

**이름(생년월일) : 김태년 (1993.05.26)**

**Edwith ID : taenyun@naver.com**

**팀 정보 : 김태년 (1993.5.26)/정찬우(1994.02.23)**

**난 네가 지난 여름 만든**

**수제맥주가 무엇인지 알고 있다**

**수제맥주 레시피로 맥주 종류 예측하기**

**& 학교생활 정보로 음주량 예측하기**

**여름 밤의 단 한가지 즐거움**



길바닥의 아스팔트조차 녹아 내리는 무더운 여름, 날이 어두워져도 내려갈 줄 모르는 열기에 잠 못 이루는 밤에 우리를 달래주는 것은 시베리아 툰드라 지대에서 온 듯한 시원한 에어컨 바람, 심장이 쫄깃쫄깃한 미드 한편, 그리고 냉장고에서 꺼낸 시원한 맥주 한 캔이다. 이 중에서 가장 중요한 것은 단연 시원한 맥주이다. 사실, 냉장고에서 방금 꺼낸 차가운 맥주 한 캔이면, 에어컨바람이 없어도, 재미있는 드라마가 없어도, 적도지방보다도 더 더운 열기에 한껏 치솟은 불쾌지수가 반은 내려간다.

맥주는 종류가 여러가지다. 사람마다 각자 좋아하는 스타일의 맥주가 있다. 어떤 사람은 시원한 라거(Lager)를 좋아하고 어떤 사람은 향이 진하게 올라오는 에일(Ale)을 좋아하기도 하며, 혹자는 진하고 쌉싸름한 IPA(Indian Pale Ale)를 선호하기도 한다. 최근 들어서는 자신만의 레시피을 통해 직접 만드는 수제맥주가 인기를 끌고 있다. 우리나라도 점점 수제맥주를 파는 맥주집이 늘어나고 있고, 집에서 직접 맥주를 만들어 마시는 사람이 늘어나고 있다.

맥주의 스타일은 어떠한 레시피를 사용하는가에 따라서 달라진다. 제조과정에서 맥아를 어떤 것을 쓰느냐, 아로마는 무엇을 쓰느냐, 그리고 어떤 양조방법(brewing method)를 쓰느냐에 라 전혀 다른 스타일의 맥주가 탄생한다. 그렇다면 어떠한 레시피를 사용하는가를 본다면 어떠한 스타일을 맥주가 만들어지는 지 예측할 수 있지 않을까?

데이터 분석에는 Brewer’s Friend(<https://www.brewersfriend.com>)에서 제공하는 Beer Recipes를 모아 놓은 Kaggle Dataset인 <Brewer’s Friend Beer Recipes> (<https://www.kaggle.com/jtrofe/beer-recipes>)를 이용하였다.

이 데이터셋에는 수제맥주 레시피 사이트에 사람들이 올려 놓은 수제 맥주 레시피에 대한 맥주 스타일 및 알코올 도수 등을 비롯한 23가지의 정보에 대한 데이터가 있으며, 73,000여 종의 수제 맥주 레시피의 데이터가 있다.

**변수 이름과 설명**

**BeerID** : 수제맥주 인덱스 번호

**Name** : 수제맥주의 이름

**URL** : 웹페이지(<https://www.brewersfriend.com>) 상의 레시피 주소

**Style** : 맥주의 종류(우리의 목표변수)

**StyleID** : 맥주의 종류에 대한 숫자형 값

**Size(L)** : 해당 레시피를 통해 만들었을 때 나오는 맥주의 양

**OG** : 발효 전 맥아즙의 비중

**FG** : 발효 후 맥아즙의 비중

**ABV** : 알코올 도수

**IBU** : 국제 쓴맛 단위(IBU, International Bittering Units)

**Color** : 표준참조방법(Standard Reference Method), 0(밝은색)에서부터 40(검은색)까지 있음

**BoilSize** : 끓이기 시작 할 때 액체의 양

**BoilTime** : 맥아즙을 끓이는 시간

**BoilGravity** : 끓임 과정 전 맥아줍의 비중

**Efficiency** : 맥주 담금 추출 효용성 – 담금 과정 중 곡물에서 당분을 추출해 내는 정도

**MashThickness** : 곡물 한 파운드 당 들어가는 물의 양

**SugarScale** : 맥아 내의 용해성 물질의 농도를 측정하는 척도 (Specific Gravity / Plato)

**BrewMethod** : 양조에 이용하는 기술

**PitchRate** : 중력단위 당 발효기에 넣는 이스트의 양 (cells/ml/deg)

**PrimaryTemp** : 발효단게의 온도

**PrimingMethod** : 정보 없음

**PrimingAmount** : 사용된 프라이밍 슈거의 양

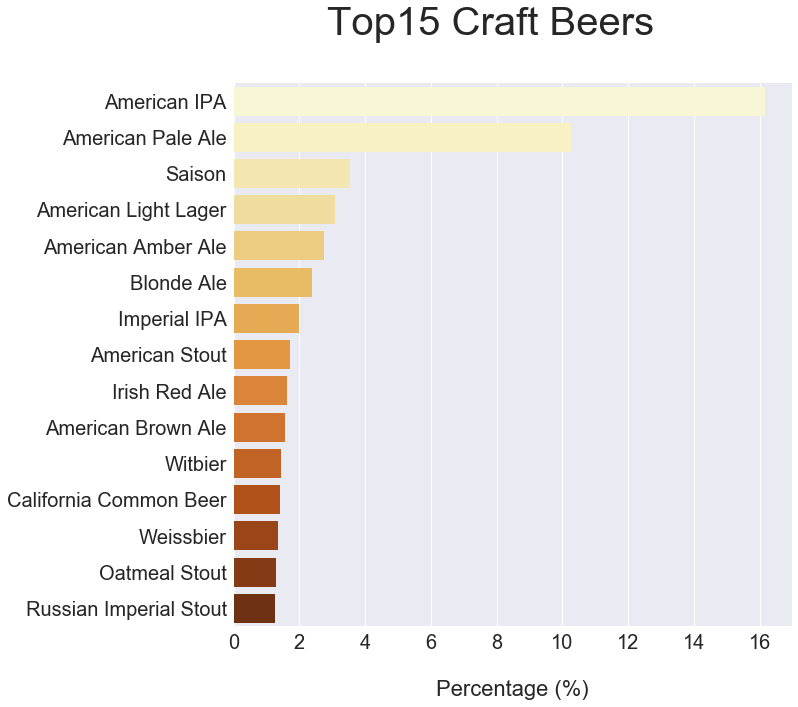
**UserId** : 사용자 ID

**데이터 탐색**

본격적인 데이터 분석에 앞서 이 데이터셋이 담고 있는 정보가 무엇인지 탐색을 해보았다.

