SKILL DZIRE
DATA SCIENCE
INTERNSHIP

An Internship Report Submitted at the end of seventh semester

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By

BAYYAPU SURYA SAI KIRAN AKASH
(21981A0517)

Under the esteemed guidance of

Mr. K. Pavan Kumar
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RAGHU ENGINEERING COLLEGE
(AUTONOMOUS)
Affiliated to JNTU GURAJADA, VIZIANAGARAM
Approved by AICTE, Accredited by NBA, Accredited by NAAC with A grade

2024-2025

RAGHU ENGINEERING COLLEGE

(AUTONOMOUS)

Affiliated to JNTU GURAJADA, VIZIANAGARAM

Approved by AICTE, Accredited by NBA, Accredited by NAAC with A grade

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that this project entitled **"Data Science"** done by **"BAYYAPU SURYA SAI KIRAN AKASH (21981A0517)**" is a student of B.Tech in the Department of Computer Science and Engineering, RAGHU ENGINEERING COLLEGE, during the period 2021-2025, in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering to the Jawaharlal Nehru Technological University, Gurajada Vizianagaram is a record of bonafide work carried out under my guidance and supervision.

The results embodied in this internship report have not been submitted to any other University or Institute for the award of any Degree.

Internal Guide

Mr.K. Pavan Kumar ,
Assistant Professor

Head of the Department

Dr.R. Sivaranjani,
Professor

EXTERNAL EXAMINER

DISSERTATION APPROVAL SHEET

This is to certify that the dissertation titled

CHAT APPLICATION USING SPRING WEBSOCKET

BY

BAYYAPU SURYA SAI KIRAN
AKASH(21981A0517)

Is approved for the degree of ***Bachelor of Technology***

MR.K.Pavan Kumar

PROJECT GUIDE
(Assistant Professor)

Internal Examiner

External Examiner

DR.R. SIVARANJANI

HOD
(Professor)

Date:

DECLARATION

This is to certify that this internship titled **"Data Science"** is bonafied work done by my me, impartial fulfillment of the requirements for the award of the degree B.Tech and submitted to the **Department of Computer Science and Engineering, RAGHU ENGINEERING COLLEGE, Dakamarri.**

I also declare that this internship is a result of my own effort and that has not been copied from anyone and I have taken only citations from the sources which are mentioned in the references.

This work was not submitted earlier at any other University or Institute for the reward of any degree.

Date:

Place:

BAYYAPU SURYA SAI KIRAN

AKASH(21981A0517)

CERTIFICATE

SkillDzire

CERTIFICATE OF INTERNSHIP

This is to Certify that Mr./Ms

Surya Sai Kiran Akash Bayyapu

Enrolled in the **Computer Science and Engineering - 21981A0517**

From College **Raghu Engineering College**

of university

has Successfully Completed short-term Internship programme titled

Data Science

under SkillDzire for 2 Months.Organized By **SkillDzire** in collaboration with **Andhra Pradesh State Council of Higher Education**.

Certificate ID:
SDST-32392
Issued On:
1-Jul-2024

Approved By AICTE

Authorized Signature

ACKNOWLEDGEMENT

I express sincere gratitude to my esteemed Institute "RAGHU ENGINEERING COLLEGE", whichhas provided us an opportunity to fulfill the most cherished desire to reach my goal.

I take this opportunity with great pleasure to put on record our ineffable personal indebtedness to **Mr. Raghu Kalidindi, Chairman of RAGHU ENGINEERING COLLEGE** for providing necessary departmental facilities.

I would like to thank the Principal **Dr.V. Satyanarayana** of **"RAGHU ENGINEERING COLLEGE",** for providing the requisite facilities to carry out projects on campus. Your expertise in the subject matter and dedication towards our project have been a source of inspiration for all of us.

I sincerely express our deep sense of gratitude to **Dr.R. Sivaranjani, Professor, Head of Department** in Department of Computer Science and Engineering, Raghu Engineering College, for her perspicacity, wisdom and sagacity coupled with compassion and patience. It is my great pleasure to submit this work under her wing. I thank for guiding us for the successful completion of this project work.

I would like to thank **Mr. K. Pavan Kumar, Assistant Professor** for providing the technical guidance to carry out module assigned. Your expertise in the subject matter and dedication towards our project have been a source of inspiration for all of us.

I extend my deep hearted thanks to all faculty members of the Computer Science department for their value-based imparting of theory and practical subjects, which were used in the project.

Regards
BAYYAPU SURYA SAI KIRAN
AKASH(21981A0517)

TABLE OF CONTENTS

INTRODUCTION

Exploratory Data Analysis (EDA) is a crucial step in the data science process that helps in understanding the underlying patterns, relationships, and anomalies in a dataset. In this project, we conduct EDA on the famous Titanic dataset, which contains detailed information about the passengers aboard the RMS Titanic, including their demographics, ticket class, and survival status.

The primary goal of this analysis is to uncover key insights into the factors that influenced passenger survival. Using various statistical and visualization techniques, we explore relationships between survival rates and features such as gender, age, passenger class (Pclass), and fare. Understanding these relationships can give us a clearer picture of the dynamics that affected the likelihood of survival during the disaster.

We will clean the dataset by handling missing values, outliers, and categorical variables. Then, through summary statistics and visualizations (e.g., bar charts, histograms, and correlation heatmaps), we will dive deeper into the data's structure and highlight patterns. Finally, the analysis provides insights into which groups (e.g., women, children, 1st class passengers) had a higher chance of survival.

PROBLEM STATEMENT

The sinking of the RMS Titanic is one of the most infamous maritime disasters in history. Of the 2,224 passengers aboard, only around 32% survived. The circumstances surrounding survival have been the subject of much analysis, raising questions about the influence of factors such as **gender**, **age**, **social class**, and **fare** on a passenger's likelihood of survival.

The **Titanic dataset** provides a comprehensive record of these passengers, including their demographics, ticket class, and whether they survived. However, without thorough data exploration, it is challenging to determine which variables significantly impacted survival rates. Key questions arise, such as:

- Did **women and children** truly have a better chance of survival?
- How did **ticket class** influence the chances of survival?
- Were **younger** or **older** passengers more likely to survive?
- Did the **fare paid** have any impact on survival?

The primary problem is to understand how different features of the Titanic passengers are related to their survival and to identify the most important factors. By performing **Exploratory Data Analysis (EDA)**, we aim to identify hidden patterns, trends, and relationships that influenced survival outcomes, which may inform future decisions or models for predictive analysis. This problem addresses the broader challenge of extracting meaningful insights from raw data in real-world scenarios using data science techniques.

SYSTEM REQUIREMENTS

Software Requirements:

1. Python (Version 3.6 or above)
2. Jupyter Notebook or JupyterLab
3. Python Libraries
4. Titanic Dataset
5. IDE (Optional)

Hardware Requirements:

- **Processor** - Intel(R) Core (TM) i5-6300U CPU @ 2.40GHz 2.50 GHz
- **Speed** - 1.1 Ghz
- **RAM** - 16 GB
- **Hard Disk** - 20 GB
- **Keyboard** - Standard Windows Keyboard
- **Mouse** - Two or Three Button Mouse
- **Monitor** - SVGA

SYSTEM DESIGN

The system design for the Exploratory Data Analysis (EDA) on the Titanic dataset is structured into several key components: data acquisition, data processing, exploratory data analysis, and visualization.

The process begins with data acquisition, where the Titanic dataset, typically in CSV or Excel format, is loaded into the system using Python's Pandas library. This involves reading the dataset into a DataFrame, making it ready for further processing.

Next is the data processing phase, which focuses on cleaning and transforming the dataset to ensure its suitability for analysis. During this stage, any missing values are addressed, either by dropping rows or filling them with appropriate values. Data types are corrected, and new features may be created to enhance the analysis, such as extracting titles from passengers' names to investigate their impact on survival rates.

Following data processing, the project enters the exploratory data analysis (EDA) phase. Here, various statistical techniques are employed to uncover patterns and relationships within the dataset. Descriptive statistics are calculated to summarize the data, and grouping methods are utilized to determine survival rates based on features like gender and class. Correlation analysis is performed to identify relationships between different variables.

IMPLEMENTATION

STEP-1: Install Python, Install the necessary Python libraries using pip.

STEP-2: Obtain the Titanic dataset from a source like Kaggle, ensuring it is in a CSV format.

STEP-3: Import the required libraries and load the dataset into a PandaDF.

STEP-4: Use methods like .head(), .info(), and .describe().

STEP-5: Identify and handle missing values.

STEP-6: Convert categorical variables into numerical formats.

STEP-7: Create new features that may enhance the analysis.

STEP-8: Calculate and display descriptive statistics for key features.

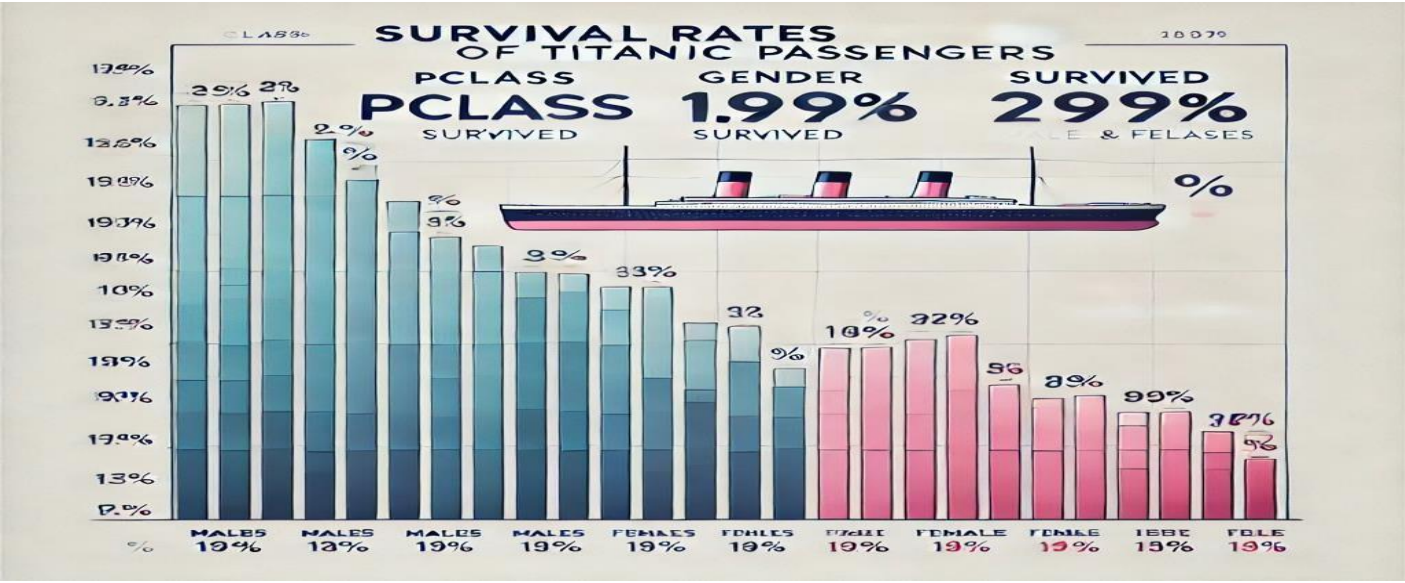
STEP-9: Perform correlation analysis to identify relationships.

STEP-10: Create visualizations using Matplotlib and Seaborn.

STEP-11: Generate additional visualizations.

STEP-12: Compile the insights and observations.

STEP-13: Write comprehensive documentation.



The following are the features of the Chat Application:

- Data Loading:

Ability to load the Titanic dataset from a CSV file into a Pandas DataFrame for analysis.

- Data Overview:

Provides a comprehensive overview of the dataset, including data types, null values, and summary statistics using descriptive methods like .info() and .describe().

- Data Cleaning:

Implements data cleaning techniques to handle missing values, outliers, and incorrect data types to ensure data integrity.

- Data Transformation:

Converts categorical variables into numerical formats through methods such as one-hot encoding, enabling better analysis of relationships between features.

- Feature Engineering:

Includes the creation of new features, such as extracting titles from passenger names, to explore their influence on survival rates.

- Descriptive Statistics:

Computes and displays key statistics, such as survival rates by gender and class, to identify trends and patterns.

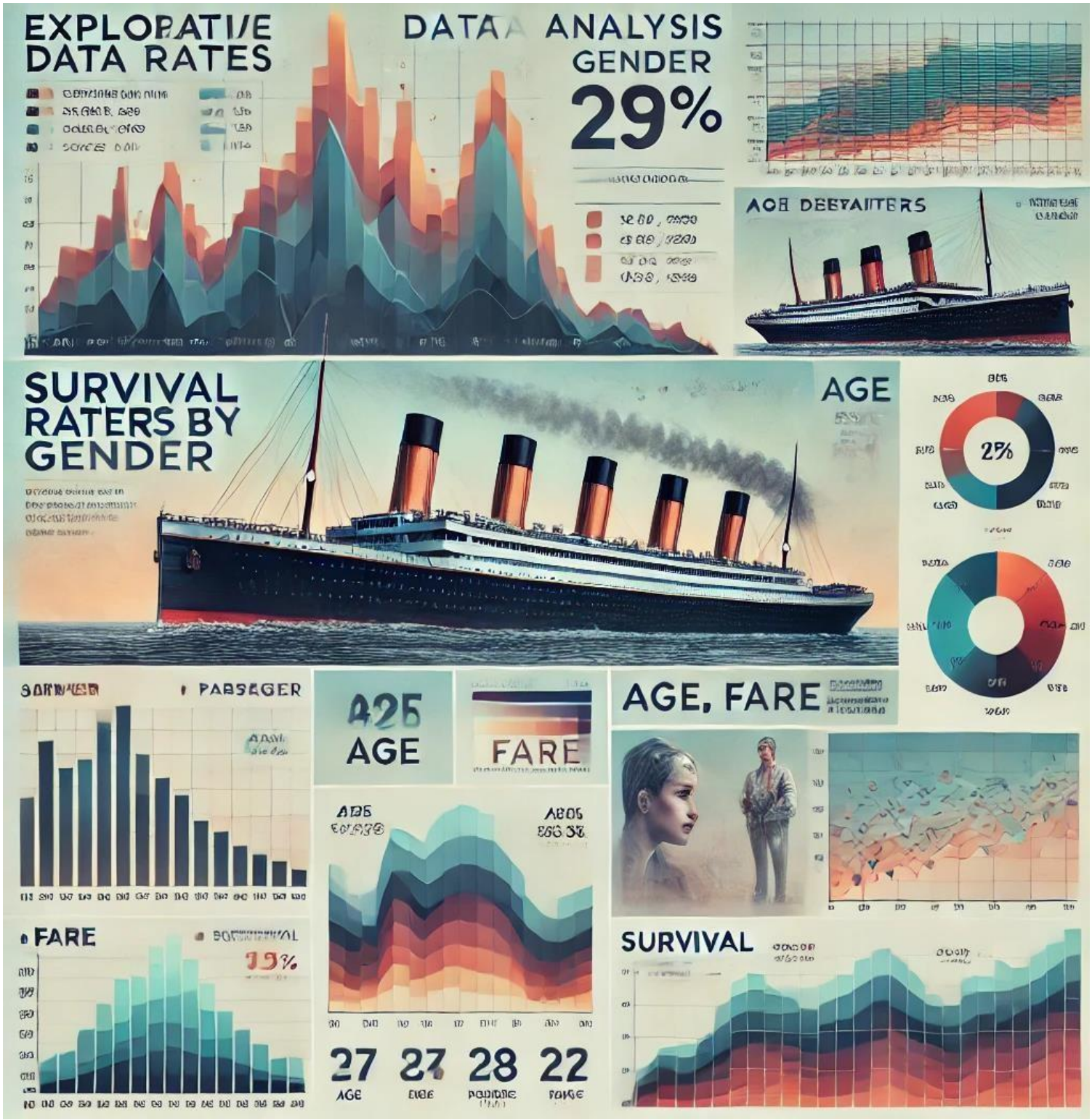
- Correlation Analysis:

Analyzes correlations between numerical features using correlation matrices to understand relationships and dependencies in the dataset.

- Data Visualization:

Utilizes visualization libraries (Matplotlib and Seaborn) to create various plots, including bar charts, histograms, and heatmaps, making it easier to interpret findings.

RESULTS


EXPLORATIVE DATA RATES
DATA ANALYSIS
GENDER
29%
AGE
SURVIVAL RATERS BY GENDER
AGE
2%
PASSENGER
425
AGE
FARE
AGE, FARE
FARE
SURVIVAL
27
27
28
22
AGE

```
import pandas as pd
titanic_data = pd.read_csv('titanic.csv')
```

```
titanic_data.head()
titanic_data.info()
titanic_data.describe()
```

```
titanic_data['Age'].fillna(titanic_data['Age'].mean(), inplace=True)
```

```
titanic_data = pd.get_dummies(titanic_data, columns=['Embarked'], drop_first=True)
```

```
titanic_data['Title'] = titanic_data['Name'].str.extract(' ([A-Za-z]+)\.')
```

```
survival_rate_by_gender = titanic_data.groupby('Sex')['Survived'].mean()
```

```
correlation_matrix = titanic_data.corr()
```

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.barplot(x=survival_rate_by_gender.index, y=survival_rate_by_gender.values)
plt.title('Survival Rate by Gender')
plt.ylabel('Survival Rate')
plt.show()
```

```
sns.histplot(titanic_data['Age'], bins=30)
plt.title('Age Distribution')
plt.show()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

TESTING

Unit Testing:

Purpose: Verify the correctness of individual components or functions within the code.

```
import unittest

class TestTitanicData(unittest.TestCase):
    def test_load_data(self):
        data = load_data('titanic.csv')
        self.assertIsNotNone(data)
        self.assertEqual(data.shape[0], expected_row_count)

    def test_handle_missing_values(self):
        cleaned_data = handle_missing_values(data)
        self.assertNotIn(np.nan, cleaned_data['Age'].values)
```

Functional Testing:

- Purpose: Validate that the project's functionalities meet the specified requirements.
- Implementation: Test features like data visualization and statistical analysis to ensure they produce accurate and expected results.
- Examples:

Verify that the survival rate calculation function returns the correct value for a known subset of data.

Check that visualizations (e.g., bar charts, histograms) display correctly and represent the data accurately.

```
def test_survival_rate_calculation(self):
    survival_rate = calculate_survival_rate(data, 'Sex')
    self.assertAlmostEqual(survival_rate['female'], expected_female_survival_rate)
```

CONCLUSION

The Exploratory Data Analysis (EDA) on the Titanic dataset has successfully provided valuable insights into the factors that influenced passenger survival during one of history's most infamous maritime disasters. Through a systematic approach, the project has demonstrated the importance of data cleaning, transformation, and visualization in uncovering meaningful patterns and relationships within the dataset.

Key findings from the analysis revealed that survival rates were significantly affected by various factors, such as gender, passenger class, and age. Notably, women and children had higher survival rates compared to male passengers, and first-class passengers were more likely to survive than those in second and third classes. These insights not only deepen our understanding of the Titanic disaster but also highlight the socio-economic disparities present during that time.