**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**



**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

ФАКУЛЬТЕТ Информатика, искусственный интеллект и системы управления

КАФЕДРА Системы обработки информации и управления

**Методические указания к лабораторным работам по курсу «Машинное обучение»**

**Лабораторная работа №1**

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

Выполнила Зозуля О.А. (ИУ5-25М)

Москва, 2023 г.

# ЗАДАНИЕ

* Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов [здесь.](https://github.com/ugapanyuk/courses_current/wiki/DSLIST)

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

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

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. 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.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

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

# РЕШЕНИЕ

# Лабораторная работа №1  
Датасет: самые популярные игры платформы Steam  
#%% md  
## Импортирование необходимых библиотек  
#%%  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import matplotlib.colors as colors  
import os  
from pathlib import Path  
#%% md  
## Некоторые дополнительные функции  
#%%  
def data\_to\_int(raw\_data: pd.DataFrame, columns: list):  
 for column in columns:  
 raw\_data[column] = raw\_data[column].astype(int)  
  
def remove\_commas(raw\_data: pd.DataFrame, columns: list):  
 for column in columns:  
 raw\_data[column] = raw\_data[column].apply(lambda i: ''.join(i.split(",")))  
#%% md  
## Открытие файла с данными  
#%%  
file\_name = 'data.csv'  
file\_path = f'{Path(os.getcwd()).parent}//data//{file\_name}'  
data = pd.read\_csv(file\_path)  
  
remove\_commas(data, ["Current", "All\_time peak", "24th Peak"])  
data\_to\_int(data, ["No.", "Current", "All\_time peak", "24th Peak"])  
  
print(data)  
#%% md  
## Построение графиков  
#%%  
figure, axis = plt.subplots(1, 3, figsize=(15, 5))  
  
axis[0].plot(  
 data["No."],  
 data["Current"],  
 color="Red",  
 label='За всё время'  
)  
axis[0].set\_title('Максимальное количество игроков на данный момент')  
  
axis[1].plot(  
 data["No."],  
 data["24th Peak"],  
 color="Green",  
 label='24 Пик'  
)  
axis[1].set\_title('24-ое максимальное количество игроков')  
  
axis[2].plot(  
 data["No."],  
 data["All\_time peak"],  
 color="Blue",  
 label='За всё время'  
)  
axis[2].set\_title('Максимальное количество игроков за всё время')  
  
plt.tight\_layout(pad=5)  
plt.show()  
#%%  
plt.title("Неправильный график Heatmap.")  
  
plt.imshow(  
 X=[data["No."], data["Current"], data["All\_time peak"]],  
 interpolation="nearest",  
 aspect='auto',  
 cmap='autumn',  
 norm=colors.Normalize(data["No."].min(), data["No."].max())  
)  
  
plt.show()  
#%%  
plt.title("Stacked area plot первых ста игр.")  
  
plt.stackplot(data["No."][:100],  
 data["Current"][:100],  
 data["24th Peak"][:100],  
 data["All\_time peak"][:100],  
 labels=["Current", "24th Peak", 'All\_time peak'])  
plt.legend(loc='upper right')  
  
plt.show()  
#%%  
plt.title("Violin plot первых ста игр.")  
  
plt.violinplot([data["Current"][:100],  
 data["24th Peak"][:100],  
 data["All\_time peak"][:100],  
 ])  
  
plt.show()  
#%%  
plt.title("Неправильная гистограмма первых 10 игр.")  
  
plt.hist2d(  
 x=data["All\_time peak"][:10],  
 y=data["24th Peak"][:10],  
 weights=data["No."][:10],  
 cmap='autumn',  
)  
  
plt.show()  
#%%  
plt.title("Bar plot первых 10 игр.")  
  
plt.bar(  
 data["Name"][:10],  
 data["Current"][:10],  
 color='maroon',  
)  
  
fig = plt.gcf()  
fig.set\_size\_inches(22, 10)  
fig.savefig('test2png.png', dpi=100)  
  
plt.show()  
#%% md  
  
#%% md  
# Вывод  
Количество игроков в самых популярных играх стремится к нулю по экспоненте, по принципу Парета 80% игроков сконцентрированы в первых 20% играх.

















