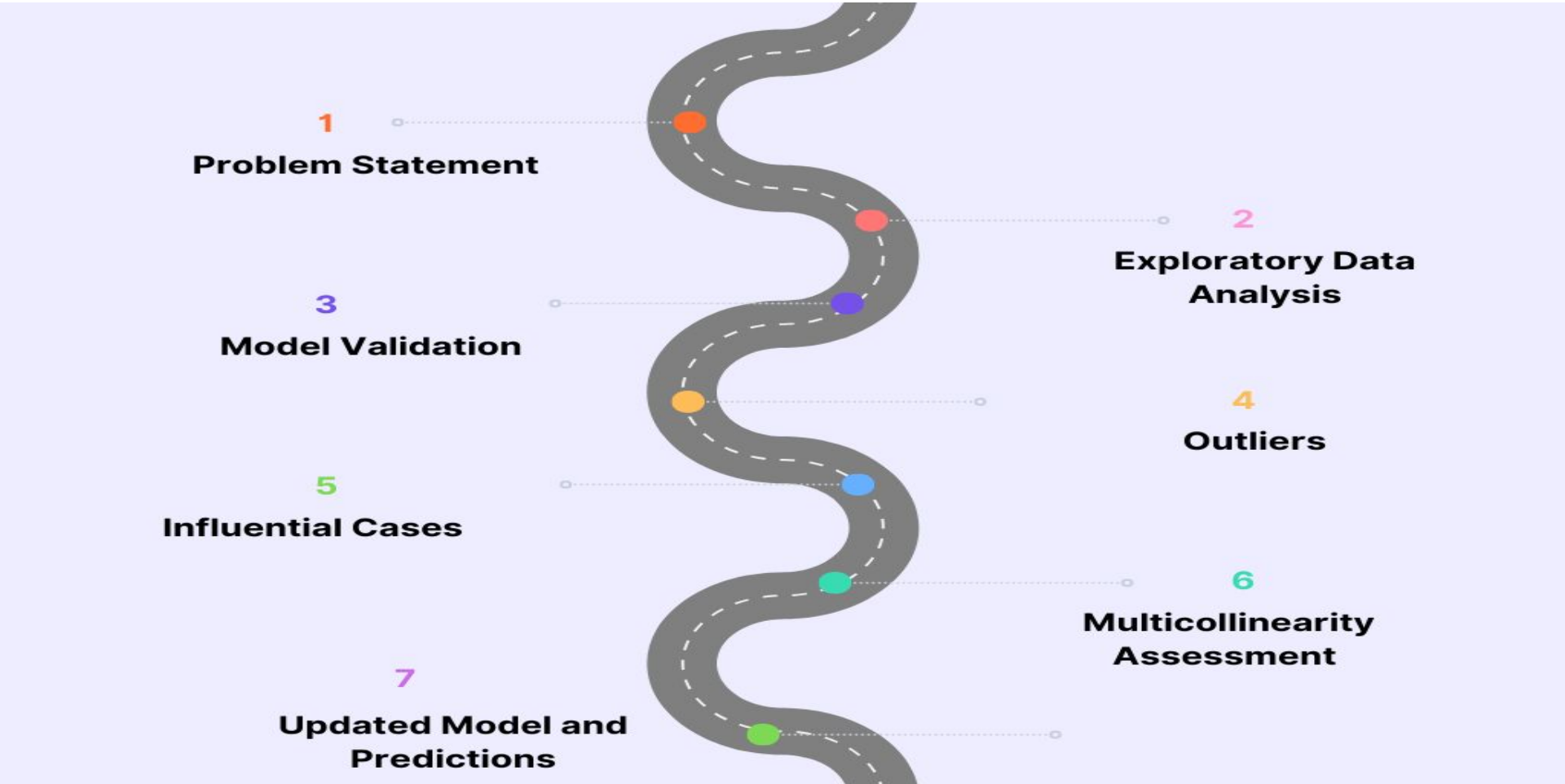
Case Study

Diagnostic Checks: Influential Outliers and Multicollinearity Assessment

Analysis Roadmap



Problem Statement and Objectives

Build a predictive model for website developer data that minimizes overfitting, addresses multicollinearity, and manages outliers to ensure reliable performance.

- Analyze the dataset and identify the best model for prediction
- Evaluate the impact of outliers and influential cases on the model’s performance
- Assess multicollinearity among predictors and implement corrections if required
- Validate the model’s performance using metrics such as Adjusted R Squared and BIC
- Address any overfitting concerns and give recommendations for improving the model

Exploratory Data analysis

In this analysis, the dependent variable (**Y**) is Websites Delivered. The independent variables (**X**) include Backlog of Orders, Team Number, Team Experience, Process Change, Year, and Quarter. These predictors are used to model and explain the variation in the number of websites delivered.

Identification number	Websites delivered	Backlog of orders	Team number	Team experience	Process change	Year	Quarter
1	1	12	1	3	0	2001	1
2	2	18	1	6	0	2001	2
3	7	26	1	9	0	2001	3
4	2	28	1	12	0	2001	4
5	1	36	1	15	0	2002	1
6	10	45	1	18	1	2002	2
7	10	36	1	21	1	2002	3
8	1	18	2	3	0	2001	1
9	1	25	2	6	0	2001	2
10	6	28	2	9	0	2001	3
11	5	28	2	12	0	2001	4
12	11	38	2	15	0	2002	1
13	15	38	2	18	1	2002	2
14	13	34	2	21	1	2002	3
15	3	18	3	3	0	2001	1
16	3	23	3	6	0	2001	2
17	13	29	3	9	0	2001	3
18	6	24	3	12	0	2001	4
19	2	32	3	15	0	2002	1
20	11	41	3	18	1	2002	2
21	3	33	3	21	1	2002	3
22	7	21	4	3	0	2001	1
23	12	22	4	6	0	2001	2
24	7	25	4	9	0	2001	3
25	1	27	4	12	0	2001	4

```
The Structure of the dataframe:
> str(df)
'data.frame':   73 obs. of  7 variables:
 $ Y : int  1 2 7 2 1 10 10 1 1 6 ...
 $ X1: int  12 18 26 28 36 45 36 18 25 28 ...
 $ X2: int  1 1 1 1 1 1 1 2 2 2 ...
 $ X3: int  3 6 9 12 15 18 21 3 6 9 ...
 $ X4: int  0 0 0 0 0 1 1 0 0 0 ...
 $ X5: int  2001 2001 2001 2001 2002 2002 2002 2001 2001 2001 ...
 $ X6: int  1 2 3 4 1 2 3 1 2 3 ...
```

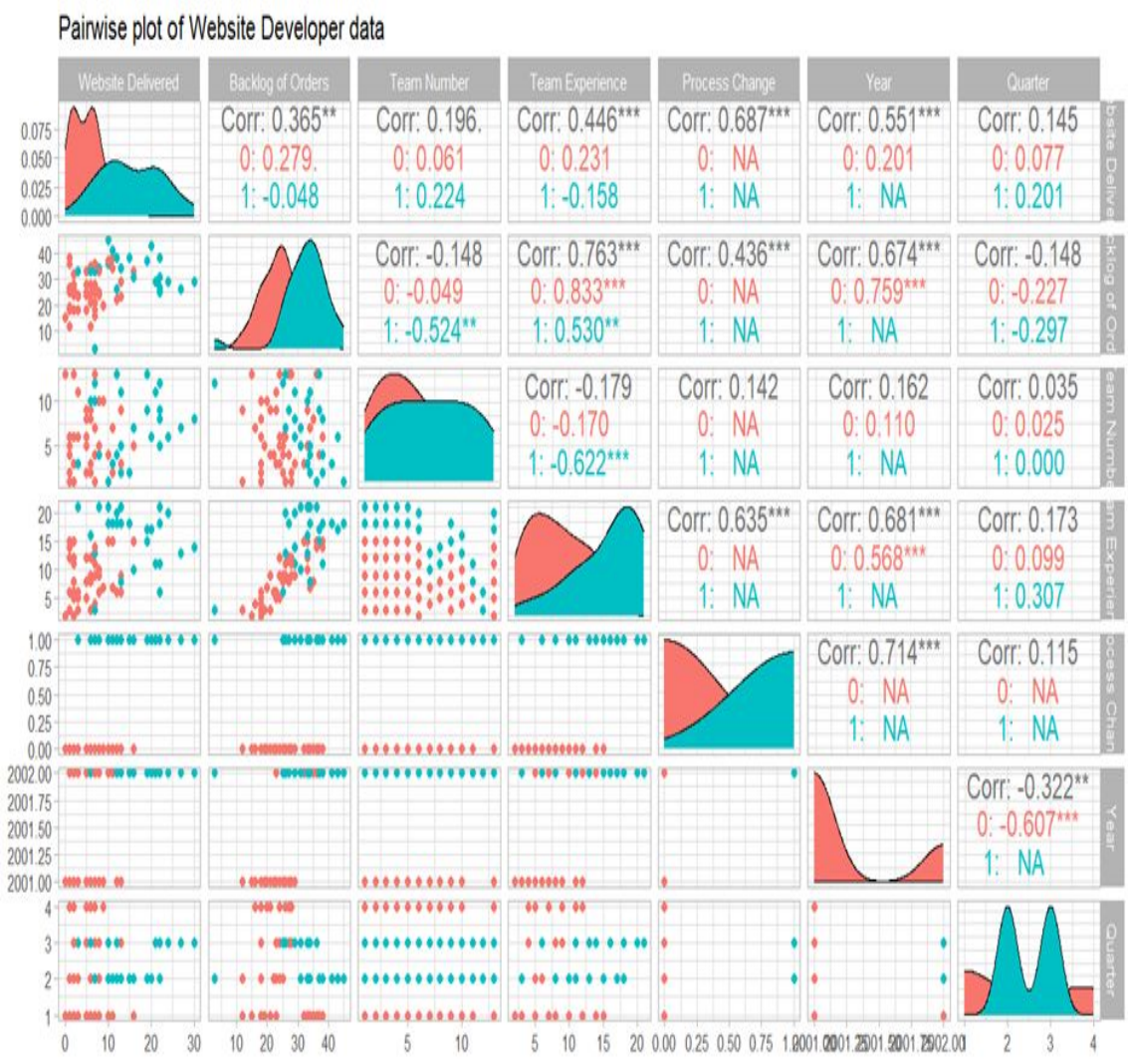
No. of Null values in each of the columns:

```
> colSums(is.na(df))
id  Y X1 X2 X3 X4 X5 X6
 0  0  0  0  0  0  0  0
```

[1] "The Summary of the dataframe:"

```
> summary(df)
       Y               X1              X2              X3              X4              X5              X6
 Min.   : 0.000   Min.   : 3.00   Min.   : 1.000   Min.   : 2.00   Min.   :0.0000   Min.   :2001   Min.   :1.000
 1st Qu.: 3.000   1st Qu.:23.00   1st Qu.: 3.000   1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:2001   1st Qu.:1.000
 Median : 7.000   Median :28.00   Median : 6.000   Median :11.00   Median :0.0000   Median :2002   Median :2.000
 Mean   : 9.041   Mean   :27.82   Mean   : 6.288   Mean   :10.85   Mean   :0.3562   Mean   :2002   Mean   :2.342
 3rd Qu.:13.000   3rd Qu.:34.00   3rd Qu.: 9.000   3rd Qu.:15.00   3rd Qu.:1.0000   3rd Qu.:2002   3rd Qu.:3.000
 Max.   :30.000   Max.   :45.00   Max.   :13.000   Max.   :21.00   Max.   :1.0000   Max.   :2002   Max.   :4.000
```

- Websites delivered ranges from 0 to 30 with an average of 9apprx.
- Backlog of orders (X1) varies between 3 and 45, with a median of 28. Team experience (X3) is between 2 to 21 months with an average of 11 months.
- Process change (X4) is a binary variable (0 or 1). Two years and 4 quarters are covered in this dataset (X5 and X6)
- There are no null values in any of the columns.



