Анализ данных

Лабораторные работы

Храмов Д.А.

17.11.2019

[Лабораторная работа №1. Векторы и матрицы 1](#_Toc32952737)

[Лабораторная работа №2. Списки и таблицы 4](#_Toc32952741)

[Лабораторная работа № 3. Разведочный анализ данных 10](#_Toc32952745)

[Лабораторная работа № 4. Кластерный анализ: метод иерархической кластеризации 11](#_Toc32952749)

[Лабораторная работа № 5. Кластерный анализ: метод k-средних 12](#_Toc32952753)

[Лабораторная работа № 6. Линейная регрессия 12](#_Toc32952757)

[Лабораторная работа № 7. Прогнозирование временных рядов 14](#_Toc32952761)

[Лабораторная работа № 8. Классификация: метод k ближайших соседей 15](#_Toc32952765)

[Лабораторная работа № 9. Классификация: деревья классификации и регрессии 15](#_Toc32952769)

1. В пояснительной записке к работе обязательно указывать: фамилию, имя, номер работы, номер группы, номер варианта.
2. Задание записывать полностью.
3. Решение оформить в виде R-скрипта.
4. Привести вывод скрипта и необходимые иллюстрации (черно-белые).
5. В конце пояснительной записки привести результат выполнения sessionInfo() (информация об использованной версии R, загруженных пакетах и операционной системе).

# Лабораторная работа №1. Векторы и матрицы

## Вариант 1

**1.** Создайте вектор и

1. выведите на экран элементы v с 5-го по 9-й;
2. создайте вектор u, добавив в конец вектора v числа 3.5 и 5;
3. удалите элементы v с 1-го по 3-й;
4. замените 7-й элемент v значением NA.

**2.** Функция runif(n) создает вектор, состоящий из случайных чисел, равномерно распределенных на интервале .

1. Создайте вектор x, состоящий из десяти чисел, равномерно распределенных на интервале .
2. Выделите из x элементы, удовлетворяющие условию и составьте из них вектор y. Используйте логическую индексацию.
3. Подсчитайте число элементов y.
4. Составьте из x и y вектор z следующим образом: .
5. Составьте вектор w из всех элементов z, за исключением 2-го, 3-го и 4-го.

Перед запуском runif задайте затравку генератора случайных чисел с помощью функции set.seed(m). m - номер студента в списке группы.

**3.** Постройте с помощью функции rep последовательности:

1. 6,6,6,6,6,6
2. 5,8,5,8,5,8,5,8
3. 5,5,5,5,8,8,8,8

**4.** С помощью функции seq постройте последовательности:

1. чисел от 1 до 10;
2. чисел от 3 до 21 с шагом 2;
3. чисел от -1 до 1 количеством 20.

**5.** Создайте из векторов и

1. матрицу A
2. матрицу B

**6.** Постройте график функции на интервале .

## Вариант 2

**1.** Создайте вектор и

1. выведите на экран элементы v с 2-го по 7-й;
2. создайте вектор w, добавив в конец вектора v вектор ;
3. удалите элементы v с 8-го по 10-й;
4. замените 5-й элемент w значением NA.

**2.** Функция runif(n) создает вектор, состоящий из случайных чисел, равномерно распределенных на интервале .

1. Создайте вектор x, состоящий из десяти чисел, равномерно распределенных на интервале .
2. Выделите из x элементы, удовлетворяющие условию и составьте из них вектор y. Используйте логическую индексацию.
3. Подсчитайте число элементов y.
4. Составьте из x и y вектор z следующим образом: .
5. Составьте вектор w из всех элементов z, за исключением 2-го, 3-го и 4-го.

Перед запуском runif задайте затравку генератора случайных чисел с помощью функции set.seed(m). m - номер студента в списке группы.

**3.** Постройте с помощью функции rep последовательности:

1. 3, повторенное 10 раз
2. 1,8,1,8,1,8,1,8
3. 4,4,4,8,8,8,-1,-1,-1

**4.** С помощью функции seq постройте последовательности:

1. чисел от 1 до 15;
2. чисел от 5 до 65 с шагом 5;
3. чисел от 0 до 5.5 количеством 20.

**5.** Создайте из векторов и

1. матрицу A
2. матрицу B

**6.** Постройте график функции на интервале .

## Вариант 3

**1.** Создайте вектор и

1. выведите на экран элементы v с 8-го по 11-й;
2. создайте вектор w, добавив в конец вектора v число -6;
3. удалите элементы v с 1-го по 3-й;
4. замените элемент v c 5-го по 8-й значениями NA.

**2.** Функция runif(n) создает вектор, состоящий из случайных чисел, равномерно распределенных на интервале .

1. Создайте вектор x, состоящий из десяти чисел, равномерно распределенных на интервале .
2. Выделите из x элементы, удовлетворяющие условию и составьте из них вектор y. Используйте логическую индексацию.
3. Подсчитайте число элементов y.
4. Составьте из x и y вектор z следующим образом: .
5. Составьте вектор w из всех элементов z, за исключением 2-го, 3-го и 4-го.

Перед запуском runif задайте затравку генератора случайных чисел с помощью функции set.seed(m). m - номер студента в списке группы.

**3.** Постройте с помощью функции rep последовательности:

1. 1, повторенное 9 раз
2. -1,1,-1,1,-1,1,-1,1
3. 3,3,3,1,1,1,2,2,2

**4.** С помощью функции seq постройте последовательности:

1. чисел от 1 до 25;
2. чисел от 2 до 16 с шагом 2;
3. чисел от 0 до 3 количеством 30.

**5.** Создайте из векторов и

1. матрицу С
2. матрицу D

**6.** Постройте график функции на интервале .

# Лабораторная работа №2. Списки и таблицы

## Вариант 1

**1.** Создайте список my.list, используя в качестве элементов:

v1 <- c("A", "B", "C")  
v2 <- seq(1, 3, 0.5)  
v3 <- c(FALSE, TRUE)

1. Присвойте элементам списка имена text, number и logic.
2. Выведите на экран структуру списка my.list.
3. Выведите на экран компоненты с 3-го по 5-й из элемента number списка my.list.
4. Выведите на экран все компоненты элемента number, кроме 2-го.
5. Выведите на экран первый элемент my.list.
6. Вычислите длину элемента number списка.
7. Определите класс списка my.list и классы его элементов. *Совет*: для определения классов элементов списка используйте функцию class совместно с lapply или sapply.

*Совет*: для обращения к элементам списка по именам применяют знак $ (список$элемент), а по номерам – двойные квадратные скобки (список[[номер\_элемента]]).

**2.** Создайте список mylist, состоящий из трех элементов: последовательности чисел от 1 до 5; строки "My Text"; функции, реализующей умножение x \* sin(x). Присвойте элементам списка имена a, b и c соответственно.

* Можно ли выполнить следующие действия:

1. mylist[1] + 1
2. mylist$a + 1

* Добавьте в mylist элемент d, равный "Hello!" и выведите на экран структуру обновленного списка mylist.
* Удалите элемент b. Выведите на экран структуру полученного в результате списка. *Совет*: для удаления элемента name списка list нужно приравнять этот элемент НИЧТО (NULL): list$name <- NULL.

**3.** Создайте таблицу df на основе следующих данных:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Height | Weight | Age | Hand |
| John | 180 | 75 | 18 | L |
| Mary | 163 | 48 | 17 | R |
| Sam | 177 | 72 | 17 | R |

Колонки таблицы: Height (рост в см), Weight (вес в кг), Age (возраст в годах), Hand (доминантная рука: R=правая, L=левая).

1. Задайте имена строк df: c("John", "Mary", "Sam").
2. Выведите на экран заголовки колонок таблицы df.
3. Выведите на экран колонку с информацией о росте (Height).
4. Выведите на экран статистическую сводку по росту.
5. Выведите на экран данные по росту (Height), весу (Weight) и возрасту (Age) Мэри.
6. Каков будет результат выполнения команды df[5,5]?
7. Выведите на экран рост тех, чей вес превышает 50 кг.

*Советы*: Имена строк задают функцией rownames или аргументом row.names функции data.frame. Статистическую сводку по данным, содержащимся в векторе x (среднее, максимальное и минимальное значения, медиану и пр.), можно узнать с помощью summary(x).

**4.** Создайте таблицу City на основе следующих данных по численности мужского (Male) и женского (Female) населения Нью-Йорка, Парижа и Шанхая:

|  |  |  |
| --- | --- | --- |
| City | Sex | Number |
| New York | Male | 3794204 |
| New York | Female | 4214074 |
| Paris | Male | 1032971 |
| Paris | Female | 1160060 |
| Shanghai | Male | 8776560 |
| Shanghai | Female | 8623440 |

* Выведите на экран данные по численности населения (колонку Number) двумя способами: с помощью имени колонки и с помощью ее номера.
* К какому классу относится результат выполнения команд:

1. City[,3]
2. City[3]
3. City[[3]]

* Извлеките из таблицы данные о численности мужского населения городов.
* Извлеките из таблицы названия городов, численность женского населения которых превосходит 4000000.

**5.** Создайте таблицы df1, df2, df3 с элементами:

df1: a = 1:5, b = 6:10, c = letters[1:5]  
df2: d = 6:10, e = 11:15, f = letters[6:10]  
df3: a = 6:10, b = 11:15, c = letters[6:10]

* Объедините df1 и df2 по столбцам.
* Объедините df1 и df3 по строкам.
* Можно ли объединить по строкам df1 и df2?

**6.** Набор данных mtcars содержит параметры 32 автомобилей по данным журнала *Motor Trend* (США) за 1974 год. Для доступа к данным просто наберите mtcars в R.

В mtcars содержатся данные о пробеге на одном галлоне топлива (mpg), числе цилиндров двигателя (cyl), объеме двигателя в кубических дюймах (disp), массе автомобиля в тоннах (wt) и другая информация.

1. Создайте новую таблицу, в которую включите все данные по автомобилям с 6-цилиндровым двигателем.
2. Сколько машин имеют объем двигателя больше 300 in?
3. Какую среднюю мощность двигателя (hp – от *horsepower*) имеют автомобили с 4, 6 и 8 цилиндрами?
4. Постройте график зависимости пробега на одном галлоне от мощности двигателя. Выделите разными цветами группы машин, различающиеся по числу цилиндров.

## Вариант 2

**1.** Создайте из векторов:

список Date с элементами year, month и day.

1. Выведите на экран элемент year списка Date.
2. Выведите на экран компоненты с 4-го по 7-й из элемента day списка Date.
3. Выведите на экран все компоненты элемента month, кроме 2-го.
4. Запишите команду, которая заменит значения элемента year в списке Date на c(2010:2020).
5. Вычислите длину элемента year, получившуюся после замены.
6. Удалите 4-й элемент из month в Date.

*Совет*: для обращения к элементам списка по именам применяют знак $ (список$элемент), а по номерам – двойные квадратные скобки (список[[номер\_элемента]]).

**2.**

1. Для списка mylist <- list(a=1:10, b="Good morning", c="Hi") запишите команды R, которые добавляют 1 к каждому элементу первого вектора из mylist.
2. Для списка b <- list(a=1:10, c="Hello", d="AA") запишите команды, выбирающие вcе элементы, кроме 2-го, из первого вектора в b.
3. Запишите команды, добавляющие в список x <- list(a=5:10, c="Hello", d="AA") новый элемент z со значением "NewItem".
4. Рассмотрим список y <- list("a", "b", "c"). Запишите команды, которые задают новые имена ("one", "two", "three") элементам y.

**3.** Создайте таблицу my.data, колонки которой содержат:

* вектор из строк "London", "Paris", "Casablanca", "Roma", "Madrid";
* вектор из чисел от 5 до 9;
* логический вектор из значений: TRUE, TRUE, FALSE, FALSE, FALSE.

Присвойте столбцам таблицы имена col1, col2 и col3 соответственно.

Присвойте строкам таблицы имена "1st", "2nd", "3rd", "4th", "5th".

**4.** Создайте на основе данных mtcars таблицу auto: первый столбец – марки автомашин (хранятся в именах строк mtcars), второй столбец – пробег на одном галлоне топлива (mpg), третий столбец – объем двигателя в кубических дюймах (disp).

Набор данных mtcars содержит параметры 32 автомобилей по данным журнала *Motor Trend* (США) за 1974 год. Для доступа к данным наберите mtcars в R.

1. Преобразуйте данные пробега mpg из миль/галлон в километр/литр (1 миля = 1.609 км, 1 галлон = 3.785 л). Сохраните преобразованные данные во втором столбце auto. Замените название этого столбца на kmpl.
2. Преобразуйте объем двигателя (disp) из кубических дюймов в литры (1 дюйм = 2.54 см, 1000 см = 1 л). Сохраните преобразованные данные в столбце vol. Удалите столбец disp из таблицы.
3. Выясните, какая марка машин имеет наименьший расход топлива, а какая наибольший. *Совет*: воспользуйтесь функцией summary или функциями min, max.
4. Постройте график зависимости пробега на одном литре топлива от объема двигателя в литрах.

**5.** Создайте таблицу City на основе следующих данных по численности мужского (Male) и женского (Female) населения Нью-Йорка, Парижа и Шанхая:

|  |  |  |
| --- | --- | --- |
| City | Sex | Number |
| New York | Male | 3794204 |
| New York | Female | 4214074 |
| Paris | Male | 1032971 |
| Paris | Female | 1160060 |
| Shanghai | Male | 8776560 |
| Shanghai | Female | 8623440 |

1. Извлеките из таблицы и выведите на экран данные по численности женского населения городов.
2. Извлеките из таблицы названия городов, численность женского населения которых превосходит 2 млн.
3. Каково общее число мужчин во всех указанных городах?

**6.** Создайте таблицы df1, df2, df3 с элементами:

df1: a = 1:5, b = LETTERS[1:5]  
df2: number = 6:10, letter = letters[6:10]  
df3: a = 6:10, b = letters[6:10]

1. Объедините df1 и df2 по столбцам.
2. Объедините df1 и df3 по строкам.

## Вариант 3

**1.** Создайте список mylist, состоящий из элементов:

1. последовательности чисел от 1 до 5;
2. "Hello!";
3. последовательности символов a, b, ..., f

Присвойте элементам списка имена a, b и c соответственно.

* Удалите элемент b. Выведите на экран структуру полученного в результате списка.
* Добавьте в mylist элемент d, равный "Hi!" и выведите на экран структуру обновленного списка mylist.
* Можно ли выполнить следующие действия:

1. mylist[1] + 1
2. mylist$a + 1

*Совет*: для удаления элемента name списка list нужно приравнять этот элемент НИЧТО (NULL): list$name <- NULL.

**2.**

1. Рассмотрим список x <- list("a", "b", "c"). Запишите команды, которые задают новые имена ("one", "two", "three") элементам x.
2. Для списка my <- list(a=1:10, b="Good morning", c="See you") запишите команды R, которые добавляют 1 к каждому элементу первого вектора из my.
3. Для списка x <- list(a=1:10, c="Hello", d="AA") запишите команды, выбирающие вcе элементы, кроме 2-го, из первого вектора в x.
4. Запишите команды, добавляющие в список y <- list(a=5:10, c="Hello", d="Hi") новый элемент b со значением "Wow".

**3.** Создайте таблицу df на основе следующих данных:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Height | Weight | Age | Hand |
| John | 180 | 75 | 17 | L |
| Mary | 163 | 48 | 17 | R |
| Sam | 177 | 82 | 47 | R |

Колонки таблицы: Height (рост в см), Weight (вес в кг), Age (возраст в годах), Hand (доминантная рука: R=правая, L=левая).

1. Задайте имена строк df: c("John", "Mary", "Sam").
2. Выведите на экран заголовки колонок таблицы df.
3. Выведите на экран колонку с информацией о росте (Height).
4. Выведите на экран статистическую сводку по росту.
5. Выведите на экран данные по росту (Height), весу (Weight) и возрасту (Age) Джона и Сэма.
6. Выведите на экран рост тех, чей вес превышает 80 кг.

*Советы*: Имена строк задают функцией rownames или аргументом row.names функции data.frame. Статистическую сводку по данным, содержащимся в векторе x (среднее, максимальное и минимальное значения, медиану и пр.), можно узнать с помощью summary(x).

**4.** Вычислите температуру по Фаренгейту, соответствующую 20–25 градусам Цельсия, по формуле:

Градусы Цельсия задаются с шагом 1.

Сохраните полученные значения температур в таблице C2F\_Conversion с заголовками столбцов Celsius и Fahrenheit и выведите таблицу на экран.

**5.** Набор данных mtcars содержит параметры 32 автомобилей по данным журнала *Motor Trend* (США) за 1974 год. Данные касаются пробега на одном галлоне топлива (mpg), числа цилиндров двигателя (cyl), объема двигателя в кубических дюймах (disp), массы автомобиля в тоннах (wt) и др. Для доступа к данным наберите mtcars в R.

1. Создайте новую таблицу, в которую включите все данные по автомобилям с 8-цилиндровым двигателем.
2. Какую среднюю мощность двигателя (hp – от *horsepower*) имеют автомобили с 4, 6 и 8 цилиндрами?
3. Сколько машин имеют объем двигателя больше 350 in?
4. Постройте график зависимости пробега на одном галлоне от мощности двигателя. Выделите разными цветами группы машин, различающиеся по числу цилиндров.

**6.** Для таблиц:

a <- data.frame(a=c(0,1,2), b=c(3,4,5), c=c(6,7,8)) b <- data.frame(a=c(9,10,11), c=c(12,13,14))

1. Объедините таблицы a и b по строкам. Для этого добавьте к таблице b одноименный столбец, заполненный пробелами в данных (NA).
2. Объедините таблицы a и b по столбцам, результат сохраните в таблице c. Выведите заголовки столбцов c (names). Переименуйте столбцы c в a, b, ..., e.

# Лабораторная работа № 3. Разведочный анализ данных

**Задание**

1. Подсчитать описательные статистики, характеризующие численность населения городов Украины (summary).
2. Подсчитать среднее арифметическое численности населения городов (mean). Сколько процентов городов Украины имеет население меньше, чем население “средне-арифметического” города?
3. Постройте гистограмму численности населения городов Украины.
4. Есть ли в выборке выбросы? Если есть, то удалите их из выборки и снова рассчитайте среднее арифметическое. На сколько оно изменилось?
5. Чему равна медиана численности населения городов? Как он изменилась после удаления выбросов?
6. Какая мера центральной тенденции выборки – среднее арифметическое или медиана – более устойчиво к присутствию выбросов?
7. Какие меры центральной тенденции выборки вы знаете?

## Вариант 1

Набор данных: *Население городов Украины по данным переписи 1959 г.* (town\_ua\_1959.csv).

## Вариант 2

Набор данных: *Население городов Украины по данным переписи 2001 г.* (town\_ua\_2001.csv).

## Вариант 3

Набор данных: *Население городов Украины по данным переписи 1989 г.* (town\_ua\_1989.csv).

# Лабораторная работа № 4. Кластерный анализ: метод иерархической кластеризации

Необходимо провести кластерный анализ предложенного набора данных с помощью метода иерархической кластеризации.

**Задание**

1. Выполнить кластерный анализ.
   1. Отобрать необходимые переменные.
   2. Выполнить стандартизацию (если это нужно).
   3. Задать расстояние между объектами.
   4. Задать расстояние между кластерами.
   5. Визуализировать результаты при помощи дендрограммы.
2. Определить число кластеров. Обосновать свой выбор, проверив другие варианты.
3. Интерпретировать результаты кластерного анализа:
   * что представляют собой типичный элемент кластера?
   * чем объекты в каждом из получившихся кластеров похожи друг на друга?
   * чем объекты из разных кластеров отличаются друг от друга?

## Вариант 1

Набор данных: *Экономика городов*.

## Вариант 2

Набор данных: *Занятость в странах Европы*.

## Вариант 3

Набор данных: *USArrests* (Тяжкие преступления).

*Внимание!* В файлах README к наборам данных могут содержаться дополнительные вопросы. Ответьте на них.

# Лабораторная работа № 5. Кластерный анализ: метод k-средних

Необходимо провести кластерный анализ предложенного набора данных с помощью метода k-средних.

**Задание**

1. Выполнить кластерный анализ.
   1. Отобрать необходимые переменные.
   2. Выполнить стандартизацию (если это нужно).
   3. Определить число кластеров при помощи графика “каменистая осыпь” (“локоть”) или путем проб и ошибок. Обосновать свой выбор.
   4. Выполнить кластеризацию с помощью kmeans(). Предварительно задать затравку генератора случайных чисел: set.seed(N), где N – номер студента в списке группы.
   5. Визуализировать результаты при помощи многомерного шкалирования (функция cmdscale).
2. Интерпретировать результаты кластерного анализа:
   * что представляет собой типичный элемент кластера?
   * чем объекты из разных кластеров похожи друг на друга?
   * чем объекты из разных кластеров различаются?

## Вариант 1

Набор данных: *USArrests* (Тяжкие преступления).

## Вариант 2

Набор данных: *Boston Housing*.

## Вариант 3

Набор данных: *Расходы на питание во Франции*.

*Внимание!* В файлах README к наборам данных могут содержаться дополнительные вопросы. Ответьте на них.

# Лабораторная работа № 6. Линейная регрессия

## Вариант 1

Набор данных: *Boston Housing* (Цены на жилье в пригородах Бостона).

**Задание**

1. Постройте гистограмму распределения цен на жилье. Следует ли преобразовать его к нормальному?
2. Постройте корреляции между признаками. Позволяет ли корреляция исключить какие-то признаки?
3. Разделите данные на обучающую и тестовую выборки. Предварительно задайте затравку генератора случайных чисел: set.seed(N), где N – ваш номер в списке группы.
4. Постройте модель линейной регрессии.
5. Ответьте на вопросы:
   * Являются ли коэффициенты модели статистически значимыми?
   * Чему равен коэффициент детерминации? Что означает полученное значение этого коэффициента?
   * Какую интерпретацию вы могли бы дать свободному члену уравнения регрессии?
6. Исключите незначимые коэффициенты модели и повторите пункты 4-5.
7. Выберите модель, дающую минимальную среднеквадратичную ошибку на тестовой выборке.
8. Отобразите на графике остатки для выбранной модели.
9. В каких районах цены на жилье существенно отличаются от спрогнозированных с помощью линейной модели (являются выбросами)?

## Вариант 2

Набор данных: *Prestige* (Престиж профессии в Канаде).

**Задание**

1. Проследите как доход (income) зависит от образования (education). Остальные переменные можно не рассматривать.
2. Постройте гистограммы распределений образования и дохода.
3. Постройте модель линейной регрессии.
4. Ответьте на вопросы:
   * Являются ли коэффициенты модели статистически значимыми?
   * Чему равен коэффициент детерминации? Что означает такое значение этого коэффициента?
   * Какую интерпретацию вы могли бы дать свободному члену уравнения регрессии? Подсказка: центрируйте переменную с помощью scale(x, center = TRUE, scale = F). Что произойдет, если не выполнить центрирование?
5. Отобразите на графике остатки модели.
6. Какие профессии имеют доход, существенно выходящий за рамки описанного линейной моделью (как в большую, так и в меньшую стороны)?

## Вариант 3

Набор данных: *Abalone* (Морское ушко)

**Задание**

1. Разделите данные на обучающую и тестовую выборки. Предварительно задайте затравку генератора случайных чисел: set.seed(N), где N – ваш номер в списке группы.
2. Постройте график pairs для всего набора данных.
3. Постройте корреляции между признаками.
4. Постройте модель линейной регрессии.
5. Ответьте на вопросы:
   * Являются ли коэффициенты модели статистически значимыми?
   * Чему равен коэффициент детерминации? Что означает такое значение этого коэффициента?
6. Отобразите на графике остатки модели.

plot(model, pch=16, which=1)

model - это ваша линейная модель, полученная с помощью функции lm.

1. Оцените величину среднеквадратичной ошибки на тестовой выборке.
2. Попробуйте добиться уменьшения среднеквадратичной ошибки, используя в модели квадраты (poly(var,2)) и четвертые степени (poly(var,4)) переменных (var). Подсказку дает график pairs.

# Лабораторная работа № 7. Прогнозирование временных рядов

**Задание**

1. Построить график исходного временного ряда. Преобразовать даты в тип Date.
2. Ответить на вопросы:
   1. Есть ли у ряда тренд?
   2. Есть ли сезонность? И какая она?
   3. Меняет ли ряд свой характер?
   4. Есть ли в данных выбросы?
3. Построить линейную регрессионную модель.
4. Выполнить прогноз на заданный промежуток времени.
5. Построить графики исходного ряда и подогнанных данных с добавлением спрогнозированных значений.
6. Вывести спрогнозированные значения в виде таблицы.
7. Какова ошибка прогноза? Сравните ее с величиной тестовых данных.

*Внимание!* Период прогнозирования указан в файлах README к наборам данных.

## Вариант 1

Набор данных: *Вина Авcтралии* (rose).

## Вариант 2

Набор данных: *Australian monthly electricity production*.

## Вариант 3

Набор данных: *Minimum Daily Temperatures*.

# Лабораторная работа № 8. Классификация: метод k ближайших соседей

**Задание**

1. Выделить обучающую и тестовую выборки. Предварительно задать затравку генератора случайных чисел: set.seed(N), где N – номер студента в списке группы.
2. Определить наилучшее значение k.
3. Оценить качество прогноза на тестовой выборке с помощью матрицы неточностей (table, caret::confusionMatrix).
4. Указать процент ошибок, допущенных классификатором на тестовой выборке.

*Внимание!* Стандартизуйте данные в случае необходимости.

## Вариант 1

Набор данных: *BreastCancer* (Рак груди).

## Вариант 2

Набор данных: *Sonar, Mines vs. Rocks* (Сонар: металл или камень?).

## Вариант 3

Набор данных: *Glass* (Определение типа стекла по его составу).

# Лабораторная работа № 9. Классификация: деревья классификации и регрессии

**Задание**

1. Выделить обучающую и тестовую выборки.
2. Построить дерево принятия решений (rpart.plot, rattle).
3. Оценить качество прогноза на тестовой выборке с помощью матрицы неточностей (table, caret::confusionMatrix).
4. Указать процент ошибок, допущенных классификатором на тестовой выборке.

## Вариант 1

Набор данных: *Census Income* (Классификация по доходам).

## Вариант 2

Набор данных: *Credit Approval* (Кредитный скоринг).

## Вариант 3

Набор данных: *Landsat Satellite* (Классификация земного покрова).