1. ВВЕДЕНИЕ

Вода является одним из основных природных ресурсов, определяющих смягчение экологических и социальных проблем. Её качество оказывает воздействие на такие сферы, как здравоохранение, промышленность и сельское хозяйство. Качество воды оказывает значительное влияние на здоровье человека, живых организмов и экосистемы в целом. Загрязнение воды может привести к различным заболеваниям. Многие отрасли зависят от качества воды, включая сельское хозяйство, которое требует чистой воды для орошения.

Понимание различных уровней качества воды может помочь в принятии решений по управлению ресурсами. Изменение параметров качества воды требует регулярного мониторинга и анализа.

Таким образом, тема данной курсовой работы является актуальной для многих сфер – от экологии и здравоохранения до промышленности. Использование современных методов анализа данных, таких как иерархическая кластеризация и алгоритм k-means может помочь в понимании структуры и закономерностей данных о качестве воды.

Цель курсовой работы — разработка и реализация модели кластеризации методом иерархической кластеризации и k-means для автоматизации мониторинга качества воды.

Задачи, решаемые в данной курсовой работе:

- Теоретический анализ предметной области: обоснование важности анализа данных и рассмотрение ключевых типов информации, применяемых в исследованиях.
- Анализ данных: сбор и предобработка данных (обработка пропусков, выбросов, дубликатов), исследование взаимосвязей параметров.
- Кластеризация: применение иерархической кластеризации и k-means.
- Визуализация: построение интерактивных графиков.
- Интерпретация результатов: сравнение эффективности методов.

Методы исследования, используемые в работе:

- Исследовательский анализ данных ;
- Методы первичной обработки данных;
- Статистический анализ: описательные статистики;
- Методы машинного обучения: Иерархическая кластеризация, K-means (подбор числа кластеров через "метод локтя" и “метод силуэта”);
- Средства визуализации и отчётности.

В работе использованы инструменты: R (ggplot2, cluster, gridExtra, factoextra), Glarus BI и программные среды: RStudio, Glarus BI.

Курсовая работа состоит из нескольких частей: введение, теоретическая часть, в которой рассматриваются основы предметной области и типы данных. Практическая часть, где проводится анализ данных, строятся модели и визуализируются результаты, заключение, в котором подводятся итоги исследования и предлагаются перспективы для дальнейшей работы.

2. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ИССЛЕДОВАНИЯ

2.1. Описание используемых данных

В практической работе будет проанализировано качество воды.

В качестве исследуемых данных взят набор данных «Water Quality».

Показатели качества вода набора данных «Water Quality»:

1. Значение pH:

PH является важным параметром при оценке кислотно-щелочного баланса воды. Это также показатель кислотного или щелочного состояния воды. ВОЗ рекомендовала максимально допустимый уровень pH от 6,5 до 8,5. Текущие диапазоны исследований составляли 6,52–6,83, что соответствует стандартам ВОЗ.

2. Твердость:

Жесткость в основном обусловлена солями кальция и магния. Эти соли растворяются в геологических отложениях, через которые проходит вода. Продолжительность контакта воды с материалом, создающим жесткость, помогает определить, насколько высока жесткость сырой воды. Первоначально жесткость определялась как способность воды осаждать мыло, обусловленная содержанием кальция и магния.

3. Твердые вещества (Общее количество растворенных твердых веществ - TDS):

Вода обладает способностью растворять широкий спектр неорганических и некоторых органических минералов или солей, таких как калий, кальций, натрий, бикарбонаты, хлориды, магний, сульфаты и т.д. Эти минералы придают воде нежелательный вкус и разбавленный цвет. Это важный параметр при использовании воды. Вода с высоким значением TDS указывает на высокую минерализацию. Желаемый предел для TDS

составляет 500 мг / л, а максимальный предел составляет 1000 мг / л, который предписан для употребления в пищу.

4. Хлорамины:

Хлор и хлорамин являются основными дезинфицирующими средствами, используемыми в системах общественного водоснабжения. Хлорамины чаще всего образуются при добавлении аммиака к хлору для обработки питьевой воды. Содержание хлора в питьевой воде до 4 миллиграммов на литр (мг/л или 4 частей на миллион (ppm)) считается безопасным.

5. Сульфат:

Сульфаты - это природные вещества, которые содержатся в минералах, почве и горных породах. Они присутствуют в окружающем воздухе, грунтовых водах, растениях и продуктах питания. Основное коммерческое применение сульфатов - в химической промышленности. Концентрация сульфатов в морской воде составляет около 2700 миллиграммов на литр (мг/л). В большинстве источников пресной воды он колеблется от 3 до 30 мг / л, хотя в некоторых географических точках наблюдаются гораздо более высокие концентрации (1000 мг /л).

6. Проводимость:

Чистая вода не является хорошим проводником электрического тока, скорее это хороший изолятор. Увеличение концентрации ионов увеличивает электропроводность воды. Как правило, количество растворенных твердых веществ в воде определяет электропроводность. Электрическая проводимость (EC) фактически измеряет ионный процесс раствора, который позволяет ему передавать ток. Согласно стандартам ВОЗ, значение EC не должно превышать 400 МКС / см.

7. Органический углерод:

Общий органический углерод (TOC) в исходных водах поступает из разлагающихся природных органических веществ (NOM), а также из синтетических источников. TOC - это показатель общего количества углерода в органических соединениях в чистой воде. Согласно US EPA < 2 мг / л в

качестве TOC в очищенной / питьевой воде и <4 мг / л в исходной воде, которая используется для лечения.

8. Тригалометаны:

ТГМ - это химические вещества, которые могут быть обнаружены в воде, обработанной хлором. Концентрация ТГМ в питьевой воде варьируется в зависимости от уровня содержания органических веществ в воде, количества хлора, необходимого для очистки воды, и температуры обрабатываемой воды. Уровень ТГМ до 80 промилле считается безопасным в питьевой воде.

9. Мутность:

Мутность воды зависит от количества твердого вещества, присутствующего во взвешенном состоянии. Это показатель светоизлучающих свойств воды, и тест используется для определения качества сбрасываемых отходов по отношению к коллоидному веществу. Среднее значение мутности, полученное для кампуса Wondo Genet (0,98 NTU), ниже рекомендованного ВОЗ значения в 5,00 NTU.

10. Пригодность для питья:

Указывает, безопасна ли вода для потребления человеком, где 1 означает Пригодную для питья, а 0 означает Непригодную для питья.

2.2. Исследовательский анализ данных (EDA)

Для анализа в R был импортирован датасет (Рисунок 5). На Рисунке 6 отображена его структура.

```
> df <- read.csv("D:/Файлы_учёба/4_семестр/ЯПСОД/курсовая/water_potability.csv", header=TRUE)
> head(df)
        ph Hardness   Solids Chloramines  Sulfate Conductivity
1       NA 204.8905 20791.32    7.300212 368.5164     564.3087
2 3.716080 129.4229 18630.06    6.635246       NA     592.8854
3 8.099124 224.2363 19909.54    9.275884       NA     418.6062
4 8.316766 214.3734 22018.42    8.059332 356.8861     363.2665
5 9.092223 181.1015 17978.99    6.546600 310.1357     398.4108
6 5.584087 188.3133 28748.69    7.544869 326.6784     280.4679
  Organic_carbon Trihalomethanes Turbidity Potability
1      10.379783        86.99097  2.963135          0
2      15.180013        56.32908  4.500656          0
3      16.868637        66.42009  3.055934          0
4      18.436524       100.34167  4.628771          0
5      11.558279        31.99799  4.075075          0
6       8.399735        54.91786  2.559708          0
```

Рисунок 5 – Импортированный датасет

```
> str(df)
'data.frame':   3276 obs. of  10 variables:
 $ ph              : num  NA 3.72 8.1 8.32 9.09 ...
 $ Hardness        : num  205 129 224 214 181 ...
 $ Solids          : num  20791 18630 19910 22018 17979 ...
 $ Chloramines     : num  7.3 6.64 9.28 8.06 6.55 ...
 $ Sulfate         : num  369 NA NA 357 310 ...
 $ Conductivity    : num  564 593 419 363 398 ...
 $ Organic_carbon  : num  10.4 15.2 16.9 18.4 11.6 ...
 $ Trihalomethanes: num  87 56.3 66.4 100.3 32 ...
 $ Turbidity       : num  2.96 4.5 3.06 4.63 4.08 ...
 $ Potability      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Рисунок 6 – Структура датасета

Все столбцы, кроме последнего имеют тип numeric, столбец Potability имеет тип integer. Это свидетельствует о том, что информация представлена в виде чисел, что упрощает её обработку для последующего статистического анализа и моделирования. Спомощью функции is.na() выведена информация о пропусках (Рисунок 7).

```
> colSums(is.na(df))
             ph        Hardness          Solids     Chloramines
            491               0               0               0
        Sulfate    Conductivity  Organic_carbon Trihalomethanes
            781               0               0             162
      Turbidity      Potability
              0               0
```

Рисунок 7 – Информация о пропусках

Столбцы ph, Sulfate и Trihalomethanes имеют пропуски, которые необходимо удалить, чтобы обеспечить чистоту и точность данных для анализа. Кроме того, следует удалить дубликаты. Процесс очистки данных показан на Рисунке 8.

```
> df<-df[!duplicated(df),]
> df$ph[is.na(df$ph)]<-mean(df$ph, na.rm = TRUE)
> df$Trihalomethanes[is.na(df$Trihalomethanes)]<-mean(df$Trihalomethanes, na.rm = TRUE)
> df$Sulfate[is.na(df$Sulfate)]<-mean(df$Sulfate, na.rm = TRUE)
> df$Potability <- NULL
```

Рисунок 8 – Очистка данных

Результат устранения пропусков представлен на Рисунке 9.

```
> colSums(is.na(df))
             ph        Hardness          Solids     Chloramines         Sulfate    Conductivity
              0               0               0               0               0               0
 Organic_carbon Trihalomethanes       Turbidity
              0               0               0
```

Рисунок 9 – Проверка успешности удаления пропусков

Для повышения качества кластеризации и обеспечения устойчивости алгоритмов к аномальным значениям необходимо удалить выбросы в данных, чтобы они не искажали результаты. Для определения наличия выбросов в данных были построены диаграммы boxplot (ящики с усами) (Рисунки 10-11).

```
> plots <- list()
> for (col in names(df)) {
+   plots[[col]] <- ggplot(df, aes_string(y = col)) +
+     geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
+     ggtitle(paste(col)) +
+     theme_minimal() +
+     ylab("") +
+     xlab("")
+ }
> do.call(grid.arrange, c(plots, ncol = 3))
```

Рисунок 10 – Код диаграмм boxplot

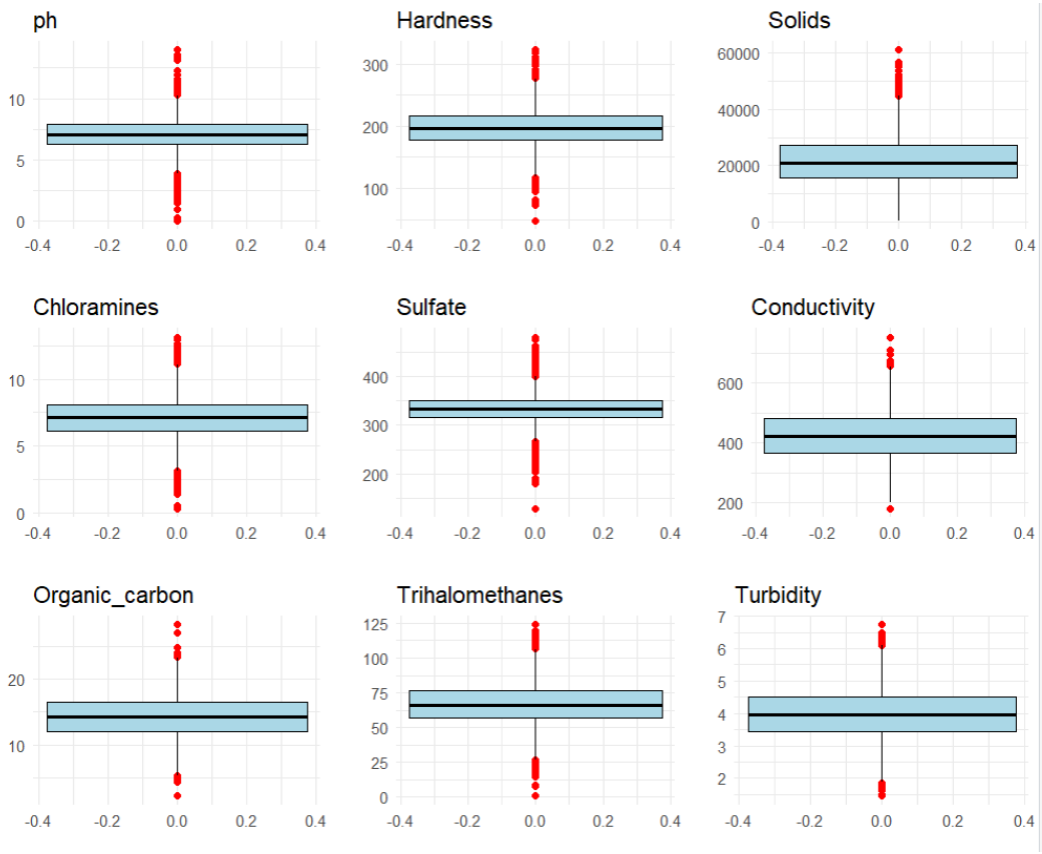


Рисунок 11 – Ящики с усами

Данные содержат выбросы. Так как алгоритм k-means неустойчив к выбросам, удалим их (Рисунки 12-13).

```
> remove_outliers_iqr <- function(df, columns = names(df)) {
+   for (col in columns) {
+     if (is.numeric(df[[col]])) {#
+       Q1 <- quantile(df[[col]], 0.25, na.rm = TRUE)
+       Q3 <- quantile(df[[col]], 0.75, na.rm = TRUE)
+       IQR <- Q3 - Q1
+       lower_bound <- Q1 - 0.5 * IQR
+       upper_bound <- Q3 + 0.5 * IQR
+       df <- df[df[[col]] >= lower_bound & df[[col]] <= upper_bound | is.na(df[[col]]), ]
+     }
+   }
+   return(df)
+ }
> df <- remove_outliers_iqr(df)
```

Рисунок 12 – Удаление выбросов

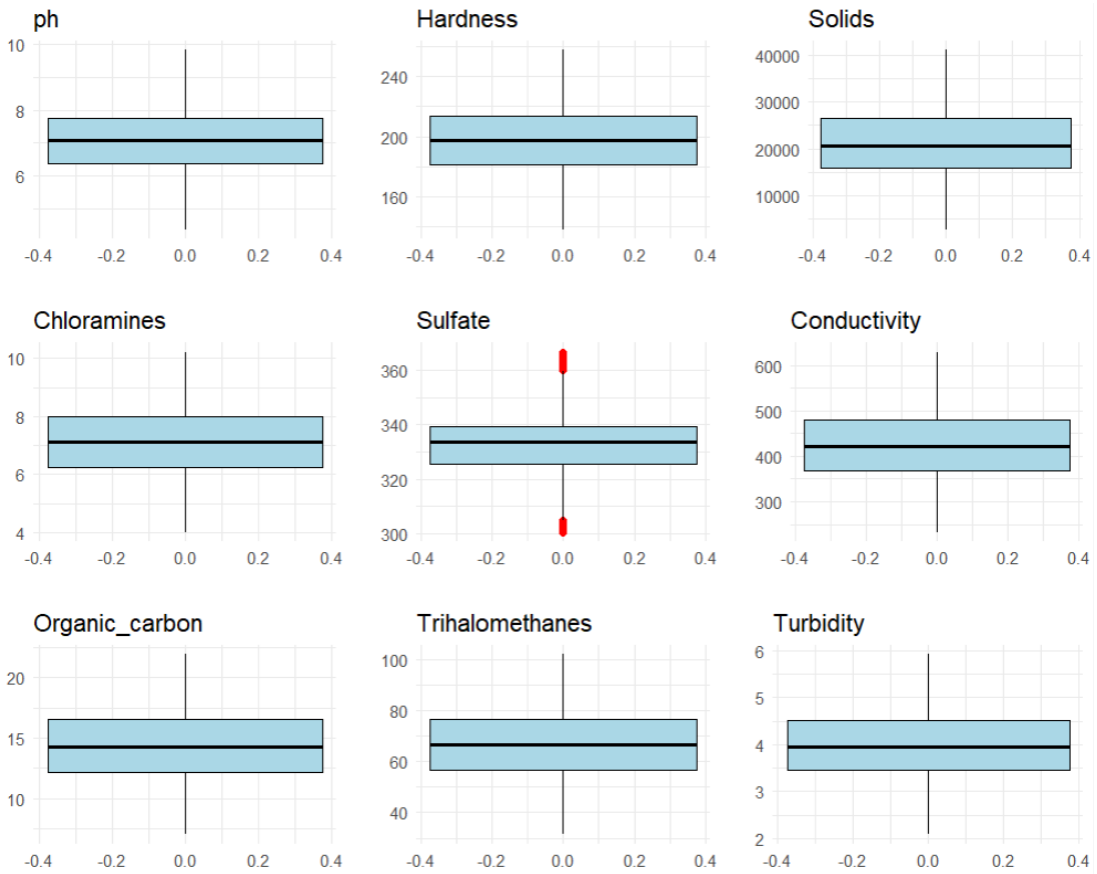


Рисунок 13 – Результат удаления выбросов

В столбце Sulfate не удалось избавиться от всех выбросов, но это не критично, так как вода может иметь такой уровень сульфатов. Выбросы в остальных столбцах удалены успешно.

Гистограммы распределения параметров воды представлены на Рисунке 14.

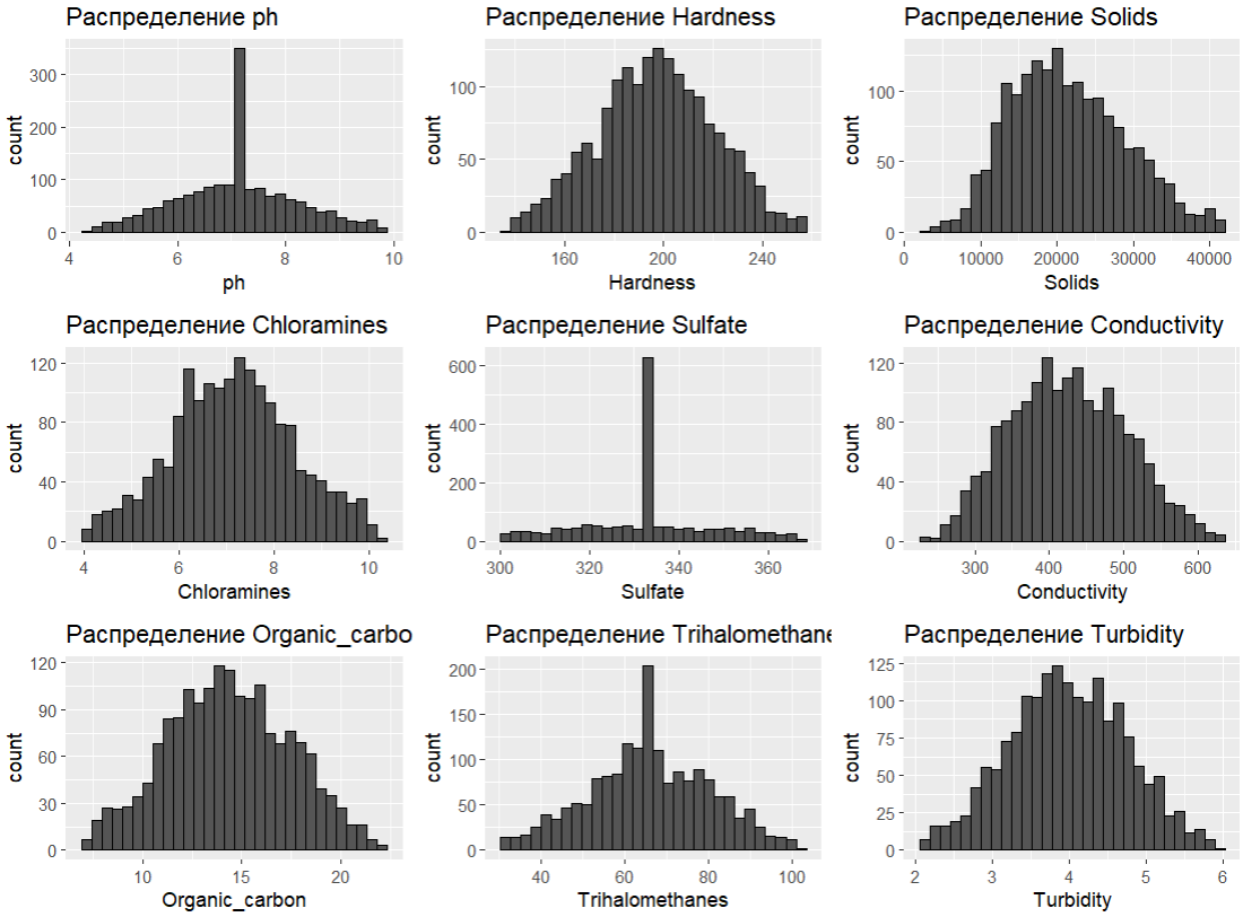
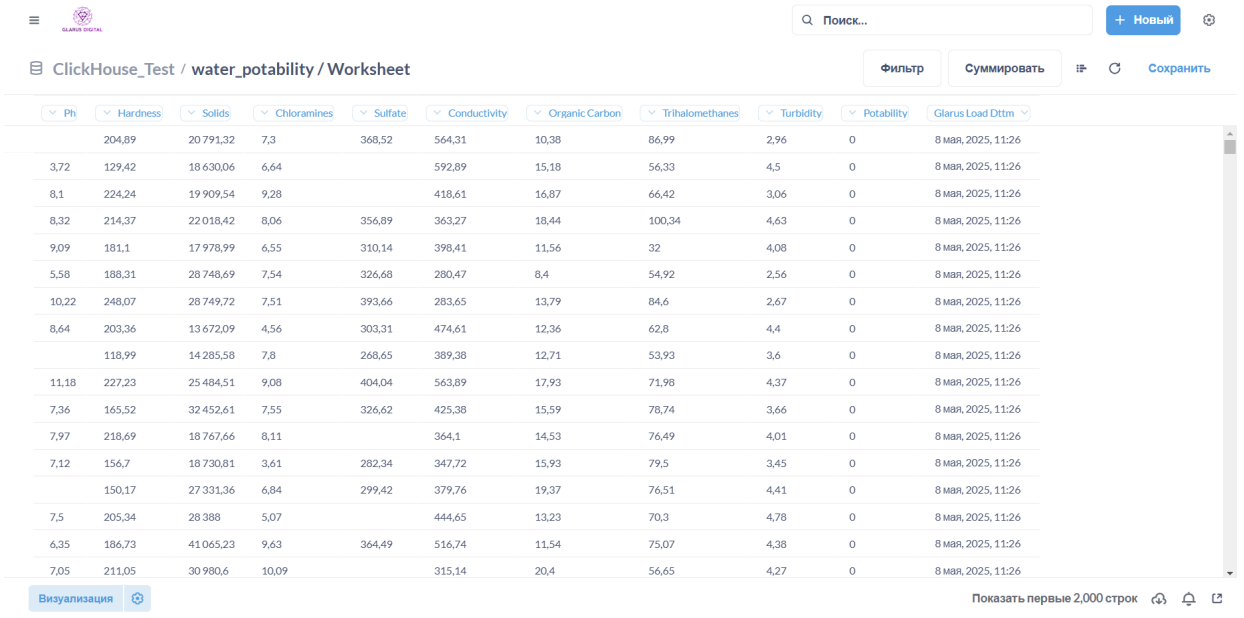


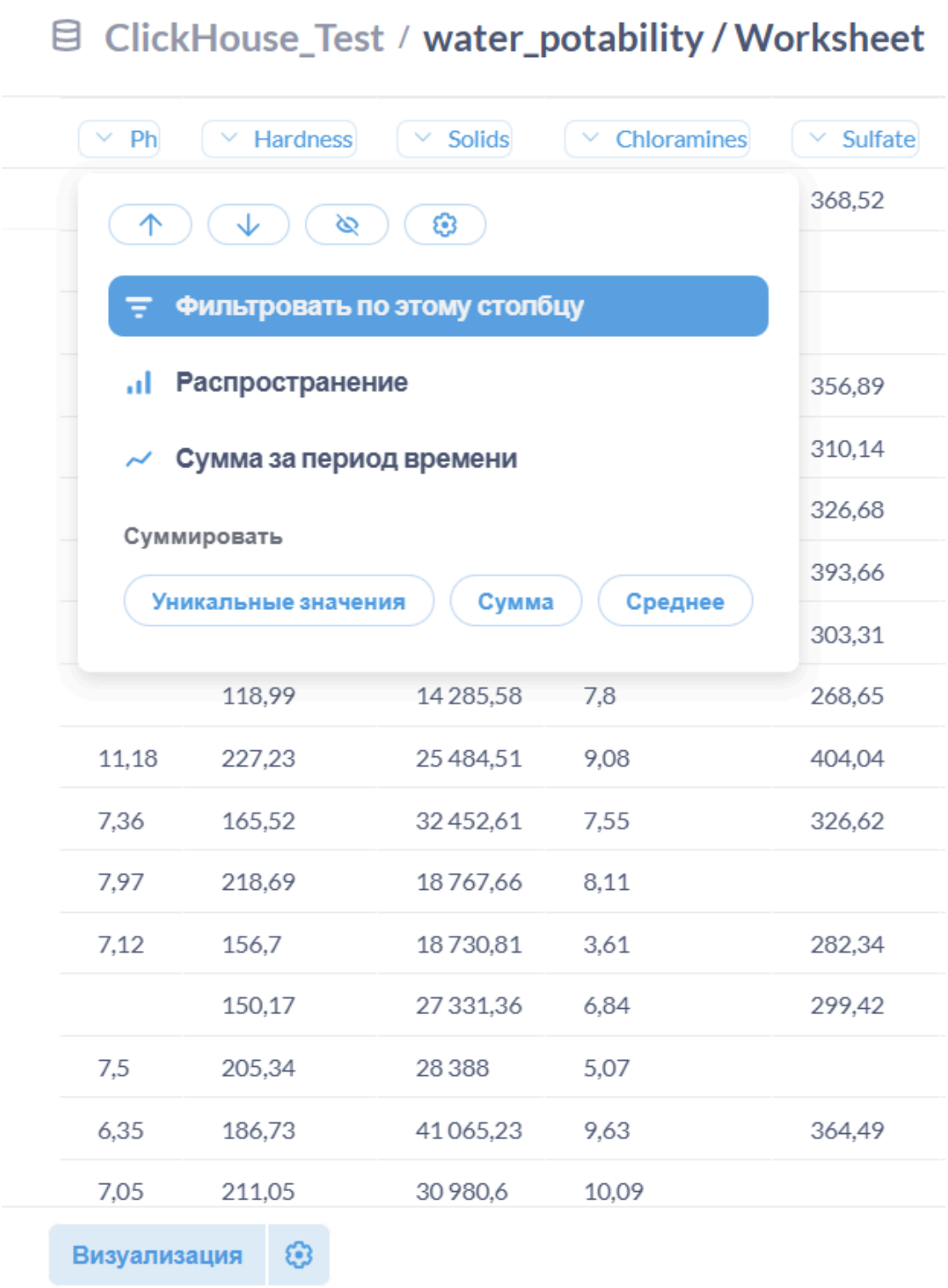
Рисунок 14 – Распределение параметров воды

В Glarus BI также был загружен датасет, удалены пропущенные значения (Рисунки 15-20).



Ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic Carbon	Trihalomethanes	Turbidity	Potability	Glarus Load Dttm
	204,89	20 791,32	7,3	368,52	564,31	10,38	86,99	2,96	0	8 мая, 2025, 11:26
3,72	129,42	18 630,06	6,64		592,89	15,18	56,33	4,5	0	8 мая, 2025, 11:26
8,1	224,24	19 909,54	9,28		418,61	16,87	66,42	3,06	0	8 мая, 2025, 11:26
8,32	214,37	22 018,42	8,06	356,89	363,27	18,44	100,34	4,63	0	8 мая, 2025, 11:26
9,09	181,1	17 978,99	6,55	310,14	398,41	11,56	32	4,08	0	8 мая, 2025, 11:26
5,58	188,31	28 748,69	7,54	326,68	280,47	8,4	54,92	2,56	0	8 мая, 2025, 11:26
10,22	248,07	28 749,72	7,51	393,66	283,65	13,79	84,6	2,67	0	8 мая, 2025, 11:26
8,64	203,36	13 672,09	4,56	303,31	474,61	12,36	62,8	4,4	0	8 мая, 2025, 11:26
	118,99	14 285,58	7,8	268,65	389,38	12,71	53,93	3,6	0	8 мая, 2025, 11:26
11,18	227,23	25 484,51	9,08	404,04	563,89	17,93	71,98	4,37	0	8 мая, 2025, 11:26
7,36	165,52	32 452,61	7,55	326,62	425,38	15,59	78,74	3,66	0	8 мая, 2025, 11:26
7,97	218,69	18 767,66	8,11		364,1	14,53	76,49	4,01	0	8 мая, 2025, 11:26
7,12	156,7	18 730,81	3,61	282,34	347,72	15,93	79,5	3,45	0	8 мая, 2025, 11:26
	150,17	27 331,36	6,84	299,42	379,76	19,37	76,51	4,41	0	8 мая, 2025, 11:26
7,5	205,34	28 388	5,07		444,65	13,23	70,3	4,78	0	8 мая, 2025, 11:26
6,35	186,73	41 065,23	9,63	364,49	516,74	11,54	75,07	4,38	0	8 мая, 2025, 11:26
7,05	211,05	30 980,6	10,09		315,14	20,4	56,65	4,27	0	8 мая, 2025, 11:26

Рисунок 15 – Импортированный датасет



Ph	Hardness	Solids	Chloramines	Sulfate
				368,52
				356,89
				310,14
				326,68
				393,66
				303,31
	118,99	14 285,58	7,8	268,65
11,18	227,23	25 484,51	9,08	404,04
7,36	165,52	32 452,61	7,55	326,62
7,97	218,69	18 767,66	8,11	
7,12	156,7	18 730,81	3,61	282,34
	150,17	27 331,36	6,84	299,42
7,5	205,34	28 388	5,07	
6,35	186,73	41 065,23	9,63	364,49
7,05	211,05	30 980,6	10,09	

Рисунок 16 – Удаление пропусков в столбце ph

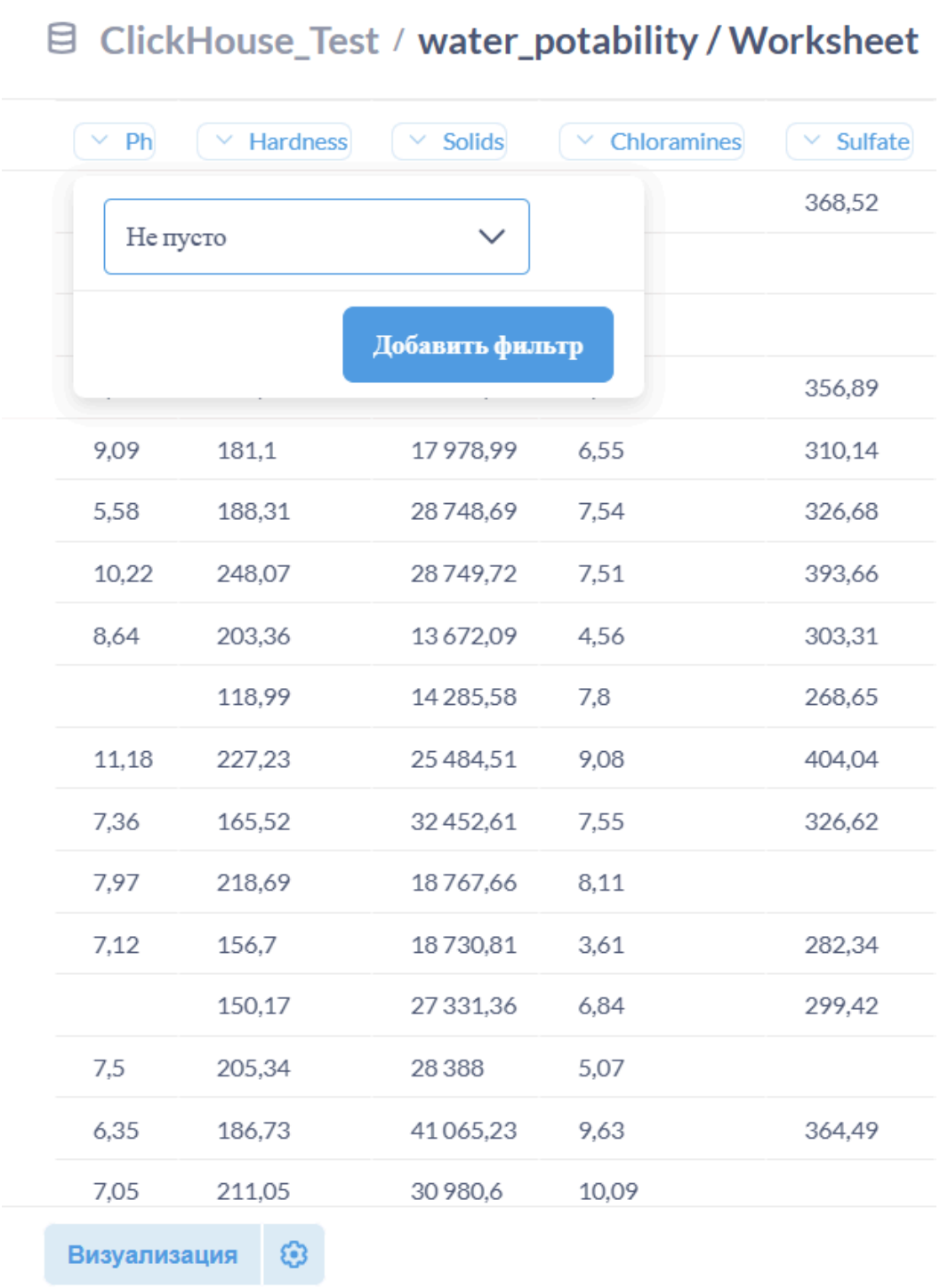


Рисунок 17 – Удаление пропусков в столбце ph

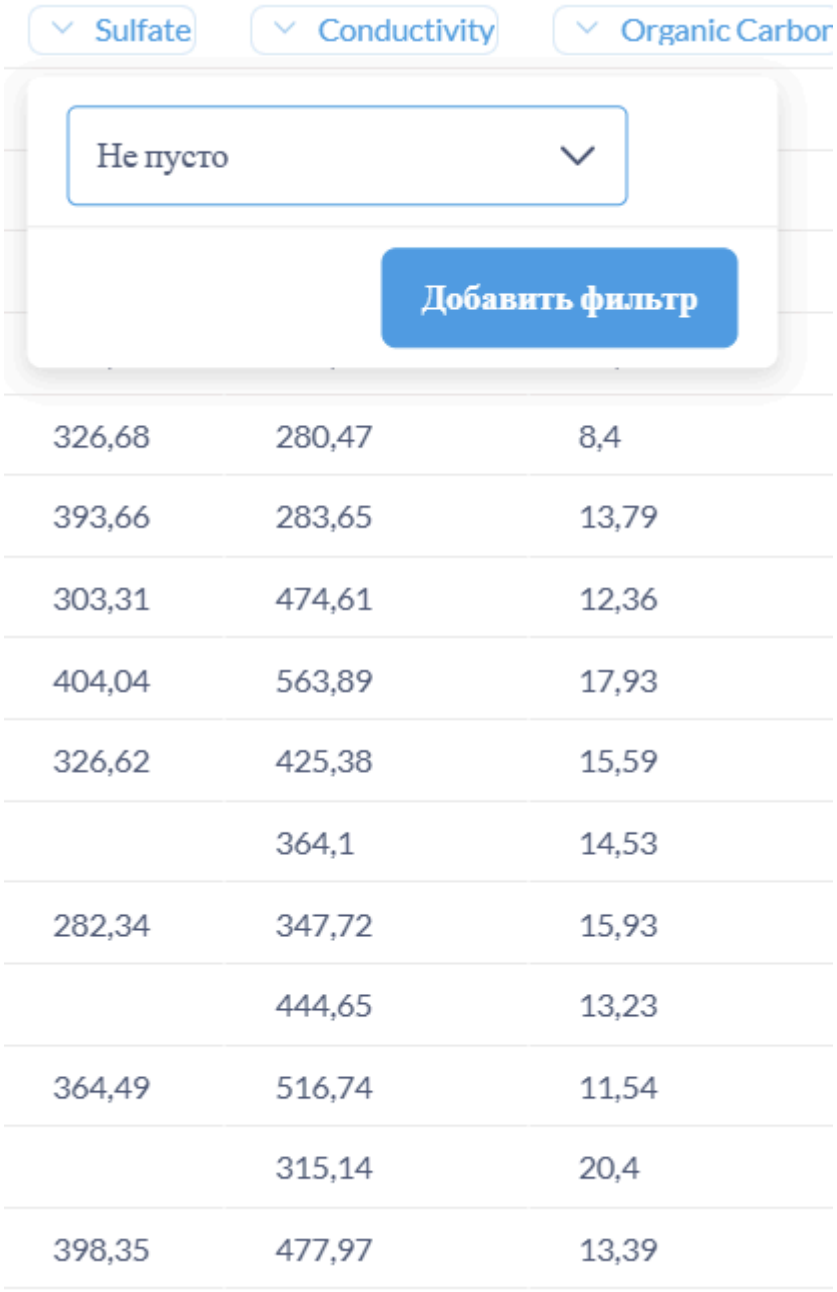


Рисунок 18 – Удаление пропусков в столбце Sulfate

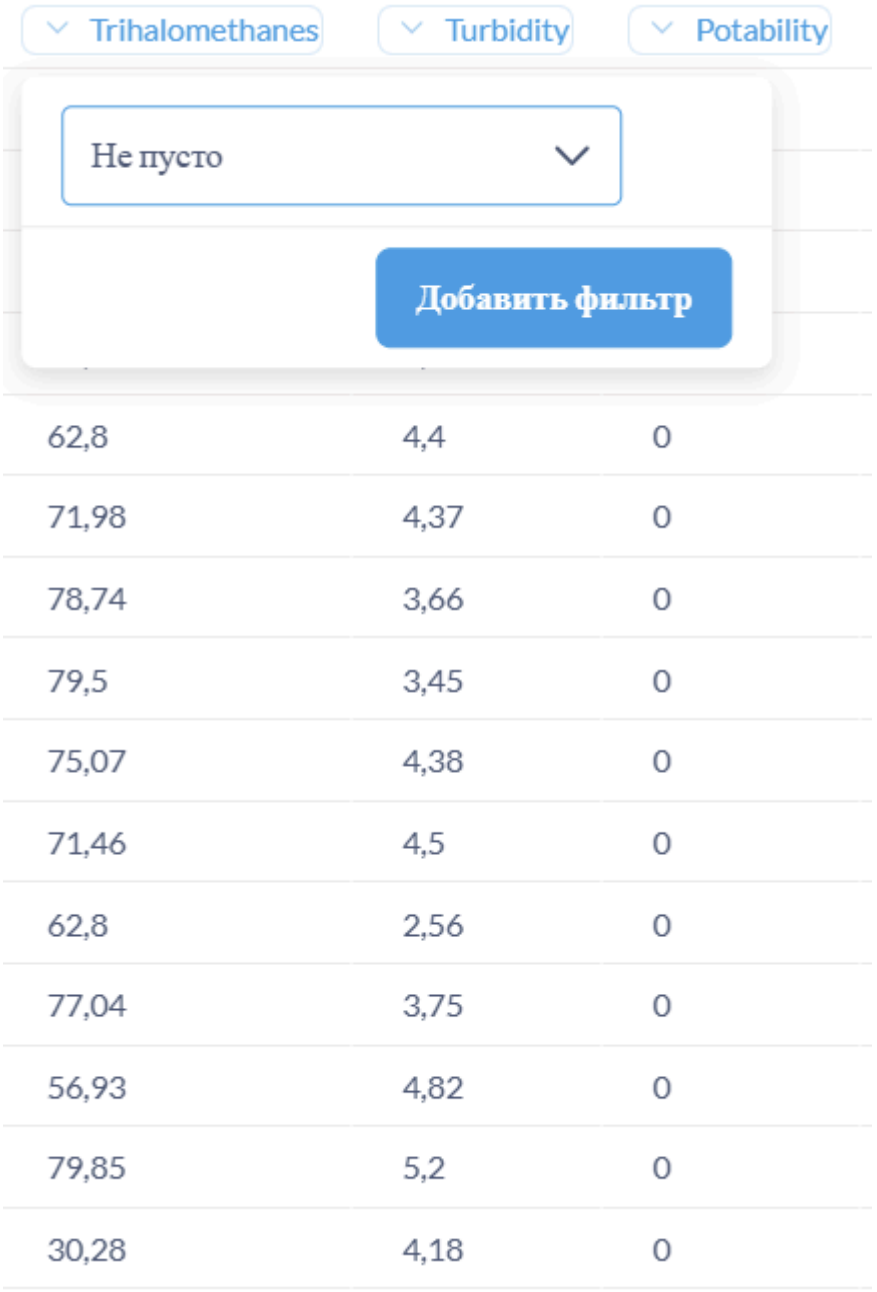


Рисунок 19 – Удаление пропусков в столбце Trihalomethanes

Рисунок 20 – Датасет, очищенный от пропусков

2.3 Применение методов статистического анализа

На рисунке 21 показана подробная статистика датасета. Кроме того, построена матрица корреляции параметров воды (Рисунки 22-23).

```
> summary(df)
       ph              Hardness        Solids         Chloramines    
 Min.   : 0.000   Min.   : 47.43   Min.   :  320.9   Min.   : 0.352  
 1st Qu.: 6.093   1st Qu.:176.85   1st Qu.:15666.7   1st Qu.: 6.127  
 Median : 7.037   Median :196.97   Median :20927.8   Median : 7.130  
 Mean   : 7.081   Mean   :196.37   Mean   :22014.1   Mean   : 7.122  
 3rd Qu.: 8.062   3rd Qu.:216.67   3rd Qu.:27332.8   3rd Qu.: 8.115  
 Max.   :14.000   Max.   :323.12   Max.   :61227.2   Max.   :13.127  
 NA's   :491                                                        
    Sulfate        Conductivity   Organic_carbon  Trihalomethanes 
 Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   :  0.738  
 1st Qu.:307.7   1st Qu.:365.7   1st Qu.:12.07   1st Qu.: 55.845  
 Median :333.1   Median :421.9   Median :14.22   Median : 66.622  
 Mean   :333.8   Mean   :426.2   Mean   :14.28   Mean   : 66.396  
 3rd Qu.:360.0   3rd Qu.:481.8   3rd Qu.:16.56   3rd Qu.: 77.337  
 Max.   :481.0   Max.   :753.3   Max.   :28.30   Max.   :124.000  
 NA's   :781                                     NA's   :162      
   Turbidity       Potability     
 Min.   :1.450   Min.   :0.0000  
 1st Qu.:3.440   1st Qu.:0.0000  
 Median :3.955   Median :0.0000  
 Mean   :3.967   Mean   :0.3901  
 3rd Qu.:4.500   3rd Qu.:1.0000  
 Max.   :6.739   Max.   :1.0000
```

Рисунок 21 – Статистика

```
> cor_matrix <- cor(df)
> corrplot(cor_matrix, tl.col = "black")
```

Рисунок 22 – Матрица корреляции. Код

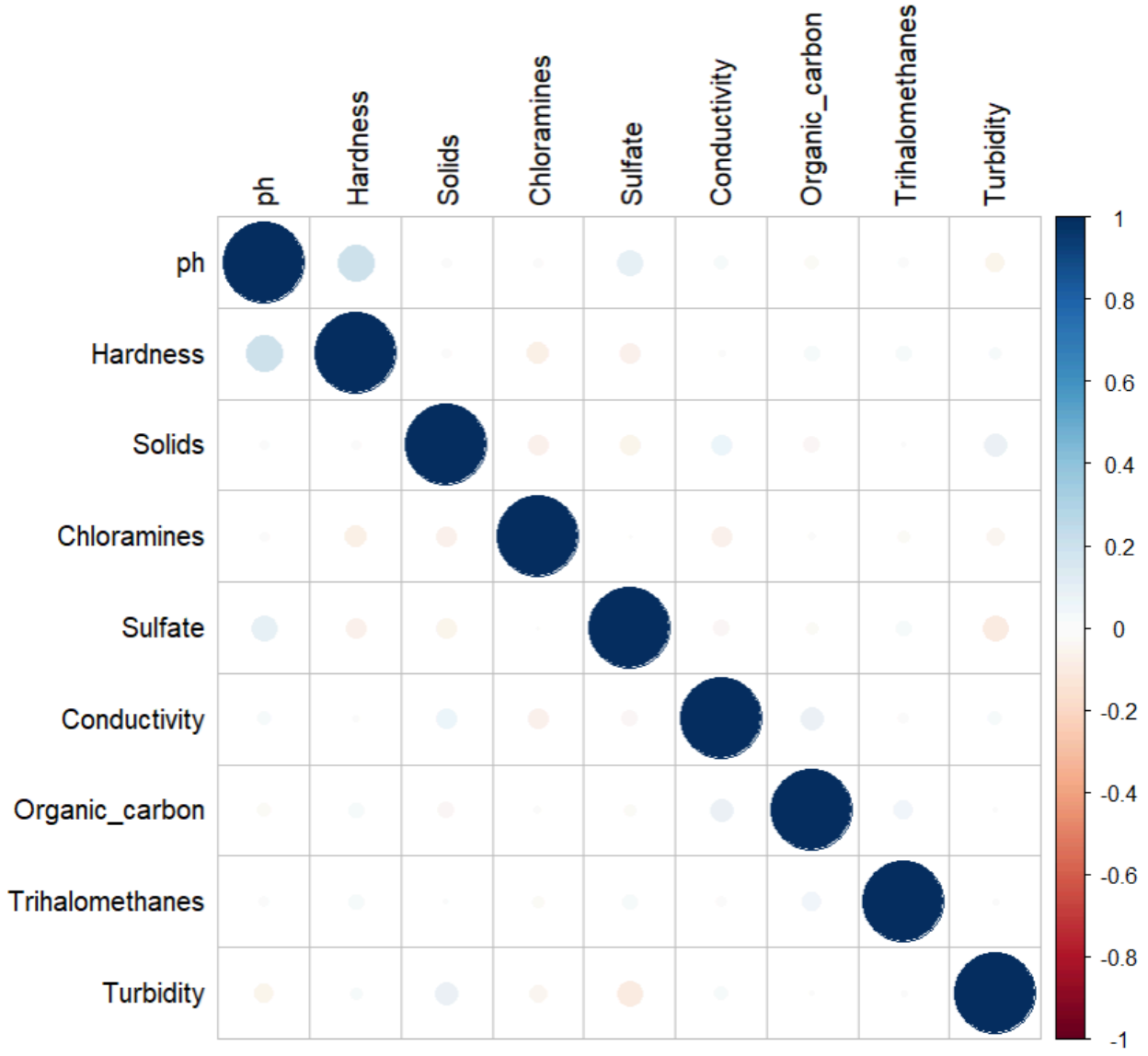


Рисунок 23 – Матрица корреляции. Код

Параметры имеют слабую корреляцию, по модулю не выше 0.3, значит связь между этими переменными крайне мала.

Для применения алгоритмов кластеризации данные были масштабированы (Рисунок 24).

```
> df_scaled <- scale(df_numeric)
```

Рисунок 24 – Масштабирование данных

2.4 Машинное обучение в анализе данных

Данные готовы к запуску кластеризации. Применение алгоритма иерархической кластеризации и вывод результата в виде таблицы с количеством объектов в кластерах показано на Рисунке 25.

```
> dist_matrix <- dist(df_scaled, method="euclidean")
> hc <- hclust(dist_matrix, method="ward.D2")
> clusters <- cutree(hc, k=2)
> table(clusters)
clusters
   1    2
1137  613
```

Рисунок 25 – Иерархическая кластеризация

При задании параметра k равным 2 алгоритм выделяет 2 кластера, содержащих 1137 и 613 объектов соответственно.

Для применения алгоритма k-means необходимо определить оптимальное количество кластеров, для этого использован метод локтя (Рисунки 26-27).

```
> fviz_nbclust(df_scaled, kmeans, method = "wss") +
+   ggtitle("K-means: метод локтя")
```

Рисунок 26 – Метод локтя. Код

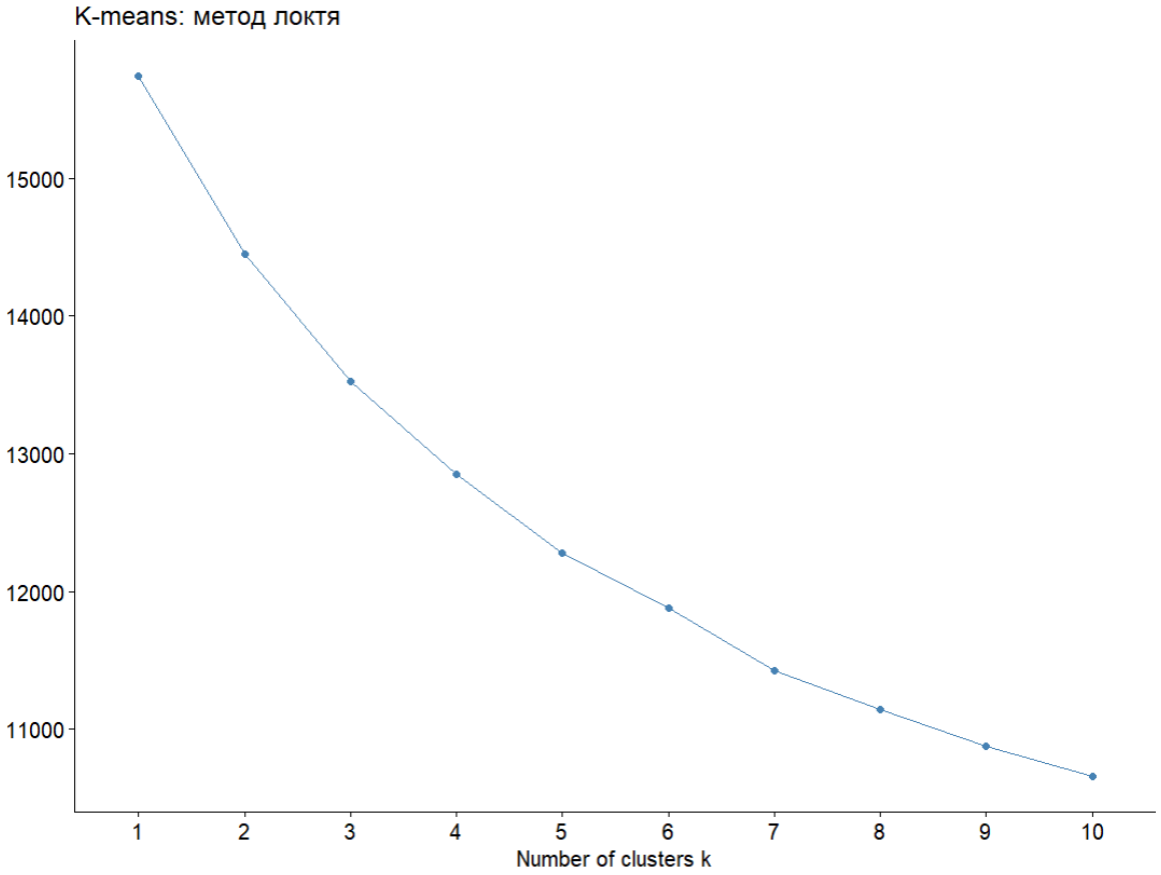


Рисунок 27 – Метод локтя. График

Метод локтя оказался не эффективным, поэтому был применён метод силуэта (Рисунки 28-29).

```
> fviz_nbclust(df_scaled, kmeans, method = "silhouette") +
+   ggtitle("K-means: метод силуэта")
```

Рисунок 28 – Метод силуэта. Код

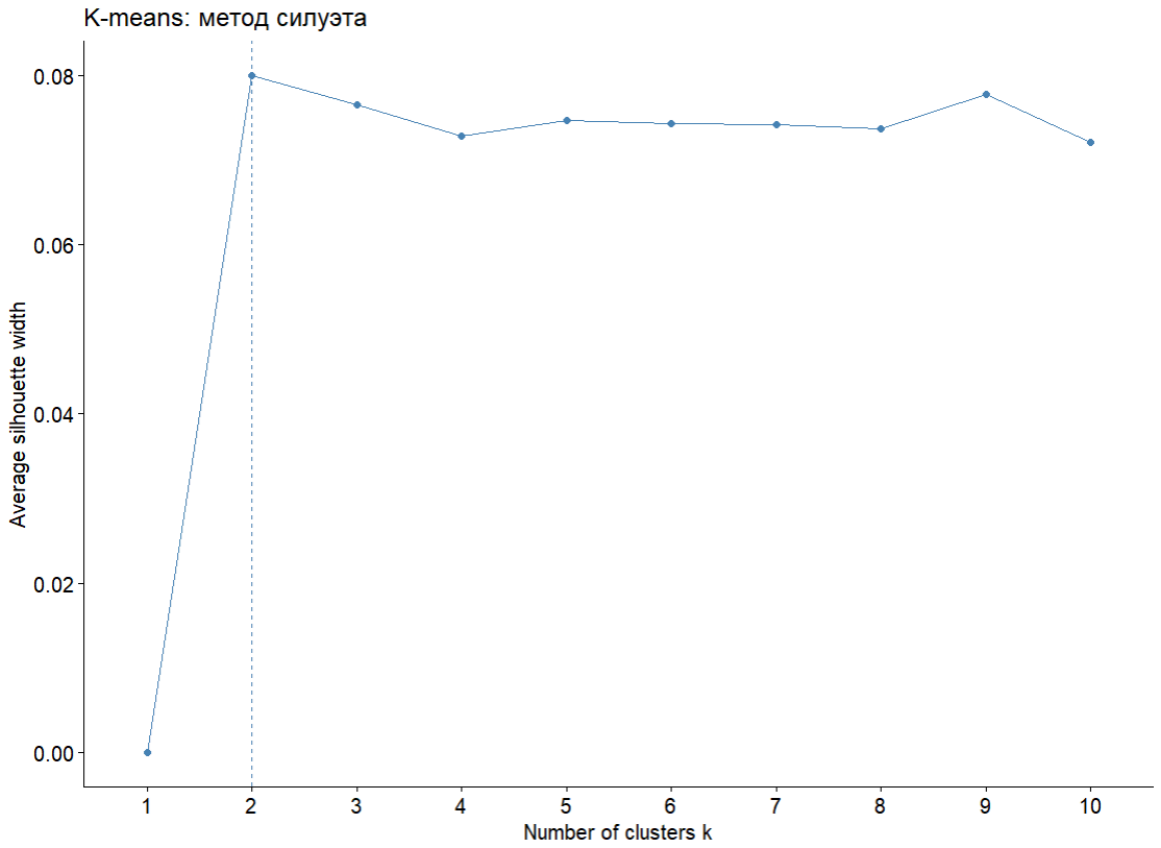


Рисунок 29 – Метод силуэта. График

Оптимальное значение k - количество кластеров равно 2. Скриншот запуска алгоритма k-means с параметром k=2 представлен на Рисунке 30.

```
> set.seed(42)
> k <- 2
> kmeans_result <- kmeans(df_scaled, centers = k, nstart = 10)
```

Рисунок 30 – Алгоритм k-means

Результаты работы алгоритма представлены на Рисунках 31-33.

```
> kmeans_result$cluster
  1   2   3   4   5   6   7   8   9  10  11  12  13  14
  1   2   1   1   1   2   1   1   2   1   2   1   2   2
 15  16  17  18  19  20  21  22  23  24  25  26  27  28
  2   2   2   1   1   2   1   2   2   2   2   2   2   1
 29  30  31  32  33  34  35  36  37  38  39  40  41  42
  1   2   1   1   1   1   1   2   2   1   2   2   1   1
 43  44  45  46  47  48  49  50  51  52  53  54  55  56
  1   1   2   2   1   1   2   1   2   2   2   1   1   2
 57  58  59  60  61  62  63  64  65  66  67  68  69  70
  1   2   1   1   2   2   1   1   2   2   2   2   1   2
 71  72  73  74  75  76  77  78  79  80  81  82  83  84
  1   2   2   1   1   2   2   2   2   1   2   1   1   1
 85  86  87  88  89  90  91  92  93  94  95  96  97  98
  2   1   2   1   1   1   1   2   1   1   1   2   1   2
 99 100 101 102 103 104 105 106 107 108 109 110 111 112
  2   1   2   1   1   2   2   1   1   1   2   1   1   1
113 114 115 116 117 118 119 120 121 122 123 124 125 126
  1   1   1   1   1   1   2   2   2   2   1   1   1   1
127 128 129 130 131 132 133 134 135 136 137 138 139 140
  1   1   1   2   2   2   2   1   2   2   1   1   1   2
141 142 143 144 145 146 147 148 149 150 151 152 153 154
  1   2   2   1   2   1   2   1   1   2   1   2   2   2
155 156 157 158 159 160 161 162 163 164 165 166 167 168
  1   1   1   1   1   1   1   1   2   2   1   1   2   1
169 170 171 172 173 174 175 176 177 178 179 180 181 182
  1   1   2   1   1   1   1   2   2   1   2   2   1   2
183 184 185 186 187 188 189 190 191 192 193 194 195 196
  2   1   2   1   2   1   2   2   1   1   2   2   2   1
197 198 199 200 201 202 203 204 205 206 207 208 209 210
  1   2   1   2   2   1   1   1   1   1   1   1   1   1
211 212 213 214 215 216 217 218 219 220 221 222 223 224
  2   1   2   1   1   2   1   2   1   2   2   2   2   1
225 226 227 228 229 230 231 232 233 234 235 236 237 238
  1   1   1   2   1   1   2   2   2   1   2   2   2   1
239 240 241 242 243 244 245 246 247 248 249 250 251 252
  2   1   1   1   1   1   1   1   2   1   1   1   1   1
253 254 255 256 257 258 259 260 261 262 263 264 265 266
  1   2   1   1   2   2   1   1   1   2   1   1   1   1
267 268 269 270 271 272 273 274 275 276 277 278 279 280
```

Рисунок 31 – Векторы принадлежности кластеру

```
> kmeans_result$centers
          ph   Hardness     Solids Chloramines    Sulfate Conductivity Organic_carbon
1  0.4786544  0.4792466 -0.4322960 -0.08219319  0.2035193  -0.07564423     -0.1238103
2 -0.4976221 -0.4982377  0.4494266  0.08545027 -0.2115842   0.07864179      0.1287165
  Trihalomethanes  Turbidity
1      0.06043247 -0.1500752
2     -0.06282723  0.1560223
```

Рисунок 32 – Координаты центроидов

```
> kmeans_result$tot.withinss
[1] 14392.74
```

Рисунок 33 – Суммарная внутрикластерная сумма квадратов

2.5 Визуализация данных

Визуализация работы алгоритмов показана на Рисунках 34-37.

```
> plot(hc, main="Иерархическая кластеризация", xlab="", sub="")
```

Рисунок 34 – Дендрограмма. Код

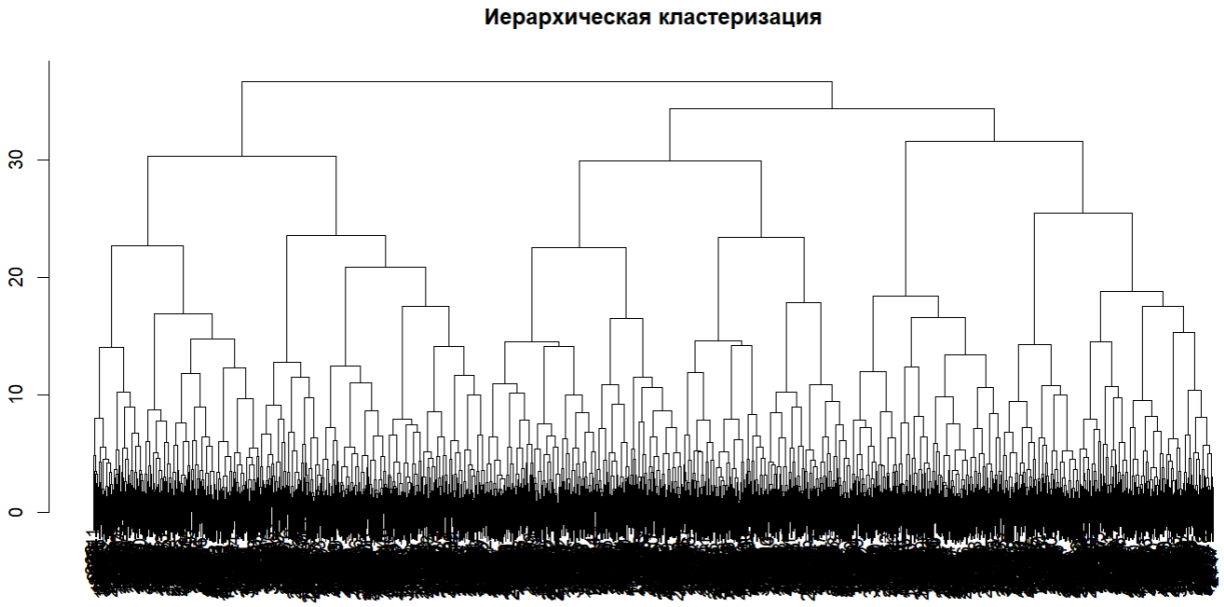


Рисунок 35 – Дендрограмма. График

```
> plot(hc, main="Иерархическая кластеризация", xlab="", sub="")
> pca_res <- prcomp(df_scaled)
> df_pca <- data.frame(pca_res$x[,1:2], Cluster = factor(kmeans_result$cluster))
> ggplot(df_pca, aes(x=PC1, y=PC2, color=Cluster)) + geom_point() + ggtitle("K-means Clustering")
```

Рисунок 36 – Визуализация работы k-means. Код

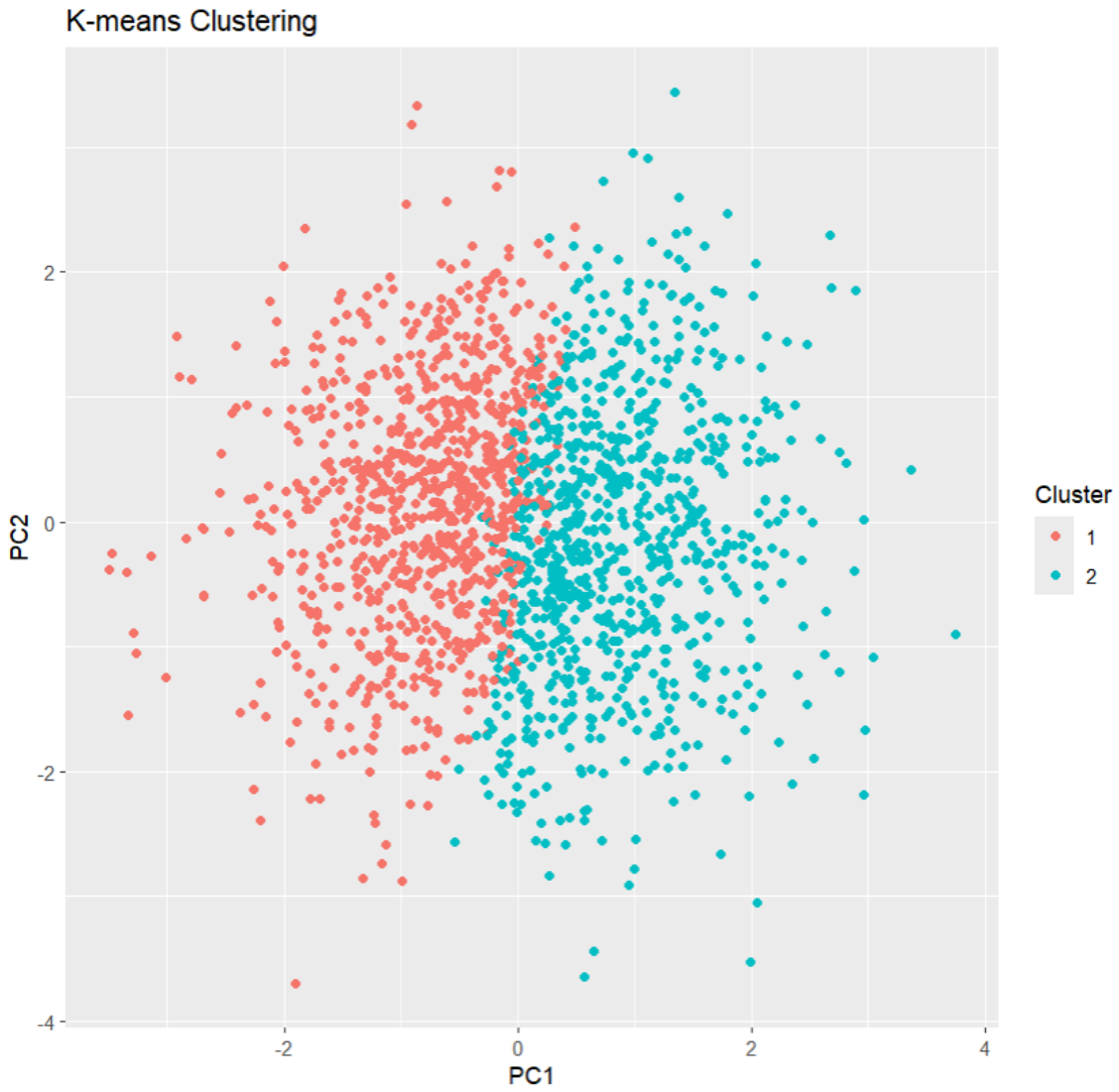


Рисунок 37 – Визуализация работы k-means. График

На графике чётко обозначены два кластера, полученные в результате применения алгоритма kmeans к обработанным данным. Использование метода главных компонент (PCA) позволило упростить многомерное пространство признаков до двух осей (PC1 и PC2), обеспечивая наглядное разделение наблюдений.

Кластер 1 (красный) — характеризуется более низкими значениями по первой главной компоненте (PC1). Исходя из анализа центроидов кластеров, этот кластер содержит в себе воду плохого качества.

Кластер 2 (голубой) — сосредоточен на правой стороне графика и включает наблюдения с высокими значениями PC1. Исходя из анализа центроидов кластеров, этот кластер содержит в себе воду хорошего качества.

3. АВТОМАТИЗАЦИЯ И ОТЧЁТНОСТЬ В АНАЛИЗЕ ДАННЫХ

3.1 Генерация отчётов в R

Процесс создания отчёта с помощью RMarkdown показан на Рисунках 38-41.

```
---
title: "Использование методов иерархической кластеризации и алгоритма k-means для определения качества воды"
author: "Ильина Ксения"
date: "2025-05-15"
output:
  html_document:
    toc_depth: 2
    toc_float: true
    theme: cosmo
    highlight: tango
---
```

Рисунок 38 – Оформление заголовка

```
## 1. Введение
Этот отчет представляет собой анализ данных о качестве воды с использованием методов кластеризации. Целью исследования является реализация модели кластеризации методом иерархической кластеризации и k-means для автоматизации мониторинга качества воды.
```

Рисунок 39 – Оформление введения

```
## 2. Исследовательский анализ данных (EDA) и очистка данных
### 2.1. Структура данных
```{r 1}
str(df)
```
2.2. Информация о пропусках
```{r 2}
colSums(is.na(df))
```
2.2. Обработка данных
```{r 3}
df<-df[!duplicated(df),]
df$ph[is.na(df$ph)]<-mean(df$ph, na.rm = TRUE)
df$Trihalomethanes[is.na(df$Trihalomethanes)]<-mean(df$Trihalomethanes, na.rm = TRUE)
df$Sulfate[is.na(df$Sulfate)]<-mean(df$Sulfate, na.rm = TRUE)
df$Potability <- NULL
```
```

Рисунок 40 – Оформление основной части работы

```
> library(rmarkdown)
> rmarkdown::render("Report.Rmd")


processing file: Report.Rmd

output file: Report.knit.md

"D:/Program/R_Studio/RStudio/resources/app/bin/quarto/bin/tools/pandoc" +RTS -K512m -RTS Report.knit.md --to html4 --
from markdown+autolink_bare_uris+tex_math_single_backslash --output Report.html --lua-filter "D:\Program\R\R-4.4.2\li
brary\rmarkdown\rmarkdown\lua\pagebreak.lua" --lua-filter "D:\Program\R\R-4.4.2\library\rmarkdown\rmarkdown\lua\latex
-div.lua" --embed-resources --standalone --variable bs3=TRUE --section-divs --table-of-contents --toc-depth 2 --varia
ble toc_float=1 --variable toc_selectors=h1,h2 --variable toc_collapsed=1 --variable toc_smooth_scroll=1 --variable t
oc_print=1 --template "D:\Program\R\R-4.4.2\library\rmarkdown\rmd\h\default.html" --no-highlight --variable highlight
js=1 --variable theme=bootstrap --mathjax --variable "mathjax-url=https://mathjax.rstudio.com/latest/MathJax.js?confi
g=TeX-AMS-MML_HTMLorMML" --include-in-header "C:\Users\user\AppData\Local\Temp\RtmpeiqQ9R\rmarkdown-str489435db23f2.h
tml"

Output created: Report.html
```

Рисунок 41 – Генерация отчёта

Сгенерированный отчёт показан на Рисунках 42-52.

ЗАКЛЮЧЕНИЕ

В процессе работы был проведён всесторонний анализ данных о качестве воды, что позволило глубже понять состояние водных ресурсов и выявить ключевые факторы, влияющие на их качество. Исследование охватило все этапы аналитического процесса, начиная с первичной предобработки данных и заканчивая визуализацией результатов.

На первом этапе был детально описан используемый датасет, включая его характеристики, источники данных и структуру. Это позволило сформировать чёткое представление о том, какие именно параметры были включены в анализ и как они могут быть интерпретированы.

Далее был проведён исследовательский анализ данных (EDA), который включал очистку данных от пропусков, дубликатов и выбросов. Этот этап был критически важен, поскольку от качества данных напрямую зависит точность и надёжность последующих моделей. Очистка данных позволила устранить потенциальные искажения и обеспечить корректность последующих этапов анализа.

После очистки данных был выполнен корреляционный анализ, который позволил выявить взаимосвязи между различными параметрами качества

воды. Это дало возможность понять, какие факторы могут оказывать наибольшее влияние на общее состояние водных ресурсов и как они могут взаимодействовать друг с другом.

Затем были построены модели иерархической кластеризации и k-means. Для метода k-средних использовались методы локтя и силуэта, чтобы определить оптимальное количество кластеров. Эти методы позволили выбрать наиболее подходящее количество кластеров, что обеспечило более точную и релевантную сегментацию данных.

Данные были разбиты на две группы: вода хорошего качества и вода плохого качества. Это разделение позволило более наглядно продемонстрировать различия между этими категориями и выявить ключевые особенности каждой из них.

Для визуализации результатов были использованы различные инструменты: гистограммы распределения переменных качества воды, дендрограмма и результат работы k-means. Эти визуализации позволили наглядно представить полученные данные и сделать выводы о состоянии водных ресурсов.

В R сгенерирован отчёт с использованием RMarkdown, в котором подробно описан процесс работы, включая все этапы анализа, используемые методы и полученные результаты. Это позволило сделать исследование прозрачным и доступным для других исследователей и специалистов.

Таким образом, проведённое исследование подтвердило актуальность применения методов кластеризации в анализе качества воды. Полученные результаты демонстрируют их практическую значимость для экологии, здравоохранения и промышленности. Это подчёркивает важность дальнейшего развития и применения аналитических методов для улучшения состояния водных ресурсов и обеспечения их устойчивого использования.

Отчет по работе,сгенерированный при помощи RMarkdown

Использование методов иерархической кластеризации и алгоритма k-means для определения качества воды

Ильина Ксения

2025-05-15

1. Введение

Этот отчет представляет собой анализ данных о качестве воды с использованием методов кластеризации. Целью исследования является реализация модели кластеризации методом иерархической кластеризации и k-means для автоматизации мониторинга качества воды.

2. Исследовательский анализ данных (EDA) и очистка данных

2.1. Структура данных

```
str(df)
```

```
## 'data.frame':    3276 obs. of  10 variables:
##  $ ph             : num  NA 3.72 8.1 8.32 9.09 ...
##  $ Hardness       : num  205 129 224 214 181 ...
##  $ Solids         : num  20791 18630 19910 22018 17979 ...
##  $ Chloramines    : num  7.3 6.64 9.28 8.06 6.55 ...
##  $ Sulfate        : num  369 NA NA 357 310 ...
##  $ Conductivity   : num  564 593 419 363 398 ...
##  $ Organic_carbon : num  10.4 15.2 16.9 18.4 11.6 ...
##  $ Trihalomethanes: num  87 56.3 66.4 100.3 32 ...
##  $ Turbidity      : num  2.96 4.5 3.06 4.63 4.08 ...
##  $ Potability     : int  0 0 0 0 0 0 0 0 0 0 ...
```

Рисунок 42 – Готовый отчёт. Часть 1

2.2. Информация о пропусках

```
colSums(is.na(df))
```

```
##              ph        Hardness          Solids     Chloramines         Sulfate    Conductivity  Organic_carbo
n
##             491               0               0               0             781               0
0
## Trihalomethanes       Turbidity      Potability
##             162               0               0
```

2.2. Обработка данных

```
df<-df[!duplicated(df),]
df$ph[is.na(df$ph)]<-mean(df$ph, na.rm = TRUE)
df$Trihalomethanes[is.na(df$Trihalomethanes)]<-mean(df$Trihalomethanes, na.rm = TRUE)
df$Sulfate[is.na(df$Sulfate)]<-mean(df$Sulfate, na.rm = TRUE)
df$Potability <- NULL
```

Рисунок 43 – Готовый отчёт. Часть 2

2.3. Информация о выбросах

```
plots <- list()
for (col in names(df)) {
  plots[[col]] <- ggplot(df, aes_string(y = col)) +
    geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
    ggtitle(paste(col)) +
    theme_minimal() +
    ylab("") +
    xlab("")
    #coord_flip()
}
do.call(grid.arrange, c(plots, ncol = 3))
```

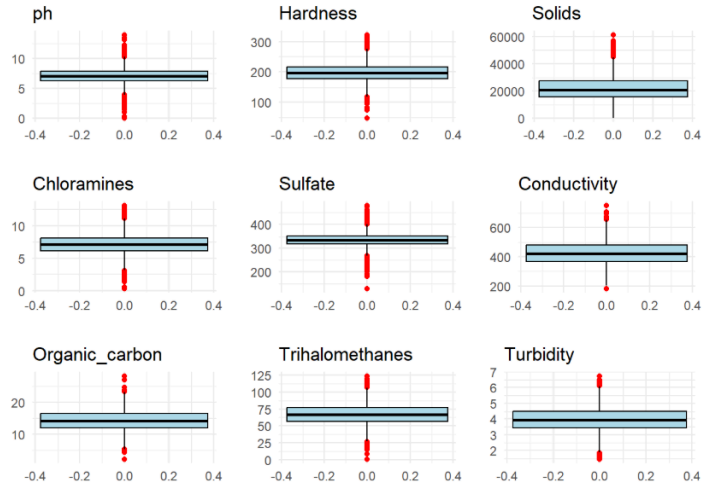


Рисунок 44 – Готовый отчёт. Часть 3

2.4. Удаление выбросов

```
remove_outliers_iqr_1 <- function(df, columns = names(df)) {
  for (col in columns) {
    if (is.numeric(df[[col]])) {#
      Q1 <- quantile(df[[col]], 0.25, na.rm = TRUE)
      Q3 <- quantile(df[[col]], 0.75, na.rm = TRUE)
      IQR <- Q3 - Q1
      lower_bound <- Q1 - 1.2 * IQR
      upper_bound <- Q3 + 1.2 * IQR
      df <- df[df[[col]] >= lower_bound & df[[col]] <= upper_bound | is.na(df[[col]]), ]
    }
  }
  return(df)
}
df <- remove_outliers_iqr_0.5(df)
```

3. Статистический анализ

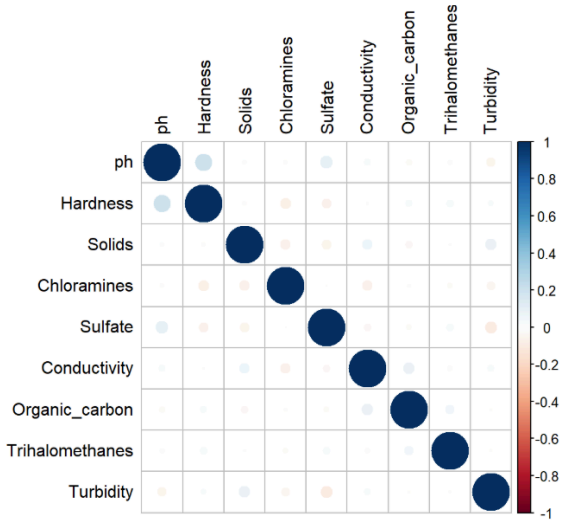
3.1. Визуализация распределения переменных

```
plots <- list()
for (col in names(df)) {
  plots[[col]] <- ggplot(df, aes_string(x = col)) +
    geom_histogram( color = "black", bins = 30) +
    ggtitle(paste("Распределение", col))
}
do.call(grid.arrange, c(plots, ncol = 3))
```

Рисунок 45 – Готовый отчёт. Часть 4

3.2. Анализ корреляции

```
cor_matrix <- cor(df)
corrplot(cor_matrix, tl.col = "black")
```



3.3. Масштабирование

```
df_scaled <- scale(df_numeric)
```

Рисунок 46 – Готовый отчёт. Часть 5

4. Машинное обучение

4.1. Иерархическая кластеризация

Построение дерева

```
dist_matrix <- dist(df_scaled, method="euclidean")
hc <- hclust(dist_matrix, method="ward.D2")
```

Выделение 2 кластеров

```
clusters <- cutree(hc, k=2)
table(clusters)
```

```
## clusters
##    1    2
## 1137  613
```

4.2. k-means

4.2.1. Подбор оптимального количества кластеров

Метод локтя

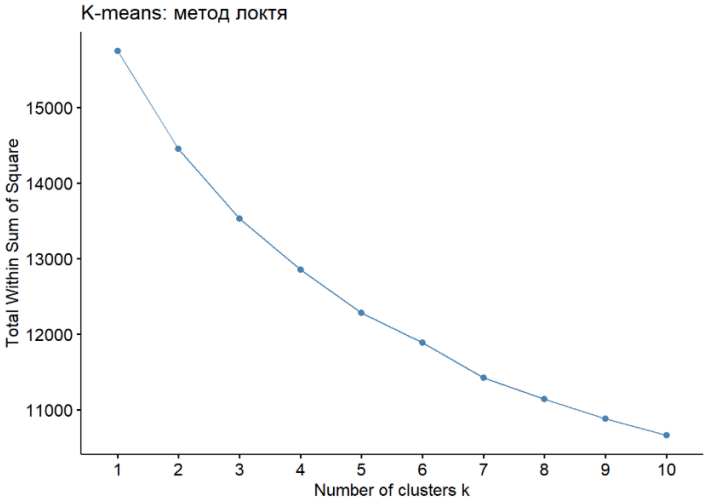
```
fviz_nbclust(df_scaled, kmeans, method = "wss") +
  ggtitle("K-means: метод локтя")
```

Рисунок 47 – Готовый отчёт. Часть 6

4.2.1. Подбор оптимального количества кластеров

Метод локтя

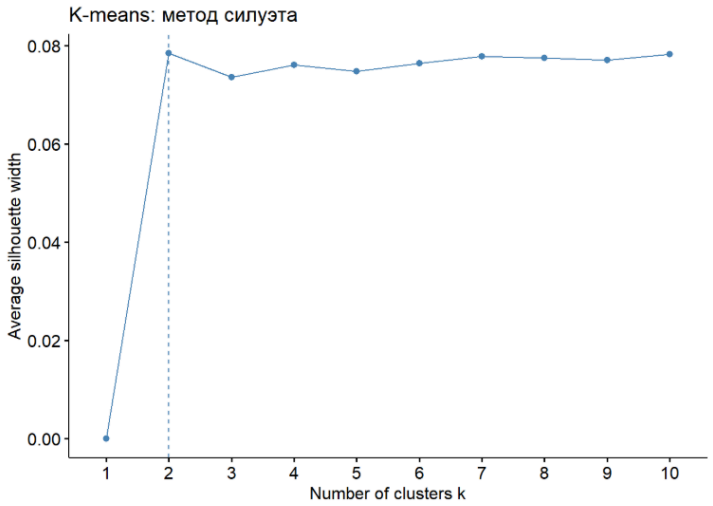
```
fviz_nbclust(df_scaled, kmeans, method = "wss") +
  ggtitle("K-means: метод локтя")
```



Метод силуэта

```
fviz_nbclust(df_scaled, kmeans, method = "silhouette") +
  ggtitle("K-means: метод силуэта")
```

Рисунок 48 – Готовый отчёт. Часть 7



4.2.2. Запуск алгоритма

```
set.seed(42)
k <- 2
kmeans_result <- kmeans(df_scaled, centers = k, nstart = 10)
```

4.2.3. Анализ результатов

```
kmeans_result$cluster # вектор принадлежности кластеру
```

```
##   3   4   5   6   8  11  12  15  16  17  20  22  23  25  26  28  31  32  35  36  41  42
##   1   1   1   2   1   2   1   2   2   2   1   2   1   2   2   2   1   1   1   2   1   2
##  45  46  47  50  54  58  60  64  66  69  70  74  75  76  79  82  83  86  87  91  92  93
##   2   2   1   1   1   2   2   1   2   1   2   1   1   2   2   2   1   1   1   1   1   1
##  95  96  97 100 102 103 106 107 108 109 111 115 116 122 128 130 132 136 137 138 139 140
##   1   2   1   1   1   2   1   1   1   2   1   1   2   2   1   2   2   2   2   2   1   1
## 142 144 145 146 147 149 151 154 155 156 158 160 161 163 165 167 168 169 171 172 174 177
```

Рисунок 49 – Готовый отчёт. Часть 8

1. Введение
2. Исследовательский анализ данных (EDA) и очистка данных
3. Статистический анализ
4. Машинное обучение
5. Визуализация

```
kmeans_result$centers # координаты центроидов
```

```
##          ph   Hardness     Solids Chloramines    Sulfate Conductivity Organic_carbon Trihalomethanes  Turbidi
ty
## 1  0.4786544  0.4792466 -0.4322960 -0.08219319  0.2035193  -0.07564423     -0.1238103      0.06043247 -0.15007
52
## 2 -0.4976221 -0.4982377  0.4494266  0.08545027 -0.2115842   0.07864179      0.1287165     -0.06282723  0.15602
23
```

```
kmeans_result$tot.withinss # суммарная внутрикластерная сумма квадратов
```

```
## [1] 14392.74
```

5. Визуализация

5.1. Иерархическая кластеризация

```
plot(hc, main="Иерархическая кластеризация", xlab="", sub="")
```

Рисунок 50 – Готовый отчёт. Часть 9

1. Введение
2. Исследовательский анализ данных (EDA) и очистка данных
3. Статистический анализ
4. Машинное обучение
5. Визуализация

5. Визуализация

5.1. Иерархическая кластеризация

```
plot(hc, main="Иерархическая кластеризация", xlab="", sub="")
```

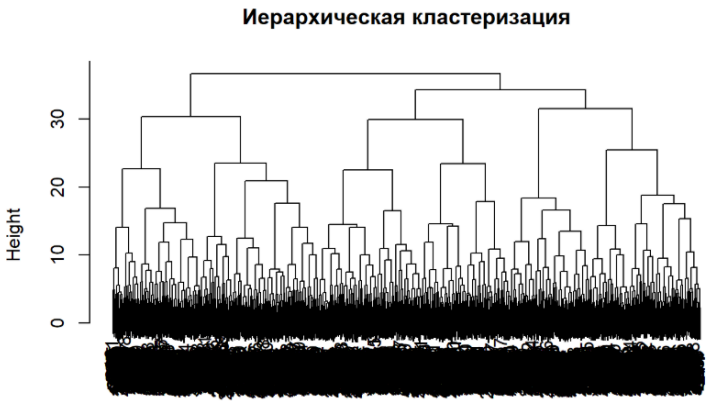


Рисунок 51 – Готовый отчёт. Часть 10

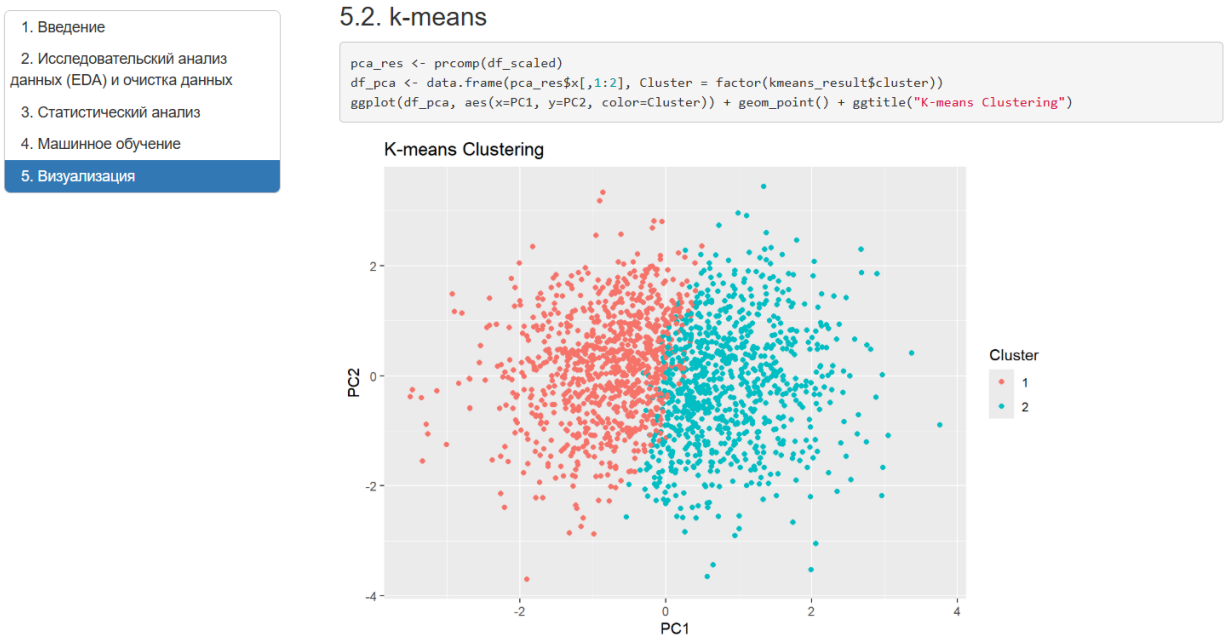


Рисунок 52 – Готовый отчёт. Часть 11