Salary Data Cleaning

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
df = pd.read_csv("Levels_Fyi_Salary_Data[1].csv")
```

In [3]:

```
df.shape
```

Out[3]:

```
(62642, 29)
```

In [4]:

```
df.head(10)
```

Out[4]:

	timestamp	company	level	title	totalyearlycompensation	location	yearsofexperience	yearsatcompany	tag	basesalary	...	Doctorate_Degre
0	6/7/2017 11:33:27	Oracle	L3	Product Manager	127000	Redwood City, CA	1.5	1.5	NaN	107000.0	...	(
1	6/10/2017 17:11:29	eBay	SE 2	Software Engineer	100000	San Francisco, CA	5.0	3.0	NaN	0.0	...	(
2	6/11/2017 14:53:57	Amazon	L7	Product Manager	310000	Seattle, WA	8.0	0.0	NaN	155000.0	...	(
3	6/17/2017 0:23:14	Apple	M1	Software Engineering Manager	372000	Sunnyvale, CA	7.0	5.0	NaN	157000.0	...	(
4	6/20/2017 10:58:51	Microsoft	60	Software Engineer	157000	Mountain View, CA	5.0	3.0	NaN	0.0	...	(
5	6/21/2017 17:27:47	Microsoft	63	Software Engineer	208000	Seattle, WA	8.5	8.5	NaN	0.0	...	(
6	6/22/2017 12:37:51	Microsoft	65	Software Engineering Manager	300000	Redmond, WA	15.0	11.0	NaN	180000.0	...	(
7	6/22/2017 13:55:26	Microsoft	62	Software Engineer	156000	Seattle, WA	4.0	4.0	NaN	135000.0	...	(
8	6/22/2017 23:08:16	Microsoft	59	Software Engineer	120000	Redmond, WA	3.0	1.0	NaN	0.0	...	(
9	6/26/2017 21:25:45	Microsoft	63	Software Engineer	201000	Seattle, WA	12.0	6.0	NaN	157000.0	...	(

10 rows × 29 columns

In [5]:

```
df.isnull().sum()
```

Out[5]:

```
timestamp                      0
company                        5
level                        119
title                          0
totalyearlycompensation        0
location                       0
yearsofexperience              0
yearsatcompany                 0
tag                          854
basesalary                     0
stockgrantvalue                0
bonus                          0
gender                     19540
otherdetails               22505
cityid                         0
dmaid                          2
rowNumber                      0
Masters_Degree                 0
Bachelors_Degree               0
Doctorate_Degree               0
Highschool                     0
Some_College                   0
Race_Asian                     0
Race_White                     0
Race_Two_Or_More               0
Race_Black                     0
Race_Hispanic                  0
Race                       40215
Education                  32272
dtype: int64
```

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62642 entries, 0 to 62641
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   timestamp                62642 non-null  object 
 1   company                  62637 non-null  object 
 2   level                    62523 non-null  object 
 3   title                    62642 non-null  object 
 4   totalyearlycompensation  62642 non-null  int64  
 5   location                 62642 non-null  object 
 6   yearsofexperience        62642 non-null  float64
 7   yearsatcompany           62642 non-null  float64
 8   tag                      61788 non-null  object 
 9   basesalary               62642 non-null  float64
 10  stockgrantvalue          62642 non-null  float64
 11  bonus                    62642 non-null  float64
 12  gender                   43102 non-null  object 
 13  otherdetails             40137 non-null  object 
 14  cityid                   62642 non-null  int64  
 15  dmaid                    62640 non-null  float64
 16  rowNumber                62642 non-null  int64  
 17  Masters_Degree           62642 non-null  int64  
 18  Bachelors_Degree         62642 non-null  int64  
 19  Doctorate_Degree         62642 non-null  int64  
 20  Highschool               62642 non-null  int64  
 21  Some_College             62642 non-null  int64  
 22  Race_Asian               62642 non-null  int64  
 23  Race_White               62642 non-null  int64  
 24  Race_Two_Or_More         62642 non-null  int64  
 25  Race_Black               62642 non-null  int64  
 26  Race_Hispanic            62642 non-null  int64  
 27  Race                     22427 non-null  object 
 28  Education                30370 non-null  object 
dtypes: float64(6), int64(13), object(10)
memory usage: 13.9+ MB
```

In [7]:

```
df.duplicated().sum()
```

Out[7]:

```
0
```

Time Stamp

In [8]:

```
df["timestamp"].isnull().sum()
```

Out[8]:

```
0
```

In [9]:

```
df["timestamp"].value_counts()
```

Out[9]:

```
2/25/2020 13:25:07     3
1/10/2019 21:44:02     3
10/25/2019 10:26:31    3
8/18/2019 4:59:01      2
9/6/2019 6:49:56       2
                      ..
5/14/2020 1:52:06      1
5/14/2020 2:40:13      1
5/14/2020 3:03:59      1
5/14/2020 4:12:43      1
1/29/2019 5:12:59      1
Name: timestamp, Length: 62561, dtype: int64
```

In [10]:

```
import datetime
from datetime import datetime
```

In [11]:

```
df["timestamp"] = pd.to_datetime(df["timestamp"]).dt.date
```

In [12]:

```
df.rename(columns = {"timestamp":"Date"}, inplace = True)
```

Company

In [13]:

```
df["company"].isna().sum()
```

Out[13]:

```
5
```

In [14]:

```
df.dropna(subset = "company", inplace=True)
```

In [15]:

```
df["company"].value_counts()
```

Out[15]:

```
Amazon                      8126
Microsoft                   5216
Google                      4330
Facebook                    2990
Apple                       2028
                            ...
Samsung research America       1
Bny Mellon                     1
yelp                           1
Bloomberg lp                   1
tableau software               1
Name: company, Length: 1631, dtype: int64
```

In [16]:

```
df.rename(columns= {"company":"Company"}, inplace = True)
```

Level

In [17]:

```
df["level"].value_counts()
```

Out[17]:

```
L4                             5014
L5                             4871
L3                             3337
L6                             2871
Senior Software Engineer       1443
                               ... 
Mid Market                        1
Gr 7                              1
Senior BA                         1
Consulting Analyst                1
Bioinformatics Scientist II       1
Name: level, Length: 2923, dtype: int64
```

In [18]:

```
df["level"].isnull().sum()
```

Out[18]:

```
119
```

In [19]:

```
df["level"].fillna(df["level"].mode()[0], inplace= True)
```

In [20]:

```
df["level"].mode()
```

Out[20]:

```
0    L4
Name: level, dtype: object
```

In [21]:

```
df.rename(columns= {"level":"Level"}, inplace = True)
```

Title

In [22]:

```
df["title"].isnull().sum()
```

Out[22]:

```
0
```

In [23]:

```
df["title"].value_counts()
```

Out[23]:

```
Software Engineer               41227
Product Manager                  4673
Software Engineering Manager     3568
Data Scientist                   2578
Hardware Engineer                2200
Product Designer                 1516
Technical Program Manager        1381
Solution Architect               1157
Management Consultant             976
Business Analyst                  885
Marketing                         710
Mechanical Engineer               490
Sales                             461
Recruiter                         451
Human Resources                   364
Name: title, dtype: int64
```

In [24]:

```
df.rename(columns= {"title" : "Title"}, inplace = True)
```

Total year compensation

In [25]:

```
df["totalyearlycompensation"].isnull().sum()
```

Out[25]:

```
0
```

In [26]:

```
df["totalyearlycompensation"].value_counts()
```

Out[26]:

```
200000     1196
150000     1106
250000      907
180000      904
160000      874
           ... 
155500        1
160500        1
1355000       1
865000        1
814000        1
Name: totalyearlycompensation, Length: 893, dtype: int64
```

In [27]:

```
df.rename(columns = {"totalyearlycompensation" : "Yearly Compensation"}, inplace = True)
```

Location

In [28]:

```
df["location"].isnull().sum()
```

Out[28]:

```
0
```

In [29]:

```
df["location"].value_counts()
```

Out[29]:

```
Seattle, WA                      8701
San Francisco, CA                6797
New York, NY                     4562
Redmond, WA                      2649
Mountain View, CA                2275
                                 ... 
San Fernando, LB, Philippines       1
Suwanee, GA                         1
Oxford, MS                          1
Wayne, PA                           1
Hilbert, WI                         1
Name: location, Length: 1050, dtype: int64
```

In [30]:

```
df.rename(columns = {"location" : "Location"}, inplace = True)
```

Years of Experince

In [31]:

```
df["yearsofexperience"].isnull().sum()
```

Out[31]:

```
0
```

In [32]:

```
for values in df["yearsofexperience"].values:
   print(values)
```

```
6.0
3.0
2.0
1.0
0.0
1.0
0.0
7.0
13.0
8.0
5.0
6.0
2.0
4.0
10.0
2.0
14.0
8.0
0.0
```

In [33]:

```
pd.set_option("display.max_rows", None)
df["yearsofexperience"].value_counts(sort= True, ascending= True)
```

Out[33]:

```
10.50       1
0.60        1
69.00       1
6.75        1
45.00       1
1.40        1
0.58        1
0.30        1
1.60        1
0.80        2
5.50        2
42.00       2
3.80        2
41.00       2
0.25        2
8.50        2
4.50        2
11.50       2
39.00       3
36.00       3
7.50        3
34.00       4
38.00       4
0.50        4
6.50        4
37.00       5
3.50        5
31.00       7
1.50       10
32.00      11
2.50       12
40.00      12
33.00      12
29.00      22
35.00      30
27.00      39
28.00      42
26.00      52
24.00     116
30.00     121
23.00     146
21.00     187
22.00     223
19.00     299
25.00     394
17.00     510
18.00     649
16.00     761
13.00    1165
11.00    1182
14.00    1182
20.00    1910
9.00     2051
12.00    2165
15.00    2931
7.00     3451
8.00     3622
6.00     3990
1.00     4070
0.00     4603
10.00    4760
4.00     4896
2.00     5528
3.00     5529
5.00     5885
Name: yearsofexperience, dtype: int64
```

In [34]:

```
df.rename(columns = {"yearsofexperience" : "Experience(years)"}, inplace = True)
```

Years at company

In [35]:

```
df["yearsatcompany"].isnull().sum()
```

Out[35]:

```
0
```

In [36]:

```
for values in df["yearsatcompany"].values:
    print(values)
```

```
2.0
3.0
2.0
0.0
0.0
1.0
0.0
7.0
3.0
8.0
1.0
4.0
2.0
2.0
1.0
2.0
12.0
4.0
0.0
```

In [37]:

```
df.rename(columns = {"yearsatcompany" : "Years at Company"}, inplace = True)
```

Tags

In [38]:

```
df["tag"].isnull().sum()
```

Out[38]:

```
853
```

In [39]:

```
df["tag"].value_counts()
```

```
Transfer Pricing                                                                            1
Youtube                                                                                     1
Strategy & Consulting                                                                       1
AAA Games                                                                                   1
Planning and Control                                                                        1
Oculus                                                                                      1
Trading Infrastructure                                                                      1
Tech recruiting                                                                             1
E-Commerce                                                                                  1
Customs                                                                                     1
bioinformatics                                                                              1
Partners                                                                                    1
ATE Test Engineer                                                                           1
UX Writing                                                                                  1
Systems Architecture                                                                        1
Subscription                                                                                1
Web Browser Developer                                                                       1
System Testing                                                                              1
Merchant acquisition                                                                        1
```

In [40]:

```
df["tag"].fillna(value = "No Tags", inplace = True)
```

In [41]:

```
df.rename(columns = {"tag" : "Tag"}, inplace = True)
```

Base Salary

In [42]:

```
df["basesalary"].isnull().sum()
```

Out[42]:

```
0
```

In [43]:

```
df["basesalary"].value_counts()
```

```
91000.0       101
186000.0      100
181000.0       98
171000.0       96
79000.0        95
81000.0        95
36000.0        93
270000.0       93
74000.0        92
193000.0       92
73000.0        91
89000.0        90
48000.0        89
191000.0       88
196000.0       88
41000.0        87
25000.0        87
194000.0       86
42000.0        84
43000.0        83
```

In [44]:

```
df["basesalary"].min()
```

Out[44]:

```
0.0
```

In [45]:

```
df["basesalary"].replace(140000.0, df["basesalary"].min(), inplace = True)
```

In [46]:

```
df.rename(columns = {"basesalary" : "Base Salary"}, inplace = True)
```

Stock Grant Value

In [47]:

```
df["stockgrantvalue"].isnull().sum()
```

Out[47]:

```
0
```

In [48]:

```
for values in df["stockgrantvalue"].values:
    print(values)
```

```
0.0
30000.0
0.0
5000.0
0.0
0.0
0.0
0.0
50000.0
262000.0
20000.0
125000.0
18500.0
280000.0
0.0
18000.0
130000.0
0.0
120000.0
```

In [49]:

```
df.rename(columns = {"stockgrantvalue" : "Stock Grant"}, inplace = True)
```

Bonus

In [50]:

```
df["bonus"].isnull().sum()
```

Out[50]:

```
0
```

In [51]:

```
df["bonus"].value_counts()
```

```
275000.00        1
252000.00        1
246000.00        1
156000.00        1
205000.00        1
109000.00        1
188000.00        1
169000.00        1
1600.00          1
142000.00        1
27500.00         1
1000000.00       1
164000.00        1
149000.00        1
520000.00        1
184000.00        1
290000.00        1
153000.00        1
202000.00        1
14200.00         1
```

In [52]:

```
df.rename(columns= {"bonus" : "Bonus"}, inplace = True)
```

Gender

In [53]:

```
df["gender"].isnull().sum()
```

Out[53]:

```
19539
```

In [54]:

```
df["gender"].value_counts()
```

Out[54]:

```
Male                               35698
Female                              6999
Other                                400
Title: Senior Software Engineer        1
Name: gender, dtype: int64
```

In [55]:

```
df["gender"].replace("Title: Senior Software Engineer", "Missing", inplace = True)
df["gender"].fillna("Missing", inplace = True)
```

In [56]:

```
df.rename(columns= {"gender" : "Gender"}, inplace = True)
```

Other details

In [57]:

```
df["otherdetails"].isnull().sum()
```

Out[57]:

```
22502
```

In [59]:

```
df["otherdetails"].value_counts()
```

```
$15k signing bonus, Title: Software Engineer, Race: Asian, Academic Level: Bachelor's degree
1
Title: Resident Engineer Ii, Race: White, Academic Level: Bachelor’s degree
1
Title: Security Engineer, Race: American Indian or Alaska Native
1
sdsa, Title: Software Engineer, Race: Asian, Academic Level: Master’s degree
1
150,000 sign on, Title: Sr. Engineering Manager, Race: White, Academic Level: Doctorate (PhD)
1
Title: Sde2, Race: Two or More Races, Academic Level: Doctorate (PhD)
1
10k relocation, 20k sign in bonus, 35k RSU over 4 years, Title: Tech Lead, Race: Hispanic / Latino, Academic Level: Mast
er’s degree
1
Title: Pm 3
1
industry hire
1
Name: otherdetails, dtype: int64
```

In [60]:

```
df.rename(columns= {"otherdetails" : "Other Details"}, inplace = True)
```

City ID

In [61]:

```
df["cityid"].isnull().sum()
```

Out[61]:

```
0
```

In [62]:

```
df["cityid"].value_counts()
```

```
8931        1
7479        1
7160        1
6611        1
20661       1
8554        1
3748        1
10285       1
11018       1
5028        1
34700       1
18094       1
6762        1
4960        1
8360        1
13149       1
7786        1
38771       1
9509        1
28230       1
```

In [63]:

```
df.rename(columns= {"cityid" : "CIty ID"}, inplace = True)
```

Dmaid

In [64]:

```
df["dmaid"].isnull().sum()
```

Out[64]:

```
2
```

In [65]:

```
df["dmaid"].value_counts()
```

```
516.0        1
522.0        1
574.0        1
537.0        1
760.0        1
698.0        1
576.0        1
705.0        1
746.0        1
546.0        1
734.0        1
632.0        1
540.0        1
610.0        1
656.0        1
503.0        1
687.0        1
693.0        1
651.0        1
658.0        1
```

In [66]:

```
df["dmaid"].fillna(0.0, inplace = True)
```

In [67]:

```
df.rename(columns= {"dmaid" : "Dmaid"}, inplace = True)
```

Row Number

In [68]:

```
df["rowNumber"].isnull().sum()
```

Out[68]:

```
0
```

In [69]:

```
df["rowNumber"].value_counts()
```

```
28471    1
28473    1
28475    1
28476    1
28477    1
28481    1
28501    1
28482    1
28483    1
28484    1
28487    1
28488    1
28491    1
28494    1
28495    1
28496    1
28497    1
28498    1
28499    1
5424     1
```

In [70]:

```
df.rename(columns= {"rowNumber" : "Row Number"}, inplace = True)
```

Masters Degree

In [71]:

```
df["Masters_Degree"].isnull().sum()
```

Out[71]:

```
0
```

In [72]:

```
df["Masters_Degree"].value_counts()
```

Out[72]:

```
0    47246
1    15391
Name: Masters_Degree, dtype: int64
```

Bachelors_Degree

In [73]:

```
df["Bachelors_Degree"].isnull().sum()
```

Out[73]:

```
0
```

In [74]:

```
df["Bachelors_Degree"].value_counts()
```

Out[74]:

```
0    50034
1    12603
Name: Bachelors_Degree, dtype: int64
```

Doctorate_Degree

In [75]:

```
df["Doctorate_Degree"].isnull().sum()
```

Out[75]:

```
0
```

In [76]:

```
df["Doctorate_Degree"].value_counts()
```

Out[76]:

```
0    60834
1     1803
Name: Doctorate_Degree, dtype: int64
```

High_School

In [77]:

```
df["Highschool"].isnull().sum()
```

Out[77]:

```
0
```

In [78]:

```
df["Highschool"].value_counts()
```

Out[78]:

```
0    62317
1      320
Name: Highschool, dtype: int64
```

In [79]:

```
df.rename(columns= {"Highschool" : "High_School"}, inplace = True)
```

Some_College

In [80]:

```
df["Some_College"].isnull().sum()
```

Out[80]:

```
0
```

In [81]:

```
df["Some_College"].value_counts()
```

Out[81]:

```
0    62282
1      355
Name: Some_College, dtype: int64
```

Race_Asian

In [82]:

```
df["Race_Asian"].isnull().sum()
```

Out[82]:

```
0
```

In [83]:

```
df["Race_Asian"].value_counts()
```

Out[83]:

```
0    50865
1    11772
Name: Race_Asian, dtype: int64
```

Race_Two_or_More

In [84]:

```
df["Race_Two_Or_More"].isnull().sum()
```

Out[84]:

```
0
```

In [85]:

```
df["Race_Two_Or_More"].value_counts()
```

Out[85]:

```
0    61833
1      804
Name: Race_Two_Or_More, dtype: int64
```

Race_Black

In [86]:

```
df["Race_Black"].isnull().sum()
```

Out[86]:

```
0
```

In [87]:

```
df["Race_Black"].value_counts()
```

Out[87]:

```
0    61947
1      690
Name: Race_Black, dtype: int64
```

Race_Hispanic

In [88]:

```
df["Race_Hispanic"].isnull().sum()
```

Out[88]:

```
0
```

In [89]:

```
df["Race_Hispanic"].value_counts()
```

Out[89]:

```
0    61508
1     1129
Name: Race_Hispanic, dtype: int64
```

Race

In [90]:

```
df["Race"].isnull().sum()
```

Out[90]:

```
40212
```

In [91]:

```
df["Race"].value_counts()
```

Out[91]:

```
Asian          11772
White           8031
Hispanic        1128
Two Or More      804
Black            690
Name: Race, dtype: int64
```

In [92]:

```
df["Race"].fillna("No_Race", inplace = True)
```

Education

In [93]:

```
df["Education"].isnull().sum()
```

Out[93]:

```
32269
```

In [94]:

```
df["Education"].value_counts()
```

Out[94]:

```
Master's Degree      15391
Bachelor's Degree    12599
PhD                   1703
Some College           355
Highschool             320
Name: Education, dtype: int64
```

In [95]:

```
df["Education"].fillna("Missing", inplace = True)
```