Введение в визуализацию данных

Андрей Куртасов
Системный аналитик

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

МФТИ

Проведем опрос

Как успехи с итоговыми заданиями по модулю «Python для анализа данных»?

Чем сегодня займемся?

МФТИ

1. Рассмотрим понятие визуализации данных и некоторые практические рекомендации.
2. Познакомимся с возможностями библиотек Matplotlib и Seaborn для построения различных видов графиков.

МФТИ

Визуализация данных

Сущность визуализации данных

- Визуализация данных — это представление данных в графическом виде.
- Создаются диаграммы, графики и карты, обеспечивающие доступный способ увидеть и понять тенденции, выбросы и закономерности в данных.
- Результаты визуализации могут быть интерактивными.

Как визуализация данных помогает в машинном обучении?

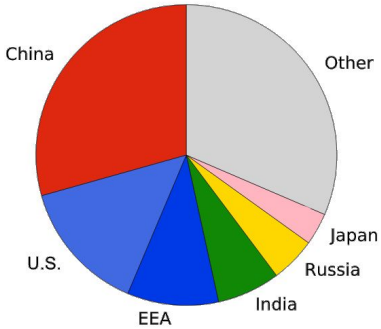
МФТИ

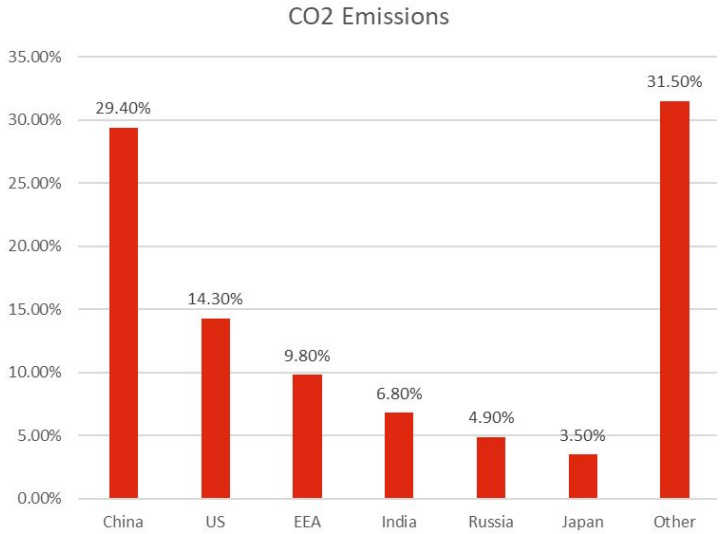
1. Выявление закономерностей, трендов и выбросов для предварительной обработки датасетов.
2. Выявление потенциальных корреляций между переменными для отбора признаков.
3. Мониторинг точности прогнозирования для дополнительной настройки модели в процессе использования.
4. Демонстрация результатов машинного обучения и прогнозной аналитики конечным пользователям.

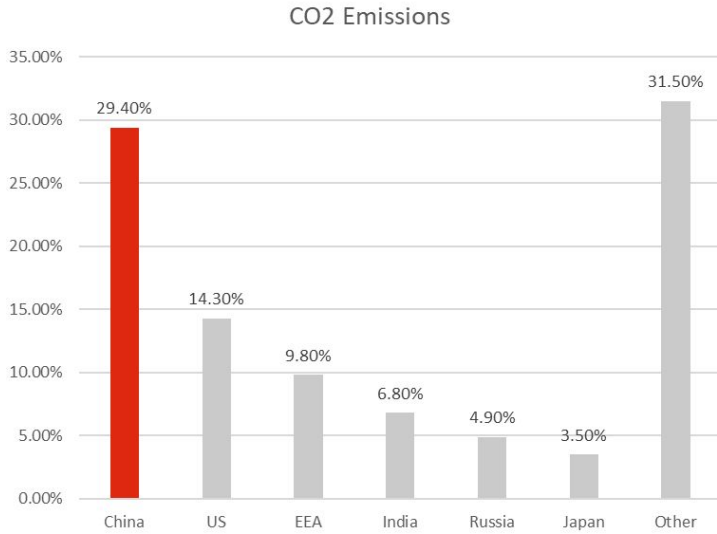
Главные требования к визуализации

- Наглядность
- Простота
- Понятность
- Соответствие цели исследования

Какая диаграмма нагляднее?







Для наглядного представления информации используют:

- разные формы
- раскраску элементов;
- выделение конкретных элементов цветом;
- надписи.

Изображения: https://www.polymersearch.com/blog/

Как построить график?

1. Определиться с типом графика:
   a. гистограмма — распределение числовой переменной;
   b. круговая диаграмма – отображение частей целого;
   c. столбчатая диаграмма — численная характеристика категориальной переменной;
   d. диаграмма рассеяния – взаимосвязь между двумя численными переменными;
   e. тепловая карта – отображение корреляции между переменными в датасете;
   f. и т. д.

Как построить график?

2. Найти подходящий график в одной из доступных библиотек.
   Например: https://seaborn.pydata.org/examples/index.html.
3. Передать в функцию библиотеки данные. Убедиться, что данные визуализируются корректно.
4. Настроить внешний вид графика:
   a. шкалы;
   b. цвета;
   c. подписи осей и элементов данных.

Соображения по внешнему виду графиков

- Учитывать особенности восприятия информации
  - Не отображать слишком много объектов одновременно
  - Прежде чем добавлять объект, спросите себя, дает ли он существенный прирост информации
- Учитывать data-ink ratio
- Учитывать теорию цвета
- Использовать с осторожностью:
  - 3D
  - Тени
  - Анимация
- Интерактивная графика – хороший способ оптимизировать визуализацию (но не всегда доступный)

$$\frac{data\text{–}ink}{total\text{–}ink} = \frac{Elements\ conveying\ data\ information}{All\ elements\ in\ the\ chart}$$

МФТИ

Вопросы? Комментарии? Напишите в чат или поднимите руку

МФТИ

Визуализация данных с помощью Python

Библиотеки Python для визуализации данных

МФТИ

В среде Python существует множество библиотек для визуализации:

- `Matplotlib` – наиболее популярная библиотека на основе `NumPy`
- `Seaborn` – расширение `Matplotlib` (статистические графики и др.)
- `Bokeh`, `plotly.py` – интерактивные графики
- `Altair` – генерирует графику согласно спецификации Vega-Lite (специальные JSON-структуры)
- `NumPy-stl` – создание 3D-изображений подобно CAD-системам

Упражнение 1: визуализация распределения числовой переменной

Дан датасет о блокбастерах:

https://github.com/sit-2021-int214/021-Worldwide-Blockbusters-2019-1977/raw/main/blockbusters_clean.csv.

1. Постройте гистограмму распределения рейтинга IMDB с помощью Matplotlib.
2. Количество столбцов: 15.
3. Подписи осей: X – количество фильмов, Y – рейтинг.
4. Цвет столбцов: голубой.

Упражнение 1a: добавим среднее и медиану

С помощью функции `axvline()` проведите:

a. пунктирную линию синего цвета для обозначения среднего значения;
b. пунктирную линию зеленого цвета для обозначения медианы.

Упражнение 2: взаимосвязь двух числовых переменных

Выведите график рассеяния по двум переменным:
бюджет фильма и мировые кассовые сборы.
Подпишите оси.

Рекомендуется испробовать возможности как Matplotlib,
так и Seaborn.

Упражнение 3: визуализация категориальных переменных

1. Ознакомьтесь с галереей примеров Seaborn: https://seaborn.pydata.org/examples/index.html
2. Нужно добавить на график рассеяния «бюджет – сборы» информацию о жанрах фильмов (колонка `genre_1`).
   Как это можно сделать? Поделитесь идеями.

Упражнение 4: круговая диаграмма

1. По колонке `mpaa_rating` постройте круговую диаграмму, показывающую соотношение между рейтингами.
2. На диаграмму добавьте процентные соотношения и число фильмов, относящихся к каждому из рейтингов.

МФТИ

Свободная дискуссия

Ваши вопросы? Пожелания?

До встречи!