1. Code Snippet

Univariate EDA on OTT data

```
import matplotlib.pyplot as plt
import seaborn as sns

# Set style for plots
sns.set(style="whitegrid")

# 1. Distribution of `type` (Movie/TV Show)
plt.figure(figsize=(6,4))
sns.countplot(data=data_ott, x='type', palette='Set2')
plt.title('Distribution of OTT Content Type (Movie/TV Show)')
plt.show()
```

```
# 2. Distribution of `release_year`
plt.figure(figsize=(10,6))
sns.histplot(data=data_ott, x='release_year', bins=30, kde=True, color='skyblue')
plt.title('Distribution of Release Year')
plt.show()
```

```
# 3. Top 10 countries with most content
plt.figure(figsize=(10,6))
top_countries = data_ott['country'].value_counts().head(10)
sns.barplot(y=top_countries.index, x=top_countries.values, palette='Set3')
plt.title('Top 10 Countries by Content Count')
plt.show()
```

```
# 4. Rating distribution
plt.figure(figsize=(10,6))
sns.countplot(data=data_ott, y='rating',
order=data_ott['rating'].value_counts().index, palette='coolwarm')
plt.title('Distribution of Ratings')
plt.show()
```

```
# 5. Duration distribution
plt.figure(figsize=(10,6))
sns.histplot(data=data_ott, x='duration', kde=False, color='salmon')
plt.xticks(rotation=90)
plt.title('Distribution of Duration')
plt.show()
```

Univariate EDA on Job Posting Data

```
# Fraudulent Job Posts: Count of real vs. fraudulent jobs.
sns.countplot(data=data_job_posting, x='fraudulent', palette='Set1')
plt.title('Fraudulent vs Non-Fraudulent Jobs')
plt.show()
```

```
# Telecommuting Jobs: Distribution of jobs allowing remote work.
sns.countplot(data=data_job_posting, x='telecommuting', palette='Set2')
plt.title('Distribution of Telecommuting Jobs')
plt.show()
```

```
# Salary Range Distribution: A boxplot of the salary ranges, once they are parsed
into a numerical format.
sns.boxplot(data=data_job_posting, x='salary_range', palette='Set3')
plt.xticks(rotation=90)
plt.title('Salary Range Distribution')
plt.show()
```

Multivariate EDA for OTT dataset

```
 # Content Type vs. Release Year (OTT Dataset):
sns.countplot(data=data_ott, x='release_year', hue='type', palette='Set2')
plt.title('Content Type by Release Year')
plt.show()
```

Multivariate EDA for Job Posting dataset

```
# Fraudulent Jobs vs. Company Profile Completeness (Job Posting Dataset)
sns.boxplot(data=data_job_posting, x='fraudulent', y='company_profile',
palette='coolwarm')
plt.title('Company Profile Completeness in Fraudulent vs Non-Fraudulent Jobs')
plt.show()
```

2. Results

Initial

OTT Data

- Shape: 8,807 rows and 12 columns.
- Columns Overview:
  - show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description
- Data Types: Mostly string (object) except for release_year (integer).
- Missing Values:
  - The columns director, cast, and country contain missing values.
  - date_added, rating, and duration also have a few missing entries.
- Potential Issues:
  - Missing values in several columns.
  - The date_added column likely contains dates as strings, so conversion to a proper date format may be needed.
  - Categorical variables like type, country, rating may require encoding for analysis.

Job Posting Data

- Shape: 17,880 rows and 18 columns.
- Columns Overview:
  - job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, telecommuting, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent
- Data Types: A mix of numerical and categorical variables.
- Missing Values:
  - The columns department, salary_range, company_profile, requirements, benefits, employment_type, required_experience, required_education, industry, and function contain missing data.
- Potential Issues:
  - High amount of missing values in several columns.
  - Binary columns like telecommuting, has_company_logo, and fraudulent are already in numeric format.
  - Columns like salary_range may need to be parsed into numerical values for analysis.

Final

OTT Dataset:

- **Content Type Distribution:** The dataset shows a breakdown of content into two types: Movies and TV Shows.
- **Release Year Distribution:** The content spans several decades, with a concentration of shows and movies being released in more recent years (2010–2020).

- **Top Contributing Countries:** The top countries producing the most content include the United States, India, and the United Kingdom.
- **Rating Distribution:** Content is spread across various ratings (e.g., TV-MA, TV-14, PG-13), with a significant portion aimed at mature audiences.
- **Duration Distribution:** Movies and TV Shows differ in their duration, with most TV shows having multiple seasons and movies varying in runtime.

Job Posting Dataset:

- **Fraudulent vs Non-Fraudulent Jobs:** The dataset contains both legitimate and fraudulent job postings, and the percentage of fraudulent jobs can be quantified.
- **Telecommuting Jobs:** The number of remote jobs is captured, indicating the percentage of jobs that allow telecommuting.
- **Salary Range:** Many entries do not provide a salary range, but from the available data, there is a broad distribution of salaries.
- **Experience and Education Requirements:** A wide range of experience levels and educational qualifications are listed, from entry-level to senior-level roles, and from high school diplomas to advanced degrees.
- **Industry and Function:** The dataset shows which industries and functions are most common in the job market, with Marketing, Sales, and Healthcare appearing frequently.

3. Observations / Findings / Inferences

OTT Dataset:

Observation: Movies dominate the platform compared to TV shows.
Inference: There is a higher demand for movies or more movies are produced and added compared to TV shows.

Observation: A significant portion of the content is released between 2010 and 2020.
Inference: There has been a rapid expansion of digital streaming platforms in the last decade, leading to more content being produced and released.

Observation: Countries like the United States and India lead in content production.
Inference: These countries have large entertainment industries that dominate OTT platforms. Additionally, the rise of regional content is evident with countries like India being a significant contributor.

Observation: The ratings distribution shows a large number of mature-rated content (e.g., TV-MA).
Inference: OTT platforms are more inclined to offer content for mature audiences, possibly due to fewer censorship restrictions compared to traditional broadcasting.

Job Posting Dataset:

Observation: A notable percentage of job postings are flagged as fraudulent.
Inference: There is a significant presence of job scams, which highlights the need for stricter vetting processes on job platforms to prevent fraudulent listings.

Observation: A growing number of jobs allow for telecommuting.

Inference: The rise in remote jobs reflects changes in work culture, possibly accelerated by technological advances and global events like the COVID-19 pandemic.

Observation: Salary ranges are often missing or incomplete.
Inference: Many companies prefer not to disclose salary ranges in job postings, potentially to keep flexibility during negotiations or to remain competitive.

Observation: Education and experience requirements are diverse, ranging from entry-level to senior positions.
Inference: Job postings cater to a wide audience, from fresh graduates to experienced professionals. However, certain industries (e.g., healthcare, technology) may demand higher qualifications.

Observation: Some industries and functions are more prone to fraudulent postings.
Inference: Fraudulent job postings tend to appear in industries with high demand and low barriers to entry (e.g., customer service, marketing), as these positions are easier to exploit.