Startups

Predicting Startups Profit

Introduction

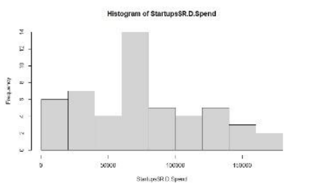
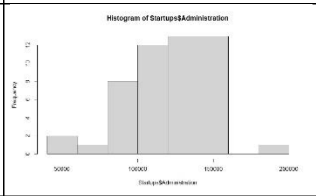
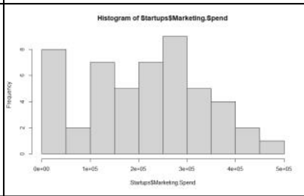
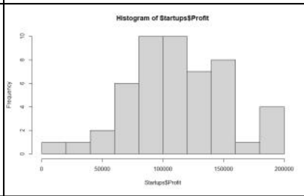
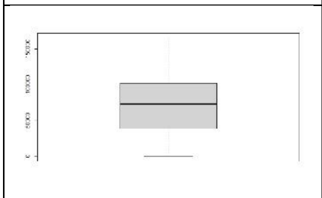
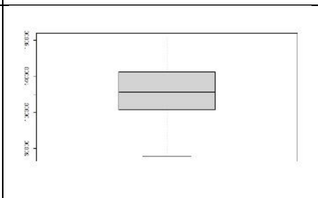
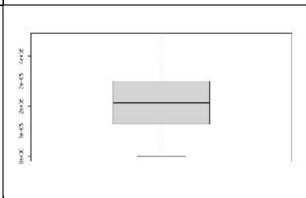
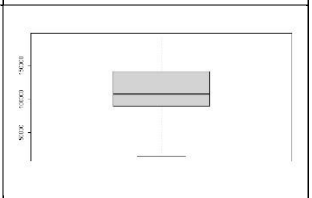
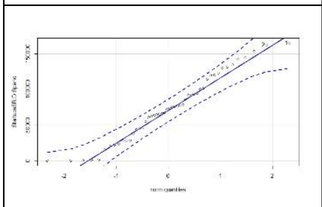
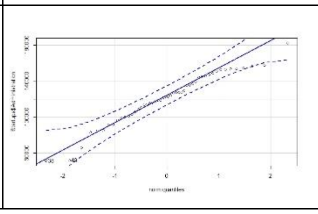
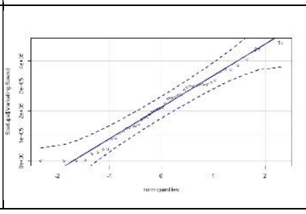
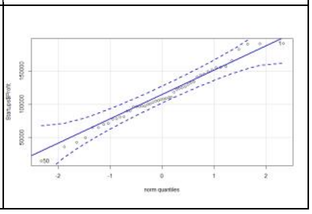
It is hard to say which startup will succeed and which one is doomed to failure. However, a study of the financial component of a startup can partially answer this question. That is why we set a goal to study the financial success of startups, namely, to calculate the impact of changes in the cost structure of organisations (marketing, R&D, and administrative costs) on their future profits. Such analysis can also be the basis for making a number of managerial decisions, such as "creating a startup's cost structure at the initial stage to maximise profits", "choosing the state of startup creation and registration", etc.

The work is carried out by using statistical analysis with R programming language in RStudio skills. The dataset has been downloaded from the website https://www.kaggle.com.

Table of contents

Primary Data Analysis and Descriptive Statistics

Costs and Profit	R&D	Administration	Marketing	Profit
Measures of location				
Mean	73722	121345	211025	112013
Median	73051	122700	212716	107978
Quantiles	Q1 = 39936 Q3 = 101603	Q1 = 103731 Q3 = 182646	Q1 = 129300 Q3 = 299469	Q1 = 90139 Q3 = 139766
Measures of variation				
Range	165349.2	131362.4	471784.1	177580.4
Interquartile range	61666.43	41111.3	170169	49627.07
Variance	2107017150	784997271	14954920097	1624588173
Standard deviation	45902.26	28017.8	122290.3	40306.18
Relative measures of variation				
Coefficient of variation	0.6226431	0.2308944	0.579506	0.359836
The distribution shape				
Skewness (A)	0.1542932	- 0.4600745	- 0.04372111	0.02191219
Kurtosis (E)	- 0.891987	- 0.03664891	- 0.814161	- 0.2871546
Plots of the variables				

R&D	Administration	Marketing	Profit
			
			
			

Identification of Outliers

In this section, we use two algorithms for identifying outliers (Tukey's range test and z-score method) on the dataset and compare their performance results. Comparing the two methods of outlier identification, we conclude that the z-score method is 'softer', as it considers an outlier only those values that are very different from the other values in the sample, while the Tukey's method is more rigorous and usually identifies more outliers. We have identified one outlier in the dataset using Tukey's range test; however, we checked the value that was counted as an outlier (a startup with a value in the Profit column equal to 14681.40) and consider this deviation to be insignificant.

From Sample to Population

The estimation of distribution parameters can be either point or interval. The former is carried out in a majority of cases, while the latter depends on the type of distribution (it must be normal for confidence intervals to be valid).

Determining the Distribution Type

To check the normality of the distribution, we used the Pearson and Shapiro-Wilk criteria. The tests showed that the values of R&D Costs, Administration Costs, Marketing Costs and Profit are distributed normally.

Point Distribution Estimates

Costs and Profit	R&D	Administration	Marketing	Profit
Expected Value	73721.62	121344.6	211025.1	112012.6
Variance	2107017150	784997271	14954920097	1624588173

Interval Distribution Estimates (significance level 0.05)

Costs and Profit	R&D	Administration	Marketing	Profit
Expected Value	(60676.34; 86766.89)	(113382.1; 129307.2)	(176270.6; 245779.6)	(100557.7; 123467.5)
Variance	(1470240556; 3271878108)	(547757679; 1218981718)	(10435287646; 23222723015)	(1133609861; 2522739066)

We also construct confidence intervals for the probability of a startup being located in a particular state using the prop.test function:

California	(0.2243695; 0.4784617)
Florida	(0.2075822; 0.4581030)
New York	(0.2243695; 0.4784617)

From the data we can conclude that the sample was collected by region (i.e., not randomly) so that the shares of the number of startups in each region were close to each other.

Hypothesis Testing

Parametric Statistical Tests (significance level 0.05)

Hypothesis	Conclusion
The average profit of a startup is \$112013 (the expected value of the general population is 112013). *H0: μ = 112013, H1: μ ≠ 112013 (p-value = 0.9999, H0 is accepted).*	The average profit of a startup is \$112013
The variance of profit relative to the mean is equal to 1624588173 (variance is equal to 1624588173). *H0:* σ^2 = 1624588173, *H1:* σ^2 ≠ 1624588173 *(p-value = 0.9463, H0 is accepted).*	The spread of values is 1624588173 and the deviation from the mean value of startup profits is \$40306.18
The average value of administration costs is equal to (or less than) the average value of marketing costs. *H0: μ1 = μ2, H1: μ1 ≠ μ2 (p-value = 3.617e-06, H0 is rejected).*	Administrative costs are lower than marketing costs
The average value of R&D costs is equal to (or less than) the average value of administration costs. *H0: μ1 = μ2, H1: μ1 ≠ μ2 (p-value = 2.599e-09, H0 is rejected).*	R&D costs are lower than administrative costs

The shares of successful startups with Profit > $100000 are equal in California and New York. *H0: μ1 = μ2, H1: μ1 ≠ μ2 (p-value = 0.4897, H0 is accepted).*	The shares of successful startups are equal

Nonparametric Statistical Tests

Hypothesis	Conclusion
The distributions of different types of costs are the same. *H0: F1(x) = F2(x), H1: F1(x) ≠ F2(x) (p-value for all the tests is less than the significance level, H0 is rejected).*	The distributions of different types of costs vary
The distributions of profit are similar across states (Kraskell-Wallis/Mood criterion). *H0: all samples are drawn from one population; H1: at least one sample is drawn from a different population.*	The distributions of startup profits are similar across the states
The distributions of costs are similar across states (Kraskell-Wallis/ Mood criterion). *H0: all samples are drawn from one population; H1: at least one sample is drawn from a different population.*	The distributions of startup profits are similar across the states

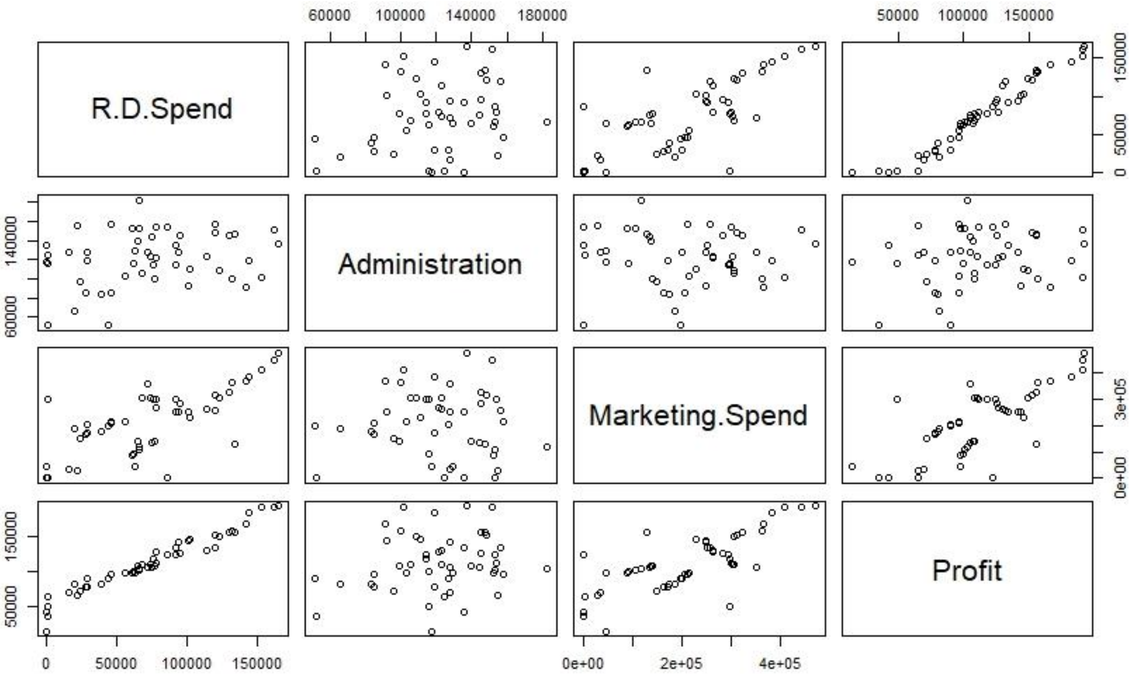
Analysis of Variance

By conducting an analysis of variance, we confirmed that the profits and different types of costs of startups are independent from the state in which they were established.

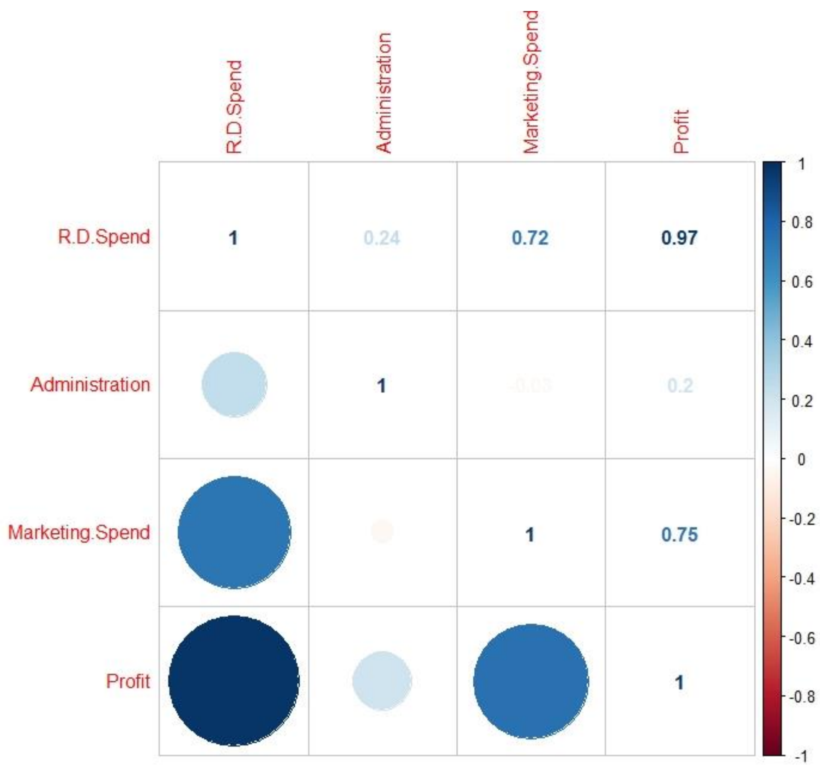
Construction of Linear Regression Models

Dependency Analysis

In order to analyse the dependencies, we have constructed pairwise scatter diagrams. The figure below shows that the strongest relationship is between R&D costs and profit (direct linear relationship). The relationship between R&D Costs and Marketing Costs, Marketing Costs and Profit is slightly weaker.



Let us confirm this by analysing the correlation matrix:



The statements described above are confirmed. Thus, it is possible to build a regression model.

Paired Model (significance level 0.01)

Starting with the paired model, for its construction we decided to take Profit as a dependent variable and R&D Costs as an independent variable, since we found the strongest relationship between them.

$$y = 4.903^{+4} + 8.543^{-1} * x$$

Checking the fulfilment of the basic assumptions of regression analysis:

- Residuals are normally distributed;
- The expected value of the residuals is zero;
- The random variables have the same variance;
- The Breusch-Pagan criterion did not confirm the absence of autocorrelation.

Checking the model for its statistical significance and quality:

- The coefficients are statistically significant;
- The model is statistically significant;
- The average approximation error is 11.07014.

The average approximation error is greater than 10, which means that this model can only be trusted with caution. It is possible that the error would be smaller if more explanatory variables were added, and therefore a multiple regression model could be built.

Multiple Model (significance level 0.01)

Для выбора объясняющих переменных для множественной модели мы воспользовались функцией regsubsets и определили, что наиболее точную модель получим при использовании двух объясняющих переменных (R.D.Spend и Marketing.Spend).

$$y = 4.698^{+4} + 7.966^{-1} * R\&D\ Costs + 2.991^{-2} * Marketing\ Costs$$

Checking the fulfilment of the basic assumptions of regression analysis:

- Residuals are normally distributed;
- The expected value of the residuals is zero;
- The random variables have the same variance;
- The Breusch-Pagan criterion did not confirm the absence of autocorrelation.

Checking the model for its statistical significance and quality:

- The coefficient of Marketing Costs is statistically insignificant, all the other coefficients are statistically significant;
- The model is statistically significant;
- The average approximation error is 10.60871;
- There is no multicollinearity between independent variables (VIF).

The average approximation error has become smaller, but still exceeds the threshold value of 10. Therefore, this model does not fulfil certain criteria, but still, as we believe, it is applicable in practice, taking into account the possible error.

Multiple Model Considering Qualitative Variables

$$y = 4.875^{+4} + 8.53^{-1} * R\&D\ Costs + 1.164^{+3} * z1 + 9.597 * z2$$

California: $y = 4.875^{+4} + 8.53^{-1} * R\&D\ Costs$

Florida: $y = 4.875^{+4} + 8.53^{-1} * R\&D\ Costs + 1.164^{+3}$

New York: $y = 4.875^{+4} + 8.53^{-1} * R\&D\ Costs + 9.597$

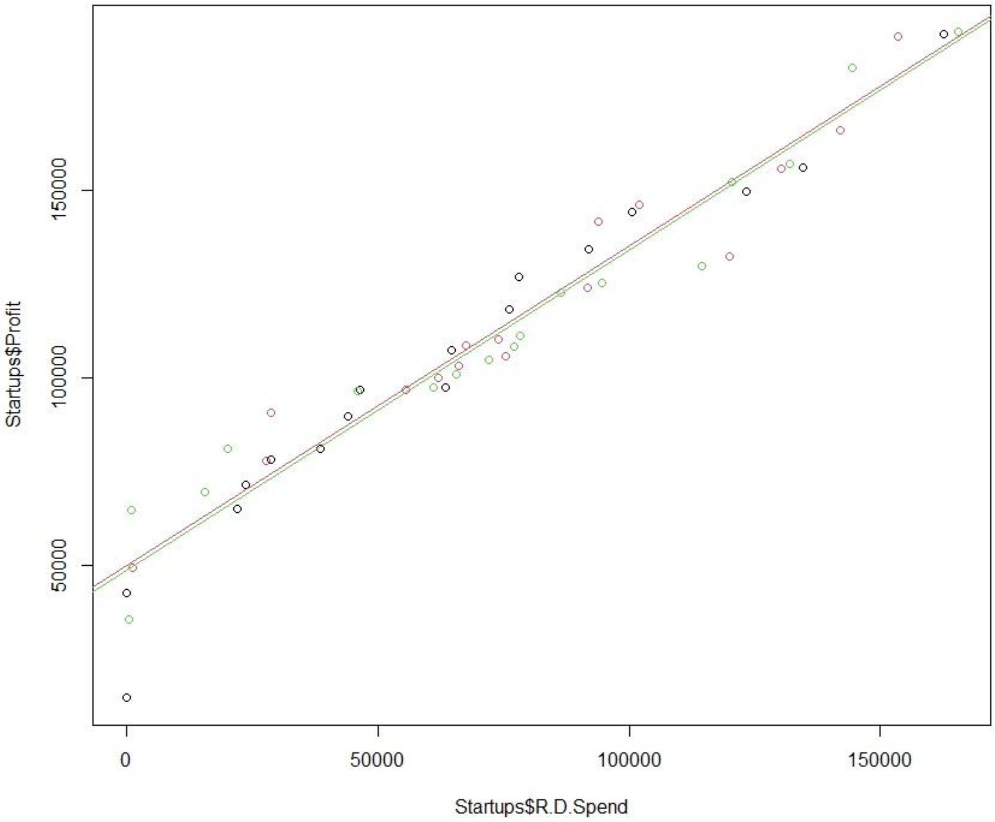
Checking the fulfilment of the basic assumptions of regression analysis:

- Residuals are normally distributed;
- The expected value of the residuals is zero;
- The random variables have the same variance;
- The Breusch-Pagan criterion did not confirm the absence of autocorrelation.

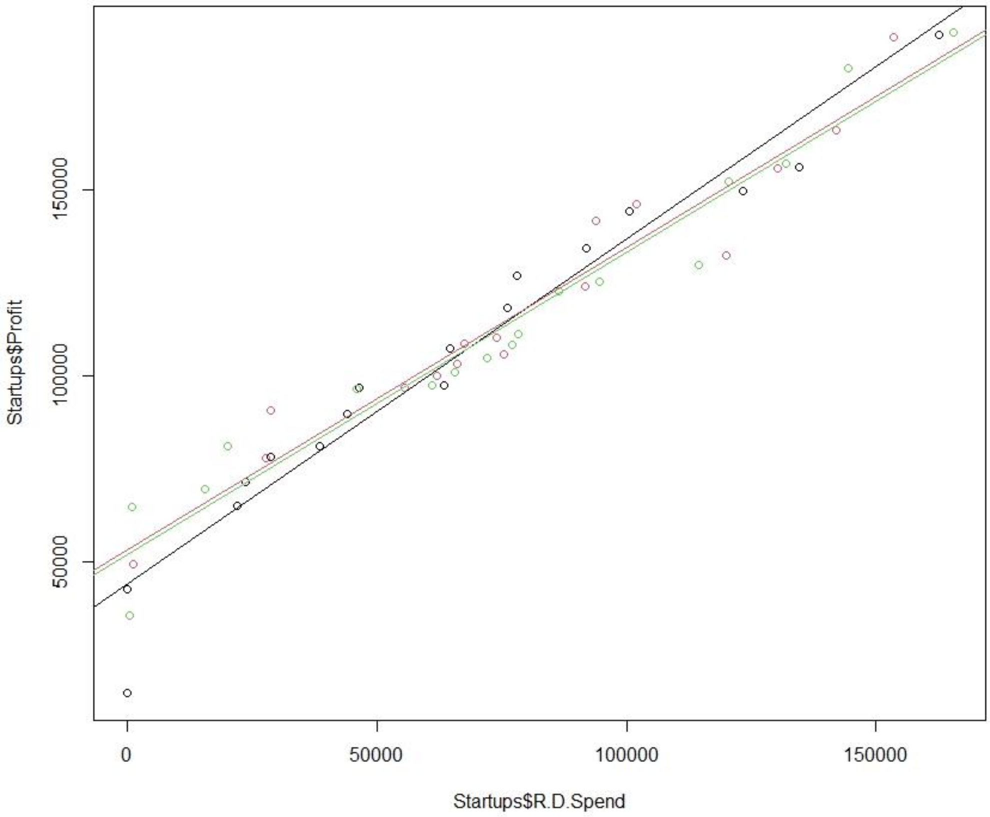
Checking the model for its statistical significance and quality:

- The coefficient of State is statistically insignificant, all the other coefficients are statistically significant;
- The model is statistically significant;
- The average approximation error is 11.08748;
- There is no multicollinearity between independent variables (VIF)

As we can see, the average approximation error has increased, which raises suspicions about the quality of the model. Therefore, we decided to compare the graphs:



Based on the graph, we can assume that the independent variable State is not statistically significant. Thus, let us check this by building separate models.



Based on the graphs of the three different models based on the state of location of startups, we can assume that the difference between them is not statistically significant. Let us test this assumption using Chow's criterion.

H0: the difference of coefficients of individual regression models is statistically insignificant (p-value is greater than the significance level, H0 is accepted).

The implementation of Chow's criterion confirmed the hypothesis that the difference in the coefficients of the models is statistically insignificant. Thus, we conclude that the model taking into account qualitative variables is not applicable in practice.

Conclusion

In the performed project, we have collected, processed, analysed and visualised data, proposed relevant hypotheses and made a reasoned choice of a suitable alternative. In the course of our work, we achieved the goal we set in the introduction and studied the financial success of start-ups, namely, we calculated the impact of changes in the cost structure of startups (marketing costs, R&D costs, and administrative costs) on their profits in the three states (California, New York, and Florida) of the United States.

The most important part of our work was the construction of regression models. We concluded that the most suitable model for use was the multiple regression model that takes into account such independent variables as R&D Costs and Marketing Costs, as it meets the largest number of criteria and its average approximation error is the smallest. To give an example, with marketing costs of $9500 and R&D costs of $21000, the possible revenue is $63988.25. However, it can vary between $44875.64 and $83100.87. Nevertheless, it is worth remembering that there are no perfect models, and even when using this one it is worth being careful about the obtained results.

References

1. Startup - Multiple Linear Regression / kaggle // URL: https://www.kaggle.com/karthickveerakumar/startup-logistic-regression
2. Jank W. Business Analytics for Managers. Springer-Verlag New York, 2011.
3. Linear Regression // URL: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
4. America's Best Startup Employers / Forbes // URL: https://www.forbes.com/americas-best-startup-employers/#47e0c95e6527
5. Digital Dolina // URL: https://digital-dolina.ru/