Тема дипломной работы
«Сравнение библиотек для визуализации данных»

Восприятие информации в нонешнем калейдоскопе событий это сложный и многогранный процесс, зависящий как от физиологических, так и от психологических факторов, а также от контекста, в котором информация представляется.

Исследования показывают, что от 55% (согласно Альберту Мейерсу) и до 90% (согласно Генри К. Линдгрену, 1962) человек воспринимает через зрение.

Визуализация данных - это удобный и быстрый способ оценить данные в виде разнообразных зависимостей, представленных в виде точек, графиков, диаграмм, схем, деревьев и особенно 3D объектов. Кроме аналитиков, работающих с данными (Data Science), визуализация данных также полезна для анализа и поиска зависимостей перед обучением нейросетей (Machine Learning).

Python обладает довольно впечатляющим количеством пакетов (далее - объединенных в библиотеки для решения определенных задач) из самых разных областей человеческой деятельности. В официальном Python хранилище PyPI (www.pypi.org) зарегистрировано более 590 тысяч пакетов.
По разным оценкам проектов, связанных с визуализацией данных, существует от 35 самых известных (https://www.awesomepython.org/?c=viz) и до 500 небольших пакетов (https://www.pythonrepo.com/search?q=visualizations) .

Обзор проекта

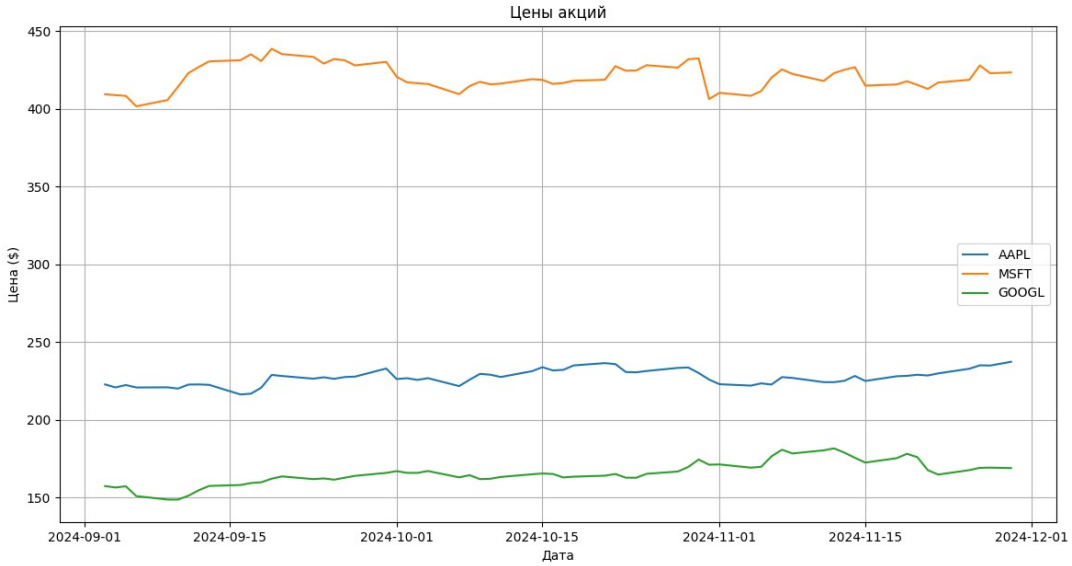
В проекте изучаются и реализуются возможности эффективной визуализации данных. Предлагается посмотреть на результаты и сравнить код для первых трех библиотек визуализации данных из списка: Matplotlib, Seaborn, Plotly.

Matplotlib

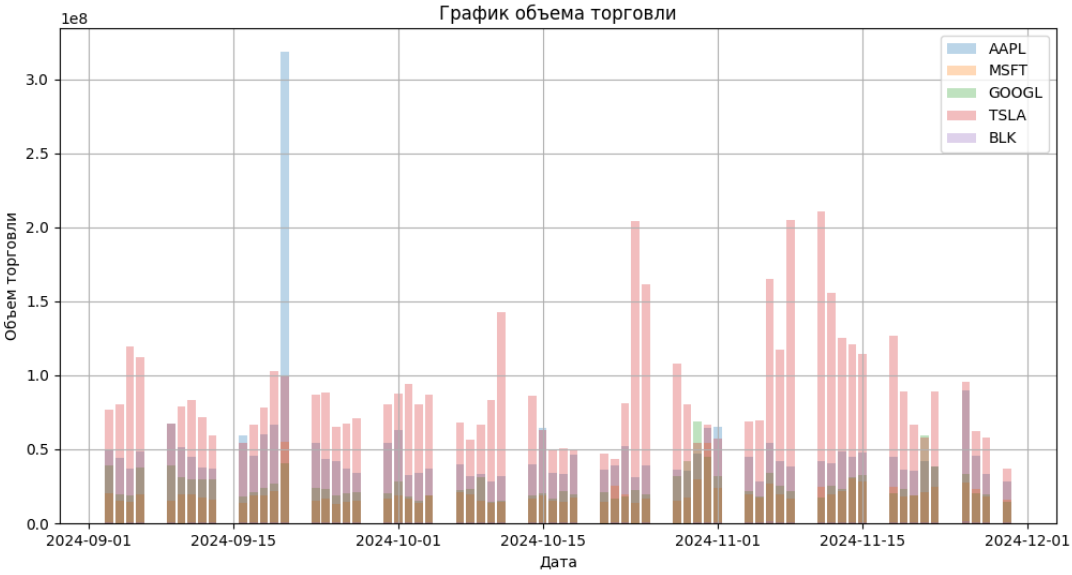
Matplotlib это базовый вариант библиотеки для визуализации данных, поддерживает широкий спектр типов графиков, таких как линейные графики, диаграммы рассеяния, столбчатые диаграммы, гистограммы, круговые диаграммы, полосы погрешностей, коробчатые диаграммы; позволяет создавать несколько подплотов в рамках одного рисунка, что позволяет проводить сложные визуальные сравнения.
Хорошо интегрируется с другими научными библиотеками, такими как NumPy, Pandas.

Ниже представлены следующие скриншоты из рабочего кода:

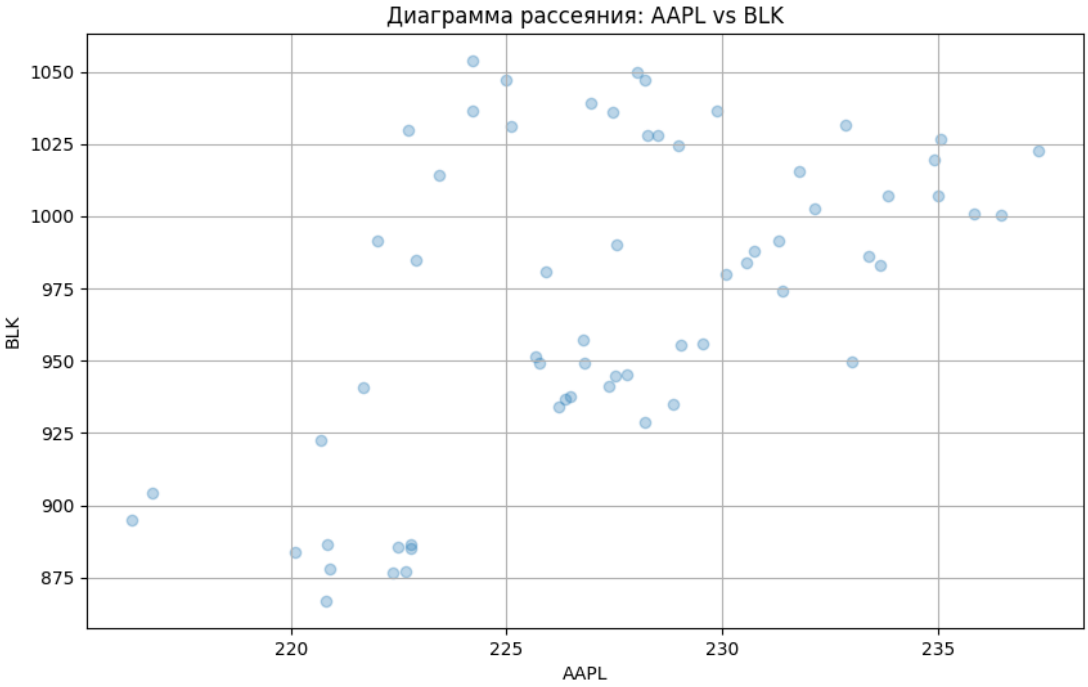
Matplotlib: Линейный график (line plot):
график для отображения акций на бирже.



