🧠 Week 8 Student Guide – Exploratory Data Analysis (EDA) Project

This guide prepares you for the **Week 8 Assignment**, which requires you to perform a **complete Exploratory Data Analysis (EDA)** on a simulated basketball box score dataset. This week brings together skills from previous weeks: data manipulation (Pandas), statistics (descriptive stats), and visualization (Matplotlib/Seaborn).

💻 Part 1: EDA Fundamentals Review

Exploratory Data Analysis (EDA) is the critical process of analyzing data to summarize its main characteristics, often using visual methods. In basketball, EDA helps us understand performance patterns, identify outliers, and see relationships between different statistics.

1. Descriptive Statistics

The `.describe()` method in Pandas provides a powerful numerical summary of your data.

- **Count:** The number of valid (non-missing) entries.
- **Mean/Median:** Measures of central tendency (average and middle value).
- **Std (Standard Deviation):** Measures the spread or variability of the data (how consistent the performance is).
- **Min/Max/Quartiles (25%, 50%, 75%):** Describe the range and distribution of the data.

2. Visualization (Univariate & Bivariate)

Plot Type	Goal in EDA	Key Pandas/Seaborn Code	Example Insight
Histogram	Show the **distribution** of a single metric (e.g., Points) to see if it's normally distributed (bell-curve) or skewed.	`sns.histplot(df['PTS'], kde=True)`	"Most games fall in the 15-20 point range."
Box Plot	**Compare** a single metric (e.g., Points) across **categories** (e.g., Position) to see differences in performance and outliers.	`sns.boxplot(data=df, x='Position', y='PTS')`	"Centers (C) have a higher median Rebound total than Guards (G)."

Plot Type	Goal in EDA	Key Pandas/Seaborn Code	Example Insight
Heatmap	Show the **correlation** (relationship) between many pairs of numerical metrics.	sns.heatmap(df[['PTS', 'REB']].corr(), annot=True, cmap='coolwarm')	"Points (PTS) have a strong positive correlation with Free Throw % (FT%)."

🏀 Part 2: Week 8 Assignment Tasks

The assignment uses a provided synthetic dataset (clean_box_scores.csv) with 1000 rows of game data.

1. Load Data (Code Cell 4)

Your first step is to load the data that was generated for the assignment.

Python
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('clean_box_scores.csv')
```
2. Summary Statistics (Step 1)

Use .describe() to get a statistical overview of the key performance and efficiency metrics.

Python
```
# Focus on key statistics and shooting percentages
df.describe()
```
3. Visualize Distributions (Step 2)

Create a **histogram** to understand the common range of one of the metrics, such as Points (PTS).

Python
```
plt.figure(figsize=(10,4))
sns.histplot(df['PTS'], kde=True, bins=20)
plt.title('Distribution of Points Scored')
plt.show()
```
4. Explore Correlations (Step 3)

Create a **heatmap** of the correlation matrix to easily identify which stats are related to each other (e.g., PTS, REB, AST, and shooting percentages).

Python

```
plt.figure(figsize=(8,6))
# Calculate the correlation for the metrics and visualize using a heatmap
sns.heatmap(df[['PTS','REB','AST','FG%','3P%','FT%']].corr(),
        annot=True, # Show the correlation values on the heatmap
        cmap='coolwarm')
plt.title('Correlation Heatmap of Key Metrics')
plt.show()
```

5. Position-Based Analysis (Extra Visualization)

While not explicitly listed as a step, the tutorial included a box plot for position-based analysis, which is a key part of EDA for basketball. You can add this step to meet the visualization requirement.

- **Goal:** Compare a stat (e.g., Points) across the categorical Position column.

Python

```
plt.figure(figsize=(8,6))
sns.boxplot(data=df, x='Position', y='PTS', palette='Set2')
plt.title('Points Distribution by Position')
plt.show()
```

6. Write a Summary of Insights (Step 5)

Based on your charts and statistics, write a 3–5 sentence summary.

- **Reference the Mean/Std:** What's the average scoring/rebounding game like?
- **Reference the Correlation Heatmap:** Which metrics showed the strongest relationships (e.g., high positive correlation)? *Hint: PTS is often strongly correlated with shooting efficiency.*
- **Reference the Box Plot:** How does one position (e.g., 'G' vs 'C') differ from another in a key metric?