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 потребительских расходов населения»

Выполнила:
студентка 4-го курса, гр. БИВТ-21-4
Савенко Е. И.

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

Москва 2025

Оглавление

ВВЕДЕНИЕ

Экономическое поведение населения оказывает значительное влияние на макро- и микроэкономические процессы. Одним из ключевых аспектов такого поведения являются потребительские расходы, которые формируют спрос на товары и услуги, определяют уровень экономической активности и служат индикатором благосостояния общества. Разработка эффективных методов прогнозирования потребительских расходов позволяет более точно оценивать будущие экономические тенденции, что особенно важно для государственных органов, бизнеса и финансовых учреждений. В данной курсовой работе будет разработана модель прогнозирования потребительских расходов.

Целью работы является создание модели, позволяющей прогнозировать сумму потребительских расходов, опираясь на статистические методы анализа данных.

Актуальность темы связана с необходимостью повышения эффективности экономического планирования, улучшения стратегий управления бизнесом и разработки государственных мер по регулированию экономики.

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 прогнозирования потребительских расходов населения.

Объект исследования – потребительские расходы населения как элемент социально-экономической системы.

1. АНАЛИЗ ХАРАКТЕРИСТИК ОБЪЕКТА ИССЛЕДОВАНИЯ

1.1 Описание объекта исследования

Объектом исследования являются потребительские расходы населения. Для анализа была взята статистика, представленная Росстатом. Первичные статистические данные о посещаемости были собраны за период с января 2019 года по декабрь 2024 года включительно [1]. На графике, представленном на рисунке 1, можно проследить зависимость суммы потребительских расходов от времени. Более детальная информация с исходными данными приведена в Приложении А.

Рисунок 1 – Зависимость количества потребительских расходов от времени.

1.2 Анализ объекта исследования с помощью статистических показателей

Вычислив абсолютный прирост, который составил 5040 миллиардов рублей, можно констатировать наличие возрастающей тенденции. Средний темп прироста, равный 9,5 %, показывает, насколько в среднем увеличилась сумма потребительских расходов за анализируемые пять лет.

Прогноз на 2024 год, сделанный на основе среднего абсолютного прироста, показал 71182 миллиардов рублей, в то время как прогноз на основе среднего темпа прироста составил 72435 миллиардов рублей.

Метод аналитического выравнивания с использованием уравнения прямой: $y = 5354{,}5 \cdot t + 37408{,}9$ даёт прогноз на 2024 год на уровне 69536 миллиардов рублей.

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

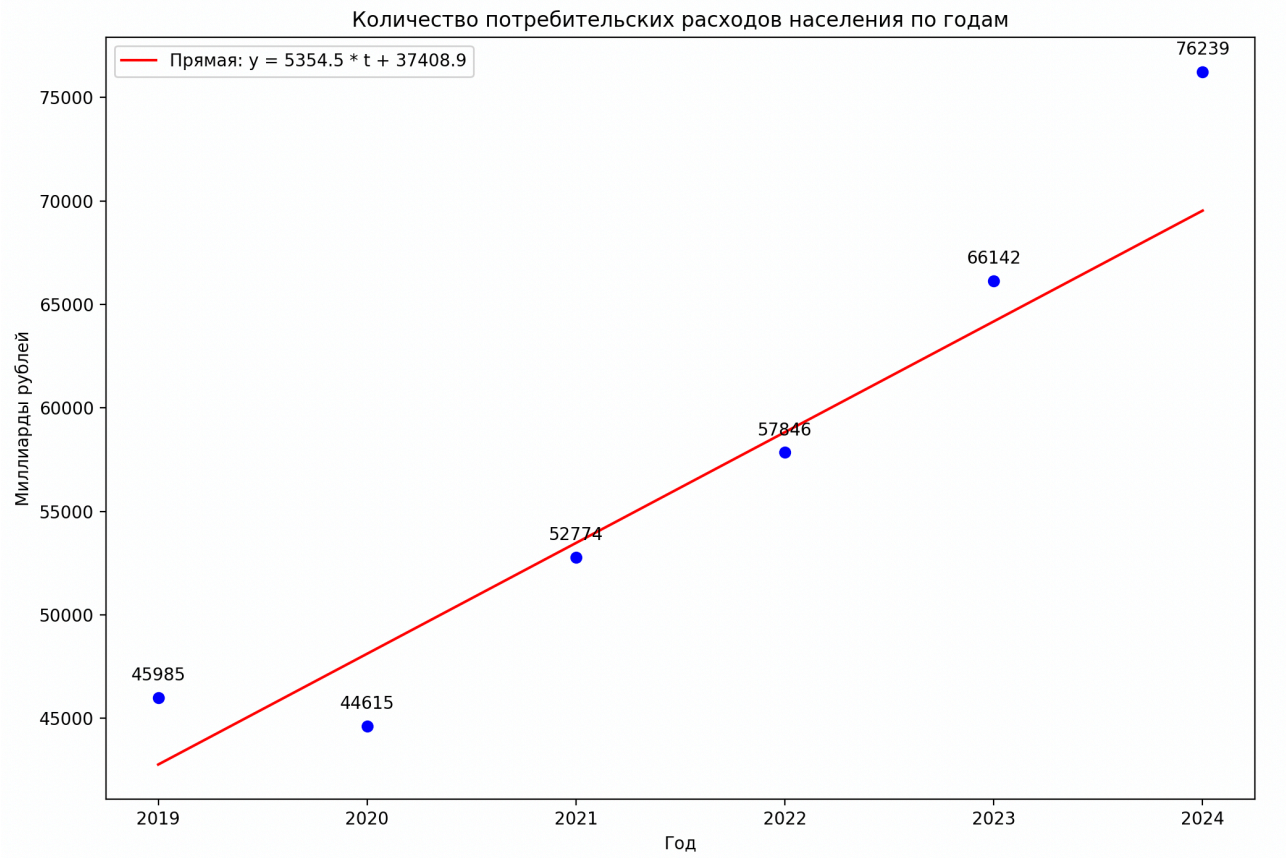

Рисунок 2 – Предсказание суммы потребительских расходов на 2024

Подробные статистические показатели представлены в Приложении А.

1.3 Выявление причинно-следственных связей

Изучение причинно-следственных связей для прогнозирования потребительских расходов населения включает анализ ряда ключевых факторов.

- Одним из важнейших аспектов являются доходы, которые напрямую влияют на потребительские расходы. Этим фактором можно объяснить изменения в уровне жизни и спросе на товары, а также корреляцию между экономическим ростом и потребительскими расходами.
- Уровень безработицы является важным индикатором экономической нестабильности и может существенно повлиять на потребительские расходы. Высокий уровень безработицы может привести к снижению доходов населения, что, в свою очередь, уменьшает возможности для потребления.

- Общее количество населения напрямую связано с объемом потребительских расходов, поскольку большее число людей означает больший рынок для товаров и услуг.
- Инфляция оказывает значительное влияние на потребительские расходы, так как она повышает цены на товары и услуги, что, в свою очередь, снижает покупательную способность населения. Высокая инфляция может привести к тому, что даже при стабильных доходах люди будут вынуждены сокращать расходы, выбирая более дешевые товары или сокращая потребление. С другой стороны, умеренная инфляция может стимулировать потребление, если люди ожидают роста цен в будущем.
- Не менее важен временной аспект, включающий сезонные колебания, которые могут значительно изменять сумму расходов. Например, в месяцы перед праздниками или перед сменой сезона может значительно возрасти количество расходов.

Проведение комплексного анализа этих факторов способствует более глубокому пониманию причинно-следственных связей в прогнозировании потребительских расходов. Это позволяет выявлять уязвимости в экономике и разрабатывать эффективные стратегии для стимулирования или коррекции потребительского 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 потребительские расходы населения. Для реализации этой цели будет применен специализированный набор данных, который содержит информацию о таких параметрах, как доходы населения, численность населения и другие ключевые переменные, влияющие на покупательскую активность.

Первым этапом станет подготовка и очистка данных, а также выявление факторов, оказывающих наибольшее влияние на потребительские расходы. 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 потребительских расходов 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 потребительских расходов. 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, который позволит заранее прогнозировать потенциальную сумму потребительских расходов населения. 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 суммы потребительских расходов населения от ряда потенциально полезных факторов $x_1, \dots\ x_5$, на основе наблюдений за шесть лет.

x_1 – количественная дискретная переменная, представляющая среднедушевые денежные доходы населения, измеряемая в миллиардах рублей в месяц. [2]

x_2 – количественная дискретная переменная, отражающая численность безработных в возрасте 15-72 лет, выраженную в тысячах человек в месяц. [3]

x_3 – количественная дискретная переменная, указывающая на численность населения, измеряемая в миллионах человек в год. [4]

x_4 – количественная дискретная переменная, равная инфляции, выраженная в процентах в месяц. [5]

x_5 – количественная дискретная переменная, означающая номер месяца по счету.

y – выходная количественная дискретная переменная, отражающая сумму потребительских расходов населения, измеряемая в миллиардах рублей в месяц.

Первичные статистические данные по каждому фактору приведены в Приложении Б.

2.2 Проверка гипотезы о нормальном распределении выходной величины

Для проверки гипотезы о нормальном распределении использовались два метода: «Правило трёх сигм» и критерий Пирсона.

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 $(a-3\sigma, a+3\sigma)$.

Для начала разделим данные на интервалы, гистограмма приведена на рисунке 3. Применяя правило трех сигм к нашим данным, были получены значения 68.06, 98.61 и 100, что указывает на то, что выходная величина близка к нормальному распределению.

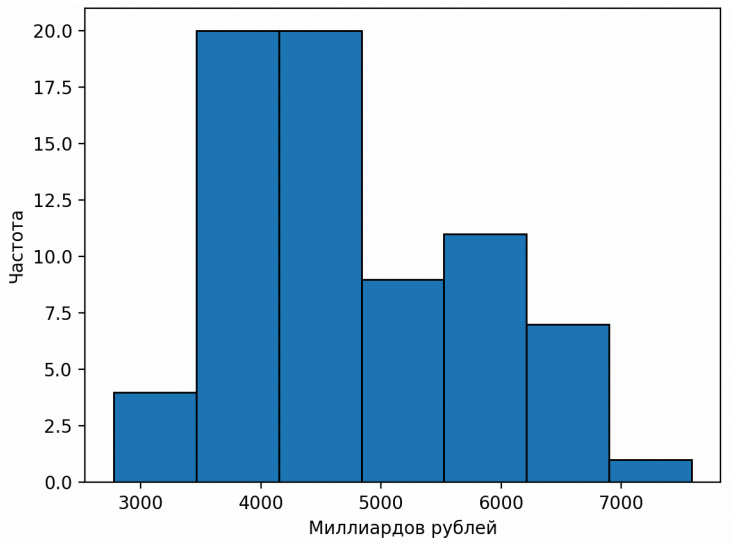

Рисунок 3 – Гистограмма потребительских расходов

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 1,2 раза больше критического, что говорит о том, что распределение не является нормальным по критерию Пирсона.

Таким образом, данные о сумме потребительских расходов не подчиняются нормальному распределению по критерию Пирсона и подчиняется по правилу трех сигм. Несоответствие нормальному распределению 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 4).

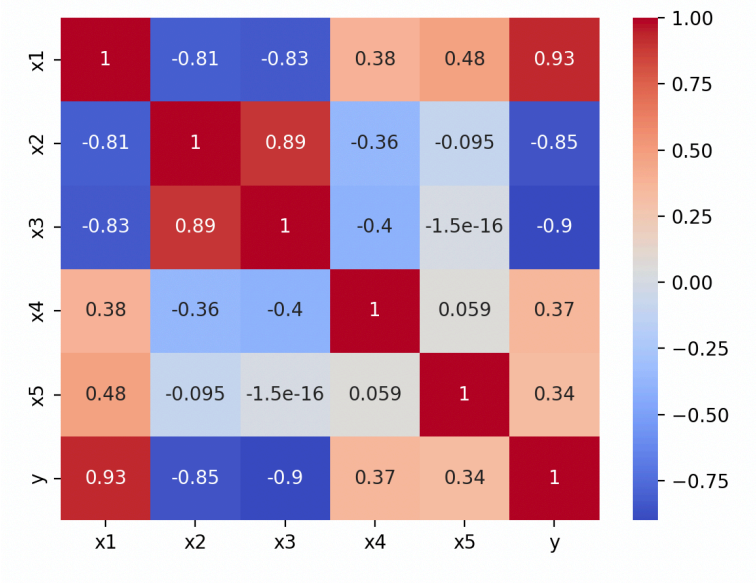

Рисунок 4 – Корреляционная матрица

Все 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 сумма потребительских расходов населения. Независимыми переменными являются 5 признаков: доходы населения, численность безработицы, численность населения, инфляция, номер месяца.

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:

$$Y = 150400 + \ 0{,}028 * X_1 + 0 * X_2 - 1000{,}68 * X_3 + 0 * X_4 + 52{,}5748 \ * X_5$$

3. ИССЛЕДОВАНИЕ МОДЕЛИ

3.1 Анализ статистической значимости уравнения регрессии

Проведем дисперсионный анализ с использованием критерия Фишера (F-критерия). Он помогает определить, есть ли статистически значимое влияние независимых переменных на зависимую переменную y. 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 (случайным шумом):

$$S_y = S_e$$

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 $s^2_{\text{факт}} > s_e^2$. 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:

$$F = \frac{s^2_{\text{факт}}}{s_e^2}$$

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 $F = 41{,}426,\ \alpha_F = 0$ при $F_{\text{кр}} = 2{,}852, \alpha = 0.05$, 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. $s_j = s\sqrt{[(C^TC)^{-1}]_{jj}}$, где $[(C^TC)^{-1}]_{jj}$ – j-й диагональный элемент матрицы $(C^TC)^{-1}$, $s = \sqrt{s_e^2}$.

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

$$\left(\beta_j - t_\gamma s_j;\ \beta_j + t_\gamma s_j\right),$$

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	1,42E+05	2,93E+04	4,844	0	8,29E+04	2,01E+05
1	x1	0,0266	0,012	2,153	0,037	0,002	0,052
2	x2	-0,0841	0,125	-0,673	0,504	-0,336	0,168
3	x3	-940,7479	198,232	-4,746	0	-1340,258	-541,238
4	x4	-3,7077	11,572	-0,32	0,75	-27,03	19,614
5	x5	51,8012	22,581	2,294	0,027	6,292	97,31

Критическое значение равно $T_{\text{кр}} = 2{,}12$. Работая с уровнем значимости $\alpha = 0.05$, заметны коэффициенты, которые являются статистически незначимыми, то есть они оказывают незначительно влияние на нашу модель. Видим, что коэффициенты для X_2 и для X_4 являются статистически не значимыми, т.к. значение их уровня значимости больше заданного, а также они не входят в доверительные интервалы и не входят в критическую область Т-критерия. 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. Матрица $(C^T\ C)$ может оставаться невырожденной, но 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 0.00096, что близко к нулю. Из анализа матрицы парных корреляций (рисунок 4) можно увидеть значительную корреляцию между признаками x_1, x_2 и x_3. Это свидетельствует о наличии мультиколлинеарности в модели. Рекомендуется исключить некоторые факторы для оптимизации модели.

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 $R^2_{adj} = 0{,}916$.

Шаги при удалении:

1. Удаление «x_4» со Р-значением 0,75 приводит к $R^2_{adj} = 0{,}921$
2. Удаление «x_2» со Р-значением 0,48 приводит к $R^2_{adj} = 0{,}924$
3. Удаление «x_5» со Р-значением 0,022 приводит к $R^2_{adj} = 0{,}919$

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

Шаги при добавлении:

1. Добавление «x_3» с приводит к $R^2_{adj} = 0{,}833$
2. Добавление «x_5» приводит к $R^2_{adj} = 0{,}919$
3. Добавление «x_1» приводит к $R^2_{adj} = 0{,}924$
4. Добавление «x_2» приводит к $R^2_{adj} = 0{,}921$

Таким образом, в двух случаях была нужно исключать факторы x_2 и x_4.

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 и Numpy, 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 matplotlib.pyplot as plt
from tabulate import tabulate
import os
import statsmodels.api as sm
from scipy.stats import f, chi2, t
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

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: lambda x: float(x.replace(',',
'.'))}) for column, path in data_paths.items()}
    df = pd.DataFrame(data)
    df['timestamp'] = pd.period_range(start='2019-01', end='2024-12', 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(len(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('Рублей')
    plt.ylabel('Частота')
    plt.title('Гистограмма потребительских расходов населения')
    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. Проверяет нормальность
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
    if sigma_68 > 68 and sigma_95 > 95 and sigma_99 > 99.7:
        is_norm_distribution = 'Распределение нормальное по правилу 3-х сигм'
    else:
        is_norm_distribution = 'Распределение не нормальное по правилу 3-х сигм'
    return sigma_68, sigma_95, sigma_99, is_norm_distribution
```

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):
        mean_value = mean_midpoints(intervals, frequencies)
        return np.sum([(((left + right) / 2) - mean_value) ** 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)])
    #  Количество оцениваемых параметров для нормального распределения описывается двумя
параметрами (среднее и стандартное отклонение)
    r = 2
    degrees_of_freedom = len(intervals) - r - 1
    alpha = 0.05
    chi_squared_critical = chi2.ppf(1 - alpha, degrees_of_freedom)
    if chi_squared_observed < chi_squared_critical:
        is_chi_norm_distribution = 'Нет оснований отвергать гипотезу H0 по критерию
Пирсона'
    else:
        is_chi_norm_distribution = 'Распределение не является нормальным по критерию
Пирсона'
    return chi_squared_observed, chi_squared_critical,  chi_squared_observed /
chi_squared_critical, is_chi_norm_distribution
```

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 - 1)
    return F_statistic, critical_value, p_value
```

Листинг 7. Критерий Фишера

Проверка значимости уравнения регрессии с помощью закона распределения Стьюдента.

```
def student_test(model, num_samples, num_features):
    """Определяет критическое значение критерия Стьюдента."""
    k = num_samples - num_features - 1
    alpha = 0.05
    critical_value = t.ppf(1 - alpha / 2, k)
    return critical_value
```

Листинг 8. Закон распределения Стьюдента

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 9. Функция вывода оценок

Основная часть программы.

```
def main():
    # Загрузка данных
    data_paths = {
        "x1": './punkt2/x1.txt',
        # "x2": './punkt2/x2.txt',
        "x3": './punkt2/x3.txt',
```

```
        # "x4": './punkt2/x4.txt',
        "x5": './punkt2/x5.txt',
        "y": './punkt2/y.txt'
        }
    df = load_data(data_paths)
    target_values = df['y'].values.astype(int)

    # Интервальное распределение
    print('Интервальное распределение')
    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"),
end='\n\n')

    # Построение гистограммы
    plot_histogram(target_values, num_intervals)

    # Нормальное распределение по теореме 3-х сигм
    print('Нормальное распределение по теореме 3-х сигм')
    sigma_68, sigma_95, sigma_99, is_norm_distribution =
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"))
    print(is_norm_distribution, end='\n\n')

    # Нормальное распределение по критерию Пирсона
    print('Нормальное распределение по критерию Пирсона')
    chi_squared_observed, chi_squared_critical, chi_squared_ratio,
is_chi_norm_distribution = 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"), end='\n')
    print(is_chi_norm_distribution, end='\n\n')

    # Анализ корреляций
    analyze_correlation(df)

    # Составляем матрицу признаков и вектор ответов
    X = df.drop('y', axis=1)
    y = df['y']

    # Разделение данных на тренировочные и тестовые
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=24)

    # Добавление константы к матрице признаков для МНК
    X_train_mn = sm.add_constant(X_train)
    X_test_mn = sm.add_constant(X_test)

    # Реализация МНК для тренировочных данных
    model = sm.OLS(y_train, X_train_mn).fit()

    # Получение прогнозов для тестовых данных
    y_predicted = model.predict(X_test_mn)

    # Выполнение теста Фишера
    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()

    # Вывод результатов регрессии
    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]']

    # Выполнение теста Стьюдента
    print('Анализ статистической значимости коэффициентов уравнения регрессии:')
    print(tabulate(coefficients_table, headers='keys', tablefmt='fancy_grid'))
    critical_value_s = student_test(model, len(y_test), X_test_mn.shape[1] - 1)
    print("Критическое значение T-критерия:", critical_value_s)
```

```
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 10. Основная функция.

4.3 Численное исследование результатов моделирования

Оценка модели была произведена по 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)	66234
Средняя абсолютная ошибка (MAE)	194,577
Коэффициент детерминации	0,935
Адаптивный коэффициент детерминации	0,924

Исходя из таблицы 2 коэффициент детерминации сильный, это свидетельствует о хорошем качестве предсказаний. MSE принимает высокое значение. MAE примерно равно 195, в контексте рассматриваемой задачи это значит, что предсказанное значение потенциальной суммы потребительских расходов может отличаться от истинного значения на $\pm$195.

Исходя из коэффициентов детерминации модель достаточно точная, однако значения MSE и MAE высоки. Средняя ошибка в 195 млрд руб. может быть значительной для прогноза. поэтому следует произвести улучшение модели, чтобы добиться приемлемого качества предсказания. Для этого можно добавить предыдущие значения расходов как дополнительные факторы или использовать более точные модели прогнозирования.

ВЫВОДЫ

В результате проведенного исследования были выявлены статистические связи между потребительскими расходами населения и различными экономическими факторами. Анализ данных позволил выделить ключевые переменные, оказывающие влияние на уровень потребительских расходов. При формализации и классификации этих переменных была проверена гипотеза о нормальности распределения исходных данных.

Корреляционный анализ подтвердил наличие значимых статистических взаимосвязей между переменными, а построение регрессионной модели позволило выделить 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 потребительских расходов населения является эффективным инструментом для предсказания будущих тенденций в экономике. Полученные результаты исследования могут стать основой для разработки стратегий и принятия обоснованных управленческих решений в области экономического планирования и социальной политики.

СПИСОК ЛИТЕРАТУРЫ

1. СберИндекс. Платформа для анализа потребительских расходов. URL: https://sberindex.ru/ru/dashboards/consumer-spending (дата обращения: 6 марта 2025).
2. Федеральная служба государственной статистики (Росстат). Официальный сайт. Раздел «Показатели по потребительским расходам». URL: https://rosstat.gov.ru/folder/13397 (дата обращения: 6 марта 2025).
3. Федеральная служба государственной статистики (Росстат). Официальный сайт. Раздел «Рабочая сила, занятость и безработица». URL: https://rosstat.gov.ru/labour_force (дата обращения: 6 марта 2025).
4. Федеральная служба государственной статистики (Росстат). Официальный сайт. Раздел «Индекс потребительских цен». URL: https://rosstat.gov.ru/folder/12781 (дата обращения: 6 марта 2025).
5. Уровень инфляции. Инфляция в России: таблицы. URL: https://уровень-инфляции.рф/таблицы-инфляции (дата обращения: 6 марта 2025).
6. Апалькова Т.Г., Ашихмина Е.А., Дормидонтова В.А. Применение языков программирования R и Python для проверки гипотезы о равенстве дисперсий по критерию Фишера в экономических исследованиях // Научный журнал "Управленческий учет". 2023. № 2. URL: https://journals.fa.ru (дата обращения: 6 марта 2025).
7. Математика и профиль. Критерий согласия. URL: http://mathprofi.ru/kriteriy_soglasiya.html (дата обращения: 6 марта 2025).
8. Чулков, Н. Г. "Методы статистического анализа данных" — М.: Экономика, 2010. — 304 с.

ПРОГРАММНЫЕ ПРИЛОЖЕНИЯ

Программной реализации алгоритма:

URL: …

ПРИЛОЖЕНИЕ А

Статистика потребительских расходов с 2019 по 2024 года.

	Потребительские расходы (млрд. руб.)					
месяц\год	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
январь	3 470,37	3 670,85	3 813,76	4 452,33	4 828,12	5545,74
февраль	3 411,94	3 652,71	3 796,68	4 488,85	4 804,22	5675,06
март	3 689,42	3 937,55	4 159,84	5 016,01	5 196,07	6073,22
апрель	3 664,45	2 774,98	4 213,02	4 615,77	5 182,53	6015,49
май	3 708,02	2 927,01	4 250,64	4 639,64	5 297,55	6135,44
июнь	3 764,55	3 389,56	4 305,72	4 678,49	5 411,14	6227,46
июль	3 850,89	3 752,63	4 409,06	4 783,55	5 605,84	6483,1
август	3 954,66	3 941,49	4 584,20	4 926,37	5 835,10	6672,9
сентябрь	3 909,61	3 935,07	4 586,14	4 824,98	5 733,08	6577,3
октябрь	3 962,60	4 010,43	4 648,62	4 878,46	5 842,14	6644,2
ноябрь	3 989,86	3 976,76	4 585,23	4 904,31	5 767,85	6604,4
декабрь	4 608,66	4 646,24	5 421,44	5 637,07	6 638,83	7585,6
итого	45 985,03	44 615,28	52 774,35	57 845,83	66 142,47	76 239,91

ПРИЛОЖЕНИЕ Б

Статистические характеристики данных о потребительских расходах населения.

Исходные данные

Год	Потребительские расходы населения, млрд. руб.
2019	45985
2020	44615
2021	52774
2022	57846
2023	66142
2024	76239

Базисный анализ

Год	Потребительские расходы населения, млрд. руб.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2019	45985			
2020	44615	-1370	97	-3
2021	52774	6789	114,8	14,8
2022	57846	11861	125,8	25,8
2023	66142	20157	143,8	43,8
2024	76239	30254	165,8	65,8

Цепной анализ

Год	Потребительские расходы населения, млрд. руб.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2019	45985			
2020	44615	-1370	97	-3
2021	52774	8159	118,3	18,3
2022	57846	5072	109,6	9,6
2023	66142	8296	114,3	14,3
2024	76239	10097	115,3	15,3

Средний уровень ряда: 53473 (млрд. руб.)

Средний абсолютный прирост: 5040 (млрд. руб.)

Средний темп роста: 109.5%

Средний темп прироста: 9.5%

Прогноз по среднему абсолютному приросту: 71182 (млрд. руб.)

Прогноз по среднему темпу роста: 72435 (млрд. руб.)

Прогноз по МНК: 69536 (млрд. руб.)

Относительная погрешность по среднему абсолют. приросту: 6.63%

Относительная погрешность по среднему темпу роста: 4.99%

Относительная погрешность по МНК: 8.79%

Уравнение линейного тренда: y = 5354.5 * t + 37408.9

ПРИЛОЖЕНИЕ В

Первичные статистические данные о факторах.

Дата	x1	x2	x3	x4	x5
31.01.2019	30075	3689	147,8	5	1
28.02.2019	30075	3693	147,8	5,24	2
31.03.2019	30075	3634	147,8	5,27	3
30.04.2019	34424	3596	147,8	5,17	4
31.05.2019	34424	3512	147,8	5,13	5
30.06.2019	34424	3453	147,8	4,66	6
31.07.2019	35009	3395	147,8	4,59	7
31.08.2019	35009	3343	147,8	4,33	8
30.09.2019	35009	3353	147,8	3,99	9
31.10.2019	41481	3384	147,8	3,77	10
30.11.2019	41481	3475	147,8	3,54	11
31.12.2019	41481	3512	147,8	3,05	12
31.01.2020	31810	3514	147,9	2,42	1
29.02.2020	31810	3484	147,9	2,31	2
31.03.2020	31810	3485	147,9	2,55	3
30.04.2020	33202	3756	147,9	3,1	4
31.05.2020	33202	4121	147,9	3,03	5
30.06.2020	33202	4500	147,9	3,21	6
31.07.2020	35123	4652	147,9	3,37	7
31.08.2020	35123	4749	147,9	3,57	8
30.09.2020	35123	4807	147,9	3,67	9
31.10.2020	43355	4796	147,9	3,98	10
30.11.2020	43355	4730	147,9	4,42	11
31.12.2020	43355	4610	147,9	4,91	12
31.01.2021	32694	4487	147,4	5,19	1
28.02.2021	32694	4374	147,4	5,67	2
31.03.2021	32694	4253	147,4	5,78	3
30.04.2021	38308	4111	147,4	5,52	4
31.05.2021	38308	3917	147,4	6,01	5
30.06.2021	38308	3762	147,4	6,51	6
31.07.2021	40721	3606	147,4	6,47	7

31.08.2021	40721	3494	147,4	6,69	8
30.09.2021	40721	3389	147,4	7,41	9
31.10.2021	47754	3334	147,4	8,14	10
30.11.2021	47754	3296	147,4	8,4	11
31.12.2021	47754	3282	147,4	8,39	12
31.01.2022	39950	3288	147	8,74	1
28.02.2022	39950	3231	147	9,16	2
31.03.2022	39950	3178	147	16,7	3
30.04.2022	46141	3081	147	17,83	4
31.05.2022	46141	3043	147	17,11	5
30.06.2022	46141	3003	147	15,9	6
31.07.2022	46165	2972	147	15,09	7
31.08.2022	46165	2944	147	14,3	8
30.09.2022	46165	2926	147	13,67	9
31.10.2022	56942	2925	147	12,63	10
30.11.2022	56942	2887	147	11,97	11
31.12.2022	56942	2847	147	11,92	12
31.01.2023	45397	2779	146,4	11,76	1
28.02.2023	45397	2728	146,4	10,97	2
31.03.2023	45397	2659	146,4	3,51	3
30.04.2023	49810	2570	146,4	2,3	4
31.05.2023	49810	2494	146,4	2,5	5
30.06.2023	49810	2408	146,4	3,24	6
31.07.2023	51665	2344	146,4	4,3	7
31.08.2023	51665	2298	146,4	5,13	8
30.09.2023	51665	2286	146,4	6	9
31.10.2023	65470	2269	146,4	6,68	10
30.11.2023	65470	2257	146,4	7,47	11
31.12.2023	65470	2249	146,4	7,42	12
31.01.2024	52399	2224	146,1	7,44	1
29.02.2024	52399	2183	146,1	7,67	2
31.03.2024	52399	2095	146,1	7,69	3
30.04.2024	58791	2032	146,1	7,82	4
31.05.2024	58791	1997	146,1	8,29	5

30.06.2024	58791	1942	146,1	8,58	6
31.07.2024	63463	1910	146,1	9,13	7
31.08.2024	63463	1855	146,1	9,04	8
30.09.2024	63463	1843	146,1	8,62	9
31.10.2024	77679	1806	146,1	8,53	10
30.11.2024	77679	1791	146,1	8,88	11
31.12.2024	77679	1779	146,1	9,51	12