

СОДЕРЖАНИЕ

## ВВЕДЕНИЕ ....................................................................................... 31 МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ............................ 4

1.1 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ............................................................... 4

1.2 РЕГРЕССИОННЫЙ АНАЛИЗ ...................................................................... 5 1.3 КАНОНИЧЕСКИЙ АНАЛИЗ .................................................................... 6

1.4 МЕТОДЫ СРАВНЕНИЯ СРЕДНИХ .......................................................... 7 1.5 КЛАСТЕРНЫЙ АНАЛИЗ ............................................................................ 8 1.6 ДИСКРИМИНАНТНЫЙ АНАЛИЗ ............................................................ 9

1.7 ДЕРЕВЬЯ КЛАССИФИКАЦИИ .............................................................. 10

1.8 АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ И КЛАССИФИКАЦИЯ ...................... 12

1.9 МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ ..................................................... 13

### 1.10 МОДЕЛИРОВАНИЕ СТРУКТУРНЫМИ УРАВНЕНИЯМИ, ПРИЧИННОЕ

МОДЕЛИРОВАНИЕ ................................................................................... 14

1.11 МЕТОДЫ АНАЛИЗА ВЫЖИВАЕМОСТИ.............................................. 15 1.12 ВРЕМЕННЫЕ РЯДЫ ........................................................................... 17

1.13 ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТОВ ................................................. 18 1.14 КАРТЫ КОНТРОЛЯ КАЧЕСТВА .......................................................... 19

[2 АНАЛИЗ ДАННЫХ .................................................................... 21ЗАКЛЮЧЕНИЕ 26](#_Toc48027)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27](#_Toc48028)

## ПРИЛОЖЕНИЕ А ..........................................................................28

## ВВЕДЕНИЕ

На сегодняшний день человечество шагнуло далеко в своём развитии. Благодаря статистике можно спрогнозировать почти любое событие. Начиная с бытового и заканчивая медицинским исследованием. Так, к примеру, можно провести анализ курса лечения тем или иным препаратом. Спрогнозировав приблизительные результаты курса лечения. Стоит ли его применять к пациенту или нет.

Целью курсовой работы является анализ рынка недвижимости на базе приложения Statistica 10.

Задачей является изучение методов статистического анализа данных, построение простейшей модели и проведение анализа.

Курсовая работа состоит из двух глав: методов статистического анализа и анализа данных.

В первой главе мною рассмотрены виды статистического анализа. В ней можно ознакомиться поближе с методами, которые применяются при обработке данных. Первая глава, в свою очередь разбита на пункты, где подробно рассказывается об анализах. Всего в ней четырнадцать пунктов.

Во второй главе, по таблице, составленной мной по квартирам в новостройках Краснодара, мы проведём анализ данных. В этой главе два пункта, регрессионный и корреляционный анализы данных.

В заключении подведены основные итоги курсовой работы.

## 1 Методы статистического анализа

Статистикой (от лат. Status-состояние дел) называется отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических знаний, изучение количественной стороны общественных явлений в числовой форме.

Впервые в науку это определение ввёл германский учёный Готфрид Ахенвалль в 1746 году. Он предложил заменить преподававшееся в то время в германских университетах «Государствоведение» на «Статистику», что и послужило началом развития статистики как науки и учебной дисциплины.

Но статистический учет был и до того, как его ввёл Ахенвалль, например, в древнем Китае велась перепись населения, а в Древнем Риме вёлся учет имущества граждан.

Ниже приведён список методов статистического анализа: корреляционный анализ, регрессионный анализ, канонический анализ, методы сравнения средних, кластерный анализ, дискриминантный анализ, деревья классификаций, анализ главных компонент классификация, многомерное шкалирование, моделирование структурными уравнениями или причинное моделирование, методы анализа выживаемости, временные ряды, планирование экспериментов, карты контроля качества.

### 1.1 Корреляционный анализ

Суть данного анализа заключается в том, что между случайными величинами может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. Также между переменными может возникать другая связь, которая заключается в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Данный вид связи называется стохастической. И проявляется она тогда, когда имеются общие случайные факторы, влияющие на обе переменные. Мерой зависимости здесь выступает коэффициент корреляции(r), который варьируется от -1 до +1. Если коэффициент отрицательный, то с увеличением одной переменной значения другой убывают.

Если переменные независимы, коэффициент равен 0. В противном случае, между переменными существует зависимость. Чем ближе коэффициент корреляции к 1, тем сильнее зависимость. R достигает своих предельных значений только тогда, когда зависимость между переменными линейная.

Корреляционный анализ позволяет установить силу и направление стохастической взаимосвязи между случайными величинами. Если переменные измерены хотя-бы в интервальной шкале и имеют нормальное распределение, то анализ осуществляется посредством вычисления коэффициента корреляции Пирсона, иначе используются корреляции

Спирмена, тау Кендала, или Гамма.

### 1.2 Регрессионный анализ

Метод в статистике, где исследуется влияние одной или нескольких независимых переменных (*X*1, …, *X*n) на зависимую переменную *Y*.

Данный анализ позволяет моделировать, проверять и исследовать пространственные отношения, помогает объяснить факторы, которые стоят за наблюдаемыми закономерностями.

Сферы жизни, где может быть использован регрессионный анализ многочисленны. Это прогнозирование выпадения осадков, степень влияния тех или иных факторов на то или иное заболевание и прочее.

Наиболее известным методом регрессионного анализа является МНК (метод наименьших квадратов). Он применяется для решения различных задач, которые основаны на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.

Метод позволяет построить глобальную модель процесса, подлежащие анализу или изучению. Также он создает уравнение регрессии в котором отражается процесс, происходящий при анализе.

Также в регрессионном анализе присутствует еще один метод. Это ГВР (географически взвешенная регрессия). Данный метод все чаще используется в географии и других дисциплинах.

Метод создаёт локальную модель прогнозируемой или изучаемой переменной, применяя уравнение регрессии к каждому объекту в наборе данных.

### 1.3 Канонический анализ

Данный вид анализа предназначен для анализа зависимостей между двумя списками признаков, которые характеризуют объекты. Так, к примеру, можно изучить зависимость между различными неблагоприятными факторами и появлением определённой группы симптомов, связь между показателями клинико-лабораторных синдромов больного. Канонический анализ является обобщением множественной корреляции как меры связи между одной переменной и множеством других переменных.

Как мы знаем, множественная корреляция есть максимальная корреляция между одной переменной и линейной функцией других переменных.

Данная концепция была обобщена на случай связи между множествами переменных, признаков, которые в той или иной мере характеризуют объекты. Для анализа нам достаточно рассмотреть наиболее коррелированные комбинации из каждого множества.

Так, к примеру, первое множество переменных состоит из признаков (y1, …, yn), во втором множестве (х1, …, хn), тогда их взаимосвязь можно оценить, как корреляцию между линейными комбинациями *а1y1 + …+ anyn, b1x1 + …+ bnxn*, которая называется канонической корреляцией.

Задача канонического анализа состоит в том, чтобы каноническая корреляция была максимальной.

### 1.4 Методы сравнения средних

Довольно часто в прикладных задачах попадаются случаи, когда один средний результат одной серии экспериментов отличается от среднего результата другой серии.

Поскольку средние-это результат измерений, то, в большинстве своём, они различаются. Только остаётся вопрос, можно ли объяснить расхождение средних неизбежными случайными ошибками эксперимента или вызвано конкретными причинами.

Поскольку речь идёт о сравнении двух средних, можно использовать критерий Стьюдента. Он является параметрическим, поскольку предполагается, что признак имеет нормальное распределении в каждой серии экспериментов. В последнее время, широко стал применяться непараметрический критерий сравнения средних.

Сравнение средних результатов является одним из способов выявления зависимостей между переменными признаками, которые, в свою очередь, характеризуют исследуемую совокупность объектов.

Если при разбиении объектов на подгруппы при помощи категориальной независимой переменной верна гипотеза о неравенстве средних некоторой зависимой переменной в подгруппах, а это значит, что существует стохастическая взаимосвязь между зависимой переменной и категориальным предиктором.

Например, благодаря этому, установлено, что верна гипотеза о равенстве средних показателей физического и умственного развития детей в группах тех матерей, которые на протяжении беременности употребляли или не употребляли табачные изделия, это означает, что существует зависимость между курением матери в период беременности и умственным и интеллектуальным развитием ребёнка.

Наиболее общий метод сравнения средних-дисперсионный анализ. В терминологии дисперсионного анализа фактором называется категориальный предиктор.

Дисперсионный анализ можно определить, как параметрический, статистический метод, который предназначен для оценки влияния различных факторов на результат эксперимента и для планирования будущих экспериментов. Поэтому в дисперсионном анализе можно исследовать зависимость количественного признака от одного или нескольких количественных факторов. Для рассмотрения одного фактора используют однофакторный анализ, в противном случае многофакторный.

### 1.5 Кластерный анализ

Данный анализ является методом классификационного анализа. Его основное значение – разбиение множества исследуемых объектов и признаков на однородные, в том или ином смысле, группы или кластеры. Предполагается, что исходные данные для анализа будут довольно объёмными. Другими словами, может быть большим как количество объектов исследования, так и признаков, характеризующих эти объекты.

Немаловажным достоинством кластерного анализа является то, что он даёт возможность производить разбиение объектов не по одному, а по ряду признаков. Помимо всего, в отличие от большинства математикостатистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество данных почти произвольной природы.

Поскольку кластеры являются своего рода группами однородности, то задачей кластерного анализа является то, что на основании признаков объектов разбивает их множество на n кластеров так, чтобы каждый объект принадлежал одной группе разбиения. При этом объекты, принадлежащие одному кластеру, должны быть сходными, а объекты, которые принадлежат разным кластерам, разнородными.

Если объекты кластеризации представить, как точки в многомерном пространстве признаков, то сходство между ними можно представить через понятие расстояния между точками, так как интуитивно понятно, что чем меньше расстояние между объектами, тем они более однородны.

### 1.6 Дискриминантный анализ

Данный анализ включает в себя статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает обучающими выборками. Дискриминантный анализ является многомерным, поскольку использует несколько признаков объекта, количество которых может быть довольно большим. Цель метода состоит в том, чтобы на основе изменения измерения различных признаков объекта классифицировать его, другими словами, отнести к одному из заданных классов некоторым оптимальным способом.

Предполагается, что исходные данные с признаками объектов содержат группирующую переменную, которая в свою очередь определяет принадлежность объекта к той или иной группе. Поэтому в дискриминантном анализе предусмотрена проверка непротиворечивости классификации, проведенной методом, с исходной эмпирической классификацией. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. В общем случае задачу дискриминации можно сформулировать следующим образом.

Пусть результатом наблюдения над объектом является построение nмерного случайного вектора X=(X1, X2, …, Xn), где X1, X2, …, Xn - признаки объекта. Необходимо установить правило, по которому значениям координат вектора X объект относят к одной из возможных совокупностей i, где i= 1, …, n. Методы дискриминации можно условно разделить на параметрические и непараметрические. В параметрических известно, что распределение векторов признаков в каждой совокупности нормально, не нет сведений о параметрах этих распределений.

Непараметрические методы не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений.

Если выполняются условия применимости дискриминантного анализа – независимые переменные – предикторы должны быть измерены как минимум в интервальной шкале, их распределение должно соответствовать нормальному закону, необходимо воспользоваться классическим дискриминантным анализом, в противном случае – методом общие модели дискриминантного анализа.

### 1.7 Деревья классификации

Данный анализ является методом классификационного анализа, который позволяет предсказывать принадлежность объектов к тому или иному классу в зависимости от соответствующих значений признаков, которые характеризуют объекты. Принято называть признаки независимыми переменными, а саму переменную, указывающую на принадлежность объектов к классам, называют зависимой.

Деревья классификации способны выполнять одномерное ветвление по переменным различных типов: категориальным, порядковым, интервальным. Чего нет в классическом дискриминантном анализе. Также не накладываются какие-либо ограничения на закон распределения количественных переменных.

Как и в дискриминантном анализе, данный метод даёт возможность анализировать вклады отдельных переменных в процедуру классификации.

Также, деревья классификации могут быть очень сложными, что довольно часто случается. Но с другой стороны, благодаря использованию специальных графических процедур, упрощается интерпретация результатов даже для очень сложных деревьев.

Большая популярность деревьев объясняется тем, что результаты можно представить графически и простотой интерпретации. Но самым важным отличительным свойством деревьев классификации является их иерархичность и широкая применимость.

Структура данного метода такова, что пользователь, по управляемым параметрам, может строить деревья произвольной сложности, при малых ошибках классификации. Но по сложному дереву, из-за большой совокупности решающих правил, затруднительно классифицировать новый объект. Поэтому при построении дерева классификации пользователь должен найти разумный компромисс между сложностью дерева и трудоёмкостью процедуры классификации.

Широкая сфера применимости деревьев делает их весьма привлекательным инструментом анализа данных, но не следует полагать, что его рекомендуется использовать вместо традиционных методов классификационного анализа. Как-раз наоборот, если выполнены более строгие теоретические предположения, налагаемые традиционными методами, и выборочное распределение обладает некоторыми специальными свойствами, к примеру соответствие распределения переменных по нормальному закону. То более результативным будет использование именно традиционных методов.

Но, как метод разведочного анализа или как последнее средство, когда все традиционные методы отказывают, по мнению многих исследователей, деревья классификации не имеют себе равных.

### 1.8 Анализ главных компонент и классификация

Довольно часто на практике возникает необходимость обработки большой размерности данных. Метод, который мы рассматриваем, позволяет решить эту задачу и служит для достижения двух целей:

* уменьшение общего числа переменных, другими словами редукция данных. Это необходимо для того, чтобы получить «главные» и «некоррелирующие» переменные.
* классификация наблюдений и переменных, при помощи строящегося факторного пространства;

Анализ имеет некоторое сходство с факторным анализом в постановочной части решаемых задач, но с другой стороны, имеется также и ряд значительных отличий:

* не используются итеративные методы для извлечения факторов при анализе главных компонент.
* вместе с активными переменными и наблюдениями, используемыми для извлечения главных компонент, можно задать вспомогательные переменные и наблюдения, после, вспомогательные переменные и наблюдения проектируются на факторное пространство, вычисленное на основе активных переменных и наблюдений.
* все вышеперечисленные возможности позволяют использовать данный метод как хорошее средство для классификации переменных и наблюдений одновременно;

Решение основной задачи метода достигается созданием векторного пространства латентных (скрытых) переменных (факторов) с размерностью меньше исходной. Исходная размерность определяется числом переменных для анализа в исходных данных.

### 1.9 Многомерное шкалирование

Данный метод подходит как альтернатива факторному анализу, где достигается сокращение числа переменных, путём выделения скрытых факторов, которые объясняют связи между наблюдаемыми переменными.

Целью данного метода является поиск и интерпретация скрытых переменных, дающих возможность пользователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков. Показателями сходства объектов на практике могут служить расстояния или степени связи между ними.

В многомерном шкалировании сходства между переменными выражаются с помощью матрицы коэффициентов корреляции. В этом анализе в качестве исходных данных можно использовать абсолютно произвольный тип матрицы сходства, такие как расстояния, корреляции и прочее.

Но несмотря на своё сходство, шкалирование и факторный анализ имеют ряд существенных различий. К примеру факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. Шкалирование таких ограничений не накладывает, оно может быть применимо, если задана матрица попарных сходств объектов.

В терминах различий получаемых результатов факторный анализ стремится извлечь больше факторов – скрытых переменных по сравнению с многомерным шкалированием. Поэтому шкалирование чаще приводит к проще интерпретируемым решениям. Но немало важным является то, что этот метод можно применять к любым типам расстояний или сходств, а факторный анализ, в свою очередь, требует, чтобы в качестве исходных данных была использована корреляционная матрица переменных или по файлу исходных данных сначала была вычислена матрица корреляций.

Основное предположение шкалирования состоит в том, что имеется некоторое метрическое пространство существенных базовых характеристик, которые неявно и послужили основой для полученных эмпирических данных о близости между парами объектов. А это значит, что объекты можно представить в виде точек в этом пространстве. Также предполагают, что более близким объектам соответствуют меньшие расстояния в пространстве базовых характеристик.

Исходя из этого, многомерное шкалирование является совокупностью методов анализа эмпирических данных о близости объектов, с помощью которых определяется размерность пространства существенных для данной содержательной задачи характеристик измеряемых объектов и конструируется конфигурация точек в этом пространстве.

### 1.10 Моделирование структурными уравнениями, причинное моделирование

Данный вид нового статистического анализа, уже получившего признание, является совокупностью многомерного статистического анализа, анализа корреляционных структур и новейших вычислительных алгоритмов.

Эта необычайно мощная техника многомерного анализа включает методы из различных областей статистики, множественная регрессия и факторный анализ получили здесь естественное развитие и объединение.

Объектом моделирования структурными уравнениями являются сложные системы, внутренняя структура которых не известна, также называется «черный ящик». Наблюдая параметры системы при помощи метода, можно исследовать структуру, установить причинно-следственные взаимосвязи между элементами системы.

Постановку задачи можно сформулировать следующим образом. Пусть имеются переменные, для которых известны статистические моменты, такие как матрица выборочных коэффициентов корреляции или ковариации. Эти переменные называются явными. Они могут являться характеристиками сложной системы.

Реальные связи между наблюдаемыми явными переменными могут быть достаточно сложными, но мы предположим, что имеется некоторое число скрытых переменных, которые с известной степенью точности объясняют структуру этих связей. Поэтому с помощью латентных переменных строится модель связей между явными и неявными переменными.

В некоторых задачах скрытые переменные можно рассматривать как причины, а явные как следствия, поэтому, эти модели называются причинными. Также допускается, что латентные переменные могут быть связаны между собой. Допускается достаточно сложная структура связей, однако её тип постулируется – это связи, которые описываются линейными уравнениями. Одни параметры линейных переменных известны, другие нет, и являются свободными параметрами.

Главная идея состоит в том, что является возможной проверка связи переменных A и B линейной зависимостью B=xA, анализируя их дисперсии и ковариации. Идея основана на простом свойстве среднего и дисперсии: если умножить каждое число на некоторую константу С, среднее значение также умножится на С, при этом стандартное отклонение также умножится на модуль С.

### 1.11 Методы анализа выживаемости

Изначально данный анализ был развит в медицинских, биологических исследованиях и страховании, но позже его стали широко применять в социальных и экономических науках, а также в промышленности в инженерных задачах.

Допустим изучается эффективность лечения или какого-либо препарата. Наиболее важной и объективной информацией будет являться средняя продолжительность жизни пациента с момента начала лечения или начала приёма препарата или же средняя продолжительность ремиссии. Для описания средних значений можно было бы использовать стандартные параметрические и непараметрические методы. Но в анализируемых данных есть значительная особенность – найдутся пациенты, которые в течение всего времени наблюдения выжили, а у некоторых заболевание находится в стадии ремиссии. Также может появиться группа больных, с которыми оборвались контакты до завершения эксперимента. Используя стандартные методы анализа пришлось бы исключать эту группу, тем самым потерять собранную информацию. Но большинство пациентов являются выжившими, другими словами они вылечились, то есть данный метод лечения оказался действенным.

Информация, когда нет данных о наступлении интересующего нас события называется неполной, в противном случае полной.

Наблюдения, которые содержат неполную информацию, называются цензурированными наблюдениями. Эти наблюдения типичны, когда наблюдаемая величина представляет время до наступления некоторого критического события, а продолжительность наблюдения ограничена по времени. Использование цензурированных наблюдений составляет специфику рассматриваемого метода, то есть анализа выживаемости.

В данном методе исследуются вероятностные характеристики интервалов времени между последовательным возникновением критических событий. Такие исследования называются анализом длительностей до момента прекращения, которые можно определить, как интервалы времени между началом и прекращением наблюдения за объектом, при котором объект перестаёт отвечать заданным для наблюдения свойствам.

Целью исследования является определение условных вероятностей, которые связаны с длительностями до момента прекращения. Построение таблиц времён жизни, подгонка распределения выживаемости, оценивание функции выживания с помощью процедуры Каплана – Мейера относятся к описательным методам исследования цензурированных данных.

Некоторые из приведенных выше методов позволяют сравнивать выживаемость в двух и более группах. Также анализ выживаемости содержит регрессионные модели для оценивания зависимостей между многомерными непрерывными переменными со значениями, аналогичными временам жизни.

Общие модели дискриминантного анализа. Если не выполняются условия применимости дискриминантного анализа, предикторы должны быть измерены как минимум в интервальной шкале, их распределение должно соответствовать нормальному закону, необходимо воспользоваться методом общие модели дискриминантного анализа.

Данный анализ имеет такое имя потому, что в нём для анализа дискриминантных функций используется общая линейная модель. В данном модуле анализ дискриминантных функций рассматривается как общая многомерная линейная модель, в которой категориальная зависимая переменная представляется векторами с кодами, обозначающими различные группы для каждого наблюдения.

Метод общего дискриминантного анализа имеет ряд значительных преимуществ перед классическим дискриминантным анализом. К примеру, не устанавливается никаких ограничений на тип используемого предиктора или на тип определяемой модели, возможен пошаговый выбор предикторов и выбор наилучшего подмножества предикторов, в случае наличия в файле данных кросс–проверочной выборки выбор наилучшего подмножества предикторов можно провести на основе долей ошибочной классификации для кросс-проверочной выборки и так далее.

### 1.12 Временные ряды

Данный тип анализа является наиболее интенсивно развивающимся и перспективным направлением статистики. Под временным или динамическим рядом понимается последовательность наблюдений некоторого признака X в последовательные равностоящие моменты t. Отдельные наблюдения называются уровнями ряда и обозначаются как xt, t=1, …, n.

При исследовании временного ряда обычно выделяют несколько составляющих:

Xt=ut + yt + ct + et, t=1, …, n,

где переменные обозначают следующее: ut – тренд, плавно меняющаяся компонента, описывающая чистое

влияние долговременных факторов; yt - сезонная компонента, отражающая повторяемость процессов в

течение не очень длительного периода; ct - циклическая компонента, отражающая повторяемость процессов в

течение длительных периодов времени свыше одного года; t – случайная компонента, которая отражает влияние не поддающихся

учёту и регистрации случайных факторов.

Первые три компоненты представляют собой детерминированные составляющие. Случайная составляющая, в свою очередь, образована в результате суперпозиции большого числа внешних факторов, оказывающих каждый в отдельности незначительное влияние на изменение значений признака Х. Анализ и исследование временного ряда, позволяют строить модели для прогнозирования значений признака Х на будущее время, если известна последовательность наблюдений в прошлом.

### 1.13 Планирование экспериментов

Планирование экспериментов являет собой искусство располагать наблюдения в определённом порядке или проводить специально спланированные проверки с целью полного использования возможностей этих методов.

Метод широко применяется в настоящее время как в науке, так и в различных областях практической деятельности. В основном, практическая цель состоит в том, чтобы показать статистическую значимость эффекта воздействия определённого фактора на изучаемую зависимую переменную.

Обычно основная цель планирования экспериментов заключается в извлечении максимального количества объективной информации о влиянии изучаемых факторов на интересующий исследователя показатель с помощью наименьшего числа дорогостоящих наблюдений.

Но, к несчастью, на практике в большинстве случаев, недостаточное внимание уделяется планированию исследований. Собирают данные, после уже проводят статистическую обработку и анализ. Но сам по себе правильно проведенный статистический анализ недостаточен для достижения научной достоверности, так как качество любой информации, которую мы получаем в результате анализа данных, зависит от качества самих данных. Поэтому планирование экспериментов находит все большее применение в прикладных исследованиях. Целью методов планирования экспериментов является изучение влияния определенных факторов на исследуемый процесс и поиск оптимальных уровней факторов, определяющих требуемый уровень течения данного процесса.

### 1.14 Карты контроля качества

В современном мире довольно актуальным стал вопрос качества как товаров, так и услуг, предоставляемых населению. От успешного решения этого вопроса в значительной степени зависит благополучие фирмы, организации или учреждения.

Качество продукции или услуг сформировывается в процессе научных исследований, конструкторских и технологических разработок, обеспечивается хорошей организацией производства и услуг. Однако изготовление продукции и оказание услуг, независимо от их вида, всегда связано с определённым непостоянством условий производства и предоставления. Что приводит к некоторой вариабельности признаков их качества.

Отчего актуальными являются вопросы разработки методов контроля качества, позволяющие вовремя обнаружить признаки нарушения технологического процесса или оказания услуг. Вдобавок, для достижения и поддержания высокого уровня качества, который удовлетворит потребителя нужны методы, направленные не на устранение дефектов готовой продукции и несоответствий услуг, а на предупреждение и прогнозирование причин их появления.

Карта контроля является инструментом, который позволяет отслеживать ход протекания процесса и воздействовать на него, предупреждая его отклонения от предъявленных к процессу требований.

Инструментарий данного метода широко использует статистические методы, основанные на теории вероятностей и математической статистики.

Применение статистических методов позволяет при ограниченных объёмах анализируемых изделий с заданной степенью точности и достоверности судить о состоянии качества выпускаемой продукции. Также обеспечивает прогнозирование, оптимальное регулирование проблем в области качества, принятие правильных управленческих решений не на основе «шестого чувства», а при помощи научного изучения и выявления закономерностей в накапливаемых массивах числовой информации.

## 2 Анализ данных

Во второй главе я собрал данные с новостроек двух застройщиков Краснодара: Юг Строй Империал и ССК.

Для проведения анализа воспользуемся следующей таблицей данных:

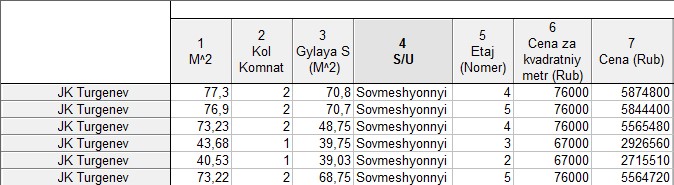


Рисунок 1 – Фрагмент таблицы дынных В исходной таблице указаны параметры квартир застройщиков: квадратные метры или площадь квартиры (M^2); количество комнат в квартире (Kol Komnat); жилая площадь (Gylaya S); тип санузла (S/U); этаж (Etaj); стоимость одного квадратного метра (Cena za kadratniy metr); общая стоимость квартиры (Cena);

Переменные выбраны не случайно, они нужны для проведения анализов. Также они являются основными критериями при формировании цены на квартиру.

Проведем несколько анализов с этими данными.

Первым проведём корреляционный анализ.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | | Correlations (Kurs)  Marked correlations are significant at p < ,05000 N=50 (Casewise deletion of missing data) | | | | | |
| Means | Std.Dev. | M^2 | Kol Komnat | Etaj | Cena |
|  | M^2 | 86 | 44 | 1,000000 | 0,859492 | 0,661253 | 0,955330 |
| Kol Komnat | 2 | 1 | 0,859492 | 1,000000 | 0,597360 | 0,723432 |
| Etaj | 11 | 6 | 0,661253 | 0,597360 | 1,000000 | 0,558958 |
| Cena | 5856099 | 4094702 | 0,955330 | 0,723432 | 0,558958 | 1,000000 |

Рисунок 2 – Таблица корреляции

Как мы можем видеть по данным полученной таблицы, каждая переменная статистически значима от другой. При этом, значимость переменной от самой себя не наблюдается. Чем ближе коэффициент к 1, тем лучше взаимосвязь.

Самой слабой является взаимосвязь между ценой и этажом. А самой сильной, взаимосвязь цены квадратных метров.

Построим двухмерную модель по переменным Цена и Квадратные метры.

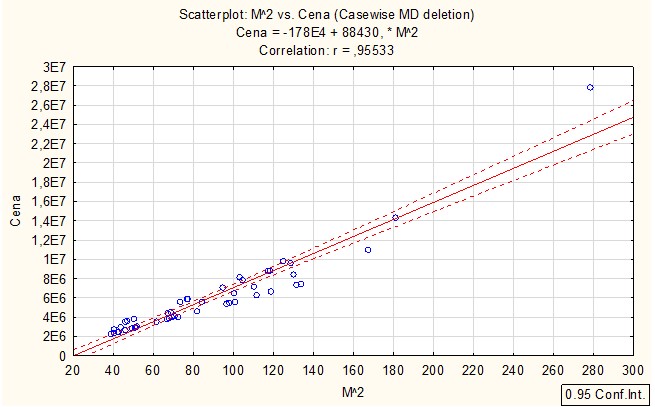


Рисунок 3 – Диаграмма рассеяния от Цены и Площади

Как мы можем видеть, значения находятся довольно близко к корреляционной прямой, что говорит о хорошем подборе значений.

Теперь составим по переменным цена и этаж.

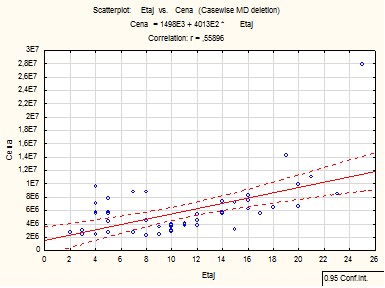


Рисунок 4 – Диаграмма рассеяния от Цены и Этажа

В данном случае, значения имеют довольно сильный разлёт. Значения, в большинстве своём находятся далеко от прямой.

Исходя из полученных диаграмм, можно сделать вывод, что лучшая корреляция происходит между переменными цена и квадратные метры.

Следующим анализом, который мы проведём будет регрессионный анализ.

Построим модель парной регрессии.

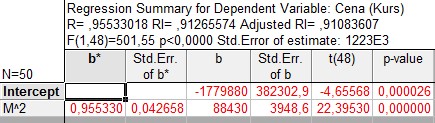


Рисунок 5 – Парная регрессия

В информационной части содержатся сведения о результатах анализа, а именно:

Dependent (имя зависимой переменной), в данном случае это Cena;

№ of cases = 50 (Число наблюдений, по которым построена

регрессионная модель);

Multiple Regression = 0.9553(Коэффициент множественной корреляции);

R-square = R2 = 0.9127(коэффициент детерминации);

Adjusted R2 = 0.9108(скорректированный коэффициент детерминации);

Std.Error of estimate = 1223E3(стандартная ошибка оценки). Эта статистика – мера рассеяния наблюдаемых значений относительно регрессионной прямой;

Intercept = -1779880(оценка свободного члена b0 регрессии), если выбрана регрессия, включающая свободный член;

Std.Error = 382302.9(стандартная ошибка оценки свободного члена b0); t = -4.65568(значение t-критерия и уровень значимости p) для проверки

гипотезы о равенстве нулю свободного члена b0;

F = 501.55, df = 1, 48, p<0.0000 (Значение F-критерия, число степеней свободы и уровень значимости p) используются в качестве общего F-критерия для проверки гипотезы о зависимости предикторов и отклика.

Уравнение регрессии:

Цена=-1779880+Квадратные\_метры\*88430

Исходя из таблицы регрессии можно сделать вывод, что при изменении площади квартиры в среднем на 1 единицу, цена на квартиру увеличивается на 88430 руб.

Как мы можем наблюдать, коэффициент R2>0.75. Получается, что зависимость сильная.

Далее возьмём Общую стоимость как зависимую переменную, а в качестве независимых будут квадратные метры, этаж и количество комнат.

Используя данный метод, можно выявить функциональную зависимость стоимости квартиры от вышеупомянутых переменных.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| N=50 | Regression Summary for Dependent Variable: Cena (Kurs) R= ,97775992 RІ= ,95601446 Adjusted RІ= ,95314584  F(3,46)=333,27 p<0,0000 Std.Error of estimate: 8863E2 | | | | | |
| b\* | Std.Err. of b\* | b | Std.Err. of b | t(46) | p-value |
| Intercept |  |  | -980053 | 306391,8 | -3,19869 | 0,002500 |
| M^2 | 1,339219 | 0,064860 | 123964 | 6003,8 | 20,64773 | 0,000000 |
| Kol Komnat | -0,361521 | 0,060670 | -1353876 | 227206,7 | -5,95879 | 0,000000 |
| Etaj | -0,110645 | 0,041340 | -79434 | 29678,4 | -2,67649 | 0,010275 |

Рисунок 6 – Таблица регрессии

Для полученной таблицы также распишем значения.

Dependent, как и в предыдущем случае, это Cena;

№ of cases = 50;

Multiple Regression = 0.9777;

R-square = R2 = 0.9560;

Adjusted R2 = 0.9531;

Std.Error of estimate = 8863E2;

Intercept = -980053; Std.Error = 306391.8; t = -3.19869;

F = 333.27, df = 3, 46, p<0.0000.

Как мы можем наблюдать, коэффициент R2>0.75. Получается, что зависимость сильная.

Уравнение регрессии:

Цена=-980053+123964\*Квадратные\_метры-

1353876\*Количество\_комнат-79434\*Этаж

Исходя из таблицы регрессии можно сделать вывод, что при изменении площади квартиры в среднем на 1 единицу и других неизменных переменных, цена на квартиру увеличивается на 123964 руб.

# ЗАКЛЮЧЕНИЕ

В результате проделанной работы, была достигнута поставленная цель, а именно, проведён анализ данных, предоставленных в таблице, при помощи

Statistica 10.

Построив несколько моделей и проведя анализы, мы можем увидеть значимость переменных относительно друг друга и какие переменные наиболее значимые.

В процессе работы были построена таблица регрессионного анализа, таблица и графики корреляционного анализа.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Халафян А.А. Statistica 6, Статистический анализ данных, Москва,

изд. «Бином», 2007. – 508 с.

1. Доугерти К. Введение в эконометрику, Москва, изд. «М», 1999. – 416 с.
2. Kubsu, Виды статистического анализа,URL: <http://statlab.kubsu.ru/node/4>(дата обращения: 15.12.2017).
3. Wikipedia, URL:

[https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7)

[D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9\_%](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7)

[D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7)  (дата обращения:

10.12.2017)

1. URL:[http://desktop.arcgis.com/ru/arcmap/10.3/tools/spatial-statisticstoolbox/regression-analysis-basics.htm](http://desktop.arcgis.com/ru/arcmap/10.3/tools/spatial-statistics-toolbox/regression-analysis-basics.htm) (дата обращения: 13.12.2017)

## Приложение А

Данные, используемые при выполнении анализов, предоставленные в таблице Excel.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Площадь (М^2) | Количест во комнат | Жилая площадь  (М^2) | Санузел | Этаж (Номер) | Цена за квадратн ый метр  (Рублей) | Цена (Рублей) |
| ЖК  Тургенев | 77,3 | 2 | 70,8 | Совмеще нный | 4 | 76000 | 5874800 |
| ЖК  Тургенев | 76,9 | 2 | 70,7 | Совмеще нный | 5 | 76000 | 5844400 |
| ЖК  Тургенев | 73,23 | 2 | 48,75 | Совмеще нный | 4 | 76000 | 5565480 |
| ЖК  Тургенев | 43,68 | 1 | 39,75 | Совмеще нный | 3 | 67000 | 2926560 |
| JK Turgenev | 40,53 | 1 | 39,03 | Совмеще нный | 2 | 67000 | 2715510 |
| ЖК  Тургенев | 73,22 | 2 | 68,75 | Совмеще нный | 5 | 76000 | 5564720 |
| ЖК  Тургенев | 278,45 | 5 | 215,1 | Совмеще нный | 25 | 99990 | 27842215,  5 |
| ЖК  Тургенев | 128,77 | 3 | 122,9 | Совмеще нный | 4 | 75000 | 9657750 |
| ЖК  Тургенев | 118,05 | 3 | 108,4 | Раздельн ый | 7 | 75000 | 8853750 |
| ЖК  Тургенев | 117,37 | 3 | 111,5 | Совмеще нный | 8 | 75000 | 8802750 |
| ЖК  Тургенев | 104,4 | 3 | 98,8 | Раздельн ый | 5 | 75000 | 7830000 |
| ЖК  Тургенев | 94,56 | 3 | 84,16 | Раздельн ый | 4 | 75000 | 7092000 |
| ЖК  Тургенев | 46 | 1 | 44,4 | Совмеще нный | 9 | 77000 | 3542000 |
| ЖК  Тургенев | 47,1 | 1 | 44,4 | Совмеще нный | 11 | 77000 | 3626700 |
| ЖК  Тургенев | 50 | 1 | 46,6 | Совмеще нный | 12 | 77000 | 3850000 |
| ЖК  Тургенев | 180,85 | 3 | 165,25 | Совмеще нный | 19 | 79000 | 14287150 |
| ЖК  Тургенев | 130,03 | 3 | 121,09 | Совмеще нный | 23 | 65000 | 8451950 |
| ЖК  Тургенев | 167,18 | 4 | 155,85 | Раздельн ый | 21 | 66000 | 11033880 |
| ЖК  Тургенев | 124,71 | 3 | 120,47 | Совмеще нный | 20 | 79000 | 9852090 |
| ЖК  Тургенев | 110,4 | 3 | 99,95 | Совмеще нный | 15 | 65000 | 7176000 |
| ЖК  Тургенев | 100,18 | 3 | 94,19 | Совмеще нный | 18 | 65000 | 6511700 |
| ЖК  Тургенев | 103,29 | 2 | 97,72 | Совмеще нный | 16 | 79000 | 8159910 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ЖК  Тургенев | 84,24 | 2 | 80,59 | Совмеще нный | 17 | 66000 | 5559840 |
| ЖК  Тургенев | 68,86 | 2 | 64,68 | Раздельн ый | 8 | 66000 | 4544760 |
| ЖК  Тургенев | 67,43 | 2 | 63,4 | Раздельн ый | 5 | 66000 | 4450380 |
| ЖК  Сердце  Столицы | 118,91 | 5 | 114,31 | Раздельн ый | 20 | 56000 | 6658960 |
| ЖК  Сердце  Столицы | 133,91 | 4 | 129,31 | Раздельн ый | 16 | 56000 | 7498960 |
| ЖК  Сердце  Столицы | 111,8 | 4 | 107,3 | Раздельн ый | 16 | 56000 | 6260800 |
| ЖК  Сердце  Столицы | 131,59 | 3 | 127,2 | Совмеще нный | 14 | 56000 | 7369040 |
| ЖК  Сердце  Столицы | 100,63 | 3 | 96,15 | Совмеще нный | 14 | 56000 | 5635280 |
| ЖК  Сердце  Столицы | 97,76 | 3 | 93,16 | Совмеще нный | 14 | 56000 | 5474560 |
| ЖК  Сердце  Столицы | 96,43 | 3 | 92,03 | Раздельн ый | 12 | 56000 | 5400080 |
| ЖК  Сердце  Столицы | 81,85 | 3 | 77,45 | Раздельн ый | 12 | 56000 | 4583600 |
| ЖК  Сердце  Столицы | 72,5 | 2 | 68,1 | Раздельн ый | 11 | 56000 | 4060000 |
| ЖК  Сердце  Столицы | 70,52 | 2 | 66,12 | Раздельн ый | 11 | 58000 | 4090160 |
| ЖК  Сердце  Столицы | 70,33 | 2 | 65,93 | Раздельн ый | 11 | 58000 | 4079140 |
| ЖК  Сердце  Столицы | 68,89 | 2 | 64,49 | Совмеще нный | 11 | 58000 | 3995620 |
| ЖК  Сердце  Столицы | 67,73 | 2 | 63,33 | Раздельн ый | 10 | 58000 | 3928340 |
| ЖК  Сердце  Столицы | 66,71 | 2 | 62,31 | Раздельн ый | 10 | 58000 | 3869180 |
| ЖК  Сердце  Столицы | 61,69 | 1 | 57,29 | Совмеще нный | 10 | 58000 | 3578020 |
| ЖК  Сердце  Столицы | 51,25 | 1 | 46,85 | Совмеще нный | 10 | 58000 | 2972500 |
| ЖК  Сердце  Столицы | 50,67 | 1 | 46,27 | Совмеще нный | 10 | 58000 | 2938860 |
| ЖК  Сердце  Столицы | 49,19 | 1 | 44,79 | Совмеще нный | 10 | 58000 | 2853020 |
| ЖК  Сердце  Столицы | 46,13 | 1 | 41,73 | Совмеще нный | 5 | 58000 | 2675540 |
| ЖК  Сердце  Столицы | 45,82 | 1 | 41,42 | Совмеще нный | 7 | 58000 | 2657560 |
| ЖК  Сердце  Столицы | 42,79 | 1 | 40,54 | Совмеще нный | 9 | 58000 | 2481820 |
| ЖК  Сердце  Столицы | 42,33 | 1 | 40,13 | Совмеще нный | 3 | 58000 | 2455140 |
| ЖК  Сердце  Столицы | 40,55 | 1 | 38,95 | Совмеще нный | 4 | 58000 | 2351900 |
| ЖК  Сердце  Столицы | 38,99 | 1 | 36,9 | Совмеще нный | 8 | 58000 | 2261420 |
| ЖК  Сердце  Столицы | 51,85 | 1 | 49,68 | Совмеще нный | 15 | 59000 | 3059150 |