Customer Acquisition Strategy Analysis – Red Ventures

Mitali Soni, Lkshme Srada Srinivas

```
setwd("~/Downloads")
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

Understanding the business problem is the core of any analysis. This can be done with the help of Exploratory Data Analysis and Visualization techniques. Through analysis of the Red Ventures dataset, we explore patterns necessary to better understand the variables and the relationship between them.

As a first step, we checked the dimensions and the summary of the dataset to understand the volume and basic statistics of the data.

```
# EDA
redventuredf<-read_excel("Red Ventures Telecom Chapter 4.xlsx")
redventuredf<-clean_names(redventuredf)

names(redventuredf)
```

```
##  [1] "session_id"           "session_start_time"   "browser_name"
##  [4] "sessions"             "total_orders"         "cart_order"
##  [7] "phone_order"          "city"                 "connection_speed"
```

```
## [10] "country"              "device_type"          "distinct_page_views"
## [13] "isp_name"             "landing_page_raw"     "manufacturer"
## [16] "metro_name"           "os_name"              "order_monthly_charge"
## [19] "state"                "zip_code"             "traffic_source"
```

```
str(redventuredf)
```

```
## tibble [20,256 x 21] (S3: tbl_df/tbl/data.frame)
##  $ session_id          : chr [1:20256] "03b1d69a-6463-40a3-b67b-a434d02d673b" "03b1d69a-6463-40a3-b6
##  $ session_start_time  : POSIXct[1:20256], format: "2018-02-28 13:49:47" "2018-02-28 13:49:47" ...
##  $ browser_name        : chr [1:20256] "Chrome" "Chrome" "Safari" "Chrome" ...
##  $ sessions            : num [1:20256] 1 1 1 1 1 1 1 1 1 1 ...
##  $ total_orders        : num [1:20256] 1 1 1 1 1 1 1 1 1 1 ...
##  $ cart_order          : num [1:20256] NA NA NA NA NA NA NA NA NA NA ...
##  $ phone_order         : num [1:20256] 1 1 1 1 1 1 1 1 1 1 ...
##  $ city                : chr [1:20256] "Glendale" "Glendale" "Riverside" "San Antonio" ...
##  $ connection_speed    : chr [1:20256] "Broadband" "Broadband" "Xdsl" "Cable" ...
##  $ country             : chr [1:20256] "US" "US" "US" "US" ...
##  $ device_type         : chr [1:20256] "Desktop" "Desktop" "Desktop" "Desktop" ...
##  $ distinct_page_views : num [1:20256] 3 3 2 1 2 2 2 2 2 2 ...
##  $ isp_name            : chr [1:20256] "Verizon Business" "Verizon Business" "Att Internet Services"
##  $ landing_page_raw    : chr [1:20256] "https://phonecompanyx.com/" "https://phonecompanyx.com/" "ht
##  $ manufacturer        : chr [1:20256] "Unknown" "Unknown" "Apple" "Unknown" ...
##  $ metro_name          : chr [1:20256] "Phoenix (Prescott)" "Phoenix (Prescott)" "Los Angeles" "San A
##  $ os_name             : chr [1:20256] "Windows 10" "Windows 10" "OS X" "Windows 7" ...
##  $ order_monthly_charge: num [1:20256] 60 60 80 25 25 ...
##  $ state               : chr [1:20256] "AZ" "AZ" "CA" "TX" ...
##  $ zip_code            : num [1:20256] 85301 85301 92507 78203 90212 ...
##  $ traffic_source      : chr [1:20256] "Natural Search" "Natural Search" "Direct Entry" "Direct Entry
```

```
summary(redventuredf)
```

```
##   session_id        session_start_time               browser_name
##  Length:20256       Min.   :2018-02-17 18:00:21.00   Length:20256
##  Class :character   1st Qu.:2018-02-21 14:19:38.25   Class :character
##  Mode  :character   Median :2018-02-25 12:49:09.50   Mode  :character
##                     Mean   :2018-02-25 06:47:37.36
##                     3rd Qu.:2018-02-28 19:21:27.75
##                     Max.   :2018-03-04 16:27:32.00
##
##     sessions  total_orders     cart_order     phone_order        city
##  Min.   :1   Min.   :1       Min.   :1       Min.   :1       Length:20256
##  1st Qu.:1   1st Qu.:1       1st Qu.:1       1st Qu.:1       Class :character
##  Median :1   Median :1       Median :1       Median :1       Mode  :character
##  Mean   :1   Mean   :1       Mean   :1       Mean   :1
##  3rd Qu.:1   3rd Qu.:1       3rd Qu.:1       3rd Qu.:1
##  Max.   :1   Max.   :1       Max.   :1       Max.   :1
##              NA's   :19593   NA's   :19951   NA's   :19898
##  connection_speed     country           device_type        distinct_page_views
##  Length:20256       Length:20256       Length:20256       Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.000
##                                                           Mean   :1.153
```

```
##                                                        3rd Qu.:1.000
##                                                        Max.   :6.000
##
##    isp_name         landing_page_raw   manufacturer        metro_name
##  Length:20256       Length:20256       Length:20256       Length:20256
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    os_name          order_monthly_charge    state              zip_code
##  Length:20256       Min.   : 24.99       Length:20256       Min.   : 1027
##  Class :character   1st Qu.: 24.99       Class :character   1st Qu.:33542
##  Mode  :character   Median : 24.99       Mode  :character   Median :56063
##                     Mean   : 45.36                          Mean   :57201
##                     3rd Qu.: 59.99                          3rd Qu.:90045
##                     Max.   :117.99                          Max.   :99743
##                     NA's   :19593
##  traffic_source
##  Length:20256
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# finding missing values

colSums(is.na(redventuredf))
```

```
##           session_id   session_start_time         browser_name
##                    0                    0                    0
##             sessions         total_orders           cart_order
##                    0                19593                19951
##          phone_order                 city     connection_speed
##                19898                    0                    0
##              country          device_type  distinct_page_views
##                    0                    0                    0
##             isp_name     landing_page_raw         manufacturer
##                    0                    0                    0
##           metro_name              os_name order_monthly_charge
##                    0                    0                19593
##                state             zip_code       traffic_source
##                    0                    0                    0
```

In our EDA process, we discovered that some columns have missing values and there were inconsistencies in the capitalization of character-type entries. 21 variables are present, with 20256 values in each. A significant observation was that the ratio of actual conversions/orders to total number of sessions is 3.3%, which formed the premise for our analysis in terms of improving customer acquisition. Some of our findings are: California has the highest number of sessions, while traffic is more through the 'direct entry' source. and desktops are the most used device type. We explore relationships between variable further after cleaning the data.

#3. Is there any cleaning or transformation required? If so perform the necessary operations. In our efforts to clean the data, we replaced missing values with 0 and converted all character-type values to lowercase for uniformity. We also checked the number of unique values in each column in order to identify variables of interest. We converted these variable columns to factor type to aid us in the analysis, while also extracting days from the Session Start Time column, which later helped us check traffic on each day of the week. Our variables of interest are Sessions, Total Orders, States, Traffic Source, Day of the week, Device Types and Browser Names.

```
#Data Cleaning
# replacing na in cart order, phone order, Total orders and Order Monthly charge with 0

redventuredf<-redventuredf %>% replace_na(list(cart_order = 0, phone_order = 0,
order_monthly_charge = 0, total_orders = 0))
colSums(is.na(redventuredf))
```

```
##            session_id   session_start_time         browser_name
##                     0                    0                    0
##              sessions         total_orders           cart_order
##                     0                    0                    0
##          phone_order                 city     connection_speed
##                     0                    0                    0
##              country          device_type  distinct_page_views
##                     0                    0                    0
##             isp_name     landing_page_raw         manufacturer
##                     0                    0                    0
##           metro_name              os_name order_monthly_charge
##                     0                    0                    0
##                state             zip_code       traffic_source
##                     0                    0                    0
```

```
# Converting values in categorical columns to lower case

redventuredf<-as.data.frame(lapply(redventuredf,function(x) if (is.character(x)) tolower(x) else x))

# Adding a day column for day of the week

redventuredf<-redventuredf %>% mutate(day=wday(session_start_time, label=TRUE, abbr=TRUE))%>%arrange(da

# Converting columns of interest to factors

redventuredf[,c(3,8,10:11,15:17,19:22)]<-lapply(redventuredf[,c(3,8,10:11,15:17,19:22)],factor)


# Structure Summary

dim(redventuredf)
```

```
## [1] 20256    22
```

```
summary(redventuredf)
```

```
##   session_id        session_start_time                          browser_name
```

```
##  Length:20256       Min.   :2018-02-17 18:00:21.00   chrome           :7662
##  Class :character   1st Qu.:2018-02-21 14:19:38.25   safari           :4777
##  Mode  :character   Median :2018-02-25 12:49:09.50   internet explorer:2121
##                     Mean   :2018-02-25 06:47:37.36   chrome mobile    :1982
##                     3rd Qu.:2018-02-28 19:21:27.75   firefox          :1607
##                     Max.   :2018-03-04 16:27:32.00   edge             :1507
##                                                      (Other)          : 600
##     sessions  total_orders       cart_order       phone_order
##  Min.   :1   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:1   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :1   Mean   :0.03273   Mean   :0.01506   Mean   :0.01767
##  3rd Qu.:1   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##           city       connection_speed   country     device_type
##  tampa      :  579   Length:20256       us:20256   desktop:14271
##  ft wayne   :  330   Class :character              mobile : 4827
##  los angeles:  329   Mode  :character              tablet : 1158
##  dallas     :  257
##  reading    :  242
##  plano      :  234
##  (Other)    :18285
##  distinct_page_views   isp_name         landing_page_raw      manufacturer
##  Min.   :1.000       Length:20256       Length:20256       unknown  :10785
##  1st Qu.:1.000       Class :character   Class :character   apple    : 5665
##  Median :1.000       Mode  :character   Mode  :character   microsoft: 1511
##  Mean   :1.153                                             samsung  : 1371
##  3rd Qu.:1.000                                             lg       :  362
##  Max.   :6.000                                             motorola :  154
##                                                            (Other)  :  408
##                      metro_name           os_name     order_monthly_charge
##  los angeles              : 3193   windows 10 :6668   Min.   :  0.000
##  tampa-st. pete (sarasota): 2014   windows 7  :4095   1st Qu.:  0.000
##  dallas-ft. worth         : 1633   ios        :3624   Median :  0.000
##  hartford & new haven     :  840   android    :2348   Mean   :  1.485
##  new york                 :  792   os x       :2235   3rd Qu.:  0.000
##  seattle-tacoma           :  734   windows 8.1: 675   Max.   :117.990
##  (Other)                  :11050   (Other)    : 611
##      state         zip_code            traffic_source   day
##  ca     :3876   19606  :  214   direct entry  :12251   Sun:2734
##  fl     :2417   18106  :  132   natural search: 2612   Mon:2887
##  tx     :2371   46825  :  121   other         :  129   Tue:3134
##  ct     :1071   76244  :   97   paid search   : 5264   Wed:3078
##  in     : 918   98052  :   78                          Thu:3056
##  il     : 904   92584  :   74                          Fri:2929
##  (Other):8699   (Other):19540                          Sat:2438
```

```
str(redventuredf)
```

```
## 'data.frame':    20256 obs. of  22 variables:
##  $ session_id          : chr  "178df541-6cdd-4dc4-af9e-f85b8ae5b334" "20b785f2-920a-4222-b517-407ece
##  $ session_start_time  : POSIXct, format: "2018-02-18 18:47:47" "2018-02-25 17:50:04" ...
##  $ browser_name        : Factor w/ 20 levels "amazon silk",..: 3 8 17 17 3 3 17 9 3 3 ...
```

```
##  $ sessions            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ total_orders        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ cart_order          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ phone_order         : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ city                : Factor w/ 2649 levels "aberdeen","abilene",..: 2025 189 370 2060 669 669 258
##  $ connection_speed    : chr  "cable" "broadband" "cable" "cable" ...
##  $ country             : Factor w/ 1 level "us": 1 1 1 1 1 1 1 1 1 1 ...
##  $ device_type         : Factor w/ 3 levels "desktop","mobile",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ distinct_page_views : num  1 2 5 2 2 2 3 3 3 3 ...
##  $ isp_name            : chr  "time warner cable internet llc" "mci communications services inc. dba
##  $ landing_page_raw    : chr  "https://phonecompanyx.com/" "https://phonecompanyx.com/" "https://phon
##  $ manufacturer        : Factor w/ 41 levels "acer","alcatel",..: 38 38 5 5 38 38 5 38 38 38 ...
##  $ metro_name          : Factor w/ 196 levels "abilene-sweetwater",..: 155 103 103 173 78 78 78 173 1
##  $ os_name             : Factor w/ 18 levels "android","chrome os",..: 12 12 7 7 11 11 7 11 12 12 ..
##  $ order_monthly_charge: num  25 25 25 60 60 ...
##  $ state               : Factor w/ 51 levels "ak","al","ar",..: 44 5 5 10 7 7 7 10 48 48 ...
##  $ zip_code            : Factor w/ 5104 levels "1027","1056",..: 3749 4227 4286 1767 116 116 143 167
##  $ traffic_source      : Factor w/ 4 levels "direct entry",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ day                 : Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Finding no. of unique values per column

lapply(redventuredf, function(x) length(unique(x)))
```

```
## $session_id
## [1] 20189
##
## $session_start_time
## [1] 19973
##
## $browser_name
## [1] 20
##
## $sessions
## [1] 1
##
## $total_orders
## [1] 2
##
## $cart_order
## [1] 2
##
## $phone_order
## [1] 2
##
## $city
## [1] 2649
##
## $connection_speed
## [1] 8
##
## $country
## [1] 1
##
```

```
## $device_type
## [1] 3
##
## $distinct_page_views
## [1] 6
##
## $isp_name
## [1] 905
##
## $landing_page_raw
## [1] 1670
##
## $manufacturer
## [1] 41
##
## $metro_name
## [1] 196
##
## $os_name
## [1] 18
##
## $order_monthly_charge
## [1] 6
##
## $state
## [1] 51
##
## $zip_code
## [1] 5104
##
## $traffic_source
## [1] 4
##
## $day
## [1] 7
```

```
# Finding order conversion percentage
sum(redventuredf[5])
```

```
## [1] 663
```

```
count(redventuredf[1])
```

```
##       n
## 1 20256
```

```
sum(redventuredf[5])/count(redventuredf[1]) * 100
```

```
##          n
## 1 3.273104
```

#4.1 Is there a relationship between number of sessions and the traffic source for each of the week? How can Red Ventures leverage this information?

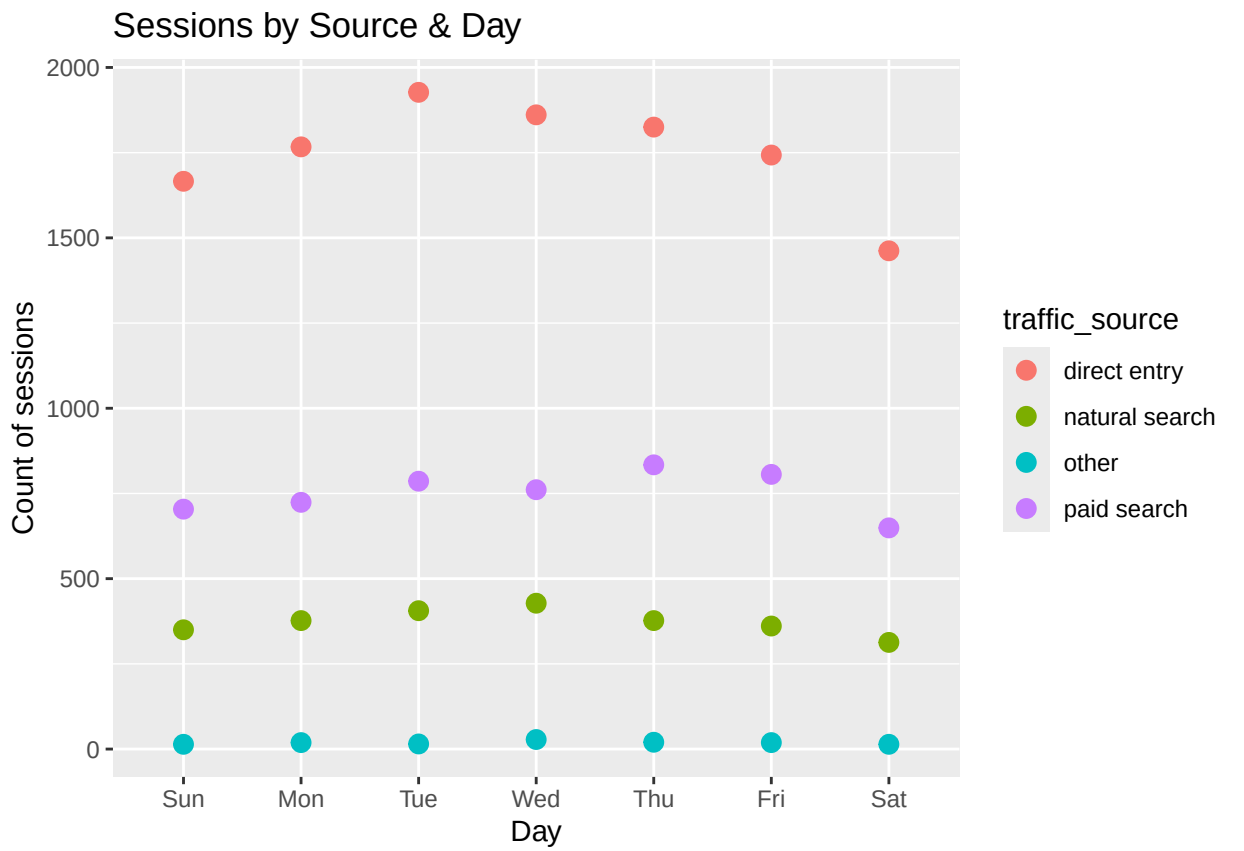
Knowing session counts by traffic source and day of the week can provide crucial insights for Red Ventures to boost the effectiveness of current marketing campaigns. By identifying which days of the week account for more traffic through specific sources, advertising /marketing budget can be allocated accordingly and more specifically. From our analysis and visualization through a scatter plot, we have identified that the highest source of traffic is 'Direct' search, followed by 'Paid' search and then 'Organic'. Within individual sources, Tuesdays are the highest for direct traffic while Thursdays followed by Fridays are the highest for paid searches. Elevated levels of direct traffic indicates good brand awareness and recognition. Red Ventures can leverage their brand recall, and amplify promotional activities such as exclusive deals and product releases on the best performing day - Tuesday. A second strategy could be to increase ad budgets on bids on Thursdays and Fridays, when paid searches peak. On days with lower traffic, spend can be reduced to ensure marketing budgets are allocated more effectively. Organic search, being lower than the above sources mean that Search Engine Optimization strategies are not in place. They can be improved through steps like keyword research, and on-page optimization of meta tags, headers, alt text etc. Along with this, the "Other" category of traffic source seems to be an untapped territory for Red Ventures. Social media presence, email marketing, referrals, blogs, influencer marketing etc. can be explored further.

```
# Creating a dataframe grouped by day and source to check session count

sessionsbysource_day<-redventuredf %>% select(session_id,day,traffic_source) %>%
group_by(day,traffic_source)%>%
summarise(count=n())
```

```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
## argument.
```

```
# Creating scatter plot
ggplot(data=sessionsbysource_day, aes(x=day, y=count, color=traffic_source))+geom_point(size=3)+
ggtitle("Sessions by Source & Day")+
xlab("Day")+ylab("Count of sessions")
```



#4.2 What is the underlying pattern between the total number of orders and the states from which they are placed? How can this help Red Ventures tailor marketing strategies specific to different regions?

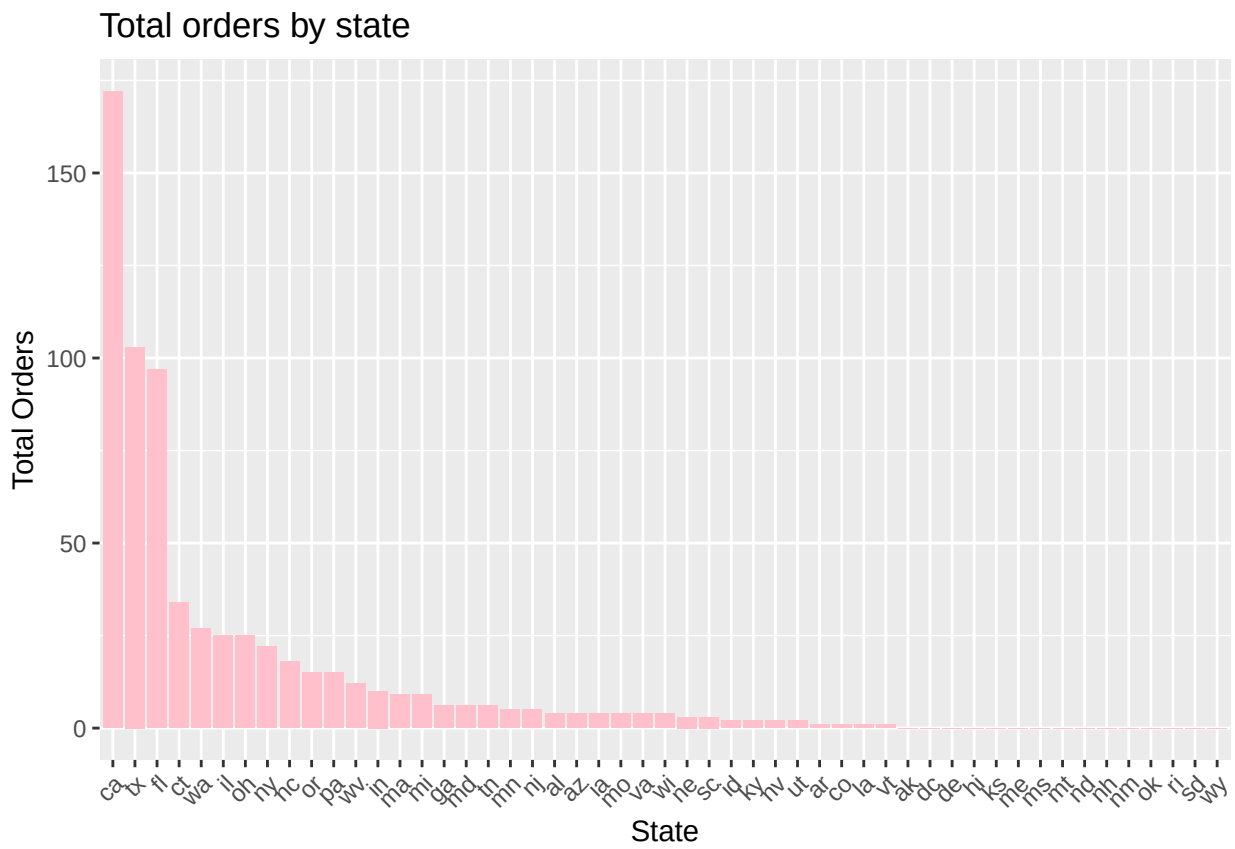
Having adequate knowledge about which region orders come from, can greatly strengthen marketing efforts for a company like Red Ventures, which focuses on customer acquisition across various industries and localities. From our data and the column chart, it is apparent that most transactions come from California, Texas, Florida, Connecticut and Washington. This means that marketing and promotional activities echo better with the above-mentioned regions, indicating that more ad spend can be allocated towards areas with a higher likelihood of conversion and profitability. Analyzing states with lower market share can be a signal to explore new opportunities for customer acquisition through revised marketing strategies. Adjustments in current strategy might be necessary to understand the reason for under-performance. This involves steps like A/B testing various strategies or benchmarking to compare with competitors' performance in such low-conversion states.

```
# Creating a df grouped by state to check order count

orderbystate<-redventuredf %>% select(state,total_orders) %>% group_by(state) %>%
 summarise(total=sum(total_orders))%>% arrange(desc(total))

# Create a column chart

ggplot(orderbystate) + geom_col(aes(x=reorder(state,desc(total)),y=total), fill='pink')+
ggtitle("Total orders by state")+ xlab("State") + ylab("Total Orders") +     theme(axis.text.x=element_
```



#4.3 How do device types and browsers influence the number of sessions?

Grouping the customers based on the type of device along with the browser they use to access The Red Ventures website, would be beneficial to segmenting the audience and optimizing their experience through targeted marketing strategies. This analysis can provide insights into demographics, preferred content format and usage patterns across different devices and browsers. On analysis of the data and preparation of a bar chart, we can see that desktops are the most used device type, followed by mobile phones and tablets. With respect to browsers used, Chrome is the most popular among desktop browsers whereas, among mobile and tablet users, Safari is favored most. Since there is a stark difference in the number of desktop versus mobile/tablet users, the Red Ventures website can be optimized for mobile usage. Since mobile/ tablet users may prefer quick load times and succinct content (video snippets or images).On the other hand, desktop users may prefer detailed and richer content. Hence, a targeted recommendation system such as mobile-friendly layouts and concise content for the former category and in-depth articles or interactive features for the latter can reduce bounce rates while enhancing customer experience.Since we know which browsers are utilized most, ensuring cross-browser compatibility is also essential as certain features render differently across each browser.

```
# Creating df grouped by device type and browser to check session count

sessionbydevice_browser<-redventuredf %>% select(device_type,browser_name,session_id) %>%
  mutate(browser_name = str_replace_all(browser_name, c(
  "chrome mobile" = "chrome",
  "internet explorer mobile" = "internet explorer",
  "opera mobile" = "opera",
  "unknown" = "other"))) %>%
  group_by(device_type,browser_name) %>% summarise(count=n())
```

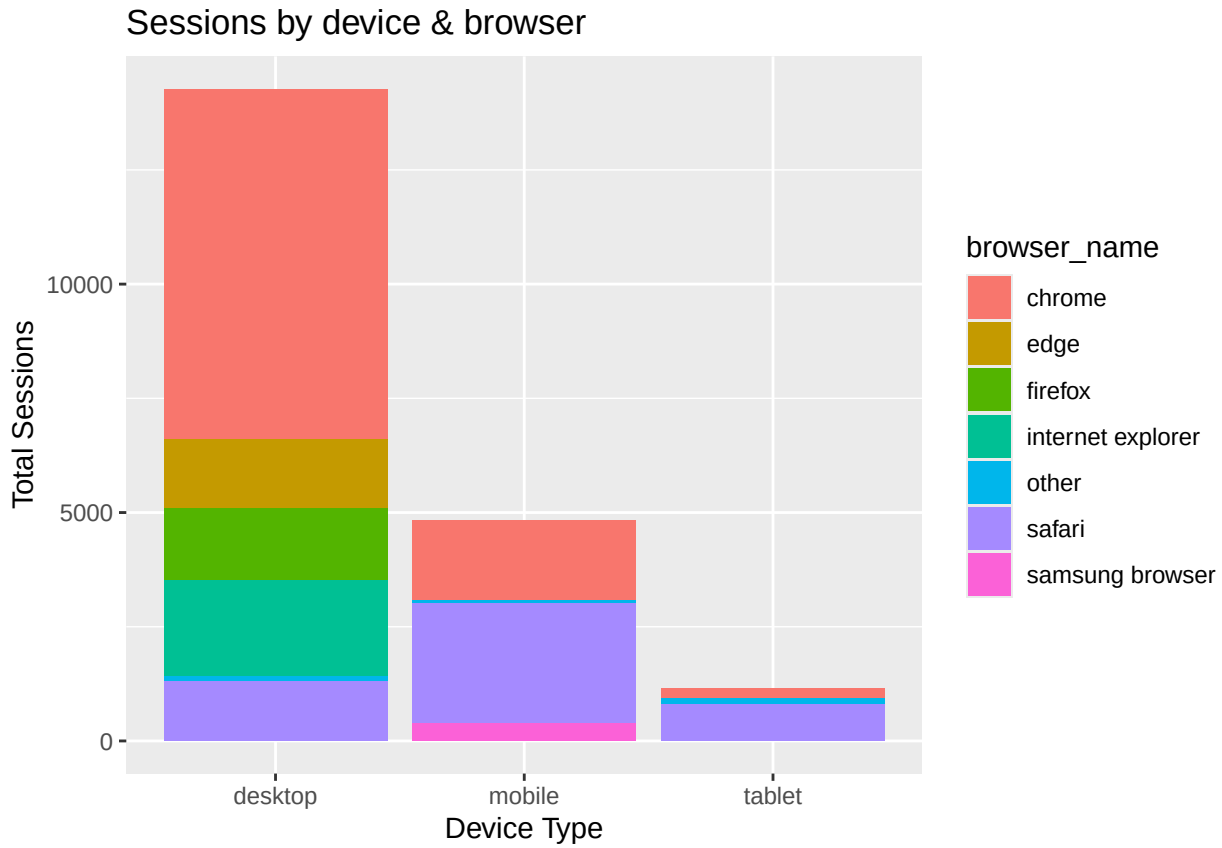
```
## `summarise()` has grouped output by 'device_type'. You can override using the
## `.groups` argument.
```

```
sessionbydevice_browser <- sessionbydevice_browser %>%
mutate(browser_name = ifelse(count<100,"other",browser_name)) %>% group_by(device_type,browser_name) %>%
summarise(count=sum(count))%>%arrange(device_type,desc(count))
```

```
## `summarise()` has grouped output by 'device_type'. You can override using the
## `.groups` argument.
```

```
# Creating a stacked chart

ggplot(sessionbydevice_browser, aes(x=reorder(device_type,desc(count)),y=count,fill=browser_name))+
geom_col(position = "stack")+ ggtitle("Sessions by device & browser")+
xlab("Device Type") + ylab("Total Sessions")
```



The exploratory data analysis for Red Ventures reveals that desktops are the primary device used, with Chrome as the leading browser, while mobile/tablet users prefer Safari. To optimize user experience, the company should enhance its website for both device types, offering rich content for desktop users and mobile-friendly layouts for mobile/tablet users. Marketing efforts should focus on high-conversion states like California and Texas, while exploring opportunities in under-performing regions. Additionally, Red Ventures can boost traffic by launching promotions on peak days, improving SEO strategies, and tapping into social media and email marketing channels for broader reach. To further leverage this data, the company can create features and dependent variables for predictive and prescriptive analysis, enabling data-driven decision-making that enhances customer targeting and improves marketing effectiveness.