  
      2019.5   , 2026.    , 2018.    , 2025.    , 2020.5   , 2020.3333,  
      2019.3333, 2019.0833, 2019.034  ])
```

- contract\_until 데이터 중, 년도를 표기하는 데에 있어 숫자가 아닌 (문자 + 숫자) 행이 존재 (통일 필요) → 전부 float으로 변환

	age	contract_until
0	31	2021.0
1	27	2020.0
2	31	2021.0
3	32	2020.0
4	25	2021.0



	age	contract_until
0	31	3.0
1	27	2.0
2	31	3.0
3	32	2.0
4	25	3.0

- contract\_until의 의미를 직관적으로 활용하기 위해, (계약만료 년도 - 현재 년도(2018년))을 계산하여 전처리함

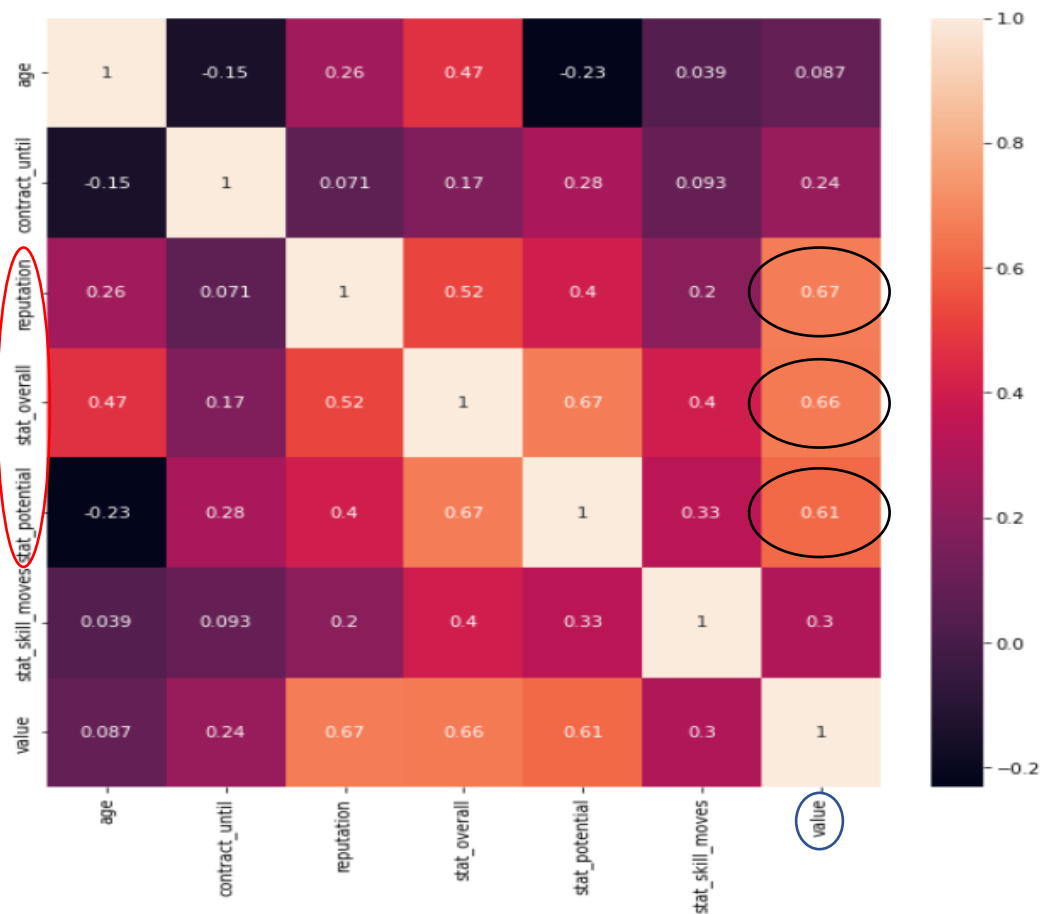
---

## 3. EDA(탐색적 자료분석)

---

## 상관관계 분석

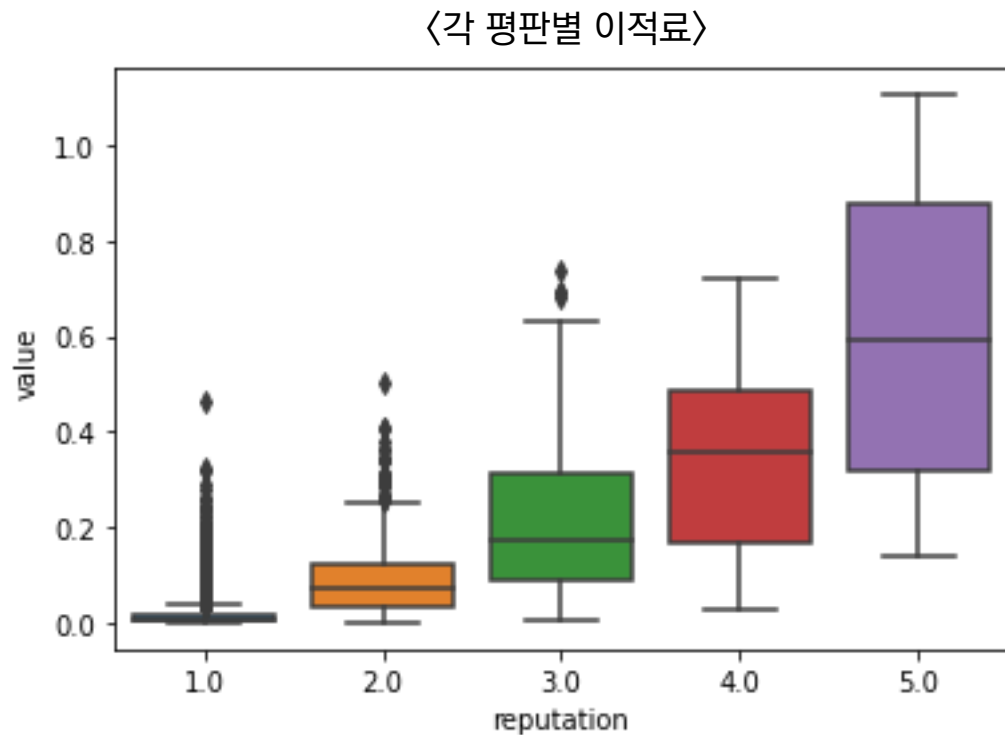
〈변수간 상관관계〉



- Value와 관계가 높은 변수는 reputation, stat\_overall, stat\_potential
- 이 변수 3가지로도 이적료 값을 어느정도 예측할 수 있을 것
- 단순히 나이가 어리다는 것보다는, 스탯 잠재력이 얼마나 큰지가 더 중요
- 배경지식에서 언급했던, 계약 잔여기간 · 나이 · 기량 순으로 이적료가 책정 되는 것이 아니라, 평판 · 기량 · 잠재력 순으로 이적료가 매겨짐

# EDA

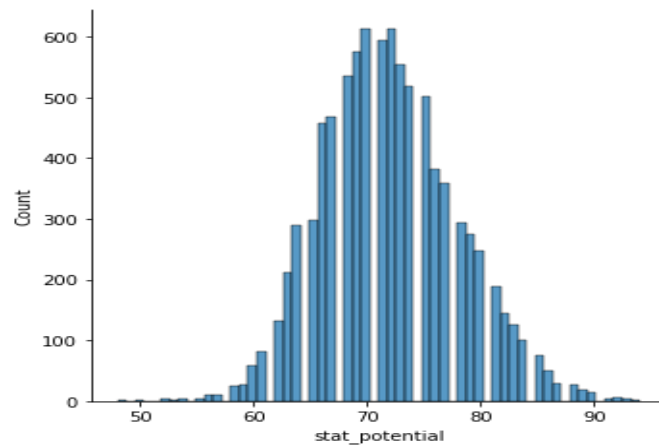
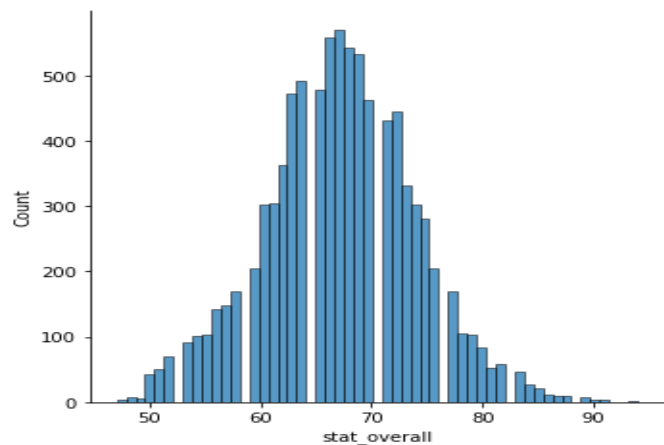
## ▪ Boxplot (reputation)



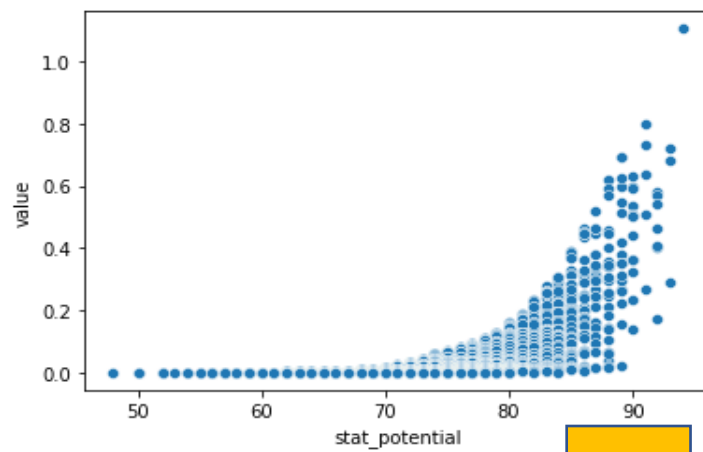
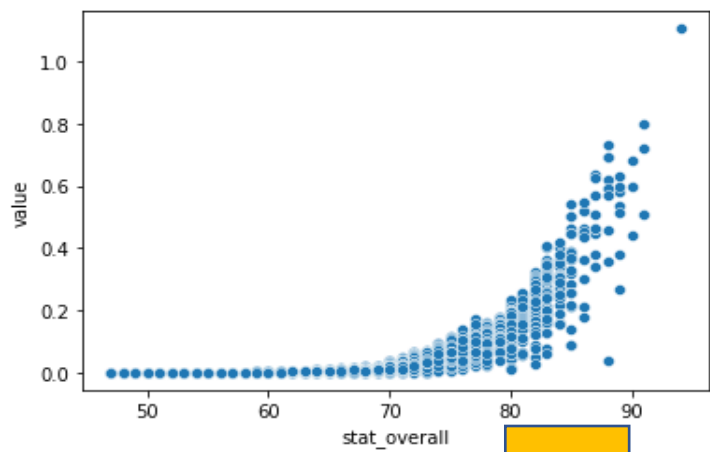
- 평판이 높을 수록, 평균적으로 이적료가 높다는 것을 Boxplot을 통해 시각화
- 평판이 높지 않아도, 다른 요소들로 인해 이적료가 높은 경우도 다수 존재
- 이상값 처리 : 자연발생이라고 판단 → 향후 모델을 만들었을 때, 현상/예측을 잘 설명할 수 없을 수도 있을 가능성 존재

# EDA

- Histogram, Scatterplot (stat\_overall, stat\_potential)



→ 분포도



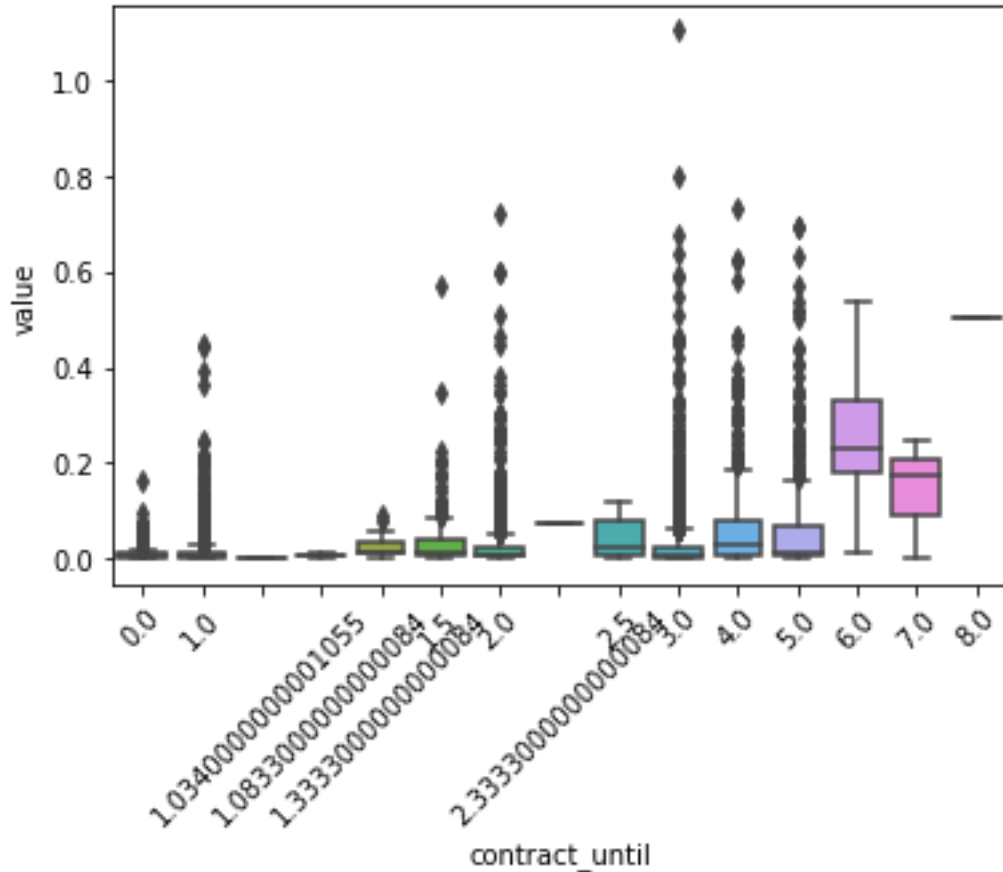
→ 산점도 (value와의 관계)

\*잠재 역량 보다는 실제로 역량을 발휘하는 것이 몸값  
인상에 더 효과적인 것으로 판단됨

# EDA

- 만료 계약기간이 조금 남을 수록 이적료가 싸질 것 → △

〈남은 계약기간에 따른 이적료〉

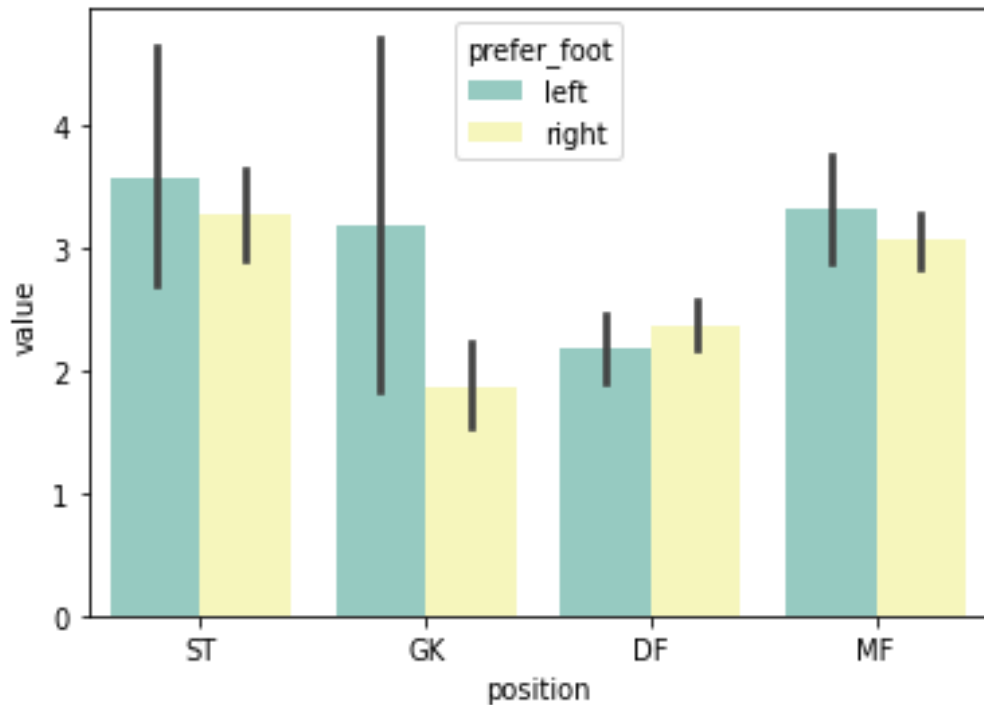


- 이적료와 계약기간의 상관관계가 0.24였던 것처럼, boxplot으로 비교를 해 보아도 가정을 뒷받침할 근거가 명확하지 않아 보임
- 하지만, 계약기간이 0년 남아, 곧 FA가 될 선수들의 이상치들은 다른 기간에 비해 현저히 낮은 것을 나타냄

# EDA

- 이적료 상위권은 공격수 비율이 많을 것 → ○

〈포지션/선호 발 위치 별 이적료〉

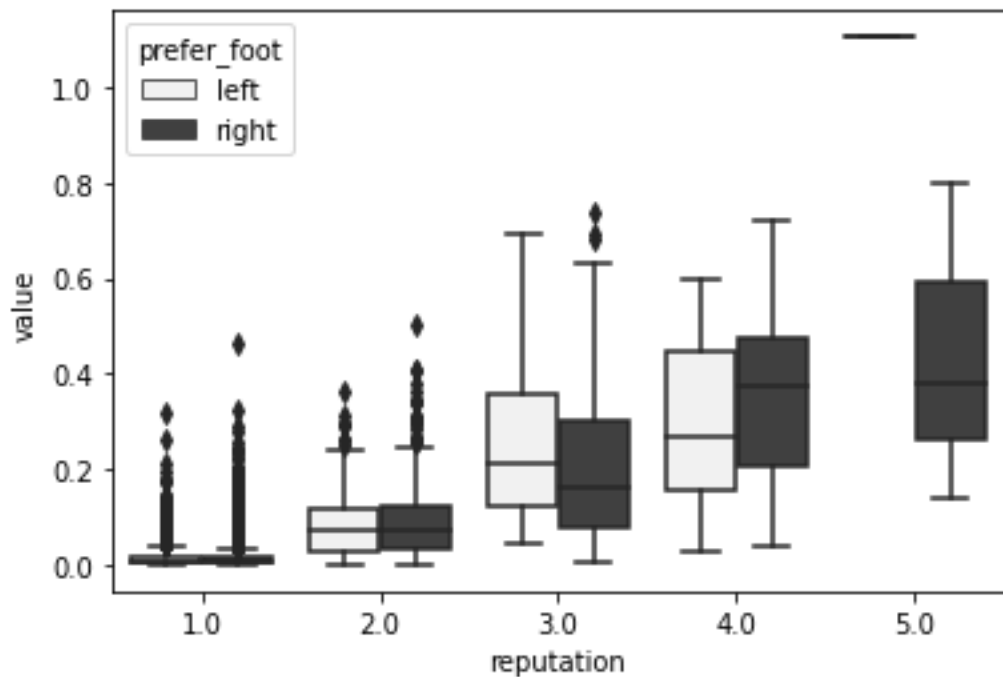


- 수비수 - 골키퍼 - 미드필더 - 공격수 순으로 평균 이적료가 높아지고 있는 것을 보임
- 수비 수를 제외한 나머지 포지션에서는 왼발을 사용하는 선수의 평균 이적료가 더 높은 것을 나타냄 → but, 빈도가 적은 왼발에 고 이적료 선수가 몰렸을 가능성이 존재함

# EDA

- 같은 조건의 선수라면, 분포가 적은 왼발을 선호하는 선수가 더 가치 있을 것이라 예상 → X

〈인기도/선호 발 위치 별 이적료〉



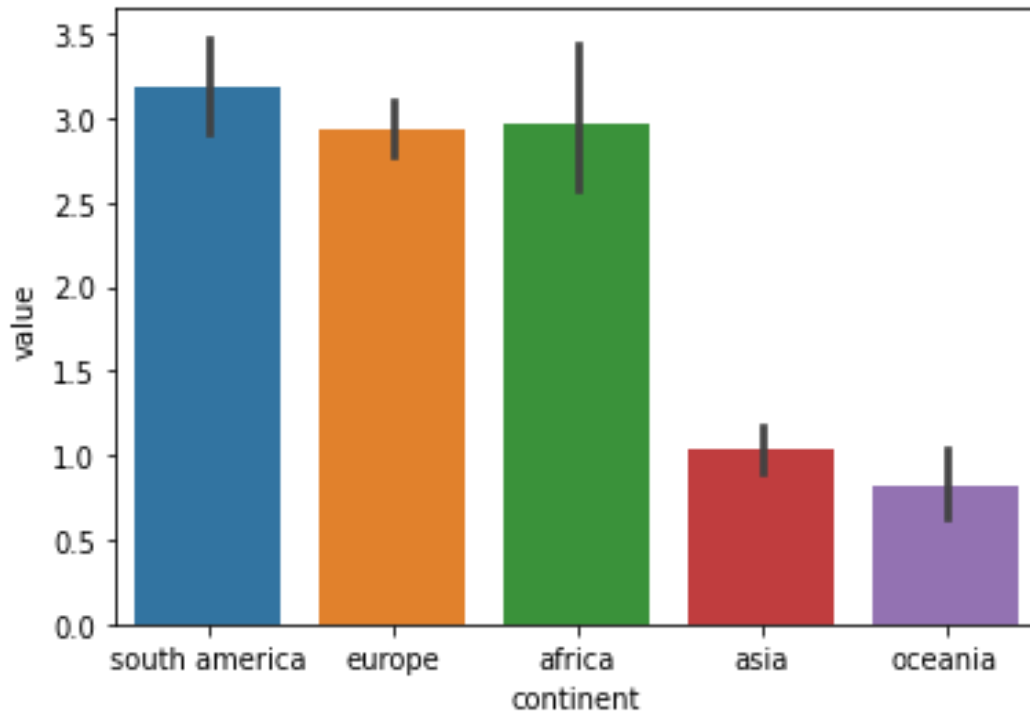
- 그저 왼발과 오른발을 사용하는 선수들의 평균 이적료 차이를 비교하기엔 빈도수 차이가 크기 때문에, 이적료에 가장 크게 기여하는 평판 변수가 같을 때의 선호 발 위치 따른 이적료 차이를 비교함
- 가장 분포가 많은 평판 1.0을 보면, 왼발 오른발에 따른 이적료 차이는 거의 없다고 판단됨



# EDA

## ▪ 대륙별 평균 이적료

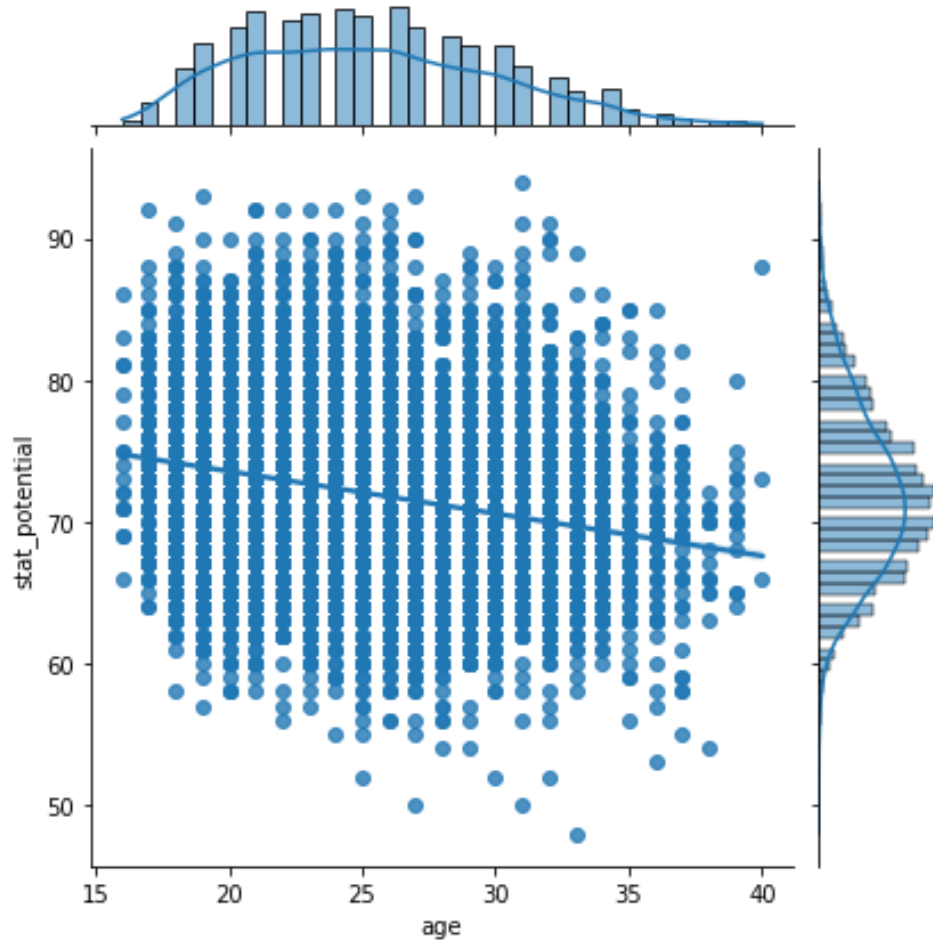
〈대륙에 따른 이적료〉



- 평균 이적료는 남아메리카 - 아프리카 - 유럽순으로 높지만, 유럽 선수의 빈도 수가 압도적으로 많기 때문에 유럽 선수의 이적료는 대부분 높음
- 아프리카와 아시아 선수의 빈도 수는 비슷하지만, 아프리카 선수의 평균 이적료가 압도적으로 높은 것을 볼 수 있음

# EDA

## ▪ Age와 stat\_potential과의 상관관계 시각화



- 나이가 어릴 수록 기존 스탯 대비 잠재 스탯이 어느정도 더 높아질 가능성이 있다는 것으로 판단됨 (음의 상관관계)
- 나이와 잠재 스탯의 상관관계 지수는  $-0.23$

---

## 4. 머신러닝 모델

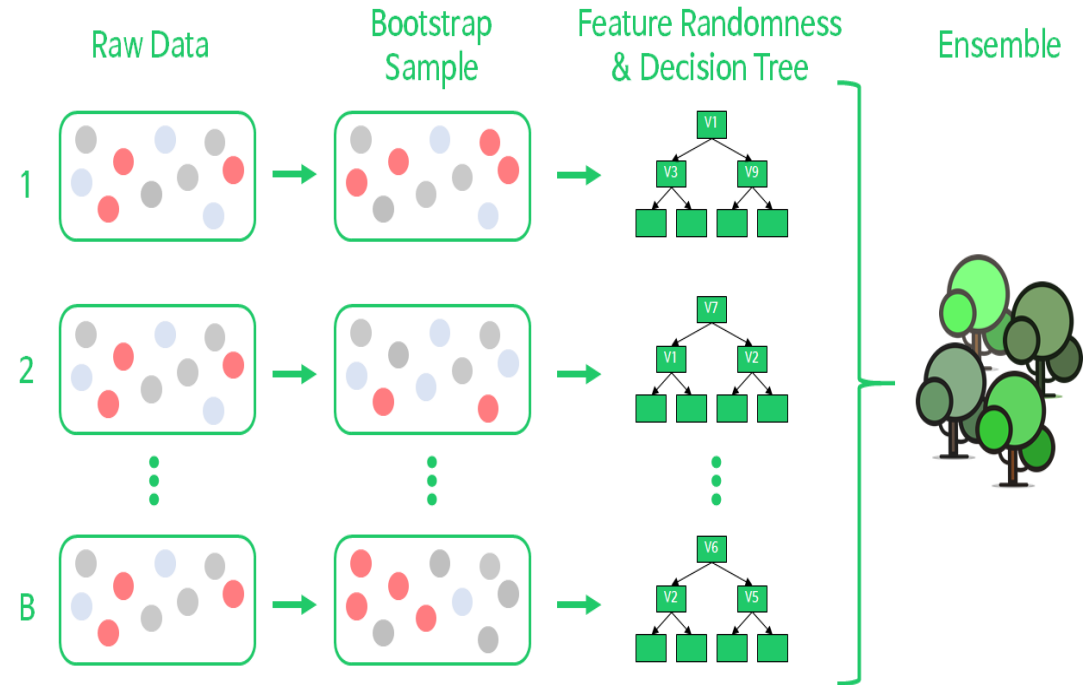
---

# 예측 머신러닝 모델

- 랜덤 포레스트 모델 (Random Forest)

- 여러가지 x변수들(기량, 평판, 나이 등)을 통해 해당 선수의 이적료를 예측하는 Task

- 랜덤 포레스트 모델은 앙상블 기법 중 하나
- 앙상블 기법 : 여러 모델들의 예측을 평균 or 다수결하여 가장 좋은 예측 정확성을 향상시키는 방법
- 각 트리들의 예측 정확도를 종합하여 최대화 하는 방식
- 정확도가 매우 높아, 현업에서 많이 사용되는 것으로 알려짐



# 랜덤 포레스트 모델

```
features = ['reputation', 'stat_overall', 'stat_potential']  
tr_xxdata = tr_data[features]  
tr_yydata = tr_data['value']  
ts_xxdata = ts_data[features]
```

```
from sklearn.ensemble import RandomForestRegressor  
  
model = RandomForestRegressor(n_estimators=200, max_depth=5, random_state=0)  
model.fit(X_train, y_train)
```

```
y_predict = model.predict(ts_xxdata)
```

```
ts_data['value'] = y_pred
```

- 모델을 학습 시키는 데에 있어, 앞서 언급했던 이적료 값에 가장 영향을 미치는 3개 변수 선택
- 트리의 깊이를 5로 제한하고, 200개의 샘플링을 통해 학습 진행
- 학습된 모델에 제공받은 test data의 종속변수 3개 대입
- Test data에 예측 이적료 값 column 추가

# 랜덤 포레스트 모델

- 제공받은 Test data

id	name	age	continent	contract_until	position	prefer_foot	reputation	stat_overall	stat_potential	stat_skill_moves	
1	Cristiano Ronaldo	33	europe	2022	ST	right	5	94	94	5	
2	Neymar Jr	26	south america	2022	ST	right	5	92	93	5	
4	K. De Bruyne	27	europe	2023	MF	right	4	91	92	4	
5	E. Hazard	27	europe	2020	ST	right	4	91	91	4	

예측

value
94127748.83732148
71501479.43154414
61274391.76521444
61506981.050928734

# 랜덤 포레스트 모델

파쑈리 컨퍼런스.jpynb

X

해외축구선수이적료(파쑈리)

X

test\_real.csv

X

Delimiter:

	id	name	age	contract_until	reputation	stat_overall	stat_potential	stat_skill_moves	continent_africa	continent_asia	continent_europe	continent_oceania	continent_south_amer...	position_DF	position_GK	position_MF	position_ST	prefer_foot_left	prefer_foot_right	value
1	1	Cristiano Ronaldo	33	4.0	5.0	94	94	5.0	0	0	1	0	0	0	0	0	1	0	1	94127748.83732148
2	2	Neymar Jr	26	4.0	5.0	92	93	5.0	0	0	0	0	1	0	0	0	1	0	1	71501479.43154414
3	4	K. De Bruyne	27	5.0	4.0	91	92	4.0	0	0	1	0	0	0	0	1	0	0	1	61274391.76521444
4	5	E. Hazard	27	2.0	4.0	91	91	4.0	0	0	1	0	0	0	0	0	1	0	1	61506981.050928734
5	6	L. Modrić	32	2.0	4.0	91	91	4.0	0	0	1	0	0	0	0	1	0	0	1	61506981.050928734
6	10	R. Lewandowski	29	3.0	4.0	90	90	4.0	0	0	1	0	0	0	0	0	1	0	1	54930646.61369193
7	11	T. Kroos	28	4.0	4.0	90	90	3.0	0	0	1	0	0	0	0	1	0	0	1	54930646.61369193
8	15	P. Dybala	24	4.0	3.0	89	94	4.0	0	0	0	0	1	0	0	0	1	1	0	71351246.18154357
9	17	A. Griezmann	27	5.0	4.0	89	90	4.0	0	0	1	0	0	0	0	1	0	1	0	54712798.12884345
10	23	S. Agüero	30	3.0	4.0	89	89	4.0	0	0	0	0	1	0	0	0	1	0	1	51116859.16380524
11	25	K. Mbappé	19	4.0	3.0	88	95	5.0	0	0	1	0	0	0	0	1	0	0	1	72087796.66780052
12	28	J. Rodríguez	26	1.5	4.0	88	89	4.0	0	0	0	0	1	0	0	1	0	1	0	52603284.42806194
13	31	C. Eriksen	26	2.0	3.0	88	91	4.0	0	0	1	0	0	0	0	1	0	0	1	60773823.19937339
14	35	Marcelo	30	4.0	4.0	88	88	5.0	0	0	0	0	1	1	0	0	0	1	0	40467921.58901108
15	39	Thiago Silva	33	2.0	4.0	88	88	2.0	0	0	0	0	1	1	0	0	0	0	1	40467921.58901108
16	40	S. Handanović	33	3.0	3.0	88	88	1.0	0	0	1	0	0	0	1	0	0	0	1	50378711.89647984
17	43	M. Icardi	25	3.0	3.0	87	90	3.0	0	0	0	0	1	0	0	0	1	0	1	56981571.695679456
18	48	C. Immobile	28	5.0	3.0	87	87	3.0	0	0	1	0	0	0	0	0	1	0	1	44406166.16393798
19	51	J. Vertonghen	31	1.0	3.0	87	87	3.0	0	0	1	0	0	1	0	0	0	1	0	44406166.16393798
20	55	L. Sané	22	3.0	2.0	86	92	4.0	0	0	1	0	0	0	0	0	1	1	0	54647505.145931765
21	56	Bernardo Silva	23	4.0	2.0	86	91	4.0	0	0	1	0	0	0	0	0	1	1	0	54620361.83378414
22	57	Ederson	24	7.0	2.0	86	90	1.0	0	0	0	0	1	0	1	0	0	1	0	53185457.39871359
23	61	Roberto Firmino	26	5.0	3.0	86	87	4.0	0	0	0	0	1	0	0	1	0	0	1	42178772.045172155
24	62	R. Varane	25	4.0	3.0	86	91	2.0	0	0	1	0	0	1	0	0	0	0	1	54581080.58378414
25	69	Aspilicueta	28	4.0	3.0	86	86	2.0	0	0	1	0	0	1	0	0	0	0	1	38824677.62108679
26	70	L. Bonucci	31	5.0	3.0	86	86	2.0	0	0	1	0	0	1	0	0	0	0	1	38824677.62108679
27	71	T. Alderweireld	29	2.0	3.0	86	87	2.0	0	0	1	0	0	1	0	0	0	0	1	42178772.045172155
28	74	M. Özil	29	3.0	4.0	86	86	4.0	0	0	1	0	0	0	0	1	0	1	0	35993077.48450801

---

## 5. 결론

---



# 결론

---

- Train data를 가지고 여러 EDA를 통해 인사이트를 도출한 결과, reputation, stat\_overall, stat\_potential 변수가 이적료에 가장 큰 영향을 미치는 것 파악
- 위의 변수들을 통해 예측 모델인 랜덤 포레스트 모델 학습
- 학습한 모델을 통해 제공받은 Test data 선수들의 이적료 예측

# 아쉬운 점

---

- 데이터 스케일링, 변수 추출 등을 상세하게 했으면 더 정확한 예측값이 나왔을 것 → 추가 공부 필요
- 다양한 머신러닝 기법들을 적용해서 각각을 비교했으면 더 좋았을 것 → 추가 공부 필요
- 여러 부분의 논리 구조가 조금 명확하지 않았던 것 같음

---

# 감사합니다

---