Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №8**

**Дисциплина: Обработка больших данных**

**Тема: «Корреляционный и регрессионный анализ данных»**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.Д. Воробьев

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т.А.Приходько

**Цель:** ознакомиться с понятием корреляционный и регрессионный анализ данных, некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться оценивать связь между переменными и оценивать степень этой связи.

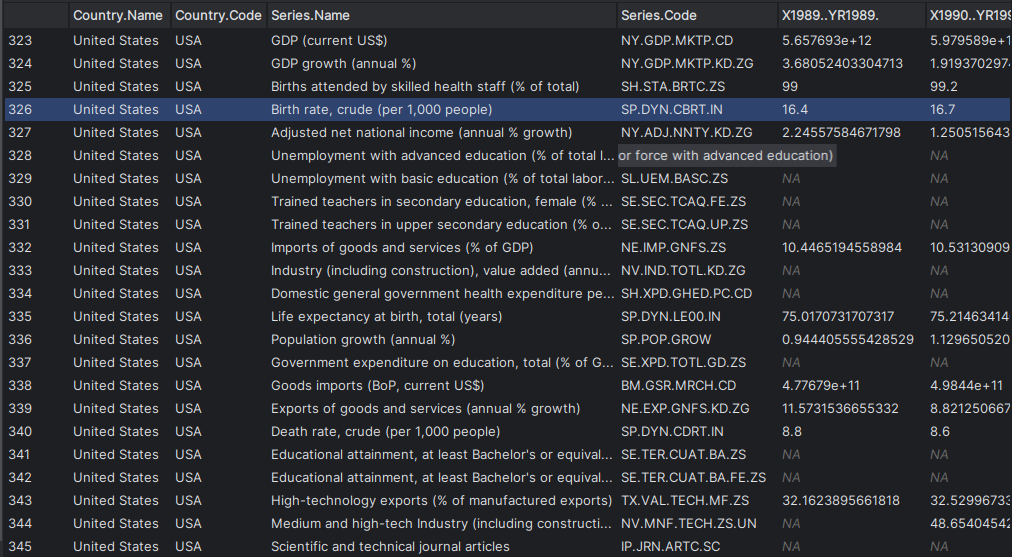
**Ход работы:**

Из исходного csv-файла были импортированы данные в Intellij IDEA. Таблица представлена на рисунке 1.



Рисунок 1 – Исходная таблица с данными

В данном наборе данных представлены данные о двадцати двух характеристиках восемнадцати стран мира с 1989 года по 2018 год. Удалим все страны, кроме страны согласно варианту (США), а также повторяющиеся строчки. На рисунке 2 изображена таблица получившихся данных.

РРисунок 2 – Измененная таблица с данными

Согласно первому заданию, построим кривую прироста ВВП США в процентах и USD. Соответствующие графики изображены на рисунках 3,4.

