**SkillCraft Data 탐색적 자료분석 – 김한범**

**🡪 SkillCraft Data를 통해 알고 싶은 것 :**

**등급(LeagueIndex)차이에 영향을 주는 변수들은 무엇인가?**

**(1) 데이터 구조 간략히 파악(str, summary)**

🡪 Age, HoursPerWeek, TotalHours 변수는 ‘?’마크 때문에 chr로 인식한다.

따라서, csv파일을 불러올 떄 na.strings를 통해 ‘?’마크를 na로 인식하게 한다.

str(skill)

'data.frame': 3395 obs. of 20 variables:

$ GameID : int 52 55 56 57 58 60 61 72 77 81 ...

$ LeagueIndex : int 5 5 4 3 3 2 1 7 4 4 ...

$ Age : int 27 23 30 19 32 27 21 17 20 18 ...

$ HoursPerWeek : int 10 10 10 20 10 6 8 42 14 24 ...

$ TotalHours : int 3000 5000 200 400 500 70 240 10000 2708 800 ...

$ APM : num 144 129 70 108 123 ...

$ SelectByHotkeys : num 0.00352 0.0033 0.0011 0.00103 0.00114 ...

$ AssignToHotkeys : num 0.00022 0.000259 0.000336 0.000213 0.000327 ...

$ UniqueHotkeys : int 7 4 4 1 2 2 6 6 2 8 ...

$ MinimapAttacks : num 1.10e-04 2.94e-04 2.94e-04 5.33e-05 0.00 ...

$ MinimapRightClicks : num 0.000392 0.000432 0.000461 0.000543 0.001329 ...

$ NumberOfPACs : num 0.00485 0.00431 0.00293 0.00378 0.00237 ...

$ GapBetweenPACs : num 32.7 32.9 44.6 29.2 22.7 ...

$ ActionLatency : num 40.9 42.3 75.4 53.7 62.1 ...

$ ActionsInPAC : num 4.75 4.84 4.04 4.92 9.37 ...

$ TotalMapExplored : int 28 22 22 19 15 16 15 45 29 27 ...

$ WorkersMade : num 0.001397 0.001193 0.000745 0.000426 0.001174 ...

$ UniqueUnitsMade : int 6 5 6 7 4 6 5 9 7 6 ...

$ ComplexUnitsMade : num 0 0 0 0 0 ...

$ ComplexAbilitiesUsed: num 0.00 2.08e-04 1.89e-04 3.84e-04 1.93e-05 ...

🡪 LeagueIndex가 int로 입력되어 있다. 이후 분석을 위해 LeagueIndex 변수를Factor로 바꾸어 주고 label을 붙이는 작업을 한다.

summary(skill)

GameID LeagueIndex Age HoursPerWeek TotalHours

Min. : 52 Platinum:811 Min. :16.00 Min. : 0.00 Min. : 3.0

1st Qu.: 2464 Diamond :806 1st Qu.:19.00 1st Qu.: 8.00 1st Qu.: 300.0

Median : 4874 Master :621 Median :21.00 Median : 12.00 Median : 500.0

Mean : 4805 Gold :553 Mean :21.65 Mean : 15.91 Mean : 960.4

3rd Qu.: 7108 Silver :347 3rd Qu.:24.00 3rd Qu.: 20.00 3rd Qu.: 800.0

Max. :10095 Bronze :167 Max. :44.00 Max. :168.00 Max. :1000000.0

(Other) : 90 NA's :55 NA's :56 NA's :57

APM SelectByHotkeys AssignToHotkeys UniqueHotkeys

Min. : 22.06 Min. :0.000000 Min. :0.0000000 Min. : 0.000

1st Qu.: 79.90 1st Qu.:0.001258 1st Qu.:0.0002042 1st Qu.: 3.000

Median :108.01 Median :0.002500 Median :0.0003526 Median : 4.000

Mean :117.05 Mean :0.004299 Mean :0.0003736 Mean : 4.365

3rd Qu.:142.79 3rd Qu.:0.005133 3rd Qu.:0.0004988 3rd Qu.: 6.000

Max. :389.83 Max. :0.043088 Max. :0.0017522 Max. :10.000

MinimapAttacks MinimapRightClicks NumberOfPACs GapBetweenPACs

Min. :0.000e+00 Min. :0.0000000 Min. :0.000679 Min. : 6.667

1st Qu.:0.000e+00 1st Qu.:0.0001401 1st Qu.:0.002754 1st Qu.: 28.958

Median :3.993e-05 Median :0.0002815 Median :0.003395 Median : 36.724

Mean :9.831e-05 Mean :0.0003874 Mean :0.003463 Mean : 40.362

3rd Qu.:1.189e-04 3rd Qu.:0.0005141 3rd Qu.:0.004027 3rd Qu.: 48.291

Max. :3.019e-03 Max. :0.0040408 Max. :0.007971 Max. :237.143

ActionLatency ActionsInPAC TotalMapExplored WorkersMade UniqueUnitsMade

Min. : 24.09 Min. : 2.039 Min. : 5.00 Min. :7.698e-05 Min. : 2.000

1st Qu.: 50.45 1st Qu.: 4.273 1st Qu.:17.00 1st Qu.:6.830e-04 1st Qu.: 5.000

Median : 60.93 Median : 5.096 Median :22.00 Median :9.052e-04 Median : 6.000

Mean : 63.74 Mean : 5.273 Mean :22.13 Mean :1.032e-03 Mean : 6.534

3rd Qu.: 73.68 3rd Qu.: 6.034 3rd Qu.:27.00 3rd Qu.:1.259e-03 3rd Qu.: 8.000

Max. :176.37 Max. :18.558 Max. :58.00 Max. :5.149e-03 Max. :13.000

ComplexUnitsMade ComplexAbilitiesUsed

Min. :0.000e+00 Min. :0.000e+00

1st Qu.:0.000e+00 1st Qu.:0.000e+00

Median :0.000e+00 Median :2.028e-05

Mean :5.943e-05 Mean :1.419e-04

3rd Qu.:8.554e-05 3rd Qu.:1.814e-04

Max. :9.023e-04 Max. :3.084e-03

🡪 3395개의 관측값, 20개의 변수들로 구성되어 있다.

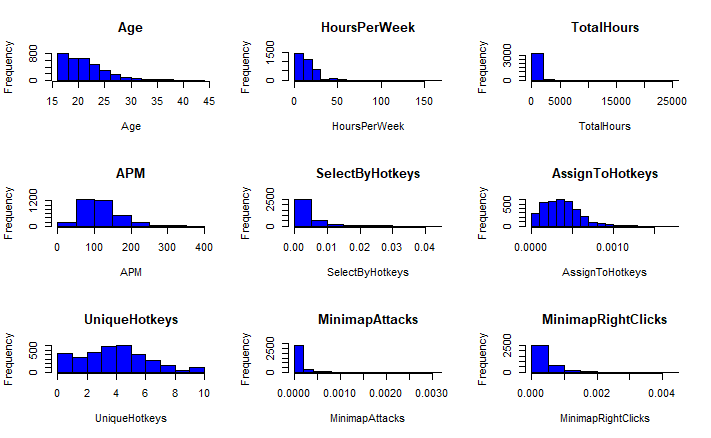
🡪 NA가 들어 있는 변수들 : Age(55), HoursPerWeek(56), TotalHours(57)

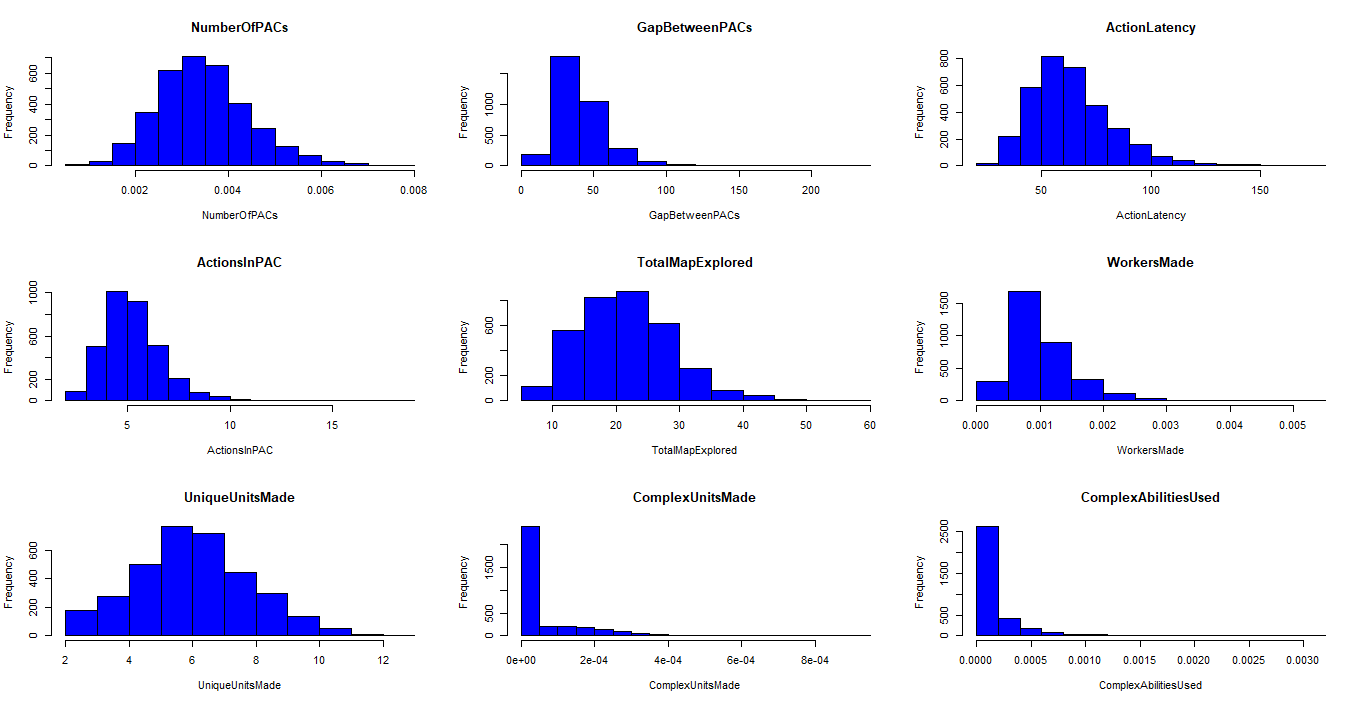
🡪 Outlier가 들어 있는 변수들 : TotalHours(1000000.0)

🡪 NA값은 제거하면 profession이 모두 제거된다.

따라서 Outlier값만 평균값으로 대체하고 탐색적 자료 분석을 진행한다.

**(2) 18개 변수들의 히스토그램(hist)**

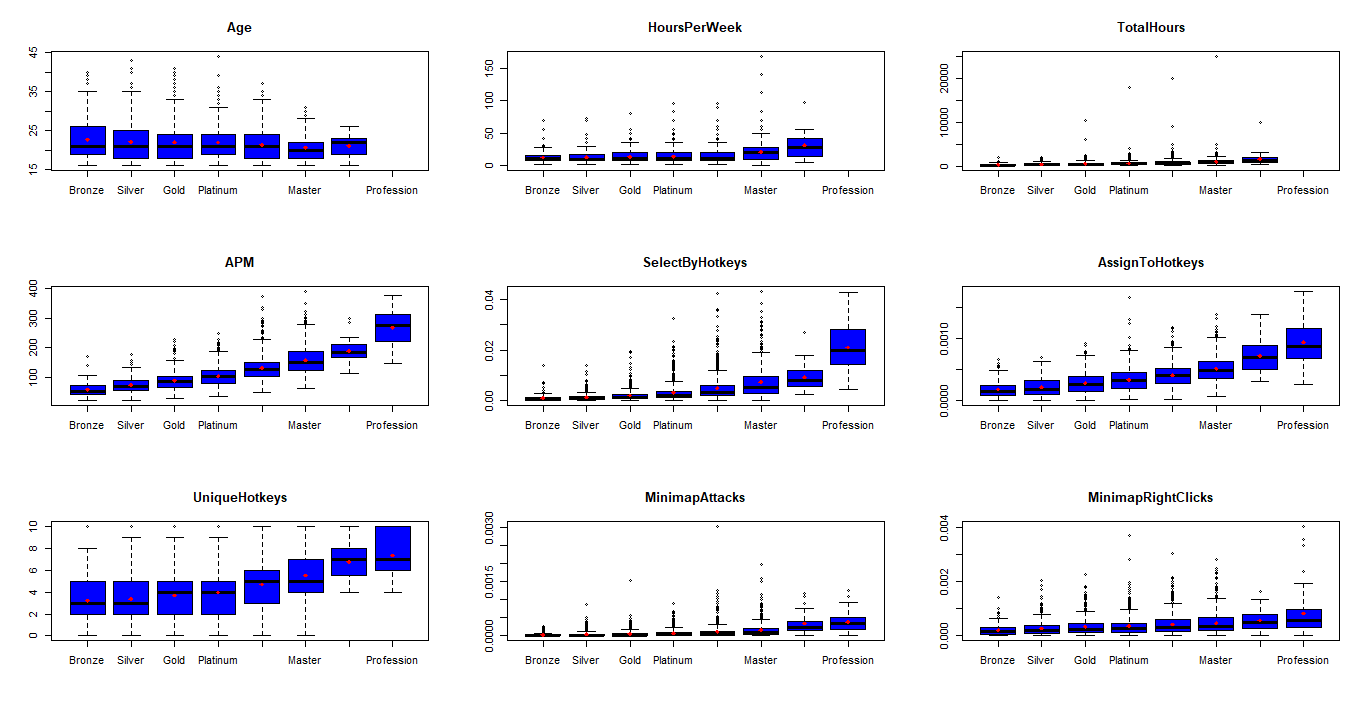
🡪 각 변수들의 분포 모양을 살펴보기 위해 히스토그램을 그려본다.

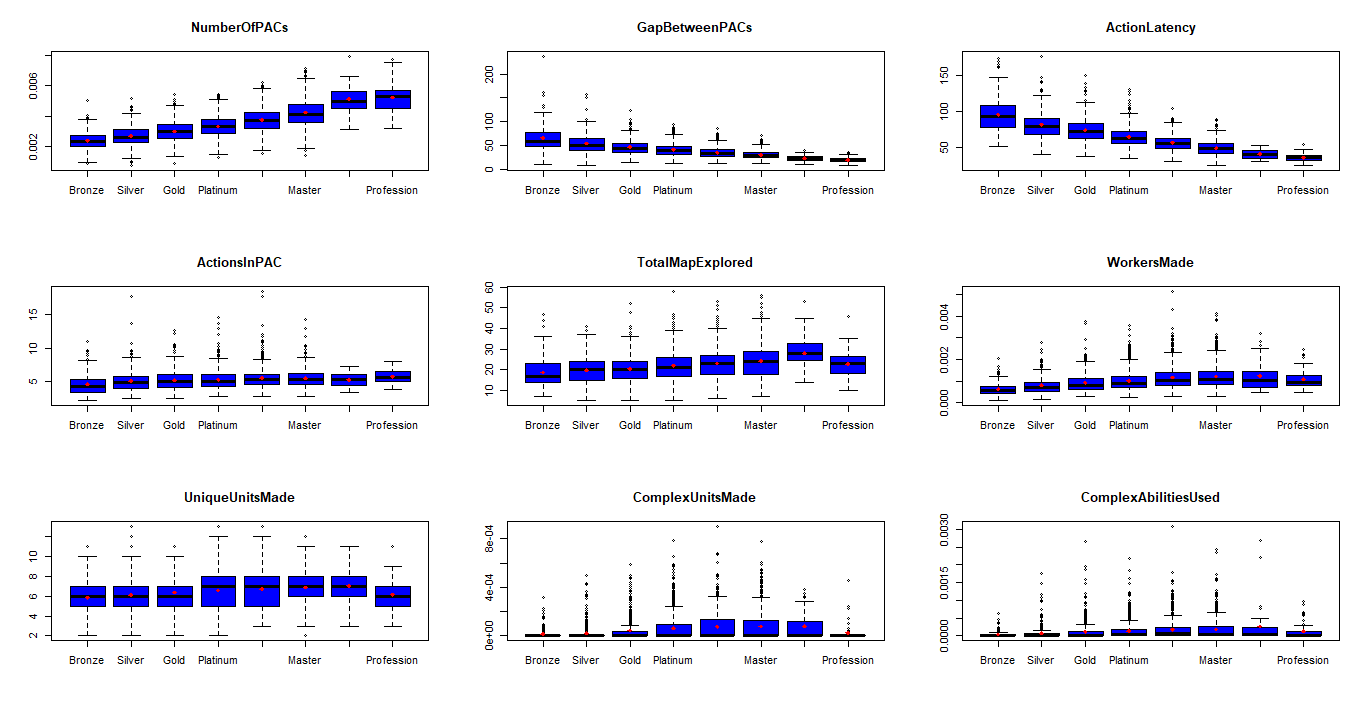
****

**🡪** Positively Skewed된 분포 형태가 많이 보인다.

🡪 x와 y의 scale이나 계급구간을 맞추어 주어 더 확연한 히스토그램을 그릴 필요가 있어 보인다

**(3) LeagueIndex에 따른 18개 변수들의 boxplot (boxplot)**

🡪 LeagueIndex를 기준으로 하는 boxplot을 그려본다

--> LeagueIndex에 따른 평균값차이가 확연히 드러나는 변수들은 다음과 같다.

(APM / Selectbyhotkeys / AssigntoHotkeys / NumberofPacs / ActionLatency)

🡪 APM, Selectbyhotkeys, UniqueHotkeys의 경우 profession에서 분산값이 크다는 것을 알 수 있다

🡪 outlier로 보이는 관측값들이 변수들마다 존재하지만, scale에 의한 영향도 고려해야 되므로

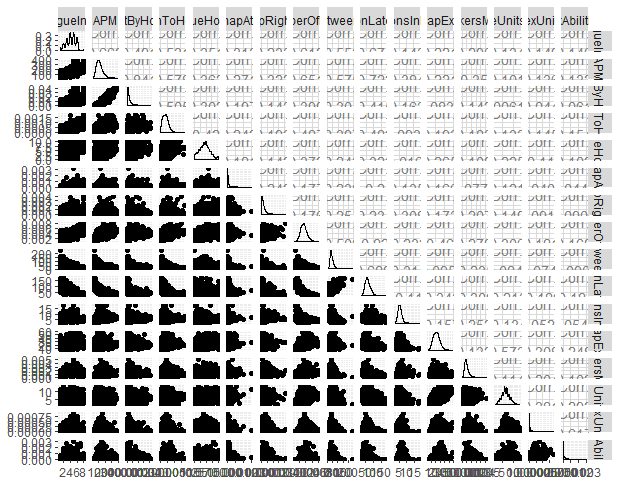
Outlier로 보이는 값들에 대한 처리는 더 고민해봐야 한다.

**(4) LeagueIndex에 따른 18개 변수들의 산점도와 상관계수(ggpairs, cor, corrplot)**

🡪 Age, HoursPerWeek, TotalHours 변수들은 Professional의 값이 없기에 산점도, 상관계수 분석에선 제외한다.

🡪 Professional의 상관관계를 구하기 위해 LeagueIndex를 interger로 코딩한다.

🡪 LeagueIndex를 기준으로 다른 변수들과의 상관관계 값을 완전히 신뢰할 수는 없다. LeagueIndex 간의 구간 차이를 동일하다고 가정하고 계산한 값이기 때문이다. 하지만 대강의 관계 파악을 위해 분석해보기로 한다.



🡪 각각의 산점도가 잘 보이지 않는다. 수치로 확인해 봐야겠다.

LeagueIndex[,1]

LeagueIndex 1.0000000 APM 0.6623089 SelectByHotkeys 0.4905987

AssignToHotkeys 0.5311251 UniqueHotkeys 0.3512135 MinimapAttacks 0.3125443

MinimapRightClicks 0.2332314 NumberOfPACs 0.6121925 GapBetweenPACs -0.5532456

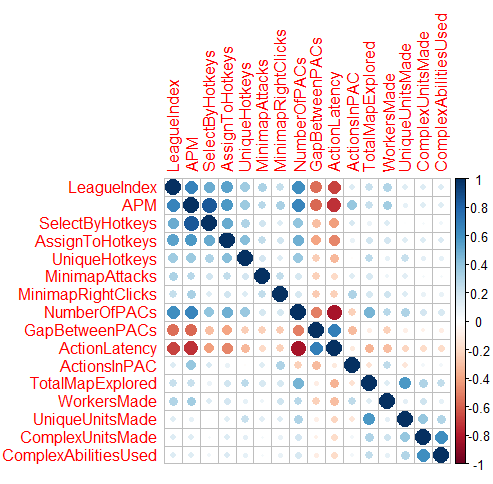
ActionLatency -0.6739388 ActionsInPAC 0.1431214 TotalMapExplored 0.2206764

WorkersMade 0.2956427 UniqueUnitsMade 0.1335075 ComplexUnitsMade 0.1488341

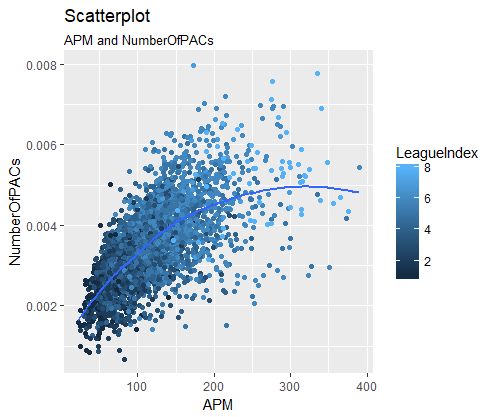
ComplexAbilitiesUsed 0.1457330

🡪 주요 양의 상관관계 변수들 : APM, SelectByHotkeys, AssignToHotkeys, NumberOfPACs

주요 음의 성관관계 변수들 : GapBetweenPACs, ActionLatency



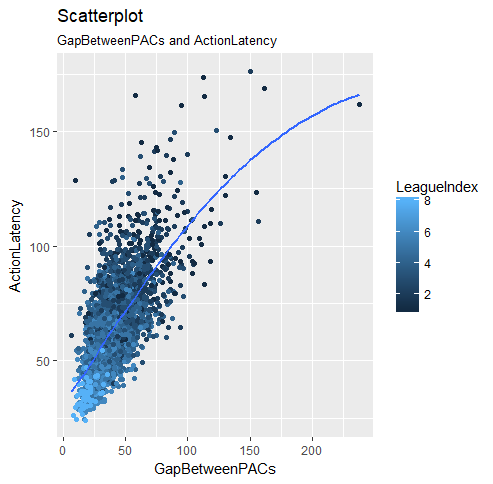
🡪 변수들간의 상관관계가 존재함을 알 수 있다.

 **(5) APM, NumberOfPACs, GapBetweenPACs, ActionLatency 변수 추가 분석(ggplot)**

🡪 양의 상관관계를 보인 두 변수와 음의 상관관계를 보인 두 변수에 대해 더 알아보고자 한다.

🡪 두 값은 양의 상관관계를 갖고 있고 두 값이 높을수록 LeagueIndex도 높음을 알 수 있다

🡪 그러나, NumberOfPACS은 APM이 어느 정도 높아지고 나선 그 영향력이 줄어듬을 알 수 있다.

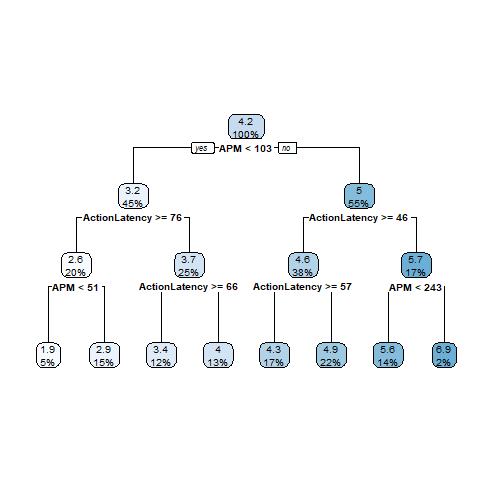


🡪 음의 상관관계의 두 변수들은 보다 경향성이 확연하다. 두 변수의 값이 동시에 작을수록 높은 수준의 Player에 속한다.

**(6) SkillCraft Decision Tree 추가 분석(rpart)**

🡪 작성자는 스타크래프트 게임을 접한지 19년째다. 그래서 이 과제와는 별개로 스타크래프트 실력에 영향을 주는 요인들에 대해 생각해 본적이 많았다. 작성자는 속칭 Physical이란 말로 불리는 손속도(APM)와 각 상황에서의 판단능력이 스타크래프트 실력에 가장 큰 영향력을 준다고 생각했다.

이 생각이 적절한지에 대해 주어진 과제자료를 통해 조금이나마 다가가고 싶었다.



🡪 Decision Tree 를 그려본 결과는 위와 같다. APM이 높고, ActionLatency가 낮을수록 스타크래프트 실력 등급이 높음을 확인할 수 있다. APM과 실력이 비례관계인 것은 당연한 얘기인 것 같고, ActionLateny는 상황에서의 판단 빠르기를 볼 수 있는 측정 지표라면, 작성자가 기존에 생각했던 것과 얼추 들어 맞는 것 같다.

**Wine Quality Data 탐색적 자료분석 – 김한범**

**🡪 Wine Quality Data를 통해 알고 싶은 것 :**

**등급(quality)차이에 영향을 주는 변수들은 무엇인가?**

**1. Red Wine**

**(1) 데이터 구조 간략히 파악(str)**

'data.frame': 1599 obs. of 12 variables:

$ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...

$ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...

$ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...

$ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...

$ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...

$ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...

$ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...

$ density : num 0.998 0.997 0.997 0.998 0.998 ...

$ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...

$ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...

$ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...

$ quality : int 5 5 5 6 5 5 5 7 7 5 ...

-> 1599개의 관측치, 12개의 변수로 이루어져 있다. Is.na로 확인한 결과 결측치는 없다.

-> 변수 12개 중 quality를 제외한 11개 변수는 연속형 변수.

**(2) Quality 변수의 빈도 파악 (table, barplot 이용)**

> table(wine\_r$quality)

3 4 5 6 7 8

10 53 681 638 199 18



summary(wine\_r$quality)

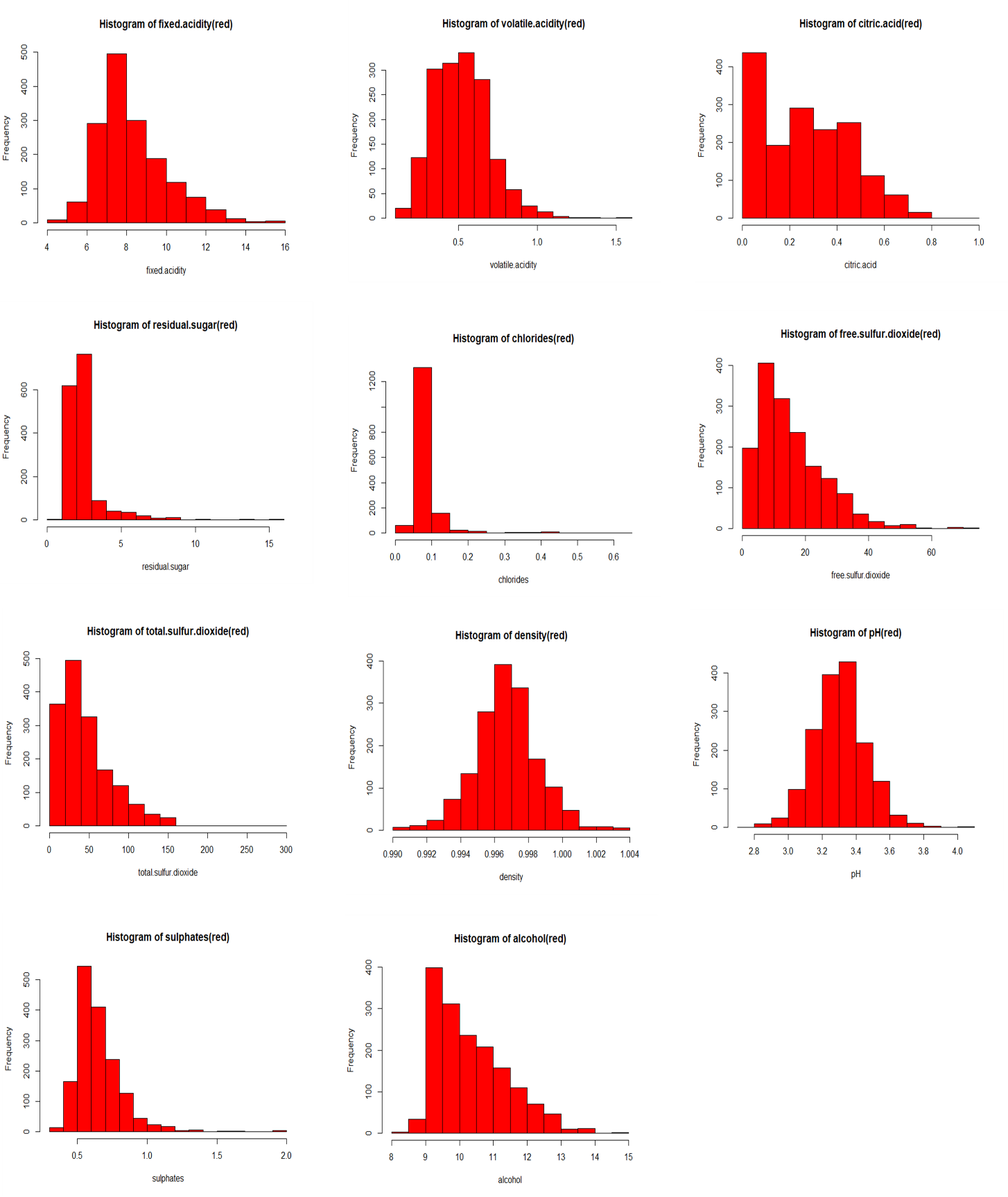
Min. 1st Qu. Median Mean 3rd Qu. Max.

3.000 5.000 6.000 5.636 6.000 8.000

🡪 quality 5,6이 전체의 82.4% 차지함을 알 수 있다.

**(3) quality에 따른 11개 변수들의 기술 통계 summary 및 히스토그램 (dplyr, hist)**

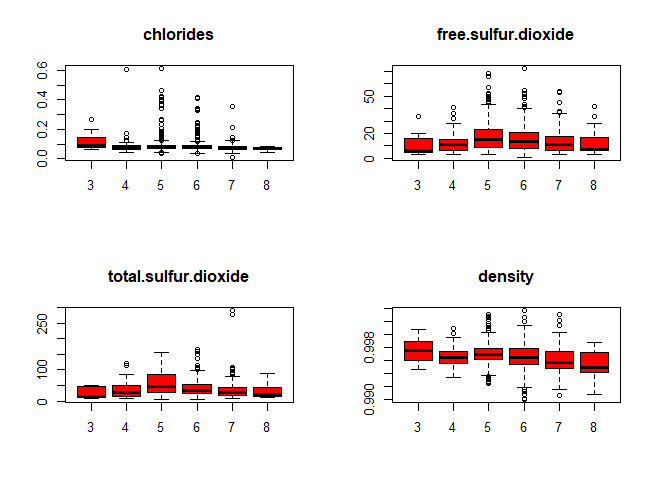
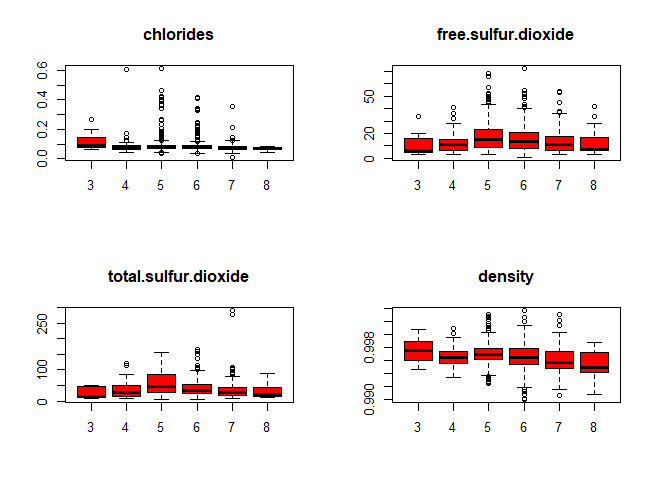
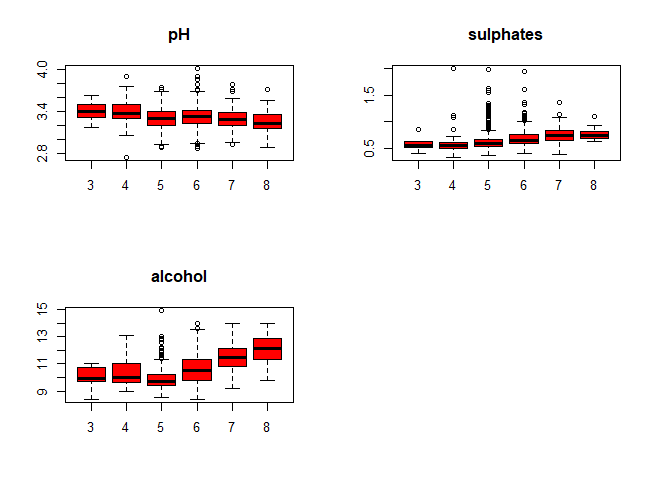
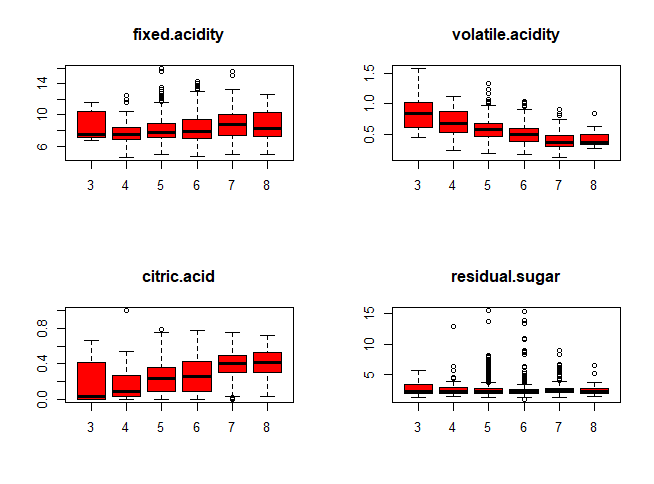


****

🡪 롱테일 분포를 보이는 변수들과 skewed된 분포가 많다.

🡪 scale 변환이 필요해 보이는 변수들이 있는 것 같다.(residual.sugar, chlorides)

**(4) quality에 따른 11개 변수들의 boxplot (boxplot)**



🡪 boxplot을 통해 아웃라이어 값들을 확인할 수 있다. 하지만 scale에 따라 다를 수 있으니 outlier처리에 관해선 보류한다.

🡪 boxplot을 통해서도 quality와 관계가 있어 보이는 변수들을 눈으로 확인 할 수 있다(volatile.acidity, alcohol)

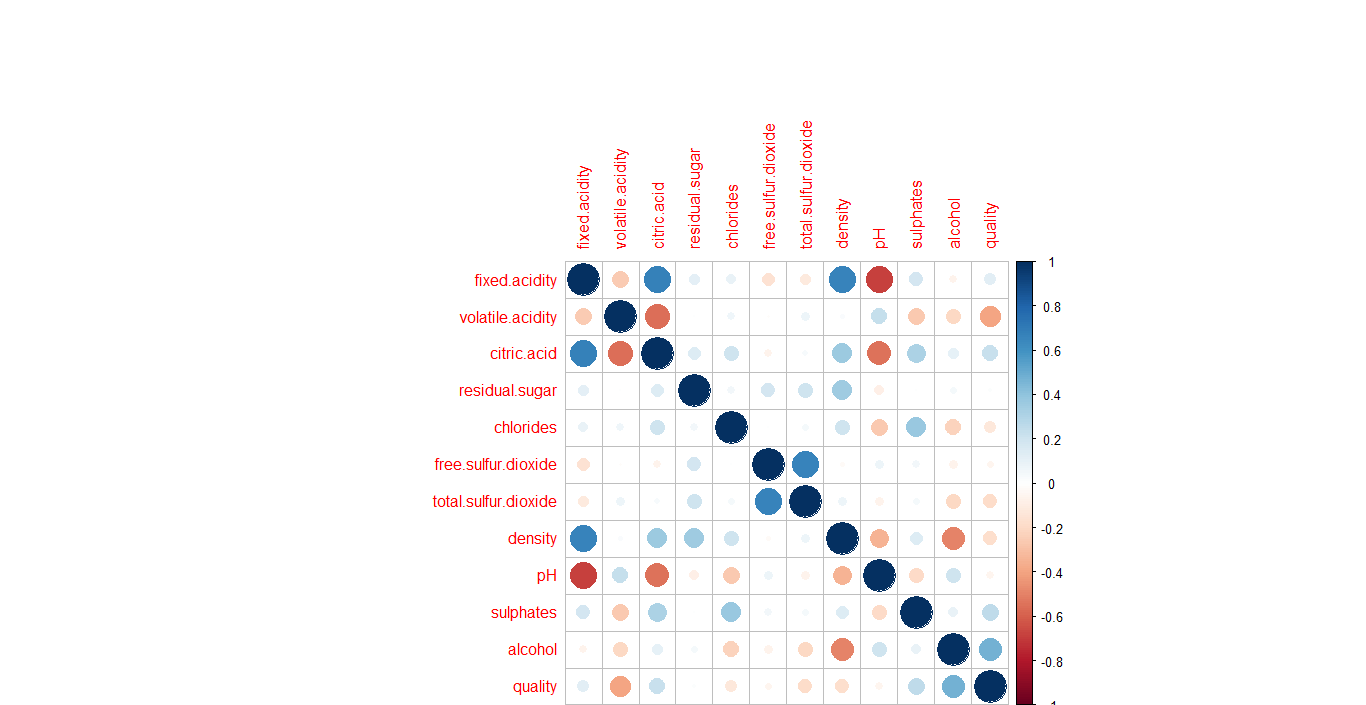
**(5) quality에 따른 11개 변수들의 산점도와 상관계수(cor, corrplot)**

quality[,1]

fixed.acidity 0.12405165 volatile.acidity -0.39055778 citric.acid 0.22637251

residual.sugar 0.01373164 chlorides -0.12890656 free.sulfur.dioxide -0.05065606

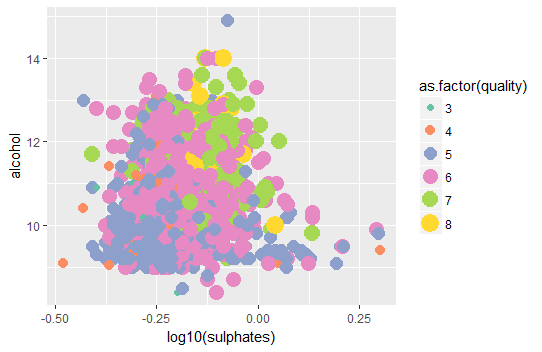
total.sulfur.dioxide -0.18510029 density -0.17491923 pH -0.05773139

sulphates 0.25139708 alcohol 0.47616632

🡪 quality에 양의 상관관계를 보이는 변수는 alcohol, sulphates 음의 상관관계를 보이는 것은 volatile.acidity라는 것을 알 수 있다.

🡪 quality 이외의 다른 변수들의 상관관계를 보면 (density, fixed.acidity), (citric.acid, fixed.acidity),음의 상관관계는 (fixed.acidity, pH), (volatile.acidity, citric.acid) 가 있다.

**(6) alcohol, sulphates, volatile.acidity 변수 추가 분석(ggplot, )**



🡪 alcohol과 sulphates과 동시에 높은 관찰값들은 quality의 값도 높음을 알 수 있다

🡪 반면, sulphates 값이 높고 alcohol 값이 낮다면 quality가 높다고 할 수 없다.

**2. White WIne**

**(1) 데이터 구조 간략히 파악(str)**

str(wine\_w)

'data.frame': 4898 obs. of 12 variables:

$ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...

$ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...

$ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...

$ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...

$ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...

$ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...

$ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...

$ density : num 1.001 0.994 0.995 0.996 0.996 ...

$ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...

$ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...

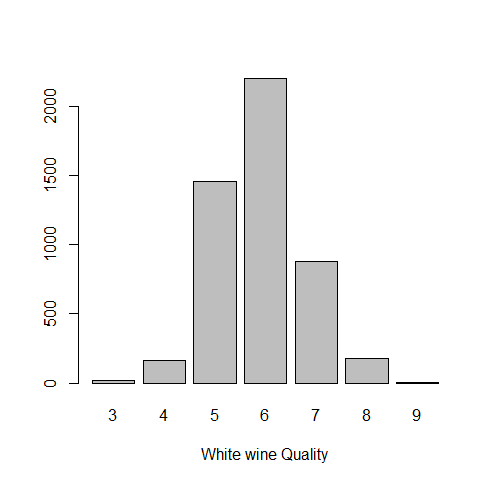
$ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...

$ quality : int 6 6 6 6 6 6 6 6 6 6 ...

🡪 4898개의 관측값들, 12개의 변수임을 알 수 있다.

🡪 quality 변수가 integer임을 알 수 있다

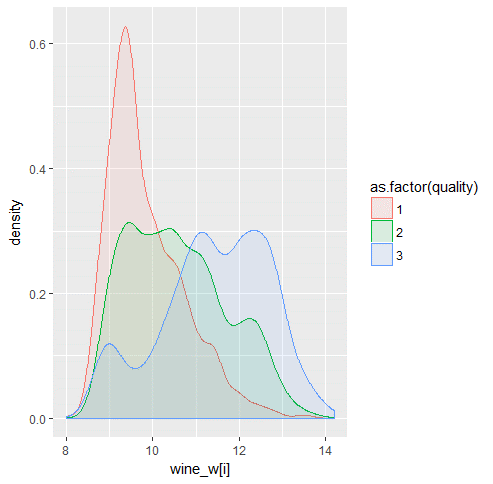
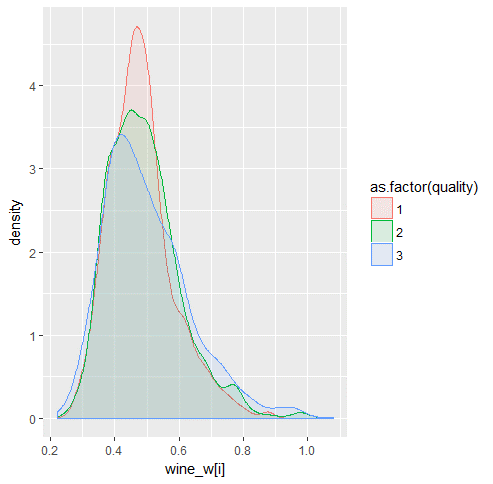
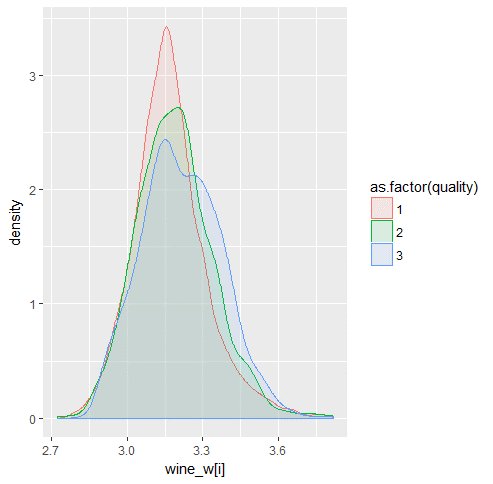
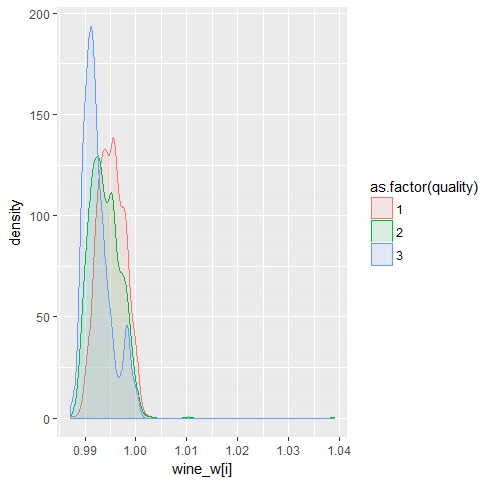
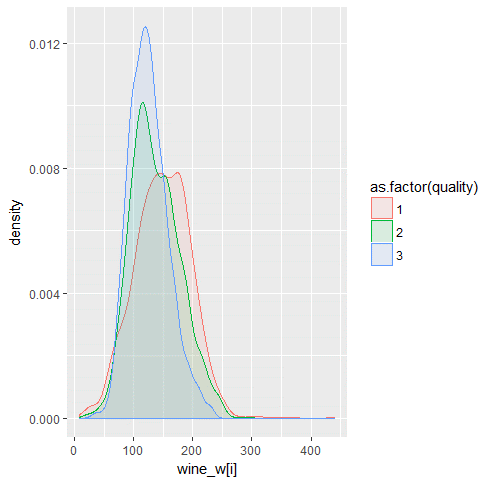
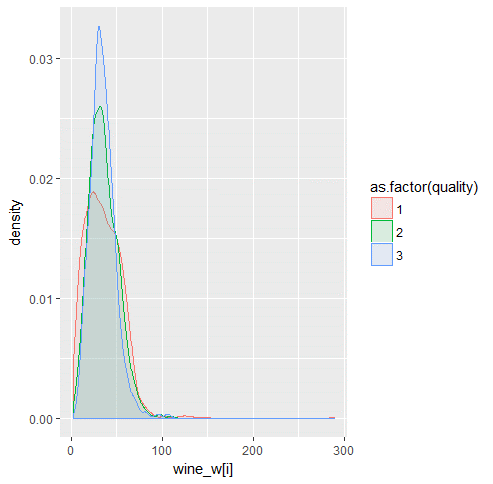
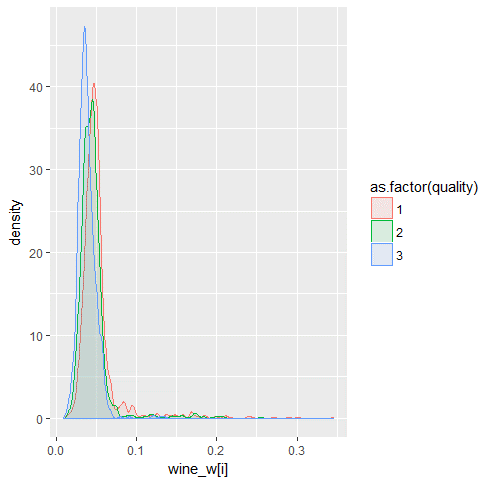
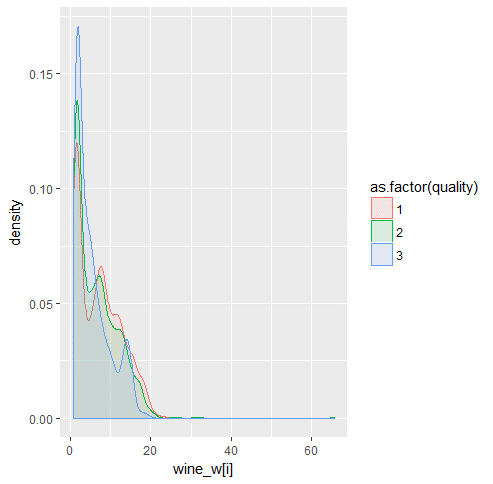
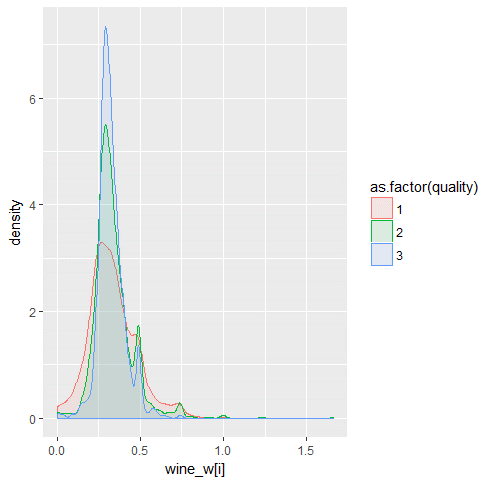
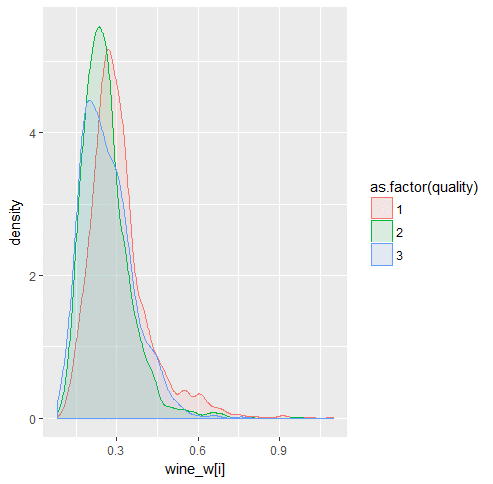
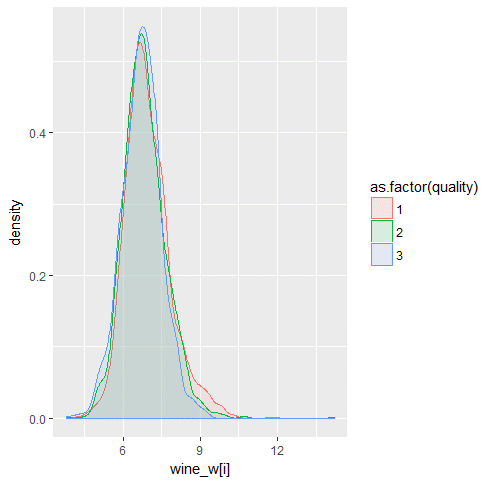
**(2) quality 변수의 빈도 파악 및 새 변수 생성(barplot)**



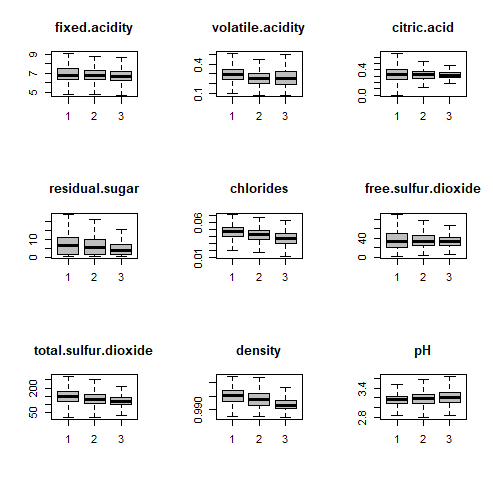
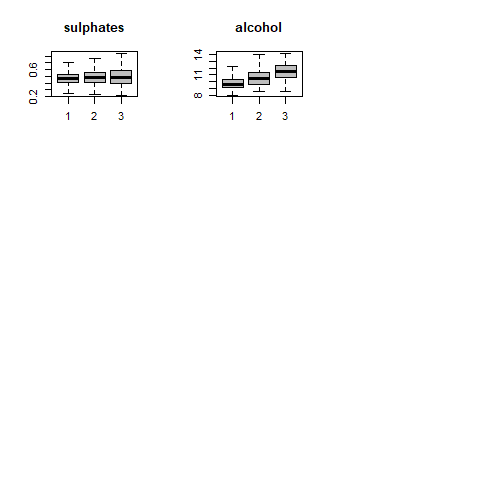
**🡪** 3,4와 8,9 값을 가진 변수들이 너무 적음을 알 수 있다. 따라서 quality 변수를 조정해주고자 한다.

🡪 3,4,5 = bad(1), 6=medium(2), 7,8,9 = good(3) 다시 변수를 조정해주었다.

**(3) quality에 따른 11개 변수들의 density 히스토그램 (ggplot)**

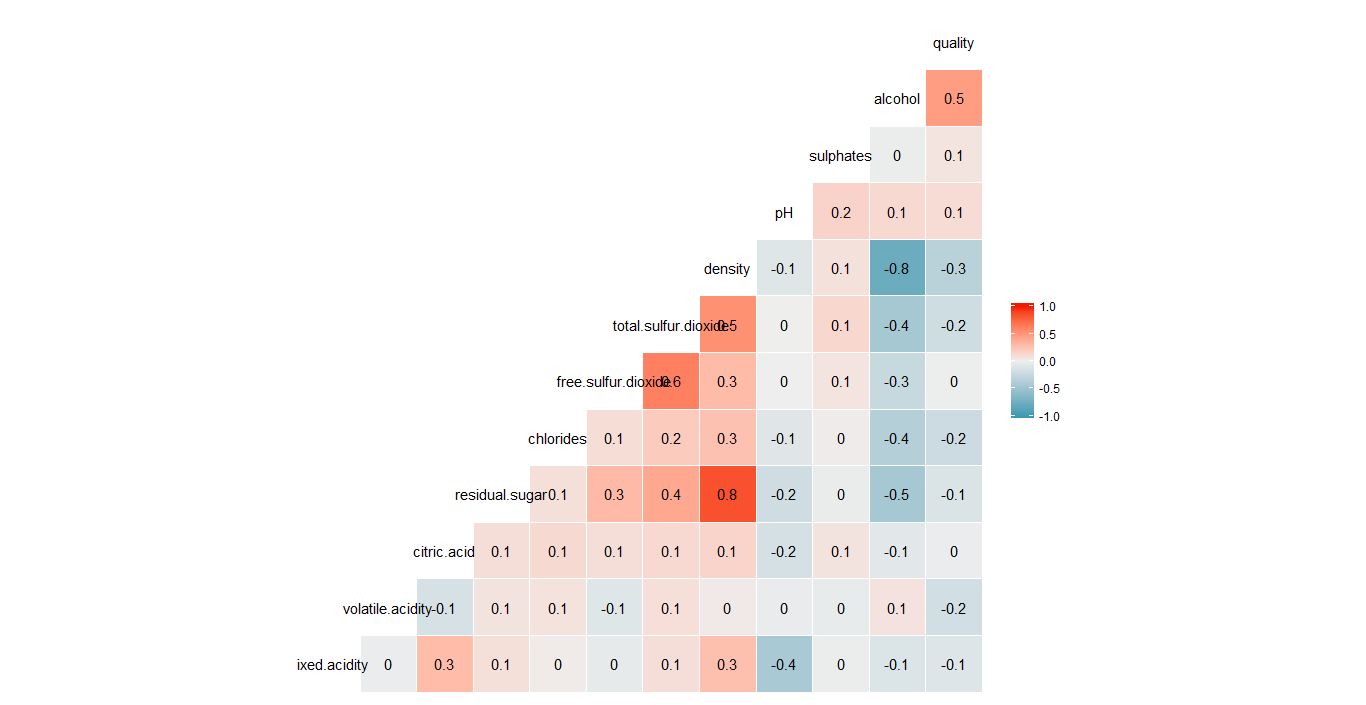


**(4) quality에 따른 11개 변수들의 boxplot (boxplot)**



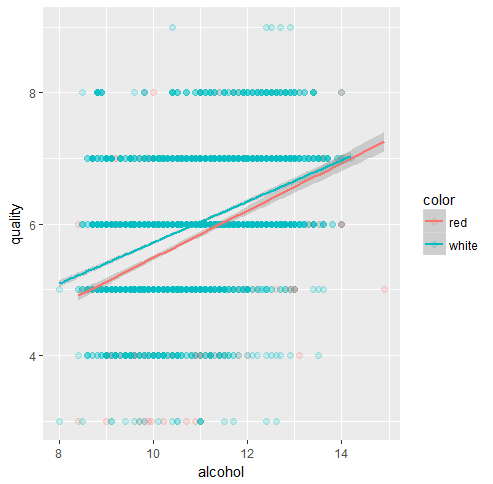
🡪 Outlier들의 영향으로 quality에 따른 boxplot이 제대로 그려지지 않아 boxplot 옵션 중에 outline=F를 해주어서 boxplot 모양을 유지하였다.

🡪 alcohol이 quality 차이에 영향을 주는 것으로 보인다.

**(5) quality에 따른 11개 변수들의 산점도와 상관계수(cor, corrplot)**

🡪 quality와 alcohol은 양의 상관관계를 갖고 있음을 알 수 있다.

🡪 alcohol과 density는 음의 상관관계가 강하다는 것을 알 수 있다.

**(6) White , Red Wine 에서 alcohol 변수 영향 확인하기 (ggplot)**

**🡪** Red, White Wine 모두 alcohol이 높을수록 높은 등급을 받은 것을 알 수 있다.