R Project

Krishna Hemant, Snehal Rajwar, Ashok Thiruvengadam

01/03/2022

```
sampled <- read.csv("sampled.csv")
```

R Markdown

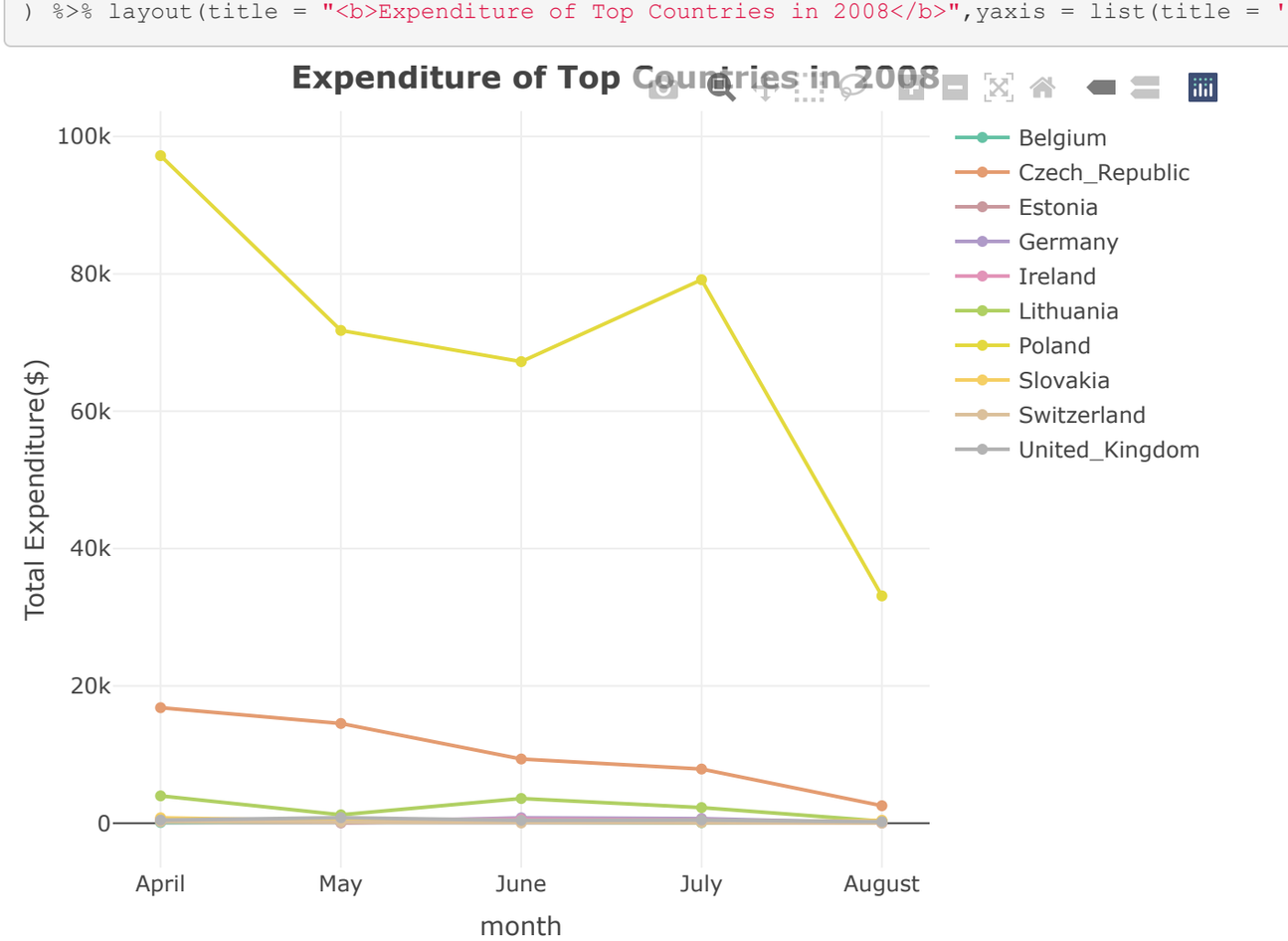
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

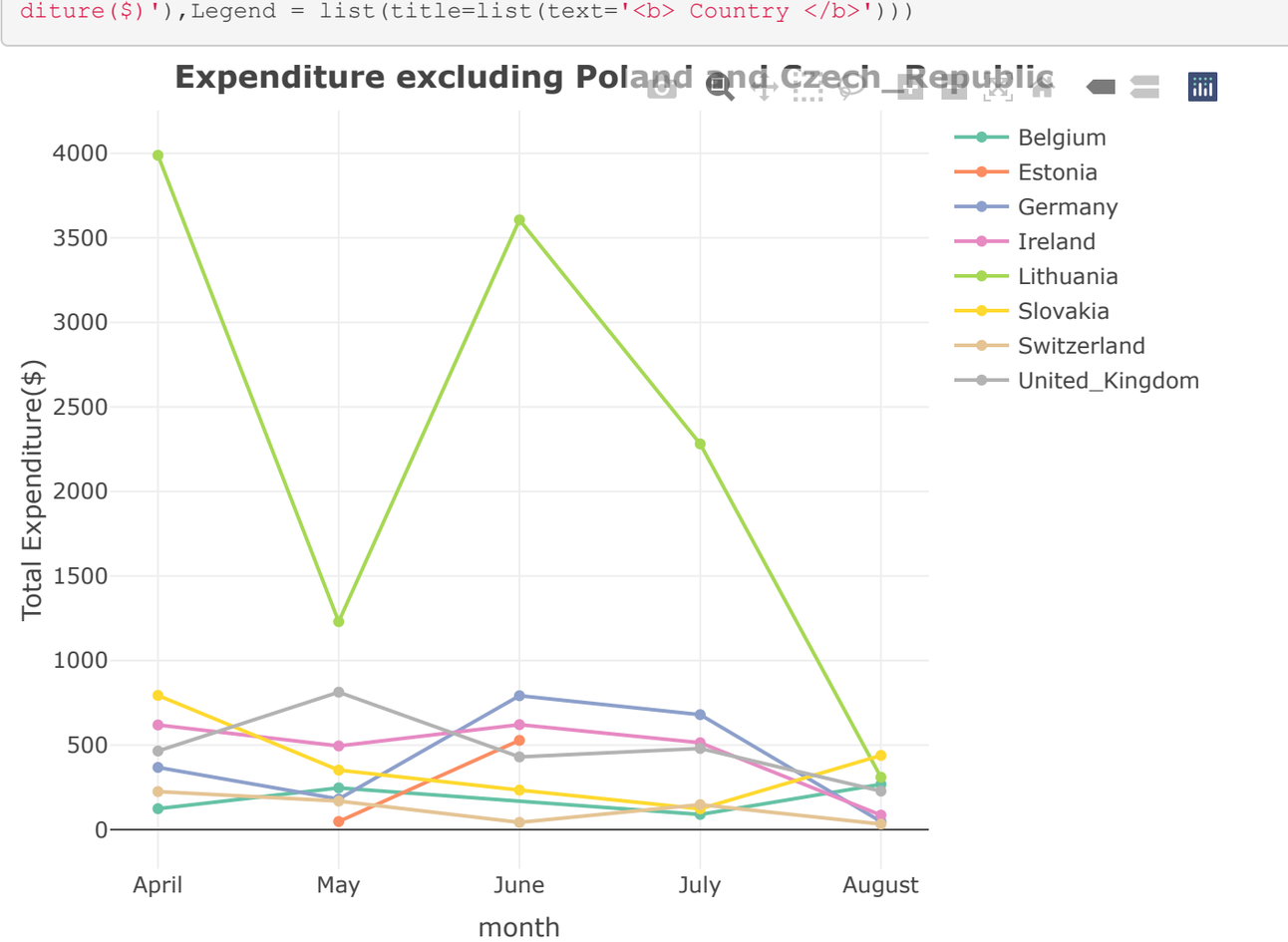
```
library(ggplot2)
library(dplyr)
library(tidyr)
library(gridExtra)
library(lubridate)
library(plotly)
library(treemapify)
library(scales)
library(ggalluvial)
library(RColorBrewer)
library(forcats)
library(treemap)
library(hrbrthemes)
library(readr)
library(readxl)
library(magrittr)
library(tidyverse)
library(sf)
library(rnaturalearth)
library(rnaturalearthdata)
library(worldmet)
```

```
#Forming the sampled data frame
df <- sampled %>% select(country,price) %>% group_by(country) %>% summarise(total_price = sum(price)) %>% arrange
(-total_price)
df1<- sampled %>% select(month,country,price) %>% group_by(month,country) %>% summarise(total_price = sum(price))

#Filtering domain names from countries and separating both
domains <- c('net(*.net)','com(*.com)','int(*.int)','org(*.org)','biz(*.biz)')

domain_df <- df %>% filter((df$country %in% domains))

country_total <- df %>% filter(!(df$country %in% domains))
```

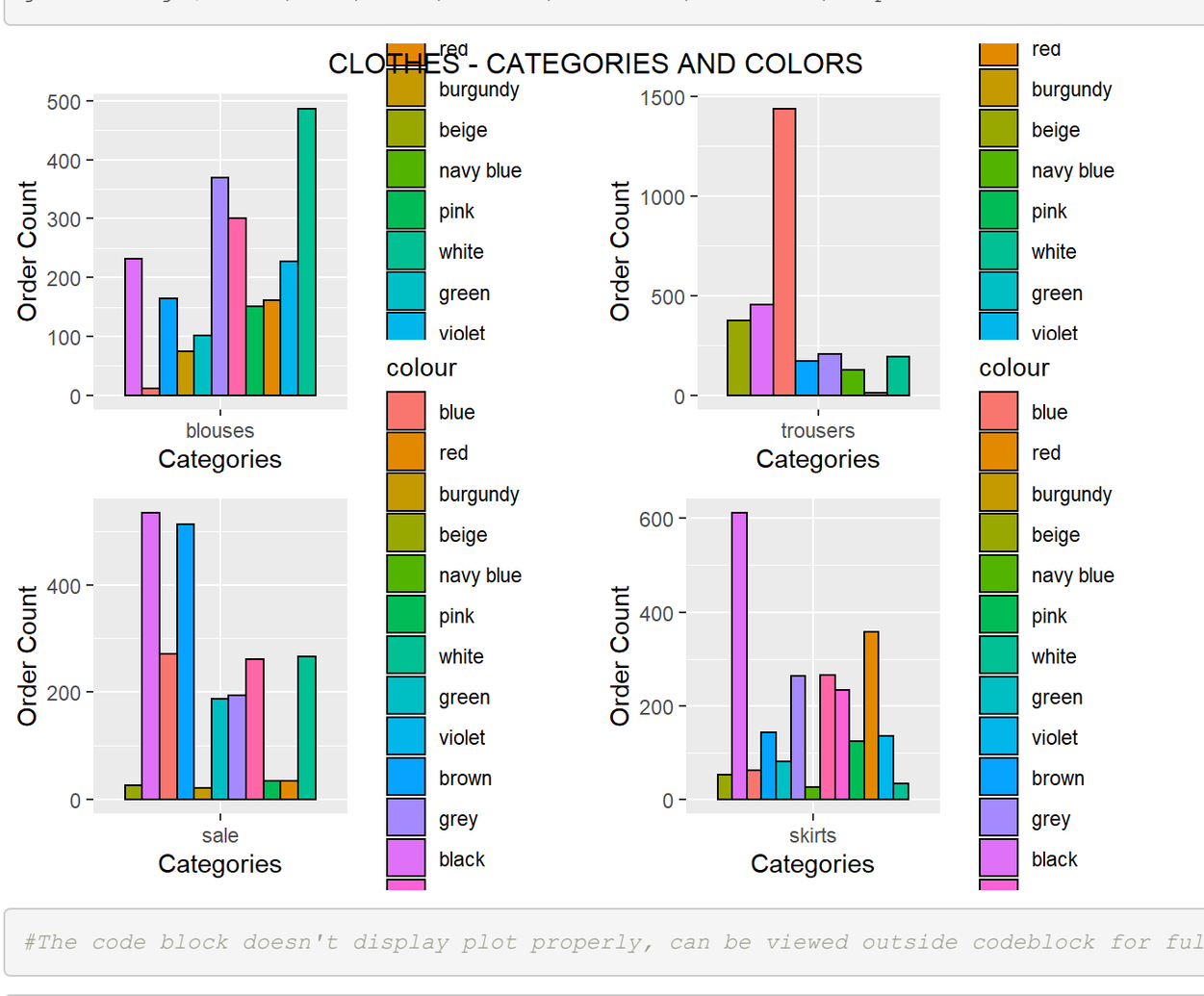
```
#line plot for various countries over four months of 2008 and the price spent
h <- head(country_total,10)
vec <- as.vector(h$country)

#sorting months in order
dd<- ungroup(df1)
x <- c("April","May",'June','July','August')
dd <- dd %>% filter(dd$country %in% vec) %>% mutate(month =  factor(month, levels = x)) %>% arrange(month)


#A line plot to show the total amount spent by the top 10 countries through the months of 2008

#This data in the plot shows that Poland has the highest expenditure since it's a Polish E-commerce company follo
wed by neighbors Czech_Republic and Lithuania. United kingdom seems to be an exception since it's a booming econo
my.
