FIFA 데이터 분석 진행 4

이보람

# 2월 4주차-(2) 2.24

## [일반 회귀 분석 vs 분위 회귀 분석]

### 1. 분위회귀분석(Quantile regression)

* 일반 회귀 분석은 종속 변수의 평균을 추정하는 방식이지만, 분위회귀분석(quantile regression)은 특정 분위값을 추정하는 방식이다.
* 종속 변수를 기준으로 분석 대상을 줄을 세우고 이를 n등분 했을 때,

나뉜 구간에서 독립 변수들이 지니는 효과의 크기가 모두 다를 수 있음을 상정하고 분석하는 방법이다.

* Wage가 가장 높은 선수를 1등으로 삼아 그 뒤로 가장 낮은 선수까지 일렬로 줄을 세우고,

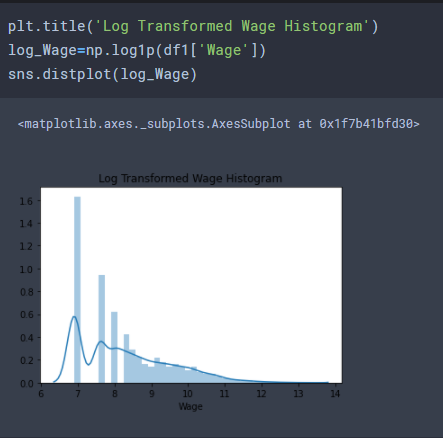
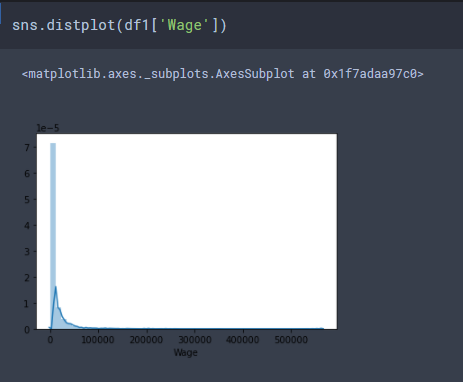
이 줄을 N등분 해서 회귀 분석을 하는 것이다.



**Wage 1위 €565000 – L.Messi**

### 2. Fifa 데이터셋, 일반 회귀 분석의 문제점

* 종속변수의 정규성이 떨어진다.
* Max outlier 수치가 꽤 많다.



로그를 해도 이 모양…….

### 3. 분위 회귀 분석(Quantile regression) 적용이 용이한 경우

-1. 데이터 크기가 충분히 클 때

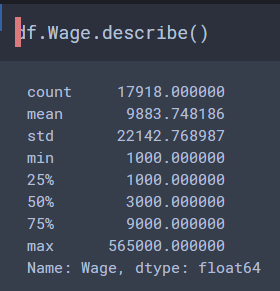
-2. 종속변수의 분포에 이질성이 관찰된 경우

-3. 독립변수(x), 종속변수(y)에 미치는 효과를 구간별로 상세히 알아내야 할 필요가 있을 경우

▷ 일반 회귀 분석의 가장 큰 단점 중 하나는 ‘특정 독립변수의 영향이 모든 구간에서 동일함을 상정한다는 것이다’.

하지만, Quantile regression을 활용함으로써 각각의 구간별로 최적화된 의사결정을 내릴 수 있다.

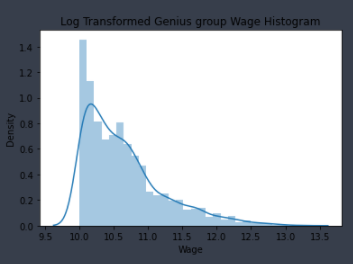
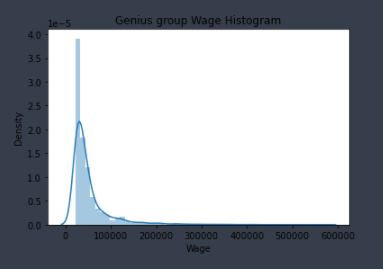
### 4. Wage 이상치(outlier) 탐색 및 종속변수 분리



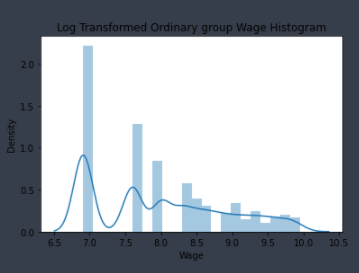
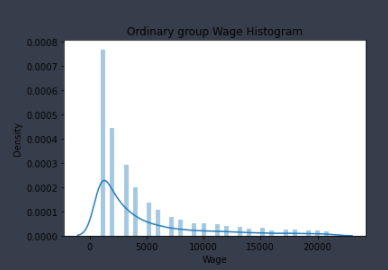
결과: Wage가 과도하게 많은 선수들은 총 2031명, Wage 1000인 선수들은 총 4873명이다.

* Wage 21000 이상인 선수들은 Genius그룹으로, 이하는 ordinary 그룹으로 종속변수를 나눔

1. Genius 그룹, Wage



1. Ordinary 그룹, Wage



# 2월 4주차-(1) 2.23

## [진행내용]

### 1. 구단 가치

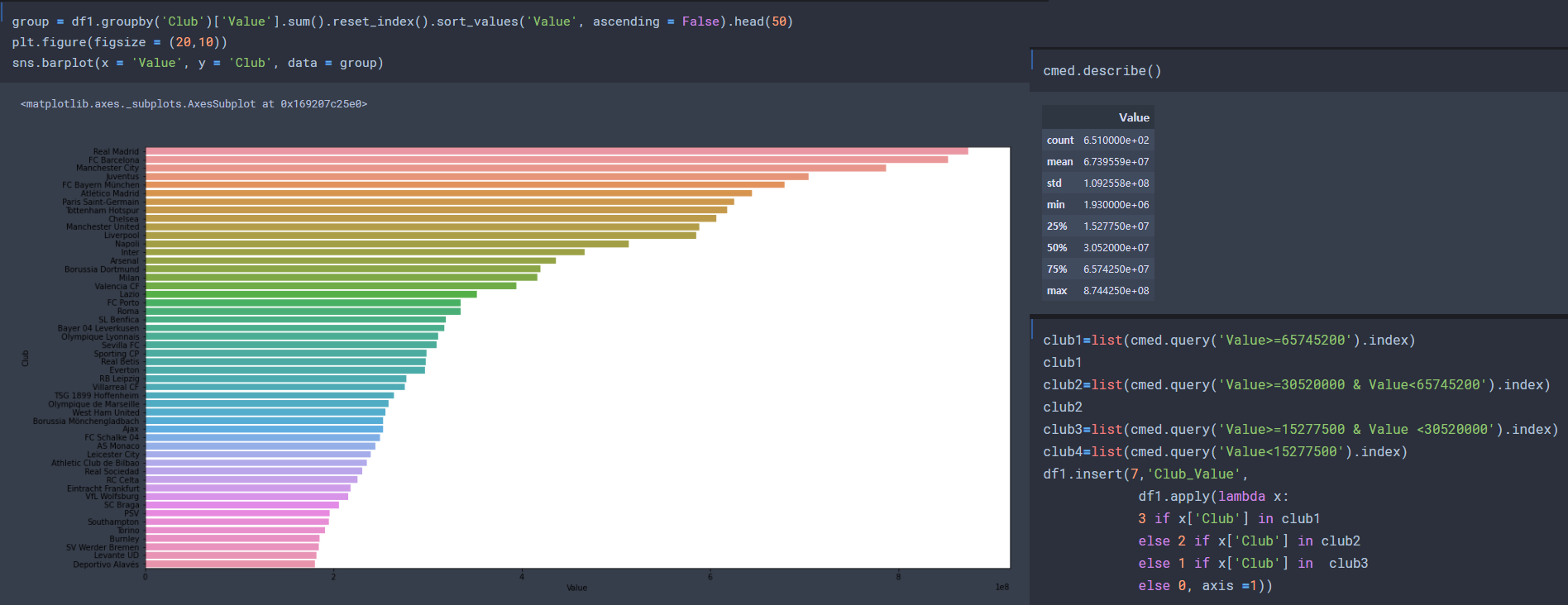
- 축구 클럽의 재정 부분을 말하는 것이 아니라 구단 선수들의 개별 가치를 합산한 것이다.

선수를 싸게 사와서 기량을 높혀 선수 가치를 올리면 구단 가치가 올라가는 것이라고 가정하였고,

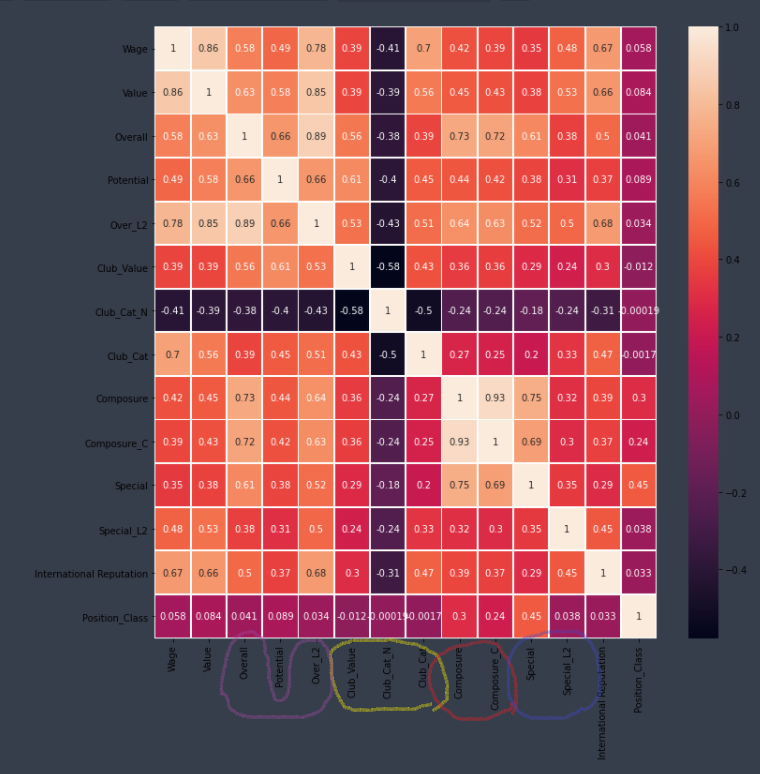
어느 팀이 값어치(Value) 높은 선수들을 많이 보유했는지 보여준다

**구단 가치 TOP10** (단위: 유로, 천만단위에서 반올림)

|  |  |
| --- | --- |
| CLUB | Value Sum |
| Real Madrid | 8.7억 |
| FC Barcelona | 8.5억 |
| Manchester City | 7.9억 |
| Juventus | 7.0억 |
| FC Bayern München | 6.7억 |
| Atlético Madrid | 6.4억 |
| Paris Saint-Germain | 6.3억 |
| Tottenham Hotspur | 6.2억 |
| Chelsea | 6.1억 |
| Manchester United | 5.9억 |

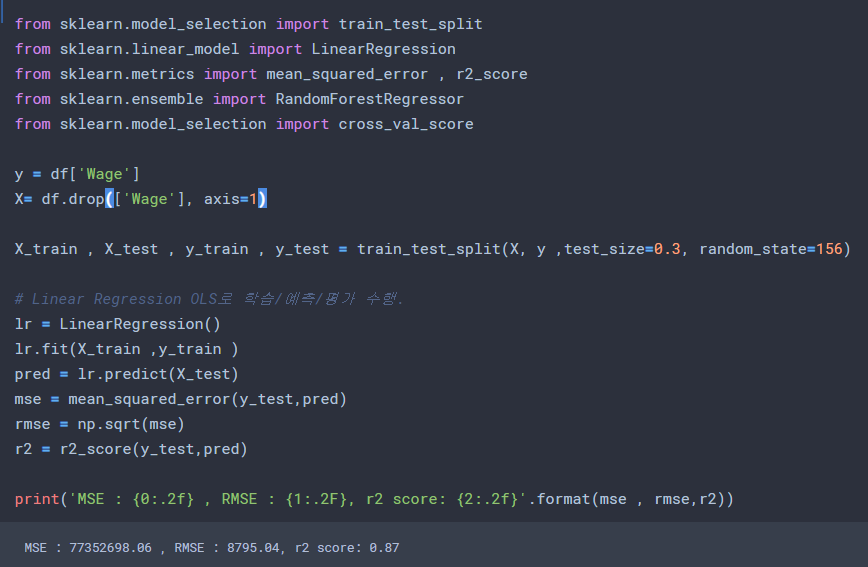


### 히트맵으로 선택할 변수, 상관관계 살펴보기

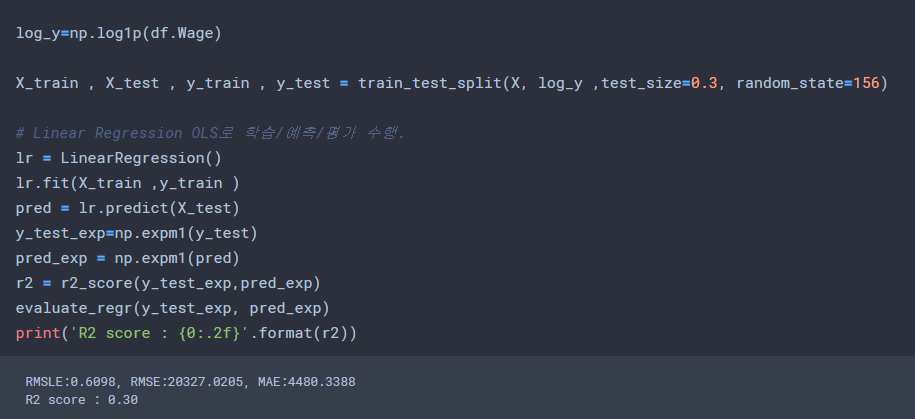


### 독립변수 선택 -> 더미-> 회귀 평가

* 가. LinearRegression



* 나. 종속변수 log화 및 평가



R2스코어가 너무 떨어져서, 로그화를 할 필요가 없다고 판단

* 다. 교차 검증 및 규제 실행(라쏘 alpha값을 살펴볼 때 r2스코어도 함께 체크해야 함)

## [새로운 과제]

### 종속변수 나누기

* Messi 처럼 이상치가 높은 선수들이 매우 높은 Wage를 받으면서 정규성이 확보가 잘 되지 않는다.
* 이상치로 볼만한 특별한 선수 그룹, 평범한 그룹, Wage를 거의 최하로 받는 그룹을 나누어서

모델을 만들어볼 예정이다.

# 2월 3주차

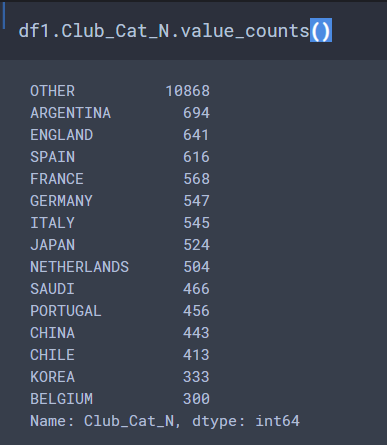
## [진행 내용]

1. **~~(과제4. Club 범주화를 추가해서 비교해보기)~~**

나라별, 리그별로 범주화를 해보았는데 Club종류가 너무 많다.

팀 프로젝트 자료들을 토대로 재작성과 추가를 하였지만, 기타 항목으로 범주화 된 클럽들이

많아 정확도가 오르더라도 아쉽다고 느꼈다.



1. **~~(과제1. Value와 관련된 파생 변수),~~****~~Club\_Value~~**

~~~~

축구 클럽에 대한 정보를 수집하던 중, 재미있는 데이터를 발견했다.

여기서 제시한 **구단 가치**란, 축구 클럽의 재정 부분을 말하는 것이 아니라

구단 선수들의 개별 가치를 합산한 것이다.

선수를 싸게 사와서 기량을 높혀 선수 가치를 올리면 구단 가치가 올라가는 것이라고

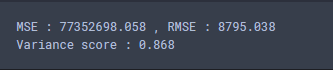
가정하였고, 어느 팀이 값어치(Value) 높은 선수들을 많이 보유했는지 보여준다.

하여, Club\_Value 파생변수를 추가했다.

# 2월 2주차(~2.14)

## [진행 내용]

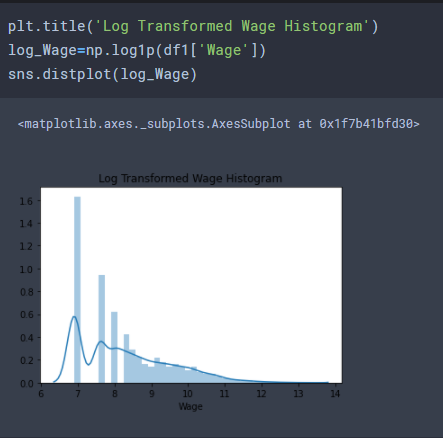
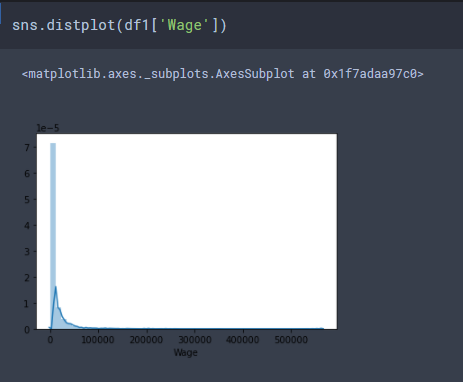
1. MSE scores, 평균 RMSE의 값이 매우 크다.(Wage log전이다 보니..) 과적합 모델, 규제나 데이터 변환이 필요하다



1. Wage 로그화

표준정규분포 모양에 가깝게 만들기 위해 종속변수 로그화.

여전히 예쁜 모양은 아니다;



1. **~~(과제2. Overall, age, value 등 범위를 나눌 때, 명쾌한 기준이 있었으면 좋겠다. )~~**

파생변수 이전의 데이터로 표준화 스케일링이나 최소-최대 스케일링

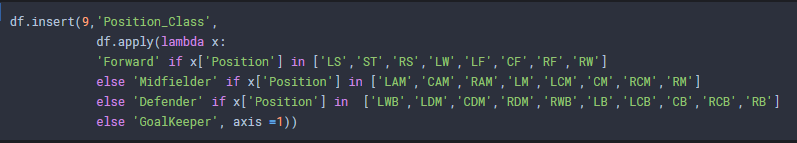
스케일링 개념이 덜 잡혀서 완료하지는 못함.

# 2월 1주차

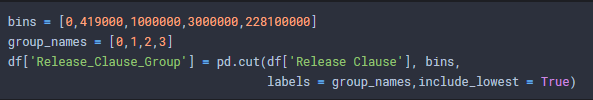
## [코드 & 사용자 함수]

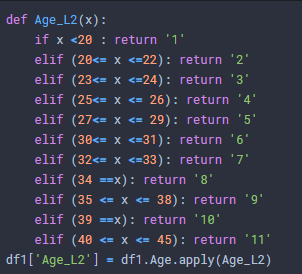
데이터 전처리를 위해 사용한 것(개인적으로 기억하고픈 것 포함)

1. **그룹화1:** Position\_class, Nationality, Club\_cat 등

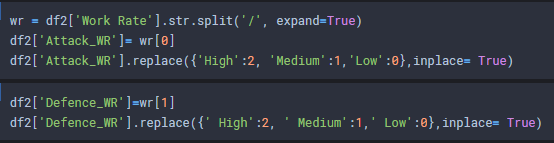


1. **그룹화2:** age, overall 등

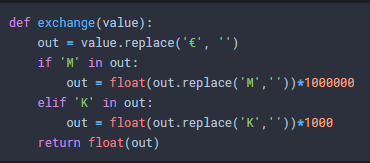


****

1. **변환1:** work rate



1. **변환2:** Wage, Value, Release Clause



5. **하는 중:** joined

## [진행 내용, 과제]

### 진행 내용

* + - 1. 전처리, null 값 처리하기

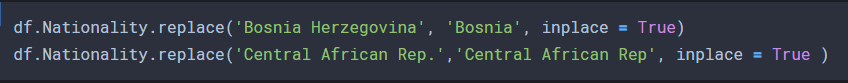


Club~ ability 계열의 모든 수치가 nan인 행 삭제

* + - 1. 불필요 칼럼 선 삭제

'Unnamed: 0','ID','Photo', 'Flag', 'Club Logo', 'Body Type', 'Real Face','Position','Jersey Number', 'Height', 'Weight'

* + - 1. Nationality 칼럼에서 중복인 것 정리하기



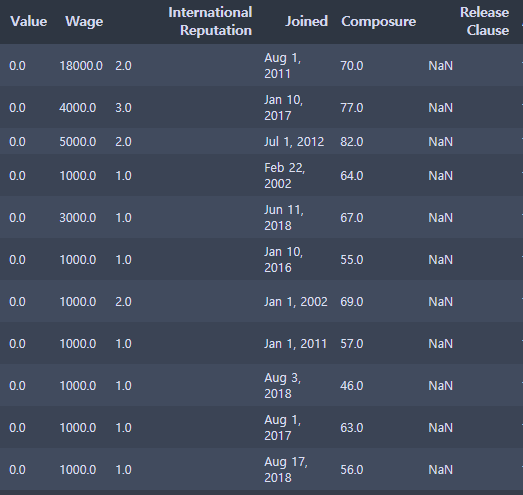
* + - 1. 파생변수 생성
* Nationality > Classified
* 포지션 종류 > Position\_Class
* Club > Club\_Cat (Wage 중앙값으로 정렬 후, 그룹화)

텍스트, 스크린샷, 모니터, 벽이(가) 표시된 사진

자동 생성된 설명

* Age > Age\_L2
* Overall > Over\_L2
* Special > Special\_L2
* Work Rate > ‘Attack\_WR’, ‘Defence\_WR’
  + - 1. 각종 변수 숫자화해서 상관관계를 1차로 살펴봄(히트맵)
      2. Release Clause 가 nan인 경우:

- Joined 는 있지만 Value가 0 (11명)



- Joined 가 nan(1265) > 어떻게 처리할지 고민

### 과제

해결할 것, 가설 내용

1. **~~Value와 관련된 파생 변수~~**
2. **~~Overall, age, value 등 범위를 나눌 때, 명쾌한 기준이 있었으면 좋겠다.~~**

~~(현재는 여러가지로 만들어서 정확도가 올라가는 것으로 선택했음)~~

1. **모델 내 독립변수로 선택할 것과 선택하지 않을 것들의 기준을 세우고 싶다.**

( 공부하기)

1. **~~Club 범주화를 추가해서 비교해보기(나라별- 현재는 Wage 기준으로 범주화)~~**
2. **개인 가설)** 근속연수가 늘어날수록 연봉이 오르는 일반 회사를 생각하자면,
   1. Joined도 일정 범위까지는 Wage에 영향을 주지 않을까?
   2. Joined를 개선해서 살펴보기
3. 시간이 남는다면) **Club~skill종류 칼럼들에서 null값을 가진 48개의 행을 제거했는데, 다른 방안이 있는지 살펴보기**
4. **모델 예측 성능 평가 내용 추가하기**