SuperStore(Retail) Data Analysis

1. Import libarary

In [1]:

```
library(tidyr)
library(dplyr)
library(tidyverse)
library(readxl)
```

```
Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

── Attaching packages ─────────────────────────────────────── tidyvers
e 1.2.1 ──
✔ ggplot2 2.2.1     ✔ purrr   0.2.4
✔ tibble  1.4.2     ✔ stringr 1.3.0
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ────────────────────────────────────────── tidyverse_conf
licts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

2. Load data

In [2]:

```
dataset <- read_excel('../../data/data-data-anlaysis/super-store-data.xls')
```

```
Warning message in read_fun(path = path, sheet_i = sheet, limits = lim
its, shim = shim, :
“Coercing text to numeric in L2236 / R2236C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L5276 / R5276C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L8800 / R8800C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9148 / R9148C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9149 / R9149C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9150 / R9150C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9388 / R9388C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9389 / R9389C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9390 / R9390C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9391 / R9391C12: '05408'”Warning message
in read_fun(path = path, sheet_i = sheet, limits = limits, shim = shi
m, :
“Coercing text to numeric in L9743 / R9743C12: '05408'”
```

3-1. Understanding the data structure

In [3]:

```
str(dataset)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       9994 obs. of  21 varia
bles:
 $ Row ID       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Order ID     : chr  "CA-2016-152156" "CA-2016-152156" "CA-2016-1386
88" "US-2015-108966" ...
 $ Order Date   : POSIXct, format: "2016-11-08" "2016-11-08" ...
 $ Ship Date    : POSIXct, format: "2016-11-11" "2016-11-11" ...
 $ Ship Mode    : chr  "Second Class" "Second Class" "Second Class" "S
tandard Class" ...
 $ Customer ID  : chr  "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer Name: chr  "Claire Gute" "Claire Gute" "Darrin Van Huff"
"Sean O'Donnell" ...
 $ Segment      : chr  "Consumer" "Consumer" "Corporate" "Consumer"
...
 $ Country      : chr  "United States" "United States" "United States"
"United States" ...
 $ City         : chr  "Henderson" "Henderson" "Los Angeles" "Fort Lau
derdale" ...
 $ State        : chr  "Kentucky" "Kentucky" "California" "Florida"
...
 $ Postal Code  : num  42420 42420 90036 33311 33311 ...
 $ Region       : chr  "South" "South" "West" "South" ...
 $ Product ID   : chr  "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-100
00240" "FUR-TA-10000577" ...
 $ Category     : chr  "Furniture" "Furniture" "Office Supplies" "Furn
iture" ...
 $ Sub-Category : chr  "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product Name : chr  "Bush Somerset Collection Bookcase" "Hon Deluxe
Fabric Upholstered Stacking Chairs, Rounded Back" "Self-Adhesive Addre
ss Labels for Typewriters by Universal" "Bretford CR4500 Series Slim R
ectangular Table" ...
 $ Sales        : num  262 731.9 14.6 957.6 22.4 ...
 $ Quantity     : num  2 3 2 5 2 7 4 6 3 5 ...
 $ Discount     : num  0 0 0 0.45 0.2 0 0 0.2 0.2 0 ...
 $ Profit       : num  41.91 219.58 6.87 -383.03 2.52 ...
```

3-2. Observing data

In [4]:

```
dataset %>% head()
```

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...
1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...
2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...
3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...
4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...
5	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...
6	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	...

4. The process of data analysis

데이터 분석은 대체로 다음과 같은 5단계의 과정을 거친다.

- 1) 데이터의 관찰 -> 2) 가설의 설정 -> 3) 가설 검증 -> 4) 결론 도출 -> 5) 공유

이 과정에서 때로는 다른 데이터를 더 확보해서 분석에 포함시키기도 하고, 데이터의 결함과 미비한 점을 파악함으로써 데이터 관리자와 공급자와 커뮤니케이션 함.

마지막으로 공유의 단계에서는 데이터와 관련된 다른 사람들과 의사소통 하며 더 나은 의사결정을 돕게한다.
retail 데이터 셋을 관찰한 결과 아래처럼 9개의 가설을 세웠다. 이를 차례로 검증해 보겠다.

생각해 볼 수 있는 궁금증들...

1. Ship Date를 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 Date를 기반으로 동시구매가 많이 일어나는 상품은 무엇인가?
9. 특정상품의 판매시기와 지역별 수요를 파악해보자

In [ ]:

5. Data Analysis

5-1. 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

- "Ship Date를 기반으로 배송이 가장 오래걸리는 상품은 무엇인가?"라는 질문의 의도는 주문과 배송완료 사이의 시간, 그러니까 소비자의 입장에서 생각하것이지만
- (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-Category별로 Ship Date와 Order Date의 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

- leadTime이라는 변수를 생성하고 각각의 Sub-Category에 대해서 leadTime의 평균과 분산을 구한다.
- 각각의 Sub-Category에 대해서 box-plot을 그린다.
- leadTime이 가장 긴 20개의 관찰값을 출력한다.

Change Columns name

- 컬럼을 선택하는데 공백이 있어서 에러가 뜨는 경우가 발생
- 공백을 underbar(_)로 변경

In [5]:

```
colnames(dataset)
```

'Row ID' 'Order ID' 'Order Date' 'Ship Date' 'Ship Mode' 'Customer ID'
'Customer Name' 'Segment' 'Country' 'City' 'State' 'Postal Code' 'Region'
'Product ID' 'Category' 'Sub-Category' 'Product Name' 'Sales' 'Quantity' 'Discount'
'Profit'

In [6]:

```
colnames(dataset) <- c("row_id", "order_id", "order_date", "ship_date", "ship_mode",
                       "segment", "country", "city", "state", "postal_code", "regio
                       "sub_category", "product_name", "sales", "quantity", "discou

colnames(dataset)
```

'row_id' 'order_id' 'order_date' 'ship_date' 'ship_mode' 'customer_id'
'customer_name' 'segment' 'country' 'city' 'state' 'postal_code' 'region'
'product_id' 'category' 'sub_category' 'product_name' 'sales' 'quantity' 'discount'
'profit'

Task 1

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

In [7]:

```
dataset$lead_time <- dataset$ship_date - dataset$order_date
```

In [8]:

```
dataset %>% select(ship_date, order_date, lead_time) %>% head()
```

ship_date	order_date	lead_time
2016-11-11	2016-11-08	259200 secs
2016-11-11	2016-11-08	259200 secs
2016-06-16	2016-06-12	345600 secs
2015-10-18	2015-10-11	604800 secs
2015-10-18	2015-10-11	604800 secs
2014-06-14	2014-06-09	432000 secs

In [9]:

```
library(lubridate)
```

```
Attaching package: ‘lubridate’

The following object is masked from ‘package:base’:

    date
```

In [10]:

```
task1 <- dataset %>%
        group_by(sub_category) %>%
        summarise(avg_lt = mean(lead_time), sd_lt = sd(lead_time)) %>%
        arrange(desc(avg_lt))

task1 %>% head()
```

sub_category	avg_lt	sd_lt
Art	350267.3 secs	150888.8 secs
Binders	347528.8 secs	150421.4 secs
Supplies	346964.2 secs	161572.7 secs
Envelopes	346960.6 secs	149741.5 secs
Labels	345837.4 secs	151424.9 secs
Phones	345697.2 secs	147584.5 secs

In [11]:

```
class(task1$avg_lt)
```

'difftime'

- avg_lt와 sd_lt를 구했지만, 단위가 초(second)로 되어있음.
- `class(task1$avg_lt)` 를 실행해보니 `class(task1$avg_lt)` 라고 함.
- 그러므로 google에서 "convert difftime second to days"를 검색하여 아래와 같은 해결책을 얻고 task1을 완료.

In [12]:

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

task1$sd_lt <- task1$sd_lt %>%
                as.numeric(units = "days") %>%
                round(2)

print(task1)
```

```
# A tibble: 17 x 3
   sub_category avg_lt sd_lt
   <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     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         3.89  1.8
13 Paper          3.89  1.75
14 Accessories    3.89  1.72
15 Bookcases      3.81  1.68
16 Machines       3.75  1.99
17 Copiers        3.62  1.88
```

- 품목별로 평균적인 leadTime의 크기가 크지 않음.
- 표준편차의 차이도 별로 크지 않음.
- 평균과 표준편차 외에 전체적인 분포를 task2에서 살펴보도록 함.

In [ ]:

```

```

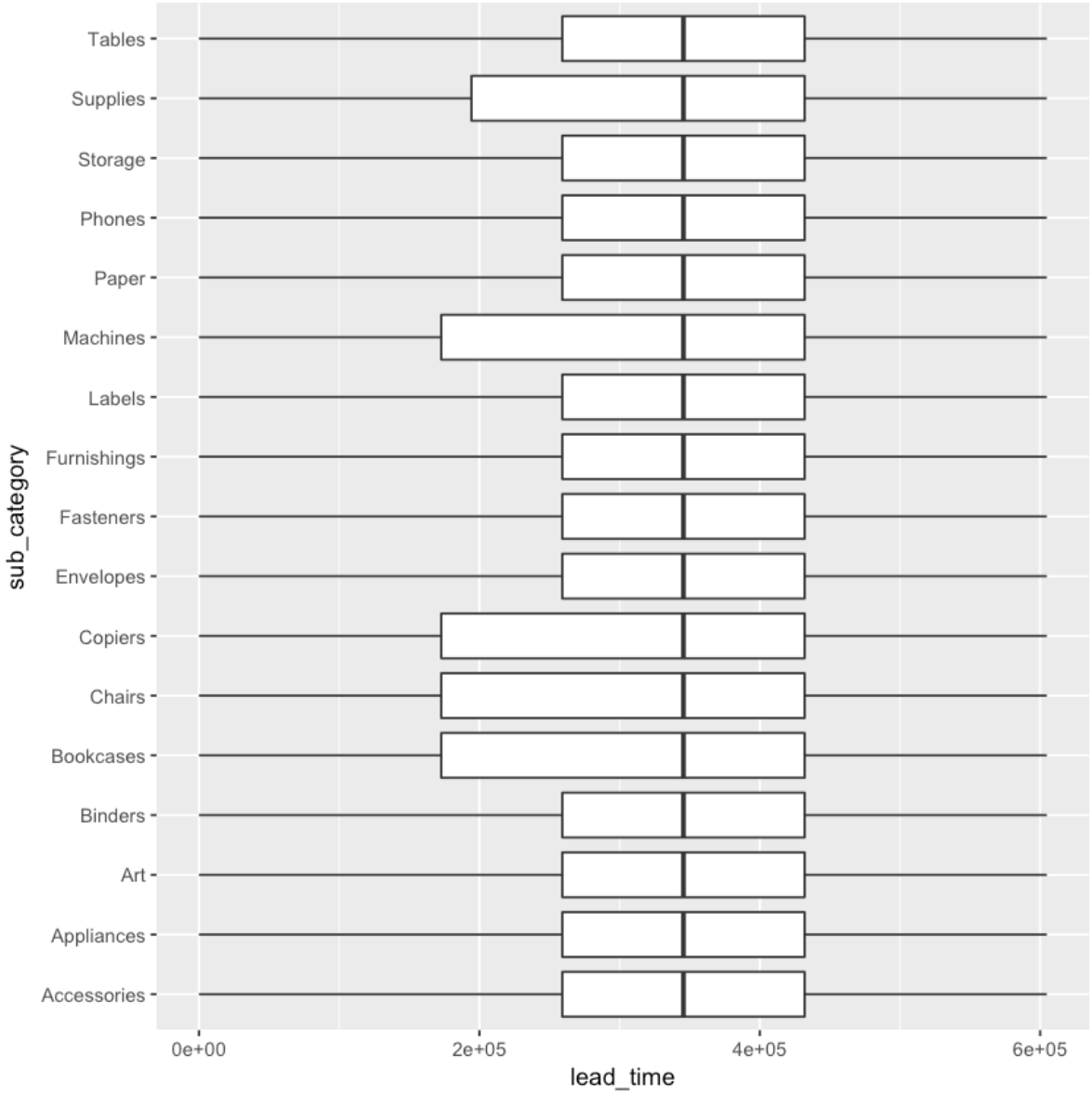
Task 2

각각의 Sub-Category에 대해서 box-plot을 그린다.

In [13]:

```
ggplot(dataset) +
    geom_boxplot(aes(x = sub_category, y = lead_time)) +
    coord_flip()
```

```
Don't know how to automatically pick scale for object of type difftim
e. Defaulting to continuous.
```



In [14]:

```
dataset$lead_time %>% head()
```

```
Time differences in secs
[1] 259200 259200 345600 604800 604800 432000
```

In [15]:

```
class(dataset$lead_time)
```

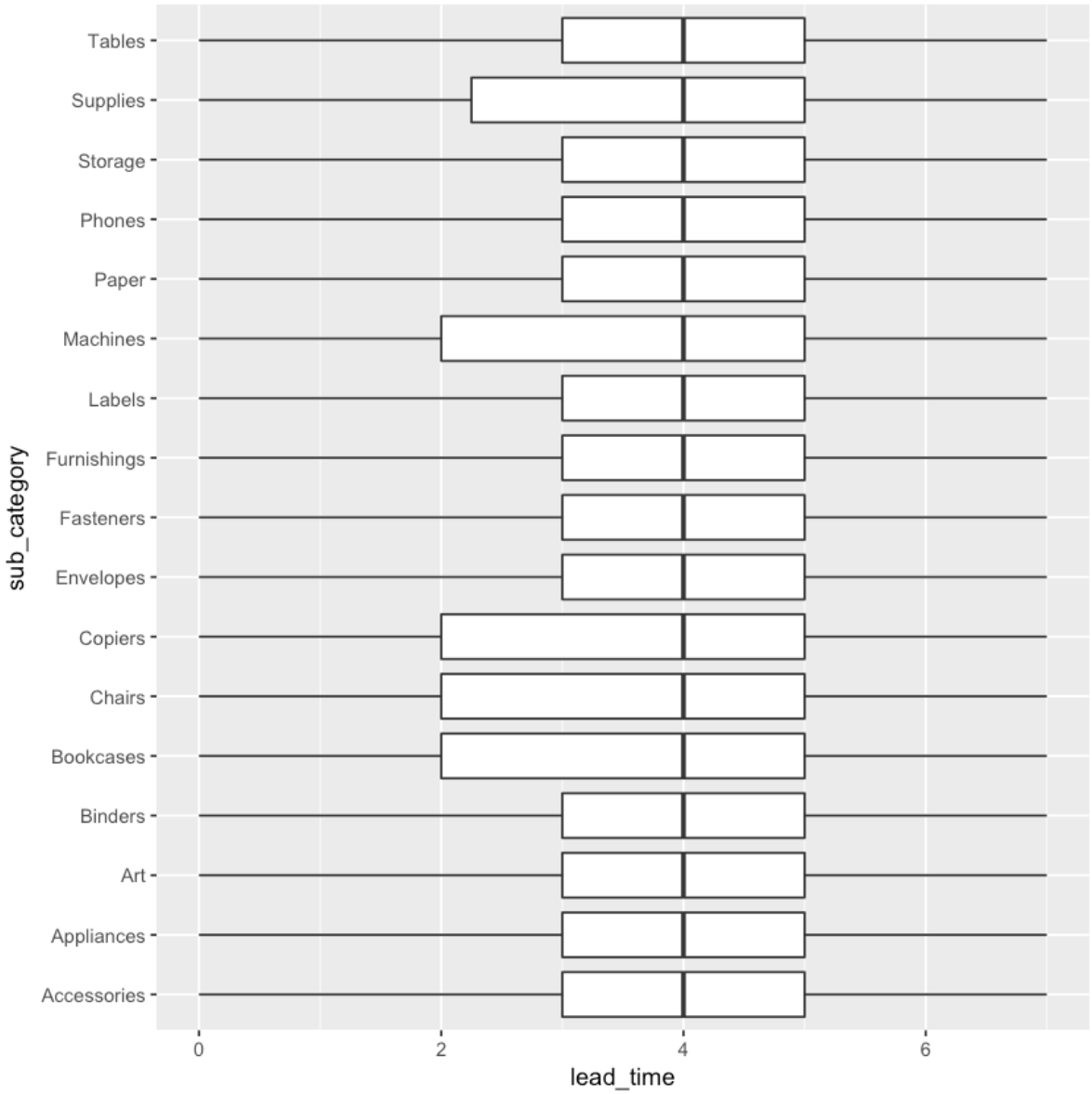
'difftime'

- leadTime의 수치가 뭔가 괴상하게 나옴.
- 원인을 살펴보니 leadTime변수 자체가 seconds단위로 되어 있음.

- 위에서 days로 바꾸었던 방법을 이용해서 해결.

In [16]:

```
dataset$lead_time <- dataset$lead_time %>% as.numeric(units = "days")

ggplot(dataset) +
    geom_boxplot(aes(x = sub_category, y = lead_time)) +
    coord_flip()
```



In [17]:

```
nrow(dataset)

max(dataset$lead_time)

sum(dataset$lead_time==max(dataset$lead_time))

sum(dataset$lead_time==max(dataset$lead_time))/nrow(dataset)*100
```

9994

7

621

6.21372823694217

Pythonic 한 방법들이 있는지?

In [18]:

```
# max = max(dataset$lead_time)
# print('ㅇ'+ max)
```

위의 그림으로 부터 lead_time에 대해서 아래와 같은 결론을 내릴 수 있음.

- 모든 상품에 대해서 중간값이 4일이다.
- Furnishing, Machines, Chairs, Bookcases와 같이 크기가 클 수 있는 제품의 경우에도 leadTime이 특별히 길다고 말할 수 없다.
- 전체 관찰값의 갯수는 `nrow(dataset)` : 9,994이다.
- 그런데 `max(dataset$leadTime)` :7이다.
- leadTime이 7인 경우는 총
  - `sum(dataset$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 leadTime을 7일 내로 100% 처리했으며, 6일내로 94% 처리했다.
- lead_time이 대체로 잘 관리되고 있음을 알 수 있다.

In [ ]:

```

```

Task 3

leadTime이 가장 긴 20개의 관찰값을 출력한다.

- 만약에 lead_time의 분포가 long-tail의 모습을 보인다면, 즉 몇몇 제품의 lead_time이 이상하게 높았다면, 이런 케이스를 나열하고 분석하는 것이 Task 3의 목적.
- 그러나 Task 2의 분석 결과를 보면 leadTime이 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 leadTime이 7일인 경우는 어떤 상품들이 많이 있는지 확인.

아래와 같은 관찰을 수행할 수 있다.

- Task 3-1: lead_time이 7인 주문을 category와 sub_category로 나누어서 갯수를 관찰한다. category로 나눈 것은 pie-chart로 그리고 sub-category로 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, lead_time이 7일인 Furniture 주문의 갯수를 보는 것이 아니라, 전체 Furniture 주문 중에서 lead_time이 7일인 경우의 비율을 계산하는 것이 바람직하다.

Task 3-2 : 전체 Furniture 주문 중에서 leadTime이 7일인 경우의 비율을 계산하는 것

In [19]:

```
# For Category
task3_2a <- dataset %>%
    group_by(category) %>%
    summarise(max_lead_time_percent = 100*sum(lead_time==7)/length(lead_time)) %>%
    arrange(desc(max_lead_time_percent))

print(task3_2a)
```

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

In [20]:

```
task3_2a
```

category	max_lead_time_percent
Office Supplies	6.372386
Furniture	6.082037
Technology	5.847320

In [21]:

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

print(task3_2b)
```

```
# A tibble: 17 x 2
   sub_category max_lead_time_percent
   <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 leadTime이 7일에 해당했음.
- 그러므로 발주 process 효율성의 향상을 위해서는 Supplies물품을 주문 받은 이후 배송을 시행하기 까지 왜 오래 걸리는지 알아볼 필요가 있다는 결론을 내릴 수 있음!

In [ ]:

```
```

5-2. 마진이 가장 많이 남는 상품은 무엇인가?

Background & Strategy

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

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

Tasks Specification

- Task 1. 각 Category와 Sub-Category에 대해서 Sales와 Profit을 각각 Aggregate한다. 그리고 Profit을 Sales로 나누어서 profitRatio을 구한다.
- Task 2. 분기를 나타내는 변수를 생성하고 위의 분석을 반복한다.

Task 1

In [22]:

```
colnames(dataset)
```

'row_id' 'order_id' 'order_date' 'ship_date' 'ship_mode' 'customer_id'
'customer_name' 'segment' 'country' 'city' 'state' 'postal_code' 'region'
'product_id' 'category' 'sub_category' 'product_name' 'sales' 'quantity' 'discount'
'profit' 'lead_time'

In [23]:

```
task1a <- dataset %>%
    group_by(category) %>%
    summarise(sales = sum(sales), profit = sum(profit)) %>%
    mutate(profit_ratio = round(profit/sales, 2)) %>%
    arrange(desc(profit_ratio))

print(task1a)
```

```
# A tibble: 3 x 4
  category          sales  profit profit_ratio
  <chr>             <dbl>   <dbl>        <dbl>
1 Office Supplies 719047. 122491.         0.17
2 Technology      836154. 145455.         0.17
3 Furniture       742000.  18451.         0.02
```

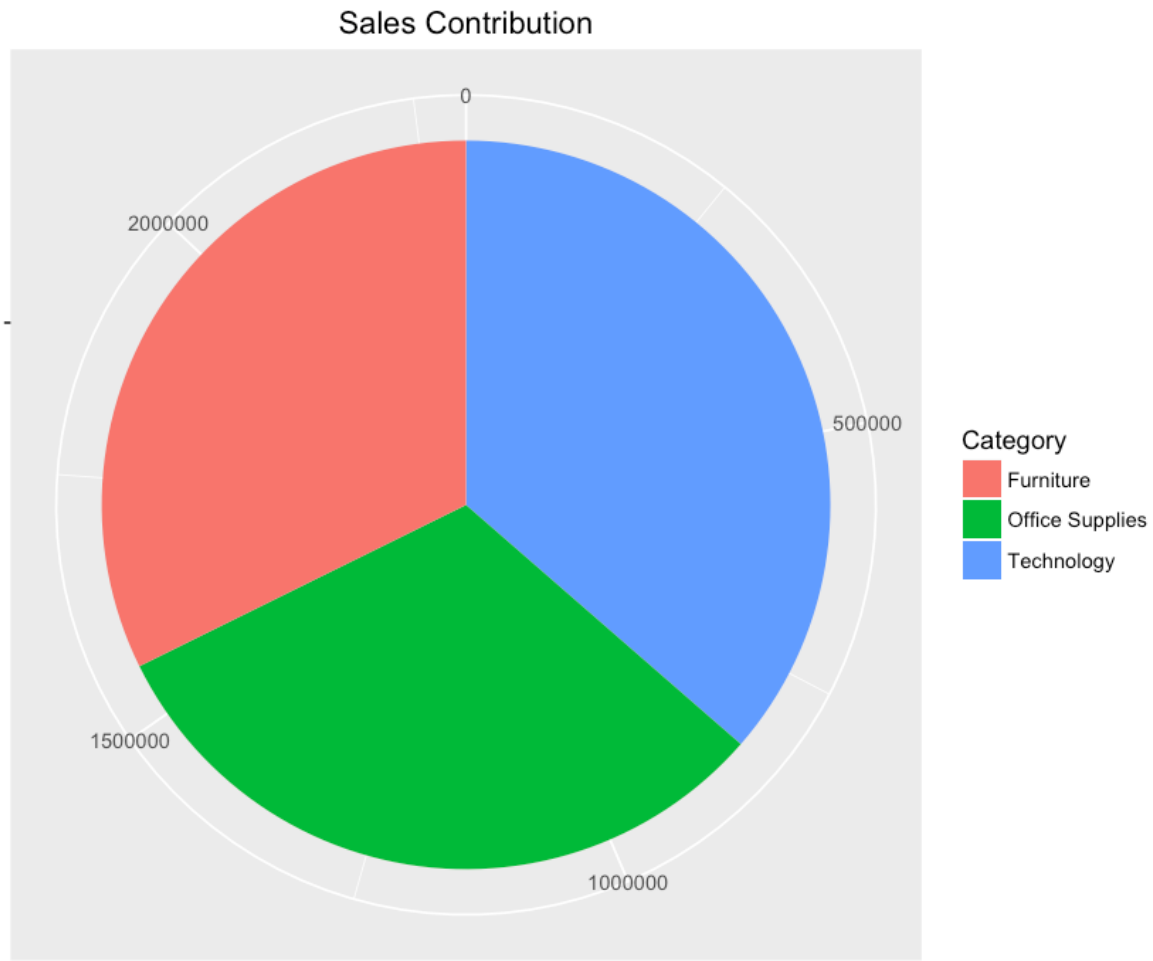
Sales

In [24]:

```
# 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")
```

In [25]:

```
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)
```



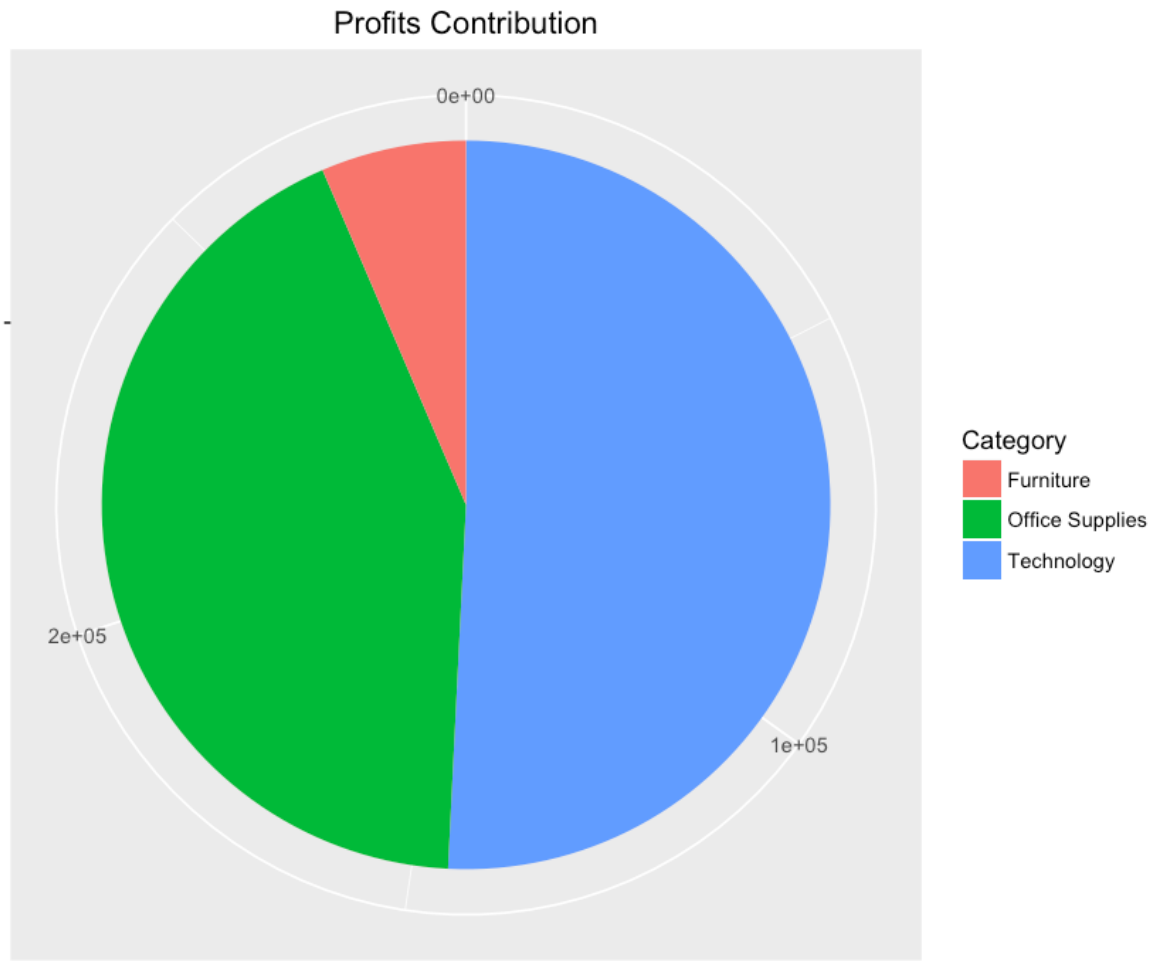
Profits

In [26]:

```
# 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))
```

In [27]:

```
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 보임.
- 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 골치 아픈 문제가 발생. 분류가 심지어 이익률도 2% ㅠㅠ

Furniture에 대해서 떠오르는 질문은 다음과 같음.

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 함.

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

In [28]:

```
task1b <- dataset %>%
    group_by(sub_category) %>%
    summarise(sales = sum(sales), profit = sum(profit)) %>%
    mutate(profit_ratio = round(profit/sales, 2)) %>%
    arrange(desc(profit_ratio))

task1b
```

sub_category	sales	profit	profit_ratio
Labels	12486.31	5546.2540	0.44
Paper	78479.21	34053.5693	0.43
Envelopes	16476.40	6964.1767	0.42
Copiers	149528.03	55617.8249	0.37
Fasteners	3024.28	949.5182	0.31
Accessories	167380.32	41936.6357	0.25
Art	27118.79	6527.7870	0.24
Appliances	107532.16	18138.0054	0.17
Binders	203412.73	30221.7633	0.15
Furnishings	91705.16	13059.1436	0.14
Phones	330007.05	44515.7306	0.13
Storage	223843.61	21278.8264	0.10
Chairs	328449.10	26590.1663	0.08
Machines	189238.63	3384.7569	0.02
Bookcases	114880.00	-3472.5560	-0.03
Supplies	46673.54	-1189.0995	-0.03
Tables	206965.53	-17725.4811	-0.09

위의 표는 이처럼 Diverging bar로 표현할 수도 있음

In [29]:

```
task1b %>% head()
```

sub_category	sales	profit	profit_ratio
Labels	12486.31	5546.2540	0.44
Paper	78479.21	34053.5693	0.43
Envelopes	16476.40	6964.1767	0.42
Copiers	149528.03	55617.8249	0.37
Fasteners	3024.28	949.5182	0.31
Accessories	167380.32	41936.6357	0.25

In [30]:

```
task1b$profit_hl <-
    ifelse(task1b$profit_ratio < mean(task1b$profit_ratio),
           "below average", "above average")

task1b <- task1b %>% arrange(profit_ratio)
```

In [31]:

```
task1b
```

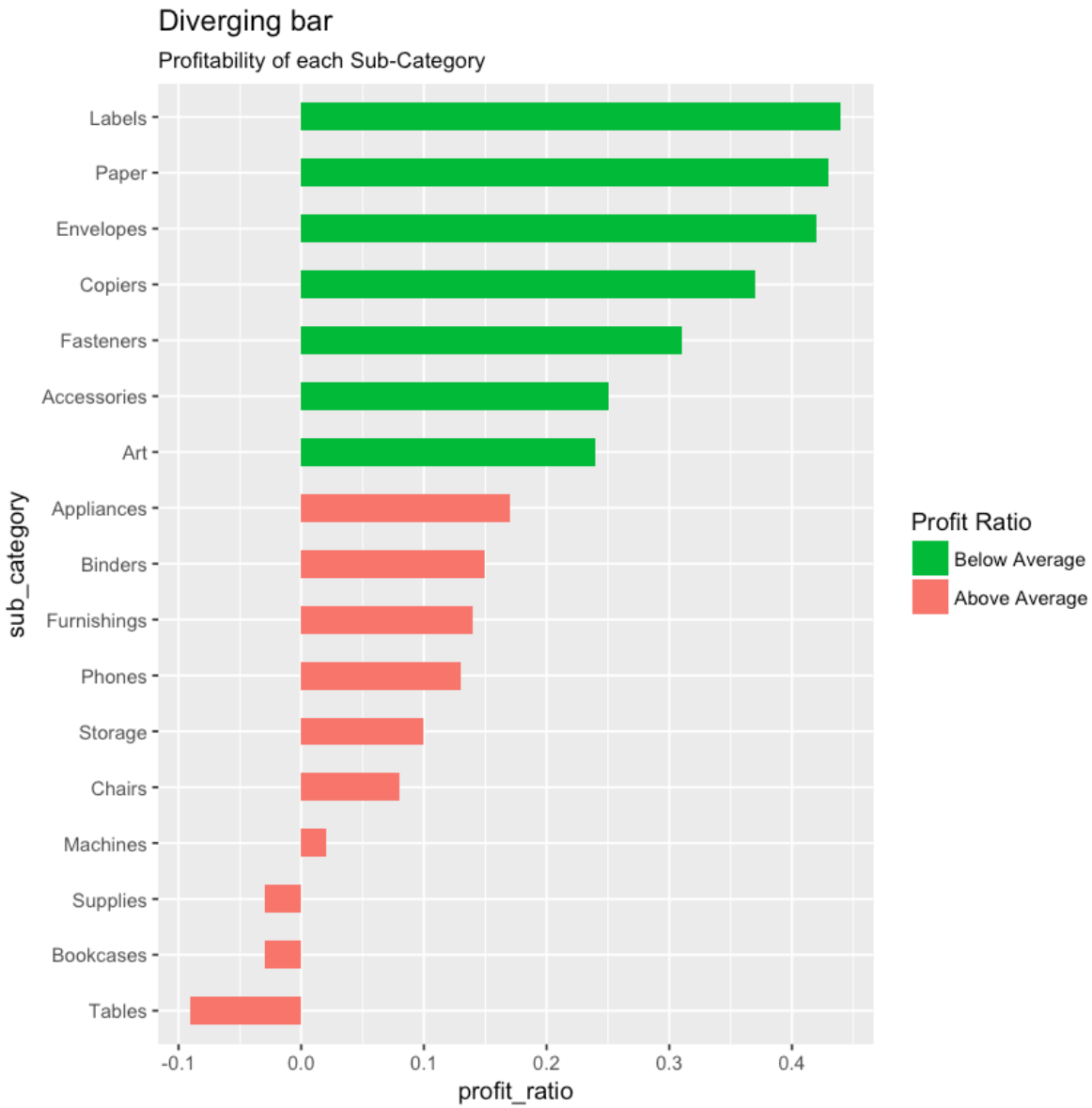
sub_category	sales	profit	profit_ratio	profit_hl
Tables	206965.53	-17725.4811	-0.09	below average
Bookcases	114880.00	-3472.5560	-0.03	below average
Supplies	46673.54	-1189.0995	-0.03	below average
Machines	189238.63	3384.7569	0.02	below average
Chairs	328449.10	26590.1663	0.08	below average
Storage	223843.61	21278.8264	0.10	below average
Phones	330007.05	44515.7306	0.13	below average
Furnishings	91705.16	13059.1436	0.14	below average
Binders	203412.73	30221.7633	0.15	below average
Appliances	107532.16	18138.0054	0.17	below average
Art	27118.79	6527.7870	0.24	above average
Accessories	167380.32	41936.6357	0.25	above average
Fasteners	3024.28	949.5182	0.31	above average
Copiers	149528.03	55617.8249	0.37	above average
Envelopes	16476.40	6964.1767	0.42	above average
Paper	78479.21	34053.5693	0.43	above average
Labels	12486.31	5546.2540	0.44	above average

In [32]:

```
# 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 = profit_ratio, label = profit_ratio)) +
  geom_bar(stat = 'identity', aes(fill = profit_hl), 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()
```

In [33]:

```
print(a)
```



혹은 좀 더 modern look을 제공하는 아래와 같은 "Diverging Lollipip Chart"도 가능

In [34]:

```
# a <- ggplot(task1b,
#             aes(x = sub_category, y = profit_ratio, labe = profit_ratio)) +
#     geom_point(stat = 'identity', fill = "black", size = 8)

# a
```

In [35]:

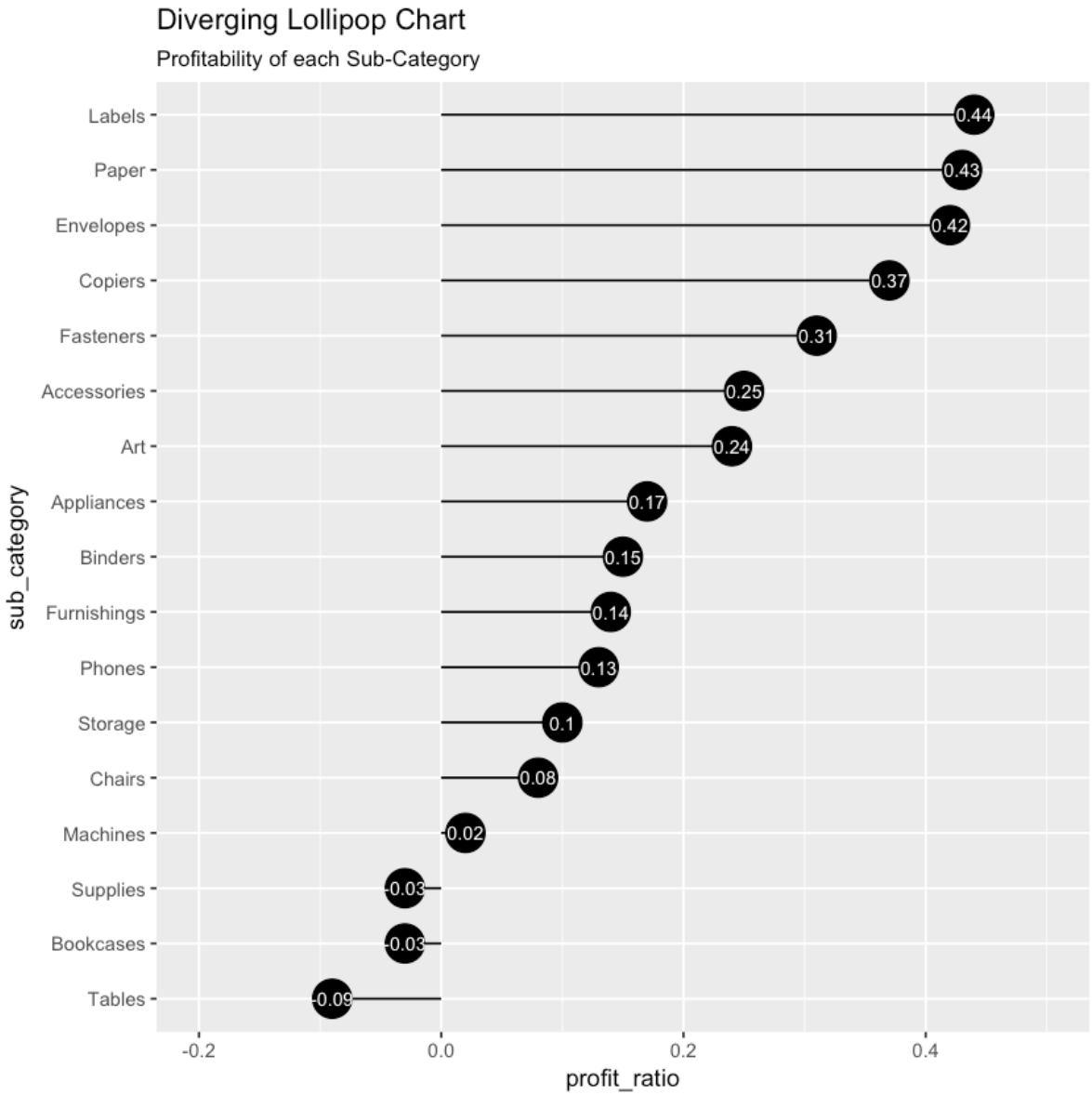
```
# a <- ggplot(task1b,
#             aes(x = sub_category, y = profit_ratio, labe = profit_ratio)) +
#     geom_point(stat = 'identity', fill = "black", size = 8) +
#     geom_segment(aes(y = 0, x = sub_category,
#                      yend = profit_ratio, xend = sub_category),
#                  color = 'black')

# a
```

In [36]:

```
# a <- ggplot(task1b,
#             aes(x = sub_category, y = profit_ratio, label = profit_ratio)) +
#   geom_point(stat = 'identity', fill = "black", size = 8) +
#   geom_segment(aes(y = 0, x = sub_category,
#                    yend = profit_ratio, xend = sub_category),
#                color = "black") +
#   geom_text(color = "white", size = 3)

# a
```

In [37]:

```
a <- ggplot(task1b,
            aes(x = sub_category, y = profit_ratio, label = profit_ratio)) +
  geom_point(stat = 'identity', fill = "black", size = 8) +
  geom_segment(aes(y = 0, x = sub_category,
                   yend = profit_ratio, 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()

a
```



- 이익률 하위 부분의 Storage, Chairs, Bookcases, Tables 모두 가구류에 해당.
  - 내가 만약에 이 기업을 경영자라면 해당 소형 가구라인의 유지를 전면적으로 고민할 것
- 이익률이 높은 Sub-Category들의 경우에는 이익률은 높지만 실제 이익의 총량은 얼마 안되는 품목들도 많이 있음.
  - Labels, Envelopes, Fastener, Art의 경우에는 이익 자체가 크지 않음. (봉투를 팔아서 돈을 벌면 얼마나 벌겠습니까...)
- 같은 table을 이익 순으로 정렬하는 것이 다른 시각을 제공할 수 있음.

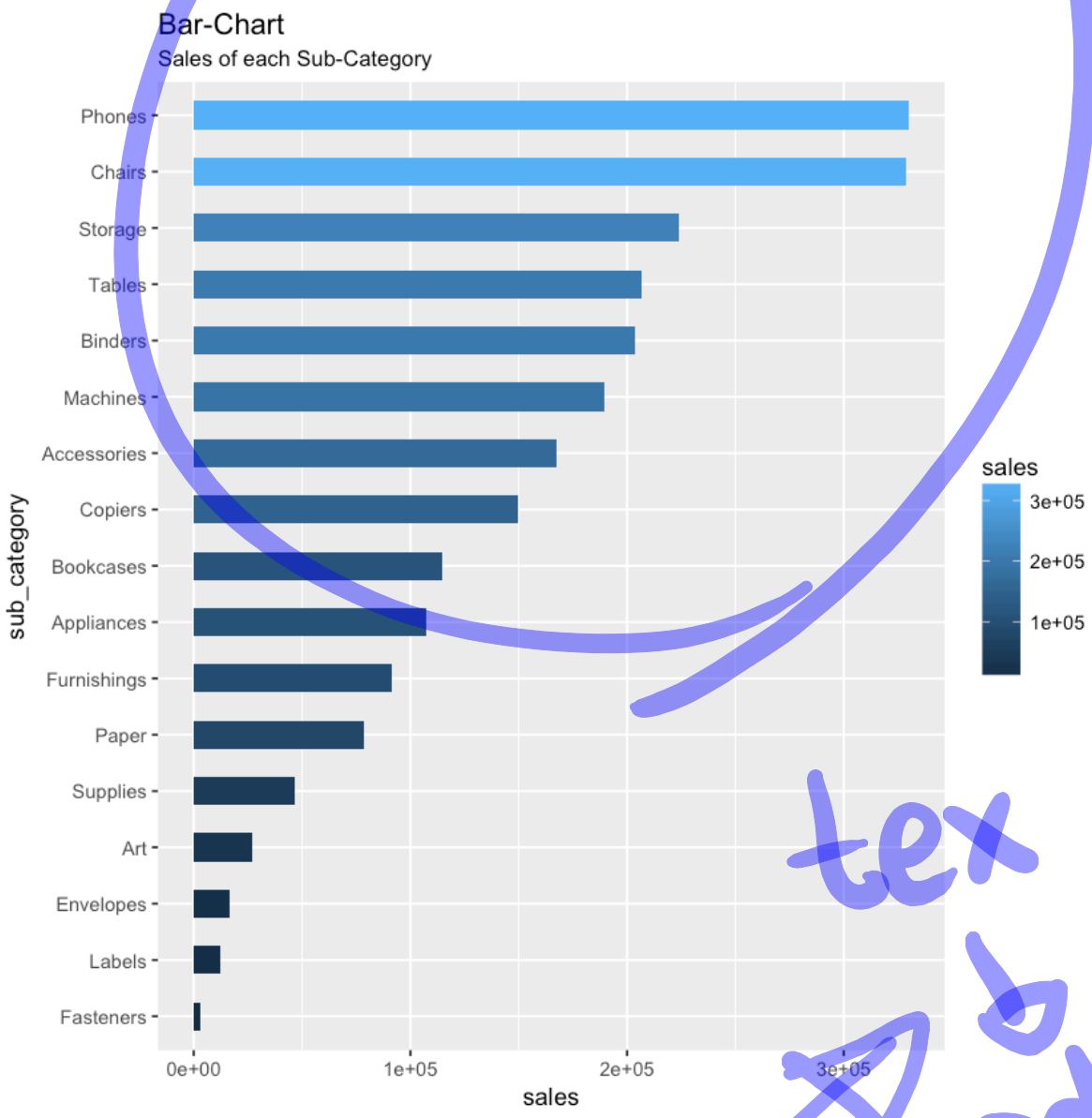
In [38]:

```
task1b %>% arrange(desc(profit, sales))
```

sub_category	sales	profit	profit_ratio	profit_hl
Copiers	149528.03	55617.8249	0.37	above average
Phones	330007.05	44515.7306	0.13	below average
Accessories	167380.32	41936.6357	0.25	above average
Paper	78479.21	34053.5693	0.43	above average
Binders	203412.73	30221.7633	0.15	below average
Chairs	328449.10	26590.1663	0.08	below average
Storage	223843.61	21278.8264	0.10	below average
Appliances	107532.16	18138.0054	0.17	below average
Furnishings	91705.16	13059.1436	0.14	below average
Envelopes	16476.40	6964.1767	0.42	above average
Art	27118.79	6527.7870	0.24	above average
Labels	12486.31	5546.2540	0.44	above average
Machines	189238.63	3384.7569	0.02	below average
Fasteners	3024.28	949.5182	0.31	above average
Supplies	46673.54	-1189.0995	-0.03	below average
Bookcases	114880.00	-3472.5560	-0.03	below average
Tables	206965.53	-17725.4811	-0.09	below average

In [39]:

```
a <- ggplot(task1b,
    aes(x = reorder(task1b$sub_category, sales), y = sales, label = sales)) +
    geom_bar(stat = 'identity', aes(fill = sales), width = .5) +
    labs(x = "sub_category",
         title = "Bar-Chart",
         subtitle = "Sales of each Sub-Category") +
    coord_flip()

a
```



Task 2

2. 분기를 나타내는 변수를 생성하고 위의 분석을 반복한다.

In [40]:

```
colnames(dataset)
head(dataset)
```

'row_id' 'order_id' 'order_date' 'ship_date' 'ship_mode' 'customer_id' 'customer_name' 'segment' 'country' 'city' 'state' 'postal_code' 'region' 'product_id' 'category' 'sub_category' 'product_name' 'sales' 'quantity' 'discount' 'profit' 'lead_time'

row_id	order_id	order_date	ship_date	ship_mode	customer_id	customer_name	segment	co
1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	U
2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	U
3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	U
4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	U
5	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	U
6	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	U

In [41]:

```
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))

str(task2)
head(task2)
```

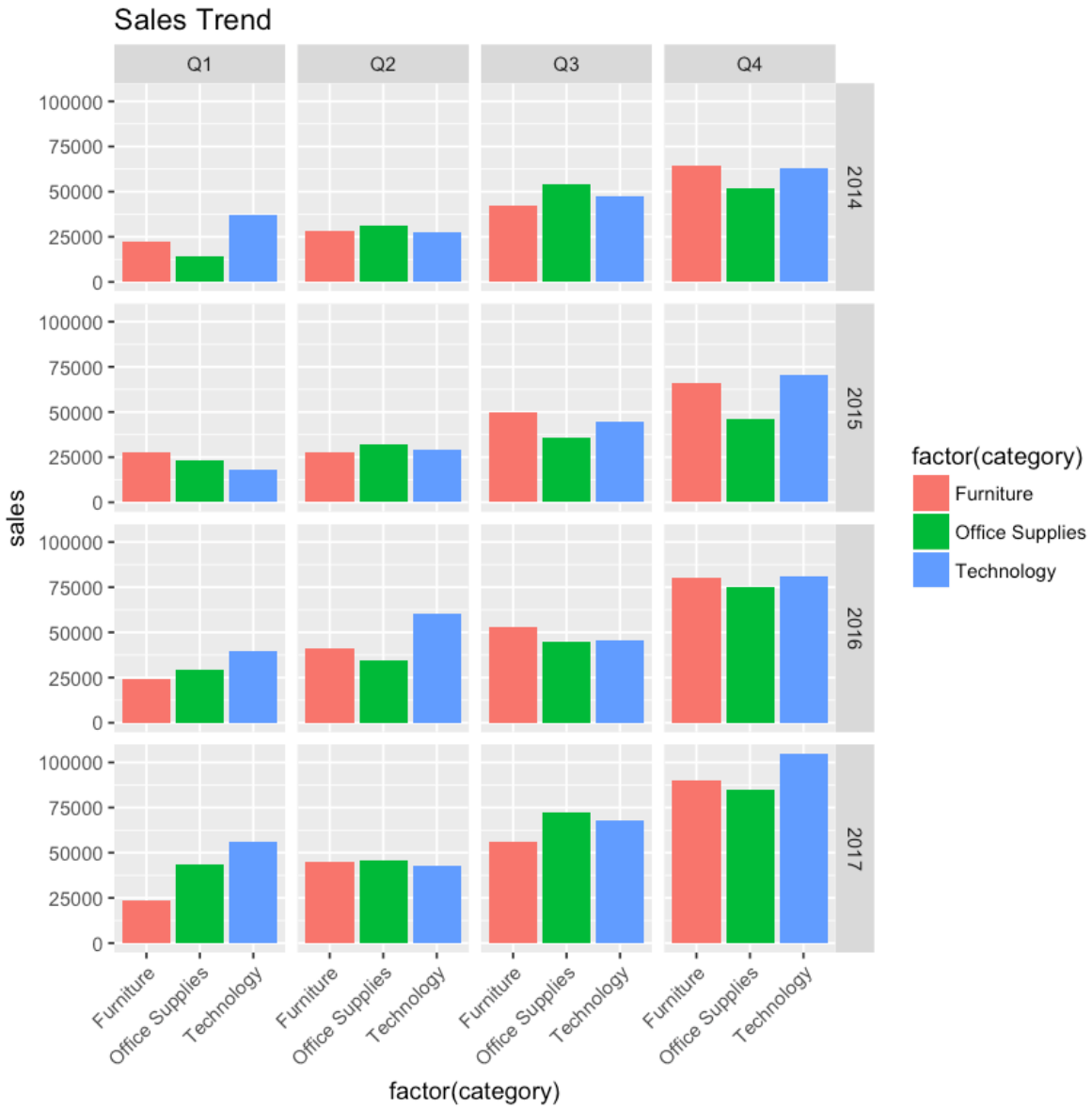
```
Classes ‘grouped_df’, ‘tbl_df’, ‘tbl’ and 'data.frame': 48 obs. of  5
variables:
 $ year    : Factor w/ 4 levels "2014","2015",..: 1 1 1 1 1 1 1 1 1 1
...
 $ quarter : Factor w/ 4 levels "Q1","Q2","Q3",..: 1 1 1 2 2 2 3 3 3 4
...
 $ category: chr  "Furniture" "Office Supplies" "Technology" "Furnitur
e" ...
 $ sales   : num  22656 14529 37263 28064 31244 ...
 $ profit  : num  -202 2235 1778 801 5779 ...
 - attr(*, "vars")= chr  "year" "quarter"
 - attr(*, "drop")= logi TRUE
```

year	quarter	category	sales	profit
2014	Q1	Furniture	22656.14	-202.4968
2014	Q1	Office Supplies	14528.68	2235.4549
2014	Q1	Technology	37262.97	1778.2709
2014	Q2	Furniture	28063.75	800.8178
2014	Q2	Office Supplies	31243.74	5778.8456
2014	Q2	Technology	27231.28	4624.4058

Sales Trend

In [42]:

```
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 있음.
  - 시계열의 이런 주기성을 계절성(seasonality)라고 함.
- 시계열 자료에서 경향의 구성 요소는 크게 3가지로 생각할 수 있음.
  - 1) 트렌드
  - 2) 계절성
  - 3) 그외의 잡음
- 우선 계절성은 어떤 고정된 길이의 시간에 따라서 주기적인 모습(cyclic pattern)을 보이는 것을 의미함.

Seasonality를 대처하는 방법

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

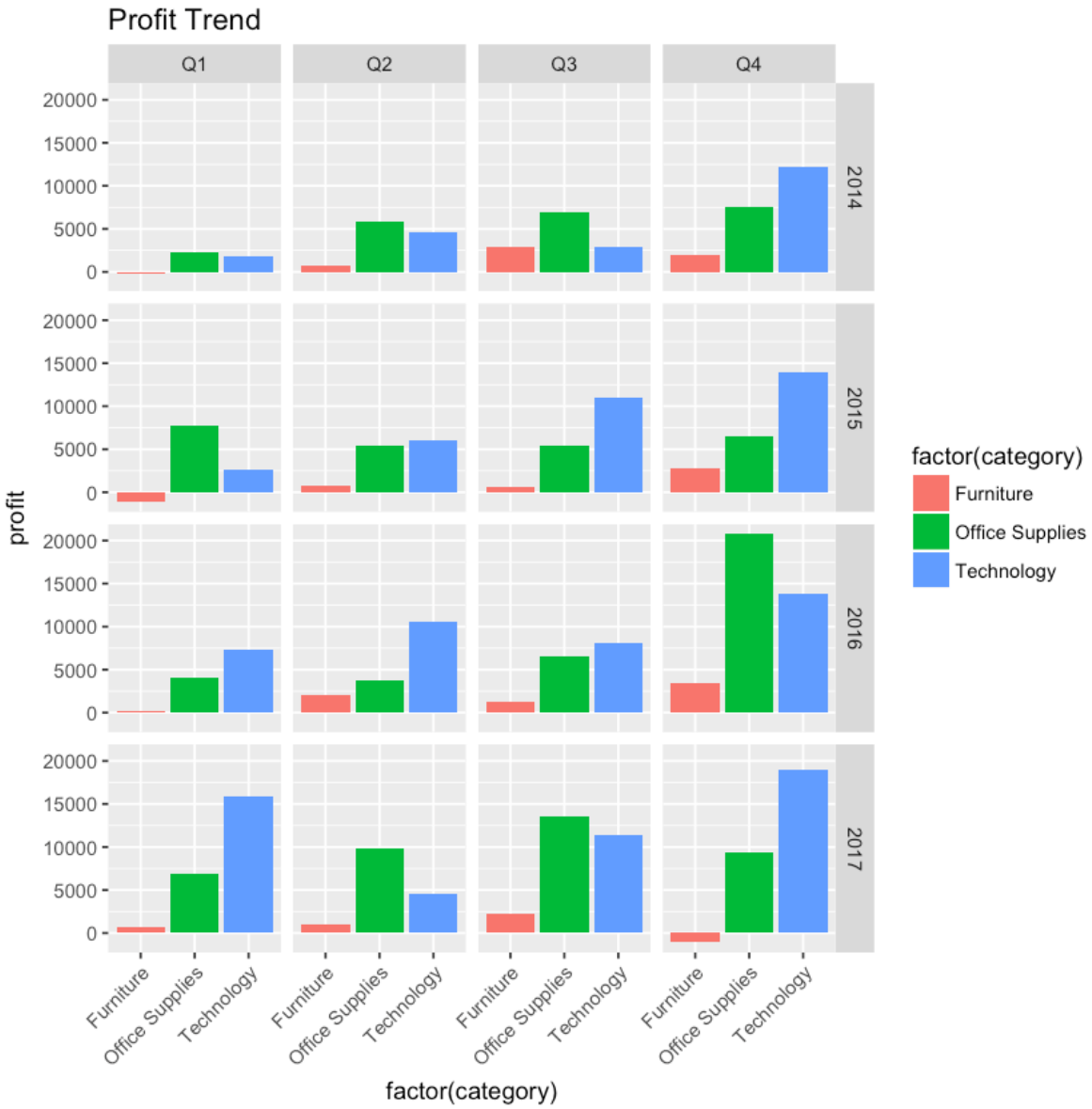
해당 기업의 자료

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

Profit Trend

In [43]:

```
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 분석
  - 앞에서 살펴본 매출과 비슷한 패턴을 보이는 것을 확인할 수 있음.
- 2014년 1분기에는 전년과 전전년 동분기에 비해서 Technology 제품에 대해서 큰 수익을 거둔것을 살펴볼 수 있음.
  - 만약에 조금 더 분석을 해본다면 2012년, 2013년, 2014년의 1분기에 대해서 각각 어떤 상품들이 팔렸는지 알고 싶음.
  - 예를 들어서 2014년 1분기에 아이폰의 새로운 버전이 나왔고 그것을 해당 쇼핑몰에서 많이 판매하였다면, 그것이 매출에 크게 기여하였다라고 말할 수 있음.
- 앞의 분석에서 문제로 제기했던 Furniture의 경우에는 2014년도 4분기에는 전년과 전전년 동분기에 대비해서 순이익이 적었음.
  - 이것 역시 이유를 더 살펴보고 2015년에의 Furniture 관련 전략을 수립할 필요가 있어보임.

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 Date를 기반으로 동시구매가 많이 일어나는 상품은 무엇인가?
- 9) 특정상품의 판매시기와 지역별 수요를 파악해보자

Task 3

2014년, 2015년, 2016년, 2017년의 1분기에 대해서 각각 어떤 상품들이 팔렸는지 알아보기

In [44]:

```
task3 <- 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, product_id, product_
  group_by(year, quarter, category, product_id, product_name, customer_id, customer_
  summarise(sales = sum(sales), profit = sum(profit))

task3$year <- factor(task3$year)
task3$quarter <- factor(paste0("Q", task3$quarter))

str(task3)
head(task3)
```

```
Classes ‘grouped_df’, ‘tbl_df’, ‘tbl’ and 'data.frame': 9982 obs. of
9 variables:
 $ year         : Factor w/ 4 levels "2014","2015",..: 1 1 1 1 1 1 1 1
1 1 ...
 $ quarter      : Factor w/ 4 levels "Q1","Q2","Q3",..: 1 1 1 1 1 1 1
1 1 1 ...
 $ category     : chr  "Furniture" "Furniture" "Furniture" "Furniture"
...
 $ product_id   : chr  "FUR-BO-10001337" "FUR-BO-10001972" "FUR-BO-100
01972" "FUR-BO-10003034" ...
 $ product_name : chr  "O'Sullivan Living Dimensions 2-Shelf Bookcase
s" "O'Sullivan 4-Shelf Bookcase in Odessa Pine" "O'Sullivan 4-Shelf Bo
okcase in Odessa Pine" "O'Sullivan Elevations Bookcase, Cherry Finish"
...
 $ customer_id  : chr  "GA-14725" "JS-15595" "TS-21340" "BD-11605" ...
 $ customer_name: chr  "Guy Armstrong" "Jill Stevenson" "Toby Swindel
l" "Brian Dahlen" ...
 $ sales        : num  206 302 181 334 62 ...
 $ profit       : num  -12.1 -199.62 -320.6 3.93 -53.29 ...
 - attr(*, "vars")= chr  "year" "quarter" "category" "product_id" ...
 - attr(*, "drop")= logi TRUE
```

year	quarter	category	product_id	product_name	customer_id	customer_name	sales	
2014	Q1	Furniture	FUR-BO-10001337	O'Sullivan Living Dimensions 2-Shelf Bookcases	GA-14725	Guy Armstrong	205.666	-12.
2014	Q1	Furniture	FUR-BO-10001972	O'Sullivan 4-Shelf Bookcase in Odessa Pine	JS-15595	Jill Stevenson	302.450	-199.
2014	Q1	Furniture	FUR-BO-10001972	O'Sullivan 4-Shelf Bookcase in Odessa Pine	TS-21340	Toby Swindell	181.470	-320.
2014	Q1	Furniture	FUR-BO-10003034	O'Sullivan Elevations Bookcase, Cherry Finish	BD-11605	Brian Dahlen	333.999	3.
2014	Q1	Furniture	FUR-BO-10003433	Sauder Cornerstone Collection Library	BS-11590	Brendan Sweed	61.960	-53.

year	quarter	category	product_id	product_name	customer_id	customer_name	sales	
2014	Q1	Furniture	FUR-BO-10003966	Sauder Facets Collection Library, Sky Alder Finish	LC-17050	Liz Carlisle	290.666	3.

2014~2017년 Q1의 Technology 카테고리 제품별 판매량

In [45]:

```
a <- task3 %>%
  filter(category == 'Technology' & quarter == 'Q1') %>%
  group_by(product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt))

a %>% head()
```

product_name	cnt
AT&T 17929 Lendline Telephone	5
Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	5
Aastra 57i VoIP phone	3
Cisco SPA525G2 IP Phone - Wireless	3
Hewlett Packard 310 Color Digital Copier	3
Kensington SlimBlade Notebook Wireless Mouse with Nano Receiver	3

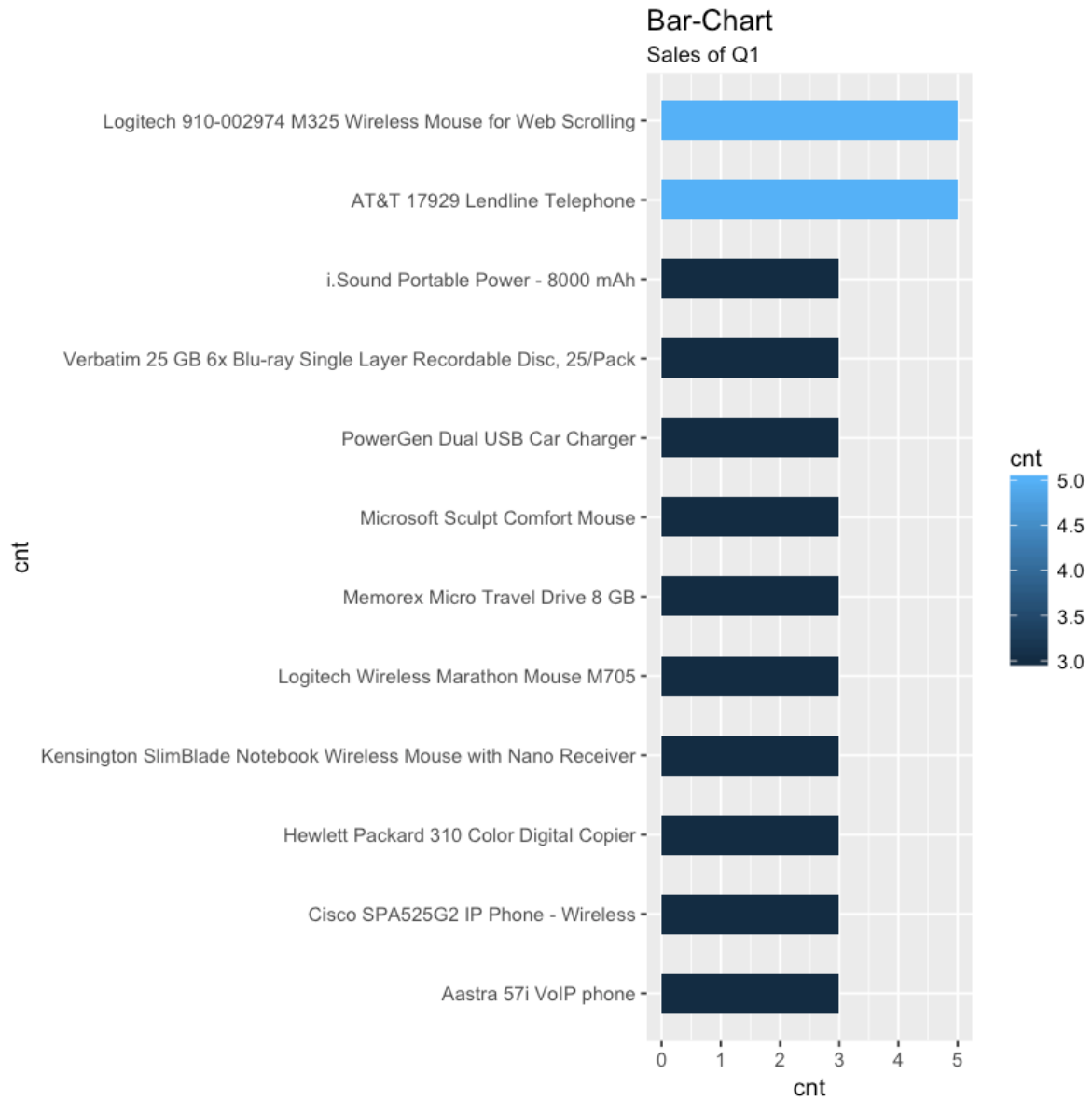
In [46]:

```
a <- a %>%
  filter(cnt >= 3)

a
```

product_name	cnt
AT&T 17929 Lendline Telephone	5
Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	5
Aastra 57i VoIP phone	3
Cisco SPA525G2 IP Phone - Wireless	3
Hewlett Packard 310 Color Digital Copier	3
Kensington SlimBlade Notebook Wireless Mouse with Nano Receiver	3
Logitech Wireless Marathon Mouse M705	3
Memorex Micro Travel Drive 8 GB	3
Microsoft Sculpt Comfort Mouse	3
PowerGen Dual USB Car Charger	3
Verbatim 25 GB 6x Blu-ray Single Layer Recordable Disc, 25/Pack	3
i.Sound Portable Power - 8000 mAh	3

In [47]:

```
ggplot(a,
  aes(x = reorder(product_name, cnt), y = cnt, label = cnt)) +
  geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
  labs(x = "cnt",
       title = "Bar-Chart",
       subtitle = "Sales of Q1") +
  coord_flip()
```



2014년 Q1의 Technology 카테고리 제품별 판매량

In [48]:

```
task3 %>%
  filter(category == 'Technology' & quarter == 'Q1' & year == 2014) %>%
  group_by(year, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head
```

year	product_name	cnt
2014	Enermax Aurora Lite Keyboard	2
2014	Hewlett-Packard Deskjet 6540 Color Inkjet Printer	2
2014	Maxell DVD-RAM Discs	2
2014	AT&T 17929 Lendline Telephone	1
2014	AT&T 841000 Phone	1
2014	AT&T CL82213	1

2015년 Q1의 Technology 카테고리 제품별 판매량

In [49]:

```
task3 %>%
  filter(category == 'Technology' & quarter == 'Q1' & year == 2015) %>%
  group_by(year, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head
```

year	product_name	cnt
2015	Square Credit Card Reader	2
2015	AT&T 17929 Lendline Telephone	1
2015	Anker 24W Portable Micro USB Car Charger	1
2015	Anker Ultrathin Bluetooth Wireless Keyboard Aluminum Cover with Stand	1
2015	Bady BDG101FRU Card Printer	1
2015	Belkin F8E887 USB Wired Ergonomic Keyboard	1

2016년 Q1의 Technology 카테고리 제품별 판매량

In [50]:

```
task3 %>%
  filter(category == 'Technology' & quarter == 'Q1' & year == 2016) %>%
  group_by(year, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head
```

year	product_name	cnt
2016	AT&T 17929 Lendline Telephone	3
2016	Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	3
2016	Jabra BIZ 2300 Duo QD Duo Corded Headset	2
2016	Verbatim 25 GB 6x Blu-ray Single Layer Recordable Disc, 25/Pack	2
2016	i.Sound Portable Power - 8000 mAh	2
2016	AT&T 1070 Corded Phone	1

2017년 Q1의 Technology 카테고리 제품별 판매량

In [51]:

```
task3 %>%
  filter(category == 'Technology' & quarter == 'Q1' & year == 2017) %>%
  group_by(year, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head
```

year	product_name	cnt
2017	Cisco SPA525G2 IP Phone - Wireless	2
2017	Geemarc AmpliPOWER60	2
2017	HP Officejet Pro 8600 e-All-In-One Printer, Copier, Scanner, Fax	2
2017	Hewlett Packard 310 Color Digital Copier	2
2017	Hewlett Packard LaserJet 3310 Copier	2
2017	Jawbone MINI JAMBOX Wireless Bluetooth Speaker	2

(2014~2017)년도별 Q1의 Technology 카테고리 제품별 판매량

In [162]:

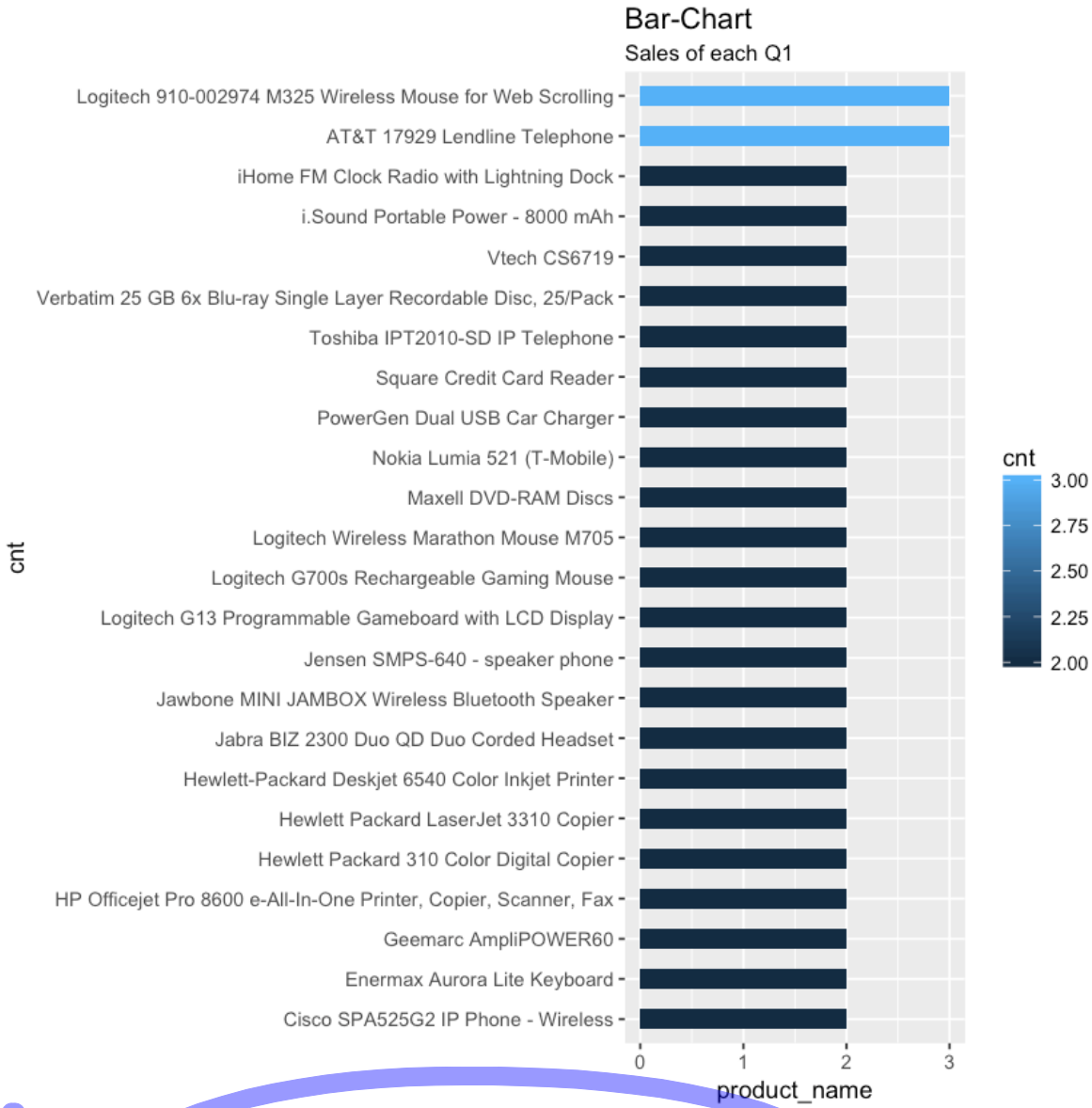
```
a <- task3 %>%
  filter(quarter == 'Q1' & category == 'Technology') %>%
  select(year, quarter, category, product_name) %>%
  group_by(year, quarter, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(year, desc(cnt)) %>%
  filter(cnt >= 2)

a
```

```
Adding missing grouping variables: `product_id`, `customer_id`
```

year	quarter	product_name	cnt
2014	Q1	Enermax Aurora Lite Keyboard	2
2014	Q1	Hewlett-Packard Deskjet 6540 Color Inkjet Printer	2
2014	Q1	Maxell DVD-RAM Discs	2
2015	Q1	Square Credit Card Reader	2
2016	Q1	AT&T 17929 Lendline Telephone	3
2016	Q1	Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	3
2016	Q1	Jabra BIZ 2300 Duo QD Duo Corded Headset	2
2016	Q1	Verbatim 25 GB 6x Blu-ray Single Layer Recordable Disc, 25/Pack	2
2016	Q1	i.Sound Portable Power - 8000 mAh	2
2017	Q1	Cisco SPA525G2 IP Phone - Wireless	2
2017	Q1	Geemarc AmpliPOWER60	2

In [55]:

```
ggplot(a,
  aes(x = reorder(a$product_name, cnt), y = a$cnt, label = cnt)) +
  geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
  labs(x = "cnt",
       y = "product_name",
       title = "Bar-Chart",
       subtitle = "Sales of each Q1")+
  coord_flip()
```



5-3. 가장 많은 상품을 구매한 고객은 누구이며 언제 구매하였는가?

Task 1

- 가장 많은 상품을 구매한 고객

In [56]:

```
# str(dataset)
# head(dataset)
```

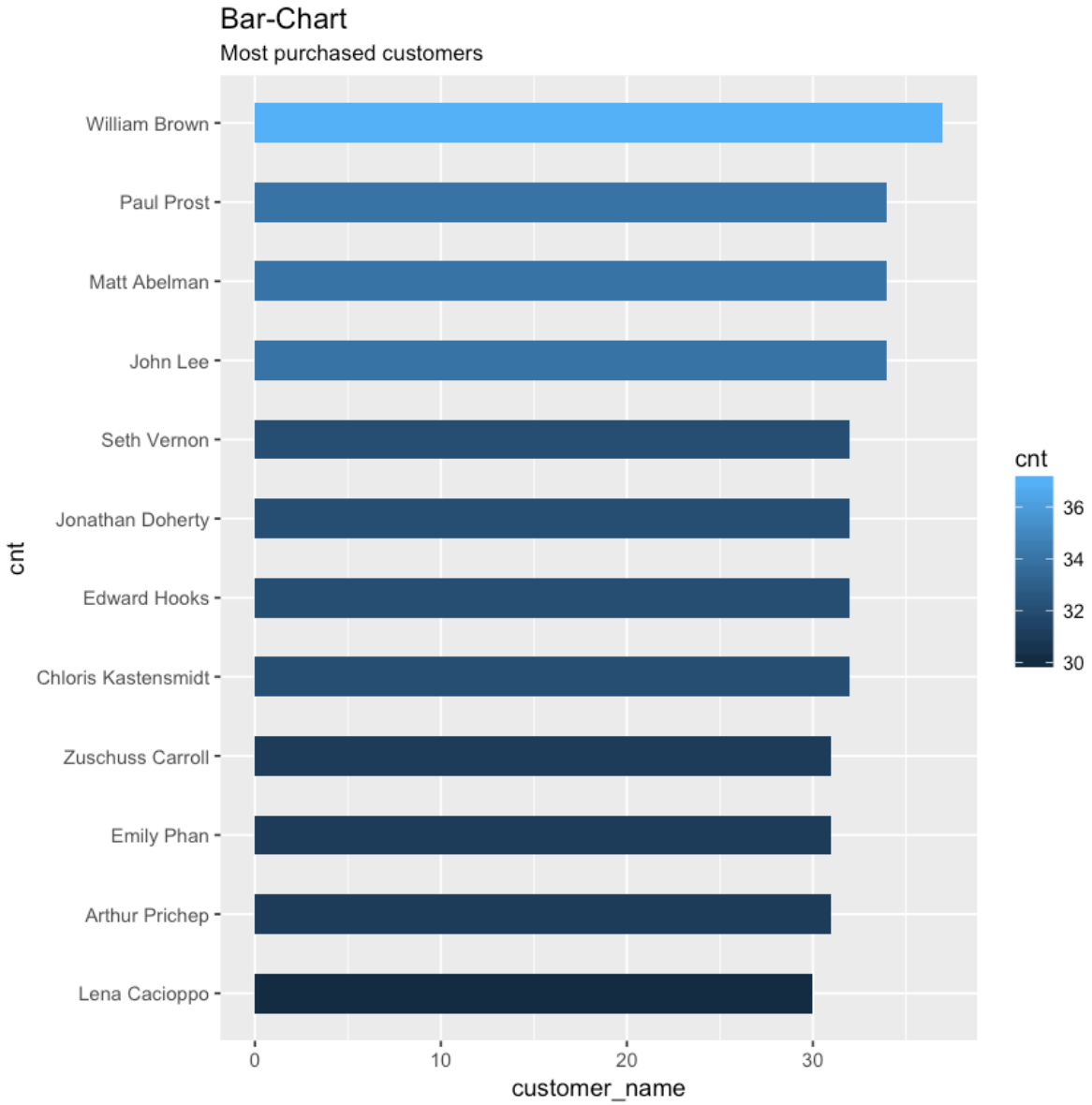
In [57]:

```
vip <- dataset %>% group_by(customer_id, customer_name) %>% summarise(cnt = n()) %>%
vip
```

customer_id	customer_name	cnt
WB-21850	William Brown	37
JL-15835	John Lee	34
MA-17560	Matt Abelman	34
PP-18955	Paul Prost	34
CK-12205	Chloris Kastensmidt	32
EH-13765	Edward Hooks	32
JD-15895	Jonathan Doherty	32
SV-20365	Seth Vernon	32
AP-10915	Arthur Prichep	31
EP-13915	Emily Phan	31
ZC-21910	Zuschuss Carroll	31
LC-16870	Lena Cacioppo	30

In [58]:

```
ggplot(vip,
  aes(x = reorder(customer_name, cnt), y = cnt, label = cnt)) +
  geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
  labs(x = "cnt",
       y = "customer_name",
       title = "Bar-Chart",
       subtitle = "Most purchased customers") +
  coord_flip()
```



In [59]:

```
# vip$customer_name <- factor(vip$customer_name)
# vip$cnt <- factor(vip$cnt)
str(vip)
```

```
Classes ‘grouped_df’, ‘tbl_df’, ‘tbl’ and 'data.frame': 12 obs. of  3
variables:
 $ customer_id  : chr  "WB-21850" "JL-15835" "MA-17560" "PP-18955" ...
 $ customer_name: chr  "William Brown" "John Lee" "Matt Abelman" "Paul
Prost" ...
 $ cnt          : int  37 34 34 34 32 32 32 32 31 31 ...
 - attr(*, "vars")= chr "customer_id"
 - attr(*, "drop")= logi TRUE
 - attr(*, "indices")=List of 12
  ..$ : int 8
  ..$ : int 4
  ..$ : int 5
  ..$ : int 9
  ..$ : int 6
  ..$ : int 1
  ..$ : int 11
  ..$ : int 2
  ..$ : int 3
  ..$ : int 7
  ..$ : int 0
  ..$ : int 10
 - attr(*, "group_sizes")= int  1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "biggest_group_size")= int 1
 - attr(*, "labels")='data.frame':      12 obs. of  1 variable:
  ..$ customer_id: chr  "AP-10915" "CK-12205" "EH-13765" "EP-13915"
...
  ..- attr(*, "vars")= chr "customer_id"
  ..- attr(*, "drop")= logi TRUE
```

?

- Diverging Lollipop Chart를 활용하려면?

In [60]:

```
a <- ggplot(vip,
            aes(x = reorder(customer_name, cnt), y = cnt, label = cnt)) +
            geom_point(stat = 'identity', fill = "black", size = 8) +
            geom_segment(aes(y = 0, x = cnt,
                             yend = customer_name, xend = cnt),
                         colamor = "black") +
            geom_text(color = "white", size = 3) +
      labs(x = "cnt",
           y = "customer_name",
           title = "Diverging Lollipop Chart",
           subtitle = "VIP") +
      ylim(-0.2, 0.5) +
      coord_flip()
a
```

```
6. withCallingHandlers({
 .     rpr <- mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler)
7. mime2repr[[mime]](obj)
8. repr_text.default(obj)
9. paste(capture.output(print(obj)), collapse = "\n")
10. capture.output(print(obj))
11. evalVis(expr)
12. withVisible(eval(expr, pf))
13. eval(expr, pf)
14. eval(expr, pf)
15. print(obj)
16. print.ggplot(obj)
17. ggplot_build(x)
18. layout$train_position(data, scale_x(), scale_y())
19. f(..., self = self)
```

Task 2

- 구매가 많이 일어난 도시와 구매가 적게 일어난 도시
  - 추후에 마케팅에 활용

?

- 총 531개의 도시가 존재
  - city라는 컬럼을 생성해서 총 city의 수를 세려면?

In [61]:

```
a <- dataset %>% group_by(city) %>% summarise(cnt = n())
str(a)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       531 obs. of  2 variabl
es:
 $ city: chr  "Aberdeen" "Abilene" "Akron" "Albuquerque" ...
 $ cnt : int  1 1 21 14 16 4 7 2 10 27 ...
```

In [62]:

```
good_score_city <- dataset %>% group_by(city) %>% summarise(cnt = n()) %>% arrange(c
bad_score_city <- dataset %>% group_by(city) %>% summarise(cnt = n()) %>% arrange(de


str(good_score_city)
head(good_score_city)
str(bad_score_city)
head(bad_score_city)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       13 obs. of  2 variable
s:
 $ city: chr  "New York City" "Los Angeles" "Philadelphia" "San Franci
sco" ...
 $ cnt : int  915 747 537 510 428 377 314 222 170 163 ...
```

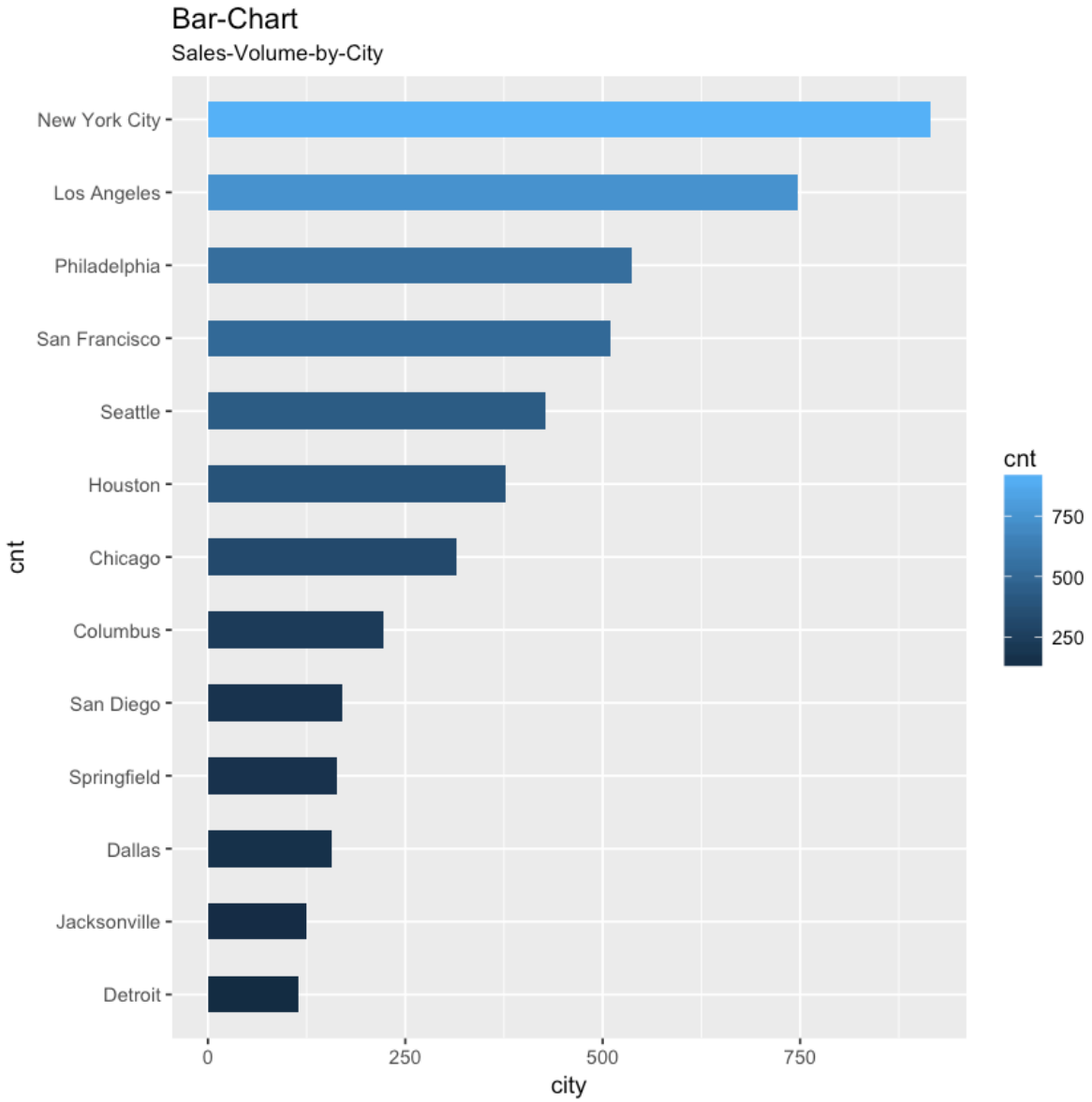
city	cnt
New York City	915
Los Angeles	747
Philadelphia	537
San Francisco	510
Seattle	428
Houston	377

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       70 obs. of  2 variable
s:
 $ city: chr  "Aberdeen" "Abilene" "Antioch" "Arlington Heights" ...
 $ cnt : int  1 1 1 1 1 1 1 1 1 1 ...
```

city	cnt
Aberdeen	1
Abilene	1
Antioch	1
Arlington Heights	1
Atlantic City	1
Bartlett	1

In [63]:

```
ggplot(good_score_city,
  aes(x = reorder(city, cnt), y = cnt, label = cnt)) +
  geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
  labs(x = "cnt",
       y = "city",
       title = "Bar-Chart",
       subtitle = "Sales-Volume-by-City") +
  coord_flip()
```



Task 3

- 가장 많이 구매한 고객은 언제 구매하였는가?

In [64]:

```
dataset %>%
  filter(customer_id == "WB-21850") %>%
  select(row_id, customer_id, customer_name, order_id, order_date, category, sub_cat
  arrange(order_date)
```

row_id	customer_id	customer_name	order_id	order_date	category	sub_category	product_id	product_name
4309	WB-21850	William Brown	CA-2014-125829	2014-11-04	Technology	Phones	TEC-PH-10001079	Polycom SoundPoint Pro SE-225 Corded phone
4310	WB-21850	William Brown	CA-2014-125829	2014-11-04	Furniture	Tables	FUR-TA-10002041	Bevis Round Conference Table Top, X-Base
4311	WB-21850	William Brown	CA-2014-125829	2014-11-04	Office Supplies	Binders	OFF-BI-10001036	Cardinal EasyOpen D-Ring Binders
4312	WB-21850	William Brown	CA-2014-125829	2014-11-04	Office Supplies	Paper	OFF-PA-10000223	Xerox 2000
4313	WB-21850	William Brown	CA-2014-125829	2014-11-04	Technology	Machines	TEC-MA-10003246	Hewlett-Packard Deskjet D4360 Printer

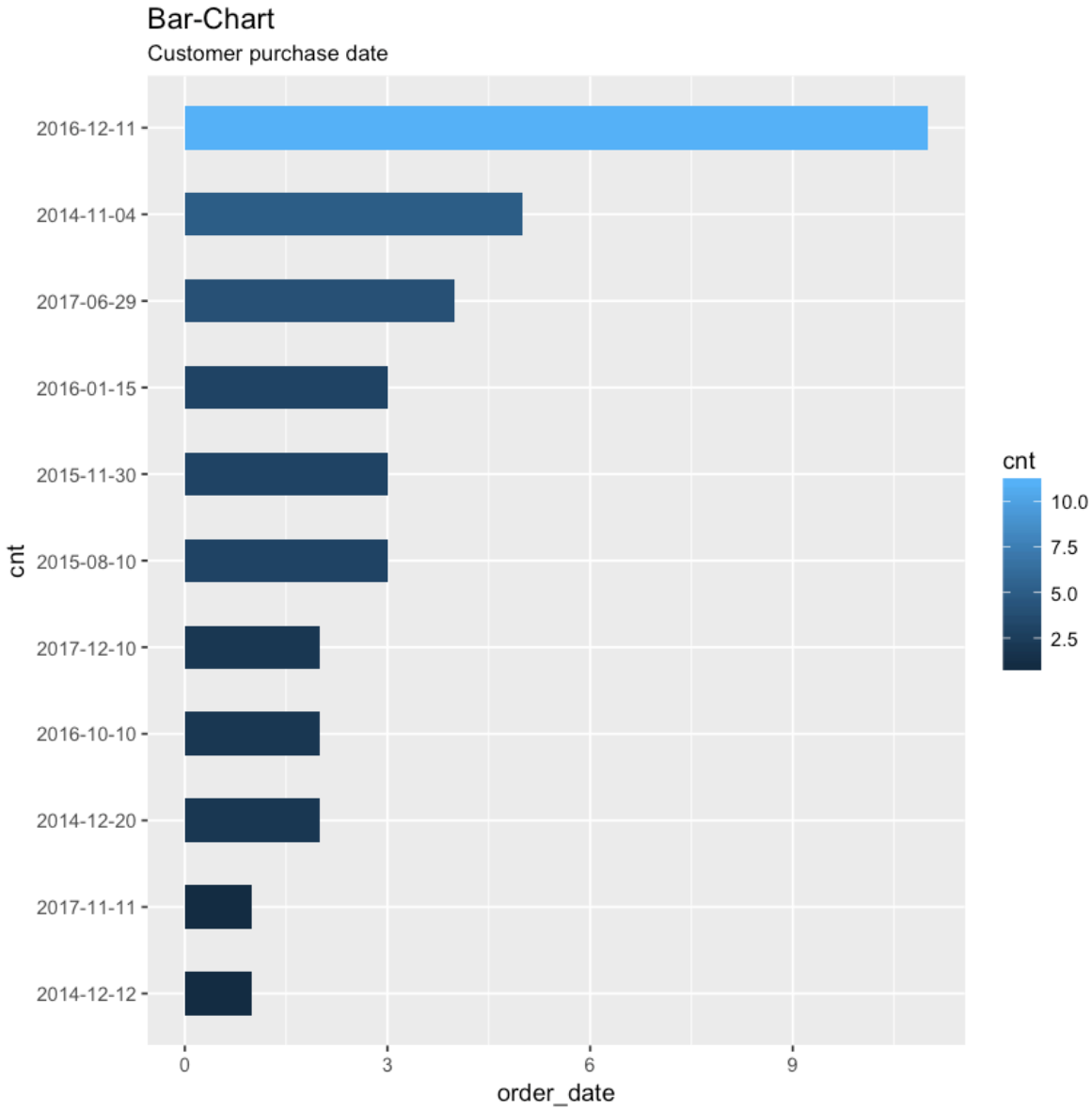
In [65]:

```
a <- dataset %>%
  filter(customer_id == "WB-21850") %>%
  group_by(order_date, customer_id, customer_name) %>%
  summarise(cnt = n()) %>%
  arrange(order_date)
a
```

order_date	customer_id	customer_name	cnt
2014-11-04	WB-21850	William Brown	5
2014-12-12	WB-21850	William Brown	1
2014-12-20	WB-21850	William Brown	2
2015-08-10	WB-21850	William Brown	3
2015-11-30	WB-21850	William Brown	3
2016-01-15	WB-21850	William Brown	3
2016-10-10	WB-21850	William Brown	2
2016-12-11	WB-21850	William Brown	11
2017-06-29	WB-21850	William Brown	4
2017-11-11	WB-21850	William Brown	1
2017-12-10	WB-21850	William Brown	2

In [66]:

```
ggplot(a,
  aes(x = reorder(order_date, cnt), y = cnt, label = cnt)) +
  geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
  labs(x = "cnt",
       y = "order_date",
       title = "Bar-Chart",
       subtitle = "Customer purchase date") +
  coord_flip()
```



?

- 날짜 순서로 정렬하려면?

In [67]:

```
# ggplot(a,
#   aes(x = order_date, y = cnt, label = cnt)) +
#   geom_bar(stat = 'identity', aes(fill = cnt), width = .5) +
#   labs(x = "cnt",
#        y = "order_date",
#        title = "Bar-Chart",
#        subtitle = "Customer purchase date") +
#   coord_flip()
```

?

Task 4

- 재구매일 계산

In [68]:

```
dataset %>%
  filter(customer_id == "WB-21850") %>%
  group_by(order_date, customer_id, customer_name) %>%
  summarise(cnt = n()) %>%
#   mutate(reorder_date = order_date - order)
  arrange(order_date)
```

order_date	customer_id	customer_name	cnt
2014-11-04	WB-21850	William Brown	5
2014-12-12	WB-21850	William Brown	1
2014-12-20	WB-21850	William Brown	2
2015-08-10	WB-21850	William Brown	3
2015-11-30	WB-21850	William Brown	3
2016-01-15	WB-21850	William Brown	3
2016-10-10	WB-21850	William Brown	2
2016-12-11	WB-21850	William Brown	11
2017-06-29	WB-21850	William Brown	4
2017-11-11	WB-21850	William Brown	1
2017-12-10	WB-21850	William Brown	2

?

Task 5

- 가장 많이 구매한 고객의 재구매 카테고리
- 재구매가 많이 일어난 카테고리

In [69]:

```
dataset %>%
  filter(customer_id == "WB-21850") %>%
  group_by(sub_category) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt))
```

sub_category	cnt
Binders	10
Phones	4
Accessories	3
Appliances	3
Furnishings	3
Paper	3
Art	2
Fasteners	2
Storage	2
Tables	2
Chairs	1
Envelopes	1
Machines	1

In [70]:

```
dataset %>%
  filter(customer_id == "WB-21850") %>%
  group_by(order_date, customer_id, customer_name, sub_category) %>%
  summarise(cnt = n()) %>%
  arrange(order_date)
```

order_date	customer_id	customer_name	sub_category	cnt
2014-11-04	WB-21850	William Brown	Binders	1
2014-11-04	WB-21850	William Brown	Machines	1
2014-11-04	WB-21850	William Brown	Paper	1
2014-11-04	WB-21850	William Brown	Phones	1
2014-11-04	WB-21850	William Brown	Tables	1
2014-12-12	WB-21850	William Brown	Furnishings	1
2014-12-20	WB-21850	William Brown	Accessories	1
2014-12-20	WB-21850	William Brown	Appliances	1
2015-08-10	WB-21850	William Brown	Appliances	1
2015-08-10	WB-21850	William Brown	Phones	2
2015-11-30	WB-21850	William Brown	Binders	1
2015-11-30	WB-21850	William Brown	Fasteners	1
2015-11-30	WB-21850	William Brown	Phones	1
2016-01-15	WB-21850	William Brown	Binders	1
2016-01-15	WB-21850	William Brown	Envelopes	1
2016-01-15	WB-21850	William Brown	Paper	1
2016-10-10	WB-21850	William Brown	Binders	1
2016-10-10	WB-21850	William Brown	Furnishings	1
2016-12-11	WB-21850	William Brown	Accessories	1
2016-12-11	WB-21850	William Brown	Appliances	1
2016-12-11	WB-21850	William Brown	Art	1
2016-12-11	WB-21850	William Brown	Binders	4
2016-12-11	WB-21850	William Brown	Chairs	1
2016-12-11	WB-21850	William Brown	Paper	1
2016-12-11	WB-21850	William Brown	Storage	1
2016-12-11	WB-21850	William Brown	Tables	1
2017-06-29	WB-21850	William Brown	Accessories	1
2017-06-29	WB-21850	William Brown	Art	1
2017-06-29	WB-21850	William Brown	Binders	1
2017-06-29	WB-21850	William Brown	Storage	1
2017-11-11	WB-21850	William Brown	Fasteners	1
2017-12-10	WB-21850	William Brown	Binders	1
2017-12-10	WB-21850	William Brown	Furnishings	1

In [71]:

```
dataset %>%
  group_by(customer_id, sub_category) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head
```

customer_id	sub_category	cnt
MA-17560	Paper	11
WB-21850	Binders	10
EA-14035	Paper	8
LB-16795	Binders	8
AB-10060	Paper	7
AP-10915	Art	7

Task 6

- 구매 건수가 많이 일어나는 카테고리

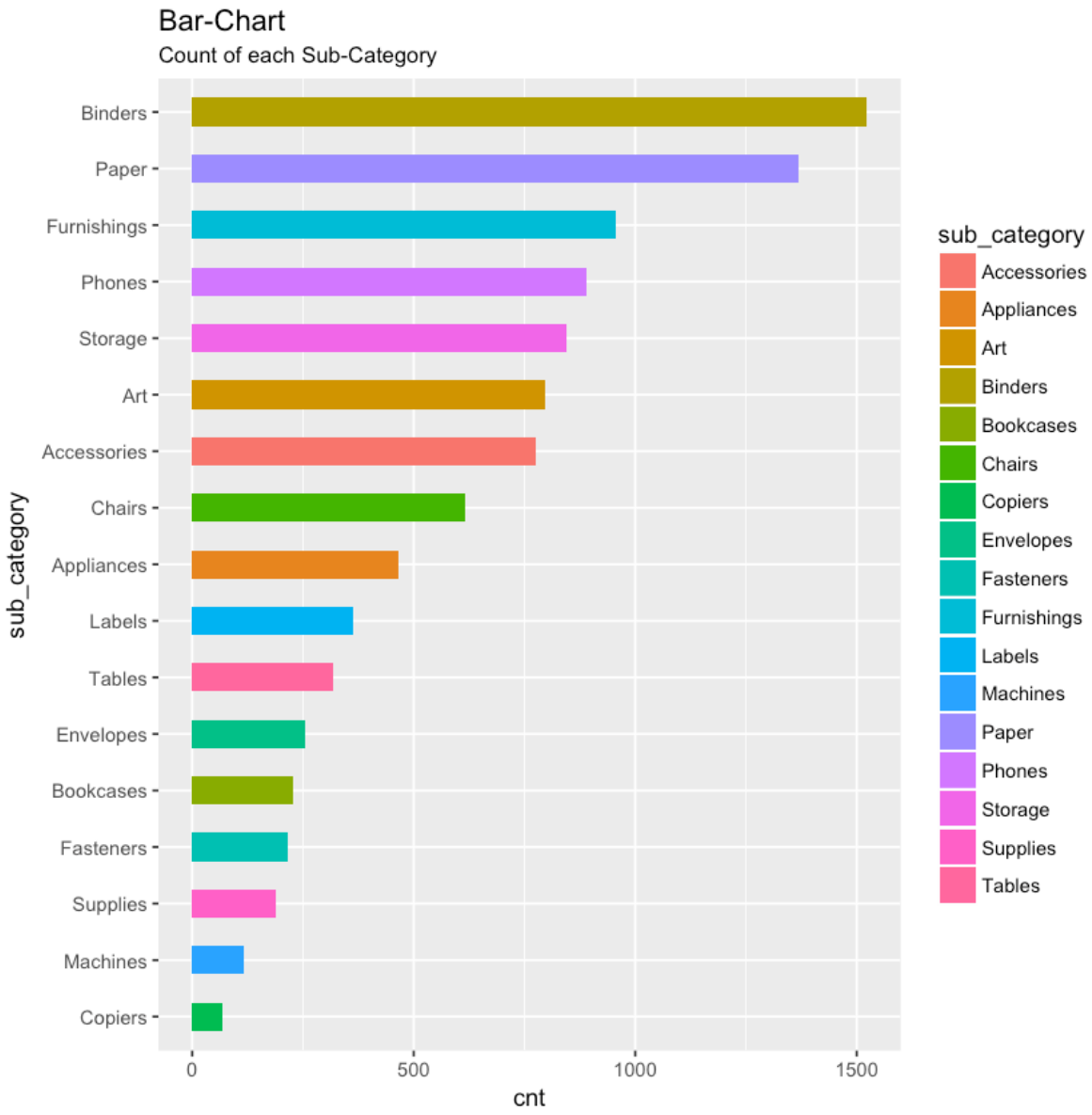
In [72]:

```
a <- dataset %>%
  group_by(sub_category) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt))
a
```

sub_category	cnt
Binders	1523
Paper	1370
Furnishings	957
Phones	889
Storage	846
Art	796
Accessories	775
Chairs	617
Appliances	466
Labels	364
Tables	319
Envelopes	254
Bookcases	228
Fasteners	217
Supplies	190
Machines	115
Copiers	68

In [73]:

```
a <- ggplot(a,
    aes(x = reorder(sub_category, cnt), y = cnt, label = cnt)) +
    geom_bar(stat = 'identity', aes(fill = sub_category), width = .5) +
    labs(x = "sub_category",
         title = "Bar-Chart",
         subtitle = "Count of each Sub-Category") +
    coord_flip()

a
```



5-4. 가장 판매가 부진한 상품은 무엇이고 이유는 무엇인가?

Task 1

- 가장 판매가 부진한 상품

In [84]:

```
a <- dataset %>%
  group_by(product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt))
a %>% head()
```

product_name	cnt
Staple envelope	48
Easy-staple paper	46
Staples	46
Avery Non-Stick Binders	20
Staples in misc. colors	19
KI Adjustable-Height Table	18

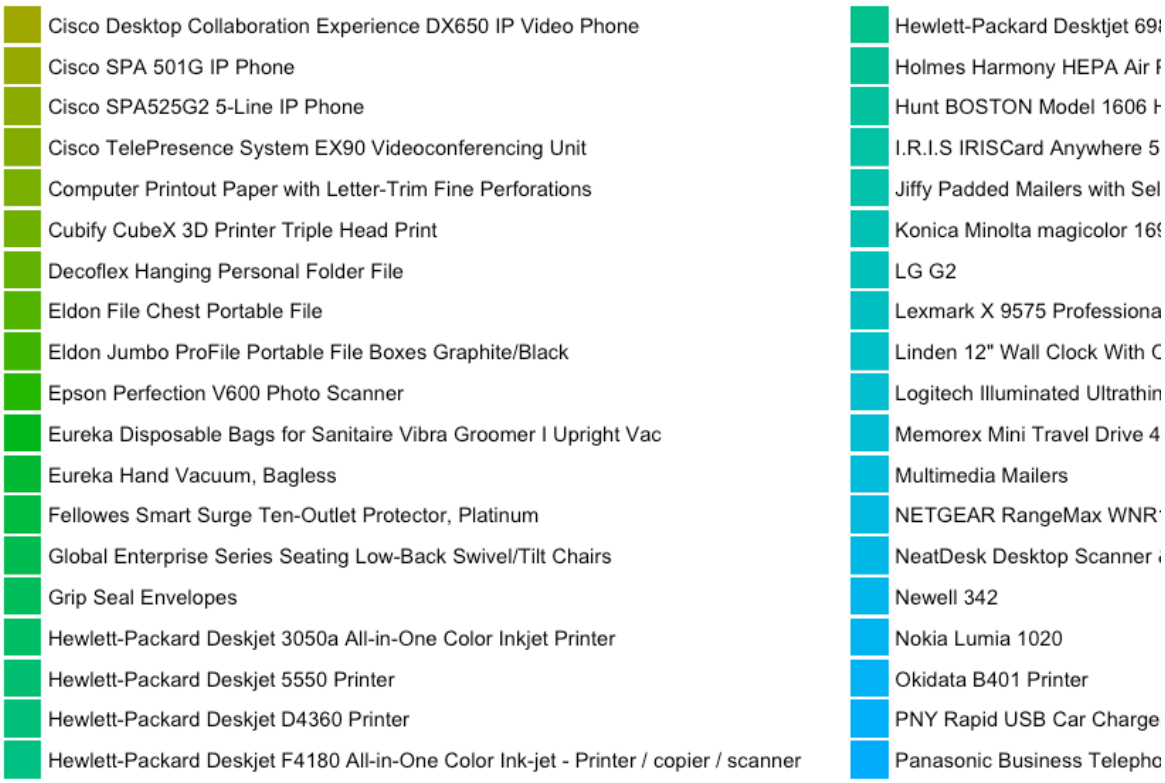
In [85]:

```
bad_score_product <- a %>% filter(cnt == 1)
bad_score_product %>% head()
```

product_name	cnt
4009 Highlighters	1
AT&T EL51110 DECT	1
Acco Glide Clips	1
Avaya IP Phone 1140E VoIP phone	1
Avery 484	1
Avery 5	1

In [83]:

```
a <- ggplot(bad_score_product,
    aes(x = reorder(product_name, cnt), y = cnt, label = cnt)) +
    geom_bar(stat = 'identity', aes(fill = product_name), width = .5) +
    labs(x = "sub_category",
         title = "Bar-Chart",
         subtitle = "bad_score_product") +
    coord_flip()

a
```



Task 2

- 이유는?
- 가장 판매가 부진한 상품이 있는 sub_category를 확인

In [87]:

```
a <- dataset %>%
  group_by(sub_category, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt))
a %>% tail()
```

sub_category	product_name	cnt
Phones	RCA ViSYS 25425RE1 Corded phone	1
Phones	Xiaomi Mi3	1
Storage	Decoflex Hanging Personal Folder File	1
Storage	Eldon File Chest Portable File	1
Storage	Eldon Jumbo ProFile Portable File Boxes Graphite/Black	1
Tables	Barricks Non-Folding Utility Table with Steel Legs, Laminate Tops	1

In [91]:

```
a <- a %>% filter(cnt == 1)
a
```

sub_category	product_name	cnt
Chairs	Global Enterprise Series Seating Low-Back Swivel/Tilt Chairs	1
Envelopes	Grip Seal Envelopes	1
Envelopes	Jiffy Padded Mailers with Self-Seal Closure	1
Envelopes	Multimedia Mailers	1
Envelopes	Park Ridge Embossed Executive Business Envelopes	1
Fasteners	Acco Glide Clips	1
Furnishings	Linden 12" Wall Clock With Oak Frame	1
Furnishings	Ultra Commercial Grade Dual Valve Door Closer	1
⋮	⋮	⋮
Paper	Rediform S.O.S. Phone Message Books	1

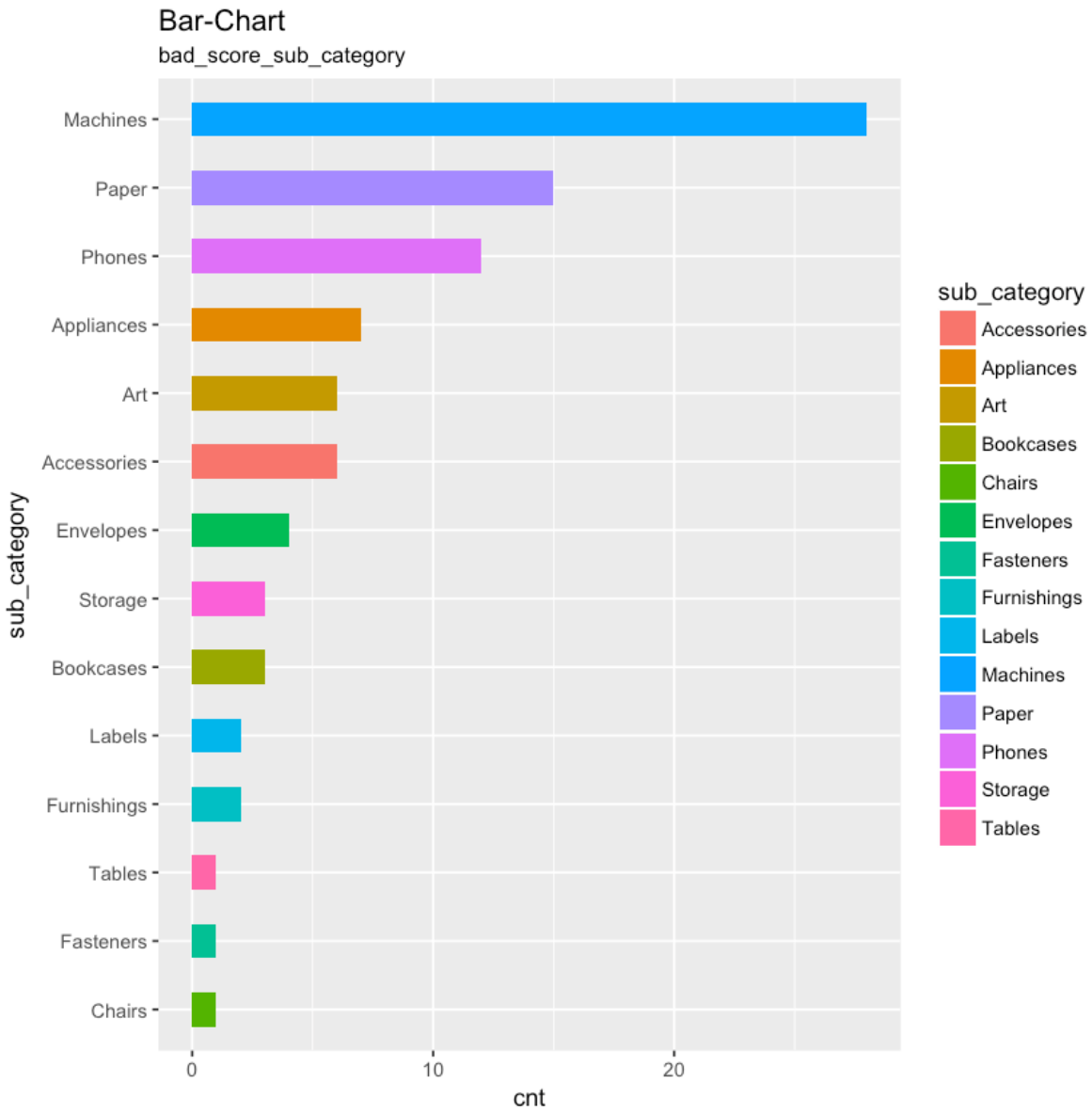
In [97]:

```
b <- a %>% group_by(sub_category) %>% summarise(cnt = n()) %>% arrange(desc(cnt))
b
```

sub_category	cnt
Machines	28
Paper	15
Phones	12
Appliances	7
Accessories	6
Art	6
Envelopes	4
Bookcases	3
Storage	3
Furnishings	2
Labels	2
Chairs	1
Fasteners	1
Tables	1

In [98]:

```
a <- ggplot(b,
    aes(x = reorder(sub_category, cnt), y = cnt, label = cnt)) +
    geom_bar(stat = 'identity', aes(fill = sub_category), width = .5) +
    labs(x = "sub_category",
         title = "Bar-Chart",
         subtitle = "bad_score_sub_category") +
    coord_flip()

a
```



이유는 Machne 같은 경우는 기술이 빠르게 변화하니까?

paper는?

In [ ]:

```

```

5-5. 많이 팔리는 상품의 가격수준과 일정가격 이하 상품의 판매량은

어떠한가?

Task 1

- 많이 팔리는 상품의 가격 수준

전체 기준으로 많이 팔린 상품

In [107]:

```
dataset %>%
  group_by(product_id, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head()
```

product_id	product_name	cnt
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15
TEC-AC-10003628	Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	15
FUR-CH-10002880	Global High-Back Leather Tilter, Burgundy	14
FUR-CH-10003774	Global Wood Trimmed Manager's Task Chair, Khaki	14
OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal Lined Pattern	14
FUR-CH-10004287	SAFCO Arco Folding Chair	13

새로운 컬럼을 생성한 dataframe 생성

- per_price, per_profit

In [135]:

```
dataset2 <- dataset %>%
  mutate(per_price = sales / quantity, per_profit = profit / quantity)

head(dataset2)
```

row_id	order_id	order_date	ship_date	ship_mode	customer_id	customer_name	segment	country	city
1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles
4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale

method1

- per_price가 group_by에 들어가면서 결과같이 바뀜

In [138]:

```
dataset2 %>%
  group_by(product_id, product_name, per_price) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head()
```

product_id	product_name	per_price	cnt
OFF-PA-10001970	Xerox 1881	12.280	10
FUR-FU-10000010	DAX Value U-Channel Document Frames, Easel Back	4.970	9
OFF-AR-10004078	Newell 312	4.672	9
OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal Lined Pattern	16.784	9
OFF-PA-10000157	Xerox 191	19.980	9
OFF-ST-10003123	Fellowes Bases and Tops For Staxonsteel/High-Stak Systems	33.290	9

?

- 위와 아래가 다름... 어떤 기준으로 접근해야 좋을까?

method 2

- 전체 기준으로 많이 팔린 것들

In [139]:

```
dataset2 %>%
  group_by(product_id, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head()
```

product_id	product_name	cnt
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15
TEC-AC-10003628	Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	15
FUR-CH-10002880	Global High-Back Leather Tilter, Burgundy	14
FUR-CH-10003774	Global Wood Trimmed Manager's Task Chair, Khaki	14
OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal Lined Pattern	14
FUR-CH-10004287	SAFCO Arco Folding Chair	13

?

Solution ... inner_join?

- method2의 product_id와 dataset2의 product_id를 inner_join해서 per_price 가져오기?

In [144]:

```
a <- dataset2 %>%
  group_by(product_id, product_name) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head(10)

dim(a)
a
```

10 3

product_id	product_name	cnt
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15
TEC-AC-10003628	Logitech 910-002974 M325 Wireless Mouse for Web Scrolling	15
FUR-CH-10002880	Global High-Back Leather Tilter, Burgundy	14
FUR-CH-10003774	Global Wood Trimmed Manager's Task Chair, Khaki	14
OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal Lined Pattern	14
FUR-CH-10004287	SAFCO Arco Folding Chair	13
FUR-TA-10001095	Chromcraft Round Conference Tables	13
OFF-BI-10000145	Zipper Ring Binder Pockets	13
OFF-BI-10000301	GBC Instant Report Kit	13
OFF-BI-10000977	Ibico Plastic Spiral Binding Combs	13

In [150]:

```
colnames(dataset2)
```

'row_id' 'order_id' 'order_date' 'ship_date' 'ship_mode' 'customer_id'
'customer_name' 'segment' 'country' 'city' 'state' 'postal_code' 'region'
'product_id' 'category' 'sub_category' 'product_name' 'sales' 'quantity' 'discount'
'profit' 'lead_time' 'per_price' 'per_profit'

In [152]:

```
dataset2 %>% group_by(product_id, product_name, per_price, per_profit) %>% summarise
```

product_id	product_name	per_price	per_profit	cnt
OFF-PA-10001970	Xerox 1881	12.280	5.7716	10
OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal Lined Pattern	16.784	5.8744	9
FUR-FU-10002364	Eldon Expressions Wood Desk Accessories, Oak	7.380	2.1402	8
OFF-AP-10001492	Acco Six-Outlet Power Strip, 4' Cord Length	8.620	2.2412	8
OFF-LA-10002762	Avery 485	12.530	5.8891	8
OFF-PA-10000157	Xerox 191	19.980	9.3906	8
OFF-PA-10001838	Adams Telephone Message Book W/Dividers/Space For Phone Numbers, 5 1/4"X8 1/2", 300/Messages	5.880	2.8812	8
OFF-PA-	Adams Phone Message Book, 200 Message Capacity, 8 1/16”	6.880	3.1648	8

In [ ]:

```

```

In [158]:

```
dataset2 %>% group_by(product_id, product_name, per_price, per_profit) %>% summarise
```

product_id	product_name	per_price	per_profit	mean(per_price)	cnt
FUR-BO-10000112	Bush Birmingham Collection Bookcase, Dark Cherry	91.6860	-13.0980	91.6860	1
FUR-BO-10000330	Sauder Camden County Barrister Bookcase, Planked Cherry Finish	102.8330	-1.2098	102.8330	2
FUR-BO-10000330	Sauder Camden County Barrister Bookcase, Planked Cherry Finish	120.9800	16.9372	120.9800	1
FUR-BO-10000362	Sauder Inglewood Library Bookcases	119.6860	-11.9686	119.6860	1
FUR-BO-10000362	Sauder Inglewood Library Bookcases	136.7840	5.1294	136.7840	1
FUR-BO-10000362	Sauder Inglewood Library Bookcases	145.3330	13.6784	145.3330	1
FUR-BO-10000362	Sauder Inglewood Library Bookcases	170.9800	39.3254	170.9800	2
FUR-BO-	O'Sullivan 2-Shelf Heavy-Duty Bookcases	14.5740	-26.2332	14.5740	1

In [ ]:

```

```

In [ ]:

```

```

In [149]:

```
a %>% left_join(dataset2, by = "product_id")
```

product_id	product_name.x	cnt	row_id	order_id	order_date	ship_date	ship_mode	customer_id	customer_na
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15	540	CA-2015-134894	2015-12-07	2015-12-11	Standard Class	DK-12985	Darren Kout
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15	800	CA-2015-101910	2015-11-27	2015-12-03	Standard Class	CD-11920	Carlos D
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15	2957	CA-2017-123638	2017-06-27	2017-07-04	Standard Class	MA-17995	Michelle Arr
FUR-CH-10002647	Situations Contoured Folding Chairs, 4/Set	15	3145	US-2016-148110	2016-09-05	2016-09-11	Standard Class	AR-10825	Anthony Raw
FUR-CH-	Situations Contoured			CA-		2016-09-			

In [ ]:

```

```

In [ ]:

```

```

In [112]:

```
b <- a %>%
  filter(cnt >= 4)

dim(b)
b
```

OFF-AR-10000588	Newell 345	59.520	3	4	19.840	
OFF-AR-10001615	Newell 34	59.520	3	4	19.840	
OFF-AR-10002067	Newell 334	99.200	5	4	19.840	
OFF-AR-10003183	Avery Fluorescent Highlighter Four-Color Set	8.016	3	4	2.672	
OFF-AR-10003582	Boston Electric Pencil Sharpener, Model 1818, Charcoal Black	56.300	2	4	28.150	
OFF-BI-10002103	Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl	13.904	2	4	6.952	
OFF-BI-10004318	Ibico EB-19 Dual Function Manual Binding System	276.784	2	4	138.392	
OFF-BI-10004364	Storex Dura Pro Binders	3.564	3	4	1.188	
OFF-EN						

In [114]:

```
mean(b$price)
```

44.7389268292683

In [ ]:

```

```

?

많이 팔리는 상품의 가격 수준 시각화 (Box-Plot)

- 위에 profit에 비교했을 때, Storage가 부정적으로 나왔는데... 잘못 해석될 가능성이 농후한 분석

In [115]:

```
a <- dataset %>%
  group_by(sub_category, product_id, product_name, sales, quantity) %>%
  summarise(cnt = n()) %>%
  mutate(price = sales/quantity) %>%
  arrange(desc(cnt))
a %>% head()
```

sub_category	product_id	product_name	sales	quantity	cnt	price
Paper	OFF-PA-10001838	Adams Telephone Message Book W/Dividers/Space For Phone Numbers, 5 1/4"X8 1/2", 300/Messages	11.76	2	5	5.88
Accessories	TEC-AC-10004510	Logitech Desktop MK120 Mouse and keyboard Combo	98.16	6	4	16.36
Appliances	OFF-AP-10004532	Kensington 6 Outlet Guardian Standard Surge Protector	61.44	3	4	20.48
Art	OFF-AR-10000588	Newell 345	59.52	3	4	19.84
Art	OFF-AR-10001615	Newell 34	59.52	3	4	19.84
Art	OFF-AR-10002067	Newell 334	99.20	5	4	19.84

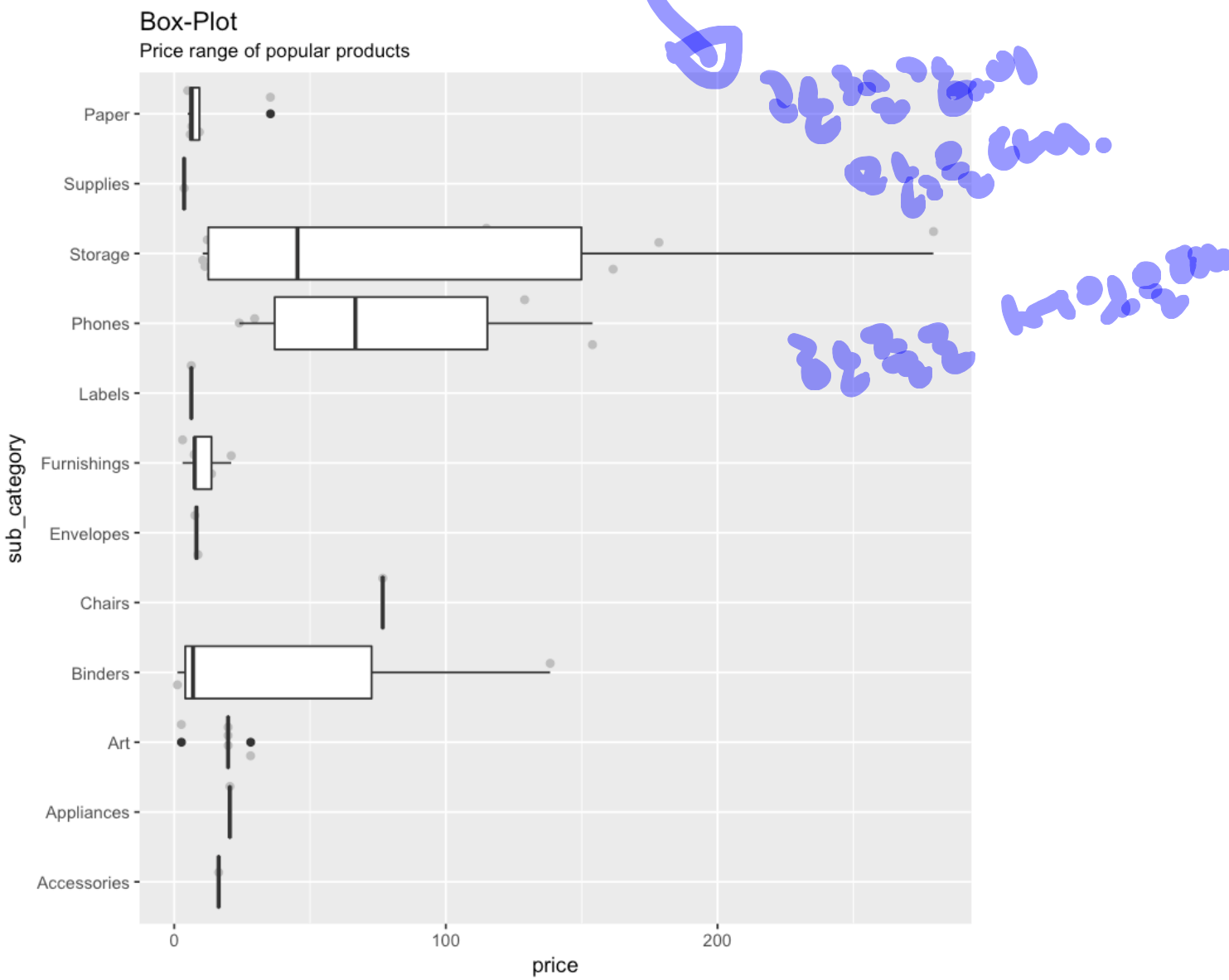
In [116]:

```
b <- a %>%
  filter(cnt >= 4)

dim(b)
b
```

41 7

In [126]:

```
ggplot(b, aes(x = reorder(sub_category, cnt), y = price)) +
  geom_jitter(col='grey') +
  geom_boxplot() +
  labs(x = "sub_category",
         title = "Box-Plot",
         subtitle = "Price range of popular products") +
  coord_flip()
```



Task 2

- 일정가격 이하 상품의 판매량은 어떠한가?

  ?
- 어떻게 접근해야할까?
- 그룹을 나눠서 (계급구간) count를 해야할까?

In [128]:

```
colnames(dataset)
```

'row_id' 'order_id' 'order_date' 'ship_date' 'ship_mode' 'customer_id'
'customer_name' 'segment' 'country' 'city' 'state' 'postal_code' 'region'
'product_id' 'category' 'sub_category' 'product_name' 'sales' 'quantity' 'discount'
'profit' 'lead_time'

In [133]:

```
head(dataset)
```

tomer_id	customer_name	segment	country	city	...	region	product_id	category	sub_c
CG-12520	Claire Gute	Consumer	United States	Henderson	...	South	FUR-BO-10001798	Furniture	Bo
CG-12520	Claire Gute	Consumer	United States	Henderson	...	South	FUR-CH-10000454	Furniture	
DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	West	OFF-LA-10000240	Office Supplies	
SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	South	FUR-TA-10000577	Furniture	
SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	South	OFF-ST-10000760	Office Supplies	
BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	...	West	FUR-FU-10001487	Furniture	Fur

In [134]:

```
dataset2 <- dataset %>%
  mutate(per_price = sales / quantity, per_profit = profit / quantity)

head(dataset2)
```

tomer_name	segment	country	city	...	category	sub_category	product_name	sales
Claire Gute	Consumer	United States	Henderson	...	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600
Claire Gute	Consumer	United States	Henderson	...	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.9400
arrin Van Huff	Corporate	United States	Los Angeles	...	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.6200
an O'Donnell	Consumer	United States	Fort Lauderdale	...	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775
an O'Donnell	Consumer	United States	Fort Lauderdale	...	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.3680
sina Hoffman	Consumer	United States	Los Angeles	...	Furniture	Furnishings	Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood	48.8600

In [131]:

```
dataset %>% group_by(product_name, sales) %>% summarise(cnt = n()) %>% arrange(desc(
```

Eldon Expressions Wood Desk Accessories, Oak	14.760	6
Adams Telephone Message Book W/Dividers/Space For Phone Numbers, 5 1/4"X8 1/2", 300/Messages	11.760	5
#10- 4 1/8" x 9 1/2" Recycled Envelopes	17.480	4
Acme Value Line Scissors	7.300	4
Advantus Rolling Storage Box	51.450	4
Avaya 4621SW VoIP phone	177.480	4
Avery 51	18.900	4
Avery Fluorescent Highlighter Four-Color Set	8.016	4
Bevis Steel Folding Chairs	230.280	4
Boston Electric Pencil Sharpener, Model 1818, Charcoal Black	56.300	4
Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl	13.904	4
DAX Wood Document Frame	27.460	4

In [132]:

```
dataset %>% group_by(product_name, sales, quantity) %>% summarise(cnt = n()) %>% ar
```

product_name	sales	quantity	cnt
Adams Telephone Message Book W/Dividers/Space For Phone Numbers, 5 1/4"X8 1/2", 300/Messages	11.760	2	5
#10- 4 1/8" x 9 1/2" Recycled Envelopes	17.480	2	4
Acme Value Line Scissors	7.300	2	4
Advantus Rolling Storage Box	51.450	3	4
Avaya 4621SW VoIP phone	177.480	3	4
Avery 51	18.900	3	4
Avery Fluorescent Highlighter Four-Color Set	8.016	3	4
Bevis Steel Folding Chairs	230.280	3	4
Boston Electric Pencil Sharpener, Model 1818, Charcoal Black	56.300	2	4
Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl	13.904	2	4
DAX Wood Document Frame	27.460	2	4

5-6. Discount가 많을 수록 매출이 늘어나는가?

In [ ]:

```

```

5-7. 지역별로 가장 많이 팔리는 상품은 무엇인가?

In [ ]:

```

```

5-8. Order Date를 기반으로 동시구매가 많이 일어나는 상품은 무엇인가?

In [ ]:

```

```

5-9. 특정상품의 판매시기와 지역별 수요를 파악해보자

Task 1

- 특정 상품의 판매시기
- 어떻게 접근을 해야할까?

In [173]:

```
dataset %>% group_by(sub_category) %>% summarise(n_distinct=n()) %>% arrange(desc(n_
```

sub_category	n_distinct
Binders	1523
Paper	1370
Furnishings	957
Phones	889
Storage	846
Art	796
Accessories	775
Chairs	617
Appliances	466
Labels	364
Tables	319
Envelopes	254
Bookcases	228
Fasteners	217
Supplies	190
Machines	115
Copiers	68

Machines가 가장 많이 팔리는 시기는?

In [182]:

```
# dataset %>%
#   filter(sub_category == 'Machines') %>%
#   group_by(sub_category, order_date) %>%
#   summarise(cnt = n()) %>%
#   arrange(desc(cnt)) %>%
#   head(10)
```

In [180]:

```
dataset %>%
  filter(sub_category == 'Machines') %>%
  mutate(year_month = substr(order_date, 1, 7)) %>%
  select(year_month, sub_category) %>%
  group_by(sub_category, year_month) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  head(10)
```

sub_category	year_month	cnt
Machines	2014-09	12
Machines	2015-11	6
Machines	2016-11	6
Machines	2017-11	6
Machines	2015-12	5
Machines	2016-05	5
Machines	2016-06	5
Machines	2017-09	5
Machines	2014-03	4
Machines	2016-03	4

In [ ]:

```
```

Task 2

- 지역별 수요

In [161]:

```
dataset %>%
  group_by(city) %>%
  summarise(cnt = n()) %>%
  arrange(desc(cnt)) %>%
  filter(cnt >= 100)
```

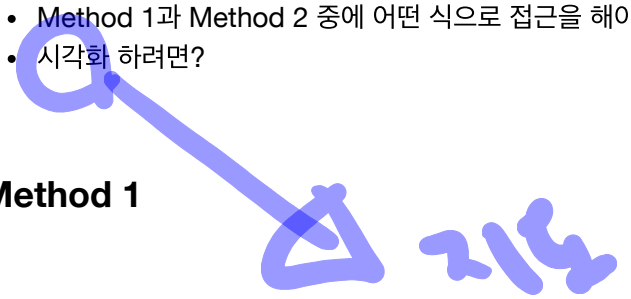
city	cnt
New York City	915
Los Angeles	747
Philadelphia	537
San Francisco	510
Seattle	428
Houston	377
Chicago	314
Columbus	222
San Diego	170
Springfield	163
Dallas	157
Jacksonville	125
Detroit	115

Task 3

- 지역별 특정 제품 수요
- seasonality와 연관지어도 좋을듯?

?

- Method 1과 Method 2 중에 어떤 식으로 접근을 해야할까?
- 시각화 하려면?



Method 1

In [165]:

```
dataset %>%
  filter(city == "New York City") %>%
  group_by(sub_category) %>%
  summarise(cnt = n ()) %>%
  arrange(desc(cnt))
```

sub_category	cnt
Binders	145
Paper	124
Phones	89
Storage	82
Furnishings	78
Art	70
Accessories	64
Chairs	62
Appliances	36
Labels	36
Bookcases	29
Envelopes	23
Tables	23
Supplies	19
Fasteners	17
Machines	12
Copiers	6

Method 2

In [168]:

```
dataset %>%
  filter(city == "New York City") %>%
  group_by(sub_category, product_name) %>%
  summarise(cnt = n ()) %>%
  arrange(desc(cnt)) %>%
head(10)
```

sub_category	product_name	cnt
Envelopes	Staple envelope	7
Paper	Easy-staple paper	7
Binders	Acco Pressboard Covers with Storage Hooks, 9 1/2" x 11", Executive Red	4
Binders	GBC Premium Transparent Covers with Diagonal Lined Pattern	4
Accessories	Enermax Aurora Lite Keyboard	3
Accessories	Logitech LS21 Speaker System - PC Multimedia - 2.1-CH - Wired	3
Accessories	WD My Passport Ultra 2TB Portable External Hard Drive	3
Art	Dixon Prang Watercolor Pencils, 10-Color Set with Brush	3
Art	Sanford Pocket Accent Highlighters	3
Binders	3M Organizer Strips	3

In [ ]: