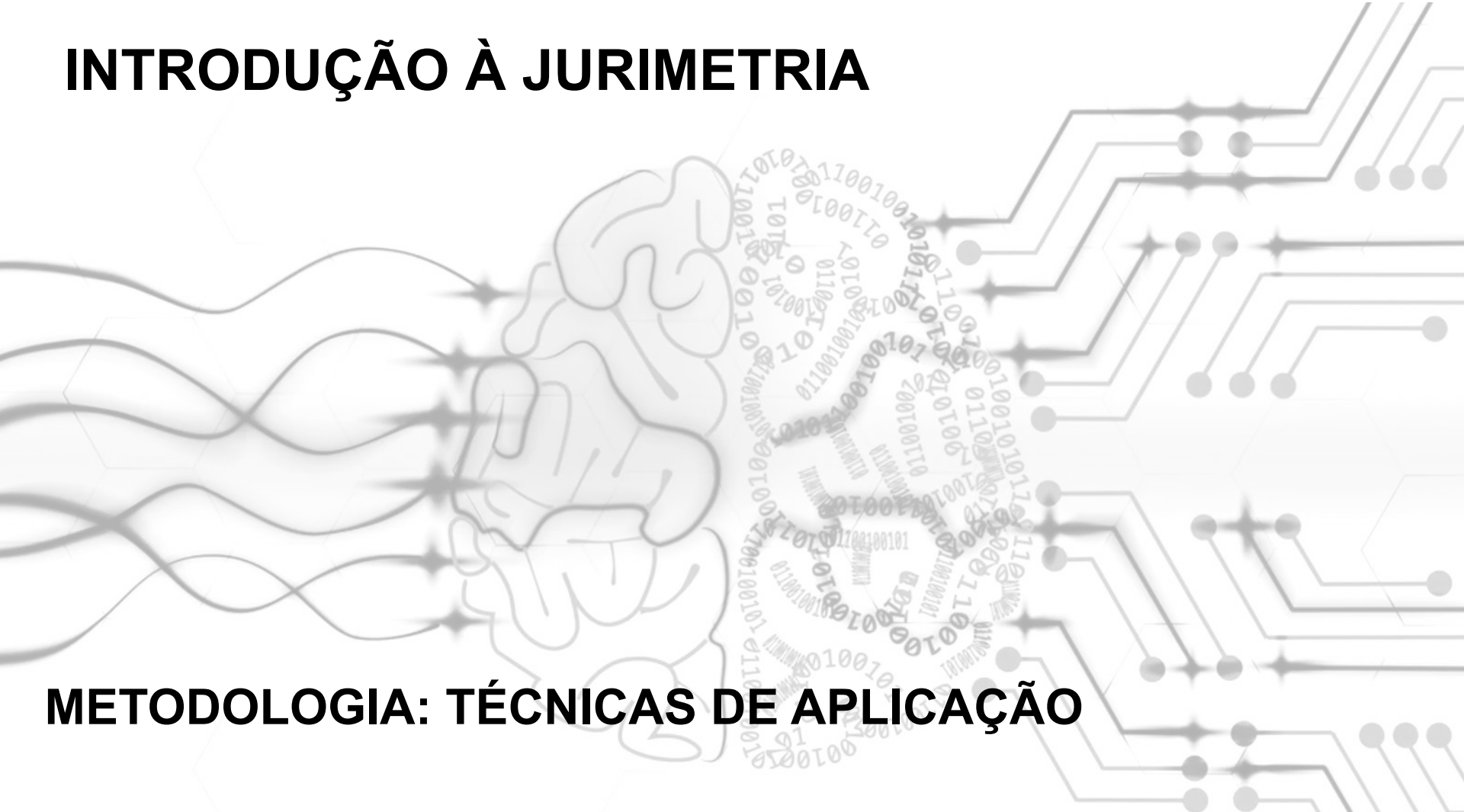
INTRODUÇÃO À JURIMETRIA



METODOLOGIA: TÉCNICAS DE APLICAÇÃO

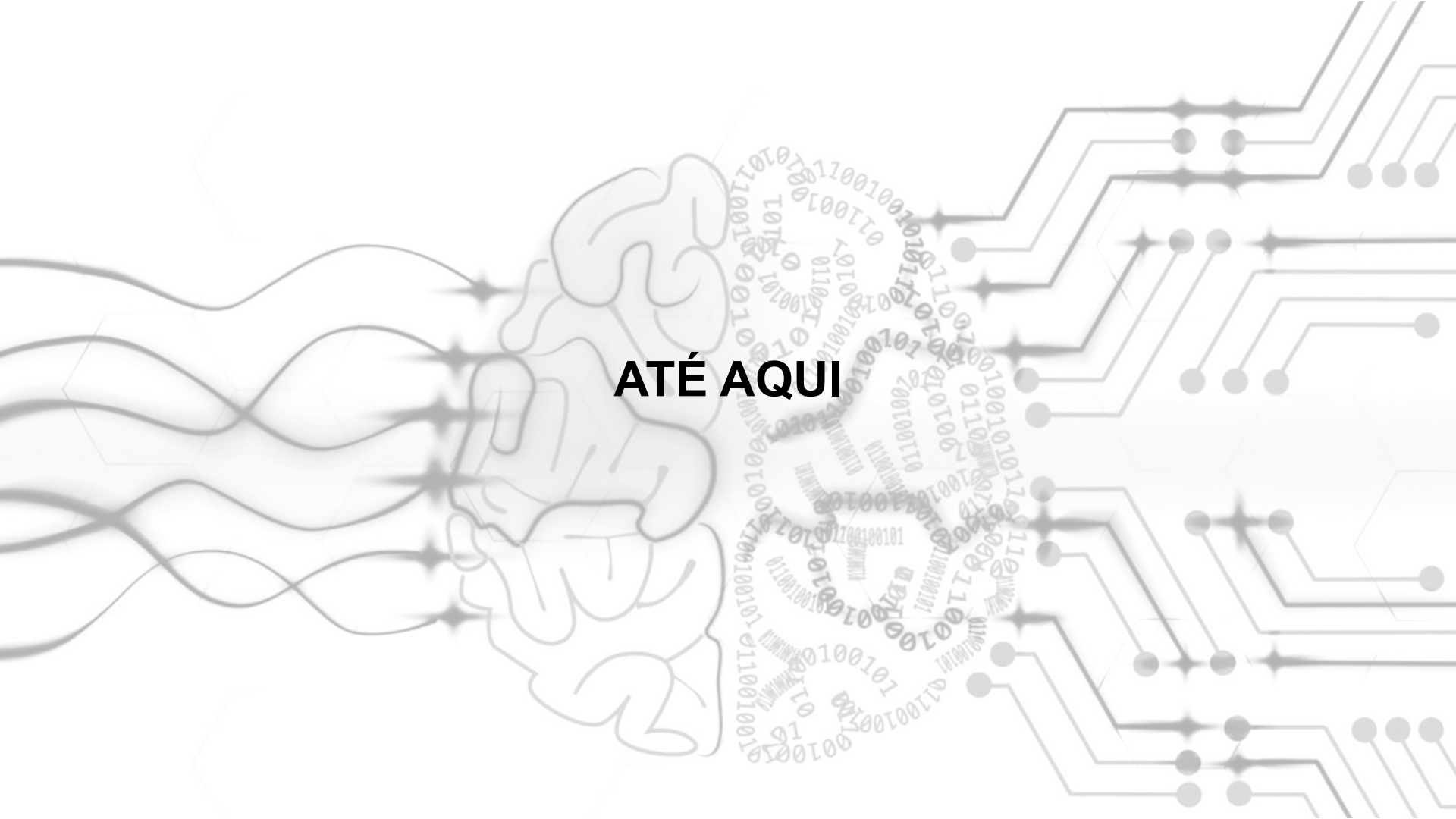
ASSOCIAÇÃO BRASILEIRA DE JURIMETRIA:

https://livro.abj.org.br

https://abj.org.br/cases/

AULA DE HOJE: **GITHUB**

https://github.com/BrunoDaleffi/curso_espge_ago_23/tree/master/aulas


ATÉ AQUI

Jurimetria

Definição: Aplicação de métodos quantitativos, principalmente estatística, ao direito.

A JURIMETRIA É UMA ÁREA DO CONHECIMENTO!

Objetivos: A jurimetria visa melhorar a precisão, eficácia e eficiência na tomada de decisões jurídicas e na formulação de políticas legais por meio da análise quantitativa e da interpretação de dados jurídicos.

A importância da jurimetria

1. A Jurimetria fornece uma abordagem quantitativa e analítica para a compreensão do direito e do sistema jurídico.
    - Análises Descritivas
    - Análise de perfil de julgamento
    - Priorização de casos
    - Redução do tempo de tramitação
    - Identificação de padrões
    - Mapeamento de jurisprudência
    - Entre outros

A importância da jurimetria

2. A Jurimetria tem um papel significativo na formulação e aplicação de políticas públicas.
   - Análise de Impacto Regulatório (Lei 13.974/19 Capítulo IV)
     - Que objetiva avaliar as consequências econômicas, sociais e ambientais de propostas regulatórias para orientar decisões do regulador
   - Avaliação de Resultado Regulatório (Decreto nº 10.411/20)
     - Mensura o desempenho de regulamentações implementadas, verifica seus objetivos alcançados e identifica áreas de melhoria.

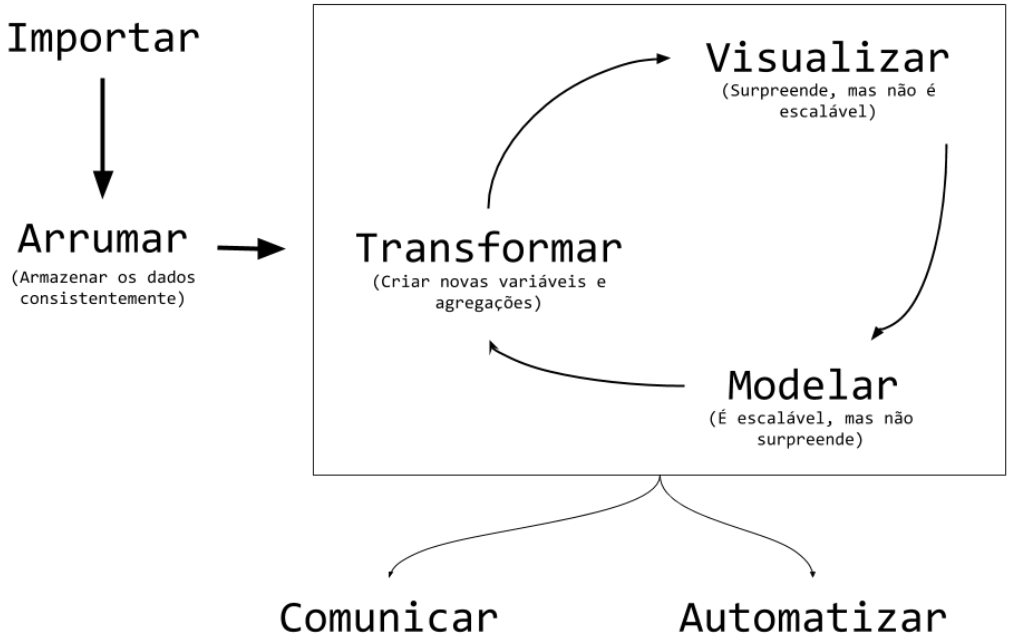
Como a Estatística ajuda o Direito?

- Através de análises descritivas
- Através de medidas-resumo
- Através das análises de impactos que ajudam na reformulação de leis
- Através das análises internas que embasam decisões estratégicas, como
  - priorização de acordos
  - provisionamento
  - mudança de políticas internas etc

Como o Direito ajuda a Estatística?

- Tomada constante de decisões metodológicas
- Tradução Juridiquês -> Estatistiquês, ou operacionalização de conceitos.
- Quando esses dados se tratam de dados jurídicos, essas decisões precisam ser embasadas pelo direito.
- Assim, o Direito ajuda a Estatística, principalmente, por meio de conhecimentos de direito processual

O ciclo da ciência de dados



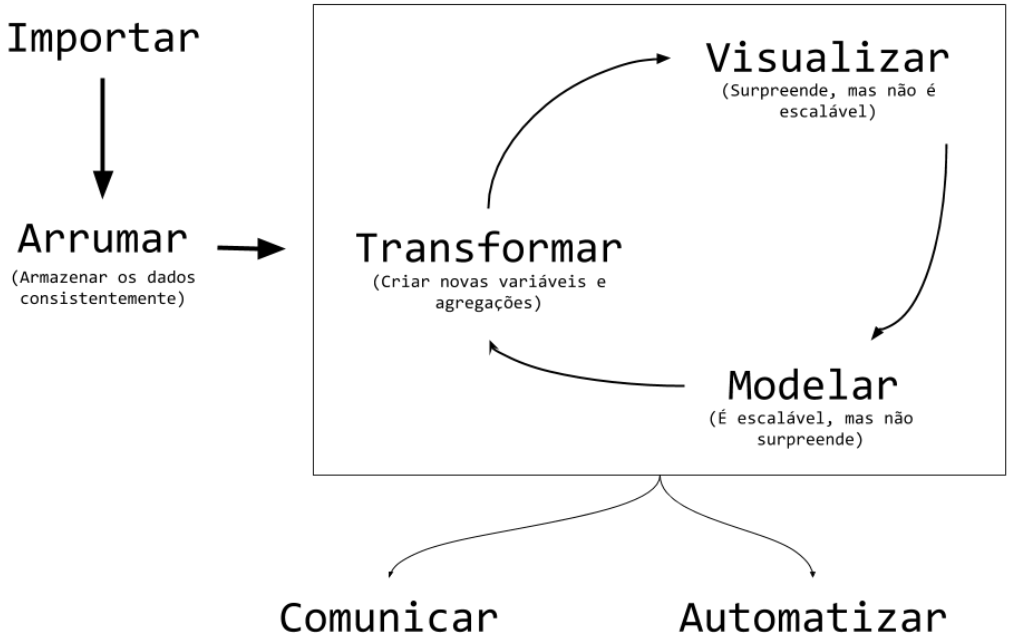
Os caminhos de uma pesquisa

1. **Definição dos objetivos**
   1.1. Análise Confirmatória vs Análise Exploratória
   1.2. Tradução: Linguagem jurídica -> linguagem analítica
2. **Listagem dos processos**
   2.1. Definição do escopo temporal, temático e regional
      2.1.1. TPUs e cifra oculta
3. **Extração e estruturação dos dados**
   3.1. Extração automática
   3.2. Extração manual
4. **Tratamento dos dados**
   4.1. Controle de inconsistências

Os caminhos de uma pesquisa

5. **Análise Descritiva**
   - 5.1. Visualização analítica (gráficos e tabelas)
   - 5.2. Criação de hipóteses
6. **Análise Inferencial**
   - 6.1. Modelos estatísticos e testes de hipótese
7. **Comunicação dos resultados**
   - 7.1. Tradução:
     Linguagem analítica ->
     linguagem jurídica
   - 7.2. Legal Design/Visual Law

O ciclo da ciência de dados



Como o Direito participa da Jurimetria?

1. **Listagem de processos**
   a. Direito dá o sistema e indica de onde os dados podem ser extraídos (rôle + competência)
2. **Extração dos dados**
   a. Extração automática: Direito ajuda a entender os dados que precisam ser extraídos (Documentos, movimentações, termos de busca)
   b. Extração manual: (A) Jurista ajuda montar o formulário (elencar respostas possíveis, controle lógico do formulário) e (B) jurista ajuda a classificar
3. **Tratamento dos dados**
   a. Controle de inconsistência: Jurista, por conhecer o processo, ajuda a mapear as possíveis inconsistências
4. **Análises**
   a. Operacionalização de conceitos: Jurista pode indicar qual é a melhor forma de fazer as transformações
   b. Estatístico entende os valores centrais; o jurista entende os valores extremos

Como o Direito participa da Jurimetria?

Antes de prosseguirmos com as etapas de análise é importante ter em mente:

1. Para que a Jurimetria funcione, precisamos da parceria entre um especialista em Estatística e um em Direito. Eles unem forças em todas as etapas do projeto.

2. O especialista em Direito orienta o estatístico sobre as particularidades jurídicas: seus termos, quais informações coletar e como torná-las úteis.

3. Por outro lado, o estatístico usa seu olhar analítico para decidir a melhor forma de avaliar os dados, mantendo o foco nos objetivos do estudo.

4. Eles são uma dupla dinâmica! Começam traduzindo termos jurídicos para a linguagem da estatística e finalizam revertendo as análises estatísticas para uma linguagem jurídica compreensível.

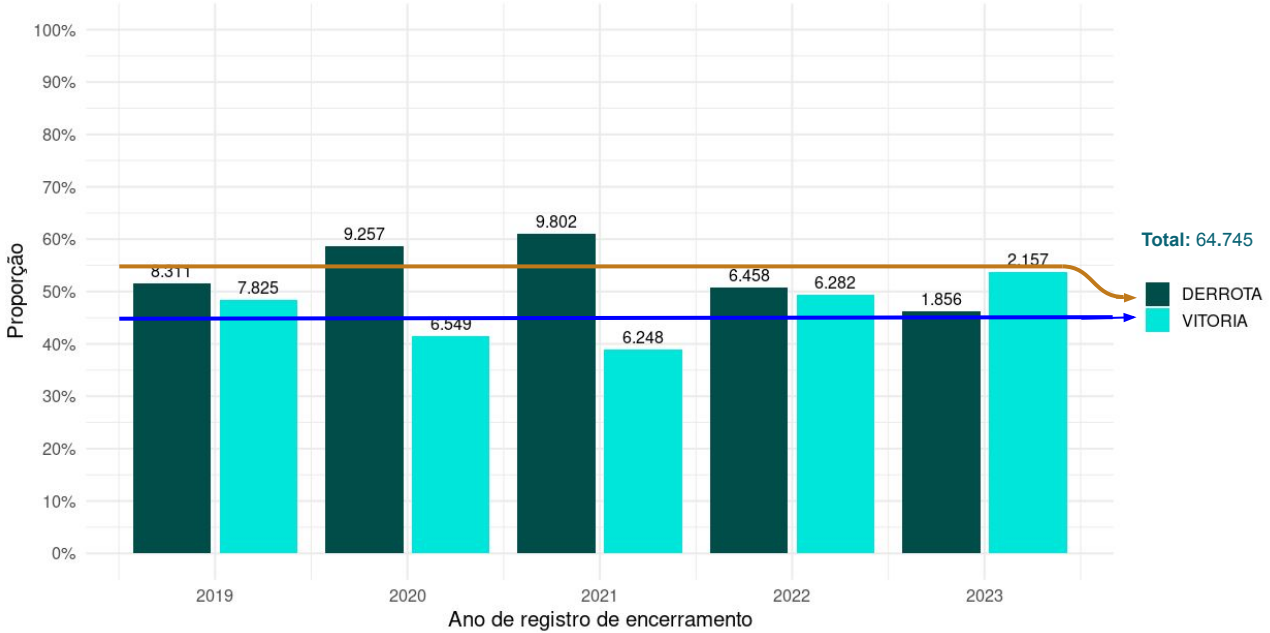
5. ANÁLISE DESCRITIVA

Introdução à análise descritiva

Análise Descritiva: É o primeiro estágio na avaliação de um conjunto de dados, usando técnicas para destacar suas características principais. Seu objetivo não é inferir conclusões, mas sim tornar os dados facilmente entendíveis.

Introdução à análise descritiva

Exemplo:



O gráfico ao lado mostra pra gente:

- Total de processos encerrados por ano
- Taxa de Derrota/Vitória por ano
- Número total de processos encerrados em todo o período
- Taxa de derrota média

Importância

- **Visão geral rápida:** A análise descritiva fornece uma visão geral imediata do que está contido no conjunto de dados.
- **É a base para análises futuras:** Antes de mergulhar em análises mais complexas, é vital entender a natureza básica dos seus dados.
- **Tomada de decisão:** Insights iniciais podem direcionar e influenciar decisões em estágios iniciais de investigações ou pesquisas.
- **Comunicação:** É mais fácil comunicar estatísticas e visualizações simples e diretas a um público que pode não ter familiaridade técnica.

As principais visualizações analíticas

Tipos básicos de gráficos

- **Histogramas:** Usados para visualizar a distribuição de um conjunto de dados.
- **Gráficos de Barra e Coluna:** Ideais para comparar quantidades de diferentes categorias.
- **Gráficos de Dispersão (Scatter Plots):** Mostram a relação entre duas variáveis quantitativas.
- **Box Plots (ou diagrama de caixa):** Representam a distribuição de dados quantitativos e ajudam a identificar outliers.
- **Gráficos de Linhas:** Úteis para visualizar tendências ao longo do tempo ou sequências ordenadas.

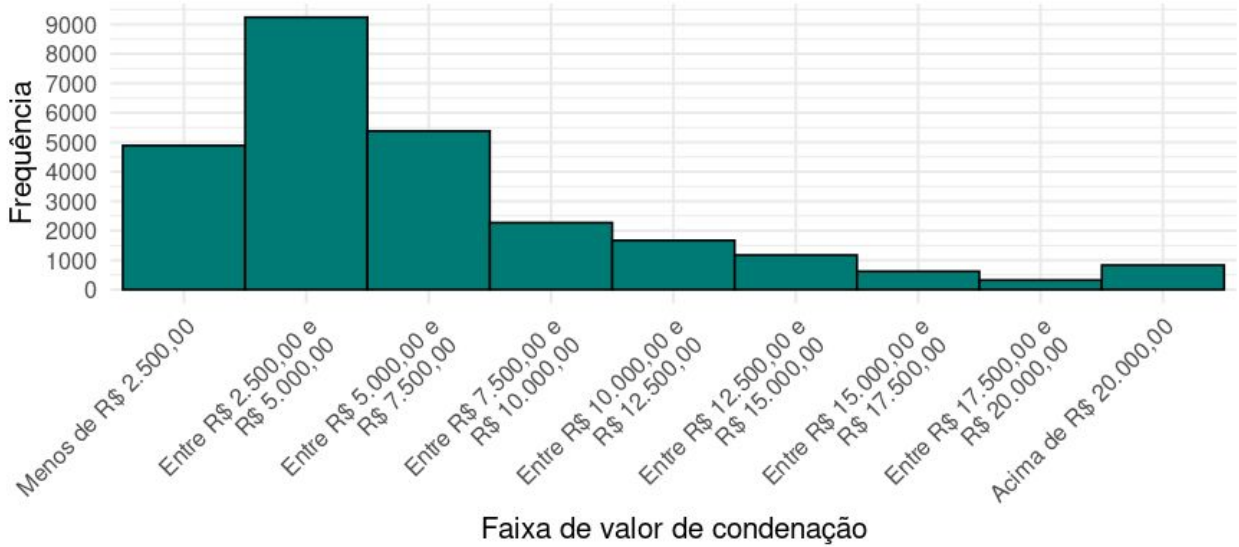
Tipos básicos de tabelas:

- **Tabelas de Frequência:** Mostram quantas vezes cada valor ou categoria aparece em um conjunto de dados.
- **Tabelas Cruzadas (Crosstabs):** Apresentam a relação entre duas variáveis categóricas, mostrando a frequência conjunta.

Tipos básicos de gráficos

Histogramas

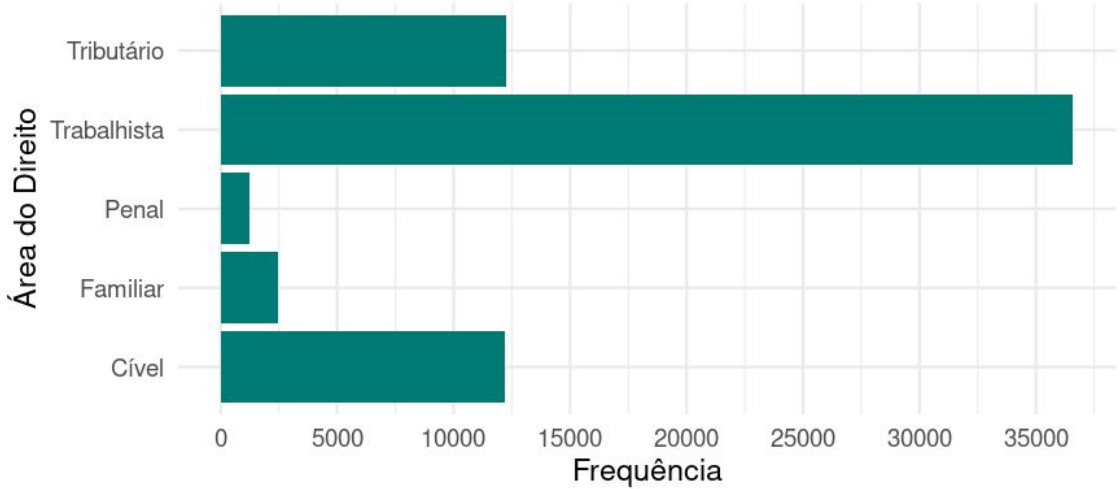
Descrição: Representação gráfica da distribuição de um conjunto de dados. Mostra a frequência com que diferentes classes de valores ocorrem.
Uso: Ideal para visualizar a distribuição de uma única **variável quantitativa contínua**, por exemplo, o valor dos processos judiciais.



Tipos básicos de gráficos

Gráficos de Barra/Coluna

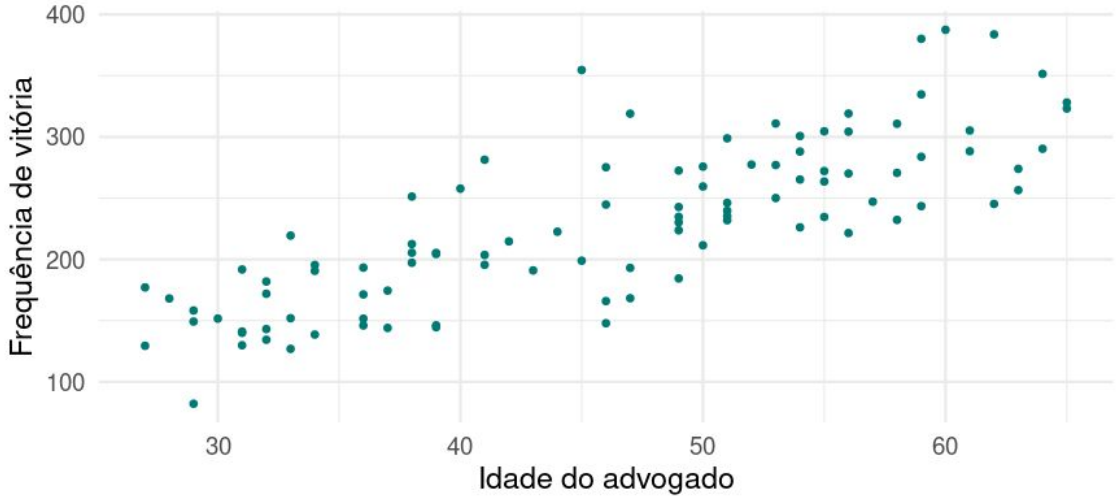
Descrição: Representa dados categóricos com barras retangulares, onde o comprimento de cada barra é proporcional ao valor que representa.
Uso: Ideal para comparar valores em diferentes categorias **(variável qualitativa)**, como o número de casos julgados por diferentes áreas do Direito.



Tipos básicos de gráficos

Gráficos de Dispersão

Descrição: Representa a relação entre duas **variáveis quantitativas**, mostrando cada par de valores como um ponto no gráfico.
Uso: Útil para identificar relações ou correlações entre variáveis, por exemplo, a relação entre a idade de um advogado e o número de casos ganhos.

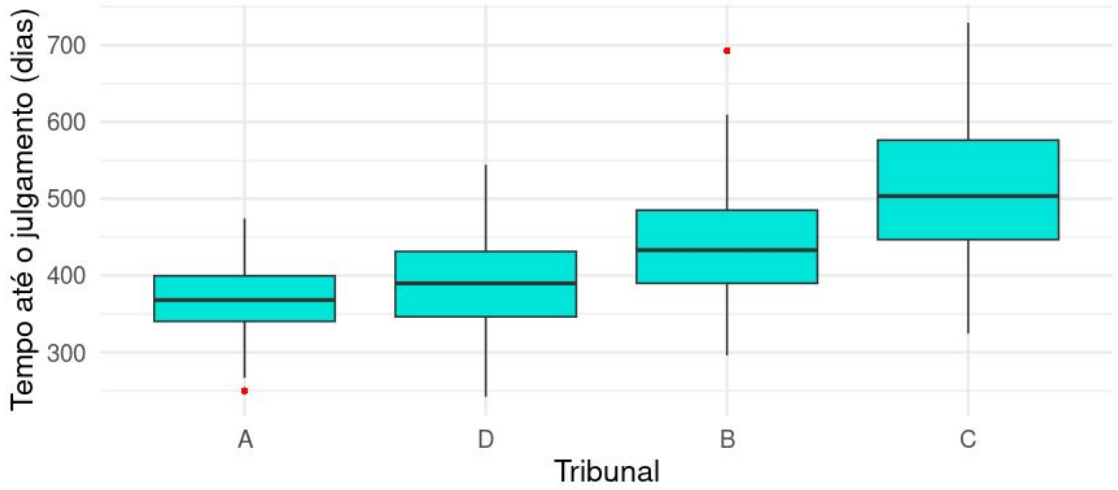


Tipos básicos de gráficos

Gráficos de Caixa (Box Plots)

Descrição: Mostra a distribuição de uma variável, indicando a mediana, quartis e possíveis outliers

Uso: Ótimo para visualizar a dispersão e simetria dos dados e para comparar a distribuição de várias categorias, como a duração de processos em diferentes tribunais.



Tipos básicos de gráficos

Gráficos de Linhas

Descrição: Mostra a evolução de uma ou mais variáveis ao longo do tempo.
Uso: Indicado para analisar tendências ao longo do tempo, como o aumento ou diminuição do número de ações judiciais por mês ou ano.

Tipos básicos de tabelas

Tabelas de Frequência

Descrição: Lista categorias e a quantidade de observações para cada categoria.
Uso: Utilizada para uma rápida visualização da distribuição de uma variável categórica, como o tipo de decisão em um conjunto de processos judiciais.

Resultado	Volume de processos
Favorável	1.250
Desfavorável	850
Em revisão	250
Anulado	40

Tipos básicos de tabelas

Tabelas Cruzadas (Crosstabs)

Descrição: Mostra a relação entre duas variáveis categóricas, apresentando a frequência conjunta.
Uso: Para identificar padrões ou associações entre duas variáveis, como a relação entre o tipo de crime e a sentença dada.

	Sentença		
Tipo de crime	**Condenado**	**Absolvido**	**Aguardando Julgamento**
Roubo	500	150	50
Fraude	200	300	100
Homicídio	100	30	20
Lesão Corporal	150	90	60

Observações iniciais e insights

Observações Iniciais: Antes de mergulhar mais fundo na análise dos dados, é fundamental realizar uma observação inicial para compreender a estrutura, características e padrões do conjunto de dados. Essas observações preliminares, muitas vezes conduzidas durante a fase de análise descritiva, servem como base para identificar tendências, anomalias ou características notáveis.

Insights: São os entendimentos que emergem dessas observações iniciais. Eles atuam como uma luz que ilumina áreas de interesse ou preocupação e podem levar a novas questões, investigações mais aprofundadas ou à formulação de hipóteses.

Importância: Esses insights iniciais são a base sobre a qual se constrói a pesquisa empírica. Eles fornecem uma direção, ajudam a refinar o foco da investigação e podem, muitas vezes, revelar padrões ou tendências que não eram inicialmente evidentes.

Traduzindo os insights em hipóteses (pré-operacionalização)

De Observações a Questões: Após as observações iniciais e ao reconhecer padrões ou tendências, deve-se formular questões. Por exemplo, ao observar que certos juízes parecem sentenciar de forma mais rigorosa em certos tipos de crimes, uma possível questão seria: "Existe uma relação entre o juiz e a severidade da sentença para esse tipo de crime?"

Formulando Hipóteses: Uma hipótese é uma declaração testável derivada de observações e questões. No exemplo anterior, uma hipótese poderia ser: "Juízes mais experientes tendem a dar sentenças mais rigorosas em casos de crimes financeiros."

Importância: A transição de observações para questões e depois para hipóteses é fundamental para estruturar a pesquisa. Fornecendo um caminho claro para testes empíricos e análises mais aprofundadas. O próximo passo é a operacionalização das hipóteses, conforme visto na aula anterior.

A diferença entre hipóteses nulas e alternativas

Hipótese Nula (H0): É uma afirmação de que não há diferença ou efeito e serve como uma linha de base para o teste. Por exemplo, "Não há diferença na severidade das sentenças entre juízes experientes e novatos em casos de crimes financeiros."

Hipótese Alternativa (Ha ou H1): É o oposto da hipótese nula e é o que você deseja testar ou provar. Seguindo o exemplo, "Juízes mais experientes dão sentenças mais rigorosas em casos de crimes financeiros do que juízes novatos."

Importância: As hipóteses nulas e alternativas formam a base dos testes estatísticos. Através do teste, busca-se rejeitar a hipótese nula em favor da alternativa. Esta estrutura permite uma abordagem sistemática e objetiva para determinar se os dados fornecem evidências suficientes para apoiar a hipótese de pesquisa.

Formulando uma hipótese

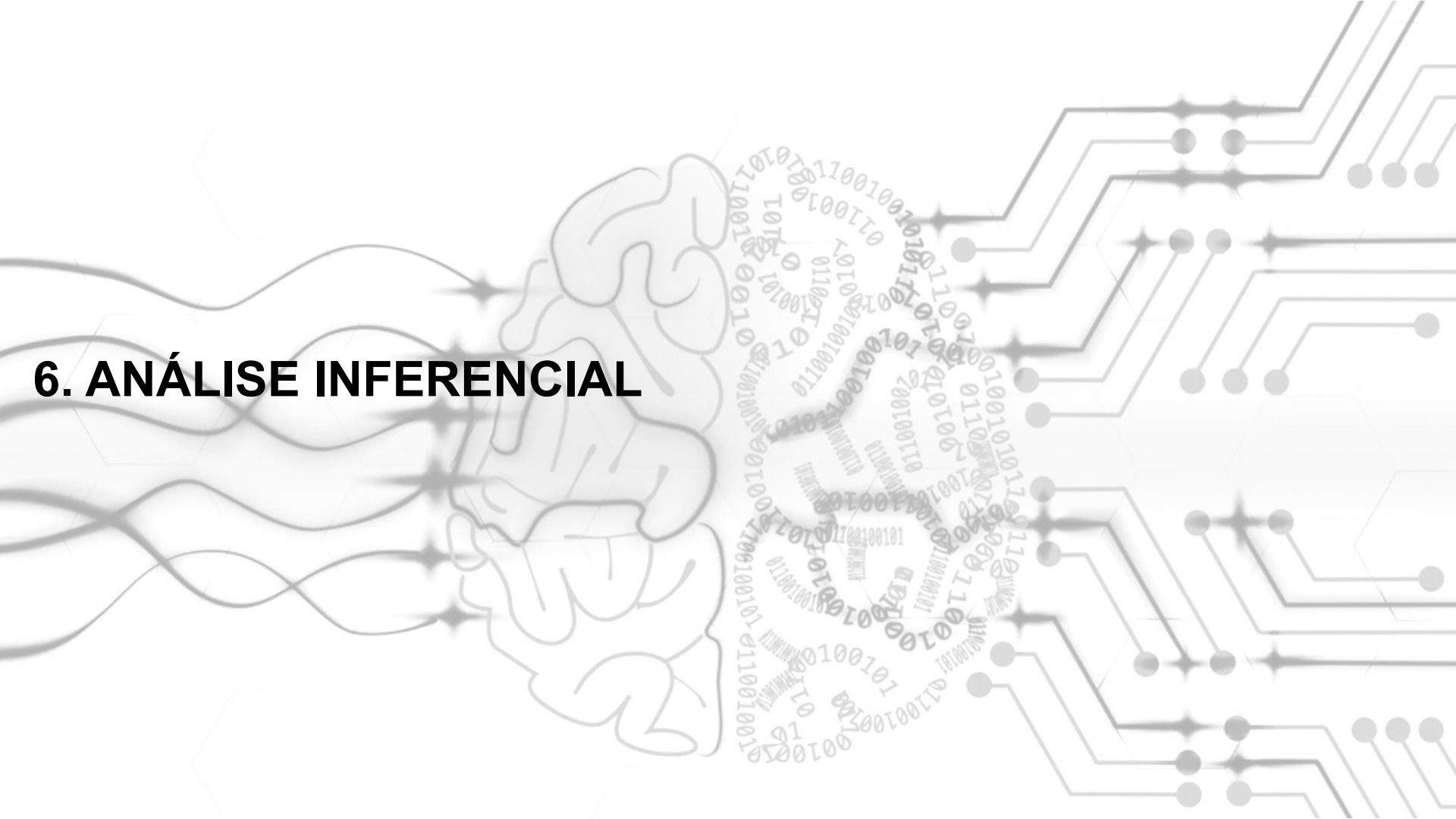
Problema: Alegação de que os pareceres técnicos emitidos pelos NatJus resultam em economia para o SUS, ao orientar decisões judiciais que evitam o fornecimento de medicamentos não comprovados ou de custo excessivo.
Questão: Os pareceres técnicos dos NatJus estão efetivamente associados a uma economia para o SUS?
Operacionalização: Diferença entre o **custo médio** de medicamentos prescritos em decisões judiciais com pareceres dos NatJus e o **custo médio** de medicamentos prescritos em decisões sem pareceres dos NatJus.
Hipótese Nula: Custo médio sem natjus ≥ **Custo médio** com Natjus

Em outras palavras: O custo médio de medicamentos prescritos em decisões judiciais sem pareceres dos NatJus é maior ou igual ao custo médio de medicamentos prescritos em decisões sem pareceres dos NatJus.

Possíveis resultados da hipótese

	A VERDADE REAL DE H_0	
DECISÃO	**H_0 é VERDADEIRA**	**H_0 é FALSA**
Não rejeitar H_0	Decisão Correta	Erro Tipo II
Rejeitar H_0	Erro Tipo I	Decisão Correta

6. ANÁLISE INFERENCIAL



Introdução à análise inferencial

Análise Inferencial: É uma abordagem estatística utilizada para fazer afirmações sobre uma população, com base nas informações obtidas de uma amostra dessa população. Ela nos permite "inferir" ou "deduzir" características ou padrões de um grupo maior (a população) utilizando os dados de um subconjunto menor (a amostra).

Fundamentos da inferência estatística

População: Refere-se ao conjunto completo de indivíduos, objetos ou dados de interesse. É o "todo" que queremos estudar ou entender. Por exemplo, se quisermos entender as opiniões de todos os advogados em um país sobre uma nova lei, a população seria **todos** os advogados desse país.

Amostra: É um subconjunto da população, selecionado de forma que represente adequadamente a população completa. No exemplo anterior, seria inviável consultar cada advogado, então poderíamos escolher um **grupo aleatório** (amostra) de advogados e inferir as opiniões da população completa a partir dessa amostra.

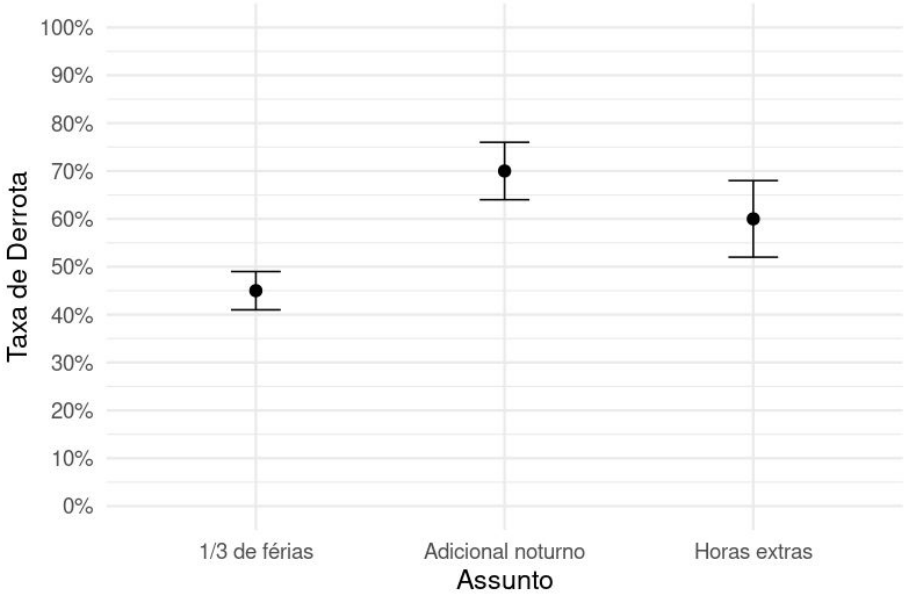
Margem de erro ou nível de confiança

Margem de Erro: É uma medida de quanto esperamos que os resultados da amostra difiram dos verdadeiros valores da população.

Nível de Confiança: Indica a probabilidade de que a margem de erro contenha o verdadeiro valor da população. Comumente, estudos no âmbito jurídico usam um nível de confiança de 95%, o que significa que há 95% de probabilidade de que o intervalo definido pela margem de erro contenha a verdadeira taxa de derrota na população.

Margem de erro ou nível de confiança

- Para os casos relacionados a **"Horas extras"**, cerca de 60% resultaram em derrota para o demandante. Considerando uma margem de erro de 8%, isso implica que a verdadeira taxa de derrota pode oscilar **entre 52% e 68%**.



- Em relação aos processos de "1/3 de férias", a análise mostra que aproximadamente 45% terminaram em derrota para o demandante. Com uma margem de erro de 4%, isso significa que o valor real da taxa de derrota pode variar entre 41% e 49%.
- Por último, nos casos ligados ao **"Adicional noturno"**, a taxa de derrota foi de 70%. Se levarmos em conta uma margem de erro de 6%, a verdadeira taxa pode estar **entre 64% e 76%**.

Modelagem estatística

Um modelo estatístico é uma representação matemática ou equação que descreve o relacionamento entre duas ou mais variáveis. Ele fornece uma estrutura para entender como variáveis específicas impactam ou são associadas a outras, permitindo previsões e inferências sobre os dados. O propósito principal de um modelo estatístico é simplificar a realidade complexa de maneira a torná-la analisável e compreensível.

$$\bar{Y}_i = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3}}$$

Modelos mais comuns

Regressão Linear: Estima o relacionamento entre duas ou mais variáveis. Em sua forma mais simples (regressão linear simples), relaciona uma variável independente a uma variável dependente.
Regressão Logística: Usada quando a variável resposta (ou dependente) é binária ou categórica (por exemplo, sim/não, verdadeiro/falso). É comumente empregada para prever a probabilidade de um evento ocorrer.
Análise de Variância (ANOVA): Comparar médias entre três ou mais grupos. Por exemplo, avaliar se três diferentes técnicas de argumentação têm efeitos distintos sobre a decisão de um juiz.
Modelos de Séries Temporais: Analisar dados que são coletados ao longo do tempo. Isso pode incluir análise de tendências ou previsões de eventos futuros.
Modelos de Sobrevivência: Estudam o tempo até a ocorrência de um evento. No contexto jurídico, pode ser utilizado para analisar o tempo até a resolução de um caso.

O GPT também é um modelo estatístico

Cada Palavra é um Número: Quando você escreve algo para o GPT, ele transforma cada palavra em um número (ou conjunto de números).

Entendendo o Contexto: Se você perguntasse a um colega advogado sobre "princípio da inocência", ele entenderia que você está falando de direito penal. Da mesma forma, o GPT usa o contexto das palavras ao redor para entender melhor o que você quer dizer.

Gerando Respostas: Usando os "neurônios artificiais" e todo o seu treinamento, o GPT gera uma resposta que acredita ser a mais relevante para sua pergunta.

O GPT erra! Existe uma probabilidade associada, não só à próxima palavra escolhida, como também no contexto e no seu significado.

Diferenças entre Análise Descritiva e Análise Inferencial

Análise Descritiva

- **Objetivo:** Descrever e sumarizar dados.
- **Métodos:** Uso de médias, medianas, modas, gráficos, tabelas, entre outros.
- **Aplicação:** Se você deseja entender a estrutura e os padrões básicos dos seus dados, usa-se a análise descritiva.
- **Limitações:** Não faz previsões ou inferências. Só descreve o que está no conjunto de dados específico.

Análise Inferencial

- **Objetivo:** Fazer previsões/inferências sobre uma população com base em uma amostra.
- **Métodos:** Uso de testes de hipótese, regressão, análise de variância, entre outros.
- **Aplicação:** Se você deseja fazer uma previsão sobre um grupo maior com base em um subconjunto (amostra) ou testar uma teoria, usa-se a análise inferencial.
- **Limitações:** A precisão das inferências pode ser afetada por tamanho de amostra, vieses e outros fatores.

Diferenças entre Análise Descritiva e Análise Inferencial

Em Resumo: Enquanto a análise descritiva fornece um retrato dos seus dados, a análise inferencial tenta fazer previsões ou inferências além desses dados. Ambas são ferramentas cruciais no arsenal de qualquer pesquisador, especialmente em campos interdisciplinares como a jurimetria.

7. COMUNICAÇÃO DOS RESULTADOS

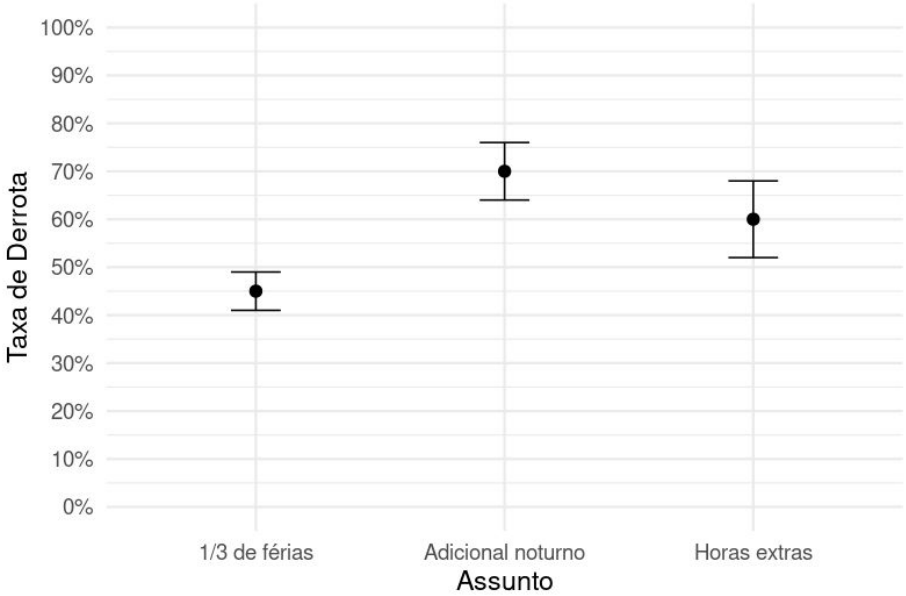
Tradução: Linguagem analítica -> linguagem jurídica

Objetivo da tradução: Transformar termos técnicos em informações acessíveis.
Exemplo: Converter um "p-valor de 0,05" em "evidência estatisticamente significativa".

Interpretação de resultados: Coeficientes, p-valores, margem de erro… e o que significam.
Exemplo: Coeficiente positivo indica correlação direta.

Criação de narrativas: Construir uma história baseada nos dados.
Exemplo: Mostrar uma tendência de crescimento de certos tipos de casos ao longo dos anos.

Linguagem Estatística



Horas extras: Taxa de derrota de 60% com margem de erro de +/- 8%.
1/3 de férias: Taxa de derrota de 45% com margem de erro de +/- 4%.
Adicional noturno: A maior taxa de derrota de 70% com margem de erro de +/- 6%.

Essas margens de erro correspondem a um nível de confiança de 95%, o que significa que estamos 95% confiantes de que as taxas reais estão dentro desses intervalos.

Importante notar que os intervalos de confiança para "Horas extras" e "Adicional noturno" se sobrepõem, o que sugere que a diferença entre essas categorias pode não ser tão significativa quanto os números brutos podem sugerir.

Conclusão:

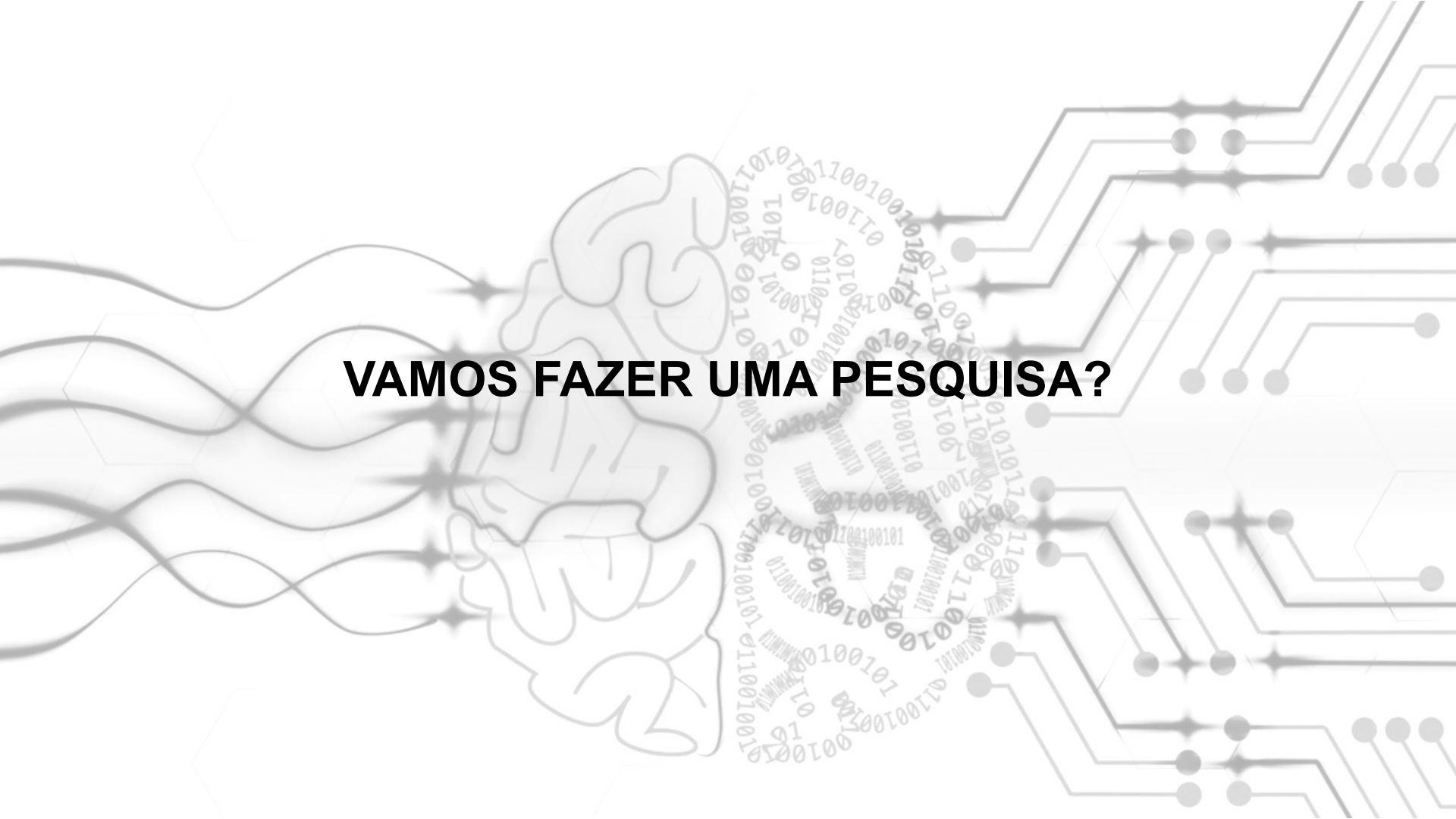
As taxas nos ajudam a entender melhor os desafios em cada categoria. Contudo, é essencial considerar as margens de erro e sobreposições ao tomar decisões estratégicas.

Linguagem Jurídica

Em nossa análise sobre os processos trabalhistas, identificamos que os pleitos relacionados a "Horas extras" não são bem-sucedidos em 60% dos casos. Já os processos que tratam do direito a "1/3 de férias" apresentam um índice de insucesso de 45%. Por outro lado, as ações que envolvem o "Adicional noturno" tendem a ser desfavoráveis ao requerente em 70% das vezes.

É fundamental destacar que, embora os números indiquem um maior insucesso no tema de "Adicional noturno" quando comparado a "Horas extras", a diferença entre eles pode não ser tão pronunciada quanto parece. Essa observação sugere que, na prática, os desafios em ganhar casos nas duas categorias podem ser mais similares do que os números isolados mostram.

Em resumo, ao planejar estratégias para futuros litígios, é crucial considerar essas tendências, ajustando nossas expectativas e abordagens de acordo com cada tema em questão.


VAMOS FAZER UMA PESQUISA?

O problema

Uma renomada empresa do setor educacional possui muitos processos judiciais em andamento. Ela contratou um especialista em jurimetria para entender melhor quais desses processos têm maior risco de resultar em perdas financeiras. O objetivo é criar uma estratégia para priorizar acordos nos casos mais arriscados financeiramente.

O cronograma

Fases do projeto:

1. **Planejamento**
    a. **Definição do escopo e objetivos**
    b. **Identificação das fontes/bases de dados**
    c. **Desenho de uma estratégia eficiente da coleta de dados**
    d. **Seleção de metodologias a serem testadas**
    e. **Definição do cronograma definitivo**
2. **Coleta dos dados**
    a. **Obtenção dos dados**
    b. **Validação e tratamento das informações**
    c. **Estruturação das bases de dados**
3. **Desenvolvimento dos modelos**
    a. **Análise descritiva**
    b. **Seleção de variáveis candidatas**
    c. **Treinamento dos modelos**
    d. **Teste de validação e ajustes**
    e. **Aplicação dos resultados na base de dados**
4. **Automatização da periodicidade das análises.**

Passo 1: Definição dos objetivos

Objetivo/Conceito: Criar uma estratégia para priorizar acordos nos casos mais arriscados financeiramente.

Questões-chave Utilizar, dentre outras informações, a base processual interna do contencioso Cível da companhia e responder: Como as diferentes características dos processos influenciam no **resultado**, **tempo** e **valor** dos processos? (Vejam, a questão-chave já “operacionaliza” alguns conceitos.)

Operacionalização:
i) um modelo estatístico capaz de estimar a probabilidade de perda dos processos ativos com base nas características do caso.
ii) um modelo estatístico capaz de estimar o tempo até o encerramento dos processos.
iii) um modelo estatístico capaz de estimar o valor de condenação a ser pago, em caso de perda.

Passo 2: Listagem dos processos

Pergunta: Como deve ser feita a listagem dos processos nesse caso?

As bases de dados disponíveis para o estudo estão alocadas no sistema de gerenciamento processual da companhia. Além das informações dos processos, também podem ser relevantes informações cadastrais e de inadimplência dos alunos (RH) e de pagamentos (contabilidade).

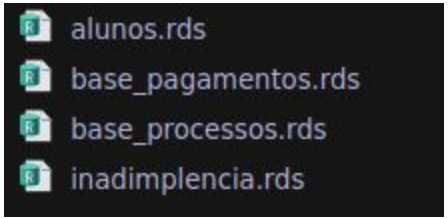
Passo 3: Extração das informações

Pergunta: E a extração? Deve ser feita de maneira manual ou automática?

Como os dados estão cadastrados internamente na companhia, sempre é mais interessante coletar os dados de fontes internas. Porque i) são mais ricas (por exemplo, informações sobre valor de condenação já aparecem de maneira estruturada) e ii) a extração muitas vezes já está automatizada pelo próprio sistema de gerenciamento

Passo 3: Extração das informações

As informações brutas foram extraídas de todos os sistemas disponíveis e resultaram em 4 bases de dados:



Alunos:
1.2 milhões de linhas
Unidade amostral:
Aluno - Curso
Unidades observacionais:

- Informações pessoais
- Ano de ingresso
- Status
- etc

Pagamentos:
90k linha
Unidade amostral:
Processo - pagamento
Unidades observacionais:

- Processo
- Beneficiário
- Informações bancárias
- data do pagamento
- informações processuais
- etc

Inadimplência:
35 milhões de linhas
Unidade amostral:
Aluno - curso - mês
Unidades observacionais:

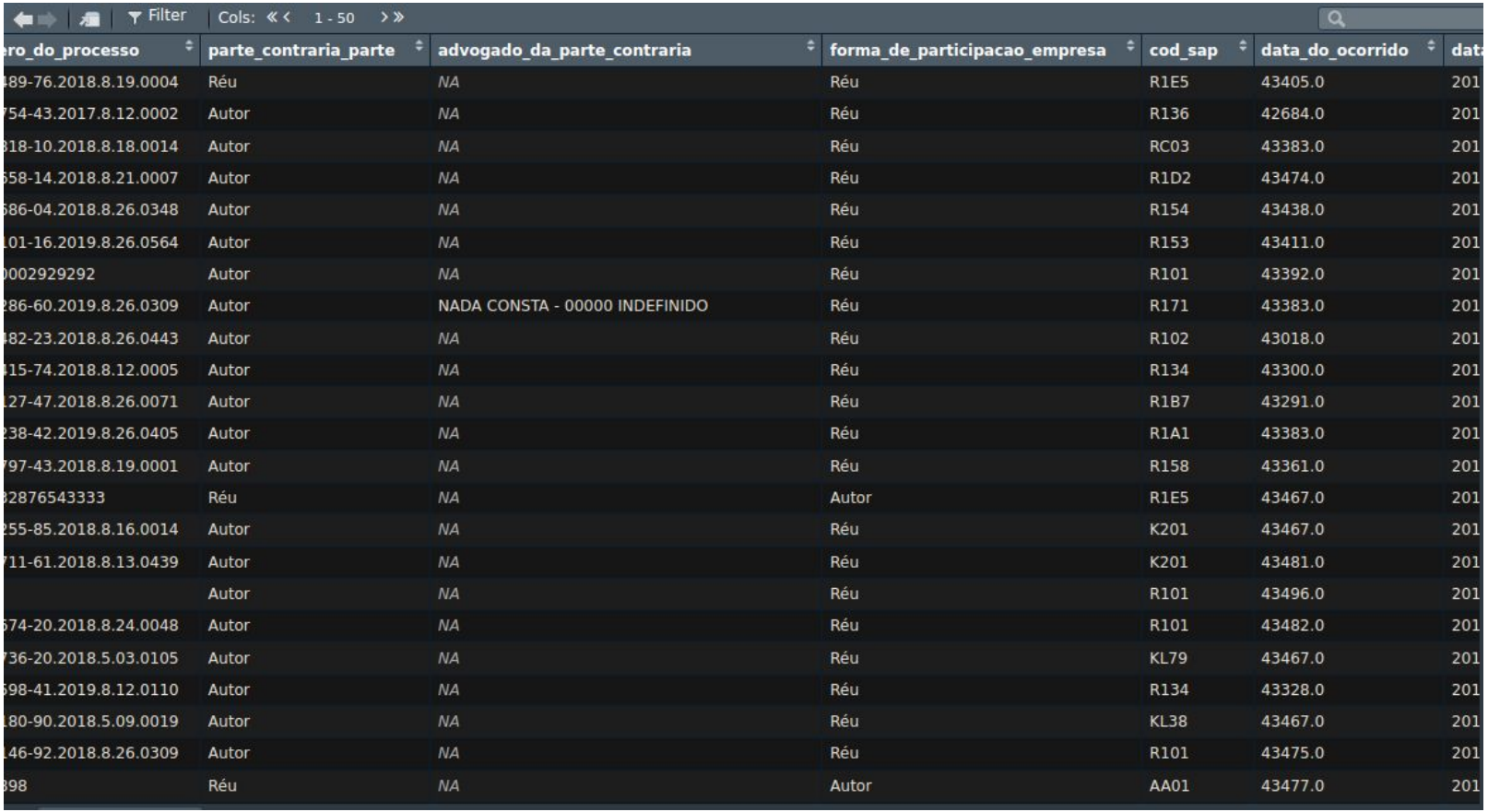
- Informações pessoais
- Saldo devedor
- Nome negativado?
- etc

Passo 3: Extração das informações

Processos:
150 mil linhas
Unidade amostral:
Processo
Unidades observacionais:

- Partes
- Advogados
- Empresa
- Participação da empresa
- filial
- datas (distribuição, julgamento, encerramento)
- comarca
- foro
- vara
- status
- esfera
- etc

Passo 3: Extração das informações



Filter | Cols: « < 1 - 50 > »

ero_do_processo	parte_contraria_parte	advogado_da_parte_contraria	forma_de_participacao_empresa	cod_sap	data_do_ocorrido	dat
489-76.2018.8.19.0004	Réu	*NA*	Réu	R1E5	43405.0	201
754-43.2017.8.12.0002	Autor	*NA*	Réu	R136	42684.0	201
818-10.2018.8.18.0014	Autor	*NA*	Réu	RC03	43383.0	201
658-14.2018.8.21.0007	Autor	*NA*	Réu	R1D2	43474.0	201
686-04.2018.8.26.0348	Autor	*NA*	Réu	R154	43438.0	201
101-16.2019.8.26.0564	Autor	*NA*	Réu	R153	43411.0	201
0002929292	Autor	*NA*	Réu	R101	43392.0	201
286-60.2019.8.26.0309	Autor	NADA CONSTA - 00000 INDEFINIDO	Réu	R171	43383.0	201
482-23.2018.8.26.0443	Autor	*NA*	Réu	R102	43018.0	201
415-74.2018.8.12.0005	Autor	*NA*	Réu	R134	43300.0	201
127-47.2018.8.26.0071	Autor	*NA*	Réu	R1B7	43291.0	201
238-42.2019.8.26.0405	Autor	*NA*	Réu	R1A1	43383.0	201
797-43.2018.8.19.0001	Autor	*NA*	Réu	R158	43361.0	201
32876543333	Réu	*NA*	Autor	R1E5	43467.0	201
255-85.2018.8.16.0014	Autor	*NA*	Réu	K201	43467.0	201
711-61.2018.8.13.0439	Autor	*NA*	Réu	K201	43481.0	201
	Autor	*NA*	Réu	R101	43496.0	201
674-20.2018.8.24.0048	Autor	*NA*	Réu	R101	43482.0	201
736-20.2018.5.03.0105	Autor	*NA*	Réu	KL79	43467.0	201
698-41.2019.8.12.0110	Autor	*NA*	Réu	R134	43328.0	201
180-90.2018.5.09.0019	Autor	*NA*	Réu	KL38	43467.0	201
146-92.2018.8.26.0309	Autor	*NA*	Réu	R101	43475.0	201
398	Réu	*NA*	Autor	AA01	43477.0	201

Passo 4: Tratamento

Pergunta: Por onde vocês começariam? Quais são as inconsistências mais graves que poderiam existir, dado o objetivo da análise?

Não existe uma regra de “por onde começar”. Entretanto, as variáveis resposta são sempre as informações mais importantes. São elas que conduzem a análise. Se a variável resposta estiver inconsistente a ponto de não poder ser estudada, então a análise está condenada.

Passo 4: Tratamento

O modelo irá classificar Vitoria/Derrota. O que fazer?

resultado	n
<chr>	<int>
Procon - resposta apresentada	35066
NA	29033
Procedente Em Parte	18841
Acordo	18177
Procedente	16524
Improcedente	12768
Extinto sem resolução de mérito	8572
PROCON - Fundamentada Atendida	3646
Desistência Da Ação	2188
Procedente em parte - apenas OBF	1636
Administrativo - arquivado mediante pagamento de AI	483
PROCON - Fundamentada Não Atendida	319
PROCON - Não Fundamentada	319
Extinto Por Inércia Da Parte Contrária	260
Administrativo - arquivado sem constatação de irregularidades	224
Extinto Pela Prescrição	204
Arquivamento promovido pelo MP	187
Parcelamento / Anistia	41
PROCON - Fundamentada Não Atendida (SEM SANÇÃO)	8
Acordo, antes da decisão	6
Procon ¿ resposta apresentada	1

status	resultado	n
<chr>	<chr>	<int>
Ativo	NA	20130
Encerrado	NA	6825
Removido	NA	1168
Baixa Provisória	NA	904
Pré-Cadastro	NA	6

Passo 4: Tratamento

Se o resultado for “Improcedente” e a empresa estiver no polo passivo -> **Vitória**
Se o resultado for “Improcedente” e a empresa estiver no polo ativo -> **Derrota**
Se o resultado for “Procedente” e a empresa estiver no polo ativo -> **Vitória**
Se o resultado for “Procedente” e a empresa estiver no polo passivo -> **Derrota**
E se o resultado for **Acordo**?
E se o resultado for “extinto sem resolução do mérito”?

Depende.

Opção 1: Excluir da base o que não é vitória/derrota
Opção 2: Definir melhor o que é vitória e derrota

Passo 4: Tratamento

```
processo_uf = limpa_string(estado),
processo_comarca = limpa_string(comarca),
processo_status = limpa_string(status),
processo_esfera = limpa_string(esfera),
processo_tipo_acao = limpa_string(acao),
processo_fase = limpa_string(fase),
processo_escritorio = limpa_string(escritorio_externo),
processo_risco = limpa_string(risco),
processo_objeto = limpa_string(objeto),
processo_objeto_da_acao = limpa_string(objeto_da_acao),
processo_subobjeto = limpa_string(sub_objeto),
processo_aluno_tem_razao = limpa_string(aluno_tem_razao),
processo_resultado = limpa_string(resultado),
processo_resultado_trat1 = dplyr::case_when(
  processo_resultado %in% c("EXTINTO SEM RESOLUCAO DO MERITO","EXTINTO POR INERCIA DA PARTE CONTRARIA", "EXTINTO PE
  forma_de_participacao_empresa == 'Réu' & processo_resultado %in% c("IMPROCEDENTE") ~ 'VITORIA',
  forma_de_participacao_empresa == 'Autor' & processo_resultado %in% c("PROCEDENTE") ~ 'VITORIA',
  forma_de_participacao_empresa == 'Réu' & processo_resultado %in% c("PROCEDENTE") ~ 'DERROTA',
  forma_de_participacao_empresa == 'Autor' & processo_resultado %in% c("IMPROCEDENTE") ~ 'DERROTA',
  processo_resultado %in% c("PROCEDENTE EM PARTE", "PROCEDENTE EM PARTE - APENAS OBF") ~ 'FAVORAVEL EM PARTE',
  is.na(processo_resultado) ~ NA_character_,
  processo_resultado %in% c("ACORDO","ACORDO, ANTES DA DECISAO", "PARCELAMENTO / ANISTIA") ~ 'ACORDO',
  TRUE ~ 'OUTRO'

),
processo_resultado_da_acao = limpa_string(resultado_da_acao),
processo_resultado_da_acao_trat1 = dplyr::case_when(
  processo_resultado_da_acao %in% c("ACAO DESFAVORAVEL A CIA","ACAO DESFAVORAVEL - APENAS OBF") ~ 'DERROTA',
  processo_resultado_da_acao %in% c("ACAO FAVORAVEL A CIA") ~ 'VITORIA',
```

Passo 4: Consolidação da base de dados final

Filtros:

- Área do direito: Cível
- Unidade de negócio: diferente de Vazio
- Esfera: Cível, Judicial, Juizado Especial Cível, Juizado Especial Federal, Justiça Comum e Justiça Federal
- Registro: Data de encerramento não vazia e posterior a 2022
- Polo: Companhia no polo passivo,
- Tipo de ação: Que envolve condenação em valores (lista enviada pelo time do jurídico)
- Responsabilidade: Excluir “0%”, “Vendedor” e Vazio
- Percentual de responsabilidade: Diferente de “0%”
- Filial: Apenas o grupo indicado pela equipe do jurídico

Resposta:

Se tiver pagamento referente à condenação/acordo registrado na base de pagamentos **Então** Derrota
Se não, **Então** Vitória

TOTAL DE OBSERVAÇÕES NA BASE: 32.682

Passo 4: Consolidação da base de dados final

Risco.

Além desta, existem ainda outras duas bases: **Tempo** e **Valor**.

processo_id	resposta	ano_cnj	processo_uf	processo_comarca	processo_esfera	processo_tipo_acao
1295	DERROTA	2017	RJ	OUTRO	JUDICIAL	ACAO DECLARATORIA
1317	DERROTA	2017	SP	OUTRO	JUSTICA COMUM	ACAO DECLARATORIA
1429	DERROTA	2013	MT	OUTRO	JUSTICA COMUM	ACAO ORDINARIA
1499	VITORIA	2014	GO	ANAPOLIS	JUSTICA COMUM	ACAO DE OBRIGACAO DE
1509	DERROTA	2015	MT	CUIABA	JUSTICA COMUM	ACAO DE INDENIZACAO I
1535	DERROTA	2015	MS	CAMPO GRANDE	JUSTICA COMUM	ACAO DE INDENIZACAO I
1537	DERROTA	2015	MG	OUTRO	JUSTICA COMUM	ACAO DE OBRIGACAO DE
1538	DERROTA	2015	RS	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
1543	DERROTA	2014	MT	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
1548	VITORIA	2015	MT	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
1562	DERROTA	2015	OUTRO	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
1567	DERROTA	2014	MG	OUTRO	JUSTICA COMUM	ACAO DECLARATORIA
1571	VITORIA	2016	RS	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
1572	DERROTA	2015	RJ	SAO GONCALO	JUSTICA COMUM	ACAO DE OBRIGACAO DE
1578	DERROTA	2016	SC	OUTRO	JUSTICA COMUM	ACAO DE OBRIGACAO DE
1579	DERROTA	2015	PR	OUTRO	JUSTICA COMUM	OUTRO
1580	DERROTA	2016	GO	ANAPOLIS	JUIZADO ESPECIAL CIVEL	ACAO DE INDENIZACAO I
1616	VITORIA	2016	ES	OUTRO	JUSTICA COMUM	ACAO DE OBRIGACAO DE
1625	DERROTA	2016	MA	OUTRO	JUSTICA COMUM	ACAO DECLARATORIA
1640	DERROTA	2016	PA	OUTRO	JUSTICA COMUM	ACAO DECLARATORIA
1649	DERROTA	2016	OUTRO	OUTRO	JUSTICA COMUM	ACAO DE INDENIZACAO I
983	DERROTA	2018	RJ	NITEROI	JUSTICA COMUM	OUTRO

Passo 5: Análise Descritiva

Recapitulando

Objetivo: Criar modelos estatísticos para prever resultado, tempo e valor dos processos.

Pergunta: Dados esses objetivos, o que é interessante resumir da base de dados e apresentar em uma análise descritiva?

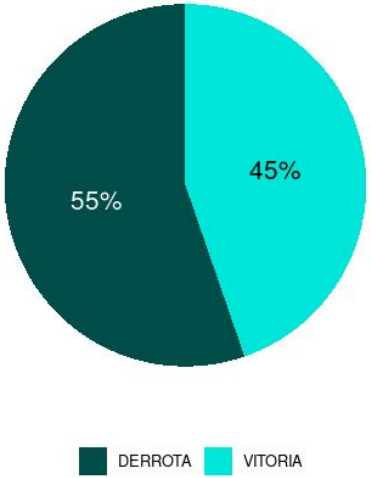
Passo 5: Análise Descritiva

A primeira coisa que devemos pensar é em nossa variável resposta, isto é, o que estamos querendo estudar.

Por exemplo: “Resultado do processo” é a principal informação que deve ser analisada, se queremos criar um modelo para “resultado”.
Analogamente, se queremos criar um modelo para prever o “valor de condenação”, a principal informação é justamente o “valor de condenação” dos processos.
Tudo o que você quer “prever” é a sua **Variável resposta.**
Tudo o que você **relacionar** com a variável resposta, chamamos de **variável explicativa.**

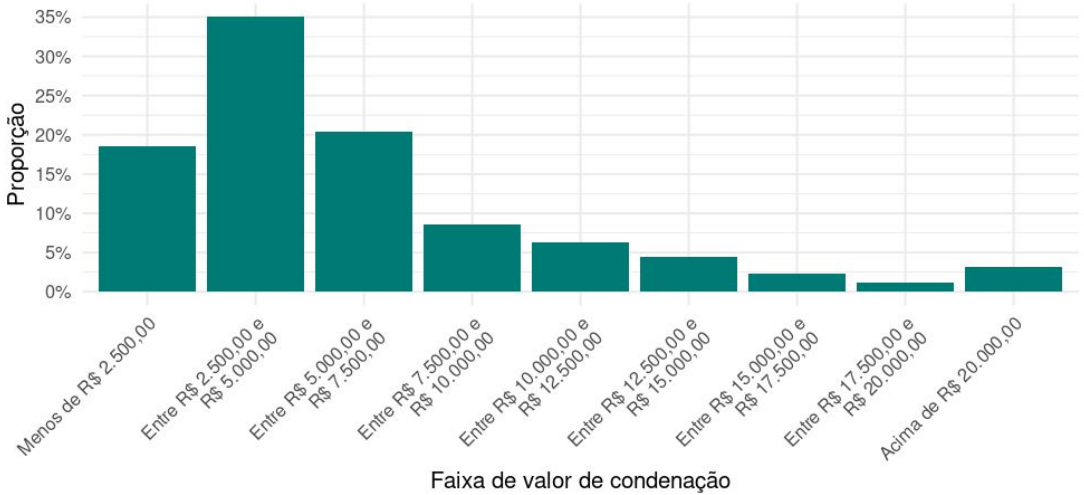
A **Variável Resposta** é explicada pelas **variáveis explicativas**

Passo 5: Análise Descritiva



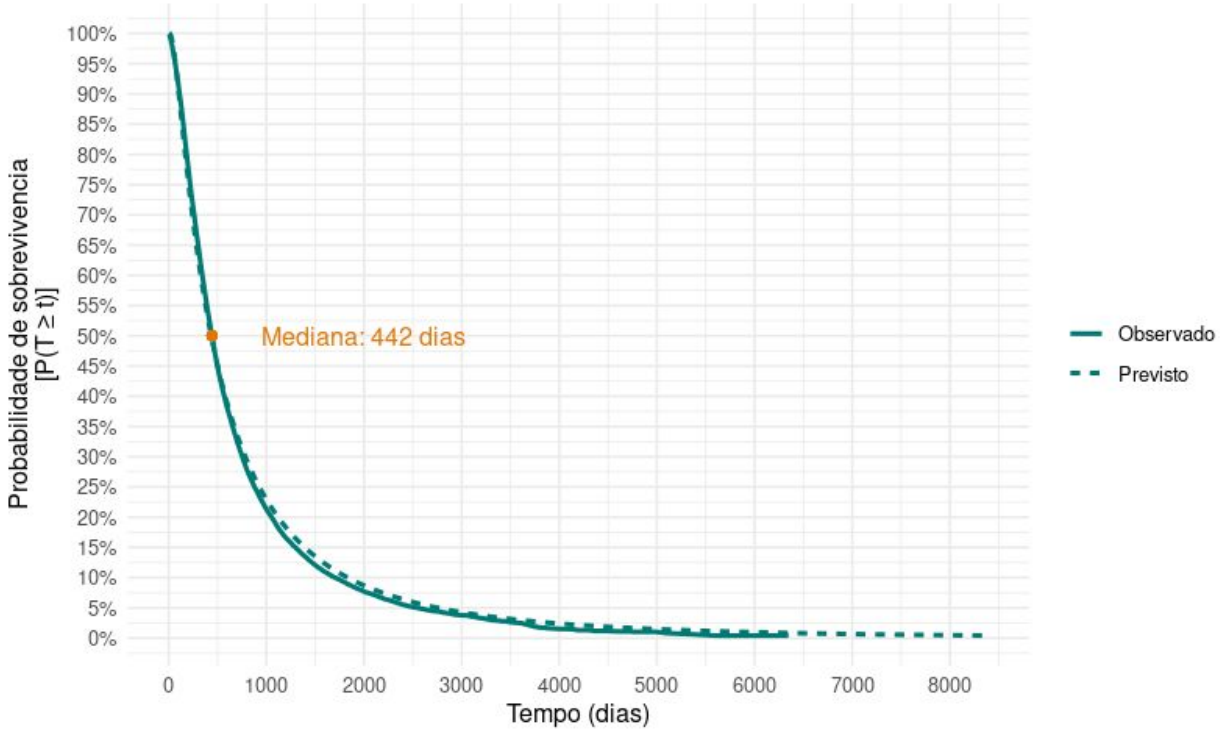
Distribuição do resultado

Distribuição do valor de condenação



Passo 5: Análise Descritiva

Distribuição do tempo até o encerramento

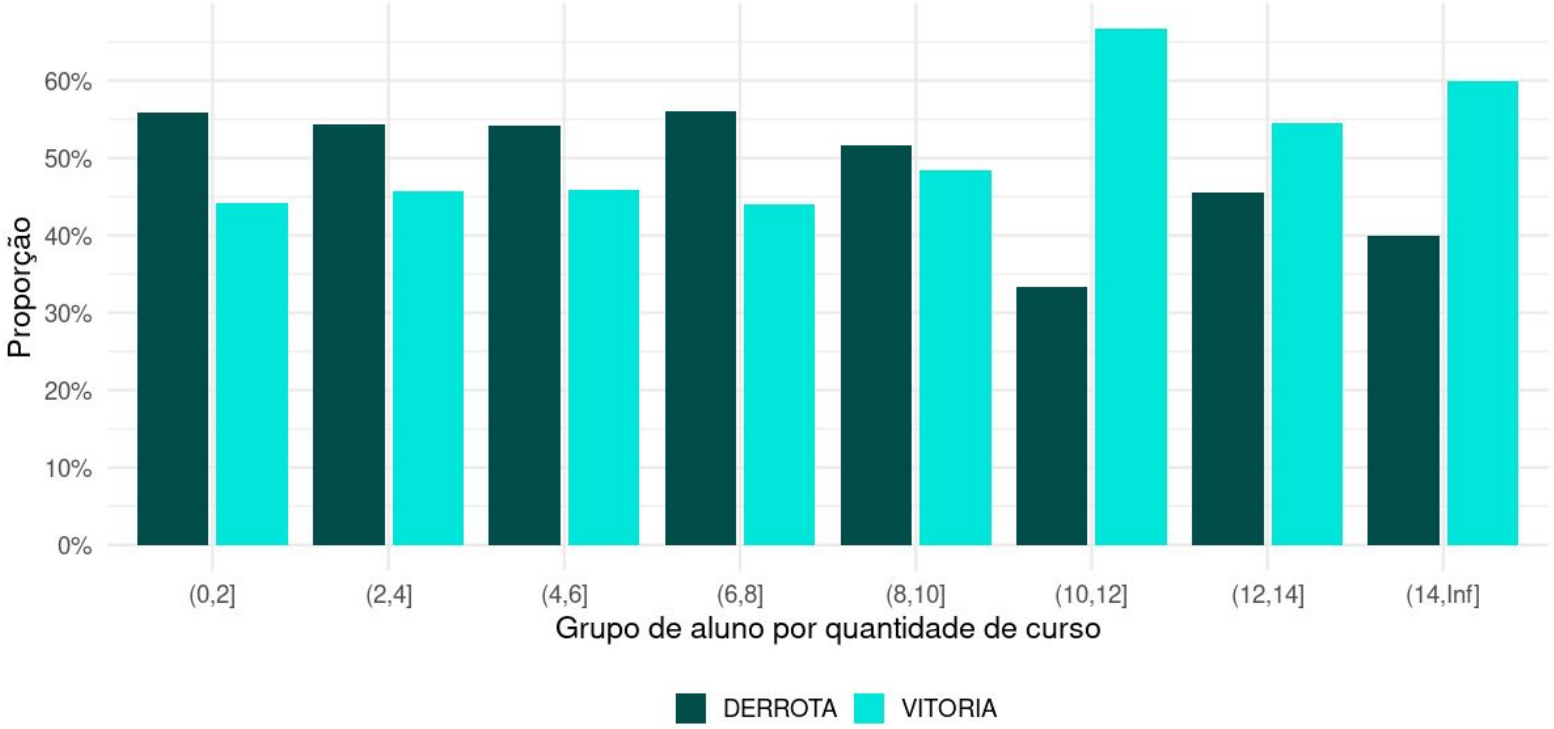


Passo 5: Análise Descritiva

Resultado x Comarca

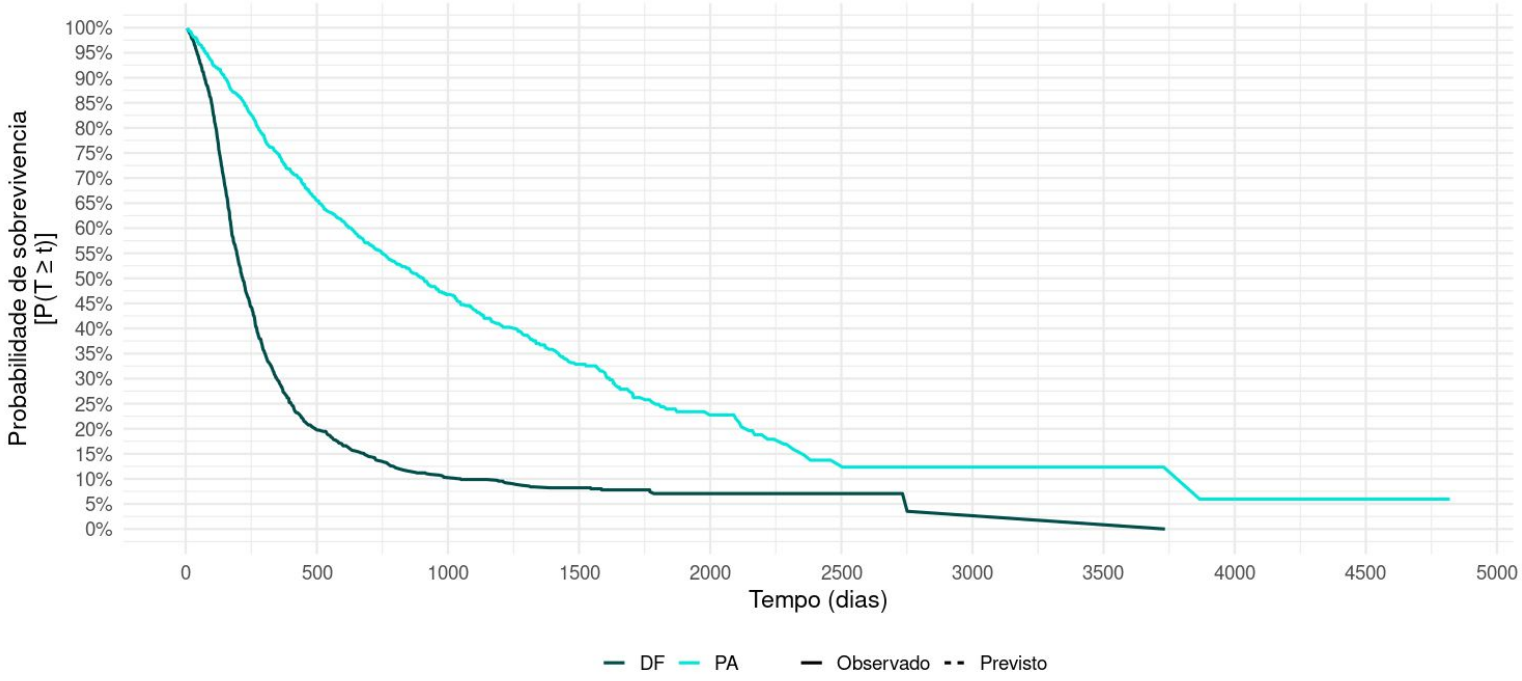
Passo 5: Análise Descritiva

Resultado x grupo de alunos por quantidade de cursos



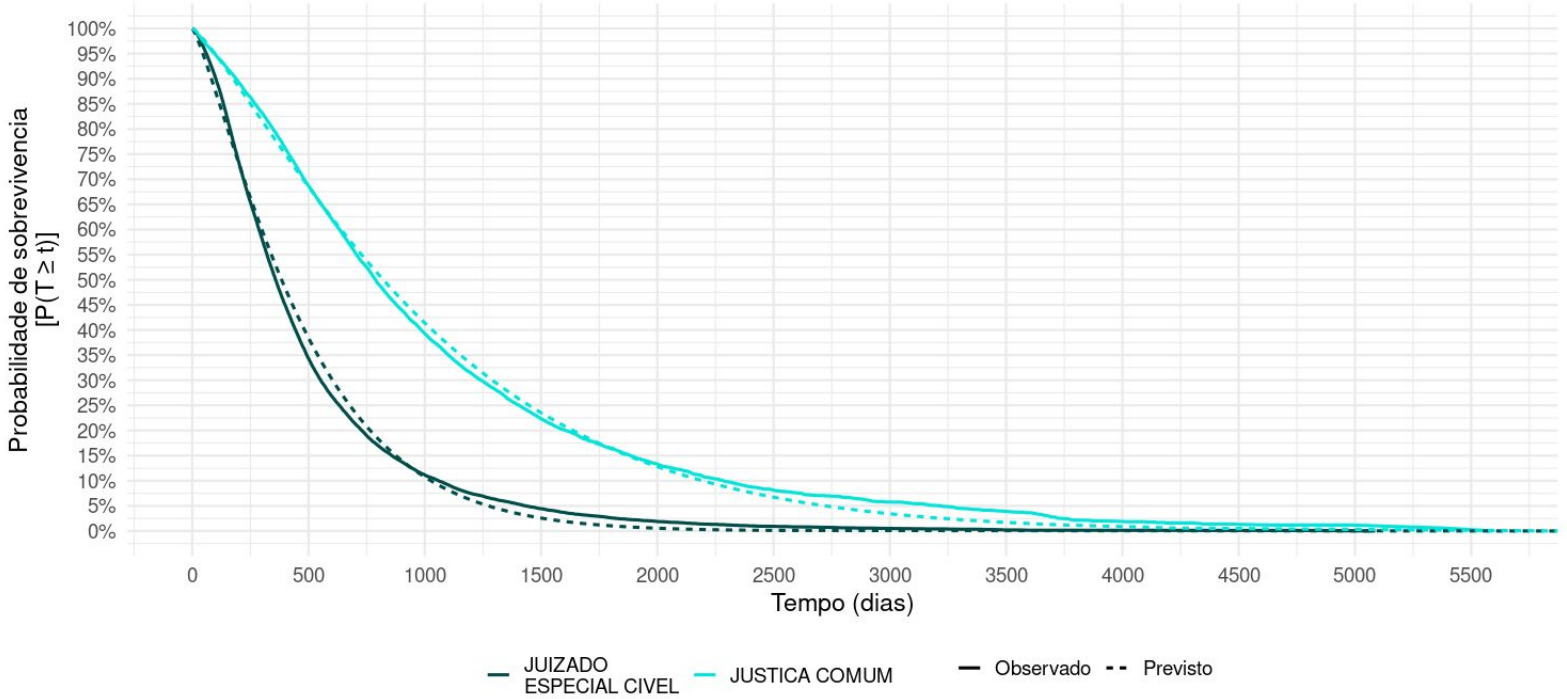
Passo 5: Análise Descritiva

Tempo: DF x PA



Passo 5: Análise Descritiva

Tempo: JEC x Justiça Comum



Passo 6: Análise Inferencial

Foram ajustados 3 modelos para a **Análise de Resultado.**
A ideia é escolher o modelo que apresentou melhor performance.

accuracy	precision	recall	AUC	modelo
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
0.739	0.766	0.761	0.808	modelo_random_forest
0.706	0.739	0.725	0.774	modelo_gradient_boost
0.697	0.736	0.706	0.771	modelo_logistico

Uma observação importante acerca de modelos estatísticos:
A acurácia nem sempre é a melhor medida de performance

Passo 6: Análise Inferencial

Importância das variáveis

variavel	importancia
aluno_saldo_dev_total	541.394746
processo_valor_causa	342.313647
ano_cnj	315.992196
aluno_dias_saldo_dev_max	314.130643
processo_objeto_OUTRO	283.727590
processo_objeto_HONORARIOS..INDENIZACAO.POR....	200.846248
processo_classificacao_Massivo	177.889927
processo_objeto_INDENIZACAO.POR.DANOS.MORAIS	152.198272
processo_objeto_INDENIZACAO.POR.DANOS.MORAIS...	134.917965
diff_aluno_ano_ingresso_min_max	126.558720
processo_objeto_OBRIGACAO.DE.FAZER	113.308627
aluno_n_cursos	107.954439
cpf_tem_processo_anterior_SIM	82.599017
processo_esfera_JUSTICA.COMUM	74.049896

Passo 6: Análise Inferencial

Conclusões do modelo de resultado:

O modelo de Random Forest desenvolvido para prever o resultado de um processo obteve um desempenho considerável em todas as métricas principais. Com uma precisão de 76.61%, ele indica uma boa capacidade de identificar verdadeiros positivos dentre todos os positivos previstos. Juntamente com um recall de 76.06%, fica evidente que o modelo também é bem-sucedido em capturar a grande maioria dos positivos reais. A acurácia de 73.89% mostra que o modelo faz previsões corretas em cerca de três quartos das vezes. Uma característica marcante é o valor da Área Sob a Curva (AUC) que é de 80.81%, indicando uma boa capacidade discriminativa do modelo em separar classes positivas de negativas.

	Prev	
Obs	DERROTA	VITORIA
DERROTA	3681	835
VITORIA	1304	2339

Passo 6: Análise Inferencial

Sobre a importância das variáveis:

As variáveis mais relevantes que afetam a previsão do modelo são:

aluno_saldo_dev_total com uma importância de 541.39: Esta variável é a mais crítica, sugerindo que o saldo total devido por um aluno tem uma influência significativa sobre o resultado do processo.
processo_objeto_OUTRO com uma importância de 283.73: A natureza do objeto do processo, quando categorizado como 'OUTRO', também tem uma relevância considerável na previsão.
ano_cnj e **processo_valor_causa** têm importâncias de 315.99 e 342.31, respectivamente, destacando a influência do ano do CNJ e do valor da causa do processo nas previsões.
processo_objeto_HONORARIOS..INDENIZACAO.POR.DANOS.MORAIS com uma importância de 200.85 sugere que processos relacionados a honorários e indenizações por danos morais são particularmente influentes.
processo_classificacao_Massivo com uma importância de 177.89 destaca que a classificação dos processos como 'Massivo' tem um papel crucial nas previsões.
Dentre outras variáveis notáveis, observamos que **cpf_tem_processo_anterior_SIM** e algumas variáveis relacionadas à esfera e tipo do processo também têm importâncias significativas.

Passo 6: Análise Inferencial

Vale ressaltar que a importância das variáveis em um modelo de Random Forest não indica causalidade direta, mas sim a influência dessas variáveis nas previsões do modelo. A presença de uma variável de alta importância sugere que mudanças em seus valores são propensas a influenciar a previsão, enquanto uma baixa importância sugere que a variável tem uma influência mínima na decisão do modelo, considerando os dados de treinamento.

Concluindo, o modelo apresenta um desempenho robusto, e as variáveis destacadas acima são críticas para a sua capacidade preditiva. Aprofundar a análise sobre essas variáveis e entender sua relação com o resultado do processo pode oferecer insights valiosos para futuras estratégias e tomadas de decisão.

Passo 6: Análise Inferencial

processo_id	previsao
1	0.7339932
2	0.5379220
3	0.5232110
4	0.7715883
5	0.7287297
6	0.5590175
7	0.5078800
8	0.7269206
9	0.9024332
10	0.5937252
11	0.7805429
12	0.5661360

Cases

Vamos analisar alguns cases da ABJ

https://abj.org.br/cases/