Report

데이터마이닝 3.3장, 3.4장 연습문제

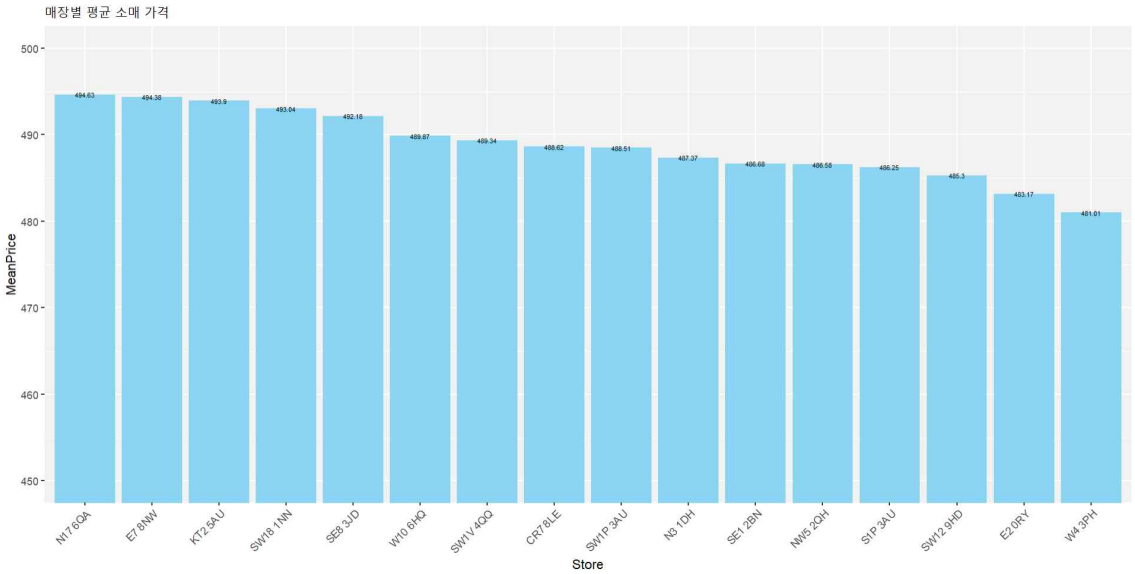
32201805 박정민

DKU 단국대학교 DANKOOK UNIVERSITY

3.3

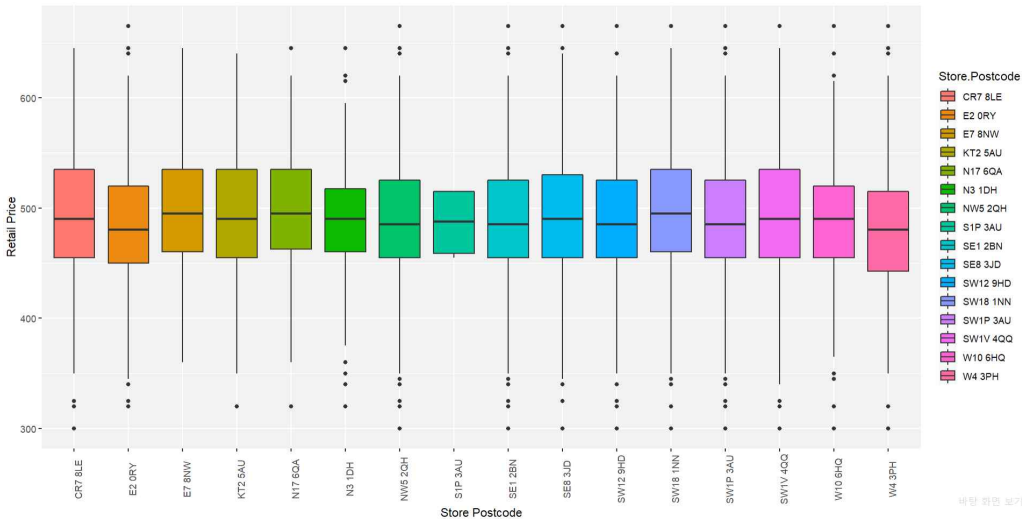
런던 컴퓨터 체인점의 노트북 판매실적: 막대차트와 박스플롯. LaptopSalesJanuary 2008.csv 파일은 런던 소재의 한 컴퓨터 체인점의 2008년 1월 매출데이터이다. 이것은 2008년도 전체 매출데이터의 일부이다.

a. 매장별 평균 소매가격을 보여주는 막대차트를 그리시오. 평균 소매가격이 가장 높은 매장은 어느 곳인가? 반대로 가장 낮은 평균 소매가격은 어떤 매장인가?

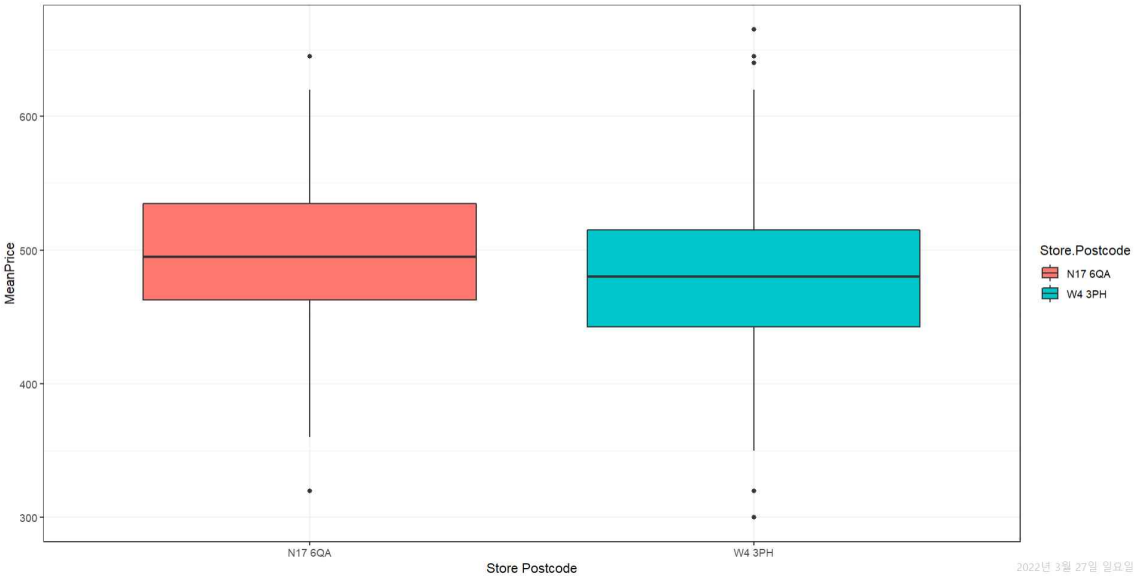


- 위 그래프는 매장별 평균 소매가격을 내림차순으로 나타낸 막대그래프이다. 평균 소매가격이 가장 높은 매장은 464.64로 N17 6QA가 가장 높고 484.01로 W4 3PH가 가장 낮은 평균 소매가격을 가지고 있다.

b. 매장별 소매가격을 더 잘 비교하려면 병렬 박스플롯을 그리시오. (a)에서 찾은 두 매장의 가격을 비교해 보시오. 두 매장의 가격분포가 어떤 차이점이 있는가?



- 위 그래프는 매장별 소매가격을 병렬 박스플롯으로 나타낸 것이다. (a)에서 찾은 평균 소매가격이 가장 높고 가장 낮은 매장 두 곳의 병렬 박스플롯을 그리기 위해서 두 곳의 매장을 추출하여 그린다.



- 위에 두 매장의 박스플롯을 그려보았다. 두 매장의 가장 큰 차이는 평균 소매가격이고 N17 6QA가 W4 3PH보다 높은 평균가격을 가지고 있다는 것을 박스의 위치에 따라 알 수 있다. 또한, 두 매장의 약 50%의 데이터의 양은 박스플롯의 IQR을 보면 비슷하다는 것을 볼 수 있다.

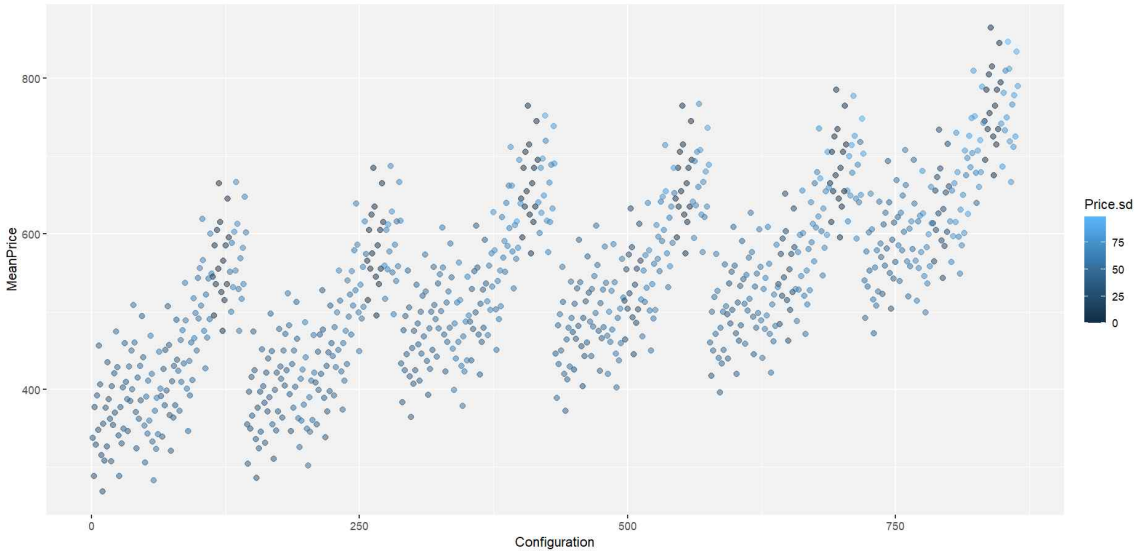
3.4

런던 컴퓨터 체인점의 노트북 판매실적: 대화형 시각화. 이 과제는 대화형 시각화 툴을 사용하도록 준비되었다. LaptopSales.txt라는 파일을 30만 행에 달하는 콤마로 분리된 파일이다. 이 데이터는 2009년 가을 거행된 콘테스트를 위해 ENBIS(the European Network for Business and Industrial Statistics)가 제공하였다.

시나리오: 여러분이 노트북 컴퓨터를 판매하는 AceII이라는 회사의 데이터 분석가라고 가정하자. 여러분에게 제품과 판매에 대한 데이터가 제공되었다. 여러분에게 주어진 과제는 2009년도 AceII사의 예상매출을 극대화하는 제품전략과 가격정책을 만드는 것이다. 대화형 시각화 툴을 사용하여 아래 질문에 답하시오.

a. 가격에 관한 질문

i. 실제로 노트북은 얼마에 판매되었는가?

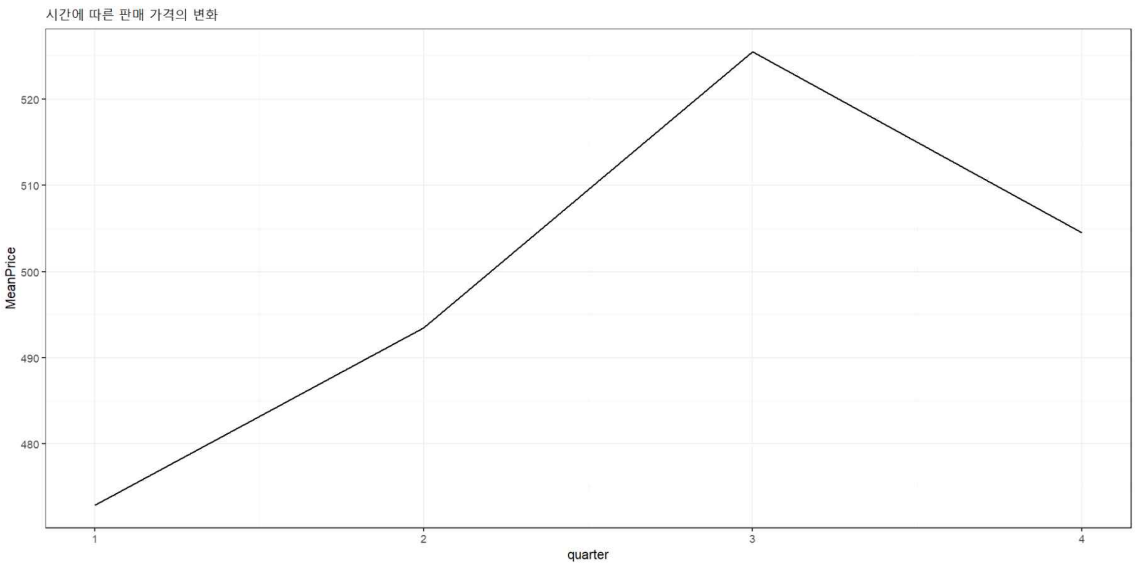


- 실제로 노트북이 얼마에 판매되었는지 보기 위해서 Configuration 별 평균 판매가격에 대한 산점도 그래프를 그렸다. 점의 색깔이 짙을수록 편차가 작다. 전체적으로 Configuration에 따른 평균 가격이 6개의 덩어리로 나눠져있어 같은 덩어리안에 있는 Configuration은 평균 가격의 차이가 적다. Configuration이 0에서 864로 갈수록 평균가격이 상승하는 것을 볼 수 있고 이 그래프의 Y축은 Configuration의 소매가격들의 평균값을 나타내어 점의 색깔이 진할수록 편차가 작아 실제 판매된 노트북 가격에 가깝다고 볼 수 있다.

ii. 시간에 따라서 판매 가격이 변화하였는가?

- 우선, 각 시간에 따른 판매 가격의 변화를 보기 위해 Laptop.csv의 data, month, mday 변수를 날짜 데이터로 바꾸어주었다.

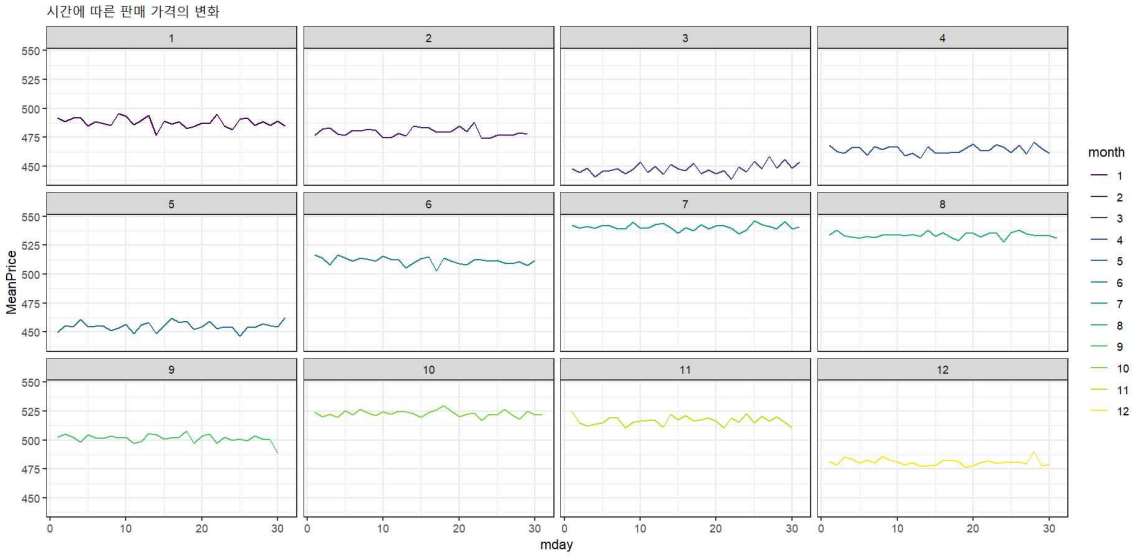
(1) **사분기**별로 판매 가격의 변화를 보자.



- 위 그래프는 사분기별 판매 가격의 평균을 나타내었다. 그래프는 3사분기에서 가장 높은 가

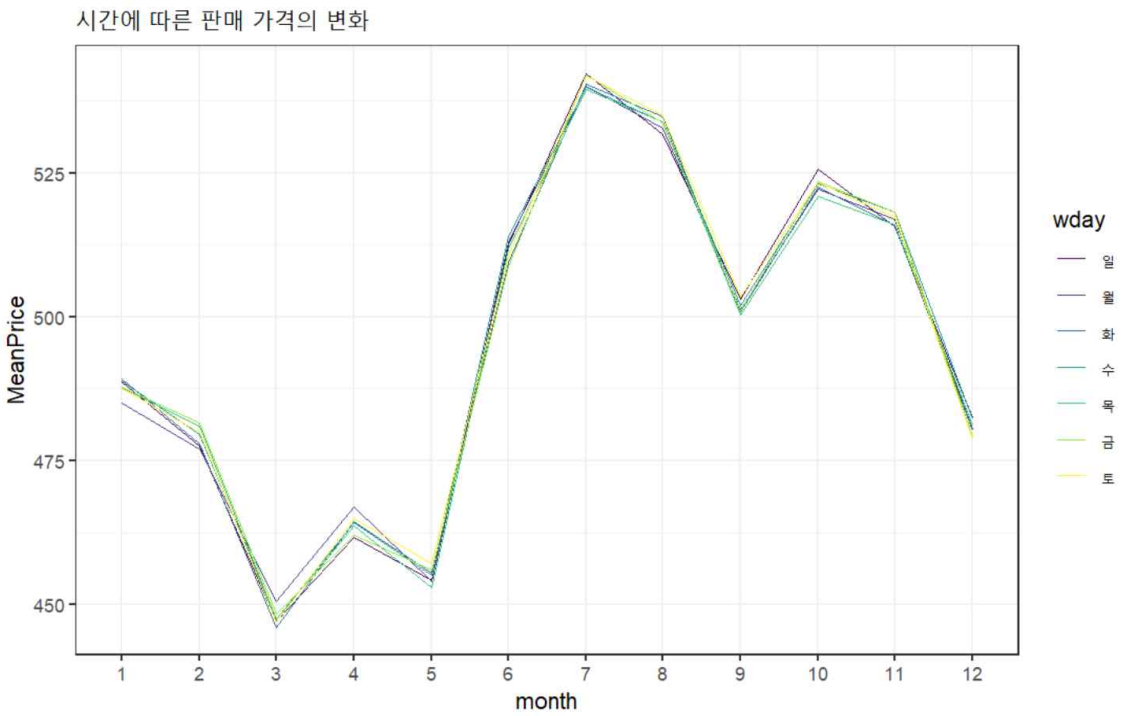
격을 나타내고 1사분기에서 가장 낮은 가격을 나타내었다.

(2) **달**별로 판매 가격의 변화를 보자.



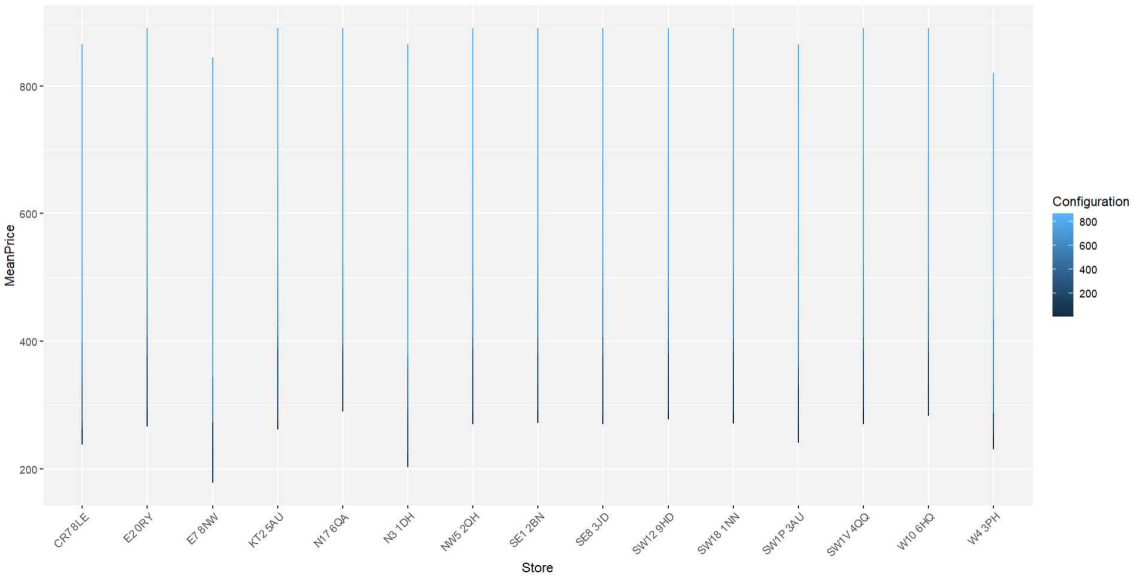
- 위 그래프를 보면 달별로 판매 가격의 평균을 볼 수 있다. 12달 중 7월과 8월이 가장 높은 가격을 나타내고 있고 3월과 5월이 가장 낮은 가격을 나타내고 있다. 앞서 사분기별 평균가격을 나타내었던 그래프와 같은 결과를 나타낸다.

(3) **요일**별로 판매 가격의 변화를 보자.

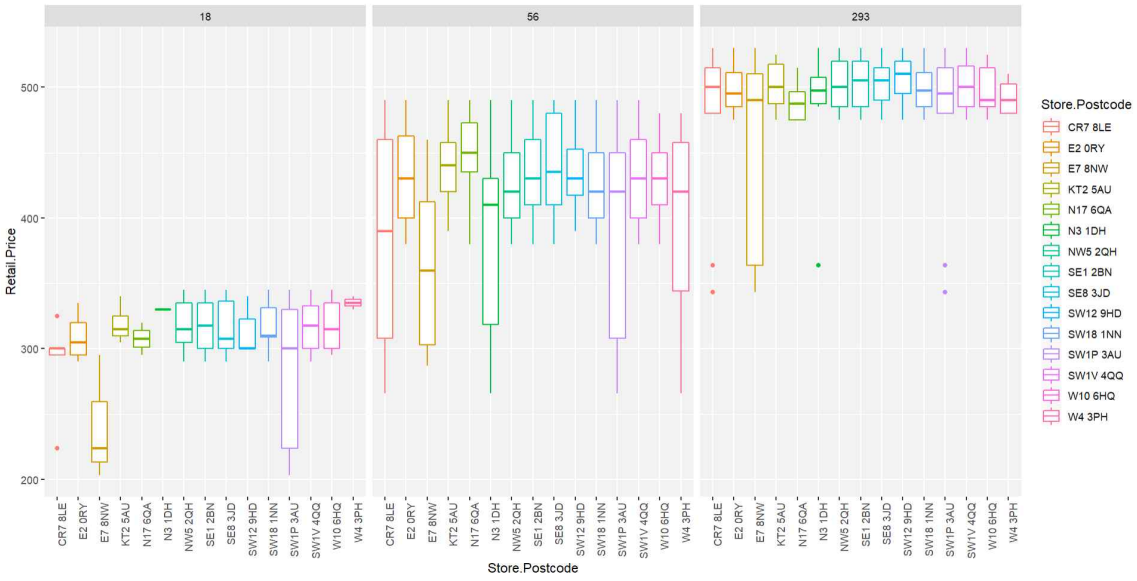


- 위 그래프는 요일별로 색깔을 다르게 지정하여 달별 평균가격을 나타낸다. 7개의 요일은 같은 경향을 띠고 있다. 따라서, 요일별로는 사분기와 달보다는 평균가격에 영향을 끼치지는 않는다는 것을 알 수 있다.

iii. 판매 가격은 각 매장별로 일관성이 있는가?

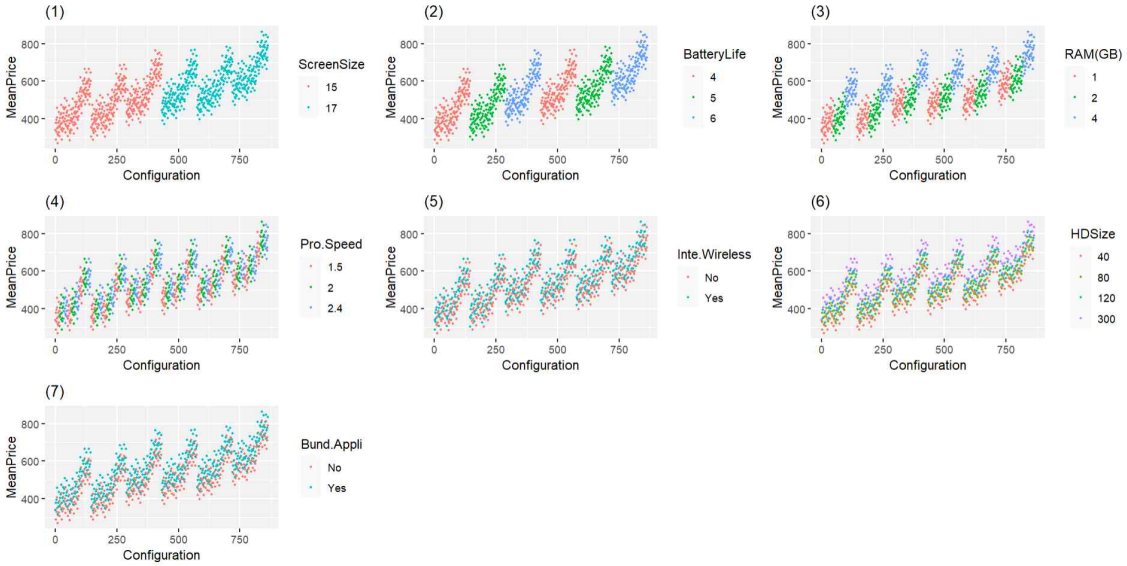


- 위 그래프는 매장에 따른 평균 소매가격을 그린 그래프이다. Congifuration은 그래프의 색깔이 연할수록 높은 값을 띤다. 따라서, Configuration이 높을수록 평균소매가격은 높다고 할 수 있지만 판매 가격이 각 매장별로 일관성이 있다는 것은 알기가 어렵다. 따라서, Configuration에서 3개를 랜덤추출하여 매장별 소매 가격을 나타내는 박스플롯을 그려보았다. 랜덤추출한 3개의 Configuration은 '56', '18', '293' 이다.



- 위 그래프를 보면 박스플롯의 크기나 중앙값, IQR 모두 일관적이지 않아 매장별 소매가격의 분포에 일관성이 없음을 한눈에 볼 수 있다.

iv. 판매 가격은 컴퓨터 사양에 따라 어떻게 다른가?

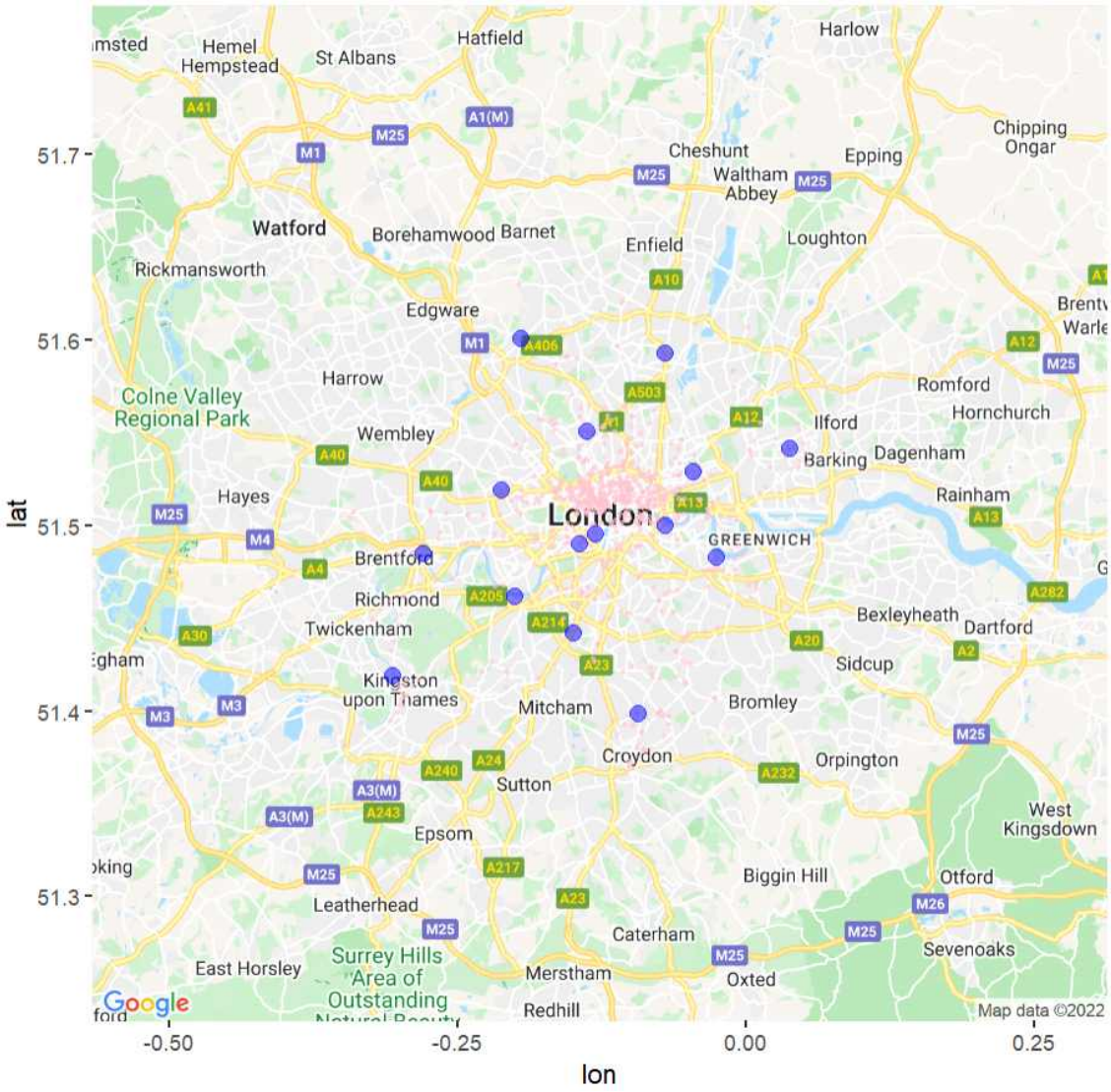


- 7개의 컴퓨터 사양에 따른 평균 판매가격에 대한 그래프를 산점도로 나타내었다. 7개의 그래프 모두 전체적으로 6개의 덩어리로 나누어져있다.

(1) ScreenSize에 대한 평균가격은 Configuration값이 0~375에는 대부분 15inch를 나타내고 375~864는 17inch를 나타내는 ScreenSize가 분포한다. 이 그래프를 통해 17이 15보다 평균가격이 높다는 것을 알 수 있다.
(2) BatteryLife에 대한 평균가격은 한 덩어리안에는 같은 BatteryLife가 존재하고 '4', '5', '6'인 BatteryLife가 두 번 반복된다. Configuration이 0~400인 곳만 본다면 BatteryLife가 높을수록 평균가격이 상승한다.
(3) RAM에 대한 평균가격은 한 덩어리안에 아래쪽에는 1이 위치하고 위쪽에는 4가 위치한다. 따라서, RAM에 따라 평균가격이 상승한다고 볼 수 있다.
(4) Processor.Speeds는 덩어리끼리 같은 구성을 띤다. 한 덩어리안에서는 빨간색점에서 파란색점으로, 파란색점에서 초록색점으로 갈수록 평균가격이 상승하여 Processor.Speeds가 클수록 평균가격이 상승한다.
(5) (7)과 같은 경향의 그래프이다. 한 덩어리안에서 파란색 점이 빨간색 점보다 위에 분포함을 볼 수 있어 파란색 점들이 빨간색 점들보다 높은 평균 가격을 띠고 있다는 것을 볼 수 있다.
(6) HD.Size..GB.에 따른 평균가격을 나타내는 그래프이다. 덩어리마다 같은 분포를 띠고 있고 한 덩어리안에는 빨간점이 가장 밑에 분포하고 보라색점이 가장 위쪽에 분포한다. 이는 HD.Size도 높을수록 평균가격이 높다는 것을 알 수 있다.
(7) (5)번 참고

b. 매장에 관한 질문

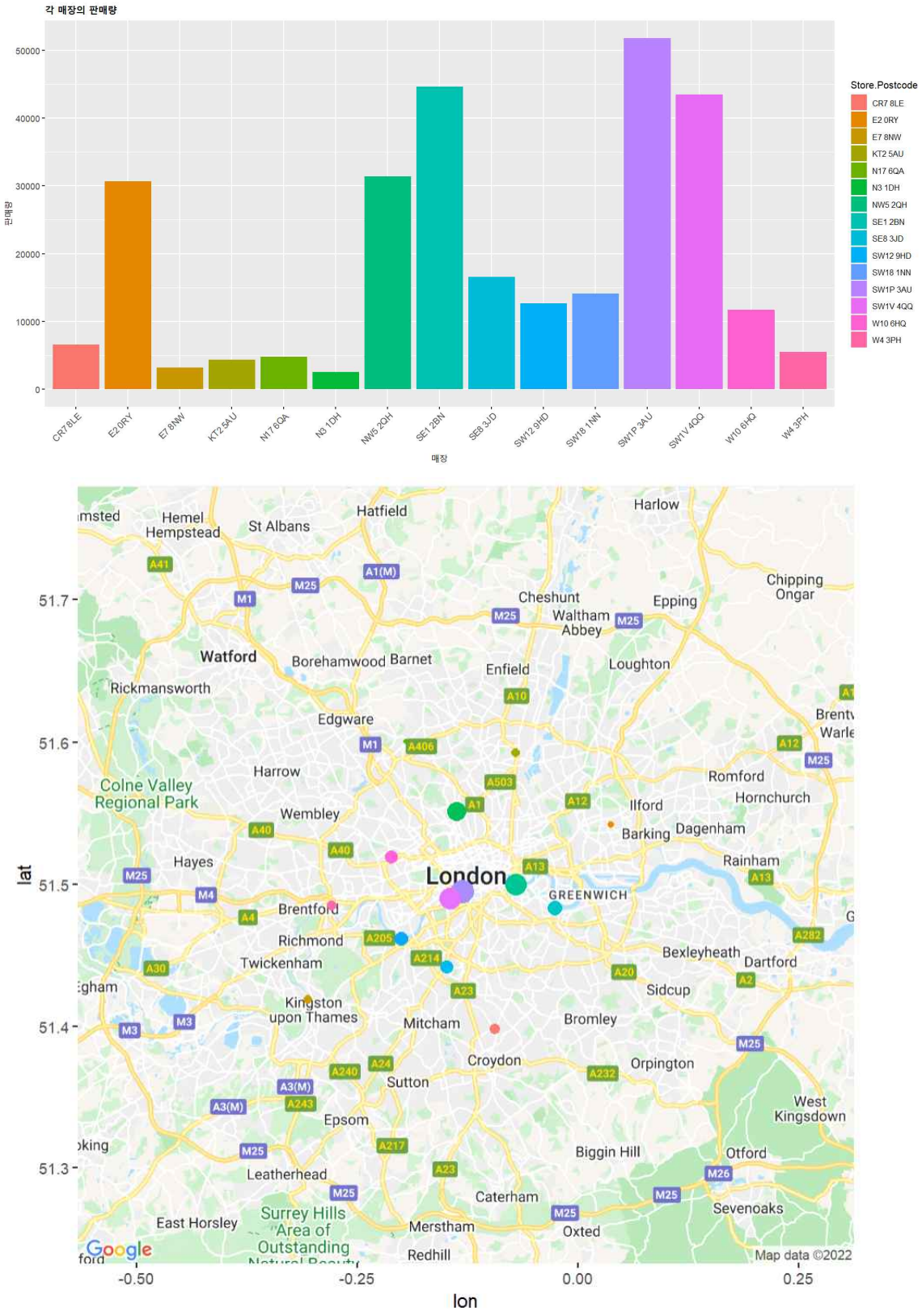
i.매장의 위치와 고객의 위치는 어디인가?



- 위에 지도 그래프에 고객의 위치는 분홍색 동그란 점으로, 매장의 위치는 파란색 동그란 점으로 표시하였다. 고객의 위치는 London의 중심부에 대부분 많이 몰려있는 것을 볼 수 있다.

ii. 어느 매장이 가장 많이 판매하는가?

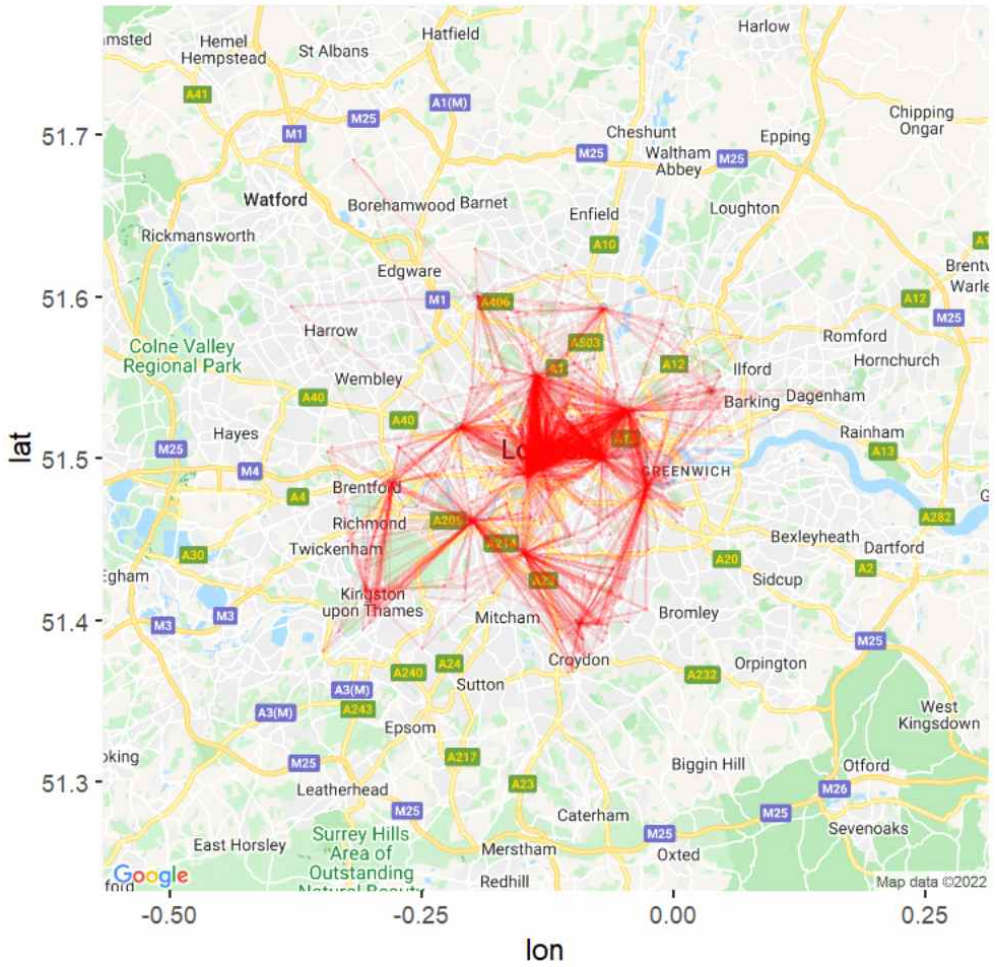
- 아래 그래프는 각 매장의 판매량을 막대 그래프로 나타낸 것이다. 가장 많이 판매한 매장은 SW1P 3AU이고, SE1 2BN, SW1V 4QQ순으로 많이 팔렸고 N3 1DH 매장이 가장 적게 판매하였다.



- 각 매장별 판매량을 지도에 나타낸 것이다. 매장별로 점의 색깔이 다르고 점의 사이즈가 클수록 판매량이 많다. London의 중심부에 대부분 크기가 큰 점들이 모여있고 퍼져갈수록 점

의 크기가 작아지는 걸 알 수 있다.

iii. 고객은 노트북을 사기 위해서 얼마나 이동하는가?

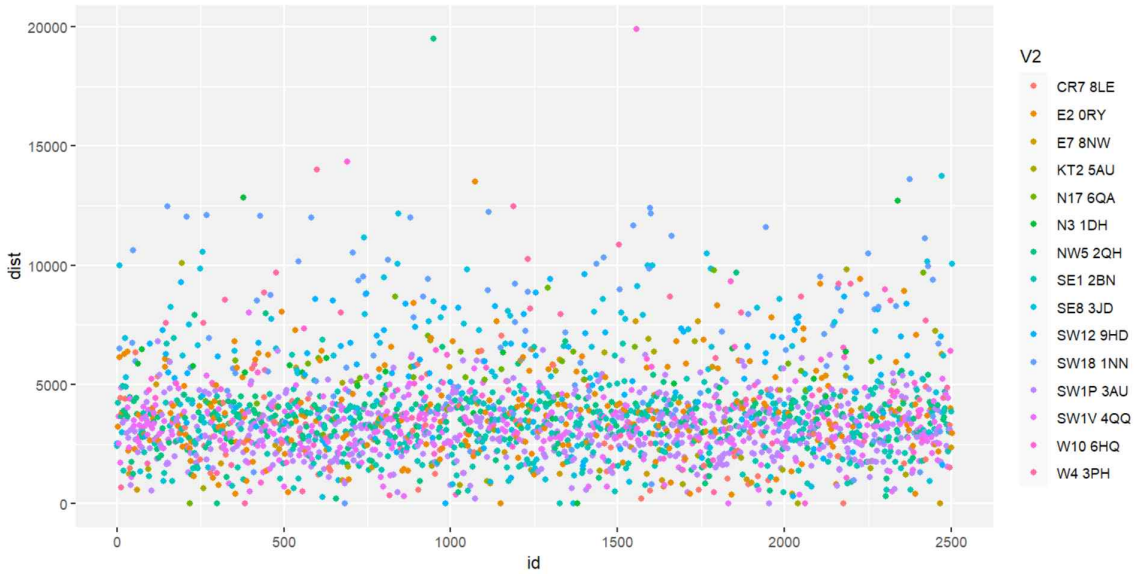


- 위 그래프는 고객들의 위치를 빨간색 점으로 나타내고 매장들의 위치를 회색 점으로 나타내었다. 고객들의 위치를 각 매장에 이은 선이 빨간색 선이므로 이 선이 고객의 이동거리를 나타낸다. 이 지도를 보면 London의 중심부가 주변부보다 이동거리가 많고 London 중심부의 아래쪽이 위쪽보다 이동거리가 많다.

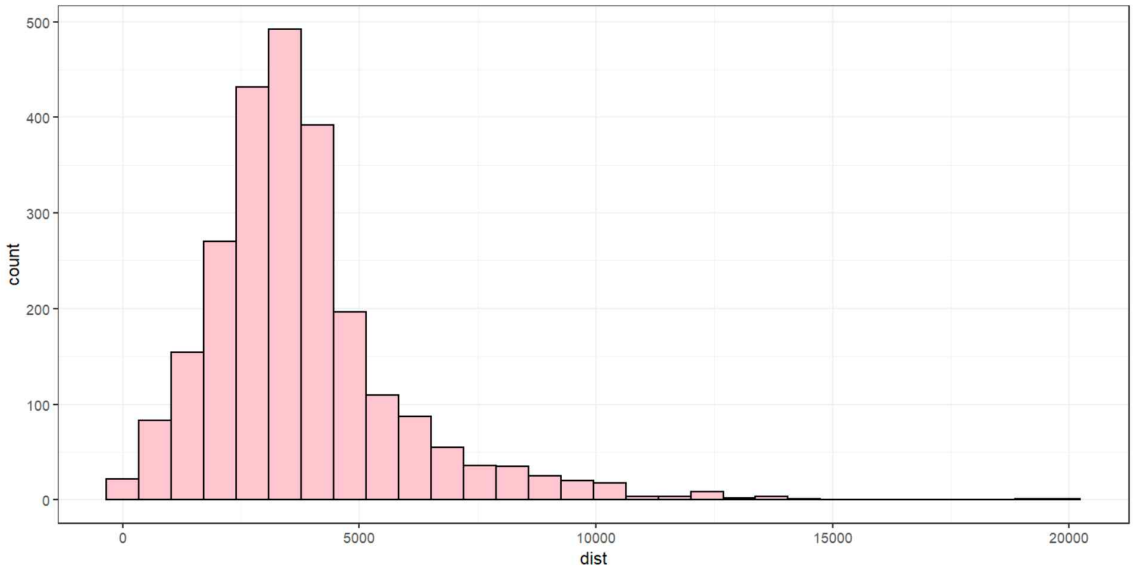
iv. 고객이 매장을 방문하기 위해 얼마나 이동하는지 다른 방법을 통해 알아보자. 고객과 매장 간의 이동거리를 담은 새로운 데이터 열을 만들어서 이 문제를 풀어보자.

```
library("geosphere")
c$dist=distGeo(b3a, b3b)
c$id<-c(1:2502)
c<-as.data.frame(c)
c
```

V1(Customer.Postcode), V2(Store.Postcode) 가 데이터프레임으로 들어있는 c에 고객과 매장사이의 이동거리를 담은 새로운 데이터 열과 id를 데이터프레임 c에 추가한 코드이다. b3a 변수는 고객의 위치를 나타내고 b3b 변수는 매장의 위치를 나타낸 것이다.



- 위 그래프를 보면 산점도를 통해 고객이 매장을 방문하기 위해 이동하는 거리가 모두 20000m이내라는 것을 알 수 있고 대부분 이동거리가 0~5000m에 분포해있다.

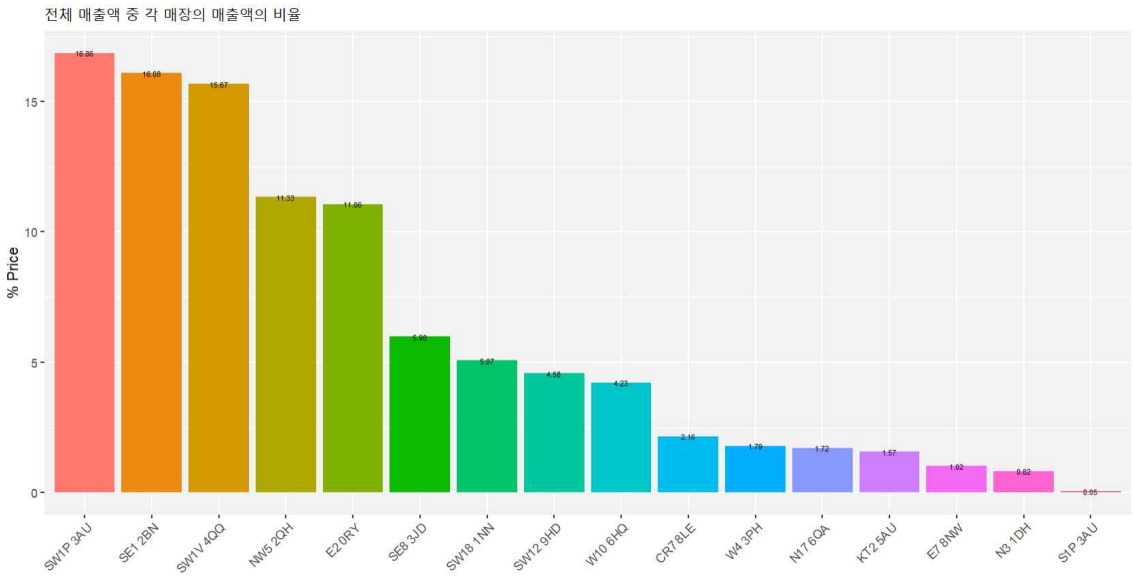


- 이동거리에 따른 히스토그램을 그려본 것이다. 위의 산점도 그래프와 동일하게 이동거리가 0~10000인 곳에서 대부분이 나타나있고 10000이상인 곳에는 거의 이동거리가 나타나지 않았다.

c. 판매에 대한 질문

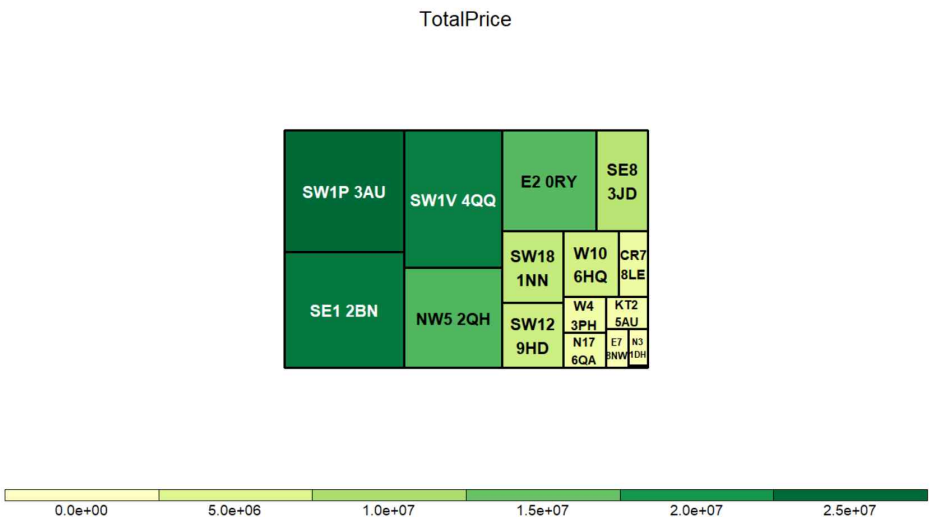
i. 각 매장에 매출액이 Acell사 전체 매출액과 어떤 연관성이 있는가?

① 막대그래프



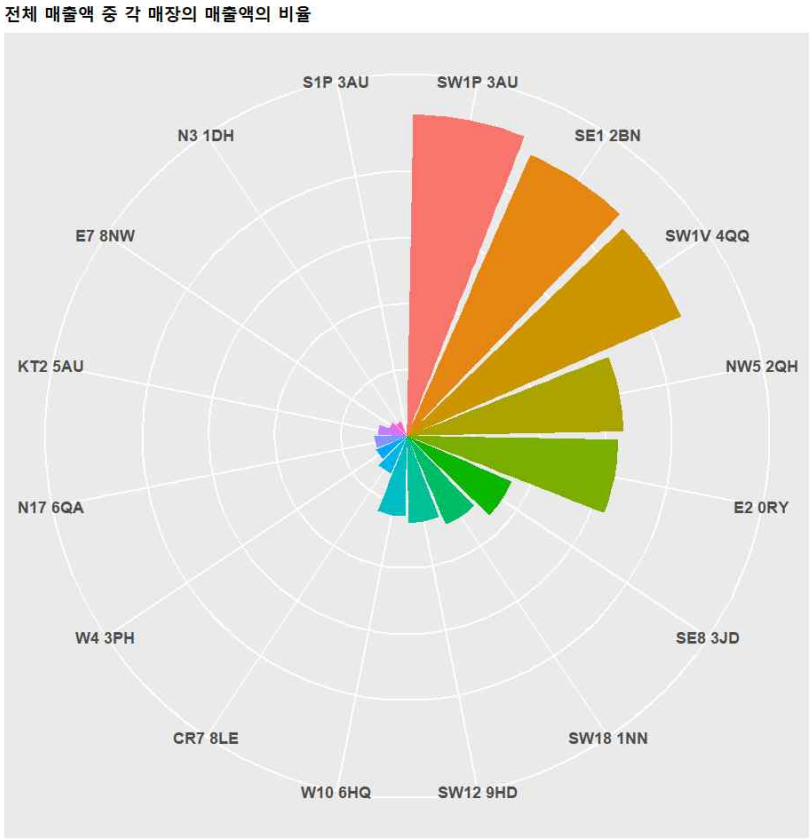
- 위 그래프는 전체 매출액 중 각 매장의 매출액의 비율을 막대그래프로 나타낸 것이다. Store.Postcode를 설명변수로 전체 매출액을 각 매장의 매출액의 총합으로 나눈 것의 비율을 반응변수로 하여 내림차순으로 그래프를 그려주었다. 가장 많은 비율을 차지하는 매장 3곳은 'SW1 3AU', 'SE1 2BN', 'SW1V 4QQ'이고, 가장 적은 비율을 차지하는 매장 3곳은 S1P 3AU, N3 1DH, E78NW이다.

② treemap 그래프



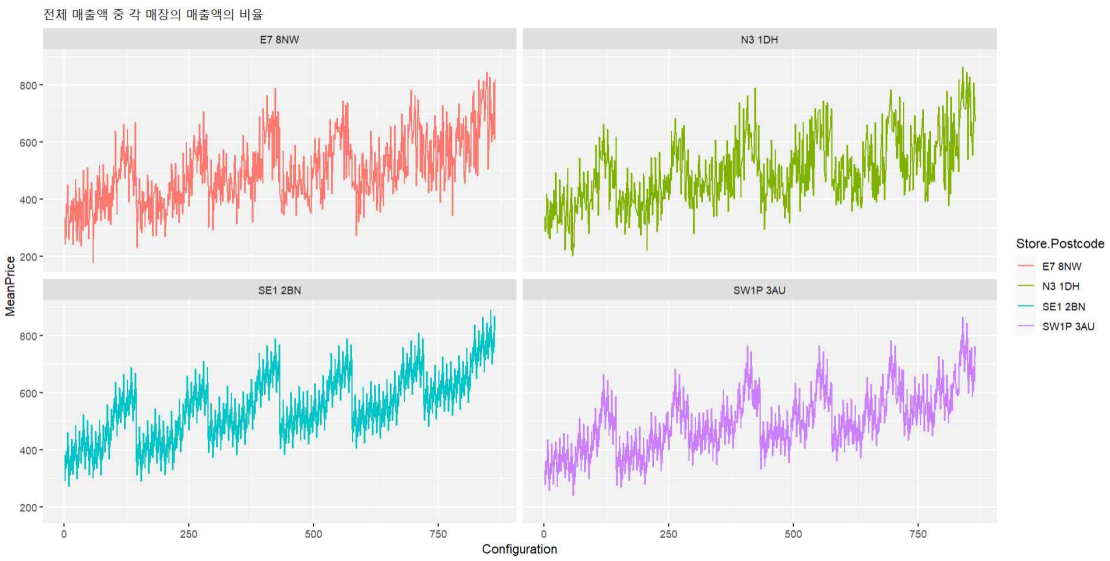
- 위 그래프는 각 매장의 매출액의 비율을 treemap으로 나타낸 그래프이다. 공간을 많이 차지할수록 진한 녹색을 띠고 적게 차지할수록 노란색을 띤다. 따라서, 이 그래프를 통해 앞서 얻었던 비율과 동일한 결과를 얻을 수 있다.

③ 원그래프



- 위 그래프도 전체 매출액 중 각 매장의 매출액의 비율을 한눈에 알아볼 수 있게 원그래프로 표현한 것이다. 앞서 나왔던 결과와 같이 SW1P 3AU라는 매장의 매출이 가장 높고 'SE1 2BN', 'SW1V 4QQ'순으로 내려간다.

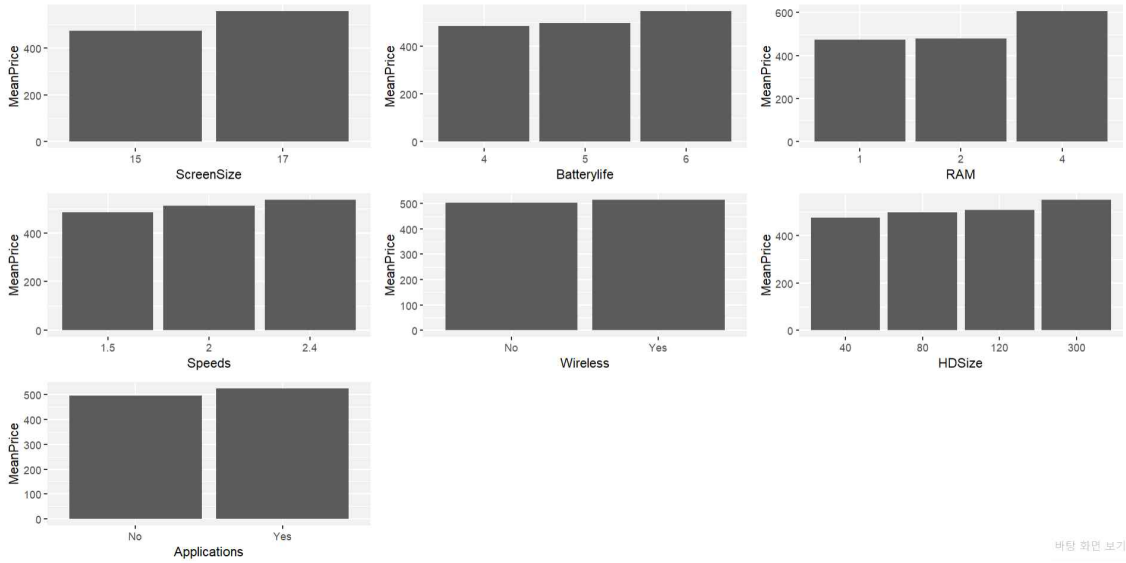
ii. 이 연관성은 컴퓨터 사양에 따라 영향을 받는가?



- 위 그래프의 위쪽 그래프 2개는 총 매출액이 낮은 매장 2곳이고 아래쪽 그래프는 총 매출액이 높은 매장 2곳이다. 앞에 그래프에서 매출액이 가장 작은 매장으로 나온 S1P 3WU는 규모가 작고 결측치가 많아 분석에서 제외하였다. 네 개의 그래프 모두 Configuration의 값이 0~500일 때에는 비슷한 경향을 보이지만 500이상일 때에는 아래 두 그래프가 위에 있는 두 그래프보다 향상하는 경향이 더 잘 보인다. 따라서, 가게 매출액이 컴퓨터 사양에 영향을 받는다고 할 수 있다.

d. 노트북 사양에 관한 질문

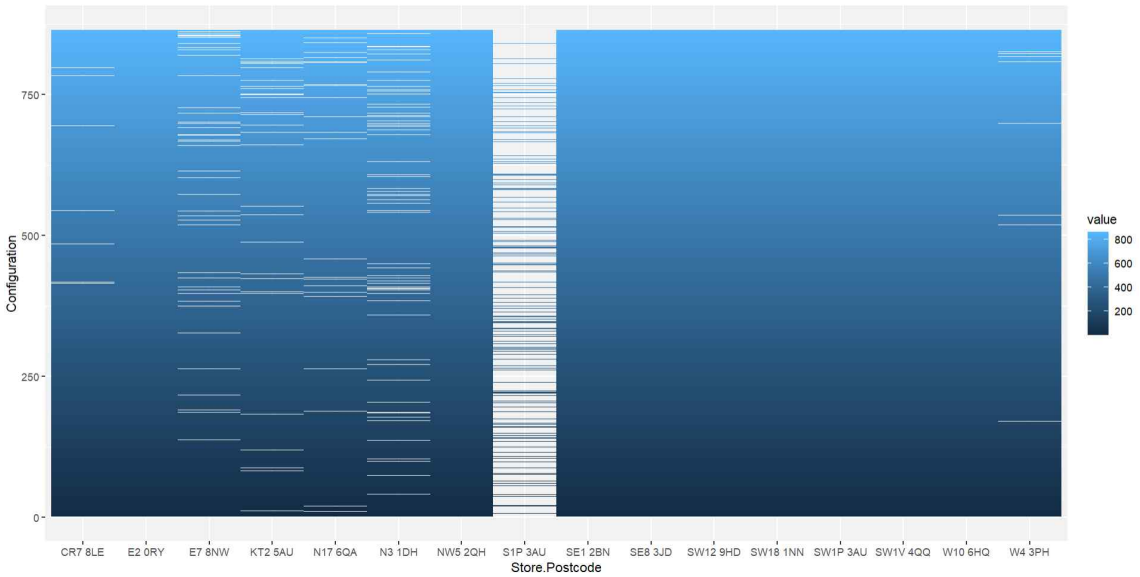
i. 각 노트북 사양의 상세한 내용은 무엇인가? 이것은 판매가격과 어떤 관련성이 있는가?



- 위 그래프는 각 노트북 사양에 따른 평균 소매 가격을 나타낸 막대 그래프이다. 겉보기에는 차이가 별로 나지 않아보이지만 첫 번째 그래프를 보면 ScreenSize가 17inch인 것은 500 후반이고 15inch는 400후반인 것을 보아 2배보다 100정도 큰 차이가 난다. 다른 그래프들도 노트북 사양의 크기가 커질수록(좋아질수록) 평균소매가격이 늘어난다는 것을 알 수 있고 'Bundled.Applications.'과 'Integrated.Wireless.'는 없는 것보다 있는 것이 평균 소매가격이 더 크다.

ii. 각 매장들이 모든 노트북 사양을 판매하는가?

- 아래 그래프는 모든 노트북 사양에 대한 Heatmap 그래프를 그린 것이다. 그래프에서 색깔이 없는 빈 공간은 결측값을 나타내므로 매장 'S1P 3AU'에서 가장 많은 결측치를 나타낸다. 따라서, 각 매장들은 모든 노트북 사양을 판매하지는 않는다.


Configuration
750
500
250
0
CR7 8LE
E2 0RY
E7 8NW
KT2 5AU
N17 6QA
N3 1DH
NW5 2QH
S1P 3AU
SE1 2BN
SE8 3JD
SW12 9HD
SW18 1NN
SW1P 3AU
SW1V 4QQ
W10 6HQ
W4 3PH
Store.Postcode
value
800
600
400
200