Multi-Linear Regression

Example- Startup Dataset

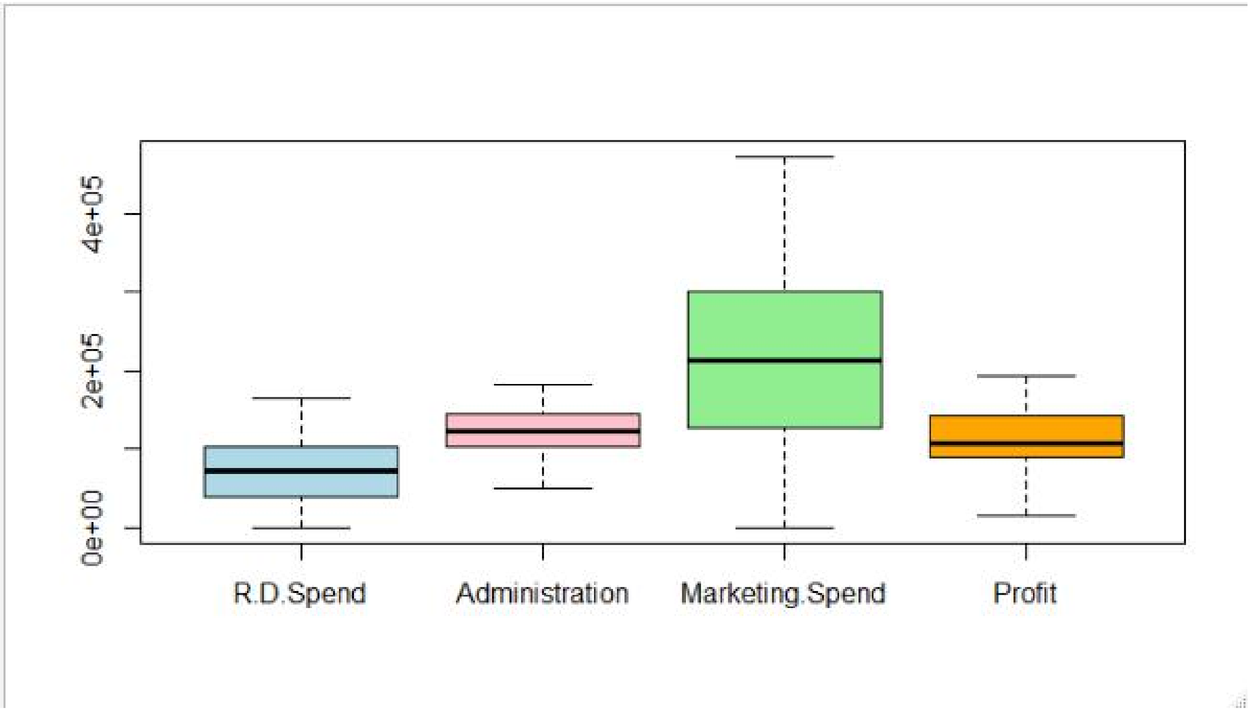
Target variable is Profit

Summary ➔

R.D.Spend	Administration	Marketing.Spend	State	Profit
Min. : 0	Min. : 51283	Min. : 0	California:17	Min. : 14681
1st Qu.: 39936	1st Qu.:103731	1st Qu.:129300	Florida :16	1st Qu.: 90139
Median : 73051	Median :122700	Median :212716	New York :17	Median :107978
Mean : 73722	Mean :121345	Mean :211025		Mean :112013
3rd Qu.:101603	3rd Qu.:144842	3rd Qu.:299469		3rd Qu.:139766
Max. :165349	Max. :182646	Max. :471784		Max. :192262

