**GeekBrains**

|  |  |
| --- | --- |
| **Направление** | Аналитика |

**ДИПЛОМНАЯ РАБОТЫ**

**Тема: Анализ выхода из строя оборудования на машиностроительном предприятии**

|  |  |
| --- | --- |
| Студентка: | Бабаева О.С. |

Москва

2024

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc185781447)

[1 Теоретические основы сбора и анализа данных 5](#_Toc185781448)

[1.1 Сбор данных 5](#_Toc185781449)

[1.1.1 Источники сбора данных 6](#_Toc185781450)

[1.1.2 Способы и инструменты сбора данных 8](#_Toc185781451)

[1.1.3 Хранение данных 10](#_Toc185781452)

[1.2 Подготовка данных 12](#_Toc185781453)

[1.2.1 Методы очистки данных 13](#_Toc185781454)

[1.2.2 Инструменты очистки данных 14](#_Toc185781455)

[1.2.3 Методы обработки данных 15](#_Toc185781456)

[1.2.4 Инструмент обработки данных 18](#_Toc185781457)

[1.3 Анализ данных 19](#_Toc185781458)

[1.3.1 Методы анализа данных 19](#_Toc185781459)

[1.3.2 Инструменты анализа данных 23](#_Toc185781460)

[2 Практическая часть работы 25](#_Toc185781461)

[2.1 Сбор данных 25](#_Toc185781462)

[2.1.1 Сбор и обработка данных, выгруженных из ПО предприятия 25](#_Toc185781463)

[2.1.2 Парсинг погодных данных 26](#_Toc185781464)

[2.1.3 Объединение данных 26](#_Toc185781465)

[2.2 Визуализация данных 27](#_Toc185781466)

[2.2.1 Работа с данными в Power Query 27](#_Toc185781467)

[2.3 Прогнозирование поломок оборудования 29](#_Toc185781468)

[2.2.1 Модель ARIMA 29](#_Toc185781469)

[2.2.2 Прогноз поломок оборудования на машиностроительном предприятии 33](#_Toc185781470)

[ЗАКЛЮЧЕНИЕ 35](#_Toc185781471)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 36](#_Toc185781472)

[ПРИЛОЖЕНИЕ А Отчет в ПО «ОРО» 37](#_Toc185781473)

[ПРИЛОЖЕНИЕ Б Объединение и преобразование заявок в один файл 38](#_Toc185781474)

[ПРИЛОЖЕНИЕ В Парсинг сайта погоды 44](#_Toc185781475)

[ПРИЛОЖЕНИЕ Г Объединение данных 49](#_Toc185781476)

[ПРИЛОЖЕНИЕ Д Объединение данных 55](#_Toc185781477)

[ПРИЛОЖЕНИЕ Е Модель ARIMA для прогноза поломок 62](#_Toc185781478)

# ВВЕДЕНИЕ

**Актуальность темы исследования.** В современном мире для любого предприятия необходимо не только поддерживать свою конкурентоспособность, но и укреплять её. Прогнозирование и моделирование результатов деятельности могут стать инструментами для повышения конкурентоспособности.

**Целью** дипломной работы является сбор, анализ и обработка данных по оборудованию машиностроительного предприятия с целью дальнейшего прогнозирования его работы и подготовка визуального отчета (дашборда) для цифровой системы предприятия.

Для достижения поставленной цели в дипломной работе необходимо выполнить следующие задачи:

* собрать необходимую информацию для работы;
* проанализировать и обработать полученные данные;
* построить визуальный отчет по имеющимся данным;
* проверить зависимость выхода из строя оборудования от температуры окружающей среды;
* обучить модель ARIMA для прогнозирования поломок во времени.

Практическая значимость результатов исследования заключается в том, что работа с данными с помощью применения современных методов прогнозирования становятся все более популярным направлением, но многие предприятия все еще не используют в своей работы данный инструментарий и имеют большие сложности по работе с данными.

**Апробация результатов исследования.** Результаты работы будут применимы на машиностроительном предприятии г. Москва.

# 1 Теоретические основы сбора и анализа данных

Из-за стремительного развития информационных технологий объём данных, которые хранятся в электронном виде, быстро растёт. Информация существует в разных форматах: текстовых, графических, аудио- и видеофайлах, гипертекстовых документах, реляционных базах данных и так далее. Но большая часть этих данных бесполезна для человека в связи с тем, что он не может обработать такой большой объём информации. Поэтому возникает проблема: как извлечь из большого объёма данных сведения, полезные для пользователя.

Анализ данных состоит из трех этапов (рисунок 1).



Рисунок 1 – Этапы анализа данных

## Сбор данных

Сбор данных играет важную роль в области науки о данных и инженерии.

Сбор данных – процесс сбора данных из различных источников на определенную тему. Источниками могут служить опросы, фокус-группы, интервью, анкеты, наблюдения и существующие базы данных. Собранная информация затем может быть организована в таблицы или диаграммы для дальнейшего анализа.

Данные могут быть структурированными, полуструктурированными и неструктурированными (рисунок 2). Форматы хранения и передачи данных с разной степенью структурированности представлены на рисунке 3.

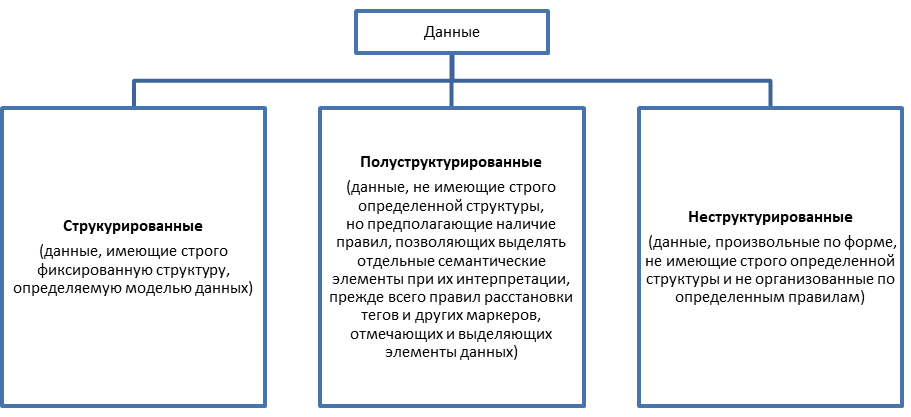


Рисунок 2 – Виды данных

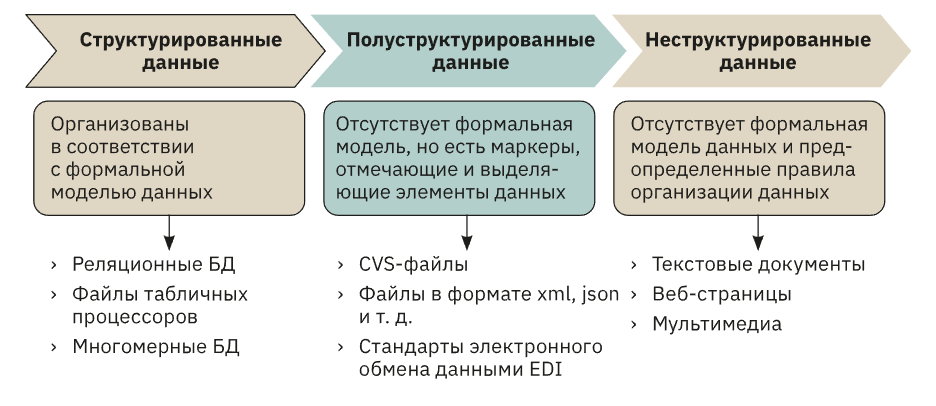


Рисунок 3 – Форматы данных

### Источники сбора данных

Источники данных изображены на рисунке 4.

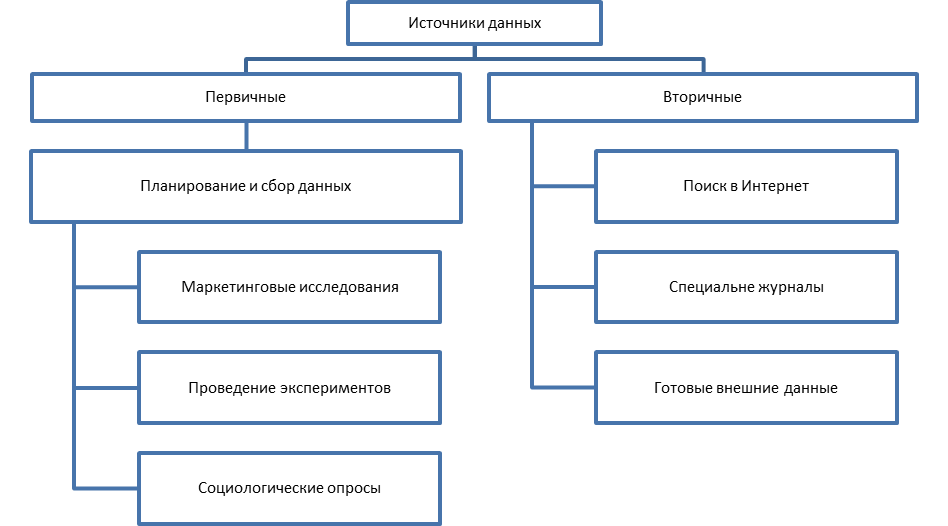


Рисунок 4 – Источники данных

Данные, впервые собираемые исследователем из первичных источников с нуля, называются первичными данными. Эти данные собираются непосредственно из источника происхождения. Это данные в режиме реального времени, которые всегда соответствуют потребностям исследователя. Первичные данные доступны в необработанном виде. Исследователю приходится тратить длительный период времени на сбор первичных данных, и, следовательно, это также дорого. Однако точность и достоверность первичных данных превышают вторичные. Некоторые примеры источниками для сбора первичных данных являются наблюдения, опросы, эксперименты, личные интервью, анкеты и т.д.

Уже существующие данные, которые ранее были собраны кем-либо другим для других целей, называются вторичными данными. В него не включены данные в режиме реального времени, поскольку исследование этой информации уже проводилось. Однако стоимость сбора вторичных данных меньше. Поскольку данные уже собирались в прошлом, их можно найти в уточненном виде. Точность и достоверность вторичных данных относительно ниже, чем первичных. Шансы найти точную информацию или данные, соответствующие потребностям исследователя, меньше. Однако время, необходимое для сбора вторичных данных, невелико и, следовательно, является быстрым и легким процессом. Некоторые примерами источниками для сбора вторичных данных являются книги, журналы, внутренние отчеты, правительственные отчеты, статьи, веб-сайты, правительственные публикации и т.д.

### Способы и инструменты сбора данных

На протяжении веков люди вручную собирали информацию. Даже в современном мире, когда сильно развита цифровизация, мы продолжаем заполнять бумажные документы, вручную вбиваем цифры в файл Excel, фиксируя события и наблюдения. Но работы с бумажными документами и ручной ввод данных отнимают много времени и сил, а кроме того, могут привести к ошибкам из-за человеческого фактора.

В небольших компаниях до сих пор используют ручной сбор данных, но крупные организации стараются этот процесс автоматизировать, используя следующие методы сбора информации.

Рассмотрим следующие методы сбора данных:

* программный интерфейс - Application programming interface (API);
* парсинг данных;
* загрузка файлов (CSV, Excel, XML и т.д.);
* интеграция с CRM-системами и т.д.

Рассмотрим подробнее несколько из вышеописанных методов.

**API (Application Programming Interface)**

API (Application Programming Interface) — это набор определений, протоколов и инструментов для создания и взаимодействия программного обеспечения. В контексте аналитики данных, API позволяет разработчикам и аналитикам обмениваться данными между различными системами, приложениями и сервисами.

Шаги использования API в аналитике данных (рисунок 5).

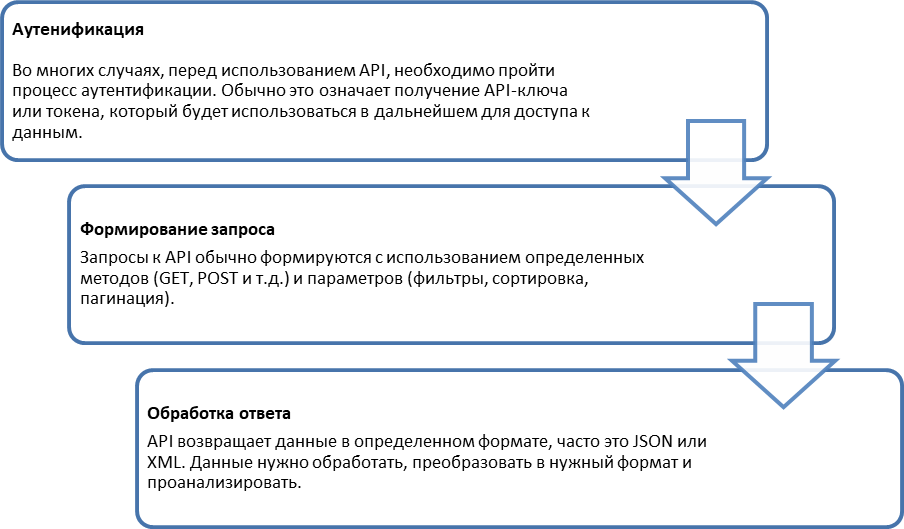


Рисунок 5 – Шаги использования API

Использования API помогает автоматизировать сбор данных, безопасность, конфиденциальность, улучшает качество полученных данных, а главное – экономит ресурсы.

**Парсинг сайтов**

Это могут быть любые выгрузки из ПО и сервисов для дальнейшего анализа в формате, который в дальнейшем можно обработать.

Парсинг — это процесс автоматического сбора и структурирования данных с помощью скриптов (парсеров). Другое название этого процесса — веб-скрейпинг.

Парсеры работают на разных языках программирования — Python, JavaScript, PHP 5 и т.д.

Парсинг нужен для того, чтобы облегчить и ускорить выполнение рутинных задач. Представьте, сколько времени потребуется человеку, чтобы собрать и систематизировать информацию о тысяче статей с веб-сайта в таблице — это может занять несколько часов. В то же время парсер способен выполнить эту работу за считанные минуты. Парсер значительно ускоряет рабочий процесс и позволяет сократить количество ошибок по сравнению с ручным трудом.

Этапы парсинга представлены на рисунке 6.

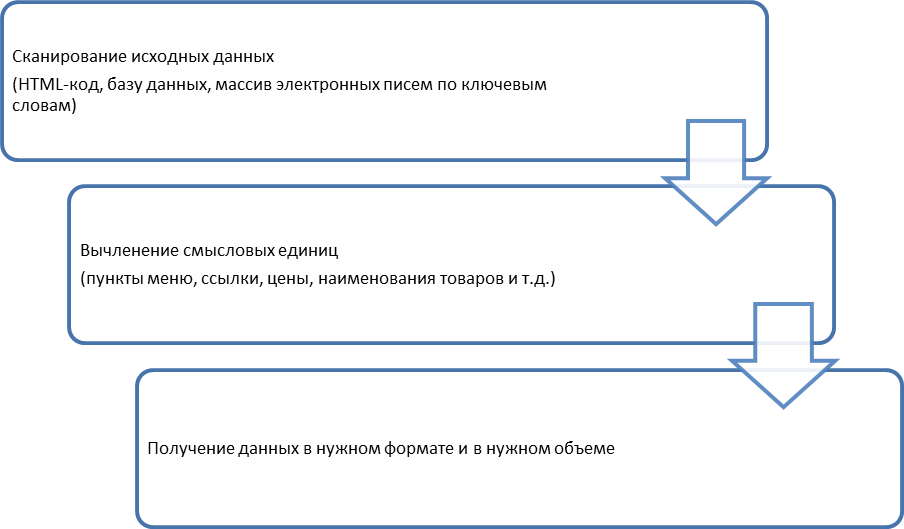


Рисунок 6 – Этапы парсинга

### Хранение данных

Существует два типа хранилищ данных: облачное и традиционное (локальное).

**Облачное хранилище**

Облачное хранилище – это вариант хранения, при котором мы используем удаленные диски для хранения данных в облачном хранилище, используемом клиентом. Для хранения данных на удаленном сервере, принадлежащем поставщику услуг, также используется сеть. Пользователь использует эти параметры хранилища для определения емкости, пропускной способности и удаленного доступа.

Особенности облачного хранилища:

* облачное хранилище предлагает множество вариантов защиты данных.
* Эти варианты хранения данных легко доступны с любого устройства, подключенного к Интернету.
* Ошибки в облачном хранилище легко отслеживаются.

**Традиционное хранилище**

Традиционное хранилище – это вариант хранения, при котором мы используем локальные физические диски для хранения данных в основном местоположении клиента. Пользователь обычно использует дисковое оборудование для хранения данных, которое используется для копирования, управления данными и интеграции их в программное обеспечение.

Особенности традиционного хранилища:

* Традиционные хранилища быстры, так как независимы от скорости Интернета.
* Безопасность настраивается пользователем вручную.
* Пользователи имеют возможность восстанавливать данные в любое время без проблем с доступом.
* Резервное копирование и модификация данных на месте просты.

Различия между этими двумя системами представлены на рисунке 7.



Рисунок 7 – Различия облачного и традиционного хранения

## 1.2 Подготовка данных

Специалисты по работе с данными тратят около 80% всего времени на то, чтобы привести данные к нужному виду для дальнейшего анализа и работы с ними. Именно поэтому данный процесс является очень важной частью в анализе данных.

Обработка данных (или предобработка) — это значимый процесс, в ходе которого сырые данные, содержащие ошибки, дубликаты и пропуски, доводятся до пригодного для анализа состояния. От качества этих данных зависит качество информации, на основе которой принимаются решения.

Начало формы

Конец формы

### 1.2.1 Методы очистки данных

Исправление или удаление ошибочных, неполных и некорректных данных называется очисткой данных. Вот несколько подходов, которые могут вам в этом помочь:

* удаление дубликатов;
* заполнение пропущенных значений;
* обработка выбросов;
* стандартизация и нормализация.

**Удаление дубликатов**

Повторяющиеся данные способны негативно повлиять на результаты анализа. Чтобы исключить дублирующиеся записи, можно использовать такие функции, как drop\_duplicates() в Python. Дубликаты могут появляться из-за разных факторов, например, при ошибочном вводе данных или объединении нескольких источников информации. Удаление дубликатов помогает повысить качество данных и точность анализа.

**Заполнение пропущенных значений**

Отсутствующие данные можно заменить на средние, медианные или модальные значения. В Pandas для этого используется метод fillna(). Пропуски могут появляться по разным причинам: из-за ошибок при сборе информации или из-за её отсутствия. Замена пропущенных значений помогает избежать ошибок в анализе и повышает качество данных.

**Обработка выбросов**

Аномальные значения способны исказить результаты анализа. Чтобы выявить и обработать выбросы, можно использовать такие методы, как межквартильный размах (IQR) или Z-оценка. Выбросы могут появляться по разным причинам: из-за ошибок при вводе данных или из-за необычных явлений. Устранение выбросов помогает повысить качество данных и точность анализа.

**Стандартизация и нормализация**

Эти методы позволяют привести данные к общему виду, что особенно важно для алгоритмов машинного обучения. В Python для этого можно использовать библиотеку sklearn.preprocessing. Применение стандартизации и нормализации помогает повысить качество данных и точность анализа.

### 1.2.2 Инструменты очистки данных

Инструменты для очистки данных представлены на рисунке 8.

Рисунок 8 – Инструменты очистки данных

### 1.2.3 Методы обработки данных

После того, как мы очистим данные, необходимо их обработать. Вот несколько подходов, которые могут вам в этом помочь:

* агрегация данных;
* трансформация данных;
* фильтрация данных;
* слияние данных.

**Агрегация данных**

Объединение и обобщение данных из разных источников в один набор называется агрегацией данных. Этот процесс подготавливает информацию для анализа, позволяя увидеть закономерности и идеи, которые невозможно обнаружить при рассмотрении отдельных точек данных. К примеру, можно вычислить средний объём продаж по месяцам. В Pandas для этого применяют метод groupby(). Агрегация данных позволяет сделать анализ проще и понятнее.

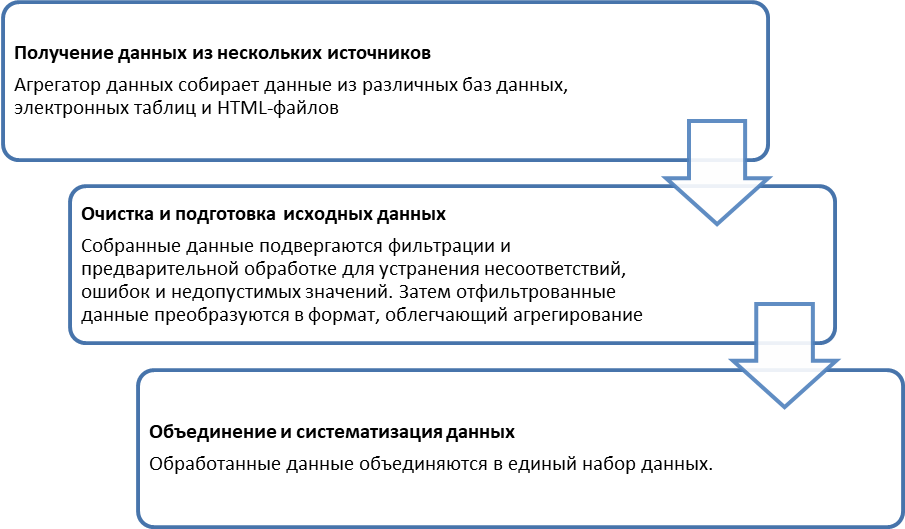
Процесс агрегации состоит из трех этапов, представленных на рисунке 9.

Рисунок 9 – Этапы агрегации

**Трансформация данных**

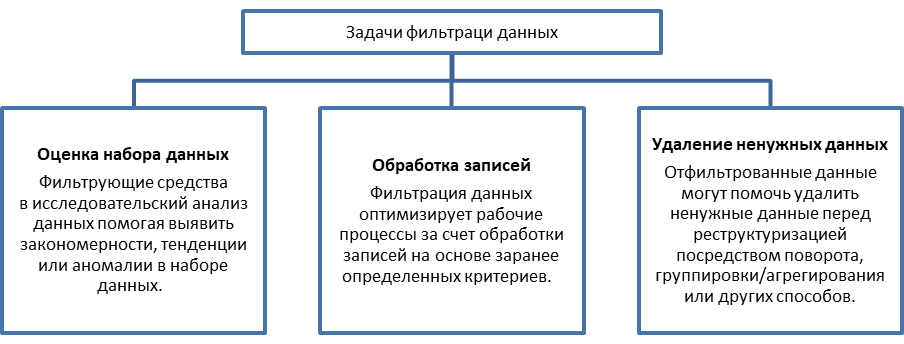
Трансформация данных — это преобразование и согласование наборов данных друг с другом или с определённой схемой. Она заключается в оптимизации представлений и форматов данных с точки зрения решаемых задач и целей анализа. Например, категориальные данные можно преобразовать в числовые с помощью метода get\_dummies() в Pandas. Преобразование данных способствует повышению их качества и делает анализ более точным.

Основная задача данного метода – преобразовать данные в такой вид, чтобы их можно было бы максимально эффективно использовать.

Методы трансформации данных: нормализация, преобразование типов и форматов, сортировка, группировка и т.д.

Данные операции производят после получения данных из разных источников и обеспечивает их дальнейшую обработку.

**Фильтрация данных**

Фильтрация данных — это процесс выделения самых необходимых данных из большого набора данных с использованием определённых условий или критериев. Она дает возможность быстро анализировать нужные данные, избегая просмотра всего набора, что делает анализ более целенаправленным и эффективным.

Задачи фильтрации представлены на рисунке 10.

Рисунок 10 – Задачи фильтрации

Основные методы фильтрации данных представлены на рисунке 11.

Рисунок 11 – Методы фильтрации

**Слияние данных**

Слияние данных — это процесс, при котором объединяются различные источники информации для получения более согласованной, точной и полезной информации по сравнению с данными из одного отдельного источника.

Этот процесс включает в себя добавление новых деталей к уже имеющимся данным, интеграцию новых случаев и удаление дублирующей или некорректной информации. В Pandas для этой цели применяются методы merge() и concat(). Объединение данных способствует повышению их качества и делает анализ более точным.

### 1.2.4 Инструмент обработки данных

Инструменты для обработки данных представлены на рисунке 12.

Рисунок 12 – Инструменты обработки данных

Обработка данных является ключевым этапом в аналитике. Применяя методы и инструменты, описанные выше, мы можем эффективно подготовить данные и добиться точных результатов. Качество данных играет важнейшую роль в аналитике, так как именно от него зависит надежность получаемых результатов. Некачественные данные могут привести к ошибочным выводам и негативно сказаться на бизнесе.

## 1.3 Анализ данных

Исследование данных для обнаружения полезной информации, выводов и помощи в принятии решений называется анализом данных. В мире, где информация имеет огромное значение, способность анализировать данные становится всё более востребованной. Анализ данных охватывает множество методов и подходов, которые позволяют исследовать информацию с разных сторон. Эти методы помогают выявлять скрытые тенденции, предсказывать будущие события и принимать взвешенные решения. Важно помнить, что каждый метод имеет свои особенности и ограничения, поэтому выбор метода зависит от конкретной задачи и типа данных.

### 1.3.1 Методы анализа данных

Методы анализа данных значительно эволюционировали, предоставляя комплексный инструментарий для понимания, интерпретации и прогнозирования шаблонов данных. Эти методы имеют решающее значение для извлечения полезной информации из данных, позволяя организациям принимать обоснованные решения.

Основные методы анализа данных:

* описательный анализ;
* качественный анализ;
* прогнозный анализ;
* диагностический анализ;
* регрессионный анализ;
* факторный анализ;
* кластерный анализ и т.д.

**Описательный анализ**

Описательный анализ рассматривается как начальный этап аналитического процесса и обычно направлен на то, чтобы ответить на вопросы о произошедших событиях. Он включает в себя систематизацию данных, работу с ними и их интерпретацию из различных источников для получения ценной информации.

Пример:

* Показатели продаж: компания розничной торговли может использовать описательную статистику, чтобы понять средний объем продаж по магазинам или определить, какие товары пользуются наибольшим спросом.
* Опросы удовлетворенности клиентов: Анализ данных опроса для поиска наиболее распространенных ответов или средних баллов.

**Качественный анализ**

Методы качественного анализа данных не предполагают прямого измерения, поэтому этот подход применяется, когда организации нужно принимать решения на основе субъективной интерпретации. Например, качественные данные могут использоваться для оценки отзывов клиентов, влияния вопросов опроса, эффективности публикаций в социальных сетях, анализа конкретных изменений или особенностей продукта и многого другого. Качественный анализ также помогает систематизировать данные по темам или категориям, которые впоследствии можно автоматизировать.

Пример:

* Анализ рынка: Бизнес может проанализировать, почему продажи продукта резко выросли в определенном квартале, изучив маркетинговую активность, изменения цен и тенденции рынка.

Медицинская диагностика: Врачи используют диагностический анализ для понимания причины симптомов на основе результатов лабораторных исследований и данных пациента.

**Прогнозный анализ**

Прогностический анализ данных даёт возможность прогнозировать будущие события, отвечая на вопрос «что произойдёт?». Чтобы применять этот метод, нужно использовать результаты описательного анализа данных, методы исследовательского и диагностического анализа, а также сочетать машинное обучение и искусственный интеллект. Благодаря этому подходу можно получить представление о будущих тенденциях и выявить потенциальные проблемы и слабые места в наборе данных. Также с помощью точных данных компании могут находить и разрабатывать идеи для улучшения операционных процессов и конкурентного преимущества.

Пример:

* Кредитный рейтинг: Финансовые учреждения используют прогностические модели для оценки вероятности дефолта клиента по кредиту.
* Прогнозирование погоды: Метеорологи используют прогностические модели для прогнозирования погодных условий на основе исторических данных о погоде.

**Диагностический анализ**

Зная причину произошедшего, нетрудно определить способы, которые к этому привели. Например, с помощью диагностического анализа можно понять, из-за чего снизились продажи, и в итоге изучить конкретные факторы, которые привели к потерям.

Пример:

* Анализ запасов: Проверка, коррелирует ли снижение продаж с нехваткой или переизбытком запасов.
* Эффективность продвижения: Анализ влияния различных рекламных кампаний, чтобы определить, какие из них не смогли привлечь клиентов.

**Регрессионный анализ**

Суть этого метода заключается в анализе исторических данных для понимания влияния изменений независимых переменных на зависимую переменную. Выявляя взаимосвязи между переменными и прошлыми событиями или инициативами, вы можете предсказать вероятные результаты в будущем. Этот подход помогает вам принимать взвешенные решения и выбирать наиболее эффективный путь.

Пример:

* Оценка рыночных тенденций: Оценка того, как изменения в экономической среде (например, процентные ставки) влияют на цены на недвижимость.
* Прогнозное ценообразование: Использование исторических данных для прогнозирования будущих ценовых тенденций на основе текущей динамики рынка.

**Факторный анализ**

Факторный анализ данных направлен на выявление закономерностей среди связанных переменных с помощью скрытых факторов. Ина-че говоря, он помогает выделить независимые переменные, что полезно для улучшения определённых областей.

Пример:

* Улучшение обслуживания: Определение ключевых факторов, таких как время ожидания, поведение персонала и результат лечения, которые влияют на удовлетворенность пациентов.
* Распределение ресурсов: Использование этих данных для улучшения областей, которые существенно влияют на удовлетворенность пациентов.

**Кластерный анализ**

Кластерный анализ позволяет наглядно представить данные и выявить в них общие закономерности. Он часто применяется, когда нужно сделать информацию более понятной или провести её анализ, а также если категории данных неоднозначны. В ходе этого процесса похожие наблюдения группируются в кластеры, что помогает присвоить группам названия и определить их категории.

Пример:

* Сегментация рынка: Разделение клиентов на группы, которые демонстрируют схожее поведение и предпочтения для более целенаправленного маркетинга.
* Настройка кампании: Разработка уникальных маркетинговых стратегий для каждого кластера для максимального привлечения и конверсий.

### 1.3.2 Инструменты анализа данных

Чтобы успешно работать с большими объёмами информации, аналитикам нужно использовать ряд особых инструментов и технологий. Эти ресурсы помогают быстро и эффективно извлекать, обрабатывать и понимать данные, что позволяет достигать поставленных целей в кратчайшие сроки. Важно помнить, что каждый инструмент имеет свою терминологию и назначение, поэтому для их использования требуется определённая подготовка и знания.

Инструмент представлены на рисунке 13.

Рисунок 13 – Инструменты анализа данных

# 2 Практическая часть работы

Данная работы будет производиться на базе машиностроительного предприятия г. Москва.

В настоящий момент предприятие находится на этапе цифровизации, что заставляет пересмотреть многие бизнес-процессы.

В рамках проекта по оборудованию было предложено собрать и проанализировать имеющиеся данные по выходу из строя оборудования, найти закономерности (в том числе от температуры окружающей среды) и создать проект по визуализации (дашборд).

## 2.1 Сбор данных

### 2.1.1 Сбор и обработка данных, выгруженных из ПО предприятия

На машиностроительном предприятии имеется ПО «ОРО», где фиксируются поломки оборудования с номером заявки, датой подачи, наименованием оборудования и другими характеристиками. Скриншот представлен в приложении А.

Данные находятся в разных разделах. В процессе работы было выгружено 9 файлов в формате .xlsx.

Рассмотрим данный процесс пошагово.

1. Импортируем все заявки (1…9.xlsx) и объединим в один файл (Приложение Б, рисунок Б.1, Б.2).
2. Заполняем пустые значения в столбце «Дата закрытия заявки» (Приложение Б, рисунок Б.3).
3. Рассчитаем длительность ремонта оборудования (Приложение Б, рисунок Б.4).
4. Изменим регистр по нескольким столбцам.
5. Удалим дубликаты данных (Приложение Б, рисунок Б.5).
6. Удалим пустые значения по двум столбцам (Приложение Б, рисунок Б.6).

Полученные данные сохраняем в CSV-файл.

### 2.1.2 Парсинг погодных данных

Для проверки зависимости частоты выхода из строя оборудования от температуры воздуха необходимо найти погодные данные.

Для выполнения поставленной задачи были проанализированы сайты с погодой по Москве и выбран источник <https://ginfo.ru>.

На сайте имеется информация о погоде с 2019 по 2024 год.

Процедура парсинга сайта:

1. Парсинг данных (Приложение В, рисунок В.1).
2. Преобразование данных (Приложение В, рисунок В.2, В.3, В.4, В.5).

Полученные данные сохраняем в CSV-файл.

### 2.1.3 Объединение данных

Произведем объединение файлов с заявками и температурой.

Шаги:

1. Импортируем файлы request.csv и weather.csv (Приложение Г, рисунок Г.1).
2. Изменение типа данных даты (Приложение Г, рисунок Г.2).
3. Объединение таблиц по дате (Приложение Г, рисунок Г.3).
4. В связи с тем, что данные с температурными показателями с 2019 года, необходимо удалить значения без температуры (Приложение Г, рисунок Г.4). После удаление остается 10535 строк.
5. Выделим из столбца «Инв.номер» категорию и номер (Приложение Г, рисунок Г.5).
6. Изменим типы данных (Приложение Г, рисунок Г.6).
7. Изменим расположение столбцов (Приложение Г, рисунок Г.7).

Полученные данные сохраняем в CSV-файл.

## 2.2 Визуализация данных

Визуализация данных — важный инструмент для анализа и интерпретации информации. Она превращает сложные массивы данных в понятные графики, диаграммы и таблицы. В данной работе мы воспользуемся инструментом Power BI от Microsoft, который предоставляет широкие возможности для визуализации данных.

### 2.2.1 Работа с данными в Power Query

Загрузим полученную таблицу join.csv с заявками и температурой в Power BI и преобразуем данные в Power Query (Приложение Д, рисунок Д.1).

Первым шагом мы должны преобразовать данные в таблице.

Затем создаем новые столбцы с необходимой информацией для будущего визуального отчета:

* Проверяем наименование цехов и удаляем несколько нечищенных данных через фильтр.
* Выделяем месяц из даты (январь, февраль и т.д.).
* Выделяем ЧПУ и универсальное оборудование (Приложение Д, рисунок Д.2).

**2.2.2 Создание дашборда в Power BI**

Power BI — это инструмент визуализации данных и бизнес-аналитики от Microsoft, который преобразует данные из разных источников для создания различных отчетов бизнес-аналитики. Microsoft Power BI позволяет компаниям легко выявлять тенденции, отслеживать производительность и принимать решения на основе данных.

Он может принимать входные данные из таблиц Excel, баз данных или даже облачного хранилища. Вы можете подключать источники данных, получать информацию и делиться ею с другими. Power BI - самый популярный инструмент бизнес-аналитики.

Первый шаг – это проработка проекта визуального отчета. Желание руководства – видеть данные о количестве поломок для каждого цеха и категории с удобной навигацией. Параметры для отчетности:

* количество поломок за год и по месяцам;
* количество поломок по возрасту;
* количество поломок по категориям;
* количество поломок по моделям;
* количество поломок по ЧПУ и универсальному оборудованию;
* количество поломок по температуре окружающей среды (по возможности).

Данные пожелания были учтены в процессе выполнения работы.

Проект дашборда представлен в приложении Д.

В настоящий момент прорабатывается вопрос дополнить дашборд общей информацией об оборудовании (не по поломкам). Собираются данные.

**2.2.3 Рекомендации**

В процессе выполнения работы были выявлены недостатки текущей системы сбора информации по оборудованию.

В настоящий момент пишется техническое задание на доработку ПО «ОРО» по следующим направлениям:

* Добавить столбец «Дата реагирования на поломку», для возможности отслеживания скорости реагирования службой механика.
* Добавить столбец «Категория поломки»: механическая, гидравлическая и т.д. Названия категорий должны быть унифицированы, заполняться из списка, чтобы избежать ошибок.
* Добавить столбец «ЗиП», где из существующего списка необходимо выбрать одно или несколько приспособлений, вышедших из строя. Это позволит в дальнейшем прогнозировать поломку и грамотно формировать страховой запас ЗиП.

Данные изменения позволят более эффективно использовать возможности данного ПО.

## 2.3 Обучение модели ARIMA прогнозу поломок

### 2.2.1 Модель ARIMA

Анализ временных рядов имеет большое значение во многих сферах, включая финансы, экономику и метеорологию. Модель авторегрессионного интегрированного скользящего среднего (ARIMA) — один из основных инструментов для предсказания будущих значений на основе исторических закономерностей в данных временных рядов. Но чтобы прогнозы были точными, важно правильно подобрать параметры для модели ARIMA.

Модель объединяет три ключевых компонента для моделирования данных (рисунок 14).

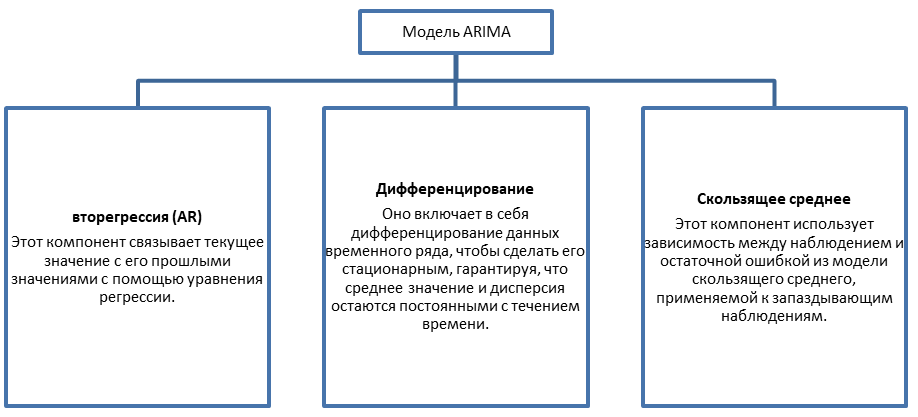


Рисунок 14 – Модель ARIMA

* Оценка параметров: оценка параметров p, d и q включает анализ графиков автокорреляционной функции (ACF) и частичной автокорреляционной функции (PACF) данных временных рядов. ACF помогает определить порядок MA (q), в то время как PACF помогает определить порядок AR (p).
* Подгонка модели: после определения параметров модель ARIMA подгоняется к данным. Это предполагает минимизацию ошибки (часто с использованием таких методов, как оценка максимального правдоподобия) для получения наиболее подходящих коэффициентов для условий авторегрессии и скользящего среднего.
* Прогнозирование: после подгонки модели ее можно использовать для прогнозирования будущих значений путем итерации с течением времени.

Параметры модели:

* p (порядок изменений): представляет количество членов авторегрессии и обозначается буквой p. Это относится к количеству прошлых наблюдений, которые напрямую влияют на текущее значение.
* d (порядок различий): представляет количество различий, необходимых для того, чтобы сделать временной ряд стационарным. Это включает вычисление различий между последовательными наблюдениями.
* q (порядок MA): обозначаемый q, он представляет количество запаздывающих ошибок прогноза в уравнении прогнозирования.

Выбор подходящих значений для этих параметров существенно влияет на способность модели к прогнозированию. Однако определение правильных значений часто является сложной задачей.

Методы выбора модели для ARIMA представлен на рисунке 15.

Рисунок 15 - Методы выбора модели

### 2.2.2 Обучение модели

Для возможности прогнозировать поломку оборудования во времени было принято воспользоваться моделью ARIMA и обучить ее на имеющемся датасете.

Перед тем, как приступить к обучению, мы проверим наши данные на пропуски (Приложение Е, рисунок Е.1).

Затем из имеющегося датасета мы выделим необходимые значения: дату подачи заявки и количество поломок в каждую дату (Приложение Е, рисунок Е.2).

Проверим, есть ли зависимость в нашем датасете количества поломок от температуры (при составлении дашборда это визуально подтвердилось). На получившемся графике также можно проследить эту закономерность, но не так явно (Приложение Е, рисунок Е.3).

Следующий шаг – проверка на стационарность. Для этого воспользуемся тестом Дики-Фуллера (Приложение Е, рисунок Е.4). Наши данные являются стационарными.

Теперь можно приступать к построению и оценке модели (Приложение Е, рисунок Е.5). Проанализируем полученные данные:

* Коэффициенты: (P>|z|) меньше 0,05.
* AIC: 12361.004 – не самая хорошая подгонка модели.
* L1: 0,09 – тоже не лучший показатель (больше 0,05).
* JB: 0,00 – должно быть больше 0,05 для того, чтобы остатки были нормальны.
* H: 1,30, что больше о,05 – это говорит о том, что остатки гетероскедастичны, это хорошо.

Построим график (Приложение Е, рисунок Е.6). По нему мы видим, что модель не особо точна.

Попробуем автоматическую прогонку модели. Ее метрики выглядят лучше, но все еще не идеальны (Приложение Е, рисунок Е.7). Оценим эту модель с помощью графиков (Приложение Е, рисунок Е.8). По графикам мы видим, что распределение не является нормальным, ее еще надо улучшать.

# ЗАКЛЮЧЕНИЕ

Предприятию, чтобы сохранять конкуренцию на рынке и эффективно использовать имеющиеся ресурсы необходимо как можно больше использовать возможности цифровизации и в частности, уметь работать с большими данными.

В результате выполнения дипломной работы были решены следующие задачи:

* проведен анализ данных по выходу из строя оборудования;
* построен дашборд;
* изучена зависимость выхода из строя оборудования от температуры внешней среды;
* поработали с моделью ARIMA для прогнозирования поломок и пришли к выводу, что необходимо дальше продолжать работу в этом направлении.

Проведенная работы показала, что имеется зависимость выхода из строя оборудования от времени года, а также рост поломок с каждым годом.

Были разработаны рекомендации по корректировки работы ПО «ОРО» для более качественного сбора данных о поломках оборудования машиностроительного предприятия.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ильяшенко О. Ю. Роль BI–систем в совершенствовании процессов обработки и анализа бизнес информации [Текст]: учебник / И.В. Ильин, Д.Д. Болобонов. – Наука и бизнес: пути развития, №6, 2017.
2. Немуров Е.В., Золотухина Е.Б. Актуальность внедрения BI систем на предприятиях в условиях современного рынка [Текст] // Международный научно-технический журнал «Теория. Практика. Инновации». – 2018.
3. Агамиров Л. В. Статистические методы анализа результатов научных исследований : учебное пособие / Л. В. Агамиров. - Москва: Изд-во МЭИ, 2018. - 71 с.
4. Белоусов П. А., Марухина О. В., Скоморохов А. О. Машинное обучение и большие данные : учебное пособие / П. А. Белоусов, О. В. Марухина, А. О. Скоморохов. - Санкт-Петербург : ГУАП, 2021. - 119 с.
5. Болдырев А. В. Технологии хранения данных / А. В. Болдырев. - Ростов-на-Дону: ДГТУ, 2019. - 77 с.
6. Келлехер Д., Тирни Б. Наука о данных: базовый курс / Джон Келлехер, Брендан Тирни. - Москва : Альпина Паблишер, 2020. – 220 с.

# ПРИЛОЖЕНИЕ А Отчет в ПО «ОРО»

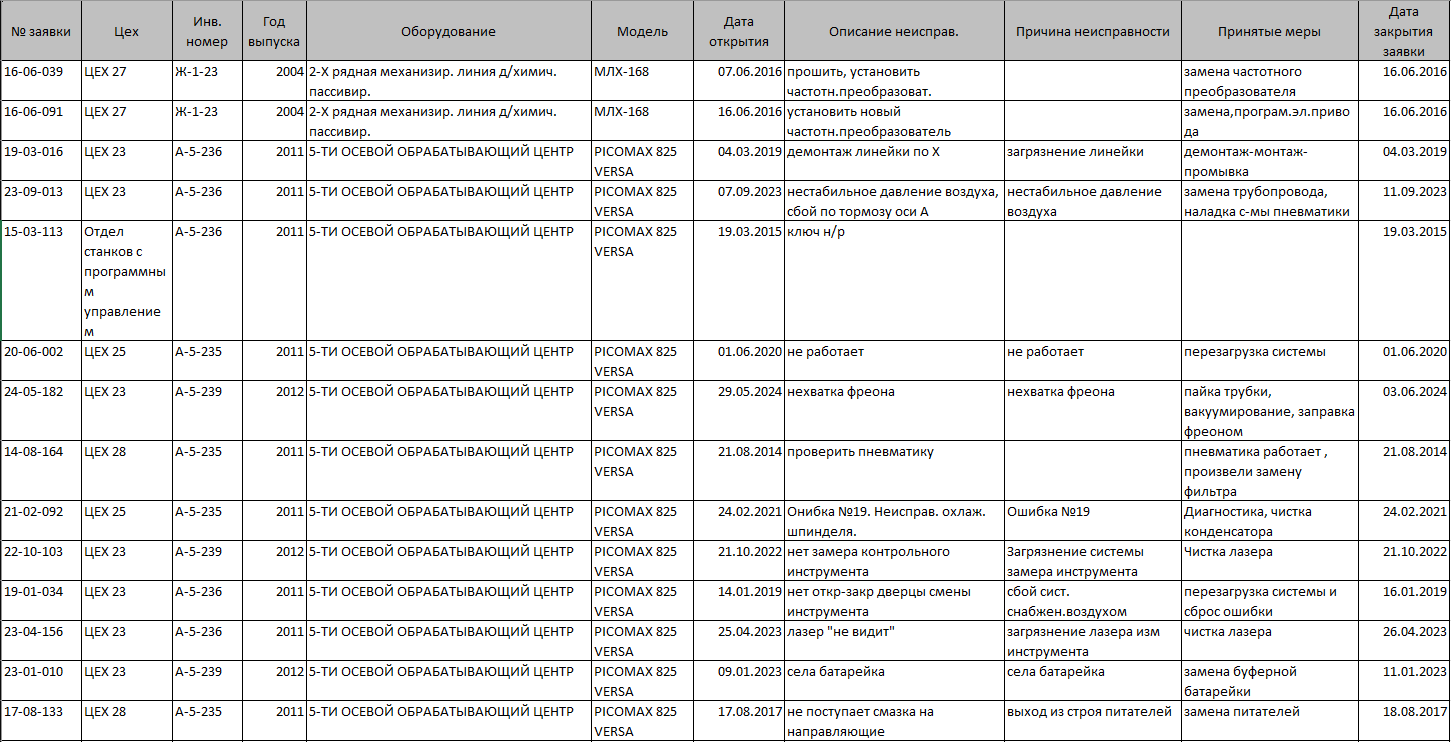


Рисунок А.1 - Отчет

# ПРИЛОЖЕНИЕ Б Объединение и преобразование заявок в один файл

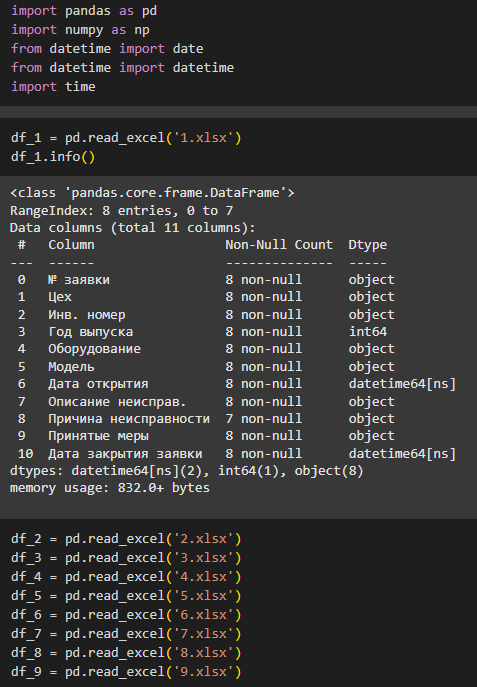


Рисунок Б.1 – Импорт файлов

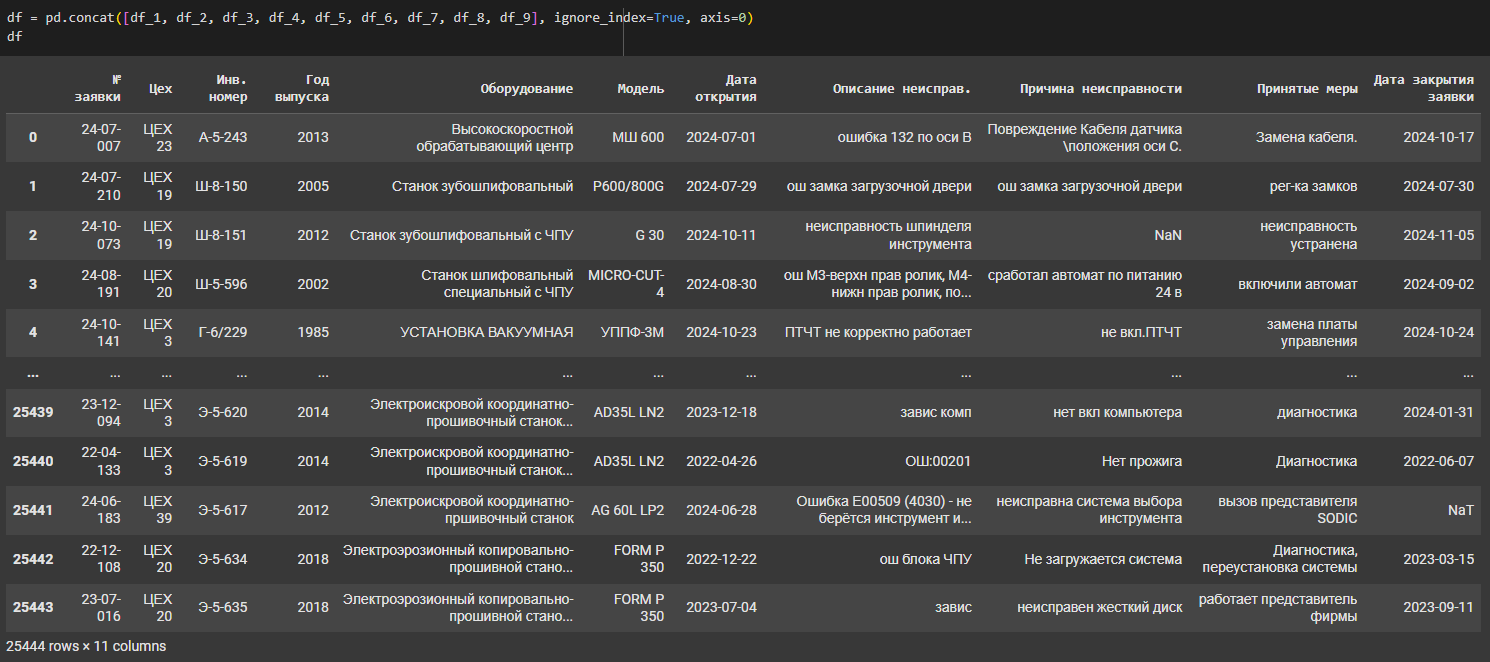
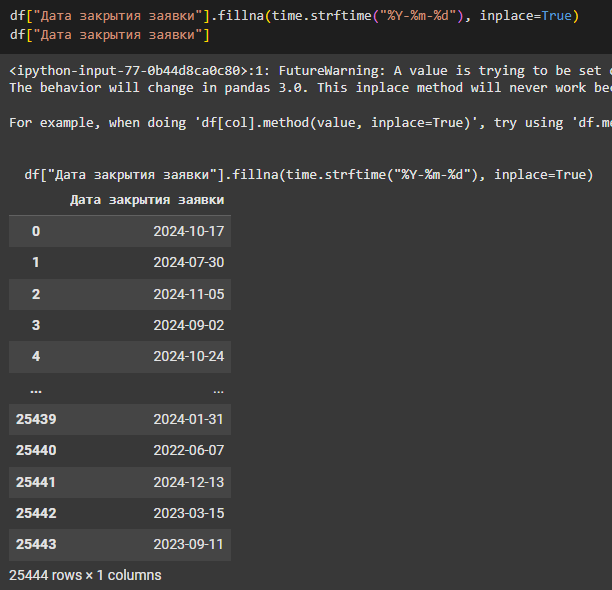


Рисунок Б.2 – Объединение файлов в один файл

Рисунок Б.3 – Заполнение пропусков даты закрытия заявки

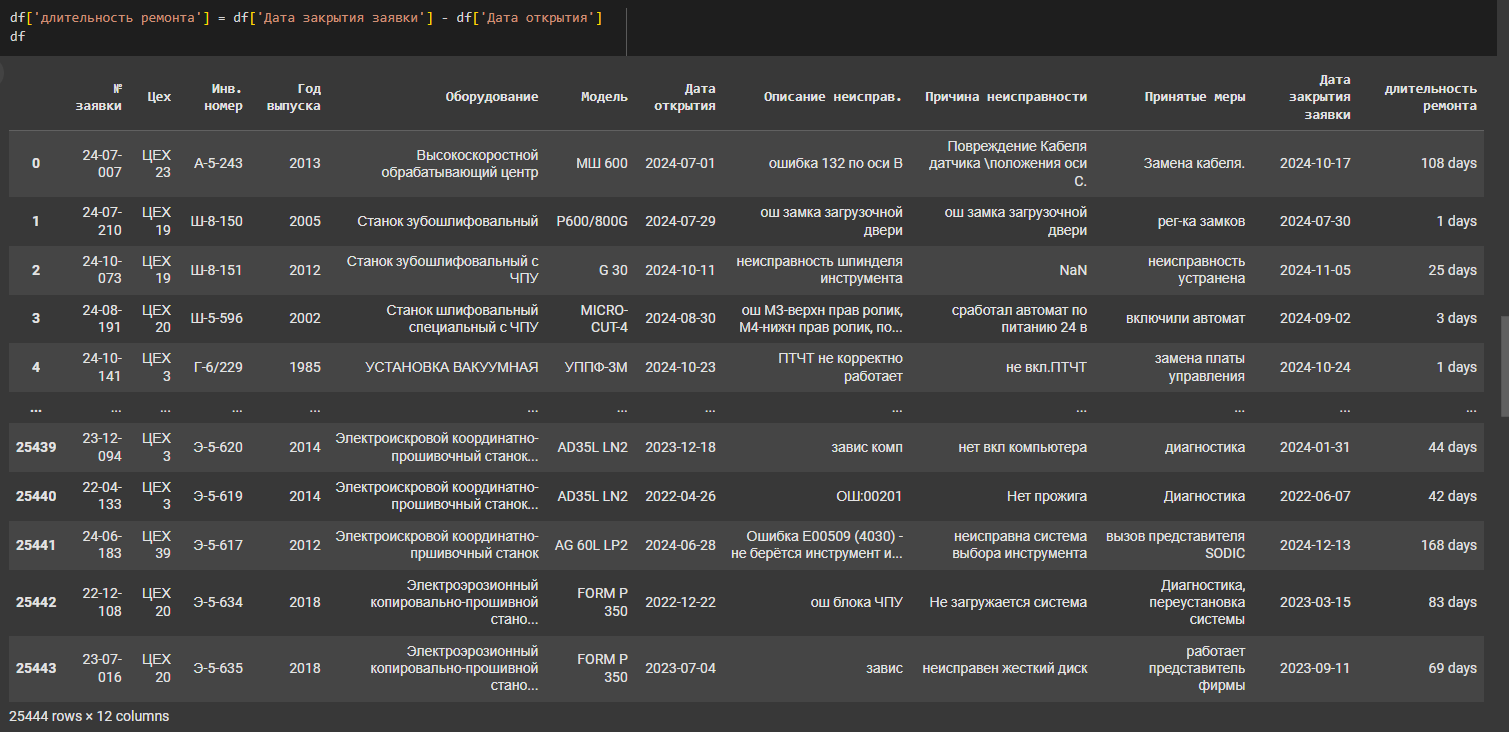


Рисунок Б.4 – Расчет продолжительности ремон

Рисунок Б.4 – Расчет длительности ремонта

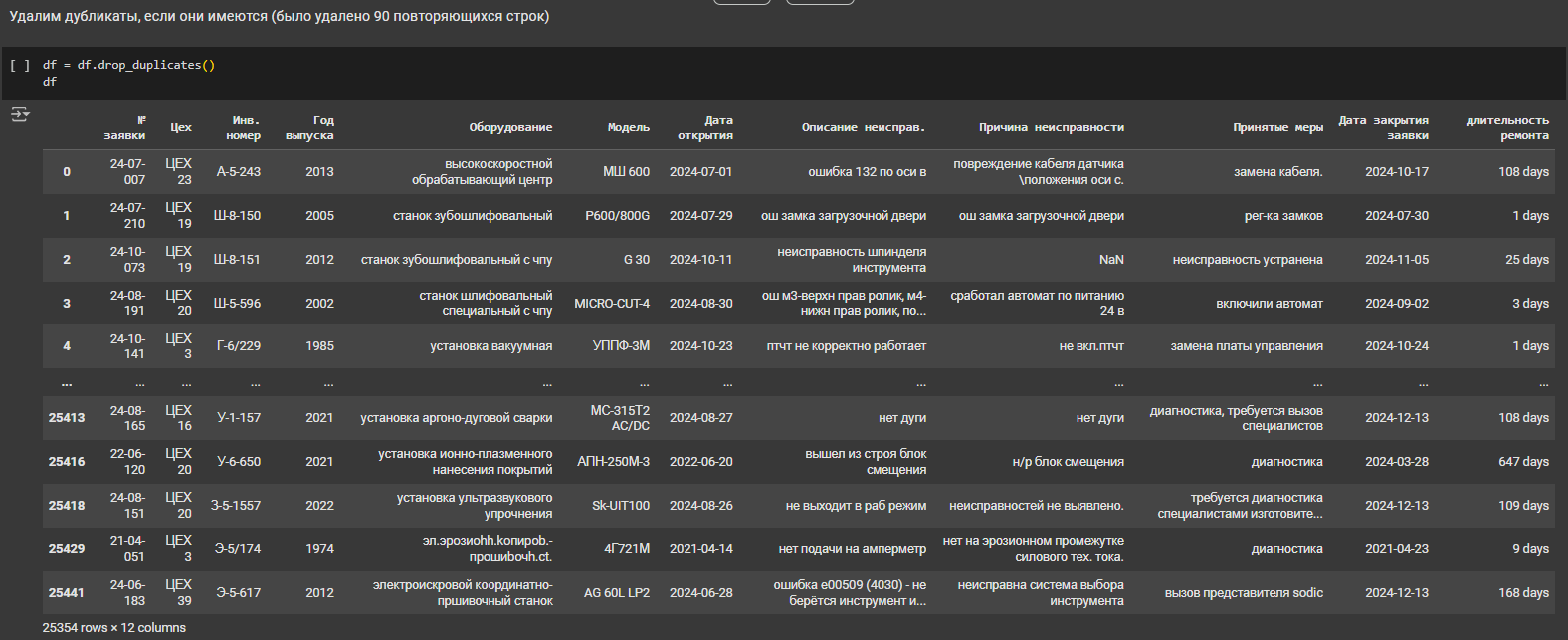


Рисунок Б.5 – Удаление дубликатов

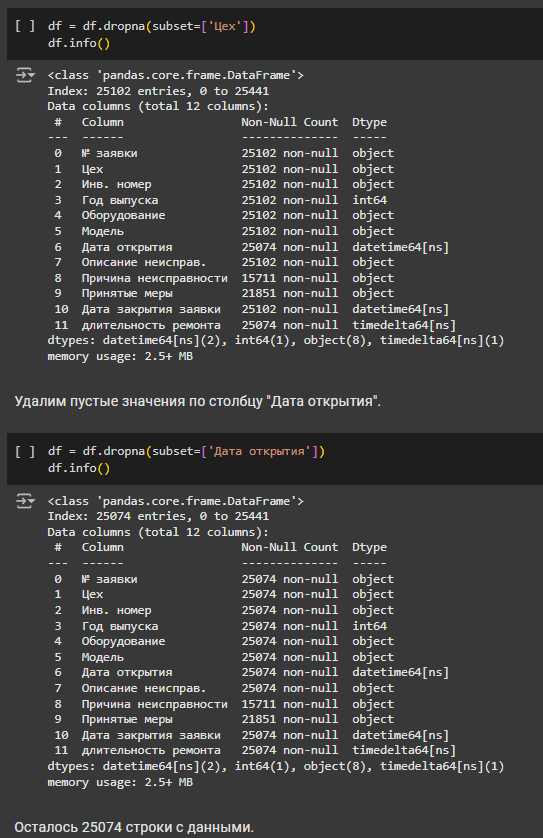
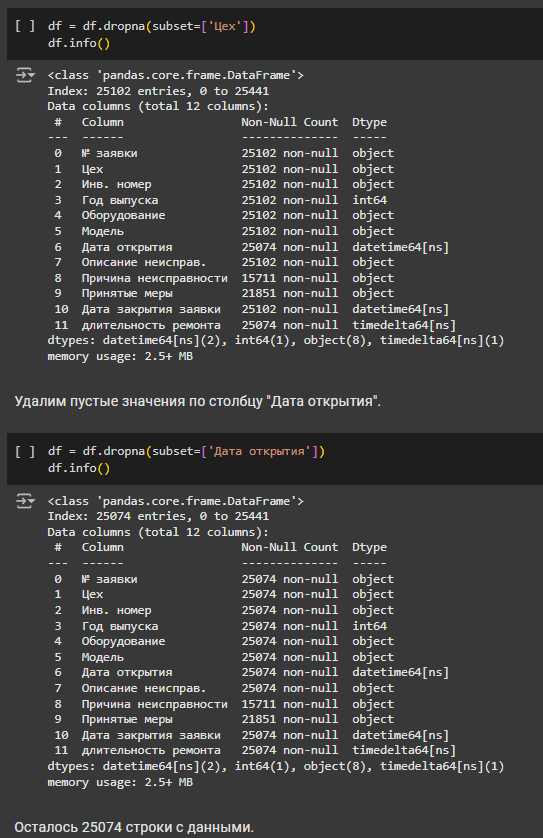
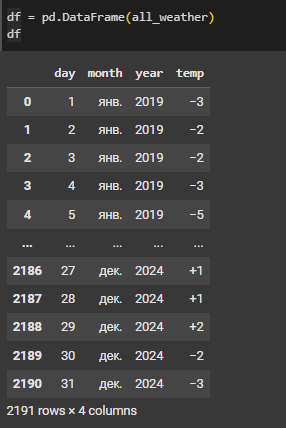


Рисунок Б.6 – Удаление пустых значений по двум столбцам

# ПРИЛОЖЕНИЕ В Парсинг сайта погоды



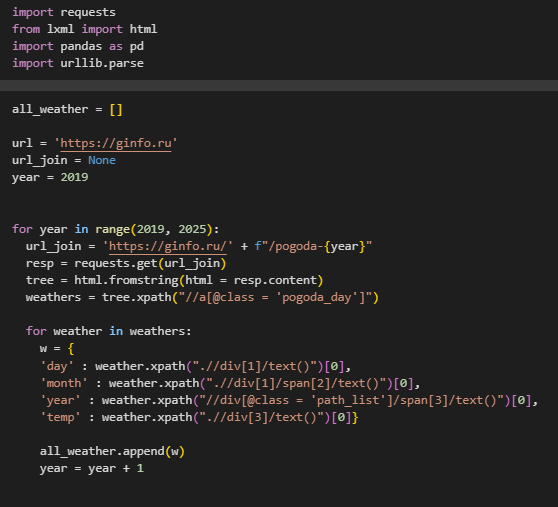


Рисунок В.1 – Парсинг данных

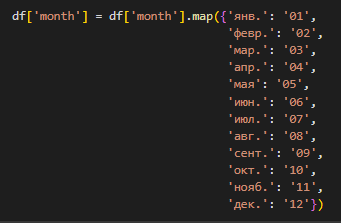
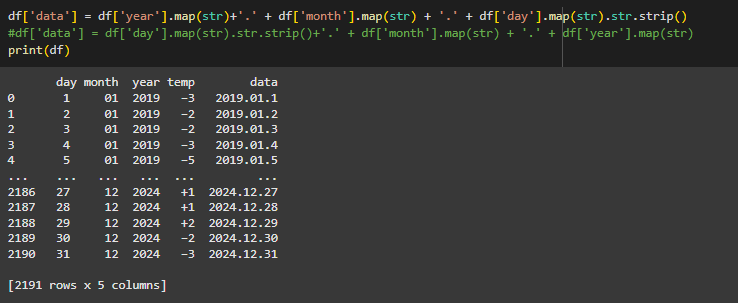


Рисунок В.2 – Преобразование даты

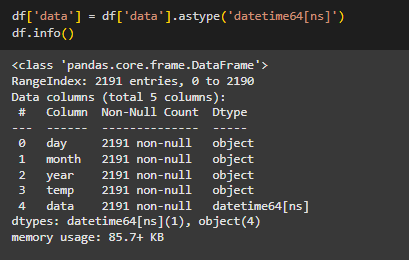


Рисунок В.3 – Преобразование даты

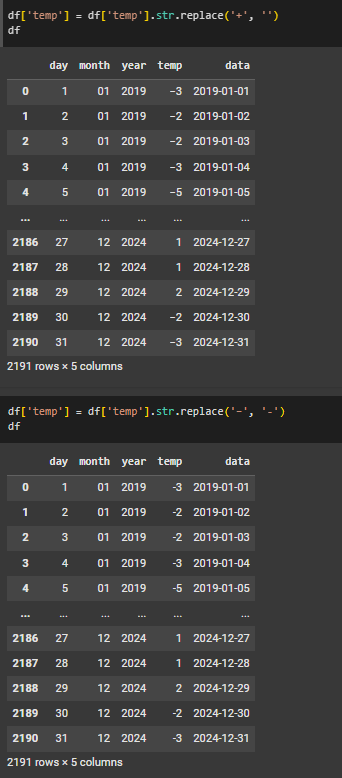
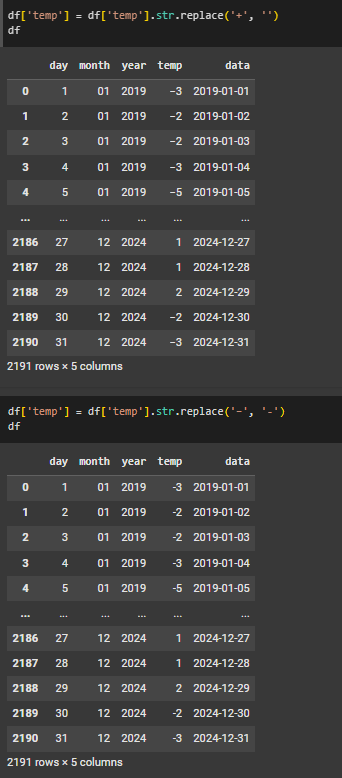


Рисунок В.4 – Преобразование температуры

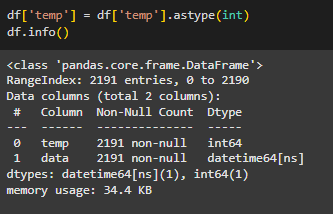


Рисунок В.5 – Преобразование температуры

# ПРИЛОЖЕНИЕ Г Объединение данных

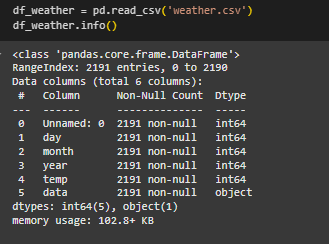
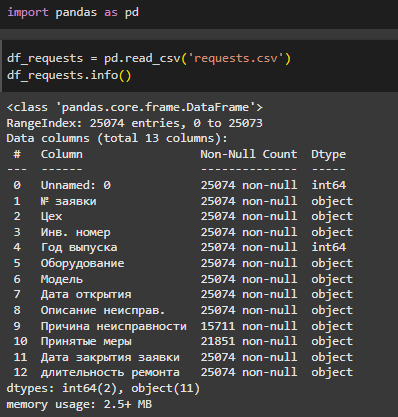
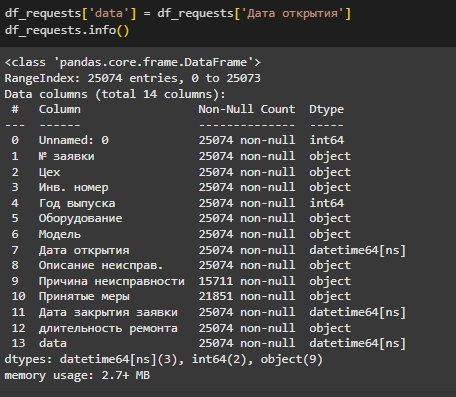


Рисунок Г.1 – Импорт данных



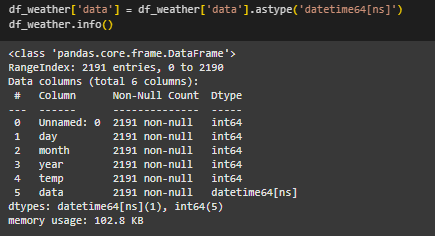


Рисунок Г.2 – Изменение типа данных даты

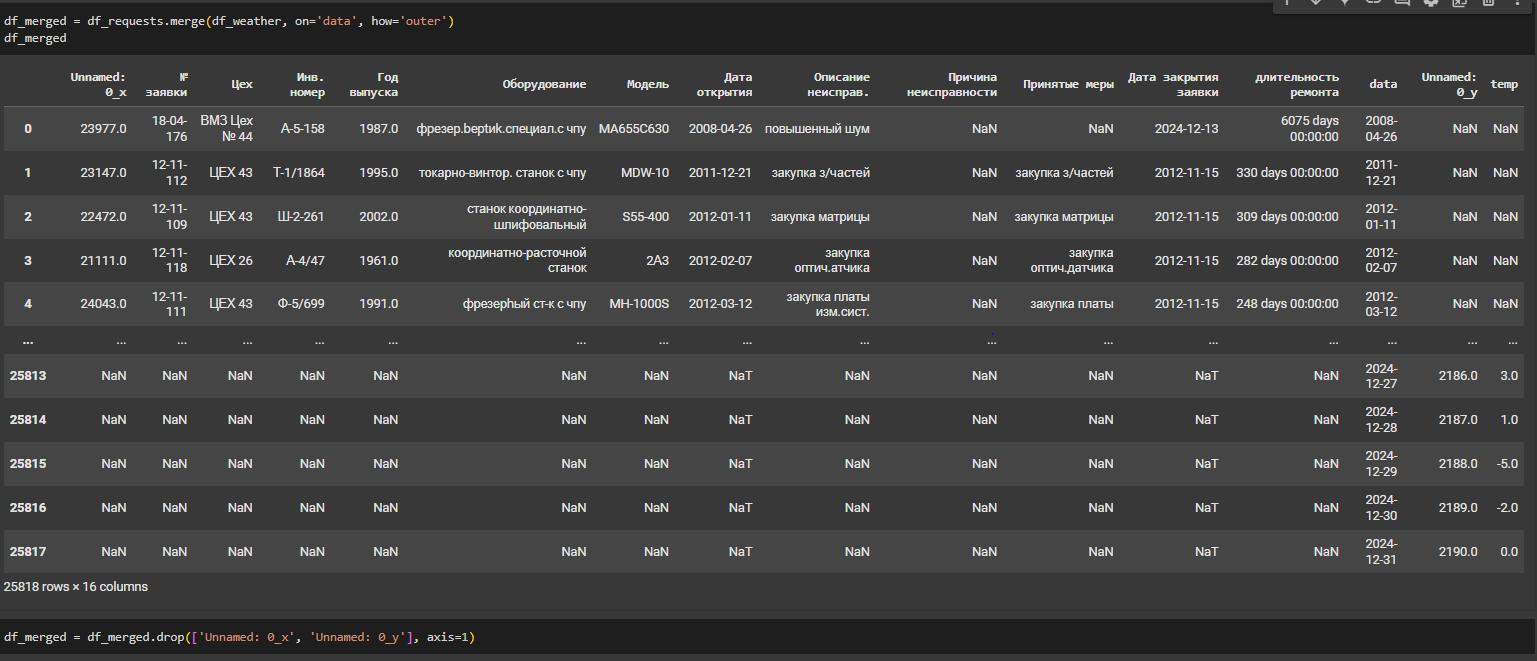


Рисунок Г.3 – Объединение таблиц по дате

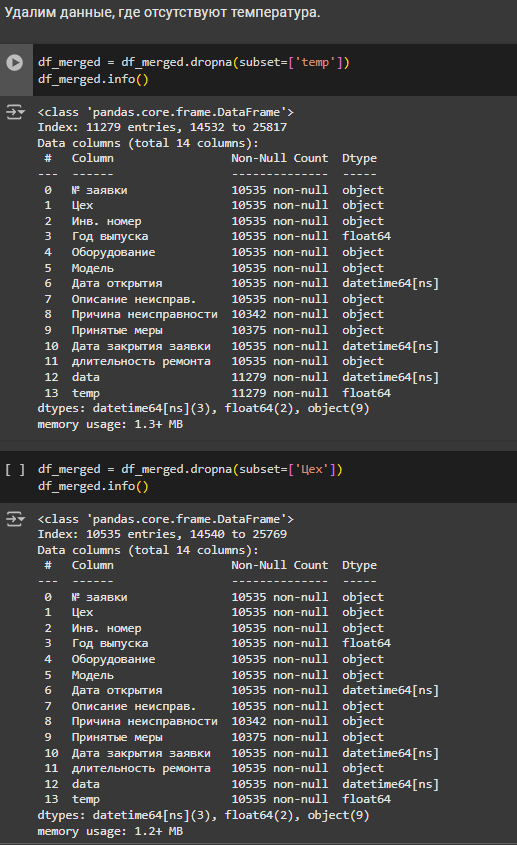
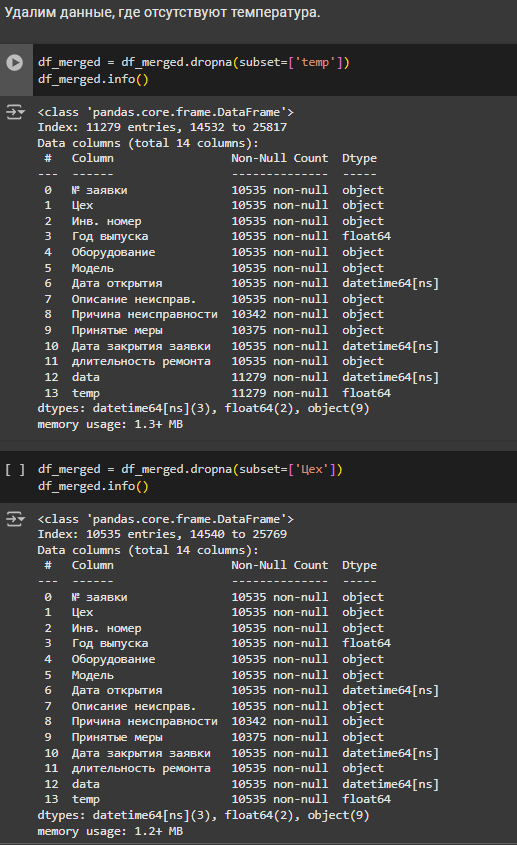


Рисунок Г.4 – Удаление пустых значений

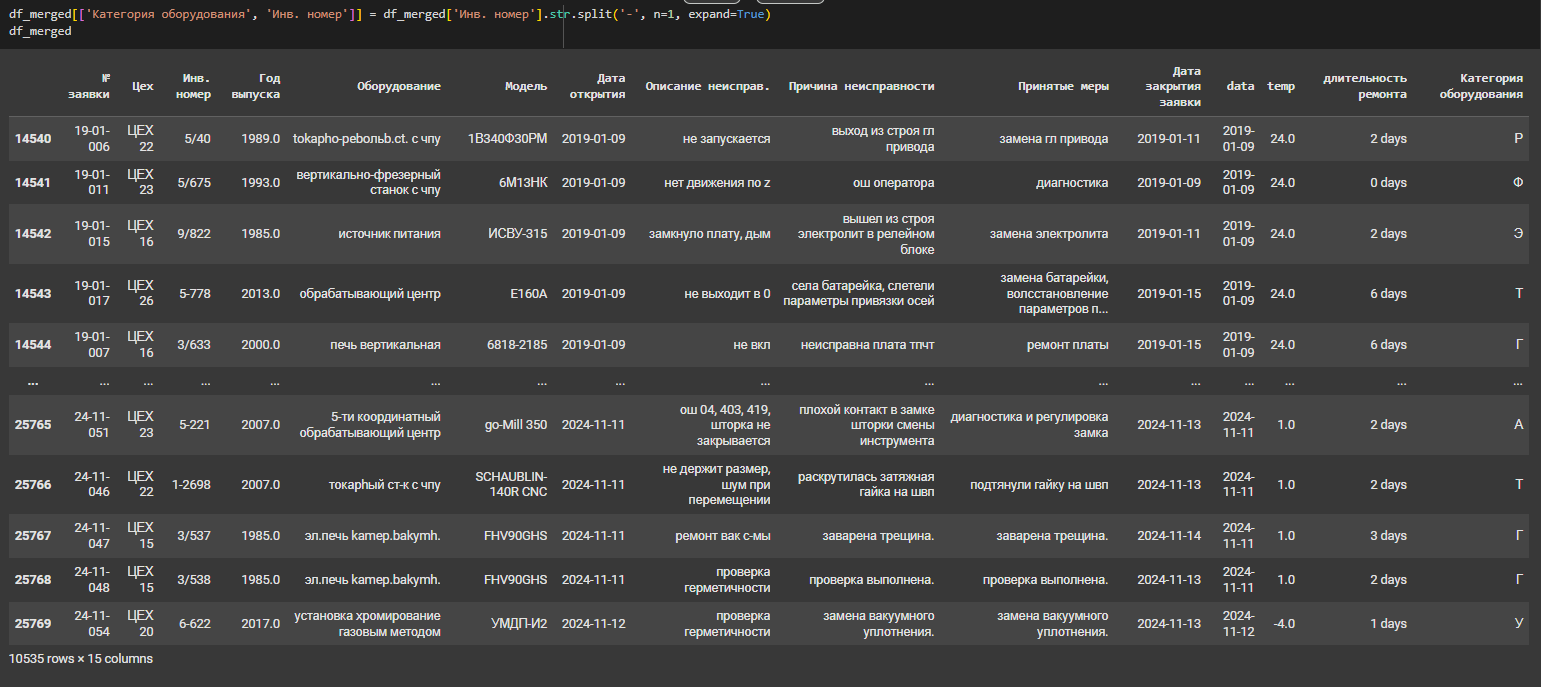


Рисунок Г.5 – Преобразование данных

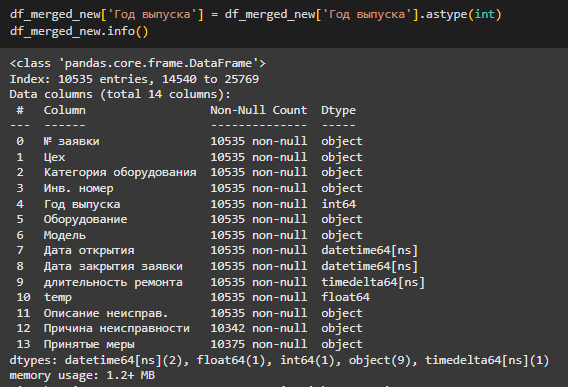


Рисунок Г.7– Преобразование данных

# ПРИЛОЖЕНИЕ Д Объединение данных

Рисунок Д.1 – Преобразование данных в Power Query



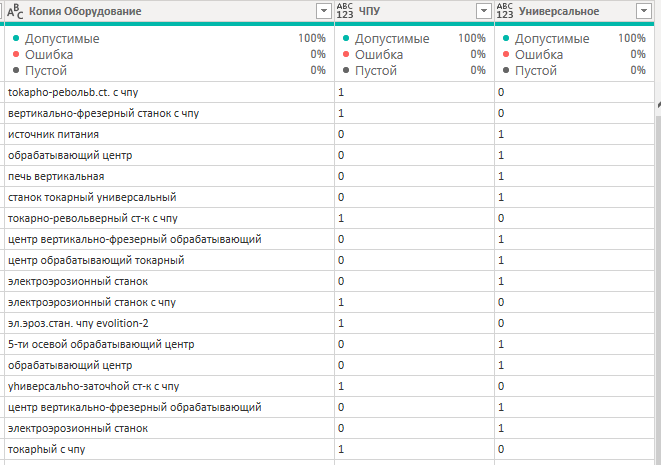


Рисунок Д.2 – Выделение универсального оборудования и оборудования с ЧПУ

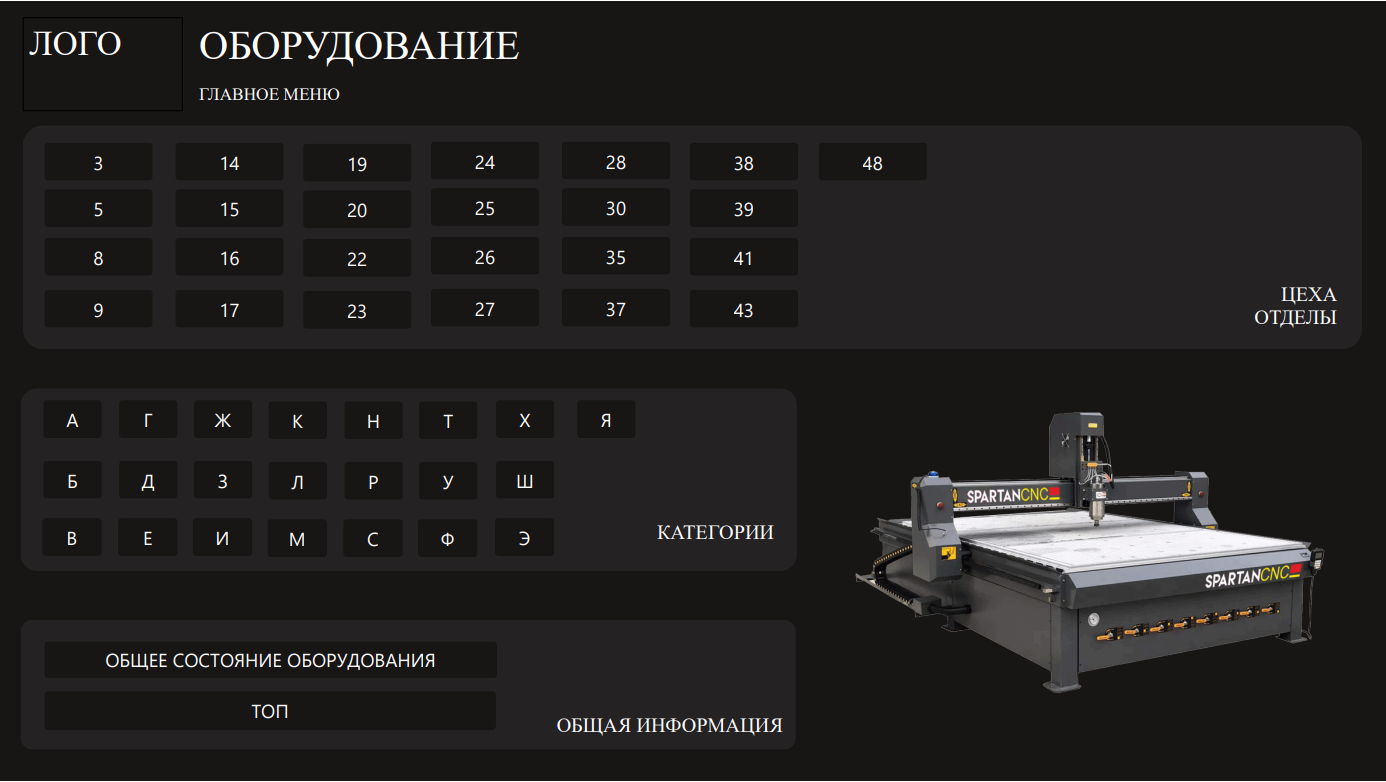


Рисунок Д.3 – Главное меню дашборда



Рисунок Д.4 – Страница «Общее состояние»

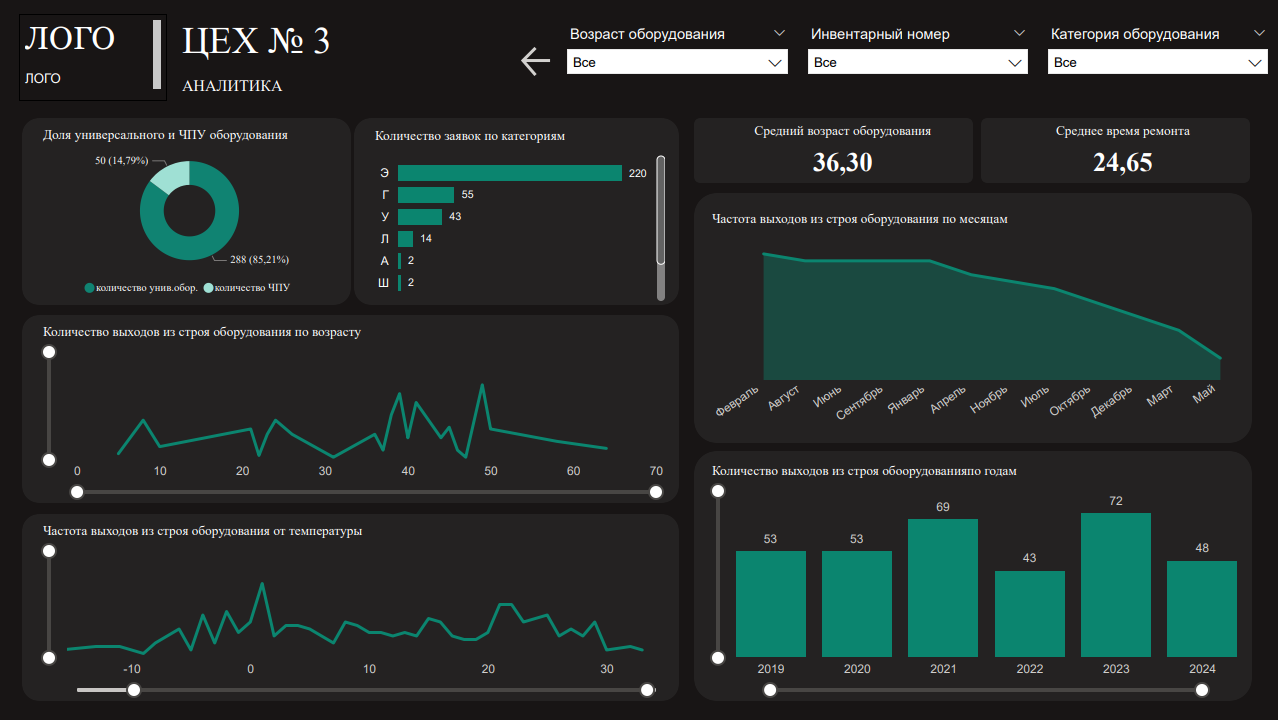


Рисунок Д.5 – Дашборд по отдельным цехам

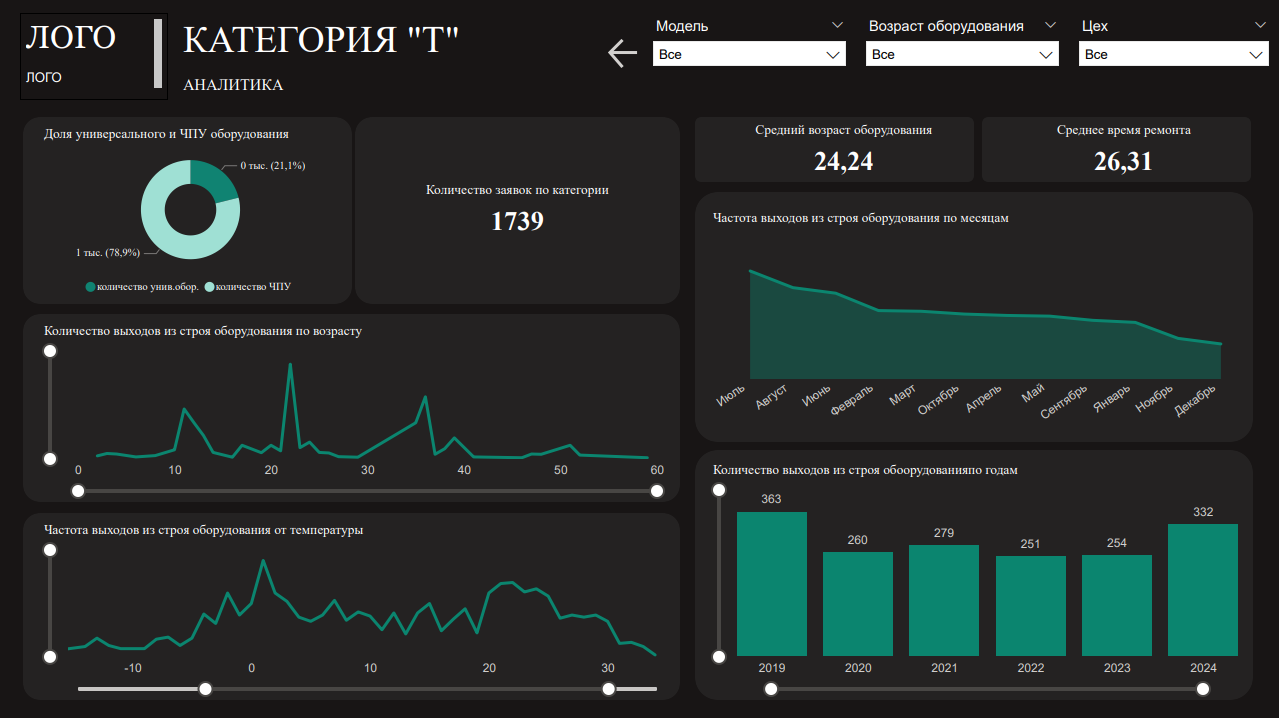


Рисунок Д.6 – Дашборд по отдельным категориям



Рисунок Д.6 – Страница «ТОП по поломкам»

# ПРИЛОЖЕНИЕ Е Модель ARIMA для прогноза поломок

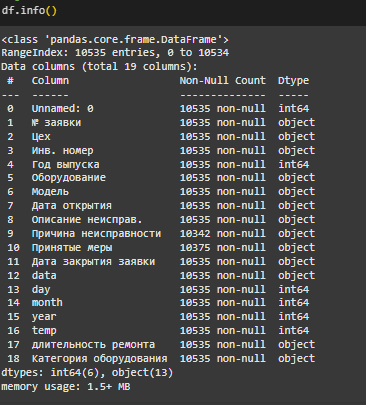
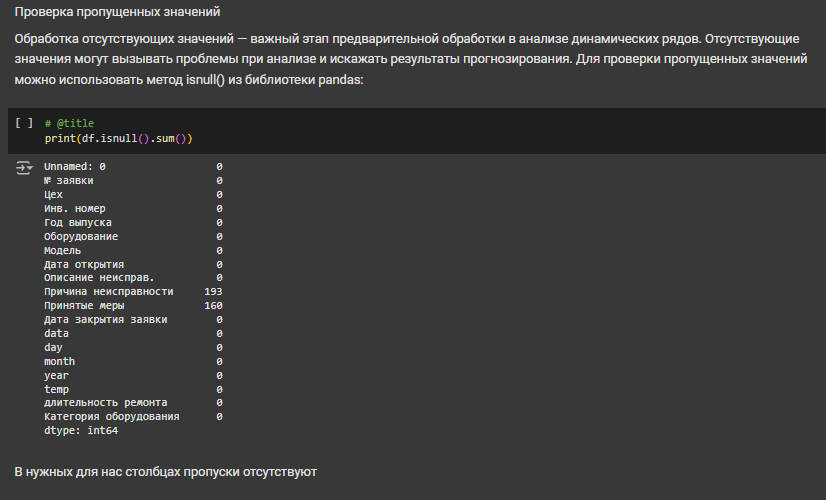


Рисунок Е.1 – Проверка пропущенных значений

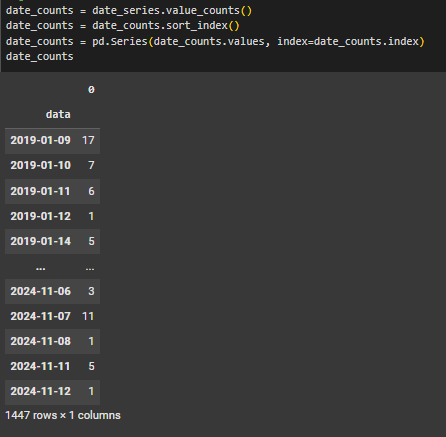
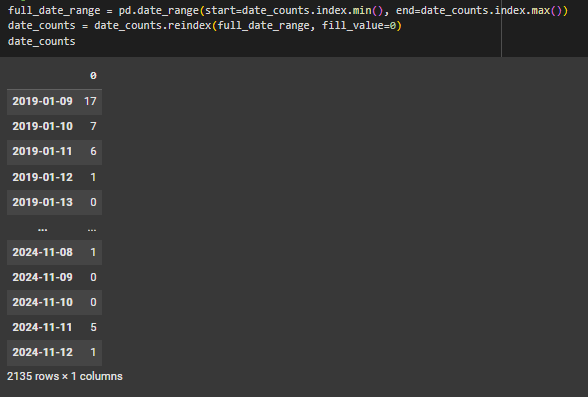


Рисунок Е.2 – Подготовка данных к обучению



Рисунок Е.3 – Зависимость поломок оборудования от температуры

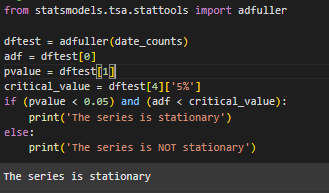
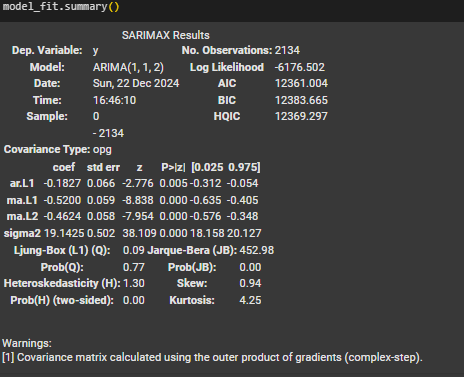


Рисунок Е.4 – Проверка на стационарность



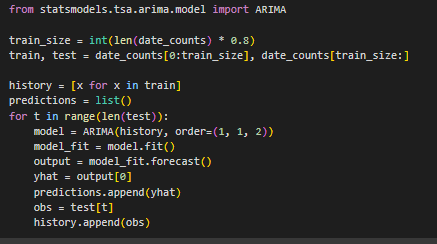


Рисунок Е.5 – Создание и оценка модели

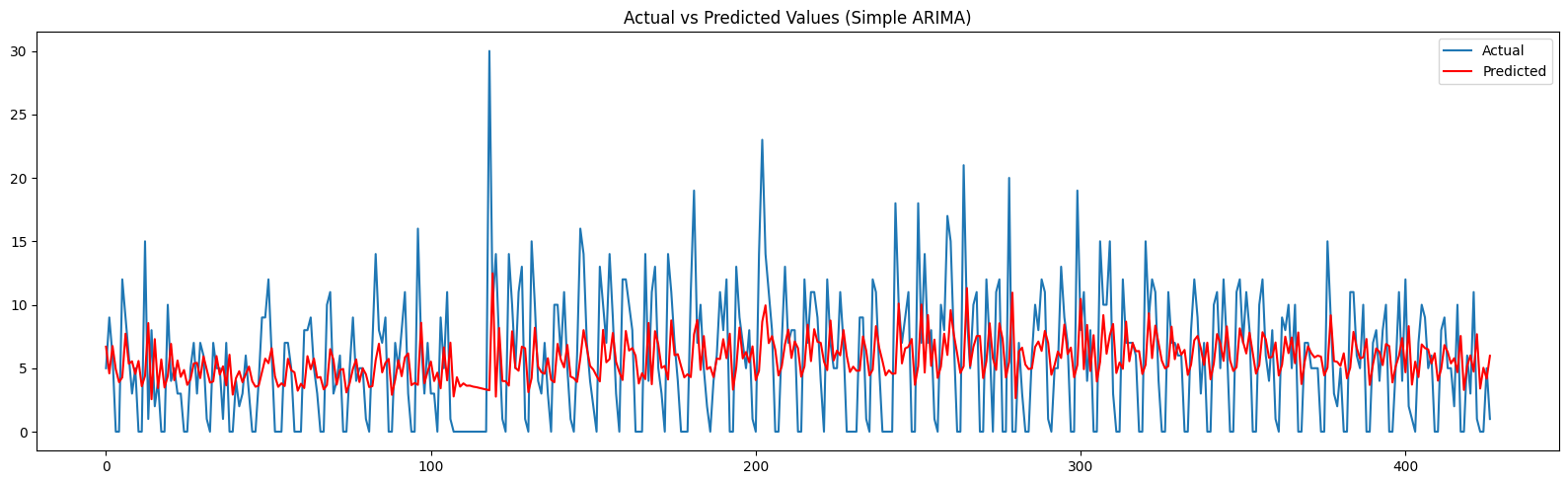


Рисунок Е.6 – График Simple ARIMA

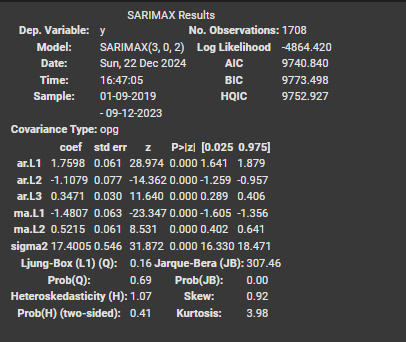


Рисунок Е.7 - Создание и оценка модели

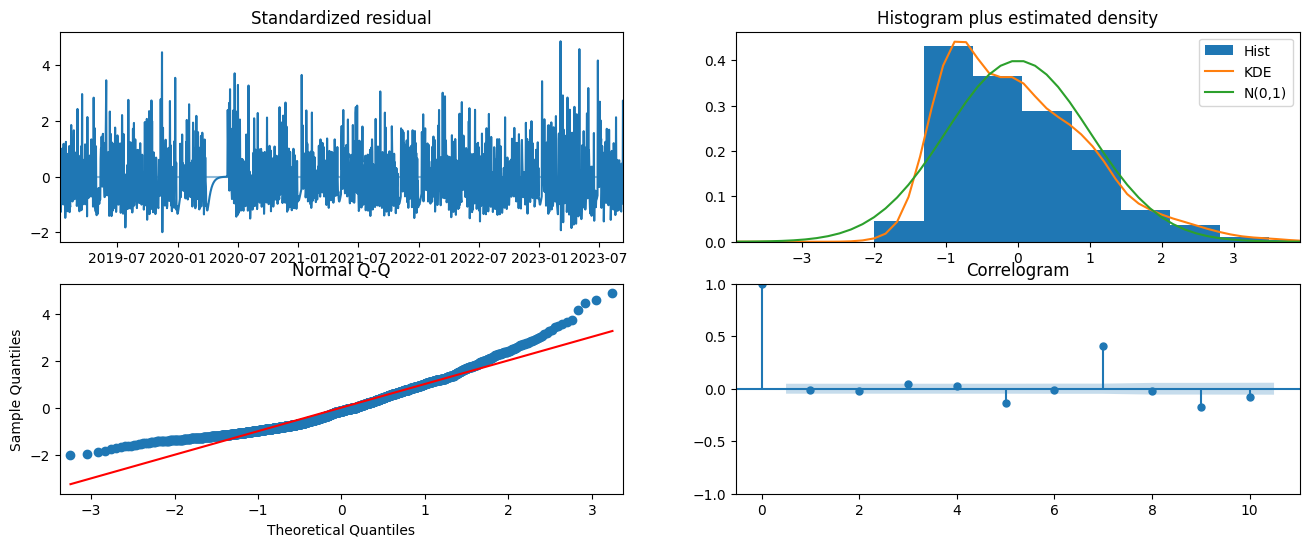


Рисунок Е.8 – График модели