**МИНОБРНАУКИ РОССИИ САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №1**

**по дисциплине «Машинное обучение»**

**Тема: «Исследование набора данных»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чишко В. О.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Татчина Я. А.

Санкт-Петербург

2024

**Цель работы:** Провести статистический анализ данных выбранного датасета.

**Краткое описание датасета**

Данные о продажах видеоигр в разных регионах. Представлена информация продаж Америки, Европы, Японии, остальных и общемировая. Датасет взят с сайта www.kaggle.com, вот ссылка на сам датасет: https://www.kaggle.com/datasets/gregorut/videogamesales/data Данные реальны.

Представлены как числовые, так и строковый, но работать будет работать исключительно с числовыми:

* Rank - топ по продажам – (числовой)
* Name - наименование видеоигры – (строковый)
* Platform - платформа, на которой игра выпускалась – (строковый)
* Year - год выхода – (числовой)
* Genre - жанр – (строковый)
* Publisher - издатель – (строковый)
* NA\_Sales - продажи в Америке (в миллионах) – (числовой)
* EU\_Sales - продажи в Европе (в миллионах) – (числовой)
* JP\_Sales - продажи в Японии (в миллионах) – (числовой)
* Other\_Sales - продажи в остальных частях всета (в миллионах) – (числовой)
* Global\_Sales - Общемировые продажи – (числовой)

Этот датасет был считан с помощью библиотеки pandas, и с ним была произведена дальнейшая работа.

**Определение параметров**

Для каждого числового атрибута были определены среднее значение и СКО. Для этого использовались функции библиотеки numpy: np.mean() и np.std() (данные представлены в файле с программой).

Также для каждого числового атрибута была построена гистограмма распределения значений при помощи функции библиотеки matplotlib plt.hist() (гистограммы представлены в файле с программой).

Проверка на наличие выбросов проводилась в функции findOutliers с использованием квартилей. У всех атрибутов данного датасета отсутствуют выбросы.

Проверка на наличие пропущенных значений проводилась при помощи функции библиотеки pandas pd.isna().sum(), которая выводит количество пропущенных значений для данного атрибута. У всех атрибутов данного датасета отсутствуют пропущенные значения.

**Корреляция между параметрами**

Была определена корреляция для всех пар числовых атрибутов. Эти данные представлены в корреляционной матрице (данные представлены в файле с программой).

По полученным данным видно, следующее: корреляция между продажами в разных регионах очень сильная.

Для демонстрации наличия корреляции между параметрами в программе представлены графики рассеивания всех продаж друг с другом. На графиках прекрасно видно наличие взаимосвязи между атрибутами (графики представлены в файле с программой).

**Вывод**

В данной работе я провёл статистический анализ датасета о продажах видеоигр в разных регионах.

Определив корреляцию атрибутов, стало понятно, что в данном датасете между ними присутствует сильная связь.