Visualization

제출일	2021, 05, 04
작성자	정승호

목 차

서론 3

본론

1. 데이터를 이용한 시각화와 패키지 성능 비교 3

결론 12

서론

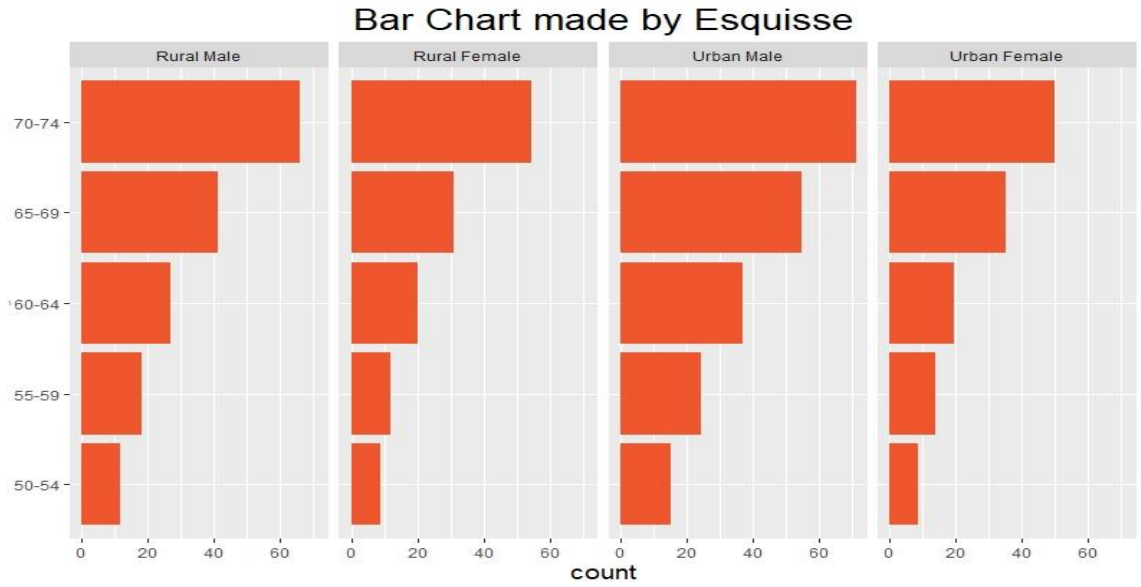
이 레포트는 빅데이터 시각화에 관련된 내용으로서, 시각화 패키지를 사용해 다양한 데이터의 시각화를 R code로 구현하고 각 패키지의 효율을 간단히 비교하였다.

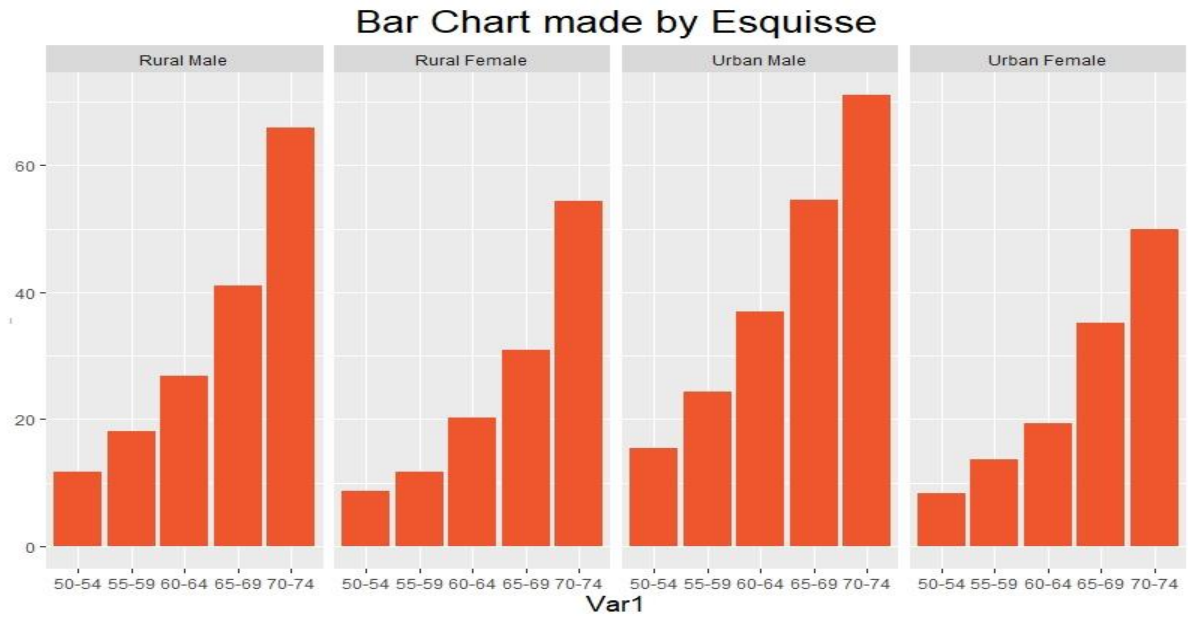
본론

1. 여러 데이터와 시각화 패키지를 사용해 그래프를 그리고 패키지 성능 비교

1)막대그래프

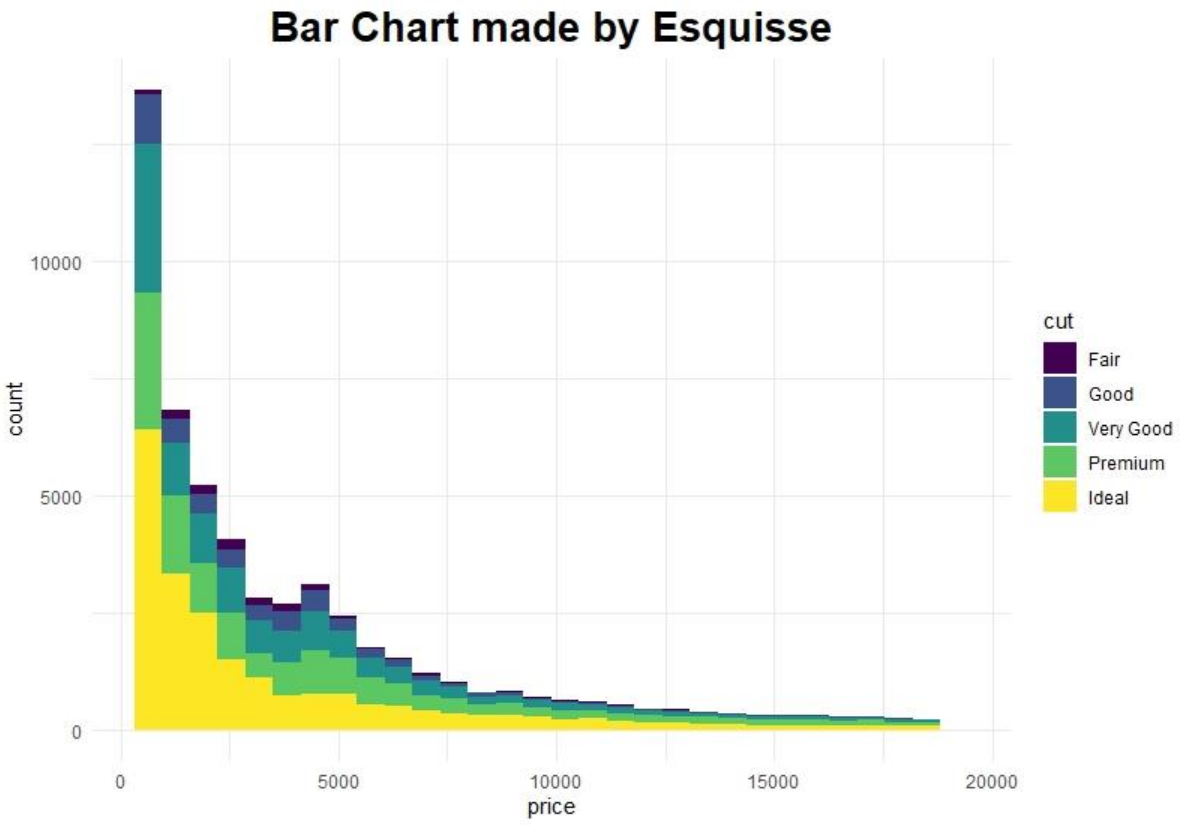
lattice와 ggplot2 어느쪽을 사용하던 간단히 코딩을 통해 시각화가 가능하다. 조금 복잡하게 나아갈 경우 ggplot2를 통해 시각화 하는것이 다양한 미적요소를 추가하거나 그래프 구성을 향상시키는데 효율적이다.





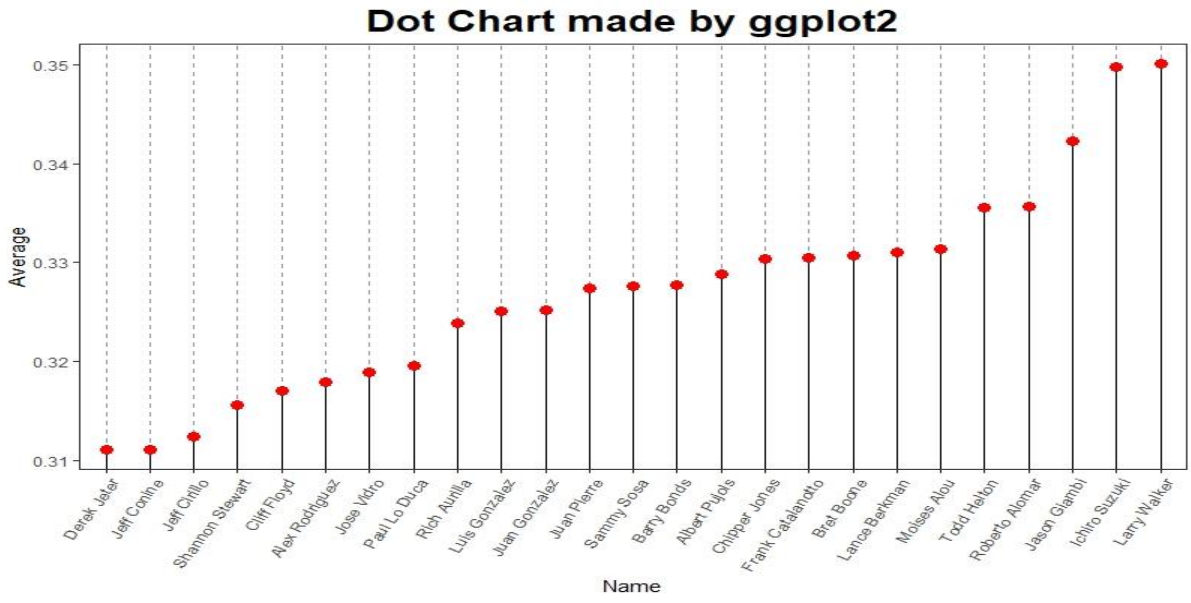
2) 누적 막대차트

내장 패키지 graphics와 ggplot2를 사용해 각각 시각화했으며 두 패키지 모두 시각화 구현상에 문제점은 발생하지 않는다. graphics 패키지가 좀 더 직관적으로 코드를 작성할 수 있지만, ggplot2는 esquisse를 통해 시각화 할 경우 미적요소 맵핑을 쉽게 추가할 수 있으며, 누적 요소를 다른 방식으로 표현할 수 있는 등 ggplot2가 좀 더 유연성이 뛰어나다.



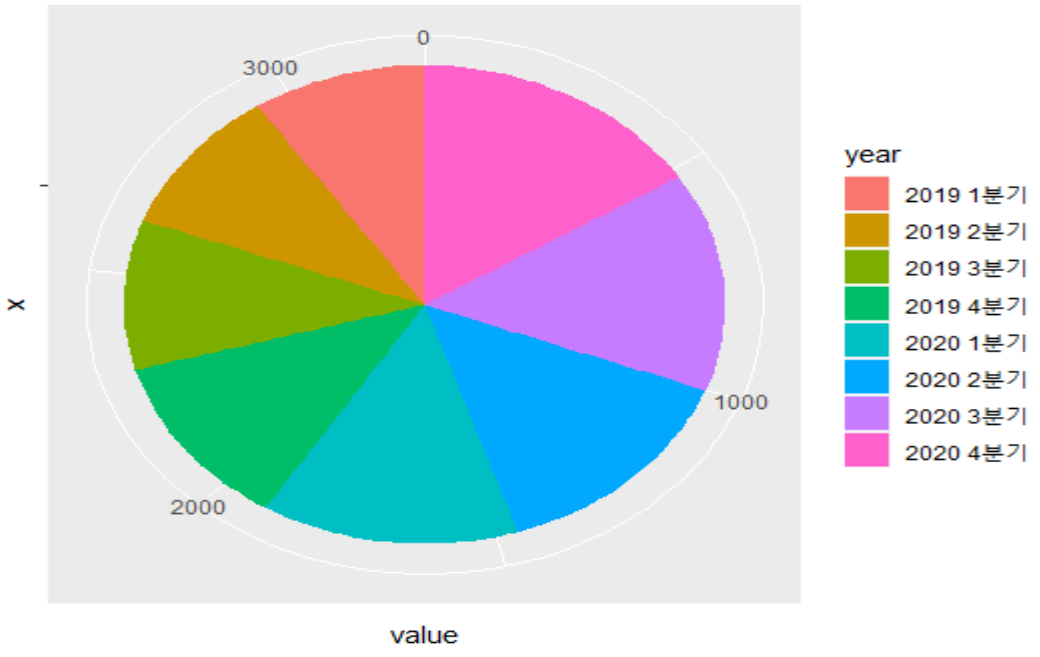
3.3. 점 그래프

lattice와 ggplot2로 시각화를 구현했다. lattice 패키지를 이용할 경우 점선을 연결해 추세를 알아보기 쉬우며, ggplot2를 이용할 경우 추가 매핑을 통해 점과 x축 범주를 연결해 각 점들의 값을 확인하기 쉽게 구현할 수 있다.



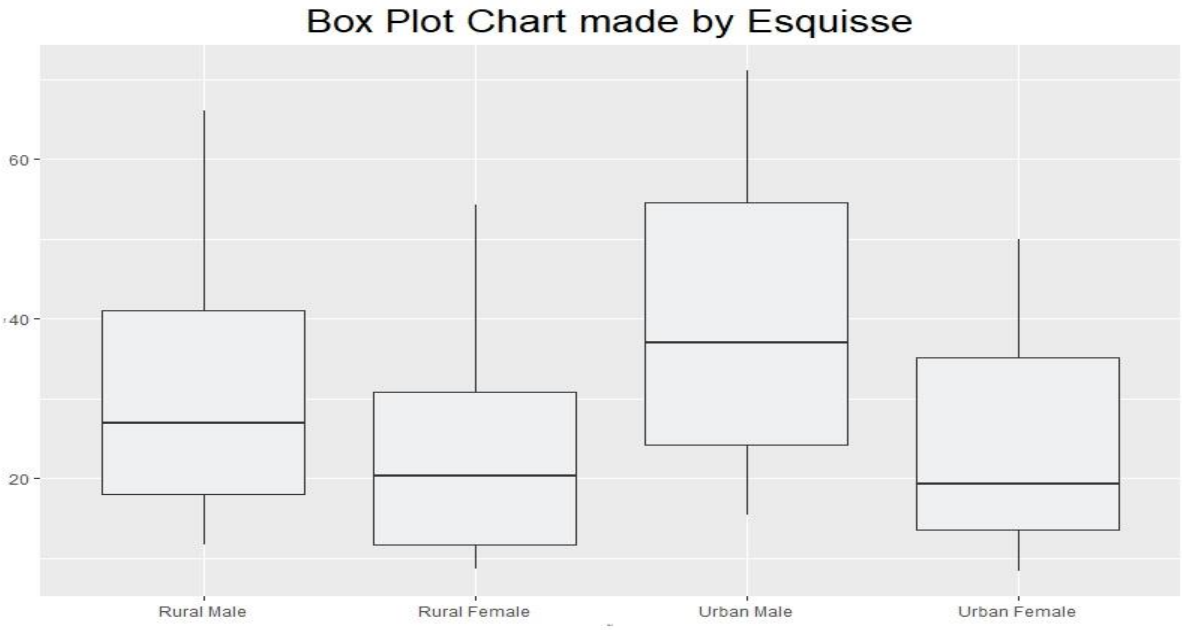
3.4. 원형 차트

graphics와 ggplot2를 이용하여 각각 시각화 하였다. 우선 ggplot2는 원형차트를 구현하기 있어서 먼저 막대차트를 구현해야 한다는 점이 아쉬운 부분이 있었다. 시각화된 두 이미지를 통해서는 데이터의 가독성이 크게 차이나지 않는 것을 확인할 수 있다.



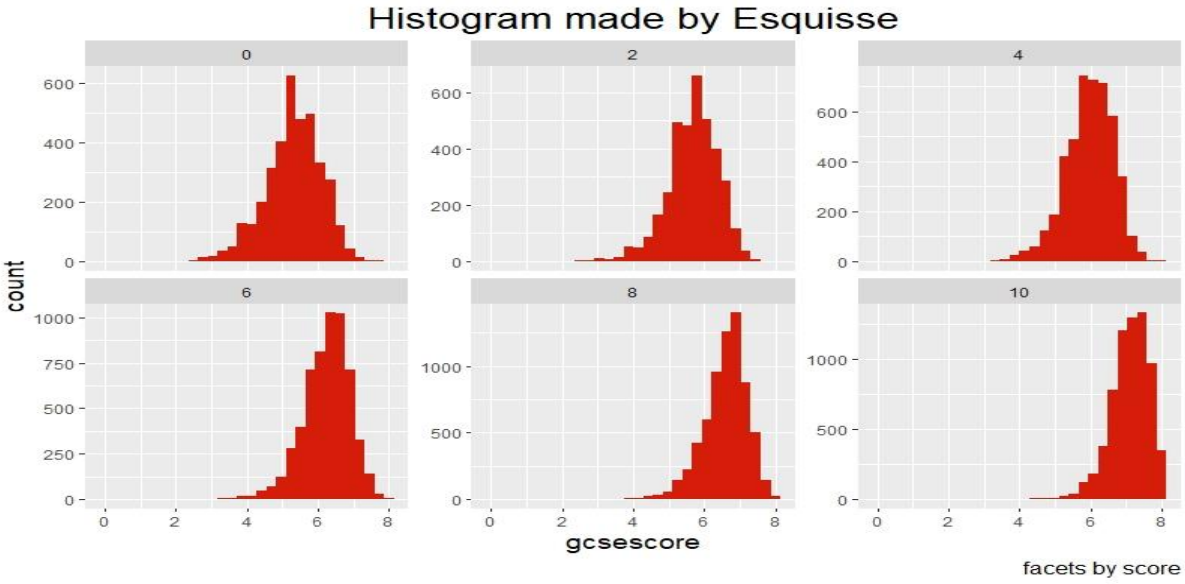
3.5. 상자 그래프

graphics와 ggplot2를 시각화에 이용했다. 두 패키지 모두 구현상 어렵거나 복잡한 점은 없다. 박스 파라미터 등 추가 맵핑요소의 경우도 문제없이 구현할 수 있는데, esquisse를 이용할 경우 단순 박스차트를 구성하기는 아주 편리하지만 추가 맵핑요소의 경우 직접 추가 코딩을 해주어야 하는 경우가 발생한다.



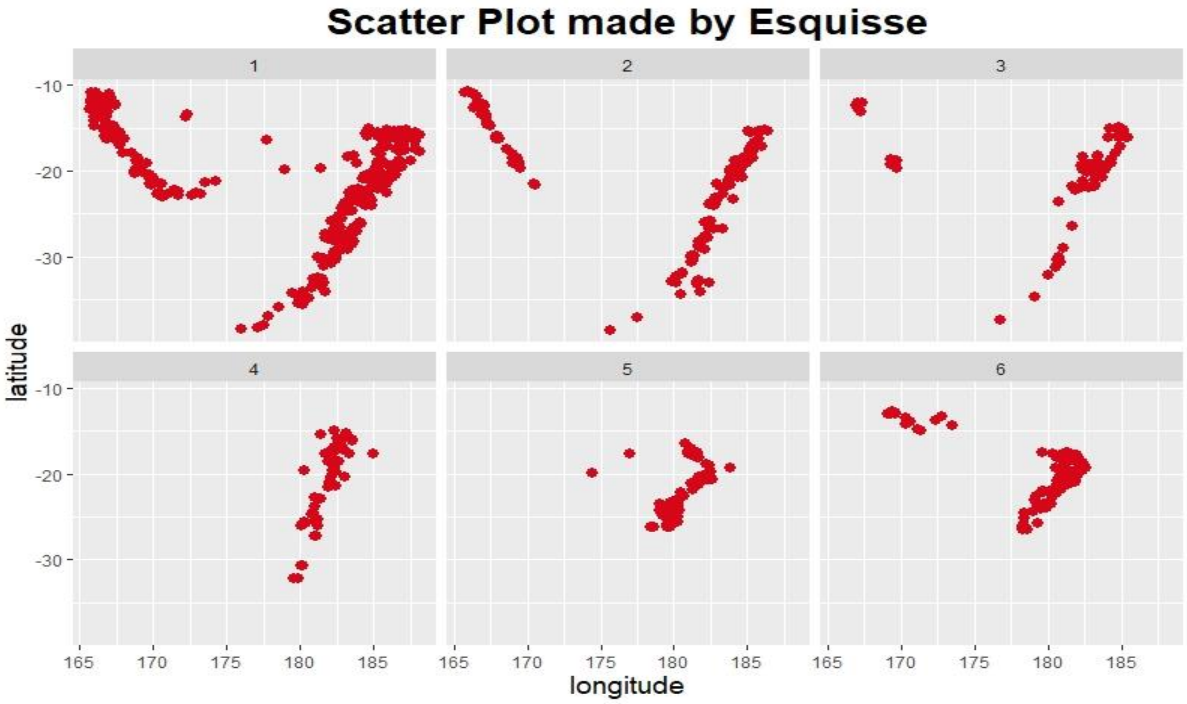
3.6. 히스토그램

기본 제공되는 패키지의 hist()함수, lattice와 esquisse를 이용했다. 단순 히스토그램 구현에는 모두 문제가 없지만, 추가적으로 밀도 곡선등을 구현하기에는 esquisse를 통한 ggplot 패키지 활용은 조금 곤란하며, 직접 추가 코딩을 해주어야 한다.



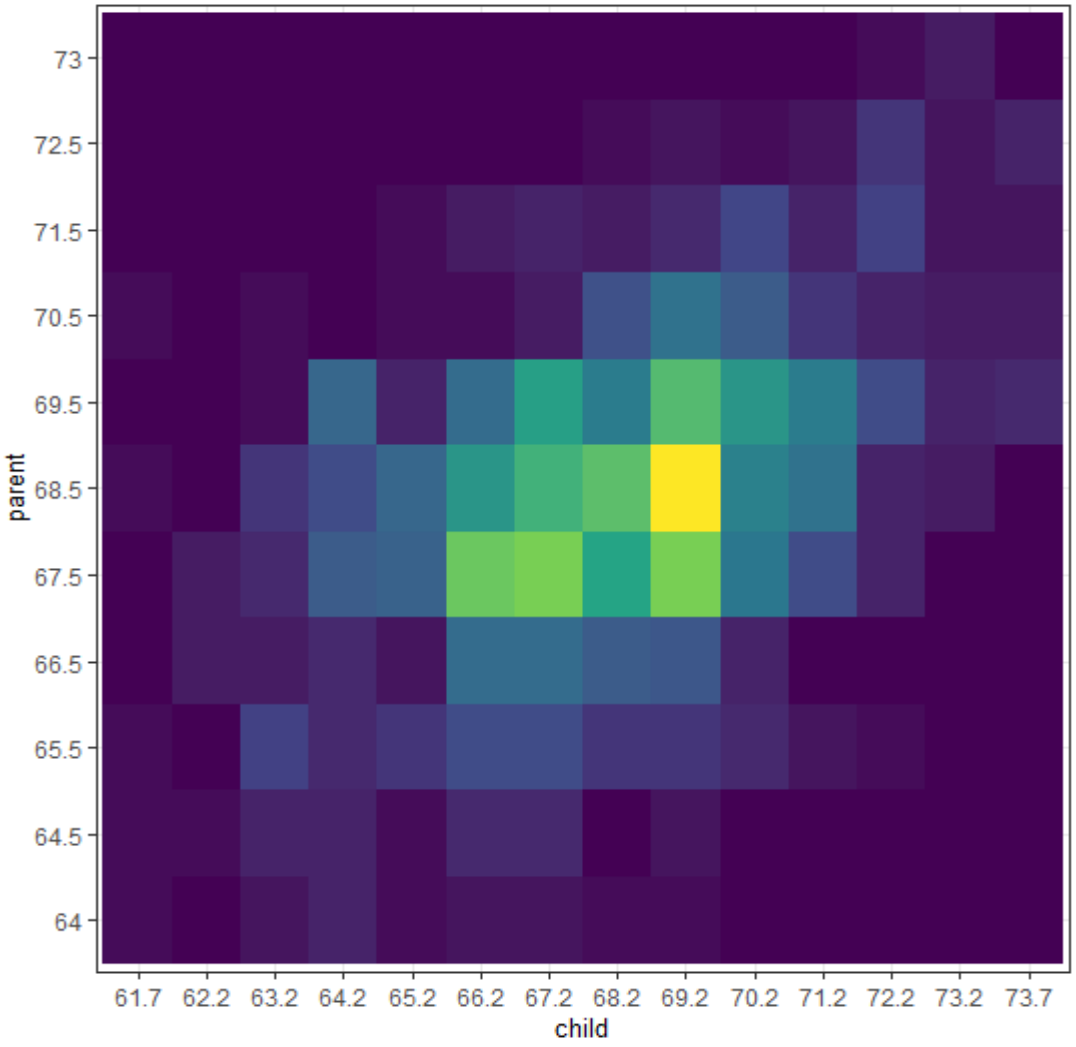
3.7. 산점도

Graphic과 esquisse의 ggplot을 이용하여 시각화하였다. 산점도를 시각화 하기 전에 반드시 데이터를 샘플링하는 과정이 필요로 하고 범주화를 통해 미리 준비해줘야한다. 산점도의 경우 데이터 전처리를 거치는 과정이 길어질 경우엔 lattice 패키지를 이용하는 것이 더 효율적일수도 있다. 두 이미지를 비교하였을 때 ggplot2의 미적맵핑요소에 가독성이 더 편리한 점을 확인할 수 있다.



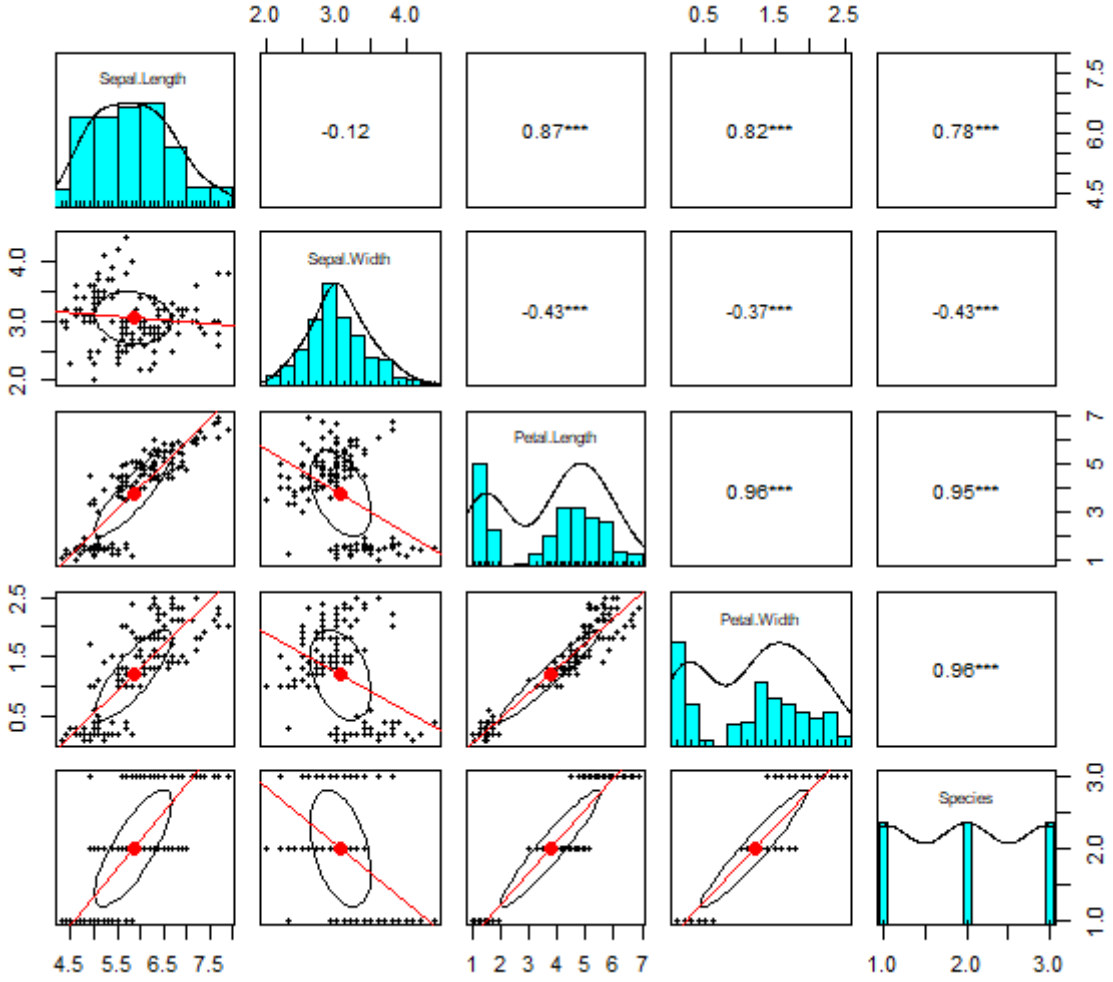
3.8. 중첩자료

graphics와 esquisse패키지를 시각화에 이용했다. 두 패키지 모두 오픈소스가 있었고 구현하는데 어려움이 없었다. 두 구현 방법 모두 parent, child를 축으로 기법을 적용하였는데 이미지와 같이 esquisse는 graphics에 비해 데이터 가독성이 떨어지는 것을 볼 수 있다.



