УДК 004

Использование Google Colab для статистического анализа на языке R

Романов Даниил Алексеевич
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Целью данного исследования является проведение статистического анализа датасета Titanic Dataset на языке R в Google Colab. Для достижения этой цели были использованы методы визуализации данных и агрегации данных. Из результатов исследования можно сделать выводы о характеристиках пассажиров на Титанике, их выживаемости и зависимости между возрастом и стоимостью билета.
Ключевые слова: Google Colab, R, Titanic Dataset, статистический анализ

Using Google Colab for statistical analysis in R

Romanov Daniil Alekseevich
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this study is to conduct a statistical analysis of the Titanic Dataset in the R language in Google Colab. To achieve this goal, data visualization and data aggregation methods were used. From the results of the study, conclusions can be drawn about the characteristics of passengers on the Titanic, their survival rate and the relationship between age and ticket price.
Keywords: Google Colab, R, Titanic Dataset, statistical analysis

1 Введение

1.1 Актуальность

Датасет Titanic Dataset является одним из наиболее известных и используемых наборов данных в машинном обучении и статистике. Исследование этого датасета позволяет лучше понять характеристики пассажиров на Титанике, их выживаемость и взаимосвязи между различными параметрами. Это может быть полезно для понимания того, какие факторы могут влиять на выживаемость в критических ситуациях.

1.2 Обзор исследований

А. Г. Буховцев посвятил своё исследование описанию методов и инструментов статистического анализа данных в системе R. В работе приводятся примеры использования R для анализа различных типов данных и описываются основные функции и возможности этой системы. Статья может

быть полезна для статистиков и исследователей, использующих R при работе с данными [1].

В. С. Мамедов провёл анализ средств языка программирования R. Он описал основные функции и возможности языка, а также привёл примеры использования R для анализа данных. Статья может быть полезна как начинающим, так и опытным пользователям R при работе с данными [2].

Исследование А. В. Золотарюка посвящено языку программирования R и его среде. В работе описываются основные функции и возможности языка, а также рассматриваются особенности работы с данными в R. Исследование может быть полезно для начинающих пользователей R при изучении языка и его применении для анализа данных [3].

1.3 Цель исследования

Целью данного исследования является проведение статистического анализа датасета Titanic Dataset на языке R в Google Colab. Это позволит более подробно изучить характеристики пассажиров на Титанике, их выживаемость и взаимосвязи между различными параметрами.

2 Материалы и методы

Для работы понадобится онлайн среда программирования Google Colab [4] и датасет Titanic Dataset [5].

3 Результаты и обсуждение

Первым шагом программы является установка необходимых библиотек: tidyverse, ggplot2, dplyr и readr. Эти библиотеки используются для работы с данными и построения графиков (рис.1).

```
# Установка необходимых библиотек
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("readr")
# Загрузка библиотек
library(tidyverse)
library(ggplot2)
library(dplyr)
library(readr)
```

Рисунок 1 - Установка и загрузка необходимых библиотек

Далее происходит загрузка датасета Titanic Dataset с помощью функции read_csv из библиотеки readr. Для просмотра первых строк датасета используется функция head, а для получения общей информации о датасете - функция summary. Это является первым шагом в анализе данных и позволяет определить дальнейшие шаги в работе с датасетом. (рис.2).

```
# Загрузка датасета
titanic <- read_csv("/content/titanic.csv")
# Просмотр первых 6 строк датасета
head(titanic)
# Просмотр информации о датасете
summary(titanic)
```

Рисунок 2 - Загрузка датасета и просмотр информации о нём

После загрузки датасета, можно просмотреть информацю о нём (рис.3).

```
   Survived          Pclass          Name                Sex
Min.   :0.0000   Min.   :1.000   Length:887         Length:887
1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
Median :0.0000   Median :3.000   Mode  :character   Mode  :character
Mean   :0.3856   Mean   :2.306
3rd Qu.:1.0000   3rd Qu.:3.000
Max.   :1.0000   Max.   :3.000
     Age        Siblings/Spouses Aboard Parents/Children Aboard
Min.   : 0.42   Min.   :0.0000          Min.   :0.0000
1st Qu.:20.25   1st Qu.:0.0000          1st Qu.:0.0000
Median :28.00   Median :0.0000          Median :0.0000
Mean   :29.47   Mean   :0.5254          Mean   :0.3833
3rd Qu.:38.00   3rd Qu.:1.0000          3rd Qu.:0.0000
Max.   :80.00   Max.   :8.0000          Max.   :6.0000
     Fare
Min.   :  0.000
1st Qu.:  7.925
Median : 14.454
Mean   : 32.305
3rd Qu.: 31.137
Max.   :512.329
```

Рисунок 3 – Информация о датасете

Из полученной информации можно сделать следующие выводы: Из результатов выполнения этого кода мы можем сделать следующие выводы:

1. Датасет содержит 887 наблюдений (строк) и 8 переменных (столбцов).

2. В датасете есть 2 переменные типа character - Name и Sex, и 6 переменных типа double - Survived, Pclass, Age, Siblings/Spouses Aboard, Parents/Children Aboard, Fare.

3. Переменная Survived - это бинарная переменная, которая указывает, выжил пассажир (1) или нет (0).

4. Переменная Pclass - это категориальная переменная, которая указывает на класс пассажира (1 - первый класс, 2 - второй класс, 3 - третий класс).

5. Переменная Age - это количественная переменная, которая указывает на возраст пассажира.

6. Переменные Siblings/Spouses Aboard и Parents/Children Aboard - это количественные переменные, которые указывают на количество братьев/сестер/супругов и родителей/детей на борту соответственно.

7. Переменная Fare - это количественная переменная, которая указывает на стоимость билета.

8. Минимальный возраст пассажира составляет 0.42 года, а максимальный - 80 лет.

9. Средний возраст пассажиров был около 29 лет, медиана - 28 лет. На борту были и дети, и пожилые люди.

10. Среднее значение переменной Survived равно 0.3856, что означает, что выжило около 38,56% пассажиров.

11. Средняя стоимость билета составляет 32.305 долларов, а медианная стоимость - 14.454 доллара.

Для построения гистограммы возраста пассажиров используется функция ggplot из библиотеки ggplot2. В этой функции задаются параметры для построения графика: x = Age (возраст), binwidth = 5 (ширина интервала гистограммы), fill = "blue" (цвет заливки столбцов) и color = "white" (цвет границ столбцов). Заголовок и подписи осей добавляются с помощью функций labs и theme (рис.4).

```
# Построение гистограммы возраста пассажиров
ggplot(titanic, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "white") +
  labs(title = "Гистограмма возраста пассажиров", x = "Возраст", y = "Количество")
```

Рисунок 4 - Построение гистограммы возраста пассажиров

Из гистограммы видно что большинство пассажиров возраста 20 - 30 лет (рис.5).

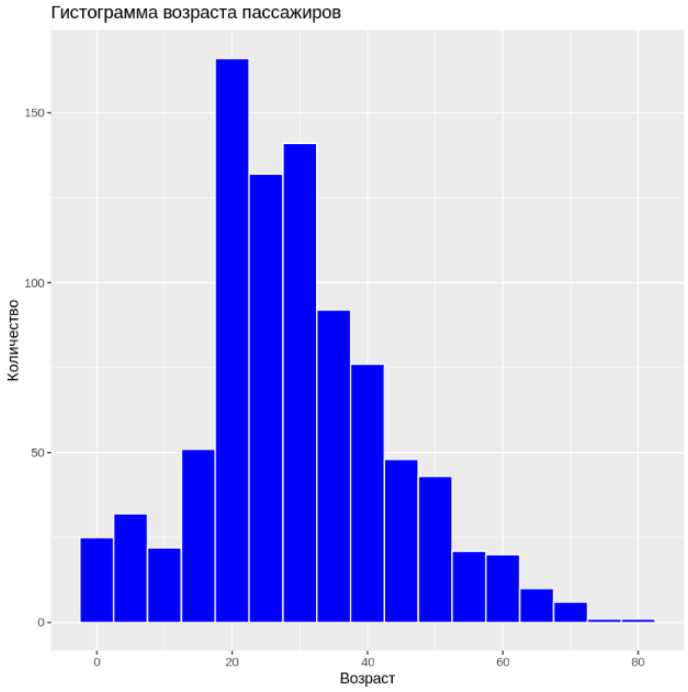

Рисунок 5 - Гистограмма возраста пассажиров

Для построения графика выживших и погибших пассажиров по классу используется функция ggplot из библиотеки ggplot2. В этой функции задаются параметры для построения графика: x = Pclass (класс), fill = factor(Survived) (фактор выживания), position = "dodge" (расположение столбцов). Заголовок и подписи осей добавляются с помощью функций labs и theme. Цвета заливки столбцов и соответствующие им метки задаются с помощью функции scale_fill_manual (рис.6).

```
# Построение графика выживших и погибших пассажиров по классу
ggplot(titanic, aes(x = Pclass, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  labs(title = "Выжившие и погибшие пассажиры по классу", x = "Класс", y = "Количество") +
  scale_fill_manual(values = c("#FF0000", "#00FF00"), labels = c("Погибшие", "Выжившие"))
```

Рисунок 6 - Построение графика выживших и погибших пассажиров по классу

По графику видно что в третьем классе погибших почти вдвое больше чем выживших. В первом классе другая ситуация, выживших больше чем погибших (рис.7).

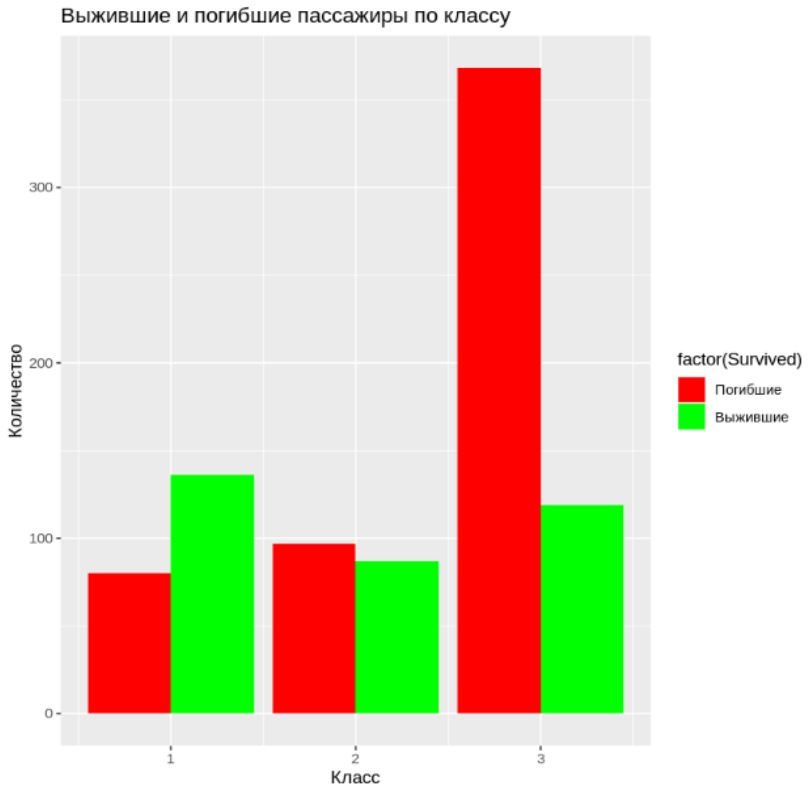

Рисунок 7 - График выживших и погибших пассажиров по классу

Для построения диаграммы рассеяния возраста и стоимости билета используется функция ggplot из библиотеки ggplot2. В этой функции задаются параметры для построения графика: x = Age (возраст), y = Fare (стоимость билета), size = 2 (размер точек), color = "blue" (цвет точек). Заголовок и подписи осей добавляются с помощью функций labs и theme (рис.8).

```
# Построение диаграммы рассеяния возраста и стоимости билета
ggplot(titanic, aes(x = Age, y = Fare)) +
  geom_point(size = 2, color = "blue") +
  labs(title = "Диаграмма рассеяния возраста и стоимости билета", x = "Возраст", y = "Стоимость билета")
```

Рисунок 8 - Построение диаграммы рассеяния

После выполнения данного участка кода получаем следующую диаграмму (рис.9).

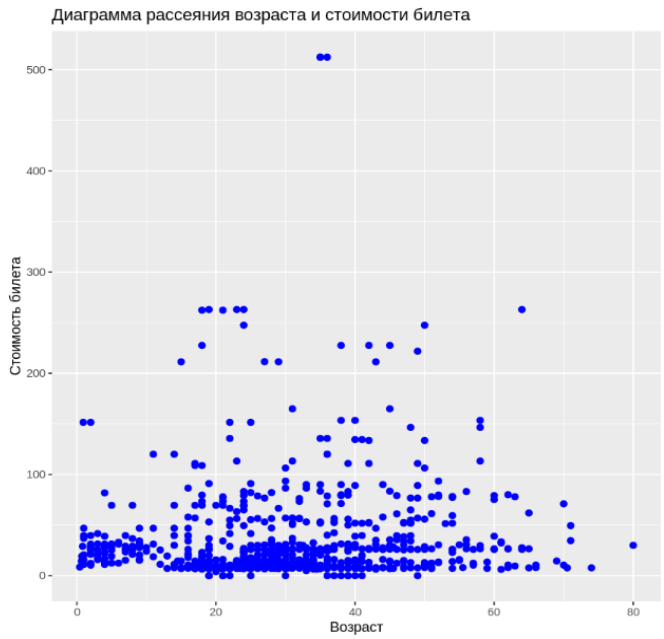

Рисунок 9 - Диаграмма рассеяния возраста и стоимости билета

Для подсчета средней стоимости билета для каждого класса используется функция group_by из библиотеки dplyr. Сначала происходит группировка по классу (Pclass), а затем выполняется агрегация с помощью функции summarize, которая подсчитывает среднее значение стоимости билета (mean_fare) (рис.10).

```
# Подсчет средней стоимости билета для каждого класса
titanic %>%
  group_by(Pclass) %>%
  summarize(mean_fare = mean(Fare))
```

Рисунок 10 - Подсчёт средней стоимости билета для каждого класса

Результатом выполнения функции group_by является следующая таблица (рис.11).

A tibble: 3 × 2

Pclass	mean_fare
<dbl>	<dbl>
1	84.15469
2	20.66218
3	13.70771

Рисунок 11 - Таблица средней стоимости билетов для каждого класса

Используем пакет ggplot2, предоставляемый пакетом tidyverse, для построения графика выживаемости пассажиров на Титанике. Создаем функцию ggplot(), чтобы указать, какую таблицу и какие переменные нужно использовать для создания графика, а затем добавляем геометрию с помощью функции geom_bar(), чтобы построить гистограмму. Для добавления осей и заголовка графика используем функцию labs() (рис.12).

```
# Постороение графика выживаемости пассажиров
ggplot(titanic, aes(x = factor(Survived))) +
  geom_bar() +
  labs(title = "Выживаемость пассажиров", x = "Выжил/не выжил", y = "Количество")
```

Рисунок 12 – Построение графика выживаемости пассажиров

Результатом выполнения функции будет являться график, который визуально отражает количество погибших и выживших (рис.13).

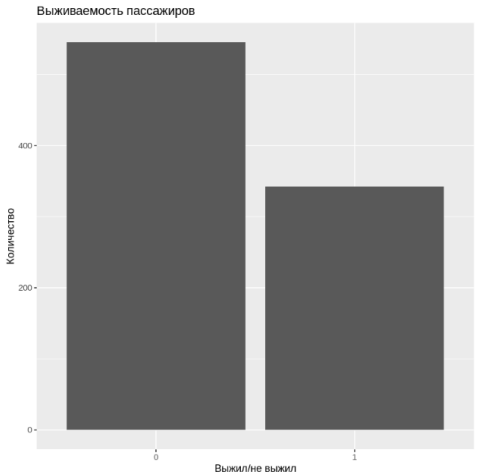

Рисунок 13 - График выживаемости пассажиров

С помощью функции ggplot() строим график зависимости выживаемости от возраста. Здесь используется функция geom_histogram(), чтобы построить гистограмму значений возраста. Затем группируем данные по выживаемости, используя функцию fill = factor(Survived), чтобы не выжившие и выжившие отображались разными цветами (рис.14).

```
# Построение графика зависимости выживаемости от возраста
ggplot(titanic, aes(x = Age, fill = factor(Survived))) +
  geom_histogram(binwidth = 5) +
  labs(title = "Выживаемость в зависимости от возраста", x = "Возраст", y = "Количество") +
  scale_fill_manual(values = c("red", "green"), name = "Выжил/не выжил")
```

Рисунок 14 – Построение графика зависимости выживаемости от возраста

Результатом выполнения данного куска кода будет график, отражающий взаимосвязь между возрастом и выживаемостью. Из этого графика видно, что большинство выживших были в возрастной категории от 20 до 30 лет, но и погибших в этой категории было больше всего (рис.15).

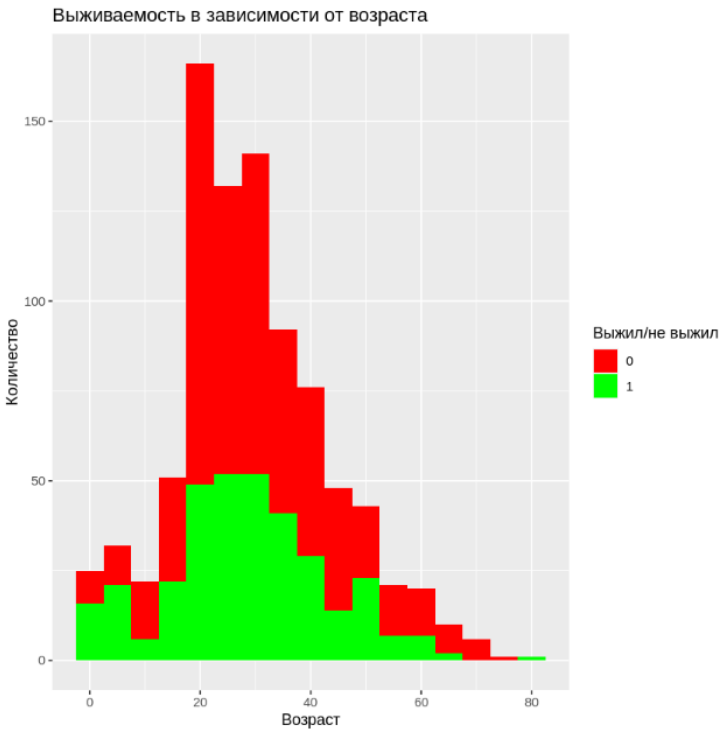

Рисунок 15 – График зависимости выживаемости от возраста

Построим график зависимости выживаемости от пола, для этого используем ggplot(). Здесь используется geom_bar() с параметром position = "dodge", чтобы бары для выживших и невыживших пассажиров отображались рядом друг с другом. Используя функцию scale_fill_manual(), указываем цвета для выживших и невыживших пассажиров (рис.16).

```
# Построение графика зависимости выживаемости от пола
ggplot(titanic, aes(x = factor(Sex), fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  labs(title = "Выживаемость в зависимости от пола", x = "Пол", y = "Количество") +
  scale_fill_manual(values = c("red", "green"), name = "Выжил/не выжил")
```

Рисунок 16 – Построение графика зависимости выживаемости от пола

В результате получаем график, который отражает зависимость между выживаемостью и полом. По графику видно, что среди женщин, выживших гораздо больше чем среди мужчин (рис.17).

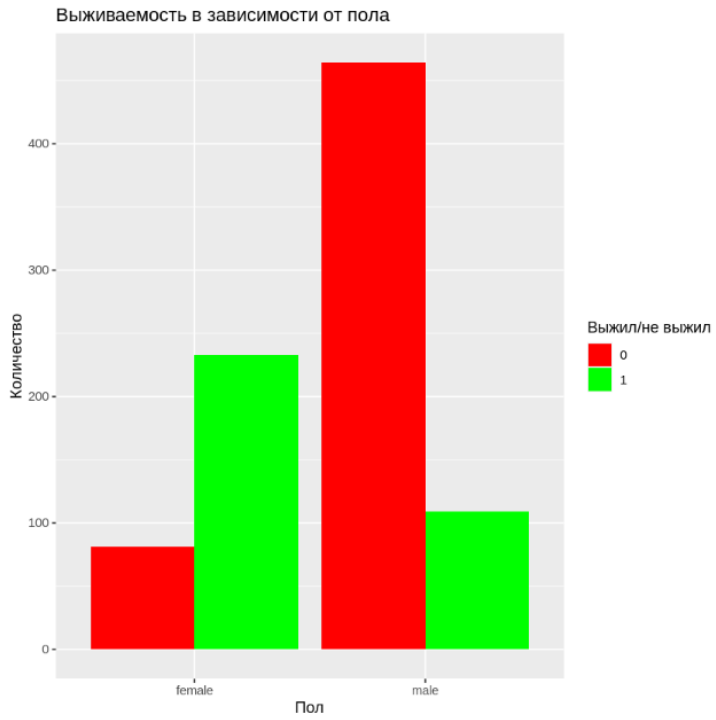

Рисунок 17 – График зависимости выживаемости от пола

Выводы

В результате выполнения программы на языке R в Google Colab был проведен статистический анализ датасета Titanic Dataset. Были использованы методы визуализации данных и агрегации данных. Из результатов анализа были сделаны выводы о характеристиках пассажиров на Титанике, их выживаемости и зависимости между возрастом и стоимостью билета.

В результате выполнения программы, можем получить следующие выводы:

- Большинство пассажиров Титаника путешествовали в 3-м классе.
- Средний возраст пассажиров был около 29 лет, медиана - 28 лет. На борту были и дети, и пожилые люди.
- Всего на борту Титаника находилось 891 пассажир, из них большая часть погибших чем выживших.
- Женщины имели более высокий процент выживаемости, чем мужчины.
- Пассажиры первого класса имели более высокий процент выживаемости, чем пассажиры второго и третьего классов.
- Среди выживших пассажиров было больше женщин и пассажиров первого класса, чем среди погибших пассажиров.

Эти выводы могут быть полезными для дальнейшего анализа данных и выявления закономерностей. Например, на основе этих выводов можно сделать предположение о том, что на выживаемость пассажиров Титаника влияли такие факторы, как пол, класс и возраст. Также можно использовать

эти выводы для разработки более точной модели выживаемости пассажиров на основе машинного обучения.

Ознакомиться с кодом программы и проверить его работоспособность можно по данной ссылке [6].

Библиографический список

1. Статистический анализ данных в системе R. Учебное пособие / А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г. Воронеж: ВГАУ, 2010. 124 с.
2. Мамедов В.С. Анализ средств языка программирования R // Молодой исследователь Дона. 2017. № 2500-1779. С. 49-54.
3. Золотарюк А. В. Язык и среда программирования R. Учебное пособие. М.: ООО «Научно-издательский центр ИНФРА-М», 2019. 162 с.
4. Google Colab URL: https://colab.research.google.com
5. Titanic Dataset URL: https://drive.google.com/file/d/1dVPOlPkcBVhXOy7vyIHyrnCcVUCkr2XS/view?usp=sharing
6. Код программы URL: https://colab.research.google.com/drive/1nQhQ9NbMUEF_MyK9KG5Q131X6KQVck1q?usp=sharing