Лабораторная работа №1: Описательная статистика, визуализация, предварительная обработка данных

**Часть 1**

Найдено:

* Данные с сайта РосСтат, распределение которых близко к нормальному: "Распределение населения по возрастным категориям".
* Данные с сайта РосСтат, распределение которых близко к равномерному: "Число организаций, выполнявших научные исследования и разработки, по секторам деятельности по Российской Федерации".

Изучено:

* Распределение с медианой, смещенной относительно среднего на ~15% размаха: "Индексы цен производителей по товарам и товарным группам до 2010 г (процент)".

Реализовано:

* Посчитаны описательные характеристики выбранных данных:
  + Среднее значение
  + Мода
  + Медиана
* Визуальное представление данных:
  + Построен график для нормального распределения, демонстрирующая - форму кривой Гаусса (приблизительно).

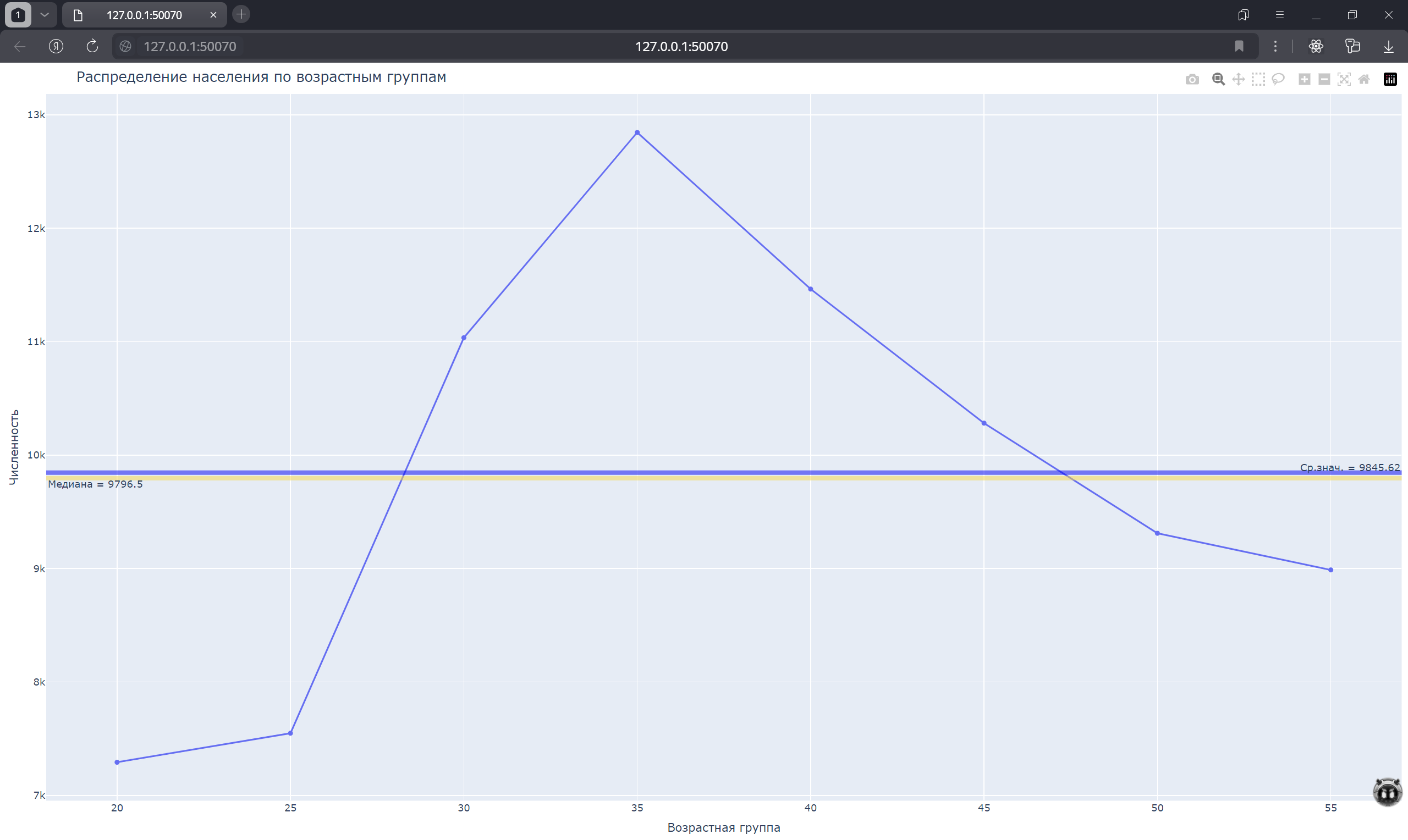


Рис.1 – График распределения, приближенного к нормальному

* + Построен график для равномерного распределения, показывающий равномерное распределение точек.

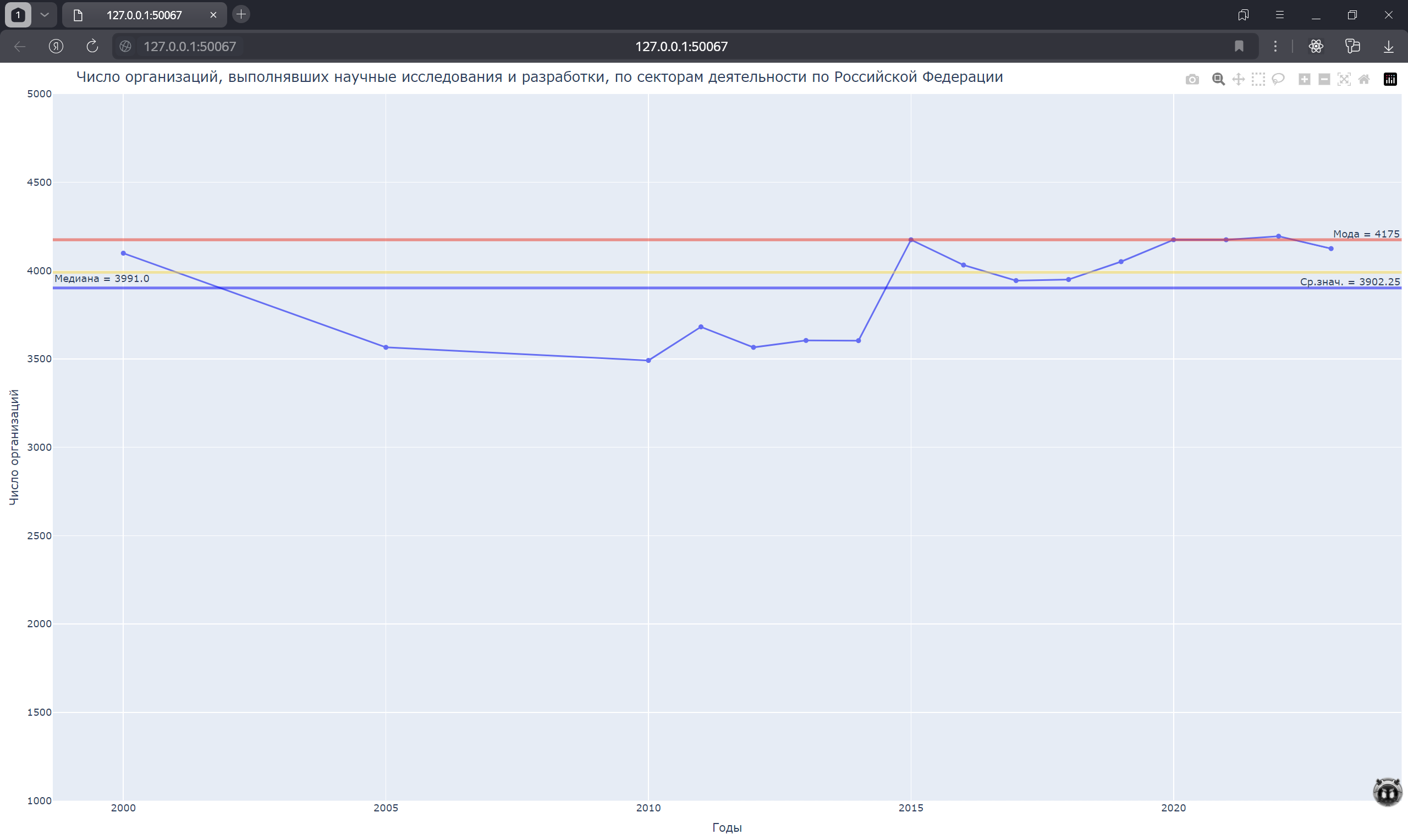


Рис.2 – График распределения, приближенного к равномерному

* + Построен график для распределения со смещенной относительно среднего на ~15% медианой

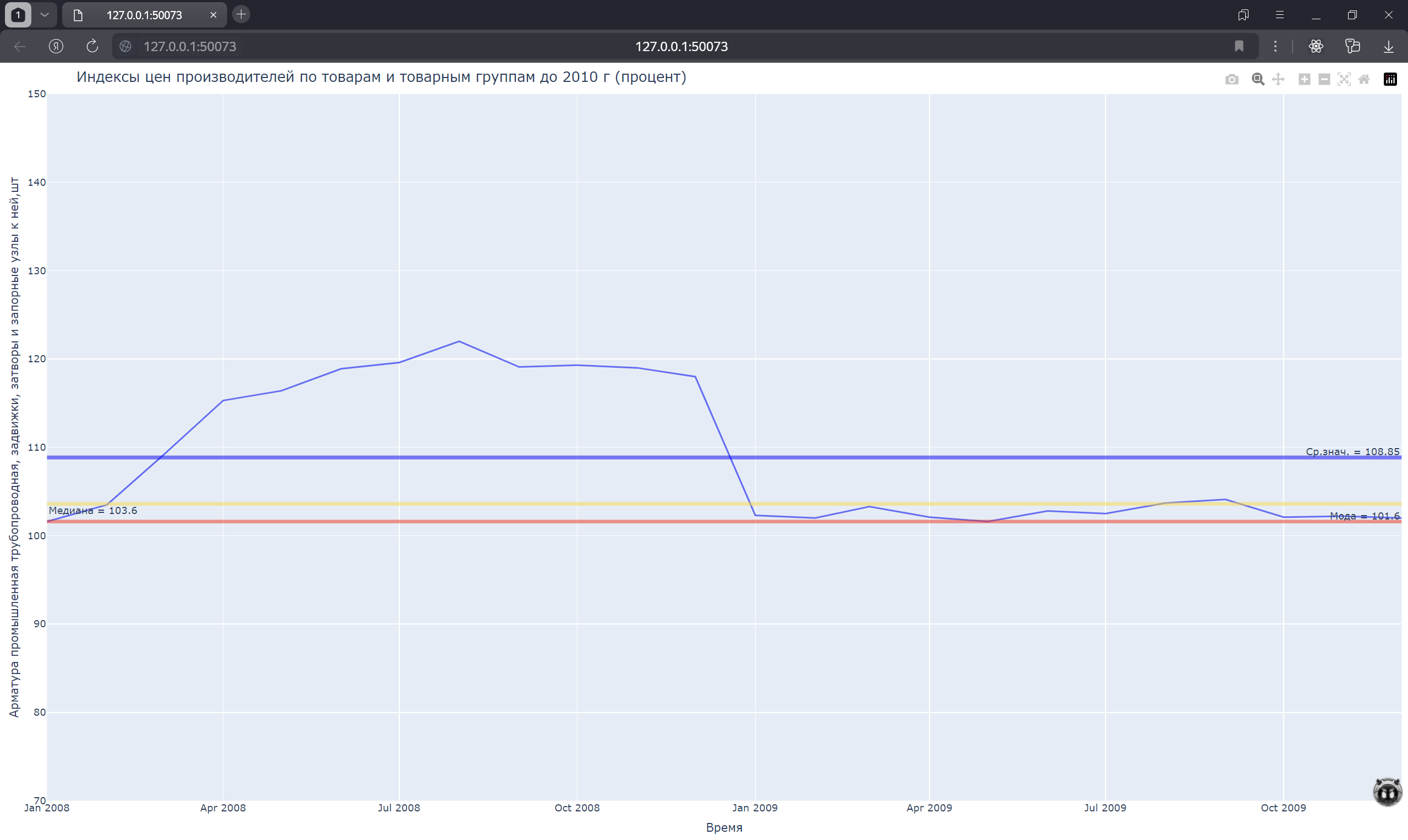


Рис.3 – График распределения со смещенной медианой

**Часть 2**

Найдено:

* Датасет с аномальными значениями: "Среднемесячная номинальная начисленная заработная плата работников по полному кругу организаций в целом по экономике по субъектам Российской Федерации с 2018 года, рублей".

Изучено и реализовано:

* Построена диаграмма размаха «ящик с усами», выявившая аномальные значения.

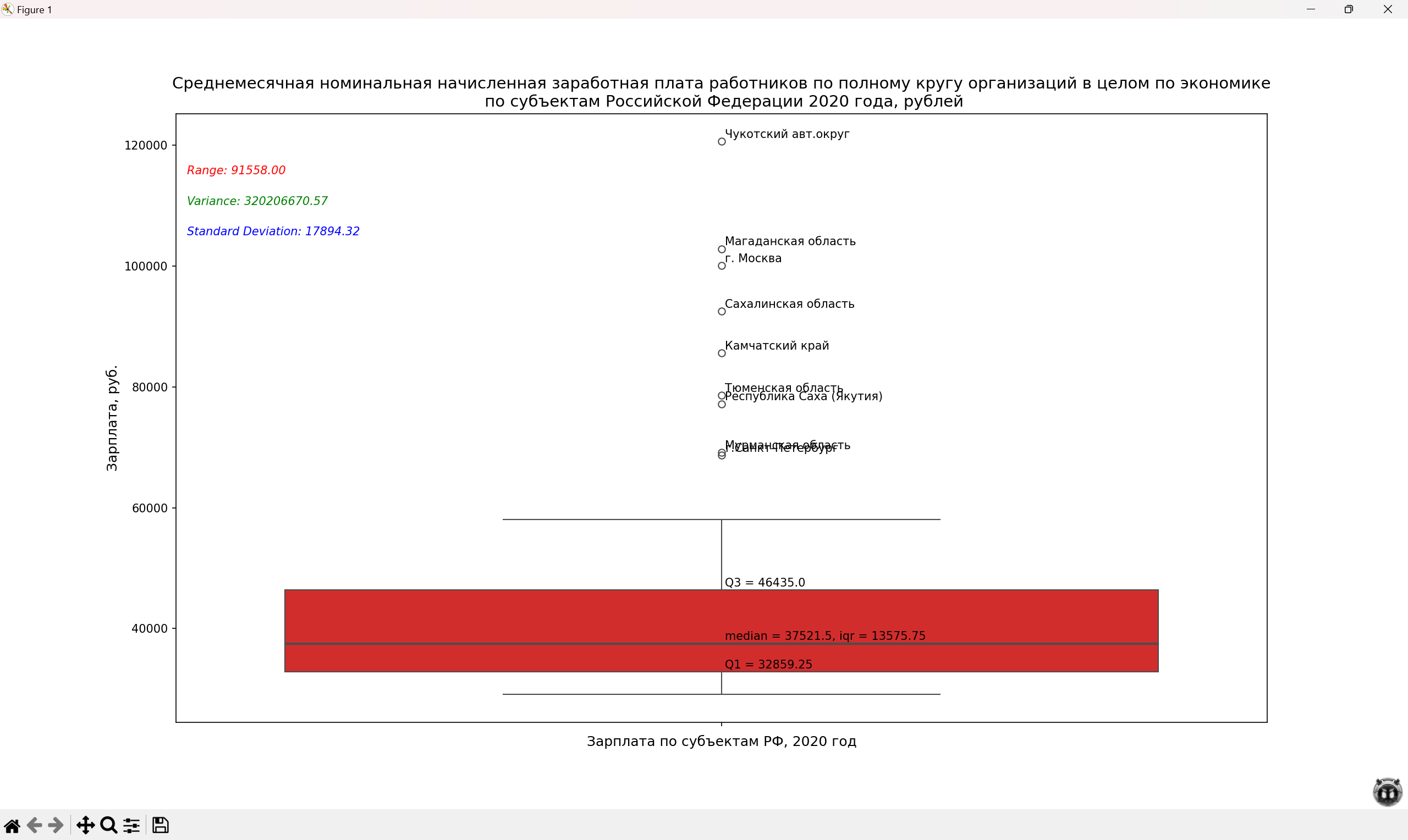


Рис.4 – Диаграмма «ящик с усами»

* При построении диаграммы были просчитаны:
  + Квартили
  + Межквартильный интервал
* Посчитаны описательные характеристики выбранных данных:
  + Размах
  + Дисперсия
  + Стандартное отклонение

Гипотеза о природе аномальных значений:  
Высокая средняя зарплата в Чукотском автономном округе и Магаданской области может быть обусловлена географической удаленностью и суровым климатом, требующими компенсации для привлечения работников, наличием ресурсной экономики с высокими зарплатами в добывающей промышленности, низкой плотностью населения, создающей конкуренцию за рабочую силу, и возможными государственными субсидиями, поддерживающими доходы в этих регионах.

Заключение: в ходе лабораторной работы были успешно найдены и проанализированы различные типы распределений данных, а также выявлены аномалии, что позволяет сделать выводы о возможных причинах их возникновения и направлениях для дальнейшего исследования.