МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

Домашнее задание

по дисциплине Практикум по программированию

Студента Сафронова Александра Александровича
фамилия, имя, отчество полностью

Курс 2 Группа МО-211

Направление 02.03.03. Математическое обеспечение и администрирование информационных систем
код, наименование

Руководитель ассистент
должность, ученая степень, звание

Гуненков М. Ю.
фамилия, инициалы, дата, подпись

Выполнил
дата, подпись студента(ки)

Итоговый рейтинг	

Омск 2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Объемы накопленных данных в настоящее время настолько внушительны, что человеку просто не по силам проанализировать их самостоятельно, хотя необходимость проведения такого анализа вполне очевидна, ведь в этих "сырых данных" заключены знания, которые могут быть использованы при принятии решений, формировании статистических отчетов или составлении моделей машинного обучения.

В ходе изучения курса были использованы следующие библиотеки для языка программирования *Python*:

1. *NumPy* — библиотека с открытым исходным кодом с поддержкой многомерных массивов (включая матрицы) и высокоуровневых математических функций, предназначенных для работы с многомерными массивами.
2. *Matplotlib* — это библиотека для визуализации данных. В ней можно построить двумерные (плоские) и трехмерные графики.
3. *SymPy* — это библиотека *Python* с открытым исходным кодом, используемая для символьных вычислений. Она предоставляет возможности компьютерной алгебры в виде отдельного приложения.
4. *SciPy* — библиотека с открытым исходным кодом, предназначенная для выполнения научных и инженерных расчётов.
5. *Pandas* — программная библиотека для обработки и анализа данных. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.
6. *Seaborn* — библиотека для создания статистических графиков на *Python*. Она построена на основе *matplotlib* и тесно интегрируется со структурами данных *pandas*.[2]

Эти библиотеки позволяют проводить обработку, анализ и визуализацию данных, строить статистику на их основе.

1 Поиск и загрузка данных

Датасет «*Left 4 Dead 2 20,000+ Player's Statistics*» был выбран на сайте *kaggle.com*, специализирующемся на исследовании данных и машинном обучении.[1]

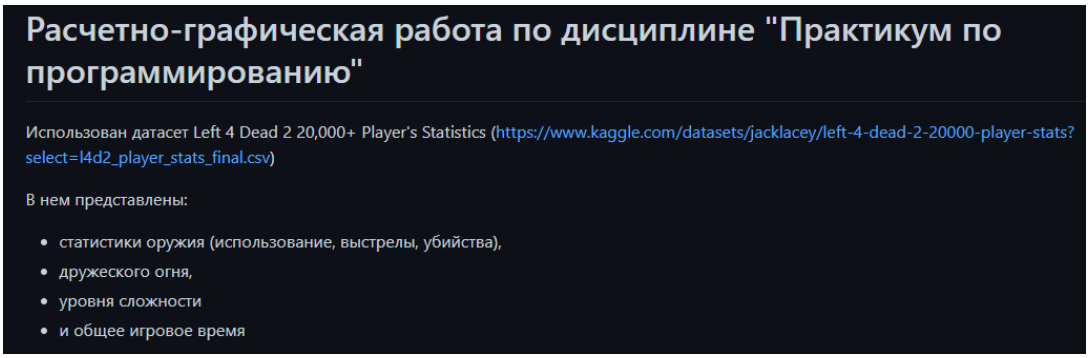

Рисунок 1 – файл *README.md*

Датасет был загружен в ноутбук командой *read_csv()* библиотеки pandas.

```
import pandas as pd

df = pd.read_csv('l4d2_player_stats_final.csv', sep=',')
```

Рисунок 2 – загрузка датасета

Данный датасет выглядит как набор из 20830 строк и 113 столбцов, разделенных запятой, в которых описаны данные игроков компьютерной онлайн-игры «*Left 4 Dead 2*», такие как процент использования всех видов оружия, количества убийств и попаданий из них, сложности, на которой пользователь играл, статистики стрельбы по дружественным персонажам и проведенного времени в игре.

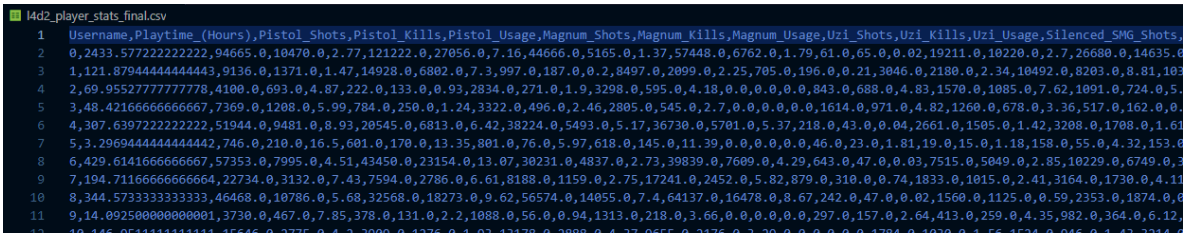

Рисунок 3 – небольшая часть датасета в формате *.csv*

	Username	Playtime_(Hours)	Pistol_Shots	Pistol_Kills	Pistol_Usage	Magnum_Shots	Magnum_Kills	Magnum_Usage	Uzi_Shots	Uzi_Kills	...	Knife_Usage	Molotovs_Thrown	Molotov_Kills	Pipe_Bombs_Thro
0	0	2433.577222	94665.0	10470.0	2.77	121222.0	27056.0	7.16	44666.0	5165.0	...	0.47	11166.0	99278.0	581
1	1	121.879444	9136.0	1371.0	1.47	14928.0	6802.0	7.30	997.0	187.0	...	0.03	788.0	10141.0	97
2	2	69.955278	4100.0	693.0	4.87	222.0	133.0	0.93	2834.0	271.0	...	0.00	23.0	130.0	44
3	3	48.421667	7369.0	1208.0	5.99	784.0	250.0	1.24	3322.0	496.0	...	0.00	135.0	1090.0	10
4	4	307.639722	51944.0	9481.0	8.93	20545.0	6813.0	6.42	38224.0	5493.0	...	0.00	613.0	4797.0	51
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
20825	20825	34.481389	10455.0	1494.0	10.70	2738.0	839.0	6.01	4613.0	595.0	...	0.00	24.0	212.0	7
20826	20826	19.644722	3834.0	766.0	11.30	3023.0	1135.0	16.75	1899.0	397.0	...	0.00	46.0	368.0	8
20827	20827	13.125278	3650.0	677.0	10.43	943.0	362.0	5.58	1401.0	85.0	...	0.03	34.0	311.0	8
20828	20828	11.973333	1982.0	239.0	3.62	1423.0	407.0	6.16	2200.0	516.0	...	0.00	44.0	260.0	3
20829	20829	31.906667	6104.0	950.0	6.25	1361.0	495.0	3.26	3686.0	746.0	...	0.00	49.0	348.0	10

Рисунок 4 – небольшая часть датасета, выведенного в виде таблицы

2 Разведывательный анализ данных

2.1 Гистограмма распределения числового признака

Гистограмма — способ представления табличных данных в графическом виде — в виде столбчатой диаграммы. Количественные соотношения некоторого показателя представлены в виде прямоугольников, площади которых пропорциональны. На гистограмме, приведенной ниже, приведено количественное распределение людей по признаку использования пистолета в процентах, по которому можно сделать вывод, что преимущественное большинство игроков используют пистолет менее чем 20% от общего времени игры.

