Raport 1

Eksploracja danych

Olga Foriasz 277529, Szymon Smoła 282252

2025-03-31

Spis treści

1 Krótki opis zagadnienia

W sprawozdaniu będziemy przeprowadzać analizę wybranych danych przedstawionych w zbiorze danych WA_Fn-UseC_-Telco-Customer-Churn.csv, oraz szczegółową interpretację otrzymanych wyników. Zbiór danych zawiera informacje o klientach pewnej firmy telekomunikacyjnej. Dane pochodzą ze strony Kaggle (źródło: https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Podczas analizy będziemy chcieli znaleźć odpowiedzi na poniżej przedstawione pytania:

- Co sprawia, że klienci pozostają lojalni i korzystają ponownie z usług?
- Dlaczego pewni klienci rezygnują z usług i co za to odpowiada? Czy da się rozpoznać, co łączy wszystkich rezygnujących?
- Co należałoby poprawić, aby zwiększyć procent lojalnych klientów?

2 Opis eksperymentów/analiz

- W raporcie zostaną przeprowadzone analizy, w jaki sposób poszczególne dane wpływają na decyzje klientów.
- W analizie wykorzystane zostaną metody graficzne. Na wykresach kołowych oraz słupkowych ukazane zostaną wybrane zmienne, przedstawiające ciekawe obserwacje dotyczące klientów. Dzięki nim będzie możliwe zauważenie ogromnych dysproporcji w niektórych kluczowych obserwacjach. Dodatkowo na wykresach korelacji będziemy chcieli odszukać zależności liniowych pomiędzy zmiennymi ciągłymi.

3 Wyniki

Nasze dane składały się początkowo z 21 cech i 7043 przypadków, ale nie wszystkie cechy będą nam potrzebne w analizie. Wykorzystamy zatem 20 cech ze wszystkich, które posiadamy.

Dane składają się z 3 kolumn o wartościach numerycznych, 17 o wartościach factor oraz 1 odpowiedzialnej za ID klienta. W analizie nie przyda się nam informacja o ID klienta, więc można ją usunąć z ramki danych, która będzie później użyta w dalszej części analizy.

Brakujące wartości zostały oznaczone poprzez NA, a rekordy którym one odpowiadały zostały usunięte z ramki danych i teraz analizowanych przypadków jest 7032.

Nie zaobserwowano żadnych niestandardowych notacji w danych.

3.1 Fragmenty R-kodów

Na podstawie danych, wyznaczamy podstawowe wskaźniki dla zmiennych ilościowych

Tabela 1: Podstawowe wskaźniki sumaryczne dla zmiennych czas trwania, miesięczne opłaty i całkowite opłaty

	czas trwania	miesięczne opłaty	całkowite opłaty
Min.	1.000	18.250	18.800
1st Qu.	9.000	35.588	401.450
Median	29.000	70.350	1397.475
Mean	32.422	64.798	2283.300
3rd Qu.	55.000	89.862	3794.738
Max.	72.000	118.750	8684.800

Tabela 1 przedstawia podstawowe wskaźniki dla zmiennych ilościowych. Możemy odczytać, że dla zmiennej pierwszej - czas trwania usługi - mediana oraz średnia jest w miarę do siebie zbliżona. Odpowiednio wynoszą 29 i niemal 32,5 miesiąca. Da się także zaobserowawać duży rozrzut pomiędzy minimum a maksimum czasu trawnia usługi. Minimalny czas wynosi 1 miesiąc, podczas gdy maksymalny 72 - czyli 6 lat. Podobne wnisoki możemy wyciągnąc dla zmiennych Oplaty_miesieczne oraz Oplaty_laczne. Minimum dla obu zmiennych jest do siebie zbliżone, w przeciwieństwie do maksimum, które dla łącznych opłat jest niemalże 73 razy większe. Jak możemy zaobserwować wartości wahają się pomiędzy 1-72 miesięcy dla długości trwania usługi, od lekko ponad 18 do prawie 119 dla miesięcznych opłatoraz od ponad 18 do aż niemalże 8685 dla opłat łącznych. Można także zauważyć, że żadna ze zmiennych nie ma rozkładu symetrycznego, jednak najbardziej zbliżony do niego jest histogram ukazujący zmienną odpowiadającą czasowi trwania usługi, dla każdego klienta. Największą zmiennością charakteryzują się opłaty łączne. Duży procent klientów, jest nimi przez niewielki okres czasu, a tylko dla nielicznych opłaty łączne sięgają górnych kresów - tj. około 8000.

Oraz podstawowe wskaźniki sumaryczne dla zmiennych jakościowych:

Tabela 2: Podstawowe wskaźniki sumaryczne dla zmiennych jakościowych

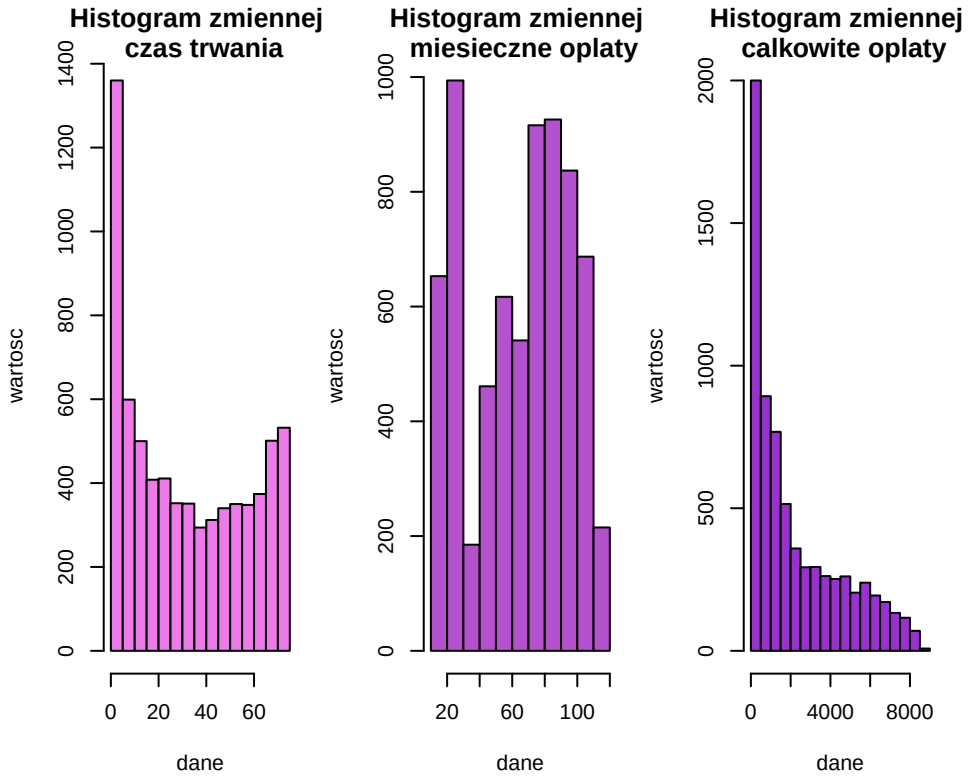
Zmienna	Wartość	Liczność
płeć	Kobieta	3483
płeć	Mężczyzna	3549
senior	Nie	5890
senior	Tak	1142
partner	Nie	3639
partner	Tak	3393
osoby zależne	Nie	4933
osoby zależne	Tak	2099
usługa telefoniczna	Nie	680
usługa telefoniczna	Tak	6352
wiele linii telefonicznych	Nie	3385
wiele linii telefonicznych	Brak usługi telefonicznej	680
wiele linii telefonicznych	Tak	2967
usługa internetowa	Cyfrowa linia abonencka	2416
usługa internetowa	Światłowód	3096
usługa internetowa	Brak	1520
ochrona online	Nie	3497
ochrona online	Brak usługi internetowej	1520
ochrona online	Tak	2015
kopia zapasowa online	Nie	3087
kopia zapasowa online	Brak usługi internetowej	1520
kopia zapasowa online	Tak	2425
ochrona urządzenia	Nie	3094
ochrona urządzenia	Brak usługi internetowej	1520
ochrona urządzenia	Tak	2418
wsparcie techniczne	Nie	3472
wsparcie techniczne	Brak usługi internetowej	1520
wsparcie techniczne	Tak	2040
telewizja strumieniowa	Nie	2809
telewizja strumieniowa	Brak usługi internetowej	1520
telewizja strumieniowa	Tak	2703
filmy strumieniowe	Nie	2781
filmy strumieniowe	Brak usługi internetowej	1520
filmy strumieniowe	Tak	2731
umowa	Z miesiąca na miesiąc	3875
umowa	Jednoletnia	1472
umowa	Dwuletnia	1685
faktura elektroniczna	Nie	2864
faktura elektroniczna	Tak	4168
metoda płatności	Czek elektroniczny	2365

Zmienna	Wartość	Liczność
metoda płatności	Czek wysyłany pocztą	1604
metoda płatności	Przelew bankowy	1542
metoda płatności	Karta kredytowa	1521
rezygnacja z usług	Nie	5163
rezygnacja z usług	Tak	1869

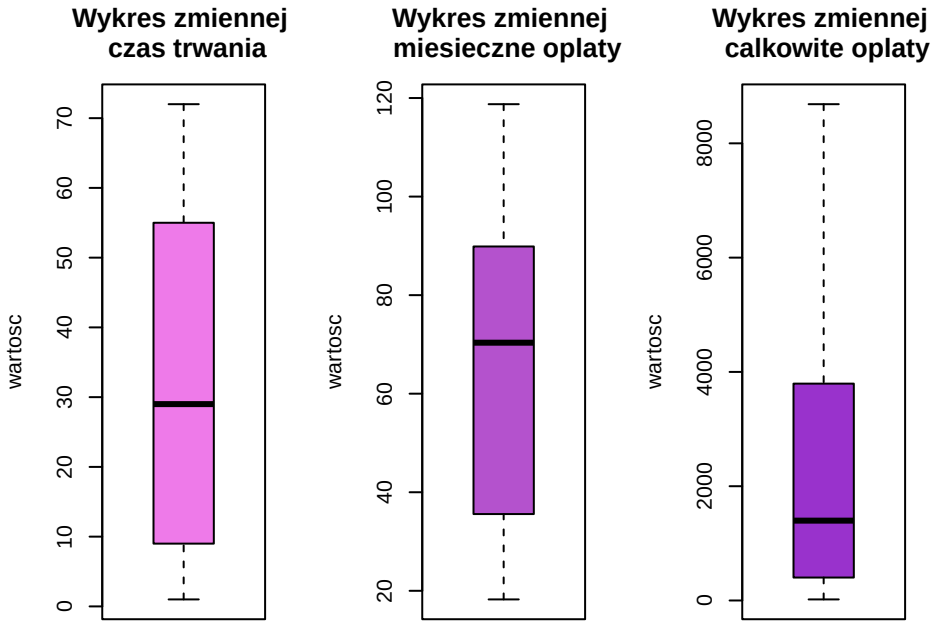
Z tabeli numer 2 można odczytać dane na temat każej zmiennej jakościowej, jednak nie wszystkie przedstawiają dane przydatne do analizy. Z tego względu zajmiemy się omówieniem ciekawszych z nich oraz takich, które mogą zostać użyte do wyciągnięcia interesujących nas wniosków. Pierwszą informacją jest to, że płeć klientów nie ma znaczenia - to znaczy, że firma dociera do obiorców zarówno płci męskiej, jak i damskiej. Następnie na podstawie zmiennej Seniorzy możemy zauważyć, że znaczna większość klientów to ludzie nie posiadający statusu seniora - jest ich dokładnie 5890, podczas gdy seniorów raptem 1142. Ciekawe obserwacje dotyczą także typu umowy - najpopularniejszą jest umowa miesięczna. Może to świadczyć, że klienci cenią sobie mozliwość bycia elsatycznym oraz brak długotrwałych zobowiązań. Co z kolei wiąże się z tym, że mogą być prędzej skorzy do odejścia niż klienci z dłuższymi umowami. Warte zauważenia jest także, to najpopularniejszą metodą płatności jest czek elektroniczny. Podczas, gdy 3 pozostałe metody płatności są wybierane z tą samą częstotliwością. Ostatnią kluczową informacją, która jest także celem naszej analizy, jest zmienna Rezygnacja_z_usługi. Ponad 25% klientów rezygnuje z usług.

