#Import the dataset

```
Aisles = read.csv("/Users/shimonyagrawal/Desktop/Instacart-market-basket-analysis/aisles.csv")
Departments = read.csv("/Users/shimonyagrawal/Desktop/Instacart-market-basket-analysis/departments.csv")
Order_Prior = read.csv("/Users/shimonyagrawal/Desktop/Instacart-market-basket-analysis/order_products__
Orders = read.csv("/Users/shimonyagrawal/Desktop/Instacart-market-basket-analysis/orders.csv")
Products = read.csv("/Users/shimonyagrawal/Desktop/Instacart-market-basket-analysis/products.csv")
```

#Install packages

```
tinytex::install_tinytex()
```

```
## Warning: Detected an existing tlmgr at /usr/local/bin/tlmgr. It seems TeX
## Live has been installed (check tinytex::tinytex_root()). You are recommended
## to uninstall it, although TinyTeX should work well alongside another LaTeX
## distribution if a LaTeX document is compiled through tinytex::latexmk().

## TinyTeX installed to /Users/shimonyagrawal/Library/TinyTeX
```

```
install.packages("DBI")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
install.packages("odbc")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
install.packages("lubridate")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
install.packages("GGally")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("forecast")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("readr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("treemap")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("scales")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```r
install.packages("tidyr")
```

```
##
##   There is a binary version available but the source version is later:
##       binary source needs_compilation
## tidyr  1.0.3  1.1.0              TRUE

## installing the source package 'tidyr'
```

```r
install.packages("arules")
```

```
##
##   There is a binary version available but the source version is later:
##        binary source needs_compilation
## arules  1.6-5  1.6-6              TRUE

## installing the source package 'arules'
```

```
install.packages("arulesViz")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
install.packages("methods")
```

```
## Warning: package 'methods' is not available (for R version 4.0.0)

## Warning: package 'methods' is a base package, and should not be updated
```

```
install.packages("plyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rj/t11km2gs693dyq4szgcrm4vr0000gn/T//RtmpXHDz95/downloaded_packages
```

```
library(DBI)
library(odbc)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------- tidyver
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------- tidyverse_con
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union

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

```
library (GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

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

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(ggplot2)
library(readr)
library(dplyr)
library(treemap)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(tidyr)
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'arules'
```

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

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
```

```
## Loading required package: grid

## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

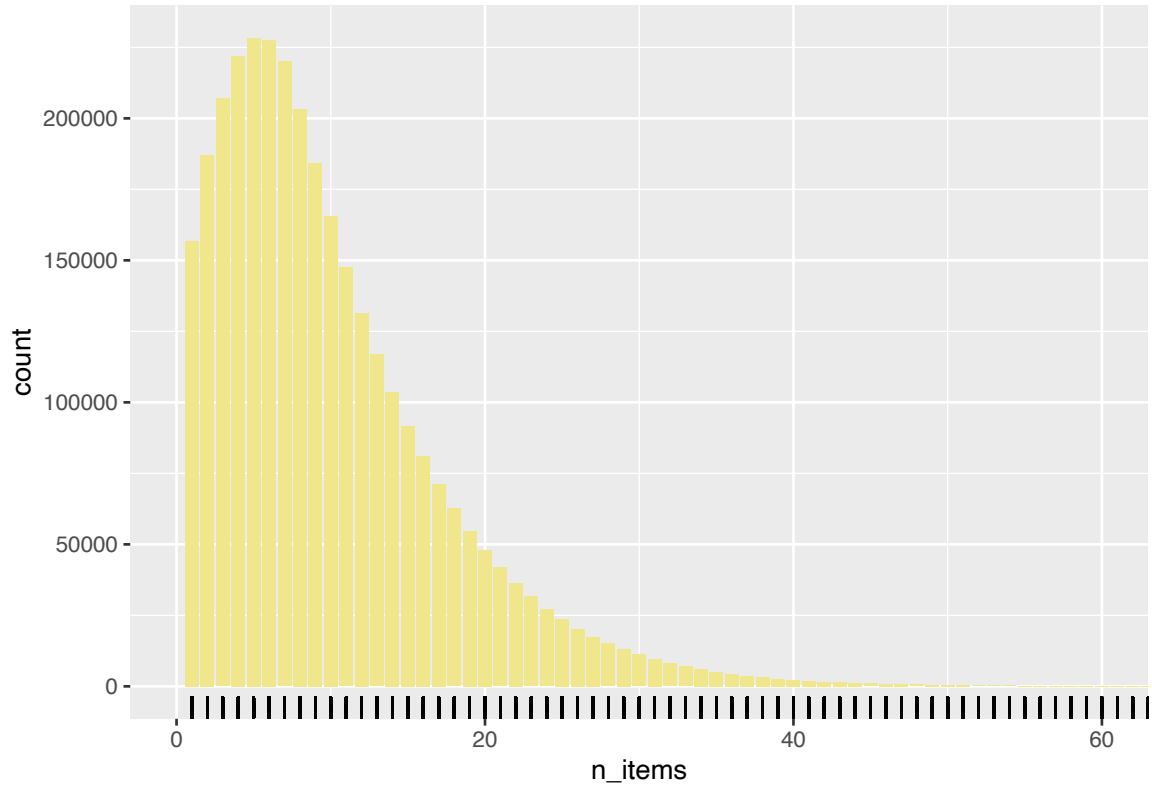
```
library(methods)
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

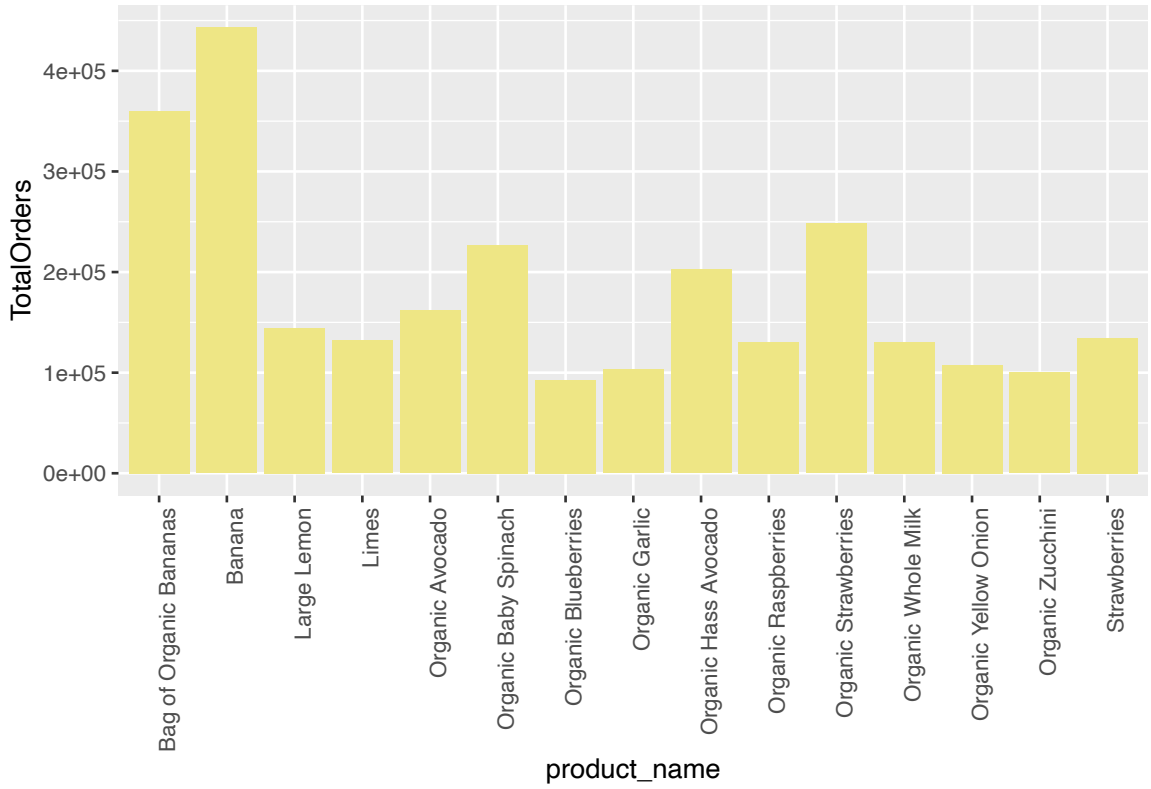
```
#Create a main data frame of the data set with relevant information

eda_df <- Products %>%
  left_join(Aisles, by = "aisle_id") %>%
  left_join(Departments, by = "department_id") %>%
  left_join(Order_Prior, by = "product_id") %>%
  left_join(Orders, by = "order_id") %>%
  drop_na()

view(eda_df)
```

```r
#How many items do people order in a single purchase?

Order_Prior %>%
  group_by(order_id) %>%
  dplyr::summarize(n_items = last(add_to_cart_order)) %>%
  ggplot(aes(x=n_items)) +
  geom_histogram(stat = "count", fill = "khaki") +
  geom_rug() +
  coord_cartesian(xlim = c(0,60))
```

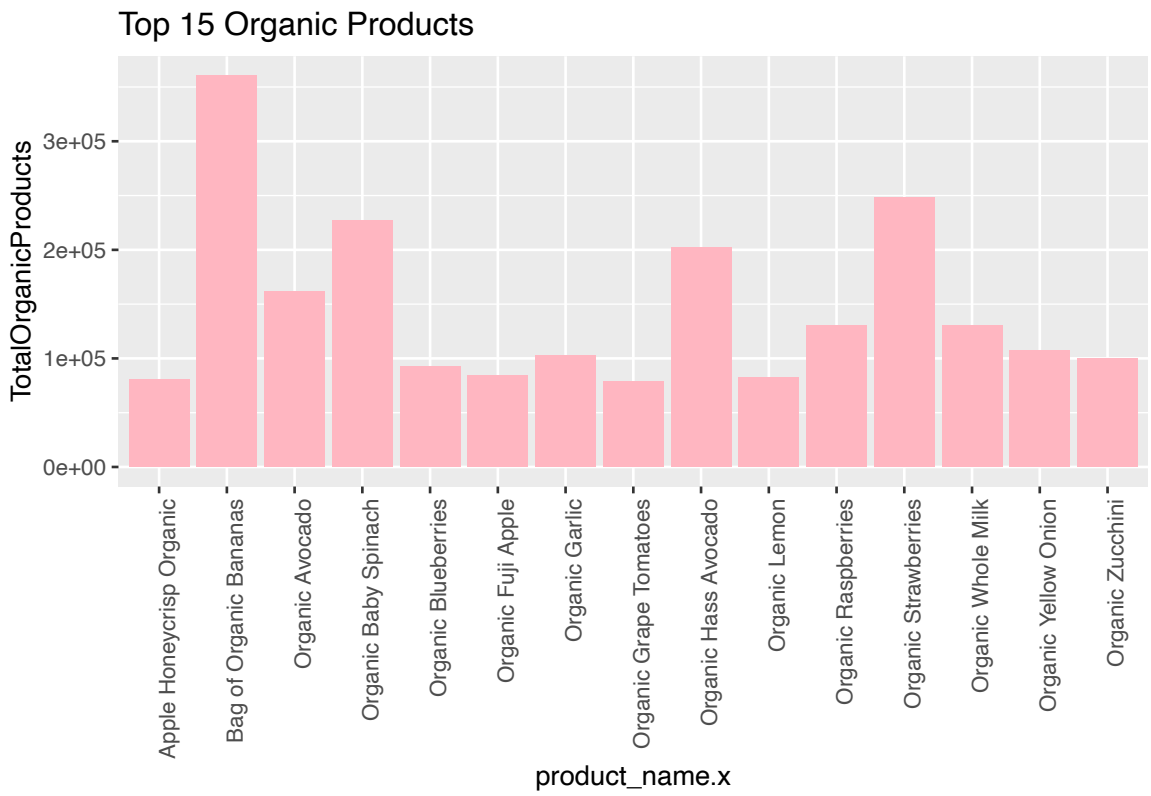
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



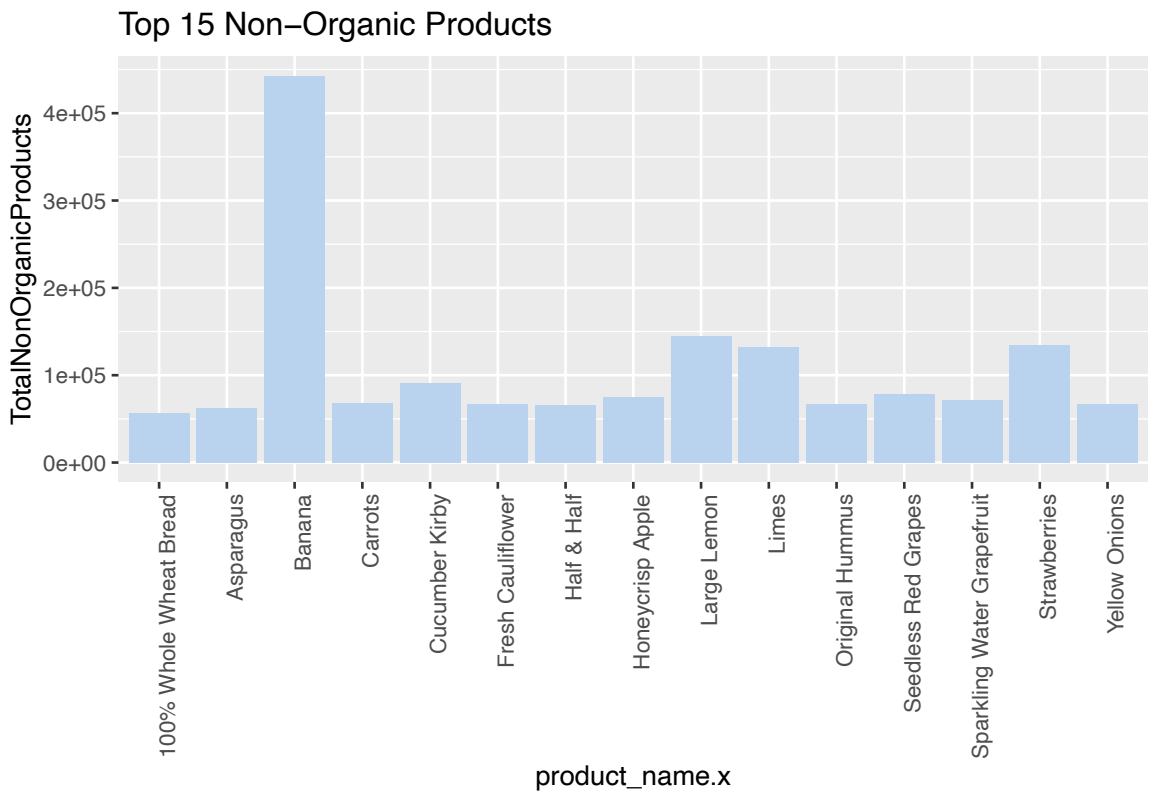

```r
#What are Instacart's most ordered items?

MostSold = eda_df %>%
  select(product_id, product_name) %>%
  group_by(product_name) %>%
  dplyr::summarise(TotalOrders = n())

Top15Sold <- MostSold %>%
  filter(TotalOrders > 92957)

ggplot(data=Top15Sold, aes(x = product_name, y = TotalOrders)) +
  geom_bar(stat = "identity", fill = "khaki2") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



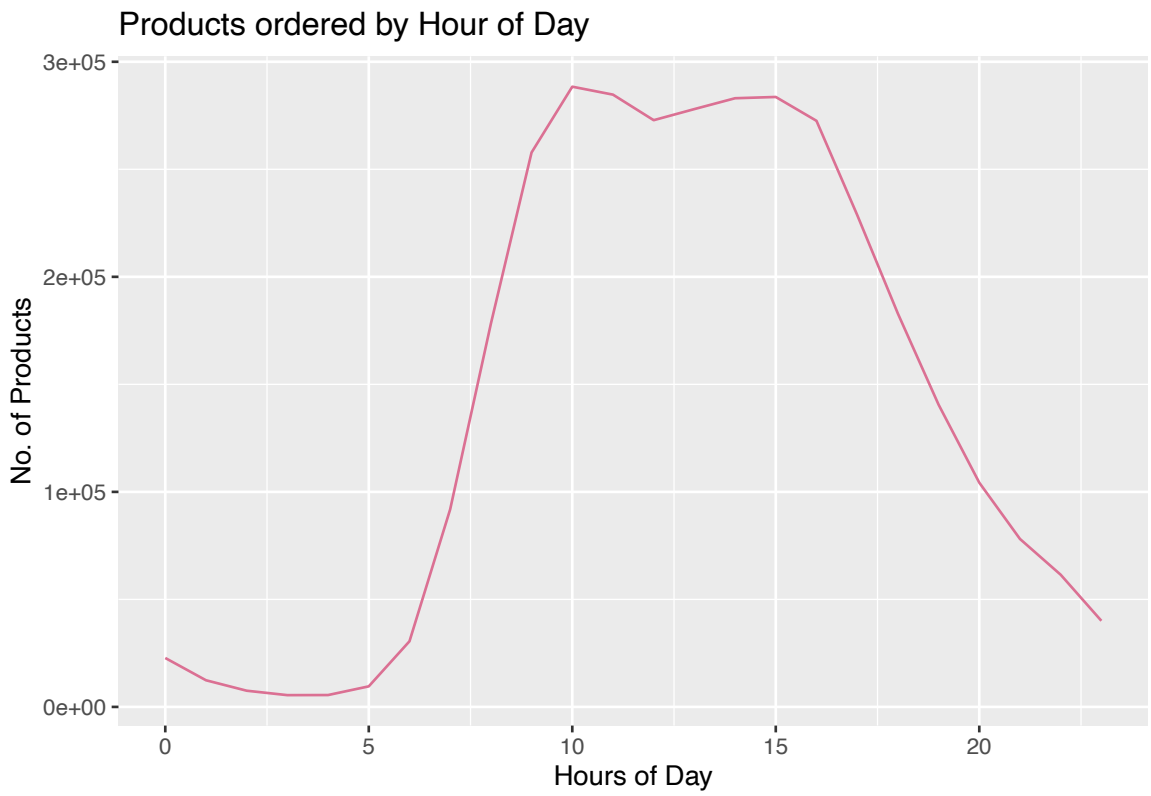

```
#What products do people order the most: organic vs non organic products?

Organic_Nonorganic <- Products %>%
  mutate(organic=ifelse(str_detect(str_to_lower(Products$product_name),'organic'),"organic","not organi

OrganicPro <- Organic_Nonorganic %>%
  filter(organic == "organic") %>%
  left_join(eda_df, by = "product_id") %>%
  select (product_id, product_name.x) %>%
  group_by(product_name.x) %>%
  dplyr::summarise(TotalOrganicProducts = n())

Top15OrganicProducts <- OrganicPro %>%
  filter(TotalOrganicProducts > 78805)

ggplot(data=Top15OrganicProducts, aes(x=product_name.x, y=TotalOrganicProducts))+
  geom_bar(stat="identity", fill="lightpink")+
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=1))+
  ggtitle("Top 15 Organic Products")
```



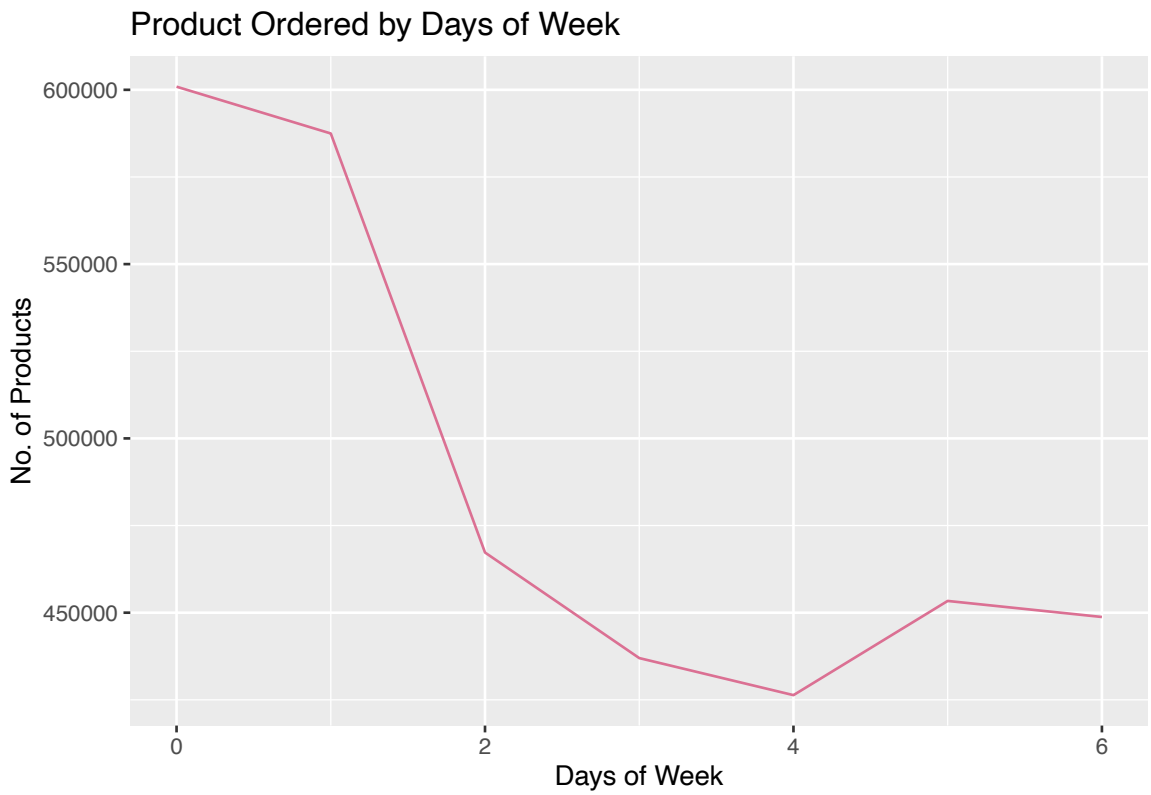

```r
NonorganicPro <- Organic_Nonorganic %>%
  filter(organic == "not organic") %>%
  left_join(eda_df, by = "product_id") %>%
  select (product_id, product_name.x) %>%
  group_by(product_name.x) %>%
  dplyr::summarise(TotalNonOrganicProducts = n())

Top15NonorganicProducts <- NonorganicPro %>%
  filter(TotalNonOrganicProducts > 56768)

ggplot(data=Top15NonorganicProducts, aes(x=product_name.x, y=TotalNonOrganicProducts))+
  geom_bar(stat="identity", fill="slategray2")+
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=1))+
  ggtitle("Top 15 Non-Organic Products")
```




```
#What hour of day and day of week most products are ordered?

#hour of day
OrderByHour <- Orders %>%
  select(order_id, order_hour_of_day)%>%
  group_by(order_hour_of_day)%>%
  dplyr::count()

ggplot(data=OrderByHour, aes(x=order_hour_of_day, y= n))+
  geom_line(color = "palevioletred") +
  ggtitle("Products ordered by Hour of Day") +
  xlab("Hours of Day") +
  ylab("No. of Products")
```




```
MeanOrderHour <- Orders %>%
  dplyr::summarise(AvgHour = mean(order_hour_of_day))

#day of week
OrderByDay <- Orders %>%
  select (order_id, order_dow) %>%
  group_by(order_dow) %>%
  dplyr::count ()

ggplot(data=OrderByDay, aes(x=order_dow, y=n))+
  geom_line(color = "palevioletred") +
  ggtitle ("Product Ordered by Days of Week") +
  xlab("Days of Week") +
  ylab ("No. of Products")
```



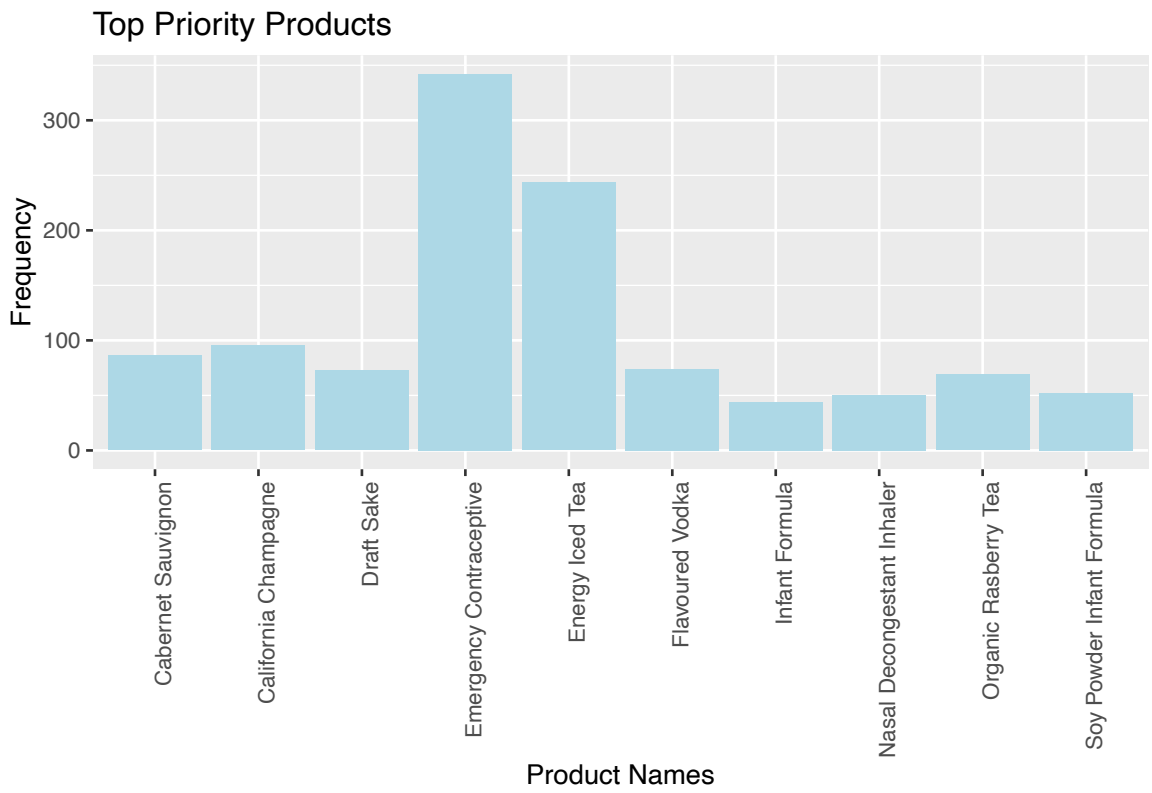

```
MeanOrderWeek <- Orders %>%
  dplyr::summarise(AvgDay = mean(order_dow))
```

```
#What items are first added to the cart by Instacart's users?

priority <- Order_Prior %>%
  group_by(product_id, add_to_cart_order) %>%
  dplyr::summarize(count = n()) %>%
  mutate(average=count/sum(count)) %>%
  filter(add_to_cart_order==1, count>10) %>%
  arrange(desc(average)) %>%
  mutate(average1= average*100000) %>%
  left_join(Products, by="product_id") %>%
  select(product_name, average1, average, count) %>%
  ungroup() %>%
  top_n(10, wt=average1)
```

```
## Adding missing grouping variables: `product_id`
```

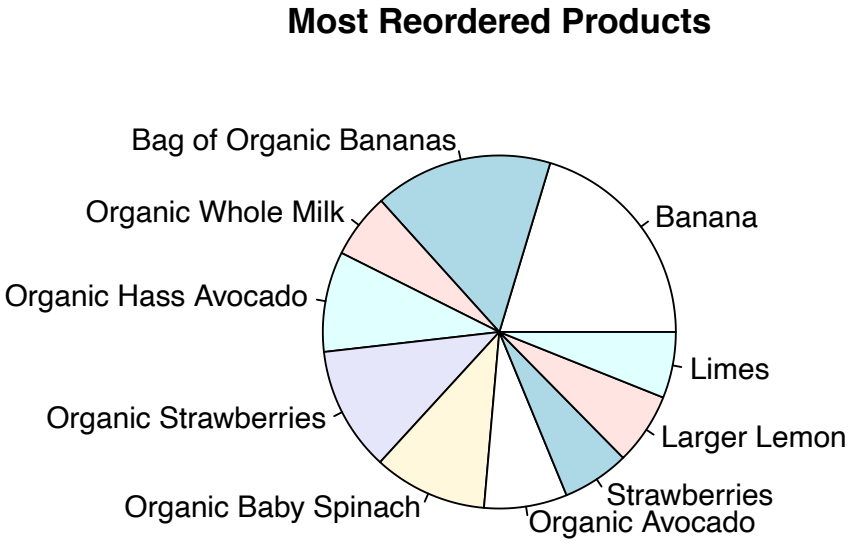
```
products <- c("Emergency Contraceptive", "Energy Iced Tea" ,"California Champagne",
              "Cabernet Sauvignon", "Flavoured Vodka", "Draft Sake", "Organic Rasberry Tea",
              "Soy Powder Infant Formula", "Nasal Decongestant Inhaler", "Infant Formula")

count <- c(37,51,14,14,50,25,27,24,30,29)
```

```
ggplot(data = priority,
       aes(x = c("Emergency Contraceptive", "Energy Iced Tea" ,"California Champagne",
                 "Cabernet Sauvignon", "Flavoured Vodka", "Draft Sake", "Organic Rasberry Tea",
                 "Soy Powder Infant Formula", "Nasal Decongestant Inhaler", "Infant Formula"),
           y= average1)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=1))+
  ggtitle("Top Priority Products") +
  xlab ("Product Names") +
  ylab ("Frequency")
```




```
#What are Instacart's most reordered products?

MostReordered_Number<- Order_Prior%>%
  group_by(product_id)%>%
  dplyr::summarize(proportion_reordered = mean(reordered), n=n())%>%
  top_n(10, wt=n)%>%
  arrange(desc(proportion_reordered))%>%
  left_join(Products,by="product_id")

#pie chart
names <- c("Banana", "Bag of Organic Bananas", "Organic Whole Milk", "Organic Hass Avocado",
           "Organic Strawberries", "Organic Baby Spinach", "Organic Avocado",
           "Strawberries", "Larger Lemon", "Limes")
n <- c(472565, 379450, 137905,213584,264683,    241921,176815,142951,152657,140627)
```

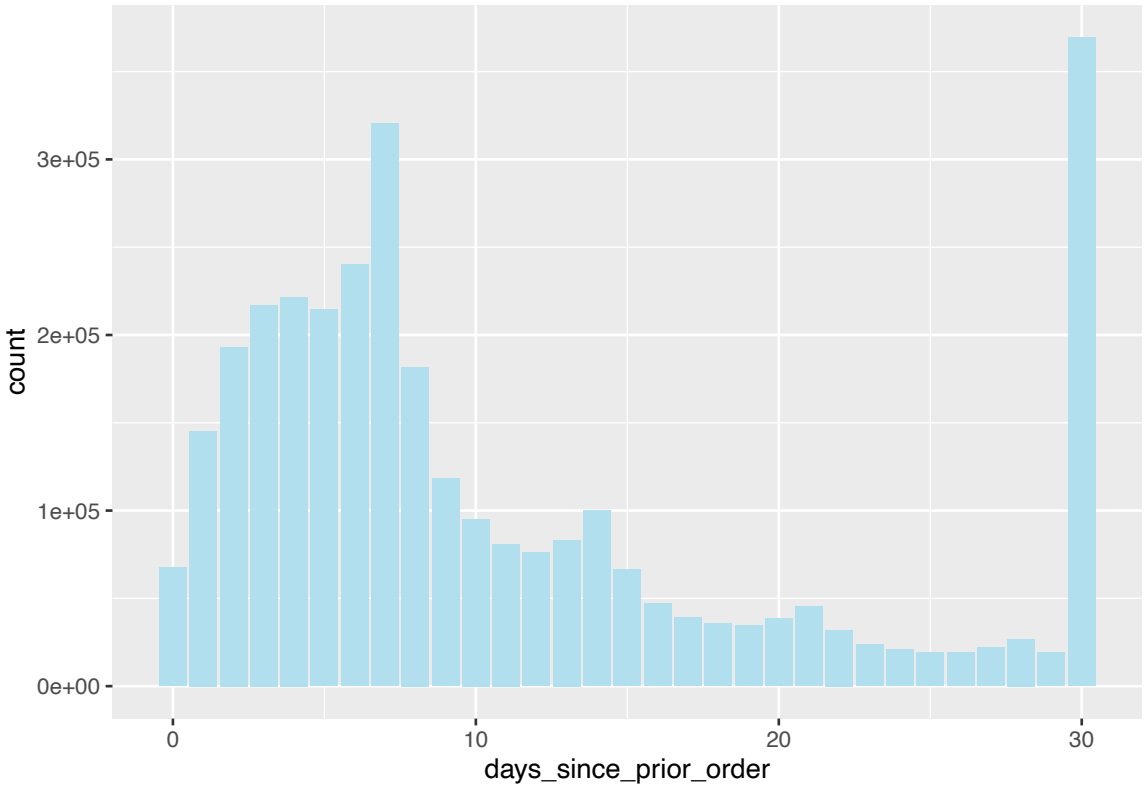
```
pie(n, names, main = "Most Reordered Products")
```




```
#When do Instacart's users place the next order?

Orders %>%
  ggplot(aes(x=days_since_prior_order)) +
  geom_histogram(stat="count",fill="lightblue2")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Warning: Removed 206209 rows containing non-finite values (stat_count).
```


3e+05
2e+05
1e+05
0e+00
count
0
10
20
30
days_since_prior_order