Zaawansowana analityka biznesowa – siła modeli predykcyjnych

226160-D

# Segmentacja klientów zakładu ubezpieczeniowego i określenie czynników ryzyka


Weronika Banasiak

Numer albumu 72698

Joanna Hawrysz

Numer albumu 122952

Ksenia Soroka

Numer albumu 89716

Joanna Zakęs

Numer albumu 100930


WARSZAWA 2024

## Spis treści

## 1. Cel projektu

Celem projektu była segmentacja ubezpieczonych i określenie głównych czynników ryzyka determinujących wartość zgłaszanych szkód.

Do celów szczegółowych zadania zaliczono:

1. Poznanie rozkładu zmiennych, ich analizę graficzną i zbadanie relacji pomiędzy nimi, w celu selekcji zmiennych na potrzeby modelu;
2. Zbudowanie modelu segmentacyjnego k-średnich na podstawie wyselekcjonowanych zmiennych;
3. Interpretacja biznesowa utworzonego modelu;
4. Identyfikacja anomalii – prób wyłudzeń odszkodowania.

## 2. Eksploracyjna analiza danych

Celem analizy eksploracyjnej było poznanie rozkładu i innych cech zmiennych, które pozwolą na wyłonienie tych spośród nich, które będą najbardziej przydatne w budowie modelu. W pierwszej kolejności użyto metody .info() oraz n.unique() (w pętli), żeby poznać typ zmiennych, liczbę braków danych oraz liczbę unikatowych wartości, które przyjmuje każda zmienna. Braki danych występowały tylko dla zmiennych acc_type i police_report_avlb.

Ze względu na występowanie zbyt dużej liczby unikatowych wartości i brak przydatności w analizie, usunięto zmienne policy_id (1000 unikatowych wartości) i zip_code (995 unikatowych wartości). Następnie sprawdzono, czy któryś rekord jest duplikatem innego – duplikatów nie znaleziono. Dokonano również zmiany typu zmiennych dotyczących dat. Po wstępnej eksploracji i przygotowaniu zbioru danych możliwa była właściwa część eksploracyjnej analizy danych.

Na potrzeby analizy eksploracyjnej zmienne podzielono na trzy rodzaje:

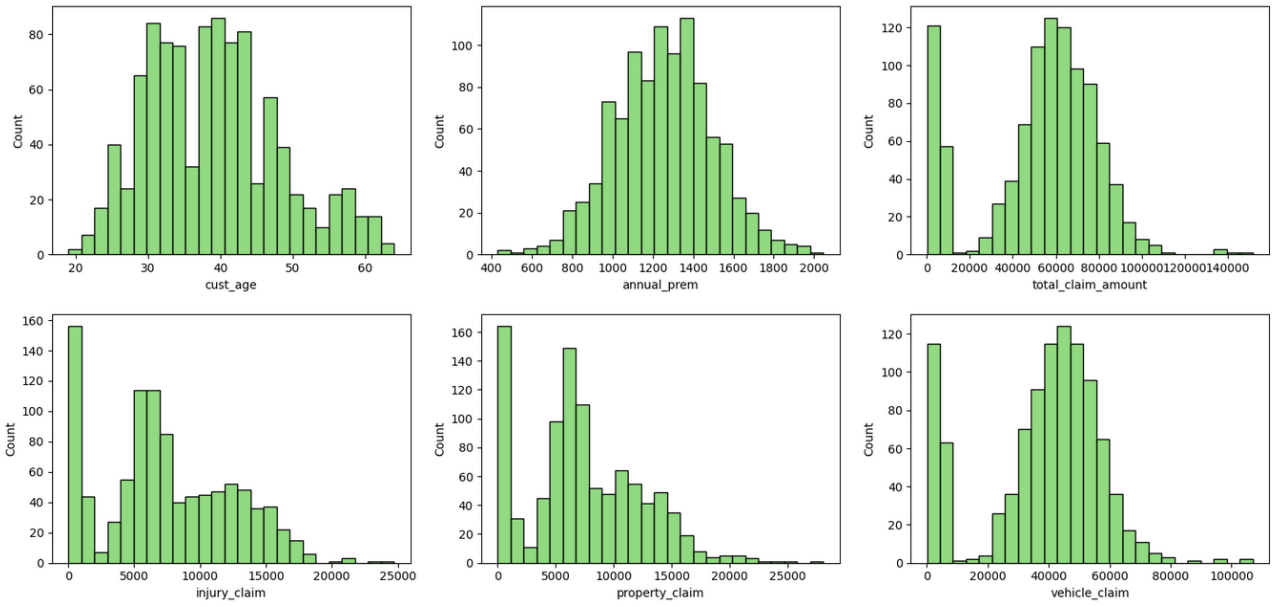
- „zmienne ilościowe” – zmienne liczbowe ciągłe;
- „zmienne mieszane” – zmienne liczbowe dyskretne;
- „zmienne jakościowe” – zmienne, które nie są liczbami.

Dla zmiennych ilościowych obliczono podstawowe miary statystyczne (Rysunek 1) i histogramy (Rysunek 2). Zauważono podobieństwo rozkładów par zmiennych: injury_claim z property_claim oraz

total_claim_amount z vehicle_claim. Podobieństwo to postanowiono sprawdzić na dalszym etapie analizy za pomocą korelacji.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **cust_age** | 1000.0 | 38.94800 | 9.140287 | 19.00 | 32.0000 | 38.00 | 44.000 | 64.0 |
| **annual_prem** | 1000.0 | 1254.51615 | 244.167395 | 431.44 | 1087.7175 | 1255.31 | 1413.805 | 2045.7 |
| **total_claim_amount** | 1000.0 | 52944.20100 | 26880.796007 | 100.00 | 41812.5000 | 58055.00 | 70592.500 | 151632.0 |
| **injury_claim** | 1000.0 | 7457.53900 | 4931.310627 | 0.00 | 4295.0000 | 6775.00 | 11330.000 | 24726.0 |
| **property_claim** | 1000.0 | 7429.98200 | 4908.125387 | 0.00 | 4445.0000 | 6750.00 | 10885.000 | 28054.0 |
| **vehicle_claim** | 1000.0 | 38056.68000 | 19210.616078 | 70.00 | 30292.5000 | 42100.00 | 50822.500 | 106960.0 |

*Rysunek 1. Podstawowe miary statystyczne zmiennych ilościowych*

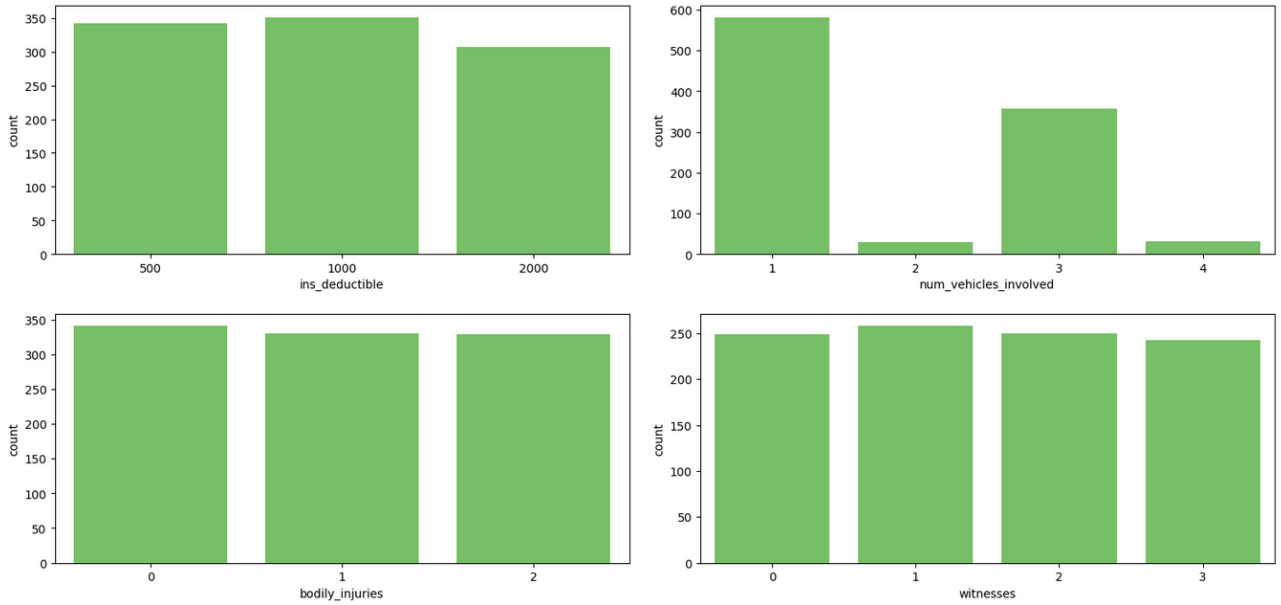



*Rysunek 2. Histogramy zmiennych ilościowych*

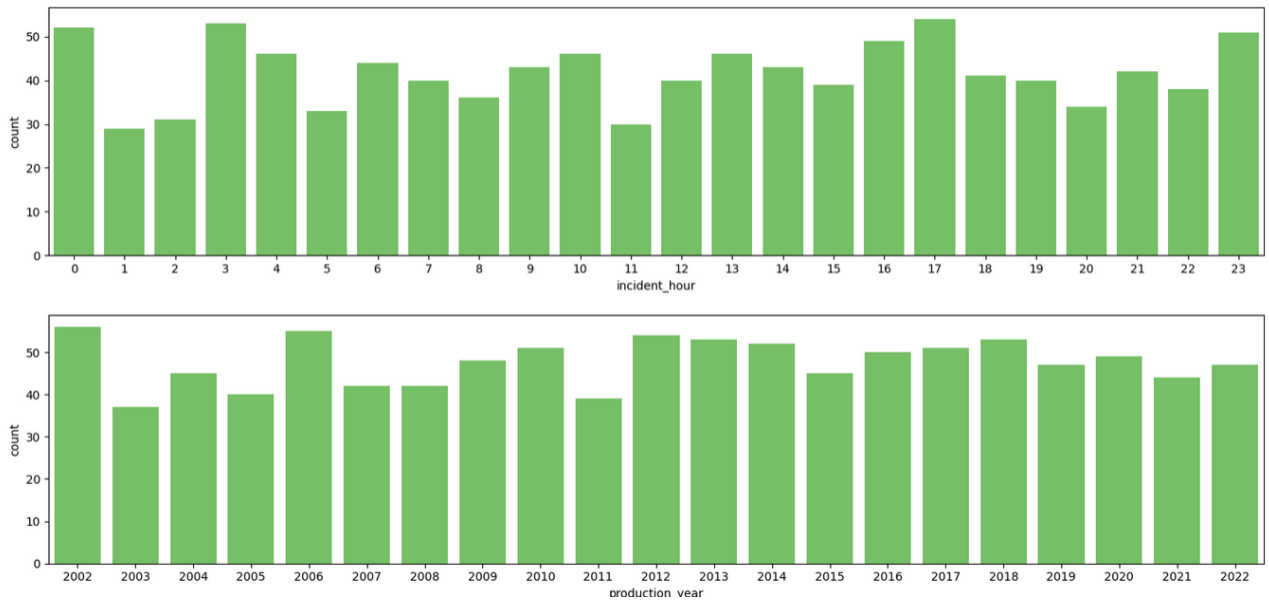
Działania te powtórzono dla grupy zmiennych mieszanych, jednakże zamiast histogramów utworzono wykresy liczności (Rysunki 3-5). Zauważono odchylenia standardowe zmiennych: num_vehicles_involved, bodily_injuries i witnesses w wartości około 1. Ich wariancje – kwadraty odchylenia standardowego – również muszą wynosić około 1. Oznacza to, że nie należy użyć tych zmiennych w modelu, ponieważ zbyt mało różnicują obserwacje, aby utworzyć dobrej jakości model klastrujący.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **ins_deductible** | 1000.0 | 1136.000 | 611.864673 | 500.0 | 500.0 | 1000.0 | 2000.0 | 2000.0 |
| **incident_hour** | 1000.0 | 11.644 | 6.951373 | 0.0 | 6.0 | 12.0 | 17.0 | 23.0 |
| **num_vehicles_involved** | 1000.0 | 1.839 | 1.018880 | 1.0 | 1.0 | 1.0 | 3.0 | 4.0 |
| **bodily_injuries** | 1000.0 | 0.988 | 0.818857 | 0.0 | 0.0 | 1.0 | 2.0 | 2.0 |
| **witnesses** | 1000.0 | 1.487 | 1.111335 | 0.0 | 1.0 | 1.0 | 2.0 | 3.0 |
| **production_year** | 1000.0 | 2012.126 | 6.015506 | 2002.0 | 2007.0 | 2012.0 | 2017.0 | 2022.0 |

*Rysunek 3. Podstawowe miary statystyczne zmiennych mieszanych*

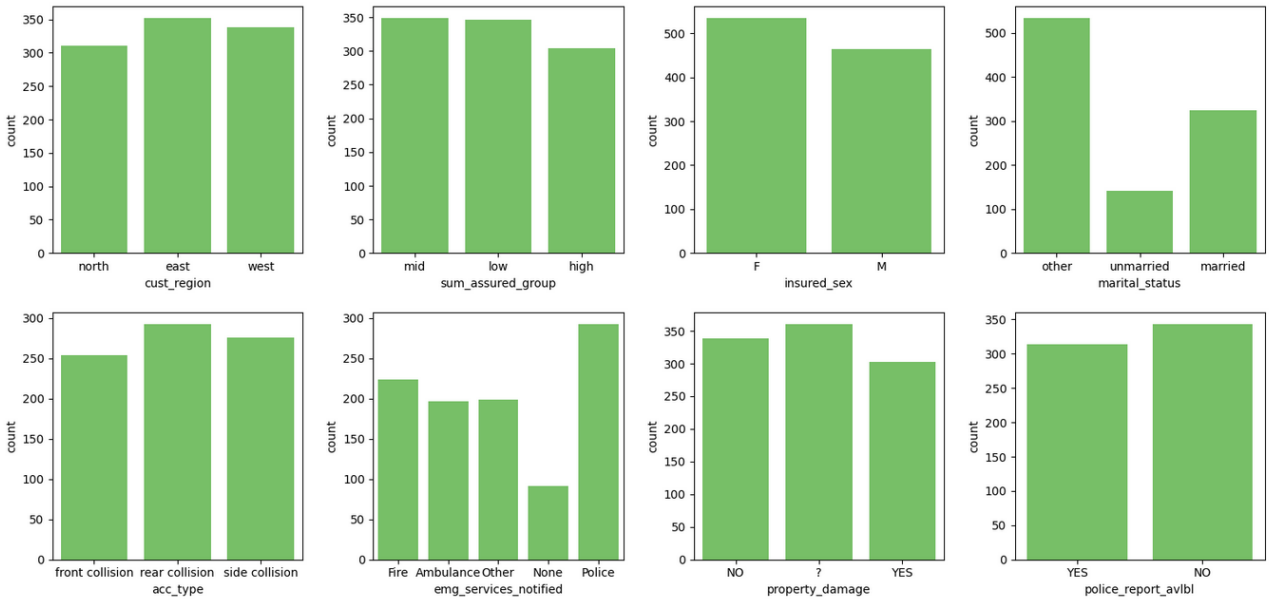



*Rysunek 4. Wykresy liczebności zmiennych mieszanych cz.1*

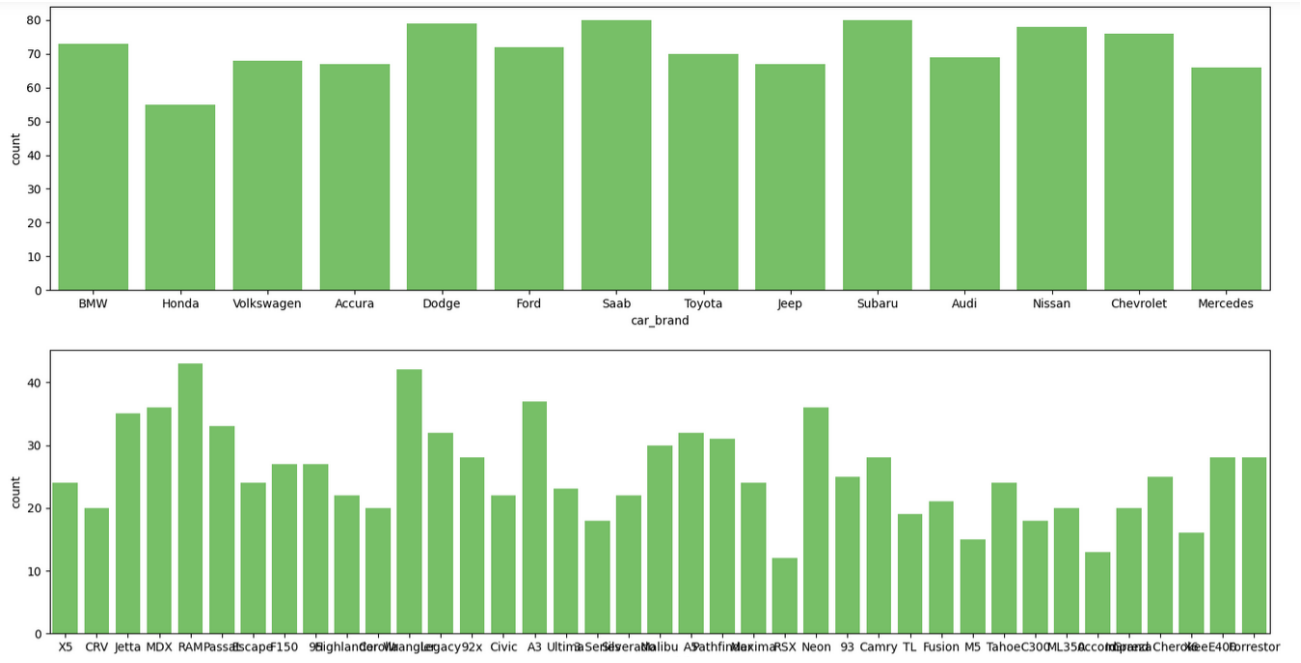



*Rysunek 5. Wykresy liczebności zmiennych mieszanych cz.2*

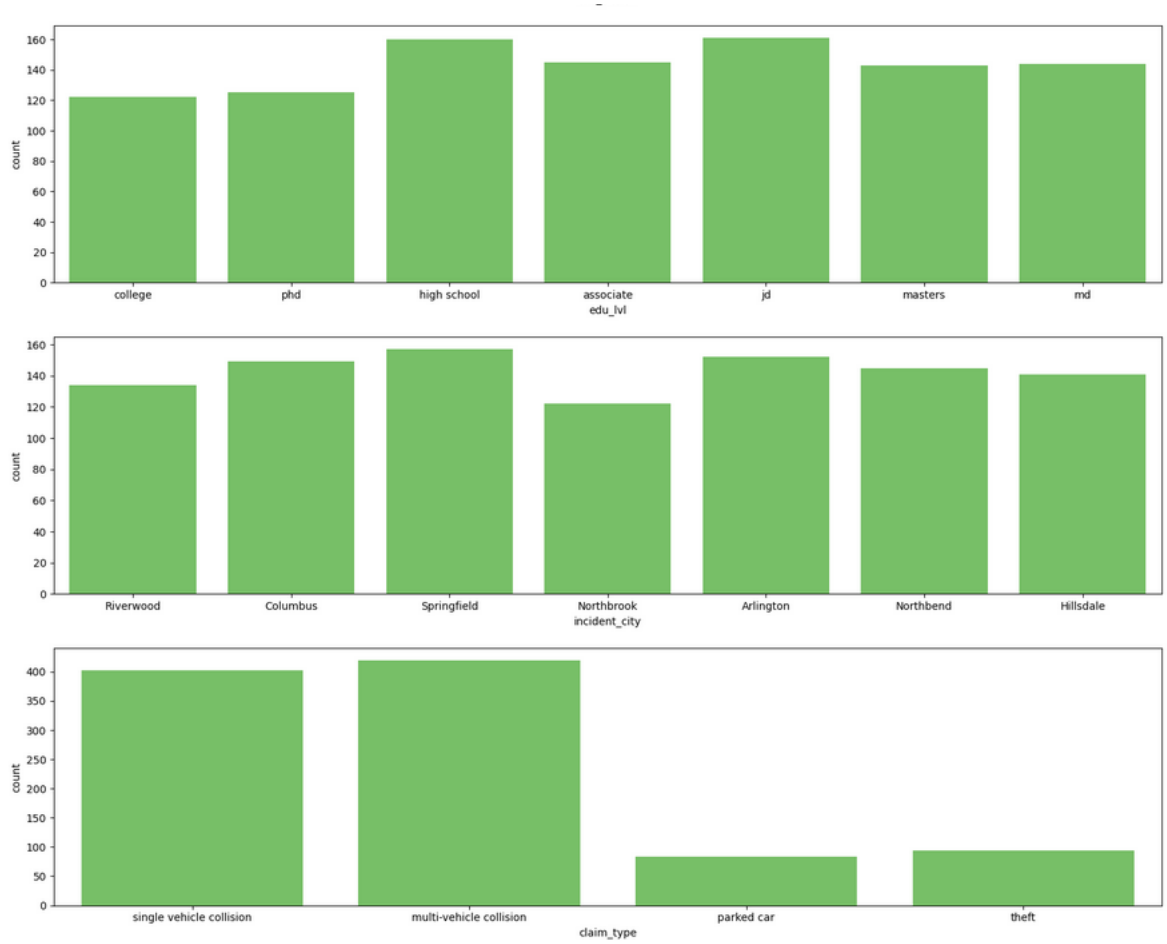
Ostatnia grupa zmiennych, również została przeanalizowana tak jak grupa zmiennych mieszanych. Natomiast ze względu, że nie są one liczbami, metoda .describe() nie dała ciekawych wyników. Na Rysunkach 6-8 przedstawiono rozkłady poszczególnych zmiennych.




*Rysunek 6. Wykresy liczebności zmiennych jakościowych cz.1*

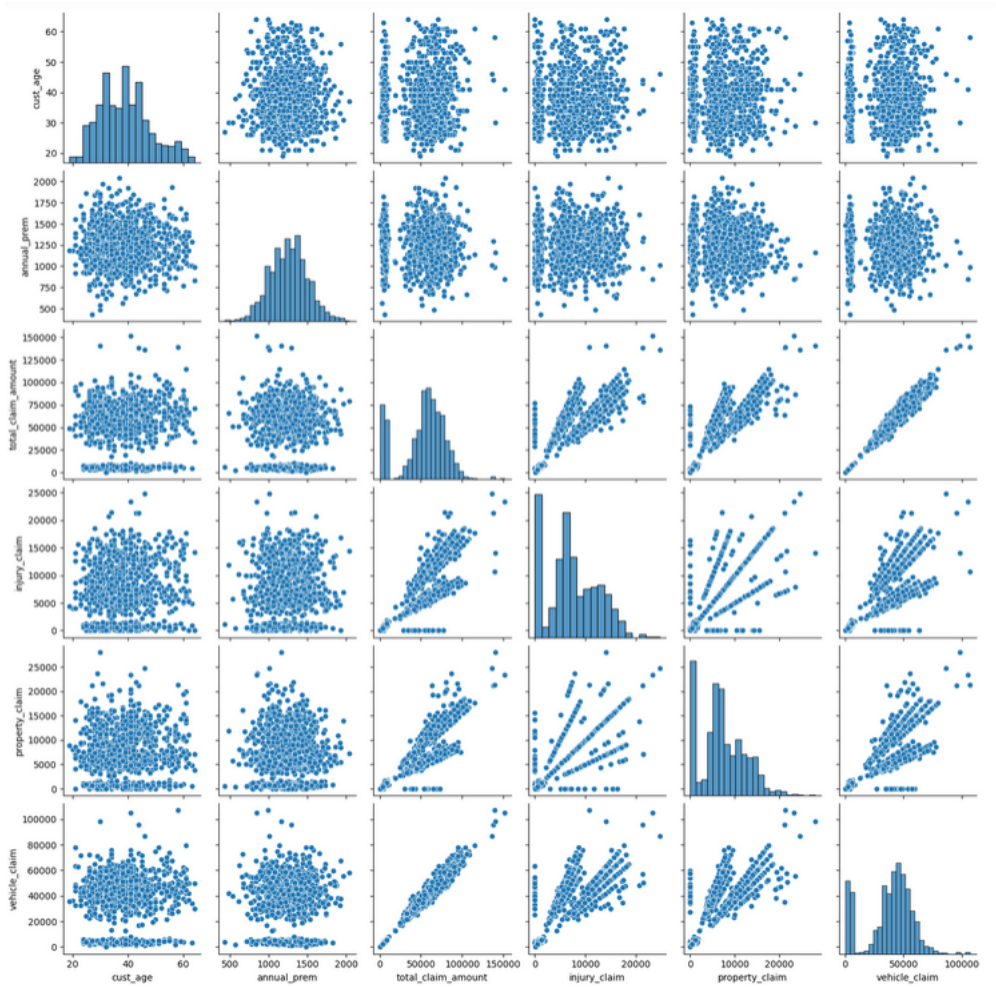



*Rysunek 7. Wykresy liczebności zmiennych jakościowych cz.2*

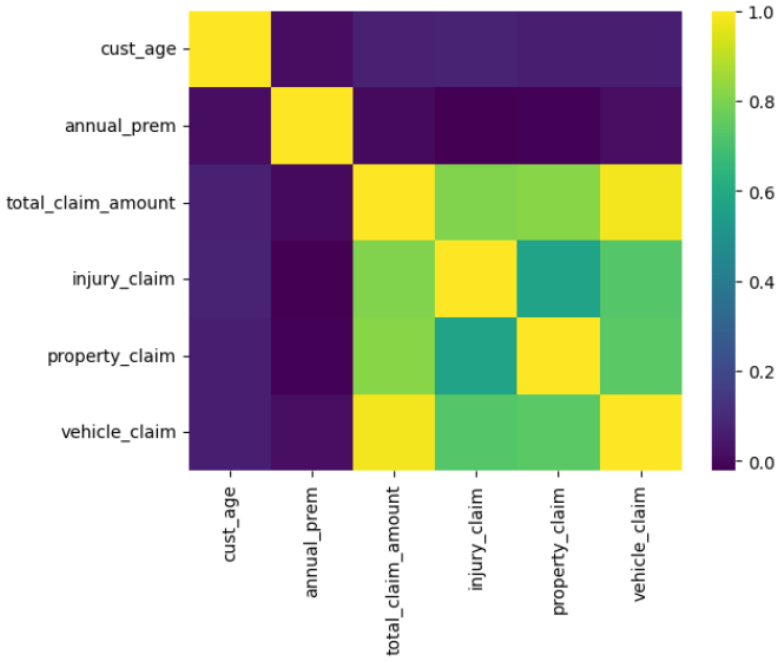



*Rysunek 8. Wykresy liczebności zmiennych jakościowych cz.3*

Badając zależności pomiędzy zmiennymi ilościowymi, zdecydowano się na wykonanie wykresów punktowych wszystkich zmiennych (pairplot), żeby przyjrzeć się relacji zidentyfikowanych wcześniej par zmiennych (Rysunek 9). Zauważono ciekawe rozkłady punktów, które mogą sugerować wysoką korelację lub nawet współliniowość, dla wszystkich zmiennych związanych z claim. Jest to logiczne – injury_claim, vehicle_claim i property_claim są składnikami total_claim_amount. Dla pewności wykonano również macierz korelacji pomiędzy zmiennymi i zwizualizowano ją funkcją heatmap (Rysunek 10). Jak podejrzewano, wszystkie 4 zmienne dotyczące claim są ze sobą silnie skorelowane (wartości korelacji od 0,8 do 0,98).

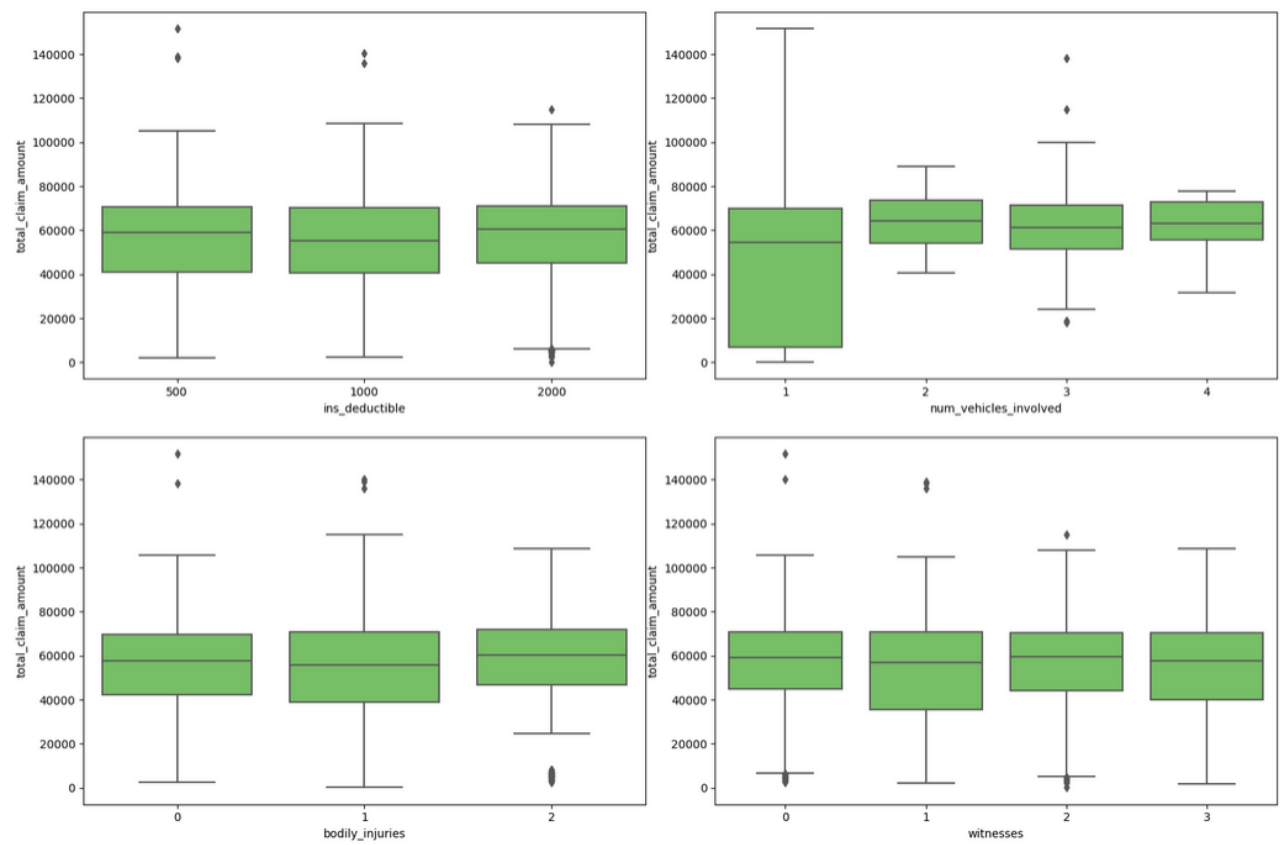



*Rysunek 9. Pairplot zmiennych ilościowych*

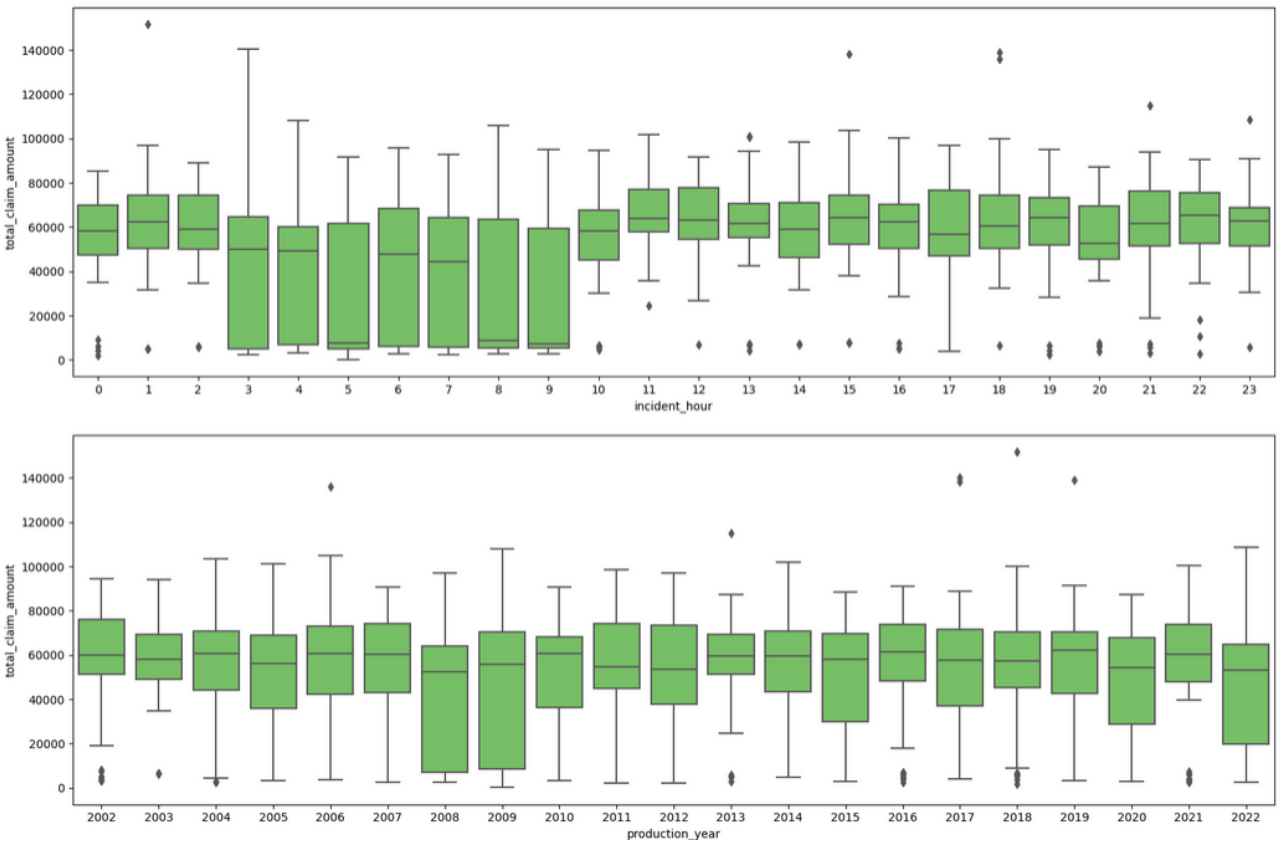



*Rysunek 10. Heatmap zmiennych ilościowych*

W przypadku zależności dla zmiennych mieszanych zdecydowano się na wykresy pudełkowe, żeby zobaczyć, jak kształtuje się total_claim_amount względem poszczególnych wartości zmiennej (Rysunek 11 i 12). Co zauważono wcześniej, mała wariancja powoduje brak zróżnicowania dla różnych wartości wybranych zmiennych – jedyną wyraźnie różną od pozostałych jest wartość 1 dla zmiennej num_vehicles_involved. Natomiast w przypadku zmiennych incydent_hour i production_year widoczne jest większe zróżnicowanie. Warto zwrócić uwagę szczególnie na godziny pomiędzy 3 a 9.



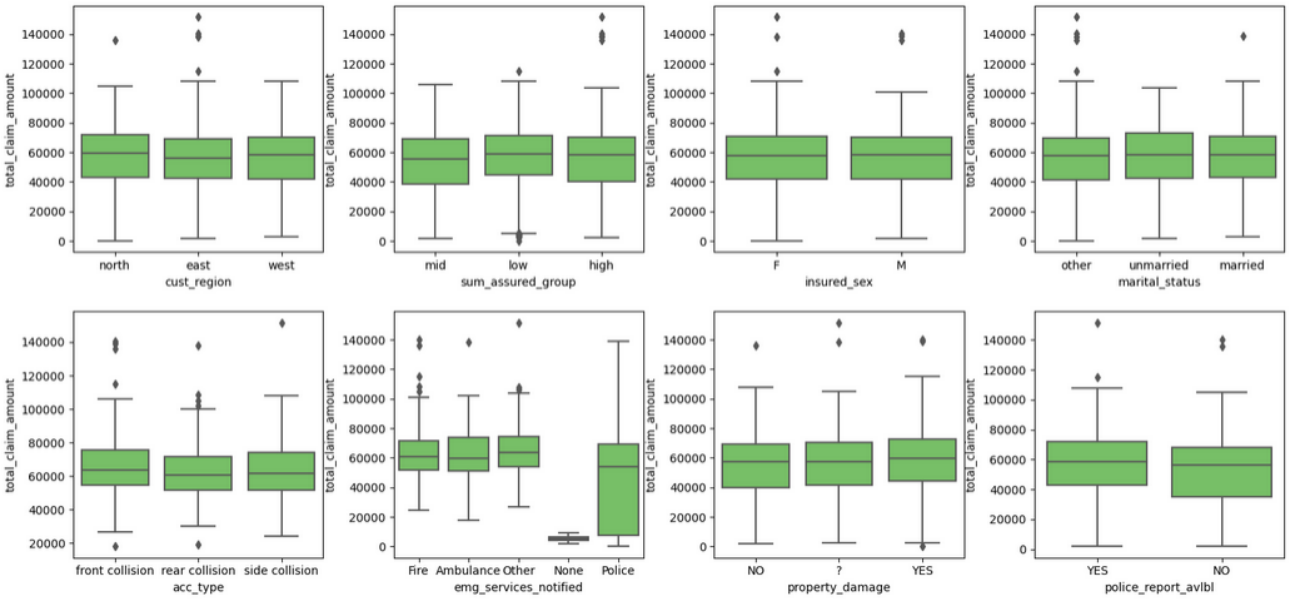

*Rysunek 11. Wykresy pudełkowe zmiennych mieszanych cz.1*



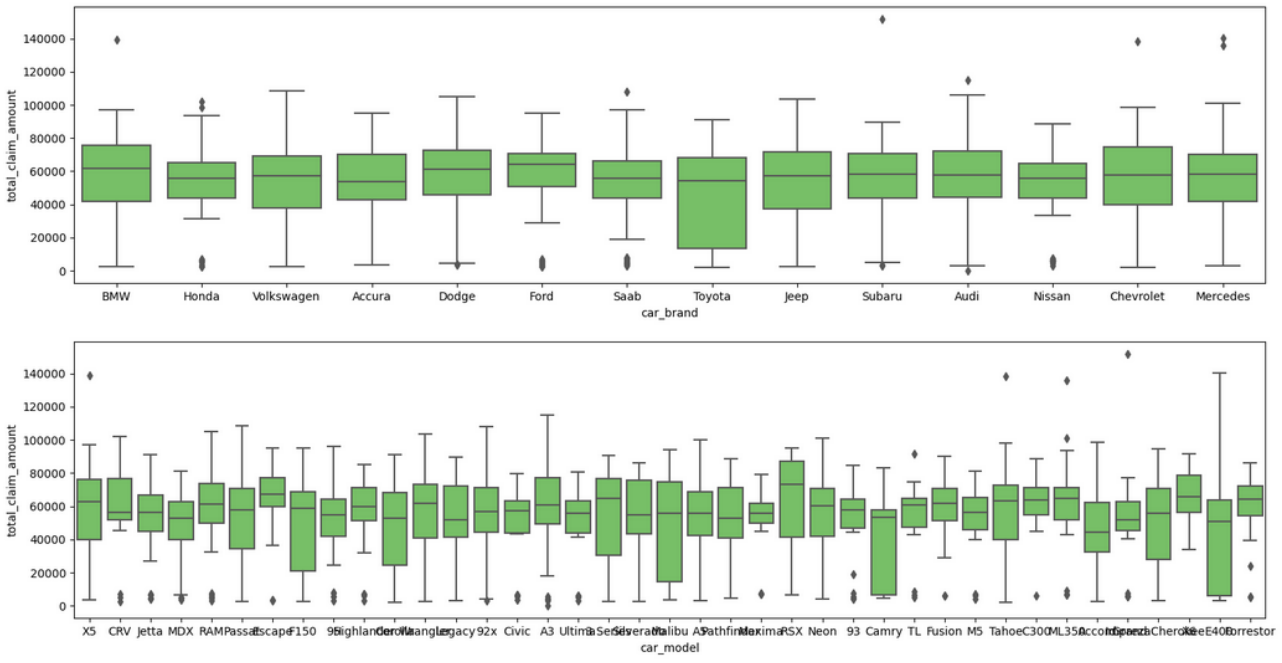

*Rysunek 12. Wykresy pudełkowe zmiennych mieszanych cz.2*

Analogiczne działania powtórzono dla grupy zmiennych jakościowych (Rysunki 13-15). Większość zmiennych nie różnicuje znacząco total_claim_amount. Natomiast ciekawe wyniki dają emg_services_notified oraz claim_type. Pomimo, że car_brand i car_model wyraźnie różnicują zbiór danych, to nie są brane pod uwagę jako zmienne do modelu, aby uniknąć redundancji – zmienna annual_prem, czyli składka roczna jest wyliczana właśnie m.in. na podstawie marki i modelu samochodu (jak również i roku produkcji).

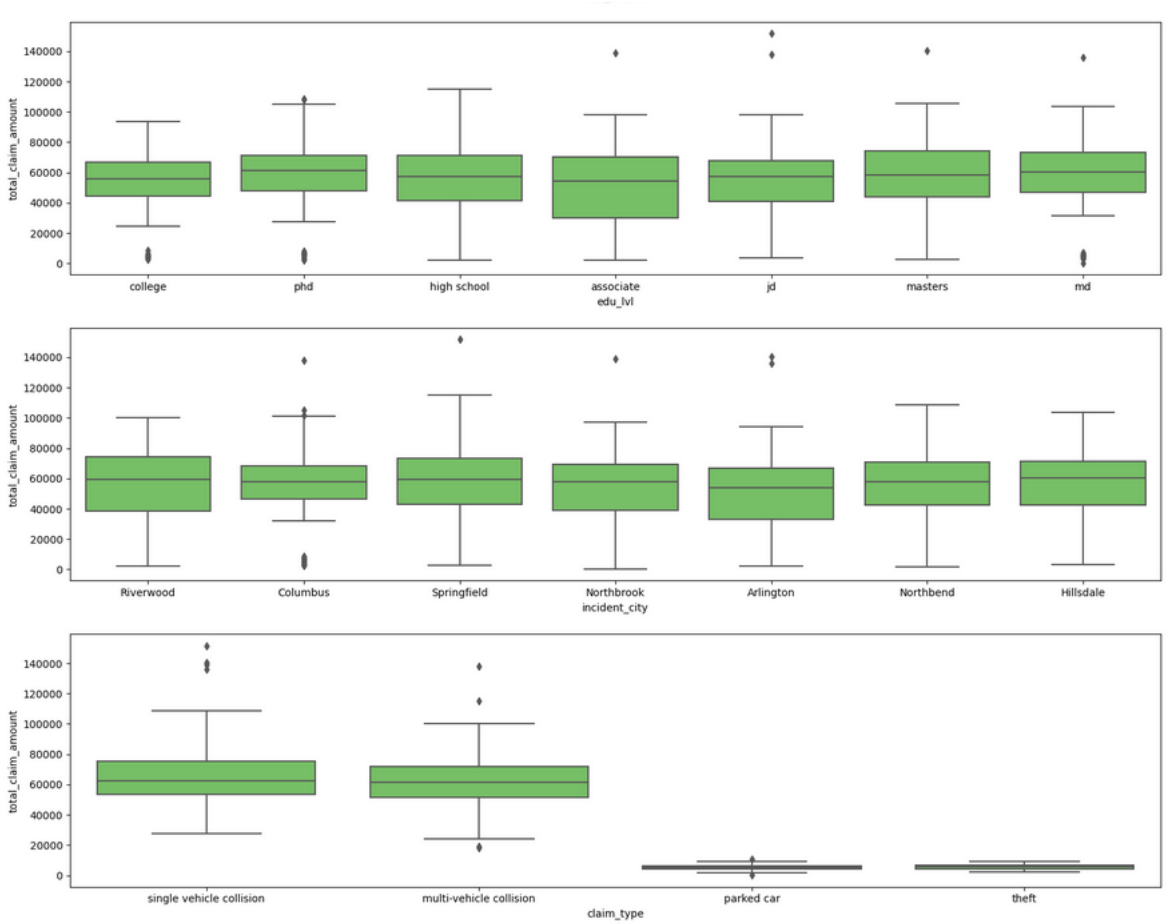
Warto również zaznaczyć, że zmienne dotyczące klienta jak na przykład płeć czy stan cywilny, nie powinny zostać użyte w modelach, niezależnie od tego jaki jest ich potencjał segmentowania klientów. Ich użycie jest dyskryminujące i nieetyczne. Jednakże w tym przypadku, nie ma takiego problemu – zmienne te nie różnicują znacznie total_claim_amount.




*Rysunek 13. Wykresy pudełkowe zmiennych jakościowych cz.1*




*Rysunek 14. Wykresy pudełkowe zmiennych jakościowych cz.2*




*Rysunek 15. Wykresy pudełkowe zmiennych jakościowych cz.3*

## 3. Modyfikacja i tworzenie zmiennych

W tym rozdziale zostały wymienione i opisane wszystkie zmodyfikowane i nowo utworzone zmienne. Na potrzeby dalszej analizy (m.in. sprawdzenia potencjalnej przydatności budowie modeli) utworzono następujące zmienne:

- 1_vehicle_involved, również zmienna binarna, przyjmuje wartość 1, jeśli w wypadku uczestniczył tylko 1 pojazd (zwykle niższe szkody), a 0 w pozostałych przypadkach;
- is_accident – zmienna binarna przyjmująca wartość 1, jeżeli zdarzenie było wypadkiem, 0 w przeciwnym przypadku;
- emg_services_notified – przekodowana zmienna o tej samej nazwie, kategorie „Fire”, „Ambulance” oraz „Other” zgrupowano w jedną kategorię ze względu na porównywalny poziom szkód. Kategorii „None” oraz „Police” nie zmieniano;
- months_between – liczba miesięcy pomiędzy rozpoczęciem polisy a wypadkiem;
- coverage_start_year – rok rozpoczęcia ochrony ubezpieczeniowej;

- is_not_first_vehicle – zmienna binarna przyjmująca wartość 1, jeżeli rok produkcji jest późniejszy niż rok powstania polisy, 0 w przeciwnym przypadku;
- vehicle_age_start – wiek samochodu w momencie rozpoczęcia ochrony ubezpieczeniowej; samochody wyprodukowane po rozpoczęciu ochrony mają przypisaną wartość 0;
- vehicle_age_acc – wiek samochodu w momencie wypadku;
- cust_age_acc – wiek klienta w momencie zdarzenia;
- police_report_NAN – przekodowana zmienna police_report, „?" zastąpiono „NAN";
- acc_type_NAN – przekodowana zmienna acc_type, jeśli zdarzenie danych nie było wypadkiem, dodano kategorię „NAN";
- injury_claim_proc, property_claim_proc, vehicle_claim_proc - zmienne określające udział zmiennych injury_claim, property_claim i vehicle_claim w zmiennej total_claim_amount;
- incident_time_39 – czy wypadek wydarzył się w godzinach 3-9, zmienna binarna;
- incident_day_of_week – dzień tygodnia zdarzenia.

Na podstawie wyników analizy eksploracyjnej oraz analizy eksperckiej do modelu k-means zdecydowano się na użycie następujących zmiennych:

- cust_age;
- months_between;
- incident_hour;
- total_claim_amount;
- injury_claim.

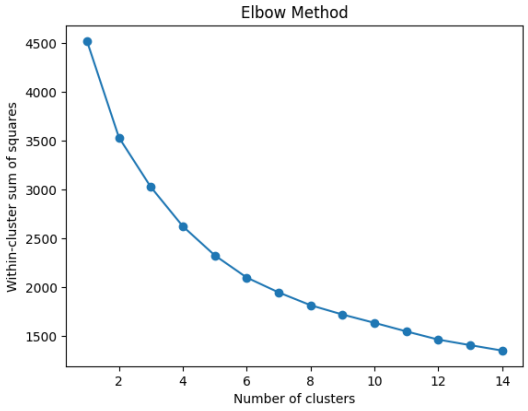
Kryteriami doboru zmiennych była wcześniej wspomniana analiza eksploracyjna i ekspercka. Dodatkowo zdecydowano się na nieuwzględnianie zmiennych binarnych. Pomimo ich przekodowania przy pomocy One Hot Encoder i pierwotnego planu użycia w modelu, pozostawienie ich spowodowało to trudności w budowie i interpretacji. Usunięto również obserwacje odstające i przeskalowano zmienne.

## 4. Model segmentacyjny

Metoda k- średnich jest metodą wyodrębniania grup (klastrów) obserwacji podobnych do siebie pod względem określonych cech (zmiennych). Metoda ta, zaliczająca się do metod *unsupervised learning,* iteracyjnie przesuwa centra klastrów tak, aby klastry stworzone wokół tych punktów charakteryzowały się jak najmniejszą wariancją.

Jako że algorytm ten polega na wyznaczaniu skupisk na podstawie obliczonych odległości, do zadania użyto jedynie zmiennych numerycznych przedstawionych już wcześniej, a więc wieku, czasu zdarzenia od podpisanej polisy, godziny zdarzenia, całkowitej kwoty ubezpieczenia oraz kwoty wartości szkody na osobie. Dzięki takiemu podziałowi, wybrano już takie zmienne, które niosą dla analizy największą wartość informacyjną.

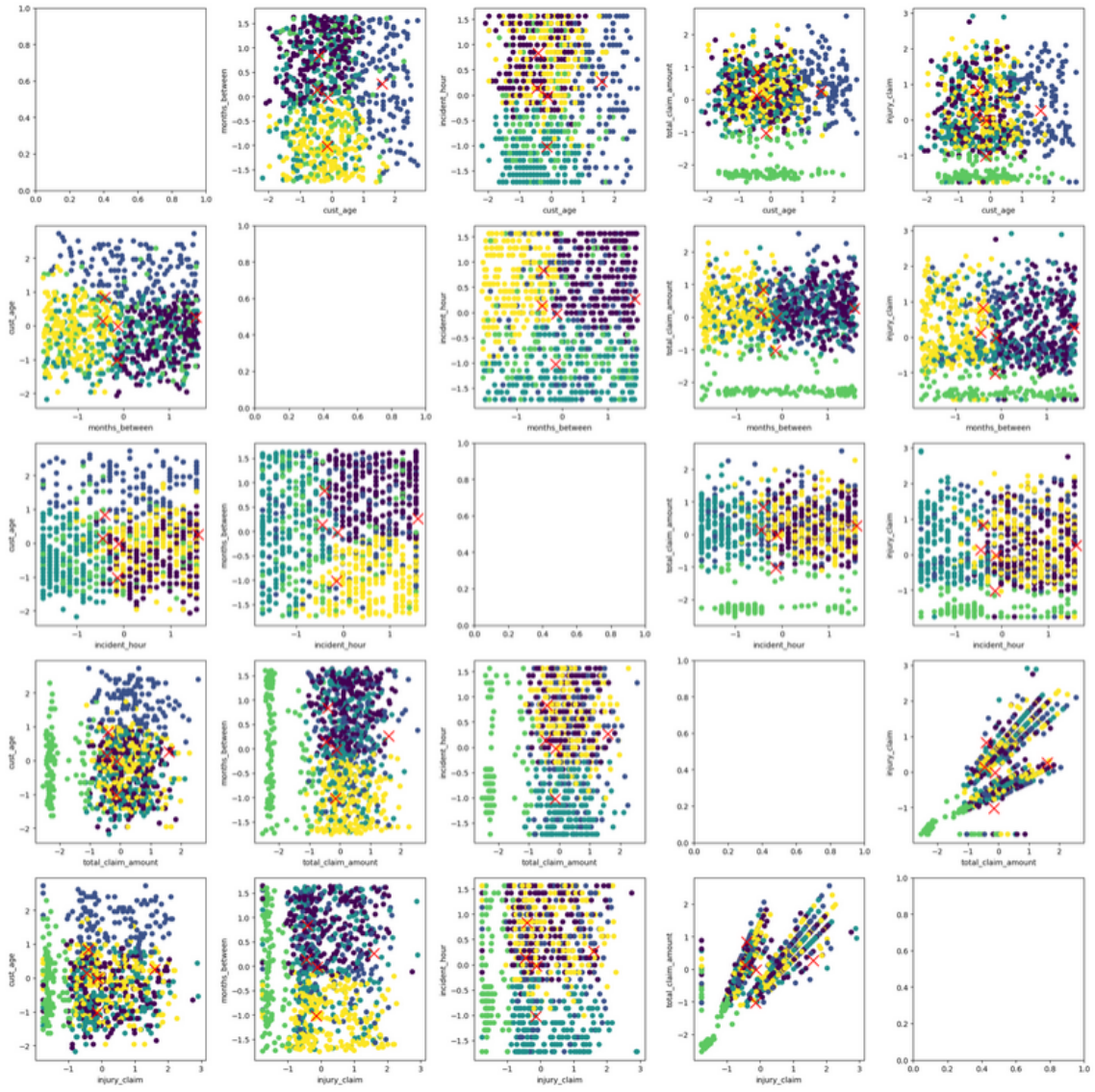
Po wybraniu zmiennych postanowiono wybrać liczbę klastrów. Do tego celu posłużono się wykresem osypiska, przedstawionym poniżej.




*Rysunek 16. Wykres osypiska*

Można zauważyć, że nie ma jednego jednoznacznego punktu, który dzieliłby wykres. To znaczy, nie ma jednoznacznie punktów, które można podzielić na takie z wysokim spadkiem i niskim spadkiem. Najbliżej taki punkt można wskazać dla liczby klastrów równej 3,4 lub 5. Aby doprecyzować liczbę klastrów, zdecydowano się skorzystać z automatycznej metody wbudowanej w pakiet w programie Python: KneeLocator. Metoda ta wskazała na optymalną liczbę klastrów równą 5 i od takiej też liczby klastrów rozpoczęto analizę.

Jako środki klastrów wybrano za pomocą wbudowanej w pakiet sklearn metody ‘k-means++’. Metoda ta, poprzez wybór na podstawie empirycznego rozkładu prawdopodobieństwa przyspiesza zbieżność, w porównaniu do powszechnie używanej metody losowego wyboru punktów.




*Rysunek 17. Kombinacje zmiennych z oznaczeniem kolorystycznym klastrów (5)*

Rysunek 17 przedstawia kombinacje wszystkich wybranych zmiennych ze sobą wraz z oznaczeniem kolorystycznym klastrów. Ze względu na wielowymiarowość analizy i brak możliwości pokazania wyników na jednym wspólnym obrazku, uznano, iż jest to najlepszy sposób szybkiej oceny wizualnej wybranej metody i ilości klastrów. Środki klastrów wyróżniono jako czerwone znaczniki. Jak widać, z powodu dużej ilości zmiennych, a także dużej ilości klastrów, interpretacja jest utrudniona. Punkty co prawda są podzielone na klastry, jednak mieszają się one ze sobą, powodując, że nie da się znaleźć sensownej interpretacji i różnicowania tych grup. Środki klastrów niejednokrotnie są bardzo blisko siebie, co dodatkowo utrudnia badanie związków.

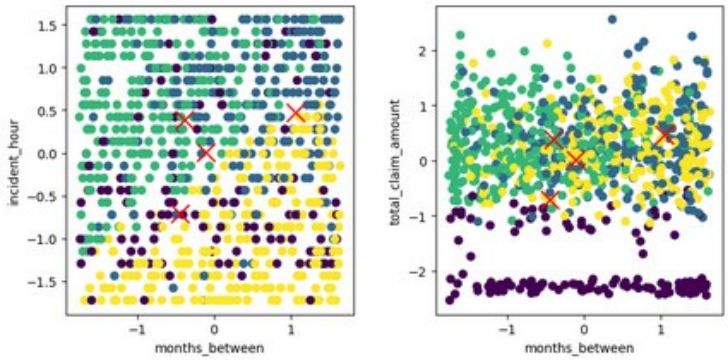
Biorąc pod uwagę powyższe, zdecydowano się na dodatkową analizę inercji oraz wypróbowanie jeszcze dwóch konfiguracji: 3 i 4 klastrów. Pomimo tego, że metoda KneeLocator wskazała 5 klastrów, to potraktowano ją jako przewodnik i punkt startowy, a nie ostatecznie wybraną opcję.

| **Liczba klastrów** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Inercja** | 4520 | 3532.95 | 3024.24 | 2623.01 | 2325.60 | 2098.74 |

*Rysunek 18. Wartość inrecji w zależności od liczby klastrów*

Jak widać, w przypadku dwóch początkowych wartości (2 i 3 klastry) spadek jest znaczący. Następnie z 3 do 4 i 4 do 5 klastrów występuje podobna procentowa wartość spadku, a tempo spadku inercji zaczyna maleć i między pięcioma i sześcioma klastrami nie ma już tak dużej różnicy. Wartości te stanowią potwierdzenie analizy wizualnej i nie wykluczają istotności trzech lub czterech klastrów.

Po szczegółowej analizie wykresów dla trzech i czterech klastrów, ustalono, że najlepiej sytuację oddają 4 klastry. Trzy klastry pozwalały na podział na różne grupy, ale niektóre z tych grup traciły na swoim pierwotnym znaczeniu, ponieważ konieczne było dostosowanie niektórych obserwacji do klastrów, które wcześniej były rozpoznawane jako oddzielne. W rezultacie, pewne zmienne mogły stracić swoją istotność w wyodrębnieniu klastrów, ponieważ posłużyły do przypisania obserwacji w bardziej ogólny sposób. Z kolei cztery klastry klarownie wyodrębniły grupy. Pomimo że cześć punktów w przypadku niektórych zmiennych jest zlokalizowana we wspólnej chmurze, to już w przypadku innej zmiennej następuje wyraźnie rozróżnienie. Oznacza to, że obserwacje są zróżnicowane wielowymiarowo niekoniecznie jednocześnie przez jedną zmienną.




*Rysunek 19. Kombinacje zmiennych z oznaczeniem kolorystycznym klastrów (4)*

Taka sytuacja jest przedstawiona na Rysunku 18., gdzie kolor fioletowy na pierwszej części nie występuje jako oddzielna chmura punktów, ale jest rozproszony po całym polu. Z kolei w przypadku zmiennej total_claim_amount jest on ulokowany typowo w dolnej części wykresu podczas gdy pozostałe zmienne są bardziej zbite w chmurę.

## 5. Analiza biznesowa

Gdy wyodrębniono już finalny model z odpowiednią liczbą klastrów dokonano ich analizy pod względem wykorzystanych zmiennych numerycznych, a także sprawdzono jak w tych grupach rozkładają się zmienne jakościowe.

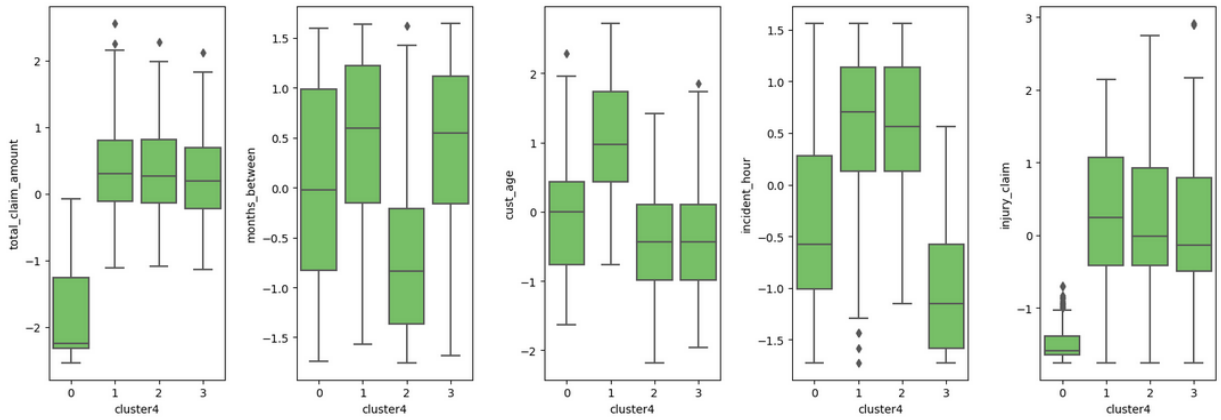
Początkowo jednak każdej obserwacji przypisano odpowiedni klaster oraz sprawdzono ich ilość. Charakteryzacja klastrów znajduje się w tabeli poniżej.

| Klaster | n | Średnia-cust_age | Średnia-months_between | Średnia-incident_hour | Średnia-total_claim_amount | Średnia-injury_claim |
|---|---|---|---|---|---|---|
| 0 | 128 | -0.10708 | 0.009905 | -0.361451 | -1.890799 | -1.464616 |
| 1 | 229 | 1.057647 | 0.469204 | 0.569260 | 0.357253 | 0.348425 |
| 2 | 292 | -0.444697 | -0.708385 | 0.589176 | 0.334598 | 0.227115 |
| 3 | 255 | -0.386836 | 0.384834 | -1.004448 | 0.245132 | 0.162212 |

*Rysunek 20. Liczba obserwacji oraz średnia wartość zmiennych w poszczególnych klastrach*

Z racji tego, że zmienne zostały już wcześniej przekształcone, można interpretować średnie jako środki klastrów, a więc także jako przeciętne wartości dla danego klastra. W ten sposób łatwo można stwierdzić, że klaster pierwszy charakteryzuje się wyższym wiekiem w stosunku do ogółu, natomiast pozostałe klastry stosunkowo niższym, z których najmłodszy jest klaster drugi. W przypadku incident_hour zmienna blisko wartości 0 oznacza kierowcę prowadzącego mniej więcej w południe natomiast pozostałe wartości to skrajny ranek (po północy) lub wieczór (przed północą). Oznacza to, że grupa z klastra trzeciego prowadziła pojazd po północy natomiast klaster zerowy już nad ranem, ale prawdopodobnie po wschodzie słońca. W przypadku pozostałych klastrów zdarzenie nastąpiło po południu lub wieczorem.

Aby dodatkowo łatwiej interpretować grupy, wykonano wykresy pudełkowe dla poszczególnych zmiennych, pokazane na Rysunku 19.




*Rysunek 21. Wykresy pudełkowe zmiennych w zależności od klastra*

Dzięki temu, od razu w sposób wizualny możemy określić który klaster wyróżnia się pod względem której zmiennej. Klaster zerowy charakteryzuje się przede wszystkim niską wartością szkody na osobie i całkowitą. Klaster drugi z kolei to osoby, u których minęło niewiele czasu pomiędzy podpisaniem polisy a zdarzeniem. Wykresy trzeci i czwarty uszczegółowiają to, co było już wskazane w poprzednim punkcie. Dodatkową pomocą w opisie klastrów na bazie wykresów pudełkowych mogą okazać się statystyki opisowe. Zostały one jednak pominięte na tym etapie.

Dodatkową wartością jest opis danych klastrów poprzez zmienne jakościowe, które nie brały udziału w analizie i dzięki temu opisanie jakie typy osób należą częściej do danych klastrów. W tabelach poniżej pokazano tylko te zmienne jakościowe, które wydają się wyróżniać powstałe grupy. Całość dostępna jest w pełnym raporcie wraz z kodem programu. Podane w tabelach wartości podane są w procentowych udziałach wewnątrz klastrów.

| Numer klastra | Is accident | |
|---|---|---|
| | nie | tak |
| 0 | 67,96% | 32.03% |
| 1 | 0 | 100% |
| 2 | 0 | 100% |
| 3 | 0 | 100% |

*Rysunek 22. Procentowy udział wartości zmiennej is_accident w poszczególnych klastrach*

Do ciekawych wniosków można dojść zestawiając ją z innymi zmiennymi: 1_vehicle_involved oraz emg_services_notified.

| Numer klastra | 1_vehicle_involved | |
|---|---|---|
| | Więcej niż 1 pojazd | 1 pojazd |
| 0 | 19,53% | 80,47% |
| 1 | 52,40% | 47,60% |
| 2 | 51,71% | 48,29% |
| 3 | 47,84% | 52,16% |

*Rysunek 23. Procentowy udział wartości zmiennej 1_vehicle_involved w poszczególnych klastrach*

Wyraźnie widać, że segment pierwszy wybija się tym, że w większości nie są to wypadki, dotyczą jednego pojazdu i nie powiadamia się innych służb (ale powiadamia się policję). W przypadku pozostałych, tak niska wartość powiadamiania policji prawdopodobnie wynika z konieczności powiadamiania innych służb, chociażby wezwania karetki.

| Numer klastra | emg_services_notified | |
|---|---|---|
| | policja | inne |
| 0 | 78,91% | 21,09% |
| 1 | 25,33% | 74,67% |
| 2 | 23,29% | 76,71% |
| 3 | 25,10% | 74,90% |

*Rysunek 24. Procentowy udział wartości zmiennej emg_services_notified w poszczególnych klastrach*

Aby podsumować, można przygotować krótką charakterystykę każdego klastra:

Klaster 0 – Jest to najmniej liczna grupa, ale bardzo charakterystyczna poprzez kradzieże lub zdarzenia związane z zaparkowanym autem charakteryzujące się raczej niską całkowitą kwotą szkody i szkody na osobie, do której nie ma konieczności wzywania innych służb niż policja. Zdarzenia te raczej nie charakteryzują kierowców, gdyż są wydarzeniami losowymi o czym świadczy bardzo szeroki wykres pudełkowy przy zmiennej mówiącej jak dużo czasu minęło od podpisania polisy do zdarzenia i zmiennej dotyczącej godziny zdarzenia (z naciskiem na poranek, gdy już jest jasno), a także raczej przeciętny wiek kierowcy wraz z dość szerokimi wąsami na wykresie pudełkowym. Robocza nazwa segmentu: “zdarzenia losowe nie-wypadki”.

Klaster 1 – najmniej liczny z pozostałych trzech klastrów charakteryzujących wypadki. W tym przypadku rolę odgrywa wiek kierowcy, który jest znacznie wyższy od pozostałych. W tej grupie częściej nikt nie ulega szkodzie na osobie (41% versus około 30% w innych klastrach – zmienna bodily_injuries), a od podpisania polisy minęło już dużo czasu. Kwota szkody jest przeciętna, być może minimalnie wyższa dla tego klastra. Wypadek miał miejsce po południu lub wieczorem. Robocza nazwa segmentu: “wypadki ostrożnych dojrzałych kierowców”.

Klaster 2 – klaster wyróżnia się spośród pozostałych tym, że minęło zdecydowanie mniej czasu od podpisania polisy do zdarzenia niż w przypadku innych segmentów, wiek jest przeciętny (lecz dolny wąs wskazuje, że częściej mogą to być trochę młodsi kierowcy), tak samo jak kwota polisy. Można powiedzieć, że kierowcy w tej grupie są, pod względem zachowania, przeciwieństwem kierowców klastra numer 1 a jest to najliczniejszy spośród klastrów charakteryzujących wypadki. Robocza nazwa segmentu: “wypadki ryzykownych roztargnionych kierowców”.

Klaster 3 – ostatni klaster to grupa kierowców o raczej przeciętnym wieku, jeżdżących w nocy (mniej więcej od północy do czasu aż wzejdzie słońce). Raczej nic więcej nie jest dla nich charakterystyczne. Robocza nazwa segmentu: “wypadki nocnych kierowców”.

## 6. Wykrywanie anomalii

W celu wykrycia nietypowych zgłoszeń mogących wskazywać na próbę wyłudzenia odszkodowania, wykorzystano algorytm *IsolationForest.* Jest to to algorytm uczenia nienadzorowanego, który specjalizuje się w wykrywaniu anomalii poprzez izolowanie nietypowych punktów danych. Konstruuje wiele drzew decyzyjnych, z których każde drzewo jest rozwijane poprzez losowe wybieranie zmiennej, a następnie losowe wybieranie wartości jej podziału. Anomalie charakteryzują się tym, że występują rzadko i różnią się od typowych obserwacji. Zwykle izolowane są zatem blisko korzenia drzewa. Algorytm określa *score* anomalii na podstawie długości ścieżek w drzewach. Krótsze ścieżki wskazują na wyższe prawdopodobieństwo anomalii. Jego skuteczność jest zauważalna w obsłudze danych wielowymiarowych, dzięki stosunkowo płytkiej naturze drzew.

Do trenowania modelu wykorzystano zmienne ilościowe i jedną zmienną binarną:

- total_claim_amount,
- annual_prem,
- vehicle_age_acc,
- is_not_first_vehicle,
- vehicle_claim_proc,
- injury_claim_proc.

Zmienne te zostały wykorzystane ze względu na ich charakterystykę, potencjalny wpływ na skłonność do prób nieuczciwego uzyskania odszkodowania.

W całym zbiorze wyodrębniono 5 obserwacji wykazujących się odmienną charakterystyką, podejrzanych o bycie próbą oszustwa. Były to obserwacje o następujących charakterystykach:

| | policy_id | total_claim_amount | annual_prem | vehicle_age_acc | is_not_first_vehicle | vehicle_claim_proc | injury_claim_proc |
|---|---|---|---|---|---|---|---|
| **55** | 211077 | 43280 | 1925.98 | 20 | 1 | 0.875000 | 0.0 |
| **88** | 73573 | 37280 | 1273.50 | 20 | 0 | 1.000000 | 0.0 |
| **95** | 430152 | 4400 | 1006.90 | 16 | 0 | 0.875000 | 0.0 |
| **394** | 208883 | 33930 | 833.13 | 18 | 1 | 0.888889 | 0.0 |
| **749** | 383388 | 39690 | 1096.10 | 18 | 0 | 1.000000 | 0.0 |

Są to obserwacje w większości poniżej 1. kwartyla całkowitej wartości szkody oraz procentowego udziału wartości szkody na osobie w całkowitej wartości szkody, a także powyżej 3. kwartyla zmiennej określającej wiek pojazdu w momencie zdarzenia i udział szkody związanej z uszkodzeniem pojazdu w całkowitej wartości szkody. Obserwacje są zróżnicowane pod względem wysokości składki rocznej oraz tego, czy pojazd, którego dotyczy zdarzenie, był pierwszym zarejestrowanym na daną polisę. Można zatem podejrzewać, że są to pojazdy na tyle stare, że klientom firmy ubezpieczeniowej bardziej opłacało się uzyskać odszkodowanie, potencjalnie umyślnie wywołując szkodę, niż sprzedać pojazd, nawet pomimo tego, że szkody te nie są wyjątkowo wysokie na tle wszystkich obserwacji. Można też zauważyć, że zdarzenia te dotyczą głównie uszkodzenia pojazdu, a nie ucierpiały w nich żadne osoby. To również może wskazywać na celowe uszkodzenie przy zachowaniu ostrożności, aby nikomu nie stała się krzywda.

## 7. Podsumowanie

Projekt miał na celu segmentację ubezpieczonych oraz identyfikację czynników ryzyka związanych ze zgłaszanymi szkodami. Przeprowadzono analizę eksploracyjną danych, w tym rozkłady zmiennych ilościowych, mieszanych i jakościowych. Przeprowadzono modyfikację i tworzenie nowych zmiennych. Następnie zbudowano model segmentacyjny k-średnich, wybierając istotne zmienne na podstawie analizy eksploracyjnej. W ostatniej części zidentyfikowano anomalie.

Eksploracyjna analiza danych obejmowała poznanie rozkładu zmiennych oraz ich relacji. Zastosowano różne metody analizy danych, w tym obliczenie podstawowych miar statystycznych, histogramów i wykresów pudełkowych. Wybrano odpowiednie zmienne do modelu segmentacyjnego, uwzględniając analizę eksploracyjną i ekspercką.

Model k-średnich został dostosowany do czterech klastrów. Analiza klastrów wykazała różnice w wieku kierowców, czasie od podpisania polisy do zdarzenia oraz kwotach szkód. Opracowano również charakterystyki dla każdego klastra, identyfikując specyficzne cechy dla poszczególnych segmentów.

W dalszej części raportu skonstruowano algorytm IsolationForest w celu wykrywania anomalii w zgłoszeniach, które mogą wskazywać na próbę oszustwa. Uwzględniono zmienne ilościowe oraz jedną zmienną binarną. Zidentyfikowano pięć obserwacji uznanych za podejrzane o oszustwo, ze względu na ich nietypową charakterystykę.

Podsumowując, projekt skoncentrował się na segmentacji ubezpieczonych, analizie ryzyka, tworzeniu modelu segmentacyjnego k-średnich oraz wykrywaniu anomalii w zgłoszeniach. Wnioski z analizy danych mogą być użyteczne w doskonaleniu procesów ubezpieczeniowych i minimalizacji ryzyka dla firm ubezpieczeniowych.