Data Science

Statistical Data Modelling

Task 2:

Use Logistic Regression to solve the case.: Lead Scoring for xyz college

Introduction:

XYZ college has a database of potential students who have expressed interest in taking up courses offered by the college. The college wishes to identify leads that are more likely to convert into students. The dataset provided contains information about various leads, including their details, interaction with the college, and the stage of their lead status.

Objective:

To assign a lead score to each lead, based on their probability of becoming a student.

Data Description:

The dataset "Logistic Regression.zip" contains two files, "leads.csv" contain 9,000 rows, each corresponding to a unique lead. The dataset includes 37 columns, with details about the lead's information, communication history, and the stage of their lead status. The descriptions of variables are provided in "Lead data dictionary" file.

The college requires you to build a model wherein you need to assign a lead score to each of the leads such that the students with higher lead score have a higher conversion chance and lower lead score have a lower conversion chance. The objective of this assignment is to assign a lead score to each lead, based on their likelihood of closing, with the possible values of "High", "Medium", and "Low". The methodology involves data cleaning, data exploration, feature engineering, model building, and lead scoring.

Create necessary dummy variables. Delete unnecessary variables. The final lead score will help xyz College identify leads that are more likely to convert into paying students, thereby improving their conversion rate and revenue. You can use R or Python to run logistic regression.

Report should include the following:

- Logistic regression equation as per data result.
- Interpretation of the results.
- Validity of the model and summary of the findings.

For categorization and predictive analytics, this kind of statistical model—also referred to as the logit model—is frequently used. Logistic regression uses a dataset of independent variables to estimate the likelihood of an event occurring, such as voting or not. Because the result is a probability, the dependent variable has a range of 0 to 1.

Data source: leads.csv (provided by the institute)

Loading packages and reading the data source file.

```
# Logistic regression

# install the package
installed.packages("dplyr")

# loading package
library(dplyr)
library(psych)
library(tidyverse)
library(ROCR)
library(caTools)

leadsdata <- read.csv("D:\\Leads.csv")
```

Inspecting the data:

Descriptive statistics.

```
> leadsdata <- read.csv("D:\\Leads.csv")
> # DATA INSPECTION ###############################
> # describes function from psych library
> describe(leadsdata)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Prospect.ID*	1	9240	4620.50	2667.50	4620.5	4620.50	3424.81	1	9240	9239	0.00	-1.20	27.75
Lead.Number	2	9240	617188.44	23406.00	615479.0	616667.19	30265.06	579533	660737	81204	0.14	-1.21	243.50
Lead.Origin*	3	9240	1.70	0.63	2.0	1.64	0.00	1	5	4	0.49	0.09	0.01
Lead.Source*	4	9240	8.92	3.46	8.0	8.58	4.45	1	22	21	0.77	0.81	0.04
Do.Not.Email*	5	9240	1.08	0.27	1.0	1.00	0.00	1	2	1	3.11	7.67	0.00
Do.Not.Call*	6	9240	1.00	0.01	1.0	1.00	0.00	1	2	1	67.94	4614.00	0.00
Converted	7	9240	0.39	0.49	0.0	0.36	0.00	0	1	1	0.47	-1.78	0.01
TotalVisits	8	9103	3.45	4.85	3.0	2.83	2.97	0	251	251	19.91	852.82	0.05
Total.Time.Spent.on.Website	9	9240	487.70	548.02	248.0	409.70	367.68	0	2272	2272	0.96	-0.40	5.70
Page.Views.Per.Visit	10	9103	2.36	2.16	2.0	2.11	1.48	0	55	55	2.87	42.33	0.02
Last.Activity*	11	9240	9.61	3.76	9.0	9.82	4.45	1	18	17	-0.12	-1.22	0.04
Country*	12	9240	10.96	6.73	14.0	11.25	0.00	1	39	38	0.15	1.72	0.07
Specialization*	13	9240	9.98	6.12	10.0	10.05	10.38	1	20	19	-0.10	-1.33	0.06
How.did.you.hear.about.X.Education*	14	9240	5.54	2.86	7.0	5.59	0.00	1	11	10	-0.50	-0.70	0.03
What.is.your.current.occupation*	15	9240	4.59	2.33	6.0	4.76	0.00	1	7	6	-0.85	-1.18	0.02
What.matters.most.to.you.in.choosing.a.course*	16	9240	1.71	0.46	2.0	1.76	0.00	1	4	3	-0.89	-1.10	0.00
Search*	17	9240	1.00	0.04	1.0	1.00	0.00	1	2	1	25.63	654.86	0.00
Magazine*	18	9240	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
Newspaper.Article*	19	9240	1.00	0.01	1.0	1.00	0.00	1	2	1	67.94	4614.00	0.00
X.Education.Forums*	20	9240	1.00	0.01	1.0	1.00	0.00	1	2	1	96.09	9233.00	0.00
Newspaper*	21	9240	1.00	0.01	1.0	1.00	0.00	1	2	1	96.09	9233.00	0.00
Digital.Advertisement*	22	9240	1.00	0.02	1.0	1.00	0.00	1	2	1	48.02	2304.50	0.00
Through.Recommendations*	23	9240	1.00	0.03	1.0	1.00	0.00	1	2	1	36.28	1314.72	0.00
Receive.More.Updates.About.Our.Courses*	24	9240	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
Tags*	25	9240	11.49	10.54	9.0	10.99	11.86	1	27	26	0.29	-1.67	0.11
Lead.Quality*	26	9240	2.50	1.76	1.0	2.29	0.00	1	6	5	0.63	-1.15	0.02
Update.me.on.Supply.Chain.Content*	27	9240	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
Get.updates.on.DM.Content*	28	9240	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
Lead.Profile*	29	9240	4.26	2.19	5.0	4.42	1.48	1	7	6	-0.69	-1.30	0.02
City*	30	9240	3.57	2.13	2.0	3.47	1.48	1	8	7	0.37	-1.46	0.02
Asymmetrique.Activity.Index*	31	9240	2.04	1.01	2.0	2.00	1.48	1	4	3	0.15	-1.61	0.01
Asymmetrique.Profile.Index*	32	9240	1.85	0.87	2.0	1.81	1.48	1	4	3	0.32	-1.51	0.01
Asymmetrique.Activity.Score	33	5022	14.31	1.39	14.0	14.32	1.48	7	18	11	-0.38	1.23	0.02
Asymmetrique.Profile.Score	34	5022	16.34	1.81	16.0	16.29	1.48	11	20	9	0.22	-0.62	0.03
I.agree.to.pay.the.amount.through.cheque*	35	9240	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00
A.free.copy.of.Mastering.The.Interview*	36	9240	1.31	0.46	1.0	1.27	0.00	1	2	1	0.81	-1.35	0.00
Last.Notable.Activity*	37	9240	8.69	3.16	9.0	8.67	5.93	1	16	15	0.01	-1.17	0.03

```
>
```

Viewing data types of each column.

```
> str(leadsdata)
'data.frame':   9240 obs. of  37 variables:
 $ Prospect.ID                                  : chr  "7927b2df-8bba-4d29-b9a2-b6e0beafe620" "2a272436-5132-4136-86fa-dcc88c88f482" "8cc8c611-a219-4f35-ad23-fdfd
2656bd8a" "0cc2df48-7cf4-4e39-9de9-19797f9b38cc" ...
 $ Lead.Number                                  : int  660737 660728 660727 660719 660681 660680 660673 660664 660624 660616 ...
 $ Lead.Origin                                  : chr  "API" "API" "Landing Page Submission" "Landing Page Submission" ...
 $ Lead.Source                                  : chr  "Olark Chat" "Organic Search" "Direct Traffic" "Direct Traffic" ...
 $ Do.Not.Email                                 : chr  "No" "No" "No" "No" ...
 $ Do.Not.Call                                  : chr  "No" "No" "No" "No" ...
 $ Converted                                    : int  0 0 1 0 1 0 1 0 0 0 ...
 $ TotalVisits                                  : int  0 5 2 1 2 0 2 0 2 4 ...
 $ Total.Time.Spent.on.Website                  : int  0 674 1532 305 1428 0 1640 0 71 58 ...
 $ Page.Views.Per.Visit                         : num  0 2.5 2 1 1 0 2 0 2 4 ...
 $ Last.Activity                                : chr  "Page Visited on Website" "Email Opened" "Email Opened" "Unreachable" ...
 $ Country                                      : chr  "" "India" "India" "India" ...
 $ Specialization                               : chr  "Select" "Select" "Business Administration" "Media and Advertising" ...
 $ How.did.you.hear.about.X.Education           : chr  "Select" "Select" "Select" "Word Of Mouth" ...
 $ What.is.your.current.occupation              : chr  "Unemployed" "Unemployed" "Student" "Unemployed" ...
 $ What.matters.most.to.you.in.choosing.a.course: chr  "Better Career Prospects" "Better Career Prospects" "Better Career Prospects" "Better Career Prospects" ...
 $ Search                                       : chr  "No" "No" "No" "No" ...
 $ Magazine                                     : chr  "No" "No" "No" "No" ...
 $ Newspaper.Article                            : chr  "No" "No" "No" "No" ...
 $ X.Education.Forums                           : chr  "No" "No" "No" "No" ...
 $ Newspaper                                    : chr  "No" "No" "No" "No" ...
 $ Digital.Advertisement                        : chr  "No" "No" "No" "No" ...
 $ Through.Recommendations                      : chr  "No" "No" "No" "No" ...
 $ Receive.More.Updates.About.Our.Courses       : chr  "No" "No" "No" "No" ...
 $ Tags                                         : chr  "Interested in other courses" "Ringing" "Will revert after reading the email" "Ringing" ...
 $ Lead.Quality                                 : chr  "Low in Relevance" "" "Might be" "Not Sure" ...
 $ Update.me.on.Supply.Chain.Content            : chr  "No" "No" "No" "No" ...
 $ Get.updates.on.DM.Content                    : chr  "No" "No" "No" "No" ...
 $ Lead.Profile                                 : chr  "Select" "Select" "Potential Lead" "Select" ...
 $ City                                         : chr  "Select" "Select" "Mumbai" "Mumbai" ...
 $ Asymmetrique.Activity.Index                  : chr  "02.Medium" "02.Medium" "02.Medium" "02.Medium" ...
 $ Asymmetrique.Profile.Index                   : chr  "02.Medium" "02.Medium" "01.High" "01.High" ...
 $ Asymmetrique.Activity.Score                  : int  15 15 14 13 15 17 14 15 14 13 ...
 $ Asymmetrique.Profile.Score                   : int  15 15 20 17 18 15 20 15 14 16 ...
 $ I.agree.to.pay.the.amount.through.cheque     : chr  "No" "No" "No" "No" ...
 $ A.free.copy.of.Mastering.The.Interview       : chr  "No" "No" "Yes" "No" ...
 $ Last.Notable.Activity                        : chr  "Modified" "Email Opened" "Email Opened" "Modified" ...
>
```

Finding the number of columns and rows (dimension).

```
> # no of rows and columns
> dim(leadsdata)
[1] 9240   37
>
```

Displaying the summary of the dataset.

```
> # summary of the dataset
> summary(leadsdata)
 Prospect.ID         Lead.Number     Lead.Origin        Lead.Source        Do.Not.Email       Do.Not.Call          Converted       TotalVisits
 Length:9240        Min.   :579533   Length:9240        Length:9240        Length:9240        Length:9240        Min.   :0.0000   Min.   :  0.000
 Class :character   1st Qu.:596485   Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000   1st Qu.:  1.000
 Mode  :character   Median :615479   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Median :  3.000
                    Mean   :617188                                                           Mean   :0.3854   Mean   :  3.445
                    3rd Qu.:637387                                                           3rd Qu.:1.0000   3rd Qu.:  5.000
                    Max.   :660737                                                           Max.   :1.0000   Max.   :251.000
                                                                                                              NA's   :137
 Total.Time.Spent.on.Website Page.Views.Per.Visit Last.Activity       Country          Specialization     How.did.you.hear.about.X.Education
 Min.   :   0.0              Min.   : 0.000       Length:9240        Length:9240        Length:9240        Length:9240
 1st Qu.:  12.0              1st Qu.: 1.000       Class :character   Class :character   Class :character   Class :character
 Median : 248.0              Median : 2.000       Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 487.7              Mean   : 2.363
 3rd Qu.: 936.0              3rd Qu.: 3.000
 Max.   :2272.0              Max.   :55.000
                             NA's   :137
 What.is.your.current.occupation What.matters.most.to.you.in.choosing.a.course    Search           Magazine         Newspaper.Article  X.Education.Forums
 Length:9240                     Length:9240                                     Length:9240        Length:9240        Length:9240        Length:9240
 Class :character                Class :character                                Class :character   Class :character   Class :character   Class :character
 Mode  :character                Mode  :character                                Mode  :character   Mode  :character   Mode  :character   Mode  :character



  Newspaper         Digital.Advertisement Through.Recommendations Receive.More.Updates.About.Our.Courses     Tags           Lead.Quality
 Length:9240        Length:9240           Length:9240             Length:9240                           Length:9240        Length:9240
 Class :character   Class :character      Class :character        Class :character                      Class :character   Class :character
 Mode  :character   Mode  :character      Mode  :character        Mode  :character                      Mode  :character   Mode  :character



 Update.me.on.Supply.Chain.Content Get.updates.on.DM.Content Lead.Profile          City           Asymmetrique.Activity.Index Asymmetrique.Profile.Index
 Length:9240                       Length:9240               Length:9240        Length:9240        Length:9240                 Length:9240
 Class :character                  Class :character          Class :character   Class :character   Class :character            Class :character
 Mode  :character                  Mode  :character          Mode  :character   Mode  :character   Mode  :character            Mode  :character

 Asymmetrique.Activity.Score Asymmetrique.Profile.Score I.agree.to.pay.the.amount.through.cheque A.free.copy.of.Mastering.The.Interview Last.Notable.Activity
 Min.   : 7.00               Min.   :11.00              Length:9240                               Length:9240                           Length:9240
 1st Qu.:14.00               1st Qu.:15.00              Class :character                          Class :character                      Class :character
 Median :14.00               Median :16.00              Mode  :character                          Mode  :character                      Mode  :character
 Mean   :14.31               Mean   :16.34
 3rd Qu.:15.00               3rd Qu.:18.00
 Max.   :18.00               Max.   :20.00
 NA's   :4218                NA's   :4218
> |
```

Convert all column names into lowercase.

```
> # make column names to lowercase
> colnames(leadsdata) <- tolower(names(leadsdata))
> colnames(leadsdata)
 [1] "prospect.id"                                "lead.number"                                "lead.origin"
 [4] "lead.source"                                "do.not.email"                               "do.not.call"
 [7] "converted"                                  "totalvisits"                                "total.time.spent.on.website"
[10] "page.views.per.visit"                       "last.activity"                              "country"
[13] "specialization"                             "how.did.you.hear.about.x.education"         "what.is.your.current.occupation"
[16] "what.matters.most.to.you.in.choosing.a.course" "search"                                  "magazine"
[19] "newspaper.article"                          "x.education.forums"                         "newspaper"
[22] "digital.advertisement"                      "through.recommendations"                    "receive.more.updates.about.our.courses"
[25] "tags"                                       "lead.quality"                               "update.me.on.supply.chain.content"
[28] "get.updates.on.dm.content"                  "lead.profile"                               "city"
[31] "asymmetrique.activity.index"                "asymmetrique.profile.index"                 "asymmetrique.activity.score"
[34] "asymmetrique.profile.score"                 "i.agree.to.pay.the.amount.through.cheque"   "a.free.copy.of.mastering.the.interview"
[37] "last.notable.activity"
> |
```

Checking any duplicate values.

```
> duplicated(leadsdata)
   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [26] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [51] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [76] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [101] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [126] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [151] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [176] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [201] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [226] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [251] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [276] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [326] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
> # extract duplicates
> leadsdata[duplicated(leadsdata)]
data frame with 0 columns and 9240 rows
> |
```

Finding NA values.

```
> # check NA values in all columns
> colSums(is.na(leadsdata))
                                  prospect.id                                   lead.number                                   lead.origin
                                            0                                             0                                             0
                                  lead.source                                  do.not.email                                  do.not.call
                                            0                                             0                                             0
                                    converted                                  totalvisits                     total.time.spent.on.website
                                            0                                           137                                             0
                           page.views.per.visit                                last.activity                                       country
                                          137                                             0                                             0
                                 specialization                 how.did.you.hear.about.x.education                  what.is.your.current.occupation
                                            0                                             0                                             0
what.matters.most.to.you.in.choosing.a.course                                        search                                      magazine
                                            0                                             0                                             0
                              newspaper.article                             x.education.forums                                    newspaper
                                            0                                             0                                             0
                          digital.advertisement                         through.recommendations                 receive.more.updates.about.our.courses
                                            0                                             0                                             0
                                           tags                                  lead.quality                       update.me.on.supply.chain.content
                                            0                                             0                                             0
                     get.updates.on.dm.content                                  lead.profile                                          city
                                            0                                             0                                             0
                  asymmetrique.activity.index                   asymmetrique.profile.index                     asymmetrique.activity.score
                                            0                                             0                                          4218
                   asymmetrique.profile.score      i.agree.to.pay.the.amount.through.cheque                 a.free.copy.of.mastering.the.interview
                                          4218                                             0                                             0
                        last.notable.activity
                                            0
```

```
> # to find total no of NA values
> sum(colSums(is.na(leadsdata)))
[1] 8710
>
```

Replacing NA values with the mean value.

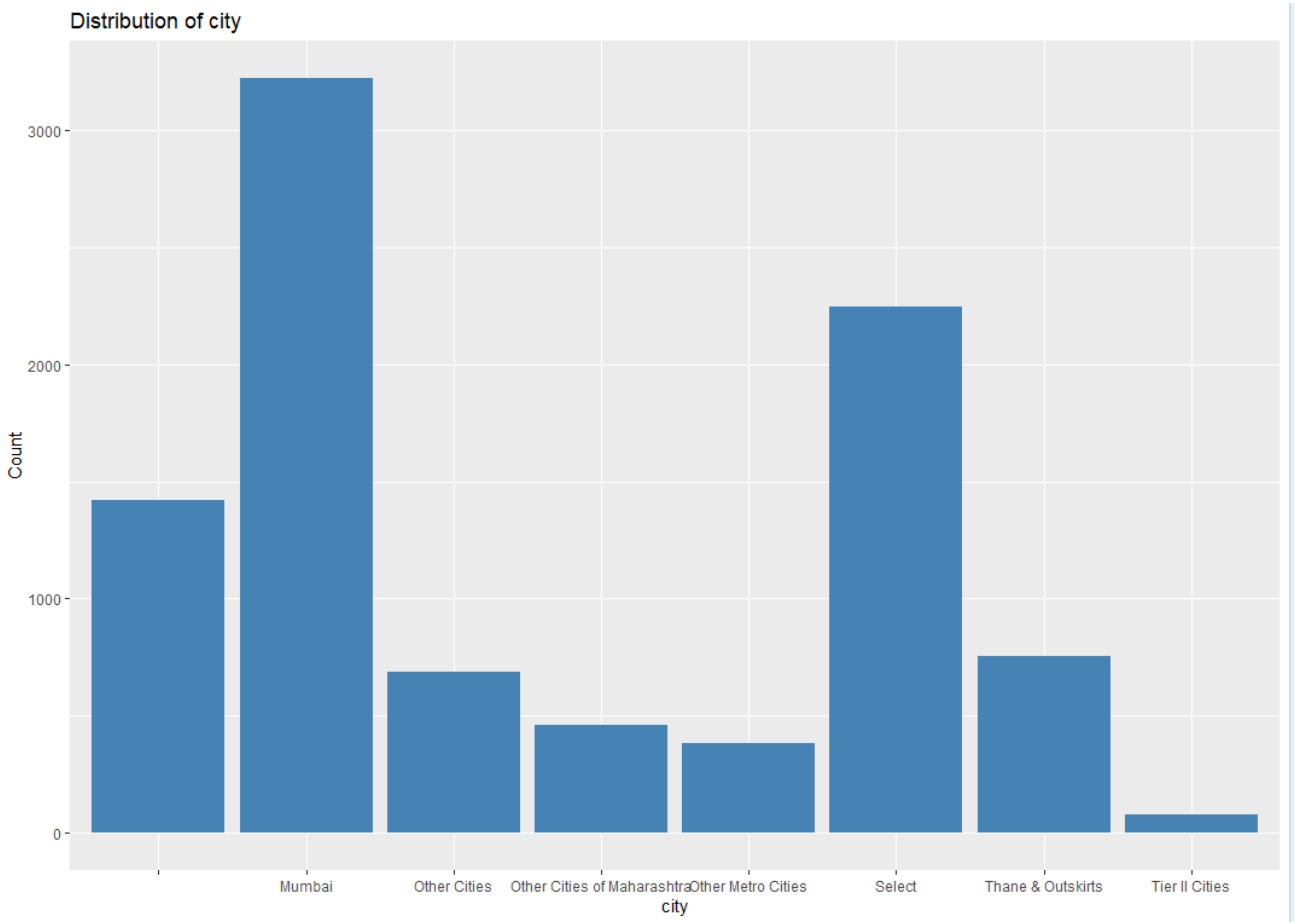
```
> # for the below 2 columns, calculate the mean without the NA and replacing the NA values with the mean
> leadsdata$totalvisits = ifelse(is.na(leadsdata$totalvisits),
+                                mean(leadsdata$totalvisits, na.rm = TRUE),
+                                leadsdata$totalvisits)
> leadsdata$page.views.per.visit = ifelse(is.na(leadsdata$page.views.per.visit),
+                                         mean(leadsdata$page.views.per.visit, na.rm = TRUE),
+                                         leadsdata$page.views.per.visit)
> |
```

Checking unique values in the city column.

```
> unique(leadsdata$city)
[1] "Select"                   "Mumbai"                   ""                         "Thane & Outskirts"        "Other Metro Cities"
[6] "Other Cities"             "Other Cities of Maharashtra" "Tier II Cities"
```

Generating a bar plot to find the distribution of the city column.

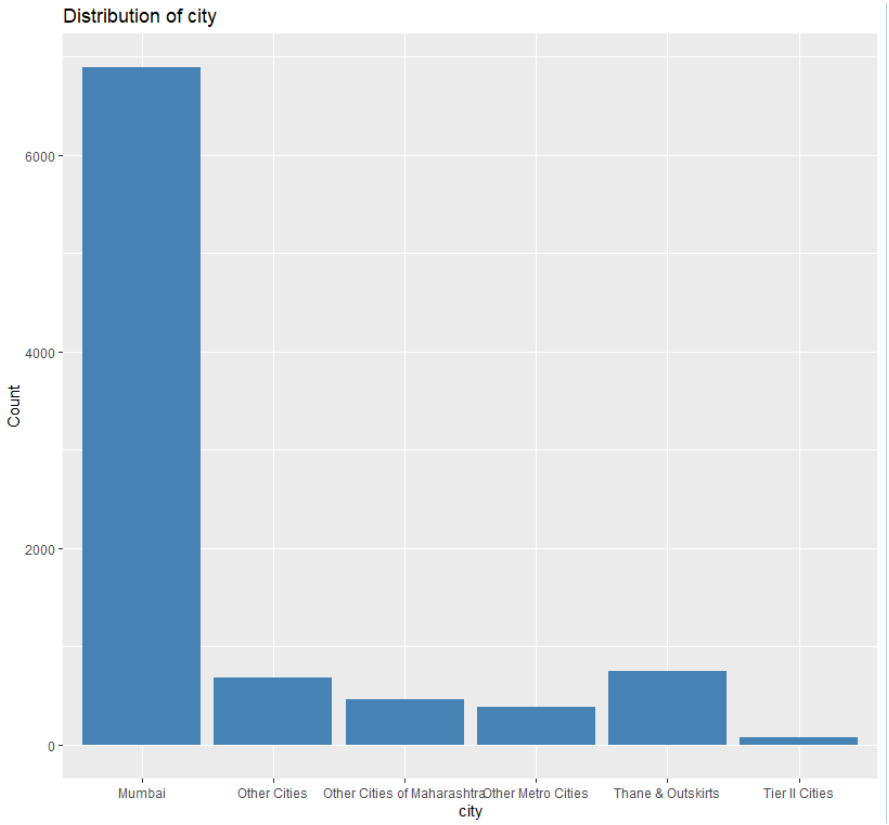
```
> # Create a bar plot to check the max count
> ggplot(leadsdata, aes(x = city)) + geom_bar(fill = "steelblue") + labs(title = "Distribution of city", x = "city", y = "Count")
> |
```



Since the Mumbai city has a maximum count, replacing the empty cells and “Select” cells with Mumbai city name.

```
> leadsdata$city <- ifelse(trimws(leadsdata$city) == "Select", "Mumbai", leadsdata$city)
> leadsdata$city <- ifelse(trimws(leadsdata$city) == "", "Mumbai", leadsdata$city)
>
```

After the replacement.,

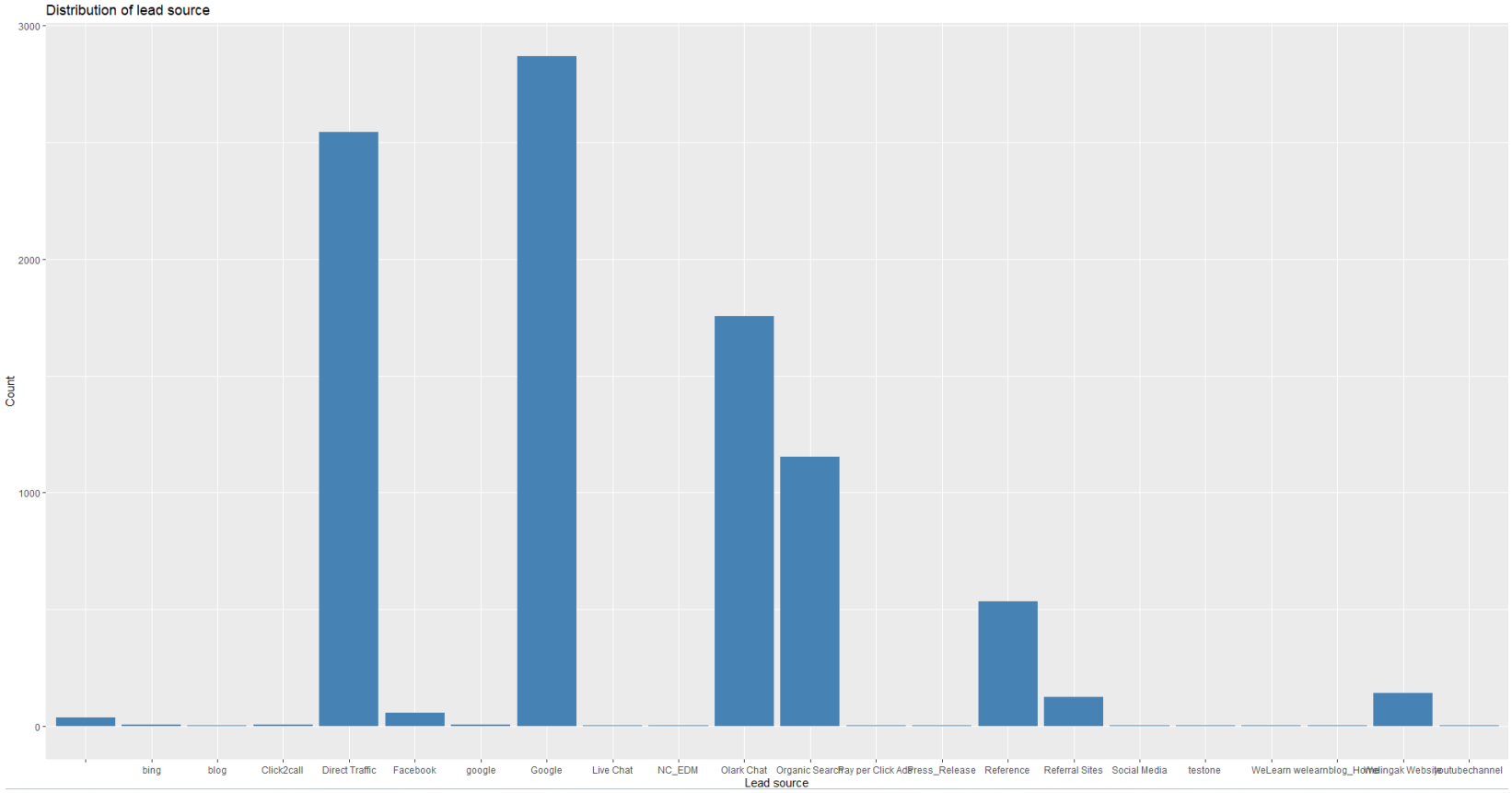


Replacing the lead profile's empty cells and "Select" cells with "Other Leads".

```
> leadsdata$lead.profile <- ifelse(trimws(leadsdata$lead.profile) == "Select", "Other Leads", leadsdata$lead.profile)
> leadsdata$lead.profile <- ifelse(trimws(leadsdata$lead.profile) == "", "Other Leads", leadsdata$lead.profile)
> unique(leadsdata$lead.profile)
[1] "Other Leads"                 "Potential Lead"              "Lateral Student"             "Dual Specialization Student"
[5] "Student of SomeSchool"
>
```

Generating a bar plot of lead source column.

```
> # Create a bar plot of lead.source
> ggplot(leadsdata, aes(x = lead.source)) + geom_bar(fill = "steelblue") + labs(title = "Distribution of lead source", x = "Lead source", y = "Count")
>
```



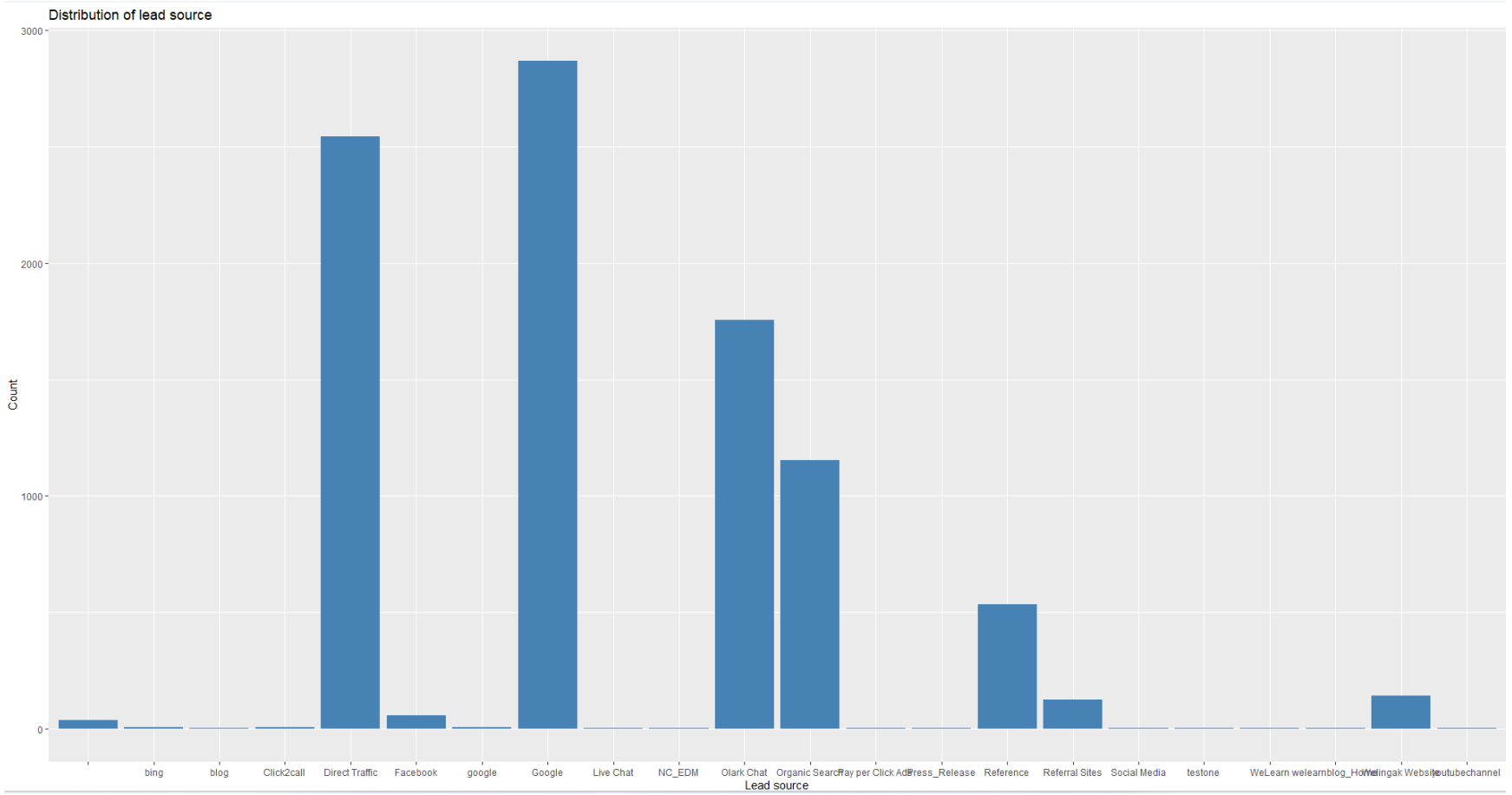
Calculate the frequency for each category in the lead source column.

```
> # Calculate the frequency count for each category
> frequency <- table(leadsdata$lead.source)
> # Find the category with the highest count (mode) / max frequency value gives the mode value
> mode <- names(frequency)[which.max(frequency)]
> # Print the mode
> print(mode)
[1] "Google"
>
```

Filling missing values with the above mode.

Re generating the bar plot to view the difference.

```
> # filling missing values in lead.source
> leadsdata$lead.source[is.na(leadsdata$lead.source)] <- mode
> # Create a bar plot of lead.origin after removing missing values
> ggplot(leadsdata, aes(x = lead.source)) + geom_bar(fill = "steelblue") + labs(title = "Distribution of lead source", x = "Lead source", y = "Count")
>
```



Replacing NA values with the median: asymmetrique.activity.score column.

```
> unique(leadsdata$asymmetrique.activity.score)
 [1] 15 14 13 17 16 11 12 10  9  8 18 NA  7
> table(leadsdata$asymmetrique.activity.score)

   7    8    9   10   11   12   13   14   15   16   17   18
   1    4    9   57   95  196  775 1771 1293  467  349    5
> # Replace NA values with median (numeric fields)################################################
> # Calculate the median
> median <- median(leadsdata$asymmetrique.activity.score,na.rm = TRUE)
> # filling missing values
> leadsdata$asymmetrique.activity.score <- ifelse(is.na(leadsdata$asymmetrique.activity.score),median, leadsdata$asymmetrique.activity.score)
> unique(leadsdata$asymmetrique.activity.score)
 [1] 15 14 13 17 16 11 12 10  9  8 18  7
> table(leadsdata$asymmetrique.activity.score)

   7    8    9   10   11   12   13   14   15   16   17   18
   1    4    9   57   95  196  775 5989 1293  467  349    5
> |
```

Replacing empty cells.

```
> unique(leadsdata$last.activity)
 [1] "Page Visited on Website"      "Email Opened"                "Unreachable"                 "Converted to Lead"           "Olark Chat Conversation"
 [6] "Email Bounced"                "Email Link Clicked"          "Form Submitted on Website"   "Unsubscribed"                "Had a Phone Conversation"
[11] "View in browser link Clicked" ""                            "Approached upfront"          "SMS Sent"                    "Visited Booth in Tradeshow"
[16] "Resubscribed to emails"       "Email Received"              "Email Marked Spam"
> leadsdata$last.activity <- ifelse(trimws(leadsdata$last.activity) == "", "other", leadsdata$last.activity)
> |
```

```
> unique(leadsdata$country)
 [1] ""                    "India"               "Russia"              "Kuwait"              "Oman"                "United Arab Emirates" "United States"
 [8] "Australia"           "United Kingdom"      "Bahrain"             "Ghana"               "Singapore"           "Qatar"                "Saudi Arabia"
[15] "Belgium"             "France"              "Sri Lanka"           "China"               "Canada"              "Netherlands"          "Sweden"
[22] "Nigeria"             "Hong Kong"           "Germany"             "Asia/Pacific Region" "Uganda"              "Kenya"                "Italy"
[29] "South Africa"        "Tanzania"            "unknown"             "Malaysia"            "Liberia"             "Switzerland"          "Denmark"
[36] "Philippines"         "Bangladesh"          "Vietnam"             "Indonesia"
> leadsdata$country <- ifelse(trimws(leadsdata$country) == "", "other", leadsdata$country)
> |
```

```
> unique(leadsdata$tags)
 [1] "Interested in other courses"                "Ringing"                                    "Will revert after reading the email"
 [4] ""                                           "Lost to EINS"                               "In confusion whether part time or DLP"
 [7] "Busy"                                       "switched off"                               "in touch with EINS"
[10] "Already a student"                          "Diploma holder (Not Eligible)"              "Graduation in progress"
[13] "Closed by Horizzon"                         "number not provided"                        "opp hangup"
[16] "Not doing further education"                "invalid number"                             "wrong number given"
[19] "Interested  in full time MBA"               "Still Thinking"                             "Lost to Others"
[22] "Shall take in the next coming month"        "Lateral student"                            "Interested in Next batch"
[25] "Recognition issue (DEC approval)"           "Want to take admission but has financial problems" "University not recognized"
> leadsdata$tags <- ifelse(trimws(leadsdata$tags) == "", "Other", leadsdata$tags)
>
```

Re-checking NA values' availability.

```
> # find NA values
> colSums(is.na(leadsdata))
                                prospect.id                                 lead.number                                 lead.origin
                                          0                                           0                                           0
                                lead.source                                do.not.email                                do.not.call
                                          0                                           0                                           0
                                  converted                                totalvisits                  total.time.spent.on.website
                                          0                                           0                                           0
                     page.views.per.visit                                last.activity                                     country
                                          0                                           0                                           0
                           specialization                     how.did.you.hear.about.x.education                what.is.your.current.occupation
                                          0                                           0                                           0
what.matters.most.to.you.in.choosing.a.course                                      search                                    magazine
                                          0                                           0                                           0
                          newspaper.article                           x.education.forums                                   newspaper
                                          0                                           0                                           0
                      digital.advertisement                       through.recommendations receive.more.updates.about.our.courses
                                          0                                           0                                           0
                                       tags                                lead.quality           update.me.on.supply.chain.content
                                          0                                           0                                           0
                  get.updates.on.dm.content                                lead.profile                                        city
                                          0                                           0                                           0
                asymmetrique.activity.index                 asymmetrique.profile.index                 asymmetrique.activity.score
                                          0                                           0                                           0
                asymmetrique.profile.score      i.agree.to.pay.the.amount.through.cheque       a.free.copy.of.mastering.the.interview
                                       4218                                           0                                           0
                     last.notable.activity
                                          0
```

Dropping unwanted / unimportant columns.

```
> # drop unwanted columns / so many NA values
> leadsdata$asymmetrique.profile.score <- NULL
> # drop unwanted columns / unique values: id and number
> leadsdata$prospect.id <- NULL
> leadsdata$lead.number <- NULL
```

```
> # drop dependent variables / occupation and specialization columns are inter dependent columns
> # find unique values
> unique(leadsdata$specialization)
 [1] "Select"                         "Business Administration"          "Media and Advertising"             ""
 [5] "Supply Chain Management"        "IT Projects Management"           "Finance Management"                "Travel and Tourism"
 [9] "Human Resource Management"      "Marketing Management"             "Banking, Investment And Insurance" "International Business"
[13] "E-COMMERCE"                     "Operations Management"            "Retail Management"                 "Services Excellence"
[17] "Hospitality Management"         "Rural and Agribusiness"           "Healthcare Management"             "E-Business"
> unique(leadsdata$what.is.your.current.occupation)
[1] "Unemployed"          "Student"             ""                    "Working Professional" "Businessman"          "Other"                "Housewife"
> # # drop unwanted column
> leadsdata$specialization <- NULL
> leadsdata$what.is.your.current.occupation <- NULL
>
```

```
> # drop unwanted columns / columns that has only one categorical variable
> unique(leadsdata$magazine)
[1] "No"
> unique(leadsdata$receive.more.updates.about.our.courses)
[1] "No"
> leadsdata$magazine <- NULL
> leadsdata$receive.more.updates.about.our.courses <- NULL
> # drop unwanted column
> leadsdata$how.did.you.hear.about.x.education <- NULL
> leadsdata$what.matters.most.to.you.in.choosing.a.course <- NULL
> # drop unwanted column
> unique(leadsdata$get.updates.on.dm.content)# only one categorical value / needs to drop
[1] "No"
> leadsdata$get.updates.on.dm.content <- NULL
>
```

```
> # drop unwanted column
> unique(leadsdata$update.me.on.supply.chain.content)# only one categorical value / needs to drop
[1] "No"
> leadsdata$update.me.on.supply.chain.content <- NULL
> # drop unwanted column
> unique(leadsdata$i.agree.to.pay.the.amount.through.cheque)# only one categorical value hence needs to drop the column
[1] "No"
> leadsdata$i.agree.to.pay.the.amount.through.cheque <- NULL
> # drop unwanted column
> leadsdata$a.free.copy.of.mastering.the.interviewlast.notable.activity <- NULL
>
```

Feature encoding:

Converting categorical variables.

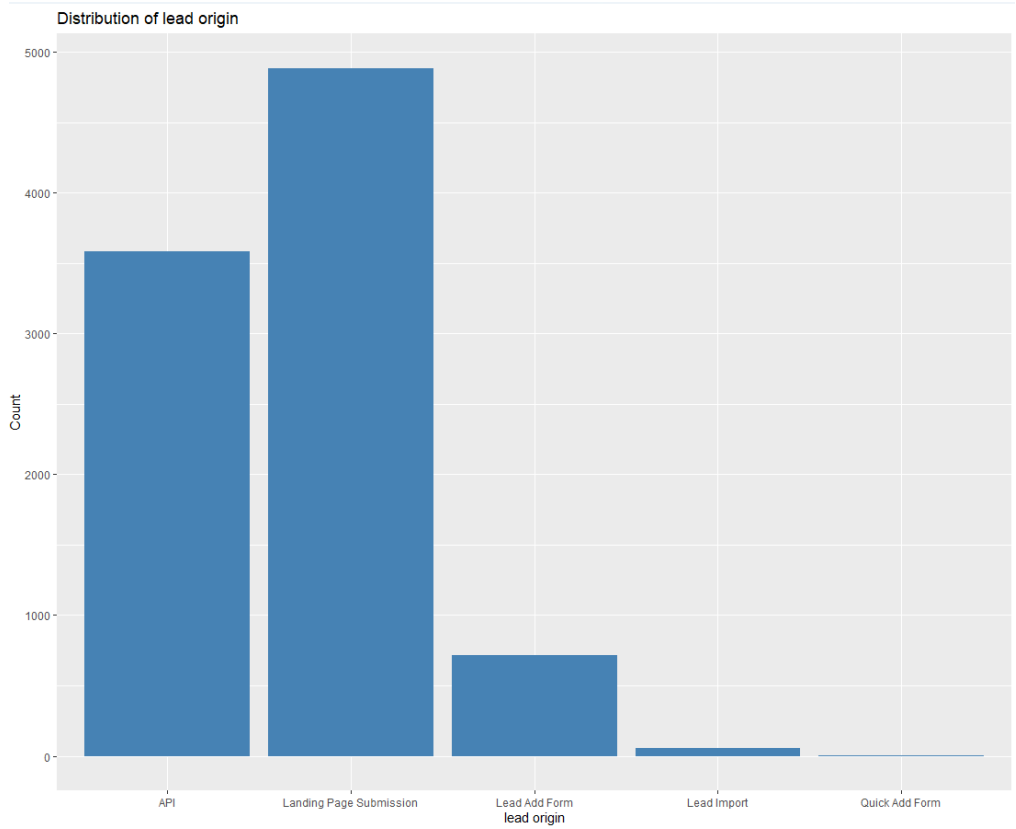
```
> unique(leadsdata$lead.quality)
[1] "Low in Relevance"  ""                  "Might be"          "Not Sure"          "Worst"             "High in Relevance"
> #Converting Ordinal to Factor
> leadsdata$lead.quality = factor(leadsdata$lead.quality, levels = c("Worst","Low in Relevance","Not Sure","Might be","High in Relevance",""), labels = c(1,2,3,4,5,5))
> unique(leadsdata$lead.quality)
[1] 2 5 4 3 1
Levels: 1 2 3 4 5
>
```

```
> unique(leadsdata$asymmetrique.activity.index)
[1] "02.Medium" "01.High"   "03.Low"    ""
> leadsdata$asymmetrique.activity.index = factor(leadsdata$asymmetrique.activity.index, levels = c("01.High","02.Medium","03.Low",""), labels = c(1,2,3,3))
> unique(leadsdata$asymmetrique.activity.index)
[1] 2 1 3
Levels: 1 2 3
>
```

```
> unique(leadsdata$asymmetrique.profile.index)
[1] "02.Medium" "01.High"   "03.Low"    ""
> leadsdata$asymmetrique.profile.index = factor(leadsdata$asymmetrique.profile.index, levels = c("01.High","02.Medium","03.Low",""), labels = c(1,2,3,3))
> unique(leadsdata$asymmetrique.profile.index)
[1] 2 1 3
Levels: 1 2 3
>
```

Finding the distribution with lead.origin.

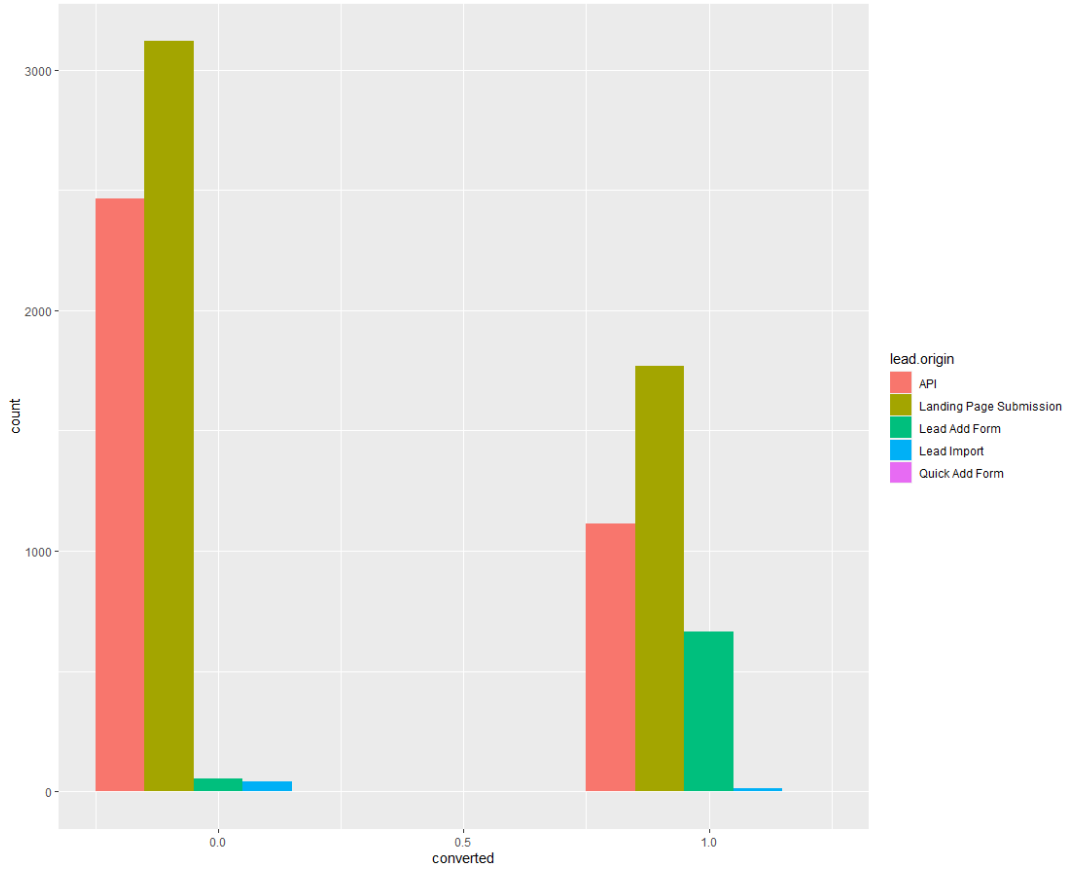
```
> # COMPARE LEAD ORIGIN WITH "CONVERTED"
> unique(leadsdata$lead.origin)
[1] "API"                    "Landing Page Submission" "Lead Add Form"           "Lead Import"             "Quick Add Form"
> ggplot(leadsdata, aes(x = lead.origin)) + geom_bar(fill = "steelblue") + labs(title = "Distribution of lead origin", x = "lead origin", y = "Count")
>
```



It is observable that the Landing page submission and API has high distribution count.

Comparing the lead origin classes with converted rate.

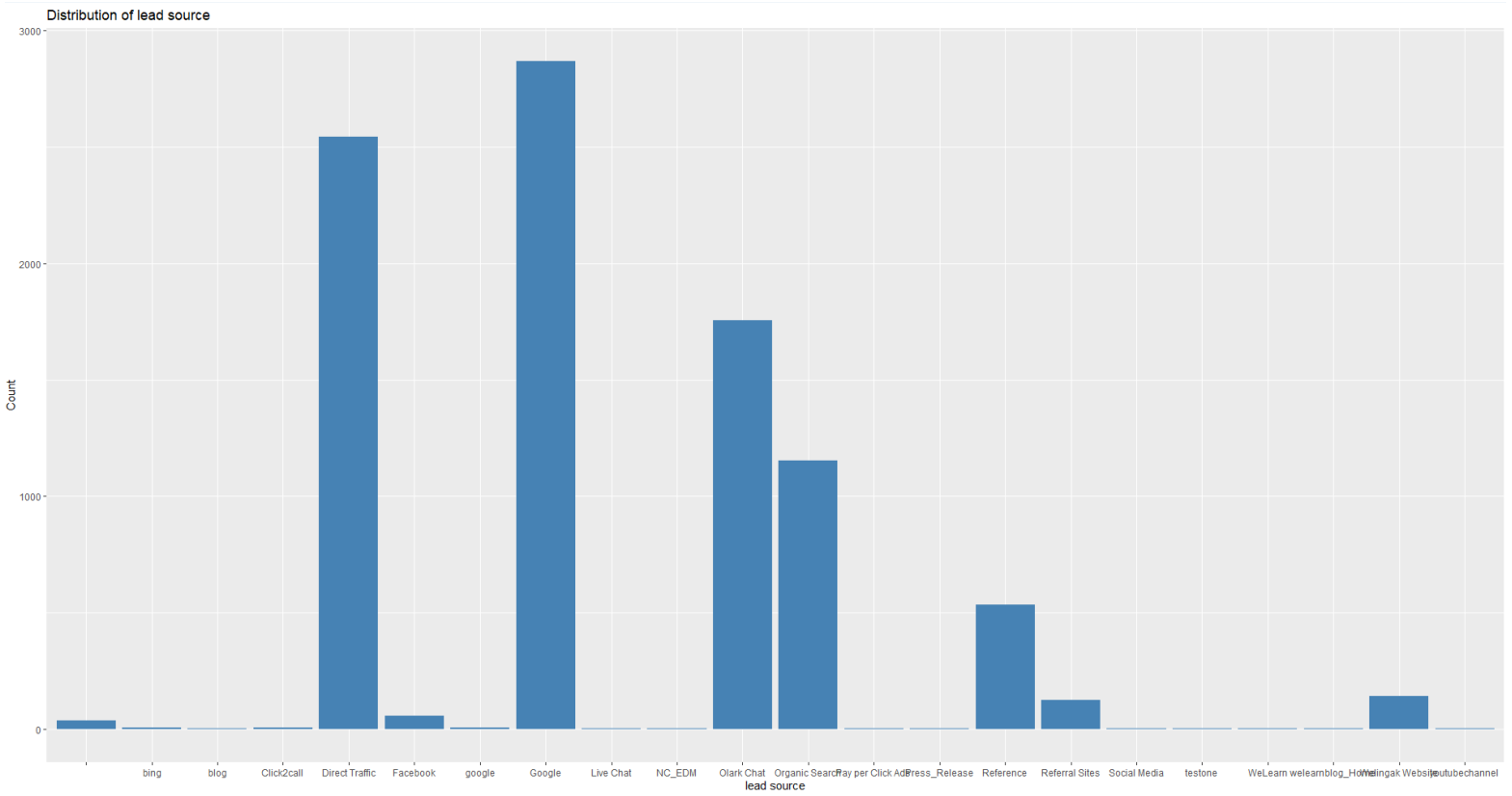
```
> p <- leadsdata %>%
+   select(converted, lead.origin) %>%
+   drop_na() %>%
+   ggplot(mapping = aes(x = converted, fill = lead.origin))
> p + geom_histogram(binwidth = 0.5, position = "dodge")
>
```



The graph indicates that the conversion rate for landing page and API submissions is less than 50% (minimal impact on lead conversion rate). On the other hand, lead conversion rate is greatly impacted by lead add form.

Viewing maximum frequency of the lead source classes.

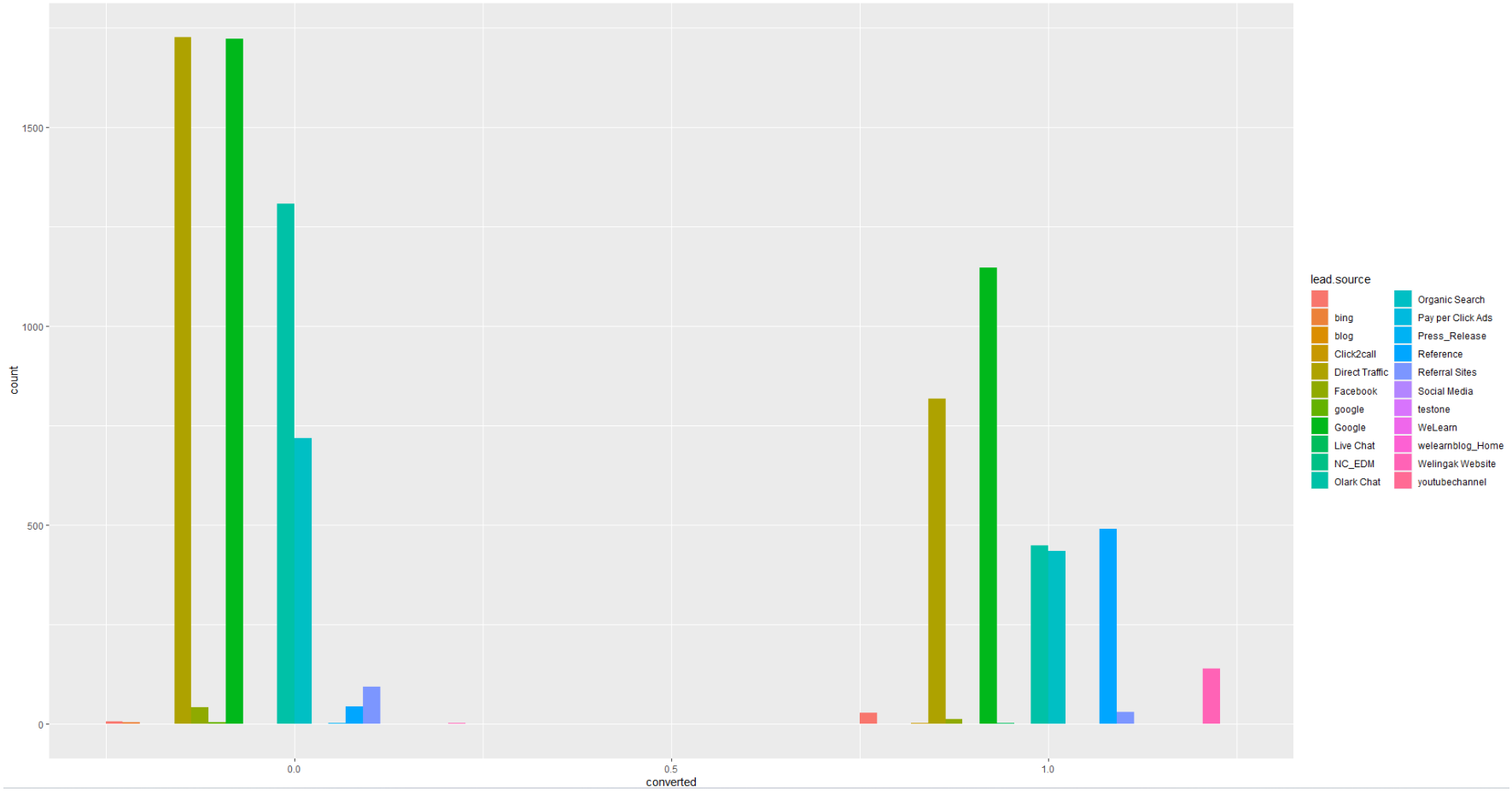
```
> unique(leadsdata$lead.source)
 [1] "Olark Chat"        "Organic Search"    "Direct Traffic"    "Google"            "Referral Sites"    "Welingak Website"  "Reference"         "google"
 [9] "Facebook"          ""                  "blog"              "Pay per Click Ads" "bing"              "Social Media"      "WeLearn"           "Click2call"
[17] "Live Chat"         "welearnblog_Home"  "youtubechannel"    "testone"           "Press_Release"     "NC_EDM"
> ggplot(leadsdata, aes(x = lead.source)) + geom_bar(fill = "steelblue") + labs(title = "Distribution of lead source", x = "lead source", y = "Count")
>
```



Based on the graph, direct traffic and google has maximum frequencies.

Comparing the lead source classes with converted rate.

```
> p <- leadsdata %>%
+   select(converted, lead.source) %>%
+   drop_na() %>%
+   ggplot(mapping = aes(x = converted, fill = lead.source))
> p + geom_histogram(binwidth = 0.5, position = "dodge")
>
```

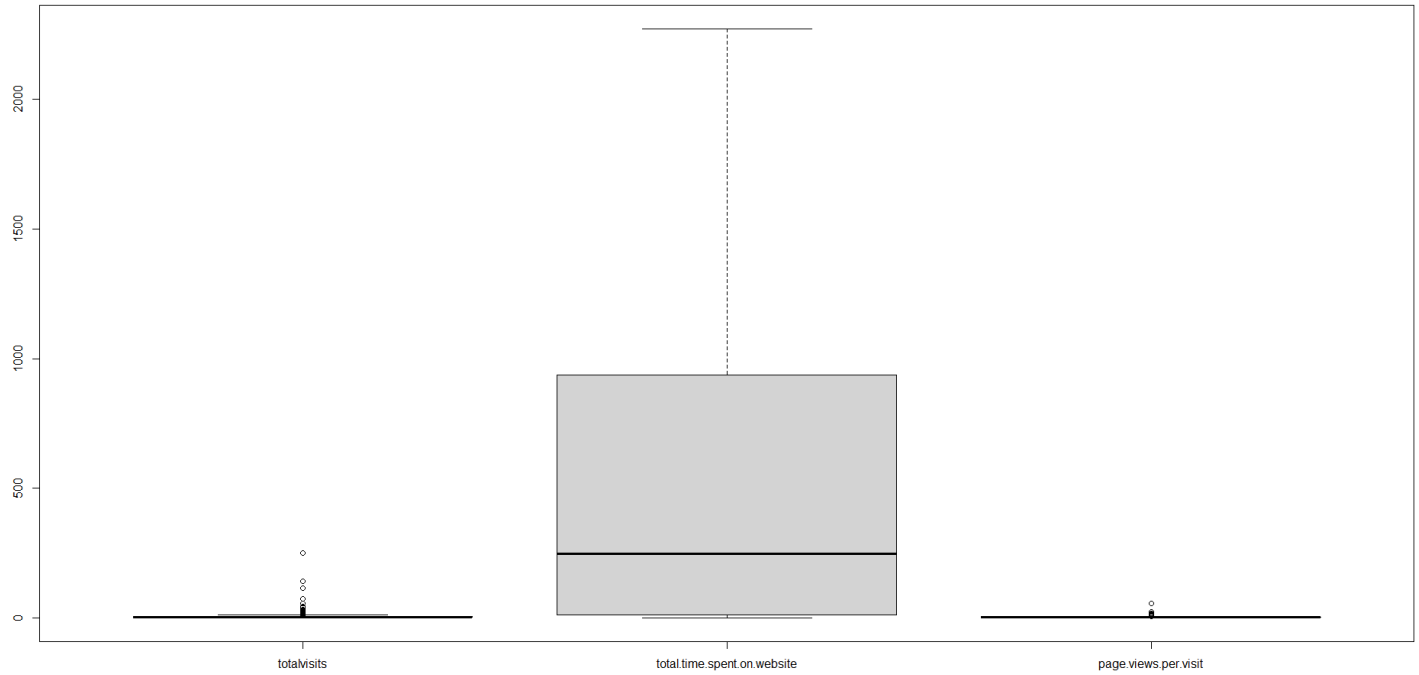


According to the comparison above, the highest amount of leads are produced by Google and Direct traffic. Additionally, there is a high conversion rate for leads and references (recommendations) obtained via the Welingak website.

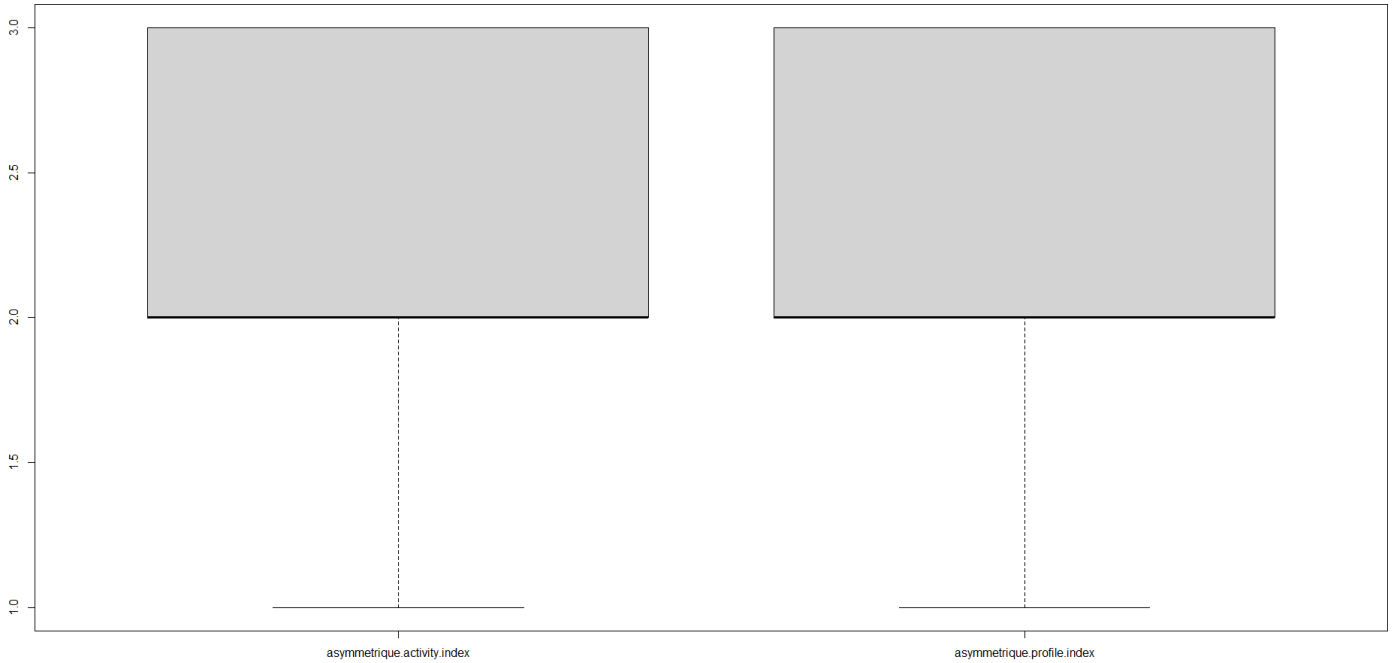
Similarly, it can be examined the influence of every feature on the lead conversion rate individually.

Finding outliers:

```
> #Plotting boxplot of continuous variable before winsorizing
> boxplot(leadsdata[c("totalvisits", "total.time.spent.on.website", "page.views.per.visit")])
> |
```




```
> boxplot(leadsdata[c("asymmetrique.activity.index","asymmetrique.profile.index")])
> |
```

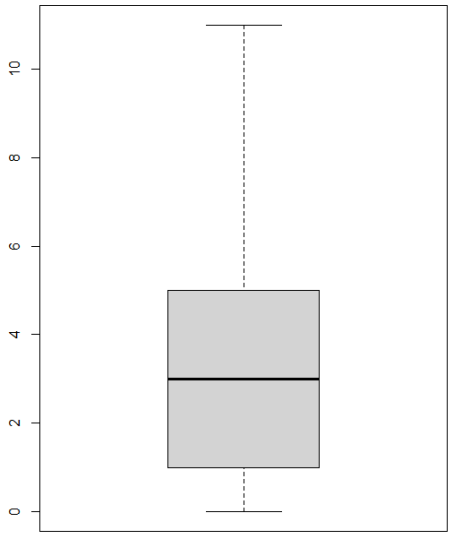


According to the above box plots, two columns have outliers., total.visits and page.views.per.visit columns.

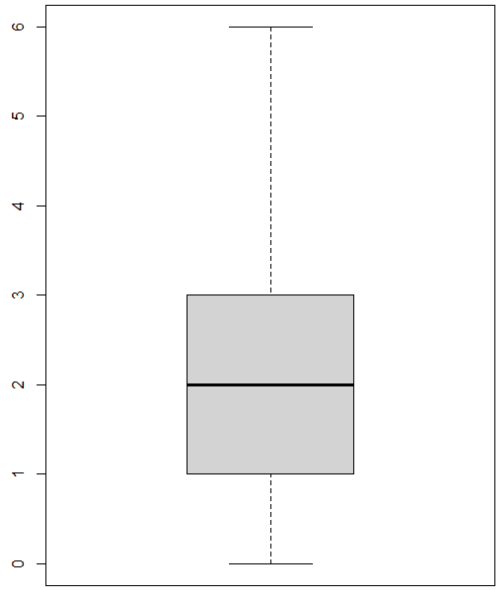
Creating a function to winsorize data.

```
> # Create a Function to Winsorize Data
> winsor <- function(x, multiplier) {
+   if(length(multiplier) != 1 || multiplier <= 0) {
+     stop("bad value for 'multiplier'")}
+
+   quartile1 = summary(x)[2] # Calculate lower quartile
+   quartile3 = summary(x)[5] # Calculate upper quartile
+   iqrange = IQR(x) # Calculate interquartile range
+
+   y <- x
+   boundary1 = quartile1 - (iqrange * multiplier)
+   boundary2 = quartile3 + (iqrange * multiplier)
+
+   y[ y < boundary1 ] <- boundary1
+   y[ y > boundary2 ] <- boundary2
+   y
+ }
>
```

```
> #Winsorizing data for total visits
> leadsdata$totalvisits <- winsor(leadsdata$totalvisits, 1.5)
> leadsdata$page.views.per.visit <- winsor(leadsdata$page.views.per.visit, 1.5)
> #Boxplot after winsorizing
> with(leadsdata, boxplot(totalvisits))
```




```
> with(leadsdata, boxplot(page.views.per.visit))
> |
```



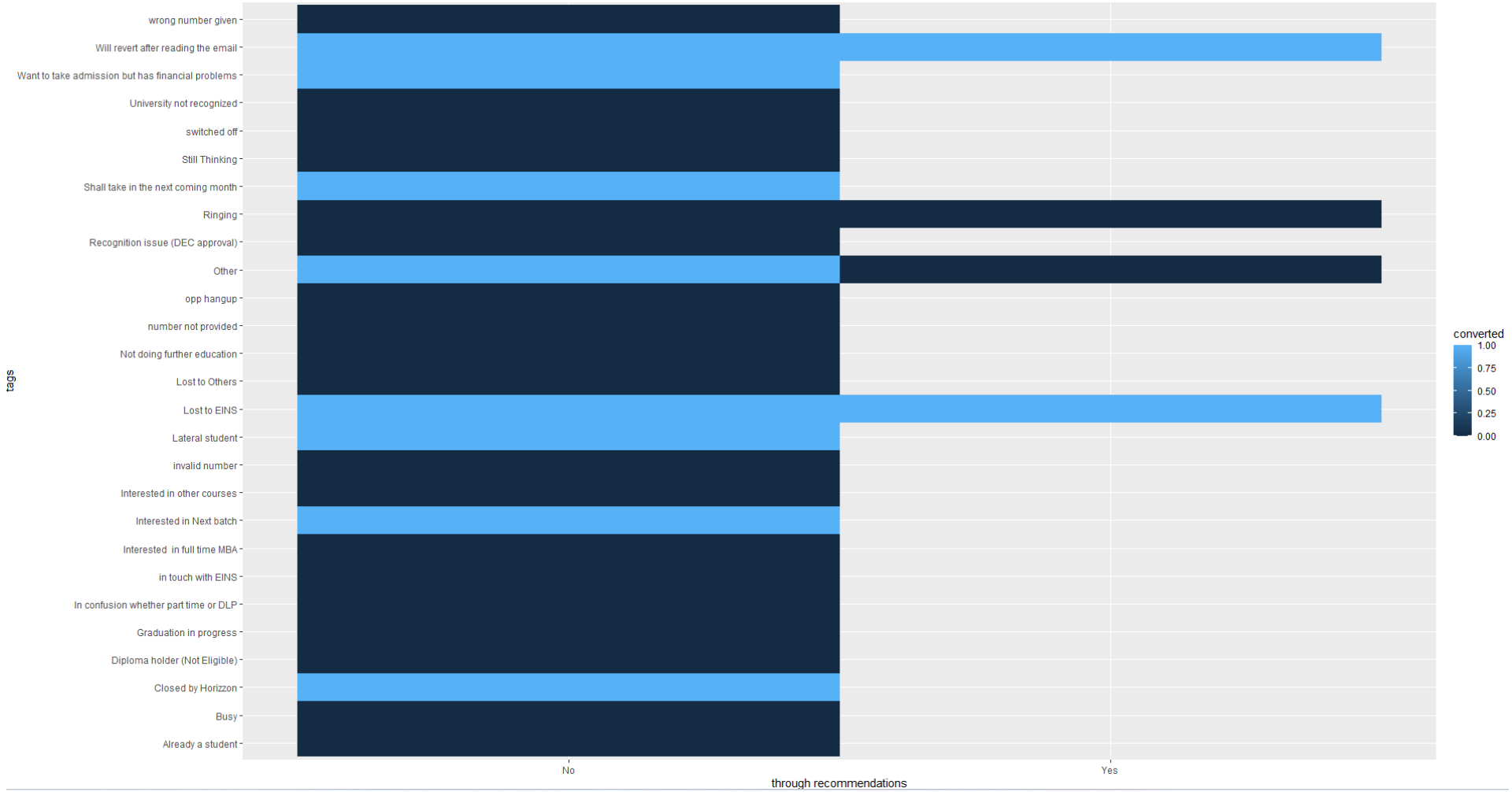
Based on the above 2 box plots, no more outliers. Heavily right skewed.

Another method to handle outliers.

```r
# Handling outliers method 2#########################

# Create box plots for the numerical columns
boxplot(leadsdata[c("totalvisits", "page.views.per.visit")])

# Define the columns for which you want to calculate the bounds
columns <- c("totalvisits", "page.views.per.visit")

# Define the threshold for outliers
outlier_threshold <- 1.5

# Create empty vectors to store the lower and upper bounds
lower_bounds <- c()
upper_bounds <- c()

# Calculate the IQR and bounds for each column using a for loop
for (column in columns) {
  column_iqr <- IQR(data[[column]])
  column_lower_bound <- quantile(data[[column]], 0.75) - outlier_threshold * column_iqr
  column_upper_bound <- quantile(data[[column]], 0.25) + outlier_threshold * column_iqr

  lower_bounds <- c(lower_bounds, column_lower_bound)
  upper_bounds <- c(upper_bounds, column_upper_bound)
}

# remove outliers for each column using a for loop
for (i in 1:length(columns)) {
  column <- columns[i]
  lower_bound <- lower_bounds[i]
  upper_bound <- upper_bounds[i]
  data <- data[data[[column]] >= lower_bound & data[[column]] <= upper_bound, ]
}

# Create a combined box plot
boxplot_data <- leadsdata[, columns, drop = FALSE]
boxplot(boxplot_data, names = columns, main = "Box Plot of Numerical Columns")
#####
```

Examining the impact on the conversion rate according to the feature column that says, "through recommendations."

```
> unique(leadsdata$through.recommendations)
[1] "No"  "Yes"
> library("ggplot2")
> mine.heatmap <- ggplot(data = leadsdata, mapping = aes(x = through.recommendations, y = tags , fill = converted)) +
+   geom_tile() + xlab(label = "through recommendations")
> mine.heatmap
>
```



It can be said that recommendations have not been that much affected on the conversion rate.

Split data to train the model.

```
> # set seeds / when splitting data the same way it will split even when the user runs this model in a future date.
> set.seed(100)
> # split the data
> split <- sample.split(leadsdata, SplitRatio = 0.8)
> split
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
> # split the dataset with respect to a single column
> # TRUE values will goes to the training dataset
> # FALSE values will goes to the test dataset
> # train the model using the train dataset
> # check the accuracy using the test dataset
> train_reg <- subset(leadsdata, split == "TRUE")
> test_reg <- subset(leadsdata, split == "FALSE")
```

According to the model, 80% of the data will be used for training and 20% will be utilized for testing.

```
> # training the model
> # Converted = binary classification / target variable
> # Totalvisits + Page.Views.Per.Visit = independent variables
> # the model calculates the probability of given values
> # pass the train dataset when building the model
> logistic_model <- glm(converted ~ totalvisits +  page.views.per.visit + total.time.spent.on.website,
+ data = train_reg, family = "binomial")
> logistic_model

Call:  glm(formula = converted ~ totalvisits + page.views.per.visit +
    total.time.spent.on.website, family = "binomial", data = train_reg)

Coefficients:
                (Intercept)                  totalvisits        page.views.per.visit  total.time.spent.on.website
                  -0.858326                     0.034045                    -0.248679                     0.001632

Degrees of Freedom: 7391 Total (i.e. Null);  7388 Residual
Null Deviance:      9846
Residual Deviance: 8752         AIC: 8760
>
```

+0.034045 is the coefficient of Total Visits.

- 0.248679 is the coefficient of Page Views Per Visit.

+0.001632 is the coefficient of Total Time Spent on Website.

Null Deviance: The model's ability to predict the response variable using just the intercept is demonstrated by the null deviation.

Residual Deviance: Once the predictors have been fitted, the residual deviance determines the deviation of the model. It represents the residual deviation following the incorporation of the predictors. The model's unacceptability is demonstrated by the substantial residual deviation.

AIC: A mathematical technique called the Akaike information criterion (AIC) is used to assess how well a model matches the data it was created from. AIC is a statistical tool used to analyze various models and identify the best fit for the data. Lower AIC values suggest better-fitting models.

Displaying the dimension of the training set and testing set.

```
> # dimension
> dim(train_reg)
[1] 7392   25
> dim(test_reg)
[1] 1848   25
>
```

Viewing the summary of the model.

```
> summary(logistic_model)

Call:
glm(formula = converted ~ totalvisits + page.views.per.visit +
    total.time.spent.on.website, family = "binomial", data = train_reg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2355  -0.8407  -0.7100   1.0296   2.1086

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -8.583e-01  4.304e-02 -19.943   <2e-16 ***
totalvisits                 3.404e-02  1.371e-02   2.484    0.013 *
page.views.per.visit       -2.487e-01  2.334e-02 -10.656   <2e-16 ***
total.time.spent.on.website  1.632e-03  5.472e-05  29.822   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9846.4  on 7391  degrees of freedom
Residual deviance: 8751.6  on 7388  degrees of freedom
AIC: 8759.6

Number of Fisher Scoring iterations: 4

> |
```

Estimate: The intercept and x slope values are represented by the estimate.

Standard error: This is the error of the intercept and slope.

Pr (>|z|): These values of page.views.per.visit and total.time.spent.on.website are very closer to zero. It suggests that these two features are more important than the others.

Number of Fisher Scoring iterations: The number of iterations is running in the background of the model when calculating coefficients.

Predicting test data.

```
> # predict test data based on model
> # use the test dataset
> predict_reg <- predict(logistic_model, test_reg, type = "response")
> predict_reg
        7        12        15        22        24        32        37        40        47        49        57        62        65        72        74
0.8003751 0.7194294 0.4284837 0.3922932 0.3872925 0.2717805 0.7170748 0.1576168 0.4025208 0.4632870 0.2202133 0.2976893 0.6513188 0.2043233 0.1442292
       82        87        90        97        99       107       112       115       122       124       132       137       140       147       149
0.3177789 0.1687551 0.6184473 0.3884768 0.4699999 0.2508695 0.5560233 0.3218804 0.2444079 0.2738306 0.3443917 0.2240888 0.1941158 0.4181041 0.2826790
      157       162       165       172       174       182       187       190       197       199       207       212       215       222       224
0.2817973 0.8342189 0.3437870 0.1387837 0.2481505 0.2976893 0.5451742 0.1797752 0.2388906 0.2976893 0.3677598 0.8991175 0.2744805 0.2947878 0.5568761
      232       237       240       247       249       257       262       265       272       274       282       287       290       297       299
0.1986859 0.2976893 0.6064408 0.1980963 0.2976893 0.2976893 0.1802990 0.3475936 0.2261143 0.3036373 0.2290854 0.2976893 0.2957873 0.1926437 0.3579066
      307       312       315       322       324       332       337       340       347       349       357       362       365       372       374
0.7379225 0.1669318 0.1758410 0.6419977 0.2976893 0.3448417 0.3151705 0.3090696 0.2091402 0.2212411 0.6625167 0.2990562 0.5640316 0.7478895 0.2976893
      382       387       390       397       399       407       412       415       422       424       432       437       440       447       449
0.2976893 0.1939160 0.2614650 0.2384374 0.2976893 0.7245684 0.3471499 0.2885415 0.2390474 0.2124675 0.2043542 0.1935409 0.4240395 0.2976893 0.2976893
      457       462       465       472       474       482       487       490       497       499       507       512       515       522       524
0.2321499 0.2685442 0.3068671 0.1338531 0.2903512 0.3621090 0.4938883 0.2976893 0.1525857 0.2037933 0.1877500 0.2688467 0.2976893 0.1724042 0.7396826
      532       537       540       547       549       557       562       565       572       574       582       587       590       597       599
0.2976893 0.8091070 0.1565243 0.6039948 0.7844008 0.6806865 0.7641625 0.6470555 0.2664749 0.2764790 0.3516011 0.2976893 0.2469686 0.7850594 0.5800101
      607       612       615       622       624       632       637       640       647       649       657       662       665       672       674
0.1850279 0.1838007 0.3080050 0.1088737 0.2976893 0.8199040 0.2873799 0.8649182 0.1155368 0.1927088 0.2421696 0.7610690 0.2197295 0.4324845 0.6916194
      682       687       690       697       699       707       712       715       722       724       732       737       740       747       749
0.3069827 0.6254211 0.2727317 0.2733795 0.2644678 0.1802706 0.2683749 0.1591394 0.2255268 0.3218724 0.3245881 0.7462268 0.7984049 0.2976893 0.3315409
      757       762       765       772       774       782       787       790       797       799       807       812       815       822       824
0.2182025 0.2227517 0.1628868 0.1158707 0.3347054 0.1260058 0.6079976 0.5726518 0.3350478 0.1911255 0.1640024 0.3581494 0.1488263 0.3290818 0.2068628
      832       837       840       847       849       857       862       865       872       874       882       887       890       897       899
0.6412238 0.7150568 0.2976893 0.2976893 0.7702823 0.2360916 0.2743530 0.2976893 0.2111640 0.3093980 0.4535869 0.4242817 0.4198805 0.2637319 0.2650329
```

```
> test_reg
                lead.origin      lead.source do.not.email do.not.call converted totalvisits total.time.spent.on.website page.views.per.visit
7   Landing Page Submission           Google           No          No         1    2.000000                       1640              2.00000
12  Landing Page Submission   Direct Traffic           No          No         1    8.000000                       1343              2.67000
15  Landing Page Submission   Direct Traffic          Yes          No         0    1.000000                        481              1.00000
22                      API           Google           No          No         0    4.000000                        377              1.33000
24  Landing Page Submission           Google           No          No         0    4.000000                        771              4.00000
32                      API           Google           No          No         0    3.000000                         88              1.50000
37  Landing Page Submission           Google           No          No         0    4.000000                       1622              4.00000
40  Landing Page Submission           Google           No          No         1    4.000000                         25              4.00000
47  Landing Page Submission   Direct Traffic           No          No         0    2.000000                        547              2.00000
49                      API           Google           No          No         0    6.000000                       1225              6.00000
57  Landing Page Submission   Referral Sites           No          No         1   11.000000                        436              6.00000
62                      API       Olark Chat           No          No         0    0.000000                          0              0.00000
65  Landing Page Submission   Direct Traffic           No          No         1    4.000000                       1435              4.00000
72  Landing Page Submission   Direct Traffic           No          No         0    4.000000                        219              4.00000
74                      API   Organic Search           No          No         0    6.000000                        224              6.00000
82            Lead Add Form Welingak Website           No          No         1    3.445238                        346              2.36282
87                      API   Referral Sites           No          No         0    4.000000                         75              4.00000
90  Landing Page Submission   Organic Search           No          No         0    2.000000                       1085              2.00000
97  Landing Page Submission           Google           No          No         0    2.000000                        511              2.00000
99                      API           Google           No          No         1    5.000000                       1110              5.00000
107                     API           Google           No          No         0    9.000000                        125              3.00000
112 Landing Page Submission           Google           No          No         1    4.000000                       1190              4.00000
115                     API           Google           No          No         1    5.000000                        727              5.00000
122 Landing Page Submission           Google           No          No         0    8.000000                        277              4.00000
124                     API           Google           No          No         0    2.000000                         39              1.00000
132                     API           Google           No          No         0    1.000000                        263              1.00000
137 Landing Page Submission   Direct Traffic           No          No         0    4.000000                        291              4.00000
140 Landing Page Submission   Referral Sites           No          No         0    7.000000                         41              3.50000
147 Landing Page Submission   Direct Traffic           No          No         0    3.000000                        718              3.00000
149 Landing Page Submission           Google           No          No         0    5.000000                        232              2.50000
157                     API   Referral Sites           No          No         0    4.000000                        174              2.00000
162                     API       Olark Chat           No          No         1   10.000000                       1815              3.33000
165 Landing Page Submission           Google           No          No         1    6.000000                        919              6.00000
172 Landing Page Submission           Google           No          No         0    5.000000                         65              5.00000
174                     API           Google           No          No         0    7.000000                        615              6.00000
182                     API       Olark Chat           No          No         0    0.000000                          0              0.00000
187                     API   Organic Search           No          No         1    8.000000                        877              2.67000
```

```
> length(predict_reg)
[1] 1848
> length(test_reg)
[1] 25
>
```

Changing probabilities.

```
> # changing probabilities
> # if the prediction value is greater than 0.5 then the class is 1, otherwise the class is 0
> predict_reg <- ifelse(predict_reg > 0.5, 1, 0)
> predict_reg
   7   12   15   22   24   32   37   40   47   49   57   62   65   72   74   82   87   90   97   99  107  112  115  122  124  132  137  140  147  149  157
   1    1    0    0    0    0    1    0    0    0    0    0    1    0    0    0    0    1    0    0    0    1    0    0    0    0    0    0    0    0    0
 162  165  172  174  182  187  190  197  199  207  212  215  222  224  232  237  240  247  249  257  262  265  272  274  282  287  290  297  299  307  312
   1    0    0    0    0    1    0    0    0    0    1    0    0    1    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    1    0
 315  322  324  332  337  340  347  349  357  362  365  372  374  382  387  390  397  399  407  412  415  422  424  432  437  440  447  449  457  462  465
   0    1    0    0    0    0    0    0    1    0    1    1    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0
 472  474  482  487  490  497  499  507  512  515  522  524  532  537  540  547  549  557  562  565  572  574  582  587  590  597  599  607  612  615  622
   0    0    0    0    0    0    0    0    0    0    0    1    0    1    0    1    1    1    1    1    0    0    0    0    0    1    1    0    0    0    0
 624  632  637  640  647  649  657  662  665  672  674  682  687  690  697  699  707  712  715  722  724  732  737  740  747  749  757  762  765  772  774
   0    1    0    1    0    0    0    1    0    0    1    0    1    0    0    0    0    0    0    0    0    0    1    1    0    0    0    0    0    0    0
 782  787  790  797  799  807  812  815  822  824  832  837  840  847  849  857  862  865  872  874  882  887  890  897  899  907  912  915  922  924  932
   0    1    1    0    0    0    0    0    0    0    1    1    0    0    1    0    0    0    0    0    0    0    0    0    0    0    1    0    0    1    0
 937  940  947  949  957  962  965  972  974  982  987  990  997  999 1007 1012 1015 1022 1024 1032 1037 1040 1047 1049 1057 1062 1065 1072 1074 1082 1087
   1    1    0    1    0    0    1    1    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
1090 1097 1099 1107 1112 1115 1122 1124 1132 1137 1140 1147 1149 1157 1162 1165 1172 1174 1182 1187 1190 1197 1199 1207 1212 1215 1222 1224 1232 1237 1240
   1    0    1    0    0    0    1    0    0    0    1    1    0    0    0    1    1    0    0    1    1    0    0    0    0    1    0    1    0    0    0
1247 1249 1257 1262 1265 1272 1274 1282 1287 1290 1297 1299 1307 1312 1315 1322 1324 1332 1337 1340 1347 1349 1357 1362 1365 1372 1374 1382 1387 1390 1397
   0    1    0    0    1    0    0    1    1    1    1    0    0    1    1    0    0    1    0    1    0    0    0    0    1    1    0    1    0    0    1
1399 1407 1412 1415 1422 1424 1432 1437 1440 1447 1449 1457 1462 1465 1472 1474 1482 1487 1490 1497 1499 1507 1512 1515 1522 1524 1532 1537 1540 1547 1549
   1    0    1    0    0    1    0    1    0    1    0    1    0    0    1    1    1    0    1    0    0    1    1    0    0    1    0    0    0    1    1
1557 1562 1565 1572 1574 1582 1587 1590 1597 1599 1607 1612 1615 1622 1624 1632 1637 1640 1647 1649 1657 1662 1665 1672 1674 1682 1687 1690 1697 1699 1707
```

Evaluating the model and generating the confusion matrix.

```
> # evaluating model accuracy
> # use confusion matrix
> test_reg$converted
   [1] 1 1 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 1 0 1
  [76] 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 0 0 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 1 0 1 1 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0
 [151] 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 1 0 0 0 0 0 1 1 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 1 1 0 1 0 0 0 1 0 1 1 0 0 0 0
 [226] 0 0 1 1 1 0 0 1 1 1 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 0 1 0 1 0 0 0 0 1 1 0 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 0 1 0 0 0 1 1 0 0 0 0
 [301] 1 1 0 1 0 0 1 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 0 0 0
 [376] 1 1 0 0 1 0 0 0 0 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 0 1 1 0 0 1 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 0 1 1 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 1 1 0 1 1 0
 [451] 1 0 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 1 1 0 1 1 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 1 1 0 1 1 1 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1
 [526] 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 0 1 0 0 0 1 1 1 0 1 0 1 0 1 1 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 1 1 1 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0
 [601] 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 0 0 0 1 1 1 1 1 0 0 1 1 0 0 1 0 0 0 0 0 0 1
 [676] 1 0 0 1 0 1 1 0 1 0 0 1 1 1 1 1 1 0 1 1 0 1 0 1 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 1 1 1 1 1 1 0 0
 [751] 0 1 0 0 0 1 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 1 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1 0 1
 [826] 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 0 1 1 0 0 1 0 0 1 0 1 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1 1
 [901] 0 0 0 0 0 1 1 0 1 1 1 1 1 0 1 0 1 1 1 1 1 0 0 1 0 1 0 1 0 0 0 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 1 1 1 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0 1 1 0 0 1
 [976] 1 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0
 [ reached getOption("max.print") -- omitted 848 entries ]
> table(test_reg$converted, predict_reg)
   predict_reg
      0    1
  0 1008  118
  1  378  344
> |
```

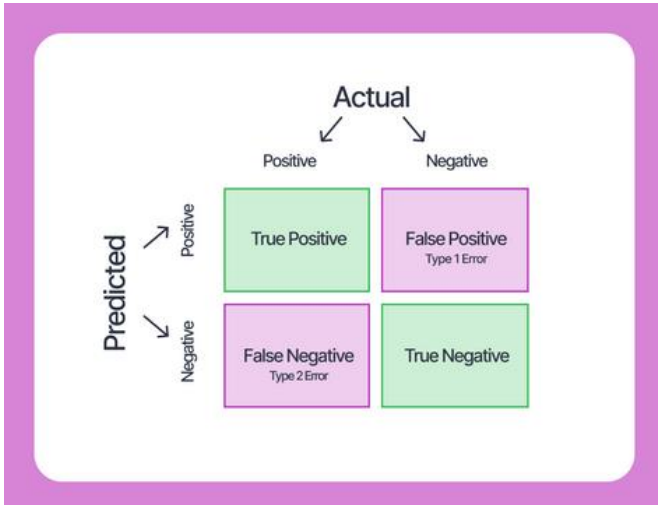
The model predicted 1008 times as 0 and actual also 0.

The model predicted 378 times as 1 but actual 0.

The model predicted 118 times as 0 and actual 1.

The model predicted 344 times as 1 and actual also 1.

Confusion Matrix:



Source: www. v7labs.com

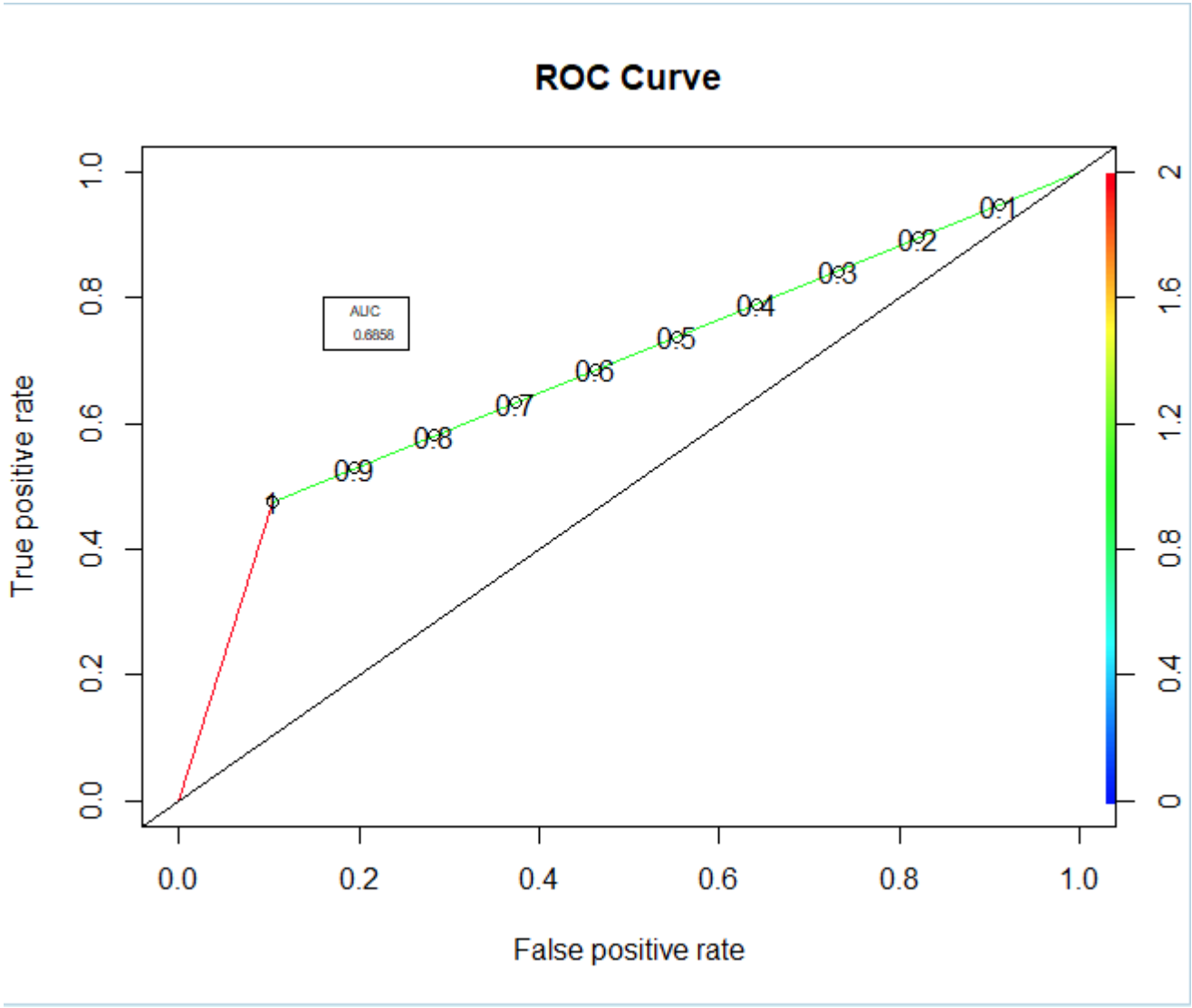
```
> missing_classerr <- mean(predict_reg != test_reg$converted)
> print(paste('Accuracy = ', 1 - missing_classerr))
[1] "Accuracy =  0.731601731601732"
>
```

The accuracy of the model: approximately 73%

Plotting ROC curve & finding AUC value.

```
> # ROC-AUC curve
> # true positive vs false positive
> ROCPred <- prediction(predict_reg, test_reg$converted)
> # tpr = true positive rate (y axis)
> # fpr = false positive rate (x axis)
> ROCPer <- performance(ROCPred, measure = "tpr", x.measure = "fpr")
> # auc = area under the curve
> auc <- performance(ROCPred, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.6858293
> # rounding to 4 decimal places
> auc <- round(auc, 4)
> auc
[1] 0.6858
>
```

```
> # plot the curve
> plot(ROCPer)
> # add colors and cutoff points
> plot(ROCPer, colorize = TRUE, print.cutoffs.at = seq(0.1, by = 0.1), main = "ROC Curve")
> abline(a = 0, b = 1)
> legend(.16, .8, auc, title = "AUC", cex = 0.5)
>
```



ROC Curve: Receiver Operating Characteristic curve is a graph that displays a classification model's performance across all classification. The model works best when the curve is distant from the diagonal.

This curve plots two parameters:

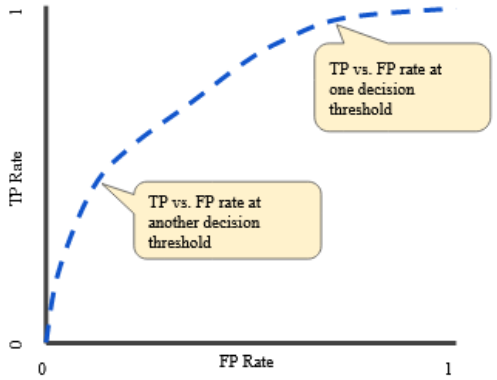
True Positive Rate: Sensitivity

$$TPR = \frac{TP}{TP + FN}$$

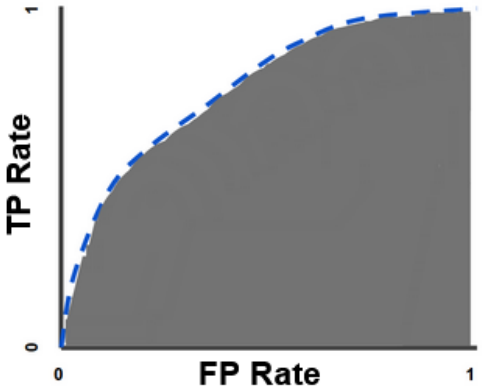
False Positive Rate: 1 - Specificity

$$FPR = \frac{FP}{FP + TN}$$

ROC Curve: Receiver Operating Characteristic curve



AUC: Area Under the ROC Curve



Source: https://developers.google.com

Applying grades based on predicted values.

```
summary(predict_reg)
#min = 0.1061
#max = 0.9221
```

```
> summary(predict_reg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1061  0.2682  0.2969  0.3818  0.4960  0.9221
>
```

Defining a function to assign grades based on some predefined value ranges.

```r
# Function to assign grades based on predicted values
assign_grade <- function(predict_reg) {
  if (predict_reg >= 0.9999) {
    return("A")
  } else if (predict_reg >= 0.7500) {
    return("B")
  } else if (predict_reg >= 0.5000) {
    return("C")
  } else if (predict_reg >= 0.2500) {
    return("D")
  } else if (predict_reg >= 0.0001) {
    return("E")
  } else {
    return("F")
  }
}
```

```
> # Function to assign grades based on predicted values
> assign_grade <- function(predict_reg) {
+   if (predict_reg >= 0.9999) {
+     return("A")
+   } else if (predict_reg >= 0.7500) {
+     return("B")
+   } else if (predict_reg >= 0.5000) {
+     return("C")
+   } else if (predict_reg >= 0.2500) {
+     return("D")
+   } else if (predict_reg >= 0.0001) {
+     return("E")
+   } else {
+     return("F")
+   }
+ }
>
```

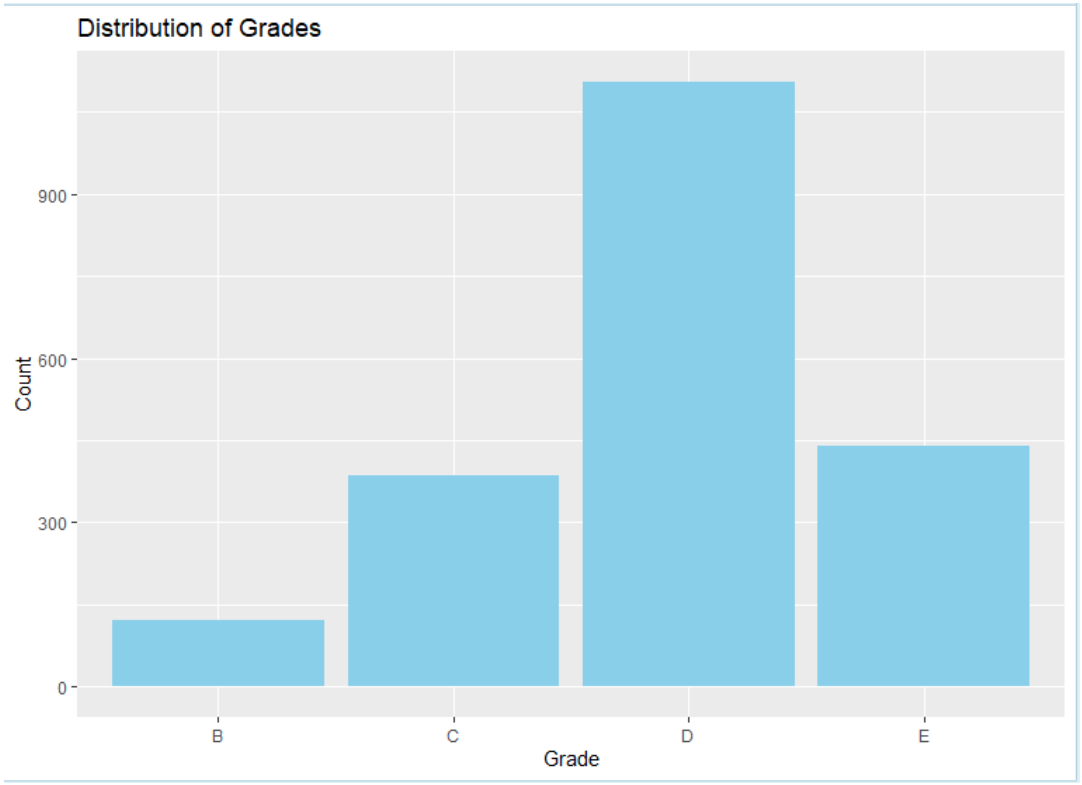
Applying the function to the predicted values.

```r
# Applying the function to the predicted values
grades <- sapply(predict_reg, assign_grade)

# Displaying the result
result <- data.frame(Prediction = predict_reg, Grade = grades)
print(result)
```

```
> # Applying the function to the predicted values
> grades <- sapply(predict_reg, assign_grade)
> # Displaying the result
> result <- data.frame(Prediction = predict_reg, Grade = grades)
> print(result)
     Prediction Grade
7     0.8059069     B
12    0.7263248     C
15    0.4302527     D
22    0.3942156     D
24    0.3899423     D
27    0.6068840     C
34    0.2670208     D
39    0.3143507     D
42    0.1132508     E
49    0.4683301     D
51    0.2969239     D
54    0.3653618     D
61    0.2022928     E
66    0.3062713     D
69    0.2969239     D
```

Visualizing distribution of grades.

```r
# drawing a histogram
unique(result$Grade)
ggplot(result, aes(x = Grade)) + geom_bar(fill = "skyblue") + labs(title = "Distribution of Grades", x = "Grade", y = "Count")
```



As per the histogram, grade D has a high frequency.

References

Alex. (2019, Jun 1). *Linear Regression Summary(lm): Interpretting in R*. Retrieved from boostedml.com: https://boostedml.com/2019/06/linear-regression-in-r-interpreting-summarylm.html

amazon.com. (n.d.). *What's the Difference Between Linear Regression and Logistic Regression?* Retrieved from amazon.com: https://aws.amazon.com/compare/the-difference-between-linear-regression-and-logistic-regression/

Arvind Shukla. (2023, Aug 7). *Neural Networks are Decision Trees*. Retrieved from www.linkedin.com: https://www.linkedin.com/pulse/neural-networks-decision-trees-arvind-shukla/

developers.google.com. (n.d.). *Classification: ROC Curve and AUC*. Retrieved from developers.google.com: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Dr. Osman Dag. (2022, Mar 4). *How to Remove Outliers from Data in R* . Retrieved from universeofdatascience.com: https://universeofdatascience.com/how-to-remove-outliers-from-data-in-r/

Gustav Willig. (2023, Jan 17). *Decision Tree vs Logistic Regression*. Retrieved from gustavwillig.medium.com: https://gustavwillig.medium.com/decision-tree-vs-logistic-regression-1a40c58307d0

John . (2020, Jan 19). *How to Remove Outliers in R*. Retrieved from www.r-bloggers.com: https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/

makemeanalyst.com. (n.d.). *Normal Probability Plot in R using ggplot2*. Retrieved from makemeanalyst.com: https://makemeanalyst.com/statistics-with-r/normal-probability-plot-in-r-using-ggplot2/

methodenlehre.github.io. (n.d.). *Graphics with ggplot2*. Retrieved from methodenlehre.github.io: https://methodenlehre.github.io/SGSCLM-R-course/graphics-with-ggplot2.html

Niam Zaki Zamani. (2021, Jan 18). *Linear Regression on Student Grade Prediction*. Retrieved from rpubs.com: https://rpubs.com/niamzaki/student_grade_prediction

Peter Bruce, Andrew Bruce. (n.d.). *Chapter 4. Regression and Prediction*. Retrieved from www.oreilly.com: https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html

Rebecca C. Steorts. (n.d.). *Comparison of Linear Regression with K-Nearest*. Retrieved from www2.stat.duke.edu: https://www2.stat.duke.edu/~rcs46/lectures_2017/03-lr/03-knn.pdf

Rohit Kundu. (2022, Sep 13). *Confusion Matrix: How To Use It & Interpret Results [Examples]*. Retrieved from www.v7labs.com: https://www.v7labs.com/blog/confusion-matrix-guide

Safa Mulani. (2022, Aug 3). *Outlier Analysis in R - Detect and Remove Outliers*. Retrieved from www.digitalocean.com: https://www.digitalocean.com/community/tutorials/outlier-analysis-in-r

stackoverflow.com. (2023, May 3). *Difference between Logistic Regression and Decision Trees*. Retrieved from stackoverflow.com: https://stackoverflow.com/questions/76161673/difference-between-logistic-regression-and-decision-trees

typeset.io. (n.d.). *What is the difference between KNN regression and linear regression?* . Retrieved from typeset.io: https://typeset.io/questions/what-is-the-difference-between-knn-regression-and-linear-1pad331c0a

www.ibm.com. (n.d.). *What is logistic regression?* . Retrieved from www.ibm.com: https://www.ibm.com/topics/logistic-regression

www.rdocumentation.org. (n.d.). *plot: Generic X-Y Plotting*. Retrieved from www.rdocumentation.org: https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/plot

www.statisticssolutions.com. (n.d.). *Correlation (Pearson, Kendall, Spearman)*. Retrieved from www.statisticssolutions.com: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

www.sthda.com. (n.d.). *QQ-plots: Quantile-Quantile plots - R Base Graphs* . Retrieved from www.sthda.com: http://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs