ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

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

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1 |
| ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | В.С. Тихонов |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург

2022

**Цель работы:** осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы этих данных.

Вариант 21. Ссылка на репозиторий: <https://github.com/Deiwas/analysis_lab1>

1. Датасет был загружен с помощью библиотеки pandas в Jupyter-ноутбук. После вывел первые 20 строк с помощью head (рис. 1).

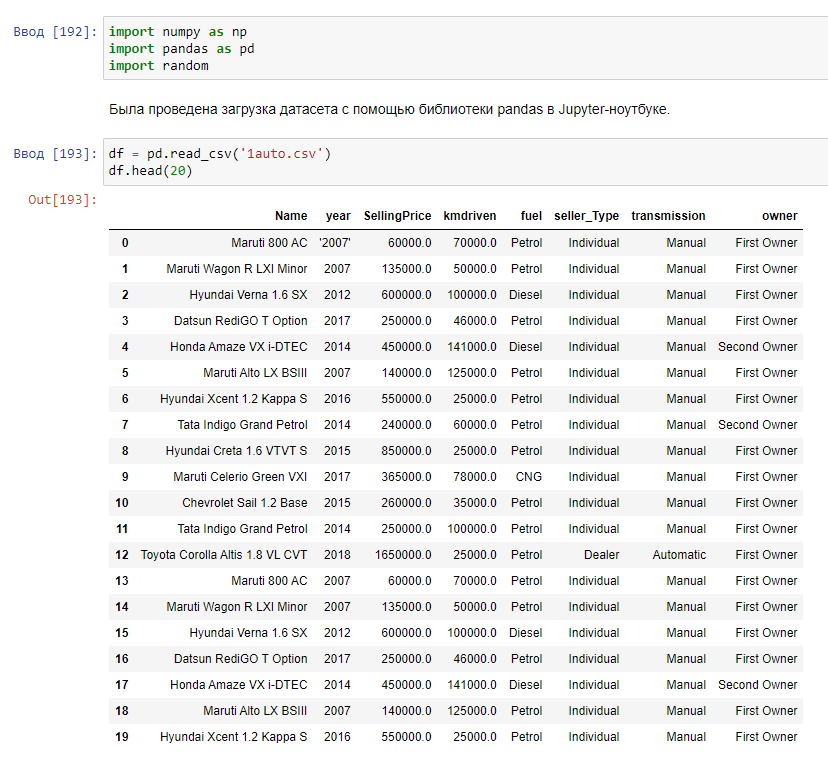


Рисунок 1 – Загруженный датасет

1. В данном датасете приведены данные об автомобилях. Первый столбец представляет собой название авто, следующий столбец – год производства, затем стоимость продажи автомобиля, его пробег, тип топлива автомобиля, тип трансмиссии, и два столбца информации о продавце.
2. Проведена оценка данных методом .info (рис. 2).

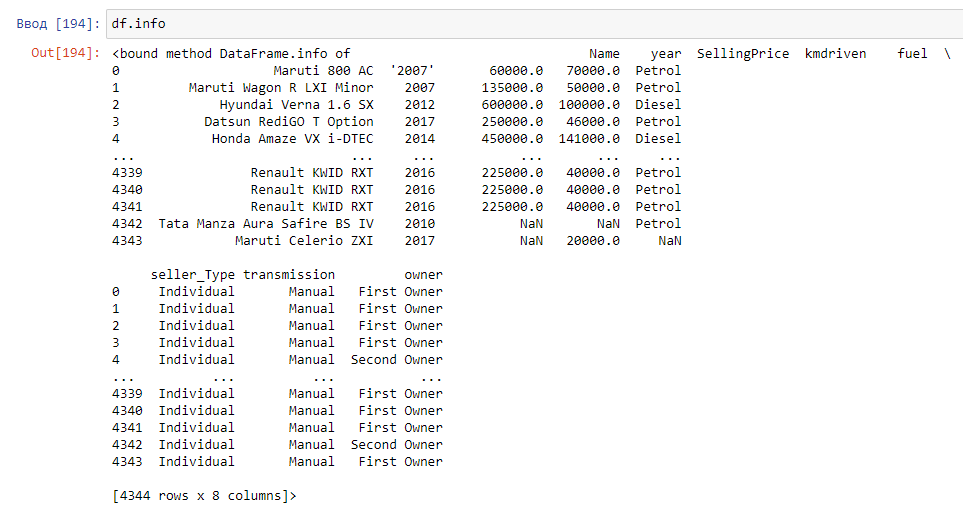


Рисунок 2 – Метод .info

1. Вывел на экран названия столбцов с помощью метода .columns (рис. 3), все названия написаны без пробелов и передают суть столбца.

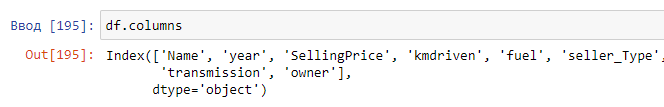


Рисунок 3 – Информация о названиях столбцов

1. Проведена проверка на наличие дубликатов в столбцах датафрейма (рис. 4). После избавился от дубликатов с помощью метода drop\_duplicates (рис. 5).

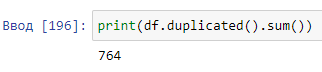


Рисунок 4 – Обнаружены дубликаты

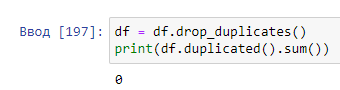


Рисунок 5 – Дубликаты удалены

1. После удаления дубликатов проведена повторная индексация датафрейма (рис. 6).

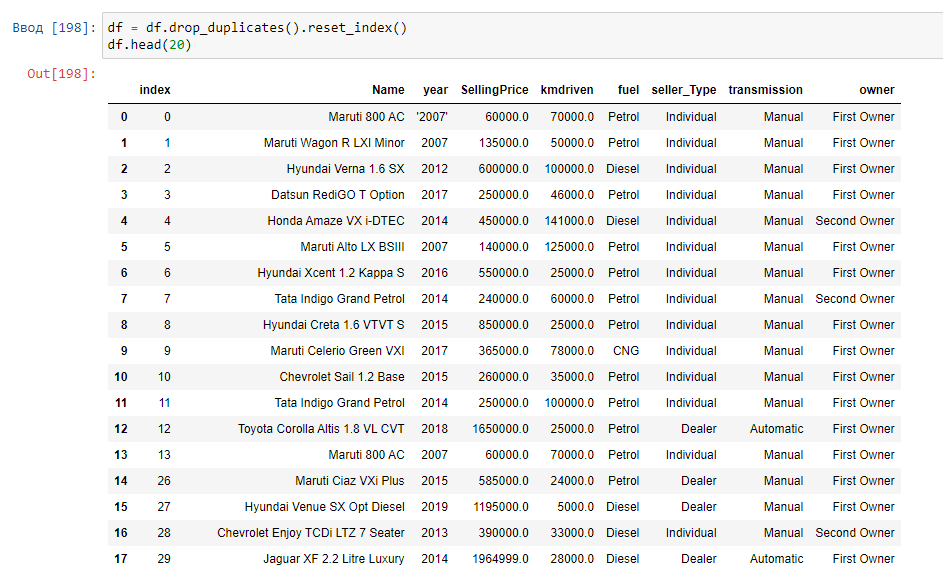


Рисунок 6 – Датафрейм с новыми индексами

1. Проведен поиск пропусков в таблице. Поскольку SellingPrice является важным столбцом датафрейма, все строчки без него будут удалены. Это решает и остальные проблемы наличия пустых ячеек датафрейма (рис. 7).

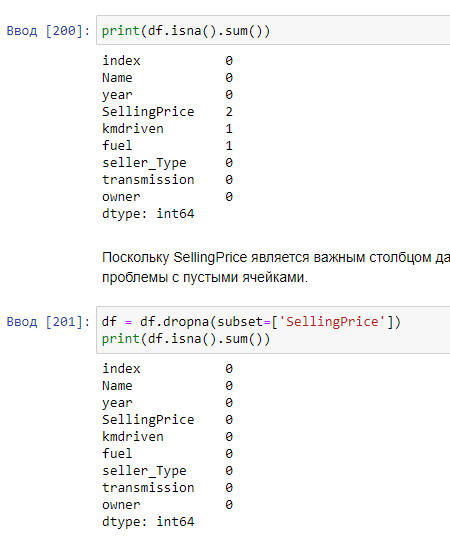


Рисунок 7 – Удаление пустых ячеек в датафрейме

1. Для работы с типами данных было решено перевести тип данных столбца года в числовой. Для осуществления перевода пришлось написать метод для удаления одинарных кавычек из столбца (рис. 8).

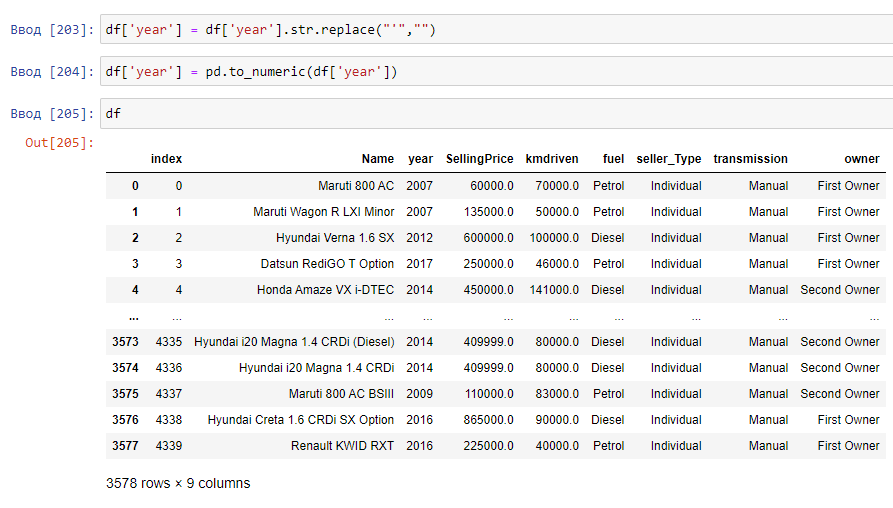


Рисунок 8 – Изменение типа данных столбца

1. Осуществлена постройка сводной таблицы по данным этого датафрейма (рис. 9).

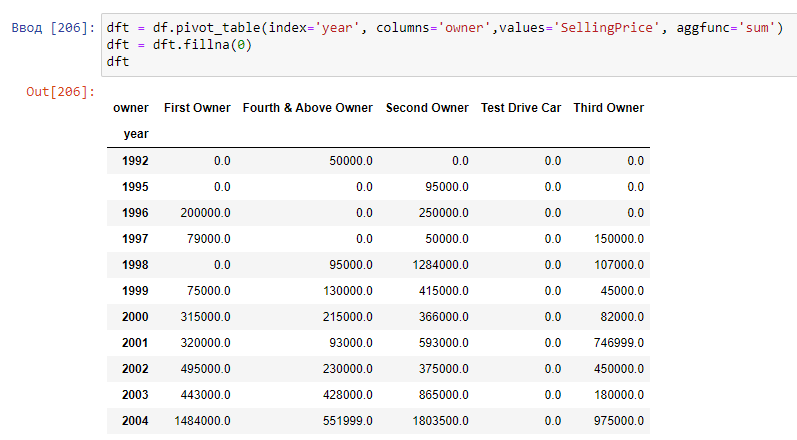


Рисунок 9 – Первая сводная таблица

1. Построил еще две сводные таблицы (рис. 10, 11).

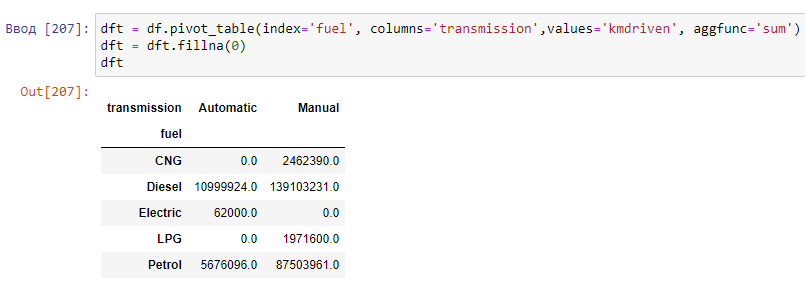


Рисунок 10 – Вторая сводная таблица

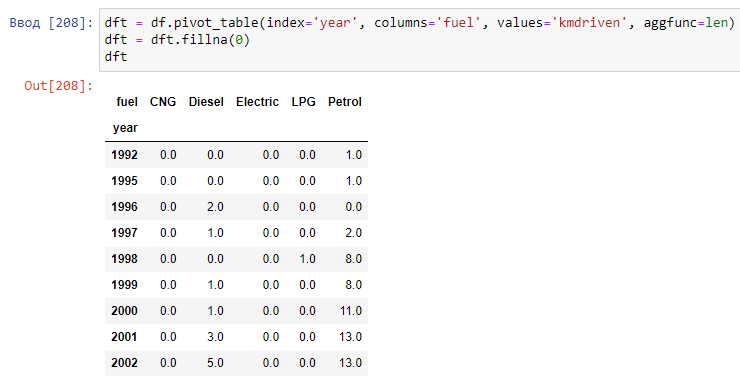


Рисунок 11 – Третья сводная таблица

1. Выделение пяти годов с максимальными суммарными продажами (рис. 12).

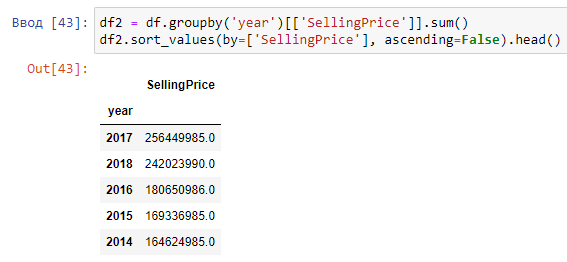


Рисунок 12 – Пять годов с максимальными продажами

**Вывод:** в ходе выполнения данной лабораторной работы были получены навыки обработки данных и их первичного анализа. Для обработки при помощи библиотеки pandas был загружен набор данных с автомобилями и информацией о них, включая предыдущих владельцев, стоимость продажи и пробег автомобиля. Были удалены дубликаты, исправлены пропуски в данных изменен тип данных одного из столбцов. Также были придуманы три сводные таблицы: сводная таблица стоимости автомобилей по годам в зависимости от владельца показывает, что преобладают продажи первого и второго владельцев, при этом тест-драйв машины не показывают спроса ни в одном году. Сводная таблица пробега в зависимости от топлива и трансмиссии показывает преобладание дизельного и бензинного топлива, а также ручной трансмиссии. Исключением являются электрические автомобили, где преобладает автоматическая трансмиссия. Последняя сводная таблица показывает полное преобладание продажи автомобилей с бензинным топливом.

Список источников:

1. Методические указания к выполнению данной работы: https://pro.guap.ru/get-task/8bc94a9491c58f462cefbcfd5c771b9a
2. Информация о Jupyter: https://thecode.media/jupyter/