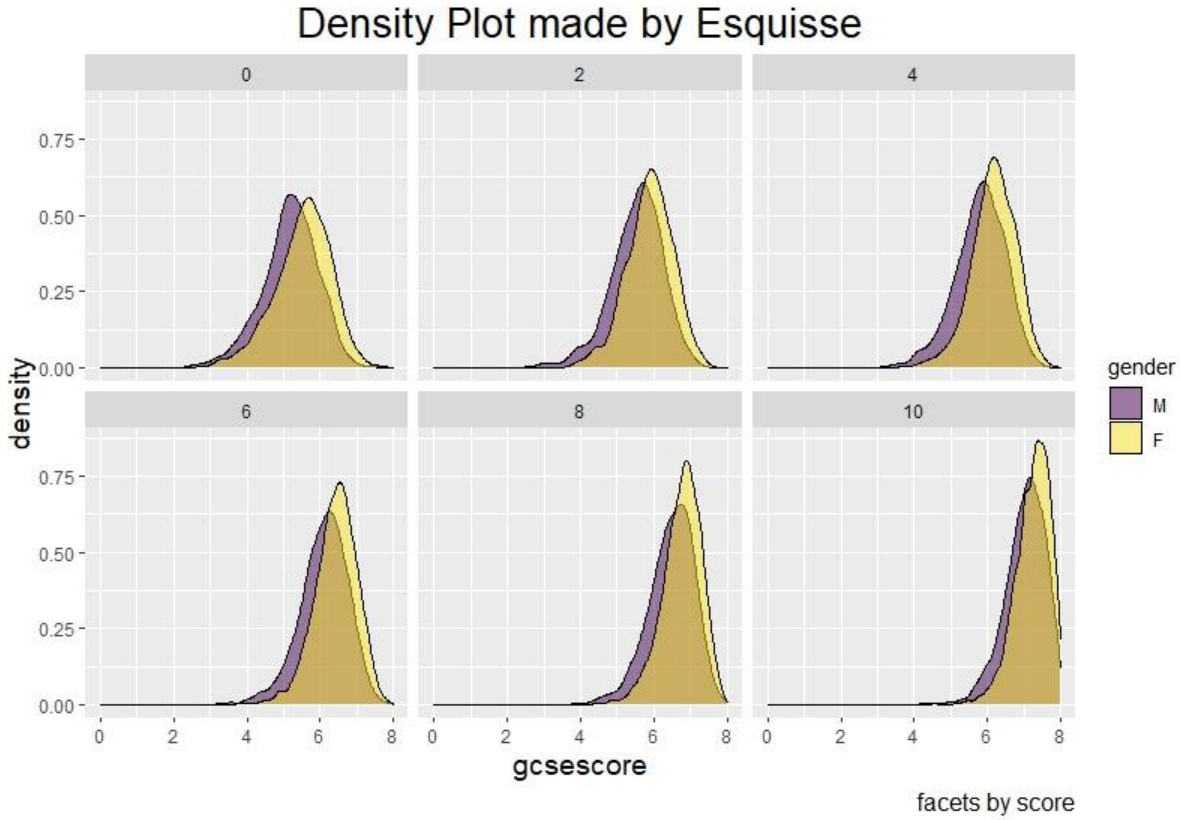
3.9. 변수간의 비교

graphics와 psych 패키지의 pairs.panel() 함수를 이용하면 산점행렬도를 쉽게 그릴 수 있다 pairs() 함수보다 더 다양한 옵션을 제공한다. 이미지를 통해 pairs() 함수와 비교하면 많은 옵션들이 그래프에 추가된 것을 확인할 수 있다.



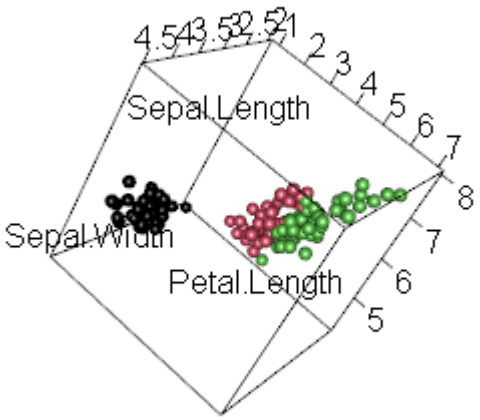
3.10. 밀도그래프

lattice와 ggplot2를 이용해 시각화했다. lattice 패키지의 경우 밀도곡선 형태가 디폴트로 제공되기때문에 추가적으로 투명도를 지정해주는 등 별도의 코딩이 필요없다. ggplot2는 esquisse 패키지를 이용하는 경우 밀도 그래프가 영역을 차지하는 형태로 제공되며, 투명도를 별도 코딩해주어야 시각화 되었을 때 변수간 밀도 차이를 식별하기 편리하다.



3.11. 3d 산점도

scatterplot3d, lattice, rgl 세 패키지를 사용했으며, 세 패키지 모두 무난히 구현이 가능하다. 3d 구현화에 최적화되어있는 rgl 패키지는 여러 각도로 산점도를 살펴볼 수 있는 추가 콘솔창을 제공하며, 변수색 지정등 코딩도 타 패키지에 비해 간단하기 때문에 사용하기 편리하다.



12.지도 시각화

ggmap과 ggplot2을 이용해 시각화했다. ggmap의 경우 지도 시각화에 특화되어 있기 때문에 좀 더 다양한 미적 맵핑요소가 제공되며, 간편하게 코딩할 수 있고, 지도 정보를 받아오기도 용이하다. ggplot2는 지도 시각화를 위해 더 많은 데이터가 요구되며, 전처리작업도 좀 더 번거롭게 요구되고 코딩 역시 ggmap에 비해 직관적이지 못하다. 때문에 ggmap에 비해 지도 시각화에서는 효율적이지 않다.

결론

다양한 패키지들에서 시각화 기능이 제공되지만, ggplot2 패키지가 가장 범용성이 높으며 다양한 매핑요소들이 지원되어 효율적으로 시각화에 활용할 수 있다. esquisse 패키지 등 ggplot2를 이용하기 좋은 패키지도 지원되고 있다. 또한 rgl이나 ggmap등 특정 시각화에 특화된 패키지들 역시 해당 부분 시각화에는 간단한 코딩과 다양한 매핑요소 제공 등 강점이 존재하기 때문에, 시각화가 필요한 데이터의 특성에 맞는 패키지를 활용하는 것이 요구된다.