МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Воронежский государственный технический университет»

УТВЕРЖДАЮ
Декан факультета информационных технологий и компьютерной безопасности
_____ Гусев П.Ю.
«21» декабря 2021 г.

РАБОЧАЯ ПРОГРАММА

дисциплины

«Язык R и базовая статистика»

Направление подготовки 09.04.01 Информатика и вычислительная техника

Профиль Искусственный интеллект

Квалификация выпускника магистр

Нормативный период обучения 2 года / 2 года и 5 м.

Форма обучения очная / заочная

Год начала подготовки 2022

Автор программы _____ /Гусев П.Ю./

Заведующий кафедрой Компьютерных интеллектуальных технологий проектирования _____ /М.И. Чижов/

Руководитель ОПОП _____ /М.И. Чижов/

Воронеж 2021

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

1.1. Цели дисциплины «Язык R и базовая статистика» изучение основ программирования на интерпретируемом языке R, а также его применения для статистической обработки данных. Повышение интереса к разработке программного обеспечения, формирование творческого подхода к программированию. Приобретение навыков, позволяющих будущим специалистам вести успешную разработку специализированного программного обеспечения в тех областях и сферах деятельности, в которых они будут трудиться.

Изучение дисциплины должно способствовать формированию у студентов основ научного мышления, в том числе: владение основными методами, способами и инструментами создания программного обеспечения, использования для решения практических задач

1.2. Задачи освоения дисциплины

- ознакомление с тенденцией развития программного обеспечения;
- изучения основных структур языка R;
- изучение операций над структурами в языке R;
- изучение лексики языка R;
- обучение применению языка R для статистической обработки данных;
- обучение применению языка R для визуализации обрабатываемых данных.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Язык R и базовая статистика» относится к дисциплинам части, формируемой участниками образовательных отношений (дисциплина по выбору) блока Б1.

Учебная дисциплина относится к дисциплинам по выбору основной профессиональной образовательной программы (далее – ОПОП) направления подготовки 09.04.01 Информатика и вычислительная техника направленности (профилю) Искусственный интеллект. В качестве входных требований выступают сформированные ранее компетенции обучающихся в области программирования, теории вероятностей и математической статистики. Результаты освоения учебной дисциплины могут быть использованы при прохождении производственной практики (НИР и преддипломной практики), а также при выполнении магистерской диссертации.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Процесс изучения дисциплины «Язык R и базовая статистика» направлен на формирование следующих компетенций:

УК-1 - Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

ОПК-9 - Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта

ОПК-10 - Способен адаптировать и применять на практике классические и новые научные принципы и методы исследований для решения задач в области создания и применения технологий и систем искусственного интеллекта и методы исследований

Компетенция	Результаты обучения, характеризующие сформированность компетенции
УК-1	Знать методы описания исходных данных для статистического анализа
	Уметь применять методы базовой статистики
	Владеть навыками формализации задач для статистического анализа
ОПК-9	Знать базовые конструкции языка R
	Уметь реализовывать алгоритмы обработки данных на языке R
	Владеть навыками выявления требований заказчика к результатам решения задач искусственного интеллекта
ОПК-10	Знать сквозные цифровые технологии искусственного интеллекта; современный опыт применения систем искусственного интеллекта
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта
	Владеть навыками определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры анализа и обработки данных

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины «Язык R и базовая статистика» составляет 3 з.е.

Распределение трудоемкости дисциплины по видам занятий

очная форма обучения

Виды учебной работы	Всего часов	Семестры
		3
Аудиторные занятия (всего)	36	36
В том числе:		
Лекции	16	16
Лабораторные работы (ЛР)	20	20
Самостоятельная работа	72	72
Виды промежуточной аттестации - зачет	+	+
Общая трудоемкость: час	108	108

зач.ед.	3	3

заочная форма обучения

Виды учебной работы	Всего часов	Семестры
		3
Аудиторные занятия (всего)	16	16
В том числе:		
Лекции	8	8
Лабораторные работы (ЛР)	8	8
Самостоятельная работа	88	88
Часы на контроль	4	4
Виды промежуточной аттестации - зачет	+	+
Общая трудоемкость: час	108	108
зач.ед.	3	3

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1 Содержание разделов дисциплины и распределение трудоемкости по видам занятий

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Знакомство с R. Визуализация данных	Идеология R. Структуры данных (векторы, матрицы, массивы данных, таблицы данных, факторы, списки). Загрузка данных. Пакеты. Числовые и текстовые функции. Циклы и условия. Пользовательские функции.	2	-	12	20
2	Визуализация данных	Столбчатые диаграммы. Гистограммы. Диаграммы. плотности распределений. Диаграммы размахов. Диаграммы рассеяния	2	4	12	20
3	Разведочный анализ данных	Оценки центрального положения и вариабельности. Исследование распределений данных. Корреляция. Тест Стьюдента. Непараметрические тесты межгрупповых различий. Визуализация групповых различий	2	4	12	18
4	Проверка статистических гипотез.	A/B-тестирование. Проверка статистических гипотез. Статистическая значимость и p-значение. Проверка на основе t-статистики. Дисперсионный анализ. Проверка на основе статистики хи-квадрат	4	4	12	18
5	Регрессия и классификация в R	Простая линейная регрессия. Множественная линейная регрессия. Предсказания на основе регрессии. Оценка качества предсказания. Перекрестная проверка. Нелинейная регрессия. Наивный байесовский алгоритм. Логистическая. регрессия. Предсказание. значений в логистической. регрессии. Оценка качества предсказания	4	4	12	16

6	Статистическое машинное обучение.	Алгоритм k ближайших соседей. Древовидные алгоритмы (дерево решений, случайный лес). Алгоритм бустинга. Переобучение. Подбор гиперпараметров алгоритм	2	4	12	16
		Итого	**16**	**20**	**72**	**108**

заочная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Знакомство с R. Визуализация данных	Идеология R. Структуры данных (векторы, матрицы, массивы данных, таблицы данных, факторы, списки). Загрузка данных. Пакеты. Числовые и текстовые функции. Циклы и условия. Пользовательские функции.	1	-	14	18
2	Визуализация данных	Столбчатые диаграммы. Гистограммы. Диаграммы. плотности распределений. Диаграммы размахов. Диаграммы рассеяния	1	1	14	18
3	Разведочный анализ данных	Оценки центрального положения и вариабельности. Исследование распределений данных. Корреляция. Тест Стьюдента. Непараметрические тесты межгрупповых различий. Визуализация групповых различий	1	1	14	18
4	Проверка статистических гипотез.	A/B-тестирование. Проверка статистических гипотез. Статистическая значимость и p-значение. Проверка на основе t-статистики. Дисперсионный анализ. Проверка на основе статистики хи-квадрат	2	2	14	18
5	Регрессия и классификация в R	Простая линейная регрессия. Множественная линейная регрессия. Предсказания на основе регрессии. Оценка качества предсказания. Перекрестная проверка. Нелинейная регрессия. Наивный байесовский алгоритм. Логистическая. регрессия. Предсказание. значений в логистической. регрессии. Оценка качества предсказания	2	2	16	16
6	Статистическое машинное обучение.	Алгоритм k ближайших соседей. Древовидные алгоритмы (дерево решений, случайный лес). Алгоритм бустинга. Переобучение. Подбор гиперпараметров алгоритм	1	2	16	16
		Итого	**8**	**8**	**88**	**104**

5.2 Перечень лабораторных работ

Лабораторная работа № 1. Установка RStudio. Работа с основными типами данных. Создание датафреймов и загрузка внешних данных. Загрузка пакетов и написание функций. Построение графиков с использованием пакета ggplot2

Лабораторная работа № 2. Расчет описательных статистик и визуализация распределений

Лабораторная работа № 3. Тестирование гипотез

Лабораторная работа № 4. Предсказание на основе уравнения линейной регрессии

Лабораторная работа № 5. Предсказание значений в логистической регрессии. Использование алгоритма XGBoost для предсказания значений

6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины не предусматривает выполнение курсового проекта (работы) или контрольной работы.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«не аттестован».

Компе-тенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Аттестован	Не аттестован
УК-1	Знать методы описания исходных данных для статистического анализа	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь применять методы базовой статистики	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками формализации задач для статистического анализа	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ОПК-9	Знать базовые конструкции языка R	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь реализовывать алгоритмы обработки данных на языке R	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками выявления требований заказчика к результатам решения задач искусственного интеллекта	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ОПК-10	Знать сквозные цифровые технологии искусственного интеллекта; современный опыт применения систем	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

	искусственного интеллекта			
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры анализа и обработки данных	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 3 семестре для очной формы обучения, 3 семестре для заочной формы обучения по двухбалльной системе:

«зачтено»

«не зачтено»

Компе-тенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Зачтено	Не зачтено
УК-1	Знать методы описания исходных данных для статистического анализа	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь применять методы базовой статистики	Решение стандартных практических задач	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками формализации задач для статистического анализа	Решение прикладных задач в конкретной предметной области	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены
ОПК-9	Знать базовые конструкции языка R	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь реализовывать алгоритмы обработки данных на языке R	Решение стандартных практических задач	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками выявления требований заказчика к результатам решения задач	Решение прикладных задач в конкретной предметной области	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены

	искусственного интеллекта			
ОПК-10	Знать сквозные цифровые технологии искусственного интеллекта; современный опыт применения систем искусственного интеллекта	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта	Решение стандартных практических задач	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры анализа и обработки данных	Решение прикладных задач в конкретной предметной области	Продемонстрирова н верный ход решения в большинстве задач	Задачи не решены

7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

7.2.1 Примерный перечень заданий для подготовки к тестированию

1. Какая из следующих функций используется для просмотра набора данных в формате электронной таблицы?

A) disp()

+B) View()

C) seq()

D) все ответы верны

2. Какая из следующих команд разделит окно графика на окна 4 X 3 и графики входят в столбец окна.

A) par(split=c(4,3))

B) par(mfcol=c(4,3))

C) par(mfrow=c(4,3))

D) par(col=c(4,3))

3. Какая из следующих команд поможет нам переименовать второй столбец в фрейме данных с именем «table» с ’alpha’ на ’beta’?

A) colnames(table)[2]=’beta’

B) colnames(table)[which(colnames==’alpha’)]=’beta’

C) setnames(table,’alpha’,’beta’)

D) Все ответы верны

4. Какая из следующих команд удалит объект R / переменную с именем «santa» из рабочей области?

A) remove(santa)

B) rm(santa)

+C) a) и b)

D)Нет верного ответа

5. «dplyr» — один из самых популярных пакетов, используемых в R для управления данными, и он содержит 5 основных функций для обработки данных. Что из следующего не является одной из основных функций пакета dplyr?

A) select()

B) filter()

C) arrange()

+D) summary()

6. Каковы будут выходные данные следующих команд?

A<-paste(“alpha”,”beta”,”gamma”,sep=” ”)

B←paste(“phi”,”theta”,”zeta”,sep=””)

parts←strsplit(c(A,B),split=” ”)

A) alpha

+B) beta

C) gamma

D) phi

E) theta

F) zeta

7. Что будет выведено следующей командой?

grepl(“neeraj”,c(“dheeraj”,”Neeraj”,”neeraj”,”is”,”NEERAJ”))

A) [FALSE TRUE TRUE FALSE TRUE]

B) [FALSE TRUE TRUE FALSE FALSE]

+C) [FALSE FALSE TRUE FALSE FALSE]

D) Все ответы не верны

8. Что будет выведено следующей командой?

A<-c("I can use because thrice in a sentence because because is a special word.")

+A) gsub(“because”,”since”,A)

B) sub(“because”,”since”,A

C) regexec(“because”,”since”,A)

D) Все ответы не верны

9. Какая из следующих команд поможет нам удалить повторяющиеся строки на основе обоих столбцов?

A) df[!duplicated(df),]

B) unique(df)

C) dplyr::distinct(df)

+D) Все ответы верны

10. Группировка является важным действием в Data Analytics и помогает нам обнаружить некоторые интересные тенденции, которые могут быть не видны в необработанных данных. Предположим, у вас есть набор данных, созданный с помощью следующих строк кода.

table<-data.table(foo=c("A","B","A","A","B","A"),bar=1:6)

Какая из следующих команд поможет нам вычислить среднее значение bar, сгруппированного по переменной foo?

A) aggregate(bar~foo,table,mean)

B) table::df[,mean(bar),by=foo]

C) dplyr::table%>%group_by(foo)%>%summarize(mean=mean(bar))

+D) Все ответы верны

7.2.2 Примерный перечень заданий для решения стандартных задач

1. Ниже приведен набор данных, необходимо построить гистограмму для переменной «Значение».

Parameter	State	Value	Dependents
Alpha	Active	50	2
Beta	Active	45	5
Beta	Passive	25	0
Alpha	Passive	21	0
Alpha	Passive	26	1
Beta	Active	30	2
Beta	Passive	18	0

Какая из следующих команд выполнит эту задачу?

A) hist(dataframed$Value)

B) ggplot2::qplot(dataframed$Value,geom=”Histogram”)

C)ggplot2::ggplot(data=dataframed,aes(dataframe$Value))+geom_histogram()

D) Все ответы верны

2.

Parameter	State	Value	Usage
Alpha	Active	50	0
Beta	Active	45	1
Beta	Passive	25	0
Alpha	Passive	21	0
Alpha	Passive	26	1
Beta	Active	30	1
Beta	Passive	18	0

Некоторые алгоритмы, такие как XGBOOST, работают только с числовыми данными. В этом случае категориальные переменные, присутствующие в наборе данных, сначала преобразуются в переменные DUMMY, которые представляют наличие или отсутствие уровня категориальной переменной в наборе данных. Например, после создания

фиктивной переменной для функции «Параметр» набор данных выглядит так, как показано ниже.

Parameter_Alpha	Parameter_Beta	State	Value	Usage
1	0	Active	50	0
0	1	Active	45	1
0	1	Passive	25	0
1	0	Passive	21	0
1	0	Passive	26	1
0	1	Active	30	1
0	1	Passive	18	0

Какая из следующих команд позволяет достичь этого?

A) dummies:: dummy.data.frame(dataframe,names=c('Parameter'))

B) dataframe$Parameter_Alpha=0
dataframe$Gende_Beta=0
dataframe$Parameter_Alpha[which(dataframe$Parameter=='Alpha')]=1
dataframe$Parameter_Beta[which(dataframe$Parameter=='Alpha')]=0
dataframe$Parameter_Alpha[which(dataframe$Parameter=='Beta']=0
dataframe$Parameter_Beta[which(dataframe$Parameter=='Beta']=1

C) contrasts(dataframe$Parameter)

+D) A) и B)

3.

Column1	Column2	Column3	Column4	Column5	Column6	
Name1	Alpha	12	24	54	0	Alpha
Name2	Beta	16	32	51	1	Beta
Name3	Alpha	52	104	32	0	Gamma
Name4	Beta	36	72	84	1	Delta
Name5	Beta	45	90	32	0	Phi
Name6	Alpha	12	24	12	0	Zeta
Name7	Beta	32	64	64	1	Sigma
Name8	Alpha	42	84	54	0	Mu
Name9	Alpha	56	112	31	1	Eta

Необходимо рассчитать корреляцию между «Столбцом2» и «Столбцом3» в наборе данных. Какой код из приведенных ниже достигнет цели?

A) corr(dataframe$column2,dataframe$column3)

B)
(cov(dataframe$column2,dataframe$column3))/(var(dataframe$column2)*sd(dataframe$column3))

C)
(sum(dataframe$Column2*dataframe$Column3)-
(sum(dataframe$Column)*sum(dataframe$Column3)/nrow(dataframe)))/(sqrt((sum(dataframe$Column2*dataframe$Column2)-
(sum(dataframe$Column2)^3)/nrow(dataframe))*

(sum(dataframe$Column3*dataframe$Column3)-
(sum(dataframe$Column3)^2)/nrow(dataframe))))

D) Ни один ответ не верный

4.

Parameter	State	Value	Dependents
Alpha	Active	50	2
Beta	Active	45	5
Beta	Passive	25	0
Alpha	Passive	21	0
Alpha	Passive	26	1
Beta	Active	30	2

В переменной с именем «dataframe» ассоциирован набор данных, где первая строка представляет имя столбца. Какой набор инструкций будет выбирать только те строки, для которых параметр Alpha?

A) subset(dataframe, Parameter=’Alpha’)

B) subset(dataframe, Parameter==’Alpha’)

C) filter(dataframe,Parameter==’Alpha’)

+D) Только b иии c

E) Все ответы верны

5. Приведенный ниже набор данных хранится в переменной с именем data.

A	B
1	Right
2	Wrong
3	Wrong
4	Right
5	Right
6	Wrong
7	Wrong
8	Right

Предположим, что B — категориальная переменная, и необходимо нарисовать блок-диаграмму для каждого уровня категориального уровня. Какая из приведенных ниже команд выполнит это?

A) boxplot(A,B,data=data)

+B) boxplot(A~B,data=data)

C) boxplot(A|B,data=data)

D) Все ответы не верны

6. Набор данных «df» содержит следующие данные:

Dates
2017-02-28
2017-02-27
2017-02-26
2017-02-25
2017-02-24
2017-02-23

2017-02-22

2017-02-21

После прочтения приведенных выше данных необходимо получить следующий вывод данных

Dates

28 Tuesday Feb 17

27 Monday Feb 17

26 Sunday Feb 17

25 Saturday Feb 17

24 Friday Feb 17

23 Thursday Feb 17

22 Wednesday Feb 17

21 Tuesday Feb 17

Какая из следующих команд даст желаемый результат?

A) format(df,"%d %A %b %y")

B) format(df,"%D %A %b %y")

C) format(df,"%D %a %B %Y")

+D) Все ответы не верны

7. Во время выбора функций с использованием следующего набора данных (именованной таблицы) «Столбец1» и «Столбец2» оказались несущественными. Следовательно, мы не хотели бы использовать эти два критерия в нашей прогностической модели.

Column1	**Column2**	**Column3**	**Column4**	**Column5**	**Column6**	
Name1	Alpha	12	24	54	0	Alpha
Name2	Beta	16	32	51	1	Beta
Name3	Alpha	52	104	32	0	Gamma
Name4	Beta	36	72	84	1	Delta
Name5	Beta	45	90	32	0	Phi
Name6	Alpha	12	24	12	0	Zeta
Name7	Beta	32	64	64	1	Sigma
Name8	Alpha	42	84	54	0	Mu
Name9	Alpha	56	112	31	1	Eta

Какая из следующих команд выберет все строки от столбца 3 до столбца 6 для приведенного ниже фрейма данных с именем таблицы?

A) dplyr::select(table,Column3:Column6)

B) table[,3:6]

C) subset(table,select=c('Column3','Column4','Column5','Column6'))

+D) Все ответы верны

8.

	Column1	**Column2**	**Column3**	**Column4**	**Column5**	**Column6**
Name1	Alpha	12	24	54	0	Alpha
Name2	Beta	16	32	51	1	Beta
Name3	Alpha	52	104	32	0	Gamma
Name4	Beta	36	72	84	1	Delta

Name5	Beta	45	90	32	0	Phi
Name6	Alpha	12	24	12	0	Zeta
Name7	Beta	32	64	64	1	Sigma
Name8	Alpha	42	84	54	0	Mu
Name9	Alpha	56	112	31	1	Eta

Какая из следующих команд выберет строки, имеющие значения «Альфа» в «Столбце 1» и значение меньше 50 в «Столбце 4»? Фрейм данных хранится в переменной с именем table.

A) dplyr::filter(table,Column1==’Alpha’, Column4<50)

B) dplyr::filter(table,Column1==’Alpha’ & Column4<50)

+C) a) и b)

D) Все ответы не верны

9.

	Column1	**Column2**	**Column3**	**Column4**	**Column5**	**Column6**
Name1	Alpha	12	24	54	0	Alpha
Name2	Beta	16	32	51	1	Beta
Name3	Alpha	52	104	32	0	Gamma
Name4	Beta	36	72	84	1	Delta
Name5	Beta	45	90	32	0	Phi
Name6	Alpha	12	24	12	0	Zeta
Name7	Beta	32	64	64	1	Sigma
Name8	Alpha	42	84	54	0	Mu
Name9	Alpha	56	112	31	1	Eta

Какая из следующих инструкций будет сортировать фрейм данных на основе «Столбец2» в порядке возрастания и «Столбец3» в порядке убывания?

A) dplyr::arrange(table,desc(Column3),Column2)

B) table[order(-Column3,Column2),]

+C) a) и b)

D) Все ответы не верны

10. Какая из следующих команд преобразует следующий набор данных с именем maverick в показанный внизу?

Входящий Dataframe – “maverick”

Grade	Male	Female
A	10	15
B	20	15
A	30	35

Обработанный dataframe

Grade	Sex	Count
A	Male	10
A	Female	15
B	Male	30
B	Female	15
A	Male	30

A Female 35

+A) tidyr::Gather(maverick, Sex,Count,-Grade)

B) tidyr::spread(maverick, Sex,Count,-Grade

C) tidyr::collect(maverick, Sex,Count,-Grade)

D) Все ответы не верны

7.2.3 Примерный перечень заданий для решения прикладных задач

1. Дана следующая функция

```
f <- function(x) {
        g <- function(y) {
                y + z
        }
        z <- 4
        x + g(x)
}
```

Если мы выполним следующие команды (написанные ниже), что будет на выходе?

```
z <- 10
f(4)
```

+A) 12

B) 7

C) 4

D) 16

2. В наборе данных ирисов есть разные виды цветов, такие как Setosa, Versicolor и Virginica, с их длиной чашелистика. Необходимо понять распределение длины чашелистиков у всех видов цветов. Один из способов сделать это — визуализировать это отношение с помощью графика, показанного ниже.

Какую функцию можно использовать для построения графика, показанного выше?

A) xyplot()

+B) stripplot()

C) barchart()

D) bwplot()

3.
Alpha	125.5	0
Beta	235.6	1
Beta	212.03	0
Beta	211.30	0
Alpha	265.46	1

Dataframe.csv

Какая из следующих команд правильно прочитает приведенный выше CSV-файл с 5 строками в наборе данных?

A) csv(‘Dataframe.csv’)

B) csv(‘Dataframe.csv’,header=TRUE)

C) dataframe(‘Dataframe.csv’)

+D) csv2(‘Dataframe.csv’,header=FALSE,sep=’,’)

4. Формат файла Excel является одним из наиболее распространенных форматов, используемых для хранения наборов данных. Ниже представлен файл excel, в котором были введены данные на третьем листе.

Alpha 125.5 0
Beta 235.6 1
Beta 212.03 0
Beta 211.30 0
Alpha 265.46 1
Dataframe.xlsx

4) Какой из следующих кодов будет считывать вышеуказанные данные на третьем листе в кадр данных в R?

A) Openxlsx::read.xlsx(“Dataframe.xlsx”,sheet=3,colNames=FALSE)

B) Xlsx::read.xlsx(“Dataframe.xlsx”,sheetIndex=3,header=FALSE)

C)XLConnect::readWorksheetFromFile(“Dataframe.xlsx”,sheet=3,header=FALSE)

+D)Все ответы верны

5. A 10 Sam
B 20 Peter
C 30 Harry
D ! ?
E 50 Mark
Dataframe.csv

5) Отсутствующие значения в этом CSV-файле представлены восклицательным знаком («!») и вопросительным знаком («?»). Какой из приведенных ниже кодов правильно прочитает приведенный выше CSV-файл в R?

A) csv(‘Dataframe.csv’)

B) csv(‘Dataframe.csv’,header=FALSE, sep=’,’,na.strings=c(‘?’))

+C) csv2(‘Dataframe.csv’,header=FALSE,sep=’,’,na.strings=c(‘?’,’!’))

D) dataframe(‘Dataframe.csv’)

6. Column 1 Column 2 Column 3
Row 1 15.5 14.12 69.5
Row 2 18.6 56.23 52.4
Row 3 21.4 47.02 63.21
Row 4 36.1 56.63 36.12
Dataframe.csv

6) Приведенный выше CSV-файл имеет название строк, а также название столбцов. Какой из следующих кодов правильно прочитает приведенный выше CSV-файл в R?

A) delim(‘Train.csv’,header=T,sep=’,’,row.names=TRUE)

B) csv2(‘Train.csv’,header=TRUE, row.names=TRUE)

C) dataframe(‘Train.csv’,header=TRUE,sep=’,’)

+D) csv(‘Train.csv’,,header=TRUE,sep=’,’)

7. Column 1 Column 2 Column 3
Row 1 15.5 14.12 69.5

Row 2 18.6 56.23 52.4
Row 3 21.4 47.02 63.21
Row 4 36.1 56.63 36.12

Dataframe.csv

Какой из следующих кодов будет читать только первые две строки CSV-файла?

+A) csv('Dataframe.csv',header=TRUE,row.names=1,sep=',',nrows=2

B) csv2('Dataframe.csv',row.names=1,nrows=2)

C) delim2('Dataframe.csv',header=T,row.names=1,sep=',',nrows=2)

D) dataframe('Dataframe.csv',header=TRUE,row.names=1,sep=',',skip.last=2)

8.

Dataframe1

Feature1	**Feature2**	**Feature3**	**Feature4**
A	1000	25.5	10
B	2000	35.5	34
C	3000	45.5	78
D	4000	55.5	3

Dataframe2

Feature1	**Feature2**	**Feature3**
E	5000	65.5
F	6000	75.5
G	7000	85.5
H	8000	95.5

Здесь хранятся два кадра данных Dataframe1 и Dataframe2, показанные выше. Какой из следующих кодов выдаст вывод, показанный ниже?

Feature1	**Feature2**	**Feature3**
A	1000	25.5
B	2000	35.5
C	3000	45.5
D	4000	55.5
E	5000	65.5
F	6000	75.5
G	7000	85.5
H	8000	95.5

A) merge(dataframe[,1:3],dataframe2)

B) merge(dataframe1,dataframe2)[,1:3]

C) merge(dataframe1,dataframe2,all=TRUE)

+D) Both 1 and 2

E) All of the above

9.

	V1	V2
1	121.5	461
2	516	1351
3	451	6918
4	613	112
5	112.36	230
6	25.23	1456
7	12	457

Dataframe

Набор данных был прочитан в R и сохранен в переменной «dataframe». Какой из приведенных ниже кодов выдаст сводку (среднее, моду, медиану) всего набора данных в одной строке кода?

A) summary(dataframe)

B) stats(dataframe)

C) summarize(dataframe)

D) summarise(dataframe)

+E) Нет верного ответа

10. Набор данных был прочитан в R и сохранен в переменной «dataframe». Отсутствующие значения были прочитаны как NA.

A	10	Sam
B	NA	Peter
C	30	Harry
D	40	NA
E	50	Mark

Dataframe

Какой из следующих кодов не даст количество пропущенных значений в каждом столбце?

A) colSums(is.na(dataframe))

B) apply(is.na(dataframe),2,sum)

C) sapply(dataframe,function(x) sum(is.na(x))

+D) table(is.na(dataframe))

7.2.4 Примерный перечень вопросов для подготовки к зачету

1. Идеология R. Структуры данных (векторы, матрицы, массивы данных, таблицы данных, факторы, списки).

2. Загрузка данных. Пакеты. Числовые и текстовые функции. Циклы и условия.

3. Пользовательские функции

4. Визуализация данных. Диаграммы. Гистограммы.

5. Визуализация данных Диаграммы плотности распределений. Диаграммы размахов. Диаграммы рассеяния

6. Оценки центрального положения и вариабельности.

7. Исследование распределений данных.

8. Корреляция. Тест Стьюдента.

9. Непараметрические тесты межгрупповых различий.

10. Визуализация групповых различий

11. A/B-тестирование. Проверка статистических гипотез.

12. Статистическая значимость и p-значение.

13. Проверка на основе t-статистики.

14. Дисперсионный анализ.

15. Проверка на основе статистики хи-квадрат

16. Простая линейная регрессия.

17. Множественная линейная регрессия.

18. Предсказания на основе регрессии.

19. Оценка качества предсказания. Перекрестная проверка.

20. Нелинейная регрессия

21. Наивный байесовский алгоритм.

22. Логистическая регрессия. Предсказание значений в логистической регрессии. Оценка качества предсказания

23. Алгоритм k ближайших соседей.
24. Древовидные алгоритмы (дерево решений, случайный лес).
25. Алгоритм бустинга.
26. Переобучение. Подбор гиперпараметров алгоритма

7.2.5 Примерный перечень заданий для подготовки к экзамену

Не предусмотрено учебным планом

7.2.6. Методика выставления оценки при проведении промежуточной аттестации

Зачет проводится по тест-билетам, каждый из которых содержит 10 вопросов и задачу. Каждый правильный ответ на вопрос в тесте оценивается 1 баллом, задача оценивается в 10 баллов (5 баллов верное решение и 5 баллов за верный ответ). Максимальное количество набранных баллов – 20.

1. Оценка «Не зачтено» ставится в случае, если студент набрал менее 11 баллов.

2. Оценка «Зачтено» ставится в случае, если студент набрал от 11 до 20 баллов.)

7.2.7 Паспорт оценочных материалов

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции	Наименование оценочного средства
1	Знакомство с R. Визуализация данных	УК-1, ПК-2, ПК-8	Тест
2	Визуализация данных	УК-1, ПК-2, ПК-8	Тест
3	Разведочный анализ данных	УК-1, ПК-2, ПК-8	Тест
4	Проверка статистических гипотез.	УК-1, ПК-2, ПК-8	Тест
5	Регрессия и классификация в R	УК-1, ПК-2, ПК-8	Тест
6	Статистическое машинное обучение.	УК-1, ПК-2, ПК-8	Тест

7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Тестирование осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

Решение стандартных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

Решение прикладных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных

задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

8 УЧЕБНО МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ)

8.1 Перечень учебной литературы, необходимой для освоения дисциплины

1 Python и анализ данных [Электронный ресурс] / Маккинли Уэс, А. Слинкина ; Уэс Маккинли; пер. А. Слинкина. - Python и анализ данных ; 2024-10-28. - Саратов : Профобразование, 2019. - 482 c.

2 Анализ данных [Электронный ресурс] : Учебно-методическое пособие / Г. В. Шнарева, Ж. Г. Пономарева ; Г. В. Шнарева, Ж. Г. Пономарева. - Анализ данных ; 2024-12-06. - Симферополь : Университет экономики и управления, 2019. - 129 c.

8.2 Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень лицензионного программного обеспечения, ресурсов информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем:

Лицензионное ПО:

- Microsoft Word
- PyCharm с плагином для R

Свободное программное обеспечение:

- LibreOffice

Отечественное ПО:

- СУБД Линтер

Ресурсы информационно-телекоммуникационной сети «Интернет»:

- http://www.edu.ru/
- Образовательный портал ВГТУ

Информационные справочные системы:

- http://window.edu.ru
- https://wiki.cchgeu.ru/

9 МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

Специализированная лекционная аудитория, оснащенная оборудованием для лекционных демонстраций и проекционной аппаратурой.

Учебные лаборатории (г. Воронеж, ул. Плехановская, д. 11):

- 202/2.
- 215/2.

Дисплейный класс, оснащенный компьютерными программами для проведения лабораторного практикума.

Кабинеты, оборудованные проекторами и интерактивными досками.

10. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

По дисциплине «Язык R и базовая статистика» читаются лекции, проводятся лабораторные работы.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, не нашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

Вид учебных занятий	Деятельность студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии.
Лабораторная работа	Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных для подготовки к ним необходимо: следует разобрать лекцию по соответствующей теме, ознакомится с соответствующим разделом учебника, проработать дополнительную литературу и источники, решить задачи и выполнить другие письменные задания.
Самостоятельная работа	Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие: - работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций; - выполнение домашних заданий и расчетов; - работа над темами для самостоятельного изучения; - участие в работе студенческих научных конференций, олимпиад; - подготовка к промежуточной аттестации.
Подготовка к промежуточной аттестации	Готовиться к промежуточной аттестации следует систематически, в течение всего семестра. Интенсивная подготовка должна начаться не позднее, чем за месяц-полтора до промежуточной аттестации. Данные перед зачетом три дня эффективнее всего использовать для повторения и систематизации материала.