Employee Payroll Data

June 11, 2023

[1]:
```
import os
import glob
import numpy as np
import pandas as pd
import shutil

pd.set_option('display.max_columns',None)
```

[2]:
```
#To read the csv file
df = pd.read_csv(r'ds_salaries.csv')
```

[3]:
```
#To display the first five rows
df.head()
```

[3]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L

[4]:
```
#To display the number of rows and columns in the dataset
df.shape
```

[4]: (607, 12)

[5]:
```
#To show the datatypes that are available in the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          607 non-null    int64 
 1   work_year           607 non-null    int64 
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64 
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64 
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64 
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
```

[6]:
```
#To show information contained in each column of the dataset
df.describe()
```

[6]:

	Unnamed: 0	work_year	salary	salary_in_usd	remote_ratio
count	607.000000	607.000000	6.070000e+02	607.000000	607.00000
mean	303.000000	2021.405272	3.240001e+05	112297.869852	70.92257
std	175.370085	0.692133	1.544357e+06	70957.259411	40.70913
min	0.000000	2020.000000	4.000000e+03	2859.000000	0.00000
25%	151.500000	2021.000000	7.000000e+04	62726.000000	50.00000
50%	303.000000	2022.000000	1.150000e+05	101570.000000	100.00000
75%	454.500000	2022.000000	1.650000e+05	150000.000000	100.00000
max	606.000000	2022.000000	3.040000e+07	600000.000000	100.00000

[ ]:

[7]:
```
#To filter data for a specific year which in this case is 2020
df[df['work_year'] == 2020]
```

[7]:

	Unnamed: 0	work_year	experience_level	employment_type	\
0	0	2020	MI	FT	
1	1	2020	SE	FT	
2	2	2020	SE	FT	
3	3	2020	MI	FT	

```
4            4       2020               SE              FT
..         …       …                …               …
67          67       2020               SE              FT
68          68       2020               EN              FT
69          69       2020               SE              FT
70          70       2020               MI              FT
71          71       2020               MI              FT

                    job_title  salary salary_currency  salary_in_usd  \
0              Data Scientist   70000             EUR          79833
1   Machine Learning Scientist  260000             USD         260000
2            Big Data Engineer   85000             GBP         109024
3         Product Data Analyst   20000             USD          20000
4    Machine Learning Engineer  150000             USD         150000
..                         …     …               …              …
67        Data Science Manager  190200             USD         190200
68             Data Scientist  105000             USD         105000
69             Data Scientist   80000             EUR          91237
70             Data Scientist   55000             EUR          62726
71             Data Scientist   37000             EUR          42197

   employee_residence  remote_ratio company_location company_size
0                  DE             0               DE            L
1                  JP             0               JP            S
2                  GB            50               GB            M
3                  HN             0               HN            S
4                  US            50               US            L
..                 …           …                 …           …
67                 US           100               US            M
68                 US           100               US            S
69                 AT             0               AT            S
70                 FR            50               LU            S
71                 FR            50               FR            S

[72 rows x 12 columns]
```

[8]:
```
#To filter data that shows the MI experience level and employment type is FT
df[(df['experience_level'] == 'MI') & (df['employment_type'] == 'FT')]
```

[8]:
```
     Unnamed: 0  work_year experience_level employment_type  \
0             0       2020               MI              FT
3             3       2020               MI              FT
7             7       2020               MI              FT
8             8       2020               MI              FT
11           11       2020               MI              FT
..          …       …                …               …
567         567       2022               MI              FT
```

```
586         586       2022              MI              FT
598         598       2022              MI              FT
599         599       2022              MI              FT
606         606       2022              MI              FT

                job_title    salary salary_currency  salary_in_usd  \
0          Data Scientist     70000             EUR          79833
3     Product Data Analyst     20000             USD          20000
7          Data Scientist  11000000             HUF          35735
8    Business Data Analyst    135000             USD         135000
11         Data Scientist   3000000             INR          40481
..                    …       …              …             …
567          Data Analyst     50000             GBP          65438
586          Data Analyst     35000             GBP          45807
598        Data Scientist    160000             USD         160000
599        Data Scientist    130000             USD         130000
606          AI Scientist    200000             USD         200000

    employee_residence  remote_ratio company_location company_size
0                   DE             0                DE            L
3                   HN             0                HN            S
7                   HU            50                HU            L
8                   US           100                US            L
11                  IN             0                IN            L
..                 …           …              …          …
567                 GB             0                GB            M
586                 GB             0                GB            M
598                 US           100                US            M
599                 US           100                US            M
606                 IN           100                US            L

[206 rows x 12 columns]
```

[9]:
```
#To filter data that shows the MI experience level or employment type is FT
df[(df['experience_level'] == 'MI') | (df['employment_type'] == 'FT')]
```

[9]:
```
     Unnamed: 0  work_year experience_level employment_type  \
0             0       2020              MI              FT
1             1       2020              SE              FT
2             2       2020              SE              FT
3             3       2020              MI              FT
4             4       2020              SE              FT
..          …       …              …             …
602         602       2022              SE              FT
603         603       2022              SE              FT
604         604       2022              SE              FT
605         605       2022              SE              FT
```

```
606         606       2022                MI              FT

                     job_title  salary salary_currency  salary_in_usd  \
0                Data Scientist   70000             EUR          79833
1    Machine Learning Scientist  260000             USD         260000
2              Big Data Engineer   85000             GBP         109024
3          Product Data Analyst   20000             USD          20000
4     Machine Learning Engineer  150000             USD         150000
..                          …     …              …             …
602              Data Engineer  154000             USD         154000
603              Data Engineer  126000             USD         126000
604               Data Analyst  129000             USD         129000
605               Data Analyst  150000             USD         150000
606               AI Scientist  200000             USD         200000

    employee_residence  remote_ratio company_location company_size
0                   DE             0               DE            L
1                   JP             0               JP            S
2                   GB            50               GB            M
3                   HN             0               HN            S
4                   US            50               US            L
..                 …           …                …           …
602                 US           100               US            M
603                 US           100               US            M
604                 US             0               US            M
605                 US           100               US            M
606                 IN           100               US            L

[595 rows x 12 columns]
```

[10]:
```
#To show specific colums(work_year, experience_level,job_title, employment_type␣
 ↪and salary)
df.loc[:5:,['work_year', 'experience_level', 'job_title', 'employment_type',␣
 ↪'salary']]
df.iloc[:5,[1,2,3,4,5]]
```

[10]:
```
   work_year experience_level employment_type                  job_title  \
0       2020               MI              FT              Data Scientist
1       2020               SE              FT  Machine Learning Scientist
2       2020               SE              FT           Big Data Engineer
3       2020               MI              FT        Product Data Analyst
4       2020               SE              FT   Machine Learning Engineer

   salary
0   70000
1  260000
2   85000
```

```
3    20000
4   150000
```

[11]:
```
#To know how many work years are there in the dataset
df.work_year.unique()
```

[11]:
```
array([2020, 2021, 2022], dtype=int64)
```

[12]:
```
#To know the salary currencies that are in the dataset
df.value_counts('salary_currency')
```

[12]:
```
salary_currency
USD    398
EUR     95
GBP     44
INR     27
CAD     18
JPY      3
PLN      3
TRY      3
CNY      2
DKK      2
BRL      2
HUF      2
MXN      2
SGD      2
AUD      2
CHF      1
CLP      1
dtype: int64
```

[13]:
```
#To check duplicates
df[df.duplicated(['salary_currency'])]
```

[13]:
```
     Unnamed: 0  work_year experience_level employment_type  \
3             3       2020               MI              FT
4             4       2020               SE              FT
5             5       2020               EN              FT
6             6       2020               SE              FT
8             8       2020               MI              FT
..          ...        ...              ...             ...
602         602       2022               SE              FT
603         603       2022               SE              FT
604         604       2022               SE              FT
605         605       2022               SE              FT
606         606       2022               MI              FT
```

```
                    job_title  salary salary_currency  salary_in_usd  \
3         Product Data Analyst   20000             USD          20000
4    Machine Learning Engineer  150000             USD         150000
5                 Data Analyst   72000             USD          72000
6          Lead Data Scientist  190000             USD         190000
8        Business Data Analyst  135000             USD         135000
..                         …     …               …             …
602              Data Engineer  154000             USD         154000
603              Data Engineer  126000             USD         126000
604               Data Analyst  129000             USD         129000
605               Data Analyst  150000             USD         150000
606               AI Scientist  200000             USD         200000

    employee_residence  remote_ratio company_location company_size
3                   HN             0                HN            S
4                   US            50                US            L
5                   US           100                US            L
6                   US           100                US            S
8                   US           100                US            L
..                 …             …                …             …
602                 US           100                US            M
603                 US           100                US            M
604                 US             0                US            M
605                 US           100                US            M
606                 IN           100                US            L

[590 rows x 12 columns]
```

[14]:
```
#To groupby different categories
for i, x in df.groupby('work_year'):
     x.to_csv(r"C:\Users\nichola.ondiek\Python\New folder\{}.csv".format(i),␣
 ↪index=False)
```

[15]:
```
#Splitting the excel by a column called salary currency into multiple sheets
column_name = 'salary_currency'

unique_values = df[column_name].unique()
writer = pd.ExcelWriter(r'C:\Users\nichola.ondiek\Python\New folder\output.
 ↪xlsx')

for unique_value in unique_values:
    frame = df[df['salary_currency'] == unique_value]
    frame.to_excel(writer, sheet_name=unique_value)
writer.save()
```

[16]:
```
#Checks if there are any blank values in any column and displays as a␣
 ↪percentage of the total that is rounded off to 4 decimal places
```

```
df.isna().mean().round(4) * 100
```

[16]:
```
Unnamed: 0            0.0
work_year             0.0
experience_level      0.0
employment_type       0.0
job_title             0.0
salary                0.0
salary_currency       0.0
salary_in_usd         0.0
employee_residence    0.0
remote_ratio          0.0
company_location      0.0
company_size          0.0
dtype: float64
```

[17]:
```
#Creates folders based on csv file names
path = r'C:\Users\nichola.ondiek\Python\New folder'

for file_path in glob.glob(os.path.join(path, '*.csv')):
    new_dir = file_path.rsplit('.', 1)[0]
    try:
        os.mkdir(os.path.join(path, new_dir))
    except WindowsError:
        # Handle the case where the target dir already exist.
        pass
    shutil.move(file_path, os.path.join(new_dir, os.path.basename(file_path)))
```

[ ]:

[ ]:

[ ]: