

## **8 Week 8: Exploratory Data Analysis (EDA) Project**

This code loads a large simulated box score dataset and performs EDA using statistics and visualizations.

# Week 8 Assignment – Exploratory Data Analysis Project

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
# --- Data Loading (Assumes 'clean_box_scores.csv' was generated in the setup cell) ---
```

```
df = pd.read_csv('clean_box_scores.csv')
```

```
# --- Step 1: Summary Statistics ---
```

```
print("--- 1. Descriptive Statistics ---")
```

```
print(df.describe())
```

```
# --- Step 2: Visualize Distributions (Histogram) ---
```

```
plt.figure(figsize=(10, 4))
```

```
sns.histplot(df['PTS'], kde=True, bins=20, color='darkblue')
```

```
plt.title('Distribution of Points Scored')
```

```
plt.show()
```

```
# --- Step 3: Explore Correlations (Heatmap) ---
```

```
print("\n--- 3. Correlation Heatmap ---")
```

```
# Select only numerical features and calculate correlation
```

```
numerical_df = df[['PTS', 'REB', 'AST', 'FG%', '3P%', 'FT%']]

plt.figure(figsize=(8, 6))

sns.heatmap(numerical_df.corr(), annot=True, cmap='coolwarm', fmt=".2f")

plt.title('Correlation Heatmap of Key Metrics')

plt.show()
```

```
# --- Categorical Analysis (Box Plot) ---
```

```
plt.figure(figsize=(8,6))

sns.boxplot(data=df, x='Position', y='REB', palette='Set2')

plt.title('Rebound Distribution by Position')

plt.show()
```

```
# --- Step 5: Write a Summary of Insights (Example) ---
```

```
print("\n--- 5. Summary of Insights ---")

print("The average game performance is approximately 18 points, 6 rebounds, and 4  
assists. The correlation heatmap reveals that Points (PTS) are most strongly related to Field  
Goal Percentage (FG%), indicating that efficient scoring drives volume. The box plot  
confirms traditional roles, with Centers (C) showing the highest median for Rebounds,  
although all positions exhibit a wide statistical spread.")
```