# STAT 218 - Analytics Project I

*Inigo Benavides and Rommel Bartolome*

*February 22, 2019*

**Abstract**

We create two models: (1) a linear regression model to predict the water rate and (2) a logistic regression model to predict whether a given district has wastage rating of less than or equal to 25%. We engineered a `desirability score` to find the best model that balances the tradeoff between interpretability and predictive power. For both, we ensured first that all model assumptions are valid. We found that the simple linear model `wd_rate ~ nrwpcent + sprw` is the most desirable model with 0.2806605 test MSE. For the logistic regression model, `nrwpcent_class ~ REGION.III + cities + REGION.VII + sprw + REGION.XI + surw` has been found to be the most desirable model. However, it appears to only have 46% accuracy.

## Introduction

Here in this project, we build two models:

- A linear regression model to predict the `wd_rate` (water rate in Pesos) for differt water districts in the Philippines given a number of features.
- A logistic regression model to predict whether a given water district has `nrwpcent` less than or equal to 25% or greater.

Below we employ various techniques to search for a good fitting model, subject to the following:

- Regression assumptions should be validated with tests at 10% significance
- VIF <= 10
- Cannot remove more than 2% of sample size
- Should have best performance on test set, using RMSE <= 15% and error rate less than 10% for categorical models.

### Loading The Data

First we load all necessary libraries for this project. We then load our sample data set and remove the indexed `X1` column, then split into `df_train` for the first 250 rows then `df_test` for the last 50.

```
knitr::opts_chunk$set(echo = TRUE)
library("tidyverse")
```

```
## -- Attaching packages ----------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------- t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("caret")
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library("GGally")
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library("car")
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library("pROC")
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
setwd("~/Stat 218/Analytics Project 1")
df <- read_csv("data_BaBe.csv") %>%
  select(-c(X1)) %>%
  mutate(REGION=as.factor(REGION),
         WD.Area=as.factor(WD.Area),
         Mun1=as.factor(case_when(Mun1 > 0 ~ 1, TRUE ~ 0)),
         conn_log=log(conn),
         vol_nrw_log=log(vol_nrw),
         wd_rate_log=log(wd_rate),
         conn_p_area_squared=conn_p_area^2,
         nrwpcent_class=case_when(nrwpcent <= 25 ~ 1, TRUE ~ 0) # Engineer categorical variable
         )
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   REGION = col_character(),
##   WD.Area = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
# Engineer Mun1 as factor (1 ~ greater than 0, 0 ~ equal to 0)
df_dummies <- dummyVars(~ ., data=df, fullRank=TRUE) %>% predict(df) %>% as.data.frame()
df_dummies_train <- df_dummies[1:250,]
df_dummies_test <- df_dummies[251:300,]
# Train test split first 250 vs. last 50
df_train <- df[1:250,]
df_test <- df[251:300,]
df %>% head
```

```
## # A tibble: 6 x 24
##   REGION WD.Area  conn conn_p_area wd_rate vol_nrw nrwpcent cities Mun1
##   <fct>  <fct>   <dbl>       <dbl>   <dbl>   <dbl>    <dbl>  <dbl> <fct>
## 1 IV     Area 3   2330       22.7     184    51125       26      0 0
## 2 VI     Area 5  23390        6.08    392. 1408272       31      1 0
## 3 V      Area 4   5268       15.8     368.  368233       40      1 0
## 4 I      Area 1   3255       30.9     283   473606       56      0 1
## 5 II     Area 1   1420       22.3     198.   34465       20      0 0
## 6 IX     Area 9   3484       16.1     348.  208936       36      1 0
## # ... with 15 more variables: Mun2 <dbl>, Mun3 <dbl>, Mun4 <dbl>,
## #   Mun5 <dbl>, gw <dbl>, sprw <dbl>, surw <dbl>, elevar <dbl>,
## #   coastal <dbl>, emp <dbl>, conn_log <dbl>, vol_nrw_log <dbl>,
## #   wd_rate_log <dbl>, conn_p_area_squared <dbl>, nrwpcent_class <dbl>
```

## Exploratory Data Analysis

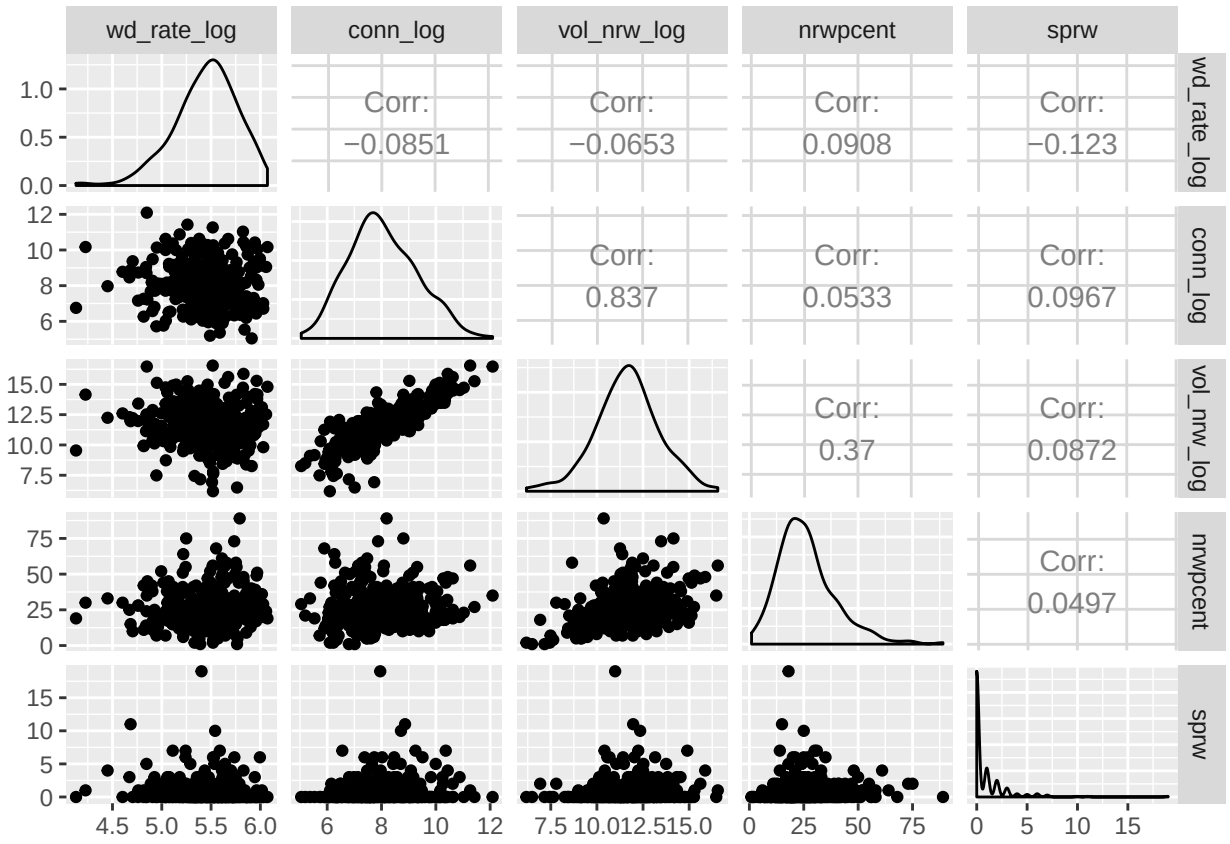
First, which features have the highest correlation with `wd_rate`?

```
cor_matrix <- cor(df_dummies)['wd_rate',] %>% sort()
cor_matrix
```

```
##          REGION.VII      WD.Area.Area 2          REGION.III
##        -0.182740530        -0.150529894        -0.136878748
##                sprw      WD.Area.Area 8          REGION.XII
##        -0.120221429        -0.107408409        -0.089546663
##      nrwpcent_class                conn      WD.Area.Area 6
##        -0.086727179        -0.082385530        -0.076713355
##             coastal            conn_log              elevar
##        -0.073113360        -0.072441415        -0.062493997
##           REGION.XI                  gw         vol_nrw_log
##        -0.060639040        -0.048053821        -0.046842750
##            REGION.V      WD.Area.Area 4                 emp
##        -0.044888088        -0.044888088        -0.040474465
##                Mun4      WD.Area.Area 9              cities
##        -0.032396553        -0.032329079        -0.026534085
##             vol_nrw         conn_p_area                Mun3
##        -0.021870343        -0.016304541        -0.013905236
##                Mun2         REGION.VIII            REGION.X
##        -0.007665626        -0.003789065        -0.002850639
##                Mun5 conn_p_area_squared           REGION.IV
##         0.002648359         0.005802799         0.007427909
##      WD.Area.Area 3      WD.Area.Area 5                surw
##         0.023625071         0.027072948         0.037031650
##           REGION.IX       REGION.CARAGA      WD.Area.Area 7
##         0.050598212         0.073613384         0.076892511
```

```
##          REGION.II          REGION.VI         REGION.CAR
##        0.082929788        0.093431792        0.096479294
##          nrwpcent               Mun1.1          REGION.I
##        0.104999052        0.128373151        0.158123929
##        wd_rate_log            wd_rate
##        0.974764184        1.000000000
```

```
df %>%
  select(c(wd_rate_log, conn_log, vol_nrw_log, nrwpcent, sprw)) %>%
  ggpairs(progress=FALSE)
```




Above, we inspect the correlation of `wd_rate` against `conn_log`, `vol_nrw_log`, `nrwpcent`, and `sprw`. We find a strong correlation between `conn_log` and `vol_new_log`.

## Regression model

We first fit on all the features:

```
# df_train %>% select(-c(wd_rate_log)) %>% lm(wd_rate_log ~ ., data=.) %>% summary()
df_train_regression <- df_train %>% select(-c(nrwpcent_class, wd_rate_log))
df_test_regression <- df_test %>% select(-c(nrwpcent_class, wd_rate_log))
full_model <- lm(wd_rate ~ ., data=df_train_regression)
full_model %>% summary
```

```
##
## Call:
## lm(formula = wd_rate ~ ., data = df_train_regression)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -172.700  -41.882    2.047   40.591  215.437
##
## Coefficients: (5 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.823e+02  8.707e+01   2.093  0.03752 *
## REGIONCAR            1.179e+02  6.744e+01   1.747  0.08201 .
## REGIONCARAGA         9.210e+01  6.016e+01   1.531  0.12729
## REGIONI              1.114e+02  5.821e+01   1.915  0.05689 .
## REGIONII             1.015e+02  5.997e+01   1.693  0.09198 .
## REGIONIII            1.007e+02  7.970e+01   1.263  0.20787
## REGIONIV             9.397e+01  5.785e+01   1.624  0.10577
## REGIONIX             1.022e+02  6.781e+01   1.508  0.13307
## REGIONV              8.210e+01  5.869e+01   1.399  0.16330
## REGIONVI             1.142e+02  6.648e+01   1.717  0.08737 .
## REGIONVII            2.416e+01  6.559e+01   0.368  0.71296
## REGIONVIII           1.202e+02  6.645e+01   1.808  0.07196 .
## REGIONX              9.669e+01  5.709e+01   1.694  0.09179 .
## REGIONXI             7.950e+01  6.336e+01   1.255  0.21097
## REGIONXII            2.429e+01  6.432e+01   0.378  0.70603
## WD.AreaArea 2       -3.806e+01  5.322e+01  -0.715  0.47528
## WD.AreaArea 3               NA         NA      NA       NA
## WD.AreaArea 4               NA         NA      NA       NA
## WD.AreaArea 5       -1.547e+01  3.375e+01  -0.458  0.64725
## WD.AreaArea 6               NA         NA      NA       NA
## WD.AreaArea 7        3.614e+01  4.421e+01   0.817  0.41464
## WD.AreaArea 8               NA         NA      NA       NA
## WD.AreaArea 9               NA         NA      NA       NA
## conn                -1.086e-03  1.176e-03  -0.923  0.35706
## conn_p_area          2.172e-03  1.054e-01   0.021  0.98359
## vol_nrw             -3.090e-06  7.586e-06  -0.407  0.68418
## nrwpcent             7.277e-01  4.569e-01   1.593  0.11267
## cities               4.130e+01  2.009e+01   2.055  0.04107 *
## Mun11                4.499e+01  1.608e+01   2.798  0.00562 **
## Mun2                 2.644e+01  1.676e+01   1.577  0.11617
## Mun3                 1.109e+01  1.603e+01   0.692  0.48983
## Mun4                 5.069e+00  1.018e+01   0.498  0.61923
## Mun5                 1.423e+01  1.341e+01   1.061  0.28994
## gw                  -2.426e-01  3.761e-01  -0.645  0.51951
## sprw                -6.831e+00  2.587e+00  -2.641  0.00889 **
## surw                 1.105e+01  1.069e+01   1.033  0.30273
## elevar              -8.907e-05  9.611e-05  -0.927  0.35515
## coastal             -1.302e+01  1.127e+01  -1.155  0.24923
## emp                  2.180e-01  2.226e-01   0.979  0.32857
## conn_log             2.551e+00  1.143e+01   0.223  0.82354
## vol_nrw_log         -7.266e+00  6.828e+00  -1.064  0.28848
## conn_p_area_squared  1.387e-05  1.083e-04   0.128  0.89819
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.41 on 213 degrees of freedom
## Multiple R-squared:  0.2114, Adjusted R-squared:  0.07813
## F-statistic: 1.586 on 36 and 213 DF,  p-value: 0.02472
```

```r
model_coefs <- summary(full_model)$coeff[-1, 4]
significant_predictors <- model_coefs[model_coefs < 0.10]

significant_predictors
```

```
##   REGIONCAR     REGIONI    REGIONII    REGIONVI  REGIONVIII     REGIONX
## 0.082007861 0.056894478 0.091978592 0.087373921 0.071956468 0.091786356
##      cities       Mun11        sprw
## 0.041069799 0.005615375 0.008890042
```

After fitting the full model, we find that only 9 features have significant betas to within 10%.

```r
# Test set performance benchmark
evaluateRMSE <- function(model, df_set) {
  predictions <- model %>% predict(df_set) %>% as.vector()
  obs <- df_set$wd_rate %>% as.vector()
  rmse <- sqrt(mean((predictions - obs)^2)) / (mean(obs))
  return(rmse)
}

evaluateRMSE(full_model, df_test_regression)
```

```
## [1] 0.2638752
```

### Recursive Feature Selection

After checking the full model, we employ forward, backward and stepwise feature selection:

```r
# Run backward elimination
initial_model <- lm(wd_rate ~ 1, data = df_train_regression)

back_elim <- step(object = full_model, scope = list(lower = initial_model), direction = "backward")
```

```
## Start:  AIC=2161.1
## wd_rate ~ REGION + WD.Area + conn + conn_p_area + vol_nrw + nrwpcent +
##     cities + Mun1 + Mun2 + Mun3 + Mun4 + Mun5 + gw + sprw + surw +
##     elevar + coastal + emp + conn_log + vol_nrw_log + conn_p_area_squared
##
##                        Df Sum of Sq     RSS    AIC
## - WD.Area               3      6741 1062575 2156.7
## - conn_p_area           1         2 1055836 2159.1
## - conn_p_area_squared   1        81 1055915 2159.1
## - REGION                9     69898 1125732 2159.1
## - conn_log              1       247 1056081 2159.2
## - vol_nrw               1       822 1056657 2159.3
## - Mun4                  1      1228 1057062 2159.4
## - gw                    1      2063 1057897 2159.6
## - Mun3                  1      2372 1058206 2159.7
## - conn                  1      4223 1060057 2160.1
## - elevar                1      4257 1060091 2160.1
## - emp                   1      4753 1060587 2160.2
## - surw                  1      5290 1061125 2160.3
## - Mun5                  1      5579 1061413 2160.4
## - vol_nrw_log           1      5613 1061447 2160.4
## - coastal               1      6617 1062451 2160.7
## <none>                              1055834 2161.1
```

```
## - Mun2                 1     12335 1068169 2162.0
## - nrwpcent             1     12577 1068411 2162.1
## - cities               1     20940 1076774 2164.0
## - sprw                 1     34563 1090397 2167.2
## - Mun1                 1     38804 1094638 2168.1
##
## Step:  AIC=2156.69
## wd_rate ~ REGION + conn + conn_p_area + vol_nrw + nrwpcent +
##     cities + Mun1 + Mun2 + Mun3 + Mun4 + Mun5 + gw + sprw + surw +
##     elevar + coastal + emp + conn_log + vol_nrw_log + conn_p_area_squared
##
##                       Df Sum of Sq     RSS    AIC
## - conn_p_area          1         8 1062583 2154.7
## - conn_p_area_squared  1       188 1062763 2154.7
## - vol_nrw              1       257 1062832 2154.8
## - Mun4                 1       715 1063290 2154.8
## - conn_log             1       757 1063333 2154.9
## - gw                   1      1855 1064430 2155.1
## - Mun3                 1      2070 1064645 2155.2
## - elevar               1      4327 1066902 2155.7
## - emp                  1      4806 1067382 2155.8
## - conn                 1      5447 1068022 2156.0
## - Mun5                 1      5503 1068078 2156.0
## - surw                 1      5744 1068319 2156.0
## - vol_nrw_log          1      7287 1069862 2156.4
## - coastal              1      8495 1071070 2156.7
## <none>                             1062575 2156.7
## - nrwpcent             1     13636 1076211 2157.9
## - Mun2                 1     14670 1077245 2158.1
## - cities               1     22033 1084608 2159.8
## - REGION              14    142734 1205309 2160.2
## - sprw                 1     31827 1094402 2162.1
## - Mun1                 1     39638 1102213 2163.8
##
## Step:  AIC=2154.69
## wd_rate ~ REGION + conn + vol_nrw + nrwpcent + cities + Mun1 +
##     Mun2 + Mun3 + Mun4 + Mun5 + gw + sprw + surw + elevar + coastal +
##     emp + conn_log + vol_nrw_log + conn_p_area_squared
##
##                       Df Sum of Sq     RSS    AIC
## - vol_nrw              1       280 1062862 2152.8
## - Mun4                 1       732 1063315 2152.9
## - conn_p_area_squared  1       785 1063367 2152.9
## - conn_log             1       826 1063409 2152.9
## - gw                   1      1854 1064437 2153.1
## - Mun3                 1      2125 1064708 2153.2
## - elevar               1      4354 1066936 2153.7
## - emp                  1      5001 1067584 2153.9
## - conn                 1      5457 1068039 2154.0
## - Mun5                 1      5589 1068172 2154.0
## - surw                 1      5814 1068397 2154.1
## - vol_nrw_log          1      7283 1069866 2154.4
## - coastal              1      8488 1071071 2154.7
## <none>                             1062583 2154.7
```

```
## - nrwpcent             1     13633 1076215 2155.9
## - Mun2                 1     14795 1077377 2156.1
## - cities               1     22817 1085400 2158.0
## - REGION              14    145703 1208286 2158.8
## - sprw                 1     31891 1094474 2160.1
## - Mun1                 1     41206 1103788 2162.2
##
## Step:  AIC=2152.75
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun3 +
##     Mun4 + Mun5 + gw + sprw + surw + elevar + coastal + emp +
##     conn_log + vol_nrw_log + conn_p_area_squared
##
##                       Df Sum of Sq     RSS    AIC
## - Mun4                 1       737 1063600 2150.9
## - conn_p_area_squared  1      1088 1063950 2151.0
## - conn_log             1      1266 1064128 2151.1
## - gw                   1      1593 1064456 2151.1
## - Mun3                 1      2169 1065031 2151.3
## - elevar               1      4175 1067037 2151.7
## - emp                  1      4740 1067602 2151.9
## - Mun5                 1      5456 1068319 2152.0
## - surw                 1      6358 1069220 2152.2
## <none>                             1062862 2152.8
## - vol_nrw_log          1      8641 1071503 2152.8
## - coastal              1      8973 1071836 2152.9
## - conn                 1      9334 1072196 2152.9
## - nrwpcent             1     13552 1076414 2153.9
## - Mun2                 1     14650 1077512 2154.2
## - cities               1     23029 1085892 2156.1
## - REGION              14    145456 1208318 2156.8
## - sprw                 1     31641 1094503 2158.1
## - Mun1                 1     40964 1103826 2160.2
##
## Step:  AIC=2150.93
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun3 +
##     Mun5 + gw + sprw + surw + elevar + coastal + emp + conn_log +
##     vol_nrw_log + conn_p_area_squared
##
##                       Df Sum of Sq     RSS    AIC
## - conn_p_area_squared  1      1036 1064635 2149.2
## - gw                   1      1309 1064909 2149.2
## - conn_log             1      1542 1065142 2149.3
## - Mun3                 1      1904 1065503 2149.4
## - elevar               1      3916 1067516 2149.8
## - emp                  1      4533 1068133 2150.0
## - Mun5                 1      5535 1069134 2150.2
## - surw                 1      6744 1070344 2150.5
## - coastal              1      8477 1072077 2150.9
## <none>                             1063600 2150.9
## - vol_nrw_log          1      8812 1072412 2151.0
## - conn                 1      9239 1072839 2151.1
## - Mun2                 1     13917 1077516 2152.2
## - nrwpcent             1     13966 1077565 2152.2
## - cities               1     22397 1085997 2154.1
```

```
## - REGION              14    148310 1211909 2155.6
## - sprw                 1     31130 1094730 2156.1
## - Mun1                 1     42435 1106034 2158.7
##
## Step:  AIC=2149.17
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun3 +
##     Mun5 + gw + sprw + surw + elevar + coastal + emp + conn_log +
##     vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - gw           1      1203 1065838 2147.4
## - conn_log     1      1569 1066204 2147.5
## - Mun3         1      1742 1066377 2147.6
## - elevar       1      4085 1068720 2148.1
## - Mun5         1      5307 1069942 2148.4
## - emp          1      5639 1070274 2148.5
## - surw         1      6365 1071001 2148.7
## <none>                     1064635 2149.2
## - coastal      1      8881 1073516 2149.2
## - conn         1      8936 1073571 2149.3
## - vol_nrw_log  1      9206 1073841 2149.3
## - Mun2         1     13393 1078028 2150.3
## - nrwpcent     1     14128 1078763 2150.5
## - cities       1     21551 1086186 2152.2
## - REGION      14    147294 1211929 2153.6
## - sprw         1     32222 1096857 2154.6
## - Mun1         1     41785 1106420 2156.8
##
## Step:  AIC=2147.45
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun3 +
##     Mun5 + sprw + surw + elevar + coastal + emp + conn_log +
##     vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - conn_log     1      1549 1067388 2145.8
## - Mun3         1      1593 1067432 2145.8
## - elevar       1      4027 1069865 2146.4
## - Mun5         1      5177 1071015 2146.7
## - emp          1      5373 1071212 2146.7
## - surw         1      7203 1073042 2147.1
## - coastal      1      8483 1074321 2147.4
## <none>                     1065838 2147.4
## - vol_nrw_log  1      9304 1075142 2147.6
## - conn         1     10363 1076202 2147.9
## - Mun2         1     12427 1078265 2148.3
## - nrwpcent     1     14644 1080482 2148.9
## - cities       1     21356 1087194 2150.4
## - REGION      14    146587 1212426 2151.7
## - sprw         1     32540 1098378 2153.0
## - Mun1         1     40682 1106520 2154.8
##
## Step:  AIC=2145.82
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun3 +
##     Mun5 + sprw + surw + elevar + coastal + emp + vol_nrw_log
```

```
##
##               Df Sum of Sq     RSS    AIC
## - Mun3         1      1973 1069361 2144.3
## - elevar       1      4623 1072011 2144.9
## - emp          1      6603 1073991 2145.4
## - Mun5         1      6641 1074028 2145.4
## - surw         1      7890 1075278 2145.7
## <none>                     1067388 2145.8
## - coastal      1      8635 1076023 2145.8
## - vol_nrw_log  1      9969 1077357 2146.1
## - conn         1     10206 1077594 2146.2
## - nrwpcent     1     13336 1080723 2146.9
## - Mun2         1     15167 1082555 2147.3
## - REGION      14    145044 1212431 2149.7
## - cities       1     27949 1095336 2150.3
## - sprw         1     31983 1099371 2151.2
## - Mun1         1     53461 1120849 2156.0
##
## Step:  AIC=2144.28
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun5 +
##     sprw + surw + elevar + coastal + emp + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - elevar       1      3608 1072969 2143.1
## - Mun5         1      5878 1075239 2143.7
## - emp          1      6815 1076176 2143.9
## - surw         1      7176 1076537 2143.9
## - coastal      1      7792 1077153 2144.1
## <none>                     1069361 2144.3
## - vol_nrw_log  1      9735 1079096 2144.5
## - conn         1     10316 1079678 2144.7
## - Mun2         1     13212 1082573 2145.3
## - nrwpcent     1     14380 1083741 2145.6
## - REGION      14    144059 1213421 2147.9
## - cities       1     26361 1095722 2148.4
## - sprw         1     30535 1099897 2149.3
## - Mun1         1     54613 1123974 2154.7
##
## Step:  AIC=2143.12
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + Mun5 +
##     sprw + surw + coastal + emp + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - Mun5         1      5418 1078388 2142.4
## - emp          1      6953 1079922 2142.7
## - surw         1      7208 1080177 2142.8
## - coastal      1      8403 1081372 2143.1
## <none>                     1072969 2143.1
## - vol_nrw_log  1      9701 1082670 2143.4
## - conn         1     10172 1083141 2143.5
## - Mun2         1     13893 1086863 2144.3
## - nrwpcent     1     14038 1087007 2144.4
## - cities       1     25078 1098048 2146.9
## - REGION      14    148453 1221423 2147.5
```

```
## - sprw         1     32618 1105588 2148.6
## - Mun1         1     53684 1126654 2153.3
##
## Step:  AIC=2142.38
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + sprw +
##     surw + coastal + emp + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - emp          1      6615 1085003 2141.9
## - surw         1      6923 1085311 2142.0
## <none>                     1078388 2142.4
## - coastal      1      8880 1087268 2142.4
## - conn         1      9334 1087722 2142.5
## - vol_nrw_log  1     11573 1089961 2143.1
## - Mun2         1     12210 1090598 2143.2
## - nrwpcent     1     14816 1093203 2143.8
## - cities       1     22691 1101079 2145.6
## - REGION      14    144487 1222875 2145.8
## - sprw         1     32791 1111179 2147.9
## - Mun1         1     49553 1127941 2151.6
##
## Step:  AIC=2141.91
## wd_rate ~ REGION + conn + nrwpcent + cities + Mun1 + Mun2 + sprw +
##     surw + coastal + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - conn         1      3057 1088060 2140.6
## - surw         1      5267 1090270 2141.1
## <none>                     1085003 2141.9
## - vol_nrw_log  1      9828 1094831 2142.2
## - coastal      1     10167 1095169 2142.2
## - Mun2         1     12235 1097238 2142.7
## - nrwpcent     1     15090 1100092 2143.4
## - REGION      14    149429 1234432 2146.2
## - cities       1     28250 1113252 2146.3
## - sprw         1     31712 1116715 2147.1
## - Mun1         1     49966 1134969 2151.2
##
## Step:  AIC=2140.61
## wd_rate ~ REGION + nrwpcent + cities + Mun1 + Mun2 + sprw + surw +
##     coastal + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - surw         1      3590 1091650 2139.4
## <none>                     1088060 2140.6
## - coastal      1     10112 1098172 2140.9
## - Mun2         1     12436 1100496 2141.4
## - nrwpcent     1     17306 1105366 2142.6
## - vol_nrw_log  1     21200 1109259 2143.4
## - cities       1     27389 1115449 2144.8
## - REGION      14    151671 1239731 2145.2
## - sprw         1     30501 1118560 2145.5
## - Mun1         1     51438 1139498 2150.2
##
```

```
## Step:  AIC=2139.43
## wd_rate ~ REGION + nrwpcent + cities + Mun1 + Mun2 + sprw + coastal +
##     vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## - coastal      1      8758 1100408 2139.4
## <none>                     1091650 2139.4
## - Mun2         1     12003 1103653 2140.2
## - nrwpcent     1     17364 1109013 2141.4
## - vol_nrw_log  1     20164 1111814 2142.0
## - REGION      14    150363 1242013 2143.7
## - cities       1     28437 1120087 2143.9
## - sprw         1     29121 1120771 2144.0
## - Mun1         1     50492 1142142 2148.7
##
## Step:  AIC=2139.43
## wd_rate ~ REGION + nrwpcent + cities + Mun1 + Mun2 + sprw + vol_nrw_log
##
##               Df Sum of Sq     RSS    AIC
## <none>                     1100408 2139.4
## - Mun2         1     12817 1113225 2140.3
## - nrwpcent     1     14764 1115171 2140.8
## - vol_nrw_log  1     15979 1116387 2141.0
## - cities       1     25861 1126268 2143.2
## - sprw         1     28809 1129217 2143.9
## - REGION      14    153609 1254017 2144.1
## - Mun1         1     51445 1151852 2148.9
```

After running backward elimination, we find a model with AIC of 2139.43. We will now run the forward elimination and the stepwise elimination:

```
# Run forward elimination
forward_elim <- step(object = initial_model, scope = list(upper = full_model), direction = "forward")
```

```
## Start:  AIC=2148.47
## wd_rate ~ 1
##
##                       Df Sum of Sq     RSS    AIC
## + sprw                 1     27063 1311828 2145.4
## + WD.Area              8     96307 1242584 2145.8
## + nrwpcent             1     21568 1317323 2146.4
## + REGION              14    148622 1190269 2147.1
## + Mun1                 1     11916 1326975 2148.2
## <none>                             1338891 2148.5
## + conn                 1      9988 1328903 2148.6
## + coastal              1      9525 1329366 2148.7
## + elevar               1      7823 1331068 2149.0
## + vol_nrw              1      6430 1332461 2149.3
## + conn_log             1      4727 1334164 2149.6
## + emp                  1      4028 1334863 2149.7
## + gw                   1      3670 1335221 2149.8
## + conn_p_area          1      2919 1335972 2149.9
## + vol_nrw_log          1      2865 1336026 2149.9
## + Mun3                 1      1754 1337137 2150.2
## + conn_p_area_squared  1       679 1338212 2150.3
```

```
## + cities               1       646 1338245 2150.3
## + Mun5                 1       257 1338634 2150.4
## + Mun2                 1        54 1338837 2150.5
## + surw                 1        31 1338859 2150.5
## + Mun4                 1         6 1338885 2150.5
##
## Step:  AIC=2145.37
## wd_rate ~ sprw
##
##                       Df Sum of Sq     RSS    AIC
## + WD.Area              8    104936 1206892 2140.5
## + nrwpcent             1     22570 1289258 2143.0
## + REGION              14    147267 1164561 2143.6
## + Mun1                 1     12126 1299703 2145.1
## <none>                             1311828 2145.4
## + conn                 1     10211 1301617 2145.4
## + vol_nrw              1      6771 1305057 2146.1
## + elevar               1      5764 1306064 2146.3
## + coastal              1      5462 1306366 2146.3
## + conn_p_area          1      5231 1306597 2146.4
## + gw                   1      3912 1307916 2146.6
## + emp                  1      3382 1308447 2146.7
## + conn_log             1      2800 1309028 2146.8
## + vol_nrw_log          1      1732 1310096 2147.0
## + conn_p_area_squared  1      1669 1310159 2147.1
## + Mun3                 1      1216 1310612 2147.1
## + surw                 1       484 1311344 2147.3
## + Mun5                 1       421 1311407 2147.3
## + Mun2                 1       197 1311631 2147.3
## + Mun4                 1       110 1311718 2147.3
## + cities               1        83 1311745 2147.3
##
## Step:  AIC=2140.52
## wd_rate ~ sprw + WD.Area
##
##                       Df Sum of Sq     RSS    AIC
## + Mun1                 1     22798 1184094 2137.8
## <none>                             1206892 2140.5
## + nrwpcent             1      7788 1199104 2140.9
## + coastal              1      4738 1202154 2141.5
## + vol_nrw              1      4624 1202268 2141.6
## + Mun3                 1      3327 1203565 2141.8
## + surw                 1      3158 1203734 2141.9
## + conn                 1      2979 1203913 2141.9
## + Mun4                 1      2350 1204542 2142.0
## + gw                   1      1865 1205028 2142.1
## + elevar               1      1050 1205842 2142.3
## + emp                  1       777 1206116 2142.4
## + conn_p_area          1       406 1206486 2142.4
## + vol_nrw_log          1       295 1206597 2142.5
## + Mun5                 1       104 1206789 2142.5
## + conn_p_area_squared  1        88 1206805 2142.5
## + cities               1        47 1206846 2142.5
## + conn_log             1        11 1206882 2142.5
```

```
## + Mun2                 1         0 1206892 2142.5
## + REGION               9     58280 1148612 2146.2
##
## Step:  AIC=2137.76
## wd_rate ~ sprw + WD.Area + Mun1
##
##                       Df Sum of Sq     RSS    AIC
## <none>                             1184094 2137.8
## + nrwpcent             1      6729 1177365 2138.3
## + cities               1      4290 1179804 2138.8
## + surw                 1      4281 1179813 2138.8
## + vol_nrw              1      3392 1180702 2139.0
## + coastal              1      3115 1180979 2139.1
## + gw                   1      2517 1181576 2139.2
## + conn                 1      2041 1182053 2139.3
## + Mun2                 1      1741 1182352 2139.4
## + elevar               1      1307 1182787 2139.5
## + Mun3                 1       757 1183337 2139.6
## + vol_nrw_log          1       623 1183471 2139.6
## + Mun5                 1       582 1183512 2139.6
## + conn_p_area          1       385 1183709 2139.7
## + emp                  1       291 1183803 2139.7
## + Mun4                 1       202 1183892 2139.7
## + conn_log             1       153 1183941 2139.7
## + conn_p_area_squared  1        47 1184047 2139.8
## + REGION               9     58927 1125166 2143.0
```

```
# Run stepwise selection
step_sel <- step(object = initial_model, scope = list(upper = full_model), direction = "both")
```

```
## Start:  AIC=2148.47
## wd_rate ~ 1
##
##                       Df Sum of Sq     RSS    AIC
## + sprw                 1     27063 1311828 2145.4
## + WD.Area              8     96307 1242584 2145.8
## + nrwpcent             1     21568 1317323 2146.4
## + REGION              14    148622 1190269 2147.1
## + Mun1                 1     11916 1326975 2148.2
## <none>                             1338891 2148.5
## + conn                 1      9988 1328903 2148.6
## + coastal              1      9525 1329366 2148.7
## + elevar               1      7823 1331068 2149.0
## + vol_nrw              1      6430 1332461 2149.3
## + conn_log             1      4727 1334164 2149.6
## + emp                  1      4028 1334863 2149.7
## + gw                   1      3670 1335221 2149.8
## + conn_p_area          1      2919 1335972 2149.9
## + vol_nrw_log          1      2865 1336026 2149.9
## + Mun3                 1      1754 1337137 2150.2
## + conn_p_area_squared  1       679 1338212 2150.3
## + cities               1       646 1338245 2150.3
## + Mun5                 1       257 1338634 2150.4
## + Mun2                 1        54 1338837 2150.5
## + surw                 1        31 1338859 2150.5
```

```
## + Mun4                 1         6 1338885 2150.5
##
## Step:  AIC=2145.37
## wd_rate ~ sprw
##
##                       Df Sum of Sq     RSS    AIC
## + WD.Area              8    104936 1206892 2140.5
## + nrwpcent             1     22570 1289258 2143.0
## + REGION              14    147267 1164561 2143.6
## + Mun1                 1     12126 1299703 2145.1
## <none>                             1311828 2145.4
## + conn                 1     10211 1301617 2145.4
## + vol_nrw              1      6771 1305057 2146.1
## + elevar               1      5764 1306064 2146.3
## + coastal              1      5462 1306366 2146.3
## + conn_p_area          1      5231 1306597 2146.4
## + gw                   1      3912 1307916 2146.6
## + emp                  1      3382 1308447 2146.7
## + conn_log             1      2800 1309028 2146.8
## + vol_nrw_log          1      1732 1310096 2147.0
## + conn_p_area_squared  1      1669 1310159 2147.1
## + Mun3                 1      1216 1310612 2147.1
## + surw                 1       484 1311344 2147.3
## + Mun5                 1       421 1311407 2147.3
## + Mun2                 1       197 1311631 2147.3
## + Mun4                 1       110 1311718 2147.3
## + cities               1        83 1311745 2147.3
## - sprw                 1     27063 1338891 2148.5
##
## Step:  AIC=2140.52
## wd_rate ~ sprw + WD.Area
##
##                       Df Sum of Sq     RSS    AIC
## + Mun1                 1     22798 1184094 2137.8
## <none>                             1206892 2140.5
## + nrwpcent             1      7788 1199104 2140.9
## + coastal              1      4738 1202154 2141.5
## + vol_nrw              1      4624 1202268 2141.6
## + Mun3                 1      3327 1203565 2141.8
## + surw                 1      3158 1203734 2141.9
## + conn                 1      2979 1203913 2141.9
## + Mun4                 1      2350 1204542 2142.0
## + gw                   1      1865 1205028 2142.1
## + elevar               1      1050 1205842 2142.3
## + emp                  1       777 1206116 2142.4
## + conn_p_area          1       406 1206486 2142.4
## + vol_nrw_log          1       295 1206597 2142.5
## + Mun5                 1       104 1206789 2142.5
## + conn_p_area_squared  1        88 1206805 2142.5
## + cities               1        47 1206846 2142.5
## + conn_log             1        11 1206882 2142.5
## + Mun2                 1         0 1206892 2142.5
## - WD.Area              8    104936 1311828 2145.4
## - sprw                 1     35691 1242584 2145.8
```

```
## + REGION               9     58280 1148612 2146.2
##
## Step:  AIC=2137.76
## wd_rate ~ sprw + WD.Area + Mun1
##
##                        Df Sum of Sq     RSS    AIC
## <none>                              1184094 2137.8
## + nrwpcent              1      6729 1177365 2138.3
## + cities                1      4290 1179804 2138.8
## + surw                  1      4281 1179813 2138.8
## + vol_nrw               1      3392 1180702 2139.0
## + coastal               1      3115 1180979 2139.1
## + gw                    1      2517 1181576 2139.2
## + conn                  1      2041 1182053 2139.3
## + Mun2                  1      1741 1182352 2139.4
## + elevar                1      1307 1182787 2139.5
## + Mun3                  1       757 1183337 2139.6
## + vol_nrw_log           1       623 1183471 2139.6
## + Mun5                  1       582 1183512 2139.6
## + conn_p_area           1       385 1183709 2139.7
## + emp                   1       291 1183803 2139.7
## + Mun4                  1       202 1183892 2139.7
## + conn_log              1       153 1183941 2139.7
## + conn_p_area_squared   1        47 1184047 2139.8
## - Mun1                  1     22798 1206892 2140.5
## + REGION                9     58927 1125166 2143.0
## - sprw                  1     38485 1222579 2143.8
## - WD.Area               8    115609 1299703 2145.1
```

After using these techniques, we find the "best" one:

```
fitstat <- function(x) {
  xsum <- summary(x)
  resid <- x$residuals
  fit <- x$fitted.values
  test_rmse <- evaluateRMSE(x, df_test_regression)
  return(c(R2 = xsum$r.squared,
           R2Adj = xsum$adj.r.squared,
           AIC = AIC(x),
           BIC = BIC(x),
           MSE = mean(resid^2),
           MAPE = mean(abs(resid/fit)),
           RMSE_test = test_rmse
  ))}

sapply(list('Full Model' = full_model,
            'Forward Elimination' = forward_elim,
            'Backward Elimination' = back_elim,
            'Stepwise Selection' = step_sel), fitstat)
```

```
##             Full Model Forward Elimination Backward Elimination
## R2        2.114113e-01        1.156158e-01            0.1781201
## R2Adj     7.812876e-02        7.861232e-02            0.1063402
## AIC       2.872564e+03        2.849226e+03         2850.9018232
## BIC       3.006380e+03        2.891484e+03         2928.3739634
```

```
## MSE       4.223337e+03        4.736375e+03         4401.6300904
## MAPE      2.099369e-01        2.257036e-01            0.2144362
## RMSE_test 2.638752e-01        2.558602e-01            0.2496572
##           Stepwise Selection
## R2              1.156158e-01
## R2Adj           7.861232e-02
## AIC             2.849226e+03
## BIC             2.891484e+03
## MSE             4.736375e+03
## MAPE            2.257036e-01
## RMSE_test       2.558602e-01
```

In summary, the backward elimination procedure found the model with the lowest test RMSE of 0.2496. However, if we inspect the significance of the best model, we find that only 8 of the 21 features are significant.

```
back_elim %>% summary
```

