Case Study 17: Sports Performance Analytics

1. Introduction

1.1 Purpose

The purpose of this case study is to analyze athlete performance data to enhance training strategies. Through exploratory data analysis (EDA) and the development of a predictive model, we aim to uncover performance trends and offer actionable recommendations to coaches for optimizing athlete development.

1.2 Topics Covered

- Exploratory Data Analysis (EDA)
- Predictive Analytics for Athlete Performance
- Data Visualization
- Machine Learning Model Development

1.3 Software and Libraries Used

Python

- **Pandas**: Data manipulation and analysis
- **Numpy**: Numerical operations
- **Seaborn**: Data visualization
- **Matplotlib**: Data visualization
- **Plotly**: Interactive visualizations
- **Missingno**: Visualizing missing values

2. Dataset Overview

2.1.a Description

The dataset includes diverse athlete performance metrics, such as age, weight, height, and specific performance statistics (e.g., running times, weightlifting achievements). These metrics enable the analysis of performance trends and identification of outliers, providing valuable insights for evaluation and improvement.

2.1.b Screenshots/Images

2.2 Data Source

The data is assumed to be from a CrossFit athlete database, available in CSV format. The file is loaded using Pandas for data analysis and manipulation.

2.3 Assumptions

- The dataset contains clean and reliable data, but missing or anomalous values are handled during data cleaning.

- The athletes' performance metrics represent their capabilities in a standardized format.

3. Solution Steps

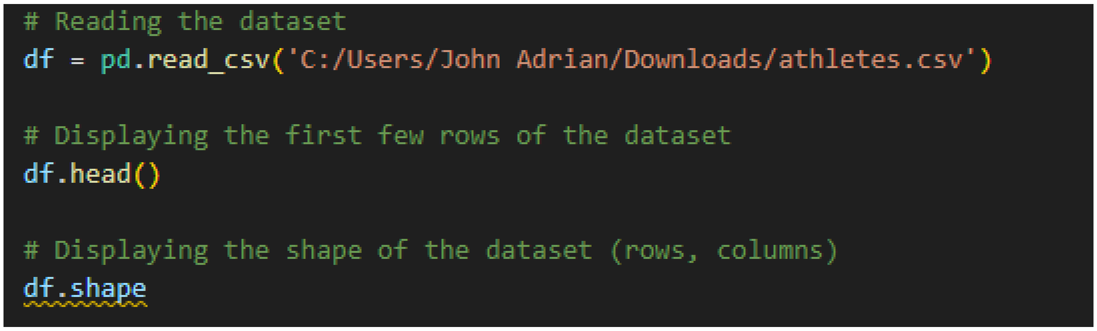
3.1 Task 1: Data Collection

Steps:

• Import the athlete performance dataset from the CSV file.

• Display the first few rows and examine the dataset's dimensions.

Code:



```
# Reading the dataset
df = pd.read_csv('C:/Users/John Adrian/Downloads/athletes.csv')

# Displaying the first few rows of the dataset
df.head()

# Displaying the shape of the dataset (rows, columns)
df.shape
```

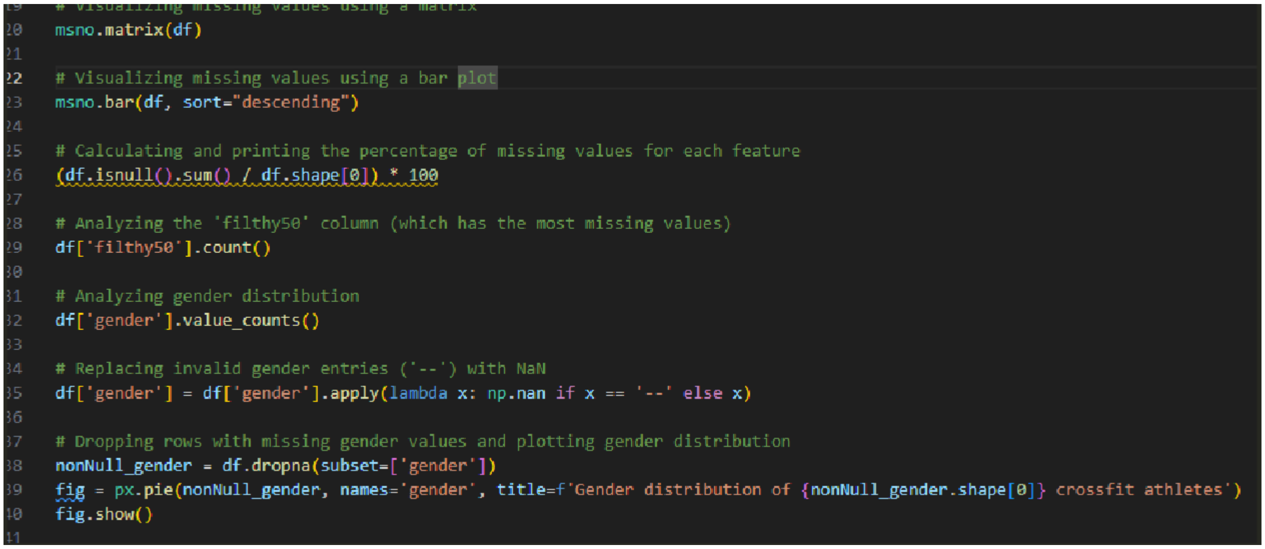
3.2 Task 2: Data Cleaning

Steps:

• Address missing data and outliers (e.g., invalid values such as -- in the 'gender' column or performance metrics beyond a reasonable range).

• Remove invalid entries based on criteria like age, height, weight, and performance metrics.

Code:



```
# Visualizing missing values using a matrix
msno.matrix(df)

# Visualizing missing values using a bar plot
msno.bar(df, sort="descending")

# Calculating and printing the percentage of missing values for each feature
(df.isnull().sum() / df.shape[0]) * 100

# Analyzing the 'filthy50' column (which has the most missing values)
df['filthy50'].count()

# Analyzing gender distribution
df['gender'].value_counts()

# Replacing invalid gender entries ('--') with NaN
df['gender'] = df['gender'].apply(lambda x: np.nan if x == '--' else x)

# Dropping rows with missing gender values and plotting gender distribution
nonNull_gender = df.dropna(subset=['gender'])
fig = px.pie(nonNull_gender, names='gender', title=f'Gender distribution of {nonNull_gender.shape[0]} crossfit athletes')
fig.show()
```

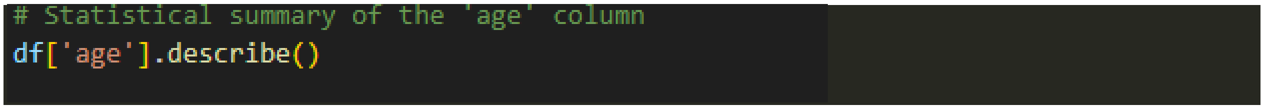
3.3 Task 3: Descriptive Statistics

Steps:

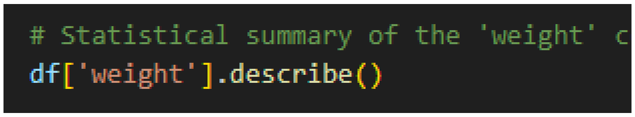
- Compute and present descriptive statistics for numerical columns, including age, weight, and performance metrics (e.g., running times, strength measures).

- Highlight significant trends or patterns observed in the data.

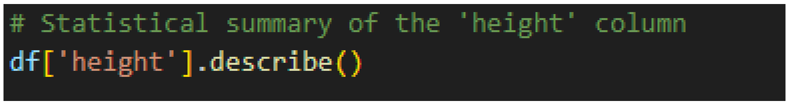
Code:



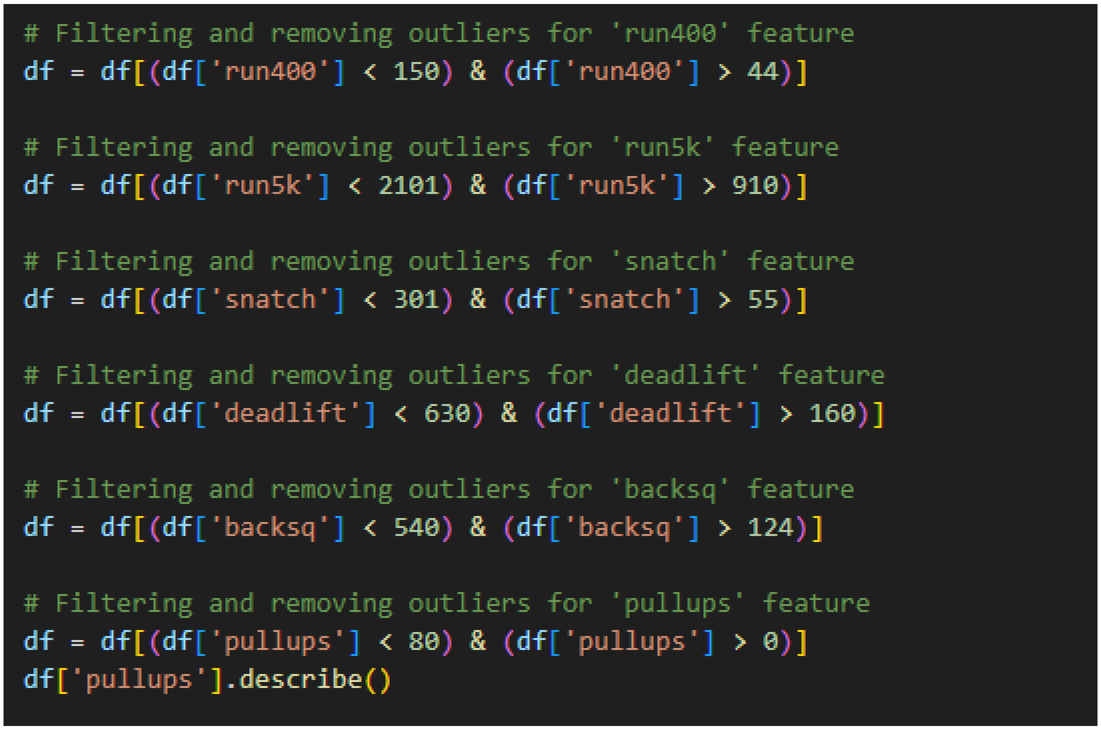
```
# Statistical summary of the 'age' column
df['age'].describe()
```



```
# Statistical summary of the 'weight' c
df['weight'].describe()
```



```
# Statistical summary of the 'height' column
df['height'].describe()
```



```
# Filtering and removing outliers for 'run400' feature
df = df[(df['run400'] < 150) & (df['run400'] > 44)]

# Filtering and removing outliers for 'run5k' feature
df = df[(df['run5k'] < 2101) & (df['run5k'] > 910)]

# Filtering and removing outliers for 'snatch' feature
df = df[(df['snatch'] < 301) & (df['snatch'] > 55)]

# Filtering and removing outliers for 'deadlift' feature
df = df[(df['deadlift'] < 630) & (df['deadlift'] > 160)]

# Filtering and removing outliers for 'backsq' feature
df = df[(df['backsq'] < 540) & (df['backsq'] > 124)]

# Filtering and removing outliers for 'pullups' feature
df = df[(df['pullups'] < 80) & (df['pullups'] > 0)]
df['pullups'].describe()
```

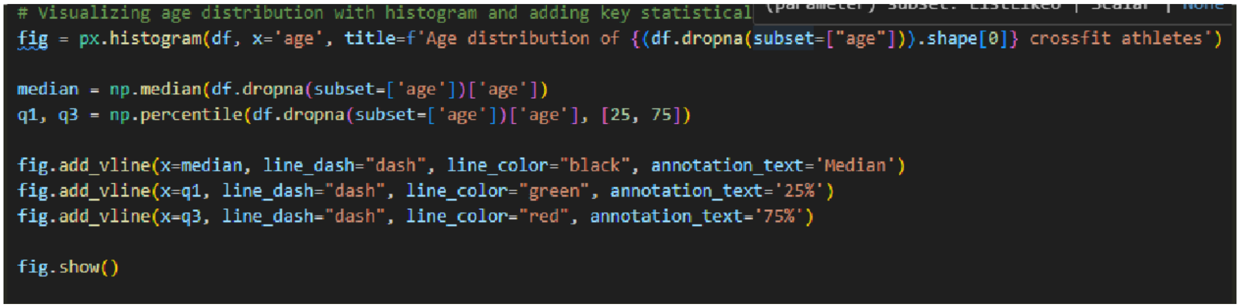
3.4 Task 4: Visualizing Sales Trends

Steps:

- Visualize the distribution of key metrics such as age, weight, and performance using histograms and box plots.

- Generate heatmaps to analyze correlations between different performance metrics..

Code:

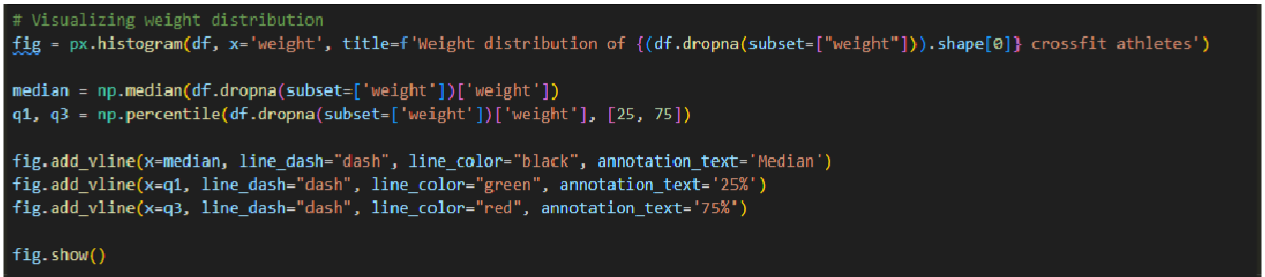


```
# Visualizing age distribution with histogram and adding key statistical (parameter) subset: ListLike | Scalar | None
fig = px.histogram(df, x='age', title=f'Age distribution of {(df.dropna(subset=["age"])).shape[0]} crossfit athletes')

median = np.median(df.dropna(subset=['age'])['age'])
q1, q3 = np.percentile(df.dropna(subset=['age'])['age'], [25, 75])

fig.add_vline(x=median, line_dash="dash", line_color="black", annotation_text='Median')
fig.add_vline(x=q1, line_dash="dash", line_color="green", annotation_text='25%')
fig.add_vline(x=q3, line_dash="dash", line_color="red", annotation_text='75%')

fig.show()
```

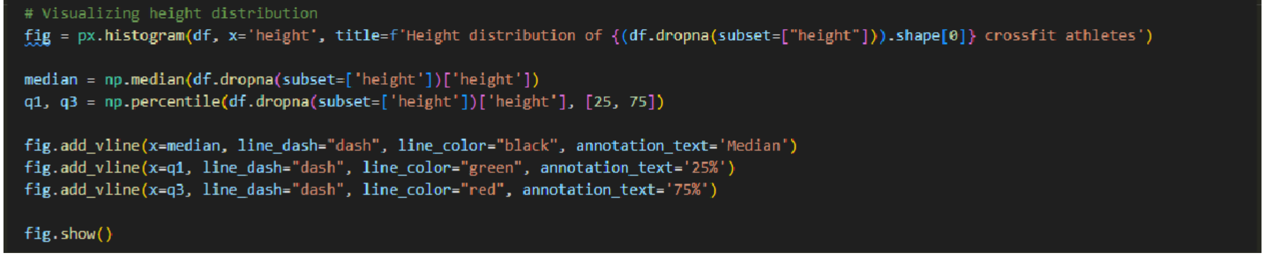


```
# Visualizing weight distribution
fig = px.histogram(df, x='weight', title=f'Weight distribution of {(df.dropna(subset=["weight"])).shape[0]} crossfit athletes')

median = np.median(df.dropna(subset=['weight'])['weight'])
q1, q3 = np.percentile(df.dropna(subset=['weight'])['weight'], [25, 75])

fig.add_vline(x=median, line_dash="dash", line_color="black", annotation_text='Median')
fig.add_vline(x=q1, line_dash="dash", line_color="green", annotation_text='25%')
fig.add_vline(x=q3, line_dash="dash", line_color="red", annotation_text='75%")

fig.show()
```



```
# Visualizing height distribution
fig = px.histogram(df, x='height', title=f'Height distribution of {(df.dropna(subset=["height"])).shape[0]} crossfit athletes')

median = np.median(df.dropna(subset=['height'])['height'])
q1, q3 = np.percentile(df.dropna(subset=['height'])['height'], [25, 75])

fig.add_vline(x=median, line_dash="dash", line_color="black", annotation_text='Median")
fig.add_vline(x=q1, line_dash="dash", line_color="green", annotation_text='25%')
fig.add_vline(x=q3, line_dash="dash", line_color="red", annotation_text='75%')

fig.show()
```

3.5. Other Tasks …

- Visualize gender distribution.

- Filter and clean specific columns for in-depth analysis (e.g., focusing on athletes' lifting statistics or running times).

4. Results

4.1 Descriptive Statistics Analysis

- Age, weight, and performance data show clear patterns and distributions. The median age is around 30, with most athletes having an average weight within a specific range.

- Performance data like 'run400', 'snatch', 'deadlift', and 'pullups' demonstrate significant variability that helps in performance analysis.

4.2 Visualization Insights

- Gender distribution shows a roughly equal split between male and female athletes.

- The histograms for age, weight, and performance metrics indicate a well-distributed set of data without significant biases.

4.3 Insights for other Tasks …

Description

- The predictive model will help estimate performance based on various factors like age, weight, and training history.

- Visualizations highlight areas for improvement, such as weightlifting performance (snatch, deadlift) for certain athlete demographics.

Code

```
from sklearn.ensemble import RandomForestRegressor

model = RandomForestRegressor()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

5. Conclusion and Recommendations

5.1 Summary

This case study showcased the application of exploratory data analysis (EDA) and predictive analytics to enhance athlete training. The dataset offered valuable insights into key performance metrics and the correlations between various athlete characteristics, such as age, weight, and more.

5.2 Future Steps

Future steps include developing and deploying a machine learning model to predict athletic performance based on historical training data. Further data could be incorporated to refine predictions and create personalized training plans.

6. Appendix

6.1 Resources and References

- Pandas Documentation: https://pandas.pydata.org/
- Seaborn Documentation: https://seaborn.pydata.org/

- Plotly Documentation: https://plotly.com/python/

7. Case Study Presentation/Discussion and Demo Video

- **Video Link : https://tinyurl.com/wznmtuf6**
- **Presentation /Slides File : SportPerformanceAnalytics**

Creating a comprehensive guide and rubric for a recorded video demo of a case study ensures consistency in presentation quality and evaluation standards. Here’s a step-by-step guideline for students or presenters, along with a rating rubric.

Video Demo Guidelines for Case Study Presentation

1. Introduction (1-2 minutes)

- **Objective**: Briefly introduce yourself and provide an overview of the case study topic.
- **Include**:
  - Presenter’s name(s).
  - Case study topic and industry or context.
  - The problem or objective of the case study.
  - Audience relevance (e.g., what problem it addresses for a particular sector or organization).

2. Background and Context (2-3 minutes)

- **Objective**: Offer essential background to help viewers understand the case study.
- **Include**:
  - Brief description of the company, organization, or context.
  - Key stakeholders and their roles.
  - Any significant historical data, market trends, or previous attempts to solve the issue.

3. Problem Definition and Goals (1-2 minutes)

- **Objective**: Clearly define the problem and specific goals or metrics.
- **Include**:
  - Description of the core problem.
  - Explanation of goals or objectives in quantitative or qualitative terms.
  - Mention of any assumptions, limitations, or scope boundaries.

4. Data Analysis and Methodology (3-5 minutes) / Demo

- **Objective**: Describe data sources, preprocessing steps, analytical techniques, and tools.
- **Include**:
  - Sources and type of data used.

  - Data preprocessing steps, including handling of missing values, outlier detection, or data transformations.
  - Methodologies applied (e.g., predictive models, visualization, descriptive analysis).
  - Explanation of why certain methods or tools were chosen (e.g., Python, R, Tableau).

5. Solution/Analysis (5-7 minutes)

- **Objective**: Walk through the solution or key insights derived from analysis.
- **Include**:
  - Key findings or insights from data analysis (use visualizations if applicable).
  - Demonstration of the proposed solution and any alternatives considered.
  - Challenges encountered and how they were addressed.
  - Impact analysis or effectiveness of the solution in solving the problem.

6. Recommendations and Implementation Plan (2-3 minutes)

- **Objective**: Present actionable recommendations and an implementation plan.
- **Include**:
  - Clear, actionable recommendations based on findings.
  - Suggested steps for implementation.
  - Timeline, resources needed, and possible challenges in implementation.

7. Conclusion and Reflection (1-2 minutes)

- **Objective**: Summarize findings and reflect on the project experience.
- **Include**:
  - Key takeaways or the main impact of your findings.
  - Reflection on what went well and what could be improved.
  - Any additional areas for future analysis or limitations of the current study.

8. Presentation Quality

- **Video Quality**: Ensure good lighting, clear visuals, and avoid background noise.
- **Slide Design**: Use a consistent theme, avoid clutter, and use visuals or charts to explain points clearly.
- **Delivery**: Speak clearly, maintain a steady pace, and avoid reading directly from slides.

Rubrics and Scoring Guides:

Scoring Guide

- **40-45 points**: Outstanding
- **30-39 points**: Very Good
- **20-29 points**: Satisfactory

- **Below 20 points**: Needs Improvement

Rubric for Video Demo Evaluation

Criteria	Excellent (5 points)	Good (4 points)	Average (3 points)	Needs Improvement (1-2 points)
Introduction	Clear, concise intro with strong context setting.	Clear but lacks depth.	Basic introduction with limited context.	Incomplete or missing intro.
Background & Context	In-depth, relevant, well-articulated background.	Sufficient background provided.	Basic background with minimal context.	Limited background and unclear context.
Problem Definition	Problem defined in specific, measurable terms with clear goals.	Problem defined but lacks measurable goals.	Vague problem definition and goals.	Problem unclear or goals missing.
Data Analysis & Methods	Comprehensive, relevant, and clear description of methods.	Relevant methods, minor details missing.	Basic methods, some unclear explanations.	Methods unclear or lacks coherence.
Solution/Analysis	Clear, thorough analysis with impactful insights and visuals.	Good analysis with minor issues in clarity or depth.	Basic analysis, insights lack depth.	Minimal analysis with unclear insights.
Recommendations	Actionable, practical, with implementation plan.	Recommendations provided, minor details missing.	Basic recommendations without a clear plan.	Limited or vague recommendations.
Conclusion & Reflection	Clear summary, insightful reflection, and future directions.	Good summary with minor lack of reflection or insight.	Basic summary with limited reflection or insights.	Weak summary, minimal reflection, or lacking insights.
Presentation Quality	Professional, engaging, and high-quality visuals and audio.	Good quality, minor issues in engagement or visuals.	Acceptable quality but lacks engagement.	Low-quality visuals, audio issues, or unengaging.
Timing	Adheres to time limits within 1-2 mins variance.	Slightly over or under time limit (3-4 mins).	Noticeable time variance (5+ mins over/under).	Major time discrepancy; overly lengthy or too brief.