Matplotlib: Диаграмма стобчатая (bar chart),
в данном случае показывает объемы торгов на бирже



Matplotlib: Диаграммы рассеяния (scatter plot):
сравнения количественных значений между различными категориями.

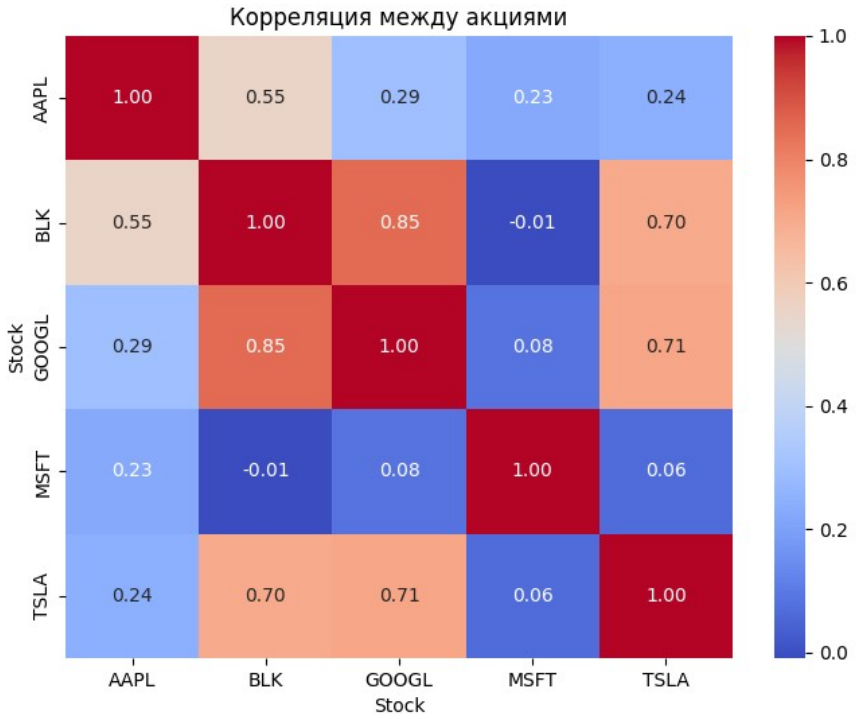


Seaborn

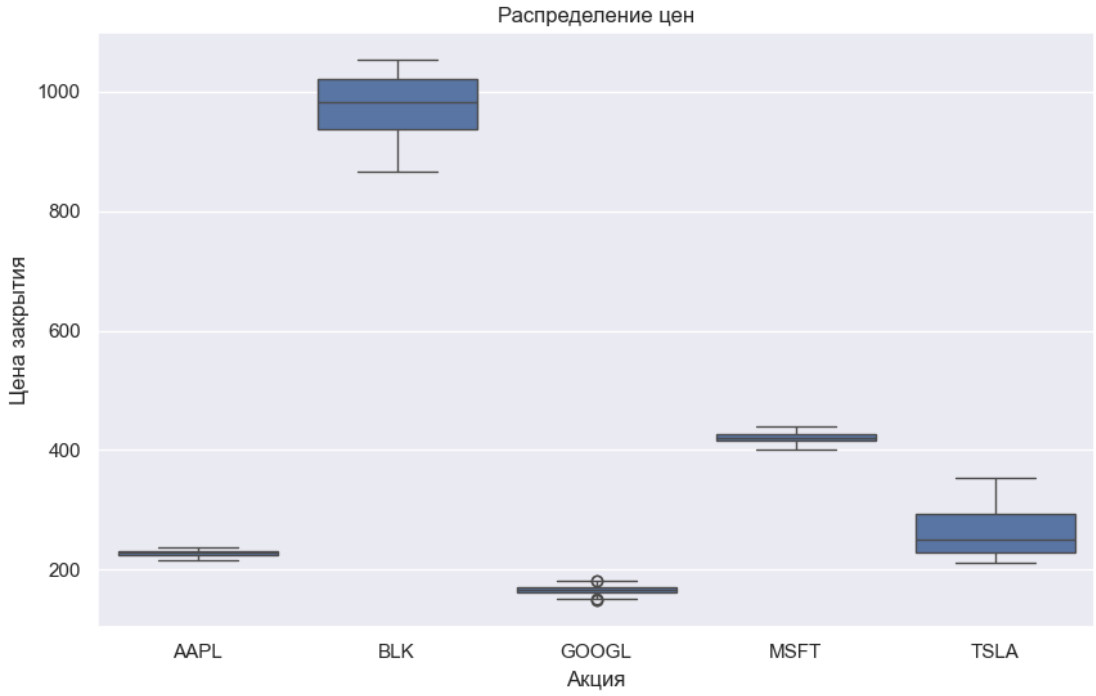
высокоуровневая библиотека для визуализации данных, основанная на Matplotlib. Seaborn предоставляет более простой и красивый интерфейс для создания сложных графиков, упрощает процесс создания привлекательных и информативных статистических графиков, при этом позволяя пользователям использовать функциональность Matplotlib.
В библиотеке Seaborn есть более 20 наборов данных(dataset).

В проекте представлены следующие примеры:

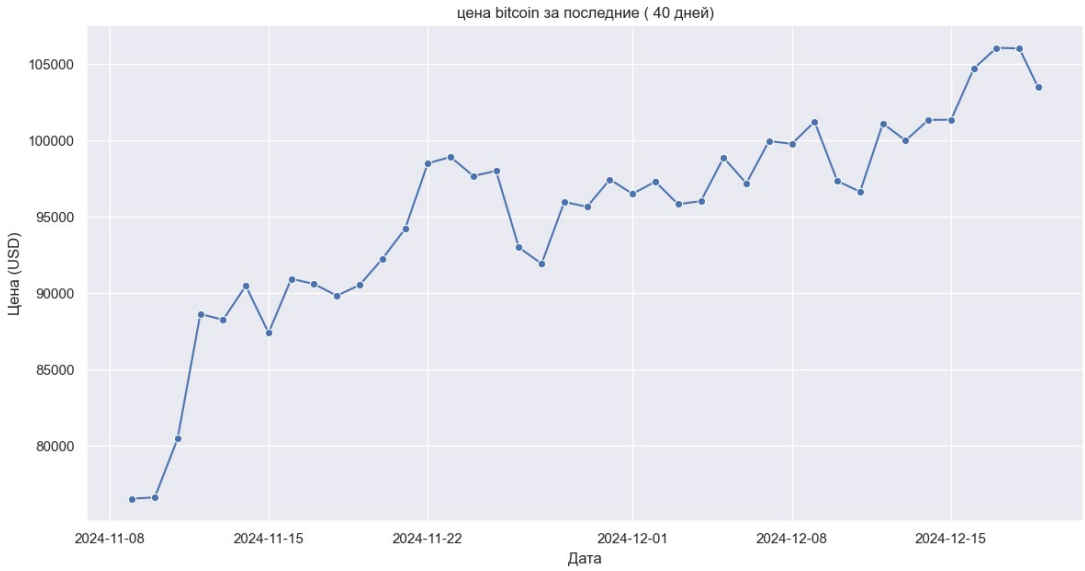
Seaborn : тепловая карта (heatmap):
Позволяют визуализировать зависимости числовых данных.



Seaborn : Диаграмма размаха:
“Ящик с усами” способ группировки данных часто применяемый на финансовых рынках.



Seaborn : Линейный график (line plot): курс Bitcoin’а за последние 40 дней.



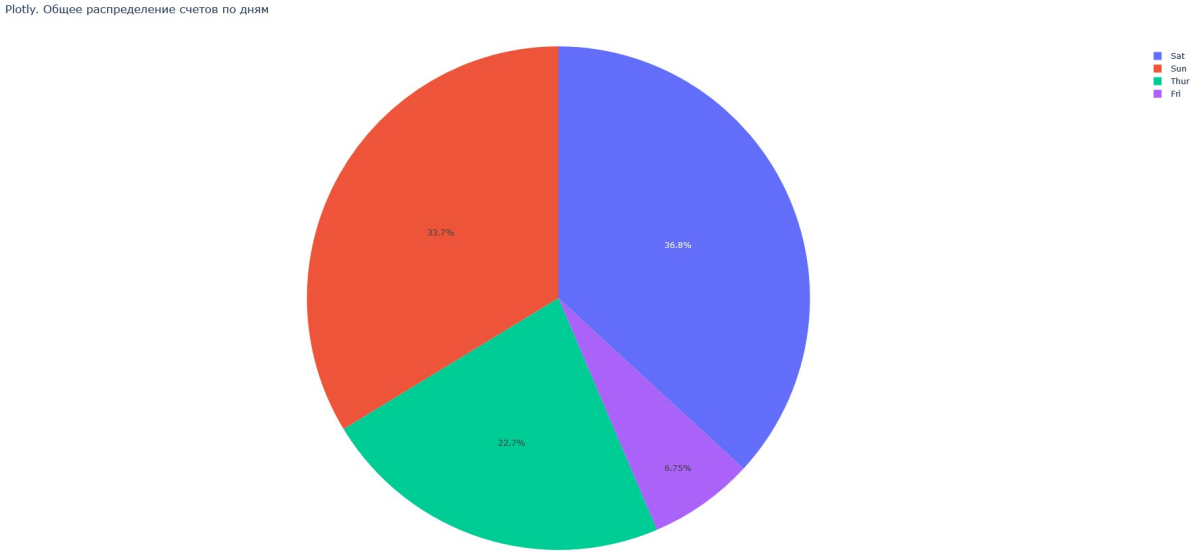
Plotly

мощная библиотека для создания интерактивных веб-графиков, отлично справляется с интерактивными визуализациями, подходящими для веб-приложений и приборных панелей, что делает его идеальным для бизнес-аналитики, но требует больше системных ресурсов.

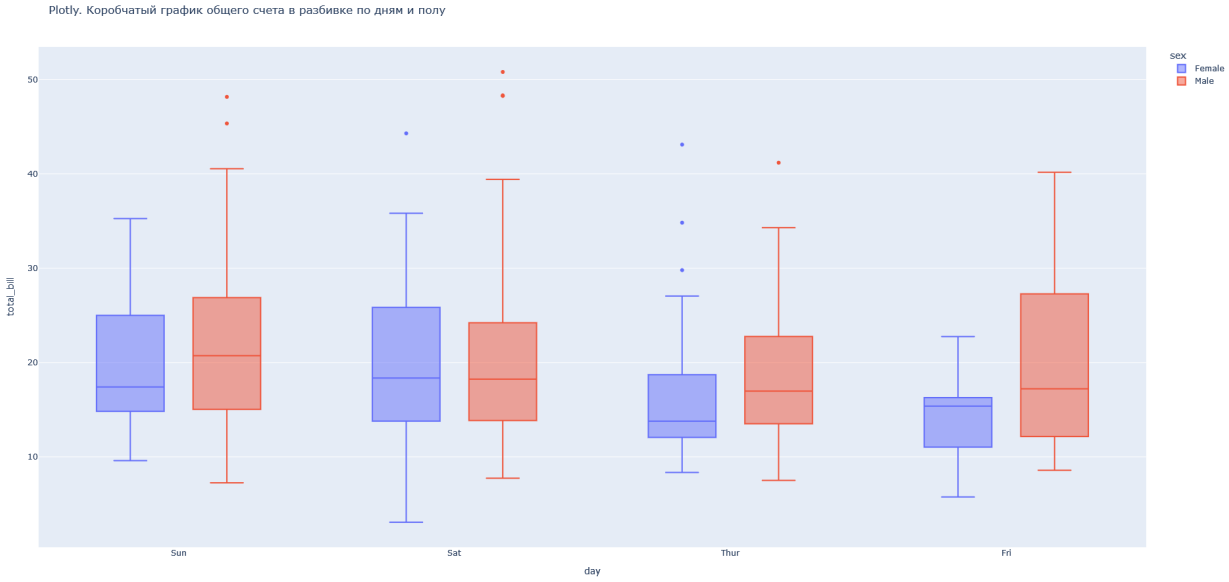
В проекте представлены следующие примеры:

Базовый линейный график (line plot):
Простой интерактивный график для анализа зависимостей.

Круговые диаграммы (pie chart):
Визуализация пропорций различных категорий.

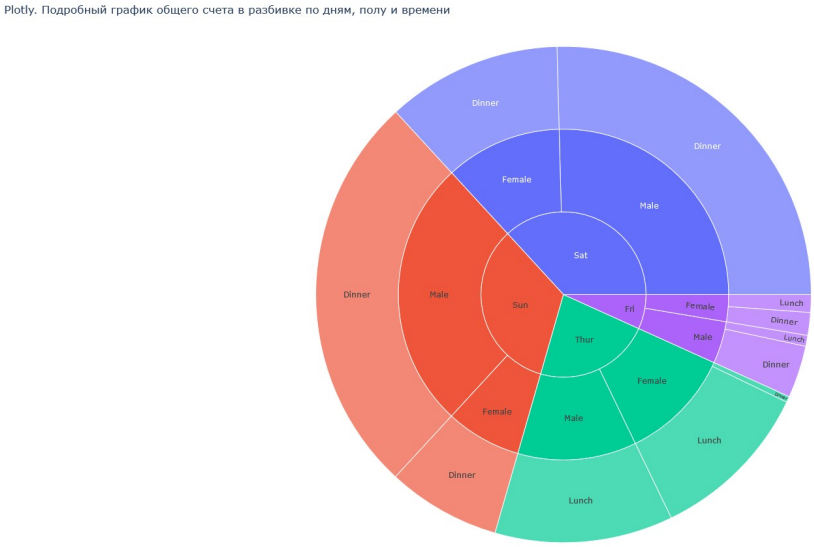


Коробчатые графики (box plot):
Анализ распределения данных с учетом выбросов.

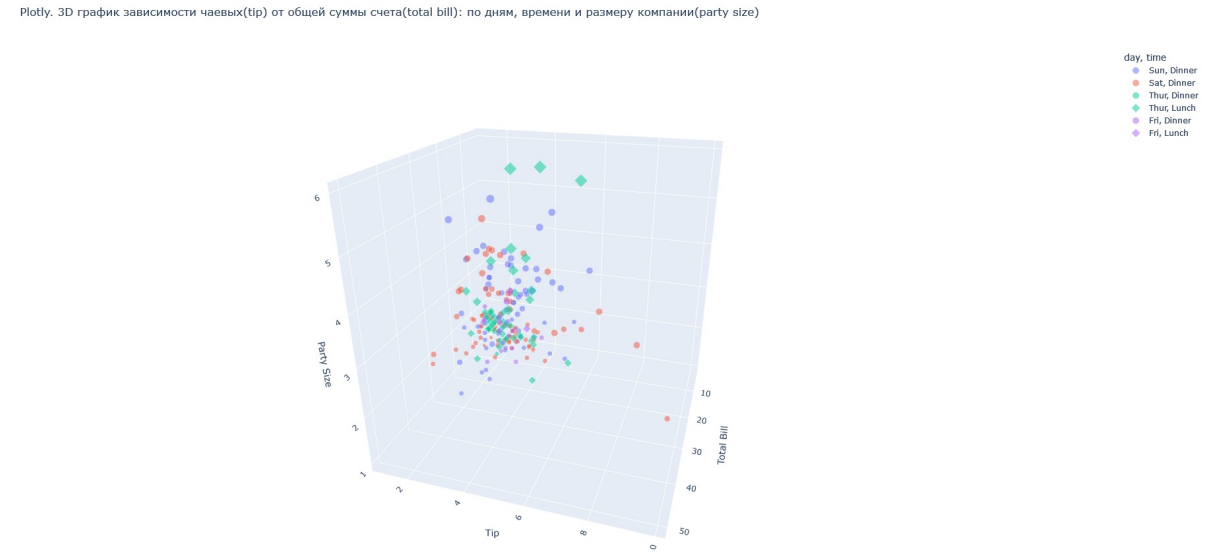


Диаграммы рассеяния (scatter plot):
Интерактивные графики для анализа взаимосвязей между переменными.

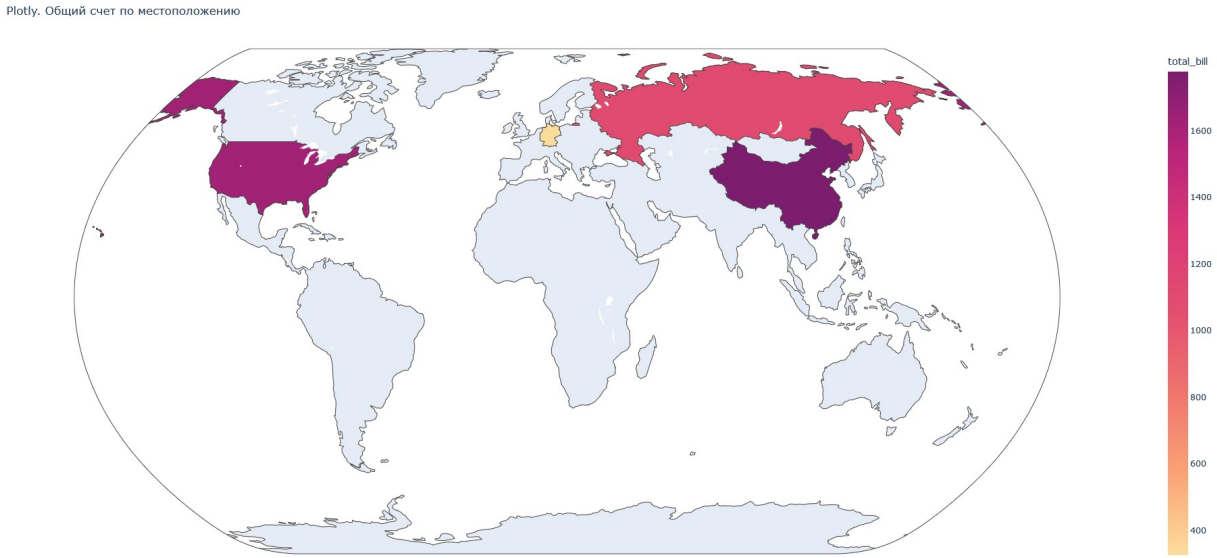
Подробная круговая диаграмма (sunburst chart) общей суммы в разбивке по деталям: дням, полу и времени



Трехмерная интерактивная диаграмма рассеяния (3D scatter plot) эффектный 3D график, показывающий зависимость чаевых (tip) от общей суммы счета (total bill), разбитый по данным из датасета: времени и размеру компании



Карта хороплет (choropleth map):
Визуализация географических данных на карте, в 13 строчках кода.



Структура проекта

Проект включает следующие компоненты:
1_matplotlib.py
2_seaborn.py
3_plotly.py
requirements.txt

Заключение

Проект демонстрирует большой спектр возможностей для визуализации данных с использованием трех библиотек Python.

Matplotlib
предоставляет набор мощных базовых инструментов, требующих существенной настройки для создания эффектных графиков. Производительность Matplotlib обычно быстрее для простых графиков из-за своего низкоуровневого характера.

Seaborn
упрощает создание более сложных и красивых визуализаций:
функции Seaborn "ориентированы на данные", т.е. они понимают структуру вашего набора данных и могут автоматически извлекать метки осей, легенды и другие элементы графика; в Matplotlib эти элементы необходимо явно определять. Встроенная поддержка доверительных

интервалов автоматически рассчитывается и отображаются, что в Matplotlib потребовало бы ручного расчета и построения.
Seaborn может быть медленнее для больших наборов данных из-за своих высокоуровневых абстракций и дополнительных вычислений.

Plotly
позволяет создавать интерактивные веб-графики, но не подойдет для больших наборов данных.

Большой набор библиотек для визуализации в Python поможет пользователям выбрать подходящий инструмент для своих задач в области визуализации данных.

Дополнительная сравнительная таблица:

	Основные характеристики	Сильные стороны	Слабые строны	Используют для
Matplotlib	Широкие возможности построения графиков, поддержка различных бэкендов, высокая настраиваемость графики	Высокая степень кастомизации, отлично подходит для 2D-графиков, является основой для других библиотек, таких как Seaborn	Более сложная кривая обучения для сложных графиков, настройки по умолчанию могут быть менее визуально привлекательными	Базовое черчение, научное черчение, пользовательские визуализации
Seaborn	Построен на базе Matplotlib, высокоуровневый интерфейс для построения привлекательных статистических графиков, поддержка визуализации категориальных данных	Простой синтаксис, красивые стили по умолчанию, встроенные темы	Ограничен статистической графикой, меньше контроля, чем в Matplotlib	Визуализация статистических данных, исследование наборов данных
Plotly	Интерактивные графики, веб-визуализация, поддержка широкого спектра типов графиков, включая 3D	Высокая интерактивность, легкий обмен визуальными изображениями, хорошо подходит для приборных панелей	Может быть медленнее для больших наборов данных, требует больше системных ресурсов	Интерактивные информационные панели, веб-приложения, бизнес-аналитика
Bokeh	Библиотека интерактивных графиков, ориентирована на современные веб-браузеры, может работать с большими наборами данных	Интерактивные визуализации, удобные для работы в Интернете, хорошо подходят для потоковых данных	Сложнее в настройке, меньшее сообщество, чем у Matplotlib	Интерактивные веб-приложения, Визуализация больших наборов данных
ggplot	Основана на грамматике графики, интуитивно понятный синтаксис, построение сложных графиков из простых компонентов	Интуитивный и декларативный стиль, отлично подходит для создания сложных визуализаций	Менее гибкий для определенных типов графиков, все еще развивается как библиотека	Статистические визуализации, Журналистика данных
Altair	Декларативная библиотека статистической визуализации, хорошо интегрируется с Pandas, основана на Vega и Vega-Lite	Чистый синтаксис, сильная интеграция с Jupyter, хорошо подходит для исследовательского анализа данных	Ограниченная интерактивность, не может работать с очень большими наборами данных	Эксплораторный анализ данных, Быстрые визуализации
	Высокоуровневая	Фокусируется на типах и семантике данных,	Требует понимания типов данных	Многомерные

Holoviews	библиотека для легкого построения сложных визуализаций, хорошо интегрируется с Bokeh, может выводиться на различные бэкенды	хорошо справляется с большими массивами данных		визуализации данных, визуализация больших наборов данных
Pygal	Библиотека для построения графиков на основе SVG, легкая и простая в использовании, хорошо подходит для простых и быстрых визуализаций	Интерактивный вывод, хорошо подходит для веб-контекста, прост в освоении	Ограниченная интерактивность по сравнению с другими, менее гибкая	Простые диаграммы для веб-приложений, быстрые сводки данных

Приложение 1.
Версия Python 3.9.13

Список необходимых библиотек:
pip install seaborn
pip install plotly
pip install statsmodels
pip install mplcursors
pip install mpl-interactions
pip install yfinance
pip install requests

Автор:
Сергей Корниенко
2024.12.18