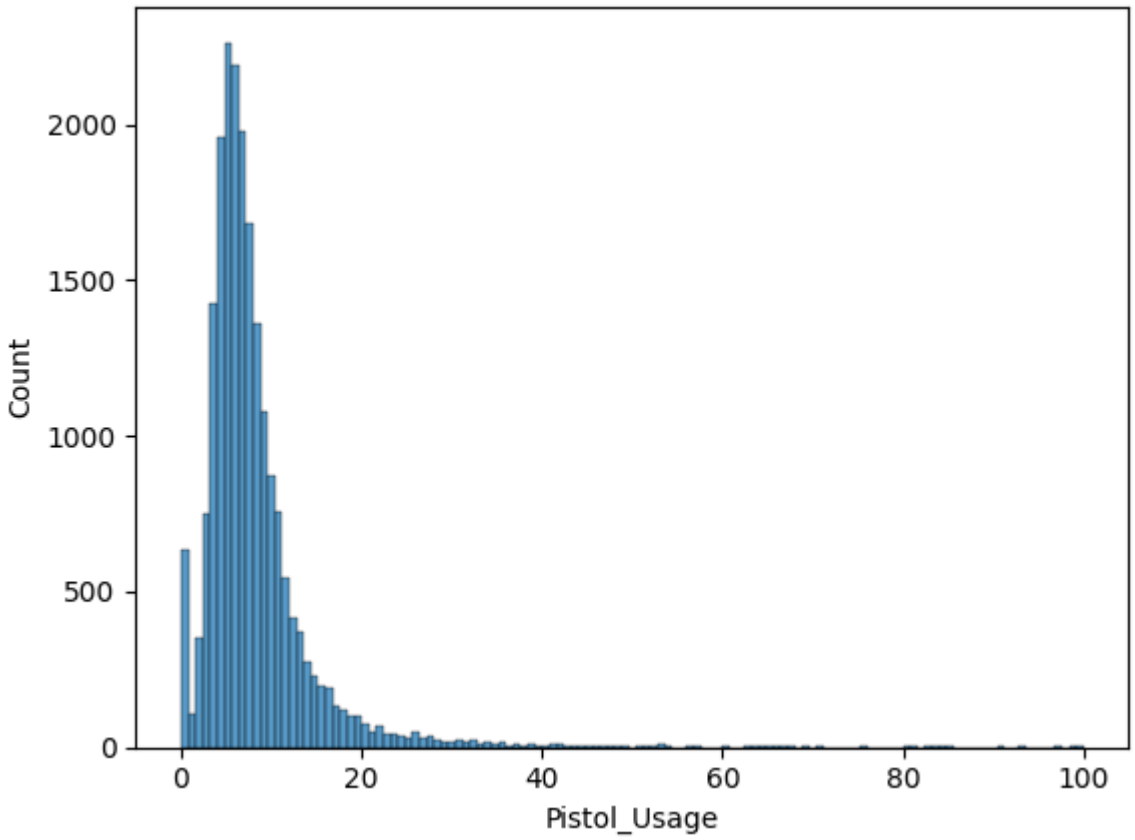

Рисунок 5 – гистограмма столбца *Pistol_Usage*

2.2 Диаграмма «ящик с усами» числового признака

Диаграмма «ящик с усами» — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы. На диаграмме, приведенной ниже, приведено распределение людей по проведенному в игре времени, по которому видно, что в датасете есть две аномалии на значениях приблизительно 260000 и 55000, которые могут помешать в будущем при анализировании данных.

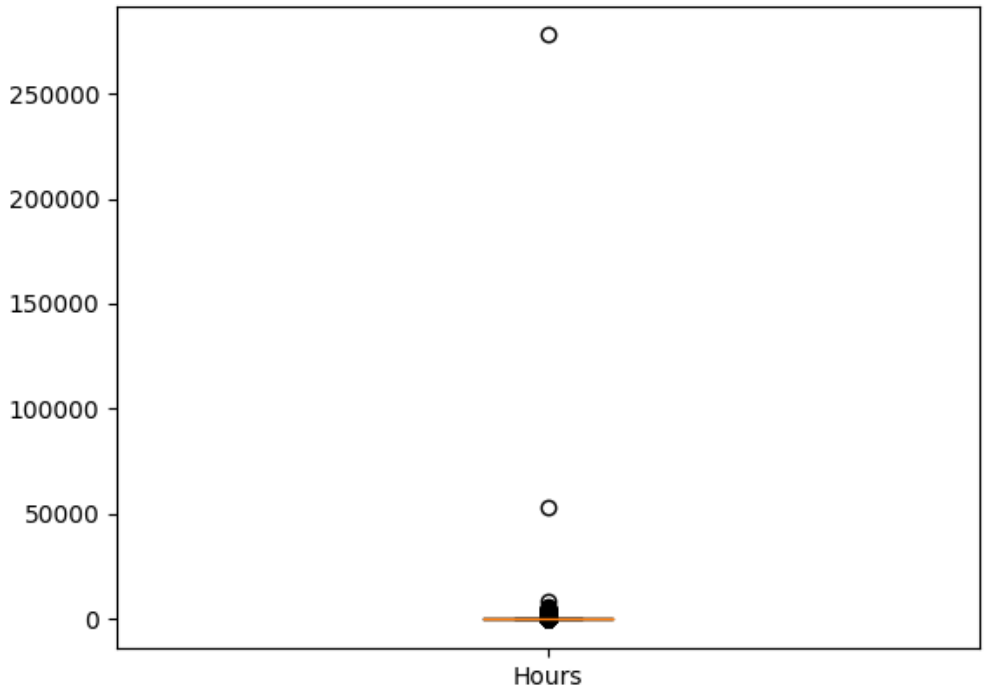

Рисунок 6 – Диаграмма «ящик с усами» столбца *Hours_without_reallife*

2.3 Круговая диаграмма номинативного признака

Круговая диаграмма — это круговая статистическая диаграмма, которая разделена на срезы, чтобы проиллюстрировать числовую пропорцию. На круговой диаграмме длина дуги каждого среза пропорциональна величине,

которую он представляет. Круговая диаграмма ниже демонстрирует распределение игроков по сложности. По диаграмме четко видно, что основными режимами сложности у игроков являются *Normal* и *Expert*.

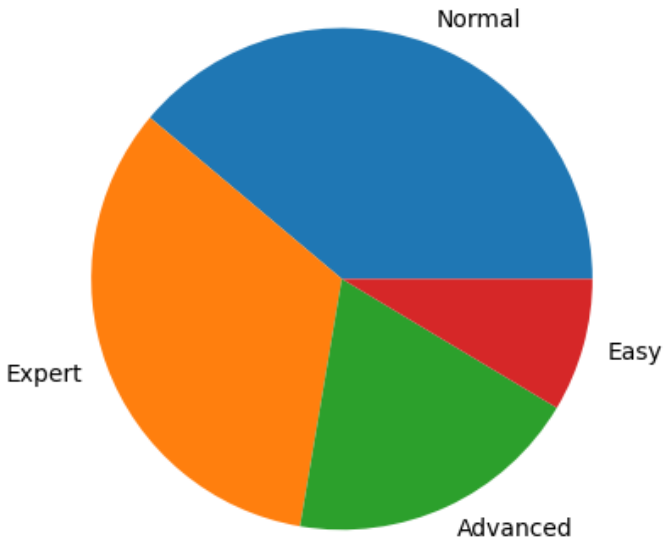

Рисунок 7 – круговая диаграмма столбца *Difficulty*

2.4 Тепловая карта со значениями взаимной корреляции между всеми парами признаков набора данных

Тепловая карта — графическое представление данных, где индивидуальные значения в таблице отображаются при помощи цвета.[4] На тепловой карте данного датасета можно выявить несколько особенностей, например, высокую зависимость параметров *Scout* и *AWP*, или *Military_Sniper_Rifle* и *Katana*. Это говорит о том, что пользователи, использующие винтовку *Scout*, также хорошо играют с винтовкой *AWP*, а игроки, использующие военную снайперскую винтовку, предпочитаю катану как оружие ближнего боя.

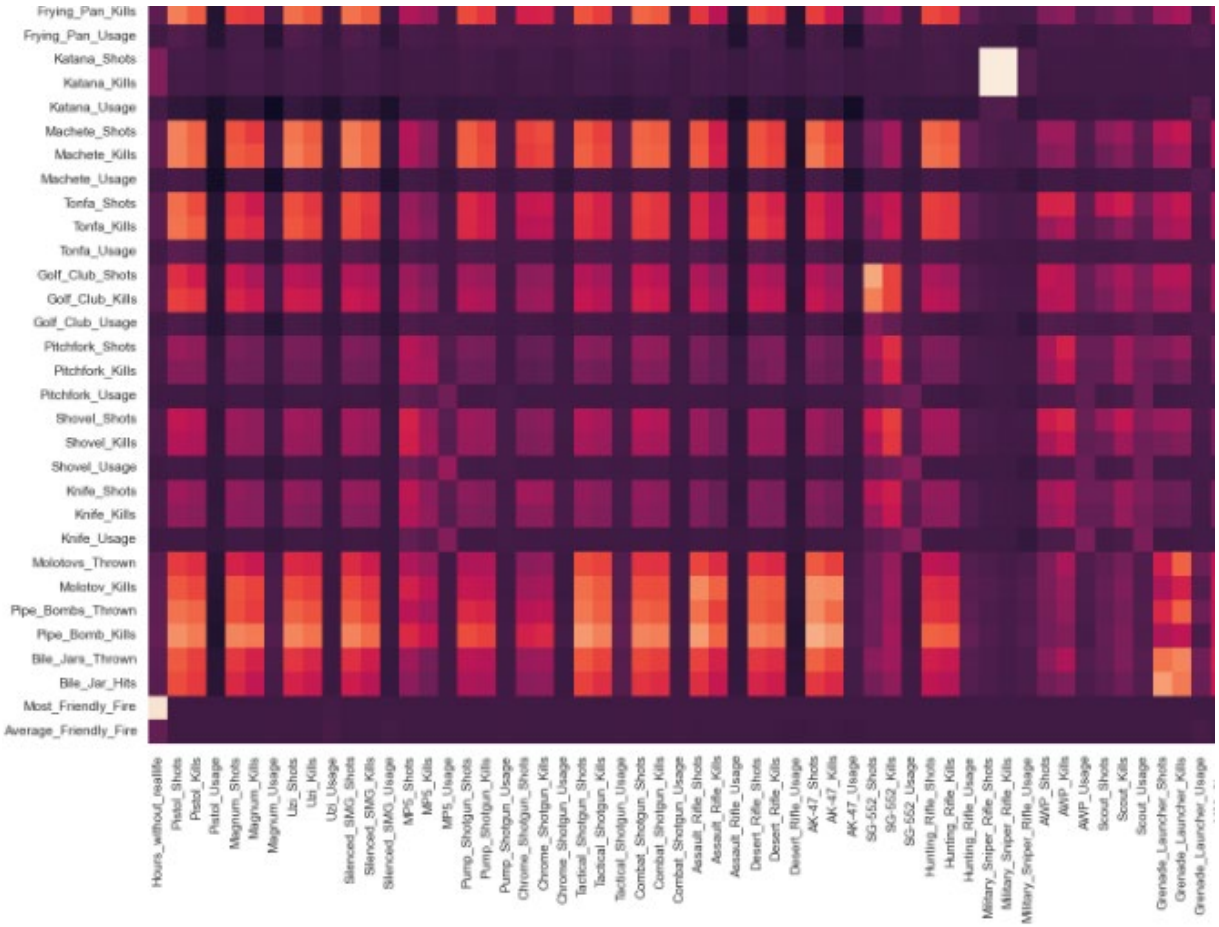

Рисунок 8 – малый фрагмент тепловой карты датасета

2.5 Диаграмма *countplot* с группировкой по двум номинативным признакам

CountPlot - столбчатая диаграмма, чаще всего используется для категориальных признаков в данных. Показывает, сколько строчек в датасете имеют каждое из выбранного значения категориального признака. Диаграмма ниже показывает, что с ростом уровня сложности увеличивается число убийств союзников.

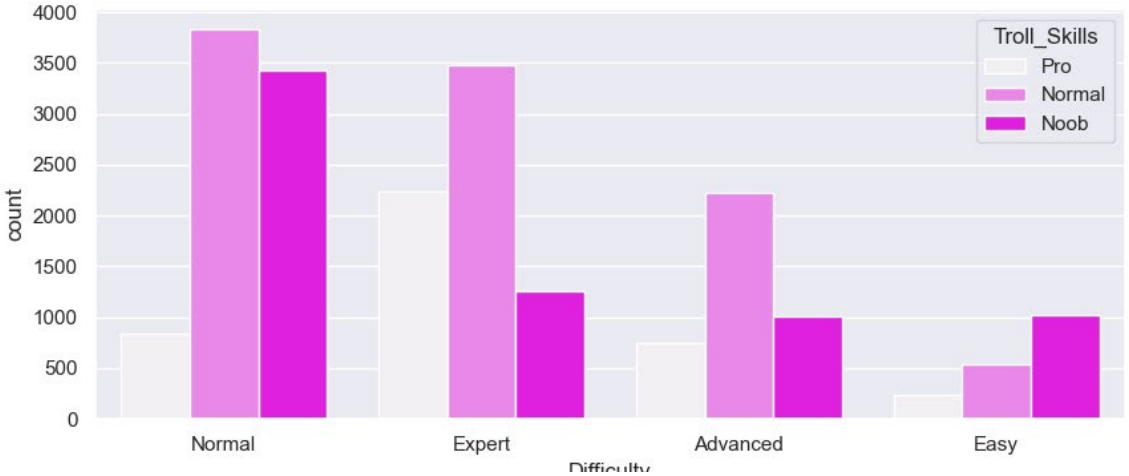

Рисунок 9 – Диаграмма *countplot* по столбцам *Troll_Skills* и *Difficulty*

3 Предварительная обработка данных

Для успешного анализа данных необходимо, чтобы все ячейки значений были заполнены. Для этого необходимо определить тип значений в столбце и, исходя из этого, подобрать метод заполнения пропущенных.

Правила заполнения ячеек значений:

1. Если значением признака является целое число, заполнить значением медианы по данному столбцу;
2. Если значением признака является действительное число, заполнить средним значением по данному столбцу;
3. Иначе заполнить значением моды по данному столбцу.

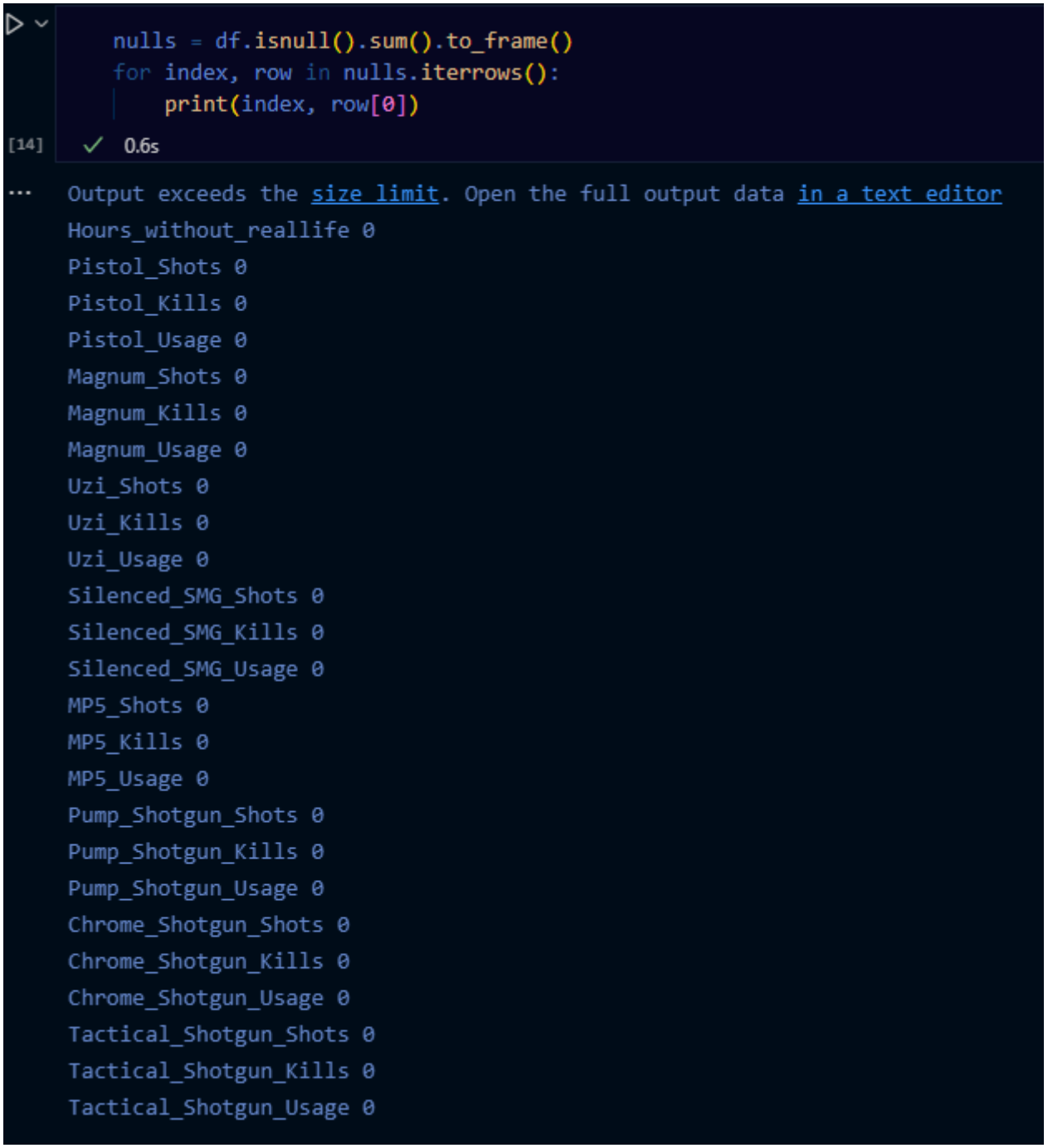

Рисунок 10 – фрагмент проверки наличия пропусков

В данном датасете пропущенных данных не оказалось, поэтому этап заполнения был пропущен.

Также было применено *one-hot* кодирование, то есть преобразование категориальных переменных в численные путем создания столбцов под каждую категорию и заполнения их значениями 0 и 1 в зависимости от категории каждой строчки.[3]

```
s = pd.Series(df.Difficulty)
pd.get_dummies(s)
```

[17] ✓ 0.3s

	Advanced	Easy	Expert	Normal
0	0	0	0	1
1	0	0	1	0
2	0	0	1	0
3	0	0	1	0
4	0	0	1	0
...	...	...	...	...
20825	0	0	0	1
20826	0	0	0	1
20827	0	0	1	0
20828	1	0	0	0
20829	1	0	0	0

Рисунок 11 – пример *one-hot* кодирования

Предобработанные данные были сохранены в формате *.csv* в той же директории, что и изначальный датасет.

```
df.to_csv('14d2_mega_final_demo_remix.csv')
```

✓ 1.6s

Рисунок 12 – сохранение итогового датасета

ЗАКЛЮЧЕНИЕ

В рамках задания были изучены библиотеки для языка *Python*, позволяющие проводить обработку, анализ и визуализацию данных, и закреплены навыки работы с ними. Был выбран и загружен подходящий датасет, данные которого были визуализированы пятью различными видами диаграмм для разведывательного анализа. По каждой из визуализаций сделаны соответствующие выводы. Также были проведены проверка данных датасета на пропуски и разбиение категориальных параметров на численные в рамках предварительной обработки данных. Полученный датасет был сохранен. Все вышеупомянутые шаги были сопровождены рисунками для более подробного описания. Задание было выполнено в полном объеме.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Left 4 Dead 2 20,000+ Player's Statistics | Kaggle. URL: https://www.kaggle.com/datasets/jacklacey/left-4-dead-2-20000-player-stats?select=l4d2_player_stats_final.csv (дата обращения: 18.12.22).

2 Визуализация данных в Seaborn. URL: https://nagornyy.me/it/vizualizatsiia-dannykh-v-seaborn/ (дата обращения: 18.12.22).

3 Быстрое кодирование (One-Hot Encoding). URL: https://www.helenkapatsa.ru/bystroie-kodirovaniie/ (дата обращения: 18.12.22).

4 Тепловая карта – Википедия. URL: https://ru.wikipedia.org/wiki/Тепловая_карта (дата обращения: 18.12.22).