Ларионова Амина Павловна ИУ5-63Б

РК №1 по ТМО по теме "Технологии разведочного анализа и обработки данных"

14 вариант 2 задача 6 набор данных

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
data= pd.read_csv('HRDataset_v14.csv', sep=",")
```

```
data.shape
```

```
(311, 36)
```

```
data.dtypes
```

```
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
```

```
# проверим есть ли пропущенные значения
data.isnull().sum()
```

```
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
```

```
# Первые 5 строк датасета
data.head()
```

```
              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]
```

Обработка пропусков данных

Для категориального признака:

Так как пропуски данных встречаются в колонке DateofTermination (дата прекращения работы), то вариант с импьютацией не подходит, так как каждая дата- это уникальное значение. Поэтому воспользуемся стратегией заполнения пропущенных значений нулями.

```
data_new1 = data.fillna(0)
data_new1.head()
```

```
              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]
```

```
data_new1.isnull().sum()
```

```
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
```

Заметим,что числовой признак с пропусками ManagerID тоже заполнился нулями.

```
data_new1= data_new1 ['DateofTermination']
data_new1.head()
```

```
0            0
1    6/16/2016
2    9/24/2012
3            0
4     9/6/2016
Name: DateofTermination, dtype: object
```

Итого: все пропуски в данном признаке заполнились нулями.

Для числового признака ManagerID. Воспоьзуемся импьютацией.

```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 311
```

```
# Выберем числовые колонки с пропущенными значениями
# Цикл по колонкам датасета
num_cols = []
for col in data.columns:
    # Количество пустых значений
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[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, dt, temp_null_count, temp_perc))
```

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

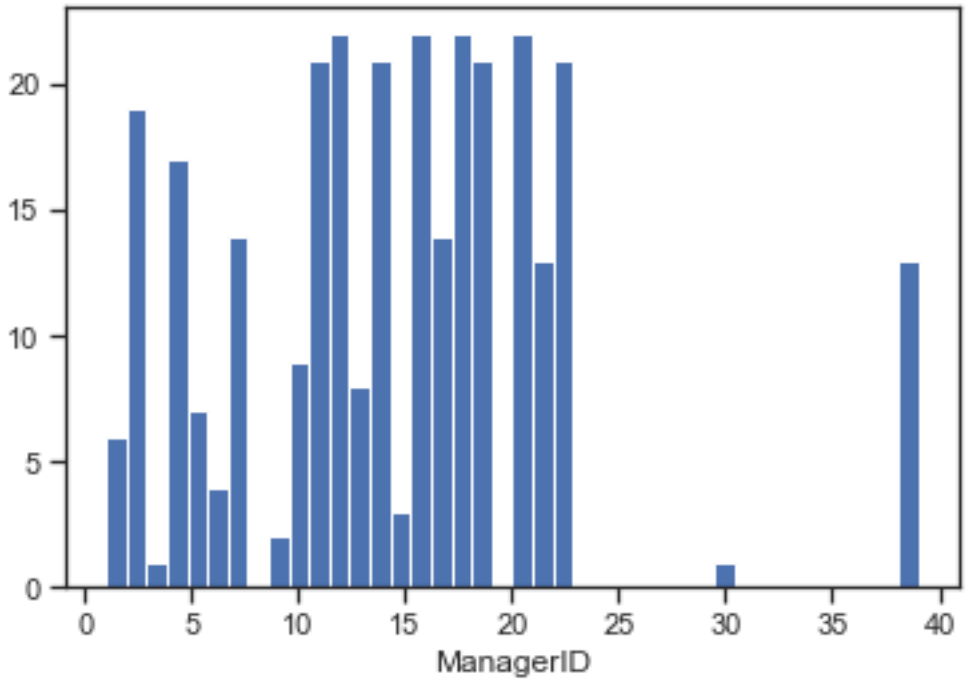
```
# Фильтр по колонкам с пропущенными значениями
data_num = data[num_cols]
data_num
```

```
     ManagerID
0         22.0
1          4.0
2         20.0
```

```
3         16.0
4         39.0
..         ...
306       20.0
307       12.0
308        2.0
309        4.0
310       14.0

[311 rows x 1 columns]
```

```
# Гистограмма по признакам
for col in data_num:
    plt.hist(data[col], 40)
    plt.xlabel(col)
    plt.show()
```



Будем использовать встроенные средства импьютации библиотеки scikit-learn.

```
data_num_ManagerID = data_num[['ManagerID']]
data_num_ManagerID.head()
```

```
   ManagerID
0       22.0
1        4.0
2       20.0
3       16.0
4       39.0
```

```
imp_num =
SimpleImputer(missing_values=np.nan,strategy='most_frequent')
data_num_imp = imp_num.fit_transform(data_num_ManagerID)
data_num_imp
```

```
array([[22.],
       [ 4.],
       [20.],
       [16.],
       [39.],
       [11.],
       [10.],
       [19.],
       [12.],
       [ 7.],
       [14.],
       [20.],
       [ 4.],
       [18.],
       [22.],
       [18.],
       [18.],
       [16.],
       [ 4.],
       [12.],
       [11.],
       [19.],
       [12.],
       [22.],
       [16.],
       [ 4.],
       [ 3.],
       [ 2.],
       [14.],
       [ 1.],
       [12.],
       [20.],
       [17.],
       [11.],
       [19.],
       [ 5.],
       [ 2.],
       [10.],
       [18.],
       [ 4.],
       [17.],
       [22.],
       [ 5.],
       [16.],
       [12.],
```

[21.],
[11.],
[19.],
[6.],
[12.],
[14.],
[12.],
[14.],
[20.],
[2.],
[2.],
[18.],
[4.],
[22.],
[7.],
[15.],
[7.],
[16.],
[20.],
[12.],
[39.],
[10.],
[17.],
[18.],
[11.],
[13.],
[19.],
[17.],
[12.],
[14.],
[7.],
[5.],
[21.],
[2.],
[19.],
[20.],
[18.],
[18.],
[22.],
[22.],
[16.],
[10.],
[16.],
[12.],
[39.],
[11.],
[7.],
[12.],
[11.],
[19.],

[21.],
[5.],
[9.],
[21.],
[6.],
[14.],
[18.],
[22.],
[16.],
[17.],
[39.],
[11.],
[21.],
[4.],
[7.],
[19.],
[12.],
[12.],
[7.],
[14.],
[14.],
[20.],
[20.],
[2.],
[18.],
[13.],
[17.],
[22.],
[18.],
[39.],
[11.],
[19.],
[18.],
[17.],
[30.],
[4.],
[2.],
[1.],
[19.],
[16.],
[12.],
[11.],
[2.],
[12.],
[14.],
[20.],
[19.],
[18.],
[22.],
[4.],

[12.],
[16.],
[15.],
[22.],
[21.],
[9.],
[39.],
[11.],
[19.],
[12.],
[17.],
[7.],
[2.],
[14.],
[14.],
[1.],
[20.],
[13.],
[20.],
[2.],
[16.],
[17.],
[2.],
[22.],
[18.],
[6.],
[22.],
[16.],
[16.],
[39.],
[11.],
[19.],
[12.],
[11.],
[12.],
[10.],
[14.],
[19.],
[20.],
[21.],
[18.],
[22.],
[2.],
[12.],
[14.],
[5.],
[20.],
[18.],
[7.],
[16.],

[22.],
[20.],
[13.],
[39.],
[11.],
[19.],
[12.],
[21.],
[14.],
[16.],
[20.],
[39.],
[21.],
[18.],
[22.],
[17.],
[16.],
[10.],
[4.],
[39.],
[11.],
[11.],
[19.],
[2.],
[12.],
[4.],
[12.],
[14.],
[21.],
[20.],
[18.],
[20.],
[2.],
[13.],
[22.],
[17.],
[16.],
[12.],
[14.],
[20.],
[11.],
[19.],
[12.],
[4.],
[4.],
[13.],
[14.],
[20.],
[5.],
[5.],

[10.],
[18.],
[22.],
[18.],
[4.],
[16.],
[12.],
[7.],
[11.],
[4.],
[1.],
[22.],
[16.],
[15.],
[1.],
[19.],
[6.],
[7.],
[12.],
[2.],
[14.],
[20.],
[21.],
[1.],
[18.],
[17.],
[2.],
[22.],
[16.],
[10.],
[13.],
[39.],
[16.],
[21.],
[11.],
[39.],
[12.],
[21.],
[14.],
[7.],
[10.],
[11.],
[7.],
[17.],
[20.],
[2.],
[17.],
[4.],
[18.],
[22.],

```
       [19.],
       [ 2.],
       [16.],
       [13.],
       [ 7.],
       [39.],
       [11.],
       [19.],
       [19.],
       [12.],
       [14.],
       [20.],
       [12.],
       [ 2.],
       [ 4.],
       [14.]])
```

```
np.unique(data_num_imp)
```

```
array([ 1.,  2.,  3.,  4.,  5.,  6.,  7.,  9., 10., 11., 12., 13.,
14.,
       15., 16., 17., 18., 19., 20., 21., 22., 30., 39.])
```

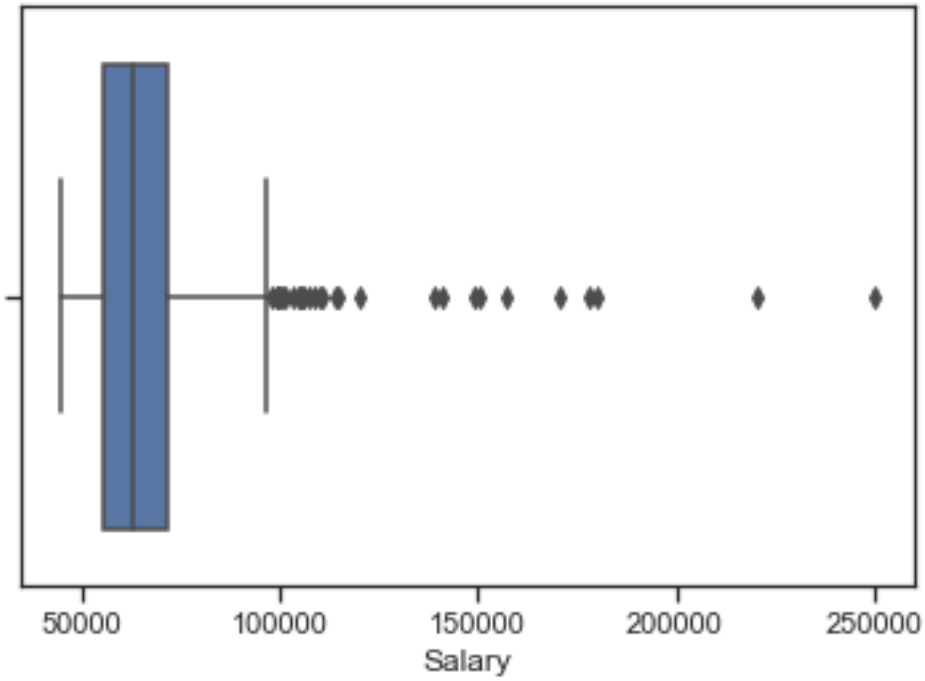
Все пропуски в числовом признаке заполнились самым частым значением,которое встречалось в этом признаке (12).

Дополнительное задание.Для произвольной колонки данных построить график "Ящик с усами (boxplot). Данный график отображает одномерное распределение вероятности.

Для признака Salary. По горизонтали.

```
sns.boxplot(x=data['Salary'])
```

```
<AxesSubplot:xlabel='Salary'>
```



Для признака Position.По веритикали.

```
# По вертикали
sns.boxplot(y=data['PositionID'])
```

```
<AxesSubplot:ylabel='PositionID'>
```