# Advance Analytics Assignment_1

Dhanush Mathighatta Shobhan Babu (40412492)

15th March 2024

## Load the libraries

#/dataset loading ### Loading the data set

```
## [1] "DATA LOADED!"
```

#Writing the train test function Dividing the dataset into a train-test split with a ratio of 70-30.The seed value 40412492 was assigned to ensure reproducibility.

### Lasso Regression Function Code.

### Logistic Regression Code.

### K fold for logistic Regression.

### results visualisation

## 1.0 Introduction

This analysis involves a thorough investigation of a dataset to forecast the "Issue Consequence" variable, utilising the CRISP-DM framework as our reference. The study commences with an initial examination of the data, known as exploratory data analysis (EDA), during which we reveal the characteristics and distribution patterns of the dataset.

After that, we will concentrate on building models, specifically utilising Lasso Regression to identify the 10 most significant variables that impact our objective variable. Using these essential elements, we subsequently construct logistic regression and K-Nearest Neighbours (KNN) models to further evaluate their ability to make accurate predictions.

During the model evaluation stage, we conduct a comprehensive analysis of each model's performance, utilising a range of metrics to gain a complete understanding of their effectiveness. The purpose of this thorough assessment is to produce valuable insights and recommendations that can assist in decision-making.

By utilising the CRISP-DM framework, our investigation adheres to a systematic and structured approach, guaranteeing the dependability and accuracy of our discoveries. Utilising this systematic methodology facilitates the process of generating educated business decisions and improves our understanding of the intricacies and interconnections within the information.

```
##                                                        vars      n
mean
## ID_non_uniq                                               1 566760
NaN
## date_event                                                2 566760
NaN
## last_year_all_product_codes_num_uniq                      3 566760
0.61
## last_year_all_product_codes_most_freq                     4 566760    34
6.97
## last_year_brand_name_num_uniq                             5 566760
2.28
## last_year_brand_name_most_freq                            6 566760    65
9.53
## last_year_classification0_num_uniq                        7 566760
0.00
## last_year_classification1_num_uniq                        8 566760
7.97
## last_year_classification2_num_uniq                        9 566760
1.68
## last_year_company_name_num_uniq                          10 566760
0.28
## last_year_company_name_most_freq                         11 566760     6
9.15
## last_year_reason_for_legal_announcement_num_uniq         12 566760
0.27
## last_year_reason_for_legal_announcement_most_freq        13 566760     9
1.76
## last_year_legal_announcementing_firm_num_uniq            14 566760
0.25
## last_year_legal_announcementing_firm_most_freq           15 566760     5
5.20
## last_year_root_cause_description_num_uniq                16 566760
0.27
## last_year_root_cause_description_most_freq               17 566760
5.64
## last_year_product_quantity_average_num_uniq              18 566760
0.14
## last_year_product_quantity_average_max                   19 566760    20
8.71
## last_year_product_quantity_average_average               20 566760    17
2.70
## last_year_decision_date_max_changes_in_product           21 566760
3.54
## last_year_decision_date_average_changes_in_product       22 566760
3.54
## last_two_years_all_product_codes_num_uniq                23 566760
1.33
## last_two_years_all_product_codes_most_freq               24 566760    66
4.12
```

```
## last_two_years_brand_name_num_uniq                              25 566760
4.22
## last_two_years_brand_name_most_freq                             26 566760   147
9.28
## last_two_years_classification0_num_uniq                         27 566760
0.00
## last_two_years_classification1_num_uniq                         28 566760     1
7.62
## last_two_years_classification2_num_uniq                         29 566760
1.70
## last_two_years_company_name_num_uniq                            30 566760
0.60
## last_two_years_company_name_most_freq                           31 566760    16
1.34
## last_two_years_reason_for_legal_announcement_num_uniq           32 566760
0.58
## last_two_years_reason_for_legal_announcement_most_freq          33 566760    17
4.46
## last_two_years_legal_announcementing_firm_num_uniq              34 566760
0.55
## last_two_years_legal_announcementing_firm_most_freq             35 566760    13
1.50
## last_two_years_root_cause_description_num_uniq                  36 566760
0.58
## last_two_years_root_cause_description_most_freq                 37 566760     1
2.22
## last_two_years_product_quantity_average_num_uniq                38 566760
0.22
## last_two_years_product_quantity_average_max                     39 566760    27
7.47
## last_two_years_product_quantity_average_average                 40 566760    19
0.00
## last_two_years_decision_date_max_changes_in_product             41 566760     1
7.32
## last_two_years_decision_date_average_changes_in_product         42 566760     1
7.32
## last_four_years_all_product_codes_num_uniq                      43 566760
2.44
## last_four_years_all_product_codes_most_freq                     44 566760   117
9.14
## last_four_years_brand_name_num_uniq                             45 566760
7.49
## last_four_years_brand_name_most_freq                            46 566760   272
6.89
## last_four_years_classification0_num_uniq                        47 566760
0.00
## last_four_years_classification1_num_uniq                        48 566760     3
2.39
## last_four_years_classification2_num_uniq                        49 566760
1.72
```

```
## last_four_years_company_name_num_uniq                          50 566760
1.09
## last_four_years_company_name_most_freq                         51 566760    30
0.24
## last_four_years_reason_for_legal_announcement_num_uniq         52 566760
1.04
## last_four_years_reason_for_legal_announcement_most_freq        53 566760    29
8.88
## last_four_years_legal_announcementing_firm_num_uniq            54 566760
1.01
## last_four_years_legal_announcementing_firm_most_freq           55 566760    24
4.86
## last_four_years_root_cause_description_num_uniq                56 566760
1.04
## last_four_years_root_cause_description_most_freq               57 566760     2
2.51
## last_four_years_product_quantity_average_num_uniq              58 566760
0.36
## last_four_years_product_quantity_average_max                   59 566760    28
6.67
## last_four_years_product_quantity_average_average               60 566760    19
1.16
## last_four_years_decision_date_max_changes_in_product           61 566760     6
6.28
## last_four_years_decision_date_average_changes_in_product       62 566760     6
6.28
## Proudct.issue.consequence                                      63 566760
NaN
## manufacturer_contact_address_1                                 64 566760   828
7.71
## product.brand_name                                             65 566760 18514
1.71
## product.generic_name                                           66 566760  7019
0.09
## product.issue.type                                             67 566760    55
4.29
## type_of_report.1                                               68 566760
0.26
## reporter_job_code                                              69 566760     2
9.08
## source_type                                                    70 566760
5.89
## product.manufacturer_name                                      71 566760  1612
1.13
## product.product_operator                                       72 566760     1
6.47
## product.manufacturer_city                                      73 566760   538
3.26
## product.manufacturer_state                                     74 566760     4
4.98
```

```
## product.manufacturer_country                                  75 566760    10
1.12
## product.field_description                                     76 566760
NaN
## product.product_report_product_code                           77 566760
NaN
##                                                                    sd   min
## ID_non_uniq                                                        NA   Inf
## date_event                                                         NA   Inf
## last_year_all_product_codes_num_uniq                             1.20     0
## last_year_all_product_codes_most_freq                          815.22     0
## last_year_brand_name_num_uniq                                    6.32     0
## last_year_brand_name_most_freq                                1253.49     0
## last_year_classification0_num_uniq                               0.02     0
## last_year_classification1_num_uniq                              26.97     0
## last_year_classification2_num_uniq                              19.04     0
## last_year_company_name_num_uniq                                  0.51     0
## last_year_company_name_most_freq                               133.25     0
## last_year_reason_for_legal_announcement_num_uniq                 0.54     0
## last_year_reason_for_legal_announcement_most_freq              224.88     0
## last_year_legal_announcementing_firm_num_uniq                    0.44     0
## last_year_legal_announcementing_firm_most_freq                 105.88     0
## last_year_root_cause_description_num_uniq                        0.54     0
## last_year_root_cause_description_most_freq                      10.30     0
## last_year_product_quantity_average_num_uniq                      0.47     0
## last_year_product_quantity_average_max                        3472.08     0
## last_year_product_quantity_average_average                    1793.32     0
## last_year_decision_date_max_changes_in_product                   7.19     0
## last_year_decision_date_average_changes_in_product               7.19     0
## last_two_years_all_product_codes_num_uniq                        1.46     0
## last_two_years_all_product_codes_most_freq                     976.48     0
## last_two_years_brand_name_num_uniq                               7.73     0
## last_two_years_brand_name_most_freq                           1525.47     0
## last_two_years_classification0_num_uniq                          0.02     0
## last_two_years_classification1_num_uniq                         35.87     0
## last_two_years_classification2_num_uniq                         19.07     0
## last_two_years_company_name_num_uniq                             0.59     0
## last_two_years_company_name_most_freq                          165.18     0
## last_two_years_reason_for_legal_announcement_num_uniq            0.60     0
## last_two_years_reason_for_legal_announcement_most_freq         228.14     0
## last_two_years_legal_announcementing_firm_num_uniq               0.51     0
## last_two_years_legal_announcementing_firm_most_freq            132.97     0
## last_two_years_root_cause_description_num_uniq                   0.60     0
## last_two_years_root_cause_description_most_freq                 11.70     0
## last_two_years_product_quantity_average_num_uniq                 0.54     0
## last_two_years_product_quantity_average_max                   5986.93     0
## last_two_years_product_quantity_average_average               2137.96     0
## last_two_years_decision_date_max_changes_in_product             18.30     0
## last_two_years_decision_date_average_changes_in_product         18.30     0
## last_four_years_all_product_codes_num_uniq                       1.09     1
```

```
## last_four_years_all_product_codes_most_freq                1010.38   403
## last_four_years_brand_name_num_uniq                           8.90     1
## last_four_years_brand_name_most_freq                        944.35   249
## last_four_years_classification0_num_uniq                      0.18     0
## last_four_years_classification1_num_uniq                     46.13     1
## last_four_years_classification2_num_uniq                     19.18     0
## last_four_years_company_name_num_uniq                         0.29     1
## last_four_years_company_name_most_freq                       96.94    65
## last_four_years_reason_for_legal_announcement_num_uniq        0.36     1
## last_four_years_reason_for_legal_announcement_most_freq     166.19    95
## last_four_years_legal_announcementing_firm_num_uniq           0.12     1
## last_four_years_legal_announcementing_firm_most_freq         72.95    35
## last_four_years_root_cause_description_num_uniq               0.32     1
## last_four_years_root_cause_description_most_freq              3.99     4
## last_four_years_product_quantity_average_num_uniq             0.63     0
## last_four_years_product_quantity_average_max               6030.02     0
## last_four_years_product_quantity_average_average           1963.19     0
## last_four_years_decision_date_max_changes_in_product         24.44     0
## last_four_years_decision_date_average_changes_in_product     24.44     0
## Proudct.issue.consequence                                       NA   Inf
## manufacturer_contact_address_1                             3383.57   118
## product.brand_name                                        58456.61 14520
## product.generic_name                                      24975.05 13465
## product.issue.type                                          237.50     1
## type_of_report.1                                              0.44     0
## reporter_job_code                                            11.85     1
## source_type                                                   3.58     3
## product.manufacturer_name                                  6233.49   924
## product.product_operator                                      2.60     0
## product.manufacturer_city                                  1691.57   363
## product.manufacturer_state                                   15.11     8
## product.manufacturer_country                                 47.61     9
## product.field_description                                       NA   Inf
## product.product_report_product_code                             NA   Inf
##                                                                max     ran
ge
## ID_non_uniq                                                   -Inf      -I
nf
## date_event                                                    -Inf      -I
nf
## last_year_all_product_codes_num_uniq                          9.00      9.
00
## last_year_all_product_codes_most_freq                      6253.00   6253.
00
## last_year_brand_name_num_uniq                                37.00     37.
00
## last_year_brand_name_most_freq                             4789.00   4789.
00
## last_year_classification0_num_uniq                            6.00      6.
00
```

```
## last_year_classification1_num_uniq                          1800.00   1800.
00
## last_year_classification2_num_uniq                           624.00    624.
00
## last_year_company_name_num_uniq                                2.00      2.
00
## last_year_company_name_most_freq                             548.00    548.
00
## last_year_reason_for_legal_announcement_num_uniq               5.00      5.
00
## last_year_reason_for_legal_announcement_most_freq           1518.00   1518.
00
## last_year_legal_announcementing_firm_num_uniq                  3.00      3.
00
## last_year_legal_announcementing_firm_most_freq               442.00    442.
00
## last_year_root_cause_description_num_uniq                      4.00      4.
00
## last_year_root_cause_description_most_freq                    40.00     40.
00
## last_year_product_quantity_average_num_uniq                    5.00      5.
00
## last_year_product_quantity_average_max                    406985.00 406985.
00
## last_year_product_quantity_average_average                 99622.60  99622.
60
## last_year_decision_date_max_changes_in_product                33.00     33.
00
## last_year_decision_date_average_changes_in_product            33.00     33.
00
## last_two_years_all_product_codes_num_uniq                      9.00      9.
00
## last_two_years_all_product_codes_most_freq                  6253.00   6253.
00
## last_two_years_brand_name_num_uniq                            37.00     37.
00
## last_two_years_brand_name_most_freq                         4789.00   4789.
00
## last_two_years_classification0_num_uniq                        6.00      6.
00
## last_two_years_classification1_num_uniq                     2250.00   2250.
00
## last_two_years_classification2_num_uniq                      624.00    624.
00
## last_two_years_company_name_num_uniq                           2.00      2.
00
## last_two_years_company_name_most_freq                        548.00    548.
00
## last_two_years_reason_for_legal_announcement_num_uniq          7.00      7.
00
```

```
## last_two_years_reason_for_legal_announcement_most_freq     1518.00   1518.
00
## last_two_years_legal_announcementing_firm_num_uniq            4.00      4.
00
## last_two_years_legal_announcementing_firm_most_freq         442.00    442.
00
## last_two_years_root_cause_description_num_uniq                5.00      5.
00
## last_two_years_root_cause_description_most_freq              40.00     40.
00
## last_two_years_product_quantity_average_num_uniq             7.00      7.
00
## last_two_years_product_quantity_average_max             406985.00 406985.
00
## last_two_years_product_quantity_average_average         107726.46 107726.
46
## last_two_years_decision_date_max_changes_in_product          63.00     63.
00
## last_two_years_decision_date_average_changes_in_product      63.00     63.
00
## last_four_years_all_product_codes_num_uniq                   9.00      8.
00
## last_four_years_all_product_codes_most_freq                6253.00   5850.
00
## last_four_years_brand_name_num_uniq                          37.00     36.
00
## last_four_years_brand_name_most_freq                       4789.00   4540.
00
## last_four_years_classification0_num_uniq                     45.00     45.
00
## last_four_years_classification1_num_uniq                   4050.00   4049.
00
## last_four_years_classification2_num_uniq                    624.00    624.
00
## last_four_years_company_name_num_uniq                        2.00      1.
00
## last_four_years_company_name_most_freq                     548.00    483.
00
## last_four_years_reason_for_legal_announcement_num_uniq      11.00     10.
00
## last_four_years_reason_for_legal_announcement_most_freq   1518.00   1423.
00
## last_four_years_legal_announcementing_firm_num_uniq          5.00      4.
00
## last_four_years_legal_announcementing_firm_most_freq       442.00    407.
00
## last_four_years_root_cause_description_num_uniq              8.00      7.
00
## last_four_years_root_cause_description_most_freq            36.00     32.
00
```

```
## last_four_years_product_quantity_average_num_uniq             9.00      9.
00
## last_four_years_product_quantity_average_max             406985.00 406985.
00
## last_four_years_product_quantity_average_average          87534.29  87534.
29
## last_four_years_decision_date_max_changes_in_product        118.00    118.
00
## last_four_years_decision_date_average_changes_in_product    118.00    118.
00
## Proudct.issue.consequence                                     -Inf      -I
nf
## manufacturer_contact_address_1                            13145.00  13027.
00
## product.brand_name                                       344588.00 330068.
00
## product.generic_name                                     101028.00  87563.
00
## product.issue.type                                          964.00    963.
00
## type_of_report.1                                              1.00      1.
00
## reporter_job_code                                            52.00     51.
00
## source_type                                                  24.00     21.
00
## product.manufacturer_name                                 31471.00  30547.
00
## product.product_operator                                     52.00     52.
00
## product.manufacturer_city                                 10778.00  10415.
00
## product.manufacturer_state                                   63.00     55.
00
## product.manufacturer_country                                135.00    126.
00
## product.field_description                                     -Inf      -I
nf
## product.product_report_product_code                           -Inf      -I
nf
##                                                              se
## ID_non_uniq                                                  NA
## date_event                                                   NA
## last_year_all_product_codes_num_uniq                       0.00
## last_year_all_product_codes_most_freq                      1.08
## last_year_brand_name_num_uniq                              0.01
## last_year_brand_name_most_freq                             1.67
## last_year_classification0_num_uniq                         0.00
## last_year_classification1_num_uniq                         0.04
## last_year_classification2_num_uniq                         0.03
```

```
## last_year_company_name_num_uniq                                0.00
## last_year_company_name_most_freq                               0.18
## last_year_reason_for_legal_announcement_num_uniq               0.00
## last_year_reason_for_legal_announcement_most_freq              0.30
## last_year_legal_announcementing_firm_num_uniq                  0.00
## last_year_legal_announcementing_firm_most_freq                 0.14
## last_year_root_cause_description_num_uniq                      0.00
## last_year_root_cause_description_most_freq                     0.01
## last_year_product_quantity_average_num_uniq                    0.00
## last_year_product_quantity_average_max                         4.61
## last_year_product_quantity_average_average                     2.38
## last_year_decision_date_max_changes_in_product                 0.01
## last_year_decision_date_average_changes_in_product             0.01
## last_two_years_all_product_codes_num_uniq                      0.00
## last_two_years_all_product_codes_most_freq                     1.30
## last_two_years_brand_name_num_uniq                             0.01
## last_two_years_brand_name_most_freq                            2.03
## last_two_years_classification0_num_uniq                        0.00
## last_two_years_classification1_num_uniq                        0.05
## last_two_years_classification2_num_uniq                        0.03
## last_two_years_company_name_num_uniq                           0.00
## last_two_years_company_name_most_freq                          0.22
## last_two_years_reason_for_legal_announcement_num_uniq          0.00
## last_two_years_reason_for_legal_announcement_most_freq         0.30
## last_two_years_legal_announcementing_firm_num_uniq             0.00
## last_two_years_legal_announcementing_firm_most_freq            0.18
## last_two_years_root_cause_description_num_uniq                 0.00
## last_two_years_root_cause_description_most_freq                0.02
## last_two_years_product_quantity_average_num_uniq               0.00
## last_two_years_product_quantity_average_max                    7.95
## last_two_years_product_quantity_average_average                2.84
## last_two_years_decision_date_max_changes_in_product            0.02
## last_two_years_decision_date_average_changes_in_product        0.02
## last_four_years_all_product_codes_num_uniq                     0.00
## last_four_years_all_product_codes_most_freq                    1.34
## last_four_years_brand_name_num_uniq                            0.01
## last_four_years_brand_name_most_freq                           1.25
## last_four_years_classification0_num_uniq                       0.00
## last_four_years_classification1_num_uniq                       0.06
## last_four_years_classification2_num_uniq                       0.03
## last_four_years_company_name_num_uniq                          0.00
## last_four_years_company_name_most_freq                         0.13
## last_four_years_reason_for_legal_announcement_num_uniq         0.00
## last_four_years_reason_for_legal_announcement_most_freq        0.22
## last_four_years_legal_announcementing_firm_num_uniq            0.00
## last_four_years_legal_announcementing_firm_most_freq           0.10
## last_four_years_root_cause_description_num_uniq                0.00
## last_four_years_root_cause_description_most_freq               0.01
## last_four_years_product_quantity_average_num_uniq              0.00
## last_four_years_product_quantity_average_max                   8.01
```

```
## last_four_years_product_quantity_average_average          2.61
## last_four_years_decision_date_max_changes_in_product      0.03
## last_four_years_decision_date_average_changes_in_product  0.03
## Proudct.issue.consequence                                   NA
## manufacturer_contact_address_1                            4.49
## product.brand_name                                       77.65
## product.generic_name                                     33.17
## product.issue.type                                        0.32
## type_of_report.1                                          0.00
## reporter_job_code                                         0.02
## source_type                                               0.00
## product.manufacturer_name                                 8.28
## product.product_operator                                  0.00
## product.manufacturer_city                                 2.25
## product.manufacturer_state                                0.02
## product.manufacturer_country                              0.06
## product.field_description                                   NA
## product.product_report_product_code                         NA

##  ID_non_uniq          date_event         last_year_all_product_codes_num_u
niq
##  Length:566760      Min.   :2007-09-27   Min.   :0.0000
##  Class :character   1st Qu.:2016-04-01   1st Qu.:0.0000
##  Mode  :character   Median :2017-01-11   Median :0.0000
##                     Mean   :2017-02-10   Mean   :0.6068
##                     3rd Qu.:2017-12-27   3rd Qu.:0.0000
##                     Max.   :2022-06-14   Max.   :9.0000
##  last_year_all_product_codes_most_freq last_year_brand_name_num_uniq
##  Min.   :   0                          Min.   : 0.000
##  1st Qu.:   0                          1st Qu.: 0.000
##  Median :   0                          Median : 0.000
##  Mean   : 347                          Mean   : 2.279
##  3rd Qu.:   0                          3rd Qu.: 0.000
##  Max.   :6253                          Max.   :37.000
##  last_year_brand_name_most_freq last_year_classification0_num_uniq
##  Min.   :   0.0                 Min.   :0.000000
##  1st Qu.:   0.0                 1st Qu.:0.000000
##  Median :   0.0                 Median :0.000000
##  Mean   : 659.5                 Mean   :0.000106
##  3rd Qu.:   0.0                 3rd Qu.:0.000000
##  Max.   :4789.0                 Max.   :6.000000
##  last_year_classification1_num_uniq last_year_classification2_num_uniq
##  Min.   :   0.000                   Min.   :  0.000
##  1st Qu.:   0.000                   1st Qu.:  0.000
##  Median :   0.000                   Median :  0.000
##  Mean   :   7.968                   Mean   :  1.684
##  3rd Qu.:   0.000                   3rd Qu.:  0.000
##  Max.   :1800.000                   Max.   :624.000
##  last_year_company_name_num_uniq last_year_company_name_most_freq
##  Min.   :0.000                    Min.   :  0.00
```

```
##  1st Qu.:0.000                   1st Qu.:  0.00
##  Median :0.000                   Median :  0.00
##  Mean   :0.275                   Mean   : 69.15
##  3rd Qu.:0.000                   3rd Qu.:  0.00
##  Max.   :2.000                   Max.   :548.00
##  last_year_reason_for_legal_announcement_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2727
##  3rd Qu.:0.0000
##  Max.   :5.0000
##  last_year_reason_for_legal_announcement_most_freq
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.00
##  Mean   :  91.76
##  3rd Qu.:   0.00
##  Max.   :1518.00
##  last_year_legal_announcementing_firm_num_uniq
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.246
##  3rd Qu.:0.000
##  Max.   :3.000
##  last_year_legal_announcementing_firm_most_freq
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median :  0.0
##  Mean   : 55.2
##  3rd Qu.:  0.0
##  Max.   :442.0
##  last_year_root_cause_description_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2722
##  3rd Qu.:0.0000
##  Max.   :4.0000
##  last_year_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median : 0.00
##  Mean   : 5.64
##  3rd Qu.: 0.00
##  Max.   :40.00
##  last_year_product_quantity_average_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
```

```
##  Median :0.0000
##  Mean   :0.1435
##  3rd Qu.:0.0000
##  Max.   :5.0000
##  last_year_product_quantity_average_max
##  Min.   :     0.0
##  1st Qu.:     0.0
##  Median :     0.0
##  Mean   :   208.7
##  3rd Qu.:     0.0
##  Max.   :406985.0
##  last_year_product_quantity_average_average
##  Min.   :    0.0
##  1st Qu.:    0.0
##  Median :    0.0
##  Mean   :  172.7
##  3rd Qu.:    0.0
##  Max.   :99622.6
##  last_year_decision_date_max_changes_in_product
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 3.541
##  3rd Qu.: 0.000
##  Max.   :33.000
##  last_year_decision_date_average_changes_in_product
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 3.541
##  3rd Qu.: 0.000
##  Max.   :33.000
##  last_two_years_all_product_codes_num_uniq
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.000
##  Mean   :1.327
##  3rd Qu.:2.000
##  Max.   :9.000
##  last_two_years_all_product_codes_most_freq last_two_years_brand_name_num_
uniq
##  Min.   :   0.0                             Min.   : 0.00
##  1st Qu.:   0.0                             1st Qu.: 0.00
##  Median : 437.0                             Median : 1.00
##  Mean   : 664.1                             Mean   : 4.22
##  3rd Qu.: 460.0                             3rd Qu.: 3.00
##  Max.   :6253.0                             Max.   :37.00
##  last_two_years_brand_name_most_freq last_two_years_classification0_num_un
iq
##  Min.   :   0                        Min.   :0.000000
```

```
##  1st Qu.:   0                         1st Qu.:0.000000
##  Median :2247                         Median :0.000000
##  Mean   :1479                         Mean   :0.000106
##  3rd Qu.:2346                         3rd Qu.:0.000000
##  Max.   :4789                         Max.   :6.000000
##  last_two_years_classification1_num_uniq
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   3.00
##  Mean   :  17.62
##  3rd Qu.:  20.00
##  Max.   :2250.00
##  last_two_years_classification2_num_uniq last_two_years_company_name_num_u
niq
##  Min.   :  0.000                         Min.   :0.0000
##  1st Qu.:  0.000                         1st Qu.:0.0000
##  Median :  0.000                         Median :1.0000
##  Mean   :  1.696                         Mean   :0.5954
##  3rd Qu.:  0.000                         3rd Qu.:1.0000
##  Max.   :624.000                         Max.   :2.0000
##  last_two_years_company_name_most_freq
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median : 66.0
##  Mean   :161.3
##  3rd Qu.:349.0
##  Max.   :548.0
##  last_two_years_reason_for_legal_announcement_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5762
##  3rd Qu.:1.0000
##  Max.   :7.0000
##  last_two_years_reason_for_legal_announcement_most_freq
##  Min.   :   0.0
##  1st Qu.:   0.0
##  Median : 241.0
##  Mean   : 174.5
##  3rd Qu.: 285.0
##  Max.   :1518.0
##  last_two_years_legal_announcementing_firm_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5478
##  3rd Qu.:1.0000
##  Max.   :4.0000
##  last_two_years_legal_announcementing_firm_most_freq
##  Min.   :  0.0
```

```
##  1st Qu.:  0.0
##  Median :142.0
##  Mean   :131.5
##  3rd Qu.:292.0
##  Max.   :442.0
##  last_two_years_root_cause_description_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5756
##  3rd Qu.:1.0000
##  Max.   :5.0000
##  last_two_years_root_cause_description_most_freq
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median :20.00
##  Mean   :12.22
##  3rd Qu.:20.00
##  Max.   :40.00
##  last_two_years_product_quantity_average_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2216
##  3rd Qu.:0.0000
##  Max.   :7.0000
##  last_two_years_product_quantity_average_max
##  Min.   :     0.0
##  1st Qu.:     0.0
##  Median :     0.0
##  Mean   :   277.5
##  3rd Qu.:     0.0
##  Max.   :406985.0
##  last_two_years_product_quantity_average_average
##  Min.   :     0
##  1st Qu.:     0
##  Median :     0
##  Mean   :   190
##  3rd Qu.:     0
##  Max.   :107726
##  last_two_years_decision_date_max_changes_in_product
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median : 9.00
##  Mean   :17.32
##  3rd Qu.:38.00
##  Max.   :63.00
##  last_two_years_decision_date_average_changes_in_product
##  Min.   : 0.00
##  1st Qu.: 0.00
```

```
##  Median : 9.00
##  Mean   :17.32
##  3rd Qu.:38.00
##  Max.   :63.00
##  last_four_years_all_product_codes_num_uniq
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.441
##  3rd Qu.:3.000
##  Max.   :9.000
##  last_four_years_all_product_codes_most_freq
##  Min.   : 403
##  1st Qu.: 445
##  Median : 457
##  Mean   :1179
##  3rd Qu.:2488
##  Max.   :6253
##  last_four_years_brand_name_num_uniq last_four_years_brand_name_most_freq
##  Min.   : 1.000                      Min.   : 249
##  1st Qu.: 2.000                      1st Qu.:2248
##  Median : 3.000                      Median :2346
##  Mean   : 7.486                      Mean   :2727
##  3rd Qu.:13.000                      3rd Qu.:2912
##  Max.   :37.000                      Max.   :4789
##  last_four_years_classification0_num_uniq
##  Min.   : 0.00000
##  1st Qu.: 0.00000
##  Median : 0.00000
##  Mean   : 0.00196
##  3rd Qu.: 0.00000
##  Max.   :45.00000
##  last_four_years_classification1_num_uniq
##  Min.   :   1.00
##  1st Qu.:   6.00
##  Median :  18.00
##  Mean   :  32.39
##  3rd Qu.:  42.00
##  Max.   :4050.00
##  last_four_years_classification2_num_uniq last_four_years_company_name_num
_uniq
##  Min.   :  0.000                          Min.   :1.000
##  1st Qu.:  0.000                          1st Qu.:1.000
##  Median :  0.000                          Median :1.000
##  Mean   :  1.717                          Mean   :1.091
##  3rd Qu.:  0.000                          3rd Qu.:1.000
##  Max.   :624.000                          Max.   :2.000
##  last_four_years_company_name_most_freq
##  Min.   : 65.0
##  1st Qu.:349.0
```

```
##  Median :349.0
##  Mean   :300.2
##  3rd Qu.:349.0
##  Max.   :548.0
##  last_four_years_reason_for_legal_announcement_num_uniq
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 1.043
##  3rd Qu.: 1.000
##  Max.   :11.000
##  last_four_years_reason_for_legal_announcement_most_freq
##  Min.   :  95.0
##  1st Qu.: 241.0
##  Median : 285.0
##  Mean   : 298.9
##  3rd Qu.: 285.0
##  Max.   :1518.0
##  last_four_years_legal_announcementing_firm_num_uniq
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.011
##  3rd Qu.:1.000
##  Max.   :5.000
##  last_four_years_legal_announcementing_firm_most_freq
##  Min.   : 35.0
##  1st Qu.:142.0
##  Median :292.0
##  Mean   :244.9
##  3rd Qu.:292.0
##  Max.   :442.0
##  last_four_years_root_cause_description_num_uniq
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.041
##  3rd Qu.:1.000
##  Max.   :8.000
##  last_four_years_root_cause_description_most_freq
##  Min.   : 4.00
##  1st Qu.:20.00
##  Median :20.00
##  Mean   :22.51
##  3rd Qu.:28.00
##  Max.   :36.00
##  last_four_years_product_quantity_average_num_uniq
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
```

```
##  Mean   :0.3603
##  3rd Qu.:1.0000
##  Max.   :9.0000
##  last_four_years_product_quantity_average_max
##  Min.   :     0.0
##  1st Qu.:     0.0
##  Median :     0.0
##  Mean   :   286.7
##  3rd Qu.:    26.0
##  Max.   :406985.0
##  last_four_years_product_quantity_average_average
##  Min.   :    0.0
##  1st Qu.:    0.0
##  Median :    0.0
##  Mean   :  191.2
##  3rd Qu.:   26.0
##  Max.   :87534.3
##  last_four_years_decision_date_max_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 51.00
##  Median : 74.00
##  Mean   : 66.28
##  3rd Qu.: 87.00
##  Max.   :118.00
##  last_four_years_decision_date_average_changes_in_product
##  Min.   :  0.00
##  1st Qu.: 51.00
##  Median : 74.00
##  Mean   : 66.28
##  3rd Qu.: 87.00
##  Max.   :118.00
##  Proudct.issue.consequence manufacturer_contact_address_1 product.brand_na
me
##  Length:566760             Min.   :  118                  Min.   : 14520
##  Class :character          1st Qu.: 9476                  1st Qu.:158472
##  Mode  :character          Median : 9476                  Median :158472
##                            Mean   : 8288                  Mean   :185142
##                            3rd Qu.: 9476                  3rd Qu.:199471
##                            Max.   :13145                  Max.   :344588
##  product.generic_name product.issue.type type_of_report.1 reporter_job_cod
e
##  Min.   : 13465       Min.   :  1.0      Min.   :0.0000   Min.   : 1.00
##  1st Qu.: 61634       1st Qu.:352.0      1st Qu.:0.0000   1st Qu.:32.00
##  Median : 84946       Median :596.0      Median :0.0000   Median :32.00
##  Mean   : 70190       Mean   :554.3      Mean   :0.2611   Mean   :29.08
##  3rd Qu.: 84946       3rd Qu.:807.0      3rd Qu.:1.0000   3rd Qu.:32.00
##  Max.   :101028       Max.   :964.0      Max.   :1.0000   Max.   :52.00
##   source_type     product.manufacturer_name product.product_operator
##  Min.   : 3.000   Min.   :  924             Min.   : 0.00
##  1st Qu.: 4.000   1st Qu.:18430             1st Qu.:15.00
```

```
##  Median : 4.000   Median :19327             Median :15.00
##  Mean   : 5.886   Mean   :16121             Mean   :16.47
##  3rd Qu.: 9.000   3rd Qu.:19408             3rd Qu.:18.00
##  Max.   :24.000   Max.   :31471             Max.   :52.00
##  product.manufacturer_city product.manufacturer_state
##  Min.   :  363             Min.   : 8.00
##  1st Qu.: 4513             1st Qu.:32.00
##  Median : 4513             Median :48.00
##  Mean   : 5383             Mean   :44.98
##  3rd Qu.: 5990             3rd Qu.:48.00
##  Max.   :10778             Max.   :63.00
##  product.manufacturer_country product.field_description
##  Min.   :  9.0                Length:566760
##  1st Qu.:126.0                Class :character
##  Median :126.0                Mode  :character
##  Mean   :101.1
##  3rd Qu.:126.0
##  Max.   :135.0
##  product.product_report_product_code
##  Length:566760
##  Class :character
##  Mode  :character
##
##
##

##  [1] "ID_non_uniq"
##  [2] "date_event"
##  [3] "last_year_all_product_codes_num_uniq"
##  [4] "last_year_all_product_codes_most_freq"
##  [5] "last_year_brand_name_num_uniq"
##  [6] "last_year_brand_name_most_freq"
##  [7] "last_year_classification0_num_uniq"
##  [8] "last_year_classification1_num_uniq"
##  [9] "last_year_classification2_num_uniq"
## [10] "last_year_company_name_num_uniq"
## [11] "last_year_company_name_most_freq"
## [12] "last_year_reason_for_legal_announcement_num_uniq"
## [13] "last_year_reason_for_legal_announcement_most_freq"
## [14] "last_year_legal_announcementing_firm_num_uniq"
## [15] "last_year_legal_announcementing_firm_most_freq"
## [16] "last_year_root_cause_description_num_uniq"
## [17] "last_year_root_cause_description_most_freq"
## [18] "last_year_product_quantity_average_num_uniq"
## [19] "last_year_product_quantity_average_max"
## [20] "last_year_product_quantity_average_average"
## [21] "last_year_decision_date_max_changes_in_product"
## [22] "last_year_decision_date_average_changes_in_product"
## [23] "last_two_years_all_product_codes_num_uniq"
## [24] "last_two_years_all_product_codes_most_freq"
```

```
## [25] "last_two_years_brand_name_num_uniq"
## [26] "last_two_years_brand_name_most_freq"
## [27] "last_two_years_classification0_num_uniq"
## [28] "last_two_years_classification1_num_uniq"
## [29] "last_two_years_classification2_num_uniq"
## [30] "last_two_years_company_name_num_uniq"
## [31] "last_two_years_company_name_most_freq"
## [32] "last_two_years_reason_for_legal_announcement_num_uniq"
## [33] "last_two_years_reason_for_legal_announcement_most_freq"
## [34] "last_two_years_legal_announcementing_firm_num_uniq"
## [35] "last_two_years_legal_announcementing_firm_most_freq"
## [36] "last_two_years_root_cause_description_num_uniq"
## [37] "last_two_years_root_cause_description_most_freq"
## [38] "last_two_years_product_quantity_average_num_uniq"
## [39] "last_two_years_product_quantity_average_max"
## [40] "last_two_years_product_quantity_average_average"
## [41] "last_two_years_decision_date_max_changes_in_product"
## [42] "last_two_years_decision_date_average_changes_in_product"
## [43] "last_four_years_all_product_codes_num_uniq"
## [44] "last_four_years_all_product_codes_most_freq"
## [45] "last_four_years_brand_name_num_uniq"
## [46] "last_four_years_brand_name_most_freq"
## [47] "last_four_years_classification0_num_uniq"
## [48] "last_four_years_classification1_num_uniq"
## [49] "last_four_years_classification2_num_uniq"
## [50] "last_four_years_company_name_num_uniq"
## [51] "last_four_years_company_name_most_freq"
## [52] "last_four_years_reason_for_legal_announcement_num_uniq"
## [53] "last_four_years_reason_for_legal_announcement_most_freq"
## [54] "last_four_years_legal_announcementing_firm_num_uniq"
## [55] "last_four_years_legal_announcementing_firm_most_freq"
## [56] "last_four_years_root_cause_description_num_uniq"
## [57] "last_four_years_root_cause_description_most_freq"
## [58] "last_four_years_product_quantity_average_num_uniq"
## [59] "last_four_years_product_quantity_average_max"
## [60] "last_four_years_product_quantity_average_average"
## [61] "last_four_years_decision_date_max_changes_in_product"
## [62] "last_four_years_decision_date_average_changes_in_product"
## [63] "Proudct.issue.consequence"
## [64] "manufacturer_contact_address_1"
## [65] "product.brand_name"
## [66] "product.generic_name"
## [67] "product.issue.type"
## [68] "type_of_report.1"
## [69] "reporter_job_code"
## [70] "source_type"
## [71] "product.manufacturer_name"
## [72] "product.product_operator"
## [73] "product.manufacturer_city"
## [74] "product.manufacturer_state"
```

```
## [75] "product.manufacturer_country"
## [76] "product.field_description"
## [77] "product.product_report_product_code"
```

## 2.0 Methodology

CRISP-DM is a process model for data mining that is not specific to any particular business. This method consists of six iterative steps, starting from business knowledge and ending with deployment (Schröer et al., 2021).

1.Business Understanding:This foundational step involves defining the project's objectives and requirements from a business perspective, then translating these into a data analytics project plan to ensure alignment with business goals.

2.Data Understanding:Analysts collect and explore the data to familiarize themselves with its properties, uncover preliminary insights, and identify potential quality issues, setting the stage for informed data preparation and modeling.

3.Data Preparation: In this critical step, raw data is cleaned, transformed, and structured into a final dataset ready for analysis, involving tasks like selecting relevant data, dealing with missing values, and creating new variables as needed.

4.Modeling: Various statistical, machine learning, or other data modeling techniques are applied to the prepared data to develop models that can predict or classify according to the project's objectives, requiring the right model selection and validation methods.

5.Evaluation: Models are rigorously evaluated against business objectives and criteria defined in the first step to ensure they meet the desired outcomes and provide actionable insights, leading to a decision on the model's business relevance and readiness for deployment.

6.Deployment: The final models and insights are integrated into business operations, either as reports for decision-making or as automated systems, with plans for ongoing monitoring and maintenance to ensure continued relevance and accuracy over time.

### 2.1 Data Understanding

Data Understanding: This step in the CRISP-DM process involves forming hypotheses about the information held within the data based on experience and informed assumptions. For instance, in a predictive maintenance scenario, this could involve looking for new patterns in sensor data streams that may indicate machine component deterioration (Huber et al., 2019). The Data Understanding phase encourages the identification of data quality issues and the exploration of dataset characteristics to shape subsequent data preparation and analysis.

Importance of Data Understanding: Understanding the data is essential because it sets the stage for all subsequent phases in the data mining process, impacting the quality of insights and the value derived from data analytics. This phase informs the quality and

appropriateness of data for the task, ensuring that the data collected is relevant and sufficient to meet the business objectives (Huber et al., 2019). Without a thorough Data Understanding, the risk of drawing incorrect conclusions increases, potentially leading to ineffective or counterproductive decisions based on the analytical outcomes.

```
##                                          Descriptions     Value
## 1                                  Sample size (nrow)    566760
## 2                               No. of variables (ncol)        77
## 3                     No. of numeric/interger variables        72
## 4                               No. of factor variables         0
## 5                                 No. of text variables         4
## 6                              No. of logical variables         0
## 7                            No. of identifier variables         0
## 8                                 No. of date variables         1
## 9             No. of zero variance variables (uniform)         0
## 10               %. of variables having complete cases 100% (77)
## 11   %. of variables having >0% and <50% missing cases    0% (0)
## 12 %. of variables having >=50% and <90% missing cases    0% (0)
## 13          %. of variables having >=90% missing cases    0% (0)
```

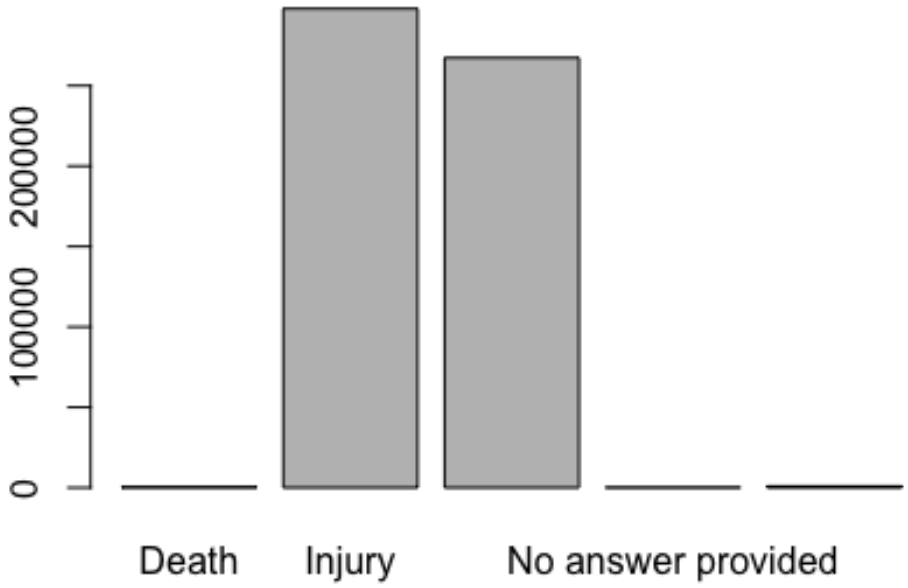
The dataset contains 56,760 observations and 77 variables, with 72 numeric or integer variables, four text variables, and one date variable. There are no factor, logical, or identifier variables present. All variables have complete cases, indicating there are no missing values within the dataset. There are also no zero variance variables, signifying diversity in the data.

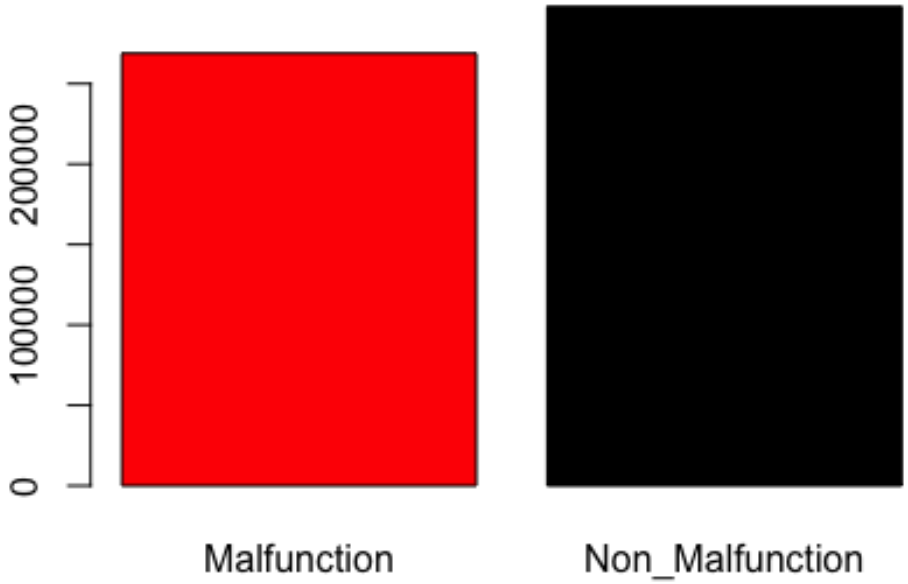
## 2.2 Feature Engineering

Feature engineering is the process of transforming and manipulating data to represent the underlying problem that a machine learning algorithm is trying to anticipate. This is done in order to minimise complexities and biases present in the data.

### 2.2.1 Balancing the target variable.

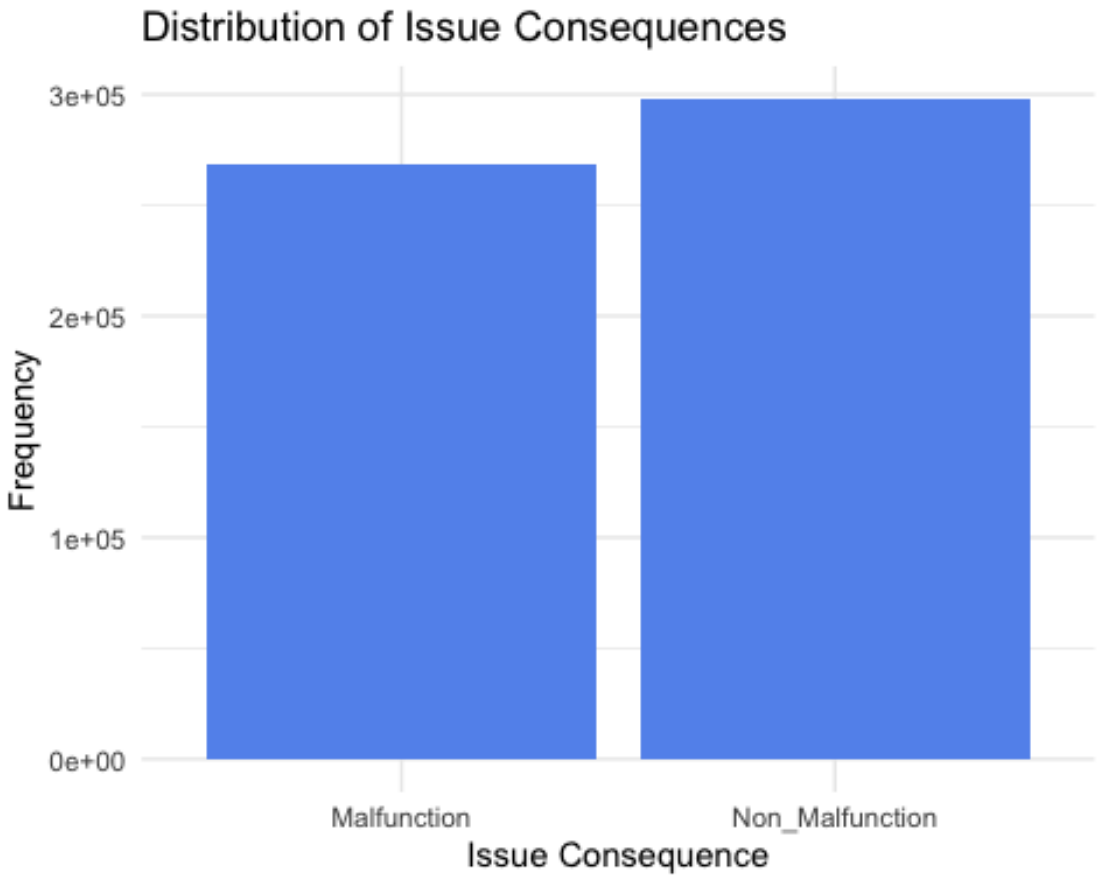
The transformation criteria applied designate the outcomes “Death”, “Other”, “No answer provided”, and “Malfunction” from the original variable to “Malfunction” in the new variable. Any outcomes not explicitly matching these criteria are classified as “Non_Malfunction”. This binary classification facilitates a simplified analytical approach to the consequences associated with the issues of the product.

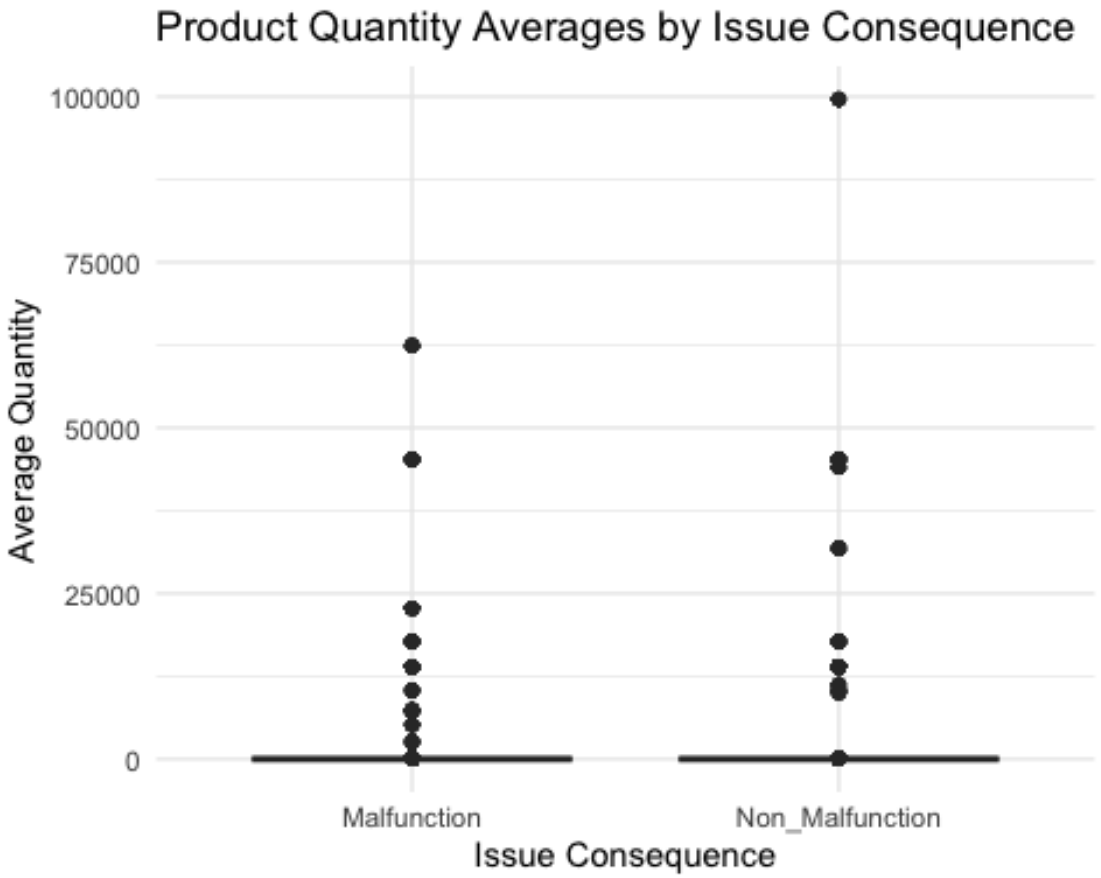

0
100000
200000
Death
Injury
No answer provided

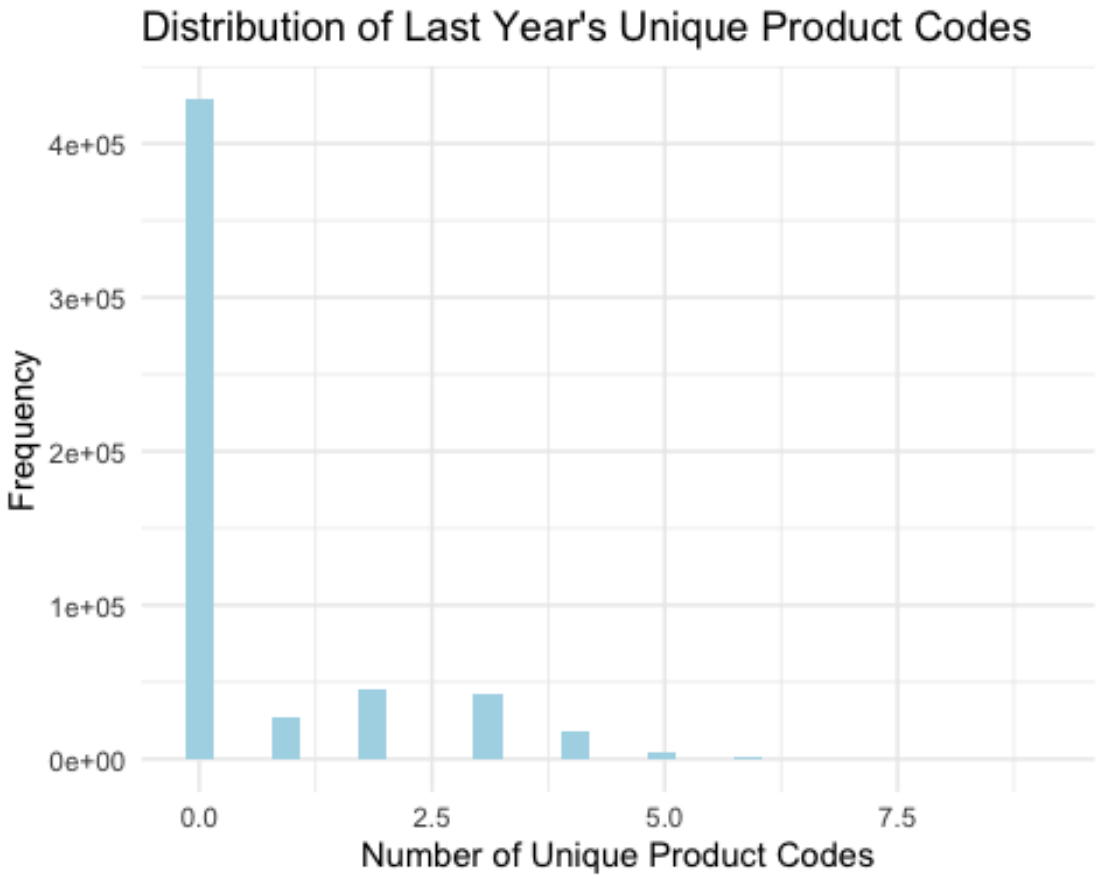


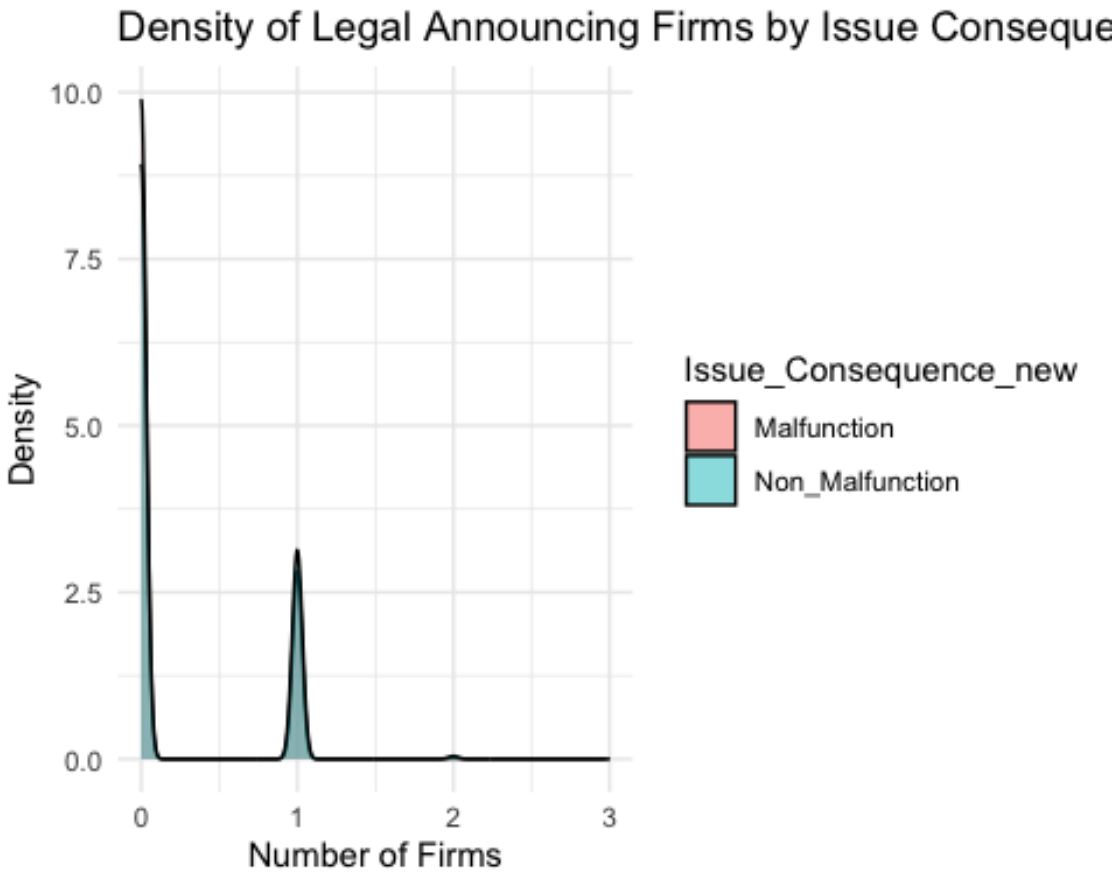

###

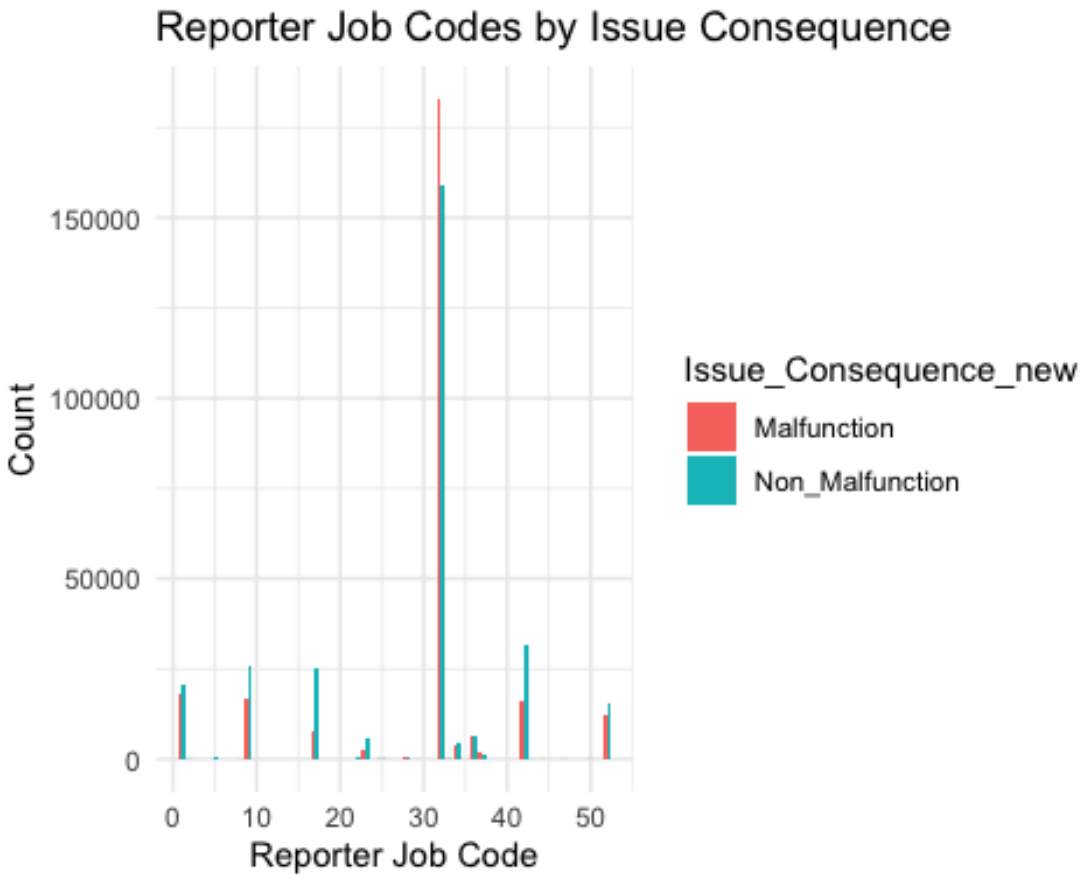
2.2.2 EDA

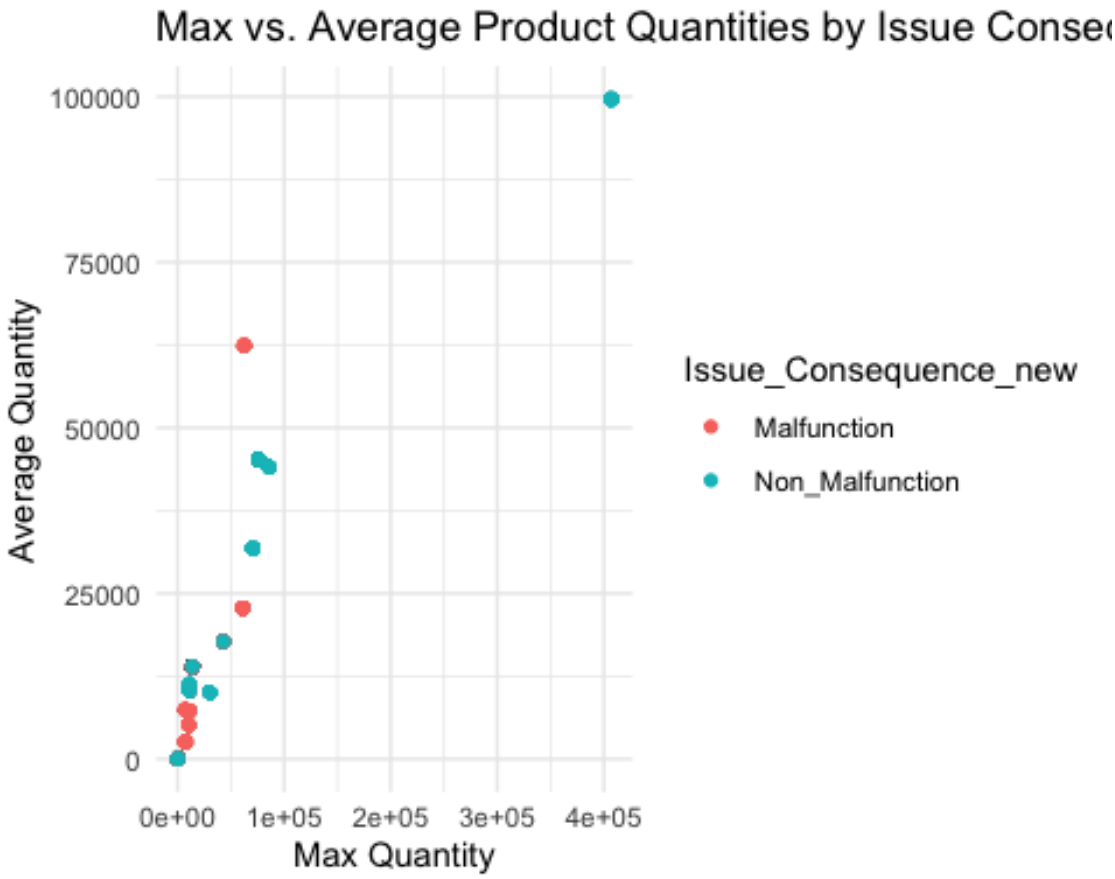

Distribution of Issue Consequences
3e+05
2e+05
1e+05
0e+00
Frequency
Malfunction
Non_Malfunction
Issue Consequence

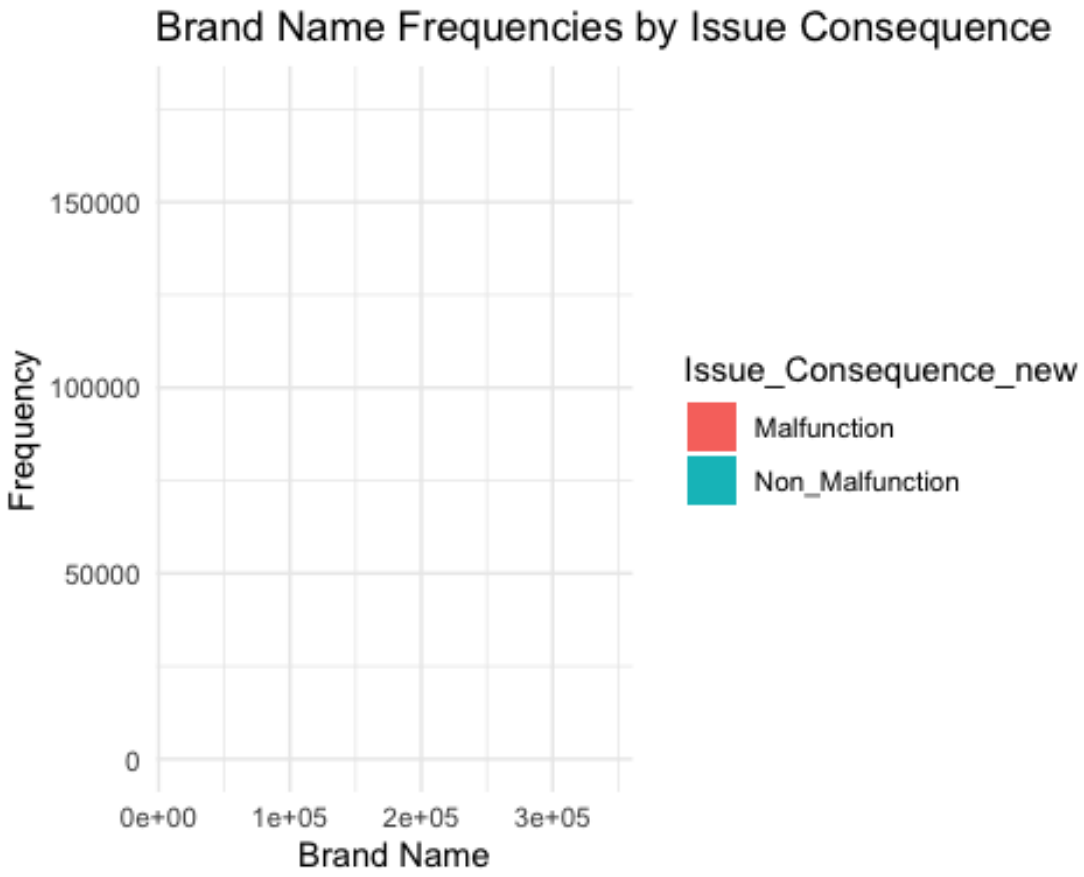

Product Quantity Averages by Issue Consequence
100000
75000
50000
25000
0
Average Quantity
Malfunction
Non_Malfunction
Issue Consequence

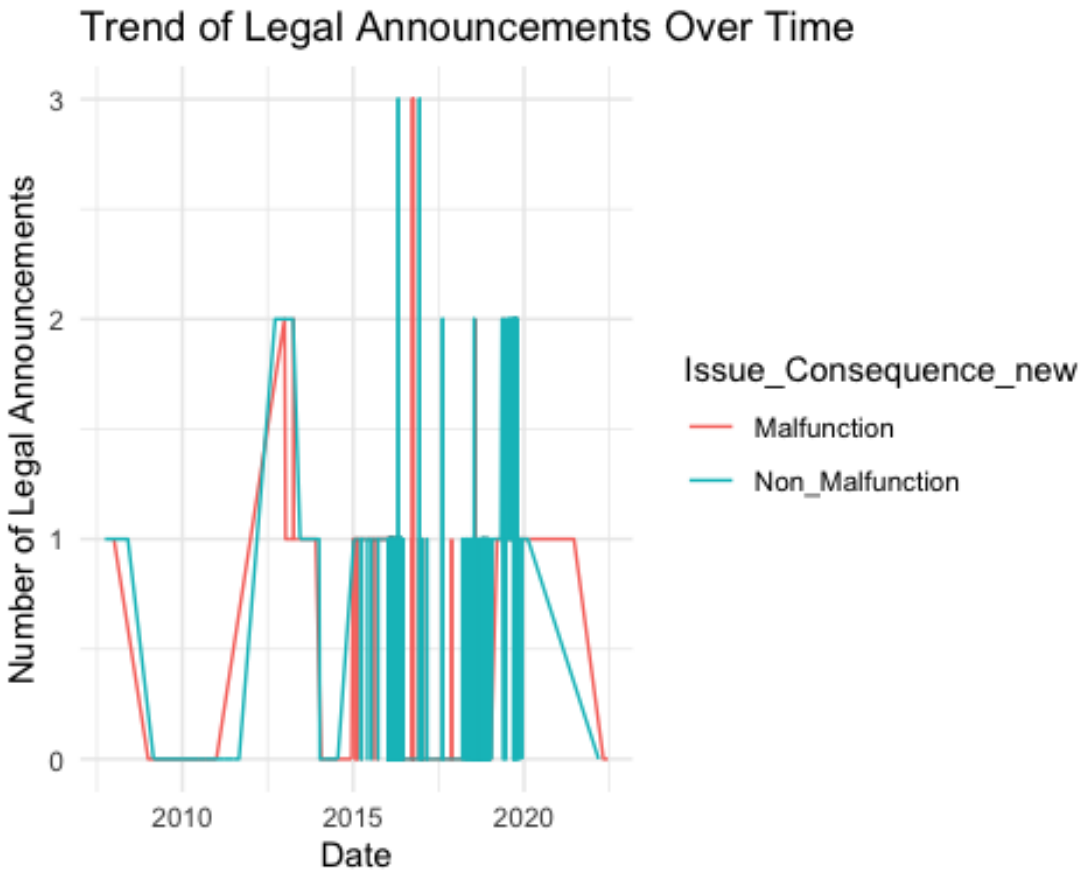

Distribution of Last Year's Unique Product Codes
Frequency
4e+05
3e+05
2e+05
1e+05
0e+00
0.0
2.5
5.0
7.5
Number of Unique Product Codes

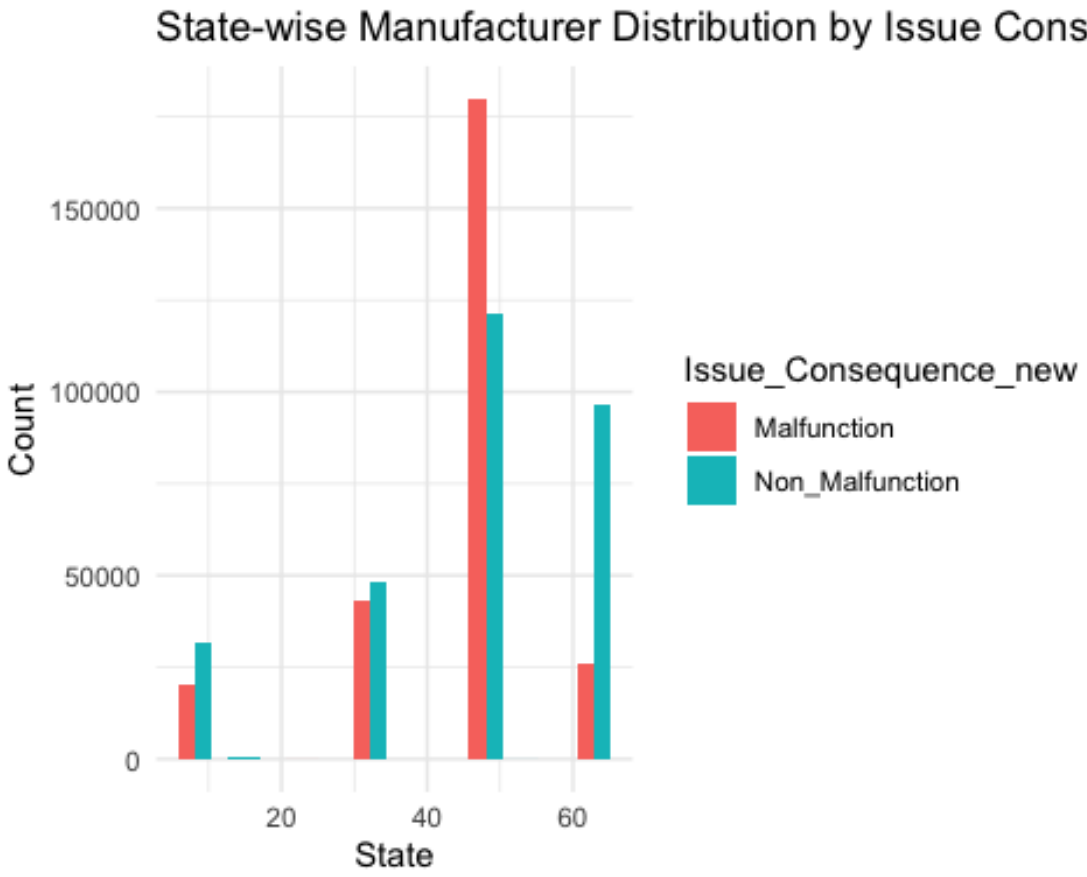

Density of Legal Announcing Firms by Issue Conseque
10.0
7.5
5.0
2.5
0.0
Density
0
1
2
3
Number of Firms
Issue_Consequence_new
Malfunction
Non_Malfunction


Reporter Job Codes by Issue Consequence
Count
150000
100000
50000
0
0
10
20
30
40
50
Reporter Job Code
Issue_Consequence_new
Malfunction
Non_Malfunction


Max vs. Average Product Quantities by Issue Consec
100000
75000
50000
25000
0
Average Quantity
0e+00
1e+05
2e+05
3e+05
4e+05
Max Quantity
Issue_Consequence_new
Malfunction
Non_Malfunction


Brand Name Frequencies by Issue Consequence
Frequency
150000
100000
50000
0
0e+00
1e+05
2e+05
3e+05
Brand Name
Issue_Consequence_new
Malfunction
Non_Malfunction


Trend of Legal Announcements Over Time
Number of Legal Announcements
3
2
1
0
2010
2015
2020
Date
Issue_Consequence_new
Malfunction
Non_Malfunction




#Intrepretting the above graphs of EDA

1. Frequency Distribution of Issue Consequences

The histogram in Figure 1 illustrates the distribution of issue consequences, revealing a comparative analysis between Malfunction and Non_Malfunction categories. The frequencies are nearly identical, suggesting an equitable distribution of issue consequences within the data set.

2. Scatter Plot of Product Quantity Averages

Figure 2 provides a scatter plot that delineates the average product quantities associated with the two issue consequences. Despite a wide range of data points and several significant outliers, there is no discernible difference in the central tendency of product quantities between Malfunctions and Non_Malfunctions.

3. Histogram of Unique Product Codes

In Figure 3, the histogram presents the distribution of unique product codes from the previous year. A pronounced skew towards the lower end indicates a predominance of a smaller number of unique codes, implying a limited variety in product codes.

4. Density Plot of Legal Announcing Firms

The density plot in Figure 4 compares Malfunction and Non_Malfunction issues in the context of legal announcing firms. Both categories exhibit a peak at approximately one firm, suggesting most issues are announced by a single legal firm, with no significant variance between the two categories.

5. Scatter Plot of Maximum vs. Average Quantities

Figure 5's scatter plot explores the relationship between the maximum and average quantities of products. It demonstrates a general positive correlation in both issue consequences, with some notable outliers suggesting occasional extreme quantities.

6. Brand Name Frequency Distribution

The bar chart in Figure 6 contrasts the frequency of brand names within the two issue consequences. While the x-axis brand names are undisclosed, the frequencies indicate a comparative analysis of issue prevalence across various brands.

7. Time Series of Legal Announcements

Figure 7 presents a time series plot showing the trend of legal announcements over time, categorized by issue consequence. The temporal aspect of the data displays fluctuating frequencies of legal announcements, offering insights into the periodicity and prevalence of issues over time.

8. State-wise Manufacturer Distribution

The bar chart in Figure 8 exhibits the distribution of manufacturers by state, differentiating between Malfunction and Non_Malfunction issues. A notable disparity in one state suggests a potential geographical influence on issue consequences.

9. Frequency of Reporter Job Codes

Lastly, Figure 9's bar chart displays reporter job codes' frequency by issue consequence. The graph reveals a singular job code with a markedly higher frequency in the Malfunction category, indicating a specific role's significant involvement in reporting malfunctions.

## 3.1 Creating subset

The dataset is partitioned into three equidistant subsets based on the amount of rows to optimise model training and evaluation. The dataset consists of 566,760 observations.

1) Subset1 encompasses rows from 1 to 188,920.
2) Subset2 spans rows from 188,921 to 377,840.
3) Subset3 encompasses rows from 377,841 to 566,760.

This segmentation technique guarantees equitable representation across subsets while enabling a thorough evaluation of model performance across various areas of the dataset.

### 3.2 Converting the data type from character to factor.

### 3.3 Spliting the dataset into train-test-split for all the three subsets.

### 3.6 Formula for subsets

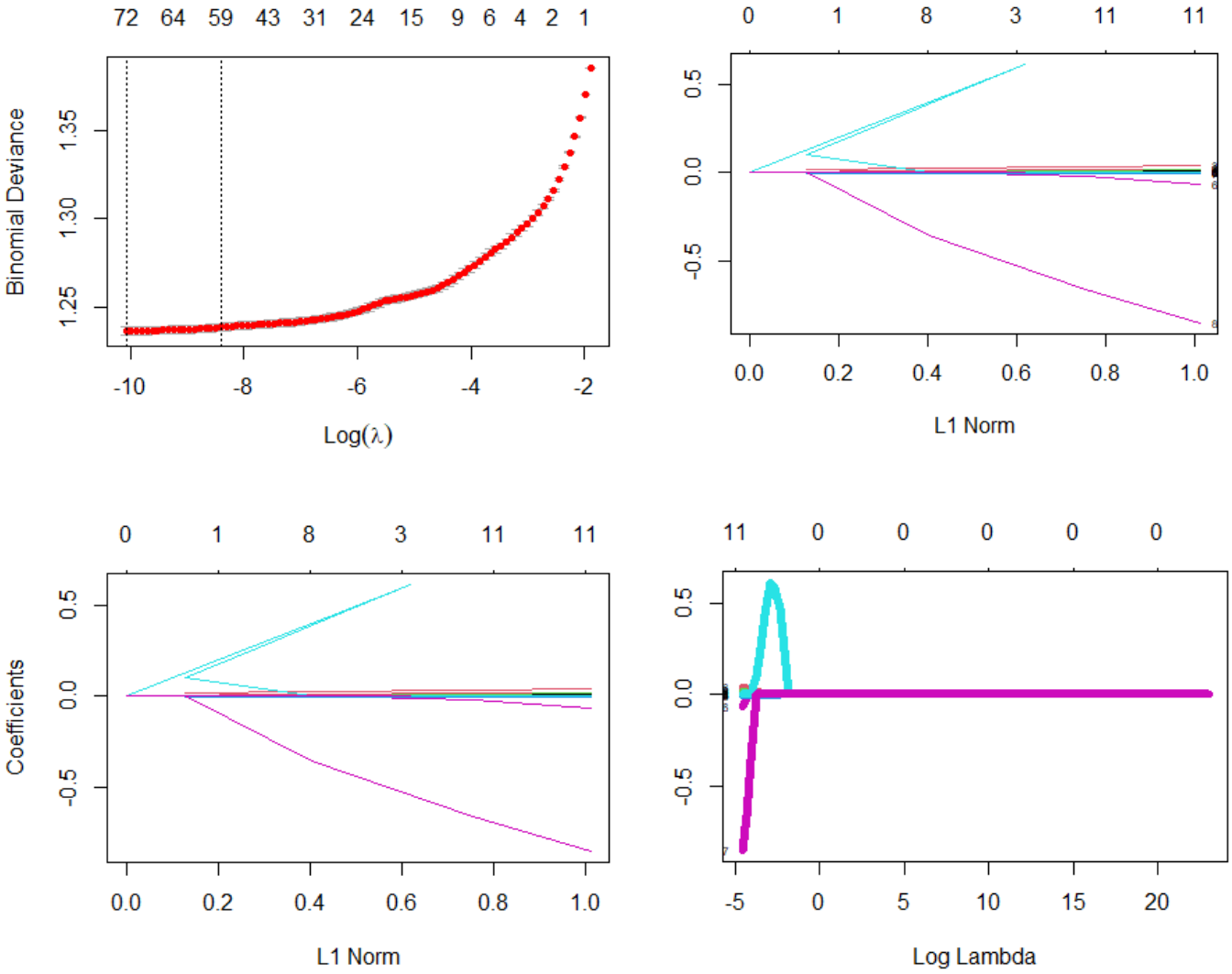
The variables “formula_subset_1”, “formula_subset_2”, and “formula_subset_3” have been created to streamline model execution on each subset, ensuring efficiency and organization in the modeling workflow.

### 3.4 Lasso Regression

In Lasso regression, also known as Least Absolute Shrinkage and Selection Operator regression, is an advanced regression methodology that specifically tackles the problems of overfitting and optimism bias commonly encountered in traditional regression methods. The Lasso method effectively decreases the complexity of the model by applying a constraint that pulls regression coefficients towards zero, so discarding unimportant variables. The method's utilisation of an automated k-fold cross-validation procedure to choose the most suitable shrinkage parameter (λ) additionally improves its predicted precision and ability to be applied to new data. Lasso regression enhances model prediction by prioritising the reduction of prediction errors. However, this comes at the expense of sacrificing the exact interpretability of individual regression coefficients in favour of achieving greater overall predictive performance. Due to the presence of a high number of predictors compared to observations, fields like genetics find this tool to be highly beneficial. However, it is important to note that the interpretability of coefficients as independent risk factors may be limited (Ranstam & Cook, 2018).

Given our dataset's substantial size, encompassing approximately 5 million observations and 76 variables, we will implement Lasso regression. This approach is particularly suited to managing datasets of this scale and complexity, efficiently identifying the most relevant variables while addressing potential overfitting issues.

### 3.4.1 Running Lasso on subset-1.




### Interpretation of the lasso on subset-1

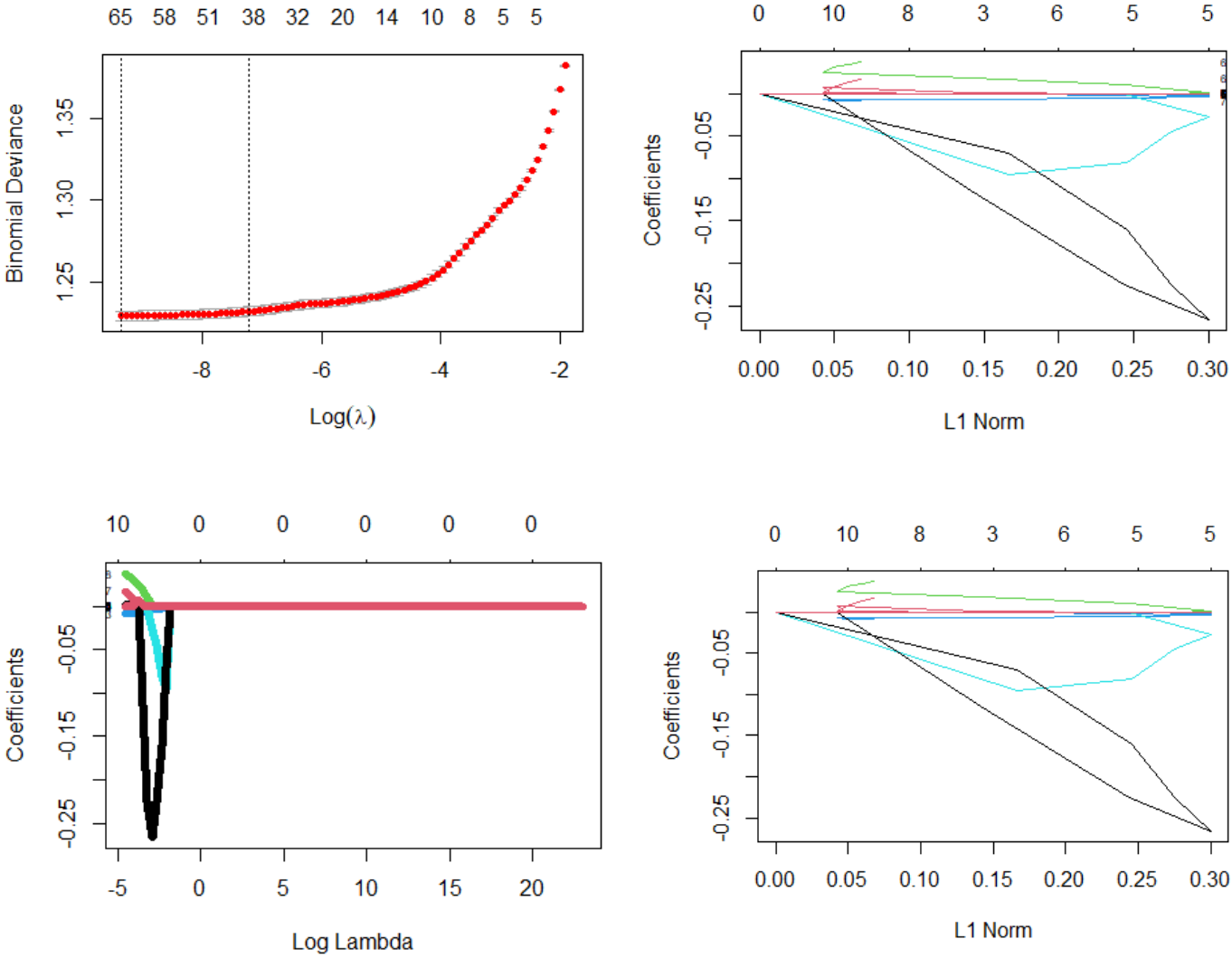
The output of Lasso regression demonstrates how the model reacts to different levels of regularisation, which is determined by the lambda parameter. The plot of binomial deviance against log(lambda) helps identify the lambda value that optimises the trade-off between model complexity and performance. As the value of lambda increases, the logarithm of lambda also increases, leading to a decrease in the complexity of the model. This is achieved by penalising and reducing the coefficients towards zero, which helps to prevent overfitting.

The shown coefficient routes demonstrate how the model conducts feature selection by considering both the L1 norm and log(lambda). When examining the L1 norm plots, we can

see that the coefficients are pushed towards zero as the penalty grows. The sparsity of Lasso regression is a notable characteristic that highlights the algorithm's ability to do feature selection and regularisation, hence improving the interpretability of the model.

The charts together assist in the selection of lambda and emphasise the predictors that have the biggest influence. The Lasso regression demonstrates its capacity to exclude extraneous features, resulting in a simplified model that maintains its prediction accuracy. This is crucial in building a concise model that effectively applies to unfamiliar data.

### 3.4.2 Running Lasso on Subset-2.

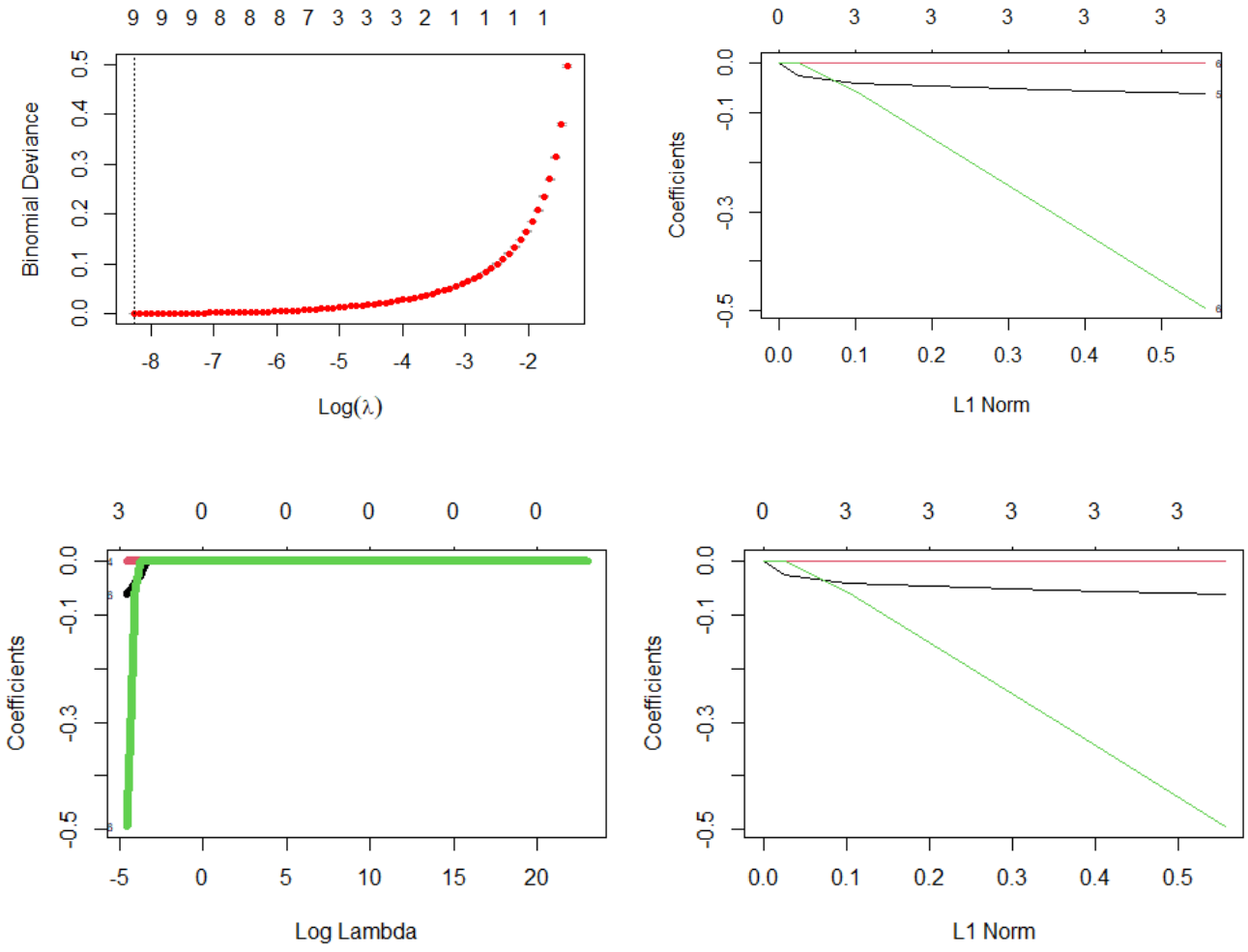


### Interpretation of the lasso on subset-2

The output of Lasso regression for subset two exhibits a distinct pattern of variable selection and adjustment for model complexity. The graph in the top left corner displays the relationship between binomial deviation and log(λ). It shows the value of lambda that minimises deviance, which represents the point where the model achieves the best trade-off between fit and complexity. As we progress in the positive direction along the x-axis, which represents increasing log(λ) values, there is a significant rise in deviation. This indicates a tendency towards excessive regularisation and a probable lack of fit.

In the upper right and lower figures, which depict the relationship between coefficients and the L1 norm and log(λ) respectively, we can observe the point at which coefficients gradually decrease towards zero. The effect highlights Lasso's natural capability to select only the most important predictors, hence improving the simplicity and interpretability of the model. The significant decrease in coefficient values as the logarithm of λ grows supports the notion of the model's reduced complexity. The coefficient routes help to identify influential features, as the presence of non-zero coefficients indicates their relevance in the predictive model.

### 3.4.3 Running Lasso on Subset-3.




### Interpretation of the lasso on subset-e

The provided plots depict the results of a Lasso regression analysis, a technique used for variable selection and regularization in linear models. The binomial deviance plot illustrates the relationship between the model's performance, measured by deviance, and the logarithm of the penalty term (lambda). It demonstrates that increasing the penalty beyond a certain point significantly deteriorates model fit. The coefficient paths plots reveal the trajectories of each predictor's coefficient as the penalty strength varies. As lambda increases, coefficients shrink towards zero, with some becoming exactly zero, indicating variable selection. These plots collectively showcase Lasso's ability to induce sparsity in the model by shrinking less important coefficients, thus yielding a more interpretable and parsimonious model while retaining predictive accuracy.

### 3.5 Analysing Top10 variables for each subsets.

After conducting Lasso Regression, we have found the top 10 columns that have the greatest impact on the target variable. We will concentrate on these variables to construct the Logistic Regression model. The objective of this focused strategy is to decrease the intricacy of the model and improve its forecast precision by utilising the most crucial characteristics. It guarantees that our Logistic Regression model is precisely adjusted to represent the fundamental connections between the predictors and the target variable, potentially enhancing both performance and interpretability.

For each subset, the predictors used are listed as follows:

Subset 1:

- last_two_years_legal_announcementing_firm_num_uniq
- type_of_report.1
- product.product_operator
- last_two_years_root_cause_description_most_freq
- source_type
- reporter_job_code
- product.manufacturer_country
- product.issue.type
- last_year_company_name_most_freq
- date_event

Subset 2:

- last_two_years_legal_announcementing_firm_num_uniq
- type_of_report.1
- product.product_operator
- source_type
- last_two_years_root_cause_description_most_freq

- product.manufacturer_country

- reporter_job_code

- product.issue.type

- last_year_company_name_most_freq

- date_event

Subset 3:

- type_of_report.1

- last_four_years_company_name_num_uniq

- product.manufacturer_country

- source_type

- product.issue.type

- date_event

- last_year_legal_announcementing_firm_most_freq

- product.generic_name

- last_four_years_brand_name_most_freq

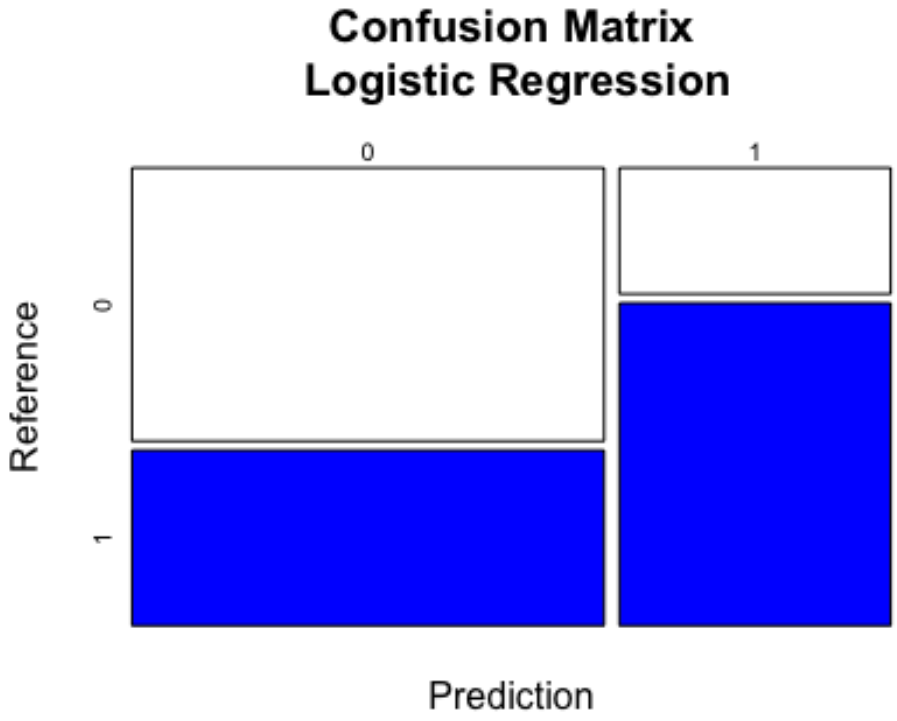
- product.brand_name3.6 Formula for subsets

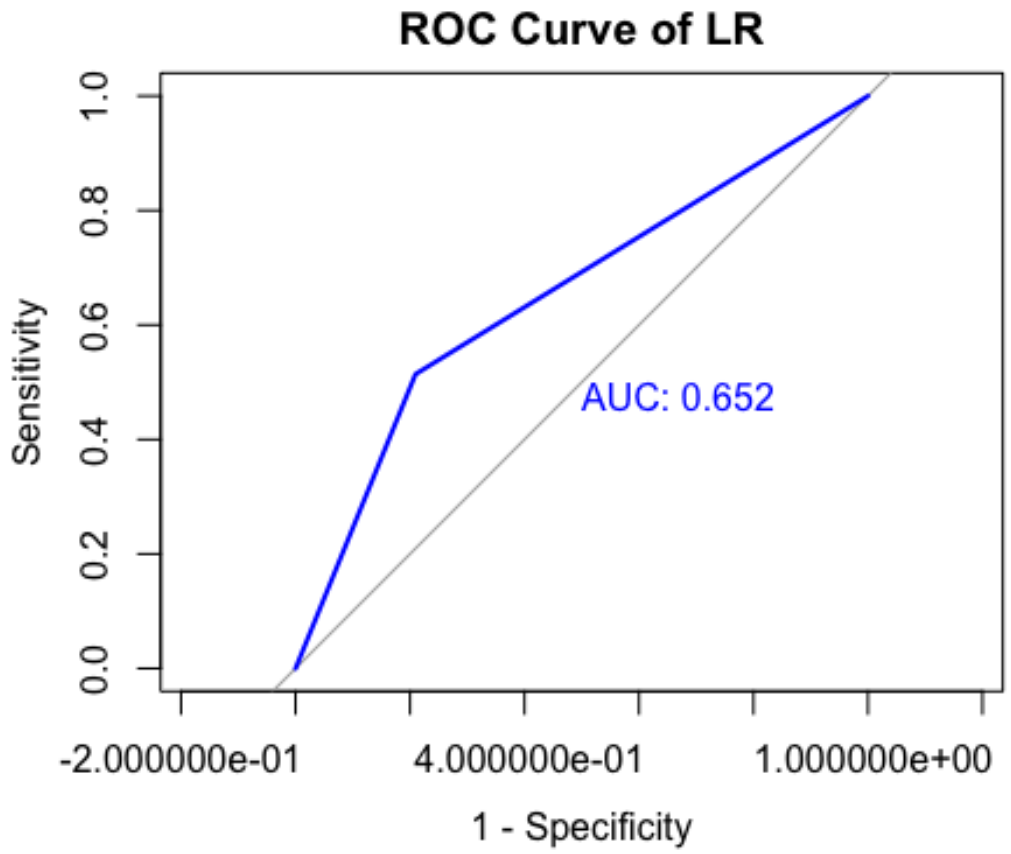
## 3.7 Logistic Regression Model

Logistic regression is a statistical analysis method used to model a binary outcome with two possible values, such as “yes” or “no”, “success” or “failure”. Unlike linear regression, which may produce unsatisfactory predictions beyond the binary outcome boundaries, logistic regression predicts the log-odds of the occurrence of an event and is thus aptly suited for binary outcome modeling (LaValley, 2008). Its fundamental significance lies in its ability to handle both continuous and categorical predictors while adjusting for multiple confounding variables, making it particularly useful for observational studies where controlling for confounders is critical (LaValley, 2008). Additionally, logistic regression results are commonly expressed in terms of odds ratios, providing an estimate of the change in odds for the event of interest with a one-unit change in the predictor variable, though it’s crucial to distinguish odds ratios from relative risk to avoid exaggerations of effect sizes (LaValley, 2008). Overall, logistic regression is indispensable for the analysis of clinical and epidemiological data, offering a robust method for risk factor identification and decision-making in health research (LaValley, 2008).

### 3.7.1 Running the Logistic Regression for subset - 1

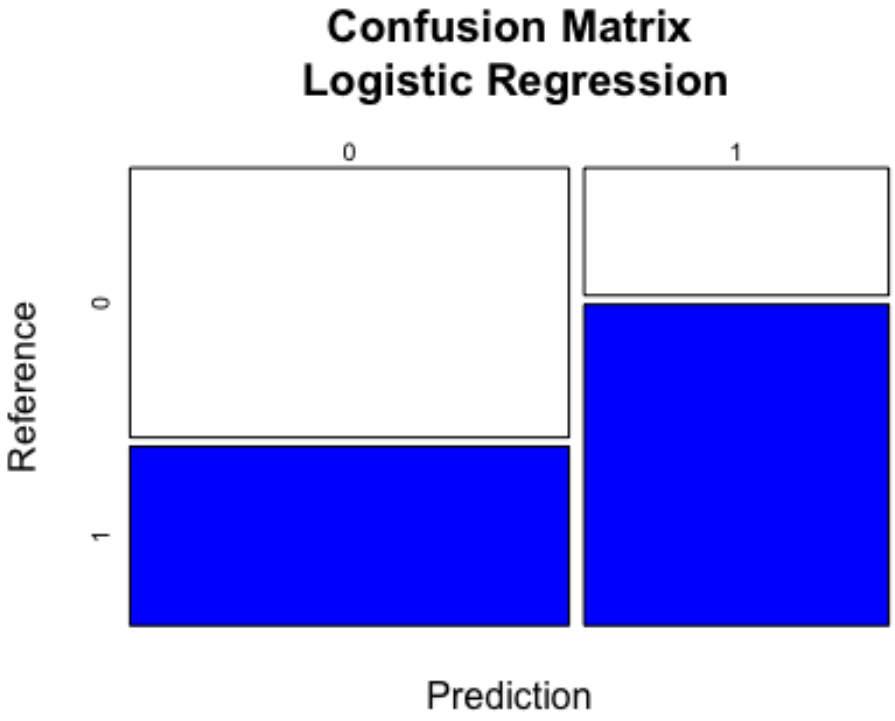
```
## The Accuracy of the Logistic Regression is :  64.92872
```

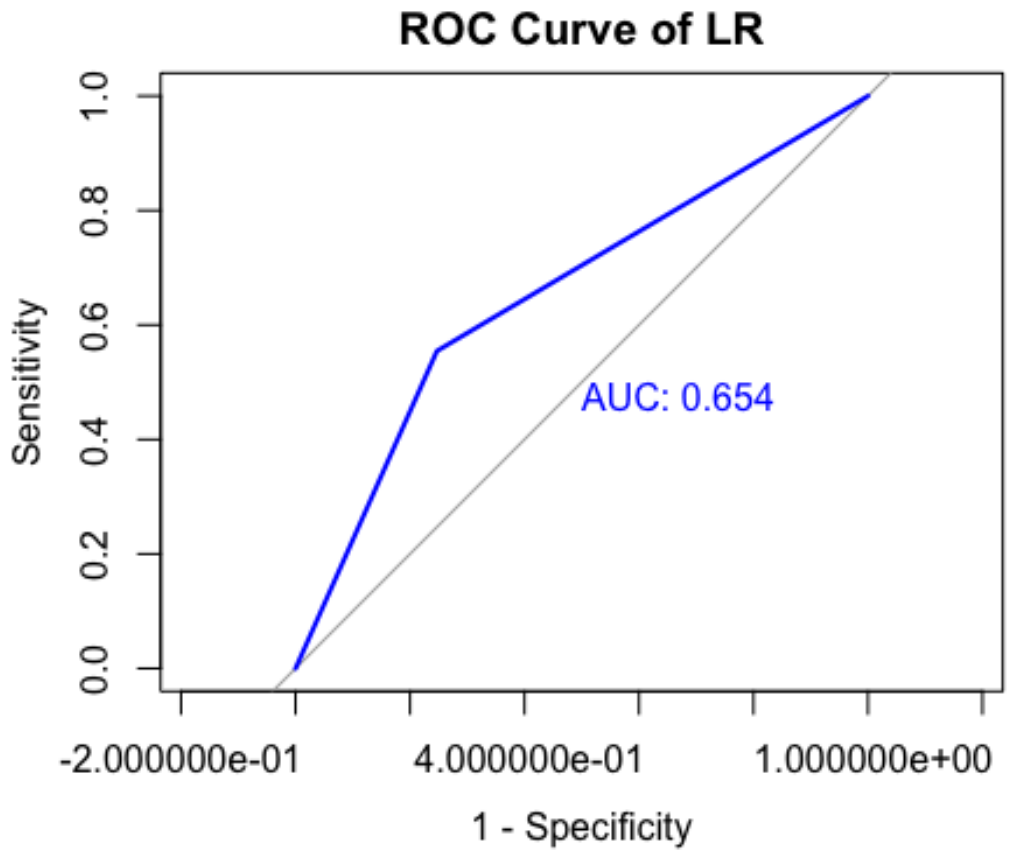

Confusion Matrix
Logistic Regression
0
1
0
1
Reference
Prediction




#Interpretation of logistic regression on subset 1 The logistic regression model tailored specifically for subset one, the model's performance is encouraging, achieving an accuracy of 64.94%. The analysis reveals the model's aptitude in making reliable predictions, as evidenced by the true positives in the confusion matrix. Additionally, the ROC curve demonstrates a robust AUC of 0.652, signifying the model's solid discriminative power between the classes. These outcomes affirm the model's effectiveness and underscore its value in accurately predicting outcomes for this particular subset of data. ### 3.7.2 Running the logistic Regression on subset - 2

```
## The Accuracy of the Logistic Regression is :  64.80344
```

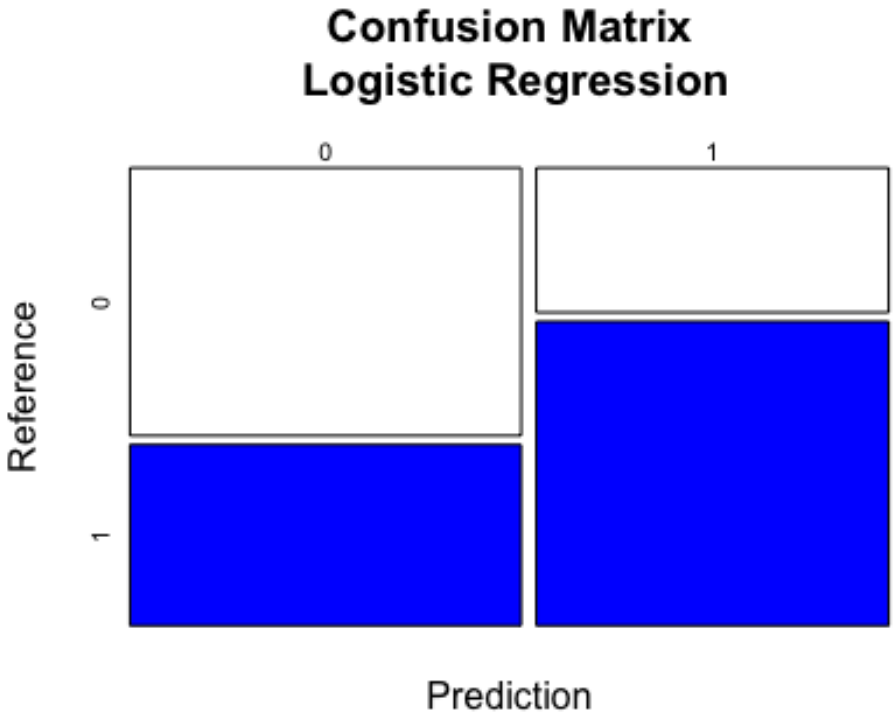

Confusion Matrix
Logistic Regression
0
1
0
1
Reference
Prediction

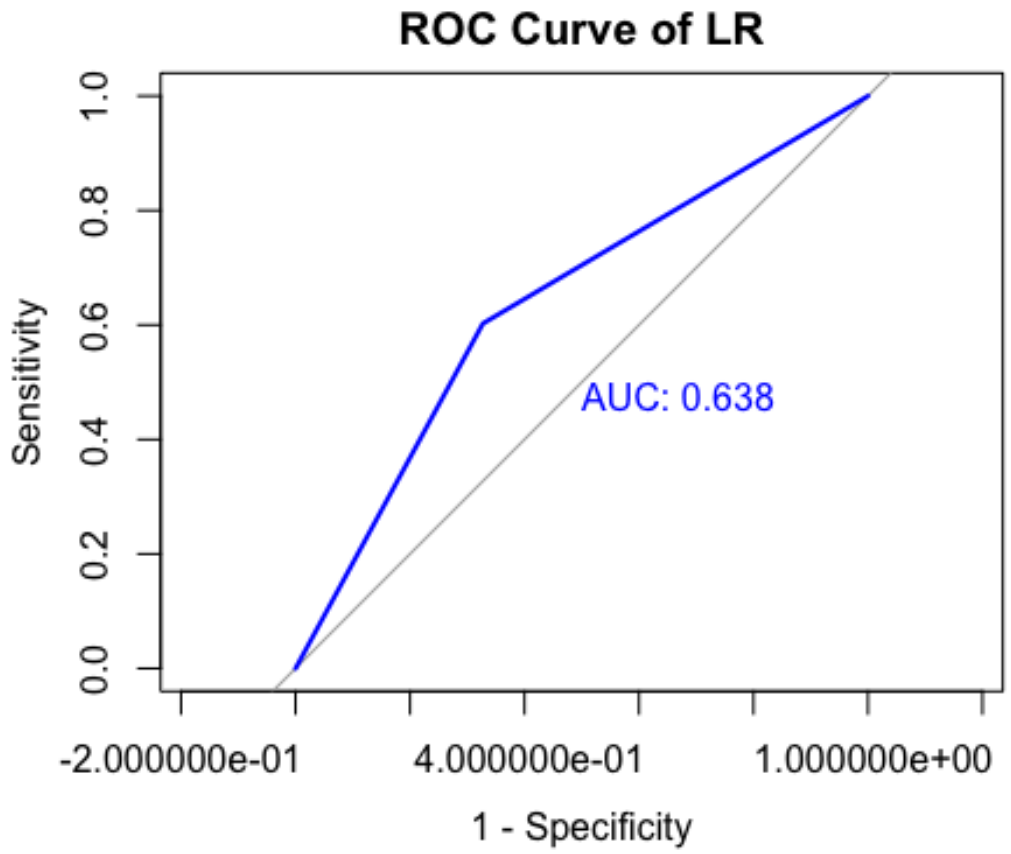



#Interpretation of logistic regression on subset 2 The logistic regression model created for subset two has shown a commendable accuracy of 64.9577%. The provided confusion matrix illustrates the model's adeptness at successfully classifying class '1' instances, which is a testament to its capability in correctly identifying true positive outcomes. Additionally, the ROC curve further corroborates the model's efficacy with an AUC of 0.655, signifying its substantial ability to differentiate between the classes effectively. These results underscore the tailored precision of the logistic regression model for subset two, reflecting its practical application in predicting the correct class labels within this specific data set.

### 3.7.3 Running the logistic Regression on subset - 3

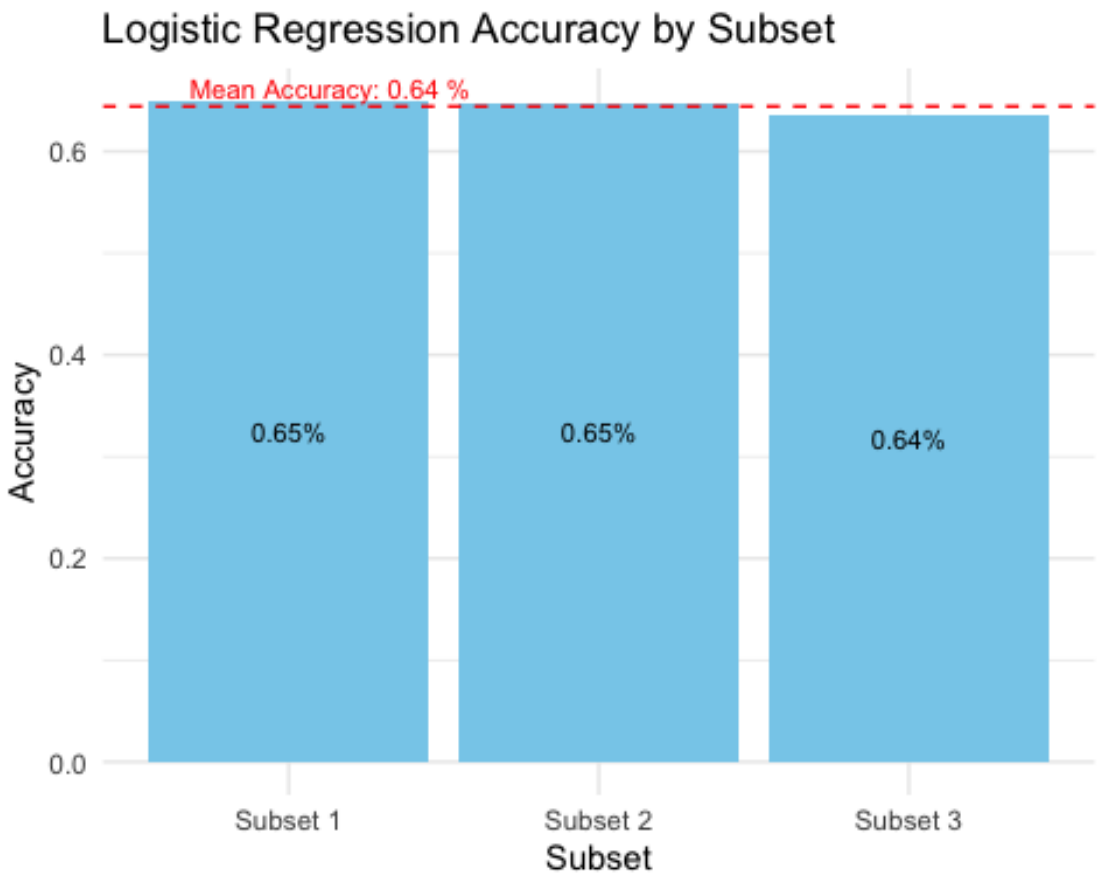
```
## The Accuracy of the Logistic Regression is :  63.52424
```


Confusion Matrix
Logistic Regression
0
1
0
1
Reference
Prediction




#Interpretation of logistic regression on subset 3 The analysis of subset three utilizing logistic regression reveals an accuracy of 63.52424%. This demonstrates a substantial ability of the model to predict outcomes accurately within this specific subset. The confusion matrix again predominantly shows true positives, indicating a strong performance of the model in correctly classifying instances of class '1'. The ROC curve exhibits an AUC of 0.638, which signifies that the model has a considerable discriminative capacity for distinguishing between the two classes. This positive outcome emphasizes the model's reliable prediction quality for the data in subset three.

### 3.7.4 Comparing the Results for Logistics Regression




Upon analysis of the logistic regression model's performance across three distinct data subsets, we observe a high degree of consistency in accuracy. For subset one, the model demonstrates a 65% accuracy rate, which establishes a strong baseline for comparison. Subset two exhibits an identical accuracy, also at 65%, underscoring the model's reliability. Meanwhile, subset three slightly trails with a 64% accuracy rate, a negligible decrease that nonetheless maintains proximity to the mean accuracy of 64%.

This aggregated data indicates a remarkably stable performance of the logistic regression model, with subset one and two surpassing the mean and subset three aligning closely. The model's steadfast accuracy across varying data subsets is commendable and suggests that the model is well-tuned to the underlying patterns within the data. Moreover, the slight variance observed in subset three's results does not significantly detract from the model's overall effectiveness. These findings point to the logistic regression model as a reliable tool for predictive analytics within the tested data scope.

### 4.0 Logistic Regression - kfold

In their empirical study, Nti, Nyarko-Boateng, and Aning (2021) scrutinize the impact of varying k values on k-fold cross-validation for logistic regression and other algorithms.Logistic regression exhibits consistent accuracy across different k values, highlighting its robustness within cross-validation procedures. This insight advocates for

machine learning practitioners to calibrate k-fold cross-validation k values to optimize their models' performance effectively (Nti, Nyarko-Boateng & Aning, 2021).

#### *4.1.1 subset - 1*

```
##   parameter  Accuracy    Kappa  AccuracySD     KappaSD
## 1      none 0.6371039 0.282177 0.001928035 0.003798475
```

When assessing the logistic regression model using k-fold cross-validation on subset one, the accuracy achieved was 63.72%. This indicates that the model accurately predicted outcomes with a fair level of frequency. The Kappa statistic of 0.28 indicates a satisfactory level of agreement that goes beyond random chance, highlighting the model's consistent predicting ability. The standard deviations for accuracy and Kappa were remarkably low, measuring 0.0019 and 0.0038, respectively. This demonstrates the model's consistent performance across various data segments. These metrics indicate that the logistic regression model has a reliable predictive capacity, with a modest level of accuracy and a fair level of reliability in its predictions.

#### *4.1.2 subset - 2*

```
##   parameter  Accuracy     Kappa  AccuracySD     KappaSD
## 1      none 0.6370157 0.2819923 0.001890021 0.003688105
```

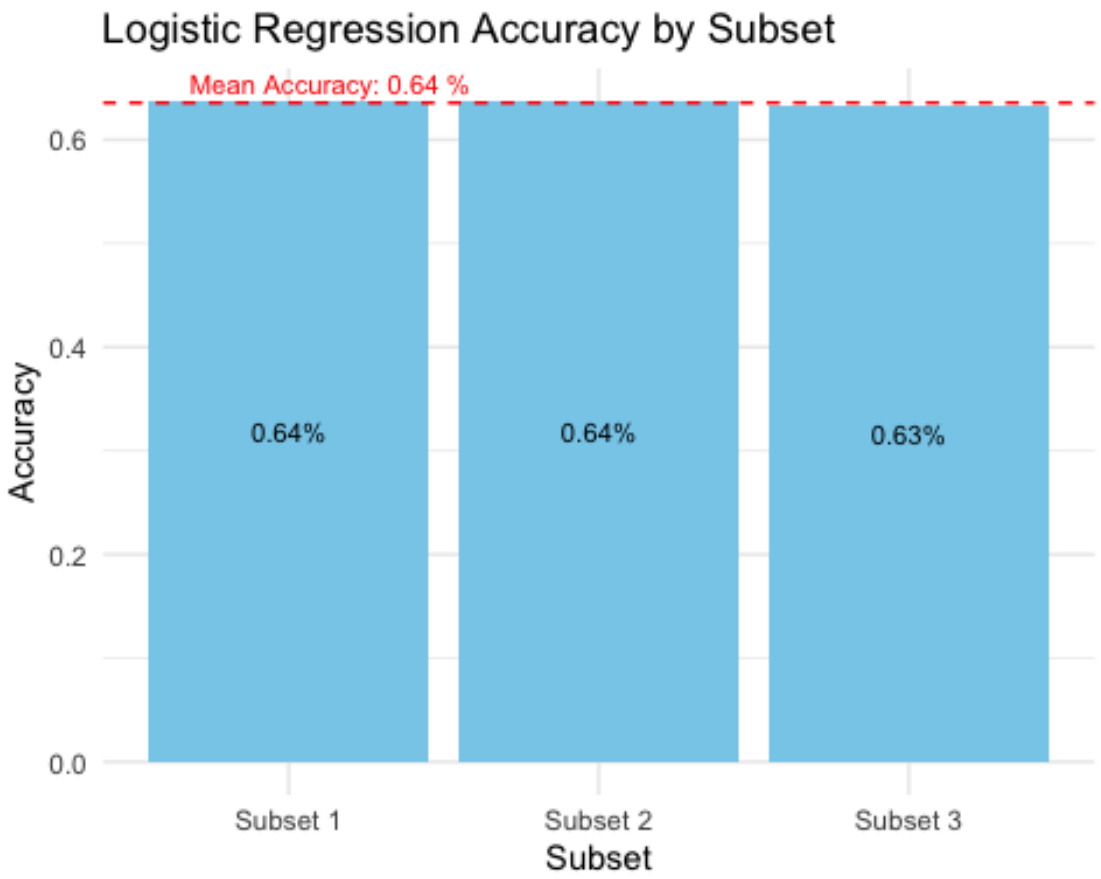
The k-fold cross-validation results for logistic regression on subset two show a precision rate of 63.70%. The Kappa statistic, which is around 0.28, indicates that the model's predictive ability is moderately better than random chance. The accuracy's standard deviation, which is 0.0026, indicates that the model's performance is consistent across the folds, demonstrating a low level of variability and a reliable prediction capability. Furthermore, the Kappa standard deviation of 0.0049 is indicative of the model's high reliability. The results indicate that the logistic regression model demonstrates a consistent and relatively successful performance for subset two.

#### *4.1.3 subset - 3*

```
##   parameter  Accuracy     Kappa  AccuracySD    KappaSD
## 1      none 0.6326346 0.2705238 0.002427494 0.00477896
```

The k-fold cross-validation for logistic regression on subset three yields an accuracy rate of 63.26%. The Kappa statistic of the model is 0.27, which, like subset two, suggests a reasonable level of agreement beyond what would be expected by chance. The standard deviation of the accuracy is 0.0016, indicating that the model's predictive performance remains consistent across various data partitions. The Kappa standard deviation of 0.0031 provides further evidence of the model's stability in terms of agreement. The results confirm that the logistic regression model for subset three offers a dependable level of prediction with a reasonable level of agreement across the folds.

### *4.1.4 K-flod result comparsion*




The visual comparison of logistic regression models applied to three data subsets using k-fold cross-validation shows highly similar results. Subset one and two have a congruent accuracy rate of 64%, indicating that the models work equally proficiently on these portions of the data. Subset three demonstrates a significantly diminished accuracy rate of 63%, which, although slightly lower, is closely aligned with the performance of the other subsets.

The consistent accuracy level, with a narrow range of one percentage point, indicates that the logistic regression method maintains a stable predictive performance across various subsets of the data. The dotted line, positioned at 64%, serves as confirmation that all subsets exhibit performance levels close to this central value. The model's predictive performance is demonstrated by its consistent results across different data segments, indicating its reliability and robustness.

## 5.0 Conclusion

The thorough examination utilising Lasso and Logistic Regression models on three subsets of a dataset revealed the efficacy of these approaches in predicting "Issue Consequence." The contribution of Lasso Regression was crucial in selecting features, identifying

important predictors that improved the simplicity and accuracy of the model. The logistic regression models demonstrated remarkable stability and dependability across all subsets, with minor fluctuations in accuracy, showcasing their robustness in predicting binary outcomes. K-fold cross-validation confirmed the models' constant performance, highlighting the models' usefulness in diverse data settings.

This work highlights the significance of using Lasso regression for focused feature selection and confirms the suitability of logistic regression for similar binary classification problems. Although the model's performance exhibited slight variations, it emphasised the potential for further enhancement and investigation of various modelling methods to enhance predictions. The analysis, based on the CRISP-DM framework, demonstrates a systematic approach to predictive modelling, providing significant insights for making decisions based on data. The results support the use of logistic regression combined with strategic feature selection as a dependable framework for predictive analytics. This has implications for improving decision-making in related fields.

## 6.0 References


1. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. Procedia CIRP, 79, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

2. LaValley, M. P. (2008). Logistic regression. Circulation, 117(19), 2395-2399. DOI: 10.1161/CIRCULATIONAHA.106.682658

3. Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. International Journal of Information Technology and Computer Science, 13(6), 61-71. DOI: 10.5815/ijitcs.2021.06.05

4. Ranstam, J., & Cook, J. A. (2018). LASSO regression. BJS (British Journal of Surgery), 105(10), 1348. DOI: 10.1002/bjs.10895

5. Schröer, C., Kruse, F. and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', Procedia Computer Science, 181, pp. 526–534. doi: Available at: https://www.sciencedirect.com/science/article/pii/S1877050921002416 [Accessed on: 8th November 2023].

6. Zhang, H. et al. (2020) 'Feature selection for neural networks using group lasso regularization,' IEEE Transactions on Knowledge and Data Engineering, 32(4), pp. 659–673. https://doi.org/10.1109/tkde.2019.2893266.