M45-retail-pkc

LS

2018-07-21

- 6. Discount가 많을 수록 매출이 늘어나는가?
- 7. 지역별로 가장 많이 팔리는 상품은 무엇인가? (ok)

```
library(readxl)
dataset <- read_excel("retail.xlsx")
```

6. Discount가 많을 수록 매출이 늘어나는가?

Background & Strategy

1. 가장 먼저 떠올릴 수 있는 생각은 Discount 와 Sales 간의 관계를 파악해 보는 것입니다. 상관계수 같은 수치를 볼 수도 있고 x, y 축에 시각화 해 볼 수도 있을 것 같았습니다. `reail` 데이터셋을 관찰해보면 일단 위로 매출을 집계한다면 비는 날짜가 생기거나 편차가 큰 편이었습니다. 그래서 월 단위 집계를 하고 월간 매출 할인액 합계 변수를 생성했습니다. 이렇게 생성한 월간 합계 값의 관계를 비교해 보고자 했습니다.
2. `retail` 데이터셋의 `Discount` 변수는 할인의 율을 나타내고 있으며, 연속형이 아닌 0.1 0.15 0.2 같은 factor 형이었습니다. 이는 특정 할인 level 에 대한 특성을 관찰해 볼 수도 있고 level 끼리 묶어서 비교해 보기에도 좋을 것 같았습니다.

Tasks Specification

1. 월단위 매출, 할인매출 변수를 생성하고 월간 매출액과 할인매출액을 scatter plot 으로 확인해 본다.
2. 할인 level 을 low, medium, high 로 그룹화하는 변수를 생성하고, 연도별, 월별 할인 효과가 매출에 미치는 영향을 trend 로 확인해 본다.

Sim: 시간별로 범주화하고, 할인 level 이라는 특성을 주어서 범주화하여 데이터를 이해하겠다는 전략 아주 좋습니다.

Task 1

먼저 `Discount` 변수를 살펴보니 대략 절반 정도의 거래건에서 Discount 가 되었으며, 그 중 70% 정도는 20% 정도의 할인율로 판매되었음을 알 수 있었습니다.

```
table(dataset$Discount)
```

```
##
##    0  0.1 0.15  0.2  0.3 0.32  0.4 0.45  0.5  0.6  0.7  0.8
## 4798   94   52 3657  227   27  206   11   66  138  418  300
```

할인매출액 변수를 생성하고 할인 level 을 0% ~ 20% 까지는 'low', 20% ~ 50% 까지는 'medium', 그 이상은 'high' level 로 구분했습니다. `year` 와 `month` 변수를 생성했습니다.

Sim: 할인 level 이 0 에서 0.8 까지 분포하는데 "low", "medium", "high" 로 구분하여 categorical 변수를 정의하는 방식은 데이터의 특성을 묶음으로 처리하려는 좋은 의도입니다. 다만, 전체 경우의 절반을 차지하는 0 을 따로 분류하는게 어떨까 합니다. 예를 들면 "noSale" 로 정의하는 것도 괜찮아 보입니다. 현재의 경우에는 "low" 에 전체 케이스의 85% 가 포함됩니다. 그 보다는 "noSale" 에 50%, "low" 에 35% 를 포함시키는게 어떨까 싶습니다.

```
m6 <- dataset
m6$DiscountAmount <- m6$Sales * m6$Discount
m6$DiscountLevel <-
  ifelse(m6$Discount == 0, "non",
         ifelse(m6$Discount <= 0.2, "low",
                ifelse(m6$Discount <= 0.5, "medium", "high")))
m6$year  <- substr(m6$`Order Date`, 1, 4)
m6$month <- substr(m6$`Order Date`, 6, 7)
```

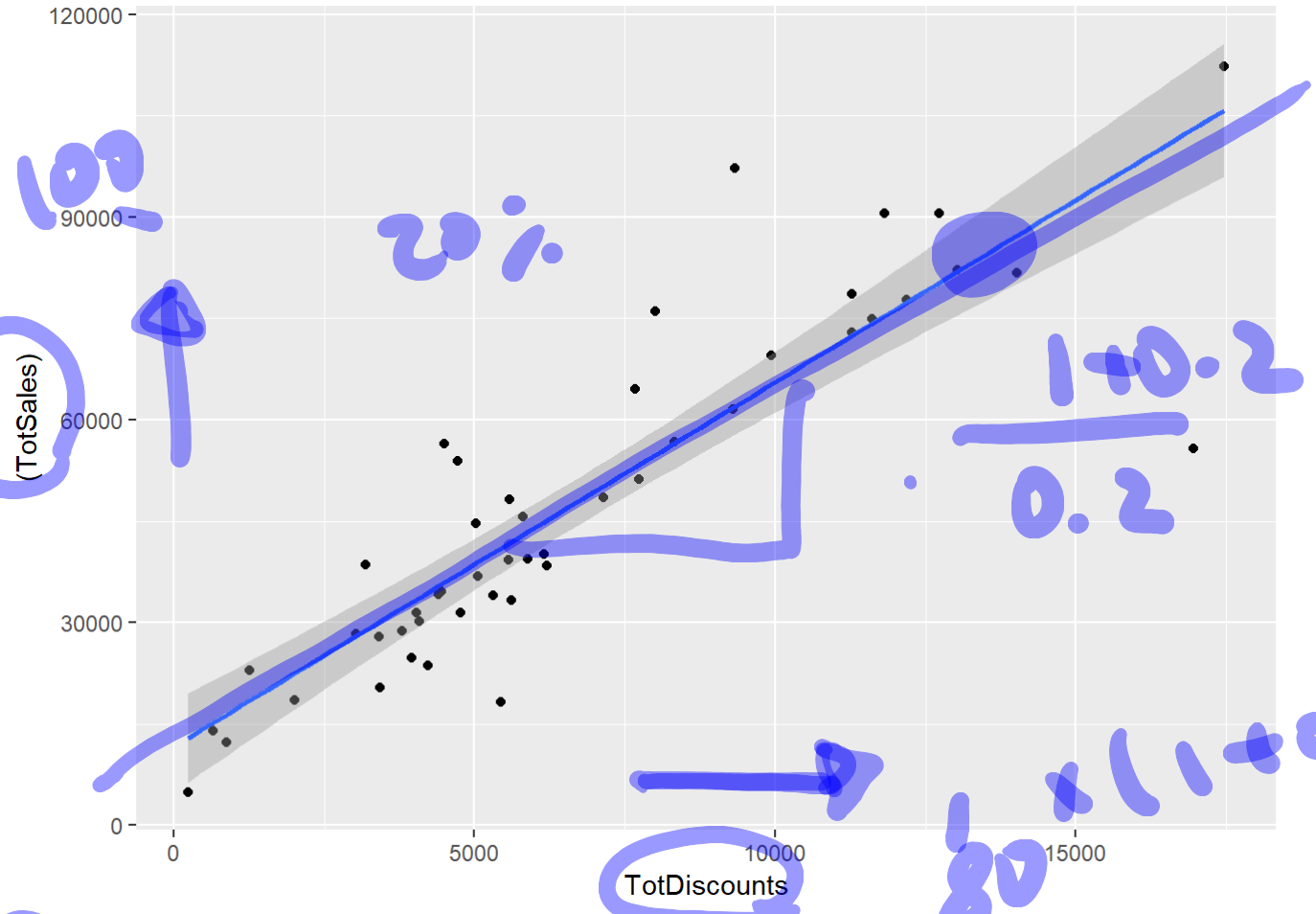
월별 매출액 합계와 할인매출액 합계를 다음과 같이 구했습니다.

```
task1 <- m6 %>%
  select(year, month, Sales, DiscountAmount) %>%
  group_by(year, month) %>%
  summarise(TotSales = sum(Sales),
            TotDiscounts = sum(DiscountAmount))
print(task1)
```

```
## # A tibble: 48 x 4
## # Groups:   year [?]
##    year  month TotSales TotDiscounts
##    <chr> <chr>    <dbl>        <dbl>
##  1 2011  01       13947         654.
##  2 2011  02        4809         235.
##  3 2011  03       55689       16954.
##  4 2011  04       28294        3039.
##  5 2011  05       23648        4233.
##  6 2011  06       34598        4447.
##  7 2011  07       33948        5311.
##  8 2011  08       27908        3413.
##  9 2011  09       81787       14023.
## 10 2011  10       31448        4031.
## # ... with 38 more rows
```

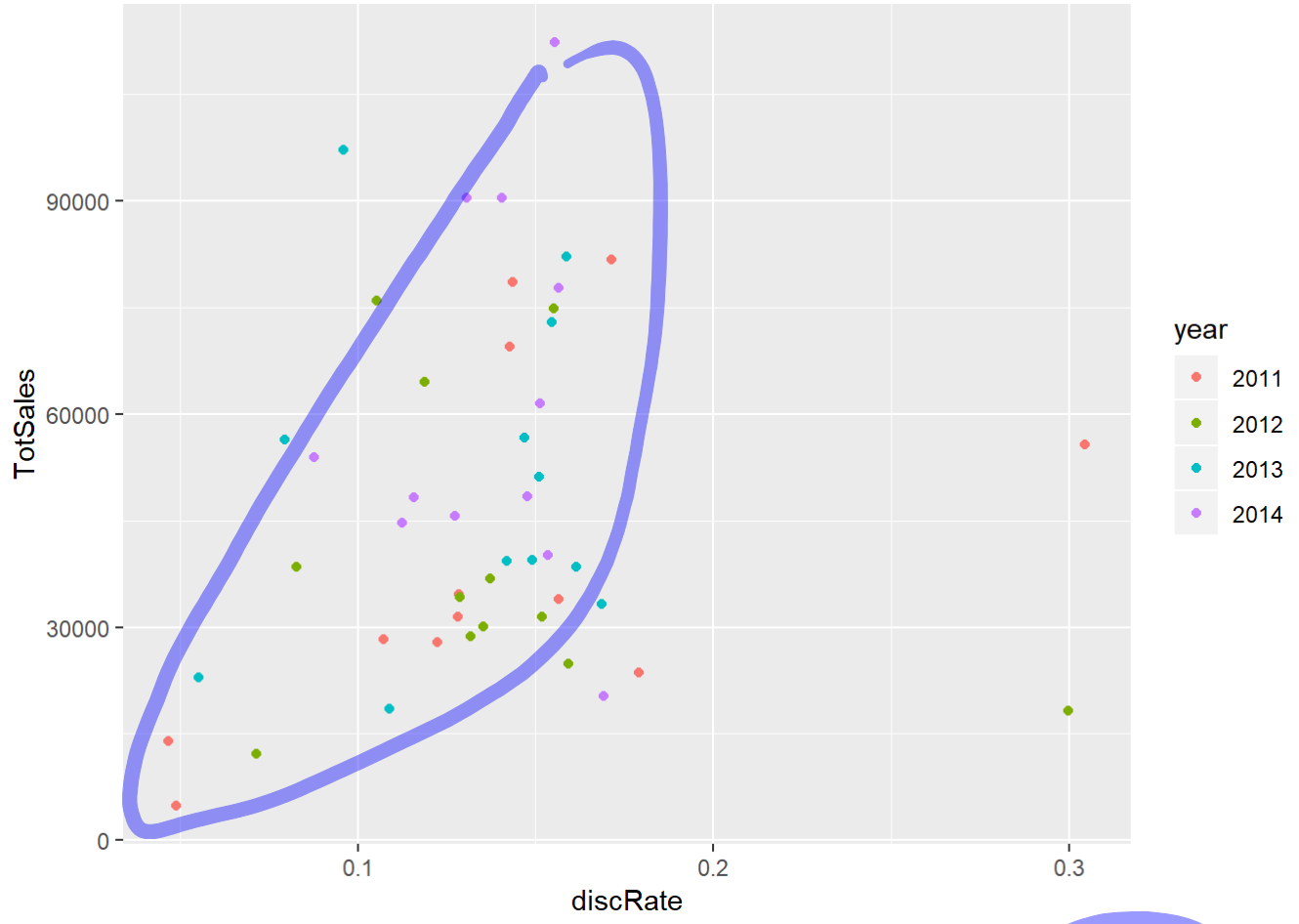
이제 할인매출액과 매출액의 관계를 plot 해 보았습니다. 몇몇 outlier 를 제외하고는 "할인을 많이 한 달에는 매출합계가 높았다." 라는 주장을 할 수 있을 것 같습니다.

```
ggplot(task1, aes(x= TotDiscounts, y= (TotSales))) +  # 경향성이 있음!
  geom_point() +
  geom_smooth(method = "lm")
```



Sim: `TotSales` 는 매출액을 의미하고 `TotDiscounts` 는 총 할인액을 의미합니다. 각 상품에 대해서 `listing price` (정가) 와 `discount price` (할인가) 가 있습니다. 예를 들어 정가가 100불짜리인 물건을 20% 할인하여 80불에 판매한다면 할인액은 20불이고, 모든 상품에 대해서 총합을 구하면 이것이 `TotDiscounts` 가 됩니다. 즉, `TotSales` 와 할인율 (Discount Rate)를 곱하면 `TotDiscounts` 가 됩니다. 그러므로 위에 그려진 scatter plot은 기울기가 1/discount rate입니다. 원 질문인 "Discount가 많을 수록 매출이 늘어나는가?"의 의도와 가까운 답변은 discount rate이 늘어가면 `TotSales` 가 늘어가는 것을 궁금해 하는 것으로 보입니다.

```
a <- task1 %>% mutate(discRate = TotDiscounts/TotSales)
ggplot(data = a, aes(x = discRate, y = TotSales, color = year)) +
  geom_point()
```



Sim: 어떤가요? `discRate` 이 올라가면 `TotSales` 가 높아지나요? 대체로 그런것 같지만 확신할 수준은 아닌 것 같습니다. 확신할 수 없는 이유는 무엇인가요? 확신할 수 있는 아웃풋을 찾으려면 어떻게 해야할까요?

1. *우선 확신할 수 없는 이유는 관찰값의 갯수가 적기 때문입니다.*
2. *현재 4년간의 데이터에 대해서 12개월이므로 48개의 점을 찍습니다.*
3. *48개의 점으로 결론이 내려지지 않는다면, 주간으로 나누어서 관찰값을 구한다면 이는 약 200개 정도의 점이 되므로 결론을 내리기가 더 수월해질 것입니다.*
4. *한 가지 유의해야 할 점은 소비시장에서 업계가 공통으로 가지는 세일기간이 있다는 것입니다. 지난 시간에 살펴본 바와 같이 미국 소비시장에서는 특히 연말에 매출이 많은데 그 기간이 세일 기간입니다.*
5. *데이터 분석의 결론은 올바른 의사결정으로 이어져야 하는 경우가 많습니다. 할인율이 높을 때 매출이 많았다는 것이 할인율을 높이면 매출이 높아진다는 것을 의미하는 것은 아닙니다. 모두가 세일을 하는 할인 시즌을 감안해야 합니다.*
6. *이를 위해서는 데이터셋 전체를 regular season과 promotion season으로 나누어서 분석을 할 필요가 있을 것 같습니다. "black friday", "x-mas", "new year"등의 굵직굵직한 이벤트의 관찰값을 제외하던지, 아니면 4Q만 제외하고 분석하는 것도 가능할 것 같습니다.*
7. *여기 예제에서 보다시피 데이터 사이언스 과정에서 논리적인 접근이나 구현 뿐 아니라 도메인 지식은 매우 중요합니다. 항상 비지니스 프로세스의 본질과 현상, 협업해야 할 대상이 되는 동료들을 이해하려는 꾸준한 노력이 필요하고, 이런 이유 때문에 이런 분석 직무가 조직내에서 가치가 높고 중요하다고 생각됩니다.*

하지만 좀 부족한 면이 있었습니다. 이 업체의 경우 총 매출건수 중 절반 가량을 상시적인 할인으로 잡는 것으로 보이기 때문에 매출이 높은 달은 당연히 할인매출의 합계도 높을 것 같다는 생각입니다. 게다가 가구나 Technology 처럼 건단가가 높은 제품들이기 때문에 의미있는 분석이라는 결론도 내리기 힘들 것 같습니다.

Sim: 저는 시간의 변화에 따라서 특성이 바뀔 것이라는 생각을 가지고 위의 문단들을 작성했습니다. 시간의 변화에 따른 변화는 제가 항상 생각하고 있는 일반적인 내용이라 여기에 적용을 해보았습니다. 박광춘님은 Category 별로 현상이 다르다라는 생각을 가지고 이 문제에 접근해 생각을 찾아냈습니다. 이 문제에 대해서 깊이 생각해 보셨음을 보여주는 대목이라 생각됩니다.

따라서 생각한 전략은 월간 매출액 대비 할인'율'을 생성하고, 전년동월 대비 1) 매출 증감 2) 할인폭 증감 의 관계를 보는 것이었습니다.

`task1` 의 테이블을 이렇게 생겼습니다. 여기서 전년도 대비한 매출액의 증감, 할일율의 증감을 구해야 하는데, 생각을 해 보니 쉽지가 않았습니다.

```
print(task1)
```

```
## # A tibble: 48 x 4
## # Groups:   year [4]
##    year  month TotSales TotDiscounts
##    <chr> <chr>    <dbl>        <dbl>
##  1 2011  01       13947         654.
##  2 2011  02        4809         235.
##  3 2011  03       55689       16954.
##  4 2011  04       28294        3039.
##  5 2011  05       23648        4233.
##  6 2011  06       34598        4447.
##  7 2011  07       33948        5311.
##  8 2011  08       27908        3413.
##  9 2011  09       81787       14023.
## 10 2011  10       31448        4031.
## # ... with 38 more rows
```

수업에서 배운 `Tidy` data 만들기 방법을 사용해 보기로 했습니다. 우선 `TotSales` 를 `value` 로 두고 월별, 연도별로 `spread` 하여 다음 테이블을 구했습니다.

```
task1_a <- task1 %>%
  select(year, month, TotSales) %>%
  spread(key= "year", value= "TotSales") %>%
  rename(raw_2011 = '2011',
         raw_2012 = '2012',
         raw_2013 = '2013',
         raw_2014 = '2014')
print(task1_a)
```

```
## # A tibble: 12 x 5
##    month raw_2011 raw_2012 raw_2013 raw_2014
##    <chr>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 01       13947    18172    18548    44708
##  2 02        4809    12214    22868    20288
##  3 03       55689    38472    51186    53916
##  4 04       28294    34198    39255    40116
##  5 05       23648    30137    56697    45650
##  6 06       34598    24800    39438    48259
##  7 07       33948    28765    38441    48430
##  8 08       27908    36899    33269    61527
##  9 09       81787    64601    72913    90508
## 10 10       31448    31406    56468    77804
## 11 11       78634    75979    82193   112335
## 12 12       69545    74917    97247    90475
```

Sim: Hadly Wickham의 정의에 따르면 tidy 데이터 셋은 각 row가 하나의 관찰값을 의미해야 합니다. 원래 제공된 `dataset` 이 tidy하며 현재 작업하신 내용은 데이터셋을 목적에 맞게 untidy하게 만든 것에 해당합니다. 일반적으로 생각하기 어려운 방법인데도 이처럼 tidy set을 untidy가 바꾸는 접근을 생각해 내신게 놀랍습니다.

그 다음 2011 년도는 전년도 데이터가 없으므로 제외시키고, 2012, 2013, 2014 년도의 전년 동월 데이터와 차이를 구해 이 컬럼만 다시 선택했습니다.

```
task1_b <- task1_a %>%
  mutate( '2012' = raw_2012 - raw_2011,
          '2013' = raw_2013 - raw_2012,
          '2014' = raw_2014 - raw_2013) %>%
  select(month, '2012', '2013', '2014')
print(task1_b)
```

```
## # A tibble: 12 x 4
##    month `2012` `2013` `2014`
##    <chr>  <dbl>  <dbl>  <dbl>
##  1 01      4225    376  26160
##  2 02      7405  10654  -2580
##  3 03    -17217  12714   2730
##  4 04      5904   5057    861
##  5 05      6489  26560 -11047
##  6 06     -9798  14638   8821
##  7 07     -5183   9676   9989
##  8 08      8991  -3630  28258
##  9 09    -17186   8312  17595
## 10 10       -42  25062  21336
## 11 11     -2655   6214  30142
## 12 12      5372  22330  -6772
```

그다음 다시 `gather` 로 원래 형태로 만들어 주어 결과적으로 년, 월별 전년도 대비 매출액 증감을 구했습니다.

```
task1_c <- task1_b %>%
  gather(colnames(task1_b)[-1],
         key = "year",
         value = "Sales_growth")
print(task1_c)
```

```
## # A tibble: 36 x 3
##    month year  Sales_growth
##    <chr> <chr>        <dbl>
##  1 01    2012          4225
##  2 02    2012          7405
##  3 03    2012        -17217
##  4 04    2012          5904
##  5 05    2012          6489
##  6 06    2012         -9798
##  7 07    2012         -5183
##  8 08    2012          8991
##  9 09    2012        -17186
## 10 10    2012           -42
## # ... with 26 more rows
```

Sim: tidy한 데이터 셋을 untidy하게 바꾸어 작업하고 다시 tidy하게 바꾸는 작업을 하였습니다. `Sale_growth` 라는 변수이름을 정의할때 작년 대비라는 의미가 파악되게 이름을 지으려면 좋겠는데 마땅한 이름을 짓기가 어려운 것 같습니다. "전년 동월 대비 매출 증감"을 구글 번역은 "YoY increase / decrease"라고 합니다. `salesGrowthYoY` 라고 지으면 어떨까요? `YoY` 는 Year over Year라고 하여 전년대비라는 뜻입니다.

이제 년, 월별 할인율의 증감을 구하면 되는데, 위의 매출 증감과 똑같은 방식이므로 `dplyr` 로 한번에 구했습니다.

그런 다음 두 테이블을 join 하여 원하는 형태의 테이블을 얻었습니다.

```
task1$DCRate <- round(task1$TotDiscounts / task1$TotSales, 4) * 100

task1_d <- task1 %>%
  select(year, month, DCRate) %>%
  spread(key= "year", value= "DCRate") %>%
  rename(raw_2011 = '2011',
         raw_2012 = '2012',
         raw_2013 = '2013',
         raw_2014 = '2014') %>%
  mutate( '2012' = raw_2012 - raw_2011,
          '2013' = raw_2013 - raw_2012,
          '2014' = raw_2014 - raw_2013) %>%
  select(month, '2012', '2013', '2014') %>%
  gather(colnames(task1_b)[-1],
         key = "year",
         value = "DCRate_growth")
print(task1_d)
```

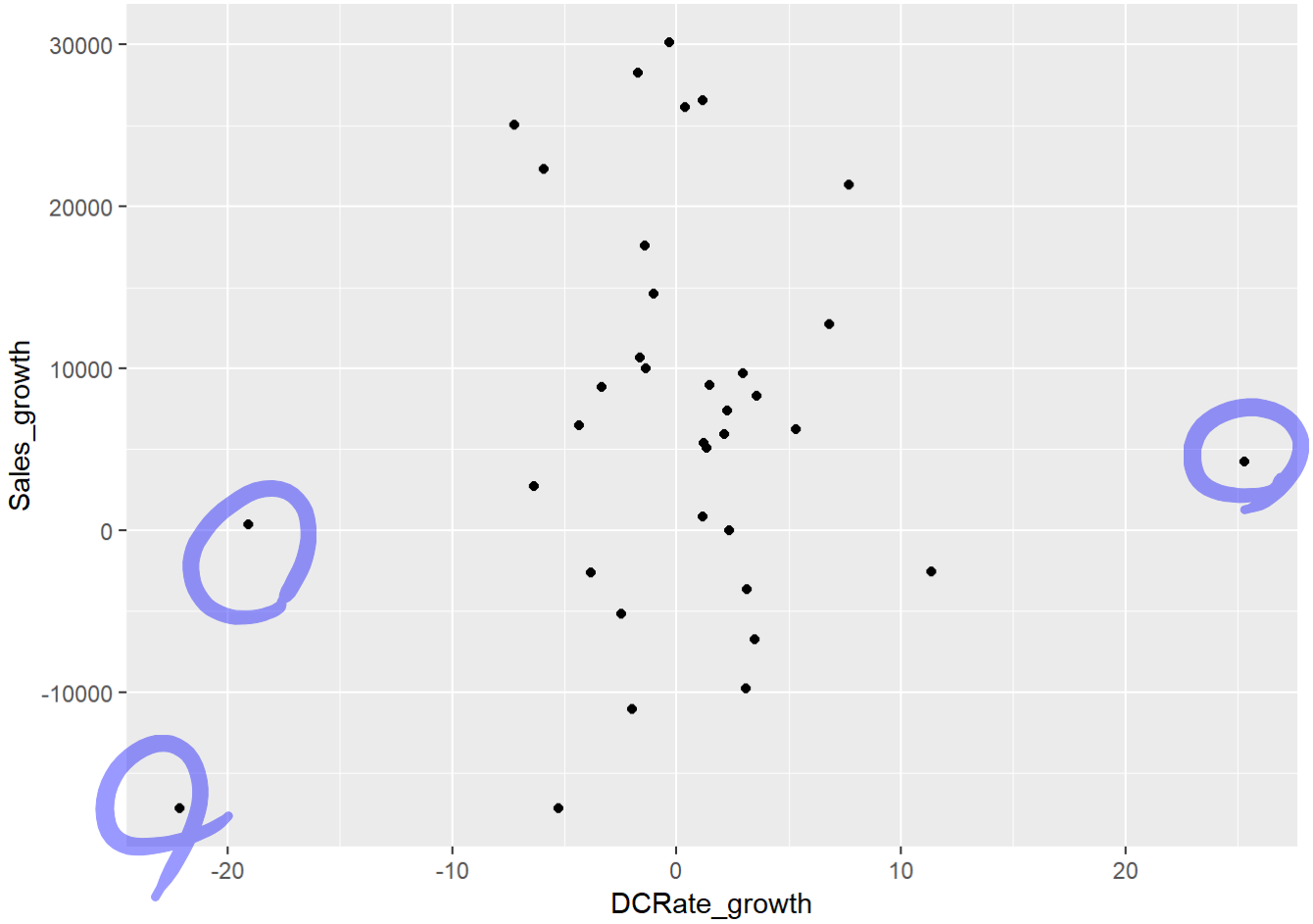
```
## # A tibble: 36 x 3
##    month year  DCRate_growth
##    <chr> <chr>         <dbl>
##  1 01    2012          25.3
##  2 02    2012           2.27
##  3 03    2012         -22.2
##  4 04    2012           2.12
##  5 05    2012          -4.35
##  6 06    2012           3.09
##  7 07    2012          -2.46
##  8 08    2012           1.49
##  9 09    2012          -5.26
## 10 10    2012           2.35
## # ... with 26 more rows
```

```
task1_e <- inner_join(task1_c, task1_d, by= c("year", "month"))
print(task1_e)
```

```
## # A tibble: 36 x 4
##    month year  Sales_growth DCRate_growth
##    <chr> <chr>        <dbl>         <dbl>
##  1 01    2012          4225          25.3
##  2 02    2012          7405           2.27
##  3 03    2012        -17217         -22.2
##  4 04    2012          5904           2.12
##  5 05    2012          6489          -4.35
##  6 06    2012         -9798           3.09
##  7 07    2012         -5183          -2.46
##  8 08    2012          8991           1.49
##  9 09    2012        -17186          -5.26
## 10 10    2012            -42           2.35
## # ... with 26 more rows
```

드디어 두 변수간 scatter plot 을 그려 봤습니다. 이 결로과 봤을 때, 직전년도 대비하여 월간 할인액을 세게 올린다고 해서 매출이 눈에띄게 오른다는 말은 할 수 없을 것 같습니다.

```
ggplot(task1_e, aes(x= DCRate_growth, y= Sales_growth)) +
  geom_point()
```



Sim: 제가 개인적으로 ggplot을 수업을 하고 나서 생기게 된 변화는 `aes` 를 지정할 때, color feature를 반드시 사용하려고 한다는 점입니다. color를 안쓰면 뭔가 아까운 생각이 들더군요. 위의 그림에서는 20%/-20% growth가 식별되는 color feature가 궁금합니다.

이번에는 같은 자료를 월별로 데이터가 보이도록 살펴보고자 했습니다. 선생님께 배운 `grid.arrange()` 로 두 자료를 비교가 가능한 배치를 만들었습니다. 2012년 01월의 경우 할인을 전년대비 쎄게 했는데 매출 증가는 미미한 정도였습니다. 2014년 01월의 경우 2012년에는 할인을 크게 하고 2013년에는 평소 수준으로 다시 줄였으므로 2014년 할인은 평소 수준인데 매출은 크게 증가했습니다. 이를 보면 할인 강도 보다는 그냥 seasonal 한 월 매출이 나오는 것 같습니다.

```
g_1 <- ggplot(task1_e, aes(x= month)) +
  geom_bar(aes(y= Sales_growth, fill=Sales_growth >= 0), stat= "identity") +
  facet_grid(rows= vars(year)) +
  scale_fill_discrete(guide=FALSE)

g_2 <- ggplot(task1_e, aes(x= month)) +
  geom_bar(aes(y= DCRate_growth, fill=DCRate_growth >= 0), stat= "identity") +
  facet_grid(rows= vars(year)) +
  scale_fill_discrete(guide=FALSE)

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(g_2, g_1, nrow= 1, ncol= 2,
             top= "Annualy Discount and Sales Growth")
```

Task 2

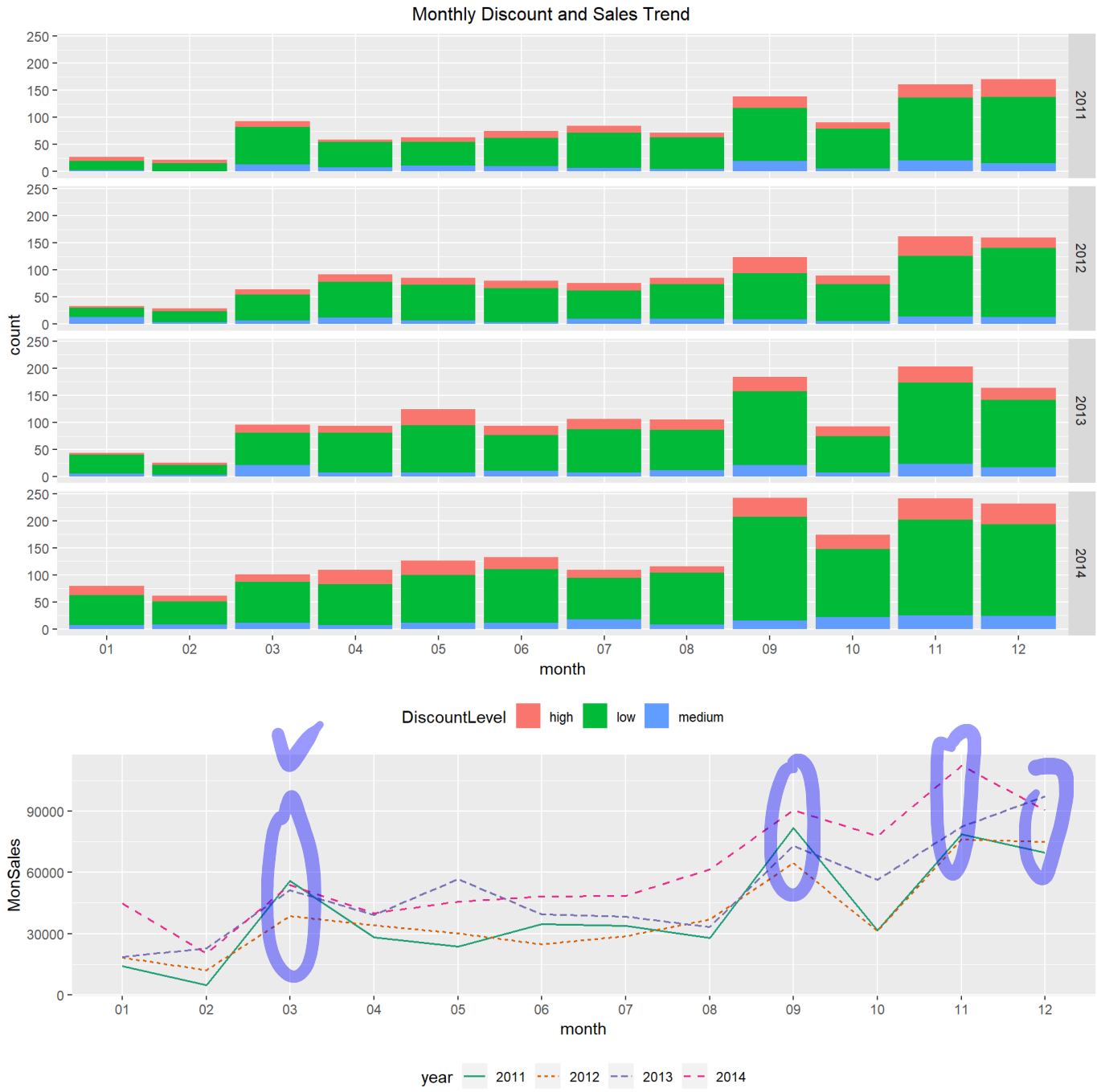
이번에는 월별 할인품목의 건수가 매출에 어떤 영향을 미쳤는지 trend 로 확인해 보고자 했습니다.

추가로 앞서 생성한 `DiscountLevel` 변수로 구분하여 할인 level 이 차지하는 비중도 볼 수 있도록 했습니다.

```
task2_a <- m6 %>%
  filter(DiscountLevel %in% c("low", "medium", "high")) %>%
  ggplot(aes(x= month)) +
  geom_bar(aes(fill= DiscountLevel)) +
  facet_grid(rows= vars(year)) +
  theme(legend.position = "bottom")

task2_b <- m6 %>%
  group_by(year, month) %>%
  summarise(MonSales = sum(Sales)) %>%
  ggplot(aes(x= month, y= MonSales, color= year, group= year)) +
  geom_line(size= 0.6, aes(linetype= year)) +
  scale_color_brewer(palette="Dark2") +
  theme(legend.position = "bottom")
```

```
library(gridExtra)
grid.arrange(task2_a, task2_b, nrow=2, heights= 2:1,
             top= "Monthly Discount and Sales Trend")
```



Sim: 덕분에 `heights= 2:1` 과 같은 parameter를 사용가능한 것을 알게 되었습니다. 앞으로 수업 자료 준비에도 큰 도움이 되지 않을까 생각됩니다. 감사합니다!

- 4년간의 그림을 살펴보면 이 회사는 3월, 9월, 11월, 12월이 되면 할인을 세게 주는 것 같습니다. 할인 강도가 high 로 주는 경우는 9~12 월 연말에 집중되어 있는 것을 볼 수 있습니다.

*Sim: 아주 흥미로운 결론이네요. 사내 내부적으로 분기별로 매출액을 가지고 보너스를 준다던지 하는 규정이 있는 걸까요?**

- 2013년도 5월에는 전월대비 higt 강도 할인 품목을 많이 늘려서 매출 상승 효과를 많이 봤고, 다음 연도에는 전체적인 할인품목 수만 늘렸더니 매출이 오르긴 했지만 전년도 보다는 상승이 미미했습니다. 이를 보면 할인 강도가 high 인 품목을 늘릴 때는 매출 상승 효과가 괜찮은 편인 것 같습니다. (당연한 말이긴 합니다.)
- 2013년 12월의 경우 예산이 모자랐는지 할인 전월 대비 감소시켰는데 매출은 전월대비 올랐습니다. 이를 보면 할인 외에 매출에 영향을 미치는 변수들이 많이 있음을 짐작할 수 있었습니다. 따라서 단순히 할인의 강도나 양이 셀 때는 매출에 긍정 영향을 미치긴 하지만 이 하나의 변수가 매출 변동을 정량적으로 설명하긴 부족하다는 생각으로 마무리 했습니다.

Sim: 아마도 12월 세일의 경우에는 소비자 경기랑도 연관이 있을수 있지 않나 생각이 드네요. 그런데 그때가 크게 불황이었는지는 모르겠습니다. 소비자 경기 지수라던지 정부 발표자료 보다는 아마존 같은 유사한 비지니스 이면서 큰 기업의 재무재표와 비교를 해보는 것도 가능한 접근이라 생각됩니다.

7. 지역별로 가장 많이 팔리는 상품은 무엇인가? (ok)

Background & Strategy

- `retail` 데이터셋에는 `Country`, `Region`, `State`, `City` 단위로 배송 지역을 구분하고 있습니다.
- 또한 `Product Name` , `Sub-Category` , `Category` 단위로 품목을 구분하고 있습니다.

이 중 분석 단위를 무엇으로 결정할 지는 1) 데이터를 개별 그룹으로 묶었을 때 레코드의 숫자가 일정 수준 이상이어야 할 것, 2) 결과를 표로 나타낼 지, plot 으로 표현할 지 를 고려하여야 할 것입니다. 가령 `City` 단위가 너무 많다면 지역명을 x 축 혹은 y축에 배치하기 힘들어 지겠지요.

Sim: 아주 좋습니다. factor 변수에서 가질수 있는 값의 갯수를 고려해서 아웃풋을 생각해야 합니다. plotting은 지도에 나타내는 것도 좋은 방법이구요.

Tasks Specification

1. 지역 단위는 `State` , `City` 순으로, 판매 단위는 `Sub-Category` , `Product Name` 순으로 Aggregation 하여 결과를 비교해 본다.
2. 적절한 단위를 선택하여 결과를 표현한다.

Task 1

`City` 단위로 묶어 record 의 수를 출력해 봅니다.
4 년간의 자료에서 10건 이하 판매량이 나오는 지역이 있다면 의미가 없는 내용입니다.
또한 City 단위로 묶을 경우 결과의 표현도 어려워 집니다.

```
dataset %>%
  group_by(Country, Region, State, City) %>%
  summarise(n= length(`Sub-Category`)) %>%
  arrange((n)) %>%
  head(10)
```

```
## # A tibble: 10 x 5
## # Groups:   Country, Region, State [1]
##    Country       Region  State    City                 n
##    <chr>         <chr>   <chr>    <chr>            <int>
##  1 United States Central Illinois Arlington Heights     1
##  2 United States Central Illinois Champaign             1
##  3 United States Central Illinois Danville              1
##  4 United States Central Illinois Glenview              1
##  5 United States Central Illinois Normal                1
##  6 United States Central Illinois Oak Park              1
##  7 United States Central Illinois Orland Park           1
##  8 United States Central Illinois Palatine              1
##  9 United States Central Illinois Romeoville            1
## 10 United States Central Illinois Tinley Park           1
```

`State` 단위에서 49 개 그룹이 생기므로 이것을 사용하기로 합니다.

```
dataset %>%
  group_by(Country, Region, State) %>%
  summarise(n= length(`Sub-Category`)) %>%
  arrange((n)) %>%
  head(10)
```

```
## # A tibble: 10 x 4
## # Groups:   Country, Region [3]
##    Country       Region  State                 n
##    <chr>         <chr>   <chr>             <int>
##  1 United States West    Wyoming               1
##  2 United States East    West Virginia         4
##  3 United States Central North Dakota          7
##  4 United States East    Maine                 8
##  5 United States East    District of Columbia 10
##  6 United States East    Vermont              11
##  7 United States Central South Dakota         12
##  8 United States West    Montana              15
##  9 United States West    Idaho                21
## 10 United States Central Kansas               24
```

그렇다면 판매 품목의 단위는 무엇으로 해야 하는가?

`Product Name` 과 `Sub-Category` 두 가지로 나누어 비교해 보기로 합니다.

```
task1_a <- dataset %>%
  group_by(Country, Region, State, `Product Name`) %>%
  summarise(n= sum(Quantity)) %>%
  arrange(desc(n))

task1_b <- dataset %>%
  group_by(Country, Region, State, `Sub-Category`) %>%
  summarise(n= sum(Quantity)) %>%
  arrange(desc(n))

print(head(task1_a))
```

```
## # A tibble: 6 x 5
## # Groups:   Country, Region, State [3]
##   Country    Region  State   `Product Name`                              n
##   <chr>      <chr>   <chr>   <chr>                                   <dbl>
## 1 United St~ East    New Yo~ Staple envelope                            44
## 2 United St~ Central Texas   Staples                                    39
## 3 United St~ West    Califo~ Eldon Shelf Savers Cubes and Bins           29
## 4 United St~ West    Califo~ 4009 Highlighters by Sanford               27
## 5 United St~ East    New Yo~ Dual Level, Single-Width Filing Carts      26
## 6 United St~ West    Califo~ Wilson Jones Clip & Carry Folder Binde~    26
```

```
print(head(task1_b))
```

```
## # A tibble: 6 x 5
## # Groups:   Country, Region, State [2]
##   Country       Region State      `Sub-Category`     n
##   <chr>         <chr>  <chr>      <chr>          <dbl>
## 1 United States West   California Paper           1091
## 2 United States West   California Binders         1057
## 3 United States West   California Furnishings      751
## 4 United States East   New York   Binders          696
## 5 United States West   California Phones           691
## 6 United States West   California Art              644
```

State 별 가장 많이 팔린 Sub-Catogory , Product Name 을 정리한 표를 다음과 같이 생성하여, 봅니다.

```
task1_a %>%
  filter(n == max(n)) %>%
  arrange(Country, Region, State, -n)
```

```
## # A tibble: 58 x 5
## # Groups:   Country, Region, State [49]
##    Country    Region  State   `Product Name`                              n
##    <chr>      <chr>   <chr>   <chr>                                   <dbl>
##  1 United St~ Central Illino~ Staples in misc. colors                    16
##  2 United St~ Central Indiana Belkin F9H710-06 7 Outlet SurgeMaster~     13
##  3 United St~ Central Iowa    Hon Olson Stacker Stools                   10
##  4 United St~ Central Kansas  Binney & Smith Crayola Metallic Color~      6
##  5 United St~ Central Kansas  "Executive Impressions 10\" Spectator~      6
##  6 United St~ Central Michig~ Avery Heavy-Duty EZD View Binder with~     15
##  7 United St~ Central Minnes~ Adams Telephone Message Book w/Freque~     11
##  8 United St~ Central Missou~ Black & Decker Filter for Double Acti~     10
##  9 United St~ Central Nebras~ Electrix Architect's Clamp-On Swing A~     14
## 10 United St~ Central North ~ Binney & Smith inkTank Desk Highlight~     10
## # ... with 48 more rows
```

- Product Name 으로 분류할 경우 State 가장 많이 팔린 품목이 두가지씩 나오는 경우가 있는데, tie 가 존재할 경우였습니다. 이 경우 항목이 너무 세부적이라는 판단으로 (4년간 판매량이 10건 이하) 단위를 Sub-Category 로 집계하는 것이 현명해 보입니다.

Sim: 네. 저도 저번주에 Product Name 으로 해보다가 같은 이유로 그냥 나중에는 Category 와 Sub-Category 로 했던것 같습니다. ㅋㅋㅋ

```
task1_b %>%
  filter(n == max(n)) %>%
  arrange(Country, Region, State, -n)
```

```
## # A tibble: 51 x 5
## # Groups:   Country, Region, State [49]
##    Country       Region  State        `Sub-Category`     n
##    <chr>         <chr>   <chr>        <chr>          <dbl>
##  1 United States Central Illinois     Binders          327
##  2 United States Central Indiana      Paper            100
##  3 United States Central Iowa         Binders           34
##  4 United States Central Kansas       Art               14
##  5 United States Central Michigan     Binders          192
##  6 United States Central Minnesota    Binders           55
##  7 United States Central Missouri     Paper             42
##  8 United States Central Nebraska     Paper             21
##  9 United States Central North Dakota Art               18
## 10 United States Central Oklahoma     Binders           42
## # ... with 41 more rows
```

- 51개 주에 대한 Best Selling 품목 정보를 얻었습니다.

Sim: 51개 주를 한번에 파악한다는 것은 어려운 일일수도 있다는 생각이 듭니다.

1) 51개 중에서 매출 비중이 높은 10개 정도를 추려내서 하는 방법도 가능할 것 같고 (이들 20% 주들이 전체 매출의 80%를 차지할 것이라고 누군가 20-80 법칙에서 그랬다고 합니다.) 아니면 `Region` 으로 나누어서 보는 것도 의미있을 수도 있다 생각합니다. 그런데 `Region` 으로 나누었을때 서부 중부 동부 남부가 소비패턴이 다를 것이다? 이는 봐야 알겠지만 `Region` 보다는 지역의 소득 수준이나 인구 구성에 연관이 더 많을 것 같기는 합니다. 이는 우리나라에도 마찬가지로 해당될 내용같습니다.

2) 아니면 지도에다가 매출액을 size로 하는 bubble chart를 그리는 것도 좋을 것이라 생각됩니다. bubble chart는 중요한게 크게 표현된다는 큰 매력이 있으니까요.