plot_ly(
        data = dd,
        x = ~month,
        y = ~total_price,
        color = ~country,
        type = "scatter",
        mode = "lines+markers"
) %>% layout(title = "<b>Expenditure of Top Countries in 2008</b>",yaxis = list(title = ' Total Expenditure($)'))
```




```
#Without Poland and Czech_Republic
# This plot shows a more zoomed in version of the previous plot excluding the top two countries.Since the data is
from 2008 it can be observed that the financial crash affected counties over various quarters especially in Q4 wh
ere sales fell to the lowest

dd1 <- dd %>% filter(!(dd$country %in% c('Poland','Czech_Republic')))

plot_ly(
        data = dd1,
        x = ~month,
        y = ~total_price,
        color = ~country,
        type = "scatter",
        mode = "lines+markers"
) %>% layout(title = "<b>Expenditure excluding Poland and Czech_Republic</b>", yaxis = list(title = ' Total Expen
diture($)'),Legend = list(title=list(text='<b> Country </b>')))
```

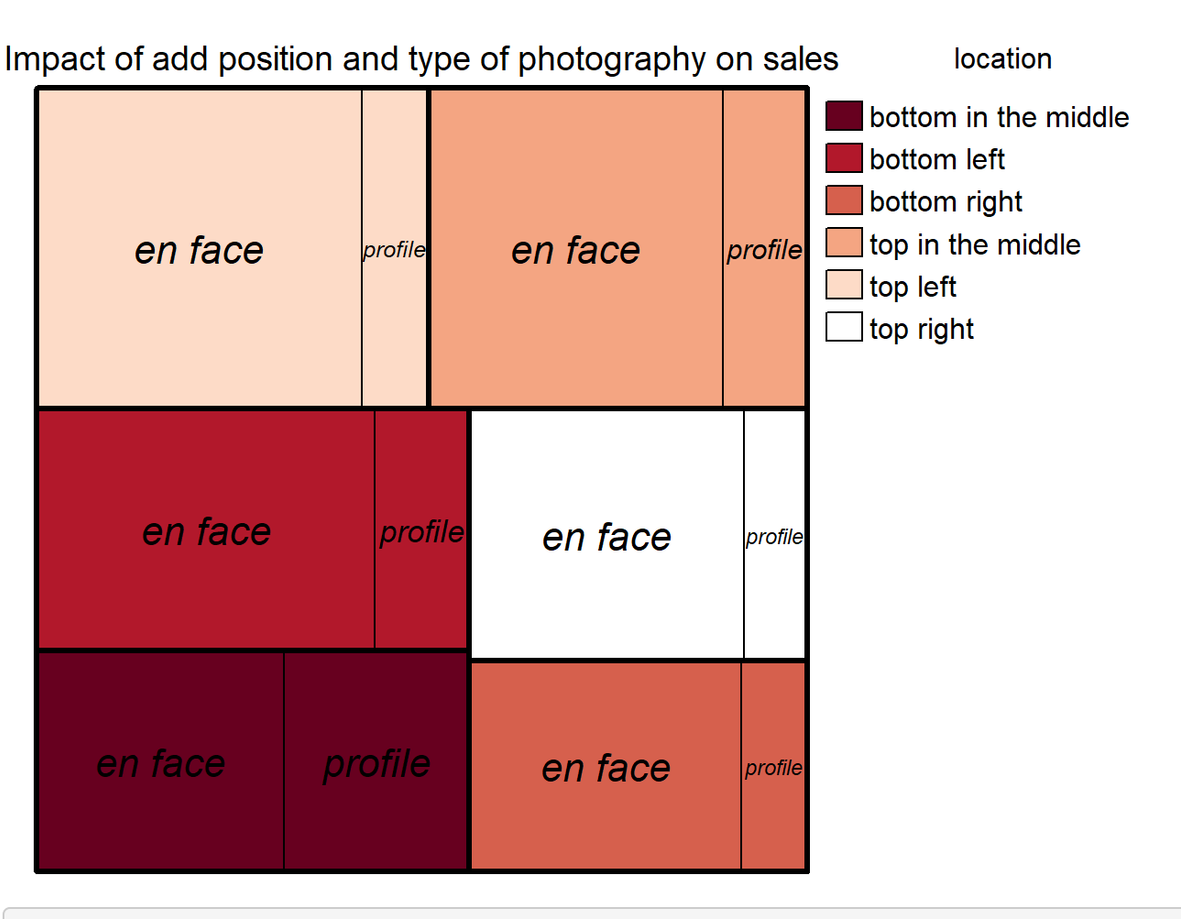



```
#Bar plot, categories vs colors
#Data frame for bar



def <- sampled %>% group_by(page.1..main.category.,colour) %>% summarise(count = n()) %>% arrange(count)
def<- def %>% group_by(page.1..main.category.) %>% arrange(count)

gg_color_hue <- function(n) {
  hues = seq(15, 375, length = n + 1)
  hcl(h = hues, l = 65, c = 100)[1:n]
}

s = unique(def$colour)
cols = setNames(gg_color_hue(length(s)), s)


blouse_df <- filter(def, page.1..main.category. == 'blouses') %>% arrange(count)
sale_df <- filter(def, page.1..main.category. == 'trousers')%>% arrange(count)
skirt_df <- filter(def, page.1..main.category. == 'sale')%>% arrange(count)
trouser_df <- filter(def, page.1..main.category. == 'skirts')%>% arrange(count)

#White blouses seem to have the most orders, followed by Grey
blouse<- ggplot(blouse_df, aes(x = page.1..main.category., y = count, fill = colour )) +
        geom_bar(stat="identity", colour = 'black', position = 'dodge')  +xlab("Categories") + ylab("Order Count"
) +scale_fill_manual(values = cols)

#Brown stands out for sale than other categories with a lot of orders
sale <- ggplot(sale_df, aes(x = page.1..main.category., y = count, fill = colour)) +
  geom_bar(stat="identity", colour = 'black', position = 'dodge')  +xlab("Categories") + ylab("Order Count") +sca
le_fill_manual(values = cols)

#Black seems to be a preferred color for skirts, and red too has a high preference
skirt <- ggplot(skirt_df, aes(x = page.1..main.category., y = count, fill = colour )) +
  geom_bar(stat="identity", colour = 'black', position = 'dodge') +xlab("Categories") + ylab("Order Count") +scal
e_fill_manual(values = cols)

#Blue trousers has the highest sales by a huge margin as it would have mostly been Jeans, followed by black and B
eige
trouser <- ggplot(trouser_df, aes(x = page.1..main.category., y = count, fill = colour )) +
  geom_bar(stat="identity", colour = 'black', position = 'dodge') +xlab("Categories") + ylab("Order Count") +scal
e_fill_manual(values = cols)

grid.arrange(blouse,sale,skirt,trouser, nrow = 2, ncol = 2, top = "CLOTHES - CATEGORIES AND COLORS")
```

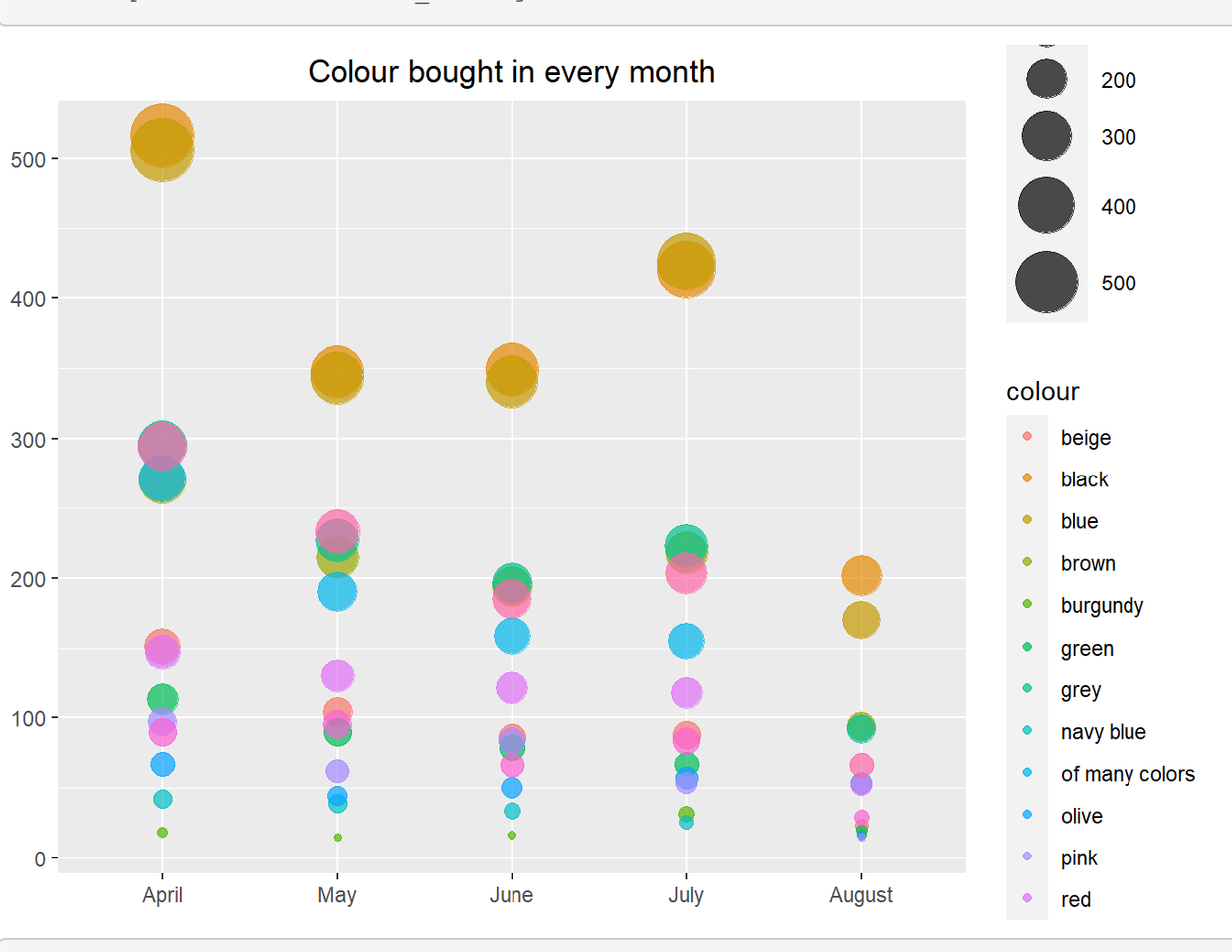



```
#The code block doesn't display plot properly, can be viewed outside codeblock for full result
```

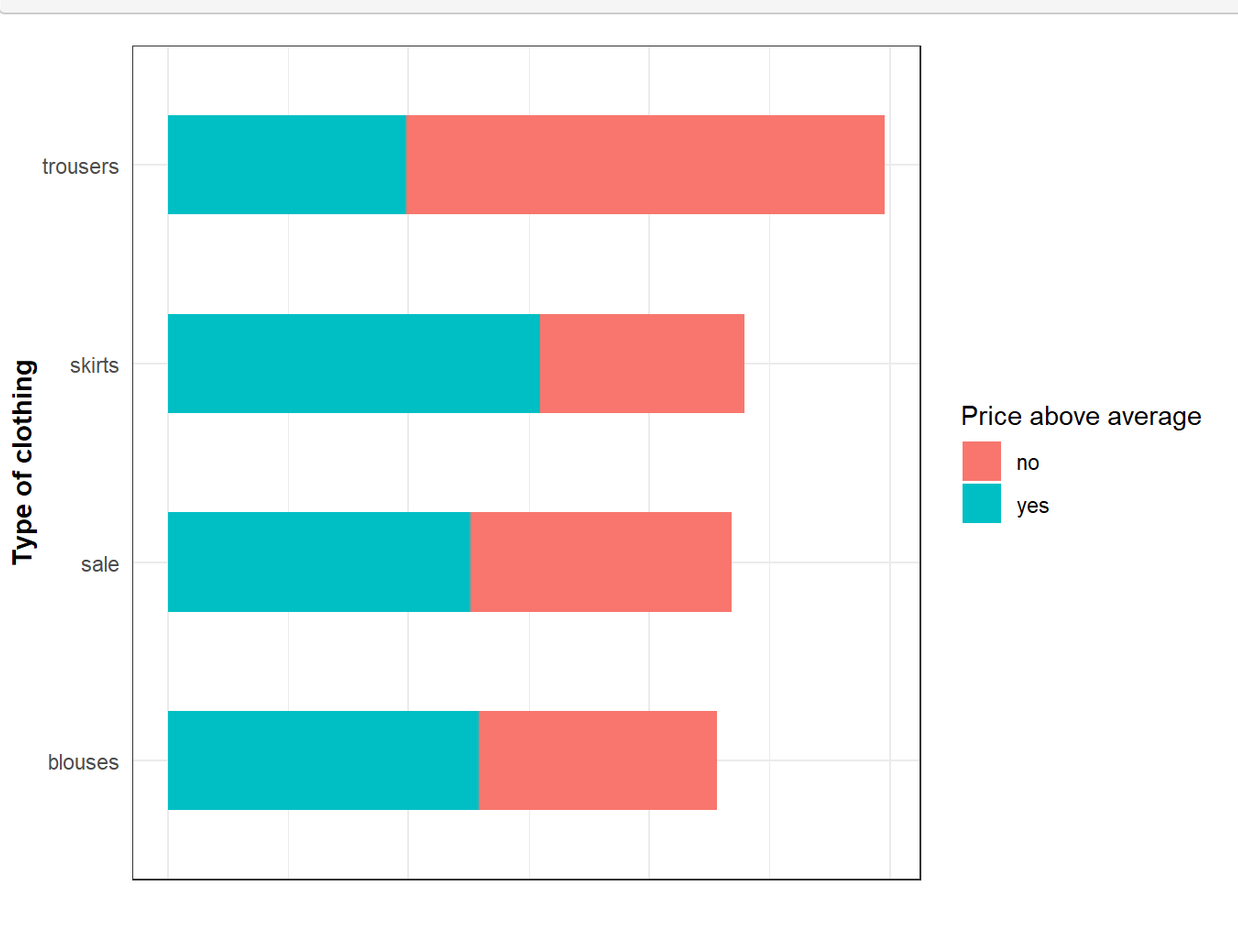
```
#The placement of advertisement along with type of photograph used is effecting the sales of the product
#en face which is a head shot or face-focused photography get's the maximum attention of the user
#The larger box tells us the total sales for that location of ad on the webpage and the division classifies
#based on the type of photography
#Top left seems to be the best position of the photo with the worst position being bottom right
data<-sampled%>%group_by(location,model.photography)%>%
  summarise (total_value=sum(price))


  treemap(data, #Your data frame object
          index=c("location","model.photography"),
          vSize = "total_value",
          vColor= "location",
          type="categorical",
          fontsize.labels=c(0,16),
          fontcolor.labels=c("white","Black"),
          fontface.labels=c(2,3),
          bg.labels=c("transparent"),
          align.labels=list(
            c("center", "center"),
            c("center", "center")
          ),
          overlap.labels=0.5,
          inflate.labels=F,


          palette = "RdGy",
          title="Impact of add position and type of photography on sales",
          fontsize.title = 14

  )
```




```
#Black and blue are consistently most bought by all the customers throughout the months which means launching mor
e styles
#in the color have higher chances of increasing sales.
#Colors like white,beige, are more popular in summer probably because of being cooler and absorbing less heat
#Besides some consistent fashion favorite must haves like brown,black,blue .The change climate affects the choice
of color
#which the company can keep in mind in launching products every month and have higher chances of success.

data_df<-sampled%>%group_by(month,colour)%>% summarise(total=n())

  x<- c("April","May",'June','July','August')
data_df <- data_df %>%  mutate(month = factor(month, levels = x)) %>% arrange(month)

ggplot(data_df, aes(x = data_df$month , y = data_df$total,
                      size = data_df$total,
                      color = colour))+
  theme(axis.title.y = element_blank() + theme(axis.title.x=element_blank())+

  geom_point(alpha = 0.7)+


  scale_size(range = c(0.8, 12), name = "Total clothes ")+
  ggtitle("Colour bought in every month")+

  # code to center the title which is left aligned
  # by default
  theme(plot.title = element_text(hjust = 0.5))
```

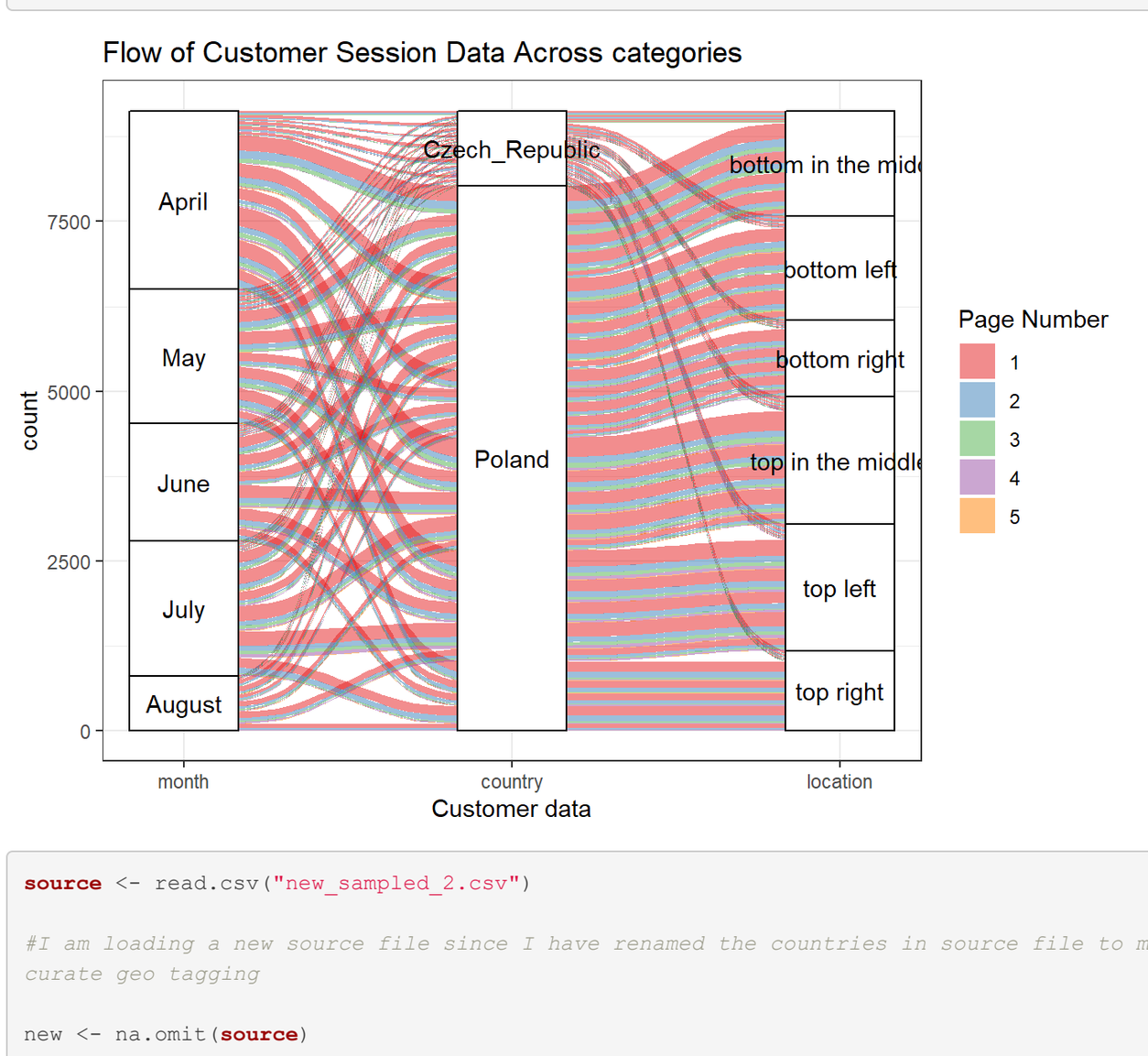



```
theme_set(theme_bw())
#67% of products in the category-trousers are sold below the average category price being a widely sold item.
#the individual profit margins are lower because of the demand increasing the overall profit
#skirts have the highest number of product above average price
#The surprising observation about e-commerce market manipulation is that the prices during "sale" are above usual
average and customers are illusioned
#into purchasing them.
