ds-job-roles-uk

November 12, 2024

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df=pd.read_csv('/content/drive/MyDrive/Kaggle Projects/Raw_Dataset.csv')
```

[3]:
```
df.head()
```

[3]:
```
         Company  Company Score                                   Job Title  \
0     Razorpoint            3.4                       Junior Data Scientist
1  tower Hamlets            3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW            4.0                              Data Scientist
3  NatWest Group            4.6                              Data Scientist
4          iwoca            3.9                        Data Scientist - Ops

               Location  Date                          Salary  \
0  Manchester, England    3d              £35K (Employer est.)
1      London, England    5d  £31.00 Per Hour (Employer est.)
2  Nottingham, England  30d+      £50K - £65K (Employer est.)
3  Edinburgh, Scotland    2d     £41K - £54K (Glassdoor est.)
4      London, England    7d      £60K - £90K (Employer est.)

                                              Skills
0  Data mining, Big data, R, Data analysis skills…
1                       R, SQL, JavaScript, Python
2                  SQL, Maths, Data science, Python
3  Software deployment, Data analysis skills, Sta…
4              MATLAB, R, Maths, C, Machine learning
```

[4]:
```
df.describe(include='all')
```

[4]:
```
                      Company  Company Score       Job Title         Location  \
count                     750     697.000000             750              750
unique                    454            NaN             563               87
top      JPMorgan Chase & Co             NaN  Data Scientist  London, England
freq                       13            NaN              67              445
```

```
mean                    NaN       3.847633             NaN              NaN
std                     NaN       0.461629             NaN              NaN
min                     NaN       1.700000             NaN              NaN
25%                     NaN       3.600000             NaN              NaN
50%                     NaN       3.900000             NaN              NaN
75%                     NaN       4.100000             NaN              NaN
max                     NaN       5.000000             NaN              NaN

        Date                          Salary  \
count    750                             635
unique    30                             447
top      30d+  £70K - £110K (Glassdoor est.)
freq     426                               8
mean     NaN                             NaN
std      NaN                             NaN
min      NaN                             NaN
25%      NaN                             NaN
50%      NaN                             NaN
75%      NaN                             NaN
max      NaN                             NaN

                                                  Skills
count                                                742
unique                                               660
top     Machine learning, Natural language processing,…
freq                                                   9
mean                                                 NaN
std                                                  NaN
min                                                  NaN
25%                                                  NaN
50%                                                  NaN
75%                                                  NaN
max                                                  NaN
```

[5]:
```
!pip install ydata_profiling
!pip install ipywidgets
```

```
Collecting ydata_profiling
  Downloading ydata_profiling-4.12.0-py2.py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: scipy<1.14,>=1.4.1 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (1.13.1)
Requirement already satisfied: pandas!=1.4.0,<3,>1.1 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (2.2.2)
Requirement already satisfied: matplotlib<3.10,>=3.5 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (3.8.0)
Requirement already satisfied: pydantic>=2 in /usr/local/lib/python3.10/dist-
packages (from ydata_profiling) (2.9.2)
```

```
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (6.0.2)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (3.1.4)
Collecting visions<0.7.7,>=0.7.5 (from
visions[type_image_path]<0.7.7,>=0.7.5->ydata_profiling)
  Downloading visions-0.7.6-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: numpy<2.2,>=1.16.0 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (1.26.4)
Collecting htmlmin==0.1.12 (from ydata_profiling)
  Downloading htmlmin-0.1.12.tar.gz (19 kB)
  Preparing metadata (setup.py) … done
Collecting phik<0.13,>=0.11.1 (from ydata_profiling)
  Downloading
phik-0.12.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(5.6 kB)
Requirement already satisfied: requests<3,>=2.24.0 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (2.32.3)
Requirement already satisfied: tqdm<5,>=4.48.2 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (4.66.6)
Requirement already satisfied: seaborn<0.14,>=0.10.1 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (0.13.2)
Collecting multimethod<2,>=1.4 (from ydata_profiling)
  Downloading multimethod-1.12-py3-none-any.whl.metadata (9.6 kB)
Requirement already satisfied: statsmodels<1,>=0.13.2 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (0.14.4)
Requirement already satisfied: typeguard<5,>=3 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (4.4.1)
Collecting imagehash==4.3.1 (from ydata_profiling)
  Downloading ImageHash-4.3.1-py2.py3-none-any.whl.metadata (8.0 kB)
Requirement already satisfied: wordcloud>=1.9.3 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (1.9.3)
Collecting dacite>=1.8 (from ydata_profiling)
  Downloading dacite-1.8.1-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: numba<1,>=0.56.0 in
/usr/local/lib/python3.10/dist-packages (from ydata_profiling) (0.60.0)
Collecting PyWavelets (from imagehash==4.3.1->ydata_profiling)
  Downloading pywavelets-1.7.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x
86_64.whl.metadata (9.0 kB)
Requirement already satisfied: pillow in /usr/local/lib/python3.10/dist-packages
(from imagehash==4.3.1->ydata_profiling) (10.4.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
jinja2<3.2,>=2.11.1->ydata_profiling) (3.0.2)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (1.3.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-
```

```
packages (from matplotlib<3.10,>=3.5->ydata_profiling) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (4.54.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (1.4.7)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (24.1)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (3.2.0)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from
matplotlib<3.10,>=3.5->ydata_profiling) (2.8.2)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in
/usr/local/lib/python3.10/dist-packages (from numba<1,>=0.56.0->ydata_profiling)
(0.43.0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas!=1.4.0,<3,>1.1->ydata_profiling) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-
packages (from pandas!=1.4.0,<3,>1.1->ydata_profiling) (2024.2)
Requirement already satisfied: joblib>=0.14.1 in /usr/local/lib/python3.10/dist-
packages (from phik<0.13,>=0.11.1->ydata_profiling) (1.4.2)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic>=2->ydata_profiling)
(0.7.0)
Requirement already satisfied: pydantic-core==2.23.4 in
/usr/local/lib/python3.10/dist-packages (from pydantic>=2->ydata_profiling)
(2.23.4)
Requirement already satisfied: typing-extensions>=4.6.1 in
/usr/local/lib/python3.10/dist-packages (from pydantic>=2->ydata_profiling)
(4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.24.0->ydata_profiling) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests<3,>=2.24.0->ydata_profiling) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.24.0->ydata_profiling) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.24.0->ydata_profiling) (2024.8.30)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.10/dist-
packages (from statsmodels<1,>=0.13.2->ydata_profiling) (0.5.6)
Requirement already satisfied: attrs>=19.3.0 in /usr/local/lib/python3.10/dist-
```

```
packages (from
visions<0.7.7,>=0.7.5->visions[type_image_path]<0.7.7,>=0.7.5->ydata_profiling)
(24.2.0)
Requirement already satisfied: networkx>=2.4 in /usr/local/lib/python3.10/dist-
packages (from
visions<0.7.7,>=0.7.5->visions[type_image_path]<0.7.7,>=0.7.5->ydata_profiling)
(3.4.2)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from patsy>=0.5.6->statsmodels<1,>=0.13.2->ydata_profiling) (1.16.0)
Downloading ydata_profiling-4.12.0-py2.py3-none-any.whl (390 kB)
                         390.6/390.6 kB
19.9 MB/s eta 0:00:00
Downloading ImageHash-4.3.1-py2.py3-none-any.whl (296 kB)
                         296.5/296.5 kB
23.8 MB/s eta 0:00:00
Downloading dacite-1.8.1-py3-none-any.whl (14 kB)
Downloading multimethod-1.12-py3-none-any.whl (10 kB)
Downloading
phik-0.12.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (686 kB)
                         686.1/686.1 kB
29.8 MB/s eta 0:00:00
Downloading visions-0.7.6-py3-none-any.whl (104 kB)
                         104.8/104.8 kB
9.9 MB/s eta 0:00:00
Downloading
pywavelets-1.7.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.5
MB)
                         4.5/4.5 MB
57.4 MB/s eta 0:00:00
Building wheels for collected packages: htmlmin
  Building wheel for htmlmin (setup.py) … done
  Created wheel for htmlmin: filename=htmlmin-0.1.12-py3-none-any.whl size=27081
sha256=e9641d58f63f1ed5d5933e206255eec496ae8d50d5370a76d74ad52e8c618963
  Stored in directory: /root/.cache/pip/wheels/dd/91/29/a79cecb328d01739e64017b6
fb9a1ab9d8cb1853098ec5966d
Successfully built htmlmin
Installing collected packages: htmlmin, PyWavelets, multimethod, dacite,
imagehash, visions, phik, ydata_profiling
Successfully installed PyWavelets-1.7.0 dacite-1.8.1 htmlmin-0.1.12
imagehash-4.3.1 multimethod-1.12 phik-0.12.4 visions-0.7.6
ydata_profiling-4.12.0
Requirement already satisfied: ipywidgets in /usr/local/lib/python3.10/dist-
packages (7.7.1)
Requirement already satisfied: ipykernel>=4.5.1 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets) (5.5.6)
Requirement already satisfied: ipython-genutils~=0.2.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets) (0.2.0)
Requirement already satisfied: traitlets>=4.3.1 in
```

```
/usr/local/lib/python3.10/dist-packages (from ipywidgets) (5.7.1)
Requirement already satisfied: widgetsnbextension~=3.6.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets) (3.6.10)
Requirement already satisfied: ipython>=4.0.0 in /usr/local/lib/python3.10/dist-
packages (from ipywidgets) (7.34.0)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets) (3.0.13)
Requirement already satisfied: jupyter-client in /usr/local/lib/python3.10/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (6.1.12)
Requirement already satisfied: tornado>=4.2 in /usr/local/lib/python3.10/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (6.3.3)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.10/dist-packages (from ipython>=4.0.0->ipywidgets)
(75.1.0)
Collecting jedi>=0.16 (from ipython>=4.0.0->ipywidgets)
  Downloading jedi-0.19.2-py2.py3-none-any.whl.metadata (22 kB)
Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-
packages (from ipython>=4.0.0->ipywidgets) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-
packages (from ipython>=4.0.0->ipywidgets) (0.7.5)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipython>=4.0.0->ipywidgets)
(3.0.48)
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-
packages (from ipython>=4.0.0->ipywidgets) (2.18.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-
packages (from ipython>=4.0.0->ipywidgets) (0.2.0)
Requirement already satisfied: matplotlib-inline in
/usr/local/lib/python3.10/dist-packages (from ipython>=4.0.0->ipywidgets)
(0.1.7)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-
packages (from ipython>=4.0.0->ipywidgets) (4.9.0)
Requirement already satisfied: notebook>=4.4.1 in
/usr/local/lib/python3.10/dist-packages (from
widgetsnbextension~=3.6.0->ipywidgets) (6.5.5)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in
/usr/local/lib/python3.10/dist-packages (from
jedi>=0.16->ipython>=4.0.0->ipywidgets) (0.8.4)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (3.1.4)
Requirement already satisfied: pyzmq<25,>=17 in /usr/local/lib/python3.10/dist-
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (24.0.1)
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.10/dist-
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (23.1.0)
Requirement already satisfied: jupyter-core>=4.6.1 in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (5.7.2)
Requirement already satisfied: nbformat in /usr/local/lib/python3.10/dist-
```

```
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (5.10.4)
Requirement already satisfied: nbconvert>=5 in /usr/local/lib/python3.10/dist-
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (7.16.4)
Requirement already satisfied: nest-asyncio>=1.5 in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.6.0)
Requirement already satisfied: Send2Trash>=1.8.0 in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.8.3)
Requirement already satisfied: terminado>=0.8.3 in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.18.1)
Requirement already satisfied: prometheus-client in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.21.0)
Requirement already satisfied: nbclassic>=0.4.7 in
/usr/local/lib/python3.10/dist-packages (from
notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.1.0)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python3.10/dist-packages (from jupyter-
client->ipykernel>=4.5.1->ipywidgets) (2.8.2)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.10/dist-packages (from
pexpect>4.3->ipython>=4.0.0->ipywidgets) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.10/dist-
packages (from prompt-
toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython>=4.0.0->ipywidgets) (0.2.13)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.10/dist-packages (from jupyter-
core>=4.6.1->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (4.3.6)
Requirement already satisfied: notebook-shim>=0.2.3 in
/usr/local/lib/python3.10/dist-packages (from
nbclassic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets)
(0.2.4)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-
packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (4.12.3)
Requirement already satisfied: bleach!=5.0.0 in /usr/local/lib/python3.10/dist-
packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (6.2.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.10/dist-
packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.7.1)
Requirement already satisfied: jupyterlab-pygments in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.3.0)
Requirement already satisfied: markupsafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
```

```
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (3.0.2)
Requirement already satisfied: mistune<4,>=2.0.3 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (3.0.2)
Requirement already satisfied: nbclient>=0.5.0 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.10.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (24.1)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.5.1)
Requirement already satisfied: tinycss2 in /usr/local/lib/python3.10/dist-
packages (from
nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.4.0)
Requirement already satisfied: fastjsonschema>=2.15 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (2.20.0)
Requirement already satisfied: jsonschema>=2.6 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (4.23.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.1->jupyter-
client->ipykernel>=4.5.1->ipywidgets) (1.16.0)
Requirement already satisfied: argon2-cffi-bindings in
/usr/local/lib/python3.10/dist-packages (from
argon2-cffi->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (21.2.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-
packages (from bleach!=5.0.0->nbconvert>=5->notebook>=4.4.1->widgetsnbextension~
=3.6.0->ipywidgets) (0.5.1)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook>=4.4.1->widgetsnbextension~=3
.6.0->ipywidgets) (24.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=2.6->nbformat->noteboo
k>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=2.6->nbformat->noteboo
k>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (0.35.1)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook>=4.4.1->widgetsnbextension~=3
.6.0->ipywidgets) (0.20.1)
Requirement already satisfied: jupyter-server<3,>=1.8 in
/usr/local/lib/python3.10/dist-packages (from notebook-shim>=0.2.3->nbclassic>=0
.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.24.0)
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.10/dist-
packages (from argon2-cffi-
```

```
bindings->argon2-cffi->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets)
(1.17.1)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-
packages (from beautifulsoup4->nbconvert>=5->notebook>=4.4.1->widgetsnbextension
~=3.6.0->ipywidgets) (2.6)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-
packages (from cffi>=1.0.1->argon2-cffi-
bindings->argon2-cffi->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets)
(2.22)
Requirement already satisfied: anyio<4,>=3.1.0 in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-s
him>=0.2.3->nbclassic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywid
gets) (3.7.1)
Requirement already satisfied: websocket-client in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-s
him>=0.2.3->nbclassic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywid
gets) (1.8.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-shim>=0.2.3->nb
classic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (3.10)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-shim>=0.2.3->nb
classic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-shim>=0.2.3->nb
classic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets) (1.2.2)
Downloading jedi-0.19.2-py2.py3-none-any.whl (1.6 MB)
                           1.6/1.6 MB
54.0 MB/s eta 0:00:00
Installing collected packages: jedi
Successfully installed jedi-0.19.2
```

[6]:
```
from ydata_profiling import ProfileReport # Import ProfileReport class
```

[7]:
```
profile = ProfileReport(df,title='Profile Summary')
profile.to_notebook_iframe() # Corrected the typo here
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[8]:
```
# Replacing pound sign '£' to ' ' And Replacing 'k' to '000'

df['Salary']=df['Salary'].str.replace("£",'').str.replace("K","000")
df
```

[8]:

	Company	Company Score	Job Title	Location	Date	Salary
0	Razorpoint	3.4	Junior Data Scientist	Manchester, England	3d	35000 (Employer est.)
1	tower Hamlets	3.7	Assistant Data Scientist (Graduate) | R-2375	London, England	5d	31.00 Per Hour (Employer est.)
2	TW	4.0	Data Scientist	Nottingham, England	30d+	50000 - 65000 (Employer est.)
3	NatWest Group	4.6	Data Scientist	Edinburgh, Scotland	2d	41000 - 54000 (Glassdoor est.)
4	iwoca	3.9	Data Scientist - Ops	London, England	7d	60000 - 90000 (Employer est.)
..	…	…	…	…	…	…
745	Deloitte	NaN	Manager, AI Architect, Strategy, Governance & …	London, England	30d+	NaN
746	Amazon Development Centre (London) Limited	NaN	Sr. Applied Scientist AGI, Contextual Ads	London, England	30d+	NaN
747	sennder	NaN	(Senior) Artificial Intelligence Engineer	London, England	30d+	NaN
748	Mott MacDonald	NaN	Senior Software Engineer / Machine Learning En…	London, England	30d+	NaN
749	Illumina	NaN	Senior Deep Learning / AI Engineer	Cambridge, East of England, England	23d	NaN

```
746                              NaN
747                              NaN
748                              NaN
749                              NaN

                                                  Skills
0    Data mining, Big data, R, Data analysis skills…
1                            R, SQL, JavaScript, Python
2                     SQL, Maths, Data science, Python
3    Software deployment, Data analysis skills, Sta…
4                 MATLAB, R, Maths, C, Machine learning
..                                                  …
745                                               NaN
746                                               NaN
747                                               NaN
748                                               NaN
749                                               NaN

[750 rows x 7 columns]
```

[9]:
```
df.head(5)
```

[9]:
```
          Company  Company Score                                    Job Title  \
0     Razorpoint             3.4                          Junior Data Scientist
1  tower Hamlets             3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW             4.0                                Data Scientist
3  NatWest Group             4.6                                Data Scientist
4          iwoca             3.9                         Data Scientist - Ops

              Location  Date                          Salary  \
0  Manchester, England    3d           35000 (Employer est.)
1      London, England    5d  31.00 Per Hour (Employer est.)
2  Nottingham, England  30d+   50000 - 65000 (Employer est.)
3  Edinburgh, Scotland    2d  41000 - 54000 (Glassdoor est.)
4      London, England    7d   60000 - 90000 (Employer est.)

                                          Skills
0  Data mining, Big data, R, Data analysis skills…
1                        R, SQL, JavaScript, Python
2                 SQL, Maths, Data science, Python
3  Software deployment, Data analysis skills, Sta…
4              MATLAB, R, Maths, C, Machine learning
```

[10]:
```
# Replacing (Employer est.) to '' AND (Glassdoor est) to ''.

#df['Salary']=df['Salary'].str.replace(r'\(Employer est.\)',"",regex=True)
#df['Salary']=df['Salary'].str.replace(r'\(Glasdoor est.\)',"",regex=True)
```

```
# OR.........One line code........
df['Salary']=df['Salary'].str.replace(r'\(Employer est.\)',"",regex=True).str.
 ↪replace(r'\(Glassdoor est.\)',"",regex=True)
```

[11]:
```
df.head(5)
```

[11]:
```
          Company  Company Score                                      Job Title  \
0     Razorpoint             3.4                           Junior Data Scientist
1  tower Hamlets             3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW             4.0                                 Data Scientist
3  NatWest Group             4.6                                 Data Scientist
4          iwoca             3.9                          Data Scientist - Ops

              Location  Date           Salary  \
0  Manchester, England    3d           35000
1      London, England    5d  31.00 Per Hour
2  Nottingham, England  30d+   50000 - 65000
3  Edinburgh, Scotland    2d   41000 - 54000
4      London, England    7d   60000 - 90000

                                          Skills
0  Data mining, Big data, R, Data analysis skills…
1                         R, SQL, JavaScript, Python
2                  SQL, Maths, Data science, Python
3  Software deployment, Data analysis skills, Sta…
4              MATLAB, R, Maths, C, Machine learning
```

[12]:
```
# import regex module....
import re
```

[13]:
```
# write a function to calculate per hour salary to yearly salary.

def convert_to_yearly(salary):
  if "Per Hour" in str(salary):
    hourly_rate=re.findall(r'\d+\.?\d*',salary)
    if hourly_rate:
      return float(hourly_rate[0])*2080 # converting to yearly salary
  else:
    return salary

# applying the function to dataframe df
df["Salary"]=df["Salary"].apply(convert_to_yearly)
```

[14]:
```
df.head()
```

[14]:
```
          Company  Company Score                                  Job Title  \
0     Razorpoint            3.4                        Junior Data Scientist
1  tower Hamlets            3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW            4.0                                Data Scientist
3  NatWest Group            4.6                                Data Scientist
4          iwoca            3.9                         Data Scientist - Ops

              Location  Date          Salary  \
0  Manchester, England    3d          35000
1      London, England    5d         64480.0
2  Nottingham, England  30d+  50000 - 65000
3  Edinburgh, Scotland    2d  41000 - 54000
4      London, England    7d  60000 - 90000

                                          Skills
0  Data mining, Big data, R, Data analysis skills…
1                         R, SQL, JavaScript, Python
2                   SQL, Maths, Data science, Python
3  Software deployment, Data analysis skills, Sta…
4              MATLAB, R, Maths, C, Machine learning
```

[15]:
```
# converting salary column dtype to str

df["Salary"]=df["Salary"].astype(str)
```

[16]:
```
#creating 2 new columns as salary_min & salary_max

df[["salary_min","salary_max"]]=df["Salary"].str.split("-",expand=True)
```

[17]:
```
df.head()
```

[17]:
```
          Company  Company Score                                  Job Title  \
0     Razorpoint            3.4                        Junior Data Scientist
1  tower Hamlets            3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW            4.0                                Data Scientist
3  NatWest Group            4.6                                Data Scientist
4          iwoca            3.9                         Data Scientist - Ops

              Location  Date          Salary  \
0  Manchester, England    3d          35000
1      London, England    5d         64480.0
2  Nottingham, England  30d+  50000 - 65000
3  Edinburgh, Scotland    2d  41000 - 54000
4      London, England    7d  60000 - 90000

                                          Skills salary_min salary_max
0  Data mining, Big data, R, Data analysis skills…     35000        None
```

```
1                         R, SQL, JavaScript, Python    64480.0       None
2                   SQL, Maths, Data science, Python     50000      65000
3  Software deployment, Data analysis skills, Sta…     41000      54000
4              MATLAB, R, Maths, C, Machine learning     60000      90000
```

[18]:
```
# Removing the white spaces in salary_min & salry _max columns (white spaces␣
 ↪before & after -)

df["salary_min"]=df["salary_min"].str.strip()
df["salary_max"]=df["salary_max"].str.strip()
```

[19]:
```
# Removing the None values from salary_min & salry_max columns

df["salary_min"]=pd.to_numeric(df["salary_min"],errors="coerce")
df["salary_max"]=pd.to_numeric(df["salary_max"],errors="coerce")
```

[20]:
```
# Ensuring location column dtype is str type

df["Location"]=df["Location"].astype(str)
```

[21]:
```
# Split location column into 'City', 'Country' based on , comma delimiter

df[["City","Country"]]=df["Location"].str.split(",",n=1,expand=True)
```

[22]:
```
df.head(5)
```

[22]:
```
          Company  Company Score                                  Job Title  \
0     Razorpoint            3.4                        Junior Data Scientist
1  tower Hamlets            3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW            4.0                              Data Scientist
3  NatWest Group            4.6                              Data Scientist
4          iwoca            3.9                   Data Scientist - Ops

              Location  Date          Salary  \
0  Manchester, England    3d          35000
1      London, England    5d         64480.0
2  Nottingham, England  30d+  50000 - 65000
3  Edinburgh, Scotland    2d  41000 - 54000
4      London, England    7d  60000 - 90000

                                          Skills  salary_min  salary_max  \
0  Data mining, Big data, R, Data analysis skills…     35000.0         NaN
1                         R, SQL, JavaScript, Python     64480.0         NaN
2                   SQL, Maths, Data science, Python     50000.0     65000.0
3  Software deployment, Data analysis skills, Sta…     41000.0     54000.0
4              MATLAB, R, Maths, C, Machine learning     60000.0     90000.0
```

```
          City    Country
0  Manchester    England
1      London    England
2  Nottingham    England
3   Edinburgh   Scotland
4      London    England
```

[23]:
```
# Using Onehot encoding (i.e 0 and 1) if skill is Yes then 1, No 0 will be encoded to all skills......AND split column skills by , comma
skills_split=df["Skills"].str.get_dummies(sep=",")
```

[24]:
```
skills_split.head()
```

[24]:
```
     AI   APIs   ATS   AWS   Account management   Accounting   Adobe Flash  \
0    0      0     0     0                     0            0             0
1    0      0     0     0                     0            0             0
2    0      0     0     0                     0            0             0
3    0      0     0     0                     0            0             0
4    0      0     0     0                     0            0             0

    Agile   Alteryx   Analysis skills  …  Supervising experience  \
0       0         0                 0  …                       0
1       0         0                 0  …                       0
2       0         0                 0  …                       0
3       0         0                 0  …                       0
4       0         0                 0  …                       0

   Supply chain  System design  Tableau  Teaching  TensorFlow  \
0             0              0        0         0           0
1             0              0        0         0           0
2             0              0        0         0           0
3             0              0        0         0           0
4             0              0        0         0           0

   Test-driven development  Ukrainian  Underwriting  XML
0                        0          0             0    0
1                        0          0             0    0
2                        0          0             0    0
3                        0          0             0    0
4                        0          0             0    0

[5 rows x 315 columns]
```

[25]:
```
# Combining the Two data Frames i.e, new data frame name df_with_skills ==> old dataframes=(df+skills_split)

df_with_skills=pd.concat([df,skills_split],axis=1)
```

[26]:
```
# new dataframe df_with_skills

df_with_skills.head()
```

[26]:
```
          Company  Company Score                                    Job Title  \
0     Razorpoint            3.4                          Junior Data Scientist
1  tower Hamlets            3.7  Assistant Data Scientist (Graduate) | R-2375
2             TW            4.0                                Data Scientist
3  NatWest Group            4.6                                Data Scientist
4          iwoca            3.9                         Data Scientist - Ops

              Location  Date          Salary  \
0  Manchester, England    3d          35000
1      London, England    5d         64480.0
2  Nottingham, England  30d+  50000 - 65000
3  Edinburgh, Scotland    2d  41000 - 54000
4      London, England    7d  60000 - 90000

                                          Skills  salary_min  salary_max  \
0  Data mining, Big data, R, Data analysis skills…     35000.0         NaN
1                        R, SQL, JavaScript, Python     64480.0         NaN
2                  SQL, Maths, Data science, Python     50000.0     65000.0
3  Software deployment, Data analysis skills, Sta…     41000.0     54000.0
4              MATLAB, R, Maths, C, Machine learning     60000.0     90000.0

         City  … Supervising experience  Supply chain  System design  \
0  Manchester  …                       0             0              0
1      London  …                       0             0              0
2  Nottingham  …                       0             0              0
3   Edinburgh  …                       0             0              0
4      London  …                       0             0              0

   Tableau  Teaching  TensorFlow  Test-driven development  Ukrainian  \
0        0         0           0                        0          0
1        0         0           0                        0          0
2        0         0           0                        0          0
3        0         0           0                        0          0
4        0         0           0                        0          0

   Underwriting  XML
0             0    0
1             0    0
2             0    0
3             0    0
4             0    0

[5 rows x 326 columns]
```

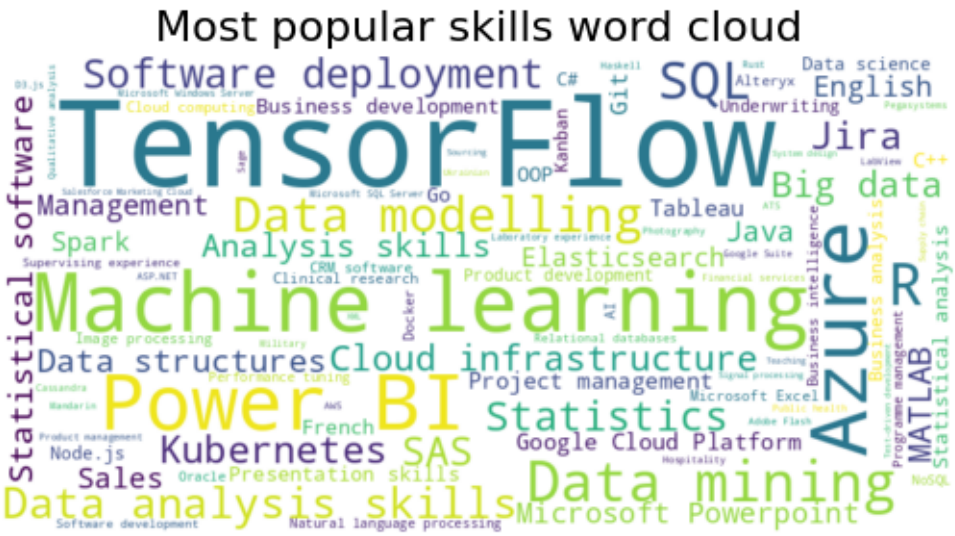
[27]:
```
# Creating new variable skill_frequencies to sum the data of all skills (i.e 1
 ↪and 0)
# iloc helps us to locate the column
# : represetns find the all the rows
# I want all the columns starting from column 'AI' upto the End so using : in
 ↪end
# using sum() to sum the all values

skill_frequencies=df_with_skills.iloc[:,df_with_skills.columns.get_loc('AI'):].
 ↪sum()
```

[28]:
```
skill_frequencies
```

[28]:
```
AI                          2
ASP.NET                     1
ATS                         1
AWS                         1
Adobe Flash                 1
                          ...
TensorFlow                165
Test-driven development     1
Ukrainian                   1
Underwriting                3
XML                         1
Length: 94, dtype: int64
```

[29]:
```
# Import world cloud module

from wordcloud import WordCloud
```

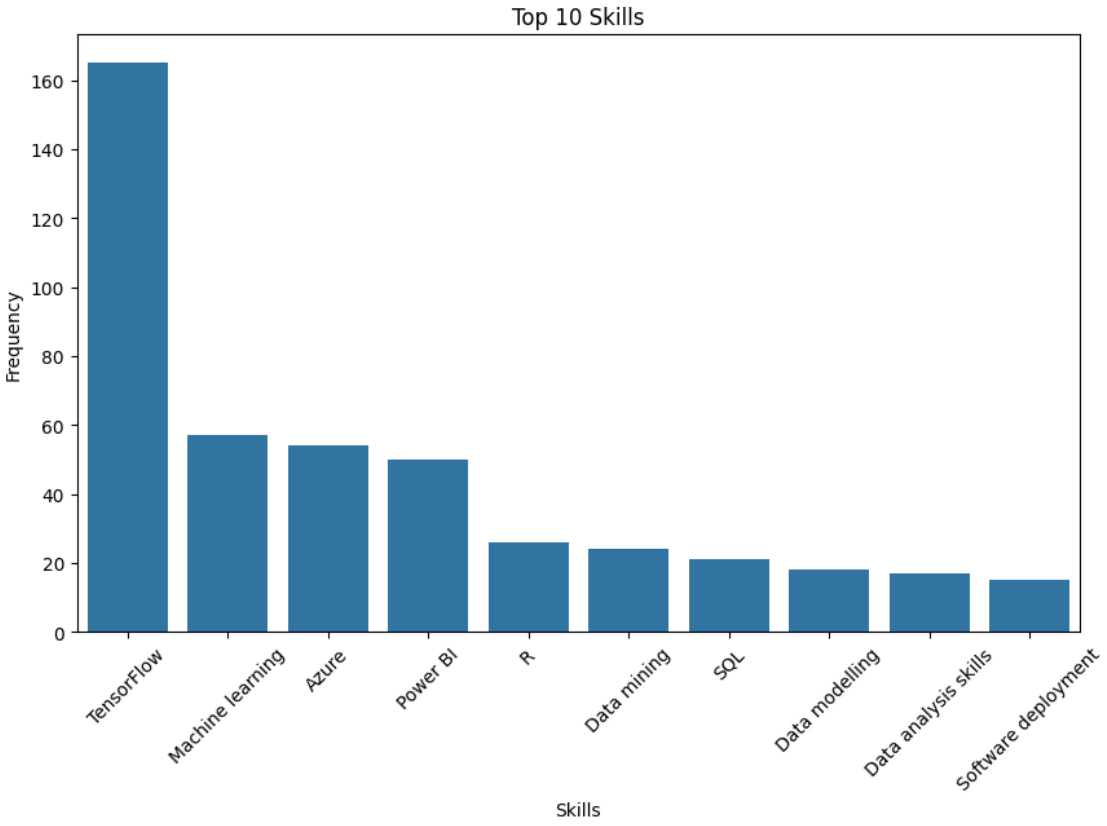
[30]:
```
#create wordcloud

# with width,height & bacground white.

#.generate_from_frequencies function here pass our variable skill_frequencies.

wordcloud=WordCloud(width=800,height=400,background_color='white').
 ↪generate_from_frequencies(skill_frequencies)

plt.imshow(wordcloud,interpolation='bilinear')
plt.axis('off')
plt.title('Most popular skills word cloud',fontsize=16)
plt.show()
```



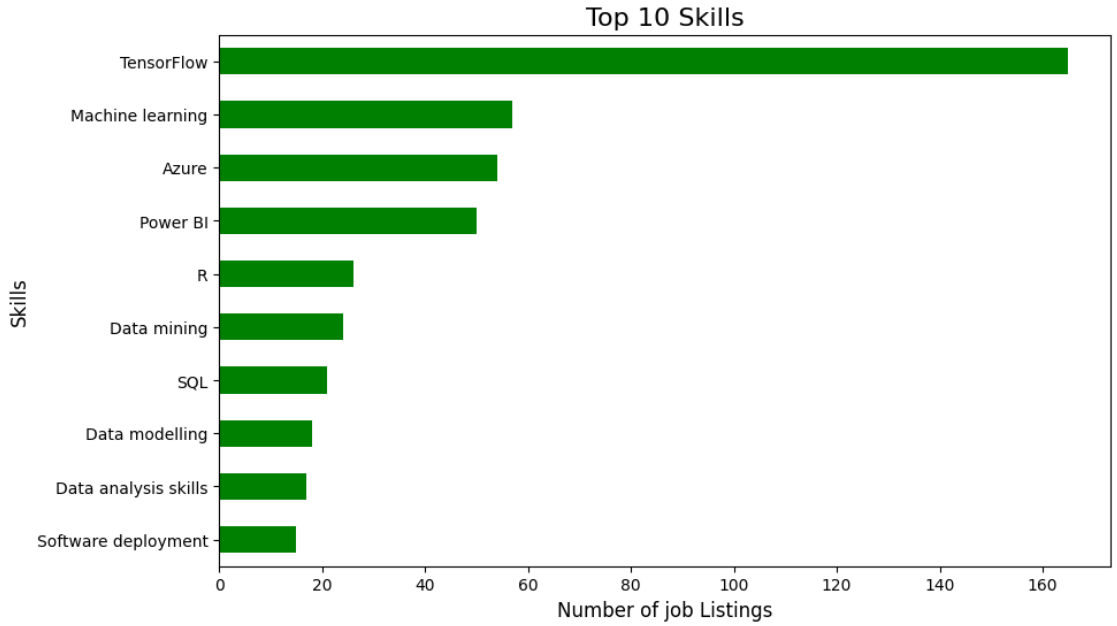
[31]:
```
# Creating a new variable skills_column, then using ilco : to select all skills
 ↪till the end,
# and sorting the data in decending order i.e Top skills will come first.

skills_column=df_with_skills.iloc[:,df_with_skills.columns.get_loc('AI'):].
 ↪sum().sort_values(ascending=False)
```

[32]:
```
# creating new variable and Selecting Top 10 skills

top_10_skills=skills_column.head(10)
```

[33]:
```
# Plotting the data in Vertical bar chart

plt.figure(figsize=(10,6))
sns.barplot(x=top_10_skills.index,y=top_10_skills.values)
plt.xlabel('Skills')
plt.ylabel('Frequency')
plt.title('Top 10 Skills')
plt.xticks(rotation=45)
plt.show()
```



[34]:
```
# Plotting the data in bar chart Horizontally using (barh)

plt.figure(figsize=(10,6))
top_10_skills.plot(kind='barh',color='green') #kind=barh means bar chart h=␣
 ↪horizontally.
plt.xlabel('Number of job Listings',fontsize=12)
plt.ylabel('Skills',fontsize=12)
plt.title('Top 10 Skills',fontsize=16)
plt.gca().invert_yaxis()
plt.show()
```



[35]:
```
# filtering data for Data Analyst  - DA Roles
# creating new datframe as data_analyst_jobs

data_analyst_jobs = df_with_skills[df_with_skills['Job Title'].str.
 ↪contains('Data Analyst',case=False,na=False)]

# creating new dataframe as data_analyst_skills
# using iloc,: to select all data starting from AI till End...

data_analyst_skills = data_analyst_jobs.iloc[:,data_analyst_jobs.columns.
 ↪get_loc('AI'):].sum().sort_values(ascending=False)
```

[36]:
```
data_analyst_jobs
```

[36]:

	Company	Company Score	\
35	Hadrian	3.7	
43	Windranger Labs	4.8	
48	Animoca Brands Limited	3.6	
53	Razorpoint	4.4	
67	Creditsafe	3.2	
129	Bannatyne	3.9	
132	Global 4 Communications Limited	3.8	
148	Diamond Light Source	3.8	
159	TW	3.3	
181	InPost UK	3.8	
189	FM Conway	3.8	

```
232                          Geomiq            4.6

                                          Job Title  \
35               Commercial Data Analyst - Internship
43                                       Data Analyst
48                  Tokenomics Data Analyst/Scientist
53                               Junior Data Analyst
67                                       Data Analyst
129                                      Data Analyst
132                                      Data Analyst
148  Data Analyst Scientist - Diamond Light Source …
159                                 Lead Data Analyst
181                                  Big Data Analyst
189                     Health and Safety Data Analyst
232                        Data Analyst - Supply Chain

                                    Location  Date           Salary  \
35                           London, England    2d   44000 - 67000
43                                    Remote   10d   60000 - 80000
48                                    Remote  30d+   35000 - 60000
53                           London, England   20d     FCFA 102000
67                            Cardiff, Wales    4d           67000
129  Darlington, North East England, England   11d   39000 - 61000
132                          Horsham, England  30d+   80000 - 83000
148                           Didcot, England    8d           55000
159                       Birmingham, England  30d+  74000 - 120000
181                           London, England    9d   51000 - 90000
189                         Sevenoaks, England  30d+   64000 - 85000
232                           London, England  30d+   40000 - 47000

                                                Skills  salary_min  \
35   Relational databases, R, SQL, Statistical anal…     44000.0
43               R, Git, SQL, Machine learning, Python     60000.0
48                     Machine learning, Data science     35000.0
53   TensorFlow, Azure, R, Google Cloud Platform, T…         NaN
67           Spark, R, Data analysis skills, NoSQL, Git     67000.0
129           Power BI, Azure, Big data, Spark, Tableau     39000.0
132  R, SQL, Statistical analysis, Maths, Machine l…     80000.0
148  Azure, R, Data analysis skills, Git, Google Cl…     55000.0
159  Azure, Management, R, Microsoft SQL Server, NoSQL     74000.0
181              Machine learning, Data science, Python     51000.0
189    TensorFlow, Spark, SQL, Maths, Machine learning     64000.0
232  Statistics, Business processes, Machine learni…     40000.0

     salary_max        City  … Supervising experience  Supply chain  \
35      67000.0      London  …                       0             0
43      80000.0      Remote  …                       0             0
```

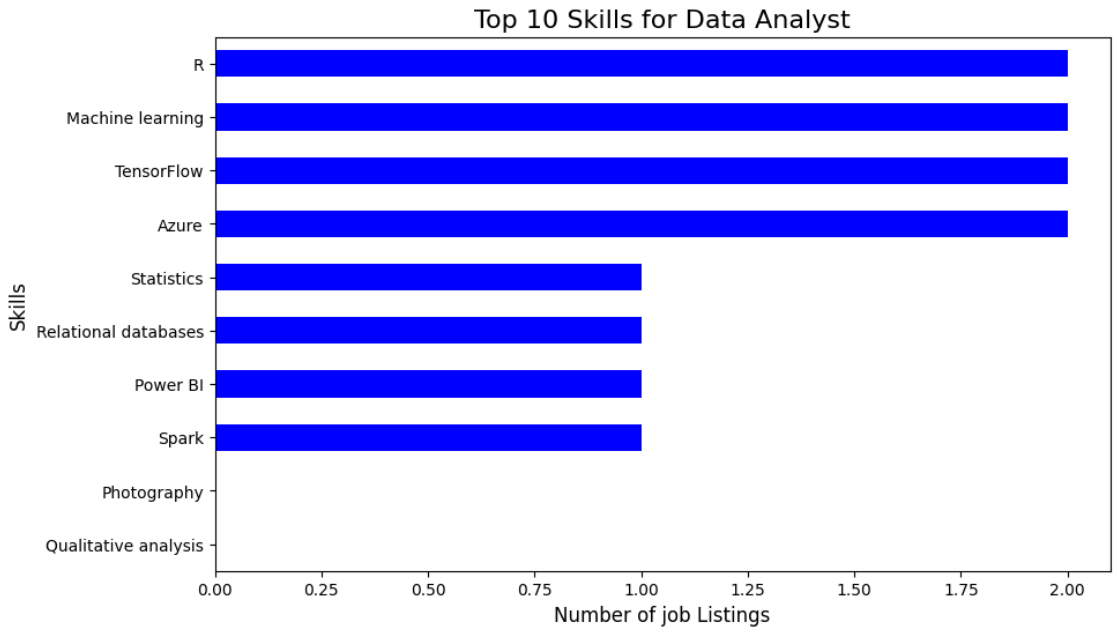
```
48      60000.0      Remote  …                    0             0
53          NaN      London  …                    0             0
67          NaN     Cardiff  …                    0             0
129     61000.0  Darlington  …                    0             0
132     83000.0     Horsham  …                    0             0
148         NaN      Didcot  …                    0             0
159    120000.0  Birmingham  …                    0             0
181     90000.0      London  …                    0             0
189     85000.0   Sevenoaks  …                    0             0
232     47000.0      London  …                    0             0

     System design  Tableau  Teaching  TensorFlow  Test-driven development  \
35               0        0         0           0                        0
43               0        0         0           0                        0
48               0        0         0           0                        0
53               0        0         0           1                        0
67               0        0         0           0                        0
129              0        0         0           0                        0
132              0        0         0           0                        0
148              0        0         0           0                        0
159              0        0         0           0                        0
181              0        0         0           0                        0
189              0        0         0           1                        0
232              0        0         0           0                        0

     Ukrainian  Underwriting  XML
35           0             0    0
43           0             0    0
48           0             0    0
53           0             0    0
67           0             0    0
129          0             0    0
132          0             0    0
148          0             0    0
159          0             0    0
181          0             0    0
189          0             0    0
232          0             0    0

[12 rows x 326 columns]
```

[37]:
```
top_10_data_analyst_skills = data_analyst_skills.head(10)
top_10_data_analyst_skills
```

[37]:
```
R                     2
Machine learning      2
TensorFlow            2
```

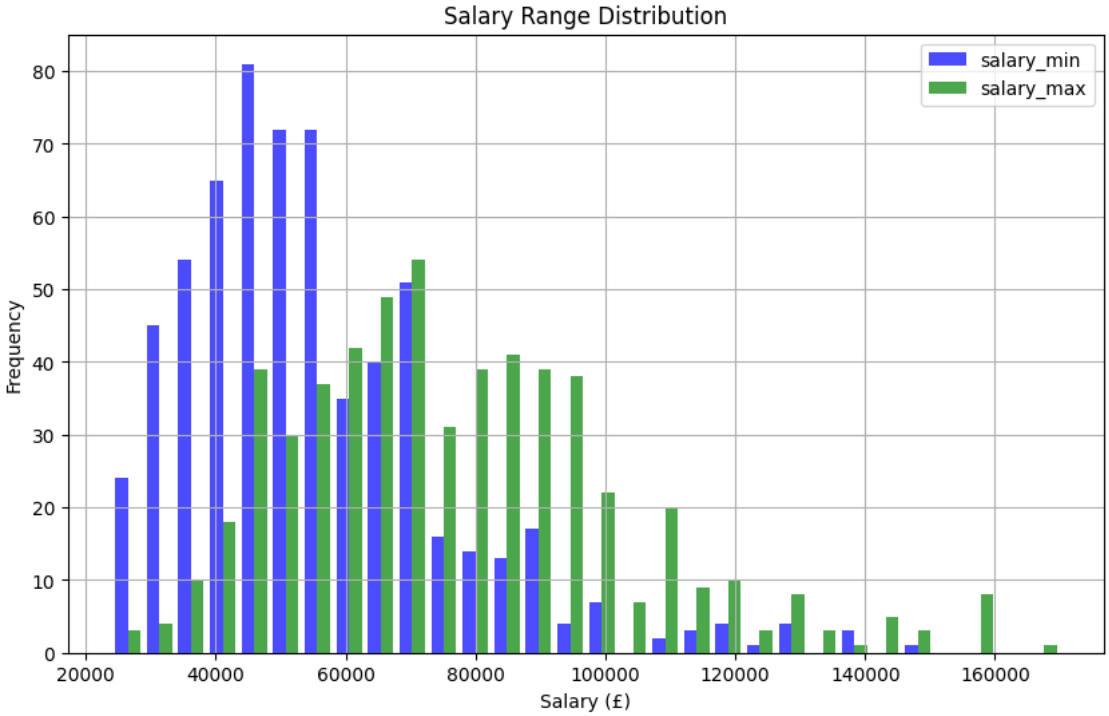
```
Azure                   2
Statistics              1
Relational databases    1
Power BI                1
Spark                   1
Photography             0
Qualitative analysis    0
dtype: int64
```

[38]:
```
plt.figure(figsize=(10,6))
top_10_data_analyst_skills.plot(kind='barh',color='blue') #kind=barh gives the␣
 ↪top skills on top
plt.xlabel('Number of job Listings',fontsize=12)
plt.ylabel('Skills',fontsize=12)
plt.title('Top 10 Skills for Data Analyst',fontsize=16)
plt.gca().invert_yaxis()
plt.show()
```



[39]:
```
# Creating Salary distribution
# creating bins=30,(more bins give more data granuality)
#label bins as salary_min,salary_max,
#color blue,green
#alpha 0.7 is transparency


plt.figure(figsize=(10,6))
```

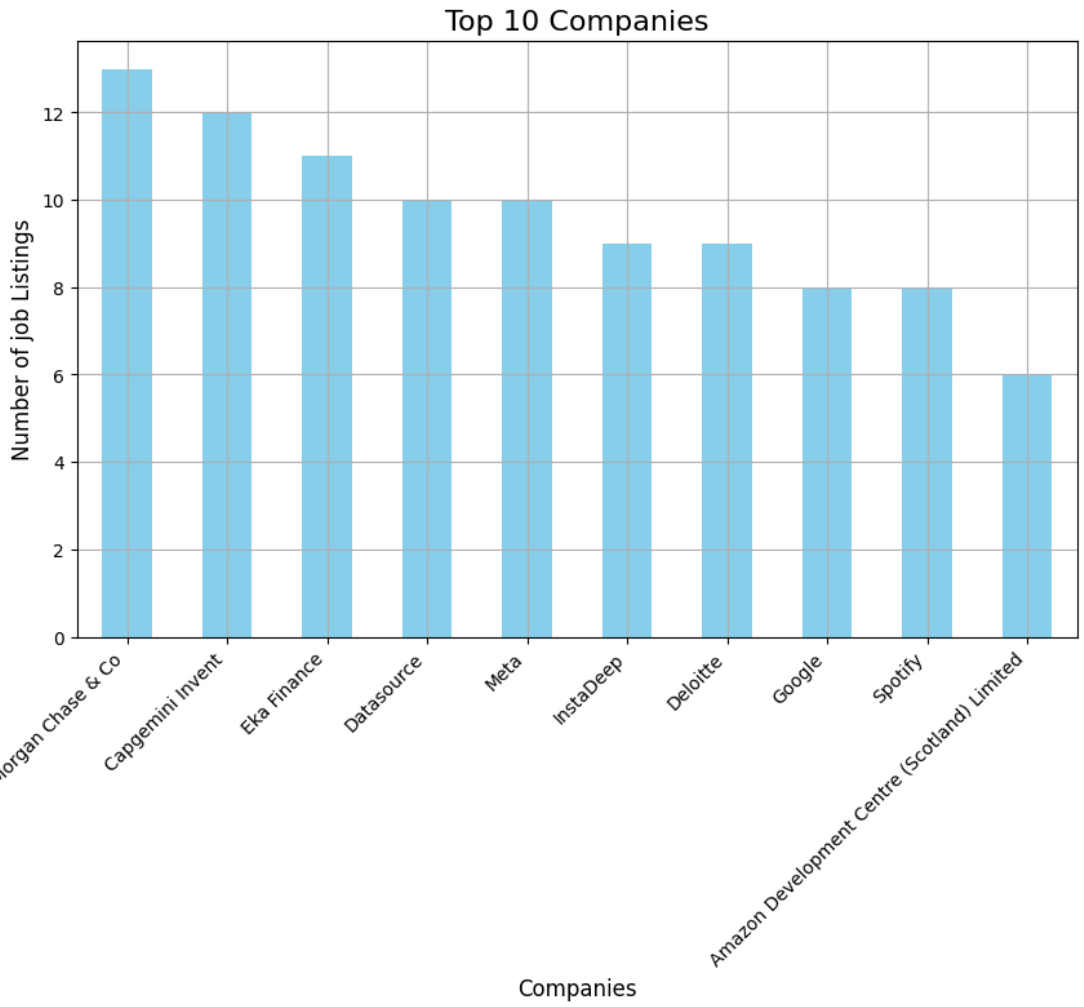
```
plt.hist([df["salary_min"],df['salary_max']],␣
 ↪bins=30,label=['salary_min','salary_max'], color=['blue','green'],alpha=0.7)
plt.xlabel('Salary (£)')
plt.ylabel('Frequency')
plt.title('Salary Range Distribution')
plt.legend()
plt.grid(True)
plt.show()
```



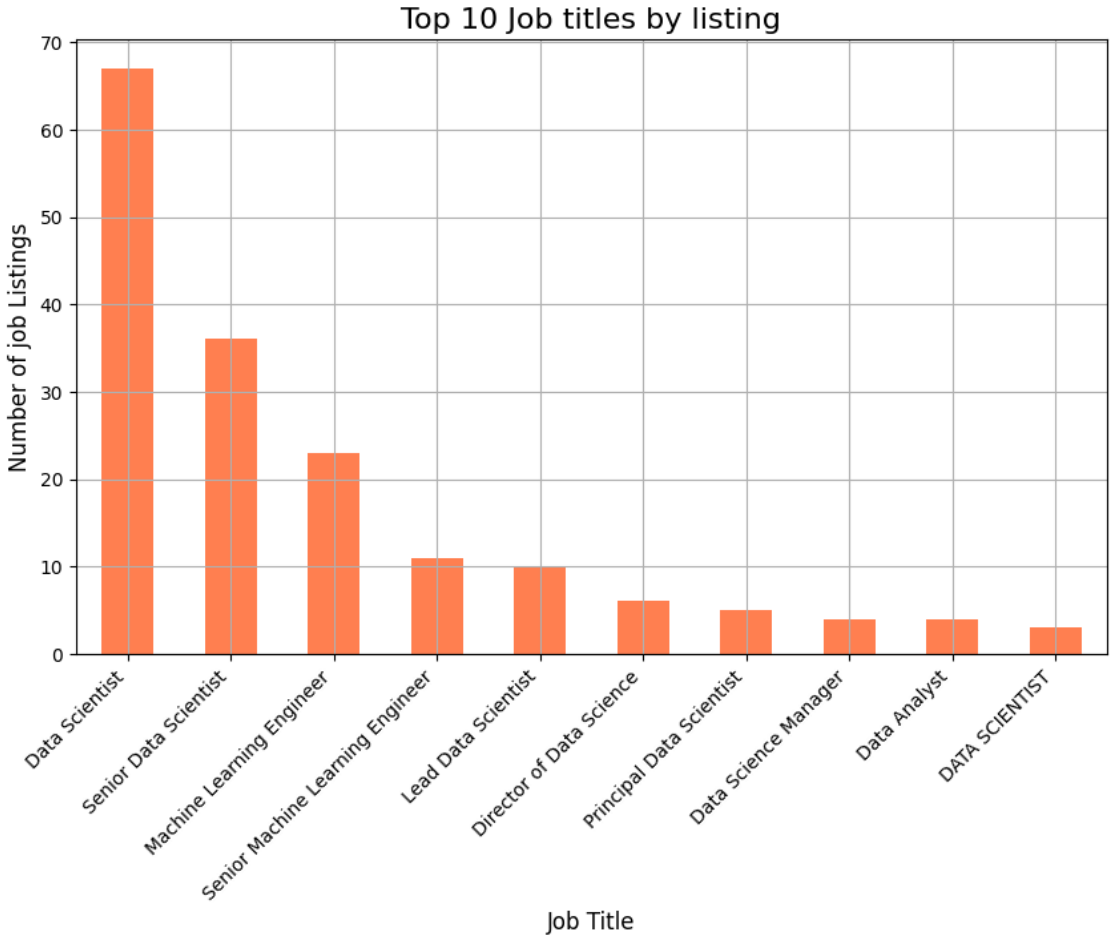
[40]:
```
# creating new varible as top_companies

top_companies = df_with_skills['Company'].value_counts().head(10)

# creating new  variable as top_job_titles

top_job_titles = df_with_skills['Job Title'].value_counts().head(10)
```

[41]:
```
# plot for top_companies

plt.figure(figsize=(10,6))
top_companies.plot(kind='bar',color='skyblue')
plt.ylabel('Number of job Listings',fontsize=12)
plt.xlabel('Companies',fontsize=12)
```

```
plt.title('Top 10 Companies',fontsize=16)
plt.grid(True)
plt.xticks(rotation=45, ha='right')
plt.show()
```



[42]:
```
# plot for top_job_titles

plt.figure(figsize=(10,6))
top_job_titles.plot(kind='bar',color='coral') #kind=barh gives the top skills␣
↪on top
plt.ylabel('Number of job Listings',fontsize=12)
plt.xlabel('Job Title',fontsize=12)
plt.title('Top 10 Job titles by listing',fontsize=16)
plt.grid(True)
plt.xticks(rotation=45, ha='right')
plt.show()
```



[43]:
```
# Replacing 'DATA SCIENTIST' titles to 'Data Scientist'  &  Assigning back to original column "Job Title" itself

df_with_skills['Job Title']= df_with_skills['Job Title'].str.replace('DATA SCIENTIST','Data Scientist', case=False)
```

[43]:
```
```