TASK - 1 CODE

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

# Create directories
os.makedirs("data", exist_ok=True)
os.makedirs("plots", exist_ok=True)

# Load dataset
df = pd.read_csv("machine-readable-business-employment-data-dec-2024-quarter.csv")

# Drop completely empty columns
df.drop(columns=["Series_title_4", "Series_title_5"], inplace=True)

# Drop duplicate rows
df.drop_duplicates(inplace=True)

# Drop rows with missing 'Data_value'
df.dropna(subset=["Data_value"], inplace=True)

# Save cleaned data
df.to_csv("data/cleaned_business_employment_data.csv", index=False)

# ---------- EDA ----------

# Summary statistics
print("Summary Statistics:")
print(df.describe())

# Top job types
plt.figure(figsize=(10, 5))
df["Series_title_1"].value_counts().plot(kind="bar", title="Job Types")
plt.tight_layout()
plt.savefig("plots/job_types.png")
plt.close()

# Jobs by industry
industry_jobs = (
    df.groupby("Series_title_2")["Data_value"]
    .sum()
    .sort_values(ascending=False)
    .head(10)
)

plt.figure(figsize=(10, 6))
industry_jobs.plot(kind="barh", title="Top 10 Industries by Total Jobs")
plt.xlabel("Total Jobs")
plt.tight_layout()
plt.savefig("plots/top_industries.png")
plt.close()

# Trend over time for a selected industry
industry = "Construction"
df_subset = df[df["Series_title_2"] == industry]

plt.figure(figsize=(10, 5))
```

```
df_subset.groupby("Period")["Data_value"].sum().plot(title=f"{industry} Jobs Over Time")
plt.ylabel("Filled Jobs")
plt.tight_layout()
plt.savefig("plots/construction_trend.png")
plt.close()

print("✅ Cleaning complete and plots saved to 'plots/' directory.")
```