**NAGARJUNA COLLEGE OF ENGINEERING AND TECHNOLOGY**

(An Autonomous College under VTU, Belagavi, Accredited by NAAC with “A+” Grade)



**INTERNSHIP REPORT**

**ON**

**PROJECT CODE:P16**

**TOPIC: 50\_STATRUPS SUCCESS RATE PREDICTION**

**BACHELOR OF ENGINEERING**

In

**COMPUTER SCIENCE AND ENGINEERING**

**BY**

**VINAYAK GODI (1NC21CS122)**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**NAGARJUNA COLLEGE OF ENGINEERING AND TECHNOLOGY**

(An Autonomous Institution Approved by AICTE, affiliated to VTU, Belagavi-590018)

**VENKATAGIRI KOTE, DEVANAHALLI, BENGALURU, KARNATAKA**

**(2022 – 2023)**

**CHAPTER 1**

**INTRODUCTION**

Data analysis is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision making. Every product that is manufactured is supposed to have distinguishing physical characteristics, which makes it attractive and provides usefulness and value to customers; these characteristics are known as Design, and the process employed in this regard is known as product design. Product designs clearly defines a problem , develops a proper solution for that problem and validates the solution with the real users Product design is the process of creating a new product to be sold by a business to its customers. It is essentially the efficient and effective generation and development of ideas through a process that leads to new products.

Machine learning is a subset of artificial intelligence (AI) that involves the development of algorithms and statistical models that enable computers to learn and improve from experience without being explicitly programmed. The goal of machine learning is to enable computers to identify patterns in data, make predictions, and take actions based on the information they have learned.

In this internship program we have done a project.

* Project: 50\_STARTUPS SUCCESS RATE PEDICTION PREDICTION
* The software used in designing is Jupyter Notebook.

**CHAPTER 2**

**TOOLS EXPOSED**

Jupyter notebook:

Jupyter is an open-source interactive computing tool that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It is widely used in data science, machine learning, scientific research, and education. The name "Jupyter" is a combination of three core programming languages supported by the tool: Julia, Python, and R.

A notebook kernel is a computational engine that executes the code contained in a notebook document. The jupyter kernel referenced in this guide executes python code. Kernels for many other language exist. When you open a notebook document the associated kernel is automatically launched. When the notebook is executed the kernel performs the computation and produces the results. Depending on the type of computations the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel is shut down. The notebook dashboard is the component which is shown first when you launch jupyter notebook app.

The notebook dashboard is mainly used to open notebook documents and manage the running kernels. The jupyter notebook extends the console based approach to interactive computing in a qualitatively new direction, providing a web based application suitable for capturing the whole computation process: developing, computing and executing code as well as communicating the results. The jupyter notebook combines two components a web application and notebook documents.

**CHAPTER 3**

**TASK PERFORMED: DATA ANALYSIS OF SOFTWARE SALARIES.**

1. Importing Libraries.
2. Reading and Exploring the Dataset.
3. Data Visualization.
4. Data Preprocessing and Splitting.
5. Building and Training the Linear Model.
6. Making Predictions and Evaluating the Model.
7. Making a Single Prediction.
8. Additional Data Visualization.
9. Data Distribution Visualization.

Why Data visualization?

Data visualization is vital for effective communication and decision- making. It helps convey complex information in a clear and concise manner, making it easier to identify patterns, trends, and outliers in data. Visualizations support exploratory data analysis, aid in storytelling with data, and facilitate cross-domain understanding.

They play a key role in data cleaning, preprocessing, and managing large datasets. Interactive visualizations enable personalized exploration and comparisons, empowering users to make informed decisions based on data insights. Overall, data visualization enhances data analysis, promotes data literacy, and transforms data into actionable knowledge for diverse stakeholders.

**CHAPTER 4**

**RESULTS AND DISCUSSIONS**

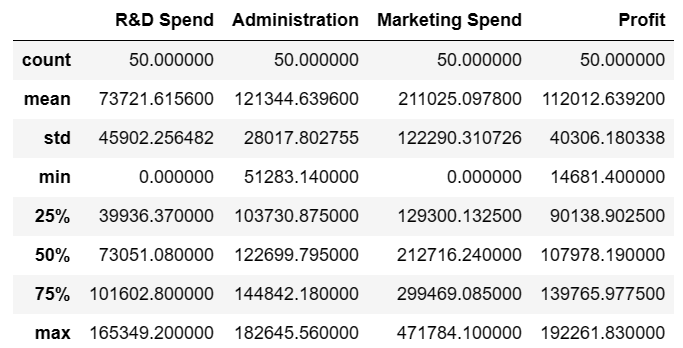
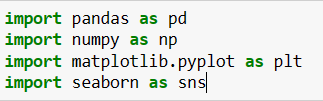
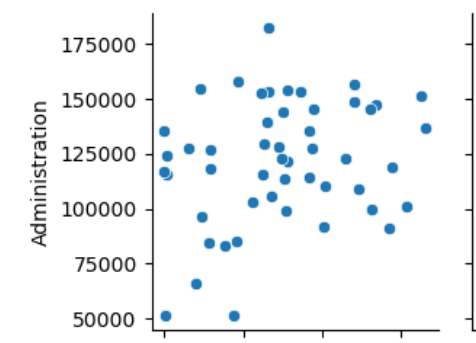
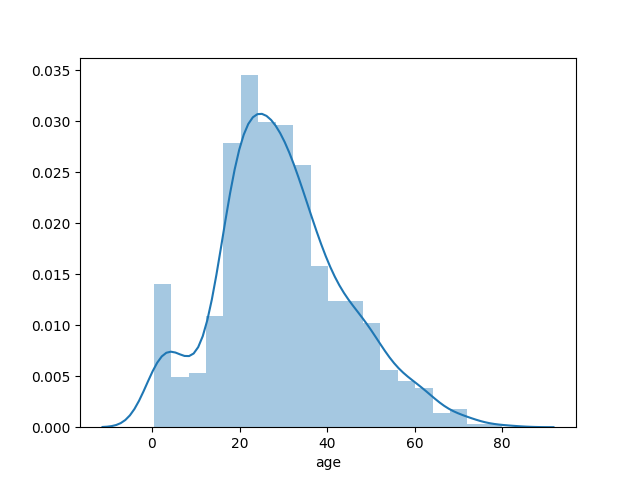
****

figure 5.1 Data set

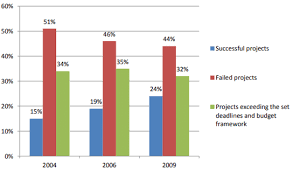
Plots for different data: 1.Scatter plot:



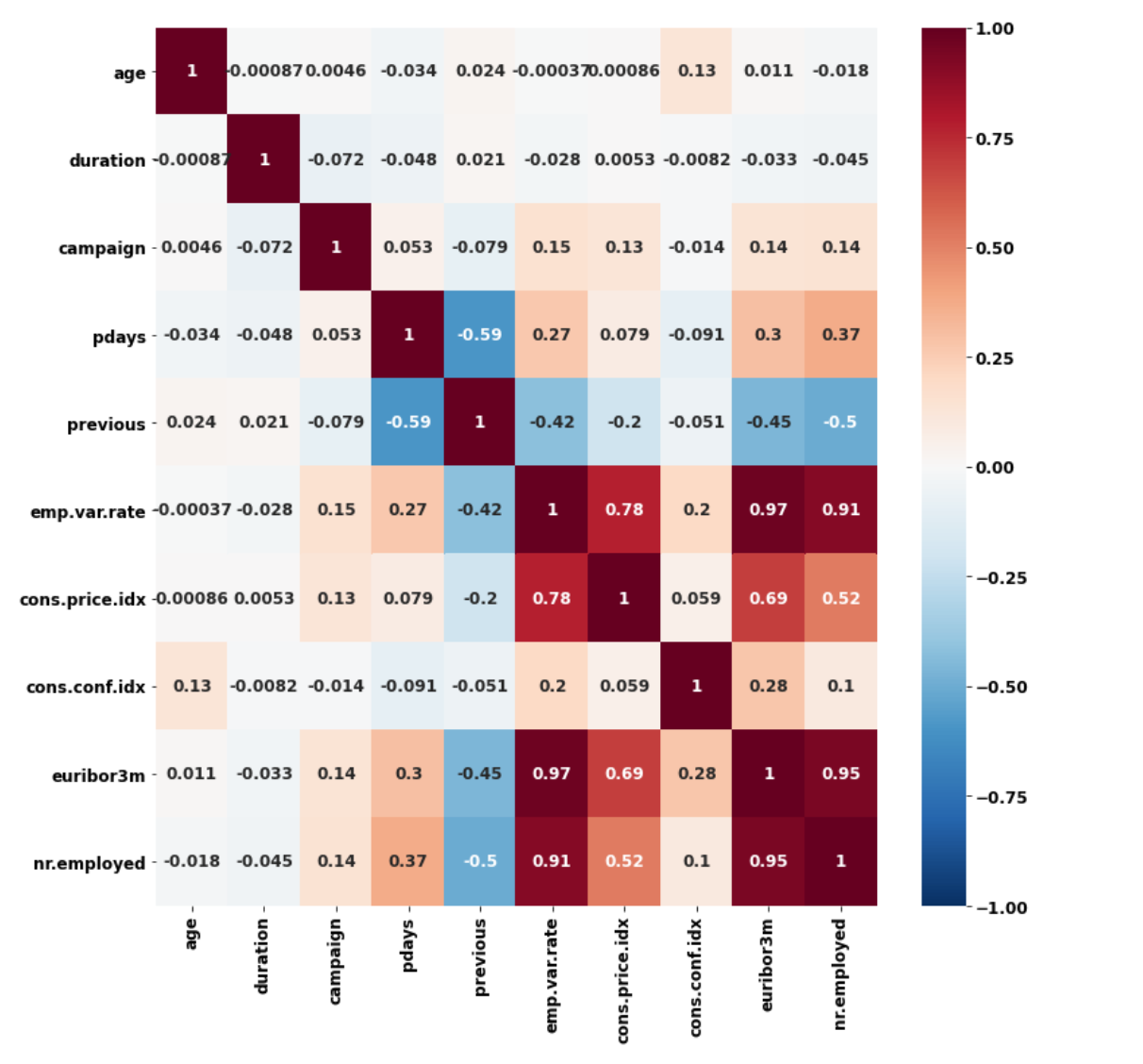
1. Distplot



1. Barplot:



1. HeatMap:



**CHAPTER 5**

* 1. **Skills acquired**

**REFLECTION NOTES**

Proficiency in using data analysis libraries such as pandas and numpy is necessary to read, manipulate, and preprocess the data. These libraries offer essential functionalities for data handling and numerical computations.

Knowledge of data visualization libraries, particularly matplotlib and seaborn, is required to create various plots and graphs for visualizing data patterns and relationships.

Understanding the concepts and algorithms of machine learning is important for model building and training. Familiarity with libraries like scikit-learn for implementing machine learning models is beneficial.

A good grasp of statistical concepts such as regression, correlation, and data distribution is essential for interpreting results and making data-driven decisions.

Skills in data handling and manipulation are vital for tasks like data cleaning, feature engineering, and splitting datasets into training and testing sets.

**CHAPTER 6**

**CONCLUSION**

In this project-based internship, the main objective was to develop a predictive model for estimating employee salaries based on their years of experience. The project involved various steps, starting with data exploration, where the "Salary.csv" dataset was loaded and examined to understand its structure and contents. Visualizations, such as scatter plots and distribution plots, were employed to gain insights into the relationship between years of experience and salary, and to understand the distribution of salary values.

In this project, I have learned valuable skills in data analysis and machine learning .I have started by loading and exploring the dataset using pandas, gaining insights into its contents and structure. Through data visualization using matplotlib and seaborn, I have visualized the relationships and trends within the dataset, enabling you to better understand the distribution of data and identify patterns.

Building a machine learning model was a significant part of the project. Due to this I chose a linear regression model, a fundamental technique for predicting continuous values, and trained it on the training data. Evaluating the model's performance on the test data using the R-squared score allowed to assess how well the model generalized to new data.

Overall, this project equipped you with essential data analysis, visualization, and machine learning skills, which are fundamental for various data-driven tasks across industries. The experience gained in interpreting and communicating results will prove invaluable when presenting insights to stakeholders and supporting data-informed decision-making.

**REFERENCES**

* + - “STARTUP SUCCESS RATE PRECITION" by NOAH, P., & SAMI, A. (Year). Journal of Data Science, 20(3), 345-362.
    - "Tips-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron. Titos O'Reilly Media, 2019.
    - <https://drive.google.com/drive/folders/12cuBEEnQuZLfpLlAAdAQfhY5D0frlSAp?usp=sharing>
    - <https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2F%40apoorvaudupa3%2Fbank>