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 №2 |
| Исследовательский 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**

Изучение связи между признаками двумерного набора данных, визуализация 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 : Использовать seaborn. По группировке - owner и количество машин каждого типа коробки передач (transmission) построить диаграмму следующего вида (см. рис. 1).

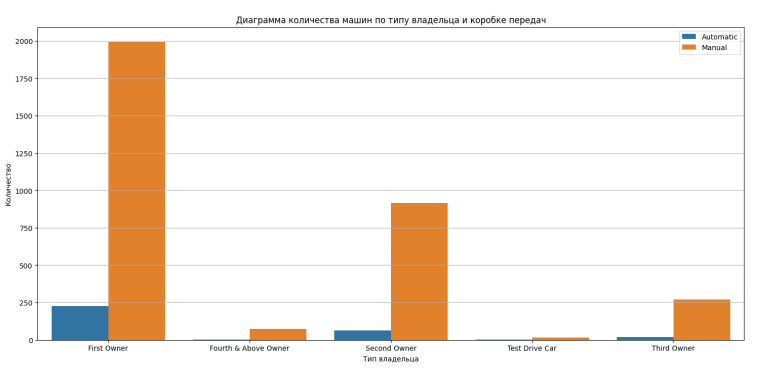


Рисунок 1 – Задание 1

Задание 2: Использовать pandas и plot. Отфильтровать набор данных и выбрать год выше 2005. По сводной таблице (pivot\_table) - отобразить среднее количество пройденных километров (kmdriven) по годам. Оставить только маркеры в виде ★ розового цвета размером 20 (см. рис. 2).

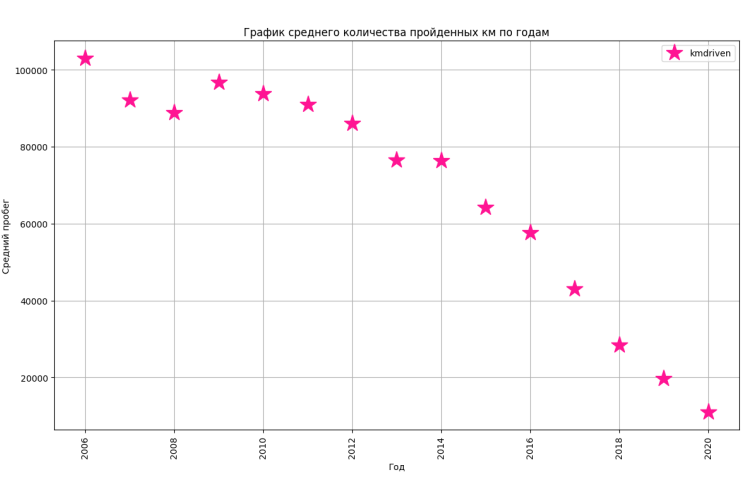


Рисунок 2 – Задание 2

Задание 3: Использовать matplotlib. Построить круговую диаграмму, которая отображает процент каждого типа топлива (fuel). Можно убрать из диаграммы тип топлива, который встречается всего 1 раз (см. рис. 3).

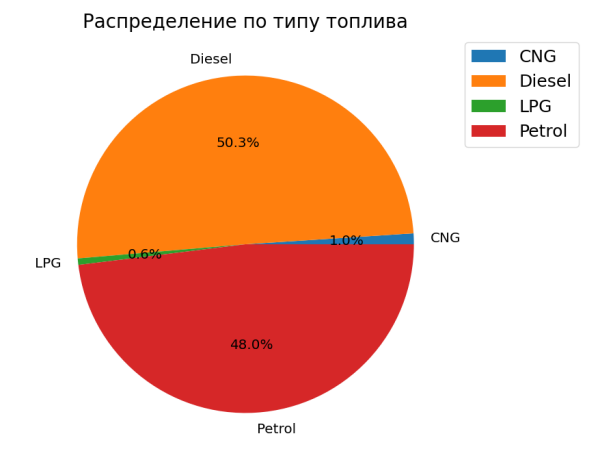


Рисунок 3 – Задание 3

**Ход работы**

Была импортирована библиотека pandas. Была создана БД из файла auto2.czv (см. рис. 4).



Рисунок 4 – Импорт auto2.czv

Начало предварительной обработки данных (как в 1 лр).

Было замечено, что часть названий отличается от общего вида. Для улучшения читаемости они были приведены к стилю snake\_case. Был использован rename (см. рис. 5).



Рисунок 5 – Изменение названий

Произошло удаление строк с неполными данными в полях selling\_price, kmdriven, fuel (см. рис. 6).



Рисунок 6 – Удаление строк с неполными данными

Производилась отчистка БД от явных дубликатов методом drop\_duplicates. Была обновлена индексация для избавления от пропусков с помощью reset\_index. Проверялось количество явных дубликатов (см. рис. 7 ).



Рисунок 7 – Отчистка БД от явных дубликатов

Были найдены неявные дубликаты в виде опечаток в уже существующих значениях. Эти данных были замечены в столбцах fuel и seller\_type методом replace (см. рис. 8).



Рисунок 8 – Нахождение неявных дубликатов

Был изменен тип данных на int, так как все значения в выбранных полях целые (см. рис. 9).



Рисунок 9 – Изменение типа данных

С помощью pandas и метода plotting.scatter\_matrix() была реализована матрица рассеивания, задан её размер. Матрица рассеивания (или scatter matrix) — это график, который помогает визуализировать взаимосвязи между несколькими переменными в наборе данных. Он представляет собой сетку из диаграмм рассеяния (scatter plots), где каждая пара переменных отображается в виде графика (см. рис. 10).

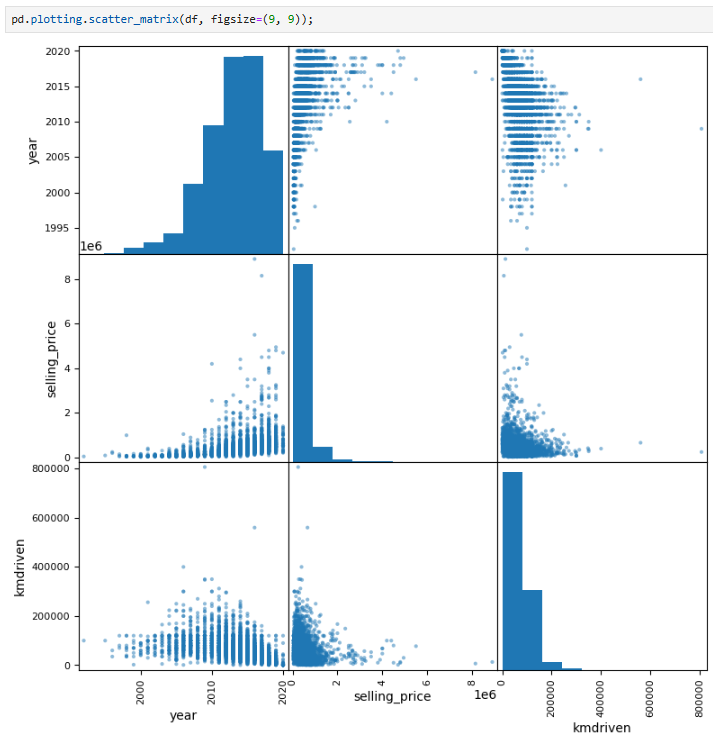


Рисунок 10 – Матрица рассеивания

Была импортирована библиотека seaborn. Seaborn — это библиотека для визуализации данных, построенная на основе matplotlib. Она предоставляет удобное представление функций для создания сложных графиков с минимальными усилиями. Для настройки матрицы из добавленной библиотеки был использован метод pairlot(). Были указаны числовые данные БД и столбец fuel для разделения данных по виду топлива с помощью параметра hue (см. рис. 11).

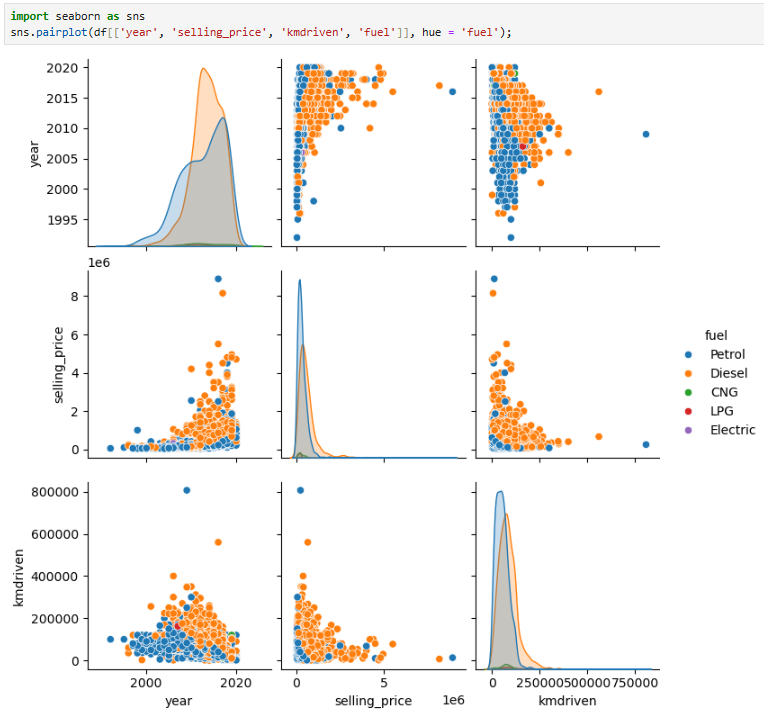


Рисунок 11 – Матрица рассеивания с помощью seaborn

По матрице рассеивания были замечены следующие связи между числовыми данными БД: Самые высокие и частые показатели пробега приходятся на промежуток с 2005 по 2010 года; C каждым новым годом цены на продажу автомобилей росли и достигли максимума в промежутке 2015-2020, ближе к 2020 году пошли на спад; График зависимости между пробегом и ценой продажи показывает уменьшение количества машин с высоким пробегом с увеличением их стоимости; Заметно доминирование дизельного топлива с годами и доминирование бензина с увеличением других показателей.

Была проведена оценка зависимости числовых данных БД с помощью коэффициентов корреляции и ковариации. Использовался метод corrwith() и параметр numeric\_only для указания только на числовые данные (см. рис. 12).

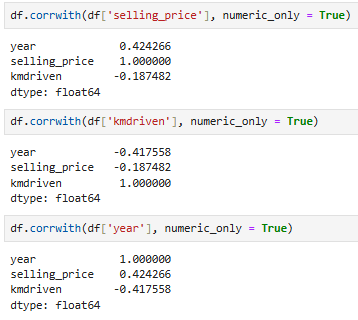


Рисунок 12 – Коэффициенты корреляции

Была замечена несильная зависимость между ценой продажи и годом выпуска, отсутствие явной зависимости между другими показателями. Зависимость между ценой продажи и годом выпуска часто объясняется несколькими факторами. Во-первых, с течением времени товары изнашиваются физически, что делает их менее привлекательными для покупателей и снижает их стоимость. Во-вторых, новые модели с улучшенными характеристиками и дизайном выходят на рынок, вызывая моральное устаревание старых версий и снижение на них спроса. Технологические инновации также играют роль: современные товары, как правило, более энергоэффективны, безопасны и экологичны, что делает старые модели менее востребованными. Кроме того, экономические факторы, такие как инфляция и рыночная конкуренция, усиливают эту тенденцию, снижая цену на старые товары в условиях высокой конкуренции и изменения покупательской способности.

Оценка была проверена с помощью коэффициентов ковариации и тепловой карты корреляции. Использовались методы cov() и heatmap() (см. рис. 13).



Рисунок 13 – Коэффициенты ковариации

Ковариация и корреляция измеряют, как две переменные изменяются относительно друг друга, но имеют ключевые отличия. Ковариация показывает направление связи (положительное или отрицательное), но её значение зависит от единиц измерения переменных, поэтому она не даёт чёткого представления о силе связи. Корреляция стандартизирована, поэтому всегда находится в диапазоне от -1 до 1, что позволяет точно оценить как направление, так и силу связи, независимо от масштаба данных. Ковариация подходит для базового анализа направления зависимости, а корреляция более удобна для интерпретации и сравнения данных. В данном случае ковариация подтвердила раннее сделанные выводы.

Для наглядности было добавлено отображение значений коэффициентов с помощью параметра annot, установленного на True (см. рис. 14).

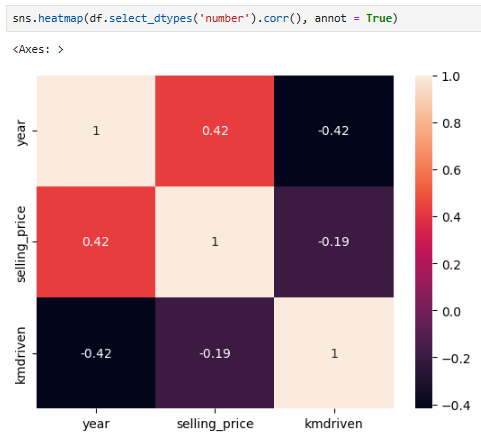


Рисунок 14 – Тепловая карта корреляции

Тепловая карта корреляции также указывает на верную интерпретацию данных. Пересечения между ценой продажи и годом выпуска подсвечены красным цветом, что подтверждает их зависимость.

Был произведен поиск взаимосвязей между данными с помощью их графического представления. Наблюдения по полученным результатам были указаны в выводе лабораторной работы.

Была построена гистограмма по группировке - owner и количество машин каждого типа коробки передач (transmission). Была импортирована библиотека matplotlib.pyplot. Использовался метод barplot из библиотеки seaborn. Методы figure(), grid(), title(), xlabel(), ylabel() из импортированной библиотеки (см. рис. 15).

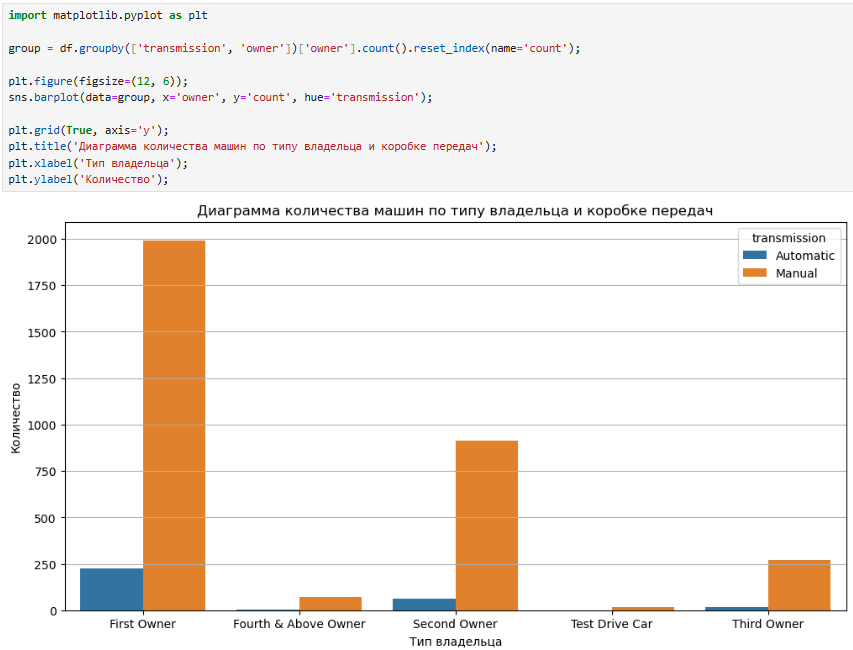


Рисунок 15 – Гистограмма по группировке

Был отфильтрован набор данных и выбран год выше 2005. Построен график зависимости по сводной таблице (pivot\_table). Было отображено среднее количество пройденных километров (kmdriven) по годам. Оставлены только маркеры в виде ★ розового цвета размером 20 с помощью plot() и методы библиотеки pandas для настройки отображения сетки и заголовков: grid(), set\_title(), set\_xlabel(), set\_ylabel(). Параметр kind отвечает за тип графика, y за определяет данные для оси Y, marker за отображения точек, color за цвет, linestyle за стиль линии, markersize за размер точек (маркеров), figsize за размер графика, legend за отображение легенды на графике (см. рис. 16).

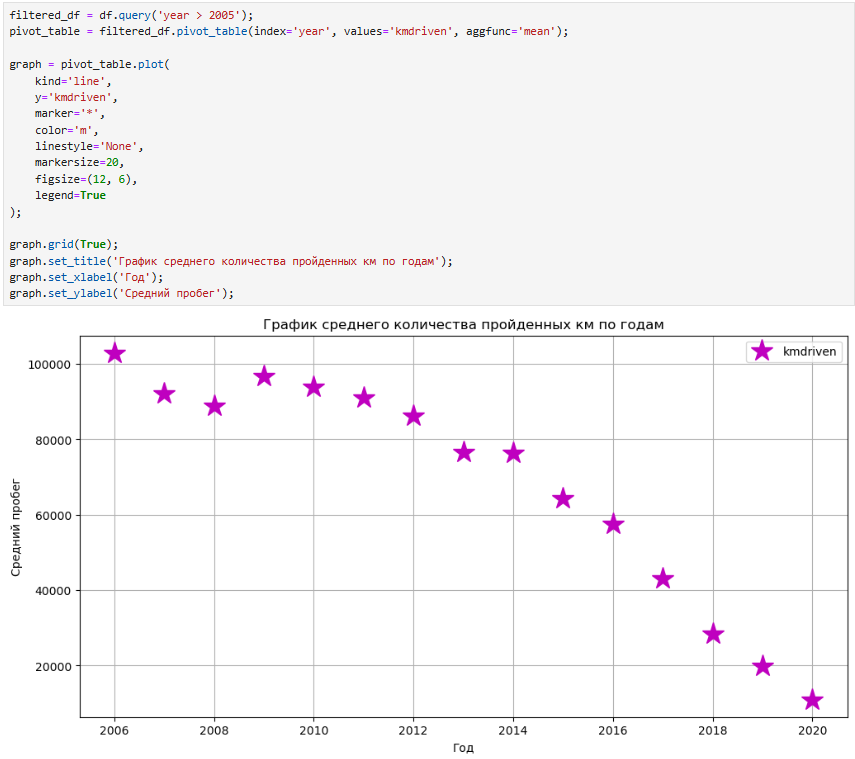


Рисунок 16 – График зависимости по сводной таблице

По выдержке данных была построена круговая диаграмма, которая отображает процент каждого типа топлива (fuel). Из диаграммы был убран тип топлива, который встречается всего 1 раз - Electric. Для более красивого отображения использовались textprops={'size': 'x-large'} и legend(fontsize = 18, bbox\_to\_anchor=(1, 1)) и размер фигуры 20 на 8. Для построения были реализованы pie() и другие методы раннее используемые библиотеки matplotlib.pyplot. Создание диаграммы происходило c параметрами: labels для названия секторов, autopct для отображения %, colors для отображения цветов (см. рис. 17).

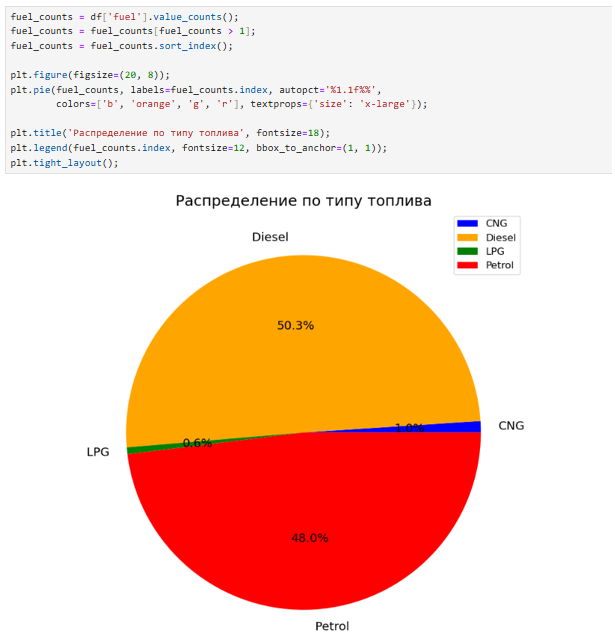


Рисунок 17 – Круговая диаграмма

Ссылка на файл:

<https://github.com/Vvvvv55531/LR2.git>

**Вывод**

В ходе выполнения лабораторной работы мной были изучены связи между признаками двумерного набора данных и произведена различного рода визуализация данных. Были улучшены навыки работы с библиотеками pandas, seaborn и matplotlib.pyplot. Реализованы диаграммы рассеяния и тепловая карта корелляции, используя plotting.scatter\_matrix() и heap() соответственно. Произведена оценка данных с помощью коэффициентов корреляции и ковариации методами corwith() и cov(). Были построены 3 вида графических представленных данных: гистограмма, график линейной зависимости и круговая диаграмм:

По гистограмме заметно преобладание дизельного топлива и бензина над всеми остальными видами топлив. Бензин и дизельное топливо могут доминировать из-за экономической выгоды их массового производства, высокой энергоемкости, которая позволяет автомобилям проходить большие расстояния на одном баке, и развитой инфраструктуры заправочных станций, делающей их доступными почти повсеместно. Химический состав этих топлив обеспечивает стабильную работу в разных климатических условиях, что делает их универсальными для различных регионов. Кроме того, автомобильная промышленность традиционно ориентирована на эти виды топлива, что поддерживает спрос и доступность

По графику линейной зависимости видно, что показатель по среднему количеству пробега падал практически с каждым следующим годом. Возможно, это происходит из-за новизны машин. У более новых автомобилей меньше времени для накопления пробега. Популярность краткосрочного лизинга также может влиять на снижение среднего пробега, так как машины возвращаются после коротких сроков эксплуатации. Кроме того, новые автомобили часто приобретаются как вторичные транспортные средства и используются реже

По круговой диаграмме видно, что количество автомобилей с механической коробкой передач преобладает среди любого типа владельцев. Автомобили с механической коробкой передач преобладают среди любого типа владельцев благодаря их экономичности и доступной стоимости, делая их привлекательными для широкого круга покупателей. Такая трансмиссия популярна среди водителей, привыкших к механике, особенно в регионах с меньшим распространением автоматических коробок передач. Механические коробки проще и дешевле в ремонте, что снижает эксплуатационные расходы и привлекает владельцев, стремящихся минимизировать затраты на обслуживание. Механика также обеспечивает больший контроль над управлением автомобилем, что ценят водители, предпочитающие активное вождение или поездки по сложным дорогам.