Практическая работа №1. Анализ данных об электромобилях

Импортируйте файл в датафрейм под названием "Electric_Vehicle_Population_Data.csv" из https://drive.google.com/file/d/172oRlGRIHck8CcuOHmLszw63VsfvigFg/view?usp=drive_link (вам необходимо скачать) и посмотрите информацию о датафрейме

Цель: Научиться загружать датасет, находить базовую информацию, выполнять фильтрацию и простую визуализацию.

Шаг 1. Загрузка данных

Задание 1: Обзор данных

1. Выведите количество строк и столбцов.
2. Выведите названия всех столбцов.
3. Какие столбцы имеют пустые ячейки? Покажите верхние 6 строк датафрейма?
5. Покажите нижние 3 строк датафрейма?
6. Заполните пустые ячейки столбца Legislative District максимальным значением данного столбца
7. Заполните пустые ячейки столбца County значением 'No data'
8. Заполните пустые ячейки столбца Electric Range средним значением столбца Base MSRP

Задание 2: Работа с пропущенными значениями

1. Посмотрите, в каких столбцах есть пустые (NaN) значения.
2. Подумайте, что с ними можно сделать:
- Удалить строки с пустыми значениями
- Заполнить их каким-то значением (например, средним числом) 3. Выполните одно из решений.

Задание 3: Работа со столбцами

1. Сколько разных производителей машин в столбце Make?
2. Сколько всего моделей TESLA?
3. Сколько машин зарегистрировано в городе SEATTLE?

Задание 4: Фильтрация данных

1. Выведите все строки, где Model Year больше 2020.

2. Отфильтруйте только те машины, у которых тип — Battery Electric Vehicle (BEV).

Задание 5: Простая визуализация

1. Постройте диаграмму количества машин по годам (Model Year).

2. Постройте диаграмму топ-5 производителей.

Задание 6: Напишите краткий вывод

- Какой год был самым популярным по количеству регистраций?
- Какие марки встречаются чаще всего?
- В каких городах больше всего электромобилей?

Задание 7: Работа с пропущенными значениями

1. Посмотрите, в каких столбцах есть пустые (NaN) значения.

2. Подумайте, что с ними можно сделать:

- Удалить строки с пустыми значениями
- Заполнить их каким-то значением (например, средним числом) 3. Выполните одно из решений.

Задание 8: Сохранение обработанного датасета

1. После удаления или заполнения пустых значений сохраните очищенный датафрейм в новый CSV-файл. 2. Убедитесь, что файл сохранился корректно.

Сохраняем очищенные данные в новый файл

Проверим сохранённый файл (можно заново прочитать и посмотреть, вывести количество строк с удаленными пустыми значениями)

Задание 8. Группировка и агрегация

Цель: Научиться использовать группировку данных.

1. Посчитайте, сколько автомобилей приходится на каждого производителя (Make).
2. Найдите средний год выпуска машин у каждого производителя.
3. Определите топ-5 производителей по количеству зарегистрированных автомобилей.
4. Постройте столбчатую диаграмму с топ-5 производителей.

Вопросы к обсуждению:

- Чем отличается метод .groupby() от фильтрации?

- Какие типы агрегатных функций (среднее, максимум, минимум, сумма) чаще всего применяются?

Задание 9. Работа с датами и временем

Цель: Понять, как использовать временные данные.

1. Преобразуйте столбец с датой регистрации (Registration Date, если он есть) в формат datetime.
2. Определите, сколько машин было зарегистрировано по годам.
3. Постройте график регистрации автомобилей по годам.
4. Найдите месяц, в который чаще всего регистрируют электромобили.

Вопросы к обсуждению:

- Почему при анализе больших данных важно правильно работать с датами?
- В каких случаях анализ по месяцам или кварталам полезнее, чем по годам?

Задание 10. Визуализация распределений

Цель: Изучить распределение данных.

1. Постройте гистограмму распределения автомобилей по годам выпуска.
2. Постройте диаграмму количества автомобилей в зависимости от типа двигателя (BEV, PHEV и др.).
3. Используйте seaborn для построения boxplot (ящика с усами) по мощности (Electric Range, если есть).

Вопросы к обсуждению:

- Что позволяет понять гистограмма?
- Для чего используют boxplot?
- Как визуализация помогает находить выбросы?

Задание 11. Выбросы и аномалии

Цель: Научиться находить и анализировать выбросы.

1. Постройте диаграмму разброса (scatter plot) зависимости года выпуска и запаса хода (Electric Range).
2. Найдите автомобили с запасом хода больше 500 км (или другого явно необычного значения).
3. Решите, как обработать такие выбросы: удалить или оставить.
4. Обсудите, могут ли эти данные быть ошибкой ввода. Вопросы к обсуждению:

- Почему выбросы могут искажать анализ?
- Когда выбросы стоит удалять, а когда — учитывать?

Задание 12. Сводные таблицы (pivot tables)

Цель: Научиться создавать сводные таблицы.

1. Создайте сводную таблицу: строки — производители, столбцы — тип двигателя, значения — количество машин.
2. Определите, какой тип двигателя наиболее популярен у TESLA.
3. Постройте тепловую карту (heatmap) для сводной таблицы.

   Вопросы к обсуждению:

   - Чем полезны сводные таблицы при анализе больших данных? Какие преимущества у сводной таблицы перед обычной группировкой?