Applied Statistics Computational Project

Data

For the Data , i have taken the match scores of Virat Kohli from 2020 to 2024.

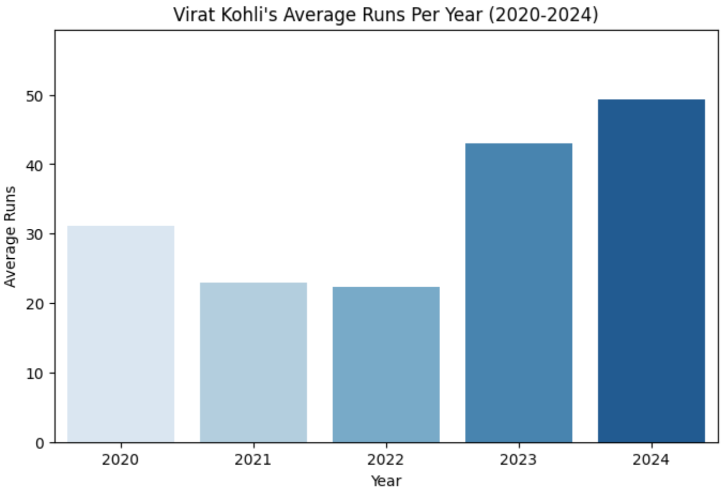
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
df = pd.read_excel('Virat␣Kohli␣ipl␣match␣scores␣2020-2024.xlsx',
    header=None)
df.columns = ["Runs"]

# Drop any NaN values (if present)
df.dropna(inplace=True)

# Convert to numeric (if needed)
df["Runs"] = pd.to_numeric(df["Runs"], errors="coerce")

# Display first few rows
print(df.head())
print(df.info())  # Check if it's numeric
```

Descriptive statistics

```
# Descriptive statistics
mean = df["Runs"].mean()
median = df["Runs"].median()
std_dev = df["Runs"].std()
variance = df["Runs"].var()
q1, q3 = np.percentile(df["Runs"], [25, 75])
iqr = q3 - q1
min_val, max_val = df["Runs"].min(), df["Runs"].max()
# Print statistics
print(f"Mean: {mean:.2f}")
print(f"Median: {median:.2f}")
print(f"Standard Deviation: {std_dev:.2f}")
print(f"Variance: {variance:.2f}")
print(f"Q1: {q1}, Q3: {q3}")
print(f"IQR: {iqr}")
print(f"Min: {min_val}, Max: {max_val}")
```

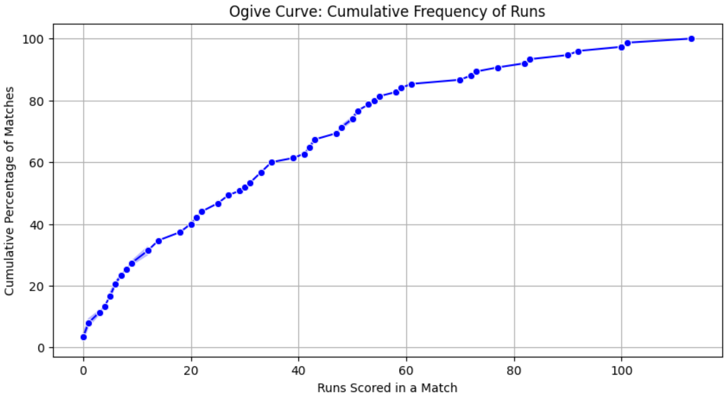
Bar diagram of runs

```
df["Year"] = ["2020"] * 15 + ["2021"] * 15 + ["2022"] * 15 + ["2023"] * 15 + ["2024"] * (len(df) - 60)

yearly_avg = df.groupby("Year")["Runs"].mean()

plt.figure(figsize=(8, 5))
sns.barplot(x=yearly_avg.index, y=yearly_avg.values, palette="Blues")
plt.title("Virat Kohli’s Average Runs Per Year (2020-2024)")
plt.xlabel("Year")
plt.ylabel("Average Runs")
plt.ylim(0, max(yearly_avg.values) + 10)
plt.show()
```



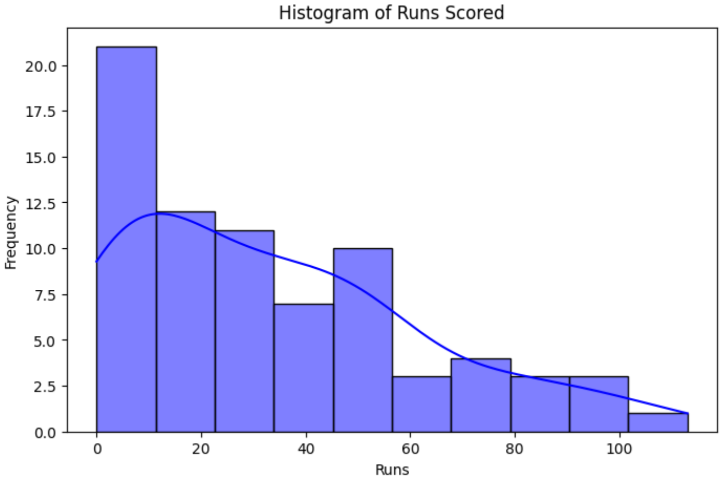
Ogive Curve

```
df_sorted = df.sort_values("Runs")
df_sorted["Cumulative Frequency"] = np.arange(1, len(df_sorted)+1) / len(df_sorted)

plt.figure(figsize=(8,5))
plt.plot(df_sorted["Runs"], df_sorted["Cumulative Frequency"], marker="o", linestyle="-", color="red")
plt.title("Ogive (Cumulative Frequency Curve)")
plt.xlabel("Runs")
plt.ylabel("Cumulative Frequency")
plt.grid()
plt.show()
```



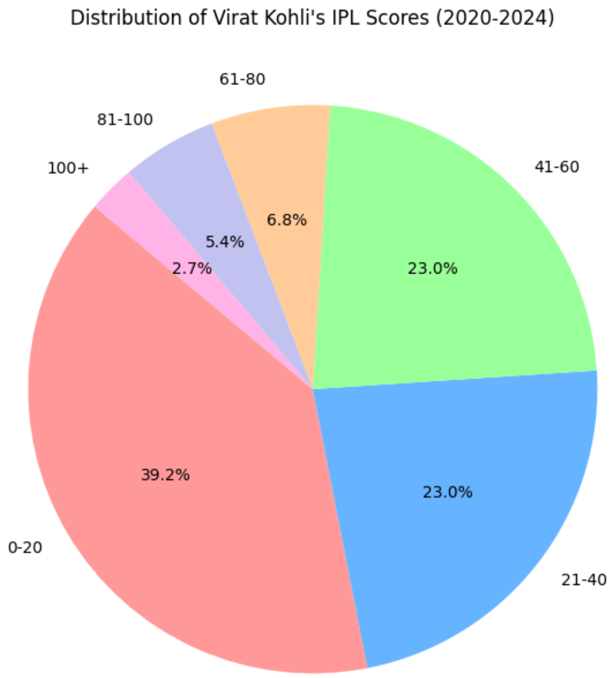
Histogram

```
plt.figure(figsize=(8,5))
sns.histplot(df["Runs"], bins=10, kde=True, color="blue")
plt.title("Histogram of Runs Scored")
plt.xlabel("Runs")
plt.ylabel("Frequency")
plt.show()
```



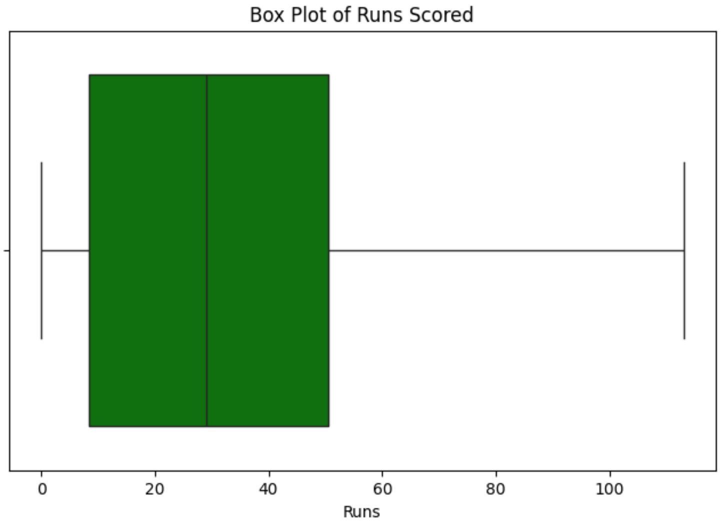
Pie chart

```
# Define score bins
bins = [0, 20, 40, 60, 80, 100, max(df["Runs"])]
labels = ["0-20", "21-40", "41-60", "61-80", "81-100", "100+"]

# Count frequency of runs in each bin
df["Score Range"] = pd.cut(df["Runs"], bins=bins, labels=labels,
    right=False)
score_distribution = df["Score Range"].value_counts().sort_index()

# Plot pie chart
plt.figure(figsize=(8, 8))
colors = ["#ff9999", "#66b3ff", "#99ff99", "#ffcc99", "#c2c2f0", "#
    ffb3e6"]
plt.pie(score_distribution, labels=score_distribution.index,
    autopct="%1.1f%%", colors=colors, startangle=140)

# Title
plt.title("Distribution of Virat Kohli’s IPL Scores (2020-2024)")
plt.show()
```



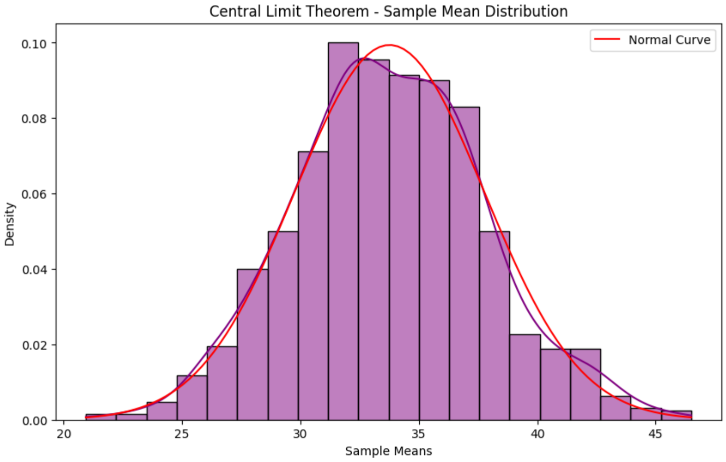
Box plot

```
plt.figure(figsize=(8,5))
sns.boxplot(x=df["Runs"], color="green")
plt.title("Box Plot of Runs Scored")
plt.show()
```



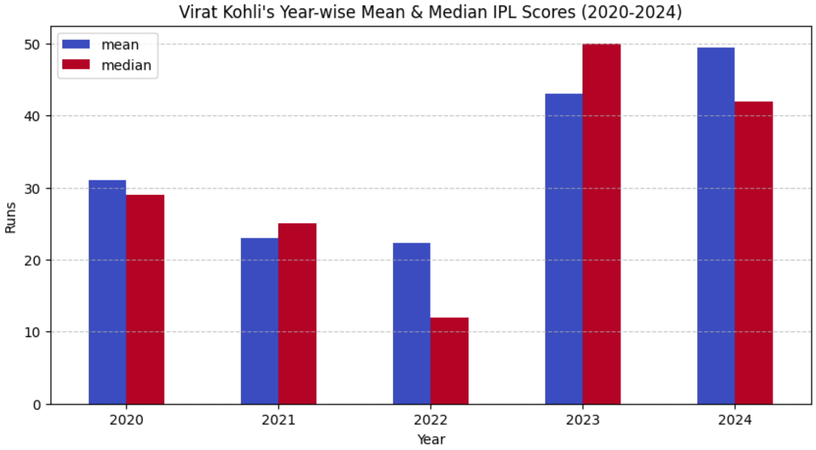
Central Limit Theorem

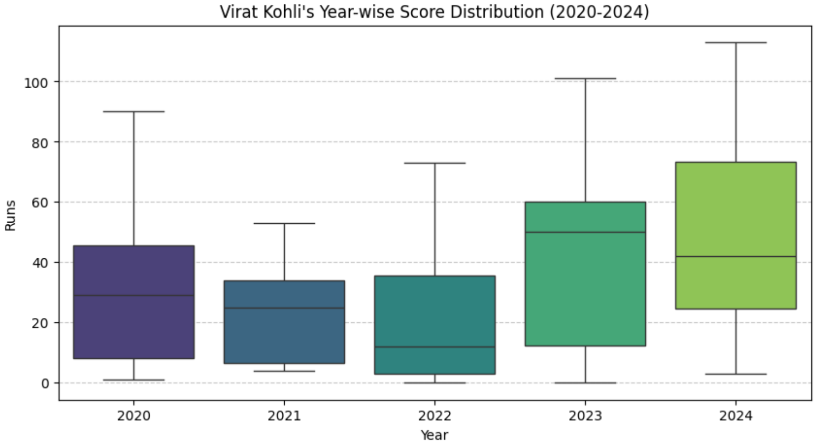
```
sample_size = 50
num_samples = 1000
sample_means = []

for _ in range(num_samples):
    sample = np.random.choice(df["Runs"], size=sample_size, replace
        =True)
    sample_means.append(np.mean(sample))


plt.figure(figsize=(10,6))
sns.histplot(sample_means, bins=20, kde=True, color="purple", stat
    ="density")
x = np.linspace(min(sample_means), max(sample_means), 100)
plt.plot(x, norm.pdf(x, np.mean(sample_means), np.std(sample_means)
    ), color='red', label="Normal Curve")
plt.title("Central Limit Theorem - Sample Mean Distribution")
plt.xlabel("Sample Means")
plt.ylabel("Density")
plt.legend()
plt.show()
```



Year wise Mean,Median,Box plot

```python
# Assign each 15 matches as a different year (assuming sequential order)
df["Year"] = np.repeat([2020, 2021, 2022, 2023, 2024], repeats=15)[:len(df)]

# Compute Descriptive Statistics for Each Year
yearly_stats = df.groupby("Year")["Runs"].describe()[["mean", "50%", "std", "min", "max"]]
yearly_stats.rename(columns={"50%": "median"}, inplace=True)

# Display statistics
print(yearly_stats)

#  **Bar Chart: Year-wise Mean & Median Scores**
plt.figure(figsize=(10,5))
yearly_stats[["mean", "median"]].plot(kind="bar", figsize=(10,5), colormap="coolwarm")
plt.title("Virat Kohli's Year-wise Mean & Median IPL Scores (2020-2024)")
plt.xlabel("Year")
plt.ylabel("Runs")
plt.xticks(rotation=0)
plt.grid(axis="y", linestyle="--", alpha=0.7)
plt.show()

# **Boxplot: Yearly Score Distribution**
plt.figure(figsize=(10,5))
sns.boxplot(x=df["Year"], y=df["Runs"], palette="viridis")
plt.title("Virat Kohli's Year-wise Score Distribution (2020-2024)")
plt.xlabel("Year")
plt.ylabel("Runs")
plt.grid(axis="y", linestyle="--", alpha=0.7)
plt.show()
```


Virat Kohli's Year-wise Mean & Median IPL Scores (2020-2024)
mean
median
Runs
0
10
20
30
40
50
2020
2021
2022
2023
2024
Year


Virat Kohli's Year-wise Score Distribution (2020-2024)
Runs
0
20
40
60
80
100
2020
2021
2022
2023
2024
Year

Observations

Bar Chart

- Shows a rise in Virat Kohli's average runs from 2020 to 2024.
- A dip is observed in 2021 and 2022, followed by an increase in 2023 and 2024.
- Its a performance improvement after a decline.

Box Plot

- The spread of runs is wide, with some matches showing high scores.
- Median is relatively low, suggesting a skewed distribution.

Central Limit Theorem (Sample Mean Distribution)

- The histogram follows a normal distribution.
- Indicates the sample means are normally distributed, supporting statistical inference.

Ogive (Cumulative Frequency Curve)

- Shows that a majority of the scores are under 50.
- The steep rise in the lower score range suggests most scores are concentrated there.
- The curve flattens out at higher values, showing fewer large scores.

Pie Chart (Score Range Distribution)

- Most scores fall in the 0-20 and 21-40 range.
- Fewer innings with scores above 80.
- Indicates consistent scoring but fewer exceptionally high innings.

Year-wise Box Plot

- 2023 and 2024 have wider interquartile ranges, indicating greater variation in scores.
- Median scores have increased in recent years.
- Upper whiskers show more high-scoring matches in 2023-24.

Year-wise Mean and Median

- Mean and median scores show a rising trend in 2023-24.
- Median is lower than the mean in earlier years, suggesting some extreme high scores affected the mean.
- 2023-24 data indicates a more stable performance with improved consistency.

Done By :

- **Puli Dinesh - AI23BTECH11019**
- **Rathod sai dhanush - Ai23Btech11021**
- **Jatavath ajay- Ai23btech11011**