2_EDA_HR_analytics_employees_engagement

December 8, 2024

1 HR Analytics - Employees Engagement

In this EDA project we will performing HR Analytics - Employees Engagement Analysis which is present in kaggle platform.In this EDA Project we will analyze and visualize our dataset.

- **The job role of a HR is not that easy as it seems from the outside**
- **The HR's have to actively participate in the recruitment process, helping employees with their issues, maintaining positive work environment, analysing the performance and efficiency, and many more**
- **Among all the job responsibilites of an HR, evaluating the performance and efficiency of the employees is considered the most difficult task**
- **The difficulty level of this task is directly proportional to the no. of employees who work under that particular HR**
- **Thus, to deal with this, we have come with an Exploratory Data Analysis (EDA) project**
- **Here, we'll be performing different analysis and visualizations using the Employees Enagagment Dataset to obtain some valuable insights**

[5]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt



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

[7]:
```
df=pd.read_csv('HRDataset_v14.csv')
df
```

[7]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	\
0	Adinolfi, Wilson K	10026	0	0	1	
1	Ait Sidi, Karthikeyan	10084	1	1	1	
2	Akinkuolie, Sarah	10196	1	1	0	

```
3                Alagbe,Trina  10088          1                 1         0
4             Anderson, Carol  10069          0                 2         0
..                         …    …        …               …        …
306             Woodson, Jason  10135          0                 0         1
307          Ybarra, Catherine  10301          0                 0         0
308            Zamora, Jennifer  10010          0                 0         0
309                Zhou, Julia  10043          0                 0         0
310              Zima, Colleen  10271          0                 4         0

     EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  Salary  …  \
0              1       5            4                       0   62506  …
1              5       3            3                       0  104437  …
2              5       5            3                       0   64955  …
3              1       5            3                       0   64991  …
4              5       5            3                       0   50825  …
..           …     …          …                     …      …  …
306            1       5            3                       0   65893  …
307            5       5            1                       0   48513  …
308            1       3            4                       0  220450  …
309            1       3            3                       0   89292  …
310            1       5            3                       0   45046  …

        ManagerName  ManagerID  RecruitmentSource PerformanceScore  \
0    Michael Albert       22.0           LinkedIn          Exceeds
1        Simon Roup        4.0             Indeed      Fully Meets
2    Kissy Sullivan       20.0           LinkedIn      Fully Meets
3      Elijiah Gray       16.0             Indeed      Fully Meets
4    Webster Butler       39.0      Google Search      Fully Meets
..              …        …                …               …
306  Kissy Sullivan       20.0           LinkedIn      Fully Meets
307  Brannon Miller       12.0      Google Search             PIP
308      Janet King        2.0  Employee Referral          Exceeds
309      Simon Roup        4.0  Employee Referral      Fully Meets
310   David Stanley       14.0           LinkedIn      Fully Meets

     EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
0                4.60               5                    0
1                4.96               3                    6
2                3.02               3                    0
3                4.84               5                    0
4                5.00               4                    0
..                …               …                    …
306              4.07               4                    0
307              3.20               2                    0
308              4.60               5                    6
309              5.00               3                    5
310              4.50               5                    0
```

```
    LastPerformanceReview_Date DaysLateLast30 Absences
0                    1/17/2019              0        1
1                    2/24/2016              0       17
2                    5/15/2012              0        3
3                     1/3/2019              0       15
4                     2/1/2016              0        2
..                         …            …       …
306                  2/28/2019              0       13
307                   9/2/2015              5        4
308                  2/21/2019              0       16
309                   2/1/2019              0       11
310                  1/30/2019              0        2

[311 rows x 36 columns]
```

[8]:
```
df.columns
```

[8]:
```
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')
```

[9]:
```
df.shape
```

[9]:
```
(311, 36)
```

311 rows 36 columns

[27]:
```
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           311 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                   311 non-null    object
 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(1), int64(16), object(19)
memory usage: 87.6+ KB
```

[28]:
```
df.dtypes
```

[28]:
```
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                      object
RecruitmentSource              object
PerformanceScore               object
EngagementSurvey              float64
EmpSatisfaction                 int64
SpecialProjectsCount            int64
LastPerformanceReview_Date     object
DaysLateLast30                  int64
Absences                        int64
dtype: object
```

cleaning the data

[29]:
```
df.isnull().sum()
```

[29]:
```
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
```

[30]:
```
df.shape
```

[30]:
```
(311, 36)
```

[31]:
```
df.fillna("0", inplace = True)
```

[32]:
```
df.isnull().sum()
```

[32]:
```
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
```

[33]:
```
df.duplicated().sum()
```

[33]:
```
np.int64(0)
```

[34]:
```
df.drop_duplicates(inplace=True)
```

[35]:
```
df.dtypes
```

[35]:
```
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                      object
RecruitmentSource              object
PerformanceScore               object
EngagementSurvey              float64
EmpSatisfaction                 int64
SpecialProjectsCount            int64
LastPerformanceReview_Date     object
DaysLateLast30                  int64
Absences                        int64
dtype: object
```

[36]:
```
df.head()
```

[36]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	…
0	Adinolfi, Wilson K	10026	0	0	1	
1	Ait Sidi, Karthikeyan	10084	1	1	1	
2	Akinkuolie, Sarah	10196	1	1	0	
3	Alagbe,Trina	10088	1	1	0	
4	Anderson, Carol	10069	0	2	0	

	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	…
0	1	5	4	0	62506	…
1	5	3	3	0	104437	…
2	5	5	3	0	64955	…
3	1	5	3	0	64991	…
4	5	5	3	0	50825	…

	ManagerName	ManagerID	RecruitmentSource	PerformanceScore
0	Michael Albert	22.0	LinkedIn	Exceeds
1	Simon Roup	4.0	Indeed	Fully Meets
2	Kissy Sullivan	20.0	LinkedIn	Fully Meets
3	Elijiah Gray	16.0	Indeed	Fully Meets
4	Webster Butler	39.0	Google Search	Fully Meets

	EngagementSurvey	EmpSatisfaction	SpecialProjectsCount
0	4.60	5	0
1	4.96	3	6
2	3.02	3	0
3	4.84	5	0
4	5.00	4	0

```
  LastPerformanceReview_Date DaysLateLast30 Absences
0                   1/17/2019              0         1
1                   2/24/2016              0        17
2                   5/15/2012              0         3
3                    1/3/2019              0        15
4                    2/1/2016              0         2

[5 rows x 36 columns]
```

EDA employees with highest salary » top 10 higest employees

[37]:
```
df.columns
```

[37]:
```
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')
```

[38]:
```
df.Salary.sort_values(ascending = False).head(10)
```

[38]:
```
150    250000
308    220450
131    180000
96     178000
55     170500
190    157000
240    150290
244    148999
243    140920
76     138888
Name: Salary, dtype: int64
```

Employees who needs the special attention Performance Improvement Plan(PIP)

[39]:
```
df.columns
```

[39]:
```
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')
```

[40]:
```
df['PerformanceScore'].unique()
```

[40]:
```
array(['Exceeds', 'Fully Meets', 'Needs Improvement', 'PIP'], dtype=object)
```

[41]:
```
df[df['PerformanceScore'] == 'PIP']
```

[41]:

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

	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	…	\
67	1	6	1	0	61568	…	
69	1	5	1	0	53189	…	
72	1	6	1	0	59231	…	
83	1	5	1	0	56847	…	
90	1	5	1	0	64057	…	
91	1	3	1	0	53366	…	
95	1	6	3	0	70187	…	
112	5	5	1	1	68898	…	
188	5	5	1	0	55800	…	
205	4	5	1	0	52674	…	
263	1	5	1	0	64021	…	
267	1	6	1	0	58273	…	
307	5	5	1	0	48513	…	

	ManagerName	ManagerID	RecruitmentSource	PerformanceScore	\
67	John Smith	17.0	Indeed	PIP	
69	Amy Dunn	11.0	Indeed	PIP	
72	John Smith	17.0	Website	PIP	
83	Michael Albert	22.0	Indeed	PIP	
90	Amy Dunn	11.0	Indeed	PIP	

```
91     Peter Monroe        7.0            LinkedIn              PIP
95    Lynn Daneault       21.0   Employee Referral             PIP
112  Brannon Miller       12.0  Diversity Job Fair             PIP
188  Brannon Miller       12.0            LinkedIn              PIP
205  Kissy Sullivan       20.0            LinkedIn              PIP
263  Brannon Miller       12.0              Indeed              PIP
267   Lynn Daneault       21.0             Website              PIP
307  Brannon Miller       12.0       Google Search              PIP

     EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
67               1.93               3                    0
69               1.12               2                    0
72               2.30               1                    0
83               3.00               1                    0
90               1.56               5                    0
91               1.20               3                    6
95               2.00               5                    0
112              3.00               3                    0
188              3.00               2                    0
205              2.33               2                    0
263              2.40               2                    1
267              1.81               2                    0
307              3.20               2                    0

    LastPerformanceReview_Date DaysLateLast30 Absences
67                   1/30/2019              6        5
69                   1/31/2019              4        9
72                   1/29/2019              2       17
83                   2/25/2019              2        5
90                    1/3/2019              6       15
91                    2/4/2019              3        2
95                   1/28/2019              4        7
112                   3/6/2011              3       10
188                  1/14/2013              6        6
205                   3/9/2018              6        3
263                  2/25/2019              6       20
267                  1/17/2019              3        5
307                   9/2/2015              5        4

[13 rows x 36 columns]
```

[42]:
```
people_pip = df[df['PerformanceScore'] == 'PIP'].Employee_Name
```

[43]:
```
len(people_pip)
```

[43]: 13

[44]:
```
people_pip
```

[44]:
```
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
```

No of absences

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

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

whether the employees are married/or not

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

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

insights » 187 employees are unmarries and 124 employees

[ ]:
```
df.columns
```

[ ]:
```
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')
```

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

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

     EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  Salary  …  \
1              5       3            3                       0  104437  …
6              1       4            3                       0   95660  …
9              1       3            3                       0   50178  …
12             1       3            4                       1   92328  …
18             4       3            3                       0  110000  …
..           …     …          …                       …     …  …
292            4       3            3                       0  113999  …
298            1       3            3                       0   84903  …
299            1       3            3                       0  107226  …
308            1       3            4                       0  220450  …
309            1       3            3                       0   89292  …
```

```
          ManagerName  ManagerID   RecruitmentSource PerformanceScore  \
1          Simon Roup        4.0              Indeed      Fully Meets
6     Alex Sweetwater       10.0           LinkedIn      Fully Meets
9        Peter Monroe        7.0              Indeed      Fully Meets
12         Simon Roup        4.0  Diversity Job Fair          Exceeds
18         Simon Roup        4.0       Google Search      Fully Meets
..                ...        ...                 ...              ...
292        Simon Roup        4.0   Employee Referral      Fully Meets
298  Brian Champaigne       13.0              Indeed      Fully Meets
299      Peter Monroe        7.0   Employee Referral      Fully Meets
308        Janet King        2.0   Employee Referral          Exceeds
309        Simon Roup        4.0   Employee Referral      Fully Meets

     EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
1                4.96               3                    6
6                3.04               3                    4
9                5.00               5                    6
12               4.28               4                    5
18               4.50               4                    5
..                ...             ...                  ...
292              4.33               3                    7
298              3.42               4                    7
299              4.20               4                    8
308              4.60               5                    6
309              5.00               3                    5

    LastPerformanceReview_Date DaysLateLast30 Absences
1                    2/24/2016              0       17
6                     1/2/2019              0       19
9                    2/18/2019              0       16
12                   2/25/2019              0        9
18                   1/15/2015              0        8
..                         ...            ...      ...
292                  2/15/2017              0        9
298                   1/4/2019              0       17
299                   2/5/2019              0        7
308                  2/21/2019              0       16
309                   2/1/2019              0       11

[70 rows x 36 columns]
```

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

[ ]:
```
61     8
299    8
243    7
254    7
```

```
25     7
      ..
126    0
127    0
128    0
129    0
310    0
Name: SpecialProjectsCount, Length: 311, dtype: int64
```

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

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

     EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  Salary  …  \
0              1       5            4                       0   62506  …
2              5       5            3                       0   64955  …
3              1       5            3                       0   64991  …
4              5       5            3                       0   50825  …
5              1       5            4                       0   57568  …
..           …     …          …                     …      …  …
304            5       5            4                       0   70507  …
305            1       5            3                       0   60446  …
306            1       5            3                       0   65893  …
307            5       5            1                       0   48513  …
310            1       5            3                       0   45046  …

        ManagerName  ManagerID RecruitmentSource PerformanceScore  \
0    Michael Albert       22.0          LinkedIn          Exceeds
2    Kissy Sullivan       20.0          LinkedIn      Fully Meets
3      Elijiah Gray       16.0            Indeed      Fully Meets
4    Webster Butler       39.0     Google Search      Fully Meets
5          Amy Dunn       11.0          LinkedIn          Exceeds
..              …        …                …              …
304  Brannon Miller       12.0          LinkedIn          Exceeds
305   David Stanley       14.0          LinkedIn      Fully Meets
306  Kissy Sullivan       20.0          LinkedIn      Fully Meets
```

```
307  Brannon Miller       12.0     Google Search              PIP
310   David Stanley       14.0          LinkedIn      Fully Meets

     EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
0                4.60               5                    0
2                3.02               3                    0
3                4.84               5                    0
4                5.00               4                    0
5                5.00               5                    0
..                ...             ...                  ...
304              5.00               3                    0
305              3.40               4                    0
306              4.07               4                    0
307              3.20               2                    0
310              4.50               5                    0

    LastPerformanceReview_Date DaysLateLast30 Absences
0                    1/17/2019              0        1
2                    5/15/2012              0        3
3                     1/3/2019              0       15
4                     2/1/2016              0        2
5                     1/7/2019              0       15
..                         ...            ...      ...
304                  1/19/2016              0        7
305                  2/21/2019              0       14
306                  2/28/2019              0       13
307                   9/2/2015              5        4
310                  1/30/2019              0        2

[241 rows x 36 columns]
```

insights out of 311 employes 70 employees have special project

Visualisation Highest salary vs lowest salary

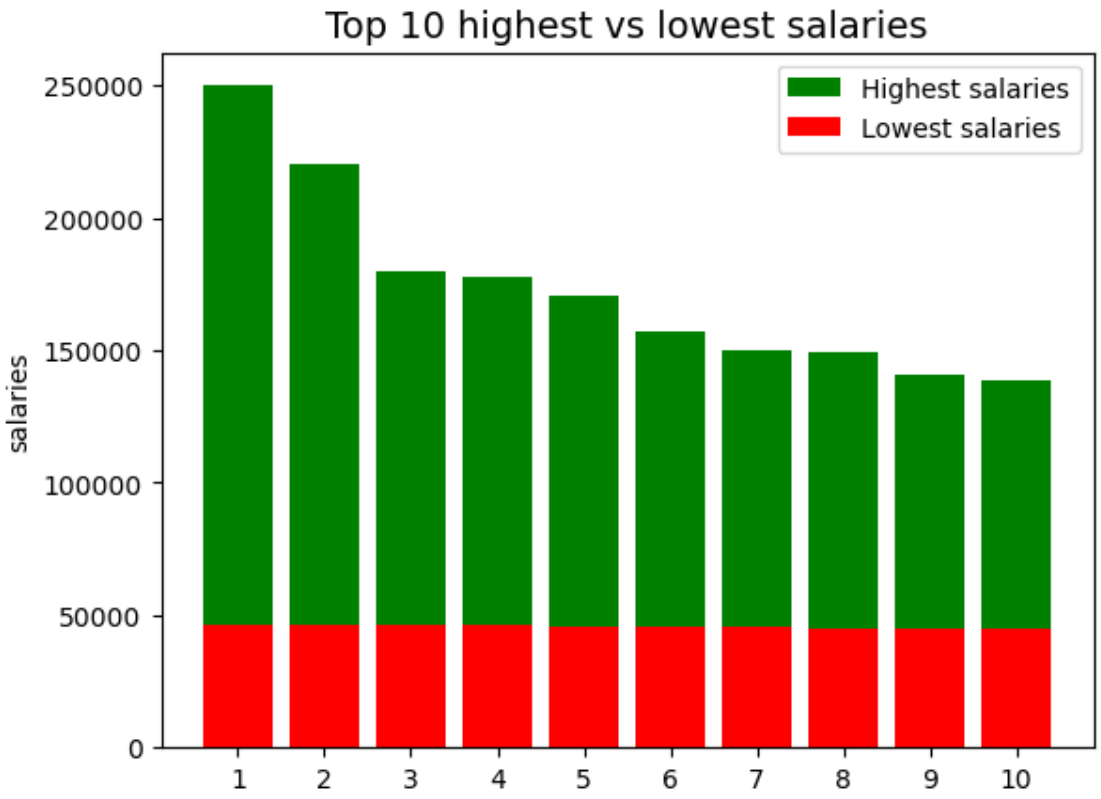
[ ]:
```
df['Salary'].sort_values(ascending = False).head(10)
```

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

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

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

[ ]:
```
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()
```



insight lowest salary are mostly in range highest salary varies

[ ]:
```
y
```

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

[ ]:
```
x
```

[ ]:
```
150     250000
308     220450
131     180000
96      178000
```

```
55     170500
190    157000
240    150290
244    148999
243    140920
76     138888
Name: Salary, dtype: int64
```

[ ]:
```
df.columns
```

[ ]:
```
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')
```

sources of recruitment

[ ]:
```
df['RecruitmentSource']
```

[ ]:
```
0               LinkedIn
1                 Indeed
2               LinkedIn
3                 Indeed
4          Google Search
             …
306             LinkedIn
307        Google Search
308    Employee Referral
309    Employee Referral
310             LinkedIn
Name: RecruitmentSource, Length: 311, dtype: object
```

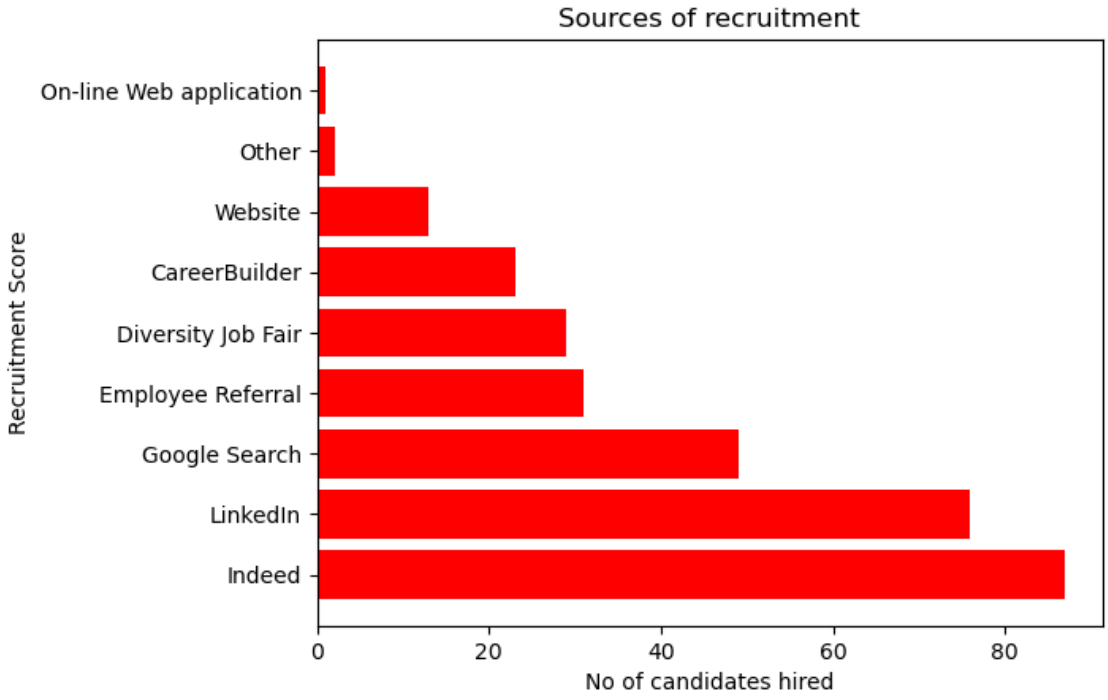
[ ]:
```
df['RecruitmentSource'].unique()
```

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

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

[ ]:
```
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
```

[ ]:
```
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()
```



Insights Indeed is the most common Indded, linkedin, google search

[ ]:
```
df.columns
```

[ ]:
```
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')
```

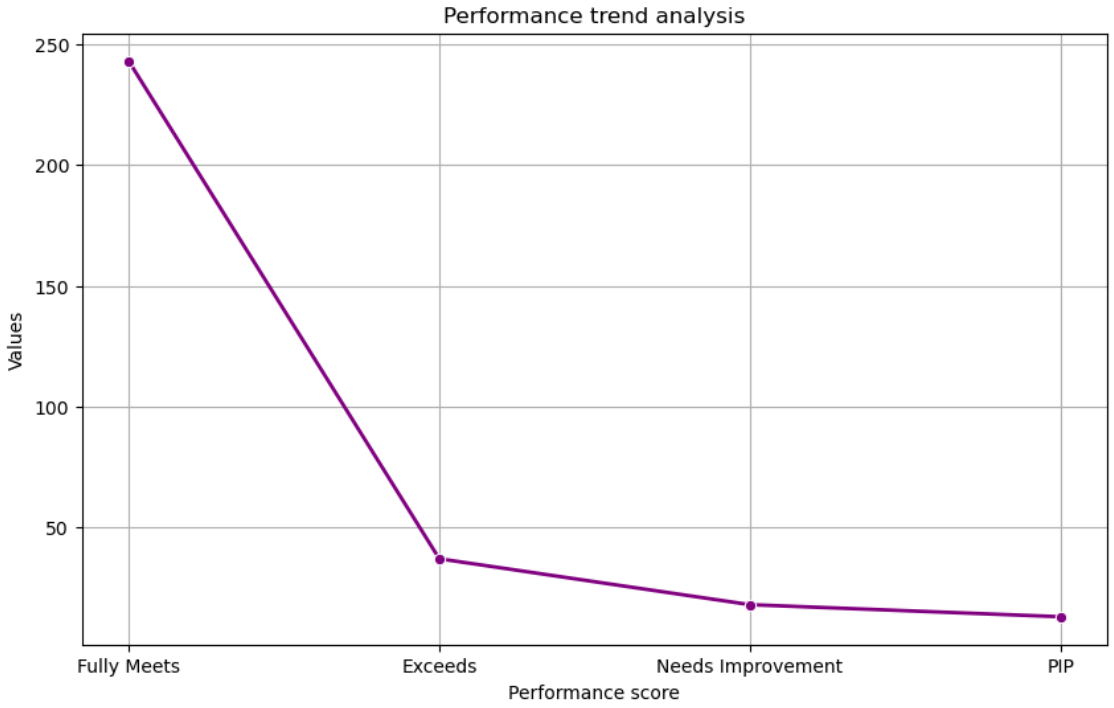
[ ]:
```
df['PerformanceScore']
```

[ ]:
```
0          Exceeds
1      Fully Meets
2      Fully Meets
3      Fully Meets
4      Fully Meets
          …
306    Fully Meets
307            PIP
308        Exceeds
309    Fully Meets
310    Fully Meets
Name: PerformanceScore, Length: 311, dtype: object
```

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

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

[ ]:
```
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()
```



insights general trend increases 50-250 mostly the score

[ ]:
```
df.columns
```

[ ]:
```
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')
```

[ ]:
```
df['EmpSatisfaction'] #scale of 1-5
```

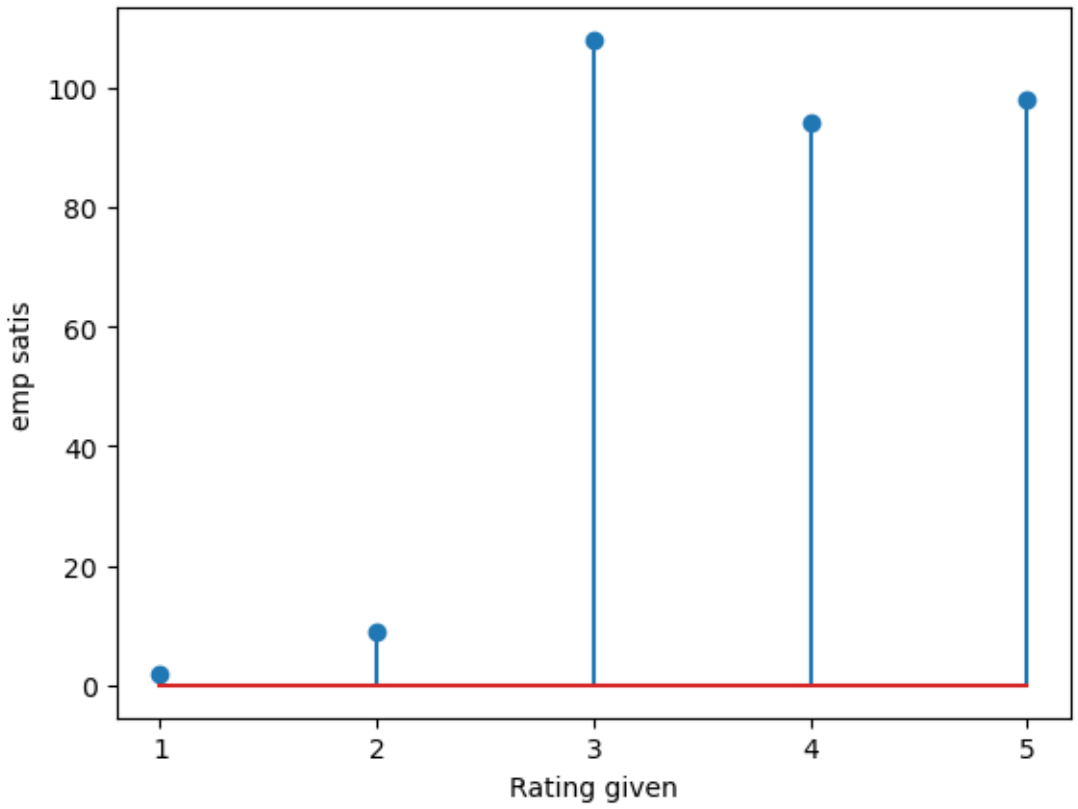
[ ]:
```
0      5
1      3
2      3
3      5
4      4
      ..
306    4
```

```
307    2
308    5
309    3
310    5
Name: EmpSatisfaction, Length: 311, dtype: int64
```

[ ]:
```
b = df['EmpSatisfaction'].value_counts()
b
```

[ ]:
```
EmpSatisfaction
3    108
5     98
4     94
2      9
1      2
Name: count, dtype: int64
```

[ ]:
```
plt.stem(b.index, b)
plt.ylabel("No of employees")
plt.xticks(b.index)
plt.xlabel("Rating given")
plt.ylabel("emp satis")
plt.show()
```



insights the most common rating 3

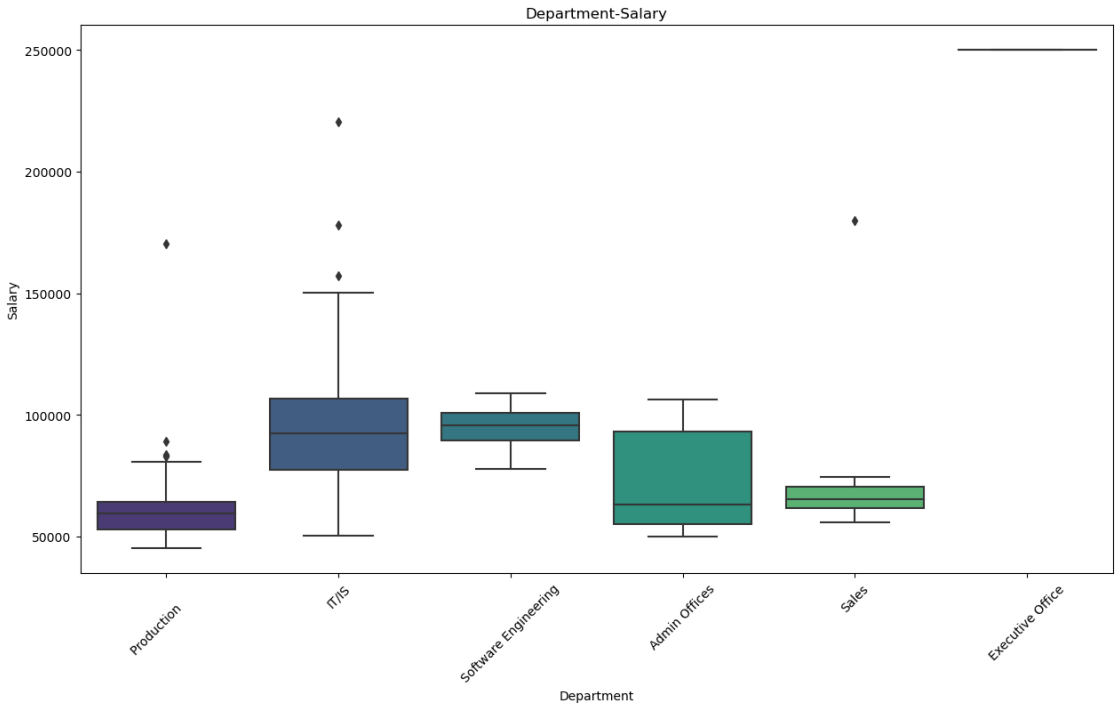
multi-variate analysuisdf

[ ]:
```
df.columns
```

[ ]:
```
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')
```

outliers in salary in each department

[ ]:
```
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()
```



insights executives are paid highest least salary is production

[ ]:
```
df.columns
```

[ ]:
```
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')
```

[ ]:
```
df.Position
```

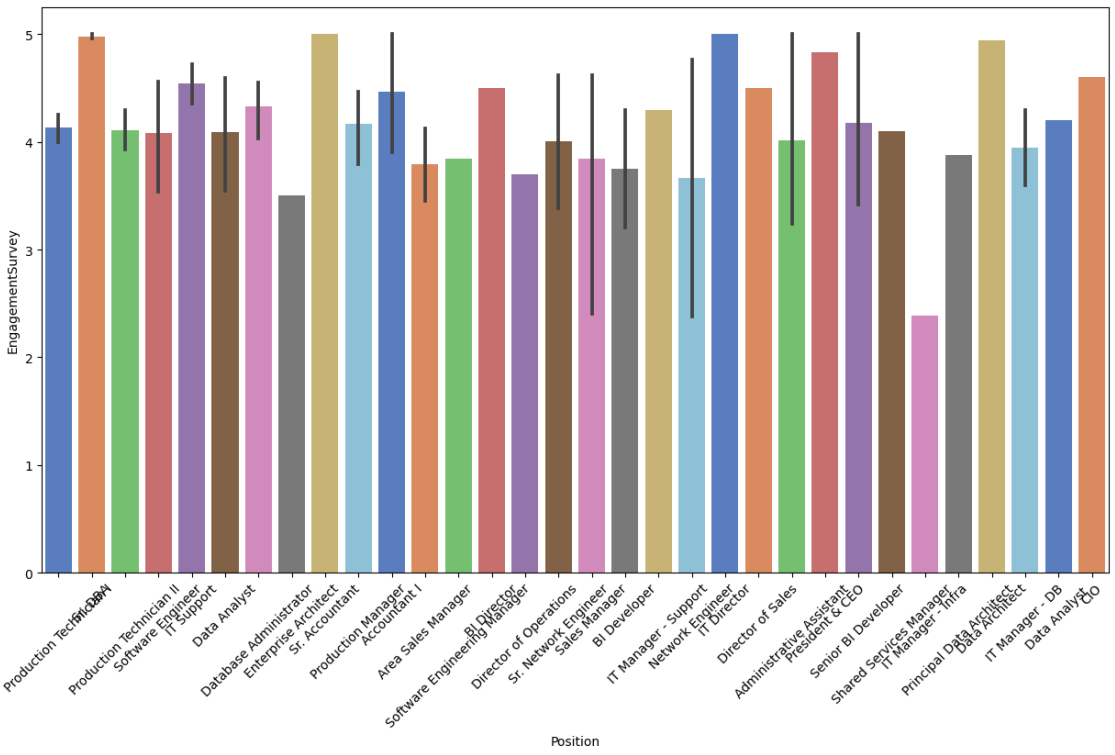
[ ]:
```
0       Production Technician I
1                     Sr. DBA
2      Production Technician II
3       Production Technician I
4       Production Technician I
                 ...
306    Production Technician II
307     Production Technician I
```

```
308                          CIO
309                Data Analyst
310     Production Technician I
Name: Position, Length: 311, dtype: object
```

[ ]:
```
df.EngagementSurvey
```

[ ]:
```
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
```

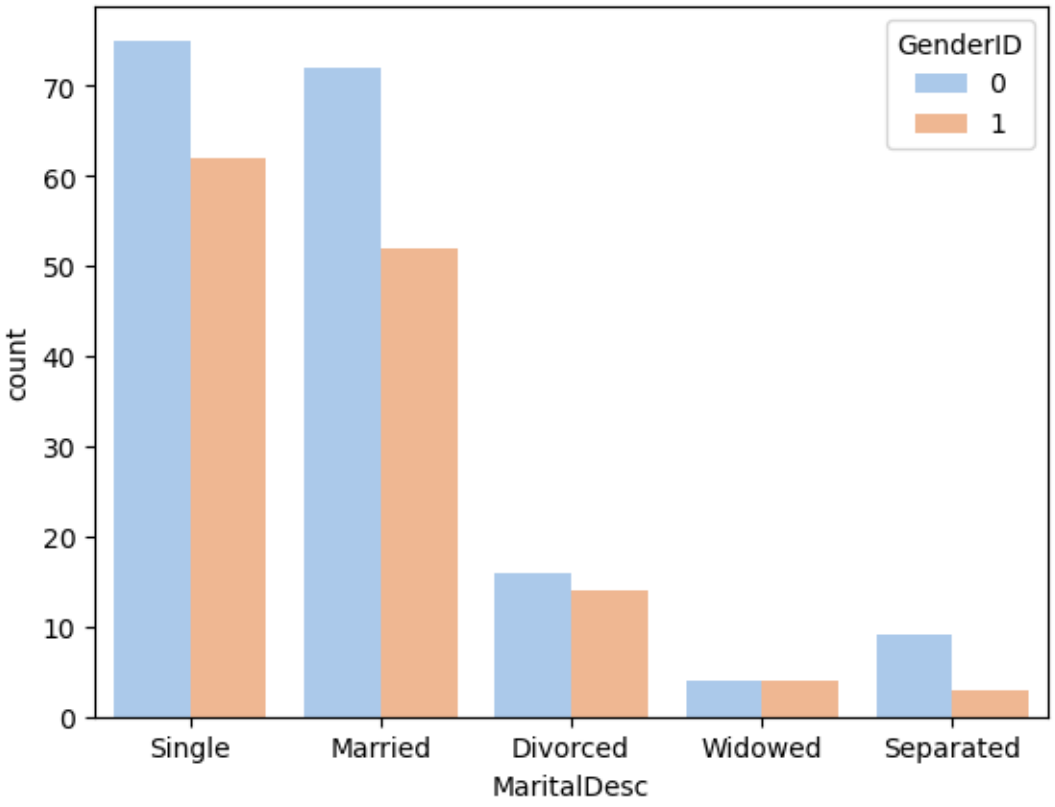
[ ]:
```
plt.figure(figsize = (15, 8))
sns.barplot(x = 'Position', y='EngagementSurvey', data = df, palette = 'muted')

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



marital status by gender

[ ]:
```
sns.countplot(x = 'MaritalDesc', hue = "GenderID", data = df, palette =␣
↪"pastel")
```

[ ]:
```
<Axes: xlabel='MaritalDesc', ylabel='count'>
```



[ ]:

[ ]:
```
df.groupby('Department')['EngagementSurvey'].mean()
```

[ ]:
```
Department
Admin Offices           4.393333
Executive Office        4.830000
IT/IS                   4.154000
Production              4.129569
Sales                   3.818710
Software Engineering    4.061818
Name: EngagementSurvey, dtype: float64
```

internal h/w

How many employees have been terminated from each position

```
[ ]: df.columns
```

```
[ ]: 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')
```

```
[ ]: df[df['Termd'] == 1].groupby('Position')['Employee_Name'].count()
```

```
[ ]: Position
     Administrative Assistant     2
     Area Sales Manager           4
     Data Analyst                 1
     Data Analyst                 1
     Database Administrator       3
     Enterprise Architect         1
     IT Manager - DB              1
     Network Engineer             1
     Principal Data Architect     1
     Production Manager           5
     Production Technician I     52
     Production Technician II    26
     Sales Manager                1
     Software Engineer            4
     Sr. DBA                      1
     Name: Employee_Name, dtype: int64
```

how many employees have been terminated for each reason

```
[ ]: df[df['Termd'] == 1].groupby('TermReason')['Employee_Name'].count()
```

```
[ ]: TermReason
     Another position                   20
     Fatal attraction                    1
     Learned that he is a gangster       1
     attendance                          7
     career change                       9
     gross misconduct                    1
     hours                               8
```

```
maternity leave - did not return     3
medical issues                       3
military                             4
more money                          11
no-call, no-show                     4
performance                          4
relocation out of area               5
retiring                             4
return to school                     5
unhappy                             14
Name: Employee_Name, dtype: int64
```

[ ]:
```
df['TermReason']
```

[ ]:
```
0      N/A-StillEmployed
1          career change
2                  hours
3      N/A-StillEmployed
4       return to school
              ...
306    N/A-StillEmployed
307     Another position
308    N/A-StillEmployed
309    N/A-StillEmployed
310    N/A-StillEmployed
Name: TermReason, Length: 311, dtype: object
```

What is the median salary of male and female employees

[ ]:
```
df.groupby('Sex')['Salary'].median()
```

[ ]:
```
Sex
F     62066.5
M     63353.0
Name: Salary, dtype: float64
```

what is the maximum no of absences taken by employees in each department

[ ]:
```
df.groupby('Department')['Absences'].max()
```

[ ]:
```
Department
Admin Offices           20
Executive Office        10
IT/IS                   20
Production              20
Sales                   20
Software Engineering    19
Name: Absences, dtype: int64
```

what is the total absences and average engaement survey score for each dept

```
[ ]: df.groupby('Department').agg({'Absences': 'sum', 'EngagementSurvey':'mean'})
```

```
[ ]:                        Absences  EngagementSurvey
     Department
     Admin Offices                78          4.393333
     Executive Office             10          4.830000
     IT/IS                       522          4.154000
     Production                 2120          4.129569
     Sales                       358          3.818710
     Software Engineering         96          4.061818
```

What is the total number of special projects and average absences for employees in each gender category?

internal homework: What is the maximum salary and minimum days late in the last 30 days for employees in each position? How many terminated employees were there in each department and what is the average employee satisfaction level among them? What is the earliest and latest date of hire for employees in each manager's team?

```
[ ]: df.groupby('Sex').agg({'SpecialProjectsCount': 'sum', 'Absences': 'mean'})
```

```
[ ]:      SpecialProjectsCount   Absences
     Sex
     F                     183  10.261364
     M                     196  10.207407
```

```
[ ]:
```