Week 2.1

Read Data

```
# Read data from xlsx
library(readxl)
superstore<-read_xlsx("Superstore.xlsx")
```

```
head(superstore)
```

```
## # A tibble: 6 x 19
##   Category City  `Country/Region` `Customer Name` Manufacturer
##   <chr>    <chr> <chr>            <chr>           <chr>
## 1 Furnitu~ Hend~ United States    Claire Gute     Bush
## 2 Furnitu~ Hend~ United States    Claire Gute     Hon
## 3 Office ~ Los ~ United States    Darrin Van Huff Universal
## 4 Furnitu~ Fort~ United States    Sean O'Donnell  Bretford
## 5 Office ~ Fort~ United States    Sean O'Donnell  Eldon
## 6 Furnitu~ Los ~ United States    Brosina Hoffman Eldon
## # ... with 14 more variables: `Order Date` <dttm>, `Order ID` <chr>, `Postal
## #   Code` <dbl>, `Product Name` <chr>, Region <chr>, Segment <chr>, `Ship
## #   Date` <dttm>, `Ship Mode` <chr>, State <chr>, `Sub-Category` <chr>,
## #   Discount <dbl>, Profit <dbl>, Quantity <dbl>, Sales <dbl>
```

Explore Data

```
summary(superstore)
```

```
##    Category             City           Country/Region     Customer Name
##  Length:9994        Length:9994        Length:9994        Length:9994
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Manufacturer         Order Date                     Order ID
##  Length:9994        Min.   :2017-01-03 00:00:00   Length:9994
##  Class :character   1st Qu.:2018-05-23 00:00:00   Class :character
##  Mode  :character   Median :2019-06-26 00:00:00   Mode  :character
##                     Mean   :2019-04-30 17:41:20
##                     3rd Qu.:2020-05-14 00:00:00
##                     Max.   :2020-12-30 00:00:00
```

```
##
##   Postal Code    Product Name          Region            Segment
##  Min.   : 1040   Length:9994        Length:9994        Length:9994
##  1st Qu.:23223   Class :character   Class :character   Class :character
##  Median :57103   Mode  :character   Mode  :character   Mode  :character
##  Mean   :55245
##  3rd Qu.:90008
##  Max.   :99301
##  NA's   :11
##    Ship Date                    Ship Mode            State
##  Min.   :2017-01-07 00:00:00   Length:9994        Length:9994
##  1st Qu.:2018-05-27 00:00:00   Class :character   Class :character
##  Median :2019-06-29 00:00:00   Mode  :character   Mode  :character
##  Mean   :2019-05-04 16:42:15
##  3rd Qu.:2020-05-18 00:00:00
##  Max.   :2021-01-05 00:00:00
##
##  Sub-Category          Discount          Profit             Quantity
##  Length:9994        Min.   :0.0000   Min.   :-6599.978   Min.   : 1.00
##  Class :character   1st Qu.:0.0000   1st Qu.:    1.729   1st Qu.: 2.00
##  Mode  :character   Median :0.2000   Median :    8.666   Median : 3.00
##                     Mean   :0.1562   Mean   :   28.657   Mean   : 3.79
##                     3rd Qu.:0.2000   3rd Qu.:   29.364   3rd Qu.: 5.00
##                     Max.   :0.8000   Max.   : 8399.976   Max.   :14.00
##
##      Sales
##  Min.   :    0.444
##  1st Qu.:   17.280
##  Median :   54.490
##  Mean   :  229.858
##  3rd Qu.:  209.940
##  Max.   :22638.480
##
```

Check data type in the environment tab

Column Operations

```
# Checking for unique values and number of unique values
unique_category<-unique(superstore$Category)
unique_category
```

```
## [1] "Furniture"       "Office Supplies" "Technology"
```

```
length(unique_category)
```

```
## [1] 3
```

```
# Column selection using index values
sample_data<-superstore[,c(1,17:19)]
sample_data
```

```
## # A tibble: 9,994 x 4
##    Category         Profit Quantity  Sales
##    <chr>             <dbl>    <dbl>  <dbl>
##  1 Furniture         41.9         2 262.
##  2 Furniture        220.          3 732.
##  3 Office Supplies    6.87        2  14.6
##  4 Furniture       -383.          5 958.
##  5 Office Supplies    2.52        2  22.4
##  6 Furniture         14.2         7  48.9
##  7 Office Supplies    1.97        4   7.28
##  8 Technology        90.7         6 907.
##  9 Office Supplies    5.78        3  18.5
## 10 Office Supplies   34.5         5 115.
## # ... with 9,984 more rows
```

```
# Creating a new column
sample_data$Avg_sales_per_unit<-sample_data$Sales/sample_data$Quantity
sample_data
```

```
## # A tibble: 9,994 x 5
##    Category         Profit Quantity  Sales Avg_sales_per_unit
##    <chr>             <dbl>    <dbl>  <dbl>              <dbl>
##  1 Furniture         41.9         2 262.               131.
##  2 Furniture        220.          3 732.               244.
##  3 Office Supplies    6.87        2  14.6                7.31
##  4 Furniture       -383.          5 958.               192.
##  5 Office Supplies    2.52        2  22.4                11.2
##  6 Furniture         14.2         7  48.9                6.98
##  7 Office Supplies    1.97        4   7.28               1.82
##  8 Technology        90.7         6 907.               151.
##  9 Office Supplies    5.78        3  18.5                6.17
## 10 Office Supplies   34.5         5 115.               23.0
## # ... with 9,984 more rows
```

```
# Subset rows where Profit >0
profit_data<-sample_data[sample_data$Profit>0,]
profit_data
```

```
## # A tibble: 8,058 x 5
##    Category        Profit Quantity   Sales Avg_sales_per_unit
##    <chr>            <dbl>    <dbl>   <dbl>              <dbl>
##  1 Furniture        41.9         2  262.               131.
##  2 Furniture       220.          3  732.               244.
##  3 Office Supplies   6.87        2   14.6                7.31
##  4 Office Supplies   2.52        2   22.4               11.2
##  5 Furniture        14.2         7   48.9                6.98
##  6 Office Supplies   1.97        4    7.28               1.82
##  7 Technology       90.7         6  907.               151.
##  8 Office Supplies   5.78        3   18.5                6.17
##  9 Office Supplies  34.5         5  115.                23.0
## 10 Furniture        85.3         9 1706.               190.
## # ... with 8,048 more rows
```

```r
# Subset rows where Profit >0 and Category=Furniture
profit_data<-sample_data[sample_data$Profit>0 & sample_data$Category=="Furniture",]
profit_data
```

```
## # A tibble: 1,374 x 5
##    Category  Profit Quantity   Sales Avg_sales_per_unit
##    <chr>      <dbl>    <dbl>   <dbl>              <dbl>
##  1 Furniture  41.9         2  262.               131.
##  2 Furniture 220.          3  732.               244.
##  3 Furniture  14.2         7   48.9                6.98
##  4 Furniture  85.3         9 1706.               190.
##  5 Furniture 240.          3 1045.               348.
##  6 Furniture  15.5         3  124.                41.4
##  7 Furniture   2.96        2    6.16               3.08
##  8 Furniture  17.1         1   90.0               90.0
##  9 Furniture   7.10        5  319.                63.9
## 10 Furniture  22.3         4   79.8               19.9
## # ... with 1,364 more rows
```

```r
rm(profit_data, sample_data)
```

```r
# Common issue of partial string match
superstore[superstore$`Product Name`=="Xerox",]
```

```
## # A tibble: 0 x 19
## # ... with 19 variables: Category <chr>, City <chr>, `Country/Region` <chr>,
## #   `Customer Name` <chr>, Manufacturer <chr>, `Order Date` <dttm>, `Order
## #   ID` <chr>, `Postal Code` <dbl>, `Product Name` <chr>, Region <chr>,
## #   Segment <chr>, `Ship Date` <dttm>, `Ship Mode` <chr>, State <chr>,
## #   `Sub-Category` <chr>, Discount <dbl>, Profit <dbl>, Quantity <dbl>,
## #   Sales <dbl>
```

```r
# We take help of packages
library(stringr)
row_index<-str_which(superstore$`Product Name`, "Xerox")
sample_data<-superstore[row_index,]
sample_data
```

```
## # A tibble: 865 x 19
##    Category City  `Country/Region` `Customer Name` Manufacturer
##    <chr>    <chr> <chr>            <chr>           <chr>
##  1 Office ~ Conc~ United States    Andrew Allen    Xerox
##  2 Office ~ Troy  United States    Ted Butterfield Xerox
##  3 Office ~ Los ~ United States    Kunst Miller    Xerox
##  4 Office ~ Los ~ United States    Jim Sink        Xerox
##  5 Office ~ Minn~ United States    Karl Braun      Xerox
##  6 Office ~ Colu~ United States    Ryan Crowe      Xerox
##  7 Office ~ Colu~ United States    Dorothy Wardle  Xerox
##  8 Office ~ Rose~ United States    Lena Creighton  Xerox
##  9 Office ~ Rose~ United States    Lena Creighton  Xerox
## 10 Office ~ San ~ United States    Sally Hughsby   Xerox
```

```
## # ... with 855 more rows, and 14 more variables: `Order Date` <dttm>, `Order
## #   ID` <chr>, `Postal Code` <dbl>, `Product Name` <chr>, Region <chr>,
## #   Segment <chr>, `Ship Date` <dttm>, `Ship Mode` <chr>, State <chr>,
## #   `Sub-Category` <chr>, Discount <dbl>, Profit <dbl>, Quantity <dbl>,
## #   Sales <dbl>
```

```
# Change case
superstore$`Country/Region`<-str_to_upper(superstore$`Country/Region`)
```

```
# Split Customer name into first name, last name
name_data<-as.data.frame(str_split(superstore$`Customer Name`, " ", simplify = T))
superstore$Firstname<-name_data$V1
superstore$Lastname<-name_data$V2
```

```
# Manipulate date
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
##
## Attaching package: 'lubridate'
```

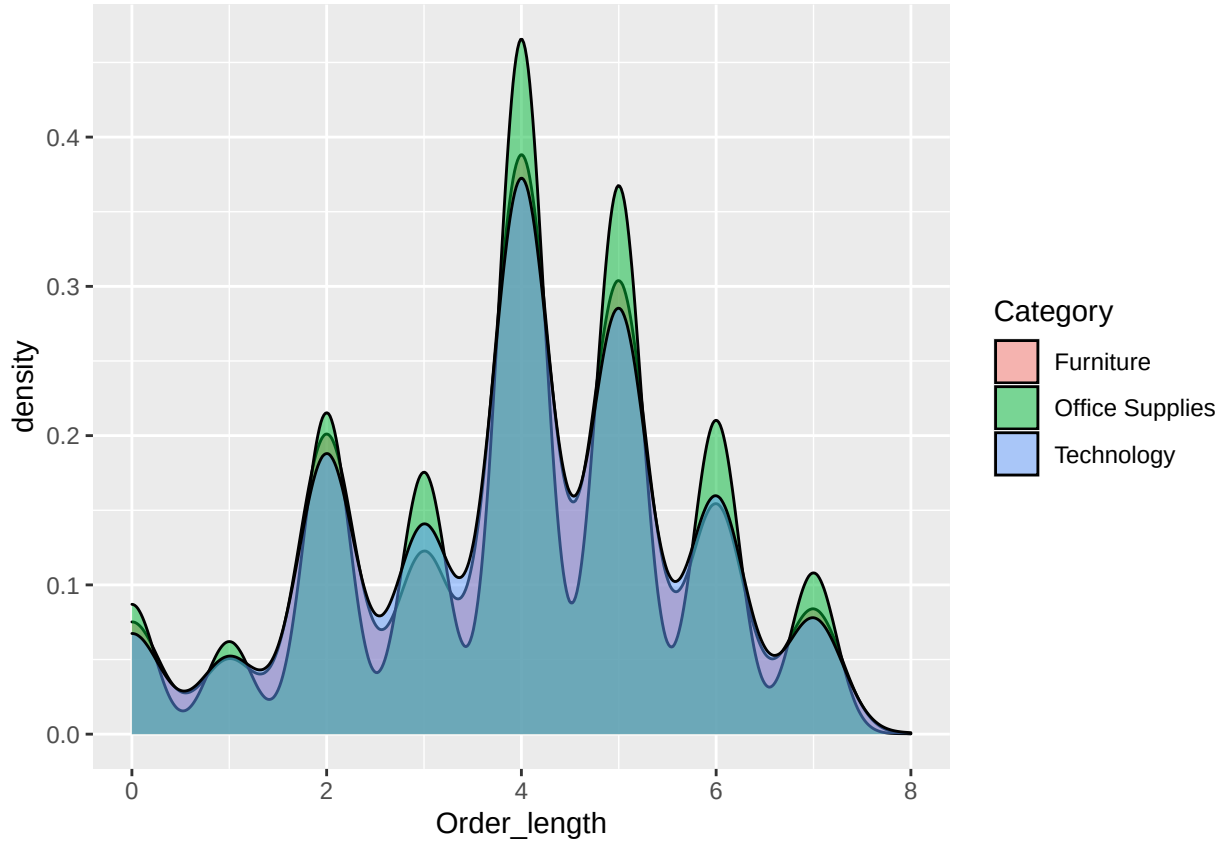
```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
superstore$`Order Date`<-as_date(superstore$`Order Date`)
superstore$`Ship Date`<-as_date(superstore$`Ship Date`)
```

```
# Find difference between order date and ship date
superstore$Order_length<-as.numeric(superstore$`Ship Date`-superstore$`Order Date`)
```

```
# Check distribution of order length
hist(superstore$Order_length)
```

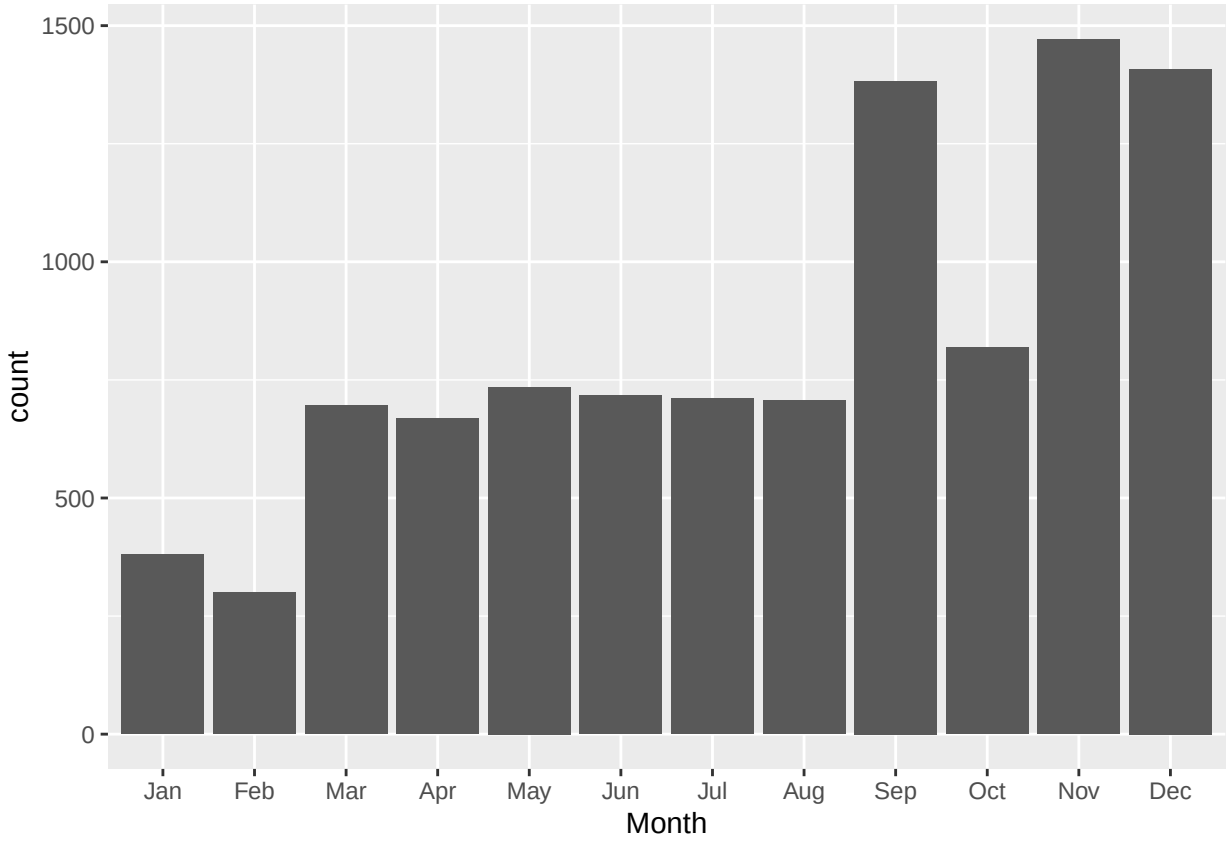

```
# Does the order length distribution vary by category?
library(ggplot2)
ggplot(superstore, aes(Order_length, fill=Category))+geom_density(alpha=0.5)
```



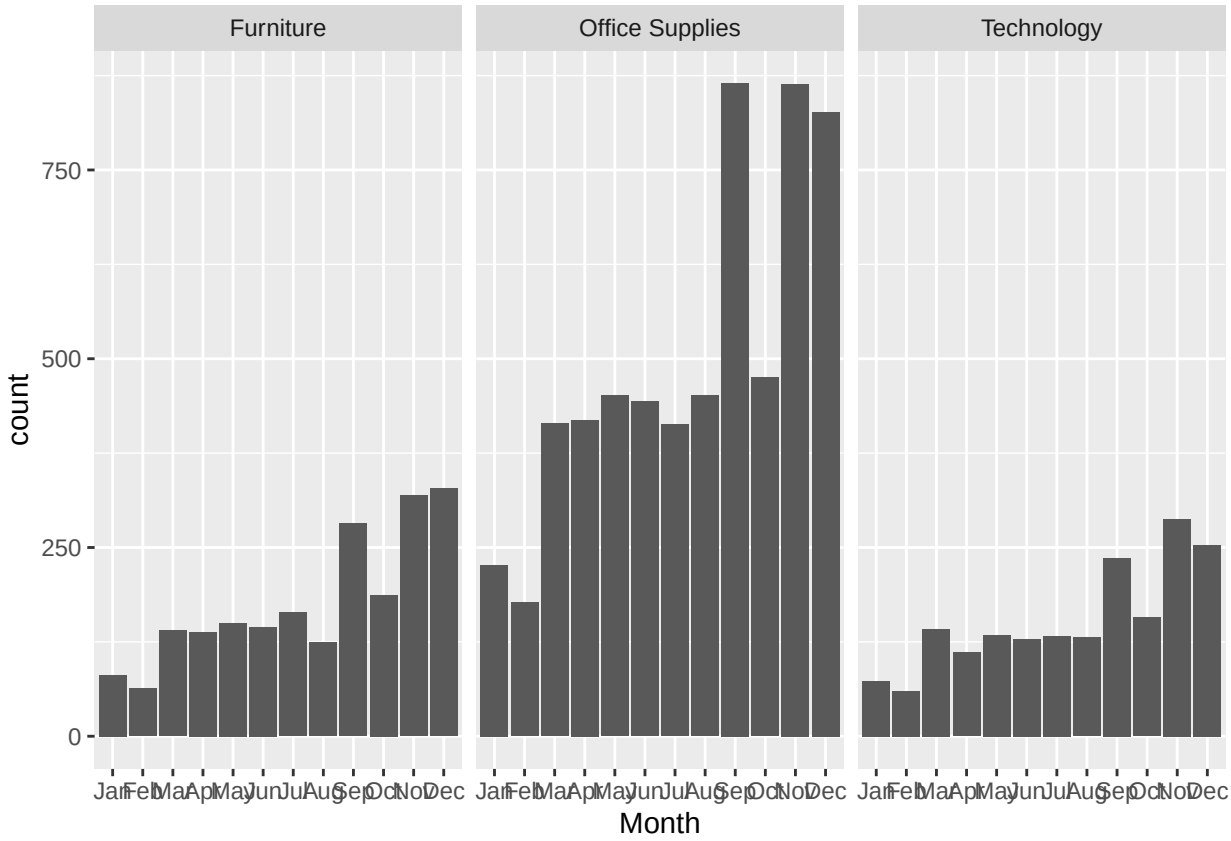

```
# Extract year from date
superstore$Year<-year(superstore$`Order Date`)
```

```
# Extract month from date
superstore$Month<-month(superstore$`Order Date`, label=TRUE)
```

```
# Distribution of sales records by month
ggplot(superstore, aes(Month))+geom_bar()
```

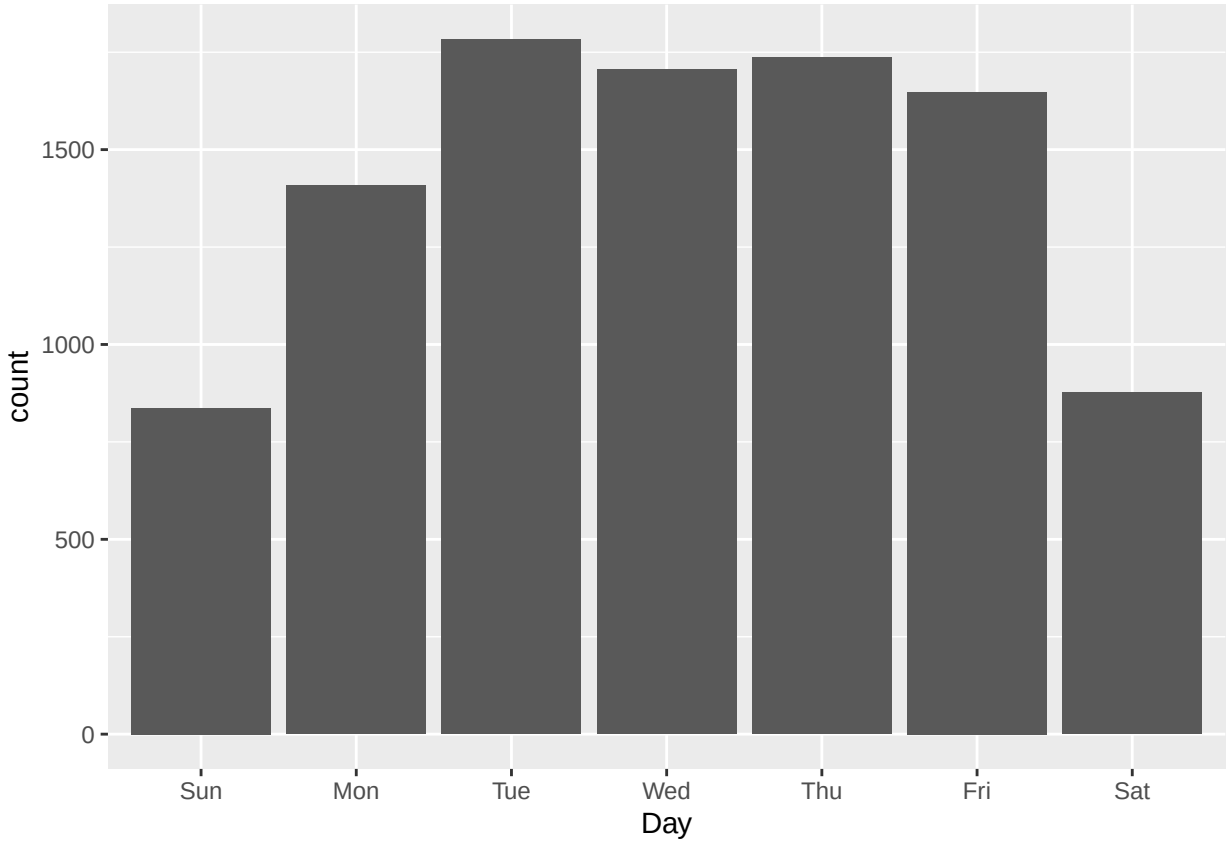



```
# Does the sales record distribution by month vary be category
ggplot(superstore, aes(Month))+geom_bar()+facet_grid(.~Category)
```

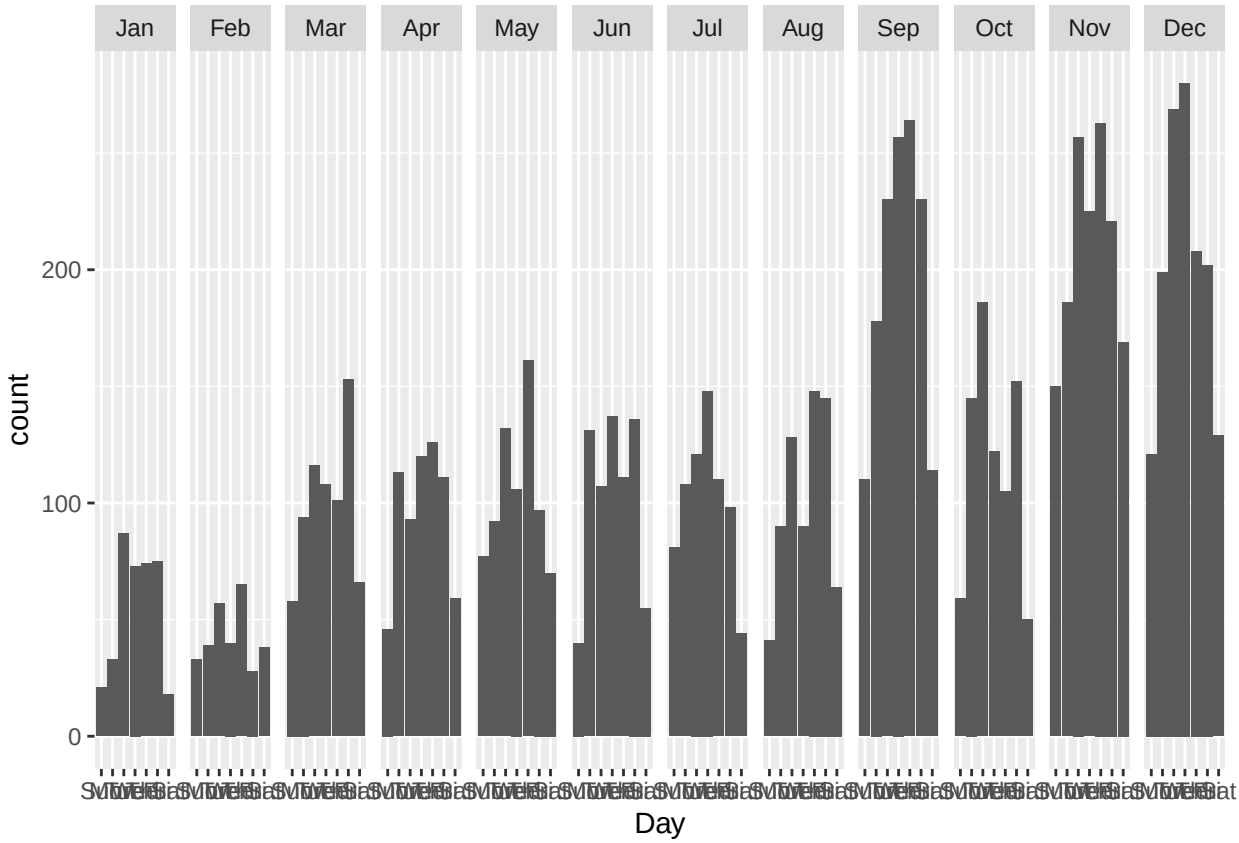



```
# Extract day from date
superstore$Day<-wday(superstore$`Order Date`, label=T)
```

```
# Distribution of sales records by day
ggplot(superstore, aes(Day))+geom_bar()
```




```
# Distribution of sales records by day and by month
ggplot(superstore, aes(Day))+geom_bar()+facet_grid(.~Month)
```



Data Aggregation

```
library(tidyr)
library(dplyr)
```

```
##
## 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
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
# Calculate number of records by category
superstore %>%
  group_by(Category) %>%
  summarise(Number_of_records=n())%>%
  arrange(desc(Number_of_records))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   Category        Number_of_records
##   <chr>                       <int>
## 1 Office Supplies              6026
## 2 Furniture                    2121
## 3 Technology                   1847
```

```r
# Calculate number of records by category and region
superstore %>%
  group_by(Category, Region) %>%
  summarise(Number_of_records=n())%>%
  arrange(desc(Number_of_records))
```

```
## `summarise()` regrouping output by 'Category' (override with `.groups` argument)
```

```
## # A tibble: 12 x 3
## # Groups:   Category [3]
##    Category        Region  Number_of_records
##    <chr>           <chr>               <int>
##  1 Office Supplies West                 1897
##  2 Office Supplies East                 1712
##  3 Office Supplies Central              1422
##  4 Office Supplies South                 995
##  5 Furniture       West                  707
##  6 Furniture       East                  601
##  7 Technology      West                  599
##  8 Technology      East                  535
##  9 Furniture       Central               481
## 10 Technology      Central               420
## 11 Furniture       South                 332
## 12 Technology      South                 293
```

```r
# One aggregation function on one variable
superstore %>%
  group_by(Category) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T)) %>%
  arrange(desc(Total_Sales))
```

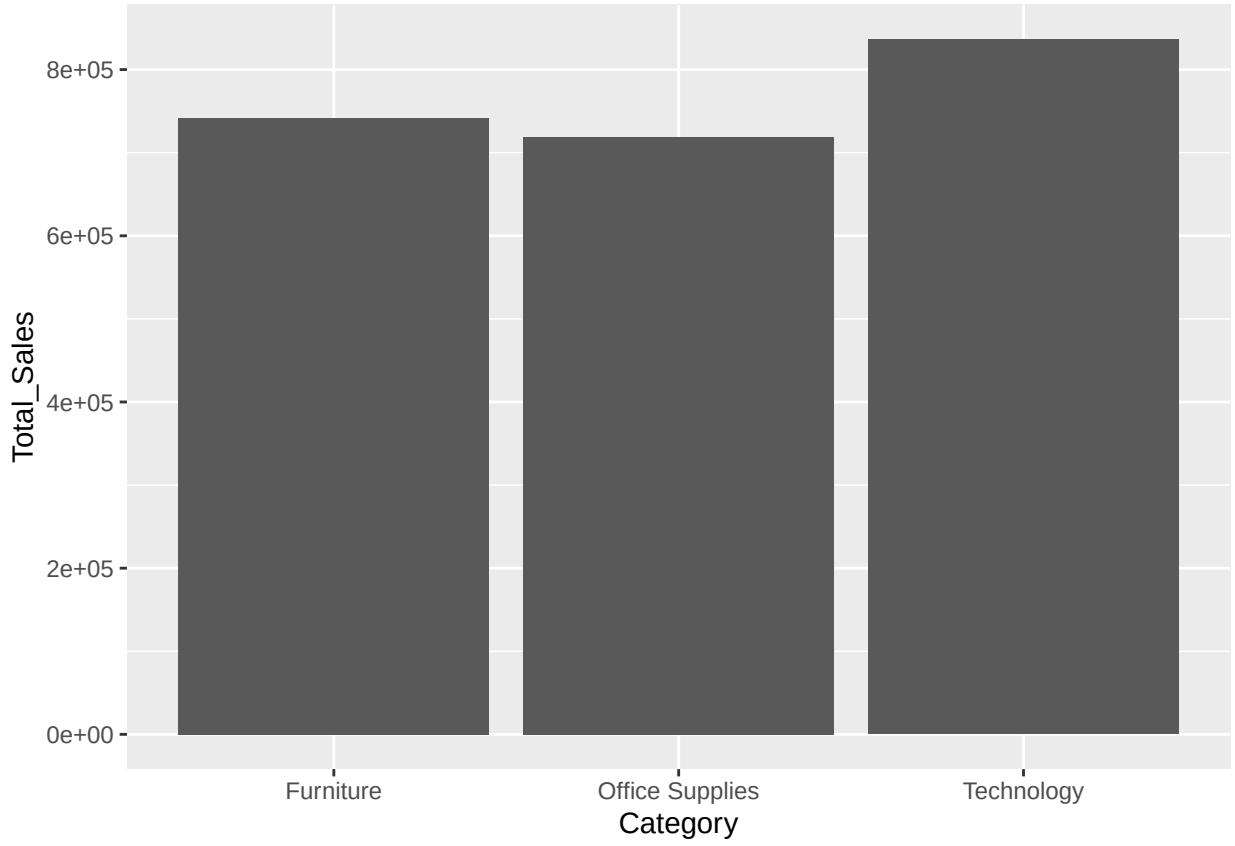
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   Category        Total_Sales
##   <chr>                 <dbl>
```

```
## 1 Technology          836154.
## 2 Furniture           742000.
## 3 Office Supplies     719047.
```

```
# One aggregation function on one variable with visualization
superstore %>%
  group_by(Category) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T)) %>%
  arrange(desc(Total_Sales)) %>%
  ggplot(aes(x=Category, y=Total_Sales))+geom_bar(stat="identity")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```




```
# One aggregation function on multiple variables
superstore %>%
  group_by(Category) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T), Total_Profit=sum(Profit,na.rm=T))%>%
  arrange(desc(Total_Sales))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 3
##   Category        Total_Sales Total_Profit
##   <chr>                 <dbl>        <dbl>
```

```
## 1 Technology          836154.      145455.
## 2 Furniture           742000.       18451.
## 3 Office Supplies     719047.      122491.
```

```
# One aggregation function on multiple variables
superstore %>%
  group_by(Category, Region) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T), Total_Profit=sum(Profit,na.rm=T))%>%
  arrange(desc(Total_Sales))
```

```
## `summarise()` regrouping output by 'Category' (override with `.groups` argument)
```

```
## # A tibble: 12 x 4
## # Groups:   Category [3]
##    Category        Region  Total_Sales Total_Profit
##    <chr>           <chr>         <dbl>        <dbl>
##  1 Technology      East         264974.       47462.
##  2 Furniture       West         252613.       11505.
##  3 Technology      West         251992.       44304.
##  4 Office Supplies West         220853.       52610.
##  5 Furniture       East         208291.        3046.
##  6 Office Supplies East         205516.       41015.
##  7 Technology      Central      170416.       33697.
##  8 Office Supplies Central      167026.        8880.
##  9 Furniture       Central      163797.       -2871.
## 10 Technology      South        148772.       19992.
## 11 Office Supplies South        125651.       19986.
## 12 Furniture       South        117299.        6771.
```

```
# Multiple aggregation functions on one variable
superstore %>%
  group_by(Category) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T), Avg_Sales=mean(Sales,na.rm=T))%>%
  arrange(desc(Total_Sales))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 3
##   Category        Total_Sales Avg_Sales
##   <chr>                 <dbl>     <dbl>
## 1 Technology          836154.      453.
## 2 Furniture           742000.      350.
## 3 Office Supplies     719047.      119.
```

```
# Multiple aggregation functions on multiple variables
superstore %>%
  group_by(Category) %>%
  summarise(Total_Sales=sum(Sales,na.rm=T), Avg_Profit=mean(Profit,na.rm=T))%>%
  arrange(desc(Total_Sales))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 3
##   Category        Total_Sales Avg_Profit
##   <chr>                 <dbl>      <dbl>
## 1 Technology          836154.      78.8
## 2 Furniture           742000.       8.70
## 3 Office Supplies     719047.      20.3
```

```
# Note that sum function is different from count function
superstore %>%
  group_by(Category) %>%
  summarise(Number_of_records=n(), Total_Sales=sum(Sales,na.rm=T))%>%
  arrange(desc(Total_Sales))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 3
##   Category        Number_of_records Total_Sales
##   <chr>                       <int>       <dbl>
## 1 Technology                   1847     836154.
## 2 Furniture                    2121     742000.
## 3 Office Supplies              6026     719047.
```

Real world application

```
nydata<-read.csv('ny_accidents.csv', na.strings = "")
```

```
# Convert date format
nydata$CRASH.DATE<-as_date(nydata$CRASH.DATE, format="%m/%d/%y")
```

```
# See how number of motor vehicle collisions change over days
nydata$Day<-wday(nydata$CRASH.DATE, label=T)
```

```
nydata %>%
  group_by(Day)%>%
  summarise(Total_Incidents=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 7 x 2
##   Day   Total_Incidents
##   <ord>           <int>
## 1 Sun            252021
## 2 Mon            250014
## 3 Tue            244362
## 4 Wed            249376
## 5 Thu            251740
## 6 Fri            252942
## 7 Sat            248507
```

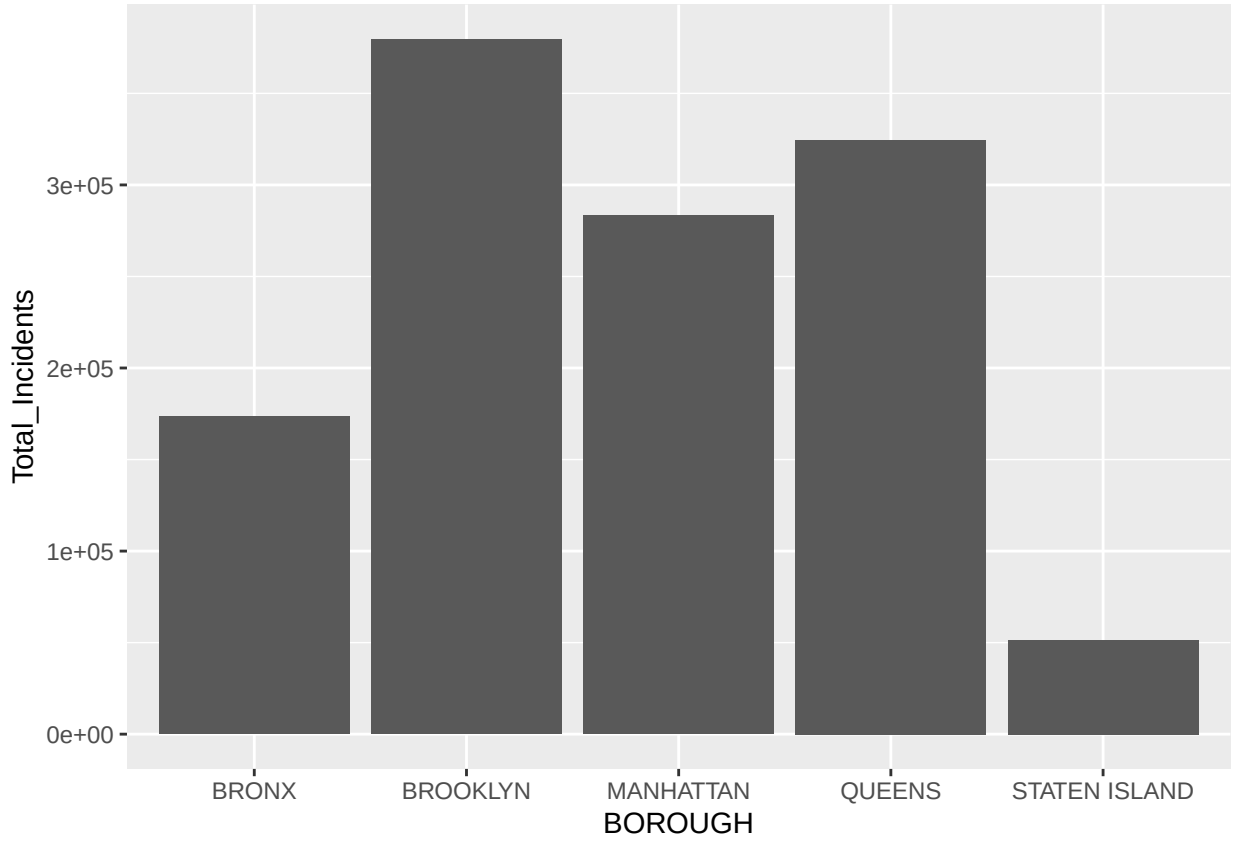
```r
# Check number of collisions by Borough
nydata %>%
  group_by(BOROUGH)%>%
  summarise(Total_Incidents=n())%>%
  arrange(desc(Total_Incidents))%>%
  drop_na()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   BOROUGH       Total_Incidents
##   <chr>                   <int>
## 1 BROOKLYN               379520
## 2 QUEENS                 324566
## 3 MANHATTAN              283140
## 4 BRONX                  173601
## 5 STATEN ISLAND           51419
```

```r
# Check number of collisions by Borough and plot
nydata %>%
  group_by(BOROUGH)%>%
  summarise(Total_Incidents=n())%>%
  arrange(desc(Total_Incidents))%>%
  drop_na()%>%
   ggplot(aes(x=BOROUGH, y=Total_Incidents))+geom_bar(stat="identity")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```




```
# Task 1: Which zip code in Bronx has the maximum number of collisions?
```

```
# What are the top three contributing factors for vehicle 1 involved in a collision?
```

```
# Which Borough has the highest number of persons injured?
```

Pivot using tidyr package

```
# wide to long form
data<-relig_income
answer<-data %>%
  pivot_longer(-religion,
               names_to = "income",
               values_to = "count",
               values_drop_na = TRUE)
```

```
data<-billboard
answer <- data %>%
  pivot_longer(-c(artist, track, date.entered),
    names_to = "week",
    values_to = "rank",
    values_drop_na = TRUE)
```

```r
# long to wide form
data<-fish_encounters
answer <- data %>%
  pivot_wider(names_from = station, values_from = seen)
```

```r
answer<-data %>%
  pivot_wider(
    names_from = station,
    values_from = seen,
    values_fill = list(seen = 0)
  )
```