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**

**Отчёт лабораторная №7**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гаранина Л.В.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

**Тема:** Корреляционный и регрессионный анализ данных.

**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.

**Задание:**

1. Определить количество переменных и объем выборки в наборе данных longley. Подвести дескриптивный анализ. Получить корреляционную матрицу и пояснить ее. Проверить данные на нормальность. Обосновать результаты графически.
2. Для примера доказать с помощью графиков, что корреляция есть.
3. Сформировать датафрейм для страны Швеция, построить кривую прироста ВВП, проанализировать взаимное влияние атрибутов, отобразить корреляцию:
4. Роста ВВП и прироста населения;
5. Прироста населения на динамику безработицы;
6. Изменения расходов на медицину и увеличения продолжительности жизни и смертность;
7. Прирост людей с высшим образованием на рост экспорта товаров и на прирост высокотехнологичного производства;
8. Расходов на образование на кумулятивный прирост бакалавров среди женщин;
9. Прирост людей с высшим образованием на развитие высоких технологий (прирост статей в научно журнале);
10. Найти и отобразить на графике наиболее коррелирующие параметры;
11. С помощью регрессионного анализа найти зависимые переменные и пояснить влияние на них независимых переменных;
12. С помощью функции predict() построить прогноз по любому атрибуту.

**Ход работы**:

Выполним задания для встроенного набора данных longley, содержащего датафрейм с 7-ю экономическими показателями, собранными в период с 1947 до 1962 года.

Чтобы проверить количество переменных используется метод ncol(), так как количество переменных это количество столбцов в таблице. Аналогично, объем выборки находится через nrow() – количество строк в таблице.

В результате нашли, что Longley содержит 7 переменных и 16 – объем данных. Далее выполним дескриптивный анализ: найдем моду, медиану и среднее. Это можно посмотреть на рисунке 1.

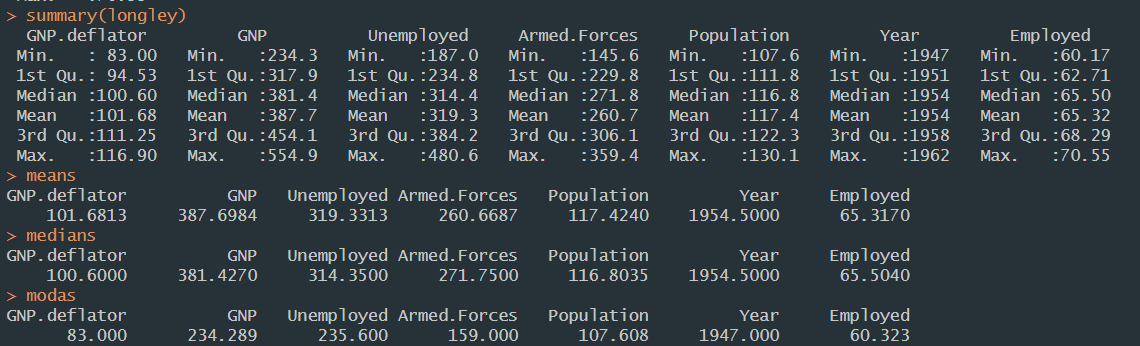


Рисунок 1 – Дескриптивный анализ

Далее получим корреляционную матрицу, изображенную на рисунке 2.

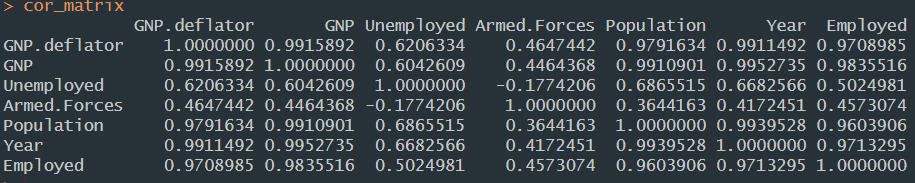


Рисунок 2 – Корреляционнаяя матрица

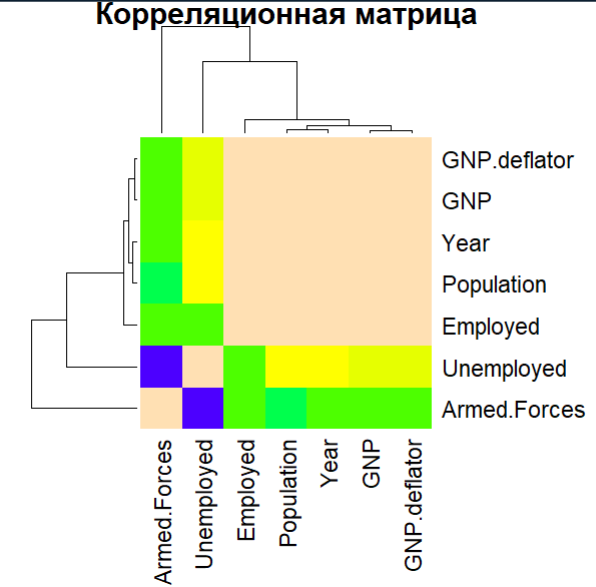


Рисунок 3 – График корреляционной матрицы

По диагонали матрицы коэффициент корреляции всегда равен 1, потому что смотрится линейная зависимость признака от самого себя. Чем ближе к единице по модулю, тем больше корреляция. На графике синий цвет – отрицательная корреляция, бежевый – очень близко к 1. Слева и сверху – дендрограммы, которые показывают, как данные были сгруппированы в процессе кластеризации. Чем они ближе, тем эти переменные более похожи.

Проверим датасет на нормальность распределения. Для этого для каждого признака построим квантильно-квантильный график. Как видно по рисунку 4, данные довольно близки к нормальным. Для этого также проведем тест Шапиро (рисунок 5).

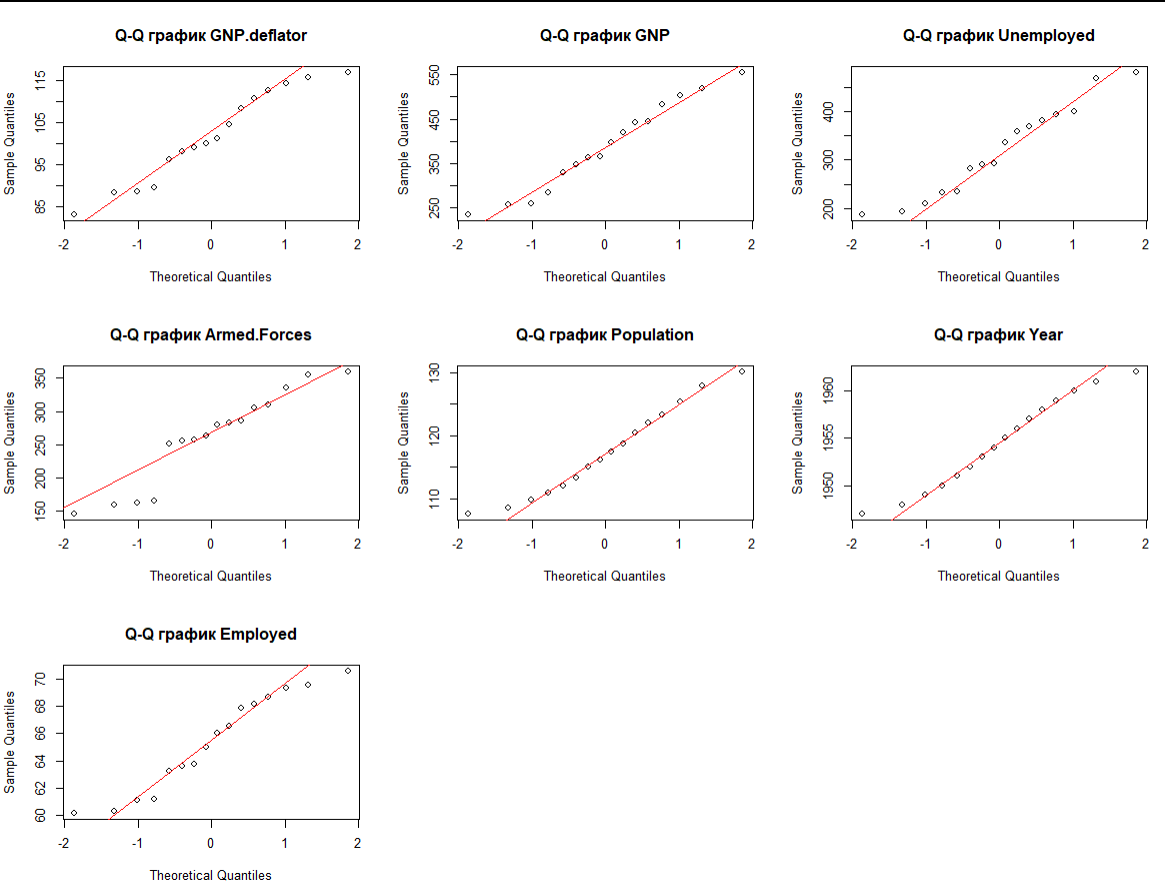


Рисунок 4 – Q-Q графики признаков

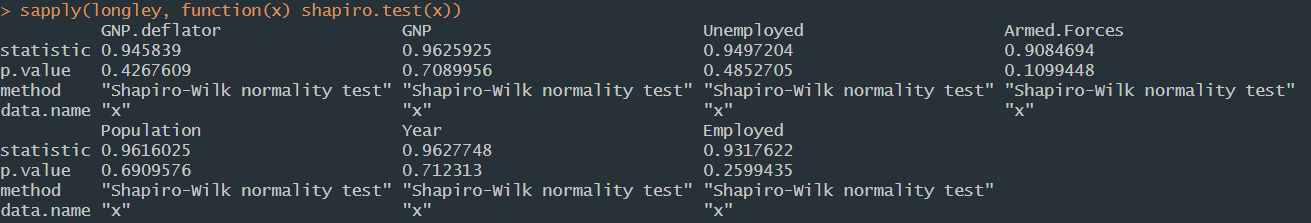


Рисунок 5 – Тест Шапиро для признаков

Здесь у всех признаков значение p-value больше 0,05, что указывает, что данные распределены нормально.

Далее для примера, указанного ниже, нужно доказать с помощью графиков, что корреляция есть.

> x<-rexp(50);

> cor(x,log(x),method="spearman")

[1] 1

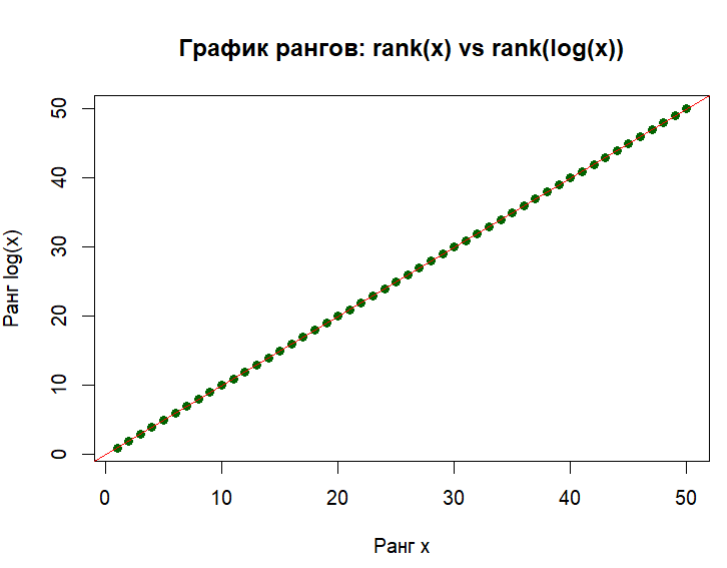


Рисунок 6 – График рангов

Этот график – способ визуализировать Спирменовскую корреляцию, которая измеряет монотонную зависимость между переменными и не обязательно линейную. Так как ранги растут одновременно и точки на графике растут вдоль диагонали, то переменные сильно связаны монотонно.

Теперь создадим датасет по стране Швеция. На рисунке 7 показана его часть.

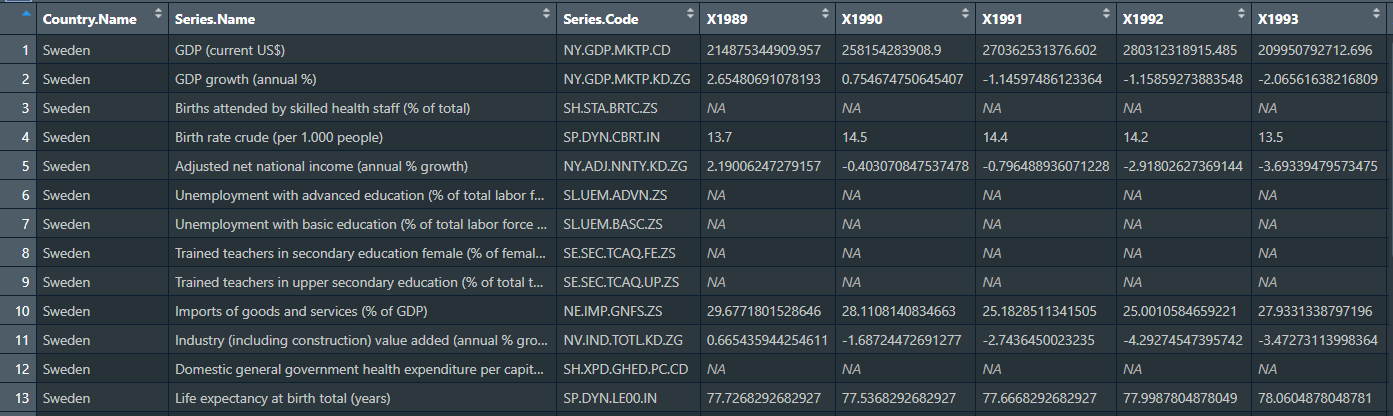


Рисунок 7 – Часть датасета по стране Швеция

Отобразим кривую прироста ВВП.

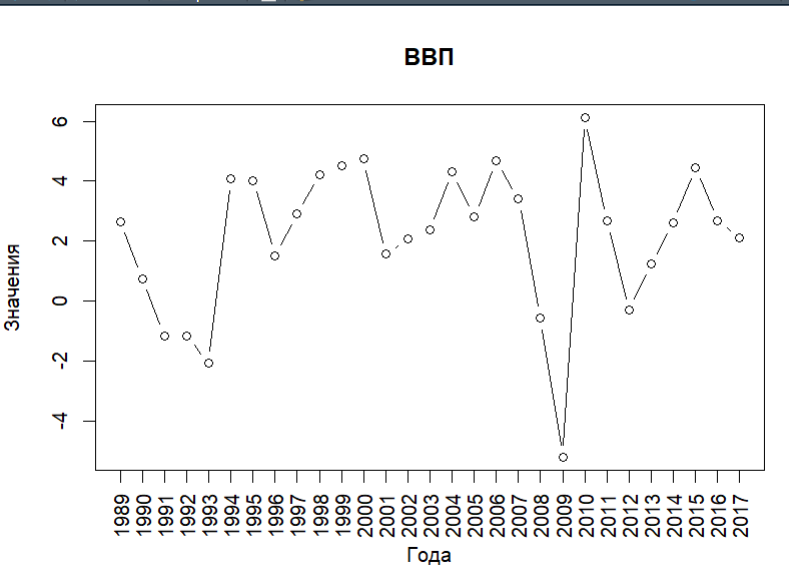


Рисунок 8 – Кривая прироста ВВП

Найдем корреляцию между определенными признаками, как указано в задании. Результаты следующие:

1. Рост ВВП и прирост населения: -0.2098522;
2. Прирост населения и безработица: 0.3443609;
3. Расходы на медицину и продолжительность жизни: 0.9338235;
4. Расходы на медицину и смертность: -0.9115507;
5. Образование и экспорт: 1;
6. Образование и высокотехнологичное производство: -0.5;
7. Расходы на образование и бакалавры-женщины: -1;
8. Высшее образование и статьи в научных журналах: -0.4.

В пунктах 1, 2, имеется довольно слабая связь, в 6, 8 – умеренная, в 3, 4 – очень сильная, а в 5 и 7 – идеальная. Причем в 1, 4, 6, 7 и 8 связь отрицательная, в остальных положительная.

Найдем и отобразим на графике наиболее коррелирующие параметры.

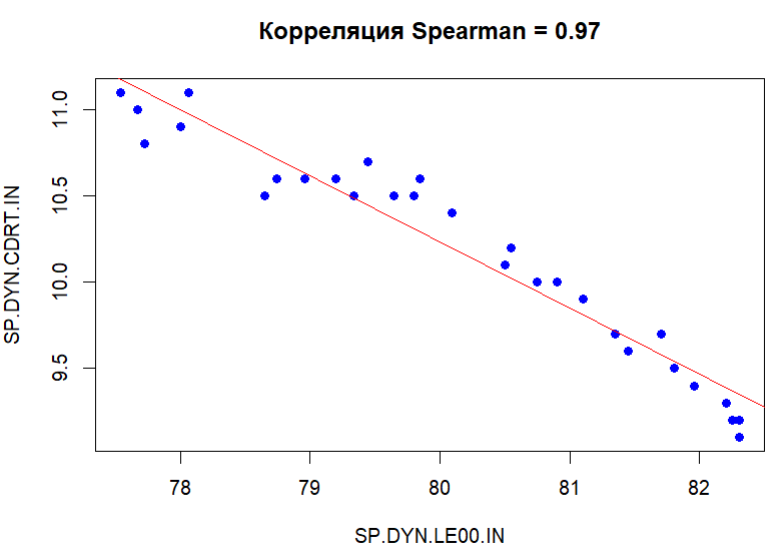


Рисунок 9 – График отрицательной корреляции

График показывает связь общего уровня смертности и ожидаемой продолжительности жизни: чем меньше смертность, тем больше ожидаемый возраст.

Теперь с помощью регрессионного анализа найдем зависимые переменные.

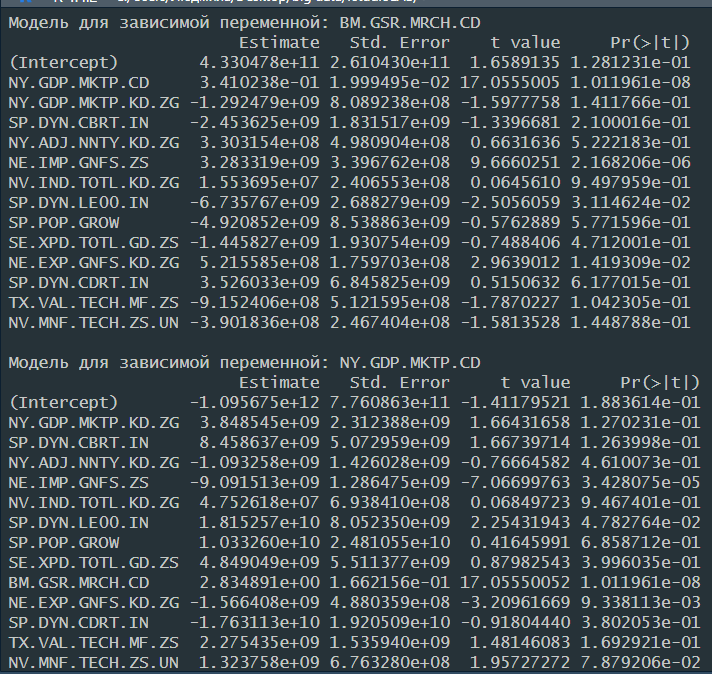


Рисунок 10 – Влияние независимых переменных на зависимые

С помощью функции predict() построим прогноз по атрибуту ВВП. Он показан на рисунке 11. Как можем видеть, прогнозируемое значение очень близко с реальными, значит мы построили хорошую модель с правильными зависимостями.

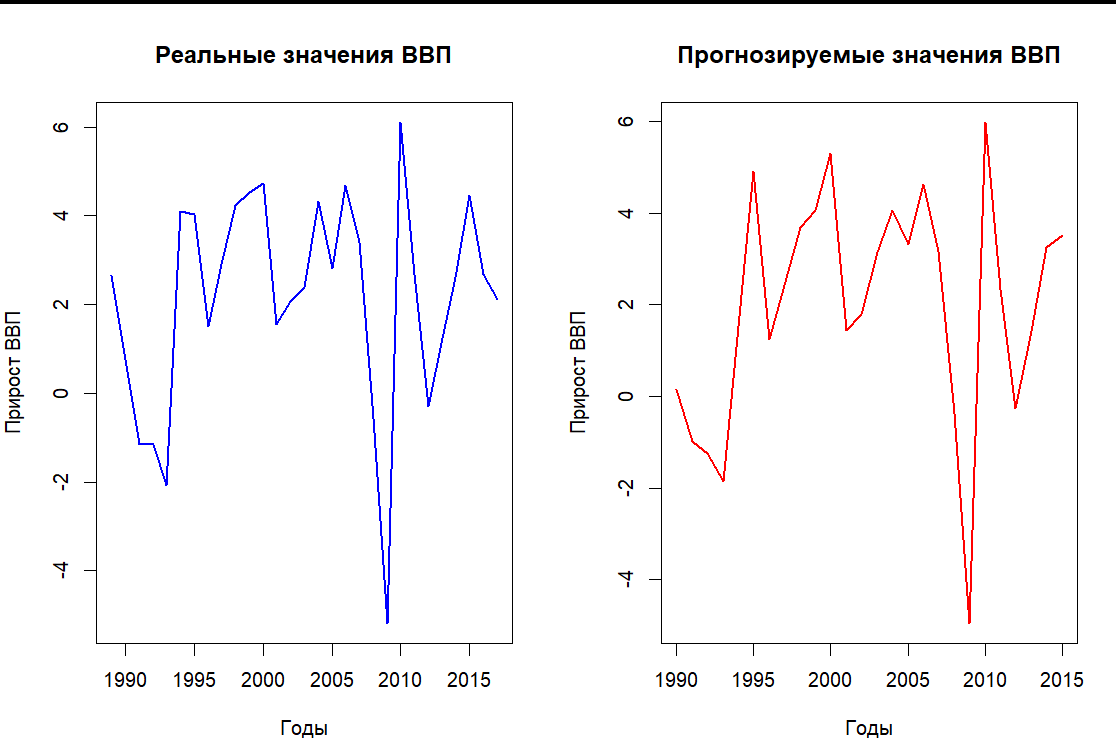


Рисунок 11 – Реальные и прогнозируемые данные

**Вывод:** проделав лабораторную работу, были изучены корреляционный и регрессионный анализ данных, а также оценка связей.

