project1

Maria Garcia

2024-06-07

##Importing the data

We will be using a data set from Kaggle that contains app data from Waze.

```
waze_data <- read_csv("/Users/mariagarcia/Desktop/DAT301/waze_app_dataset.csv")
```

```
## Rows: 14999 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): label, device
## dbl (11): ID, sessions, drives, total_sessions, n_days_after_onboarding, tot...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(waze_data)
```

```
## spc_tbl_ [14,999 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ID                      : num [1:14999] 0 1 2 3 4 5 6 7 8 9 ...
##  $ label                   : chr [1:14999] "retained" "retained" "retained" "retained" ...
##  $ sessions                : num [1:14999] 283 133 114 49 84 113 3 39 57 84 ...
##  $ drives                  : num [1:14999] 226 107 95 40 68 103 2 35 46 68 ...
##  $ total_sessions          : num [1:14999] 296.7 326.9 135.5 67.6 168.2 ...
##  $ n_days_after_onboarding: num [1:14999] 2276 1225 2651 15 1562 ...
##  $ total_navigations_fav1 : num [1:14999] 208 19 0 322 166 0 185 0 0 72 ...
##  $ total_navigations_fav2 : num [1:14999] 0 64 0 7 5 0 18 0 26 0 ...
##  $ driven_km_drives       : num [1:14999] 2629 13716 3059 914 3950 ...
##  $ duration_minutes_drives: num [1:14999] 1986 3160 1611 587 1220 ...
##  $ activity_days          : num [1:14999] 28 13 14 7 27 15 28 22 25 7 ...
##  $ driving_days           : num [1:14999] 19 11 8 3 18 11 23 20 20 3 ...
##  $ device                  : chr [1:14999] "Android" "iPhone" "Android" "iPhone" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ID = col_double(),
##   ..   label = col_character(),
##   ..   sessions = col_double(),
##   ..   drives = col_double(),
##   ..   total_sessions = col_double(),
##   ..   n_days_after_onboarding = col_double(),
##   ..   total_navigations_fav1 = col_double(),
##   ..   total_navigations_fav2 = col_double(),
##   ..   driven_km_drives = col_double(),
##   ..   duration_minutes_drives = col_double(),
##   ..   activity_days = col_double(),
```

```
##   ..   driving_days = col_double(),
##   ..   device = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(waze_data)
```

```
## # A tibble: 6 x 13
##      ID label    sessions drives total_sessions n_days_after_onboarding
##   <dbl> <chr>       <dbl>  <dbl>          <dbl>                   <dbl>
## 1     0 retained      283    226          297.                     2276
## 2     1 retained      133    107          327.                     1225
## 3     2 retained      114     95          136.                     2651
## 4     3 retained       49     40           67.6                      15
## 5     4 retained       84     68          168.                     1562
## 6     5 retained      113    103          280.                     2637
## # i 7 more variables: total_navigations_fav1 <dbl>,
## #   total_navigations_fav2 <dbl>, driven_km_drives <dbl>,
## #   duration_minutes_drives <dbl>, activity_days <dbl>, driving_days <dbl>,
## #   device <chr>
```

```
#summary(waze_data)
```

##Preprocessing 1. rename "lablel" column to "customer_status" (more meaningful) 2. remove any blank values for our target variables (customer_status)

```
colnames(waze_data)[colnames(waze_data) == "label"] <- "customer_status"
colnames(waze_data)
```

```
##  [1] "ID"                      "customer_status"
##  [3] "sessions"                "drives"
##  [5] "total_sessions"          "n_days_after_onboarding"
##  [7] "total_navigations_fav1"  "total_navigations_fav2"
##  [9] "driven_km_drives"        "duration_minutes_drives"
## [11] "activity_days"           "driving_days"
## [13] "device"
```

```
sapply(waze_data, function(x) sum(is.na(x)))
```

```
##                      ID         customer_status                sessions
##                       0                     700                       0
##                  drives          total_sessions n_days_after_onboarding
##                       0                       0                       0
##  total_navigations_fav1  total_navigations_fav2        driven_km_drives
##                       0                       0                       0
## duration_minutes_drives           activity_days            driving_days
##                       0                       0                       0
##                  device
##                       0
```

```
nrow(waze_data)
```

```
## [1] 14999
```

```
waze_data_cleaned <- waze_data[!is.na(waze_data$customer_status), ]
#waze_data_cleaned <- na.omit(waze_data)
sapply(waze_data_cleaned, function(x) sum(is.na(x)))
```

```
##                      ID         customer_status                sessions
##                       0                       0                       0
##                  drives          total_sessions n_days_after_onboarding
##                       0                       0                       0
##  total_navigations_fav1  total_navigations_fav2        driven_km_drives
##                       0                       0                       0
## duration_minutes_drives           activity_days            driving_days
##                       0                       0                       0
##                  device
##                       0
```

```
nrow(waze_data_cleaned)
```

```
## [1] 14299
```

```
write.csv(waze_data_cleaned, "waze_data_cleaned.csv", row.names = FALSE)
#waze_data_cleaned <- na.omit(waze_data)
```

##Now we will create a new dataset to further review retained vs churned customers Explore the distrobution of retrained vs. churned customers

```
churn_data <- waze_data_cleaned[!is.na(waze_data_cleaned$customer_status), ]
table(waze_data_cleaned$customer_status)
```

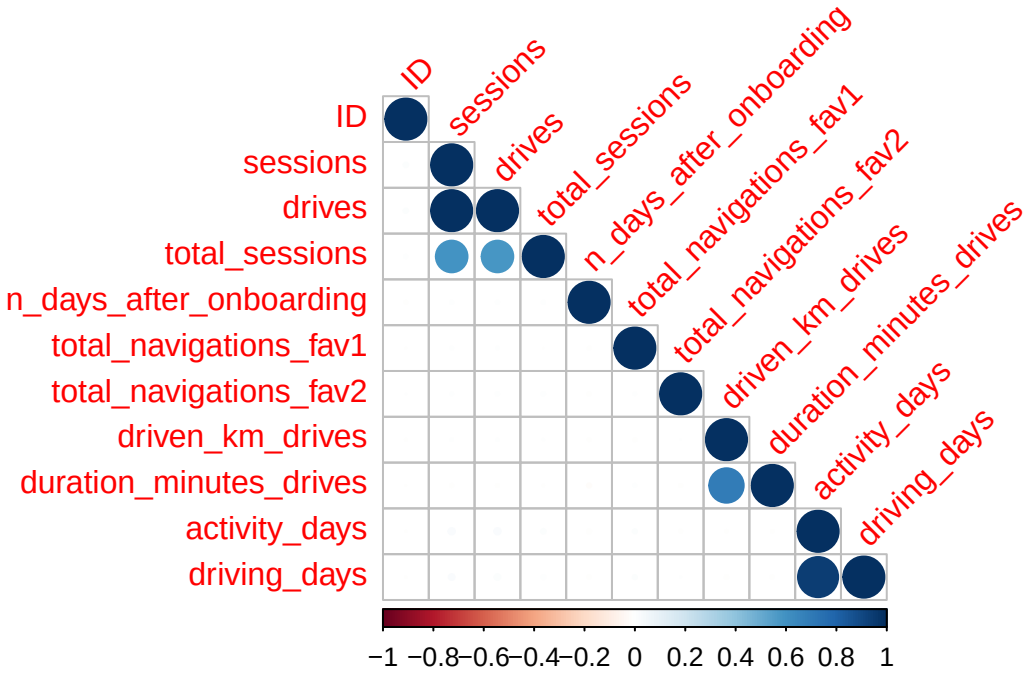
```
##
##  churned retained
##     2536    11763
```

```
prop.table(table(waze_data_cleaned$customer_status))
```

```
##
##   churned  retained
## 0.1773551 0.8226449
```

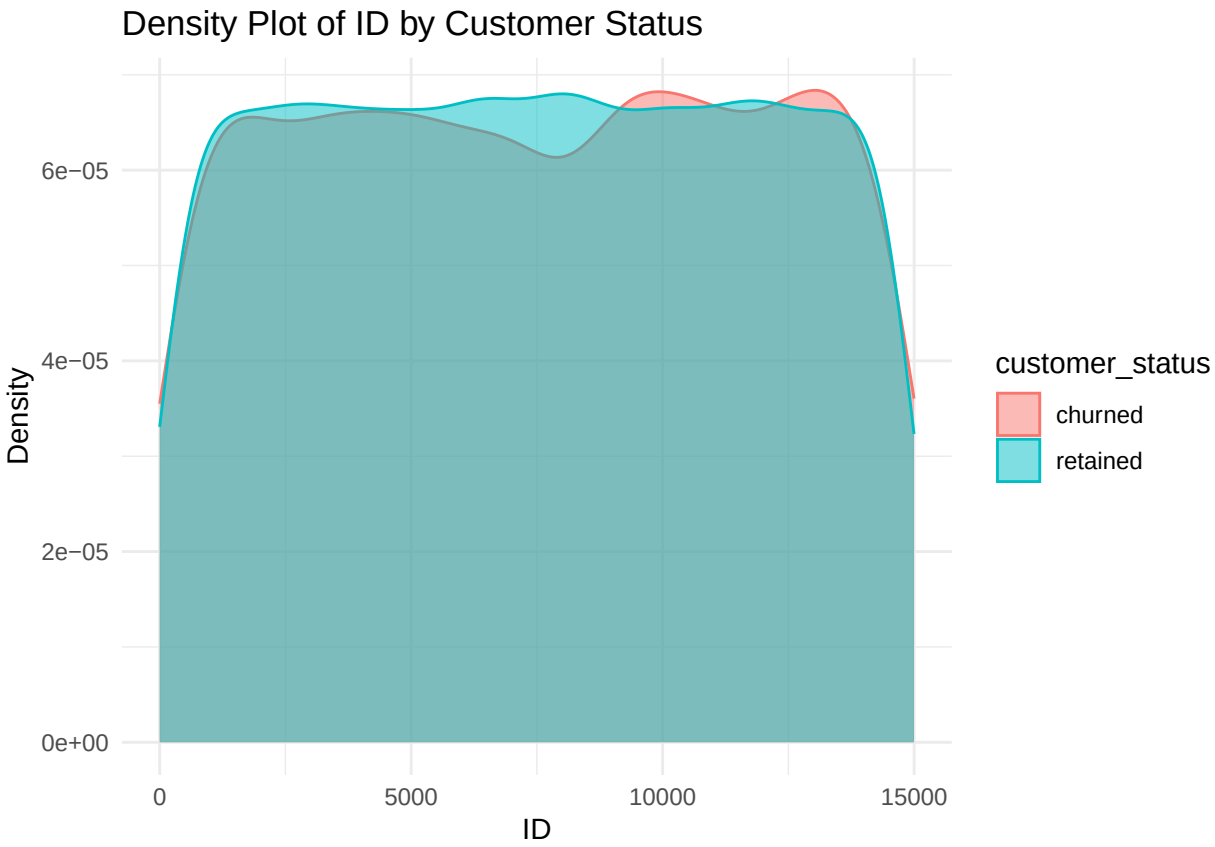
##Correlation? We know there is a significant size difference in the population of churned vs. retained. Can we use a heat map to identify any correlation with the other 12 variables?

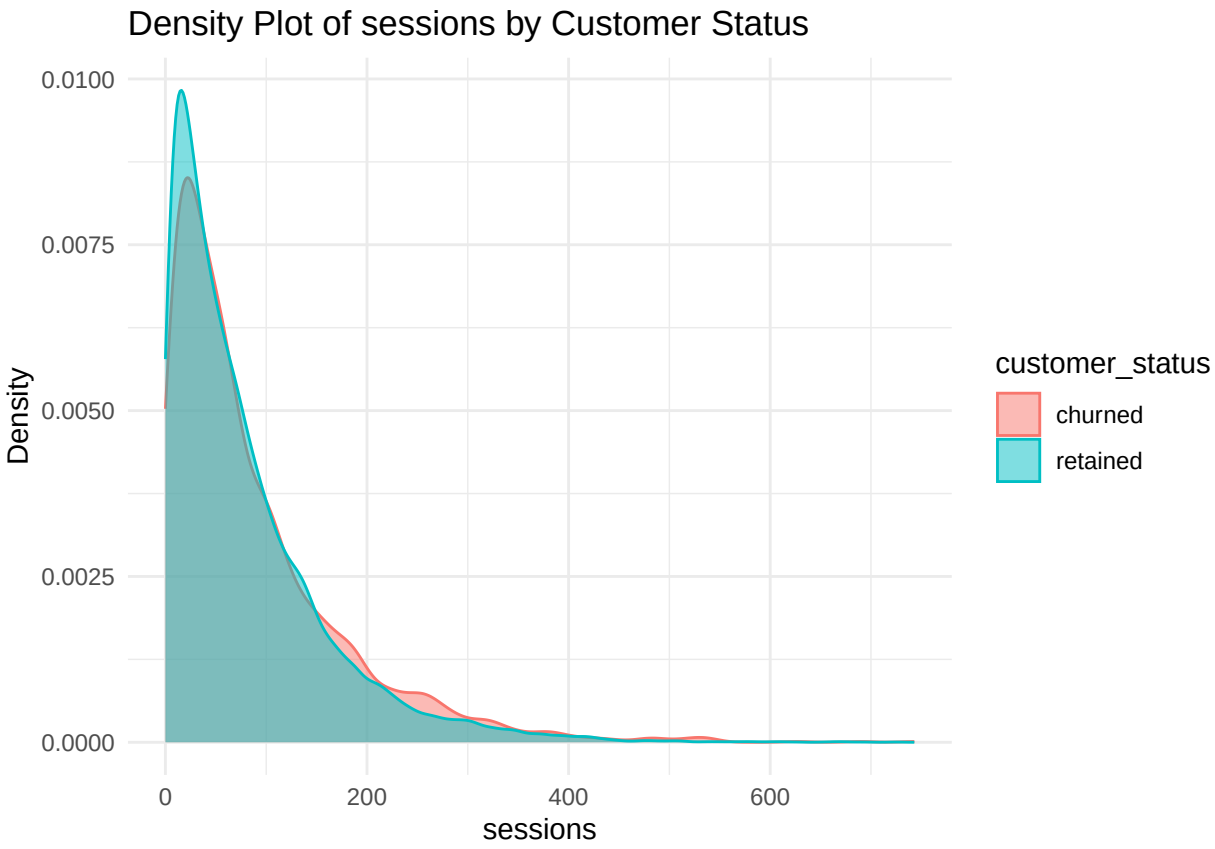
```
churn_data$customer_status <- as.factor(churn_data$customer_status)
numeric_cols <- unlist(lapply(churn_data, is.numeric))
churn_numeric <- churn_data[, numeric_cols]
cor_matrix <- cor(churn_numeric)
corrplot(cor_matrix, method = "circle", type = "lower", tl.coll = "black", tl.srt = 45)
```

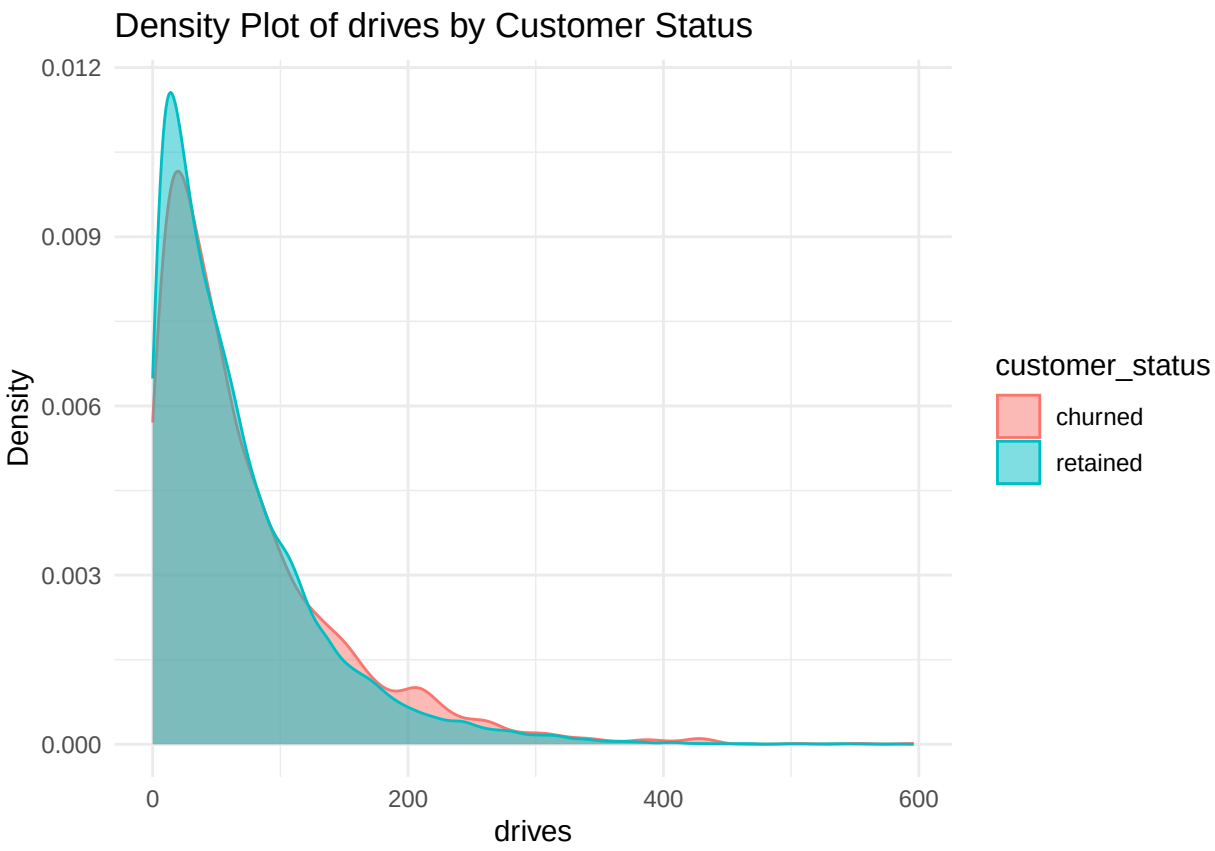


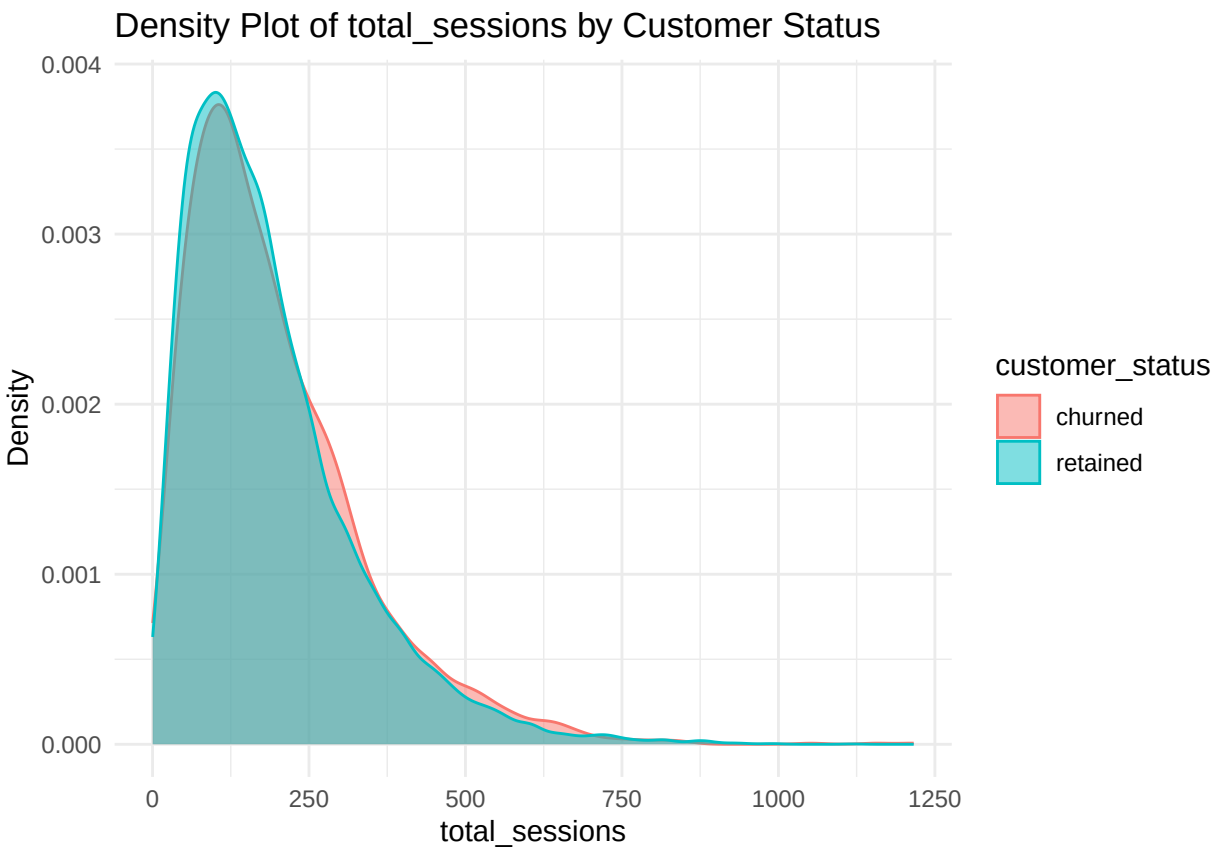
##Visualisizing correlations using density plots

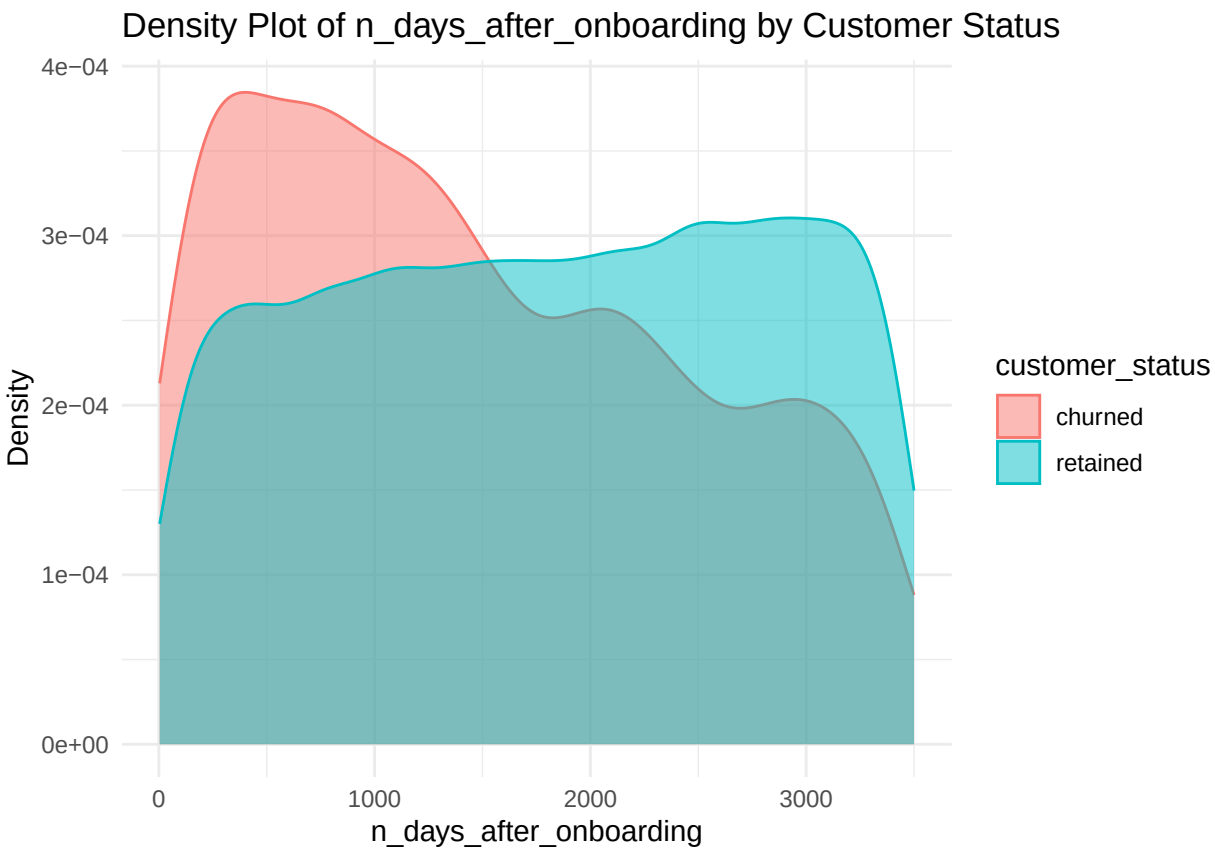
```
numerical_features <- waze_data_cleaned %>% select_if(is.numeric) %>% names()
for (feature in numerical_features) {
  p <- ggplot(waze_data_cleaned, aes(x = !!sym(feature), color = customer_status, fill = customer_statu
    geom_density(alpha = 0.5) +
    labs(title = paste("Density Plot of", feature, "by Customer Status"),
         x = feature,
         y = "Density") +
    theme_minimal()
  print(p)
}
```

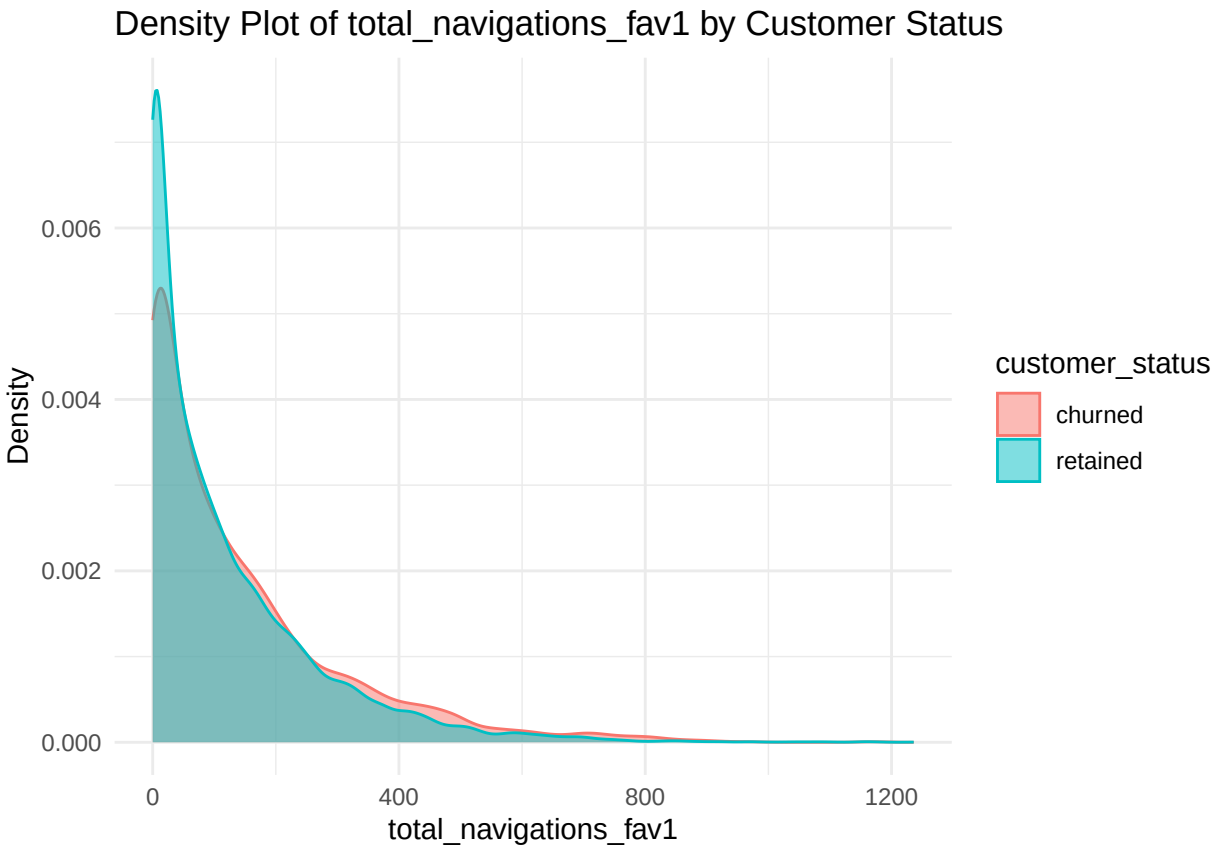

Density Plot of ID by Customer Status
Density
6e−05
4e−05
2e−05
0e+00
0
5000
10000
15000
ID
customer_status
churned
retained

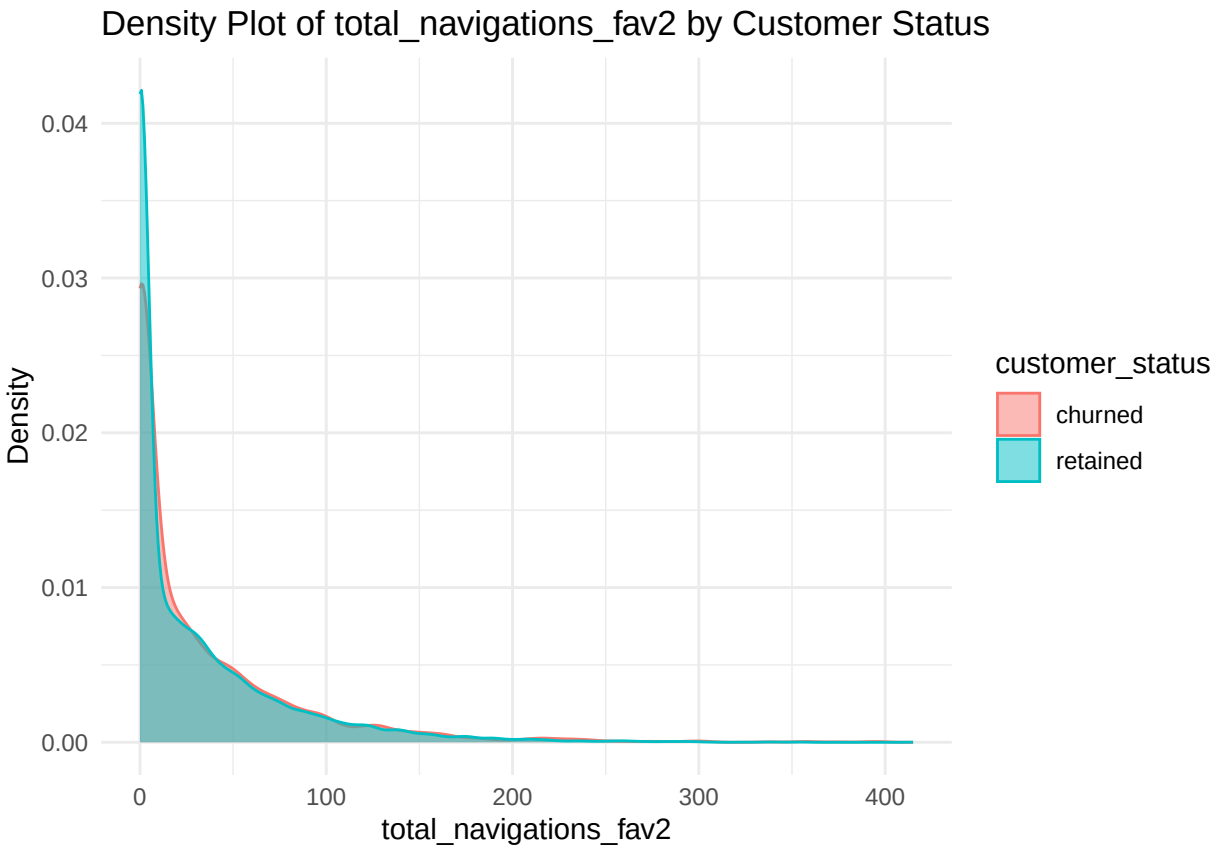

Density Plot of sessions by Customer Status
0.0100
0.0075
0.0050
0.0025
0.0000
Density
0
200
400
600
sessions
customer_status
churned
retained

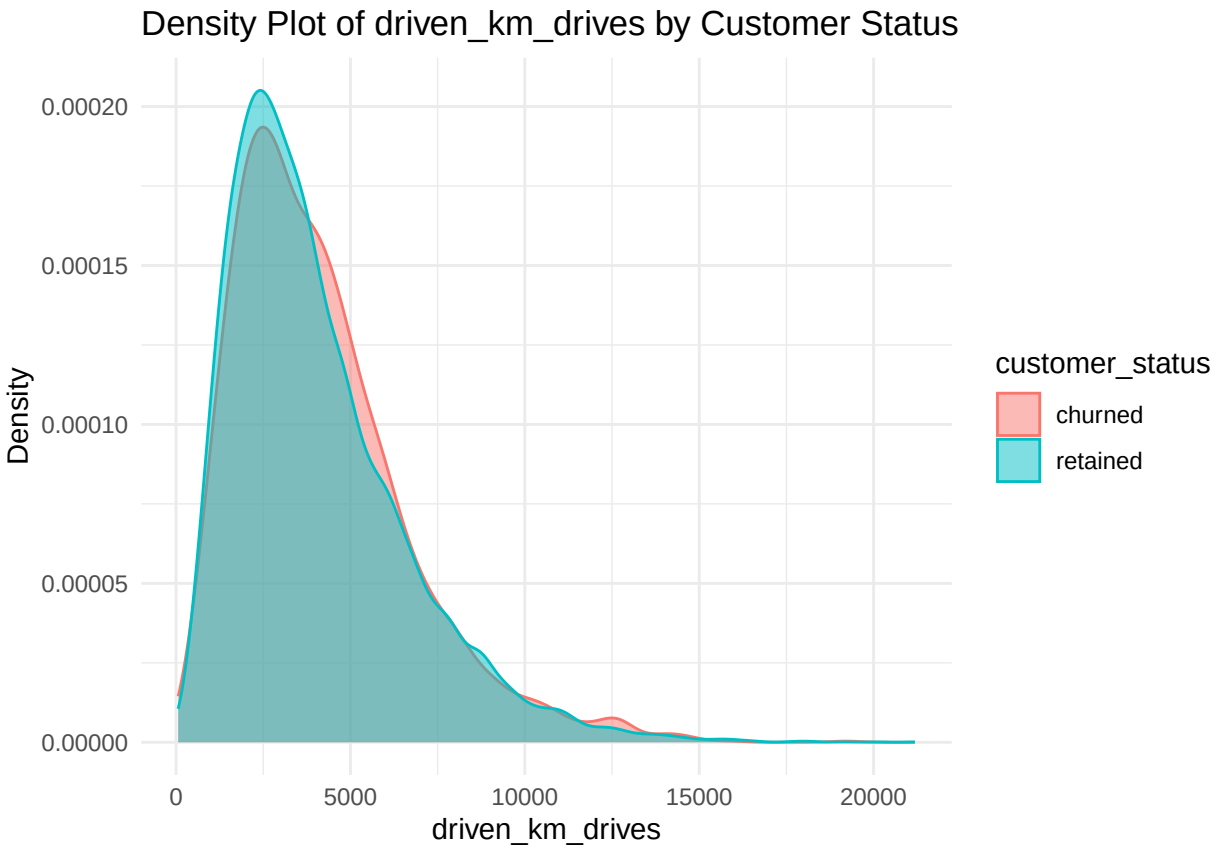

Density Plot of drives by Customer Status
0.012
0.009
0.006
0.003
0.000
Density
0
200
400
600
drives
customer_status
churned
retained

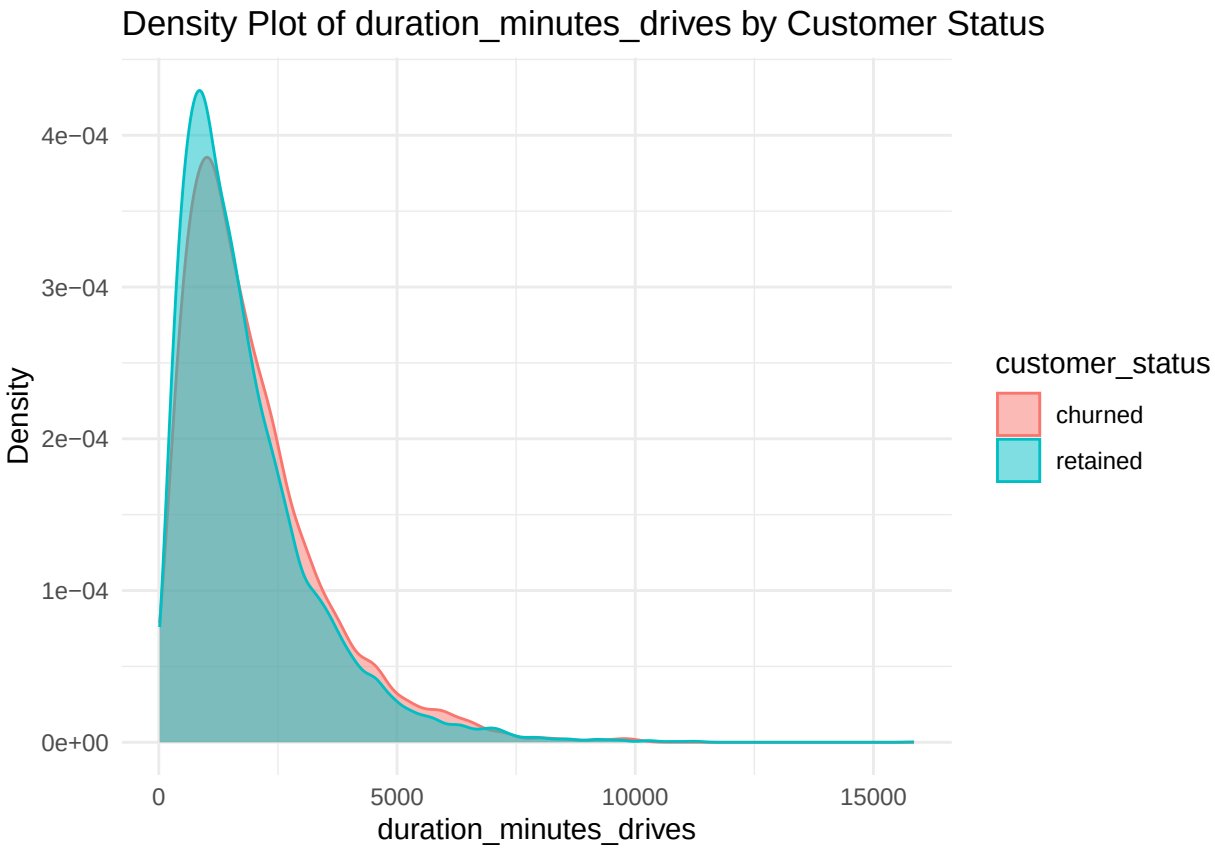

Density Plot of total_sessions by Customer Status
0.004
0.003
0.002
0.001
0.000
Density
0
250
500
750
1000
1250
total_sessions
customer_status
churned
retained

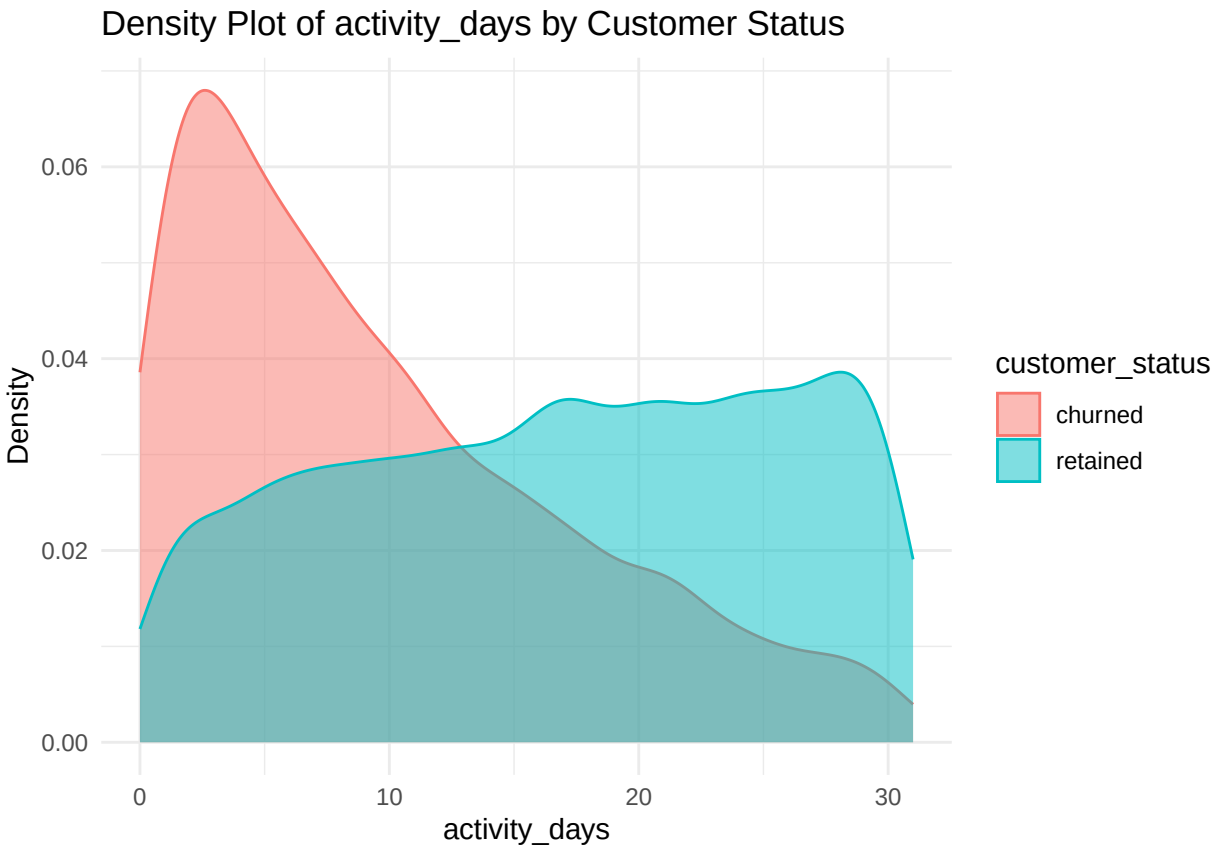

Density Plot of n_days_after_onboarding by Customer Status
Density
4e−04
3e−04
2e−04
1e−04
0e+00
0
1000
2000
3000
n_days_after_onboarding
customer_status
churned
retained

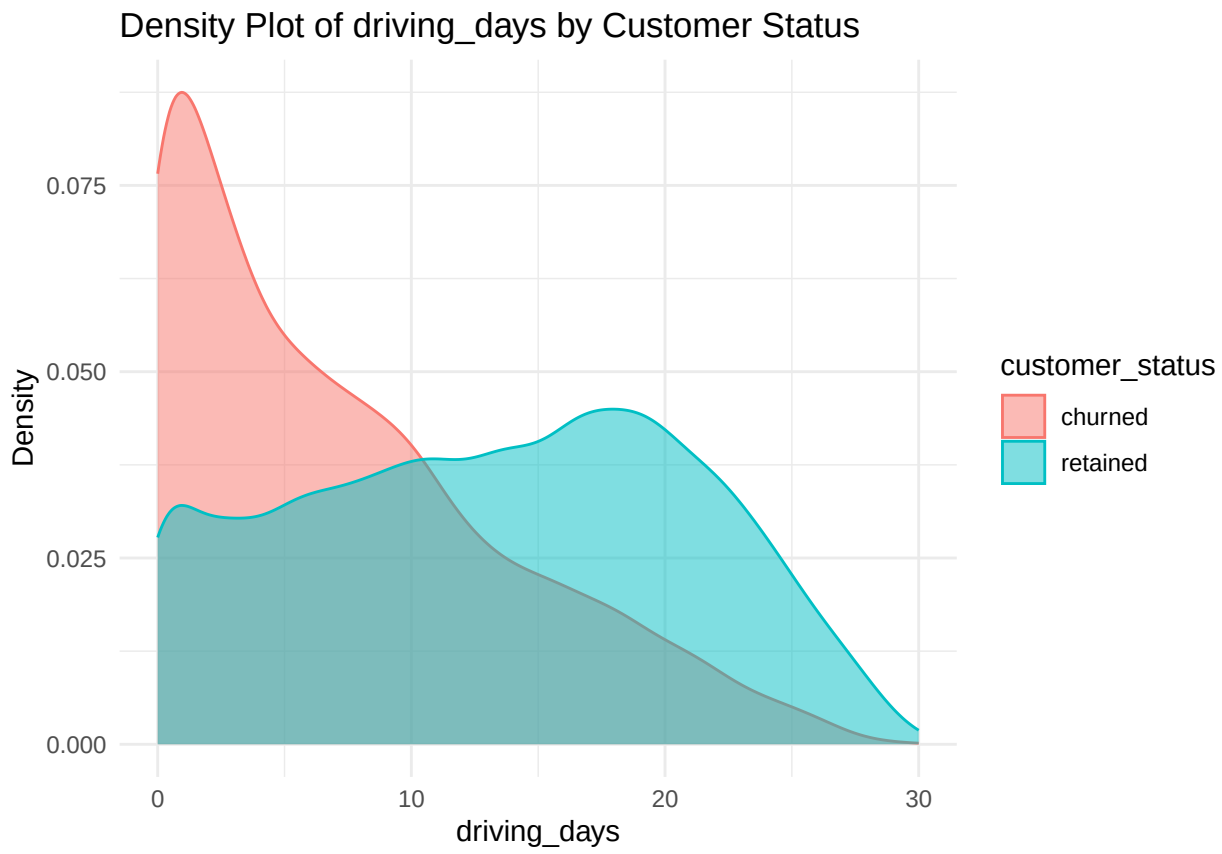

Density Plot of total_navigations_fav1 by Customer Status
Density
0.006
0.004
0.002
0.000
0
400
800
1200
total_navigations_fav1
customer_status
churned
retained


Density Plot of total_navigations_fav2 by Customer Status
Density
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
total_navigations_fav2
customer_status
churned
retained


Density Plot of driven_km_drives by Customer Status
Density
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
20000
driven_km_drives
customer_status
churned
retained


Density Plot of duration_minutes_drives by Customer Status
Density
4e−04
3e−04
2e−04
1e−04
0e+00
0
5000
10000
15000
duration_minutes_drives
customer_status
churned
retained

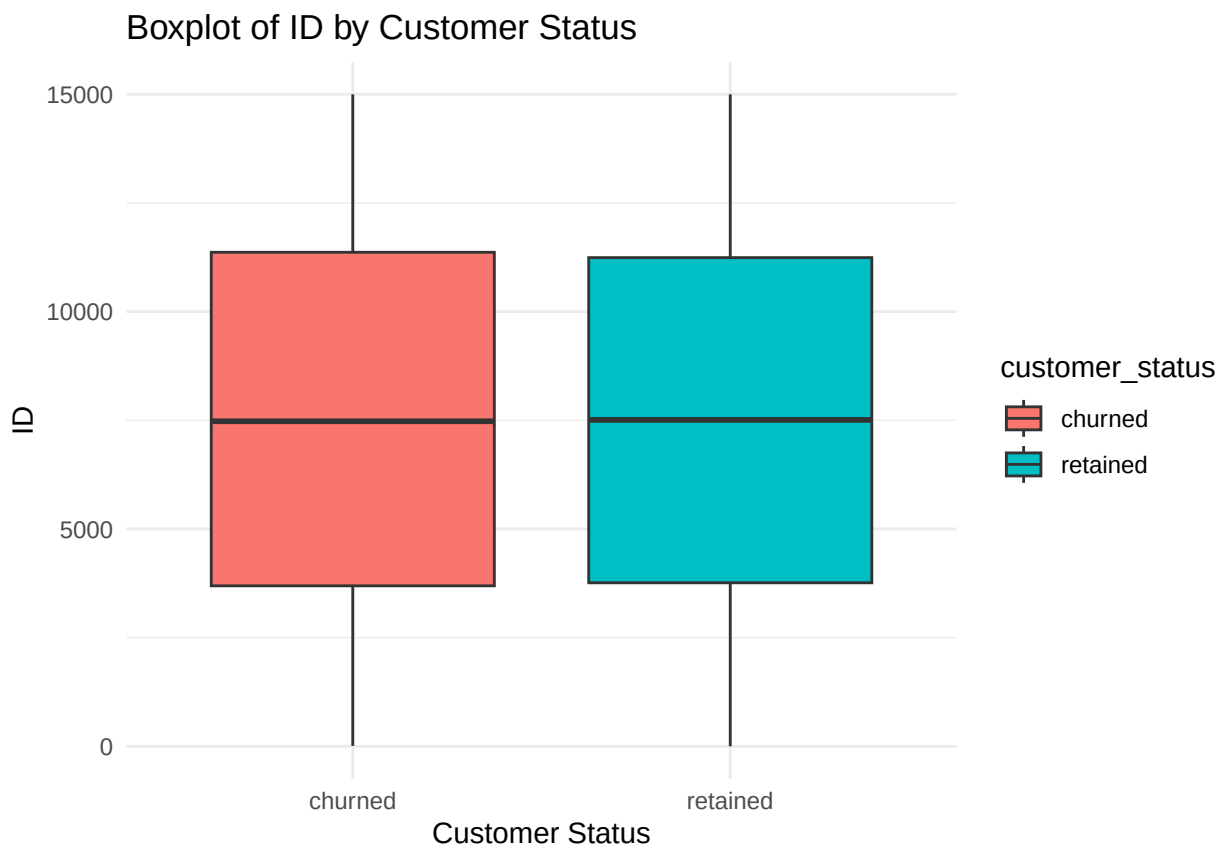

Density Plot of activity_days by Customer Status
Density
0.06
0.04
0.02
0.00
0
10
20
30
activity_days
customer_status
churned
retained

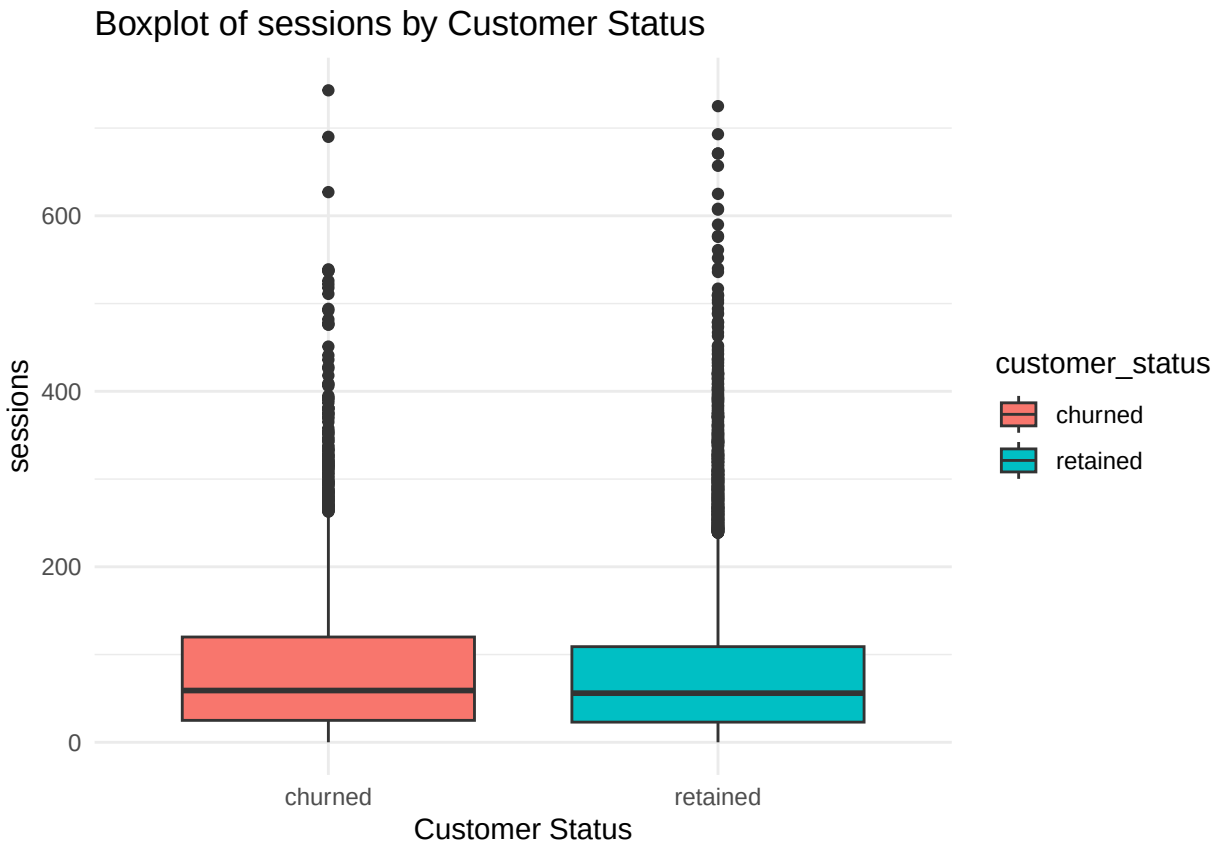


##Visualizing using boxplots

```
# Boxplot for numerical variables by customer_status
numerical_features <- waze_data_cleaned %>% select_if(is.numeric) %>% names()

for (feature in numerical_features) {
  p <- ggplot(waze_data_cleaned, aes(x = customer_status, y = !!sym(feature), fill = customer_status)) +
    geom_boxplot() +
    labs(title = paste("Boxplot of", feature, "by Customer Status"),
         x = "Customer Status",
         y = feature) +
    theme_minimal()
  print(p)
}
```

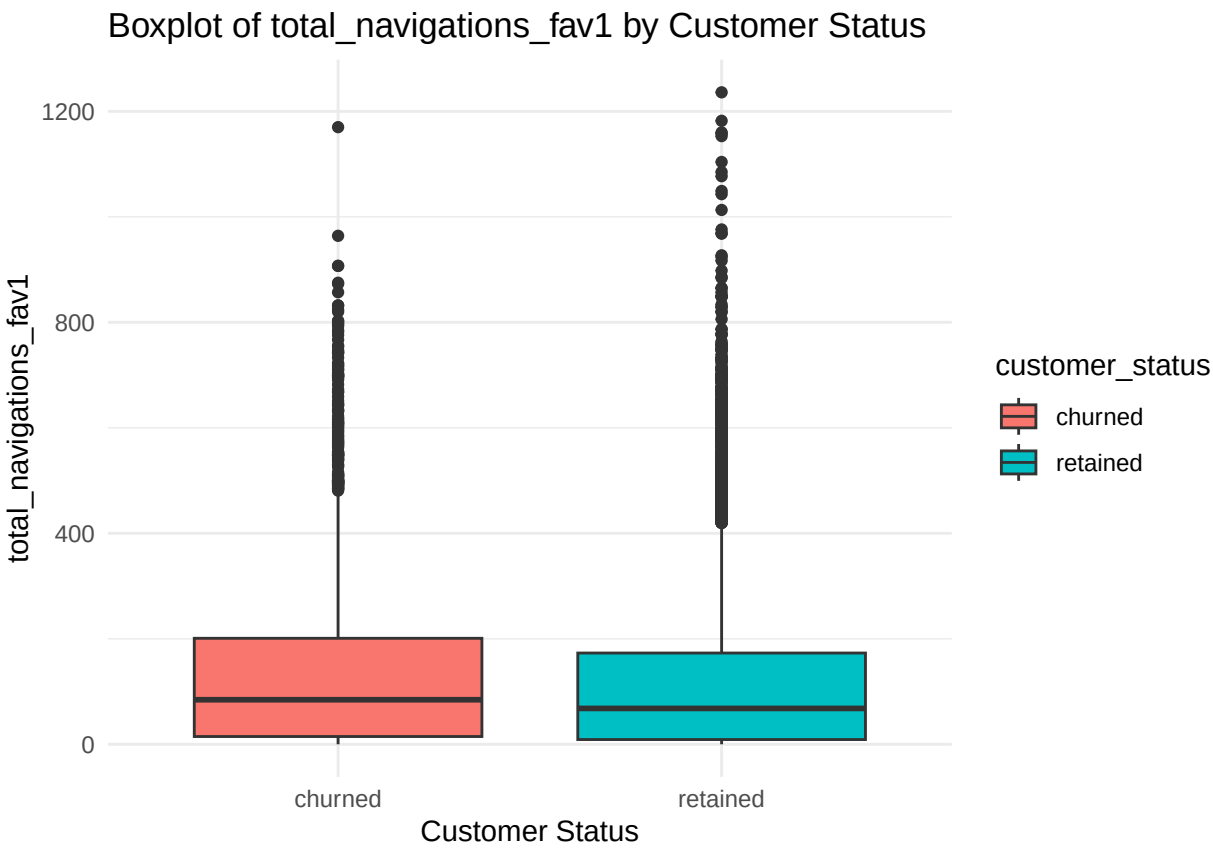

Boxplot of ID by Customer Status
15000
10000
5000
0
ID
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of sessions by Customer Status
sessions
600
400
200
0
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of drives by Customer Status
600
400
200
0
drives
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of total_sessions by Customer Status
1250
1000
750
500
250
0
total_sessions
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of n_days_after_onboarding by Customer Status
n_days_after_onboarding
3000
2000
1000
0
churned
retained
Customer Status
customer_status
churned
retained

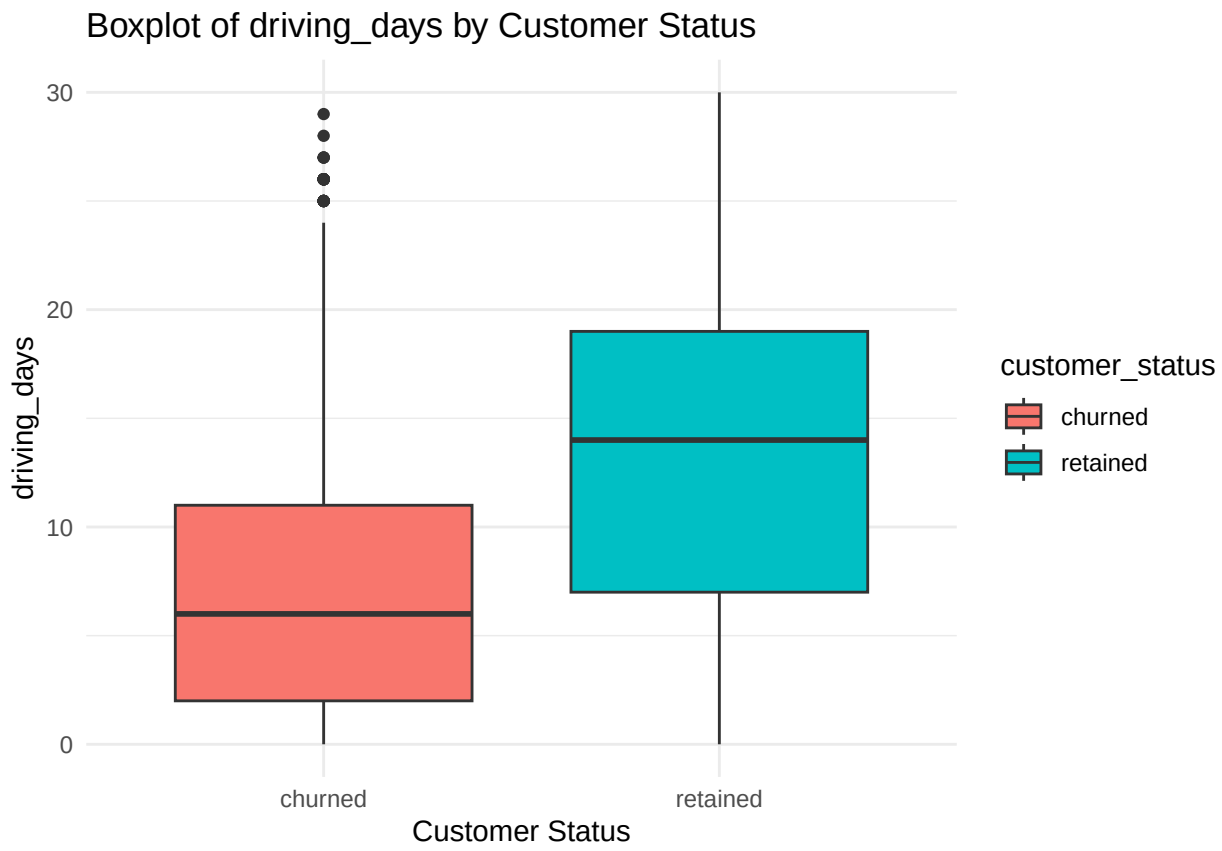

Boxplot of total_navigations_fav1 by Customer Status
total_navigations_fav1
1200
800
400
0
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of total_navigations_fav2 by Customer Status
total_navigations_fav2
400
300
200
100
0
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of driven_km_drives by Customer Status
driven_km_drives
20000
15000
10000
5000
0
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of duration_minutes_drives by Customer Status
duration_minutes_drives
15000
10000
5000
0
churned
retained
Customer Status
customer_status
churned
retained


Boxplot of activity_days by Customer Status
activity_days
30
20
10
0
churned
retained
Customer Status
customer_status
churned
retained



##Statistic Modeling

```
waze_data_cleaned$customer_status <- as.factor(waze_data_cleaned$customer_status)
model <- glm(customer_status ~ sessions + drives + total_sessions +
             n_days_after_onboarding + total_navigations_fav1 +
             total_navigations_fav2 + driven_km_drives +
             duration_minutes_drives + activity_days +
             driving_days + device,
             data = waze_data_cleaned, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = customer_status ~ sessions + drives + total_sessions +
##     n_days_after_onboarding + total_navigations_fav1 + total_navigations_fav2 +
##     driven_km_drives + duration_minutes_drives + activity_days +
##     driving_days + device, family = binomial, data = waze_data_cleaned)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.927e-02  8.562e-02  -0.225 0.821924
## sessions                 1.350e-03  3.484e-03   0.388 0.698296
## drives                  -3.440e-03  4.259e-03  -0.808 0.419248
## total_sessions          -1.317e-04  2.114e-04  -0.623 0.533457
## n_days_after_onboarding  3.891e-04  2.384e-05  16.320  < 2e-16 ***
## total_navigations_fav1  -1.099e-03  1.489e-04  -7.377 1.62e-13 ***
```

```
## total_navigations_fav2  -1.137e-03  5.012e-04  -2.268 0.023309 *
## driven_km_drives         1.471e-05  1.331e-05   1.105 0.269093
## duration_minutes_drives -8.335e-05  2.239e-05  -3.723 0.000197 ***
## activity_days            8.095e-02  9.012e-03   8.983  < 2e-16 ***
## driving_days             2.774e-02  1.043e-02   2.659 0.007827 **
## deviceiPhone             7.059e-03  4.915e-02   0.144 0.885804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13366  on 14298  degrees of freedom
## Residual deviance: 11571  on 14287  degrees of freedom
## AIC: 11595
##
## Number of Fisher Scoring iterations: 5
```