

Relatório do trabalho de Métodos Descritivos de Data Mining

Professor João Fonseca



Grupo AB

Íris Corrula – m20200405

Mariya Mukhacheva – m20200334

Índice

Introdução	1
Método de Trabalho.....	1
Análise de Variáveis.....	3
Valores Negativos.....	10
Exclusão de Outliers	10
Transformação dos Missing Values	11
Avaliação e Caracterização de Segmentos	12
Escolha do número de Clusters	14
Importância das Variáveis	17
Análise das variáveis segundo os Segmentos.....	18
Salário	18
Valor do Cliente	19
Percentagem do Salário gasto em Percentagem	20
Idade e Presença de Filhos	20
Zona de Residência, Educação e Antiguidade do Cliente.....	21
Perfil do Cliente	22
Insights.....	22
Conclusão	23
Limitações.....	23

Índice Tabelas

Tabela 1	3
Tabela 2	3
Tabela 3	4
Tabela 4	10
Tabela 5	10
Tabela 6	11
Tabela 7	11
Tabela 8	14
Tabela 9	16
Tabela 10	17
Tabela 11	17
Tabela 12	19
Tabela 13	19
Tabela 14	19
Tabela 15 Tabela 16	19
Tabela 17	20
Tabela 18	20
Tabela 19 Tabela 20	21

Tabela 21	21
Tabela 22	22

Índice Figuras

Figura 1 – Frequencia de clients por ClaimsRate.....	5
Figura 2 – Valor monetário do cliente (€).	5
Figura 3 – Valor de Seguro de saúde pago por cliente (€).	5
Figura 4 – Valor do Seguro de lar pago por cliente (€).....	6
Figura 5 – Valor de seguro de trabalho pago por clientes (€).	6
Figura 6 – Valor de Seguro de Vida pago por cliente (€).	6
Figura 7 – Valor de seguro de carro pago por cliente (€).	7
Figura 8 – Salário mensal bruto por cliente (€).	7
Figura 9 – Ano em que cada cliente se tornou cliente na seguradora	7
Figura 10 – Ano de Nascimento por cliente	8
Figura 11 – Cliente com e sem filhos.....	8
Figura 12 – Grau de educação por cliente.....	8
Figura 13 – Zona em que cada cliente vive.	9
Figura 14 – Gráficos das correlações.....	9
Figura 15 – Valor pago de Seguros de saúde e carro e salário bruto mensal por cliente	12
Figura 16 – Valor pago de Seguro de Lar, Trabalho e Vida por cliente	12
Figura 17 – Frequência de clientes por ClaimsRate	12
Figura 18 – Gráfico de Correlações	13
Figura 19 – Cluster Plot	14
Figura 20 – Distancia do centroid e número de clusters possíveis	14
Figura 21 – Tamanho dos clusters para cada cenário	15
Figura 22 – Proximidades de clusters para K=4 e K=5	15
Figura 23 – Dispersão de clusters, em média, por cada produto apresentado	15
Figura 24 – Comparação da distribuição dos segmentos obtidos com a distribuição das variáveis de produtos	18
Figura 25 – Salário bruto mensal por cliente e por escalão de rendimento	18

Introdução

No âmbito da cadeira de Data Mining I, foi-nos proposta a realização do projeto em que com base nos clientes de uma seguradora portuguesa é necessário efetuar um estudo de modo a identificar diferentes segmentos.

O grupo representa consultores especializados em Data Mining, e o grande desafio deste trabalho, é descrever e explicar os diferentes *clusters* de clientes com a respetiva abordagem de vantagens ou desvantagens de diferentes decisões de modo, a que o departamento de Marketing possa entender o comportamento dos seus clientes.

Método de Trabalho

Decidimos pôr em prática todo o conhecimento adquirido nas aulas de Data Mining I para obter os resultados coerentes e que façam sentido perante o contexto:

- 1- Começamos por analisar as variáveis com o **Input Data**. Foi Importado o *dataset* para se verificar todas as variáveis e os seus tipos e alguma possível incoerência nas mesmas. Vimos que não existem registos duplicados. E fizemos uma exploração inicial de dados;
- 2- De seguida avançamos com a interpretação dos conceitos das variáveis e a exploração dos dados através dos nodes: **StatExplore**, **Multiplot**, **Graph Explore** e **Variable Clustering** para verificar se existem:
 - a. *Missing values*
 - b. *Outliers* em variáveis numéricas, que pode ter duas interpretações:
 - i. Valores que são erros, como por exemplo ano de nascimento = 1028
 - ii. Valores que distanciam radicalmente de todos os outros, no entanto fazem sentido, como por exemplo ordenado = 55 215 €
 - c. Valores negativos
 - d. Identificar quais as variáveis com pouco valor para a análise dos segmentos, através da análise das correlações entre as variáveis;
- 3- Aplicamos os nodes: **Transform Variables e Filter** para converter os valores negativos em variáveis de prémios e fazer a filtragem dos *outliers* e dos *missings values* em variáveis categóricas, neste passo, adicionalmente, efetuamos a análise dos valores excluídos;
- 4- Após a exploração das variáveis e filtragem dos *outliers*, utilizámos o node **Impute** com o objetivo de substituir os *missing values* das variáveis numéricas com o valor médio da respetiva variável e aplicar as transformações importantes de algumas variáveis. Criaram-se variáveis para ajudar a análise com node **Transform Variables**. Adicionalmente, voltamos a olhar para as correlações entre todas as variáveis de modo a identificar as fortemente correlacionadas e excluir uma delas, visto que ter duas ou mais variáveis fortemente correlacionadas (>85%) não traz valor ao modelo de *clustering*. Para a exclusão destas variáveis utilizou-se a **Metadata** para que estas variáveis correlacionadas não façam parte da segmentação do cliente.
- 5- Prosseguimos para a aplicação o node: **Variable Clustering** para compreender a relação entre as variáveis. Nesta fase tivemos de tomar uma decisão, tendo em conta o contexto e o comportamento das variáveis e se faz sentido ter duas análises paralelas de *clustering*. Uma com base nas variáveis de pagamento de prémios e outra com base nas variáveis globais sobre o cliente. Chegou-se à conclusão que só fazia sentido aplicar apenas uma análise de *clustering* feita com base na informação de pagamento de prémios visto que não tínhamos variáveis contínuas suficientes para formar um modelo

sobre o comportamento global do cliente. De seguida aplicamos o node **SAS Code** com o código que calcula as distâncias até ao centroide em vários cenários de *clusters*;

- 6- Passamos para o node **Cluster** para efetuar a segmentação. Nesta fase fizemos a análise de vários cenários e optou-se pelo cenário com 4 segmentos, com base nos seus pagamentos de prémios de acordo com o tipo de seguro. Extraímos a informação da segmentação com as restantes variáveis e com o suporte do **Excel** fizemos uma análise detalhada de modo a destacar padrões com ajuda da aplicação de formatação condicional pelo escalão de cores;
- 7- Para concluir a análise e apresentar os resultados ao departamento de Marketing, elaboramos um mapa resumo do perfil do cliente e os seus insights em cada um dos segmentos;

Para finalizar, o nosso trabalho explicamos algumas limitações e as vantagens da aplicação da segmentação de clientes;

Nota:

Para efetuar a exploração da informação usamos as seguintes estatísticas para interpretar o seu comportamento:

Média	É a média dos dados, que é a soma de todas as observações divididas pelo número de observações. Dá nos ideia sobre o valor central da informação. É uma medida pouco resistente, é muito sensível a valores grandes e muito pequenos (<i>outliers</i>)
Mediana	É o ponto médio do conjunto de dados. Este valor do ponto médio é o ponto em que metade das observações estão acima do valor e metade das observações estão abaixo do valor. A mediana é determinada por classificar as observações e encontrar a observação que está no número $[N + 1] / 2$ na ordem de grandeza. Se o número de observações for ímpar, a mediana é o valor médio das observações que são classificadas com números de $N / 2$ e $[N / 2] + 1$. A mediana é mais resistente a números grandes, e acaba por dar-nos uma ideia mais correta sobre o valor central da informação.
Q1	Quartis são os três valores: o primeiro quartil a 25% (Q1), o segundo quartil a 50% (Q2 ou mediana) e o terceiro quartil a 75% (Q3), que dividem uma amostra de dados ordenados em quatro partes iguais.
Q3	
Q1-1,5*IQR	É o cálculo para o traçar o limite inferior e superior, os valores que são inferiores ou superiores são considerados <i>outliers</i> . IQR é a amplitude interquartil (Q3-Q1), é uma medida de dispersão, não é influenciada pelos <i>outliers</i> .
Q3+1,5*IQR	
Desvio Padrão	É a medida mais comum de dispersão, ou o quanto os dados estão dispersos sobre a média. Como o desvio padrão está nas mesmas unidades que os dados, ele é normalmente mais fácil de interpretar do que a variância. É uma medida pouco resistente, é muito sensível a valores grandes e muito pequenos (<i>outliers</i>)
Correlação	É a relação dentro de uma ampla classe de relações estatísticas que envolva dependência entre duas variáveis, varia entre -1 (correlação negativa) e 1 (correlação positiva), em que, quanto mais o seu valor estiver perto dos extremos mais forte é a relação.

