INSTITUTO UNIVERSITÁRIO DE LISBOA

Análise Não Supervisionada de Incidentes Rodoviários

Identificação de Dimensões Latentes e Padrões de Heterogeneidade

RELATÓRIO DO PROJETO

Relatório técnico referente ao Projeto da UC de Métodos de Aprendizagem Não Supervisionada

TRABALHO REALIZADO POR:

DANIEL FONSECA | 125158 | CD-PL

FRANCISCO GONÇALVES | 130649 | CD-PL

JOÃO FILIPE | 130665 | CD-PL

GUILHERME PIRES | 131658 | CD-PL

Docentes:

RICARDO NOUTEL DE MATOS CORREIA
MAFALDA PONTE

25/26

Índice

Introdução

O presente relatório foi desenvolvido no âmbito da Unidade Curricular de **Métodos de Aprendizagem Não Supervisionada** (MANS) e tem como objetivo aplicar, de forma prática, os conteúdos lecionados através da análise de uma base de dados de **incidentes rodoviários**.

Neste contexto, os incidentes rodoviários têm características diferentes entre si, tanto no que diz respeito às pessoas envolvidas como às circunstâncias em que ocorrem e às consequências que daí resultam. Esta diversidade torna a análise deste tipo de dados mais complexa, sobretudo quando se pretende ir além de uma simples descrição e perceber se existem **padrões** ou tipos de incidentes com características semelhantes.

Neste enquadramento, é adotada uma abordagem de **aprendizagem não supervisionada**, tal como explorado ao longo da unidade curricular. Uma vez que não existe uma variável-alvo definida, o objetivo passa por analisar os dados e identificar **padrões** e semelhanças entre os incidentes, sem impor classificações à partida.

Para isso, numa primeira fase, aplica-se uma **Análise de Componentes Principais** (PCA) às variáveis **ativas** da base de dados, permitindo reduzir a dimensionalidade e sintetizar a informação mais relevante. De seguida, com base nas componentes principais obtidas, realiza-se uma análise de **clustering**, com o objetivo de agrupar os incidentes de acordo com as suas semelhanças.

Os grupos identificados são posteriormente caracterizados através das variáveis de **perfil**, o que permite interpretar os clusters obtidos e distinguir diferentes tipologias de sinistros.

Assim, o trabalho procura organizar e interpretar a informação disponível, permitindo perceber melhor a diversidade existente nos incidentes rodoviários.

Dados

A base de dados analisada reporta informações sobre sinistros automóveis, agregando dados contratuais, características demográficas dos segurados e detalhes específicos sobre os acidentes participados.

A nível de limpeza, substituiu-se os valores "?" por "NA" para se fazer a contagem de valores "em falta", através do comando *sum(is.na(dadosg8)*. Verificou-se que apenas as colunas de texto continham "NA", portanto não se realizou nenhuma limpeza, dado que o PCA foi desenhado para funcionar em colunas numéricas. Também se excluiu variáveis redundantes para evitar multicolinearidade perfeita (como a soma total dos custos).

Assim, a dimensão final da amostra utilizada no estudo é constituída por **907 observações** (clientes) e 24 variáveis.

Para a construção dos eixos do PCA foram consideradas para **variáveis ativas** (INPUT) as variáveis **numéricas**:

- **Financeiras/Sinistro:** vehicle_claim, property_claim, injury_claim, number_of_vehicles_involved, incident_hour_of_the_day, bodily_injuries, witnesses.
- **Demográficas:** age, months_as_customer.
- **Veículo:** auto_year.

As restantes variáveis foram tratadas como **variáveis passivas** (PROFILE), utilizadas exclusivamente para a caracterização dos grupos identificados:

- total_claim_amount (*Tornou-se passiva porque a retirámos para evitar redundância, mas continua a ser útil para análise descritiva*)
- incident_date
- insured_sex (Género)
- insured_education_level
- insured_occupation
- insured_hobbies
- insured_relationship
- incident_type
- collision_type
- incident_severity
- authorities_contacted
- property_damage
- police_report_available
- auto_make

Através desta distinção, é possível - através das **variáveis ativas** - colocar os incidentes numa certa posição de um gráfico. Futuramente - através das **variáveis passivas** (categóricas) - sabemos que estes casos têm algo em comum: exemplos em que a causa era “Parked Car” ou “Vehicle Theft”, havendo assim a possibilidade de criar um cluster.

Alguns aspetos que tivemos em conta, previamente à aplicação do PCA, foram:

- A escala dos valores estava dispersa, por exemplo, o desvio padrão da variável “age” era 9, mas do “total_claim_amount” é 26,4. No entanto, a função *principal()* , da library psych, já contabiliza com esta situação automaticamente, dado que usa correlações.
- A assimetria (Skew) tinha valores entre -0.6 e +0.5 o que indica que não existem variáveis extremamente distorcidas.

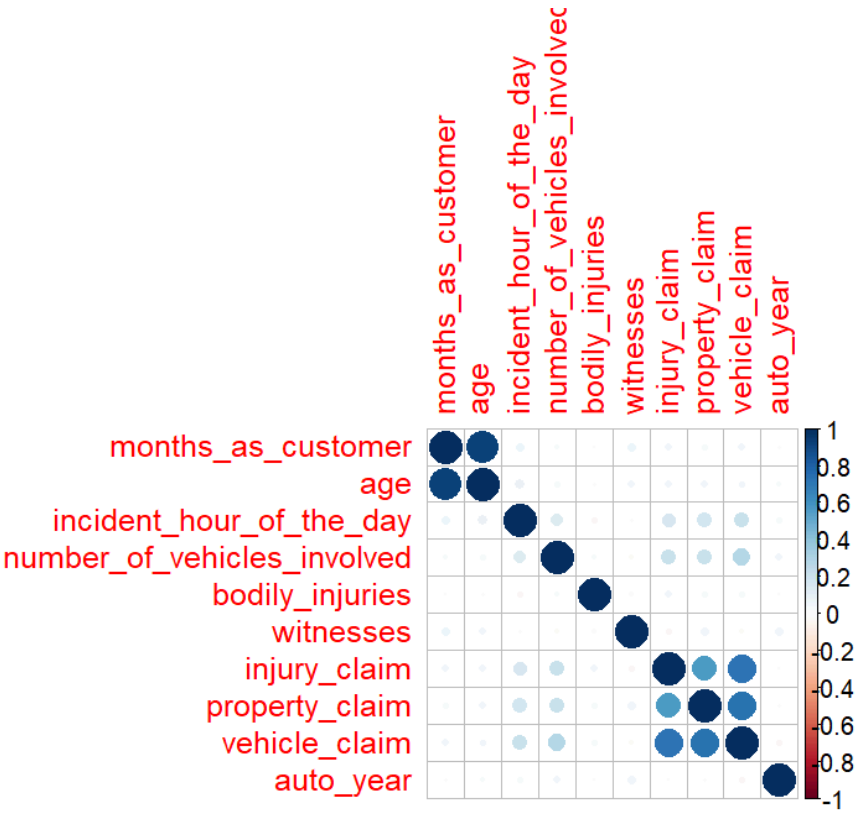

Figura 1: Gráfico de correlação entre as variáveis INPUT

A matriz de correlações (Figura 1) revela a existência de correlações positivas fortes (cor azul) dentro de grupos específicos de variáveis e uma ausência de correlações negativas (cor vermelha).

Identificámos **dois grupos**:

- Um demográfico, evidenciado pela forte correlação entre a idade do segurado (age) e a sua antiguidade (months_as_customer).
- E um de sinistralidade, onde as variáveis de custo (vehicle_claim, property_claim, injury_claim) apresentam fortes correlações entre si, indicando que a severidade de um acidente tende a refletir-se em todas as componentes de custo.

É igualmente relevante notar a fraca correlação (áreas claras/brancas) entre estes dois grupos. Isto sugere que a maturidade do cliente não nos permite, de forma linear, prever o custo do sinistro.

Identificação das dimensões da análise

De forma a sabermos quantos componentes teria o PCA, recorremos ao critério do cotovelo (Scree Plot) e à análise paralela, de forma a averiguar o que é significância real vs acaso.

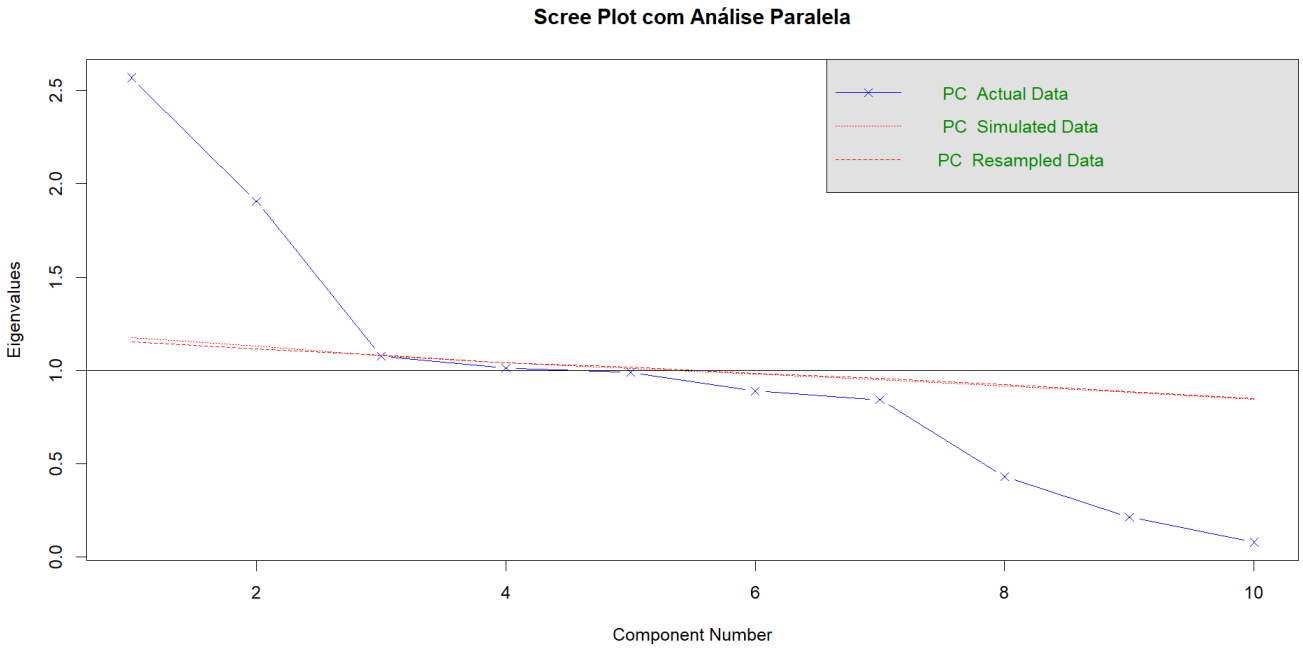

Gráfico 1: Scree Plot (Com Análise Paralela)

A análise paralela cria uma base de dados do mesmo tamanho que a que estamos a utilizar e permite-nos averiguar se "A minha Componente 1 real explica mais variância do que uma Componente criada a partir de dados aleatórios?"

- **Se SIM:** A componente é real (é um padrão, não é coincidência). **Mantém-se**
- **Se NÃO:** A componente é apenas "ruído". **Descarta-se.**

Desta forma, não usámos apenas o critério básico de usar Componentes Principais com Eigenvalue > 1 (Gráfico 1). Usámos, então, o sugerido pelo algoritmo: usar **2 componentes principais** (Figura 2).

```
Console   Terminal ×   Background Jobs ×
R 4.5.1 · C:/Users/danie/Desktop/MANS/
> resultado_parallel <- fa.parallel(dados_pca,
+                                   fa = "pc",
+                                   ylabel = "Eigenvalues",
+                                   show.legend = TRUE,
+                                   main = "Scree Plot com Análise Paralela")
Parallel analysis suggests that the number of factors =  NA  and the number of components =  2
>
```

Figura 2: Recomendação de 2 Componentes Principais pela Análise Paralela

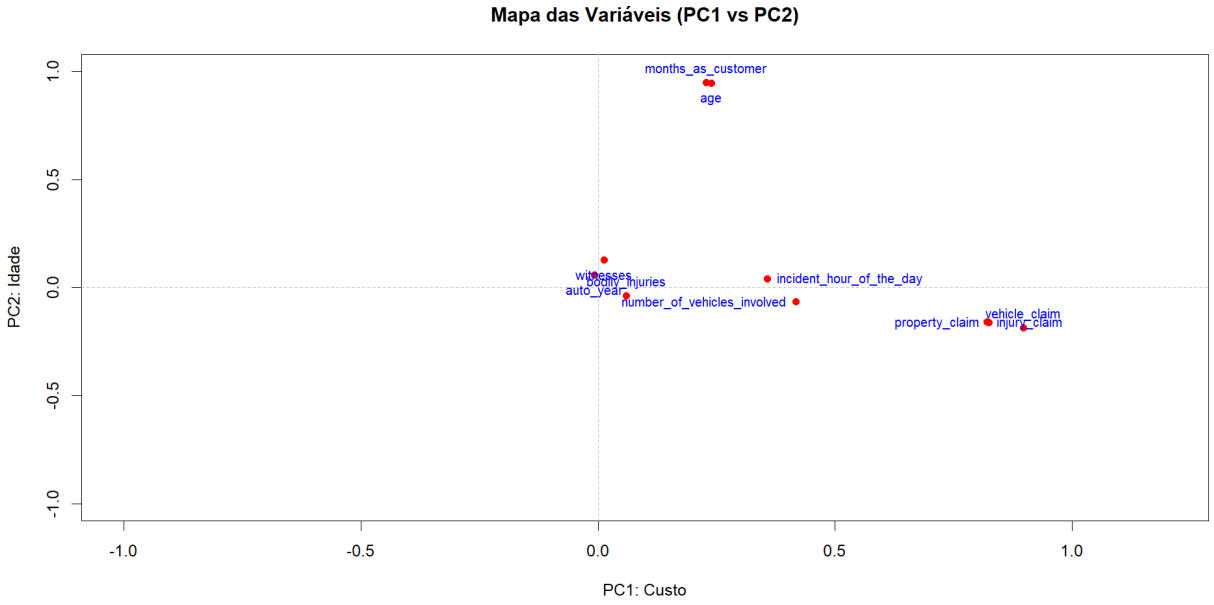

Gráfico 2: Mapa das Variáveis (PC1 vs PC2)

Através do mapa das variáveis (Gráfico 2) podemos inferir que:

- No Eixo X (PC1): As variáveis vehicle_claim, property_claim e injury_claim estão muito à direita. Logo, este eixo mede a "Severidade Financeira".
- No Eixo Y (PC2): As variáveis age e months_as_customer estão muito em cima. Logo, o Eixo Vertical mede a "Maturidade".

As **variáveis** do gráfico 2 que estão mais próximas estão **correlacionadas**, ou seja, age e months_as_customer, dizem quase a mesma coisa (quem é mais velho é cliente há mais tempo). O mesmo acontece com o grupo dos custos (vehicle_claim e property_claim).

Variáveis opostas (180 graus): Estão correlacionadas **negativamente**: Se houvesse uma variável num canto e outra no canto oposto diagonal, significaria que quando uma sobe, a outra desce. Não se evidencia este facto no gráfico, logo não há variáveis

correlacionadas negativamente, o que também fora explicado no gráfico de correlações.

Variáveis a 90 graus (ortogonais): São **independentes.** O grupo da "Idade" é ortogonal com o grupo do "Custo", o que indica que a idade do cliente e o custo do sinistro são dimensões independentes. Saber a idade não ajuda a prever o custo diretamente através de uma relação linear simples.

```
                            PC1   PC2     h2    u2 com
months_as_customer                0.95 0.9515 0.048 1.1
age                               0.95 0.9525 0.047 1.1
incident_hour_of_the_day    0.36       0.1291 0.871 1.0
number_of_vehicles_involved 0.42       0.1782 0.822 1.0
bodily_injuries                        0.0049 0.995 1.7
witnesses                              0.0163 0.984 1.0
injury_claim                0.82       0.6982 0.302 1.1
property_claim              0.82       0.7047 0.295 1.1
vehicle_claim               0.90       0.8377 0.162 1.1
auto_year                              0.0034 0.997 1.0

                       PC1  PC2
SS loadings           2.57 1.91
Proportion Var        0.26 0.19
Cumulative Var        0.26 0.45
Proportion Explained  0.57 0.43
Cumulative Proportion 0.57 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 components are sufficient.
```

Figura 3: Resultados do PCA com 2 Componentes Principais

A Componente 1 (PC1): Dimensão de Severidade Financeira, representada no eixo horizontal, explica a maior parte da variabilidade dos dados (**26%** da variabilidade).

Apresenta correlações positivas muito fortes com todas as variáveis de custo: *vehicle_claim* (0.90), *injury_claim* (0.82) e *property_claim* (0.82). Apresenta ainda uma associação moderada com a complexidade do acidente (*number_of_vehicles_involved*: 0.42).

Esta dimensão representa o **impacto económico e a gravidade do sinistro**. Valores elevados neste eixo indicam acidentes muito dispendiosos e complexos e valores baixos indicam pequenos incidentes. Assim, intitulou-se de "**Gravidade do Sinistro".**

Componente 2 (PC2): Dimensão de Maturidade e Fidelização. Esta componente vertical capturou os dados independentes do custo e explica **19%** da variabilidade dos dados. Esta dimensão é definida quase exclusivamente pelas variáveis demográficas e contratuais, apresentando correlações fortes com a idade do

segurado (*age - 0.95*) e a sua antiguidade na seguradora (*months_as_customer - 0.95*).

O PC2 representa o **ciclo de vida do cliente**. Valores elevados identificam clientes mais velhos e com longa relação com a seguradora; valores baixos identificam clientes jovens ou clientes novos. Intitulou-se o PC2 de "**Maturidade do Cliente"**.

Algumas variáveis não foram bem representadas pelas duas componentes do PCA - as com valores h2 baixo:

- bodily_injuries (h2 = 0.0049)
- witnesses (h2 = 0.0163)
- auto_year (h2 = 0.0034)

Estas variáveis foram praticamente ignoradas pelo modelo, o que significa que o número de testemunhas ou o ano do carro não estão correlacionados nem com o Custo nem com a Idade.

Embora a variância acumulada de 45% possa parecer modesta, a análise das comunalidades (h2) revela que esta percentagem se deve à elevada heterogeneidade de algumas variáveis secundárias. Enquanto as variáveis do risco (vehicle_claim, age, months_as_customer) apresentam taxas de explicação superiores a 80%, variáveis como witnesses ou auto_year (comunalidades < 0.05), não se correlacionaram com as estruturas principais.

Foi ainda testada uma solução alternativa com **três componentes principais**, o que permitiria aumentar a variância explicada acumulada para 56%.

A terceira componente (PC3) capturava as variáveis *auto_year* (0.68) e *witnesses* (0.56).

No entanto, optou-se por descartar esta opção e manter a solução de **duas componentes**, tendo em conta dois critérios:

1. A Análise Paralela indicou que a terceira componente **não** apresentava **variância** significativamente superior ao ruído aleatório.
2. A terceira dimensão **não** apresentava uma **justificação plausível**. Associar o ano do veículo a testemunhas não justifica a criação de um perfil de cliente específico para fins de segmentação.

Assim, a solução final com dois eixos (Custo e Maturidade) privilegiou a simplicidade da interpretação.

Identificação da heterogeneidade na base de dados

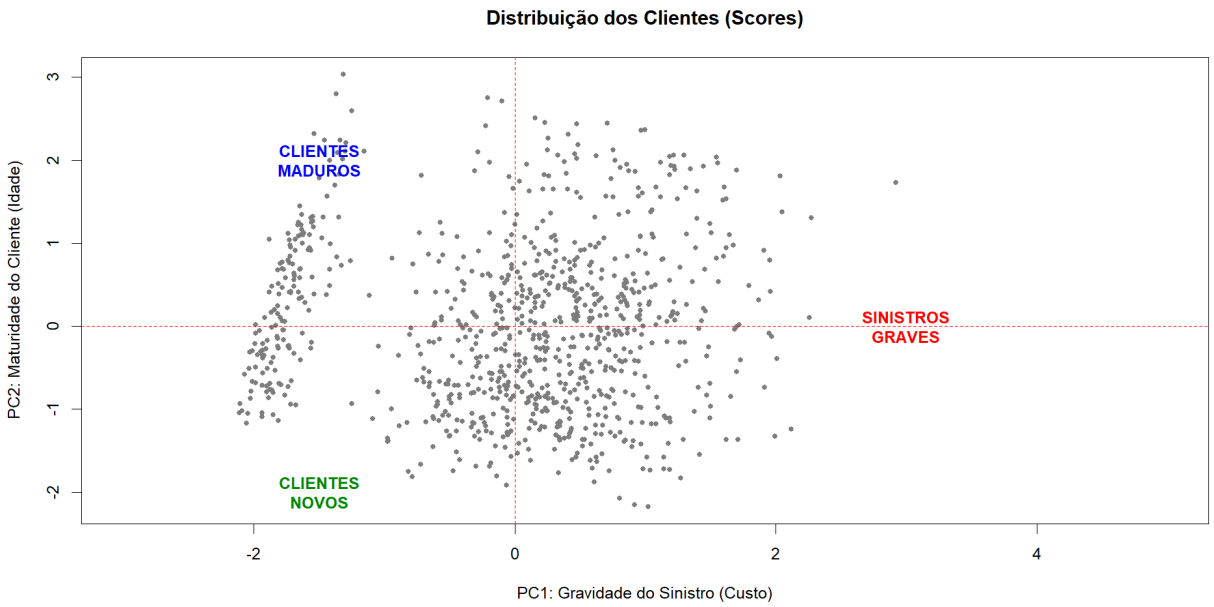

Gráfico 3: Distribuição dos clientes pré-clustering

Decidimos procurar **3 Clusters** (grupos), pois visualmente o gráfico da figura 5 parece ter três grandes zonas e também considerámos pertinente realizar uma separação entre os clientes novos e os seniores (visualmente, pelo eixo 0 do PC2).

Cluster à esquerda: Valores baixos no PC1.

Cluster à direita, em cima: Valores maiores em PC1 e altos em PC2.

Cluster à direita, em baixo: Valores maiores em PC1 e baixos em PC2.

Caracterização dos Clusters

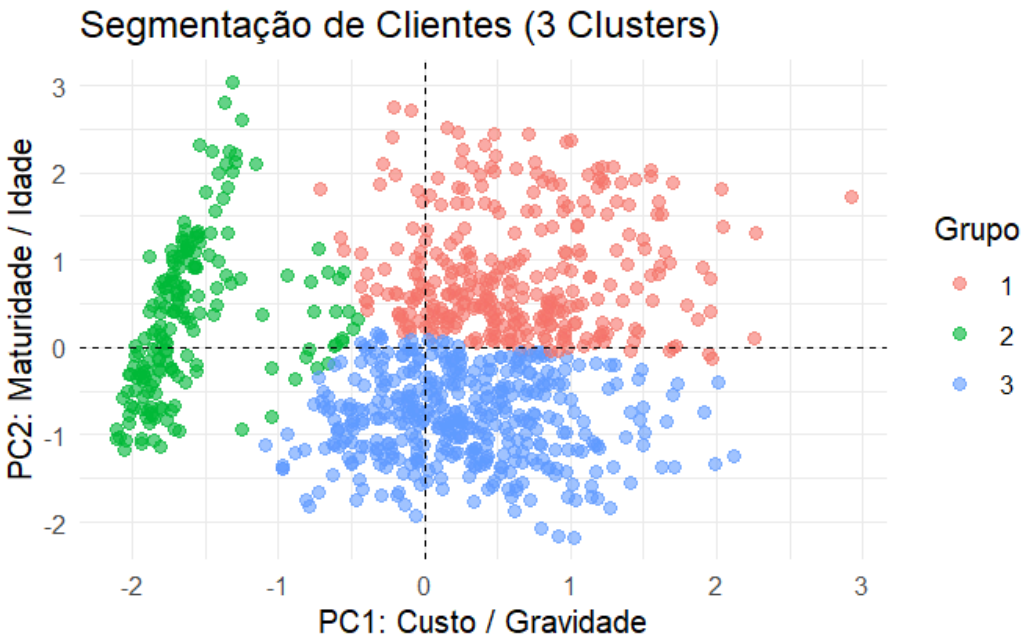

Gráfico 4: Distribuição dos clientes pós-clustering

Depois de executar a função **kmeans** atribuímos a cada **incidente rodoviário** um grupo: 1, 2 ou 3, assim, construímos o gráfico 4 em que mostra de forma ilustrativa a localização de cada um desses grupos.

Antes de proceder à análise e à nomeação de cada um dos clusters, fomos responder a três perguntas: **"Quem bateu e quem foi roubado?"**, **"Como varia a gravidade dos acidentes de grupo para grupo?"** e **"Será que um grupo tem mais homens/mulheres?"**. Também efetuámos em cada um dos grupos uma **média** de três variáveis que considerámos serem decisivas no momento da análise: **total_claim_amount**, **age** e **months_as_customer**.

cluster	total_claim_amount	age	months_as_customer
1	64263	47.5	314.2
2	9335	38.4	198.4
3	63499	32.6	122.6

Tabela 1: Médias das três variáveis em cada cluster

	Multi-vehicle Collision	Parked Car	Single Vehicle Collision	Vehicle Theft
1	207	0	201	0
2	7	75	17	88
3	163	0	149	0

Tabela 2: Tipo de Incidente por Cluster

A análise da tabela 2 revela uma forte ligação entre o tipo de incidente e o perfil de risco. Os Clusters 1 (Seniores) e 3 (Jovens) concentram exclusivamente as **colisões de trânsito** (Multi-vehicle e Single Vehicle), sendo os **grupos de alto custo**. O Cluster 2 (Incidentes Menores) isola perfeitamente os sinistros **sem condução perigosa**, agrupando 100% dos Roubos (Vehicle Theft) e 100% dos acidentes em estacionamento (Parked Car), o que explica o seu baixo custo médio.

Assim respondendo à pergunta de "**Quem bateu e quem foi roubado?**", podemos dizer que os grupos 1 e 3 foram os que bateram mais vezes e o grupo 2 o que foi roubado.

	Major Damage	Minor Damage	Total Loss	Trivial Damage
1	136	117	155	0
2	5	88	10	84
3	105	118	89	0

Tabela 3: Gravidade do Incidente por Cluster

Referente à tabela 3 observamos que os Clusters 1 e 3 demonstram a gravidade mais elevada, com custos médios de aproximadamente 64.000€ (Tabela 1).

O **Cluster 1** (Seniores) destaca-se com a Gravidade Crítica, registando o maior número de casos de "Total Loss" (Perda Total - 155 casos), sugerindo acidentes **mais destrutivos** para o veículo do que os do Cluster 3.

Em contraste, o **Cluster 2** apresenta uma gravidade e custo **muito baixos**, aproximadamente 9.000€ de custos (Tabela 1), sendo o único grupo com incidentes classificados como "Trivial Damage".

Desta forma, respondendo à pergunta de "Como varia a gravidade dos incidentes de grupo para grupo?", podemos dizer que o grupos 1 e 3 apresentam uma gravidade elevada e o grupo 2 uma gravidade baixa.

	FEMALE	MALE
1	235	173
2	103	84
3	155	157

Tabela 4: Sexo por Cluster

Verifica-se, na tabela 4, um padrão de distribuição de género notavelmente diferente entre os dois principais grupos de alto risco.

O **Cluster 1** (Seniores), demonstra uma **predominância feminina** acentuada nesta amostra (235 Mulheres vs 173 Homens).

O **Cluster 3** (Jovens), o cluster mais jovem, apresenta uma distribuição de género praticamente **equilibrada** (155 Mulheres vs 157 Homens).

O perfil do **Cluster 2** (Incidentes Menores), de baixo risco, também apresenta uma **predominância feminina** (103 Mulheres vs 84 Homens).

Estes resultados indicam padrões distintos de equilíbrio e predominância feminina nos clusters que envolvem colisões de trânsito, permitindo responder à pergunta "Será que um grupo tem mais homens/mulheres?".

Com base nas tabelas e no gráfico 4, conseguimos traçar um perfil para cada grupo.

Ao **grupo 1** denominámos "Seniores Fidelizados de alto risco". Este grupo representa os clientes com **maior maturidade** e **incidentes de elevada gravidade.** O perfil do cliente é o seguinte: mais velho, com uma média de idade de **47,5 anos** e uma relação **longa** com a seguradora (média de 314 meses ou 26 anos). Neste grupo existe uma predominância do género **feminino** (235 mulheres vs 173 homens). Os tipos de Sinistro caracterizam-se exclusivamente por **colisões** (tanto com múltiplos veículos como veículo único). Apresenta o impacto financeiro mais dispendioso, com um custo médio de sinistro de **64.263€**. Destaca-se por ter a maior frequência de "Perda Total" (155 casos), indicando acidentes com **destruição completa do veículo**.

Ao **grupo 2** demos o nome "Clientes de baixo risco". Este grupo isola os sinistros que **não envolvem condução ativa perigosa** ou grandes custos, representando o grupo de baixo valor. O cliente apresenta uma idade intermédia (média de 38,4 anos) e este grupo tem uma distribuição de **género equilibrada**. Contém **100% dos casos de**

Roubo (*Vehicle Theft*) e **100% dos acidentes com Carro Estacionado** (*Parked Car*). Não existem colisões em andamento neste grupo. É o grupo de **menor impacto financeiro**, com um custo médio drasticamente inferior em comparação com os outros grupos, **9.335€**. É o único grupo que contém incidentes classificados como "Danos Triviais".

Ao **grupo 3** denominamos **"Jovens de alto risco"** este grupo espelha a gravidade do **grupo 1**, mas num segmento demográfico muito mais jovem e recente. Tem o **perfil do cliente** mais jovem (média de **32,6 anos**) e com menor histórico na companhia (**122 meses** ou ~10 anos). A distribuição entre homens e mulheres é praticamente igual. Tal como o **grupo 1**, foca-se inteiramente em colisões (Multi-veículo e Veículo Único). O **impacto financeiro** apresenta uma **severidade financeira muito alta**, quase idêntica à dos seniores (**63.499**), embora com **menos casos de "Perda Total"** em comparação com o **grupo 1** (89 vs 155), sugerindo acidentes graves mas ligeiramente menos catastróficos.

Conclusão

Embora a Análise em Componentes Principais tenha permitido identificar eixos de interpretação lógica (Custo vs. Maturidade), a **variância total** explicada fixou-se em **apenas 45%**. Este valor modesto impõe cautela na generalização dos resultados, uma vez que mais de metade da variabilidade original dos dados, associada a variáveis como o ano do veículo e testemunhas, não foi captada pelo modelo.

Logo, apesar do PCA ter sido eficaz na extração das dimensões centrais de risco, revelou-se insuficiente para descrever a totalidade desta carteira de clientes.

A subsequente **análise de *clustering*** (k-means) validou a existência de três perfis distintos de sinistralidade, permitindo uma segmentação de risco.

Seniores Fidelizados de alto risco (Grupo 1): Clientes com longa relação contratual, predominantemente do género feminino, envolvidos em colisões de alto custo e elevada taxa de perda total.

Clientes de baixo risco (Grupo 2): Sinistros de baixo valor monetário, estritamente relacionados com roubos e danos em estacionamento, isolando ocorrências que não advêm de condução perigosa ativa.

Jovens de alto risco (Grupo 3): Clientes mais recentes e jovens, com um perfil de risco financeiro elevado, muito semelhante ao grupo 1 em termos de custos, mas demograficamente distinto.

Concluindo, esta análise demonstra que o **custo elevado não é transversal a todos os clientes**, concentrando-se em grupos específicos de **colisões (Jovens e Seniores)**, enquanto no **Grupo 2 os sinistros representam ocorrências de baixo impacto financeiro**. Estes resultados permitem à seguradora desenhar **estratégias diferenciadas**: políticas de retenção e acompanhamento para os clientes seniores de alto valor, e revisão de prémios de risco para o segmento jovem, separando eficazmente a gestão de sinistros graves da gestão de ocorrências menores e administrativas.