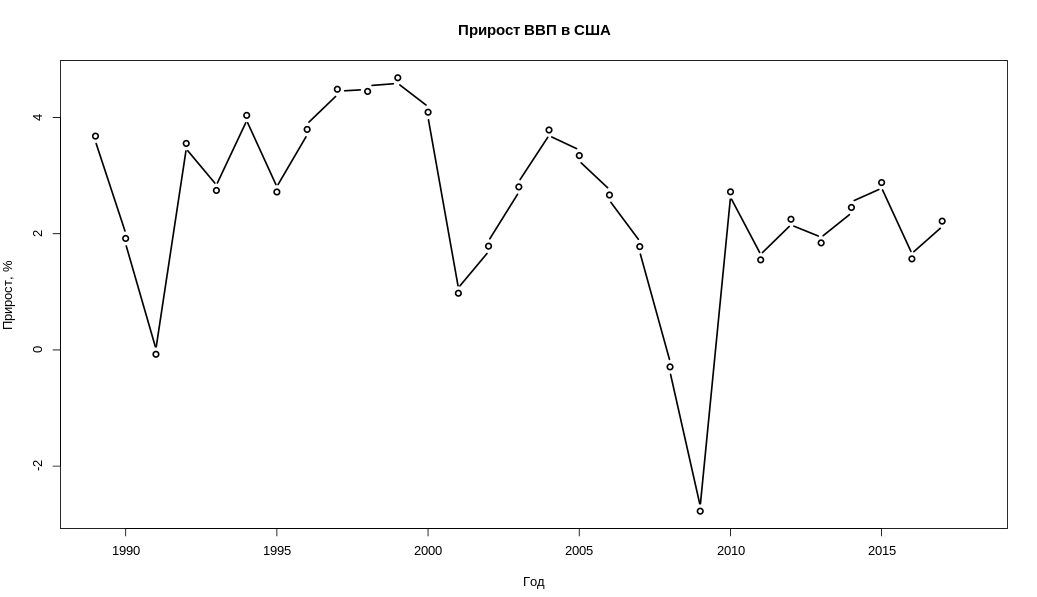


Рисунок 3 – Рост ВВП США на протяжении 30 лет в процентах относительно прошлого года

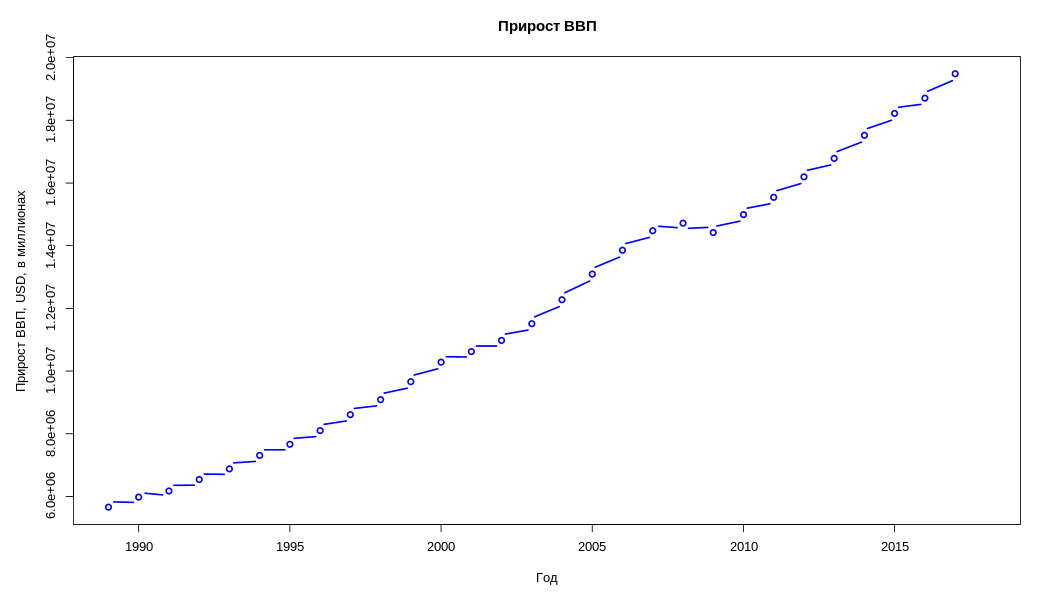


Рисунок 4 – Рост ВВП США на протяжении 30 лет в USD

Далее приступим к поиску корреляции заданных характеристик.

Сперва проанализируем рост ВВП и прирост населения. Выполним корреляцию двумя методами: Спирмена и Пирсона. Результаты работы методов отображены на рисунке 5.

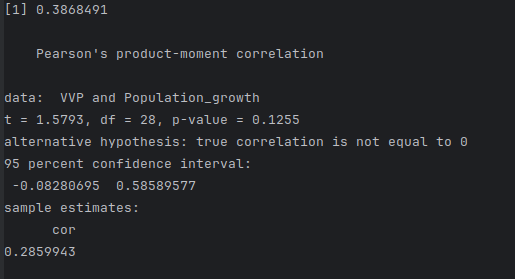


Рисунок 5 – Корреляция роста ВВП и прироста населения США

По полученным значениям видно, что корреляция между ростом ВВП и приростом населения не очень сильная. Изобразим визуально результаты исследования корреляции между характеристиками. Значения коэффициентов корреляции рисуются в виде эллипсов, отражающих форму плотности двумерного нормального распределения с данным значением корреляции между компонентами. Чем ближе значение коэффициента корреляции к +1 или -1 — тем более вытянутым становится эллипс. Наклон эллипса отражает знак. На рисунке 6 представлено представление в виде эллипсов для данных характеристик.

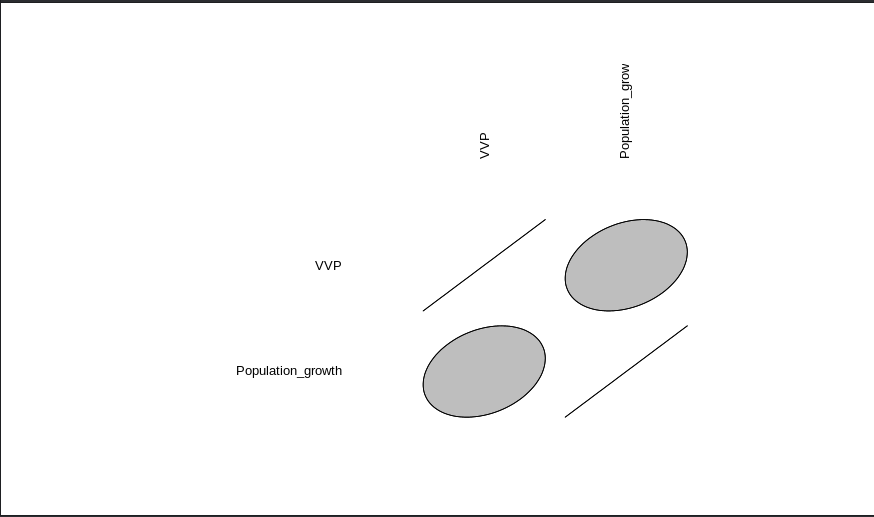


Рисунок 6 – Визуальное представление результатов исследования корреляции

Далее рассмотрим корреляцию между изменением расходов на медицину и продолжительностью жизни. На рисунке 7 отражена корреляция по аналогичным методам.

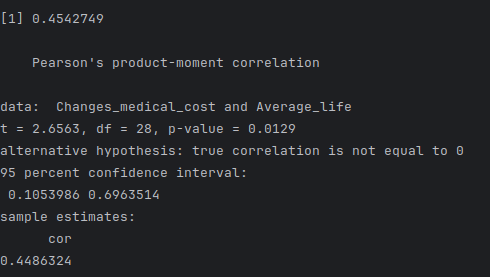


Рисунок 7 – Корреляция между изменением расходов на медицину и увеличением смертности

В это случае корреляция между характеристиками сильнее.

Затем рассмотрим между изменением расходов на медицину и смертностью. На рисунке 8 представлена корреляция для этих характеристик.

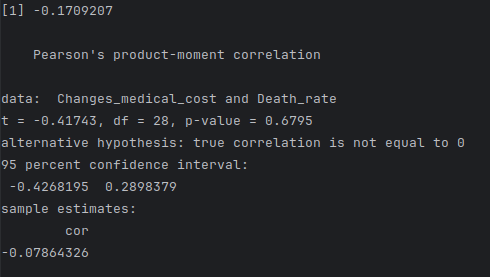


Рисунок 8 – Корреляция между изменением расходов на медицину и смертностью

Коэффициент корреляции между изменением расходов на медицину и смертностью очень слаба. Вновь изобразим результаты исследования корреляции визуально. В этот раз между тремя характеристиками. Рисунок 9 отображает эту корреляцию.

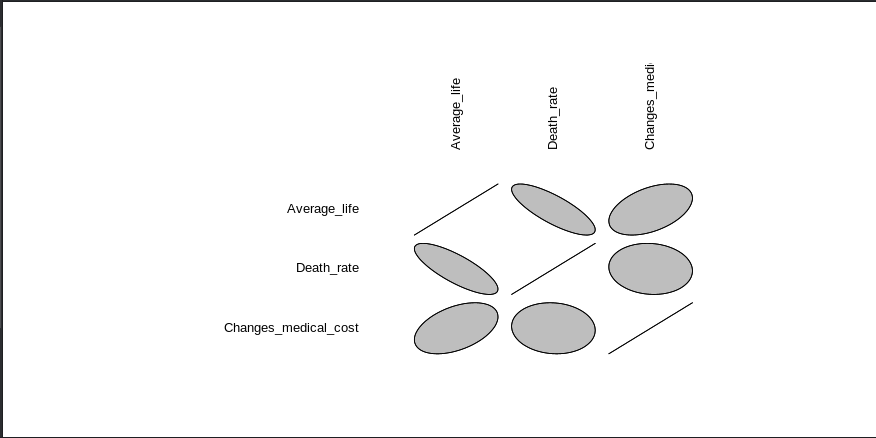


Рисунок 9 – Визуальное представление результатов исследования корреляции

В следующем задании необходимо было с помощью регрессионного анализа выявить зависимые и независимые переменные. Результат показан на рисунке 10.

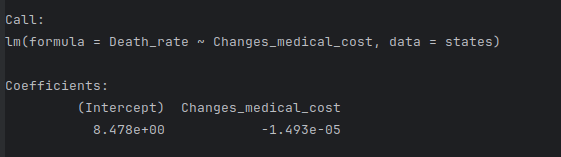


Рисунок 10 – Результат регрессионного анализа

В качестве независимой переменной были взяты расходы на медицину, а в качестве зависимой – значение смертности.

С помощью функции predict() было сделано предсказание по атрибуту значение смертности. Результат представлен на рисунке 11.

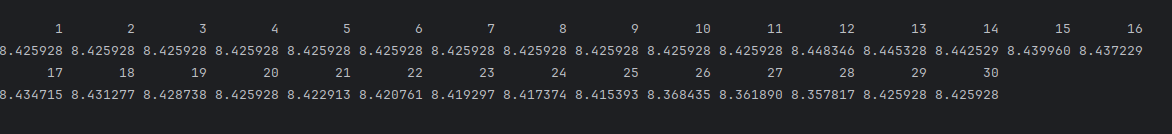


Рисунок 11 – Прогноз по атрибуту значение смертности

Для наглядности был построен график. Он представлен на рисунке 12.

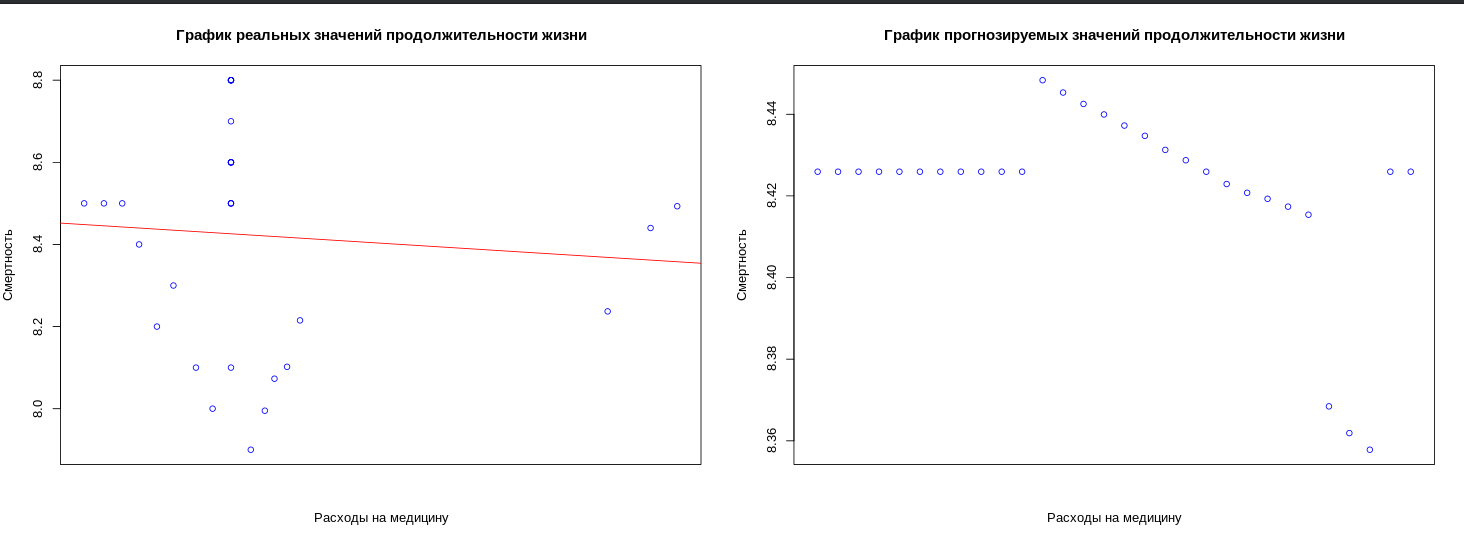


Рисунок 12 – Графики реальных (слева) и прогнозируемых (справа) значений смертности

**Вывод**

В данной работе были ознакомлены с понятием корреляционного и регрессионного анализа данных, некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Помимо этого, научились оценивать связь между переменными и оценивать степень этой связи.