Министерство науки и высшего образования Российской Федерации
Министерство науки, высшего образования и инноваций Кыргызской Республики

Межгосударственная образовательная организация высшего образования
Кыргызско-Российский Славянский университет имени
первого Президента Российской Федерации Б.Н. Ельцина

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
по дисциплине
«Анализ данных»

Уровень высшего образования: БАКАЛАВРИАТ
Направление подготовки: 38.03.01 - РФ, 580100 - КР Экономика
Профиль: «Математические методы в экономике»
Квалификация: бакалавр

Бишкек 2026

Фонд оценочных средств предназначен для контроля знаний обучающихся по направлению подготовки 38.03.01 - РФ, 580100 - КР Экономика профиль «Математические методы в экономике».

Фонд оценочных средств рассмотрен и утвержден на заседании Международной бизнес-школы протокол № _1_ от 03.09.2026 г.

Директор Международной бизнес-школы Кочербаева А.А.

Руководитель образовательной программы
Математические методы в экономике Мокроусов Н.В.

Исполнители:

Доцент Мусакулова Ж.А.

1. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ, С УКАЗАНИЕМ ЭТАПОВ ИХ ФОРМИРОВАНИЯ В ПРОЦЕССЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

В процессе освоения дисциплины «Анализ данных» обучающийся формирует профессиональную компетенцию **ПК-3: «Способен использовать специализированные программные средства для повышения офисной продуктивности и коммуникаций, проведения оценки и анализа, визуализации результатов анализа и предлагаемых решений».**

Формируемые компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенции	Виды оценочных средств / Шифр
ПК-3: Способен использовать специализированные программные средства для повышения офисной продуктивности и коммуникаций, проведения оценки и анализа, визуализации результатов анализа и предлагаемых решений	Знать (Уровень 1-3): 3.1.1. Функциональные возможности и назначение специализированных библиотек языка Python (NumPy, pandas, Matplotlib, Seaborn, statsmodels, scikit-learn) для импорта, очистки, предобработки и статистического анализа данных. 3.1.2. Правила выборки и настройки графических представлений (гистограммы, диаграммы рассеяния, boxplot, сводные сетки faceting) в зависимости от типа аналитических задач. 3.1.3. Методологию проведения разведочного анализа данных (EDA), агрегирования и построения сводных таблиц (groupby, pivot_table).	Блок A, D (Контрольные вопросы, Тесты)
	Уметь (Уровень 1-3): 3.2.1. Применять специализированный программный инструментарий (pandas, NumPy) для чтения, очистки, трансформирования, фильтрации и объединения (merge/join) табличных данных различных форматов (CSV, Excel). 3.2.2. Строить статистические и аналитические визуализации средствами Matplotlib и Seaborn, настраивать масштабы, палитры, подписи и добавлять поясняющие аннотации. 3.2.3. Формировать наглядные аналитические отчёты, сводные дашборды и презентационные графики для графической защиты предлагаемых решений.	Блок B, D (Лабораторные работы, Практикумы)
	Владеть (Уровень 1-3): 3.3.1. Навыками практической работы в специализированной среде Python для проведения первичной оценки, фильтрации и статистического анализа структурированных данных. 3.3.2. Техникой подготовки наглядных визуальных материалов (от графика для первичного исследования до лаконичных диаграмм для презентации). 3.3.3. Приёмами аргументированной коммуникации результатов анализа и обоснования выработанных рекомендаций перед руководящим составом и заказчиками.	Блок C, D (Сквозные проекты, Экзамен)

2. ТЕХНОЛОГИЧЕСКАЯ КАРТА ДИСЦИПЛИНЫ

Дисциплина: «Анализ данных» | Курс/семестр: 4 / 7 | Общая трудоемкость: 5 ЗЕТ (160 часов) | Форма контроля: Экзамен (7 семестр)
Распределение часов: Лекции – 32 ч., Практические (лабораторные) – 32 ч., Самостоятельная работа – 69 ч., Контроль – 26.7 ч.

Название модулей (разделов) дисциплины согласно РПД	Контроль	Форма контроля	Зачётный минимум	Зачётный максимум	График контроля (нед.)
Модуль 1. Введение в анализ данных и базовые инструменты (NumPy, pandas)	Текущий Рубежный	Лабораторные работы 1-4 Тест по Модулю 1	7 5	11 8	1-4 4
Модуль 2. Постановка вопросов и EDA (трансформации, объединения, строки и даты)	Текущий Рубежный	Лабораторные работы 5-8 (Кейс-стади) Тест по Модулю 2	8 5	12 8	5-8 8
Модуль 3. Визуализация и коммуникация результатов	Текущий Рубежный	Лабораторная работа 9-10 Защита аналитического дашборда	5 4	9 6	9-11 11
Модуль 4. Статистика и эксперименты (A/B, бутстрэп)	Текущий Рубежный	Лабораторные работы 11-12 Контрольная работа (A/B-тесты)	5 4	9 6	12-14 14
Модуль 5. Модели и ML-пайплайн (регрессия, классификация, кластеризация, PCA)	Текущий Рубежный	Лабораторные работы 13-15 Защита сквозного ML-проекта	6 5	9 8	15-17 17
ВСЕГО за 7 семестр (текущий + рубежный контроль)			45	70	
Промежуточный контроль (Экзамен)	Промежуточный	Итоговое компьютерное тестирование	15	30	Сессия
СЕМЕСТРОВЫЙ РЕЙТИНГ ПО ДИСЦИПЛИНЕ			60	100	

3. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ И ИНЫЕ МАТЕРИАЛЫ, НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ (ОЦЕНОЧНЫЕ СРЕДСТВА)

Блок А (ЗНАТЬ) — Контрольные вопросы по разделам дисциплины

Перечень теоретических вопросов для проверки уровня обученности «ЗНАТЬ» по 5 разделам дисциплины «Анализ данных»:

Раздел 1. Введение в анализ данных и базовые инструменты (NumPy, pandas)

1. Что лучше всего описывает роль аналитика данных в прикладном экономическом проекте?

2. Какой этап обычно следует сразу после постановки аналитического вопроса?

3. Какие два условия определяют совместимость форм массивов при векторизованных операциях broadcasting в NumPy?

4. Какое утверждение про срезы NumPy чаще всего верно: создают ли они представление (view) или копию (copy)?

5. Что показывает метод DataFrame.info() в библиотеке pandas и как он помогает оценить объем памяти?

6. Что такое индекс в pandas, каковы его типы и зачем он используется при агрегировании?

7. Какая операция является векторизованной по умолчанию в NumPy в отличие от стандартных списков Python?

8. Какой показатель описательной статистики более устойчив к выбросам: среднее арифмитическое или медиана?

9. Какой риск наиболее типичен при неверной постановке аналитической задачи?

10. Что корректнее использовать для табличных данных с разнородными типами столбцов: ndarray или DataFrame?

11. Какой подход к обработке пропущенных значений (NaN) считается ошибочным без анализа причин их возникновения?

12. Чем принципиально отличается индексация loc от iloc в pandas?

13. Какой метод удобнее для расчета нескольких агрегатов по группам: groupby().agg() или value_counts()?

14. Что означает предупреждение SettingWithCopyWarning в pandas и как его корректно избежать?

15. Какой инструмент pandas чаще всего используют для формирования сводных таблиц: pivot_table или merge?

Раздел 2. Постановка вопросов и EDA (трансформации, объединения, строки и даты)

1. Что является примером корректной, проверяемой аналитической гипотезы в экономике?

2. Почему в аналитических выводах нельзя автоматически интерпретировать статистическую корреляцию как причинно-следственную связь?

3. Какой разрез аналитики наиболее типичен для исследования временных рядов продаж?

4. Какой формат представления результата EDA наиболее полезен для бизнес-коммуникации: таблица, график или связка «таблица + график + вывод»?

5. В чем ключевое различие между операциями groupby и pivot_table?

6. Почему важно всегда анализировать объем сегмента (n) вместе со средними значениями показателей?

7. Что делает метод reset_index() после выполнения групповых агрегаций?

8. В каких ситуациях pivot_table предпочтительнее метода pivot?

9. Какой тип соединения (join) сохраняет все строки левой таблицы при слиянии Dataframe?

10. Что такое many-to-many merge и какой риск искажения данных он несет?

11. Какой способ предпочтительнее метода apply() для простых поэлементных вычислений в pandas?

12. Для чего используется метод melt() при реструктуризации таблиц?

13. Какой тест корректности объединения таблиц (merge) является базовым и обязательным?

14. Какую ошибку интерпретации может вызвать агрегирование по слишком крупным категориям?

15. Зачем приводить строковые поля к единому регистру и удалять начальные/конечные пробелы (str.strip)?

16. Что означает «ошибка парсинга даты» и как корректно ее обработать с помощью errors='coerce'?

17. Почему хранение дат в строковом формате (object) хуже, чем в специализированном типе datetime64?

Раздел 3. Визуализация и коммуникация результатов (Matplotlib & Seaborn)

1. Какой тип графика лучше всего подходит для оценки формы распределения непрерывной числовой переменной?

2. Какой график предпочтителен для сравнения медиан и квартилей по нескольким категориям (boxplot vs violinplot)?

3. Какой риск искажения восприятия возникает при обрезке оси Y для столбиковой диаграммы (bar chart)?

4. Что удобнее для построения сетки графиков: функция plt.subplots() или многократный вызов plt.figure()?

5. Какие обязательные элементы должны присутствовать на графике для его корректной интерпретации?

6. Как визуализация позволяет оперативно идентифицировать аномальные выбросы и ошибки в данных?

7. Для чего применяют доверительные интервалы и error bars на графиках сравнения средних?

8. Что такое фасетирование (faceting / FacetGrid) и когда оно полезно в анализе данных?

9. Какая цветовая палитра корректна для категориальных признаков, а какая для последовательных (sequential)?

10. Что означает принцип «визуальной честности» при выборе масштаба и осей координат?

11. Когда оправдано применение логарифмического масштаба на координатных осях?

12. Что отличает график для презентации руководству от графика для первичного исследовательского анализа?

13. Какой ключевой элемент Data Storytelling завершает аналитическое сообщение (Контекст — Вывод — Рекомендация)?

Раздел 4. Статистика и эксперименты (A/B, бутстрэп)

1. Чем отличается теоретическое понятие вероятность от статистической частоты события?

2. В чем различие между генеральной совокупностью и статистической выборкой?

3. Что означает параметр ddof=1 при расчете несмещенной оценки стандартного отклонения в Python?

4. Как корректно интерпретировать 95% доверительный интервал для разности средних?

5. Что такое нулевая гипотеза (H0) и альтернативная гипотеза (H1) в статистическом тестировании?

6. Что означает показатель p-value в классической проверке гипотез и каков порог значимости alpha?

7. Для чего применяется непараметрический метод бутстрэпа (bootstrap) при анализе выборок?

8. В чем заключается риск множественного тестирования (multiple testing) без внесения поправок (Bonferroni, FDR)?

9. Что такое ошибка II рода (False Negative) и как она связана с мощностью статистического теста?

Раздел 5. Модели и ML-пайплайн (регрессия, классификация, кластеризация, PCA)

1. Как интерпретируется коэффициент линейной регрессии при прочих равных условиях (ceteris paribus)?

2. Какая предпосылка метода наименьших квадратов (OLS) связана с постоянством дисперсии ошибок (гомоскедастичность)?

3. Какая метрика качества регрессии интерпретируется в исходных единицах измерения целевого признака (MAE, RMSE)?

4. Что показывает график остатков (residuals plot) и как с его помощью диагностировать нелинейность?

5. Что моделирует логистическая регрессия: прямое значение класса или вероятность принадлежности к классу?

6. Почему метрика Accuracy является вводящей в заблуждение при сильном дисбалансе классов?

7. В каких ситуациях PR-кривая (Precision-Recall) более информативна, чем ROC-кривая?

8. Что такое порог классификации (classification threshold) и как он влияет на баланс Precision и Recall?

9. Что показывает матрица ошибок (Confusion Matrix) и какие 4 исхода она содержит?

10. Что такое утечка данных (Data Leakage) в машинном обучении и как ее предотвратить?

11. Зачем необходим объект Pipeline в библиотеке scikit-learn?

12. Чем метод GridSearchCV отличается от RandomizedSearchCV при подборе гиперпараметров?

13. В чем заключаются основные цели применения метода главных компонент (PCA) в экономике?

14. Почему для алгоритма кластеризации k-means обязательной процедурой является масштабирование признаков (StandardScaler)?

Блок В (УМЕТЬ) — Лабораторные практические работы в Colab / Python

Комплекс лабораторных практических заданий для проверки уровня обученности «УМЕТЬ» (выполняются в среде Google Colab на языке Python):

Лабораторная работа №1-4 (Раздел 1). Вычисления в NumPy и трансформация данных в Pandas

• Задание 1. Создание и векторизованная обработка 1D/2D массивов NumPy (arange, linspace, random). Сравнение быстродействия циклов Python и векторизованных операций. Вычисление Z-score нормализации.
• Задание 2. Загрузка CSV/Excel файлов в pandas DataFrame, проведение первичной диагностики (.info(), .describe(), .isna()). Экспорт результатов в форматы CSV и Parquet.
• Задание 3. Составная фильтрация данных, расчет групповых агрегатов, ранжирование (top-N), предотвращение предупреждения SettingWithCopyWarning.

Лабораторная работа №5-8 (Раздел 2). Разведочный анализ данных (EDA) и подготовка датасетов

• Задание 1. Кейс-стади «Анализ продаж видеоигр»: очистка аномалий, построение сводных таблиц (groupby, pivot_table), проверка гипотез о различиях региональных продаж.
• Задание 2. Объединение табличных данных (inner, left join), контроль потерь строк, применение векторизованных операций взамен apply().
• Задание 3. Очистка строковых полей (str.strip, lower), парсинг и преобразование дат в datetime64, генерация 4 признаков времени (год, месяц, день недели, квартал).

Лабораторная работа №9-10 (Раздел 3). Практикум визуализации и Data Storytelling

• Задание 1. Построение комплексного дашборда 2x2 в Matplotlib & Seaborn (hist/KDE, boxplot, scatter plot, line plot) с аннотациями выбросов и экстремумов.
• Задание 2. Стилизация и логарифмирование масштабов, фасетирование по категориальным сегментам (FacetGrid).
• Задание 3. Подготовка аналитического одностраничника (Executive One-Pager) по схеме «Контекст — Вывод — Рекомендация».

Лабораторная работа №11-12 (Раздел 4). Статистический анализ, бутстрэп и симуляция A/B-тестов

• Задание 1. Расчет полных таблиц описательных статистик, оценке устойчивости среднего к выбросам (расчет до и после тримминга).
• Задание 2. Построение бутстрэп-распределения для разности медиан независимых выборок, расчет 95% доверительных интервалов.
• Задание 3. Симуляция A/B-эксперимента: задание эффекта (effect size), объема выборки и дисперсии, вычисление мощности теста и p-value.

Лабораторная работа №13-15 (Раздел 5). Построение ML-пайплайнов, регрессионный анализ и классификация

• Задание 1. Обучение OLS-регрессии в statsmodels, диагностика мультиколлинеарности через VIF (Variance Inflation Factor), анализ графиков остатков.
• Задание 2. Логистическая регрессия в scikit-learn, построение кривых ROC и PR, подбор оптимального порога классификации с учетом бизнес-затрат.
• Задание 3. Проектирование единого объекта Pipeline (StandardScaler + OneHotEncoder + Model), кросс-валидация и подбор гиперпараметров через GridSearchCV.

Блок С (ВЛАДЕТЬ) — Темы и спецификаторы сквозных практических проектов

Перечень тем сквозных практических проектов для оценки уровня «ВЛАДЕТЬ» (полный цикл EDA, моделирования и презентации результатов):

1. Прогнозирование экономических показателей и выручки предприятия (Задача регрессии: OLS, Ridge, Lasso, оценка гетероскедастичности и VIF).

2. Оценка и прогнозирование оттока клиентов (Churn Rate) в банковском секторе и сервисах (Задача классификации: Logistic Regression, Random Forest, подбор порогов, ROC-AUC / PR-AUC).

3. Исследовательский аналитический комплекс факторов спроса и коммерческого успеха продуктов (На примере датасета Video Game Sales / Ecommerce).

4. Сегментация клиентов и финансовых показателей предприятий (Обучение без учителя: PCA снижения размерности + k-means / иерархическая кластеризация).

5. Оценка эффективности продуктовых изменений и маркетинговых акций (A/B-тестирование: проверка гипотез, бутстрэп, пермутационные тесты и анализ мощности).

Блок D (ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ) — База тестовых заданий

Типовые тестовые задания с закрытой формой ответа (выбирается 1 правильный вариант) для проведения рубежного контроля и экзаменационного тестирования (30 вопросов из банка):

Вопрос №1 (Модуль 1). Какое из следующих утверждений наиболее точно описывает векторизацию в NumPy?

1) Выполнение операций через явные циклы for в Python
2) Выполнение поэлементных операций над массивами в компилируемом C-коде без явных циклов Python
3) Автоматическое преобразование двухмерных массивов в одномерные
4) Хранение данных в виде связных списков

Правильный ответ: 2

Вопрос №2 (Модуль 1). В каком случае при проведении операций над двумя массивами NumPy выполняется правило broadcasting?

1) Размерности массивов должны быть строго одинаковыми по всем осям
2) Для каждой оси их размеры либо равны, либо один из них равен 1
3) Обе матрицы должны быть квадратными
4) Один из массивов должен состоять только из нулей

Правильный ответ: 2

Вопрос №3 (Модуль 1). Что происходит при срезе (slicing) обычного ndarray в NumPy: b = a[1:5]?

1) Создается полная независимая копия данных в памяти
2) Создается представление (view), изменяющее исходный массив a при изменении b
3) Возникает ошибка индексации
4) Массив a автоматически удаляется

Правильный ответ: 2

Вопрос №4 (Модуль 1). Какой метод библиотеки pandas возвращает сводку о типах данных, количестве непустых значений и объеме занимаемой памяти?

1) df.describe()
2) df.info()
3) df.head()
4) df.shape

Правильный ответ: 2

Вопрос №5 (Модуль 1). Какое главное отличие индексатора loc от iloc в pandas?

1) loc работает по целым позициям, iloc — по меткам строк/столбцов

2) loc работает по текстовым или явным меткам строк/столбцов, iloc — по целочисленным позициям (0-indexed)

3) loc используется только для Series, iloc — только для DataFrame

4) Между ними нет никакого различия

Правильный ответ: 2

Вопрос №6 (Модуль 1). Предупреждение SettingWithCopyWarning в pandas обычно возникает при:

1) Попытке присвоить значение срезу DataFrame через цепочку индексации df[df['A']>0]['B'] = 1

2) Чтении несуществующего файла CSV

3) Делении на ноль в колонке типа float

4) Использовании функции pivot_table()

Правильный ответ: 1

Вопрос №7 (Модуль 1). Какая функция NumPy создаст массив из 100 равноотстоящих чисел в интервале от 0 до 1 включительно?

1) np.arange(0, 1, 100)

2) np.linspace(0, 1, 100)

3) np.random.rand(100)

4) np.ones(100)

Правильный ответ: 2

Вопрос №8 (Модуль 2). Что из перечисленного является первейшей задачей разведочного анализа данных (EDA)?

1) Обучение нейронной сети с миллионом параметров

2) Понимание структуры данных, поиск аномалий, проверка гипотез и генерация первичных аналитических выводов

3) Автоматическая генерация кода на языке C++

4) Настройка базы данных SQL

Правильный ответ: 2

Вопрос №9 (Модуль 2). Какой метод pandas удобнее применять для агрегирования данных по нескольким группам с расчетом нескольких различных метрик для каждого столбца?

1) df.value_counts()

2) df.groupby('cat').agg({'sales': 'sum', 'profit': ['mean', 'std']})

3) df.applymap()

4) df.sort_values()

Правильный ответ: 2

Вопрос №10 (Модуль 2). Какой тип слияния (merge) таблиц A и B сохранит абсолютно все строки из таблицы A, дополнив их совпадающими строками из B?

1) inner join

2) left join

3) right join

4) cross join

Правильный ответ: 2

Вопрос №11 (Модуль 2). Для чего используется метод pd.melt() в pandas?

1) Для удаления строк с пропусками

2) Для трансформации таблицы из широкого формата (wide) в длинный (long)

3) Для вычисления корреляционной матрицы

4) Для экспорта таблицы в Excel

Правильный ответ: 2

Вопрос №12 (Модуль 2). Какое выражение корректно переведет строковый столбец с датами '2026-11-12' в тип datetime64 в pandas?

1) df['date'] = df['date'].astype(int)

2) df['date'] = pd.to_datetime(df['date'])

3) df['date'] = df['date'].str.lower()

4) df['date'] = pd.Categorical(df['date'])

Правильный ответ: 2

Вопрос №13 (Модуль 2). Как извлечь день недели (0 - понедельник, 6 - воскресенье) из столбца 'date' типа datetime64?

1) df['date'].dt.dayofweek

2) df['date'].str.day()

3) df['date'].time.day()

4) df['date'].get_weekday()

Правильный ответ: 1

Вопрос №14 (Модуль 2). Какая операция удалит начальные и конечные пробелы в строковом столбце 'city'?

1) df['city'].str.strip()

2) df['city'].drop_spaces()

3) df['city'].clean()

4) df['city'].str.replace_all()

Правильный ответ: 1

Вопрос №15 (Модуль 3). Какой тип диаграммы является наиболее предпочтительным для анализа формы распределения одной непрерывной случайной величины?

1) Круговая диаграмма (Pie chart)

2) Гистограмма (Histogram) или график плотности (KDE)

3) Столбчатая диаграмма (Bar chart)

4) Диаграмма Гантта

Правильный ответ: 2

Вопрос №16 (Модуль 3). Что обозначает «жирная линия» внутри прямоугольника (box) на диаграмме размаха (boxplot)?

1) Среднее арифметическое значение

2) Медиану (50-й процентиль)

3) Моду распределения

4) Стандартное отклонение

Правильный ответ: 2

Вопрос №17 (Модуль 3). Какая библиотека Python построена поверх Matplotlib и предоставляет высокоуровневый интерфейс для статистической визуализации?

1) NumPy

2) Seaborn

3) SciPy

4) OpenPyXL

Правильный ответ: 2

Вопрос №18 (Модуль 3). Что такое faceting при построении графиков в Seaborn (например, через sns.relplot или catplot)?

1) Построение графика в трехмерном пространстве

2) Разбиение визуализации на сетку подграфиков (subplots) в зависимости от значений категориальной переменной

3) Изменение цвета фона на черный

4) Автоматическое сохранение графика в формат PDF

Правильный ответ: 2

Вопрос №19 (Модуль 3). Какой масштаб осей предпочтителен при визуализации показателей с разбросом в несколько порядков (например, доходы населения)?

1) Линейный масштаб

2) Логарифмический масштаб (Log scale)

3) Процентный масштаб

4) Двоичный масштаб

Правильный ответ: 2

Вопрос №20 (Модуль 3). Что из перечисленного снижает «визуальную честность» и искажает восприятие диаграммы столбцов (Bar Chart)?

1) Начало шкалы оси Y не с 0 (обрезка нуля)

2) Наличие подписей осей и единиц измерения

3) Использование контрастных цветов для выделения ключевого факта

4) Равномерный шаг сетки

Правильный ответ: 1

Вопрос №21 (Модуль 4). Почему медиана считается более устойчивой (робастной) статистикой, чем среднее арифметическое?

1) Медиана всегда больше среднего

2) На значение медианы не оказывают изолированного сильного влияния экстремальные выбросы в хвостах распределения

3) Медиану проще вычислить в уме

4) Медиана используется только для нормального распределения

Правильный ответ: 2

Вопрос №22 (Модуль 4). Что показывает p-value при проверке статистической гипотезы?

1) Вероятность того, что альтернативная гипотеза H1 верна

2) Вероятность получить такое же или еще более крайнее значение тестовой статистики при условии, что нулевая гипотеза H0 верна

3) Точный размер эффекта в популяции

4) Вероятность ошибки в исходных данных

Правильный ответ: 2

Вопрос №23 (Модуль 4). В чем заключается суть непараметрического бутстрэпа (bootstrap)?

1) В генерации новых данных с помощью нейронной сети

2) В многократном повторном извлечении выборок с возвращением (resampling with replacement) из исходного набора данных

3) В удалении всех нечисловых столбцов

4) В проверке равенства дисперсий по формуле Фишера

Правильный ответ: 2

Вопрос №24 (Модуль 4). Ошибка I рода (False Positive) при проведении A/B-тестирования — это:

1) Отклонение верной нулевой гипотезы H0 (утверждение об эффекте, когда его на самом деле нет)

2) Принятие неверной нулевой гипотезы H0 (пропуск реального эффекта)

3) Ошибка в математическом расчете среднего

4) Выбор неверного размера шрифта на графике

Правильный ответ: 1

Вопрос №25 (Модуль 4). Для чего при проведении множественного тестирования гипотез применяют поправки (например, Бонферрони или FDR)?

1) Для увеличения размера выборки

2) Для контроля и снижения групповой вероятности ошибки I рода (Family-Wise Error Rate)

3) Для ускорения работы компьютера

4) Для устранения пропусков в данных

Правильный ответ: 2

Вопрос №26 (Модуль 5). Что показывает коэффициент R^2 (коэффициент детерминации) в модели линейной регрессии?

1) Среднюю абсолютную ошибку прогноза

2) Долю дисперсии зависимой переменной, объясненную включенными в модель независимыми признаками

3) Вероятность правильной классификации

4) Наличие мультиколлинеарности

Правильный ответ: 2

Вопрос №27 (Модуль 5). Что моделирует логистическая регрессия для бинарной целевой переменной y ∈ {0, 1}?

1) Непосредственно непрерывное значение y
2) Логарифм шансов (logit) и вероятность P(y=1|X) попадания объекта в класс 1
3) Расстояние между кластерами
4) Главные компоненты матрицы X

Правильный ответ: 2

Вопрос №28 (Модуль 5). Почему метрика Accuracy может быть неинформативной при оценке качества модели?

1) При сильном дисбалансе классов (например, 99% норма, 1% отток) банальная константная модель дает Accuracy 99%, не находя целевой класс
2) Accuracy рассчитывается только для регрессии
3) Accuracy всегда равна 0
4) Accuracy не умеет работать с числовыми признаками

Правильный ответ: 1

Вопрос №29 (Модуль 5). Что представляет собой утечка данных (data leakage) при машинный обучении?

1) Похищение базы данных хакерами
2) Попадание информации из обучающей/тестовой целевой выборки в процесс препроцессинга или обучения (например, расчет StandardScaler по всей базе до разбиения)
3) Сохранение файла модели на открытый флеш-накопитель
4) Использование слишком большого числа деревьев

Правильный ответ: 2

Вопрос №30 (Модуль 5). В чем заключается главное преимущество использования Pipeline в scikit-learn?

1) Ускорение работы процессора в 10 раз
2) Объединение шагов трансформирования данных и оценки модели в единый объект с заменой fit/transform без утечки данных при кросс-валидации
3) Автоматическая отрисовка красиво оформленных графиков
4) Возможность работы без библиотеки Python

Правильный ответ: 2

Вопрос №31 (Модуль 5). Какой метод снижения размерности сохраняет максимум дисперсии данных путем ортогонального линейного преобразования?

1) K-Means
2) Метод главных компонент (PCA)
3) DBSCAN
4) Дерево решений

Правильный ответ: 2

Вопрос №32 (Модуль 5). Зачем перед обучением таких алгоритмов, как k-means, KNN или Ridge-регрессия, необходимо масштабирование признаков (StandardScaler)?

1) Без этого Python выдаст синтаксическую ошибку

2) Алгоритмы основаны на евклидовых расстояниях или L2-штрафах, поэтому признаки с большими абсолютными значениями будут доминировать

3) Масштабирование удаляет строковые столбцы

4) Это необходимо только для визуализации

Правильный ответ: 2

Вопрос №33 (Модуль 5). Какая библиотека Python специально предназначена для построения эконометрических моделей с выводом подробных сводок (summary) со статистиками t, p-value, R-squared?

1) statsmodels

2) OpenCV

3) PyGame

4) Flask

Правильный ответ: 1

Вопрос №34 (Модуль 5). Чем RandomizedSearchCV отличается от GridSearchCV при подборе гиперпараметров?

1) RandomizedSearchCV проверяет абсолютно все комбинации сетки

2) RandomizedSearchCV случайно сэмплирует заданное количество комбинаций из пространства параметров, экономя время

3) RandomizedSearchCV работает только для задач регрессии

4) Между ними нет никакой разницы

Правильный ответ: 2

Вопрос №35 (Модуль 5). Что показывает матрица ошибок (Confusion Matrix) для бинарной классификации?

1) Размеры исходного файла CSV

2) Количество True Positives (TP), False Positives (FP), True Negatives (TN) и False Negatives (FN)

3) Коэффициенты корреляции Пирсона

4) Время выполнения кода

Правильный ответ: 2

4. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ, ОПРЕДЕЛЯЮЩИЕ ПРОЦЕДУРЫ ОЦЕНИВАНИЯ ЗНАНИЙ, УМЕНИЙ И НАВЫКОВ

Лабораторная работа - один из видов практических работ. Целью лабораторной работы является углубление и закрепление теоретических знаний через развитие навыков обработки данных для решения поставленной задачи в присутствии и под руководством преподавателя.

Лабораторная работа служит для оценки освоения общепрофессиональных и профессиональных компетенций уровня «уметь» и «владеть».

Лабораторные работы включают задания по обработке количественных и качественных данных и решения исследовательских задач на их основе. Поскольку задания являются обширными, непосредственно в аудитории преподавателем разбирается постановка задачи, обосновываются и демонстрируются инструменты, необходимые для её решения, уточняются требования к оформлению результатов.

Окончательное выполнение лабораторной работы происходит в форме самостоятельной домашней работы.

Выполненная домашняя работа сдаётся по расписанию следующей лабораторной работы в виде файла. Работа проверяется преподавателем. Ошибки обсуждаются со студентом. Выставляется оценка.

Шкалы и критерии оценивания видов учебных работ

1. Критерии оценивания выполнения лабораторных практических работ (Уровень «УМЕТЬ» / «ВЛАДЕТЬ»):

Критерий оценивания	Низкий (0-30)	Фрагментарный (31-59)	Достаточный (60-84)	Высокий (85-100)	Вес (Xi)
Решение аналитической задачи	Ход решения ошибочен, кодовые блоки не исполняются	Ход решения верен, есть синтаксические и логические ошибки	Задача решена верно, есть не более 2 мелких неточностей	Задача решена полностью корректно, код векторизован и оптимизирован	0.6
Оформление и визуализация	Требования к стилю кода и графикам не выдержаны	Графики не имеют подписей осей и заголовков	Есть не более 2 мелких замечаний к оформлению графиков	Графики публикациионного качества, добавлены аннотации и выводы	0.3
Своевременность сдачи	Сдано с нарушением срока (0 баллов)	Сдано с нарушением срока (0 баллов)	Сдано своевременно (100 баллов)	Сдано своевременно (100 баллов)	0.1
ИТОГОВАЯ ОЦЕНКА			Формула расчета: Итог = 0.6 * X1 + 0.3 * X2 + 0.1 * X3		1.0

2. Шкала оценивания результатов компьютерного тестирования (Уровень «ЗНАТЬ»):

Низкий	Фрагментарный	Поверхностный	Достаточный	Высокий
0 - 30% верных ответов (0-30 баллов)	31 - 59% верных ответов (31-59 баллов)	60 - 69% верных ответов (60-69 баллов)	70 - 84% верных ответов (70-84 балла)	85 - 100% верных ответов (85-100 баллов)

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ И ВЫПОЛНЕНИЮ КОНТРОЛЬНЫХ ЗАДАНИЙ

Организация учебной работы и самостоятельной подготовки:
Курс «Анализ данных» имеет объём 5 ЗЕТ (160 часов) и преподаётся в 7 семестре. Обучение построено на сочетании лекционных занятий (32 ч.), практических лабораторных практикумов в среде Python / Google Colab (32 ч.) и самостоятельной работы студентов (69 ч.).

Рекомендуемый порядок выполнения лабораторных работ и практических заданий:
1. Ознакомиться с теоретическим материалом темы по конспекту лекций и официальной документации библиотек (pandas.pydata.org, numpy.org, seaborn.pydata.org).
2. Открыть шаблон Jupyter-ноутбука (.ipynb) в среде Google Colab, подключить необходимые библиотеки (import numpy as np, import pandas as pd, import matplotlib.pyplot as plt, import seaborn as sns).
3. Загрузить исходный датасет, провести проверку целостности данных, обработку пропусков и явных аномалий.
4. Выполнить расчет требуемых статистических показателей и группировок (groupby, pivot_table), оформить наглядные графики с подписями осей и легендой.
5. Сформулировать аналитические выводы под каждой решенной задачей в ячейках формата Markdown.
6. Сохранить файл под именем «Фамилия_Лабораторная_работа_N.ipynb» и представить его к защите преподавателю.

Балльно-рейтинговая система перевода итоговой суммы баллов в оценку:

Семестровый рейтинг (баллы)	Академическая оценка	Уровень сформированности компетенции ПК-3
85 - 100 баллов	«Отлично»	Высокий уровень
70 - 84 баллов	«Хорошо»	Достаточный уровень
60 - 69 баллов	«Удовлетворительно»	Поверхностный уровень
Менее 60 баллов	«Неудовлетворительно»	Низкий / Фрагментарный уровень