|  |  |
| --- | --- |
|  | 🞂R을 이용한 데이터 분석  Skill Craft탐색적 데이터 분석 |
|  |  |
| = | 🞂이현호 |

# 목차

[목차 - 1 -](#_Toc520149775)

[소개 - 2 -](#_Toc520149776)

[서론 - 2 -](#_Toc520149777)

[본론 - 3 -](#_Toc520149778)

[랜덤포레스트를 이용한 변수의 중요도 파악 - 3 -](#_Toc520149779)

[반응변수에 대한 설명변수의 분포 - 4 -](#_Toc520149780)

[결론 - 8 -](#_Toc520149781)

R을 이용한 데이터 분석

Skill Craft탐색적 데이터 분석

# 소개

오픈 소스 통계 소프트웨어 R로 빅데이터 통계 분석 및 시각화를 실제 데이터에 적용해봄으로써 한 눈에 빠르게 데이터의 분포 및 형태를 파악하고 데이터를 그룹화 해서 비교하고 이상값 밀 결측값을 빠르게 찾을 수 있습니다.

# 서론

Skill Craft는 게임을 하는 동안 유저에게서 생성되는 정보를 3395 \* 20.로 나눈 데이터셋입니다.

반응변수는 LeagueIndex로 설정하였고, 리그의 등급을 8개의 범주로 나눈 데이터입니다.

설명변수는 총 19개 이며, GameId는 유저들의 게임아이디 입니다. 이 아이디는 고유한 것을 알 수 있습니다.

Age는 유저들의 나이이고, HousPerWeek와 TotalHours는 유저들의 게임시간, APM은 유저들이 마우스 또는 키보드를 통해 입력한 분당 명령 수이며, SelectByHotKeys는 하나의 타임프레임에서 사용한 핫 키 사용 평균 횟수이며, AssignToHotkeys는 핫 키에 할당된 건물 또는 유닛 수이고, UniqueHotkeys는 유니크핫키의 사용 수, MinimapAttacks은 미니맵을 통한 공격횟수, MinimapRightClicks은 미니맵을 이용한 우클릭 횟수(공격이나 이동), NumberOfPACs와 GapBetweenPACs는 특정 유닛을 클릭한 후 명령을 완료하는 동작, ActionLatency은 PAC의 첫번째 동작에서 완료할 동작까지 걸리는 평균시간, ActionsInPAC는 PAC내에서 일어나는 작업의 수, TotalMapExplored는 총 맵의 정찰 범위, WorkersMade는 하나의 타임프레임 내에서 평균 일꾼 수, UniqueUnitsMade는 하나의 타임프레임 내에서 Unique한 유닛의 수(다양성), ComplexUnitsMade는 다루기 어려운 유닛의 평균 수, ComplexAbilitiesUsed는 다루기 어려운 능력입니다.

각 설명변수가 반응변수인 LeagueIndex에 어떠한 영향을 주는지 각 리그의 계급을 올리기 위해서 중요한 것은 어떤 것인지 알기 위해 탐색적 데이터 분석을 실시하겠습니다.

# 본론

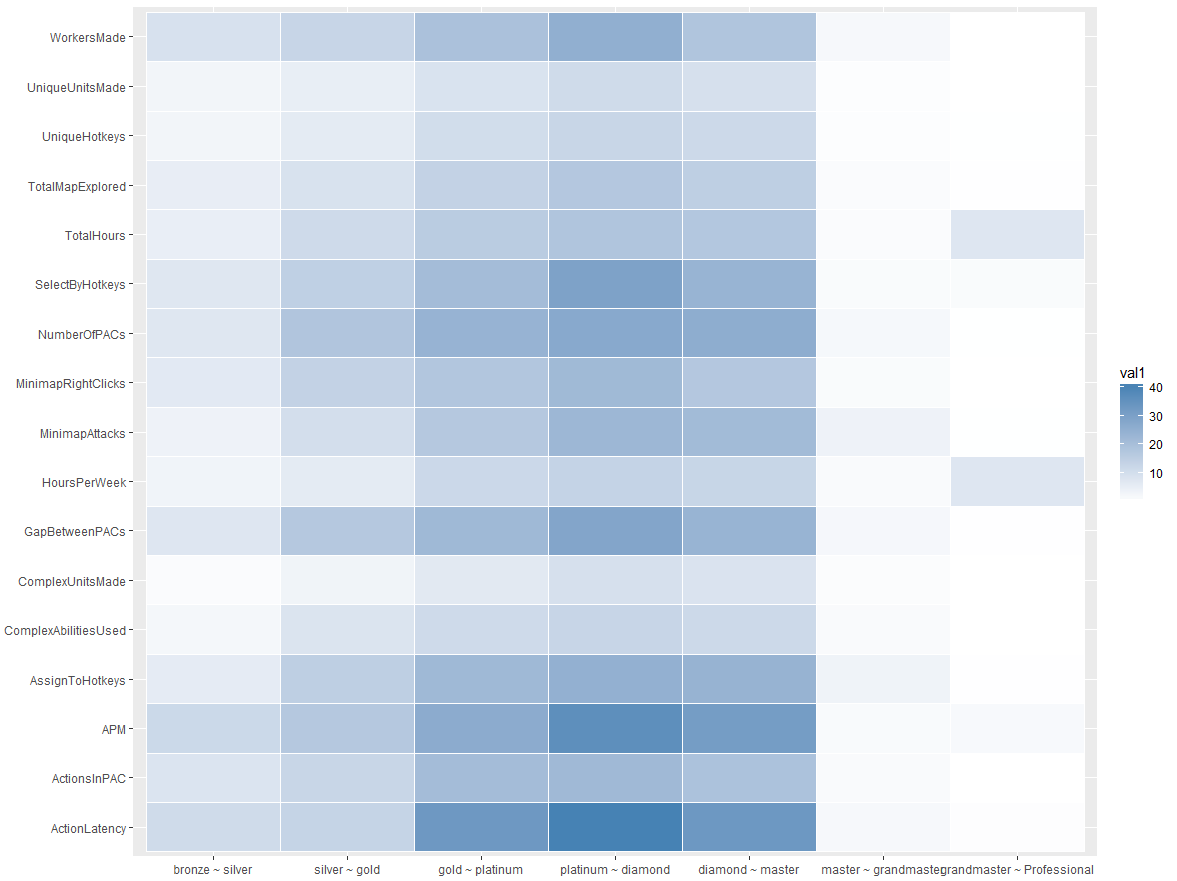
## 랜덤포레스트를 이용한 변수의 중요도 파악

각 티어간의 변수의 중요드를 알기 위해 랜덤포레스트를 이용하여 변수의 중요도를 시각화 하였습니다.

변수의 중요도를 보다 정확하게 도출하기 위해서 범주형 변수인 4개의 변수 중 GameId와 Age는 제거하고, HoursPerWeek와 TotalHours는 숫자형으로 변경하였습니다.

아래의 결과는 각각의 티어구간의 차이를 설명하는 변수들의 중요도를 시각화한 것입니다.

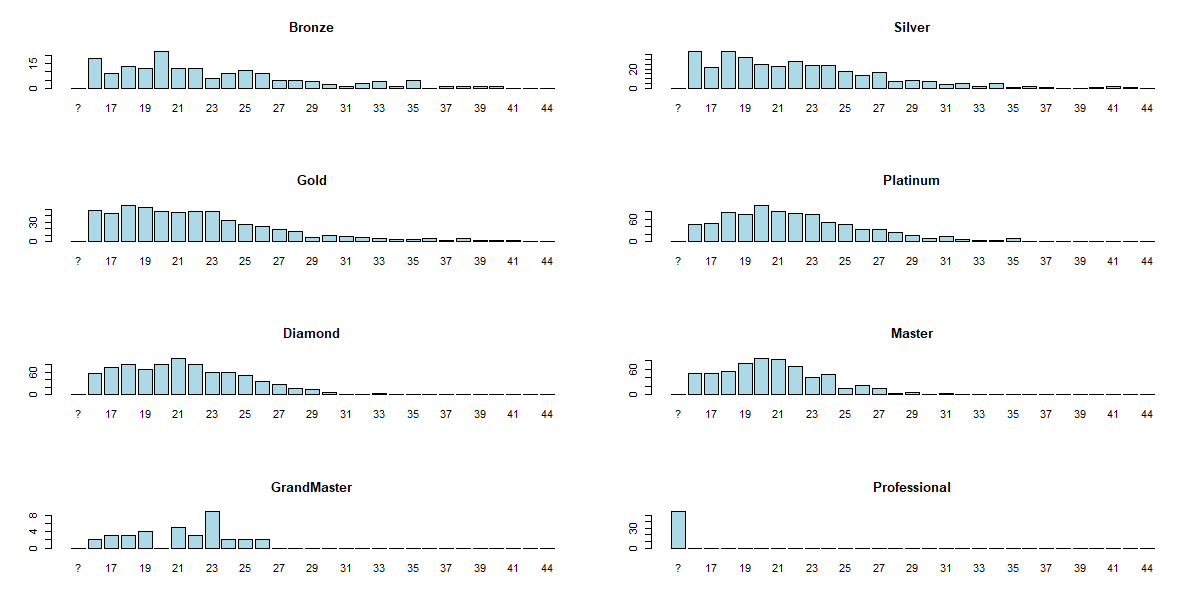
예를 들어 Bronze와 Silver의 차이를 크게 나타내는 것은 APM, ActionLatency임을 알 수 있습니다. Silver와 Gold는 경우 APM, GapBetweenPACs임을 알 수 있고, Gold와 Platinum은 APM, ActionLatency가 중요하다는 점을, 알 수 있고, Platinum과 Diamond도 마찬가지로 APM과 ActionLatency가 중요합니다.



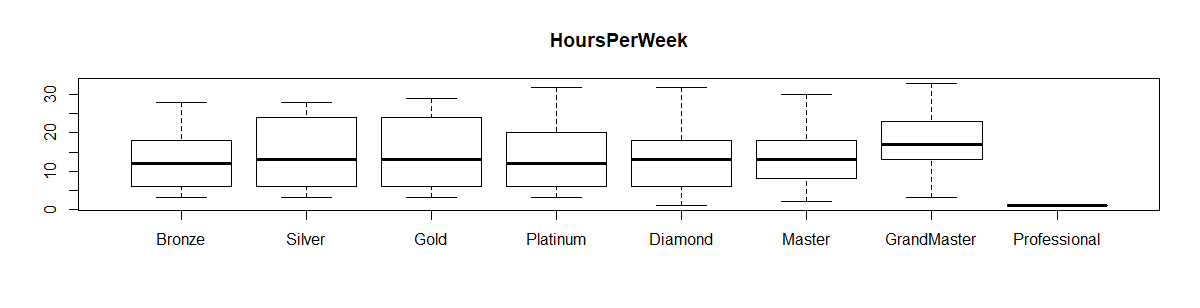
즉 브론즈에서 마스터까지는 APM과 GapBetweenPACs, 그리고 ActionLatency가 상당히 중요한 요소라는 점을 알 수 있습니다.

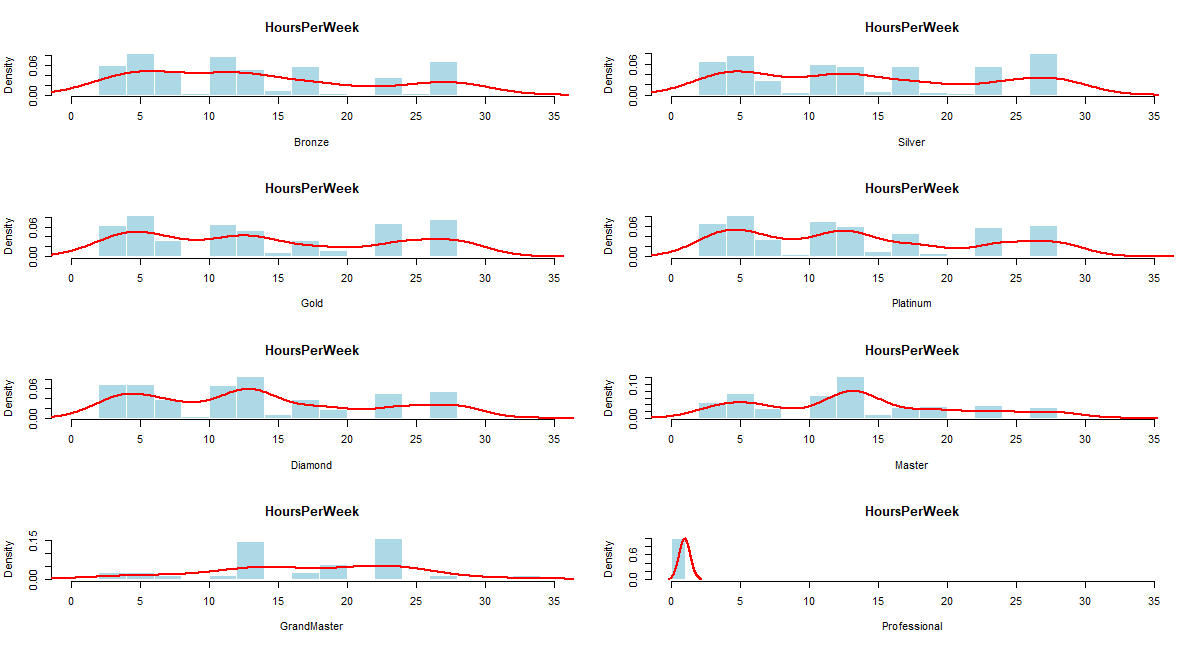
그러나 GrandMaster와 Professional의 경우에는 변수의 중요도가 매우 낮게 나타납니다. 또한 GrandMaster와 Professional의 차이가 게임시간이 가장 크다는 점은 Professional의 적은 데이터로 인해서 나타났다고 할 수 있고, GrandMaster와 Professional은 서로 비슷하다고 할 수 있습니다.

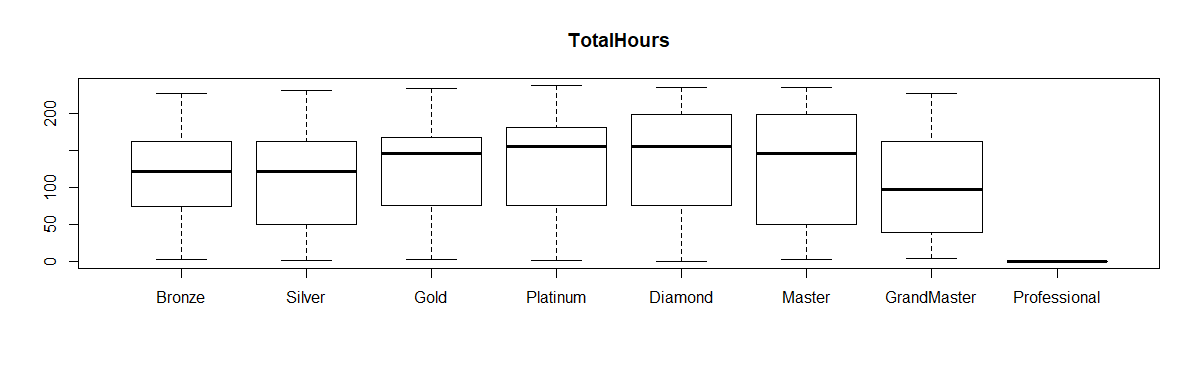
## 반응변수에 대한 설명변수의 분포

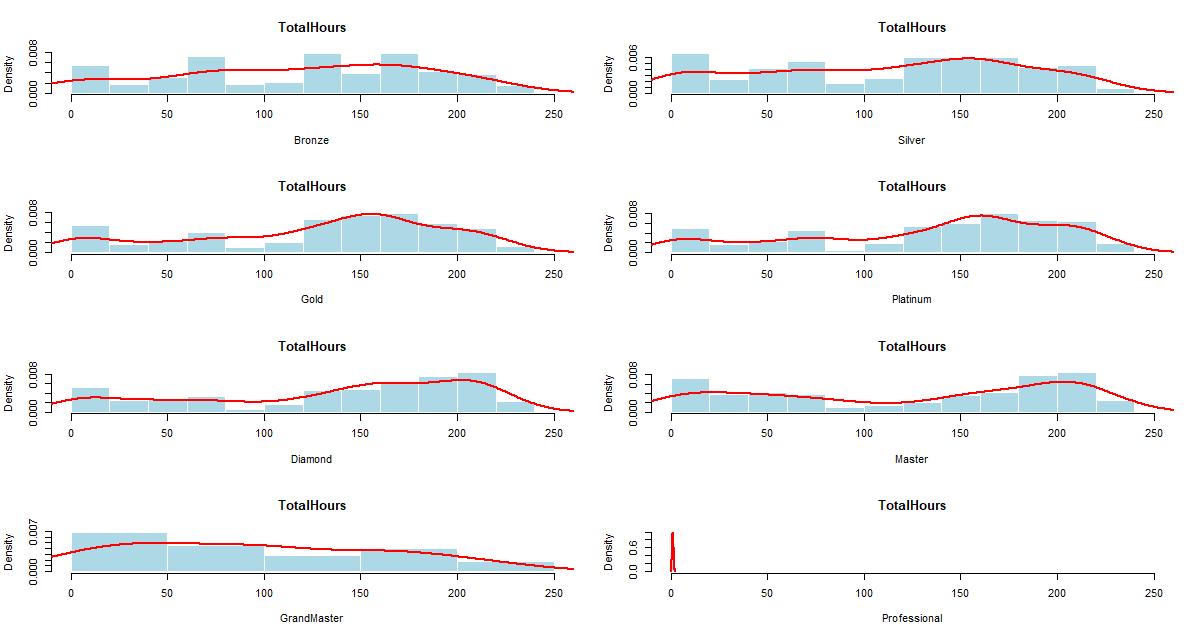


각 LeagueIndex에 대한 나이에 대한 분포를 시각화하였습니다. 분포가 상당히 비슷한 것으로 보아, 나이는 각 LeagueIndex에 영향을 주지 않는다는 점을 알 수 있습니다. Professional은 데이터가 없기 때문에 ‘?’로 표시가 되었습니다.



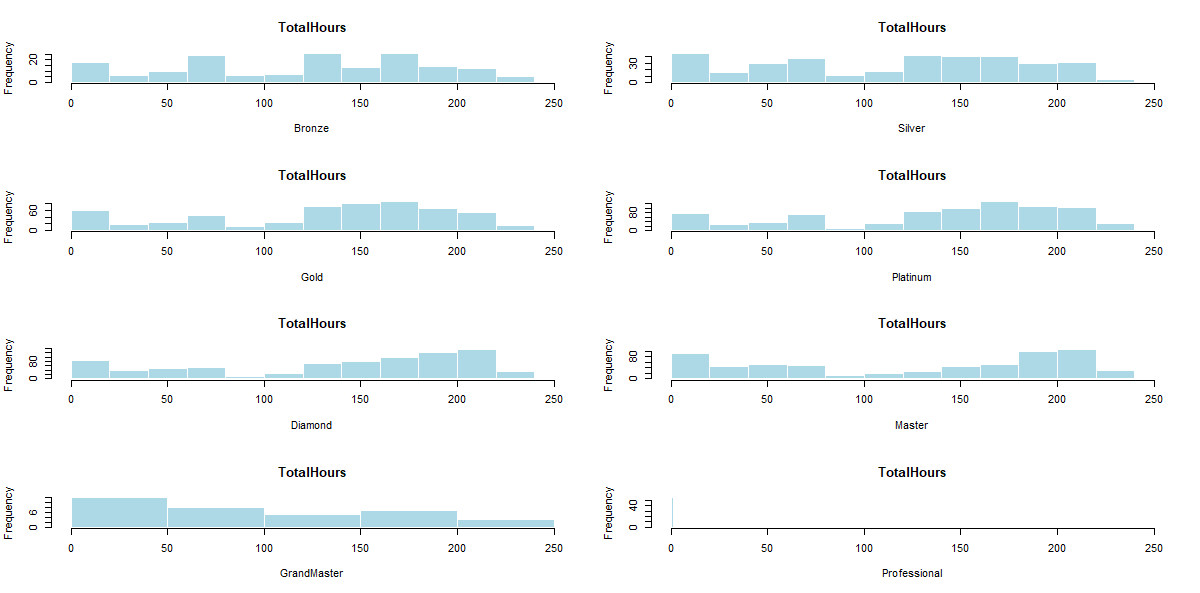




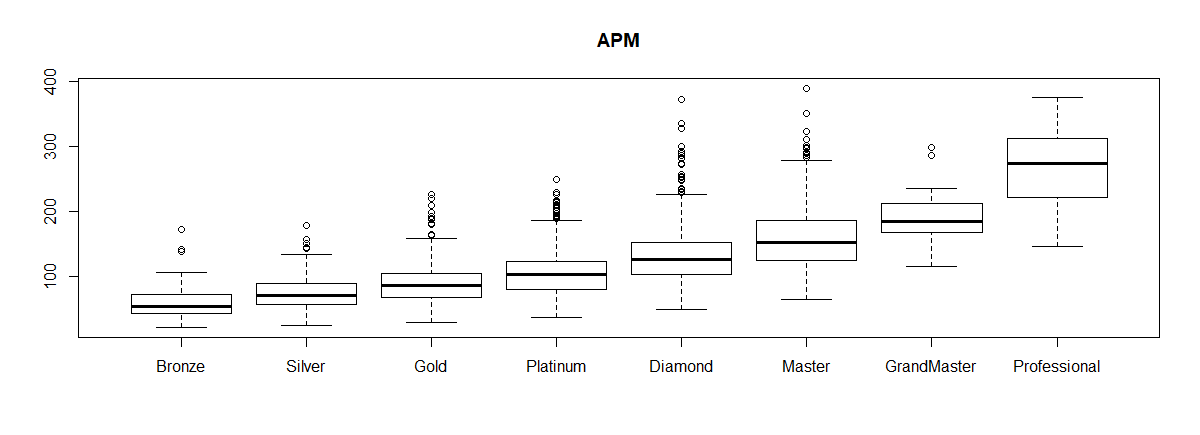


HoursPerWeek 와 TotalHours는 LeagueIndex에 관계없이 분포가 거의 비슷합니다. 게임플레이 시간은 티어에 영향을 주지 않는다고 판단할 수 있습니다.

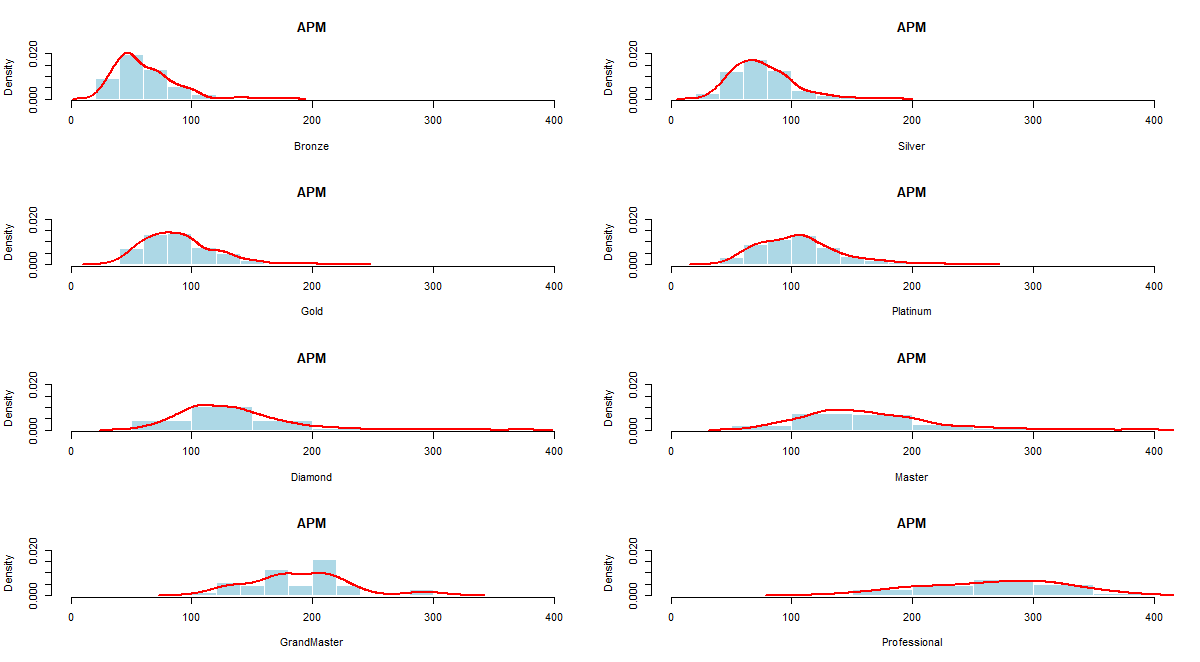
또한 분포의 차이가 큰 것이 확인이 되었고 위의 랜덤포레스트의 GrandMaster와 Professional의 변수의 중요도가 게임시간이 높게 나온 이유를 알 수 있습니다.



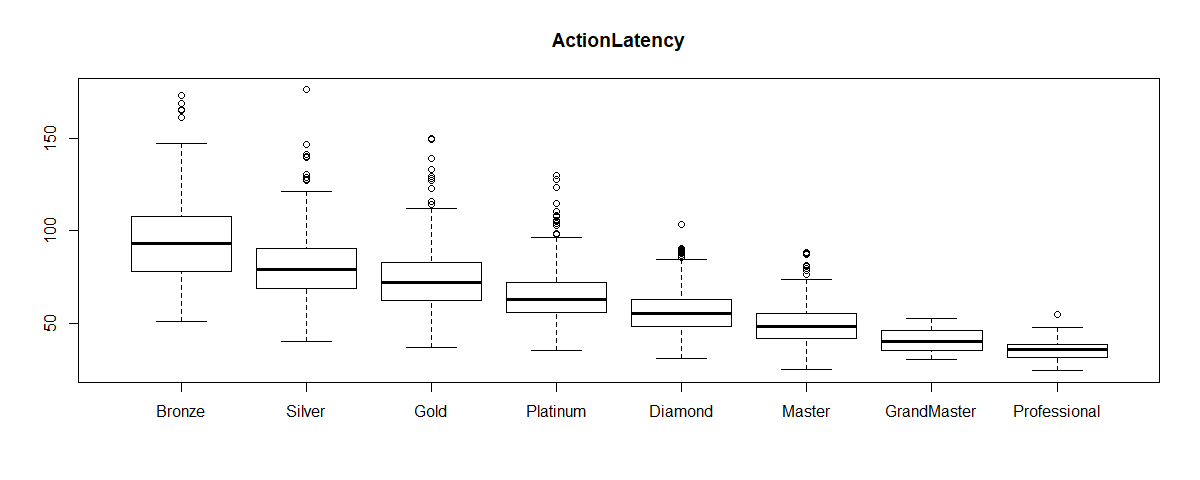
또한 상대빈도로 확인한 결과 게임을 많이 하는 등급은 Gold ~ Master구간이며 가장 많이하는 등급은 Diamond 구간임을 확인 할 수 있습니다.



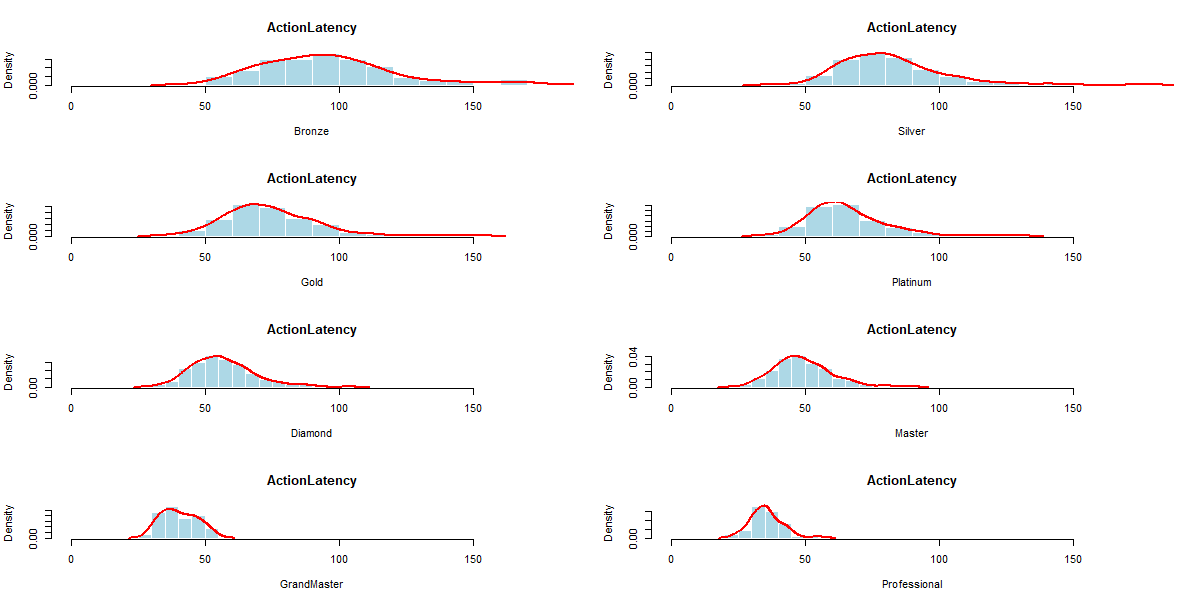
APM에 대한 boxplot을 보면 각 LeagueIndex간의 차이가 뚜렷한 것을 볼 수 있습니다. 랜덤포레스트의 변수의 중요도가 높게 나오는 이유를 알 수 있습니다.



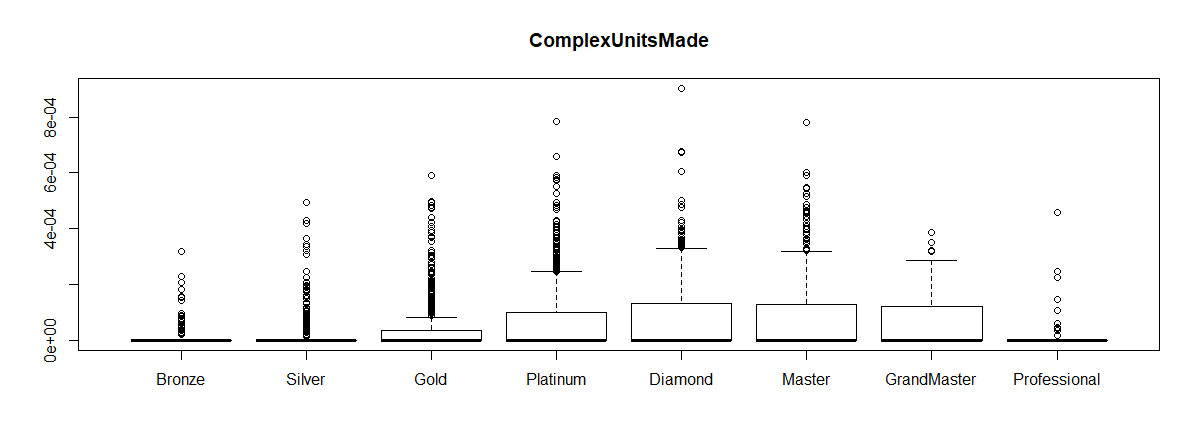
마찬가지로 분포또한 LeagueIndex 별로 중심이 점점 오른쪽으로 가는 것을 확인 할 수 있습니다.



마찬가지로 변수의 중요도가 높았던 ActionLatency의 경우에도 각 등급간의 boxplot이 확연하게 다른 것을 알 수 있습니다. 또한 분포의 경우에도 점점 왼쪽으로 이동하는 것으로 보아 상위등급으로 등급을 올리기 위해서는 ActionLatency를 점점 낮춰야 한다는 점을 볼 수 있습니다.

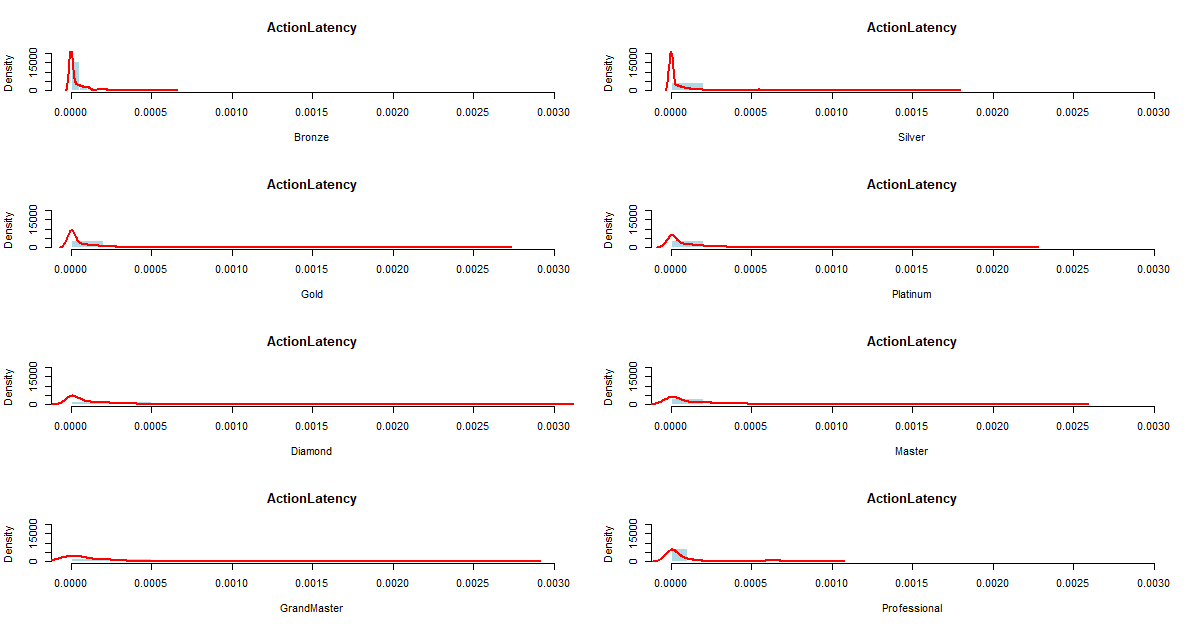


변수의 중요도가 모든 구간에서 가장 낮았던 변수는 ComplexUnitsMade임을 위의 랜덤포레스트 변수의 중요도 그림에서 알 수 있습니다. 더 정확하게 알아보기 위해 boxplot을 그려 보았습니다.



확연하게 모든 등급에서 비슷한 것을 알 수 있었습니다.

또한, 분포도 상당히 등급에 관계없이 유사함을 알 수 있습니다.



# 결론

각 등급의 실력을 객관적으로 나눌 수 있는 기준은 크게 3가지로 나뉨을 알 수 있었습니다.

APM, ActionLatency, GapBetweenPACs 이 3가지는 랜덤포레스트에서 변수의 중요도가 높은 수치로 나왔고, 각 분포와 boxplot도 유의한 차이를 보여주었습니다. 즉 각 등급에서 상위등급으로 진급하기 위해서는 위의 3가지를 훈련하는 것이 효율적이라고 판단 할 수 있습니다.

또한 영향이 없는 변수는 ComplexUnitsMade와 Age, 게임시간 등이 있었습니다. 다루기 어려운 유닛을 사용 하는 것은 등급에 상관없이 골고루 분포하는 것을 확인 할 수 있었고, 나이또한 등급별로 골고루 분포되어 있었습니다. 또한 게임시간은 각 시간대별로 골고루 분포되어 있었고 마찬가지로 등급에 영향을 미치지 않는 것을 확인 할 수 있었습니다.

또한, 랜덤포레스트의 결과와 반응변수와 설명변수의 각 관계를 탐색하는 결과가 상당히 유사함을 알 수 있었습니다. 각각의 모든 설명변수에 대해 직접적으로 접근하지 않아도 좋은 모델을 이용하여 원하는 결과를 도출 할 수 있다는 점을 알게 되었습니다.