3.2 Wykresy

Na podstawie danych wygenerowane zostały wykresy dla zmienncyh ilościowych:



Rysunek 1: Histogramy dla zmiennych ilościowych

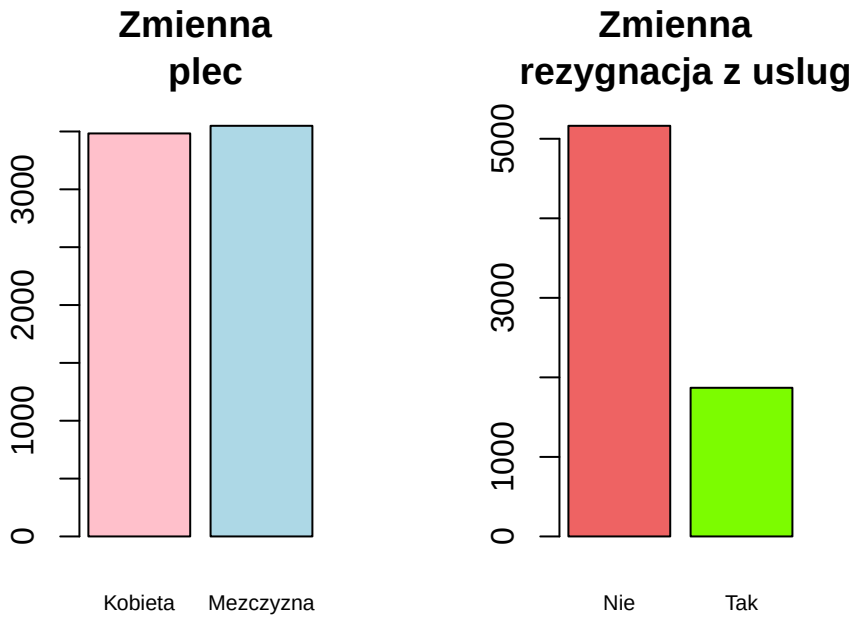


Rysunek 2: Wykresy pudełkowe dla zmiennych ilościowych

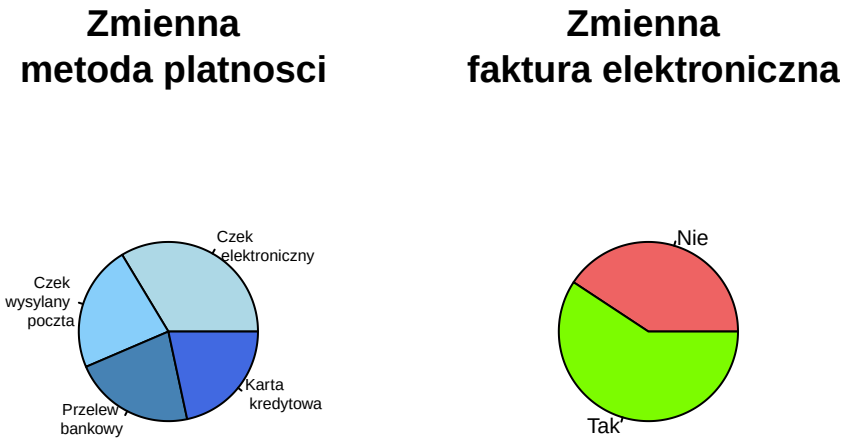
Wykresy są graficzną metodą przedstawienia danych ilościowych. Do zmiennych przedstawionych na wykresach na Rysunku 1. oraz 2. odnosi się także Tabela 1, pod którą znajduje

się krótka analiza wskaźników sumarycznych. Wykresy dla zmiennej czasu trwania usługi, potwierdzają tezę wysuniętą z odczytywania wskaźników sumarycznych. Najważniejszą obserwacją jest to, że umowy na krótki okres czasu mają zdcecydowaną przewagę nad pozostałymi możliwościami. Wykresy dla zmiennej opłat miesięcznych, pokazują że wartości najpierw rosną, następnie zaliczają ogromny spadek. Po czym znów rosną i ponownie opadają, jednak już nie tak drastycznie. Ostatnie Wykresy przedstawiające zmienną Oplaty_laczne ukazują widoczne skumulowanie wartości w dolnych przedziałach. Oznacza to, że znaczna liczba klientów płaci podobne stawki, które oscylują w dolnych granicach cen opłat łącznych.

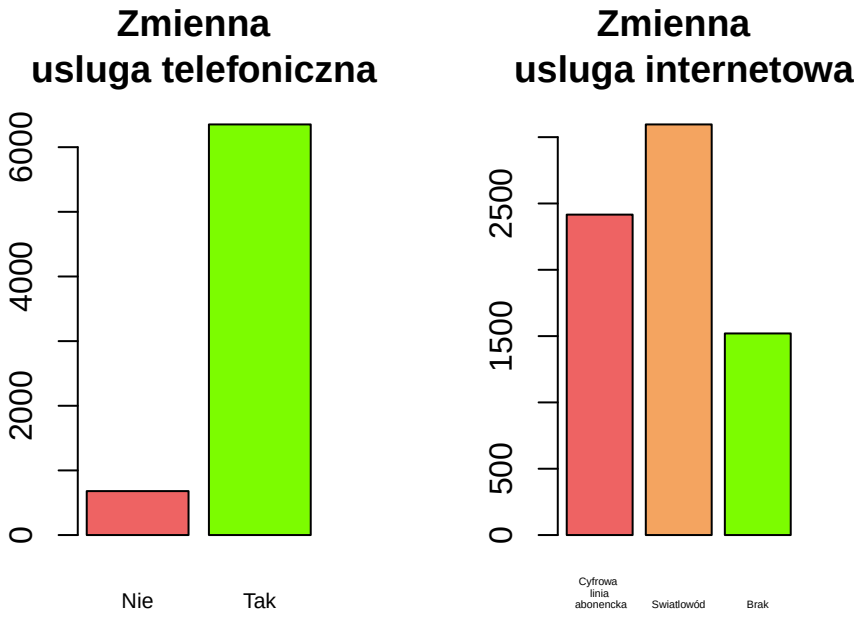
Poniżej znajdują się wybrane wykresy dla zmiennych jakościowych



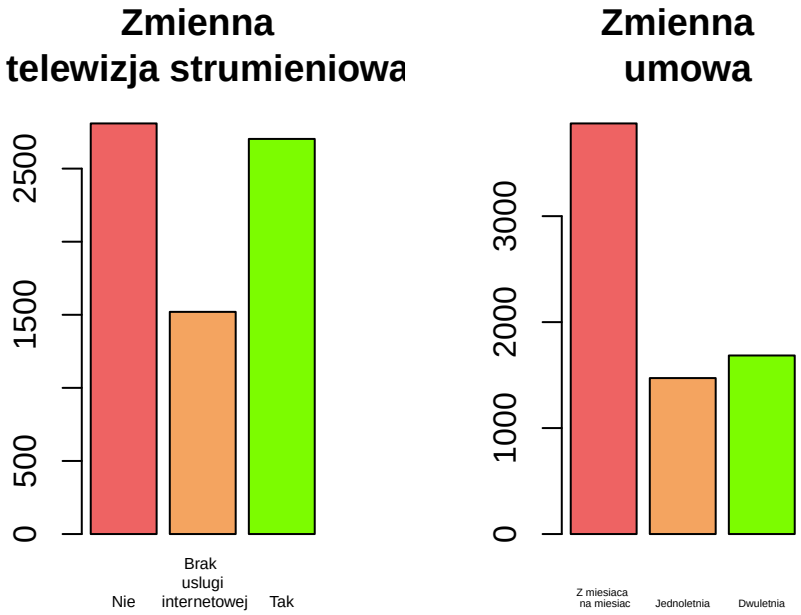
Rysunek 3: Wybrane wykresy dla zmiennych jakościowych



Rysunek 4: Wybrane wykresy dla zmiennych jakościowych



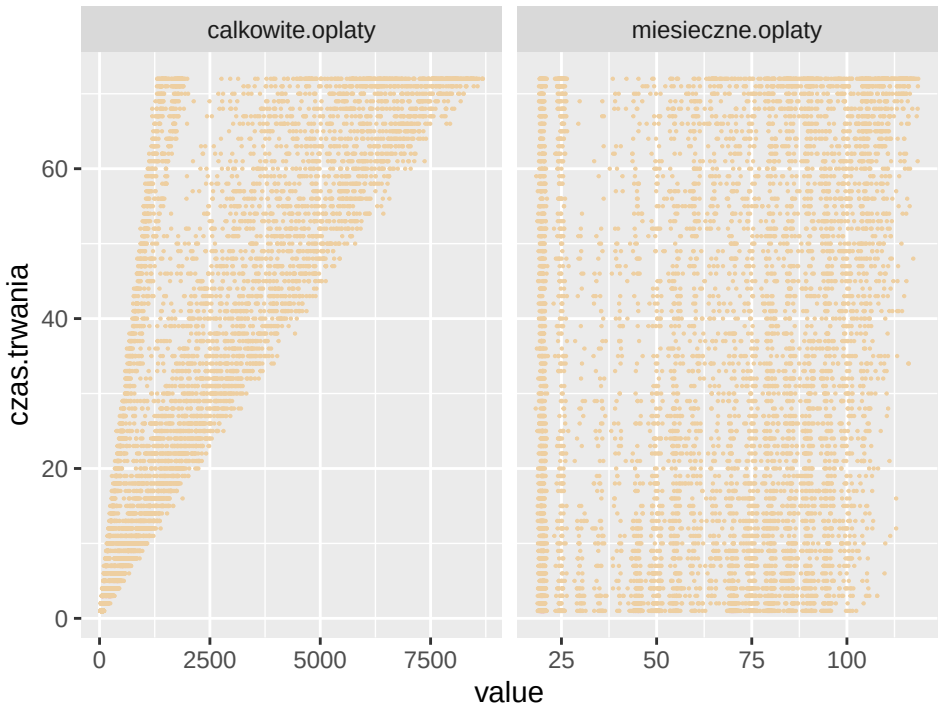
Rysunek 5: Wybrane wykresy dla zmiennych jakościowych



Rysunek 6: Wybrane wykresy dla zmiennych jakościowych

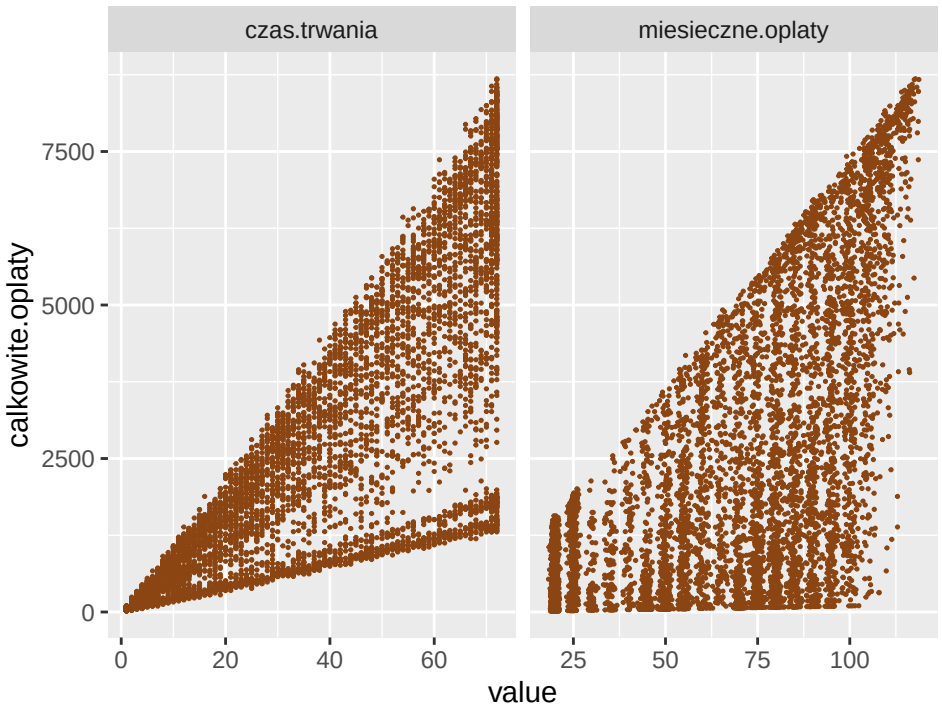
Wykresy na Rysunku 3-6 przedstawiają graficznie zmienne jakościowe. Na posdstawie wykresu, dochodzimy do analogicznych wniosków jak przy analizie Tabeli 2.

Następnie wykresy przedstawiające zależności pomiędzy parami zmiennych ciągłych



Rysunek 7: Wykresy zależności zmiennych ciągłych

Rysunek 8: Wykresy zależności zmiennych ciągłych



Rysunek 9: Wykresy zależności zmiennych ciągłych

Wykresy zależności ukazują, że pomiędzy zmiennymi nie ma zależności liniowych, najbardziej zbliżone do oczekiwanych efektów można zaobserwować na Rysunku 5. oraz 7., gdzie przedstawione są zależności zmiennej czasu trwania usługi od łącznych opłat bądź na odwrót. Tam zależność także nie jest idealnie liniowa, ze względu na ukazany duży rozrzut danych, jednak są to jedyne zmienne, których zależność choćby w pewnym stopniu zbliża się do zależności liniowej. Oczywiście jest to wyjaśnione tym, że im dłużej trwa umowa, tym większe koszty ponosi klient, czyli rosną łączne opłaty.

Tabela 3: Wybrane wskazniki sumaryczne dla zmiennych czas trwania, miesięczne opłaty i całkowite opłaty, dla kielntów, którzy odeszli

	czas trwania	miesięczne opłaty	całkowite opłaty
Min.	1.000	18.850	18.850
1st Qu.	2.000	56.150	134.500
Median	10.000	79.650	703.550
Mean	17.979	74.441	1531.796
3rd Qu.	29.000	94.200	2331.300
Max.	72.000	118.350	8684.800

Tabela 4: Wybrane wskazniki sumaryczne dla zmiennych czas trwania, miesięczne opłaty i całkowite opłaty dla klientów, którzy zostali

	czas trwania	miesięczne opłaty	całkowite opłaty
Min.	1.00	18.250	18.800
1st Qu.	15.00	25.100	577.825
Median	38.00	64.450	1683.600
Mean	37.65	61.307	2555.344
3rd Qu.	61.00	88.475	4264.125
Max.	72.00	118.750	8672.450

Podstawowe wskaźniki sumaryczne zmiennych jakościowych dla klientów, którzy odeszli:

Tabela 5: Podstawowe wskaźniki sumaryczne dla zmiennych jakościowych dla klientów, którzy odeszli

Zmienna	Wartość	Liczność
płeć	Kobieta	939
płeć	Mężczyzna	930
senior	Nie	1393
senior	Tak	476
partner	Nie	1200
partner	Tak	669
osoby zależne	Nie	1543
osoby zależne	Tak	326
usługa telefoniczna	Nie	170
usługa telefoniczna	Tak	1699
wiele linii telefonicznych	Nie	849
wiele linii telefonicznych	Brak usługi telefonicznej	170
wiele linii telefonicznych	Tak	850
usługa internetowa	Cyfrowa linia abonencka	459
usługa internetowa	Światłowód	1297
usługa internetowa	Brak	113
ochrona online	Nie	1461
ochrona online	Brak usługi internetowej	113
ochrona online	Tak	295
kopia zapasowa online	Nie	1233
kopia zapasowa online	Brak usługi internetowej	113
kopia zapasowa online	Tak	523
ochrona urządzenia	Nie	1211
ochrona urządzenia	Brak usługi internetowej	113
ochrona urządzenia	Tak	545

Zmienna	Wartość	Liczność
wsparcie techniczne	Nie	1446
wsparcie techniczne	Brak usługi internetowej	113
wsparcie techniczne	Tak	310
telewizja strumieniowa	Nie	942
telewizja strumieniowa	Brak usługi internetowej	113
telewizja strumieniowa	Tak	814
filmy strumieniowe	Nie	938
filmy strumieniowe	Brak usługi internetowej	113
filmy strumieniowe	Tak	818
umowa	Z miesiąca na miesiąc	1655
umowa	Jednoletnia	166
umowa	Dwuletnia	48
faktura elektroniczna	Nie	469
faktura elektroniczna	Tak	1400
metoda płatności	Czek elektroniczny	1071
metoda płatności	Czek wysyłany pocztą	308
metoda płatności	Przelew bankowy	258
metoda płatności	Karta kredytowa	232
rezygnacja z usług	Nie	0
rezygnacja z usług	Tak	1869

Podstawowe wskaźniki sumaryczne zmiennych jakościowych dla klientów, którzy zostali:

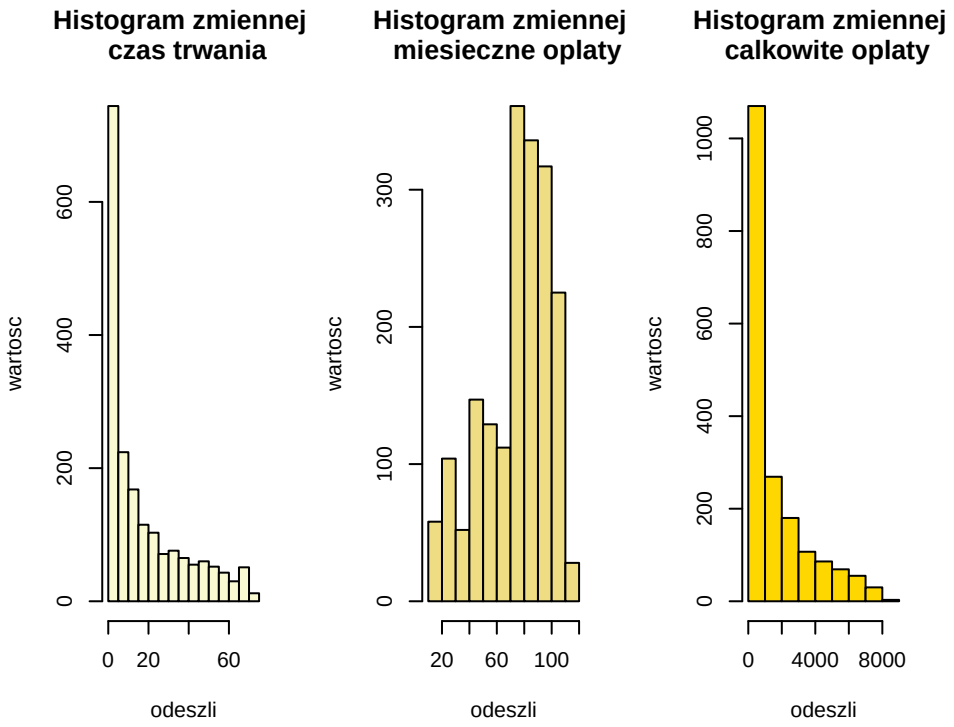
Tabela 6: Podstawowe wskaźniki sumaryczne dla zmiennych jakościowych, dla klientów, którzy zostali

Zmienna	Wartość	Liczność
płeć	Kobieta	2544
płeć	Mężczyzna	2619
senior	Nie	4497
senior	Tak	666
partner	Nie	2439
partner	Tak	2724
osoby zależne	Nie	3390
osoby zależne	Tak	1773
usługa telefoniczna	Nie	510
usługa telefoniczna	Tak	4653
wiele linii telefonicznych	Nie	2536
wiele linii telefonicznych	Brak usługi telefonicznej	510
wiele linii telefonicznych	Tak	2117
usługa internetowa	Cyfrowa linia abonencka	1957
usługa internetowa	Światłowód	1799

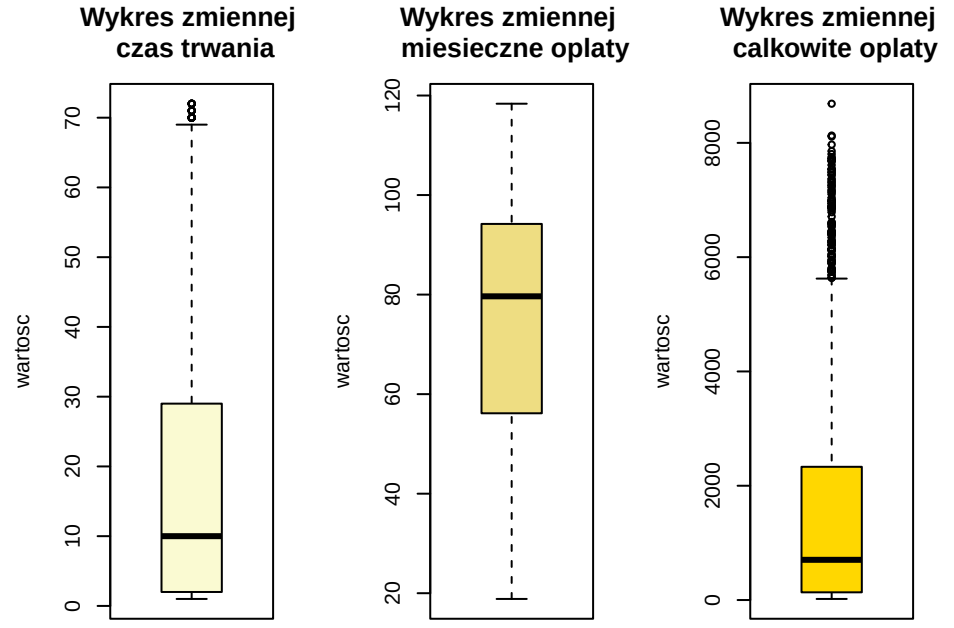
Zmienna	Wartość	Liczność
usługa internetowa	Brak	1407
ochrona online	Nie	2036
ochrona online	Brak usługi internetowej	1407
ochrona online	Tak	1720
kopia zapasowa online	Nie	1854
kopia zapasowa online	Brak usługi internetowej	1407
kopia zapasowa online	Tak	1902
ochrona urządzenia	Nie	1883
ochrona urządzenia	Brak usługi internetowej	1407
ochrona urządzenia	Tak	1873
wsparcie techniczne	Nie	2026
wsparcie techniczne	Brak usługi internetowej	1407
wsparcie techniczne	Tak	1730
telewizja strumieniowa	Nie	1867
telewizja strumieniowa	Brak usługi internetowej	1407
telewizja strumieniowa	Tak	1889
filmy strumieniowe	Nie	1843
filmy strumieniowe	Brak usługi internetowej	1407
filmy strumieniowe	Tak	1913
umowa	Z miesiąca na miesiąc	2220
umowa	Jednoletnia	1306
umowa	Dwuletnia	1637
faktura elektroniczna	Nie	2395
faktura elektroniczna	Tak	2768
metoda płatności	Czek elektroniczny	1294
metoda płatności	Czek wysyłany pocztą	1296
metoda płatności	Przelew bankowy	1284
metoda płatności	Karta kredytowa	1289
rezygnacja z usług	Nie	5163
rezygnacja z usług	Tak	0

3.3 Wykresy

Na podstawie danych klientów, którzy opuścili analizowaną firmę, wygenerowane zostały wykresy zmiennych ilościowych:

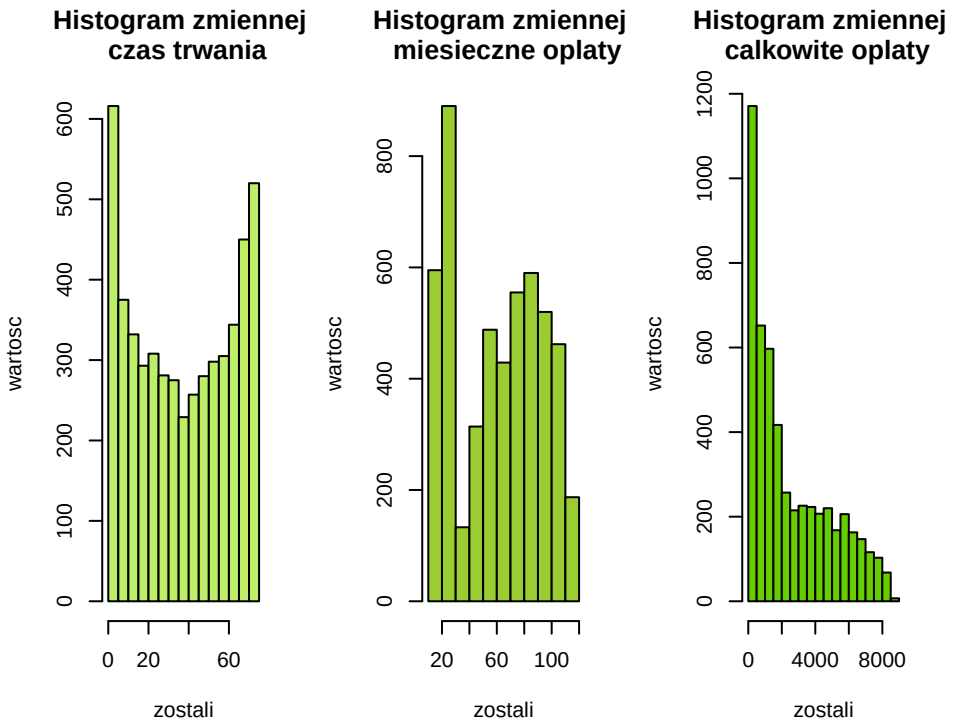


Rysunek 10: Histogramy dla zmiennych ilościowych, dla klientów, którzy odeszi

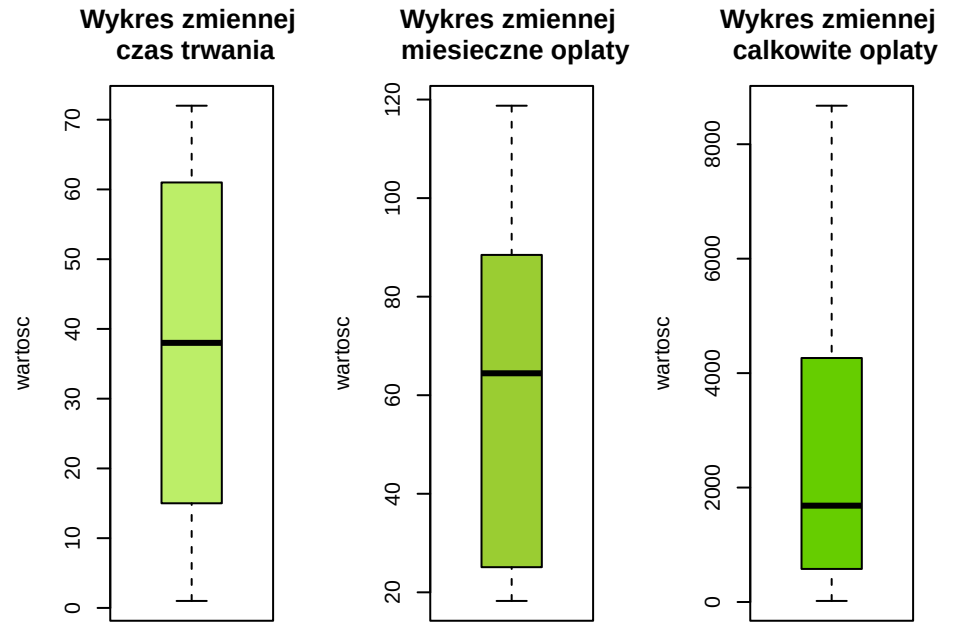


Rysunek 11: Wykresy pudełkowe dla zmiennych ilościowych, dla klientów, którzy odeszi

Podobnie na podstawie danych klientów, którzy nadal korzystają z usług analizowanej firmy, wygenerowane zostały wykresy zmiennych ilościowych:



Rysunek 12: Histogramy dla zmiennych ilościowych, dla klientów, którzy zostali

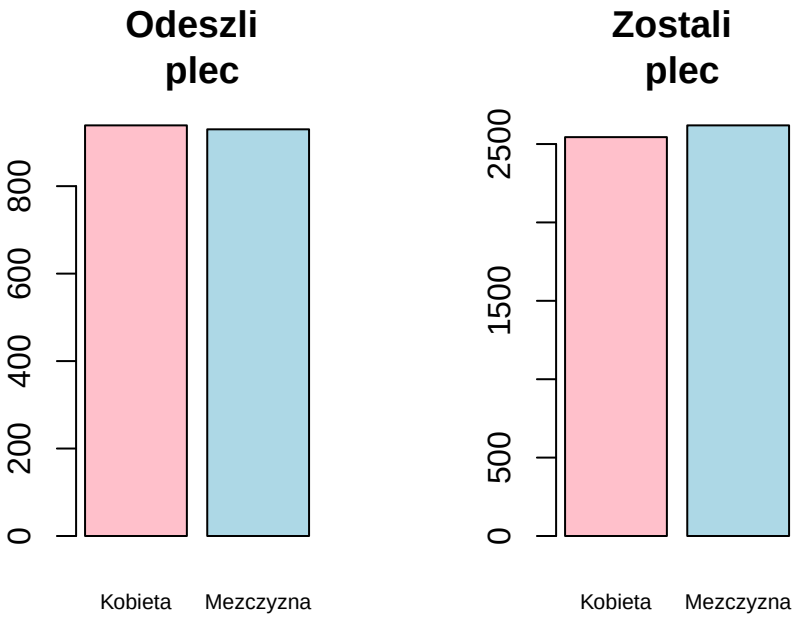


Rysunek 13: Wykresy pudełkowe dla zmiennych ilościowych, dla klientów, którzy zostali

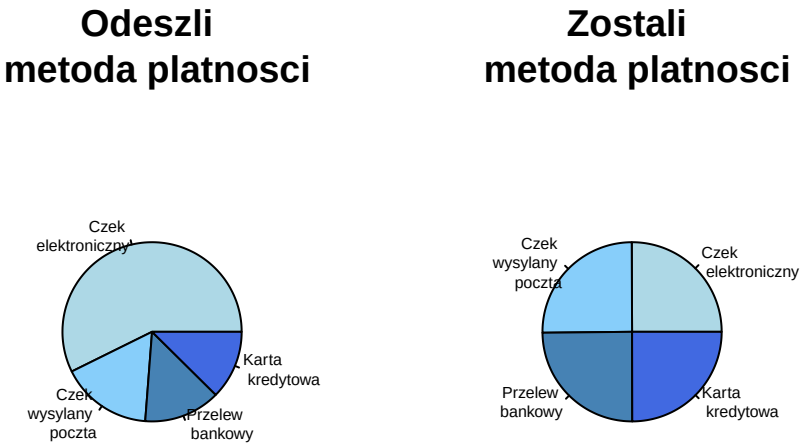
Histogramy zmiennych ilościowych odpowiedzialnych za koszta miesięcze i łączne przedstawione dla klientów, którzy odeszli i dla tych, którzy zostali przy analizowanej firmie w większości nie pokazują znacznych różnic między tymi grupami. Jedyne miejsce, w którym

te grupy odbiegają od siebie możemy zaobserwować przy niskich opłatach miesięcznych (w okolicach do 40). W takim przypadku widzimy, że znaczna część osób, która zdecydowała się na takie opłaty miesięczne nie zrezygnowała jeszcze z korzystania z usług świadczonych przez tę firmę. Natomiast histogramy zmiennej czas trwania znacznie się między sobą różnią. Ten obrazujący klientów lojalnych posiada nawet rozkład symetryczny, starych klientów (około 6 lat lojalności) jest prawie tylu co klientów nowych (do 5 miesięcy lojalności). Na ich podstawie możemy stwierdzić, że klienci, którzy odchodzą od analizowanej firmy robią to zazwyczaj przed upływem 10 miesięcy od rozpoczęcia korzystania z jej usług, a ci którzy wytrwają przy niej dłużej niż rok najprawdopodobniej zostaną w niej na znacznie dłuższy okres. [Im dłużej klient korzysta z usług firmy tym mniej prawdopodobne jest to, że z nich zrezygnuje]

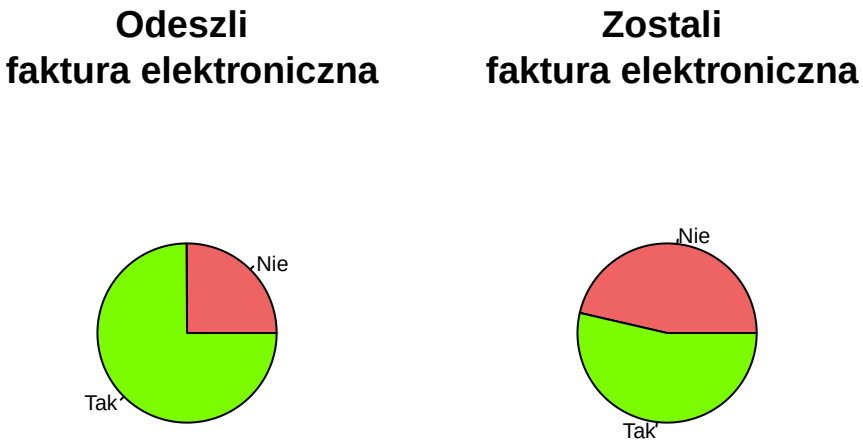
Poniżej znajduje się porównanie wykresów jakościowych klientów, którzy przestali korzystać z usług świadczonych przez firmę oraz tych którzy nadal z nich korzystają



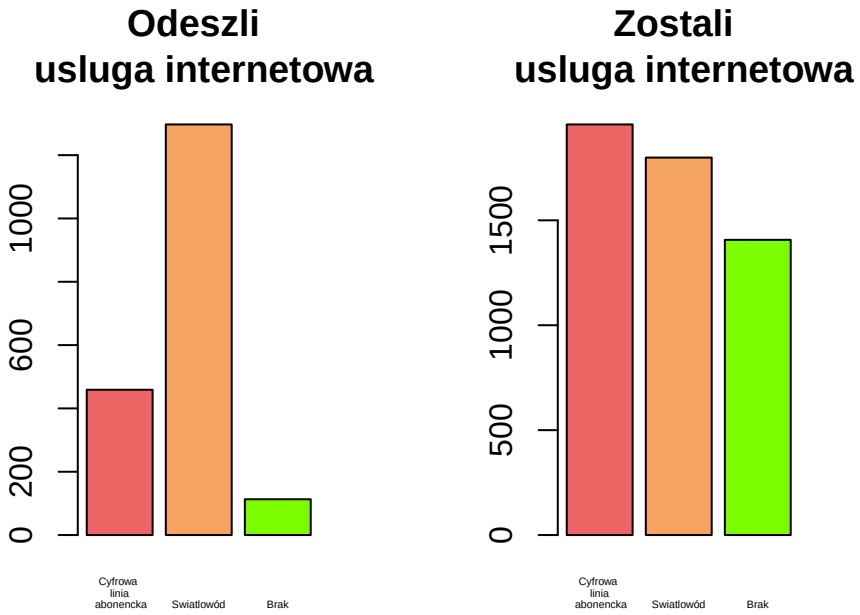
Rysunek 14: Wykres dla zmiennej jakościowej z podziałem na klientów, którzy odeszli oraz tych, którzy zostali



Rysunek 15: Wykres dla zmiennej jakościowej z podziałem na klientów, którzy odeszli oraz tych, którzy zostali



Rysunek 16: Wykres dla zmiennej jakościowej z podziałem na klientów, którzy odeszli oraz tych, którzy zostali



Rysunek 17: Wykres dla zmiennej jakościowej z podziałem na klientów, którzy odeszli oraz tych, którzy zostali

Rysunek 18: Wykres dla zmiennej jakościowej z podziałem na klientów, którzy odeszli oraz tych, którzy zostali

Z rysunku 15 jasno wynika, że pośród grupy klientów którzy zrezygnowali z usług firmy przeważającą metodą płatności były czeki elektroniczne.

Rysunek 16 mówi nam, że około 75% klientów odchodzących z analizowanej firmy korzystało z faktur online.

Na rysunku 17 widzimy, że znaczna większość klientów opuszczających firmę wybrało internet światłowodowy, natomiast między klientami lojalnymi nie ma aż tak widocznej różnicy w typach usług internetowych.

Dzięki rysunkowi 18 możemy stwierdzić, że klienci którzy wybierają kontrakty dłuższe niż te z miesiąca na miesiąc, o wiele rzadziej z nich rezygnują.

3.4 Podsumowanie

Na podstawie przeprowadzonej analizy można zauważyć, że aż 25% klientów rezygnuje. Ważną obserwacją było także to, że większość rezygnujących, to klienci, którzy zdecydowali się na umowy krótkookresowe. Natomiast płeć nie odgrywa roli, jeśli chodzi o klientów lojalnych i nielojalnych [Rysunek 14]. Zauważone zostało też, że osoby nieposiadające statusu seniora mają znaczną przewagę nad tymi posiadającymi. Na początkowych wykresach ukazany został duży rozrzut, jeśli chodzi o długość trwania usług. Co ciekawe jednak, żadna zmienna nie była z drugą w relacji liniowej. Najbliżej tego były zmienne odpowiadające długości lojalności klienta do kosztów łącznych, co jest spójne z oczekiwaniami. Następnie przeanalizowane zostały dane z podziałem na klientów, którzy odeszli oraz tych, którzy zostali. Mogliśmy dzięki temu zauważyć, że większość lojalnych klientów korzystała z wielu usług, podczas gdy osoby, które odstąpiły od umowy korzystały z mocno ograniczonej oferty.

Analizując dane dało się zobserwować wiele cech klientów. Były w tym zmienne, które charakteryzowały się tym, że klienci każdą z możlwiych opcji wybierali z tą samą czestotliwością. Wiemy także korzystając z Tabeli 2., że klienci korzystają z usług firmy bez względu na zmienne takie jak posiadanie partnera, płeć czy utrzymywanie innych. Jest do dobrą oznaką, ponieważ oznacza to, że oferta dociera do każdego typu klienta. Możemy też zauważyć, że jedną z najchetniej wybieranych usług jest internet, który zdecydowało się wykupić ponad połowa klientów. Następną najchętniej wykorzystywaną usługą jest obsługa wielu linii telefonicznych oraz kopia zapasowa online. Cenną informacją jest także fakt, że większość decyduje się na faktury elektorniczne, co może być powiązane z wiekiem klientów.

Klienci często odchodzą ze względu na krótkie okresy kontraktów (około 87% klientów, którzy odeszli miało kontrakty miesięczne [Rysunek 18]). Firma może spróbować naprawić ten problem wprowadzając więcej opcji kontraktowych, na przykład taką trwającą 3 miesiące. Takie rozwiązanie może sprawić, że klienci rzadziej będą zauważać pieniądze upływające z ich kont, a kontrakt nie będzie na tyle długi, żeby bawiali się czy z niego w pełni wykorzystają. Ewentualnie, jeśli firma nie chce wprowadzać nowych czasowych kontraktów to można rozważyć lekkie dostosowanie cen, ponieważ na rysunkach 10-13 przy miesięcznych opłatach widać, że klienci, którzy decydują się na niższe stawki chętniej zostają przy firmie. W ostateczności można pomyśleć o wprowadzeniu atrakcyjnych ofert startowych dla nowych klientów. Pozwoliłoby to uzyskać firmie nowych użytkowników, a im dłużej udałoby się ich w niej przetrzymać, tym mniej prawdopodobne będzie ich odejście (z danych przedstawiających zmienną czas trwania wynika, że działa to trochę jak siła nawyku)