In [83]:
```
import pandas as pd
import numpy as np
```

In [112…
```
df=pd.read_csv(r'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cl
df
```

Out[112...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | |
|---|---|---|---|---|---|---|---|
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | |
| **2** | 13 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **4** | 17 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **11547** | 25136 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11548** | 25137 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **11549** | 25138 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **11550** | 25141 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of LOWER quality than prop... | Employed full-time | S |
| **11551** | 25142 | I am a developer by profession | Yes | Less than once a month but more | OSS is, on average, of HIGHER | Employed full-time | |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| | | | | than once per ... | quality than pro... | |

11552 rows × 85 columns

In [113…
```
df2.columns
```

Out[113…
```
Index(['Respondent', 'MainBranch', 'Hobbyist', 'OpenSourcer', 'OpenSource',
       'Employment', 'Country', 'Student', 'EdLevel', 'UndergradMajor',
       'EduOther', 'OrgSize', 'DevType', 'YearsCode', 'Age1stCode',
       'YearsCodePro', 'CareerSat', 'JobSat', 'MgrIdiot', 'MgrMoney',
       'MgrWant', 'JobSeek', 'LastHireDate', 'LastInt', 'FizzBuzz',
       'JobFactors', 'ResumeUpdate', 'CurrencySymbol', 'CurrencyDesc',
       'CompTotal', 'CompFreq', 'ConvertedComp', 'WorkWeekHrs', 'WorkPlan',
       'WorkChallenge', 'WorkRemote', 'WorkLoc', 'ImpSyn', 'CodeRev',
       'CodeRevHrs', 'UnitTests', 'PurchaseHow', 'PurchaseWhat',
       'LanguageWorkedWith', 'LanguageDesireNextYear', 'DatabaseWorkedWith',
       'DatabaseDesireNextYear', 'PlatformWorkedWith',
       'PlatformDesireNextYear', 'WebFrameWorkedWith',
       'WebFrameDesireNextYear', 'MiscTechWorkedWith',
       'MiscTechDesireNextYear', 'DevEnviron', 'OpSys', 'Containers',
       'BlockchainOrg', 'BlockchainIs', 'BetterLife', 'ITperson', 'OffOn',
       'SocialMedia', 'Extraversion', 'ScreenName', 'SOVisit1st',
       'SOVisitFreq', 'SOVisitTo', 'SOFindAnswer', 'SOTimeSaved',
       'SOHowMuchTime', 'SOAccount', 'SOPartFreq', 'SOJobs', 'EntTeams',
       'SOComm', 'WelcomeChange', 'SONewContent', 'Age', 'Gender', 'Trans',
       'Sexuality', 'Ethnicity', 'Dependents', 'SurveyLength', 'SurveyEase'],
      dtype='object')
```

# Checking if there are any duplicate values

In [114…
```
df.duplicated().any()
```

Out[114…
```
True
```

# Getting number of unique rows

In [115…
```
df['Respondent'].nunique()
```

Out[115…
```
11398
```

# Dropping Duplicates

In [116…
```
df=df.drop_duplicates()
df
```

Out[116...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | |
|---|---|---|---|---|---|---|---|
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | |
| **2** | 13 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **4** | 17 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **11547** | 25136 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11548** | 25137 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **11549** | 25138 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **11550** | 25141 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of LOWER quality than prop... | Employed full-time | S |
| **11551** | 25142 | I am a developer by profession | Yes | Less than once a month but more | OSS is, on average, of HIGHER | Employed full-time | |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| | | | | than once per ... | quality than pro... | |

11398 rows × 85 columns

## Top 5

In [117…
```
df.head()
```

Out[117…

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | Count |
|---|---|---|---|---|---|---|---|
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | Unit Stat |
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | Ne Zeala |
| **2** | 13 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | Unit Stat |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | Unit Kingdc |
| **4** | 17 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time | Austra |

5 rows × 85 columns

# Selecting only the rows with NaN values using 'any()' method

In [118…
```
df[pd.isnull(df).any(axis=1)]
```

Out[118...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | |
|---|---|---|---|---|---|---|---|
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | |
| **2** | 13 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **4** | 17 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **11547** | 25136 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11548** | 25137 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **11549** | 25138 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **11550** | 25141 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of LOWER quality than prop... | Employed full-time | S |
| **11551** | 25142 | I am a developer by profession | Yes | Less than once a month but more | OSS is, on average, of HIGHER | Employed full-time | |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| | | | | than once per ... | quality than pro... | |

10390 rows × 85 columns

# How many missing values are there in the EdLevel column?

In [119…
```
len(df[pd.isnull(df.EdLevel)])
```

Out[119…
```
112
```

# Unique data in WorkLoc Column

In [120…
```
df['WorkLoc'].unique()
```

Out[120…
```
array(['Home', 'Office', 'Other place, such as a coworking space or cafe',
       nan], dtype=object)
```

# Removing null values and replacing them with the median ('Office')

In [121…
```
location = ['Office']
df[df['WorkLoc'].isin(location)]
```

Out[121...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | C |
|---|---|---|---|---|---|---|---|
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | |
| **5** | 19 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **6** | 20 | I am not primarily a developer, but I write co... | No | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | Li |
| **8** | 23 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **9** | 24 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **11544** | 25128 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11545** | 25133 | I am a developer by profession | No | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | E |
| **11546** | 25134 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | E |
| **11549** | 25138 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **11551** | 25142 | I am a developer by profession | Yes | Less than once a month but more | OSS is, on average, of HIGHER | Employed full-time | K |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | C |
|---|---|---|---|---|---|---|---|
| | | | | than once per ... | quality than pro... | | |

6806 rows × 85 columns

In [122…

```
df2=df.replace(np.NaN,'Office')
df2
```

Out[122...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | |
|---|---|---|---|---|---|---|---|
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **1** | 9 | I am a developer by profession | Yes | Once a month or more often | The quality of OSS and closed source software ... | Employed full-time | |
| **2** | 13 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **4** | 17 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **11547** | 25136 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11548** | 25137 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | |
| **11549** | 25138 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **11550** | 25141 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of LOWER quality than prop... | Employed full-time | S |
| **11551** | 25142 | I am a developer by profession | Yes | Less than once a month but more | OSS is, on average, of HIGHER | Employed full-time | |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| | | | | than once per ... | quality than pro... | |

11398 rows × 85 columns

In [123…
```
df2['WorkLoc'].unique()
```

Out[123…
```
array(['Home', 'Office', 'Other place, such as a coworking space or cafe'],
      dtype=object)
```

# Unique values in UndergradMajor

In [124…
```
df2['UndergradMajor'].unique()
```

Out[124…
```
array(['Computer science, computer engineering, or software engineering',
       'Office',
       'Information systems, information technology, or system administration',
       'Another engineering discipline (ex. civil, electrical, mechanical)',
       'A business discipline (ex. accounting, finance, marketing)',
       'Web development or web design', 'Mathematics or statistics',
       'A social science (ex. anthropology, psychology, political science)',
       'Fine arts or performing arts (ex. graphic design, music, studio art)',
       'A natural science (ex. biology, chemistry, physics)',
       'A humanities discipline (ex. literature, history, philosophy)',
       'I never declared a major',
       'A health science (ex. nursing, pharmacy, radiology)'],
      dtype=object)
```

# How many in health Science Major

In [125…
```
major=['A health science (ex. nursing, pharmacy, radiology)']
df2[df2['UndergradMajor'].isin(major)]
```

Out[125...

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| **567** | 1202 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time |
| **3466** | 7291 | I am a developer by profession | Yes | Once a month or more often | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **3490** | 7349 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **3772** | 7965 | I am a developer by profession | Yes | Once a month or more often | OSS is, on average, of LOWER quality than prop... | Employed full-time |
| **4112** | 8673 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **4212** | 8886 | I am a developer by profession | Yes | Never | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **4708** | 9973 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time |
| **4876** | 10345 | I am a developer by profession | Yes | Less than once per year | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **5440** | 11589 | I am a developer by profession | Yes | Less than once per year | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **5640** | 12011 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment |
|---|---|---|---|---|---|---|
| **5849** | 12475 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time |
| **7352** | 15740 | I am a developer by profession | No | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **7507** | 16097 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **7646** | 16434 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **8071** | 17364 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time |
| **8140** | 17535 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time |
| **8218** | 17695 | I am not primarily a developer, but I write co... | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time |
| **8300** | 17871 | I am a developer by profession | Yes | Less than once a month but more than once per ... | OSS is, on average, of HIGHER quality than pro... | Employed full-time |
| **8428** | 18196 | I am a developer by profession | Yes | Less than once a month but more than once per ... | The quality of OSS and closed source software ... | Employed full-time |
| **10019** | 21574 | I am not primarily a developer, but I write co... | Yes | Never | The quality of OSS and closed source software ... | Employed full-time |

| | Respondent | MainBranch | Hobbyist | OpenSourcer | OpenSource | Employment | |
|---|---|---|---|---|---|---|---|
| **10172** | 21931 | I am a developer by profession | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |
| **10872** | 23532 | I am a developer by profession | Yes | Less than once per year | OSS is, on average, of HIGHER quality than pro... | Employed full-time | |
| **11125** | 24135 | I am not primarily a developer, but I write co... | Yes | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | N |
| **11460** | 24937 | I am a developer by profession | No | Less than once per year | The quality of OSS and closed source software ... | Employed full-time | |

24 rows × 85 columns

# How many in each category

In [126…
```
df2['UndergradMajor'].value_counts()
```

Out[126…
```
UndergradMajor
Computer science, computer engineering, or software engineering          6953
Information systems, information technology, or system administration     794
Another engineering discipline (ex. civil, electrical, mechanical)       759
Office                                                                    737
Web development or web design                                             410
A natural science (ex. biology, chemistry, physics)                       403
Mathematics or statistics                                                 372
A business discipline (ex. accounting, finance, marketing)                244
A social science (ex. anthropology, psychology, political science)        210
A humanities discipline (ex. literature, history, philosophy)             207
Fine arts or performing arts (ex. graphic design, music, studio art)      161
I never declared a major                                                  124
A health science (ex. nursing, pharmacy, radiology)                        24
Name: count, dtype: int64
```

# Finding median salaray in ConvertedComp by first removing non numeric values

In [127…
```
df2['ConvertedComp'].describe()
```

Out[127…
```
count      11398
unique      3516
top       Office
freq         816
Name: ConvertedComp, dtype: object
```

In [128…
```
df2['ConvertedComp']
```

Out[128…
```
0         61000.0
1         95179.0
2         90000.0
3        455352.0
4         65277.0
           ...
11547    130000.0
11548     19880.0
11549    105000.0
11550     80371.0
11551      Office
Name: ConvertedComp, Length: 11398, dtype: object
```

In [129…
```
df_3=df2['ConvertedComp'].replace('Office',np.NaN)
df_3
```

Out[129…
```
0         61000.0
1         95179.0
2         90000.0
3        455352.0
4         65277.0
           ...
11547    130000.0
11548     19880.0
11549    105000.0
11550     80371.0
11551         NaN
Name: ConvertedComp, Length: 11398, dtype: float64
```

In [130…
```
median=df_3.median()
median
```

Out[130…
```
57745.0
```

# Replacing the null values of column with the median value we just calculated

In [131…
```
df2['ConvertedComp']=df2['ConvertedComp'].replace('Office',median)
```

In [132…
```
df2['ConvertedComp']
```

Out[132…
```
0         61000.0
1         95179.0
2         90000.0
3        455352.0
4         65277.0
           ...
11547    130000.0
11548     19880.0
11549    105000.0
11550     80371.0
11551     57745.0
Name: ConvertedComp, Length: 11398, dtype: float64
```

# TASK : Create a new normalized Column for annual salaries of all employees using 2 columns Converted Freq and Converted Comp

In [133…
```
df2.filter(items=['CompFreq','CompTotal'])
```

Out[133…

| | CompFreq | CompTotal |
|---|---|---|
| **0** | Yearly | 61000.0 |
| **1** | Yearly | 138000.0 |
| **2** | Yearly | 90000.0 |
| **3** | Monthly | 29000.0 |
| **4** | Yearly | 90000.0 |
| **…** | … | … |
| **11547** | Yearly | 130000.0 |
| **11548** | Yearly | 74400.0 |
| **11549** | Yearly | 105000.0 |
| **11550** | Yearly | 80000.0 |
| **11551** | Office | Office |

11398 rows × 2 columns

In [134…
```
df2['CompFreq']=df2['CompFreq'].replace('Office', np.NaN)
df2['CompFreq']
```

Out[134…
```
0         Yearly
1         Yearly
2         Yearly
3        Monthly
4         Yearly
          ...
11547     Yearly
11548     Yearly
11549     Yearly
11550     Yearly
11551        NaN
Name: CompFreq, Length: 11398, dtype: object
```

In [135…
```
df2['CompTotal']=df2['CompTotal'].replace('Office', np.NaN)
df2['CompTotal']
```

Out[135…
```
0         61000.0
1        138000.0
2         90000.0
3         29000.0
4         90000.0
           ...
11547    130000.0
11548     74400.0
11549    105000.0
11550     80000.0
11551         NaN
Name: CompTotal, Length: 11398, dtype: float64
```

In [136…
```
df2.filter(items=['CompFreq','CompTotal'])
```

Out[136…

| | CompFreq | CompTotal |
|---|---|---|
| **0** | Yearly | 61000.0 |
| **1** | Yearly | 138000.0 |
| **2** | Yearly | 90000.0 |
| **3** | Monthly | 29000.0 |
| **4** | Yearly | 90000.0 |
| **...** | ... | ... |
| **11547** | Yearly | 130000.0 |
| **11548** | Yearly | 74400.0 |
| **11549** | Yearly | 105000.0 |
| **11550** | Yearly | 80000.0 |
| **11551** | NaN | NaN |

11398 rows × 2 columns

In [137…
```
df2['CompFreq'].value_counts()
```

Out[137…
```
CompFreq
Yearly     6073
Monthly    4788
Weekly      331
Name: count, dtype: int64
```

In [138…
```
annualcomp=[]
for x,y in zip(df2['CompFreq'], df2['CompTotal']):
    if x=='Monthly':
        annualcomp.append(y*12)
    elif x=='Weekly':
        annualcomp.append(y*52)
    else:
        annualcomp.append(y)


df2['NormalizedAnnualCompensation']=anncomp
df2[['NormalizedAnnualCompensation']]
```

Out[138…

| | NormalizedAnnualCompensation |
|---|---|
| **0** | 61000.0 |
| **1** | 138000.0 |
| **2** | 90000.0 |
| **3** | 348000.0 |
| **4** | 90000.0 |
| **…** | … |
| **11547** | 130000.0 |
| **11548** | 74400.0 |
| **11549** | 105000.0 |
| **11550** | 80000.0 |
| **11551** | NaN |

11398 rows × 1 columns

# Median annual salary

In [139…
```
df2['NormalizedAnnualCompensation'].median()
```

Out[139…
```
100000.0
```

In [ ]: