# SKILLCRAFT EXPLANATORY DATA ANALYSIS
# 스킬크래프트 탐색적 자료 분석

SNU 4차산업혁명 아카데미
빅데이터 애널리틱스 양서윤

# 데이터 개요와 분석 목적

- 데이터 개요 : RTS 멀티플레이 온라인 게임인 스타크래프트의 유저 활동 정보를 담은 Skillcraft 데이터. ID와 나이와 같은 개별 정보부터 핫키, 미니맵, 유닛 수 등 게임에서 가능한 여러 동작들에 관한 수치가 저장되어 있다. 특히, 유저의 등급이 기록되어 있어 등급별로 게임 활동의 특징을 분석하기 용이하다.

- 데이터 변수 설명 : 총 8개의 레벨 중 플레이어의 등급을 나타내는 변수 LeagueIndex를 비롯하여 유저의 나이, 게임 시간, 핫키 사용 수, 유닛 수 등 총 20개의 변수로 구성되어 있다.

- 분석 목적 : 해당 데이터는 19개의 서로 다른 변수들의 등급 별 영향 유무와 그 강도에 대한 정보를 담고 있다. 특히 본 과제에서는 통계량과 분포의 형태를 통해 등급들과 선형 관계를 보이는 변수들, 즉 유저가 해당 변수 값을 키우면 등급도 비례해서 높아질 수 있는 변수들을 중심으로 데이터 분석 및 추론을 하는 것을 목적으로 삼았다.

- 데이터 출처: http://archive.ics.uci.edu/ml/datasets/skillcraft1+master+table+dataset

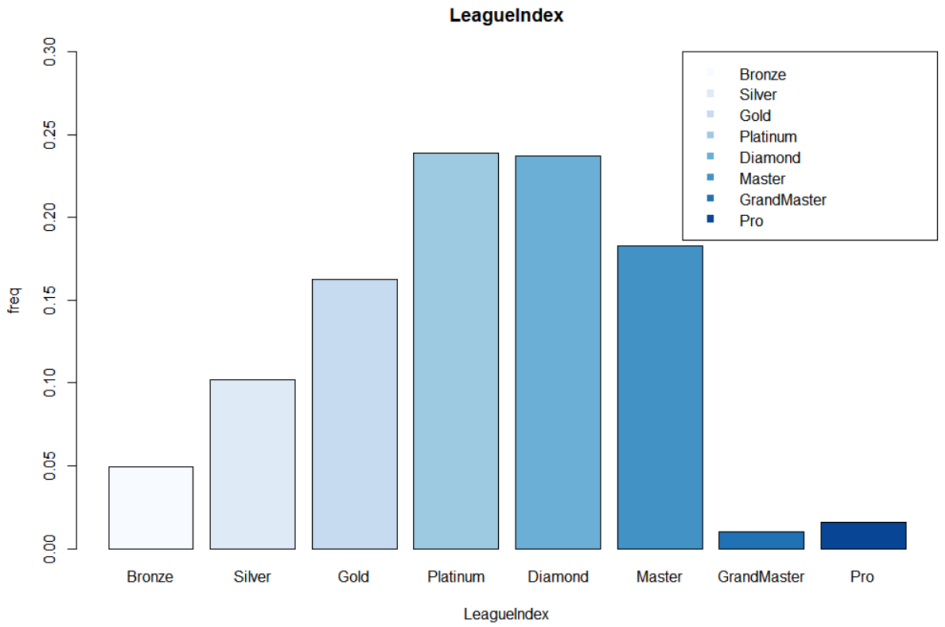
# 데이터 불러오기 및 전처리

```
# 파일 및 라이브러리 불러오기
setwd('C:/Users/yuniv/OneDrive/문서/R/explanatory-data-analysis/data')
skill <- read.csv(file = 'SkillCraft1_Dataset.csv', header=TRUE,
                  stringsAsFactors=FALSE, na.strings=c('?'))
library(dplyr)
library(RColorBrewer)

# 결측값 확인 및 NA 처리
table(is.na(skill))
skill[skill=='?'] == NA

# 데이터 요약치 확인
dim(skill)
str(skill)
names(skill)
attach(skill)
summary(skill)
```

- Skillcraft 데이터 파일이 저장된 디렉토리를 세팅하고, read.csv 함수로 데이터를 불러왔다.
- 탐색적 자료 분석에 필요한 라이브러리들(dplyr, RColorBrewer)을 불러왔다.
- 데이터 내의 결측값을 확인하고, 결측값들의 처리를 용이하게 하기 위해 NA로 변환했다.
- 데이터는 총 3395개 행과 20개 열(변수)로 구성되어 있으며, summary 함수로 변수 별 통계량을 확인한다.

# 등급 별 데이터 분류

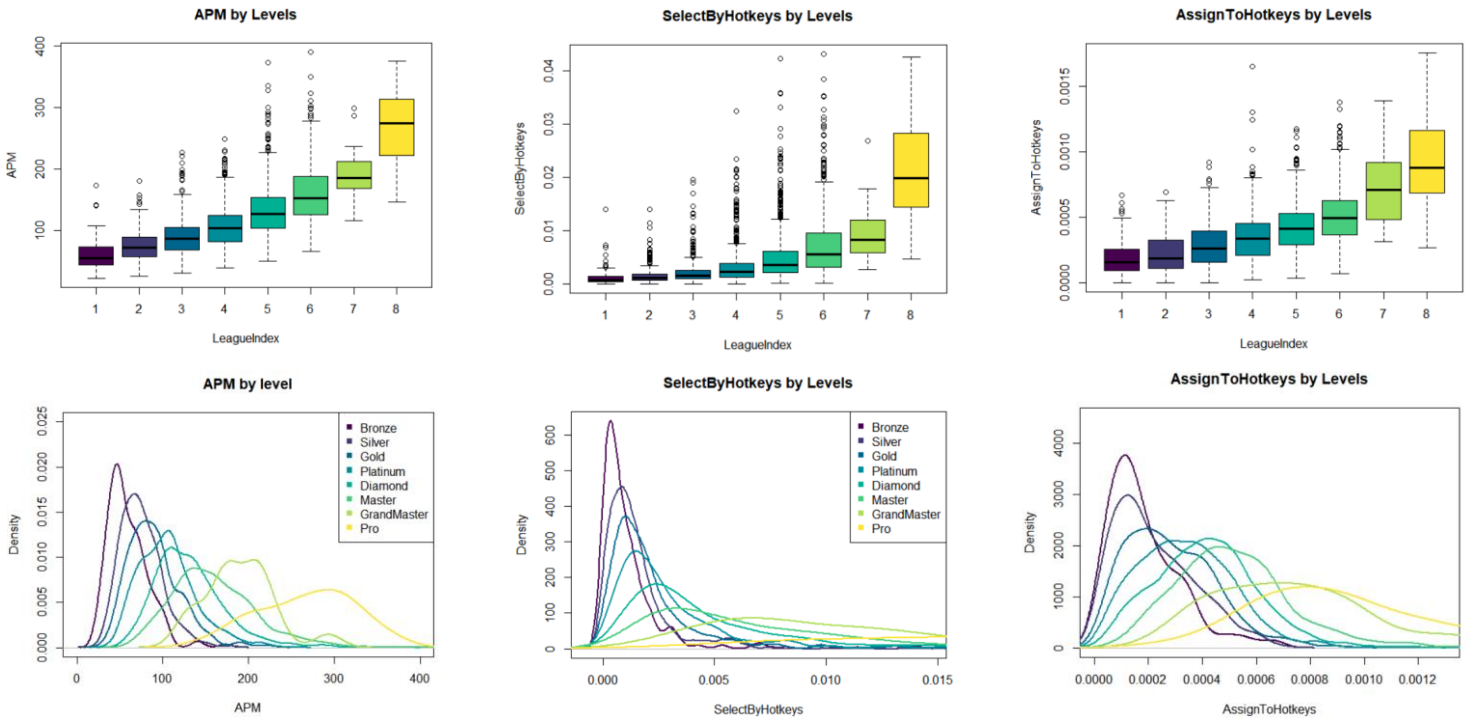
```
# 등급 지정, 등급 별 데이터 분류
level.name = c('Bronze', 'Silver', 'Gold', 'Platinum',
               'Diamond', 'Master', 'GrandMaster', 'Pro')
level1 = skill[skill$LeagueIndex == 1, ]
level2 = skill[skill$LeagueIndex == 2, ]
level3 = skill[skill$LeagueIndex == 3, ]
level4 = skill[skill$LeagueIndex == 4, ]
level5 = skill[skill$LeagueIndex == 5, ]
level6 = skill[skill$LeagueIndex == 6, ]
level7 = skill[skill$LeagueIndex == 7, ]
level8 = skill[skill$LeagueIndex == 8, ]

# 등급 8개의 색상 지정
mycol = brewer.pal(8, 'Blues')

# 등급 분포 시각화
counts = table(skill$LeagueIndex) # 레벨 별 빈도수
freq = counts/sum(counts) # 레벨 별 상대도수
barplot(freq, main = "LeagueIndex", xlab = "LeagueIndex",
        ylab = "freq", ylim = c(0,0.3),
        names.arg=level.name, col=mycol)
legend("topright", level.name, col=mycol, pch=15)
```




- 등급 별 분포를 확인하기 위해 데이터프레임의 인덱싱을 사용해 각 레벨에 해당하는 데이터를 추출하여 각각 새로운 변수에 저장하고, barplot을 이용하여 등급 별 상대빈도를 시각화 했다.
- 상식적으로 최고 수준의 플레이어 수가 그보다 낮은 등급의 플레이어 수보다 많지 않아 데이터가 위와 같이 형성되어 있다고 말할 수 있으나, 등급 별로 데이터 수의 차이가 꽤 있으므로 정확한 분석은 어려울 것으로 생각된다(특히 높은 등급인 그랜드마스터와 프로의 경우 다른 등급에 비해 데이터 수가 현저하게 적다).

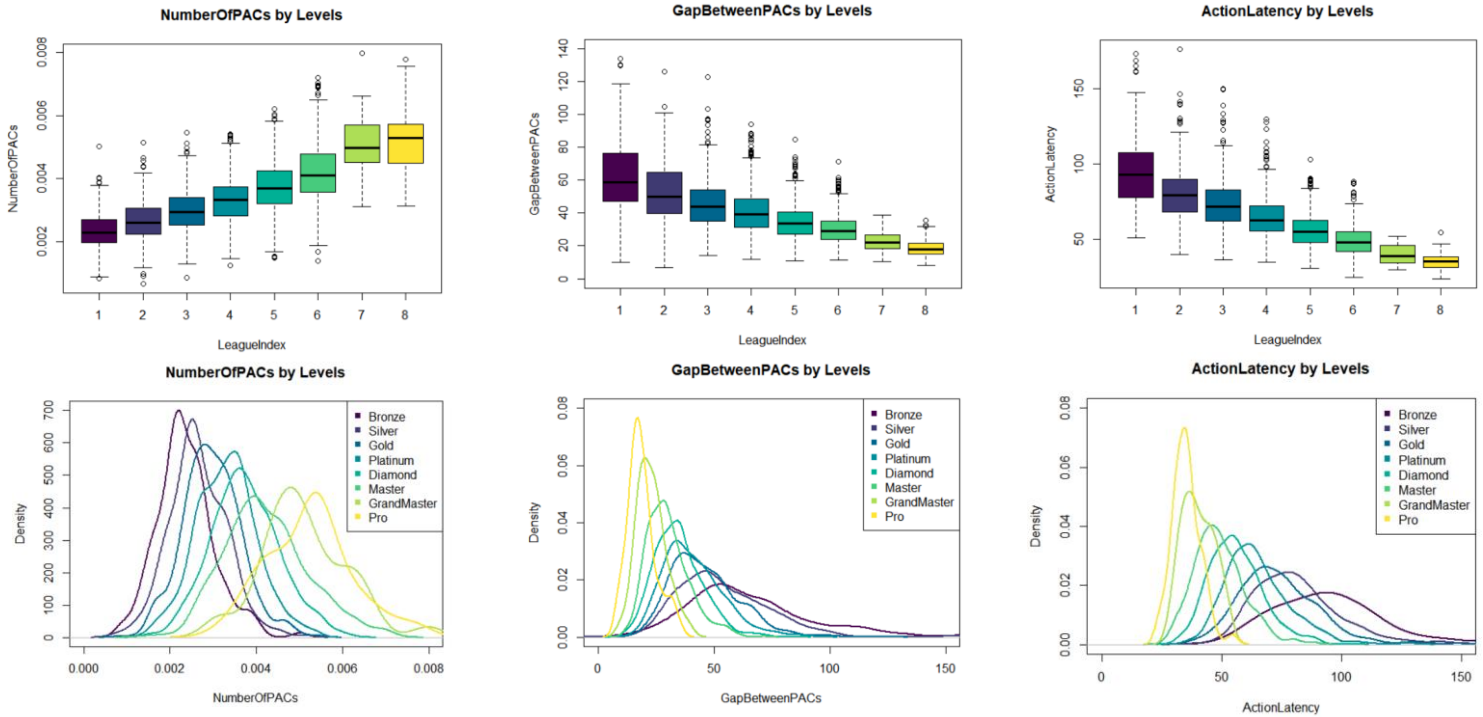
# 등급 비례 변수 – 1*

* 등급에 비례하여 그 값이 등급과 함께 계속 증가하는 변수를 “등급 비례 변수”라고 부르기로 한다.



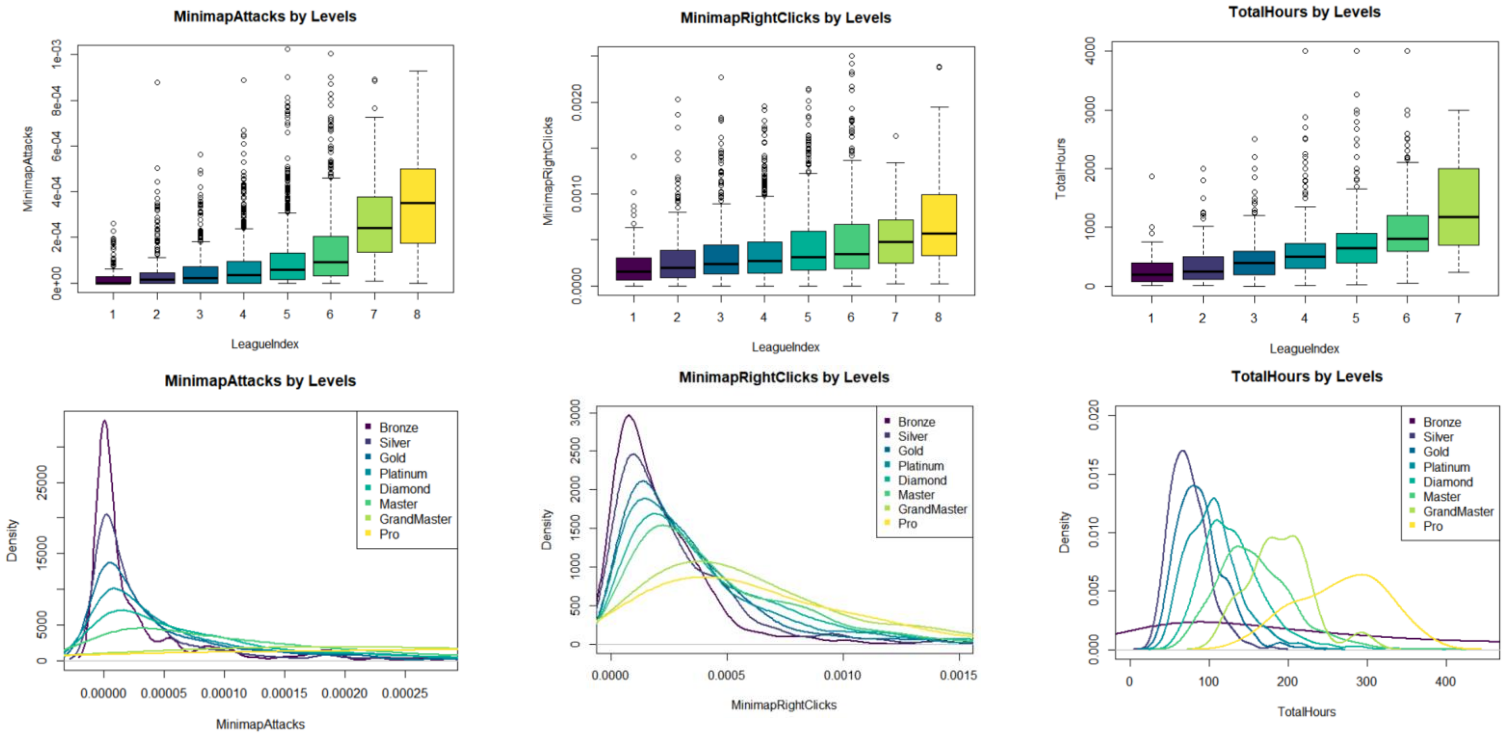

- 유저의 명령 입력과 연관된 변수들인 APM(분당 입력한 명령 수), SelectbyHotkeys(단축키 사용 평균 횟수), AssignToHotkeys(핫키에 할당된 건물, 유닛 수)는 대략 레벨에 비례하여 상승하는 추세를 보인다.
- APM과 AssignToHotkeys는 1등급부터 7등급까지 데이터 값이 고르게 증가하고 있으며, 등급이 높아질수록 유저들 사이의 차이는 커지는 즉 분산이 커지는 공통점을 가지고 있다.
- APM과 SelectbyHotkeys는 최고 등급인 8등급에서 값이 급상승하는 모습을 보인다. 두 변수는 7등급과 8등급의 차를 극명하게 보여주기 때문에, 8등급으로 레벨이 상승하는 데 중요한 요인으로 추측할 수 있다.

# 등급 비례 변수 – 2



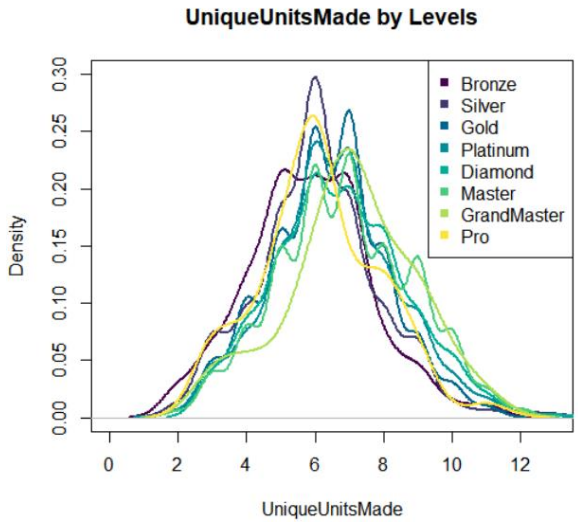

- boxplot과 density plot으로 대략적인 분포를 살펴본 결과, 특정 유닛을 클릭한 후 명령을 완료하는 동작인 PAC(Perception-Action-Cycles)도 등급과 일종의 비례 관계를 형성하는 데이터 분포를 보였다.
- PAC의 수를 나타내는 NumberOfPACs는 1에서 6등급까지 거의 비슷한 정도로 증가하다가 6, 7등급 사이에서 두드러진 상승세를 보인다. 미세한 차이이긴 하나, 각각의 밀도 함수도 평균이 조금씩 커지고 분산이 높아지는 경향을 보인다.
- 특별히 명령을 빠르게 여러 번 내리는 동작에 관한 두 변수 GapBetweenPACs와 ActionLatency는 등급과 음의 상관관계를 보인다. 등급이 올라갈수록 평균은 작아지고 분산은 작아지는 형태를 보인다.

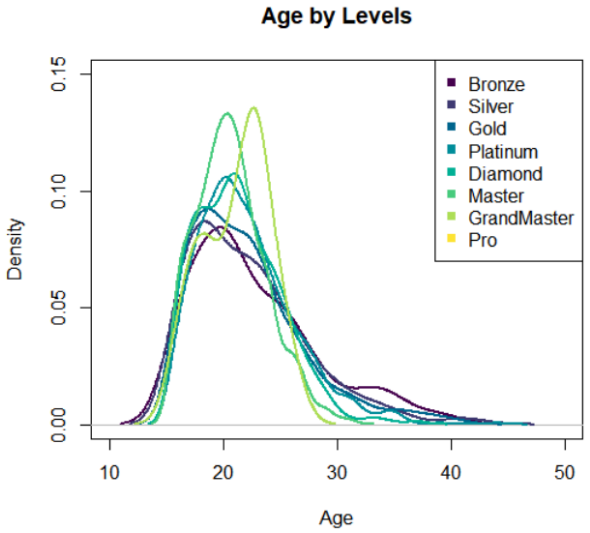
# 등급 비례 변수 – 3

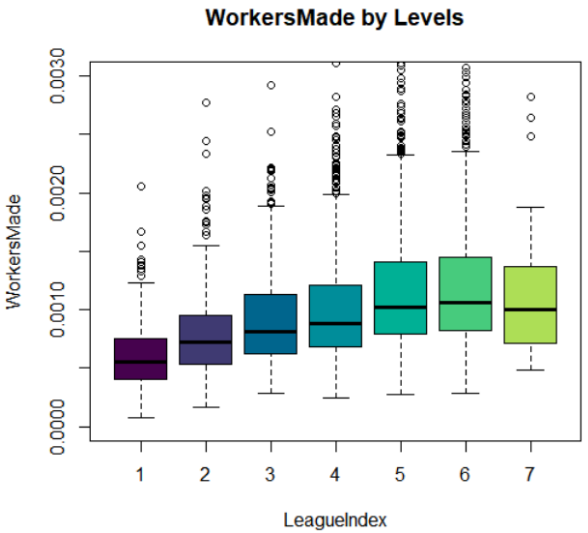



- 위 변수들은 레벨에 따라 그 값이 소폭 상승하는 추이를 보이나, 앞서 살펴본 변수들에 비해서는 레벨 간의 차이가 극명하지 않다. 레벨을 결정짓는 중요한 요인으로 작용하지는 않을 것이다.
- MinimapAttacks, MinimapRightClicks는 게임 화면 내 미니맵을 활용하는 정도를 나타낸다. 두 변수 모두 Boxplot의 이상치가 비교적 많은 것과 밀도 함수에서 모든 레벨 평균이 0과 0.0001 사이에 위치하며 분포가 서로 크게 다르지 않은 것으로 보아 뚜렷한 선형 관계를 찾기 어려울 것으로 생각된다.
- TotalHours는 등급이 높아질수록 그 값이 조금씩 증가하는 형태를 보인다.

# 등급 비례 변수 제외 예시

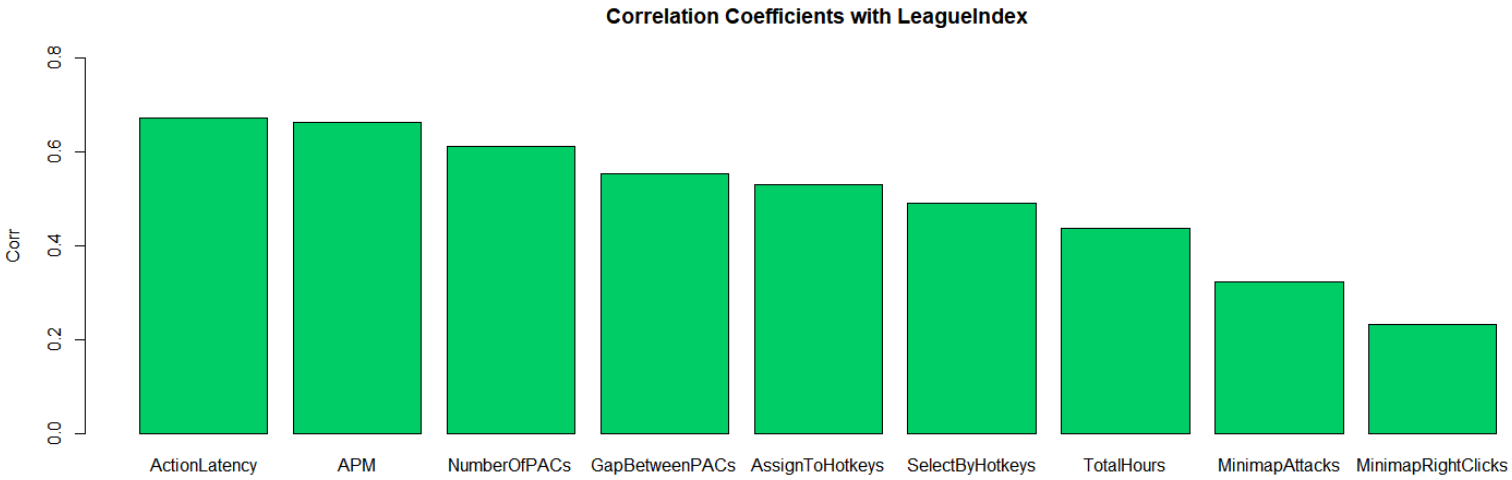



- 등급 비례 변수로 선정되지 못한 10개 변수 그래프의 일부다. UniqueUnitsMade, Age의 밀도 그래프와 같이 등급이 상승해도 그 평균과 분포가 비슷하거나, 세번째 WorkersMade와 같이 등급에 따라 상승하는 패턴을 보이다가 값이 갑자기 떨어지고 높아지는 모습을 보이는 변수는 전반적인 등급 상승의 주요 요인이라 하기에 불규칙하여 제외시켰다.

# 등급 비례 변수와 등급의 상관계수




- 상관계수는 두 변수 사이의 선형관계의 유무와 강도를 나타내는 측도다. 9개의 등급 비례 변수와 등급 사이의 상관계수를 각각 계산한 결과, ActionLatency > APM > NumberOfPACs > GapBetweenPAC > AssignToHotkeys > SelectByHotkeys > TotalHours > MinimapAttacks > MinimapRightClicks 순으로 상관관계가 높은 것으로 나타났다. 음의 상관관계를 보이는 변수가 있기 때문에, 상관계수에 절댓값을 적용하여 비교했다.
- 상관계수가 높은 상위 4개 변수는 모두 유저가 얼마나 빠른 시간 내에 많은 명령을 내리는지를 나타내는 변수들이다. 즉, 스타크래프트에서는 상황에 따라 명령을 즉시 생각해낼 수 있는 두뇌 회전과 빠른 손놀림이 등급 상승에 전반적으로 중요한 조건이라고 할 수 있다.
- 전체 등급에 비례하여 증가하는 변수만을 선택했기 때문에, 등급 별로 상승에 결정적인 영향을 미치는 변수는 알지 못하는 데에 한계가 있다. 이는 R에서 랜덤포레스트 학습 방법을 사용하면 구할 수 있다.