ИУ5-61Б Муханов Ержан

Рубежный контроль №1

Вариант 14 - 2 задача, 6 набор данных

Для студентов групп ИУ5-61Б, ИУ5Ц-81Б, ИУ5И-61Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния". Набор данных - Human Resources Data Set

Ввод [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns

df = pd.read_csv('data/HRDataset_v14.csv')
df.head(10)
```

Out[2]:

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfSc
0	Adinolfi, Wilson K	10026	0	0	1	1	5	
1	Ait Sidi, Karthikeyan	10084	1	1	1	5	3	
2	Akinkuolie, Sarah	10196	1	1	0	5	5	
3	Alagbe,Trina	10088	1	1	0	1	5	
4	Anderson, Carol	10069	0	2	0	5	5	
5	Anderson, Linda	10002	0	0	0	1	5	
6	Andreola, Colby	10194	0	0	0	1	4	
7	Athwal, Sam	10062	0	4	1	1	5	
8	Bachiochi, Linda	10114	0	0	0	3	5	
9	Bacong, Alejandro	10250	0	2	1	1	3	

10 rows × 36 columns

Ввод [3]:

```
df.describe
```

Out[3]:

```
<bound method NDFrame.describe of                Employee_Name  EmpID  Marr
iedID  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]>
```

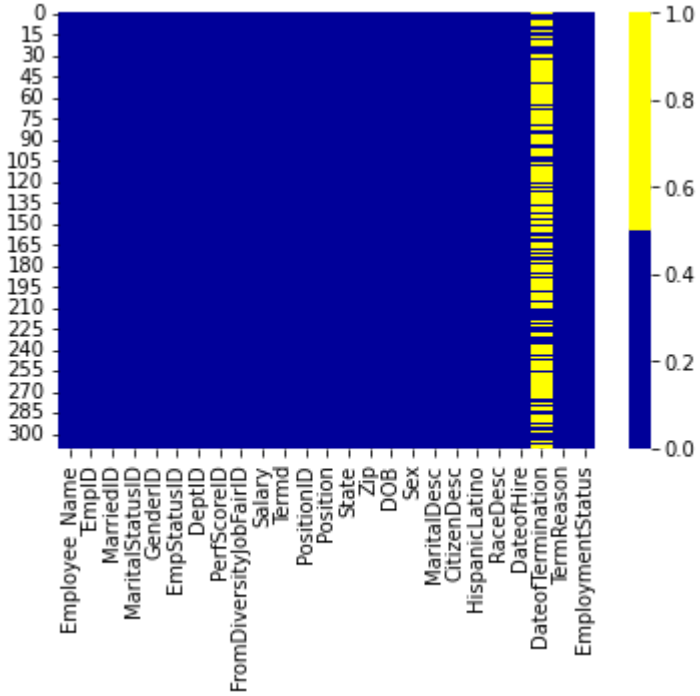
Обработка пропусков в данных

Ввод [14]:

```
cols = df.columns[:25]
# желтый - пропущенные данные, синий - не пропущенные
colours = ['#000099', '#ffff00']
sns.heatmap(df[cols].isnull(), cmap=sns.color_palette(colours))
```

Out[14]:

```
<AxesSubplot:>
```



Ввод [5]:

```
#Количество пустых ячеек в колонках:
df.isnull().sum()
```

Out[5]:

```
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
```

Ввод [6]:

```
#Типы данных в колонках:
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
```

Ввод [21]:

```
#Количество пустых числовых значений
num_cols = []
total_count = df.shape[0]
for col in df.columns:
    # Количество пустых значений
    temp_null_count = df[df[col].isnull()].shape[0]
    dt = str(df[col].dtype)
    if temp_null_count>0 and (dt=='float64' or dt=='int64'):
        num_cols.append(col)
        temp_perc = round((temp_null_count / total_count) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col,
```

```
Колонка ManagerID. Тип данных float64. Количество пустых значений 8, 2.57%.
```

Возьмем в качестве количественного признака признак EmpSatisfaction - показатель удовлетворенности работы Заменим пропуски на медианное значение:

Ввод [23]:

```
med = df['EmpSatisfaction'].median()
print(med)
df['EmpSatisfaction'] = df['EmpSatisfaction'].fillna(med)
```

```
4.0
```

Ввод [24]:

```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, round(pct_missing*100)))

for col in df.columns:
    missing = df[col].isnull()
    num_missing = np.sum(missing)
```

```
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 - 67%
TermReason - 0%
EmploymentStatus - 0%
Department - 0%
ManagerName - 0%
ManagerID - 3%
RecruitmentSource - 0%
PerformanceScore - 0%
EngagementSurvey - 0%
EmpSatisfaction - 0%
SpecialProjectsCount - 0%
LastPerformanceReview_Date - 0%
DaysLateLast30 - 0%
Absences - 0%
```

Ввод [25]:

```
print(df['EmpSatisfaction'])
```

```
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
```

В качестве категориального признака можно было бы взять, например, Position, но так как в этом столбце нет пропущенных значений.

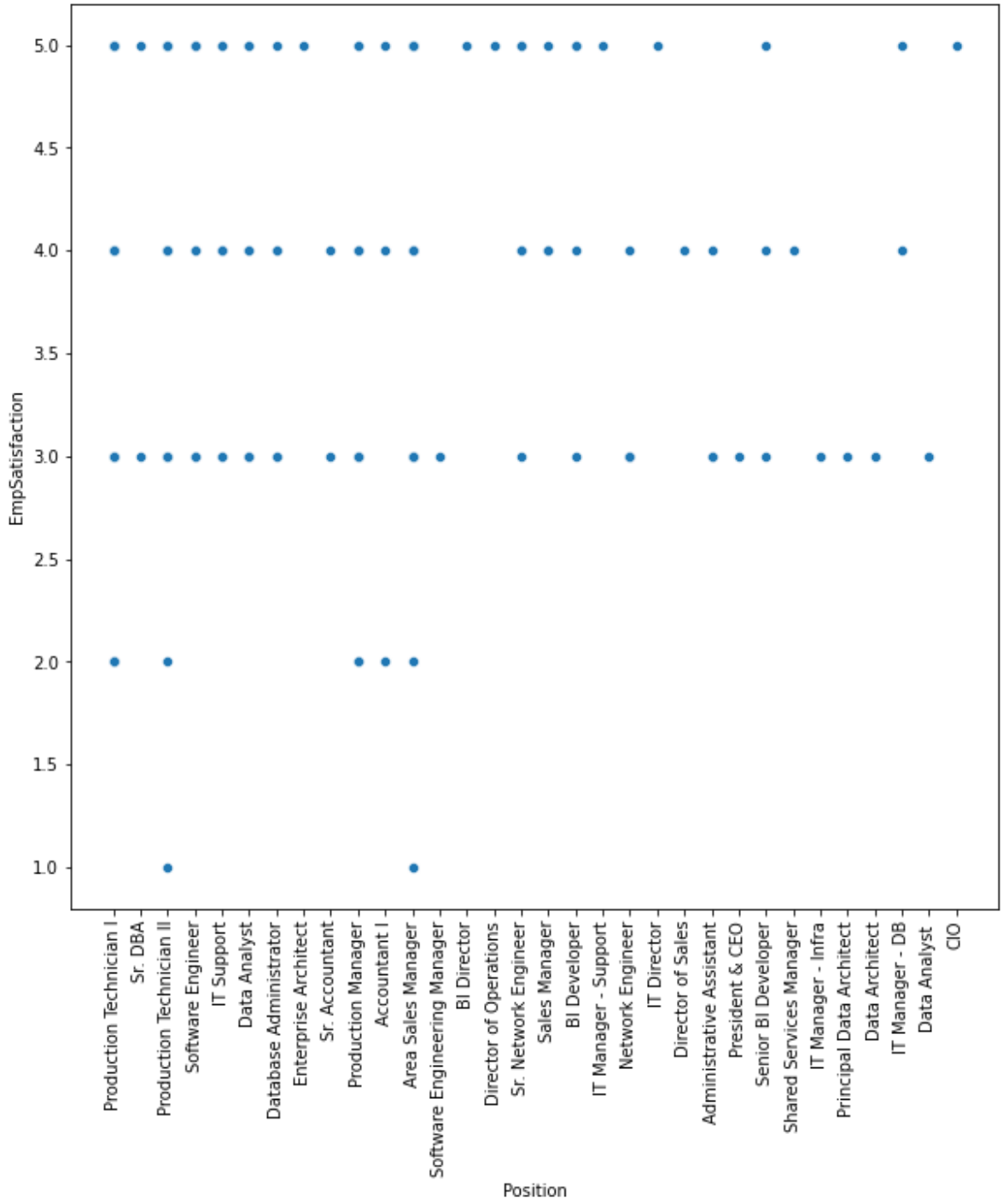
Диаграмма рассеивания

Ввод [26]:

```
fig, ax = plt.subplots(figsize=(10,10))
plt.xticks(rotation=90)
sns.scatterplot(ax=ax, x='Position', y='EmpSatisfaction', data=df)
```

Out[26]:

```
<AxesSubplot:xlabel='Position', ylabel='EmpSatisfaction'>
```



Ввод [ ]: