ex - 1b

October 31, 2025

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re

sns.set(style="whitegrid")

ROLE_KEYWORDS = {
    'Data Scientist': ['data scientist', r'\bds\b', 'machine learning␣
 ↪scientist', 'ml scientist'],
    'Data Engineer': ['data engineer', 'etl engineer', 'pipeline engineer'],
    'Data Analyst': ['data analyst', 'business analyst', 'analyst', 'bi␣
 ↪analyst'],
    'Machine Learning Engineer': ['ml engineer', 'machine learning engineer',␣
 ↪'mle'],
    'BI Developer': ['bi developer', 'business intelligence', 'power bi',␣
 ↪'tableau developer'],
    'Research Scientist': ['research scientist', 'researcher'],
    'Other': []
}

def map_title_to_role(title):
    t = title.lower()
    for role, keys in ROLE_KEYWORDS.items():
        for key in keys:
            if re.search(r'\b' + re.escape(key) + r'\b', t) or key in t:
                return role
    return 'Other'

def categorize_roles(df, title_col='job_title'):
    df = df.copy()
    df[title_col] = df[title_col].astype(str)
    df['role'] = df[title_col].apply(map_title_to_role)
    return df

def plot_role_distribution(df, title_col='role'):
    counts = df[title_col].value_counts().reset_index()
```

```
    counts.columns = ['role', 'count']
    plt.figure(figsize=(10,5))
    sns.barplot(data=counts, x='role', y='count')
    plt.xticks(rotation=45, ha='right')
    plt.title("Distribution of Data Science Roles (bar)")
    plt.tight_layout()
    plt.show()

    plt.figure(figsize=(7,7))
    plt.pie(counts['count'], labels=counts['role'], autopct='%1.1f%%',␣
 ↪startangle=140)
    plt.title("Distribution of Data Science Roles (pie)")
    plt.tight_layout()
    plt.show()

if _name_ == "_main_":
    sample_titles = [
        "Senior Data Scientist", "Junior Data Analyst", "Machine Learning␣
 ↪Engineer",
        "Data Engineer", "BI Developer (Power BI)", "Business Analyst - Data",
        "Research Scientist, ML", "Data Scientist / ML", "Analyst", "ETL␣
 ↪Engineer",
        "Data Scientist", "Data Analyst", "MLOps Engineer", "Data Engineer -␣
 ↪Big Data"
    ]
    df = pd.DataFrame({'job_title': sample_titles})
    df = categorize_roles(df)
    print(df[['job_title','role']])
    plot_role_distribution(df, title_col='role')
```