Министерство науки и высшего образования Российской Федерации

 Федеральное государственное автономное образовательное учреждение

**«Национальный исследовательский технологический университет «МИСиС»**

Институт ИКН

Курсовая работа по предмету

«Прикладной статистический анализ»

Тема:

«Разработка модели прогнозирования объемов продажи молока»

Выполнил: Кривцов К. Е. БИВТ-21-5

Руководитель: Маркарян А.О.

Москва 2023

**Оглавление**

[Введение 3](#_Toc153288749)

[Анализ характеристик объекта исследования 5](#_Toc153288750)

[Описание объекта исследования 5](#_Toc153288751)

[Анализ объекта исследования с помощью статистических показателей и выявление причинно-следственных связей 7](#_Toc153288752)

[Постановка задачи моделирования 12](#_Toc153288753)

[Моделирование статистических зависимостей 13](#_Toc153288754)

[Формализация и классификация переменных 13](#_Toc153288755)

[Корреляционный анализ 14](#_Toc153288756)

[Проверка гипотезы о нормальном распределении выходной величины 15](#_Toc153288757)

[Построение регрессионной модели 17](#_Toc153288758)

[Исследование модели 20](#_Toc153288759)

[Анализ статистической значимости уравнения регрессии 20](#_Toc153288760)

[Анализ статистической значимости коэффициентов уравнения регрессии 21](#_Toc153288761)

[Исследование мультиколлинеарности факторов 22](#_Toc153288762)

[Применение шагов регрессионного анализа для улучшения модели 23](#_Toc153288763)

[Реализация и численное исследование результатов моделирования 24](#_Toc153288764)

[Обоснование выбора и описание программного обеспечения 24](#_Toc153288765)

[Описание основных модулей программы 25](#_Toc153288766)

[Численное исследование результатов моделирования 26](#_Toc153288767)

[Выводы 27](#_Toc153288768)

[Приложение 28](#_Toc153288769)

# Введение

В условиях постоянно меняющегося бизнес-окружения и конкурентной борьбы актуальность разработки эффективных стратегий и планов развития для предприятий становится ключевым фактором для их успешной деятельности. Особенно важным аспектом этого процесса является прогнозирование объемов продаж, поскольку это позволяет предприятию адекватно реагировать на рыночные тенденции, изменения спроса и сезонные колебания.

Молочная промышленность, являясь важным сегментом пищевой индустрии, вносит существенный вклад в обеспечение населения необходимыми питательными веществами. С учетом изменяющихся сезонных предпочтений и требований потребителей, эффективное прогнозирование объемов производства молочных продуктов становится критическим элементом стратегического планирования. А с учетом растущего внимания к здоровому образу жизни и правильному питанию, молочные продукты становятся неотъемлемой частью сбалансированного рациона. Богатство белка, кальция и других важных элементов делает молочные продукты не только вкусными, но и важными для поддержания здоровья.

В данном контексте использование математических моделей, включающих линейную регрессию и динамические ряды с учетом сезонности, приобретает особую важность. Такая модель позволяет не только анализировать прошлые тенденции и прогнозировать будущие объемы продаж, но и предоставляет предприятию ценные инструменты для успешной реализации своих стратегий развития.

Ключевым аспектом эффективности модели является ее способность адаптироваться к сезонным колебаниям в спросе на молочные продукты. Это позволяет предприятию точнее прогнозировать не только общие объемы продаж, но и адаптировать производственные планы под сезонные изменения в предпочтениях потребителей. Таким образом, модель не только помогает предприятию выполнить свой текущий план производства, но и оптимизирует его, учитывая динамику рынка.

Использование такой модели становится неотъемлемой частью стратегического управления предприятием, предоставляя руководству необходимые инструменты для более точного планирования и принятия решений. В результате применения этой модели предприятие может не только уверенно справляться с текущими вызовами рынка, но и адаптироваться к будущим изменениям, обеспечивая устойчивость и успешное развитие в долгосрочной перспективе.

Целью нашего исследования является выявление стратегических направлений для оптимизации производства, удовлетворения потребительского спроса и укрепления позиций предприятий в данном секторе, а также, как итог, построение модели для прогнозирования продаж молока.

# Анализ характеристик объекта исследования

## Описание объекта исследования

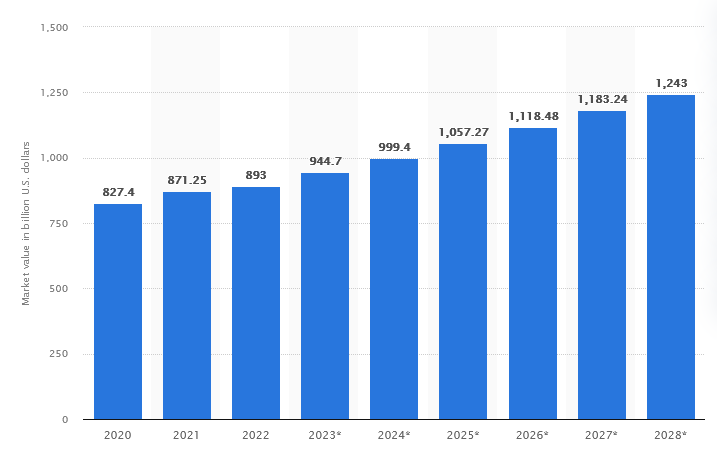
Молочная промышленность представляет собой один из ключевых сегментов в пищевой индустрии, нацеленный на производство и предоставление качественных молочных продуктов для потребителей. Эта отрасль отличается высокой степенью конкуренции, и одним из фундаментальных аспектов её успешного функционирования является эффективное управление производственными процессами, а также представляет собой важный сегмент пищевой отрасли, посвященный производству и предоставлению разнообразных молочных продуктов, от свежего молока до различных видов сыра, йогуртов и других деликатесов. Этот сектор характеризуется широким ассортиментом продукции и является неотъемлемой частью повседневного рациона потребителей.

Инвестиции в молочную промышленность привлекают внимание как внутри страны, так и за её пределами. Согласно последнему отчету аналитической компании в области продовольствия, в 2021 году инвестиции в производство молочных продуктов в России удвоилось с $4.4 млрд. до $8.4 млрд. Мировой рынок молочной продукции также продолжает демонстрировать стабильный рост. Его привлекательность для инвесторов обусловлена стабильным спросом на высококачественные молочные товары и устойчивостью самого рынка.

Молочная промышленность представляет собой важный сегмент пищевой отрасли, посвященный производству и предоставлению разнообразных молочных продуктов, от свежего молока до различных видов сыра, йогуртов и других деликатесов. Этот сектор характеризуется широким ассортиментом продукции и является неотъемлемой частью повседневного рациона потребителей.

Инвесторы в молочной промышленности внимательно следят за тенденциями потребительского спроса и стремятся диверсифицировать свои портфели продуктов. Несмотря на вызовы, представленные экономическими колебаниями, мировой кризис и другими факторами, молочная промышленность продолжает развиваться и приносить прибыль, что подтверждается графиком объема производства молочных продуктов в период с 2013 по 2023 годы.

По прогнозам аналитиков на ближайшие 5 лет в мире также ожидается рост инвестиций в молочную сферу:

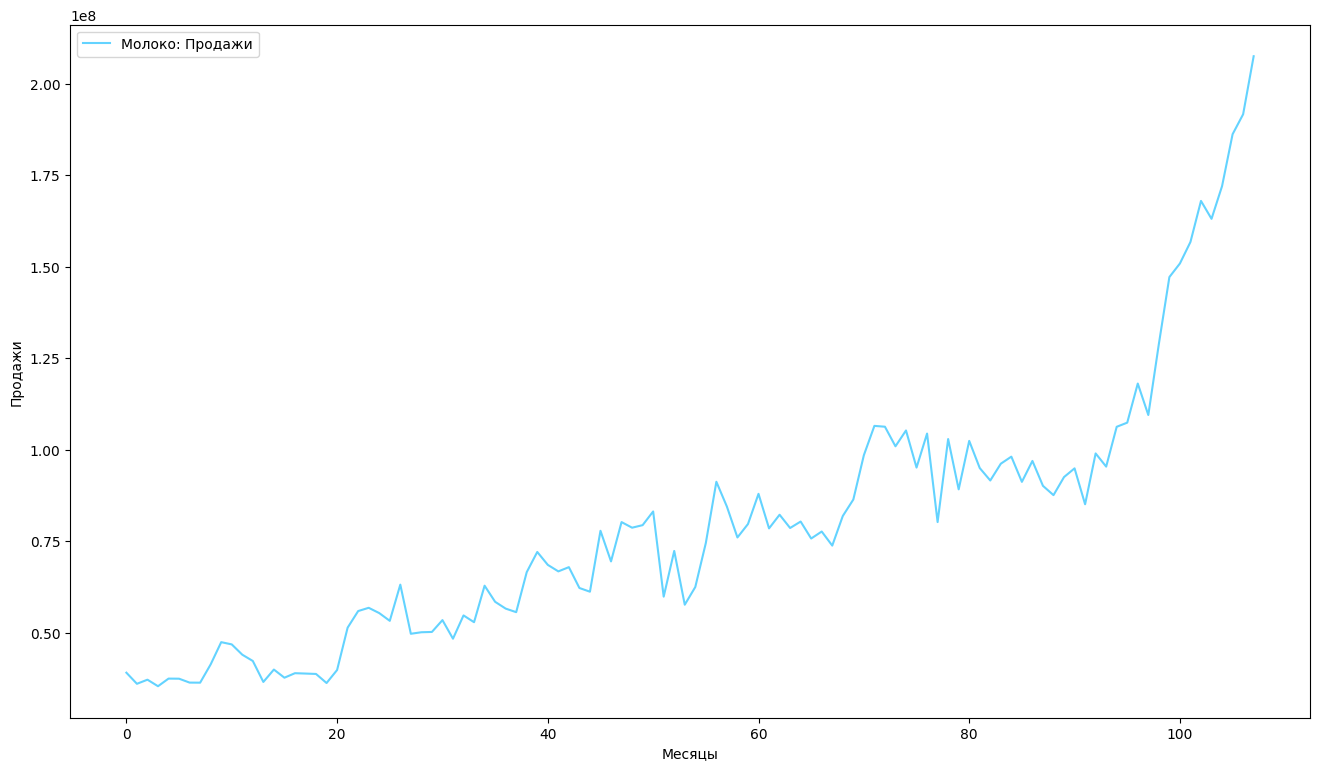


*Рис 1. Прогноз объемов инвестиций в молочную сферу до 2028 года*

Объектом нашего исследования является молочная промышленность с учетом её динамики и актуальных тенденций на мировом рынке. В центре внимания находится процесс производства, адаптация к изменяющимся потребительским предпочтениям и стратегии развития, направленные на обеспечение устойчивого роста и конкурентоспособности.

## Анализ объекта исследования с помощью статистических показателей и выявление причинно-следственных связей

Одним из ключевых инструментов, позволяющих более глубоко понять динамику молочной промышленности, является анализ графика продаж молока по месяцам. Этот подход имеет высокую важность в контексте стратегического управления производственными процессами и предоставляет ценную информацию для принятия решений.



*Рис.2 Продажи молока по месяцам за 2014 -2022 годы*

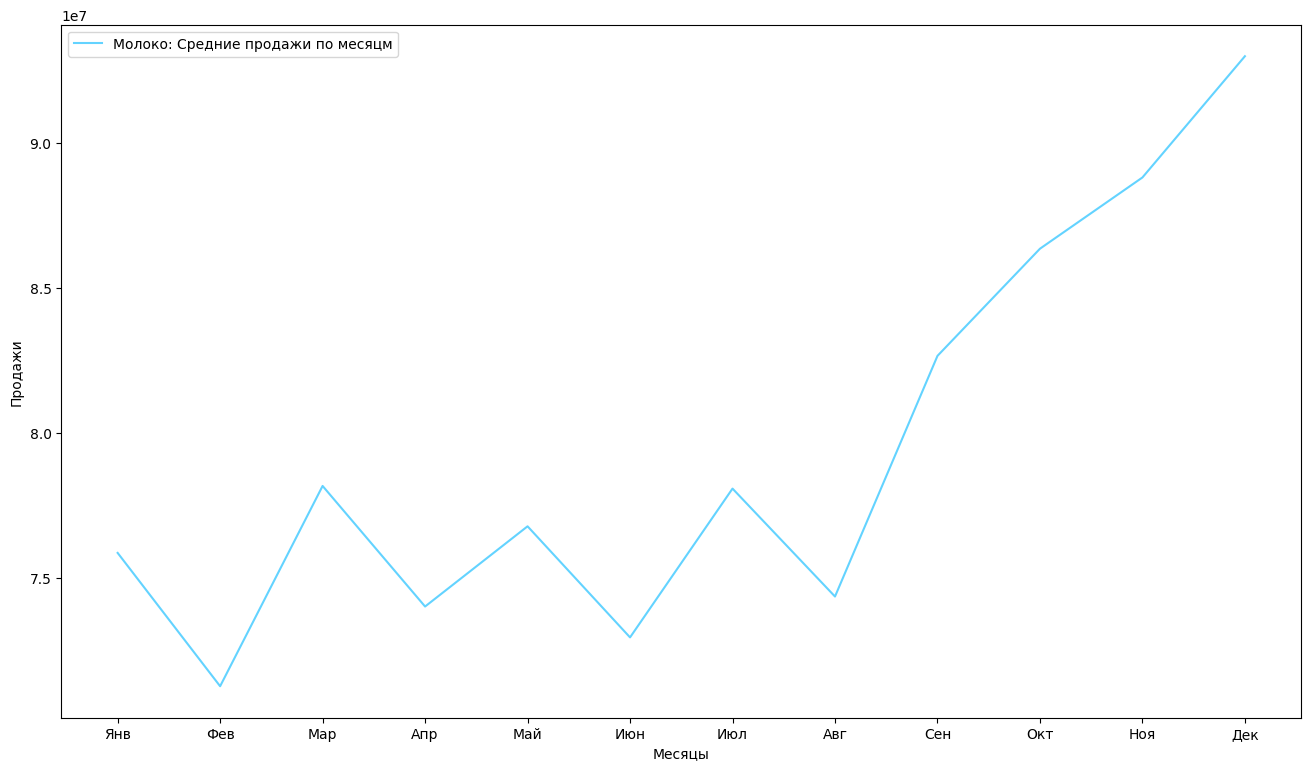
Такой график позволяет выделить тренды в потребительском поведении и понять, какие месяцы являются наиболее активными с точки зрения продаж. Это не только обеспечивает предприятия молочной промышленности информацией для оптимизации производства в соответствии с сезонными колебаниями, но и позволяет адаптировать маркетинговые стратегии для максимизации доходов в периоды повышенного спроса.

Важность рассмотрения графика усредненных месячных продаж молока выходит далеко за пределы простого отслеживания количества проданных товаров. Этот анализ становится ключевым элементом стратегического управления в контексте молочной промышленности, где сезонные колебания имеют прямое влияние на общий успех предприятия.

В молочной промышленности существует четкая зависимость между сезонами и уровнем потребительского спроса на продукцию. Усредненные месячные продажи молока позволяют выявить эти колебания и выделить тенденции, определяющие периоды пика и спада периоды в годовом цикле. Например, в летние месяцы возможно увеличение потребления молочных продуктов, обусловленное повышенным спросом на охлаждающие напитки, мороженое и другие прохладительные угощения.

Разбиение года на сезоны также позволяет выявить взаимосвязи с календарными событиями, праздниками или изменениями в потребительском поведении. Например, праздничный сезон может спровоцировать увеличение спроса на молочные продукты, так как они часто входят в состав праздничных блюд и угощений.

Такой детальный анализ усредненных месячных продаж молока предоставляет предприятию инсайты для более точного прогнозирования спроса, оптимизации производства и планирования ресурсов. Это становится стратегическим инструментом, позволяющим адаптироваться к сезонным колебаниям, минимизировать риски избыточного запаса или дефицита продукции, и, таким образом, обеспечивать устойчивое и успешное функционирование в динамичной молочной индустрии.



*Рис.3 Усредненные продажи молока по месяцам за 2014 -2022 годы*

В процессе анализа усредненных месячных продаж молока выявлены интересные закономерности.

В феврале наблюдается небольшая просадка, что может быть обусловлено зимним периодом и возможным снижением потребительского интереса к прохладным молочным напиткам.

В марте, в свою очередь, продажи растут, возможно, в связи с весенним сезоном и стремлением потребителей к более активному образу жизни после зимних месяцев.

Апрель характеризуется вторичной просадкой, за исключением 2017 года, что может свидетельствовать о специфических факторах этого года или изменениях в потребительском поведении.

Июнь, как правило, означает дополнительное снижение продаж, вероятно, связанное с началом летнего сезона, когда предпочтения потребителей могут смещаться в сторону легких и охлаждающих продуктов. С июля, самого жаркого месяца, возможно, начинается восстановление спроса, особенно на мороженое и другие освежающие молочные десерты.

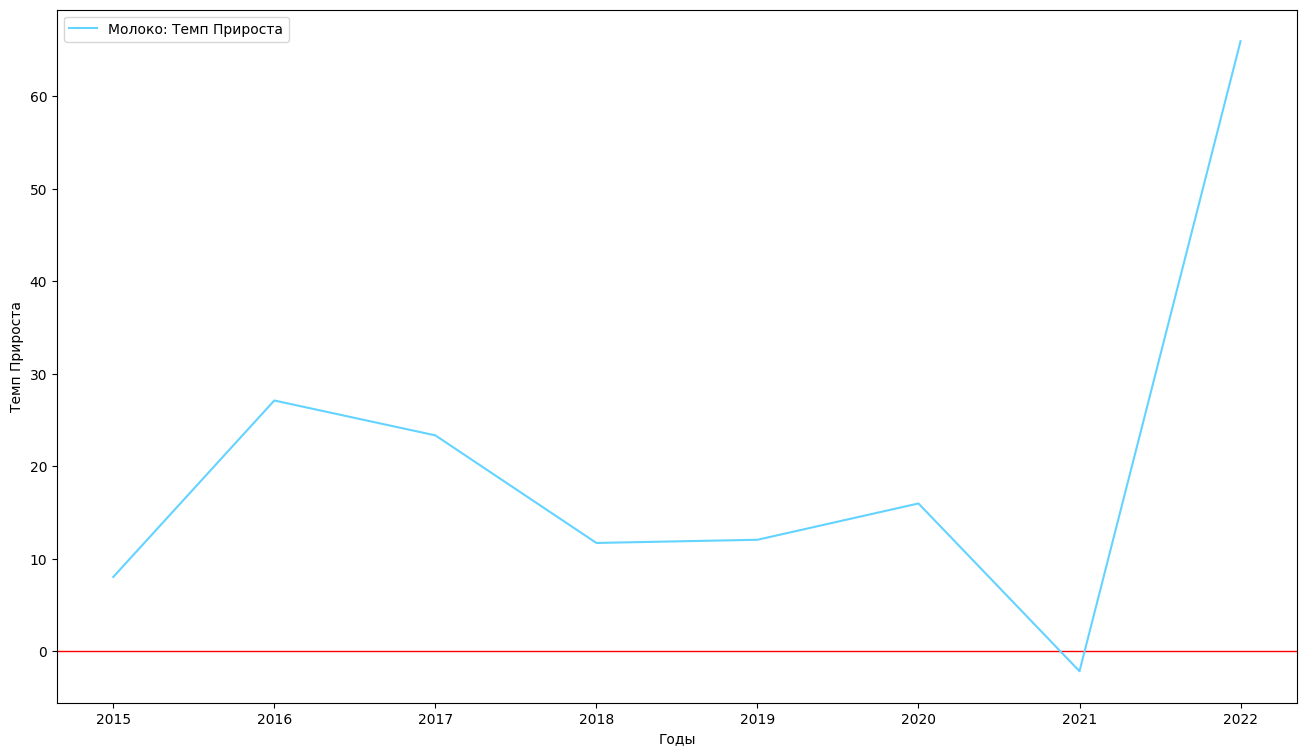
К концу года отмечается общий рост продаж молока. Это может быть связано с тем, что в холодные месяцы потребители склонны больше времени проводить дома, что влияет на увеличение потребления молочных продуктов.

Не менее важным является анализ темпа роста и темпа прироста. Единственное различие между этими показателями заключается в способе их измерения. Темп роста представляет собой процентное изменение объема продаж молока относительно предыдущего периода. Темп прироста, с другой стороны, определяется как темп роста минус 100%, что выражает изменение относительно базового уровня.

Анализ темпа роста предоставляет информацию о том, как быстро изменяется объем продаж молока в процентном соотношении к предыдущему периоду. Этот показатель помогает выделить периоды активного роста и выявить возможные точки перегиба в динамике продаж. Высокие значения темпа роста могут указывать на успешные маркетинговые стратегии, внедрение новых продуктов или повышенный спрос в определенные периоды.

Темп прироста, с другой стороны, предоставляет контекст к тому, насколько текущий уровень продаж отличается от базового уровня, выраженного в процентах. Этот показатель особенно полезен для выявления общего тренда роста или спада, а также для определения, насколько текущие продажи превосходят или отстают от среднего уровня.

Таким образом, тщательный мониторинг темпа роста и темпа прироста в сочетании с общим анализом усредненных месячных продаж создает надежную основу для принятия обоснованных решений в сфере управления продажами и разработки стратегии молочного предприятия. Эти инструменты анализа помогают не только реагировать на текущие изменения рынка, но и предвидеть потенциальные тенденции, что является ключевым фактором для достижения успешных результатов в динамичной и конкурентной молочной промышленности.



*Рис.4 Темп прироста за 2014 - 2022 годы*

Анализ графика темпа прироста продаж молока предоставляет интересную динамику в различные периоды и подчеркивает ключевые события, влияющие на отрасль.

В целом, стоит отметить, что отрицательный рост, отмеченный в 2020 году, является следствием глобальной пандемии COVID-19. Этот период характеризуется изменением потребительского поведения, экономическими трудностями и непрогнозируемыми факторами.

В то время как большинство лет отмечаются положительным ростом, выделяются два периода — 2014-2015 годы и 2021-2022 годы. В первый период можно увидеть довольно сильный рост, что может свидетельствовать о благоприятной ситуации на рынке или успешных стратегиях предприятия.

Во второй период, 2021-2022 года, отмечается очень сильный рост. Это может говорить о успешных стратегий или о способности предприятия эффективно адаптироваться к переменам в молочной промышленности.

В остальные годы также отмечается положительный рост относительно предыдущего года, но он не достигает таких высоких значений, как в 2014-2015 и 2021-2022 годах. Это может указывать на более стабильные, но менее драматичные изменения в стратегии продаж, а также на общую тенденцию к росту в отрасли.

## Постановка задачи моделирования

Для решения задачи прогнозирования продаж молока предприятием в следующем году поставлена следующая задача моделирования:

Цель: разработать модель машинного обучения, которая будет способна прогнозировать продажи молока предприятием в следующем году по месяцам.

Задачи:

Подготовка данных: сначала необходимо провести предварительную обработку данных, включая заполнение пропущенных значений, кодирование категориальных признаков, исключение выбросов и масштабирование числовых признаков.

Выбор модели: далее нужно выбрать подходящую модель машинного обучения для решения задачи классификации, где целевой переменной будет массив из 12 чисел – продаж за каждый месяц. Возможными моделями могут быть аддитивная или мультипликативная модель динамического ряда в связке с линейной, полиномиальной или степенной регрессией.

Обучение модели: на этом этапе мы будем обучать выбранную модель на обучающем наборе данных. Модель будет учиться на данных о продажах за 2014 – 2022 года.

Оценка модели: после обучения модели необходимо оценить ее производительность с использованием тестового набора данных за 2023 год. Коэффициент детерминации , ковариация и корреляция будут использоваться для оценки качества модели.

Задача моделирования направлена на помощь руководству предприятия в ретроспективном анализе принятых решений и в оценке выполнения плана дальнейшего развития предприятием.

# Моделирование статистических зависимостей

## Формализация и классификация переменных

Формализация и классификация переменных в задаче прогнозирования продаж молока предприятием может быть следующей:

Выходная величина:

Y: массив из 12 значений – годовые продажи по месяцам.

Математически статистическая зависимость может быть описана в виде:

Y = f (P, S, E)

Где

Y – уровни ряда

P – прогнозируемые степенной функцией значения

S – сезонные и циклические колебания

E – случайные колебания

Виды моделей в зависимости от связи компонент:

1. Аддитивная

Значения сезонной компоненты предполагаются постоянными для различных циклов: амплитуда колебаний приблизительно постоянна

Y = P + S + Е

1. Мультипликативная

Зависимость уровней ряда от значений сезонной компоненты: амплитуда колебаний возрастает или уменьшается

Y = P \* S \* Е

Для решения поставленной задачи выбирается аддитивная модель.

## Корреляционный анализ

Ковариация в размерных величинах равна что усложняет ее интерпретацию, особенно при измерениях переменных в разных единицах. В связи с этим оценка силы связи между переменными становится затруднительной.

Однако коэффициент корреляции Пирсона, равный является безразмерной величиной и лежит в диапазоне от -1 до 1. Это значение указывает на сильную положительную линейную связь между переменными. Такой коэффициент обычно считается высоким и говорит о том, что при увеличении значения одной переменной вероятнее всего увеличится и значение другой переменной, и наоборот. Таким образом, коэффициент корреляции Пирсона подходит для измерения линейных связей между переменными в данном контексте и является удобным инструментом для определения силы и направления линейной зависимости.

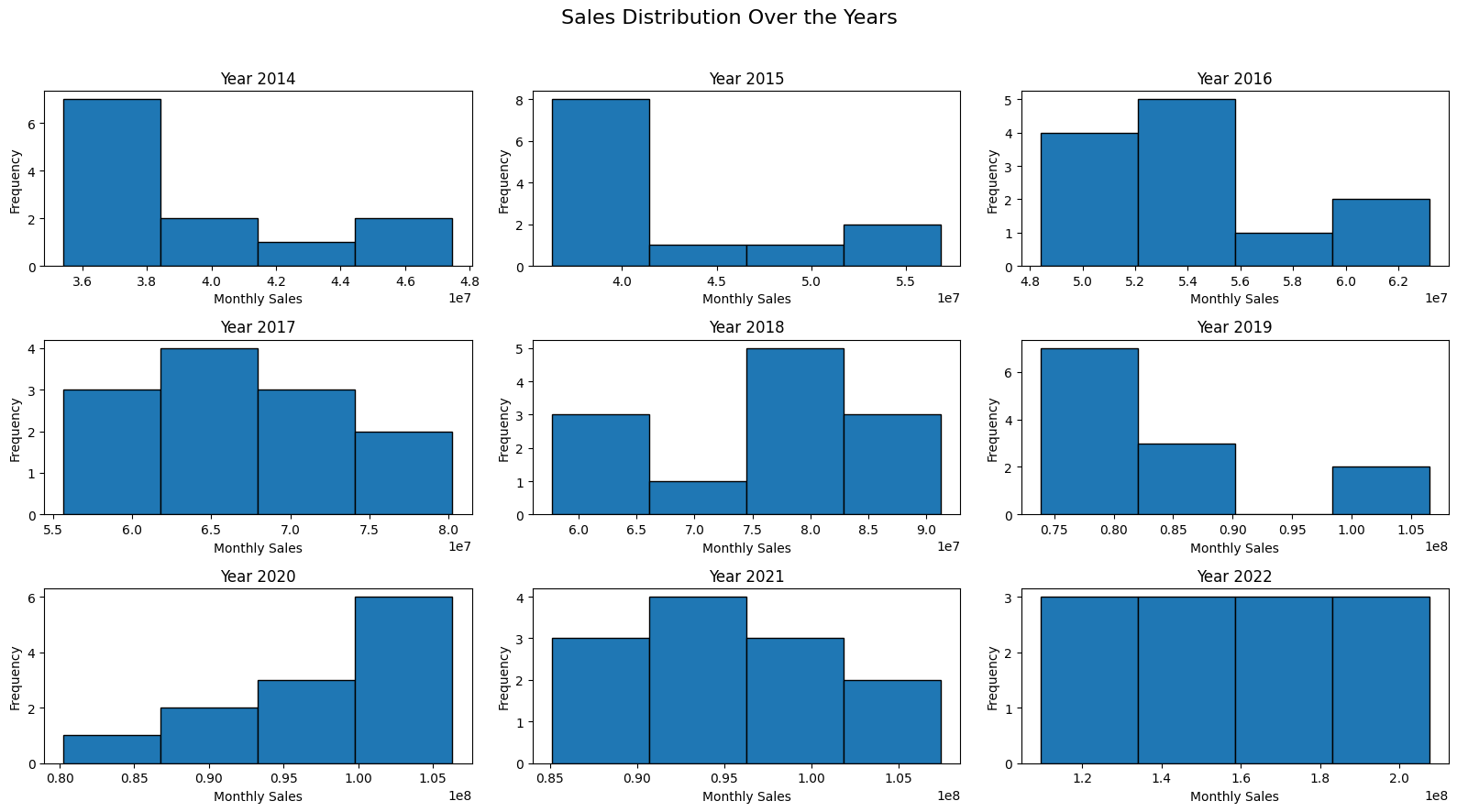
Значение Коэффициента Детерминации указывает на то, что объясненная дисперсия величины ответа моделью составляет более 97%, что подчеркивает высокую точность модели в объяснении изменчивости зависимой переменной на основе независимой. Такие высокие значения коэффициента корреляции и R^2 говорят о том, что линейная модель хорошо соответствует данным и может успешно прогнозировать значения зависимой переменной.

## Проверка гипотезы о нормальном распределении выходной величины

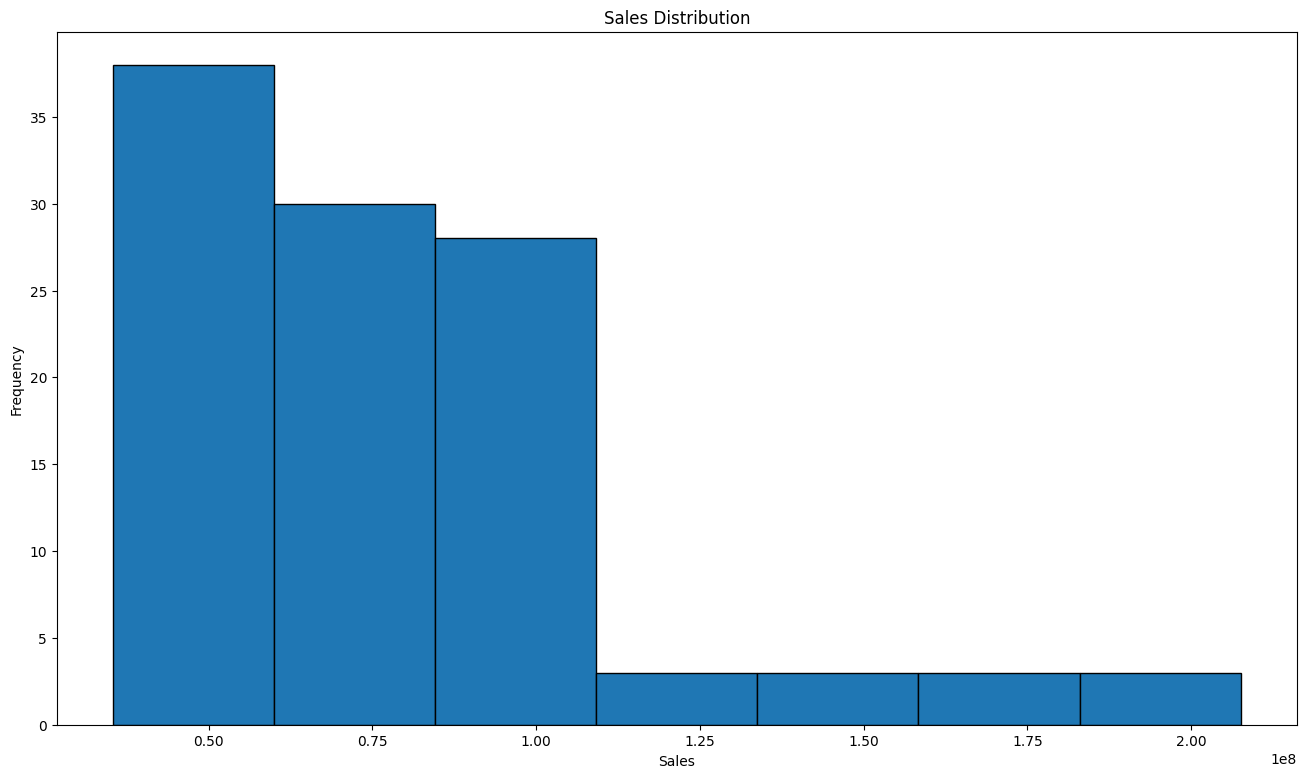
Проверка гипотезы о нормальном распределении выходной величины является важным этапом в анализе данных, так как позволяет оценить, насколько хорошо данные соответствуют теоретическому нормальному распределению. В случае продаж молока, такая проверка может предоставить ценную информацию о структуре данных и их распределении.

Нормальное распределение, или распределение Гаусса, является одним из ключевых понятий в статистике. Оно характеризуется симметричной колоколообразной формой и хорошо подходит для описания многих случайных процессов. В контексте продаж молока нормальное распределение может указывать на стабильность и предсказуемость спроса.

Для проверки нормальности распределения продаж молока, мы можем воспользоваться графическим методом, например, гистограммой.



*Рис.5 Распределение продаж за 2014 - 2022 годы по годам*



*Рис.6 Распределение продаж за 2014 - 2022 годы*

Однако, при проведении такой проверки на наших данных о продажах молока, мы наблюдаем, что распределение выходной величины не соответствует нормальному распределению. Вместо того, чтобы иметь колоколообразную форму, мы можем обнаружить отклонения, которые могут быть обусловлены сезонностью, внешними факторами или другими специфическими особенностями в данных о продажах молока. В таких случаях, более сложные методы статистического анализа могут быть применены для более точного моделирования и предсказания динамики продаж.

## Построение регрессионной модели

При разработке регрессионной модели для прогнозирования месячных продаж молока был проведен внимательный анализ доступных данных. График, отображающий динамику продаж, обнаруживает форму, которая в значительной степени соответствует степенной функции. Это решение подтверждается сравнительным анализом с полиномиальной функцией, где степенная функция продемонстрировала более высокую степень соответствия реальным данным.

Выбор степенной функции в качестве основы для регрессионной модели обусловлен не только статистическими критериями, но и логикой, которую предоставляет форма графика. Степенная функция, с ее способностью отражать экспоненциальные изменения, представляется более адекватной для моделирования роста или убывания месячных продаж в сравнении с полиномиальной аппроксимацией.

Дополнительно, наша модель будет дополнена сезонной компонентой. Это важный шаг, так как месячные продажи могут подвергаться влиянию сезонных факторов, таких как праздничные периоды, времена года и другие факторы, влияющие на потребительский спрос. Интеграция сезонной компоненты позволит учесть эти периодические колебания и, таким образом, сделать нашу модель более точной и адаптированной к реальным условиям. Степенная функция, дополненная сезонной коррекцией, создаст основу для более эффективного прогнозирования, обеспечивая более точные результаты для управленческих решений.

Процесс определения сезонных колебаний включает следующие шаги:

1. Сглаживание исходного ряда методом скользящей средней.

На первом этапе применяется метод скользящей средней для сглаживания исходного временного ряда. Этот метод позволяет убрать шумы и выделить основной тренд в данных, что упрощает дальнейший анализ.

Обозначения:

h – количество периодов

n – количество сезонов в одном периоде

m – число уровней ряда по которым вычисляется скользящая средняя

l - индекс первого скользящего среднего

k – индекс последнего скользящего среднего

Скользящее среднее вычисляется по формуле:

i = 1, 2 … n – m + 1

При выравнивании теряется несколько членов исходного ряда с обоих концов. Их можно восстановить путем экстраполяции сглаженного ряда:

1. Определение сезонной компоненты.

Следующим шагом является выделение сезонной компоненты. Мы ищем периодические колебания, которые могут быть связаны с сезонными факторами, такими как месяцы или времена года. Это может быть достигнуто путем вычитания от сглаженных данных из исходного ряда.

Находятся оценки сезонной компоненты как разность между

фактическими уровнями ряда и скользящими средними:

Определяются средние значения сезонных компонент за весь период:

Эта усредненная сезонная компонента будет использоваться для детализации степенной функции.

Для предсказания продаж конкретного месяца будет использоваться функция:

На математическом языке задача может ставится так:

для .

D представляет собой множество месяцев прогнозируемого года

C представляем продажи по месяцам с учетом сезонности в прогнозируемый год.

# Исследование модели

## Анализ статистической значимости уравнения регрессии

Критерий F используется для проверки статистической значимости регрессионной модели в целом. Если значение F-критерия значительно, это указывает на то, что объясненная изменчивость моделью является статистически значимой.

Формула для расчета F-критерия в контексте анализа регрессии:

SSR - объясненная сумма квадратов (сумма квадратов регрессии)

SSE - остаточная сумма квадратов (сумма квадратов ошибок)

n – количество наблюдений

m – количество факторов (предикторов) модели

Кроме того, значение P-значения ассоциировано с F-критерием и представляет собой вероятность получения такого или более экстремального значения F при условии, что нулевая гипотеза верна. Низкое P-значение (обычно меньше выбранного уровня значимости, например, 0.05) говорит о том, что можно отклонить нулевую гипотезу и считать модель статистически значимой.

Полученное F-критерия равно 32.72, что сравнивается с критическими значениями для уровней значимости 10% и 5%. Высокое значение F-критерия и низкое P-значение 0.0002868 указывают на статистическую значимость модели. Это означает, что модель в целом имеет существенное воздействие на зависимую переменную, и регрессионные коэффициенты не равны нулю.

## Анализ статистической значимости коэффициентов уравнения регрессии

# Реализация и численное исследование результатов моделирования

## Обоснование выбора и описание программного обеспечения

В данной секции рассмотрено обоснование выбора программного обеспечения и библиотек для проведения анализа данных, построения и обучения моделей машинного обучения.

**Python**: язык программирования Python был выбран как основной язык для анализа данных и машинного обучения. Это обосновывается его популярностью, богатой экосистемой библиотек и фреймворков, таких как Pandas, Scikit-learn, Seaborn, и других.

**Pandas**: для работы с данными. Pandas предоставляет удобные средства для загрузки, очистки, преобразования и анализа данных.

**Seaborn и Matplotlib**: для визуализации данных. Эти библиотеки позволяют строить графики и визуализации, что помогает в понимании данных.

**SciPy**: для проведения статистических тестов. SciPy содержит множество статистических функций и тестов, которые используются для проверки гипотез и анализа данных.

**Scikit-learn**: для построения и обучения моделей машинного обучения. Scikit-learn предоставляет реализации различных алгоритмов машинного обучения и инструменты для оценки их производительности.

**Другие библиотеки**: В работе также использовались библиотеки для специфических задач, такие как NumPy для работы с массивами данных, и другие.

Выбор Python и указанных библиотек обоснован их удобством, мощностью и популярностью в области анализа данных и машинного обучения. Это позволило провести исследование данных, построить модели и провести анализ результатов с высокой эффективностью и удобством.

## Описание основных модулей программы

В данной секции представлено описание основных модулей программы, используемых для анализа данных, построения моделей машинного обучения и проведения оценки результатов.

1. Модуль загрузки данных:

Основная задача этого модуля - загрузить данные из источника и подготовить их для дальнейшего анализа. В нем выполняется чтение данных, удаление или заполнение отсутствующих.

1. Модуль визуализации данных:

В данном модуле проводится визуализация данных с использованием библиотек Seaborn и Matplotlib. Создаются различные типы графиков и визуальных представлений для более глубокого понимания данных.

1. Модуль статистического анализа:

Этот модуль включает в себя статистический анализ данных с использованием библиотеки SciPy. Здесь проводятся различные статистические тесты, такие как тесты на нормальность распределения, а также анализ корреляции между признаками.

1. Модуль построения моделей машинного обучения:

Основная часть программы, где создаются и обучаются модели машинного обучения. В данном модуле используются библиотеки Scikit-learn для выбора, настройки и обучения различных алгоритмов классификации и регрессии.

1. Модуль оценки результатов:

В этом модуле проводится оценка качества моделей на основе различных метрик, например и коэффициента корреляции

## Численное исследование результатов моделирования

В данной секции представлено численное исследование результатов моделирования, проведенное в рамках данной работы.

Для анализа качества построенных моделей машинного обучения были использованы различные метрики, такие как , выборочный коэффициент корреляции (Пирсона).

Ковариация

Коэффициент Пирсона

*Прогноз сюда же*

# Выводы

Список литературы и использованных источников:

* <https://asia.nikkei.com/Spotlight/DealStreetAsia/Southeast-Asia-s-dairy-industry-attracts-big-private-equity-deals> статья про молочные инвестиции

# Приложение

* <https://github.com/AltairkaGit/MilkStats> полный код курсовой работы