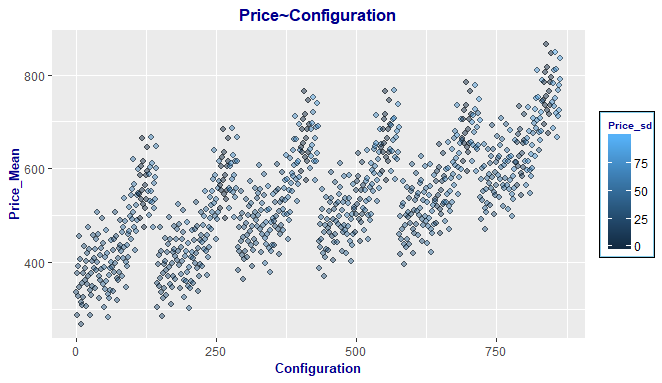
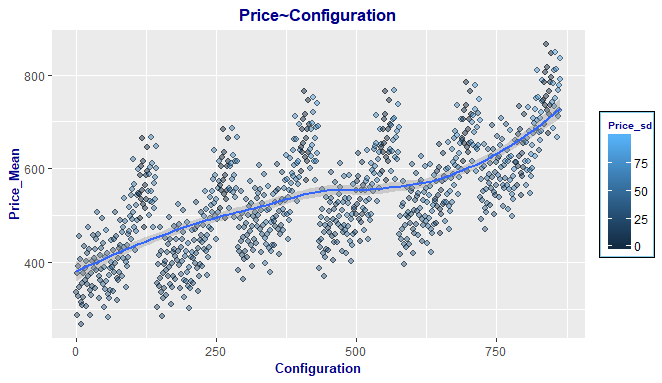
3.4

a.

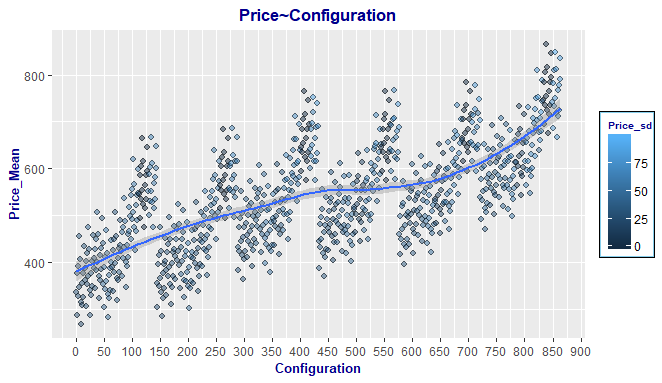
i.실제로 노트북은 얼마에 판매되었는가?



위 그림을 보면 configuration ~ Retail.Price 가 Configuration의 증가에 맞게 Retail.Price가 증가하는 비례관계임을 알 수 있다. 이 형태를 좀 더 명확하게 보기 위해 회귀선을 부여했다.



회귀선을 부여한 결과, 두 변수의 상관관계를 더 강하게 볼 수 있었다. 그렇다면 노트북의 판매가를 어떻게 말해야 할까. 그래프를 보면 Configuration에 따라 명확한 집단이 6개 보인다. 해당 구간만큼 집단을 쪼개 그 집단의 평균 판매가를 답하려 한다.



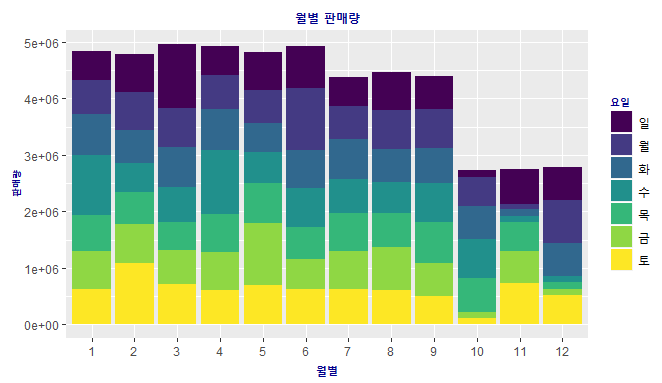
. scale\_x\_continuous 함수를 사용해 집단의 범위를 세밀하게 파악한다. 이후 집단을 X축 기준으로 나누어 6개의 집단으로 묶어준다. 6집단으로나눴을 경우 집단 간 평균 가격은 다음과 같다.

**M\_Case1 M\_Case2 M\_Case3 M\_Case4 M\_Case5 M\_Case6**

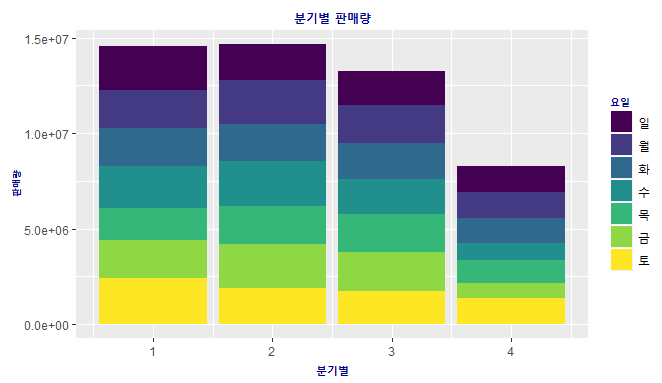
**1 443.2061 468.4685 550.6505 553.131 571.1029 652.8876**

**ii. 시간에 따라서 판매가격이 변화하였는가?**

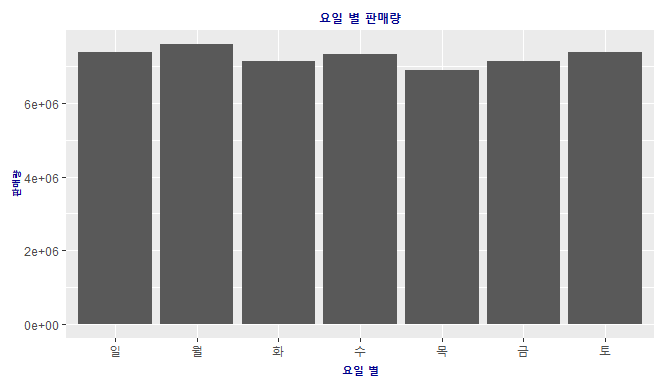
우선, 시간을 기준으로 판매가격을 볼 때, 시간의 단위를 세 가지 종류로 나누려 한다.  
월 별, 분위 별, 요일 별. 월 별, 분위 별 데이터에는 추가적으로 요일 변수를 삽입할 수 있어 삽입한 상태에서 결과를 보려 한다.



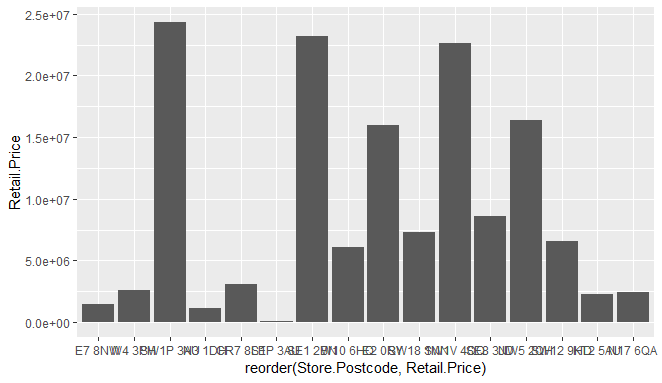
위 그림은 월별-판매가의 자료 그래프다. 해당 노트북 가격의 누적량을 통해 연초가 연말보다 더 높은 가격을 형성한다고 볼 수 있다. 10월과 12월 사이의 노트북 판매가격은 급격하게 하락한다. 10월과 12월 사이에 추수감사절, 크리스마스 등의 이벤트가 영향을 끼친 것인지 조심스럽게 추측해본다.  
요일 데이터의 경우, 1월부터 9월까지는 요일 간 판매가격에 차이가 보이지 않는다. 그러나 10월부터 12월 사이의 데이터에는 유의미한 변화가 보인다. 10월 (금, 토, 일), 11월(월, 화, 수), 12월(수, 목, 금) 판매가가 타 요일에 비해 상대적으로 저렴하다. ‘런던 컴퓨터’가 월마다 해당 요일에 행사를 열었을 거라 생각해본다.



분기별 데이터의 경우, 1, 2, 3분기에 비해 4분기 판매량이 급격히 하락했다. 월별 데이터로 파악한 결과, 10, 11, 12월의 매출 하락이 동일한 결과로 나타났다고 본다. 해당 시기의 요일 데이터는 4분기 요일(수,금)이 타 요일에 비해 판매량이 부진한 것을 제외하고는 유의미한 차이가 보이지 않았다.

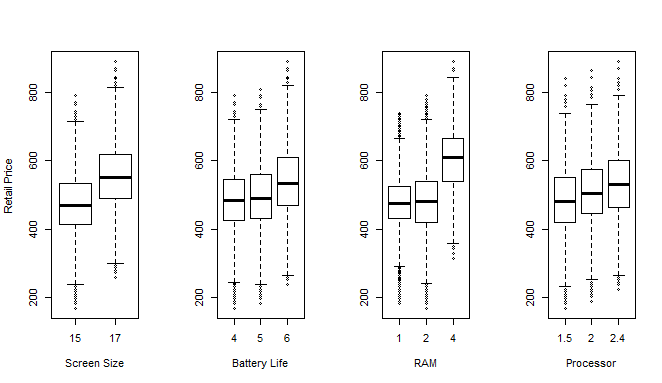


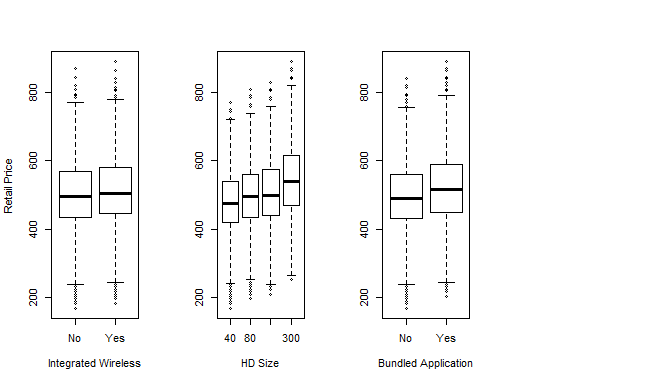
요일 판매량 만을 막대그래프로 출력한 결과, 한 해를 다루어 보았을 때 요일 간 뚜렷한 차이는 보이지 않았다. 전체적으로 판매량이 균형적이었다.



상점과 판매량의 막대그래프를 관찰한 결과, 지역 별로 판매가에 편차가 강하게 나타나는 것을 알 수 있다. 지역 별로 선호하는 모델, 가격대가 존재하는 것으로 추측된다.

# iv 판매 가격은 컴퓨터 사양에 따라 어떻게 다른가?





노트북의 각 성능과 가격을 비교해 Box Plot으로 결과를 나타냈다.

Screen Size의 경우, 15인치에 비해 17인치 노트북의 가격이 더 높았으며, 중간값 지점을 미루어 보아 그 차이가 50% 이상 존재했다.

Battery Life의 경우, 4시간과 5시간 사이에는 유의미한 차이가 보이지 않았다. 그러나 6시간의 경우 타 배터리에 비해 근소한 차이를 보였다.

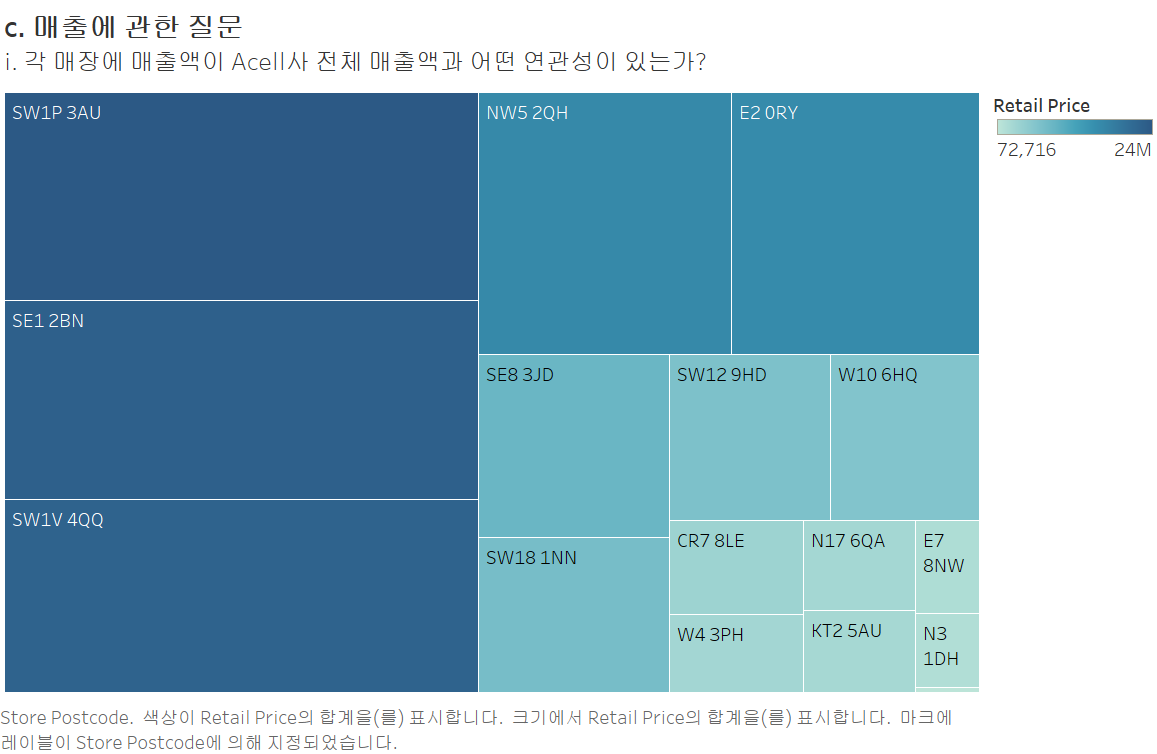
RAM의 경우, 1G RAM과 2G RAM은 판매가에 있어 동일한 형태를 보였으나, 4G RAM는 둘에 대해 50%가 훌쩍 넘는 변화를 보였다.

Processor의 경우, 1.5, 2, 2.4가 근소하지만 전체적으로 증가하는 모양새를 보였다.

Integrated Wireless의 경우, 둘 사이의 유의미한 가격 변동의 징후가 보이지 않았다.

HD Size의 경우, 전체적으로 가격이 증가하는 모양새이나 그 평균 정도가 50%를 넘지 않는다.

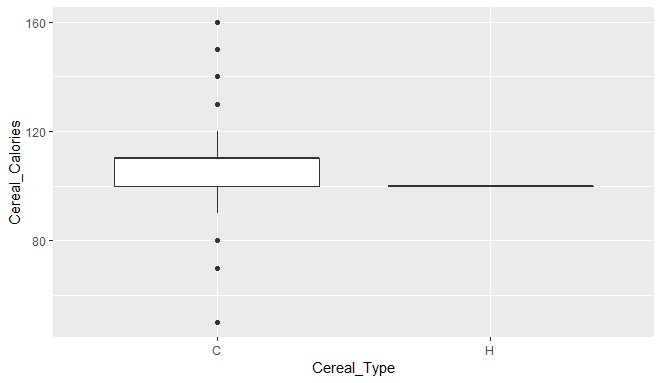
Bundied Application의 경우, 유무에 있어 근소한 차이를 보였다.



해당 트리맵을 통해 전체 매장 중 ‘SW1P 3AU’, ‘SE1 2BN’, ‘SW1V 4QQ’ 세 매장이 전체 매출의 50%를 감당하고 있는 것으로 파악된다. 또한 ‘E7 8NW’, ‘N3 1DH’, ‘N17 6QA’, ‘KT2 5AU’ 등은 매출을 모두 합쳐도 앞서 말한 세 매장 중 한 곳의 영업이익을 넘기지 못하고 있는 것을 알 수 있다.

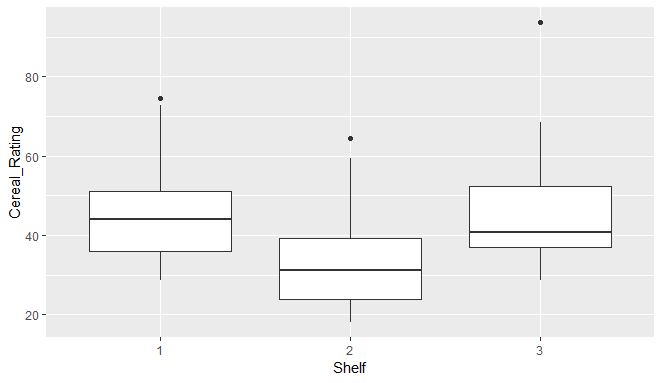
4.1

d. R을 이용하여 저온용과 고온용 시리얼에 함유된 칼로리를 비교하기 위해 박스플롯을 나란히 그리시오. 그리고 이를 통해 무엇을 알 수 있는가?



박스플롯을 통해 Cold 상태가 Hot보다 칼로리의 편차가 존재한다는 것을 확인했다. 저온용 시리얼의 경우 전체적인 전체적인 칼로리의 범위는 100 ~ 110 사이이나 분포가 40부터 160까지인 것으로 상당히 넓게 퍼져있는 것으로 보인다. 반면, 고온용 시리얼의 경우 칼로리의 분포가 거의 0에 수렴한다.

e. R을 사용하여 시리얼이 전시된 진열대 높이에 따라 소비자 평점의 박스플롯을 나란히 그리시오. 진열대 높이로부터 소비자 평점을 예측할 수 있다면, 진열대 높이의 세 가지 범주를 그래도 유지할 필요가 있겠는가?



박스플롯을 통해 진열대 높이가 시리얼에 대한 소비자 평점과 관계가 있음을 파악했다.

1층에 존재하는 시리얼이 타 층에 비해 전체적으로 높은 평점을 보유하고 있고, 그 다음이 3층, 마지막이 2층이다. 3층에 존재하는 시리얼의 경우, 타 층에 비해 이상치가 가장 높은 것으로 보인다.

진열대의 높이가 소비자의 평점에 영향을 준 것인지, 판매처가 소비자 평점을 토대로 진열을 바꾼 것인지 인과관계는 알 수 없으나, 두 변수 간에 유의미한 상관관계가 존재함을 파악할 수 있다.

f. 양적 변수에 대한 상관계수 표를 작성하시오. 또한 이들 변수에 대한 산점도 행렬을 작성하시오.

> cor(Cereal[,4:11], use="complete.obs")

calories protein fat sodium fiber

calories 1.00000000 0.03399166 0.5073732397 0.2962474981 -0.29521183

protein 0.03399166 1.00000000 0.2023533963 0.0115588913 0.51400610

fat 0.50737324 0.20235340 1.0000000000 0.0008219036 0.01403587

sodium 0.29624750 0.01155889 0.0008219036 1.0000000000 -0.07073492

fiber -0.29521183 0.51400610 0.0140358654 -0.0707349230 1.00000000

carbo 0.27060605 -0.03674326 -0.2849336855 0.3284091857 -0.37908370

sugars 0.56912054 -0.28658397 0.2871524866 0.0370589612 -0.15094850

potass -0.07136125 0.57874284 0.1996367171 -0.0394380876 0.91150392

carbo sugars potass

calories 0.27060605 0.569120535 -0.071361247

protein -0.03674326 -0.286583967 0.578742837

fat -0.28493369 0.287152487 0.199636717

sodium 0.32840919 0.037058961 -0.039438088

fiber -0.37908370 -0.150948502 0.911503921

carbo 1.00000000 -0.452069189 -0.365002934

sugars -0.45206919 1.000000000 0.001413982

potass -0.36500293 0.001413982 1.000000000

해당 상관계수 표를 통해 fiber ~ potass 변수가 가장 강한 상관성을 띄는 것을 확인했다.

ii. 이들 상관관계를 바탕으로 어떻게 변수의 개수를 축소할 수 있는가?

상관관계가 높은 변수만 추출해서 사용가능하다.