Waterman Workspaces Customer Churn Analysis: Impact of CRM Cases

Bilal Raja

2024-10-02

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

Abstract

This report discusses the customer churn analysis for Waterman Workspaces by using the customers' membership data, together with the data from the Customer Relationship Management (CRM) system to derive insights for investigating the factors that may potentially influence the customers' churn behaviour. Factors such as case escalations, frequency of cases, and time taken to resolve a case were investigated using logistic regression modelling. The purpose of our analysis is to use the findings to guide retention strategies and improve services of the organization.

Background and Motivation

Waterman Workspaces uses a membership subscription model to run their business which are defined as their 'products'. Some examples of the products include casual hire, team membership, part time membership, and suite among others. Due to an intense market competition in the domain, the organization found it pertinent to investigate customer retention on a precautionary basis in order to uncover an aspect of their business' health. Two of the most crucial areas were analyzing the customers' attendances and foot traffic and case analysis using the data from the CRM. The area to investigate was the CRM for me where the analysis focused on whether the case patterns - that include the number of cases and their resolution time - and their escalation statuses affect the likelihood of churn.

Through the CRM system, Waterman Workspaces keeps a record of the numerous interactions that take place with their customers. However, given the data, it is unclear how the information from the data source

correlate with churn. Thus, it is pertinent to investigate the dynamics between the service quality provided by the organization and churn in order to formulate actionable insights and enhance customer satisfaction.

Objectives and Significance

The objectives and significance of the report include:

- Identifying the key drivers for churn: Assess the variables that influence that lead to customers churning. The variables investigated in the report include the number of cases, their resolution time, case escalations, and whether different sites of Waterman Workspaces have greater churns. Accurate analysis gives birth to formulating targeted strategies.
- Modelling and quantifying the impact: Determining how the potential drivers affect churn likelihood via logistic regression modeling.
- Develop actionable insights for long-term success: Provide data-driven recommendations to the organization that are concise and understood with ease for the business for retention strategies. Fostering informed decision-making leads to enhanced overall business performance.

Methodology

The data was provided the host organization in the form of 2 primary datasets; CRM Cases data (`wmcases`) and membership data (`memberships`).

The CRM Cases data output is seen below where we can see the variables that come along with it followed by the variables in the memberships data.

Table 1: Table 1: List of Variables in the `wmcases` Dataset

Variables	Variables
(Do Not Modify) Case	(Do Not Modify) Row Checksum
(Do Not Modify) Modified On	Case Title
Account Number (Customer) (Account)	Customer
Case Age (Days)	Follow Up By
Case Note	Site
Priority	Status Reason
Is Escalated	Modified On
Case Age	Case Number
Case Type	Satisfaction
Sentiment Value	Service Level
SLA	Severity
Status	Description
Subject	(Do Not Modify) Case

Table 2: Table 2: List of Variables in the `memberships` Dataset

Variables	Variables
Product Name	Created On
Billing End Date	Billing Start Date
Created By	Lessee
Location	Status

Variables	Variables
Status Reason	Total Monthly Lease
Accounting Code	Lease Products
Product Category	Account Number (Lessee) (Account)
Account Category (Lessee) (Account)	Industry (Lessee) (Account)

Upon consultation with the relevant personnel in the organization, the variables and observations that were not needed were removed from our analysis. Moreover, since the key variable was the *Account Number* of the customers, due diligence was made to ensure blanks and non-customer observations were removed as well. When the irrelevant observations and variables were removed, the 2 datasets were joined using a left join by using the `Account Number` as the key variable. The left join was essential as it retained all records of the CRM data and the matching records of the memberships data so that every case lodged in the CRM system was included in the analysis even if some memberships data for customers was missing. In theory, this approach factors in customers that may have recently joined in or left already.

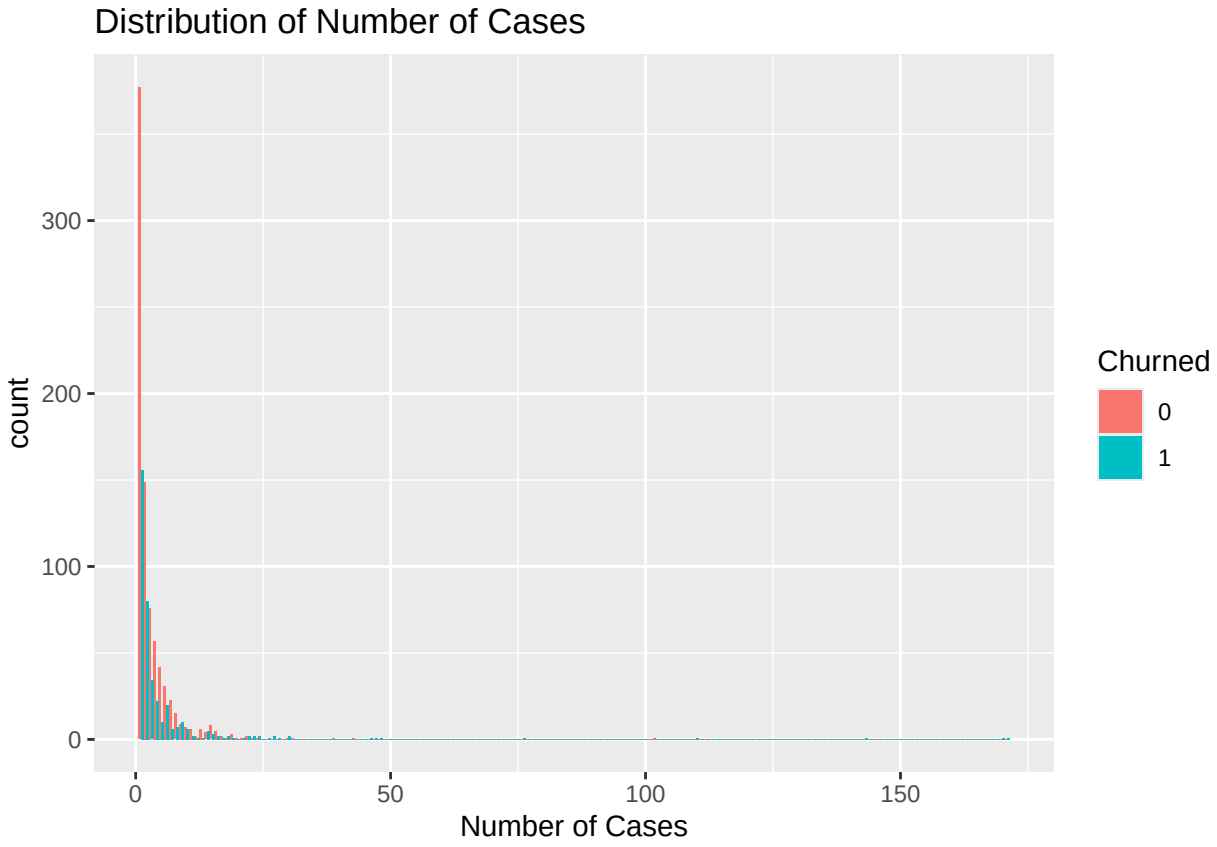
The merged data produced a large dataset with an increasing number of variables that needed to be cleaned further. As before, potential `NA` values were catered for and further nonessential variables and observations were removed. The resulting dataset had the following useful variables for our analysis:

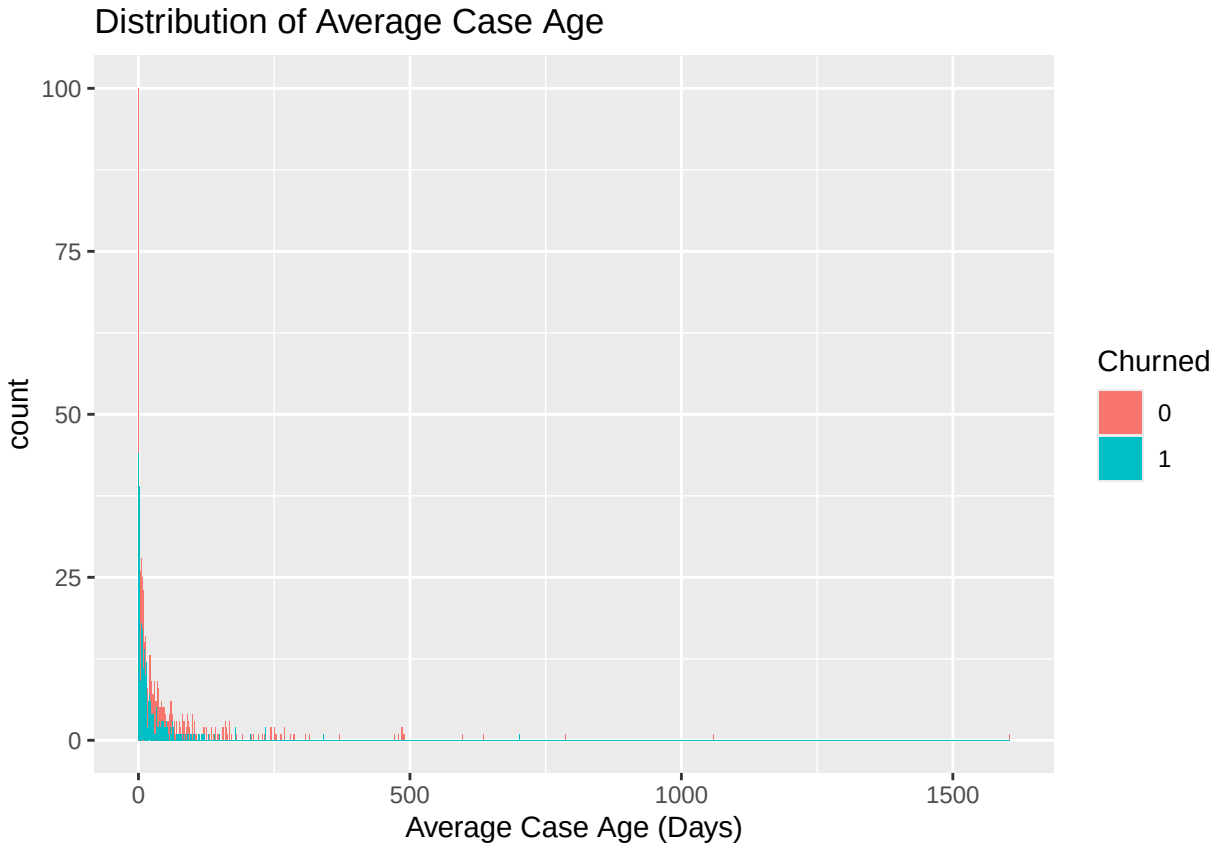
Table 3: Table 3: Variables in the `cases_joined` Dataset

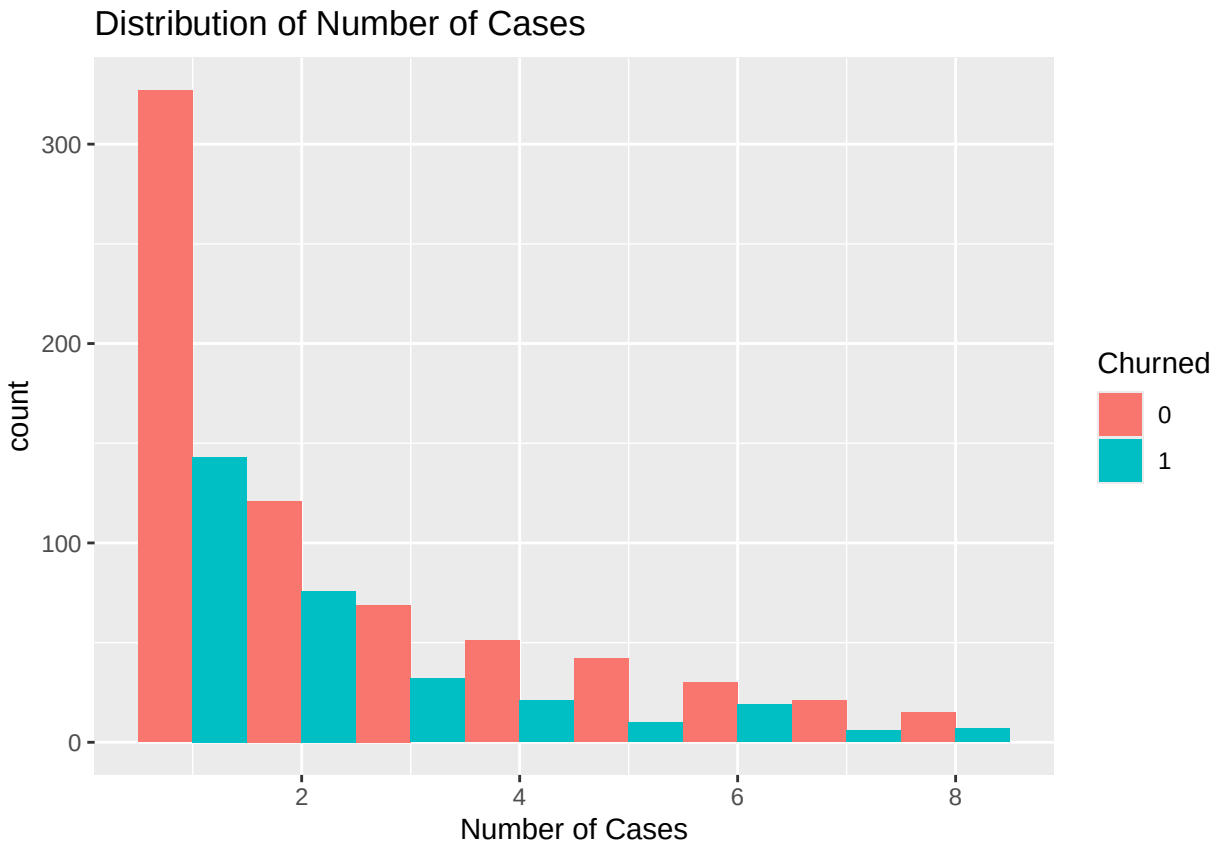
Variables	Variables
Account Number	Customer
Case Number	Case Title
Case Age	Case Age (Days)
Is Escalated	Follow Up By
Case Note	Site
Case Type	Status.x
Description	Subject
Created On	Billing End Date
Billing Start Date	Created By
Status Reason.y	Total Monthly Lease
Lease Products	Product Category
Account Category (Lessee) (Account)	Account Number

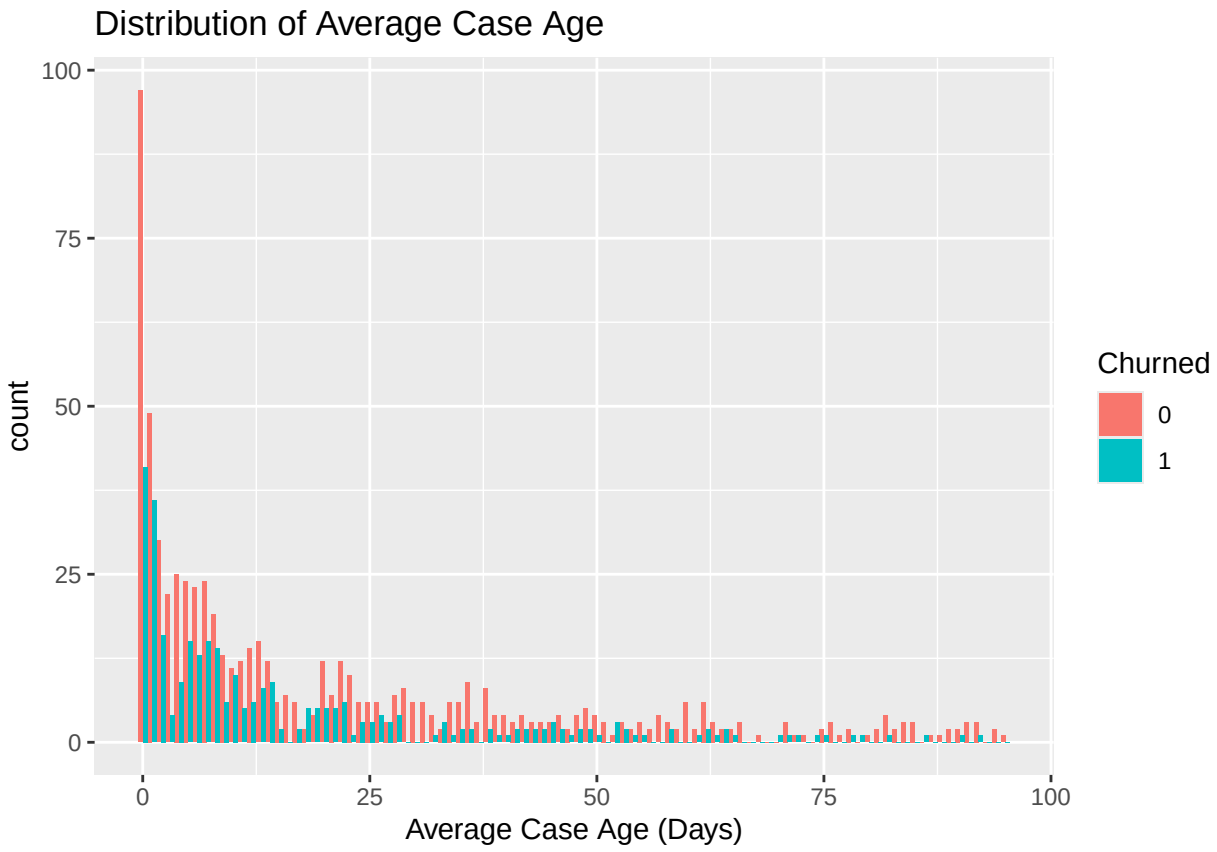
```
## # A tibble: 1,224 x 2
##    `Account Number` num_cases
##    <chr>                <int>
##  1 112222                   2
##  2 ACTBBDC3                 1
##  3 AV000014                 4
##  4 AV000036                 4
##  5 AV000041                 1
##  6 AV000046                 4
##  7 AV000047                14
##  8 AV000049                 2
##  9 AV000051                 4
## 10 AV000056                 1
## # i 1,214 more rows

## # A tibble: 1,224 x 2
##    `Account Number` avg_case_age
```

```
##    <chr>                  <dbl>
##  1 112222                 122
##  2 ACTBBDC3              1605
##  3 AV000014                62
##  4 AV000036                40
##  5 AV000041                 0
##  6 AV000046                 4
##  7 AV000047                 6
##  8 AV000049                 5.5
##  9 AV000051                21
## 10 AV000056                 2
## # i 1,214 more rows
```




Distribution of Average Case Age
count
100
75
50
25
0
0
500
1000
1500
Average Case Age (Days)
Churned
0
1


Distribution of Number of Cases
count
300
200
100
0
2
4
6
8
Number of Cases
Churned
0
1




```
##                num_cases avg_case_age is_escalated     churned
## num_cases     1.00000000  -0.01202973   0.05507895 -0.01864134
## avg_case_age -0.01202973   1.00000000  -0.15042038 -0.08756574
## is_escalated  0.05507895  -0.15042038   1.00000000  0.47719949
## churned      -0.01864134  -0.08756574   0.47719949  1.00000000
```

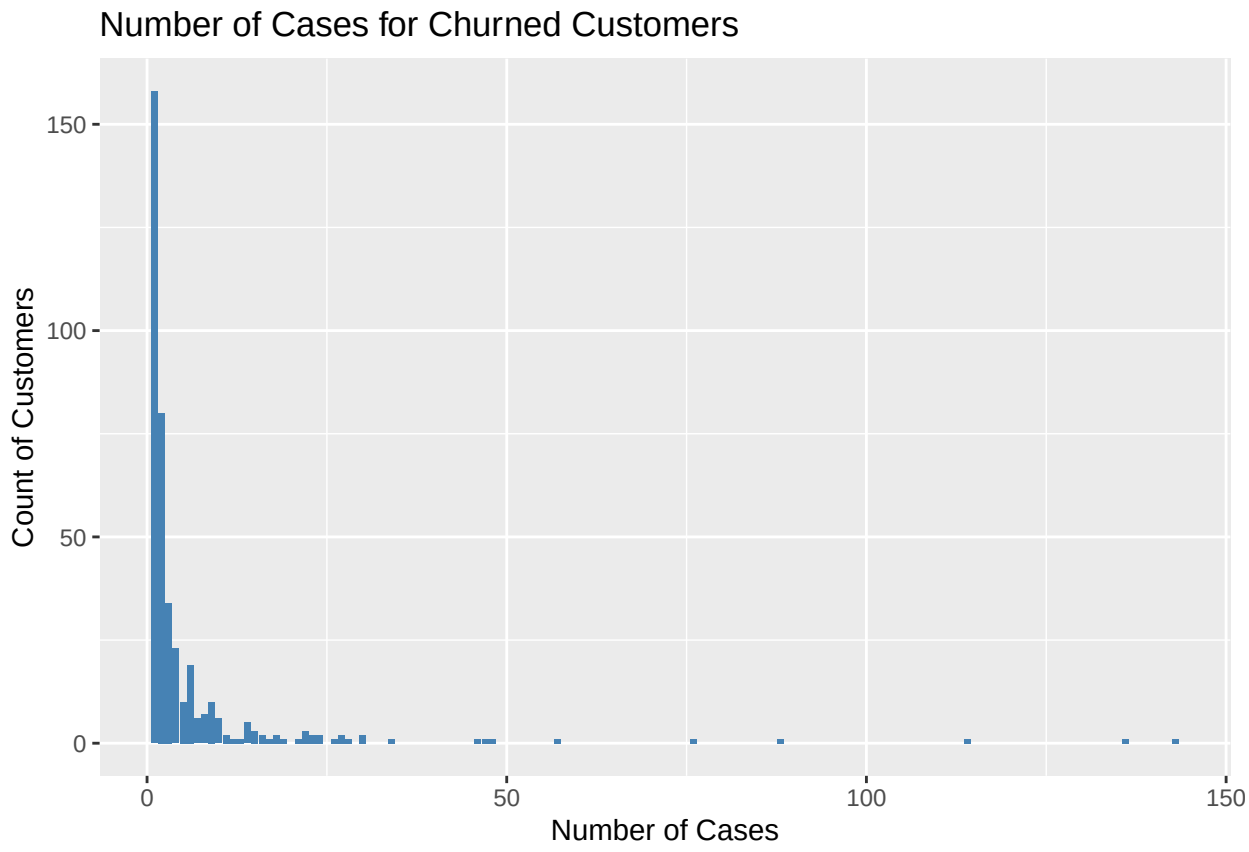
```
##
## Call:
## glm(formula = churned ~ num_cases + avg_case_age + is_escalated +
##     site, family = binomial, data = customer_features_clean)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.837e+01  4.945e+02  -0.037    0.970
## num_cases                   3.730e-02  5.021e-02   0.743    0.458
## avg_case_age                1.708e-03  3.773e-03   0.453    0.651
## is_escalated                4.609e+00  6.647e-01   6.934 4.09e-12 ***
## siteBundoora                1.967e+01  4.945e+02   0.040    0.968
## siteCamberwell              1.948e+01  4.945e+02   0.039    0.969
## siteCaribbean Park          1.728e+01  4.945e+02   0.035    0.972
## siteCasey Corporate Centre  1.593e+00  1.767e+03   0.001    0.999
## siteChadstone               1.671e+01  4.945e+02   0.034    0.973
## siteEastland                1.891e+01  4.945e+02   0.038    0.969
## siteGENERAL ENQUIRY         3.485e+01  2.450e+03   0.014    0.989
## siteNarre Warren            1.676e+01  4.945e+02   0.034    0.973
```

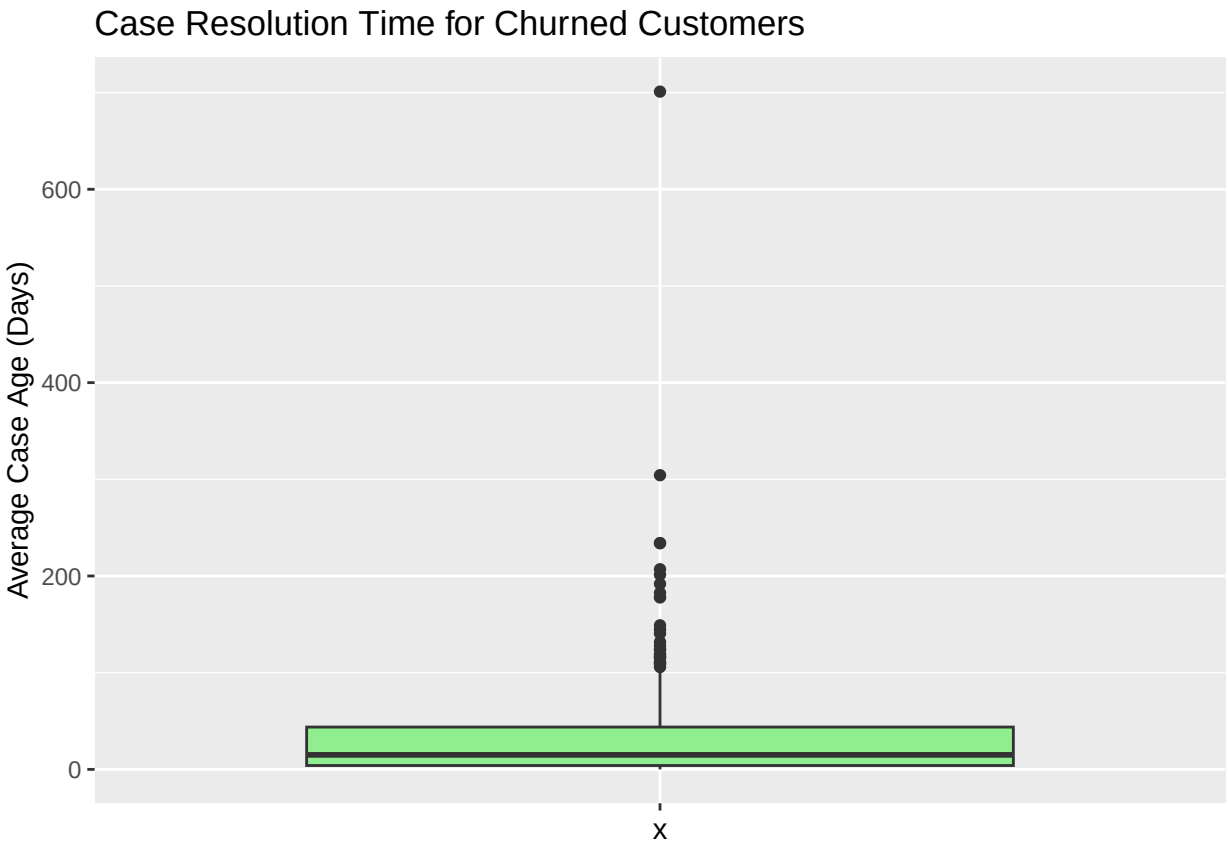
```
## siteRichmond                1.754e+01  4.945e+02   0.035    0.972
## siteSixty Four on Victor    1.384e+01  4.945e+02   0.028    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1169.46  on 913  degrees of freedom
## Residual deviance:  793.44  on 900  degrees of freedom
##   (76 observations deleted due to missingness)
## AIC: 821.44
##
## Number of Fisher Scoring iterations: 15
```

```
##
## Call:
## glm(formula = churned ~ num_cases + avg_case_age + is_escalated +
##     site + num_cases_escalated + case_age_escalated, family = binomial,
##     data = customer_features_clean)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.802e+01  5.057e+02  -0.036    0.972
## num_cases                   3.978e-02  5.062e-02   0.786    0.432
## avg_case_age                1.623e-03  3.790e-03   0.428    0.668
## is_escalated                4.775e+00  1.192e+00   4.005 6.21e-05 ***
## siteBundoora                1.931e+01  5.057e+02   0.038    0.970
## siteCamberwell              1.912e+01  5.057e+02   0.038    0.970
## siteCaribbean Park          1.693e+01  5.057e+02   0.033    0.973
## siteCasey Corporate Centre  1.231e+00  1.770e+03   0.001    0.999
## siteChadstone               1.635e+01  5.057e+02   0.032    0.974
## siteEastland                1.856e+01  5.057e+02   0.037    0.971
## siteGENERAL ENQUIRY         3.450e+01  2.452e+03   0.014    0.989
## siteNarre Warren            1.641e+01  5.057e+02   0.032    0.974
## siteRichmond                1.720e+01  5.057e+02   0.034    0.973
## siteSixty Four on Victor    1.344e+01  5.057e+02   0.027    0.979
## num_cases_escalated        -1.465e-01  3.348e-01  -0.437    0.662
## case_age_escalated          2.258e-02  5.925e-02   0.381    0.703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1169.46  on 913  degrees of freedom
## Residual deviance:  793.14  on 898  degrees of freedom
##   (76 observations deleted due to missingness)
## AIC: 825.14
##
## Number of Fisher Scoring iterations: 15
```

```
##           df      AIC
## log_model 14 821.4432
## int_model 16 825.1422
```

```
##           df      BIC
## log_model 14 888.8928
## int_model 16 902.2275
```




Case Resolution Time for Churned Customers
Average Case Age (Days)
600
400
200
0
x