Waterman Workspaces Customer Churn Analysis: CRM Cases

Bilal Raja

2024-11-01

Contents

1 Abstract **2**

2 Background and Motivation **2**

3 Objectives and Significance **2**

4 Methodology **3**

4.1 Data Cleaning and Preparation 3

4.2 Creating the Final Data Subset 5

4.3 Exploratory Data Analysis 8

4.4 Logistic Regression Modeling 13

4.5 Using Interaction Terms and Logistic Regression Model 16

4.6 Model Comparison 18

4.7 Modelling Assumptions and Limitations 19

4.8 Self Reflection 19

4.9 Software and Tools Used 20

5 Conclusion **21**

5.1 Limitatations 21

5.2 Practical Implications, Recommendations, and Future Analysis 21

6 References **22**

1 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. The GitHub repository can be accessed here.

2 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.

3 Objectives and Significance

The objectives and significance of the report include:

- **Foster a Data-Driven Culture:** Using data to derive insights cultivates a culture of making informed decisions. Our analysis served as a stepping stone for understanding Waterman Wokrspaces' customer engagement by understanding the CRM Cases data that allowed us to view the frequency of cases and their resolution times among other factors.
- **Identifying the Key Drivers for Churn:** The primary objective is to 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. Understanding these drivers will allow for a focused approach on areas most associated with customer churn.
- **Develop a Predictive Model for Churn:** Logistic modelling is used with the aim to estimate the probability of a customer churning based on the provided CRM data. This approach provides incentives to act proactively in order to retain customers and minimize future churns.
- **Informed Retention Strategies**: Our analysis will allow for pointing out at-risk customers to aid in developing tailored retention strategies. By providing customized solutions, Waterman Workspaces can engage with their customers before they cancel.
- **Enhance Business Sustainability and Gain Competitive Advantage:** The main objective of our analysis is to aid Waterman Workspaces to keep their customers happy who in return keep the business running while attracting further potential suitors. Thus, reducing customer churn not only boosts profits, but it also places Waterman Workspaces in a healthy position in the market and having a competitive edge in the co-working space industry.

4 Methodology

4.1 Data Cleaning and Preparation

4.1.1 Data Background and Review

The data was provided by the host organization in the form of 2 primary datasets; CRM Cases data (`wmcases`) and membership data (`memberships`). The data for the cases is kept in the CRM system while the membership data was kept elsewhere and undisclosed. However, they were shared via Microsoft's SharePoint folder securely. After retrieving the data, numerous meetings with the supervisor took place over the course of five to six weeks to review the data and discuss the requirement such as exploring the data fields in ways that maintained the integrity and privacy of the data. Thus, any information that would help identify a customer personally and intimately were opted to be excluded before loading.

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 (see Tables 1 and 2).

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: List of Variables in the ‘memberships‘ Dataset

Variables	Variables
Product Name	Created On
Billing End Date	Billing Start Date
Created By	Lessee
Location	Status
Status Reason	Total Monthly Lease
Accounting Code	Lease Products
Product Category	Account Number (Lessee) (Account)
Account Category (Lessee) (Account)	Industry (Lessee) (Account)

4.1.2 Data Cleaning and Joining

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 customers' account numbers. Since the mentioned account variables were differently named in the datasets, namely `Account Number (Customer) (Account)` in the CRM data and `Account Number (Lessee) (Account)`in the membership data, they were both renamed to be `Account Number` for our analysis to make a left join possible.

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.

```
## Rows: 19,585
## Columns: 37
## $ ‘Case Title‘                           <chr> "General Christmas Touch Ups", "~
## $ ‘Account Number‘                       <chr> "0001", "MS001517", "MS001517", ~
## $ Customer                               <chr> "Waterman Workspaces", "Everest ~
## $ ‘Case Age (Days)‘                      <dbl> 28, 45, 45, 45, 23, 23, 23, 3, 3~
## $ ‘Follow Up By‘                         <dttm> 2023-12-22 15:34:00, 2024-02-05~
## $ ‘Case Note‘                            <chr> NA, NA, NA, NA, NA, NA, NA, "QR ~
## $ Site                                   <chr> "Narre Warren", "Caribbean Park"~
## $ Priority                               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ ‘Status Reason.x‘                      <chr> "Problem Solved", "Problem Solve~
## $ ‘Is Escalated‘                         <chr> "No", "No", "No", "No", "No", "N~
## $ ‘Modified On‘                          <dttm> 2024-01-16 12:28:00, 2024-03-18~
## $ ‘Case Age‘                             <chr> "40140", "64837", "64837", "6483~
## $ ‘Case Number‘                          <chr> "CAS-21133-M2P6Y9", "CAS-21727-R~
## $ ‘Case Type‘                            <chr> "Normal Resolution", "Not Applic~
## $ Satisfaction                           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ ‘Sentiment Value‘                      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ ‘Service Level‘                        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ SLA                                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Severity                               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Status.x                               <chr> "Resolved", "Resolved", "Resolve~
## $ Description                            <chr> "- Touch up of paint on walls ne~
## $ Subject                                <chr> "Maintenance - Common Area", "Ge~
## $ ‘Product Name‘                         <chr> NA, "Part Time Membership", "Mem~
## $ ‘Created On‘                           <dttm> NA, 2024-06-04 16:22:17, 2024-0~
## $ ‘Billing End Date‘                     <dttm> NA, 2099-12-31 00:00:00, 2024-0~
## $ ‘Billing Start Date‘                   <dttm> NA, 2024-06-01, 2024-04-01, 202~
## $ ‘Created By‘                           <chr> NA, "James Cheng", "James Cheng"~
## $ Lessee                                 <chr> NA, "Everest Commercial", "Evere~
## $ Location                               <chr> NA, "Caribbean Park", "Caribbean~
## $ Status.y                               <chr> NA, "Active", "Inactive", "Inact~
## $ ‘Status Reason.y‘                      <chr> NA, "Leased", "Inactive - Upgrad~
## $ ‘Total Monthly Lease‘                  <dbl> NA, 200.00, 0.00, 200.00, 3845.0~
## $ ‘Accounting Code‘                      <chr> NA, "Standard Membership", "Zero~
## $ ‘Lease Products‘                       <chr> NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ ‘Product Category‘                     <chr> NA, "Part Time Membership", "Wat~
## $ ‘Account Category (Lessee) (Account)‘ <chr> NA, "Individual Membership", "In~
## $ ‘Industry (Lessee) (Account)‘          <chr> NA, "Financial & Insurance Servi~
```

As we can see using the `glimpse` function, the joining method produces 37 columns and 19585. Moreover, the variables' types look to be correct.

4.2 Creating the Final Data Subset

The merged data having produced a large dataset with an increasing number of variables, there needed to be further cleaning. As before, potential `NA` values were catered for and further nonessential variables and observations were removed. Such examples included a repeating column reflecting the account numbers, and the priority column along with the customer sentiments had no records. Such columns were removed and rows that were raised internally but Waterman themselves and test cases were removed.

The resulting dataset had the following variables listed in 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
Status Reason.x	Case Type
Status.x	Description
Subject	Created On
Billing End Date	Billing Start Date
Created By	Status Reason.y
Total Monthly Lease	Product Category

Next, a vector was created to define churn statuses that includes

- Inactive - Customer Cancelled
- Off-boarding
- Inactive

The mentioned attributes were selected upon consultations and discussions with the supervisor.

Using the defined churn statuses, a new subset of the data was formed, namely `customer_features`. The cases were grouped by their account numbers and summarized to count the number of cases per account and a similar process was followed to summarize by the average case age as well to provide insights into a the typical duration of cases for each customer. In addition, the median was taken for the case ages over a mean to handle the skewness of the data. In such scenarios, mean is sensitive to extreme data points which could provide an untrue picture of the time it takes to resolve a case. On the other hand, the median takes the ***50th percentile*** (the central point). Exactly half of the data points are less than the median, and exactly half of the data points are greater than the median. It's right in the middle and it is not affected by outliers or skewed data (Orn, 2023). Moreover, `churned` was created as a binary variable. The logic behind making it binary was that either a customer churns or does not churn and in this case, `1` meant churned and `0` meant that the customer did not churn.

Additionally, `product_category` was created which retained the first product associated with the account to reflect the main product used by the customer. The initial reasoning behind including the product category was to analyze whether different categories had varying churn rates. A similar approach was taken to incorporate the `site` variable to the subset to investigate whether the sites of the organizations have varying churn rates that could point towards service quality issues geographically.

The joined dataset was further filtered to include another binary variable `Is Escalated` that was created to provide numeric representation for the case escalations. Here, when a case is marked as "Yes" in escalations, the final value is converted to `1` and `0` otherwise. Finally, steps were taken to ensure that the `product_category` and `site` variables were converted to factor.

4.2.1 Choosing the final Variables

We used the `vis_dat` and the `vis_miss` funcitons for another observation of the types of data and the missing values in any other variables.

Figure 1 shows that little of the observations missing in site while product category had almost half of it's observations missing (about 47%). TO numerically assess, the product categories were isolated next.

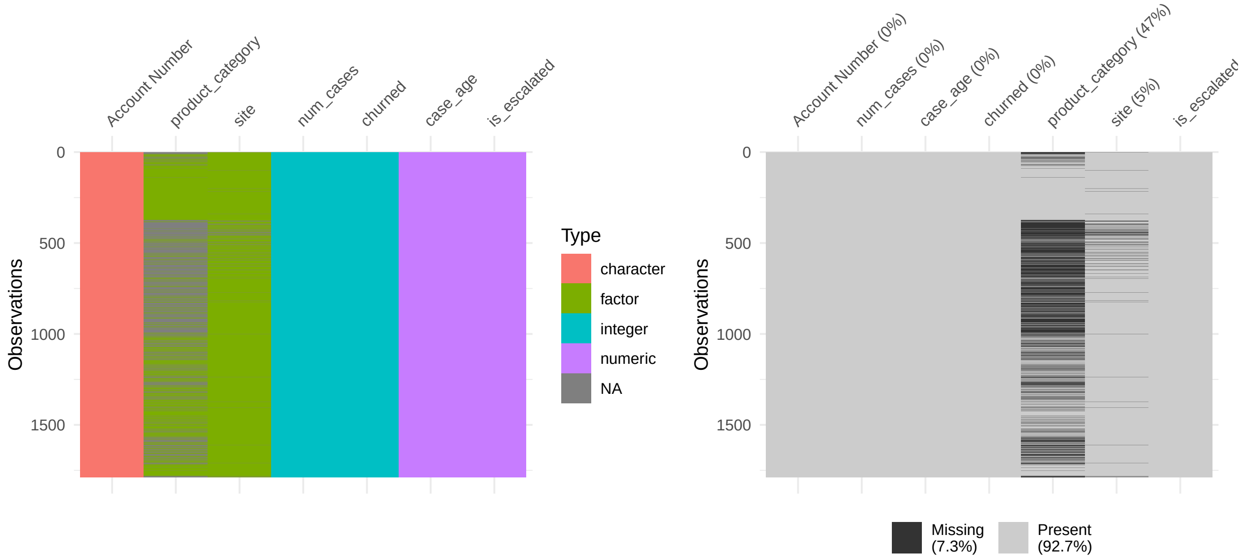

Figure 1: Comparison of dataset structure and missing values.

Table 4: Product Categories in Customer Features Dataset

product_category	Count
NA	834
Suite	272
Access Membership	235
Part Time Membership	161
Dedicated Desk	78
Anchor Suite	63
Team Membership	59
Waterman Account	23
Unlimited Membership	19
Access Membership + After Hours Pack	11
Part Time Membership + After Hours Pack	11
Retail Space	7
Virtual Office Membership	4
Landlord/Contractor	3
(NFS) Membership Plus - $350	2
Enterprise Membership	2
Casual Membership	1
Hot Desk	1
Team Membership Premium	1
Virtual Office Plus Membership	1

Table 5: Variables in the customer features Dataset

Variables
Account Number
num_cases
case_age
churned
site
is_escalated

While we are able to see what type the variables are, Figure 1 and table 4 shows that there are a significant number of missing observations in `product_category`. Due to a significantly high proportion of missing values of the product categories, the variable was not included in our analysis as the missing values would distort our results and provide little to no meaningful analysis. One thing interesting to point out was that the table also provided further information on the total number of each type of product subscribed. Having a healthier number of observations would have given an incentive to investigate how popular products' subscribers performed in our churn analysis.

Thus, as seen the table 5, the final variables that were chosen for our analysis and their explanations are as follows:

- `Account Number`: A unique identifier of a customer.
- `num_cases`: The total number of cases per customer.
- `case_age`: The median age in days taken to resolve a case.
- `churned`: A binary indicator that shows whether a customer churned or did not churn.
- `site`: The site or the location of the customer of Waterman.
- `is_escalated`: A binary indicator that shows whether a case raised by the customer was escalated.

This section of the project took the most time as it required careful considerations and numerous discussions to decide our choices of the variables and observations. Some observations were voluntarily removed while some were required to be removed without further explanation.

4.3 Exploratory Data Analysis

4.3.1 Summary Description

Table 6: Custom Summary for Selected Variables

num_cases	case_age	churned	is_escalated	site
Min. : 1.00	Min. : 0.00	Min. :0.0000	Min. :0.0000	Chadstone :512
1st Qu.: 1.00	1st Qu.: 3.00	1st Qu.:0.0000	1st Qu.:0.0000	Caribbean Park :450
Median : 2.00	Median : 11.00	Median :0.0000	Median :0.0000	Narre Warren :225
Mean : 9.27	Mean : 33.12	Mean :0.2181	Mean :0.1801	Eastland :220
3rd Qu.: 6.00	3rd Qu.: 32.00	3rd Qu.:0.0000	3rd Qu.:0.0000	Sixty Four on Victor: 90
Max. :1175.00	Max. :1605.00	Max. :1.0000	Max. :1.0000	(Other) :207
NA	NA	NA	NA	NA's : 84

Table 6 shows the overall summary of the subset. It illustrates that a customer raises 4.246 issues on average and it takes an average of 39 days to resolve the issue. Both `num_cases` and `case_age` have medians lower than the mean that signal positive skewness which will be visualized next.

4.3.2 Distribution of Case Frequency and Case Age (with outliers)

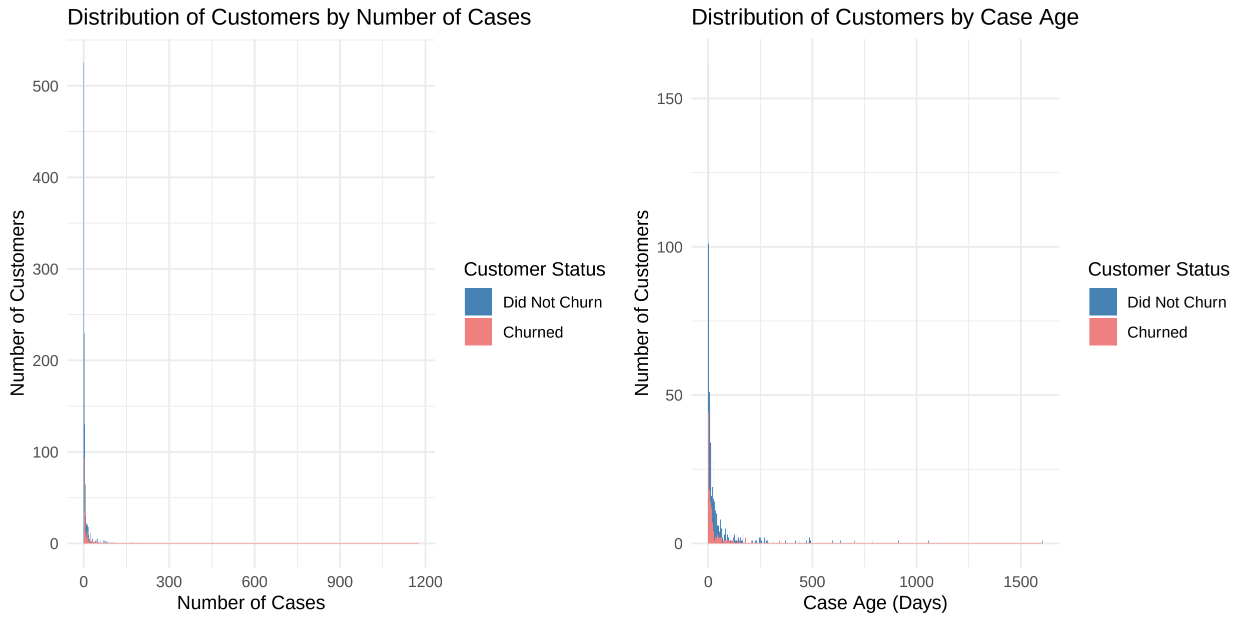

Figure 2: Outlier Distributions

Figure 2 reveals high positive skewness for the number of cases and average case age. These are such extreme data points that intuition dictated to investigate further. However, it was recommended to not include extreme outliers for business purposes. Simultaneously, Figure 2 offered an insight into how the data may be managed currently by the organization, hinting at problems that need to be addressed.

Since it was recommended to remove the outliers, the interquartile (IQR) method was employed for their removal in order to paint a clearer picture for our analysis. Using the interquartile method was preferred to any other method because of its robustness and by using the middle 50% of the data, our analysis is not influenced by the extreme values observed. Hence, the IQR method allows for preserving the underlying distribution of the data without any trade-off of the data integrity and models.

For self reflection purposes, it was realized that using the median for the average case ages previously did not remove the outliers.

4.3.3 Churn Status Comparison

Before visualizing the number of cases and the case resolution times, we first review a holistic comparison of customers churning against the customers who did not churn from the final cleaned subset of the data. Figure 3 below illustrates that There are 294 customers who churned and 1091 customers who did not churn.

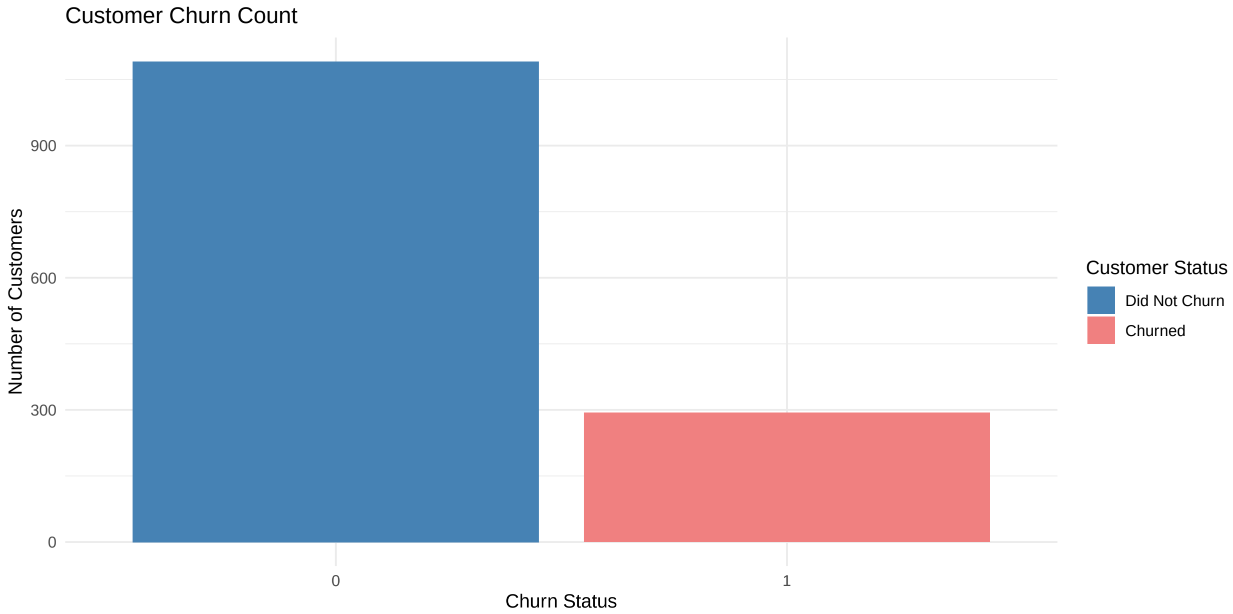

Figure 3: Analyzing Customers Churning

4.3.4 Distribution of Case Frequency and Case Age (cleaned)

Next, we plot the histograms to analyze the distributions from the cleaned data.

Figure 4 shows number of customers against the number of cases. The customers that churned and those who did not churn are segregated by separate colours for ease in visual comparison.

The figure suggests that the distribution of customers by their number of cases is positively skewed which means that there are more customers that faced issues less frequently. Moreover, it shows that as the number of cases increase, the number of customers churning decreases significantly. While this may seem counter-intuitive, it also shows that there decreasing number of customers that face more problems. In this case, it is more important to analyze the proportion of customers that churned versus the customers that did not churn given the number of cases. Even so, the number of customers that leave are still lower than the customers that do not leave at every bin of number of cases.

Figure 5 shows number of customers against the average case age. The customers that churned and those who did not churn are segregated by separate colours for ease in visual comparison.

The figure suggests that the distribution of customers against the case age is also positively skewed which means that there are more customers that had cases resolved mostly sooner rather than later. Moreover, it shows that as the case resolution time increases, the number of customers churning decreases overall. This too seems counter-intuitive and also shows that there are less customers that face longer case resolution times.

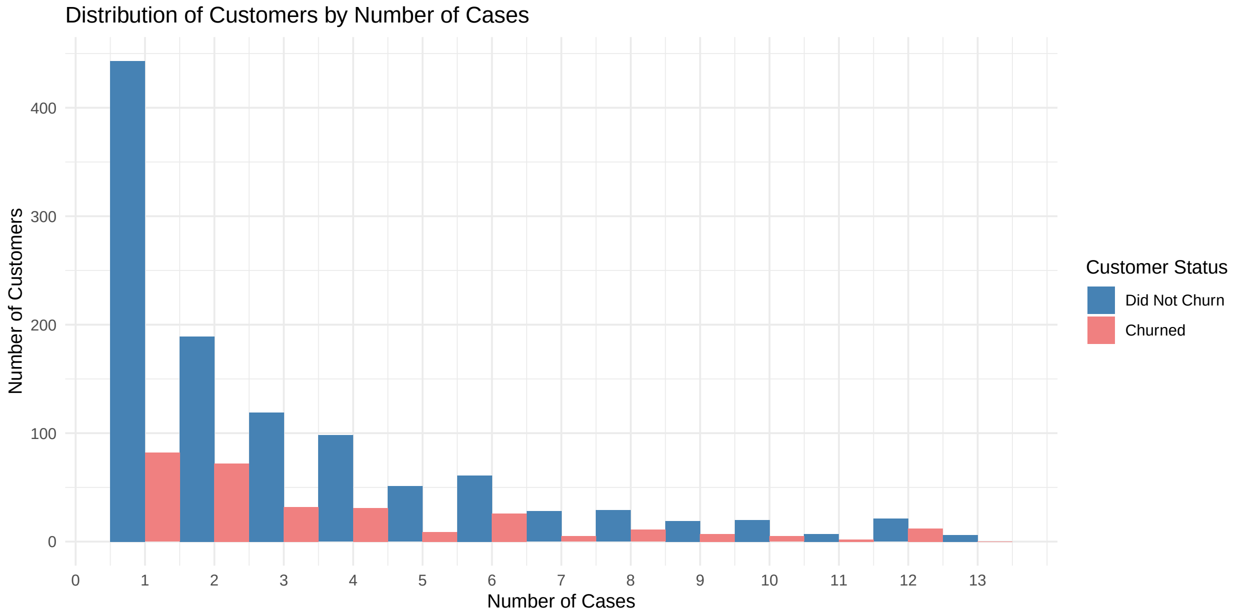

Figure 4: Number of Cases Distribution

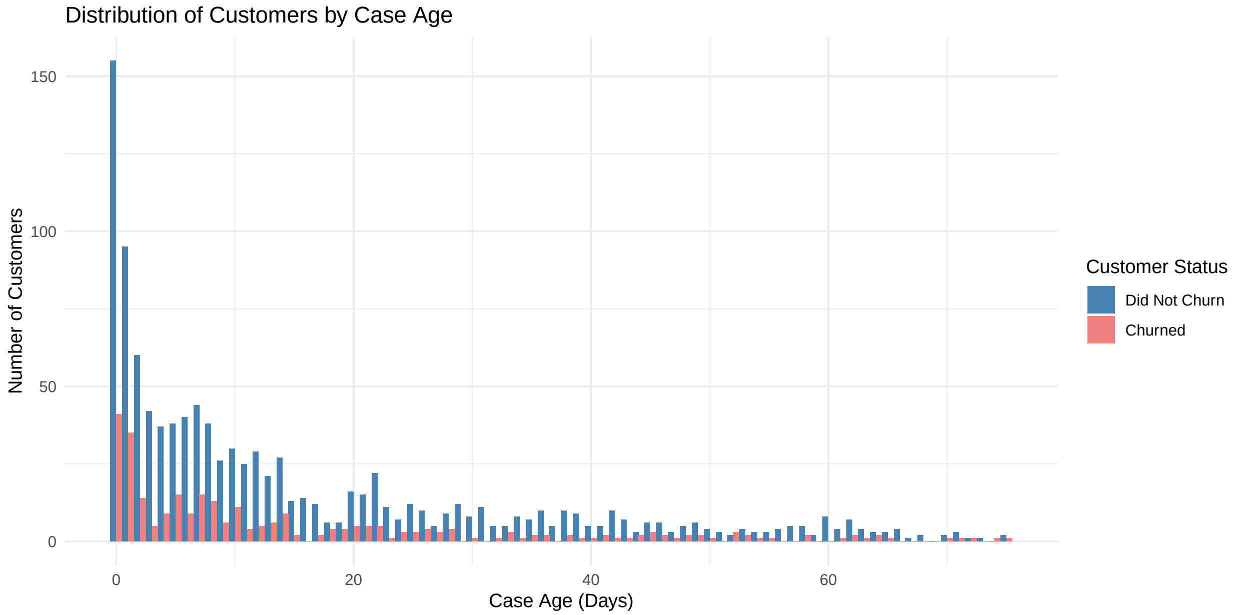

Figure 5: Case Age Distribution

Table 7: Correlation Matrix

	num_cases	case_age	is_escalated	churned
num_cases	1.0000000	0.0569439	0.1653377	0.0659591
case_age	0.0569439	1.0000000	-0.1519819	-0.0084799
is_escalated	0.1653377	-0.1519819	1.0000000	0.2129736
churned	0.0659591	-0.0084799	0.2129736	1.0000000

4.3.5 Correlation

The correlation matrix in Table 7 produces some interesting points. The correlations between all variables are very weak and inconclusive. While it was assumed that as the number of cases and case resolution times increase, the impact on churn would increase as well. However, this positive correlation is observed to be extremely weak being 0.0659591. On the other hand, cases being escalated have a positive correlation with customers churning as well. However, it too has a very weak correlation of 0.2129736

4.4 Logistic Regression Modeling

For the purpose of our analysis, we applied logistic regression to model the relationships between the explanatory variables and the outcome of customers churning. According to Peng et al. (2002) generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. In the simplest case of linear regression for one continuous predictor X and one dichotomous outcome variable Y, the plot of such data results in two parallel lines, each corresponding to a value of the dichotomous outcome. Thus, the binary nature of our outcomes encourage us to use logistic regression.

Moreover, the interpretability of logistic regression is fairly easier as it indicates how the predictors impact the outcome (in terms of log-odds). Since we take into account the probabilistic outcome, it helps us to identify at-risk customers and develop retention strategies. While more complex models could have been possible to apply, such as decision trees, logistic model was preferred to avoid unnecessary complications and provide easy to interpret results for the business audience.

The logistic regression model can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}$$

Or

$$\text{churned} \sim \beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}$$

Where: - p is the probability of customer churn. - β_0 is the intercept. - $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the predictors: number of cases, case age, escalation status, and site, respectively.

Once the log-odds are calculated, they are converted to a probability using the following equation:

$$p = \frac{e^{\beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}}}{1 + e^{\beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}}}$$

This equation produces the log-odds output between 0 and 1, which represents the probability of the customer churning.

Table 8: Logistic Regression Model Summary

Predictor	Coefficient	Standard Error	Z-Value	P-Value
Intercept	-16.9131073	520.5755434	-0.0324892	0.9740819
Number of Cases	0.0318375	0.0239921	1.3270002	0.1845086
Case Age (Days)	0.0058508	0.0039993	1.4629518	0.1434806
Is Escalated	0.7942216	0.1811311	4.3847898	0.0000116
siteBundoora	16.0117055	520.5756181	0.0307577	0.9754628
siteCamberwell	15.5292145	520.5756021	0.0298309	0.9762020
siteCaribbean Park	15.3067814	520.5755422	0.0294036	0.9765427
siteCasey Corporate Centre	0.1792426	1478.6036856	0.0001212	0.9999033
siteChadstone	15.0150774	520.5755406	0.0288432	0.9769896
siteEastland	16.0489765	520.5755502	0.0308293	0.9754057
siteGENERAL ENQUIRY	33.4008737	2455.3642797	0.0136032	0.9891465
siteNarre Warren	15.0441202	520.5755629	0.0288990	0.9769451
siteRichmond	15.2931120	520.5756571	0.0293773	0.9765637
siteSixty Four on Victor	12.3384472	520.5764957	0.0237015	0.9810907

Table 9: Model Fit - McFadden's R^2

Metric	Value
McFadden's R^2	0.0882911

4.4.1 Fitting the model

```
## fitting null model for pseudo-r2
```

Table 10: Variable Importance

Variable	Importance
num_cases	1.3270002
case_age	1.4629518
is_escalated	4.3847898
siteBundoora	0.0307577
siteCamberwell	0.0298309
siteCaribbean Park	0.0294036
siteCasey Corporate Centre	0.0001212
siteChadstone	0.0288432
siteEastland	0.0308293
siteGENERAL ENQUIRY	0.0136032
siteNarre Warren	0.0288990
siteRichmond	0.0293773
siteSixty Four on Victor	0.0237015

4.4.2 Model Interpretation and Summary

The summary of the estimates are provided in Table 8 while the McFadden's R^2 value and variable importance are displayed in tables 9 and 10 respectively.

The R^2 value of 0.088 suggests that while our model does explain some variability in the data, it still has low predictive power.

On the other hand, the variable `is_escalated` is considered the most important variable in our data.

Each coefficient in the output of the regression reflects the log-odds change in the probability of a customer churning for a 1 unit increase in a predictor while the other predictors are held constant. For testing significance of our predictors, we use the cut-off $\alpha = 0.05$ (5% significance level).

- **Intercept:** The intercept serves as baseline log-odds of churn when all predictors are valued at zero. The estimate for the intercept is -16.913 but is held insignificant at $\alpha = 0.05$ significance as the associated p-value is 0.947.
- **Number of Cases:** For each additional case associated with a customer, the log-odds of the customer churning increase by 0.0318. This suggests that as a customer accumulates more cases, or faces more issues, they may be more likely to churn. However, since the p-value > 0.05 (at 0.185), we conclude that this effect is not statistically significant.
- **Case Age (Days):** When a case is unresolved each passing day, the log-odds of a customer churning increase by 0.0059. Initial impressions mean that the longer a case takes to be resolved contribute slightly to the likelihood of a customer churning, it is statistically insignificant as the p-value > 0.05 (at 0.143).
- **Is Escalated:** When a case is escalated (value equals 1), the log-odds of a customer churning increase by 0.794. This reflects a likelihood on the higher side and is statistically significant since p-value < 0.05 (at 1.16e-05). Thus we conclude with confidence that the escalated cases are associated with a higher probability of a customer churning. This also suggest that this is a strong indicator of a customer's dissatisfaction.
- **Site:** There were several sites in our data and they had to be coded as dummy variables for our analysis purposes. Each coefficient of a site compares it with a reference category. However, none of the sites have a p-value lower than 0.05, meaning that this model suggests no individual site has any meaningful nor significant contribution to a customer's likelihood of churning.

This model does provide potentially valuable insights, but it points towards the need for further refinement. While an escalated case is picked out to be the strongest and the most significant predictor from our data, the rest of the variables not being significant means we need more meaningful predictors. That said, it can be initially concluded that investigating the escalated cases further could prove to be beneficial in retaining customers.

4.5 Using Interaction Terms and Logistic Regression Model

In order to enhance our mode, we decided to see how interaction terms behaved and affected our model. The objective ehre was to see whether the model is enhanced and if there are any changes to our results.

4.5.1 Creating Interaction Terms

The following interaction terms were defined our `customer_features_clean` dataset:

- **num_cases_escalated:** This variable represents the interaction between the number of cases variable and the escalated case variable. The significance of this variable is that it would allow us to assess whether the effect of the number of cases on churn is different when the cases are escalated as compared to when the cases are not escalated.
- **case_age_escalated:** This variable is derived from the interaction between the case age variable and the escalated case variable. This term will allow us to assess whether older cases have any influence on the customers' churn rate depending on their escalation status.

This is an attempt to analyze and reveal the complex relationships the variables have together as compared to being assessed for their main effects alone.

The new equation we use for our model is as follows:

$$\text{churned} \sim \beta_0+\beta_1\text{num\_cases}+\beta_2\text{case\_age}+\beta_3\text{is\_escalated}+\beta_4\text{site}+\beta_5\text{num\_cases\_escalated}+\beta_6\text{case\_age\_escalated}$$

Table 11: Logistic Regression Model Summary

Predictor	Coefficient	Standard Error	Z-Value	P-Value
Intercept	-16.9861131	520.5707984	-0.0326298	0.9739698
Number of Cases	0.0490863	0.0283443	1.7317874	0.0833114
Case Age (Days)	0.0053191	0.0042486	1.2519619	0.2105837
Is Escalated	0.9722940	0.2936043	3.3115790	0.0009277
siteBundoora	16.0283026	520.5708681	0.0307899	0.9754371
siteCamberwell	15.5542850	520.5708510	0.0298793	0.9761633
siteCaribbean Park	15.3325069	520.5707905	0.0294533	0.9765031
siteCasey Corporate Centre	0.1915503	1477.8322645	0.0001296	0.9998966
siteChadstone	15.0396641	520.5707892	0.0288907	0.9769517
siteEastland	16.0908478	520.5707995	0.0309100	0.9753413
siteGENERAL ENQUIRY	33.4407111	2455.3632730	0.0136195	0.9891336
siteNarre Warren	15.0737723	520.5708114	0.0289562	0.9768995
siteRichmond	15.3276762	520.5709083	0.0294440	0.9765105
siteSixty Four on Victor	12.3498880	520.5717436	0.0237237	0.9810730
num_cases_escalated	-0.0572069	0.0520342	-1.0994102	0.2715892
case_age_escalated	0.0041568	0.0127943	0.3248927	0.7452623

Table 12: Interaction Model Fit - McFadden's R^2

Metric	Value
McFadden's R^2	0.0882911

4.5.2 Model Interpretation and Summary

In summary, the addition of the interaction terms pose no new insights in our analysis as both the additions are not statistically significant at the 5% confidence level. The escalated case variable remains to be the only significant contributor our model. The results can be viewed in table 11.

```
## fitting null model for pseudo-r2
```

Moreover, As we can see in table 12, there is no real change in the R^2 value that remains at 0.088.

Table 13: Model Comparison: AIC and BIC

Model	AIC	BIC
Logistic Model	1287.547	1360.015
Interaction Model	1290.277	1373.097

4.6 Model Comparison

Finally, we compare the AIC and BIC values of the models as shown in table 13. A model with the lower AIC and BIC value is preferred and in our case, we observe that the logistic model without the interaction terms performs better in fitting on both fronts without needing to be too complex. While the logistic model with the interaction terms captured some of the complexities offered by the interaction terms, it did not improve the model fit sufficiently to justify the need to add the said interaction terms.

4.7 Modelling Assumptions and Limitations

- **Logical Representation of Customers Churning:** A customer either leaves or does not leave and thus its churn status is defined as either 1 if they leave, or 0 otherwise. It is important to note that this is an oversimplification of churn as there may be numerous attributes and factors that are not captures this way since churn behaviour can be influenced by gradual dissatisfaction which is difficult to quantify.
- **Independent and Unique Cases:** We assume that a case raised by a customer has an independent effect on churn. In reality, a customer may raise multiple cases that may be related to the original, indicating an escalated case. While we do have a column that displays whether a case is escalated or not, the analysis would not capture a cumulative effect of multiple cases that may lead to dissatisfaction.
- **Static Location:** We assume that once a customer purchases a subscription from Waterman Workspaces, the site where they are based in does not change. The static nature of the customer ensures a case is handled within a specified site. If a customer's attribute was dynamic between sites, our analysis could potentially fall to misrepresentation of their influence on churn.
- **Limited Predictors:** While the dataset has enough variables for us to perform our analysis, they are not enough for a more comprehensive and accurate analysis which is also reflected by the R-Squared value. Having more factors such as customer satisfaction ratings and feedback could enrich our model's predictive capacity.
- **Data Quality:** Product category was initially deemed as an important predictor to inspect for our analysis. However, about half of the observations were missing after forming a dataset by joining the cases and membership data. Moreover, important variables such as customer satisfaction, SLA, etc. had no records or entries. Their presence would have allowed for a richer and more detailed analysis.
- **Limited Time-series Analysis:** The data provided did not contain dates and time for when a customer raised an issue and when it was resolved. Instead, the time taken to resolve the cases in minutes and days was provided. Having dates and time would have allowed for some sort of trend analysis and also have us factor in weekends to correctly assume that a case would not be resolved during, say, Saturday and Sunday.
- **Selection Bias:** Since we are analyzing how cases may lead to a customer churning, the data used is only coming from the CRM system. Customers may show dissatisfaction without even raising any cases in the system and such behaviours are underrepresented in our analysis.

Understanding and accepting the limitations and assumptions allows for a transparent view of the difficulties encountered during the phase of the internship. Having limited variables has been very challenging in presenting more material and limited the ability to have more comprehensive modelling and discussions.

4.8 Self Reflection

The tasks and the duration of the placement have been an eventful, insightful, and humbling experiences. Not only did the experience hone my taste and interest in interrogating the numbers more in the future, but it also exposed me to a realm of data in the context for businesses. While there were numerous challenges that I have faced up till finishing writing the report, there was a lot of learning in the process.

The first month and a half was spent in data collection, cleaning, and wrangling. During this time, I learnt how and what to ask for in the spirit of conduction good analysis. Moreover, the numerous meetings held with the supervisor allowed for ease in identifying the variables and observations that were needed for the analysis and which ones needed to be removed. Moreover, as the discovery of data took place, an opportunity to get to understand the business better presented itself as I interacted with numerous coworkers in the office.

The most challenging part, however, was that there was new information regarding the data as soon as I neared the completion of my analysis initially. This meant that new data had to be loaded and understood again; essentially back to square one. The silver lining was that my initial analysis did not produce any significant estimations for customers churning. While my supervisor and the team reassured me that the task is purely for precautionary reasons and there really may be nothing to find, it did not sit right with me. The new data then shed light on new insights and I finally got a significant predictor for customer churn. The only issue was managing time and it cost me timely submissions of my presentations and report. More importantly, it has been a difficult challenge expanding on my work considering the nature of the work given the type of data and analysis required of me which, to me, lacked in more substance and content.

Finally, the units taught during my course have allowed me to demonstrate the skills in a practical and industry setting. In technical learning, being a jack of all trades (having learnt numerous units) made it possible to explore multiple options for our analysis. I have taken many lessons from this experience and I fully intend to apply them in the next phase of my life where I enter the professional realm. I will take the appreciations of data analysis, exploring my problem-solving and adaptability skills, and resiliency in the face of adversity forward.

4.9 Software and Tools Used

The software used for our analysis purposes has purely been R Language on R studio. According to the following research, it is believed that although being a little unfriendly to use, R is perhaps today's best tool for statistical data processing. It provides reliable calculations and almost immeasurable data processing capabilities, but requires at least basic knowledge of the statistical methods used (Hackenberger, 2020). This is a testament to the great capability of the programming language as it offers numerous relevant packages for the task. Moreover, R Studio provides a plethora of options when it comes to presenting a report or, in fact, making presentations. Knitting options such PDF, Word, and HTML are provided with sources such as markdown and Quarto, providing immense flexibility for the user to be creative.

The packages provided by R are easy to use and understand, offering numerous ways to process, quantify, and present data. For the purpose of our tasks and report, the following packages were used:

- **tidyverse:** A powerful collection of packages that work together seamlessly for data wrangling, visualization, and analysis. Key packages used for our tasks were `dplyr` for wrangling and manipulation, `ggplot2` for visualization, and `tidyr` for reshaping our data in the desired output (Wickham et al., 2019).
- **readxl:** This package allowed for importing Excel files that were in `.xls` or `.xl` format (Wickham & Bryan, 2019).
- **visdat:** The package assist in making visualizations that highlight data types, missing values, and data distributions. For diagnostic purposes, we used the functions `vis_dat` and `vis_miss` (Tierney, 2021).

-kableExtra: To be able to articulate outputs that were best viewed in a table format, this package was used for a more clean and appealing look. It is an enhancement on the basic `kable` function to produce more polished reports (Zhu, 2020).

-knitr: Without this package, it would not have been possible to produce this report as it compiles and integrates all the outputs within a single document (Xie, 2015).

-bookdown: This package is useful for giving structure to our report as it helps with generating sections, chapters, and citations. This makes longer documents easier to use by providing table of contents and the ability to cross reference (Xie, 2016).

-patchwork and gridExtra: Used for arranging multiple `ggplot2` outputs in the document. The packages allowed for a better side-by-side comparison of outputs where necessary (Pedersen, 2020) and (Auguie, 2017).

-broom: The package converts model outputs into tidy data frames. In our report, we used the package to display the logistic regression models' outputs for better readibility and interpretation ease by extracting and summarizing the model results (Robinson et al., 2024).

-pscl: This package provides tools for assessing model fit in logistic regression. For our case, we used it to extract the McFadden's R^2 as a measure of goodness of fit to evaluate our models (Jackman, 2020).

-caret: While this package is used extensively for evaluating and training machine learning models, it allowed us to assess the importance of the variables used in our logistic regression models (Kuhn, 2008).

5 Conclusion

In conclusion, our analysis explored the factors deemed to be critical for customers churning. We made use of the logistic regression model after intense and comprehensive data cleaning efforts. The use of the logistic regression model was to aide in ease of interpretibility. It was found that while some variability is captured given the variables by our model, there is a need for more explanatory variables. If a case is escalated, a customer is more likely to leave that a customers that do not escalate cases. This was the most statistically significant variable while site, the number of cases per customer, and the average time it took to resolve a case were not statistically significant.

5.1 Limitatations

The lack of relevant variables in our data has been a consistent complaint. Such a challenge gives birth to the lack of ability to innovate when it comes to trying methods and finding exciting solutions. For example, having customer sentiments after resolving a case could provide for an opportunity to perform sentiment analysis using textual data to give deeper meaning to our insights. Moreover, having correctly labelled case lodge and resolve dates and time would allow for time-series forecasting as well, and if there is enough data over the years, churning seasonality could be identified. This would allow for expanding on such peaks of churn rates and dive deeper into the data to investigate what may have led to such leavers. Thus, data completeness in a longer time-frame is critical for such analysis to produce meaningful insights. There are exciting possibilities if some of these limitations, if not all, are addressed.

5.2 Practical Implications, Recommendations, and Future Analysis

Even with the data quality challenges, there were some actionable insights found that Waterman Workspaces could put to good effect. By focusing on the critical metrics, the organization can reduce their churn rates to desired natural levels. Proactively engaging with the customers who have a higher number of cases lodged during their stay as a customer ensures build of trust. Furthermore, it is recommended to handle cases before there is a need for them to be escalated. However, in the event of cases being escalated, protocols should be developed to resolve them in a timely fashion. Furthermore, more relevant variables would need to be added to have better fitting models to explain the variability and to explore their significance. Data collected by sending out satisfaction surveys may be used in conjunction to enhance the findings and develop further actionable insights.

6 References

- Hackenberger, B. K. (2020). R software: unfriendly but probably the best. Croatian Medical Journal, 61(1), 66–68. https://doi.org/10.3325/cmj.2020.61.66

- Orn, A. (2023, December 3). Means and Medians: When To Use Which - Research Collective. Research-Collective.com. https://research-collective.com/means-and-medians-when-to-use-which/

- Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96(1), 3–14. https://doi.org/10.1080/00220670209598786

- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

- Wickham H, Bryan J (2023). *readxl: Read Excel Files.* R package version 1.4.3, https://CRAN.R-project.org/package=readxl.

- Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi:10.21105/joss.00355 https://doi.org/10.21105/joss.00355, http://dx.doi.org/10.21105/joss.00355.

- Zhu H (2024). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* R package version 1.4.0, https://CRAN.R-project.org/package=kableExtra.

- Xie Y (2024). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.48, https://yihui.org/knitr/. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963 Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

- Xie Y (2024). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.40, https://github.com/rstudio/bookdown. Xie Y (2016). *bookdown: Authoring Books and Technical Documents with R Markdown.* Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138700109, https://bookdown.org/yihui/bookdown.

- Pedersen T (2024). *patchwork: The Composer of Plots.* R package version 1.3.0, https://CRAN.R-project.org/package=patchwork.

- Robinson D, Hayes A, Couch S (2024). *broom: Convert Statistical Objects into Tidy Tibbles.* R package version 1.0.6, https://CRAN.R-project.org/package=broom.

- Simon Jackman (2024). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. Sydney, Australia. R package version 1.5.9. URL https://github.com/atahk/pscl/

- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05