Exploratory Data Analysis (EDA) Using NumPy, Pandas & Matplotlib

Hilal Khan

Center for Artificial Intelligence & Emerging Technologies

Agenda

- What is EDA?
- Why EDA is Important
- Dataset Overview (Titanic)
- NumPy for Numerical Analysis
- Pandas for Data Manipulation
- Handling Missing Values
- Data Visualization with Matplotlib
- Insights Conclusion

Introduction to EDA

What is Exploratory Data Analysis?

- Process of analyzing datasets
- Summarizing main characteristics
- Using statistics + visualization
- Finding patterns, anomalies, relationships

Why EDA is Important?

- Understand the dataset
- Detect missing values
- Identify outliers
- Feature selection
- Improve ML model performance

Dataset Overview

Dataset Used: Titanic Dataset

- Source: Kaggle Titanic Competition
- File: titanic.csv
- Rows: 891 passengers
- Columns: 12 features

Features Description

- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Fare
- Embarked

Loading the Data

Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Loading Dataset

```
df = pd.read_csv("titanic.csv")
df.head()
```

Basic Information

```
df.info()
df.describe()
```

NumPy for Numerical Analysis

Why NumPy?

- Efficient numerical computation
- Supports arrays
- Faster than Python lists

Convert Column to NumPy Array

```
ages = df["Age"].dropna().values
np.mean(ages)
np.median(ages)
np.std(ages)
```

Statistical Insights

- Mean Age
- Median Age
- Standard Deviation

Data Cleaning

Checking Missing Values

```
df.isnull().sum()
```

Handling Missing Age

```
df["Age"].fillna(df["Age"].mean(), inplace=True)
```

Dropping Cabin Column

```
df.drop("Cabin", axis=1, inplace=True)
```

Pandas Analysis

Group By Survival

```
df.groupby("Sex")["Survived"].mean()
```

Survival by Class

```
df.groupby("Pclass")["Survived"].mean()
```

Visualization with Matplotlib

Histogram of Age

```
plt.hist(df["Age"], bins=30)
plt.title("Age␣Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```

Bar Plot of Survival

```
df["Survived"].value_counts().plot(kind="bar")
plt.title("Survival␣Count")
plt.show()
```

Fare Distribution

```
plt.hist(df["Fare"], bins=30)
plt.title("Fare␣Distribution")
plt.show()
```

Survival by Gender Plot

```
df.groupby("Sex")["Survived"].mean().plot(kind="bar")
plt.title("Survival␣Rate␣by␣Gender")
plt.show()
```

Correlation Analysis

Correlation Matrix

```
corr = df.corr(numeric_only=True)
corr
```

Heatmap (Basic Matplotlib)

```
plt.imshow(corr)
plt.colorbar()
plt.title("Correlation␣Matrix")
plt.show()
```

Insights

Key Insights

- Females had higher survival rate
- 1st class passengers survived more
- Fare positively correlated with survival
- Age slightly negatively correlated

Conclusion

Summary

- Used NumPy for statistics
- Used Pandas for cleaning grouping
- Used Matplotlib for visualization
- Extracted meaningful insights

Next Steps

- Feature Engineering
- Encoding categorical variables
- Train ML model

Programming Tips

- Start small and build up
- Test frequently
- Clean your data before modeling
- Visualize before concluding
- Don't be afraid to experiment!

Questions?
Feel free to ask any questions about today's topics!

Thank You!

Thank you for your attention!

Keep Practicing!

Access Course Materials:

Download Course Materials

Center for Artificial Intelligence & Emerging Technologies