МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Кубанский государственный университет»**

**(ФГБОУ ВО «КубГУ»)**

**Кафедра вычислительных технологий**

**ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №6**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_К.С. Фисун

(подпись)

Факультет Компьютерных технологий и прикладной математики

Направление подготовки02.03.02 Фундаментальная информатика и информационные технологии курс 3

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

Краснодар 2022

**Лабораторная работа № 6**

**«Задачи классификации и кластеризации»**

**Цель работы:**

Закрепить знания, об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений.

**Общие сведения:**

Под**классификацией** будем понимать отнесение объектов (наблюдений, событий) к одному из **заранее известных классов**.

Классификация - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются правила причисление объекта к группе).

Классификация относится **к стратегии** **обучения с учителем** (supervised learning), которое также именуют контролируемым или управляемым обучением.

Классификация может быть **одномерной** (по одному признаку) и **многомерной** (по двум и более признакам).

Задача **кластеризации** сходна с задачей классификации, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены и формируются автоматически, поэтому синонимами термина "кластеризация" являются "автоматическая классификация", "**обучение без учителя**" и "таксономия".

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить, как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгустков точек".

Цель кластеризации - поиск таких сгустков.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести **разведочный анализ** и изучить "структуру данных".

Само понятие "кластер" определено неоднозначно: каждый алгоритм (или даже каждое исследование одного и того же алгоритма) формирует свои "кластеры". Переводится понятие кластер (cluster) как "скопление", "гроздь".

**Ход работы:**

Имеется набор данных о крушениях самолетов за период с 1907 по 2017 года (рисунок 1).

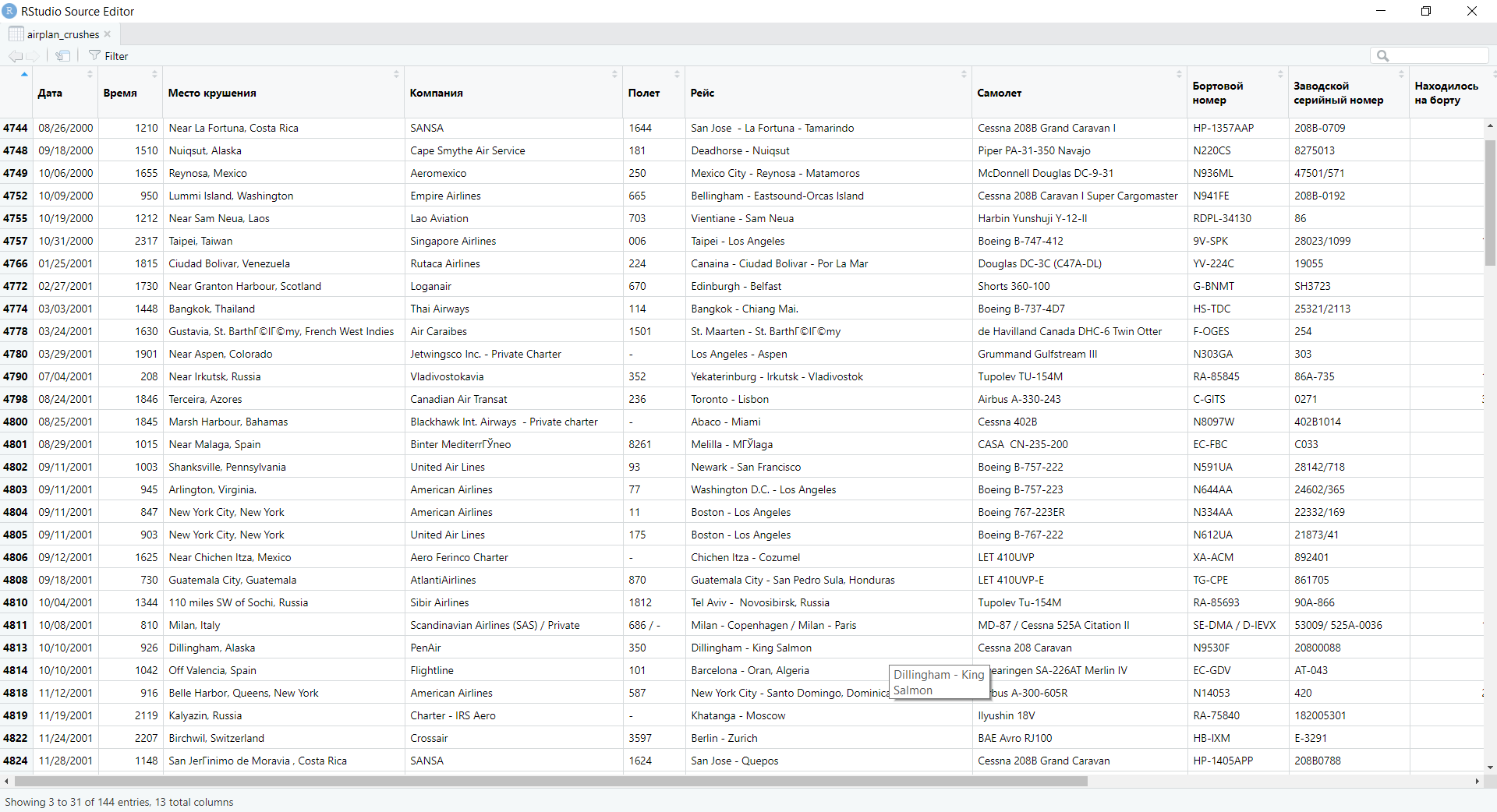


Рис. 1 – Набор данных о крушения самолетов.

Выполним иерархическую кластеризацию набора данных “Крушение самолетов”, построив дендрограмму (рисунок 2).

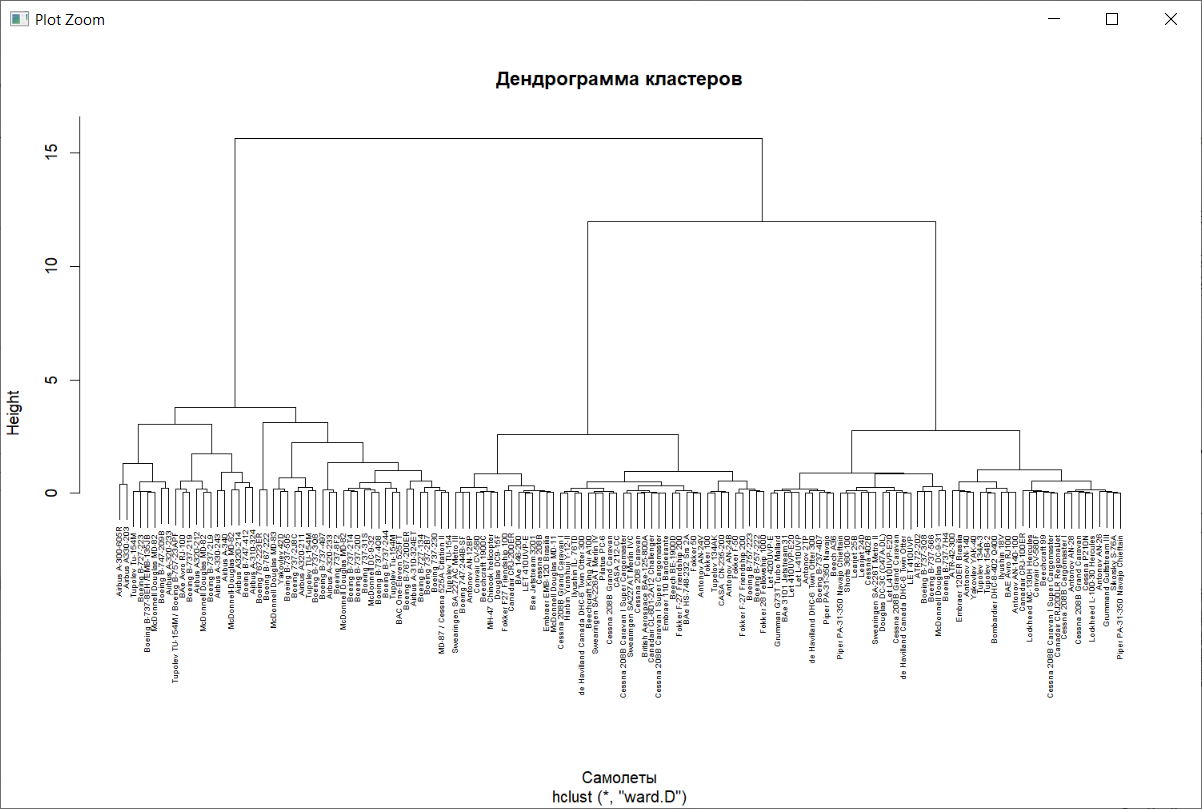


Рис. 2 – Дендрограмма кластеров.

Далее данная дендрограмма была разбита на явно выделяющиеся 3 группы (рисунок 3).

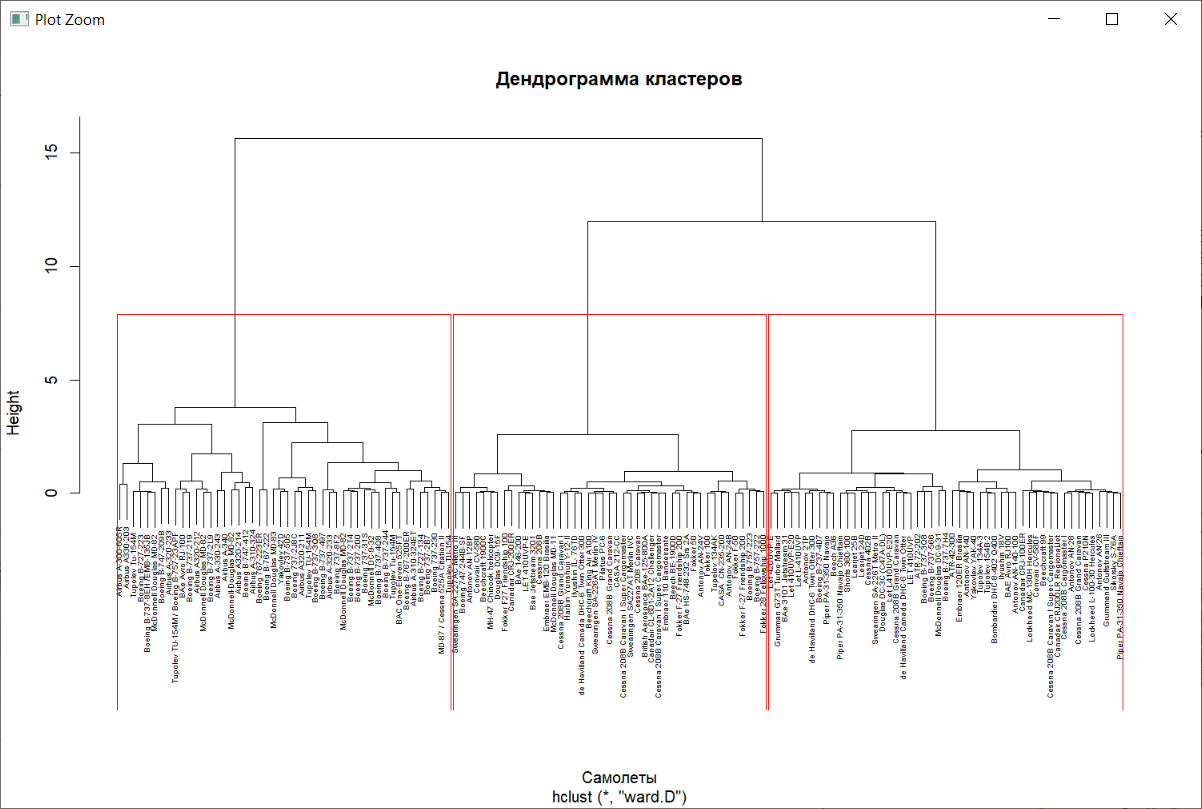


Рис. 3 – Разбитие дендрограммы на 3 группы.

Для каждой группы определяем средние значения характеристик и строим датафрейм. Из полученного датафрейма строим столбчатую диаграмму (рисунок 4). По полученным результатам удалось выяснить, что крушения поделились на три группы:

* крушения с большим количеством людей на борту;
* крушения с поздним временем и малым количеством людей;
* крушения с ранним временем и малым количеством людей

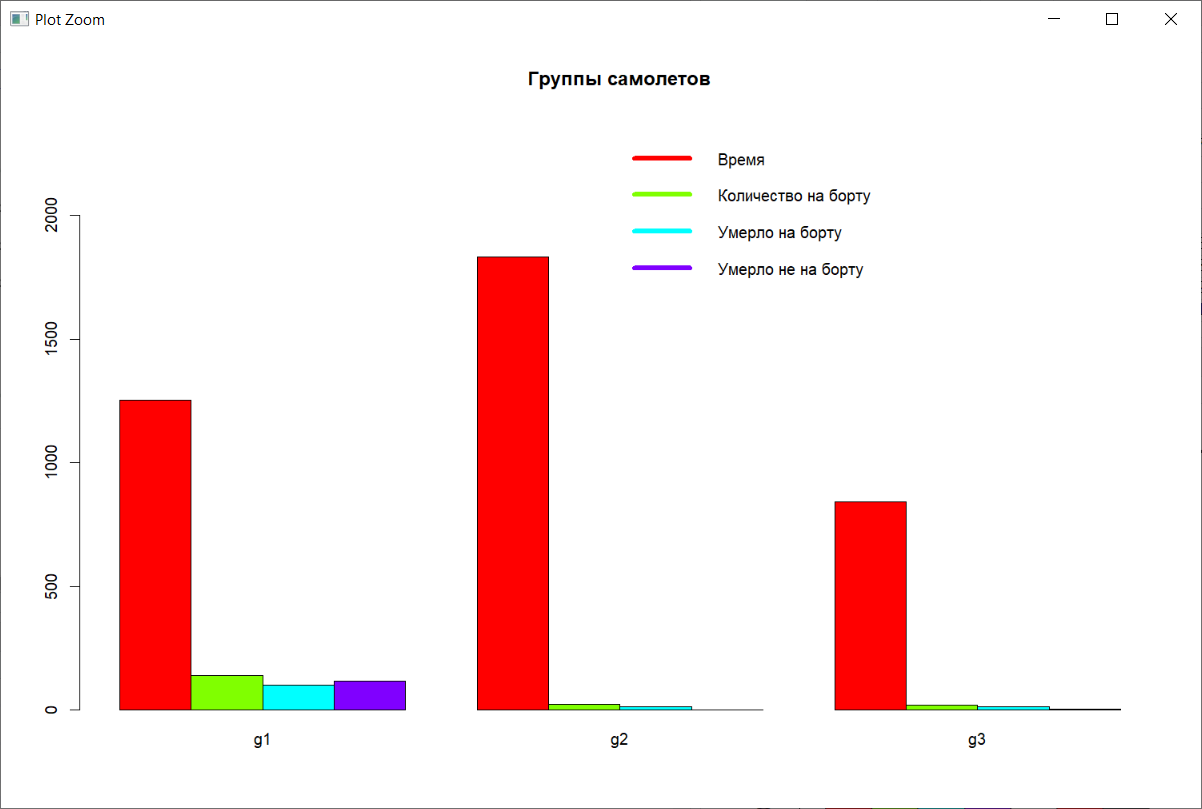


Рис. 4 - Группы крушений с легендой.

Далее была построена диаграмма каменной осыпи, исходя из который можно сделать вывод, что данные можно было разбить и на 3 и на 5 групп.

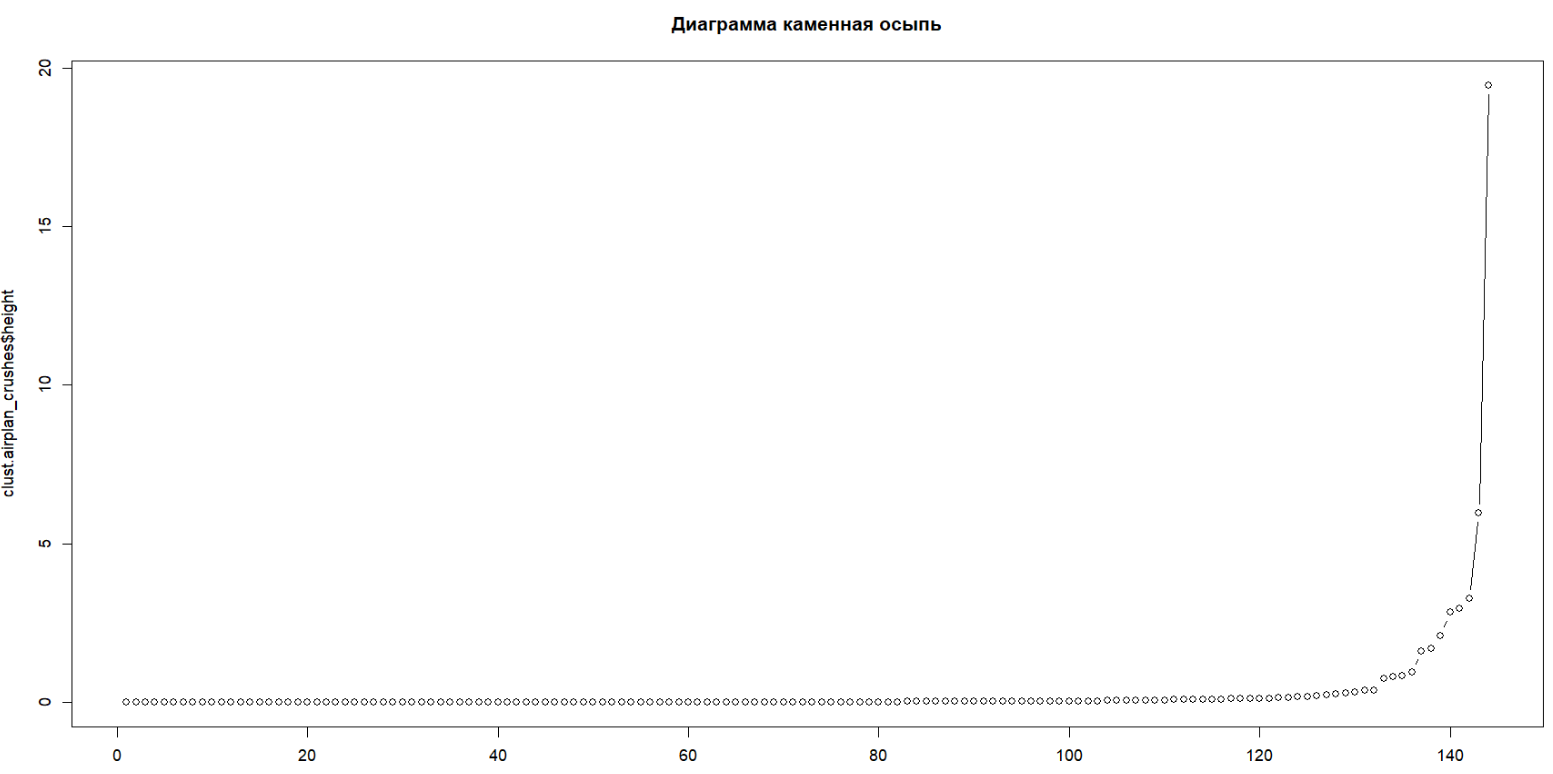


Рис. 5 – Каменная осыпь.

После была выявлена с помощью диаграммы рассеивания, зависимость количества людей на борту от летальных случаев на борту (рисунок 6) и построен boxplot времени крушений у выявленных групп (рисунок 7).

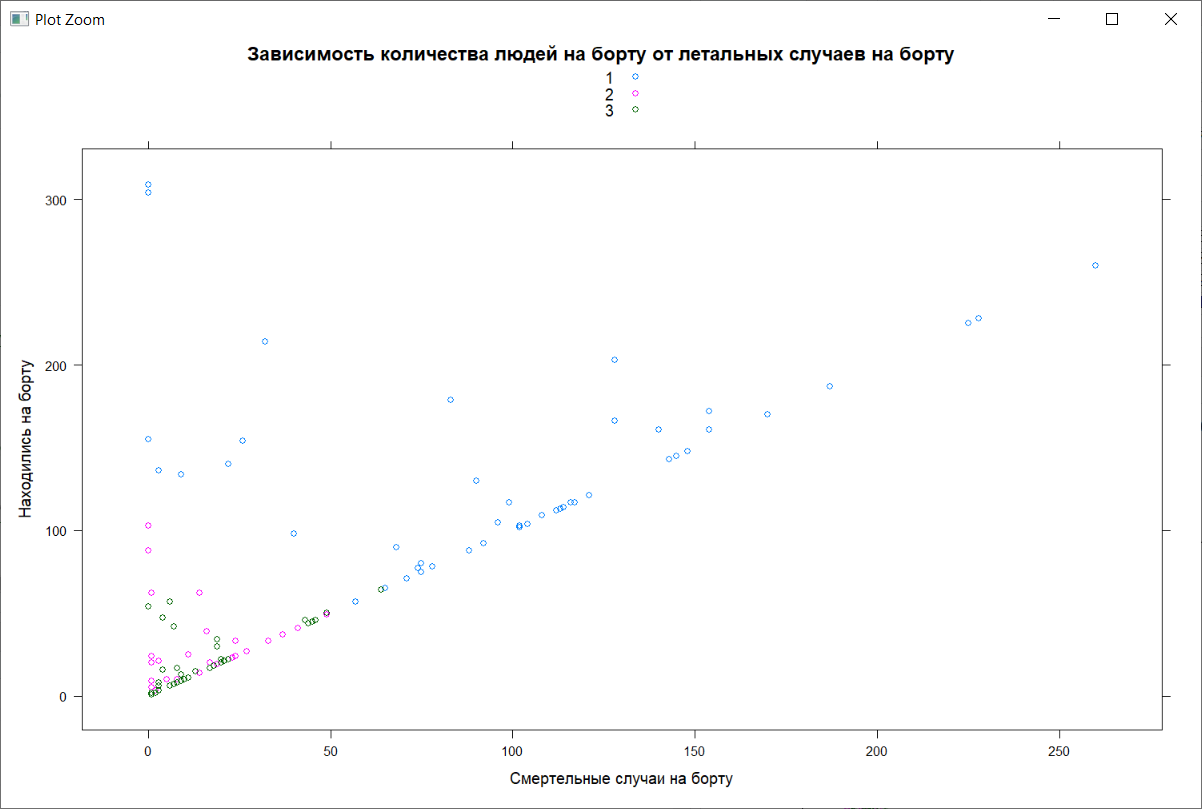


Рис. 6 – Диаграмма рассеивания.

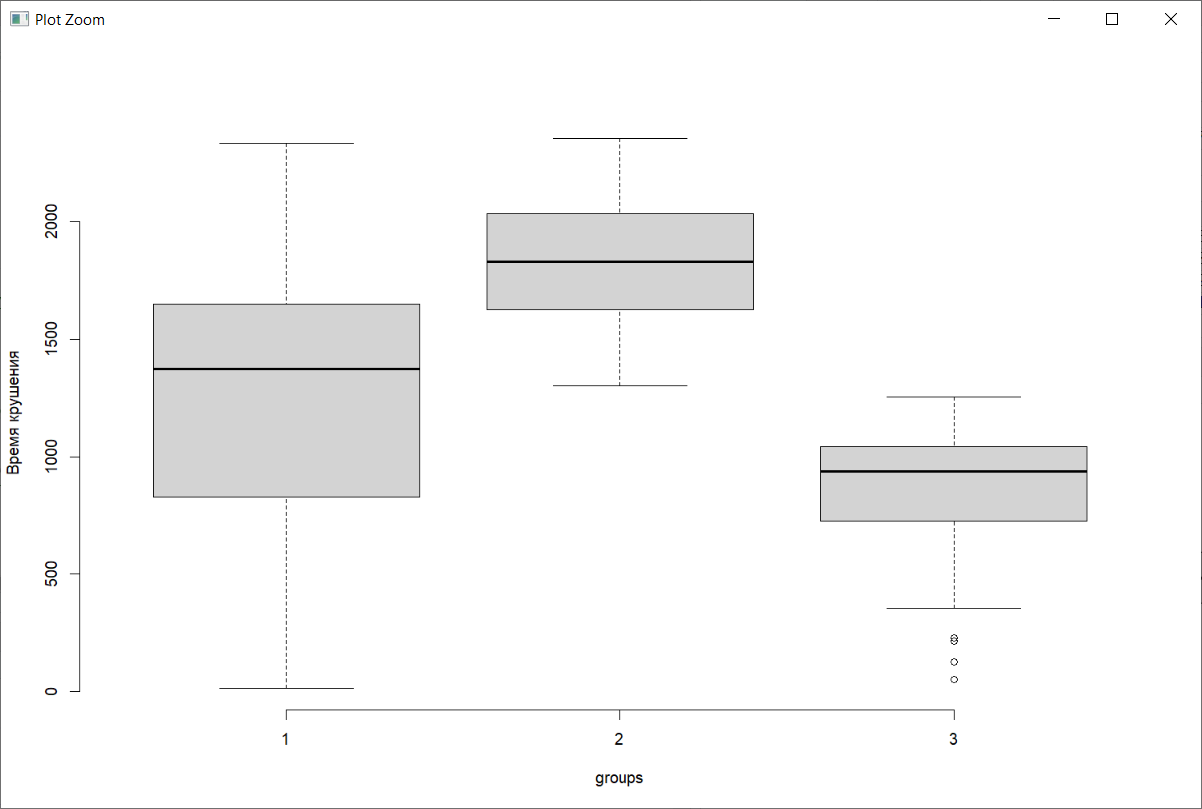


Рис. 7 – boxplot времени крушения.

Явно видна зависимость от времени крушения в каждой группе.

Cоставим по наивному Байесовскому подходу таблицы вероятностей по всем признакам (рисунок 8).

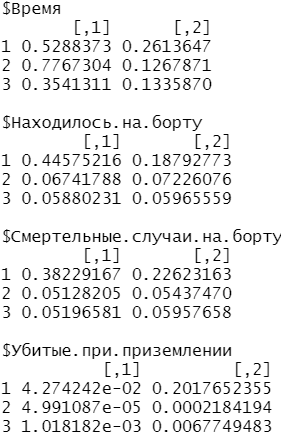


Рис. 8 - Средние значения параметров (первый столбец) и их стандартные отклонения (второй столбец) для каждого выделенного класса (фрагмент таблицы).

Для визуальной сравнительной оценки связи измеренных переменных с метками классов удобно рассмотреть ядерные функции плотности условной вероятности (рисунок 9).

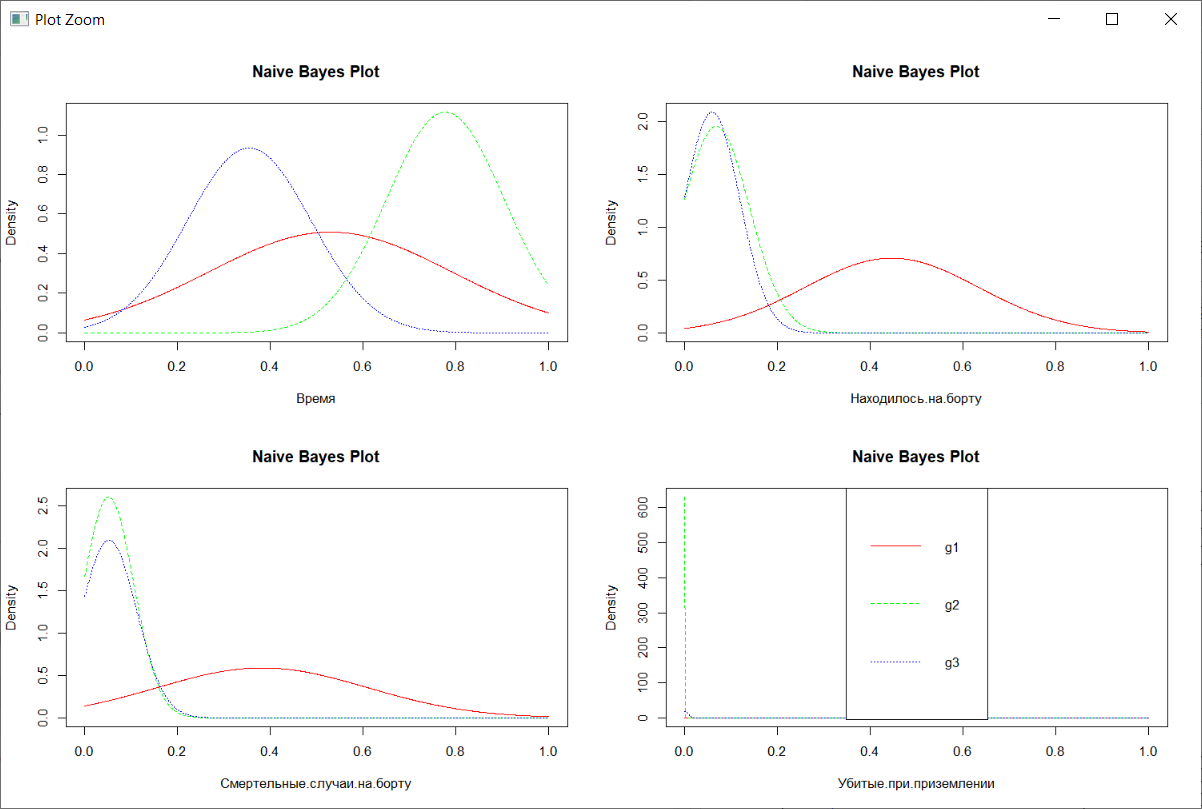


Рис. 9 - Ядерные функции плотности условной вероятности для таблицы naive\_airplan\_crushes (рисунок 9).

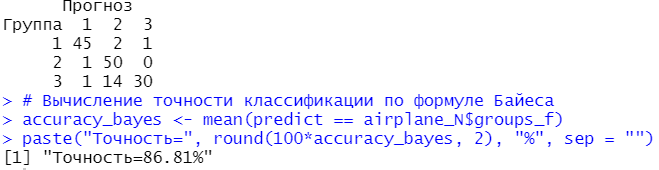


Рис. 10 - Соотношение фактического состояния и прогноза для таблицы naive\_airplan\_crushes.

Рассмотрим классификацию Decision Tree. Мы разбиваем исходные данные на две выборки обучающую и тестовую. Далее, используя логическое индексирование. разобьем исходные данные. **Для построения модели сначала необходимо задать какой параметр (целевой признак) мы хотим предсказать и от каких признаков он зависит.**

Следующим шагом строим модель, используя функцию ***ctree()***, передав ей в качестве параметра нашу формулу и выборку для обучения модели (в нашем случае это trainData).

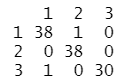


Рис. 11 – Обучение модели

Для применения модели используется функция **predict.** Далее строится так называемая ***Confusion Matrix*** (матрица ошибок или путаницы, если переводить дословно).

