Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5-24М
вань хао

Москва — 2022 г.

1. Цель лабораторной работы

Изучение различных методов визуализация данных и создание истории на основе данных.

2. Задание

2.1 Выбрать набор данных (датасет).

2.2 Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.

На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.

Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.

Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.

История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

2.3 Сформировать отчет и разместить его в своем репозитории на github.

3. Ход выполнения работы

Контекст

QS World University Rankings - это ежегодная публикация глобальных рейтингов университетов компанией Quacquarelli Symonds. Рейтинг QS получает одобрение Международной группы экспертов по ранжированию

(IREG) и считается одним из трех наиболее читаемых университетских рейтингов в мире. QS публикует свои рейтинги университетов в партнерстве с компанией Elsevier.

Содержание

Этот набор данных содержит данные по университетам с 2017 по 2022 год. Он содержит в общей сложности 15 характеристик.

университет - название университета

год - год составления рейтинга

rank_display - ранг, присвоенный университету

score - рейтинг университета, основанный на шести ключевых метриках, упомянутых выше

ссылка - ссылка на страницу профиля университета на сайте QS

страна - страна, в которой находится университет

город - город, в котором находится университет

регион - континент, в котором расположен университет

логотип - ссылка на логотип университета

тип - тип университета (государственный или частный)

research_output - качество исследований в университете

studentfacultyratio - количество студентов, приходящихся на один факультет

international_students - количество иностранных студентов, обучающихся в университете

размер - размер университета с точки зрения площади

faculty_count - количество преподавателей или академического персонала в университете

3.2. Основные характеристики набора данных

Импортируйте соответствующую библиотеку и передайте данные:

```
]: import pandas as pd
   import seaborn as sns
   import matplotlib.pyplot as plt
   %matplotlib inline
```

```
]: lab1=pd.read_csv("/Users/mohao/Downloads/ммо/lab11.csv")
```

```
]: lab1.head()
```

Отображение данных с помощью команды .head()

```
lab1.head()
```

	university	year	rank_display	score	link	country	city	region	
0	Massachusetts Institute of Technology (MIT)	2017	1	100.0	https://www.topuniversities.com/universities/m...	United States	Cambridge	North America	https://www.topuniversi
1	Stanford University	2017	2	98.7	https://www.topuniversities.com/universities/s...	United States	Stanford	North America	https://www.topuniversi
2	Harvard University	2017	3	98.3	https://www.topuniversities.com/universities/h...	United States	Cambridge	North America	https://www.topuniversi
3	University of Cambridge	2017	4	97.2	https://www.topuniversities.com/universities/u...	United Kingdom	Cambridge	Europe	https://www.topuniversi
4	California Institute of Technology (Caltech)	2017	5	96.9	https://www.topuniversities.com/universities/c...	United States	Pasadena	North America	https://www.topuniversi

Отображение типов данных с помощью команды .info()

```
lab1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6482 entries, 0 to 6481
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   university              6482 non-null   object
 1   year                    6482 non-null   int64
 2   rank_display            6414 non-null   object
 3   score                   2820 non-null   float64
 4   link                    6482 non-null   object
 5   country                 6482 non-null   object
 6   city                    6304 non-null   object
 7   region                  6482 non-null   object
 8   logo                    6482 non-null   object
 9   type                    6470 non-null   object
 10  research_output         6480 non-null   object
 11  student_faculty_ratio   6407 non-null   float64
 12  international_students  6318 non-null   object
 13  size                    6480 non-null   object
 14  faculty_count           6404 non-null   object
dtypes: float64(2), int64(1), object(12)
memory usage: 759.7+ KB
```

Отбросьте ненужные столбцы

```
lab1.drop(["link","logo"],axis=1)
```

	university	year	rank_display	score	country	city	region	type	research_output	student_faculty_ratio	internatior
0	Massachusetts Institute of Technology (MIT)	2017	1	100.0	United States	Cambridge	North America	Private	Very High	4.0	
1	Stanford University	2017	2	98.7	United States	Stanford	North America	Private	Very High	3.0	
2	Harvard University	2017	3	98.3	United States	Cambridge	North America	Private	Very High	5.0	
3	University of Cambridge	2017	4	97.2	United Kingdom	Cambridge	Europe	Public	Very high	4.0	
4	California Institute of Technology (Caltech)	2017	5	96.9	United States	Pasadena	North America	Private	Very High	2.0	

Предварительная обработка расширенных данных для последующего анализа и построения графиков

Вакантные значения для статистических данных и обработки

```
lab1.isnull().sum()
```

```
university                   0
year                         0
rank_display                68
score                     3662
link                         0
country                      0
city                       178
region                       0
logo                         0
type                        12
research_output              2
student_faculty_ratio       75
international_students     164
size                         2
faculty_count               78
dtype: int64
```

```
lab1.isnull().melt()
```

	variable	value
0	university	False
1	university	False
2	university	False
3	university	False
4	university	False
...	...	...
97225	faculty_count	False
97226	faculty_count	False
97227	faculty_count	False
97228	faculty_count	False
97229	faculty_count	False

[8]:
```
sns.displot(data=lab1.isnull().melt(),y='variable',hue='value',multiple='fill',aspect=2)
plt.title('missing values')
```

[8]: Text(0.5, 1.0, 'missing values')

Поскольку в последующем анализе предполагается анализировать только отображение города и ранга, обрабатываются только эти два столбца

```
[10]: lab1.loc[:,'city'].fillna(value='unkonw',inplace=True)
```

```
[11]: lab1.loc[:,'rank_display'].fillna(value='out of range',inplace=True)
```

```
[12]: lab1.isnull().sum()
```

```
[12]: university                   0
      year                         0
      rank_display                 0
      score                     3662
      link                         0
      country                      0
      city                         0
      region                       0
      logo                         0
      type                        12
      research_output              2
      student_faculty_ratio       75
      international_students     164
      size                         2
      faculty_count               78
      dtype: int64
```

3.3. Визуальное исследование датасета

```
[13]: sns.displot(lab1['score'])
```

```
[13]: <seaborn.axisgrid.FacetGrid at 0x7faab87d8880>
```

Рисование значков, связанных с дробями

Для того чтобы понять взаимосвязь между количеством преподавателей и оценками студентов, нам необходимо проанализировать обе категории, которые в исходных данных существуют как "объект", поэтому сначала нам необходимо изменить тип

```
4]: lab1.loc[:,'faculty_count']=lab1.loc[:,'faculty_count'].str.replace(',', '').astype(float)
    lab1.loc[:,'student_faculty_ratio'].astype(float)
    lab1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6482 entries, 0 to 6481
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   university              6482 non-null   object
 1   year                    6482 non-null   int64
 2   rank_display            6482 non-null   object
 3   score                   2820 non-null   float64
 4   link                    6482 non-null   object
 5   country                 6482 non-null   object
 6   city                    6482 non-null   object
 7   region                  6482 non-null   object
 8   logo                    6482 non-null   object
 9   type                    6470 non-null   object
 10  research_output         6480 non-null   object
 11  student_faculty_ratio   6407 non-null   float64
 12  international_students  6318 non-null   object
 13  size                    6480 non-null   object
 14  faculty_count           6404 non-null   float64
dtypes: float64(3), int64(1), object(11)
memory usage: 759.7+ KB
```

Количество преподавателей и их оценки, а также общее изменение баллов от года к году отображаются на карте

```
5]: sns.jointplot(x='student_faculty_ratio',y='faculty_count',data=lab1,kind="hex")
    sns.catplot(x="year", y="score", kind="box", data=lab1)
```

[26]: <seaborn.axisgrid.FacetGrid at 0x7faaea8191f0>

Как видно, нет определенной связи между успеваемостью студентов и количеством преподавателей, а после 2018 года наблюдается постепенное снижение показателей баллов в рейтинге qs каждого университета.

Для того чтобы более четко увидеть влияние между каждым фактором, нам необходимо вызвать sns.hetmap для корреляционного анализа

```
import re
```

```
lab1['international_students'] = lab1.international_students.str.replace(r'\W+','').astype(float)
```

```
/var/folders/9l/v94s42394sv1b6r8jrdc1w980000gn/T/ipykernel_94439/472355414.py:1: FutureWarning: The default value o
f regex will change from True to False in a future version.
  lab1['international_students'] = lab1.international_students.str.replace(r'\W+','').astype(float)
```

```
lab1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6482 entries, 0 to 6481
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   university              6482 non-null   object
 1   year                    6482 non-null   int64
 2   rank_display            6482 non-null   object
 3   score                   2820 non-null   float64
 4   link                    6482 non-null   object
 5   country                 6482 non-null   object
 6   city                    6482 non-null   object
 7   region                  6482 non-null   object
 8   logo                    6482 non-null   object
 9   type                    6470 non-null   object
 10  research_output         6480 non-null   object
 11  student_faculty_ratio   6407 non-null   float64
 12  international_students  6318 non-null   float64
 13  size                    6480 non-null   object
 14  faculty_count           6404 non-null   float64
dtypes: float64(4), int64(1), object(10)
```

19]:
```
sns.heatmap(lab1.corr(),annot=True)
plt.title('Correlation Matrix')
```

19]: Text(0.5, 1.0, 'Correlation Matrix')

Видно, что именно количество сотрудников оказывает наибольшее влияние на результаты, что соответствует здравому смыслу, в то время как количество иностранных студентов также в некоторой степени влияет на результаты, а результаты отдельных учеников не так сильно связаны с результатами рейтинга школы.

Чтобы тщательно изучить влияние количества сотрудников на количество иностранных студентов, была построена линейная регрессия этих двух переменных

```
[20]: sns.set_style('white')
      sns.lmplot(data=lab1,x='international_students',y='faculty_count',
                 aspect=2,height=8)
      plt.title('Plot of the linear regression between number of faculty and number of  international students', fontsize=
                 color="green")
      plt.xlabel('Number of International Students')
      plt.ylabel('Number of Faculty')
```

```
[20]: Text(5.924999999999997, 0.5, 'Number of Faculty')
```

В то же время, данные по школам в таблице разделены на два типа: государственные и общественные школы, и мы хотели бы изучить между этими двумя типами школ и количеством иностранных

```
sns.set_style('dark')
plt.figure(figsize=(10,12))
sns.barplot(
    data=lab1,x="year", y='international_students', hue='type',palette='viridis')
plt.legend(bbox_to_anchor=(1, 1), loc='upper right')
plt.title('Number of international students in private and public universities from 2017-2022', fontsize=20,
          color="purple")
plt.ylabel('Number of International Students')
plt.xlabel('Type of University')
```

```
Text(0.5, 0, 'Type of University')
```

Наконец, я хотел бы рассмотреть, сколько университетов из 100 лучших в мире существует в разных странах, т.е. в каких странах находится большинство университетов из 100 лучших в мире

[22]:
```
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import plotly.express as px
```

[23]:
```
graph=lab1[['international_students','country']].groupby('country').sum().sort_values(by='international_students',as
```

[27]:
```
sns.set_style('dark')
plt.figure(figsize=(10,12))
sns.barplot(
    data=graph, x='international_students', y=graph.index, palette='viridis')
plt.title('Top 100 Countries of Choice for International Students for the Years 2017-2022', fontsize=15,
          color="purple")
plt.ylabel('Countries', fontsize=20)
plt.xlabel('count', fontsize=20)
```

[27]: Text(0.5, 0, 'count')

[27]: Text(0.5, 0, 'count')

Как видите, Великобритания, США и Австралия - страны с наибольшей долей университетов из топ-100 в мире.

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets