1.IMPORTIG LIBRARIES & LOADING DATA

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import scipy as sp
import plotly.express as px
```

```
#LOADING DATA
data=pd.read_csv('/content/cleanedhrdata1.csv')
```

```
#CHECK DATA TYPES BEFORE CORRELATION ANALYSIS
data.dtypes
```

```
Employee_Name                object
EmpID                         int64
MarriedID                     int64
MaritalStatusID               int64
GenderID                      int64
EmpStatusID                   int64
DeptID                        int64
PerfScoreID                   int64
FromDiversityJobFairID        int64
Salary                      float64
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
Age                         float64
YearofHire                    int64
dtype: object
```

2.UNDERSTANDING VARIABLES AND CORRELATIONS

CORRELATION MATRIX

```
#dataframe correlation
data.corr(method='spearman')
```

```
<ipython-input-15-0b1275de715e>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In
  data.corr(method='spearman')
```

	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversit
EmpID	1.000000	0.062241	-0.024763	0.021992	0.066473	0.118753	-0.722983	
MarriedID	0.062241	1.000000	0.422250	-0.015587	0.093639	-0.089919	-0.073122	
MaritalStatusID	-0.024763	0.422250	1.000000	-0.019591	0.146495	-0.013490	0.021657	
GenderID	0.021992	-0.015587	-0.019591	1.000000	-0.039438	-0.028786	-0.031596	
EmpStatusID	0.066473	0.093639	0.146495	-0.039438	1.000000	0.043512	-0.086539	
DeptID	0.118753	-0.089919	-0.013490	-0.028786	0.043512	1.000000	-0.087443	
PerfScoreID	-0.722983	-0.073122	0.021657	-0.031596	-0.086539	-0.087443	1.000000	
FromDiversityJobFairID	0.045787	-0.015476	0.052762	0.030913	0.194942	-0.094058	0.016454	
Salary	-0.102754	0.044431	-0.032349	0.072254	-0.105372	-0.364390	0.100897	
Termd	0.090994	0.072646	0.127540	-0.018322	0.919831	0.042068	-0.110940	
PositionID	-0.019589	-0.048855	-0.027123	-0.088424	0.243606	-0.095504	-0.005835	
Zip	0.003117	-0.042588	-0.082777	0.005403	-0.040226	0.264427	-0.034639	
ManagerID	0.111075	-0.134850	-0.028867	-0.034803	0.150747	0.589086	-0.072689	

```
#converting the categorical into numerical
df_num= data
for col_name in df_num.columns:
    if(df_num[col_name ].dtype == 'object'):
     df_num[col_name] = df_num [col_name].astype('category')
     df_num[col_name] = df_num[ col_name].cat.codes
df_num
print(df_num)
```

```
     Employee_Name  EmpID  MarriedID  MaritalStatusID  GenderID  EmpStatusID  \
0                0  10026          0                0         1            1
1                1  10084          1                1         1            5
2                2  10196          1                1         0            5
3                3  10088          1                1         0            1
4                4  10069          0                2         0            5
..             ...    ...        ...              ...       ...          ...
298            298  10135          0                0         1            1
299            299  10301          0                0         0            5
300            300  10010          0                0         0            1
301            301  10043          0                0         0            1
302            302  10271          0                4         0            1

     DeptID  PerfScoreID  FromDiversityJobFairID     Salary  ...  \
0         5            4                       0  11.043018  ...
1         3            3                       0  11.556339  ...
2         5            3                       0  11.081450  ...
3         5            3                       0  11.082004  ...
4         5            3                       0  10.836144  ...
..      ...          ...                     ...        ...  ...
298       5            3                       0  11.095787  ...
299       5            1                       0  10.789587  ...
300       3            4                       0  12.303426  ...
301       3            3                       0  11.399667  ...
302       5            3                       0  10.715439  ...

     RecruitmentSource  PerformanceScore  EngagementSurvey  EmpSatisfaction  \
0                    5                 0              4.60                5
1                    4                 1              4.96                3
2                    5                 1              3.02                3
3                    4                 1              4.84                5
4                    3                 1              5.00                4
..                 ...               ...               ...              ...
298                  5                 1              4.07                4
299                  3                 3              3.20                2
300                  2                 0              4.60                5
301                  2                 1              5.00                3
302                  5                 1              4.50                5

     SpecialProjectsCount  LastPerformanceReview_Date  DaysLateLast30  \
0                       0                         105               0
1                       6                          67               0
2                       0                          13               0
3                       0                          96               0
4                       0                          63               0
..                    ...                         ...             ...
298                     0                         136               0
299                     0                          58               5
300                     6                         131               0
301                     5                         118               0
302                     0                         116               0
```

```
     Absences   Age  YearofHire
0           1  40.0        2011
1          17  48.0        2015
2           3  35.0        2011
3          15  35.0        2008
4           2  34.0        2011
```

RELATIONSHIP BETWEEN VARIABLES

```
sns.pairplot(df_num)
```

STRENGTH OF CORRELATION

```
df_num.corr()
```

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfS
Employee_Name	1.000000	-0.002574	0.026928	0.090891	0.012252	0.108170	-0.025452	0
EmpID	-0.002574	1.000000	0.062158	-0.046993	0.021896	0.070305	0.110422	-0

Calculate the Pearson correlation coefficient between each pair of variables. The Pearson correlation coefficient measures the linear relationship between two variables.

GenderID	0.012252	0.021896	0.015587	0.018441	1.000000	0.033226	0.038250	0

```
df_num.corr(method='spearman')
```

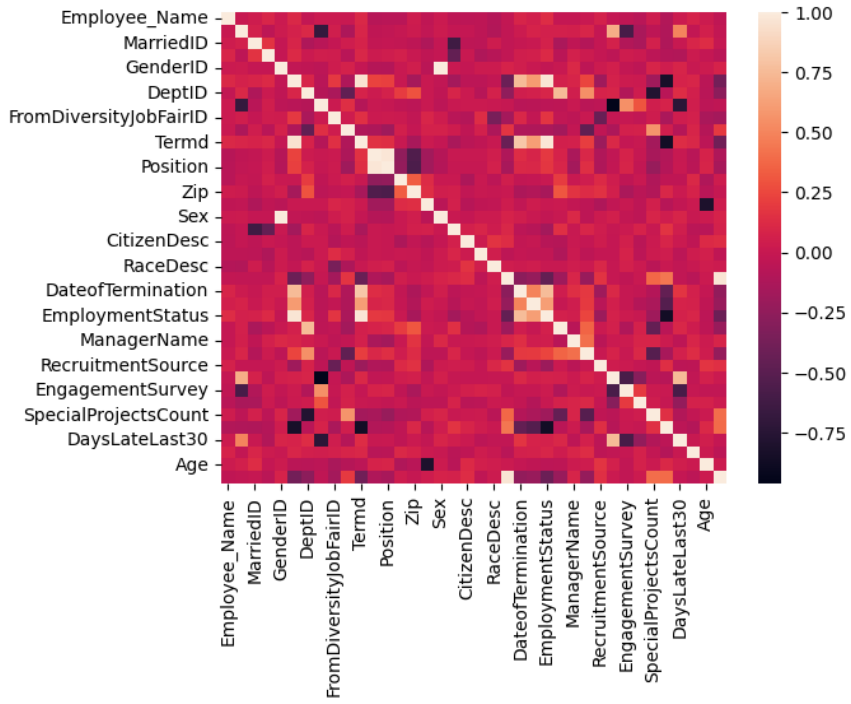
	Employee_Name	EmpID	MarriedID	MaritalStatusID	Ger
Employee_Name	1.000000	-0.002560	0.026928	0.087423	0.0
EmpID	-0.002560	1.000000	0.062241	-0.024763	0.0
MarriedID	0.026928	0.062241	1.000000	0.422250	-0.0
MaritalStatusID	0.087423	-0.024763	0.422250	1.000000	-0.0
GenderID	0.012252	0.021992	-0.015587	-0.019591	1.0
EmpStatusID	0.102911	0.066473	0.093639	0.146495	-0.0
DeptID	-0.018650	0.118753	-0.089919	-0.013490	-0.0
PerfScoreID	0.009879	-0.722983	-0.073122	0.021657	-0.0
FromDiversityJobFairID	0.014493	0.045787	-0.015476	0.052762	0.0
Salary	-0.011903	-0.102754	0.044431	-0.032349	0.0
Termd	0.105537	0.090994	0.072646	0.127540	-0.0
PositionID	-0.057656	-0.019589	-0.048855	-0.027123	-0.0
Position	-0.052463	-0.022967	-0.035058	-0.008281	-0.0
State	0.011368	0.000195	-0.001317	0.084100	0.0
Zip	0.050812	0.003117	-0.042588	-0.082777	0.0
DOB	-0.037680	-0.028296	-0.114020	-0.060534	-0.0
Sex	0.012252	0.021992	-0.015587	-0.019591	1.0
MaritalDesc	-0.049206	-0.016705	-0.615569	-0.749690	0.0
CitizenDesc	-0.058132	-0.019658	-0.046860	-0.010853	-0.0
HispanicLatino	-0.057483	-0.039976	-0.060450	-0.126284	0.0
RaceDesc	-0.097436	-0.103381	0.033375	-0.054119	0.0
DateofHire	-0.029248	-0.033599	-0.025589	-0.065657	0.0
DateofTermination	0.100355	0.067994	0.041486	0.101340	0.0
TermReason	0.034163	0.072450	0.010727	0.066624	0.0
EmploymentStatus	0.122832	0.072239	0.072922	0.139924	-0.0
Department	-0.016748	0.063844	-0.119748	-0.089490	-0.0
ManagerName	-0.010223	0.052814	-0.032991	0.043549	0.0
ManagerID	-0.009391	0.111075	-0.134850	-0.028867	-0.0
RecruitmentSource	0.065704	-0.000372	-0.031092	-0.050611	0.0
PerformanceScore	-0.007505	0.722474	0.073122	-0.021657	0.0
EngagementSurvey	-0.090935	-0.578412	-0.105975	-0.024465	0.0
EmpSatisfaction	-0.079889	-0.109929	-0.139928	-0.021905	-0.0
SpecialProjectsCount	0.028124	-0.046745	0.065632	-0.038756	0.0
LastPerformanceReview_Date	-0.103544	-0.071917	-0.094575	-0.098279	-0.0
DaysLateLast30	-0.016783	0.526484	0.012876	-0.061913	0.0
Absences	0.078593	-0.019605	0.097549	0.047923	-0.0
Age	0.035640	0.030502	0.111548	0.057660	0.0
YearofHire	-0.022426	-0.025518	-0.031589	-0.067657	0.0

38 rows × 38 columns

STRENGTH OF CORRELATION BY VISUALIZATION

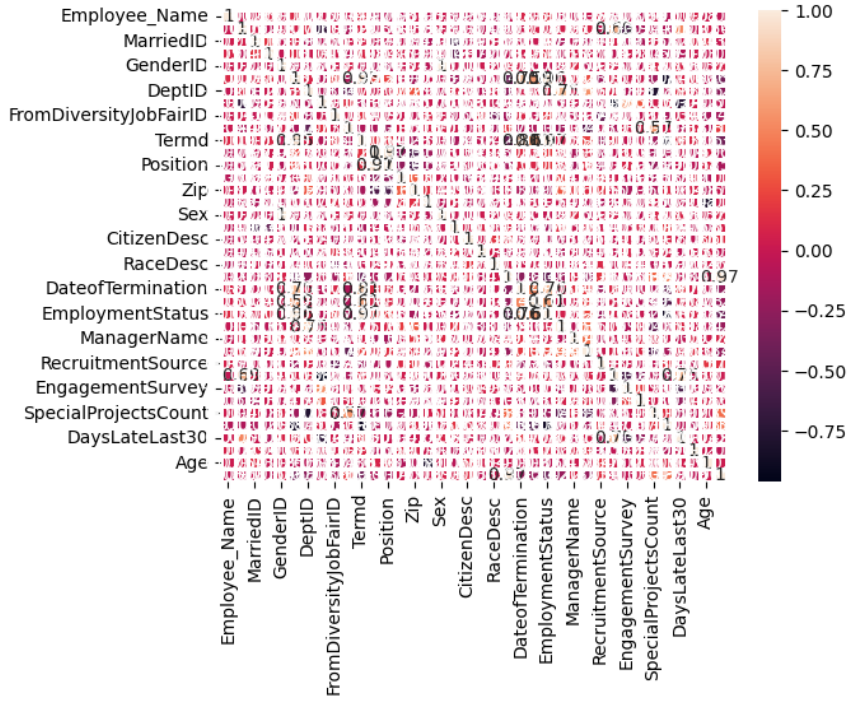
```
sns.heatmap(df_num.corr())
```

```
<Axes: >
```




```
sns.heatmap(df_num.corr(),linewidths=1,annot=True)
```

```
<Axes: >
```

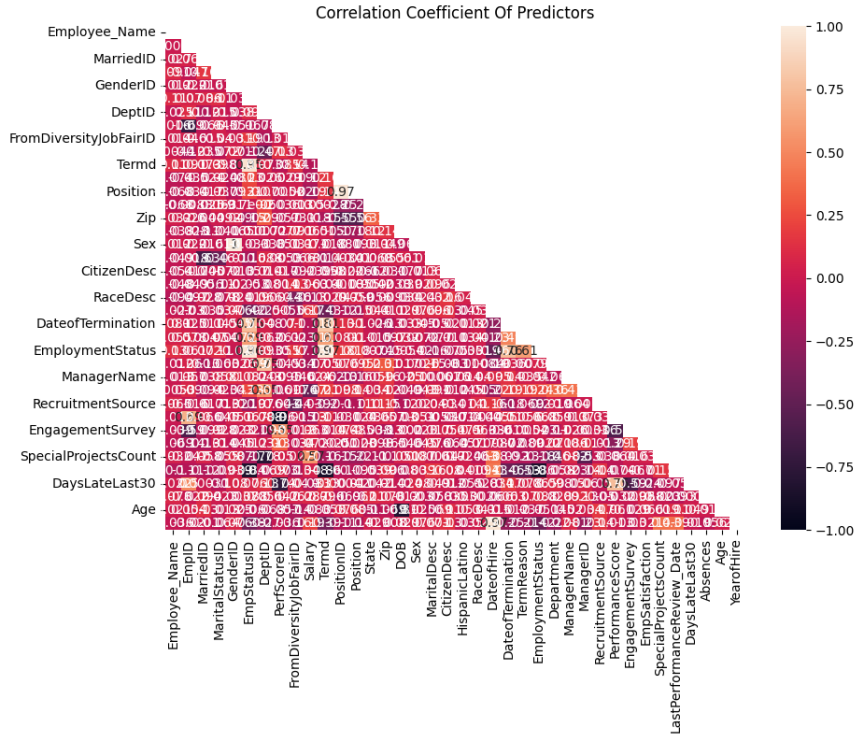



```
# set figure size
plt.figure(figsize=(10,7))

# Generate a mask to onlyshow the bottom triangle
mask = np.triu(np.ones_like(df_num.corr(), dtype=bool))

# generate heatmap
sns.heatmap(df_num.corr(), annot=True, mask=mask, vmin=-1, vmax=1)
```

```
plt.title('Correlation Coefficient Of Predictors')
plt.show()
```



PAIRS OF EACH VARIABLES

Identify the pairs of variables that are negatively correlated, positively correlated, or uncorrelated.

```
correlation_matrix=df_num.corr()
correlation_pairs=correlation_matrix.unstack()
correlation_pairs
```

```
Employee_Name  Employee_Name                1.000000
               EmpID                       -0.002574
               MarriedID                    0.026928
               MaritalStatusID              0.090891
               GenderID                     0.012252
                                              ...
YearofHire     LastPerformanceReview_Date   0.394156
               DaysLateLast30              -0.018770
               Absences                    -0.055828
               Age                          0.025356
               YearofHire                   1.000000
Length: 1444, dtype: float64
```

SORTING PAIRS

```
sorted_pairs=correlation_pairs.sort_values()
sorted_pairs
```

```
PerformanceScore            PerfScoreID                 -0.962189
PerfScoreID                 PerformanceScore            -0.962189
LastPerformanceReview_Date  EmploymentStatus            -0.864644
EmploymentStatus            LastPerformanceReview_Date  -0.864644
Termd                       LastPerformanceReview_Date  -0.863582
                                                          ...
```

```
DateofHire                  DateofHire                  1.000000
RaceDesc                    RaceDesc                    1.000000
HispanicLatino              HispanicLatino              1.000000
ManagerName                 ManagerName                 1.000000
YearofHire                  YearofHire                  1.000000
Length: 1444, dtype: float64
```

GROUPING SORTED PAIRS

Interpret the results of the Pearson correlation coefficient: values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values close to 0 indicate no correlation.

```
corr_matrix = df_num.corr()

positive_pairs = []
negative_pairs = []
no_correlation = []

for i in range(len(corr_matrix.columns)):
    for j in range(i+1, len(corr_matrix.columns)):
        col1 = corr_matrix.columns[i]
        col2 = corr_matrix.columns[j]
        correlation = corr_matrix.loc[col1, col2]
        if correlation > 0:
            positive_pairs.append((col1, col2, correlation))
        elif correlation < 0:
            negative_pairs.append((col1, col2, correlation))
        else:
            no_correlation.append((col1, col2, correlation))

df_positive = pd.DataFrame(positive_pairs, columns=['Column 1', 'Column 2', 'Correlation'])
df_negative = pd.DataFrame(negative_pairs, columns=['Column 1', 'Column 2', 'Correlation'])
df_no_correlation = pd.DataFrame(no_correlation, columns=['Column 1', 'Column 2', 'Correlation'])

print("Positive_pairs:")
print(df_positive)

print("Negative pairs:")
print(df_negative)


print("No correlation:")
print(df_no_correlation)
```

```
Positive_pairs:
                      Column 1         Column 2  Correlation
0                Employee_Name        MarriedID     0.026928
1                Employee_Name  MaritalStatusID     0.090891
2                Employee_Name         GenderID     0.012252
3                Employee_Name      EmpStatusID     0.108170
4                Employee_Name      PerfScoreID     0.015645
..                         ...              ...          ...
361  LastPerformanceReview_Date       YearofHire     0.394156
362             DaysLateLast30         Absences     0.000221
363             DaysLateLast30              Age     0.048946
364                   Absences              Age     0.011348
365                        Age       YearofHire     0.025356

[366 rows x 3 columns]
Negative pairs:
                      Column 1        Column 2  Correlation
0                Employee_Name           EmpID    -0.002574
1                Employee_Name          DeptID    -0.025452
2                Employee_Name      PositionID    -0.074049
3                Employee_Name        Position    -0.067812
4                Employee_Name           State    -0.067630
..                         ...             ...          ...
332        SpecialProjectsCount        Absences    -0.023328
333  LastPerformanceReview_Date  DaysLateLast30    -0.054385
334  LastPerformanceReview_Date        Absences    -0.092633
335             DaysLateLast30      YearofHire    -0.018770
336                   Absences      YearofHire    -0.055828

[337 rows x 3 columns]
No correlation:
Empty DataFrame
Columns: [Column 1, Column 2, Correlation]
Index: []
```

```
df_positive
```

	Column 1	Column 2	Correlation
0	Employee_Name	MarriedID	0.026928
1	Employee_Name	MaritalStatusID	0.090891
2	Employee_Name	GenderID	0.012252
3	Employee_Name	EmpStatusID	0.108170
4	Employee_Name	PerfScoreID	0.015645
...	...	...	...
361	LastPerformanceReview_Date	YearofHire	0.394156
362	DaysLateLast30	Absences	0.000221
363	DaysLateLast30	Age	0.048946
364	Absences	Age	0.011348
365	Age	YearofHire	0.025356

366 rows × 3 columns

```
df_negative
```

	Column 1	Column 2	Correlation
0	Employee_Name	EmpID	-0.002574
1	Employee_Name	DeptID	-0.025452
2	Employee_Name	PositionID	-0.074049
3	Employee_Name	Position	-0.067812
4	Employee_Name	State	-0.067630
...	...	...	...
332	SpecialProjectsCount	Absences	-0.023328
333	LastPerformanceReview_Date	DaysLateLast30	-0.054385
334	LastPerformanceReview_Date	Absences	-0.092633
335	DaysLateLast30	YearofHire	-0.018770

```
df_no_correlation
```

	Column 1	Column 2	Correlation

SORTED HIGH CORRELATED

```
high_corr=sorted_pairs[(sorted_pairs)>0.5]
high_corr
```

```
EmpID             DaysLateLast30      0.500112
DaysLateLast30    EmpID               0.500112
PerfScoreID       EngagementSurvey    0.549614
EngagementSurvey  PerfScoreID         0.549614
ManagerID         DeptID              0.550240
                                        ...
DateofHire        DateofHire          1.000000
RaceDesc          RaceDesc            1.000000
HispanicLatino    HispanicLatino      1.000000
ManagerName       ManagerName         1.000000
YearofHire        YearofHire          1.000000
Length: 76, dtype: float64
```