ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cтарший преподаватель |  |  |  | В. В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1 |
| Предварительный анализ данных 2024 по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ гр. № | 4217 |  |  |  | В. А. Милованов |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2024

**Цель работы**

Осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы в этих данных.

**Индивидуальное задание**

Вариант 4

Данные о поддержанных авто:

1. имя

2. год

3. цена продажи

4. пройдено км

5. топливо

6. тип продавца

7. коробка передач

8. тип владельца

Задание 1: Группировка - тип коробки передач (transmission) по количеству разных типов топлива (fuel) .

Задание 2: Группировка - тип коробки передач (transmission) по количеству разных типов топлива (fuel). Создать датафрейм. Переименовать столбец с количеством в “сount”. Отсортировать по убыванию столбца “count”.

Задание 3: Сводная таблица (pivot\_table) - средняя стоимость для каждого года year. Отсортировать по убыванию. Округлить до трёх знаков.

Задание 4: Сводная таблица (pivot\_table) - среднее количество пройденных километров для каждого типа коробки передач transmission - строки и типа топлива - fuel - столбцы. Отсортировать по возрастанию transmission. Округлить до двух знаков.

**Ход работы**

Импортируем библиотеку pandas. Создадим БД из файла auto.czv. Выведем результат (рис. 1).

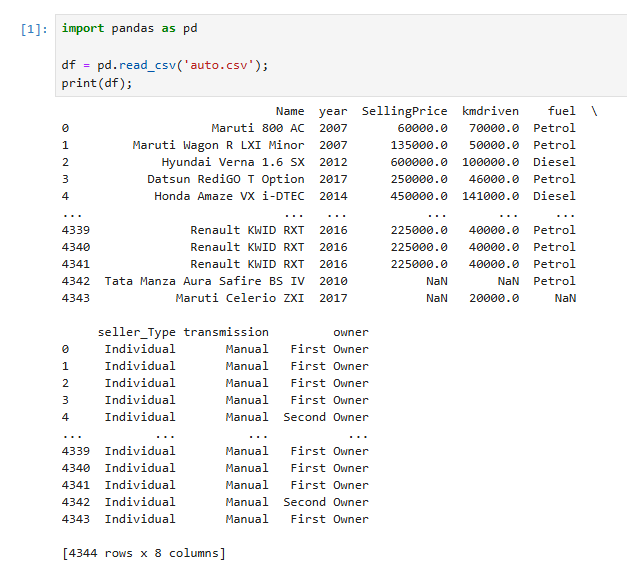


Рисунок 1 – Импорт pandas

Получим информацию о первых 20 строках БД с помощью метода head. Передадим ему число строк, нужных для вывода (рис. 2).

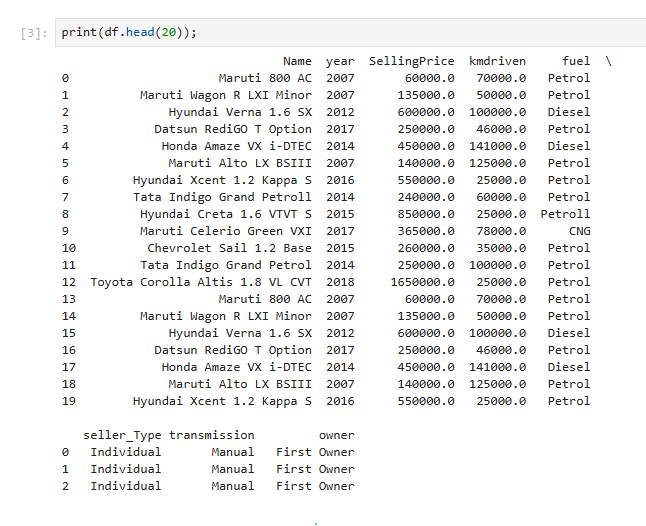


Рисунок 2 – Вывод строк

Данные представляют собой информацию о подержанных автомобилях, выставленных на продажу. В таблице содержатся следующие столбцы: Name - Название и модель автомобиля; year - Год выпуска автомобиля; SellingPrice - Цена продажи автомобиля в рублях; kmdriven - Пробег автомобиля в километрах; fuel - Тип топлива, используемый автомобилем; seller\_Type - Тип продавца; transmission - Тип коробки передач; owner - Количество предыдущих владельцев. Предметная область данных — вторичный рынок автомобилей. На основе этих данных можно анализировать цены на подержанные автомобили в зависимости от их характеристик, таких как год выпуска, пробег, тип топлива и количество владельцев. Это полезно для покупателей, которые хотят оценить, насколько справедлива цена, и для продавцов, которые хотят определить рыночную стоимость своих автомобилей.

Получим информацию по столбцам с помощью info (рис. 3).

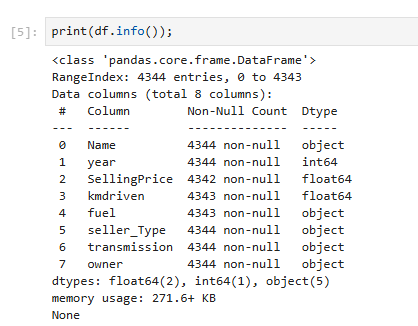


Рисунок 3 – Информация о полях

Выведем информацию по числовым данным, используя метод describe (рис. 4).

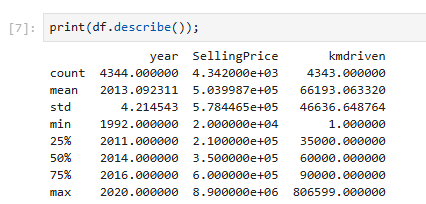


Рисунок 4 – Числовые поля

Исследуем названия столбцов при помощи colums (рис. 5).

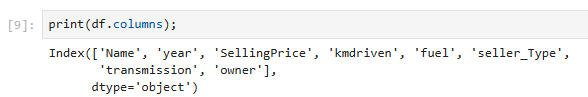


Рисунок 5 – Названия полей

Заметим, что часть названий отличается от общего вида. Для улучшения читаемости приведем их к стилю snake\_case. Используем rename (рис. 6).

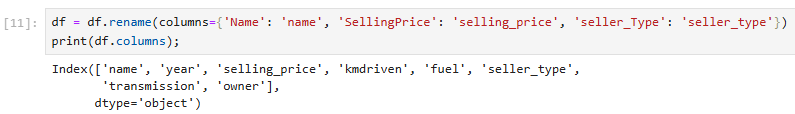


Рисунок 6 – Изменение названий полей

Проверим результат.

Перейдем к поиску пропущенных значений. Посчитаем найденные методом isna пропуски с помощью sum (рис. 7).

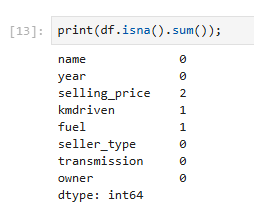


Рисунок 7 – Поиск пропусков

Избавимся от строк с неполными данными в полях selling\_price, kmdriven, fuel и посмотрим на итог (рис. 8).

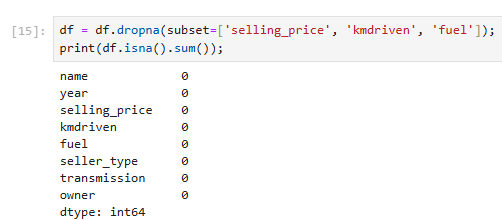


Рисунок 8 – Удаление пропусков

Выявим дубликаты. Начнем с явных. Применим метод duplicated в связке с sum (рис. 9).



Рисунок 9 – Поиск явных дубликатов

Отчистим БД от явных дубликатов методом drop\_duplicates. Обновим индексацию, чтобы избавиться от пропусков с помощью reset\_index. Снова увидим количество явных дубликатов. Убедимся в их отсутствии (рис. 10).

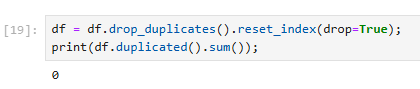


Рисунок 10 – Удаление явных дубликатов

Перейдем к неяным дубликатам. Вернем перечень уникальных значений столбца name. Заметим, что столбец состоит почти полностью из уникальных данных, поэтому перейдем к следующим полям (рис. 11).

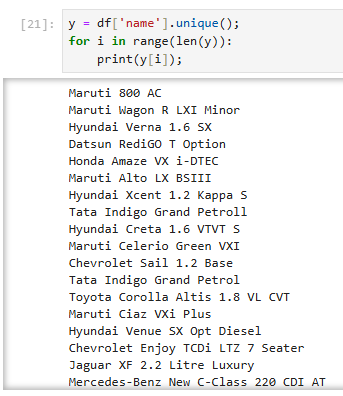


Рисунок 11 – Анализ столбца name

Исследуем уникальные значения остальных текстовых полей (рис. 12).

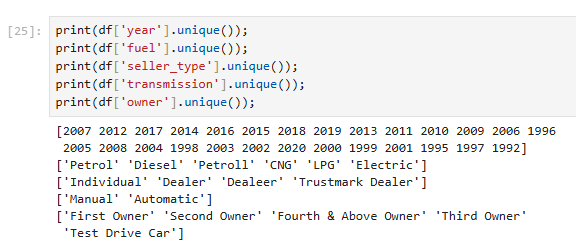


Рисунок 12 – Анализ текстовых столбцов

Найдем неявные дубликаты в виде опечаток в уже существующих значениях. Заменим эти данных в столбцах fuel и seller\_type методом replace (рис. 13).

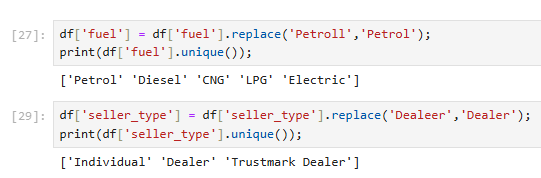


Рисунок 13 – Замена данных

Исследуем уникальные значения числовых полей (рис. 14).

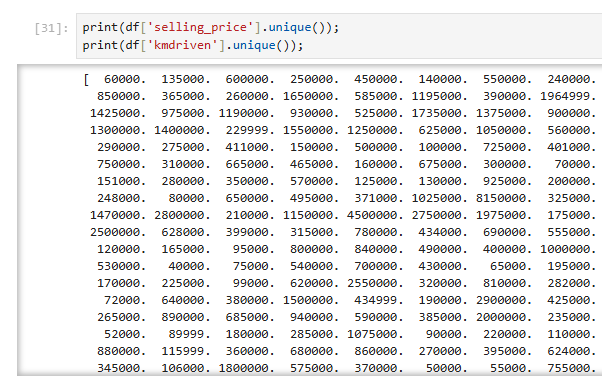


Рисунок 14 – Анализ числовых столбцов

Поменяем тип данных на int, так как все значения в выбранных полях целые (рис. 15).

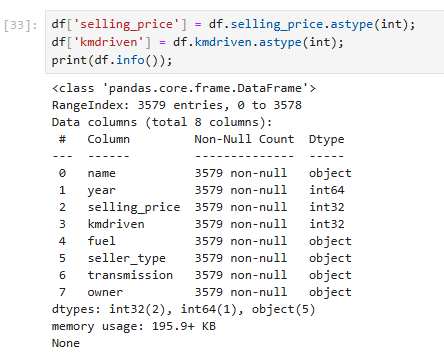


Рисунок 15 – Изменение типа данных

Перейдем к группировке. Сгруппируем тип коробки передач (transmission) по количеству разных типов топлива (fuel), используя groupby (рис. 16).

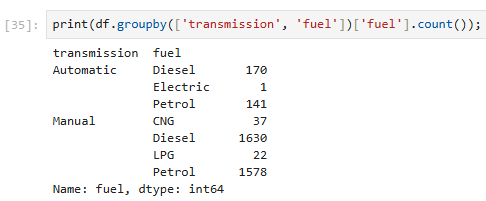


Рисунок 16 – Группировка

Создадим датафрейм из группы, поменяем название последнего поля на count (рис. 17).

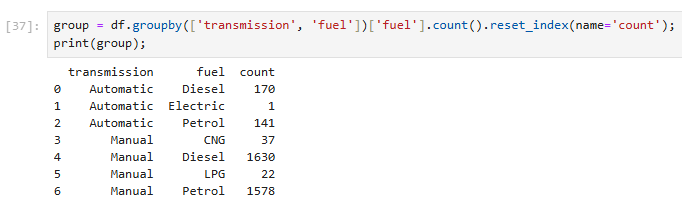


Рисунок 17 – Создание датафрейма

Перейдем к сортировке. Отсортируем датафрейм по убыванию столбца count (рис. 18).

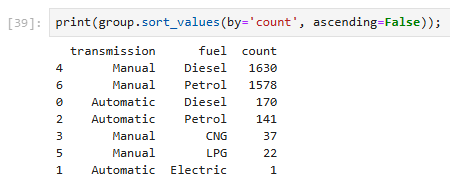


Рисунок 18 – Сортировка

Создадим сводные таблицы, используя метод pivot\_table. Первая таблица основана на средней стоимости для каждого года year. Отсортирована по

убыванию. Значения округлены до трёх знаков (рис. 19).



Рисунок 19 – Первая сводная таблица

Вторая таблица основана на среднем количестве пройденных километров для каждого типа коробки передач transmission - строки и типа топлива - fuel - столбцы. Отсортирована по возрастанию transmission. Значения округлены до двух знаков (рис. 20).

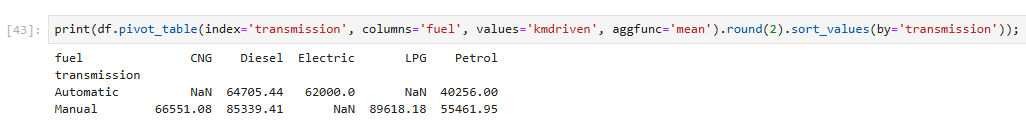


Рисунок 20 – Вторая сводная таблица

Ссылка на файл:

<https://github.com/Vvvvv55531/LR1.git>

**Вывод**

В ходе лабораторной работы мной была проведена обработка данных, содержащих информацию о подержанных автомобилях, с использованием Google Colab и библиотеки pandas. Основной задачей было изучить приведенную структуру данных и провести её предварительную обработку.

После загрузки датасета и ознакомления с его структурой был сделан анализ основных характеристик автомобилей. Далее следовала проверка на наличие пропусков и дубликатов, после чего данные были очищены и приведены в правильный и однородный вид. Были устранены ошибки в названиях столбцов, а также скорректированы типы данных.

В дальнейшем были выполнены группировки данных и создание сводных таблиц в соответствии с заданием. Эти шаги помогли систематизировать информацию и выделить ключевые показатели, что сделало данные более удобными для восприятия.

Итогом работы стало создание структурированной и подготовленной базы данных, которая может быть использована для глубокого анализа рынка подержанных автомобилей и принятия обоснованных решений.