МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой
Математических методов исследования операций
Азарнова Т.В.
18.05.2022 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.О.17. Анализ данных

1. Код и наименование направления подготовки/специальности:

38.03.05 Бизнес информатика

2. Профиль подготовки/специализация:

Бизнес-аналитика и системы автоматизации предприятий

3. Квалификация выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: Математических методов исследования операций

6. Составители программы: Азарнова Татьяна Васильевна, доктор техн. наук, профессор кафедры математических методов исследования операций

7. Рекомендована: НМС факультета Прикладной математики, информатики и механики, протокол №8 от 15.04.2022.

8. Учебный год: 2024/2025 **Семестр(-ы):** 5

9.Цели и задачи учебной дисциплины

В рамках данного курса слушатели получают знания о математическом аппарате анализа статистических данных различной природы и приобретают навыки в математическом моделировании процесса исследования, т.е. в искусстве формализации постановки реальной задачи, которое заключается в умении перевести задачу с языка проблемно-содержательного (экономического, социологического, медицинского, технического и т.п.) на язык абстрактных математических схем и моделей.

Задачи дисциплины – формирование знаний, умений и навыков по следующим направлениям: способы организации выборок; методы проверки статистических гипотез; дисперсионный анализ; факторный анализ; методы классификации; дискриминантный анализ; деревья решений; анализ временных рядов, использование современного программного обеспечения для статистического анализа данных.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к обязательным дисциплинам базового цикла. Для изучения курса необходимы базовые знания линейной алгебры, математического анализа, теории вероятностей, математической статистики, методов оптимизации.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения:

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ОПК-4	Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач	ОПК-4.1 ОПК-4.2 ОПК-4.3	ОПК-4.1 Собирает и анализирует информацию для поддержки принятия решений	знать: – методы формирования выборок; – методы предварительной обработки данных; – методы отбора информативных признаков; – методы проверки статистических гипотез; – методы оценки зависимости между категорированными, порядковыми и интервальными данными; – методы дисперсионного анализа; – методы факторного анализа; – методы классификации; – методы регрессионного анализа; – методы анализа временных рядов; уметь: – проводить опросы и анкетирование для сбора информации в задачах поддержки принятия решений; – анализировать многомерные данные и формировать систему методов их обработки; владеть (иметь навык(и)): – подбора методов сбора информации и проверки качества моделей обработки данных в задачах

			ОПК-4.2 Использует методы и программные средства обработки информации	принятия решений; – интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов. знать: – методы и программные средства сбора информации, необходимой для решения исследовательских задач; – методы программные средства анализа информации для решения поставленных научно-исследовательских задач; уметь: – применять методы и программные средства обработки и статистического анализа данных для решения поставленных задач; – использовать возможности пакетов прикладных программ для обработки информации и решения поставленных задач; владеть (иметь навык(и)): – навыками подбора методов и программных средств обработки информации при решении прикладных задачи.

12. Объем дисциплины в зачетных единицах/час (*в соответствии с учебным планом)* — 5/180.

Форма промежуточной аттестации*(зачет/экзамен) экзамен*

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость	
		Всего	По семестрам
			№ семестра
Контактная работа			
в том числе:	лекции	34	34
	практические	16	16
	лабораторные	16	16
Самостоятельная работа в том числе: курсовая работа (проект)		78	78
Форма промежуточной аттестации		36	36
Итого:		180	180

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
		1. Лекции	
1	Первичная статистическая обработка данных	Шкалы измерений. Унифицированное представление разнотипных данных. Способы заполнения пропущенных данных. Визуализация многомерных данных. Анализ резко выделяющихся наблюдений	Анализ данных 38.03.05
2	Проверка статистических гипотез в прикладных задачах	Основные типы гипотез, проверяемых в ходе статистического анализа и моделирования. Критерий знаков для одной выборки. Критерий Манна-Уитни. Критерий Уилкоксона. Критерий знаков для анализа парных повторных наблюдений. Критерий знаковых ранговых сумм Уилкоксона. Проверка гипотез, связанных с параметрами нормального распределения.	Анализ данных 38.03.05
3	Дисперсионный анализ	Однофакторный дисперсионный анализ Многофакторный дисперсионный анализ Ранговый дисперсионный анализ	Анализ данных 38.03.05
4	Анализ структуры и тесноты статистической связи между исследуемыми переменными.	Анализ тесноты связи между количественными переменными. Анализ статистической связи между порядковыми переменными. Анализ связей между классификационными переменными.	Анализ данных 38.03.05
5	Факторный анализ	Факторный анализ при жестко фиксированном количестве факторов Факторный анализ при нефиксированном количестве факторов Методы вращения и интерпретации факторов	Анализ данных 38.03.05
6	Распознавание образов и типологизация объектов в социально – экономических исследованиях.	Сущность, типологизация и прикладная направленность задач классификации объектов. Классификация при наличии обучающих выборок (дискриминантный анализ). Классификация без обучения (параметрический случай): расщепление смесей вероятностных распределений. Классификация без обучения (непараметрический случай): кластер-анализ.	Анализ данных 38.03.05
7	Временные ряды	Понятие временного ряда. Компоненты временных рядов. Стационарные временные ряды и их основные характеристики. Неслучайная составляющая временного ряда и методы его сглаживания. Модели нестационарных временных рядов и их идентификация. Прогнозирование экономических показателей, основанное на использовании временных рядов.	Анализ данных 38.03.05
		2. Лабораторные занятия	
1	Первичная статистическая обработка данных	Анализ данных в пакете статистического анализа данных по темам: -Шкалы измерений. -Унифицированное представление разнотипных	Анализ данных 38.03.05

		данных. -Способы заполнения пропущенных данных. -Визуализация многомерных данных. -Анализ резко выделяющихся наблюдений	
2	Проверка статистических гипотез в прикладных задачах	Анализ данных в пакете статистического анализа данных по темам: - Критерий знаков для одной выборки. -Критерий Манна-Уитни. -Критерий Уилкоксона. -Критерий знаков для анализа парных повторных наблюдений. -Критерий знаковых ранговых сумм Уилкоксона. -Проверка гипотез, связанных с параметрами нормального распределения.	Анализ данных 38.03.05
3	Дисперсионный анализ	Анализ данных в пакете статистического анализа данных по темам: -Однофакторный дисперсионный анализ. -Многофакторный дисперсионный анализ. -Ранговый дисперсионный анализ.	Анализ данных 38.03.05
4	Анализ структуры и тесноты статистической связи между исследуемыми переменными.	Анализ тесноты связи между количественными переменными в пакете статистического анализа данных. Анализ статистической связи между порядковыми переменными в пакете статистического анализа данных. Анализ связей между классификационными переменными в пакете статистического анализа данных.	Анализ данных 38.03.05
5	Факторный анализ	Факторный анализ при жестко фиксированном количестве факторов в пакете статистического анализа данных. Факторный анализ при нефиксированном количестве факторов в пакете статистического анализа данных. Методы вращения и интерпретации факторов в пакете статистического анализа данных.	Анализ данных 38.03.05
6	Распознавание образов и типологизация объектов в социально – экономических исследованиях.	Классификация при наличии обучающих выборок (дискриминантный анализ) в пакете статистического анализа данных. Классификация без обучения (параметрический случай): расщепление смесей вероятностных распределений в пакете статистического анализа данных. Классификация без обучения (непараметрический случай): кластер-анализ в пакете статистического анализа данных.	Анализ данных 38.03.05
7	Временные ряды	Прогнозирование экономических показателей, основанное на использовании временных рядов в пакете статистического анализа данных.	Анализ данных 38.03.05

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Первичная статистическая обработка данных	4	2	2	10	18
2	Проверка статистических гипотез в прикладных задачах	6	4	2	10	22
3	Дисперсионный анализ	2	2	2	6	12
4	Анализ структуры и тесноты	8	2	2	10	22

	статистической связи между исследуемыми переменными.					
5	Факторный анализ	2	2	2	6	12
6	Распознавание образов и типологизация объектов в социально – экономических исследованиях.	4	2	2	10	18
7	Временные ряды	8	2	4	12	26
8	Курсовая работа				14	14
	Итого	34	16	16	78	180

14. Методические указания для обучающихся по освоению дисциплины

Для лучшего усвоения материала студентам рекомендуется домашняя работа с конспектами лекций, презентациями, выполнение практических заданий для самостоятельной работы, выполнение лабораторных работ, использование рекомендованной литературы и методических материалов. В рамках общего объема часов, отведенных для изучения дисциплины, предусматривается выполнение следующих видов самостоятельных работ студентов (СРС): изучение теоретического материала, выполнение в пакете статистического анализа данных заданий по темам, изученным на лекционных и практических занятия.

При использовании дистанционных образовательных технологий и электронного обучения выполнять все указания преподавателей, вовремя подключаться к online занятиям, ответственно подходить к заданиям для самостоятельной работы.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины *(список литературы оформляется в соответствии с требованиями ГОСТ и используется общая сквозная нумерация для всех видов источников)*

а) основная литература:

№ п/п	Источник
1	Котиков, П. Е. Анализ данных : учебно-методическое пособие / П. Е. Котиков. — Санкт-Петербург : СПбГПМУ, 2019. — 48 с. — ISBN 978-5-907184-46-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/174498
2	Вольфсон, М. Б. Анализ данных : учебное пособие / М. Б. Вольфсон. — Санкт-Петербург : СПбГУТ им. М.А. Бонч-Бруевича, 2015. — 81 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/180254
3	Кузьмин, В. И. Методы анализа данных : учебное пособие / В. И. Кузьмин, А. Ф. Гадзаов. — 2-е изд., перераб. и доп. — Москва : РТУ МИРЭА, 2020. — 155 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/171433.

б) дополнительная литература:

№ п/п	Источник
4	Горелов, В. И. Анализ статистических данных : практикум : [16+] / В. И. Горелов, Т. Н. Ледащева ; Российская международная академия туризма. – Москва : Университетская книга, 2015. – 120 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=574944. – Библиогр.: с. 107. – ISBN 978-5-98699-151-1. – Текст : электронный.
5	Агалаков, С. А. Статистические методы анализа данных : учебное пособие : [16+] / С. А. Агалаков ; Омский государственный университет им. Ф. М. Достоевского. – Омск : Омский государственный университет им. Ф.М. Достоевского, 2017. – 92 с. : табл., граф., схем., ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=562918 . – Библиогр. в кн. – ISBN 978-5-7779-2187-1. – Текст : электронный.
6	Жуковский, О. И. Информационные технологии и анализ данных : учебное пособие / О. И. Жуковский ; Томский Государственный университет систем управления и радиоэлектроники (ТУСУР). – Томск : Эль Контент, 2014. – 130 с. : схем., ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=480500 . –

	Библиогр.: с. 126. – ISBN 978-5-4332-0158-3. – Текст : электронный.

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет)*:

№ п/п	Ресурс
1.	*ЭБС Лань*
2.	*ЭБС ЮРАЙТ*
3.	*edu.vsu.ru*

16. Перечень учебно-методического обеспечения для самостоятельной работы *(учебно-методические рекомендации, пособия, задачники, методические указания по выполнению практических (контрольных), курсовых работ и др.)*

№ п/п	Источник
1	Зубов, Н. Н. Статистика в биомедицине, фармации и фармацевтике : учебное пособие : [16+] / Н. Н. Зубов, В. И. Кувакин, С. З. Умаров ; под общ. ред. И. А. Наркевича. – Москва ; Берлин : Директ-Медиа, 2019. – 386 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=578236 . – Библиогр.: с. 326-327. – ISBN 978-5-4499-1173-5. – DOI 10.23681/578236. – Текст : электронный.
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. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=227280. – ISBN 978-5-7598-0866-4. – Текст : электронный.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

При реализации дисциплины могут использоваться технологии электронного обучения и дистанционные образовательные технологии на базе портала edu.vsu.ru, а также другие доступные ресурсы сети Интернет.

18. Материально-техническое обеспечение дисциплины:

Лекционная аудитория должна быть оснащенной современным компьютером с подключенным к нему проектором с видеотерминала на настенный экран. Практические и лабораторные занятия должны проводиться в специализированной аудитории, оснащенной современными персональными компьютерами и программным обеспечением в соответствии с тематикой изучаемого материала. Предполагаемое оборудование для компьютерных классов: компьютеры в составе:системный блок: процесс Intel(R) Core(TM) i3-4160 CPU @ 3.60GHz, оперативная память 8Гб, HDD 500Гб, видеокарта NVIDIA GeForce GTX 750; монитор: Acer V226HQL; мультимедиа-проектор ViewSonic PA503W. Коммутатор HP ProCurve Switch 1400-24G; мультимедийная акустическая система SVEN SPS-702

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Первичная статистическая	ОПК-4	ОПК-4.1	Задание для лабораторной работы 1

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
	обработка данных		ОПК-4.2 ОПК-4.3	
2.	Проверка статистических гипотез в прикладных задачах	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 1
3	Дисперсионный анализ	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 1
4	Анализ структуры и тесноты статистической связи между исследуемыми переменными.	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 1
5	Факторный анализ	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 2
6	Распознавание образов и типологизация объектов в социально – экономических исследованиях.	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 3
7	Временные ряды	ОПК-4	ОПК-4.1 ОПК-4.2 ОПК-4.3	Задание для лабораторной работы 4
Промежуточная аттестация форма контроля - экзамен				Тест Практическое задание

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

Тестовые задания, Лабораторные работы, Устный опрос

1. **Лабораторная работа №1**

1. В предложенном вам файле «Таблица квартир» рассмотрите переменную Цена. Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
2. В предложенном вам файле «Cars» рассмотрите переменную Цена. Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.

3. В предложенном вам файле «Tights» рассмотрите переменную Цена. Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
4. В предложенном вам файле «Таблица квартир» рассмотрите переменную Цена. Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
5. В предложенном вам файле «Продолжительность жизни» рассмотрите переменную «в городе». Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
6. В предложенном вам файле «Продолжительность жизни» рассмотрите переменную «в селе». Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
7. В предложенном вам файле «Продолжительность жизни» рассмотрите переменную «в селе». Вычислите все описательные статистики и постройте простейшие статистические графики для данной переменной, проверьте гипотезу о нормальности, используя глазомерный метод проверки на нормальность. Проинтерпретируйте полученные результаты.
8. В предложенном вам файле Adstudy (стандартные примеры) рассмотрите переменные «Gender» и «Advert». Проверьте гипотезу о независимости этих двух признаков.
9. В предложенном вам файле «школьники» рассмотрите переменные «Continuation decision» и «1 for girls, 0 for boys» . Проверьте гипотезу о независимости этих двух признаков.
10. В предложенном вам файле «школьники» рассмотрите переменные «Continuation decision» и «Number of younger siblings (at age 16)». Проверьте гипотезу о независимости этих двух признаков.
11. В предложенном вам файле «Качество работы» рассмотрите переменные «Приоритеты в работе» и «Состав семьи». Проверьте гипотезу о независимости этих двух признаков.
12. В предложенном вам файле «1960-1985» рассмотрите переменные «NOIL» и «INTER». Проверьте гипотезу о независимости этих двух признаков.
13. В предложенном вам файле NEW11 приведены данные об объемах продаж в 15 магазинах до и после рекламы. Выяснить, значим ли эффект рекламы.
14. С помощью t-теста для независимых признаков проверьте гипотезу о том, средние значения переменной « Number of O-levels obtained at national exams at age 16 (prior to continuation decision)» одинаковы для мальчиков и девочек. Файл «школьники».
15. С помощью t-теста для независимых признаков проверьте гипотезу о том, средние значения переменной «price» одинаковы для различных значений переменной «foreign/russian». Файл «Mydata».
16. С помощью t-теста для независимых признаков проверьте гипотезу о том, средние значения переменной «price» одинаковы для различных значений переменной «status». Файл «Mydata».
17. С помощью t-теста для независимых признаков проверьте гипотезу о том, средние значения переменной « Number of O-levels obtained at national exams at age 16 (prior to continuation decision)» одинаковы для мальчиков и девочек. Файл «школьники».
18. С помощью t-теста для независимых признаков проверьте гипотезу о том, средние значения переменной «цена » одинаковы для различных фирм. Файл «Tight».
19. С помощью t-теста для зависимых признаков проверьте гипотезу о том, средние значения переменной «GDP60» и переменной «GDP85» равны (нет эффекта времени) Файл «1960-1985».

20. С помощью дисперсионного анализа проверьте влияние «округа» на «уровень безработицы». Файл «Экономика». Проинтерпретируйте полученные результаты.
21. С помощью дисперсионного анализа проверьте влияние «округа» на «коэффициент расслоения». Файл «Экономика». Проинтерпретируйте полученные результаты.
22. В предложенном вам файле Employees закодируйте переменную Age тремя значениями (молодые, средний возраст, зрелый возраст). Проверьте с помощью дисперсионного анализа влияние возраста на заработок (“SALARY”). Проинтерпретируйте полученные результаты.
23. С помощью дисперсионного анализа проверьте влияние «диагональ» на «цену». Файл «LCD». Проинтерпретируйте полученные результаты.
24. С помощью дисперсионного анализа проверьте влияние «TCO» на «цену». Файл «LCD». Проинтерпретируйте полученные результаты.
25. В предложенном вам файле «Mydata» закодируйте переменную «run» тремя значениями (маленький, средний, большой). Проверьте с помощью дисперсионного анализа влияние пробега на цену. Проинтерпретируйте полученные результаты.

NUM	номер страны в базе данных Summers, Heston (1988);
NOIL (dummy)	1 для страны, не добывающей нефть, 0 - для добывающей;
INTER (dummy)	1 для страны с хорошим качеством данных, 0 - в противном случае;
OECD (dummy)	1 для страны
GDP60	ВВП на душу населения в 1960 г.;
GDP85	ВВП на душу населения в 1985 г.;
GDPGRO	средний рост ВВП на душу населения с 1960 г. по 1985 г. (в %);
POPGRO	средний рост работоспособного населения с 1960 г. по 1985 г. (в %);
IONY	средняя доля инвестиций (включая государственные) в общем объеме ВВП с 1960 г. по 1985 г. (в %);
SCH	средняя доля работоспособного населения, имеющая полное среднее образование с 1960 г. по 1985 г. (в %);
LIT	доля людей среди населения старше 15 лет, умеющих читать и писать в 1960 г.

Лабораторная работа №2

1. В предложенном файле содержатся результаты четырех тестов для пожилых людей: arith – арифметический тест, info – информационный тест, picture - тест дополнения картинок, similars – тест на подобие. Провести факторный анализ на основе данной выборки. Оставить оптимальное количество факторов, оценить качество восстановления корреляционной матрицы, используя графический анализ и методы вращения дать интерпретацию полученных факторов через исходные переменные.

Лабораторная работа №3

Выполнить следующее задание по дискриминантному анализу в пакете прикладных программ Statistica :

1. Проверить предпосылки дискриминантного анализа для предложенных данных:
   1.1 . Нормальность распределения признаков внутри групп.
   1.2 . Совпадение ковариационных матриц во всех группах.

2. Построить матрицы межгруппового и внутригруппового рассеивания.
3. Построить дискриминантные функции.
4. По стандартизованным коэффициентам проинтерпретировать существенность признаков с точки зрения различия между классами.
5. По структурным коэффициентам проинтерпретировать существенность признаков с точки зрения различия между классами.
6. Оценить требуемое количество дискриминантных функций по критерию χ^2.
7. Оценить качество дискриминантных функций по расположению классов в пространстве дискриминантных функций.
8. Построить классифицирующие функции Фишера.
9. Оценить качество классификации по классифицирующим функциям Фишера.
10. Оценить апостериорные вероятности принадлежности объектов к классам, априорные вероятности задать самостоятельно.
11. Определить принадлежность нового объекта к одному из трех классов.

Лабораторная работа №4

1. В таблице 1, приведенной ниже приведены данные по доходам населения за период с 1999 по 2001 год и сведения о приросте сбережений на вкладах и в ценных бумагах и о расходах на покупку валюты за тот же период.

Таблица 1.

	Доходы населения, млрд. р.	Прирост сбережений во вкладах и ценных бумагах, млрд. р.	Расходы на покупку валюты, млрд. р.
Январь 1999	166,2	4,3	13,8
Февраль 1999	186	8,2	13,6
Март 1999	197,9	4,4	21,8
Апрель 1999	220,5	9,3	15
Май 1999	212,5	8,3	13,6
Июнь 1999	226,5	10,2	17,4
Июль 1999	226,6	8,8	21,3
Август 1999	239,1	5,5	22,2
Сентябрь 1999	239,8	6,2	20,4
Октябрь 1999	250,8	7	18,1
Ноябрь 1999	257	8,2	21,8
Декабрь 1999	354,9	17,4	27,7
Январь 2000	215	8,6	17,2
Февраль 2000	261,3	12,3	17,5
Март 2000	286,5	13,2	22,6
Апрель 2000	291,5	11,1	18,4
Май 2000	284,5	15,6	16,5
Июнь 2000	315,1	17	18,3
Июль 2000	308,1	13,2	20,6
Август 2000	322,7	9,4	23,9
Сентябрь 2000	331,5	10,9	22,9

Октябрь 2000	325,5	7,8	24,7
Ноябрь 2000	348,5	15	22,3
Декабрь 2000	452,3	8,1	28,9
Январь 2001	290,2	13,3	20,5
Февраль 2001	337,5	12,8	20,2
Март 2001	376,1	15	21,3
Апрель 2001	395,4	17	21,2
Май 2001	372,1	11,2	22,6
Июнь 2001	428,2	17,1	23,7
Июль 2001	424,9	14,9	26,7
Август 2001	437,2	16,2	29
Сентябрь 2001	436,1	20,5	22,6
Октябрь 2001	438,6	17,5	26,2
Ноябрь 2001	448,3	17,9	31,3
Декабрь 2001	580,6	22,4	35,4

Методом экспоненциального сглаживания постройте модель одного из предложенных временных рядов. Оцените качество модели и сделайте прогноз на несколько шагов вперед.

2. Проведите сезонную декомпозицию для ряда Series_G, расположенного в папке Examples. Укажите: имеет ли предложенный временной ряд тренд, сезонную компоненту, вид сезонности, период сезонности. Проанализируйте, является ли случайная компонента данного ряда стационарным рядом.

Критерии оценки лабораторных работ:

- оценка «отлично» выставляется студенту, если все задания выполнены;
- оценка «хорошо» выставляется студенту, если все задания выполнены, но возможно, с некоторыми недочетами
- оценка «удовлетворительно» выставляется студенту, если задания выполнены частичнов и (или) с недочетами.
- оценка «неудовлетворительно», если выполнено меньше 50 % задания.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: тест, курсовая работа

Тест.

Тест используется для оценки знаний студентов на экзамете и представляет контрольно-измерительный материал промежуточной аттестации, позволяющий оценить степень сформированности знаний, умений и навыков.

Для оценивания результатов обучения на экзамене используются следующие показатели:

1) знание учебного материала и владение понятийным аппаратом статистического анализа данных;
2) умение анализировать многомерные данные;
3) умение применять методы анализа данных при решении задач в различных прикладных областях;
4) владение навыками построения и проверки качества моделей анализа данных;
7) владение навыками интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов.

Контрольно-измерительный материал №_1

1. Проинтерпретируйте основные статистические показатели, используемые в описательной статистике: *Minimum, maximum, Range, Медиана, Доверительный интервал, Variance, Lower, Upper Quartile, Std.Dev, Skewness, Kurtosis*. (1 балл).

2. Опишите t-тест для зависимых и независимых выборок. Приведите примеры задач на использование данных методов. (2 балл).
3. Приведите пример задачи на использование рангового дисперсионного анализа. (1 балл)
4. Опишите метод решения задачи дискриминации в дискриминантном анализе. (1 балл)
5. Опишите метод интерпретации факторов в факторном анализе (1 балл).
6. Сформулируйте алгоритм метода древовидной классификации. (2 балла).
7. Дайте определение временного ряда. Нарисуйте пример графика временного ряда. (1 балл)
8. Охарактеризуйте основные компоненты временных рядов. (1 балл).
9. Дайте определение автокорреляционной функции для стационарных временных рядов. (1 балл).
10. Опишите метод восходящих и нисходящих серий для проверки гипотезы о наличии тренда у ряда. (1 балл)
11. Сформулируйте алгоритм сглаживания скользящими средними для временных рядов. (2 балла).
12. Опишите подробно метод экспоненциального сглаживания для простейших временных рядов. (2 балла).

Контрольно-измерительный материал №_2

1. Проинтерпретируйте основные статистические показатели, используемые в описательной статистике: *Minimum, maximum, Range, Медиана, Доверительный интервал, Variance, Lower, Upper Quartile, Std.Dev, Skewness, Kurtosis*. (1 балл).

2. Методы проверки гипотезы о нормальном распределении признака. (1 балл).

3. Опишите общий принцип проверки статистических гипотез. (1 балл)
4. Опишите принцип построения таблиц сопряженности для проверки гипотезы о независимости двух категорированных переменных. Приведите пример задачи на использование данного метода. (2 балл).
3. Как определяется количество дискриминантных функций в дискриминантном анализе. (1 балл)
4. Опишите последовательность шагов факторного анализа (1 балл).
6. Сформулируйте алгоритм метода к-средних. (2 балла).
7. Дайте определение временного ряда. Нарисуйте пример графика временного ряда. (1 балл)
8. Охарактеризуйте основные компоненты временных рядов. (1 балл).
9. Дайте определение автокорреляционной функции для стационарных временных рядов. (1 балл).
10. Опишите метод сезонной декомпозиции временных рядов. (2 балла)
11. Сформулируйте алгоритм сглаживания скользящими средними для временных рядов. (2 балла).
12. Опишите подробно метод экспоненциального сглаживания для простейших временных рядов. (2 балла).

Для оценивания результатов обучения на экзамене используется 4-балльная шала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».
Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания	Шкала оценок
Обучающийся в полной мере владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), сдал все практические и лабораторные работы, среднее количество правильных ответов на вопросы тестов превышает 80%.	*Отлично*
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), но не сдал одну практическую или лабораторную работу, среднее количество правильных ответов на вопросы тестов находится в диапазоне 70-80%.	*Хорошо*
Обучающийся демонстрирует неуверенное владение понятийным аппаратом данной области науки (теоретическими основами дисциплины), не сдал две практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов находится в диапазоне 60-70%.	*Удовлетвори-тельно*
Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не сдал более двух практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов менее 70%.	*Неудовлетвори-тельно*

Курсовая работа – вид контрольно-измерительного материала промежуточной аттестации, позволяющий оценить способность студентов использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач.

Образец курсовой работы:

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

Факультет прикладной математики, информатики и механики
Кафедра математических методов исследования операций

«Анализ данных»

Курсовая работа

Направление 38.04.05 «Бизнес-информатика»

Зав. кафедрой:
Руководитель:
Выполнил:

Воронеж

Содержание

Введение

Обработка числовой информации в наши дни немыслима без применения компьютера. Современный специалист обязан обладать навыками компьютерной обработки данных и иметь представление о программном обеспечении, с помощью которого ее можно выполнять. Сегодня существует большое количество специализированных приложений для статистического анализа. Одним из несомненных лидеров среди таких продуктов признана программа STATISTICA фирмы StatSoft, Inc., США. Помимо очень мощного набора процедур статистического и графического анализа, эта программа обладает весьма дружественным интерфейсом, что делает ее достаточно легкой для освоения и удобной в работе.

Цель работы: провести анализ данных с помощью Statistica 10.

Задачи:

1. Изучить теоретические основы курса «Анализ данных
2. Изучить основы компьютерной обработки данных в пакете Statistica;
3. Получить описательные статистики;
4. Произвести дисперсионный, факторный, дискриминантный анализ;
5. Построить регрессионную модель, таблицы сопряженности.

1. Описание исходного набора данных

Исходный набор данных - таблица с данными об алмазах. Этот классический набор данных содержит цены и другие атрибуты почти 54 000 алмазов.

Исходные показатели описаны в таблице 1.

Название столбца	**Содержимое**
price	Цена в долларах США от $326 до $18 823
carat	Вес алмаза от 0,2 до 5,01
cut	Качество среза: 1 – ясное *(Fair);* 2- хорошее *(Good);*

	3 – очень хорошее *(Very Good);* 4 -премиум *(Premium);* 5 – идеальное *(Ideal).*
Color diamond	Цвет алмаза от J (худший) до D (лучший)
clarity	Ясность алмаза: **I1** - внутреннее включение 1 *(Included 1):* высокий уровень включений; **SI2** - мелкие внутренние включения 2 *(Slightly Included 2):* 85% азмазов этого уровня чистоты имеют видимые невооруженным глазом включения; **SI1** - мелкие внутренние включения 1 *(Slightly Included 1):* только 50% алмазов этого уровня чистоты не имеют видимых включений, бриллиант чист на глаз; **VS2** - очень мелкие внутренние включения 2 *(Very Slightly Included 2):* алмазы этого уровня чистоты будут иметь несколько включений, различимых лишь при 10-кратном увеличении. Некоторые включения видны невооруженным глазом; **VS1** - очень мелкие внутренние включения 1 *(Very Slightly Included 1):* алмазы этого уровня чистоты будут иметь несколько включений, различимых лишь при 10-кратном увеличении, но неразличимых невооруженным глазом; **VVS2** - мельчайшие внутренние включения 2 *(Very Very Slightly Included 2)*: внутри алмаза этого уровня чистоты будут незначительные включения, видимые лишь при 20-кратном и более увеличении; **VVS1** - мельчайшие внутренние включения 1 *(Very Very Slightly Included 1):* внутри алмаза этого уровня

	чистоты будет одно незначительное включение, видимое лишь при 20-кратном и более увеличении; **IF** - внутренне чистые (Internationally Flawless): редкие алмазы самого высокого уровня чистоты. Они на 100% чисты, в них нет никаких включений.
x	Длина в мм (от 0 до 10.74)
y	Ширина в мм (от 0 до 58,9)
z	Глубина в мм (от 0 до 31.8)
depth	Общая глубина в процентах $D = \frac{z}{\text{среднее значение } (x,y)}$
table	Ширина стола верхней части алмаза относительно самой широкой точки (от 43 до 95)

Табл.1. Исходные показатели

Данные выглядят следующим образом (Рис.1.):

	1 carat	2 cut	3 color	4 clarity	5 depth	6 table	7 price	8 x	9 y	10 z
1.000000	0.23	5	E	SI2	62	55	326	3.95	3.98	2.43
2.000000	0.21	4	E	SI1	60	61	326	3.89	3.84	2.31
3.000000	0.23	2	E	VS1	57	65	327	4.05	4.07	2.31
4.000000	0.29	4	I	VS2	62	58	334	4.2	4.23	2.63
5.000000	0.31	2	J	SI2	63	58	335	4.34	4.35	2.75
6.000000	0.24	3	J	VVS2	63	57	336	3.94	3.96	2.48
7.000000	0.24	3	I	VVS1	62	57	336	3.95	3.98	2.47
8.000000	0.26	3	H	SI1	62	55	337	4.07	4.11	2.53
9.000000	0.22	1	E	VS2	65	61	337	3.87	3.78	2.49
10.000000	0.23	3	H	VS1	59	61	338	4	4.05	2.39
…										
53935.000000	0.72	4	D	SI1	63	59	2757	5.69	5.73	3.58
53936.000000	0.72	5	D	SI1	61	57	2757	5.75	5.76	3.5
53937.000000	0.72	2	D	SI1	63	55	2757	5.69	5.75	3.61
53938.000000	0.7	3	D	SI1	63	60	2757	5.66	5.68	3.56
53939.000000	0.86	4	H	SI2	61	58	2757	6.15	6.12	3.74
53940.000000	0.75	5	D	SI2	62	55	2757	5.83	5.87	3.64

Рис.1. Исходные данные в пакете Statistica

2. Практическая часть

2.1. Описательные статистики

Рассчитаем описательные статистики для всех показателей. Результаты представлены на рис. 2.

Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile	Range	Std.Dev.	Skewness	Kurtosis
	Descriptive Statistics (Worksheet in diamonds2)												
carat	53940	0.798	0.700	0.300000	2604	0.2000	5.01	0.4000	1.040	4.81	0.474	1.116646	1.25664
cut	53940	3.904	4.000	5.000000	21551	1.0000	5.00	3.0000	5.000	4.00	1.117	-0.717180	-0.39797
color	53940	104.174	105.000	106.0000	11292	101.0000	107.00	102.0000	106.000	6.00	2.050	-0.315877	-1.23369
clarity	53940	103.256	103.000	102.0000	13065	101.0000	108.00	102.0000	104.000	7.00	1.767	0.692061	0.00932
depth	53940	61.799	62.000	62.00000	19849	43.0000	79.00	61.0000	63.000	36.00	1.465	-0.070331	5.32619
table	53940	57.458	57.000	56.00000	10039	43.0000	95.00	56.0000	59.000	52.00	2.234	0.797126	2.80212
price	53940	3932.800	2401.000	605.0000	132	326.0000	18823.00	950.0000	5324.500	18497.00	3989.440	1.618395	2.17770
x	53940	5.731	5.700	4.370000	448	0.0000	10.74	4.7100	6.540	10.74	1.122	0.378676	-0.61816
y	53940	5.735	5.710	4.340000	437	0.0000	58.90	4.7200	6.540	58.90	1.142	2.434167	91.21456
z	53940	3.539	3.530	2.700000	767	0.0000	31.80	2.9100	4.040	31.80	0.706	1.522423	47.08662

Рис. 2. Описательные статистики

В этой таблице рассчитаны все описательные статистики. Объясним их значение на примере price (цены):

1. **Valid N –** объем выборки (53940);
2. **Mean –** арифметическая средняя, т.е. средняя цена алмаза = 3932,8$;
3. **Median –** медиана, такое значение случайной величины, для которого выполняется равенство: $P(X > MeX) = P(x < MeX) = 1/2$. В нашем случае, 50% алмазов стоят дороже 2401$, а остальные 50% дешевле этой цены;
4. **Mode –** мода, наиболее вероятное значение изучаемой характеристики. Таким образом, мода для характеристики price = 605$;
5. **Minimum/Maximum –** минимальное/максимальное значение изучаемой характеристики. Соответственно, минимальная цена - 326$, а максимальная цена - 18823$;
6. **Lower/Upper Quartile –** нижний и верхний квартили - такое значение анализируемой величины, что вероятность попасть левее/правее этого значения равна p = 0,25/0,75, соответственно. Таким образом, с вероятностью 25% цена алмаза будет меньше 950 $ и с вероятностью 75% - больше 5324$;
7. **Range –** размах, т.е. разброс между минимальным и максимальным значением. Расстояние между минимальной (326$) и максимальной (18823$) ценой алмаза будет 18497;
8. **Std. Dev. –** стандартное отклонение - 3989,44;
9. **Skewness -** коэффициент асимметрии. У нормального закона распределения коэффициент асимметрии $\gamma=0$. В нашем примере $\gamma=1,61$, то есть распределение не симметрично, а имеет «хвост», вытянутый вправо.

10. **Kurtosis** – эксцесс, показатель, характеризующий островершинность распределения. За эталон берется нормальное распределение, где E = 0. Для цены на алмазы E = 2,17, то есть распределение островершинное.

2.2. Дисперсионный анализ

В дисперсионном анализе проверяется гипотеза о равенстве нескольких генеральных средних (математических ожиданий) H_0: $M(X1) = M(X2) = ... = M(Xn)$. Перед началом дисперсионного анализа мы убедились, что данные распределены нормально, согласно тесту Колмогорова-Смирнова (p>0,2) и глазомерному методу.

Выдвигаем гипотезу:

H_0: разница в ценах (price) между качествами среза (cut) несущественна;

H_1: разница в ценах (price) между качествами среза (cut) существенна;

Таким образом, группирующей переменное является качество среза, а зависимой – цена (рис. 3).

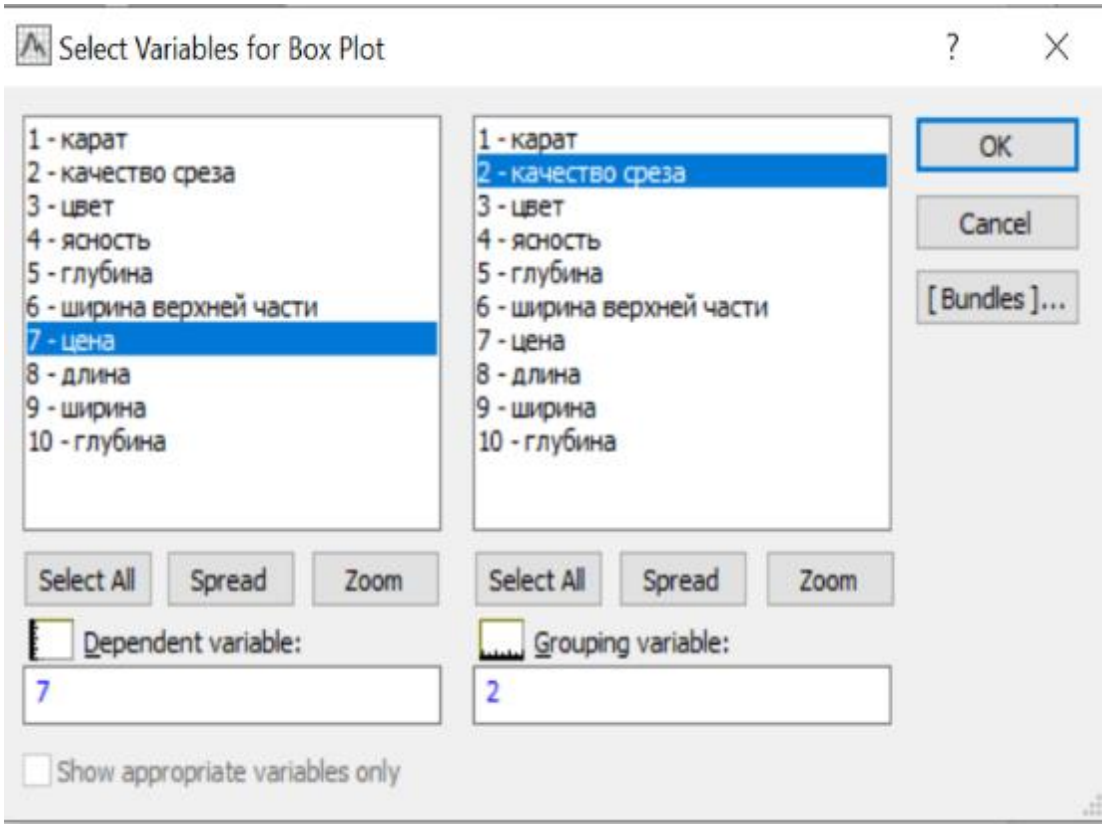

Рис. 3. Выбор переменных для дисперсионного анализа ANOVA

Посмотрим, каким образом распределились данные, при помощи Box Plot (рис. 4).

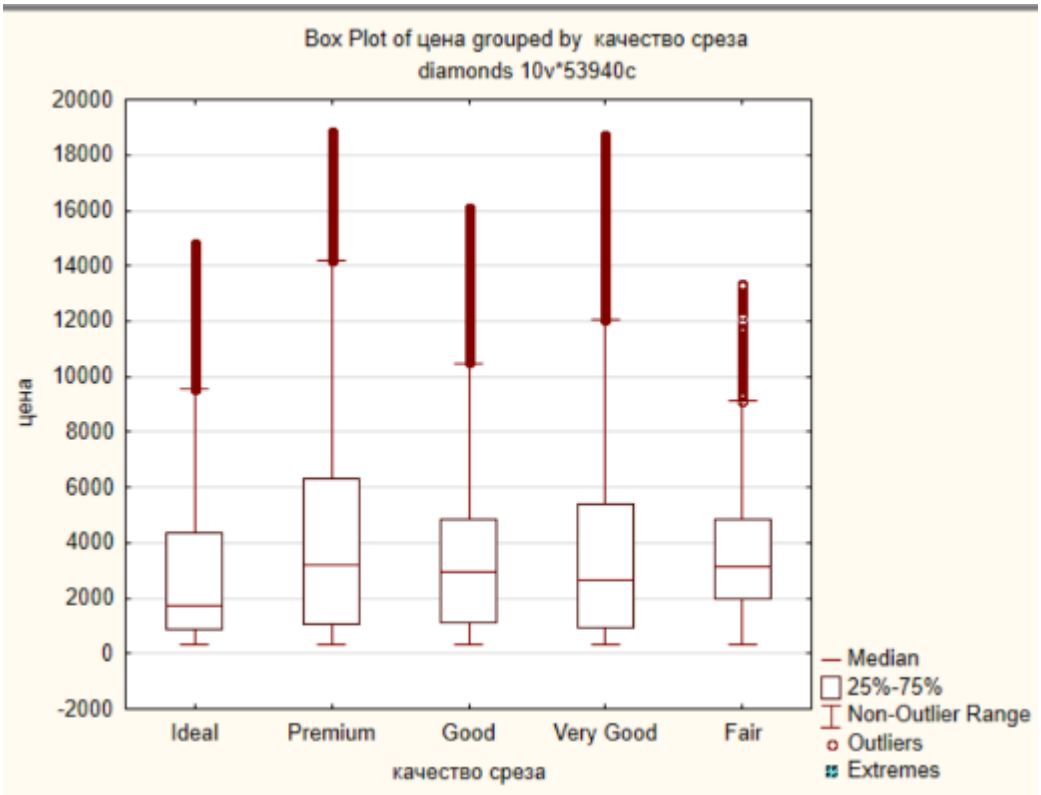

Рис. 4. Box Plot для цены и качества среза

По графику видно, что разница между средними значениями цен у разных качеств среза есть, поэтому «ящики с усами» расположились в соответствии со своим средним значением не симметрично. На графике хорошо заметны диапазоны цен: например, размах у качества среза “Premium” самый большой.

Analysis of Variance (diamonds)
Marked effects are significant at p < .05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
цена	1.883219E+10	4	4.708047E+09	7.206390E+11	53223	13539992	347.7142	0.00

Рис. 5. Анализ вариант

Согласно рис.5, а именно p <0,05, подтверждаем то, что различия срезов в цене имеются. Чтобы понять, в каких парах различия существенны, а где незначительны, проведем Post-hoc тест.

Breakdown Table of Descriptive Statistics (diamonds) N=53228 (No missing data in dep. var. list)		
качество среза	цена Means	цена N
Ideal	3110.311	21002
Premium	4584.258	13791
Good	3686.242	4819
Very Good	3975.630	12077
Fair	3841.131	1539
All Grps	3761.806	53228

Рис. 6. Подсчет количества переменных (столбец N)

Т.к. у нас разное количество переменных в каждой из категорий срезов (рис.6), будем использовать tukey HSD tor unequal N. Результаты представлены на рис. 7.

	Unequal N HSD; Variable: цена (diamonds) Marked differences are significant at p < .05000				
качество среза	**{1} M=3110.3**	{2} M=4584.3	{3} M=3686.2	{4} M=3975.6	{5} M=3841.1
Ideal {1}		0.000017	0.000017	0.000017	0.000018
Premium {2}	0.000017		0.000017	0.000017	0.000017
Good {3}	0.000017	0.000017		0.001081	0.769902
Very Good {4}	0.000017	0.000017	0.001081		0.849088
Fair {5}	0.000018	0.000017	0.769902	0.849088	

Рис. 7. Tukey HSD tor unequal N

Мы видим, что попарно цены срезов Ideal {1}, Premium {2}, Good {3}, Very Good {4} и Fair {5} существенных различий не имеют. Кроме пар Very Good {4} и Fair {5}, Good {3} и Fair {5}, , цена на которые существенно различима. Это видно и визуально на графике *Box Plot* (рис. 4).

Чтобы выразить эту разницу количественно мы должны найти разницу между средними. Например, {2} - {1} = 4584, 26 – 3110,31 = 1473, 95.

Таким образом, гипотеза H_0 подтверждена, хотя пары Good {4} - Fair {5}, Good {3} - Fair {5} ее опровергают в пользу гипотезы H_1. Это объясняет тем, что цена зависит не только от качества среза, но и от других факторов.

2.3. Факторный анализ

Задачей факторного анализа является объединение большого количества показателей, признаков, которыми характеризуется объект, в меньшее количество искусственно построенных на их основе факторов, чтобы полученная в итоге

система факторов была наиболее удобна с точки зрения содержательной интерпретации.

Отбираем все количественные переменные из файла данных, которые должны быть включены в факторный анализ. В нашем случае, это переменные carat, cut, depth, table, price, x, y, z. И начинаем анализ выбранных переменных.

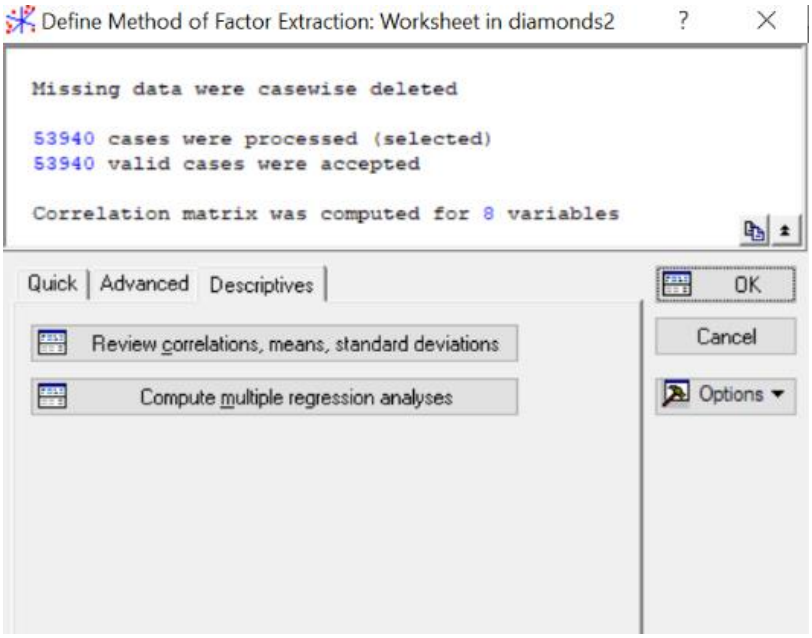

Рис. 8. Начало факторного анализа

Данное окно имеет следующую структуру. Верхняя часть окна является информационной: здесь сообщается, что пропущенные значения обработаны методом Casewise. Обработано 53940 случаев и 53940 случаев приняты для дальнейших вычислений. Корреляционная матрица вычислена для 8 переменных.

Построим корреляционную матрицу выбранных переменных (рис.9).

Correlations (Worksheet in diamonds2)
Casewise deletion of MD
N=53940

Variable	carat	cut	depth	table	price	x	y	z
carat	1.00	-0.13	0.03	0.18	0.92	0.98	0.95	0.95
cut	-0.13	1.00	-0.22	-0.43	-0.05	-0.13	-0.12	-0.15
depth	0.03	-0.22	1.00	-0.30	-0.01	-0.03	-0.03	0.09
table	0.18	-0.43	-0.30	1.00	0.13	0.20	0.18	0.15
price	0.92	-0.05	-0.01	0.13	1.00	0.88	0.87	0.86
x	0.98	-0.13	-0.03	0.20	0.88	1.00	0.97	0.97
y	0.95	-0.12	-0.03	0.18	0.87	0.97	1.00	0.95
z	0.95	-0.15	0.09	0.15	0.86	0.97	0.95	1.00

Рис. 9. Корреляционная матрица

Мы видим, что наблюдаются значения корреляции больше, чем 0,8, следовательно некоторые факторы зависят друг от друга. Но большинство факторов не коррелируют друг с другом.

Выведем результаты факторного анализа (рис. 10):

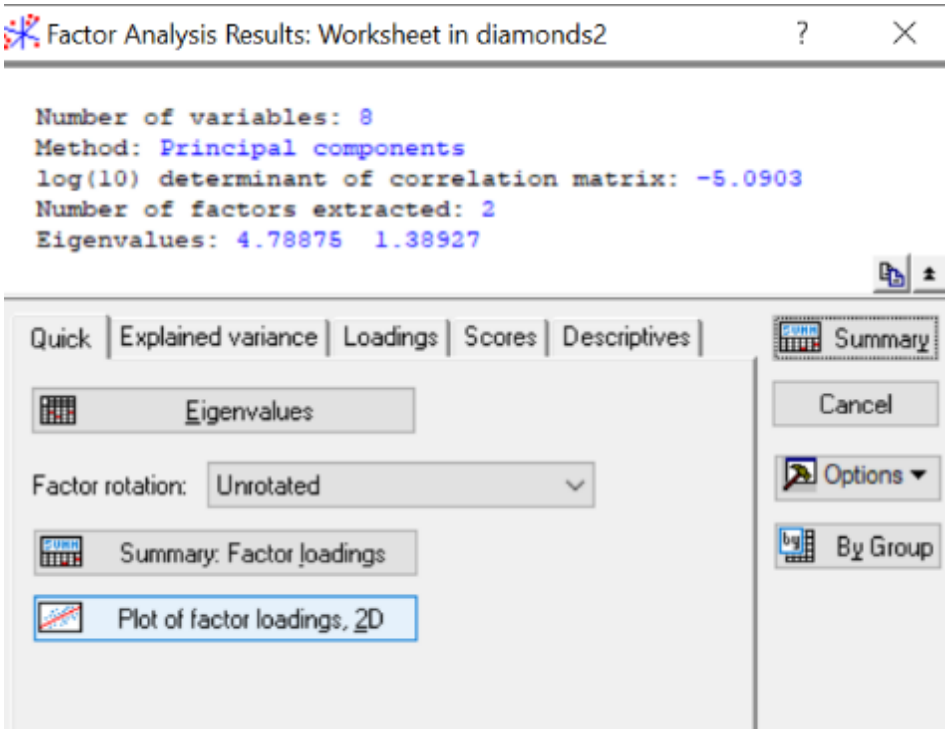

Рис. 10. Результаты факторного анализа

В верхней части окна Результаты факторного анализа дается информационное сообщение:

- **Number of variables** (число анализируемых переменных) – 8;
- **Method** (метод анализа) – главные компоненты;
- **log(10) determination of correlation matrix** (десятичный логарифм детерминанта корреляционной матрицы) – -5,0903;
- **Number of Factor extraction** (число выделенных факторов) – 2;
- **Eigenvalues** (собственные значения) – 4,78875; 1,38927.

Построим график каменистой осыпи (рис. 11).

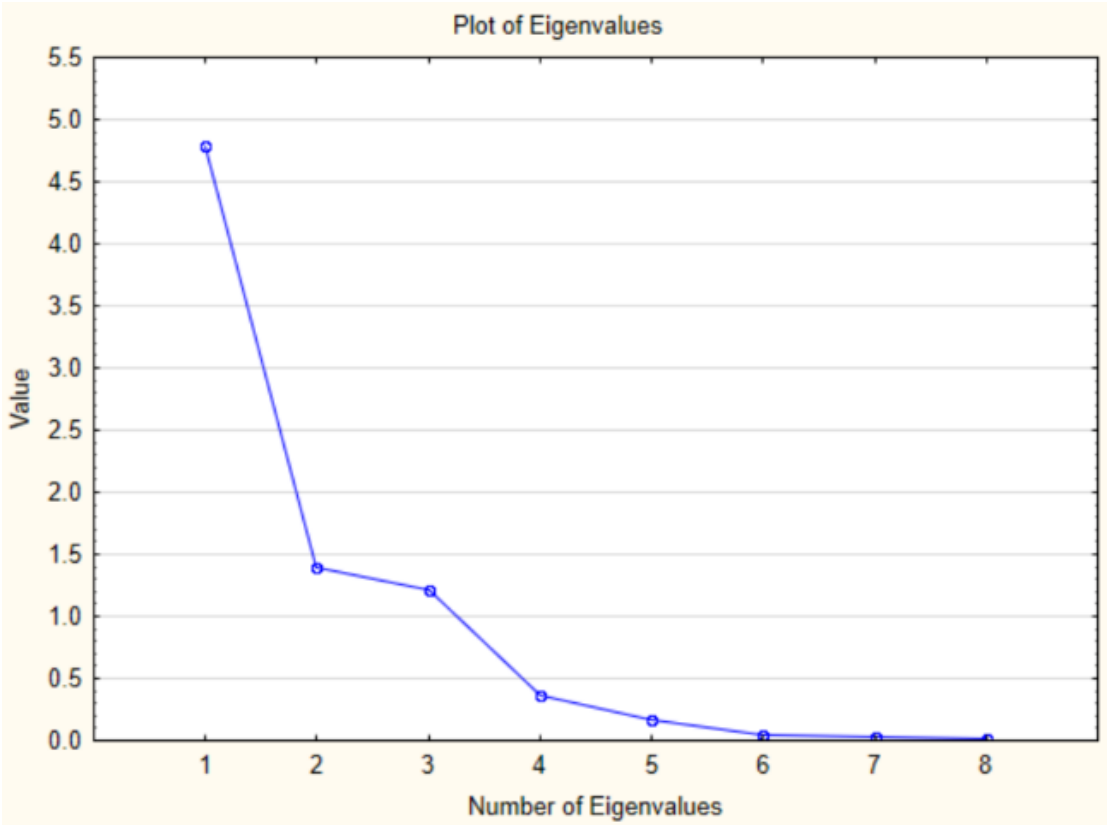

Рис. 11. График каменистой осыпи

В точках с координатами 1, 2 осыпание замедляется наиболее существенно, следовательно, теоретически можно ограничиваться двумя факторами.

На рис. 12 мы видим таблицу с текущими факторными нагрузками. Факторные нагрузки могут интерпретироваться как корреляции между соответствующими переменными и факторами – чем выше нагрузка по модулю, тем больше близость фактора к исходной переменной.

Factor Loadings (Unrotated) (Worksheet in diamonds2)
Extraction: Principal components
(Marked loadings are >.700000)

Variable	Factor 1	Factor 2
carat	-0.986602	-0.058150
cut	0.177836	-0.725199
depth	-0.006868	-0.270802
table	-0.235113	0.872984
price	-0.924841	-0.119937
x	-0.987891	-0.042271
y	-0.974794	-0.047317
z	-0.973105	-0.078286
Expl.Var	4.788750	1.389268
Prp.Totl	0.598594	0.173659

Рис. 11. Таблица факторных нагрузок

Соответственно, первый фактор более коррелирует с переменными, чем второй. Поскольку количество значений первого фактора больших, чем 0,7, больше, чем у второго.

Их трудно проинтерпретировать, возникает вопрос, какой смысл придать второму фактору. В этом случае целесообразно прибегнуть к повороту осей, надеясь получить решение, которое можно интерпретировать в предметной области. Построим график нагрузок (рис. 12).

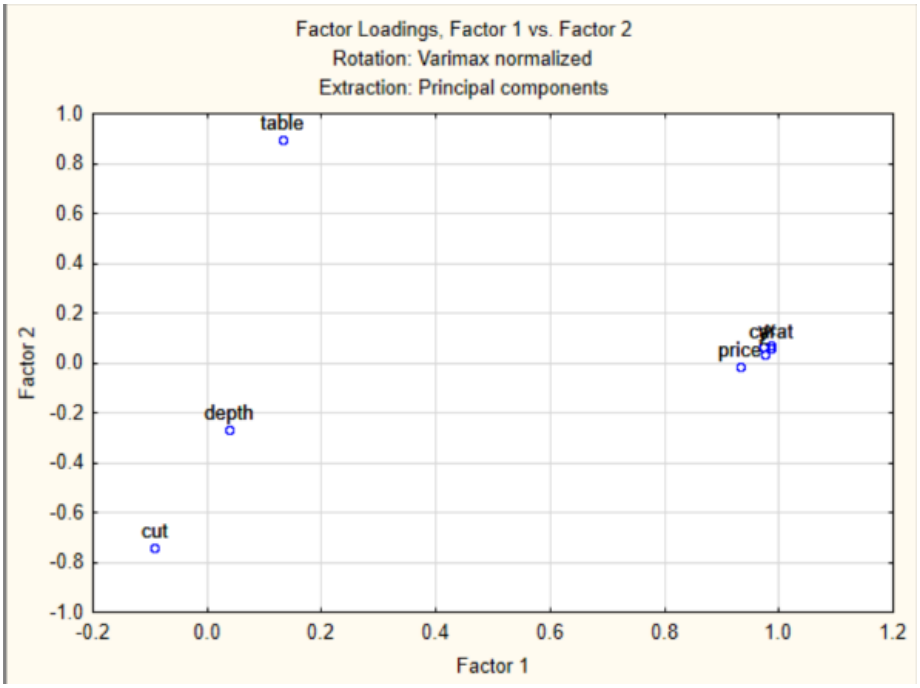

Рис. 12. График факторных нагрузок (varimax normalized)

Конечно, этот график отличается от предыдущего. Посмотрим еще нагрузки численно (рис. 13).

Factor Loadings (Varimax normalized) (Worksheet in diamonds2)
Extraction: Principal components
(Marked loadings are >.700000)

Variable	Factor 1	Factor 2
carat	0.986727	0.055989
cut	-0.093038	-0.740866
depth	0.038045	-0.268205
table	0.132894	0.894270
price	0.932502	-0.012507
x	0.986177	0.071911
y	0.973749	0.065388
z	0.975641	0.034431
Expl.Var	4.743561	1.434458
Prp.Totl	0.592945	0.179307

Рис. 13. Таблица факторных нагрузок(varimax normalized)

Теперь найденное решение уже можно интерпретировать. Факторы чаще интерпретируют по нагрузкам. Первый фактор теснее всего связан с carat, price, x, y, z Второй фактор – cut и table. Таким образом, мы произвели классификацию переменных на две группы.

Согласно графику каменистой осыпи (рис. 11), использование двух факторов – оптимально для решения задачи с алмазами. Полученная в итоге система факторов является наиболее удобной, с точки зрения содержательной интерпретации.

Дискриминантный анализ

Задача состоит в том, чтобы по исходных данным классифицировать алмазы по качеству среза.

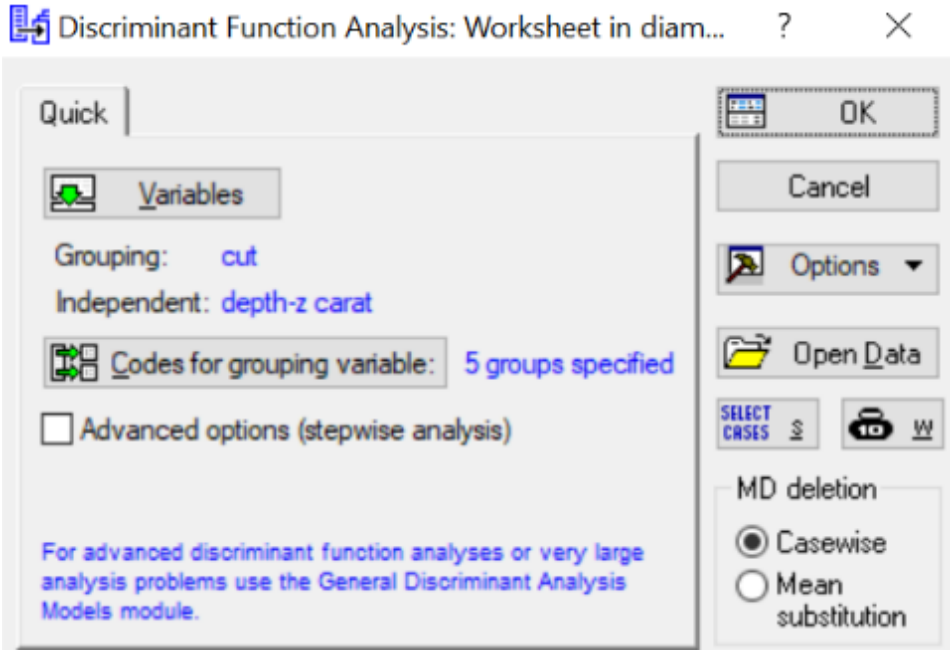

Рис. 14. Начало дискриминантного анализа

В качестве группируемой переменной выбираем cut, а в качестве независимых переменных carat , depth, table, price, x, y, z.

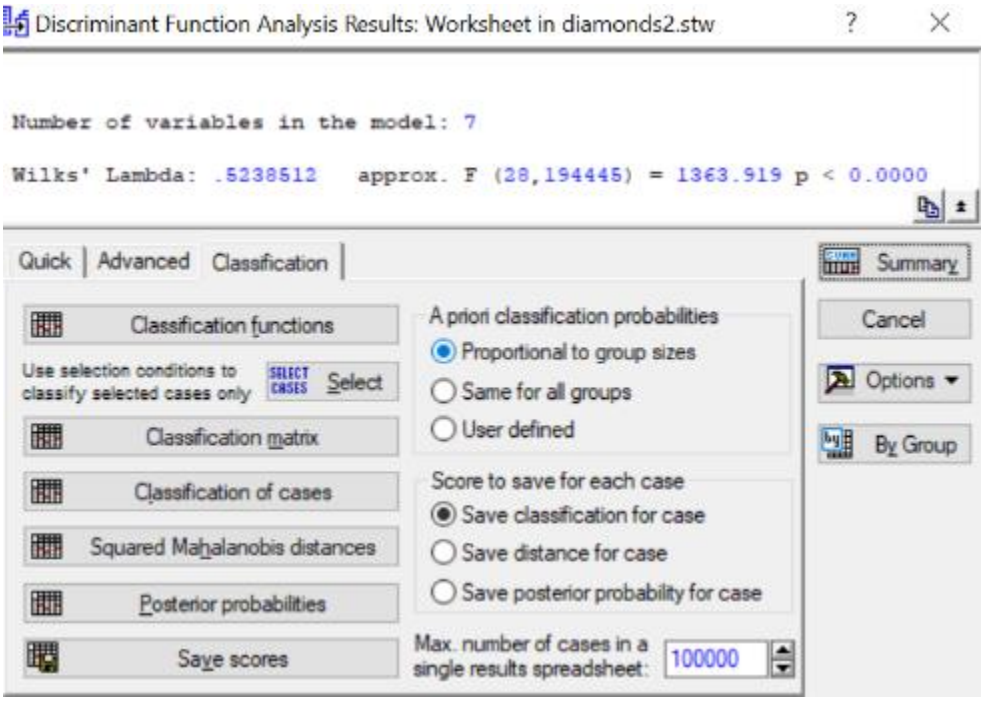

Рис. 15. Результаты анализа дискриминантных функций

На рис. 15 представлены результаты анализа дискриминантных функций:

- **Number of variables in the model** (число переменных в модели) - 7;
- **Wilks' Lambda** (значение лямбды Уилкса) – 0,524;
- **Approx. F** (28, 194) (приближенное значение F – статистики, связанной с лямбдой Уилкса) – 1363, 919;
- **p** < 0.0000 – уровень значимости F – критерия для значения 1363, 919.

Значение статистики Уилкса лежит в интервале [0,1]. Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации, а значения, лежащие около 1, свидетельствуют о плохой дискриминации. По данным показателя Wilks' Lambda (значение лямбды Уилкса), равного 0,524 и по значению F – критерия равного 1363, 919, можно сделать вывод, что данная классификация корректная.

В качестве проверки корректности обучающих выборок посмотрим результаты классификационной матрицы, нажав кнопку Classification matrix (Классификационная матрица) (рис. 16).

	Classification Matrix (Worksheet in diamonds2.stw) Rows: Observed classifications Columns: Predicted classifications					
Group	**Percent Correct**	G_1:1 p=.02985	G_2:2 p=.09095	G_3:3 p=.22399	G_4:4 p=.25567	G_5:5 p=.39954
G_1:1	60.68323	977	72	96	280	185
G_2:2	12.55605	520	616	1106	1330	1334
G_3:3	22.44662	129	445	2712	3742	5054
G_4:4	62.43927	53	158	2051	8611	2918
G_5:5	94.02812	12	8	416	851	20264
Total	61.51279	1691	1299	6381	14814	29755

Рис. 16. Классификационная матрица

Из классификационной матрицы можно сделать вывод, что не все объекты были правильно отнесены экспертным способом к выделенным группам. Есть ли переменные, неправильно отнесенные к соответствующим группам, можно посмотреть Classification of cases (Классификация случаев).

	Classification of Cases (Worksheet in diamonds2.stw) Incorrect classifications are marked with *					
Case	**Observed Classif.**	1 p=.02985	2 p=.09095	3 p=.22399	4 p=.25567	5 p=.39954
1.000000	G_5:5	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1
2.000000	G_4:4	G_4:4	G_3:3	G_5:5	G_2:2	G_1:1
*3.000000	G_2:2	G_4:4	G_3:3	G_2:2	G_5:5	G_1:1
*4.000000	G_4:4	G_3:3	G_5:5	G_4:4	G_2:2	G_1:1
*5.000000	G_2:2	G_3:3	G_4:4	G_2:2	G_5:5	G_1:1
*6.000000	G_3:3	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1
*7.000000	G_3:3	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1
*8.000000	G_3:3	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1
9.000000	G_1:1	G_1:1	G_2:2	G_3:3	G_4:4	G_5:5
*10.000000	G_3:3	G_4:4	G_3:3	G_5:5	G_2:2	G_1:1
…						
*53934.00000(	G_3:3	G_4:4	G_3:3	G_5:5	G_2:2	G_1:1
*53935.00000(	G_4:4	G_3:3	G_4:4	G_2:2	G_5:5	G_1:1
53936.000000	G_5:5	G_5:5	G_4:4	G_3:3	G_2:2	G_1:1
*53937.00000(	G_2:2	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1
*53938.00000(	G_3:3	G_2:2	G_4:4	G_3:3	G_1:1	G_5:5
53939.000000	G_4:4	G_4:4	G_5:5	G_3:3	G_2:2	G_1:1
53940.000000	G_5:5	G_5:5	G_3:3	G_4:4	G_2:2	G_1:1

Рис. 17. Классификация случаев

В таблице классификации случаев (рис. 17) некорректно отнесенные объекты помечаются звездочкой (*). Таким образом, задача получения корректных обучающих выборок состоит в том, чтобы исключить из обучающих выборок те объекты, которые по своим показателям не соответствуют большинству объектов, образующих однородную группу. В нашем случае, некорректно отнесенных объектов достаточное количество, что объясняет низкие значения коэффициента корректности. Поскольку количество ошибок велико, при выводе классификационных функций будем учитывать высокую погрешность.

Проведем классификацию объектов. Для этого вызываем метод Classification functions. Появится окно, из которого можно выписать классификационные функции для каждого класса (рис. 18).

	Classification Functions; grouping: cut (Worksheet in diamonds2.stw)				
Variable	G_1:1 p=.02985	G_2:2 p=.09095	G_3:3 p=.22399	G_4:4 p=.25567	G_5:5 p=.39954
depth	69.69	68.57	68.04	67.87	67.61
table	33.92	33.51	33.15	33.31	32.43
price	0.01	0.01	0.01	0.01	0.01
x	198.65	197.82	197.15	198.67	197.84
y	24.11	25.08	25.20	24.37	24.85
z	-166.45	-166.52	-166.02	-166.90	-166.45
carat	-329.88	-334.21	-334.72	-334.49	-335.79
Constant	-3439.13	-3341.21	-3284.75	-3284.02	-3216.74

Рис. 18. Классификационные функции

$$cut(5) = -3216{,}74 + 67{,}61 * depth + 32{,}43 * table + 0{,}01 * price + 197{,}84 * x + 24{,}85 * y\text{-}166{,}45 * x\text{-}335{,}79 * carat$$

$$cut(4) = -3284{,}02 + 67{,}87 * depth + 33{,}31 * table + 0{,}01 * price + 198{,}67 * x + 24{,}37 * y\text{-}166{,}90 * x\text{-}334{,}49 * carat$$

$$cut(3) = -3284{,}75 + 68{,}04 * depth + 33{,}15 * table + 0{,}01 * price + 197{,}15 * x + 25{,}20 * y\text{-}166{,}02 * x\text{-}334{,}72 * carat$$

$$cut(2) = -3341{,}21 + 68{,}57 * depth + 33{,}51 * table + 0{,}01 * price + 197{,}82 * x + 25{,}08 * y\text{-}166{,}52 * x\text{-}334{,}21 * carat$$

$$cut(1) = -3439{,}13 + 69{,}69 * depth + 33{,}92 * table + 0{,}01 * price + 198{,}65 * x + 24{,}11 * y\text{-}166{,}45 * x\text{-}329{,}88 * carat$$

С помощью этих функций можно будет в дальнейшем классифицировать новые случаи. Новые случаи будут относиться к тому классу, для которого классифицированное значение будет максимальное.

2.4. Множественная регрессия

Регрессионный анализ позволяет предсказать, чему в среднем будет равно значение одного признака при заданном значении другого признака.

Рассмотрим, как зависит цена алмаза от веса, общей глубины и ширины стола верхней части алмаза относительно самой широкой его точки.

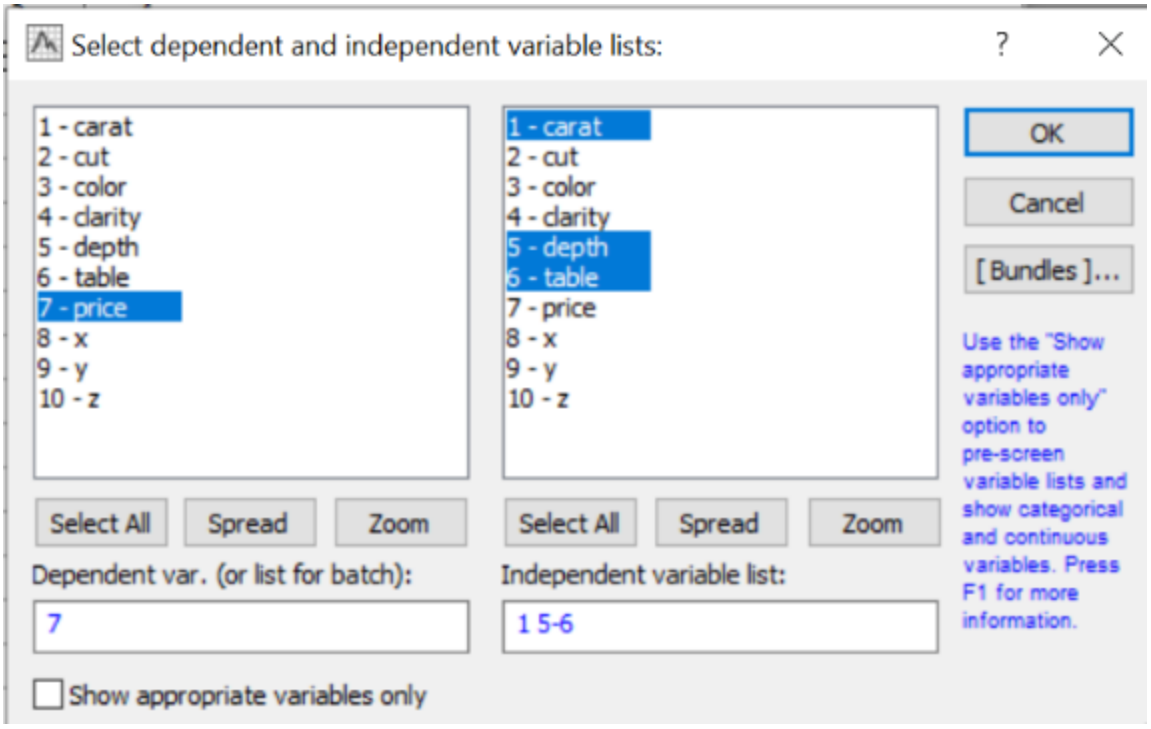

Рис. 19. Выбор переменных для множественной регрессии.

Рассчитаем корреляционную матрицу (рис. 20).

Correlations (Worksheet in diamonds2)

Variable	carat	depth	table	price
carat	1.000000	0.028224	0.181618	0.921591
depth	0.028224	1.000000	-0.295779	-0.010647
table	0.181618	-0.295779	1.000000	0.127134
price	0.921591	-0.010647	0.127134	1.000000

Рис. 20. Корреляционная матрица.

Мы видим, что не все факторы коррелируют друг с другом, т.к. не везде коэффициент парной корреляции больше 0,8.

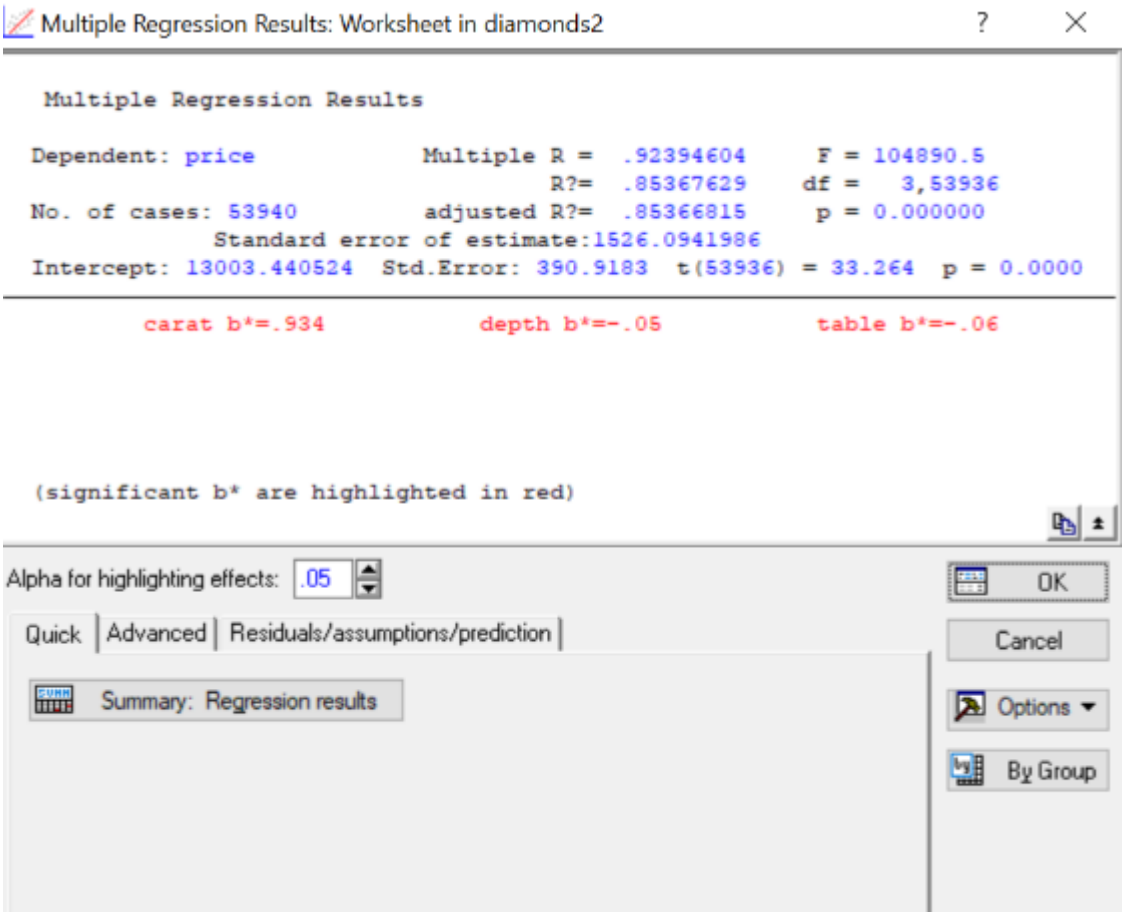

Рис. 20. Результаты регрессионного анализа

Интерпретируем результаты регрессионного анализа, представленные на рис. 20:

- **Dependent –** имя зависимой переменной – price;
- **No. of cases -** число наблюдений. N = 53940;
- **Multiple R –** коэффициент множественной корреляции. R=0,924
- **R?** - коэффициент детерминации. Он изменяется от 0 до 1 и отражает «качество» рассчитанной регрессии, показывая долю (%) общего разброса выборочных точек, которая «объясняется» построенной регрессией. Для наших данных R^2=0,85, что означает, что 85% дисперсии зависимой переменной price объясняется вариацией независимых переменных;
- **Adjusted R? –** скорректированный коэффициент детерминации. У нас от также равен 0,85.
- **F, df, p –** F-критерий, число степеней свободы, принятое при его расчете, и вероятность ошибки для нулевой гипотезы F-теста. F-тест в регрессионном анализе применяется для оценки статистической значимости модели. Т.к. p <

0,05 можно заключить, что рассчитанная регрессия удовлетворительно описывает связь между исследуемыми признаками;

- **Standart error od estimate –** параметр, отражающий степень разброса выборочных значений относительно линии регрессии = 1526,094;
- **Intercept –** значение свободного члена регрессионного уравнения = 13003, 441;
- **Std. Error –** стандартная ошибка свободного члена регрессионного уравнения = 390, 918;
- **T –** критерий Стьюдента t используется для проверки нулевой гипотезы о равенстве 0 свободного члена регрессионного уравнения. P – вероятность ошибки для этой нулевой гипотезы. T=33,264;
- **Beta -** стандартизованный коэффициент регрессии. Расчет beta коэффициентов позволяет оценить, в какой степени значения зависимой переменной определяются значениями независимой переменной. Таким образом, t(carat)=0,934, t(depth)=-0,05, t(table)= -0,6. Значит, большее влияние на результат оказывает фактор carat, т.к. его значение большее по модулю среди остальных факторов.

Построим таблицу регрессионных результатов (рис. 21). Обратим внимание на уровни значимости каждого фактора. Если данное значение больше, чем 0,05, то его следует исключить из дальнейшего рассмотрения в нашем анализе. Значит, исключаем факторы depth и table.

Regression Summary for Dependent Variable: price (Worksheet in diamonds2)
R= .92394604 R?= .85367629 Adjusted R?= .85366815
F(3,53936)=1049E2 p<0.0000 Std.Error of estimate: 1526.1

N=53940	b*	Std.Err. of b*	b	Std.Err. of b	t(53936)	p-value
Intercept			13003.44	390.9183	33.2638	0.00
carat	0.933752	0.001681	7858.77	14.1509	555.3558	0.00
depth	-0.054309	0.001731	-151.24	4.8199	-31.3776	0.00
table	-0.058515	0.001759	-104.47	3.1412	-33.2585	0.00

Рис. 21. Таблица регрессионных результатов

Выдвигаем гипотезу H_0: переменные carat, depth, table влияют на цену, соответственно, альтернативная гипотеза H_1: переменные carat, depth, table не влияют на цену.

Из рис.21 видно, что оба коэффициента регрессии не отличаются от 0 (P = 0,00), следовательно, принимается гипотеза H_0. В целом, построенная регрессионная модель хорошо описывает связь между ценой на алмазы, весом и общей глубиной (R^2 = 85%). Само же регрессионное уравнение мы можем записать следующим образом:

$$price = 7765,141*carat-151,24*depth-104,47*table-2256,36$$

Важной частью регрессионного анализа является анализ остатков (остатки представляют собой разности между наблюдаемыми значениями зависимой переменной и теми ее значениями, которые предсказываются регрессионной моделью).

Построим частотную гистограмму остатков (рис. 22).

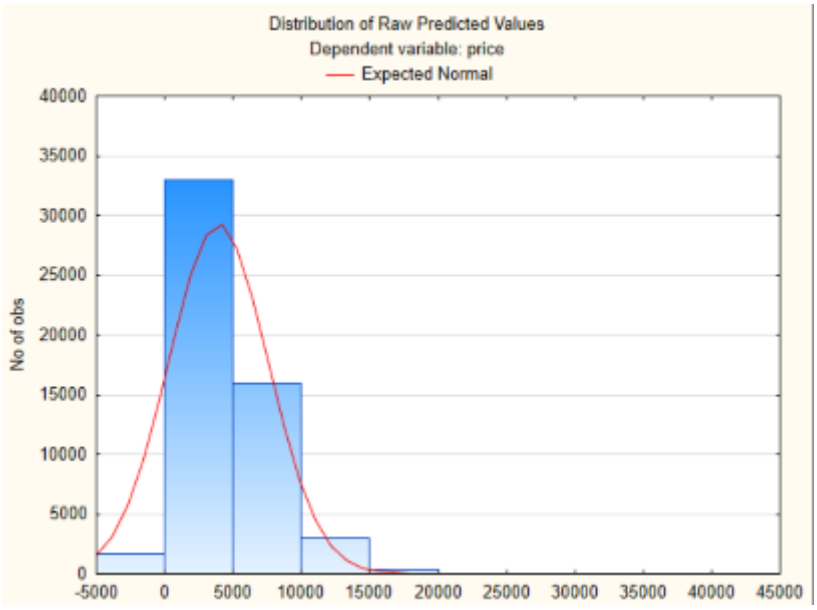

Рис. 22. Частотная гистограмма остатков

Используя глазомерный метод, мы можем утверждать, что гистограмма имеет нормальное распределение, хотя и видна небольшая асимметрия. Чтобы подтвердить или опровергнуть нормальность распределения остатков построим график нормальных вероятностей (рис. 23).

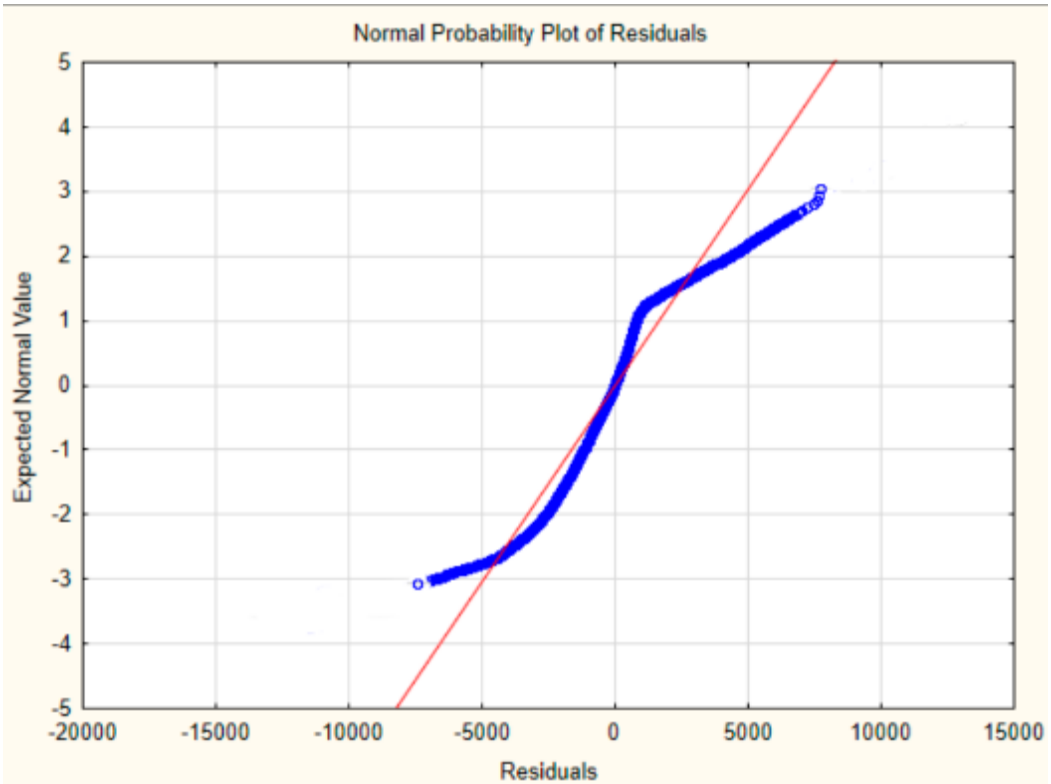

Рис. 23. Вероятностный график распределения остатков

На рис. 23 видны незначительные отклонения остатков от теоретической прямой нормального распределения. Учитывая, что большая часть прямой лежит в допустимых пределах, делает вывод, что линейный регрессионный анализ применим. И уравнение регрессии выведено верно.

2.5. Таблицы сопряженности

С помощью таблиц сопряженности выясним, есть ли зависимость между качеством среза и ясностью алмаза.

Выдвигаем гипотезу:

H_0: качество среза (cut) и ясность алмаза (clarity) зависимы;

H_1: качество среза (cut) и ясность алмаза (clarity) независимы;

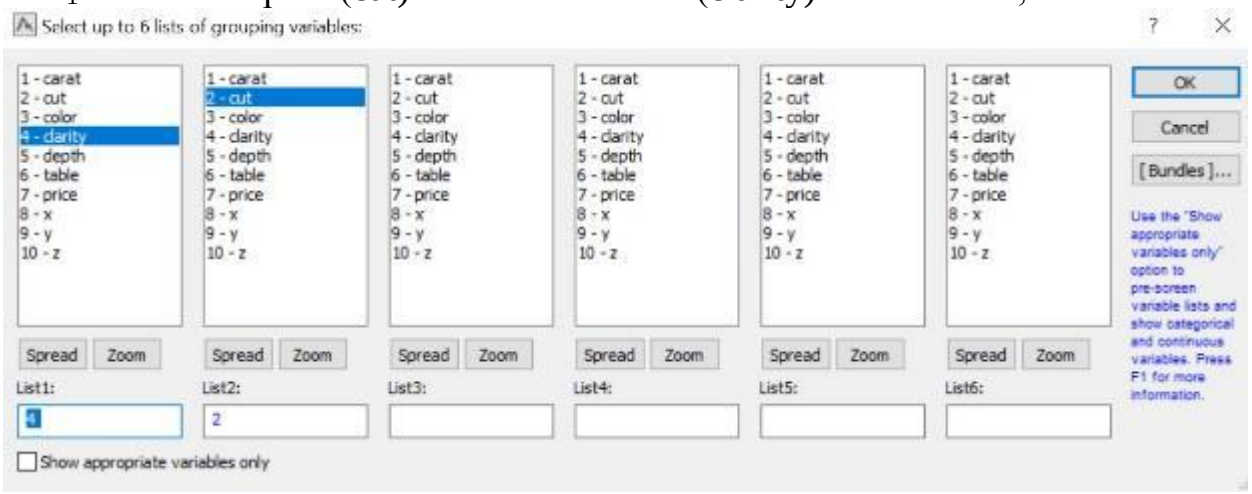

Рис. 24. Выбор переменных

Получаем следующие результаты (рис. 25-26).

Summary Frequency Table (Worksheet in diamonds2)
Marked cells have counts > 10
(Marginal summaries are not marked)

clarity	cut 1	cut 2	cut 3	cut 4	cut 5	Row Totals
SI2	466	1081	2100	2949	2598	9194
SI1	408	1560	3240	3575	4282	13065
VS1	170	648	1775	1989	3589	8171
VS2	261	978	2591	3357	5071	12258
VVS2	69	286	1235	870	2606	5066
VVS1	17	186	789	616	2047	3655
I1	210	96	84	205	146	741
IF	9	71	268	230	1212	1790
All Grps	1610	4906	12082	13791	21551	53940

Рис. 25. Сводная таблица частот

	Summary Frequency Table (Worksheet in diamonds2) Marked cells have counts > 10 (Marginal summaries are not marked)						
	clarity	cut 1	cut 2	cut 3	cut 4	cut 5	Row Totals
Count	SI2	466	1081	2100	2949	2598	9194
Column Percent		28.94%	22.03%	17.38%	21.38%	12.06%	
Row Percent		5.07%	11.76%	22.84%	32.08%	28.26%	
Total Percent		0.86%	2.00%	3.89%	5.47%	4.82%	17.04%
Count	SI1	408	1560	3240	3575	4282	13065
Column Percent		25.34%	31.80%	26.82%	25.92%	19.87%	
Row Percent		3.12%	11.94%	24.80%	27.36%	32.77%	
Total Percent		0.76%	2.89%	6.01%	6.63%	7.94%	24.22%
Count	VS1	170	648	1775	1989	3589	8171
Column Percent		10.56%	13.21%	14.69%	14.42%	16.65%	
Row Percent		2.08%	7.93%	21.72%	24.34%	43.92%	
Total Percent		0.32%	1.20%	3.29%	3.69%	6.65%	15.15%
Count	VS2	261	978	2591	3357	5071	12258
Column Percent		16.21%	19.93%	21.45%	24.34%	23.53%	
Row Percent		2.13%	7.98%	21.14%	27.39%	41.37%	
Total Percent		0.48%	1.81%	4.80%	6.22%	9.40%	22.73%
Count	VVS2	69	286	1235	870	2606	5066
Column Percent		4.29%	5.83%	10.22%	6.31%	12.09%	
Row Percent		1.36%	5.65%	24.38%	17.17%	51.44%	
Total Percent		0.13%	0.53%	2.29%	1.61%	4.83%	9.39%
Count	VVS1	17	186	789	616	2047	3655
Column Percent		1.06%	3.79%	6.53%	4.47%	9.50%	
Row Percent		0.47%	5.09%	21.59%	16.85%	56.01%	
Total Percent		0.03%	0.34%	1.46%	1.14%	3.79%	6.78%
Count	I1	210	96	84	205	146	741
Column Percent		13.04%	1.96%	0.70%	1.49%	0.68%	
Row Percent		28.34%	12.96%	11.34%	27.67%	19.70%	
Total Percent		0.39%	0.18%	0.16%	0.38%	0.27%	1.37%
Count	IF	9	71	268	230	1212	1790
Column Percent		0.56%	1.45%	2.22%	1.67%	5.62%	
Row Percent		0.50%	3.97%	14.97%	12.85%	67.71%	
Total Percent		0.02%	0.13%	0.50%	0.43%	2.25%	3.32%
Count	All Grps	1610	4906	12082	13791	21551	53940
Total Percent		2.98%	9.10%	22.40%	25.57%	39.95%	

Рис. 26. Результирующая таблица

В таблицах на рис. 25- рис. 26 мы видим значения ожидаемых частот, количества переменных по каждой группе и их процентные соотношения. Так, мы видим, что больше всего алмазом с ясностью SI1 (24,22%), и с качеством среза 5 (39,95%) – идеальное. Чтобы понять, зависимы ли они рассчитаем Х-квадрат Пирсона (рис. 27).

Statistic	Statistics: clarity(8) x cut(5) (Worksheet in diamonds2)		
	Chi-square	df	p
Pearson Chi-square	4391.398	df=28	p=0.0000
M-L Chi-square	3465.907	df=28	p=0.0000

Рис. 27. Расчёт Х-квадрата Пирсона

Согласно расчетам Х-квадрата Пирсона $p < 0,05$, следовательно, гипотеза H_0 принимается. Таким образом, качество среза (cut) и ясность алмаза (clarity) зависимы.

Заключение

В ходе проделанной работе мы на практике закрепили знания, полученные в ходе курса «Анализ данных». Считаю, что цель достигнута, анализ данных алмазов с помощью пакета Statistica проведен.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

Исходные данные были проанализированы с помощью факторного, дискриминантного, дисперсионного анализа, а также были применены описательная статистика и модель множественной регрессии, рассчитаны таблицы сопряженности. Методы многомерного статистического анализа хорошо описали реальное поведение совокупности алмазов. Выводы, сделанные на основании найденного статистического материала точны и надежны.

Для оценивания курсовой работы используется 4-балльная шала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».
Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания	Шкала оценок
Выбранные для курсовой работы данные в полной мере отвечают целям и задачам исследования. Корректно проведена предварительная обработка данных. Корректно применяются все выбранные методы исследования. По всем применяемым методам полученные хорошо с теоретическрой и прикладной постановки задачи обоснованные выводы. Сделано качественное заключение по всей работе в целом.	*Отлично*
Выбранные для курсовой работы данные в полной мере отвечают целям и задачам исследования. Корректно проведена предварительная обработка данных. Корректно применяются все выбранные методы исследования. По всем применяемым методам полученные выводы, но выводы недостаточно хорошо обоснованы. Заключение не в полной мере отражает цвыполнение целей и задач исследования.	*Хорошо*
Выбранные для курсовой работы данные в не полной мере отвечают целям и задачам исследования. Есть ошибки в одном или нескольких этапах обработки данных. Не получены обоснованные выводы по по всем результатам исследования.	*Удовлетвори-тельно*
Курсовая работа не выполнена, или выбранные для курсовой работы данные не отвечают целям и задачам исследования, или неправильно применяются методы обработки данных, или работа не содержит обоснованных выводов.	*Неудовлетвори-тельно*