**Документация на проекта**

# **Въведение**

Предоставеният R скрипт е част от проект за анализ на данни, включващ обработка и манипулиране на данни, предоставени от „Techno Magic Land“. Този код е фокусиран върху почистването и подготовката на данни, които се използват за по-нататъшен анализ.

***Цели на проекта и въпроси, представляващи интерес***

* Идентифициране на повтарящи се посетители: Анализиране на поведението, предпочитанията и взаимодействията на потребителите при множество посещения.
* Анализ на популярността на експериментите: Идентифициране на най-популярните и най-малко популярните експерименти и изследване на възможни зависимости.
* Сегментиране на потребителите (начинаещи): Групиране на тези потребители въз основа на тяхното поведение и възможно обяснение за него.

***Методология и подход***

Проектът ще използва усъвършенствани техники за обработка на данни за почистване, манипулиране и анализ на обширния набор от данни. Фокусът ще бъде върху идентифицирането на модели, корелации и прозрения, които могат да осигурят по-задълбочено разбиране на потребителското поведение и предпочитания в Techno Magic Land. Анализът ще бъде допълнен с подходящи статистически и визуални инструменти, за да се осигури цялостно разбиране на изведените от проекта резултати.

# **Описание на кода**

## **Data understanding and data preparation**

***Импортиране на данните***

* Библиотеки: readxl се използва за четене на Excel файлове.
* Зареждане на данни: Три таблици от файла на Excel TML\_vistors\_case\_study\_data.xlsx се зареждат в R като отделни dataframes – TML\_first\_sheet, TML\_second\_sheet и TML\_third\_sheet.
* Преглед на данни: Функцията View се извиква на всеки dataframe за визуална проверка на данните в R.

***Зареждане на допълнителни библиотеки***

Библиотеките ggplot2, dplyr, tidyr и cluster се зареждат за визуализация и манипулиране на данни.

***Премахване на редовете със записи на служебно издадени карти***

* Първоначален подход: Коментираният код предлага първоначална стратегия за филтриране на персонал въз основа на конкретното условие в казуса (Брой визити до издаване на лична гривна == 0).
* Преработен подход: Въз основа на информация от Алексей Потебня, четвърти лист от същия Excel файл се зарежда като TML\_third\_sheet\_filtered. В тази таблица ние ръчно филтрирахме имената на хората, на които е била издадена служебна гривна. Премахнахме всички записи, при които имената са на кирилица и написани по „правилен“ начин. Например Иван Иванов – на кирилица и с две главни букви.

***Намиране на „staff“ записи и в другите две таблици и премахването им***

Тъй като премахнахме служебните записи ръчно (чрез филтриране в Excel), трябва да намерим кои са тези премахнати записи. Това правим като намерим липсващите имена, които присъстват в нефилтрираната трета таблица, но не присъстват във филтрираната.

Потребителските имена на персонала след това се използват за филтриране на записите на персонала в първата и втората таблица.

***Обработка на липсващи стойности в първата таблица***

* Замяна на NA: Всички стойности на NA се заменят с 0.
* Корекция на дата: За „Дата на излизане“, ако стойността е 0, тя се заменя със стойността на „Дата на влизане“.
* Премахване на дати, в които има нули: Редове, където „Дата на излизане“ е 0, се идентифицират и премахват.

***Обработка на липсващи стойности във втората таблица***

* Краен час на играта: Липсващите стойности в „Край на игра“ се заменят със стойностите от „Начало на игра“.
* Премахване на NA: Редове с NA в „Край на игра“ са идентифицирани за премахване.
* Продължителност минути: Липсващите стойности в „Продължителност минути“ се заменят с 0.

***Относно третата таблица***

Избор на колона: От тази таблица решаваме да ползваме само колоната „Гривна ниво“ (ниво на гривна) от третия лист поради проблеми с данните в други колони. Колоната „Поредна гривна“ съдържа данни за партидния номер на гривната, което не служи за цел на нашия проект. Колона „Брой визити до издаване на лична гривна“ не дава същата информация като колона „Брой визити общо“. Освен това колоната „Общи точки на посещение в момента“ ни показва текущите точки на гривната, а не целите точки от всички експерименти, събрани при всяко посещение. По думите на Алексей Потебня, има „непрехвърлена“ информация за броя точки от минали посещения на някои потребители, което не е отразено в данните. Това обяснява и разминаването на стойностите в колоните брой точки и нивото на гривната. Така например има хора с много малко точки, но най-високо ниво – „Магьосник“. Затова решаваме да работим с колоната „Гривна ниво“, тъй като според неговите думи това е единствената колона, която дава правилна информация.

***Проверка за липсващи стойности***

Скриптът проверява дали има останали липсващи стойности в първата и втората таблица.

***Идентифициране на посетители, които са направили повече от едно посещение***

* Първата таблица с премахнати „staff“ филтрираме по „Дата“ и „USERNAME“, вследствие на което наблюденията в таблицата намаляват и няма повторения.
* Спрямо разясненията на Алексей Потебня, в нова таблица извеждаме само валидните посетители, тоест тези, чието посещение е над 1 час и под 4 часа.
* Намираме броя посещения на всеки един посетител и извеждаме резултата в отделна таблица.
* Извеждаме в нова таблица (customers) само посетителите с повече от едно посещение.

***Идентифициране на посетители, които са останали по-малко от 4 часa и имат повече от едно посещение***

* Смятаме времето за посещение на всеки посетител като от „Дата на излизане“ извадим „Дата на влизане“. Резултатът извеждаме в нова колона в часове, използвайки функцията difftime.
* Оставяме само посетители, чието време за посещение е под 4 часa.
* Групираме данните от филтрираната първа таблица по колоната „USERNAME“ и изчисляваме колко пъти се среща всяко уникално потребителско име. Запаметяваме резултата в таблицата username\_counts.
* Филтрираме таблицата username\_counts да включва само потребителски имена, които се срещат повече от един път.

***Интересно наблюдение***

Първоначално бяхме решили да филтрираме валидните наблюдения по време на посещението по-голямо от 1 час и по-малко от 4 часа (след иформация, която получихме от Алексей Потебня). Обаче ако филтрираме по колоната „time\_per\_visit“ (разликата между колоните „Дата на излизане" и „Дата на влизане“) хората, застояли се по-малко от час, но дошли повече от веднъж, можем да видим, че хората с първо посещение по-кратко от 1 час са склонни да се връщат (има наблюдения). Пример за това необичайно потребителско поведение е потребителското име „ivan2009“. Ето защо решаваме да работим с клиенти, които са филтрирани само по горна граница време на престой – 4 часа.

***Идентифициране на предпочитаните експерименти от типа посетители, описан в горния абзац, и средно време за всеки експеримент***

* От филтрираната вече втора таблица оставяме само посетителите с повече от едно посещение. Запазваме резултата в нова таблица filtered\_rows.
* От тази таблица намираме колко пъти е направен всеки един експеримент и запазваме резултата в нова таблица experiments\_count.
* Визуализираме резултата чрез ggplot.
* От таблицата filtered\_rows смятаме средното време на експеримент: Групираме по колоната „Експонат“ и използваме функцията mean към колоната „Продължителност минути“. Запазваме резултата в нова колона.
* Визуализираме резултата чрез ggplot.

***Обработка по времето за експеримент***

Има някои експерименти, при които продължителността на участник е повече от 4 часа, но предполагаме, че това е без значение. Филтрираме колоната по експерименти, чиято продължителност не надвишава 60 минути, тъй като предполагаме, че ще има промяна в изпробването на експерименти. Например забелязахме, че има един клиент с твърде много време, прекарано на експеримента „Опити с тънка сапунена ципа“, което разваля общите резултати. Ето защо решаваме да го премахнем, тъй като това е някаква аномалия или грешка.

* Изчисляваме максималното време за експеримента „Опити с тънка сапунена ципа“: От таблицата filtered\_rows оставяме само редовете, съдържащи този експеримент. С функцията max намираме максималната стойност в колоната „Продължителност минути“ и запазваме резултата в нова колона.
* Правим същото за експеримента „Как се получават цветовете?“.
* В таблицата filtered\_rows оставяме само експериментите, на които посетителите са прекарали под 60 минути.
* Смятаме средното време на експеримент: Групираме по колоната „Експонат“ и използваме функцията mean към колоната „Продължителност минути“. Запазваме резултата в нова колона.
* Визуализираме резултата чрез ggplot.
* Създаваме boxplot за експеримента „Как се получават цветовете?“.

***Намиране на корелационната матрица***

* В таблицата filtered\_rows създаваме нова колона „visited“ със стойност 1 за всеки ред. Това показва, че посетителят е изпробвал даден експеримент.
* Създаваме нова таблица result\_table, където експериментите са колони, а редовете са посещенията. Всеки отделен посетител има свой ред. Ако човекът е посетил даден експеримент, то в колоната на експеримента стои стойност 1. Ако човекът не е посетил даден експеримент, в колоната стои стойност 0.
* Създаваме корелационна матрица с функцията cor, използвайки данните от таблицата result\_table.
* Визуализираме матрицата с помощта на heatmap за по-голяма прегледност на получените резултати.

## **Клъстерен анализ**

***Идентифициране на посетителите с ранг „Начинаещ“***

От филтрираната трета таблица избираме само посетителите с ниво на гривната „Начинаещ“ и запазваме резултата в нова таблица beginners\_only.

***Създаване на нови таблици за по-нататъшна работа***

От филтрираните първа, втора и трета таблица създаваме нови таблици beginners\_1, beginners\_2 и beginners\_3. Чрез филтриране на данните по колона „USERNAME“ тези нови таблици съдържат само посетителите, които се съдържат в таблицата beginners\_only.

***Намиране на броя посещения на всеки посетител***

* Създаваме нова таблица visits\_beginners, която включва колоната „USERNAME“ от таблицата beginners\_1 и втора колона, в която е посочен броят посещения на всеки човек.
* Преименуваме колоните по подходящ начин.

***Намиране на средното време на всеки посетител***

* Използваме данни от таблицата beginners\_2.
* Правим групиране по „USERNAME“.
* Прилагаме функцията mean към колоната „Продължителност минути“, като изключваме всички NA стойности и запазваме резултата в нова колона „Средно време на човек“.

***Намиране на средните точки на всеки посетител***

* Използваме данни от таблицата beginners\_1.
* Правим групиране по „USERNAME“.
* Прилагаме функцията mean към колоната „Точки за посещението“, като изключваме всички NA стойности и запазваме резултата в нова колона „Среден брой точки на човек“.

***Намиране на броя направени експерименти от всеки посетител***

* Създаваме нова таблица experiments\_per\_customer, която включва колоната „USERNAME“ от таблицата beginners\_2 и втора колона, в която е посочен броят направени експерименти от всеки посетител.
* Преименуваме колоните по подходящ начин.

***Събиране на всичко в една таблица***

Групирайки по общата колона „USERNAME“ във всички от горе посочените таблици, поставяме в една нова таблица средните точки на посетител, средното време на посетител, броя посещения на посетител и броя направени експерименти от посетител.

***Превръщане на колоните в новата в „numeric“***

С помощта на функцията as.numeric, която прилагаме към всяка колона, гарантираме, че стойностите в колоните са числови. Това е задължително, за да направим успешно клъстерния анализ най-накрая.

***Клъстерен анализ***

Спрямо данните в новосъздадената таблица направихме три анализа между различните колони, защото именно те дават резултати, подходящи за интерпретиране.

1. Анализ между колоните „Среден брой точки на човек“ и „Средно време на човек“

* Решаваме да групираме данните в 5 клъстера.
* Първият клъстер включва посетители със средно 30 точки, които са прекарали средно 2-3 минути на експериментите.
* Вторият клъстер включва посетители със средно от 30 до 70 точки, които също са прекарали средно 2-3 минути на експериментите.
* Третият клъстер включва няколко посетители със средно от 85 до 125 точки, които отново са прекарали средно 2-3 минути на експериментите.
* Четвъртият клъстер включва само двама посетители със средно 26 и 42 точки, които са прекарали средно около 20 минути на експериментите.
* Петият клъстер включва двама посетители със средно около 20 точки, които са прекарали средно около 60 минути на експериментите.
* Можем да заключим, че почти всички посетители са прекарали средно 2-3 минути на експериментите и техният среден брой точки не зависи от това (тоест е различен за отделните посетители).
* Въпреки че има двама посетители, които са прекарали средно почти 1 час на експериментите, те не са получили висок брой точки.
* Въпреки че посетителите от третия клъстер са прекарали средно само 2-3 минути на експериментите, те са получили висок брой точки.
* Мнозинството от посетители са получили средно 70 точки максимум.

1. Анализ между колоните „Среден брой точки на човек“ и „Брой посещения“

* Решаваме да групираме данните в 5 клъстера.
* Първият клъстер включва посетители със средно до 45 точки, които имат едно посещение.
* Вторият клъстер включва посетители със средно от 45 до 120 точки, от които почти всички имат едно посещение.
* Третият клъстер включва посетители със средно от 10 до 50 точки, които имат две посещения.
* Четвъртият клъстер включва посетители със средно от 10 до 40 точки, които имат три, четири или пет посещения.
* Петият клъстер включва само двама посетители със средно около 20 точки, които имат едно посещение.
* Можем да заключим, че независимо от броя посещения повечето посетители имат средно до 60 точки максимум.
* Посетителите с висок брой точки имат едно посещение, защото вероятно са решили, че не биха могли да подобрят резултата си.
* Посетителите с не много висок брой точки пък са разделени на две групи – такива, които са се отказали след първото си посещение, и такива, които са направили още няколко посещения, вероятно за да подобрят резултата си.
* Всяко следващо посещение не гарантира по-висок брой получени точки.

1. Анализ между колоните „Среден брой точки на човек“ и „Брой направени експерименти“

* Решаваме да групираме данните в 4 клъстера.
* Първият клъстер включва посетители със средно до 45 точки, които са направили средно до 50 експеримента.
* Вторият клъстер включва посетители със средно от 40 до 120 точки, които са направили средно от 25 до 100 експеримента.
* Третият клъстер включва посетители със средно от 10 до 60 точки, които са направили средно от 50 до 220 експеримента.
* Четвъртият клъстер включва само двама посетители със средно около 20 точки, които са направили средно 22 и 32 експеримента.
* Можем да заключим, че повече от половината посетители не са направили много експерименти (до 50 експеримента) и също така имат малък брой точки (до 40 точки).
* Не можем да кажем, че посетители с много експерименти имат голям брой точки.
* Има само около 10 посетители, които са направили над 100 експеримента, което не е гарантирало по-голям брой точки.
* Не можем да открием голяма зависимост между „Среден брой точки“ и „Брой направени експерименти“.

# **Заключение**

Изводите, които извлякохме от данните, можем да структурираме по три линии.

Първата линия на анализ са хората, които реално съставляват основната ни извадка – след като почистихме данните от шум, извадихме хората, които са посетили Techno Magic Land повече от веднъж, а престоят им е продължил не повече от 4 часа (максимумът време, в което човек може да пребивава там). Тази извадка сама по себе си е много ценна, тъй като може да се направи по-дълбок анализ на поведението на лоялните клиенти и да се направят допускания за потенциалните такива – как могат и те да станат лоялни. Съответно маркетинг и бизнес отделите могат да изведат доста поведенческа и специфична информация, която би им послужила в разработването на успешни маркетинг и бизнес стратегии за в бъдеще.

По-нататък използвахме същата тази извадка, за да направим корелационната матрица на експериментите. Идеята беше да видим дали посещението на някой от експериментите е свързано с посещението на друг. Това изследване дава изключително ценна информация, тъй като тя може да бъде използвана за усъвършенстване на потребителската пътека вътре в Techno Magic Land. Например там, където корелацията е между 0.45 и 0.95, е налице силна корелация, т.е. зависимост на двата посетени експоната, което може да послужи за стратегическото разположение на експонатите в пространството. Също така могат да бъдат добавени някакъв тип пакети от експерименти, а експонатите в тях да бъдат подбрани именно с помощта на този способ (вместо по зони или сфера на експонатите например). Тоест създава се поле за усъвършенстване на целия бизнес като резултат от обработката и анализа на данните.

Третата линия засяга клъстерния анализ, където са анализирани т.нар. „Начинаещи“ в Techno Magic Land. Съответно сме извели средните точки на посетител, средното време на посетител, броя посещения на посетител и броя направени експерименти от посетител. Клъстерите, обособени по някои двойки характеристики, биха били изключително полезни при анализ на потребителското поведение, идентифициране на повтарящ се тип поведение и съответно усъвършенстване на подхода на компанията към всеки един клъстер. Информация за клиентите във всеки един клъстер може да бъде извадена и анализирана в по-голям детайл. Тези последващи анализи могат да спомогнат за по-успешното таргетиране на потребители и съответно това би довело до по-пълноценно извижяване.