Análise de Variáveis

As variáveis que nos foram fornecidas são referentes a **10 296** clientes de uma seguradora. As respetivas variáveis, descrição e o tipo de variável estão apresentados no seguinte quadro:

Variável	Descrição	Tipo de Variável
CustID	ID do cliente (Chave da tabela)	ID
FirstPolYear	Ano em que se tornou cliente	Continua
BirthYear	Ano do nascimento do cliente	Continua
EducDeg	Grau de educação	Nominal
MonthSal	Salário bruto mensal (€)	Continua
GeoLivArea	Área onde o cliente reside	Nominal
Children	Presença de filhos, (1-sim, 0-não)	Nominal (Binária)
CustMonVal	Valor monetário do cliente (€)	Continua
ClaimsRate	Rácio do valor monetário do cliente	Continua
PremMotor	Prémio mensal pago no seguro de carro (€)	Continua
PremHousehold	Prémio mensal pago no seguro de lar (€)	Continua
PremHealth	Prémio mensal pago no seguro de saúde (€)	Continua
PremLife	Prémio mensal pago no seguro de vida (€)	Continua
PremWork	Prémio mensal pago no seguro de trabalho (€)	Continua

Tabela 1

A primeira análise realizada às variáveis contínuas foi a análise das suas estatísticas principais, como se encontra no quadro em baixo:

Variável	CustMonVal	ClaimsRate	PremHealth	PremHousehold	PremWork	PremLife	PremMotor	MonthSal	FirstPolYear	BirthYear
Média	177,89	0,74	171,58 €	210,43 €	41,28 €	41,86 €	300,47 €	2 506,67 €	1991	1968
Mediana	186,71	0,72	162,81 €	132,80 €	25,67 €	25,56 €	298,61 €	2 501,0 €	1986	1968
Q1	-9,44	0,39	111,82 €	49,45 €	10,67 €	9,89 €	190,59 €	1 706,0 €	1980	1953
Q3	399,78	0,98	219,82 €	290,05 €	56,79 €	57,79 €	408,30 €	3 290,0 €	1992	1983
Q1-1,5*IQR	-623,27	-0,50	-50,23 €	- 311,95 €	- 58,61 €	- 61,96 €	- 135,98 €	- 670,38 €	1962	1908
Q3+1,5*IQR	1 013,61	1,87	381,85 €	650,90 €	125,97 €	129,64 €	734,87 €	5 666,63 €	2010	2028
Máximo	11 875,89	256,20	28 272,0 €	25 048,80 €	1 988,70 €	398,30 €	11 604,42 €	55 215,0 €	53 784	2001
Mínimo	-165 680,0	0,00	-2,11 €	- 75,00 €	- 12,00 €	- 7,00 €	- 4,11 €	333,0 €	1974	1028
Desvio Padrão	1 945,8	2,90	296,41	352,60	51,51	47,48	211,91	1 157,45	511,27	19,71
Outlier	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Missing Value	0	0	43	0	86	104	34	36	30	17

Tabela 2

As variáveis **PremHealth**, **PremHousehold**, **PremWork**, **PremLife** e **PremMotor** representam os prémios da seguradora. Estas variáveis apresentam valores negativos, isto porque existem clientes da seguradora que no ano anterior pagaram o prémio referente ao ano que se está a analisar. Analisando a média e mediana destas variáveis, observa-se que em todas elas a média é superior à mediana, o que nos indica que a assimetria é positiva, ou seja, existem registos com valores mais elevados. Estas situações podem-se dever ao facto de as variáveis apresentarem *outliers*. É de notar que a variável **PremMotor** tem a média e a mediana muito próximas, quer dizer que os seus dados têm distribuição simétrica. Relativamente à dispersão dos valores desta variável, pode-se considerar que o valor do desvio padrão é muito elevado, sendo este superior à média, e por isso pode-se concluir que a variável tem os seus registos muito dispersos.

A variável **CustMonVal** que representa o valor monetário do cliente, representa valores negativos, ou seja, existem clientes que não acrescentam valor à seguradora, o que por outras palavras significa que a seguradora perde mais dinheiro do que ganha com estes clientes. Analisando a média e a mediana, verifica-se que a média é inferior à mediana, pelo que esta variável tem uma assimetria negativa, ou seja, existem valores muito baixos que estão a influenciar a média. Assim sendo pode-se desde já dizer que existem clientes que não acrescentam valor monetário à seguradora, ou seja a seguradora perde muito dinheiro com alguns dos clientes aqui analisados (valor mínimo= -165 680,0). Relativamente à dispersão dos valores, o desvio padrão tem um elevado valor, acima da média da variável, sendo que os valores são dispersos negativamente (média é inferior à mediana).

A variável **ClaimsRate** representa o rácio do valor monetário do cliente, ou seja, o seu conceito está relacionado com o da variável **CustMonVal**, sendo que não apresenta qualquer valor negativo, isso é explicado pelo facto de que é um rácio (uma variável ser dividida pela outra). Observando a média e a mediana, é de averiguar que a média é próxima da mediana por isso pode-se considerar que a dispersão é simétrica, pelo que tanto existe um equilíbrio entre a média e os valores mínimos e máximos, o desvio padrão está acima da média, isso indica uma dispersão de registos com valores elevados, como os dados estão centrados podemos afirmar que esta situação está a acontecer devido à existência de um pequeno número de *outliers*.

As variáveis **MonthSal**, **FirstPolYear** e **BirthYear** têm a média e mediana muito próximas, pelo que podemos dizer que existe uma simetria nestes variáveis. Esta afirmação pode ser considerada válida através do desvio padrão que são inferiores à média, e com valores relativamente próximos, o que se pode considerar uma concentração de dados. Nenhuma destas variáveis representa valores negativos, o que faz todo o sentido uma vez que estamos perante a representação do Salário mensal bruto, o ano em que o cliente se tornou cliente e o ano em que o cliente nasceu.

Por último analisou-se os valores omissos de cada variável, e verificou-se que as variáveis **PremHealth**, **PremWork**, **PremLife**, **PremMotor**, **MonthSal**, **FirstPolYear** e **BirthYear** apresentam valores omissos. Como estes valores representam 3% dos registos totais, ou seja, a percentagem já é considerada significativa e a retirada destes valores significa a perda de muitos registos, pelo que são registos a serem modificados e permaneceram na análise futura.

Desta análise também se pode analisar os *outliers* que foi calculada através da regra do cálculo dos interquartis. Assim sendo calculou-se os limites dos interquartis em que $IQR = (Q3 - Q1)$ e o cálculo para o limite inferior é $Q1 - 1.5 * IQR$ e para o limite superior $Q3 + 1.5 * IQR$. Pode-se então verificar que todas as variáveis contêm *outliers*, sendo estes avaliados em pormenor e retirados (pelo menos os mais extremos), da análise.

A análise realizada às variáveis nominais encontra-se no quadro em baixo:

Variável	Children	EducDeg	GeoLivArea
Nº Níveis	3	5	5
Moda	1	3-BSc/MSc	4
Missing Values	21	17	1

Tabela 3

Verifica-se que em todas as variáveis categóricas existem valores omissos, sendo que o total de registos é no máximo 39. Como não temos a certeza da veracidade dos valores omissos e por apenas representarem 0.4% dos dados, decidimos retirar esses registos.

Passamos agora para a análise das representações gráficas das variáveis para conhecermos melhor o seu comportamento:

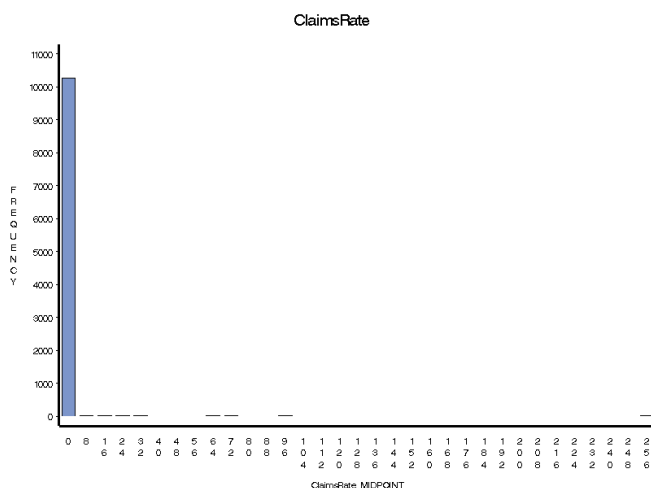


Figura 1 – Frequencia de clients por ClaimsRate

A variável **ClaimsRate** representa o valor que o cliente traz à empresa em rácio:

$$\frac{\text{Valor pago pela seguradora}}{\text{Prémio}}$$

em que se for <1 é interpretado como: cliente pagou mais em prémio do que recebeu em indenização da seguradora, quer dizer que a seguradora “ganhou” lucro com esse cliente, se for ≥ 1 é exatamente ao contrário, a seguradora “perdeu” dinheiro com esse cliente.

Pela tabela 2 e pela figura 1 podemos observar que a maioria dos valores se concentram entre o 0 e o 32. Verifica-se claramente que existem *outliers* e não verificamos *missing values*.

A variável **CustMonVal**, tal com a variável **ClaimsRate** representa o valor que o cliente traz à empresa, mas no caso da **CustMonVal** a unidade é em € e quanto maior for o valor mais seguradora ganha com o cliente:

$$(\text{Lucro anual pelo cliente}) \times (\text{Antiguidade do cliente}) - \text{Custos}$$

Pela Tabela 2 e pelo gráfico podemos observar que uma parte dos valores são negativos e maior parte dos valores estão concentrados entre -3000 e 3000€, nota-se claramente que existem valores *outliers*, que são os valores radicalmente diferentes do padrão.

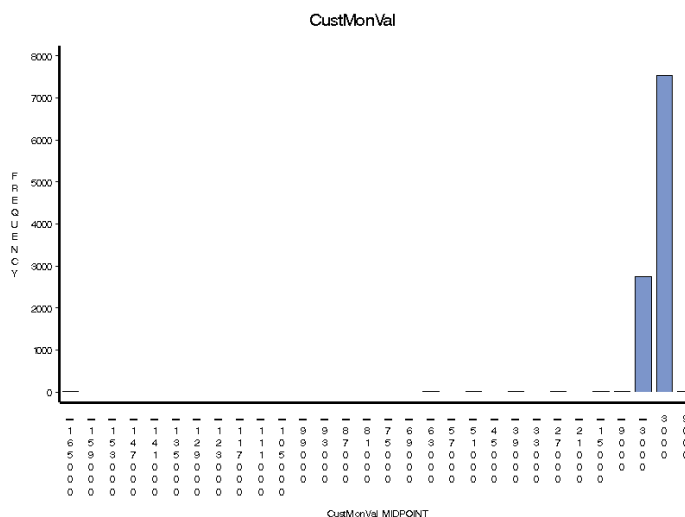


Figura 2 – Valor monetário do cliente (€).

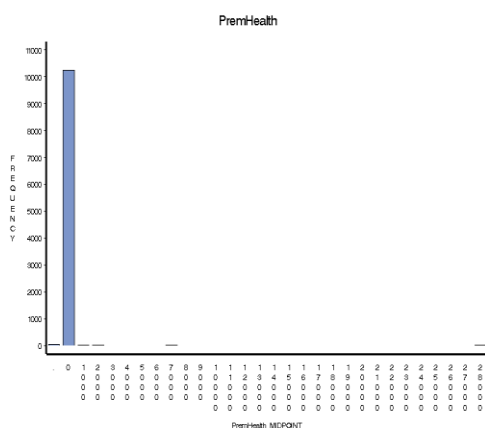


Figura 3 – Valor de Seguro de saúde pago por cliente (€).

A variável **PremHealth** representa o valor monetário de prémio pago pelo cliente, referente ao seguro de saúde. Pela tabela 2 e pelo gráfico podemos observar que a maior parte dos valores estão concentrados entre 0 e 1000€. Nota-se claramente que existem valores *outliers*, que são os valores radicalmente diferentes do padrão, e verifica-se a presença de valores nulos e valores negativos.

A variável **PremHousehold** representa o valor monetário de prémio pago, referente ao seguro de lar. Pela tabela 2 e pelo gráfico podemos observar que maior parte dos valores estão concentrados entre 0 e 1600€. Nota-se claramente que existem valores *outliers* e verifica-se a presença de valores negativos.

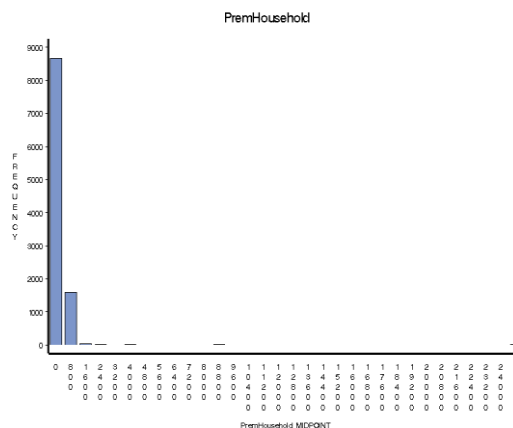


Figura 4 – Valor do Seguro de lar pago por cliente (€).

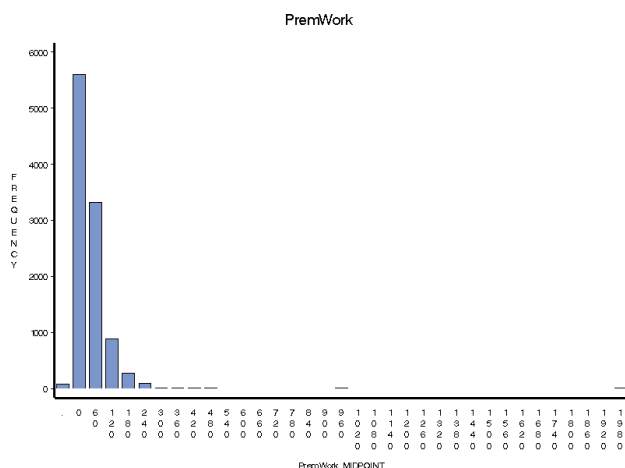


Figura 5 – Valor de seguro de trabalho pago por clientes (€).

A variável **PremWork** representa o valor monetário de prémio pago, referente ao seguro de trabalho. Pela tabela 2 e pelo gráfico podemos observar que maior parte dos valores estão concentrados entre 0 e 300€. Nota-se claramente que existem valores *outliers* e verifica-se a presença de valores negativos e nulos.

A variável **PremLife** representa o valor monetário de prémio pago pelo cliente referente ao seguro de vida. Pela tabela 2 e pelo gráfico podemos observar que maior parte dos valores estão concentrados entre 0 e 246€. Neste caso não existe evidência de afirmar a presença de valores *outliers*, no entanto, verifica-se a presença de valores negativos e nulos.

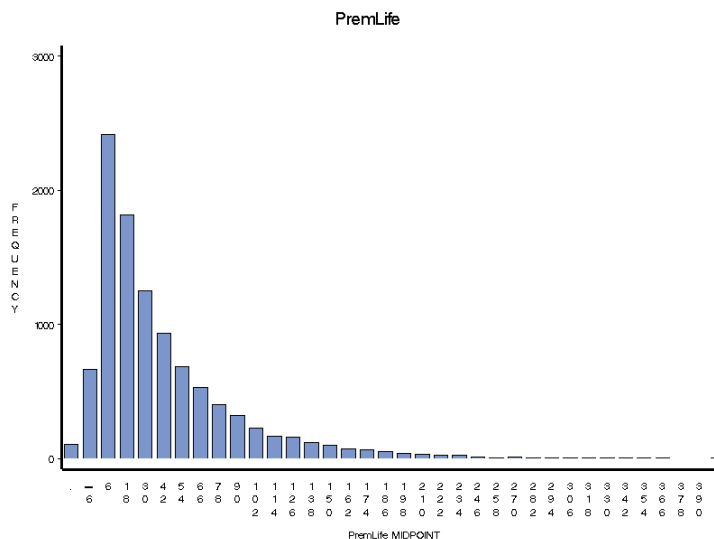


Figura 6 – Valor de Seguro de Vida pago por cliente (€).

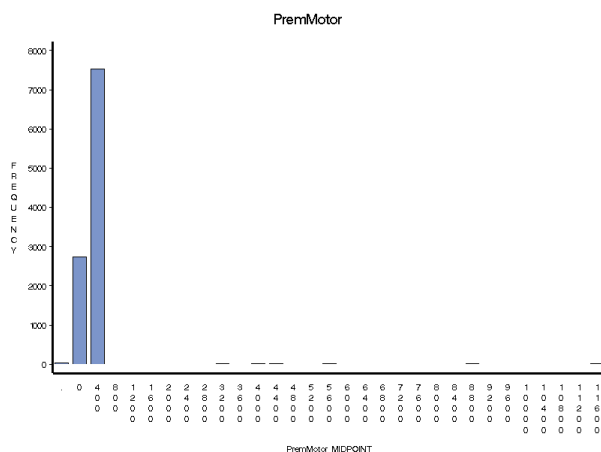


Figura 7 – Valor de seguro de carro pago por cliente (€).

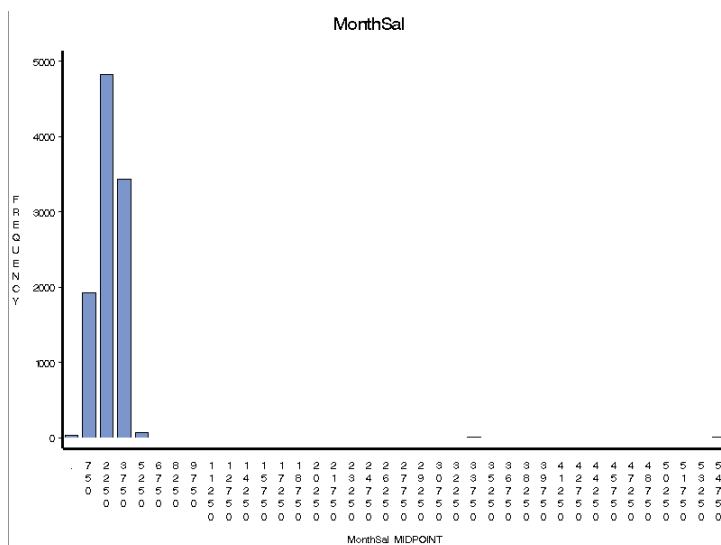


Figura 8 – Salário mensal bruto por cliente (€).

A variável **FirsPolYear** representa o 1º ano do cliente na seguradora. Pela tabela 2 e pelo gráfico podemos observar a presença de um *outlier* e de valores nulos. É de notar que a distribuição de clientes parece ser algum muito próximo com a distribuição uniforme, isto é, se retirarmos aleatoriamente qualquer cliente desta amostra a probabilidade é “igual” de ele ter qualquer ano entre 1974 e 1998. Adicionalmente, com base nesta informação podemos criar uma variável “antiguidade”, (2016- FirsPolYear).

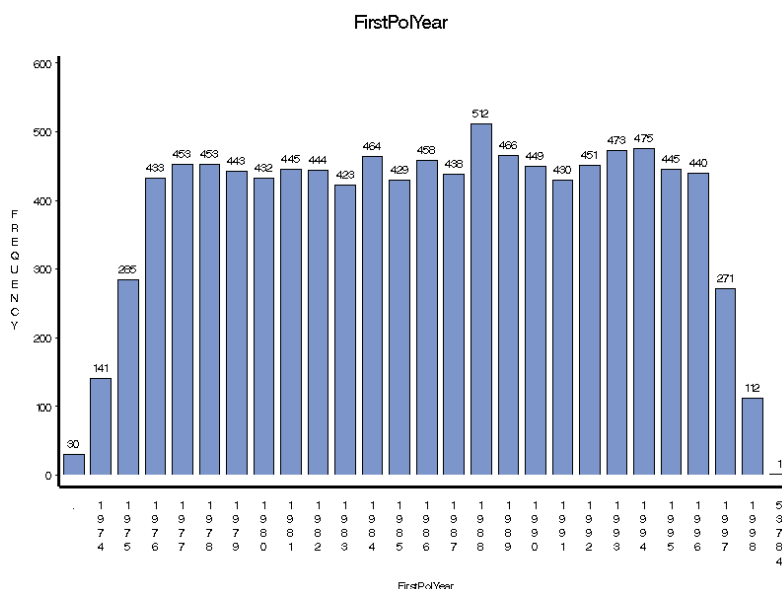


Figura 9 – Ano em que cada cliente se tornou cliente na seguradora

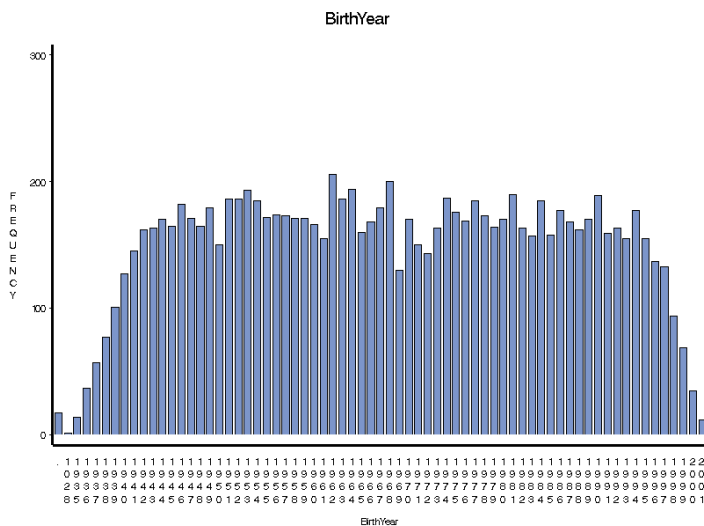


Figura 10 – Ano de Nascimento por cliente

A variável **Children** é considerada uma variável binária em que 1 significa que o cliente tem filhos e 0 que o cliente não tem filhos. Pela análise da tabela 2 e do gráfico à esquerda, pode-se dizer que existem mais clientes com filhos do que sem filhos. Também se pode ver que esta variável possui valores omissos.

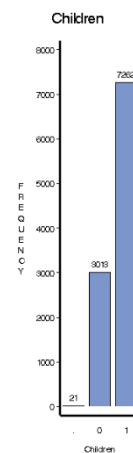


Figura 11 – Cliente com e sem filhos

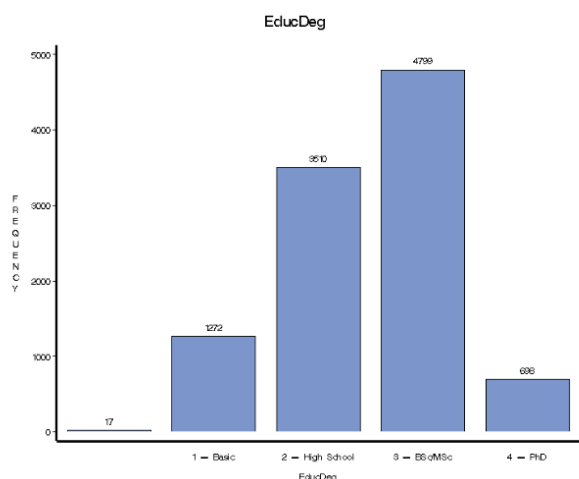


Figura 12 – Grau de educação por cliente

A variável **EducDeg** é uma variável categoria que representa o grau académico de cada cliente da seguradora. Através da tabela 2 e do gráfico apresentado podemos dizer que o nível de educação BSC/MSc (Licenciados e Mestres) é adquirido por mais clientes e que o que menos clientes têm é o PHD (Doutorados). Podemos ainda verificar que existem valores omissos e que comparados com os restantes níveis podemos verificar que é quase insignificante.

A variável **GeoLivArea** é uma variável categórica que representa geograficamente onde os clientes da seguradora residem. Pelo que se verificou tanto na tabela 2 como no gráfico representativo, existem um elevado número de cliente a viverem na zona 4 e menos cliente a viverem na zona 2. Aqui pose-se confirmar que a variável contém valores omissos, e que são insignificantes em comparação com os restantes níveis da variável

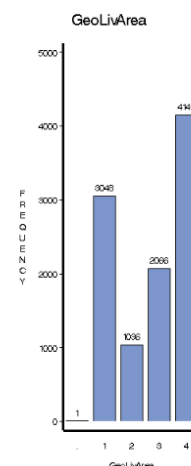


Figura 13 – Zona em que cada cliente vive.

Tendo em conta que as duas variáveis **CustMonVal** e **ClaimsRate** representam o mesmo conceito decidimos verificar a sua correlação.

	BirthYear	ClaimsRate	CustMonVal	FirstPolYear	MonthSal	PremHealth	PremHousehold	PremLife	PremMotor	PremWork
BirthYear	100%	0%	0%	-1%	-69%	0%	15%	23%	-16%	21%
ClaimsRate	0%	100%	-99%	0%	0%	1%	-1%	0%	-1%	0%
CustMonVal	0%	-99%	100%	0%	0%	0%	3%	1%	3%	2%
FirstPolYear	-1%	0%	0%	100%	1%	0%	-1%	0%	0%	0%
MonthSal	-69%	0%	0%	1%	100%	0%	-13%	-20%	14%	-17%
PremHealth	0%	1%	0%	0%	0%	100%	2%	3%	-7%	8%
PremHousehold	15%	-1%	3%	-1%	-13%	2%	100%	26%	-27%	24%
PremLife	23%	0%	1%	0%	-20%	3%	26%	100%	-41%	34%
PremMotor	-16%	-1%	3%	0%	14%	-7%	-27%	-41%	100%	-35%
PremWork	21%	0%	2%	0%	-17%	8%	24%	34%	-35%	100%

Figura 14 – Gráficos das correlações

Pela matriz de correlações podemos verificar que existe uma correlação negativa muito forte (-99%) entre as variáveis **ClaimRate** e **CustMonVal**. Por isso decidimos retirar a variável **CustMonVal** da análise, pois tem um desvio padrão maior, ou seja, os dados são mais dispersos do que em **ClaimRate**.

Aproveitando a informação sobre as restantes correlações, podemos ver que **BirthYear** e **MonthSal** também têm correlação negativa, no entanto não é tão forte (-69%). Assim sendo, foi decidido proceder com essas variáveis, e confirmar a sua correlação depois da “limpeza” (dos *outliers*). Em relação às restantes variáveis, todas elas têm correlações baixas.

Valores Negativos

Em variáveis contínuas que representam os pagamentos de prémios, sendo o caso das variáveis **“PremMotor”, “PremHousehold”, “PremHealth”, “PremLife” e “PremWork”**, em que os valores negativos dizem respeito aos pagamentos feitos nos anos anteriores, foi decidido converter esses valores negativos em positivos.

Pode-se observar as estatísticas ligeiramente alteradas:

Variável	ClaimsRate	PremHealth	PremHousehold	PremWork	PremLife	PremMotor	MonthSal	FirstPolYear	BirthYear
Média	0,74	171,58 €	216,56 €	42,09 €	42,25 €	300,47 €	2 506,67 €	1991	1968
Média (Ant.)	0,74	171,58 €	210,43 €	41,28 €	41,86 €	300,47 €	2 506,67 €	1991	1968
Mediana	0,72	162,81 €	132,80 €	25,67 €	25,56 €	298,61 €	2 501,00 €	1986	1968
Q1	0,39	111,82 €	53,90 €	10,89 €	9,89 €	190,59 €	1 706,00 €	1980	1953
Q3	0,98	219,82 €	290,05 €	56,79 €	57,79 €	408,30 €	3 290,00 €	1992	1983
Q1-1,5*IQR	-0,50	-50,23 €	-300,33 €	- 57,96 €	- 61,96 €	- 135,98 €	- 670,38 €	1962	1908
Q3+1,5*IQR	1,87	381,85 €	644,275 €	125,64 €	129,64 €	734,87 €	5 666,63 €	2010	2028
Máximo	256,20	28 272,00 €	25 048,80 €	1 988,70 €	398,30 €	11 604,42 €	55 215,00 €	53 784	2001
Mínimo	0	2,11 €	0 €	0,11 €	0,11 €	1,78 €	333,00 €	1974	1028
Mínimo (ant.)	0	-2,11 €	-75,0 €	-12,0 €	-7,0 €	-4,11 €	333,0 €	1974	1028
Desvio Padrão	2,92	296,41	348,87	50,85	47,13	211,91	1 157,45	511,27	19,71
D. Padrão (ant.)	2,90	296,41	352,60	51,51	47,48	211,91	1 157,45	511,27	19,71
Outlier	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Missing Value	0	43	0	86	104	34	36	30	17

Tabela 4

Exclusão de Outliers

Inicialmente, pensamos em aplicar “Extrem Percentils” com os percentis de 1%, no entanto isso provocaria retirada de alguns poucos registos a mais, por isso decidimos definir os limites para exclusão manualmente, retirando apenas os *outliers* mais severos com base na observação direta aos dados. É de notar que esse tipo de trabalho manual só é possível em poucas variáveis, porque se tivesse de analisar mais de 10 variáveis teria de se aplicar um método automático dos percentis extremos.

Aplicamos então os seguintes limites apresentados na tabela 5, o que provocou a exclusão de 31 registos.

Nome da variável	Limite Inferior			Limite Superior		
FirstPolYear	1974	-1	1973	1998	+1	1999
BirthYear	1935	-1	1934	2001	+1	2002
MonthSal	333	-1	332	5021	+1	5022
ClaimsRate	0		0	1,62	+1	3
PremMotor	1,78	-1	1	585,22	+1	586
PremHousehold	0		0	2223,75	+1	2225
PremHealth	2,11	-1	1	442,86	+1	444
PremLife	0,11		0	398,3	+1	399
PremWork	0,11		0	930,44	+1	931

Tabela 5

Desta forma, tendo em consideração a exclusão dos registos das variáveis nominais, os valores omissos, e a exclusão dos *outliers* em variáveis contínuas, verificou-se que no total foram retirados 69 registos da nossa base de dados, o que equivale a 0.67% do total de registo. Ficamos com **10 227** registos. Em cada um dos limites acrescentamos uma “folga” de -1 e +1 para retirar os decimais, e se estamos a aplicar esse critério em uma variável achamos que faz sentido aplicar noutras, igualmente desta forma estamos reduzir a margem de erro caso haja um erro da nossa parte em analisar os dados manualmente.

Transformação dos Missing Values

Como anteriormente se verificou, a maioria das variáveis contínuas têm presença de valores omissos e estes representam 270 registos (2.6%). Este valor de valores omissos é significativo, e ao ser retirado perdía-se uma grande quantidade de registos. Optamos por aplicar o método de substituir os valores omissos pela média da respetiva variável e desta forma não alterar as estatísticas de cada uma das variáveis. No entanto, é de referir que, se a quantidade de registos fosse muito maior também não podíamos simplesmente substituir valores omissos pela média, porque isso poderia mudar fundamentalmente a estrutura dos dados subjacentes e influenciar no comportamento dos *clusters*, e a decisão poderia estar fundamentada em valores ausentes substituídos pelas médias, ao invés dos dados originais.

Assim sendo, as limpezas dos dados são apresentadas nos quadros em baixo:

Variável	ClaimsRate	PremHealth	PremHousehold	PremWork	PremLife	PremMotor	MonthSal	FirstPolYear	BirthYear
Média	0,68	167,94 €	213,25 €	41,80 €	42,27 €	296,88 €	2 497,12 €	1986	1968
Média (Ant.)	0,74	171,58 €	216,56 €	42,09 €	42,25 €	300,47 €	2 506,67 €	1991	1968
Mediana	0,72	163,03 €	132,80 €	26,45 €	26,45 €	298,28 €	2 497,12 €	1986	1968
Mediana (Ant.)	0,72	162,81 €	132,80 €	25,67 €	25,56 €	298,61 €	2 501,00 €	1986	1968
Máximo	1,62	442,86 €	2 223,75 €	494,10 €	398,30 €	585,22 €	5 021,00 €	1998	2001
Máximo (Ant.)	256,20	28 272,00 €	25 048,80 €	1 988,70 €	398,30 €	11 604,42 €	55 215,00 €	53784	2001
Mínimo	0,00	2,11 €	- €	0,11 €	0,11 €	1,78 €	333,00 €	1974	1935
Mínimo (Ant.)	0,00	2,11 €	- €	0,11 €	0,11 €	1,78 €	333,00 €	1974	1028
Desvio Padrão	0,32	74,01	230,96	45,99	46,93	137,81	982,86	6,61	17,38
D. Padrão (Ant.)	2,92	296,41	348,87	50,85	47,13	211,91	1157,45	511,27	19,71

Tabela 6

Variavel	Children	EducDeg	GeoLivArea
Nº Níveis	2	4	4
Moda	1	3-BSc/MSc	4
Missing Values	0	0	0

Tabela 7

Depois da exclusão dos *outliers* podemos observar uma ligeira descida do valor médio, pois os *outliers* deixaram de ser influenciados pela média. Foi com base nessas médias recalculadas que os valores omissos foram substituídos em cada uma das variáveis.

Pode-se verificar que o desvio padrão de todas as variáveis foi alterado significativamente, ou seja, houve uma diminuição da dispersão dos dados, isto é, a informação para cada variável ficou mais concentrada.

Para complementar a análise pré-processual voltou-se a olhar para alguns dos gráficos das variáveis agora mais “nítidos” em termos de leitura.

Pode-se observar que as variáveis **PremHealth**, **PremMotor** e **MonthSal** têm distribuições simétricas, que estas já não contêm valores omissos e outliers.

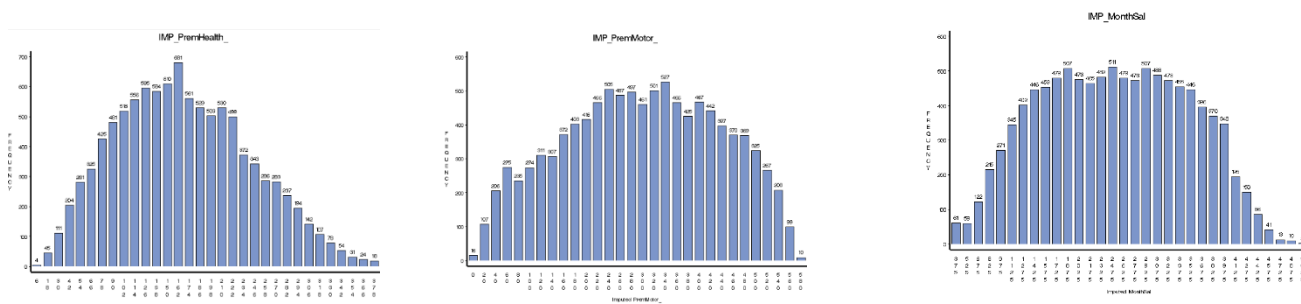


Figura 15 – Valor pago de Seguros de saúde e carro e salário bruto mensal por cliente

A distribuição das variáveis, **PremHousehold**, **PremWork**, **PremLife** apresentam assimetria positiva (ou à direita).

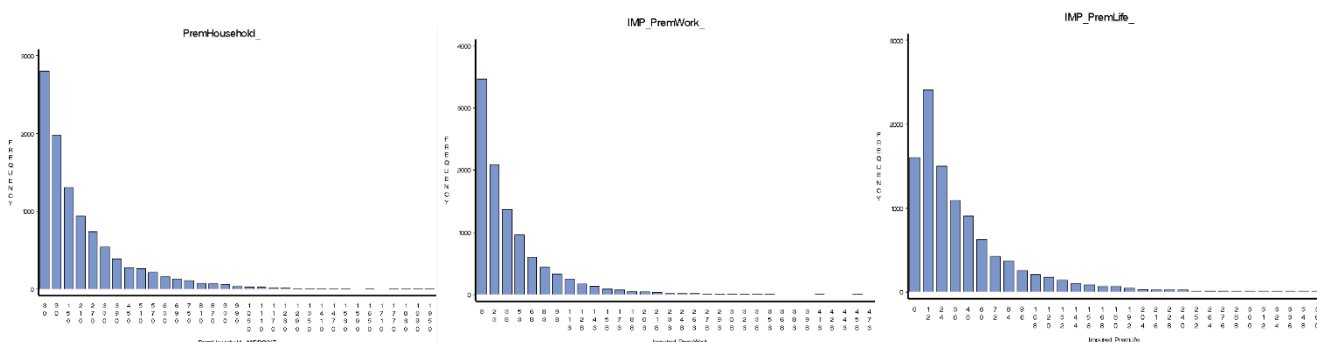


Figura 16 – Valor pago de Seguro de Lar, Trabalho e Vida por cliente

No caso da variável **ClaimsRate** podemos observar três comportamentos diferentes, os valores entre 0 e 0.85 têm um comportamento quase uniforme, depois observamos uma grande concentração a volta do 1, e de seguida vemos uma caída radical da frequência de observações a partir do 1.05. Podemos concluir que existem mais clientes a contribuírem para o lucro da seguradora do que os clientes indenizados.

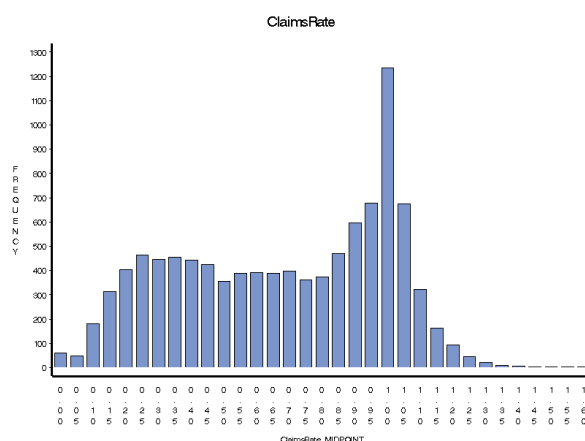


Figura 17 – Frequência de clientes por ClaimsRate

Avaliação e Caracterização de Segmentos

Ao fim de termos concluído a fase da limpeza pré-processual, chegamos a conclusão que fazia sentido criar variáveis que possam trazer valor para a análise de *clustering*:

As novas variáveis:

- ✓ **Idade** 2016 – BirthYear
Representa a Idade dos clientes no ano de 2016 (Pois estes dados pertencem a 2016)
- ✓ **Antiguidade** 2016 – FirstPolYear
Representa a Idade a que o Cliente se tornou cliente da seguradora
- ✓ **Premio_Total** PremHealth + PremLife + PremMotor + PremWork + PremHousehold
Representa o montante total pago em prémios pelo cliente
- ✓ **Percen_gasta_seguro** Premio_Total/ MonthSal
Representa a proporção do salário gasta em pagamento de prémios pelo cliente

Verificamos de novo as correlações, pois foram adicionadas mais variáveis aos dados em análise, para verificar-se se existem ou não correlações novas:

	Antiguidade	Claims Rate	Idade	Prem Health	Prem Life	Prem Motor	Prem Work	Prem Household	% gasta seguro	Premio Total	MonthSal
Antiguidade	100%	1%	-2%	-1%	-1%	1%	-1%	0%	2%	0%	-2%
ClaimsRate	1%	100%	0%	15%	8%	-12%	7%	-5%	0%	-5%	0%
Idade	-2%	0%	100%	3%	-26%	25%	-26%	-26%	-63%	-26%	92%
PremHealth	-1%	15%	3%	100%	8%	-62%	7%	7%	2%	6%	3%
PremLife	-1%	8%	-26%	8%	100%	-65%	37%	40%	39%	39%	-24%
PremMotor	1%	-12%	25%	-62%	-65%	100%	-63%	-64%	-48%	-62%	23%
PremWork	-1%	7%	-26%	7%	37%	-63%	100%	40%	41%	40%	-25%
PremHousehold	0%	-5%	-26%	7%	40%	-64%	40%	100%	62%	100%	-24%
Percen_gasta_seguro	2%	0%	-63%	2%	39%	-48%	41%	62%	100%	62%	-69%
Premio_Total	0%	-5%	-26%	6%	39%	-62%	40%	100%	62%	100%	-24%
MonthSal	-2%	0%	92%	3%	-24%	23%	-25%	-24%	-69%	-24%	100%

Figura 18 – Gráfico de Correlações

Ao analisar as correlações de cada variável, chegamos as seguintes conclusões:

- **“Idade”** está fortemente correlacionada com **“MonthSal”** (92%). Quer dizer que quanto mais velho for o cliente, mais ganha ao final do mês.
→ Excluiu-se a variável Idade para a formação de clusters.
- **“Premio_Total”** está muito influenciada pela variável **“PremHousehold”**, correlação (100%). Ou seja, à medida que o prémio pago pelo seguro de Lar aumenta, o prémio total aumenta igualmente. Assim sendo, está relacionado com o fato de os prémios de seguro de Lar serem os mais elevados.
→ Vamos excluir a variável Prémio Total para a formação de clusters.

É de destacar que a variável **“Antiguidade”** tem as correlações a tender para o 0 em relação às restantes variáveis. Tendo em consideração a figura 18, pode-se concluir que é uma variável completamente independente das restantes, e que a sua distribuição é “uniforme” pelas restantes variáveis. A variável **“ClaimRate”** de forma geral, não apresenta correlação elevada em valor absoluto, apesar de ter alguma correlação com as variáveis **“PremHealth”**, **“PremLife”** e **“PremMotor”**. Tendo em conta o conceito da variável **“ClaimRate”** e por ter alguma correlação com algumas variáveis, acreditamos que esta variável é capaz de trazer valor à segmentação, no entanto, não fazia sentido entrar no modelo de *clustering*. Por isso, resolvemos transformá-la numa variável nominal, para conseguir detetar os clientes que tendem para o valor 1 – os que

estão prestes de deixar de ser rentáveis para a seguradora. E os que têm um valor >1 – os clientes que não são rentáveis para a seguradora:

Se “ClaimsRate” > 1 então “Alerta”

Se “ClaimsRate” > 0.9 e “ClaimsRate” ≤ 1 então “Pré-Alerta”

Caso contrário “OK”

Portanto, depois de excluir algumas das variáveis, procedemos à análise do comportamento das variáveis entre si:

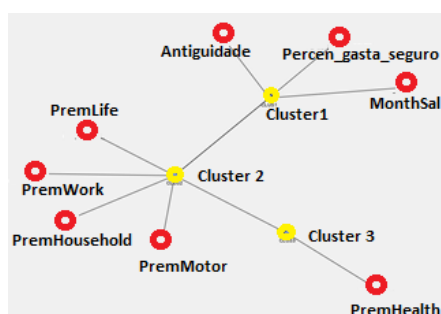


Figura 19 – Cluster Plot

Pela representação gráfica “Cluster Plot” das variáveis podemos ver que foram formados 3 grupos de variáveis, em que o primeiro grupo são as variáveis referentes às características globais do cliente, como: salário, percentagem do rendimento gasto em seguros e a sua antiguidade na seguradora.

O segundo e o terceiro grupo são referente às variáveis que são os prémios pagos por cada um dos produtos da seguradora. A variável do seguro de saúde ficou isolada num só grupo, porque é uma variável pouco relacionada com as restantes, exceto com a variável do prémio de carro.

Para prosseguir com o *clustering*, resolvemos fundir os grupos 2 e 3 e formar um grupo de prémios de produtos da seguradora. No que diz respeito ao grupo 1, tendo em consideração que a “Antiguidade” é uma variável que não vai trazer valor à segmentação, achou-se que seria interessante analisar o comportamento do salário e da percentagem do rendimento dentro de cada segmento gerado pela aplicação do *clustering* dos prémios.

Escolha do número de Clusters

Na figura 16 está apresentada a relação entre o erro, ou seja, a distância entre o *centroid*, e o número de *clusters* possíveis:

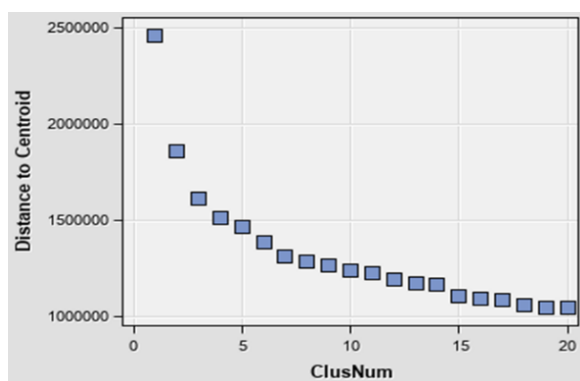


Figura 20 – Distancia do centroid e número de clusters possíveis

Cluster	Distância até Centroid	Redução
1	2453999,03	
2	1857527,69	24%
3	1607262,99	13%
4	1512038,4	6%
5	1465466,32	3%
6	1381991,82	6%

Tabela 8

Na tabela 8, a coluna “Redução”, apresenta a redução da distância até ao *Centroid* comparada com o *cluster* anterior, com base na “Redução” podemos analisar melhor qual é vantagem em termos percentuais em passar para próximo *cluster*:

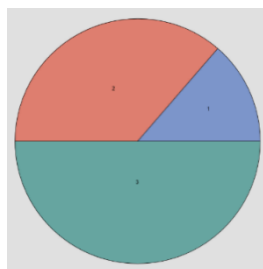
$$\text{Redução} = \frac{[\text{Distância até Centroid do (Cluster-1)}] - [\text{Distância até Centroid do (Cluster)}]}{[\text{Distância até Centroid do (Cluster-1)}]}$$

- de 1 para 2 (24%) → OK redução grande, ficamos com 2 *clusters*
- de 2 para 3 (13%) → OK redução grande, ficamos com 3 *clusters*
- de 3 para 4 (6%) → 6% já é considerado como uma pequena diferença, ou seja, seria viável ficarmos com 3 *clusters*, no entanto vamos prosseguir, ficamos com 4 *clusters*, 1ª redução mínima.
- de 4 para 5 (3%) → 3% é considerado como uma muito pequena diferença, ou seja, seria viável ficarmos com 4 *clusters*, e neste caso não faz sentido prosseguir, pois é 2ª redução mínima.

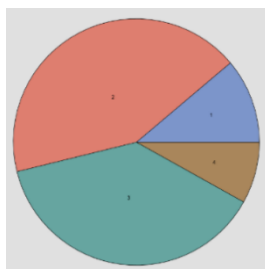
Chegou-se à conclusão de que se devia ter entre 3 a 4 *clusters*, no entanto tivemos curiosidade em testar também o modelos para 5 *clusters*.

Tamanho e a dispersão dos *clusters*:

3 Clusters (1º cenário)



4 Clusters (1º cenário)



5 Clusters (1º cenário)

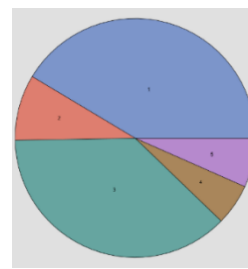


Figura 21 – Tamanho dos clusters para cada cenário

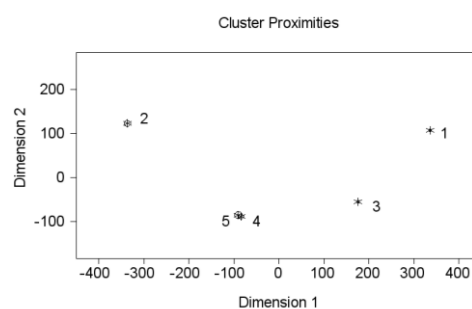
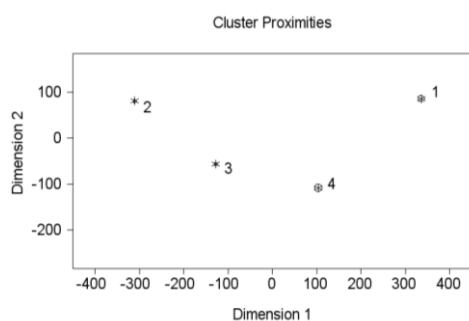


Figura 22 – Proximidades de clusters para K=4 e K=5

Pelo gráfico da dispersão dos *clusters*, podemos ver que no caso de 5 *clusters*, temos os grupos 5 e 4 muito próximos. Vamos analisar a distribuição das médias de cada produto por vários cenários de *clustering*:

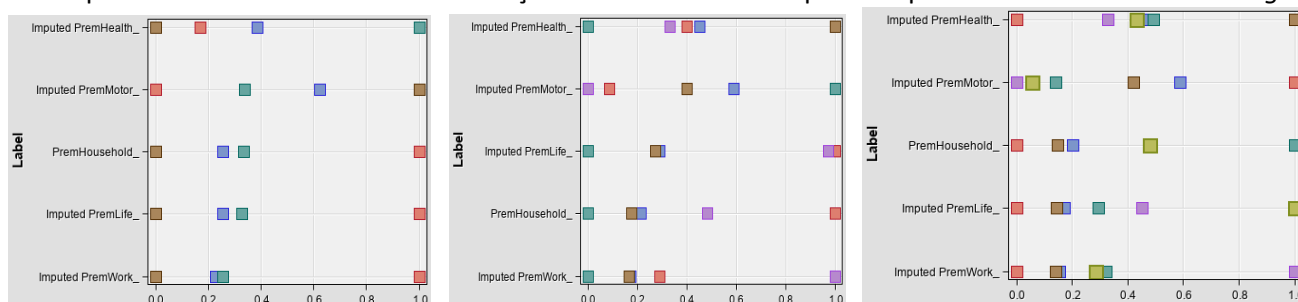


Figura 23 – Dispersão de clusters, em média, por cada produto apresentado

Segmento	Freq.	%	PremHealth	PremLife	PremMotor	PremWork	PremHousehold	Os seguros mais altos
1	1379	13%	143,45 €	112,27 €	105,95 €	118,08 €	555,60 €	Vida+Trabalho+Lar
2	3734	37%	237,06 €	49,18 €	209,16 €	44,52 €	248,98 €	Saúde
3	5114	50%	124,08 €	18,35 €	412,41 €	19,25 €	94,86 €	Carro
1	1132	11%	161,75 €	103,19 €	135,49 €	58,17 €	657,94 €	Vida+Lar
2	4381	43%	112,64 €	17,15 €	428,05 €	17,55 €	90,45 €	Carro
3	3904	38%	234,82 €	40,65 €	235,86 €	40,53 €	190,36 €	Saúde
4	810	8%	153,36 €	100,84 €	107,10 €	156,21 €	366,34 €	Vida+Trabalho
1	4228	41%	110,32 €	16,87 €	431,44 €	17,21 €	88,89 €	Carro
2	919	9%	171,31 €	60,60 €	150,46 €	67,78 €	700,31 €	Lar
3	3839	38%	233,87 €	38,21 €	242,63 €	39,43 €	178,75 €	Saúde
4	570	6%	151,09 €	83,74 €	104,59 €	174,49 €	384,72 €	Trabalho
5	671	7%	163,55 €	165,25 €	123,29 €	61,99 €	381,60 €	Vida

Tabela 9

Para ajudar a análise aplicamos formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixo em tons de verde. O escalão de cores foi aplicado para cada tipo de seguro, para comparar os valores médios dos prémios a nível do mesmo seguro.

- Tamanho dos clusters:
 - **Carro:** Podemos observar que em todos os cenários o maior *cluster* é o *cluster* dos clientes que apostam no seguro de carro. No 1º cenário o cluster representa 50% dos clientes, no 2º cenário 43% e no 3º cenário 41%. O comportamento desses clientes é igual em todos os cenários, ou seja, eles pagam um prémio elevado de carro e um prémio baixo em restantes seguros;
 - **Saúde:** O 2º maior *cluster*, é o mesmo *cluster* em todos os cenários, que é o cluster dos clientes que apostam no seguro de saúde. É de notar que a percentagem dos clientes em todos os cenários é praticamente igual, 37%-38%. No entanto o comportamento dos clientes do 1º cenário difere dos restantes cenários. No 1º cenário esses clientes pagam um prémio elevado no seguro de saúde, e pagam um prémio médio nos restantes seguros. No 2º e 3º cenário podemos ver que entre os restantes seguros destaca-se o seguro de carro. Ou seja, 2º e 3º cenário descrevem melhor o comportamento do cliente;
 - **Vida+Trabalho+Lar:** No 1º cenário os clientes que apostam juntamente em seguros de Vida, Trabalho e Lar, são unidos no mesmo grupo. No 2º cenário houve uma separação dos clientes Vida+Lar e Vida+Trabalho e no 3º cenário houve uma separação dos clientes por cada tipo de prémios. Podemos ver que o comportamento dos clientes em termos de seguros de carro e de saúde é sempre o mesmo em todos os cenários, ou seja, o de carro é mais baixo e o de saúde médio. No entanto, o comportamento dos clientes perante os restantes seguros não é nítido no 1º cenário, pois são todos elevados. No 2º e 3º cenários, já se consegue melhor, visualizar a descrição do comportamento dos clientes. É lógico que o 3º cenário, representa de forma mais nítida o comportamento dos clientes no entanto, neste último cenário, existem alguns grupos formados que têm um tamanho muito pequeno e isso iria dificultar a análise com os restantes indicadores;

- Dispersão dos *clusters*: Podemos observar que efetivamente os clientes dos grupos 4 e 5 do 3º cenário têm comportamentos muito parecidos na maior parte das variáveis, sendo que, diferem radicalmente apenas na posição perante o seguro de trabalho, e também nota-se a diferença entre os prémios de vida. Achou-se então que, entre os 3 cenários o que representa melhor o comportamento do cliente é o 2º.

Importância das Variáveis

3 clusters (1ºcenário)		4 clusters (2ºcenário)		5 clusters (3ºcenário)	
PremMotor	1,000	PremMotor	1,000	PremMotor	1,000
PremWork	0,917	PremHealth	0,950	PremHealth	0,976
PremHealth	0,916	PremLife	0,860	PremWork	0,884
PremLife	0,915	PremWork	0,855	PremLife	0,883
PremHousehold	0,911	PremHousehold	0,848	PremHousehold	0,809

Tabela 10

Para se perceber melhor qual das variáveis, influência mais a tomada de decisão na formação de *clusters*, analisou-se a importância de cada uma das variáveis dentro de cada cenário. Podemos ver em todos os cenários que a variável Prémio de Carro, e última é a de Prémio de Lar (é explicado pelo facto de que o prémio de lar, em média, é o prémio mais caro entre todos os prémios dentro de cada cluster). Podemos ver que no 1º cenário, para além do prémio de carro, as restantes variáveis referentes a outros prémios, têm pesos muito próximos na decisão da segmentação, o que não é considerado um bom indicador. O 2º e 3º cenários diferem apenas na prioridade entre o prémio de vida e prémio trabalho.

Optamos por considerar os 4 *clusters*:

Cluster	Nº Obs.	Cluster Próximo	Prémio Saúde	Prémio Vida	Prémio Carro	Prémio Trabalho	Prémio Lar	Seguro Nº1	Prémio total médio
1	1132	4	161,75 €	103,19 €	135,49 €	58,17 €	657,94 €	Vida+Lar	1 116,54 €
2	4381	3	112,64 €	17,15 €	428,05 €	17,55 €	90,45 €	Carro	665,83 €
3	3904	2	234,82 €	40,65 €	235,86 €	40,53 €	190,36 €	Saude	742,22 €
4	810	1	153,36 €	100,84 €	107,10 €	156,21 €	366,34 €	Vida+Trabalho	883,85 €

Tabela 11

A tabela 11 apresenta o prémio médio pago pelo cada tipo de seguro distribuído por clusters. Para ajudar na análise aplicou-se a formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixos em tons de verde. O escalão de cores foi aplicado para cada tipo de seguro, para comparar os valores médios dos prémios por segmento a nível do mesmo seguro. Identificamos as seguintes características para cada cluster:

- Cluster 1
- Clientes que apostam em seguro de Vida e Lar
 - Gastam menos no prémio de Carro
 - Pagam um prémio médio nos em Saúde e Trabalho
 - O prémio médio total é o mais elevado entre todos os clusters.
- Cluster 2
- Clientes que apostam em seguro de Carro
 - Gastam pouco nos seguros Vida, Lar, Saúde e Trabalho
 - O seu prémio médio total é o mais baixo entre todos os clusters
- Cluster 3
- Clientes que apostam em seguro de Saúde
 - Gastam menos nos prémios nos seguros de Lar, Vida e Trabalho
 - Pagam um prémio médio em seguro de Carro

Segmento:
Vida+Lar

Segmento:
Carro

Segmento:
Saude

- O seu prémio médio total é ligeiramente superior em relação ao Cluster 2

Cluster 4

- Clientes que apostam em seguro de Vida e Trabalho
- Gastam menos no prémio de Carro
- Pagam um prémio médio nos seguros de Lar e Saúde
- O seu prémio médio total é o 2º mais elevado entre todos os clusters

Segmento:
Vida+Trabalho

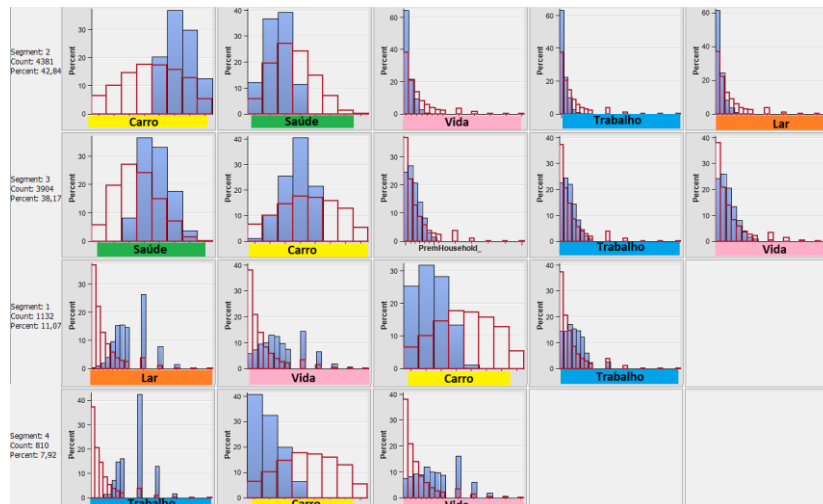


Figura 24 – Comparação da distribuição dos segmentos obtidos com a distribuição das variáveis de produtos

Análise das variáveis segundo os Segmentos

Salário

Com base na variável **MonthSal** achamos que seria importante analisar os vários escalões de rendimento dentro de cada um dos segmentos, para isso optamos por analisar o histograma:

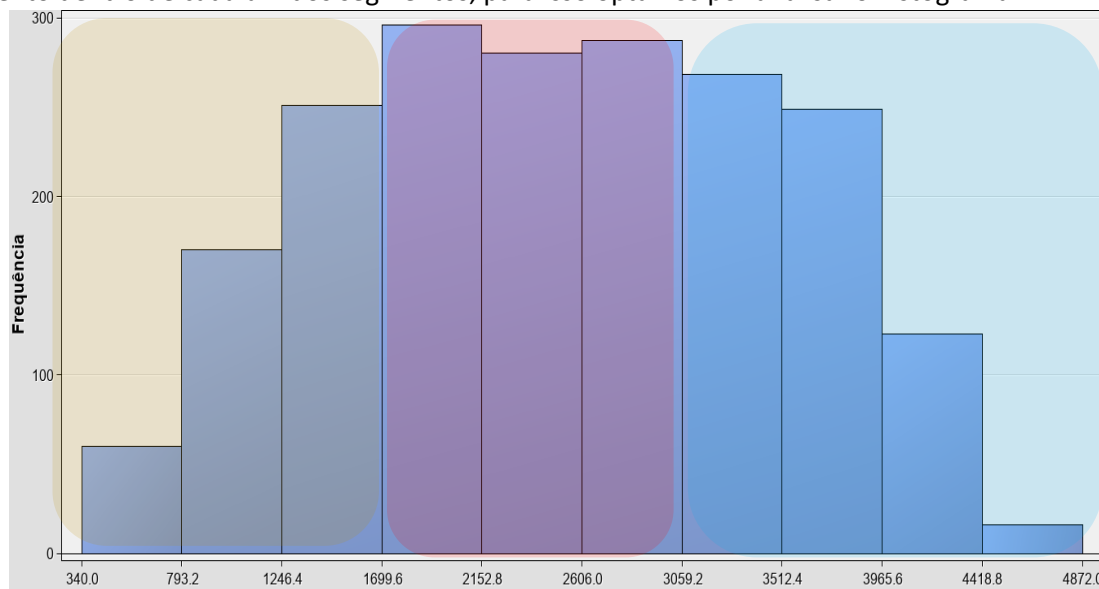


Figura 25 – Salário bruto mensal por cliente e por escalão de rendimento

Formamos os seguintes grupos com os respetivos limites:

- Salário Baixo: clientes com ordenado abaixo da média [333:1700[→ 2531 clientes (25%)
- Salário Médio: clientes com ordenado médio [1700:3060[→ 4405 clientes (43%)

- Salário Alto: clientes com ordenado acima da média [3060:5021] → 3291 clientes (32%)

Ordenado	Cluster				Total de Clientes
	Vida+Lar (1)	Carro (2)	Saude (3)	Vida+Trabalho (4)	
Salário Baixo	598	454	951	528	2531
Salário Médio	235	2733	1297	140	4405
Salário Alto	296	1194	1647	142	3291
Total de Clientes	1132	4381	3904	810	10227

Tabela 12

	Vida+Lar (1)	Carro (2)	Saude (3)	Vida+Trabalho (4)
Salário Baixo	53%	10%	24%	65%
Salário Médio	21%	62%	33%	17%
Salário Alto	26%	27%	42%	18%

Tabela 13

	Vida+Lar (1)	Carro (2)	Saude (3)	Vida+Trabalho (4)
Ordenado Médio	2 024,35 €	2 627,15 €	2 643,23 €	1 750,34 €

Tabela 14

Para ajudar a análise foi aplicada a formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixo em tons de verde. O escalão de cores foi aplicado para cada cluster na tabela12 e global nas restantes duas tabelas. Podemos observar que a maior parte dos clientes dos segmentos “Vida+Trabalho” e “Vida+Lar” têm ordenados baixos. Nos segmentos “Carro” e “Saúde” podemos ver que a maioria dos clientes têm ordenado médio ou alto.

Valor do Cliente

	OK	Pré Alerta	Alerta	Total de Clientes
Vida+Lar (1)	787	136	209	1132
Carro (2)	2654	632	1095	4381
Saude (3)	2727	474	703	3904
Vida+Trabalho (4)	544	60	206	810
Total de Clientes	6712	1302	2213	10227

Tabela 15

	OK	Pré Alerta	Alerta
Vida+Lar (1)	70%	12%	18%
Carro (2)	61%	14%	25%
Saude (3)	70%	12%	18%
Vida+Trabalho (4)	67%	7%	25%

Tabela 16

Para ajudar a análise aplicou-se a formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixo em tons de verde. O escalão de cores foi aplicado globalmente na tabela 15 e por classificação na tabela 16. Foi identificado, que, a maior parte dos clientes são rentáveis. A percentagem de “Pré-Alerta” é baixa, no entanto no segmento “Carro” é o mais destacável. Consideramos que a percentagem de clientes com “Alerta” (25%) é alta em segmentos “Carro” e “Vida+Trabalho”. Porém, o segmento “Carro” é consideravelmente maior do que o de “Vida+Trabalho” e representam quase 50% de todos os clientes com “Alerta”. Em contrapartida, os segmentos que apresentam mais clientes rentáveis, em termos proporcionais dentro de cada segmento são “Vida+Lar” e “Saúde”, adicionalmente é de referir, que apesar de tudo, o segmento “Carro” apresenta 2654 clientes rentáveis, que são 26% de todos os clientes.

Percentagem do Salário gasto em Percentagem

Apesar de existir correlação negativa, 63% entre as variáveis: o salário e a percentagem gasta em seguro, não foi considerado muito forte. No entanto, pode-se dizer que as duas variáveis são influenciadas uma pela outra, ou seja, à medida que o ordenado dos clientes aumenta a percentagem paga em seguro diminui, isso pode querer dizer que, em média todos os clientes pagam pelo menos um valor mínimo, independentemente do ordenado que recebem (prémio mínimo = 353,55€ e o ordenado mínimo = 333 €). Quando o ordenado é muito baixo esse valor mínimo representa uma maior proporção do ordenado, quando o ordenado é alto, esse valor mínimo consequentemente apresenta uma menor proporção. No entanto, achamos que seria interessante observar este indicador dentro de um grupo com os ordenados médios ou altos.

	Vida+Lar (1)		Carro (2)		Saude (3)		Vida+Trabalho (4)	
	% Gasta em Seguro (média)	Nº de Clientes	% Gasta em Seguro (média)	Nº de Clientes	% Gasta em Seguro (media)	Nº de Clientes	% Gasta em Seguro (media)	Nº de Clientes
Salário Baixo	122%	598	50%	454	63%	951	100%	528
Salário Médio	46%	235	28%	2733	33%	1297	36%	140
Salário Alto	29%	299	19%	1194	20%	1656	22%	142

Tabela 17

Para ajudar na análise aplicamos formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixos em tons de verde. O escalão de cores foi aplicado por classificação do salário para comparar as percentagens dentro do mesmo universo de ordenados. Observou-se que o segmento “Vida+Lar” tende a gastar mais do seu ordenado em termos proporcionais em relação aos outros segmentos, e o segmento “Carro” é o que gasta menos.

Idade e Presença de Filhos

Já foi referido anteriormente que a idade dos clientes é fortemente correlacionada positivamente com o ordenado mensal (92%). Ou seja, à medida que a idade aumenta, o salário do cliente também aumenta. No entanto resolveu-se realizar uma análise para perceber o comportamento da idade e da presença dos filhos distribuída por segmentos:

	Salário Baixo		Salário Médio		Salário Alto	
	% Filhos=1	Idade (% por segment)	% Filhos=1	Idade (% por segment)	%Filhos=1	Idade (% por segment)
Vida+Lar	76%	23 (53%)	76%	43 (21%)	27%	67 (26%)
Carro	95%	33 (10%)	92%	47 (62%)	62%	64 (27%)
Saude	91%	29 (24%)	86%	43 (33%)	20%	69 (42%)
Vida+Trabalho	73%	22 (65%)	78%	43 (17%)	31%	68 (18%)

Tabela 18

Para ajudar a análise foi aplicada a formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixos em tons de verde. O escalão de cores foi aplicado globalmente no caso da presença dos filhos e por segmentos na idade. Verifica-se que a proporção de clientes com filhos é maior em universos com ordenados mais baixos, e menor em universo de ordenados altos. Podemos ver que não existe diferença quase nenhuma entre os clientes com ordenados baixos e ordenados médios no que diz respeito à presença de filhos, adicionalmente é de referir que, no segmento “Carro” mesmo os clientes com ordenados altos, a maior parte, tem filhos. Em relação à idade, clientes dos segmentos “Vida+Lar” e “Vida+Trabalho” têm tendência de ter mais clientes novos, e “Carro” e “Saúde” é exatamente ao contrário.

Zona de Residência, Educação e Antiguidade do Cliente

	Zona 1	Zona 2	Zona 3	Zona 4	
Vida+Lar	30%	9%	20%	41%	100%
Carro	30%	10%	21%	40%	100%
Saude	30%	10%	19%	41%	100%
Vida+Trabalho	29%	10%	21%	40%	100%

Tabela 19

	1 - Basic	2 - High School	3 - BSc/MSc	4 - PhD	
Vida+Lar	35%	47%	17%	0%	100%
Carro	4%	25%	59%	11%	100%
Saude	8%	39%	48%	5%	100%
Vida+Trabalho	45%	43%	12%	0%	100%

Tabela 20

Segmento	Média de Antiguidade	Média de Idade
Vida+Lar	30	39
Carro	30	50
Saude	30	51
Vida+Trabalho	30	34

Tabela 21

Para ajudar a análise aplicou-se formatação condicional pelo escalão de cores, em que os valores mais altos são apresentados em tons de vermelho, os intermédios em tons de amarelo e os valores mais baixo em tons de verde. O escalão de cores foi aplicado por segmento. A tabela 19 apresenta a percentagem dos clientes por segmento e residência. Pode-se ver que a distribuição das zonas por segmentos é idêntica, ou seja, esta variável não vai poder trazer valor para a segmentação do cliente. A tabela 20 apresenta a percentagem dos clientes por segmento e escolaridade. Observa-se que os clientes doutorados (4 - PhD) são uma minoria em todos os segmentos exceto no segmento “Carro”. Os segmentos “Vida+Lar” e “Vida+Trabalho” apresentam comportamentos muito similares, sendo que a maior parte dos clientes têm escolaridade “Basic” ou “High School”. Os segmentos “Carro” e “Saude” apresentam comportamentos muito similares, pois a maioria dos clientes têm escolaridade “High School” ou “BSc/MSc”.

No que diz respeito à “Antiguidade”, apresentado na Tabela 21, verifica-se que não existe qualquer diferença entre os segmentos. É de referir que não é muito coerente com a idade dos clientes, pois apesar de a idade aumentar a antiguidade do cliente na seguradora mantém se estável. Aachamos que a variável “Antiguidade” não vai trazer valor à análise.

Perfil do Cliente





	Segmento 1	Segmento 2	Segmento 3	Segmento 4
	Vida+Lar	Carro	Saude	Vida+Trabalho
Seguro 1º				
Outros Seguros	Carro - Saude e Trabalho +	Vida, Lar, Saude e Trabalho -	Vida, Lar e Trabalho - Carro +	Carro, Saude e Lar +
Edução	Basic + High School	High School+BSc/MSc+ PhD	High School+BSc/MSc	Basic + High School
Filhos	63%	84%	59%	67%
Idade Média	39	50	51	34
Ordenado Médio	2 024 € Médio	2 627 € Alto	2 643 € Alto	1 750 € Baixo
Valor do Cliente	70% OK ✓ () 18% Alerta !	61% OK ✓ 25% Alerta !	70% OK ✓ 18% Alerta !	67% OK ✓ 25% Alerta !
% do rendimento gasta em Seguros	Gasto Alto	Gasto Baixo	Gasto Médio	Gasto Médio
Prémio Total	1 116 € Alto	665 € Baixo	742 € Baixo	883 € Gasto Médio

Tabela 22

Insights

- À medida que o cliente for mais velho maior é a probabilidade de ter ordenado alto e não ter filhos, única exceção são os clientes que apostam em seguro de carro, essas têm uma grande probabilidade de ter filhos independentemente da idade.
- Os clientes que apostam em seguro de **carro** gastam menos (em termos proporcionais por tipo de seguro) dos restantes prémios. Em média são clientes mais velhos.
- Os clientes que apostam em seguro de **saúde** apostam também no seguro de carro e gastam menos (em termos proporcionais por tipo de seguro) dos restantes prémios. Em média são clientes mais velhos.
- Os clientes do segmento “**Vida+Trabalho**” são os clientes com os ordenados mais baixos mas que apostam em todos os prémios. Em média são clientes mais novos.
- Os clientes do segmento “**Vida+Lar**” são os clientes com os ordenados médios. Em média são clientes mais novos que apesar de não apostar no seguro de carro, apostam em restantes prémios e são clientes que gastam mais em seguros entre todos.

Conclusão

O objetivo da análise das variáveis fornecidas foi para se caracterizar os vários tipos de clientes que a seguradora tem na sua base de dados, para que o departamento de marketing consiga direcionar as várias campanhas que tem para os clientes que melhor poderão tirar proveito.

A realização da caracterização dos tipos de clientes que a seguradora tem, foi calculada tendo em conta as 5 variáveis que representam os prémios que a seguradora possui, ou seja, os Prémios de Vida, Lar, Saúde, Carro e Trabalho. Com a realização do *clustering* verificou-se que seria possível ter clusters com 5 segmentos, 4 segmentos e 3 segmentos, sendo que a diferença seria o agrupamento dos prémios por segmento. Foi decidido agrupar os tipos de clientes em 4 segmentos.

Foi também averiguada a possibilidade de ao retirar-se todos os outliers e não apenas os severos. Realizou-se a análise de correlações e a classificação das variáveis que seriam excluídas do processo, e a conclusão a que se chegou foi que eram escolhidas as mesmas variáveis para o processo de *clustering*, sendo estas as que representam os prémios do seguro. Também foi averiguada se em algum momento a escolha do número de segmentos e a disposição deste se alterava com a exclusão de todos os *outliers*, e veio a ser provado que em nada alterava a nossa análise, pelo que se optou por apenas retirar os *outliers* severos, sendo que não se retiraram registos sem a mínima justificação.

Foi apresentada uma solução em que a segmentação foi decidida pelo algoritmo k-means com base nas 5 variáveis referentes aos pagamentos de prémios. De seguida, em cada um dos segmentos foi efetuada uma análise sobre as suas características globais de forma que se consiga perceber com as restantes variáveis os diferentes tipos de clientes que o departamento de marketing poderá utilizar para o lançamento de campanhas que sejam direcionadas a subconjuntos específicos aqui retratados.

Limitações

O objetivo desta análise apresentada no relatório foi analisar o comportamento do cliente perante os produtos disponíveis na seguradora em primeiro lugar, pois é o negócio da seguradora, sendo que de seguida foi analisado o perfil em termos das características gerais de clientes. Assim sendo a limitação que tivemos foi referente às características gerais do cliente, ou seja, em cada segmento existem clientes com diversos intervalos de ordenados, idades, educação... etc., o que dificultou a análise de perfil. Por outro lado, dentro de cada segmento conseguiu-se concluir que os clientes têm um comportamento muito parecido relativamente aos tipos de seguros, e também se verifica que em cada segmento existe uma tendência a nível das características gerais do cliente.

Anexo

Diagrama SAS

