СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В бесконечно развивающимся мире соответственно развитию происходит и рост и накопление данных и их источников. Повсеместная цифровизация и популярность электронных, в частности, мобильных устройств, подключённых к мировой интернет сети только добавляет масла в огонь неистовой экспансии информационных полей и данных. Каждый человек носит с собой переносной бесконечный источник данных. И кому-то эти данные всегда пригодятся для улучшения виртуальных условий и пущего их расширения. Первым шагом к улучшению данных условий всегда является тот или иной статистический анализ получаемых данных.

Одним из популярнейших видов активности, будь то в бездельном метро, в перерыве между работой или просто в свободное от дел время являются мобильные игры. И одной из популярнейших из них является игра «Clash of Clans». Разумеется, подобный ресурс является и большим кладезем полезной для аналитиков информации и в этой курсовой работе мы попытаемся раскрыть его.

Цель курсовой работы — провести статистический анализ данных на примере данных игры «Clash of Clans».

Задачи, решаемые в данной курсовой работе:

- предобработка данных;
- анализ данных;
- использование знаний математической статистики с применением современных средств обработки данных: аналитической платформы Loginom;
- закрепление знаний об оформлении документации.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Статистический анализ данных — это процесс преобразования данных в информацию с помощью различных методов и техник. Он включает в себя сбор, организацию, анализ, интерпретацию и представление данных с целью установления связей между переменными, нахождения закономерностей, обоснования гипотез и иных методов выделения новой информации из исходной. Данная работа затронет снижение размерности данных и их агрегацию с целью формирования выводов и статистик.

1.1 Факторный анализ

Рассмотрим один из наиболее сложных инструментов, применяемых в работе, — факторный анализ. Факторный анализ — многомерный метод, применяемый для изучения взаимосвязей между значениями переменных. Изначально он получил распространение в психологии. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки.

Однако широкие возможности, которые этот метод предоставляет в распоряжение исследователей для описания отношений в ситуациях с большим количеством признаков, привели к тому, что, помимо широкого традиционного использования в психологии, он эффективно используется в различных областях науки и техники, от метеорологии и коммуникационных технологий до социологии и международных отношений. [**Error! Reference source not found.**]

Рассматривая данный метод с точки зрения уменьшения размерности измерений, факторный анализ — это область математической статистики (один из разделов многомерного статистического анализа), объединяющая вычислительные методы, которые в ряде случаев позволяют получить

компактное описание исследуемых явлений на основе обработки больших массивов информации. [1.6]

Существуют следующие типы факторного анализа:

1. Детерминированный (функциональный) — результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов.
2. Стохастический (корреляционный) — связь между результативным и факторными показателями является неполной или вероятностной.
3. Прямой (дедуктивный) — от общего к частному.
4. Обратный (индуктивный) — от частного к общему.
5. Одноступенчатый и многоступенчатый.
6. Статический и динамический.
7. Ретроспективный и перспективный.

Обязательные условия факторного анализа:

1. Все признаки должны быть количественными.
2. Число признаков должно быть более чем в два раза больше числа переменных.
3. Выборка должна быть однородна.
4. Исходные переменные должны быть распределены симметрично.
5. Факторный анализ осуществляется по коррелирующим переменным.

Рассмотрим пример: экономист непосредственно наблюдает множество различных показателей статистического учета деятельности предприятий, чтобы выявить закономерности, влияющие на рост производительности труда (образовательный уровень рабочих, коэффициент сменности оборудования, электровооруженность труда, возраст оборудования, количество мест в столовых и т.п.). Так или иначе, все факторы, отражаемые этими показателями, воздействуют на изучаемый показатель — производительность труда. При этом многие из них связаны между собой, порой отражая с разных сторон те же, по существу, явления.

При помощи факторного анализа этих связей (корреляций) удается обнаружить, что на самом деле решающее влияние на рост производительности труда оказывает лишь несколько обобщенных факторов (например, размер предприятия, уровень организации труда, характер продукции), непосредственно не наблюдавшихся при исследовании. Собственно, это их действие и проявляется в учитываемых показателях.

Выявленные факторы позволяют строить аналитические модели с относительно небольшим числом независимых переменных, что упрощает их реализацию и интерпретацию пользователем, снижает вычислительные затраты и время, требуемое на получение решений, а следовательно, повышает оперативность принятия решений на основе результатов анализа. Знание этих факторов в дальнейшем также позволяет обоснованно включать их в качестве управляемых факторов (переменных) в модель экономического эксперимента, рассчитывать обобщенные индексы, характеризующие экономические явления и т.д.

В настоящее время факторный анализ широко используется в экономике, психологии, нейрофизиологии, социологии, политологии, и статистике. Также имеется ряд работ, в которых описано применение факторного анализа в геологии. Он служит для определения некоторых характеристик месторождений полезных ископаемых, форм и процессов формирования рудных тел.

Рассмотрим примеры задач из различных областей, в решении которых может помочь факторный анализ:

Расчет прибыли от продаж компании. Прибыль зависит от четырех основных факторов: объема продаж, ассортимента реализованной продукции, себестоимости продукции и цены реализации продукции. С помощью факторного анализа можно рассчитать, как каждый из перечисленных факторов влияет на величину прибыли компании, и, исходя из полученных результатов, выработать пути максимизации прибыли. Аналогичным образом можно рассчитать и проанализировать затраты на производство продукции.

Прогнозирование распространения инфекционного заболевания. В качестве исторических данных нередко используются исторические сведения о погоде (температура воздуха, влажность). Данная информация позволяет учесть факторы, благоприятно влияющие на размножение переносчиков заболевания. Поэтому каждый объект в выборке описывается десятками признаков.

Так как изменения погодных условий, как правило, происходит постепенно, это приводит к наличию сильной корреляции между факторами в указанных данных. Для решения указанной проблемы в данной задаче может быть использован факторный анализ.

При анализе сильно коррелированные друг с другом переменные объединяются в один фактор, в результате перераспределяется разброс между компонентами и получается наиболее простая и наглядная структура факторов. После объединения корреляция компонентов внутри каждого фактора друг с другом будет выше, чем их корреляция с компонентами других факторов. Эта процедура также позволяет выделить скрытые переменные, что особенно важно при анализе социальных представлений и ценностей.

Существует несколько типов факторного анализа, в нашем случае используется метод главных компонент, по причине того, что этот метод использует платформа Loginom. Цель метода заключается в снижении размерности. Задача снижения размерности набора данных состоит в описании точек данных с помощью величин количеством меньшим по сравнению с размерностью пространства. Данные величины должны быть функциями исходных координат, то есть:

$$\eta_k = F_k(\xi_1, \xi_2, \dots, \xi_m),$$

где η_k — координаты в новом пространстве;

F_k — функция от исходных переменных;

ξ_i — исходные переменные;

$k = 1..m', m' < m;$

m — исходная размерность пространства;

m' — размерность нового пространства.

Функции F_k задают отображение F из исходного пространства R^m в пространство $R^{m'}$. В методе главных компонент F — некоторое линейное ортогональное нормированное отображение, т. е.:

$$F_k(\xi_1, \xi_2, \dots, \xi_m) = c_{1k}(\xi_1 - \mu_1) + \cdots + c_{mk}(\xi_m - \mu_m),$$

где $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$;

x_{ij} — средние по набору данных значения признаков;

c_{ik} — значения матрицы перехода.

На коэффициенты c_{ij} накладываются следующие условия:

$$\sum_{k=1}^{m} c_{ik}^2 = 1, \sum_{k=1}^{m} c_{ik}\ c_{jk} = 0,$$

где $i, j = 1..m, i \neq j$.

Так же для анализа используется критерий оценки дисперсии J:

$$J = \frac{D\eta_1 + \cdots + D\eta_{m\prime}}{D\xi_1 + \cdots + D\xi_m},$$

где D — вычисление дисперсии случайной величины.

Первая главная компонента — это нормированно-центрированная линейная комбинация исходных признаков, которая среди всех прочих нормировано-центрированных линейных комбинаций обладает на данном наборе данных наибольшим значением критерия J.

k-ой главной компонентой (k = 2..m) называется такая нормировано-центрированная линейная комбинация исходных признаков, которая не коррелированна с (k–1) предыдущими главными компонентами и среди всех

прочих нормированно-центрированных линейных комбинаций, не коррелированных с предыдущими (k–1) главными компонентами, обладает на данном наборе данных наибольшим значением критерия J.

Процесс, так или иначе, сводится к выбору новой ортогональной системы координат в пространстве наблюдения, когда угол между факторами остается правильным при повороте осей координат. Направление, в котором массив данных имеет наибольший разброс, выбирается в качестве первого главного компонента, а каждый последующий выбирается так, чтобы разброс данных вдоль него являлся максимальным и чтобы он обладал свойством ортогональности другим основным компонентам, выбранным ранее.

Существует два метода вращения осей: варимакс и квартимакс.

Варимакс — наиболее распространенный на практике метод, целью которого является минимизировать количество переменных, имеющих высокие нагрузки на конкретный фактор, что в свою очередь способствует упрощению описания фактора за счет группировки вокруг него только тех переменных, которые с ним связаны в большей степени, чем с остальными.

Квартимакс действует обратным образом, минимизируя количество факторов, необходимых для объяснения конкретной переменной. Квартимакс-вращение приводит к выделению одного из общих факторов с достаточно высокими нагрузками на большинство переменных.

Поскольку факторный анализ позволяет уменьшить громоздкий набор переменных до сравнительно меньшего набора факторов, он подходит для упрощения сложных моделей.

Таким образом, за счёт уменьшения размерности вводимых данных, факторный анализ является сильным инструментов при работе с многомерными данными, таким как данные игры «Clash of Clans».

1.2 Агрегация

Для формирования статистических выводов в работе будет использоваться метод агрегирования данных. Агрегация — это процесс объединения элементов в одну систему. Она включает в себя суммирование, нахождение максимума, минимума, среднего, медианы и других метрик.

Агрегация данных применяется в различных областях, где необходимо обрабатывать большие объемы информации и получать обобщенные результаты для принятия обоснованных решений. Такими областями являются:

1. Бизнес и финансы: В сфере бизнеса агрегация применяется для анализа продаж, финансовых показателей и производственных данных. Например, хозяева компаний могут объединять ежедневные данные о продажах для извлечения еженедельных или ежемесячных отчетов о доходах. Финансовые аналитики используют агрегацию для выявления общей суммы инвестиций, объема торговли и других ключевых аспектов финансовых операций.
2. Маркетинг и аналитика: В данной области агрегация данных помогает оценить результаты маркетинговых кампаний, поведение пользователей на сайтах или в приложениях, и другие метрики. Например, путем агрегации посещений сайта веб-аналитики могут определить наиболее популярные страницы или узнать общее количество посещений за определенный период времени.
3. Медицина и наука: В этих областях агрегация данных используется для анализа результатов исследований, статистики заболеваемости, а также для мониторинга и прогнозирования эпидемий и пандемий. Например, агрегация данных о заболеваемости может помочь выявить тенденции и провести анализ эффективности медицинских программ.
4. Социальные науки и общественное мнение: В этой области агрегация используется для определения общих тенденций и

предпочтений в обществе. Например, агрегация опросов общественного мнения помогает политологам и социологам понять предпочтения избирателей, анализировать общественное мнение и тенденции социокультурного развития.

5. Телекоммуникации и сети: В данной области агрегация данных используется для анализа трафика, мониторинга производительности сети и предоставления обобщенной информации о передаче данных.

Это всего лишь несколько примеров областей, в которых агрегация данных является критически важной. В сущности, агрегация применяется везде, где необходимо обобщить большие объемы информации для выявления тенденций, принятия решений и понимания ключевых показателей.

На уровне железа, агрегация данных обычно осуществляется с использованием специализированных вычислительных устройств и процессоров. Например, в современных базах данных могут использоваться специализированные аппаратные ускорители для выполнения операций агрегации из больших объемов данных. Такие устройства способны обрабатывать множество операций параллельно, что увеличивает общую производительность в сравнении с обычными процессорами.

Также, в рамках распределенных систем, где данные хранятся на нескольких узлах, агрегация может осуществляться путем параллельной обработки данных на уровне каждого узла, а затем объединения результатов для получения общей сводной информации.

Таким образом, агрегирование данных является широкой и повсеместной задачей, с первого взгляда алгоритмически затратной, но оптимально и быстро реализованной в реальных системах. Агрегация данных поможет нам собрать статистику по исследуемым данным.

Следующий блок будет на практике сочетать описанные выше методы с целью проведения анализа.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

Выполнение анализа начинается с выбора данных. В качестве исследуемого набора на вход подаётся набор данных «Clash of Clans Dataset», представляющий из себя данные о случайной выборке кланов из игры «Clash of Clans». [1.2.1]

2.1 Предобработка данных

Данные представлены в виде таблицы со следующими столбцами данных:

- clan_name — название клана;
- clan_tag — внутри игровой идентификатор клана;
- clan_discriprion — описание клана;
- clan_type — тип клана по способу вступления. Может принимать значения open, inviteOnly и closed, обозначающие, что клан открытый, вступление осуществляется по приглашению или что клан является закрытым;
- clan_location — страна, указанная в профиле клана;
- isFamilyFriendly — дозволенность нецензурного лексикона в клане;
- другие (всего датасет насчитывает 27 столбцов).

Все операции будут выполнены на low-code платформе Loginom.

Отделим от общего числа факторов те, которые будут использоваться в анализе.

Факторы «clan_tag», «clan_name», «clan_description», «clan_badge_url», «war_frequency» не будут использоваться, так как они не являются показательными для анализа.

В данных присутствуют пустые значения и выбросы, мешающие проводить анализ. Для начала, избавимся от строк, в значениях которых

присутствуют пропуски. Для этого воспользуемся узлом «Заполнение пропусков», предназначенным для работы с пустыми ячейками. Для чистоты анализа методом обработки пропущенных значений будет их удаление. (Рисунок 1)

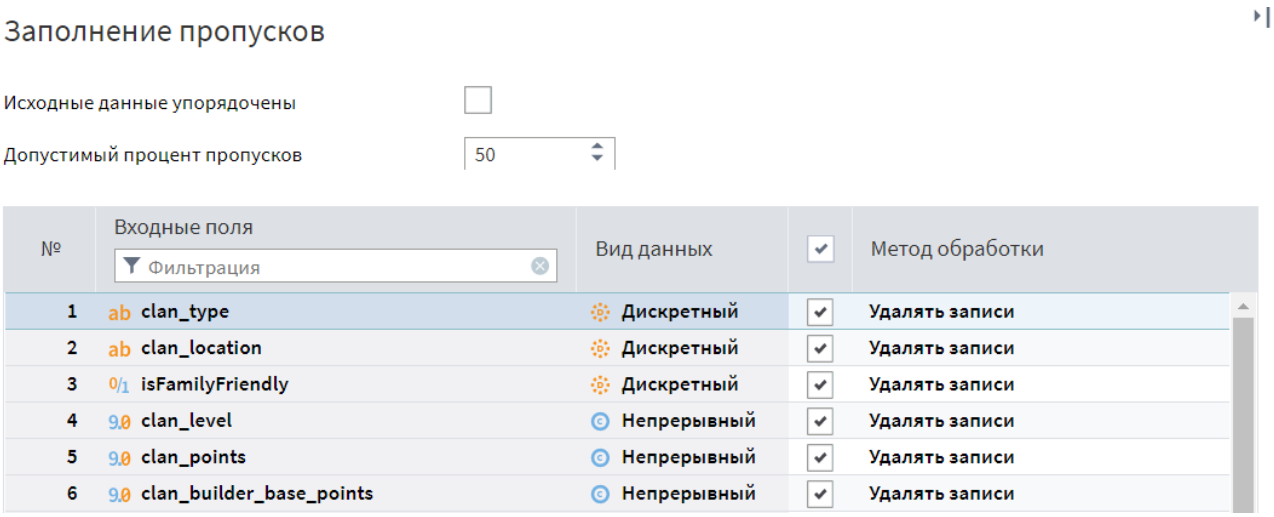


Рисунок 1 — Настройка узла «Заполнение пропусков»

Далее проведём удаление выбросов с помощью узла «Редактирование выбросов». (Рисунок 2) В нашем датасете они присутствуют в столбцах категориальных переменных, в частности в столбце «clan_type».

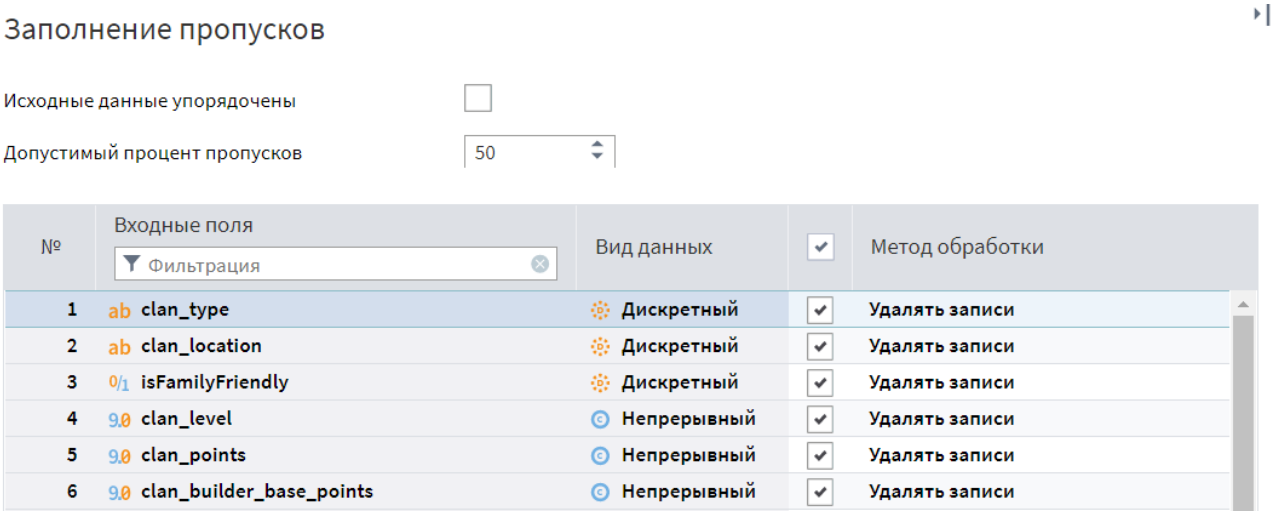


Рисунок 2 — Настройка узла «Редактирование выбросов»

Так же, анализ будет проводиться только для активных кланов. Поэтому используем узел «Фильтр строк» для отсеивания кланов с числом членов, не превышающих ноль. (Рисунок 3)

Рисунок 3 — Настройка узла «Фильтр строк»

Таким образом, мы провели предобработку данных, исключив лишние столбцы и неполные в своём содержании строки. Однако, в дата фрейме всё ещё очень много столбцов, что затрудняет формирование решений и дальнейших действий. Поэтому для комфортного продолжения работы мы применим алгоритм понижения размерности для вещественных столбцов, используя узел «Факторный анализ».

Настроим узел для проведения анализа. На рисунке 4 показана настройка входных столбцов данного узла. Столбцы данных с не заданным назначением в анализе использоваться не будут.

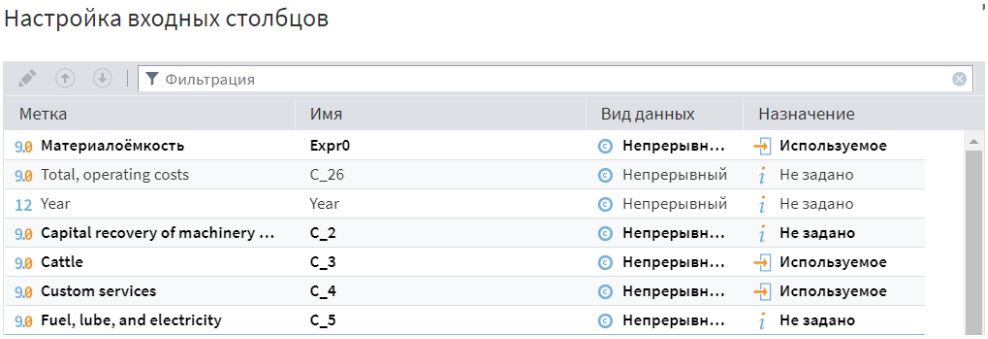


Рисунок 4 — Часть факторов, по которым будет производиться анализ

В качестве критерия значимости факторов используем собственные значения с порогом 1, вращение осуществим методом «Варимакс» для простоты интерпретации результата.

После обучения узла на нижнем выходном порте получаем таблицу факторных нагрузок исходных факторов. (**Error! Reference source not found.**5) Столбцы «Фактор1», «Фактор2», «Фактор3» и «Фактор4» — это новые факторы, которые являются суперпозицией исходных. При этом большой по модулю вес исходных факторов в новых столбцах сигнализирует о корреляции исходных факторов с результирующими.

#	ab Имя	ab Метка	9.0 Фактор1	9.0 Фактор2	9.0 Фактор3	9.0 Фактор4
1	clan_level	clan_level	0,7455822891	0,1801427234	0,4747768944	0,2204347026
2	clan_points	clan_points	0,9379832678	0,1271844569	0,162126564	0,2120554431
3	clan_builder_base_points	clan_builder_base_points	0,9310495644	0,1276573286	0,1566758716	0,2281274833
4	clan_versus_points	clan_versus_points	0,9310495644	0,1276573286	0,1566758716	0,2281274833
5	required_trophies	required_trophies	0,1751255573	0,8246213657	0,1058189784	0,1353835001
6	war_win_streak	war_win_streak	0,3722023395	0,04602697442	0,0600373932	-0,3247670799

Рисунок 5 — Фрагмент таблицы факторных нагрузок

Раскроем смысл полученных факторов. Для этого потребуется воспользоваться мерой адекватности выборки Кайзера-Мейера-Олкина — величиной, используемой для оценки применения факторного анализа. Значения, модуль которых лежит в промежутке от 0,5 до 1, имеют статистически сильную зависимость, когда как значения с модулем, не превышающим 0,5, указывают на недостаточную для формирования выводов зависимость между факторами.

Если вес по модулю больше 0,5, то фактор, на пересечении с которым находится данный вес, является значимым и включается в выборку. Группа значимых исходных факторов образует собой новый фактор «Фактор n» (n — номер нового фактора), при этом подразумевается, что исходные факторы коррелируют между собой.

Проанализируем веса в столбцах «Фактор1», «Фактор2», «Фактор3» и «Фактор4». Согласно критерию Кайзера-Мейера-Олкина, первый фактор зависим от столбцов «clan_level», «clan_points», «clan_builder_base_points», «clan_versus_points», «war_wins», «num_members», «clan_capital_points», «clan_capital_hall_level». Таким образом, он сформирован из всех столбцов, отражающих заработанные в игровом процессе очки и достижения, тем самым неся смысл опыта и могущества кланов.

Второй фактор коррелирует со столбцами «required_trophies», «required_builder_base_trophies» и «required_versus_trophies», тем самым обобщая требования для вступления в клан.

Высокие нагрузки на третий фактор имеют столбцы «war_ties», «war_wins» и «war_loses», что означает, что он несёт смысл военной активности клана.

Четвёртый фактор сформирован на основе столбцов «mean_member_level» и «mean_member_trophies», что означает, что он является обобщением игрового опыта участников клана.

В совокупности с удалением лишних столбцов и пропущенных значений, факторный анализ подводит к концу предобработку данных. Из искомых 16-ти вещественных столбцов были получены четыре, по которым в следующей главе будет проведён статистический анализ.

2.2 Анализ данных

Для сбора статистики воспользуемся визуализатором «Куб», позволяющим в удобном формате производить группировку множества строк по численным и категориальным переменным и получать агрегированную статистику.

Для начала, рассмотрим значения вычисленных в ходе факторного анализа метрик для различных значений типа клана по способу вступления. (Рисунок 6) Факторы были переназваны для удобства визуализации.

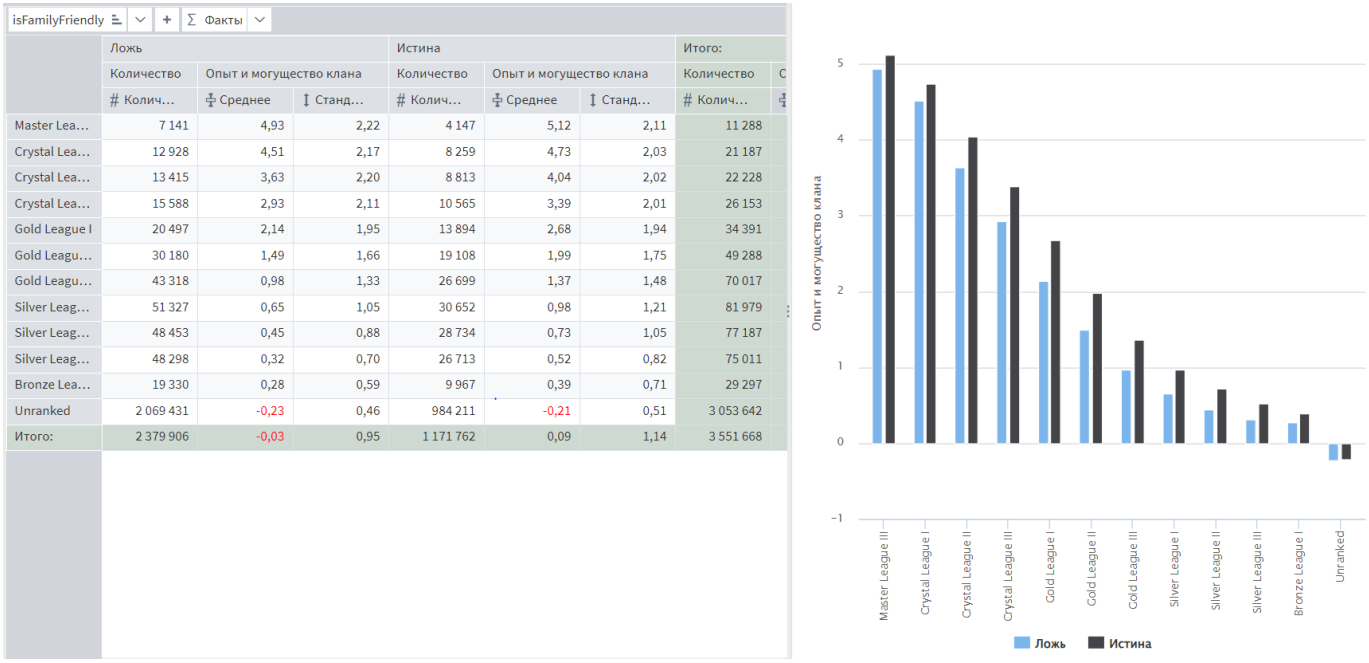
	Количество	Опыт и могущество клана		Требования к вступлению		Могущество игроков		Военная активность	
	# Колич…	Среднее	Станд…	Среднее	Станд…	Среднее	Станд…	Среднее	Станд…
closed	168 692	0,10	1,37	0,83	2,19	0,23	1,37	0,23	1,62
inviteOnly	819 229	0,23	1,41	0,05	1,07	0,29	1,09	0,19	1,47
open	2 563 747	-0,07	0,81	-0,07	0,81	-0,11	0,92	-0,07	0,73
Итого:	3 551 668	0,01	1,02	0,00	1,00	0,00	1,00	0,00	1,01

Рисунок 6 — Параметры типов клана по способу вступления

Как можем увидеть из таблицы, наиболее могущественными кланами являются кланы, вступить в которые можно только по приглашению. Однако, они обладают низкими требованиями на вступление, что объясняется тем, что данный показатель для них не применяется. Эти же кланы имеют наисильнейшую базу игроков. Наименее успешным по всем характеристикам, хотя и наименее требовательным и наиболее численным является класс кланов

с открытым вступлением. Самыми малочисленными и воинственными являются закрытые кланы.

Далее рассмотрим распределение могущества кланов по военным лигам и оценим влияние цензурности на данный параметр. (Рисунок 7)



isFamilyFriendly | Факты

	Ложь			Истина			Итого:	
	Количество	Опыт и могущество клана		Количество	Опыт и могущество клана		Количество	С
	# Колич…	Среднее	Станд…	# Колич…	Среднее	Станд…	# Колич…	
Master Lea…	7 141	4,93	2,22	4 147	5,12	2,11	11 288	
Crystal Lea…	12 928	4,51	2,17	8 259	4,73	2,03	21 187	
Crystal Lea…	13 415	3,63	2,20	8 813	4,04	2,02	22 228	
Crystal Lea…	15 588	2,93	2,11	10 565	3,39	2,01	26 153	
Gold League I	20 497	2,14	1,95	13 894	2,68	1,94	34 391	
Gold Leagu…	30 180	1,49	1,66	19 108	1,99	1,75	49 288	
Gold Leagu…	43 318	0,98	1,33	26 699	1,37	1,48	70 017	
Silver Leag…	51 327	0,65	1,05	30 652	0,98	1,21	81 979	
Silver Leag…	48 453	0,45	0,88	28 734	0,73	1,05	77 187	
Silver Leag…	48 298	0,32	0,70	26 713	0,52	0,82	75 011	
Bronze Lea…	19 330	0,28	0,59	9 967	0,39	0,71	29 297	
Unranked	2 069 431	-0,23	0,46	984 211	-0,21	0,51	3 053 642	
Итого:	2 379 906	-0,03	0,95	1 171 762	0,09	1,14	3 551 668	

Рисунок 7 — Распределение могущества кланов по военным лигам и по цензурности

Как и стоило ожидать, среднее значение могущества оказалось наиболее высоким у наиболее высоких по рейтингу кланов. Однако, так же в среднем по каждой лиге оно является более высоким у тех кланов, что придерживаются цензурной лексике. При этом, такие кланы в каждой лиге присутствуют в меньшинстве.

С военной активностью ситуация является обратной. (Рисунок 8) Наиболее активными участниками конфликтов в среднем выступают кланы, не являющиеся цензурными.

	Ложь			Истина			Итого:	
	Количество	Военная активность		Количество	Военная активность		Количество	Вое
	# Колич…	Среднее	Станд…	# Колич…	Среднее	Станд…	# Колич…	
Master Lea…	7 141	1,52	3,63	4 147	1,48	3,75	11 288	
Crystal Lea…	12 928	1,24	3,47	8 259	1,13	3,41	21 187	
Crystal Lea…	13 415	1,12	3,20	8 813	0,92	3,20	22 228	
Crystal Lea…	15 588	0,96	2,98	10 565	0,71	3,00	26 153	
Gold League I	20 497	0,81	2,73	13 894	0,59	2,68	34 391	
Gold Leagu…	30 180	0,56	2,30	19 108	0,37	2,35	49 288	
Gold Leagu…	43 318	0,31	1,92	26 699	0,18	1,87	70 017	
Bronze Lea…	19 330	0,27	1,17	9 967	0,26	1,52	29 297	
Silver Leag…	48 298	0,21	1,29	26 713	0,15	1,25	75 011	
Silver Leag…	51 327	0,19	1,52	30 652	0,09	1,51	81 979	
Silver Leag…	48 453	0,19	1,32	28 734	0,11	1,32	77 187	
Unranked	2 069 431	-0,06	0,67	984 211	-0,08	0,65	3 053 642	
Итого:	2 379 906	0,01	0,99	1 171 762	-0,01	1,05	3 551 668	

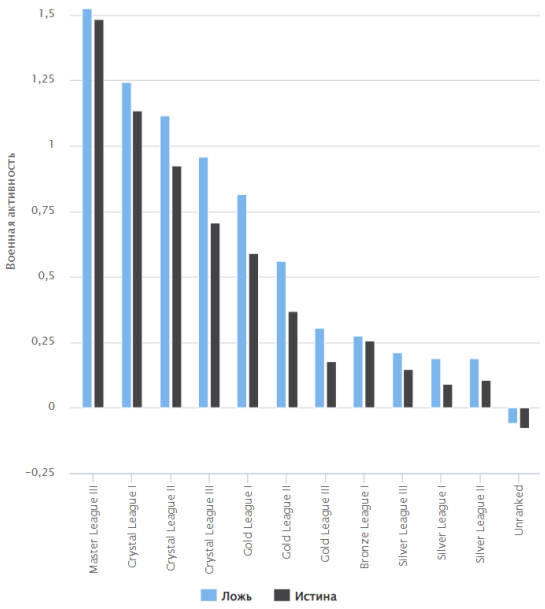


Рисунок 8 — Распределение военной активности кланов по военным лигам и по цензурности

Теперь рассмотрим соответственное распределение средней опытности игроков. (Рисунок 9)

	Ложь			Истина			Итого:	
	Количество	Могущество игроков		Количество	Могущество игроков		Количество	Могу
	# Колич…	Среднее	Станд…	# Колич…	Среднее	Станд…	# Колич…	Ср
Master Lea…	7 141	1,26	1,66	4 147	1,21	1,66	11 288	
Crystal Lea…	12 928	1,03	1,75	8 259	0,97	1,78	21 187	
Crystal Lea…	13 415	0,81	1,51	8 813	0,76	1,37	22 228	
Crystal Lea…	15 588	0,63	1,43	10 565	0,56	1,32	26 153	
Gold League I	20 497	0,57	1,39	13 894	0,44	1,29	34 391	
Gold Leagu…	30 180	0,50	1,25	19 108	0,37	1,20	49 288	
Gold Leagu…	43 318	0,43	1,03	26 699	0,31	1,20	70 017	
Silver Leag…	51 327	0,29	0,93	30 652	0,17	0,96	81 979	
Silver Leag…	48 453	0,10	0,97	28 734	-0,00	1,08	77 187	
Unranked	2 069 431	-0,04	0,96	984 211	-0,04	0,96	3 053 642	
Silver Leag…	48 298	-0,13	0,83	26 713	-0,21	0,84	75 011	
Bronze Lea…	19 330	-0,34	0,79	9 967	-0,40	0,79	29 297	
Итого:	2 379 906	0,00	1,00	1 171 762	-0,00	1,01	3 551 668	

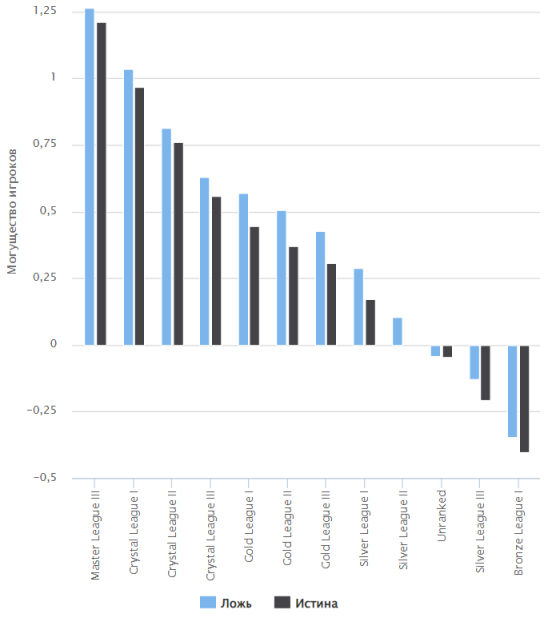


Рисунок 9 — Распределение среднего могущества игроков кланов по военным лигам и по цензурности

Игроки кланов, не поддерживающих цензуру, оказались в среднем более опытными по всем лигам, что контринтуитивно, учитывая, что наибольшим успехом в развитии кланов обладают почитатели цензурной лексики. Это

может говорить о более эффективном использовании ресурсов, военной силы и более оптимальной и продуктивной природе таких кланов в общем.

Так же заметим ассимиляцию наиболее слабых игроков в кланах наинизшей лиги. Это может быть обусловлено неопытностью в выборе клана и ранней тягой к исследованию всех аспектов игры, в том числе боевых.

Так же рассмотрим некоторые статистики по странам.

Найдём страны, в которых игра «Clash of Clans» наиболее популярна. (Рисунок 10)

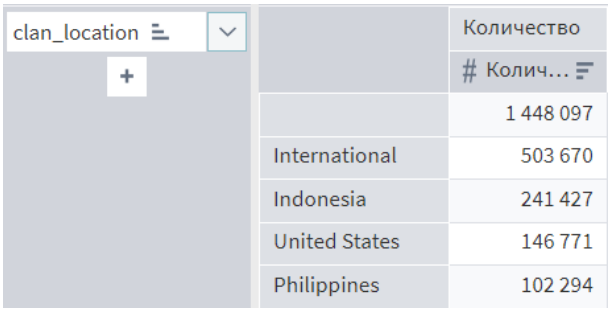


Рисунок 10 — Страны с наивысшей популярностью игры

Чаще всего в данном датасете ячейки столбцов о местоположении оказались не заполнены. Следующим по популярности является класс кланов, местоположение которых является интернациональным. И, наконец, страной, в которой игра является наиболее популярной оказалась Индонезия.

Так же найдём наиболее статистически сильную по кланам стану. (Рисунок 11) Ей оказалась маленькая африканская страна Лесото. Её средний показатель опыта и могущества равняется 1.85, когда как данный показатель у страны, занимающей второе место, равняется 1.09. Сильный отрыв обусловлен высокой концентрацией кланов по приглашению. Такие кланы наиболее успешны по характеристикам своей силы. Более глубокое исследование подтверждает уникальность данной страны, она является необычным феноменом, возможно, является неким мемом среди сильных игроков. [1.2.2]

		Количество	Опыт и могущество клана		Требования к вступлению		Военная активность		Могущество игроков	
		# Колич...	Среднее	Станд...	Среднее	Станд...	Среднее	Станд...	Среднее	Станд...
Lesotho	inviteOnly	248	3,42	2,88	-0,71	1,16	5,00	5,32	0,55	0,92
	closed	53	0,62	2,11	0,80	2,49	2,13	3,79	-0,02	1,30
	open	233	0,45	1,68	-0,08	0,91	0,38	1,90	-0,17	1,12
	Итого:	534	1,85	2,77	-0,29	1,34	2,70	4,57	0,18	1,10

Рисунок 11 — Статистика самой могущественной страны

ЗАКЛЮЧЕНИЕ

Статистический анализ наборов данных является ключевым инструментом развития информационной сферы в целом и, в частности, сферы мобильных игр. Выявление закономерностей помогает улучшению игрового опыта, что является положительной динамикой как для пользователей, так и для разработчиков. Статистические данные позволяют правильно распределять серверные мощности, принимать наиболее комфортные для всех решения в политике игры и её развитии.

Анализируя данные игры «Clash of Clans» мы провели предобработку данных, включающую понижение размерности при помощи факторного анализа, проанализировали страны и популярность игры в них, рассмотрели статистику по военным лигам и приличности кланов и сделали по ним выводы, использовав low-code платформу Loginom. Тем самым цель данной курсовой работы была выполнена.

В ходе данной курсовой работы выполнены следующие задачи:

- выполнена предобработка данных;
- проведён анализ данных;
- использованы знания математической статистики с применением современных средств обработки данных: аналитической платформы Loginom;
- закреплены знания об оформлении документации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1. Савельев В. Статистика и котики — Москва: Издательство АСТ, 2020 — 198 с.

1.2. Колоков А. Заставьте данные говорить — 2022. — 451 с.

1.3. LoginomHelp [Электронный ресурс]. — Режим доступа: https://help.loginom.ru/userguide/processors/scrutiny/factor-analysis.html.

1.4. Ранняя пенсия [Электронный ресурс] / Владислав Кошелев. Материалоёмкость продукции — Режим доступа: https://retireearly.ru/buisness/materialoemkost.

1.5. Ионин В.Г.Статистическая группировка и распознавание некоторых видов распределения вероятностей — Новосибирский государственный университет экономики и управления «НИНХ», 2014 — 13 с.

1.6. Радиченко Г.И.Распределенные вычислительные системы — Южно-Уральский государственный университет, 2012 — 176 с.

2.1 Kaggle [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/datasets/asaniczka/clash-of-clans-clans-dataset-2023-3-5m-clans

2.2 Coc-stats [Электронный ресурс]. — https://www.coc-stats.net/en/locations/32000134/clans/

2.3 Финансовый директор [Электронный ресурс] / Рамазанова Ляйсан Тимуровна. Факторный анализ материалоёмкости: формула расчёта — Режим доступа: https://www.fd.ru/articles/162069-faktornyy-analiz-materialoemkosti-formula-rascheta.

2.4 BaseGroupLabs[Электронный ресурс] — Режим доступа: https://basegroup.ru/deductor/function/algorithm/factor-analysis.

2.5 Водопьянов И. В. Калькулирование материалоёмкости продукции: учебное пособие — Москва: Проспект, 2020 — 177 с.

2.6 Ранняя пенсия [Электронный ресурс] / Владислав Кошелев. Материалоёмкость продукции — Режим доступа: https://retireearly.ru/buisness/materialoemkost.

2.7 LoginomHelp [Электронный ресурс]. — Режим доступа: https://help.loginom.ru/userguide/visualization/cube/index.html

2.8 Matprofi [Электронный ресурс] / Группировка данных. Виды группировок. Перегруппировка. — http://mathprofi.ru/gruppirovka_dannyh.html

ПРИЛОЖЕНИЯ

Приложение А — Сценарий проекта на платформе Loginom.

Приложение А

На рисунке 11 представлен итоговый сценарий проекта на аналитической платформе Loginom.

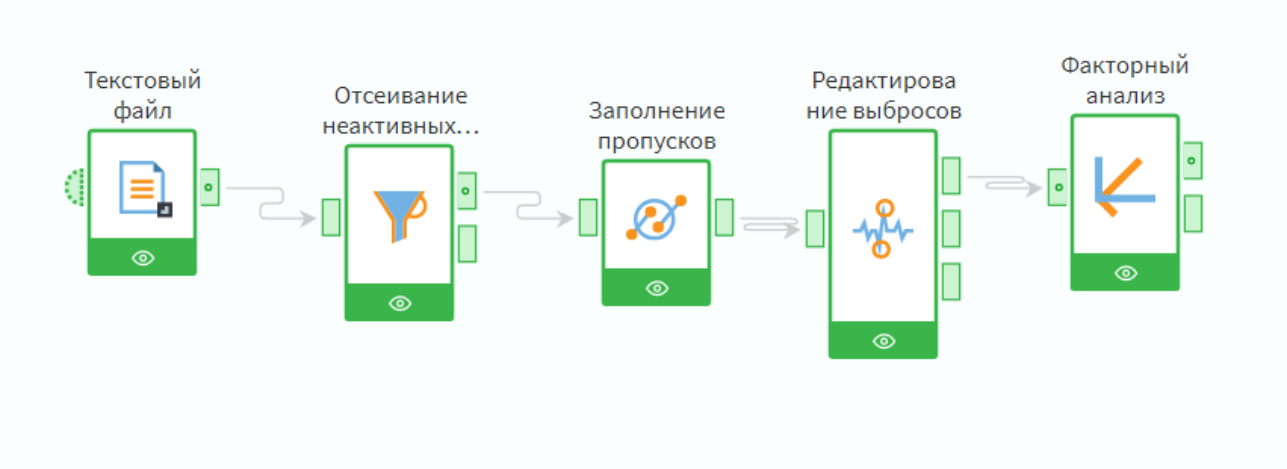


Рисунок 11 — Сценарий на платформе Loginom

Описание сценария:

1. Импорт файла с данными.
2. Отсеивание неактивных кланов.
3. Удаление строк с пропущенными значениями.
4. Удаление выбросов.
5. Факторный анализ.