In [5]:
```
import pandas as pd
import numpy as np

salary="salary_data_states.xlsx"
salary_full=pd.read_excel(salary)
```

In [6]:
```
salary_full
```

Out[6]:

PAID_WAGE_SUBMITTED_UNIT	JOB_TITLE	...	PREVAILING_WAGE_SOC_TITLE	WORK_STATE
year	SOFTWARE ENGINEER	...	Software Developers, Applications	Illinois
year	Assistant Professor of Marketing	...	Business Teachers, Postsecondary	California
year	SPECIAL EDUCATION TEACHER	...	Special Education Teachers, Kindergarten and E...	California
year	SCIENCE TEACHER	...	Biological Science Teachers, Postsecondary	Texas
year	SENIOR SOFTWARE ENGINEER	...	Software Developers, Systems Software	Oregon
...	...	...	...	.
hour	MIDDLE SCHOOL TEACHERS	...	Middle School Teachers, Except Special and Car...	Northern Mariana Islands
hour	PRESCHOOL TEACHER	...	Preschool Teachers, Except Special Education	Northern Mariana Islands
hour	TEACHER	...	Teachers and Instructors, All Other*	Northern Mariana Islands
hour	PRESCHOOL TEACHER	...	Preschool Teachers, Except Special Education	Northern Mariana Islands
hour	PRESCHOOL TEACHER	...	Preschool Teachers, Except Special Education	Northern Mariana Islands

In [7]:
```
salary_full.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167278 entries, 0 to 167277
Data columns (total 27 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   CASE_NUMBER                     167278 non-null  object 
 1   CASE_STATUS                     167278 non-null  object 
 2   CASE_RECEIVED_DATE              167278 non-null  object 
 3   DECISION_DATE                   167278 non-null  object 
 4   EMPLOYER_NAME                   167278 non-null  object 
 5   PREVAILING_WAGE_SUBMITTED       167278 non-null  float64
 6   PREVAILING_WAGE_SUBMITTED_UNIT  167278 non-null  object 
 7   PAID_WAGE_SUBMITTED             167278 non-null  float64
 8   PAID_WAGE_SUBMITTED_UNIT        167278 non-null  object 
 9   JOB_TITLE                       167278 non-null  object 
 10  WORK_CITY                       167275 non-null  object 
 11  EDUCATION_LEVEL_REQUIRED        11093 non-null   object 
 12  COLLEGE_MAJOR_REQUIRED          11051 non-null   object 
 13  EXPERIENCE_REQUIRED_Y_N         11093 non-null   object 
 14  EXPERIENCE_REQUIRED_NUM_MONTHS  4965 non-null    float64
 15  COUNTRY_OF_CITIZENSHIP          11093 non-null   object 
 16  PREVAILING_WAGE_SOC_CODE        167278 non-null  object 
 17  PREVAILING_WAGE_SOC_TITLE       167278 non-null  object 
 18  WORK_STATE                      167278 non-null  object 
 19  WORK_STATE_ABBREVIATION         167278 non-null  object 
 20  WORK_POSTAL_CODE                53674 non-null   object 
 21  FULL_TIME_POSITION_Y_N          156185 non-null  object 
 22  VISA_CLASS                      167278 non-null  object 
 23  PREVAILING_WAGE_PER_YEAR        167210 non-null  float64
 24  PAID_WAGE_PER_YEAR              167278 non-null  float64
 25  JOB_TITLE_SUBGROUP              167278 non-null  object 
 26  order                           167278 non-null  int64  
dtypes: float64(5), int64(1), object(21)
memory usage: 34.5+ MB
```

In [8]:
```
salary_full.head()
```

Out[8]:

G_WAGE_SOC_TITLE	WORK_STATE	WORK_STATE_ABBREVIATION	WORK_POSTAL_CODE	FULI
evelopers, Applications	Illinois	IL	NaN	
achers, Postsecondary	California	CA	94132.0	
ial Education Teachers, Kindergarten and E...	California	CA	NaN	
gical Science Teachers, Postsecondary	Texas	TX	NaN	
e Developers, Systems Software	Oregon	OR	NaN	

In [ ]:
```
## In Prevailing_wage_submitted_unit, some rows are year, month and bi-weekly
## calculate prevailing_wage_submitted based on yearly.
```

In [12]:
```
## find unique value of state

salary_full['WORK_STATE'].unique()

## We use full state name. No state code in the data.
```

Out[12]:
```
array(['Illinois', 'California', 'Texas', 'Oregon', 'New Jersey',
       'New York', 'Connecticut', 'Washington', 'Maryland',
       'North Carolina', 'District of Columbia', 'South Carolina',
       'Rhode Island', 'Colorado', 'Michigan', 'Missouri', 'Minnesota',
       'Wyoming', 'Louisiana', 'Pennsylvania', 'Tennessee', 'Idaho',
       'Massachusetts', 'Nebraska', 'Georgia', 'Ohio', 'Florida',
       'Indiana', 'Arizona', 'Kentucky', 'Iowa', 'Wisconsin', 'Alabama',
       'Arkansas', 'Virginia', 'New Mexico', 'West Virginia', 'Oklahoma',
       'Utah', 'Nevada', 'Mississippi', 'New Hampshire', 'Delaware',
       'Kansas', 'Alaska', 'Hawaii', 'Vermont', 'North Dakota', 'Maine',
       'Montana', 'Virgin Islands', 'South Dakota', 'Guam', 'Puerto Rico',
       'Palau', 'Guamam', 'Northern Mariana Islands'], dtype=object)
```

In [9]:
```
salary_full.describe().T
```

Out[9]:

	count	mean	std	min	25%	
PREVAILING_WAGE_SUBMITTED	167278.0	71157.518830	38746.239518	5.05	54475.00	69
PAID_WAGE_SUBMITTED	167278.0	81641.855212	41477.029632	5.05	61000.00	77
RIENCE_REQUIRED_NUM_MONTHS	4965.0	34.692044	22.317783	0.00	12.00	
PREVAILING_WAGE_PER_YEAR	167210.0	74274.868236	25356.245893	10504.00	56880.00	70
PAID_WAGE_PER_YEAR	167278.0	85532.766271	38738.466697	10500.00	63000.00	78
order	167278.0	83714.716305	48300.236431	1.00	41901.25	83

Check missing values

In [10]:
```
salary_full.isnull().sum()
```

Out[10]:
```
CASE_NUMBER                      0
CASE_STATUS                      0
CASE_RECEIVED_DATE               0
DECISION_DATE                    0
EMPLOYER_NAME                    0
PREVAILING_WAGE_SUBMITTED        0
PREVAILING_WAGE_SUBMITTED_UNIT   0
PAID_WAGE_SUBMITTED              0
PAID_WAGE_SUBMITTED_UNIT         0
JOB_TITLE                        0
WORK_CITY                        3
EDUCATION_LEVEL_REQUIRED    156185
COLLEGE_MAJOR_REQUIRED      156227
EXPERIENCE_REQUIRED_Y_N     156185
EXPERIENCE_REQUIRED_NUM_MONTHS  162313
COUNTRY_OF_CITIZENSHIP      156185
PREVAILING_WAGE_SOC_CODE         0
PREVAILING_WAGE_SOC_TITLE        0
WORK_STATE                       0
WORK_STATE_ABBREVIATION          0
WORK_POSTAL_CODE            113604
FULL_TIME_POSITION_Y_N       11093
VISA_CLASS                       0
PREVAILING_WAGE_PER_YEAR        68
PAID_WAGE_PER_YEAR               0
JOB_TITLE_SUBGROUP               0
order                            0
dtype: int64
```

In [ ]:
```
## We have missing values in several columns. We need to deal with those befo
```