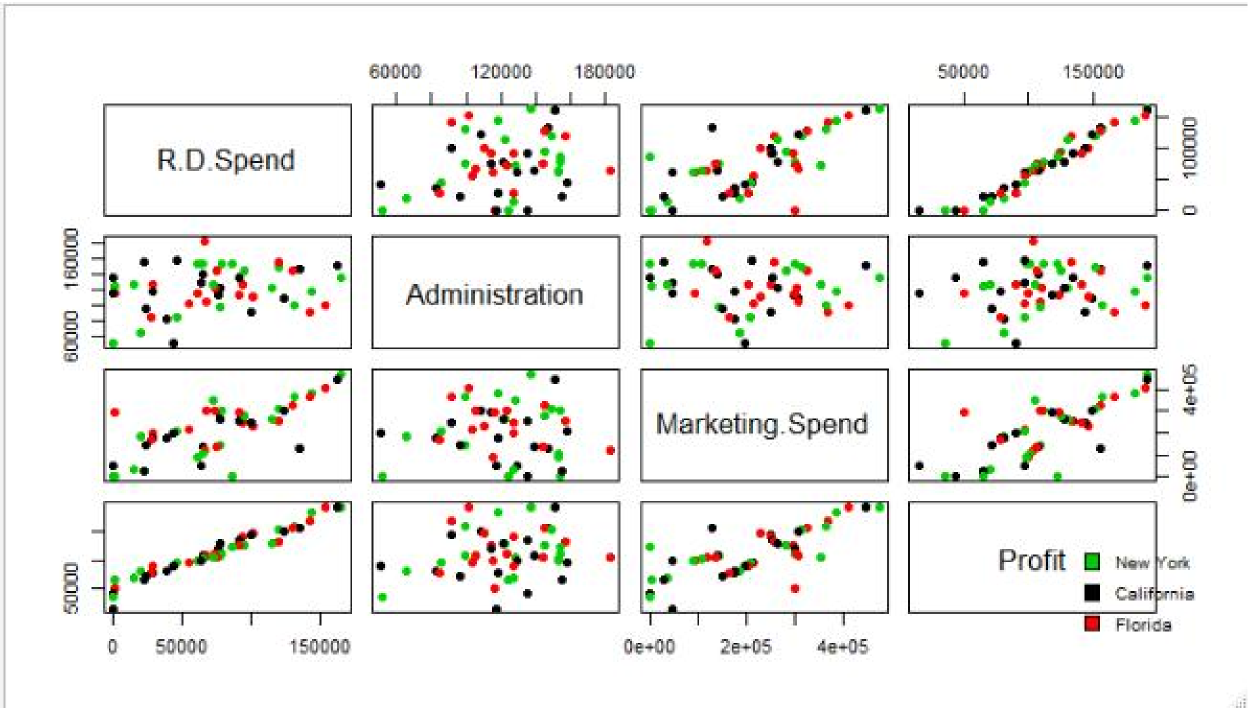
In the above summary all variables are continues except state because state is in categorial format.

Box Plot ➔



From the above box plot we can infer that no outliers are present in data.

Pairs Plot →



Correlation →

	R.D.Spend	Administration	Marketing.Spend	Profit
R.D.Spend	1	0.241955245	0.724248133	0.972900466
Administration	0.241955245	1	-0.032153875	0.200716568
Marketing.Spend	0.724248133	-0.032153875	1	0.747765722
Profit	0.972900466	0.200716568	0.747765722	1

Two variables in Administration and Marketing.Spend are negatively correlated and remaining are positively correlated with each other, so maybe there is no collinearity problem in independent variables.

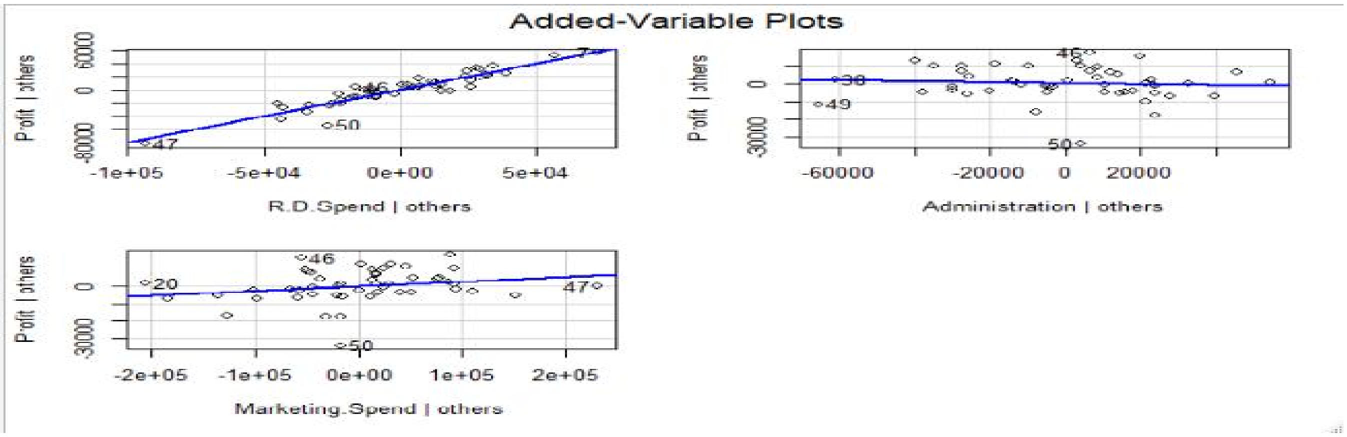
Model-1 ➔

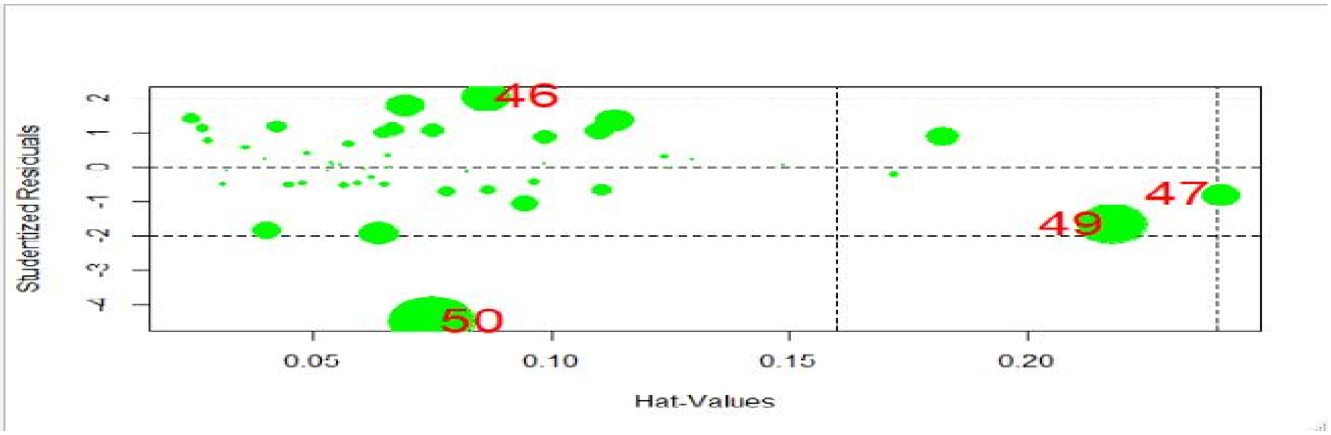
```
model.S <- lm(Profit~R.D.Spend+Administration+Marketing.Spend)
```

```
Multiple R-squared:  0.9507,   Adjusted R-squared:  0.9475
```

Based on above R^2 value, 95% of variation in the profit because Administration is not significant where as Marketing.Spend in somewhat significant in our model.

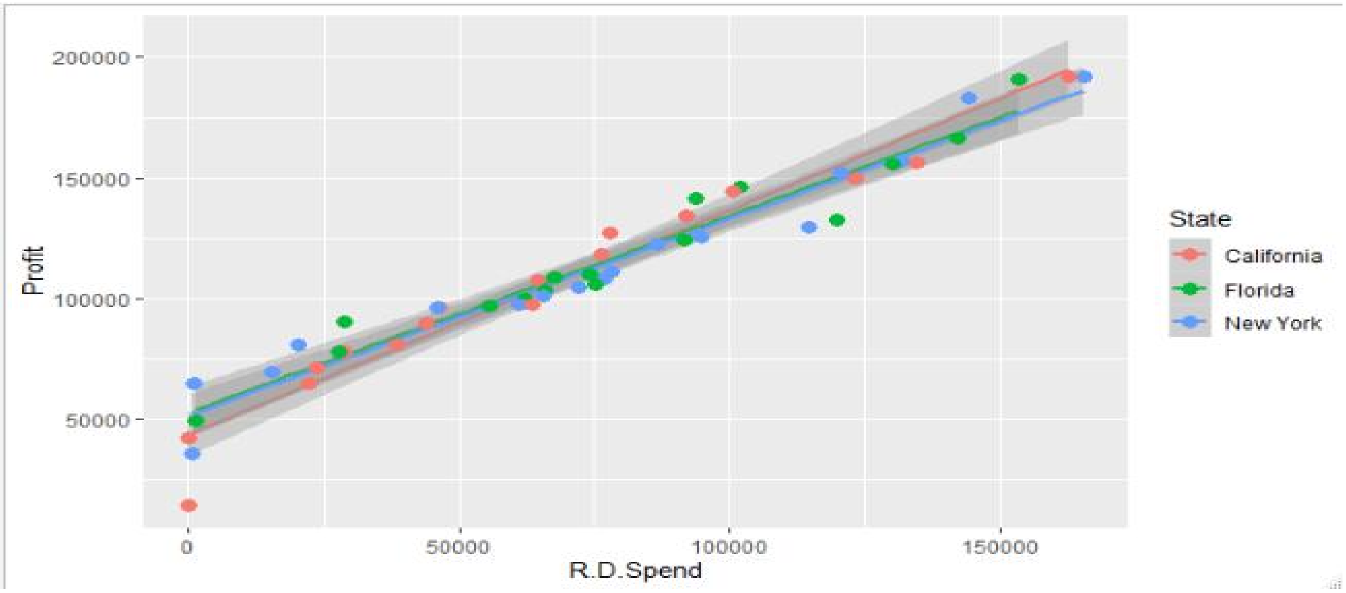
AV Plot ➔



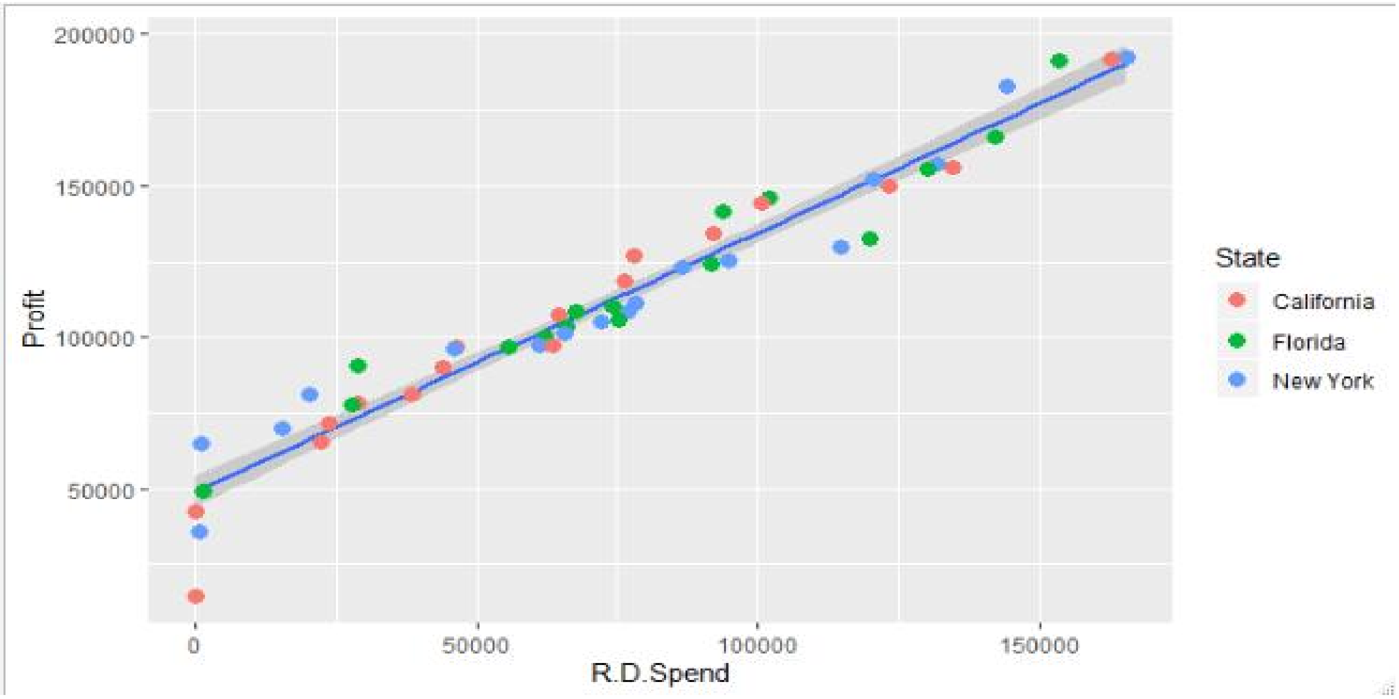


From the above plot observations 46,47,49 and 50 are influence index so we can remove them from our model to get more accuracy.

Considering State as variable



Based on above plot , all the plots are overlapping with each other and only negligible difference between them. By removing state variable we will get same accuracy then no need to consider state variable , so we will not consider State variable in our model.



Model-2 ➔

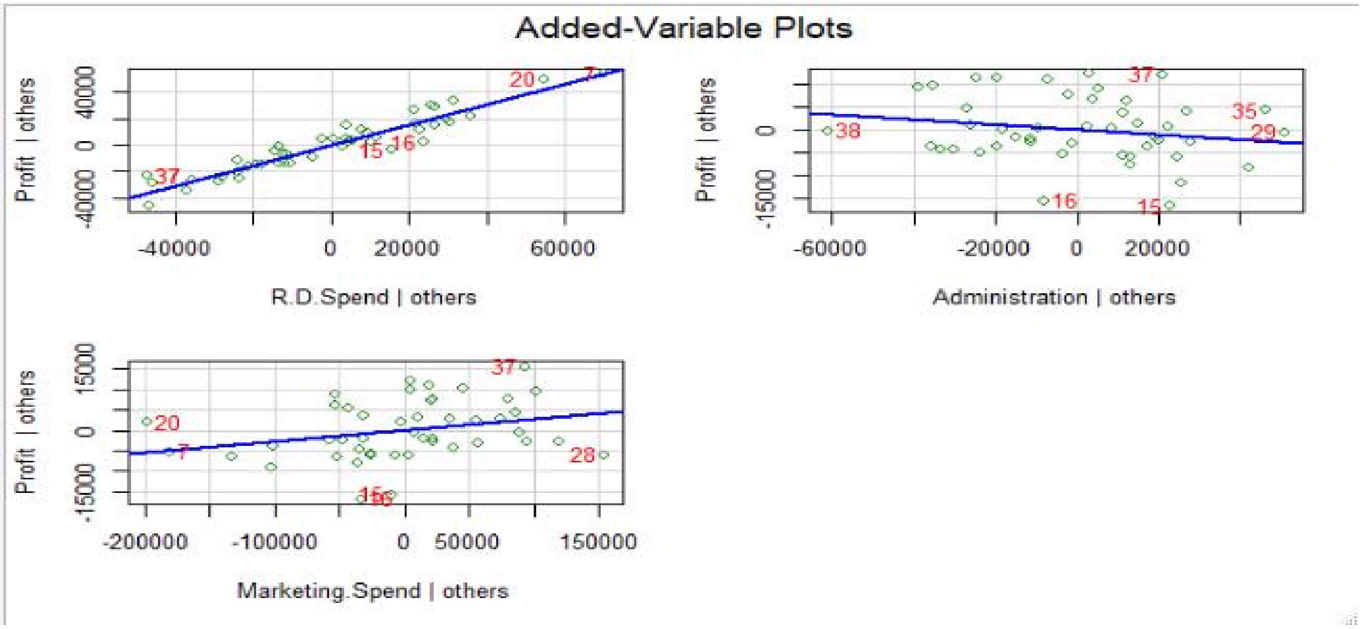
```
model.S.2 <- lm(Profit~R.D.Spend+Administration+Marketing.Spend,data = df_Sta
rtups)

Multiple R-squared:  0.9626,   Adjusted R-squared:  0.9599

rmse_2
6774.245

correlation is 0.9748282
```

AV Plot ➔

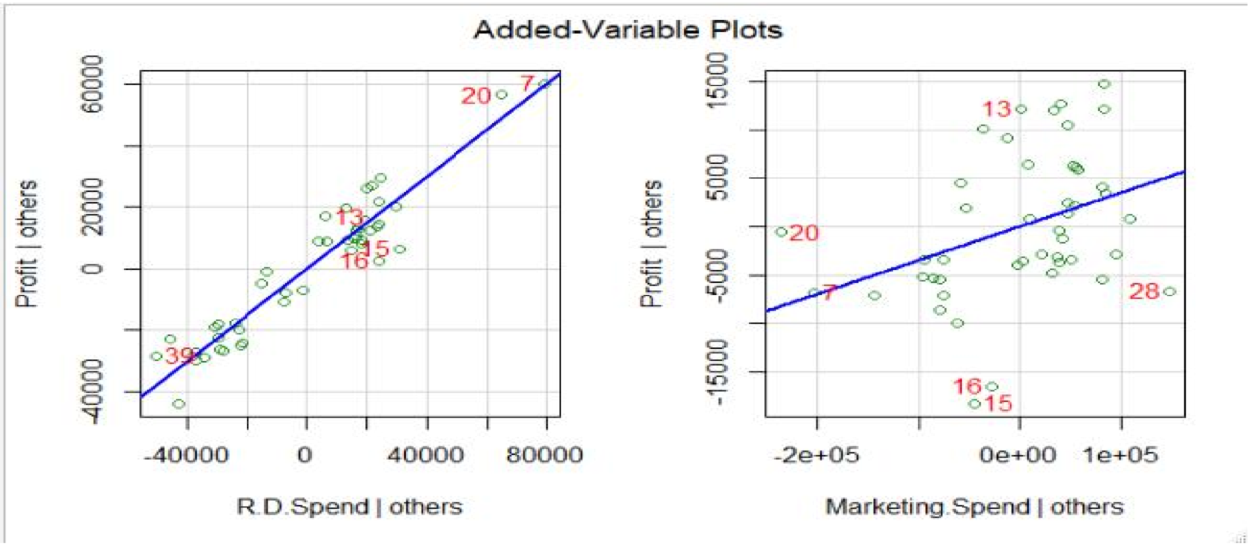


From the above information, Administrative is insignificant in our model with only 79% confidence level.

Model-3 ➔

```
model.S.3 <- lm(Profit~R.D.Spend+Marketing.Spend,data = df_Startups)
Multiple R-squared:  0.9612,	Adjusted R-squared:  0.9594
RMSE value is 6899.99
Correlation is 0.9748121
```

AV Plot ➔



After removing Administration we get probability of error for considering variable Marketing.Spend is 0.01 which is less than 0.05, so we can say it is significant variable in model.

Model No	Variable	R^2	RMSE	Cor(Y,predicted)
Model-1	All variable except State	0.9507	8855.344	0.975062
Model-2	Removed obs-46,47,49,50	0.9626	6774.245	0.9748282
Model-3	Removed variable Administration from Model-2	0.9612	6899.99	0.9748121

Based on high R^2 and correlation between predicted and actual value and low RMSE Model-2 is good model.