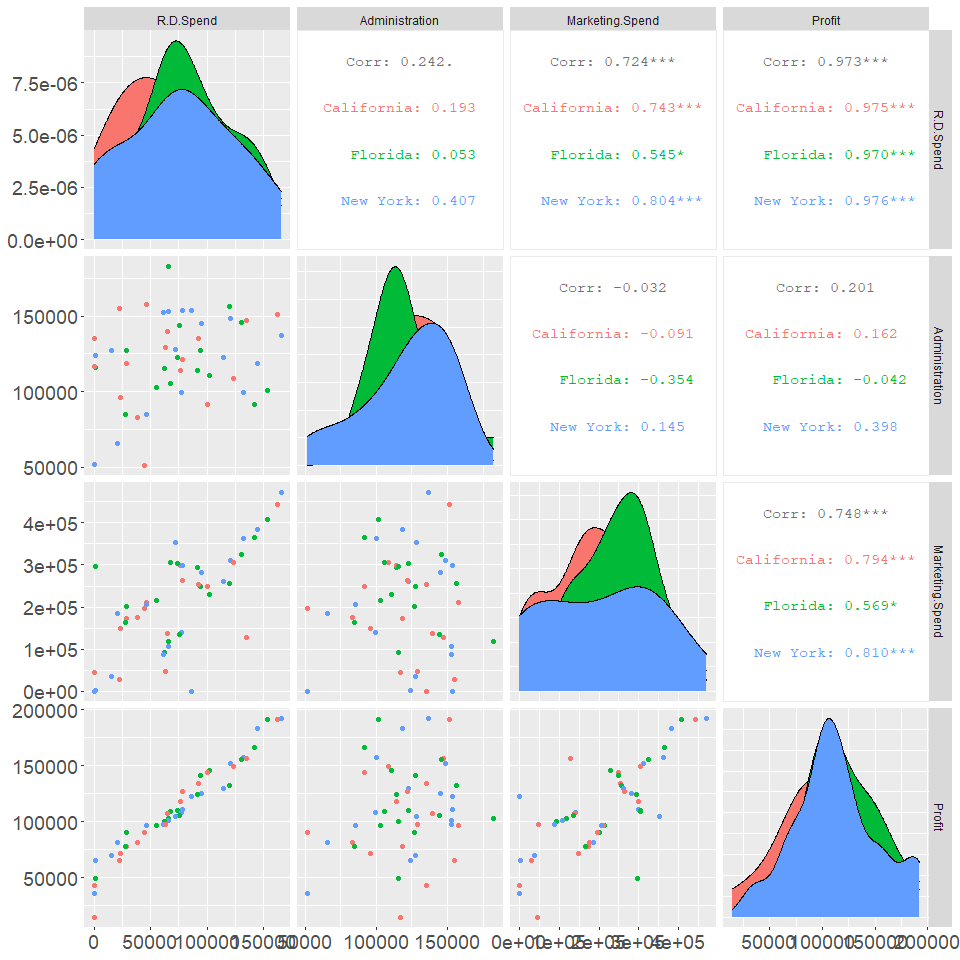
50\_Startups\_StatAnalysis

Tram Ngoc Le (Jennie)

2/28/2021

## 1. Read the data

## 2. EDA: Scatterplot Matrix

 Given the exploratory analysis, restricting our regression to three variables of R&D Spend, Marketing Spend, and State makes sense. Looking at the scatterplot matrices, the relationship between R&D Spend and Profit, as well as the relationship between Marketing Spend and Profit can be described by a line, therefore, the correlation for those variables are meaningful. On the contrary, the relationship between Profit and Administration cannot be described by a line, therefore, the correlation between Profit and Administration is meaningless.

The correlation between Profit and other 3 quantitative variables of R&D Spend, Marketing Spend, and Administration can be demonstrated as follow:

cor(Profit, R&D Spending) = 0.973 cor(Profit, Administration) = 0.201 cor(Profit, Marketing Spend) = 0.748 cor(R&D Spend, Marketing Spend) = 0.724 cor(R&D Spend, Administration) = 0.242 cor(Marketing Spend, Administration) = -0.321

The correlations of 0.973 and 0.748, which are close to 1, between Profit and R&D Spending as well as Profit and Marketing Spend, respectively, demonstrate strong, positive relationships.

The correlation of 0.724 demonstrates a moderate and positive relationship between R&D Spending and Marketing Spend. This correlation needs to be as small as possible. However, it is acceptable in this case since they are not highly correlated.

The categorical variable, “State”, can be used in our regression with Profit, R&D Spending, and Marketing Spend since each subgroup of “State” has roughly normal distribution.

## 3. Summary Plots

col.vector <- c("New York" = "cadetblue",  
 "California" = "firebrick",  
 "Florida" = "pink")  
  
# Interaction term (State vs R&D Spending)  
s.1 <- x %>% ggplot(aes(x=R.D.Spend, y=Profit, col=State))+  
 geom\_point(size=2)+  
 ggtitle("R&D Spending vs. Profit of Startups")+  
 labs(x="R&D Spend ($)", y="Profit ($)")+  
 scale\_color\_manual(values=col.vector)+  
 theme.info  
  
# Interaction term (State vs Marketing Spending)  
s.2 <- x %>% ggplot(aes(x=Marketing.Spend, y=Profit, col=State))+  
 geom\_point(size=2)+  
 ggtitle("Marketing Spending vs. Profit of Startups")+  
 labs(x="Marketing Spending ($)", y="Profit ($)")+  
 scale\_color\_manual(values=col.vector)+  
 theme.info  
  
# Add two plots into one figure  
grid.arrange(s.2, s.1, ncol=2)

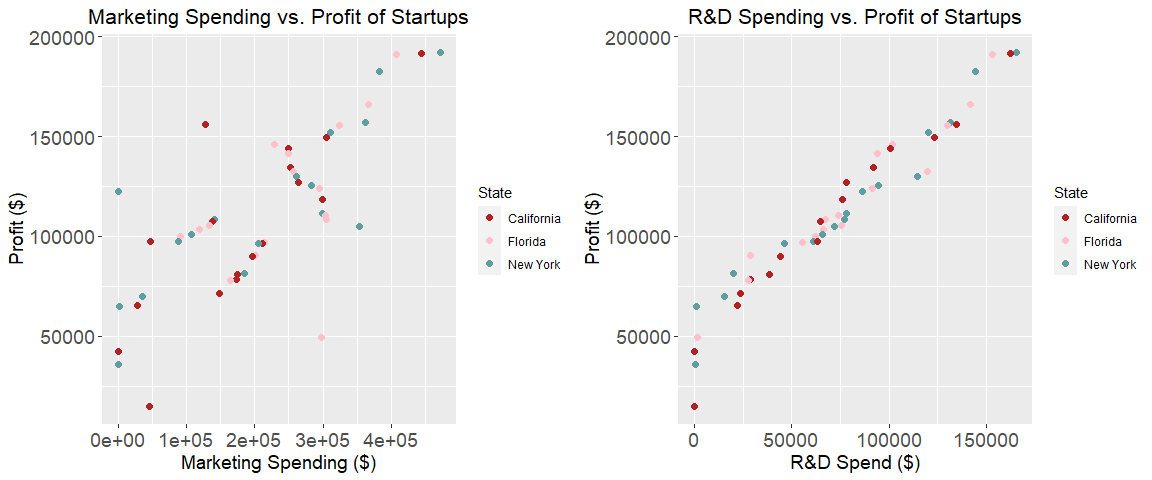


FIGURE 2.TERM INTERACTIONS (R&D SPENDING AND STATE) AND (STATE AND MARKETING SPEND)

Interaction Terms, Categorical x Numerical (State x R&D Spend) and (State x Marketing Spend)

Visually speaking, it is possible that (State x Marketing Spend) is an interaction. For (State x Marketing Spend), all the points seems to form three upward straight line that are not parallel For (State x R&D Spend), all the points look collerated and might created three parallel lines.

# Fit a model to predict profit  
lm.1 <- lm(Profit~R.D.Spend +  
 Marketing.Spend +  
 Marketing.Spend\*State,  
 data = x)  
summary(lm.1)

##   
## Call:  
## lm(formula = Profit ~ R.D.Spend + Marketing.Spend + Marketing.Spend \*   
## State, data = x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28752.5 -4190.5 -10.4 5353.7 14114.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.037e+04 4.207e+03 9.595 2.98e-12 \*\*\*  
## R.D.Spend 7.910e-01 4.112e-02 19.234 < 2e-16 \*\*\*  
## Marketing.Spend 6.792e-02 2.308e-02 2.943 0.00522 \*\*   
## StateFlorida 9.187e+03 8.001e+03 1.148 0.25725   
## StateNew York 1.050e+04 5.670e+03 1.851 0.07104 .   
## Marketing.Spend:StateFlorida -4.620e-02 3.255e-02 -1.419 0.16299   
## Marketing.Spend:StateNew York -5.512e-02 2.470e-02 -2.231 0.03093 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9051 on 43 degrees of freedom  
## Multiple R-squared: 0.9557, Adjusted R-squared: 0.9496   
## F-statistic: 154.8 on 6 and 43 DF, p-value: < 2.2e-16

Model evaluation

1. RMSE = 9051

Interpretation: Expect approximately 95% of Startup profit (y-value) to be within +/- two $9,051 from the regression line.

1. Adjusted R-square = 0.9496

Interpretation: About 94.96% of the variability in Startup profit (y-value) can be explained by this regression model containing R&D Spend, Marketing Spend, and (State and Marketing Spend)

1. Overall F-test: Test all slopes at once.

Hypothesis:

Set alpha = 0.05

H\_0: B(R&D Spend) = B(Marketing Spend) = B(State.Florida) = B(StateNewYork) = B(Marketing Spend: StateFlorida) = B(Marketing.Spend: StateNewYork)

H\_a: At least one slope (B) is not zero

# Analysis of variance (ANOVA)  
anova(lm.1)

## Analysis of Variance Table  
##   
## Response: Profit  
## Df Sum Sq Mean Sq F value Pr(>F)   
## R.D.Spend 1 7.5349e+10 7.5349e+10 919.6841 < 2e-16 \*\*\*  
## Marketing.Spend 1 3.1165e+08 3.1165e+08 3.8039 0.05767 .   
## State 2 2.3905e+05 1.1952e+05 0.0015 0.99854   
## Marketing.Spend:State 2 4.2121e+08 2.1060e+08 2.5706 0.08820 .   
## Residuals 43 3.5229e+09 8.1929e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p-value < 2e-16 < 0.05 -> fail to reject null hypothesis, the model is adaquate

1. Individual t-test:

Test each slope separately

Hypothesis tests when alpha = 0.05:

* Slope of “R.D. Spend” is statistically significant (p-value < 2e-16 < alpha)
* Slope of “Marketing.Spend” is statistically significant (p-value = 0.0052 < alpha)
* Slope of “StateFlorida” is NOT statistically significant (p-value = 0.25 > alpha)
* Slope of “StateNewYork” is NOT statistically significant (p-value = 0.07 > alpha)
* Slope of “Marketing.Spend:StateFlorida” is NOT statistically significant (p-value = 0.16 > alpha)
* Slope of “Marketing.Spend:StateNewYork” is statistically significant (p-value = 0.03 < alpha)