Titanic Dataset Analysis

An Exploratory Approach Using Python

Presentation prepared by Gyanankur Baruah

Introduction

This presentation analyzes the Titanic dataset using Python tools like pandas, Plotly, and Dash. We aim to extract meaningful insights regarding survival rates based on various factors.

01

Dataset Overview

Features of the Titanic dataset

The Titanic dataset comprises various features including passenger demographics (age, gender) and class (1st, 2nd, 3rd). Insights drawn from this data can help us understand survival patterns.

Number of records and variables

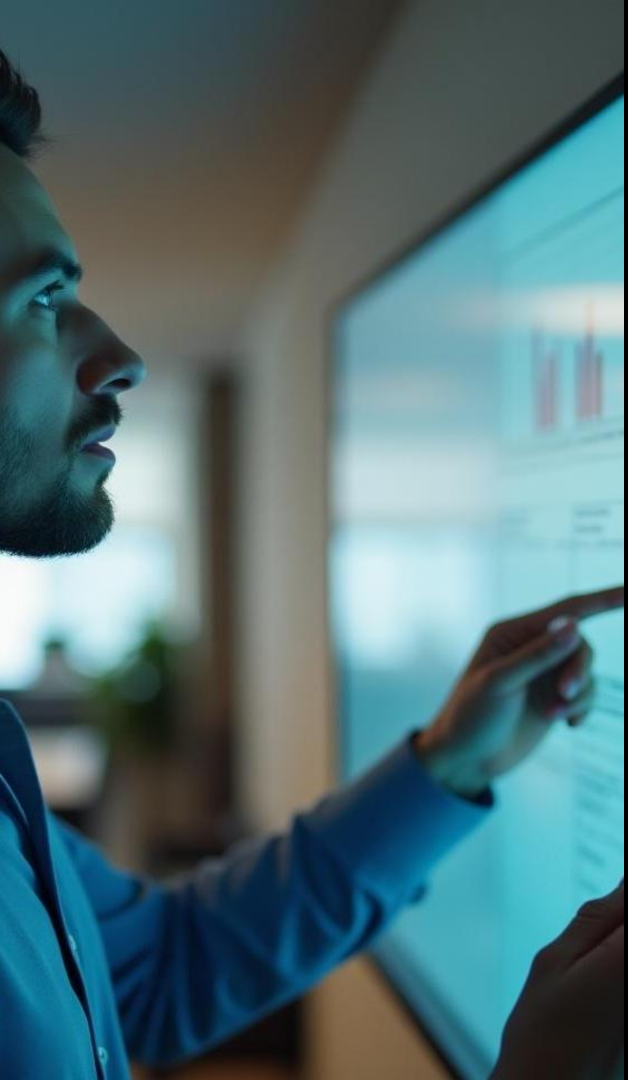
The dataset contains 887 records with attributes such as PassengerId, Name, Sex, Age, Class (Pclass), and Survived status. Certain variables have missing values, which necessitates careful handling during analysis.

Initial observations and insights

Preliminary analysis reveals potential trends, such as the survival likelihood potentially varying by gender and class. Observations highlight the need for a detailed study on these variables.

02

Data Cleaning

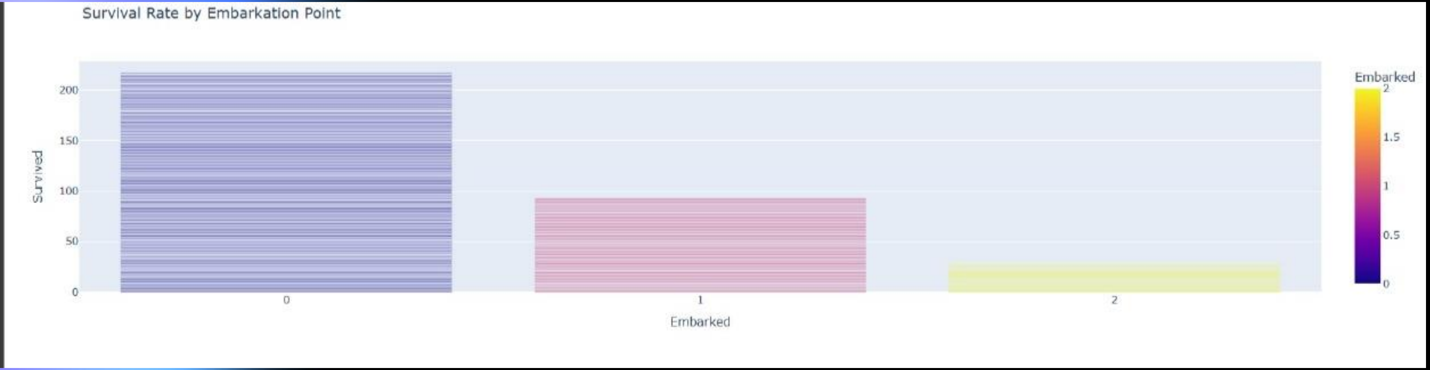


Handling missing values

Missing values in critical fields like Age and Cabin need attention. Strategies include imputation for Age based on median values and dropping Cabin due to excessive missing data, ensuring our dataset is robust.

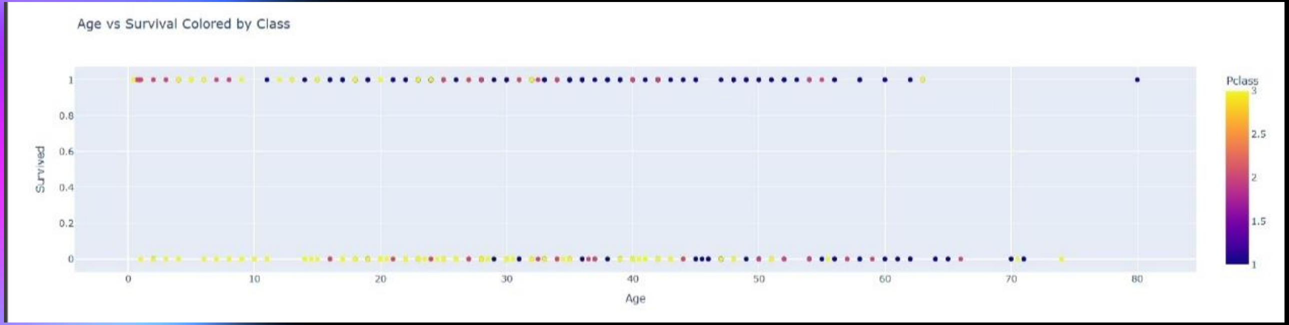
```
# Dropping Cabin column

df.drop(columns=[‘Cabin’], inplace=True)


# Filling missing Age values

df[‘Age’].fillna(df[‘Age’].median(), inplace=True)


# Encoding categorical variables

df[‘Sex’] = df[‘Sex’].map({‘male’: 0, ‘female’: 1})

df[‘Embarked’].fillna(‘S’, inplace=True)
```

Categorical data conversions



Certain columns, such as Sex and Embarked, require conversion into categorical variables for analysis. This transformation simplifies the data handling process and enhances the interpretability of results.

Dropping unnecessary columns

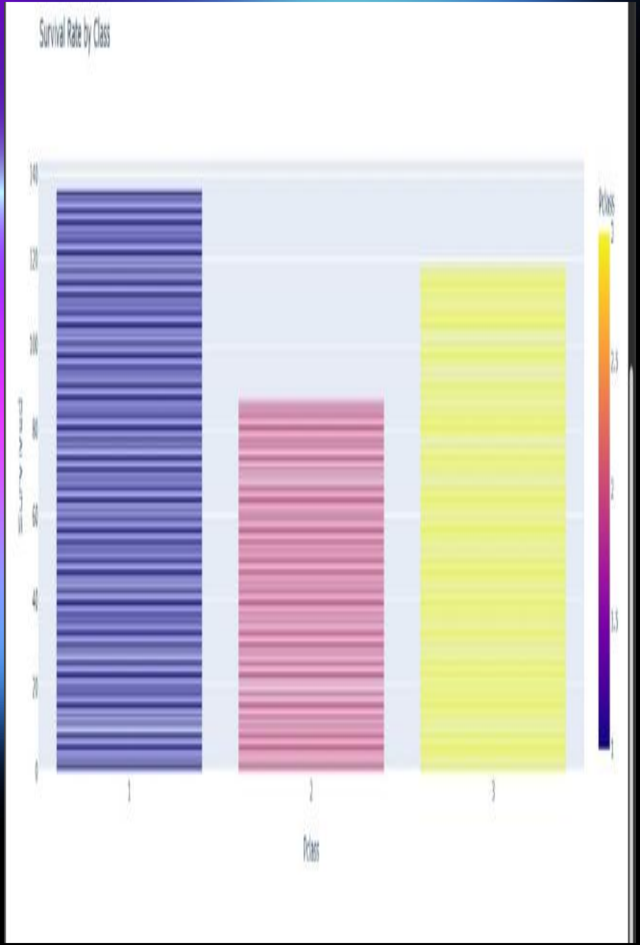


Unnecessary or redundant columns (like Cabin) will be removed to streamline the dataset. This reduces complexity and focuses analysis on the most impactful variables.

03

Survival Analysis

Survival rate by class



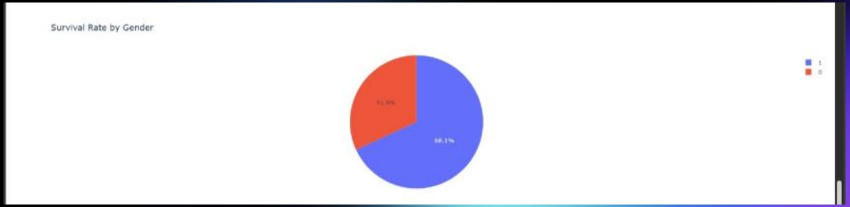
Analysis of the survival rates reveals significant disparities among classes. First class passengers had a markedly higher survival rate compared to those in second and third classes. This difference underscores the impact of socio-economic status on survival outcomes.

```
fig = px.bar(df, x=‘Pclass’, y=‘Survived’, color=‘Pclass’, title=‘Survival Rate by Passenger Class’)

fig.show()
```

Survival rate by gender

The analysis shows that women had a significantly higher survival rate compared to men. This trend suggests that gender played a critical role in survival during the disaster, highlighting societal norms regarding lifeboat allocation.

```
Fig = px.pie(df, names=‘Sex’, values=‘Survived’, title=‘Survival Rate by Gender’)

fig.show()
```



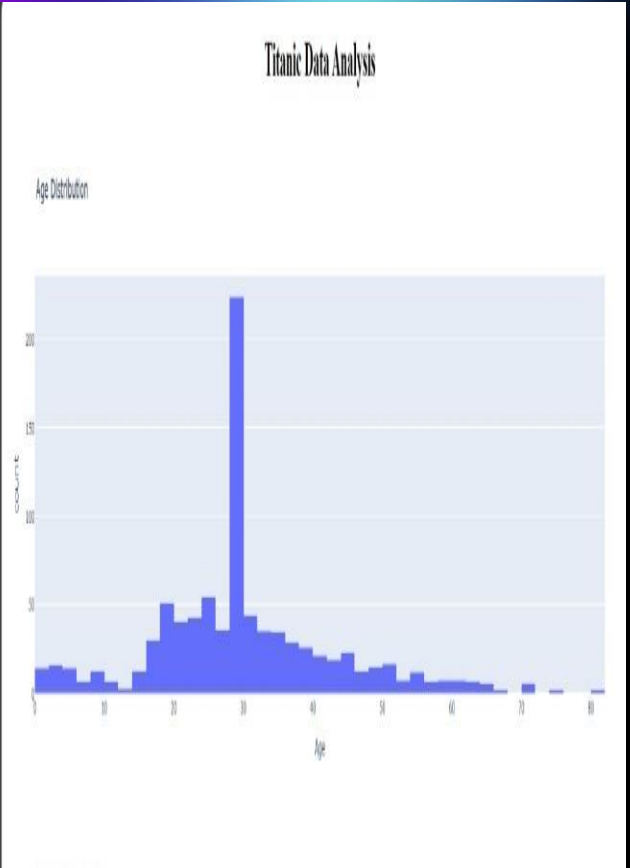
Impact of family size on survival

Passengers traveling with family (siblings/spouses or parents/children) had different survival outcomes. The data indicates that having family members aboard could enhance survival chances, illustrating the importance of family connections in crisis situations.

04

Visualizations

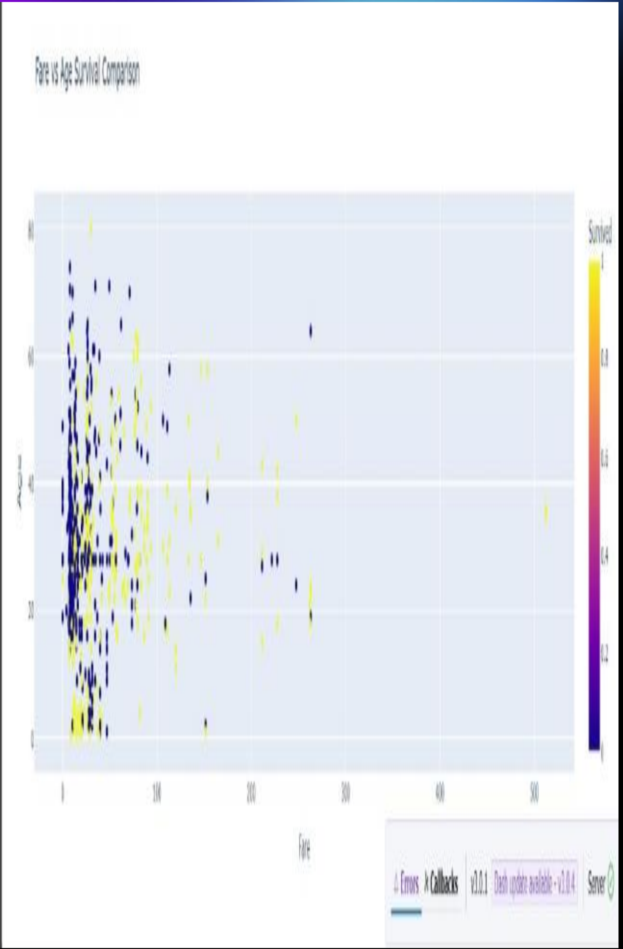
Age distribution analysis



The age distribution of passengers indicates that a majority fell between 20-40 years old. The analysis suggests that younger passengers may have been prioritized in evacuation, a point worthy of further investigation.

```
import plotly.express as px

fig = px.histogram(df, x=‘Age’, title=‘Age Distribution’)

fig.show()
```

Fare vs. age scatter plot



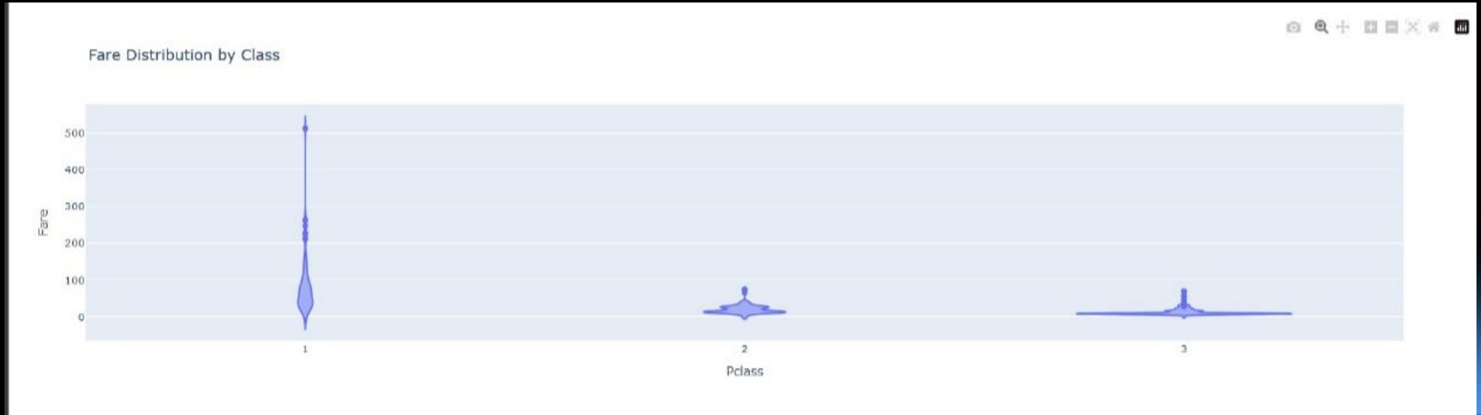
A scatter plot showcases the relationship between fare prices and passenger ages, indicating that older passengers generally paid higher fares. This trend could reflect the socio-economic factors influencing ticket purchases.

```
Fig = px.scatter(df, x=‘Fare’, y=‘Age’, color=‘Survived’, title=‘Fare vs. Age Survival Comparison’)

fig.show()
```

Fare distribution by class

A violin plot demonstrates fare distribution, revealing vast differences among classes. The first-class fares display significant variation, while second and third-class fares are more closely clustered, affirming economic disparities.

```
Fig = px.violin(df, x=‘Pclass’, y=‘Fare’, title=‘Fare Distribution by Class’)

fig.show()
```



Creating interactive dashboards

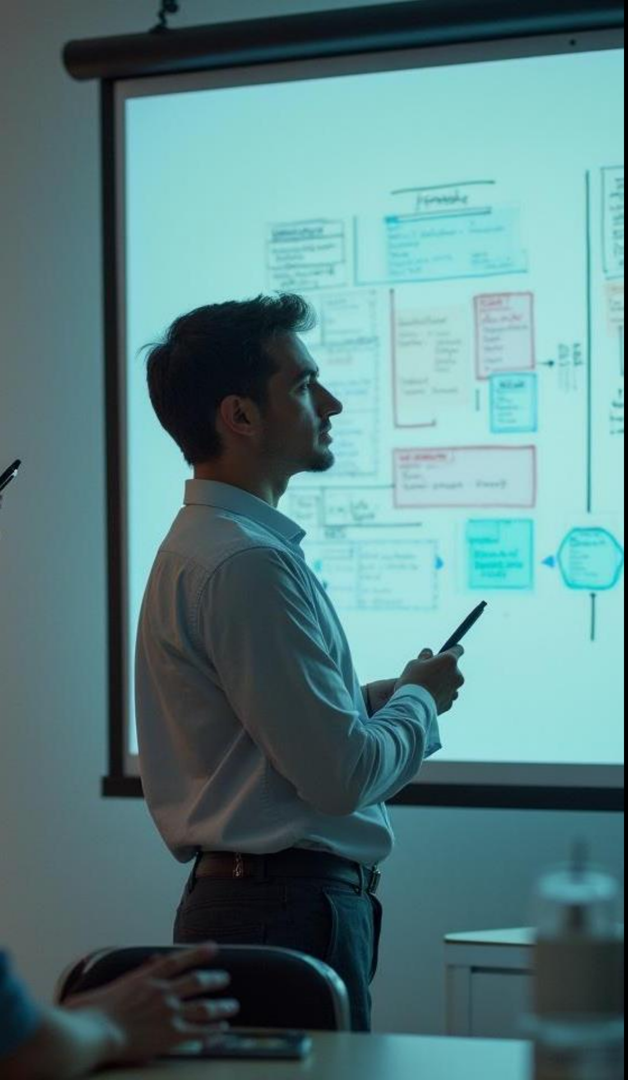
Developing interactive dashboards using tools like Dash can provide real-time insights into the dataset, allowing users to explore various aspects of data and findings dynamically.

05

Recommendations

Future work on prediction models

To improve insights further, implementing advanced machine learning models to predict survival outcomes based on the dataset can uncover deeper relationships and enhance understanding.



Enhancing data insights through feature engineering

Feature engineering can refine the dataset further, by creating new variables that capture interactions and nonlinear relationships among existing variables, which could lead to more accurate models.

Conclusions

The analysis of the Titanic dataset reveals critical insights into survival probabilities influenced by passenger class, gender, and family dynamics. Future efforts should focus on improving prediction models and enhancing visualizations for deeper engagement.

Thank you!

Do you have any questions?