TITLE:

Comprehensive Data Science Project Report: Analysis, Modeling, and Strategic Insights

Tools: Python, Pandas, NumPy, Seaborn, Matplotlib, SciPy, Scikit-learn

1. EXECUTIVE SUMMARY

This report presents a comprehensive data science project integrating data acquisition, preprocessing, exploratory analysis, statistical testing, and machine learning model development. Publicly available datasets were used to identify patterns, validate hypotheses, and generate predictive insights. The project demonstrates an end-to-end analytical workflow designed to support data-driven decision-making. Based on the findings, strategic recommendations are proposed to enhance analytical practices and improve organizational outcomes.

2. INTRODUCTION

Purpose of the Project:

The objective of this project is to apply fundamental data science techniques to real-world datasets. The report consolidates learning outcomes across multiple stages, including data preparation, visualization, statistical analysis, and predictive modeling, ultimately translating insights into actionable recommendations.

3. PROJECT OVERVIEW

Week	Focus Area	Outcome
Week 1	Data Acquisition & Cleaning	Prepared dataset for analysis
Week 2	Visualization & Storytelling	Identified patterns
Week 3	Hypothesis Testing	Validated assumptions
Week 4	Machine Learning	Built predictive model
Week 5	Strategic Reporting	Generated recommendations

4. DATA ACQUISITION, CLEANING & PREPROCESSING (WEEK 1)

The project began with acquiring the Titanic dataset from a publicly available source. Initial exploration was conducted to understand data structure, variable types, and overall quality. Several preprocessing steps were performed, including handling missing values, correcting inconsistencies, and preparing the dataset for analysis. These steps ensured reliability and improved the effectiveness of subsequent analytical processes.

Screenshots:

```
[3]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

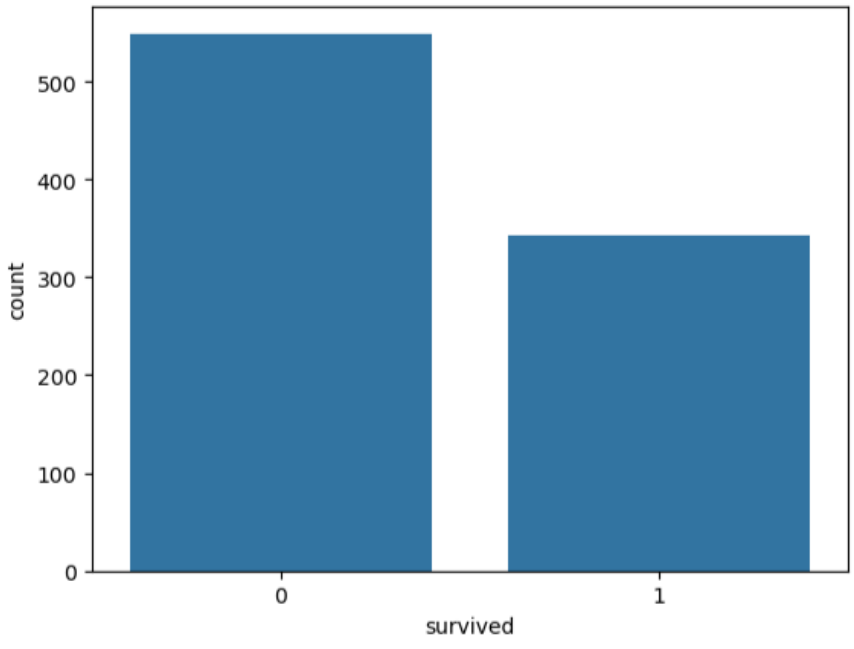
```
[8]: df.describe()
```

[8]:

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
[7]: df.isnull().sum()

[7]: survived          0
     pclass            0
     sex               0
     age             177
     sibsp             0
     parch             0
     fare              0
     embarked          2
     class             0
     who               0
     adult_male        0
     deck            688
     embark_town       2
     alive             0
     alone             0
     dtype: int64
```

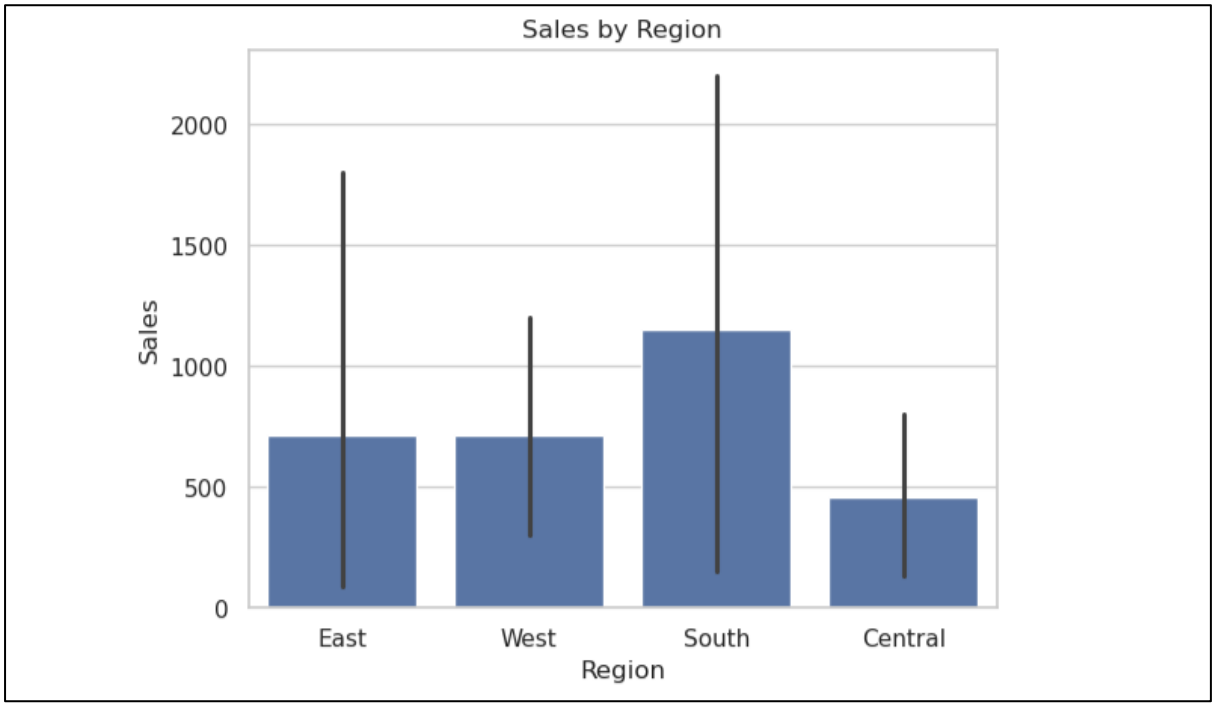
```
[14]: sns.countplot(x='survived', data=df)
      plt.show()
```

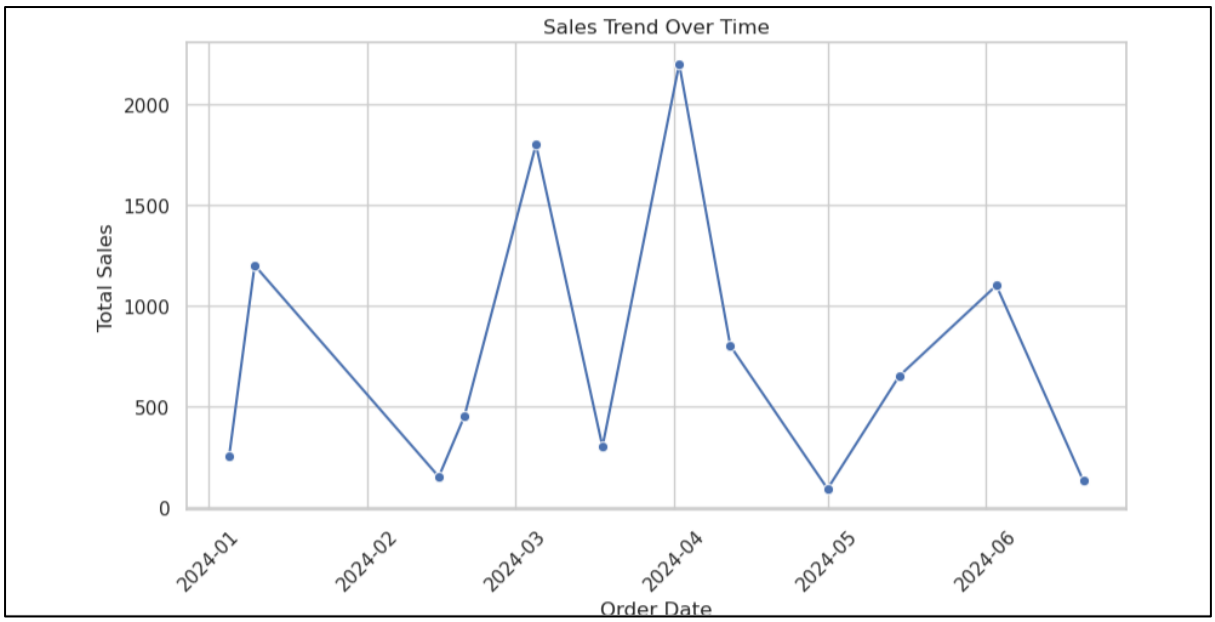


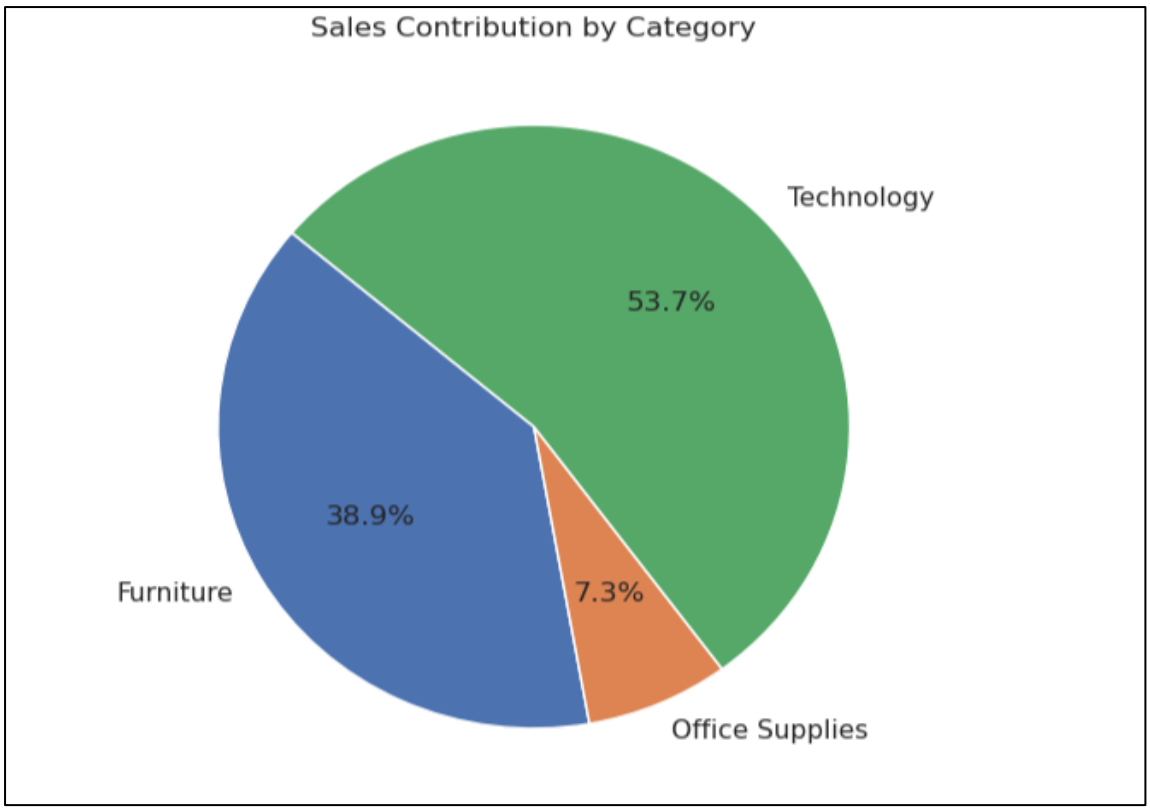
Effective preprocessing is critical because data quality directly impacts analytical accuracy and model performance.

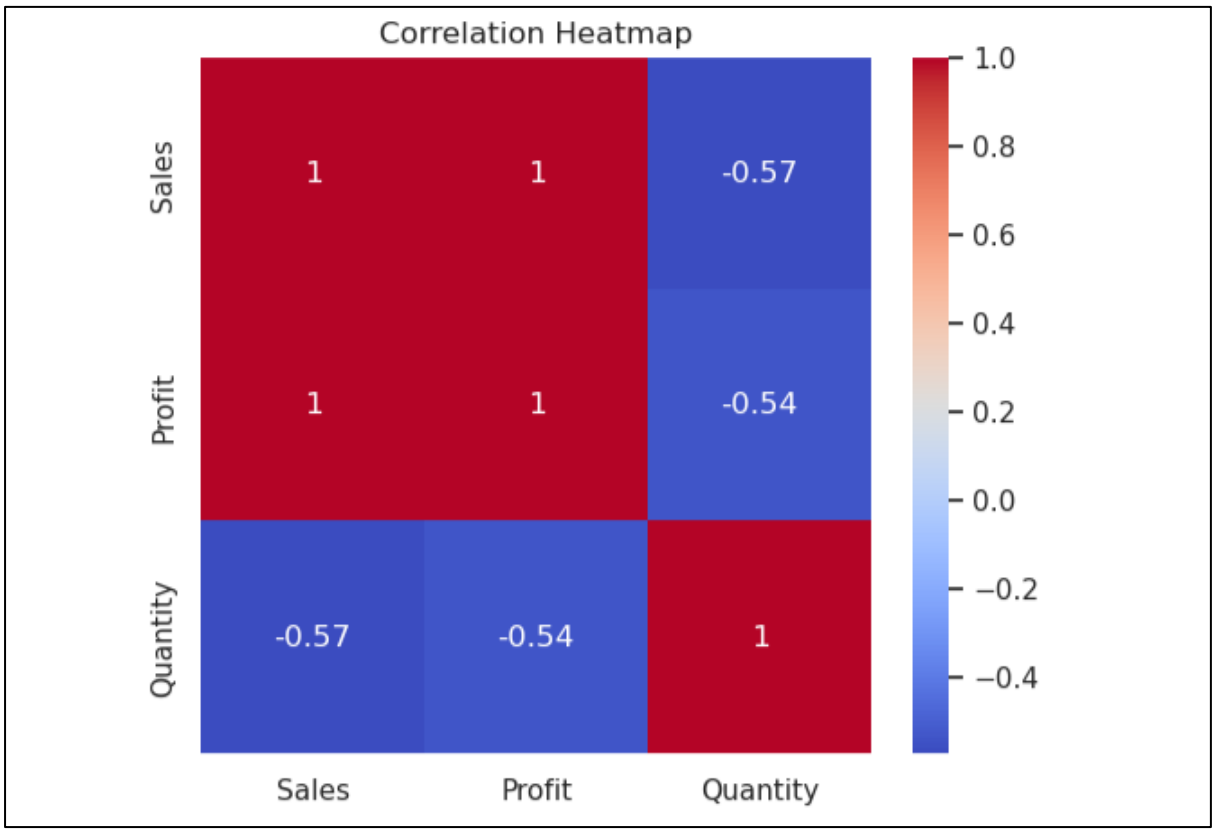
5. EXPLORATORY DATA ANALYSIS & VISUAL STORYTELLING (WEEK 2)

Advanced visualizations were developed to explore relationships within the dataset and communicate insights clearly to non-technical stakeholders. These visual narratives helped uncover trends and behavioral patterns.





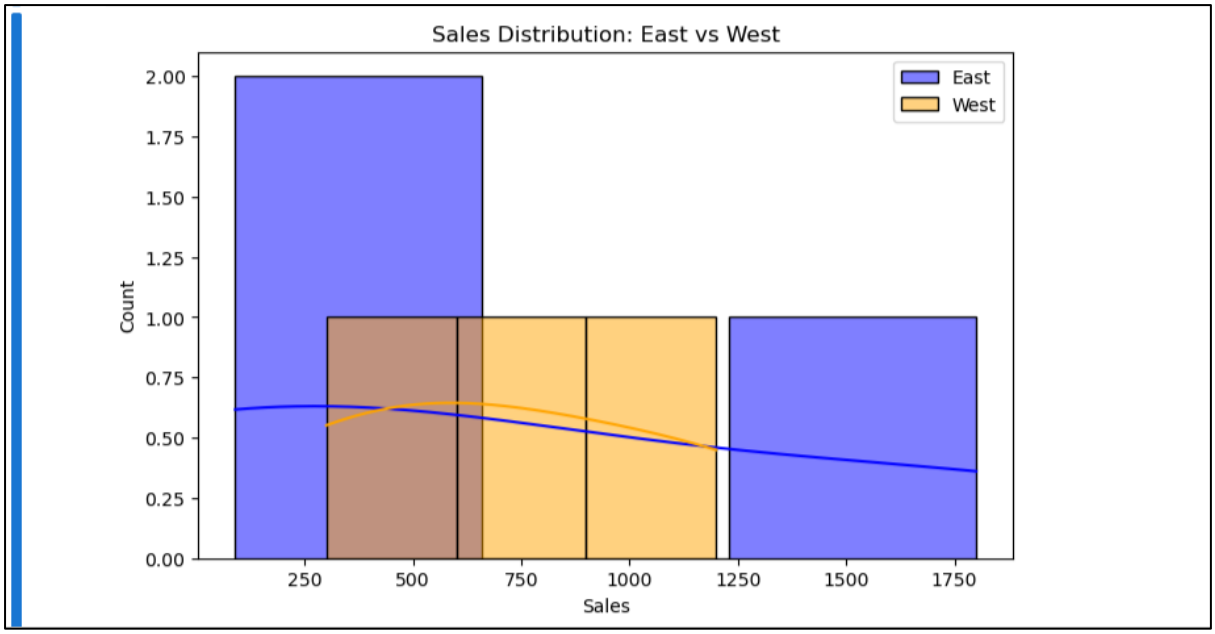

Sales Contribution by Category
Technology
53.7%
38.9%
Furniture
7.3%
Office Supplies


Correlation Heatmap
Sales
Profit
Quantity
1
1
-0.57
1
1
-0.54
-0.57
-0.54
1
Sales
Profit
Quantity
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

6. STATISTICAL ANALYSIS & HYPOTHESIS TESTING (WEEK 3)

Statistical methods were applied to determine whether observed differences were meaningful or occurred due to random variation. Hypothesis testing enhanced analytical rigor and supported evidence-based conclusions.

```
[10]: t_stat, p_value = stats.ttest_ind(east_sales, west_sales, equal_var=False)

      t_stat, p_value

[10]: (np.float64(-0.005510164608707995), np.float64(0.9959632639353717))

[11]: alpha = 0.5

      if p_value < alpha:
          print("Reject the Null Hypothesis")
      else:
          print("Fail to Reject the Null Hypothesis")

      Fail to Reject the Null Hypothesis
```



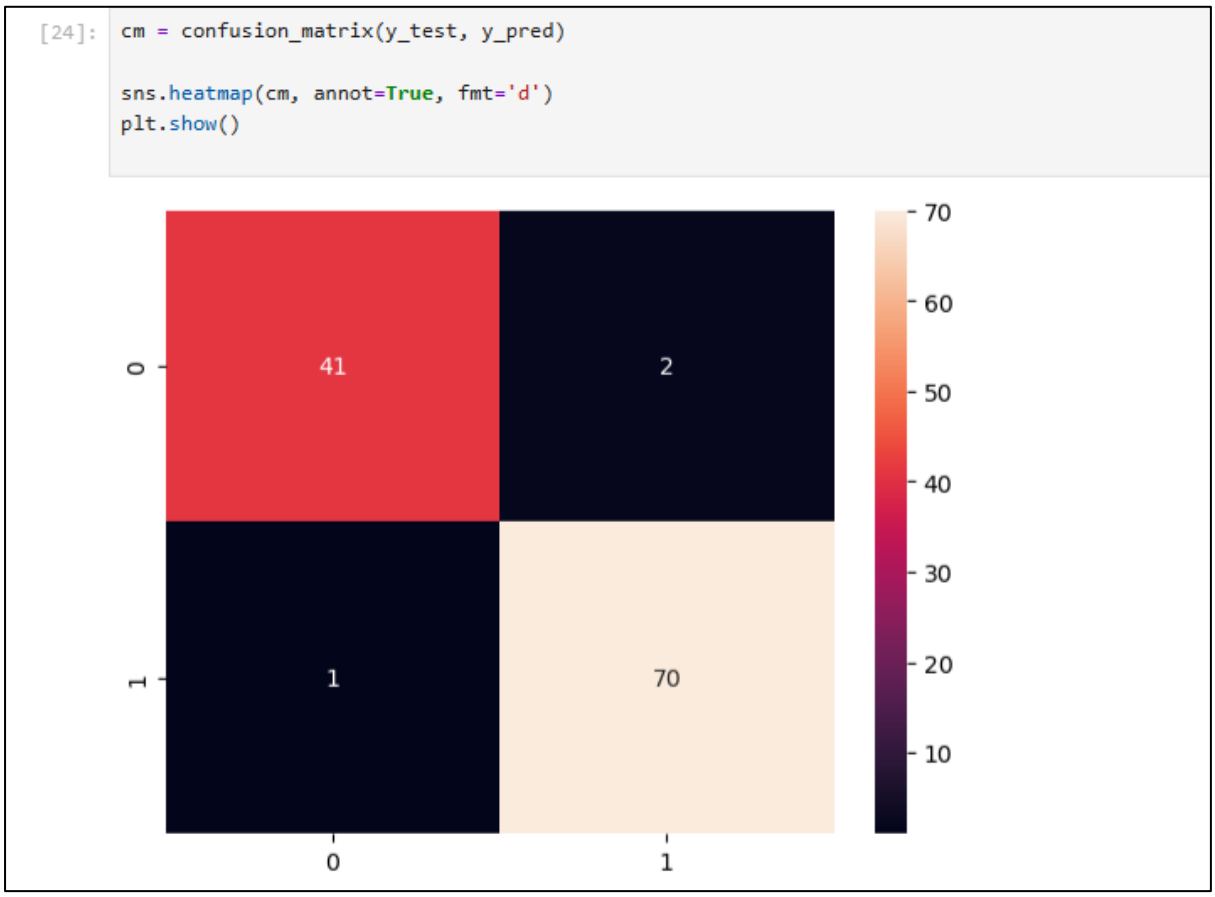
A significance level of 0.05 was used to guide decision-making.

7. MACHINE LEARNING MODEL DEVELOPMENT (WEEK 4)

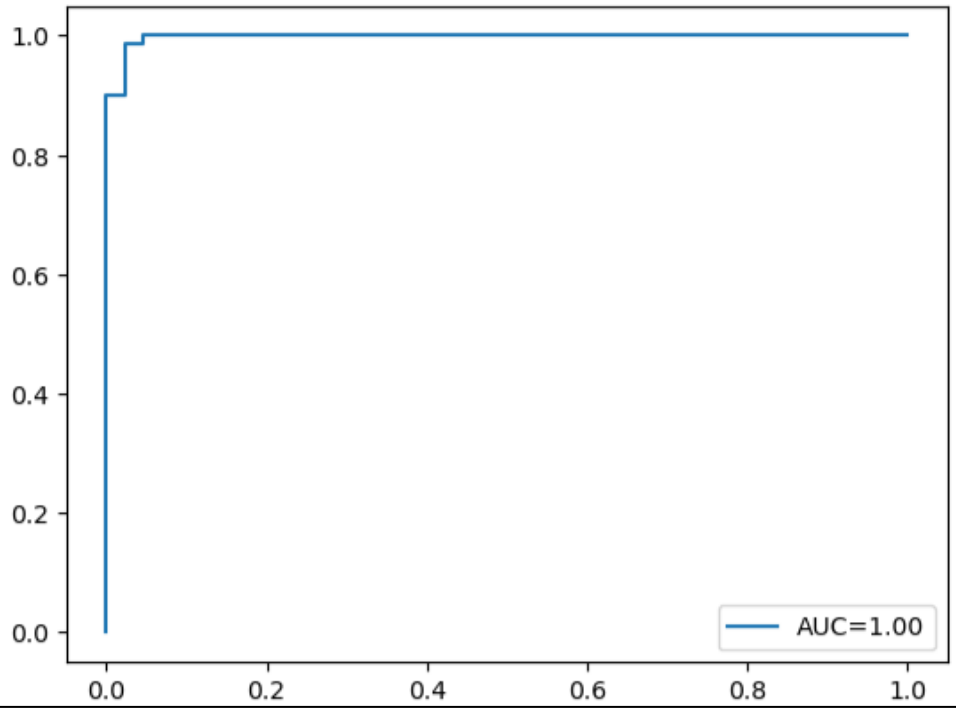
A supervised machine learning approach was implemented to build a predictive model. Logistic Regression was selected due to its interpretability and effectiveness in binary classification tasks.

Explain briefly:

- Train-test split
- Scaling
- Training
- Prediction



```
fpr, tpr, _ = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label=f"AUC={roc_auc:.2f}")
plt.legend()
plt.show()
```



8. KEY FINDINGS

- Data preprocessing significantly improved dataset usability.
- Visualization revealed meaningful patterns.
- Statistical testing validated analytical assumptions.
- The machine learning model demonstrated strong predictive capability.
- Structured workflows enhance analytical reliability.

9. STRATEGIC RECOMMENDATIONS

- Organizations should prioritize data cleaning before analysis.
- Decisions should be supported by statistical validation.
- Predictive models should be leveraged for proactive decision-making.
- Continuous data monitoring can improve long-term performance.
- Investment in data literacy can strengthen organizational strategy.

---

10. CONCLUSION

This project demonstrates the successful application of an end-to-end data science workflow. By integrating data preparation, analysis, statistical validation, and machine learning, the study highlights the transformative potential of data-driven strategies.