50\_startups prediction for Profit Model

**Step1** -> taking prediction simply on dataset

**Step2**-> Change the State categorical variable into dummy variable

**Step3**-> Remove values which is having N/A or zero as value

**Step4**-> predict model , gave me best correlation value till now.

**step5**-> Finding outlier, I have got on 47, but after prediction it doesn't improve the overall performance of model

**step6**-> Taking VIF value from model which I create in Step4 ,

vif(reg\_dummy)

R.D.Spend Administration Marketing.Spend State.California State.Florida

2.331870 1.171396 2.312491 1.372512 1.355895

The highest VIF value is for R.D.Spend, removing it to model and again made the module but no improvement in the result.

Step7-> Taking AIC value from model which I create in Step4 ,

Start: AIC=826.5

dummy\_startups$Profit ~ R.D.Spend + Administration + Marketing.Spend +

State.California + State.Florida

Df Sum of Sq RSS AIC

- State.Florida 1 1.8538e+05 2.2523e+09 824.50

- State.California 1 1.8789e+07 2.2709e+09 824.88

- Administration 1 9.1592e+07 2.3437e+09 826.33

<none> 2.2521e+09 826.50

- Marketing.Spend 1 1.0659e+08 2.3587e+09 826.63

- R.D.Spend 1 2.2134e+10 2.4386e+10 934.08

Step: AIC=824.5

dummy\_startups$Profit ~ R.D.Spend + Administration + Marketing.Spend +

State.California

Df Sum of Sq RSS AIC

- State.California 1 2.7455e+07 2.2797e+09 823.06

- Administration 1 9.1507e+07 2.3438e+09 824.33

<none> 2.2523e+09 824.50

- Marketing.Spend 1 1.0648e+08 2.3587e+09 824.63

- R.D.Spend 1 2.2184e+10 2.4436e+10 932.17

Step: AIC=823.06

dummy\_startups$Profit ~ R.D.Spend + Administration + Marketing.Spend

Df Sum of Sq RSS AIC

- Marketing.Spend 1 9.2234e+07 2.3720e+09 822.88

- Administration 1 1.0115e+08 2.3809e+09 823.06

<none> 2.2797e+09 823.06

- R.D.Spend 1 2.2411e+10 2.4690e+10 930.65

Step: AIC=822.88

dummy\_startups$Profit ~ R.D.Spend + Administration

Df Sum of Sq RSS AIC

<none> 2.3720e+09 822.88

- Administration 1 1.9401e+08 2.5660e+09 824.50

- R.D.Spend 1 5.4757e+10 5.7129e+10 967.24

Call:

lm(formula = dummy\_startups$Profit ~ R.D.Spend + Administration,

data = dummy\_startups)

Coefficients:

(Intercept) R.D.Spend Administration

6.300e+04 8.141e-01 -7.897e-02

Note:- As per this result the minimum AIC value **AIC=822.88**, model is dummy\_startups$Profit ~ R.D.Spend + Administration

**Step7**-> Making model on recommendation of AIC value but still that result don't have considerable improvement.

So as per my coding best result is from step4.

