



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO

Painel em Python do Centro de Inteligência

Kallil de Araújo Bezerra

Natal-RN, Brasil
2020

Kallil de Araújo Bezerra

Painel em Python do Centro de Inteligência

Trabalho de Conclusão de Curso apresentado ao Programa de Residência em Tecnologia da Informação do Instituto Metrópole Digital da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do título de Especialista em Tecnologia da Informação. Área de Concentração:

Orientador: Elias Jacob

Natal-RN, Brasil
2020

Kallil de Araújo Bezerra

Painel em Python do Centro de Inteligência

Trabalho de Conclusão de Curso apresentado ao Programa de Residência em Tecnologia da Informação do Instituto Metr pole Digital da Universidade Federal do Rio Grande do Norte como requisito parcial para a obten  o do t tulo de Especialista em Tecnologia da Informa  o.  rea de Concentra  o:

Trabalho aprovado. Natal-RN, Brasil, 7 de outubro de 2020:

Elias Jacob
Orientador

Professor
Examinador

Professor
Examinador

Natal-RN, Brasil
2020

Dedico este trabalho ao meu irmão, Kaio, que apesar de ser mais novo que eu sempre me ensinou, e me ensina, bastante coisa. Ele é o melhor amigo que alguém pode ter, obrigado irmãozinho.

Agradecimentos

Agradeço a todo o time da JFRN que sempre proporcionou uma ótima estrutura e um ambiente de trabalho muito bom muito bom, sempre bem equipado e atendendo às demandas de forma rápida. Agradeço, também, ao time de infraestrutura que sempre foi extremamente solícito e também ao grande time de desenvolvedores que ajudaram a ressuscitar robôs robôs, tornando-os até mais rápidos. Devo agradecer ao time de BI também, que nesses 18 meses me ensinou muito e me mostrou diversas ferramentas que eu jamais teria conhecido e aprendido a usar se não estivesse cercado de gente competente. Por último, mas não menos importante, gostaria de agradecer aos professores Eduardo Aranha, Elias Jacob e Leonardo Bezerra pelas aulas, pelo tempo dedicado a me orientar de diversas formas e pelas reflexões que foram levantadas durante as várias conversas, me senti muito sortudo de ter conhecido esses trio que me inspirou bastante a estudar mais, me esforçar mais e ser melhor de maneira geral.

I have spoken. -Kuiil

Resumo

O objetivo deste trabalho é criar um painel de visualização de dados da Justiça Federal do Rio Grande do Norte (JFRN), nesse painel devem ser mostrados quais tipos de processos são mais frequentes na JFRN de acordo com as Varas e divididos por mês. Para criar esse painel a linguagem de programação Python foi escolhida, usando as bibliotecas Dash e Plotly. Os painéis feitos pela JFRN atualmente usam softwares pagos, então esse seria o primeiro painel usando ferramentas *open source* e gratuitas, abrindo um caminho para que novos painéis usem tecnologias semelhantes.

Palavras-chave: Processo Judicial. *Business Intelligence*. Visualização de dados.

Abstract

The objective of this work is to create a panel of data visualization of the Federal Justice of Rio Grande do Norte (JFRN), in that panel it must be shown which types of processes are more frequent in JFRN according to the courts and divided by month. To create this panel, the Python programming language was chosen, using the Dash and Plotly libraries. The dashboards made by JFRN currently use paid software, so this would be the first panel using free open source tools, paving the way for new dashboards to use similar technologies.

Keywords: Judicial process. Business Intelligence. Data visualization.

Lista de ilustrações

Figura 1 – Resumo de um sistema BI	14
Figura 2 – Preço do Power BI Premium fonte: < https://powerbi.microsoft.com/en-us/ >	17
Figura 3 – Qlikview versus Qlik Sense fonte: < https://www.qlik.com/us/products/ >	18
Figura 4 – Estrutura básica do painel	21
Figura 5 – Tabela com as diferentes frequências de Assuntos	24
Figura 6 – Comportamento da 6ª e 12ª Vara no ano de 2015	26
Figura 7 – Comportamento da 4ª e 10ª Vara no ano de 2016	26
Figura 8 – Comportamento da 8ª e 14ª Vara no ano de 2018	26
Figura 9 – Distribuição normal	27
Figura 10 – Primeira versão do painel	31
Figura 11 – Sétima versão do painel	32
Figura 12 – Décima quarta versão do painel	33
Figura 13 – Trigésima versão do painel	34

Lista de abreviaturas e siglas

IMD	Instituto Metrópole Digital
UFRN	Universidade Federal do Rio Grande do Norte
JFRN	Justiça Federal do Rio Grande do Norte
TI	Tecnologia da Informação
BI	<i>Business Intelligence</i>

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta

símbolos não usados e usados não listados

Sumário

	Introdução	12
1	BI NA JUSTIÇA FEDERAL DO RN	13
1.1	Estrutura básica do BI	13
1.1.1	Visualização de dados	15
1.1.2	Análise de dados	15
1.2	Custos do BI	16
1.2.1	Exemplos de ferramentas	17
2	CONSTRUÇÃO DO PAINEL	19
2.1	Tecnologias usadas	19
2.1.1	Justificando o Python	20
2.2	Estrutura básica do painel	21
2.2.1	Plotly e Dash	21
2.2.2	Pandas	22
2.2.3	Estrutura dos dados	22
2.2.4	Análise de anomalias	23
3	DETECÇÃO DE ANOMALIAS NO PAINEL	25
3.1	Distribuição dos dados	25
3.2	Detecção de anomalias	27
4	CONCLUSÃO	29
	REFERÊNCIAS	30
	ANEXO A – EVOLUÇÃO DO PAINEL	31

Introdução

A Tecnologia da Informação (TI) vem se tornando cada vez mais importante na administração em diversas áreas, as aplicações vão desde a infraestrutura que **busca conectar os diferentes setores, mantendo a segurança da rede**, até a automação de processos. Além disso, nesse espectro de aplicações da TI podemos incluir o melhoramento da gestão usando a computação, atualmente a quantidade de dados e variáveis disponíveis para o gestores é muito grande, e é extremamente difícil de se gerenciar essa massa de dados, para isso existem várias ferramentas que têm por objetivo auxiliar na visualização e futura tomada de decisão dos administradores.

Uma área da TI que tem crescido bastante é a *Business Intelligence* (BI), que reúne uma série de conceitos que podem ser aplicados em instituições, de qualquer tamanho e área, com o objetivo de dar suporte à tomada de decisão, trazendo dados e gerando informação, que serve de base para as escolhas das estratégias de um determinado negócio.

Sistemas de Business Intelligence (BI) combinam dados operacionais com ferramentas analíticas para apresentar informações complexas e competitivas para os tomadores de decisão (NEGASH, 2003). Mesmo antes desse conceito as ferramentas para análise de dados já existiam, dentre as quais podemos citar os bancos de dados, visualização e análise de dados e as diversas análises estatísticas que encontramos.

De acordo com um estudo feito por C. Willen (WILLEN, 2002), as estratégias do BI têm sido usadas para assistir as seguintes atividades:

- Gerenciamento da performance corporativa
- Otimizar relações com clientes, monitorar atividades dos negócios e suporte às decisões tradicionais
- Uso de ferramentas BI para operações e/ou estratégias específicas
- Criação de relatórios com métricas dos negócios

Nessa lista fica claro que o BI é muito importante para entender o comportamento interno da empresa em que for aplicado, as conclusões são usadas para alocar ou realocar recursos para áreas mais importantes ou que estejam sob alta demanda.

1 BI na Justiça Federal do RN

- o uso principal tem sido voltado à extração de dados dos sistemas judiciais
- qual banco de dados?
- endereço do portal BI do TRF5

Na Justiça Federal do Rio Grande do Norte (JFRN) o BI é importante para entender o que está acontecendo nas diferentes Varas. O software usado é o Qlikview, que é capaz de montar gráficos e consultas a partir do banco de dados fornecido pelo Tribunal Regional Federal da 5ª região (TRF5), a distribuição dos painéis, gerados pelo Qlikview, é feita pelo portal BI do TRF5.

Nesse portal existem diferentes painéis, que atendem demandas distintas, sendo bastante usado por servidores e magistrados. Porém, o processo de desenvolvimento nesses painéis apresenta alguns gargalos, pois antes de desenvolver é necessário que hajam documentos que autorizem a criação de um novo painel, e após o desenvolvimento é necessário um outro documento autorizando a publicação desse painel. Durante o desenvolvimento também há algumas demandas que dependem do time de Tecnologia da Informação do Tribunal Regional Federal da 5ª Região (TRF5), essas demandas vão desde disponibilização de dados até renovação de licenças expiradas, essa dependência também gera alguns atrasos para o time local da JFRN. Cada desenvolvedor deve ter uma licença válida, que é paga, isso acaba aumentando o custo do desenvolvimento.

1.1 Estrutura básica do BI

Como foi apresentado anteriormente, o BI serve para auxiliar nas tomadas de decisão, e isso é alcançado usando os dados da instituição em que estiver sendo empregado. Os dados são armazenados em *Data Warehouses*(DW), que são armazéns de dados, em tradução livre, esses armazéns guardam dados históricos, então a partir deles é possível analisar o desenvolvimento de variáveis importantes e o comportamento delas de acordo com períodos que o usuário desejar, e assim identificar padrões, essa análise, que é simples, já poderia ser usada para prever possíveis mudanças na estratégia de negócios (NEGASH, 2003).

Após o armazém, os dados devem ser coletados e limpos, a limpeza corresponde a remoção de linhas erradas, que contenham dados errados ou faltosos, que podem atrapalhar na análise e apresentação ao gestor. Em seguida, os dados são apresentados à pessoa do negócio, que a partir das suas análises irá tomar alguma decisão que afeta a estratégia da empresa.

Depois de carregar os dados do DW, os dados devem ser limpos, essa limpeza de dados corresponde a remoção de dados que sejam ou errados ou inexistentes, que podem causar problemas na análise de dados e na visualização que será exibida para o gestor. Imagine que num conjunto de dados sobre alturas de pessoas existe alguém com 15,5m de altura, esse dado é errado e pode causar algum distúrbio na análise, alterando a média da

altura dessa amostra, por exemplo.

Após a limpeza desses dados, eles podem ser lidos pelos painéis, que geram as visualizações para os gestores, que por sua vez tomarão as decisões baseados em dados corretos.

De forma resumida, o BI usa o DW para guardar os dados, usa um conjunto de ferramentas e técnicas para limpar e extrair os dados, essa técnica também é conhecida como *Extraction, Transform, Load* (ETL), e ,finalmente, apresenta gráficos que mostram o comportamento de variáveis de interesse da instituição para o gestor, que a partir disso escolhe alguma estratégia para os rumos da empresa/órgão/setor que gerencia.

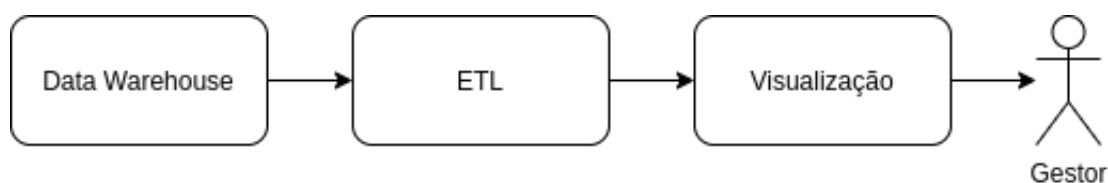


Figura 1 – Resumo de um sistema BI

Isso que foi tratado acima corresponde às etapas do processamento de dados estruturados, ou seja, dados que podem ser organizados e categorizados em linhas e colunas, e que, muitas vezes, possuem relações entre si. O processo para "manusear" dados não estruturados é um pouco diferente porque eles não são tão bem organizados, e alguns passos precisam ser inseridos nesse caminho para que eles sejam apresentados e tratados da melhor forma, evitando distorções.

É possível ver que a **Inteligência de Negócios** é formada por várias áreas diferentes dentro da Tecnologia da Informação, a seguir estão listadas algumas:

- Armazenamento na forma de *Data Warehouse*
- Visualização de dados
- Mineração de dados
- Processamento analítico na forma de *Online Analytic Processing* (OLAP)
- Gerenciamento do conhecimento
- Probabilidade
- Estatística
- Análises preditivas
- Detecção de anomalias

No decorrer do trabalho foram utilizados de forma mais frequente a análise de dados, a visualização e a **detecção de anomalias**.

faltou descrever a detecção de anomalias??

1.1.1 Visualização de dados

A visualização de dados é a representação gráfica dos dados, ela engloba a produção de imagens que comunicam o comportamento e a relação dos dados analisados. Existem diferentes formas de exibição:

- Gráficos de barras
- Histogramas
- Gráficos de linhas
- Tabelas
- Gráficos de dispersão

Na visualização de dados é importante transmitir o comportamento dos dados. Se um número é o dobro de outro, isso deve ser refletido na visualização, que não deve distorcer ou enganar. Ao mesmo tempo, a visualização deve ser fácil de entender. Boas representações visuais podem melhorar a mensagem da visualização. Se uma figura contém cores que não combinam, elementos visuais desequilibrados ou outras características que distraem o usuário, então será difícil de se interpretar corretamente os dados (WILKE, 2019). Para que os dados sejam lidos da melhor forma, também é importante que o usuário saiba realizar interpretações, e entender sobre o que é mostrado. Por exemplo, o usuário precisa ter noções de **correlação e causalidade**.

1.1.2 Análise de dados

A análise de dados engloba vários processos distintos, a limpeza, transformação e modelagem dos dados fazem parte desse procedimento que pode apresentar informações relevantes onde estiver sendo aplicado.

Para fazer uma boa análise é necessário conhecer os dados, então estudar sobre o que eles medem e sobre a realidade em que eles são aplicados, no caso da JFRN é importante entender quais são as competências das Varas. As Varas da JFRN possuem especialidades diferentes, tendo algumas competência para análise de processos Cíveis, Criminais e, ainda, os Juizados Especiais para processamento de demandas com menor complexidade. Então, como existem esses diferentes tipos de Vara, imagine que em uma delas, de competência Cível, existe uma alta frequência de Processos que pertenceriam a alguma Vara Criminal, a partir disso podemos levantar algumas hipóteses:

1. Os dados estão errados;
2. A análise dos dados está errada;

3. As pessoas desconhecem a competência da Vara e por isso cadastram processos que não deveriam ser dessa Vara.

Normalmente a análise errada é responsável por alguns problemas na visualização, e cabe ao analista de dados investigar as possíveis anomalias, e descobrir se realmente houveram erros na análise ou se aconteceu algum evento diferente na Vara, por exemplo. Em qualquer hipótese fica claro que a coordenação e comunicação entre diferentes áreas, TI e servidores da justiça, é essencial.

Também é importante notar que apesar das diversas funções que descobrem anomalias e casos extremos, ainda é importante que o analista, humano, conheça os dados e esteja sempre atualizado tanto em relação à tecnologia que será usada como também ao ramo em que estiver atuando.

1.2 Custos do BI

A Tecnologia da Informação (TI) é fundamental para o gerenciamento e manutenção correta dos negócios atualmente, ela deve ser vista como um ativo da instituição em que for aplicada, que merece investimento em *hardware*, *software* e pessoal capacitado, além dos treinamentos que devem ser dados à medida que a TI expande (NEGASH, 2003).

- **Hardware** - Os custos relacionados ao *hardware* variam de acordo com a estrutura que a instituição já tiver em mãos, se um *data warehouse* já existe, então a expansão precisa ser feita para um *data mart*, que é uma parte dedicada aos sistemas BI. Dependendo da estrutura pode ser necessária a expansão para um sistema de redes mais robusto, que suporte o tráfego dos dados.
- **Software** - Os *softwares* BI possuem um custo elevado, mas também possuem muitas funcionalidades, alguns deles como o Power BI tem assinaturas a partir de **\$60.000** por ano, como pode ser visto no site da [Microsoft](#), na cotação atual isso passa de R\$300.000, então pesquisar entre os principais fornecedores do mercado é crucial para ter um preço justo e que atenda às necessidades do cliente.
- **Implementação** - Essa categoria é uma extensão da anterior porque está diretamente relacionada. Após a aquisição da ferramenta BI e do *hardware* é necessário, também, realizar um treinamento do pessoal. Esse tipo de gasto também é recorrente, ou seja, sempre vai existir porque à medida que novas pessoas chegam e que o sistema expande, novos treinamentos devem ser realizados. Estima-se que esse tipo de manutenção corresponde a 15% dos custos.
- **Pessoal** - O custo com pessoal envolve tanto quem realmente vai trabalhar com BI como envolve quem vai dar suporte às atividades de BI. Por exemplo, o time de

na citação da tabela de preços da Microsoft, ou referências do tipo, sugiro colocar como nota de rodapé, pra ficar legível-acessível

infraestrutura deve estar preparado com as tecnologias de engenharia de dados, e manutenção do servidor que armazena os dados do BI.

1.2.1 Exemplos de ferramentas

Existem muitas ferramentas BI no mercado hoje, com o crescimento da área também houve a maior oferta de ferramentas que fazem painéis mais rápidos ou de forma mais simples, exigindo menos treino ou menos infraestrutura. Alguns serviços também oferecem a computação em nuvem, e "alugam" o poder computacional de acordo com o uso do cliente. Com a grande variedade de produtos também vem uma grande diferença de preços. Podemos citar, novamente, o Power BI, um dos líderes nesse tipo de atividade, custando (em 2020) \$4.995 por mês para o serviço *Premium*, e esse preço pode aumentar de acordo com os serviços extra que o cliente desejar incluir no pacote.

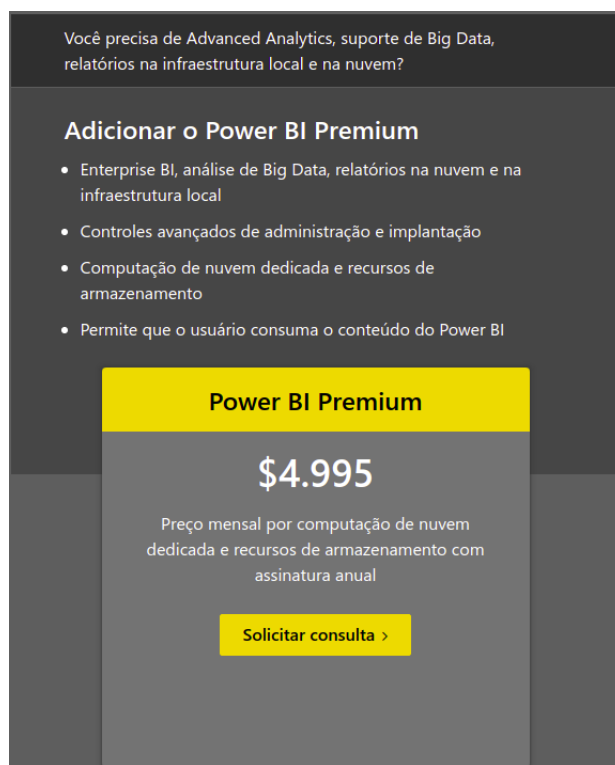


Figura 2 – Preço do Power BI Premium
fonte: <<https://powerbi.microsoft.com/en-us/>>

Além do Power BI podemos citar o Qlikview, que é usado na JFRN. Ele é desenvolvido pela Qlik, que ultimamente vem focando os esforços e investimentos no Qlik Sense, que oferece **várias vantagens**. O Qlikview foi o principal produto da Qlik durante muito tempo, a primeira versão dele data de 1994, e recebeu várias melhorias ao longo dos anos, se adequando às **novas tecnologias e incorporando funções diferentes**. Atualmente a Qlik tenta levar os clientes do QlikView para o Qlik Sense, que fornece painéis em plataformas móveis (*smartphones* por exemplo), integração com APIs e construção de visualização

quais vantagens? que novas tecnologias e funções foram incorporadas?

simplificada, através de recursos como clique-e-arraste. Os preços variam de 40 a 70 dólares, por mês por licença.

Feature & Function	Qlik Sense	QlikView
Freeform Associative Exploration	✓	✓
Augmented Intelligence	✓	
Dashboarding/Guided Analytics	✓	✓
Governed Self-service Analytics	✓	
Visual Data Prep	✓	
Advanced Data Prep	✓	✓
Broad Data Connectivity	✓	✓
Modern Platform Built on Open APIs	✓	
SaaS/Multi-Cloud	✓	
Offline Mobile	✓	

Figura 3 – Qlikview versus Qlik Sense
fonte: <<https://www.qlik.com/us/products/>>

**existem soluções gratuitas boas?
porque não entraram no levantamento de possibilidades?**

2 Construção do painel

Antes de avançar para a parte técnica, é importante explicar o que é o Centro de Inteligência da JFRN e como um painel poderia ajudar na tarefa deles. De acordo com o site do Centro de Inteligência, "O Centro de Inteligência da Justiça Federal do Rio Grande do Norte, criado pela Portaria nº 205/2017 – DF, em observância à Portaria nº 369/2017 – CJF, tem o objetivo de criar meios administrativos para prevenir demandas repetitivas, bem como de agilizar a sua tramitação processual, através do debate entre os seus componentes e os demais atores do sistema de justiça."

Esse tipo de demanda deve ser comunicado às autoridades para que a ação sobre esses processos repetitivos seja rápida, e não haja uma sobrecarga. Portanto, o painel será usado para auxiliar na análise dessas demandas, tentando acompanhar a evolução e desenvolvimento delas. Então, um painel que mostre os tipos de Assuntos mais recorrentes em cada Vara pode auxiliar o Centro de Inteligência a se preparar e comunicar embasado nos dados.

o uso de ferramentas abertas e gratuitas foi desconsiderado pq?

elas poderiam ser úteis para dar autonomia aos analistas de dados de negócio

2.1 Tecnologias usadas

Uma das tarefas que fez parte do desenvolvimento do painel foi a pesquisa e escolha da ferramenta que poderia gerar a visualização, de forma rápida e com facilidade de ser distribuída pela infraestrutura de TI da JFRN. Como foi mostrado anteriormente, as ferramentas pagas custam caro, e a estrutura de desenvolvimento de painéis do TRF5 usa QlikView, que além de demandar uma licença para desenvolvimento, também precisa de alguns documentos para a publicação do painel. Então, algumas ferramentas gratuitas foram consideradas, e nesse processo de escolha foram analisados alguns pontos:

- Pago ou gratuito: uma ferramenta paga pode exigir um custo alto para implementar e manter, então uma opção gratuita deve ser favorecida;
- Pronto ou próprio: existem softwares BI que já têm várias funções prontas, mas, nos casos do QlikView, PowerBI, Tableau, **existem limitações que impedem a quantidade de linhas que serão lidas do conjunto de dados**, por exemplo. Então, a criação das próprias análises e construção de gráficos foi o caminho escolhido.

Python é uma linguagem de programação de alto nível e de aplicações gerais, portanto, nada tem a ver como uma ferramenta pronta de BI, não tem integração automática de dados, nem criação simples de gráficos e visualizações, porém é completamente gratuita e tem um ótimo suporte da própria comunidade de usuários. De acordo com o *TIOBE index*, que é um índice de popularidade de linguagens de programação, a linguagem Python

é a terceira mais popular, ficando atrás de Java e C, a lista completa pode ser acessada na página do [TIOBE](#).

Linguagem	Popularidade
C	16.95%
Java	12.56%
Python	11.28%
C++	6.94%
C#	4.16%
Visual Basic	3.97%
JavaScript	2.14%
PHP	2.09%
R	1.99%
SQL	1.57%

Tabela 1 – Linguagens de programação mais populares

Para desenvolver o trabalho as bibliotecas Pandas, Plotly e Dash foram usadas, elas servem para analisar dados, gerar visualizações e montar o painel, respectivamente. Além do Python para construir o painel em si, foi usado o QlikView para extrair os dados e gerar um arquivo que pudesse ser lido pelo Pandas. Portanto, de forma resumida temos:

- Python

- Dash

versões da bibliotecas? do python?

- Plotly

- Pandas

- QlikView

2.1.1 Justificando o Python

Construir um painel usando Python pode apresentar alguns desafios, os desenvolvedores precisam ter noções de visualização de dados, análise de dados e programação com a linguagem. Mas, como foi dito anteriormente, a linguagem Python é bastante popular e não é difícil encontrar profissionais com esse perfil.

Outro ponto positivo de se usar Python é a replicabilidade, é possível criar painéis que atendam diversas Varas da JFRN, usando dados diferentes e gerando visualizações focadas na Vara. Essa forma de se desenvolver painéis mais simples não precisa ficar restrita ao Centro de Inteligência, ela pode expandir para atender necessidades mais simples, que não precisem do QlikView, de forma mais rápida mas atendendo às necessidades do gestor, levando em conta as características locais dos dados de onde for aplicado.

É claro que esse tipo de expansão da TI deve ser acompanhada de um time maior de profissionais, com diferentes habilidades e competências, treinamentos relacionados

mais uma justificativa: CNJ vem padronizando as soluções de IA com Python, vide SINAPSES

a Python, visualização de dados, análises de dados etc. Mas os impactos disso seriam bons, os gestores teriam melhor controle sobre seus ambientes de trabalho, com novas visualizações e dados para basear novas estratégias por exemplo.

2.2 Estrutura básica do painel

Os dados usados são de uma análise feita para o Ministério Público Federal (MPF), essa análise já apresentava a divisão dos processos por Assunto e Vara, distribuídos ao longo dos meses. A partir dos dados, as visualizações começaram a ser feitas, usando Plotly, para tentar enxergar algum comportamento que pudesse ser útil aos gestores. Após isso, as visualizações foram embarcadas no Dash, e a estrutura mínima do painel ficou da seguinte forma:



Os dados não podem ser lidos direto do DW?

fato e dimensão são estratégias não citadas anteriormente (do Qlikview)

faltou explicar??

parágrafo longo....

Figura 4 – Estrutura básica do painel

A leitura e mineração dos dados é feita pelo Pandas, e a apresentação pelo Dash. O painel que se propôs ao Centro de Inteligência não tem a estrutura clássica com diferentes tabelas (**fato e dimensão**), com as quais se geram as visualizações, no lugar disso, existe um **arquivo .csv** que contém os dados que serão usados, esses dados .csv fazem parte uma extração que veio do Processo Judicial eletrônico (PJe), e a partir desse arquivo o painel vai criar subconjuntos de acordo com o ano e órgão julgador escolhidos pelo usuário. Portanto, esse painel se aproxima mais de um visualizador de dados do que de um painel BI. Nele é possível selecionar duas variáveis: o Órgão Julgador e o Ano. A partir dessas escolhas o sistema vai fatiar os dados recebidos e mostrará algumas análises. Essas análises são mostradas em forma de tabelas condicionais, que mudam as cores das células de acordo com a frequência de aparição dos Assuntos.

2.2.1 Plotly e Dash

A Plotly é uma empresa canadense que desenvolve ferramentas para análise e visualização de dados. Os serviços essenciais são gratuitos, basta carregar a biblioteca no programa e começar a usar, isso vale para o *plotly graph objects* por exemplo, que gera gráficos interativos, e também vale para o Dash, que é um dos seus principais produtos.

Plotly além de ser o nome da empresa, também é o nome da ferramenta de visualização de dados. Ela foi usada nas primeiras versões do painel, mas como a visualização passou a se concentrar nas tabelas, acabou saindo da versão atual.

A biblioteca Dash é um *framework* usado para construir aplicações web que apresentem um visual simples de se configurar e que sirva para análises de dados, não é necessário (porém ajuda bastante) conhecer *html* ou outras tecnologias de *front-end* para montar um painel. O resultado pode ser distribuído pela internet, usando serviços como o *Heroku*, de forma gratuita.

**inserir versões das bibliotecas usadas e referencias
sugestão: usar footnote ou compilar tudo num trecho do TCC**

2.2.2 Pandas

O Pandas é essencial na execução do painel, ele carrega as ferramentas necessárias para a manipulação dos dados, como a seleção correta do Órgão Julgador escolhido, e o ano a ser visualizado. Além disso, ele também é responsável por montar os *dataframes*, que são estruturas de dados, que servem de base para as tabelas e as avaliações por cores que é mostrada na visualização final.

2.2.3 Estrutura dos dados

integrar consulta ao banco de dados!!!!

Nessa primeira versão do painel os dados virão de um arquivo *.csv* gerado a partir do Qlikview, que é o software de BI padrão da JF. Esse arquivo carrega várias colunas, entre elas podemos citar número do processo, status, classe judicial, documento da parte, data do trânsito em julgado. Porém, para fazer a análise dos dados serão usadas as seguintes colunas:

- Órgão Julgador - os órgãos julgadores são as Varas da JFRN que ficam espalhadas pelo Estado, o usuário precisa selecionar um desses órgãos para visualizar os dados.
- Data Primeira Distribuição - essa é a data em que o processo chega na JFRN, mesmo que caia numa Vara que não seja da competência dele essa data é importante para analisar que Vara o recebeu e quando ele chegou na JFRN.
- Assunto - é o tema do processo, existem diferentes categorias em que um processo pode ser categorizado, e a partir desse campo é possível contar quantos processos de cada tipo deram entrada na JFRN.
- Assunto Código - diferentes assuntos possuem diferentes códigos, e a contagem dos processos se dá usando esse campo, que agrupa os códigos que são iguais e conta o total para saber quantos deram entrada na JFRN.

A partir da escolha do Ano e Órgão Julgador, o painel irá fazer as análises e seleções relevantes, populando a tabela e mostrando ao usuário quais são os processos mais frequentes de cada mês, no Ano e Vara escolhidos.

2.2.4 Análise de anomalias

A detecção de anomalias é, uma conjunto de técnica que servem para identificar comportamentos que fogem do que é esperado. Um dos desafios do trabalho foi encontrar uma forma de se detectar os Assuntos que possuísem alta frequência de entrada na JFRN, porque, teoricamente, cada Ano e cada Vara possuem diferentes distribuições de Assuntos, e um modelo de detecção de anomalia que se encaixa bem em um determinado período, pode não se encaixar em outros. São 15 órgãos julgadores diferentes, e os anos que podem ser consultados são de 2014 até 2020, então são 90 distribuições diferentes. Apesar disso, usamos uma abordagem simples mas eficaz.

Primeiro, há uma análise da média (\bar{x}) de Assuntos que entraram na Vara, essa análise leva em conta o ano selecionado e o ano anterior, após isso, o desvio padrão (σ) é calculado e novas variáveis são geradas.

As variáveis são:

- $anom_2$ definida como:

$$anom_2 = media_{assuntos} + (2 * \sigma)$$

- $anom_1$ definida como:

$$anom_1 = media_{assuntos} + \sigma$$

- $media_{assuntos}$ que é a média simples dos assuntos, a cada dois anos:

$$media_{assuntos} = \sum_{ano}^{ano-1} \frac{assuntos}{total_{meses}}$$

Com essas variáveis encontradas, a distribuição das cores segue as regras a seguir, em que *total* significa a quantidade total de Assuntos de determinada categoria:

$$F_{cores} = \begin{cases} Vermelho & \text{se } total \geq anom_2 \\ Amarelo & \text{se } total \geq anom_1 \text{ e } total < anom_2 \\ Verde & \text{se } total \geq media_{assuntos} \text{ e } total < anom_1 \end{cases} \quad (2.1)$$

Dessa forma é possível ver quais são os Assuntos que estão entrando com alta frequência, essa visualização deve ser usada para justificar uma possível análise, feita pelo gestor, para entender se essa frequência é realmente uma anomalia, ou se isso era esperado.

Ao longo do tempo o painel sofreu diversas mudanças. Essas mudanças foram incrementais e uma das principais fontes de exemplos e usos das ferramentas do Dash foi a plataforma Medium, que apresenta vários artigos exemplificando formas de se usar o Dash e como usar melhor os recursos da biblioteca. Um desses artigos do Medium foi muito importante para a definição de uma estrutura base de desenvolvimento do painel, o texto de Ishan Mehta (MEHTA, 2020) apresenta uma proposta de estrutura que pode

ser replicada e melhorada em trabalhos futuros, e a partir dessa estrutura o painel foi montado e desenvolvido, com novas visualizações e diferentes análises.

Na figura abaixo é possível ver um exemplo da aplicação das fórmulas no painel.


	Julho: 07_total
Dívida Ativa	482
Competência da Justiça Federal	301
Cartão de Crédito	186
Profissional	97
Lei de Imprensa	87
Efeito Suspensivo / Impugnação / Embargos à Execução	46

Figura 5 – Tabela com as diferentes frequências de Assuntos

como verificar tendência nos dados? incluir gráfico de tendência histórica por assunto/vara?
não foi verificado que avaliar a tendência seria útil para a análise?

3 Detecção de anomalias no painel

Em alguns casos os dados podem apresentar registros que não parecem pertencer ao resto do conjunto, esses registros desviam muito do resto das observações e podem ser um problema na análise. Se o analista de dados conhecer o negócio, ele pode conseguir explicar a origem desses desvios. Por exemplo, imagine que uma loja registra um total de 10 vendas de produtos diariamente, e em uma determinada semana, sem explicação aparente, vendeu 1000 (diariamente também), após isso as vendas voltam para a casa dos 10 por dia. Claramente essa semana diferente deveria ser analisada para se entender as razões desse salto, mas se esse mesmo comportamento é registrado durante a *black friday*, ele pode ser, justamente, o esperado para aquele período. O *outlier*, como também são conhecidas essas anomalias, pode ter sido originada em algum *bug* do sistema e isso precisa investigado com as outras áreas do negócio, além da TI.

Por causa dessas nuances a detecção de anomalias deve ser tratada com cuidado, porque analisando apenas os dados por si só não garante que os valores que se distanciam do normal são realmente anômalos. Na prática, é esperado que os dados sejam analisados apenas pelo analista, mas é importante que haja uma integração entre as diferentes áreas do negócio, no caso da JFRN, Varas e equipe de TI. 

3.1 Distribuição dos dados

A própria natureza do das Varas na Justiça já gera uma possível frequência maior de determinados Assuntos, então é esperado que uma Vara Penal receba uma alta demanda de Assuntos da competência dela, o mesmo raciocínio pode ser aplicado na 6ª Vara que é de Execução Fiscal. Mas, também existem Assuntos que aparecem na primeira distribuição dos Processos mas que não são da competência da Vara em que foi criado, e isso, apesar de não acontecer com uma frequência alta, acontece de forma pulverizada, então existem muitos Assuntos que possuem 1, 2 ou até 3 ocorrências, e poucos Assuntos que acontecem em frequências muito altas. Esse comportamento pode ser notado abaixo:

Número alto de assuntos por processo indica necessariamente um outlier?

ou, um número baixo, de um tema incomum pode indicar essa ocorrência também?

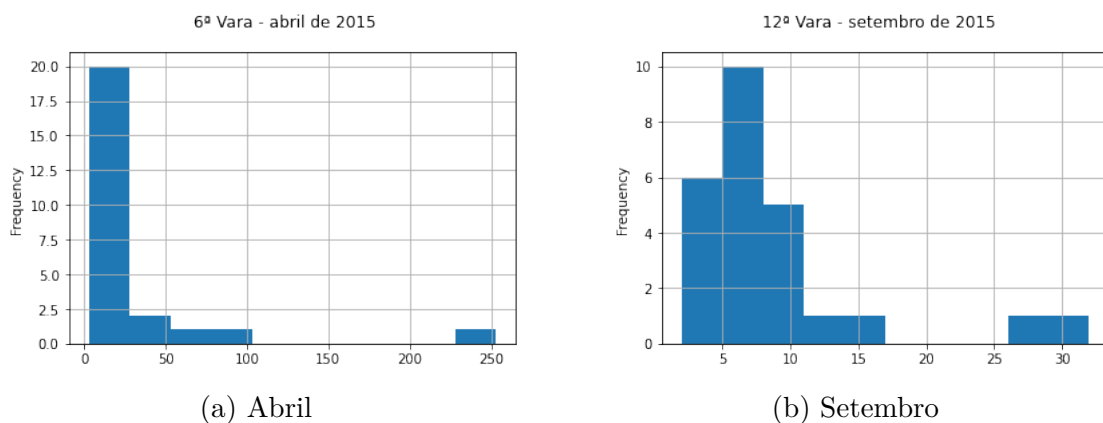


Figura 6 – Comportamento da 6ª e 12ª Vara no ano de 2015

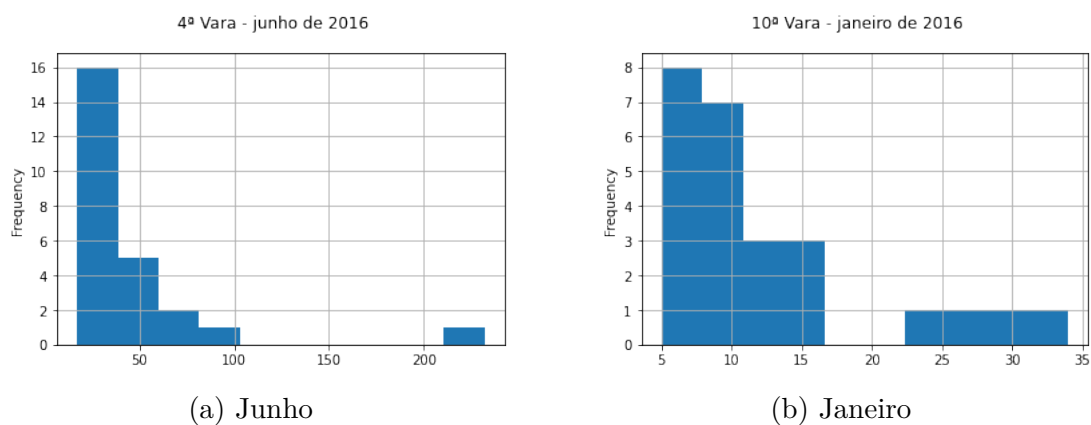


Figura 7 – Comportamento da 4ª e 10ª Vara no ano de 2016

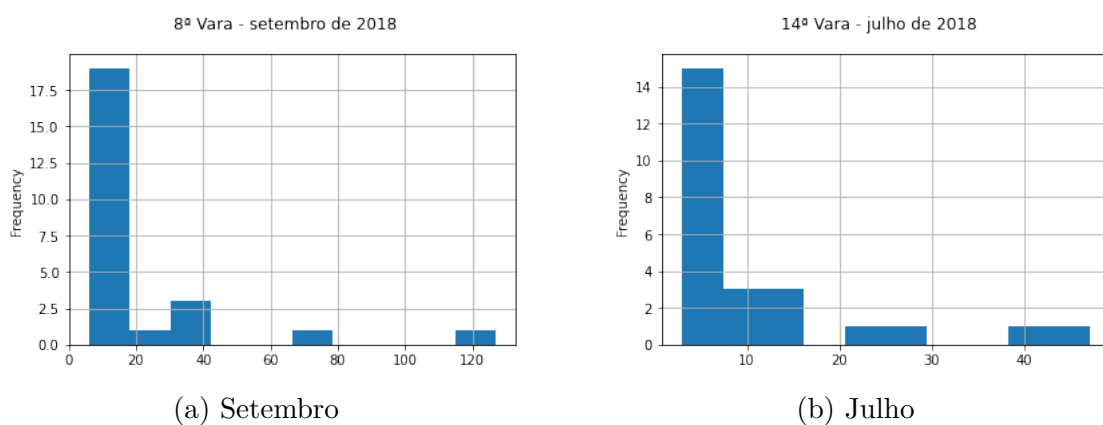


Figura 8 – Comportamento da 8ª e 14ª Vara no ano de 2018

No eixo horizontal são mostradas as frequências dos Assuntos, e no vertical a quantidade de Assuntos distintos, então essa concentração na esquerda indica que existem muitos processos com Assuntos diferentes mas em pequenas quantidades, e poucos (concentrados à direita) que possuem uma frequência alta, e são justamente esses, concentrados à direita, que pertencem à competência da Vara. Por isso que aparecem numa frequência tão superior aos outros, fazendo com que a distribuição desses processos não seja normal.

3.2 Detecção de anomalias

Como foi dito no início do capítulo, em alguns casos dados podem ser considerados diferentes demais para pertencerem a algum grupo e podem ser considerados *outliers*, e em muitos casos isso leva em consideração a distribuição normal.

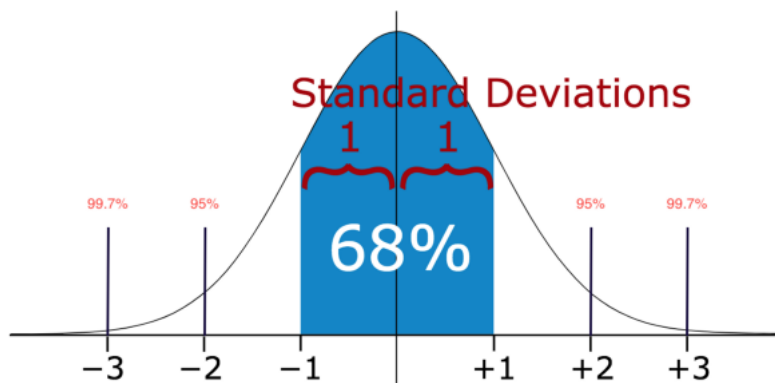


Figura 9 – Distribuição normal

Na distribuição normal 68% dos dados estão dentro de 1 desvio padrão, representado pela letra grega sigma (σ), e 95% estão dentro de 2, isso facilita a procura por *outliers* porque a partir disso é possível determinar, de forma pragmática, o que pode ser um dado errado. Após isso, o analista de dados pode tentar entender melhor o que originou essa anomalia e decidir se isso é errado de fato ou se é um fenômeno genuíno, que possui uma explicação. No caso do painel, o objetivo não é tentar remover esses dados ou fazer qualquer operação neles, mas sim mostrar esses dados para o gestor e ele decide o que será feito. Nesse ponto se torna importante a integração entre a gestão das Varas e o time de TI, porque a partir disso as duas partes conseguem entender melhor os dados que estão sendo exibidos, os possíveis erros e as causas dos *outliers*, que nem sempre são resultados errados, longe disso, os *outliers* aqui precisam ser mostrados para que sejam pesquisados e entendidos melhor.

A maioria das técnicas de detecção de *outliers* se aplicam para dados com distribuição normal, e se os dados não forem normalizados, alguns passos são acrescentados, como a transformação logarítmica, para forçar a normalização. Esses métodos não atendem à função do painel porque a distribuição apresentada não é normal, e algumas características das Varas são perdidas se forçarmos a normalização.

4 Conclusão

Existem diversas ferramentas de BI no mercado, mas elas podem ser muito caras e exigir uma mão de obra muito específica, pois possuem linguagens de desenvolvimento próprias. Além disso, a distribuição dos acessos é um pouco mais complexa porque exige uma certa burocracia para se permitir que painéis sejam publicados. Então, a biblioteca Dash é uma ótima alternativa à essas ferramentas pagas, pois além de ser gratuita, usa a linguagem de programação Python no desenvolvimento, e essa linguagem é muito difundida no mundo todo, e por ser bastante usada, é **relativamente fácil encontrar profissionais** que tenham boas noções de desenvolvimento nela.

Além disso, a possibilidade de se desenvolver painéis *in-house* e também fazer essa distribuição dentro da própria infraestrutura da JFRN permite que mais painéis sejam feitos, dando aos gestores uma melhor visão sobre as Varas e suas necessidades, assim, proporciona um melhor suporte às decisões e estratégias tomadas.

Existem dificuldades para se desenvolver esse tipo de painel, porém, após construir uma estrutura básica de desenvolvimento o trabalho fica muito mais fluido, e a preocupação passa a ser somente na colheita dos dados.

Por fim, também fica claro que a **integração dos gestores com a TI** é essencial para o bom aproveitamento das ferramentas de visualização de dados, pois o analista de dados precisa entender melhor o negócio que os dados representam e os gestores precisam saber interpretar os dados para que as decisões sejam bem embasadas.

requisitos importantes:

integração com o banco DW?
projeto no GIT?
projeto na infra da JFRN?

trabalhos futuros?

Referências

- MEHTA, I. *Python-Dash data visualization dashboard template*. 2020. Disponível em: <<https://medium.com/analytics-vidhya/python-dash-data-visualization-dashboard-template-6a5bff3c2b76>>. Acesso em: 20 ago. 2020. Citado na página 23.
- NEGASH, S. Business intelligence. *Communications of the Association for Information Systems*, v. 1, n. 1, p. 177–195, 2003. Citado 3 vezes nas páginas 12, 13 e 16.
- WILKE, C. O. *Fundamentals of Data Visualization*. [S.l.]: O'Reilly, 2019. Citado na página 15.
- WILLEN, C. Airborne opportunities. *Intelligente Enterprise*, v. 1, p. 11–12, 2002. Citado na página 12.

ANEXO A – Evolução do painel

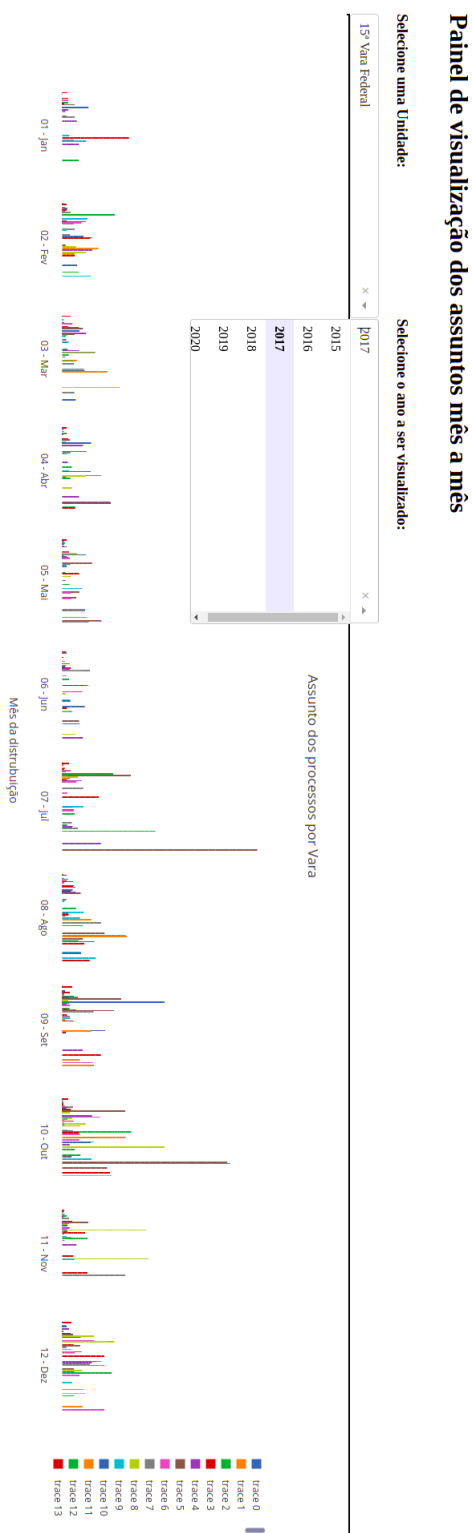
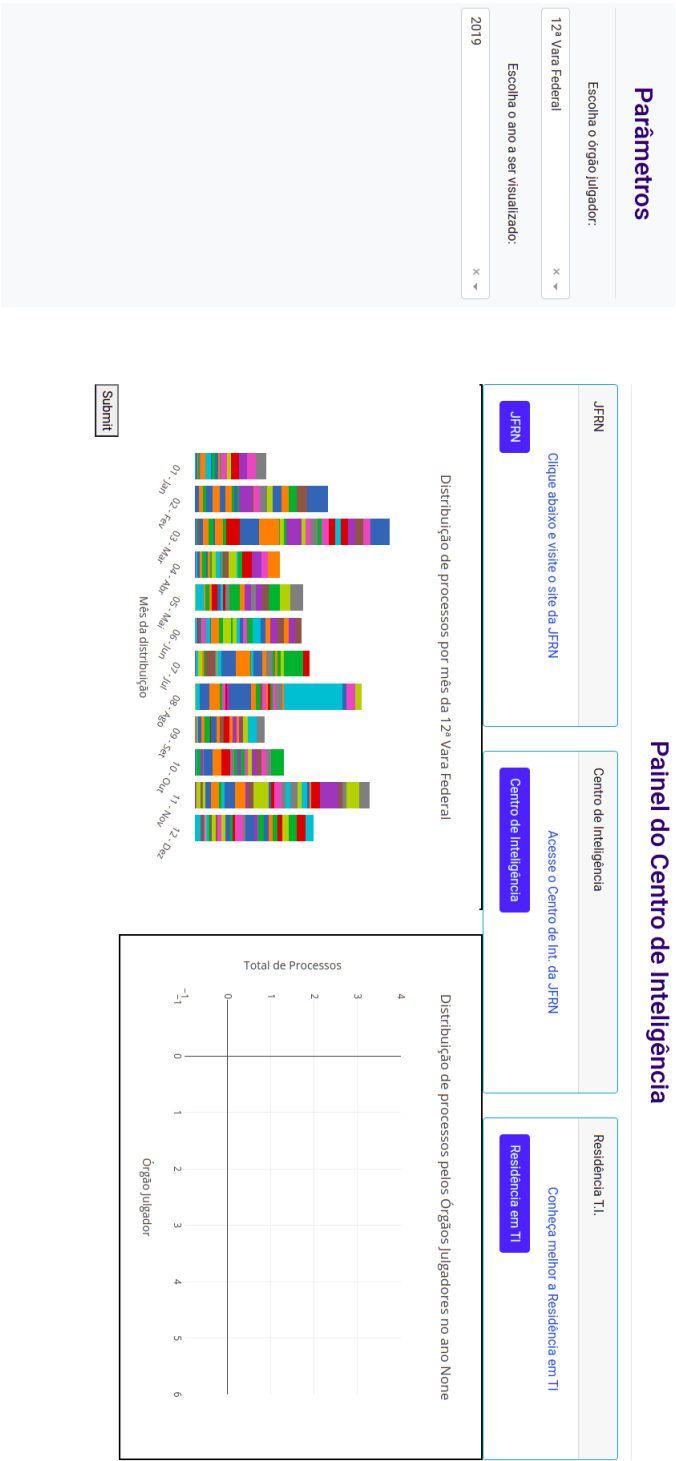


Figura 10 – Primeira versão do painel



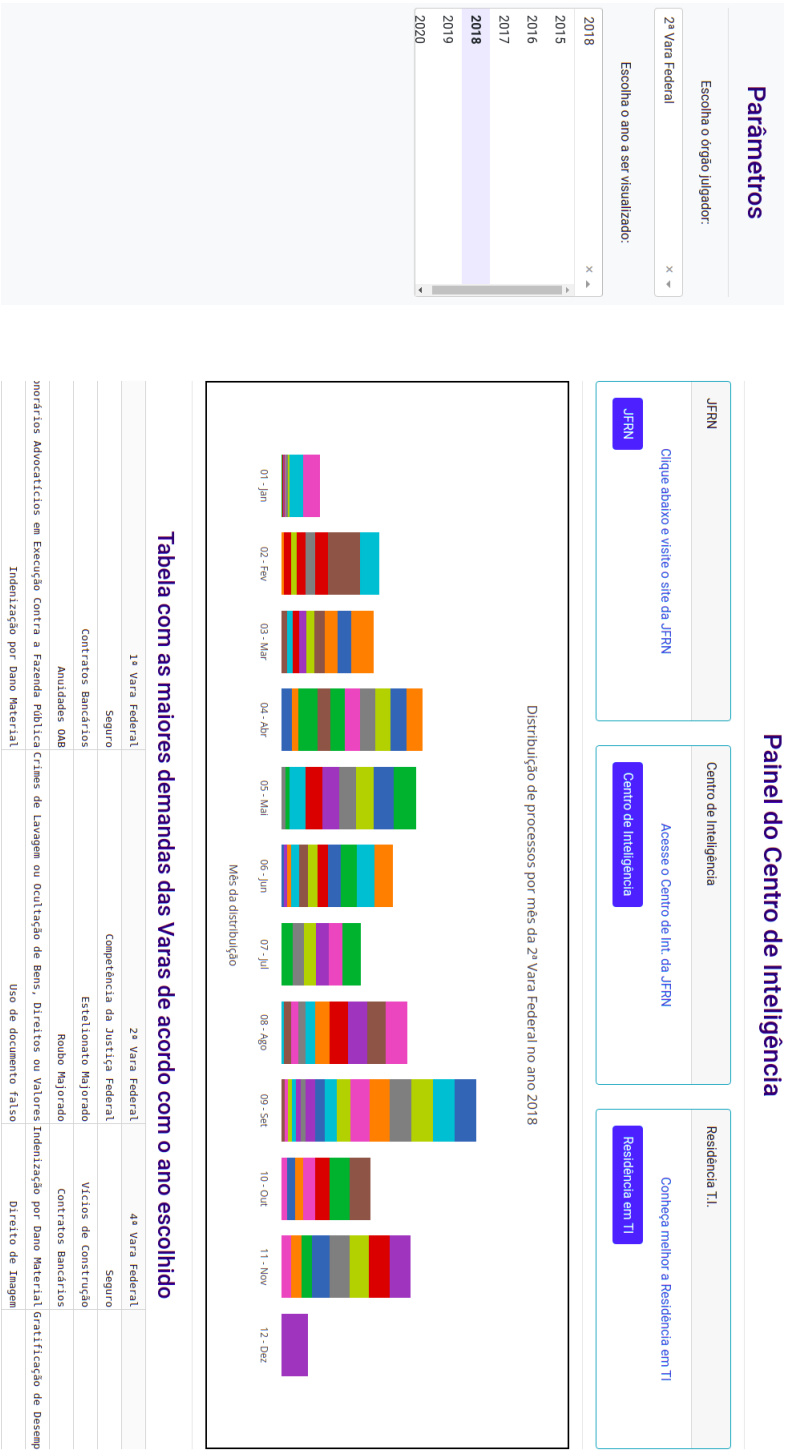


Figura 12 – Décima quarta versão do painel

