Министерство науки и высшего образования Российской Федерации

Новосибирский Государственный технический университет

Кафедра автоматизированных систем управления



**Отчет по лабораторной работе № 2**

**по дисциплине «Методы анализа данных»**

**«Классификация»**

Выполнили

студенты группы АВТ-213

Андриевский В. В.

Боровинская Д. И.

Речкин Е. А.

Преподаватель:

к. т. н. Ганелина Н. Д.

г. Новосибирск 2025 г

**Содержание**

[Описательная статистика. Корреляционный анализ 3](#_Toc209367118)

[Регрессионный анализ 15](#_Toc209367119)

[Дисперсионный анализ 22](#_Toc209367120)

# Описательная статистика. Корреляционный анализ

Цель работы: изучение методов решения задачи классификации.

Среда выполнения: Python.

Описание исходных данных: датасет Bike Sharing - содержит информацию о почасовом и ежедневном количестве арендованных велосипедов в системе Capital bikeshare за период с 2011 по 2012 год, а также соответствующие данные о погоде и сезонах.

Независимая переменная: temp - нормализованная температура в градусах Цельсия (значения получены с помощью формулы (t-t\_min)/(t\_max-t\_min), где t\_min = -8, t\_max = +39).

Зависимая переменная: registered - количество зарегистрированных пользователей.

**Дескриптивный анализ**

|  |  |  |
| --- | --- | --- |
|  | temp | registered |
| Среднее | 0,496987 | 153,7869 |
| Стандартная ошибка | 0,001461 | 1,14813 |
| Медиана | 0,5 | 115 |
| Мода | 0,62 | 4 |
| Стандартное отклонение | 0,192556 | 151,3573 |
| Дисперсия выборки | 0,037078 | 22909,03 |
| Эксцесс | -0,94184 | 2,750018 |
| Асимметричность | -0,00602 | 1,557904 |
| Интервал | 0,98 | 886 |
| Минимум | 0,02 | 0 |
| Максимум | 1 | 886 |
| Сумма | 8637,14 | 2672662 |
| Счет | 17379 | 17379 |

Оценка близости выборок к нормальной: для оценки понадобятся такие описательные статистики, как мода, медиана, средн. значение, эксцесс и коэффициент асимметрии.

Если коэффициент асимметрии и эксцесс равны нулю, то гипотеза о нормальности не отклоняется.

Также для симметричного распределения характерно равенство моды, медианы и средн. значения (а так как нормальное распределение является симметричным, то это же условие характерно и для него).

По результатам дескриптивного анализа мы видим, что:

* для переменной temp значения моды (0,62), медианы (0,5) и средн. значения (0,496987) близки. Коэффициент асимметрии (-0,00602) довольно мал и близок к нулю, однако значение эксцесса (-0,94184) близко к -1. То есть, выборка близка к нормальной, но нельзя явно говорить о том, что она является таковой.
* для переменной registered значения моды (4), медианы (115) и средн. значения (153,7869) довольно различны. Коэффициент асимметрии 1,557904) больше нуля, то есть правое плечо распределения будет длиннее левого. Значение эксцесса (2,750018) также больше нуля, это говорит о том, что гистограмма будет более вытянутая. Значит, данная выборка не является нормальной.

Построение гистограмм: количество интервалов рассчитаем по правилу Стёрджесса: k = 1 + log2(N), где N – общее число наблюдений (17379).

Величина шага (длина интервала): h = (макс. значение – мин. значение) / k

h (temp) = 0,06496495

h (registered) = 58,73361788

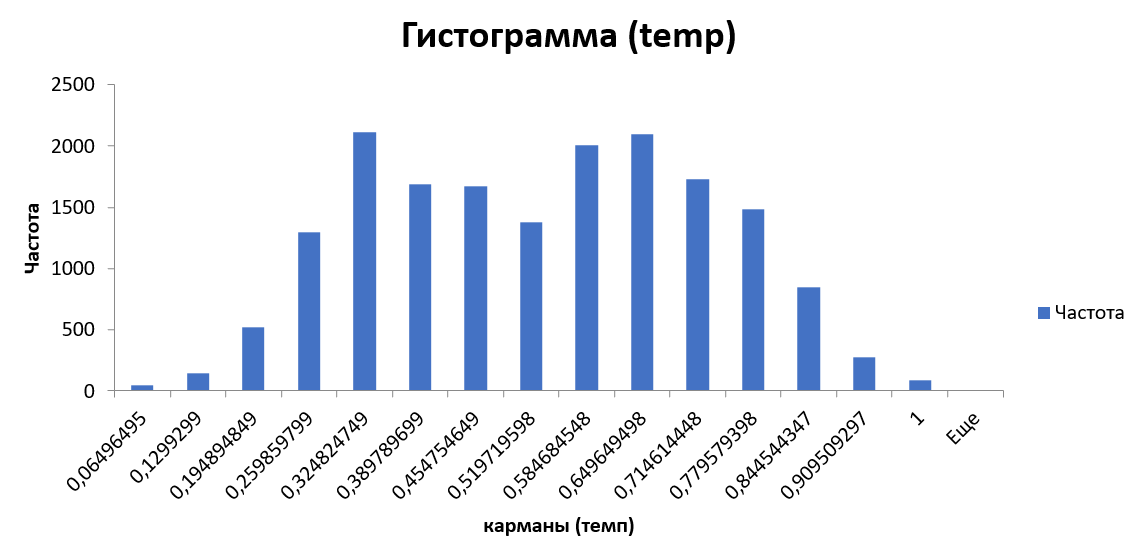


Рисунок 1 – гистограмма для выборки переменной temp

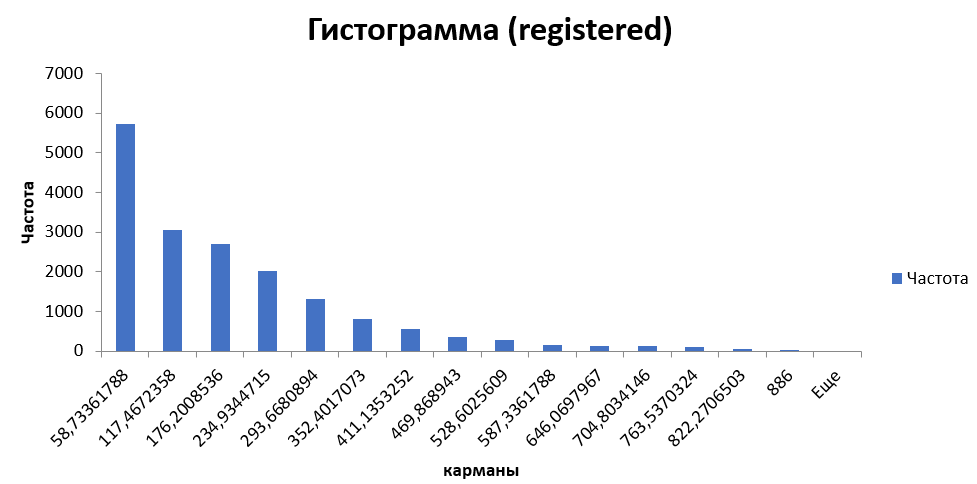


Рисунок 2 – гистограмма для выборки переменной registered

