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



import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df=pd.read_csv('HRDataset_v14.csv')
df
```

Out[2]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	Dep
0	Adinolfi, Wilson K	10026	0	0	1	1	
1	Ait Sidi, Karthikeyan	10084	1	1	1	5	
2	Akinkuolie, Sarah	10196	1	1	0	5	
3	Alagbe,Trina	10088	1	1	0	1	
4	Anderson, Carol	10069	0	2	0	5	
...	...	...	...	...	...	...	
306	Woodson, Jason	10135	0	0	1	1	
307	Ybarra, Catherine	10301	0	0	0	5	
308	Zamora, Jennifer	10010	0	0	0	1	
309	Zhou, Julia	10043	0	0	0	1	
310	Zima, Colleen	10271	0	4	0	1	

311 rows × 36 columns

In [3]:
```
#data understanding
df.columns
```

Out[3]:
```
Index(['Employee_Name', 'EmpID', 'MarriedID', 'MaritalStatusID', 'GenderID',
       'EmpStatusID', 'DeptID', 'PerfScoreID', 'FromDiversityJobFairID',
       'Salary', 'Termd', 'PositionID', 'Position', 'State', 'Zip', 'DOB',
       'Sex', 'MaritalDesc', 'CitizenDesc', 'HispanicLatino', 'RaceDesc',
       'DateofHire', 'DateofTermination', 'TermReason', 'EmploymentStatus',
       'Department', 'ManagerName', 'ManagerID', 'RecruitmentSource',
       'PerformanceScore', 'EngagementSurvey', 'EmpSatisfaction',
       'SpecialProjectsCount', 'LastPerformanceReview_Date', 'DaysLateLast30',
       'Absences'],
      dtype='object')
```

In [4]:
```
df.shape
```

Out[4]:
```
(311, 36)
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 311 entries, 0 to 310
Data columns (total 36 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Employee_Name               311 non-null    object 
 1   EmpID                       311 non-null    int64  
 2   MarriedID                   311 non-null    int64  
 3   MaritalStatusID             311 non-null    int64  
 4   GenderID                    311 non-null    int64  
 5   EmpStatusID                 311 non-null    int64  
 6   DeptID                      311 non-null    int64  
 7   PerfScoreID                 311 non-null    int64  
 8   FromDiversityJobFairID      311 non-null    int64  
 9   Salary                      311 non-null    int64  
 10  Termd                       311 non-null    int64  
 11  PositionID                  311 non-null    int64  
 12  Position                    311 non-null    object 
 13  State                       311 non-null    object 
 14  Zip                         311 non-null    int64  
 15  DOB                         311 non-null    object 
 16  Sex                         311 non-null    object 
 17  MaritalDesc                 311 non-null    object 
 18  CitizenDesc                 311 non-null    object 
 19  HispanicLatino              311 non-null    object 
 20  RaceDesc                    311 non-null    object 
 21  DateofHire                  311 non-null    object 
 22  DateofTermination           104 non-null    object 
 23  TermReason                  311 non-null    object 
 24  EmploymentStatus            311 non-null    object 
 25  Department                  311 non-null    object 
 26  ManagerName                 311 non-null    object 
 27  ManagerID                   303 non-null    float64
 28  RecruitmentSource           311 non-null    object 
 29  PerformanceScore            311 non-null    object 
 30  EngagementSurvey            311 non-null    float64
 31  EmpSatisfaction             311 non-null    int64  
 32  SpecialProjectsCount        311 non-null    int64  
 33  LastPerformanceReview_Date  311 non-null    object 
 34  DaysLateLast30              311 non-null    int64  
 35  Absences                    311 non-null    int64  
dtypes: float64(2), int64(16), object(18)
memory usage: 87.6+ KB
```

In [6]:
```
df.dtypes
```

Out[6]:
```
Employee_Name                object
EmpID                         int64
MarriedID                     int64
MaritalStatusID               int64
GenderID                      int64
EmpStatusID                   int64
DeptID                        int64
PerfScoreID                   int64
FromDiversityJobFairID        int64
Salary                        int64
Termd                         int64
PositionID                    int64
Position                     object
State                        object
Zip                           int64
DOB                          object
Sex                          object
MaritalDesc                  object
CitizenDesc                  object
HispanicLatino               object
RaceDesc                     object
DateofHire                   object
DateofTermination            object
TermReason                   object
EmploymentStatus             object
Department                   object
ManagerName                  object
ManagerID                   float64
RecruitmentSource            object
PerformanceScore             object
EngagementSurvey            float64
EmpSatisfaction               int64
SpecialProjectsCount          int64
LastPerformanceReview_Date   object
DaysLateLast30                int64
Absences                      int64
dtype: object
```

In [7]:
```
#cleaning the data

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

Out[7]:
```
Employee_Name                0
EmpID                        0
MarriedID                    0
MaritalStatusID              0
GenderID                     0
EmpStatusID                  0
DeptID                       0
PerfScoreID                  0
FromDiversityJobFairID       0
Salary                       0
Termd                        0
PositionID                   0
Position                     0
State                        0
Zip                          0
DOB                          0
Sex                          0
MaritalDesc                  0
CitizenDesc                  0
HispanicLatino               0
RaceDesc                     0
DateofHire                   0
DateofTermination          207
TermReason                   0
EmploymentStatus             0
Department                   0
ManagerName                  0
ManagerID                    8
RecruitmentSource            0
PerformanceScore             0
EngagementSurvey             0
EmpSatisfaction              0
SpecialProjectsCount         0
LastPerformanceReview_Date   0
DaysLateLast30               0
Absences                     0
dtype: int64
```

In [10]:
```
df.fillna("0", inplace = True)
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Employee_Name                0
EmpID                        0
MarriedID                    0
MaritalStatusID              0
GenderID                     0
EmpStatusID                  0
DeptID                       0
PerfScoreID                  0
FromDiversityJobFairID       0
Salary                       0
Termd                        0
PositionID                   0
Position                     0
State                        0
Zip                          0
DOB                          0
Sex                          0
MaritalDesc                  0
CitizenDesc                  0
HispanicLatino               0
RaceDesc                     0
DateofHire                   0
DateofTermination            0
TermReason                   0
EmploymentStatus             0
Department                   0
ManagerName                  0
ManagerID                    0
RecruitmentSource            0
PerformanceScore             0
EngagementSurvey             0
EmpSatisfaction              0
SpecialProjectsCount         0
LastPerformanceReview_Date   0
DaysLateLast30               0
Absences                     0
dtype: int64
```

In [12]:
```
df.duplicated().sum()
```

Out[12]:
```
np.int64(0)
```

In [13]:
```
df.drop_duplicates(inplace=True)
```

In [14]:
```
#EDA
#TOP 10 EMPLOYEES WITH HIGHEST SALARY
```

In [15]:
```
df.columns
```

Out[15]:
```
Index(['Employee_Name', 'EmpID', 'MarriedID', 'MaritalStatusID', 'GenderID',
       'EmpStatusID', 'DeptID', 'PerfScoreID', 'FromDiversityJobFairID',
       'Salary', 'Termd', 'PositionID', 'Position', 'State', 'Zip', 'DOB',
       'Sex', 'MaritalDesc', 'CitizenDesc', 'HispanicLatino', 'RaceDesc',
       'DateofHire', 'DateofTermination', 'TermReason', 'EmploymentStatus',
       'Department', 'ManagerName', 'ManagerID', 'RecruitmentSource',
       'PerformanceScore', 'EngagementSurvey', 'EmpSatisfaction',
       'SpecialProjectsCount', 'LastPerformanceReview_Date', 'DaysLateLast30',
       'Absences'],
      dtype='object')
```

In [20]:
```
df["Salary"].sort_values(ascending=False).head(10)
```

Out[20]:
```
150    250000
308    220450
131    180000
96     178000
55     170500
190    157000
240    150290
244    148999
243    140920
76     138888
Name: Salary, dtype: int64
```

In [21]:
```
#Employees who needs the special attention
#Performance Improvement Plan(PIP)
```

In [22]:
```
df['PerformanceScore'].unique()
```

Out[22]:
```
array(['Exceeds', 'Fully Meets', 'Needs Improvement', 'PIP'], dtype=object)
```

In [23]:
```
df[df['PerformanceScore'] == 'PIP']
```

Out[23]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	Dep
67	Delarge, Alex	10306	0	0	1	1	
69	Desimone, Carl	10310	1	1	1	1	
72	Dietrich, Jenna	10304	0	0	0	1	
83	Erilus, Angela	10299	0	3	0	1	
90	Fernandes, Nilson	10308	1	1	1	1	
91	Fett, Boba	10309	0	0	1	1	
95	Forrest, Alex	10305	1	1	1	1	
112	Gonzalez, Juan	10300	1	1	1	5	
188	Miller, Ned	10298	0	0	1	5	
205	O'hare, Lynn	10303	0	0	0	4	
263	Sparks, Taylor	10302	1	1	0	1	
267	Stansfield, Norman	10307	1	1	1	1	
307	Ybarra, Catherine	10301	0	0	0	5	

13 rows × 36 columns

In [24]:
```
people_pip=df[df['PerformanceScore']=='PIP'].Employee_Name
```

In [25]:
```
len(people_pip)
```

Out[25]:
```
13
```

In [26]:
```
people_pip
```

Out[26]:
```
67           Delarge, Alex
69         Desimone, Carl
72       Dietrich, Jenna
83          Erilus, Angela
90     Fernandes, Nilson
91              Fett, Boba
95           Forrest, Alex
112         Gonzalez, Juan
188            Miller, Ned
205           O'hare, Lynn
263       Sparks, Taylor
267     Stansfield, Norman
307     Ybarra, Catherine
Name: Employee_Name, dtype: object
```

In [27]:
```
#no. of abscenes
```

In [31]:
```
df['Absences'].value_counts()
```

Out[31]:
```
Absences
4     23
16    23
7     21
2     21
15    20
14    17
13    17
3     16
19    16
6     16
11    15
17    15
1     14
9     14
20    14
5     12
8     11
10    10
12     8
18     8
Name: count, dtype: int64
```

In [33]:
```
#how many employees are married and how many are not
```

In [34]:
```
df['MarriedID'].value_counts()
```

Out[34]:
```
MarriedID
0    187
1    124
Name: count, dtype: int64
```

In [35]:
```
# insights...187 employees are unmarried and 124 are married
```

In [37]:
```
#count for how many employees have special project

df[df['SpecialProjectsCount'] !=0]
```

Out[37]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	Dep
1	Ait Sidi, Karthikeyan	10084	1	1	1	5	
6	Andreola, Colby	10194	0	0	0	1	
9	Bacong, Alejandro	10250	0	2	1	1	
12	Barbossa, Hector	10012	0	2	1	1	
18	Becker, Renee	10245	0	0	0	4	
...	...	...	...	...	...	...	
292	Voldemort, Lord	10118	1	1	1	4	
298	Wang, Charlie	10172	0	0	1	1	
299	Warfield, Sarah	10127	0	4	0	1	
308	Zamora, Jennifer	10010	0	0	0	1	
309	Zhou, Julia	10043	0	0	0	1	

70 rows × 36 columns

In [38]:
```
df['SpecialProjectsCount'].sort_values(ascending = False)
```

Out[38]:
```
299    8
61     8
254    7
162    7
70     7
      ..
8      0
267    0
266    0
265    0
20     0
Name: SpecialProjectsCount, Length: 311, dtype: int64
```

In [41]:
```
df[df['SpecialProjectsCount'] == 0]
```

Out[41]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	Dep
0	Adinolfi, Wilson K	10026	0	0	1	1	
2	Akinkuolie, Sarah	10196	1	1	0	5	
3	Alagbe,Trina	10088	1	1	0	1	
4	Anderson, Carol	10069	0	2	0	5	
5	Anderson, Linda	10002	0	0	0	1	
...	...	...	...	...	...	...	
304	Winthrop, Jordan	10033	0	0	1	5	
305	Wolk, Hang T	10174	0	0	0	1	
306	Woodson, Jason	10135	0	0	1	1	
307	Ybarra, Catherine	10301	0	0	0	5	
310	Zima, Colleen	10271	0	4	0	1	

241 rows × 36 columns

In [42]:
```
#insights >> out of 311 employes 70 employees have special project
```

In [43]:
```
#visualization highest salary vs lowest salary
```

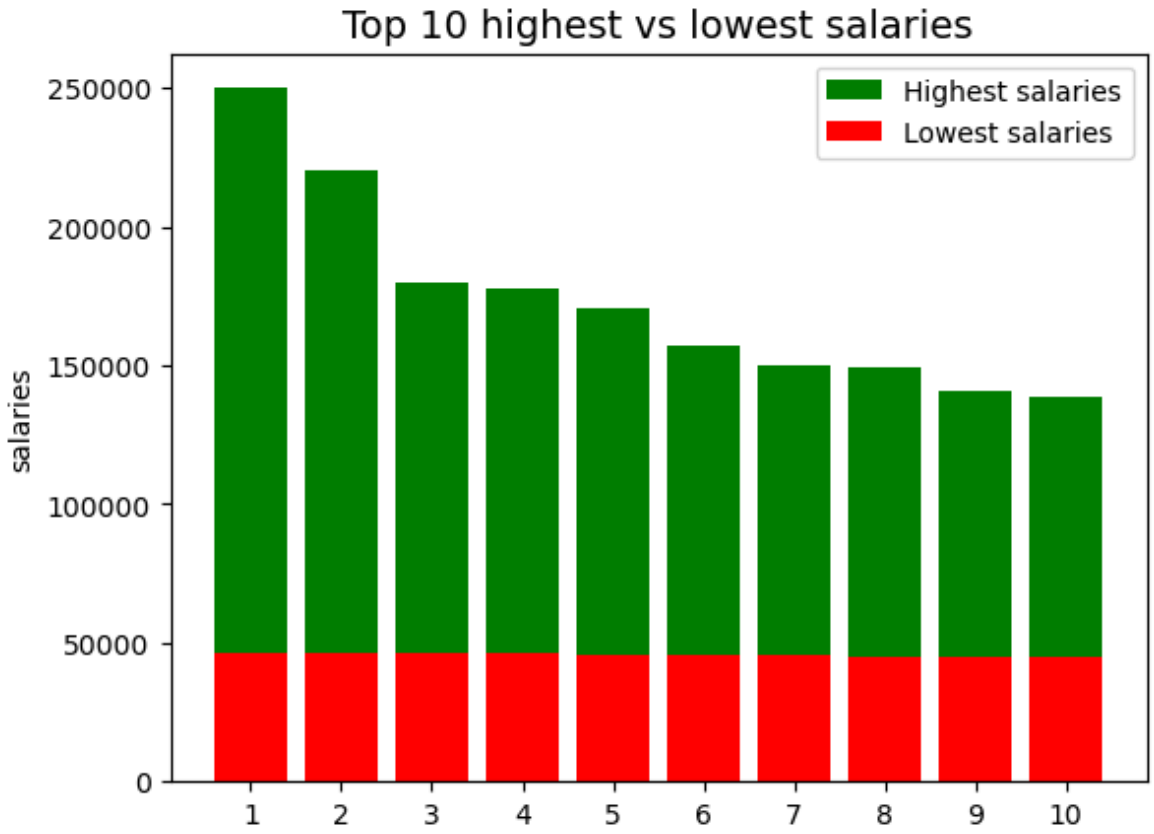
In [45]:
```
df['Salary'].sort_values(ascending=False).head(10)
```

Out[45]:
```
150    250000
308    220450
131    180000
96     178000
55     170500
190    157000
240    150290
244    148999
243    140920
76     138888
Name: Salary, dtype: int64
```

In [46]:
```
df['Salary'].sort_values(ascending=False).tail(10)
```

Out[46]:
```
226    46430
247    46428
74     46335
159    46120
216    45998
152    45433
176    45395
231    45115
140    45069
310    45046
Name: Salary, dtype: int64
```

In [47]:
```
c = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

x = df['Salary'].sort_values(ascending = False).head(10)
y = df['Salary'].sort_values(ascending = False).tail(10)


plt.bar(c, x, color = 'g', label = 'Highest salaries')
plt.bar(c, y, color = 'r', label = 'Lowest salaries')



plt.title('Top 10 highest vs lowest salaries', fontsize = 14)

plt.xticks(c)
plt.ylabel('salaries')
plt.legend()
plt.show()
```



In [48]:
```
#highest salary varies but lowest salaries are mostly in range
```

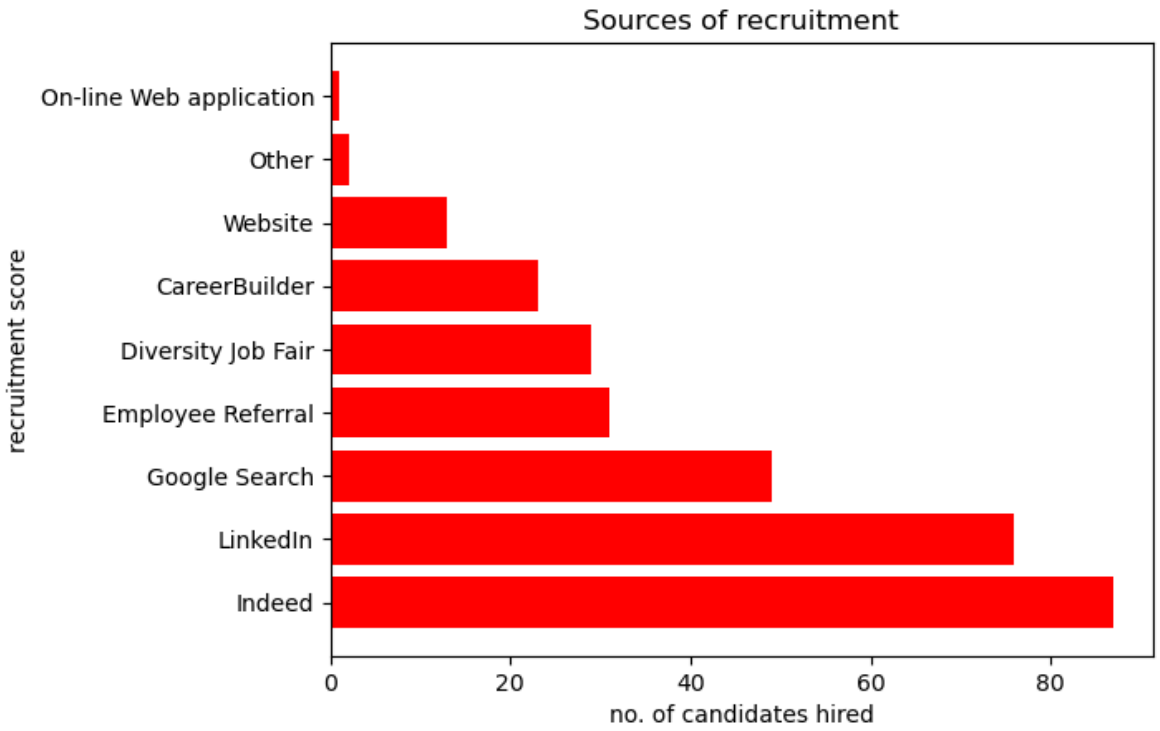
In [55]:
```
#source of recruitement
df['RecruitmentSource'].unique()
```

Out[55]:
```
array(['LinkedIn', 'Indeed', 'Google Search', 'Employee Referral',
       'Diversity Job Fair', 'On-line Web application', 'CareerBuilder',
       'Website', 'Other'], dtype=object)
```

In [56]:
```
l=df['RecruitmentSource'].value_counts()
l
```

Out[56]:
```
RecruitmentSource
Indeed                      87
LinkedIn                    76
Google Search               49
Employee Referral           31
Diversity Job Fair          29
CareerBuilder               23
Website                     13
Other                        2
On-line Web application      1
Name: count, dtype: int64
```

In [58]:
```
plt.barh(l.index ,l,color='r')
plt.title('Sources of recruitment', fontsize = 12)

plt.xlabel('no. of candidates hired')
plt.ylabel('recruitment score')
plt.show()
```



In [59]:
```
#INSIGHTS INEED IS THE MOST COMMON
```

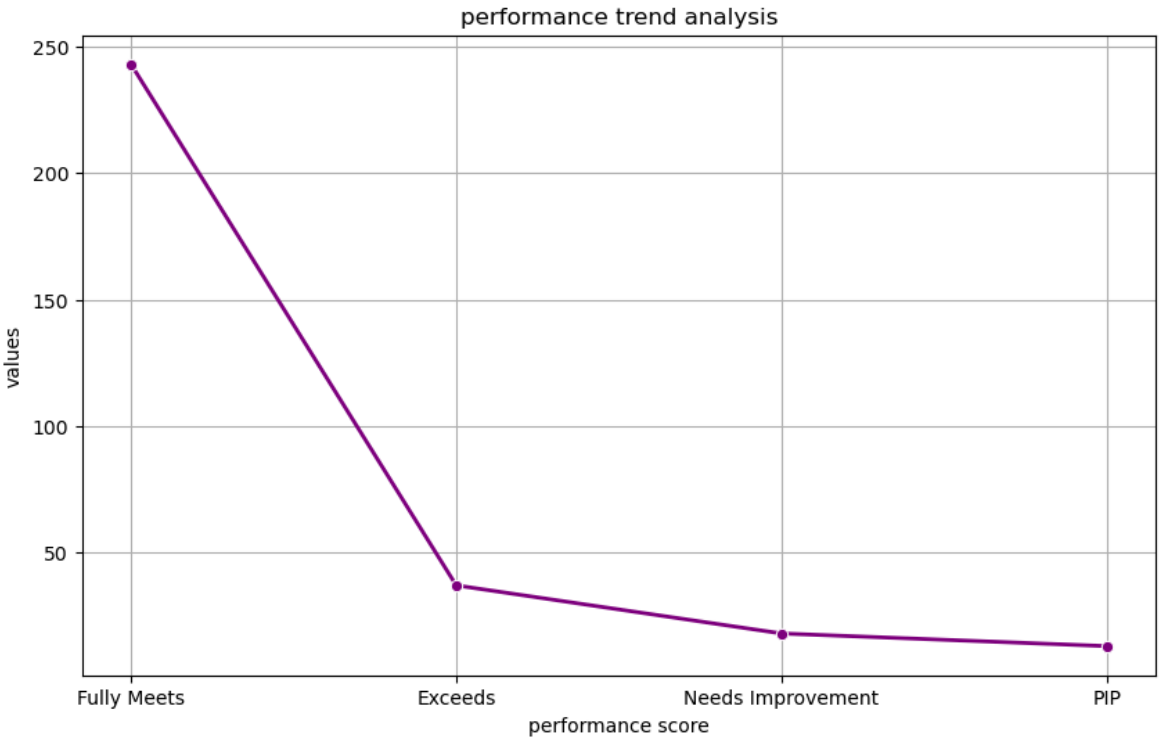
In [60]:
```
#performance trend analysis
```

In [62]:
```
z= df['PerformanceScore'].value_counts()
z
```

Out[62]:
```
PerformanceScore
Fully Meets          243
Exceeds               37
Needs Improvement     18
PIP                   13
Name: count, dtype: int64
```

In [65]:
```
plt.figure(figsize=(10,6))

sns.lineplot(data=z,marker='o',color='purple',linewidth=2)
```

```
plt.title('performance trend analysis')
plt.xlabel('performance score')
plt.ylabel('values')
plt.grid()
plt.show()
```

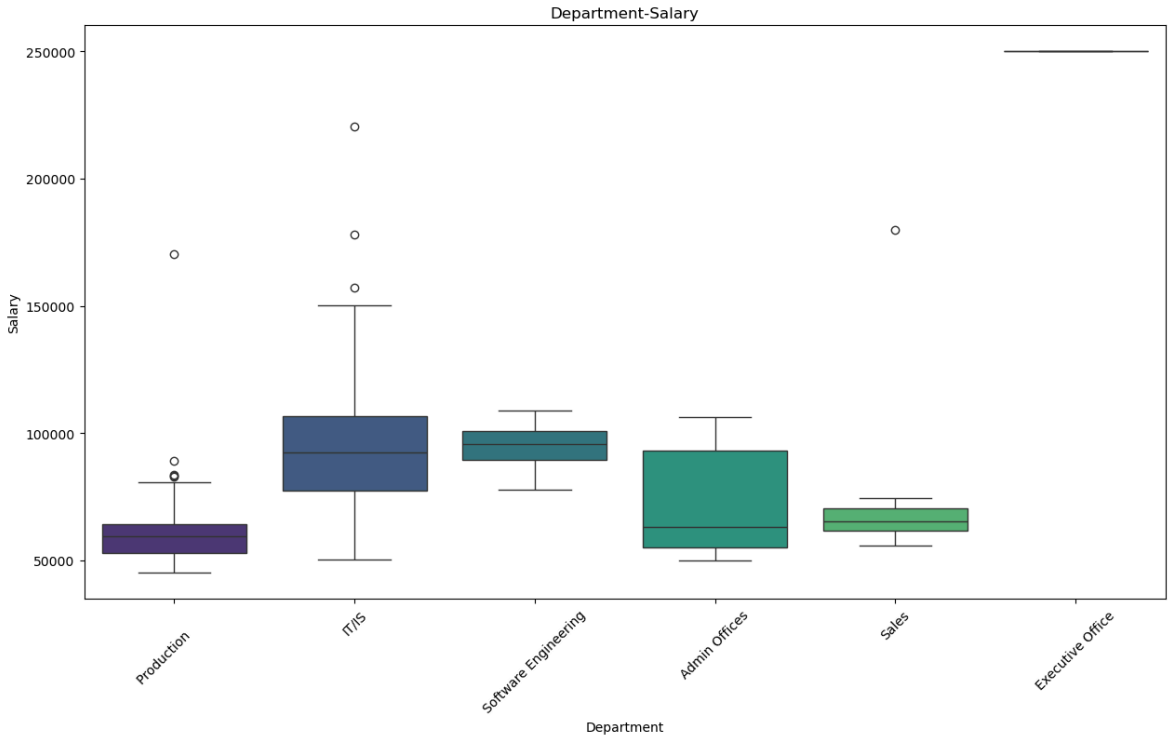


In [66]:
```
#insights
#general trend increases
#50-250 mostly the score
```

In [67]:
```
#outliers in salary in each department
```

In [68]:
```
#multivariate analysis
```

In [69]:
```
plt.figure(figsize = (15, 8))

sns.boxplot(x = 'Department', y ='Salary', data = df, palette = 'viridis')
plt.title("Department-Salary")

plt.xlabel("Department")
plt.ylabel("Salary")
plt.xticks(rotation = 45)
plt.show()
```



In [70]:
```
#insights
#executives are paid highest
#least salary is production
```

In [71]:
```
df.EngagementSurvey
```

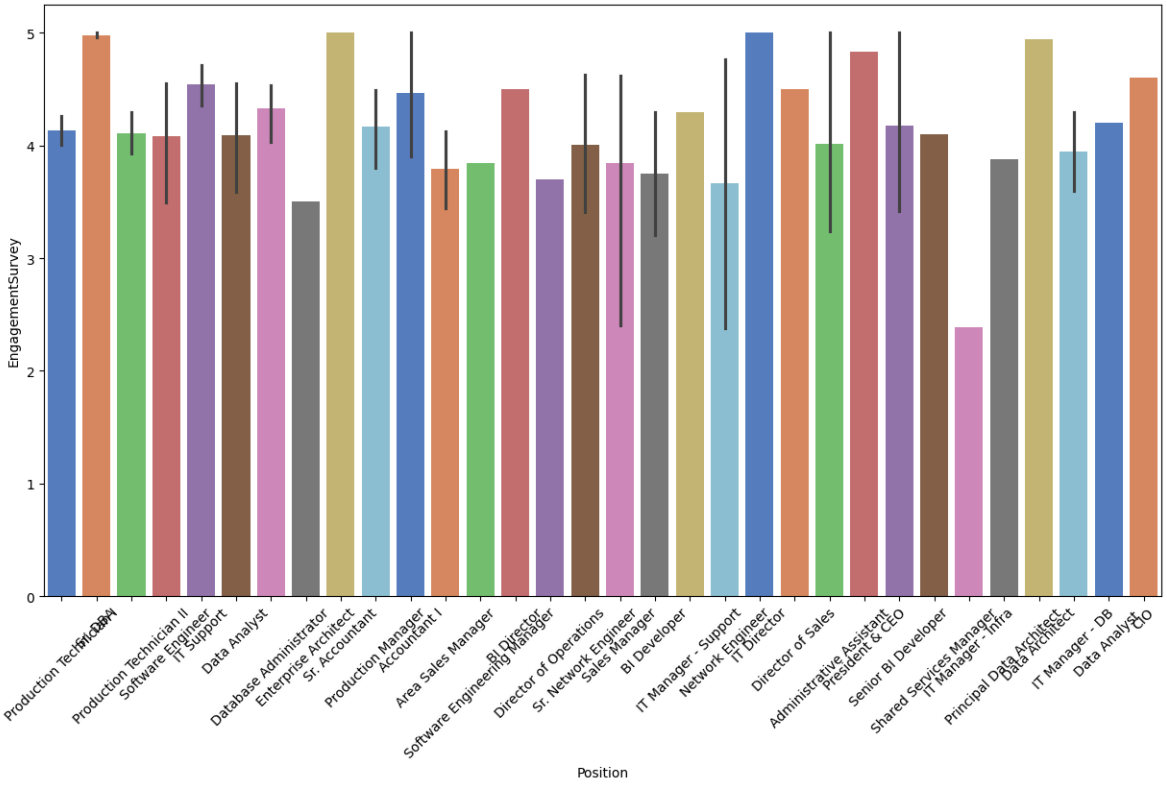
Out[71]:
```
0      4.60
1      4.96
2      3.02
3      4.84
4      5.00
       ...
306    4.07
307    3.20
308    4.60
309    5.00
310    4.50
Name: EngagementSurvey, Length: 311, dtype: float64
```

In [72]:
```
df.Position
```

Out[72]:
```
0       Production Technician I
1                      Sr. DBA
2      Production Technician II
3       Production Technician I
4       Production Technician I
                  ...
306    Production Technician II
307     Production Technician I
308                         CIO
309               Data Analyst
310     Production Technician I
Name: Position, Length: 311, dtype: object
```

In [73]:
```
plt.figure(figsize = (15, 8))
sns.barplot(x = 'Position', y='EngagementSurvey', data = df, palette = 'muted')
```

```
plt.xticks(rotation = 45)
plt.show()
```



In [74]:
```
#end of eda project by RAJNEESH SRIVASTAVA thankIng youu
```

In [1]:
```
pip install nbconvert
```

```
Defaulting to user installation because normal site-packages is not writeableNot
e: you may need to restart the kernel to use updated packages.

Requirement already satisfied: nbconvert in c:\programdata\anaconda3\lib\site-pac
kages (7.16.6)
Requirement already satisfied: beautifulsoup4 in c:\programdata\anaconda3\lib\sit
e-packages (from nbconvert) (4.12.3)
Requirement already satisfied: bleach!=5.0.0 in c:\programdata\anaconda3\lib\site
-packages (from bleach[css]!=5.0.0->nbconvert) (6.2.0)
Requirement already satisfied: defusedxml in c:\programdata\anaconda3\lib\site-pa
ckages (from nbconvert) (0.7.1)
Requirement already satisfied: jinja2>=3.0 in c:\programdata\anaconda3\lib\site-p
ackages (from nbconvert) (3.1.6)
Requirement already satisfied: jupyter-core>=4.7 in c:\programdata\anaconda3\lib
\site-packages (from nbconvert) (5.7.2)
Requirement already satisfied: jupyterlab-pygments in c:\programdata\anaconda3\li
b\site-packages (from nbconvert) (0.3.0)
Requirement already satisfied: markupsafe>=2.0 in c:\programdata\anaconda3\lib\si
te-packages (from nbconvert) (3.0.2)
Requirement already satisfied: mistune<4,>=2.0.3 in c:\programdata\anaconda3\lib
\site-packages (from nbconvert) (3.1.2)
Requirement already satisfied: nbclient>=0.5.0 in c:\programdata\anaconda3\lib\si
te-packages (from nbconvert) (0.10.2)
Requirement already satisfied: nbformat>=5.7 in c:\programdata\anaconda3\lib\site
-packages (from nbconvert) (5.10.4)
Requirement already satisfied: packaging in c:\programdata\anaconda3\lib\site-pac
kages (from nbconvert) (24.2)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\programdata\anaconda3\l
ib\site-packages (from nbconvert) (1.5.0)
Requirement already satisfied: pygments>=2.4.1 in c:\programdata\anaconda3\lib\si
te-packages (from nbconvert) (2.19.1)
Requirement already satisfied: traitlets>=5.1 in c:\programdata\anaconda3\lib\sit
e-packages (from nbconvert) (5.14.3)
Requirement already satisfied: webencodings in c:\programdata\anaconda3\lib\site-
packages (from bleach!=5.0.0->bleach[css]!=5.0.0->nbconvert) (0.5.1)
Requirement already satisfied: tinycss2<1.5,>=1.1.0 in c:\programdata\anaconda3\l
ib\site-packages (from bleach[css]!=5.0.0->nbconvert) (1.4.0)
Requirement already satisfied: platformdirs>=2.5 in c:\programdata\anaconda3\lib
\site-packages (from jupyter-core>=4.7->nbconvert) (4.3.7)
Requirement already satisfied: pywin32>=300 in c:\programdata\anaconda3\lib\site-
packages (from jupyter-core>=4.7->nbconvert) (308)
Requirement already satisfied: jupyter-client>=6.1.12 in c:\programdata\anaconda3
\lib\site-packages (from nbclient>=0.5.0->nbconvert) (8.6.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\programdata\anaconda3
\lib\site-packages (from jupyter-client>=6.1.12->nbclient>=0.5.0->nbconvert) (2.
9.0.post0)
Requirement already satisfied: pyzmq>=23.0 in c:\programdata\anaconda3\lib\site-p
ackages (from jupyter-client>=6.1.12->nbclient>=0.5.0->nbconvert) (26.2.0)
Requirement already satisfied: tornado>=6.2 in c:\programdata\anaconda3\lib\site-
packages (from jupyter-client>=6.1.12->nbclient>=0.5.0->nbconvert) (6.5.1)
Requirement already satisfied: fastjsonschema>=2.15 in c:\programdata\anaconda3\l
ib\site-packages (from nbformat>=5.7->nbconvert) (2.20.0)
Requirement already satisfied: jsonschema>=2.6 in c:\programdata\anaconda3\lib\si
te-packages (from nbformat>=5.7->nbconvert) (4.23.0)
Requirement already satisfied: attrs>=22.2.0 in c:\programdata\anaconda3\lib\site
-packages (from jsonschema>=2.6->nbformat>=5.7->nbconvert) (24.3.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in c:\program
data\anaconda3\lib\site-packages (from jsonschema>=2.6->nbformat>=5.7->nbconvert)
(2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in c:\programdata\anaconda3\li
```

```
b\site-packages (from jsonschema>=2.6->nbformat>=5.7->nbconvert) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in c:\programdata\anaconda3\lib\sit
e-packages (from jsonschema>=2.6->nbformat>=5.7->nbconvert) (0.22.3)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\site-pack
ages (from python-dateutil>=2.8.2->jupyter-client>=6.1.12->nbclient>=0.5.0->nbcon
vert) (1.17.0)
Requirement already satisfied: soupsieve>1.2 in c:\programdata\anaconda3\lib\site
-packages (from beautifulsoup4->nbconvert) (2.5)
```

In [2]:

```
  Cell In[2], line 1
    --to pdf
         ^
SyntaxError: invalid syntax
```

In [ ]: