

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по выполнению лабораторных работ
по дисциплине «Big Data: основы анализа больших данных»
для студентов специальности 38.03.01 Экономика
направленность (профиль) «Бухгалтерский учет, анализ и аудит»

Ставрополь, 2026

Методические указания по дисциплине «Big Data: основы анализа больших данных» предназначены для выполнения лабораторных заданий по изучаемой дисциплине.

Выполнение заданий направлено на формирование у студентов знаний о процессе сбора, анализа и обработки больших данных с помощью современного программного обеспечения и аналитического инструментария с целью решения основных аналитических задач и задач прогнозной аналитики, моделирование ситуаций и решений в экономике.

Содержание

Цели и задачи изучения дисциплины	4
Компетенции обучающегося, формируемые в результате освоения дисциплины	4
Лабораторная работа по теме 1. Данные в экономике, их визуализация и предварительная обработка	5
Лабораторная работа по теме 2. Прикладной анализ данных	6
Лабораторная работа по теме 3. Генеральная и выборочная совокупности	7
Лабораторная работа по теме 4. Случайные величины	8
Лабораторная работа по теме 5. Анализ описательных статистик	9
Лабораторная работа по теме 6. Предельные теоремы теории вероятностей	11
Лабораторная работа по теме 7. Технология и инструменты анализа данных	12
Лабораторная работа по теме 8. Этапы анализа данных	15
Лабораторная работа по теме 9. Сравнение средних	18
Лабораторная работа по теме 10. Корреляционный анализ	20
Лабораторная работа по теме 11. Факторный анализ и метод главных компонент	21
Лабораторная работа по теме 12. Регрессионный анализ	23
Лабораторная работа по теме 13. Анализ временных рядов	25
Лабораторная работа по теме 14. Методы data mining	27
Лабораторная работа по теме 15. Анализ данных с применением машинного обучения	29
Лабораторная работа по теме 16. Методика прогнозирования, прогнозная аналитика	31
Список рекомендуемой литературы	32

Цели и задачи изучения дисциплины

Целью освоения дисциплины является формирование части профессиональных (ПК-1) компетенций будущего экономиста по направлению подготовки 38.03.01 Экономика направление (профиль) «Бухгалтерский учет, анализ и аудит».

Основными задачами изучения дисциплины «Big data: основы анализа больших данных» являются:

- формирование представлений о содержании, формах, особенностях дисциплины « - формирование представлений о содержании, формах, особенностях дисциплины «Big Data: основы анализа больших данных»;
- изучение основных методов систематизации, обработки и анализа статистических данных;
- изучение методов выявления и оценки степени зависимости между экономическими показателями;
- формирование навыков по применению технологий обработки данных для решения прикладных задач, связанных с экономикой.

В результате освоения дисциплины обучающийся приобретает следующие профессиональные компетенции:

ПК-1: способен осуществлять обработку статистической информации: сводку и группировку данных по утвержденным методикам, формировать системы взаимосвязанных статистических и аналитических показателей, анализировать деятельность организации.

Лабораторная работа по теме 1. Данные в экономике, их визуализация и предварительная обработка

Задание 1.

На основании представленной информации составьте таблицу, содержащую понятия и определения данных в экономике.

Данные — это информация в широком смысле, включающая в себя разнообразные факты и показатели, которые могут быть выражены как в числовой форме, так и в других видах представления. В англоязычной терминологии слово "data" часто используется как синоним к "information". Таким образом, термин "data system" и его производные могут иметь неоднозначные толкования, которые охватывают как «систему данных», так и «информационную систему».

Бит является минимальной единицей двоичной меры энтропии и количества информации в вычислительных системах, соответствующей одному двоичному разряду.

В процессе информационного взаимодействия происходит трансформация данных из одной формы в другую. С учетом научно-технического развития и увеличения сложности связей в социуме, трудозатраты на обработку информации постоянно растут, что связано с усложнением механизмов управления производственными и социальными процессами, ускорением темпов возникновения и внедрения новых носителей и хранилищ информации, а также ростом объемов данных.

Процесс обработки данных включает следующие этапы.

1. Сбор данных — аккумуляирование информации с целью обеспечения ее полноты для качественного принятия решений.
2. Формализация — стандартизация данных, полученных из различных источников, для их последующего сопоставления и улучшения доступности информации.
3. Фильтрация — исключение избыточных данных, не требующихся для принятия решений, обеспечивает сокращение информационного «шума» и повышение достоверности и адекватности данных.
4. Сортировка — систематизация данных согласно критерию упрощения процесса их использования и повышения эффективности доступа к информации.
5. Группировка - агрегирование данных по специфическому признаку для улучшения удобства обращения с информацией.
6. Архивация данных — организация хранения данных в формате, обеспечивающем легкий доступ и минимизацию затрат на хранение.
7. Защита данных — комплекс мероприятий по обеспечению безопасности данных, направленных на предотвращение их утраты, несанкционированного доступа и изменений.
8. Транспортировка данных — обеспечение передачи данных между удаленными участниками информационного процесса, в том числе между серверами (источниками данных) и клиентами (потребителями данных).
9. Преобразование данных — изменение формата данных для обеспечения их совместимости с различными техническими платформами и носителями, что особенно актуально в условиях многократной передачи и использования данных в различных информационных системах.

Задание 2.

Визуализируйте табличные данные по индексам цен на рынке коммерческой недвижимости с использованием программы Microsoft Excel, сделайте вывод о полученной тенденции:

Таблица – Данные по индексам ценам

Год	Индекс, %	Год	Индекс, %
2010	156,9	2016	118,5
2011	146,3	2017	117,5
2012	113,1	2018	147,7
2013	125,1	2019	123,4
2014	122,5	2020	110,3
2015	118,8	2021	92,4

Ход выполнения задания:

На новой вкладке Microsoft Excel создадим новый лист куда перенесем данные задания. Проверим данные на соответствие единому формату и разрядности при помощи функции «Формат ячеек» -- «число» -- «выравнивание». Визуализируем данные применив к ним функцию «вставка - - график - - график с маркерами».

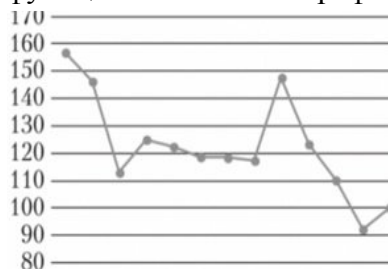


Рисунок – Визуализация данных задания

Для того чтобы сделать вывод о тенденции данных необходимо на правой кнопке мыши выбрать функцию «добавить тренд». Появившаяся пунктирная линия будет указывать на тренд, который имеется в данных.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме 2. Прикладной анализ данных

Задание 1.

Требуется создать облако для хранения файлов с большими данными, создать персональный кабинет для управления и произвести индивидуальные настройки для того чтобы хранить любую информацию. В каждом аккаунте Google по умолчанию доступно 15 ГБ для хранения данных. Это пространство используется сервисами "Google Диск", "Google Фото" и Gmail.

Шаг 1. Откройте страницу drive.google.com

Откройте сайт drive.google.com на компьютере. В разделе "Мой диск" вы найдете: загруженные или синхронизированные файлы и папки; созданные вами документы, таблицы, презентации и формы Google.

Шаг 2. Загрузите или создайте файлы Вы можете загрузить файлы с компьютера или создать их на Google Диске. Google Диск поддерживает следующие типы файлов:

документы; изображения аудио; видео. Загрузка файлов и папок. На компьютере загружать файлы в личные и общие папки можно на странице drive.google.com, а также в приложении "Google Диск".

Для удаления файлов с большим объемом данных в веб-интерфейсе отсортируйте файлы по размеру (от большего к меньшему). Переместите ненужные файлы в корзину, а затем удалите их окончательно.

Задание 2.

Используя произвольный набор данных на портале data.gov.ru импортируйте данные и сохраните их. На основе набора данных подготовьте выборку, содержащую значения выбранного показателя за 10 лет по РФ.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме 3: Генеральная и выборочная совокупности

Задание 1.

Вычислите выборочные характеристики по данным наблюдения:

Таблица – Исходные данные

№	1	2	3	4	5
X	2	6	10	14	18

Исходные данные и расчетные показатели представлены в виде двух расчетных таблиц (два способа расчета $\bar{x}$, $\text{var}(x)$).

Ход решения:

Найти выборочные характеристики можно несколькими способами:

1. Внесите данные в Excel и воспользуйтесь простыми расчётными функциями Excel задав формулу через знак «=» и выбрав соответствующие ячейки. Для нахождения суммы чисел в столбцах можно воспользоваться функцией СУММ. Для нахождения квадрата числа необходимо задать формулу в ячейке: $= (B2-B3)^2$ (номера ячеек приведены в качестве примера). Приведем промежуточные результаты расчетов.

Таблица – Расчетные данные

№	x	$(x - \bar{x})^2$
1	2	64
2	6	16
3	10	0
4	14	16
5	18	64
Итого	50	160
Среднее	10	32
	$\bar{x}$	$\text{var}(x)$
$\text{var}(x) = 160/5 = 32$		

№	x	x^2
1	2	4
2	6	36
3	10	100
4	14	196
5	18	324
Итого	50	660
Среднее	10	132
	$\bar{x}$	$\bar{x}^2$
$\text{var}(x) = \bar{x}^2 - (\bar{x})^2 = 132 - 100 = 32$		

2. Для вычисления выборочной средней и выборочной дисперсии в Excel имеются функции: $\bar{x} = \text{СРЗНАЧ}(\text{массив})$, $\text{var}(x) = \text{ДИСПР}(\text{массив})$

Задание 2.

Объем генеральной совокупности составляет 10 000 человек, а выборочной – 500 человек.

Сформируйте механическую выборку.

Ход решения:

$10\,000 / 500 = 20$. Таким образом, отбирается каждый 20 человек. После этого необходимо определить начальную точку отбора, которая выбирается случайным образом. Допустим, первым элементом выборки стал респондент под номером 651, тогда следующие будут: 671, 691, 711 и далее, но не будет сформирована выборка из 500 человек.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме 4: Случайные величины

Задание 1.

Найдите вероятности при помощи биномиального распределения.

Для поступления в магистратуру студенту необходимо выполнить 50 тестовых заданий, каждое из которых имеет четыре варианта ответа при одном правильном.

Ход решения:

Предположим, что студент не подготовился к вступительному экзамену, поэтому может только угадывать правильный ответ. Выбор наугад вариантов ответа дает 25%-ную вероятность выбора правильного ответа на каждый вопрос. Какова вероятность, что студент правильно угадает ответы на 35 тестовых вопросов из 50?

Обратимся к биномиальному распределению:

с $n=50$, $k=35$, $p=0,25$:

$$\begin{aligned} P(X=35; 50; 0,25) &= \frac{50!}{35!(50-35)!} \cdot 0,25^{35} \cdot (1-0,25)^{50-35} = \\ &= 2250829575120 \cdot 1,1319291764952112841754989892486E-23 = \\ &= 2,548E-11 = 0,00000000002548. \end{aligned}$$

Как видно из вычисленных значений, вероятность правильно угадать ответы на 35 вопросов из 50 ничтожно мала.

Задание 2. Теперь тестовых вопросов будет 10, каждый из которых будет иметь пять вариантов ответа (20%-ная вероятность). Какова вероятность, что студент правильно угадает ответы на 7 тестовых вопросов из 10?

Ход решения:

$$\begin{aligned} P(X=7; 10; 0,2) &= \frac{10!}{7!(10-7)!} \cdot 0,2^7 \cdot (1-0,2)^{10-7} = \\ &= 3628800/30240 \cdot 0,0000128 \cdot 0,512 = 0,000786432. \end{aligned}$$

Вероятность правильно угадать ответы на 7 вопросов из 10 и в данном случае крайне низка.

Такой же результат можно получить с помощью таблицы биномиального распределения.

Таблицы строятся для различных комбинаций n , k , p , что делает их довольно объемными. Поэтому приведем таблицу биномиального распределения вероятностей только для $n=10$. Поскольку $k=7$, а $p=0,2$, то на пересечении соответствующих строки и столбца находим искомое значение вероятности, равное 0,000786.

Вероятности для биномиального распределения ($n = 10$)

k	$p = 0,05$	$p = 0,1$	$p = 0,15$	$p = 0,2$	$p = 0,25$	$p = 0,3$	$p = 0,35$	$p = 0,4$	$p = 0,45$	$p = 0,5$
0	0,598737	0,348678	0,196874	0,107374	0,056314	0,028248	0,013463	0,006047	0,002533	0,000977
1	0,315125	0,38742	0,347425	0,268435	0,187712	0,121061	0,072492	0,040311	0,020724	0,009766
2	0,074635	0,19371	0,275897	0,30199	0,281568	0,233474	0,175653	0,120932	0,076303	0,043945
3	0,010475	0,057396	0,129834	0,201327	0,250282	0,266828	0,25222	0,214991	0,166478	0,117188
4	0,000965	0,01116	0,040096	0,08808	0,145998	0,200121	0,237668	0,250823	0,238367	0,205078
5	6,1E-5	0,001488	0,008491	0,026424	0,058399	0,102919	0,15357	0,200658	0,234033	0,246094
6	3,0E-6	0,000138	0,001249	0,005505	0,016222	0,036757	0,06891	0,111477	0,159568	0,205078
7	0	9,0E-6	0,000126	0,000786	0,00309	0,009002	0,021203	0,042467	0,074603	0,117188
8	0	0	8,0E-6	7,4E-5	0,000386	0,001447	0,004281	0,010617	0,02289	0,043945
9	0	0	0	4,0E-6	2,9E-5	0,000138	0,000512	0,001573	0,004162	0,009766
10	0	0	0	0	1,0E-6	6,0E-6	2,8E-5	0,000105	0,000341	0,000977

Задание 3.

Используя базу Росстата загрузите в пакет Excel выбранные данные. Чтобы загрузить можно выбрать функцию Файл---Открыть в главном окне программы Excel, а затем выбрать нужный файл. В этом пакете для построения описательных статистик необходимо воспользоваться встроенными функциями (СЧЕТ, МИН, МАКС, ДИСП, СРЗНАЧ и т.д.).

В пакете Excel на вкладке Вставка в разделе Диаграммы есть класс Гистограмма, однако он визуализирует вовсе не распределение переменной, а отдельные ее значения в виде столбцов, откладывая по оси абсцисс номер наблюдения. Чтобы построить настоящую гистограмму, необходимо воспользоваться надстройкой «Пакет анализа». Нажмите на кнопку «Анализ данных» на вкладке Данные и выберите пункт Гистограмма. В появившемся окне в поле «Входной интервал» укажите столбец необходимой переменной. Если вы выделили столбец вместе с первой строкой, содержащей имя переменной, то поставьте галочку напротив «Метки». По умолчанию результаты разбиения сохраняются на отдельном листе. Выделите два полученных столбца и только теперь выберите пункт Вставка — Диаграммы — Гистограмма.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме 5. Анализ описательных статистик

Задание 1.

Интервальный ряд роста студентов задан в таблице

Группа i	Рост см	Частота f	Накопленная частота f_i'
1	160-165	2	
2	165-170	4	
3	170-175	9	
4	175-180	6	

Ход решения:

Формула медианы: $Me = x_0 = \frac{\frac{n}{2} - f_{me-1}}{f_{me}} * i$

1. Номер медианного студента равен $(n+1)/2 = (21+1)/2 = 11$

2. Медианная группа – 3-я, так как для нее выполняется неравенство: $6 < 11 < 15$. Нижняя граница медианного интервала равна 170 см, его длина – 5 см.

$$3. \quad \text{Медиана равна: } Me = 170 + \frac{\frac{21}{2} - 6}{9} * 5 = 172,5$$

Задание 2.

Интервальный ряд роста студентов задан в таблице

Группа i	Рост см	Частота f	Накопленная частота f _i '
1	160-165	2	2
2	165-170	4	6
3	170-175	9	15
4	175-180	6	21

Вычислите следующие показатели: квартиль, модальный интервал, соберите данные показатели в Excel

Ход решения:

Квартиль делит совокупность на четыре равные части. Алгоритм расчеты квартиля и медианы аналогичны.

Для решения данной задачи определим квартильный интервал. Для его вычисления должно выполняться условие (накопленная частота) $2 < 21/4 < 6$. В нашем случае, квартиль находится во втором интервале (165-170).

$$Q = x_1 + \frac{\frac{n}{4} - f_{q1-1}}{f_{q1}} * i,$$

где x_1 – нижняя граница квартильного интервала, f_{q1} – частота в данном интервале;

f_{q1-1} – накопленная частота в предыдущем интервале, i – длина квартильного интервала.

$$Q = 165 + \frac{\frac{21}{4} - 2}{4} * 5 = 169,1$$

Мода – значение, которое встречается наиболее часто. Модальный интервал, интервал с наибольшей частотой.

Для интервального ряда используется следующий алгоритм расчета моды:

$$Mo = x_0 + \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} * i, \text{ где}$$

$f_{mo}, f_{mo-1}, f_{mo+1}$ – частота в модальном, предыдущем и последующем интервалах соответственно.

Для нашего примера:

$$Mo = 170 + \frac{9-4}{(9-4)+(9-6)} * 5 = 173,12$$

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 6. Предельные теоремы теории вероятностей.

Задание 1.

По выборочным данным о расходах сырья при производстве продукции по старой и новой технологиям получены выборочные исправленные дисперсии $S_1^2 = 2,192$, $n_1 = 13$
 $S_2^2 = 1,611$, $n_2 = 9$

Можно ли при уровне значимости 0,05 считать статистически значимыми различие между исправленными дисперсиями?

Ход решения:

Расчетное значение критерия $-F_p = \frac{S_1^2}{S_2^2} = \frac{2,192}{1,611} = 1,36$

Число степеней свободы $v_1 = n_1 - 1 = 12$, $v_2 = n_2 - 1 = 8$.

Определим в Excel Критическую точку используя формулу

$$F_{кр} = F_{РАСПОБ}(0,05; 12; 8) = 3,28$$

Поскольку $F_p = 1,36 < F_{кр} = 3,28$, то H_0 принимается, т.е. выборочное исправление дисперсии не различаются

Задание 2.

Сгенерируйте исходные данные для анализа с помощью имитационного моделирования.

Исходные данные для генерации данных:

Количество товаров в каждой категории – 3;

В каждой категории по 2 вида товара;

Количество исследуемых городов - №;

В каждом городе по 3 магазина;

Период моделирования – 4 года, начало периода 01.01.2020

Сгенерируйте базу данных, состоящую из 300 строк.

Ход решения:

При выполнении работы для имитационного моделирования используется генератор случайных чисел, который расположен в статистической надстройке «Анализ данных».

После подключения данного пакета в разделе Analysis.

После выполнения операций, сделайте зарисовку таблицы транзакций в соответствии с описанием – оформите шапку таблицы и заполните произвольными данными строки.

Задание 3.

Найдите вероятность $P(X < 2,0)$ используя таблицы стандартного распределения.

Ход решения:

Находим строку со значением 2,0 и столбец со значением 0,0. На пересечении соответствующей строки и столбца расположено искомое значение - 0,4772. Прибавим к этому значению 0,5 и получим $P(X < 2,0) = 0,9772$. Это означает, что площадь под кривой нормального распределения от минус бесконечности до значения 2,0 составляет 0,9772, то есть вероятность того, что случайная величина будет иметь значение, меньшее 2,0.

Найдем вероятность $P(X > 2,0)$. Для этого надо найти площадь под кривой для значения меньше 2,0 и вычесть ее из общей площади под кривой, которая равна 1. Площадь под кривой левее 2,0 содержит площадь от минус бесконечности до нуля и от нуля до 2,0. Если первая известна (она равна 0,5 как половина общей площади, которая составляет 1), то вторую находим по таблице на пересечении строки (2,0) и столбца

$P(X \leq 2,0) = 0,50000 + 0,4772 = 0,9772$;

$P(X > 2,0) = 1 - 0,9772 = 0,0228$.

0,0228 площади под кривой находится выше значения 2,0. Вероятность значений больше 2,0 составляет всего 0,023.

Изменим условия и определим долю значений, которая лежит левее точки -2,0, т. е. вероятность $P(X < -2,0)$. Поскольку стандартное нормальное распределение симметрично, то площадь между -2,0 и 0 равна 0,4772, так же как и площадь между 0 и 2,0. Площадь под кривой левее точки -2,0 вычисляется следующим образом: $0,50000 - 0,4772 = 0,0228$, что соответствует площади под кривой правее точки 2,0. Итак, площадь под кривой нормального распределения от минус бесконечности до точки -2,0 составляет 0,0228, т. е. это вероятность того, что случайная величина примет значение, меньшее -2,0. Доля значений, которые лежат левее точки - 2,0, не превышает 2,3 %.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 7. Технология и инструменты анализа данных.

Задание 1.

Заполните таблицу на основании изученных материалов.

Традиционная аналитика	Big Data аналитика
Постепенный анализ небольших пакетов данных	Обработка сразу всего массива доступных данных
Редакция и сортировка данных перед обработкой	Данные обрабатываются в их исходном виде
Старт с гипотезы и ее тестирования относительно данных	Поиск корреляций по всем данным до получения искомой информации
Данные собираются, обрабатываются, хранятся и лишь затем анализируются	Анализ и обработка больших данных в реальном времени, по мере поступления

Задание 2.

В рамках выполнения плана по продажам банковских услуг корпоративным клиентам отделом клиентских менеджеров Банка были представлены результаты работы по итогам 3 квартала 2024 года (таблица). Визуализируйте полученные результаты и представьте их в виде Диаграммы Воронки.

Таблица - Исходные данные

Этапы	Сумма, млн. рублей
Перспективы	929
Подтвержденные перспективы	605
Предложения	450
Переговоры	384
Конечные продажи	341

Воронкообразная диаграмма используется для визуализации постепенного сокращения данных при их переходе от одной фазы к другой в процессе.



Рисунок – Воронкообразная диаграмма

Воронкообразная диаграмма используется для визуализации прогрессивного сокращения данных при их переходе от одной фазы к другой. Данные на каждом из этих этапов представлены в виде различных частей на 100% (всего). Как и круговая диаграмма, воронкообразная диаграмма также не использует никаких осей. Например, в клиентском отделе Банка будут этапы, как показано ниже:

Перспективы → Подтвержденные перспективы → Предложения → Переговоры → Конечные продажи.

Как правило, значения уменьшаются постепенно. Многие перспективы определены, но часть из них проверены и еще меньше претендуют на предложения. Еще меньшее число приходит на переговоры, и, в конце концов, есть только несколько сделок, которые выиграны.

Решение:

-Используя MS Excel создадим исходную таблицу. Вставим столбец «Дополнительные расчеты» таблицу и внесем соответствующие формулы.

Этап	Дополнительные расчеты	Сумма, млн. руб.
Перспективы	0	929
Подтвержденные перспективы	$=(\text{НАИБОЛЬШИЙ}(\text{\$D\$3:\$D\$7;1})-\text{D4})/2$	605
Предложения	$=(\text{НАИБОЛЬШИЙ}(\text{\$D\$3:\$D\$7;1})-\text{D5})/2$	450
Переговоры	$=(\text{НАИБОЛЬШИЙ}(\text{\$D\$3:\$D\$7;1})-\text{D6})/2$	384
Конечные продажи	$=(\text{НАИБОЛЬШИЙ}(\text{\$D\$3:\$D\$7;1})-\text{D7})/2$	341

Рисунок – Расчетные формулы

-Получены следующие данные, которые будут использованы для построения диаграммы.

Этап	Дополнительные расчеты	Сумма, млн. руб.
Перспективы	0	929
Подтвержденные перспективы	162	605
Предложения	239,5	450
Переговоры	272,5	384
Конечные продажи	294	341

Рисунок – Расчетные значения

-Выберите данные и вставьте в столбчатую диаграмму с накоплением Вставка – Гистограмма с накоплением.

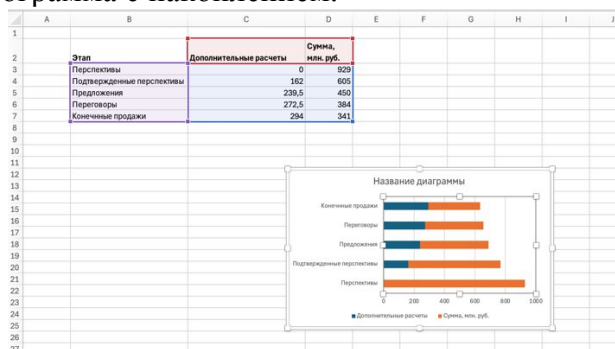


Рисунок – Гистограмма с накоплением

-Выделите на диаграмме данные из столбца дополнительные расчеты и измените цвет синий на «нет заливки» .

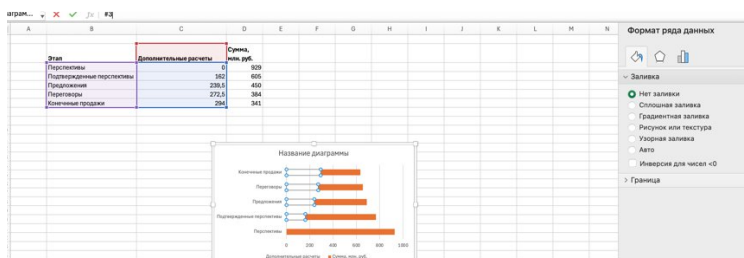


Рисунок – Форматирование диаграммы

Завершим создание диаграммы.

-Необходимо щелкнуть правой кнопкой по вертикальной оси диаграммы - -формат оси - - выбрать обратный порядок категорий.

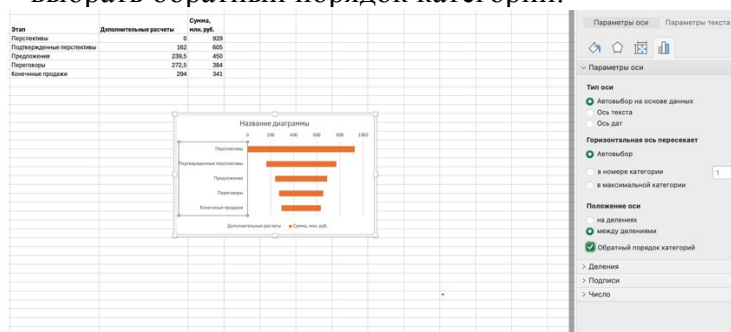


Рисунок – Форматирование диаграммы

-изменить заголовок диаграммы
 -правой кнопкой щелкнуть по диаграмме - - формат ряда данных - - параметры ряда -
 - боковой надзор - - 0%.
 -режим «Конструктор» - -добавить элемент диаграммы - - подписи данных - - по центру.



Рисунок – Диаграмма Воронка.

Задание 3

Соотнесите задачи и функции Big Data

Функция	Задача
Big Data – собственно массивы необработанных данных	Хранение и управление большими объемами постоянно обновляющейся информации
Data mining – процесс обработки и структуризации данных, этап аналитики для выявления закономерностей	Структурирование разнообразных сведений, поиск скрытых и неочевидных связей для приведения к единому знаменателю
Machine learning – процесс машинного обучения на основе обнаруженных связей в процессе анализа	Аналитика и прогнозирование на основе обработанной и структурированной информации

Задание 4.

Подготовьте эссе на одну из предложенных тем:

1. Информация как ресурс и движущая сила развития человеческого общества.
2. Big Data: сущность, настоящее и будущее
3. История развития больших данных с 1970 г по 2024 г
4. Мировой рынок Big Data
5. Влияние Big Data на современное общество

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 8. Этапы анализа данных

Задание 1.

Банк проводит выборочное исследование (опрос) в ряде компаний, бюджетных организаций. Создайте простую веб-форму для опроса, используя платформу Google Forms. Для работы с данной платформой потребуется войти в свой аккаунт Google (если его нет, необходимо создать, в дальнейшем это Вам позволит пользоваться большинством сервисов Google).

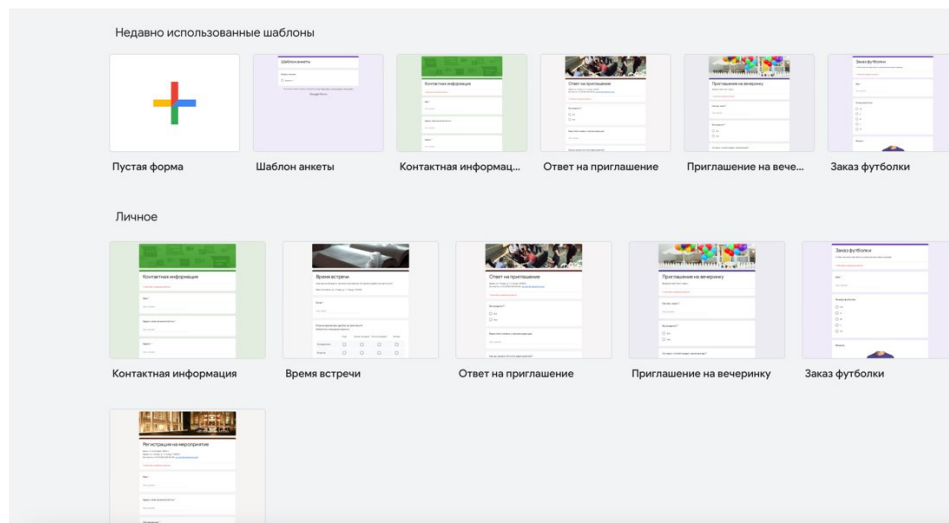
1) Создание аккаунта Google.

- Выберите тип аккаунта: для себя или для управления бизнесом (при выполнении данного задания выбираем - для себя -);
- Откройте страницу входа в аккаунт Google <https://myaccount.google.com/> - Нажмите - - создать аккаунт - - в раскрывающемся списке выберите, зачем вы создаете аккаунт. Для: личного использования; ребенка; организации;
- Введите свою ФИО и дату рождения;
- Укажите имя пользователя в соответствующем поле;
- Введите пароль и подтвердите его;
- Нажмите далее - - добавьте и подтвердите номер телефона (необязательно) - - нажмите далее.

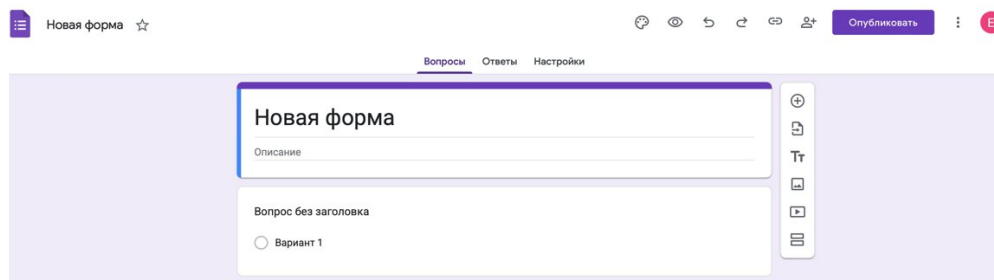
2) Перейдите на страницу Google Forms: <https://docs.google.com/forms/>

- Рассмотрите галерею предлагаемых шаблонов (в будущем Вы сможете выбирать наиболее подходящий вариант под рабочую задачу)

← Галерея шаблонов



3) Выбираем «Пустая форма», откроется окно «Новая форма»

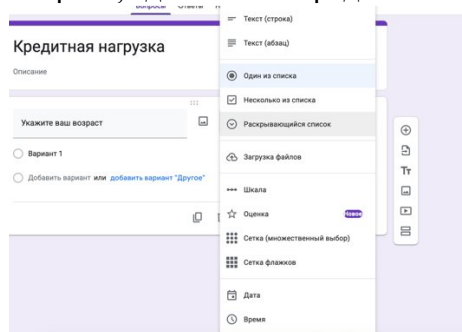


-Переименуем нашу форму «Кредитная нагрузка», это можно сделать если поставить курсор на строку «Новая форма» и начать печатать. В случае ошибок можно отменить действие, нажав соответствующую стрелку (аналогично работе в word)

4)Зададим в веб-форме следующие вопросы, которые позволят провести анализ кредитной нагрузки физических лиц, сделать выводы и разработать мероприятия, которые позволят финансовой организации сократить риски, нарастить клиентскую базу и пр.:

1. Сколько Вам лет?
2. Есть ли у вас постоянное место работы?
3. Количество работающих членов семьи?
4. Количество детей, их возраст?
- 5.Среднемесячный доход?
6. Количество имеющихся кредитов?
7. Ежемесячные платежи по кредитам?

-Ответы на вопросы, которые мы хотим получить, можно запросить в разной форме. Для этого надо задать параметры желаемого ответа. Это можно сделать справа от основного вопроса, где шаблон предлагает список:



В нашем случае ответ на вопрос будет предполагать краткий ответ , поэтому выбираем параметр - -текст (строка) – Можно предусмотреть варианты ответов на вопрос, тогда необходимо выбрать вкладку - - несколько из списка—и задать варианты. Поставим курсор на отметку - - обязательный вопрос- - , для того чтобы ответ на него обязательно присутствовал.

Шаг 5.

-Для создания следующего вопроса выберем справа «+» добавить вопрос, форма создаст следующий вопрос и т.д. На панели справа есть настройки, которыми можно управлять при формировании вопроса. Также можно использовать функционал Google Forms, который позволяет: настроить тему, увидеть количество полученных ответов, настроить предпросмотр, удалить шаблон, создать копию и пр.

-Далее необходимо открыть доступ к опросу, это можно сделать во вкладке «открыть доступ». В данной вкладке надо задать соответствующие параметры:

-Общий доступ: редакторы - - доступ ограничен- - (в данном случае пользователи не смогут редактировать предлагаемые вопросы); респонденты - - все у кого есть ссылка - - (в данном случае пользователь смогут добавить ответы на поставленные вопросы).

-Чтобы начать получать ответы веб-форму необходимо опубликовать. Для этого нажмите соответствующую кнопку «опубликовать». В параметрах опубликованной формы не забудьте указать «принимать ответы» в противном случае респонденты не смогут дать на них ответ. Ссылку на веб-форму для получения ответов можно направлять участникам опроса для этого необходимо нажать на значок «копировать ссылку для заполнения формы»

-После получения ответов в исследуемой выборке, собранные данные можно импортировать в excel для дальнейшего анализа. Для этого необходимо установить связь с таблицами Excel - - новая таблица/существующая таблица – (выберите куда будут выгружаться полученные ответы).

Задание 2.

Установите SQL Server для использования в качестве базы данных, данных для разработки и анализа в рабочей среде.

Ход решения:

Базы данных и язык SQL (Structured Query Language) играют важную роль в современном мире информационных технологий. Большинство приложений, которые мы используем ежедневно, хранят данные в базах данных, и SQL используется для доступа, извлечения и обновления этих данных.

Шаг 1: Выбор базы данных и установка СУБД

Система управления базами данных (СУБД) - это программное обеспечение, позволяющее создавать базы данных, управлять ими и взаимодействовать с ними.

Существует много различных СУБД, таких как MySQL, PostgreSQL, Oracle, Microsoft SQL Server и многие другие. Каждая из них имеет свои преимущества и недостатки.

Доступно несколько популярных СУБД, в том числе:

1. MySQL. Широко используемая система управления реляционными базами данных с открытым исходным кодом.
2. PostgreSQL. Еще одна популярная система управления реляционными базами данных с открытым исходным кодом. PostgreSQL известна своей надежностью и производительностью.
3. SQLite: легкая, бессерверная и автономная система управления базами данных, SQLite идеально подходит для небольших проектов и учебных целей.

После выбора СУБД необходимо установить ее на компьютер.

Для MySQL можно использовать официальный сайт для загрузки дистрибутива (<https://dev.mysql.com/downloads/mysql/>) и следовать инструкциям по установке.

Шаг 2: Установите клиент SQL

Клиент SQL - это инструмент, который позволяет взаимодействовать с базой данных и выполнять SQL-запросы. Существует множество клиентов для работы с базами данных, но наиболее популярными являются:

1. MySQL Workbench - <https://dev.mysql.com/downloads/workbench/>

MySQL Workbench - это графический клиент для работы с сервером MySQL.

2. Microsoft SQL Server Management Studio

<https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15>

Microsoft SQL Server Management Studio – это графический клиент для работы с сервером Microsoft SQL Server.

3. Oracle SQL

Developer - <https://www.oracle.com/tools/downloads/sqldev-downloads.html>

Oracle SQL Developer - это графический клиент для работы с сервером Oracle.

4. pgAdmin - [https://](https://www.pgadmin.org/download/)

www.pgadmin.org/download/ pgAdmin - это графический клиент для работы с сервером PostgreSQL.

5. HeidiSQL - <https://www.heidisql.com/download.php>

HeidiSQL - это бесплатный графический клиент для работы с базами данных, который поддерживает технологию SQL, а именно

MySQL, Microsoft SQL Server, PostgreSQL и SQLite.

6. DBeaver - <https://dbeaver.io/download/>

DBeaver - это бесплатный универсальный графический клиент для работы с базами данных, который поддерживает множество СУБД, включая MySQL, DretaraShl Mrarla Mirrenft

Шаг 3: Работа с базой данных

1. Создание базы данных

После установки СУБД нужно создать базу данных. Это можно сделать с помощью команды SQL CREATE DATABASE:

- код sql:

```
CREATE DATABASE mydatabase;
```

"mydatabase". Это будет пустая база данных без таблиц.

2. Создание таблицы

После создания базы данных нужно создать таблицу.

Таблица - это структурированный способ хранения данных в базе данных. Можно создать таблицу с помощью команды SQL CREATE TABLE:

- код sql

```
CREATE TABLE
```

```
customers (id INT PRIMARY KEY, name
```

```
VARCHAR(50), email VARCHAR(50) );
```

Создадим таблицу "customers" с тремя столбцами: "id", "name" и "email". Столбец "id" имеет тип данных "INT" и используется как первичный ключ (PRIMARY KEY) для таблицы. Столбцы

"name" и "email" имеют тип данных "VARCHAR(50)" и хранят текстовые значения до 50 символов.

3. Добавление данных в таблицу

После создания таблицы нужно добавить данные в нее.

Это можно сделать с помощью команды SQL

```
INSERT INTO.
```

-- код Sql

```
INSERT INTO customers (id, name, email) VALUES (1, John Smith',  
'john.smith@example.com'), (2, 'Jane Doe', 'jane.doe@example.com'),(3, 'Bob Johnson',  
'bob.johnson@example.com');
```

Добавим три записи в таблицу:

"customers" с помощью одной команды. Каждая запись имеет значение для столбцов "id", "name" и "email".

4. Извлечение данных из таблицы

После добавления данных в таблицу можно извлечь их с помощью команды SQL SELECT:

- код Sql SELECT * FROM customers;

Здесь мы выбираем все столбцы и все записи из таблицы "customers"

Результатом будет таблица со всеми добавленными ранее записями.

5. Обновление и удаление данных из таблицы

Чтобы обновить данные в таблице, нужно использовать команду SQL UPDATE:

- код sql

```
UPDATE customers SET name = 'John Doe'
```

```
WHERE id = 1;
```

Здесь мы обновляем имя клиента с id = 1 на "John Doe".

Чтобы удалить данные из таблицы, нужно использовать команду SQL DELETE.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме 9. Сравнение средних

Задание 1.

Имеются данные по удельному весу вредных выбросов в окружающую среду (тыс. т) по трем регионам РФ.

Регион	2017	2018	2019	2020	2021	2022
Воронежская область	3,5	3,9	4,0	4,2	4,9	
Ростовская область	2,9	3,6	5,1	4,2	3,5	3,1
Краснодарский край	3,9	3,5	3,7	5,2		

Проверим следующие гипотезы при $\alpha = 0,05$:

Но (нулевая гипотеза): регион не влияет на объем выбросов в окружающую среду (или групповые средние равны общей средней)

А (альтернативная гипотеза): регион влияет на объем выбросов (или групповые средние не равны общей средней)

Ход решения:

1. Найдем внутригрупповые средние по регионам. Для этого воспользуемся функцией =СРЗНАЧ

Полученные значения представим в виде рабочей таблицы

Средние	Значение
x1 (воронеж)	4,10
x2 (ростов)	3,73
x3 (краснодар)	4,08
Общая средняя	3,95

2. Общую среднюю вычислим используя функцию =СРЗНАЧ, только задав диапазон исследуемых значений *Образец формулы:* =СРЗНАЧ(С3:G3;C4:H4;C5:F5)

3. Найдем полную сумму квадратов отклонений по формуле: $SS = \sum (x - \bar{x}_{\text{общ}})^2$ для этого найдем квадраты отклонений и просуммируем их. *Образец формулы:* =СУММ(C12:G12;C13:H13;C14:F14)

Полная сумма квадратов отклонений для наших значений равна 6,537

4. Найдем внутригрупповую сумму квадратов отклонений по формуле:

$$\sum SS_w = \sum (x - \bar{x}_{\text{вн гр}})^2$$

из каждой варианты вычтем найденную внутригрупповую среднюю. Далее просуммируем найденные значения.

Внутригрупповая сумма квадратов отклонений для наших вариантов равна 6,08

5. Из полученных промежуточных значений найдем межгрупповую сумму квадратов отклонений по формуле: $SS_b = \sum n_i (\bar{x}_{\text{вн гр}} - \bar{x}_{\text{общ}})^2$ Для нашего задания $SS_b = 0,46$

6. Сделаем проверку полученных показателей. Для этого суммируем значения межгрупповой и внутригрупповой суммы квадратов ($SS_b + SS_w = SS$)

$$6,09 + 0,46 = 6,54$$

7. Вычислим число степеней свободы.

-полное число степеней свободы равно $(n-1)$, где n – общее количество исследуемых показателей (вариантов) $= (5+6+4)-1 = 14$

-межгрупповое число степеней свободы $(m-1)$, где m – число групп $= (3-1) = 2$

-внутригрупповое число степеней свободы $(n-m) = 15-3 = 12$

8. Рассчитаем средний квадрат между группами $MS_{\text{МГ}} = \frac{SS_b}{m-1} = 0,22$

9. Рассчитаем средний квадрат внутри групп $MS_{\text{ВГ}} = \frac{SS_w}{n-m} = 0,51$

10. Найдем значение F расчетное $= \frac{MS_{\text{МГ}}}{MS_{\text{ВГ}}} = 0,45$

11. Найдем значение F критич используя формулу Excel =FРАСПОБР с ,05 либо таблицу Фишера. Для наших значений (0,05; 2; 12)=3,89
 12. Сравним полученные значения: так как F расчетное F критич, то гипотеза H_0 принимается, то есть регион не влияет на объем выбросов в окружающую среду.
 13. Решим данную задачу вторым способом используя пакет «Анализ данных» - «Дисперсионный факторный анализ»
- Обратим внимание, на параметр «группирование» оно у нас выполнено по строкам, а параметр «метки» по первому столбцу.
14. Сравним полученные значения, сделаем выводы и оформим работу.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 10. Корреляционный анализ

Задача 1.

Получены данные о спросе на табачные изделия и доходе населения страны. Проанализируйте взаимосвязи между имеющимися данными.

Год	Спрос	Индекс цен, %	Уровень доходов
2014	38,7	105,1	6900
2015	33,6	125,4	6725
2016	34,4	122,5	7020
2017	37,7	118	7150
2018	39,5	115,2	7375
2019	42,7	110,4	7800
2010	40,1	114,2	8200
2011	41,1	113,8	8700

Ход решения:

1. Рассчитаем ковариацию между спросом и индексом цен с использованием функционала Excel по формуле =КОВАРИАЦИЯ.Г, где Массив1 – показатели спроса, Массив 2 – показатели Индекса цен. Полученный результат -0,137.
2. Также, для оценки степени взаимосвязи двух диапазонов показателей рассчитаем коэффициент корреляции по формуле $\frac{cov_{xy}}{\sqrt{SS_x SS_y}}$, где SS_x – дисперсия по показателю спроса, SS_y – дисперсия по показателю индекс цен.

С использованием функционала Excel данную формулу можно представить в виде:

$$\frac{(=КОВАРИАЦИЯ.Г)}{КОРЕНЬ(ДИСП.Г(массив спрос)*ДИСП.Г(массив индекс))} = -0,76 (K1)$$

Из формулы видно, что между ценам и спросом присутствует довольно сильная обратная связь, то есть с ростом цен спрос снижается.

3. Подтвердить данный вывод можно с помощью расчета коэффициента детерминации, который рассчитывается путем возведения коэффициента корреляции в квадрат. Также можно воспользоваться функцией Excel =СТЕПЕНЬ (значение коэффициента корреляции; 2).

Задача 2

Установите соответствие между наиболее распространенными подходами обработки данных (ПО) и их характеристиками

1. SQL	а. высокопроизводительная платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что эта платформа упрощает системный ландшафт, уменьшая затраты на поддержку аналитических систем.
MapReduce	б. язык структурированных запросов, позволяющий работать с базами данных. С его помощью можно создавать и модифицировать данные, а управлением массива данных занимается соответствующая система управления базами данных (СУБД).
SAP HANA	с. модель распределения вычислений. Используется для параллельных вычислений над очень большими наборами данных (петабайты и более). В программном интерфейсе не данные передаются на обработку программе, а программа – данным. Таким образом запрос представляет собой отдельную программу. Принцип работы заключается в последовательной обработке данных двумя методами.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 11. Факторный анализ и метод главных компонент

Задание 1.

По месячным данным начиная с января 2019 по июнь 2021 были оценены три регрессии:

$$Y = 18 - 0,427P + 0,000007I$$

$$\ln Y = 4,7 - 0,096P + 0,000001I$$

$$\ln Y = 12 - 3,11 \ln P + 0,0317 \ln I$$

Y – агрегированные расходы на медицинские услуги (млрд. руб.); P – индекс цен на медицинские услуги; I – среднемесячный доход (руб.). Дайте экономическую интерпретацию полученным результатам.

Ход выполнения работы:

Коэффициенты при переменной I во всех моделях незначимы, следовательно, доход не влияет на расходы на медицинские услуги. При увеличении индекса цен на медицинские услуги на единицу агрегированные расходы на медицинские услуги уменьшаются на 0,427 млрд. рублей. При увеличении индекса цен на медицинские услуги на единицу агрегированные расходы на медицинские услуги уменьшаются на 0,096 %. При увеличении индекса цен на медицинские услуги на 1 % агрегированные расходы на медицинские услуги.

Задание 2:

Эксперт считает, что уровень активности в теневой экономике Y зависит либо положительно от налогового бремени X , либо отрицательно от уровня государственных

расходов на предотвращение теневой экономической деятельности Z. Переменная Y может зависеть от обеих переменных X и Z. Полученные международные данные двух перекрестных выборок по Y, X и Z (в млн. долл. США): для групп по 30 индустриально развитых и для группы из 30 развивающихся стран. Исследователь оценивает регрессионные зависимости: $\log Y$ от $\log X$ и $\log Z$; $\log Y$ только от $\log X$; $\log Y$ только от $\log Z$ одновременно для каждой выборки, получая следующие результаты (в скобках приведены стандартные ошибки):

Таблица – Исходные данные

Параметр	Индустриально развитые страны			Развивающиеся страны		
	1	2	3	4	5	6
$\log X$	0,669	0,201	-	0,806	0,727	-
s. e.	(0,154)	(0,201)	-	(0,137)	(0,09)	-
$\log Z$	-0,646	-	-0,053	-0,091	-	0,427
s. e.	(0,162)	-	(0,124)	(0,117)	-	(0,116)
Константа	-1,137	-1,065	1,23	-1,122	-1,024	2,824
s. e.	(0,863)	(1,069)	(0,896)	(0,873)	(0,858)	(0,835)
R^2	0,44	0,10	0,01	0,71	0,70	0,33

Переменная X положительно коррелирована с Z в обеих выборках. Требуется выбрать какая из моделей лучше для индустриально развитых и для развивающихся стран, объяснить изменения в оценках коэффициентов и их стандартных отклонений в других моделях.

Ход выполнения работы:

Запишем модели, оцененные по данным для индустриально развитых стран:

$$(1) \widehat{\ln Y} = -1,137 + 0,669 \ln X - 0,646 \ln Z, R^2 = 0,44;$$

0,863 0,154 0,162

$$(2) \widehat{\ln Y} = -1,065 + 0,201 \ln X, R^2 = 0,1;$$

1,069 0,112

$$(3) \widehat{\ln Y} = 1,23 - 0,053 \ln Z, R^2 = 0,01.$$

0,896 0,124

Предположим, что лучшей является модель (1) – тогда в модели (2) коэффициент при переменной $\ln(X)$ является смещенным. Знак смещения совпадает со знаком произведения коэффициентов при не включённом факторе и ковариации не включённого и включенного факторов.

$\widehat{\beta_{\ln Z}} < 0$, $cov(X, Z) > 0$, следовательно смещение отрицательно, что можно наблюдать в модели (3): $0,201 < 0,669$. Модель (3) является неподходящей, так как $R^2 = 0,01$ – очень мал.

Таким образом, для индустриально развитых стран уровень активности в теневой экономике зависит как от налогового бремени, так и от государственных расходов на предотвращение теневой экономической деятельности.

Аналогично запишем модели, оцененные по данным для развивающихся стран:

$$(4) \widehat{\ln Y} = -1,122 + 0,806 \ln X - 0,091 \ln Z, R^2 = 0,71;$$

0,873 0,137 0,117

$$(5) \widehat{\ln Y} = -1,024 + 0,727 \ln X, R^2 = 0,7;$$

0,858 0,09

$$(6) \widehat{\ln Y} = 2,824 + 0,427 \ln Z, R^2 = 0,33.$$

0,835 0,116

Проанализировав коэффициенты и их стандартные ошибки в модели (4), мы можем заметить, что коэффициент при переменной $\ln Z$ незначим, т.е. этот фактор является в модели лишним. В этом случае оценки коэффициентов при переменной $\ln X$ в моделях (4) и (5)

должны быть близки, что можно наблюдать. В модели (6) знак оценки коэффициента при переменной $\ln Z$ является положительным, что означает положительную зависимость уровня активности теневой экономики от уровня государственных расходов на предотвращение теневой экономической деятельности. Таким образом, для развивающихся стран лучшей является модель (5), уровень активности в теневой экономике в этих странах зависит только от налогового бремени.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 12. Регрессионный анализ

Задание 1

Две группы экспертов проанализировали 12 проектов с точки зрения их эффективности.

Ранжировка 1 группы	Ранжировка 2 группы
1	2
3	3
4	1
2	4
5	6
6	5
12	9
7	7
8	8
9	10
10	12
11	11

Необходимо определить согласованность мнений экспертов в группах (то есть определить существует ли связь между мнениями экспертов в группах)

Ход решения:

Для расчета ранговой корреляции коэффициента Спирмена и доказательства согласованности мнения экспертов воспользуемся формулой:

$$r_s = 1 - \frac{\sum_{i=1}^n (x_i^{(k)} - x_i^{(j)})^2}{\frac{1}{6}(n^3 - n)} = 1 - \frac{\sum_{i=1}^n (x_i^{(k)} - x_i^{(j)})^2}{\frac{1}{6}(12^3 - 12)}$$

Сделаем промежуточные расчеты в Excel, найдем разность между рангами 1 и 2 группы (d)

Возведем полученные значения в квадрат d^2 и получим общую сумму.

Подставим в формулу, полученные значения, при $n=12$ (12 проектов).

«Соберем» формулу и получим:

$$r_s = 1 - \frac{6}{(1728 - 12)} * (1 + 0 + 9 + 4 + 1 + 1 + 9 + 0 + 0 + 1 + 4 + 0) = 0,892$$

Полученное значение коэффициента корреляции Спирмена составило 0,892, что свидетельствует о положительной ранговой связи между переменными и о согласованности мнений экспертов.

Задание 2

Изучается зависимость маркетинговых расходов (у) от доходности компаний (х). Данные по основным финансово-кредитным учреждениям в регионе приведены в таблице 1.

Таблица 1 – Исходные данные

№ п/п	Расходы на маркетинговые исследования	Доходность компаний
1	60	89
2	90	70
3	37	115
4	35	100
5	10	190
6	20	140
7	40	90
8	55	100
9	89	50
10	25	130
11	120	47
12	80	47
13	55	90
14	110	40
15	17	150

1. Построим график по эмпирическим данным, а также изучим модели регрессии, подходящие для описания данной зависимости между показателями. Сравним и оценим модели
2. Сделаем прогноз изменения маркетинговых расходов в при изменении доходности компании в следующих двух периодах до 160 и 165 соответственно

Ход решения:

1. Построим график по исходным данным примера. После переноса данных в Excel, проверим формат данных, он должен быть числовым.

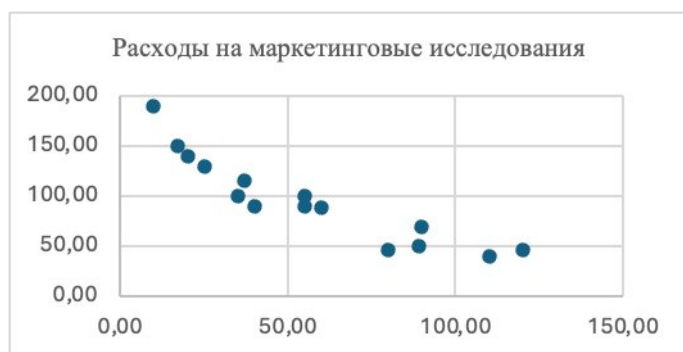


Рисунок – графическое изображение зависимости

2. Построим уравнения парной регрессии и определим наиболее подходящее. Для этого сделаем предположение о той или иной зависимости изучаемых явлений, задавшись уровнем значимости $\alpha = 0,05$ (вероятность ошибки 5%).

Для каждого уравнения регрессии рассчитаем коэффициент парной корреляции, коэффициент детерминации.

Для расчета параметров уравнения регрессии а и b, применим метод наименьших квадратов, где n – количество наблюдений

Ниже представлены формулы для трех моделей парной регрессии, воспользуемся ими для дальнейших расчетов:

Линейная регрессия: $\hat{Y} = a + bx + \varepsilon$

$$\bar{b} = \frac{\sum \bar{x}\bar{y} - \bar{x}\bar{y}}{\sum \bar{x}^2 - (\bar{x})^2} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\bar{a} = \bar{y} - \bar{b}\bar{x} = \frac{\sum y_i \sum x_i^2 - \sum x_i y_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Полиномиальная регрессия (парабола второго порядка)

Это форма регрессионного анализа, в котором взаимосвязь между независимой переменной x и зависимой переменной y моделируется как многочлен n -ой степени от x .

$$\begin{aligned} \hat{Y} &= a + bx^2 \\ \bar{b} &= \frac{\sum x^2 y - x^2 \bar{y}}{\sum x^4 - (x^2)^2} \\ \bar{a} &= \bar{y} - \bar{b} * x^2 \end{aligned}$$

Гиперболическая регрессия (гипербола) $\hat{Y} = a + \frac{b}{x} + \varepsilon$

Используется, когда заданные точки лежат вблизи гиперболы, то есть значения y убывают и стремятся к постоянной величине при увеличении значений x ($x \neq 0$).

$$\begin{aligned} \bar{b} &= \frac{\left(\frac{y}{x}\right) - \left(\frac{1}{x}\right) * \bar{y}}{\left(\frac{1}{x^2}\right) - \left(\frac{1}{x}\right)^2} \\ \bar{a} &= \bar{y} - \bar{b} \left(\frac{1}{x}\right) \end{aligned}$$

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 13. Анализ временных рядов

Задача 1

Квартальные данные о производстве электроэнергии в РФ охватывающие период 2018-2021 представлены в таблице.

Таблица – Производство электроэнергии в РФ

Год	Квартал	Порядковый номер квартала t	Производство электроэнергии y_t , млрд. кВт. ч
2018	I	1	274
	II	2	221
	III	3	218
	IV	4	279
2019	I	5	290
	II	6	232
	III	7	230
	IV	8	287
2020	I	9	292
	II	10	237
	III	11	235
	IV	12	288
2021	I	13	301

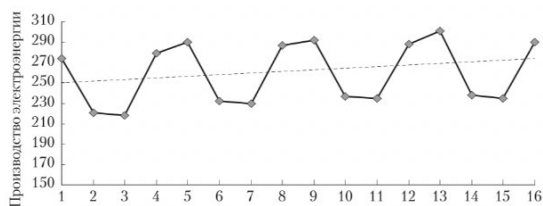
	II	14	238
	III	15	235
	IV	16	290

Требуется рассчитать прогнозную оценку производства электроэнергии в I квартале следующего года.

Ход решения:

1) Для проведения дальнейших расчетов создадим промежуточную таблицу где «переведем» кварталы года в последовательные числа от 1 до 16 (4 года в каждом по 4 квартала).

2) Начальный этап исследования связан с проведением графического анализа динамики исходного временного ряда. Построим график временного ряда («вставка» - «диаграмма») и проанализируем тенденцию.



На графике отчетливо видно внутригодовые сезонные колебания, характерные для динамики производства электроэнергии, что связано с периодическими подъемами производства в осенне-зимний период и спадами в весеннее-летние месяцы. Характер сезонных колебаний свидетельствует о целесообразности выбора для этого временного ряда модели с аддитивной сезонностью.

3) Найдем скользящую среднюю по предлагаемым формулам и заполним столбец «скользящая средняя» таблицы. Формулу можно задать, используя Excel и делая ссылки на соответствующие ячейки.

формуле:

$$y'_3 = \frac{\frac{1}{2} \cdot 274 + 221 + 218 + 279 + \frac{1}{2} \cdot 290}{4} = 250,0$$

$$y'_4 = \frac{\frac{1}{2} \cdot 221 + 218 + 229 + 290 + \frac{1}{2} \cdot 232}{4} = 253,375 \text{ и т.д.}$$

4) Вычтем из значений исходного уровня значения сглаженного ряда.

5) Рассчитаем оценки сезонной составляющей. Для этого создадим и заполним вспомогательную таблицу. Перенесем данные столбца $x_t = y_t - y'_t$ во вновь созданную таблицу соблюдая временную последовательность, начиная с 3-го квартала 2018 года. Для переноса столбцов в строки можно воспользоваться функцией Excel – специальная вставка – транспонировать и поочередно перенести данные. Для применения данной функции данные в столбце не должны содержать значения формул, поэтому их «значения» модно скопировать в свободный столбец.

Год	1 квартал	2 квартал	3 квартал	4 квартал
2018	-	-	-32,000	25,625
2019	33,750	-26,750	-30,000	26,125
2020	29,875	-25,875	-29,125	22,625
2021	35,500	-27,750	-	-
$\bar{x}_i$	33,042	-26,792	-30,375	24,792
S_i	32,875	-26,958	-30,542	24,625

Оценка сезонной составляющей осуществляется следующим образом:

Для 1 квартала: $\bar{x}_1 = \frac{33,750 + 29,875 + 35,500}{3} = 33,042$

$\bar{x}_2, \bar{x}_3, \bar{x}_4$ – рассчитываются аналогично.

6) Требуется провести корректировку полученных значений сезонной составляющей, так как их сумма отлична от нуля: $\sum_{i=1}^4 \bar{x}_i = 0,667 (33,042 - 26,792 - 30,375 + 24,792)$.

Далее определим значение поправки для расчета в дальнейшем скорректированных оценок сезонности:

$$-\bar{x} = \frac{0,667}{4} = 0,1667$$

Окончательную оценку сезонной компоненты для каждого квартала оценим по следующей формуле: $S_1 = 33,042 - 0,1667 = 32,875$ S_2, S_3, S_4 – рассчитываются аналогично.

7) Рассчитаем десезонализированный ряд (из исходных данных удаляем сезонную составляющую для соответствующего квартала).

Десезонализированный ряд y_t находится по следующей формуле:

$$274 - 32,875 = 241,125$$

$$221 - (-26,958) = 250,958 \text{ и т.д.}$$

8) Определим коэффициенты линейной трендовой модели для десезонализированного ряда.

Для этого воспользуемся инструментом Excel -- «Анализ данных»:

- y – производство электроэнергии

- t – фактор времени. (выделяем данные с шапкой таблицы, при расчете ставим знак V в ячейке «метки»).

Получена модель вида: $\hat{y}_t$

Для 1 квартала 2018 года расчет производится следующим образом:

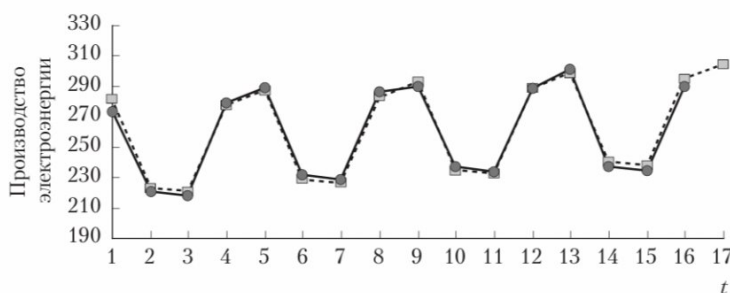
$$\hat{y} = 246,933 + 1,442t$$

9) Используя полученную модель, рассчитанную с применением метода наименьших квадратов, определим расчетные значения произведенной электроэнергии с учетом сезонного компонента и сделаем прогноз. Для этого подставим значения от 1 до 16 в полученную модель и добавим соответствующую сезонную компоненту. Для

$$\hat{y}_1 = 246,933 + 1,442 * 1 + 32,875$$

$$\hat{y}_2 = 246,933 + 1,442 * 2 \text{ и т.д.}$$

По аналогии рассчитаем значения для прогнозного периода 17 и 18 кварталов и представим данные графически:



Полученные значения характеристик точности, а также результаты визуального анализа свидетельствуют о близости расчетных и фактических значений. При этом необходимо учитывать, что использование этого подхода базируется на предположении об устойчивости сезонных эффектов во времени, а применение модели кривой роста для описания и прогнозирования тенденции предполагает сохранение инерционности развития.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 14. Методы data mining

Задание 1

Работа с датами в SQL включает в себя операции:

- извлечения,
- форматирования,
- добавления и вычитания
- и сравнения дат.

Понимание того, как каждая СУБД обрабатывает даты, поможет вам писать более эффективные и переносимые запросы.

Разберём основы работы с датой и извлечение из даты, добавление, вычитание и сравнение дат.

Напишите запросы в SQL Server, Oracle, PostgreSQL на извлечение частей даты, получение текущей даты и времени, сложение и вычитания дат, сравнения дат.

Ход решения:

ANSI Стандартные Функции

ANSI SQL стандарт определяет несколько функций для работы с датами, которые поддерживаются большинством СУБД:

CURRENT_DATE: Возвращает текущую дату.

CURRENT_TIME: Возвращает текущее время.

CURRENT_TIMESTAMP:

Возвращает текущую дату и время.

EXTRACT: Извлекает часть даты, такую как год, месяц, день и т.д.

MS SQL Server

MS SQL Server предлагает широкий набор функций для работы с датами:

Получение текущей даты и времени:

Sql

SELECT GETDATE(); -- Текущая дата и время

SELECT CURRENT_TIMESTAMP; - Текущая дата и время (ANSI стандарт)

Извлечение частей даты:

Sql

SELECT YEAR(GETDATE()) AS

CurrentYear;

SELECT MONTH(GETDATE()) AS

CurrentMonth;

SELECT DAY (GETDATE()) AS CurrentDay;

PostgreSQL также поддерживает набор функций для работы с датами:

Oracle

Oracle предоставляет свои уникальные функции для работы с датами, особенно для прибавления и вычитания дат:

Получение текущей даты и времени:

Sql

SELECT SYSDATE FROM dual; -- Текущая дата и время

SELECT CURRENT_DATE FROM dual; -- Текущая дата (ANSI стандарт)

SFI FCT CURRENT TTMESTAMP FROM dual.

Sql

SELECT SYSDATE FROM dual; -- Текущая дата и время

SELECT CURRENT_DATE FROM dual; --

Текущая дата (ANSI стандарт)

SELECT CURRENT_TIMESTAMP FROM dual;

- Текущая дата и время (ANSI стандарт)

Извлечение частей даты:

Sql
SELECT EXTRACT(YEAR FROM SYSDATE) AS
CurrentYear FROM dual;
SELECT EXTRACT(MONTH FROM SYSDATE)
AS CurrentMonth FROM dual;
SELECT EXTRACT(DAY FROM SYSDATE) AS
CurrentDay FROM dual;
MySQL
MySQL также поддерживает функции для работы с датами:
Получение текущей даты и времени:
Sql
SELECT CURDATE(); -- Текущая дата
SELECT CURTIME(); -- Текущее время
SELECT NOW(); -- Текущая дата и время
MS SQL Server
Сложение и вычитание дат:
Добавление 10 дней к текущей дате
SELECT DATEADD (day, 10, GETDATE()) AS AddDays;
- Вычитание 10 дней из текущей даты
SELECT DATEADD (day, -10, GETDATE ())
AS SubtractDays;
-- Добавление 2 месяцев к текущей дате
SELECT DATEADD (month, 2, GETDATE ())
AS AddMonths;
- Вычитание 1 года из текущей даты
SELECT DATEADD (year, -1, GETDATE ())

Задание 2

Используя Power BI подготовьте и загрузите Power BI Desktop. Ознакомьтесь с интерфейсом приложения.

После установки Power BI Desktop подключитесь к источникам данных <Получить дополнительные данные>, прокрутите список всех источников данных.

Представьте, что Вы являетесь аналитиком данных, работающим для розничной торговли солнцезащитным очками. Вы хотите помочь вашему клиенту сформировать наиболее перспективный целевой рынок, поэтому проводите анализ, где солнце светит чаще всего. Найдите и систематизируйте данную информацию используя Power BI Desktop. На вкладке главная страница Power BI Desktop» выберите <Получить данные> чтобы подключиться к источнику веб-данных.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 15. Анализ данных с применением машинного обучения

Задание 1.

Дана таблица order с колонками order_id, customer_id, order_date и total_amount.

Напишите запрос, который вернет customer_id клиентов, сделавших заказы на сумму более 1000 в течение последних 30 дней.

Order id	Customer id	Order date	Total amount
1	101	2024-08-01	1500.00
2	102	2024-08-02	800.00

3	101	2024-08-03	2000.00
4	103	2024-07-15	1200.00
5	104	2024-08-05	950.00
6	102	2024-08-06	1100.00
7	105	2024-08-07	1800.00
8	103	2024-08-08	750.00

Ход решения:

```
SELECT DISTINCT customer_id FROM orders WHERE total_amount > 1000 AND
order_date >= CURRENT_DATE - INTERVAL
'30 days';
```

```
SELECT DISTINCT customer_id FROM orders WHERE total_amount > 1000 AND order_date
BETWEEN CURRENT_DATE - 30
AND CURRENT_DATE;
```

Задание 2.

Напишите запрос, который вернет имя сотрудника с самой высокой зарплатой в каждом отделе.

Employee_id	name	department	salary
1	Helen	Sales	50000
2	Alexandr	IT	60000
3	Kira	Marketing	55000
4	Semen	Sales	52000
5	Lev	IT	65000
6	Olya	Marketing	58000
7	Georgie	Sales	51000
8	Oleg	IT	68000

Ход решения:

Создадим запрос:

```
b) SELECT name, department FROM employees e1 WHERE salary = (SELECT MAX(salary)
FROM
```

```
employees e2 WHERE e1.department = e2.department);
```

```
c) SELECT e1.name, e1.department FROM employees e1 JOIN (SELECT department,
MAX(salary) as max_salary FROM employees GROUP BY department) e2 ON e1.department =
e2.department AND e1.salary = e2.max_salary;
```

Задание 3

Создайте SQL коды для решения задач:

1) найдите магазины, в которых средний чек превышает заданную сумму (например 300 у.е.) за последние 6 месяцев.

Select

s. name, AVG(o. total_amount) AS

avg_order_value

FROM orders o

JOIN stores s ON o. store_id =

s. store_id

WHERE o. order_date >= CURRENT_DATE -

```
INTERVAL '6 months'
GROUP BY s. name
HAVING AVG (o. total_amount) > 300;
```

2) рассчитайте валовую прибыль (разницу между продажной и закупочной ценами) по категориям товаров за последний квартал

```
Select
p. category, SUM(oi quantity * (oi. unit_price - p.price)) AS gross_profit
FROM order_items oi
JOIN orders o ON oi.order_id =
o. order_id
JOIN products p ON oi. product_id =
p. product_id
WHERE o.order_date > CURRENT_DATE - INTERVAL '3 months'
GROUP BY p. category;
```

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Лабораторная работа по теме № 16. Методика прогнозирования. Прогнозная аналитика.

Задание 1.

Применяя метод Холта произведите сглаживание исходного временного ряда

Объем инвестиций	Период	Млрд., руб.
2020	январь	1000
	февраль	1200
	март	1260
	апрель	1197
	май	1790
	июнь	1969
	июль	1260
	август	1323
	сентябрь	1250
	октябрь	1790
	ноябрь	2148
	декабрь	2578
2021	январь	1050
	февраль	1260
	март	1323
	апрель	1257
	май	1880
	июнь	2068
	июль	1323
	август	1390
	сентябрь	1313
	октябрь	1880
	ноябрь	2256
	декабрь	2707

Ход решения:

Классическое экспоненциальное сглаживание работает в основном со стационарными рядами данных. При наличии тренда возникает необходимость прогнозировать с учетом трендовой составляющей.

$$\widehat{Y}_{t+p} = L_t + pT,$$

P – горизонт прогнозирования

L_t - оценка текущего уровня

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} - T_{t-1})$$

Где α – постоянная для сглаживания

T_t - оценка тренда

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

Для подбора оптимальных значений α и β необходимо провести расчет MSE для сетки значений α и β .