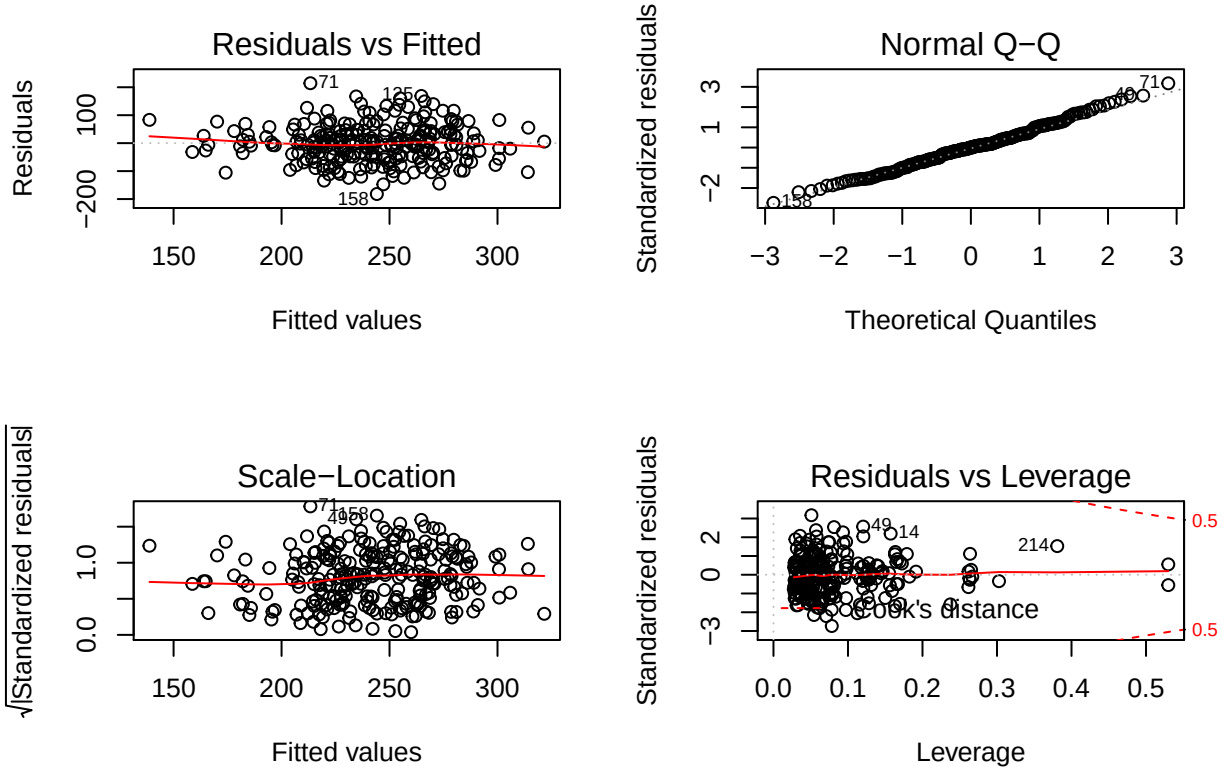
```
##
## Call:
## lm(formula = wd_rate ~ REGION + nrwpcent + cities + Mun1 + Mun2 +
##     sprw + vol_nrw_log, data = df_train_regression)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -181.927  -46.303    0.141   39.296  214.347
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  218.3429    65.8954   3.313  0.00107 **
## REGIONCAR    105.2265    61.4074   1.714  0.08796 .
## REGIONCARAGA 102.4468    53.9628   1.898  0.05889 .
## REGIONI       91.4357    52.2691   1.749  0.08157 .
## REGIONII      76.0140    53.9002   1.410  0.15982
## REGIONIII     38.3694    52.2917   0.734  0.46385
## REGIONIV      68.8834    51.2303   1.345  0.18009
## REGIONIX      82.7541    61.5143   1.345  0.17986
## REGIONV       54.8512    52.4528   1.046  0.29679
## REGIONVI      71.8292    52.0796   1.379  0.16917
## REGIONVII     -2.6772    57.1546  -0.047  0.96268
## REGIONVIII    83.2179    56.7378   1.467  0.14383
## REGIONX       83.7406    54.5612   1.535  0.12621
## REGIONXI      36.7077    55.9622   0.656  0.51252
## REGIONXII     -1.9152    57.7241  -0.033  0.97356
## nrwpcent       0.6830     0.3896   1.753  0.08097 .
## cities        36.9202    15.9148   2.320  0.02123 *
## Mun11         39.3662    12.0313   3.272  0.00123 **
## Mun2          23.2076    14.2101   1.633  0.10381
## sprw          -5.9596     2.4340  -2.449  0.01510 *
## vol_nrw_log   -6.6233     3.6321  -1.824  0.06952 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.32 on 229 degrees of freedom
## Multiple R-squared:  0.1781, Adjusted R-squared:  0.1063
## F-statistic: 2.481 on 20 and 229 DF,  p-value: 0.0006572
```

```
back_elim %>% vif
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## REGION      2.281696 14        1.029900
## nrwpcent    1.367237  1        1.169289
## cities      2.319411  1        1.522961
## Mun1        1.726566  1        1.313989
## Mun2        1.324694  1        1.150954
## sprw        1.265760  1        1.125060
## vol_nrw_log 1.861680  1        1.364434
```

In terms of enforcing the multicollinearity assumption, the model coefficients all have VIFs less than 10.

```
par(mfrow=c(2,2))
plot(back_elim)
```




To validate the regression model's assumptions, we plot the above charts, and we find that the model fits the linear regression model assumptions.

### All Possible Model's approach

In this secion, we search for a model by trying all possible combinations after filtering out insignificant features through univariate filtering. We first run a correlation plot against the predictor to prune out unrelated features, such that it is close to 0.

```
cor_matrix <- cor(df_dummies_train)['wd_rate',] %>% abs %>% sort
cor_matrix
```

```
##                Mun4            REGION.X                surw
##         0.002176644         0.004127298         0.004844616
##                Mun2                Mun5           REGION.IV
##         0.006333579         0.013843862         0.019231486
##      WD.Area.Area 9              cities conn_p_area_squared
##         0.019883513         0.021967031         0.022520354
##           REGION.IX      WD.Area.Area 5          REGION.VIII
##         0.025978328         0.029659753         0.032747400
##                Mun3      WD.Area.Area 3          vol_nrw_log
##         0.036195087         0.037391797         0.046257898
##         conn_p_area      WD.Area.Area 6                   gw
##         0.046691277         0.051211933         0.052351763
##            REGION.V      WD.Area.Area 4                 emp
##         0.053716281         0.053716281         0.054851465
##            conn_log           REGION.XI      WD.Area.Area 7
##         0.059416782         0.061981726         0.062328494
##             vol_nrw           REGION.II           REGION.VI
##         0.069297382         0.071160389         0.075435351
##              elevar          REGION.CAR       REGION.CARAGA
##         0.076438403         0.077311674         0.082147129
##             coastal                conn              Mun1.1
##         0.084343336         0.086371315         0.094338423
##          REGION.XII      nrwpcent_class      WD.Area.Area 8
##         0.099769620         0.106208521         0.114424971
##            nrwpcent          REGION.III                sprw
##         0.126920348         0.137618695         0.142171622
##            REGION.I      WD.Area.Area 2          REGION.VII
##         0.150535557         0.153733098         0.163678598
##         wd_rate_log             wd_rate
##         0.972815404         1.000000000
```

Based on the above, we remove the features whose correlation is less than 0.05: `Mun4`, `surw`, `Mun2`, `Mun5`, `cities`, `Mun3`, `conn_p_area`, `vol_nrw_log`, `conn_p_area_squared`. We also decided to remove `WD.Area` because it produces null betas.

```
cor_matrix[cor_matrix < 0.05]
```

```
##                Mun4            REGION.X                surw
##         0.002176644         0.004127298         0.004844616
##                Mun2                Mun5           REGION.IV
##         0.006333579         0.013843862         0.019231486
##      WD.Area.Area 9              cities conn_p_area_squared
##         0.019883513         0.021967031         0.022520354
##           REGION.IX      WD.Area.Area 5          REGION.VIII
##         0.025978328         0.029659753         0.032747400
##                Mun3      WD.Area.Area 3          vol_nrw_log
##         0.036195087         0.037391797         0.046257898
##         conn_p_area
##         0.046691277
```

We retain `REGION` because we find that other categoricals have correlation >= 0.05.

```
# Subset constructor
formulaConstructor <- function(predictors) {
  predictors %>% paste(collapse=" + ") %>% paste("wd_rate ~", .) %>% as.formula()
}
```

```r
# Define regression evaluation function
evaluateRegression <- function(model, data, model_name) {
  model_summary <- model %>% summary
  r_squared <- model_summary$r.squared
  adj_r_squared <- model_summary$adj.r.squared
  aic <- model %>% AIC
  bic <- model %>% BIC
  train_mse <- evaluateRMSE(model, df_train_regression)
  test_mse <- evaluateRMSE(model, df_test_regression)
  ps <- model_summary$coeff[,4]
  num_significant <- length(ps[ps < 0.1]) - 1

  if (length(ps) >= 2) {
    vifs <- model %>% vif()
    vifs <- vifs[vifs > 10] %>% length
  } else {
      vifs <- 0
    }
  eval <-
    list(r_squared, adj_r_squared, aic, bic, train_mse, test_mse, num_significant, vifs) %>%
    as.data.frame()

  names(eval) <- c('r_squared', 'adj_r_squared', 'aic', 'bic',
                   'train_mse', 'test_mse', 'num_significant', 'vifs')
  eval <- eval %>% mutate(model_name=model_name,
                          num_predictors=model$rank - 1)
  return(eval)
}

# We remove WD.Area
sub_df <- df_train_regression %>%
  select(-c(Mun2, Mun3, Mun4, Mun5, cities, surw, conn_p_area,
            vol_nrw_log, conn_p_area_squared, WD.Area))
predictors <- sub_df %>% colnames
predictors <- predictors[predictors != 'wd_rate']

# Subset constructor
formulaConstructor <- function(predictors) {
  predictors %>% paste(collapse=" + ") %>% paste("wd_rate ~", .)
}

# Define powerset function to generate all possible models
powerset = function(set){
    ps = list()
    ps[[1]] = numeric()
    for(element in set){
        temp = vector(mode="list",length=length(ps))
        for(subset in 1:length(ps)){
            temp[[subset]] = c(ps[[subset]],element)
        }
        ps=c(ps,temp)
    }
    return(ps)
```

```
}
```

```
powerset_results <- powerset(predictors) %>%
  Filter(function(x) length(x) >= 2, .) %>%
  Map(function(x) {
    model_spec <- formulaConstructor(x)
    restricted_model <- lm(model_spec[1], data=sub_df)
    restricted_model_results <- evaluateRegression(restricted_model, metrics, model_spec)
    return(restricted_model_results)
    }, .) %>%
  Reduce(function(x, y) {
    return(x %>% rbind(y))
  }, .)
```

```
powerset_results %>% head
```

```
##     r_squared adj_r_squared      aic      bic train_mse  test_mse
## 1 0.112895307  0.0560296219 2859.994 2919.859 0.2845395 0.2528277
## 2 0.112967841  0.0561068049 2859.974 2919.838 0.2845279 0.2558208
## 3 0.007529907 -0.0005062884 2862.052 2876.138 0.3009635 0.2741843
## 4 0.113082541  0.0521783382 2861.941 2925.328 0.2845095 0.2545748
## 5 0.119069395  0.0625994846 2858.248 2918.113 0.2835476 0.2561151
## 6 0.024842322  0.0169463087 2857.653 2871.739 0.2983270 0.2778805
##   num_significant vifs                        model_name num_predictors
## 1               5    1           wd_rate ~ REGION + conn             15
## 2               5    1        wd_rate ~ REGION + vol_nrw             15
## 3               0    0          wd_rate ~ conn + vol_nrw              2
## 4               5    1 wd_rate ~ REGION + conn + vol_nrw             16
## 5               8    1       wd_rate ~ REGION + nrwpcent             15
## 6               1    0         wd_rate ~ conn + nrwpcent              2
```

```
powerset_results %>% filter(num_significant == num_predictors)
```

```
##    r_squared adj_r_squared      aic      bic train_mse  test_mse
## 1 0.03706982    0.02927282 2854.498 2868.584 0.2964507 0.2806605
##   num_significant vifs                model_name num_predictors
## 1               2    0 wd_rate ~ nrwpcent + sprw              2
```

We desire a model that is both predictive in the test set and has high proportion of significant predictors. We engineer a desirability score to capture this sentiment to balance the tradeoff between interpretability and predictive power.

```
powerset_results %>%
  mutate(prop_sig=num_significant / num_predictors) %>%
  mutate(desirabiity_score = prop_sig / test_mse) %>%
  arrange(desc(desirabiity_score)) %>% head
```

```
##    r_squared adj_r_squared      aic      bic train_mse  test_mse
## 1 0.03706982    0.02927282 2854.498 2868.584 0.2964507 0.2806605
## 2 0.04931754    0.02983634 2857.298 2881.948 0.2945594 0.2672976
## 3 0.04903867    0.02955176 2857.372 2882.022 0.2946026 0.2697454
## 4 0.06386106    0.04467789 2853.444 2878.094 0.2922976 0.2717965
## 5 0.04826673    0.03272823 2855.574 2876.703 0.2947221 0.2693310
## 6 0.05096343    0.03546895 2854.865 2875.994 0.2943043 0.2700578
##   num_significant vifs                                  model_name
## 1               2    0                   wd_rate ~ nrwpcent + sprw
```

```
## 2               4    1 wd_rate ~ conn + Mun1 + sprw + emp + conn_log
## 3               4    0       wd_rate ~ conn + Mun1 + gw + sprw + emp
## 4               4    0 wd_rate ~ conn + nrwpcent + Mun1 + sprw + emp
## 5               3    0            wd_rate ~ conn + Mun1 + sprw + emp
## 6               3    0   wd_rate ~ nrwpcent + Mun1 + sprw + conn_log
##   num_predictors prop_sig desirabiity_score
## 1              2     1.00          3.563024
## 2              5     0.80          2.992919
## 3              5     0.80          2.965759
## 4              5     0.80          2.943379
## 5              4     0.75          2.784677
## 6              4     0.75          2.777183
```

By inspection, we select the most desirable since it has just 2 predictors that are all significant and a test mse of 0.2806605 and no coefficients with high VIFs.

