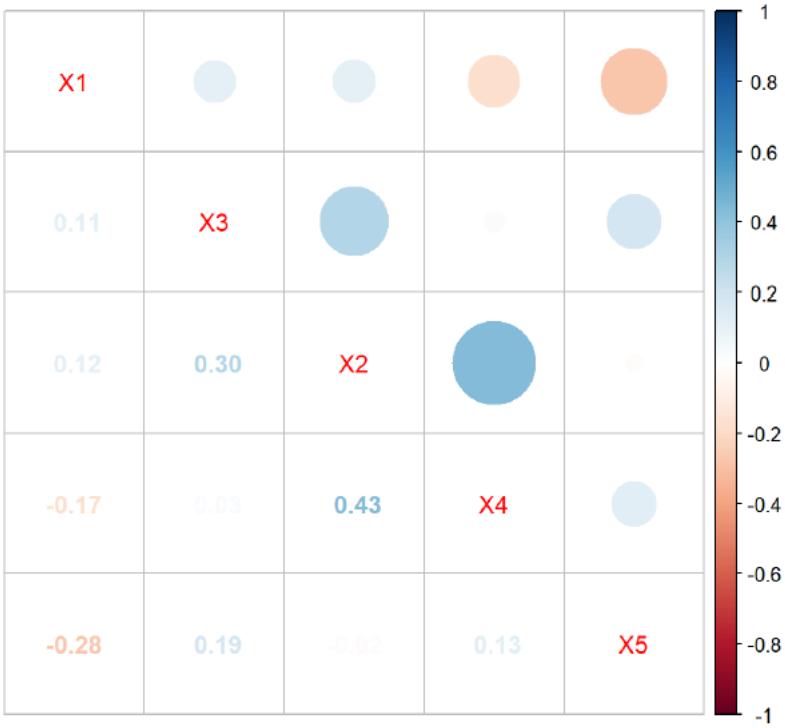
1. Health Dataset

   Analysis:

   1. No missing data
   2. Summary

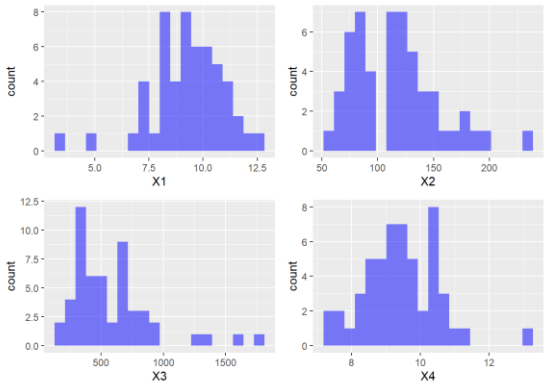
```
##        X1              X2              X3              X4
##  Min.   : 3.600   Min.   : 60.0   Min.   : 190.0   Min.   : 7.200
##  1st Qu.: 8.300   1st Qu.: 82.0   1st Qu.: 353.0   1st Qu.: 8.800
##  Median : 9.400   Median :114.0   Median : 525.0   Median : 9.500
##  Mean   : 9.306   Mean   :116.1   Mean   : 589.8   Mean   : 9.436
##  3rd Qu.:10.300   3rd Qu.:134.0   3rd Qu.: 686.0   3rd Qu.:10.300
##  Max.   :12.800   Max.   :238.0   Max.   :1792.0   Max.   :13.000
##        X5
##  Min.   : 35.0
##  1st Qu.: 80.0
##  Median :103.0
##  Mean   :110.6
##  3rd Qu.:129.0
##  Max.   :292.0
```

   3. Correlation among variables:

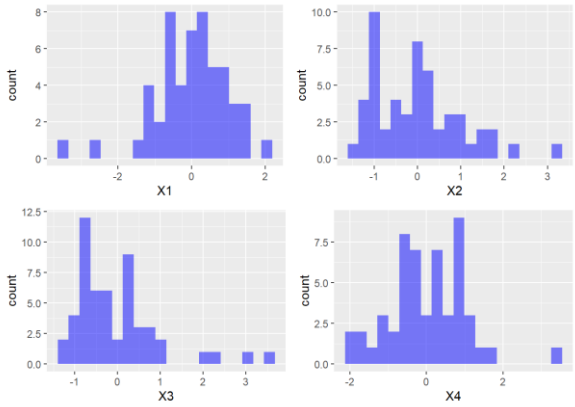


      - X1 and X4 are correlated followed by X2 and X3.

   4. Data distribution with and without scaling:



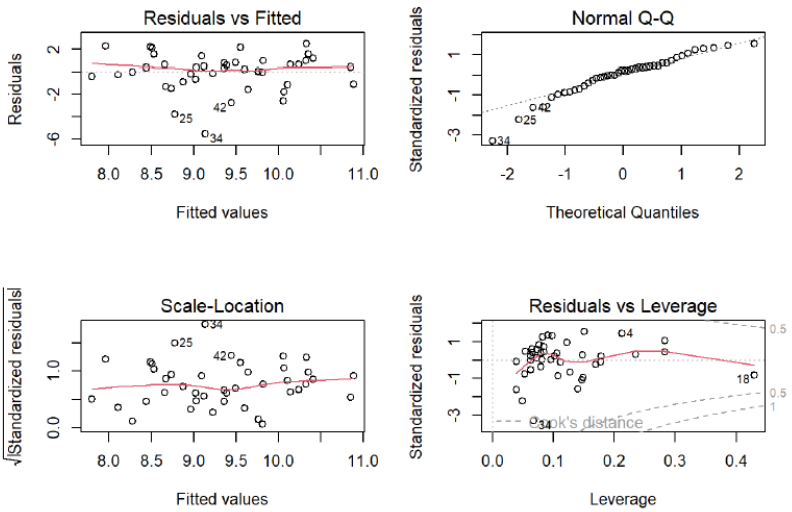
With scaling:



5. Model Building
   a) Multi-linear regression:

```
##
## Call:
## lm(formula = X1 ~ ., data = tr_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5384 -0.8124  0.3500  0.9346  2.4708
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.941468   2.376118   5.446 3.53e-06 ***
## X2           0.004479   0.008806   0.509   0.6140
## X3           0.001347   0.001028   1.311   0.1980
## X4          -0.349895   0.274152  -1.276   0.2098
## X5          -0.014802   0.007185  -2.060   0.0465 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.731 on 37 degrees of freedom
## Multiple R-squared:  0.1827, Adjusted R-squared:  0.09432
## F-statistic: 2.067 on 4 and 37 DF,  p-value: 0.1049
```

- From above, only X5 is statistically significant.
- Even F-statistic isn't that significant(p value is higher).



- Above residual plots vs fitted values shows there are lot of outliers and leverage points.
- Train MSE: 2.65, Test MSE : 1.68.

b) Polynomial regression with single predictors:
   - Just to see the behavior of predictors vs response.
   - Except X5 no predictors showed good relation with response(X1).

c) Polynomial regression with multiple predictors:

```
x2 <- tr_data$X2
log_X3 <- log(tr_data$X3)
log_X5 <- log(tr_data$X5)
log_X4 <- log(tr_data$X4)
poly_model <- lm(X1 ~ X2 + poly(log_X3,log_X4,degree = 1) + poly(log_X5,degree = 3), data = tr_d
ata)
summary(poly_model)
```

   - Summary : not much improvement compared to Linear model

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8775 -0.5405  0.0090  0.8767  3.0103
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              8.695409   1.069724   8.129 1.42e-09 ***
## X2                                       0.005517   0.008856   0.623   0.5374
## poly(log_X3, log_X4, degree = 1)1.0      2.469142   1.976271   1.249   0.2198
## poly(log_X3, log_X4, degree = 1)0.1 -2.588650   1.897594  -1.364   0.1812
## poly(log_X5, degree = 3)1               -4.239883   1.798550  -2.357   0.0241 *
## poly(log_X5, degree = 3)2                0.890678   1.720976   0.518   0.6080
## poly(log_X5, degree = 3)3                3.202922   1.662678   1.926   0.0622 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.64 on 35 degrees of freedom
## Multiple R-squared:  0.3065, Adjusted R-squared:  0.1877
## F-statistic: 2.579 on 6 and 35 DF,  p-value: 0.03564
```

   - Train MSE : 2.24 , Test MSE : 1.46.

d) GLM :
   - Performed cross-validation : LOOCV,10-fold CV and 5- fold CV.
   - Least CV error was found with LOOCV(cverror = 2.76)

e) GAM:
   - Used various GAM models and have performed ANOVA test to find best model.

```
gam.m1 = gam(X1~s(X2,4)+s(X3,4)+s(X4,4)+s(X5,4), data = tr_data)
gam.m2 = gam(X1~s(X2,2)+s(X3,2)+s(X4,5)+s(X5,5), data = tr_data)
gam.m3 = gam(X1~s(X2,2)+ s(X3,2)+s(X4,2)+s(X5,5), data = tr_data)
gam.m4 = gam(X1~lo(X2,X3,X4,X5,span=0.5), data = tr_data)
```

- AIC criterion for gam.m3 is smaller than that of gam.m2,gam.m1,gam.m4, hence we will use gam.m3 for model building.
- Train MSE : 1.8, Test MSE : 3.28 . Model highly overfitted.

f) Support Vector Regressor:

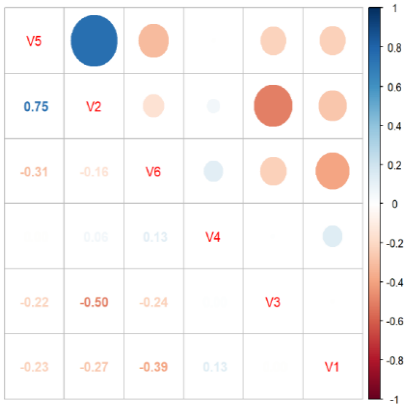
- Finding the best model by tuning. Selecting values of epsilon from 0 to 1 at gap of 0.1, Cost from 1 to 10 at gap of1, gamma from 0.1, 1, 5, 10, 100.
- Performed hyperparameter tuning to find best params
- model resulted in Train MSE : 2.0976561 Test MSE : 1.1008407 R sq. : 0.3662247 As this is not a good fit so tuning the model further manually to find the best model by changing cost, gamma and epsilon. This resulted in cost = 5 gamma = 5.5 and epsilon = 0.4 for radial kernel.
- Train MSE : 0.430 , Test MSE : .431

6) Final Model Results:

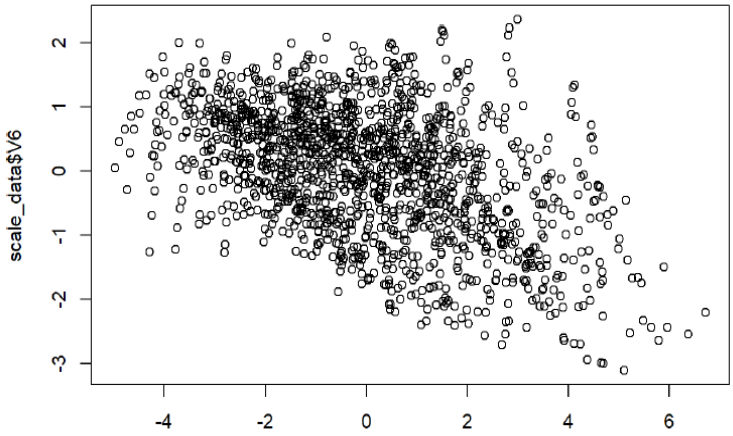
```
##         Model_Name        Train_MSE Test_MSE
## 1     Linear Model 2.64079150899683 1.589009
## 2 Polynomial Model 2.24060093752846 1.466093
## 3    Random Forest 2.24060093752846 1.466093
## 4              GAM  1.7194608785101 3.285385
## 5       GLM(LOOCV)             2.60 2.769386
## 6              SVR 0.43043609837064 0.431499
```

2.Airfoil Dataset

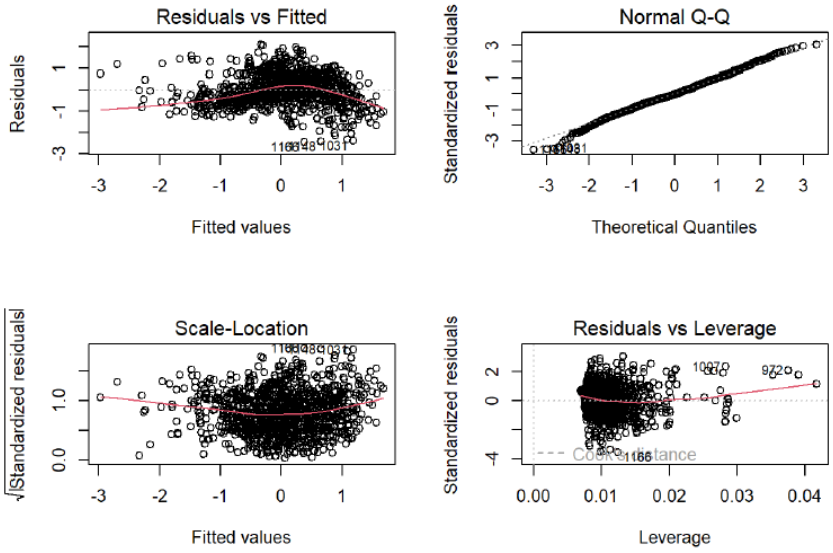
1. No missing values found.
2. DATASET INFO:
   V1 = Frequency, in Hertz.
   V2 = The angle of attack, in degrees.
   V3 = Chord length, in meters.
   V4 = Free-stream velocity, in meters per second.
   V5 = Suction side displacement thickness, in meters.
   V6 = Scaled sound pressure level, in decibels.
3. Correlation plot :



4. Plot of response vs predictor shows nonlinearity



5. Model Building
   a) Multilinear Regression:



- residual plot shows that there are leverage points as well
- Removed outliers and leverage after which RSE and R2squared improved.

b) GLM :
   - LOOCV , 5-fold and 10-fold cross validation was performed.
   - 10 fold CV gave best results.

c) Similarly various other models such as Boosting, Random forest, BART , Lasso regression was trained. Random forest gave best results

4) From boosting&randomforests V1& V5 are most significant variable which can also be cross checked using corr values.

5) Final summary on dataset 2:

- we have done EDA process like checking null values, scaling, checking outliers. And all parameters have their effect on the response it has been observed that ignorance of one parameter also causes rise in test mse. Maximum R2 is obtained when removing outliers which is 72%. The least test error was for Boosting, Bagging showing complex models performed better for this data. Two predictors(V3,V4) are kindof qualitative so had to convert them to quantitative.

6) Best results :

Random forest with train MSE : 0.16 and test MSE : 0.12