МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский технологический университет «МИСиС»»

ИКН кафедра АСУ

КУРСОВАЯ РАБОТА

по дисциплине

«Прикладной статистический анализ»

на тему

«Разработка модели прогнозирования количества покупателей в торговом центре»

Выполнил:
студент 3-го курса, гр. БИВТ-21-4
Савенко Е.И.

Научный руководитель:
К.т.н., доцент, ученый
секретарь кафедры ИКТ
Маркарян А О.

Москва 2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4

1. АНАЛИЗ ХАРАКТЕРИСТИК ОБЪЕКТА ИССЛЕДОВАНИЯ 5

1.1 Описание объекта исследования 5

1.2 Анализ объекта исследования с помощью статистических показателей 5

1.3 Выявление причинно-следственных связей 6

1.4 Постановка задачи моделирования 7

2. МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ 9

2.1 Формализация и классификация переменных 9

2.2 Проверка гипотезы о нормальном распределении выходной величины 9

2.3 Корреляционный анализ 10

2.4 Построение регрессионной модели 11

2.4.1 Структурная идентификация модели 11

2.4.2 Параметрическая идентификация модели 11

В соответствии с методом наименьших квадратов, задача заключается в аппроксимации кривой известной функцией. Вычисление параметров уравнения множественной линейной регрессии будет произведено с помощью алгоритма МНК. 11

3. ИССЛЕДОВАНИЕ МОДЕЛИ 12

3.1 Анализ статистической значимости уравнения регрессии 12

3.3 Исследование мультиколлинеарности факторов 13

3.4 Применение шагового регрессионного анализа для улучшения модели 14

4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ 16

4.1 Обоснование выбора и описание программного обеспечения 16

4.2 Описание основных модулей программы 16

4.3 Численное исследование результатов моделирования 22

ВЫВОДЫ 24

СПИСОК ЛИТЕРАТУРЫ 25

ПРОГРАММНЫЕ ПРИЛОЖЕНИЯ 26
ПРИЛОЖЕНИЕ А 27
ПРИЛОЖЕНИЕ Б 28
ПРИЛОЖЕНИЕ В 29

ВВЕДЕНИЕ

Индустрия розничной торговли претерпевает масштабные изменения, связанные с технологическими инновациями, изменениями в поведении потребителей и колебаниями рыночных тенденций. В условиях такой динамики эффективное управление торговыми центрами требует разработки точных методов прогнозирования. Одним из ключевых вопросов является прогнозирование количества посетителей в торговых центрах. В данной курсовой работе будет разработана модель для прогнозирования потока покупателей.

Целью работы является создание модели, позволяющей прогнозировать количество посетителей торгового центра, опираясь на статистические методы анализа данных.

Актуальность темы связана со стремительными изменениями в индустрии розничной торговли, где умение прогнозировать покупательский спрос становится важным элементом стратегического управления.

Задачи исследования:

- проанализировать характеристики исследуемого объекта;
- смоделировать статистические зависимости;
- изучить построенную модель;
- программно реализовать модель и провести численный анализ полученных результатов.

Предмет исследования – разработка статистической модели для прогнозирования количества посетителей торговых центров.

Объект исследования – торговые центры как элементы современной розничной торговли.

1. АНАЛИЗ ХАРАКТЕРИСТИК ОБЪЕКТА ИССЛЕДОВАНИЯ

1.1 Описание объекта исследования

Объектом исследования являются торговые центры в контексте современной розничной торговли. В качестве примера для анализа был выбран один из крупнейших торговых центров Москвы — ТРЦ «Европейский». Первичные статистические данные о посещаемости этого ТРЦ были собраны за период с января 2018 года по декабрь 2022 года включительно [1]. На графике, представленном на рисунке 1, можно проследить зависимость числа посещений от времени. Более детальная информация с исходными данными приведена в Приложении А.

Рис. 1 – Зависимость количества посещений ТРЦ «Европейский» от времени.

1.2 Анализ объекта исследования с помощью статистических показателей

Вычислив абсолютный прирост, который составил $-4\,624\,400$ человек, можно констатировать наличие убывающей тенденции. Средний темп прироста, равный -7,5%,

показывает, насколько в среднем сократилась посещаемость ТРЦ за анализируемые четыре года.

Прогноз на 2022 год, сделанный на основе среднего абсолютного прироста, показал 47 830 943 человека, в то время как прогноз на основе среднего темпа прироста составил 48 508 684 человека.

Метод аналитического выравнивания с использованием уравнения прямой: $y = -6040054.5 \cdot t + 71338859$ даёт прогноз на 2022 год на уровне 41 138 587 человек.

На рисунке 2 можно также увидеть убывающую тенденцию и заметное расхождение между линией линейной регрессии и фактическими значениями посещаемости.

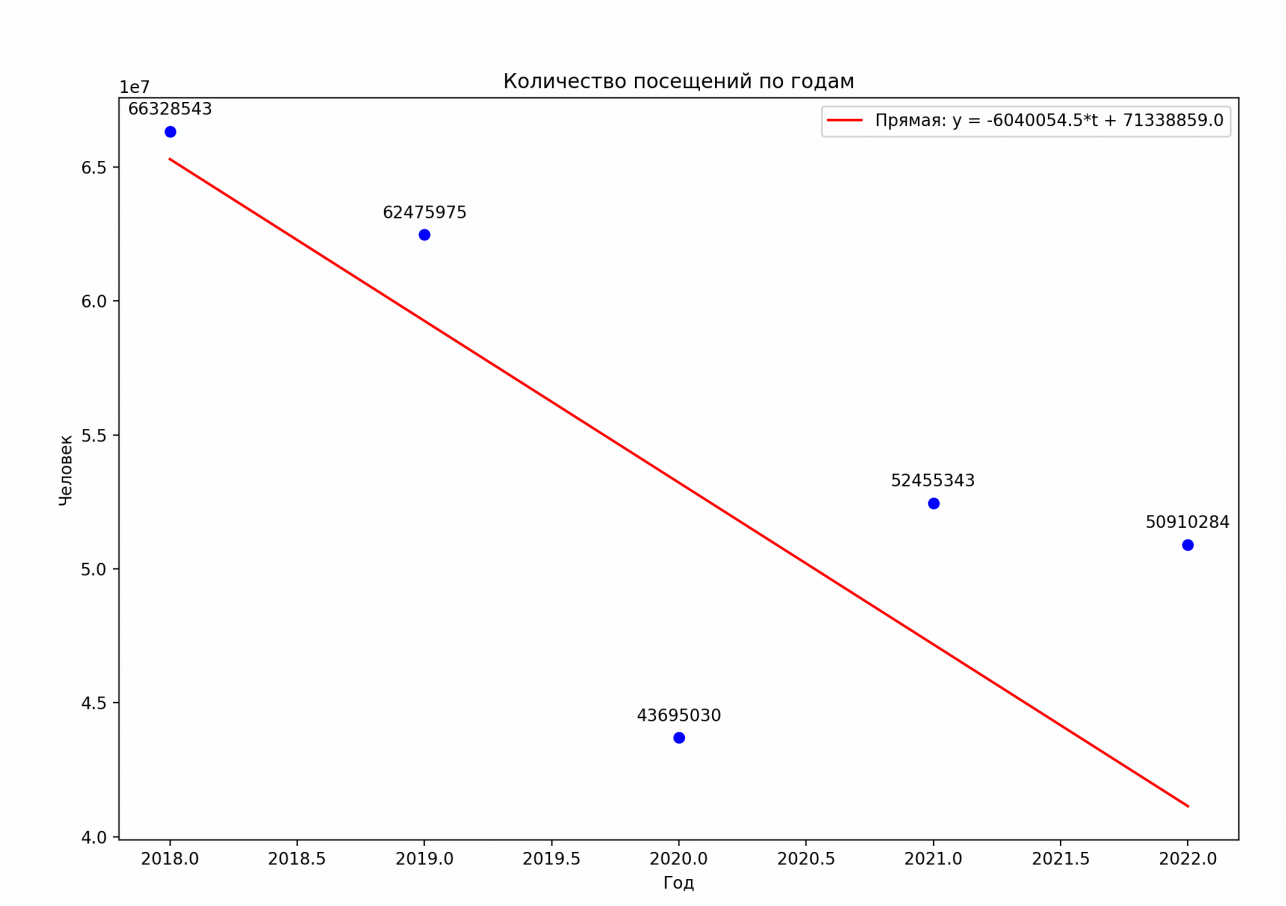

Рис. 2 – Количество посещений ТРЦ по годам

Подробные статистические показатели представлены в Приложении Б.

1.3 **Выявление причинно-следственных связей**

Изучение причинно-следственных связей для прогнозирования потока покупателей в торговом центре включает анализ ряда ключевых факторов.

- Одним из важнейших аспектов являются рекламные расходы, которые оказывают значительное влияние на привлечение посетителей. Успешные рекламные

кампании способны существенно увеличить поток клиентов и повысить их интерес к торговому центру.

- Индексы потребительских цен также играют важную роль, отражая влияние инфляции на покупательскую способность населения и служа индикатором изменений в потребительском поведении, что может повлиять на уровень посещаемости.
- Не менее важен временной аспект, включающий сезонные колебания, праздничные дни и дни недели, которые могут значительно изменять количество посетителей. Например, в праздничные дни и выходные посещаемость может значительно возрасти.
- Уровень безработицы является другим критически важным показателем. Высокий уровень безработицы может указывать на экономическую нестабильность, что, в свою очередь, негативно сказывается на посещаемости торговых центров.
- Анализ продаж за предыдущие периоды предоставляет ценную информацию. Изучение успешных периодов позволяет прогнозировать будущую посещаемость. Если в прошлом были успешные продажи, это может привлечь больше покупателей в будущем.
- Средняя зарплата населения также оказывает значительное влияние, поскольку рост доходов может привести к повышению покупательской активности и увеличению спроса на товары и услуги.
- Рост населения расширяет потенциальную клиентскую базу и может способствовать увеличению посещаемости торгового центра. Новые жители приносят с собой новые потребности, что может положительно сказаться на объемах продаж.

Проведение комплексного анализа этих факторов способствует более глубокому пониманию причинно-следственных связей в прогнозировании числа покупателей. Это позволяет эффективно управлять ресурсами и разрабатывать специальные предложения в периоды повышенного спроса. Для достижения надежных результатов необходимо проводить статистический анализ и моделирование, учитывающие взаимосвязи между перечисленными факторами.

1.4 Постановка задачи моделирования

Постановка задачи моделирования ориентирована на разработку и обучение модели, способной предсказывать количество покупателей в торговом центре. Для реализации этой цели будет применен специализированный набор данных, который содержит информацию о таких параметрах, как средняя зарплата населения, праздничные дни и другие ключевые переменные, влияющие на торговую активность.

Первым этапом станет подготовка и очистка данных, а также выявление факторов, оказывающих наибольшее влияние на число покупателей. Необходимо провести анализ структуры данных, чтобы обнаружить возможные пропуски или выбросы, способные негативно сказаться на качестве модели. Использование методов визуализации данных поможет глубже понять распределение значений и взаимосвязи между переменными.

На следующем этапе потребуется выбрать подходящий алгоритм, который сможет учесть все особенности процесса прогнозирования потока покупателей в зависимости от различных факторов. Уравнение множественной регрессии станет основой математической модели, так как оно эффективно учитывает несколько переменных, влияющих на точность предсказаний. Обучение модели будет проводиться на одной части данных, после чего её эффективность будет оценена на тестовой выборке.

Оценка качества модели будет включать анализ её точности, чувствительности и специфичности, а также использование других метрик, специально адаптированных для задач предсказания посещаемости торгового центра. Для этого будут созданы графики зависимостей, рассчитаны коэффициенты корреляции и выявлены взаимосвязи между переменными. Кроме того, необходимо определить оптимальную форму парной зависимости для более глубокого анализа данных.

С целью повышения точности прогноза будет проведён отбор незначительных переменных с применением шагового регрессионного анализа. Этот этап позволит исключить факторы, которые не имеют значительного влияния на результат, что, в свою очередь, увеличит предсказательную силу модели.

В результате завершения анализа и интерпретации полученных данных будет достигнута основная цель данного моделирования: создание инструмента для торговых компаний, который позволит заранее прогнозировать потенциальное количество посетителей в торговом центре. Это будет способствовать более эффективному планированию ресурсов и разработке целевых маркетинговых стратегий.

2. МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ

2.1 Формализация и классификация переменных

Выбранные статистические данные, отражающие зависимость количества посещений торгового центра от ряда потенциально полезных факторов $x_1, \dots x_5$, на основе наблюдений за пять лет.

x_1 – количественная дискретная переменная, представляющая количество праздничных и выходных дней в каждом месяце, измеряемая в днях.[2]

x_2 – количественная дискретная переменная, отражающая среднюю зарплату по Москве, выраженную в рублях. [6]

x_3 – количественная дискретная переменная, показывающая оборот розничной торговли непродовольственными товарами по месяцам в Москве, также в рублях. [5]

x_4 – количественная дискретная переменная, указывающая на временной аспект в виде месяца.

y – выходная количественная дискретная переменная, отражающая количество посещений торгового центра, измеряемая в числе человек.

Первичные статистические данные по каждому фактору приведены в Приложении В.

2.2 Проверка гипотезы о нормальном распределении выходной величины

Для проверки гипотезы о нормальном распределении использовались два метода: «Правило трёх сигм» и критерий Пирсона. Чтобы упростить вычисления, количество посетителей было переведено в тысячи человек и округлено до целого.

Согласно правилу трёх сигм, вероятность того, что случайная величина отклонится от своего среднего значения более чем на 3σ (три стандартных отклонения), достаточно высока. Если величина подчинена нормальному распределению $N(a, \sigma)$, то примерно 68% значений находятся в пределах $(a - \sigma, a + \sigma)$, около 95% – в интервале $(a-2\sigma, a+2\sigma)$, и 99.7% – в диапазоне $(a-3\sigma, a+3\sigma)$. Однако, применяя это правило к нашим данным, были получены значения 85.92, 95.78 и 97.19, что указывает на отсутствие нормального распределения для выходной величины.

Далее была проведена проверка гипотезы о нормальности случайной величины с использованием критерия Пирсона. Этот критерий χ^2 предназначен для оценки соответствия эмпирического распределения предполагаемому. Рассмотрим гипотезу H_0, согласно которой распределение является нормальным. Если наблюдаемое значение $\chi^2_{\text{набл}}$ не превышает

критическое значение $\chi^2_{\text{крит}}$, гипотеза H_0 может быть принята. В нашем случае наблюдаемое значение $\chi^2_{\text{набл}}$в 54 раза больше критического, что позволяет отвергнуть гипотезу о нормальном распределении.

Таким образом, данные о количестве посещений торгового центра не подчиняются нормальному распределению. Это может негативно повлиять на качество модели, поэтому рекомендуется провести выравнивание данных для улучшения точности предсказаний.

2.3 Корреляционный анализ

Корреляционный анализ представляет собой метод, позволяющий оценить степень связи между изменениями двух или более переменных. Основным показателем в этом процессе является коэффициент корреляции, наиболее распространённым является коэффициент Пирсона.

Этот коэффициент может принимать значения в диапазоне от -1 до 1, что позволяет судить о характере связи. Когда коэффициент близок к 1, это свидетельствует о сильной положительной корреляции; близкое к -1 значение указывает на выраженную отрицательную корреляцию. Коэффициент, находящийся около нуля, говорит о слабой взаимосвязи между переменными. Для наглядного представления взаимосвязей между несколькими переменными создаётся матрица корреляций, в которой представлены соответствующие коэффициенты (рисунок 3).

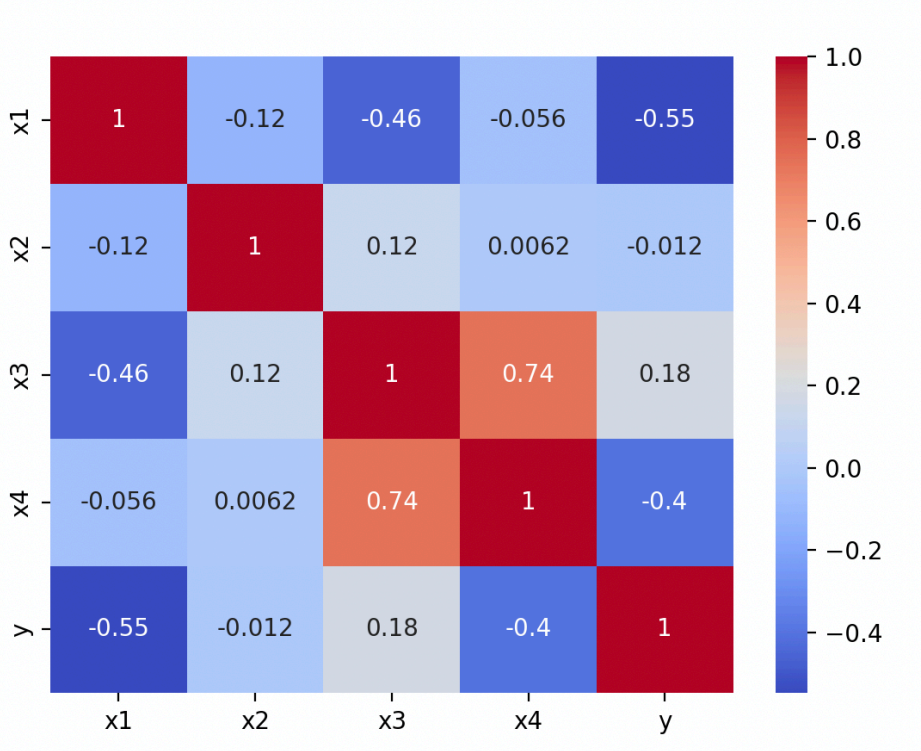

Рис. 3 – Корреляционная матрица

Можно заметить, что переменная x_2 имеет слабую корреляцию с выходной, а остальные переменные неплохо коррелируют с выходом. Корреляция между независимыми переменными называется мультиколлинеарностью, такая связь введет к неопределенности и плохим результатам предсказания.

2.4 Построение регрессионной модели

2.4.1 Структурная идентификация модели

Зависимой переменной является количество посетителей ТРЦ. Независимыми переменными являются 4 признака: количество праздничных дней, средняя заработная плата, оборота розничной торговли непродовольственных товаров, месяц по счету.

Рассмотрим уравнение множественной линейной регрессии $Y = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_n \cdot X_n + \varepsilon$, где Y – зависимая переменная, $X_1 \dots X_n$ – независимые переменные, $\beta_0, \beta_1 \dots \beta_n$ - коэффициенты регрессии, ε – случайная ошибка. Данная функциональная форма отлично подойдет для рассматриваемой задачи.

2.4.2 Параметрическая идентификация модели

В соответствии с методом наименьших квадратов, задача заключается в аппроксимации кривой известной функцией. Вычисление параметров уравнения множественной линейной регрессии будет произведено с помощью алгоритма МНК.

Уравнение множественной линейной регрессии, которое имеет вид:

$$Y = 1719000 - 107000X_1 - 1{,}58X_2 + 3{,}93X_3 - 67800X_4$$

3. ИССЛЕДОВАНИЕ МОДЕЛИ

3.1 Анализ статистической значимости уравнения регрессии

Общая сумма квадратов отклонений переменной y от среднего значения $\overline{y}$ может быть разложена на две составляющие:

$$S_y = S_{\text{факт}} + S_e,$$

где $S_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$ - общая сумма квадратов отклонений; $S_{\text{факт}} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$- сумма квадратов отклонений, объясненная регрессией; $S_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – остаточная сумма квадратов отклонений (необъясненная).

Выдвинем гипотезу о равенстве нулю коэффициентов регрессии. В том случае выходная переменная y не зависит от факторов, и вариация y обусловлена только воздействием ошибок: $S_y = S_e$. Противоположным является случай, при котором выходная переменная y функционально зависит от факторов: $S_y = S_{\text{факт}}$.

Для сравнения $S_{\text{факт}}$ и S_e их необходимо разделить на соответствующее число степеней свободы, получив таким образом средний квадрат отклонений на одну степень свободы – дисперсию: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$, $s_{\text{факт}}^2 = \frac{1}{m}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$, $s_e^2 = \frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Статистическая значимость уравнения регрессии определяется условием $s_{\text{факт}}^2 > s_e^2$. Задача сводится к проверке нулевой гипотезы $H_0: D_{\text{факт}} = D_e$ при конкурирующей гипотезе $H_1: D_{\text{факт}} > D_e$. Оценка статистической значимости уравнения регрессии выполняется с помощью F-критерия Фишера: $F = \frac{s_{\text{факт}}^2}{s_e^2}$.

Уравнение регрессии является статистически значимым, если:

1. F попадает в критическую область при заданном уровне значимости α, то есть $F > F_{\text{кр}}$

2. Уровень значимости α_F, для которого F является критической точкой (вероятность нулевой гипотезы, P-значение) меньше заданного уровня значимости α, то есть $\alpha_F < \alpha$.

Для данной модели $F = 6.601,\ \alpha_F = 0{,}002$ при $F_{\text{кр}} = 2.7694, \alpha = 0.05$, что удовлетворяет заданным условиям. Следовательно, текущее уравнение регрессии можно назвать статистически значимым.

3.2 Анализ статистической значимости коэффициентов уравнения регрессии

Для проверки значимости коэффициентов (таблица 1) формулируются гипотезы: $H_0: \beta j = 0$ (коэффициент незначим), $H_1: \beta_j \neq 0$ (коэффициент значим). В качестве критерия выбирается случайная величина T_j, распределенная по закону Стьюдента с $n - m - 1$ степенями свободы: $T_j = \frac{\beta_j}{s_j}$, где β_j – коэффициент уравнения регрессии при факторе x_j, s_j – стандартная ошибка коэффициента β_j. $s_j = s\sqrt{[(C^T C)^{-1}]_{jj}}$, где $[(C^T C)^{-1}]_{jj}$ – j-й диагональный элемент матрицы $(C^T C)^{-1}$, $s = \sqrt{s_e^2}$.

Коэффициент β_j статистически значим, то есть значимо отличается от нуля (принимается гипотеза H_1 на уровне значимости α), если:

1. T_j попадает в критическую область при заданном уровне значимости α, то есть $|T\_j| > T_{\text{кр}}$;

2. Уровень значимости α_{T_j}, для которого T_j является критической точкой (P-значение) меньше заданного уровня значимости α: $\alpha_{T_j} < \alpha$.

Интервальная оценка для коэффициентов β_j определяется с помощью доверительного интервала $(\beta_j - t_\gamma s_j;\ \beta_j + t_\gamma s_j)$, где $t_\gamma = t(a, n - m - 1)$.

Таблица 1. Статистическая значимость коэффициентов регрессии.

		coef	std err	t	P>|t|	[0.025	0.975]
0	const	1.719e+06	1.21e+06	1.425	0.163	-7.29e+05	4.17e+06
1	x1	-1.07e+05	3.67e+04	-2.915	0.006	-1.82e+05	-3.25e+04
2	x2	-1.5814	0.926	-1.707	0.097	-3.462	0.299
3	x3	3.9289	0.714	5.506	0.000	2.480	5.377
4	x4	-6.78e+04	9315.586	-7.278	0.000	-8.67e+04	-4.89e+04

Работая с уровнем значимости $\alpha = 0.05$, заметны коэффициенты, которые являются статистически незначимыми, то есть они оказывают незначительно влияние на нашу модель. Избавление от таких коэффициентов может привести к лучшим результатам предсказания.

3.3 Исследование мультиколлинеарности факторов

Мультиколлинеарность в контексте множественной регрессии подразумевает высокую степень взаимной корреляции между независимыми переменными.

Это явление может привести к следующим последствиям:

1. Хотя матрица $(C^T\ C)$ может оставаться невырожденной, её определитель будет мал, что вызывает резкое увеличение значений элементов обратной матрицы. Это, в свою очередь, приводит к значительным дисперсиям оценок коэффициентов.

2. Оценки коэффициентов становятся чувствительными к небольшим изменениям в наблюдаемых данных и размере выборки, что делает модель менее пригодной для анализа и прогнозирования.

3. t-статистики коэффициентов уменьшаются, и их оценка по t-критерию теряет свою информативность.

Если в матрице парных коэффициентов корреляции наблюдаются высокие значения между парами переменных, это указывает на наличие мультиколлинеарности. В случае, когда факторы не коррелированы, матрица парных корреляций будет единичной, а её определитель равен 1. Если же факторы взаимосвязаны, все коэффициенты корреляции будут равны единице, а определитель станет равным нулю. Таким образом, чем ближе определитель матрицы парных корреляций к нулю, тем более выражена мультиколлинеарность, и наоборот.

В данном случае определитель матрицы составляет 0.07583, что не близко к нулю. Тем не менее, анализ матрицы парных корреляций (см. рисунок 3) показывает значительную корреляцию между признаками x_4 и x_3. Это свидетельствует о наличии мультиколлинеарности в модели. Рекомендуется исключить некоторые факторы для её оптимизации.

3.4 Применение шагового регрессионного анализа для улучшения модели

Шаговый регрессионный анализ реализуется двумя способами. С помощью добавления факторов и с помощью их удаления. При добавлении определяется фактор, имеющий наиболее высокий коэффициент корреляции с выходной величиной, а после происходит пошаговое добавление остальных факторов исходя из условия увеличения скорректированного коэффициента детерминации. При удалении факторов берется модель с максимальным числом переменных, на каждом шаге проводится удаление наименее значимого фактора. Изначальный $R^2_{adj} = 0.717$.

Шаги при удалении:

1. Удаление «x_2» со Р-значением 0.097 приводит к $R^2_{adj} = 0.713$

На втором шаге и далее происходит уменьшение оценки. Следовательно, признаки, которые необходимо использовать все признаки.

Шаги при добавлении:

1. Добавление «x_3» с приводит к $R^2_{adj} = 0.115$
2. Добавление «x_4» приводит к $R^2_{adj} = 0.669$
3. Добавление «x_1» приводит к $R^2_{adj} = 0.713$
4. Добавление «x_2» приводит к $R^2_{adj} = 0.717$

Таким образом, в двух случаях была нет необходимости исключать факторы.

4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ

4.1 Обоснование выбора и описание программного обеспечения

В ходе работы основным языком программирования стал Python, что объясняется его широкими возможностями: компактным синтаксисом, богатой стандартной библиотекой и поддержкой активного сообщества разработчиков. Это значительно упрощает как разработку, так и сопровождение кода.

Для обработки и анализа данных применялась библиотека Pandas, которая предлагает удобные инструменты для работы с табличными данными. Для реализации статистических моделей использовались библиотеки Scikit-learn и Statsmodels, предоставляющие разнообразные алгоритмы для анализа и построения моделей.

Визуализация результатов и исследование структуры данных осуществлялись с помощью Matplotlib и Seaborn. Гибкость и функциональность этих библиотек позволили создать информативные графики, способствующие глубокому пониманию данных.

4.2 Описание основных модулей программы

Для начала были импортированы все необходимые библиотеки, которые будут использованы на протяжении всей работы.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy.stats import f, chi2
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error
from tabulate import tabulate
import os
```

Листинг 1. Импортирование библиотек.

Первичные статистические данные считываются из файлов и организуются в pandas DataFrame с помощью функции `load_data()`.

```
def load_data(data_paths):
    """Загружает данные из заданных путей."""
```

```
    data = {column: np.loadtxt(path, converters={0: np.int64}) for column,
path in data_paths.items()}
    df = pd.DataFrame(data)
    df['timestamp'] = pd.period_range(start='2018-01', end='2022-12',
freq='M')
    df.set_index('timestamp', inplace=True)
    return df
```

Листинг 2. Обработка первичных данных.

Определение и вывод гистограммы интервального ряда распределения для выходной величины описано в функциях `compute_discrete_distribution()`, `compute_interval_distribution()`, `plot_histogram()`.

```
def compute_discrete_distribution(data):
    """Вычисляет дискретное распределение."""
    distribution = {}
    for value in data:
        distribution[value] = distribution.get(value, 0) + 1
    return dict(sorted(distribution.items()))

def compute_interval_distribution(data):
    """Вычисляет интервальное распределение."""
    discrete_distribution = compute_discrete_distribution(data)
    num_intervals = int(np.ceil(1 + 3.222 * np.log10(max(data) -
min(data))))
    interval_length = int(np.ceil((max(data) - min(data)) / num_intervals))

    intervals = [(i, i + interval_length) for i in range(min(data),
max(data), interval_length)]
    frequencies = [0] * len(intervals)

    for i, interval in enumerate(intervals):
        for value, count in discrete_distribution.items():
            if interval[0] <= value < interval[1]:
                frequencies[i] += count

    return intervals, frequencies, num_intervals, interval_length


def plot_histogram(data, num_bins):
    """Строит гистограмму данных."""
    plt.hist(data, bins=num_bins, range=(min(data), max(data)),
edgecolor='black')
    plt.xlabel('Тыс. человек')
    plt.ylabel('Частота')
    plt.title('Гистограмма количества посещений')
    plt.show()
```

Листинг 3. Гистограммы интервального ряда распределения отклика

Были реализованы функции для проверки выходной переменной на нормальное распределение. Функция проверки правила трёх сигм выводит проценты вхождений в интервалы: одной, двух и трёх сигм `compute_normal_distribution_properties()`.

```
def compute_normal_distribution_properties(data):
    """Вычисляет свойства нормального распределения."""
    mean_value = np.mean(data)
    std_dev = np.std(data)

    sigma_68 = np.mean((mean_value - std_dev <= data) & (data <= mean_value
+ std_dev)) * 100
    sigma_95 = np.mean((mean_value - 2 * std_dev <= data) & (data <=
mean_value + 2 * std_dev)) * 100
    sigma_99 = np.mean((mean_value - 3 * std_dev <= data) & (data <=
mean_value + 3 * std_dev)) * 100

    return sigma_68, sigma_95, sigma_99
```

Листинг 4. Функция правила трёх сигм

Функция проверки нормальности распределения с помощью критерия Пирсона использует вспомогательную функцию, делящую данные на интервалы.

```
def compute_normal_distribution_pearson(data, intervals, frequencies,
interval_length):
    """Проверяет нормальность распределения по критерию Пирсона."""
    n = sum(frequencies)
    midpoints = [(left + right) / 2 for left, right in intervals]

    def mean_midpoints(intervals, frequencies):
        return np.sum([((left + right) / 2) * frequencies[i] for i, (left,
right) in enumerate(intervals)]) / sum(frequencies)

    def variance_midpoints(intervals, frequencies):
        return np.sum([(((left + right) / 2) - mean_midpoints(intervals,
frequencies)) ** 2 * frequencies[i] for i, (left, right) in
enumerate(intervals)]) / sum(frequencies)

    def std_dev_midpoints(intervals, frequencies):
        return np.sqrt(variance_midpoints(intervals, frequencies))

    standardized_midpoints = [(point - mean_midpoints(intervals,
frequencies)) / std_dev_midpoints(intervals, frequencies) for point in
midpoints]
```

```
    def laplace_function(x):
        return np.exp(- (x ** 2 / 2)) / (np.sqrt(2 * np.pi))

    theoretical_frequencies = [(interval_length * n /
std_dev_midpoints(intervals, frequencies)) * laplace_function(ui) for ui in
standardized_midpoints]

    chi_squared_observed = np.sum([(frequencies[i] - j) ** 2 / j for i, j in
enumerate(theoretical_frequencies)])

    degrees_of_freedom = 2
    chi_squared_critical = chi2.ppf(1 - 0.05, len(intervals) -
degrees_of_freedom - 1)
    return chi_squared_observed, chi_squared_critical, chi_squared_observed
/ chi_squared_critical
```

Листинг 5. Критерий хи-квадрат Пирсона

Функция, реализующая корреляционный анализ, выводит на экран матрицу парных корреляций и её детерминант.

```
def analyze_correlation(df):
    """Визуализирует матрицу корреляций."""
    sns.heatmap(df.corr(), annot=True, cmap="coolwarm", annot_kws={"size":
10})
    plt.show()
    print(f'Детерминант матрицы парных корреляций:
{np.linalg.det(df.corr().to_numpy())}')
```

Листинг 6. Функция корреляционного анализа

Проверка значимости уравнения регрессии с помощью критерия Фишера.

```
def fisher_test(y_true, X, model, num_samples, num_features):
    """Выполняет тест Фишера."""
    S2_fact = np.sum((model.predict(X) - np.mean(y_true)) ** 2) /
num_features
    S2_e = np.sum((y_true - model.predict(X)) ** 2) / (num_samples -
num_features - 1)
    F_statistic = S2_fact / S2_e
    alpha = 0.05
    critical_value = f.ppf(1 - alpha, num_features, num_samples -
num_features - 1)
    p_value = 1 - f.cdf(F_statistic, num_features, num_samples -
num_features)

    return F_statistic, critical_value, p_value
```

Листинг 7. Критерий Фишера

Для оценки качества модели была написана функция, которая выводит на экран абсолютную ошибку среднего, среднеквадратичную ошибку, коэффициент детерминации и его исправленную версию.

```
def evaluate_model(y_true, y_predicted, num_samples, num_features, model):
    """Оценивает качество модели."""
    mse = mean_squared_error(y_true, y_predicted)
    mae = mean_absolute_error(y_true, y_predicted)
    r_squared = model.rsquared
    r_squared_adj = 1 - (num_samples - 1) / (num_samples - num_features - 1)
* (1 - r_squared)

    evaluation_metrics = [
        ["Среднеквадратичная ошибка (MSE)", f"{mse:.4f}"],
        ["Средняя абсолютная ошибка (MAE)", f"{mae:.4f}"],
        ["Коэффициент детерминации", f"{r_squared:.3f}"],
        ["Адаптивный коэффициент детерминации", f"{r_squared_adj:.3f}"]
    ]

    print(tabulate(evaluation_metrics, headers=["Метрика", "Значение"],
tablefmt="fancy_grid"))
```

Листинг 8. Функция вывода оценок

Основная часть программы.

```
def main():
    data_paths = {
        "x1": './punkt2/x1.txt',
        "x2": './punkt2/x2.txt',
        "x3": './punkt2/x3.txt',
        "x4": './punkt2/x4.txt',
        "y": './punkt2/y.txt'
    }
    df = load_data(data_paths)
    target_values = (df['y'].values * 0.001).round().astype(int)

    print('Интервальное распределение', '\n')
    intervals, frequencies, num_intervals, interval_length =
compute_interval_distribution(target_values)
    interval_table = [["Интервал", "Частота"]]
    for i, interval in enumerate(intervals):
        interval_table.append([f"[{interval[0]}, {interval[1]})",
frequencies[i]])
    print(tabulate(interval_table, headers="firstrow",
tablefmt="fancy_grid"))
    print()
```

```
    plot_histogram(target_values, num_intervals)

    print('Нормальное распределение по теореме 3-х сигм')
    sigma_68, sigma_95, sigma_99 =
compute_normal_distribution_properties(target_values)

    sigma_table = [
        ["Интервал", "Процент"],
        ["1 сигма", f"{sigma_68:.2f}"],
        ["2 сигмы", f"{sigma_95:.2f}"],
        ["3 сигмы", f"{sigma_99:.2f}"]
    ]
    print(tabulate(sigma_table, headers="firstrow", tablefmt="fancy_grid"))

    if sigma_68 > 68 and sigma_95 > 95 and sigma_99 > 99.7:
        print('Распределение нормальное')
    else:
        print('Распределение не нормальное')
    print()

    print('Нормальное распределение по критерию Пирсона')
    chi_squared_observed, chi_squared_critical, chi_squared_ratio =
compute_normal_distribution_pearson(target_values, intervals, frequencies,
interval_length)

    pearson_table = [
        ["Показатель", "Значение"],
        ["Хи-квадрат наблюдаемое", f"{chi_squared_observed:.3f}"],
        ["Хи-квадрат критическое", f"{chi_squared_critical:.3f}"],
        ["Рассчитанное значение", f"{chi_squared_ratio:.3f}"]
    ]
    print(tabulate(pearson_table, headers="firstrow",
tablefmt="fancy_grid"))
    print()

    analyze_correlation(df)

    X = df.drop('y', axis=1)
    y = df['y']
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.33, random_state=42)
    X_train_mn = sm.add_constant(X_train)
    X_test_mn = sm.add_constant(X_test)
    model = sm.OLS(y_train, X_train_mn).fit()
    y_predicted = model.predict(X_test_mn)

    F_statistic, critical_value, p_value = fisher_test(y_test, X_test_mn,
model, len(y_test), X_test_mn.shape[1] - 1)

    print(f"F-критерий: {F_statistic:.3f}")
```

```
    print(f"Критическое значение: {critical_value:.3f}")
    print(f"P-значение: {p_value:.3f}")
    print()

    result_summary = model.summary()
    coefficients_table = pd.DataFrame(result_summary.tables[1].data[1:],
columns=result_summary.tables[1].data[0])
    coefficients_table.columns = [' ', 'coef', 'std err', 't', 'P>|t|',
'[0.025', '0.975]']

    print('Анализ статистической значимости коэффициентов уравнения
регрессии:')
    print(tabulate(coefficients_table, headers='keys',
tablefmt='fancy_grid'))
    print()

    coefficients_table.to_csv('./punkt2/coefficients_table.csv', index=True)

    evaluate_model(y_test, y_predicted, len(y_test), X_test_mn.shape[1] - 1,
model)
```

Листинг 9. Основная функция.

4.3 Численное исследование результатов моделирования

Оценка модели была произведена по 4 характеристикам, оценивающим качество предсказаний и модели в целом:

- MSE (Mean Squared Error) – средняя квадратичная ошибка. Измерение среднего квадрата разности между предсказанными и фактическими значениями.
- MAE (Mean Absolute Error) – средняя абсолютная ошибка. Измерение среднего значения абсолютных разностей между предсказанными и фактическими значениями.
- R^2 score – коэффициент детерминации. Измеряет долю дисперсии зависимой переменной, которая может быть объяснена моделью. Принимает значения от 0 до 1, где 1 означает идеальное предсказание.
- Adjusted R^2 score – скорректированный коэффициент детерминации, учитывающий количество предикторов в модели и корректирующий R_2 score в случае наличия избыточных предикторов.

Таблица 2. Характеристики модели.

Метрика	Значение
Среднеквадратичная ошибка (MSE)	3,93E+11
Средняя абсолютная ошибка (MAE)	530223
Коэффициент детерминации	0,777

Адаптивный коэффициент детерминации	0,717

Исходя из таблицы 2 коэффициент детерминации нормальный, это свидетельствует о не плохом качестве предсказаний. MSE принимает высокое значение. MAE примерно равно 530223, в контексте рассматриваемой задачи это значит, что предсказанной значение потенциального количества посетителей ТРЦ может отличаться от истинного значения на ± 530223.

Пусть коэффициенты детерминации не плохие, значения MSE и MAE достаточно высоки, поэтому следует произвести улучшение модели, чтобы добиться приемлемого качества предсказания.

ВЫВОДЫ

В результате проведенного исследования были обнаружены статистические связи между числом посетителей торгового центра и множеством факторов. Анализ характеристик объекта исследования позволил выделить ключевые переменные, оказывающие влияние на целевой показатель. При формализации и классификации переменных была проверена гипотеза о нормальности распределения выходных данных.

Корреляционный анализ подтвердил наличие значимых статистических взаимосвязей между переменными, а построение регрессионной модели дало возможность выявить структурные и параметрические характеристики влияющих факторов. Проверка модели подтвердила статистическую значимость регрессионного уравнения и позволила оценить значимость его коэффициентов. Применение шагового регрессионного анализа способствовало улучшению модели, оптимизации коэффициентов и устранению мультиколлинеарности факторов.

В процессе программной реализации и численного анализа результатов моделирования был обоснован выбор используемого программного обеспечения, описаны основные модули программы и проведен количественный анализ полученных результатов.

Разработанная модель прогнозирования посещаемости торгового центра является эффективным инструментом для предсказания будущих тенденций. Полученные результаты исследования окажут помощь в принятии управленческих решений в сфере торговли.

СПИСОК ЛИТЕРАТУРЫ

1. Европейский торговый центр. (н.д.). О компании. URL: https://europe-tc.ru/about/ (дата обращения: 9.10.2024)
2. Гарант.Ру. (н.д.). Календарь бухгалтера и юриста. URL: https://www.garant.ru/calendar/buhpravo/ (дата обращения: 9.10.2024)
3. Гудман, Л. А., Тьюри, Э. Статистика: принципы и практика. М.: Наука, 2017.
4. Джонсон, Р. А., Кастилло, Л. Введение в статистику. М.: Вильямс, 2016.
5. Федеральная служба государственной статистики. (н.д.). Розничная торговля. URL: https://rosstat.gov.ru/statistics/roznichnayatorgovlya (дата обращения: 9.10.2024)
6. Федеральная служба государственной статистики. (н.д.). Рынок труда, занятость и заработная плата. URL: https://rosstat.gov.ru/labor_market_employment_salaries (дата обращения: 9.10.2024)
7. Фредман, Л. Статистика для бизнеса и экономики. М.: Дело, 2018.

ПРОГРАММНЫЕ ПРИЛОЖЕНИЯ

Программной реализации алгоритма:

URL: https://github.com/darcysoul/cursovaya_psa

ПРИЛОЖЕНИЕ А

Статистика посещений ТРЦ «Европейский» с 2018 по 2022 года.

	Посещаемость (чел.)				
месяц\год	**2018**	**2019**	**2020**	**2021**	**2022**
январь	5 920 777	5 268 603	5 122 083	4 259 785	4 336 041
февраль	5 107 023	4 630 542	5 025 002	4 272 518	4 360 325
март	5 260 188	5 251 521	3 869 908	4 647 206	4 634 183
апрель	5 241 923	5 394 705	217 114	4 651 527	3 958 694
май	5 513 691	4 714 597	258 315	4 201 100	3 963 504
июнь	5 908 677	5 490 355	2 629 653	4 193 056	4 096 926
июль	5 323 940	4 641 573	3 970 811	4 217 312	4 322 628
август	5 939 614	4 793 950	4 399 658	4 471 314	4 192 339
сентябрь	4 861 658	4 936 602	3 806 806	4 233 799	3 807 424
октябрь	4 632 388	5 558 193	4 723 175	4 269 717	3 831 074
ноябрь	5 334 797	5 718 277	4 688 987	4 036 141	4 454 125
декабрь	7 283 867	6 077 057	4 983 518	5 001 868	4 953 021
итого	66 328 543	62 475 975	43 695 030	52 455 343	50 910 284

ПРИЛОЖЕНИЕ Б

Статистические характеристики данных о посещении ТРЦ.

Год	Посещения, чел.
2018	66328543
2019	62475975
2020	43695030
2021	52455343
2022	50910284

Год	Посещения, чел.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2018	66328543			
2019	62475975	-3852568.0	94.2	-5.8
2020	43695030	-22633513.0	65.9	-34.1
2021	52455343	-13873200.0	79.1	-20.9
2022	50910284	-15418259.0	76.8	-23.2

Год	Посещения, чел.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2018	66328543			
2019	62475975	-3852568.0	94.2	-5.8
2020	43695030	-18780945.0	69.9	-30.1
2021	52455343	8760313.0	120.0	20.0
2022	50910284	-1545059.0	97.1	-2.9

Средний уровень ряда: 56238723 (чел.)

Средний абсолютный прирост: -4624400 (чел.)

Средний темп роста: 92.5%

Средний темп прироста: -7.5%

Прогноз по среднему абсолютному приросту: 47830943 (чел.)

Прогноз по среднему темпу роста: 48508684 (чел.)

Прогноз по МНК: 41138587 (чел.)

Относительная погрешность по среднему абсолют. приросту: 6.05%

Относительная погрешность по среднему темпу роста: 4.72%

Относительная погрешность по МНК: 19.19%

ПРИЛОЖЕНИЕ В

Первичные статистические данные о факторах.

Год		x_1	x_2	x_3	x_4
2018	**январь**	14	70251	1202292	1
2018	**февраль**	9	80184	1165811	2
2018	**март**	11	84082	1278051	3
2018	**апрель**	9	89318	1274769	4
2018	**май**	11	81064	1309579	5
2018	**июнь**	10	90094	1349531	6
2018	**июль**	9	80999	1396384	7
2018	**август**	8	77618	1469262	8
2018	**сентябрь**	10	77274	1452049	9
2018	**октябрь**	8	791506	1453254	10
2018	**ноябрь**	9	78947	1453568	11
2018	**декабрь**	10	113989	1719395	12
2019	**январь**	14	79681	1291648	13
2019	**февраль**	8	85370	1259407	14
2019	**март**	11	95179	1379374	15
2019	**апрель**	8	102908	1368527	16
2019	**май**	13	89045	1383874	17
2019	**июнь**	11	96030	1424197	18
2019	**июль**	8	91608	1468437	19
2019	**август**	9	86734	1540480	20
2019	**сентябрь**	9	86685	1514309	21
2019	**октябрь**	8	89129	1530861	22
2019	**ноябрь**	8	88657	1542425	23
2019	**декабрь**	9	135375	1800000	24
2020	**январь**	14	88845	1367384	25
2020	**февраль**	10	92390	1356534	26
2020	**март**	12	105238	1522431	27
2020	**апрель**	30	101552	914977	28
2020	**май**	17	91824	1041553	29
2020	**июнь**	10	98700	1340047	30
2020	**июль**	9	98930	1511632	31
2020	**август**	10	90304	1608996	32
2020	**сентябрь**	8	93062	1574094	33
2020	**октябрь**	9	94065	1606022	34
2020	**ноябрь**	10	95315	1588013	35

2020	**декабрь**	8	153648	1854811	36
2021	**январь**	16	93059	1476753	37
2021	**февраль**	9	104451	1448633	38
2021	**март**	9	116355	1622265	39
2021	**апрель**	8	117769	1653338	40
2021	**май**	16	105247	1678827	41
2021	**июнь**	9	109307	1701207	42
2021	**июль**	9	108520	1765410	43
2021	**август**	9	98974	1877278	44
2021	**сентябрь**	8	102233	1849605	45
2021	**октябрь**	10	103394	1847684	46
2021	**ноябрь**	10	104878	1805998	47
2021	**декабрь**	9	172553	2190795	48
2022	**январь**	15	103124	1728245	49
2022	**февраль**	9	114701	1751505	50
2022	**март**	9	146044	1981923	51
2022	**апрель**	9	120502	1666472	52
2022	**май**	13	113671	1676048	53
2022	**июнь**	9	123689	1700349	54
2022	**июль**	10	115294	1751405	55
2022	**август**	8	109060	1847690	56
2022	**сентябрь**	8	113896	1773711	57
2022	**октябрь**	10	113163	1789380	58
2022	**ноябрь**	9	113723	1806308	59
2022	**декабрь**	9	185646	2070194	60