**СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ”**

**СТОПАНСКИ ФАКУЛТЕТ**


# Data Prep and Cluster Analysis: Usage-Based Segmentation

**Сергей Филипов - 1EB3100357**

**Мариела Бушева - 3EB3100538**

**Костадин Димитров - 3EB3100566**

**Борислав Иванов - 7EB3100561**

Дисциплина: **Наука за данните в бизнеса и финансите**

Доц. д-р **Боряна Пелова**

**Юни 2025,**

**София**

# Съдържание

**1. Запознаване с бизнеса (Business Understanding)** ..... 3
**2. Запознаване с наличните данните** ..... 4
**3. Литературен преглед** ..... 6
**4. Резултати от анализа на данните** ..... 7
**4.1. Пречистване и подготовка на данните** ..... 7
**4.2. Анализ на разпределенията и корелации** ..... 8
**4.3. Категоризация на usage-характеристики („биниране“)** ..... 10
**4.3.1. Методология и бизнес-логика на биниране** ..... 10
**4.3.2. Обобщение на разпределението по бинове на „Zero user“ и „no user“** ..... 17
**4.4. Клъстеризация с K-Modes** ..... 18
**4.4.1. Първична клъстеризация по услуги** ..... 18
**4.4.2. Финална клъстеризация** ..... 18
**4.5. Профилиране на клъстерите по демография и приходи** ..... 20
**4.6. Анализ на удовлетвореността (NPS/анкета)** ..... 23
**5. Ограничения при провеждане на анализа** ..... 27
**6. Маркетингов фокус** ..... 28
**Заключение** ..... 28
**Източници:** ..... 29

# Списък с таблици и фигури

**Figure 1.** Корелационна матрица на основните usage-показатели ..... 9
**Figure 2.** Boxplot на основните usage-показатели и разпределения на outlier-ите ..... 10
**Figure 3.** Разпределение на клиентите по бинове на гласовата активност в мобилната телефония (MTEL_VOICE_CNT) ..... 11
**Figure 4.** Разпределение на клиентите по обем на мобилния интернет трафик (MTEL_TOT_VOL_MB): хистограма и бинова категоризация ..... 11
**Figure 5.** Разпределение и категоризация на изходящите минути във фиксиран телефон (TEL_OUT_MOU) ..... 12
**Figure 6.** Разпределение и категоризация на броя гледани телевизионни програми (DTV_VIEW_NBR) ... 13
**Figure 7.** Разпределение и категоризация на гледанията на видеосъдържание по заявка (DTV_NR_VOD_1YR) ..... 13
**Figure 8.** Разпределение на клиентите по категории на свален интернет трафик (INT_VOL_DOWN) ..... 14
**Figure 9.** Разпределение на клиентите по категории на стрийминг интернет трафик (INT_VOL_STREAMING) ..... 15
**Figure 10.** Разпределение на клиентите по категории на качен интернет трафик (INT_VOL_UP) ..... 15
**Figure 11.** Разпределение на клиентите по категории на общ интернет трафик (INT_VOL_TOT) ..... 16
**Figure 12.** Разпределение на клиентите по финални клъстери (Cluster Proportions) ..... 20
**Figure 13.** Дял на клиентите по жизнен етап (lifestage) в различните финални клъстери ..... 21
**Figure 14.** Разпределение на клиентите по подкатегории жизнен етап (lifestage details) ..... 22
**Figure 15.** Разпределение на отговорите на въпрос Q1_1 ..... 23
**Figure 16.** Разпределение на отговорите на въпроси Q1_2_1 и Q1_2_2 ..... 24
**Figure 17.** Разпределение на отговорите на въпроси Q1_2_3 и Q1_2_4 ..... 24

**Figure 18.** Разпределение на отговорите на въпроси Q3_3_1 и Q4_2_9B ..... 25
**Figure 19.** Разпределение на отговорите на въпроси Q4_1_1 и Q4_1_2 ..... 26
**Figure 20.** Разпределение на отговорите на въпроси Q4_1_3 и Q5_1_1 ..... 26

**Table 1.** Обобщение на основните usage-колони по услуги след пречистване и импутация ..... 8
**Table 2.** Обобщение на "Zero user" и "no user" категориите по usage-характеристики ..... 17
**Table 3.** Описание на финалните клъстери – модални профили по услуги ..... 19
**Table 4.** Демографски профили на финалните клъстери по език и пол ..... 20

# Въведение

В дигиталната ера потребителското поведение оставя след себе си богата следа от данни – от обема мобилен интернет до броя гледани видеа по заявка. Превръщането на тази информация в полезно знание е предизвикателство, но и възможност. Настоящото изследване си поставя за цел да изведе смислена клиентска сегментация, базирана изцяло на реални поведенчески модели на използване на телекомуникационни услуги.

Използвайки обширен масив от потребителски данни, предоставени под формата на .RData файл (Digital_2015_2016.RData). В рамките на анализа се фокусирахме върху пет ключови категории услуги, отразяващи различни аспекти на потребителското дигитално поведение. Това включва: мобилна телефония (напр. брой разговори и обем мобилен трафик), фиксирана телефония (брой изходящи минути), интернет (трафик при изтегляне, качване и стрийминг), дигитална телевизия (гледаемост на линейни канали и съдържание при поискване), както и OTT съдържание, представено чрез използване на мобилни приложения като Yelo и Triiing, които предоставят телевизионен и гласов достъп извън традиционните канали. Всеки ред в базата данни представлява отделен клиент с конкретна история на употреба, която е подложена на трансформация: почистване, категоризация, конструиране на индикатори и нормализация, с цел подготвяне на данните за клъстерен анализ.

В основата на нашия подход стои алгоритъмът k-modes, избран заради способността си да борави с категорични данни и да разкрива хомогенни групи сред разнообразни профили на поведение. За всяка услуга първоначално се изгражда отделен клъстерен модел, а след това резултатите се интегрират в единна сегментация чрез финална кластеризация. Получените сегменти са впоследствие профилирани спрямо демографски показатели, клиентски жизнен цикъл, приходна стойност и резултати от анкети за удовлетвореност и препоръка (NPS).

## 1. Запознаване с бизнеса (Business Understanding)

Телекомуникационният сектор е изправен пред едно от най-бързо променящите се потребителски поведения в цифровата икономика. В този контекст, поведенческата сегментация се явява ключов инструмент. За разлика от традиционните подходи, които групират клиентите на база демография, абонаментни пакети или общи предпочитания, поведенческата сегментация изследва реалните модели на взаимодействие с услугите – например обема изразходвани мобилни данни, честотата на гледане на телевизия, или интензитета на използване на фиксирана телефония. Това позволява създаването на по-точни клиентски профили и вземане на информирани бизнес решения, основани на фактическа употреба, а не на допускания.

Според рамката CRISP-DM (Cross Industry Standard Process for Data Mining), първата фаза – Business Understanding – включва ясно дефиниране на целите и ограниченията на проекта от гледна точка на бизнеса. В конкретния случай, телекомуникационната компания цели да постигне

по-дълбоко разбиране за потребителската база чрез изграждането на стабилна, интерпретируема сегментация на клиентите въз основа на използването на четири основни типа услуги: мобилна телефония, фиксирана телефония, интернет и цифрова телевизия.

Формулираната бизнес цел е ясна: да бъдат обособени хомогенни групи клиенти с подобно поведение, които впоследствие могат да бъдат таргетирани с конкретни стратегии за маркетинг, задържане или upsell. Паралелно с това, проектът цели да улесни междуфункционалната комуникация в рамките на компанията – например между маркетинг, продажби и обслужване на клиенти – чрез създаване на лесно разбираеми профили, които да служат като база за оперативни и стратегически действия.

## 2. Запознаване с наличните данните

Работният масив „Digital_2015_2016.RData“ представлява мащабна панелна база, съдържаща 264 530 реда – всеки ред отговаря на конкретен клиент за определен месец в периода 2015–2016 година. Данните комбинират богат набор от източници, като включват: оперативни записи от системите на оператора, агрегирани поведенчески индикатори, демографски профили, маркетингови сегменти и резултати от клиентски анкети.

Типовете променливи и ключови колони са:

***Уникален идентификатор:*** CUSTOMERNUMBER – идентифицира клиентите, но се срещат и дубликати, тъй като един и същ клиент присъства в различни времеви периоди (над 81 000 дублирани стойности).

***Времеви индекси:*** MONTH_CODE (например 201504), CUSTOMER_RESPONSE_DATE – позволяват надлъжно проследяване на поведението и промените в абонаментния статус.

***Флагове за услуги:*** Променливи като HAS_DTV, HAS_INTERNET, HAS_MOBILE, HAS_VOICE, CUS_FL_DTV, CUS_FL_INT и др. дават бинарна индикация (0/1 или Yes/No) за наличие или ползване на съответната услуга.

***Поведенчески usage-сегментации:***

– YELO_APP_USAGE и TRIIING_USAGE (non user, light user, medium user, heavy user);

– Флагове като CUS_FL_MTEL (използва мобилна телефония), CUS_FL_ATV (аналогова телевизия), CUS_FL_TEL (фиксирана телефония);

– Обеми на трафик: HMS_VOL_MB_DOWN_TOTAL, INT_VOL_DOWN, MTEL_TOT_VOL_MB и други – измерват обем на използвани MB за различни услуги.

***Демографски и финансови данни:***

– CUS_AGE (от 18 до 98 г., медиана ~77), CUS_SEX, CUS_LIFESTAGE и CUS_LIFESTAGE_DETAILS – описват възрастов и семеен профил.

– CUS_REVENUE_SEGMENT, CUS_VALUE, CUS_VALUE_MARGIN, CUS_VALUE_REVENUE – стойностни индикатори за приход, печалба и общ разход по клиента. Пример: CUS_VALUE – медиана 76.91, средна стойност 79.47, но максимални стойности достигат 4523.13 (индикативно за outliers).

– Адрес и регион: CUS_ZIP, CUS_CABLE_OWNER – позволяват географски анализ.

**Технически профил:** Колони за устройства и операционни системи:

– FL_DEVICE_IPAD, FL_DEVICE_IPHONE, FL_DEVICE_IPOD, FL_OS_ANDROID, FL_OS_IOS, FL_OS_WINDOWS, FL_SMART_DEVICE – дават детайл за видовете устройства и ОС, с които клиентът използва услугите.

– Характеристики на използван хардуер – например STBTYPE (тип на set-top-box устройството, като "GEN3" и др.).

***Променливи от анкетни и поведенчески източници:*** Голям брой въпросници и скали за удовлетвореност, препоръка, навици и нагласи – например NPS, скали тип Q1_1, Q2_2_1A, LG_BE_DIG_WEBSITE_SCALE11 и др.

***Качество на данните:***

– Значителна част от колоните съдържат липсващи стойности – средно около 4% за базови флагове, но до над 60% при анкетни и маркетингови променливи.

– Ключови usage-флагове като CUS_FL_DTV, CUS_FL_INT, CUS_FL_MTEL имат липси около 4% (над 10 000 реда).

– Финасови показатели като CUS_VALUE_3M_AFTER_SURVEY имат до 31% липсващи стойности.

– Някои usage-променливи (например за приложението Triiing) и анкетираните скали са със 70–90% липси, което изисква специфична стратегия за импутация или филтриране при анализа.

***Разпределение:*** флагови usage-променливи:

– CUS_FL_DTV (използва цифрова телевизия): 238 740 с "1", 15 277 с "0"

– CUS_FL_MTEL (използва мобилна телефония): 125 262 с "1", 128 755 с "0"

– Това показва, че базата покрива различни клиентски типове – едновременно многопродуктови и "single-play" абонати.

***Сегментни и групиращи променливи:***

– SEGMENT, PRODUCTMIX, BUNDLENAME, CUS_REVENUE_SEGMENT, CUS_SERVICE_SEGMENT – позволяват отделяне на клиенти по тип услуга и абонамент.

– Маркировки за промоции, пакетни комбинации, брой линии, lifetime metrics.

***Други специфики:***

– Има над 20 usage-индикатора само за интернет (обеми по тип протокол: FTP, HTTP, streaming, mail, p2p).

– За телевизия: отделни броячи за гледаемост по канал, по вид съдържание (VOD, TVOD, SVOD), с технически lifetime за box-устройствата.

– Индикатори за churn (отлив), портвания, комуникационни и финансови churn събития по адрес.

В обобщение данните са едновременно мащабни и хетерогенни – включват както стриктно структурирани usage лога, така и анкети, финансови и демографски показатели. Основно предизвикателство е управлението на липсващи стойности. Това налага внимателна подготовка и стандартизация на променливите преди провеждането на какъвто и да е клъстерен анализ или сегментация.

## 3. Литературен преглед

В процеса на анализа и клъстеризацията на клиентското поведение и потребление в телекомуникационния сектор, особено при наличие на многомерни и смесени данни, е от ключово значение изборът на подходящи алгоритми и техники за обработка. Основните методи, използвани в настоящия проект, се опират на утвърдени подходи, описани в научната литература.

Работа с категорийни данни и K-modes клъстеризация: Тъй като голяма част от реалните бизнес данни съдържат категорийни признаци, използването на класическия алгоритъм k-means се оказва неприложимо без трансформация на данните. В тази връзка, ключов момент в проекта е приложението на алгоритъма k-modes, който представлява естествено разширение на k-means за категорийни атрибути. Алгоритъмът k-modes въвежда нова мярка за несходство, използва модите вместо средните стойности и по този начин позволява ефективно групиране на големи масиви от чисто категорийни или смесени данни. Този подход е подробно описан в статията на Zhexue Huang (1998), където се подчертава неговата скалируемост и способността му да работи с огромни по размер данни, без да се губи интерпретируемостта на резултатите. В проекта, чрез биниране на непрекъснатите променливи в категории и създаване на отделни бинове за всяка услуга, бяха създадени основните входни данни за k-modes клъстеризацията, което е в пълно съответствие с препоръчаните от Huang техники. Освен това, k-modes демонстрира предимство пред другите подходи при анализ на клиентски профили, тъй като крайният резултат – модите на клъстерите – са лесно интерпретируеми за нуждите на бизнес потребителите (Huang, 1998).

Намаляване на размерността и подготовка на признаците: Преди клъстеризацията често се налага обработка на многомерни масиви от данни и извеждане на най-съществените характеристики. В тази посока, анализа на главните компоненти (Principal Component Analysis – PCA) е един от най-широко използваните методи за намаляване на размерността, както и за идентифициране на скритата структура в данните. Класическата монография на I.T. Jolliffe (2002) разглежда подробно теоретичните основи и практическите приложения на PCA, като подчертава, че основната цел е редуциране на голям набор взаимно свързани променливи до по-малък брой некорелирани главни компоненти, които запазват възможно най-голяма част от вариацията в оригиналния масив. В нашия анализ, въпреки че финалният клъстерен модел е на база категорийни променливи, предварителната изследователска фаза включваше и анализ на корелационната структура между основните количествени характеристики чрез PCA и корелационни матрици, с цел разбиране на връзките между услугите (Jolliffe, 2002).

Клъстеризация и бизнес приложение: При анализа на големи клиентски бази в реални бизнес условия е важно избраните методи да позволяват не само статистическа коректност, но и бизнес интерпретация на резултатите. В своя статия Huang (1998) показва, че k-modes не просто групира ефективно обекти, но и прави възможно описанието на всеки клъстер чрез неговите модални характеристики – нещо изключително ценно при сегментиране на потребители според начина им на използване на услуги. Така се гарантира, че всяка група може да бъде ясно профилирана и използвана за последващи бизнес решения.

В изследването на Abdul-Rahman et al. (2021) посредством алгоритъма *k-modes* са дефинирани три ясно разграничими клиентски сегмента—„Potential High-Value Customers“, „Low-Value Customers“ и „Disinterested Customers“. След това използват *Decision Tree Classifier* (CART) с Gini критерий и 10-fold кръстосана валидация за профилиране на всеки сегмент, постигащо точност от 81.3%. Този двустепенен подход (първо клъстеризация с k-modes, след това класификация) значително улеснява интерпретацията на профилите и дава конкретни насоки за маркетингови стратегии—например таргетиране на премия клас клиенти с персонализирани

оферти или активиране на неангажирани клиенти с подходящи кампании (Abdul-Rahman et al. 2021). Методологията в твоя проект следва много сходна логика: биниране на основни количествени показатели (MOU, MB, VOD и т.н.), прилагане на k-modes за групиране и описване на модалните профили—което подсигурява директна приложимост за бизнес решения, също както в случая на Abdul-Rahman и колеги.

На база на направения литературен обзор и в съответствие с доказаните в научната литература подходи, в проекта е избран методът k-modes клъстеризация, като най-подходящ за работа с категорийни и бинирани данни, каквито се получиха след трансформация на основните количествени показатели. Използваната методология следва логиката на доказани двустепенни модели от практиката (като този на Abdul Rahman et al., 2021), прилагайки я в контекста на реални телекомуникационни данни. Предварителната обработка чрез биниране, анализ на корелации и използване на PCA улесни изграждането на смислени входни признаци за клъстеризация, а крайното групиране позволи ясно интерпретируемо профилиране на клиентите, което е напълно съвместимо с нуждите на бизнес анализа и вземането на решения.

## 4. Резултати от анализа на данните

Следвайки дефинираната методология и на база предварително обработените и категоризирани данни, в този раздел се представят емпиричните резултати от прилагането на клъстерния анализ. Извършен е детайлен преглед на получените клъстери, тяхната структура и значимост от бизнес гледна точка.

### 4.1. Пречистване и подготовка на данните

Първоначалният масив от данни съдържа 264 530 наблюдения и 471 променливи, отнасящи се до потреблението и характеристиките на клиенти на телекомуникационни услуги за периода 2015–2016 г. Още в началото бяха установени 10 513 записа с липсващи стойности по основните флагове за използване на услуги (CUS_FL_MTEL, CUS_FL_INT, CUS_FL_DTV и CUS_FL_TEL). За да се осигури качество и последователност на анализа, от по-нататъшното изследване бяха изключени всички записи, в които едновременно липсва информация за всички ключови услуги. Така работният масив е редуциран до само тези клиенти, за които има поне една активна услуга.

В хода на предварителната обработка е взето решение да се анализира само най-актуалната информация за всеки клиент. За целта всички записи бяха сортирани по клиентски номер и хронологичен код на месеца, след което за всеки клиент се запази само последният наличен запис. Това позволява реален профил на клиента към края на периода, елиминирайки потенциални повторения и исторически артефакти. Допълнително, за да се изследват само индивидуални крайни клиенти, от извадката бяха премахнати всички записи, в които клиентът е обозначен като „Business“ по променливата CUS_SEX. Този подход гарантира, че в анализа участват само физически лица, чието потребление може да бъде сравнително и сегментирано по уместен начин.

След като бяха премахнати некоректните и непълни записи, данните преминаха през процес на импутация на липсващи стойности за ключовите количествени показатели, специфични за всяка от основните телекомуникационни услуги. При всеки клиент с активна услуга всички липсващи стойности в съответните usage-колони бяха заместени с нула. Тази процедура е приложена поотделно за фиксирана телефония, мобилна телефония, дигитална телевизия и интернет, като за всяка група се използваха съответните променливи. Така например, ако даден

клиент разполага с интернет услуга, но за някой от измерителите не е отчетена стойност, това се интерпретира като реална липса на използване, а не като грешка в измерването.

След импутацията е създадено обобщение на състоянието на usage-колоните според наличието на услуга. За целта колоните бяха групирани по вид услуга, а за всяка от тях се изчислиха следните показатели: брой липсващи стойности при неактивна услуга, брой липсващи стойности при активна услуга, брой нулеви стойности при активна услуга, както и брой наблюдения с реална употреба (стойност по-голяма от нула).

**Table 1.** Обобщение на основните usage-колони по услуги след пречистване и импутация

| Услуга | Колона | NaN (flag=0) | NaN (flag=1) | Zero values (flag=1) | Non-null > 0 (flag=1) |
|---|---|---|---|---|---|
| DTV | DTV_NR_LINES | 11146 | 0 | 0 | 161537 |
| DTV | DTV_NR_BOXES | 11146 | 0 | 73 | 161464 |
| DTV | DTV_LIFETIME | 11146 | 0 | 116 | 161421 |
| INT | INT_NR_LINES | 12221 | 0 | 0 | 160462 |
| INT | INT_LIFETIME | 12221 | 0 | 150 | 160312 |
| INT | INT_VOL_TOT | 12221 | 0 | 446 | 160016 |
| INT | INT_VOL_DOWN | 12221 | 0 | 448 | 160014 |
| INT | INT_VOL_UP | 12221 | 0 | 486 | 159976 |
| INT | INT_USAGEDAYS | 12221 | 0 | 492 | 159970 |
| DTV | DTV_FL_INTERACTIVE | 11146 | 0 | 1605 | 159932 |

В Таблица 1 е показан откъс от това обобщение за десетте най-често използвани usage-показателя. Например, за променливата DTV_NR_LINES, която се отнася до броя на използваните цифрови телевизионни линии, при неактивна услуга има 11 146 липсващи стойности, докато при активна услуга липсващи стойности не се наблюдават и всички 161 537 случая са с отчетена ненулева стойност. Подобна картина се наблюдава и при други ключови показатели. При INT_VOL_DOWN (обем на свалени данни по интернет) отново се виждат 12 221 липсващи стойности при неактивни клиенти, а при активни всички имат попълнени данни, като в 160 014 случая е отчетена реална употреба (стойност над нула), докато само 448 са с нулево потребление.

Сходни зависимости се наблюдават и за други usage-променливи:

- **DTV_NR_BOXES**: 11 146 липсващи при неактивни, 0 при активни; 73 активни с нулево потребление, а в 161 464 случая има реална употреба.
- **INT_NR_LINES**: 12 221 липсващи при неактивни, 0 при активни; всички 160 462 активни случая са с реална стойност.
- **INT_LIFETIME**: 150 активни с нулево потребление, а 160 312 с реална употреба.

Това обобщение потвърждава, че след импутацията и пречистването, данните са напълно готови за следващите етапи на анализа, като не съществуват систематични пропуски или аномалии в основните usage-колони.

## 4.2. Анализ на разпределенията и корелации

В тази стъпка бяха изследвани взаимните зависимости между основните количествени показатели за потребление чрез изграждане на корелационна матрица. Анализът обхваща променливи, свързани с основните услуги – мобилна телефония (MTEL_VOICE_MOU и MTEL_TOT_VOL_MB), фиксирана телефония (TEL_OUT_MOU), дигитална телевизия

(DTV_VIEW_NBR, DTV_NR_VOD) и интернет (INT_VOL_DOWN, INT_VOL_UP, INT_VOL_STREAMING, INT_VOL_TOT).

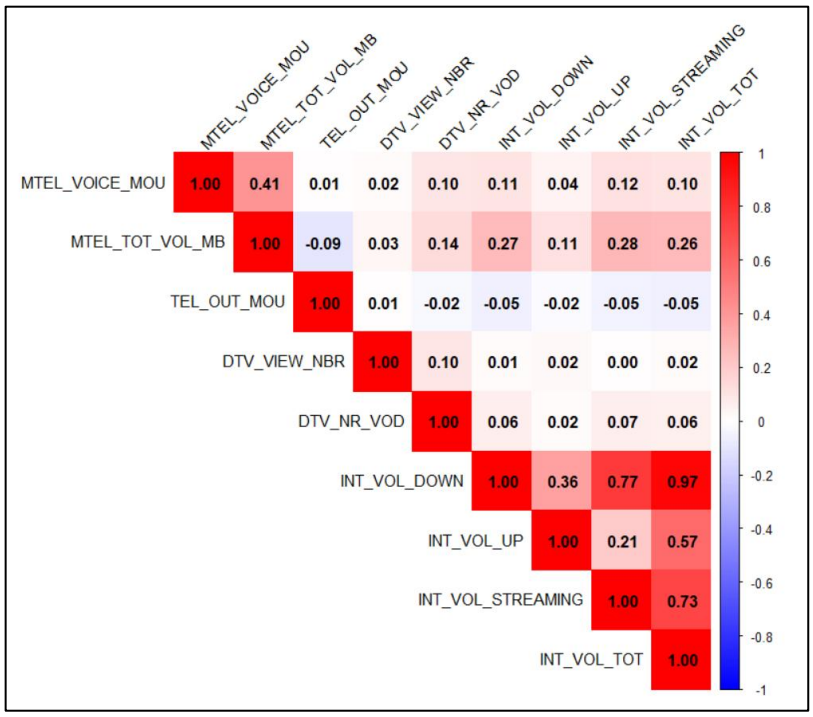


**Figure 1.** Корелационна матрица на основните usage-показатели

На Фигура 1 е визуализирана получената корелационна матрица. Най-висока е корелацията между различните показатели за интернет потребление – например INT_VOL_DOWN (сваляне), INT_VOL_UP (качване), INT_VOL_STREAMING и INT_VOL_TOT (общо потребление), които показват стойности между 0.73 и 0.97. Това е логично, тъй като общият обем на интернет трафика е функция на всички подвидове активност. От друга страна, зависимостта между потреблението на мобилни и фиксирани услуги е много слаба или почти липсваща (корелационни коефициенти в диапазона от -0.09 до 0.14), което говори за сравнително независими потребителски профили между отделните услуги. Сходно слаби са и връзките между мобилните услуги и дигиталната телевизия.

Паралелно с това е направен и анализ на разпределенията и възможните отклонения (outliers) в основните usage-показатели чрез boxplot-визуализация (Фигура 2). По всички показатели ясно се открояват единични наблюдения с изключително високи стойности, които представляват потенциални аномалии или малка група екстремни потребители. Най-отчетливо това се вижда при интернет-трафика, където някои клиенти са регистрирали високи обеми на използване, докато преобладаващата маса е концентрирана при ниски стойности. Сходни, макар и по-ограничени, отклонения се наблюдават и при мобилните услуги и дигиталната телевизия.

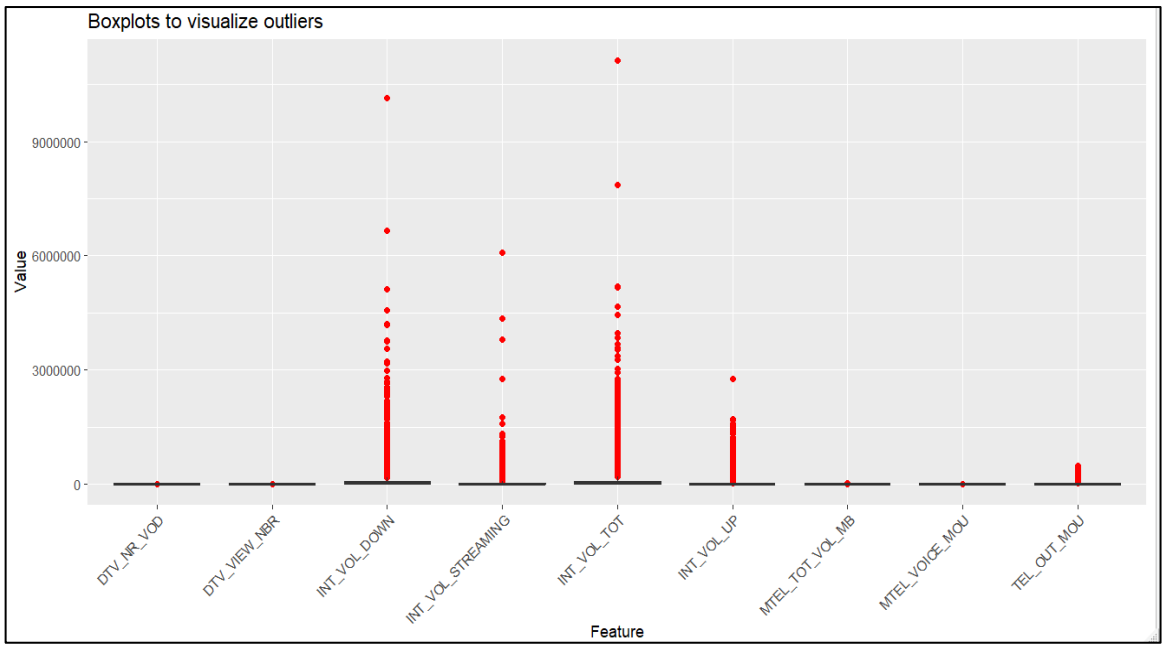


**Figure 2.** Boxplot на основните usage-показатели и разпределения на outlier-ите

Тези визуализации потвърждават както слабото преплитане между различните видове услуги, така и наличието на сравнително малка група „екстремни“ потребители, които се открояват от основната маса по отношение на своето потребление.

## 4.3. Категоризация на usage-характеристики („биниране“)

След пречистването и обработката на основните usage-колони, следващата стъпка в анализа е категоризацията на количествените характеристики в отделни групи („бинове“). Този процес не само улеснява интерпретацията на различните нива на потребление, но и е абсолютно необходим за последващата клъстеризация, тъй като алгоритъмът K-Modes изисква категорийни входни данни. В следващите подточки се описва методологията на биниране, както и резултатите от разпределението на клиентите по формираните категории.

### 4.3.1. Методология и бизнес-логика на биниране

Границите на биновете за различните usage-характеристики са определени, като се съчетава бизнес логика и емпирично наблюдение на разпределенията. За всеки показател се анализираха разпределението, медианата, квартилите и прагът на outlier-ите, което позволи гъвкаво определяне на категориите: „Zero user“, „Low“, „Medium“, „High“, „Very High“, както и специална категория „no user“ за клиенти без активна услуга.

Графиката илюстрира разпределението на клиентите според биновете на гласовата им активност в мобилната телефония. Прави впечатление, че най-голямата група са т.нар. „no user“ – клиенти, които не използват гласова услуга в мобилната мрежа (над 80 000 души). Това са предимно потребители с други услуги или абонати.

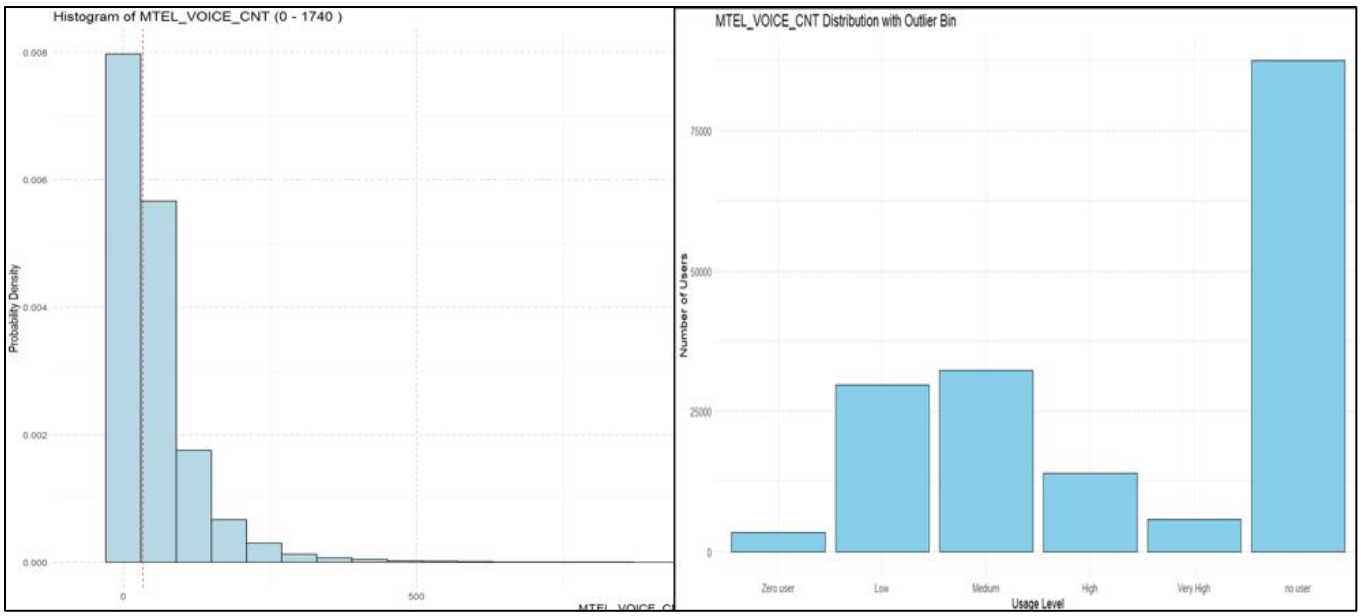


**Figure 3.** Разпределение на клиентите по бинове на гласовата активност в мобилната телефония (MTEL_VOICE_CNT)

Сред активните ползватели най-голям дял имат клиентите с „Medium“ и „Low“ usage (съответно около 30 000 за всеки бин), а най-малка е групата на „Very High“ – едва няколко хиляди абонати. Това разпределение показва отчетлива поляризация между липсващи и умерени ползватели, докато много високата гласова активност е по-рядко срещана. Отчетливо се вижда и наличието на значим брой клиенти с нулева активност („Zero user“), които са регистрирани като потребители на услугата, но не са осъществили нито едно изходящо обаждане.

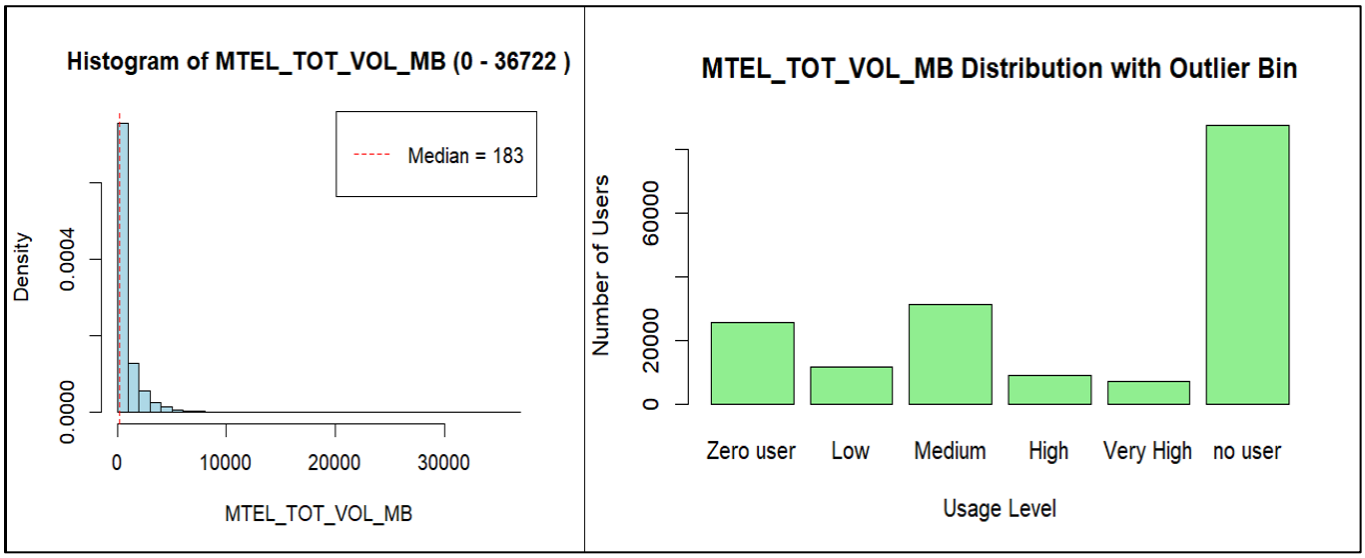


**Figure 4.** Разпределение на клиентите по обем на мобилния интернет трафик (MTEL_TOT_VOL_MB): хистограма и бинова категоризация

На лявата част на фигура 4 е представена хистограма на разпределението на обема на мобилния интернет трафик (в MB) за всички клиенти с активна мобилна услуга. Разпределението е силно дясно-асиметрично – повечето клиенти имат много нисък месечен обем (медианата е само 183 MB), докато малък брой потребители генерират значително по-висок трафик. Това ясно се вижда от „опашката“ вдясно, където има редки, но екстремни стойности.

Дясната графика показва разпределението на клиентите по usage-бинове – от „Zero user“ и „Low“ до „Very High“ и „no user“. Най-голям е делът на клиентите, които не използват мобилен

интернет („no user“), следвани от тези с „Zero user“ (регистрирани, но с 0 трафик), както и групите с „Medium“ usage. Клиентите с „Very High“ трафик са относително малко, което потвърждава силната концентрация на малко, но интензивни потребители в групата на най-високите стойности.

Това разпределение е характерно за телекомуникационните услуги – преобладават ниските нива на потребление, докато много активните потребители са малък процент, но генерират значителна част от общия трафик.

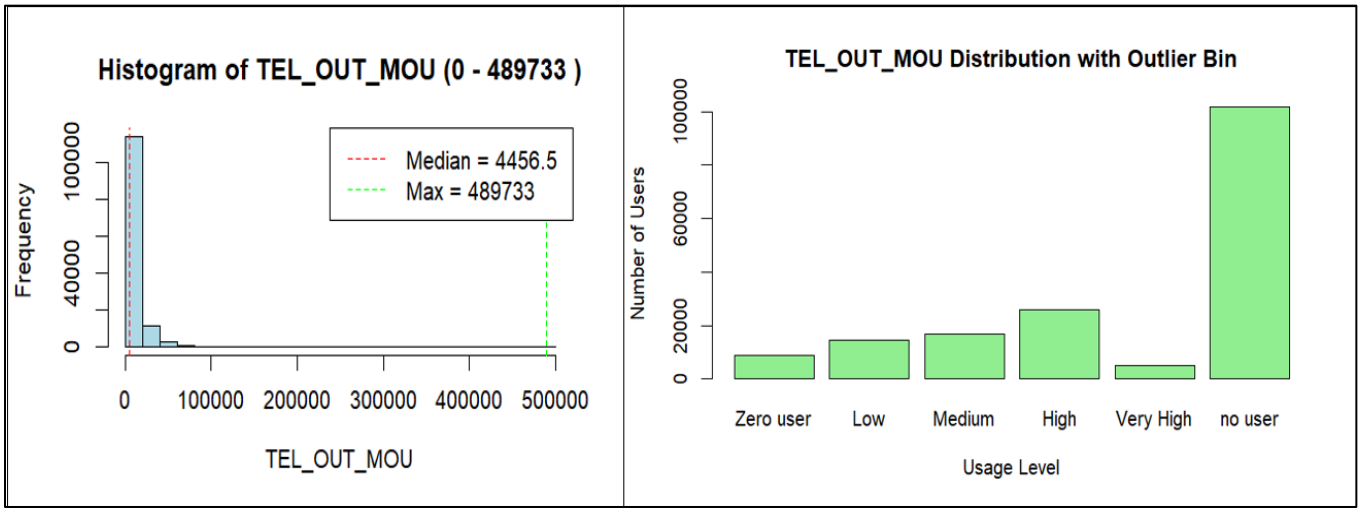


**Figure 5.** Разпределение и категоризация на изходящите минути във фиксиран телефон (TEL_OUT_MOU)

Графиките илюстрират разпределението на изходящите минути във фиксираната телефония (TEL_OUT_MOU) сред потребителите. Хистограмата вляво ясно показва силна дясна асиметрия – преобладаващата част от клиентите извършват малко на брой разговори, докато малък дял от наблюденията се отличава с изключително високи стойности (максимумът достига до 489 733 минути). Медианата (4456.5 минути) се намира много по-близо до ниските стойности, което потвърждава наличието на изолирани, но екстремни потребители.

Диаграмата вдясно показва разпределението на потребителите по дефинирани usage-категории, като най-голям дял отново са „no user“ – клиенти без отчетено потребление във фиксирана телефония за разглеждания период. Сред активните клиенти най-многобройни са тези в ниските и средните usage-групи, докато групата „Very High“ обхваща малка, но отчетлива част от клиентската база, които значително се различават от типичното поведение.

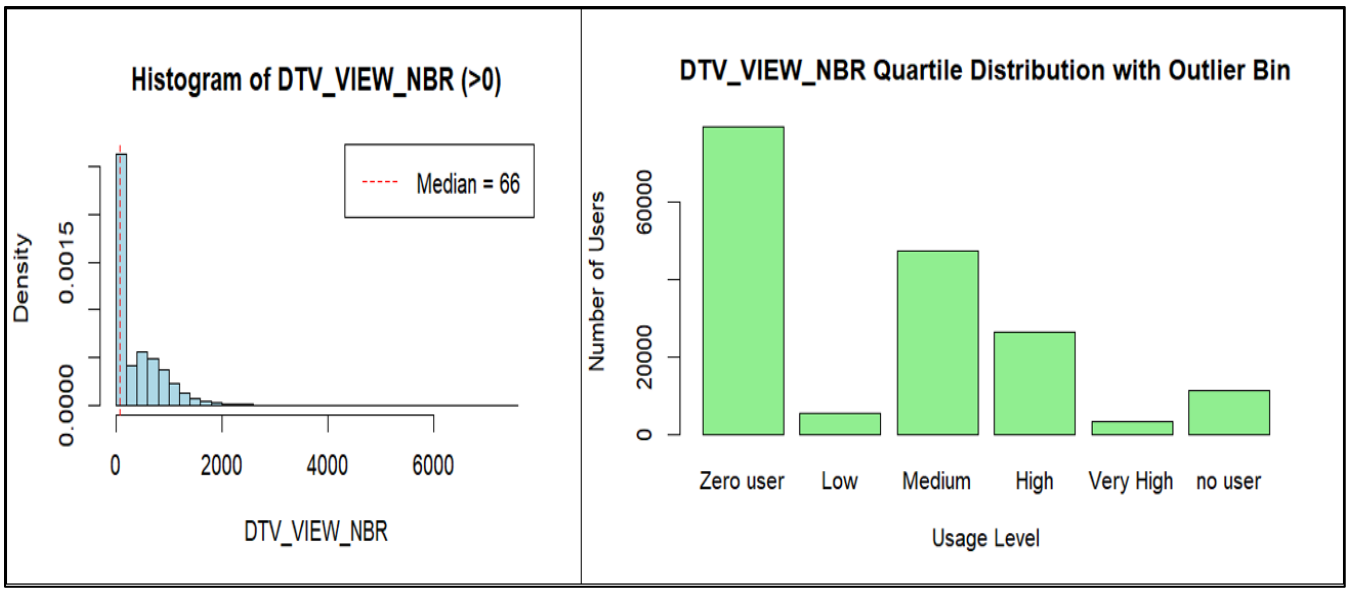


**Figure 6.** Разпределение и категоризация на броя гледани телевизионни програми (DTV_VIEW_NBR)

Двете графики визуализират начина, по който се разпределя броят гледани телевизионни програми (DTV_VIEW_NBR) сред потребителите на дигитална телевизия. Хистограмата вляво показва, че разпределението е силно дясно-скосено – повечето клиенти гледат относително малък брой програми (медианата е 66), докато по-големите стойности са рядкост и създават дълга „опашка“ в дясно.

На barplot-а вдясно е представено разпределението на клиентите по usage-категории. Най-голямата група са „Zero user“ – клиенти с отчетена услуга, но без реално използване през разглеждания период. Значителна част попадат и в категория „Medium“, докато групите „High“ и „Very High“ са много по-редки, което показва наличие на интензивно гледащи, но относително малобройни клиенти. Наблюдава се и отделна категория „no user“ – клиенти, при които няма никаква активност или не е налична телевизионна услуга.

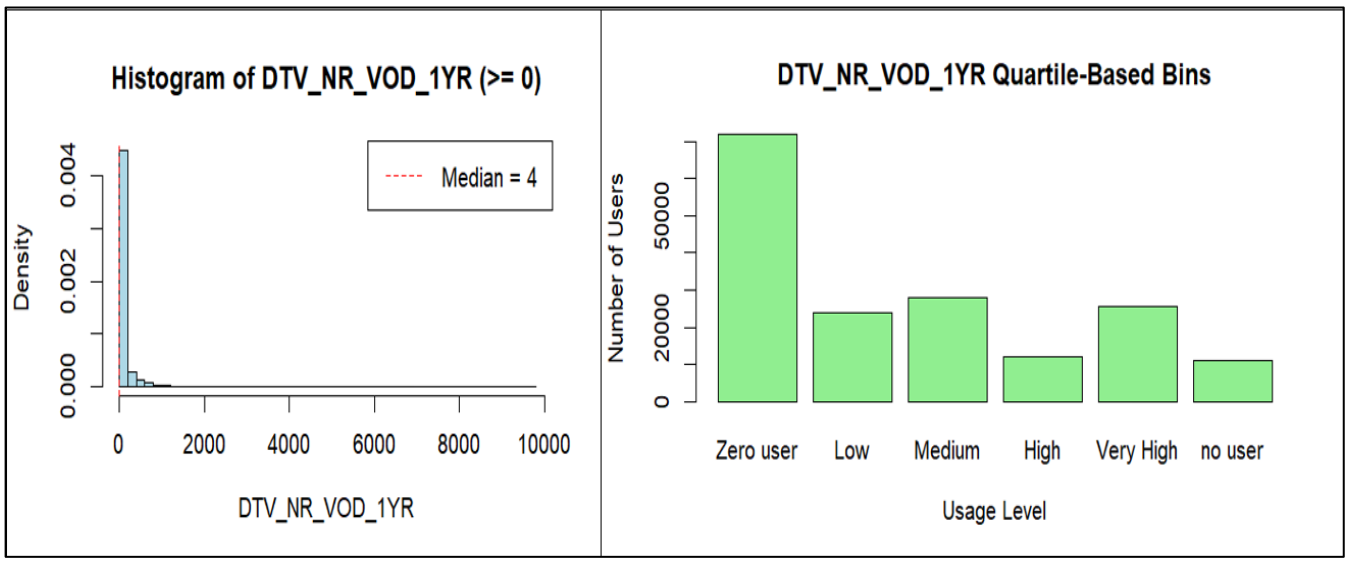


**Figure 7.** Разпределение и категоризация на гледанията на видеосъдържание по заявка (DTV_NR_VOD_1YR)

Графиките представят анализа на годишния брой гледания на видеосъдържание по заявка (Video on Demand, VOD) за потребителите на дигитална телевизия. В лявата част на фигурата хистограмата ясно показва, че голяма част от клиентите имат много нисък или нулев брой гледания – медианата е само 4, което означава, че половината от абонатите са гледали до 4 VOD

предавания за цяла година. Разпределението е силно дясно-скосено – с единични клиенти с изключително висок брой гледания.

Barplot-ът вдясно илюстрира категоризацията по usage-бинове. Отново доминират „Zero user“ – клиенти с активна услуга, но без нито едно гледане на VOD. Категориите „Medium“ и „Very High“ са също видимо представени, което говори за наличие на сегмент от потребители с висока или много висока активност, въпреки че те са сравнително малобройни. Категорията „no user“ обобщава клиентите без достъп до тази услуга.

Данните потвърждават, че за повечето домакинства услугата VOD остава рядко използвана, но има и ясно изразени групи с интензивна консумация.

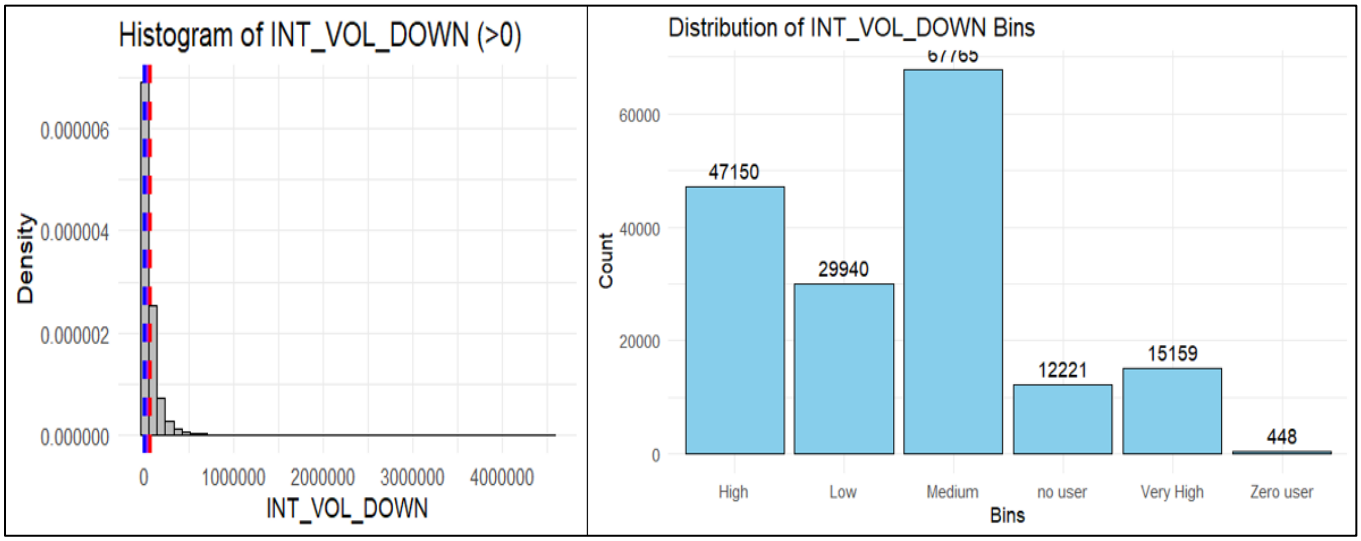


**Figure 8.** Разпределение на клиентите по категории на свален интернет трафик (INT_VOL_DOWN)

Графиката илюстрира разпределението на клиентите според количеството свалени данни (INT_VOL_DOWN) за активните интернет потребители. На хистограмата вляво ясно се вижда силно дясно-скосено разпределение – преобладаващата част от клиентите генерират ниски обеми трафик, а само малък брой използват изключително големи количества данни. Червената линия показва средната стойност на свалания трафик, а синята – модата (най-често срещаната стойност). В случая всички тези централни тенденции са разположени близо до нулата, като средната стойност е по-вдясно спрямо медианата и модата, което потвърждава наличието на малка група интензивни потребители, „дърпащи“ средното нагоре.

В дясната част на фигурата е показано бинираното разпределение по категории потребление. Най-голям е делът на „Medium“ потребителите (над 67 хиляди), следвани от „High“ и „Low“. Групата на „Very High“ и „Zero user“ е относително малка, което допълнително подчертава концентрираността на трафика в ниския и средния диапазон. Около 12 хиляди клиенти са с етикет „no user“, тоест нямат активна интернет услуга или липсват данни за тях.

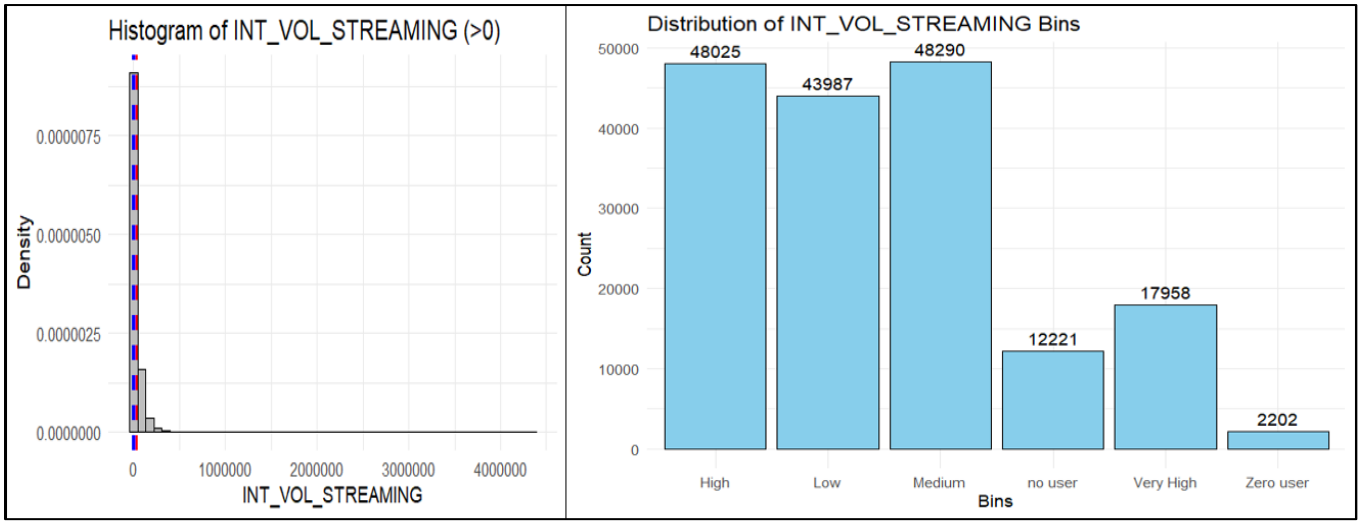


**Figure 9.** Разпределение на клиентите по категории на стрийминг интернет трафик (INT_VOL_STREAMING)

Графиката визуализира поведението на клиентите според обема на използвания стрийминг трафик (INT_VOL_STREAMING). Хистограмата вляво демонстрира ясно изразено дясно-скосено разпределение, при което най-голямата част от потребителите имат ниски стойности, а малка част — изключително високи обеми стрийминг данни. Червената линия обозначава средната стойност, лилавата — медианата, а синята — модата на разпределението. Трите индикатора са разположени много близо до нулата, което свидетелства, че болшинството клиенти гледат онлайн съдържание сравнително рядко или в малки обеми.

В дясната част е показано бинираното разпределение на абонатите по различни категории на стрийминг употреба. Най-големи са групите „Medium“, „High“ и „Low“, като всяка от тях обхваща между 43 000 и 48 000 потребители, което показва сравнително равномерно разпределение между тези три групи. Групата „Very High“ съдържа по-малък, но съществен брой интензивни стрийминг потребители (над 17 000). Категориите „Zero user“ и „no user“ са значително по-малки, с 2 202 и 12 221 клиенти съответно. Това разпределение подсказва, че стрийминг услугите са широко използвани, но само малка част от клиентите генерират наистина високи стойности, докато по-голямата част от потребителите остават в ниските и средните нива на употреба.

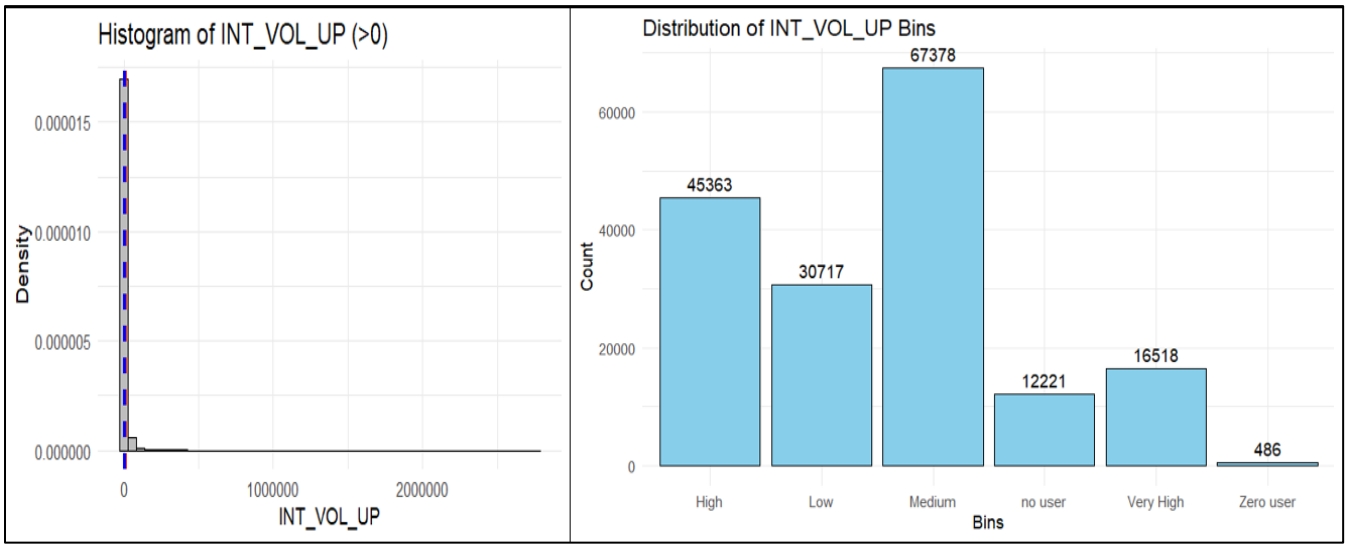


**Figure 10.** Разпределение на клиентите по категории на качен интернет трафик (INT_VOL_UP)

Графиката илюстрира разпределението на клиентите според обема на качените данни (upload трафик) — метриката INT_VOL_UP. Хистограмата вляво ясно показва дясно-скосено разпределение, където най-много клиенти качват сравнително малки обеми данни. Средната стойност (червена линия), медианата (лилава линия) и модата (синя линия) са концентрирани в най-ниската част на диапазона, което е типично за upload трафика, тъй като повечето клиенти ползват предимно download услуги.

В дясната графика са показани категориите („бинове“) по обем на качените данни. Най-голяма е групата „Medium“ с над 67 000 потребители, следвана от „High“ (45 363) и „Low“ (30 717). Групата „Very High“ обхваща над 16 500 интензивни upload потребители, а категориите „Zero user“ и „no user“ съдържат съответно 486 и 12 221 клиенти. Подобно на предходните usage метрики, тук също е видимо, че мнозинството потребители попадат в ниските и средните нива на активност, докато много малка част имат екстремни стойности. Това е логично предвид по-ограничените нужди на масовия клиент от качване на големи обеми данни.

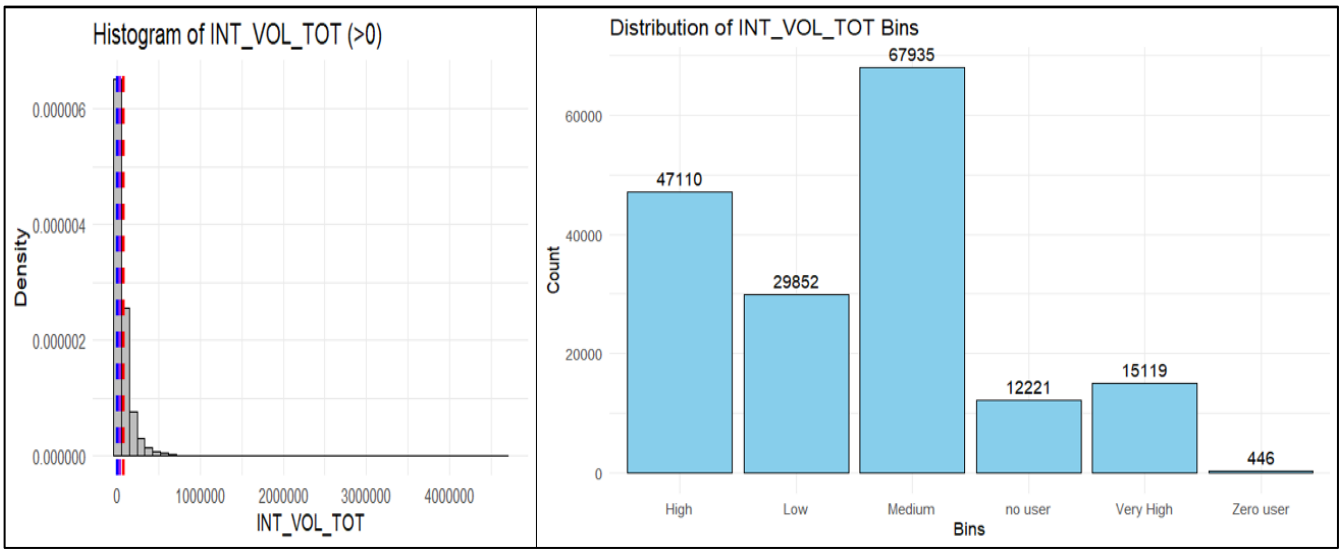


**Figure 11.** Разпределение на клиентите по категории на общ интернет трафик (INT_VOL_TOT)

Графиката илюстрира разпределението на клиентите според общия интернет трафик (download + upload) за измервания период. На хистограмата (вляво) ясно се вижда силно дясно-скосено разпределение — огромната част от клиентите имат ниски стойности на общия трафик, докато броят на клиентите с екстремно високи стойности е минимален. Медианата, средната стойност и модата са разположени в най-лявата част на разпределението, което е типичен профил за потребление на домашен интернет.

Разпределението по бинове (дясната графика) потвърждава този извод: най-голям е броят на клиентите в групата „Medium“ (67 935), следвани от „High“ (47 110) и „Low“ (29 852). Категорията „Very High“ включва 15 119 домакинства с интензивна интернет употреба. От друга страна, делът на клиентите с липсващо или нулево потребление („no user“ и „Zero user“) остава сравнително нисък. Този профил на разпределение показва ясно изразено мнозинство на средни и високи консуматори, докато клиентите с минимална активност са ограничено малко.

Проведеният емпиричен анализ на основните показатели за потребление на телекомуникационни услуги показва ясно изразена хетерогенност в профила на клиентите. Всички usage-разпределения се характеризират с рязко дясно-скосени дистрибуции — повечето клиенти попадат в групите с ниска до средна интензивност на използване, докато високите стойности са запазени за сравнително малък сегмент от клиентската база. Сегментацията по

бинове допълнително разкрива, че макар част от клиентите да не проявяват никаква активност („Zero user“ или „no user“), основната маса активно използва предлаганите услуги, като при интернет и мобилните услуги делът на високите консуматори е съществен.

Това разпределение е характерно за цифрови услуги на съвременни пазари — масовият потребител има умерено, но постоянно потребление, докато малък дял от домакинствата поддържа интензивна употреба, вероятно свързана с по-специфични нужди или дигитален начин на живот. Получените бинове по usage служат като основа за по-нататъшна клъстеризация и профилиране на клиентите по поведенчески типове.

### 4.3.2. Обобщение на разпределението по бинове на „Zero user“ и „no user“

В следващата таблица е представено резюме на разпределението на клиентите по бинове за всички ключови количествени usage-показатели след биниране. За всяка от биниранитe колони са изведени броят на ненулевите наблюдения, както и делът на „Zero user“ и „no user“ категории. Това дава добра представа за честотата на различните нива на потребление и валидността на всеки usage-показател в клъстерния анализ.

**Table 2.** Обобщение на "Zero user" и "no user" категориите по usage-характеристики

| Column | NaN_total | Non_null_total | Zero_user_count | No_user_count |
|---|---|---|---|---|
| MTEL_VOICE_CNT_BIN | 0 | 172683 | 3422 | 87572 |
| MTEL_TOT_VOL_MB_BIN | 0 | 172683 | 25744 | 87572 |
| TEL_OUT_MOU_BIN | 0 | 172683 | 8723 | 101846 |
| DTV_VIEW_NBR_BIN | 0 | 172683 | 79422 | 11146 |
| DTV_NR_VOD_1YR_BIN | 0 | 172683 | 71894 | 11146 |
| INT_VOL_DOWN_BIN | 0 | 172683 | 448 | 12221 |
| INT_VOL_UP_BIN | 0 | 172683 | 486 | 12221 |
| INT_VOL_STREAMING_BIN | 0 | 172683 | 2202 | 12221 |
| INT_VOL_TOT_BIN | 0 | 172683 | 446 | 12221 |

Прави впечатление, че броят на липсващи стойности (NaN_total) за всички usage-бинове е нулев, което показва успешна импутация и подготовка на данните. Всяка от променливите съдържа по 172 683 наблюдения, колкото е размерът на крайната извадка.

Колоната **Zero_user_count** показва броя на клиентите с регистрирана нулева употреба на конкретната услуга, докато **No_user_count** отразява клиентите, които изобщо нямат абонамент за съответната услуга. Например, при показателя MTEL_VOICE_CNT_BIN имаме 3 422 клиента с активна мобилна услуга, но без реализирано изходящо повикване, докато при MTEL_TOT_VOL_MB_BIN клиентите без мобилен интернет (no user) са 87 572, а тези с активна услуга, но нулева употреба – 25 744.

За фиксирана телефония (TEL_OUT_MOU_BIN) преобладава броят на клиентите без абонамент (101 846), докато за телевизията (DTV_VIEW_NBR_BIN и DTV_NR_VOD_1YR_BIN) делът на „Zero users“ е по-висок в сравнение с останалите услуги. При интернет услугите (INT_VOL_DOWN_BIN, INT_VOL_UP_BIN, NT_VOL_STREAMING_BIN, INT_VOL_TOT_BIN) почти всички клиенти с активен абонамент демонстрират някаква употреба, а напълно нулева активност се среща рядко.

Таблицата ясно показва, че по-голямата част от извадката е съсредоточена в групата „no user“ за всяка услуга – тоест голям дял от клиентите ползват ограничен брой услуги, а не всички едновременно. Същевременно сред активните потребители, броят на реално неупотребяващите („Zero user“) е значително по-нисък, особено при интернет услугите. Тези бинови разпределения

са изключително полезни за следващите етапи на анализ, като профилиране и клъстеризация, защото позволяват по-лесно разграничаване на клиентските сегменти според тяхната реална активност.

## 4.4. Клъстеризация с K-Modes

Важна част от анализа на клиентското поведение е групирането на клиентите според сходството на техните usage-профили. За тази цел е използван K-Modes алгоритъм, който е подходящ за работа с категорийни променливи и дава възможност за откроряване на характерни сегменти във всеки тип телекомуникационна услуга. Клъстеризацията е проведена поетапно – първо по отделни услуги, а след това чрез интегриране на получените профили в обобщени клъстерни групи.

### 4.4.1. Първична клъстеризация по услуги

В този етап на анализа е приложен K-Modes алгоритъм отделно за всяка основна телекомуникационна услуга, като за всеки клиент е определен клъстерен профил въз основа на характерните бинове за потребление. За мобилната телефония бяха изведени 8 клъстера, които отразяват различни комбинации между честота на гласови разговори и обем на използван мобилен интернет. Например, най-големият клъстер (87 572 клиенти) е групата на неактивните потребители („no user“), докато други по-малки клъстери са съсредоточени около ниски и средни нива на използване („Low voice, Medium data“), както и при „High“ и „Very High“ потребители.

При фиксираната телефония K-Modes алгоритъмът идентифицира 6 групи. В най-големия клъстер попадат отново неактивни потребители („no user“), докато останалите групи се отличават с различна степен на използване на outgoing minutes, включително отделен клъстер за „Zero user“ и специфични сегменти за „Low“, „Medium“, „High“ и „Very High“ потребители.

За интернет услугата също бяха получени 8 различни профила, които отразяват не само обема на свалените и качените данни, но и гледането на стрийминг съдържание. Тук се откроряват клъстери с „Low“ или „Medium“ активност, но и значителни групи с висока интензивност на използване на интернет ресурси („Very High“). Отново има група от неактивни клиенти, макар и с по-малък дял спрямо другите услуги.

Клъстеризацията при дигиталната телевизия разкри 6 основни групи според честотата на гледане. Най-голям е делът на „Zero user“ и „no user“ сегментите, а останалите клъстери са разпределени между „Low“, „Medium“, „High“ и „Very High“ потребители, като „High TV viewership“ ясно се отличава с повишена ангажираност.

Това многоетапно групиране по услуги позволява по-прецизно профилиране и разбиране на различните клиентски поведения, а така създадените клъстери формират база за последващ интегриран анализ и таргетиране.

### 4.4.2. Финална клъстеризация

След провеждане на първичната клъстеризация по отделни услуги, е реализирана интегрирана клъстеризация чрез K-Modes върху обобщените профили за всяка услуга. Така всеки клиент е класифициран във финален клъстер, който съчетава най-характерните му модели на потребление на мобилни, фиксирани, интернет и телевизионни услуги.

В резултат на анализа са изведени 11 интегрирани клъстера, като за всеки от тях е характерна различна комбинация от потребителски профили. Например, най-голям дял от клиентите попада в групата „no user“, характеризираща се с липса на активност във всички

услуги. В други клъстери доминират профили с висока интернет активност („high usage“, „very high streaming“), средно потребление на телевизия („medium view“) или изразени групи с ниско или средно ниво на използване.

Тази класификация позволява ясно разграничаване между различните сегменти клиенти според реалното им поведение. В таблицата по-долу са представени наименованията на профилите за всяка услуга по клъстер, което подпомага интерпретацията на резултатите и последващите маркетингови действия.

**Table 3.** Описание на финалните клъстери – модални профили по услуги

| FINAL_ CLUSTER | Cluster_MTEL _NAME | Cluster_INT _NAME | Cluster_DTV _NAME | Cluster_TEL _NAME | Описание |
|---|---|---|---|---|---|
| 1 | no user | low usage | medium view | zero user | Клиенти без мобилна или фиксирана телефония, с ниско интернет потребление и умерено гледане на телевизия; най-често по-възрастни домакинства, използващи основно телевизия и интернет за базови нужди. |
| 2 | no user | medium usage, low streaming | zero user | zero user | Домакинства с умерена интернет употреба, ограничено видео съдържание (ниско стрийминг), без фиксирана или мобилна телефония, и без гледане на телевизия; типично консервативни клиенти. |
| 3 | no user | medium usage | medium view | zero user | Клиенти със средна интернет консумация и умерено гледане на телевизия, без други активни услуги; балансиран дигитален профил, подходящ за базови онлайн и ТВ услуги. |
| 4 | no user | high usage | zero user | zero user | Активни интернет потребители, които не използват телевизия, фиксирана или мобилна телефония; ориентирани към онлайн услуги, често семейства с по-голяма дигитална ангажираност. |
| 5 | no user | high usage, very high streaming | high view | zero user | Сегмент с високо интернет потребление, много интензивно използване на видео стрийминг и телевизия, без участие във фиксирана и мобилна телефония; дигитално активни домакинства с фокус върху забавленията. |
| 6 | no user | high usage, medium streaming | medium view | zero user | Домакинства с висока интернет консумация и средно гледане на телевизия и стрийминг, без гласови услуги; профил на съвременен, мултискринингов потребител. |
| 7 | no user | medium usage | zero user | zero user | Клиенти със средно интернет потребление, които не гледат телевизия и не използват гласови услуги; най-често самостоятелни, по-млади дигитални потребители. |
| 8 | no user | medium usage, low streaming | zero user | zero user | Абонати със средна интернет активност, ограничено видео потребление, без други услуги; дигитални клиенти с фокус върху базови онлайн дейности. |
| 9 | no user | very high usage | medium view | zero user | Много активни интернет потребители със средно гледане на телевизия; дигитално напреднали домакинства, вероятно с няколко членове, интензивно използващи интернет за работа, забавление и обучение. |
| 10 | no user | high usage, very high upload | zero user | zero user | Потребители с много високо интернет използване, включително качване на големи обеми данни, без гласови и ТВ услуги; подходящи за сегменти с фрийлансъри или дистанционна работа. |
| 11 | no user | high usage | high view | zero user | Клиенти с висока интернет активност и интензивно гледане на телевизия, без използване на гласови услуги; домакинства с комплексна дигитална ангажираност и разнообразни онлайн нужди. |

Структурата на финалните клъстери предоставя база за по-детайлно профилиране на клиентите, както и за разработване на целеви маркетингови стратегии към отделните сегменти в

зависимост от модела на потребление на различните услуги. Така интегрираната клъстеризация позволява едновременно отчитане на комбинираното поведение на потребителите в мултисервисна среда.

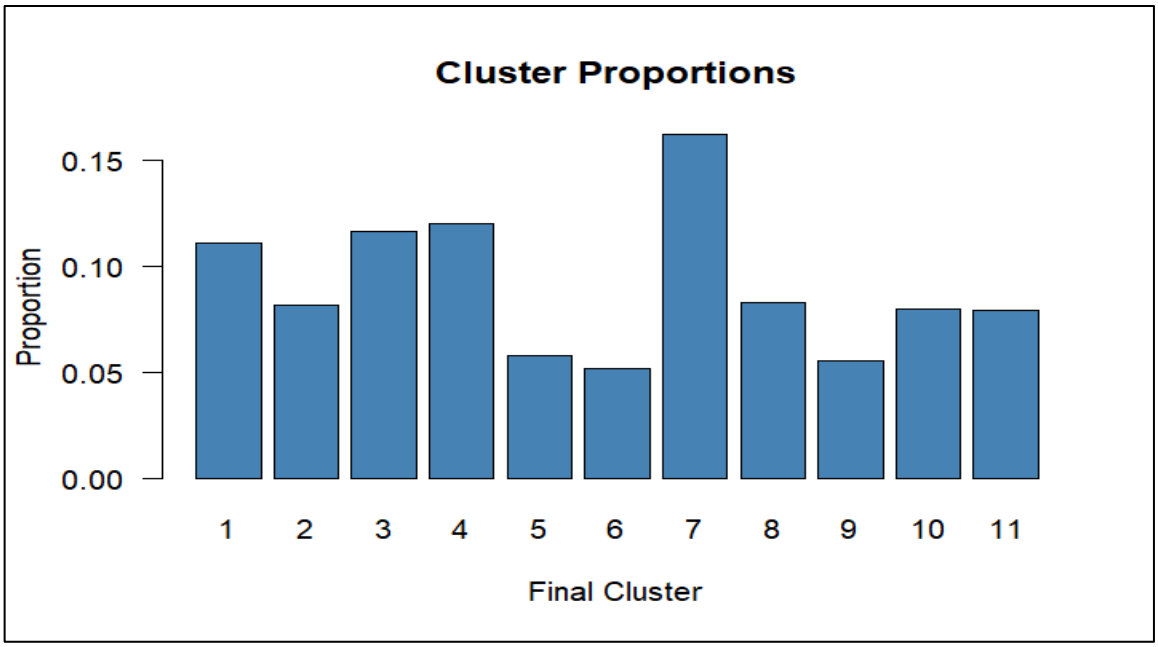


**Figure 12.** Разпределение на клиентите по финални клъстери (Cluster Proportions)

На фигурата е представено разпределението на клиентите по изведените финални клъстери. Вижда се, че няма силно доминираща група – делът на всеки клъстер варира между 5% и 16% от извадката, което показва балансирано сегментиране на клиентската база. Най-голям относителен дял има клъстер 7 (над 15%), докато клъстерите 5 и 6 са с най-малък брой потребители. Това разпределение подсказва наличие както на големи, хомогенни групи клиенти с близък профил, така и на по-малки, специфични сегменти, които могат да бъдат обект на таргетирани действия.

## 4.5. Профилиране на клъстерите по демография и приходи

В таблицата са представени процентните дялове на езиковите групи (Dutch/French) и пола (Female/Male) по всеки от финалните 11 клъстера, изведени след интегрираната клъстеризация. Прави впечатление, че във всички клъстери преобладават клиенти, чийто основен език е Dutch (над 94% във всеки клъстер), докато делът на френскоговорящите варира между 2.9% и 5.4%. Това отразява общата структура на анализираната клиентска база и показва, че езикът не е основен диференциращ фактор между клъстерите.

**Table 4.** Демографски профили на финалните клъстери по език и пол

| FINAL_CLUSTER | Dutch (%) | French (%) | Female (%) | Male (%) |
|---|---|---|---|---|
| 1 | 96.81 | 3.19 | 33.6 | 66.4 |
| 2 | 96.39 | 3.61 | 30.3 | 69.7 |
| 3 | 97.08 | 2.92 | 25.5 | 74.5 |
| 4 | 95.73 | 4.27 | 26.7 | 73.3 |
| 5 | 95.41 | 4.59 | 28.8 | 71.2 |
| 6 | 96.39 | 3.61 | 27.2 | 72.8 |
| 7 | 96.62 | 3.38 | 24.7 | 75.3 |

| | | | | |
|---|---|---|---|---|
| 8 | 96.1 | 3.9 | 30.1 | 69.9 |
| 9 | 94.98 | 5.02 | 26.2 | 73.8 |
| 10 | 94.64 | 5.36 | 26.9 | 73.1 |
| 11 | 96.05 | 3.95 | 26.7 | 73.3 |

По отношение на пола се забелязва известна хомогенност между групите, като навсякъде делът на мъжете е по-висок – между 66% и 75%. Клъстер 7 се откроява с най-висок дял на мъже (75%), докато най-балансирано е разпределението в клъстер 1 (34% жени и 66% мъже). Това предполага, че основните сегменти, получени от клъстерния анализ, са относително равномерно разпределени по език и пол, което позволява изводите от анализа да бъдат прилагани върху широк кръг клиенти без необходимост от значително допълнително сегментиране по демографски признак.

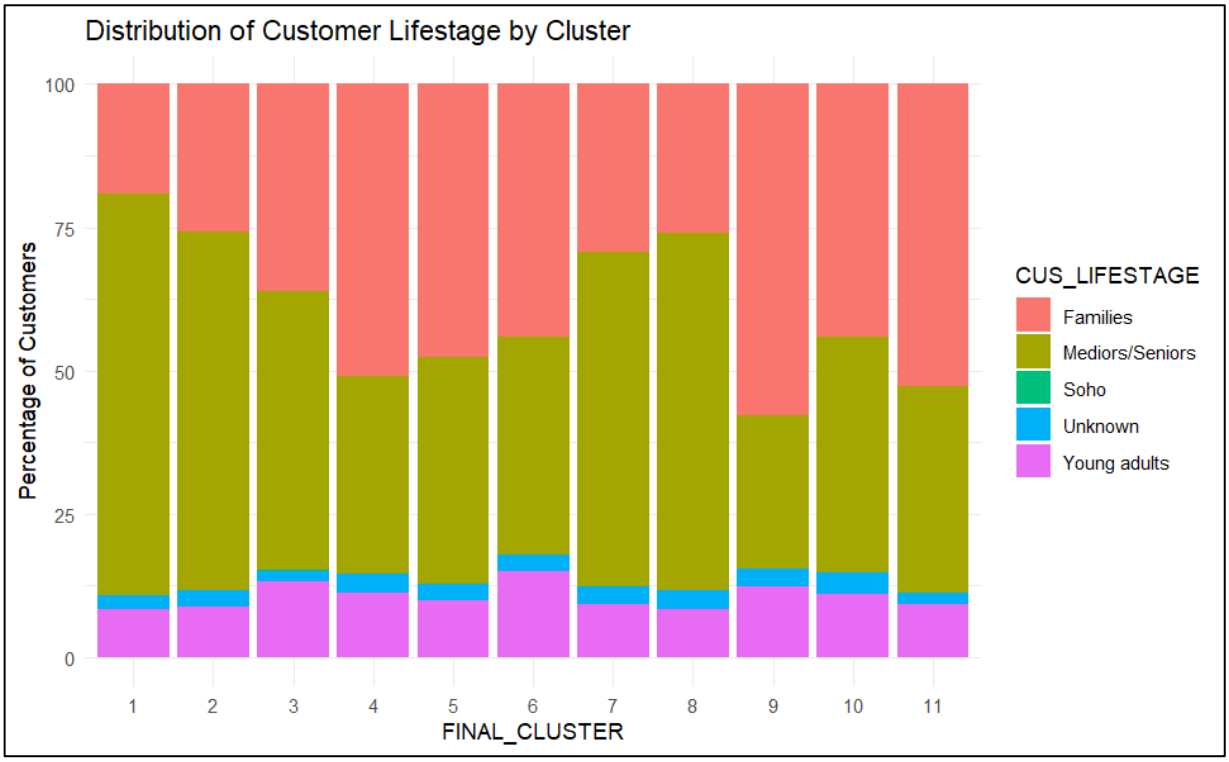


**Figure 13.** Дял на клиентите по жизнен етап (lifestage) в различните финални клъстери

Графиката представя процентното разпределение на клиентите по основни групи („Families“, „Mediors/Seniors“, „Soho“, „Young adults“, „Unknown“) във всеки от 11-те финални клъстера. Прави впечатление, че две основни групи („Families“ и „Mediors/Seniors“) доминират във всички клъстери, но техният относителен дял варира значително между отделните сегменти:

**Клъстер 1 и 2** са с преобладаващо високо присъствие на „Mediors/Seniors“ (над 70% и 62%), докато делът на „Families“ е по-нисък. Това може да се обясни с по-голямата склонност на по-възрастните клиенти към по-ниска или традиционна употреба на телекомуникационни услуги.

**Клъстер 4, 5, 6, 10 и 11** показват доминация на „Families“, което подсказва, че в тези сегменти се концентрират по-активни или разнообразни потребители, вероятно с повече членове в домакинството и по-висока необходимост от дигитални услуги.

**Групата „Young adults“** заема между 8% и 13% във всички клъстери, като няма клъстер с изключително висока концентрация на тази група.

**„Soho“** (малки бизнеси и домашни офиси) и „Unknown“ са представени с минимални дялове, което потвърждава фокуса на анализа върху крайни клиенти.

В някои клъстери има видим баланс между „Families“ и „Mediors/Seniors“, докато в други едната група рязко доминира. Това подчертава потенциала за по-таргетирани маркетингови кампании според лейфстейдж структурата на всяка група.

Лейфстейдж профилът по клъстери е ключов за по-дълбоко разбиране на клиентските сегменти. Той показва дали даден клъстер събира по-консервативни и възрастни потребители, или такива с по-модерен и семеен профил. Това позволява по-добро таргетиране на комуникационните и продуктовите предложения според реалната структура на потребителите в отделните сегменти.

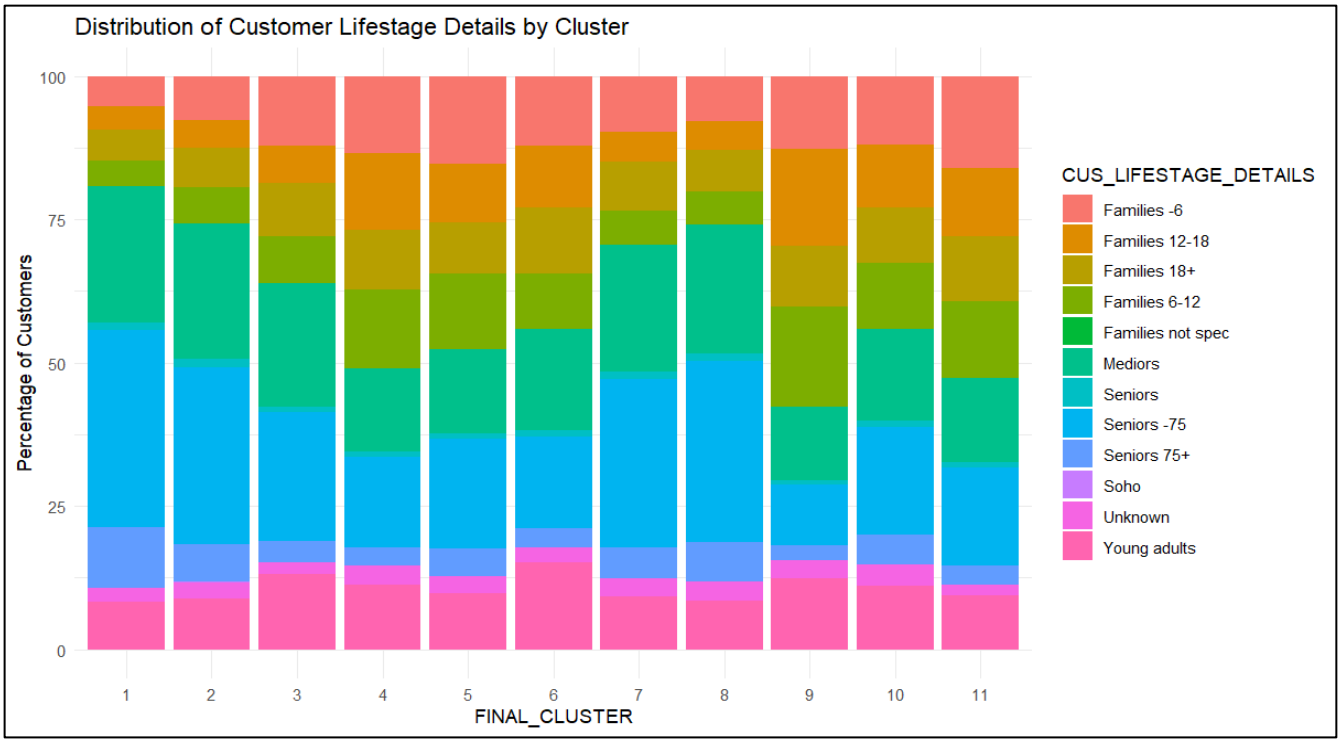


**Figure 14.** Разпределение на клиентите по подкатегории жизнен етап (lifestage details)

Графиката илюстрира процентното разпределение на клиентите по детайлни демографски подкатегории (lifestage details) във всеки от финалните клъстери, получени от интегрираната клъстеризация. Всяка цветна лента представя отделна подгрупа – например: „Families -6“ (семейства с деца под 6 г.), „Families 12-18“, „Medior“, „Seniors 75+“, „Young adults“, „Unknown“ и др. Ясно се вижда разнообразието във всеки клъстер: При някои клъстери преобладават по-млади домакинства и млади хора (напр. „Families 18+“, „Young adults“), докато при други е по-голям делът на пенсионерите („Seniors“, „Seniors 75+“) и т.н.

**Клъстери 1 и 2** имат видимо по-висок дял на по-възрастните клиенти (Medior/Seniors), докато някои от по-високите номера клъстери съдържат повече млади домакинства.

Подкатегориите **„Unknown“ и „Soho“** (вероятно малки бизнеси или случаи с неустановен профил) съставляват минимален дял във всички групи, което говори за добра покриваемост и точност на демографската информация.

**„Families -6“, „Families 6-12“, „Families 12-18“** и **„Families 18+“** позволяват много фино разбиване на семейните профили и проследяване на динамиката между клъстерите.

Тази графика дава основа за прецизно таргетиране на кампании и за определяне на най-релевантните продукти и услуги според етапа на живот на клиентите във всеки клъстер. Така например, оферти за детски услуги ще са по-релевантни за клъстери с по-висок дял на „Families -

6“ и „Families 6-12“, докато продукти за по-възрастни ще намерят отклик при групи, доминирани от „Seniors 75+“ и „Medior“

## 4.6. Анализ на удовлетвореността (NPS/анкета)

В този етап от анализа е извършено профилиране на клиентската удовлетвореност на базата на отговорите от анкетата (NPS и допълнителни въпроси). Всеки въпрос е насочен към оценка на различни аспекти на потребителското преживяване и лоялността към доставчика на телекомуникационни услуги. Отговорите са обобщени по финалните клъстери, което позволява да се идентифицират специфични сегменти с висока или ниска степен на удовлетвореност, както и потенциални рискови групи.

Въпросите от анкетата обхващат теми като препоръка към други потребители, удовлетвореност от конкретни услуги (интернет, телевизия, мобилни и фиксирани телефони), сравнителна оценка спрямо други доставчици, намерения за отказ, както и обща оценка на съотношението цена/качество.

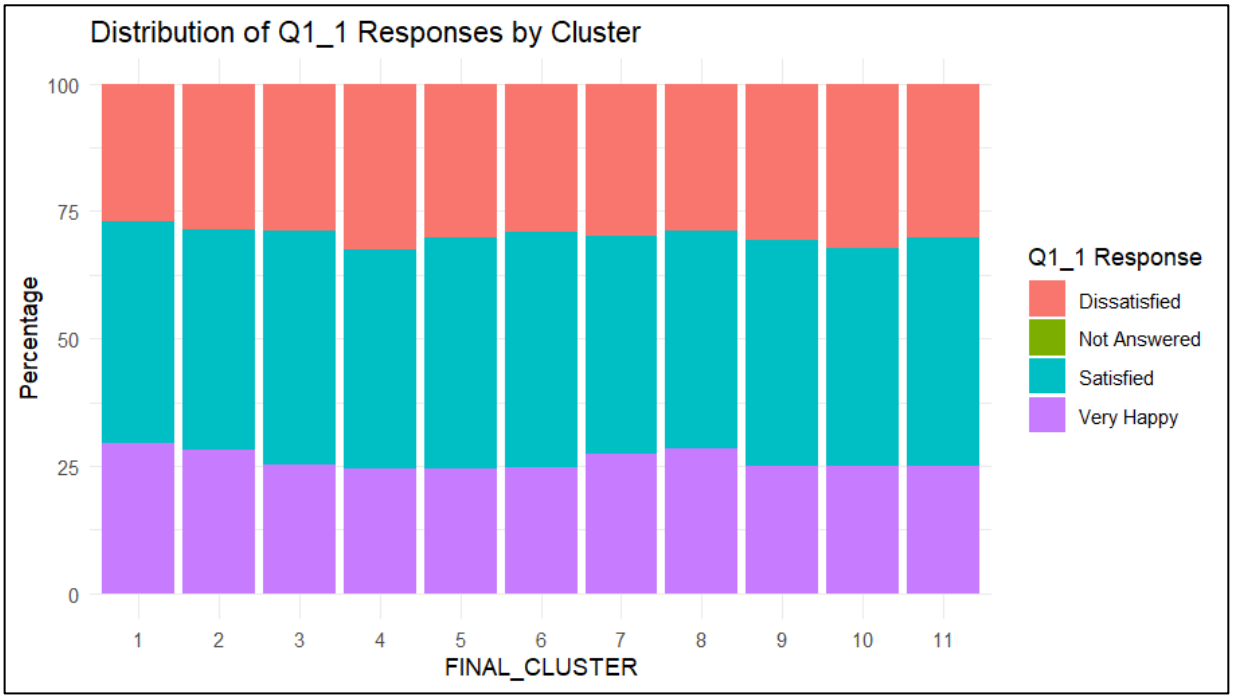


**Figure 15.** Разпределение на отговорите на въпрос Q1_1

Фигура 15 представя процентното разпределение на клиентските отговори по финален клъстер за въпрос Q1_1 ***(„Бихте ли препоръчали XYZ на семейството и/или приятели, ако ви попитат за вашето мнение?“***), който измерва т.нар. Net Promoter Score (NPS). Визуализацията е реализирана като стълбчета, където различните цветове съответстват на отделните категории от Лайкърт-скалата: червеното индикира „Недоволни“ („Dissatisfied“), жълтото – „Доволни“ („Satisfied“), а зеленото – „Много доволни“ („Very Happy“).

От графиката се наблюдава, че разпределението на отговорите е относително балансирано между клъстерите – делът на „Много доволни“ варира между 25% и 35%, а „Доволни“ формират най-голямата част във всички групи. Делът на „Недоволни“ също е значим, но не доминира в нито един клъстер. Не се откроява клъстер с отчетливо негативен профил, което сочи, че склонността към препоръка на оператора е сравнително равномерна във всички клиентски сегменти, без рязко изразени екстремни стойности. Това свидетелства за липса на сериозни негативни групи и относителна хомогенност в клиентското преживяване по отношение на NPS.

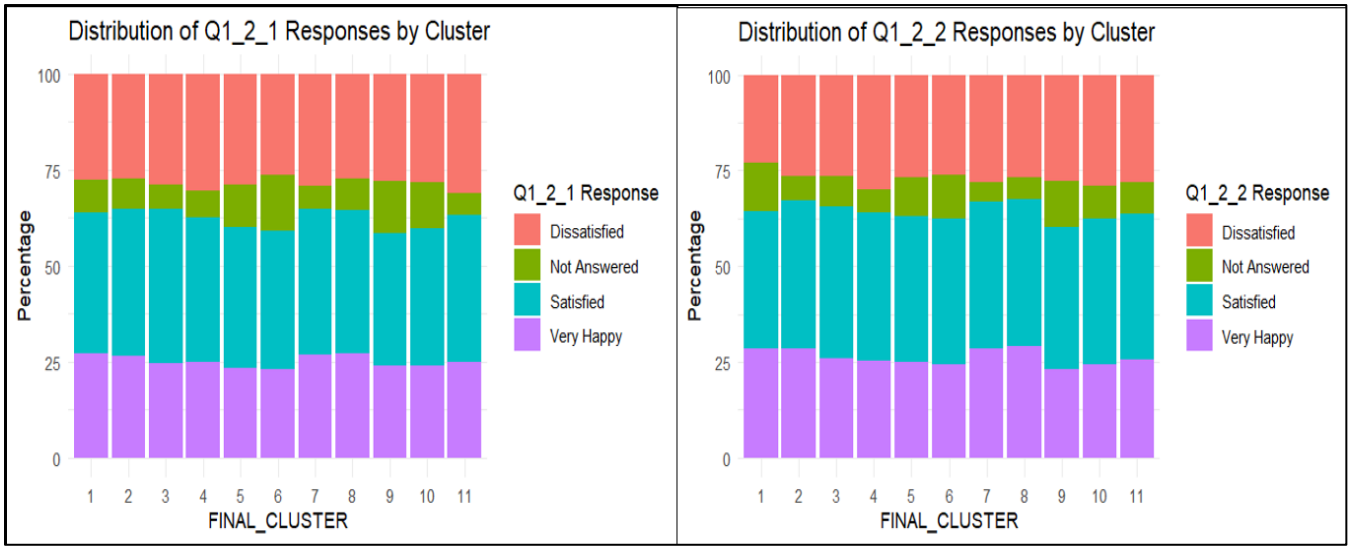


**Figure 16.** Разпределение на отговорите на въпроси Q1_2_1 и Q1_2_2

Лявата част на фигурата представя разпределението на отговорите по клъстери за въпрос Q1_2_1: *„**Бихте ли препоръчали интернет услугата на XYZ на семейството и/или приятели, ако ви попитат за вашето мнение**?“* Дясната част показва резултатите за въпрос Q1_2_2: *„**Бихте ли препоръчали дигиталната телевизия на XYZ на семейството и/или приятели, ако ви попитат за вашето мнение**?“*

Видимо, при повечето клъстери разпределението е близко до това при общия въпрос за препоръка – най-голям е делът на доволните („Satisfied“), следван от „Very Happy“. Отделни клъстери (например 7, 8, 9) се отличават с леко по-висок процент много доволни спрямо други, но липсва клъстер със сериозно негативен профил. „Not Answered“ също запазва приблизително еднакво присъствие, без рязко да доминира в отделни групи.

Особено при интернет услугата (Q1_2_1) има малки вариации между клъстерите, като някои групи с по-висока ангажираност към интернет (примерно клъстери с висок дигитален профил) демонстрират и малко по-висока склонност към препоръка. Аналогично, при дигиталната телевизия (Q1_2_2), най-доволните клиенти са равномерно разпределени, без отчетлива доминация на конкретен сегмент.

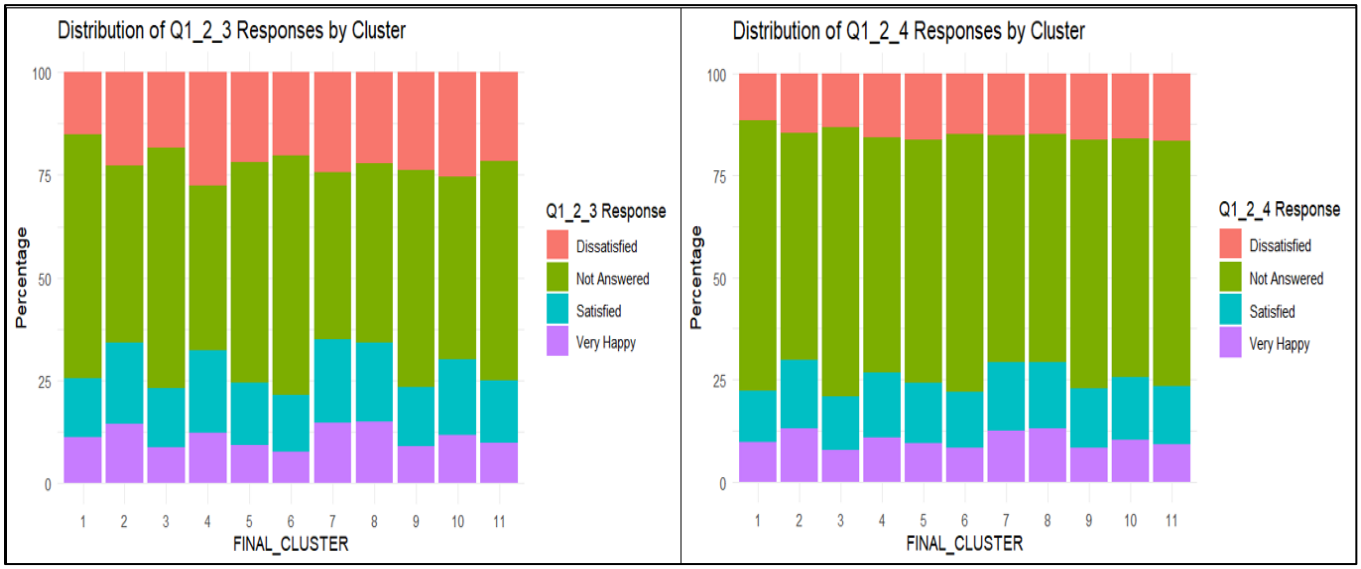


**Figure 17.** Разпределение на отговорите на въпроси Q1_2_3 и Q1_2_4

Лявата част на фигурата показва разпределението на отговорите по финални клъстери за въпрос Q1_2_3: „***Бихте ли препоръчали услугата за фиксиран телефон на XYZ на семейството и/или приятели, ако ви попитат за вашето мнение***?“, а дясната – за въпрос Q1_2_4: „***Бихте ли препоръчали мобилната услуга на XYZ на семейството и/или приятели, ако ви попитат за вашето мнение?***“.

И за двата въпроса най-голям дял във всички клъстери имат „Not Answered “, което показва, че значителна част от клиентите не са ползватели или не са ангажирани с тези услуги. Сред отговорилите, делът на „Dissatisfied“ (недоволни) при фиксираната телефония (Q1_2_3) е видимо по-висок в сравнение с останалите услуги, като почти във всички клъстери недоволните преобладават над доволните. За мобилната услуга (Q1_2_4) ситуацията е сходна: „Not Answered“ е доминираща категория, а сред отговорилите няма клъстер с отчетливо висок дял на много доволни клиенти („Very Happy“).

Тези разпределения подчертават слабата ангажираност към фиксираните и мобилни телефонни услуги в повечето клъстери, както и относително ниските нива на удовлетвореност сред ползващите. Въпреки това няма изолирани клъстери с изключително позитивен или негативен профил, като различията между сегментите остават умерени.

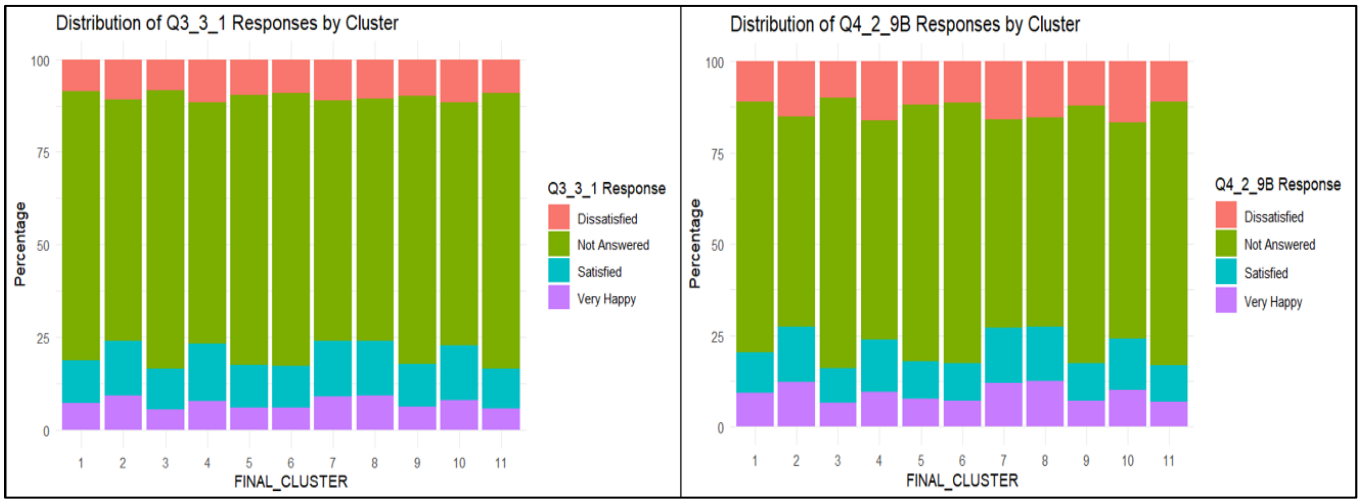


**Figure 18.** Разпределение на отговорите на въпроси Q3_3_1 и Q4_2_9B

Лявата част на фигурата показва разпределението по финални клъстери за въпрос Q3_3_1: „***Бихте ли, въз основа на фактурата си, препоръчали XYZ на семейството и/или приятели, ако ви попитат за вашето мнение?***“, а дясната – за въпрос Q4_2_9B: „***Имате ли усещането, че при XYZ можете да подадете жалба?***“.

И при двата въпроса доминира високият дял на „Not Answered“ във всички клъстери, което показва ниска ангажираност на клиентите или неотносима тематика за голяма част от базата. При останалите отговори се запазва сходна структура – делът на доволните и много доволните е значително по-нисък спрямо неотговорилите, а недоволните формират относително малка част от всички наблюдавани групи. Липсват клъстери с ясно изразен превес на негативни или позитивни мнения, като разпределението е равномерно в цялата клиентска база.

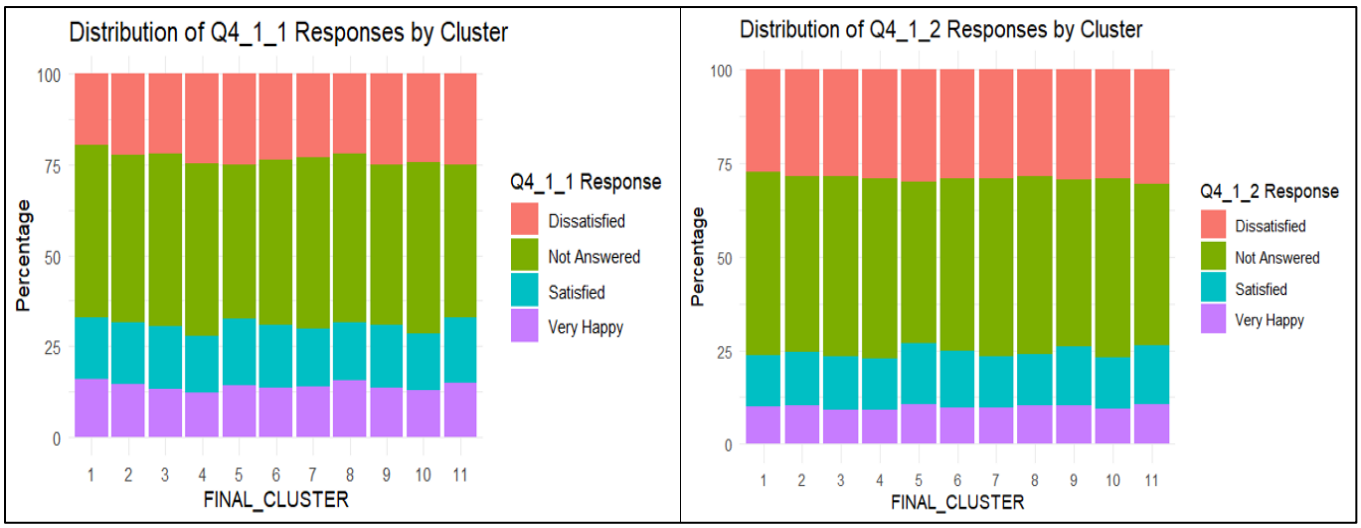


**Figure 19.** Разпределение на отговорите на въпроси Q4_1_1 и Q4_1_2

Лявата част на фигурата представя разпределението по финални клъстери за въпрос Q4_1_1: „***Ако днес трябва да изберете доставчик на телефон, интернет и телевизия, бихте ли избрали отново XYZ***?“ Дясната част показва резултатите за въпрос Q4_1_2: „***Смятате ли, че XYZ е по-добър или по-лош от другите доставчици на телефон, интернет и телевизия?***“

И при двата въпроса се наблюдава значителен дял на „Not Answered“ сред всички клъстери, което може да показва или неутралност, или ниска ангажираност по темата. Най-голямата част от отговорилите са в групата на „Satisfied“, следвани от „Very Happy“, докато делът на „Dissatisfied“ е значителен, но не доминира в нито един клъстер. В някои клъстери се забелязват леки вариации, но няма клъстер с превес на негативни или позитивни мнения.

Тази структура потвърждава, че отношението към избора на XYZ и сравнението с конкуренцията е относително равномерно разпределено сред клиентската база, без отчетливи клъстерни различия.

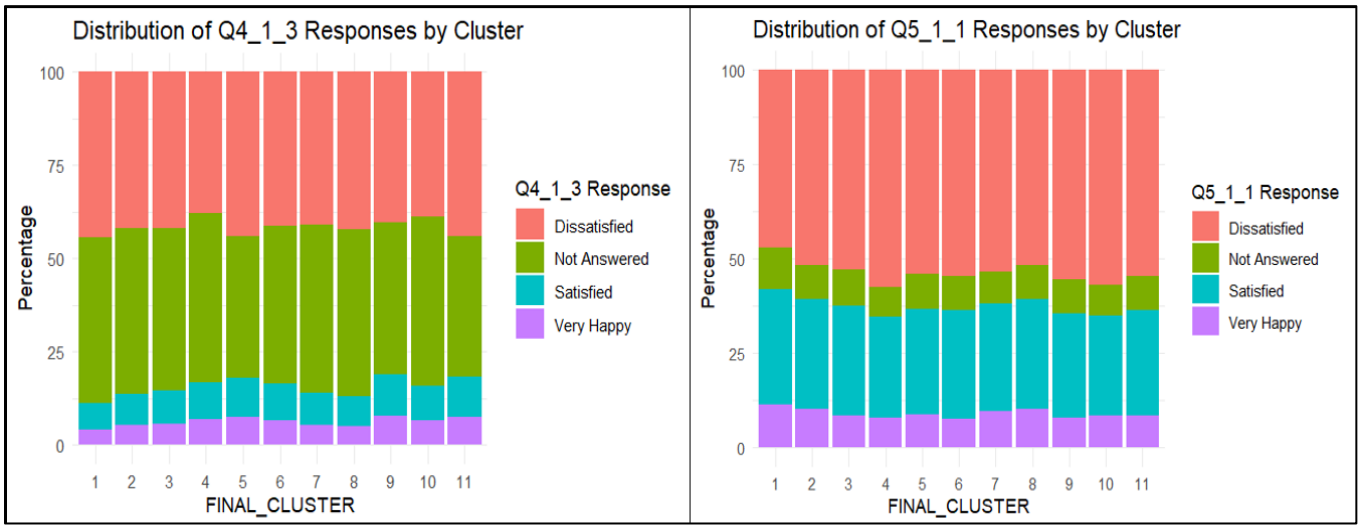


**Figure 20.** Разпределение на отговорите на въпроси Q4_1_3 и Q5_1_1

Лявата част на фигурата представя разпределението по финални клъстери за въпрос Q4_1_3: „***Обмисляте ли да прекратите някои продукти/услуги на XYZ?***“ Дясната част показва резултатите за въпрос Q5_1_1: „***До колко сте доволни от съотношението цена/качество на вашите продукти от XYZ?“***

Анализа на разпределението на отговорите по въпросите Q4_1_3 и Q5_1_1 недвусмислено показва, че във всички клиентски клъстери преобладава сериозно недоволство. На въпроса „Обмисляте ли да прекратите някои продукти/услуги на XYZ?“ най-големият дял във всяка група заема именно категорията „Dissatisfied“ (Недоволни), а значителна част от клиентите изобщо не отговарят на този въпрос. Това може да се тълкува като ниска ангажираност или липса на доверие към компанията, което само по себе си е негативен сигнал.

По отношение на въпроса за удовлетвореността от съотношението цена/качество на услугите („До колко сте доволни от съотношението цена/качество на вашите продукти от XYZ?“), картината е сходна – и тук най-голям процент от клиентите във всеки клъстер изразяват неудовлетвореност. Макар втората по големина група да са „Satisfied“, делът на доволните и особено на „Very Happy“ остава осезаемо по-нисък в сравнение с недоволните. Това ясно показва, че значителна част от клиентите изпитват разочарование от предлаганите продукти и услуги, което носи реален риск от загуба на база и затруднява изграждането на дългосрочни взаимоотношения с клиентите.

В обобщение, данните от анкетите категорично сочат, че проблемът с клиентската удовлетвореност е съществен и масов, като преобладаващата част от абонатите се чувстват неудовлетворени или не намират смисъл да дадат положителна обратна връзка. Това е сигнал за необходимост от сериозен анализ и спешни мерки за подобряване на качеството на услугите, комуникацията и ценовата политика.

## 5. Ограничения при провеждане на анализа

Въпреки сериозните усилия за подготовка и чистене на данните, анализът се сблъска с редица практически ограничения, които не могат да се игнорират. На първо място, работата с телеком данни, идващи от различни източници и системи, неминуемо води до хетерогенност – имахме и структурирани usage лога, и анкети, и демографски профили. Още в началото стана ясно, че липсващите стойности са хроничен проблем. Например, usage флаговете като CUS_FL_DTV, CUS_FL_INT и другите основни услуги имаха доста липси, което в абсолютна стойност означава над 10 000 реда със съмнителна информативност. При анкетните въпроси и маркетинговите променливи липсите бяха дори над 60 -70%, което на практика ги правеше почти неизползваеми за масови анализи без сериозен риск от изкривяване.

Второто е работата с дублирани редове – всеки клиент фигурираше с няколко записа за различни месеци. В един момент трябваше да вземем решение кой запис да използваме. Заложихме на най-новия запис за всеки клиент, за да има една snapshot картина на поведението. Това обаче означава, че губим информация за динамиката във времето и промените в потреблението, което може да е критично за някои типове анализи.

Трети важен аспект са така наречените outliers – имаше клиенти с абсурдно високи стойности на употреба, например над 4000 единици за някои usage индикатори, докато медианата е десетки пъти по-ниска. е наложително да се дефинират bins за отделните usage променливи, като най-горният bin по същество събираше всичко, което се отличава рязко от масовия случай. Така запазихме логиката на сегментацията, но все пак тези екстремни стойности влияят на статистическите разпределения.

Като заключение анализът даде стабилна и логична сегментация, но на цената на множество компромиси и внимателно подбрани подходи към липсващи стойности, аутлайъри и дублирани записи. Винаги, когато има голям, хетерогенен dataset, това е цената на реалния data science – между чистата теория и практическата „кал“.

## 6. Маркетингов фокус

Анализът на реалното използване на телекомуникационните услуги показа ясно изразена сегментация на клиентската база според поведението – от напълно пасивни до силно ангажирани потребители. Най-отчетливо се открояват три основни профила: **„no user/zero user“**, **„medium“** и „high usage“ клиенти, като тези групи се повтарят във всяка от основните услуги (мобилна телефония, интернет, телевизия и фиксирана телефония).

Въз основа на това препоръчваме силен маркетингов фокус върху персонализираното таргетиране според реалното потребление, а не универсален подход за всички. Данните от анализа показват, че:

– **Мултипродуктовите клиенти** (домакинства с млади семейства и деца) носят най-висок приход и имат най-висока склонност към up-sell и cross-sell, особено когато вече комбинират мобилен интернет и телевизия;– **„Zero user“** сегментите са подходящи за целенасочени реактивационни кампании и стимули за старт на ползване, тъй като голяма част от тях изобщо не използват наличните услуги;

– **Сегментите с ниска удовлетвореност по NPS или с ниско потребление** имат по-висока склонност към отлив и изискват специфични retention подходи;

– **Демографските анализи** (CUS_LIFESTAGE, CUS_SEX, CUS_VALUE) потвърждават, че домакинствата с млади семейства са най-печеливши, докато възрастните, живеещи сами, са по-склонни към минимален пакет.

Фокусът трябва да бъде върху **поведенческото таргетиране** – персонализирани кампании към най-ценните клиенти, стимули за повторно въвличане на пасивните, както и запазване на лоялните чрез добавена стойност според реалните им нужди и потенциал. Така се използват максимално откритите от сегментацията възможности и се повишава ефективността на маркетинга.

## Заключение

Извършеният анализ демонстрира потенциала на поведенческата сегментация въз основа на реални usage-данни в телекомуникационния сектор. Чрез интегрирана методология, включваща почистване, биниране и клъстеризация на категорийни и количествени променливи, успяхме да идентифицираме отчетливи групи клиенти с различни дигитални профили и потребителско поведение. Получените 11 клъстера предоставят надеждна основа за таргетирани маркетингови стратегии и персонализирани бизнес решения – от предлагане на подходящи услуги и оферти до оптимизация на комуникацията и повишаване на удовлетвореността.

Резултатите показват ясно изразена хетерогенност в нуждите и навиците на клиентите – от умерени потребители с фокус върху традиционните услуги до дигитално активни домакинства с интензивна консумация на интернет и стрийминг. Анализът на демографските, приходните и анкетните показатели по клъстери потвърждава, че профилирането по usage е по-информативно от класическите демографски сегментации и позволява по-прецизно разбиране на пазарната база.

Методологичният подход – следващ добрите практики от научната литература и съобразен с реалните ограничения на данните – доказва, че комбинирането на бизнес логика с аналитични техники води до практически приложими резултати. Ограниченията, свързани с липсващи стойности и специфики на данните, бяха отчетени и минимизирани чрез целенасочени preprocessing стъпки и стриктен контрол на качеството.

В заключение, настоящият проект не само валидира използването на k-modes клъстеризация за usage-базирана сегментация, но и предоставя практически инструменти за вземане на информирани решения, които могат да бъдат адаптирани и към други индустрии с подобен профил на данните. Надграждането на този подход чрез допълнителни анализи (например предиктивно моделиране или автоматизация на таргетирането) би могло да осигури още по-голяма стойност за бизнеса и крайните клиенти.

## Източници:


**[1]** Пелова, Б. (2024). Презентации по Предварителна обработка на големи масиви от данни в бизнеса. Зимен семестър, учебна година 2024/2025. Available at: elearn.uni-sofia.bg (Accessed: 19 June 2025).

**[2]** Пелова, Б. (2024) Наука за данните в бизнеса и финансите: Техники за моделиране на многомерни масиви от данни. Лекционен курс, Софийски университет „Св. Климент Охридски“: elearn.uni-sofia.bg (Accessed: 19 June 2025).

**[3]** GemSeek (2017). *Case Study: Data Prep and Cluster Analysis*. MSc Program in Business Analytics, FEBA, Sofia University, предоставено по партньорски договор № 523/6 -Nov-2017.

**[4]** Huang, Z. (1998). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore: World Scientific.

**[5]** Jolliffe, I.T. & Cadima, J. (2016) *Principal Component Analysis: A Review and Recent Developments*. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), Chapter 1. elearn.uni-sofia.bg (Accessed: 19 June 2025).

**[6]** Bogdanova, B., Marchev Jr., A., Zakov, V., Stefanov, D., & Genov, K. (2022). Prediction of the Air Pollution by Geo-locations in Sofia. *IFAC PapersOnLine*, 55(11), 190–195. Available at: https://doi.org/10.1016/j.ifacol.2022.08.071 (Accessed: 19 June 2025).

**[7]** Bogdanova, B., & Stancheva-Todorova, E. (2021). ML-based predictive modelling of stock market returns. *AIP Conference Proceedings*, 2333, 150006. Available at: https://doi.org/10.1063/5.0042805 (Accessed: 19 June 2025).

**[8]** Abdul-Rahman, S., Kamal Arifin, N. F., Hanafiah, M. & Mutalib, S. (2021) ‘Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier’, International Journal of Advanced Computer Science and Applications, 12(9), pp. 434–444. doi: 10.14569/IJACSA.2021.0120950. Available at: https://thesai.org/Downloads/Volume12No9/Paper_50-Customer_Segmentation_and_Profiling_for_Life_Insurance.pdf (Accessed: 19 June 2025).

**[9]**