P.PORTO

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO

Dinis Pinto, Rodrigo Oliveira, Nuno Cruz, Pedro Flores

ESTG, Felgueiras, Portugal, 8230619@estg.ipp.pt

ESTG, Felgueiras, Portugal, 8230131@estg.ipp.pt

ESTG, Felgueiras, Portugal, 8240330@estg.ipp.pt

ESTG, Felgueiras, Portugal, 8240407@estg.ipp.pt

Trabalho Prático

Métodos Quantitativos de Apoio à Decisão (2025/2026)

Sumário

Este trabalho prático apresenta uma análise detalhada sobre a eficiência no setor logístico, tendo como base a base de dados "Smart Logistics", aqui foram analisados os fatores que podem levar as entregas a serem feitas com atraso e quais os padrões encontrados. A metodologia base utilizada foi a CRISP-DM, garantindo assim o rigor analítico e o entendimento de todo o negócio.

Foi feita uma análise inicial para perceber o equilíbrio das variáveis, que se demonstrou positiva, foi também feita uma análise para perceber se existia a necessidade de limpar os dados, retirando dados duplicados ou substituindo dados omisso, não foi necessário, a base de dados já está limpa. De seguida foram aplicados modelos estatísticos, inicialmente desenvolvendo um modelo de regressão logística para prever a probabilidade de atraso nas entregas. Este modelo não se demonstrou muito efetivo, apesar de apresentar uma especificidade de 98% na identificação de camiões com atraso, apresentou um poder explicativo global muito baixo (Pseudo R-quadrado de apenas 0,71%), demonstrando assim que os atrasos são influenciados por fatores externos não capturados pelo modelo.

Para complementar a análise, realizou-se também uma análise de clusters hierárquicos, com foco na sazonalidade, analisando Julho e Dezembro. A análise dos resultados mostra a existência de dois perfis distintos:

- Um perfil caracterizado por maior valor financeiro e maior intensidade de utilização de ativos, que mostra uma taxa superior de atrasos.
- Outro perfil caracterizado por valores mais pequenos e níveis de inventários diferentes, com uma maior estabilidade temporal.

As conclusões deste estudo revelam que o modelo de regressão logística desenvolvido não consegue captar as variáveis que explicam as ocorrências de atraso, pode ser importante para o futuro, fazer o monitoramento de mais variáveis e talvez considerar outros modelos, como modelos de Machine Learning.

Índice

1. Introdução

Atualmente, o setor da logística é um pilar fundamental para a economia global, sendo uma parte importante no processo da globalização. Sendo assim, este setor necessita de uma eficiência operacional cada vez mais afiada, para isso, é necessário conseguir monitorizar os ativos ou camiões de forma precisa e de preferência em tempo real, para conseguir fazer adaptações necessárias de modo a obter o melhor resulta, entregas mais rápidas.

A perceção das variáveis e condições associadas ao setor logístico deve ser melhorada através de análises estatísticas, permitindo criar modelos de previsão e perceber quais os fatores mais importantes para as ocorrências de atraso, podendo assim mexer naquelas em que se pode manipular ou alterando as datas e desenvolver técnicas de mitigação de riscos de atrasos caso as variáveis identificadas como críticas estejam fora do nosso controle.

Este trabalho surge no contexto na unidade curricular de Métodos Quantitativos de Apoio à Decisão, sendo a base de dados utilizada "Smart Logistics", o principal foco desta análise, é identificar padrões que estejam associados ao atraso das entregas e perceber quais variáveis se demonstram como críticas que impactam o bom funcionamento e eficiência da operação logística.

Para garantir o rigor analítico, este trabalho segue a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining)*, esta metodologia é estruturada em seis fases, desde a compreensão do negócio até a implementação dos modelos. O trabalho tem duas fases principais, o desenvolvimento de um modelo de previsão, modelo de regressão logística, e análise de clusters.

Através desta análise, pretende-se melhorar a tomada de decisão, fornecendo dados e previsões relevantes acerca da realidade. Tentando diminuir o risco de atraso e falha no processo logístico.

2. Metodologia CRISP-DM

2.1. Compreensão do Negócio

O setor logístico necessita de garantir rapidez nas entregas, devendo evitar ao máximo atrasos, para isso é necessária uma eficiência na gestão das rotas e dos seus ativos (camiões).

O objetivo principal desta análise é perceber quais os fatores e variáveis afetam a eficiência das entregas e por consequência levam ao atraso nas mesmas, podendo assim, melhorar a gestão operacional recorrendo a previsões de atraso tendo em conta algumas variáveis.

Para responder a este objetivo, vai ser utilizado um modelo de regressão logística, para tentar estimar a probabilidade de uma dada entrega se atrasar em função das variáveis consideradas. Para além disso vai também ser feito um Clustering, para tenar descobrir padrões sazonais, entre inverno e verão. Mas antes disto tudo, será feita uma validação dos dados e identificação de possíveis enviesamentos.

O sucesso desta análise será medido pela capacidade do modelo preditivo, conseguir de facto identificar que 56% das entregas são feitas com atraso. Esta compreensão permitirá aos gestores adotar medidas e políticas de contenção destes atrasos.

2.2. Compreensão dos Dados

A base de dados analisada, conta com 1000 registos de 18 variáveis, estas variáveis representam diferentes áreas do negócio, desde o número de ativos (camiões) ao estado do trânsito.

2.2.1. Qualidade e limpeza dos dados

Antes de qualquer análise estatística, procedeu-se à verificação da qualidade dos dados.

- Verificação de dados omissos: foi feita uma procura por dados omissos ou nulos, não foram encontrados nenhuns.
- Verificação de dados duplicados: procurou-se perceber se existiam dados duplicados, à semelhança do anterior, não forma encontrados quaisquer dados duplicados.

A ausência de dados omisso ou duplicados, permitiu avançar para as próximas fases sem ser necessária usar qualquer técnica de imputação ou remoção de dados, mantendo assim a mostra total de 1000 registos.

2.2.2. Exploração das variáveis críticas

A exploração das variáveis que mais podem influenciar os atrasos logísticos, é importante para identificar se podem existir enviesamentos nos dados.

2.2.2.1 Asset_ID

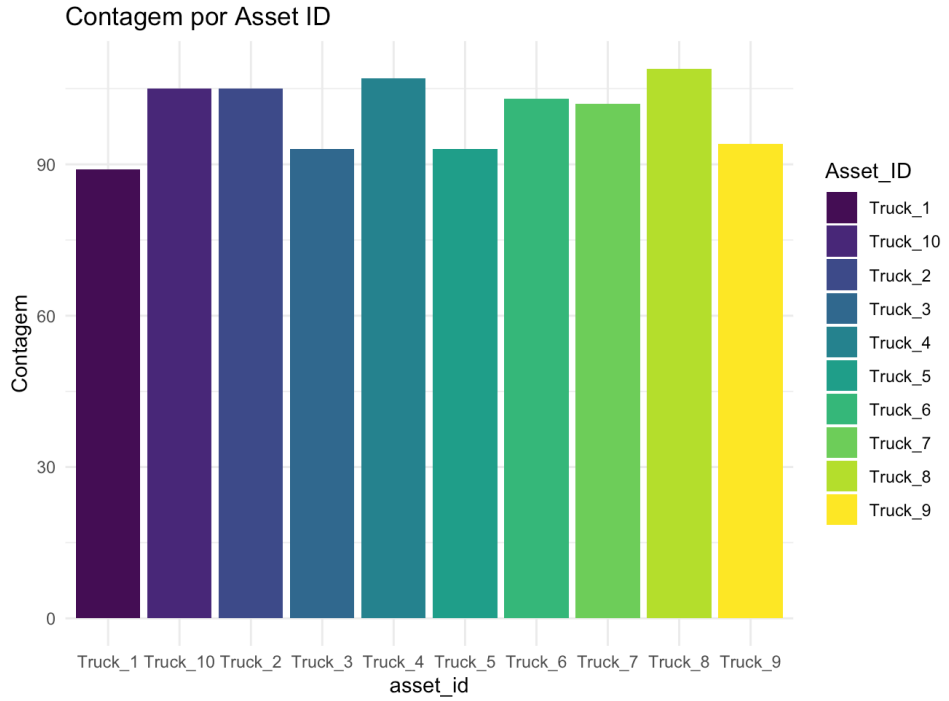

Figura 1: Contagem por Asset ID.

A variável "Asset_ID" representa o número atribuído a cada camião, é uma forma de perceber se existe algum camião que seja desproporcionalmente atrasado.

O gráfico de barras abaixo (Figura 1), faz a contagem do número de entregas por camião, como pode ser observado, o número de entregas por camião é bastante parecido, não existindo nenhum camião que faça muitas mais entregas que outro, o que evita que exista enviesamento na análise posterior de atrasos.

2.2.2.2 Inventory_Level

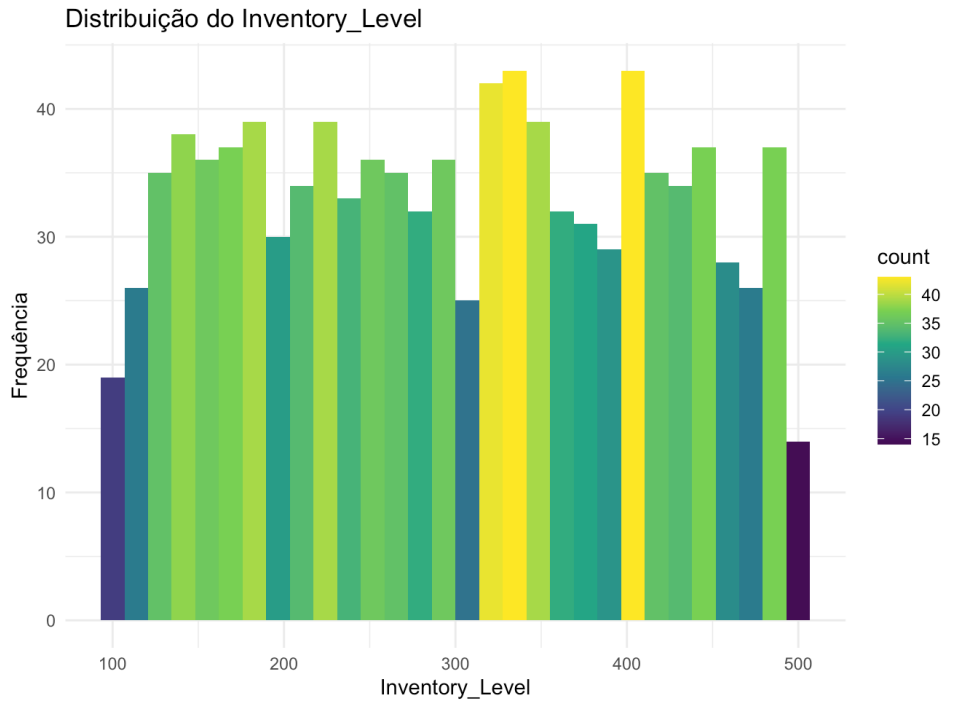

Figura 2: Distribuição do Inventory_Level.

A variável "Inventory_Level" corresponde ao nível de stock presente em armazém. O histograma (Figura 2) associado a esta variável apresente valeres entre 100 e 500, apesar de ser uma distribuição grande, a ocorrência dos valores é relativamente nivelada, com exceção do mínimo e do máximo. Esta variável poderá ser importar tene para explicar possíveis atrasos.

2.2.2.3 Temperature

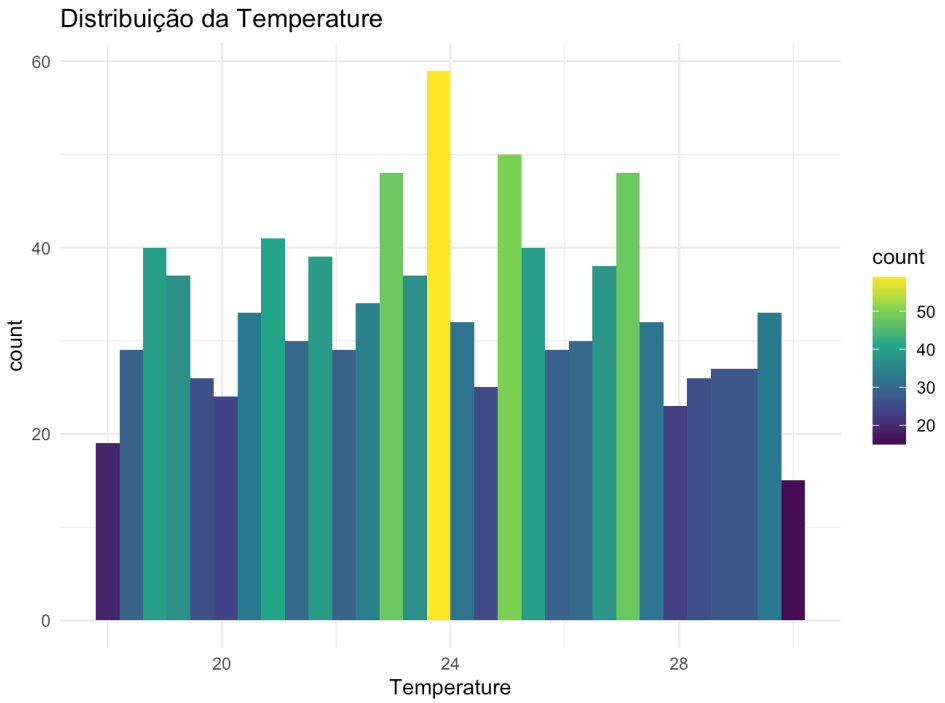

Figura 3: Distribuição da Temperature.

A "variável Temperature" representa a temperatura ambiente no dia da entrega. Pelo histograma (Figura 3) podemos perceber que a temperatura mínima foi 18 graus e máxima de 30 graus, com maior frequência temos cerca de 24 graus. Temperaturas diferentes podem influenciar a atividade logística de forma indireta, ligadas a fadiga do motorista e o transporte de bens sensíveis pode ser afetado pela temperatura, o que pode levar a adotar certos procedimentos que gastam tempo que seria utilizado na entrega.

2.2.2.4 Traffic_Status

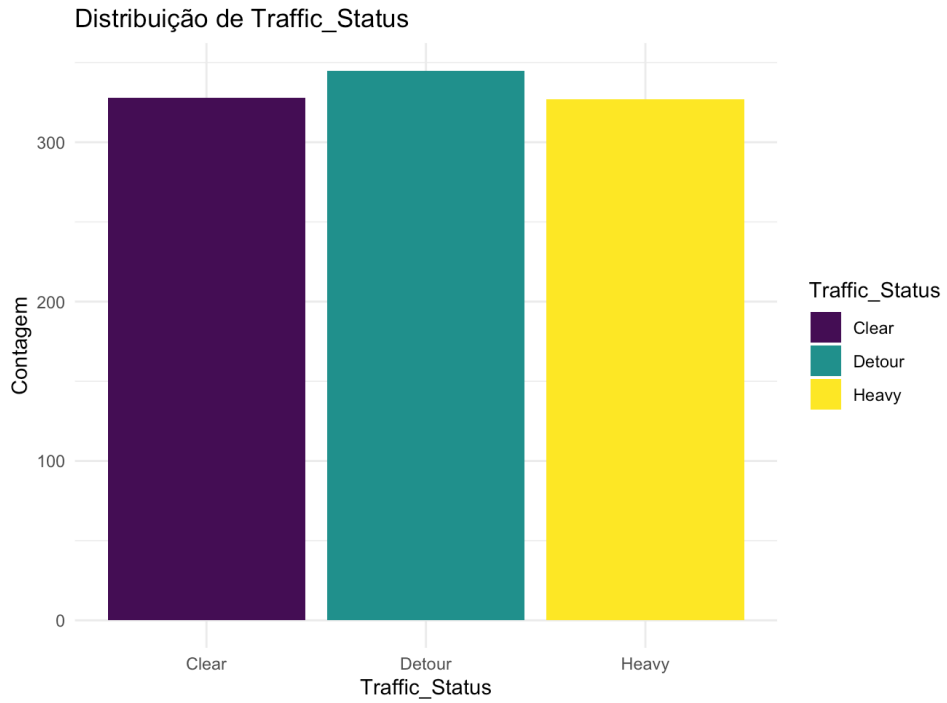

Figura 4: Distribuição de Traffic_Status.

A variável "Traffic_status" é uma variável categórica, dado que ela não apresenta valores numéricos, mas sim categorias Clear, Detour e Heavy. Esta varável representa o estado do trânsito no momento da entrega. Pelo gráfico (Figura 4) de barras podemos ver que existe uma distribuição quase uniforme, esta homogeneidade pode ser interessante pelo facto de assim existirem aproximadamente o mesmo número de ocorrências entre si e conseguimos assim medir bem se o estado do trânsito é ou não umas das causas de atraso. Se tivéssemos uma das categorias com uma ocorrência extremamente baixa, seria injusto compará-la com as outras, por esta ter uma amostra muito pequena e não representativa.

2.2.2.5 Logistics_Delay

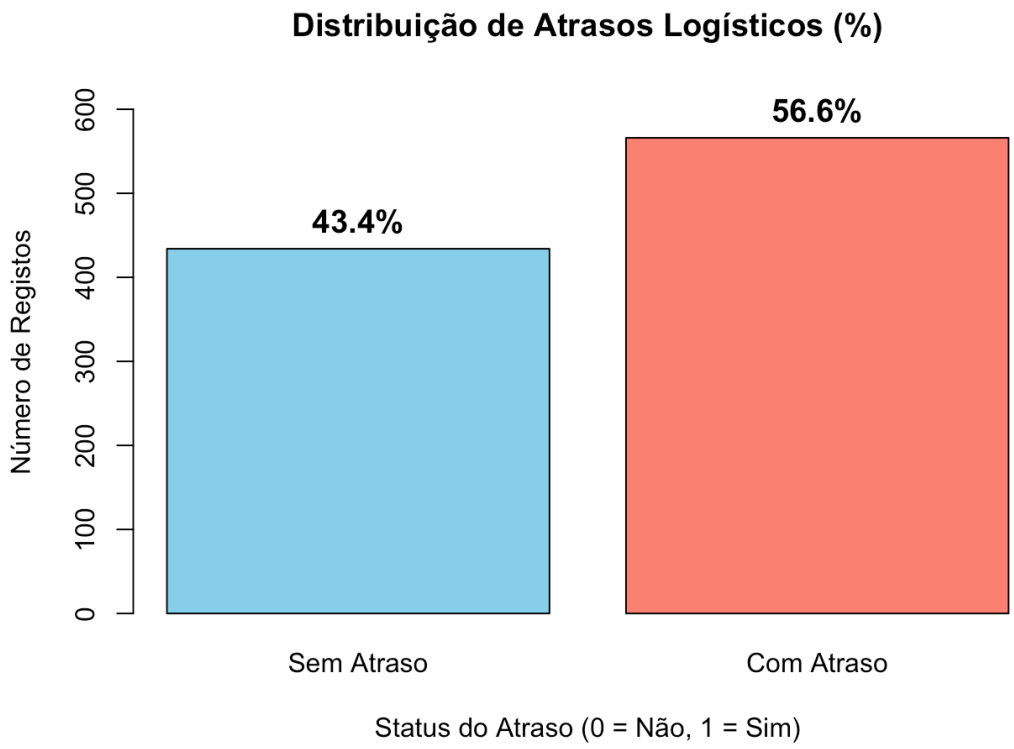

Figura 5: Distribuição de Atrasos Logísticos.

A variável "Logistics_delay", é a variável em estudo, ela é quem indica se existiu atraso (1) ou não (0). Através do gráfico (Figura 5) podemos observar que temos uma taxa de atraso, 56,6%, superior à taxa de não atraso, 43,4%, o que torna este estudo ainda mais importante, para tentar com o modelo de previsão, reduzir esta diferença e de preferência invertê-la.

2.2.2.6 Waiting time

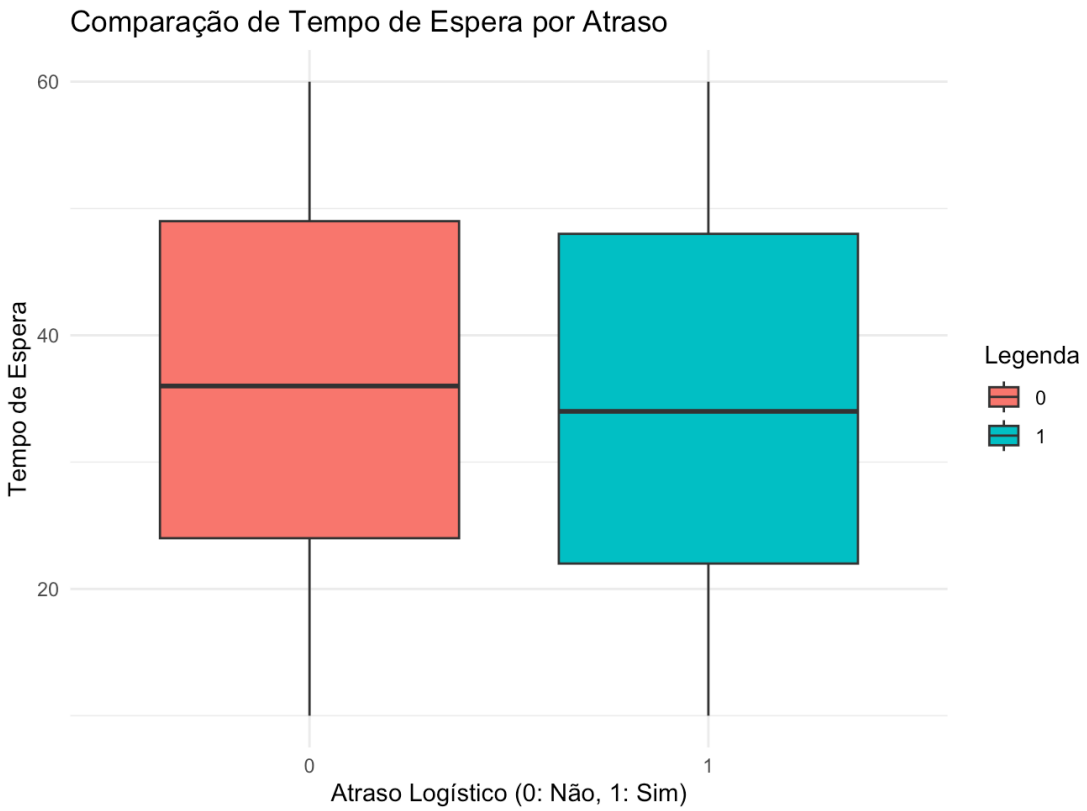

Figura 6: Comparação de Tempo de Espera por Atraso.

A variável "Waiting_Time" indica o tempo de espera que o camião teve no seu trajeto. Curiosamente, entregas sem atraso apresentaram um tempo de espera maior, o que não seria de esperar, já que se o camião esteve mais tempo parado, deveria atrasar.

2.2.3 Síntese da Análise Descritiva

A análise das variáveis com apoio dos gráficos, permitiu perceber que a base de dados é bastante equilibrada e tem uma boa variabilidade. Este equilíbrio é uma vantagem para a construção de modelos preditivos, dado que o equilíbrio e variabilidade podem evitar alguns enviesamentos indesejados. Podemos então prosseguir para as próximas etapas.

2.3. Criação de modelos de variáveis

Começamos por estudar e definir as variáveis que achamos significativas ao avaliar os atrasos da empresa. Escolhemos as variáveis "inventory_level" visto que o stock existente é fundamental para poder ou não expedir imediatamente a encomenda.

Escolhemos também "asset_id". Esta variável não é importante para verificar se há atraso ou não mas é muito importante para a empresa perceber com que frequência cada camião se atrasa.

Por outro lado, escolhemos "temperature" pelo facto de que, temperaturas baixas afetam o desempenho operacional. Em condições extremas, verifica-se uma alteração no conforto e bem-estar dos motoristas, o que pode impactar negativamente a motivação, ritmo de trabalho e capacidade de concentração.

Ecolhemos também "Traffic_status", que, como é lógico, tem interferência direta nos atrasos

Escolhemos também "Waiting_Time", porque, é o tempo que cada camião fica parado durante o seu trajeto, o que influencia no tempo total de entrega.

Escolhemos "Demand_Forecast" porque diz-nos a procura. Com a informação das previsões da procura, podemos controlar melhor os prazos e as entregas e é realmente importante para a existência ou não de atrasos.

2.4 Criação de modelo significativo

Para a criação deste modelo, utilizamos apenas as variáveis significativas

Após a criação do modelo significativo, a variável "Temperature" deixou de ser significativa.

- Modelo de regressão logística: $\frac{1}{1+e^{-(1{,}113441-0{,}008737\,WaitingTime)}}$

- Odds Ratio: $e^{-0{,}008737}$= 0,991. 1-0,991= 0,001. Se o Waiting Time mudar em 1 unidade, a probabilidade de atraso muda 0,1%.

- McFadden = 3,911e-03

- r2ML (Cox & Snell R2) = 5,3395E-03

- r2CU (Nagelkerke R2) = 7.161431e-03

Os coeficientes de determinação (Pseudo R-quadrados) obtidos indicam que o modelo possui um **poder explicativo muito baixo**. O valor de Nagelkerke (0,71%) e o de McFadden (0,39%) sugerem que as variáveis independentes incluídas no modelo explicam menos de 1% da variabilidade da ocorrência de atrasos nos camiões

2.5 Matriz de confusão

	Referência	
Previsão	0	1
0	16	11
1	418	555

Predição 0 (não atraso) ; Predição 1 (atraso)

0-0, verdadeira negativa - existem 16 camiões que foram classificados corretamente com atrasos

0-1, falso negativo - existem 11 casos que foram classificados como não atrasado, mas na verdade atrasaram

1-0, falso positivo - existem 418 classificados como atrasados, mas na verdade não atrasaram

1-1, verdadeiro positivo - existem 555 classificados como atrasados, o que corresponde à realidade

- Acurácia = 0,571 - o modelo acertou 57,1%, o que representa uma taxa de acerto moderada. No entanto a acurácia por si só, por vezes, pode ser enganosa quando há uma distribuição desbalanceada entre classe.

- Sensibilidade = 0,03687 - o modelo identificou corretamente que 3,6% dos camiões não se atrasaram. A sensibilidade representa a proporção de verdadeiros negativos que foram corretamente identificados.

- Especificidade = 0.98057 - A especificidade é a capacidade do modelo de identificar

corretamente os camiões que se atrasaram. Neste caso é extremamente alta (98%), o que significa que o modelo encontra com facilidade os camiões que se atrasam.

- Valor Preditivo Positivo (Pos Pred Value ou Precisão) = 0,59259

Este valor representa a proporção de previsões corretas para os casos classificados como "não atraso". Aproximadamente 59,2% dos camiões classificados como não atraso realmente não se atrasaram.

- Valor Preditivo Negativo (Neg Pred Value) = 0.57040

Este valor representa a proporção de clientes que foram corretamente identificados

com atraso. O valor é moderado, o que significa que o modelo é mais ou menos confiável quando um camião se atrasa.

- Kappa = 0.0196 - O valor de Kappa compara a acurácia do modelo com o que seria esperado por acaso. Um valor próximo de zero ou negativo indica que o modelo não se está a sair melhor do que uma classificação aleatória, especialmente quando há desequilíbrio nas classes

- McNemar's Test P-Value = <2e-16 - O teste de McNemar avalia se há uma diferença significativa entre as classificações corretas e incorretas nas categorias de 0 e 1. Neste caso, o p-valor extremamente baixo (<2e-16) sugere que o modelo apresenta um desequilíbrio significativo entre as predições corretas para churn e não churn.

- Balanced Accuracy = 0.50872 - A acurácia balanceada é a média entre sensibilidade e especificidade. O valor de 0.50872 indica que o modelo está praticamente no nível de uma classificação aleatória, o que é uma consequência da especificidade extremamente baixa.

- Conclusão: O modelo é muito bom para prever camiões que realmente se atrasaram (Especificidade = 98%). No entanto te um mau desempenho a prever camiões que não se atrasaram (sensibilidade = 3,6%)

2.6 Curva de ROC

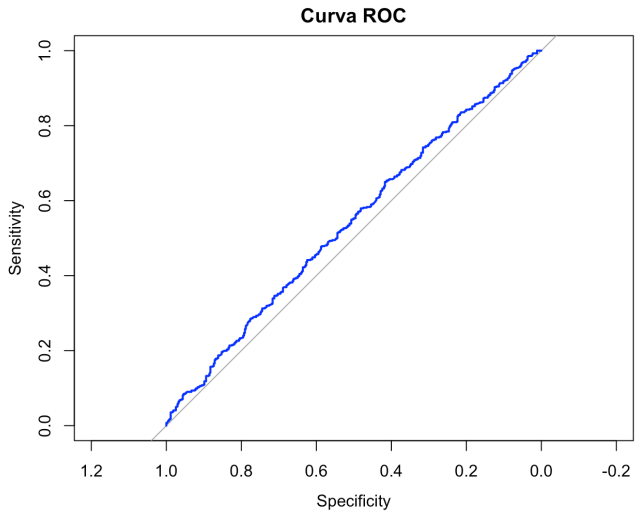

Figura 7: Curva ROC.

AUC Value = 0,5397

A curva ROC (Receiver Operating Characteristic) é um gráfico que ilustra o desempenho do modelo de classificação em diferentes thresholds. A curva representa

a sensibilidade (taxa de verdadeiros positivos) contra 1 - especificidade (taxa de falsos positivos.

A linha diagonal cinza representa o desempenho de um modelo que faz previsões aleatórias. Uma curva ROC que se aproxima dessa linha indica que o modelo não é muito melhor que uma previsão aleatória.

A AUC (Área Sob a Curva) mede a capacidade do modelo de separar as duas classes (neste caso, churn e não churn). A AUC varia entre 0 e 1:

• AUC = 1: O modelo é perfeito na distinção entre as classes.

• AUC = 0.5: O modelo não tem poder discriminativo e está basicamente a adivinhar aleatoriamente (equivalente a uma linha diagonal).

• AUC < 0.5: O modelo está a prever pior que aleatório (inverte as classes).

Neste caso a AUC = 0,5397, o que indica que o modelo tem um desempenho muito baixo para prever os camiões que se atrasam e que não se atrasam. Basicamente, está marginalmente melhor que uma classificação aleatória.

Este modelo não vai ser bom para a empresa, provavelmente as variáveis que utilizamos não foram as melhores.

3. Clustering

3.1 Distribuição de Registos por Mês.

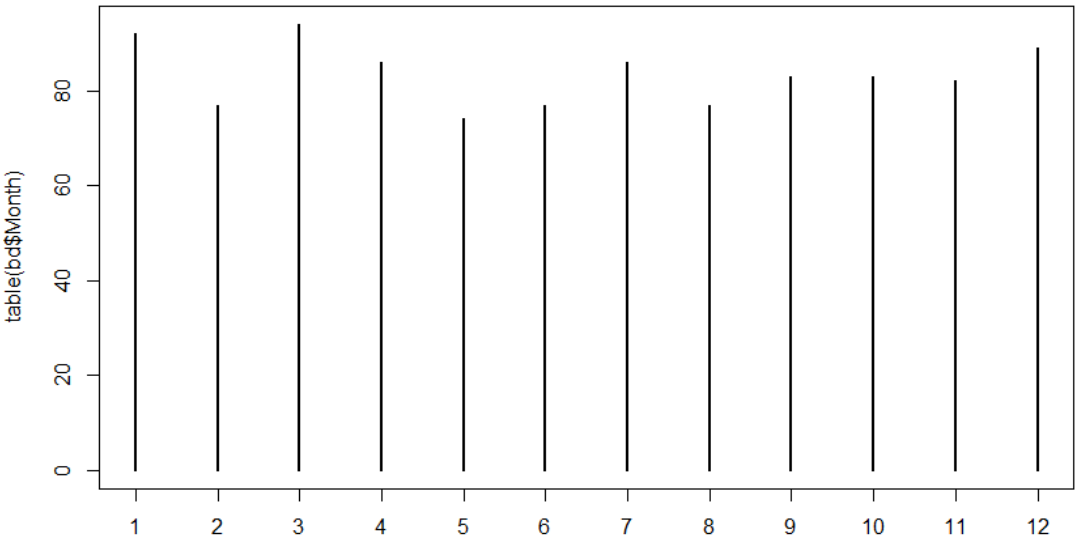

Figura 8: Distribuição de Registos por Mês.

A Figura 8 ilustra a distribuição dos registos ao longo dos 12 meses de 2024. Embora o número de observações seja relativamente equilibrado entre meses, existem ligeiras variações: por exemplo, março apresenta 94 registos (pico anual), enquanto maio apresenta 74 registos (valor mínimo). Julho e dezembro, que serão analisados em maior detalhe, têm 86 e 89 registos, respetivamente, o que garante uma amostra suficientemente representativa em ambos os casos.

Esta distribuição relativamente homogénea ao longo do ano indica que as operações logísticas decorrem de forma contínua, sem períodos de paragem prolongados. A escolha de julho (verão) e dezembro (inverno) para análise de clusters permite captar potenciais efeitos sazonais relacionados com clima, campanhas comerciais e alterações na procura.

3.2 Preparação dos Dados

Na fase de preparação, realizamos algumas transformações no script R, sendo eles :

• Criação da variável tipo_trafico a partir de Traffic_Status, codificando as categorias Clear, Heavy e Detour em valores numéricos 1, 2 e 3.

• Criação da variável Month a partir da variável Timestamp.
• Seleção de grupos mais pequenos de dados para os meses de dezembro (Month = 12) e julho (Month = 7), permitindo uma análise comparativa entre períodos sazonais.
• Padronização (standardização) das variáveis numéricas selecionadas para a análise de clusters, fazendo com que todas tenham o mesmo peso para o cálculo das distâncias.

3.3 Escolha de Clusters dezembro

```
Cluster sizes:
 2 3 4 5

Validation Measures:
                                  2       3       4       5

hierarchical Connectivity  16.8460 31.5111 33.8778 56.8587
             Dunn           0.3099  0.3264  0.3264  0.3431
             Silhouette     0.0988  0.0871  0.0657  0.0903
kmeans       Connectivity  58.8536 54.1837 68.7929 77.4254
             Dunn           0.2780  0.3042  0.2569  0.2710
             Silhouette     0.0944  0.1147  0.1199  0.1175

Optimal Scores:

             Score   Method       Clusters
Connectivity 16.8460 hierarchical 2
Dunn          0.3431 hierarchical 5
Silhouette    0.1199 kmeans       4
```

Figura 9: Escolha de Clusters dezembro.

O número de clusters para o mês de dezembro foi escolhido através de critérios internos e também através da visualização do dendrograma. Testamos também soluções entre dois e cinco clusters, para o método hierárquico e também para o k-means.

Os resultados dos critérios internos (nomeadamente medidas de compactação e separação entre grupos) mostram que os resultados não são positivos para mais de dois clusters. Pelo contrário, a solução com dois clusters mostra-se mais estável e mais fácil para interpretar do ponto de vista operacional.

Adicionalmente, a observação do dendrograma de dezembro (Figura 9) mostra um aumento acentuado da distância de ligação no momento em que se passa de dois para

um único cluster. Este salto indica que a fusão final ocorre entre grupos relativamente heterogéneos, o que reforça a adequação da escolha de dois clusters.

Assim, considerando simultaneamente os critérios quantitativos de validação e a interpretação visual do dendrograma, optou-se por uma solução com **dois clusters** para o mês de dezembro, permitindo uma segmentação clara e consistente das observações.

3.3.1 Modelação (Análise de Clusters dezembro)

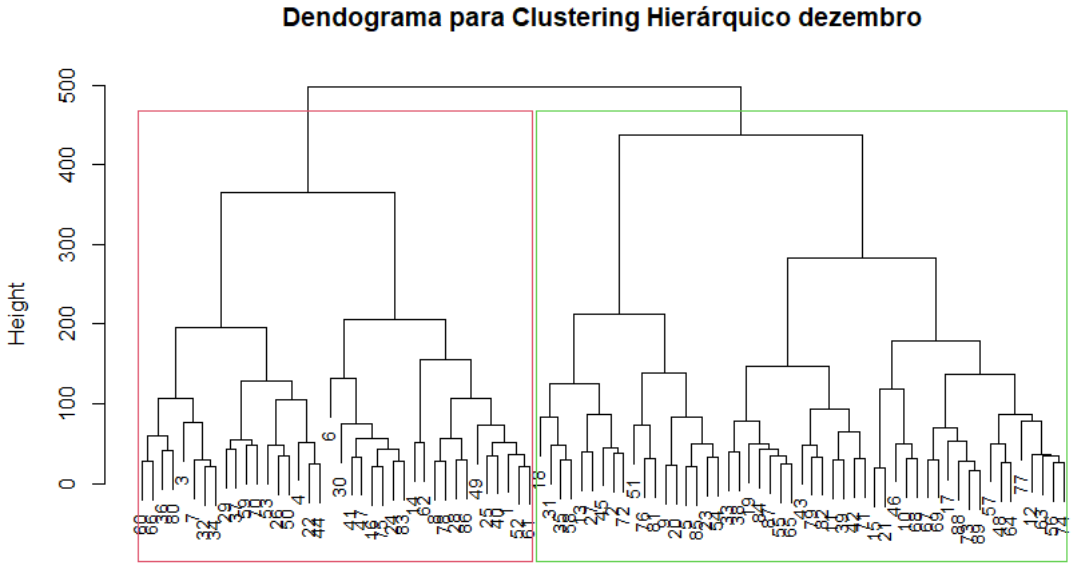

Figura 10: Dendrograma dos Clusters dezembro.

A técnica de modelação escolhida foi a análise de clusters hierárquica. Para cada um dos meses em estudo (julho e dezembro) foi construída uma matriz de distâncias euclidianas com base nas variáveis: Inventory_Level, Temperature, Humidity, tipo_trafico, Waiting_Time, User_Transaction_Amount, User_Purchase_Frequency, Asset_Utilization e Logistics_Delay. Em seguida, aplicou-se o método hierárquico aglomerativo com ligação completa (complete linkage), que tende a formar clusters compactos e bem separados.

Para o mês de dezembro, o dendrograma apresentado na Figura 3 mostra como as observações vão sendo agregadas em função da distância euclidiana entre elas. O eixo

horizontal representa as observações e o eixo vertical a distância de ligação. Observa-se um salto mais pronunciado na altura em que se passa de dois para um cluster, o que, em conjunto com os critérios internos calculados no R (via clValid), justifica a escolha de uma solução com dois clusters para este mês.

A solução final para dezembro agrupa as 89 observações em dois clusters: o Cluster 1 com 38 registos (42.7%) e o Cluster 2 com 51 registos (57.3%).

```
> table(bd_dez$Logistics_Delay, bd_dez$cluster_hc_dez) #sumario variaveis categoricas

     1  2
  0 19 22
  1 19 29
> table(bd_dez$Traffic_Status, bd_dez$cluster_hc_dez)

          1  2
  Clear  11 14
  Detour 16 28
  Heavy  11  9
> table(bd_dez$tipo_trafico, bd_dez$cluster_hc_dez)

     1  2
  1 11 14
  2 11  9
  3 16 28
```

Figura 11: Análise das variáveis categóricas por Cluster dezembro.

Através da análise da varável "Logistics_Delay", conseguimos perceber que que tanto no Cluster 1 como no Cluster 2, a taxa de atraso é bastante elevada. No caso do Cluster 1 temos 19 atrasos e 19 não atrasos, já no Cluster 2 chegamos mesmo a ter mais atrasos que não atrasos, 29 para 22. Assim, percebemos que nenhum dos clusters está livre de atrasos, mas o Cluster 2 apresenta uma predominância maior nos atrasos, cerca de 56,9%.

Pela variável "Traffic_Status" observamos que o estado Detour é o mais frequente em ambos os Clusters, porém no Cluster 2 apresenta um peso relativo e absoluto maior. No Cluster 1, cerca de 42,1% das entregas foram classificadas com Detour, no Cluster 2 já sobe para 54,9%. O estado Heavy também apresenta diferenças no peso relativo sobre os Clusters, no 1 representa 28,9% dos casos, já no 2 representa 17,6%. Por fim, o estado Clear apresenta um peso relativamente semelhante, sendo 29% para o Cluster 1 e 27% para o 2.

Pela varável, "tipo_trafico", criada anteriormente a partir de "Traffic_Status" (1=Clear, 2=Heavy, 3=Detour), vemos que realmente, não foram alterados dados da

variável, apenas foi passada de categórica para numérica, podendo assim ser manipulada pelo R.

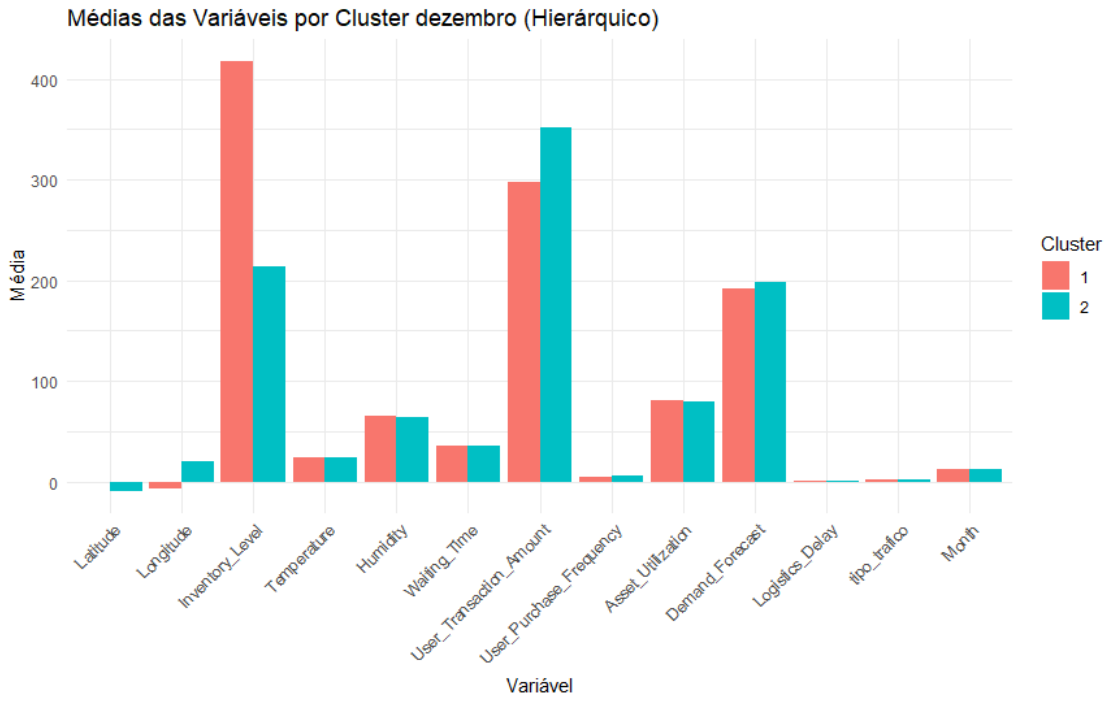

Figura 12: Médias das Variáveis por Cluster dezembro.

A Figura 12 apresenta as médias das principais variáveis por cluster para dezembro. Se compararmos os valores, verificamos que o Cluster 1 apresenta, em média, um nível inventário mais elevado (Inventory_Level ≈ 418.0) e uma utilização de ativos também superior (Asset_Utilization ≈ 81.3%). O Cluster 2, apresenta um nível de inventário médio mais baixo (≈ 214.1) mas um valor médio de User_Transaction_Amount mais elevado (≈ 352.3).

Os tempos médios de espera (Waiting_Time) são praticamente iguas entre os dois clusters (cerca de 35.5–35.8 minutos). Estes dados dizem-nos que as encomendas mais caras tendem a sofrer mais atrasos do que as outras.

As Figuras 13 a 16 mostram a distribuição de variáveis-chave por cluster no mês de dezembro. Cada boxplot permite observar as diferenças de média, assim como a dispersão de dados e ainda a existência ou não de valores extremos.

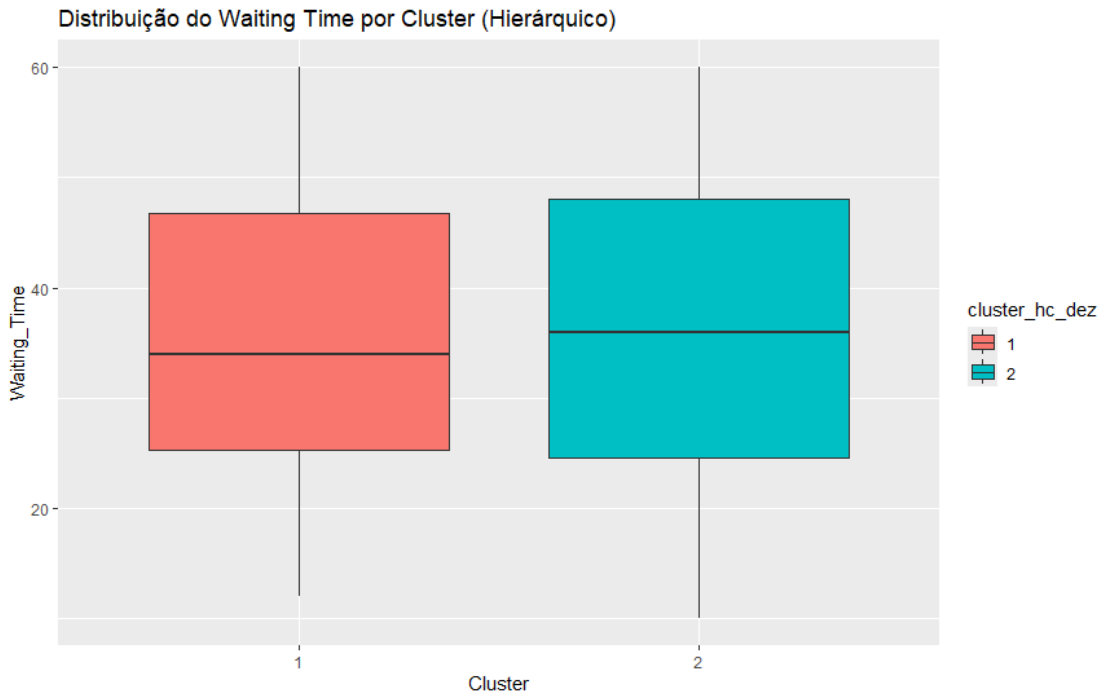

Figura 13: Distribuição de Waiting_Time por Cluster dezembro.

A Figura 13, mostra a distribuição do Waiting_Time por cluster, verifica-se que as medianas são muito similares e muito próximas, reforçando a ideia de que o tempo de espera médio é semelhante entre os grupos. Ainda assim, as diferenças na amplitude interquartil mostram que um dos grupos cumpre os prazos de forma regular enquanto o outro apresenta oscilações nos tempos de espera.

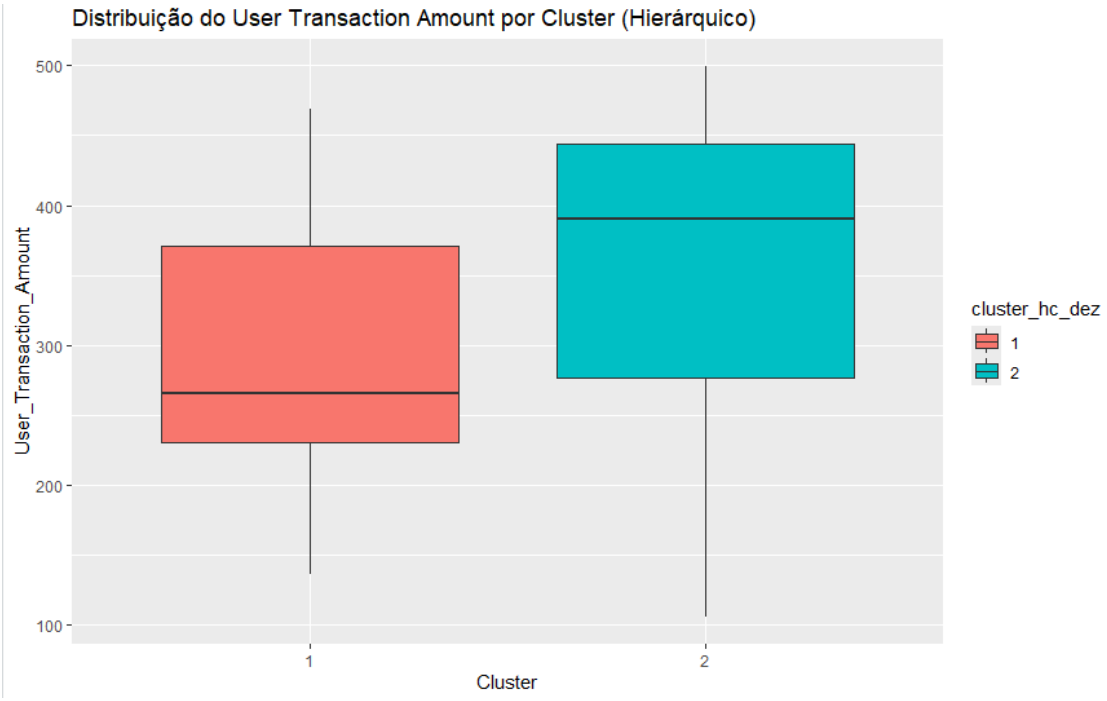

Figura 14: Distribuição de User_Transaction_Amount por Cluster dezembro.

A Figura 14 apresenta a distribuição de "User_Transaction_Amount" por cluster. É possível observar que o Cluster 2 apresenta claramente valores medianos maiores, mas curiosamente abrange também valores mais baixos, ou seja, apesar de trabalhar maioritariamente com operações de maior valor, é também ele que trabalha com as de menor valor.

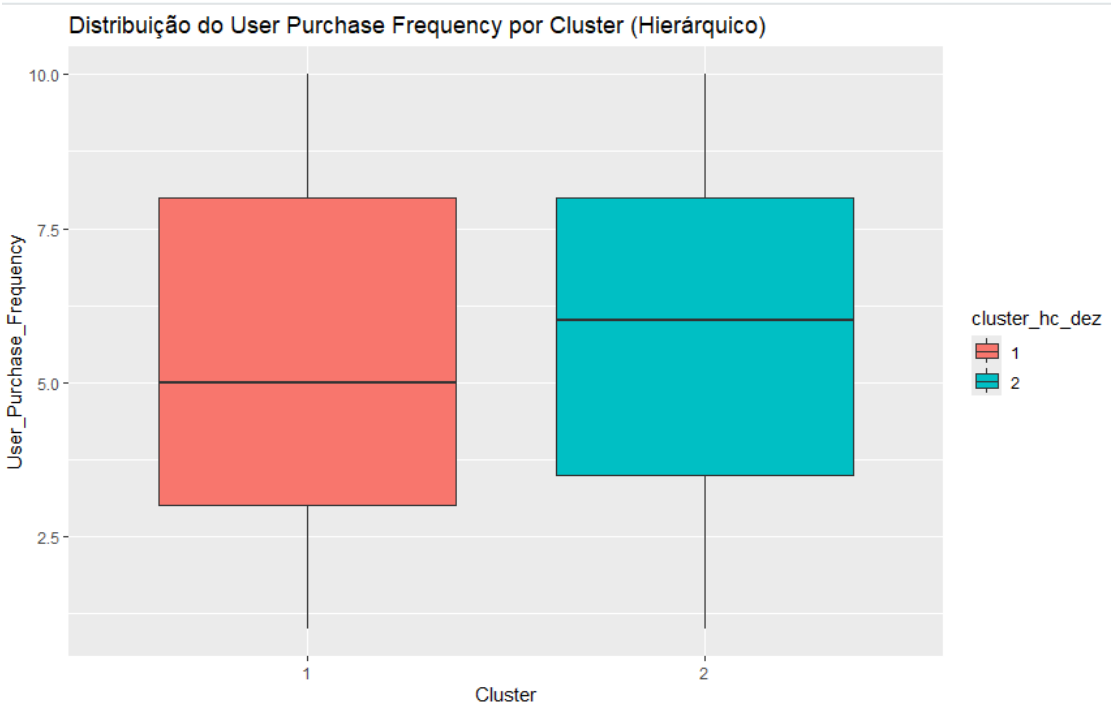

Figura 15: Distribuição de User_Purchase_Frequency por Cluster dezembro.

Na Figura 15, relativa à variável User_Purchase_Frequency, embora haja diferenças muito pouco significativas, a análise de frequência compra permite-nos identificar os clientes leais. Consequentemente, estes clientes vão ter uma expectativa de qualidade de serviço maior. Caso aconteçam falhas nas entregas, o cliente pode quebrar a sua confiança.

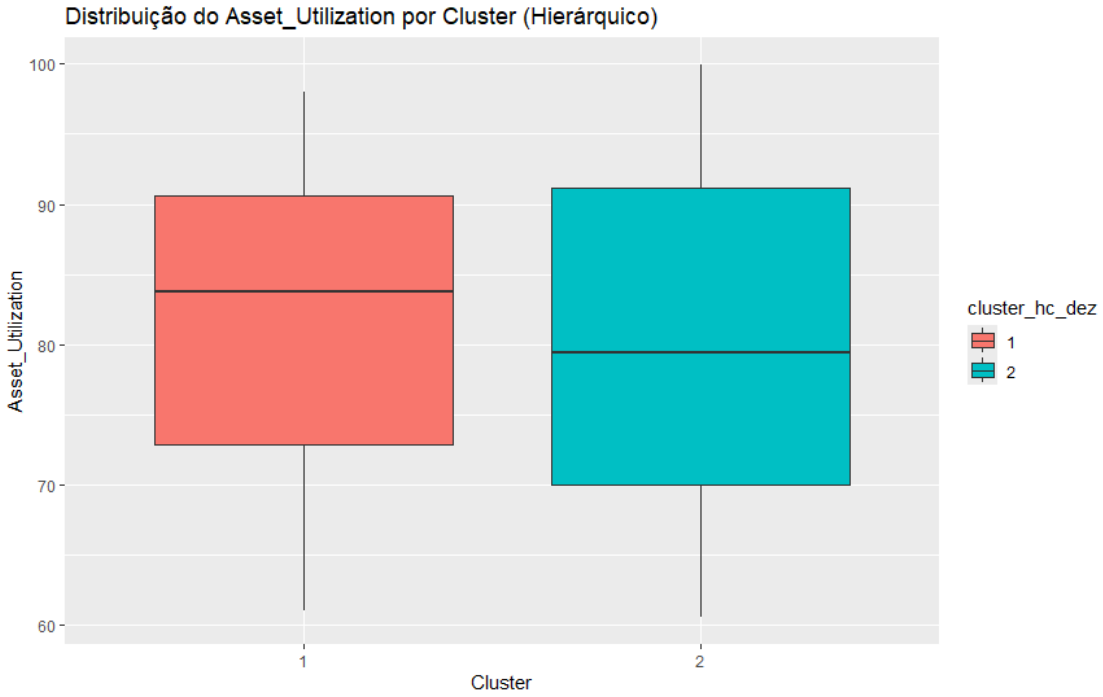

Figura 16: Distribuição de Asset_Utilization por Cluster dezembro.

Para terminar, a Figura 16 demonstra a distribuição de Asset_Utilization por cluster. As diferenças demonstradas nas medianas podem indicar que um dos clusters funciona mais próximo da saturação de capacidade, o que aumenta a probabilidade de atrasos devido a uma menor margem de manobra para lidar com imprevistos (falhas mecânicas, desvios de rota, alterações súbitas na procura, etc.).

3.4. Escolha de clusters julho

```
Cluster sizes:
 2 3 4 5

Validation Measures:
                                 2       3       4       5

hierarchical Connectivity   5.9758 14.3349 28.4663 47.2448
             Dunn           0.3602  0.3372  0.3386  0.3070
             Silhouette     0.1090  0.0448  0.0280  0.0460
kmeans       Connectivity  49.1163 64.3532 77.3730 86.1032
             Dunn           0.2515  0.2350  0.2823  0.3180
             Silhouette     0.1113  0.1253  0.1126  0.1284

Optimal Scores:

             Score  Method       Clusters
Connectivity 5.9758 hierarchical 2
Dunn         0.3602 hierarchical 2
Silhouette   0.1284 kmeans       5
```

Figura 17: Escolha de Clusters julho.

Para decidirmos o número de clusters para julho usamos os mesmos métodos que já tínhamos usado para dezembro, para que a coerência fosse garantida na comparação de períodos sazonais. Avaliamos novamente entre dois e cinco clusters.

Tal como em dezembro, concluímos que dividir os dados em mais de dois grupos não facilita a análise e só traz complexidade. A segmentação em dois clusters é o melhor neste caso.

Através da visualização do dendrograma conseguimos ver uma divisão entre os dois grupos. O gráfico mostra que os dois grupos têm realidades completamente opostas.

A resolução para julho igual à de dezembro é perfeita pois garante uma base sólida e direta entre os dois meses. Com esta coerência podemos observar que o comportamento dos clientes evoluiu de um mês para outro.

3.4.1 Modelação (Análise de Clusters julho)

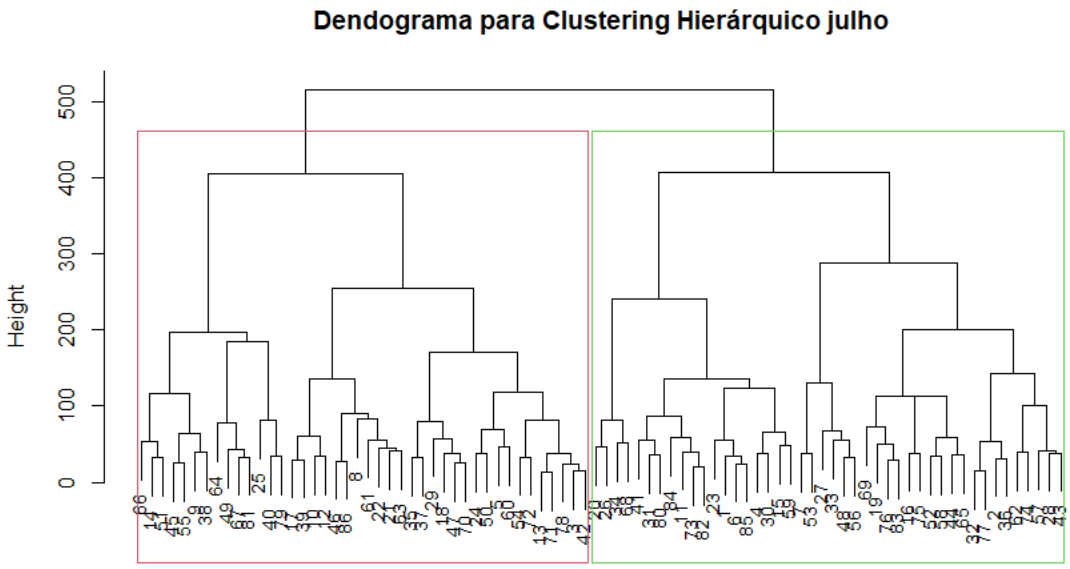

Figura 18: Dendrograma dos Clusters julho.

Para o mês de julho, o dendrograma da Figura 18 mostra um padrão semelhante, com duas grandes massas de observações que se separam a uma distância relativamente maior, o que suporta novamente a escolha de dois clusters. Neste caso, os 86 registos de julho são repartidos entre o Cluster 1 (42 observações, 48.8%) e o Cluster 2 (44 observações, 51.2%).

```
> table(bd_jul$Logistics_Delay, bd_jul$cluster_hc_jul) #sumario variaveis categoricas

    1  2
0 14 18
1 30 24
> table(bd_jul$Traffic_Status, bd_jul$cluster_hc_jul)

         1  2
Clear   12 17
Detour 18 14
Heavy   14 11
> table(bd_jul$tipo_trafico, bd_jul$cluster_hc_jul)

    1  2
1 12 17
2 14 11
3 18 14
```

Figura 19: Análise das variáveis categóricas por Cluster julho.

Através da análise da varável "Logistics_Delay", conseguimos perceber que que tanto no Cluster 1 como no Cluster 2, a taxa de atraso é bastante elevada. No caso do Cluster 1 temos 30 atrasos e 14 não atrasos, já no Cluster 2 temos 24 atrasos para 18 não atrasos. Assim, percebemos que nenhum dos clusters está livre de atrasos, mas o Cluster 1 apresenta uma predominância maior nos atrasos, cerca de 68,2%. À semelhança de Dezembro, o Cluster que apresenta maior valor de utilização e maior valor de ativo, é também aquele que apresenta uma maior taxa de atraso, mas em julho, a diferença é ainda mais pronunciada.

Relativamente à variável Traffic_Status, em julho o Cluster 2 apresenta uma maior representatividade o estado Clear (cerca de 40,5% das observações), seguida de Detour (33,3%) e Heavy (26,2%). Já no Cluster 1, observa-se um comportamento distinto: as situações Detour e Heavy ganham maior peso relativo (cerca de 40,9% e 31,8%, respetivamente), enquanto os casos Clear descem para cerca de 27,3%.

Pela varável, "tipo_trafico", criada anteriormente a partir de "Traffic_Status" (1=Clear, 2=Heavy, 3=Detour), vemos que realmente, não foram alterados dados da variável, apenas foi passada de categórica para numérica, podendo assim ser manipulada pelo R.

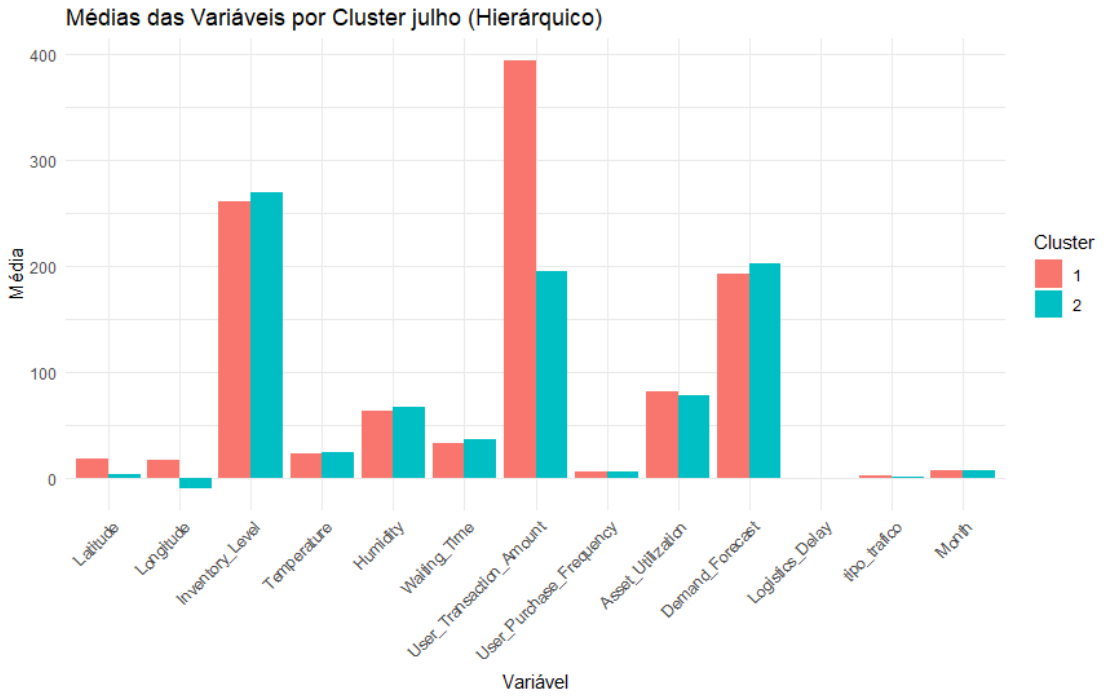

Figura 20: Médias das Variáveis por Cluster julho.

A Figura 20 apresenta as médias de algumas das variáveis, relativo ao Cluster de julho. Vamos então analisar algumas das variáveis, "Waiting_Time" podemos ver que o Cluster 2 apresenta um valor médio ligeiramente superior, 37,1 minutos. A frequência média de compra também é ligeiramente superior no Cluster 2, com 6,4 de média. Já o Cluster 1, apresenta valores superiores por exemplo nas variáveis "User_Transaction_Amount" com 384,6 de média e também na variável "Asset_Utilization" com uma média de 81,9%, sugerindo um perfil de utilização maior.

Nas Figuras 21 a 24, vão ser analisadas estas variáveis com mais detalha com auxílio de Box Plots, permitindo comparar a forma das distribuições e identificar eventuais outliers.

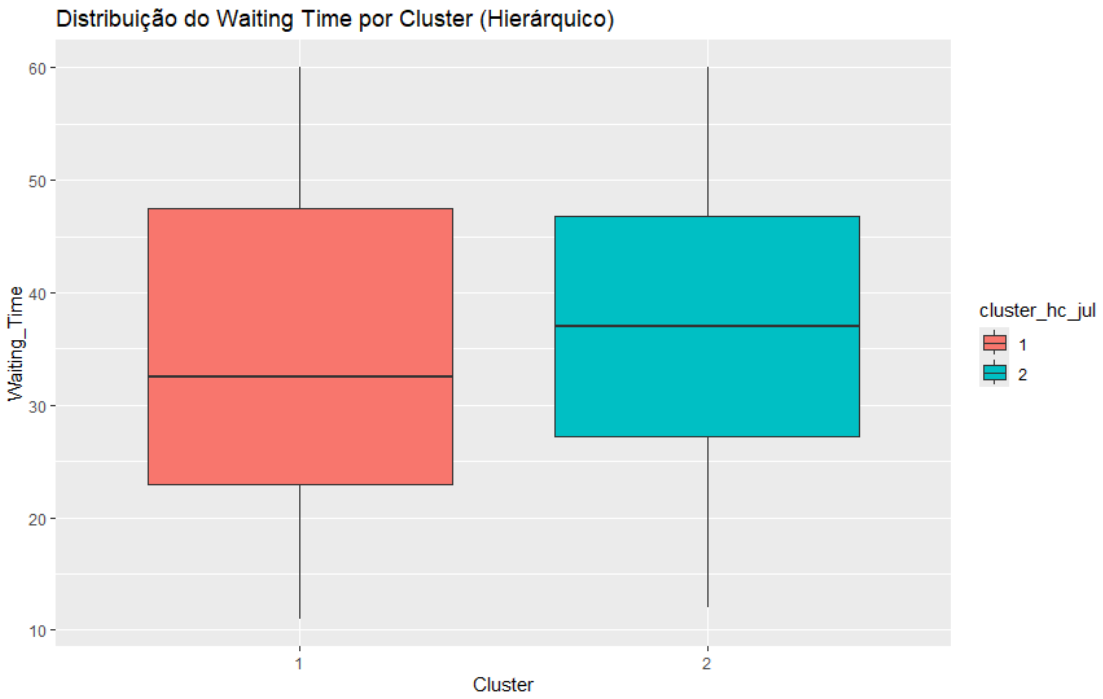

Figura 21: Distribuição de Waiting_Time por Cluster julho.

Na Figura 21, referente ao Waiting_Time, observamos se os tempos de espera no Cluster 2 apresentam diferenças do Cluster 1. O Cluster 1 apresenta, tempos de espera mais baixos derivados da sua mediana inferior, o que sugere atendimentos mais rápidos e eficientes. No entanto, observa-se alguma inconstância nos tempos de espera, o que indica que nem todos os casos são homogéneos dentro do cluster. Já o Cluster 2 apresenta tempos de espera mais elevados, pois, a sua mediana é superior, e tem uma distribuição mais deslocada para valores altos. Este cluster também apresenta uma dispersão considerável, o que evidencia maior diversidade e maior probabilidade de situações com tempos de espera prolongados. Resumindo, o Cluster 2 está ligado a situações de maior demora no atendimento, enquanto o Cluster 1 corresponde a um padrão de espera mais reduzido.

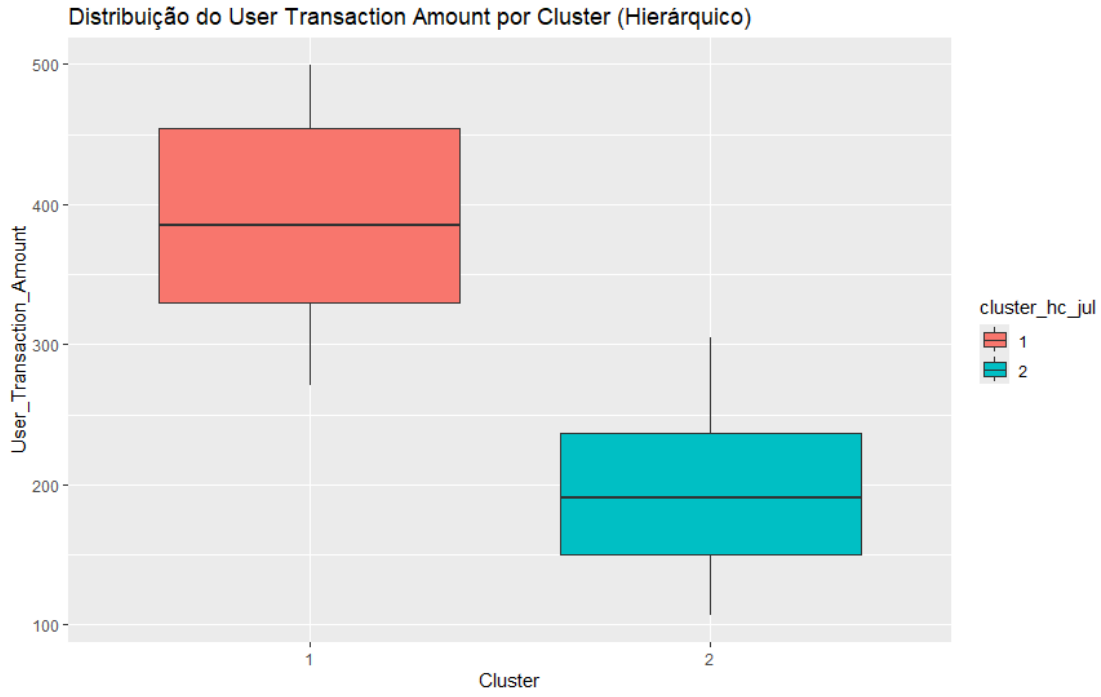

Figura 22: Distribuição de User_Transaction_Amount por Cluster julho.

A Figura 22 demonstra novamente que o Cluster 1 concentra os envios de maior valor (User_Transaction_Amount), este tem uma mediana mais elevada e uma cauda superior mais longa. Este tipo de informação é relevante, pois, do ponto de vista de negócio, sugere que atrasos em envios de maior valor tendem a concentrar-se num grupo específico, o que justificaria a implementação de certas ações com o objetivo de minimizar o atraso. (por exemplo, rotas alternativas ou prioridade na atribuição de recursos).

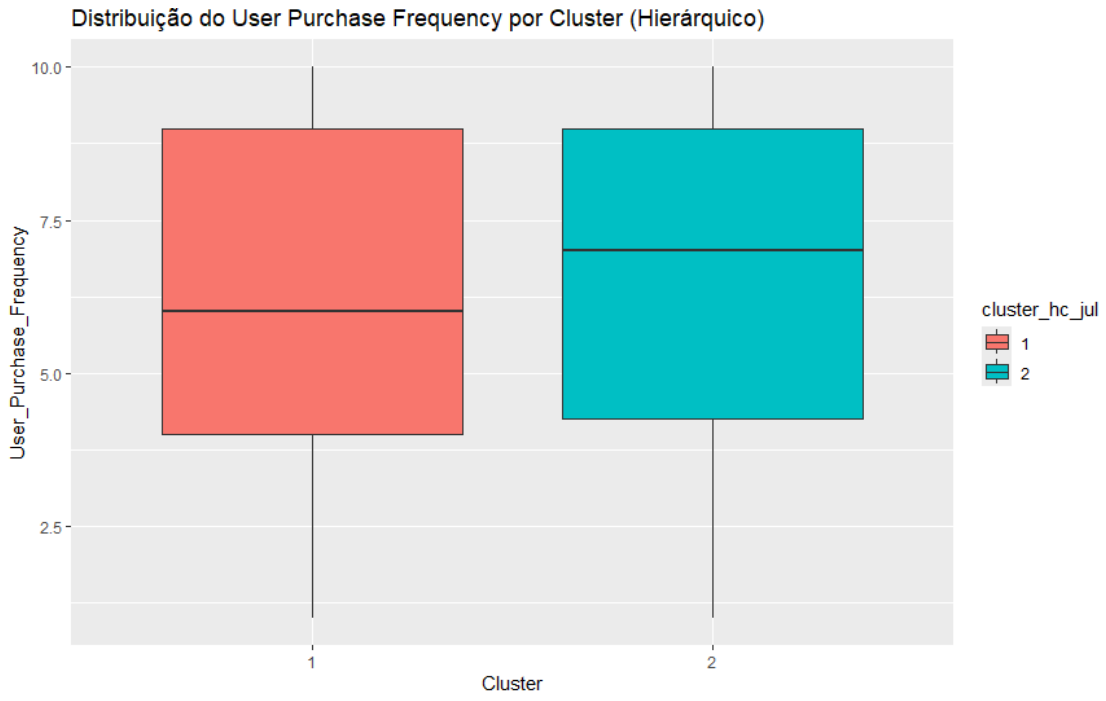

Figura 23: Distribuição de User_Purchase_Frequency por Cluster julho.

A Figura 23 permite nos avaliar se os utilizadores com maior frequência de compra se concentram num dos dois clusters. Caso se verifique que clientes mais frequentes estão predominantemente num cluster com mais atrasos, isso pode sinalizar um risco de insatisfação e perda de fidelização dos clientes, o que deve ser considerado em futuras recomendações operacionais.

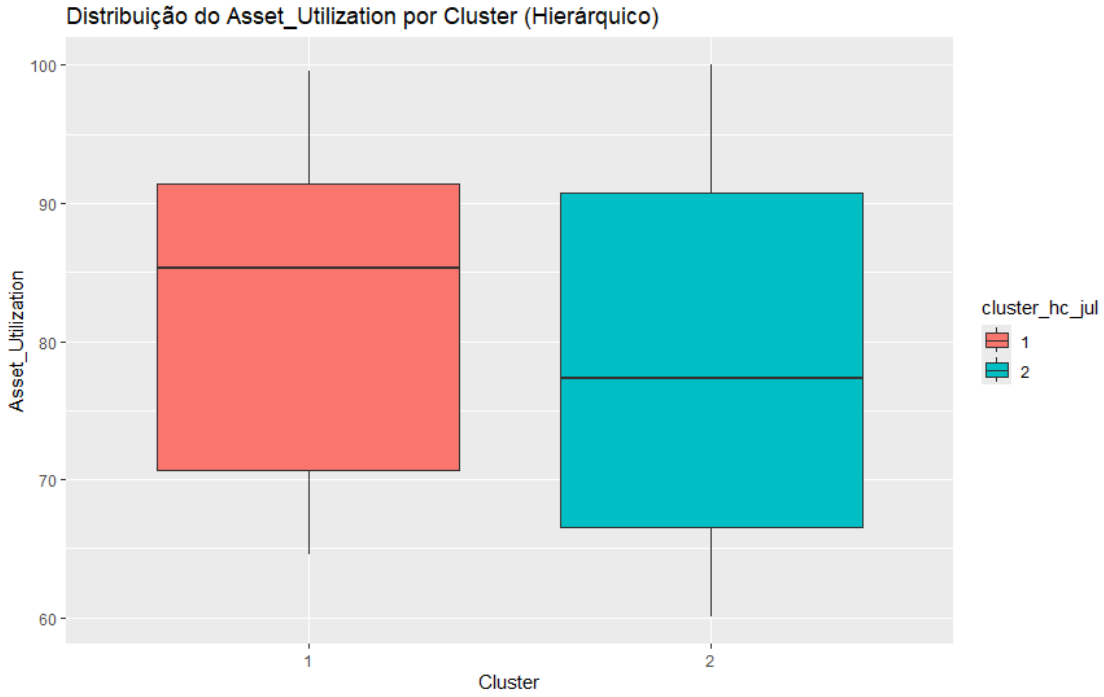

Figura 24: Distribuição de Asset_Utilization por Cluster julho.

Para terminar, a Figura 24 demonstra a distribuição de Asset_Utilization por cluster. As diferenças demonstradas nas medianas podem indicar que um dos clusters funciona mais próximo da saturação de capacidade, o que aumenta a probabilidade de atrasos devido a uma menor margem de manobra para lidar com imprevistos (falhas mecânicas, desvios de rota, alterações súbitas na procura, etc.).

4. Avaliação

Na avaliação os nossos principais objetivos são analisar a qualidade e utilidade dos resultados obtidos, tanto no modelo de regressão logística como na análise de clusters, verificando se estes respondem aos objetivos definidos.

Quanto ao modelo de regressão logística, os resultados das análises mostram limitações claras. Apesar de apresentar uma especificidade muito elevada (≈98%), indicando uma boa capacidade para identificar entregas com atraso, o modelo apresenta uma sensibilidade muito baixa (≈3,6%), o que significa que o modelo não é eficaz na identificação correta das entregas sem atraso. A acurácia balanceada próxima de (≈57,1%) muito próximo de 50% que corresponde ao valor de um modelo aleatório e o valor da AUC de (≈0,5397) que confirma que o desempenho do modelo é apenas marginalmente melhor do que uma classificação aleatória. Avaliando os valores Pseudo R-quadrados entendemos que estes valores são muito baixos o que reforça que as variáveis consideradas explicam apenas uma pequena parte da variabilidade dos atrasos.

No entanto, a análise de clusters revelou-se mais eficaz e com melhores valores. A escolha de dois clusters para julho e dezembro foi sustentada por código sustentado por dendrogramas, permitindo identificar grupos com perfis distintos. Em ambos os meses, observou-se de forma constante um cluster associado a maior valor financeiro das transações, maior utilização de ativos e maior proporção de atrasos, reforçando a coerência e utilidade dos resultados obtidos.

5. Resultados

Os resultados obtidos e analisados ao longo deste trabalho permitem avaliar o desempenho da operação logística e identificar padrões associados aos atrasos nas entregas.

A análise descritiva mostrou que a base de dados é bastante equilibrada e adequada para uma análise, não demonstrando enviesamentos. Verificou-se também que 56,6% das entregas foram feitas com atraso.

Sobre o modelo de regressão logística, demonstrou-se desempenho geral fraco. Apesar de conseguir identificar bastante bem entregas com atraso, o modelo apresentou uma capacidade discriminativa muito baixa, evidenciado por baixos valores dos Pseudo R-quadrados e um AUC próximo de 0,5. Demonstrando-se um modelo quase aleatório.

Já na análise de clusters hierárquicos revelou-se mais consistente. Foram identificados dois clusters em dois períodos diferentes, julho e dezembro. Em ambos os meses, um dos clusters apresenta uma maior utilização do ativo, maior valor das transações e maior taxa de atraso.

A comparação sazonal demonstrou que, apesar das estruturas dos clusters serem semelhantes, em julho os atrasos estão mais relacionados com as condições do trânsito e aos tempos de espera, já em dezembro aparentam estar mais relacionados com o aumento da intensidade operacional.

6. Conclusão

Este trabalho permitiu estudar a eficiência operacional no setor logístico através da base de dados *Smart Logistics.* Com a fase de limpeza de dados verificámos que não existiam dados duplicados nem dados omissos o que permitiu a passagem direta para a fase de modelação.

No entanto, com a aplicação da regressão logística percebemos que o modelo não possui robustez suficiente para prever corretamente os atrasos. Com uma métrica de desempenho muito próxima do valor aleatório (50%), concluímos que o modelo não é capaz distinguir padrões claros.

Em contraste com a análise da regressão logística, a análise de clusters hierárquicos revelou uma segmentação robusta e confiável. Através da descomposição dos dados sazonais, percebemos que quanto mais cara é a mercadoria, quanto mais ocupados estão os recursos maior é o risco de atraso das entregas. Percebemos também que as razões de atraso dos camiões em julho não são as mesmas de dezembro, concluindo assim que a eficiência da empresa não é igual durante todo o ano.

Do ponto de vista prático, concluímos que o gestor deve dar total prioridade às encomendas mais cara visto que são as que se atrasam mais. Isto pode significar a utilização dos melhores camiões, dos melhores motoristas e etc...

A empresa deve também implementar margens de segurança maiores nos meses mais críticos garantindo assim o cumprimento de todas as entregas.

Ainda assim a empresa deve fazer uma revisão das suas políticas de capacidade e de planeamento. O modelo diz que no mês de julho os ativos estão no seu limite, logo deve haver uma maior disponibilidade tanto de recursos como de ativos.

Para um modelo melhor, recomendamos adicionar algumas variáveis explicativas, como por exemplo fatores externos (meteorologia, acidentes rodoviários, características dos caminhos das rotas) e a aplicação de modelos mais avançados de Machine Learning, como árvores de decisão, que poderão analisar relações não lineares mais detalhadas e complexas. Sugerimos também a o estudo de mais períodos temporais o que poderá reforçar ainda mais o apoio à tomada de decisão no contexto logístico.

7. Apêndice

Link Google Colab do Dashboard:

https://colab.research.google.com/drive/1BLPPEkGph4ksPExVO-0R6khweZbGl1q9?usp=sharing