**Министерство науки и высшего образования Российской Федерации**
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»**

# Отчёт

по лабораторной работе №2 «Экстраполяция»

по дисциплине «**Математические модели исторических процессов**»

Автор: Малаев Степан Геннадьевич
Факультет инфокоммуникационных технологий
Группа: K33422
Преподаватель: Екатерина Ивановна

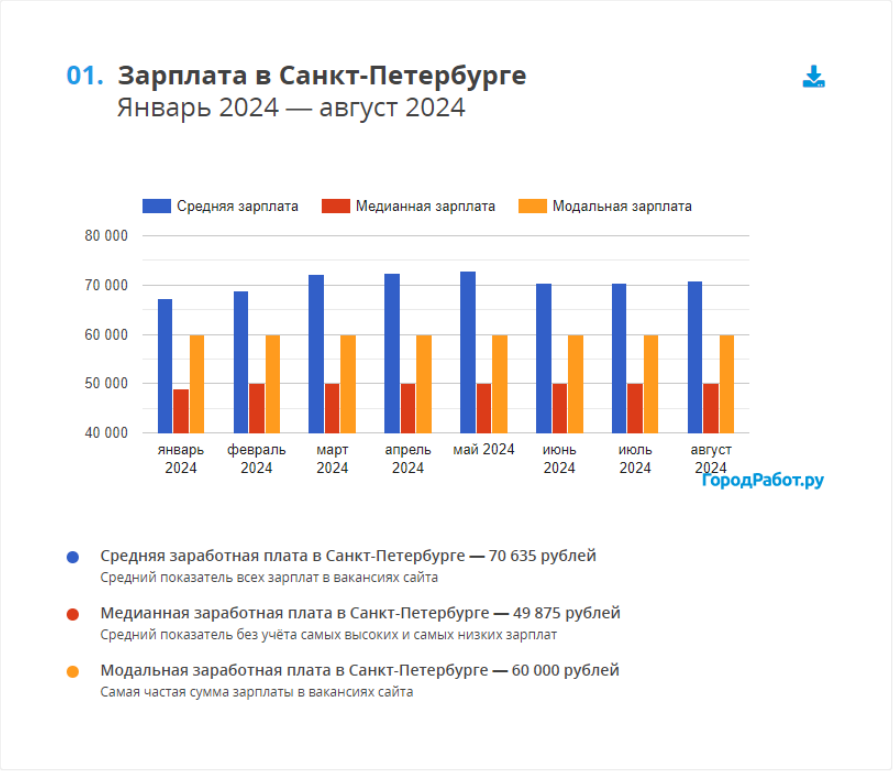
Санкт-Петербург
2024

## Выбор данных

Были собраны данные о зарплатах с сайта "ГородРабот.ру". Для анализа данных о зарплатах с данного ресурса был реализован процесс автоматизированного сбора информации посредством веб-скрейпинга. В рамках этого процесса был использован метод ротации бесплатных прокси, что позволило значительно оптимизировать выполнение запросов к сайту — более 3000 запросов были обработаны всего за 10 минут. Основная часть времени ушла на разработку сложного алгоритма, который включал в себя поиск надежных прокси-хабов, их парсинг, а также ротацию прокси для обеспечения стабильности и эффективности сбора данных. Итогом работы стала детализированная статистика, охватывающая данные о средней, модальной и медианной зарплатах по различным городам России за каждый месяц с 2017 года.

Однако, данные за 2024 год оказались не репрезентативными, поскольку информация для всех городов была одинаковой. Это вынудило меня исключить данные за 2024 год из анализа, так как их использование могло бы исказить результаты.

Ссылка на ресурс: "ГородРабот.ру"



## Обзор данных

Данные, собранные для анализа, включают следующие ключевые параметры:

1. **Город** — название населенного пункта, для которого собирается информация о зарплатах.
2. **Год** — календарный год, в который был произведен сбор данных.
3. **Месяц** — конкретный месяц в рамках года, для которого рассчитаны показатели.
4. **Мода зарплат** — наиболее часто встречающаяся зарплата среди всех проанализированных данных за данный период в конкретном городе.
5. **Средняя зарплата** — среднее арифметическое значение всех зарплат, полученных для данного периода и города.
6. **Медианная зарплата** — зарплата, находящаяся посередине распределения: 50% зарплат ниже данного значения и 50% выше.

Период для анализа определяется как связка "год + месяц", что позволяет отслеживать динамику зарплат по каждому месяцу за весь анализируемый промежуток времени с 2017 года. Данные структурированы таким образом, чтобы отражать месячные изменения зарплат по различным регионам России, обеспечивая всесторонний анализ средней, модальной и медианной зарплат в разрезе времени и географии.

Для уменьшения данных я буду рассматривать только среднее значение зарплаты.

| | year | month | region | mean | predicted | period |
|---|---|---|---|---|---|---|
| 3 | 2023 | февраль | омск | 51550.0 | False | 74 |
| 4 | 2023 | апрель | воронеж | 54240.0 | False | 76 |
| 6 | 2023 | апрель | омск | 55182.0 | False | 76 |
| 7 | 2022 | февраль | воронеж | 45625.0 | False | 62 |
| 9 | 2022 | февраль | омск | 44992.0 | False | 62 |
| 14 | 2023 | февраль | пермь | 52377.0 | False | 74 |
| 15 | 2023 | февраль | воронеж | 50355.0 | False | 74 |
| 18 | 2022 | февраль | волгоград | 44538.0 | False | 62 |
| 21 | 2022 | февраль | пермь | 44579.0 | False | 62 |
| 22 | 2023 | февраль | донецк | 64066.0 | False | 74 |

## Предобработка данных

Для предобработки данных были выполнены следующие шаги:

1. **Удаление ненужных колонок** — все незначимые для анализа и прогнозирования поля были исключены из набора данных для оптимизации процесса обработки и минимизации шума.
2. **Устранение пропусков** — пропущенные значения в ключевых параметрах, таких как зарплаты или временные метки, были корректно обработаны либо удалены, чтобы избежать искажения результатов при дальнейших расчетах.
3. **Создание целочисленных периодов** — для упрощения работы с временными рядами был создан новый параметр, представляющий собой целочисленный период, объединяющий год и месяц.

## Базисные показатели

Базисные показатели являются важными аналитическими инструментами в экономике и статистике для измерения изменений какого-либо показателя (например, зарплаты, прибыли и т.д.) по сравнению с базисным периодом (обычно самым первым периодом). Они позволяют оценить, насколько изменились данные относительно начальной точки отсчета.

**Базисный прирост**:
Позволяет отследить долгосрочные тенденции и изменения относительно первого периода, что помогает оценить общий рост или спад.

**Цепной прирост**:
Используется для измерения изменений между смежными периодами. Он особенно полезен, если нужно оценить краткосрочные изменения.

**Базисный темп роста**:
Применяется для вычисления отношения текущего значения к базисному периоду. Удобен, когда необходимо понять, на сколько процентов или во сколько раз увеличилось значение относительно базиса.

**Цепной темп роста**:
Применяется для расчета изменений между соседними периодами, как и цепной прирост, но выражается в процентном соотношении.

Для их вычисления был использован следующий алгоритм:

```
dataframes = {}

for region in regions:
    region_df = df[df["region"] == region]
    region_df = region_df.sort_values(by='period')

    region_df['base_growth'] = region_df['mean'] - region_df['mean'].iloc[0]
    region_df['chain_growth'] = region_df['mean'].diff()
    region_df['base_growth_rate'] = region_df['mean'] / region_df['mean'].iloc[0]
    region_df['chain_growth_rate'] = region_df['mean'] / region_df['mean'].shift(1)

    dataframes[region] = region_df
Executed at 2024.09.18 03:25:33 in 53ms
```

Пример результатов города Красноярск для первых 20 периодов

| Год | Месяц | Город | Зарплата | Базисный прирост | Цепной прирост | Базисный темп роста | Цепной темп роста |
|---|---|---|---|---|---|---|---|
| 2017 | январь | красноярск | 33522.0 | 0.0 | NaN | 1.000000 | NaN |
| 2017 | февраль | красноярск | 32532.0 | -990.0 | -990.0 | 0.970467 | 0.970467 |
| 2017 | март | красноярск | 33181.0 | -341.0 | 649.0 | 0.989828 | 1.019950 |
| 2017 | апрель | красноярск | 33424.0 | -98.0 | 243.0 | 0.997077 | 1.007323 |
| 2017 | май | красноярск | 33443.0 | -79.0 | 19.0 | 0.997643 | 1.000568 |
| 2017 | июнь | красноярск | 34070.0 | 548.0 | 627.0 | 1.016347 | 1.018748 |
| 2017 | июль | красноярск | 34300.0 | 778.0 | 230.0 | 1.023209 | 1.006751 |
| 2017 | август | красноярск | 34677.0 | 1155.0 | 377.0 | 1.034455 | 1.010991 |
| 2017 | сентябрь | красноярск | 36350.0 | 2828.0 | 1673.0 | 1.084363 | 1.048245 |
| 2017 | октябрь | красноярск | 36802.0 | 3280.0 | 452.0 | 1.097846 | 1.012435 |
| 2017 | ноябрь | красноярск | 36950.0 | 3428.0 | 148.0 | 1.102261 | 1.004022 |
| 2017 | декабрь | красноярск | 36452.0 | 2930.0 | -498.0 | 1.087405 | 0.986522 |
| 2018 | январь | красноярск | 36756.0 | 3234.0 | 304.0 | 1.096474 | 1.008340 |
| 2018 | февраль | красноярск | 38693.0 | 5171.0 | 1937.0 | 1.154257 | 1.052699 |
| 2018 | март | красноярск | 38036.0 | 4514.0 | -657.0 | 1.134658 | 0.983020 |
| 2018 | апрель | красноярск | 37104.0 | 3582.0 | -932.0 | 1.106855 | 0.975497 |
| 2018 | май | красноярск | 38668.0 | 5146.0 | 1564.0 | 1.153511 | 1.042152 |
| 2018 | июнь | красноярск | 35641.0 | 2119.0 | -3027.0 | 1.063212 | 0.921718 |
| 2018 | июль | красноярск | 36742.0 | 3220.0 | 1101.0 | 1.096056 | 1.030891 |
| 2018 | август | красноярск | 38071.0 | 4549.0 | 1329.0 | 1.135702 | 1.036171 |

Выводы:
В целом, зарплаты в Красноярске демонстрируют положительную динамику по сравнению с базисным периодом. Хотя в начале 2017 года наблюдаются небольшие спады, к концу периода (август 2018 года) зарплаты стабильно растут на 13-14% по сравнению с базисом.

Цепной прирост показывает, что зарплаты подвержены краткосрочным колебаниям. Особенно в 2018 году видно, что зарплаты могут как увеличиваться на несколько тысяч рублей за месяц, так и снижаться.

Можно предположить, что на зарплаты в Красноярске влияют сезонные и экономические факторы. Например, значительный рост зарплат в феврале 2018 года может быть связан с экономической активностью после зимних праздников.

## Линейная регрессия

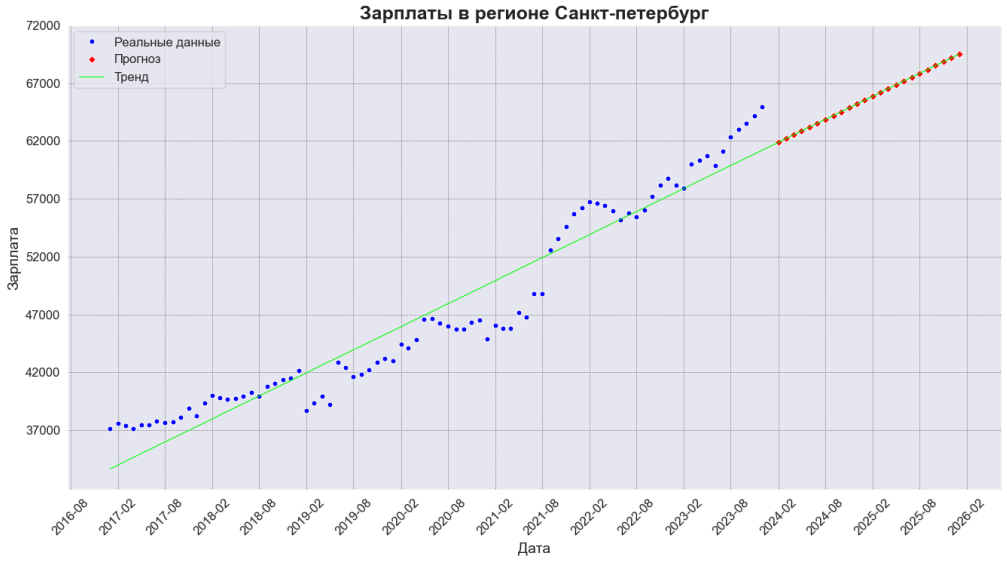
Линейная регрессия — это метод анализа данных, который используется для определения зависимости между двумя переменными. Он строит прямую линию, которая наилучшим образом описывает связь между независимой переменной (x) и зависимой переменной (y). Эта линия минимизирует сумму квадратов отклонений точек от линии (ошибки прогноза).

Линия тренда — это прямая, которая отображает общую тенденцию изменения данных на графике. В контексте линейной регрессии, линия тренда показывает, как изменяется зависимая переменная (например, зарплата) в зависимости от независимой переменной (например, время).

Простой пример: если данные показывают, что зарплаты растут с течением времени, линия тренда поможет прогнозировать, какими будут зарплаты в будущем.

Были построены линии тренда на основе исторических данных зарплат в каждом городе, что позволило выявить общую тенденцию их изменения.

Для примера рассмотрим Санкт-Петербург.




Выводы:

1. **Реальные данные** (синие точки):

Зарплаты в Санкт-Петербурге с 2016 по 2023 год демонстрируют восходящую тенденцию с периодическими колебаниями.