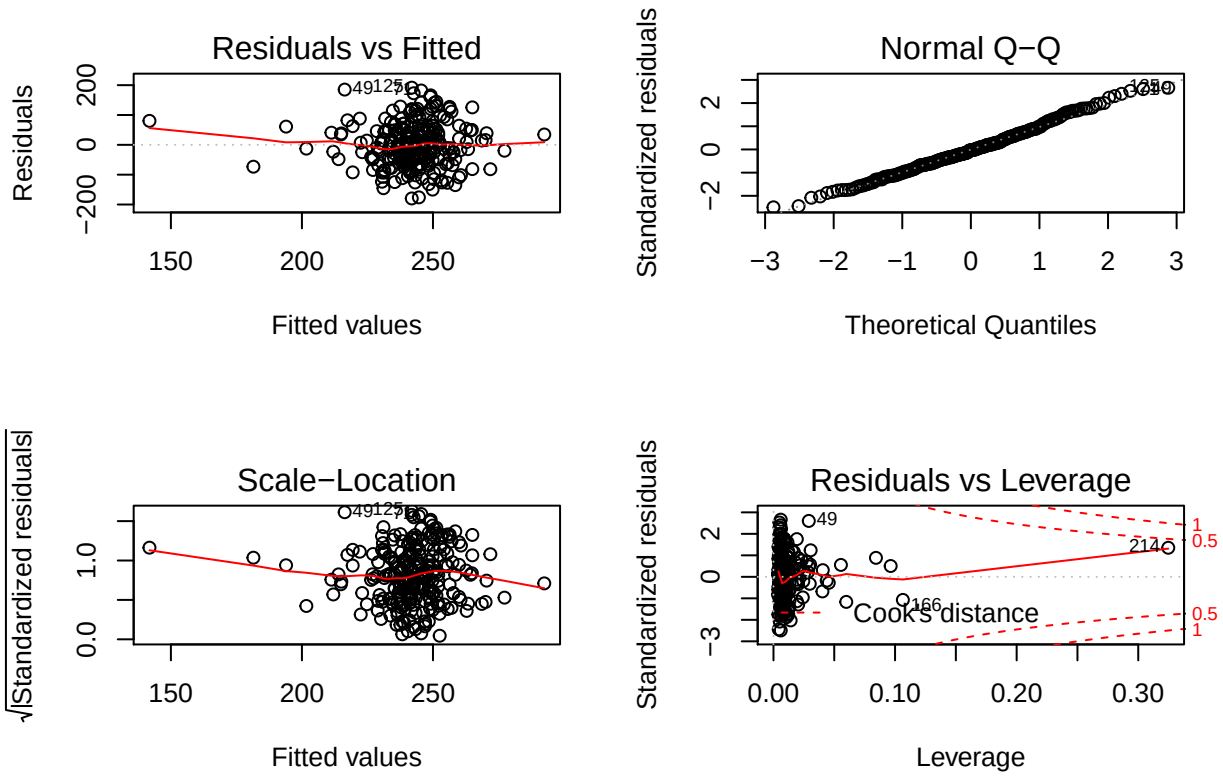
```
desirable_model <- lm(wd_rate ~ nrwpcent + sprw, data=df_train_regression)
desirable_model %>% summary
```

```
##
## Call:
## lm(formula = wd_rate ~ nrwpcent + sprw, data = df_train_regression)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -179.702  -47.857   -2.563   46.742  191.598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 228.1777    10.3555  22.034   <2e-16 ***
## nrwpcent      0.7223     0.3474   2.079   0.0386 *
## sprw         -5.2293     2.2552  -2.319   0.0212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.25 on 247 degrees of freedom
## Multiple R-squared:  0.03707,    Adjusted R-squared:  0.02927
## F-statistic: 4.754 on 2 and 247 DF,  p-value: 0.009418
```

```
desirable_model %>% vif
```

```
## nrwpcent     sprw
## 1.000413 1.000413
```

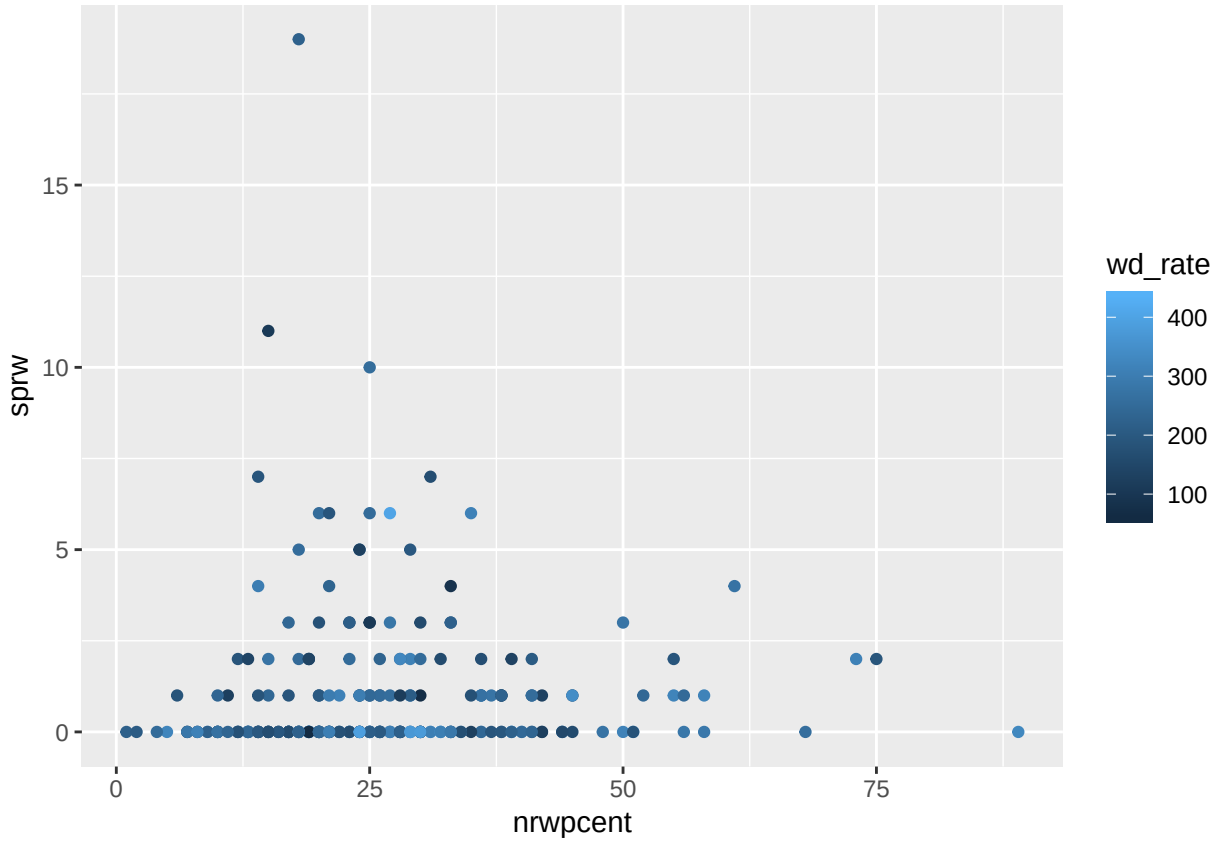
```
par(mfrow=c(2,2))
plot(desirable_model)
```



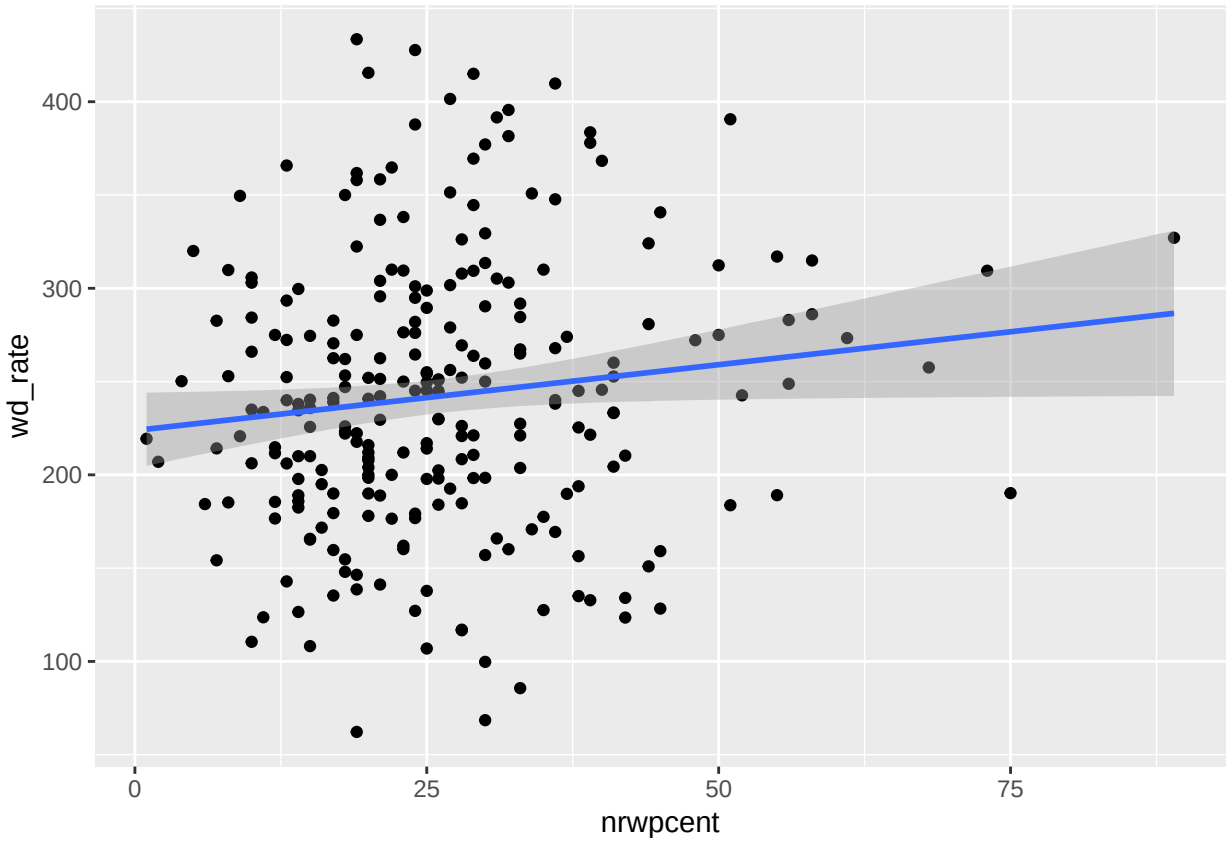

This desirable model seems to meet all the regression assumptions. We therefore make this our final model, with the following interpretation:

1. For every increase in `nrwpcent` (percent of non-revenue water from total displaced water), we can expect a 0.73 increase in the water rate price.
2. For every increase in `sprw` (number of spring water sources), we can expect a 5.22 decrease in the water rate price.

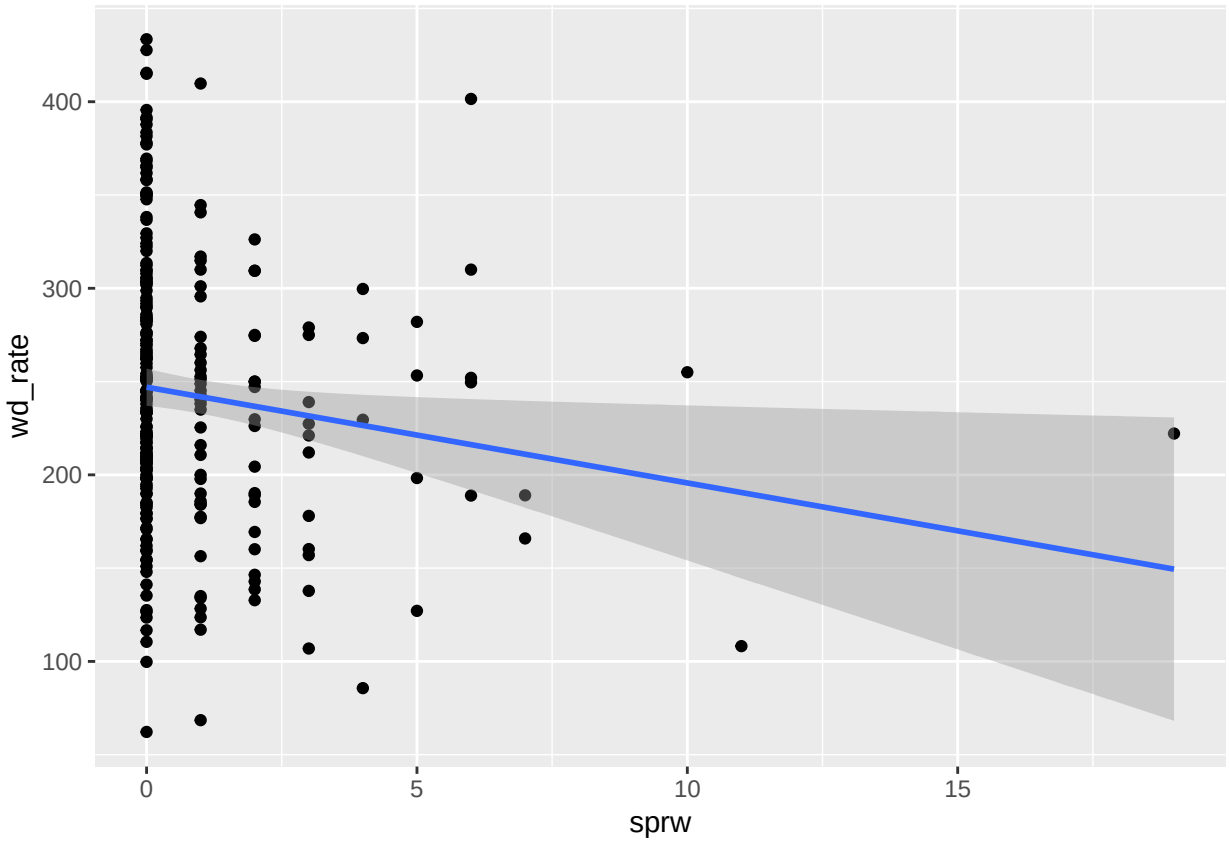
```
df_train %>% ggplot(aes(x=nrwpcent, y=sprw, color=wd_rate)) + geom_point()
```




```
df_train %>% ggplot(aes(x=nrwpcent, y=wd_rate)) + geom_point() + geom_smooth(method="lm")
```




```
df_train %>% ggplot(aes(x=sprw, y=wd_rate)) + geom_point() + geom_smooth(method="lm")
```




We inspect the scatterplot of these significant variables to the `wd_rate` and indeed see the trend that corresponds to the betas.

## Classification

In this section, we build a classification model to predict whether a given water district has `nrwpcent_class` less than or equal to 25% or greater.

We first inspect the class balance:

```
# Inspect class balance
df_train %>% group_by(nrwpcent_class) %>% summarise(freq=n())
```

```
## # A tibble: 2 x 2
##   nrwpcent_class  freq
##            <dbl> <int>
## 1              0   113
## 2              1   137
```

```
df_train_classification <- df_dummies_train %>%
  select(-c(wd_rate_log, wd_rate, vol_nrw, vol_nrw_log, nrwpcent))
df_test_classification <- df_dummies_test %>%
  select(-c(wd_rate_log, wd_rate, vol_nrw, vol_nrw_log, nrwpcent))
```

```
# We remove some features in this section
sub_df <- df_train_classification %>%
  select(-c(Mun2, Mun3, Mun4, Mun5, conn_p_area, conn_p_area_squared,
            `WD.Area.Area 2`, `WD.Area.Area 3`, `WD.Area.Area 4`, `WD.Area.Area 5`,
```

```
            `WD.Area.Area 6`, `WD.Area.Area 7`, `WD.Area.Area 8`, `WD.Area.Area 9`))
predictors <- sub_df %>% colnames
predictors <- predictors[predictors != 'nrwpcent_class']

# Subset constructor for classification
formulaConstructor <- function(predictors) {
  predictors %>% paste(collapse=" + ") %>% paste("nrwpcent_class ~", .)
}
```

Again, similar to the "all models approach" in the linear regression we made, we also create a function for the classification model:

```
# Define function to evaluate classification model
evaluateClassification <- function(model, model_name) {
  model_summary <- model %>% summary
  ps <- model_summary$coeff[,4]
  num_significant <- length(ps[ps < 0.1])
  predict_probabilities <- predict(model, df_dummies_test)
  auc <- roc(df_dummies_test$nrwpcent_class, predict_probabilities)$auc[1]

  if (length(ps) >= 2) {
    vifs <- model %>% vif()
    vifs <- vifs[vifs > 10] %>% length
  } else {
      vifs <- 0
  }

  eval <- list(auc, num_significant, vifs) %>%
    as.data.frame()

  names(eval) <- c('auc', 'num_significant', 'vifs')
  eval <- eval %>% mutate(model_name=model_name,
                          num_predictors=model$rank - 1)
  return(eval)
}
```

```
full_model_classification <- predictors %>%
  formulaConstructor%>% lm(data=sub_df, family="binomial")
full_model_classification %>% summary
```

```
##
## Call:
## lm(formula = ., data = sub_df, family = "binomial")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95257 -0.41613  0.02364  0.46112  0.91110
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.199e-02  4.713e-01   0.195   0.8454
## REGION.CAR     3.244e-01  4.056e-01   0.800   0.4246
## REGION.CARAGA -6.281e-02  3.567e-01  -0.176   0.8604
## REGION.I       8.803e-02  3.445e-01   0.256   0.7985
## REGION.II      8.877e-02  3.549e-01   0.250   0.8027
```

```
## REGION.III     5.290e-01  3.409e-01   1.552   0.1221
## REGION.IV      1.199e-01  3.372e-01   0.355   0.7226
## REGION.IX     -4.950e-02  4.090e-01  -0.121   0.9038
## REGION.V       1.427e-02  3.455e-01   0.041   0.9671
## REGION.VI      8.472e-02  3.448e-01   0.246   0.8061
## REGION.VII     4.357e-01  3.738e-01   1.165   0.2451
## REGION.VIII   -9.745e-02  3.759e-01  -0.259   0.7957
## REGION.X      -1.761e-02  3.615e-01  -0.049   0.9612
## REGION.XI      3.921e-01  3.728e-01   1.052   0.2941
## REGION.XII     4.051e-03  3.825e-01   0.011   0.9916
## conn           1.999e-06  6.516e-06   0.307   0.7593
## cities        -2.091e-01  1.033e-01  -2.025   0.0441 *
## Mun1.1        -7.470e-02  7.588e-02  -0.984   0.3259
## gw             3.241e-03  2.287e-03   1.417   0.1579
## sprw           3.724e-02  1.635e-02   2.277   0.0237 *
## surw          -6.371e-02  6.780e-02  -0.940   0.3484
## elevar         4.307e-08  5.982e-07   0.072   0.9427
## coastal       -2.744e-02  7.081e-02  -0.388   0.6987
## emp           -1.768e-03  1.378e-03  -1.283   0.2008
## conn_log       4.817e-02  4.158e-02   1.159   0.2479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4615 on 225 degrees of freedom
## Multiple R-squared:  0.2262, Adjusted R-squared:  0.1436
## F-statistic:  2.74 on 24 and 225 DF,  p-value: 5.557e-05
```

```
# Evaluate logistic regression model AUC
predict_probabilities <- predict(full_model_classification, df_dummies_test)
roc(df_dummies_test$nrwpcent_class, predict_probabilities)$auc[1]
```

```
## [1] 0.5378422
```

The full model has an AUC of 0.5378422.

We employ recursive feature selection:

```
initial_model_classification <- glm(nrwpcent_class ~ 1, data = df_dummies_train, family = "binomial")

back_elim_classification <- step(object = full_model_classification,
                                 scope = list(lower = initial_model), direction = "backward")
```

```
## Start:  AIC=-362.99
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.IX + REGION.V + REGION.VI +
##     REGION.VII + REGION.VIII + REGION.X + REGION.XI + REGION.XII +
##     conn + cities + Mun1.1 + gw + sprw + surw + elevar + coastal +
##     emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - REGION.XII     1   0.00002 47.918 -364.99
## - REGION.V       1   0.00036 47.918 -364.99
## - REGION.X       1   0.00051 47.919 -364.99
## - elevar         1   0.00110 47.919 -364.99
## - REGION.IX      1   0.00312 47.921 -364.98
## - REGION.CARAGA  1   0.00661 47.925 -364.96
```

```
## - REGION.VI      1   0.01286 47.931 -364.92
## - REGION.II      1   0.01332 47.931 -364.92
## - REGION.I       1   0.01391 47.932 -364.92
## - REGION.VIII    1   0.01431 47.932 -364.92
## - conn           1   0.02004 47.938 -364.89
## - REGION.IV      1   0.02690 47.945 -364.85
## - coastal        1   0.03198 47.950 -364.83
## - REGION.CAR     1   0.13629 48.054 -364.28
## - surw           1   0.18803 48.106 -364.01
## - Mun1.1         1   0.20641 48.125 -363.92
## - REGION.XI      1   0.23552 48.154 -363.77
## - conn_log       1   0.28586 48.204 -363.50
## - REGION.VII     1   0.28928 48.207 -363.49
## - emp            1   0.35054 48.269 -363.17
## <none>                       47.918 -362.99
## - gw             1   0.42760 48.346 -362.77
## - REGION.III     1   0.51299 48.431 -362.33
## - cities         1   0.87315 48.791 -360.48
## - sprw           1   1.10453 49.023 -359.29
##
## Step:  AIC=-364.99
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.IX + REGION.V + REGION.VI +
##     REGION.VII + REGION.VIII + REGION.X + REGION.XI + conn +
##     cities + Mun1.1 + gw + sprw + surw + elevar + coastal + emp +
##     conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - REGION.V       1   0.00074 47.919 -366.99
## - elevar         1   0.00115 47.919 -366.99
## - REGION.X       1   0.00185 47.920 -366.98
## - REGION.IX      1   0.00716 47.925 -366.95
## - conn           1   0.02009 47.938 -366.89
## - REGION.CARAGA  1   0.02016 47.938 -366.89
## - coastal        1   0.03223 47.950 -366.82
## - REGION.VIII    1   0.03644 47.955 -366.80
## - REGION.II      1   0.04017 47.958 -366.78
## - REGION.VI      1   0.04322 47.961 -366.77
## - REGION.I       1   0.04723 47.965 -366.75
## - REGION.IV      1   0.09699 48.015 -366.49
## - surw           1   0.18806 48.106 -366.01
## - Mun1.1         1   0.21196 48.130 -365.89
## - REGION.CAR     1   0.28146 48.200 -365.53
## - conn_log       1   0.28906 48.207 -365.49
## - emp            1   0.35065 48.269 -365.17
## <none>                       47.918 -364.99
## - gw             1   0.42780 48.346 -364.77
## - REGION.XI      1   0.63319 48.551 -363.71
## - REGION.VII     1   0.72142 48.640 -363.26
## - cities         1   0.89683 48.815 -362.36
## - sprw           1   1.11482 49.033 -361.24
## - REGION.III     1   1.93976 49.858 -357.07
##
## Step:  AIC=-366.99
```

```
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.IX + REGION.VI + REGION.VII +
##     REGION.VIII + REGION.X + REGION.XI + conn + cities + Mun1.1 +
##     gw + sprw + surw + elevar + coastal + emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - elevar         1    0.0009 47.920 -368.98
## - REGION.X       1    0.0059 47.925 -368.96
## - REGION.IX      1    0.0123 47.931 -368.92
## - conn           1    0.0200 47.939 -368.88
## - coastal        1    0.0316 47.950 -368.82
## - REGION.CARAGA  1    0.0427 47.962 -368.76
## - REGION.II      1    0.0574 47.976 -368.69
## - REGION.VIII    1    0.0659 47.985 -368.64
## - REGION.VI      1    0.0749 47.994 -368.60
## - REGION.I       1    0.0789 47.998 -368.58
## - surw           1    0.1875 48.106 -368.01
## - Mun1.1         1    0.2123 48.131 -367.88
## - REGION.IV      1    0.2124 48.131 -367.88
## - conn_log       1    0.2899 48.209 -367.48
## - REGION.CAR     1    0.3336 48.252 -367.25
## - emp            1    0.3506 48.269 -367.17
## <none>                       47.919 -366.99
## - gw             1    0.4294 48.348 -366.76
## - cities         1    0.9031 48.822 -364.32
## - REGION.XI      1    0.9179 48.837 -364.24
## - REGION.VII     1    1.0739 48.993 -363.45
## - sprw           1    1.1577 49.077 -363.02
## - REGION.III     1    3.9857 51.905 -349.01
##
## Step:  AIC=-368.98
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.IX + REGION.VI + REGION.VII +
##     REGION.VIII + REGION.X + REGION.XI + conn + cities + Mun1.1 +
##     gw + sprw + surw + coastal + emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - REGION.X       1    0.0056 47.925 -370.95
## - REGION.IX      1    0.0130 47.933 -370.92
## - conn           1    0.0200 47.940 -370.88
## - coastal        1    0.0314 47.951 -370.82
## - REGION.CARAGA  1    0.0433 47.963 -370.76
## - REGION.II      1    0.0570 47.977 -370.69
## - REGION.VIII    1    0.0672 47.987 -370.63
## - REGION.VI      1    0.0754 47.995 -370.59
## - REGION.I       1    0.0780 47.998 -370.58
## - surw           1    0.1876 48.107 -370.01
## - REGION.IV      1    0.2116 48.131 -369.88
## - Mun1.1         1    0.2116 48.131 -369.88
## - conn_log       1    0.2893 48.209 -369.48
## - REGION.CAR     1    0.3362 48.256 -369.24
## - emp            1    0.3507 48.271 -369.16
## <none>                       47.920 -368.98
## - gw             1    0.4290 48.349 -368.75
```

```
## - cities         1    0.9145 48.834 -366.26
## - REGION.XI      1    0.9296 48.849 -366.18
## - REGION.VII     1    1.0742 48.994 -365.44
## - sprw           1    1.1652 49.085 -364.98
## - REGION.III     1    3.9850 51.905 -351.01
##
## Step:  AIC=-370.95
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.IX + REGION.VI + REGION.VII +
##     REGION.VIII + REGION.XI + conn + cities + Mun1.1 + gw + sprw +
##     surw + coastal + emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - REGION.IX      1    0.0110 47.936 -372.90
## - conn           1    0.0202 47.946 -372.85
## - coastal        1    0.0329 47.958 -372.78
## - REGION.CARAGA  1    0.0383 47.964 -372.75
## - REGION.VIII    1    0.0619 47.987 -372.63
## - REGION.II      1    0.0692 47.995 -372.59
## - REGION.VI      1    0.0979 48.023 -372.44
## - REGION.I       1    0.0983 48.024 -372.44
## - surw           1    0.1825 48.108 -372.00
## - Mun1.1         1    0.2061 48.131 -371.88
## - REGION.IV      1    0.2704 48.196 -371.55
## - conn_log       1    0.2856 48.211 -371.47
## - emp            1    0.3547 48.280 -371.11
## - REGION.CAR     1    0.3558 48.281 -371.11
## <none>                       47.925 -370.95
## - gw             1    0.4331 48.358 -370.71
## - cities         1    0.9100 48.835 -368.25
## - REGION.XI      1    1.0089 48.934 -367.75
## - sprw           1    1.1629 49.088 -366.96
## - REGION.VII     1    1.1740 49.099 -366.90
## - REGION.III     1    4.4435 52.369 -350.79
##
## Step:  AIC=-372.9
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.VI + REGION.VII + REGION.VIII +
##     REGION.XI + conn + cities + Mun1.1 + gw + sprw + surw + coastal +
##     emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - conn           1    0.0196 47.956 -374.79
## - coastal        1    0.0317 47.968 -374.73
## - REGION.CARAGA  1    0.0334 47.970 -374.72
## - REGION.VIII    1    0.0566 47.993 -374.60
## - REGION.II      1    0.0812 48.018 -374.47
## - REGION.VI      1    0.1169 48.053 -374.29
## - REGION.I       1    0.1172 48.054 -374.29
## - Mun1.1         1    0.2016 48.138 -373.85
## - surw           1    0.2023 48.139 -373.84
## - conn_log       1    0.2880 48.224 -373.40
## - REGION.IV      1    0.3088 48.245 -373.29
## - emp            1    0.3484 48.285 -373.09
```

```
## - REGION.CAR     1    0.3693 48.306 -372.98
## <none>                       47.936 -372.90
## - gw             1    0.4264 48.363 -372.68
## - cities         1    0.9318 48.868 -370.08
## - REGION.XI      1    1.0584 48.995 -369.44
## - sprw           1    1.2052 49.142 -368.69
## - REGION.VII     1    1.2241 49.160 -368.59
## - REGION.III     1    4.7434 52.680 -351.31
##
## Step:  AIC=-374.79
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.VI + REGION.VII + REGION.VIII +
##     REGION.XI + cities + Mun1.1 + gw + sprw + surw + coastal +
##     emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - coastal        1    0.0285 47.984 -376.65
## - REGION.CARAGA  1    0.0334 47.989 -376.62
## - REGION.VIII    1    0.0603 48.016 -376.48
## - REGION.II      1    0.0846 48.041 -376.35
## - REGION.VI      1    0.1110 48.067 -376.22
## - REGION.I       1    0.1230 48.079 -376.15
## - surw           1    0.1834 48.139 -375.84
## - Mun1.1         1    0.2074 48.163 -375.72
## - REGION.IV      1    0.2987 48.255 -375.24
## - conn_log       1    0.3005 48.256 -375.23
## - REGION.CAR     1    0.3703 48.326 -374.87
## <none>                       47.956 -374.79
## - gw             1    0.4709 48.427 -374.35
## - emp            1    0.9740 48.930 -371.77
## - cities         1    0.9871 48.943 -371.70
## - REGION.XI      1    1.1540 49.110 -370.85
## - sprw           1    1.1885 49.145 -370.67
## - REGION.VII     1    1.2116 49.168 -370.56
## - REGION.III     1    4.9187 52.875 -352.38
##
## Step:  AIC=-376.65
## nrwpcent_class ~ REGION.CAR + REGION.CARAGA + REGION.I + REGION.II +
##     REGION.III + REGION.IV + REGION.VI + REGION.VII + REGION.VIII +
##     REGION.XI + cities + Mun1.1 + gw + sprw + surw + emp + conn_log
##
##                 Df Sum of Sq    RSS     AIC
## - REGION.CARAGA  1    0.0308 48.015 -378.49
## - REGION.VIII    1    0.0674 48.052 -378.30
## - REGION.II      1    0.1070 48.091 -378.09
## - REGION.VI      1    0.1137 48.098 -378.05
## - REGION.I       1    0.1653 48.150 -377.79
## - Mun1.1         1    0.2059 48.190 -377.58
## - surw           1    0.2076 48.192 -377.57
## - REGION.IV      1    0.3242 48.309 -376.96
## - conn_log       1    0.3277 48.312 -376.94
## <none>                       47.984 -376.65
## - REGION.CAR     1    0.4265 48.411 -376.43
## - gw             1    0.4865 48.471 -376.12
```

```
## - emp            1    0.9771 48.962 -373.61
## - cities         1    1.0177 49.002 -373.40
## - REGION.XI      1    1.1364 49.121 -372.79
## - sprw           1    1.1796 49.164 -372.57
## - REGION.VII     1    1.1871 49.172 -372.54
## - REGION.III     1    5.4627 53.447 -351.69
##
## Step:  AIC=-378.49
## nrwpcent_class ~ REGION.CAR + REGION.I + REGION.II + REGION.III +
##     REGION.IV + REGION.VI + REGION.VII + REGION.VIII + REGION.XI +
##     cities + Mun1.1 + gw + sprw + surw + emp + conn_log
##
##               Df Sum of Sq    RSS     AIC
## - REGION.VIII  1    0.0558 48.071 -380.20
## - REGION.II    1    0.1409 48.156 -379.75
## - REGION.VI    1    0.1610 48.176 -379.65
## - Mun1.1       1    0.1898 48.205 -379.50
## - surw         1    0.2078 48.223 -379.41
## - REGION.I     1    0.2231 48.238 -379.33
## - conn_log     1    0.3235 48.339 -378.81
## <none>                     48.015 -378.49
## - REGION.IV    1    0.4380 48.453 -378.22
## - REGION.CAR   1    0.4676 48.483 -378.06
## - gw           1    0.4881 48.503 -377.96
## - emp          1    0.9847 49.000 -375.41
## - cities       1    0.9872 49.002 -375.40
## - sprw         1    1.2128 49.228 -374.25
## - REGION.XI    1    1.2397 49.255 -374.11
## - REGION.VII   1    1.2629 49.278 -374.00
## - REGION.III   1    6.2973 54.313 -349.68
##
## Step:  AIC=-380.2
## nrwpcent_class ~ REGION.CAR + REGION.I + REGION.II + REGION.III +
##     REGION.IV + REGION.VI + REGION.VII + REGION.XI + cities +
##     Mun1.1 + gw + sprw + surw + emp + conn_log
##
##              Df Sum of Sq    RSS     AIC
## - REGION.II   1    0.1689 48.240 -381.32
## - Mun1.1      1    0.1865 48.257 -381.23
## - REGION.VI   1    0.2039 48.275 -381.14
## - surw        1    0.2119 48.283 -381.10
## - REGION.I    1    0.2715 48.343 -380.79
## - conn_log    1    0.3354 48.406 -380.46
## <none>                    48.071 -380.20
## - gw          1    0.4832 48.554 -379.69
## - REGION.CAR  1    0.4930 48.564 -379.64
## - REGION.IV   1    0.5206 48.592 -379.50
## - emp         1    0.9767 49.048 -377.17
## - cities      1    1.0331 49.104 -376.88
## - sprw        1    1.1826 49.254 -376.12
## - REGION.XI   1    1.3206 49.392 -375.42
## - REGION.VII  1    1.3572 49.428 -375.23
## - REGION.III  1    6.6454 54.716 -349.82
##
```

```
## Step:  AIC=-381.32
## nrwpcent_class ~ REGION.CAR + REGION.I + REGION.III + REGION.IV +
##     REGION.VI + REGION.VII + REGION.XI + cities + Mun1.1 + gw +
##     sprw + surw + emp + conn_log
##
##              Df Sum of Sq    RSS     AIC
## - REGION.VI   1    0.1209 48.361 -382.69
## - Mun1.1      1    0.1625 48.402 -382.48
## - REGION.I    1    0.1759 48.416 -382.41
## - surw        1    0.2402 48.480 -382.08
## - conn_log    1    0.3046 48.545 -381.74
## <none>                    48.240 -381.32
## - REGION.IV   1    0.3879 48.628 -381.32
## - REGION.CAR  1    0.4314 48.671 -381.09
## - gw          1    0.5505 48.790 -380.48
## - cities      1    0.9848 49.225 -378.27
## - emp         1    0.9991 49.239 -378.19
## - sprw        1    1.0553 49.295 -377.91
## - REGION.XI   1    1.1945 49.434 -377.20
## - REGION.VII  1    1.2484 49.488 -376.93
## - REGION.III  1    6.6397 54.880 -351.08
##
## Step:  AIC=-382.69
## nrwpcent_class ~ REGION.CAR + REGION.I + REGION.III + REGION.IV +
##     REGION.VII + REGION.XI + cities + Mun1.1 + gw + sprw + surw +
##     emp + conn_log
##
##              Df Sum of Sq    RSS     AIC
## - REGION.I    1    0.1102 48.471 -384.12
## - Mun1.1      1    0.1622 48.523 -383.86
## - conn_log    1    0.2438 48.605 -383.44
## - surw        1    0.2678 48.629 -383.31
## - REGION.IV   1    0.2963 48.657 -383.17
## - REGION.CAR  1    0.3862 48.747 -382.70
## <none>                    48.361 -382.69
## - gw          1    0.5597 48.921 -381.82
## - cities      1    0.9250 49.286 -379.96
## - emp         1    0.9398 49.301 -379.88
## - sprw        1    0.9734 49.334 -379.71
## - REGION.XI   1    1.0910 49.452 -379.12
## - REGION.VII  1    1.1597 49.520 -378.77
## - REGION.III  1    6.6203 54.981 -352.62
##
## Step:  AIC=-384.12
## nrwpcent_class ~ REGION.CAR + REGION.III + REGION.IV + REGION.VII +
##     REGION.XI + cities + Mun1.1 + gw + sprw + surw + emp + conn_log
##
##              Df Sum of Sq    RSS     AIC
## - Mun1.1      1    0.1546 48.626 -385.33
## - REGION.IV   1    0.2259 48.697 -384.96
## - conn_log    1    0.2526 48.724 -384.82
## - surw        1    0.2906 48.762 -384.63
## - REGION.CAR  1    0.3528 48.824 -384.31
## <none>                    48.471 -384.12
```

```
## - gw          1    0.5832 49.054 -383.13
## - sprw        1    0.9030 49.374 -381.51
## - cities      1    0.9039 49.375 -381.50
## - emp         1    0.9859 49.457 -381.09
## - REGION.XI   1    1.0223 49.493 -380.91
## - REGION.VII  1    1.0932 49.564 -380.55
## - REGION.III  1    6.6218 55.093 -354.11
##
## Step:  AIC=-385.33
## nrwpcent_class ~ REGION.CAR + REGION.III + REGION.IV + REGION.VII +
##     REGION.XI + cities + gw + sprw + surw + emp + conn_log
##
##              Df Sum of Sq    RSS     AIC
## - conn_log    1    0.1507 48.776 -386.55
## - REGION.IV   1    0.2240 48.850 -386.18
## - surw        1    0.2801 48.906 -385.89
## - REGION.CAR  1    0.3459 48.972 -385.56
## <none>                    48.626 -385.33
## - gw          1    0.5588 49.184 -384.47
## - cities      1    0.7510 49.377 -383.50
## - sprw        1    0.8822 49.508 -382.83
## - emp         1    0.8903 49.516 -382.79
## - REGION.XI   1    0.9976 49.623 -382.25
## - REGION.VII  1    1.0898 49.715 -381.79
## - REGION.III  1    6.5628 55.188 -355.68
##
## Step:  AIC=-386.55
## nrwpcent_class ~ REGION.CAR + REGION.III + REGION.IV + REGION.VII +
##     REGION.XI + cities + gw + sprw + surw + emp
##
##              Df Sum of Sq    RSS     AIC
## - surw        1    0.2782 49.055 -387.13
## - REGION.IV   1    0.2858 49.062 -387.09
## - REGION.CAR  1    0.3664 49.143 -386.68
## <none>                    48.776 -386.55
## - cities      1    0.6156 49.392 -385.42
## - gw          1    0.6312 49.408 -385.34
## - emp         1    0.7568 49.533 -384.70
## - REGION.XI   1    0.9310 49.707 -383.83
## - sprw        1    1.0151 49.791 -383.40
## - REGION.VII  1    1.1191 49.895 -382.88
## - REGION.III  1    8.1403 56.917 -349.97
##
## Step:  AIC=-387.13
## nrwpcent_class ~ REGION.CAR + REGION.III + REGION.IV + REGION.VII +
##     REGION.XI + cities + gw + sprw + emp
##
##              Df Sum of Sq    RSS     AIC
## - REGION.IV   1    0.2797 49.334 -387.71
## - REGION.CAR  1    0.3822 49.437 -387.19
## <none>                    49.055 -387.13
## - cities      1    0.6538 49.708 -385.82
## - gw          1    0.7394 49.794 -385.39
## - REGION.XI   1    0.9141 49.969 -384.52
```

```
## - sprw        1    0.9267 49.981 -384.45
## - emp         1    0.9866 50.041 -384.15
## - REGION.VII  1    1.1451 50.200 -383.36
## - REGION.III  1    8.0547 57.109 -351.12
##
## Step:  AIC=-387.71
## nrwpcent_class ~ REGION.CAR + REGION.III + REGION.VII + REGION.XI +
##     cities + gw + sprw + emp
##
##              Df Sum of Sq    RSS     AIC
## - REGION.CAR  1    0.3247 49.659 -388.07
## <none>                    49.334 -387.71
## - cities      1    0.7894 50.124 -385.74
## - REGION.XI   1    0.7899 50.124 -385.74
## - sprw        1    0.9237 50.258 -385.07
## - emp         1    0.9373 50.271 -385.01
## - gw          1    0.9633 50.298 -384.88
## - REGION.VII  1    1.0541 50.388 -384.43
## - REGION.III  1    7.8694 57.204 -352.71
##
## Step:  AIC=-388.07
## nrwpcent_class ~ REGION.III + REGION.VII + REGION.XI + cities +
##     gw + sprw + emp
##
##              Df Sum of Sq    RSS     AIC
## <none>                    49.659 -388.07
## - REGION.XI   1    0.7631 50.422 -386.26
## - cities      1    0.8721 50.531 -385.72
## - emp         1    0.8991 50.558 -385.58
## - sprw        1    0.9036 50.563 -385.56
## - gw          1    0.9313 50.590 -385.43
## - REGION.VII  1    1.0421 50.701 -384.88
## - REGION.III  1    7.6838 57.343 -354.10
```

```
forward_elim_classfication <- step(object = initial_model_classification,
                                   scope = list(upper = full_model_classification), direction = "forwar
```

```
## Start:  AIC=346.27
## nrwpcent_class ~ 1
##
##                 Df Deviance    AIC
## + REGION.III     1   309.27 313.27
## + cities         1   336.45 340.45
## + coastal        1   339.63 343.63
## + emp            1   341.43 345.43
## + Mun1.1         1   341.52 345.52
## + surw           1   341.78 345.78
## + REGION.VI      1   341.89 345.89
## + REGION.VIII    1   342.24 346.24
## + REGION.XII     1   342.24 346.24
## <none>               344.27 346.27
## + gw             1   342.64 346.64
## + REGION.CARAGA  1   342.69 346.69
## + REGION.IX      1   342.77 346.77
## + REGION.VII     1   342.83 346.83
```

```
## + REGION.V       1   342.96 346.96
## + conn           1   343.28 347.28
## + REGION.X       1   343.35 347.35
## + REGION.I       1   343.38 347.38
## + REGION.CAR     1   343.56 347.56
## + REGION.XI      1   343.72 347.72
## + REGION.II      1   343.84 347.84
## + sprw           1   343.88 347.88
## + conn_log       1   344.14 348.14
## + REGION.IV      1   344.18 348.18
## + elevar         1   344.26 348.26
##
## Step:  AIC=313.27
## nrwpcent_class ~ REGION.III
##
##                 Df Deviance    AIC
## + cities         1   304.12 310.12
## + emp            1   305.39 311.39
## + sprw           1   306.11 312.11
## + surw           1   306.27 312.27
## + conn           1   306.28 312.28
## + REGION.VII     1   306.59 312.59
## <none>               309.27 313.27
## + REGION.XI      1   307.81 313.81
## + REGION.CAR     1   307.96 313.96
## + REGION.VIII    1   308.23 314.23
## + REGION.XII     1   308.23 314.23
## + REGION.IV      1   308.35 314.35
## + REGION.IX      1   308.43 314.43
## + conn_log       1   308.68 314.68
## + gw             1   308.69 314.69
## + coastal        1   308.71 314.71
## + REGION.CARAGA  1   308.72 314.72
## + Mun1.1         1   308.73 314.73
## + REGION.VI      1   308.82 314.82
## + REGION.X       1   309.06 315.06
## + REGION.V       1   309.13 315.13
## + REGION.I       1   309.26 315.26
## + elevar         1   309.27 315.27
## + REGION.II      1   309.27 315.27
##
## Step:  AIC=310.12
## nrwpcent_class ~ REGION.III + cities
##
##                 Df Deviance    AIC
## + REGION.VII     1   299.06 307.06
## + sprw           1   300.02 308.02
## + REGION.XI      1   302.01 310.01
## + surw           1   302.10 310.10
## <none>               304.12 310.12
## + emp            1   302.92 310.92
## + gw             1   302.94 310.94
## + REGION.CARAGA  1   303.02 311.02
## + REGION.CAR     1   303.17 311.17
```

```
## + conn           1   303.25 311.25
## + REGION.VIII    1   303.37 311.37
## + REGION.XII     1   303.37 311.37
## + REGION.IX      1   303.54 311.54
## + REGION.IV      1   303.63 311.63
## + REGION.VI      1   303.73 311.73
## + REGION.V       1   303.85 311.85
## + coastal        1   303.88 311.88
## + conn_log       1   303.93 311.93
## + REGION.X       1   304.03 312.03
## + elevar         1   304.07 312.07
## + Mun1.1         1   304.11 312.11
## + REGION.II      1   304.11 312.11
## + REGION.I       1   304.11 312.11
##
## Step:  AIC=307.06
## nrwpcent_class ~ REGION.III + cities + REGION.VII
##
##                 Df Deviance    AIC
## + sprw           1   295.42 305.42
## + REGION.XI      1   296.43 306.43
## <none>               299.06 307.06
## + surw           1   297.18 307.18
## + gw             1   297.59 307.59
## + REGION.CARAGA  1   298.03 308.03
## + REGION.CAR     1   298.07 308.07
## + emp            1   298.19 308.19
## + REGION.IV      1   298.24 308.24
## + coastal        1   298.41 308.41
## + REGION.VIII    1   298.50 308.50
## + REGION.XII     1   298.50 308.50
## + conn           1   298.52 308.52
## + REGION.IX      1   298.62 308.62
## + conn_log       1   298.80 308.80
## + REGION.VI      1   298.87 308.87
## + REGION.V       1   298.91 308.91
## + elevar         1   298.95 308.95
## + REGION.X       1   299.03 309.03
## + Mun1.1         1   299.04 309.04
## + REGION.I       1   299.05 309.05
## + REGION.II      1   299.06 309.06
##
## Step:  AIC=305.42
## nrwpcent_class ~ REGION.III + cities + REGION.VII + sprw
##
##                 Df Deviance    AIC
## + REGION.XI      1   292.33 304.33
## + surw           1   292.88 304.88
## <none>               295.42 305.42
## + gw             1   293.85 305.85
## + REGION.CARAGA  1   294.27 306.27
## + REGION.VIII    1   294.31 306.31
## + REGION.CAR     1   294.37 306.37
## + emp            1   294.43 306.43
```

```
## + coastal        1   294.45 306.45
## + REGION.IV      1   294.66 306.66
## + REGION.V       1   294.74 306.74
## + conn           1   294.87 306.87
## + REGION.X       1   295.07 307.07
## + REGION.IX      1   295.07 307.07
## + REGION.XII     1   295.09 307.09
## + REGION.I       1   295.29 307.29
## + conn_log       1   295.36 307.36
## + Mun1.1         1   295.36 307.36
## + REGION.II      1   295.37 307.37
## + elevar         1   295.39 307.39
## + REGION.VI      1   295.39 307.39
##
## Step:  AIC=304.33
## nrwpcent_class ~ REGION.III + cities + REGION.VII + sprw + REGION.XI
##
##                 Df Deviance    AIC
## + surw           1   289.55 303.55
## <none>               292.33 304.33
## + coastal        1   290.81 304.81
## + emp            1   290.97 304.97
## + gw             1   291.07 305.07
## + REGION.IV      1   291.15 305.15
## + REGION.CAR     1   291.20 305.20
## + conn           1   291.29 305.29
## + REGION.CARAGA  1   291.30 305.30
## + REGION.VIII    1   291.36 305.36
## + REGION.V       1   291.81 305.81
## + REGION.I       1   292.04 306.04
## + REGION.IX      1   292.06 306.06
## + REGION.X       1   292.07 306.07
## + REGION.XII     1   292.10 306.10
## + REGION.II      1   292.20 306.20
## + conn_log       1   292.25 306.25
## + Mun1.1         1   292.27 306.27
## + elevar         1   292.33 306.33
## + REGION.VI      1   292.33 306.33
##
## Step:  AIC=303.55
## nrwpcent_class ~ REGION.III + cities + REGION.VII + sprw + REGION.XI +
##     surw
##
##                 Df Deviance    AIC
## <none>               289.55 303.55
## + gw             1   288.24 304.24
## + REGION.IV      1   288.32 304.32
## + REGION.CAR     1   288.48 304.48
## + coastal        1   288.53 304.53
## + REGION.CARAGA  1   288.71 304.71
## + REGION.VIII    1   288.72 304.72
## + emp            1   288.84 304.84
## + REGION.X       1   289.14 305.14
## + conn           1   289.21 305.21
```

```
## + REGION.V       1   289.24 305.24
## + conn_log       1   289.32 305.32
## + REGION.XII     1   289.33 305.33
## + REGION.I       1   289.37 305.37
## + REGION.IX      1   289.40 305.40
## + Mun1.1         1   289.46 305.46
## + REGION.II      1   289.47 305.47
## + elevar         1   289.54 305.54
## + REGION.VI      1   289.54 305.54
```

```
stepwise_classification <- step(object = initial_model_classification,
                                scope = list(upper = initial_model), direction = "both")
```

```
## Start:  AIC=346.27
## nrwpcent_class ~ 1
```

```
fitstat_classification <- function(x) {
  predict_probabilities <- predict(x, df_dummies_test)
  return(c(AUC = roc(df_dummies_test$nrwpcent_class, predict_probabilities)$auc[1]
  ))}

sapply(list('Full Model' = full_model_classification,
            'Forward Elimination' = forward_elim_classfication,
            'Backward Elimination' = back_elim_classification,
            'Stepwise Selection' = stepwise_classification), fitstat_classification)
```

```
##           Full Model.AUC  Forward Elimination.AUC Backward Elimination.AUC
##                0.5378422                0.5450886                0.5136876
##   Stepwise Selection.AUC
##                0.5000000
```

```
summary(forward_elim_classfication)
```

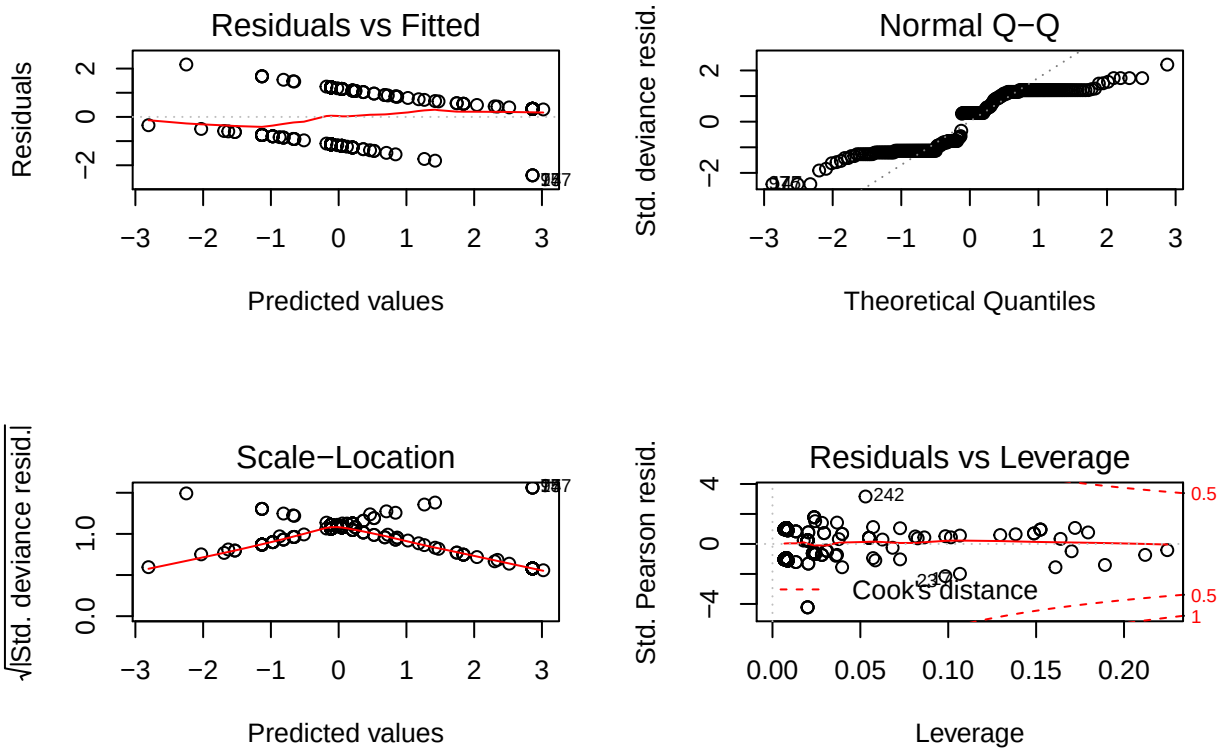
```
##
## Call:
## glm(formula = nrwpcent_class ~ REGION.III + cities + REGION.VII +
##     sprw + REGION.XI + surw, family = "binomial", data = df_dummies_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4152  -1.1299   0.3336   1.1580   2.1675
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.11276    0.17759  -0.635  0.52546
## REGION.III   2.97366    0.64049   4.643 3.44e-06 ***
## cities      -1.01907    0.36799  -2.769  0.00562 **
## REGION.VII   1.83608    0.88856   2.066  0.03879 *
## sprw         0.15866    0.07799   2.034  0.04190 *
## REGION.XI    1.37718    0.79124   1.741  0.08176 .
## surw        -0.55839    0.37290  -1.497  0.13428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 344.27  on 249  degrees of freedom
## Residual deviance: 289.55  on 243  degrees of freedom
## AIC: 303.55
##
## Number of Fisher Scoring iterations: 5
```

```
forward_elim_classfication %>% vif
```

```
## REGION.III     cities REGION.VII       sprw  REGION.XI       surw
##   1.043088   1.160175   1.090258   1.063917   1.045833   1.051823
```

```
par(mfrow=c(2,2))
plot(forward_elim_classfication)
```




```
predicted_probabilities <- forward_elim_classfication %>% predict(df_dummies_test)
predicted_probabilities[predicted_probabilities > 0.5] <- 1
predicted_probabilities[predicted_probabilities <= 0.5] <- 0
predicted_probabilities <- predicted_probabilities %>% as.vector() %>% as.factor
observed <- (df_dummies_test %>% mutate(nrwpcent_class=as.factor(nrwpcent_class)))$nrwpcent_class
confusionMatrix(data=predicted_probabilities, reference = observed)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 14 18
##          1  9  9
##
```

```
##                Accuracy : 0.46
##                  95% CI : (0.3181, 0.6068)
##     No Information Rate : 0.54
##     P-Value [Acc > NIR] : 0.8990
##
##                   Kappa : -0.0563
##  Mcnemar's Test P-Value : 0.1237
##
##             Sensitivity : 0.6087
##             Specificity : 0.3333
##          Pos Pred Value : 0.4375
##          Neg Pred Value : 0.5000
##              Prevalence : 0.4600
##          Detection Rate : 0.2800
##    Detection Prevalence : 0.6400
##       Balanced Accuracy : 0.4710
##
##        'Positive' Class : 0
##
```

Based on the above procedures, we select the model determined by our forward selection procedure since it has the highest test AUC of 0.54; however, it appears to only have 46% accuracy.