```
import requests
from bs4 import BeautifulSoup
import pandas as pd

def scrape_indeed_jobs(search_query, location):
    """Scrapes job listings from Indeed based on

    # Indeed search URL
    URL = f"https://in.indeed.com/jobs?q={search_

    # Set User-Agent to avoid blocking
    headers = {
        "User-Agent": "Mozilla/5.0 (Windows NT 10
    }

    # Send request
    response = requests.get(URL, headers=headers)
    soup = BeautifulSoup(response.text, "html.par

    jobs = []

    # Find all
    for job_card in soup.find_all("div", class_="
        # Extract Job Title
        title = job_card.find("h2", {"data-testic
        title = title.text.strip() if title else

        # Extract Company Name
        company = job_card.find("span", class_="j
        company = company.text.strip() if company

        # Extract Location
        location = job_card.find("div", {"data-te
        location = location.text.strip() if locat

        jobs.append([title, company, location])

```

Run this cell to mount your Google Drive.
Learn more
Dismiss

```
    # Convert to DataFrame
    df = pd.DataFrame(jobs, columns=["Job Title",

    return df


# Scrape Data Analyst jobs in Mumbai
job_data = scrape_indeed_jobs("data analyst", "Mu

# Display first 5 rows
print(job_data.head())

# Save to CSV
job_data.to_csv("job_data.csv", index=False)
print("Data saved successfully!")

```

```
Empty DataFrame
Columns: [Job Title, Company, Location]
Index: []
Data saved successfully!
```

```
# Install ChromeDriver and Chromium in Google Col
!apt-get update
!apt-get insta                                  um-chromedriver
!pip install s                                  river-manager

# Set ChromeDriver path for Colab
import os
os.environ["PATH"] += ":/usr/lib/chromium-browser

```

Run this cell to mount your Google Drive.
Learn more

```
Get:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,632 B]
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease [1,581 B]
Get:4 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:7 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  Packages [1,319 kB]
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:11 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,661 kB]
Get:12 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:13 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,704 kB]
Get:14 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [3,664 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,532 kB]
Get:16 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,235 kB]
Get:17 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [2,639 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [3,813 kB]
Get:19 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2,950 kB]
Fetched 28.9 MB in 3s (10.1 MB/s)
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Reading package lists... Done
```

```
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  apparmor chromium-browser libfuse3-3 liblzo2-2 snapd squashfs-tools systemd-hwe-hwdb udev
Suggested packages:
  apparmor-profiles-extra apparmor-utils fuse3 zenity | kdialog
The following NEW packages will be installed:
  apparmor chromium-browser chromium-chromedriver libfuse3-3 liblzo2-2 snapd squashfs-tools
  systemd-hwe-hwdb udev
0 upgraded, 9 newly installed, 0 to remove and 32 not upgraded.
Need to get 30.1 MB of archives.
After this operation, 123 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 apparmor amd64 3.0.4-2ubuntu2.4 [598 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 liblzo2-2 amd64 2.10-2build3 [53.7 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 squashfs-tools amd64 1:4.5-3build1 [159 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 udev amd64 249.11-0ubuntu3.12 [1,557 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/main amd64 libfuse3-3 amd64 3.10.5-1build1 [81.2 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 snapd amd64 2.66.1+22.04 [27.6 MB]
Get:7 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 chromium-browser amd64 1:85.0.4183.83-0ubuntu2.22.04.1 [49.2
Get:8 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 chromium-chromedriver amd64 1:85.0.4183.83-0ubuntu2.22.04.1
Get:9 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 systemd-hwe-hwdb all 249.11.5 [3,228 B]
Fetched 30.1 MB in 2s (17.9 MB/s)
Preconfiguring packages ...
Selecting previously unselected package apparmor.
(Reading database ... 124947 files and directories currently installed.)
Preparing to unpack .../0-apparmor_3.0.4-2ubuntu2.4_amd64.deb ...
Unpacking apparmor (3.0.4-2ubuntu2.4) ...
Selecting previously unselected package liblzo2-2:amd64.
Preparing to unpack .../1-liblzo2-2_2.10-2build3_amd64.deb ...
Unpacking liblzo2-2:amd64 (2.10-2build3) ...
Selecting previously unselected package squashfs-tools.
Preparing to unpack .../2-squashfs-tools_1%3a4.5-3build1_amd64.deb ...
Unpacking squashfs-tools (1:4.5-3build1) ...
```

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriv
from selenium.webdriver.support import expected_c
import pandas as pd
import time

# Set up Seler
options = webdriver.ChromeOptions()
options.add_argument("--headless")  # Run without
options.add_argument("--no-sandbox")
options.add_argument("--disable-dev-shm-usage")
options.add_argument("user-agent=Mozilla/5.0 (Wir

# Use Chromium in Google Colab
driver = webdriver.Chrome(options=options)

def scrape_naukri_jobs(search_query, location):
    """Scrapes job listings from Naukri.com using

    URL = f"https://www.naukri.com/{search_query}
    driver.get(URL)

```

Run this cell to mount your Google Drive.
Learn more

```
    # Wait until job cards load
    try:
        WebDriverWait(driver, 15).until(
            EC.presence_of_element_located((By.CL
        )
    except:
        print("❌ Job listings did not load!")
        return pd.DataFrame(columns=["Job Title",

    jobs = []

    # Extract job listings
    job_cards = driver.find_elements(By.CLASS_NAM
    for job in job_cards:
        try:
            title = job.find_element(By.CLASS_NAM
        except:
            title = "Not available"

        try:
            company = job.find_element(By.CLASS_N
        except
            co                    available"

        try:
            location = job.find_element(By.CLASS_
        except:
            location = "Not available"

        try:
            salary = job.find_element(By.CLASS_NA
        except:
            salary = "Not available"

        try:
            skills = ", ".join([skill.text.strip(
        except:
            skills = "Not available"
```

Run this cell to mount your Google Drive.
Learn more

```
        jobs.append([title, company, location, sa

    df = pd.DataFrame(jobs, columns=["Job Title",

    return df


# Scrape Data Analyst jobs in Mumbai from Naukri
job_data = scrape_naukri_jobs("data-analyst", "mu

# Close browser
driver.quit()

# Display first 5 rows
print(job_data.head())

# Save Data to CSV
job_data.to_csv("naukri_jobs.csv", index=False)
print("Data saved successfully!")

```

Run this cell to mount your Google Drive.
Learn more

```
0             Data Analyst (World Pane
1          GN- Public Service - Data & 
2           Data Analyst - Monitoring 
3  Opening For Data Analyst -(DME) For Mumbai loc...
4                                   Data Analyst

                      Company                     Location         Salary  \
0                   Numerator              Hybrid - Mumbai  3-3.5 Lacs PA
1                   Accenture  Mumbai, Gurugram, Bengaluru  Not disclosed
2  Aditya Birla Education Trust    Mumbai (All Areas)(Worli)    5-7 Lacs PA
3       Inland World Logistics                       Mumbai  Not disclosed
4               HH Consultancy                       Remote    7-9 Lacs PA

                                          Skills
0  Advanced Excel, R, Mac, Python, SQL, Excel, Pa...
1  data quality, data modeling, dashboards, artif...
2  Data Analysis Tools, Data Analysis, Programmin...
3  Data Analysis, Data Management, Data Reporting...
4  Data Analysis, Data Manipulation, Data Managem...
Data saved successfully!
```

```
import pandas as pd

df = pd.read_csv("naukri_jobs.csv")

# Top 5 highest-paying jobs
print(df[df['Salary'] != "Not disclosed"].sort_va

```

```
# Most common skills
from collections import Counter
skills = ", ".join(df["Skills"].dropna()).split("
top_skills = Counter(skills).most_common(10)
print("Top In-Demand Skills:", top_skills)

```

```
                Job Title                       Company  \
17           Data Analyst                         WebMD
4            Data Analyst                 HH Consultancy
6            Data Analyst                            ETG
9   Data Analyst || German                           mnc
8            Data Analyst  Good2Great Industries Pvt Ltd

                    Location           Salary  \
17        Mumbai (All Areas)  9.5-19.5 Lacs PA
4                    Remote       7-9 Lacs PA
6                Navi Mumbai       7-9 Lacs PA
9   Mumbai (All Areas), Pune      7-14 Lacs PA
8        Mumbai (All Areas)       6-8 Lacs PA

                                            Skills
17  SQL, Power BI, Data Analyst, Tableau, Bi, Data...
4   Data Analysis, Data Manipulation, Data Managem...
6   SQL, Python, Power BI, data modeling, dashboar...
9   German Language, Translation, German, Language...
8   Data Analysis, Advanced Excel, Statistical Dat...
Top In-Demand Skills: [('Data', 8), ('Data analysis', 7), ('Data Analysis', 7), ('Analysis', 6), ('Data Analyst', 5), ('Data Managem
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
inputfile= "/c                    _jobs.csv"
df=pd.read_csv

# Remove "Lacs PA" and clean the column
df["Salary"] = df["Salary"].str.replace(" Lacs PA

# Handle cases where salary is missing or "Not di
df["Salary"] = df["Salary"].apply(lambda x: x.spl
df["Salary"] = pd.to_numeric(df["Salary"], errors

# Fill missing salaries with the mean salary
df["Salary"].fillna(df["Salary"].mean(), inplace=

print(df.head())  # Check the fixed Salary column

```

Run this cell to mount your Google Drive.
Learn more

```
                                     Job Title  \
0             Data Analyst (World Panel by Kantar)
1          GN- Public Service - Data & AI - Analyst
2            Data Analyst - Monitoring & Evaluation
3  Opening For Data Analyst -(DME) For Mumbai loc...
```

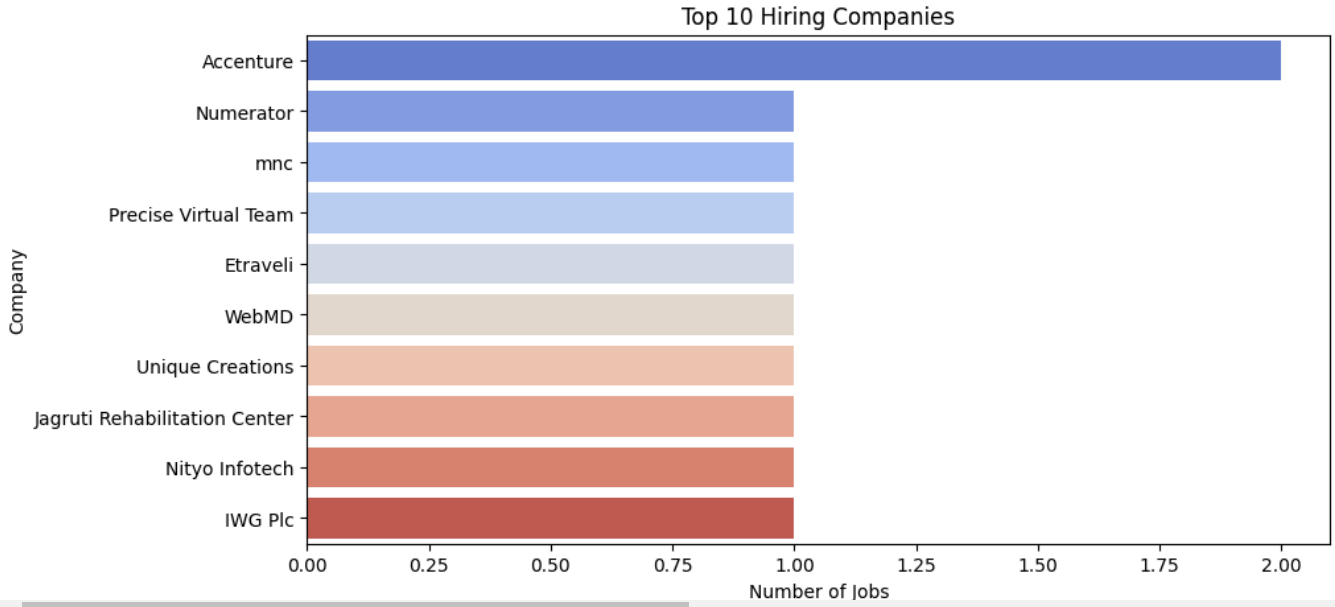
```
4                                   Data Analyst

                       Company                     Location    Salary  \
0                    Numerator              Hybrid - Mumbai  3.000000
1                    Accenture  Mumbai, Gurugram, Bengaluru  5.833333
2  Aditya Birla Education Trust    Mumbai (All Areas)(Worli)  5.000000
3        Inland World Logistics                       Mumbai  5.833333
4                HH Consultancy                       Remote  7.000000

                                           Skills
0  Advanced Excel, R, Mac, Python, SQL, Excel, Pa...
1  data quality, data modeling, dashboards, artif...
2  Data Analysis Tools, Data Analysis, Programmin...
3  Data Analysis, Data Management, Data Reporting...
4  Data Analysis, Data Manipulation, Data Managem...
<ipython-input-6-752260c4ff9f>:16: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained ass
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col


  df["Salary"].fillna(df["Salary"].mean(), inplace=True)
```

1 **Top Hiring Companies**

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

# Count the top 10 hiring companies
top_companies = df["Company"].value_counts().head

# Set figure size
plt.figure(figsize=(10, 5))

# Use Seaborn'
sns.barplot(x=                    s.values, y=top_compani

# Set labels and title
plt.xlabel("Number of Jobs")
plt.ylabel("Company")
plt.title("Top 10 Hiring Companies")

# Show the plot
plt.show()
```

Run this cell to mount your Google Drive.
Learn more

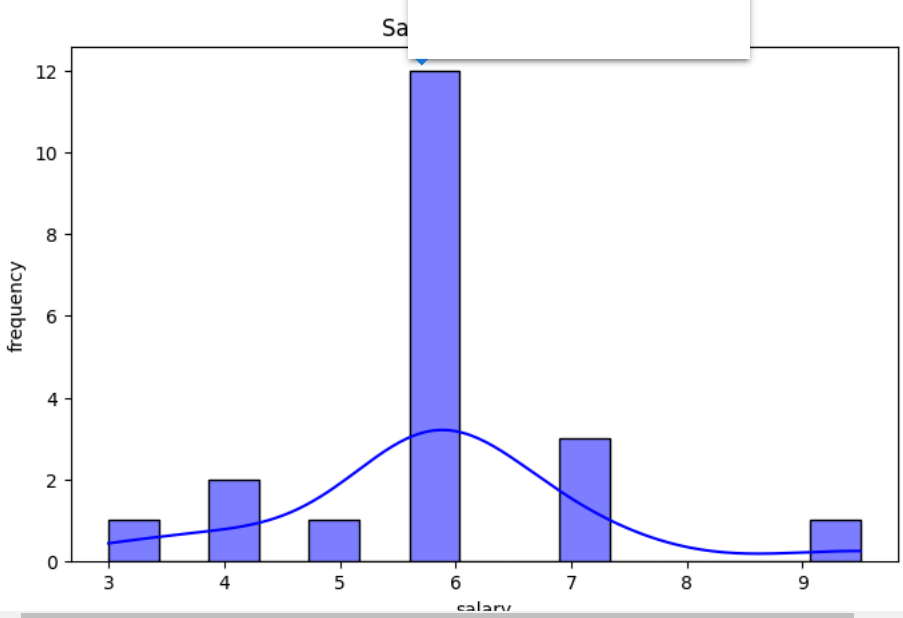
```
<ipython-input-10-8abe968e1c3c>:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=top_companies.values, y=top_companies.index, palette="coolwarm")
```



2 Salary Distribution Analysis

```
plt.figure(figsize=(8,5))
sns.histplot(df['Salary'],bins=15,kde=True,color=
plt.xlabel('salary')
plt.ylabel('frequency')
plt.title('Salary Distribution')
plt.show()
```

Run this cell to mount your Google Drive.
Learn more

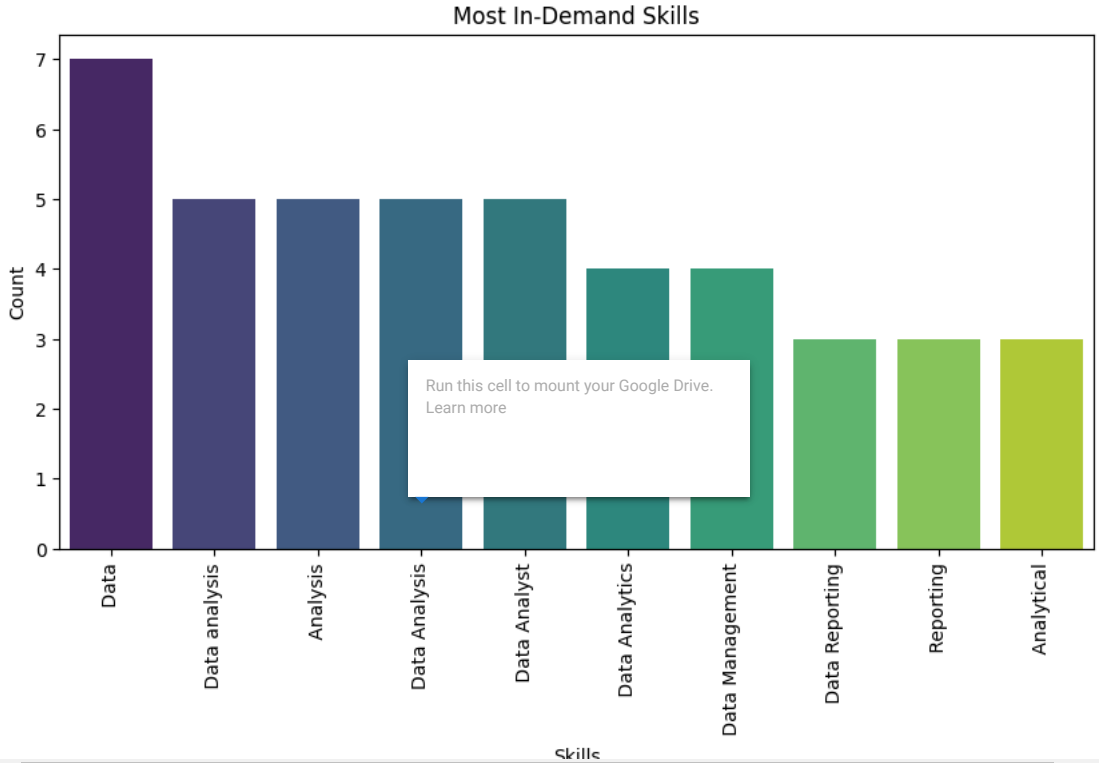


3 Most In-Demand Skills

```
from collections import Counter

all_skills = ' '.join(df['Skills'].dropna()).split
```

```
all_skills=', '.join(df['Skills'].dropna()).split
top_skills=Counter(all_skills).most_common(10)

skills_df=pd.DataFrame(top_skills,columns=['Skill

plt.figure(figsize=(10,5))
sns.barplot(x=skills_df['Skills'],y=skills_df['Co
plt.xlabel('Skills')
plt.ylabel('Count')
plt.title('Most In-Demand Skills')
plt.xticks(rotation=90)
plt.show()
```

```
<ipython-input-15-49f1bb55ffbc>:9: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `le

  sns.barplot(x=skills_df['Skills'],y=skills_df['Count'],palette="viridis")
```




```
from google.colab import drive
drive.mount('/content/drive')
```

4 Job Location Trends

```
top_locations = df["Location"].value_counts().hea

plt.figure(figsize=(10, 5))
sns.barplot(x=top_locations.values, y=top_locatic
```

```
plt.xlabel("Number of Job Postings")
plt.ylabel("Location")
plt.title("Top 10 Job Locations for Data Analysts
plt.show()
```

```
<ipython-input-16-5aff73691f4d>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=top_locations.values, y=top_locations.index, palette="mako")
```


```
# Save the cleaned dataset for visualization in P
df.to_csv("cle                    jobs.csv", index=False)

print("Cleaned data saved successfully!")
```

Run this cell to mount your Google Drive.
Learn more

```
Cleaned data saved successfully!
```

```
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_sp
from sklearn.linear_model import LinearRegressior
from sklearn.ensemble import RandomForestRegressc
from sklearn.metrics import mean_absolute_error
from sklearn.preprocessing import MultiLabelBinar
from sklearn.decomposition import PCA

```

```
# Load dataset
file_path = "/content/cleaned_naukri_jobs.csv"
df = pd.read_csv(file_path)

# One-hot encoding for categorical variables
df = pd.get_dummies(df, columns=["Company", "Loca

# Ensure Salary is numeric
df["Salary"] = pd.to_numeric(df["Salary"], errors

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

# Convert Skills into Numeric Format
df["Skills"] = df["Skills"].fillna("").apply(lamb
mlb = MultiLabelBinarizer()
skills_encoded = pd.DataFrame(mlb.fit_transform(d
df = pd.concat([df, skills_encoded], axis=1)
df.drop(columns=["Skills"], inplace=True)

# Remove constant columns (if any)
constant_colum                    col in df.columns if
df.drop(colum                    olumns, inplace=True)

# Identify and remove low-variance features
threshold = 0.95
low_variance_cols = [col for col in df.columns if
df.drop(columns=low_variance_cols, inplace=True)
print(f"Dropped {len(low_variance_cols)} low-vari

# Prepare data for modeling
X = df.drop(columns=["Salary"])
y = df["Salary"]
X_train, X_test, y_train, y_test = train_test_spl

# Train Linear Regression Model
lin_model = LinearRegression()
lin_model.fit(X_train, y_train)
```

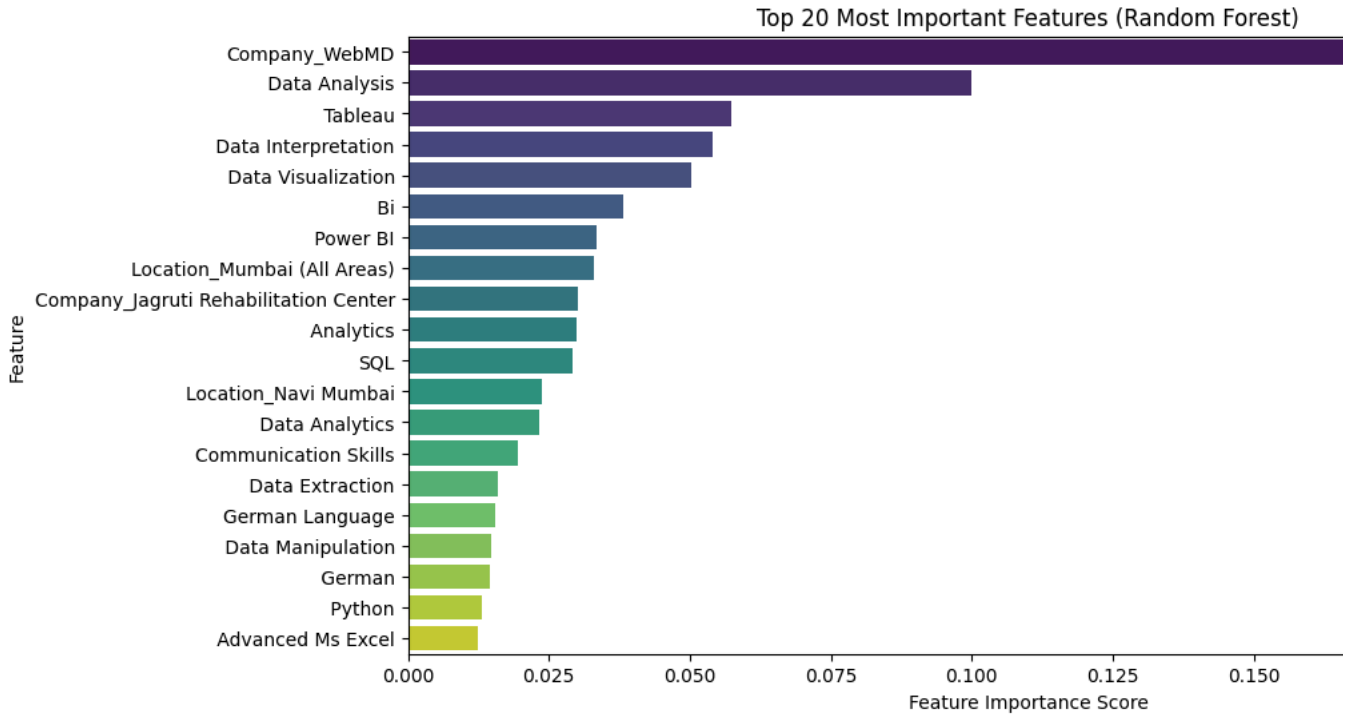
Run this cell to mount your Google Drive.
Learn more

```
y_pred_lin = lin_model.predict(X_test)
mae_lin = mean_absolute_error(y_test, y_pred_lin)
print(f"Linear Regression MAE: {mae_lin:.2f}")

# Train Random Forest Model
rf_model = RandomForestRegressor(n_estimators=100
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
print(f"Random Forest MAE: {mae_rf:.2f}")

# Feature Importance (Random Forest)
feature_importance_rf = pd.Series(rf_model.featur

# Plot Feature Importance
plt.figure(figsize=(12, 6))
sns.barplot(x=feature_importance_rf[:20].values,
plt.xlabel("Feature Importance Score")
plt.ylabel("Feature")
plt.title("Top 20 Most Important Features (Random
plt.show()

# Apply PCA fo                    ity Reduction
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

pca = PCA(n_components=0.95)  # Keep 95% variance
X_pca = pca.fit_transform(X_scaled)

print(f"Original feature count: {X.shape[1]}")
print(f"Reduced feature count after PCA: {X_pca.s

# Predict Salary for a new job with Python, SQL,
new_job_skills = pd.DataFrame([[0] * len(X.columr
for skill in ["Python", "SQL", "Power BI"]:
    if skill in new_job_skills.columns:
        new_job_skills[skill] = 1

```

Run this cell to mount your Google Drive.
Learn more

```
predicted_salary = rf_model.predict(new_job_skill
print(f"💰 Predicted Salary for Python, SQL, Pow
```

```
Dropped 0 low-variance columns.
Linear Regression MAE: 0.90
Random Forest MAE: 0.84
<ipython-input-22-ec954b8f84f4>:67: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=feature_importance_rf[:20].values, y=feature_importance_rf[:20].index, palette="viridis")
```




```
Original feature count: 148
Reduced feature count after PCA: 17
```

Run this cell to mount your Google Drive.
Learn more

```
!pip install x                    t-learn pandas numpy ma
```

```
Requirement already satisfied: xgboost in /usr/local/lib/python3.11/dist-packages (2.1.4)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (1.26.4)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (3.10.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (0.13.2)
Requirement already satisfied: nvidia-nccl-cu12 in /usr/local/lib/python3.11/dist-packages (from xgboost) (2.21.5)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from xgboost) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

```
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
```

```
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_sp
from sklearn.metrics import mean_absolute_error
from sklearn.preprocessing import MultiLabelBinar
import xgboost as xgb

# Load dataset
file_path = "/content/cleaned_naukri_jobs.csv"
df = pd.read_csv(file_path)

# One-hot encode categorical columns
df = pd.get_dummies(df, columns=["Company", "Loca

# Convert Salary column to numeric
df["Salary"] = pd.to_numeric(df["Salary"], errors

# Drop missing salary values
df.dropna(subset=["Salary"], inplace=True)

# Convert Skills into numeric format
df["Skills"] = df["Skills"].fillna("").apply(lamb

# MultiLabelBi                    skills encoding
mlb = MultiLabelBinarizer()
skills_encoded = pd.DataFrame(mlb.fit_transform(d
df = pd.concat([df, skills_encoded], axis=1)

# Drop original Skills column
df.drop(columns=["Skills"], inplace=True)

# Define features (X) and target variable (y)
X = df.drop(columns=["Salary"])
y = df["Salary"]

# Train-test split (80-20)
X_train, X_test, y_train, y_test = train_test_spl
```

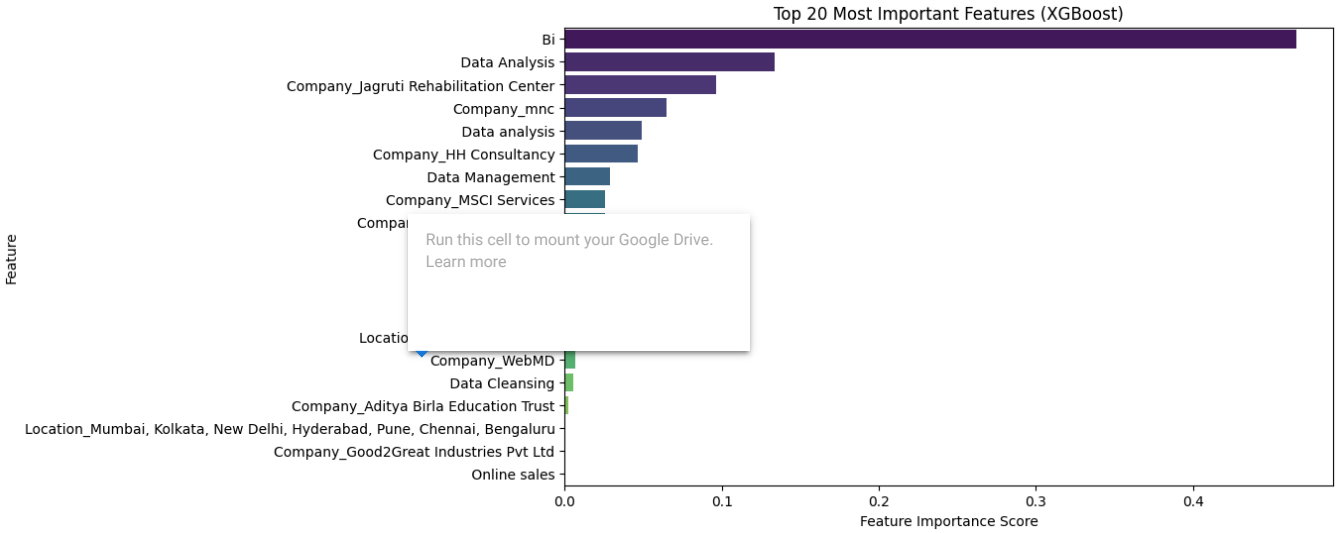
Run this cell to mount your Google Drive.
Learn more

```
# Define XGBoost model
xgb_model = xgb.XGBRegressor(objective="reg:squar

# Hyperparameter grid
param_grid = {
    "n_estimators": [100, 200, 300],  # Number of
    "learning_rate": [0.01, 0.1, 0.2],  # Step si
    "max_depth": [3, 5, 7]  # Tree depth
}

# GridSearchCV to find the best hyperparameters
grid_search = GridSearchCV(xgb_model, param_grid,
grid_search.fit(X_train, y_train)

# Best model after tuning
best_xgb_model = grid_search.best_estimator_

# Predict on test set
y_pred_xgb = best_xgb_model.predict(X_test)

# Calculate Mean Absolute Error (MAE)
mae_xgb = mean                                  ror(y_test, y_pred_xgb)
print(f" 🔹 XG                                  AE: {mae_xgb:.2f}")


# Get feature importance
feature_importance_xgb = pd.Series(best_xgb_model
feature_importance_xgb = feature_importance_xgb.s

# Plot Top 20 Features
plt.figure(figsize=(10, 6))
sns.barplot(x=feature_importance_xgb[:20].values,
plt.title("Top 20 Most Important Features (XGBoos
plt.xlabel("Feature Importance Score")
plt.ylabel("Feature")
plt.show()

# Create a new job sample with Python, SQL, Power
new_job_skills = pd.DataFrame([[0] * len(X_column
```

Run this cell to mount your Google Drive.
Learn more

```
new_job_skills = pd.DataFrame([[0]*len(X.column
# Set required skills to 1
for skill in ["Python", "SQL", "Power BI"]:
    if skill in new_job_skills.columns:
        new_job_skills[skill] = 1

# Predict salary
predicted_salary_xgb = best_xgb_model.predict(new_
print(f"💰 Predicted Salary for Python, SQL, Powe
```

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
🔹 XGBoost Model MAE: 0.76
<ipython-input-24-b72690ff7ed9>:74: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=feature_importance_xgb[:20].values, y=feature_importance_xgb[:20].index, palette="viridis")
```




```
💰 Predicted Salary for Python, SQL, Power BI: 5.85 Lacs PA
```