1. Зададим параметры α и β на уровне 0,3; L в январе 2020 = 1000 ; T в январе 2020 = 0.
2. Создадим ниже строку «экспоненциальное сглаживание с учетом тренда» где будем производить итоговые расчеты.

	A	B	C	D	E	F	G	H	I	J	K	
		январь	февраль	март	апрель	май	июнь	июль	август	сентябрь	октябрь	нс
Объем инвестиций		1000	1200	1260	1197	1790	1969	1260	1323	1250	1790	
a		0,1288359										
Экспоненциальное сглаживание		1000	1000	1025,76718	1055,94478	1074,11775	1166,34909	1269,75934	1268,50199	1275,52329	1272,23497	1
a		0,3										
b		0,3										
L		1000	1060	1107,4	1115,506	1303,01014	1453,04039	1328,78522	1306,7031	1280,08683	1431,92646	1
T		0	18	26,82	21,2058	71,095302	94,7757854	29,0665	13,7219151	1,62045901	46,68621	8
Экспоненц сглажив с учетом тренда		1000	1000	1078	1134,22	1136,7118	1374,10544	1547,81617	1357,85172	1320,42502	1281,70729	1

3. Рассчитаем значение L_t по формуле

$L_t = \alpha * Y_t + (1 - \alpha) * (L_{t-1} - T_{t-1})$. Вводим формулу в строку «L» в ячейку «февраль». Где L_{t-1} – значение в январе, T_{t-1} – значение в январе, Y_t – исходное значение инвестиций в феврале.

Параметр $\alpha=0,3$ в формуле закрепляем знаком \$\$

Пример формулы: $=B\$15 * C3 + (1 - B\$15) * (B17 - B18)$

4. Рассчитаем значение T_t по формуле

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta) * T_{t-1}$$

Вводим формулу в строку «T» в ячейку «февраль». Где L_{t-1} – значение в январе, T_{t-1} – значение в январе. Параметр $\beta=0,3$ в формуле закрепляем знаком \$\$

Пример формулы: $=B\$16 * (C17 - B17) + (1 - B\$16) * B18$

5. В строке «экспоненциальное сглаживание с учетом тренда» суммируем значения

$$\widehat{Y}_{t+p} = L_t + T$$

6. Построим график исходных значений и график полученных значений «экспоненциальное сглаживание с учетом тренда», сделаем выводы.

Задание 2

Используя данные Росстата (более 10 временных периодов), спрогнозируйте развитие агро-промышленной отрасли РФ. Для построения прогнозной модели можно опираться на ключевые макроэкономические и отраслевые показатели, используя MS Excel, ПО Statistica и другое ПО. Представьте полученные данные в виде презентации, применив Slides AI, Power Point и др.

Содержание отчета о проделанной работе

Отчет по лабораторной работе представляется в виде документа Word. В состав документа входят: название работы, цель работы, копии экрана, иллюстрирующие выполнение лабораторной работы, выводы по работе.

Список рекомендуемой литературы

Основная литература:

1. Воронова, Л. И. Big Data. Методы и средства анализа Электронный ресурс : Учебное пособие / Л. И. Воронова, В. И. Воронов. - Big Data. Методы и средства анализа, 2022-04-04. - Москва : Московский технический университет связи и информатики, 2016. - 33 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 2227-8397, экземпляров неограничено
2. Воскобойников, Ю. Е. Обработка и анализ экспериментальных данных в пакетах MathCAD и Excel : учебное пособие / Ю. Е. Воскобойников. - Обработка и анализ экспериментальных данных в пакетах MathCAD и Excel, Весь срок охраны авторского права. - Электрон. дан. (1 файл). - Новосибирск : Новосибирский государственный архитектурно-строительный университет (Сибстрин), ЭБС АСВ, 2020. - 161 с. - электронный. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-7795-0906-0, экземпляров неограничено
3. Шнарева, Г. В. Анализ данных : учебно-методическое пособие / Г. В. Шнарева, Ж. Г. Пономарева. - Анализ данных, 2024-12-06. - Электрон. дан. (1 файл). - Симферополь : Университет экономики и управления, 2019. - 129 с. - электронный. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 2227-8397, экземпляров неограничено

Дополнительная литература:

1. Горелов, В.И. Анализ статистических данных Электронный ресурс : практикум / Т.Н. Ледащева / В.И. Горелов. - Анализ статистических данных, 2020-06-16. - Москва : Российская международная академия туризма, Университетская книга, 2015. - 120 с. - Книга находится в базовой версии ЭБС IPRbooks. - ISBN 978-5-98699-151-1, экземпляров неограничено
2. Горяинова, Е. Р. Прикладные методы анализа статистических данных : учебное пособие / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. - Москва : Издательский дом Высшей школы экономики, 2012. - 312 с. - <http://biblioclub.ru/>. - ISBN 978-5-7598-0866-4, экземпляров неограничено
3. Кемаева, М. В. Количественные методы анализа экономики Электронный ресурс / Кемаева М. В. : учебно-методическое пособие. - Нижний Новгород : ННГУ им. Н. И. Лобачевского, 2017. - 49 с.
4. Селиванова, И.А. Построение и анализ алгоритмов обработки данных Электронный ресурс : учебно-методическое пособие / В.А. Блинов / И.А. Селиванова. - Построение и анализ алгоритмов обработки данных, Екатеринбург : Уральский федеральный университет, ЭБС АСВ, 2015. - 108 с. - Книга находится в базовой версии ЭБС IPRbooks. - ISBN 978-5-7996-1489-8, экземпляров неограничено

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

по организации самостоятельной работы
по дисциплине «Big Data: основы анализа больших данных»
для студентов направления подготовки 38.03.01 Экономика
направленность (профиль) «Бухгалтерский учет, анализ и аудит»

Ставрополь, 2026

Методические указания по дисциплине «Big Data: основы анализа больших данных» содержат задания для студентов, необходимые для организации самостоятельной работы.

Проработка предложенных заданий позволит студентам приобрести необходимые знания в области изучаемой дисциплины.

Предназначены для студентов направления 38.03.01 Экономика профиля «Бухгалтерский учет, анализ и аудит».

Содержание

Введение	35
План-график выполнения задания	35
Методические указания по изучению теоретического материала	36
Методические указания по подготовке к тестированию	38
Методические указания по подготовке к экзамену	39
Список рекомендуемой литературы	41

ВВЕДЕНИЕ

Данные методические указания по самостоятельной работе адресованы студентам очной и заочной форм обучения направления подготовки 38.03.01 Экономика.

В целях успешного освоения дисциплины студентам необходимо:

- изучить теоретический материал, предусмотренный практическим курсом;
- изучить теоретический материал, предназначенный для самостоятельного изучения;
- закрепить теоретический материал, выполнив практические задания;
- выполнить предусмотренные курсом практические работы.
- проводить сбор, обработку и систематизацию данных для их дальнейшего анализа и практического применения в бизнесе;
- иметь представление о построении и оценке прогнозных моделей;
- применять результаты аналитической работы с данными, интерпретировать полученные результаты;

Дисциплина изучается студентами в течение одного семестра.

1. ОБЩАЯ ХАРАКТЕРИСТИКА САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Целью учебной дисциплины является освоение основ и практических навыков аналитической работы с большими данными для решения прикладных задач в экономике.

В ходе преподавания дисциплины ставятся следующие задачи:

- формирование представлений о содержании, формах, особенностях дисциплины
- «Big Data: основы анализа больших данных»;
- изучение основных методов систематизации, обработки и анализа статистических данных;
 - изучение методов выявления и оценки степени зависимости между экономическими показателями;
 - формирование навыков по применению технологий обработки данных для решения прикладных задач, связанных с экономикой.

Самостоятельная работа по дисциплине «Big Data: Анализ данных в бизнесе» выполняется с целью получения и закрепления знаний, приобретенных при изучении теоретического материала.

2. ПЛАН – ГРАФИК ВЫПОЛНЕНИЯ ЗАДАНИЯ

Виды и содержание самостоятельной работы студента; формы контроля и сроки сдачи содержатся в рабочей программе дисциплины.

3. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ИЗУЧЕНИЮ ТЕОРЕТИЧЕСКОГО МАТЕРИАЛА

Изучение любого раздела или темы следует начинать с ознакомления с вопросами плана изучения темы. Теоретический материал представляет собой конспект лекций, содержащий необходимый набор утверждений и формул (без детальных подробностей), но с подробным обоснованием их использования при решении конкретных экономических задач. При изучении материала необходимо помимо лекционных материалов использовать рекомендуемую основную и дополнительную литературу для лучшего усвоения материала.

Осваивать теорию следует в соответствии с той последовательностью, которая представлена в плане лекции. Методика работы с литературой предусматривает ведение записи

прочитанного в виде плана - конспекта, опорного конспекта. Это позволит сделать знания системными, зафиксировать и закрепить их в памяти.

Для успешного освоения дисциплины, необходимо самостоятельно детально изучить представленные темы по рекомендуемым источникам информации:

№ п/п	Виды самостоятельной работы	Рекомендуемые источники информации (№ источника)			
		Основная	Дополнительная	Методическая	Интернет-ресурсы
1	Выполнение практических работ	1 2 3	1 2 3	1	1 2 3
2	Подготовка к лекции	1 2 3	1 2 3	1	1 2 3
3	Самостоятельное изучение литературы	1 2 3	1 2 3	1	1 2 3
4	Подготовка к зачету	1 2 3	1 2 3	1	1 2 3

4. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПОДГОТОВКЕ К ТЕСТИРОВАНИЮ

Успешное выполнение тестовых заданий является необходимым условием итоговой положительной оценки в соответствии с рейтинговой системой обучения. Выполнение тестовых заданий предоставляет студентам возможность самостоятельно контролировать уровень своих знаний, обнаруживать пробелы в знаниях и принимать меры по их ликвидации. Форма изложения тестовых заданий позволяет закрепить и восстановить в памяти пройденный материал. Тестовые задания охватывают основные вопросы по дисциплине «Эконометрика».

У студента есть возможность выбора правильного ответа или нескольких правильных ответов из числа предложенных вариантов. Для выполнения тестовых заданий студенты должны изучить лекционный материал по теме, соответствующие разделы учебников, учебных пособий и других источников.

Контрольный тест выполняется студентами самостоятельно во время практических занятий.

Тестовые задания

1. В какой вкладке размещены инструменты «Анализ данных» в MS Excel

- а) вставка
- б) вид
- в) данные

2. Прогнозные значения эндогенных переменных можно получить только с помощью экзогенных переменных

- а) да
- б) нет

3. Из перечисленных методов в статистике не используется метод

- а) корреляции
- б) повторного обследования
- в) тройных интервалов
- г) средних

3. Какие диаграммы можно использовать для построения графиков функций
- а) точечные
 - б) лепестковые
 - в) поверхностные
4. Что целесообразно создавать при необходимости сравнения значений нескольких наборов данных:
- а) поверхностную диаграмму
 - б) графики
 - в) гистограммы
5. Какой вид диаграмм можно использовать при большом количестве данных, необходимости отображения динамики изменения данных во времени:
- а) циклическая гистограмма
 - б) круговая гистограмма
 - в) график
6. Применение скользящей средней привело к потере пяти уровней в начале и пяти уровней в конце временного ряда. Следовательно, использовалась длина интервала сглаживания l :
- а) 5
 - б) 11
 - в) 15
 - г) 6
7. Датасет – это...
8. При использовании модели экспоненциального сглаживания $\overline{S_t = \alpha y + \beta S_{t-1}}$ параметр сглаживания α не может быть равен:
- а) -0,5
 - б) 0,1
 - в) 1,5
 - г) -1,0
9. Наиболее часто применяют три типа условного форматирования:
- а) прямая, обратная, пропорциональная шкала
 - б) секторальные, точечные, круговые диаграммы
 - в) цветовые шкалы, наборы значков, гистограммы
10. Что относится к количественным метрикам в BI
- а) количество операций
 - б) количество экземпляров
 - в) количество рабочих мест
 - г) количество поставщиков
11. Выбросы в анализе данных – это...
12. Регрессионная модель является линейной относительно переменных, если выполняются условия:
- а) переменные находятся в первой степени
 - б) модель должна быть аддитивной
 - в) модель должна быть мультипликативной

13. Выберите названия тех числовых характеристик данных измерения, которые отвечают за разброс данных вокруг среднего значения.

- а) дисперсия
- б) среднее квадратическое отклонение
- в) среднее измерения
- г) размах измерения
- д) медиана измерения

14. Общий вид множественного регрессионного уравнения для выборочных данных можно представить следующим математическим выражением

$$Y_i = f(X_{1i}, X_{2i}, X_{3i}) + e_i = Y_{pi} + e_i$$

- а) нет
- б) да

15. Наиболее часто используется две меры отклонения от средней тенденции:

- а) размах и стандартное отклонение
- б) дисперсия и стандартная ошибка
- в) мода и медиана

16. Качественные признаки данных могут быть:

- а) последовательными
- б) номинальными
- в) порядковыми
- г) условными

17. Сто студентов прошли тестирование по математике и физике. Пусть переменная x – рейтинг студентов по математическому анализу, а переменная y – рейтинг по физике. Коэффициент корреляции Спирмена для переменных x и y оказался равным 0,6. Это позволяет сделать вывод, что на уровне значимости 0,05:

- а) между показателями x и y существует монотонная положительная связь
- б) между показателями x и y существует монотонная отрицательная связь
- в) нет оснований для отклонения гипотезы о независимости показателей x и y
- г) не позволяет сделать вывод о зависимости или независимости показателей x и y

18. Для машинного обучения подходят данные а) бинарные

- в) предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов
- г) числовые типа `int`
- д) любых форматов в цифровом виде

19. Регрессионная модель с переменной структурой - это такая модель, которая учитывает неоднородность количественной зависимой переменной с помощью качественного фактора (фиктивной переменной)

- а) да
- б) нет

20. Наиболее часто применяют три типа условного форматирования:

- а) прямая, обратная, пропорциональная шкала
- б) секторальные, точечные, круговые диаграммы

в) цветовые шкалы, наборы значков, гистограммы

Критерии оценивания компетенций

Оценка «отлично» выставляется студенту, если количество правильных ответов на тестовые задания превышает 90 %.

Оценка «хорошо» выставляется студенту, если количество правильных ответов на тестовые задания превышает 70 %).

Оценка «удовлетворительно» выставляется студенту, если количество правильных ответов на тестовые задания превышает 50 %).

Оценка «неудовлетворительно» выставляется студенту, если количество правильных ответов на тестовые задания составляет менее 50 %, либо ответы заимствованы.

Оформление ответов на тесты

Ответы на тесты оформляются на студентом на отдельном листе самостоятельно. В правом углу проставляется ФИО и группа, далее следует номер теста и выбранный вариант ответа.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПОДГОТОВКЕ К ЗАЧЕТУ

Успеваемость студентов по дисциплине оценивается в ходе текущего контроля и промежуточной аттестации.

Текущий контроль

Текущее контрольное мероприятие считается сданным, если студент получил за него не менее 60 % от установленного для этого контроля максимального балла.

Промежуточная аттестация

Промежуточная аттестация в форме **экзамена**. Компетенции (знания, умения и навыки) студентов оцениваются по пятибалльной системе: «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно». В ходе собеседования студенту предлагается теоретический вопрос и практическая задача. Для подготовки к ответу отводится 45 минут. При подготовке к ответу студенту предоставляется право пользования методическими материалами. При проверке практического задания, оцениваются: последовательность и рациональность выполнения, полнота ответа, правильность сделанных выводов.

Вопросы для собеседования:

Классификация статистических данных

Предварительный анализ данных: общая характеристика

Классификация данных по числу переменных

Типы переменных и уровни их измерения

Характеристика генеральной совокупности и ее распределение

Группировка дискретных количественных данных

Данные в экономике: объекты, признаки, таблицы

Инструменты описательной статистики в Microsoft Excel

Анализ описательных статистик

Условное форматирование, графики и диаграммы рассеяния: обзор и сфера применения

Статистические зависимости и проверка данных

Технологии хранения данных

Закон больших чисел: сущность, основные теоремы

Корреляционный анализ взаимосвязи количественных признаков

Определение случайной величины, функция распределения случайной величины

Определение вероятности: статистическая вероятность, классическая вероятностная схема

Дискретные случайные величины, часто встречающиеся в экономической практике

Меры связи случайных величин

Проверка статистических гипотез с помощью доверительных интервалов

Нелинейные модели регрессии и их линейаризация
 Множественная линейная модель регрессии
 Анализ временных данных. Методы сглаживания временных и моделирования тенденции развития
 Методы data-mining
 Модели линейной регрессии
 Основы кластеризации данных: принципы и методы
 Метод Алгоритмы построения деревьев решений
 Основы и принципы машинного обучения, предварительная обработка данных
 Однофакторный и двухфакторный дисперсионный анализ
 Факторный анализ и метод главных компонент
 Анализ данных с применением машинного обучения

Практические задания

1. В течение 14 дней записывалось сколько кредитов на покупку автомобилей было выдано в банке в каждый из этих дней: 5, 6, 6, 7, 11, 7, 5, 2, 8, 7, 4, 8, 9, 10. Вычислим статистические характеристики данных о количестве выданных кредитов с использованием MS Excel и инструмента «Описательная статистика»

2. Заполните следующую таблицу:

№ п/п	Этапы добычи данных	Характеристика этапа
1	Постановка задачи	
2	Подготовка данных	
3	Изучение данных	
4	Построение моделей	
5	Исследование и проверка моделей	
6	Развертывание и обновление моделей	

3. Постройте графики зависимости цен закрытия и объемов торгов от времени по данным котировки акций ВТБ, Сбербанка, Газпрома?

4. Визуализируйте табличные данные по индексам цен на рынке коммерческой недвижимости с использованием программы Microsoft Excel, сделайте вывод о полученной тенденции

Год	Индекс, %	Год	Индекс, %
2010	156,9	2016	118,5
2011	146,3	2017	117,5
2012	113,1	2018	147,7
2013	125,1	2019	123,4
2014	122,5	2020	110,3
2015	118,8	2021	92,4

5. Дайте прогноз темпов роста экономики региона на предстоящий год предварительно проанализировав тенденции за предшествующий период и сгладив временной ряд используя метод экспоненциального сглаживания?

6. По 50 наблюдениям была оценена парная регрессия $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ b получены оценки коэффициента наклона и стандартного отклонения. При условии значимости 5% и 1 % проверьте основную гипотезу $H_0 : \beta_1 = 0$ при следующих результатах оценивания и основных гипотез:

- а) $\hat{\beta}_1 = 0,30, \hat{\sigma}_{\beta_1} = 0,12, H: \beta_1 \neq 0$
 б) $\hat{\beta}_1 = 0,30, \hat{\sigma}_{\beta_1} = 0,12, H: \beta_1 > 0$

Вычислите значение тестовой статистики для предложенных случаев. Сделайте выводы.

7. Требуется создать облако для хранения файлов с большими данными, создать персональный кабинет для управления и произвести индивидуальные настройки.

8. С вероятностью 0,954 определите средний выпуск продукции и долю промышленных предприятий с выпуском продукции до 10 млн. р. на основании следующих статистических данных, полученных в результате 10%-ой бесповторной выборки

Выпуск продукции, млн. руб.	До 3	3-5	5-10	10-30	Более 30
Кол-во предприятия	5	15	24	40	16

9. Установите соответствие между наиболее распространенными подходами обработки данных (ПО) и их характеристиками

1. SQL	а. высокопроизводительная платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что эта платформа упрощает системный ландшафт, уменьшая затраты на поддержку аналитических систем.
MapReduce	б. язык структурированных запросов, позволяющий работать с базами данных. С его помощью можно создавать и модифицировать данные, а управлением массива данных занимается соответствующая система управления базами данных (СУБД).
SAP HANA	с. модель распределения вычислений. Используется для параллельных вычислений над очень большими наборами данных (петабайты и более). В программном интерфейсе не данные передаются на обработку программе, а программа – данным. Таким образом запрос представляет собой отдельную программу. Принцип работы заключается в последовательной обработке данных двумя методами.

10. Требуется построить вариационный дискретный ряди получить различные изображения данных – гистограмму, полигон, кумуляту частот и относительных частот ряда распределения

5	1	3	4	5	6
1	3	4	6	8	1
4	6	3	3	2	2
2	6	3	3	7	2
0	0	4	1	2	4

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Основная литература

5. Воронова, Л. И. Big Data. Методы и средства анализа Электронный ресурс : Учебное пособие / Л. И. Воронова, В. И. Воронов. - Big Data. Методы и средства анализа, 2022-04-04. - Москва : Московский технический университет связи и информатики, 2016. - 33 с. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 2227-8397, экземпляров неограничено
6. Воскобойников, Ю. Е. Обработка и анализ экспериментальных данных в пакетах MathCAD и Excel : учебное пособие / Ю. Е. Воскобойников. - Обработка и анализ экспериментальных данных в пакетах MathCAD и Excel, Весь срок охраны авторского права. - Электрон. дан. (1 файл). - Новосибирск : Новосибирский государственный архитектурно-строительный университет (Сибстрин), ЭБС АСВ, 2020. - 161 с. - электронный. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 978-5-7795- 0906-0, экземпляров неограничено
7. Шнарева,, Г. В. Анализ данных : учебно-методическое пособие / Г. В. Шнарева, Ж. Г. Пономарева. - Анализ данных, 2024-12-06. - Электрон. дан. (1 файл). - Симферополь : Университет экономики и управления, 2019. - 129 с. - электронный. - Книга находится в премиум-версии ЭБС IPR BOOKS. - ISBN 2227-8397, экземпляров неограничено

Дополнительная литература

1. Горелов, В.И. Анализ статистических данных Электронный ресурс : практикум / Т.Н. Ледашева / В.И. Горелов. - Анализ статистических данных, 2020-06-16. - Москва : Российская международная академия туризма, Университетская книга, 2015. - 120 с. - Книга находится в базовой версии ЭБС IPRbooks. - ISBN 978-5-98699-151-1, экземпляров неограничено
2. Горяинова, Е. Р. Прикладные методы анализа статистических данных : учебное пособие / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. - Москва : Издательский дом Высшей школы экономики, 2012. - 312 с. - <http://biblioclub.ru/>. - ISBN 978-5-7598-0866-4, экземпляров неограничено
3. Кемаева, М. В. Количественные методы анализа экономики Электронный ресурс / Кемаева М. В. : учебно-методическое пособие. - Нижний Новгород : ННГУ им. Н. И. Лобачевского, 2017. - 49 с.
4. Селиванова, И.А. Построение и анализ алгоритмов обработки данных Электронный ресурс : учебно-методическое пособие / В.А. Блинов / И.А. Селиванова. - Построение и анализ алгоритмов обработки данных, Екатеринбург : Уральский федеральный университет, ЭБС АСВ, 2015. - 108 с. - Книга находится в базовой версии ЭБС IPRbooks. - ISBN 978-5-7996-1489-8, экземпляров неограничено