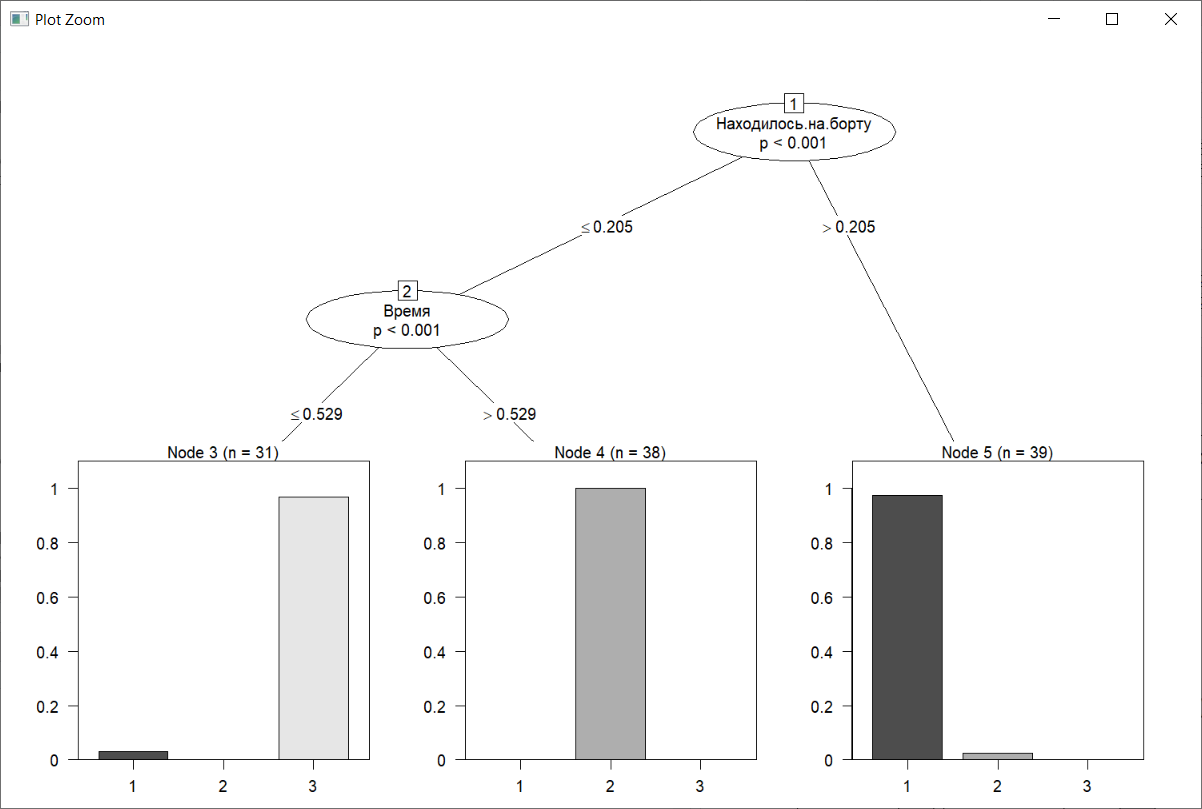


Рис. 12 – Классификация крушения с помощью двоичного дерева (стадия обучения).

Чтобы применить ***обученную модель*** для произвольной выборки, необходимо использовать функцию **predict** с параметром **newdata**.

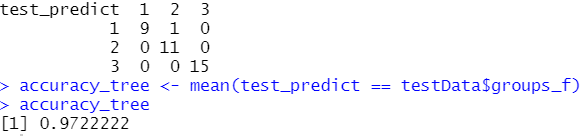


Рис. 13 - Классификация крушений с помощью двоичного дерева.

Рассмотрим алгоритм случайного леса. Главная идея алгоритма - обучение ансамбля в действии. Алгоритм Random Forest потому и называется "Случайный Лес", что для полученных данных он создает множество деревьев приятия решений и потом усредняет результат их предсказаний. Важным моментом тут является элемент случайности в создании каждого дерева - если мы создадим много одинаковых деревьев с разной точностью, то результат их усреднения будет обладать точностью одного дерева.

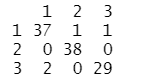


Рис. 14 – Обучение модели с помощью случайного леса.

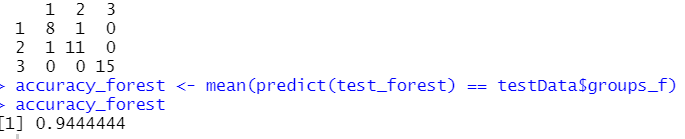


Рис. 15 – Применение на тестовой выборке.

Вывод: Закрепил знания, об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научился визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений. Выявил точность для каждого метода и выяснил, что деревья классифицирует более точно чем Байесовский подход.