В среднем, зарплаты постепенно увеличиваются, что говорит о росте благосостояния или индексации доходов.

2. **Линия тренда** (зеленая линия):

Линия тренда демонстрирует устойчивый рост зарплат с небольшим увеличением наклона. Зарплаты в Санкт-Петербурге с течением времени имеют тенденцию к увеличению.

Наклон линии тренда предполагает, что средний уровень роста зарплат сохраняется на постоянном уровне, несмотря на небольшие колебания.

3. **Прогнозируемые данные** (красные точки):

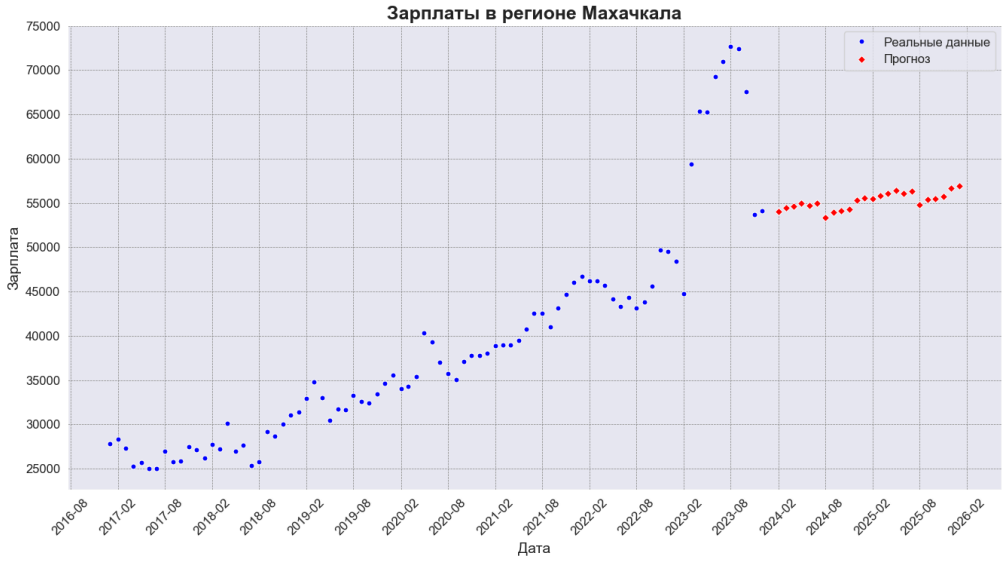
Прогноз на два года вперед (с 2023 по 2025 годы) показывает дальнейший рост зарплат, если текущий тренд сохранится.

## Экспоненциальное сглаживание

Был выбран такой метод прогнозирования как: Экспоненциальное сглаживание. Это - метод прогнозирования временных рядов, который сглаживает данные, чтобы выявить тренды и сделать более точные прогнозы. Его основная задача — учесть прошлые данные, придавая им уменьшающийся вес по мере удаления от текущего момента.

Были построены графики для всех городов.

Рассмотрим Махачкалу.




Выводы:

1. **Реальные данные** (синие точки):

Зарплаты в Махачкале с 2016 года демонстрируют положительный тренд, начиная с уровня около 25 000 рублей и постепенно увеличиваясь до 60 000 рублей к 2022 году.

В 2018 и 2020 годах видны периоды стабильности или небольших колебаний, однако общий тренд все равно направлен вверх.

В 2021–2022 годах наблюдаются значительные колебания зарплат, вероятно, связанные с экономическими факторами.

2. **Прогнозируемые данные** (красные точки):
Прогноз показывает, что в ближайшие два года зарплаты стабилизируются на уровне около 62 000 - 65 000 рублей.
Прогнозируемые данные демонстрируют небольшие колебания, но не предполагается значительного роста или падения.

Графики для всех городов можно найти в открытом репозитории проекта

## Вывод

В ходе данной лабораторной работы был проведен анализ динамики экономических показателей с использованием методов экстраполяции и трендового анализа.

Для каждого динамического ряда была построена линия тренда, которая позволяет оценить общие тенденции изменения показателя во времени.

Метод линейной регрессии использовался для вычисления коэффициентов тренда, что позволило спрогнозировать будущие значения.

Были рассчитаны такие ключевые метрики, как базисный прирост, цепной прирост, базисный и цепной темпы роста. Эти показатели позволили проанализировать, насколько изменялся показатель в каждом периоде относительно базисного и предыдущего периодов.

Используя построенные уравнения тренда, были спрогнозированы значения на 2–3 периода вперед. Прогноз позволил оценить, как будут изменяться показатели в будущем при сохранении текущих тенденций.