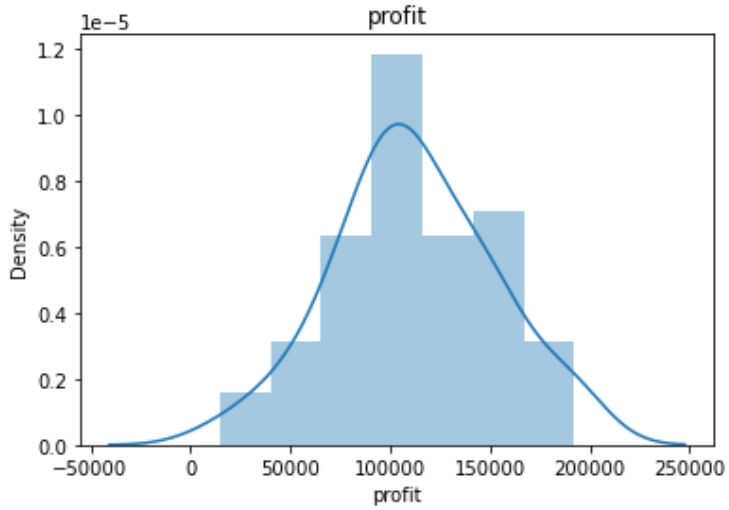
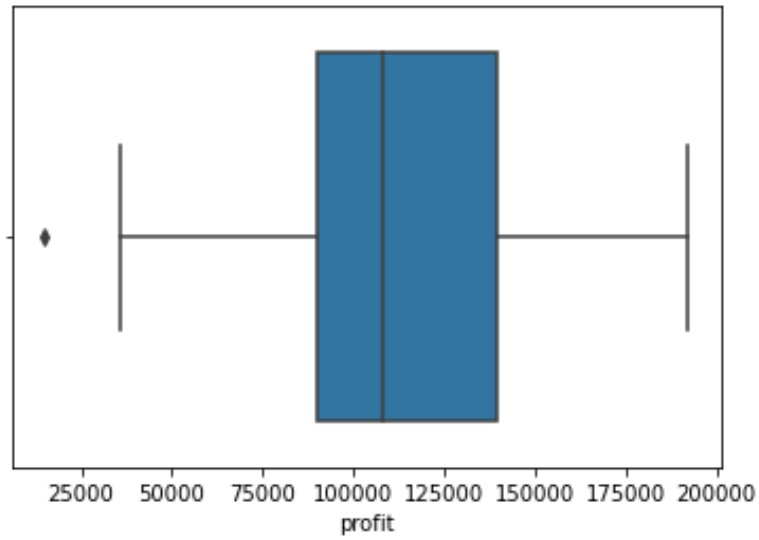
1. Variables:

- Numerical variable: R&D Spend, Administration, Marketing Spend, Profit

- Categorical variable: State

2. Distribution of Profit (output)

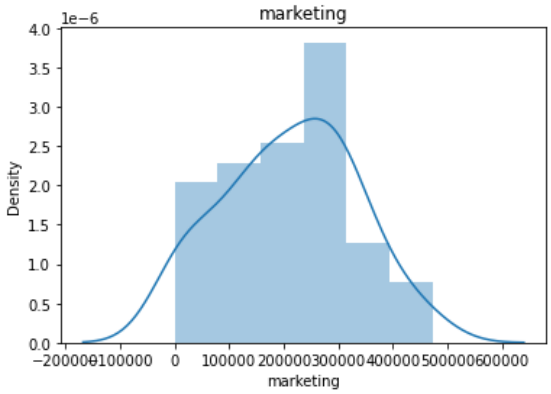




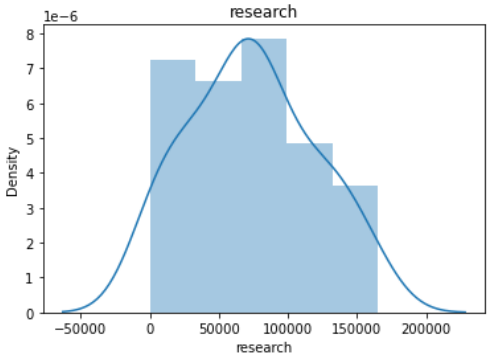
Profit shows a normal distribution with 1 outlier.

3. Distribution of numerical variables

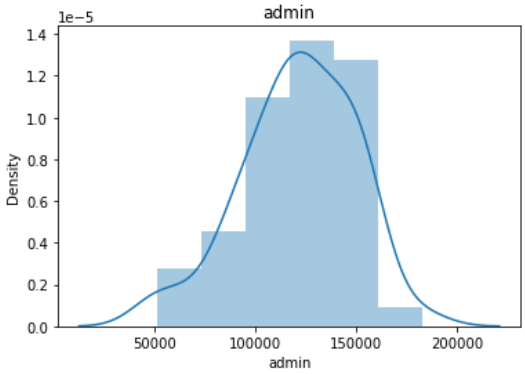
i) Marketing Spend



ii) Research Spend – Highest frequency around 50k – 100k



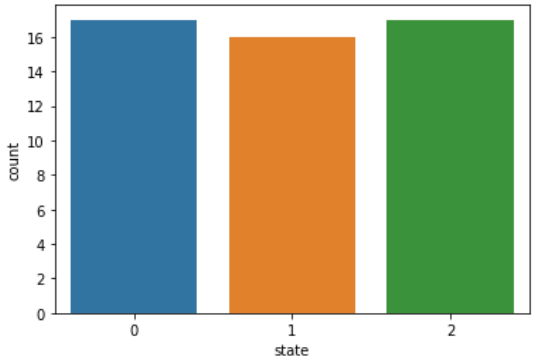
ii) Administration

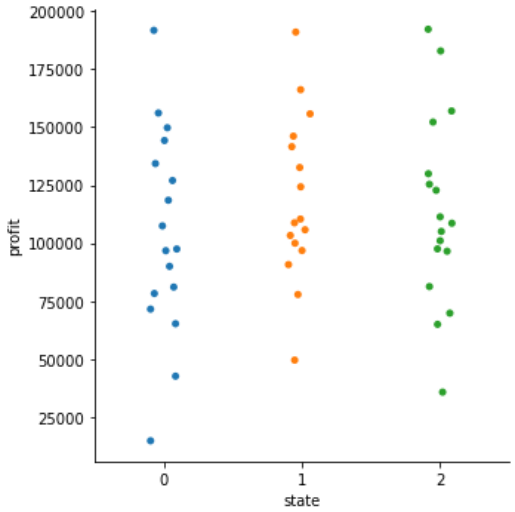


4. Multicollinearity between variables based on heatmap:

- Profit is highly correlated with Research Spend and Marketing Spend
- Correlation between Research and Marketing is 0.72

5. Categorical Variable – State (California – 0, Florida – 1, New York – 2)





Profit range are quite the same across all state, so the feature can be removed from model.

Model with state included	$R^2 = 0.946$
Model without state included	$R^2 = 0.948$
Model after removing outlier row (Profit)	$R^2 = 0.959$