M45-retail (CASE STUDY)

LearningSpoonsR

2018-07-14

0. 시작하기

데이터의 확보

- Tableau라는 비쥬얼라이제이션 업체의 교육용 자료입니다.
- 데이터를 받았을 때에 가장 처음에 해야할 작업은 데이터가 어떻게 구성이 되어있는지를 확인하는 것입니다.
- 파일로 받은 경우에는 엑셀이나 메모장으로 데이터 파일을 열어 확인합니다.

잘 정리된 데이터

- **Tidy data**란 각 column이 변수에 해당하고 각 row가 1개의 관찰을 의미하는 잘 정리된 데이터 구조를 이야기 합니다.
- 이는 R자료 구조에서는 `data.frame`에 해당하고 파이썬에서는 `pandas.DataFrame`에 해당합니다.
- 시계열 자료인 경우에는 각각 `xts`이거나 `pandas.Series`입니다.

Preprocessing

- **Tidy data**가 아닌 경우에는 주어진 데이터를 우선 tidy data를 만들어야 하며, 이를 전처리(Preprocessing)과정이라고 합니다.
- 실제 업무에서 접하게 되는 데이터는 `tidy`한 경우가 잘 없습니다.
- 그렇기 때문에 데이터를 `tidy`하게 바꾸는 preprocessing과정이 전체 데이터 분석의 시간의 상당수를 차지하는 경우가 많습니다.

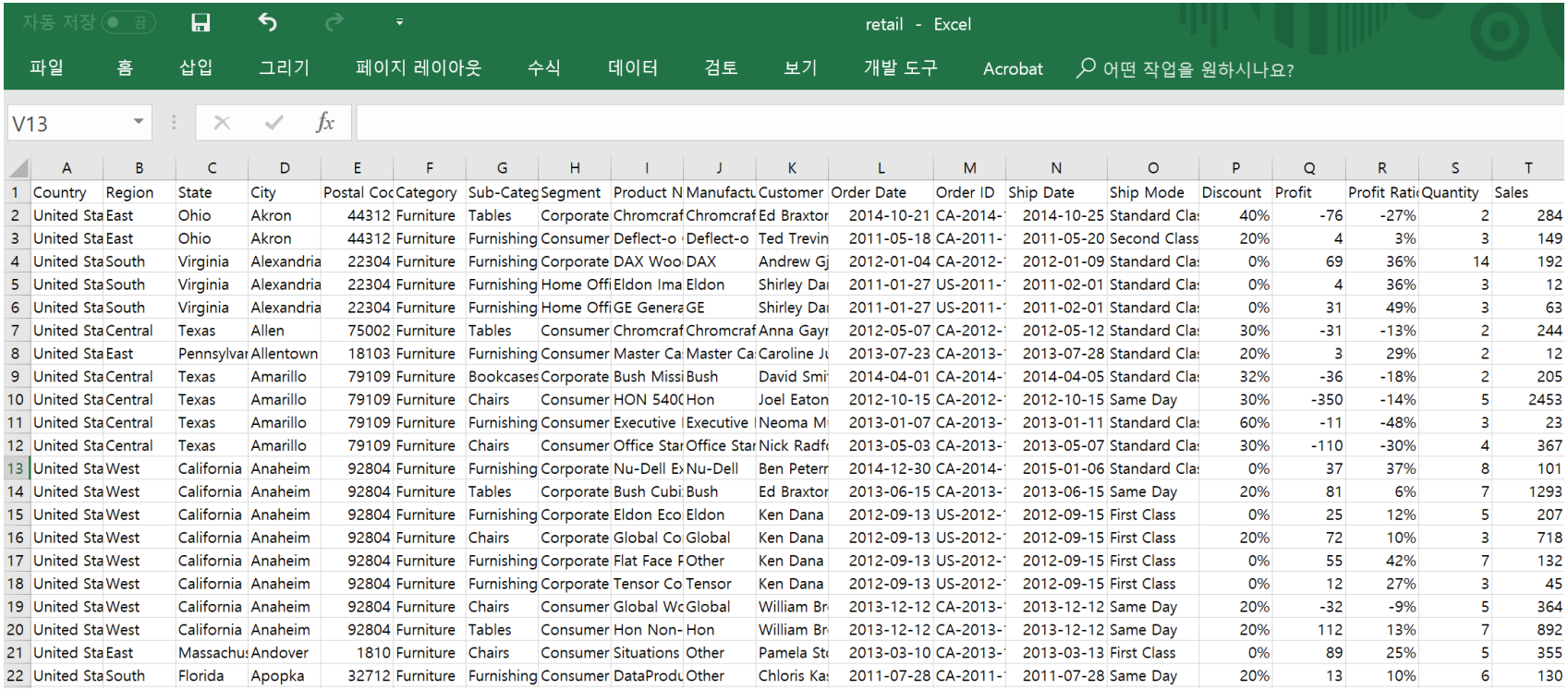
Preprocessing

- Preprocessing을 하는 과정은 다양하고 복잡하고 지저분한 raw 데이터 만큼이나 다양하고 복잡하고 세밀한 관찰력과 때로는 창의력을 요하기도 합니다.
- 이 과정이 흔히 노동집약적이기도 하지만, 이 과정이 데이터를 보는 눈과 데이터 구조를 이해하는 데에 큰 도움을 주기도 합니다.
- 또한 preprocessing과정을 열심히 수행하면서 얻게되는 경험과 프로그래밍 실력은 데이터 분석가로서의 자신감을 높여줍니다.

M51-tidyr에서는 Preprocessing과정에서

- 1. tidy하지 않은 데이터를 tidy하게 바꾸는 명령어와
- 2. 두 개 이상의 데이터 프레임을 합치는 법을 배웁니다.
- 1. 에서는 `tidyr` 패키지의 몇 가지 명령을 사용해서 `tidy`하지 않은 데이터 셋 중에서 전형적인 경우에 대처하는 법을 배웁니다.
- 2. 실제 데이터 분석을 하는 경우에 상당수는 1가지 데이터가 아닌 2개 이상의 데이터를 합치는 경우가 많습니다. 예를 들어 날씨에 따른 매출의 변화를 관찰하고 싶다면 기상청에서 제공하는 날씨데이터를 구해서 매출 데이터랑 결합을 시켜야 할 것입니다.

retail.xls



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Country	Region	State	City	Postal Coc	Category	Sub-Categ	Segment	Product N	Manufactu	Customer	Order Date	Order ID	Ship Date	Ship Mode	Discount	Profit	Profit Ratio	Quantity	Sales
2	United Sta	East	Ohio	Akron	44312	Furniture	Tables	Corporate	Chromcraf	Chromcraf	Ed Braxtor	2014-10-21	CA-2014-	2014-10-25	Standard Cla	40%	-76	-27%	2	284
3	United Sta	East	Ohio	Akron	44312	Furniture	Furnishing	Consumer	Deflect-o	Deflect-o	Ted Trevin	2011-05-18	CA-2011-	2011-05-20	Second Class	20%	4	3%	3	149
4	United Sta	South	Virginia	Alexandria	22304	Furniture	Furnishing	Corporate	DAX Woo	DAX	Andrew Gj	2012-01-04	CA-2012-	2012-01-09	Standard Cla	0%	69	36%	14	192
5	United Sta	South	Virginia	Alexandria	22304	Furniture	Furnishing	Home Offi	Eldon Ima	Eldon	Shirley Dai	2011-01-27	US-2011-	2011-02-01	Standard Cla	0%	4	36%	3	12
6	United Sta	South	Virginia	Alexandria	22304	Furniture	Furnishing	Home Offi	GE Genera	GE	Shirley Dai	2011-01-27	US-2011-	2011-02-01	Standard Cla	0%	31	49%	3	63
7	United Sta	Central	Texas	Allen	75002	Furniture	Tables	Consumer	Chromcraf	Chromcraf	Anna Gayr	2012-05-07	CA-2012-	2012-05-12	Standard Cla	30%	-31	-13%	2	244
8	United Sta	East	Pennsylvar	Allentown	18103	Furniture	Furnishing	Consumer	Master Ca	Master Ca	Caroline Ju	2013-07-23	CA-2013-	2013-07-28	Standard Cla	20%	3	29%	2	12
9	United Sta	Central	Texas	Amarillo	79109	Furniture	Bookcases	Corporate	Bush Missi	Bush	David Smi	2014-04-01	CA-2014-	2014-04-05	Standard Cla	32%	-36	-18%	2	205
10	United Sta	Central	Texas	Amarillo	79109	Furniture	Chairs	Consumer	HON 5400	Hon	Joel Eaton	2012-10-15	CA-2012-	2012-10-15	Same Day	30%	-350	-14%	5	2453
11	United Sta	Central	Texas	Amarillo	79109	Furniture	Furnishing	Consumer	Executive	Executive	Neoma M	2013-01-07	CA-2013-	2013-01-11	Standard Cla	60%	-11	-48%	3	23
12	United Sta	Central	Texas	Amarillo	79109	Furniture	Chairs	Consumer	Office Star	Office Star	Nick Radfe	2013-05-03	CA-2013-	2013-05-07	Standard Cla	30%	-110	-30%	4	367
13	United Sta	West	California	Anaheim	92804	Furniture	Furnishing	Corporate	Nu-Dell Ex	Nu-Dell	Ben Peterr	2014-12-30	CA-2014-	2015-01-06	Standard Cla	0%	37	37%	8	101
14	United Sta	West	California	Anaheim	92804	Furniture	Tables	Corporate	Bush Cubi	Bush	Ed Braxtor	2013-06-15	CA-2013-	2013-06-15	Same Day	20%	81	6%	7	1293
15	United Sta	West	California	Anaheim	92804	Furniture	Furnishing	Corporate	Eldon Eco	Eldon	Ken Dana	2012-09-13	US-2012-	2012-09-15	First Class	0%	25	12%	5	207
16	United Sta	West	California	Anaheim	92804	Furniture	Chairs	Corporate	Global Co	Global	Ken Dana	2012-09-13	US-2012-	2012-09-15	First Class	20%	72	10%	3	718
17	United Sta	West	California	Anaheim	92804	Furniture	Furnishing	Corporate	Flat Face F	Other	Ken Dana	2012-09-13	US-2012-	2012-09-15	First Class	0%	55	42%	7	132
18	United Sta	West	California	Anaheim	92804	Furniture	Furnishing	Corporate	Tensor Co	Tensor	Ken Dana	2012-09-13	US-2012-	2012-09-15	First Class	0%	12	27%	3	45
19	United Sta	West	California	Anaheim	92804	Furniture	Chairs	Consumer	Global Wc	Global	William Br	2013-12-12	CA-2013-	2013-12-12	Same Day	20%	-32	-9%	5	364
20	United Sta	West	California	Anaheim	92804	Furniture	Tables	Consumer	Hon Non-	Hon	William Br	2013-12-12	CA-2013-	2013-12-12	Same Day	20%	112	13%	7	892
21	United Sta	East	Massachus	Andover	1810	Furniture	Chairs	Consumer	Situations	Other	Pamela St	2013-03-10	CA-2013-	2013-03-13	First Class	0%	89	25%	5	355
22	United Sta	South	Florida	Apopka	32712	Furniture	Furnishing	Consumer	DataProdu	Other	Chloris Ka	2011-07-28	CA-2011-	2011-07-28	Same Day	20%	13	10%	6	130

Figure 1: retail.xls

현재 데이터는 tidy??

- 다행히도 현재 사용할 데이터는 **tidy**한 데이터입니다. **tidy**하지 않은 데이터를 처리한 예제는 `M44[0]-Preprocessing.Rmd`에서 찾을 수 있습니다.
- `M44[0]-Preprocessing.Rmd`에서는 흔한 raw 데이터를 불러와서 변수의 이름을 붙이고 **tidyr**패키지의 명령을 이용해서 **tidy**한 dataset으로 만들어내서 csv파일로 저장하는 프로세스를 보여줍니다.
- **네모난 tidy한** dataset이 되었기에 `M44[1]` 이후에는 각종 데이터 분석을 할 수 있습니다.
- 아래 명령을 통해 데이터를 불러옵니다.

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

데이터 구조를 파악합니다.

```
str(dataset)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    9994 obs. of  20 variables:
##  $ Country      : chr  "United States" "United States" "United States" "United States"
##  $ Region       : chr  "East" "East" "South" "South" ...
##  $ State        : chr  "Ohio" "Ohio" "Virginia" "Virginia" ...
##  $ City         : chr  "Akron" "Akron" "Alexandria" "Alexandria" ...
##  $ Postal Code  : num  44312 44312 22304 22304 22304 ...
##  $ Category     : chr  "Furniture" "Furniture" "Furniture" "Furniture" ...
##  $ Sub-Category : chr  "Tables" "Furnishings" "Furnishings" "Furnishings" ...
##  $ Segment      : chr  "Corporate" "Consumer" "Corporate" "Home Office" ...
##  $ Product Name : chr  "Chromcraft Rectangular Conference Tables" "Deflect-o Glass Clear
##  $ Manufacturer : chr  "Chromcraft" "Deflect-o" "DAX" "Eldon" ...
##  $ Customer Name: chr  "Ed Braxton" "Ted Trevino" "Andrew Gjertsen" "Shirley Daniels" .
##  $ Order Date   : POSIXct, format: "2014-10-21" "2011-05-18" ...
##  $ Order ID     : chr  "CA-2014-147277" "CA-2011-164224" "CA-2012-104241" "US-2011-1555
##  $ Ship Date    : POSIXct, format: "2014-10-25" "2011-05-20" ...
##  $ Ship Mode    : chr  "Standard Class" "Second Class" "Standard Class" "Standard Class
##  $ Discount     : num  0.4 0.2 0 0 0 0.3 0.2 0.32 0.3 0.6 ...
##  $ Profit       : num  -76 4 69 4 31 -31 3 -36 -350 -11 ...
##  $ Profit Ratio : num  -0.27 0.03 0.36 0.36 0.49 -0.13 0.29 -0.18 -0.14 -0.48 ...
##  $ Quantity     : num  2 3 14 3 3 2 2 2 5 3 ...
##  $ Sales        : num  284 149 192 12 63 ...
```

```
dim(dataset)
```

```
## [1] 9994   20
```

```r
head(dataset)
```

```
## # A tibble: 6 x 20
##   Country Region State City  `Postal Code` Category `Sub-Category` Segment
##   <chr>   <chr>  <chr> <chr>         <dbl> <chr>    <chr>           <chr>
## 1 United~ East   Ohio  Akron         44312 Furnitu~ Tables          Corpor~
## 2 United~ East   Ohio  Akron         44312 Furnitu~ Furnishings     Consum~
## 3 United~ South  Virg~ Alex~         22304 Furnitu~ Furnishings     Corpor~
## 4 United~ South  Virg~ Alex~         22304 Furnitu~ Furnishings     Home O~
## 5 United~ South  Virg~ Alex~         22304 Furnitu~ Furnishings     Home O~
## 6 United~ Centr~ Texas Allen         75002 Furnitu~ Tables          Consum~
## # ... with 12 more variables: `Product Name` <chr>, Manufacturer <chr>,
## #   `Customer Name` <chr>, `Order Date` <dttm>, `Order ID` <chr>, `Ship
## #   Date` <dttm>, `Ship Mode` <chr>, Discount <dbl>, Profit <dbl>, `Profit
## #   Ratio` <dbl>, Quantity <dbl>, Sales <dbl>
```

데이터 분석의 과정

- 데이터 분석은 대체로 1) 데이터의 관찰 -> 2) 가설의 설정 -> 3) 가설 검증 -> 4) 결론 도출 -> 5) 공유의 과정으로 이루어집니다.
- 이 과정에서 때로는 다른 데이터를 더 확보해서 분석에 포함시키기도 하고, 데이터의 결함과 미비한 점을 파악함으로써 데이터 관리자와 공급자와 커뮤니케이션 합니다.
- 마지막으로 공유의 단계에서는 데이터와 관련된 다른 사람들과 의사소통 하며 더 나은 의사결정을 돕게합니다.
- `retail` 데이터 셋을 관찰한 결과 아래처럼 9개의 가설을 세웠습니다. 이를 차례로 검증해 보겠습니다.

생각해 볼 수 있는 궁금증들…

1. `Ship Date`를 기반으로 배송이 가장 오래걸리는 상품은 무엇인가?
2. 마진이 가장 많이 남는 상품은 무엇인가?
3. 가장 많은 상품을 구매한 고객은 누구이며 언제 구매하였는가?
4. 가장 판매가 부진한 상품은 무엇이고 이유는 무엇인가?
5. 많이 팔리는 상품의 가격수준과 일정가격 이하 상품의 판매량은 어떠한가?
6. Discount가 많을 수록 매출이 늘어나는가?
7. 지역별로 가장 많이 팔리는 상품은 무엇인가?
8. `Order Date`를 기반으로 동시구매가 많이 일어나는 상품은 무엇인가?
9. 특정상품의 판매시기와 지역별 수요를 파악해보자

1. Ship Date를 기반으로 배송이 가장 오래걸리는 상품은 무엇인가?

Background & Strategy

Background

- 상품은 주문(Order)-출고(Ship)-배송완료(Delivery)의 3개의 시점이 있음.
- 주문과 출고사이의 시간(lead time)은 판매자의 역량이며,
- 출고와 배송완료 시점 사이의 시간(delivery time)은 delivery mode에 따라 결정됨.

Development

- 질문자의 의도는 주문과 배송완료 사이의 시간, 그러니까 소비자의 입장에서 생각하고 있는 것이지만
- (처음 떠오르는 질문은 이렇게 분석 의도와 목적과 합치하지 않는 경우가 많기에, 검증가능한 가설의 형태로 만들기 위해서 생각과 논의가 필요함)
- 업체 입장에서는 lead time이 업체 운영의 효율성과 관계된 것임.
- 그러므로 각 `Sub-Category`별로 `Ship Date`와 `Order Date`의 차이를 계산하여 평균과 분산을 구해보아야 함.

More Development

- Average보다는 lead time의 long tail이 중요한 수치임.
- Long tail이란 소비자의 불만을 야기할 정도로 소비자가 기대하는 평균 lead time에 비해서 아주 큰 값을 의미함.
- 오퍼레이션의 관점에서 이는 근로자의 휴가나 파업등의 이슈, 혹은 장비의 결함과 break down등의 상황에 해당함.
- Lead time이 아닌 delivery time의 경우에는 명절이나 연휴 시즌으로 인한 지연, 기상 조건에 따른 지연이 이에 해당함.
- Lead time이 이상값에 해당하는 경우에는 따로 데이터를 추출해서 요인을 Case별로 분석해봐야 함.

Tasks Specification

1. `leadTime`이라는 변수를 생성하고 각각의 `Sub-Category`에 대해서 `leadTime`의 평균과 분산을 구한다.
2. 각각의 `Sub-Category`에 대해서 box-plot을 그린다.
3. `leadTime`이 가장 긴 20개의 관찰값을 출력한다.

Task 1

leadTime이라는 변수를 생성하고 각각의 Sub-Category에 대해서 leadTime의 평균과 분산을 구한다.

```
dataset$leadTime <- dataset$`Ship Date` - dataset$`Order Date`
activate("lubridate")
task1 <- dataset %>%
  group_by(`Sub-Category`) %>%
  summarise(avgLT = mean(leadTime), sdLT = sd(leadTime)) %>%
  arrange(desc(avgLT))
head(task1)
```

```
## # A tibble: 6 x 3
##   `Sub-Category` avgLT             sdLT
##   <chr>          <time>            <time>
## 1 Art            350158.793969849 150923.279807913
## 2 Binders        347585.554826001 150339.074322692
## 3 Supplies       346964.210526316 161572.709673627
## 4 Envelopes      346960.62992126  149741.534450274
## 5 Labels         346074.725274725 151492.240638765
## 6 Phones         345891.563554556 147470.320524176
```

- avgLT와 sdLT를 구했지만, 단위가 초(second)로 되어있습니다.
- class(task1$avgLT)를 실행해보니 difftime라고 합니다.
- 그러므로 google에서 “convert difftime second to days”를 검색하여 아래와 같은 해결책을 얻고 task1을 완료합니다.

```r
task1$avgLT <-
  task1$avgLT %>%
  as.numeric(units = "days") %>%
  round(2)
task1$sdLT  <-
  task1$sdLT  %>%
  as.numeric(units = "days") %>%
  round(2)
```

```r
print(task1)
```

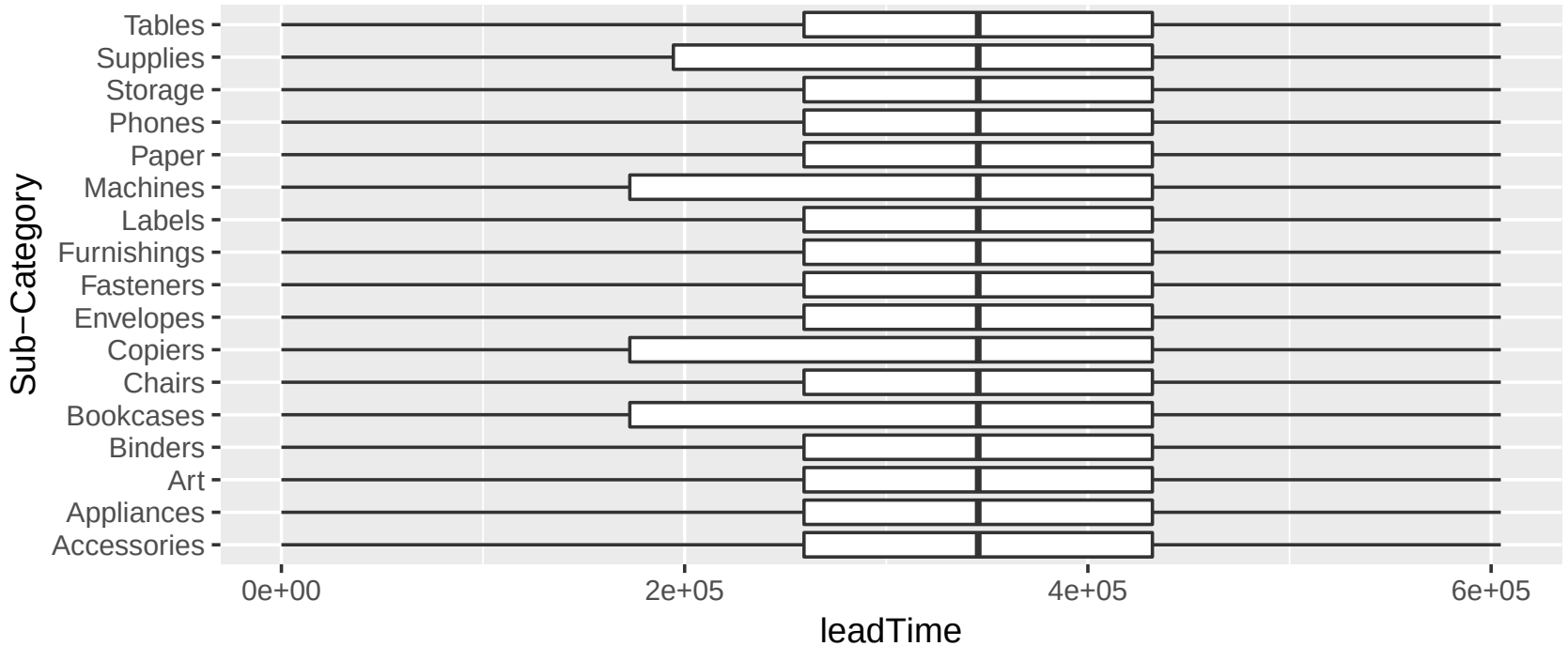
```
## # A tibble: 17 x 3
##    `Sub-Category` avgLT  sdLT
##    <chr>          <dbl> <dbl>
##  1 Art             4.05  1.75
##  2 Binders         4.02  1.74
##  3 Supplies        4.02  1.87
##  4 Envelopes       4.02  1.73
##  5 Labels          4.01  1.75
##  6 Phones          4     1.71
##  7 Appliances      3.99  1.69
##  8 Fasteners       3.98  1.76
##  9 Storage         3.98  1.76
## 10 Furnishings     3.96  1.74
## 11 Chairs          3.9   1.79
## 12 Tables          3.9   1.8
## 13 Accessories     3.89  1.72
## 14 Paper           3.89  1.75
## 15 Bookcases       3.81  1.68
## 16 Machines        3.75  1.99
## 17 Copiers         3.62  1.88
```

- 품목별로 평균적인 `leadTime`의 크기가 크지 않습니다.
- 표준편차의 차이도 별로 크지 않습니다.
- 평균과 표준편차외에 전체적인 분포를 **task2**에서 살펴보도록 합니다.

Task 2

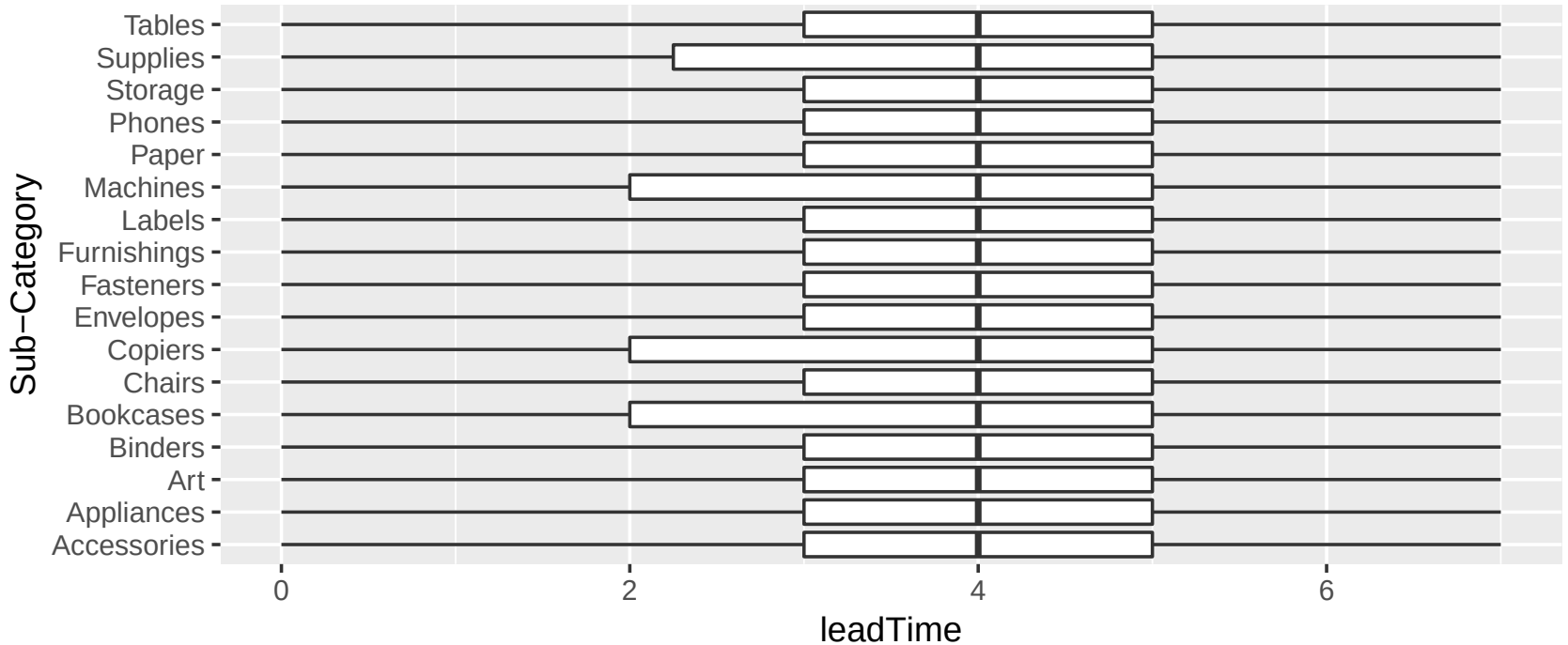
각각의 Sub-Category에 대해서 box-plot을 그린다.

```
ggplot(dataset) +
  geom_boxplot(aes(x = `Sub-Category`, y = leadTime)) +
  coord_flip()
```



- `leadTime`의 수치가 뭔가 괴상하게 나옵니다.
- 원인을 살펴보니 `leadTime`변수 자체가 seconds단위로 되어 있습니다.
- 위에서 days로 바꾸었던 방법을 이용해서 해결합니다.

```r
dataset$leadTime <- dataset$leadTime %>% as.numeric(units = "days")
ggplot(dataset) +
  geom_boxplot(aes(x = `Sub-Category`, y = leadTime)) +
  coord_flip()
```



위의 그림으로 부터 `leadTime`에 대해서 아래와 같은 결론을 내릴 수 있습니다.

- 모든 상품에 대해서 중간값이 4일이다.
- Furnishing, Machines, Chairs, Bookcases와 같이 크기가 클 수 있는 제품의 경우에도 `leadTime`이 특별히 길다고 말할 수 없다.
- 전체 관찰값의 갯수는 `nrow(dataset)`: 9994이다. 그런데 `max(dataset$leadTime)`: 7이다. `leadTime`이 7인 경우는 총 `sum(dataset$leadTime==max(dataset$leadTime))`: 621건에 해당하며 이는 `sum(dataset$leadTime==max(dataset$leadTime))/nrow(dataset)*100`: 6.2137282% 이다.
- 즉, `nrow(dataset)`건의 주문을 처리하면서 `leadTime`을 7일 내로 100% 처리했으며, 6일내로 94% 처리했다.
- `leadTime`이 대체로 잘 관리되고 있음을 알 수 있다.

Task 3

`leadTime`이 가장 긴 20개의 관찰값을 출력한다.

- 만약에 `leadTime`의 분포가 long-tail의 모습을 보인다면, 즉 몇몇 제품의 `leadTime`이 이상하게 높았다면, 이런 케이스를 나열하고 분석하는 것이 Task 3의 목적이다.
- 그러나 Task 2의 분석 결과를 보면 `leadTime`이 대체로 잘 관리되고 있음을 알 수 있다.
- 그러므로 원래 의도한 Task 3를 수행하는 것은 큰 의미가 없다.
- 그렇기 때문에 Task 3의 원래 목적은 살리면서 내용을 변경하여 전체 상품의 6%에 해당하는 `leadTime`이 7일인 경우는 어떤 상품들이 많이 있는지 알아보자.

아래와 같은 관찰을 수행할 수 있다.

- Task 3-1: `leadTime`이 7인 주문을 `Category`와 `Sub-Category`로 나누어서 갯수를 관찰한다. `Category`로 나눈 것은 pie-chart로 그리고 `Sub-Category`로 나눈 것은 항목의 갯수가 많으므로 pie-chart가 아닌 표로 정리한다.
- **Task 3-2**: 그러나 Task 3-1의 관찰은 갯수를 보는 것이 아니라 주문 갯수에 대비한 비율로 보아야 한다. 즉, `leadTime`이 7일인 Furniture 주문의 갯수를 보는 것이 아니라, 전체 Furniture 주문 중에서 `leadTime`이 7일인 경우의 비율을 계산하는 것이 바람직하다.

아래 코드로 Task 3-2를 수행해 봅니다.

Task 3-2: 전체 Furniture 주문 중에서 **leadTime**이 7일인 경우의 비율을 계산하는 것

```
# For Category
task3_2a <- dataset %>%
  group_by(`Category`) %>%
  summarise(maxLeadTimePercent = 100*sum(leadTime==7)/length(leadTime)) %>%
  arrange(desc(maxLeadTimePercent))
print(task3_2a)
```

```
## # A tibble: 3 x 2
##   Category        maxLeadTimePercent
##   <chr>                        <dbl>
## 1 Office Supplies               6.37
## 2 Furniture                     6.08
## 3 Technology                    5.85
```

```r
# For Sub-Category
task3_2b <- dataset %>%
  group_by(`Sub-Category`) %>%
  summarise(
    maxLeadTimePercent =
      100*sum(leadTime==7)/length(leadTime)) %>%
  arrange(desc(maxLeadTimePercent))
```

```r
print(task3_2b)
```

```
## # A tibble: 17 x 2
##    `Sub-Category` maxLeadTimePercent
##    <chr>                       <dbl>
##  1 Supplies                    10.5
##  2 Machines                     8.70
##  3 Fasteners                    7.83
##  4 Art                          7.54
##  5 Binders                      7.35
##  6 Tables                       6.90
##  7 Chairs                       6.48
##  8 Appliances                   6.01
##  9 Envelopes                    5.91
## 10 Copiers                      5.88
## 11 Phones                       5.85
## 12 Labels                       5.77
## 13 Furnishings                  5.75
## 14 Accessories                  5.42
## 15 Storage                      5.32
## 16 Bookcases                    5.26
## 17 Paper                        4.82
```

- Sub-Category가 Supplies인 항목의 경우에는 주문의 10%이상의 **leadTime**이 7일에 해당했습니다.
- 그러므로 발주 process 효율성의 향상을 위해서는 Supplies물품을 주문 받은 이후 배송을 시행하기 까지 왜 오래 걸리는지 알아볼 필요가 있다는 결론을 내릴 수 있습니다!

2. 마진이 가장 많이 남는 상품은 무엇인가?

Background & Strategy

- 기업의 입장에서는 매출과 이익이 모두 중요합니다.
- 각 Category와 Sub-Category에 대해서 기업의 매출과 이익에 얼마나 기여했는지 알아봅니다.
- 그리고 매출 대비 이익률을 알아볼 수 있습니다. (Profit-Revenue-Ratio)

- 또한 기업의 매출과 이익이 계속적으로 성장하고 있는지를 알아보는 것도 중요합니다.
- 그렇기 때문에 분기 단위로 나누어 위의 분석을 수행합니다.

Tasks Specification

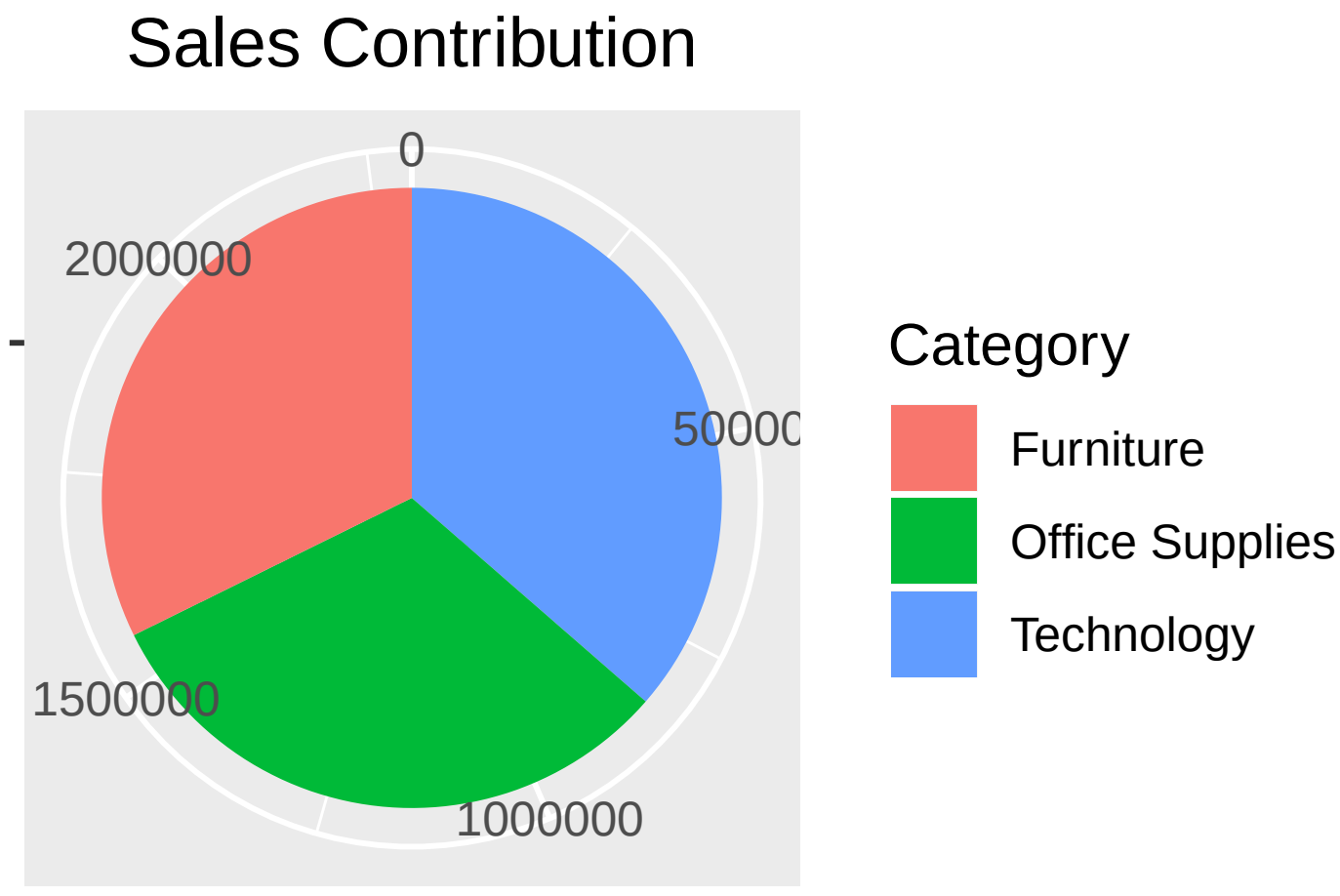
- Task 1. 각 `Category`와 `Sub-Category`에 대해서 `Sales`와 `Profit`을 각각 Aggregate한다. 그리고 `Profit`을 `Sales`로 나누어서 `profitRatio`을 구한다.
- Task 2. 분기를 나타내는 변수를 생성하고 위의 분석을 반복한다.

Task 1

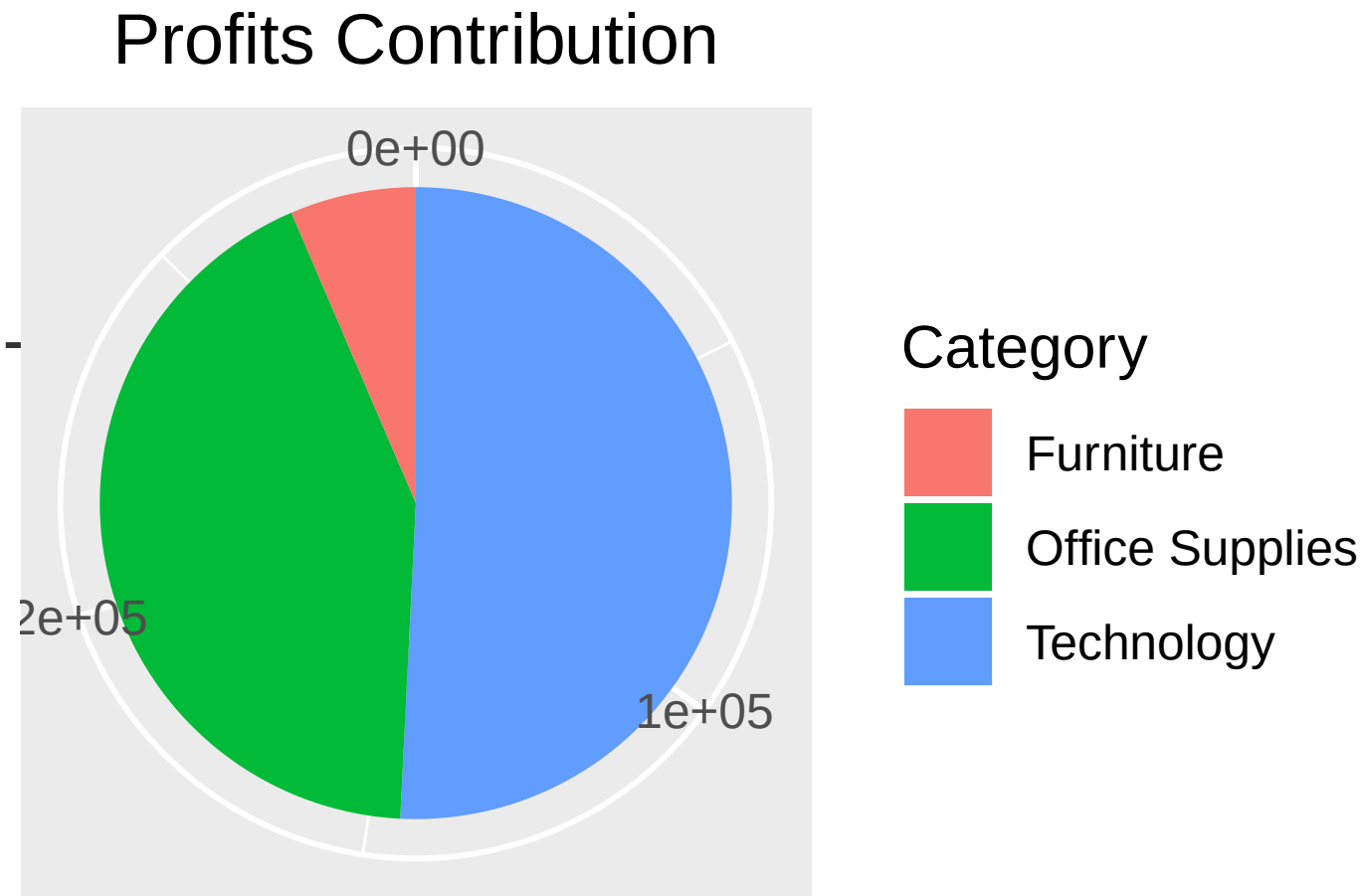
```
task1a <- dataset %>%
  group_by(Category) %>%
  summarise(Sales = sum(Sales), Profit = sum(Profit)) %>%
  mutate(profitRatio = round(Profit/Sales,2)) %>%
  arrange(desc(profitRatio))
print(task1a)
```

```
## # A tibble: 3 x 4
##   Category         Sales Profit profitRatio
##   <chr>            <dbl>  <dbl>       <dbl>
## 1 Office Supplies 719127 122474        0.17
## 2 Technology      836221 145429        0.17
## 3 Furniture       742006  18444        0.02
```

```r
# Reference: `M91.piechart`
ggplot(task1a, aes(x = "", y = Sales, fill = factor(Category))) +
  geom_bar(width = 1, stat = "identity") +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(fill = "Category", x = NULL, y = NULL,
       title = "Sales Contribution") +
  coord_polar(theta = "y", start = 0)
```



```r
ggplot(task1a, aes(x = "", y = Profit, fill = factor(Category))) +
  geom_bar(width = 1, stat = "identity") +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(fill = "Category", x = NULL, y = NULL,
       title = "Profits Contribution") +
  coord_polar(theta = "y", start = 0)
```



- Furniture의 경우에는 무려 74만불의 매출을 내고도 2만불에 못 미치는 이익률을 보였습니다.
- Furniture와 같이 1) 부피와 무게가 커서 다루는데에 인력과 시간이 많이 필요하고, 2) 재고 상태로 보유하는 것이 저장, 운반, 감가상각의 관점에서 비용이 크고, 3) 반품이 생기면 완전 골치아픈 ㅠㅠ 분류가 심지어 이익률도 2%입니다. ㅠㅠ

Furniture에 대해서 떠오르는 질문은 다음과 같습니다.

1. Sub-Category를 보아도 다 이런 식인가? 대형 가구와 중소형 가구의 이익률이 다를 수도 있지 않나?
2. 처음부터 지금까지 Furniture 비지니스는 계속 이랬나? 최근의 IKEA의 습격을 당해서 마진율이 내려간 것인가?
3. 과연 Furniture 비지니스를 계속해야 하는가? 접지 않는다면 어떤 대안으로 돌파가 가능한가?에 대해서 전략을 세우고 결론을 내려야 합니다.

- 2의 분석의 경우에는 IKEA의 동향에 대한 데이터를 추가로 확보하고 IKEA 매장의 근교 지역과 아닌 지역을 구분해서 비교하는 분석을 실시해야 할 것입니다.

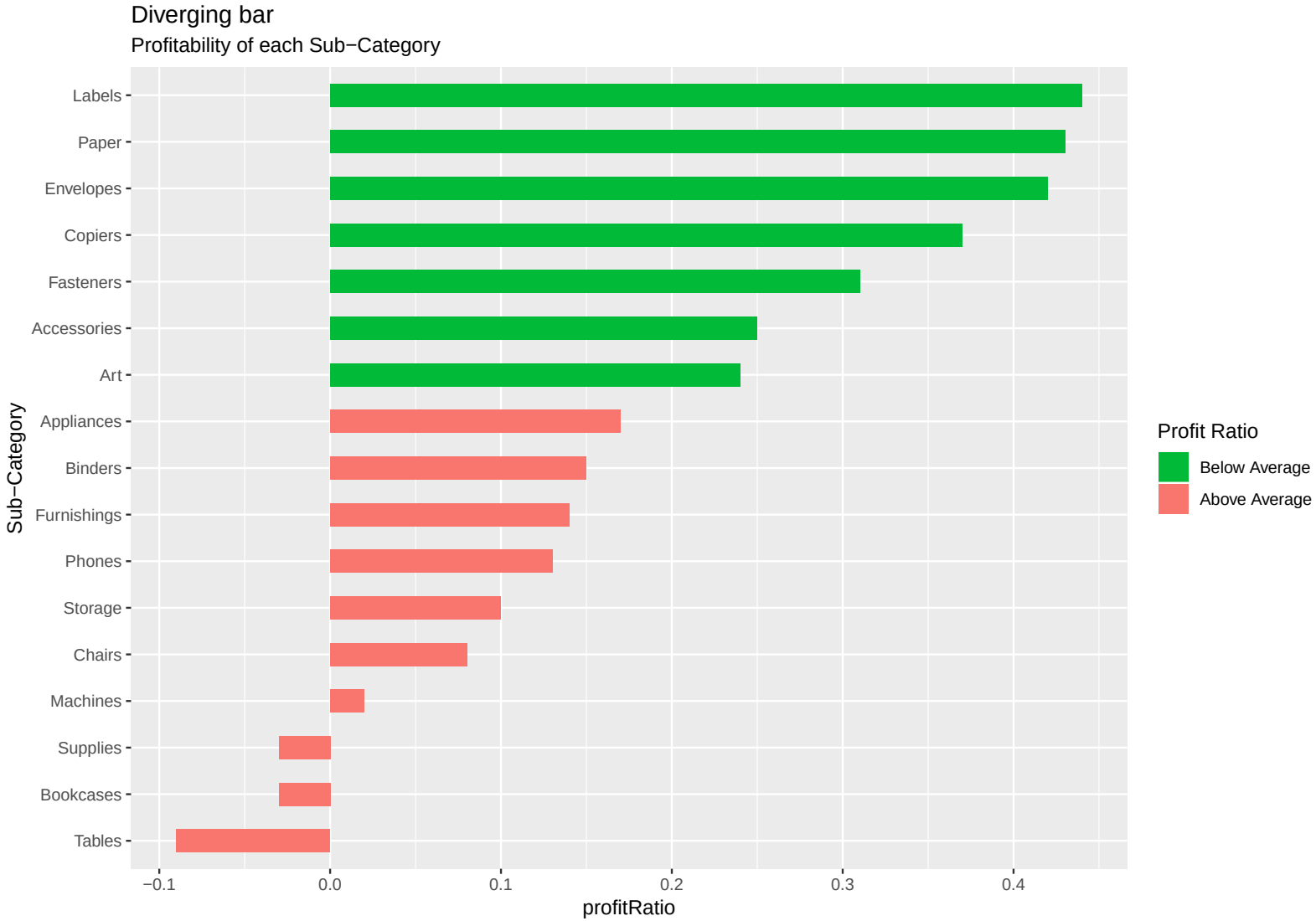
```r
task1b <- dataset %>%
  group_by(`Sub-Category`) %>%
  summarise(Sales = sum(Sales), Profit = sum(Profit)) %>%
  mutate(profitRatio = round(Profit/Sales,2)) %>%
  arrange(desc(profitRatio))
task1b
```

```
## # A tibble: 17 x 4
##    `Sub-Category`  Sales Profit profitRatio
##    <chr>           <dbl>  <dbl>       <dbl>
##  1 Labels          12507   5558        0.44
##  2 Paper           78475  34053        0.43
##  3 Envelopes       16477   6956        0.42
##  4 Copiers        149530  55618        0.37
##  5 Fasteners        3024    952        0.31
##  6 Accessories    167401  41932        0.25
##  7 Art             27137   6530        0.24
##  8 Appliances     107538  18132        0.17
##  9 Binders        203428  30200        0.15
## 10 Furnishings     91705  13070        0.14
## 11 Phones         330047  44492        0.13
## 12 Storage        223862  21280        0.1
## 13 Chairs         328454  26586        0.08
## 14 Machines       189243   3387        0.02
## 15 Bookcases      114879  -3479       -0.03
## 16 Supplies        46679  -1187       -0.03
## 17 Tables         206968 -17733       -0.09
```

위의 표는 이처럼 Diverging bar로 표현할 수도 있습니다.

```
# Reference: `M91.2.Deviation`
task1b$profitHL <-
  ifelse(task1b$profitRatio < mean(task1b$profitRatio),
         "below average", "above average")
task1b <- task1b %>% arrange(profitRatio)
# Convert to factor to preserve sorted order in plot.
task1b$`Sub-Category` <-
  factor(task1b$`Sub-Category`, levels = task1b$`Sub-Category`)
a <- ggplot(task1b,
            aes(x = `Sub-Category`, y = profitRatio, label = profitRatio)) +
  geom_bar(stat = 'identity', aes(fill = profitHL), width = .5) +
  scale_fill_manual(
    name = "Profit Ratio",
    labels = c("Below Average", "Above Average"),
    values = c("below average" = "#f8766d",
               "above average" = "#00ba38")) +
  labs(title = "Diverging bar",
       subtitle = "Profitability of each Sub-Category") +
  coord_flip()
```

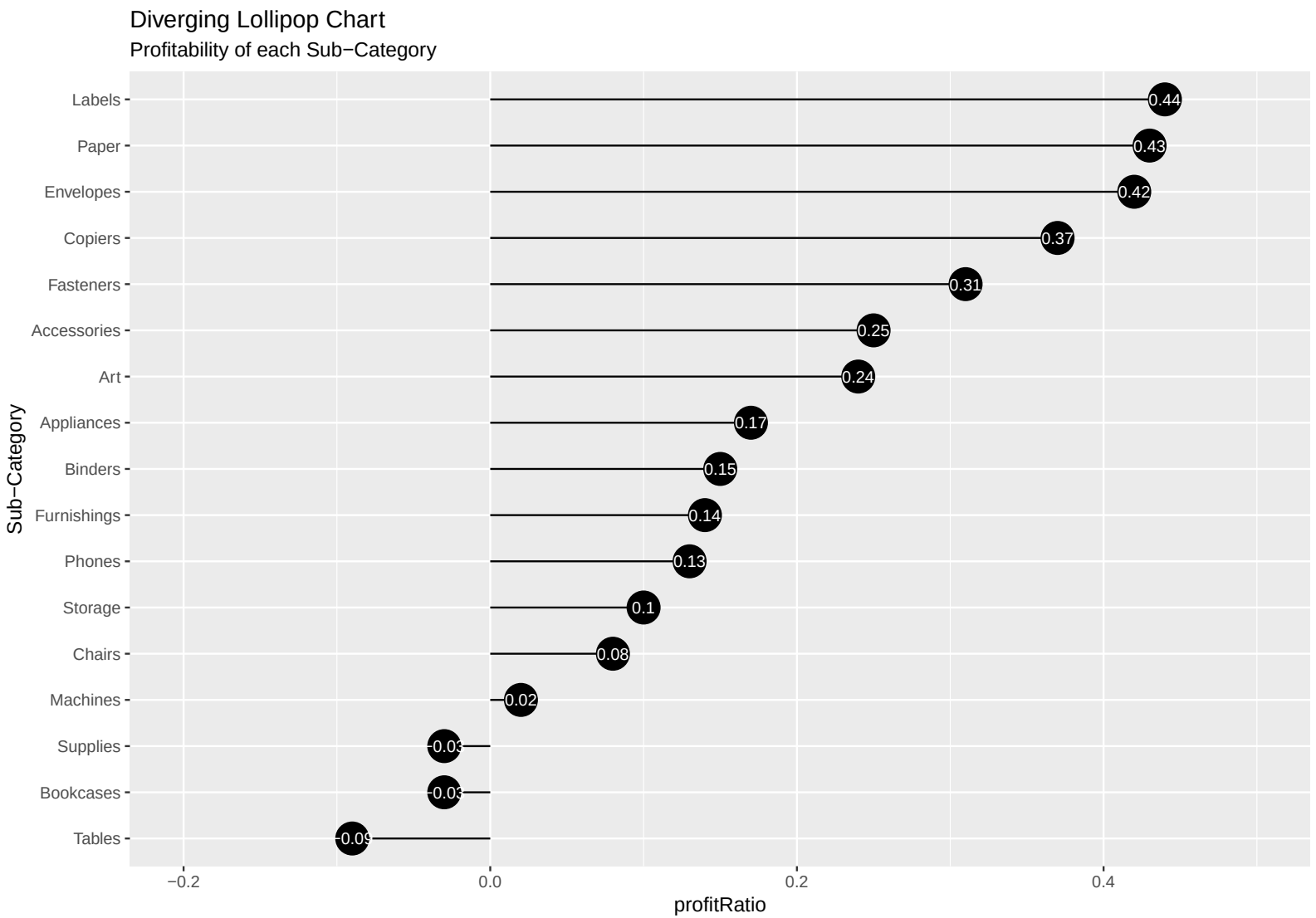
```
print(a)
```


Diverging bar
Profitability of each Sub−Category
Labels
Paper
Envelopes
Copiers
Fasteners
Accessories
Art
Appliances
Binders
Furnishings
Phones
Storage
Chairs
Machines
Supplies
Bookcases
Tables
Sub−Category
−0.1
0.0
0.1
0.2
0.3
0.4
profitRatio
Profit Ratio
Below Average
Above Average

혹은 좀 더 modern look을 제공하는 아래와 같은 “Diverging Lollipip Chart”도 가능합니다.

```
# Reference: `M91.2.Deviation`
a <- ggplot(task1b,
            aes(x = `Sub-Category`, y = profitRatio, label = profitRatio)) +
  geom_point(stat = 'identity', fill = "black", size = 8) +
  geom_segment(aes(y = 0, x = `Sub-Category`,
                   yend = profitRatio, xend = `Sub-Category`),
               color = "black") +
  geom_text(color = "white", size = 3) +
  labs(title = "Diverging Lollipop Chart",
       subtitle = "Profitability of each Sub-Category") +
  ylim(-0.2, 0.5) +
  coord_flip()
```

```
print(a)
```


Diverging Lollipop Chart
Profitability of each Sub−Category
Labels
Paper
Envelopes
Copiers
Fasteners
Accessories
Art
Appliances
Binders
Furnishings
Phones
Storage
Chairs
Machines
Supplies
Bookcases
Tables
0.44
0.43
0.42
0.37
0.31
0.25
0.24
0.17
0.15
0.14
0.13
0.1
0.08
0.02
−0.2
0.0
0.2
0.4
Sub−Category
profitRatio

- 이익률 하위 부분의 Storage, Chairs, Bookcases, Tables 모두 가구류에 해당합니다.
- 제가 만약에 이 기업을 경영자라면 해당 소형 가구라인의 유지를 전면적으로 고민할 것 같습니다.

- 이익률이 높은 Sub-Category들의 경우에는 이익률은 높지만 실제 이익의 총량은 얼마 안되는 품목들도 많이 있습니다.
- Labels, Envelopes, Fastener, Art의 경우에는 이익 자체가 크지 않습니다. (봉투를 팔아서 돈을 벌면 얼마나 벌겠습니까…)
- 같은 table을 이익순으로 정렬하는 것이 다른 시각을 제공할 수 있습니다.

```
task1b %>% arrange(desc(Profit, Sales))
```

```
## # A tibble: 17 x 5
##    `Sub-Category`  Sales Profit profitRatio profitHL
##    <fct>           <dbl>  <dbl>       <dbl> <chr>
##  1 Copiers        149530  55618        0.37 above average
##  2 Phones         330047  44492        0.13 below average
##  3 Accessories    167401  41932        0.25 above average
##  4 Paper           78475  34053        0.43 above average
##  5 Binders        203428  30200        0.15 below average
##  6 Chairs         328454  26586        0.08 below average
##  7 Storage        223862  21280        0.1  below average
##  8 Appliances     107538  18132        0.17 below average
##  9 Furnishings     91705  13070        0.14 below average
## 10 Envelopes       16477   6956        0.42 above average
## 11 Art             27137   6530        0.24 above average
## 12 Labels          12507   5558        0.44 above average
## 13 Machines       189243   3387        0.02 below average
## 14 Fasteners        3024    952        0.31 above average
## 15 Supplies        46679  -1187       -0.03 below average
## 16 Bookcases      114879  -3479       -0.03 below average
## 17 Tables         206968 -17733       -0.09 below average
```

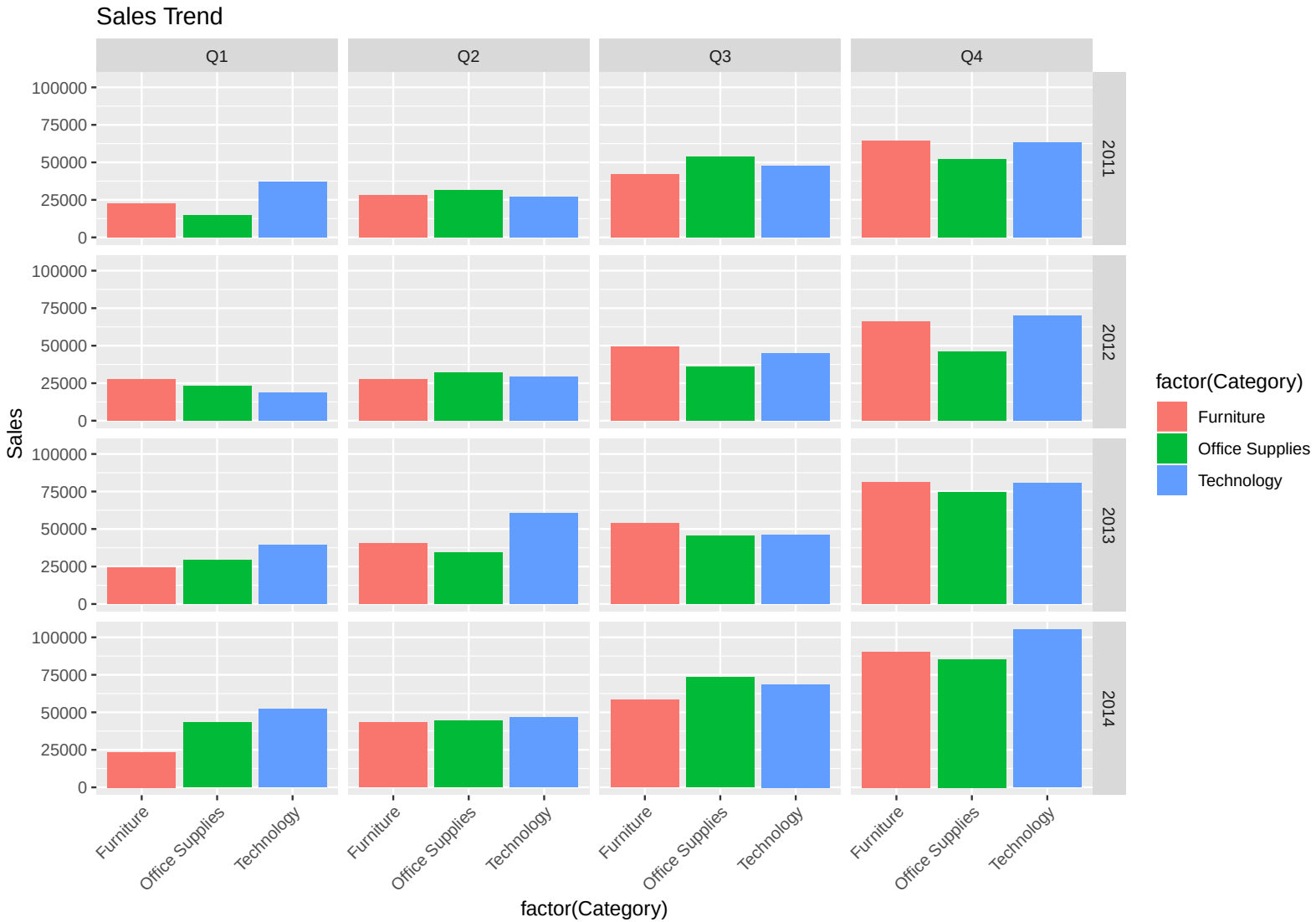
Task 2

2. 분기를 나타내는 변수를 생성하고 위의 분석을 반복한다.

```
task2 <- dataset %>%
  mutate(year    = substr(`Order Date`, 1, 4),
         quarter = ceiling(as.numeric(substr(`Order Date`, 6, 7))/3)) %>%
  select(year, quarter, Category, `Sub-Category`, Profit, Sales) %>%
  group_by(year, quarter, Category) %>%
  summarise(Sales = sum(Sales), Profit = sum(Profit))
task2$year    <- factor(task2$year)
task2$quarter <- factor(paste0("Q", task2$quarter))
head(task2)
```

```
## # A tibble: 6 x 5
## # Groups:   year, quarter [2]
##   year  quarter Category        Sales Profit
##   <fct> <fct>   <chr>           <dbl>  <dbl>
## 1 2011  Q1      Furniture       22658   -206
## 2 2011  Q1      Office Supplies 14526   2225
## 3 2011  Q1      Technology      37261   1781
## 4 2011  Q2      Furniture       28061    801
## 5 2011  Q2      Office Supplies 31245   5780
## 6 2011  Q2      Technology      27234   4620
```

```r
ggplot(task2, aes(x = factor(Category), y = Sales, fill = factor(Category))) +
  geom_bar(stat = 'identity') +
  facet_grid(year~quarter) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Sales Trend")
```



Seasonality

- 특이하게도 대체적으로 1분기, 2분기, 3분기, 4분기로 갈수록 매출이 급격하게 늘어나는 것을 볼 수 있습니다.
- 시계열의 이런 주기성을 계절성(seasonality)라고 합니다.
- 시계열 자료에서 경향의 구성 요소는 크게 1) 트렌드, 2) 계절성, 3) 그외의 잡음으로 생각할 수 있습니다.
- 우선 계절성은 어떤 고정된 길이의 시간에 따라서 주기적인 모습(cyclic pattern)을 보이는 것을 의미합니다.

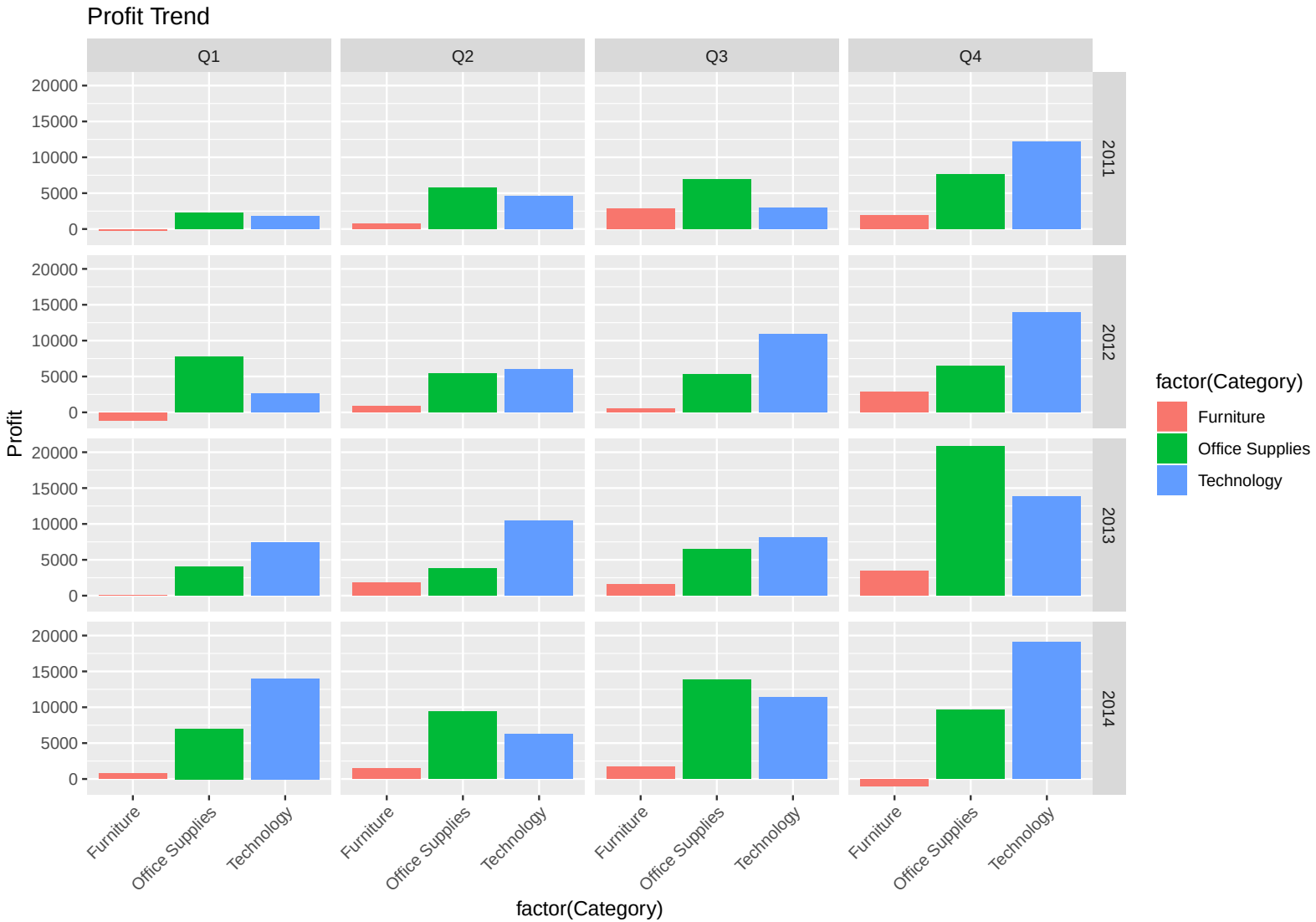
Seasonality를 대처하는 방법

- 특히, 인간의 삶과 밀접한 연관이 있는 시계열 데이터는 대부분 계절성이 있습니다.
- 교통 수단의 이용량의 경우에는 출퇴근 시간과 낮시간의 패턴에 계절성이 있고, 1주일에 대해서 요일별로의 계절성이 있습니다. 그리고 매년 명절이 찾아옵니다.
- 미국 소비자의 쇼핑 패턴을 보면 대부분의 소비가 겨울에 집중되는 것을 알 수 있습니다.그렇기 때문에 기업의 매출과 이익의 성장을 단순히 “전월대비”로 볼게 아니라, “전년동월대비” 관점으로 보아야합니다.

해당 기업의 자료

- 위의 자료에서는 우선 매출량의 트렌드는 긍정적입니다.
- 연도가 지나면서 점점 매출이 늘어나고 있습니다.
- 그리고 retail 상품이기에 계절성이 매우 뚜렷한 특징을 보이고 있습니다.
- 만약에 B2C 비지니스가 아닌 제조업체 등의 B2B 비지니스였다면, 이렇게 뚜렷한 계절성을 보이지는 않았을 것입니다.

```r
ggplot(task2, aes(x = factor(Category), y = Profit, fill = factor(Category))) +
  geom_bar(stat = 'identity') +
  facet_grid(year~quarter) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Profit Trend")
```



- 기업의 순이익과 직결되는 Profit에 대한 시계열 분석입니다.
- 앞에서 살펴본 매출과 비슷한 패턴을 보이는 것을 확인할 수 있습니다.
- 2014년 1분기에는 전년과 전전년 동분기에 비해서 Technology 제품에 대해서 큰 수익을 거둔것을 살펴볼 수 있습니다.
- 만약에 조금더 분석을 해본다면 2012년, 2013년, 2014년의 1분기에 대해서 각각 어떤 상품들이 팔렸는지 알고 싶습니다.
- 예를 들어서 2014년 1분기에 아이폰의 새로운 버전이 나왔고 그것을 해당 쇼핑몰에서 많이 판매하였다면, 그것이 매출에 크게 기여하였다라고 말할수 있겠네요.
- 앞의 분석에서 문제로 제기했던 Furniture의 경우에는 2014년도 4분기에는 전년과 전전년 동분기에 대비해서 순이익이 적었습니다.
- 이것 역시 이유를 더 살펴보고 2015년에의 Furniture 관련 전략을 수립할 필요가 있어보입니다.

Rest…

3. 가장 많은 상품을 구매한 고객은 누구이며 언제 구매하였는가?
4. 가장 판매가 부진한 상품은 무엇이고 이유는 무엇인가?
5. 많이 팔리는 상품의 가격수준과 일정가격 이하 상품의 판매량은 어떠한가?
6. Discount가 많을 수록 매출이 늘어나는가?
7. 지역별로 가장 많이 팔리는 상품은 무엇인가?
8. `Order Date`를 기반으로 동시구매가 많이 일어나는 상품은 무엇인가?
9. 특정상품의 판매시기와 지역별 수요를 파악해보자