Дипломная работа

по теме:

«Сравнение различных библиотек для визуализации данных:

Matplotlib, **Seaborn** и **Plotly**»

Выполнил студент:
Лоскутов Д.В.

г. Звенигово
2024

СОДЕРЖАНИЕ

1. Введение

1.1. Обоснование выбора темы:

Выбор мной данной темы обуславливался тем, что я, проходя обучение в общеобразовательной школе, изучал язык программирования BASIC, на бытовых персональных компьютерах Spectrum (процессор Z80), осваивал в числе прочего графические операции SCREEN,LINE,CIRCLE и прочие. В дальнейшем пользовался графическими редакторами: ***COREL, PhotoShop, Компас3D***. Графика и процесс визуализации мне близки, а когда это становится на качественно новый уровень, с применением новейших языков программирования, выбор очевиден, кроме всего прочего в современном мире, где объём информации растёт с каждым днём, визуализация данных становится всё более значимой. Ежедневно генерируются терабайты данных из различных источников, таких как социальные сети, датчики, финансовые операции и медицинские исследования. В условиях этого информационного изобилия визуализация данных становится критически важным инструментом для анализа и интерпретации информации.

Визуализация данных позволяет аналитикам и исследователям быстро выявлять закономерности, тренды и аномалии, которые могут быть неочевидны при анализе необработанных данных. Например, графики и диаграммы могут помочь понять, как различные факторы влияют на продажи или как меняются показатели здоровья в зависимости от времени. Визуализация помогает преобразовать сложные наборы данных в более понятные форматы, что упрощает процесс анализа.

Кроме того, визуализация данных способствует улучшению коммуникации между специалистами из разных областей. Графическое представление информации позволяет легко донести идеи и выводы до коллег и заинтересованных сторон, даже если они не обладают глубокими знаниями в области анализа данных. Это особенно важно в междисциплинарных командах, где участники могут иметь разный профессиональный опыт.

Визуализация данных играет ключевую роль в процессе принятия решений. На основе визуализированных данных менеджеры могут принимать более обоснованные решения. Например, в бизнесе визуализация продаж по регионам помогает определить, где следует увеличить маркетинговые усилия или где необходимо оптимизировать запасы. В здравоохранении визуализация эпидемиологических данных может помочь в планировании ресурсов и реагировании на вспышки заболеваний.

Кроме того, визуализация данных позволяет отслеживать эффективность принятых решений. Сравнение визуализированных данных до и после внедрения изменений даёт возможность оценить их влияние и, при необходимости, скорректировать стратегию. Это создаёт цикл непрерывного улучшения, где визуализация данных становится неотъемлемой частью процесса анализа и принятия решений.

В мире анализа данных существует множество инструментов и библиотек, которые позволяют создавать наглядные и информативные графики. Среди них наиболее популярными являются ***Matplotlib, Seaborn и Plotly***. Каждая из этих библиотек имеет свои особенности и преимущества, что делает их подходящими для различных задач визуализации.

1.2.Определение цели и задач исследования

Основная цель работы: Провести сравнительный анализ библиотек визуализации данных ***Matplotlib, Seaborn*** и ***Plotly***, чтобы определить их сильные и слабые стороны, а также выбрать наиболее подходящую библиотеку для различных задач визуализации.

Задачи работы:

1. Изучить функциональность каждой из библиотек, их возможности и особенности.
2. Создать визуализации с использованием каждой библиотеки на одном и том же наборе данных для сопоставимости.

3. Провести сравнительный анализ созданных визуализаций по критериям эстетики, информативности и удобства использования.
4. Подготовить рекомендации по выбору библиотеки в зависимости от конкретных потребностей пользователя.
5. Привести примеры из различных областей.

2. Основные понятия и определения

Обзор основных понятий в области визуализации данных:

- Визуализация данных (**Data Visualization**): Процесс представления данных в графической форме для облегчения анализа и интерпретации информации.
- Библиотека визуализации (**Visualization Library**): Набор инструментов и функций, предназначенных для создания графиков и диаграмм. В данном проекте рассматриваются ***Matplotlib, Seaborn*** и ***Plotly***.
- ***Matplotlib***: Библиотека для создания статических, а также интерактивных графиков. Она обеспечивает гибкость и контроль над визуализациями.
- ***Seaborn***: Библиотека, основанная на ***Matplotlib***, которая упрощает создание статистических графиков и предлагает более эстетичные визуализации с готовыми стилями.
- ***Plotly***: Библиотека для создания интерактивных графиков, позволяющая пользователям взаимодействовать с визуализациями, что особенно полезно в веб-приложениях.
- Типы графиков (**Types of Plots**): Различные формы представления данных, такие как линейные графики, столбчатые диаграммы, диаграммы рассеяния и тепловые карты.
- Интерактивность (**Interactivity**): Способность визуализации реагировать на действия пользователя, такие как наведение курсора и щелчки.
- Эстетика графиков (Plot Aesthetics): Оформление графиков, включая шрифты, размеры, цветовые схемы и оформление осей.

- Метрики визуализации (**Visualization Metrics**): Параметры, используемые для оценки качества визуализаций, такие как читаемость и информативность.
- График (**Plot**): Визуальное представление данных, которое может принимать различные формы, такие как линии, точки, столбцы и т.д.
- Оси (**Axes**): Линии, которые определяют границы графика (обычно ось X и ось Y), на которых отображаются данные.
- Подписи **(Labels**): Текстовые метки, которые описывают оси, заголовок графика и отдельные элементы графика.
- Легенда (**Legend**): Область графика, которая объясняет, что означают различные элементы графика (например, цвета или маркеры для разных категорий данных).
- Стили (**Styles**): Наборы параметров, которые определяют внешний вид графиков, включая цвет, шрифт и толщину линий.
- Данные (**Data**): Информация, которую мы хотим визуализировать, представленная в различных форматах, таких как списки, массивы или ***DataFrame*** (в библиотеке **pandas**) - двумерная структура данных (строки, столбцы подобно таблицам в ***SQL*** или ***Excel***) **.**
- Подграфики (**Subplots**): Возможность размещения нескольких графиков в одном окне или области для сравнения различных наборов данных.
- Форматирование (**Formatting**): Настройка внешнего вида графиков, включая цвет, шрифт и стиль линий, что позволяет улучшить визуальную привлекательность.
- Экспорт (**Export**): Процесс сохранения графиков в различных форматах (например, ***PNG***, ***JPEG***, ***PDF***) для использования вне среды программирования.

- Анимация (**Animation**): Процесс создания движущихся графиков, который помогает визуализировать изменения данных во времени.
- Тепловая карта (**Heat map**): Визуализация данных, где значения представлены цветами, что позволяет быстро идентифицировать паттерны и аномалии.
- Сетчатая система (**Grid System**): Система координат, которая помогает организовать расположение графиков и элементов на графике.
- Кастомизация (**Customization**): Процесс изменения стандартных параметров графиков для достижения желаемого внешнего вида и функциональности.
- Точки данных (**Data Points**): Индивидуальные элементы данных, которые отображаются на графике.
- Трендовая линия (**Trend Line**): Линия, которая показывает общее направление данных на графике, обычно используется для выявления тенденций.
- Диапазон (**Range**): Разница между максимальным и минимальным значением данных на оси графика.

3. Обзор библиотек *Matplotlib, Seaborn и Plotly*

Библиотеки ***Matplotlib***, ***Seaborn*** и ***Plotly*** являются отличными библиотеками для построения графиков данных, но они, в основном, могут строить только статические графики. Статический график в **Python** - это график, который не изменяется и представляет собой обычную картинку. Такие графики хорошо подходят для объяснения концепций в документе, на слайдах или в презентации. В большинстве случаев для передачи информации достаточно статических графиков.

- ***Matplotlib*** является одной из самых популярных библиотек для визуализации данных в **Python**. Это универсальная библиотека, которая

работает в **Python** на **Windows**, **macOS** и **Linux**. ***Matplotlib*** используют для визуализации данных любой сложности. Библиотека позволяет строить разные варианты графиков: линейные, трёхмерные, диаграммы рассеяния и другие, а также комбинировать их. Дополнительные библиотеки позволяют расширить возможности анализа данных.

- ***Seaborn*** — это библиотека, построенная на основе ***Matplotlib*** , которая упрощает создание сложных статистических графиков, тесно интегрируется со структурами данных **pandas**. Она предлагает более высокоуровневый интерфейс и множество встроенных стилей и цветовых палитр, что делает визуализацию более эстетически привлекательной.

- ***Plotly*** — это мощная библиотека для создания интерактивных графиков, которая поддерживает различные языки программирования, включая **Python**, позволяет аналитикам сосредоточиться на интерпретации данных, а не на технических аспектах визуализации, благодаря своей простоте использования и мощным функциям. ***Plotly*** – это, по сути, онлайн-библиотека, в которой хранятся ваши визуализации данных, однако она также предоставляет автономный пакет данных, который можно использовать для рисования интерактивных графиков в автономном режиме.

3.1. Функциональность библиотек *Matplotlib, Seaborn* и *Plotly*

Matplotlib

1. Позволяет создавать практически любые виды графиков и диаграмм благодаря низкоуровневому интерфейсу.
2. Поддерживает множество типов графиков, включая линейные графики, гистограммы, 3D графики, контурные графики, круговые диаграммы и др.
3. Обширные возможности настройки всех элементов графиков, таких как оси, легенды, цвета, линии, маркеры и шрифты.
4. Возможность создания анимаций и динамических визуализаций.

5. .Поддержка сохранения графиков в различных форматах (**PNG**, **PDF**, **SVG** и прочие.).
6. Может быть сложным для новичков из-за низкоуровневого интерфейса и большого количества настроек.
7. По умолчанию графики могут выглядеть не очень привлекательно, что требует дополнительных усилий для улучшения дизайна.

Seaborn

1. ***Seaborn*** построен на основе ***Matplotlib*** и делает процесс создания графиков проще и быстрее.
2. Хорошо подходит для создания статистических графиков, таких как коробчатые диаграммы и тепловые карты.

Пример коробчатой диаграммы: (график, использующийся в статистике, компактно изображающий одномерное распределение вероятностей) на базе датасета «***Tips***» (Fig.1):

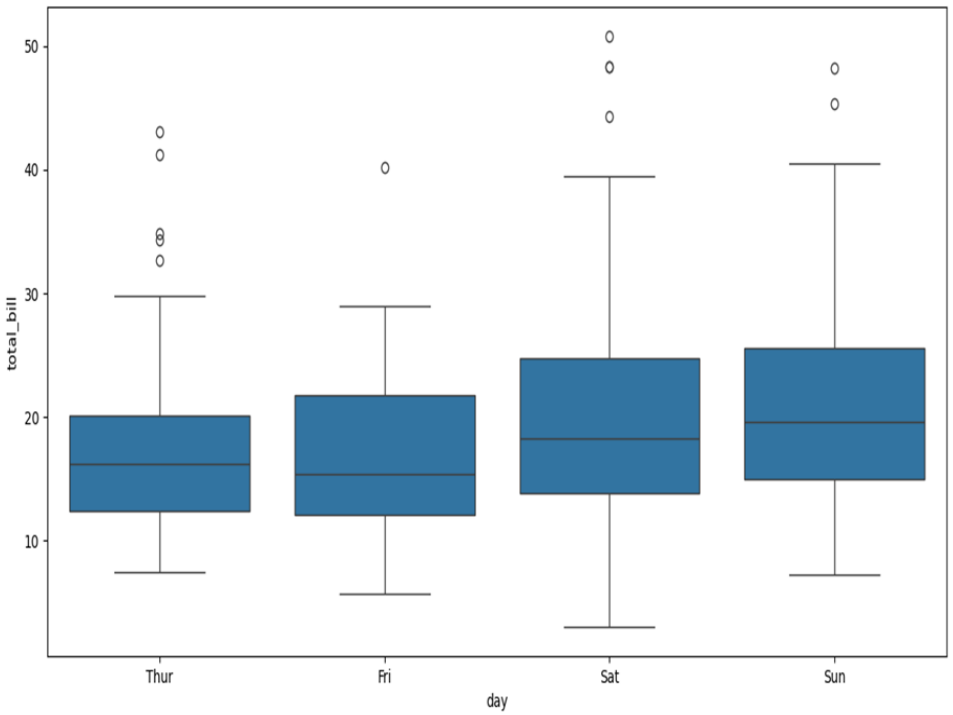
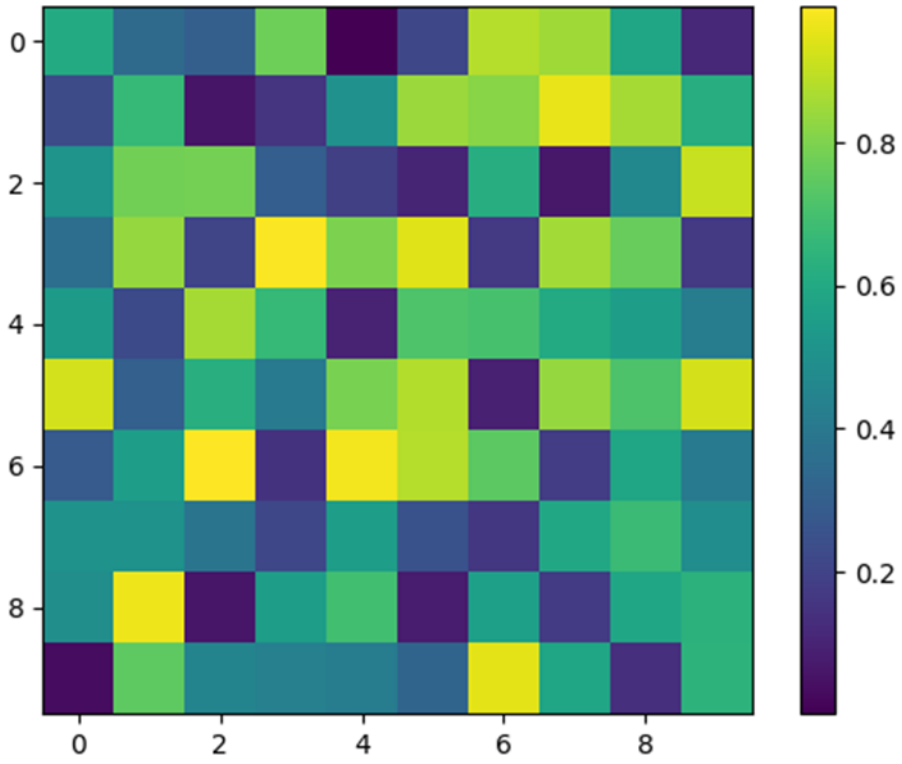


Fig 1.Общая сумма чека, распределенная по дням

Пример тепловой карты:



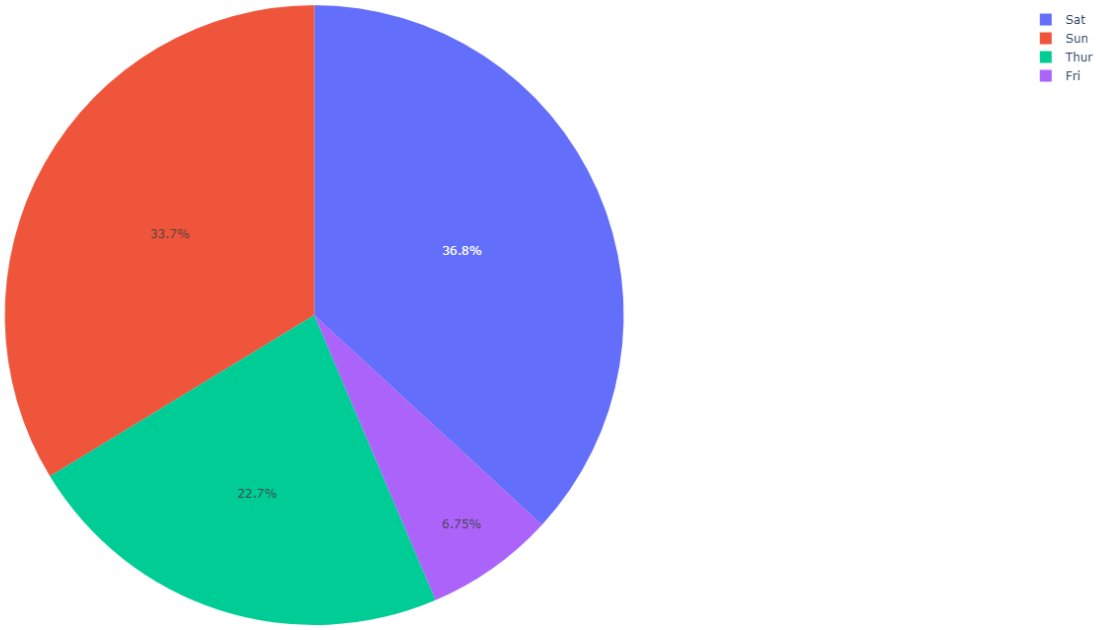
3. Имеет предустановленные стили, которые делают графики более привлекательными без лишних усилий.
4. Легко работает с данными из ***Pandas***, что упрощает анализ.
5. Комбинирует информацию о распределении данных и их плотности, что позволяет лучше понять распределение.
6. Ограничена в настройках по сравнению с ***Matplotlib***, что может быть недостатком для сложных визуализаций. Поскольку ***Seaborn*** построен на ***Matplotlib***, пользователи могут столкнуться с ограничениями, если захотят использовать функционал ***Matplotlib***, который не поддерживается в ***Seaborn***.

Plotly

1. Позволяет создавать интерактивные графики, которые можно масштабировать, перемещать и наводить курсор для получения дополнительной информации.

2. Поддерживает создание графиков, обновляющихся в реальном времени, что полезно для мониторинга данных.
3. Обширный набор графиков, включая 3D графики, карты, графики с областями, пузырьковые графики и специализированные визуализации.

Круговая диаграмма построена на базе *damacema* «**Tip**s».Показывает общую сумму чека посетителей ресторана по определенным дням.



4. Возможность сохранять и делиться графиками через облачные сервисы, такие как ***Plotly Chart Studio***.
5. Легко интегрируется в веб-приложения и поддерживает создание дашбордов.
6. Интерактивные графики могут требовать больше ресурсов, что может быть проблемой для больших наборов данных.
7. Для достижения более сложных визуализаций может потребоваться больше времени на изучение и настройку.

8. Некоторые функции могут требовать подписки на платные версии ***Plotly***, что может быть ограничивающим фактором для некоторых пользователей.

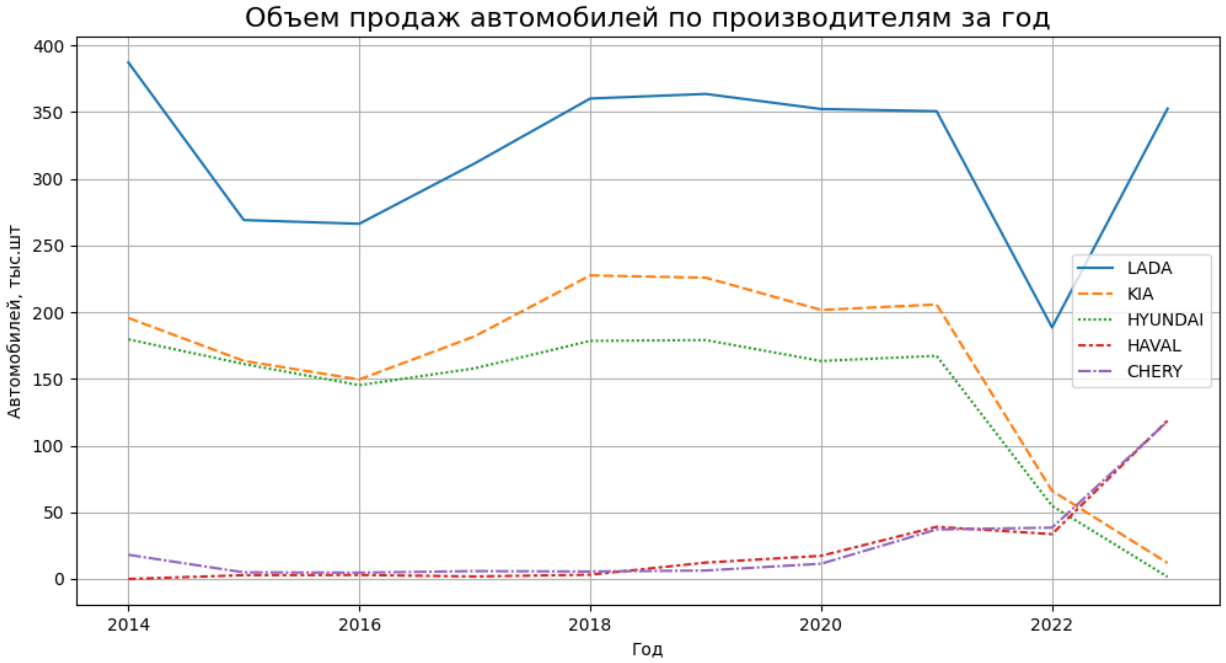
4. Сравнение библиотек *Matplotlib, Seaborn* и *Plotly* на примере датасет* «autosales»

***** датасет «autosales.csv» создан мной на основе данных сайта www.chinamobile.ru

Matplotlib

Ниже представлен линейный график (**lineplot**) объемов продаж автомобилей производителей по годам . По оси ординат назначено число продаж автомобилей различных производителей, по оси абсцисс - год продаж.

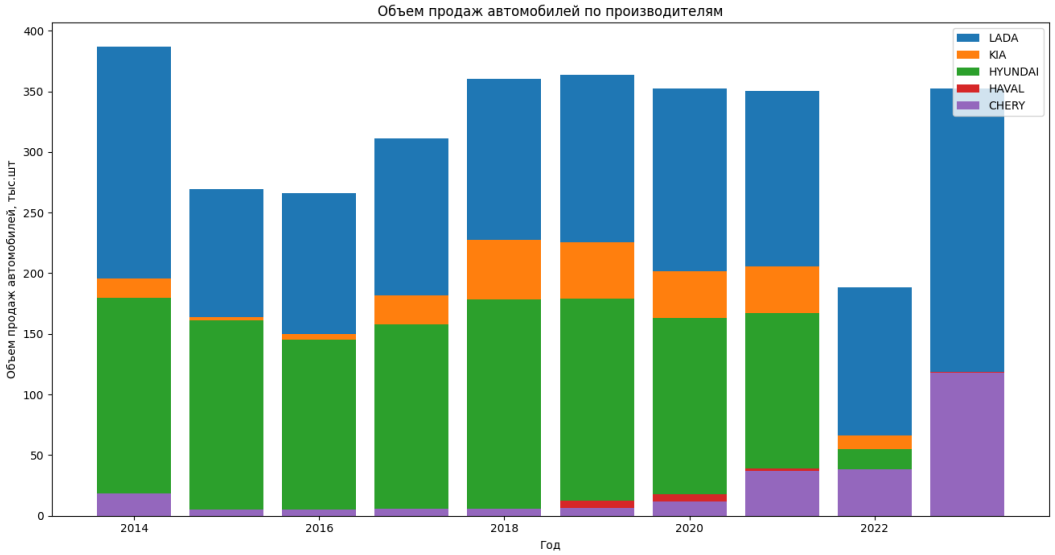
Хотя это и самый элементарный тип графика, он даёт возможность определить направление изменения показателей и обнаружить ключевые моменты роста или снижения.



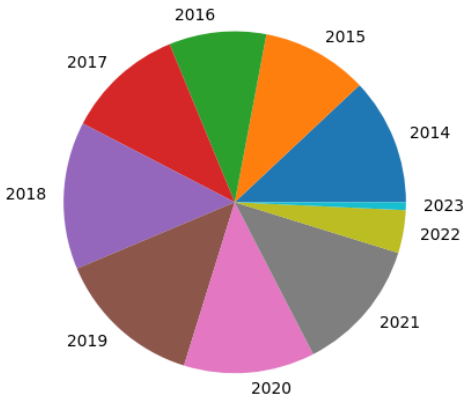
На графике можно определить снижение продаж автомобилей брендов KIA и HYUNDAI на фоне санкционной политики, в связи с началом Специальной Военной Операции. Виден и общий «провал» продаж

автомобилей (2022 г.) в связи с вышеуказанными событиями. На графике так же виден заход на рынок производителей легковых автомобилей из КНР.

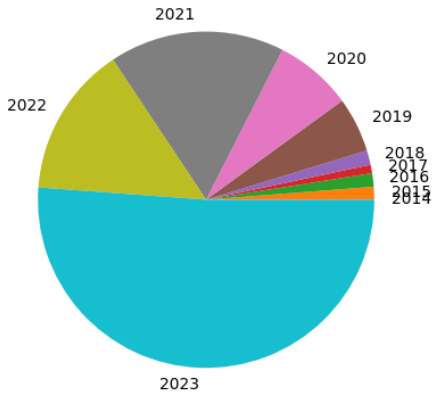
Гистограмма построена на основе тех же данных, что и приведенный выше линейный график.



и круговая диаграмма KIA и HAVAL:



KIA



HAVAL

Seaborn

Линейный график (darkgrid) с применением библиотеки Seaborn практически повторяет график, выполненный в ***Matplotlib***, т. к. он построен на основе библиотеки ***Matplotlib***.

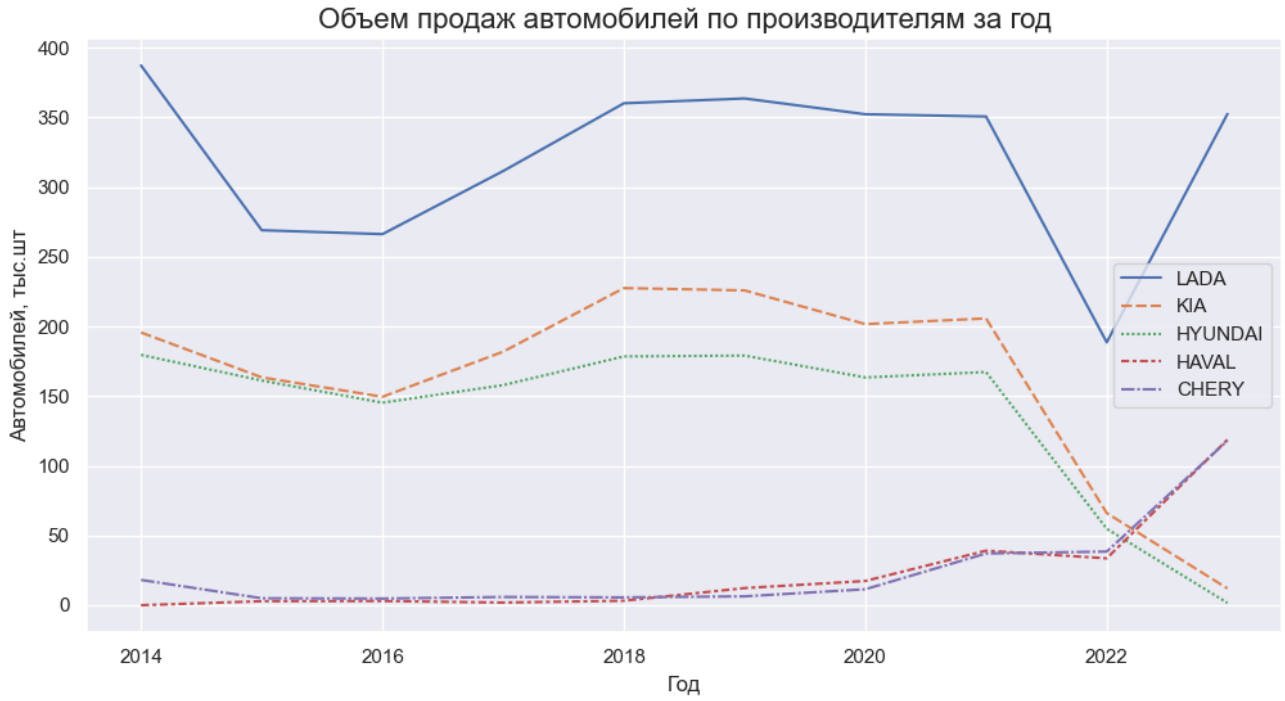
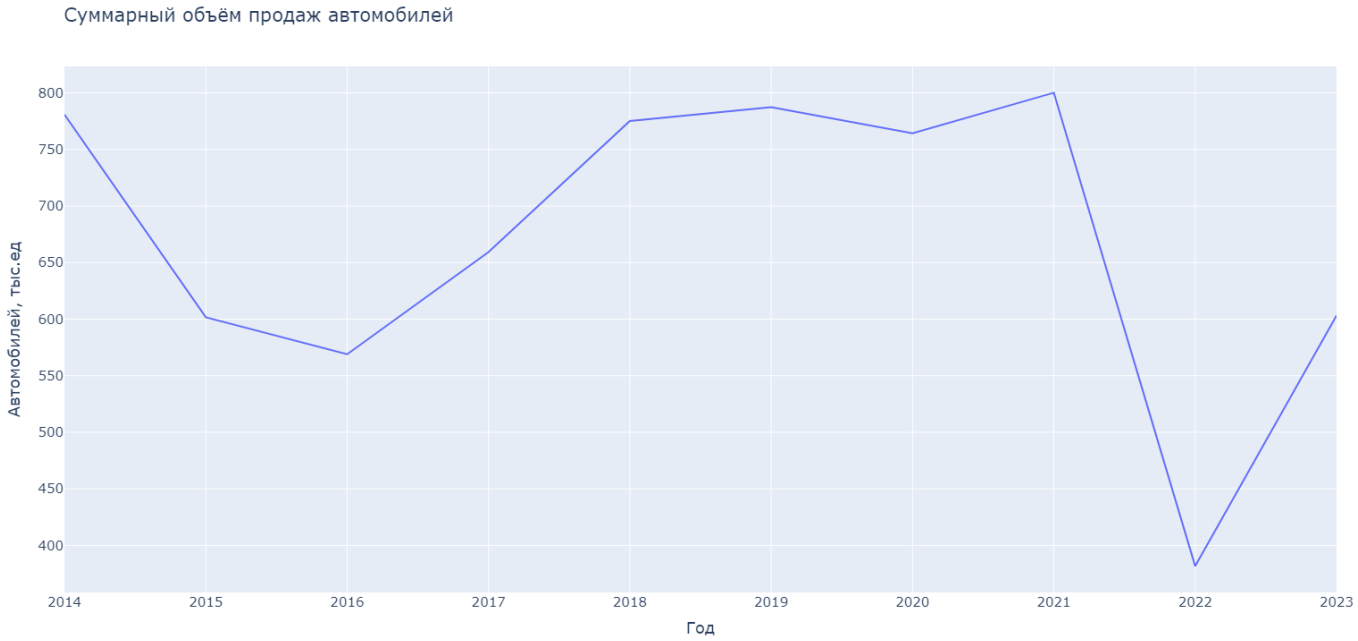


Диаграмма рассеяния с применением библиотеки Seaborn :

Plotly

Линейный график, выполненный с применением библиотеки ***Plotly*** визуально не выглядит информативным и интуитивно доступным, как графики в ***Matplotlib*** и ***Seaborn***, где графики изображены более динамично и понятно, но его выгодным отличием является интерактивность. (Здесь тоже видно как «просел» рынок продаж автомобилей в 2022 году).



5. Результат сравнения.

Характеристика	***Matplotlib***	***Seaborn***	***Plotly***
Тип графиков	Статические и интерактивные	Статистические	Интерактивные, 3D, карты
Легкость использования	Средняя (может быть сложной для новичков)	Высокая (интуитивно понятный интерфейс)	Высокая (простота для создания интерактивных графиков)
Эстетика	Низкая по умолчанию, требует настройки	Высокая (предустановленные стили и палитры)	Высокая (интерактивные и визуально привлекательные)
Кастомизация	Высокая (гибкость в настройках)	Средняя (ограничена по сравнению с ***Matplotlib***)	Средняя (достаточно для большинства задач)
Интерактивность	Ограниченная (в основном статические графики)	Ограниченная (статические графики)	Высокая (возможность взаимодействия с графиками)
Поддержка 3D	Да (но требует дополнительных усилий)	Нет	Да
Работа с данными	Хорошая интеграция с **NumPy** и **Pandas**	Отличная интеграция с **Pandas**	Хорошая интеграция с **Pandas** и другими источниками
Анимация	Да	Нет	Да
Экспорт	Поддержка различных форматов (**PNG**, **PDF**, **SVG**)	Поддержка форматов через ***Matplotlib***	Поддержка экспорта в **HTML** и облачные сервисы
Области применения	Научные исследования, инженерия, отчеты	Научные исследования, анализ данных	Бизнес-аналитика, веб-разработка, интерактивные отчеты
Преимущества	Гибкость, мощные инструменты для визуализации	Упрощение создания красивых статистических графиков	Интерактивность, возможность работы с большими данными
Недостатки	Сложность для новичков, требует настройки	Ограниченные возможности кастомизации	Требует больше ресурсов, некоторые функции платные

6. Рекомендации по выбору библиотек:

Matplotlib

- Если нужны высококачественные статические графики для научных публикаций, отчетов или презентаций.
- Настройка каждого аспекта графики, включая цвета, шрифты, размеры осей, линии и маркеры.
- При обработке больших наборов данных и визуализации без значительных затрат на производительность.
- Если необходимо сделать сложные графики, например, многоуровневые графики, графики с несколькими осями или графики с дополнительными пояснениями.
- Если вы работаете с библиотеками для анализа данных, такими как ***NumPy*** и ***Pandas***.
- Поддерживает создание анимаций, что позволяет визуализировать изменения данных во времени и делать графики более динамичными и информативными.

Seaborn

- Удобные функции для создания статистических графиков, таких как распределения, корреляционные матрицы и категориальные графики. Он автоматически обрабатывает статистические параметры и предоставляет красивые стили по умолчанию.
- Хорошо интегрируется с ***Pandas DataFrame***, что упрощает визуализацию данных. ***Pandas*** позволяет эффективно обрабатывать и манипулировать большими наборами данных.
- Предлагает множество встроенных тем и палитр, которые упрощают настройку графиков и делают их более привлекательными без необходимости вручную настраивать каждый элемент.

- Предоставляет мощные инструменты для сравнения групп, такие как **facetgrid**, который позволяет разбивать данные на подгруппы и визуализировать их в одном графике.
- Имеет простой и интуитивно понятный синтаксис, что делает его доступным для пользователей, которые только начинают работать с визуализацией данных.

Plotly:

- Предоставляет мощные инструменты для создания интерактивных графиков, позволяя пользователям взаимодействовать с визуализациями, масштабировать, наводить курсор для получения дополнительной информации и переключаться между различными представлениями данных.
- Эффективно обрабатывает большие наборы данных и позволяет визуализировать их без значительного снижения производительности. Это особенно полезно для анализа больших объемов данных, таких как временные ряды или многомерные наборы данных.
- Поддерживает множество типов графиков, включая 3D-графики, графики с несколькими осями и комбинированные графики. Это позволяет создавать сложные визуализации, которые могут быть полезны для глубокого анализа данных.
- Хорошо интегрируется с веб-приложениями и позволяет легко встраивать графики в HTML-страницы, что делает его отличным выбором для веб-визуализации и создания интерактивных дашбордов.
- Предоставляет удобные инструменты для создания графиков, которые позволяют легко сравнивать разные наборы данных, например, через наложение графиков или использование под графиков.
- Предлагает мощные инструменты для создания карт и визуализации геоданных, что делает его отличным выбором для проектов, связанных с географическим анализом.

7. Примеры применения в различных областях

Образование:

Учебные заведения активно используют визуализацию данных для анализа успеваемости студентов и оценки эффективности образовательных программ. Например, с помощью библиотек, таких как ***Matplotlib*** и ***Seaborn***, можно создавать графики, показывающие распределение оценок по предметам. Это помогает выявить слабые места в учебных планах и определить, какие темы требуют дополнительного внимания. Кроме того, дашборды могут включать визуализации, отображающие динамику успеваемости студентов по семестрам, что позволяет преподавателям и администраторам принимать обоснованные решения о необходимости изменения подходов к обучению.

Финансовые услуги:

В финансовом секторе визуализация данных играет ключевую роль в анализе рыночных тенденций и оценке инвестиционных портфелей. Финансовые аналитики используют интерактивные графики, созданные с помощью ***Plotly,*** для отображения изменений цен акций, объемов торгов и других финансовых показателей в реальном времени. Например, можно создать многослойный график, который показывает, как различные факторы, такие как экономические отчеты или изменения в политике, влияют на цены акций. Это помогает инвесторам принимать более информированные решения и своевременно реагировать на изменения на рынке.

Торговля и ритейл:

В ритейле визуализация данных используется для анализа покупательского поведения и оптимизации товарных запасов. Ритейлеры применяют тепловые карты и диаграммы для отображения данных о продажах по регионам и времени. Например, визуализация может показывать, какие товары наиболее популярны в разных магазинах или в определенные сезоны, что позволяет компаниям адаптировать свои стратегии маркетинга и управления запасами.

Также можно использовать графики для анализа эффективности рекламных акций, сравнивая продажи до и после кампании.

Спорт и фитнес:

В спортивной аналитике визуализация данных помогает тренерам и спортсменам анализировать производительность и прогресс тренировок. С помощью библиотек, таких как ***Matplotlib*** и ***Plotly***, можно создавать графики, отображающие результаты тестов на выносливость, силу и другие показатели. Например, тренеры могут визуализировать изменения в результатах спортсменов на протяжении сезона, что позволяет выявить тренды и корректировать тренировочные программы. Также, анализируя данные о физических нагрузках и восстановлении, можно оптимизировать режим тренировок для достижения максимальных результатов.

Государственное управление:

Государственные учреждения используют визуализацию данных для анализа социальных и экономических показателей, таких как уровень безработицы, доступ к образованию и здравоохранению. С помощью интерактивных карт и графиков можно визуализировать распределение ресурсов и демографические изменения в разных регионах. Например, карты могут показывать уровень безработицы по округам, что помогает в планировании программ поддержки и улучшения экономической ситуации. Визуализация данных также может быть использована для мониторинга эффективности государственных программ и оценки их воздействия на общество.