МИНОБРНАУКИ РОССИИ
федеральное государственное автономное образовательное
учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
Институт дополнительного образования
Высшая инженерная школа

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«Прогнозирование продаж с помощью адаптивных методов»

по программе профессиональной переподготовки:
«Анализ данных на языке Python»

Выполнила: Пилюганов В.А
Руководитель: к.э.н Заграновская А.В

Санкт-Петербург
2022

Актуальность исследования

В современных условиях, когда компании имеет накопленные исторические данные по продажам становится актуальна задача прогнозирования временного ряда на основе его исторических значений.

Область исследования

Объект исследования: Объем продаж бытовой техники в компании

Предмет исследования: Адаптивные методы прогнозирования

Этапы работы

- первичный анализ временных рядов, выявление аномалий, выявление тренда ;
- прогноз объема продаж с использованием следующих методов – тренд сезонная модель, модель с фиктивными переменными, модель Хольта – Уинтерса, модель Тейла-Вейджа, модель SARIMA;
- оценка полученных прогнозных значений;
- выбор подходящей модели

Исходные данные

```
[ ]  data = pd.read_csv('/content/data.csv') #загрузка файла
     data.head() #вывожу на экран первые пять строк
```

	Date	USD
0	2022-04-24 00:00:00	1165.1715
1	2022-04-23 00:00:00	689.4045
2	2022-04-22 00:00:00	523017.0405
3	2022-04-21 00:00:00	800922.1275
4	2022-04-20 00:00:00	204632.6085

```
data.dtypes
```

```
Date    datetime64[ns]
USD            float64
dtype: object
```

```
[46] data.shape
```

```
(3299, 2)
```

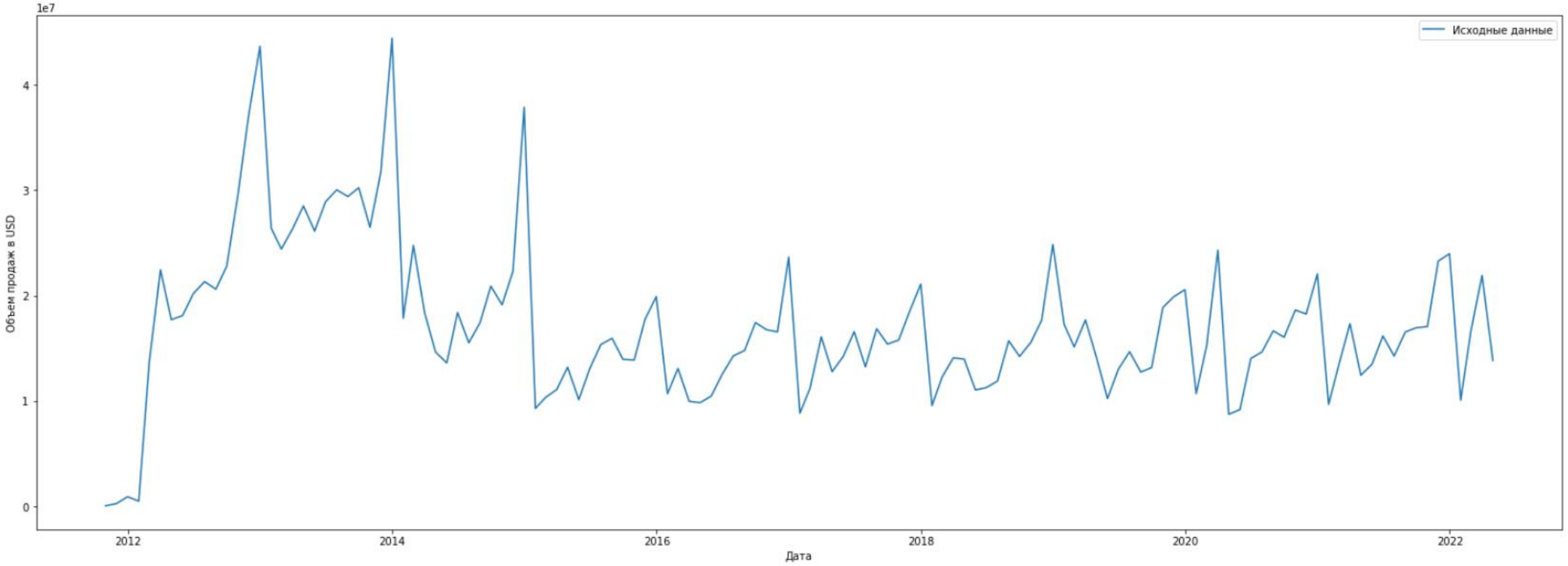
```
data.index.min()
```

```
Timestamp('2011-10-14 00:00:00')
```

```
data.index.max()
```

```
Timestamp('2022-04-24 00:00:00')
```

Предобработка

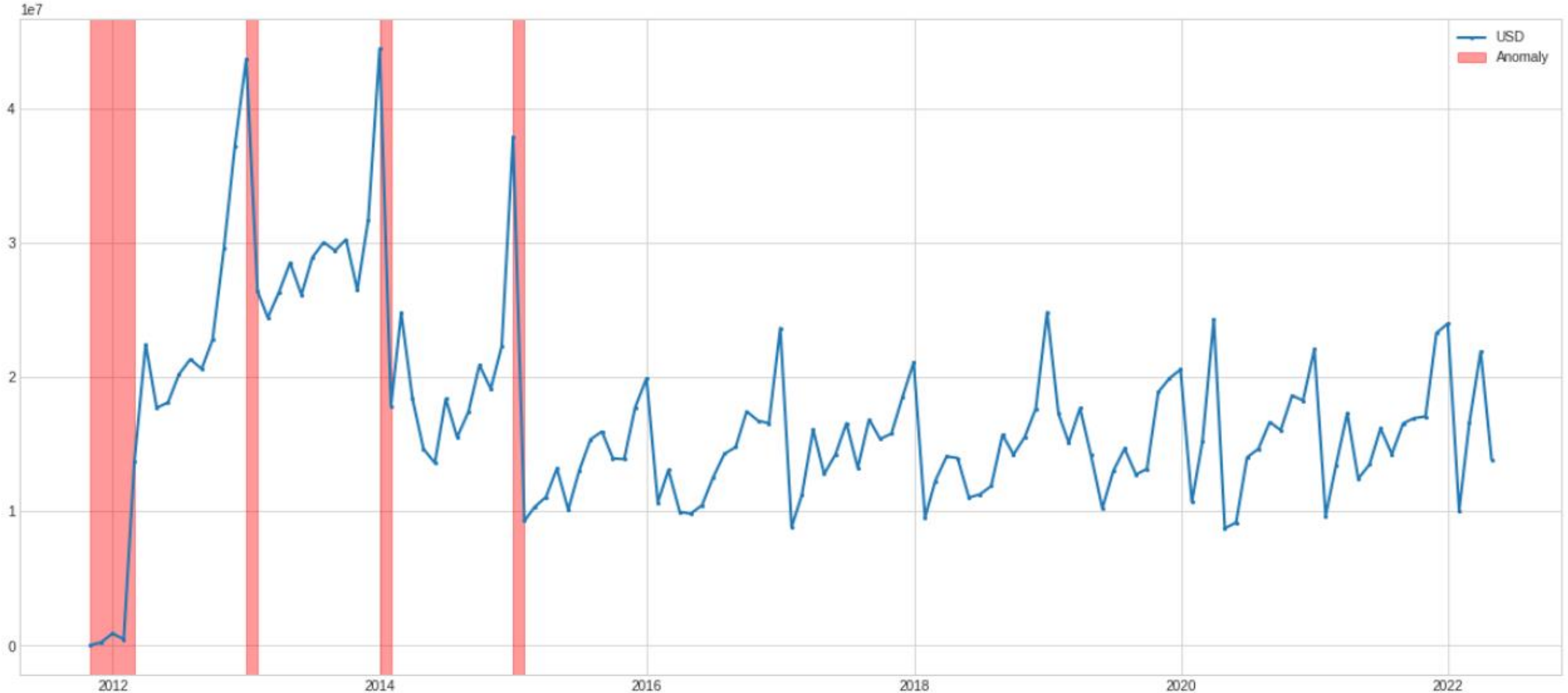



```
data = data.resample('M').sum()
```

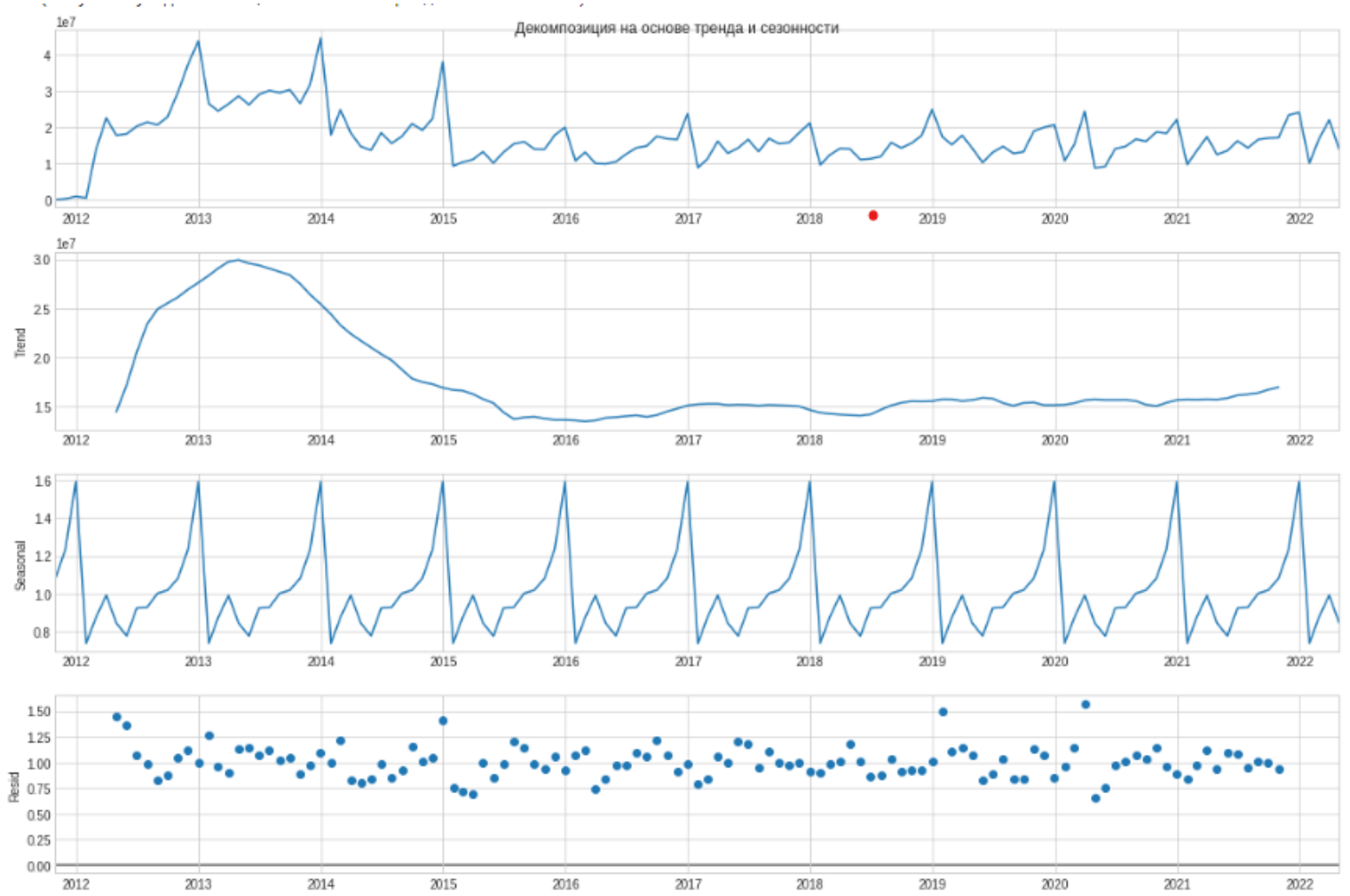
```
В исследуемом датасете:
строк - 88
столбцов - 1
Минимальная дата - 2015-01-31 00:00:00
Максимальная дата - 2015-01-31 00:00:00
```

Первичный анализ

Выявление аномалий



Декомпозиция на основе тренда и сезонности



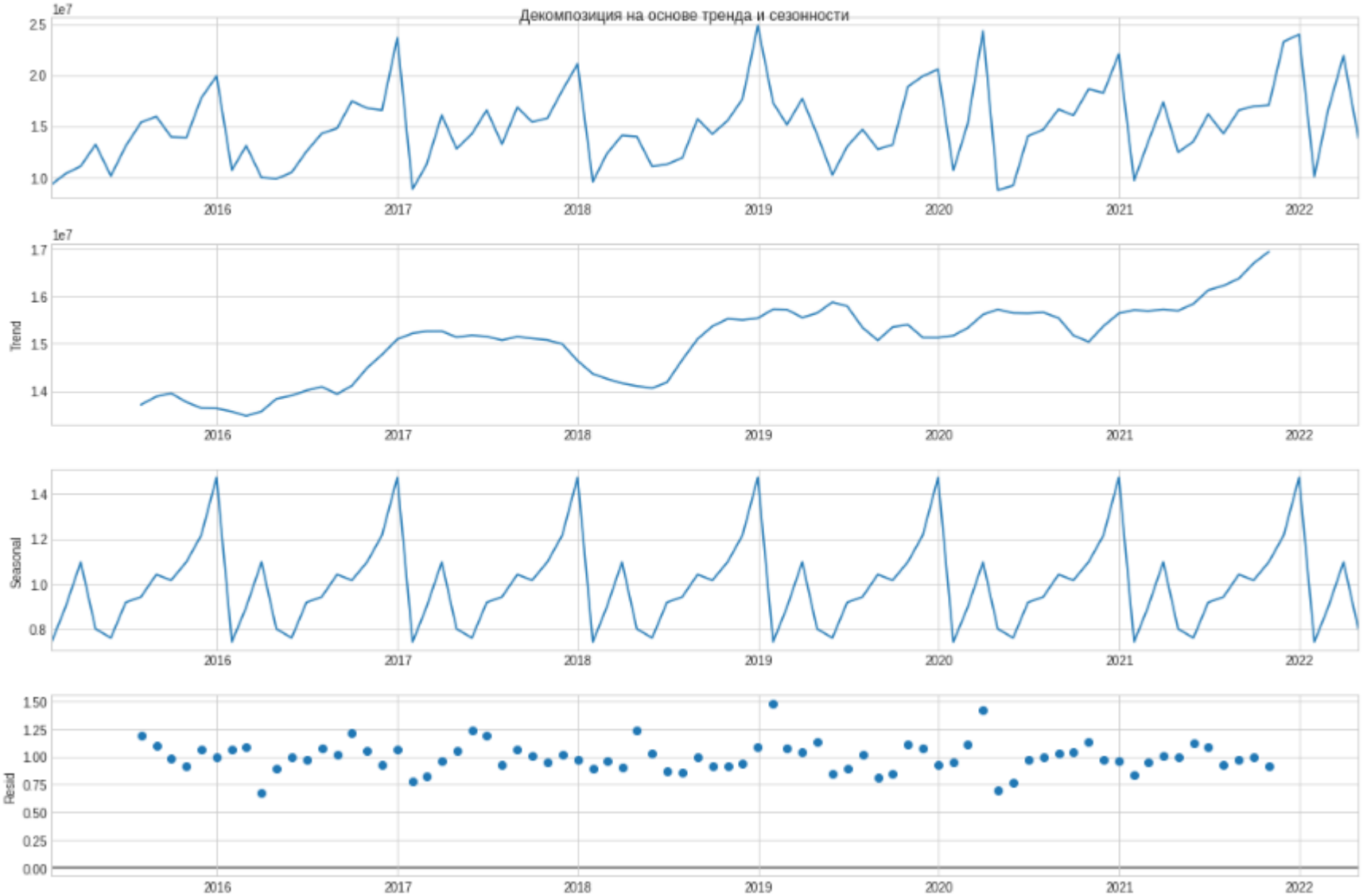

```
result_M = seasonal_decompose(data, model='multiplicative')


plt.rcParams.update({'figure.figsize': (15, 10)})
result_M.plot().suptitle('Декомпозиция на основе тренда и сезонности')
```

Тренд.
Видно, что тренд «ломается» после 2013 года.

Сезонность.
Отчетливая повторяющаяся закономерность, наблюдаемая через равные промежутки времени из-за различных сезонных факторов.

Очистка от аномалий

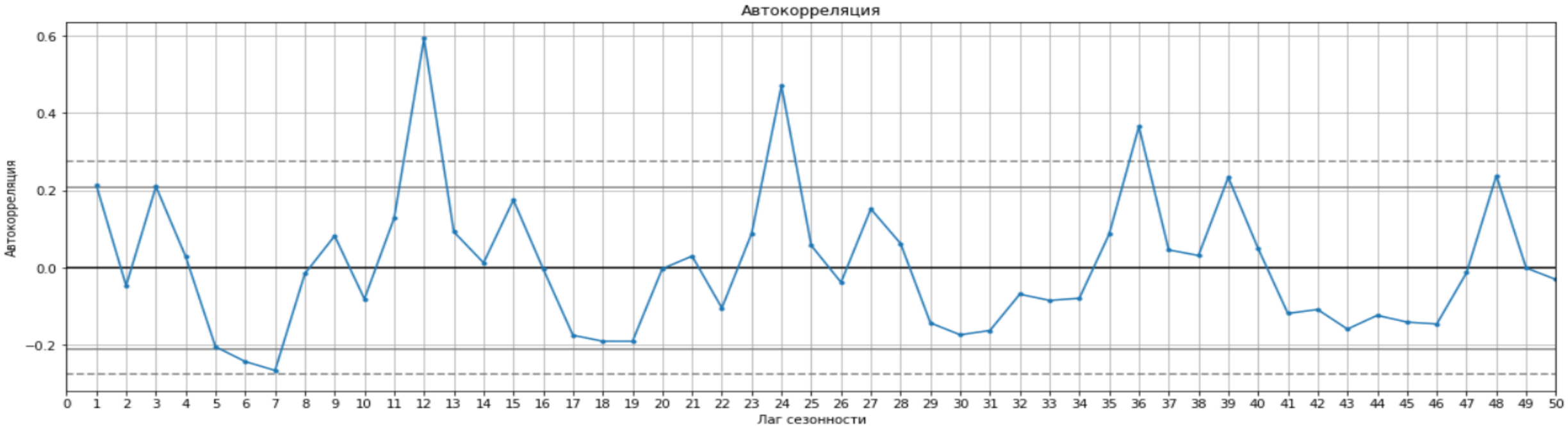



```
[ ]  data = data['2015-01-01':]
```

Тренд.
Виден слабый тренд на рост

Сезонность.
Отчетливая повторяющаяся закономерность, наблюдаемая через равные промежутки времени из-за различных сезонных факторов.

Наличие сезонности

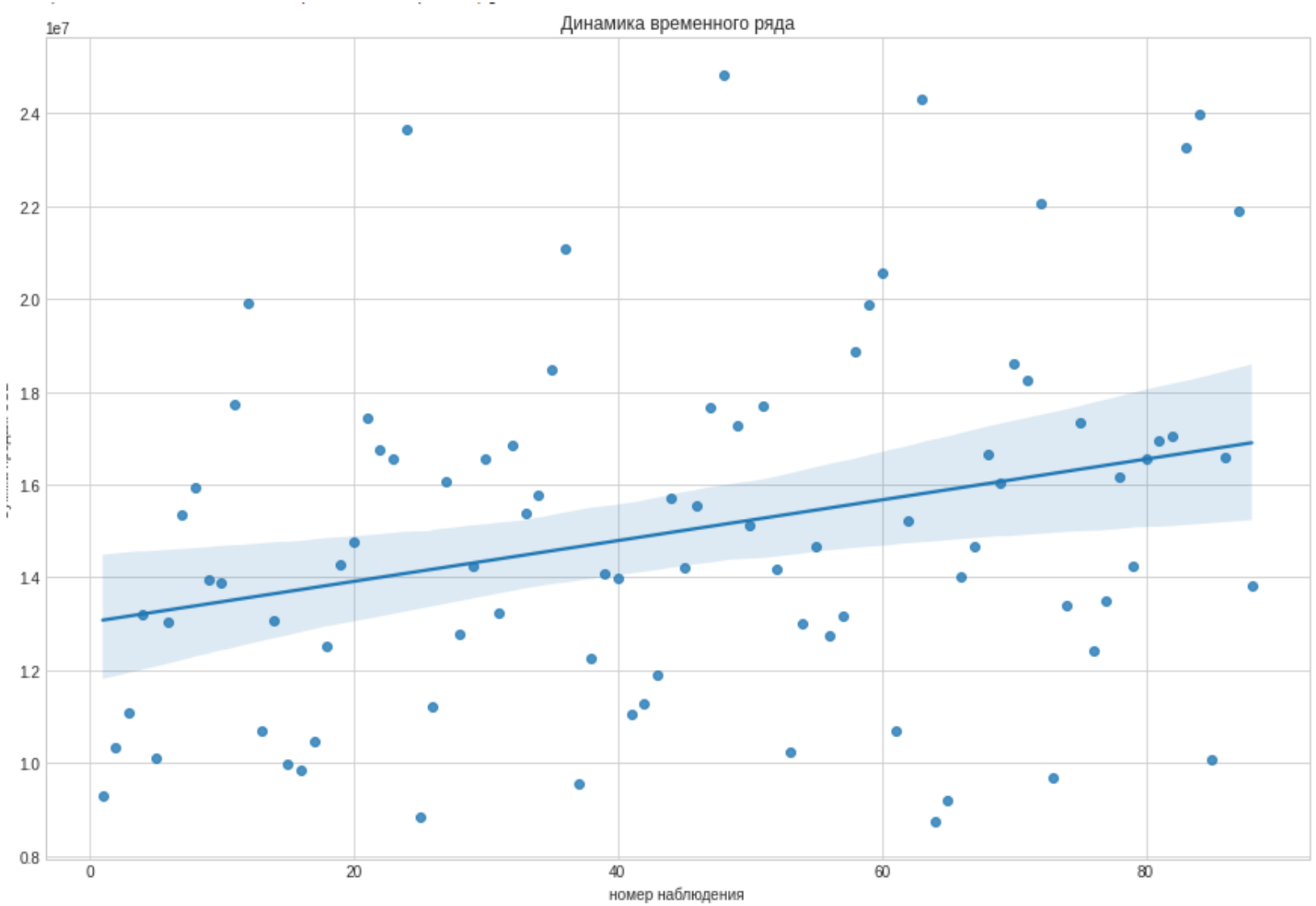


Отчетливая повторяющаяся закономерность, наблюдаемая через равные промежутки времени из-за различных сезонных факторов.

```
fig, ax = plt.subplots(figsize = (20, 6))

autocorrelation_plot(data, ax = ax, marker = '.')
ax.xaxis.set_major_locator(plt.MultipleLocator(1))
ax.set_xlim(0, 50)
plt.xlabel('Лаг сезонности')
plt.ylabel('Автокорреляция')
plt.title('Автокорреляция')
plt.show()
```

Проверка на наличие тренда



Визуально, можно предположить, что есть тренд, проверяем с помощью теста Дики-Фуллера

Проверка на наличие тренда

```
Тест на стационарность:
        Т-статистика = -1.016
        Р-значение = 0.747
Критические значения :
        1%: -3.520713130074074 - Данные не стационарны с вероятностью 99% процентов
        5%: -2.9009249540740742 - Данные не стационарны с вероятностью 95% процентов
        10%: -2.5877813777777776 - Данные не стационарны с вероятностью 90% процентов
```

Ряд не стационарен. Присутствует тренд

Тренд сезонная модель

Суть данного подхода заключается в следующем: каждое значение временного ряда ($\hat{Y}_\tau$)
в момент времени τ раскладывается на три составляющие:

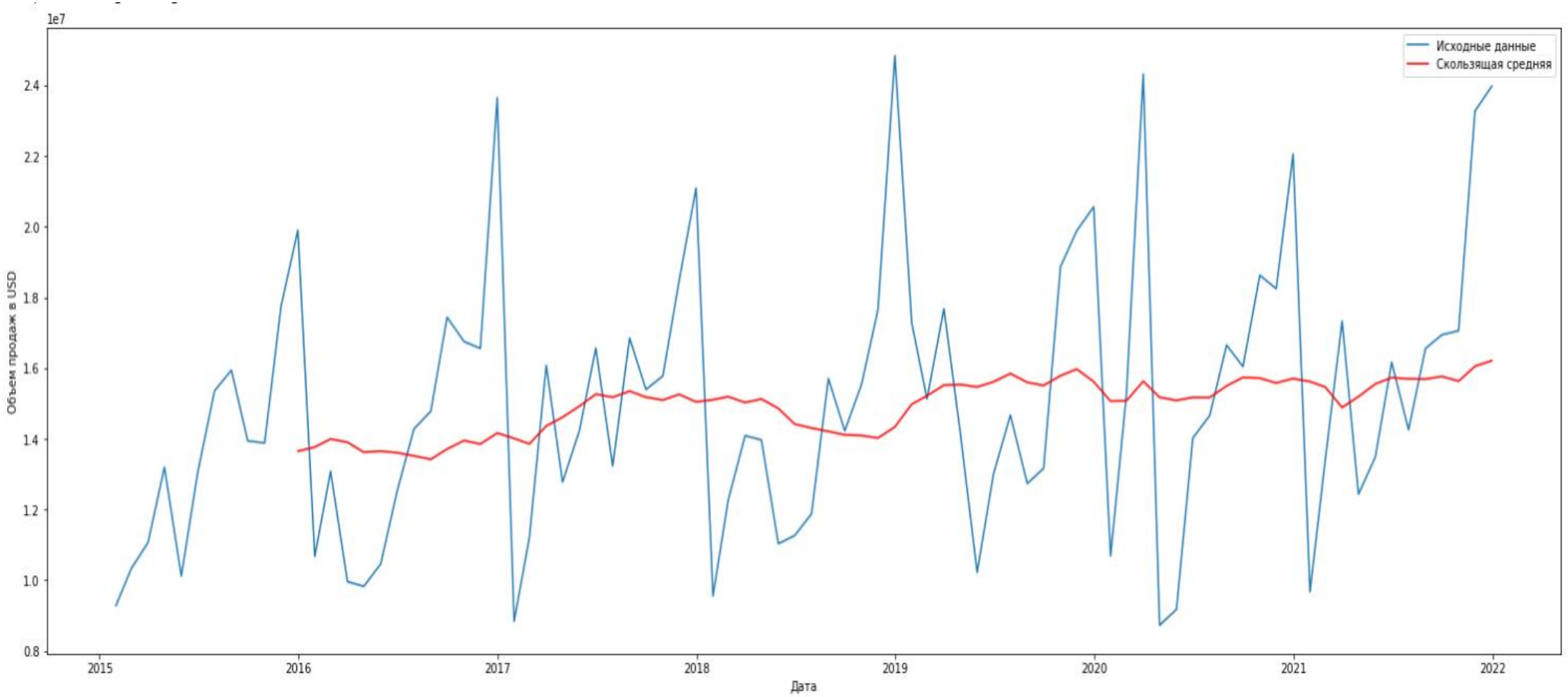
- тренд (T_τ),
- сезонную (S_τ),
- случайную компоненту (E_τ).

Алгоритм построения модели:

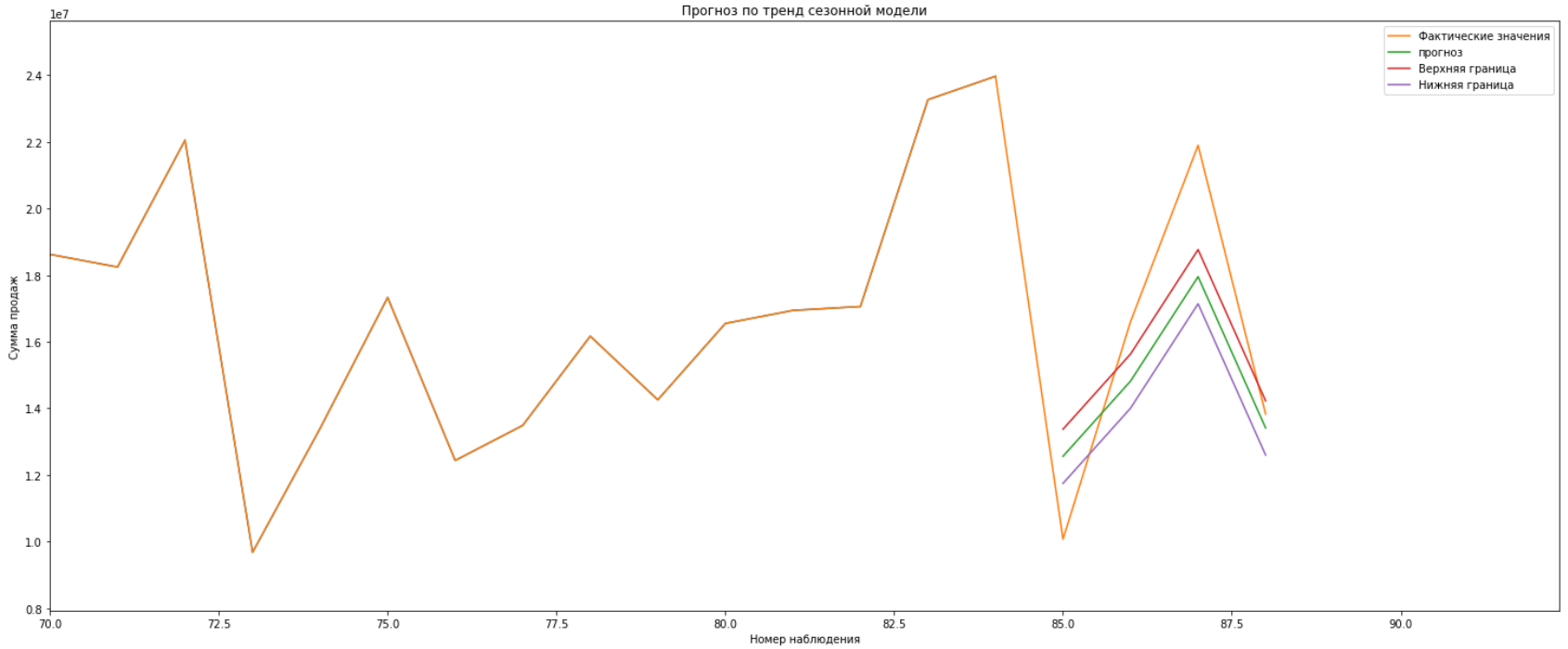
- Выравниваем ряд с помощью, скользящей средней.
- Рассчитываем значение сезонной компоненты S_τ.
- Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных.
- Рассчитываем значения T_τ с использованием полученного уравнения тренда.
- Используя полученные значения S_τ и T_τ, находим прогнозные значения уровней временного ряда.
- Оцениваем качество модели.

Тренд сезонная модель

Выравниваю ряд с помощью, скользящей средней. Ширина окна 12 месяцев



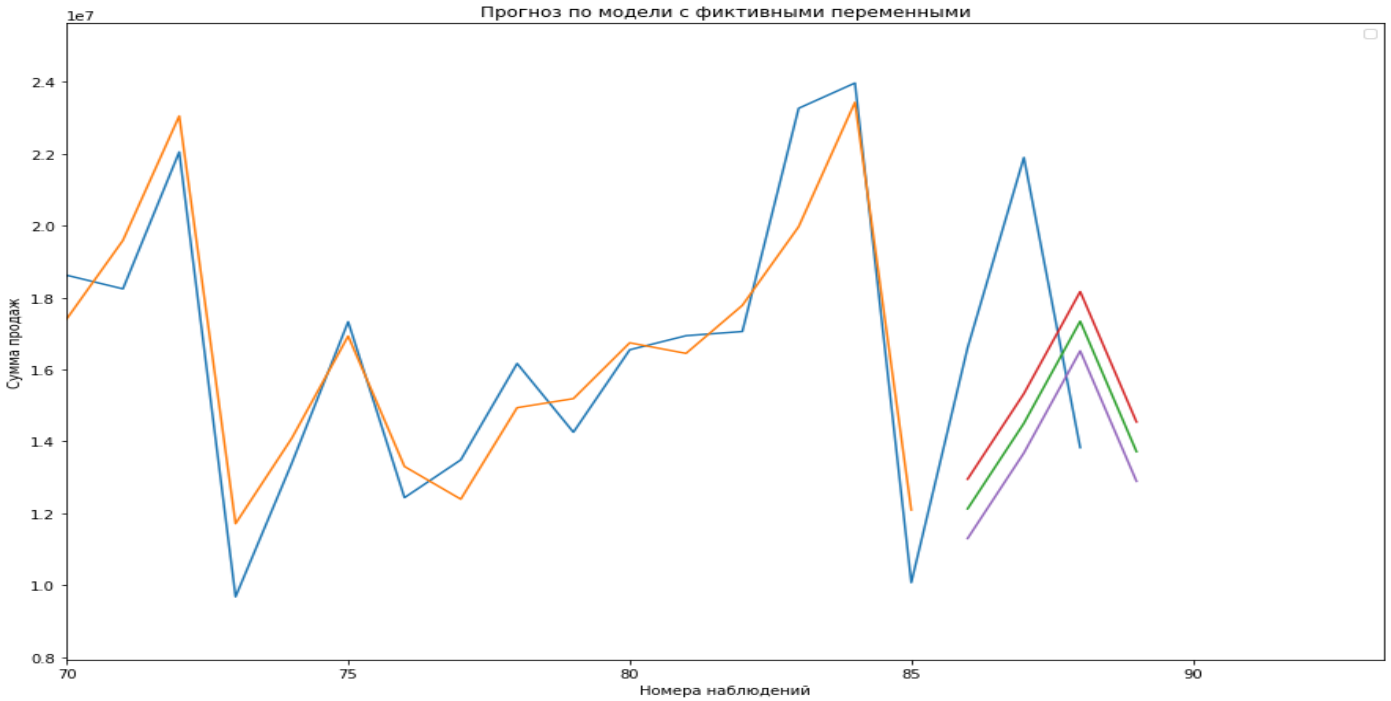
Тренд сезонная модель



Дата	Прогнозное значение	Фактическое значение
A	1	2
2022-01-31	1.26e+07	1.01e+07
2022-02-28	1.48e+07	1.66e+07
2022-03-31	1.80e+07	2.19e+07
2022-04-30	1.34e+07	1.38e+07

```
R2: 0.6623
MSE: 6284296611921.871
RMSE: 2506849.9380541053
MAD: 2159665.7115
MAPE: 0.1413
MPE: 0.0177
```

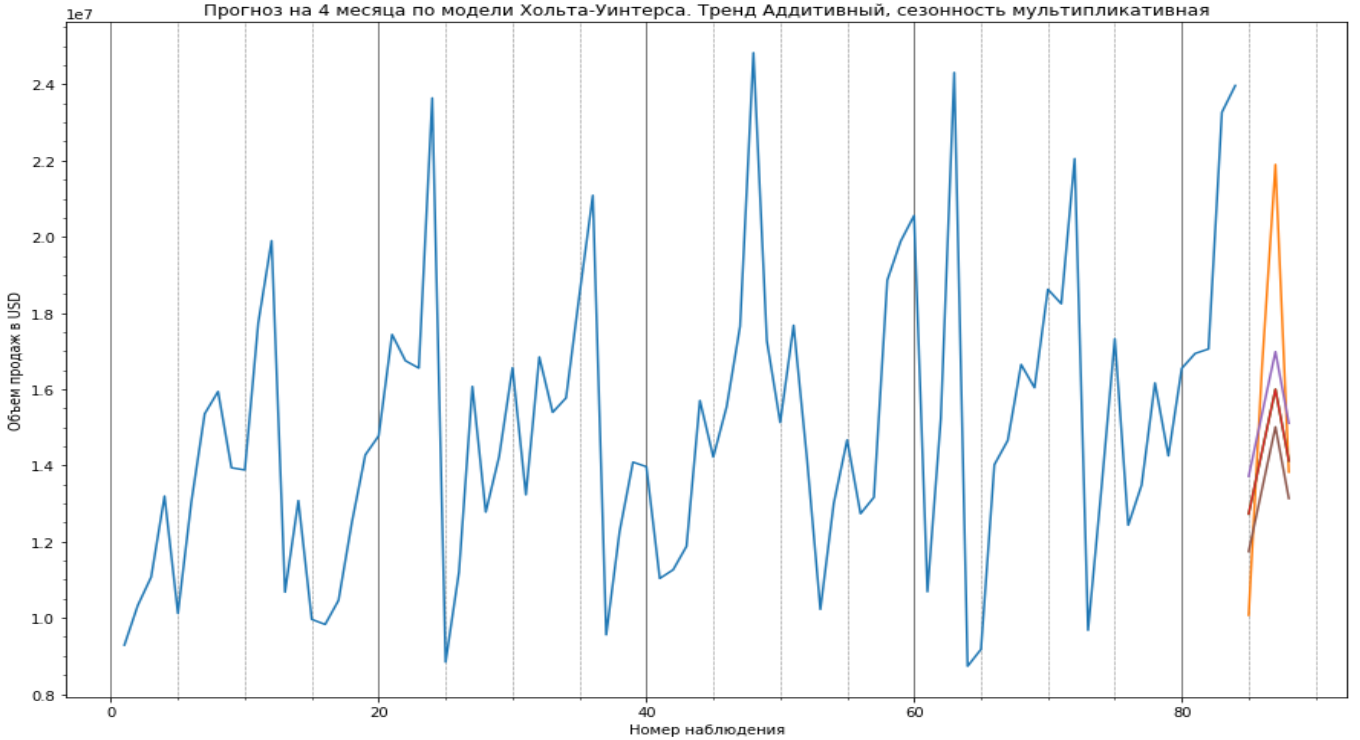
Модель с фиктивными переменными



Дата	Прогнозное значение	Фактическое значение
A	1	2
2022-01-31	1.21e+07	1.01e+07
2022-02-28	1.45e+07	1.66e+07
2022-03-31	1.73e+07	2.19e+07
2022-04-30	1.37e+07	1.38e+07

```
R2: 0.605
MSE: 7350789447918.339
RMSE: 2711233.9345615935
MAD: 2205568.976
MAPE: 0.1367
MPE: 0.0347
```

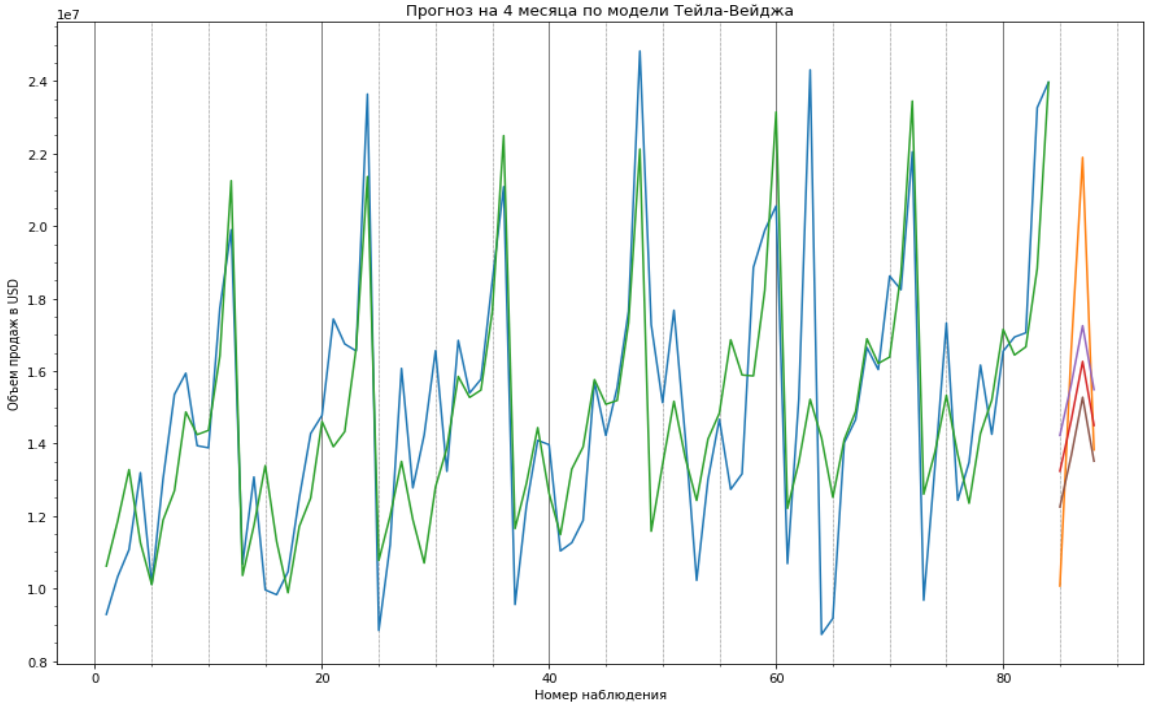
Модель Хольта Уинтерса



Дата	Прогнозное значение	Фактическое значение
A	1	2
2022-01-31	1.27e+07	1.01e+07
2022-02-28	1.43e+07	1.66e+07
2022-03-31	1.60e+07	2.19e+07
2022-04-30	1.41e+07	1.38e+07

R2: 0.3695
MSE: 11733598205679.25
RMSE: 3425434.017125312
MAD: 2771401.9882
MAPE: 0.1723
MPE: 0.0296

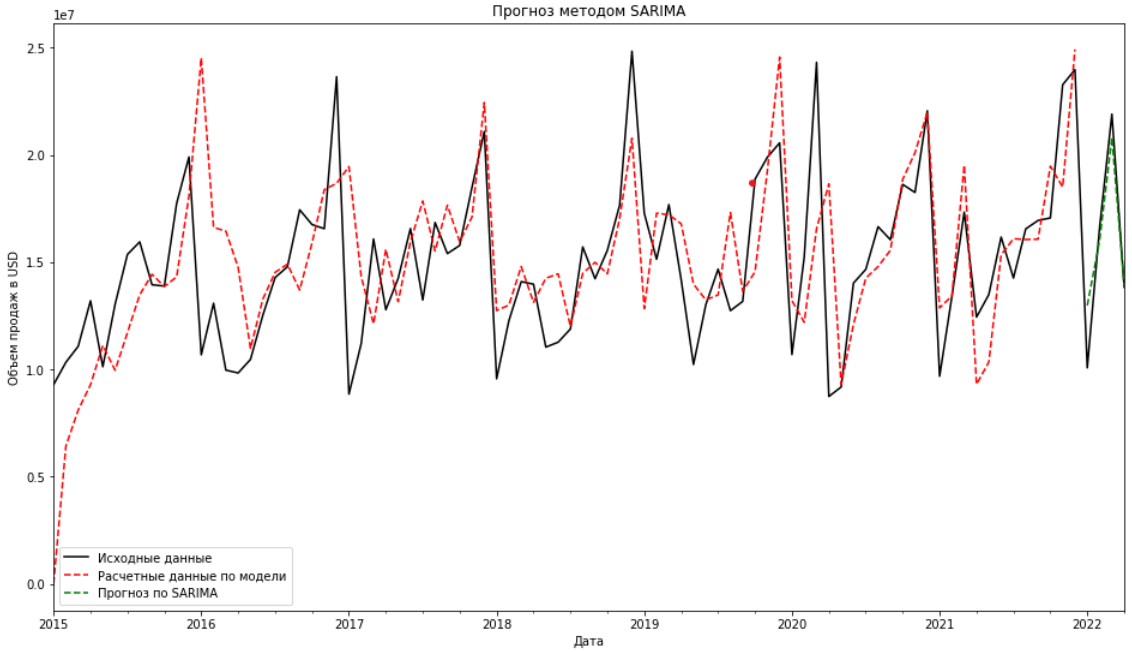
Модель Тейла Вейджа



Дата	Прогнозное значение	Фактическое значение
A	1	2
2022-01-31	1.32e+07	1.01e+07
2022-02-28	1.47e+07	1.66e+07
2022-03-31	1.63e+07	2.19e+07
2022-04-30	1.45e+07	1.38e+07

```
R2: 0.3817
MSE: 11505403622742.66
RMSE: 3391961.6187012875
MAD: 2856460.3858
MAPE: 0.1845
MPE: 0.0028
```

Модель SARIMA



Дата	Прогнозное значение	Фактическое значение
A	1	2
2022-01-31	1.33e+07	1.01e+07
2022-02-28	1.62e+07	1.66e+07
2022-03-31	2.02e+07	2.19e+07
2022-04-30	1.47e+07	1.38e+07

```
R2: 0.8658
MSE: 2497451991620.617
RMSE: 1580332.873675865
MAD: 1163189.8309
MAPE: 0.0947
MPE: -0.0518
```

Выбор адекватного метода

Модель прогнозирования	R^2	MAD	MSE	RMSE	MPE	MAPE
Тренд сезонная модель	0.6623	2159665.7115	6284296611921.871	2506849.93805	0.1413	0.0177
Модель с фиктивными переменными	0.605	2205568.976	7350789447918.339	2711233.93456	0.1367	0.0347
Модель Хольта-Уинтерса	0.3695	2771401.9882	11733598205679.25	3425434.01712	0.0296	0.1723
Модель Тейла-Вейджа	0.38	2856460.3858	11505403622742.66	3391961.6187	0.0028	0.1845
Модель SARIMA	0.8658	1163189.8309	2497451991620.617	158332.8736	-0.0518	0.0947

Из данной таблицы сложно определить, какой из методов является наиболее точен в прогнозных значениях, каждому из методов были выставлен ранг по каждому методу измерения ошибок прогнозирования, больший балл «5» получал самый точный и по убыванию до «1» который получал метод прогнозирования с самыми большими отклонениями. После, все баллы, были просуммированы. На основании полученных результатов модели прогнозирования были ранжированы по приоритету выбора, то есть от наиболее точного к наименее точному методу прогнозирования, это позволит определить точность каждого из методов, который были рассмотрены в рамках данного исследования.

Выбор адекватного метода

Модель прогнозирования	R^2	MAD	MSE	RMSE	MPE	MAPE	Сумма баллов	Приоритет выбора
Тренд сезонная модель	4	4	1	4	1	1	15	4
Модель с фиктивными переменными	3	3	2	3	2	2	15	4
Модель Хольта-Уинтерса	1	2	4	1	4	4	16	3
Модель Тейла-Вейджа	2	1	3	2	5	5	18	2
Модель SARIMA	5	5	5	5	3	3	26	1

По итогам работы можно утверждать, что модель, SARIMA, является наиболее точной по сравнению с остальными методами прогнозирования. Также модели прогнозирования такие как: тренд сезонная, модель с фиктивными переменными показали достаточную точность прогнозирования и также могут быть использованы при проведении процедуры прогнозирования, для дополнительной проверки, средний коэффициент детерминации для данных моделей был выше значения 0,6. Это означает, что факторы, которые были выбраны адекватно описывают изменение продаж.

Спасибо за внимание!