МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ,
МИНИСТЕРСТВО НАУКИ, ВЫСШЕГО ОБРАЗОВАНИЯ И ИННОВАЦИЙ
КЫРГЫЗСКОЙ РЕСПУБЛИКИ

МОО ВО Кыргызско-Российский Славянский университет
имени первого Президента Российской Федерации Б.Н. Ельцина

УТВЕРЖДАЮ
Директор Социо-Гуманитарного Института
Голобородько И. Л.
_____ 2026 г.

Анализ данных

рабочая программа дисциплины (модуля)

Закреплена за школой **Международная бизнес-школа**

Учебный план Направление 38.03.01 - РФ, 580100 - КР Экономика
Профиль "Математические методы в экономике"

Квалификация **бакалавр**
Форма обучения **очная**

Общая трудоемкость **5 ЗЕТ**
Часов по учебному плану 180
в том числе:
аудиторные занятия 64
самостоятельная работа 89

Виды контроля в семестрах:
экзамен 7

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	7 (4.1)		Итого	
Недель	18			
Вид занятий	УП	РП	УП	РП
Лекции	32	32	32	32
Практические	32	32	32	32
Контактная работа в период экзаменационной сессии	0,3	0,3	0,3	0,3
В том числе инт.	14	14	14	14
В том числе в форме практ.подготовки	32	32	32	32
Итого ауд.	64	64	64	64
Контактная работа	64,3	64,3	64,3	64,3
Сам. работа	89	89	89	89
Часы на контроль	26,7	26,7	26,7	26,7
Итого	180	180	180	180

Программу составил(и):
Доцент, к.т.н., Мусакулова Жылдыз Абдыманаповна _____

Согласовано РОП:
Мокроусов Николай Васильевич _____

Рабочая программа дисциплины
Анализ данных

разработана в соответствии с ФГОС 3++:
Федеральный государственный образовательный стандарт высшего образования - бакалавриат по направлению подготовки 38.03.01 Экономика (приказ Минобрнауки России от 12.08.2020 г. № 954)

составлена на основании учебного плана:

Направление 38.03.01 - РФ, 580100 - КР Экономика
Профиль "Математические методы в экономике"
утвержденного учёным советом вуза от 26.06.2026 протокол № 11.

Рабочая программа одобрена на заседании Международной бизнес-школы

Протокол от 03.09.2026 г. № 1
Срок действия программы: 2026-2030 уч.г.
Руководитель подразделения: Кочербаева Айнура Анатольевна

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ
к рабочей программе дисциплины (модуля)

№ п/п	Основание внесения изменения	Раздел/ пункт	Содержание изменения
1.			
2.			
3.			
4.			
5.			

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ	
1.1	Основной целью освоения дисциплины «Анализ данных» является формирование у обучающихся комплексов профессиональных компетенций в области сбора, очистки, исследовательского анализа (EDA), статистической проверки гипотез, визуализации и машинного обучения для решения прикладных экономических и аналитических задач.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП	
Цикл (раздел) ООП:	Б1.В
2.1	**Требования к предварительной подготовке обучающегося:**
2.1.1	Б1.В.03. Базы данных
2.1.2	Б1.В.18. Объектно-ориентированное программирование и визуальное моделирование
2.2	**Дисциплины и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:**
2.2.1	Б2.О.02(Пд). Преддипломная практика
2.2.2	Б1.В.14. Системы искусственного интеллекта

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)	
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	Функциональные возможности и назначение специализированных библиотек языка Python (NumPy, pandas, Matplotlib, Seaborn, statsmodels, scikit-learn) для импорта, очистки, предобработки и статистического анализа данных.
Уровень 2	Правила выборки и настройки графических представлений (гистограммы, диаграммы рассеяния, boxplot, сводные сетки faceting) в зависимости от типа аналитических задач и распределений переменных.
Уровень 3	Методологию проведения разведочного анализа данных (EDA), агрегирования и построения сводных таблиц (groupby, pivot_table).
Уметь:	
Уровень 1	Применять специализированный программный инструментарий (pandas, NumPy) для чтения, очистки, трансформирования, фильтрации и объединения (merge/join) табличных данных различных форматов (CSV, Excel).
Уровень 2	Строить статистические и аналитические визуализации средствами Matplotlib и Seaborn, настраивать масштабы, палитры, подписи и добавлять поясняющие аннотации.
Уровень 3	Формировать наглядные аналитические отчёты, сводные дашборды и презентационные графики для графической защиты предлагаемых решений.
Владеть:	
Уровень 1	Навыками практической работы в специализированной среде Python для проведения первичной оценки, фильтрации и статистического анализа структурированных данных.
Уровень 2	Техникой подготовки наглядных визуальных материалов (от графика для первичного исследования до лаконичных диаграмм для презентации).
Уровень 3	Приёмами аргументированной коммуникации результатов анализа и обоснования выработанных рекомендаций перед руководящим составом и заказчиками.

В результате освоения дисциплины обучающийся должен

3.1	**Знать:**
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 и распределений переменных.
3.1.3	Методологию проведения разведочного анализа данных (EDA), агрегирования и построения сводных таблиц (groupby, pivot_table).
3.2	**Уметь:**
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.
3.3	**Владеть:**
3.3.1	Практической работы в специализированной среде Python для проведения первичной оценки, фильтрации и статистического анализа структурированных данных.
3.3.2	Техники подготовки наглядных визуальных материалов (от графика для первичного исследования до лаконичных диаграмм для презентации).
3.3.3	Приёмов аргументированной коммуникации результатов анализа и обоснования выработанных рекомендаций перед руководящим составом и заказчиками.

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Код занятия	Наименование разделов и тем / вид занятия/	Семестр / Курс	Часов	Компетен-ции	Литература	Инте ракт.	Пр. подг.	Примечание
	Раздел 1. Введение в анализ данных и базовые инструменты (NumPy, pandas).							
1.1	Введение в анализ данных /Лек/	7	2	ПК-3	Л1.1Л2.1 Л2.2Л3.1 Э1 Э2 Э3			
1.2	Введение в анализ данных: от данных к экономическому выводу /Пр/	7	2	ПК-3	Л1.3Л2.1Л3.2 Э1 Э2 Э3		2	
1.3	Архитектура вычислений в NumPy: Базовые структуры, векторизация и производительность при работе с многомерными данными /Лек/	7	2	ПК-3	Л1.1 Л1.2 Л1.3Л2.1 Л2.2Л3.2 Э1 Э2 Э3			
1.4	NumPy: базовые операции с массивами, формы, индексация / Пр/	7	2	ПК-3	Л1.2Л2.2Л3.2 Э1 Э2 Э3		2	
1.5	NumPy: Математика данных. Broadcasting, Векторизация и Базовая статистика /Лек/	7	2	ПК-3	Л1.1 Л1.3Л2.2Л3.1 Э1 Э2 Э3			
1.6	NumPy: broadcasting, векторизация и базовая статистика /Пр/	7	2	ПК-3	Л1.1Л2.1Л3.1 Э1 Э2 Э3		2	
1.7	Библиотека Pandas: от структур данных до агрегации /Лек/	7	2	ПК-3	Л1.2Л2.1Л3.2 Э1 Э2 Э3			
1.8	Загрузка, исследование, преобразование и сохранение данных в pandas /Пр/	7	2	ПК-3	Л1.3Л2.1Л3.2 Э1 Э2 Э3		2	
1.9	Работа с многомерными массивами в NumPy /Ср/	7	5	ПК-3	Л1.2Л2.2Л3.1 Э1 Э2 Э3			
1.10	Загрузка, первичная диагностика и экспорт данных в Pandas /Ср/	7	5	ПК-3	Л1.2Л2.2Л3.1 Л3.2 Э1 Э2 Э3			
1.11	Фильтрация, трансформация и базовое агрегирование в Pandas / Ср/	7	5	ПК-3	Л1.2Л2.1Л3.2 Э1 Э2 Э3			
	Раздел 2. Постановка вопросов и EDA (трансформации, объединения, строки и даты)							
2.1	Аналитическое мышление и первичный анализ данных (EDA) /Лек/	7	2	ПК-3	Л1.3Л2.2Л3.2 Э1 Э3 Э4			

2.2	Постановка аналитических вопросов, первичная EDA и анализ продаж видеоигр /Пр/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.1 Л3.2 Э3 Э4	2	2	
2.3	EDA: Группировки, сводные таблицы и агрегирование данных в Pandas /Лек/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.2 Э3 Э4 Э6	2		
2.4	EDA: Группировки, сводные таблицы и агрегирование данных в Pandas /Пр/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.1 Л3.2 Э1 Э3 Э6		2	
2.5	Продвинутый EDA в Pandas: Интеграция, Трансформация и Контроль качества данных /Лек/	7	2	ПК-3	Л1.3Л2.1Л3.2 Э1 Э5 Э6			
2.6	EDA: объединение таблиц, применение apply и изменение структуры данных в Pandas /Пр/	7	2	ПК-3	Л1.2Л2.1Л3.1 Л3.2 Э1 Э5 Э6		2	
2.7	Строки и Даты в Pandas: Стандартизация, признаки времени и предотвращение утечек данных в машинном обучении /Лек/	7	2	ПК-3	Л1.2 Л1.3Л2.1Л3.1 Л3.2 Э1 Э3			
2.8	Обработка строковых и временных данных. Контроль качества данных в Pandas /Пр/	7	2	ПК-3	Л1.2Л2.1Л3.1 Л3.2 Э1 Э3 Э6		2	
2.9	Кейс-стади: исследовательский анализ продаж игр /Ср/	7	5	ПК-3	Л1.1 Л1.2 Л1.3Л2.1Л3.1 Л3.2 Э3 Э5 Э6			
2.10	Группировки, сводные таблицы и расчет долей /Ср/	7	5	ПК-3	Л1.1 Л1.2 Л1.3Л2.1Л3.1 Л3.2 Э1 Э3			
2.11	Объединение таблиц, изменение структуры и векторизация /Ср/	7	5	ПК-3	Л1.1 Л1.2 Л1.3Л2.1 Л2.2Л3.1 Л3.2 Э1 Э5 Э6			
2.12	Обработка текстов и временных рядов /Ср/	7	5	ПК-3	Л1.1 Л1.2 Л1.3Л2.1 Л2.2Л3.1 Л3.2 Э1 Э3 Э5			
	Раздел 3. Визуализация и коммуникация результатов							
3.1	Визуализация данных в Python: Четыре функции визуализации в аналитике /Лек/	7	2	ПК-3	Л1.1Л2.1Л3.2 Э1 Э7			
3.2	Визуализация данных с использованием Matplotlib и Seaborn /Пр/	7	2	ПК-3	Л1.1Л2.1Л3.1 Л3.2 Э1 Э7		2	
3.3	Продвинутая визуализация и оформление результатов /Лек/	7	2	ПК-3	Л1.1 Л1.2Л2.1Л3.1 Л3.2 Э1 Э3 Э7	2		
3.4	Визуализация для коммуникации: сторителлинг и аннотации /Пр/	7	2	ПК-3	Л1.1 Л1.2 Л1.3Л2.1 Л2.2Л3.1 Л3.2 Э7		2	

3.5	Продвинутая стилизация и фасетирование визуализаций / Ср/	7	5	ПК-3	Л1.1 Л1.2Л2.1Л3.1 Л3.2 Э1 Э3 Э7			
3.6	Data Storytelling и адаптация графиков под аудиторию /Ср/	7	5	ПК-3	Л1.1Л2.1Л3.1 Л3.2 Э7			
	Раздел 4. Статистика и эксперименты (A/B, бутстрэп)							
4.1	Вероятности и распределения, выборки, описательная статистика /Лек/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.1 Э1 Э2 Э3			
4.2	Описательная статистика и распределения в Python /Пр/	7	2	ПК-3	Л1.1Л2.1Л3.1 Э1 Э3 Э7	2	2	
4.3	Доверительные интервалы, проверка гипотез, A/B-тесты / Лек/	7	2	ПК-3	Л1.1Л2.1Л3.1 Л3.2 Э1 Э7			
4.4	Статистика и анализ распределений в Python /Пр/	7	2	ПК-3	Л1.1Л2.1Л3.1 Л3.2 Э3 Э5 Э6		2	
4.5	Бутстрэп, симуляции и тесты гипотез /Лек/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.1 Э1			
4.6	Бутстрэп, тесты и симуляция A/B-тестирования /Пр/	7	2	ПК-3	Л1.1 Л1.3Л2.1Л3.1 Л3.2 Э1 Э3 Э5	2	2	
4.7	Описательная статистика и анализ распределений в Python / Ср/	7	5	ПК-3	Л1.1Л2.1Л3.1 Э1 Э5			
4.8	Оценка доверительных интервалов методом бутстрэпа, проведение непараметрических тестов и симуляция A/B-экспериментов. /Ср/	7	4	ПК-3	Л1.1 Л1.2 Л1.3Л2.2Л3.1 Э1 Э3 Э7			
	Раздел 5. Модели и ML-пайплайн (регрессия, классификация, кластеризация, PCA)							
5.1	Линейная регрессия: постановка, интерпретация, диагностика / Лек/	7	2	ПК-3	Л1.1 Л1.2 Л1.3Л2.2Л3.1 Л3.2 Э3 Э5 Э6			
5.2	Линейная регрессия (OLS) и диагностика остатков в statsmodels /Пр/	7	2	ПК-3	Л1.1Л2.2Л3.1 Э3 Э5 Э6	2	2	
5.3	Классификация: логистическая регрессия и метрики /Лек/	7	2	ПК-3	Л1.1 Л1.3Л2.2Л3.1 Л3.2 Э4 Э5 Э6			
5.4	Логистическая регрессия и настройка порогов классификации /Пр/	7	2	ПК-3	Л1.1Л2.1 Л2.2Л3.1 Э4 Э5 Э6	2	2	
5.5	ML-пайплайн: подготовка признаков, валидация, CV /Лек/	7	2	ПК-3	Л1.1Л2.2Л3.2 Э4 Э5 Э6			
5.6	Построение воспроизводимых ML-пайплайнов и подбор гиперпараметров /Пр/	7	2	ПК-3	Л1.1 Л1.2 Л1.3Л2.2Л3.1 Э4 Э5 Э6		2	

5.7	Логистическая регрессия и настройка порогов классификации /Ср/	7	5	ПК-3	Л1.2 Л1.3Л2.1Л3.1 Э4 Э5 Э6			
5.8	Создание единых конвейеров обработки и обучения в scikit-learn (Pipeline) /Ср/	7	5	ПК-3	Л1.1Л2.2Л3.1 Л3.2 Э4 Э5 Э6			
5.9	Решение сквозной кейс-задачи: от первичного EDA и валидации качества до обучения модели, сравнения альтернатив на отложенной выборке /Ср/	7	5	ПК-3	Л1.1 Л1.2 Л1.3Л2.1 Л2.2Л3.1 Э4 Э5 Э6			
5.10	Консультация /КрЭк/	7	0,3	ПК-3	Л1.1 Л1.3Л2.2Л3.2 Э4 Э5 Э6			

5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

5.1. Контрольные вопросы и задания

Раздел 1. Тестовые вопросы

Что лучше всего описывает роль аналитика данных в проекте?
Какой этап обычно следует сразу после постановки аналитического вопроса?
Какие два условия определяют совместимость форм при broadcasting в NumPy?
Какое утверждение про срезы NumPy чаще всего верно (view или copy)?
Что показывает метод DataFrame.info() в pandas?
Что такое индекс в pandas и зачем он может быть полезен?
Какая операция является векторизованной по умолчанию в NumPy?
Какой показатель описательной статистики более устойчив к выбросам: среднее или медиана?
Какой риск наиболее типичен при неверной постановке задачи анализа?
Что корректнее использовать для табличных данных с разными типами столбцов: ndarray или DataFrame?
Какой подход к пропускам чаще всего считается ошибочным без анализа причин их возникновения?
Чем отличается loc от iloc?
Какой метод удобнее для нескольких агрегатов по группам: groupby().agg() или value_counts()?
Что означает предупреждение SettingWithCopy и как его избежать?
Какой инструмент pandas чаще всего используют для сводных таблиц: pivot_table или merge?
Что нужно сделать перед сравнением сегментов, если в данных много пропусков по ключевым полям?

Раздел 2. Тестовые вопросы

Что является примером корректной аналитической гипотезы?
Почему в выводах нельзя автоматически интерпретировать корреляцию как причинность?
Какой разрез наиболее типичен для анализа продаж по времени?
Какой формат результата EDA наиболее полезен для коммуникации: таблица, график или связка «таблица + график + вывод»?
В чем ключевое различие groupby и pivot_table?
Почему важно всегда смотреть на размер группы (n) вместе со средними значениями?
Что делает reset_index() после агрегирования?
Когда pivot_table предпочтительнее pivot?
Какой join сохраняет все строки левой таблицы?
Что такое many-to-many merge и какой риск он несет?
Какой способ обычно предпочтительнее apply для простых поэлементных вычислений?
Для чего используется метод melt?
Какой тест корректности merge является наиболее базовым и обязательным?
Какую ошибку интерпретации может вызвать агрегирование по слишком крупным сегментам?
Зачем приводить строковые поля к единому регистру и убирать пробелы по краям?
Что означает «ошибка парсинга даты» и как корректно ее обработать?
Какой признак времени нельзя извлечь из даты напрямую без внешних данных?
Почему хранить даты строкой обычно хуже, чем в типе datetime?

Раздел 3. Тестовые вопросы

Какой график лучше всего подходит для распределения числовой переменной?
Какой график лучше всего подходит для сравнения медиан по категориям?
Какой риск возникает при обрезке оси Y для bar / line графика?

Что удобнее для построения сетки графиков: subplots или многократный figure()?
Что обязательно должно быть на графике для корректной интерпретации метрики?
Как визуализация может помочь найти выбросы?
Для чего применяют error bars или интервалы неопределенности на графике?
Что такое faceting и когда он полезен?
Какая палитра корректнее для категориальных данных?
Что означает «визуальная честность» при выборе масштаба?
Когда оправдан логарифмический масштаб?
Что чаще всего отличает график для презентации от графика для анализа?
Какой элемент сторителлинга завершает историю в аналитике?
В чем риск cherry-picking при визуализации?
Какой прием лучше всего помогает быстро выделить одну категорию среди многих?
Как проверить, что график понятен без доклада?

Раздел 4. Тестовые вопросы

Чем отличается вероятность от частоты?
Что такое генеральная совокупность и выборка?
Какая статистика более устойчива к выбросам: среднее или медиана?
Что означает ddof=1 при расчете стандартного отклонения?
Как корректно интерпретировать 95% доверительный интервал?
Что такое нулевая гипотеза?
Что означает p-value в классической проверке гипотез?
Для чего применяют бутстрэп?
В чем риск multiple testing без поправок?
Что такое ошибка II рода?

Раздел 5. Тестовые вопросы

Что означает коэффициент в линейной регрессии при прочих равных условиях?
Какая предпосылка OLS связана с постоянной дисперсией ошибок?
Какая метрика чаще всего интерпретируется как «средняя абсолютная ошибка»?
Что показывает график остатков?
Что моделирует логистическая регрессия: значения класса или вероятность?
Почему accuracy может быть плохой метрикой при сильном дисбалансе классов?
Когда PR-кривая обычно информативнее ROC-кривой?
Что такое порог классификации и как он влияет на precision / recall?
Что показывает матрица ошибок (confusion matrix)?
Что такое утечка данных (data leakage) в классификации?
Почему нельзя подбирать гиперпараметры на test выборке?
Что дает кросс-валидация по сравнению с одним train/valid split?
Зачем нужен Pipeline в scikit-learn?
Чем GridSearchCV отличается от RandomizedSearchCV?
Какой пример является утечкой данных при препроцессинге?
Какой подход корректен для сравнения двух моделей?
В чем цель метода PCA?
Почему для k-means часто нужно масштабирование признаков?
Что измеряет silhouette score (в общих чертах)?
Какие элементы делают итоговый анализ воспроизводимым?

5.2. Темы курсовых работ (проектов)

Темы сквозных проектов.
1. Прогнозирование экономических показателей и выручки предприятия (Задача регрессии)
2. Оценка и прогнозирование оттока клиентов (Churn Rate) в сервисах / банках (Задача классификации)
3.Исследовательский анализ коммерческого успеха продуктов и факторов спроса (На примере датасета Video Game Sales)
4. Сегментация клиентов и финансовых показателей предприятий (Обучение без учителя)
5. Оценка эффективности продуктовых изменений и маркетинговых акций (A/B-тестирование)

5.3. Фонд оценочных средств

Фонд оценочных средств включает лабораторные практические задания и тестовые вопросы по каждому разделу.

Лабораторные работы к Разделу 1 (NumPy & Pandas)

Базовые массивы NumPy: Создание 1D и 2D массивов различными способами (arange, linspace, zeros, ones, random). Индексация, срезы, сравнение работы copy и view. Поэлементные и матричные операции, вычисление суммы по осям, mean, std. Нормализация признаков матрицы (z-score) по столбцам.
Сравнение производительности: Сравнение времени выполнения циклов Python и векторизованных операций NumPy на задаче суммы квадратов.

Первичный обзор и сохранение данных в Pandas: Загрузка CSV/Excel с настройкой разделителей и типов. Проведение диагностики (info, describe, isna). Запись результатов в форматы CSV и Parquet/Feather со сравнением объема файлов и скорости чтения.
Фильтрация и группировка в Pandas: Составная фильтрация с обработкой пропусков. Создание признаков через assign или прямое присвоение (строковые, временные, арифметические). Расчет групповых агрегатов (минимум 3 метрики). Ранжирование внутри групп (топ-3 по категории). Настройка assert-проверок на корректность размеров и отсутствия отрицательных значений.

Лабораторные работы к Разделу 2 (EDA & Слияния)

Кейс-стади «Анализ продаж видеоигр»: Очистка датасета от аномалий. Формулировка 3 аналитических вопросов с проведением мини-EDA (таблица + график + вывод). Построение сводных таблиц по платформам, жанрам и годам. Статистическая проверка гипотезы о различии региональных продаж по жанрам. Подготовка мини-отчета.
Агрегирование и изменение формы: Построение двух агрегированных таблиц (одиночный и составной ключи). Сравнение работы groupby().agg() и pivot_table. Вычисление внутригрупповых долей. Контроль сходимости групповых сумм с итоговыми показателями.
Объединение таблиц и оптимизация: Слияние таблиц методами inner и left с анализом потерь данных. Поиск и исправление несовпавших ключей. Перевод данных из широкого формата в длинный (melt) и обратно (pivot_table). Переписывание функции apply на векторизованную операцию.
Качество строковых и временных полей: Очистка строк (trim, регистр, нормализация категорий). Преобразование строк в datetime с обработкой невалидных значений. Создание 4 признаков времени (год, месяц, день недели, квартал, флаг выходного дня). Формирование итогового отчета качества данных (доли пропусков, дубликатов, невалидных дат).

Лабораторные работы к Разделу 3 (Визуализация)

Практикум Matplotlib & Seaborn: Построение комплексного набора из 4 графиков разных типов (hist/box/violin, bar, scatter, line). Оформление стилей, подписей осей, единиц измерения и заголовков. Построение сравнения категорий по времени. Аннотирование выбросов и максимумов. Сборка дашборда-сетки 2x2.
Публикационная подготовка графиков: Рефакторинг сырых диаграмм: настройка логарифмического масштаба, правил форматирования чисел (проценты, денежные суммы). Построение фасетных сеток по сегментам. Создание лаконичной версии графика для презентации и детальной версии для анализа.
Сторителлинг и коммуникация: Создание аналитического одностраничника для выбранной целевой аудитории: 1–2 сфокусированных графика с аннотациями и текстовым блоком «Контекст — Вывод — Рекомендация».

Лабораторные работы к Разделу 4 (Статистика & Эксперименты)

Описательный статистический анализ: Расчет полных таблиц точечных статистик. Построение и сравнение графиков распределения (гистограмма vs KDE). Оценка устойчивости среднего к выбросам (расчет до и после удаления аномалий). Проведение стандартизации и нормализации признаков.
Бутстрэп и симуляции A/B-тестирования: Построение бутстрэп-распределения для разницы средних/медиан двух независимых групп. Реализация пермутационного теста. Проведение симуляционного A/B-теста с заданием величины эффекта, размера выборки и дисперсии для вычисления мощности. Оценка стойкости к дисбалансу выборок и выбросам.

Лабораторные работы к Разделу 5 (Моделирование & ML)

Регрессионный анализ в statsmodels: Обучение OLS-модели, детальный разбор сводки коэффициентов и R^2. Построение графиков остатков, проверка гетероскедастичности тестами и графически. Диагностика мультиколлинеарности через VIF. Сравнение нескольких спецификаций моделей.
Классификация и подбор порогов в scikit-learn: Обучение логистической регрессии, получение прогнозов вероятностей. Построение кривых ROC и PR. Подбор оптимального порога классификации с учетом минимизации специфических ошибок бизнеса. Построение и анализ матрицы ошибок.
Проектирование ML-пайплайнов: Создание единого объекта Pipeline («препроцессинг + модель»). Настройка подбора гиперпараметров через GridSearchCV / RandomizedSearchCV на кросс-валидации. Проверка отсутствия утечки данных на всех шагах. Формирование сравнительной таблицы моделей.
Итоговый аналитический проект: Полный цикл обработки данных: EDA, очистка, построение baseline-модели, проектирование улучшенного пайплайна, финальная кросс-валидация на отложенном holdout-сете. Оформление воспроизводимого отчета с рекомендациями и описанием ограничений.

5.4. Перечень видов оценочных средств

1. Текущий контроль успеваемости
Лабораторные (практические) работы в Google Colab
Контрольные вопросы для самопроверки
Самостоятельные кейс-стади и мини-EDA
Фонд тестовых заданий (Тестовые вопросы)

2. Промежуточная аттестация
Защита сквозного практического проекта (Итоговый практикум)
Итоговое письменное/компьютерное тестирование

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

6.1. Рекомендуемая литература

6.1.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год
Л1.1	Маккинли У.	Python и анализ данных: учебное пособие	М.: ДМК Пресс 2015
Л1.2	Гаско, Р.	Простой Python просто с нуля	Москва: СОЛОН-ПРЕСС 2023
Л1.3	Монгуш, Ч.М.	Введение в анализ данных: учебное пособие для студентов	Кызыл: Издательство Тувинского государственного университета 2022

6.1.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год
Л2.1	Шолле Ф.	Глубокое обучение на Python: учебное пособие	СПб.: Питер 2018
Л2.2	Петрунин Ю.Ю.	Искусственные нейронные сети в экономике: математический инструмент, модель или методология?	

6.1.3. Методические разработки

	Авторы, составители	Заглавие	Издательство, год
Л3.1	Маккинли, У., Слинкина, А.	Python и анализ данных	Саратов: Профобразование 2024
Л3.2	Титов, А.Н., Тазиева, Р.Ф.	Обработка данных в Python. Основы работы с библиотекой Pandas: учебно-методическое пособие	Казань: Издательство КНИТУ 2022

6.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	Официальный сайт библиотеки pandas (pandas documentation) : сайт. — URL: https://pandas.pydata.org/docs/	URL: https://pandas.pydata.org/docs/
Э2	Официальная документация библиотеки NumPy (NumPy User Guide) : сайт. — URL: https://numpy.org/doc/	URL: https://numpy.org/doc/
Э3	Google Colaboratory (Облачная интерактивная среда выполнения Python) : сайт. — URL: https://colab.research.google.com/	URL: https://colab.research.google.com/
Э4	Образовательная платформа «Юрайт» : образовательная платформа / ЭБС. — Москва : ООО «Электронное издательство Юрайт». —	URL: https://urait.ru/
Э5	World Bank Open Data (Открытые данные Всемирного банка) : сайт. — URL: https://data.worldbank.org/	URL: https://data.worldbank.org/
Э6	Kaggle Datasets: Открытый репозиторий датасетов и аналитических ноутбуков : сайт. — URL: https://www.kaggle.com/datasets/	URL: https://www.kaggle.com/datasets/
Э7	Официальная документация библиотеки Seaborn (Statistical Data Visualization) : сайт. — URL: https://seaborn.pydata.org/	URL: https://seaborn.pydata.org/

6.3. Перечень информационных и образовательных технологий

6.3.1 Компетентностно-ориентированные образовательные технологии

6.3.1.1	Для проведения практических занятий, лабораторных работ в интерактивной среде и выполнения сквозного аналитического проекта используется следующий комплект лицензионного и свободно распространяемого ПО:
6.3.1.2	Операционная система: Microsoft Windows 10/11 (или Linux / macOS).
6.3.1.3	Веб-браузеры: Google Chrome / Yandex Browser / Mozilla Firefox (с поддержкой современных веб-стандартов для стабильной работы в облачных средах).
6.3.1.4	Google Colab (Google Research) — основная облачная платформа исполнения Jupyter-ноутбуков (.ipynb).
6.3.1.5	Специализированные библиотеки и пакеты анализа данных Python.
6.3.1.6	Офисные программы: Microsoft Office (Word, Excel) / LibreOffice / Яндекс

6.3.2 Перечень информационных справочных систем и программного обеспечения

6.3.2.1	Национальный статистический комитет Кыргызской Республики (Нацстатком КР) — [stat.kg]
6.3.2.2	Федеральная служба государственной статистики (Росстат) — [rosstat.gov.ru]
6.3.2.3	World Bank Open Data (Всемирный банк) — [data.worldbank.org]
6.3.2.4	IMF Data (Международный валютный фонд) — [data.imf.org]
6.3.2.5	
6.3.2.6	Kaggle Datasets — Крупнейший мировой репозиторий датасетов для проведение разведочного анализа данных (EDA), машинного обучения и проведения A/B-экспериментов.
6.3.2.7	UCI Machine Learning Repository — Эталонная база данных калиброванных наборов данных для тестирования алгоритмов регрессии, классификации и кластеризации.

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)	
7.1	1. Учебные помещения и аудиторный фонд
7.2	Аудитория для проведения лекционных занятий.
7.3	Компьютерный класс (лаборатория) для проведения практических занятий и самостоятельной работы.
7.4	2. Программное обеспечение (ПО)
7.5	Google Colab (Google Research) — основная облачная платформа исполнения Jupyter-ноутбуков (.ipynb).
7.6	Специализированные библиотеки и пакеты анализа данных Python.
7.7	Офисные программы: Microsoft Office (Word, Excel) / LibreOffice / Яндекс

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

Методические указания для обучающихся по освоению дисциплины «Анализ данных»

1. Общие рекомендации по организации учебного процесса

Дисциплина «Анализ данных» предназначена для формирования у студентов бакалавриата направления 38.03.01 (РФ) / 580100 (КР) «Экономика» (профиль «Математические методы в экономике») системного понимания полного цикла работы с экономическими данными: от сбора сырых транзакций и первичной диагностики до обучения воспроизводимых моделей машинного обучения и формулирования обоснованных управленческих решений.

Учебный процесс построен на сочетании теоретических лекционных занятий, интерактивных лабораторных практикумов в среде Google Colab и самостоятельной научно-исследовательской работы. Комплексное освоение дисциплины требует от обучающегося последовательного перехода от усвоения базовых вычислительных парадигм к решению сквозных экономических кейсов.

Для успешного освоения курса обучающимся рекомендуется придерживаться следующей ритмичности:

- Еженедельно: изучать лекционный материал, разбирать демонстрационные скрипты, выполнять соответствующие лабораторные работы в Colab и проходить самопроверку по контрольным вопросам.
- Ежемесячно: проводить аудиторную и самостоятельную сшивку пройденных модулей, контролировать качество собственных аналитических скриптов и формировать разделы итогового практического проекта.

2. Методические указания по работе с лекционным материалом

Лекции по дисциплине «Анализ данных» являются методологической основой курса. Они задают теоретический вектор, раскрывают архитектуру вычислений и формируют аналитическое мышление.

2.1. Подготовка и работа на лекционных занятиях

- Предварительный обзор: перед лекционным занятием ознакомьтесь с аннотацией темы и ключевыми терминами предыдущей лекции.
- Активное конспектирование: записывайте не только математические формулы и конструкции кода Python, но и экономические интерпретации, примеры ограничений и аналитические пастки (например, разницу между корреляцией и причинно-следственной связью, эффект Парадокса Симпсона).
- Разбор архитектурных схем: уделяйте особое внимание схемам движения данных (пайплайнам), принципам управления памятью в NumPy (ndarray, view vs copy) и логике паттернов группировки (Split-Apply-Combine).

2.2. Самостоятельная проработка лекций

После каждой лекции необходимо:

1. Запустить и самостоятельно воспроизвести в Google Colab все кодовые примеры, показанные лектором.
2. Проверить понимание математических механизмов: самостоятельно расписать на бумаге правила выравнивания форм при автоматическом распространении массивов (broadcasting) или матричную логику перемножения осей (axis=0 и axis=1).
3. Ответить на контрольные вопросы для самопроверки, приведенные в конце каждого раздела учебно-методического комплекса.

3. Методические указания по выполнению лабораторных (практических) работ в Google Colab

Лабораторный практикум является ключевым элементом освоения дисциплины. Все практические работы выполняются в интерактивной облачной среде Google Colab на языке Python с использованием библиотеки pandas, вычислительного ядра NumPy, модулей визуализации Matplotlib/Seaborn, а также пакетов statsmodels и scikit-learn.

3.1. Пошаговый алгоритм выполнения практической работы

Каждая лабораторная работа выполняется в виде отдельного Jupyter-ноутбука (.ipynb) и включает следующие обязательные шаги:

1. Инициализация среды: создание нового ноутбука в Colab, импорт необходимых библиотек в первой кодовой ячейке:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Настройка отображения графиков
sns.set_theme(style="whitegrid", palette="deep")
```

2. Первичная загрузка и проверка данных:

- Загрузка данных из файлов (CSV, Excel, Parquet) с явным указанием кодировки (utf-8, cp1251), разделителей (sep=';') и типов данных.
- Выполнение «правила 5 минут» для диагностики: проверка размеров таблицы (df.shape), типов столбцов (df.info()),

наличия пропусков (df.isna().sum()), полных дубликатов (df.duplicated().sum()) и базовых описательных статистик (df.describe()).
3. Очистка и трансформация (Data Cleaning & Feature Engineering):
• Прецизионный парсинг дат через pd.to_datetime(..., format='%d.%m.%Y', errors='coerce') с контролем появления невалидных значений NaT.
• Унификация строковых полей с использованием аксессора .str: удаление краевых пробелов .str.strip(), приведение к единому регистру, замена повторяющихся пробельных символов через регулярные выражения replace(r'\s+', ' ', regex=True).
• Явная обработка пропусков и нормализация категорий на основе предметного смысла.
4. Векторизованный анализ и группировка:
• Расчет экономических показателей (выручка, средний чек, доли категорий) без использования циклов for.
• Выполнение группировок через groupby().agg() с именованной агрегацией или построение сводных таблиц pivot_table().
5. Визуализация и оформление ответа:
• Построение аналитических графиков с обязательным указанием названий осей, единиц измерения, понятных заголовков и аннотированием ключевых экстремумов.
6. Аналитический вывод (Markdown):
• Каждая работа завершается текстовым разделом «Итоговый аналитический вывод» объемом от 5 до 10 предложений с экономической интерпретацией полученных цифр.
3.2. Стандарты написания аналитического кода Python
При написании программного кода студенты должны строго соблюдать следующие профессиональные требования:
• Принцип векторизации: категориально запрещается использовать циклы for и while для поэлементной обработки массивов NumPy или столбцов pandas. Все математические операции, фильтрации и условия должны выполняться через векторизованные функции или np.select() / np.where().
• Контроль памяти и предупреждений (View vs Copy):
• При создании срезов отфильтрованных данных необходимо явно применять метод .copy(), чтобы избежать системного предупреждения SettingWithCopyWarning и предотвратить непреднамеренное изменение исходного DataFrame.
• Помните: обычный срез массива NumPy создает представление (view), изменение которого меняет исходную матрицу.
• Численная стабильность при работе с float:
• Категорически запрещается сравнивать вещественные числа через точное равенство a == b. Используйте функцию np.isclose(a, b) или np.allclose().
• Правильная агрегация:
• Различайте count() (считает непустые значения в столбце) и size() (считает общее физическое количество строк в группе).
• В сводных таблицах pivot_table() параметр fill_value=0 можно применять только тогда, когда подтверждено физическое отсутствие транзакций в данном сегменте.
4. Методические указания по организации самостоятельной работы и выполнению сквозного практического проекта
Самостоятельная работа студентов направлена на углубление практических навыков и выполнение индивидуального или группового Сквозного практического проекта (Итогового практикума).
4.1. Этапы выполнения сквозного проекта
Сквозной проект представляет собой исследование реального датасета (например, продажная аналитика, финансовые показатели компаний, отток клиентов) и включает 6 последовательных этапов:
1. Постановка проблемы и формулирование аналитических вопросов:
• Перевод размытой бизнес-проблемы (например, «Упала прибыль») в конкретную систему измеримых аналитических вопросов с указанием объекта, метрики, временных границ и проверяемых гипотез (H_0 и H_1).
2. Очистка данных и контроль качества (EDA QA):
• Составление отчета качества данных: выявление пропусков, дубликатов, аномалий и логических нестыковок.
3. Разведочный анализ данных и построение дашборда:
• Расчет описательных статистик (среднее, медиана, квантили, стандартное отклонение).
• Компоновка аналитического дашборда формата 2×2 из графиков Matplotlib/Seaborn (распределение, сравнение категорий, диаграмма рассеяния, динамический временной ряд).
4. Интеграция таблиц и реструктуризация:
• Объединение фактов и справочников через merge() с обязательным аудитом ключей (validate="many_to_one", indicator=True).
• Трансформация структуры данных между «широким» и «длинным» форматами (melt / pivot_table).
5. Построение ML-пайплайна (`scikit-learn Pipeline`):
• Масштабирование признаков (StandardScaler) и кодирование категорий внутри единого конвейера.
• Обучение регрессионной или классификационной модели, подбор гиперпараметров на кросс-валидации (GridSearchCV).
• Защита от утечки данных (Data Leakage): использование строго хронологического разделения выборок (chronological split) при наличии временных меток; запрет на масштабирование признаков до разбиения на train/test.
6. Защита проекта и подготовка аналитического отчета:
• Составление емкого отчета для руководства.
4.2. Формат и структура итогового аналитического отчета
Итоговый вывод проекта должен содержать от 7 до 10 предложений и строгую структуру:
1. Характеристика датасета: объем наблюдений, период и изучаемый объект.
2. Главный результат: ключевой драйвер или лидер по выручке/объему.
3. Региональный/сегментный срез: лидеры и отстающие группы с указанием долей.

4. Временная динамика: выявленные тренды, сезонные пики или спады.
5. Экономическая интерпретация: причины наблюдаемых изменений (разделяя факты и предположения).
6. Управленческие рекомендации: конкретные шаги для бизнеса и ограничения проведенного исследования.
5. Методические указания по подготовке к текущему контролю успеваемости и зачету/экзамену
5.1. Подготовка к текущему контролю
Текущий контроль включает защиту Colab-ноутбуков с лабораторными работами и прохождение промежуточного тестирования.
• Защита лабораторных работ: студент должен уметь устно объяснить каждую строку своего кода, обосновать выбор функций pandas/NumPy, пояснить смысл параметров (axis, how, validate, errors), а также прокомментировать графики.
• Подготовка к тестам: регулярное повторение контрольных вопросов самопроверки по 5 разделам дисциплины (всего 80 вопросов в фонде).
5.2. Подготовка к промежуточной аттестации (Зачет / Экзамен)
Промежуточная аттестация проводится в форме компьютерного тестирования и защиты сквозного практического проекта.
При подготовке к зачету/экзамену обучающимся рекомендуется:
1. Повторить теоретические концепции: границы применения описательных статистик (когда среднее искажается выбросами и требуется медиана), предпосылки OLS-регрессии, смысл p-value и ошибок I/II рода, правила интерпретации матриц ошибок (Confusion Matrix) и кривых ROC-AUC / PR-AUC.
2. Проверить владение синтаксисом основных функций и методов Python.
3. Провести аудит собственного практического проекта на соответствие чек-листу качества.