- Team Experience (0.446) and Process Change (0.687) show positive correlation with Websites Delivered, suggesting they may be important predictors.
- Team Experience and Process Change have a high correlation (0.635), indicating potential multicollinearity that could affect regression models.
- Year and Process Change are strongly correlated (0.714).
- Team Number shows little correlation with most variables which indicates it may not significantly impact Websites Delivered.

Data Pre-Processing

Standardization

- Process of re-scaling features so they have
- Mean = 0 and Standard Deviation = 1
- Ensures all the features contribute equally to the model

Encoding

- Converted categorical data into numerical data i.e. 1,0
- Created binary columns for each category by taking a reference column

Interaction and Quadratic terms

- Interaction terms represents the combined effect of two variables on the target variables (e.x.) X1 * X3
- Quadratic terms represents the squared effect of a variable on the target variable (e.x.) X1^2

Before Pre-processing

```
The Structure of the dataframe:
> str(df)
'data.frame':   73 obs. of  7 variables:
 $ Y : int  1 2 7 2 1 10 10 1 1 6 ...
 $ X1: int  12 18 26 28 36 45 36 18 25 28 ...
 $ X2: int  1 1 1 1 1 1 1 2 2 2 ...
 $ X3: int  3 6 9 12 15 18 21 3 6 9 ...
 $ X4: int  0 0 0 0 0 1 1 0 0 0 ...
 $ X5: int  2001 2001 2001 2001 2002 2002 2002 2001 2001 2001 ...
 $ X6: int  1 2 3 4 1 2 3 1 2 3 ...
```

After Pre-processing

```
structure of the dataframe with main effect, interaction and quadratic terms:
> str(df)
'data.frame':   73 obs. of  76 variables:
 $ Y          : int  1 2 7 2 1 10 10 1 1 6 ...
 $ X1         : num  -1.9836 -1.2314 -0.2284 0.0223 1.0253 ...
 $ X3         : num  -1.386 -0.856 -0.327 0.203 0.733 ...
 $ X21        : num  1 1 1 1 1 1 1 0 0 0 ...
 $ X22        : num  0 0 0 0 0 0 0 1 1 1 ...
 $ X23        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X24        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X25        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X26        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X27        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X28        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X29        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X210       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X211       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X212       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X213       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X41        : num  0 0 0 0 0 1 1 0 0 0 ...
 $ X52002     : num  0 0 0 0 1 1 1 0 0 0 ...
 $ X62        : num  0 1 0 0 0 1 0 0 1 0 ...
 $ X63        : num  0 0 1 0 0 0 1 0 0 1 ...
 $ X64        : num  0 0 0 1 0 0 0 0 0 0 ...
 $ X1_X3      : num  2.74925 1.05439 0.07459 0.00454 0.75144 ...
 $ X1_X21     : num  -1.9836 -1.2314 -0.2284 0.0223 1.0253 ...
 $ X1_X22     : num  0 0 0 0 0 ...
 $ X1_X23     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X1_X24     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ X1_X41     : num  0 0 0 0 0 ...
 $ X1_X52002  : num  0 0 0 0 1.03 ...
 $ X1_X62     : num  0 -1.23 0 0 0 ...
 $ X1_X63     : num  0 0 -0.228 0 0 ...
 $ X1_X64     : num  0 0 0 0.0223 0 ...
 $ X3_X21     : num  -1.386 -0.856 -0.327 0.203 0.733 ...
 $ X3_X22     : num  0 0 0 0 0 ...
 $ X3_X23     : num  0 0 0 0 0 0 0 0 0 0 ...
```

Model Selection

Full Model : Model got trained using all predictor values

Summary:

```
Residual standard error: 5.099 on 18 degrees of freedom
Multiple R-squared:  0.8705,    Adjusted R-squared:  0.4819
F-statistic:  2.24 on 54 and 18 DF,  p-value: 0.03086
```

- High R^2 suggests the predictors explain most of the variability in Y
- The Adjusted R^2 is much lower, indicating potential overfitting due to irrelevant predictors

Backward Elimination

```
#Backward Selection
b_model <- step(full_model, direction = "backward")
```

```
Step:  AIC=176
Y ~ X1 + X3 + X21 + X22 + X23 + X24 + X25 + X26 + X27 + X28 +
    X29 + X210 + X211 + X212 + X41 + X52002 + X62 + X63 + X1_X3 +
    X1_X21 + X1_X22 + X1_X23 + X1_X24 + X1_X41 + X1_X52002 +
    X1_X62 + X1_X63 + X1_X64 + X3_X21 + X3_X22 + X3_X23 + X3_X24 +
    X3_X41 + X3_X52002 + X3_X62 + X3_X63 + X3_X64 + X21_X41 +
    X21_X52002 + X21_X62 + X22_X41 + X22_X52002 + X22_X62 + X23_X41 +
    X23_X52002 + X23_X62 + X23_X63 + X24_X41 + X24_X52002 + X41_X62

           Df Sum of Sq     RSS    AIC
- X3        1     0.020 207.78 174.01
- X21_X41   1     0.022 207.78 174.01
- X22_X41   1     0.041 207.80 174.01
- X25       1     0.074 207.83 174.02
```

