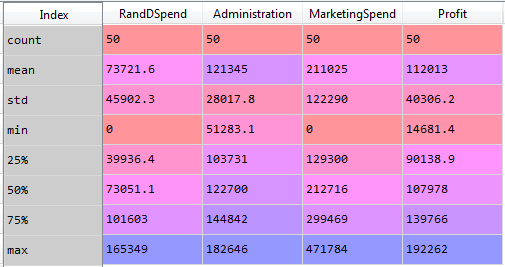
**Case Study of 50StartUps**

**Problem Statement:**

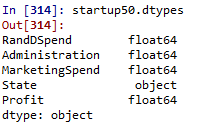
Prepare a prediction model for profit of 50\_startups data. Do transformations for getting better predictions of profit

**Exploratory data analysis:**

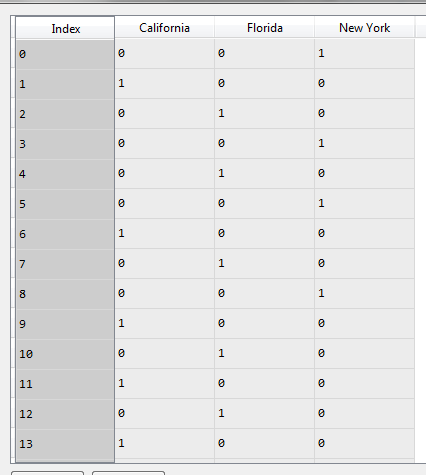


The above is the descriptive statistics of the 50 Startups dataset. There are 50 observations and 5 variables in the data frame. The missing variable in the above descriptive analysis is the “State” variables. Which is a categorical variable.

The following are the variable types:

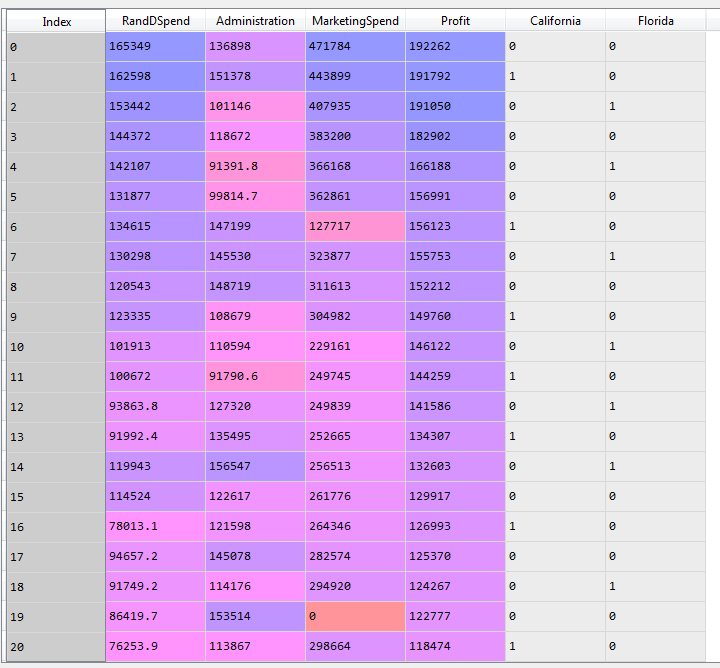


The categorical variable “State” is converted to the dummy variables of three named “California”, “Florida” and “New York” as follows:

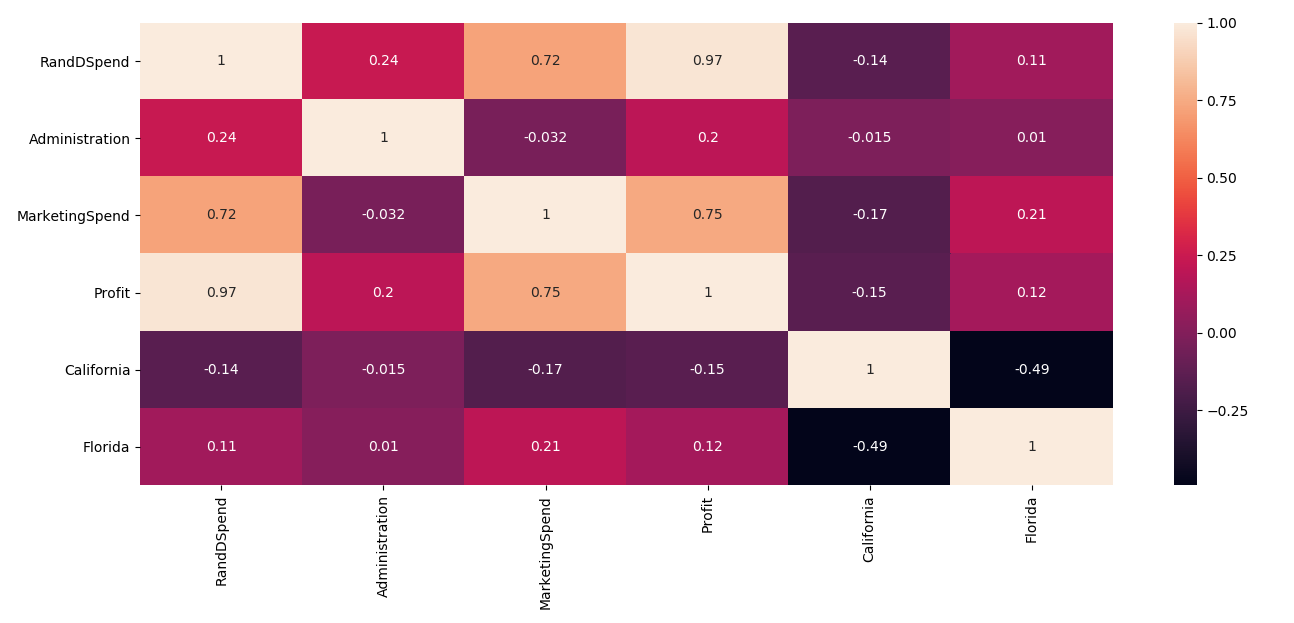


The variable “New York” is dropped including the categorical variable “State”. And the rest of the dummy variables “Florida” and “California” is concatenated to the main data frame of 50Startups.

Finally, the dataset looks like the following:



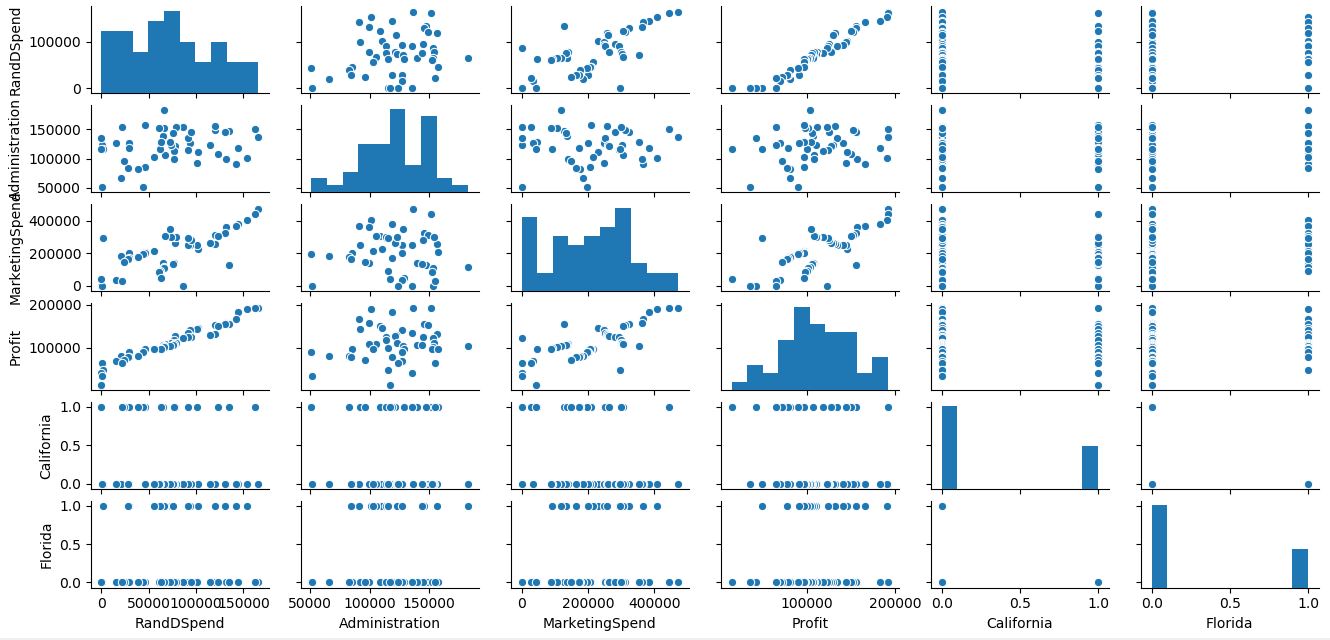
The heat map of the correlation between different variables



From the above heat map we can estimate the following:

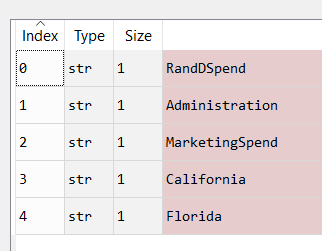
* High negative correlation between Profit and R&D Spend
* The correlation between independent variables are no significant enough to consider Multi Co linearity problem in the given dataframe.

The following is the pairplot of the dataframe:



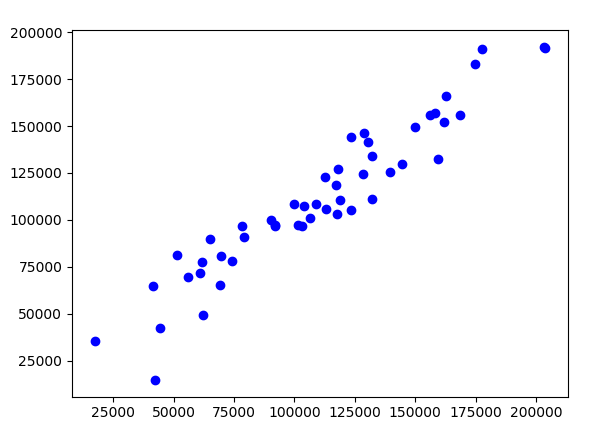
Building the sequential neural network model:

The following variables are set as the predictors:



And the target variable is set to “Profits” .

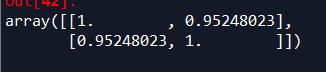
We get the prediction values of the Profits variable and we evaluate the model by looking at the following scatter plot:



The above scatter plot is between the predicted values and the existing values of the variable Profits.

* There is an upward trend which means that it is positively co-related
* The points are tightly packed which means that there is less difference between the predicted and existing values
* The spread of the points is very narrow

The following is the co-relation value:



The co-relation value is 0.95, this means that the model is 95% accurate.