E commerce analysis

Anupriya Kushwanshi

May 19, 2019

The aim of this project is to analyze users and page views dataset from an e-commerce company and provide recommendations to the product team to boost their sales strategy.

I also aim to find out if there are any issues with the dataset which needs attention.

Problems with the users' dataset:

. Duplicate rows in the users' table. I removed all the duplicates before loading the dataset in R.
. Some users have been registered as male as well as a female in the users' table. There could be several reasons behind this:
o There is a possibility that the website is asking the users' gender as soon as they enter. Some people don't want to give out their personal details hence, they may be choosing random values every time.
o The website may be tracking the users' behavior and assigning them genders. For example: if a female is browsing the items in the male section and moves back to the female section, the system may be assigning female-male-female gender to the same person in different sessions.
Cleaning the data was important in such a case. I assigned "NaN" to the gender of such rows and removed one row. By removing both the rows I would be losing information and by keeping both the rows, I would be compromising with data cleanliness.

Recommendation:
. My recommendation to the product manager is to use strong keywords and write correct information about the products. The paid and organic search are doing well in redirecting web traffic.
. Target the ads to the right user demographic to reduce churn rate.
. If the product price or shipping price is high and this is a reason for bouncing from the payment page, send offers to the users. For example: "Get 15% off on your total bill when you order within 15 minutes".
. Send shopping reminders and targeted ads to the users who left items in their cart because they may not have payment details handy at the time of check out.
. I assume the website accepts cookies. It would be good to check if the user searched for any website for price comparison.

– Loading the cleaned dataset

```
page_view <- read.csv("page_views.csv", header = T)
users <- read.csv("users_removed_duplicate_gender_reversal.csv", header = T)
```

– Exploring the data

```
head(page_view)
```

```
##   user_id homepage pymt pymt_confirmation search_page
## 1  144912        1    0                 0           0
## 2   60659        1    0                 0           0
## 3  140860        1    0                 0           1
## 4  206992        1    0                 0           1
## 5  320259        1    0                 0           0
## 6   17641        1    0                 0           0
```

```
users <- na.omit(users)
head(users)
```

```
##   user_id      date    sex  device         origin
## 1       0 4/19/2015 Female Desktop    paid_search
## 3       1 3/23/2015    NaN  Mobile    paid_search
## 5       4  1/6/2015    NaN Desktop organic_search
## 6       6 4/10/2015 Female  Mobile    paid_search
## 8       7 3/19/2015    NaN Desktop    paid_search
## 9       8 4/28/2015 Female Desktop        unknown
```

Checking null values for page view

```
colSums(is.na(page_view))
```

```
##           user_id          homepage              pymt pymt_confirmation
##                 0                 0                 0                 0
##       search_page
##                 0
```

Checking null values for user

```
if(ncol(users==6)){
  users <- users[,-6]
}
colSums(is.na(users))
```

```
## user_id    date     sex  device  origin
##       0       0       0       0       0
```

I reordered the page_view table as per the flow of pages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## 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
```

```
nrow(page_view)
```

```
## [1] 202653
```

```
page_view <- select(page_view, "user_id", "homepage", "search_page", everything())
```

A page load does not necessarily represent a unique user. The same user may visit the same page twice in one session or even several times in the given time duration. Hence, I created a dataframe of unique users for further analysis by creating pivot table in MS Excel and loading in RStudio.

```
unique_user <- page_view %>% distinct(user_id)
count(distinct(unique_user))
```

```
## # A tibble: 1 x 1
##        n
##    <int>
## 1 138191
```

I created a table grouped by user id and loaded the same for analysis.

```
uniq_pg_view <- read.csv("unique_page_views.csv")
dim(uniq_pg_view)
```

```
## [1] 138191      5
```

```
names(uniq_pg_view) <- c("user_id", "pymt_confirmation","search_page", "homepage", "pymt")
uniq_pg_view <- select(uniq_pg_view, "user_id", "homepage", "search_page","pymt", "pymt_confirma
tion")
uniq_pg_view <- na.omit(uniq_pg_view)
head(arrange(uniq_pg_view,desc(homepage)))
```

```
##   user_id homepage search_page pymt pymt_confirmation
## 1   16439        8           7    2                 0
## 2  187433        8           4    0                 0
## 3  232609        8           4    2                 0
## 4  309187        8           2    0                 0
## 5    4040        7           2    0                 0
## 6    8124        7           5    0                 0
```

I saw that some users may visit homepage 8 times, search page 7 times, payment page 2 times but still doesn't buy a product.

```
head(arrange(uniq_pg_view,desc(pymt_confirmation)))
```

```
##   user_id homepage search_page pymt pymt_confirmation
## 1     401        2           2    2                 2
## 2     495        2           2    2                 2
## 3   12935        3           2    2                 2
## 4   14884        3           2    2                 2
## 5   17519        3           3    2                 2
## 6   19412        2           2    2                 2
```

Whereas some users visit every page once and still purchase the product.

```
homepage <- sum(uniq_pg_view$homepage,na.rm=TRUE)
search_page <- sum(uniq_pg_view$search_page,na.rm=TRUE)
pymt <-  sum(uniq_pg_view$pymt,na.rm=TRUE)
pymt_confirm <- sum(uniq_pg_view$pymt_confirmation,na.rm=TRUE)
df_pg_visit <- cbind(homepage,search_page,pymt,pymt_confirm)
df_pg_visit <- as.data.frame(df_pg_visit)
df_pg_visit
```

```
##   homepage search_page  pymt pymt_confirm
## 1   202653      101105 13459          990
```

Remodeling the data for a better view, understanding, and visualization.

```
library(tidyverse)
```

```
## -- Attaching packages -----------------------------------------------------------------------
------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1     v readr   1.2.1
## v tibble  2.0.1     v purrr   0.2.5
## v tidyr   0.8.2     v stringr 1.3.1
## v ggplot2 3.1.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
## -- Conflicts ----------------------------------------------------------------------------
- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df_pg_visit <- gather(df_pg_visit, Pages, Views)
df_pg_visit
```

```
##          Pages  Views
## 1     homepage 202653
## 2  search_page 101105
## 3         pymt  13459
## 4 pymt_confirm    990
```

Visualising the number of visit for each page

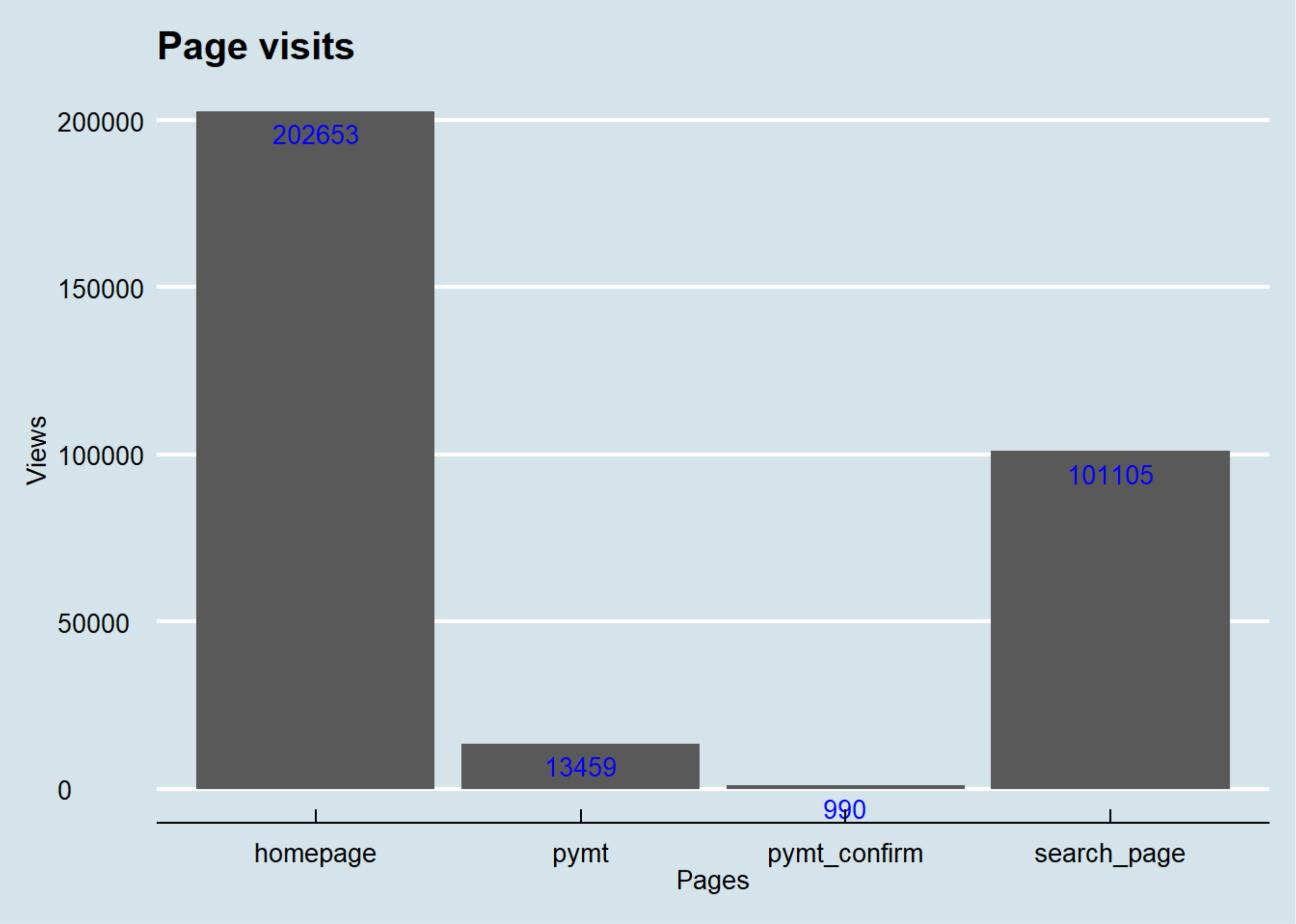
```
library(ggplot2)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.5.3
```

```
100*(990/202653)
```

```
## [1] 0.4885198
```

```
ggplot(data=df_pg_visit, aes(x=Pages, y=Views)) +
geom_bar(stat="identity")+ geom_text(aes(label=Views), vjust=1.6, color="blue", size=3.5)+
ggtitle("Page visits")+ ggthemes::theme_economist()
```



The overall conversion rate is 0.48%.According to wordstream.com “Across industries, the average landing page conversion rate was 2.35%, yet the top 25% are converting at 5.31% or higher. Ideally, you want to break into the top 10% - these are the landing pages with conversion rates of 11.45% or higher”

A high churn can can be a sign of a number of other things:

. The content is not relevant to the keywords or other acquisition methods used to get visitors to the site.
. The content is not engaging readers. - Is the language and sentiment of the content correct?
. Are the ads being targeted to right audience?
. Page load time is too much - I can talk more about this by analysing site speed report
. The user may refresh a page multiple times, which would also be counted as a pageview - Is it content loading issue?
. The website is being hit with some malicious attacks or even a hack. The site may be infected with a form of malware keeping users out.

To get more insights about the conversion funnel, I analysed the percentage of users actually making a purchase after going to payment page.

```
purchase <- (990/13459)*100
 purchase
```

```
## [1] 7.355673
```

Only 7.35% users are actually buying the product after visiting payment page. The reasons behind this may be:

. Item not shippable in users’ area
. Users’ preferred mode of payment is not supported
. Users’ coupons/referral code/promo code not applicable
Product price or shipping price may be very high - send offers like if you order within 1 hour get 15% off.
Users may not have payment details at the time of check out - send reminders targeted ads

```
library(dplyr)
nrow(uniq_pg_view)
```

```
## [1] 138191
```

```
uniq_pg_view$user_id <- as.integer(uniq_pg_view$user_id)
user_page_join <- inner_join(users,uniq_pg_view, by="user_id")
if(ncol(user_page_join==10)){
  user_page_join <- user_page_join[,-10]
}
colSums(is.na(user_page_join))
```

```
##           user_id              date               sex            device
##                 0                 0                 0                 0
##            origin          homepage       search_page              pymt
##                 0                 0                 0                 0
## pymt_confirmation
##                 0
```

```
dim(user_page_join)
```

```
## [1] 151949      9
```

```
colSums(is.na(user_page_join))
```

```
##           user_id              date               sex            device
##                 0                 0                 0                 0
##            origin          homepage       search_page              pymt
##                 0                 0                 0                 0
## pymt_confirmation
##                 0
```

```
user_page_join <- na.omit(user_page_join)
dim(user_page_join)
```

```
## [1] 151949      9
```

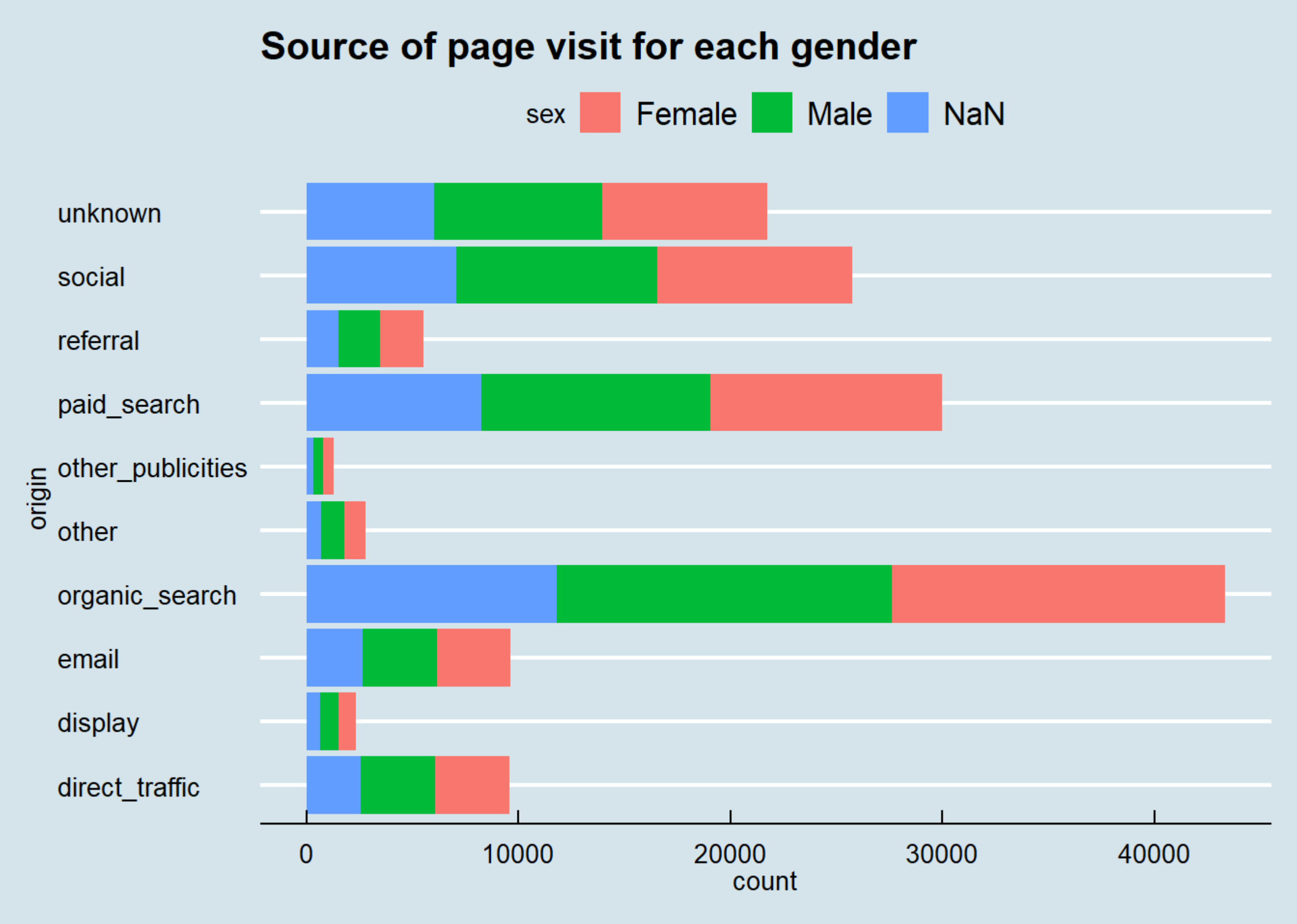
```
colnames(user_page_join)
```

```
## [1] "user_id"           "date"              "sex"
## [4] "device"            "origin"            "homepage"
## [7] "search_page"       "pymt"              "pymt_confirmation"
```

```
head(arrange(user_page_join, desc(pymt_confirmation)))
```

```
##   user_id      date    sex  device         origin homepage search_page
## 1     401 2/16/2015 Female  Mobile         social        2           2
## 2     495 2/21/2015   Male  Mobile        unknown        2           2
## 3   12935 2/22/2015    NaN Desktop organic_search        3           2
## 4   14884 3/29/2015    NaN  Mobile          email        3           2
## 5   17519  4/9/2015    NaN Desktop       referral        3           3
## 6   19412 1/26/2015   Male  Mobile          email        2           2
##   pymt pymt_confirmation
## 1    2                 2
## 2    2                 2
## 3    2                 2
## 4    2                 2
## 5    2                 2
## 6    2                 2
```

```
library(ggplot2)
ggplot(user_page_join, aes(x=origin, fill=sex)) + geom_bar()+ggtitle("Source of page visit for e
ach gender")+ ggthemes::theme_economist()+coord_flip()
```

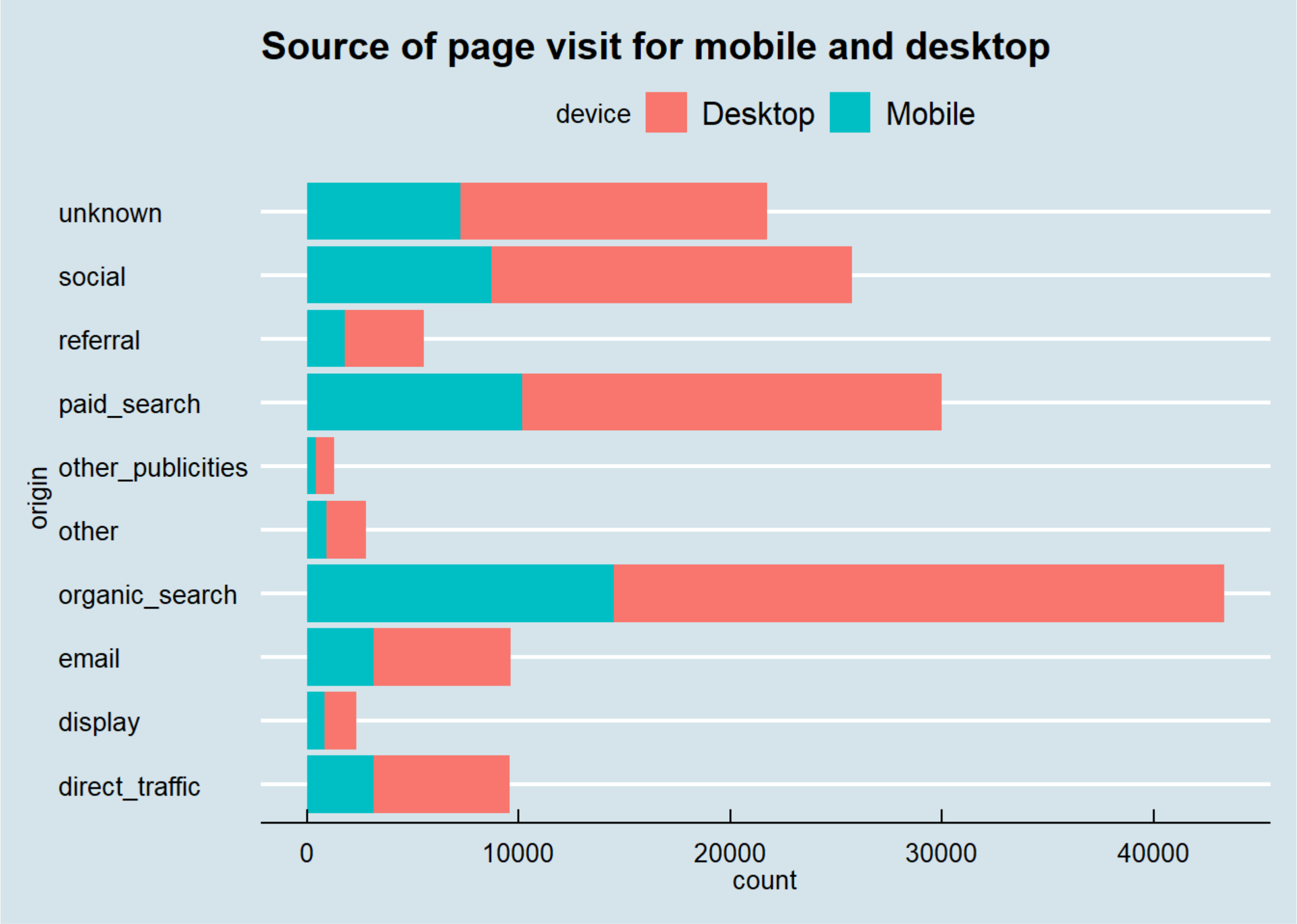


Most of the views on the website are through organic and paid search which is a good thing. This means there is good brand familiarity. But at the same time, conversion rate is so low. The potential reasons for this could be:

. The company has used excessive keywords that results in redirection of users on the website but these keywords are wrong or irrelevant.
. People don’t understand the product. Users come to the website anticipating something else that is irrelevant.

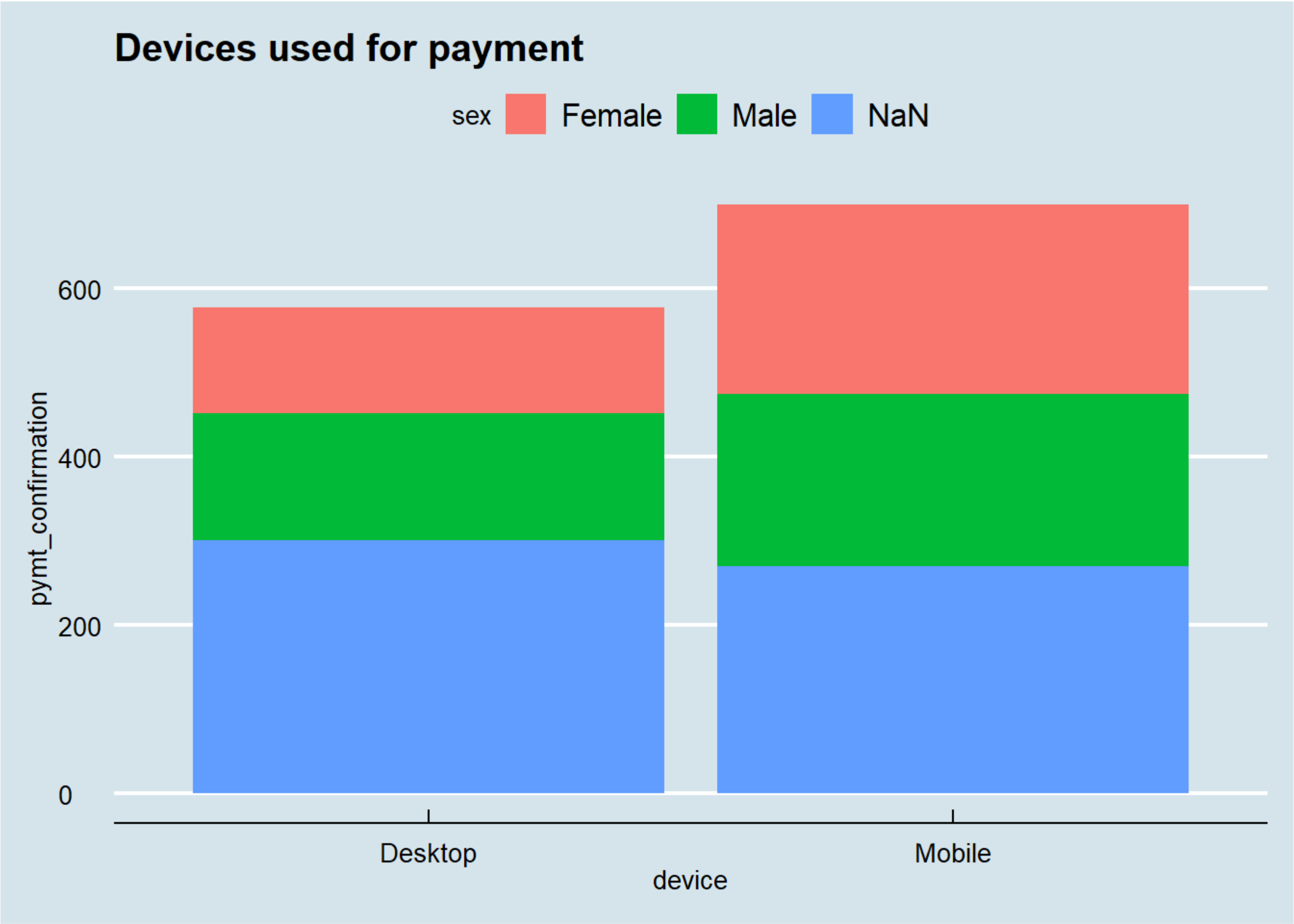
My recommendation to the product manager is to use strong keywords and write correct information about the products. The paid and organic search are doing well in redirecting web traffic.

```
ggplot(user_page_join, aes(x=origin, fill=device)) + geom_bar()+ggtitle("Source of page visit fo
r mobile and desktop")+ ggthemes::theme_economist()+coord_flip()
```



```
ggplot(user_page_join, aes(x=device, y = pymt_confirmation, fill = sex)) + geom_histogram(stat =
"identity")+ggtitle("Devices used for payment")+ ggthemes::theme_economist()
```

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



```
summary(user_page_join$device[user_page_join$pymt_confirmation!=0])
```

```
##         Desktop  Mobile
##       0     520     609
```

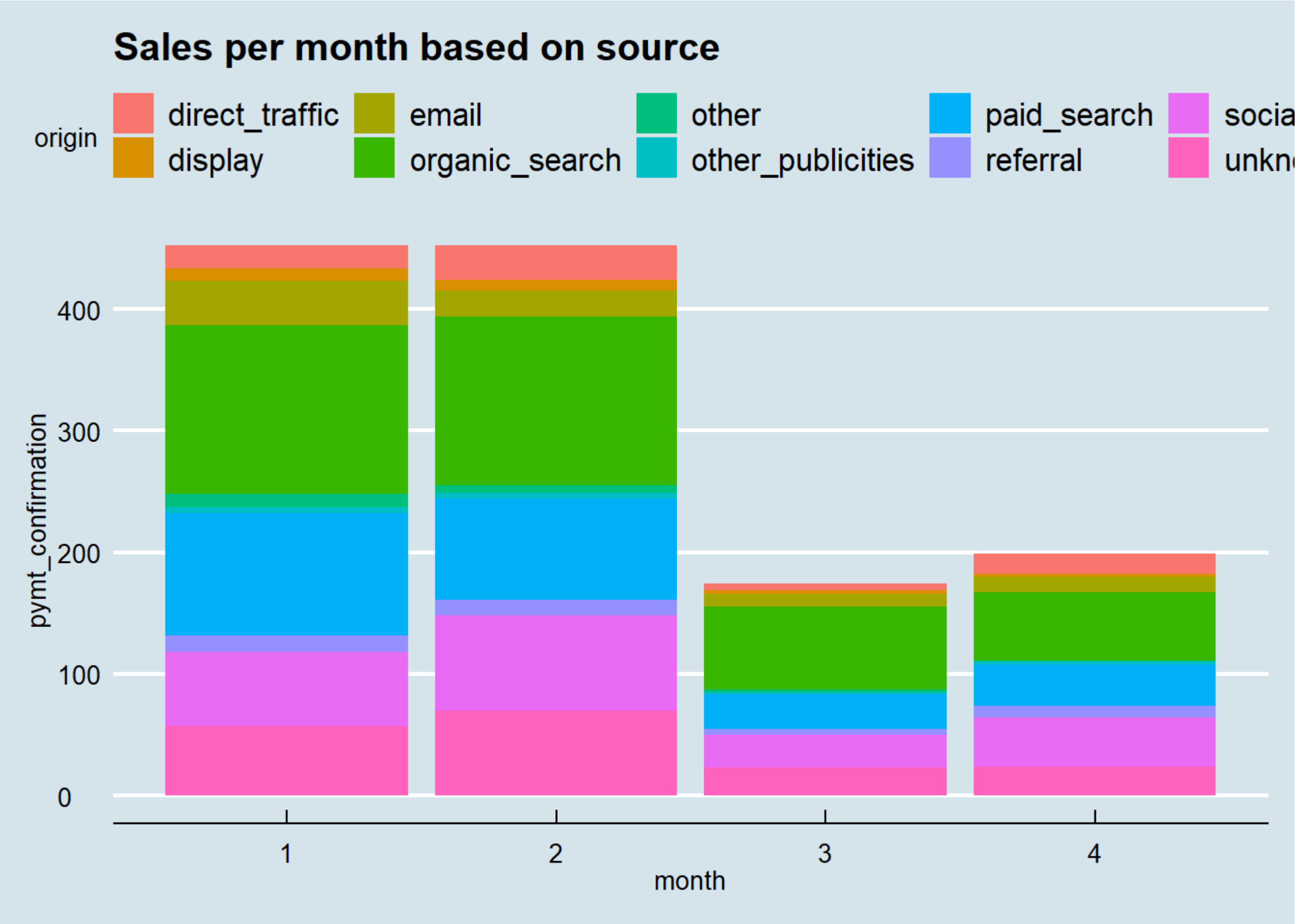
```
users_delimited <- read.csv("users_delimited.csv")
users_delimited <- na.omit(users_delimited)
user_page_join_delim <- inner_join(uniq_pg_view, users_delimited, "user_id"="user_id")
```

```
## Joining, by = "user_id"
```

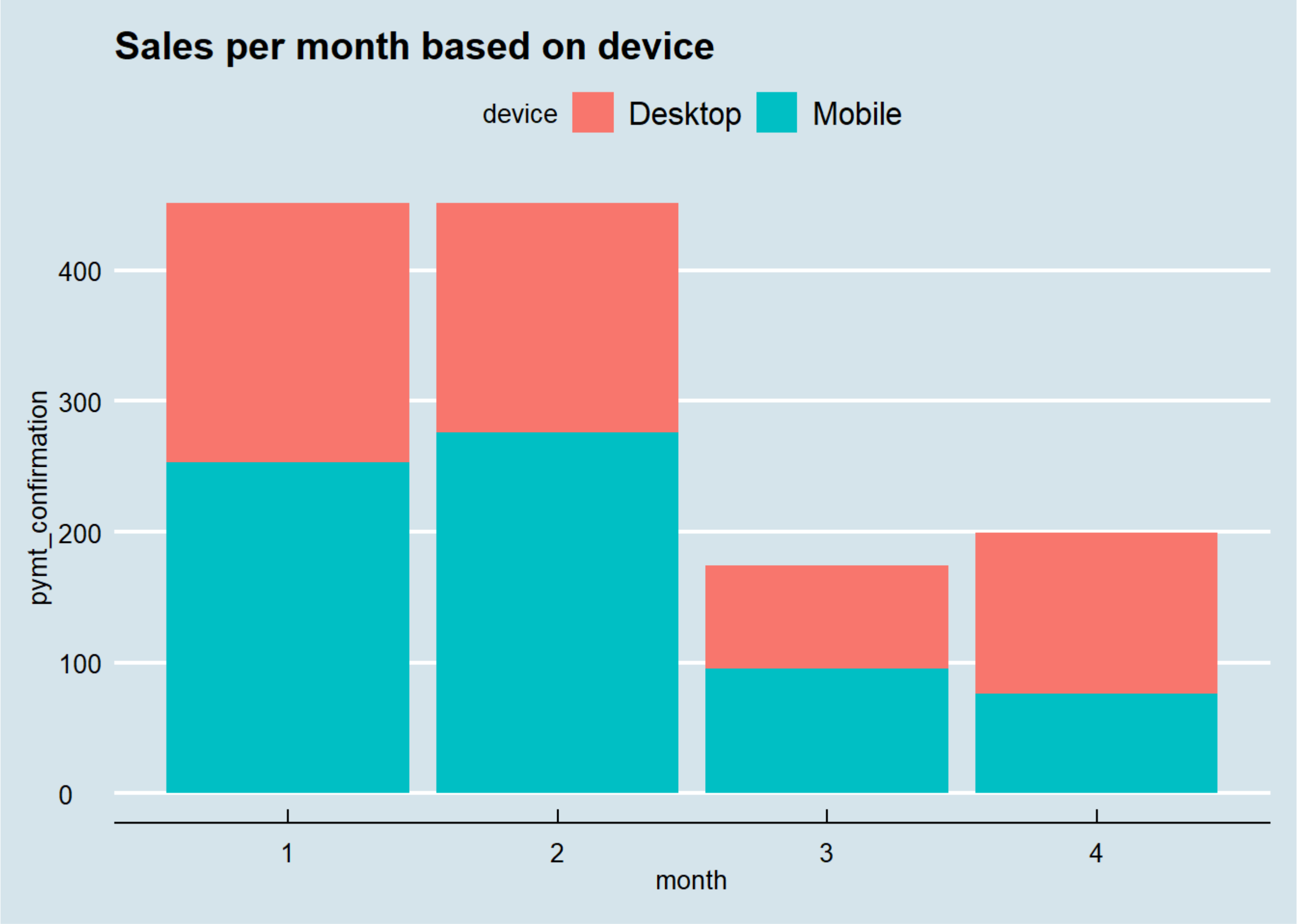
```
head(select(user_page_join_delim, "day","month","year", "user_id", everything()))
```

```
##   day month year user_id homepage search_page pymt pymt_confirmation
## 1  19     4 2015       0        1           0    0                 0
## 2  23     3 2015       1        2           1    0                 0
## 3   6     1 2015       4        2           1    0                 0
## 4  10     4 2015       6        1           0    0                 0
## 5  19     3 2015       7        2           0    0                 0
## 6  28     4 2015       8        1           1    0                 0
##      sex  device         origin
## 1 Female Desktop    paid_search
## 2    NaN  Mobile    paid_search
## 3    NaN Desktop organic_search
## 4 Female  Mobile    paid_search
## 5    NaN Desktop    paid_search
## 6 Female Desktop        unknown
```

```
ggplot(user_page_join_delim, aes(x=month, y = pymt_confirmation, fill=origin)) + geom_bar(stat =
"identity")+ggtitle("Sales per month based on source")+ ggthemes::theme_economist()
```



```
ggplot(user_page_join_delim, aes(x=month, y = pymt_confirmation, fill=device)) + geom_bar(stat =
"identity")+ggtitle("Sales per month based on device")+ ggthemes::theme_economist()
```



~~~
```r
ggplot(user_page_join_delim, aes(x=month, y = pymt_confirmation, fill=sex)) + geom_bar(stat = "identity")+ggtitle("Sales per month based on gender")+ ggthemes::theme_economist()
```

<img src="Nurx_Code_Challenge_files/figure-html/unnamed-chunk-22-1.png" width="672" />
~~~