Iteration 1:

```
Step:  AIC=174.01
Y ~ X1 + X21 + X22 + X23 + X24 + X25 + X26 + X27 + X28 + X29 +
    X210 + X211 + X212 + X41 + X52002 + X62 + X63 + X1_X3 + X1_X21 +
    X1_X22 + X1_X23 + X1_X24 + X1_X41 + X1_X52002 + X1_X62 +
    X1_X63 + X1_X64 + X3_X21 + X3_X22 + X3_X23 + X3_X24 + X3_X41 +
    X3_X52002 + X3_X62 + X3_X63 + X3_X64 + X21_X41 + X21_X52002 +
    X21_X62 + X22_X41 + X22_X52002 + X22_X62 + X23_X41 + X23_X52002 +
    X23_X62 + X23_X63 + X24_X41 + X24_X52002 + X41_X62

           Df Sum of Sq     RSS    AIC
- X21_X41   1     0.005 207.78 172.01
- X22_X41   1     0.035 207.81 172.02
- X1_X3     1     0.094 207.87 172.04
- X25       1     0.203 207.98 172.07
- X22_X52002 1    0.920 208.70 172.27
```

Forward Selection

```
#Forward Selection
null_model <- lm(Y~1, data = train_df) # Null model
f_model <- step(null_model, direction = 'forward', scope = list(upper = full_model, lower = null_model))
```

```
Step:  AIC=193.82
Y ~ X41

           Df Sum of Sq    RSS    AIC
+ X28       1   268.405 1261.8 184.63
+ X1_X3     1   217.384 1312.8 186.93
+ X23_X41   1   139.205 1391.0 190.28
+ X23_X52002 1  124.400 1405.8 190.90
+ X1_X63    1   123.990 1406.2 190.91
+ X21       1    96.720 1433.5 192.03
```

Iteration 1:

```
Step:  AIC=144.1
Y ~ X41 + X28 + X1_X3 + X1_X63 + X23_X52002 + X210 + X41_X63 +
    X1 + X21_X52002 + X3_X24 + X24_X52002 + X27 + X23 + X1_X23 +
    X23_X41 + X29 + X25

           Df Sum of Sq    RSS    AIC
<none>                  374.01 144.10
+ X22_X41   1    9.6495 364.36 144.59
+ X3_X63    1    6.9136 367.09 145.02
+ X23_X62   1    6.7669 367.24 145.04
+ X211      1    6.5377 367.47 145.08
+ X3_X52002 1    5.2990 368.71 145.27
+ X22_X62   1    3.6708 370.34 145.53
```

Backward Elimination

Final Iteration:

```
Step:  AIC=149.93
Y ~ X1 + X21 + X23 + X27 + X28 + X29 + X210 + X211 + X212 + X52002 +
    X62 + X63 + X1_X22 + X1_X23 + X1_X41 + X1_X52002 + X1_X62 +
    X1_X63 + X3_X21 + X3_X22 + X3_X23 + X3_X24 + X3_X62 + X3_X63 +
    X3_X64 + X21_X62 + X22_X62 + X23_X41 + X23_X62 + X24_X41 +
    X24_X52002 + X41_X62

            Df Sum of Sq    RSS    AIC
<none>                  246.55 149.93
- X21        1      9.38 255.93 150.10
- X3_X64     1     12.31 258.86 150.76
- X3_X24     1     23.21 269.76 153.15
```

Backward Selection Model Summary:

```
Residual standard error: 3.669 on 45 degrees of freedom
Multiple R-squared:  0.8323,    Adjusted R-squared:  0.7317
F-statistic: 8.272 on 27 and 45 DF,  p-value: 4.522e-10
```

Forward Selection

Final Iteration:

```
Step:  AIC=144.1
Y ~ X41 + X28 + X1_X3 + X1_X63 + X23_X52002 + X210 + X41_X63 +
    X1 + X21_X52002 + X3_X24 + X24_X52002 + X27 + X23 + X1_X23 +
    X23_X41 + X29 + X25

            Df Sum of Sq    RSS    AIC
<none>                  374.01 144.10
+ X22_X41    1    9.6495 364.36 144.59
+ X3_X63     1    6.9136 367.09 145.02
+ X23_X62    1    6.7669 367.24 145.04
+ X211       1    6.5377 367.47 145.08
```

Forward Selection Model Summary:

```
Residual standard error: 3.644 on 54 degrees of freedom
Multiple R-squared:  0.8015,    Adjusted R-squared:  0.7354
F-statistic: 12.12 on 18 and 54 DF,  p-value: 4.705e-13
```

Which Model to select, Backward or Forward ?

```
The Adjusted R square for backward selection model:
 0.7316964
```

```
BIC for Backward selection model:
 486.0681
```

```
The Adjusted R square for forward selection model:
 0.7353676
```

```
BIC for forward selection model:
 459.7576
```

BIC(Bayesian Information Criterion)

- Measures the goodness of fit while penalizing the number of predictors
- BIC is similar to AIC but imposes a higher penalty for the number of predictors

*** Based on the above adjusted R square value and BIC value, selected **forward selection model** as the best subset model ***

Outlier Analysis

- Outlying on X observation
- Outlying on Y observation

Outlying on X-observation

- Extreme point that lies beyond the range of our independent variable
- Leverage measures how much the i-th observation influences its own predicted value.

Interpreting of Leverage

If $hii > 2p/n$ = outlier

```
> threshold
[1] 0.5205479
```

```
        1         2         3         4         5         6         7         8         9        10
0.33061498 0.25721058 0.27387483 0.29330242 0.39685305 0.42215809 0.44281110 0.22129586 0.20262285 0.19370830
       11        12        13        14        15        16        17        18        19        20
0.18493263 0.21044067 0.24158530 0.28135641 0.09499134 0.05943915 0.11485720 0.08376334 0.38981873 0.40262659
       21        22        23        24        25        26        27        28        29        30
0.40489956 0.33708176 0.15011308 0.08118484 1.00000000 0.16701174 0.27658752 0.48491415 0.09256803 0.06682303
       31        32        33        34        35        36        37        38        39        40
0.09673622 0.08376334 0.12465875 0.14562834 0.18702900 0.21827423 0.19050681 0.18632292 0.19206321 0.20134153
       41        42        43        44        45        46        47        48        49        50
0.29564993 0.28053024 0.33017205 0.30024911 0.34280921 0.35982139 0.30418468 0.28338106 0.33814249 0.34339670
       51        52        53        54        55        56        57        58        59        60
0.14791521 0.11685614 0.05214877 0.06827300 0.12050879 0.20592646 0.53956685 0.22812172 0.23185826 0.30114922
       61        62        63        64        65        66        67        68        69        70
0.38353451 0.07386134 0.34492043 0.26074361 0.78286001 0.47705233 0.25441213 0.18599660 0.20062654 0.17549175
       71        72        73
0.20150861 0.23814414 0.44441523
```

```
> leverage[x_outliers]
       25        57        65
1.0000000 0.5395669 0.7828600
```

Outlying on Y-observation

- Extreme points that lies beyond the range of response variable (Y).
- Large residuals and pulls our model towards it.

Studentized Residual

Calculation of size of residuals after removing the extreme points on Y direction for all observations.

```
[1] "Studentized Residuals:"
> print(studentized_residuals)
          1           2           3           4           5           6           7           8           9
 0.13255577  0.10229123  0.58934834 -0.83472768 -0.25263787 -0.45561218  0.72939930  0.07430247 -0.28196472
         10          11          12          13          14          15          16          17          18
 0.41467497 -0.01547412  0.78922209 -0.39670253 -0.62033868 -0.18906986 -0.53286484  1.77620083 -0.54735573
         19          20          21          22          23          24          25          26          27
 0.85055256  0.69293277 -1.58270104 -0.99583742  0.91557944 -1.02441979         NaN  0.67351607 -0.41236981
         28          29          30          31          32          33          34          35          36
-0.56161075 -0.74779612  0.89951103 -0.44758509 -0.83528365  1.64885523 -1.62423354  1.01631910  0.37116086
         37          38          39          40          41          42          43          44          45
 0.31665956 -0.55405868 -0.91916051 -2.34293974  0.94509340  2.37096443 -0.41055971  0.32215201 -0.45048768
         46          47          48          49          50          51          52          53          54
 0.53985580 -1.23227807 -1.58999667  0.99615357  1.94896872  0.59372310  0.68554184 -0.12026121 -1.06664902
         55          56          57          58          59          60          61          62          63
 0.53097349  0.53245308 -0.78989438  2.74146509  1.64384622 -1.17240626 -2.96960469 -0.19546766  0.13757808
         64          65          66          67          68          69          70          71          72
 0.09340197  0.06696759 -0.36801809 -0.01409730  0.29626773 -0.87914240  0.88012610 -0.01856013  1.25127753
         73
-1.83716649
```

```
> cat("Observation outside of -3 and 3: ", y_outliers, "\n")
Observation outside of -3 and 3:
> studentized_residuals[y_outliers]
named numeric(0)
```

Are these data points Influential ?

Influential Cases

DFFITS: DFFITS measures the influence of each observation on the fitted values.

The ith case is influential

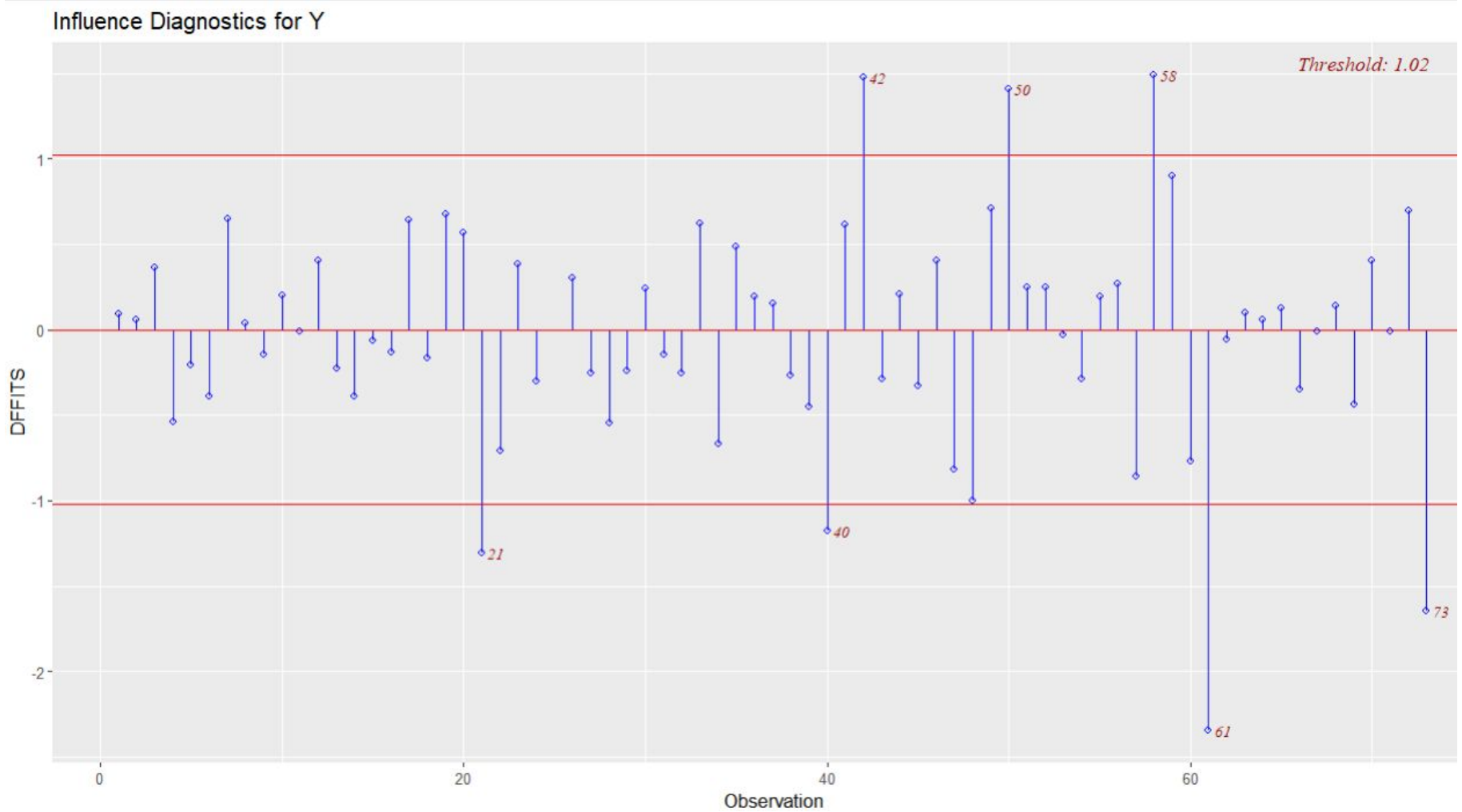
```
> threshold_dffits
[1] 1.020341
```

- If $|DFFITS\ i| > 1$ small data sets
- If $DFFITS\ i > 2\sqrt{p/n}$ large data sets

```
> dffits_values
          1           2           3           4           5           6           7           8           9
 0.09315838  0.06019355  0.36194484 -0.53775671 -0.20492821 -0.38942906  0.65023952  0.03960989 -0.14213698
         10          11          12          13          14          15          16          17          18
 0.20325248 -0.00737082  0.40744753 -0.22389627 -0.38815081 -0.06125449 -0.13395530  0.63982986 -0.16549803
         19          20          21          22          23          24          25          26          27
 0.67983430  0.56887833 -1.30550158 -0.71011074  0.38479085 -0.30450986         NaN  0.30157965 -0.25498237
         28          29          30          31          32          33          34          35          36
-0.54491409 -0.23883979  0.24070636 -0.14647479 -0.25255568  0.62223613 -0.67057600  0.48746904  0.19612646
         37          38          39          40          41          42          43          44          45
 0.15361770 -0.26513244 -0.44815140 -1.17637896  0.61230716  1.48050068 -0.28824704  0.21102309 -0.32535945
         46          47          48          49          50          51          52          53          54
 0.40473493 -0.81476129 -0.99985605  0.71202283  1.40945559  0.24737110  0.24936989 -0.02820834 -0.28873651
         55          56          57          58          59          60          61          62          63
 0.19654702  0.27114841 -0.85508318  1.49036154  0.90313316 -0.76962076 -2.34232134 -0.05520082  0.09983014
         64          65          66          67          68          69          70          71          72
 0.05547092  0.12715599 -0.35149816 -0.00823485  0.14161972 -0.44043168  0.40604648 -0.00932380  0.69957968
         73
-1.64311447
```

```
> dffits_values[high_influence_dffits]
       21        40        42        50        58        61        73
-1.305502 -1.176379  1.480501  1.409456  1.490362 -2.342321 -1.643114
```

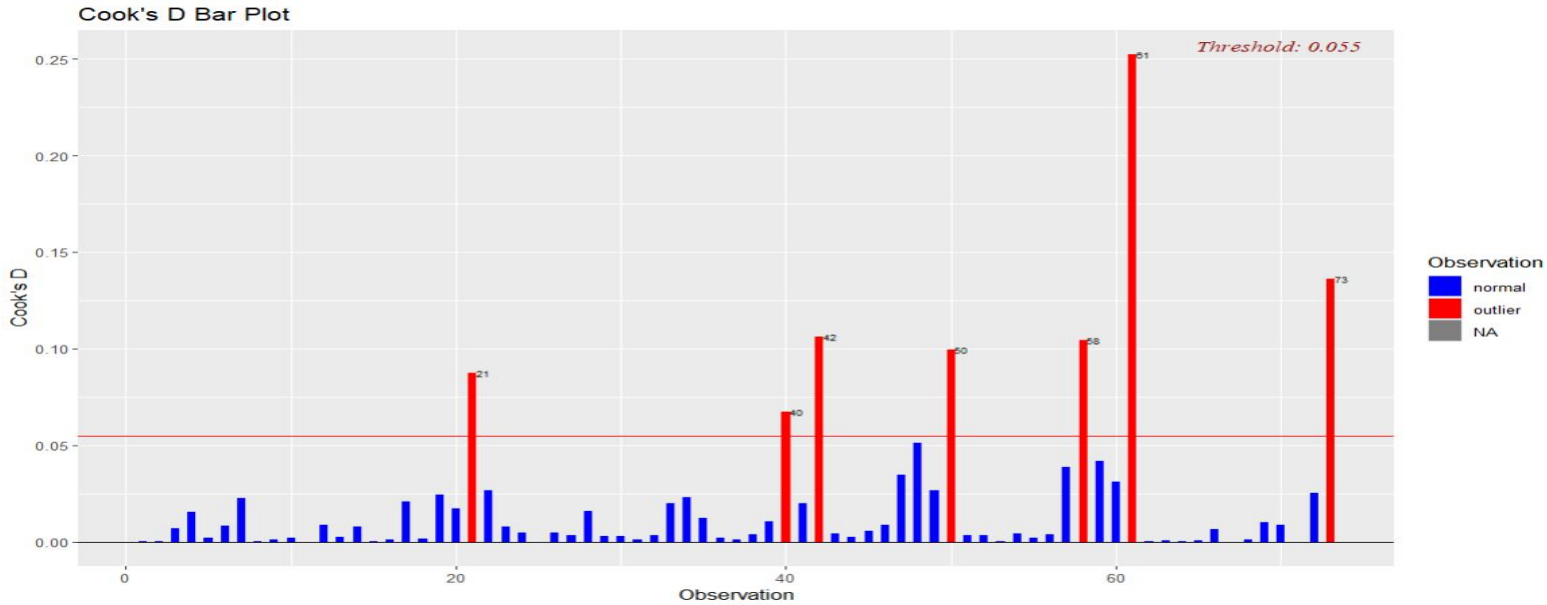
The above points are considered influential. Removing these points would significantly affect the model's predictions.


Influence Diagnostics for Y
Threshold: 1.02
42
50
58
21
40
61
73
DFFITS
1
0
-1
-2
0
20
40
60
Observation

Cook’s Distance: Cook's Distance measures the influence of each data point on the overall fit of the model.

Threshold: Di > 4/n

```
> cooks_distance[high_influence_cooks]
        21         40         42         50         58         61         73
0.08726966 0.06724465 0.10626755 0.09940455 0.10431719 0.25224019 0.13610932
```



DFBETAS: DFBETAS measures the impact of each observation on the individual regression coefficients.

The ith case is influential

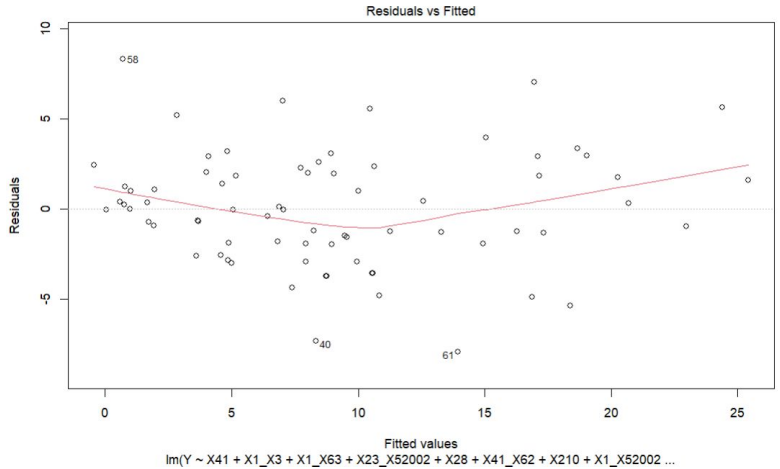
- If $|(DFBETAS)k(i)|>1$ for small data sets
- If $(DFBETAS)k(i) > 2/\sqrt{n}$ for large data sets

```
> dfbetas_values[high_influence_dfbetas, "X41"]
        33         50         58         61         66
-0.2534916  0.4788303 -0.3597691 -0.7512500 -0.3130811
```

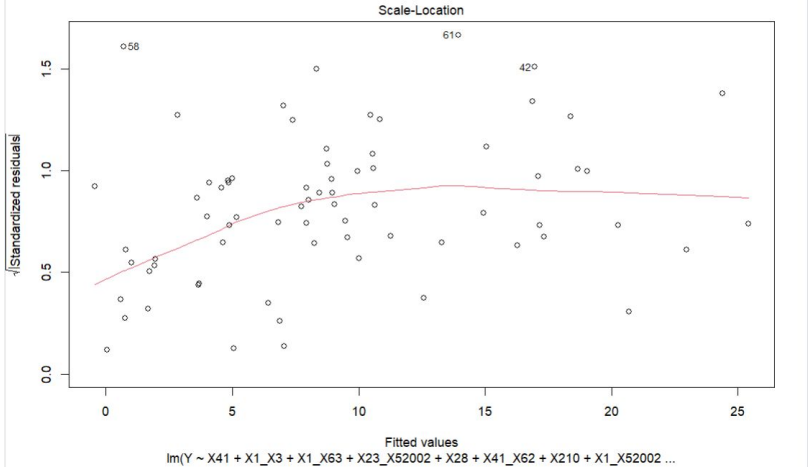
The observations with influential values from all three methods (DFFITS, Cook's Distance, and DFBETAS) are those with the below mentioned values and they should be further scrutinized for removal or adjustment in the model.

```
> cat("The influential data points are:\n" , unique_points)
The influential data points are:
 21 40 42 50 58 61 73 33 66
```

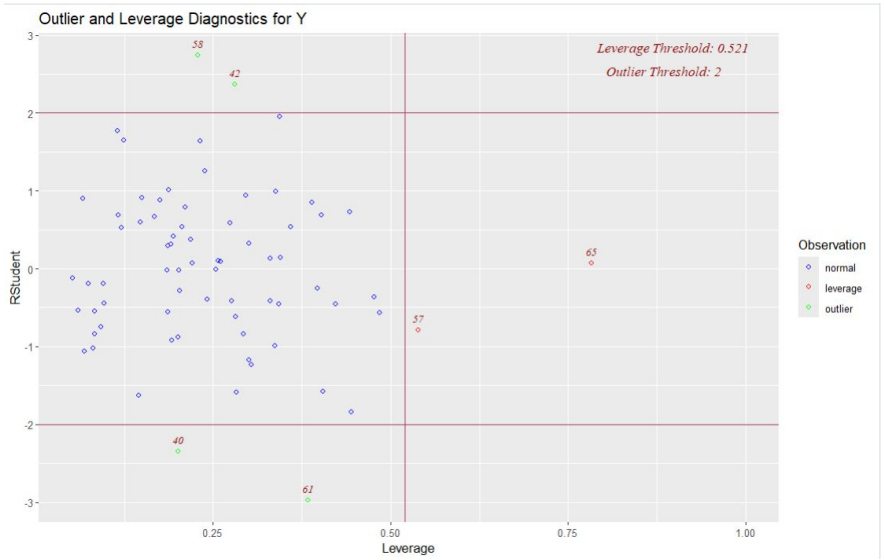
Residuals vs fitted



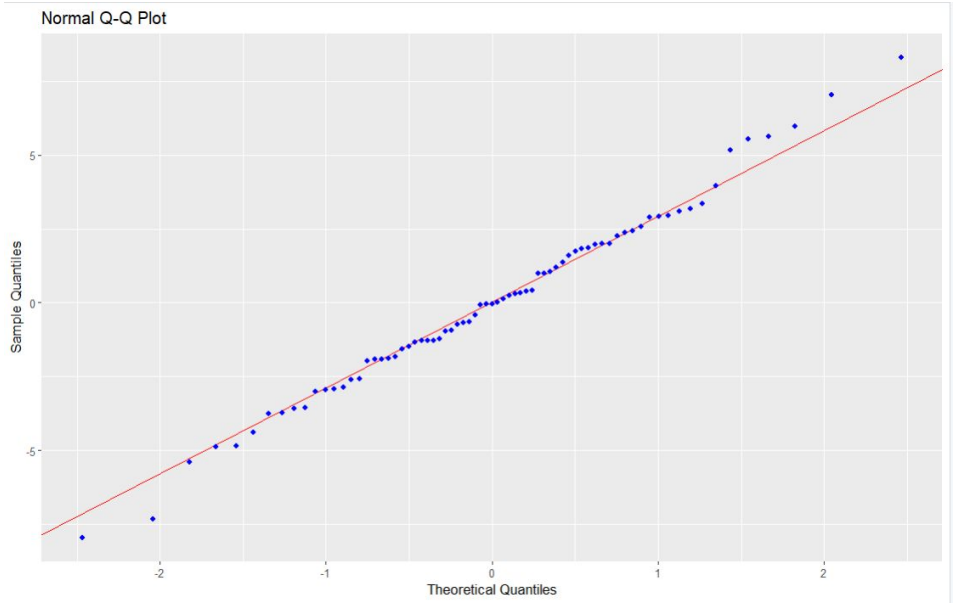
Scale-Location



Outlier and Leverage



Normal Q-Q Plot



Multicollinearity

Variation Inflation Factor: Measures how much the variance of a regression coefficient is inflated due to correlations between predictors.

```
Variance Inflation Factor values for the forward selection without outliers are:
> vif_values <- vif(f_model_wo)
> vif_values
      X41      X1_X3     X1_X63     X41_X62     X23_X63      X3_X24  X1_X52002 X21_X52002       X210      X1_X41 X23_X52002        X27
 4.812984   2.079909   2.119220   8.657483   1.067392   4.227509  10.113997   3.679348   1.329681  36.303683   2.100114   1.349069
     X212       X29    X24_X62    X24_X41    X22_X41     X1_X22 X24_X52002    X21_X41    X23_X41       X26    X24_X64    X21_X64
15.843940   1.175622   1.926560   6.048704   2.362823   2.020224   4.733817   5.562264   2.468381   1.211079   1.083315   1.047339
   X21_X62
 1.452099
> cat("The observation with maximimum VIF value is" , names(which.max(vif_values)), "with a VIF value of", max(vif_values))
The observation with maximimum VIF value is X1_X41 with a VIF value of 36.30368
```

Multicollinearity contd.

Removing Observation with High VIF Value: **X1_X41**

Summary:

```
Residual standard error: 2.283 on 39 degrees of freedom
Multiple R-squared:  0.9215,    Adjusted R-squared:  0.8731
F-statistic: 19.06 on 24 and 39 DF,  p-value: 8.269e-15
```

```
> vif(updated_f_model_wo)
      X41      X1_X3     X1_X63    X41_X62    X23_X63     X3_X24  X1_X52002 X21_X52002       X210 X23_X52002       X27       X212
 4.097608   1.841562   2.109006   4.880871   1.058723   4.219293   6.031185   3.380255   1.258346   2.070264   1.138513   4.515860
      X29    X24_X62    X24_X41    X22_X41     X1_X22 X24_X52002    X21_X41    X23_X41       X26    X24_X64    X21_X64    X21_X62
 1.162158   1.912602   5.106640   1.993966   1.852294   4.383697   3.871222   2.221632   1.114644   1.075011   1.040433   1.444689
>
```

```
The observation with maximimum VIF value is X1_X52002 with a VIF value of 6.031185
```

** Repeating the iteration until all the predictors VIF Values are lesser than 2**

Final Model

Final Model:

- It's the best subset model select from model selection
- Removed influential outliers in both X & Y direction
- Removed correlated columns with VIF value > 4

Summary of the Final Model:

```
Residual standard error: 3.428 on 55 degrees of freedom
Multiple R-squared:  0.7502,    Adjusted R-squared:  0.7139
F-statistic: 20.65 on 8 and 55 DF,  p-value: 5.066e-14
```

Anova of the Final Model:

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
X41        1 1408.96 1408.96 119.9117 1.953e-15 ***
X1_X3      1  213.73  213.73  18.1902 7.924e-05 ***
X1_X63     1   80.92   80.92   6.8865   0.01122 *
X23_X63    1   57.33   57.33   4.8792   0.03137 *
X3_X24     1   45.37   45.37   3.8613   0.05447 .
X210       1   41.57   41.57   3.5378   0.06528 .
X27        1   44.45   44.45   3.7829   0.05690 .
X29        1   48.66   48.66   4.1415   0.04668 *
Residuals 55  646.25   11.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

K-Cross Validation

```
Linear Regression

64 samples
 8 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 52, 51, 51, 50, 52
Resampling results:

  RMSE      Rsquared   MAE
  3.974736  0.6313245  3.099063

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Recommendations

- Predictor variables are more than observed variables, it would have cause overfitting
- Adding more data points to increase the observed variables number to improve the model’s generalization
- Dimensionality reduction through effective feature selection: Remove unnecessary or correlated variables

References

- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models* (4th ed.). McGraw-Hill/Irwin

- Nahhas, R. W. (2024, October 13). 5.22 Influential observations | Introduction to Regression Methods for Public Health Using R. https://www.bookdown.org/rwnahhas/RMPH/mlr-influence.html

- *How do outliers impact linear regression evaluation?* (2023, November 8).

  https://www.linkedin.com/advice/3/how-do-outliers-impact-linear-regression-dz0ff