

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

Казахский национальный исследовательский
технический университет имени К.И.Сатпаева

Институт автоматизации и информационных технологий

Кафедра кибербезопасность, обработка и хранение
информации

Сулиев Жалил Ниязович

Разработка компонента «Интерпретация данных» информационной системы в
аналитике данных

ДИПЛОМНАЯ РАБОТА

Специальность 5В070300 – Информационные системы

Алматы, 2022

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

Казахский национальный исследовательский технический
университет имени К.И.Сатпаева

Институт автоматизации и информационных технологий
Кафедра кибербезопасность, обработка и хранение информации

ДОПУЩЕН К ЗАЩИТЕ

Заведующий кафедрой
КОиХИ

доктор Ph.D, ассоц. профессор
Р.Ж.Сатыбалдиева



« 19 »

05 20__ г.

ДИПЛОМНАЯ РАБОТА

На тему: Разработка компонента «Интерпретация данных» информационной
системы в аналитике данных

Специальность 5В070300 – Информационные системы

Выполнил: Сулнев Жалил Ниязович

Рецензент

Доктор Ph.D, стар. преподаватель

Бимұрат Жанар

« 16 » 05 2022г.

Научный руководитель

Ассоц. проф. канд. тех. наук

Жумагалиев Б.И.

« 16 » 05 2022г.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

Казахский национальный исследовательский технический
университет имени К.И.Сатпаева

Институт автоматки и информационных технологий
Кафедра кибербезопасность, обработка и хранение информации

5В070300 – Информационные системы



УТВЕРЖДАЮ

Заведующий кафедрой КОиХИ
доктор Ph.D, ассоц. профессор
Р.Ж.Сатыбалдиева

« 19 » 05 2022г.

ЗАДАНИЕ

на выполнение дипломной работы

Обучающемуся: Судиеву Жалиду Ниязовичу

Тема: Разработка компонента «Интерпретация данных» информационной системы в аналитике данных.

Утверждена приказом Ректора Университета № 489-П/О от 24.12.2021г.

Срок сдачи законченной работы 24.05.2022г.

Исходные данные к дипломному проекту: материалы обследования проекта, данные описания проекта, документация по среде разработки, статистические данные, сбор теоретического материала.

Краткое содержание дипломной работы:

- а) Обзор;
- б) Постановка задачи;
- в) Исследовательская часть;
- г) Проектирование системы;
- д) Описание программного обеспечения;
- е) Заключение на основе полученных данных.

Перечень графического материала (с точным указанием обязательных чертежей):
представлены 17 слайдов презентации работы

Рекомендуемая основная литература: из 11 наименований



ГРАФИК

подготовки дипломной работы (проекта)


Наименование разделов, перечень разрабатываемых вопросов	Сроки представления научному руководителю	Примечание
1. Обзор	15.02.2022	выполн.
2. Постановка задачи	28.02.2022	выполн.
3. Исследовательская часть	20.03.2022	выполн.
4. Проектирование и разработка системы	21.04.2022	выполн.
5. Описание программного обеспечения	30.05.2022	выполн.
6. Заключение на основе полученных данных	05.05.2022	выполн.

Подписи

консультантов и нормоконтролера на законченную дипломную работу (проект) с указанием относящихся к ним разделов работы (проекта)

Наименование разделов	Консультанты, Ф.И.О. (уч. степень, звание)	Дата подписания	Подпись
Основная часть	Жумагалиев Б.И. (канд.тех.наук, ассоц.профессор)	05.05.2022 г	
Нормоконтроль	Аристомбаева М.Т. (маг.тех.наук, лектор)	16.05.2022 г	

Научный руководитель:  Жумагалиев Б.И.

Задание принял к исполнению обучающийся:  Сулиев Ж.Н.

Дата

"1" февраля 2022

АНДАТПА

Дипломдық жұмыстың мақсаты – деректерді талдаудағы ақпараттық жүйенің «Деректерді интерпретациялау» компонентін әзірлеу және шешім қабылдау үшін RStudio графикалық мүмкіндіктерін практикалық қолдану. Жұмыс барысында медицина саласында, соның ішінде инсультпен ауыратын адамдар туралы ақпарат жинау, зерттеу және талдау жүргізілді. Дипломдық жұмыс ауруға бейім адамдар туралы ақпараттарды, әртүрлі графиктер мен регрессиялық модельді құрайды. RStudio, UML диаграммалары және т.б. құралдары қолданылды. Дипломдық жұмыстың өзектілігі өте жоғары. Себебі деректерді талдау шешім қабылдаудағы маңызды кезеңдердің бірі болып табылады, талдау нәтижесінде инсульттің пайда болуының бірқатар себептері және оның алдын-алу бойынша ұсыныстар анықталды.

ANNOTATION

The purpose of the thesis is to develop the "Data Interpretation" component of an information system in data analytics and the practical application of the graphical capabilities of RStudio for decision making. In the course of the work, information was collected, research and analysis were carried out in the field of medicine, namely people suffering from a stroke. The thesis contains information about people who are susceptible to the disease, various graphs and a regression model. Tools were used, such as RStudio, UML diagrams, etc. The relevance of the thesis is extremely high. data analysis is one of the most important steps in decision making, thanks to the analysis, a number of causes of stroke and recommendations for preventing its occurrence were identified.

АННОТАЦИЯ

Целью дипломной работы является разработка компонента «Интерпретация данных» информационной системы в аналитике данных и практическое применение графических возможностей RStudio для принятия решений. В ходе работы был проведен сбор информации, исследование и анализ в области медицины, а именно людей страдающих инсультом. Дипломная работа содержит в себе информацию о людях, которые подвержены заболеванию, различные графики и регрессионную модель. Были применены инструменты, такие как RStudio, UML-диаграммы и др. Актуальность дипломной работы крайне высока, так как анализ данных является одним из самых важных этапов в принятии решения, благодаря анализу был определен ряд причин возникновения инсульта и рекомендации по предотвращению его возникновения.

СОДЕРЖАНИЕ

	Введение	9
1	Исследовательская часть	10
1.1	Методы сбора информации и инструменты анализа	10
1.2	Системный анализ	11
1.3	Линейный регрессионный анализ	12
1.4	Постановка задачи дипломной работы	14
2	Проектная часть	16
2.1	Актуальность задачи	16
2.2	Разработка и анализ системы с помощью RStudio	16
2.3	Очистка данных	17
2.4	Построение гистограмм	18
2.5	Гистограммы с наложением нормального распределения	22
2.6	Описание и возможности Voxplot	24
2.7	Линейная регрессия	26
3	Описание программного обеспечения	33
3.1	Среда разработки R и RStudio	33
	Заключение	34
	Перечень принятых сокращений, терминов	35
	Список использованной литературы	36
	Приложение	

ВВЕДЕНИЕ

В современном мире информационные технологии занимают высокое место в жизни людей и являются неотъемлемой частью существования. С каждым годом ИТ совершенствуются и значительно упрощают человеческую жизнь. Сложно представить сферу деятельности, где не применялись бы информационные технологии, будь это образование, спорт, бизнес и т.д.

Одной из сфер, на которую информационные технологии оказали огромное влияние является непосредственно – медицина. Благодаря современным технологиям удалось совершить огромный скачок развития, начиная обычными приложениями заканчивая машинами способные выявлять множество заболеваний.

В работе рассматривается подход к обработке и анализу данных в информационных системах (ИС) на основе статистических методов с применением языка R. Рассматриваются вопросы эффективности использования языка R информационных системах. Также возможности языка R: визуализация, форматы данных, поддержка внутренних скриптов, дополнительные функции. Отмечается возможность встраивания модулей на языке R в программное обеспечение подсистем информационных систем различного назначения.

Целью дипломной работы является решение проблем людей путем сбора информации, исследования и анализа данных, с применением языка R для принятия правильных решений путем визуализации и графических возможностей RStudio. Интерпретация данных таким образом помогает воспринять информацию легче и понятней.

Разработка осуществлялась с использованием современных технологий и визуализацией. Анализ был произведен исключительно в рамках дипломного проекта.

1 ИСЛЕДОВАТЕЛЬСКАЯ ЧАСТЬ

1.1 Методы сбора информации и инструменты анализа

Разработка и применение медицинских технологий является широко распространенным вопросом текущего века. Технологии для облегчения и упрощения жизни пациентам развиваются быстрым темпом. На сегодняшний день создаются все условия для получения качественной медицинской помощи, начиная удаленной записью на прием до предсказания подверженности заболевания посредством сбора информации.

Сбор информации – это процесс сбора данных с целью исследования и анализа данных для дальнейшего использования в той или иной области, также сбор информации можно охарактеризовать как процесс, который конвертирует данные в более понятную для восприятия информацию с помощью графиков или таблиц. Поиск информации является важным информационным процессом, ведь от точности и скорости данных зависит качество и своевременное принятие решений. Сбор информации в любых сферах осуществляется примерно по одному и тому же принципу, будь это образование или бизнес.

В зависимости от сферы деятельности и области применения, можно определить несколько видов сбора данных, к примеру в статических информационных системах это:

- сбор данных с первичных источников и документов;
- заполнение различных форм или шаблонов опроса сбора данных;
- сбор данных с помощью организаций, занимающихся разнообразными видами формами отчетности.
- открытые или закрытые опросы или интервью в различных формах;
- исследование факторов влияния друг на друга, такой способ часто называют – эксперимент;
- контент-анализ документов;
- наблюдение или изучение поведения.

На рисунке 1 видно какие методы сбора информации бывают:

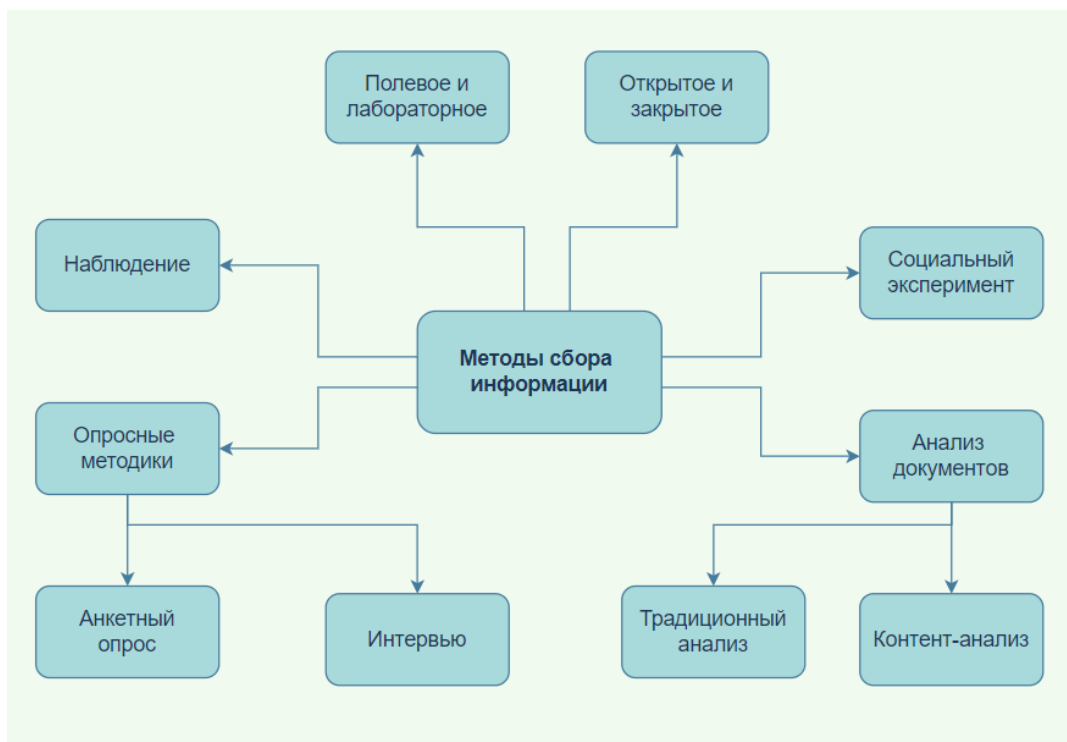


Рисунок 1 – Методы сбора информации (UML – схема)

1.2 Системный анализ

Системный анализ обычно используется для изучения сложных объектов, требующих тщательного анализа и конкретных подходов. Смысл системного анализа заключается в том, что при анализе или решении проблем он не рассматривается как общая система, а скорее принято разбивать задачи на несколько подсистем, которые представляют собой общую систему.

Фактически, сложные объекты часто называют такими объектами из-за нехватки данных, что обычно ухудшает общую картину восприятия информации. Другими словами, системный анализ — это область, которая изучает принципы, поход и исследования сложных объектов и разбивает их на более мелкие аналитические системы (подсистемы) для анализа одной и той же системы.

В медицине существуют огромные БД и сложные системы из-за высокой потребности в крупномасштабных исследованиях и проектировании систем в условиях нехватки неясной информации, нехватки ресурсов и временных ограничений, был разработан и применен только системный анализ. Этот анализ характеризуется тем, что он не учитывает ни одну, но, прежде всего, обширные и сложные системы.

Сложная система — это система, подсистемы которой принадлежат нескольким сложным системам.

Характеристики, которые могут характеризовать сложные системы, включают:

- большой размер;

- сложная структура;
- повышение уровня неопределенности в системе;
- цикл огромных информационных потоков.

Результаты сканирования системы также должны быть выполнены. Сегодня много времени и внимания уделяется внедрению результатов, полученных с помощью исследований и анализа. Это связано с тем, что на практике реализация результата зависит от разных типов систем. Маловероятно, что метод внедрения в одной системе подойдет для другой сложной и менее структурированной системы. Поэтому к этому процессу нужно подходить с особым трепетом. На рисунке 2 изображена структура системного анализа:

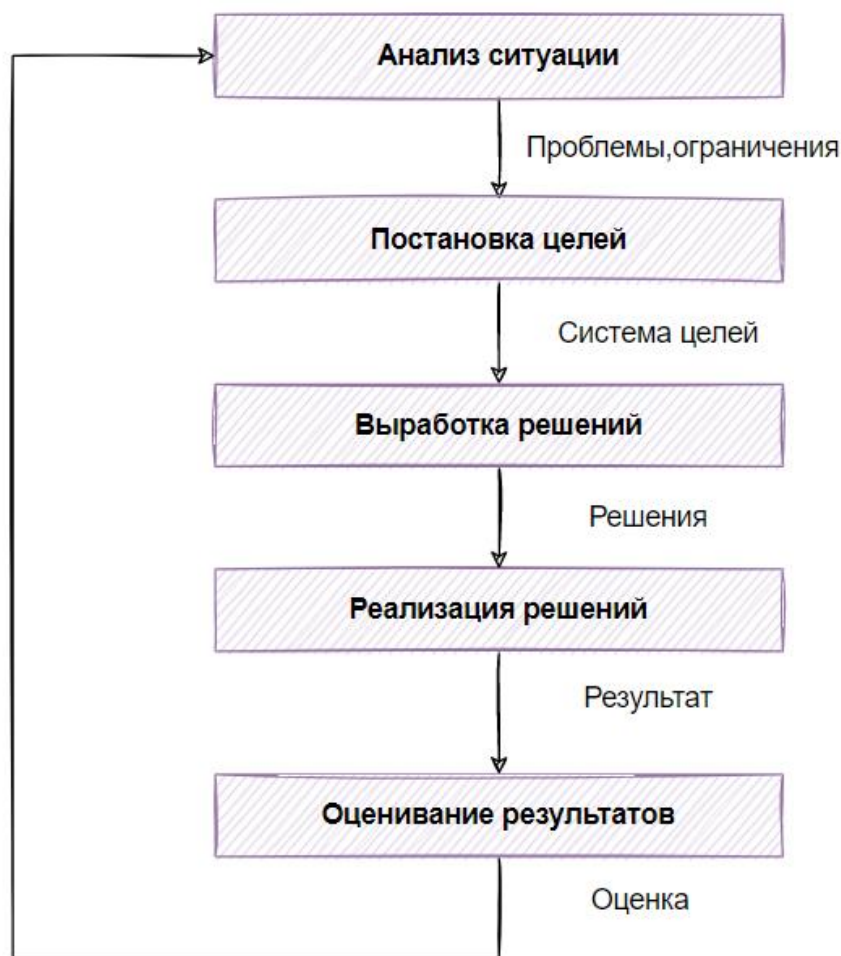


Рисунок 2 – Системный анализ (UML – схема)

1.3 Линейный регрессионный анализ

Регрессионный анализ – это взаимосвязь между двумя переменными, одна из которых является зависимой, а другая количественной. Зависимая переменная непосредственно зависит от одной или нескольких

количественных переменных, которые в свою очередь независимые. Зависимую переменную в регрессионном анализе называют результирующей, а переменные факторы принято называть предикторами или переменными объяснения.

Среднее значение итоговой переменной и средние значения предикторов имеют взаимосвязь, и записывается все в виде уравнения регрессии. Уравнение регрессии — это некая функция, которая исходя из статистических данных зависимой и управляемых переменных, подбирается определенным математическим образом. Зачастую подбирается (используется) линейная функция и при таком раскладе речь заходит об линейном регрессионном анализе [1]. Связи бывают различных видов рисунок 3:

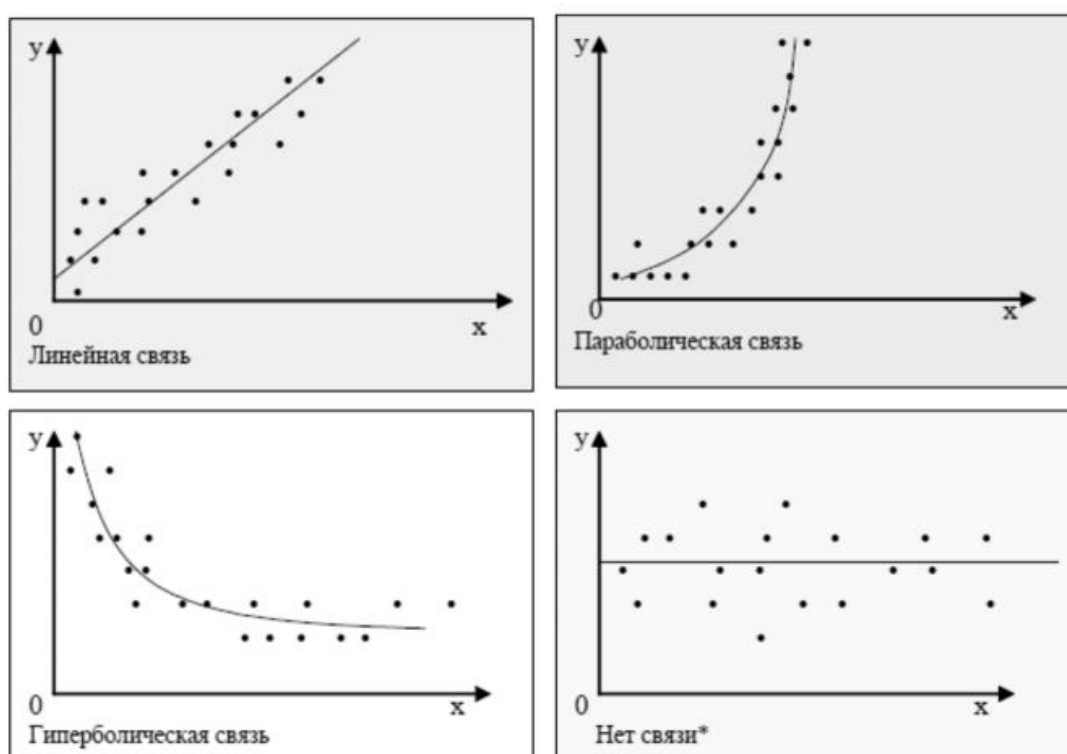


Рисунок 3 – Виды регрессионного анализа

Регрессионный анализ напрямую связан с корреляционным анализом. Корреляционный анализ исследует направление и связь между управляемыми (количественными переменными). Корреляционный анализ и регрессионный анализ фактически исследуют одну и ту же форму зависимости между регрессорами (управляемыми переменными). На самом деле данные два метода изучают одну и ту же взаимосвязь, но с разных ракурсов, дополняя друг друга на выходе, которого мы получаем качественную понятную информацию. В первую очередь применяется корреляционный анализ, а после следует применение регрессионного анализа. После того как мы убедимся в наличии взаимосвязи между переменными, используя корреляционный

анализ, после этого можно приступать к выражению формы этой связи при помощи регрессионного анализа.

Цель регрессионного анализа, заключается в предсказании ожидаемого среднего значения результирующей переменной, используя уравнение регрессии.

Главными задачами регрессионного анализа являются:

- определить вид и форму зависимости;
- определение и анализ параметров уравнения регрессии;
- проверить значимости определенных коэффициентов уравнения регрессии;
- проверить значимости отдельных коэффициентов уравнения;
- исследовать характеристики точности модели;
- построить точечные и интервальные прогнозы результирующей переменной.

Регрессионный и корреляционный анализ показывает только количественные зависимости между переменными. Причины, по которым так происходит и следствие этих действий регрессионный анализ отобразить не может. Гипотезы о причинах и следствия должны выражаться и строиться, исключительно путем теоретического анализа содержания тестируемого (изучаемого) явления или в нашем случае пациента.

1.4 Постановка задачи дипломной работы

Задачей дипломной работы является сбор, анализ и обработка информации. Данная работа в первую очередь направлена на решение проблем людей. Сбор и анализ данных, который будет проводиться, даст возможность визуализировать и выявить факторы риска заболевания людей, которые страдают таким заболеванием как инсульт. И с целью предотвращения заболевания инсультом, в данной работе будет представлена модель для предсказания заболеет человек или нет, исходя из факторов на основе искусственного интеллекта, который смотря на данные сможет выдавать prediction с точностью более 80 процентов.

В Казахстане каждый год насчитывается около 40 тысяч инцидентов связанные с болезнями кровообращения (инфаркты и инсульты). Почти половина заканчиваются летальным исходом, а если нет, то остаются глубокими инвалидами.

По данным Всемирной Организации Здравоохранения (ВОЗ), инсульт является второй ведущей причиной смерти в мире и составляет 11% всех смертей. Каждый год около 15 миллионов человек сталкиваются с инсультом, 6 миллионов из них заканчивают летальным исходом и еще порядка 5 миллионов остаются с постоянной инвалидностью [2]. Многие из этих инсультов можно предотвратить путем формирования здоровых привычек, а наблюдение за теми, кто подвергается наибольшему риску, может значительно улучшить результаты. По этим причинам эта область требует дальнейшего

изучения, чтобы предотвратить воздействие этого события на большее количество жизней, чем необходимо. Исходя из этого работа посвящена применению языка R для того, чтобы провести не только количественный, но и качественный визуальный анализ данных.

2 ПРОЕКТНАЯ ЧАСТЬ

2.1 Актуальность задачи

В последние годы число людей, страдающих инсультом чрезвычайно высокое. По результатам источника WHO Global Health Estimates, за 2019 год инсульт занял второе место по частоте летальных исходов рис.4, а на 2021 год это число не удалось минимизировать. В связи с этим актуальность данной работы несомненно высока, ведь в ней решаются проблемы людей [2].

Основные причины смерти в мире



Источник: WHO Global Health Estimates.

Рисунок 4 – Основные причины смерти в мире

2.2 Разработка и анализ системы с помощью RStudio

Обработка и анализ данных в значительной степени реализуется на основе статистических методов [1-3]. В информационных системах эти методы реализуются в виде программных модулей, в отдельных случаях в виде отдельных подсистем. Отметим, наличие на сегодня значительного количества эффективных инструментов для проведения вычислений и визуализации результатов, в подавляющем большинстве случаев, обработка производится с применением статистических пакетов, применение которых в

составе программного обеспечения подсистем информационных систем затруднено в связи с закрытостью исходного кода. Отметим также, что данные представлены в ограниченном наборе форматов, что также определяет определенные сложности при проектировании программного обеспечения. Положительным является возможность создания приложений на основе скриптов в среде R.

Вместе с тем, принятие решений на основе анализа данных требует определенной подготовки, связанной с необходимостью учета таких особенностей статистики, как принятие гипотез, определение уровня надежности, учет статистической значимости, мощности критериев и других характеристик. То есть для лиц, принимающих решение требуется определенная теоретическая подготовка, в тоже время, возможности языка R по визуализации результатов анализа значительно облегчает задачу оценки и выбора решения при решении задач анализа и прогнозирования. Отметим также, что R — это свободная программная среда с открытым исходным кодом.

Рассмотрим применение программной среды R для визуализации результатов обработки данных, важность этого метода обоснована выше. Отметим, что R обрабатывает и данные в распространенных форматах, например в очень распространенном формате *.xls.

2.3 Очистка данных

Чтобы правильно проанализировать свои данные и построить модель для прогнозирования, первым шагом является очистка данных. Большая часть данных уже очищены и готовы к анализу. Однако в наборе данных есть три столбца, над которыми следует поработать. Во-первых, один пациент, который идентифицировал себя как «неизвестный», был заменен на женщину. Это потому, что в наборе данных было больше женщин, чем мужчин. Также столбец индекс массы тела (ИМТ) и столбец больных, которые курят следовало оптимизировать. В столбце индекса массы тела отсутствует 3,93% записей, а в столбце предпочтений в отношении курения отсутствует 30% данных. Для четкости отображения данных я решил поместить среднее значение 28,89 в недостающие значения в столбце индекса массы тела. Этот подход оказал минимальное влияние на изменчивость, поскольку был небольшой процент отсутствующих записей. Кроме того, столбец с указанием того, курит пациент или нет, кажется не очень удобным, и нужно было задуматься о том, как обрабатывать нецифровые данные. Тем не менее, с целью устранить эти несоответствия данных, пришлось взять пропорции данных о курении которые отсутствовали и применить те же пропорции, чтобы дополнить недостающие данные. Результаты этой стратегии были значительно более благоприятными, чем полное исключение переменных или исключение элементов с неизвестными, поскольку существовала потенциальная связь между курением и инсультами. Однако недостатком

этого подхода является то, что он не учитывает потенциальную связь между курением и инсультом (или другими переменными), то есть использует случайное распределение. На рисунке 5 представлен фрагмент кода работы с данными.

```

24 # Импортируем датасет и даем короткое название data для удобства
25 data <- data.stroke.data
26 # Вытаскиваем информацию о данных
27 summary(data)
28 # Просмотр классов данных
29 str(data)
30 # Просмотр всех отдельных категориальных переменных
31 lapply(subset(data, select = c(gender, ever_married, work_type, Residence_type, bmi, smoking_status)), unique)
32
33 ## Работа со столбцом bmi (ИМТ)
34 # Проверим тип данных
35 class(data$bmi)
36 # Преобразовываем ИМС в числовое значение с помощью numeric
37 data$bmi <- as.numeric(data$bmi)
38 # Делаем проверку снова
39 class(data$bmi)
40 # Просмотр суммарной статистики данных
41 summary(data$bmi)
42 # Заменяем N/A значения в столбце ИМТ на среднее
43 data$bmi[is.na(data$bmi)] <- mean(data$bmi, na.rm=TRUE)
44 summary(data$bmi)
45
46 ## Работа со столбцом пол
47 # Подсчет уникальных переменных в столбце пол
48 table(data$gender)
49 # Заменяем единственное значение "other" на значение "Female" т.к. женщин в наборе данных больше
50 data$gender <- ifelse(data$gender == "other", "Female", data$gender)
51 table(data$gender)
52
53 ## Работа со столбцом Smoking Status
54 # Подсчет уникальных переменных в столбце пол
55 table(data$smoking_status)
56 # Рассчитаем вероятность бывших курильщиков FS, нынешних курильщиков S и некурящих NS, учитывая, что в столбце smoke_status есть т
57 prob.FS <- 864 / (864 + 1845 + 773)
58 prob.NS <- 1845 / (864 + 1845 + 773)
59 prob.S <- 773 / (864 + 1845 + 773)
60
61
62 ## На всякий случай копируем данные
63 data2 <- data
64 # Замена «unknown» в smoke_status на другие 3 переменные в соответствии с их весом
65 data2$rand <- runif(nrow(data2))
66 data2 <- data2%>mutate(Probability = ifelse(rand <= prob.FS, "formerly smoked", ifelse(rand <= (prob.FS+prob.NS), "never smoked",
67 data2 <- data2%>mutate(smoking_status = ifelse(smoking_status == "Unknown", Probability, smoking_status))
68 # Проверка уникальных значений нового столбца статуса курения и их количество
69 table(data2$smoking_status)
70 # Удаляем ненужные столбцы, которые ни на что не влияют
71 health <- subset(data2, select = -c(rand, Probability, smoking_status))
72 colnames(health)[12] <- "smoking_status"
73 # просмотр первых 10 строк
74 head(health, 10)
75 # «health» – это окончательный измененный набор данных, который будет использоваться для раздела EDA ниже.

```

Рисунок 5 – Очистка данных в RStudio (Фрагмент кода)

2.4 Построение гистограмм

Гистограммы являются простыми и легкими в чтении график, но дающие не мало информации. Гистограммы очень популярны и не редко используются в анализе данных, визуализация является важным инструментом для освоения информации и в принятии дальнейших решений. Для применения гистограмм были использованы такие библиотеки как: lattice, ggplot2 из пакета graphics [3].

В соответствии с рисунком 6.1 определение пола пациентов выявило что, пациентов женского пола больше, чем мужчин. Одна запись, которая была указана как «Другое», была добавлена в раздел «Женщины», поскольку большинство пациентов — женщины.

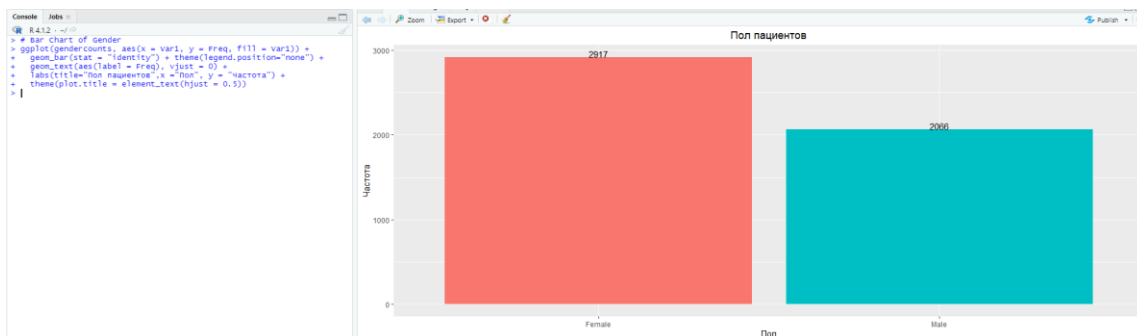


Рисунок 6.1 - Пол пациентов (График)

В ходе визуализации инсульт статуса пациентов выявилось что, количество пациентов, у которых не было инсульта, на 4485 превышает количество пациентов, у которых он был в соответствии с рисунком 6.2. В связи с этим можем сделать вывод, что не все пациенты имеют инсульт, но они все еще могут быть подвержены данному заболеванию.

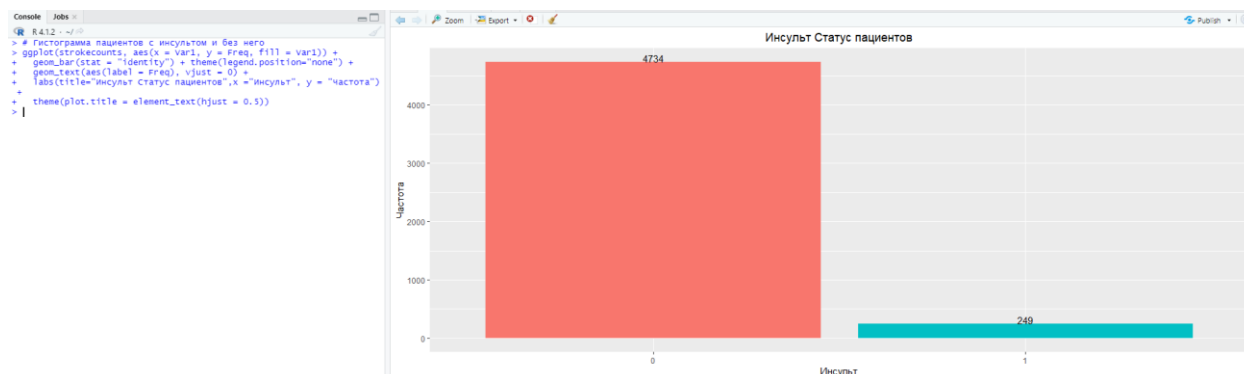


Рисунок 6.2 – Инсульт статус пациентов (График)

График статуса гипертонии показывает, что число пациентов без артериальной гипертонии значительно превышает число пациентов с артериальной гипертонией, но разрыв немного меньше, чем разрыв среди жертв инсульта в соответствии с рисунком 6.3. Делая вывод хочется сказать, что процент пациентов не имеющих инсульт, но имеющих артериальную гипертонию составляет 4,24% в то время как процент пациентов имеющих инсульт и артериальную гипертонию составляет 15,5%. Это означает что артериальная гипертония оказывает положительное влияние на появление инсульта.

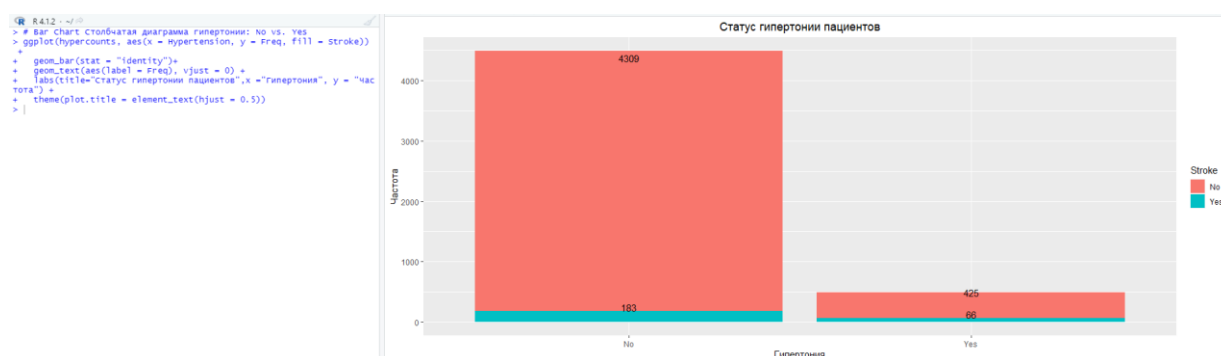


Рисунок 6.3 – Статус гипертонии пациентов (График)

Исследование показало, что разрыв между пациентами с сердечными заболеваниями и без них больше напоминает разрыв между пациентами с

инсультами и без них в соответствии с рисунком 6.4. Исходя из этого можем сделать вывод, что пациенты с сердечными заболеваниями скорее всего имеют инсульт.

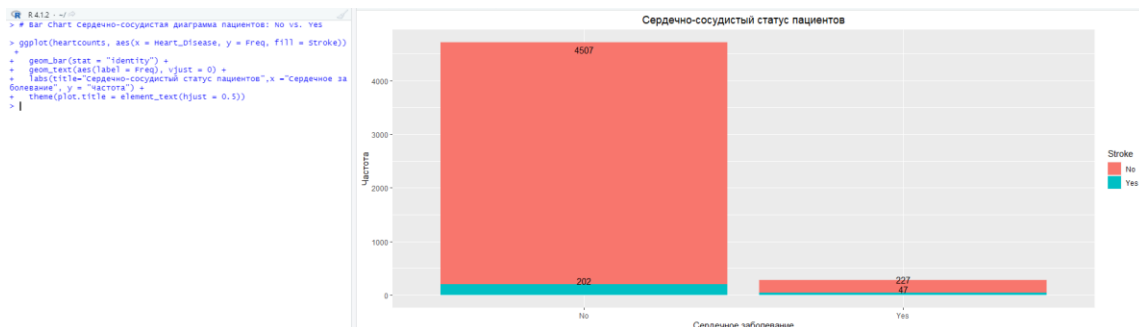


Рисунок 6.4 – Сердечно – сосудистый статус пациентов (График)

Визуализация показала, что примерно равное количество пациентов работают на государственных должностях, а именно 636, работают не по найму (на себя) 802 и являются детьми 671. Большинство пациентов работают в частных компаниях 2852, а небольшое количество никогда не работало 802 в соответствии с рисунком 6.5.

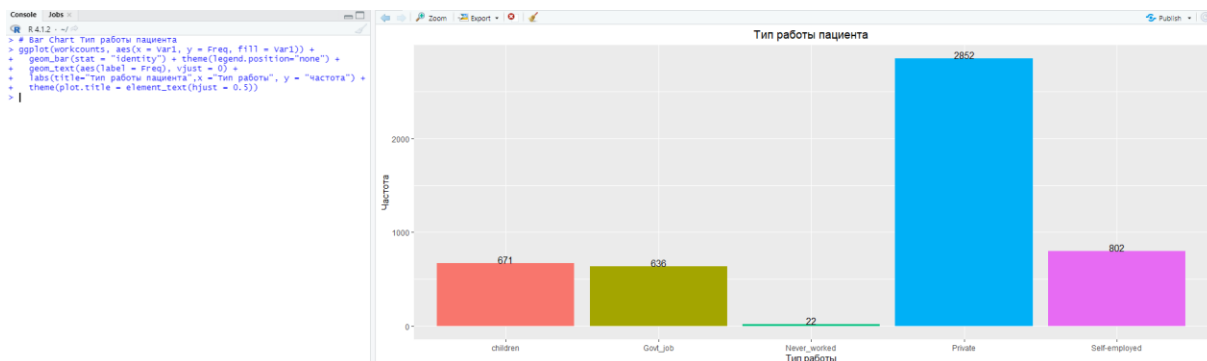


Рисунок 6.5 – Тип работы пациентов (График)

На гистограмме видно, что пациенты, которые состоят в браке превышают пациентов, которые не состоят в браке почти в два раза в соответствии с рисунком 6.6.

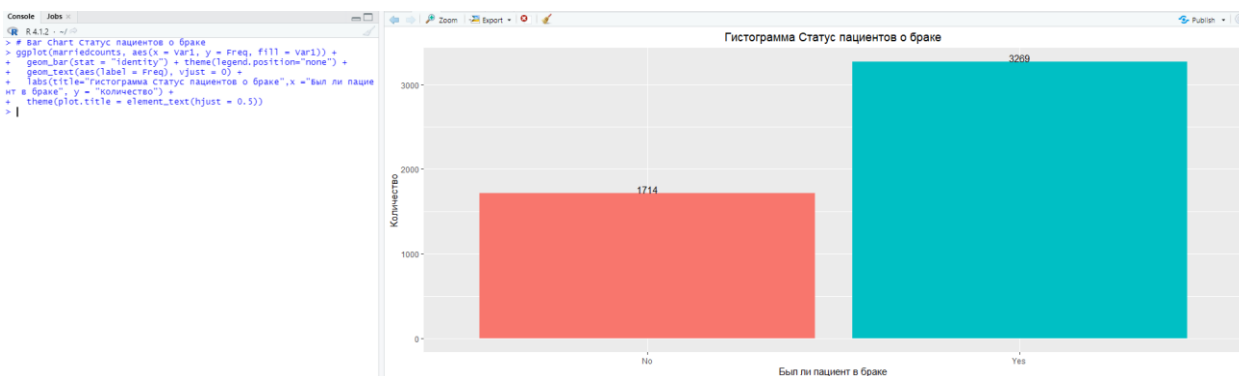


Рисунок 6.6 – Статус брака пациентов (График)

В ходе анализа выяснилось, что больные практически равномерно распределены между сельскими и городскими жителями. Из них сельскими являются 2453 пациента и 2530 проживающих в городских условиях в соответствии с рисунком 6.7.

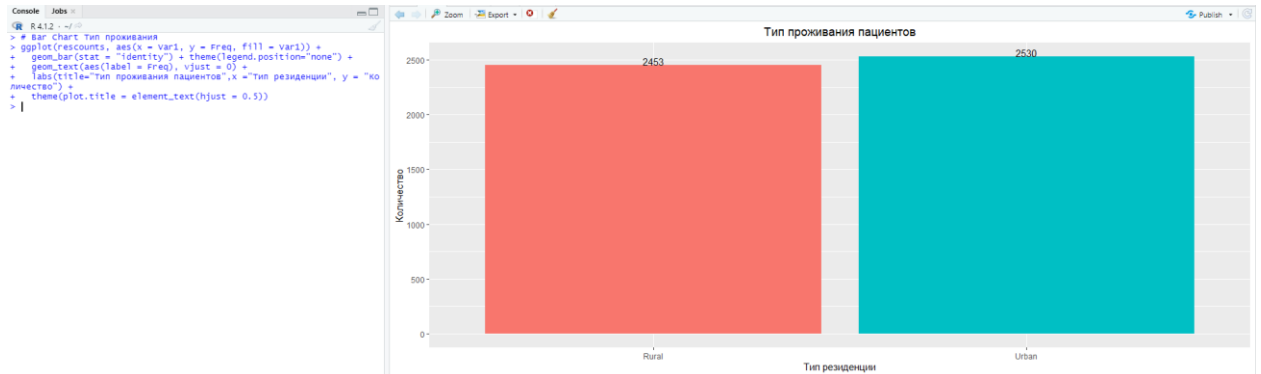


Рисунок 6.7 – Тип проживания пациентов (График)

Как мы уже узнали в ходе анализа были пациенты о которых не было известно курят они или нет. Таким образом неизвестные данные были случайным образом добавлены к трем категориям выше на основе вероятности. Большинство пациентов либо никогда не курили в соответствии с рисунком 6.8. Данные для бывших и нынешних курильщиков аналогичны.

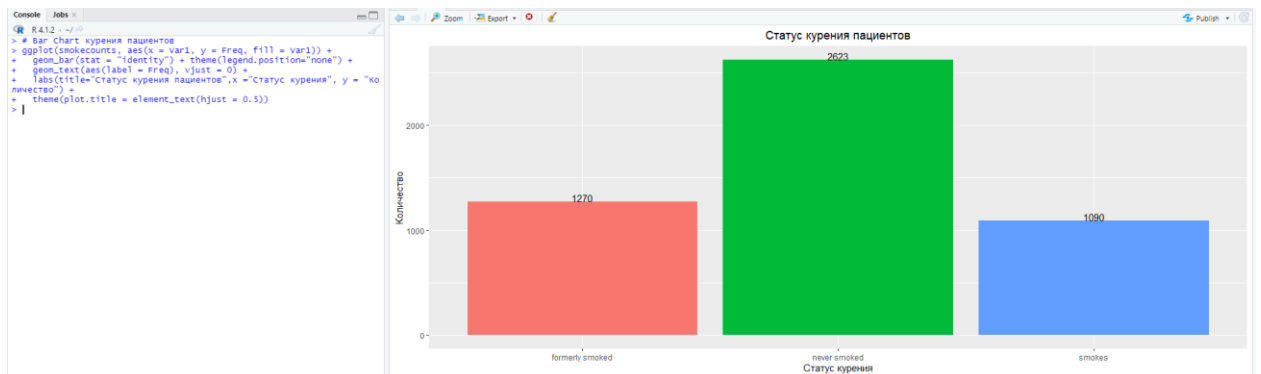


Рисунок 6.8 – Статус курения пациентов (График)

Для получения большей информации используя функцию `summary()` мы получаем полную информацию данных на текущий момент исследования в соответствии с рисунком 7.

```
welcome! want to learn more? See two factoextra-related books at https://goo.gl/ve3wBa
> summary(data2)
  id          gender          age      hypertension      heart_disease      ever_married
Min.   : 67      Length:4983      Min.   : 0.08      Min.   :0.00000      Min.   :0.00000      Length:4983
1st Qu.:17703    Class :character      1st Qu.:25.00    1st Qu.:0.00000      1st Qu.:0.00000      Class :character
Median :36896    Mode  :character      Median :45.00    Median :0.00000      Median :0.00000      Mode  :character
Mean   :36497                                Mean  :43.26     Mean  :0.09853      Mean  :0.05499
3rd Qu.:54669                                3rd Qu.:61.00    3rd Qu.:0.00000      3rd Qu.:0.00000
Max.   :72940                                Max.   :82.00     Max.   :1.00000      Max.   :1.00000

  work_type      Residence_type      avg_glucose_level      bmi      smoking_status      stroke
Length:4983      Length:4983      Min.   : 55.12      Min.   :10.30      Length:4983      Min.   :0.00000
Class :character      Class :character      1st Qu.: 77.17      1st Qu.:23.80      Class :character      1st Qu.:0.00000
Mode  :character      Mode  :character      Median : 91.85      Median :28.40      Mode  :character      Median :0.00000
Mean   :106.07      Mean   :28.92      3rd Qu.:113.98      3rd Qu.:32.80      Mean   :0.04997
Max.   :271.74      Max.   :97.60      Max.   :1.00000      Max.   :1.00000

  rand          Probability      smoking_status
Min.   :0.0000122      Length:4983      Length:4983
1st Qu.:0.2460686      Class :character      Class :character
Median :0.4935885      Mode  :character      Mode  :character
Mean   :0.4964257
3rd Qu.:0.7501778
Max.   :0.9999525
```

Рисунок 7 – Информация о данных по функции summary()

2.5 Гистограммы с наложением нормального распределения

Рассмотрим применение программной среды R для визуализации кривой распределения данных. Эта функция программной среды R также реализована.

R позволяет отображать как одномерные, так объемные графики необходимые для визуализации анализа данных.

В ходе исследования было выявлено что, возраст пациентов в исследовании близок к нормальному распределению со средним значением 43,26 по функции summary(). Основываясь на информации из функции summary() ранее и на приведенной выше диаграмме, большинству пациентов около 40 лет в соответствии с рисунком 8.1.

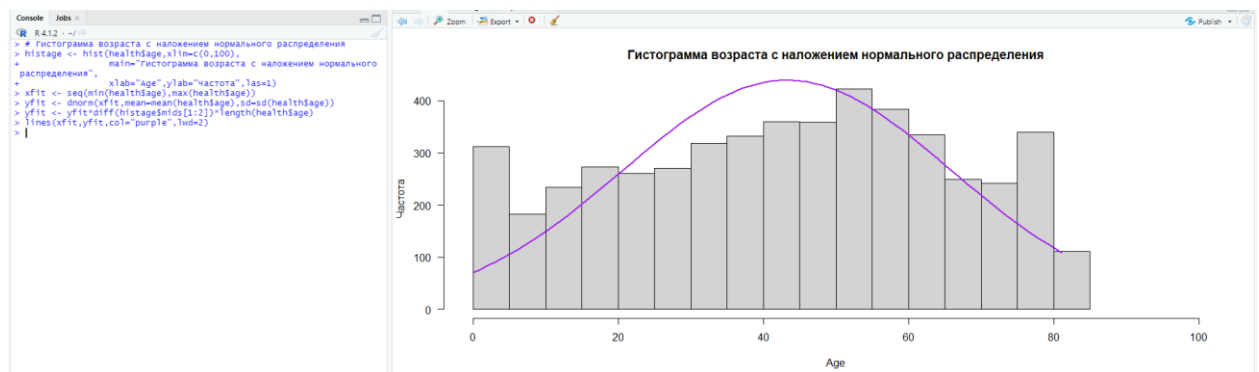


Рисунок 8.1 – Возраст с наложением нормального распределения (График)

Благодаря гистограмме удалось выявить что, средние уровни глюкозы у пациентов в исследовании искажены вправо, со средним значением 106,07 из функции summary() ранее. Делая вывод хочется подметить что, данный показатель не сильно отличается от нормы, в то время как уровень сахара в

человеческой крови должен располагаться в диапазоне от 80-100 миллиграммов на децилитр в соответствии с рисунком 8.2.

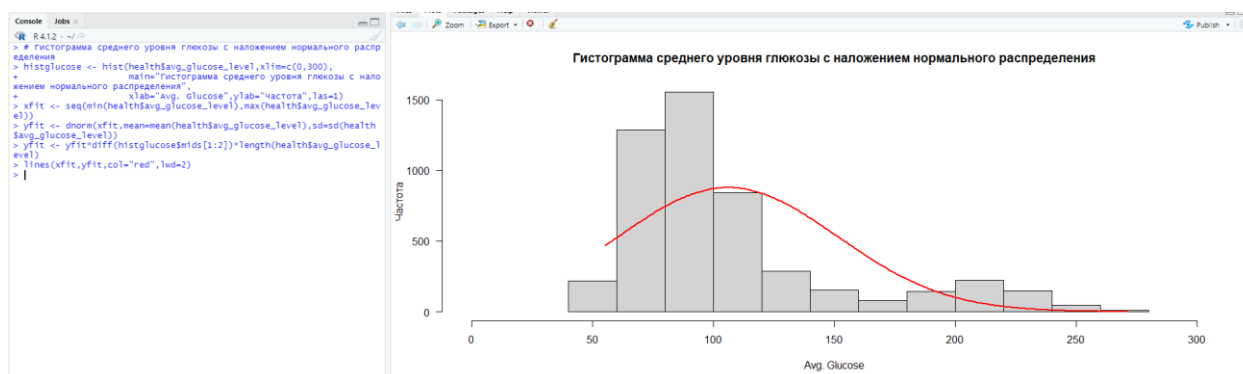


Рисунок 8.2 – Средний уровень глюкозы с наложением нормального распределения (График)

В ходе исследования, выяснилось, что данные для индекса массы тела пациента искажены вправо со средним значением 28,89 из функции summary() выше после модификации. Исходя из этой информации можно сделать вывод что среднее значение индекса массы тела отличается от нормы [4]. Пациенты в основном страдают избыточным весом, а некоторые и вовсе ожирением I и II степени. Напомним, что нормальными показателями считаются от 18,5 до 25 в соответствии с рисунком 8.3.

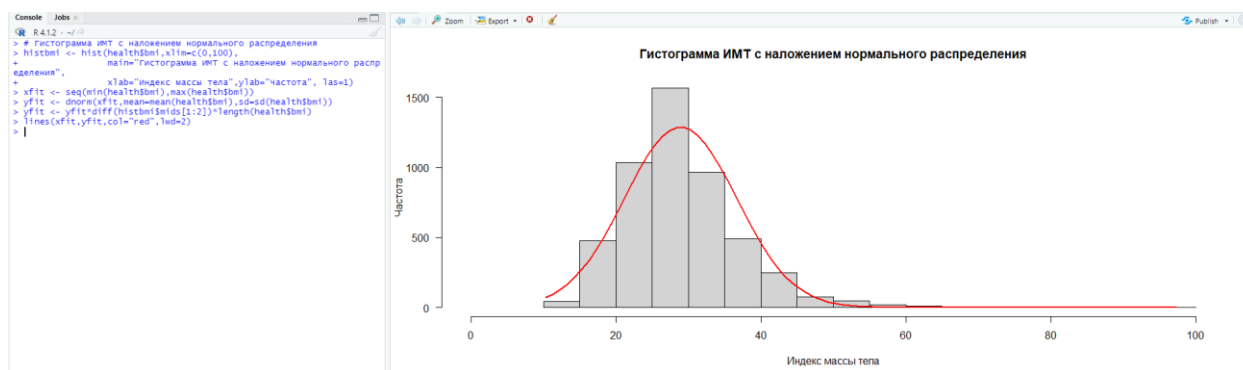


Рисунок 8.3 – Индекс массы тела с наложением нормального распределения (График)

2.6 Описание и возможности Boxplot

Графическое отображение основных характеристик выборки бокс плотом (boxplot). Боксплоты были специально придуманы известным статистиком Джоном Тьюки, для того чтобы быстро, эффективно и наглядно отражать основные робастные (устойчивые) характеристики выборки.

Боксплот диаграмма, или другими словами ящик имеющий усы называется так из-за своего графического строения: медиана которая отображается в качестве линии, находится внутри прямоугольника «ящик», размеры которого зависят от показателей разброса или точности оценки главного параметра. Вдобавок к этому у прямоугольника имеются «усы», также зависящие от длины показателей разброса или точности другими словами их называются размахом. Графики этого типа очень удобные и пользуются большой популярностью, поскольку содержат в себе очень полную статистическую характеристику анализируемой совокупности. Более того, диаграммы «ящика с усами» комфортно использовать для визуальной экспресс-оценки разницы между двумя и более группами [5].

Для того чтобы в R построить диаграмму «ящика с усами» применяется функция `boxplot()`. Визуализация, которую можно получить при помощи данной функции изображено ниже. Рассмотрим информацию отображаемую боксплотом рисунок 9:

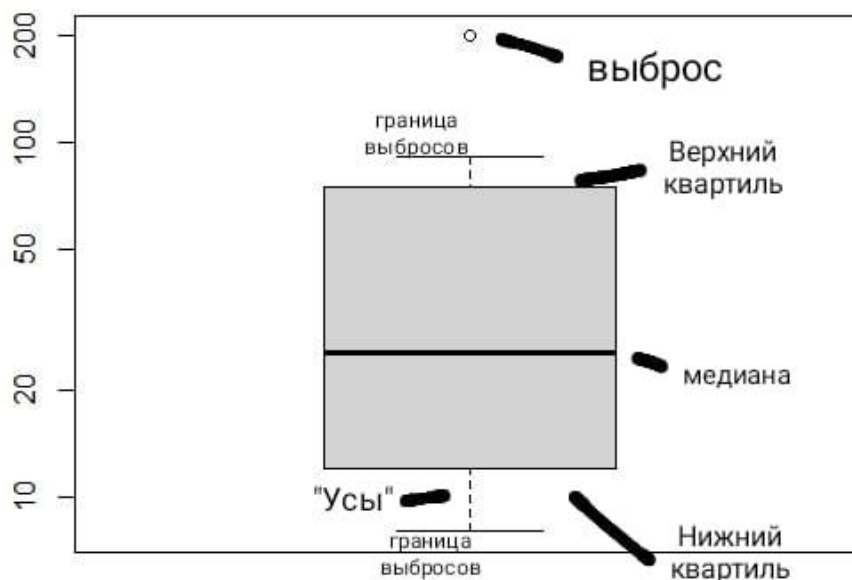


Рисунок 9 – Интерпретация боксплота

Медиана – это значение элемента в центре ранжированной выборки. Медиана нивелирует имеющиеся нетипичные выбросы, поэтому в центре показатель медианы, а не стандартное среднее. Межквартильный разброс – Interquartilerange (IQR). В дескриптивной статистике IQR, является показателем статистической дисперсии, равной разнице между 75-м и 25-м перцентилями или между верхним и нижним квартилями. IQR – это разница между первым и третьим квартилем, визуально это наблюдается на прямоугольной диаграмме данных (боксплоте). Этот показатель, определяется как 25% усеченный диапазон. Эти интервалы характеризуют масштаб исследуемых данных.

Для исследователя квартили удобный инструмент визуализации в боксплоте набора (выборки) данных. С учетом этого, межквартильный разброс (IQR) является наглядным показателем волатильности анализируемых данных.

Техника боксплотов позволяет наглядно отображать имеющиеся в рядах данных выбросы. В боксплотах обычно применяют следующие расчетные формулы для выбросов. Выбросами считаются варианты данных вне следующих границ:

- 25% перцентиль минус $1.5 \times \text{IQR}$;
- 75% перцентиль плюс $1.5 \times \text{IQR}$.

Если нет таких значений, то длина «уса» определяется минимальным и максимальным значением. При сильной волатильности, значения за пределами усов отмечаются отдельно, и характеризуются также, как выбросы.

Построим график уровня глюкозы у пациентов с инсультом и без него. Диаграмма с усами показывает относительно схожий средний уровень глюкозы у пациентов, перенесших инсульт, и у пациентов, не перенесших инсульт, с большим количеством высоких выбросов среди жертв, не перенесших инсульт в соответствии с рисунком 10.1.

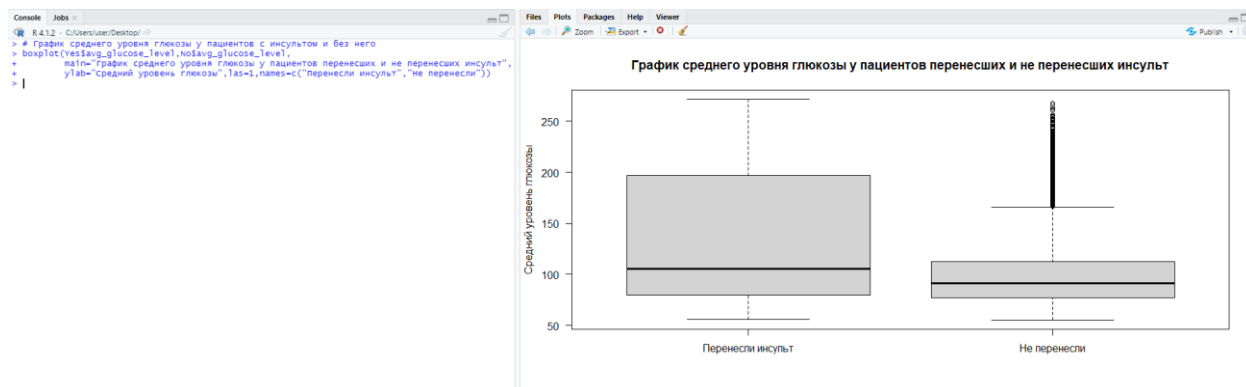


Рисунок 10.1 – Уровень глюкозы у пациентов перенесших инсульт и не перенесших (График)

Построение боксплот графика индекса массы тела у пациентов с инсультом и без него. На диаграмме показано относительно одинаковое среднее значение индекса массы тела у пациентов, перенесших инсульт, и у пациентов, не перенесших инсульт, с несколькими высокими выбросами среди жертв инсульта и большим количеством высоких выбросов среди жертв, не перенесших инсульт в соответствии с рисунком 10.2.

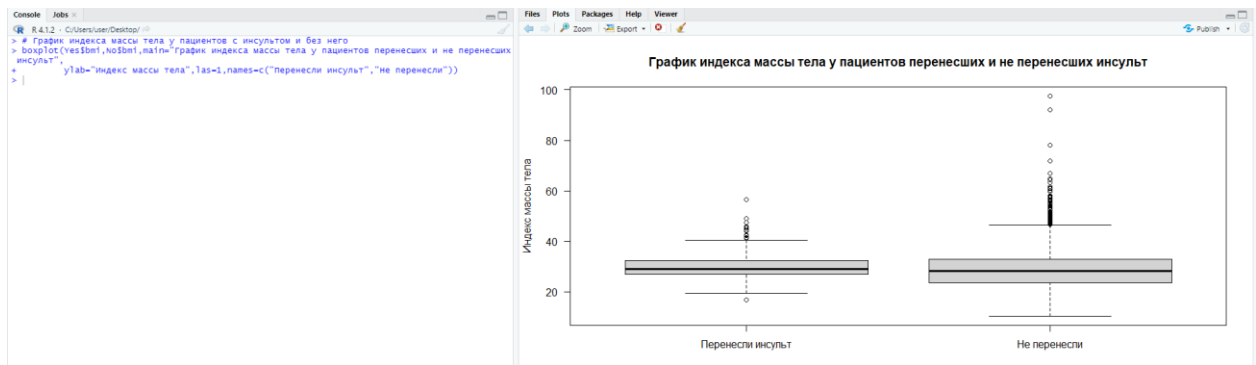


Рисунок 10.2 – Индекс массы тела у пациентов перенесших инсульт и не перенесших (График)

2.7 Линейная регрессия

Регрессионным анализом – называется способ изучения статистической взаимосвязи между двумя переменными, одна из которых является зависимой, а другая количественной. Зависимая переменная непосредственно зависит от одной или нескольких количественных переменных, которые в свою очередь независимы. Зависимую переменную в регрессионном анализе называют результирующей, а переменные факторы принято называть предикторами или переменными объяснения.

Среднее значение результирующей переменной и средние значения предикторов имеют взаимосвязь, и выражается это все в виде уравнения регрессии. Уравнение регрессии – это такая функция, которая исходя из статистических данных зависимой и управляемых переменных, подбирается определенным математическим образом. Зачастую подбирается (используется) линейная функция и при таком раскладе уже говорят о линейном регрессионном анализе.

Регрессионный анализ напрямую связан с корреляционным анализом. Корреляционный анализ исследует направление и связь между управляемыми (количественными переменными). Корреляционный анализ и регрессионный анализ фактически исследуют одну и ту же форму зависимости между регрессорами (управляемыми переменными). На самом деле данные два метода изучают одну и ту же взаимосвязь, но с разных ракурсов, дополняя друг друга на выходе, которого мы получаем качественную понятную информацию. В первую очередь применяется корреляционный анализ, а после следует применение регрессионного анализа. После того как мы убедимся в наличии взаимосвязи между переменными, используя корреляционный анализ, после этого можно приступать к выражению формы этой связи при помощи регрессионного анализа.

Цель регрессионного анализа, заключается в предсказании ожидаемого среднего значения результирующей переменной, используя уравнение регрессии.

Главными задачами регрессионного анализа являются:

- определить вид и форму зависимости;
- оценка параметров уравнения регрессии;
- исследовать характеристики точности модели;
- построить точечные и интервальные прогнозы результирующей переменной [6].

Регрессионный и корреляционный анализ показывает только количественные зависимости между переменными. Причины, по которым так происходит и следствие этих действий регрессионный анализ отобразить не может. Гипотезы о причинах и следствия должны выражаться и строиться, исключительно путем теоретического анализа содержания тестируемого (изучаемого) явления или в нашем случае пациента.

На приведенной ниже диаграмме в соответствии с рисунком 11, показана корреляция всех числовых переменных в очищенных данных. Значения в диагональных ячейках — это минимальное и максимальное значения. Например, минимальный ИМТ равен 10,3, тогда как самый высокий ИМТ равен 97,6. Из коррелограммы и таблицы корреляции видно, что все числовые переменные положительно коррелируют с предикторной переменной [stroke]. Возраст имеет самый высокий индекс корреляции с инсультом.

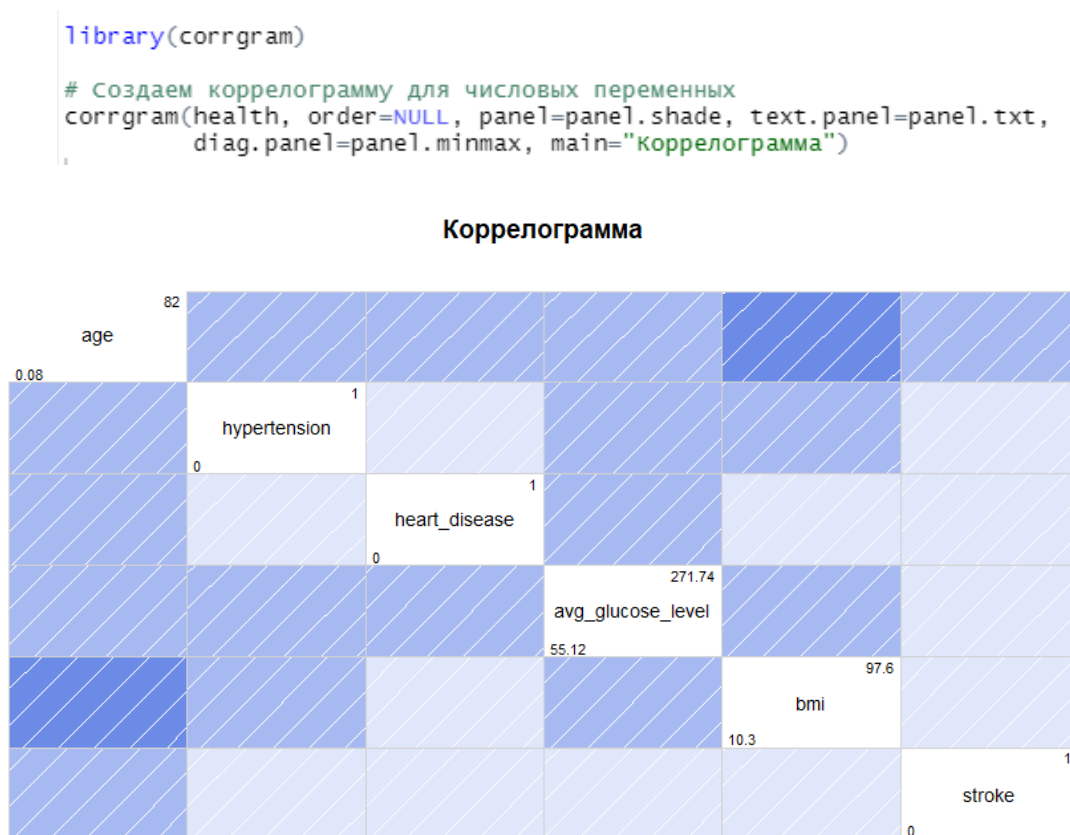


Рисунок 11 – Корреляция (График)

Результат линейной регрессии:

```
## Модель 1: обобщенная линейная модель – логистическая регрессия – DummyVar

glm_fit <- glm(stroke~., data=healthdummytrain, family = binomial)
summary(glm_fit)

stepdummy <- stepAIC(glm_fit)

summary(stepdummy)

# Делаем предикшн
probabilities <- stepdummy %>% predict(healthdummytest, type = "response")
predicted_classes <- ifelse(probabilities > 0.5, 1, 0)
# Точность модели
print(paste("Model Accuracy : ", mean(predicted_classes == healthdummytest$stroke)))

# Anova Table
anova(stepdummy)

> glm_fit <- glm(stroke~., data=healthdummytrain, family = binomial)
> summary(glm_fit)

Call:
glm(formula = stroke ~ ., family = binomial, data = healthdummytrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1774  -0.3235  -0.1614  -0.0807   3.6487

Coefficients: (5 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.986912   0.739086 -10.806 < 2e-16 ***
age           0.075208   0.007146  10.525 < 2e-16 ***
hypertension  0.415587    0.198414   2.095 0.036212 *
heart_disease 0.209405    0.235058   0.891 0.373002
avg_glucose_level 0.004827   0.001435   3.365 0.000765 ***
bmi          -0.005821   0.013943  -0.417 0.676355
gender_Female -0.002585    0.171640  -0.015 0.987984
gender_Male   NA             NA         NA      NA
ever_married_No 0.100076    0.274635   0.364 0.715561
ever_married_Yes NA             NA         NA      NA
work_type_children 0.963666    1.139327   0.846 0.397653
work_type_Govt_job 0.220309    0.288619   0.763 0.445273
work_type_Never_worked -8.931887  388.129498 -0.023 0.981640
work_type_Private 0.510509    0.202882   2.516 0.011860 *
`work_type_Self-employed` NA             NA         NA      NA
Residence_type_Rural -0.169026    0.167977  -1.006 0.314297
Residence_type_Urban NA             NA         NA      NA
`smoking_status_formerly smoked` -0.009679   0.234569  -0.041 0.967086
`smoking_status_never smoked` -0.180794   0.220805  -0.819 0.412904
smoking_status_smokes NA             NA         NA      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1358.8  on 3487  degrees of freedom
Residual deviance: 1075.5  on 3473  degrees of freedom
AIC: 1105.5

Number of Fisher Scoring iterations: 14

>
> stepdummy <- stepAIC(glm_fit)
Start: AIC=1105.55
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
  bmi + gender_Female + gender_Male + ever_married_No + ever_married_Yes +
  work_type_children + work_type_Govt_job + work_type_Never_worked +
  work_type_Private + `work_type_Self-employed` + Residence_type_Rural +
  Residence_type_Urban + `smoking_status_formerly smoked` +
  `smoking_status_never smoked` + smoking_status_smokes
```

Рисунок 12.1 – Вычисления регрессионной модели

```

Step: AIC=1097.71
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
      bmi + work_type_children + work_type_Govt_job + work_type_Private +
      Residence_type_Rural + `smoking_status_never smoked`

      Df Deviance   AIC
- bmi                1  1075.9 1095.9
- work_type_Govt_job 1  1076.3 1096.3
- work_type_children 1  1076.4 1096.4
- heart_disease      1  1076.5 1096.5
- Residence_type_Rural 1  1076.8 1096.8
- `smoking_status_never smoked` 1  1076.8 1096.8
<none>                1075.7 1097.7
- hypertension       1  1080.0 1100.0
- work_type_Private   1  1082.3 1102.3
- avg_glucose_level  1  1086.6 1106.6
- age                1  1227.7 1247.7

Step: AIC=1095.9
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
      work_type_children + work_type_Govt_job + work_type_Private +
      Residence_type_Rural + `smoking_status_never smoked`

      Df Deviance   AIC
- work_type_Govt_job 1  1076.5 1094.5
- work_type_children 1  1076.7 1094.7
- heart_disease      1  1076.8 1094.8
- `smoking_status_never smoked` 1  1076.9 1094.9
- Residence_type_Rural 1  1077.0 1095.0
<none>                1075.9 1095.9
- hypertension       1  1080.0 1098.0
- work_type_Private   1  1082.5 1100.5
- avg_glucose_level  1  1086.7 1104.7
- age                1  1231.9 1249.9

Step: AIC=1094.46
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
      work_type_children + work_type_Private + Residence_type_Rural +
      `smoking_status_never smoked`

      Df Deviance   AIC
- work_type_children 1  1077.1 1093.1
- heart_disease      1  1077.3 1093.3
- `smoking_status_never smoked` 1  1077.5 1093.5
- Residence_type_Rural 1  1077.6 1093.6
<none>                1076.5 1094.5
- hypertension       1  1080.5 1096.5
- work_type_Private   1  1082.8 1098.8
- avg_glucose_level  1  1087.3 1103.3
- age                1  1234.1 1250.1

Step: AIC=1093.13
stroke ~ age + hypertension + heart_disease + avg_glucose_level +
      work_type_Private + Residence_type_Rural + `smoking_status_never smoked`

```

Рисунок 12.2 – Вычисления регрессионной модели

```

> summary(stepdummy)

call:
glm(formula = stroke ~ age + hypertension + avg_glucose_level +
     work_type_Private, family = binomial, data = healthdummytrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1099 -0.3271 -0.1672 -0.0788  3.9230

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.133317   0.482303  -16.863 < 2e-16 ***
age           0.074104   0.006421   11.541 < 2e-16 ***
hypertension  0.387138   0.196490    1.970 0.048807 *
avg_glucose_level 0.004844   0.001381    3.507 0.000453 ***
work_type_Private 0.430842   0.172837    2.493 0.012675 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1358.8  on 3487  degrees of freedom
Residual deviance: 1080.4  on 3483  degrees of freedom
AIC: 1090.4

Number of Fisher Scoring iterations: 7

>
> # Делаем предикшн
> probabilities <- stepdummy %>% predict(healthdummytest, type = "response")
> predicted_classes <- ifelse(probabilities > 0.5, 1, 0)
> # Точность модели
> print(paste("Model Accuracy : ", mean(predicted_classes == healthdummytest$stroke)))
[1] "Model Accuracy : 0.947157190635451"
>
>
> # Anova Table
> anova(stepdummy)
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev
NULL                    3487    1358.8
age                      1    254.339   3486    1104.5
hypertension             1     5.560   3485    1098.9
avg_glucose_level       1    12.188   3484    1086.7
work_type_Private       1     6.358   3483    1080.4
> |

```

Рисунок 12.3 – Вычисления и результат регрессионной модели

Общая линейаризованная модель. Результаты модели были довольно интересными и дали новое понимание по сравнению с исследовательским анализом данных. Модель показала, что возраст, гипертония, средний уровень глюкозы и то, работал ли человек в частной компании или нет, оказывали статистически значимое влияние на вероятность инсульта у человека. Новой информацией здесь является взаимосвязь между работой в частной компании и возможностью инсульта. На самом деле модель обнаружила, что вероятность инсульта значительно возрастает при работе в частной компании по сравнению с другими формами работы. Скорее всего, это не означает, что люди не должны работать в частных компаниях, а скорее указывает на связь между стрессом и инсультом в соответствии с рисунком 12.1-3.

Исследовательский анализ данных показал, что существует значительная связь между старением и инсультами, а также связь между гипертонией и высоким уровнем глюкозы с инсультом. Это позволяет очень легко принять эти результаты в качестве факторов риска.

Хотя тип работы еще не отображается в других моделях, это не означает, что его можно игнорировать. Как было сказано ранее, связь, скорее всего, связана не с типом работы, а с количеством стресса, который испытывает человек. Хорошим методом сокращения числа инсультов, вероятно, будет сосредоточение внимания на методах снижения стресса для тех, кто работает на стрессовой работе, что может быть более распространено в частном секторе, чем в государственном, что приводит к этой корреляции, наблюдаемой в обобщенной линейной модели. Методы снижения стресса могут варьироваться от стрессового мяча, лекарств, крика в подушку, обращения за профессиональной помощью или физических упражнений. Любой из этих и многих других методов может помочь снизить риск инсульта.

Не рекомендуется человеку увольняться с работы с целью снижения вероятности инсульта, однако, если кто-то находится в ситуации, когда он считает, что его работа вызывает больше стресса, чем он может вынести, человеку следует подумать о смене работы. В целях снижения других факторов риска рекомендуется использовать метод упражнений для снижения стресса. Упражнения могут помочь снизить индекс массы тела, а при правильном питании средний уровень глюкозы также снизится.

3 Описание программного обеспечения

3.1 Среда разработки R и RStudio

R — язык программирования для анализа и обработки данных и работы с визуализацией, данный язык достаточно популярен и знаменит тем что в

статистической обработке данных позволяет получить качественную графику. Я выбрал данный язык программирования для статистического анализа из-за нескольких преимуществ. Во-первых, мне понравился синтаксис программы и т.к. я раньше не сталкивался с R, я довольно быстро его освоил. R это бесплатная, свободная программная среда вычислений большим количеством инструментов. Язык является альтернативной реализацией S, хотя между языками есть не малые отличия, но в большинстве своём код на языке S также работает в среде R. R и S разработаны в Bell Labs [7].

Преимущества языка R которые мне удалось выделить:

- удобный;
- быстрый;
- надежный;
- простой;
- многофункциональный.

Недостатков как таковых нет, но возможно стоит выделить не полную русификацию программы, в связи с этим пользователь не знающий английский язык, может столкнуться с трудностями в переводе.

R — язык программирования способный интерпретировать исходный код, главным методом которого является командный или консольный интерпретатор. Язык чувствителен в регистрах поэтому его можно назвать регистрозависимым, касательно синтаксиса он похож, с одной стороны, на функциональные языки типа Scheme, с другой — на типичные современные сценарные языки, имеющий простой и обычный синтаксис с небольшим набором необходимых конструкций. Язык объектный: каждый объект программы имеет атрибут или набор атрибутов — именованный список значений, определяющих его. Также широко используется как статистическое ПО для обработки и анализа данных став фактически стандартом для статистического программного обеспечения.

В данной работе использовался также CSV-файл (файлы данных с разделителями-запятыми) — это файлы особого типа, которые можно создавать и редактировать в Excel. В CSV-файлах данные хранятся не в столбцах, а разделенные запятыми. Текст и числа, сохраненные в CSV-файле, можно легко переносить из одной программы в другую. CSV файл это файл имеющий расширение .csv с возможностью редактирования в нем содержится информация, но в отличии от обычного excel файла здесь каждая строка — это отдельная строка таблицы, а столбцы разделены друг от друга благодаря специальному символу к примеру точкой или запятой [8].

ЗАКЛЮЧЕНИЕ

Применение языка R, позволяет пользователю провести не только количественный, но визуальный анализ данных, что является неоспоримым преимуществом использования данной технологии. Визуальный экспресс-анализ данных занимает важное место при решении задач в различных

областях человеческой деятельности. Так, например врачу, который проводит обследование сложно понимать большое количество цифр, но благодаря визуальному экспрессу анализу, благодаря графику, врач принимает быстрое и правильное решение. В ряде случаев визуальный анализ является определяющим при принятии решений, оценке результатов деятельности и экспериментов. С учетом этого, разработка компонента с помощью интерпретируемого языка программирования R в обработке данных в информационных системах является актуальным.

На примере мы убедились как с помощью анализа и визуализации данных с помощью гистограмм, боксплотов, корреляции и общей линеаризованной модели возможно получить нужную аналитическую информацию и самое главное понятную за счет визуализации, что в свою очередь открывает новые возможности для дальнейшего интеллектуального анализа.

ПЕРЕЧЕНЬ ПРИНЯТЫХ СОКРАЩЕНИЙ, ТЕРМИНОВ

ИС – Информационная система.

R — Язык программирования способный интерпретировать исходный код.

БД – База данных.

IQR – межквартильный разброс или диапазон.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1 НОУ Интуит | Курс | Статистика // электронная версия на сайте // https://intuit.ru/studies/professional_skill_improvements/20689/courses/837/lecture/31370?page=2

2 Всемирная Организация Здравоохранения ссылка на сайт // <https://www.who.int/ru/news-room/fact-sheets/detail/the-top-10-causes-of-death>.

3 Приложение к книге «Статистический анализ и визуализация данных с помощью R» электронная версия на сайте // <https://github.com/ranalytics/r-tutorials>

4 Сулиев Ж.Н. Жумагалиев Б.И., Статья «Обработка и анализ данных в информационных системах с применением языка R» 2022. // <https://orcid.org/my-orcid?orcid=0000-0001-5127-3517>

5 Базовые графические возможности R: диаграммы размахов // электронная версия // https://r-analytics.blogspot.com/2011/11/r_08.html

6 Петров, В.Г. Суфиянов. Наглядная статистика. Используем R! электронная версия на сайте // <https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf>.

7 R (язык программирования) электронная версия на сайте // [https://ru.m.wikipedia.org/wiki/R_\(язык_программирования\)](https://ru.m.wikipedia.org/wiki/R_(язык_программирования))

8 CSV-файлы электронная версия на сайте // <https://gresgroup.ru/kak-izmenit-razdelitel-v-csv-fayle/>

9 Введение в R. Заметки по R: среда программирования для анализа данных и графики., У.Н. Венэблз, Д.М. Смит., электронная версия на сайте // http://www.ievbras.ru/ecostat/Kiril/R/Biblio_N/R_Rus/Venables.pdf

10 А. Б. Шипунов, Е. М. Балдин: Анализ данных с R (сборник статей, изданных в журнале Linux Format, электронная версия на сайте // <https://sociology.knu.ua/sites/default/files/course/materials/r1.pdf>

11 R for Data Science by Hadley Wickham & Garrett Grolemund электронная версия на сайте // <https://r4ds.had.co.nz/index.html>

Приложение А (обязательное)

Текст программы

```
library(ggplot2)
library(RColorBrewer)
library(ROSE)
library(hrbrthemes)
library(rpart)
library(rpart.plot)
library(data.tree)
library(caTools)
```

```
library(plyr)
library(dplyr)
library(caret)
library(cvms)
library(tibble)
library(MASS)
library(olsrr)
library(pROC)
library(DescTools)
library(randomForest)
library(factoextra)
```

```
# Импортируем датасет и даем короткое название data для удобства
data <- data.stroke.data
# Вытаскиваем информацию о данных
summary(data)
# Просмотр классов данных
str(data)
# Просмотр всех отдельных категориальных переменных
lapply(subset(data, select = c(gender, ever_married, work_type, Residence_type,
bmi, smoking_status)), unique)
```

```
## Работа со столбцом bmi (ИМТ)
# Проверяем тип данных
class(data$bmi)
# Преобразовываем ИМС в числовое значение с помощью numeric
data$bmi <- as.numeric(data$bmi)
# Делаем проверку снова
class(data$bmi)
# Просмотр суммарной статистики данных
summary(data$bmi)
# Заменяем N/A значения в столбце ИМТ на среднее
data$bmi[is.na(data$bmi)] <- mean(data$bmi,na.rm=TRUE)
summary(data$bmi)
```

```
## Работа со столбцом Пол
# Подсчет уникальных переменных в столбце пол
table(data$gender)
# Заменяем единственное значение "other" на значение "Female" т.к. женщин в
наборе данных больше
data$gender <- ifelse(data$gender == "Other", "Female", data$gender)
table(data$gender)
```

```

## Работа со столбцом Smoking Status
# Подсчет уникальных переменных в столбце Пол
table(data$smoking_status)
# Рассчитаем вероятность бывших курильщиков FS, нынешних курильщиков
S и некурящих NS, учитывая, что в столбце smoke_status есть только эти три
категории
prob.FS <- 864 / (864 + 1845 + 773)
prob.NS <- 1845 / (864 + 1845 + 773)
prob.S <- 773 / (864 + 1845 + 773)

## На всякий случай копируем данные
data2 <- data
# Замена «unknown» в smoke_status на другие 3 переменные в соответствии с
их весом
data2$rand <- runif(nrow(data2))
data2 <- data2%>%mutate(Probability = ifelse(rand <= prob.FS, "formerly
smoked", ifelse(rand <= (prob.FS+prob.NS), "never smoked", ifelse(rand <= 1,
"smokes", "Check"))))
data2 <- data2%>%mutate(smoking.status = ifelse(smoking_status == "Unknown",
Probability, smoking_status))
# Проверка уникальных значений нового столбца статуса курения и их
количество
table(data2$smoking.status)
# Удаляем ненужные столбцы, которые ни на что не влияют
health <- subset(data2, select = -c(rand,Probability,smoking_status))
colnames(health)[12] <- "smoking_status"
# просмотр первых 10 строк
head (health,10)
# «health» — это окончательный измененный набор данных, который будет
использоваться для раздела EDA ниже.

# Построим график
ggplot(health, aes(color=smoking_status, x=age, y=bmi)) + geom_point()

class(health$id)
health$id <- NULL
class(health$id)
head(health$id, 3)

# Subset Data into Yes and No (stroke)
Yes <- subset(health, stroke == '1')

```

```
No <- subset(health, stroke == '0')
```

```
## Построение гистограмм
```

```
## Инсульт статус пациентов график 1
```

```
strokecounts <- as.data.frame(table(health$stroke))
```

```
# Поскольку «Female» имеет большинство значений, заменим num на char  
strokecounts$Var1 <- ifelse(strokecounts$Var1 == 0, "No", 'Yes')
```

```
# Гистограмма пациентов с инсультом и без него
```

```
ggplot(strokecounts, aes(x = Var1, y = Freq, fill = Var1)) +  
  geom_bar(stat = "identity") + theme(legend.position="none") +  
  geom_text(aes(label = Freq, vjust = 0) +  
  labs(title="Инсульт Статус пациентов", x = "Инсульт", y = "Частота") +  
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Статус гипертонии пациентов график 2
```

```
hypercounts <- as.data.frame(table(health$hypertension, health$stroke))
```

```
# Заменяем num на char
```

```
hypercounts$Var1 <- ifelse(hypercounts$Var1 == 0, "No", 'Yes')  
hypercounts$Var2 <- ifelse(hypercounts$Var2 == 0, "No", 'Yes')
```

```
# Заменяем заголовки
```

```
colnames(hypercounts)[1] <- 'Hypertension'  
colnames(hypercounts)[2] <- 'Stroke'
```

```
# Bar Chart Столбчатая диаграмма гипертонии: No vs. Yes
```

```
ggplot(hypercounts, aes(x = Hypertension, y = Freq, fill = Stroke)) +  
  geom_bar(stat = "identity")+  
  geom_text(aes(label = Freq, vjust = 0) +  
  labs(title="Статус гипертонии пациентов", x = "Гипертония", y = "Частота") +  
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Сердечно-сосудистый статус пациентов график 3
```

```
heartcounts <- as.data.frame(table(health$heart_disease, health$stroke))
```

```
# заменяем num на char
```

```
heartcounts$Var1 <- ifelse(heartcounts$Var1 == 0, "No", 'Yes')  
heartcounts$Var2 <- ifelse(heartcounts$Var2 == 0, "No", 'Yes')
```

```

# меняем заголовки
colnames(heartcounts)[1] <- 'Heart_Disease'
colnames(heartcounts)[2] <- 'Stroke'

# Bar Chart Сердечно-сосудистая диаграмма пациентов: No vs. Yes
ggplot(heartcounts, aes(x = Heart_Disease, y = Freq, fill = Stroke)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Freq, vjust = 0) +
    labs(title="Сердечно-сосудистый статус пациентов", x = "Сердечное
заболевание", y = "Частота") +
    theme(plot.title = element_text(hjust = 0.5))

```

```

## График 4 - Гендер
# Create gender counts table
gendercounts <- as.data.frame(table(health$gender))

```

```

# Bar Chart of Gender
ggplot(gendercounts, aes(x = Var1, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity") + theme(legend.position="none") +
  geom_text(aes(label = Freq, vjust = 0) +
    labs(title="Пол пациентов", x = "Пол", y = "Частота") +
    theme(plot.title = element_text(hjust = 0.5))

```

```

## График 5 - Тип работы
# Create work type counts table
workcounts <- as.data.frame(table(health$work_type))

```

```

# Bar Chart Тип работы пациента
ggplot(workcounts, aes(x = Var1, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity") + theme(legend.position="none") +
  geom_text(aes(label = Freq, vjust = 0) +
    labs(title="Тип работы пациента", x = "Тип работы", y = "Частота") +
    theme(plot.title = element_text(hjust = 0.5))

```

```

## График 6 - Статус пациентов о браке
# Create ever married counts table
marriedcounts <- as.data.frame(table(health$ever_married))

```

```

# Bar Chart Статус пациентов о браке
ggplot(marriedcounts, aes(x = Var1, y = Freq, fill = Var1)) +

```

```

geom_bar(stat = "identity") + theme(legend.position="none") +
geom_text(aes(label = Freq), vjust = 0) +
labs(title="Гистограмма Статус пациентов о браке",x ="Был ли пациент в
браке", y = "Количество") +
theme(plot.title = element_text(hjust = 0.5))

```

```

## Тип места жительства
# Create residence type counts table
rescounts <- as.data.frame(table(health$Residence_type))

```

```

# Bar Chart Тип проживания
ggplot(rescounts, aes(x = Var1, y = Freq, fill = Var1)) +
geom_bar(stat = "identity") + theme(legend.position="none") +
geom_text(aes(label = Freq), vjust = 0) +
labs(title="Тип проживания пациентов",x ="Тип резиденции", y =
"Количество") +
theme(plot.title = element_text(hjust = 0.5))

```

```

## Статус курения
# Create smoking status counts table
smokecounts <- as.data.frame(table(health$smoking_status))

```

```

# Bar Chart курения пациентов
ggplot(smokecounts, aes(x = Var1, y = Freq, fill = Var1)) +
geom_bar(stat = "identity") + theme(legend.position="none") +
geom_text(aes(label = Freq), vjust = 0) +
labs(title="Статус курения пациентов",x ="Статус курения", y =
"Количество") +
theme(plot.title = element_text(hjust = 0.5))

```

```

## Построение гистограмм

```

```

# Гистограмма возраста с наложением нормального распределения
histage <- hist(health$age,xlim=c(0,100),
main="Гистограмма возраста с наложением нормального
распределения",
xlab="Age",ylab="Частота",las=1)
xfit <- seq(min(health$age),max(health$age))
yfit <- dnorm(xfit,mean=mean(health$age),sd=sd(health$age))
yfit <- yfit*diff(histage$mids[1:2])*length(health$age)
lines(xfit,yfit,col="purple",lwd=2)

```

```

summary(data2)

```



```

# Гистограмма среднего уровня глюкозы с наложением нормального
распределения
histglucose <- hist(health$avg_glucose_level,xlim=c(0,300),
  main="Гистограмма среднего уровня глюкозы с наложением
нормального распределения",
  xlab="Avg. Glucose",ylab="Частота",las=1)
xfit <- seq(min(health$avg_glucose_level),max(health$avg_glucose_level))
yfit <- dnorm(xfit,mean=mean(health$avg_glucose_level),sd=sd(health$avg_glucose_level))
yfit <- yfit*diff(histglucose$mids[1:2])*length(health$avg_glucose_level)
lines(xfit,yfit,col="red",lwd=2)

```

```

# Гистограмма ИМТ с наложением нормального распределения
histbmi <- hist(health$bmi,xlim=c(0,100),
  main="Гистограмма ИМТ с наложением нормального
распределения",
  xlab="Индекс массы тела",ylab="Частота", las=1)
xfit <- seq(min(health$bmi),max(health$bmi))
yfit <- dnorm(xfit,mean=mean(health$bmi),sd=sd(health$bmi))
yfit <- yfit*diff(histbmi$mids[1:2])*length(health$bmi)
lines(xfit,yfit,col="red",lwd=2)

```

```

# График среднего уровня глюкозы у пациентов с инсультом и без него
boxplot(Yes$avg_glucose_level,No$avg_glucose_level,
  main="График среднего уровня глюкозы у пациентов перенесших и не
перенесших инсульт",
  ylab="Средний уровень глюкозы",las=1,names=c("Перенесли
инсульт","Не перенесли"))

```

```

summary(Yes$avg_glucose_level)
summary(No$avg_glucose_level)

```

```

# График индекса массы тела у пациентов с инсультом и без него
boxplot(Yes$bmi,No$bmi,main="График индекса массы тела у пациентов
перенесших и не перенесших инсульт",
  ylab="Индекс массы тела",las=1,names=c("Перенесли инсульт","Не
перенесли"))

```

```

summary(Yes$bmi)
summary(No$bmi)

library(corrgram)

# Создаем коррелограмму для числовых переменных
corrgram(health, order=NULL, panel=panel.shade, text.panel=panel.txt,
         diag.panel=panel.minmax, main="Коррелограмма")

# Просмотр значений корреляции числовых переменных с двумя десятичными
точками
round(cor(subset(health,                               select=c(age,hypertension,
heart_disease,avg_glucose_level, bmi, stroke))),2)

library(corr)

# Преобразование категориальных переменных в числовые переменные
# Новый набор данных называется onehot
dmy <- dummyVars(" ~ .", data = health)
onehot <- data.frame(predict(dmy, newdata = health))
# Просмотр заголовков нового набора данных
names(onehot)

# Таблица корреляции
cor_onehot <- correlate(onehot)

# Извлечение корреляции, связанной с инсультом
cor_onehot%>% focus(stroke)
# Построим корреляцию между инсультом и всеми остальными
cor_onehot %>%
  focus(stroke) %>%
  mutate(rowname = reorder(term, stroke)) %>%
  ggplot(aes(term, stroke)) +
  geom_col() + coord_flip() +
  theme_bw()

## Данные, воспроизведенные с использованием исходных данных
health$gender <- as.factor(health$gender)
health$stroke <- as.factor(health$stroke)
table(health$gender)

```

```

class(health$stroke)

set.seed(100)
health$AgeGroup <-NULL
sample = sample.split(health$stroke, SplitRatio = 0.7)
train = subset(health, sample==TRUE)
test = subset(health, sample==FALSE)
health$gender <- as.factor(health$gender)
table(health$gender)

#Создаем фиктивные переменные
library(fastDummies)
# дублируем данные
healthdummy <- health
# Создание фиктивных столбцов
healthdummy
dummy_cols(healthdummy,select_columns=c("gender","ever_married",
"work_type","Residence_type","smoking_status"),
          remove_selected_columns = TRUE)

# проверяем данные
summary(healthdummy)

str(healthdummy)

set.seed(123)
trainIndex <- sample(x=nrow(healthdummy),size=nrow(healthdummy)*0.7)
healthdummytrain <- healthdummy[trainIndex,]
healthdummytest <- healthdummy[-trainIndex,]

# замена factor на num
healthdummy$stroke <- ifelse(healthdummy$stroke == '1', 1, 0)
table(healthdummy$stroke)

## Модель 1: обобщенная линейная модель — логистическая регрессия —
DummyVar

glm_fit <- glm(stroke~., data=healthdummytrain, family = binomial)

```

```

summary(glm_fit)

stepdummy <- stepAIC(glm_fit)

summary(stepdummy)

# Делаем предикшн
probabilities <- stepdummy %>% predict(healthdummytest, type = "response")
predicted_classes <- ifelse(probabilities > 0.5, 1, 0)
# Точность модели
print(paste("Model Accuracy : ", mean(predicted_classes ==
healthdummytest$stroke)))

# Anova Table
anova(stepdummy)

```

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

Казахский национальный исследовательский
технический университет имени К.И.Сатпаева

Отзыв научного руководителя

Дипломная работа

Сулиев Жалил Ниязович

5В070300 – Информационные системы

Тема: Разработка компонента «Интерпретация данных» информационной системы в аналитике данных

Дипломная работа представляет собой выпускную квалификационную работу по специальности 5В070300 – «Информационные системы». Пояснительная записка состоит из введения, 3 глав, заключения, списка использованных источников и приложения.

Автор дипломной работы поставленные задачи полностью выполнил и показал владение современными технологиями в предметной области.

Дипломная работа выполнена на достаточном профессиональном уровне и содержит все необходимые сведения для такого рода работ. К замечаниям следует отнести незначительные стилистические ошибки.

Считаю, что дипломная соответствует требованиям, предъявляемым к выпускным квалификационным работам по специальности 5В070300 – «Информационные системы». Автор работы Сулиев Жалил Ниязович заслуживает присвоения академической степени бакалавра.

Научный руководитель
Ассоц.проф., канд.техн.наук



Жумагалиев Б.И.

« 16 » мая 2022 г.

РЕЦЕНЗИЯ

на дипломную работу студента 4 курса «Казахского национального
исследовательского технического университета им. К. И. Сатпаева»
специальности 5В070300 «Информационные системы»

Сулиева Жалил Ниязовича

на тему: «Разработка компонента «Интерпретация данных» информационной
системы в аналитике данных»

Интерпретация данных всегда являлась актуальным направлением в изучении данных в разных областях науки, потому что позволяет получить решения для проблем, связанных с классифицированием и прогнозированием, там, где теоремы классической математики не могут нам помочь. Основным направлением дипломной работы является интерпретация статистических данных, связанных с изучением случаев инсультов у пациентов с сердечно-сосудистыми заболеваниями. Поэтому тема дипломной работы весьма актуальна, т. к. анализ и интерпретация медицинских данных в будущем может помочь выявлять проблемы со здоровьем у пациентов с заболеваниями сердца и сосудов.

Дипломная работа состоит из введения, трёх разделов, заключения, списка используемых источников и приложения.

Во введении автор сформулировал цель и определил задачи, которые предстоит решить в работе, а также объект и предметную область исследования.

В первой главе автор рассматривает методы сбора информации, системный и линейный регрессионный анализ для решения поставленных задач.

Вторая глава посвящена разработке программы для построения графиков, позволяющих произвести анализ и интерпретацию статистических данных.

В третьей главе описываются преимущества и недостатки R – языка программирования для анализа и разработки данных.

В заключении приведены выводы о проделанной работе.

Замечание по проекту:

В тексте дипломного проекта есть грамматические ошибки, но они не ухудшают качество работы. Интерпретация полученных результатов не полная, возможно оттого, что автор не является специалистом в медицинской сфере.

В целом работа соответствует требованиям, предъявляемым к выпускным квалификационным работам, и рекомендуется к защите, а ее автор Сулиев Ж. Н. заслуживает присвоения академической степени бакалавра по специальности 5В070300 – «Информационные системы».

Рецензент:


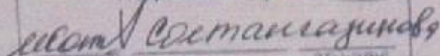
Ph.D. в информационных системах,

старший преподаватель

«АУЭС им. Гумарбека Даукеспау»

Колтабанов Растиым

Подпись заверяю



16 05 2022 г.

Бимурат Ж.

Протокол анализа Отчета подобия заведующего кафедрой

Заведующий кафедрой заявляет, что ознакомился (-ась) с Полным отчетом

подобия, который был сгенерирован Системой выявления и предотвращения плагиата в отношении работы:

Автор: Судиев Жалил Ниязович

Название: Дипломная работа Судиев Жалил

Координатор: Жумагалиев Биржан Изимович

Коэффициент подобия 1: 6.24

Коэффициент подобия 2: 1.45

Тревога: _____

После анализа Отчета подобия заведующий кафедрой констатирует следующее:

обнаруженные в работе заимствования являются добросовестными и не обладают признаками плагиата. В связи с чем, работа признается самостоятельной и допускается к защите;

обнаруженные в работе заимствования не обладают признаками плагиата, но их чрезмерное количество вызывает сомнения в отношении ценности работы, по существу, и отсутствием самостоятельности ее автора. В связи с чем, работа должна быть вновь отредактирована с целью ограничения заимствований;

обнаруженные в работе заимствования являются недобросовестными и обладают признаками плагиата, или в ней содержатся преднамеренные искажения текста, указывающие на попытки сокрытия недобросовестных заимствований. В связи с чем, работа не допускается к защите.

Обоснование:

Обнаруженные в работе заимствования являются незначительными в пределах нормы. В связи с чем, признаю работу самостоятельной и допускаю к защите.

«16» май 2022 г.

Дата/м./г.



Ф.И.О., подпись зав. кафедрой

Утверждено решением Правления от 01.04.2022 г. протокол №5

Протокол анализа Отчета подобия Научным руководителем

Заявляю, что я ознакомился (-ась) с Полным отчетом подобия, который был сгенерирован Системой выявления и предотвращения плагиата в отношении работы:

Автор: Сулиев Жалил Ниязович

Название: Дипломная работа Сулиев Жалил

Координатор: Жумагалиев Биржан Изимович

Коэффициент подобия 1: 6.24

Коэффициент подобия 2: 1.45

Тревога: _____

После анализа Отчета подобия констатирую следующее:

обнаруженные в работе заимствования являются добросовестными и не обладают признаками плагиата. В связи с чем, признаю работу самостоятельной и допускаю ее к защите;

обнаруженные в работе заимствования не обладают признаками плагиата, но их чрезмерное количество вызывает сомнения в отношении ценности работы, по существу, и отсутствием самостоятельности ее автора. В связи с чем, работа должна быть вновь отредактирована с целью ограничения заимствований;

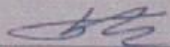
обнаруженные в работе заимствования являются недобросовестными и обладают признаками плагиата, или в ней содержатся преднамеренные искажения текста, указывающие на попытки сокрытия недобросовестных заимствований. В связи с чем, не допускаю работу к защите.

Обоснование:

Обнаруженные в работе заимствования являются незначительными в пределах нормы. В связи с чем, признаю работу самостоятельной и допускаю к защите

«16» май 2022 г.

Дата/м./г.


Подпись Научного руководителя

Утверждено решением Правления от 01.04.2022 г. протокол №5