Направление Data Analyst, компания P&G

Добро пожаловать на виртуальную стажировку компании Procter & Gamble! Предлагаем тебе примерить роль аналитика данных в IT-подразделении международной компании сектора FMCG[1], где объединяются бизнес, технологии и инновации.

Мы рекомендуем выполнять задания в указанном порядке, так как они взаимосвязаны между собой и объединены общей темой анализа временных рядов.

Выполнение всего блока заданий займет у тебя не более 60-80 минут.

По результатам выполнения заданий ты научишься следующему:

1. Анализировать тренды и сезонность, работая с временными рядами.
2. Исследовать дата-сет на периодичность и строить график автокорреляции.
3. Использовать простые средства моделирования наподобие ARIMA[2] для прогнозирования временных рядов.

Рекомендуемый тайминг:

1. 10-15 минут на первое задание.
2. 10-15 минут на второе задание.
3. 35-40 минут на третье задание.

Информация о загрузке решения:

Данный проект содержит несколько подзадач. Можно загрузить файл, содержащий решение только части заданий, но по возможности старайся сделать их все.

Желаем удачи!

Дата-сет

Для выполнения заданий виртуальной стажировки предлагаем воспользоваться открытым дата-сетом shampoo_sales.csv, содержащим данные о ежемесячном объеме продаж шампуня за трехлетний период (всего 36 наблюдений).

[1] FMCG (fast -moving consumer goods) — товары повседневного спроса, включающие продукты легкой и пищевой промышленности, а также косметику, предметы личной гигиены, моющие средства и пр.
[2] ARIMA (аббревиатура от AutoRegressive Integrated Moving Average) — одна из популярных экстраполяционных моделей.

Задание 1. Определение трендов и сезонности во временных рядах

Сегодня в качестве Data Analyst компании P&G тебе предстоит проанализировать тренды и сезонность, работая с временными рядами. Утром ты получил письмо от руководителя с инструкцией по выполнению задания.

Привет!

Мы столкнулись с необходимостью анализа данных по продажам наших шампуней. Прежде чем использовать реальные данные, мы просим тебя научиться анализировать открытый дата-сет shampoo_sales.csv на наличие определенных трендов и сезонности в представленном временном ряде.
Мы уже прочитали файл с открытыми данными и построили динамику продаж шампуня в зависимости от месяца.

Код:

```
import csv
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# читаем input-файл
df = pd.read_csv('shampoo_sales.csv')
sales = df[['Sales']]
sales.plot(figsize=(12,10),
linewidth=5, fontsize=20)
plt.xlabel('Month', fontsize=20)
plt.ylabel('Sales per month',
fontsize=20)
plt.show()
```

График:

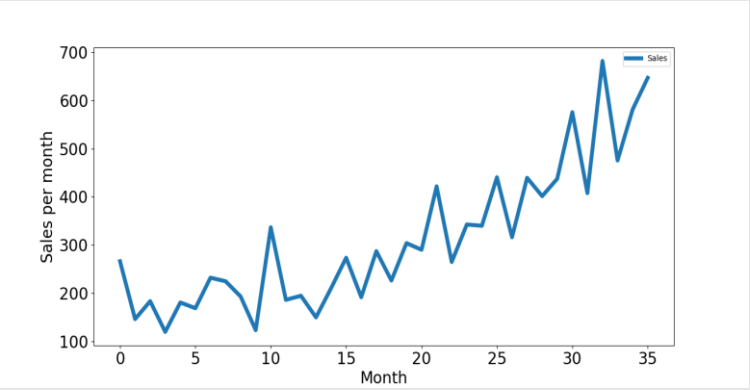

Рисунок 1. Зависимость объемов продаж шампуня от времени

Твоя задача — **дополнить код и получить еще два графика, которые нужны для анализа трендов и сезонности.**
Существует несколько способов определения трендов во временных рядах, один из которых связан с использованием скользящего среднего (a rolling average). Это когда мы берем для каждой точки временного ряда среднее значение точек по обе стороны от нее. Такой подход позволяет сглаживать шум и влияние сезонности, выделять определенный тренд в данных, если он есть.
Что касается поиска сезонных закономерностей в данных, то, наоборот, нужно удалить тренд временного ряда, чтобы было легче поймать сезонность. Для графического представления сезонности можно находить разность между последовательными точками данных, называемую разностью первого порядка (first-order difference).

Hints. В Pandas есть функции rolling и diff, которые ты должен использовать для построения графиков. В коде комментарием укажи, такой тренд наблюдается для данных по продаже шампуня, если выбрать размер окна равным шести, а также сделай вывод про сезонность.

Ждем твоего решения. Наша IT-команда очень рассчитывает на тебя ☺

Полезные материалы

Статья об анализе временных рядов: Анализ временных рядов - тренд, сезонность, шум - Электронный учебник K-tree (k-tree.ru).

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

Задание 2. Исследование периодичности и построение графика автокорреляции

После успешного завершения анализа трендов и сезонности временных рядов перед тобой стоит задача построить график автокорреляции.

На почте ты обнаружил новое письмо от IT-команды.

> Добрый день!
>
> Твоя помощь с анализом трендов и сезонности была очень своевременной.
>
> Теперь нам нужно погрузиться в вопрос корреляции временного ряда. Мы просим тебя **построить график автокорреляции[3] для объема продаж шампуней**, когда по оси x отложены значения продаж, а по оси y — то, как ряд коррелировал с самим собой при наличии лага[4]. Нам важно посмотреть, для каких лагов у нас идет положительная корреляция, а когда начинается негативная.
>
> Hints. Можешь добавить в свой код, использованный для первого задания, функцию autocorrelation_plot (разберись, как сделать это правильно, и не забудь про визуализацию plt.show()). Также прямо в коде укажи комментарием, когда у нас положительная, а когда отрицательная корреляция.
>
> Спасибо!

Полезные материалы

Статья о том, что такое автокорреляция: Нежное введение в автокорреляцию и частичную автокорреляцию (machinelearningmastery.ru).

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

[3] Автокорреляция элементов временного ряда — корреляционная зависимость между последовательными элементами временного ряда.
[4] Лаг — число периодов, по которым рассчитывается коэффициент автокорреляции между парами элементов ряда.

Задание 3. Построй модель скользящего прогноза ARIMA

Поздравляем, ты справляешься с работой аналитика данных!

Тем временем на почте появилось новое письмо с инструкцией, что делать дальше.

```
Привет!

В качестве последнего задания на сегодня в роли аналитика данных в P&G тебе предстоит построить модель скользящего прогноза ARIMA.

ARIMA (аббревиатура от AutoRegressive Integrated Moving Average) широко используется в качестве статистического метода прогнозирования временных рядов. В принципе, ARIMA подходит для прогнозирования, но в таком случае придется добавлять много спецификаций при обращении к функции predict.

Хорошей альтернативой станет модификация модели – скользящий прогноз ARIMA (rolling forecast ARIMA model), когда дата-сет разделен на обучающий и тестовый наборы данных, и последний из них используется для генерации прогноза.

Мы хотим, чтобы ты создал новый или дополнил уже использованный код и построил модель скользящего прогноза ARIMA, при этом 60% данных составили бы обучающий набор, а оставшиеся 40% – тестовый. Не забудь вывести график, сравнивающий реальные данные из тестовой выборки с теми, что были спрогнозированы моделью.

Hints. Можно разделить дата-сет на обучающий и тестовый наборы следующим образом:
size = int(len(sales) * 0.6)
train_set, test_set = sales.values[0:size], sales.values[size:len(sales)]
history = [x for x in train_set]
predictions = list()

Успехов!
```

Полезные материалы

- Статья об ARIMA: Модель ARIMA демонстрация в Python - pythobyte.com,
- Видео из онлайн-курса об использовании ARIMA: Прогноз значения показателей деятельности компании методом Arima (finoko.ru).

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива, содержащего все необходимые файлы.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.