Московский Государственный Технический Университет им. Н. Э. Баумана

Факультет «Информатика и Системы управления»

Кафедра «Автоматизированные системы обработки информации и управления»

Дисциплина «Технологии машинного обучения»

Отчёт по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных.»

Выполнил:

Студент группы ИУ5ц-83Б

Костников И.А.

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

1 Цель работы

Изучение различных методов визуализация данных.

2 Краткое описание

Построение основных графиков, входящих в этап разведочного анализа данных. **Дополнительная информация представлена во втором файле**

3 Текст программы

Текст программы представлена во втором файле

4 Экранные формы с примерами выполнения программы.

```
[62]
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
[64]
Depart = pd.read_csv('Departament.csv', sep=";")
StCoun = pd.read_csv('Student_Counceling.csv', sep=";")
StPerf = pd.read_csv('Student_Performance.csv', sep=";")
Employee = pd.read_csv('Employee.csv', sep=";")
```

```
[65]
print("Размер таблицы Departament: ", Depart.shape)
print("Размер таблицы Student_Counceling: ", StCoun.shape)
print("Размер таблицы Student_Performance: ", StPerf.shape)
print("Размер таблицы Employee: ", Employee.shape)
```

```
Размер таблицы Departament:  (40, 3)
Размер таблицы Student_Counceling:  (3819, 5)
Размер таблицы Student_Performance:  (209611, 6)
Размер таблицы Employee:  (998, 4)
```

```
[66]
print("Departament:")
Depart.head()
```

```
Departament:
```

	Department_ID	Department_Name	DOE
0	IDEPT1096	School of Management	07.03.2000
1	IDEPT1142	Centre for Aerospace Systems Design and Engine...	25.07.1966
2	IDEPT1367	Computer Centre (CC)	05.04.2001
3	IDEPT1378	Industrial Design Centre	16.02.1993
4	IDEPT1388	Centre for Policy Studies (CPS)	05.01.1999

```
[70]
print('Количество пустых ячеек в таблице Department:')
for col in Depart.columns:
    temp_null_count = Depart[Depart[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
Количество пустых ячеек в таблице Department:
Department_ID - 0
Department_Name - 0
DOE - 0
```

```
[71]
print('Количество пустых ячеек в таблице Student_Counceling:')
for col in StCoun.columns:
    temp_null_count = StCoun[StCoun[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
Количество пустых ячеек в таблице Student_Counceling:
Student_ID - 0
DOA - 0
DOB - 0
Department_Choices - 0
Department_Admission - 0
```

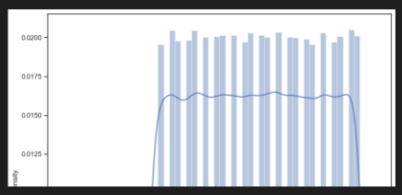
```
[72]
print('Количество пустых ячеек в таблице Student_Performance:')
for col in StPerf.columns:
    temp_null_count = StPerf[StPerf[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
Количество пустых ячеек в таблице Student_Performance:
ID - 0
Student_ID - 0
Semster_Name - 0
Paper_ID - 0
Paper_Name - 0
Marks - 0
```

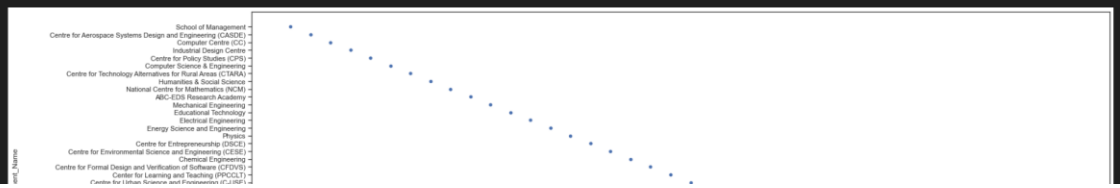
Визуальное исследование

```
[78]
fig, ax = plt.subplots(figsize=(10,10))
print('Распределение успеваемости студентов во всех вузах:')
sns.distplot(StPerf['Marks'])
```

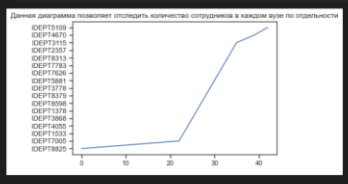
```
Распределение успеваемости студентов во всех вузах:

<AxesSubplot:xlabel='Marks', ylabel='Density'>
```



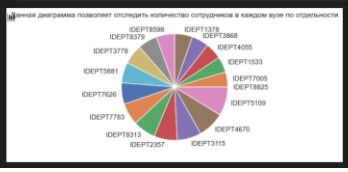
```
[79]
fig, ax = plt.subplots(figsize=(25,10))
sns.scatterplot(ax=ax, x='DOE', y='Department_Name', data=Depart)

<AxesSubplot:xlabel='DOE', ylabel='Department_Name'>
```



```
[82]
plt.plot(ox,oy)
plt.title('Данная диаграмма позволяет отследить количество сотрудников в каждом вузе по отдельности')
plt.show()
```



```
[91]
plt.pie(ox, labels=oy)
plt.title('Данная диаграмма позволяет отследить количество сотрудников в каждом вузе по отдельности')
plt.show()
```



5 Вывод

В данной лабораторной работе я научился работать библиотекой Pandas. Загружать и работать с базой данных