Курс «Базовая обработка данных на языке Python»

автор: Киреев В.С., к.т.н., доцент

Лабораторная работа № 4

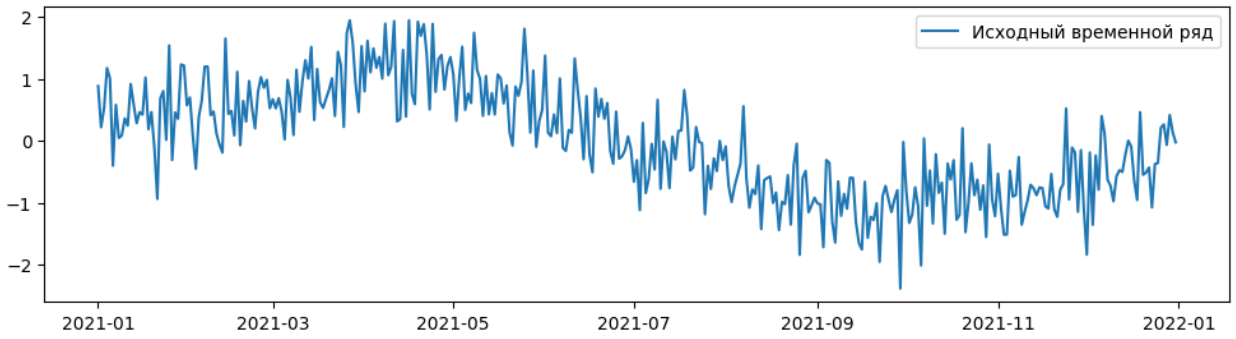
Тема: «Работа с временными рядами в Python. Подготовка данных, декомпозиция. Использование библиотеки statsmodels»

Цель работы: изучить основы работы с временными рядами в Python.

Теоретическая справка

Определение временного ряда

Временной ряд - это последовательность точек данных, собранных, записанных или измеренных через последовательные, равномерно распределенные временные интервалы. Каждая точка данных представляет собой наблюдения или измерения, проводимые в течение определенного времени, например, цены на акции, показания температуры или показатели продаж.



Данные временных рядов часто представляют в виде линейного графика, где время изображается на горизонтальной оси x, а значения переменной - на вертикальной оси y. Такое графическое представление облегчает визуализацию тенденций, закономерностей и колебаний переменной во времени, помогая в анализе и интерпретации данных.

Компоненты временного ряда

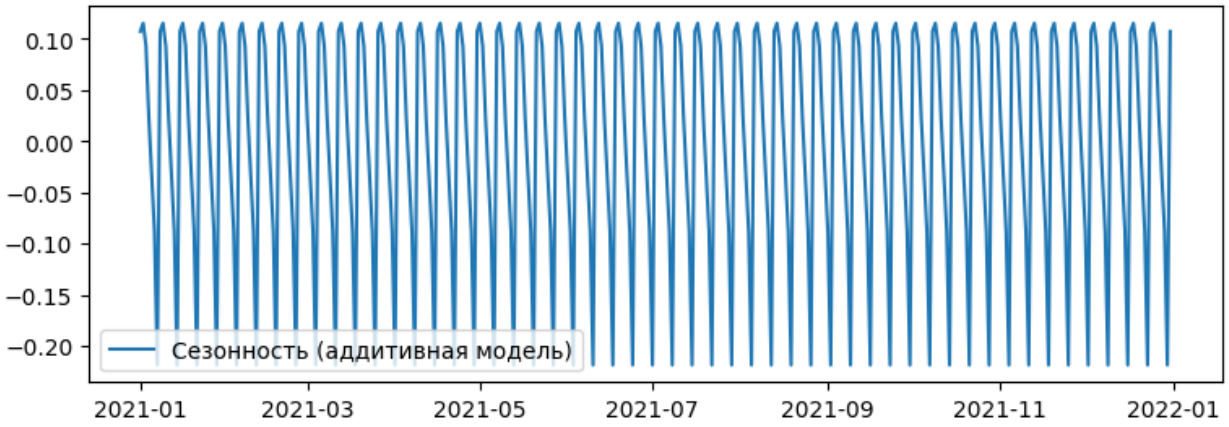
Существует четыре основных компоненты временного ряда :

Тренд : Тренд представляет собой долгосрочное движение или направленность данных с течением времени. Он фиксирует общую тенденцию ряда к увеличению, уменьшению или сохранению стабильности. Тренды могут быть линейными, указывающими на последовательное увеличение или уменьшение, или нелинейными, показывающими более сложные закономерности.

Для выделения тренда можно использовать сглаживание, например суммирование значений ряда за какой-то значимый временной период:

```
from statsmodels.tsa.seasonal import seasonal_decompose
result_add = seasonal_decompose(ts, model='additive')
```

Сезонность : Сезонность относится к периодическим колебаниям или моделям, которые происходят через регулярные интервалы во временном ряду. Эти циклы часто повторяются ежегодно, ежеквартально, ежемесячно или еженедельно и обычно зависят от таких факторов, как времена года, праздники или деловые циклы.



Циклические колебания: Циклические колебания — это долгосрочные колебания во временном ряду, которые не имеют фиксированного периода, как сезонность. Эти колебания представляют собой экономические или деловые циклы, которые могут длиться несколько лет и часто связаны с расширениями и сокращениями экономической активности.

Нерегулярность (или шум): Нерегулярность, также известная как шум или случайность, относится к непредсказуемым или случайным колебаниям в данных, которые нельзя отнести к тренду, сезонности или циклическим колебаниям. Эти колебания могут быть результатом случайных событий, ошибок измерений или других непредвиденных факторов. Нерегулярность затрудняет выявление и моделирование базовых закономерностей в данных временных рядов.

Виды разложений временных рядов:

- При аддитивном разложении временной ряд выражается в виде суммы его компонентов: Он подходит, когда величина сезонности не меняется в зависимости от величины временного ряда.
- При мультипликативном разложении временной ряд выражается как произведение его компонентов: Это целесообразно, когда величина сезонности масштабируется с величиной временного ряда.

Библиотека statsmodels

Библиотека Statsmodels используется для оценки и интерпретации различных статистических моделей на Python. Модуль tsa в Statsmodels фокусируется на анализе временных рядов и предоставляет инструменты для декомпозиции временных рядов, подбора моделей и проведения статистических тестов для данных временных рядов. Он широко используется для прогнозирования временных рядов, эконометрики и статистического анализа временных данных.

```
ts.resample('2W').sum()
```

Модель скользящего среднего (Moving Average, MA)

Модели скользящего среднего — это тип модели анализа временных рядов , обычно используемый в эконометрике для прогнозирования тенденций и понимания закономерностей в данных временных рядов. В моделях скользящего среднего текущее значение временного ряда зависит от линейной комбинации прошлых ошибок белого шума временного ряда. В анализе временных рядов скользящее среднее обозначается буквой « q », которая представляет порядок модели скользящего среднего, или, говоря простыми словами, мы можем сказать, что текущее значение временного ряда будет зависеть от прошлых ошибок q.

Самостоятельное задание

1. Загрузите файлы из предоставленного архива в единый датафрейм вида: DATE, SBER, LKOH, GAZP,MOEX, USD0. Где значения в последних 5-ти колонках являются значениями курса акций LOW (минимальные значения за день), а значения DATE – датами с типом TIMESTAMP. Используйте модуль **OS.**
2. Сделайте DATE – индексом, переиндексируйте на полный период (для этого нужно восстановить пропущенные даты) и восстановите пропущенные значения в каждом ряде.
3. Визуализировать линейными графиками в одной фигуре для рисования все ряды. Используйте метод **MATPLOTLIB.PYPLOT.SUBPLOTS**.
4. Нормализовать временные ряды и убедиться на визуализации что они стали сопоставимы. Используйте деление значений ряда на базовое значение (самое раннее значение в ряду).
5. Удалить или сгладить аномалии (выбросы) во временных рядах. Используйте логарифмирование.
6. Удалить тренд и визуализировать линейным графиком. Используйте метод **DIFF** или **SHIFT**.
7. Удалить сезонность в каждом ряде и визуализировать линейным графиком. Используйте скользящее среднее и метод **ROLLING**.
8. Выделить тренд, сезонность, остатки, с помощью аддитивной модели: используя **STATSMODELS.TSA.SEASONAL_DECOMPOSE**
9. Рассчитать корреляцию между всеми временными рядами (**CORR**).
10. Построить кластерную карту (**SEABORN.CLUSTERMAP**) для п.9. вида: