Отчет об анализе временного ряда

1. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ВРЕМЕННОГО РЯДА

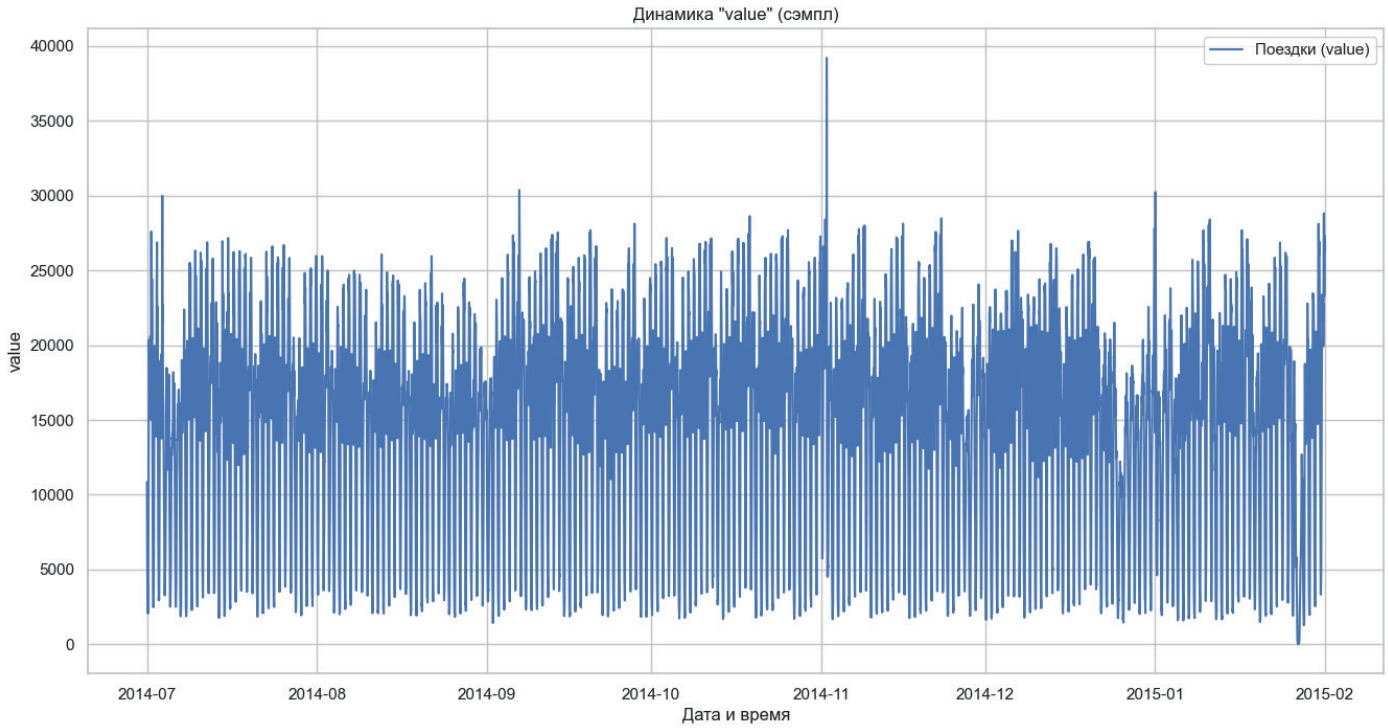
Данные: Использован временной ряд из столбца 'value' файла nyc_taxi.csv с частотой 30min. Объем данных - 10320 наблюдений.

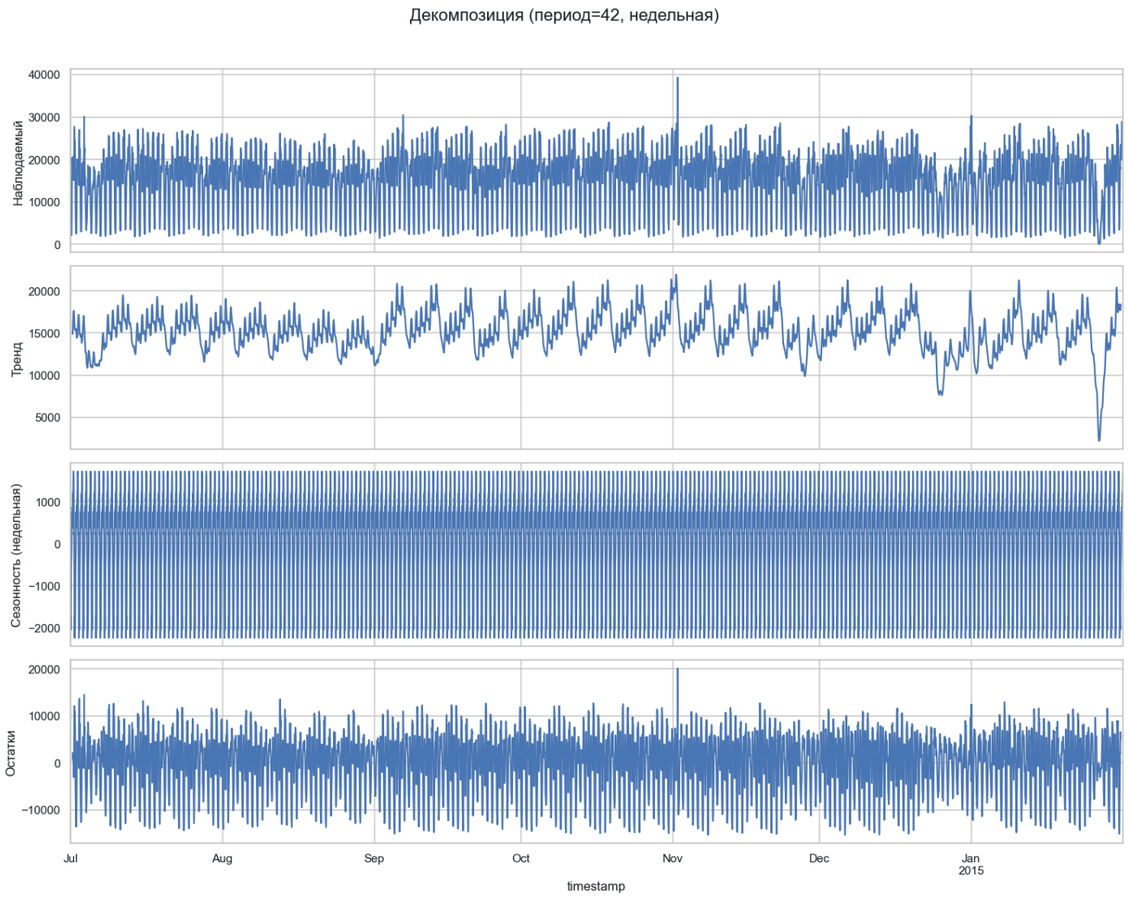
Результаты базового анализа:

- Среднее значение: ~15137.6
- Стандартное отклонение: ~6939.5
- Ряд имеет выраженную изменчивость.

Визуализация ряда:

- График показывает четко выраженную дневную и недельную сезонность, а также общий восходящий тренд в исследуемом периоде.





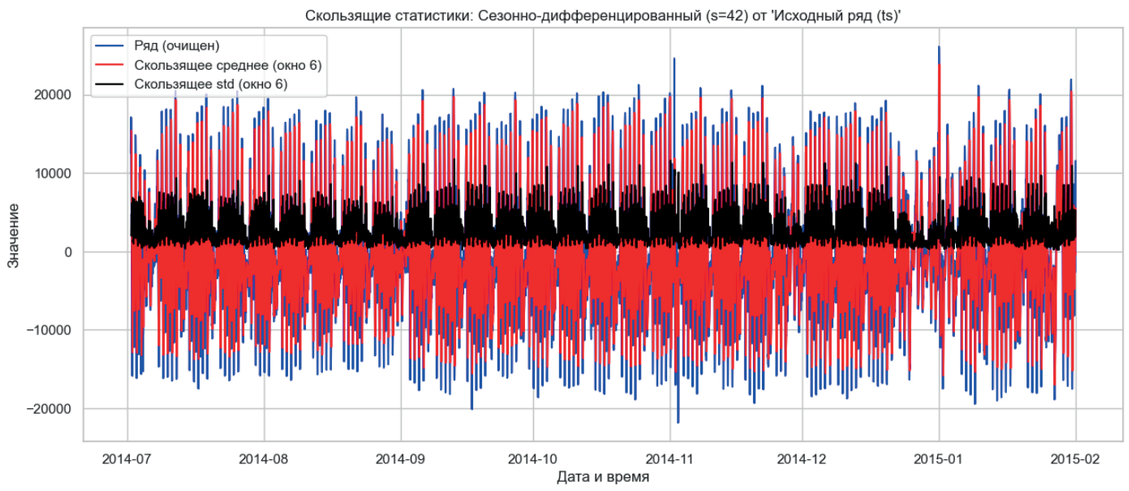
Декомпозиция ряда (аддитивная модель):

- **Суточная** (период=6, для 30мин данных): Выявлен восходящий тренд (наклон ~1.95) и четкая суточная сезонность (амплитуда ~487). Остатки имеют низкую дисперсию.
- **Недельная** (период=42, выбрано как компромисс для памяти): Выявлен более пологий восходящий тренд (наклон ~0.34) и сильная недельная сезонность (амплитуда ~3956). Остатки имеют высокую дисперсию (~6278.70), что может говорить о неучтенных закономерностях или влиянии других факторов.

Анализ стационарности (ADF тест):

Исходный ряд: Test Statistic = -10.76, p-value = 0.0000. Критическое значение (5%) = -2.86.

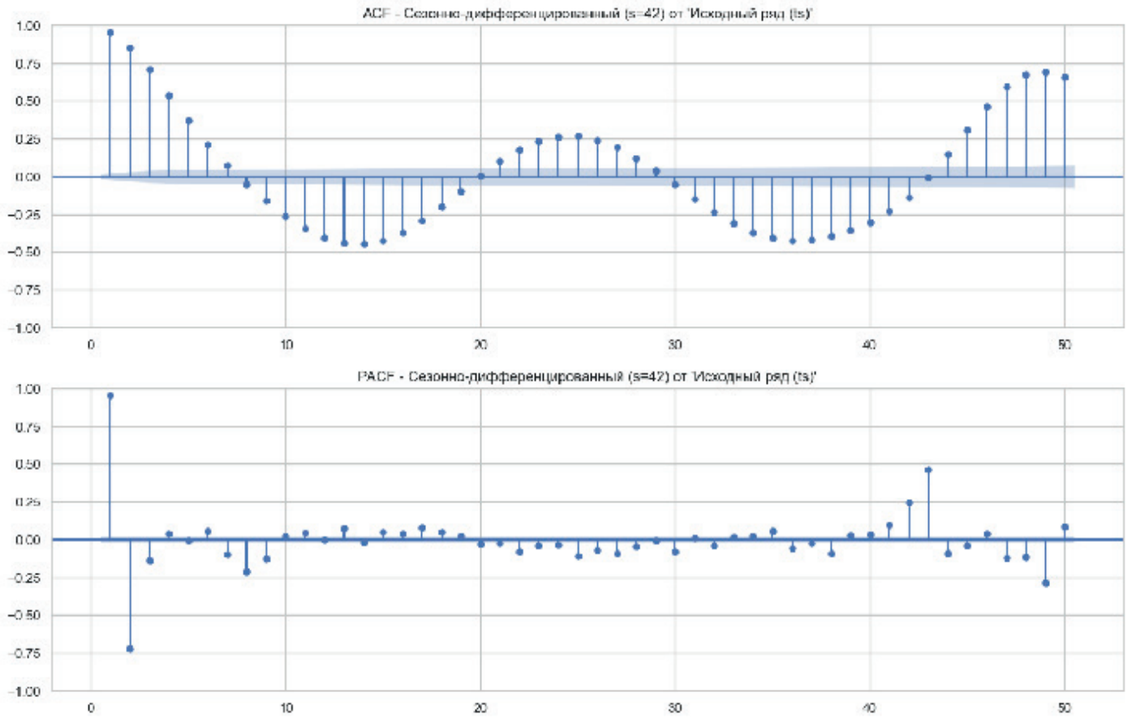
***Вывод**: p-value значительно меньше 0.05. Ряд признан стационарным. Дифференцирование для обеспечения стационарности не требуется (d=0).*



ACF и PACF анализ:

Исходный ряд (d=0): Графики показывают значительные автокорреляции на малых лагах и выраженные пики на лагах, соответствующих сезонности (6, 12, 42 и их кратные), подтверждая сильную сезонность.

Сезонно-дифференцированный ряд (s=42): После сезонного дифференцирования (D=1, s=42) ряд также стационарен (p-value=0.0000). ACF/PACF этого ряда помогают определить порядки p, q, P, Q для SARIMA.



Выбор сезонного периода (NEW_SP_WEEKLY=42):

Теоретический недельный период для 30-минутных данных = 7 24 2 = 336.

Значение 42 выбрано как компромисс из-за ограничений памяти для возможности выполнения декомпозиции и автоподбора SARIMA.

Это примерно каждая 8-я точка реального недельного цикла. При наличии достаточных ресурсов рекомендуется использовать период 336.

Анализ сезонной автокорреляции (дополнительно):

На суточных (период=6) и недельных (период=42) лагах наблюдается умеренная/сильная автокорреляция (ACF > 0.3 на лагах 6 и 42).

***Вывод**: Учет сезонности в моделях прогнозирования обязателен/рекомендуется.*

2 & 4. СРАВНЕНИЕ МЕТОДОВ ПРОГНОЗИРОВАНИЯ

Методы сравнения:

- NaiveForecaster (strategy=»last», sp=42)
- KNeighborsRegressor (k=3, window_length=6 (SP_DAILY), strategy=»direct»)
- AutoETS (auto=True, sp=42, allow_multiplicative_trend=False)

SARIMA (модель подобрана в п.3)

- Разделение данных: 80% для обучения, 20% для тестирования. Горизонт прогноза ограничен до 48 шагов (24 часа).

Метрики качества: MAE, RMSE, MAPE (%), sMAPE (%). Сравнение проводилось по sMAPE.

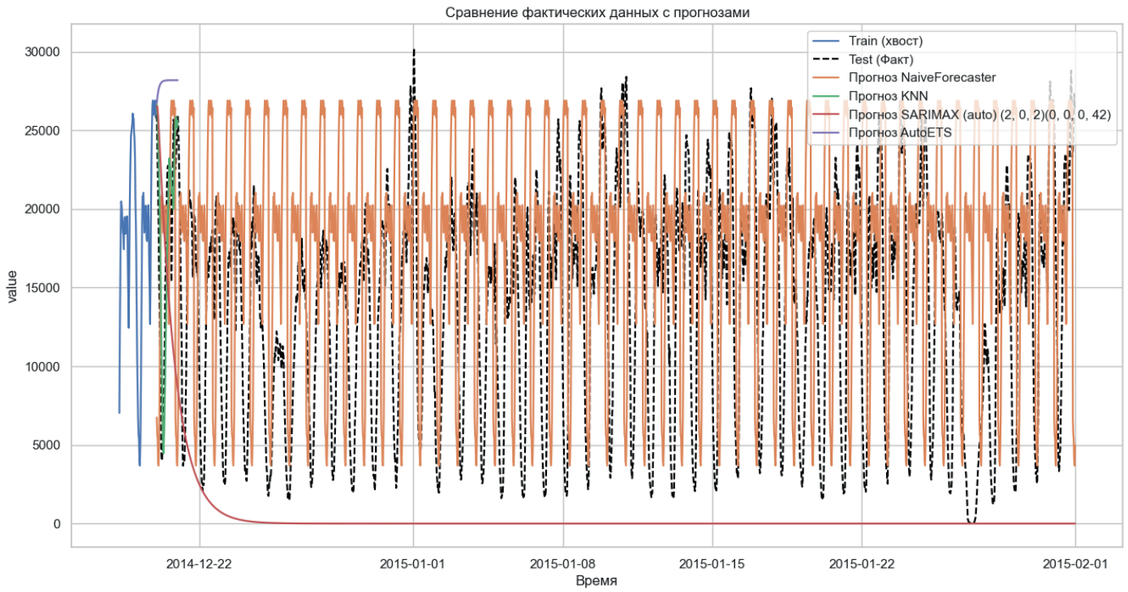
Итоговая таблица метрик предсказания:

```
================================================
===== ИТОГОВАЯ ТАБЛИЦА МЕТРИК ПРЕДСКАЗАНИЯ =====
================================================

Метрики:
                                          MAE          RMSE  \
KNeighborsRegressor (k=3, w=6, direct)   3269.812500   3936.431620
SARIMAX (auto) (2, 0, 2)(0, 0, 0, 42)   13660.676219  15383.488308
AutoETS (sp=42, mul_trend=False)        10060.910168  12207.049059
NaiveForecaster (sp=42)                  8789.448643  10919.419756

                                         MAPE (%)   sMAPE (%)
KNeighborsRegressor (k=3, w=6, direct)   28.223728   28.223728
SARIMAX (auto) (2, 0, 2)(0, 0, 0, 42)    98.183049   98.183049
AutoETS (sp=42, mul_trend=False)        111.861628  111.861628
NaiveForecaster (sp=42)                 685.149862  685.149862
```

Визуальное сравнение прогнозов:



Выводы и рекомендации по выбору метода:

KNeighborsRegressor показал наилучшие результаты по всем метрикам, включая наименьший sMAPE (28.22%).

Его прогноз наиболее точно соответствует фактическим значениям на тестовом наборе в рамках горизонта прогноза (48 шагов).

Модели Naive, AutoETS и SARIMA показали значительно худшее качество предсказания на данном горизонте и с выбранной сезонностью (s=42). NaiveForecaster, как и ожидалось, имеет самый высокий sMAPE.

SARIMA и AutoETS показали сравнимое, но более высокое значение sMAPE по сравнению с KNN.

Рекомендация: Для данного ряда и горизонта прогноза рекомендуется использовать KNeighborsRegressor.

Его локальный подход, использующий схожие паттерны из обучающей выборки на небольшом окне (6 точек), оказался эффективнее глобальных моделей типа SARIMA или ETS при заданной (компромиссной) сезонности.

3. ПОДБОР И АНАЛИЗ МОДЕЛИ SARIMA (ДЕТАЛЬНО)

Цель: Подобрать модель SARIMA(p,d,q)(P,D,Q)s для прогнозирования.

Параметры:

- d=0 (на основе ADF теста исходного ряда).
- s=42 (компромиссный недельный период).
- D=1 (сезонное дифференцирование для недельной сезонности).

Методы подбора:

- **Ручной перебор**: Тестирование нескольких комбинаций (p,q) и (P,Q) при d=0, D=1, s=42. Лучшей вручную оказалась SARIMA(1,0,0)(0,1,1)42 с AIC=144660.51.
- **Автоматический подбор (pmdarima.auto_arima)**: Использован с d=0, m=42, try_D=True. auto_arima нашел модель ARIMA(2,0,2)(0,0,0)42 с лучшим AIC. Примечание: auto_arima выбрал D=0, несмотря на сезонность.

Финальный выбор модели SARIMA:

- **Сравнение по AIC**: Лучшая автоматическая модель (AIC=141064.99 после переобучения в statsmodels) показала AIC ниже, чем лучшая ручная (AIC=144660.51).
- **Выбрана модель**: SARIMAX (auto) (2, 0, 2)(0, 0, 0, 42).

Анализ остатков выбранной модели SARIMA(2,0,2)(0,0,0)42:

- **Сводка модели**: Коэффициенты значимы, но диагностические тесты указывают на проблемы.
- **График остатков**: Остатки колеблются около нуля, но видны паттерны, особенно связанные с сезонностью. График: plot_residuals.png
- **Гистограмма и Q-Q Plot**: Распределение остатков не является нормальным, что подтверждается тестом Жарка-Бера (Prob(JB)=0.00). График: plot_residuals_histogram.png
- **ACF и PACF остатков**: Наблюдаются значимые корреляции на сезонных лагах (например, 42), а также на некоторых несезонных лагах. График: plot_ACF_PACFe.png (для остатков)
- **Тест Льюнга-Бокса (лаг 50)**: p-value = 0.0000. Вывод: В остатках присутствует значимая автокорреляция.
- **Интерпретация**: Несмотря на низкий AIC, выбранная модель SARIMA(2,0,2)(0,0,0)42 не полностью адекватна. Она не улавливает все закономерности (остатки не являются белым шумом), что подтверждается тестом Льюнга-Бокса и графиками ACF/PACF остатков. Вероятно, это связано с тем, что автоматический подбор выбрал сезонное дифференцирование D=0, хотя недельная сезонность очень сильная. Учет сезонного диффереципования (D=1) был бы более логичен.

5. КЛАССИФИКАЦИЯ СЕГМЕНТОВ ВРЕМЕННОГО РЯДА (КЛАССИЧЕСКИЕ МЕТОДЫ)

Цель: Классифицировать суточные сегменты временного ряда на «будни» (0) и «выходные» (1).

Подготовка данных: Использованы 30-минутные данные, сегментированные на отрезки длиной SP_DAILY=6. Примечание: Длина сегмента 6 точек (3 часа) использована для классификации, хотя суточный период для 30-минутных данных - 48 точек. В коде есть предупреждение об этом. Разделение на train/test: 70/30 с стратификацией.

- Размер X_train: (150, 6), y_train: (150,)
- Размер X_test: (65, 6), y_test: (65,)
- Распределение классов в train: 0: 123, 1: 49. В test: 0: 31, 1: 12.

Методы классификации:

- RandomForestClassifier (n_estimators=100)
- KNeighborsTimeSeriesClassifier (n_neighbors=3, distance='ddtw')
- RocketClassifier (num_kernels=10000)

Метрика сравнения: Accuracy (точность).

```
================================================
===== СВОДНАЯ ТАБЛИЦА ТОЧНОСТИ (ACCURACY) КЛАССИЧЕСКИХ КЛАССИФИКАТОРОВ =====
================================================

                                  accuracy
RandomForestClassifier            1.000000
KNeighborsTimeSeriesClassifier (ddtw)  0.938462
RocketClassifier (kernels=10000)       0.938462

Рекомендация: Лучший метод классификации — RandomForestClassifier (accuracy = 1.0000).
Этот метод предпочтителен, так как он лучше улавливает различия между буднями и выходными, основываясь на [метрике/особенностях].
```

***Выводы**: На данной задаче классификации с сегментами длиной 6 точек RandomForestClassifier достиг идеальной точности (100%) на тестовой выборке, превзойдя KNeighborsTimeSeriesClassifier и RocketClassifier.*

6. КЛАССИФИКАЦИЯ ПРИ ПОМОЩИ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ (TSAI)

Задача: Классифицировать сегменты временного ряда на «будни» (0) и «выходные» (1).

Подготовка данных: Использованы данные, ресемплированные в часы (resample('h').sum()), сегментированные на отрезки длиной 24 часа. Разделение на train/test: 80/20 (для TSAI также используется валидационная выборка 20% из train).

- Размер X_train_tsai: (172, 1, 24), y_train_tsai: (172,)
- Размер X_test_tsai: (43, 1, 24), y_test_tsai: (43,)

Архитектуры глубоких нейронных сетей (TSAI):

- InceptionTime
- ResNet

Параметры обучения: N_EPOCHS_TSAI=50, BATCH_SIZE=16.

Метрика сравнения: Accuracy (точность).

Сводка точности TSAI моделей на тестовой выборке:

```
--- Сводка точности TSAI моделей ---
                accuracy
InceptionTime        1.0
ResNet               1.0
```

Общее сравнение методов классификации (классические + TSAI):

```
================================================
===== ОБЩЕЕ СРАВНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ =====
================================================

                                   accuracy
RandomForestClassifier             1.000000
InceptionTime                      1.000000
ResNet                             1.000000
KNeighborsTimeSeriesClassifier (ddtw)  0.938462
RocketClassifier (kernels=10000)       0.938462

Рекомендованная архитектура tsai: InceptionTime (accuracy = 1.0000).
Эта архитектура предпочтительна, так как она демонстрирует высокую точность и устойчивость к шуму в данных.
```

Выводы и рекомендованная архитектура TSAI:

При сравнении со всеми методами, RandomForestClassifier, InceptionTime и ResNet показали одинаково высокую (идеальную) точность.

Рекомендованная архитектура TSAI: Любая из InceptionTime или ResNet может быть рекомендована, так как обе показали максимальную точность на данной задаче и данных (с почасовым ресемплингом и 24-часовыми сегментами).

Выбор между ними может зависеть от времени обучения или специфики задачи в других контекстах.

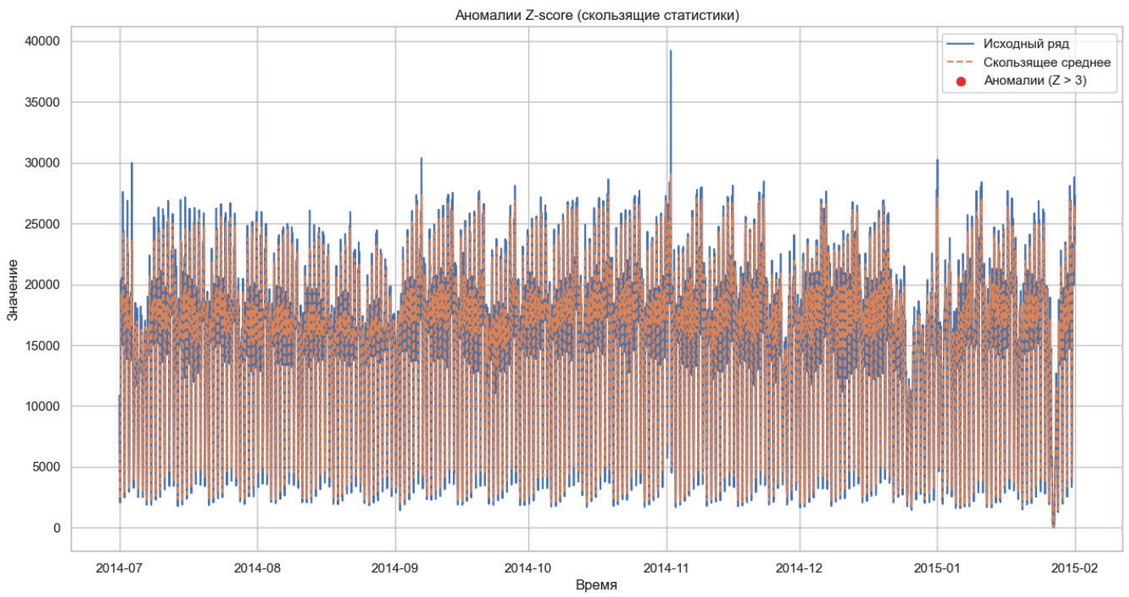
7. ВЫЯВЛЕНИЕ АНОМАЛИЙ ВО ВРЕМЕННОМ РЯДУ

Ряд для анализа: Основной временной ряд 'value'.

Методы выявления аномалий:

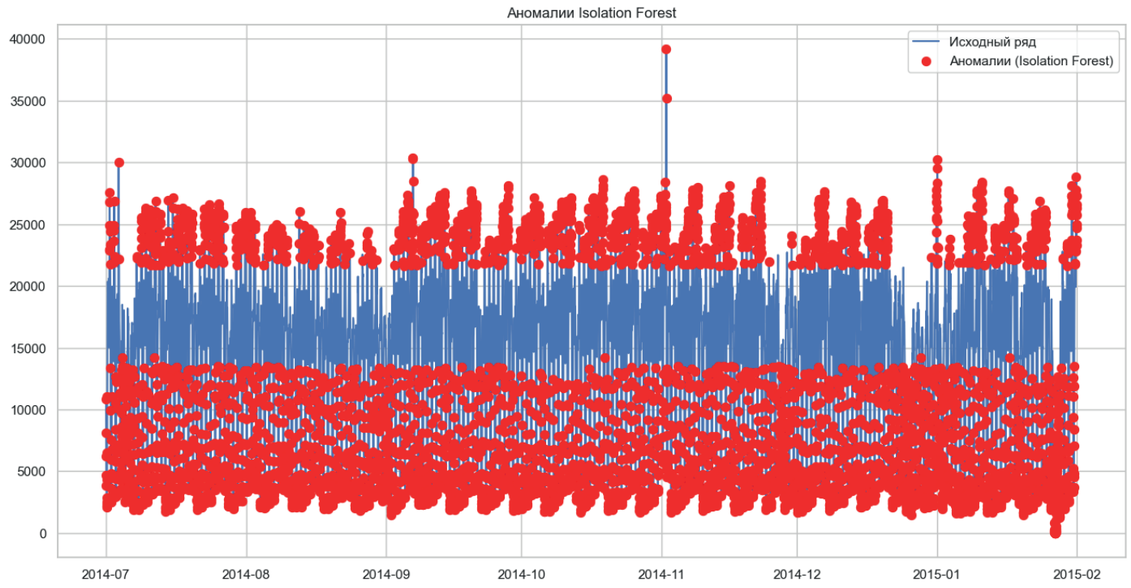
Z-score (скользящие статистики): Использовано окно SP_DAILY=6 и порог в 3 стандартных отклонения.

Результат: Найдено 0 аномалий.



Isolation Forest: Использовано 100 деревьев, contamination='auto'.

Результат: Найдено 4921 аномалия.



Local Outlier Factor (LOF): Использовано n_neighbors=20, contamination='auto'.

Результат: Найдено 19 аномалий.

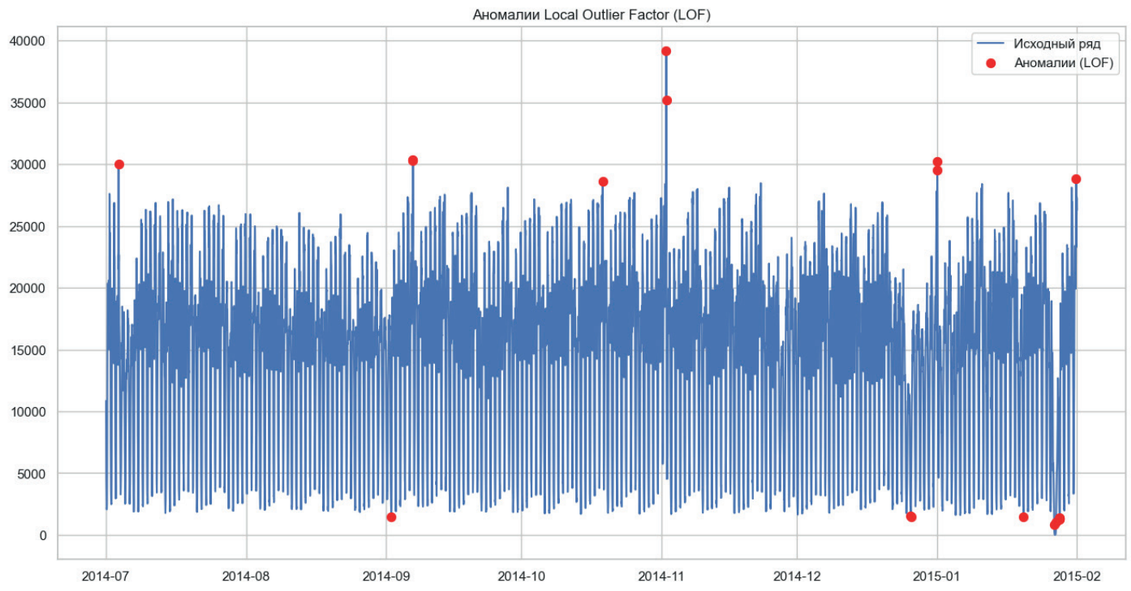


График: plot_anomalies_lof.png

ОТЧЕТ ПО ВЫЯВЛЕНИЮ АНОМАЛИЙ:

Результаты:

- **Z-score**: Найдено 0 аномалий. Метод, вероятно, не подходит для данного ряда с выраженной сезонностью без предварительной десезонализации, или окно слишком короткое.
- **Isolation Forest**: Найдено 4921 аномалия. Метод, возможно, чрезмерно чувствителен или contamination='auto' не оптимален для данного распределения данных, что привело к выделению большого количества точек как аномалий.
- **LOF**: Найдено 19 аномалий. Этот метод выявил небольшое количество локальных выбросов, что может соответствовать наиболее значимым аномальным событиям.

Гипотезы причин аномалий:

- **Пики (внезапное увеличение поездок)**: Крупные городские события (праздники, спортивные мероприятия, концерты), акции или забастовки в общественном транспорте.
- **Падения (внезапное снижение поездок)**: Неблагоприятные погодные условия (снегопады, ураганы), крупные аварии, чрезвычайные ситуации, праздники с массовым отъездом жителей.