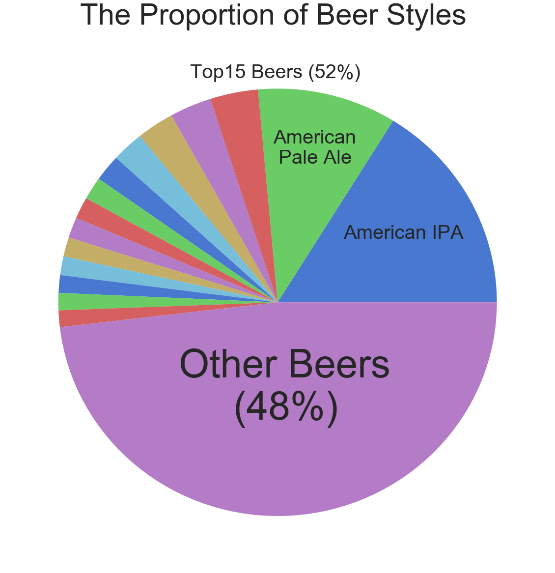
우선 이 데이터셋은 73,861 rows x 23 columns으로 이루어져 있다. 즉, 이 데이터셋은 73,861 개의 맥주 레시피에 대한 23개의 정보를 담고 있다. 또한 이 데이터셋 내의 맥주의 종류는 총 176개이다. 사람들이 어떤 종류의 맥주를 가장 많이 만들었는지 파악하기 위해서 워드 클라우드를 이용하여 맥주 스타일을 분석하였다. 표지에서 보았던 워드 클라우드가 바로 그 결과이다. 이 워드 클라우드를 보면 사람들이 가장 만이 만든 맥주는 아메리칸 아이피에이(American IPA), 아메리칸 페일 에일(American Pale Ale), 아메리칸 라이트 라거(American Light Lager) 등 인 것을 알 수 있다.

워드 클라우드는 정보를 한 눈에 직관적으로 보기는 쉽지만 정확한 수치 등은 알기 어렵다. 이 워드 클라우드 내에서는 가장 많이 만들어지는 세 종류의 맥주가 무엇인지는 어느정도 알 수 있지만, 그 다음으로 많이 만들어지는 맥주가 무엇인가는 잘 알기 힘들다. 이를 보완하기 위하여 특정 맥주 종류가 전체 레시피 중 얼마나 차지하는지를 알아보기 위한 분석을 실시하였다.



이번에는 사람들이 가장 많이 만든 상위 15개 맥주 스타일에 대하여 막대 그래프로 나타내었다. 드 클라우드에서 보았을 때는 사람들이 가장 많이 만든 3대 맥주가 아메리칸 아이피에이, 아메리칸 페일 에일, 아메리칸 라이트 라거로 보였지만, 막대 그래프를 통해 보면 사람들이 아메리칸 라이트 라거보다 새종(Saision)을 더 많이 만들었다는 것을 알 수 있다. 하지만 가장 많이 만든 두 맥주, 아메리칸 아이피에이(16.2%)와 아메리칸 페일 에일(10.2%)에 비하면 아메리칸 라이트 라거는 3%, 새종은 3.5%로 그 차이는 유의미한 정도로 크지는 않았다.

막대 그래프를 통해서 사람들이 가장 많이 만드는 맥주 상위 2개가 전체의 4분의 1인 26.5%를 차지한다는 것을 알 수 있다. 다음으로는 사람들이 가장 많이 만드는 상위 맥주 15개가 전체의 얼마를 차지하는지 쉽게 볼 수 있도록 하기 위하여 원형그래프를 그려보았다.



원형 그래프를 보면 상위 15개 맥주가 전체의 반 이상인 52%를 차지하고 있는 것을 알 수 있다. 즉, 전체 73,861 개의 레시피 중 52%인 38,302개의 레시피는 15개의 스타일 중 하나이고, 나머지 48%인 35,559개의 레시피는 161개의 스타일 중 하나인 것이다.

데이터 탐색 결과, 상위 15개의 맥주가 전체 데이터의 반 이상을 차지하기 때문에 이번 분석에서는 177개의 모든 맥주 스타일에 대하여 분석하는 대신 상위 15개의 맥주에 대해서만 집중하기로 하였다.

**데이터 결측값 확인**

먼저 데이터 전처리에 앞서서 맥주 스타일 예측에 필요 없는 5개의 변수들을 제외하였다. 제외한 변수들은 다음과 같다.

**제외한 변수들**

**BeerID** : 이 변수는 단순 인덱스이고 각 인스턴스의 고유값이므로 제외하였다.

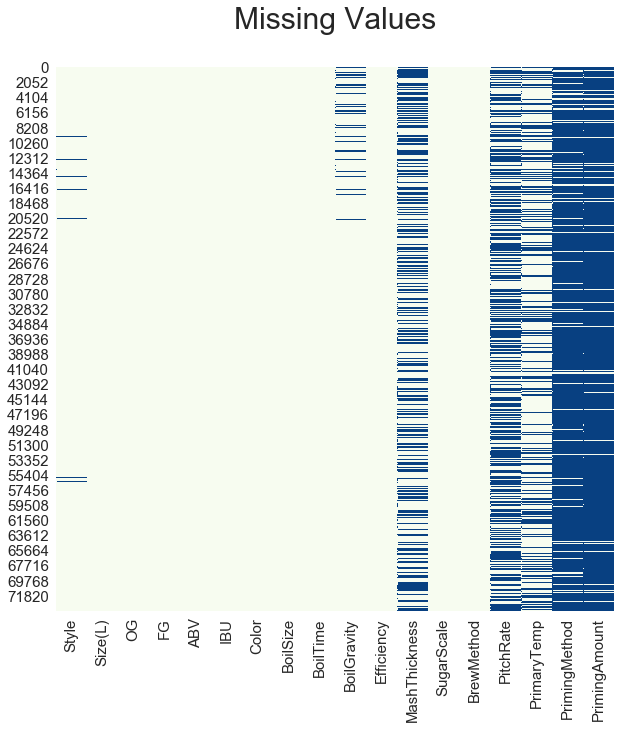
**Name** : 이 변수는 레시피를 올린 사람이 명명한 맥주의 이름이다. 이 변수는 각 인스턴스의 고유값이고, 단순 인덱스로 작용하기 때문에 제외하였다.

**URL** : 이 변수는 레시피의 URL 정보를 담고 있는 각 인스턴스의 고유값이므로 제외하였다.

**StyleID** : 이 변수는 각 맥주 스타일에 지정된 숫자형 인덱스이다. 따라서 ‘Style’ 변수와 100%일치하므로 제외하였다.

**UserId** : 이 변수는 사용자 ID로, 맥주 스타일을 예측 하는데는 상관이 없기 때문에 제외하였다.

그 후 각 변수들에 결측값이 얼마나 있는 지 분석하였다.



이 그래프를 보면 ‘Style’에 약간의 결측값이 존재하고, ‘BoilGarvity’도 어느정도 결측값이 존재하며, MashThickness와 PitchRate, PrimaryTemp, PrimingMethod, PrimingAmount는 대부분의 값이 결측값인 것을 확인 할 수 있다. 다음은 갹 변수들의 결측값이 얼마나 있는지를 표로 나타낸 것이다.

Missing Values

Style 596

Size(L) 0

OG 0

FG 0

ABV 0

IBU 0

Color 0

BoilSize 0

BoilTime 0

BoilGravity 2990

Efficiency 0

MashThickness 29864

SugarScale 0

BrewMethod 0

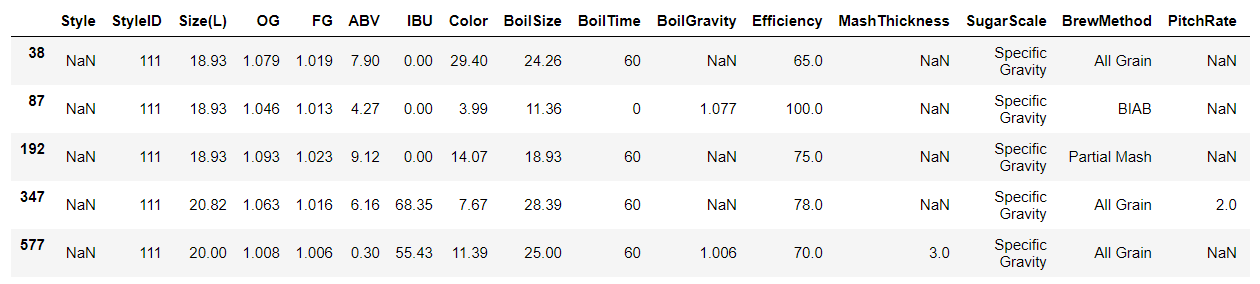
PitchRate 39252

PrimaryTemp 22662

PrimingMethod 67095

PrimingAmount 69085

결측값들을 채우기 위해서 먼저 Style의 결측값이 어떠한 모습인지 살펴 보았다.



이를 보면 Style이 결측값인 인스턴스들은 모두 StyleID가 111임을 알 수 있다. StyleID값과에 따른 맥주스타일을 나타낸 데이터셋을 보면 111은 NaN에 배정되어 있다. 이 값이 NaN값인 것은 다른 176개의 종류 중 다른 어느 종류에도 속하지 않기 때문에 NaN을 배정한 것이라고 생각한다면, 결측값에 NaN 대신 ‘Other Beers’라고 채워 넣는 것이 합리적일 것이다. 물론 이번 분석에서는 상위 15개의 맥주 스타일에 대해서만 분석할 것이기 때문에, 실제 분석에서는 ‘Style’이 ‘Other Beers’인 인스턴스들은 제외될 것이다.

Style의 결측값은 해결하였지만, 여전히 다른 6개의 변수들에도 결측값이 존재한다. 다음으로는 결측값이 존재하는 다른 6개의 변수들 중 어떤 변수를 제외시키고, 어떤 변수를 남겨야 할 것인지를 결정하도록 할 것이다. 먼저 결측값이 있는 변수들은 다음과 같다.

**결측값이 있는 변수들**

**BoilGravity** : 끓임 과정 전 맥아즙의 비중

**MashThickness** : 곡물 한 파운드 당 들어가는 물의 양

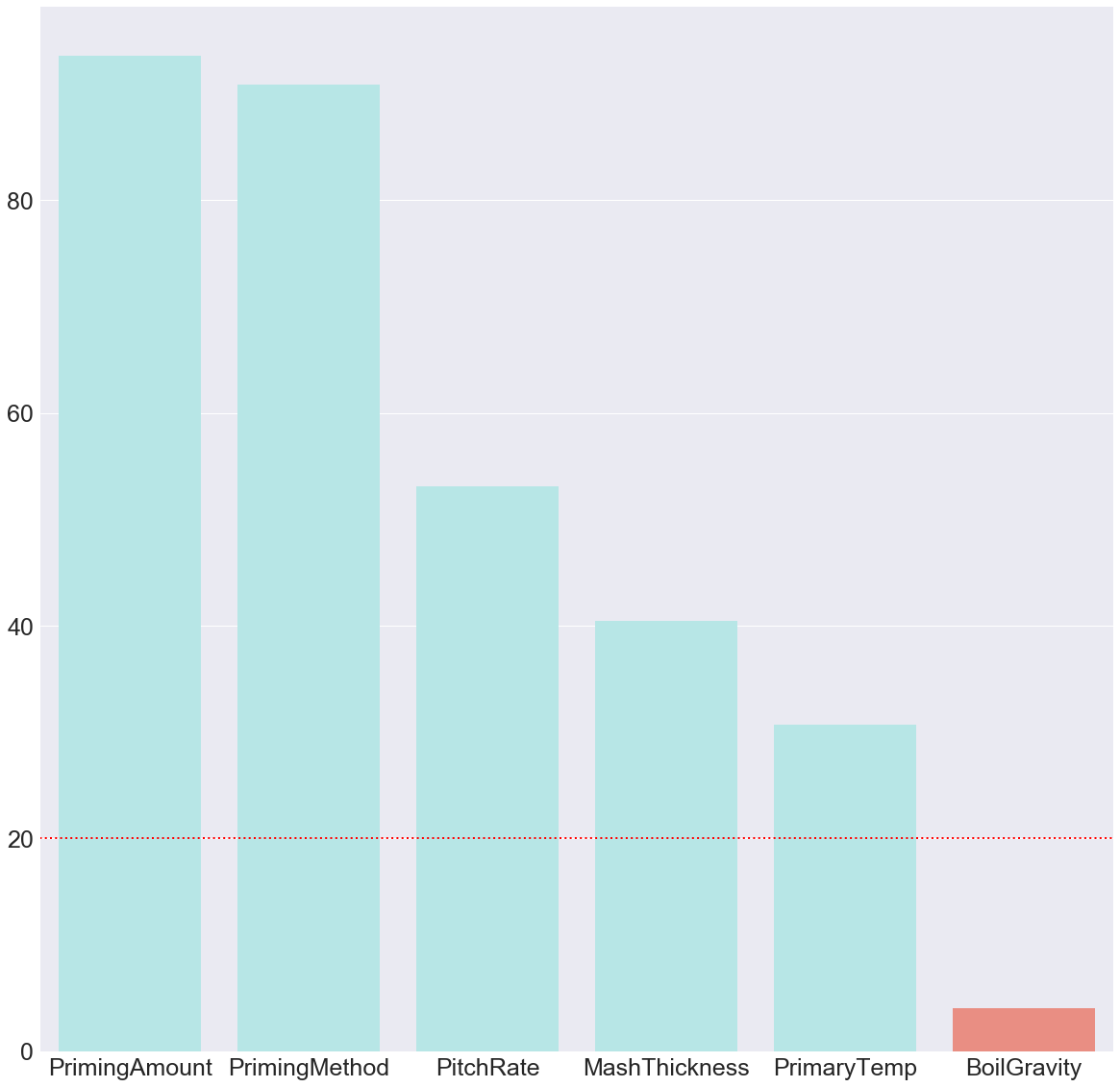
**PitchRate** : 중력단위 당 발효기에 넣는 이스트의 양 (cells/ml/deg)

**PrimaryTemp** : 발효단게의 온도

**PrimingMethod** : 정보 없음

**PrimingAmount** : 사용된 프라이밍 슈거의 양

다음은 결측값의 비율 얼마나 있는지를 나타낸 막대 그래프이다. 이를 보면 PrimingAmount와 PrimingMethod는 90% 이상이 결측값인 것을 알 수 있다.



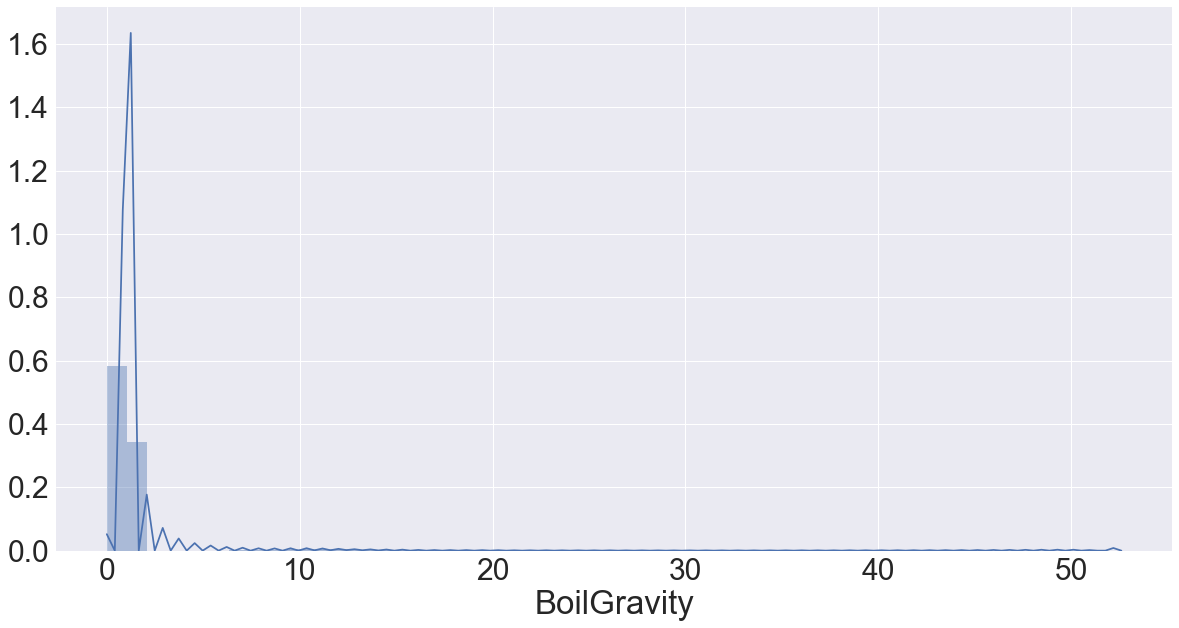
PrimingAmount는 사용된 프라이밍 슈거의 양을 나타낸다. 프라이밍 슈거는 맥주를 병에 넣을 때 넣는 소량의 몰트, 옥수수, 설탕이다. 프라이밍 슈거를 넣으면 병안에서 새롭게 발효가 시작되어 탄산가스가 만들어지고 맥주의 거품이 된다.[[1]](#footnote-1) 프라이밍 슈거를 넣는 것은 맥주를 레시피에 따라 맥주를 완성 한 후에 추가적으로 넣은 것이기 때문에 사실상 맥주 스타일을 결정하는데는 크게 영향을 끼치지 못한다. 게다가 93.5%가 결측값이기 떄문에 결측값을 적절한 값으로 채우는 것도 불가능하다.

PrimingMethod의 경우에는 변수 설명에서도 아예 어떠한 성질인지가 설명이 되어 있지 않다. 게다가 결측값도 90.8%이므로 마찬가지로 제외하는 것이 타당해보인다.

다음으로 MashThickness, PitchRate 그리고 PrimaryTemp에 대해서 살펴보자. MashThickness는 곡물 한 파운드 당 들어가는 물의 양이고, PitchRate는 중력단위 당 발효기에 넣는 이스트의 양이며, PrimaryTemp는 발효단게의 온도이다. 이들 모두 맥주 양조 과정에서 중요한 부분이고, 맥주의 스타일을 결정하는데 중요한 역할을 하는 변수들이다. 하지만 문제가 되는 것은 이들의 결측값이 30-50%라는 것이다.

맥주의 4대 주요 원료는 맥아, 홉, 효모 물이기 때문에 PitchRate가 나타내는 발효기에 이스트의 양은 맥주를 양조하는데 매우 중요한 요소 중 하나이다. 하지만 53.1%가 결측 값이기 때문에 사실상 적절한 결측값을 찾는 것이 불가능하다. 결측값이 많지 않다면 적절한 값을 찾아 결측값을 채우는 것이 맥주 스타일 예측에 도움이 될테지만, 결측값이 많다면 그만큼 적절한 값은 찾는 것도 어렵고 결과 예측에 도움이 되지 않을 것이다.

따라서 이번 분석에서는 결측값의 비율이 20%이상이면 제외하기로 한다. 그렇게 된다면 결측값이 있는 나머지 6개의 변수들 중 결측값이 4%에 불과한 BoilGravity만 남게 된다. BoilGravity는 끓임 과정 전 맥아의 비중이다. 맥아는 맥주의 4대 주요 원료 중 하나이기 때문에 이 결측값에 적절한 값을 넣는 것이 매우 중요하다.



그래프를 보면 BoilGravity의 값은 대부분 1 주변에 몰려 있지만, 최대값은 50이 넘는 것을 이상한 형태인 것을 알 수 있다. 이는 평균, 최대값, 최소값, 중간값 등이 기술되어 있는 다음 표를 보면 더욱 잘 드러난다.

Boil Gravity

mean 1.353955

std 1.930989

min 0.000000

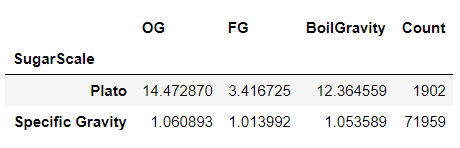
25% 1.040000

50% 1.047000

75% 1.060000

max 52. 600000

전체의 75%까지도 1.06이지만 최대값은 52.6인 것이다. 어째서 이러한 결과가 나왔는지 분석을 해보니 ‘SugarScale’이 어떤 값이냐에 따라 BoilGravity와 OG, FG의 값의 차이가 크게 나는 것을 알 수 있다. 이는 BoilGravity, OG, FG 세 변수가 모두 맥아즙과 관련된 것이고, SugarScale은 맥아즙 내의 당분의 양을 측정하는 다른 방법에 관련된 변수이기 때문이다.

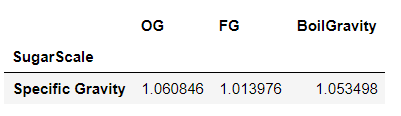


즉, BoilGravity의 값의 대부분이 1 주변에 모여 있는데, 일부 값들이 차이나게 큰 것은 Plato와 Specific Gravity라는 다른 용해질 측정방법 때문인 것이다.

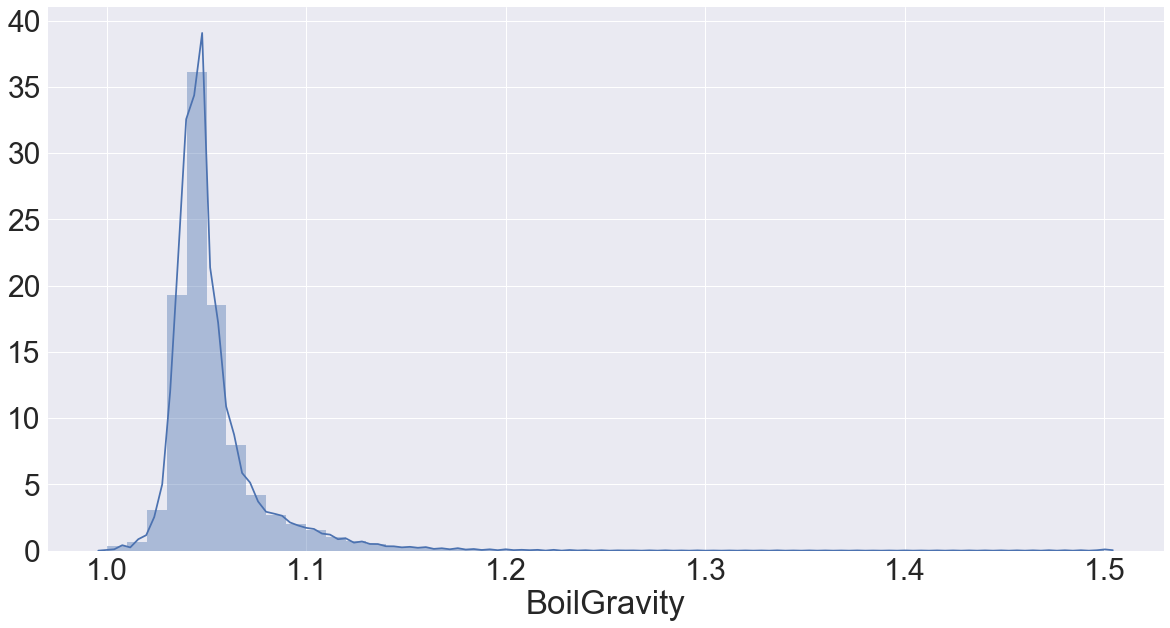
Plato는 용해질의 무게 대비 추출물(혹은 당분)의 무게를 백분율로 나타낸 것이다. 즉, 10°P는 용해질의 무게 중10%가 추출물이라는 것이다. 이 방법은 양조 시 계산이 용이하다. 반면, Specific Gravity는 추출물의 무게 대신 같은 부피의 물에 대비해서 같은 부피의 용해질이 얼마나 무거운지에 대한 것이다. 따라서 1.060 부피의 맥아즙은 같은 부피의 물보다 6%더 무겁다는 것을 의미한다. Plato를 SG변환하는 공식은 다음과 같다.

SG = 1+ (plato / (258.6 – ( (plato/258.2) \*227.1) ) ) *[[2]](#footnote-2)*

위 공식을 이용해 SugarScale을 SG로 통일하였다. Plato를 SG로 변환한 뒤의 OG,FG, BoilGravity의 평균은 다음과 같다.



또한 BoilGravity의 분포는 다음과 같다.



그래프를 보면 BoilGravity의 값이 1.0부터 1.5 사이이며, 대부분 1.0과 1.1 사이에 위치한, 왼쪽으로 편향된 분포인 것을 알 수 있다. 다음은 Plato를 SG로 변환 한 뒤의 평균, 최대값, 최소값, 중간값 등이다.

BoilGravity

mean 1.053498

std 0.027301

min 1.000000

25% 1.040000

50% 1.047000

75% 1.058000

max 1.500000

그래프가 왼쪽으로 편향되어 있기 때문에 결측값에 평균값인 1.053을 채워 넣는 것보다는 중간값인 1.047을 이용하는 것이 더 적절해 보인다.

Missing Values

Style 0

Size(L) 0

OG 0

FG 0

ABV 0

IBU 0

Color 0

BoilSize 0

BoilTime 0

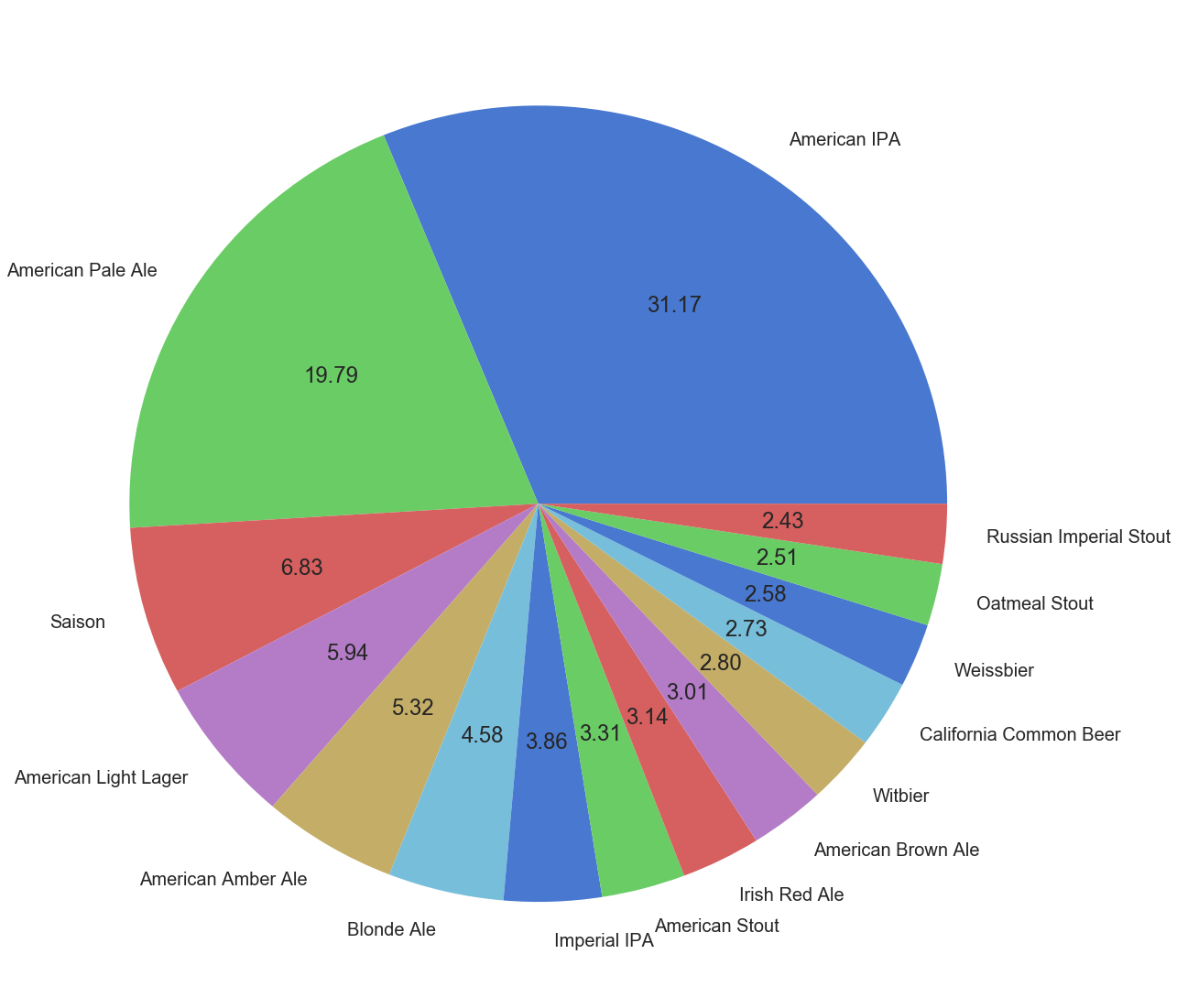
BoilGravity 0

Efficiency 0

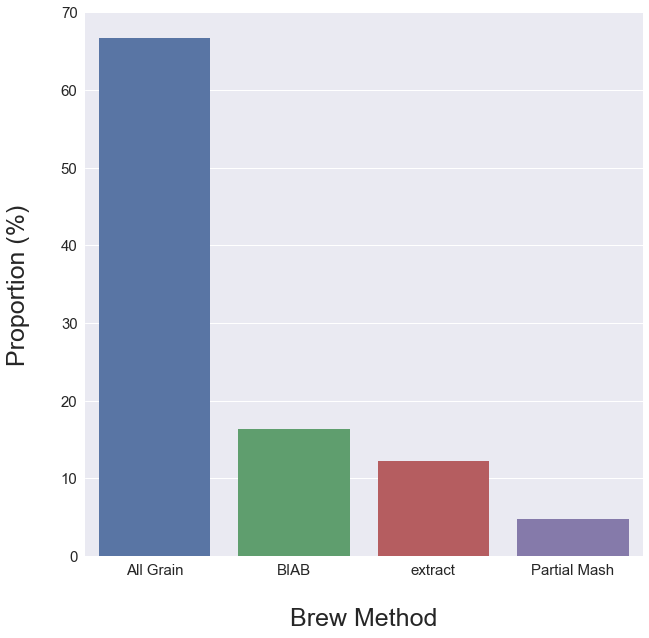
BrewMethod 0

이제 모든 변수의 결측값이 0이 되었다.

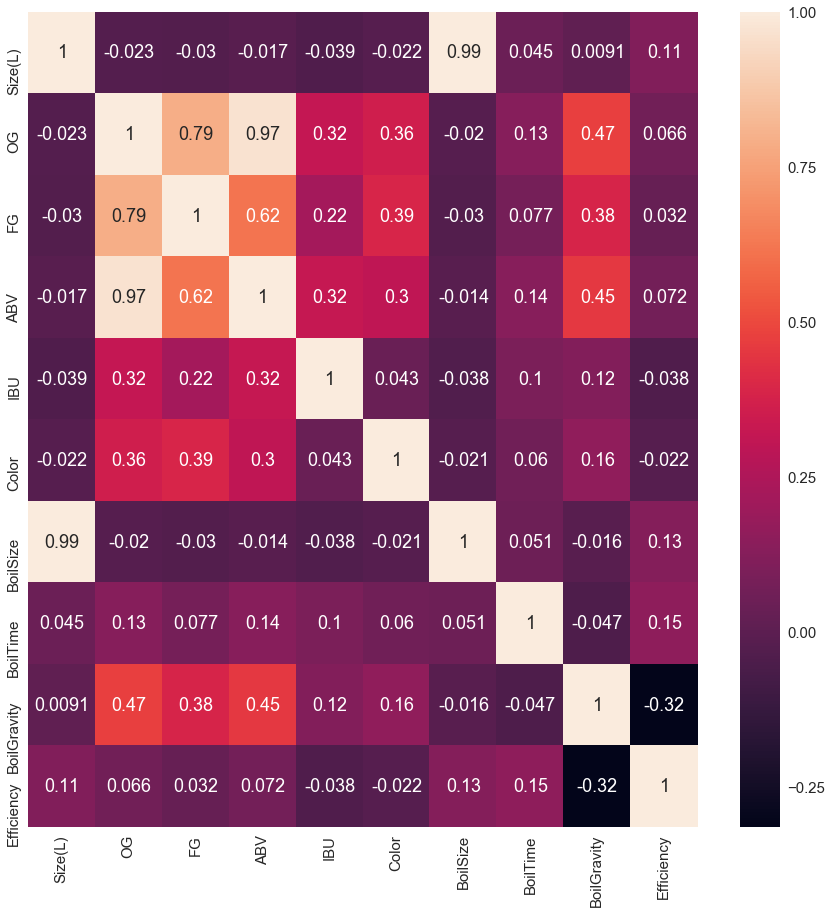
**데이터 전처리**



상술하였듯이 이번 분석에서는 상위15개의 맥주 스타일에만 집중하고 그 외의 맥주 스타일은 제외하도록 한다. 상위 15개의 맥주 스타일만 남기게 되니 전체 데이터의 인스턴스는 최초 데이터셋의 52%인 38,502개가 남았다.

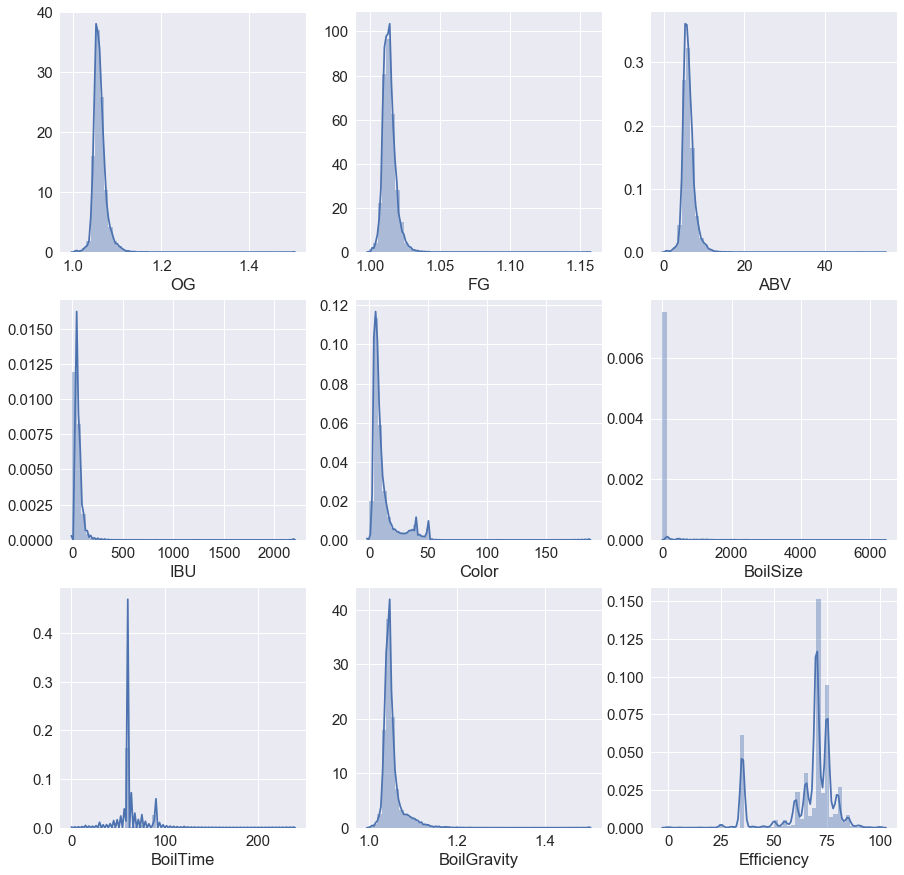


남아 있는 변수들 중 숫자형이 아닌 변수는 목표 변수인 Style을 제외하면 BrewMethod가 유일하다BrewMethod가 어떻게 이루어져 있는지 그래프를 통해 보도록 하자. 위 그래프는 상위 15개 맥주의 양조 방법(Brew Method)의 비율이다. 이를 보면 모든 방법 중 All Grain 방법이 전체의 65%정도를 차지하고 나머지 방법들이 비슷한 비율로 차지하고 있는 것을 볼 수 있다.



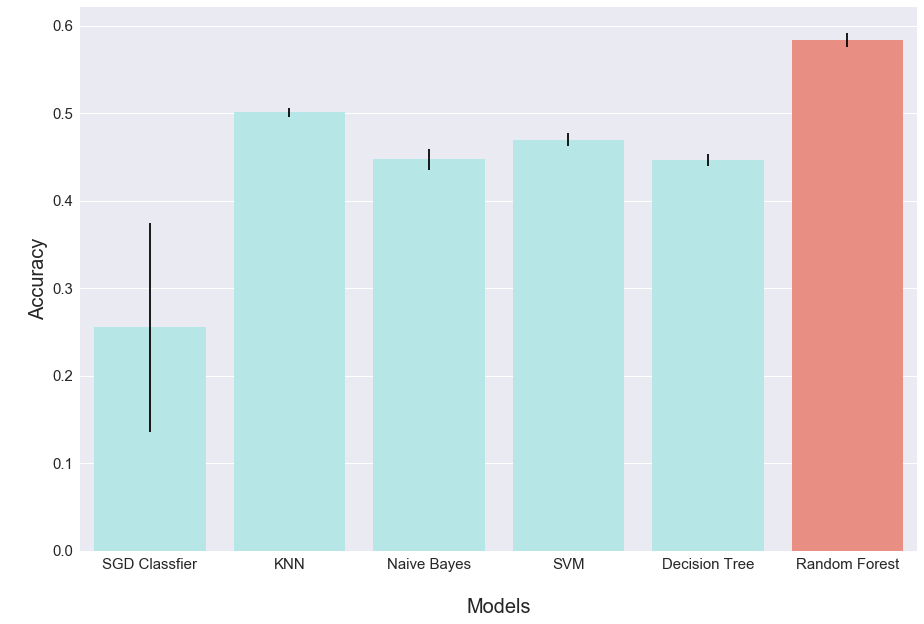
이번엔 숫자형 변수들 간에 상관관계를 히트맵을 통해 보도록 하자. 히트맵을 보면 OG, FG, ABV 간에 중간에서 강한 정도의 상관 관계가 있는 것을 알 수 있다. OG와 FG는 발효 전후의 맥아즙의 비중이기 때문에 0.79의 강한 상관 관계를 보인다. 또한 맥아즙이 발효를 통해 알코올이 생성되기 때문에 ABV 역시 OG(0.97), FG(0.62)와 중간에서 강한 정도의 상관 관계를 가진다.

또한 Size(L)과 BoilSize는 0.99로 거의 같은 값을 보인다. BoilSize는 끓이기 시작할 때 액체의 양이고, Size(L)은 그 레시피를 통해 완성된 맥주의 양이기 때문에 사실상 같은 값이다. 이렇게 서로 상관관계가 매우 높은 값이 중복으로 있게 된다면 결과 예측에 문제가 발생할 수 있다. 그렇기 때문에 중복된 값 중 하나를 제외하는 것이 결과를 좋게 만드는 방법이 될 수 있다. 이번 분석에서 실시하고자 하는 것은 레시피를 통해서 맥주 스타일을 분석해내는 것이기 때문에 레시피 과정에 포함된 BoilSize보다는 레시피를 통한 결과인 Size(L)을 제외하는 것이 더 적절해 보인다.

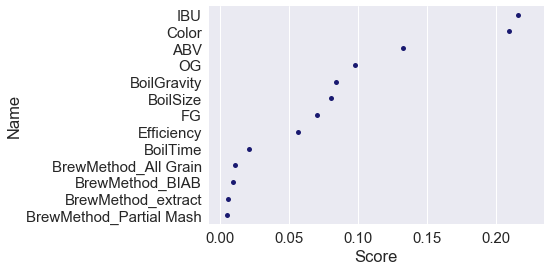


위 그래프는 최종적으로 남은 숫자형 변수들의 분포도 이다.

**머신 러닝**



지금까지 전처리한 변수들을 이용해서 레시피만을 가지고 맥주의 스타일을 예측할 수 있는지 알아보자. 레시피를 통해 맥주의 스타일을 예측하는 것은 분류과제이기 때문에, Stochastic Gradient Descent Classifier와 K Nearest Neighbor, Naïve Bayes, Support Vector Machine, Decision Tree 그리고 Random Forest 등 총 6개의 알고리즘을 가지고 시험을 해보았다. 검증방식은 K-Fold 교차검증을 실시하였으며, 총 5개의 Fold를 만들어 검증을 실시하였다. 검증 결과 Random Forest가 58.3%의 정확도로 다른 모델들 보다 평균 10% 더 정확한 예측을 하였다.



Random Forest의 어떤 Feature가 레시피를 통해 맥주 스타일을 예측하는데 기여가 컸는지를 확인하기 위해 Feature Importance를 확인해보았다. 그 결과 IBU가 21.6% 기여하였고, Color가 20.9%, 그리고 ABV가 13.2% 순이었다. 이 세 피처가 전체의 모델의 55.6%를 설명한다.

이것은 맥주 스타일을 예측하는데 가장 중요한 것이 맥주가 얼마나 쓴지, 맥주의 도수가 얼마나 되는 지, 맥주의 색깔은 어떠한 지라는 것을 보여준다.

**자기야, 이 맥주는 종류가 뭐야?**



마치 거리에 나가면 야자수가 자라고 있을 것 같은 무더운 여름날의 밤, 당신은 시원한 맥주를 마시고 있다. 그런데 이번에는 시원한 에어컨 바람 밑에서 미드를 보면서가 아니다. 이번엔 당신이 마음에 두고 있는 사람과 함께 세계맥주할인점에서 맥주를 마시고 있다. 보고만 있어도 입가에 미소가 지어지는 그 사람이 맥주 한 병을 냉장고에서 꺼내 온다. 그리고는 당신에게 이 맥주가 무슨 종류인지에 대해서 묻는다. 그런데 어쩌나, 그 맥주는 당신이 처음 보는 맥주이다. 역시 맥주의 세계는 넓고도 광활하구나. 하고 당신은 생각한다. 그러나 당신은 전혀 당황하지 않는다. 그 맥주의 쓴맛의 정도, 알코올 도수, 그리고 색깔만 유심히 본다면, 당신은 그 맥주가 어떠한 맥주인지 맞출 수 있다. 당신은 그 사람에게 그 맥주가 어떤 종류인지 말해준다. 그리고 그 사람은 당신을 존경어린 눈으로 쳐다본다. 아니, 어쩌면 속으로는 주당이라고 생각하고 있을 수도….

**술은 적당히…**

무더운 여름날 맥주를 마시는 것은 기분 좋은 일은 없다. 하지만 모든 것에는 적당히라는 것이 있는 법이다. 술을 지나치게 마시면, 공부에도 영향을 미치고, 가족관계나 건강, 이성관계에도 영향을 미칠 수 있다. 혹은 반대로 이성관계를 망쳐서, 혹은 공부가 안 되어서, 혹은 가족관계가 좋지 못해서 술을 마실 수도 있다. 그 인과관계가 어찌되었든, 음주와 학업성적, 이성관계, 가족관계 등의 여러 변수들은 상관관계가 존재한다. 즉, 이러한 변수들을 살펴봄으로써 당신이 음주를 얼마나 하는 지를 알아 볼 수 있다는 것이다. 지금까지 맥주의 종류를 어떻게 예측하는가에 대해서 알아보았다면, 이번에는 당신이 음주를 얼마나 자주하는 지를 예측할 수 있는지에 대해서 알아보도록 하자.

1. https://terms.naver.com/entry.nhn?docId=386604&cid=50346&categoryId=50346 [↑](#footnote-ref-1)
2. <https://www.brewersfriend.com/plato-to-sg-conversion-chart/> [↑](#footnote-ref-2)