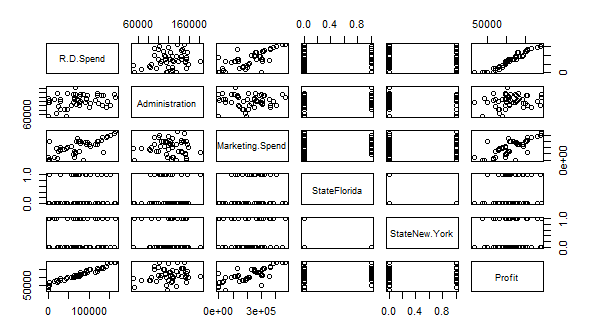
1. **50 Startups:**

df <- read.csv("50\_Startups.csv")  
View(df)

#One Hot Encoding to Categorical Variable State  
install.packages("caret")  
library(lattice)  
library(ggplot2)  
library(caret)  
dummy <- dummyVars(" ~ .", data=df)  
newdata <- data.frame(predict(dummy, newdata = df))  
new\_df <- newdata[-4]  
View(new\_df)

#EDA  
pairs(new\_df)  
  
cor(new\_df)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | R.D.Spend | Administration | Marketing.Spend | StateFlorida | StateNew.York | Profit |
| R.D.Spend | 1 | 0.241955245 | 0.72424813 | 0.10571106 | 0.039068162 | 0.9729005 |
| Administration | 0.24195525 | 1 | -0.03215388 | 0.01049309 | 0.005145226 | 0.2007166 |
| Marketing.Spend | 0.72424813 | -0.032153875 | 1 | 0.20568545 | -0.0336698 | 0.7477657 |
| StateFlorida | 0.10571106 | 0.010493089 | 0.20568545 | 1 | -0.492365964 | 0.1162443 |
| StateNew.York | 0.03906816 | 0.005145226 | -0.0336698 | -0.49236596 | 1 | 0.0313676 |
| Profit | 0.97290047 | 0.200716568 | 0.74776572 | 0.11624426 | 0.0313676 | 1 |

#Multi Linear Model  
model <- lm(Profit~R.D.Spend+Administration+Marketing.Spend+StateFlorida+StateNew.York, data=new\_df)  
summary(model)  
Call:  
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +   
 StateFlorida + StateNew.York, data = new\_df)

Residuals:  
 Min 1Q Median 3Q Max   
-33504 -4736 90 6672 17338

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.013e+04 6.885e+03 7.281 4.44e-09 \*\*\*  
R.D.Spend 8.060e-01 4.641e-02 17.369 < 2e-16 \*\*\*  
Administration -2.700e-02 5.223e-02 -0.517 0.608   
Marketing.Spend 2.698e-02 1.714e-02 1.574 0.123   
StateFlorida 1.988e+02 3.371e+03 0.059 0.953   
StateNew.York -4.189e+01 3.256e+03 -0.013 0.990   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom  
Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452   
F-statistic: 169.9 on 5 and 44 DF, p-value: < 2.2e-16

#We will drop Marketing spend since we see a high collinearity between R.D Spend and Marketing Spend  
model <- lm(Profit~R.D.Spend+Administration+StateFlorida+StateNew.York, data=new\_df)  
summary(model)

Call:  
lm(formula = Profit ~ R.D.Spend + Administration + StateFlorida +   
 StateNew.York, data = new\_df)

Residuals:  
 Min 1Q Median 3Q Max   
-33758 -4784 -116 6383 17032

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.460e+04 6.371e+03 8.571 5.17e-11 \*\*\*  
R.D.Spend 8.609e-01 3.112e-02 27.665 < 2e-16 \*\*\*  
Administration -5.270e-02 5.042e-02 -1.045 0.301   
StateFlorida 1.091e+03 3.377e+03 0.323 0.748   
StateNew.York -3.934e+01 3.309e+03 -0.012 0.991   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9593 on 45 degrees of freedom  
Multiple R-squared: 0.948, Adjusted R-squared: 0.9434   
F-statistic: 205 on 4 and 45 DF, p-value: < 2.2e-16

#We will drop States columns since the p value is very high for these 2.  
model <- lm(Profit~R.D.Spend+Administration, data=new\_df)  
summary(model)

Call:  
lm(formula = Profit ~ R.D.Spend + Administration, data = new\_df)

Residuals:  
 Min 1Q Median 3Q Max   
-34006 -4801 -303 6034 17843

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.489e+04 6.017e+03 9.122 5.7e-12 \*\*\*  
R.D.Spend 8.621e-01 3.016e-02 28.589 < 2e-16 \*\*\*  
Administration -5.300e-02 4.940e-02 -1.073 0.289   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

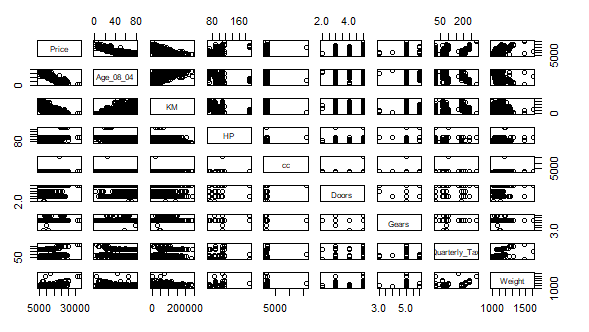
Residual standard error: 9402 on 47 degrees of freedom  
Multiple R-squared: 0.9478, Adjusted R-squared: 0.9456   
F-statistic: 426.8 on 2 and 47 DF, p-value: < 2.2e-16

|  |  |  |
| --- | --- | --- |
|  | R^2 | Adjusted R^2 |
| Model 1 | 0.9508 | 0.9452 |
| Model 2 | 0.948 | 0.9434 |
| Model 3 | 0.9478 | 0.9456 |

Since we see a better Adjusted R2 value, we will consider model 3.

1. **ToyotaCorolla:**

df <- read.csv("ToyotaCorolla.csv")  
df <- df[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]  
View(df)

#EDA  
pairs(df)  
  
cor(df)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Price** | Age\_08\_04 | KM | HP | cc | Doors | Gears | Quarterly\_Tax | Weight |
| **Price** | **1.0000** | **-0.8766** | **-0.5700** | **0.3150** | **0.1264** | **0.1853** | **0.0631** | **0.2192** | **0.5812** |
| Age\_08\_04 | **-0.8766** | 1.0000 | 0.5057 | -0.1566 | -0.0981 | -0.1484 | -0.0054 | -0.1984 | -0.4703 |
| KM | **-0.5700** | 0.5057 | 1.0000 | -0.3335 | 0.1027 | -0.0362 | 0.0150 | 0.2782 | -0.0286 |
| HP | **0.3150** | -0.1566 | -0.3335 | 1.0000 | 0.0359 | 0.0924 | 0.2095 | -0.2984 | 0.0896 |
| cc | **0.1264** | -0.0981 | 0.1027 | 0.0359 | 1.0000 | 0.0799 | 0.0146 | 0.3070 | 0.3356 |
| Doors | **0.1853** | -0.1484 | -0.0362 | 0.0924 | 0.0799 | 1.0000 | -0.1601 | 0.1094 | 0.3026 |
| Gears | **0.0631** | -0.0054 | 0.0150 | 0.2095 | 0.0146 | -0.1601 | 1.0000 | -0.0055 | 0.0206 |
| Quarterly\_Tax | **0.2192** | -0.1984 | 0.2782 | -0.2984 | 0.3070 | 0.1094 | -0.0055 | 1.0000 | 0.6261 |
| Weight | **0.5812** | -0.4703 | -0.0286 | 0.0896 | 0.3356 | 0.3026 | 0.0206 | 0.6261 | 1.0000 |

#Multi Linear Model  
model <- lm(Price~Age\_08\_04+KM+HP+cc+Doors+Gears+Quarterly\_Tax+Weight, data=df)  
summary(model)

Call:  
lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears +   
 Quarterly\_Tax + Weight, data = df)

Residuals:  
 Min 1Q Median 3Q Max   
-9366.4 -793.3 -21.3 799.7 6444.0

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -5.573e+03 1.411e+03 -3.949 8.24e-05 \*\*\*  
Age\_08\_04 -1.217e+02 2.616e+00 -46.512 < 2e-16 \*\*\*  
KM -2.082e-02 1.252e-03 -16.622 < 2e-16 \*\*\*  
HP 3.168e+01 2.818e+00 11.241 < 2e-16 \*\*\*  
cc -1.211e-01 9.009e-02 -1.344 0.17909   
Doors -1.617e+00 4.001e+01 -0.040 0.96777   
Gears 5.943e+02 1.971e+02 3.016 0.00261 \*\*   
Quarterly\_Tax 3.949e+00 1.310e+00 3.015 0.00262 \*\*   
Weight 1.696e+01 1.068e+00 15.880 < 2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1342 on 1427 degrees of freedom  
Multiple R-squared: 0.8638, Adjusted R-squared: 0.863   
F-statistic: 1131 on 8 and 1427 DF, p-value: < 2.2e-16