**Листинг программы:**

# 7.1

ncol(longley)

nrow(longley)

summary(longley)

means = apply(longley, 2, mean)

medians = apply(longley, 2, median)

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

modas = apply(longley, 2, getmode)

cor\_matrix <- cor(longley)

heatmap(cor\_matrix, main="Корреляционная матрица", col=topo.colors(10), scale="none", margins = c(8, 8))

par(mfrow=c(3, 3))

for (col in colnames(longley)) {

qqnorm(longley[[col]], main=paste("Q-Q график", col))

qqline(longley[[col]], col="red")

}

sapply(longley, function(x) shapiro.test(x))

# 7.2

# Строим матрицу диаграмм рассеяния

x<-rexp(50);

cor(x,log(x),method="spearman")

y <- log(x)

rank\_x <- rank(x)

rank\_y <- rank(y)

plot(rank\_x, rank\_y,

main = "График рангов: rank(x) vs rank(log(x))",

xlab = "Ранг x",

ylab = "Ранг log(x)",

pch = 19, col = "darkgreen")

abline(lm(rank\_y ~ rank\_x), col = "red")

# 7.

library(car)

library(stringr)

# 1.

setwd("C:/Users/Людмила/Desktop/big data/rstudioLAB")

df\_Sweden<-read.csv("Lab8\_Data.csv", sep=",", header=TRUE, encoding="UTF-8")

df\_Sweden[df\_Sweden == ".."] <- NA

# 2.

extract\_data <- function(df, codes) {

subset <- df[df$Series.Code %in% codes, 4:ncol(df)]

subset <- subset[, colSums(is.na(subset)) == 0]

# Преобразуем в числовую матрицу без изменения имён столбцов и строк:

numeric\_subset <- matrix(as.numeric(as.matrix(subset)), nrow = nrow(subset))

rownames(numeric\_subset) <- rownames(subset)

colnames(numeric\_subset) <- colnames(subset)

return(numeric\_subset)

}

# a) Рост ВВП и прирост населения.

GDP\_and\_Population <- extract\_data(df\_Sweden, c("NY.GDP.MKTP.KD.ZG", "SP.POP.GROW"))

cor(GDP\_and\_Population[1,], GDP\_and\_Population[2,], method = "spearman")

x <- as.numeric(gsub("^X", "", colnames(GDP\_and\_Population)))

y <- as.numeric(GDP\_and\_Population[1,])

plot(x, y, ylim=c(min(y), max(y)), type="b", lty=1, pch=1, xlab="Года", ylab="Значения", main="ВВП", xaxt="n")

axis(1, at=x, labels=x, las=2)

# b) Прирост населения и безработица

pop\_unemp <- extract\_data(df\_Sweden, c("SP.POP.GROW", "SL.UEM.ADVN.ZS"))

cor(pop\_unemp[1,], pop\_unemp[2,], method = "spearman")

# c) Расходы на медицину, продолжительность жизни и смертность

health <- extract\_data(df\_Sweden, c("SH.XPD.GHED.PC.CD", "SP.DYN.LE00.IN", "SP.DYN.CDRT.IN"))

cor(health[1,], health[2,], method = "spearman")

cor(health[1,], health[3,], method = "spearman")

# d) Образование, экспорт и высокотехнологичное производство

edu\_prod <- extract\_data(df\_Sweden, c("SE.TER.CUAT.BA.ZS", "NE.EXP.GNFS.KD.ZG", "NV.MNF.TECH.ZS.UN"))

cor(edu\_prod[1,], edu\_prod[2,], method = "spearman")

cor(edu\_prod[1,], edu\_prod[3,], method = "spearman")

# e) Расходы на образование и бакалавры-женщины

edu\_female <- extract\_data(df\_Sweden, c("SE.XPD.TOTL.GD.ZS", "SE.TER.CUAT.BA.FE.ZS"))

cor(edu\_female[1,], edu\_female[2,], method = "spearman")

# f) Высшее образование и статьи в научных журналах

edu\_articles <- extract\_data(df\_Sweden, c("SE.TER.CUAT.BA.ZS", "IP.JRN.ARTC.SC"))

cor(edu\_articles[1,], edu\_articles[2,], method = "spearman")

# g) Найти и отобразить на графике наиболее коррелирующие параметры.

data\_all <- df\_Sweden[, -c(1,2)]

for (i in 2:ncol(data\_all)) {

data\_all[[i]] <- as.numeric(data\_all[[i]])

}

data\_matrix <- as.data.frame(t(data\_all[, -1]))

colnames(data\_matrix) <- data\_all$Series.Code

data\_matrix <- data\_matrix[, colSums(is.na(data\_matrix)) <= 8]

data\_matrix <- data\_matrix[, colSums(is.na(data\_matrix)) < nrow(data\_matrix)]

cor\_matrix <- cor(data\_matrix, use = "pairwise.complete.obs", method = "spearman")

off\_diag <- row(cor\_matrix) != col(cor\_matrix)

max\_cor\_value <- max(abs(cor\_matrix[off\_diag]), na.rm = TRUE)

which\_max <- which(abs(cor\_matrix) == max\_cor\_value & off\_diag, arr.ind = TRUE)

var1 <- colnames(cor\_matrix)[which\_max[1, "col"]]

var2 <- rownames(cor\_matrix)[which\_max[1, "row"]]

cat("Наиболее скоррелированные параметры:", var1, "и", var2, "\n")

plot(data\_matrix[[var1]], data\_matrix[[var2]],

xlab = var1, ylab = var2,

main = paste("Корреляция Spearman =", round(max\_cor\_value, 2)),

pch = 19, col = "blue")

abline(lm(data\_matrix[[var2]] ~ data\_matrix[[var1]]), col = "red")

# h) С помощью регрессионного анализа найдите зависимые переменные и поясните влияние на них независимых переменных.

results <- list()

for (target in colnames(data\_matrix)) {

predictors <- setdiff(colnames(data\_matrix), target)

formula <- as.formula(paste(target, "~", paste(predictors, collapse = "+")))

model <- lm(formula, data = data\_matrix)

r2 <- summary(model)$r.squared

results[[target]] <- r2

}

sorted\_r2 <- sort(unlist(results), decreasing = TRUE)

top\_targets <- names(sorted\_r2)[1:3]

for (target in top\_targets) {

predictors <- setdiff(colnames(data\_matrix), target)

formula <- as.formula(paste(target, "~", paste(predictors, collapse = "+")))

model <- lm(formula, data = data\_matrix)

cat("\nМодель для зависимой переменной:", target, "\n")

print(summary(model)$coefficients)

}

# i) С помощью функции predict() постройте прогноз по любому понравившемуся Вам атрибуту.

target <- "NY.GDP.MKTP.KD.ZG" # Прирост ВВП

predictors <- setdiff(colnames(data\_matrix), target)

formula <- as.formula(paste(target, "~", paste(predictors, collapse = "+")))

model <- lm(formula, data = data\_matrix, na.action = na.exclude)

required\_vars <- all.vars(formula)

data\_complete <- data\_matrix[complete.cases(data\_matrix[, required\_vars]), ]

predictions <- predict(model, newdata = data\_complete)

years <- as.numeric(gsub("^X", "", names(predictions)))

actual <- data\_matrix[[target]]

x\_filtered <- 1989:2017

actual\_filtered <- actual[!is.na(actual)]

par(mfrow=c(1, 2))

plot(x\_filtered, actual\_filtered, type="l", col="blue", lwd=2,

xlab="Годы", ylab="Прирост ВВП", main="Реальные значения ВВП")

plot(years, predictions, type="l", col="red", lwd=2,

xlab="Годы", ylab="Прирост ВВП", main="Прогнозируемые значения ВВП")