MATH2269 Semester 2 2020 - Final Project

Sam Holt, Phil Steinke, Elleni Toumpas
RMIT

September 28, 2020

Abstract

Insert Abstract

Keywords: JAGS, multiple linear regression analysis, prediction

1 Introduction

2 Analysis

2.1 A descriptive look

2.1.1 Dimensions

Table 1: Dimenions

columns	rows
28	25626

Table 2: Inspecting the first six rows of the data.

id	rainfall_mm	rf_cum_2_day	rf_cum_3_day	rf_cum_4_day	rf_cum_5_day
1	0	0	0	0	0.2
2	0	0	0	0	0.2
3	0	0	0	0	0.2
4	0	0	0	0	0.2
5	0	0	0	0	0.2

Table 3: Inspecting the first six rows of the data.

rf_cum_6_day	rf_cum_7_day	date_day	date_local_time	site
1	25.6	2016-01-01	2016-01-01 00:00:00	Brooklyn
1	25.6	2016-01-01	2016-01-01 01:00:00	Brooklyn
1	25.6	2016-01-01	2016-01-01 02:00:00	Brooklyn
1	25.6	2016-01-01	2016-01-01 03:00:00	Brooklyn
1	25.6	2016-01-01	2016-01-01 04:00:00	Brooklyn

Table 4: Inspecting the first six rows of the data.

temperature	pm10	pm10a	wd	ws	dow	hour	winddire
23.9	30.5	30.5	263	1.6	Friday	0	W
23.5	25.8	25.8	282	1.3	Friday	1	WNW
23.1	22.2	22.2	288	1.1	Friday	2	WNW
22.1	23.2	23.2	340	0.9	Friday	3	NNW
21.4	24.3	24.3	328	0.8	Friday	4	NNW

Table 5: Inspecting the first six rows of the data.

years	yn80	roll_temp	yn50	north	north1	yn60	weekdays	mornings
2016	FALSE	NA	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2016	FALSE	NA	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
2016	FALSE	NA	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
2016	FALSE	NA	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
2016	FALSE	NA	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE

2.1.2 Datatypes

The datatypes for each variable can be found below (after variable datatypes were set on import). Numerical variables that are actually binary values or an identifier were transformed to a factor.

2.1.3 Data Preprocessing

Removing unique identifier As confirmed above, the **id** is a unique identifier, and it is therefore removed from the dataset.

Target variable frequency Before moving on to the outlier, impossible value and missing values check phases, it is important to get an understanding of the frequency of each level in the target level. This will assist when deciding on the best method for fixing possible issues in the following phases. If one of the target levels are only represented by a small number of records then simply removing the records that fail any of the further checks will be ill-advised.

We can see from the frequency of the target variables that there is a severely imbalanced dataset. If the decision had to be made on how to resolve issues in records containing a positive target class, it is advisable not to remove these records.

Table 6: Data types

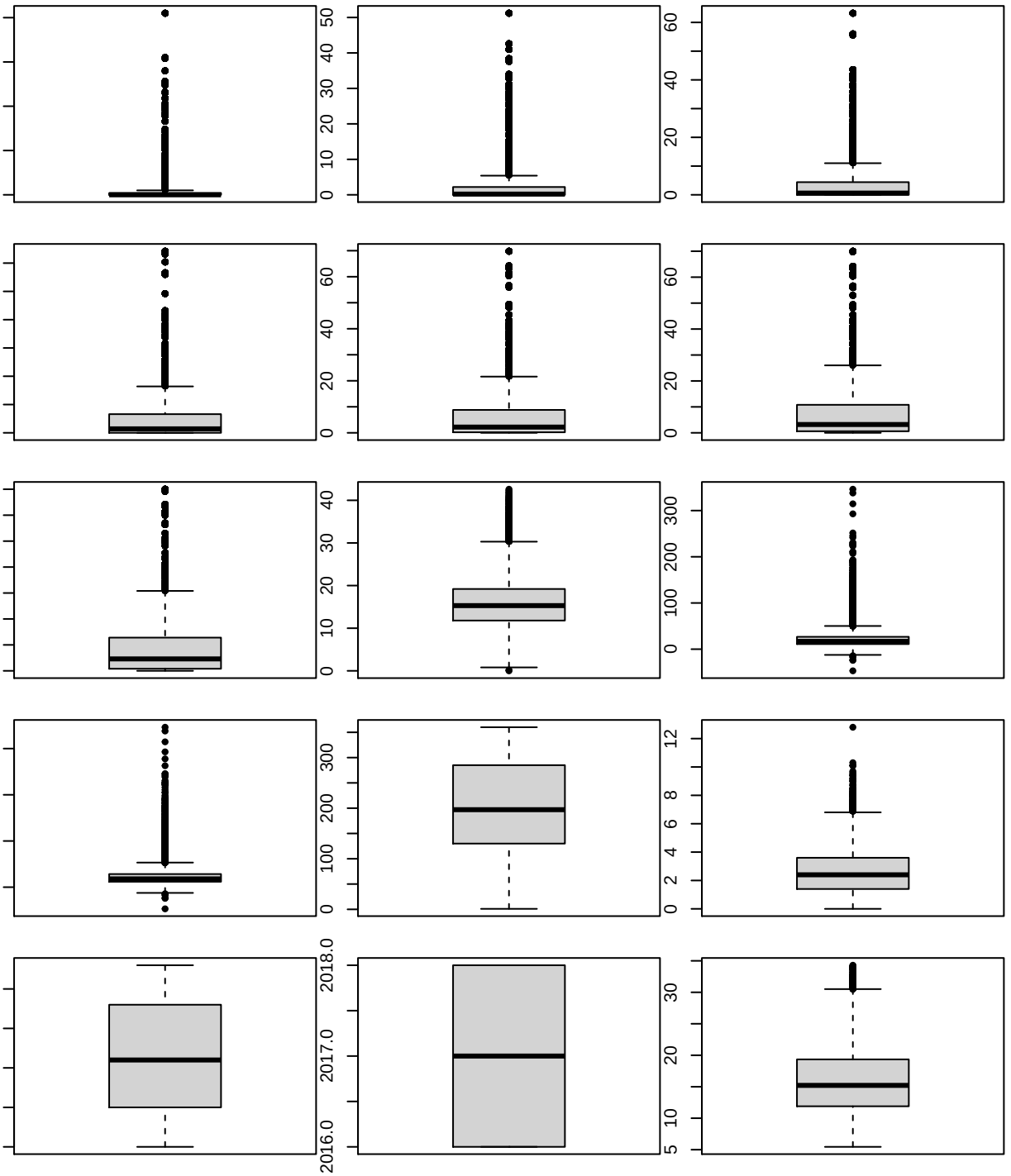
	class
id	numeric
rainfall_mm	numeric
rf_cum_2_day	numeric
rf_cum_3_day	numeric
rf_cum_4_day	numeric
rf_cum_5_day	numeric
rf_cum_6_day	numeric
rf_cum_7_day	numeric
date_day	Date
date_local_time	POSIXct POSIXt
site	character
temperature	numeric
pm10	numeric
pm10a	numeric
wd	numeric
ws	numeric
dow	character
hour	numeric
winddire	character
years	numeric
yn80	logical
roll_temp	numeric
yn50	logical
north	logical
north1	logical
yn60	logical
weekdays	logical
mornings	logical

Table 7: The count of id variable filtered to show only values that are not unique.

id	count

Table 8: The frequency of different levels in the target variable.

yn80	count
FALSE	24988
TRUE	638



Outliers

Impossible values For the numerical values in the dataset an impossible value check is performed.

Table 9: The number of rows with impossible values.

column	nrows
rainfall_mm	0
rf_cum_2_day	0
rf_cum_3_day	0
rf_cum_4_day	0
rf_cum_5_day	0
rf_cum_6_day	0
rf_cum_7_day	0
rf_cum_5_day	0
date_day	0
temperature	0
pm10	118
pm10a	118
wd	0
ws	0
hour	0
years	0
roll_temp	0

2.1.4 Missing values

Checking the missing values we can see that there are 23 rolling temperate missing records.

Inspecing the missing records, we can see that all missing values occur in the end part of the dataset with no records including a positive target class.

```
data_cleaned %>%
  filter(is.na(roll_temp)) %>%
  select(date_day, date_local_time, roll_temp, yn80) %>%
  format_table(p_caption = "Missing roll temperate values")
```

It is therefore sufficient to simply remove these records from the dataset.

```
#### INSERT SAM'S CODE TO REMOVE MISSING VALUES
```

```
#### Do we need to create any new features from the data set?
```

Feature engineering

Categorical Features To check whether there are errors (including typos or unexpected values) in the categorical features each variable is arranged in order and then inspected by the researchers. In the list of possible values printed below there seems to be no incorrect values.

```
for (col in colnames(data_cleaned)){

  if(class(data_cleaned[[col]])[1] %in% c('factor', 'ordered', 'character')){

    paste0("Unique values for ",col) %>% cat()
    cat("\n")
```

Table 10: Count of missing values by variable

	Number missing values
rainfall_mm	0
rf_cum_2_day	0
rf_cum_3_day	0
rf_cum_4_day	0
rf_cum_5_day	0
rf_cum_6_day	0
rf_cum_7_day	0
date_day	0
date_local_time	0
site	0
temperature	0
pm10	0
pm10a	0
wd	0
ws	0
dow	0
hour	0
winddire	0
years	0
yn80	0
roll_temp	23
yn50	0
north	0
north1	0
yn60	0
weekdays	0
mornings	0

Table 11: Missing roll temperate values

date_day	date_local_time	roll_temp	yn80
2016-01-01	2016-01-01 00:00:00	NA	FALSE
2016-01-01	2016-01-01 01:00:00	NA	FALSE
2016-01-01	2016-01-01 02:00:00	NA	FALSE
2016-01-01	2016-01-01 03:00:00	NA	FALSE
2016-01-01	2016-01-01 04:00:00	NA	FALSE
2016-01-01	2016-01-01 05:00:00	NA	FALSE
2016-01-01	2016-01-01 06:00:00	NA	FALSE
2016-01-01	2016-01-01 07:00:00	NA	FALSE
2016-01-01	2016-01-01 08:00:00	NA	FALSE
2016-01-01	2016-01-01 09:00:00	NA	FALSE
2016-01-01	2016-01-01 10:00:00	NA	FALSE
2018-12-31	2018-12-31 12:00:00	NA	FALSE
2018-12-31	2018-12-31 13:00:00	NA	FALSE
2018-12-31	2018-12-31 14:00:00	NA	FALSE
2018-12-31	2018-12-31 15:00:00	NA	FALSE
2018-12-31	2018-12-31 16:00:00	NA	FALSE
2018-12-31	2018-12-31 17:00:00	NA	FALSE
2018-12-31	2018-12-31 18:00:00	NA	FALSE
2018-12-31	2018-12-31 19:00:00	NA	FALSE
2018-12-31	2018-12-31 20:00:00	NA	FALSE
2018-12-31	2018-12-31 21:00:00	NA	FALSE
2018-12-31	2018-12-31 22:00:00	NA	FALSE
2018-12-31	2018-12-31 23:00:00	NA	FALSE

```
    data_cleaned %>%
      arrange(get(col)) %>%
      select(col) %>%
      unique() %>%
      pull() %>%
      as.character() %>%
      paste0(collapse = ", ") %>%
      stringr::str_trunc(width = 800, side = "right", ellipsis = "... (truncated)") %>%
      cat()

    cat("\n")
    cat("\n")

  }
}
```

```
## Unique values for site
## Brooklyn
##
## Unique values for dow
## Friday, Monday, Saturday, Sunday, Thursday, Tuesday, Wednesday
##
## Unique values for winddire
## E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW, W, WNW, WSW
```

```
#### Do we need to encode categorical variables?
```

Any categorical descriptive feature encoded

Summary statistics A quick look at the custom summary statistics can be found below. For factors we can see the most common level, with the count of appearances for that mode level. For the Date variables we can we can see the min, max and mode levels. For the numeric and integer variables we can see the mean, median, standard deviation, minimum and maximum values.

```
results_df <- exploratory_summarize(data_cleaned, col == 'id')
results_df %>%  format_table(p_caption = "Exploratory dataset")
```

```
plots <- list()

for(col in colnames(data_cleaned)){
  plots[[col]] <- univariate_distribution_plot(data_cleaned[[col]], col)
}
```

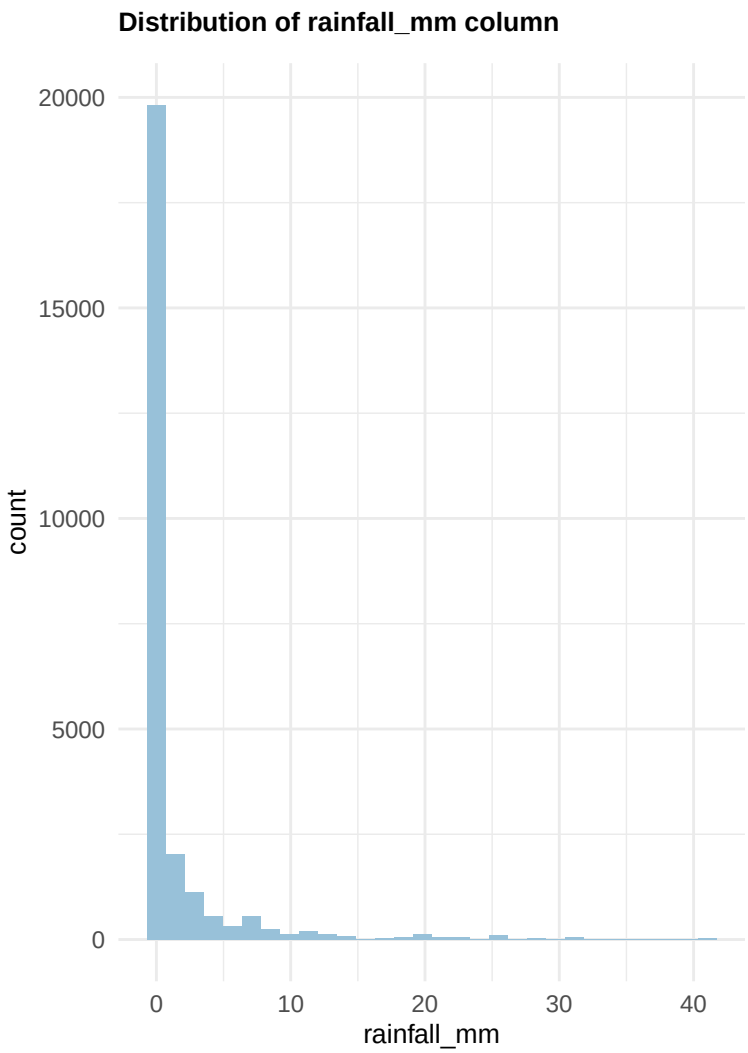
Univariate distribution

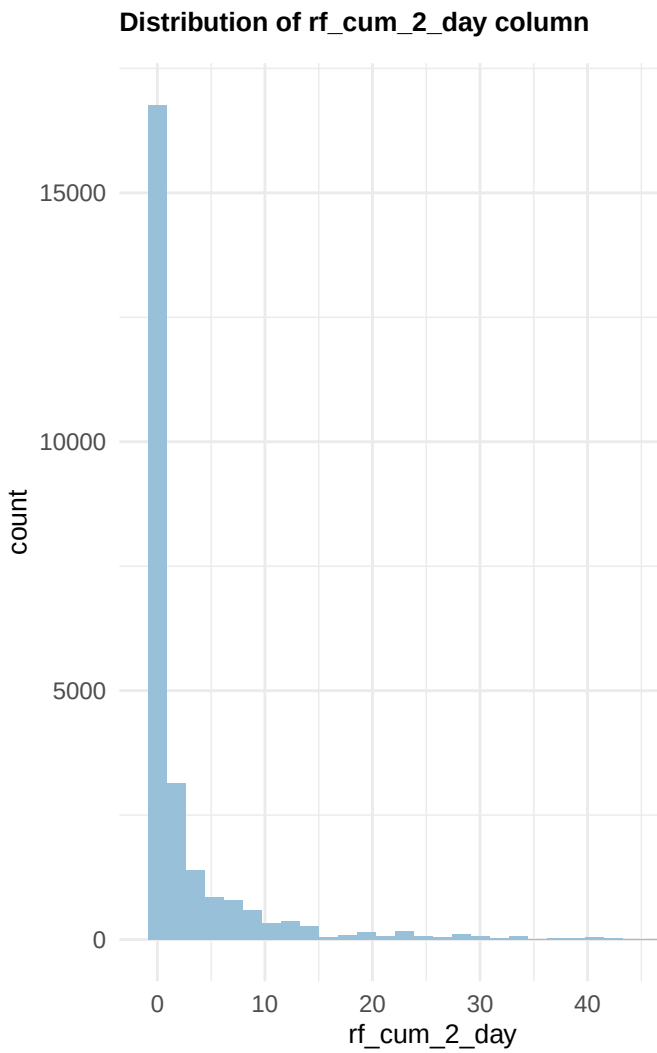
```
## [1] "date_local_time couldn't be plotted."
## [1] "site couldn't be plotted."
## [1] "dow couldn't be plotted."
## [1] "winddire couldn't be plotted."
## [1] "yn80 couldn't be plotted."
## [1] "yn50 couldn't be plotted."
## [1] "north couldn't be plotted."
## [1] "north1 couldn't be plotted."
## [1] "yn60 couldn't be plotted."
## [1] "weekdays couldn't be plotted."
## [1] "mornings couldn't be plotted."
```

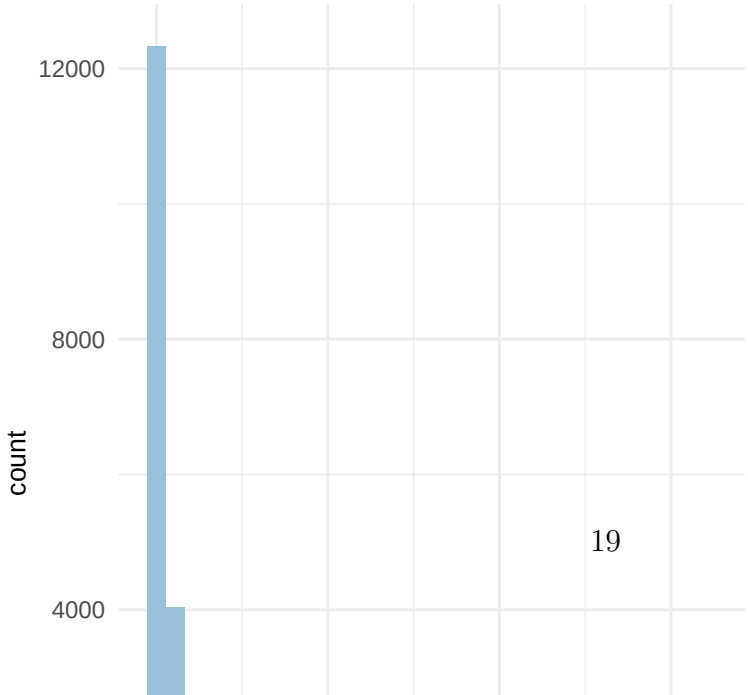
Table 12: Exploratory dataset

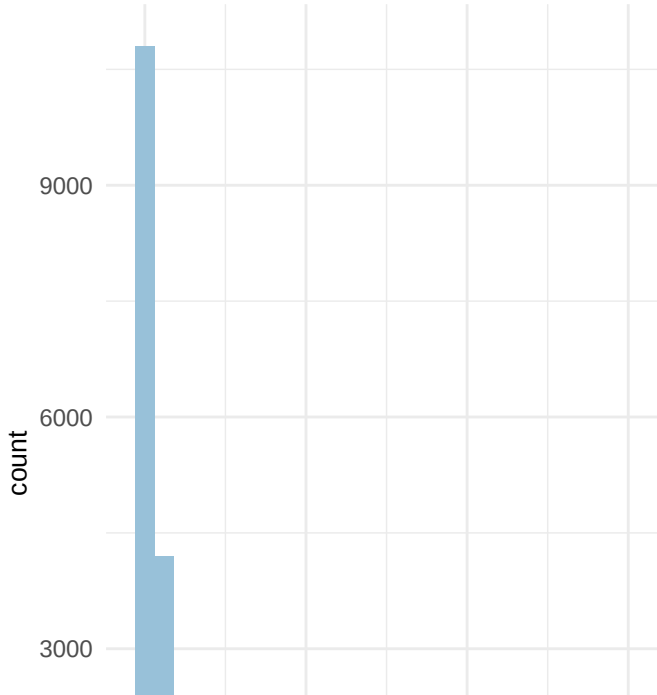
name	type	missing_values	mean	median	sd	mode	min	max	nlevs
rainfall_mm	numeric	0	1.327	0	3.87	NA	0	41	NA
rf_cum_2_day	numeric	0	2.652	0.2	5.995	NA	0	51.2	NA
rf_cum_3_day	numeric	0	3.948	0.6	7.609	NA	0	63.2	NA
rf_cum_4_day	numeric	0	5.241	1.4	8.925	NA	0	64.2	NA
rf_cum_5_day	numeric	0	6.543	2.2	10.046	NA	0	69.8	NA
rf_cum_6_day	numeric	0	7.857	3.2	11.051	NA	0	70	NA
rf_cum_7_day	numeric	0	9.198	4.6	11.97	NA	0	70	NA
date_day	Date	0	NA	NA	NA	2018-10-16	2016-01-01	2018-12-31	NA
date_local_1	POSIXct	0	NA	NA	NA	NA	NA	NA	NA
date_local_2	POSIXt	0	NA	NA	NA	NA	NA	NA	NA
site	character	0	NA	NA	NA	NA	NA	NA	NA
temperature	numeric	0	15.887	15.3	5.783	NA	0	42.6	NA
pm10	numeric	0	22.326	16.9	19.705	NA	-46.9000015259	346.2	NA
pm10a	numeric	0	23.744	17.966	20.888	NA	-46.9000015259	346.2	NA
wd	numeric	0	193.817	197	110.853	NA	1	360	NA
ws	numeric	0	2.632	2.4	1.499	NA	0	12.8000001907	NA
dow	character	0	NA	NA	NA	NA	NA	NA	NA
hour	numeric	0	11.491	11	6.933	NA	0	23	NA
winddire	character	0	NA	NA	NA	NA	NA	NA	NA
years	numeric	0	2017	2017	0.818	NA	2016	2018	NA
yn80	logical	0	NA	NA	NA	NA	NA	NA	NA
roll_temp	numeric	23	15.882	15.213	4.926	NA	5.45000004768372	30.2958333333	NA
yn50	logical	0	NA	NA	NA	NA	NA	NA	NA
north	logical	0	NA	NA	NA	NA	NA	NA	NA
north1	logical	0	NA	NA	NA	NA	NA	NA	NA
yn60	logical	0	NA	NA	NA	NA	NA	NA	NA
weekdays	logical	0	NA	NA	NA	NA	NA	NA	NA
mornings	logical	0	NA	NA	NA	NA	NA	NA	NA

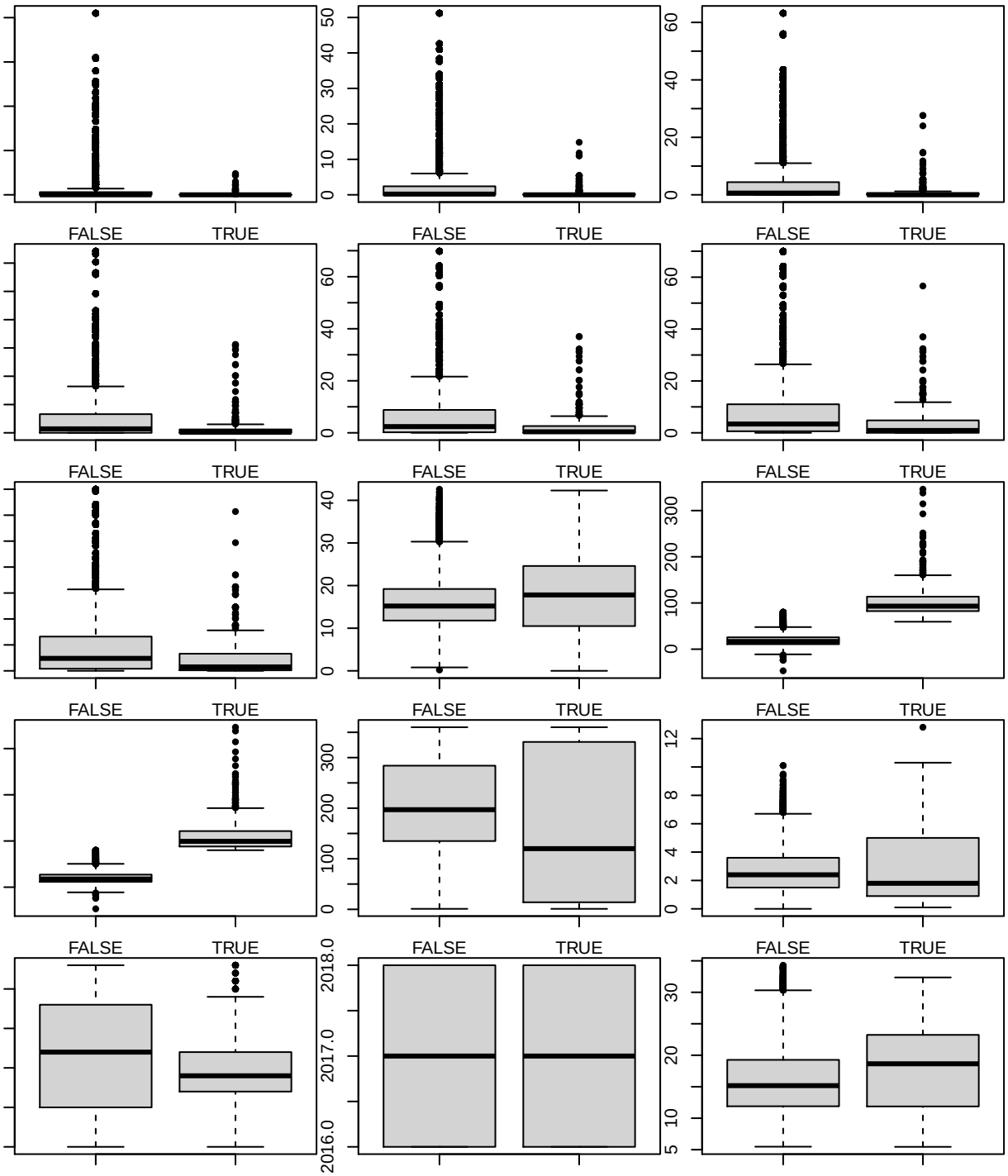
```
grid.arrange(grobs = plots, ncol = 3)
```


Distribution of rainfall_mm column
20000
15000
10000
5000
0
count
0
10
20
30
40
rainfall_mm


Distribution of rf_cum_2_day column
15000
10000
5000
0
count
0
10
20
30
40
rf_cum_2_day


Distribution of rf_cum_4_day column
12000
8000
4000
count


Distribution of rf_cum_5_day column
9000
6000
3000
count


FALSE
TRUE
0
10
20
30
40
50
60
100
200
300
2016.0
2017.0
2018.0
2
4
6
8
12
5

2.1.5 Likelihood

2.1.6 Dependant vs independant bivariate visualisation

2.1.7 Correlation matrix of predictors

2.2 Subsampling

2.2.1 Subsample size trials

2.2.2 Subsample selection trials

2.3 Mathematical model

2.4 Prior specification

2.5 Model

2.6 Experiments to improve model efficiency

2.6.1 Isolated experiments on adapt steps

2.6.2 Isolated experiments on burn in steps

2.6.3 Isolated experiments on thinning steps

2.6.4 Isolated experiments on number of saved steps

2.6.5 Isolated experiments with initial values

2.7 Model fine-tuning

2.8 Prior sensitivity analysis

2.9 Posterior Inferences

2.10 Results

2.11 Prediction

3 Conclusion

4 Reference