# Data Prep
  # load data
data2<-sampled %>% group_by(page.1..main.category.,price.2) %>% summarise(total=n())
data2<-data2%>%  group_by(page.1..main.category.) %>%
  mutate(Percentage_above_average_price = 100*total/sum(total))
  data2$Percentage_above_average_price<- format(round(data2$Percentage_above_average_price, 2), nsmall = 2)
  data2$Percentage_above_average_price<-lapply(data2$Percentage_above_average_price, function(x)  paste(x,"%"))


# Diverging Barcharts
ggplot(data2, mapping =aes(x=data2$page.1..main.category., y=data2$total, label=Percentage_above_average_price))
 +
  geom_bar(stat='identity', mapping = aes(fill=price.2), width=.5) + xlab('Type of clothing')+ylab(element_blank
())+ theme(axis.text.x = element_blank(),axis.title.y = element_text(face="bold"),
                                                                   axis.ticks =element_blank())+ coord_flip
()+ scale_fill_discrete(name = "Price above average")
```




```
#Refer to the table data2 in environment section for exact percentage values for 'Above average price'
```

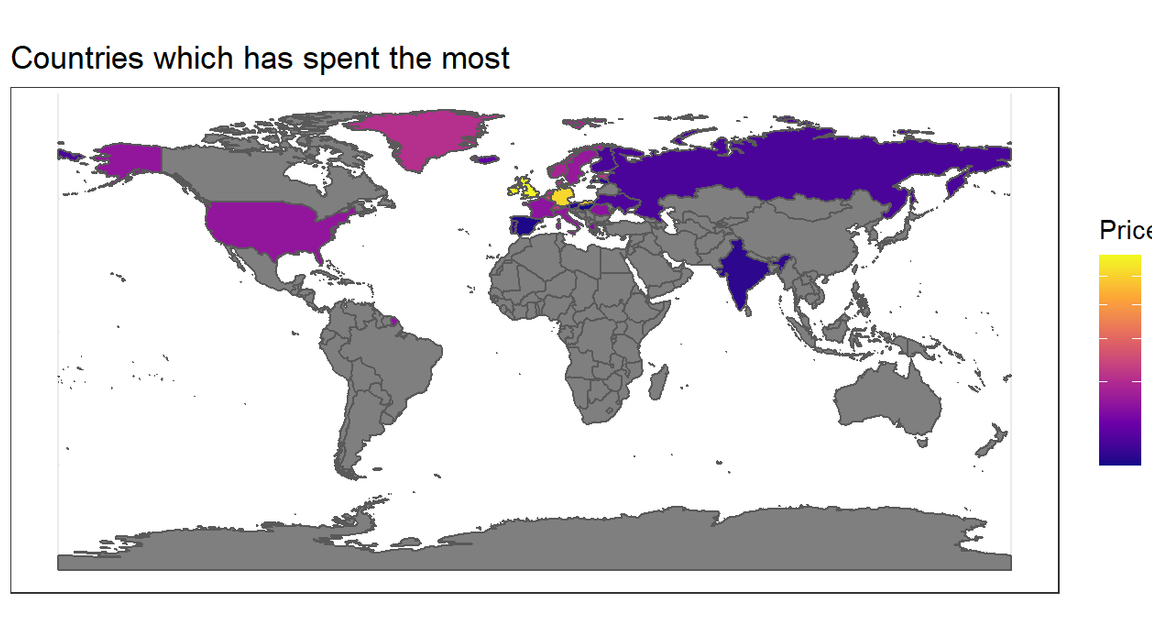
```
#Plot 3 Alluvial chart of order flow through various categories

al_df <-  sampled %>% filter(sampled$country %in% c('Poland','Czech_Republic')) %>% mutate(month =  factor(month,
levels = x)) %>% arrange(month) %>% group_by(month,country,location,page) %>% select('month','country','location'
,'page')  %>% summarise(count = n())



#This chart shows that most of the data has flown during Q2 of the year(Financial crisis reducing sales in the ne
xt quarter), and most of the people haven't been browsing past page 1 mainly.
#The highest clicks seem to be on the images top in the middle and top left of the webpage
#The image in the top right of page 2 seems to be of interest to people as it has large traffic
ggplot(data = al_df,
       aes(axis1 = month, axis2 = country, axis3 = location,
           y = count)) +
        scale_x_discrete(limits = c('month','country','location'), expand = c(.1, .05)) +
        xlab("Customer data") +
        geom_alluvium(aes(fill = as.factor(page))) +
        geom_stratum() + geom_text(stat = "stratum", aes(label = after_stat(stratum))) +   scale_fill_brewer(type
= "qual", palette = "Set1") + labs(fill = "Page Number") + ggtitle("Flow of Customer Session Data Across categori
es")
```




```
source <- read.csv("new_sampled_2.csv")

#I am loading a new source file since I have renamed the countries in source file to match the map package for ac
curate geo tagging

new <- na.omit(source)

theme_set(theme_bw())

#Creating the country_pr data frame by summarizing the total price spent by countries

country_pr <- new %>%
  group_by(country) %>%
  summarise(Tot_price = sum(price))

world <- ne_countries(scale = "medium", returnclass = "sf")
class(world)
```

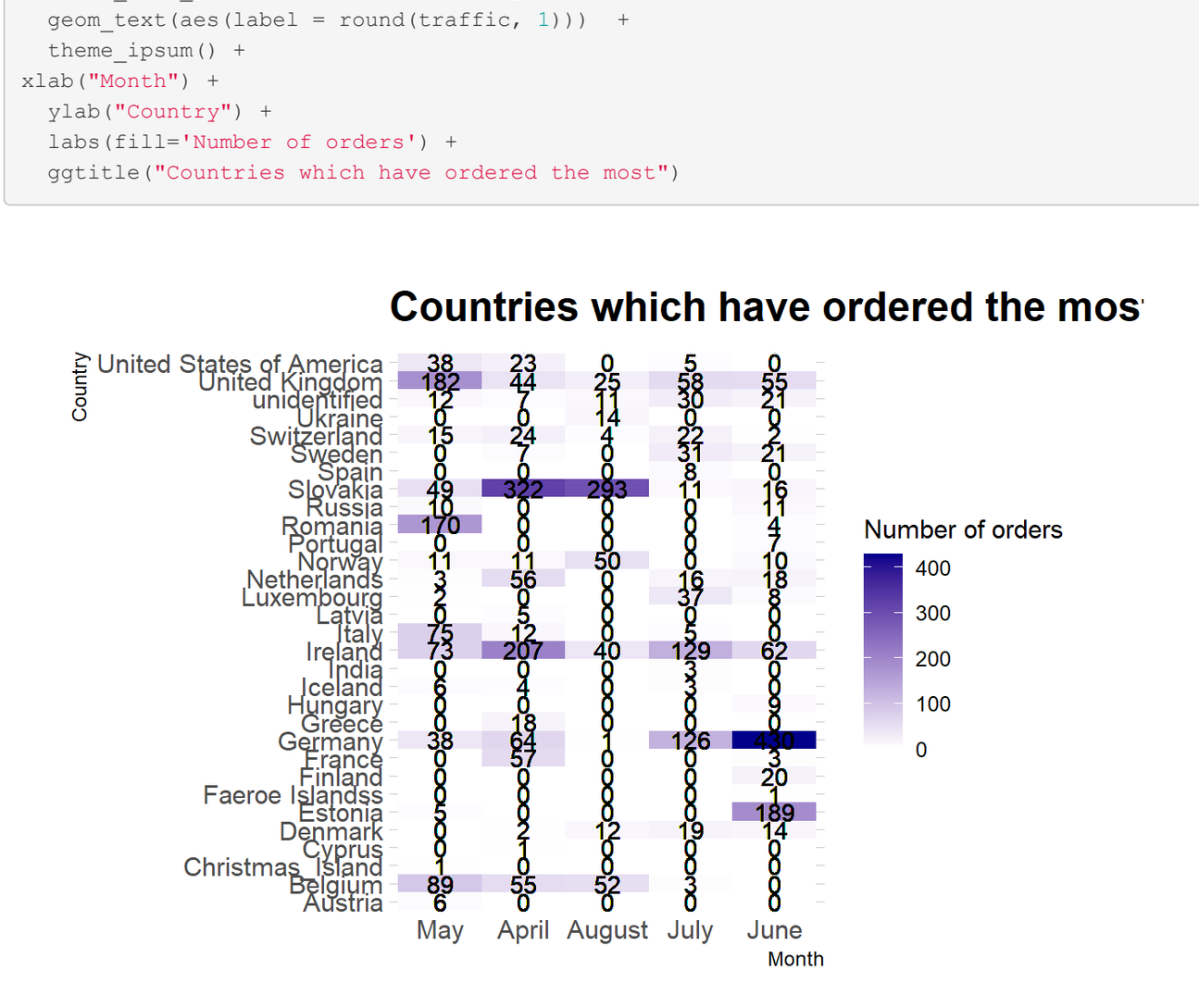
```
## [1] "sf"         "data.frame"
```

```
#Then I am joining the country_pr with the map data

new_join<-merge(world, country_pr, by.x = "sovereignt", by.y = "country", all.x = TRUE)

#Filtering domain names from countries and separating both
#Removed Poland, Czech Republic and Lithuania

n_join <- filter(new_join, sovereignt != "Poland"
      & sovereignt != "Czech Republic" & sovereignt != "Lithuania"
     & sovereignt != "com(*.com)"  & sovereignt != "net(*.net)"
     & sovereignt != "biz(*.biz)" & sovereignt != "int(*.int)")

#From the map we can see that UK has spent the most followed by UK, Ireland, Germany, Slovakia, Estonia, Romania,
Belgium, Greenland, USA, rest of Europe and India as the color turns dark blue.

ggplot(data = n_join) +
  geom_sf(aes(fill = Tot_price)) +
  scale_fill_viridis_c(option = "plasma", trans = "sqrt") +
labs(fill='Price$') +
  ggtitle("Countries which has spent the most")
```



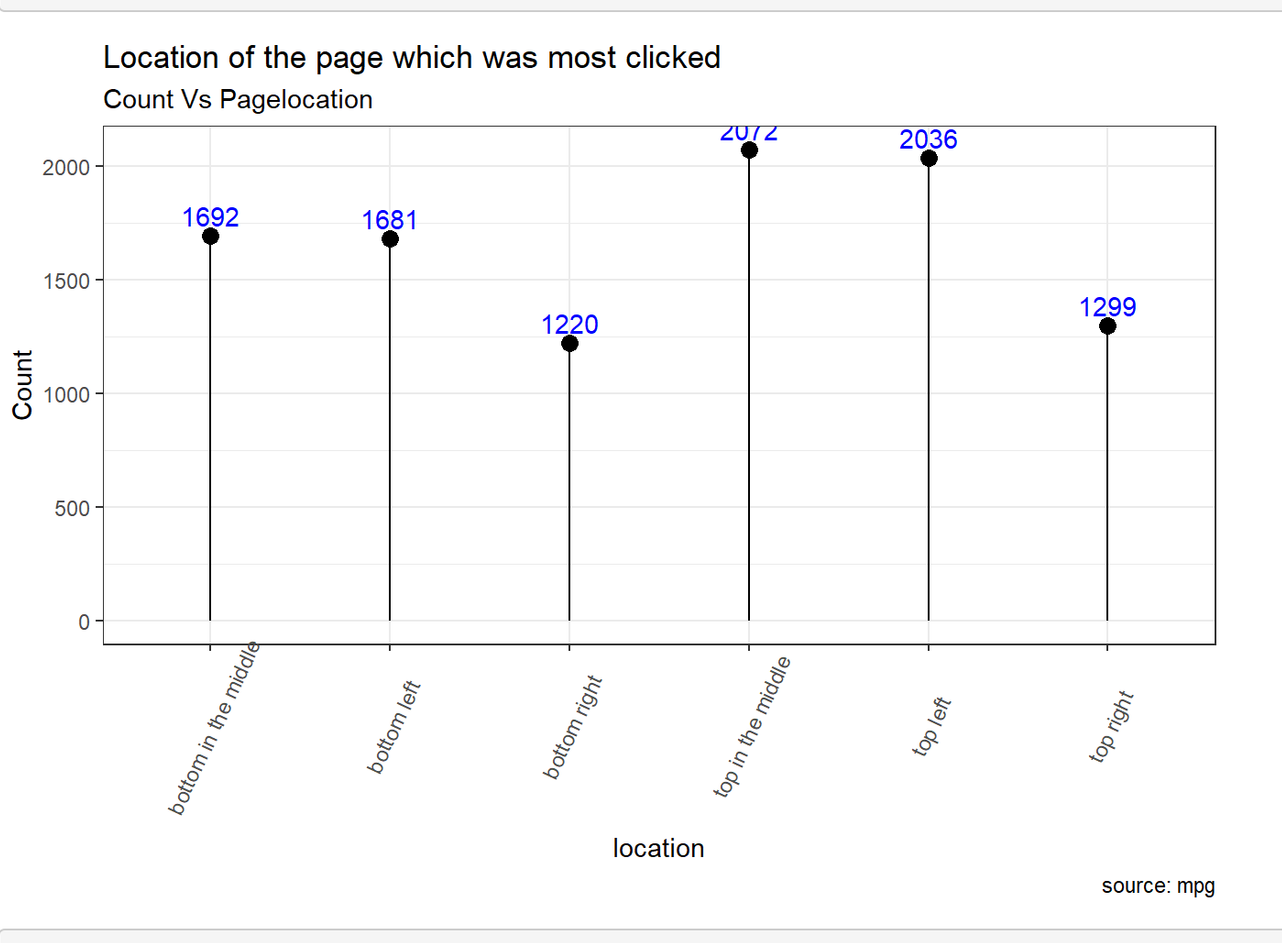

```
# This plot shows the countries which has spent the most
```

```
#Creating the new_data data frame by summarizing the number of orders spent by countries.

new_data <- new %>%
  group_by(country,month)%>%
  summarise(traffic = sum(order))


data <- expand.grid(month=new_data$month, country=new_data$country)

left_join<-merge(data, new_data, by = c("month", "country"), all.x = TRUE)

left_join[is.na(left_join)] = 0

#Filtering domain names from countries and separating both
#Removed Poland, Czech Republic and Lithuania
l_join <- filter(left_join, country != "Poland" & country != "Czech Republic" & country != "Lithuania"
                & country != "net(*.net)" & country != "com(*.com)" & country != "biz(*.biz)" & country != "int
(*.int)")

#After Poland, Czech Republic and Lithuania, people from Slovakia have made 322 orders followed by Switzerland wi
th 207 orders in April. In May United Kingdom has made the most number of orders. Germany has 430 orders in June.
And 129 orders by Ireland in July.  In August Again Slovakia has topped the list with 293 orders.

ggplot(l_join, aes(month, country, fill= traffic)) +
  geom_tile() +
  scale_fill_gradient(low="white", high="darkblue") +
  geom_text(aes(label = round(traffic, 1)))  +
  theme_ipsum() +
xlab("Month") +
  ylab("Country") +
  labs(fill='Number of orders') +
  ggtitle("Countries which have ordered the most")
```



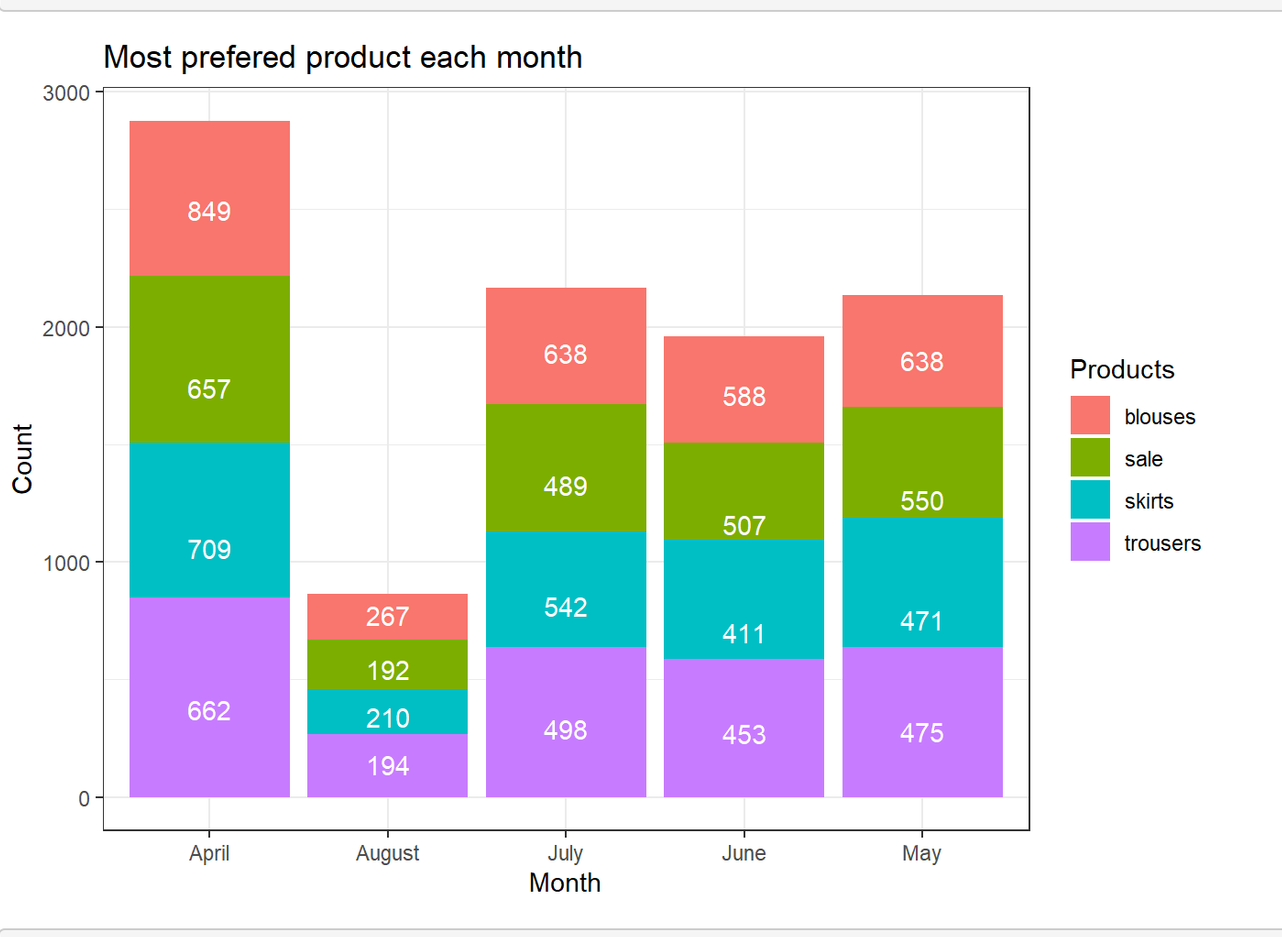

```
# this plot depicts the countries which has ordered the most
#This also depicts the internet traffic from each country

#The code block doesn't display plot properly, can be viewed outside codeblock for full result
```

```
#Creating the lol_chart data frame by summarizing the number of clicks on different parts of the webpage.

lol_chart <- new %>%
  group_by(location) %>%
  summarise(Count = n())

#The highest clicks seem to be on the images top in the middle and top left of the webpage. This shows that custo
mers have paid more attention to top in the middle and top left of the webpage.

ggplot(lol_chart, aes(x=location, y=Count)) +
  geom_point(size=3) +
  geom_segment(aes(x=location,
                   xend=location,
                   y=0,
                   yend=Count)) +
  labs(title="Location of the page which was most clicked",
       subtitle="Count Vs Pagelocation",
       caption="source: mpg") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  geom_text(aes(label = Count), vjust = -0.5, colour = "Blue")
```




```
#this plot shows the part of the webpage which was most clicked
```

```
#Creating the stacked_bar data frame by summarizing the number of clicks on different parts of the webpage.
stacked_bar <- new %>%
  group_by(month, page.1..main.category.) %>%
  summarise(Count = n()) %>%
  mutate(label_y = cumsum(Count) - 0.5 * Count)

#This stacked bar chart depicts the most purchased products; blouses are the most purchased product in April, Ma
y, June, July and August followed by skirts in April, July and August and other sale items in June and July

ggplot(stacked_bar, aes(fill=page.1..main.category., y=Count, x=month)) +
  geom_bar(position="stack", stat="identity") +
  xlab("Month") +
  ylab("Count") +
  labs(fill='Products') +
  ggtitle("Most prefered product each month") +
  geom_text(aes(y = label_y, label = Count), vjust = 0, colour = "white")
```




```
#this plot depicts the most purchased product of the month
```