MIT | Academy of Engineering

EDS Project on:

Exploring the Fifa dataset and Salary dataset.

Guided By: S.P Kale (MITAOE)

Presented By:

239 : Atharv Lokhande (20220105 0009)

225:Vaibhav Jadhav

FYBTech,MITAOE

Introduction

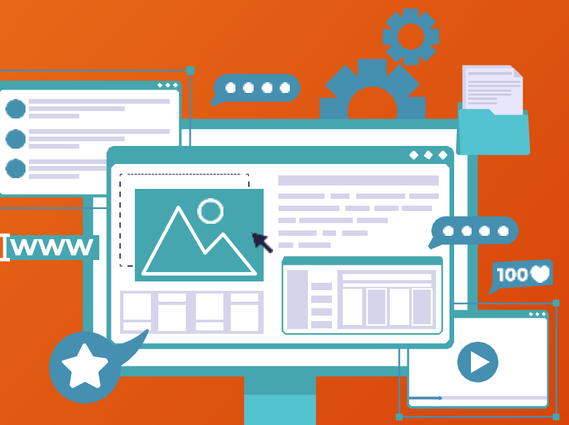
Data analytics encompasses a range of techniques and approaches to analyze data, including statistical analysis, data mining, predictive modeling, machine learning, and data visualization. By applying these techniques, analysts can discover patterns, trends, correlations, and anomalies within the data, which can provide valuable insights and drive informed decision-making. The goal of data analytics is to extract actionable information from complex and often unstructured data sets. It enables organizations to gain a deeper understanding of their operations, customers, and market trends, which can lead to improved efficiency, better customer experiences, and competitive advantages.

Motivation

FIFA datasets provide a wealth of information about football (soccer) players, teams, and matches. By analyzing this data using Python, you can gain insights into player performance, team strategies, match outcomes, and various other aspects of the game. This can be valuable for sports analytics professionals, coaches, and enthusiasts who want to understand and improve player and team performance.

Details of Dataset

- Name: Fifa Dataset
- Number of features: 14
- Number of records: 200



Data Manipulation

Data manipulation refers to the process of transforming and manipulating data to extract useful information, create new variables, and prepare data for analysis. It involves various operations such as filtering, sorting, aggregating, merging, and transforming data. Data manipulation is a critical step in the data analysis workflow and is commonly performed using p

```
# 3. Player whose salary is Minimum
min_salary_player = df.loc[df['SAL'] ==
df['SAL'].min(), 'PN'].iloc[0]
print("3. Player whose minimum Auction price: ",
min_salary_player)

# 4. Minimum Salary
min_salary = df['SAL'].min()
print("4. Minimum Auction price: ", min salary)

# 5. Player whose salary is Maximum
max salary player = df.loc[df['SAL'] ==
df['SAL'].max(), 'PN'].iloc[0]
print("5. Player whose maximum Auction price: ",
max_salary_player)
```

```
3. Player whose minimum Auction price:  MULLER
4. Minimum Auction price:  30000000
5. Player whose maximum Auction price:  Mbappe
```

```python
#14.print maximum goals scored by each team
max_goals_by_team = df.groupby('TN')['GS'].max()
print("14.Maximum goals scored by each team:\n",
max_goals_by_team)
```

```
14.Maximum goals scored by each team:
 TN
ACM              220
Al Nassar         55
Bayren FC         45
Kerla Blaster     99
MC                40
PSG               76
RM               330
Name: GS, dtype: int64
```

```python
#15.print number of players with match records
available
matches_count = df['MP'].count()
print("15. Number of players with match records
available: ", matches_count)
```

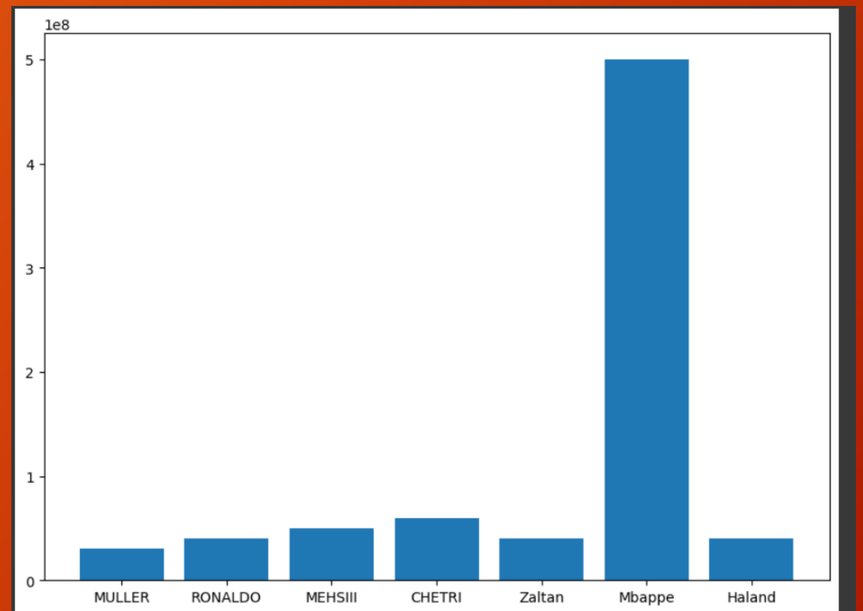
```
Name: GS, dtype: int64
15. Number of players with match records available:  7
16.Correlation between player salaries and goals scored:
```

Data Visualization

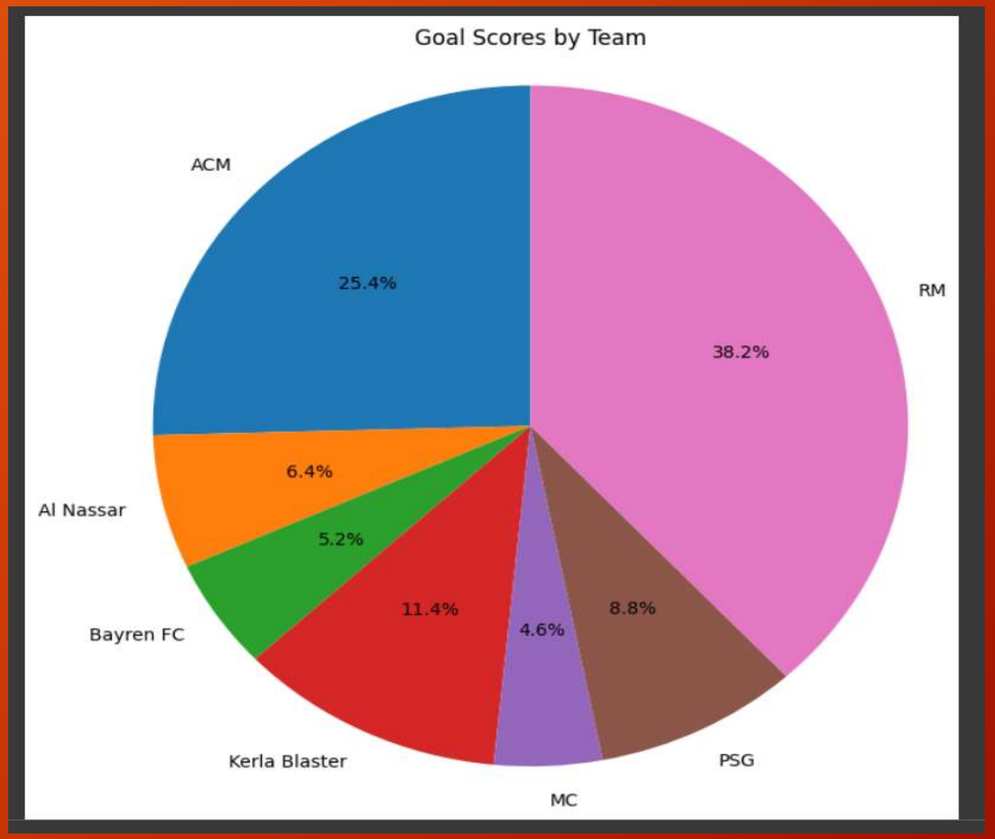
Data visualization is the process of representing data and information

- visually through charts, graphs, maps, and other graphical elements. It is a powerful technique that allows us to effectively communicate complex concepts, patterns, and trends in a visual format. Data visualization transforms complex data into visual representations that enhance understanding, reveal patterns, and support decision-making.

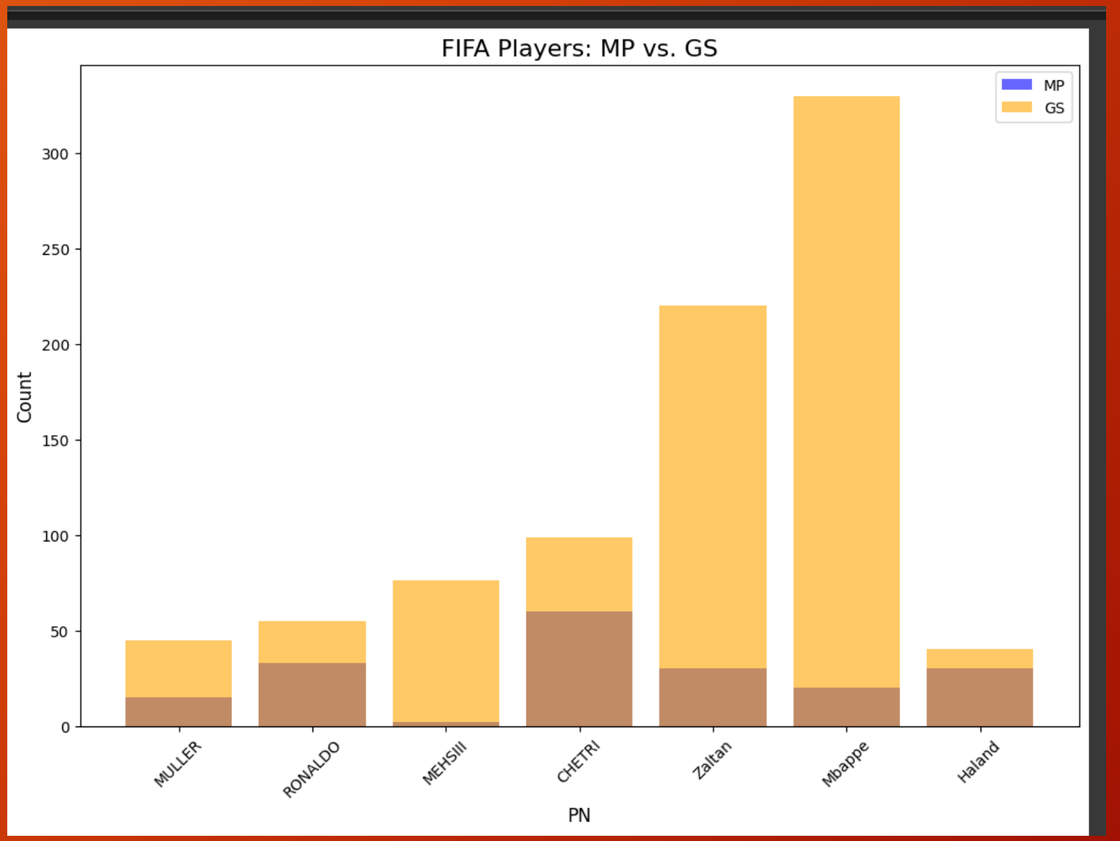
```python
import pandas as pd
from matplotlib import pyplot as plt

# Read CSV into pandas
data = pd.read_csv("/content/FIFA (1).csv")
data.head()
df = pd.DataFrame(data)

name = df['PN'].head(12)
price = df['SAL'].head(12)

# Figure Size
fig = plt.figure(figsize=(10, 7))

# Horizontal Bar Plot
plt.bar(name[0:10], price[0:10])

# Show Plot
plt.show()
```



```python
import pandas as pd
import matplotlib.pyplot as plt

# Read CSV into pandas
data = pd.read_csv("/content/FIFA (1).csv")
df = pd.DataFrame(data)

# Group the data by team and calculate the total goals scored
team_goals = df.groupby('TN')['GS'].sum()

# Get the team names and goal scores
team_names = team_goals.index.tolist()
goal_scores = team_goals.values.tolist()

# Create the pie chart
fig, ax = plt.subplots(figsize=(8, 8))
ax.pie(goal_scores, labels=team_names, autopct='%1.1f%%', startangle=90)

# Add a title
ax.set_title('Goal Scores by Team')

# Equal aspect ratio ensures that pie is drawn as a circle
ax.axis('equal')

# Show the pie chart
plt.show()
```

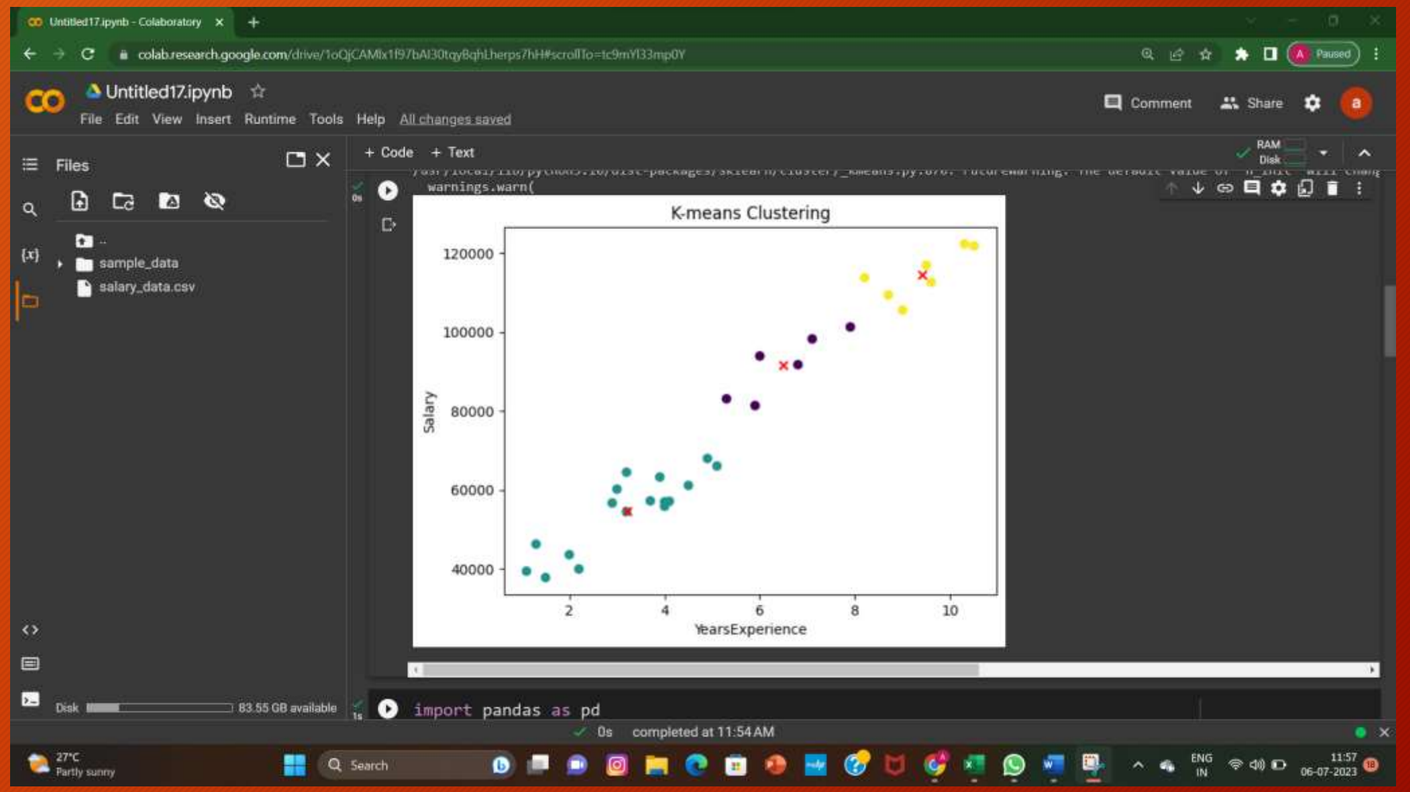


```python
import pandas as pd
import matplotlib.pyplot as plt

# Read CSV into pandas
data = pd.read_csv("/content/FIFA (1).csv")
df = pd.DataFrame(data)

# Extract the desired columns
name = df['PN'].head(12)
matches_played = df['MP'].head(12)
goals_scored = df['GS'].head(12)

# Set the figure size
fig, ax = plt.subplots(figsize=(12, 8))

# Plot the bars for matches played
ax.bar(name, matches_played, label='MP', color='blue', alpha=0.6)

# Plot the bars for goals scored
ax.bar(name, goals_scored, label='GS', color='orange', alpha=0.6)

# Set the title and labels
ax.set_title('FIFA Players: MP vs. GS', fontsize=16)
ax.set_xlabel('PN', fontsize=12)
ax.set_ylabel('Count', fontsize=12)

# Add a legend
ax.legend()

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.show()
```



Predictive Technique (K Means)

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Read the dataset from a CSV file
df = pd.read_csv("/content/salary_data.csv")

# Extract the relevant columns for clustering
X = df[['YearsExperience', 'Salary']].values

# Perform k-means clustering
k = 3  # Number of clusters
kmeans = KMeans(n clusters=k)
kmeans.fit(X)

# Get the cluster labels and centroids
labels = kmeans.labels
centroids = kmeans.cluster_centers_

# Visualize the clusters
plt.scatter(X[:, 0], X[:, 1], c=labels, cmap='viridis')
plt.scatter(centroids[:, 0], centroids[:, 1],
marker='x', color='r')
plt.xlabel('YearsExperience')
plt.ylabel('Salary')
plt.title('K-means Clustering')
plt.show()
```



Application

Data analytics is a field that involves examining and interpreting large sets of data to uncover meaningful insights and make informed decisions. In today's data-driven world, organizations and businesses collect vast amounts of data from various sources, such as customer interactions, financial transactions, social media, and sensor data.

Data analytics encompasses a range of techniques and approaches to analyze data, including statistical analysis, data mining, predictive modeling, machine learning, and data visualization. By applying these techniques, analysts can discover patterns, trends, correlations, and anomalies within the data, which can provide valuable insights and drive informed decision-making.

The goal of data analytics is to extract actionable information from complex and often unstructured data sets. It enables organizations to gain a deeper understanding of their operations, customers, and market trends, which can lead to improved efficiency, better customer experiences, and competitive advantages.

Conclusion

- In conclusion, our analysis of the Titanic dataset has provided valuable insights into the passengers and the factors influencing their survival.
- We discovered significant correlations between survival and variables such as age, gender, passenger class, and family size.
- The analysis highlighted the importance of preparedness, class disparities, and gender biases during this tragic event.
- Through data cleaning, preprocessing, visualization, and modeling, we were able to extract meaningful information.