Pesquisa sobre Percepção em Visualizações

Projeto Final de Visualização de Dados: IA369W

Filipe Antonio de Barros Reis - RA 091202
Guilherme Augusto Câmara - RA 106694
Gabriel Lisbôa Guimarães Divino - RA 116932

Sobre

Esse projeto foi desenvolvido como parte da disciplina de visualização de dados (IA369W), lecionada pela professora Paula Dornhofer Paro Costa para pós-graduação na Faculdade de Engenharia Elétrica (FEEC) da Universidade Estadual de Campinas durante o primeiro semestre de 2017.

O projeto tem como objetivo conduzir um estudo prático sobre percepção humana aplicada à visualização de dados, possibilitando uma comparação dos resultados com referências estudadas na disciplina. Os resultados apresentados são preliminares, visto que uma pesquisa maior e mais controlada seria necessária para resultados definitivos, o que foge do escopo do projeto.

Motivação do projeto

A motivação desse projeto provém da necessidade de entender a dificuldade de pessoas de áreas diversas a encontrar as melhores formas de expressão para suas visualizações. O resultado dessa pesquisa pode servir como um guia de boas práticas atualizado para a confecção de visualizações.

Ademais, esse projeto também é motivado pela baixa disponibilidade resultados de pesquisas sobre formas mais efetivas para a visualização de dados, apesar da existência de guias de boas práticas, como em "Automating the Design of Graphical Presentations of Relational Information" por Jock D. Mackinlay e “Towards Visualization Recommendation – A Semi-Automated Domain-Specific Learning Approach” por Pawandeep Kaur. Os resultados devem confirmar as regras ou apontar para mudanças surgidas com o avanço tecnológico.

Por fim, existe a motivação de fazer uma contribuição inicial para a criação de uma ferramenta de geração de gráficos de forma automática.

Abordagem Proposta

A abordagem proposta é uma pesquisa da percepção humana sobre diferentes parâmetros de visualizações para um gráfico de dispersão. Foi utilizado um algoritmo visando mostrar para avaliação do usuário pares de visualizações. Nesses pares, apenas um parâmetro é variado por exposição, sendo os parâmetros considerados:

- Mapas de cores;
- Tamanho dos pontos;
- Formato dos pontos.

Por outro lado, foram utilizados três conjuntos de dados, criados a partir de crescimento linear, decrescimento linear e valores constantes, todos somados de um ruído aleatório.

Para determinar os parâmetros ótimos, utilizamos um sistema de pontuação que pondera quantas vezes um determinado parâmetro foi escolhido em relação ao total de vezes que foi mostrado.

Os participantes da pesquisa foram selecionados de forma a mantermos um grupo heterogêneo e sem especialistas, para evitar resultados enviesados pelas boas práticas atuais. Cada participante respondeu a um número constante e pré definido (20) de pares de visualizações, de modo a todos os participantes avaliarem ao menos 25% das possíveis visualizações.

Etapas do Desenvolvimento

O desenvolvimento do projeto deu-se de acordo com as seguintes etapas:

- Definição das Ferramentas utilizadas: optou-se por usar backend em python e html com javascript para o frontend. Para gerar as visualizações, foi definido utilizar matplotlib devido à sua grande versatilidade de parâmetros para as visualizações.
- Definição e criação dos conjuntos de dados a serem utilizados;
- Definição da quantidade de parâmetros: devido ao tempo restrito disponível para realizar a pesquisa, o número de opções para cada parâmetro precisou ser limitado, já que para termos significância estatística sem um grande número de participantes, o espaço amostral deveria ser reduzido.
- Seleção dos parâmetros a serem considerados:
    - Tamanho dos pontos: foram definidos três tamanhos de pontos, 100, 300 e 750 (unidades de tamanho do matplotlib), já que podem ser considerados tamanhos pequeno, médio e grande;
    - Formato dos pontos: Círculo, triângulo equilátero voltado para cima e triângulo equilátero voltado para baixo. O círculo foi escolhido por ser um marcador comum, enquanto que os triângulos foram escolhidos não só para avaliar um formato diferente mas para tentar perceber a influência da combinação deles com o formato dos dados, ou seja, ver se existe uma relação direta entre triângulo para cima e dados linearmente crescentes;
    - Mapa de Cores: utilizamos três mapas de cores contínuos, um com cores do verde para o azul, outro do amarelo para vermelho e um último com escalas de cinza. Com esses parâmetros podemos não só avaliar as opções de cores como também a preferência do uso de cores contra escalas de cinza. Por fim, pretendemos analisar se há alguma associação perceptiva entre dados crescentes e cor vermelha ou dados decrescentes e azul.

Figura 1 - Mapas de cores utilizados

- Disponibilização de uma versão piloto para análises iniciais e possibilitar o refinamento dos objetivos das visualizações e pesquisa;
- Realização da pesquisa, seguida da análise dos resultados.

Criação e Avaliação de Dados e Resultados

Os três conjuntos de dados utilizados nas visualizações foram criados seguindo uma relação básica (crescimento e decrescimento linear e constante) e adicionando ruído grande o bastante para que uma tendência pudesse ser vista mas que os dados ainda possuíssem certo realismo. Tais dados são gerados durante a primeira inicialização do projeto e se mantêm constantes ao longo da pesquisa;

Para a obtenção dos resultados, conforme cada gráfico era analisado, duas linhas com o resultado eram criadas, ambas contendo os parâmetros utilizados em cada visualização, além de um indicativo de escolha;

Por outro lado, a pontuação é definida como o número de vezes em que o gráfico foi escolhido sobre o número de vezes em que foi mostrado. Tal pontuação foi utilizada para evitar penalizações das visualizações não escolhidas, visto que a preferência de uma opção frente a outra não indica que a não escolhida é de fato ruim.

$$Pontuação = \frac{Escolhas}{Total\ de\ visualizações\ do\ parâmetro}$$

Por fim, a exploração e análise dos resultados foi feita considerando a pontuação de cada uma das opções de visualização e estatísticas provenientes para determinar não só os parâmetros mais votados a partir da análise individual de cada parâmetro sendo variado, como também a visualização com mais votos para cada tipo de dados, sem considerar qual parâmetro foi variado.

Metodologia e Ferramentas utilizadas

- Criação da interface (Frontend); A interface da pesquisa foi criada utilizando HTML 5 e Javascript e jqWidgets para a interfaces de interação e pensada de forma a ser o mais simples possível para evitar distorções provenientes de elementos distratores presentes na tela;
- Criação do servidor(Backend): O servidor foi responsável por gerar as visualizações aleatoriamente seguindo as regras já mencionadas e armazenando os resultados em um arquivo csv de log foi criado utilizando Python 2.7 e principalmente as bibliotecas:
    - WebPy: responsável pela interação frontend e gerenciamento das conexões, que poderiam ser simultâneas;
    - Simple JSON: responsável por gerenciar os arquivos JSON utilizados para a transmissão dos dados entre front e backend;
    - Numpy: biblioteca científica utilizada para manipulação numérica dos dados;

    - Matplotlib: responsável pela geração dos gráficos e utilizada devido à vasta gama de parâmetros das visualizações. Os gráficos foram gerados no formato PNG e transformados em um arquivo JSON utilizando a biblioteca cStringIO.;
    - Implementação e manutenção do projeto online utilizando p serviço de hospedagem/servidores Cloud9;
- Análise dos resultados: feita utilizando as bibliotecas numpy e pandas para a manipulação dos dados e matplotlib para a geração das visualizações.

Análise dos Resultados

De forma a inicializar a análise dos dados, o primeiro passo realizado foi analisar a variação dos parâmetros do gráfico para o conjunto das três distribuições de dados (linear crescente, linear decrescente e constante).

Com isso, os resultados apresentados a seguir estão separados nas categorias de Formato dos Marcadores, Tamanho dos Marcadores, Mapa de Cores e Conjunto mais Votado, logo abaixo.

Formato dos Marcadores

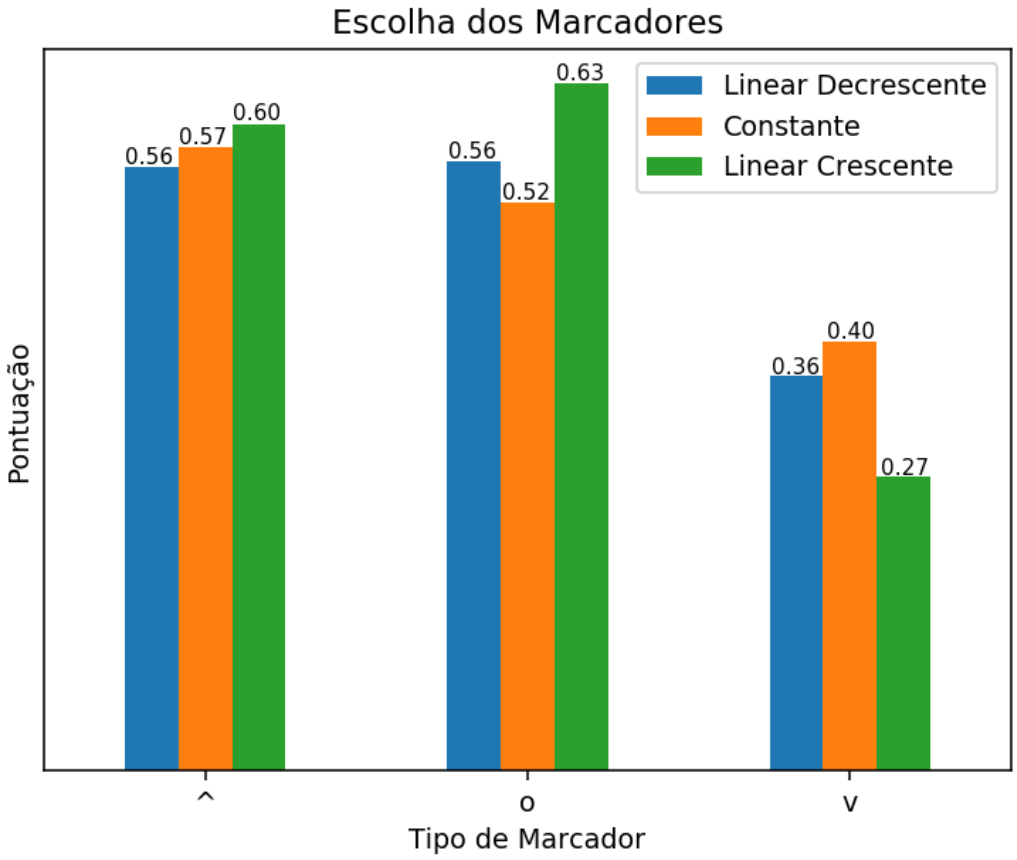

Figura 2 - Pontuação para os diferentes tipos de marcadores e dados, onde os indicadores significam: ^ - Triângulo equilátero apontado para cima; o - Circunferência; v - Triângulo equilátero apontado para baixo.

Para o gráfico com variação linear crescente, o tipo de marcador no formato 'o' foi o mais escolhido pelos participantes que realizaram nossa pesquisa, com 0.63 pontos. Já no gráfico constante, o triângulo superior foi o marcador preferido, com 0.57 pontos. Uma peculiaridade foi o empate técnico entre os marcadores 'o' e '^' na escolha durante o gráfico linear decrescente, ambos com 0.56. O marcador 'v' foi o menos escolhido dentre os tipos com no máximo 0.40 pontos.

Tamanho dos marcadores

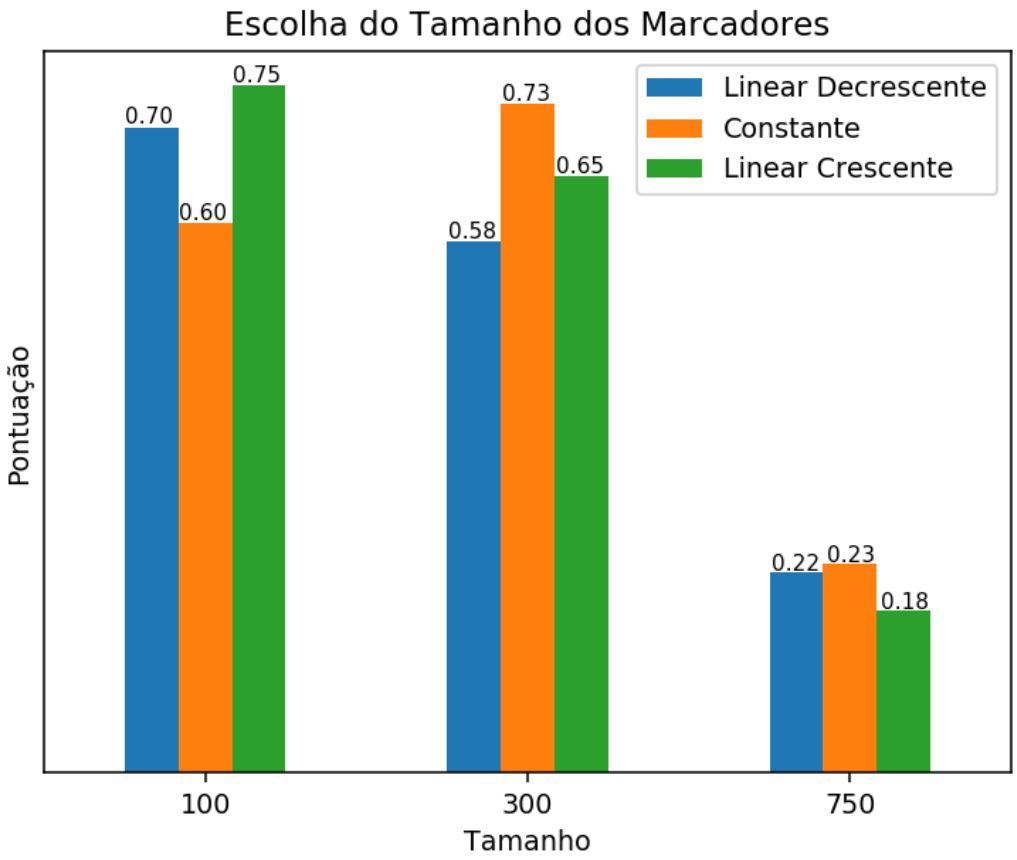

Figura 3 - Pontuação para os diferentes tamanhos de marcadores.

Quanto aos tamanhos, houve praticamente um empate em qual foi o mais popular. Apesar do tamanho 100 ter sido mais escolhido quando houve alguma tendência linear nos dados, com pontuações de 0.70 no decrescente e 0.75 no crescente. Contudo, quando essa alteração não foi apresentada, o tamanho 300 foi escolhido com pontuação 0.73. O tamanho 750 foi o pior dos parâmetros, dada a sua baixa pontuação nos três tipos de gráficos.

Mapa de cores dos marcadores

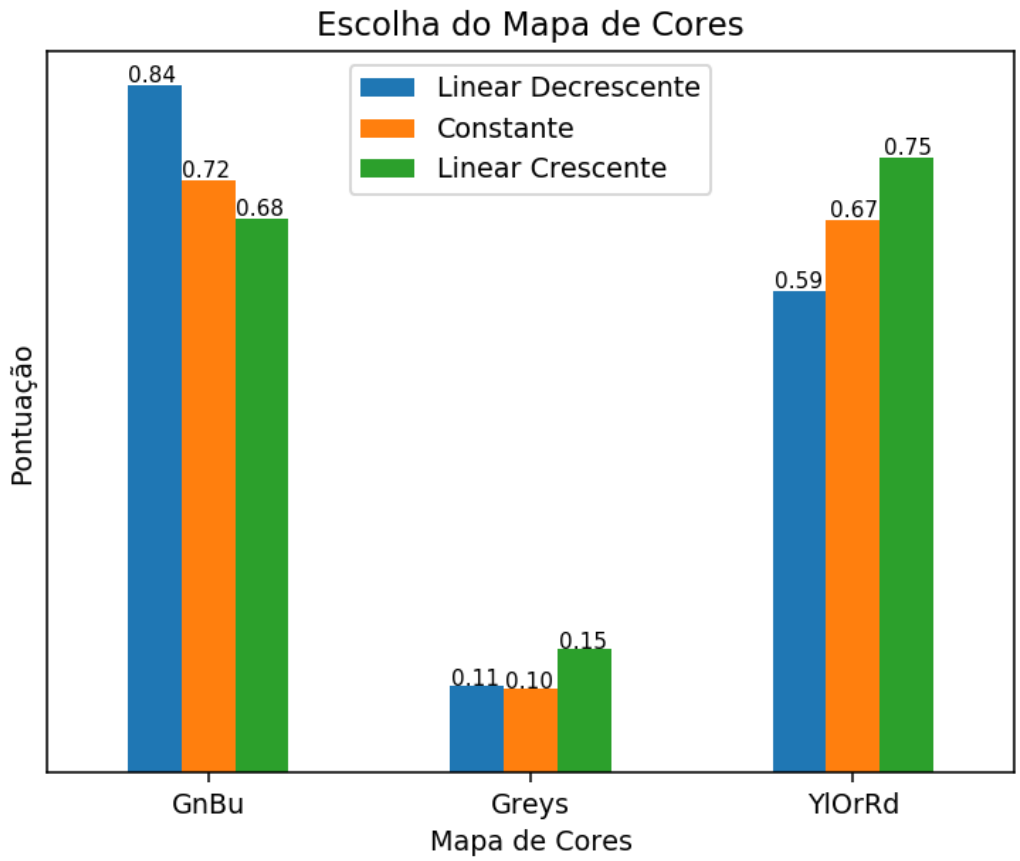

Figura 4 - Pontuação para os diferentes mapas de cores, onde: GnBu - Colormap sequencial de Verde para Azul; Greys - Colormap sequencial de tonalidades de Cinza; YlOrRd - Colormap sequencial de Amarelo para Vermelho.

O gráfico Linear Decrescente, o vitorioso foi o GnBu. Coincidentemente, para o gráfico Linear Crescente, o seu complementar (YlOrRd) apresentou uma maior adoção, enquanto houve uma leve escolha geral à adoção do GnBu como cor padrão pros gráficos. O Greys obtiveram a pior pontuação, obtendo um valor inferior a 0.2 em todos os parâmetros. Desta maneira, foi possível comprovar que, apesar de as cores Verde para Azul terem uma tendência maior de adoção, as cores Amarelo para Vermelho aparentam demonstrar a noção de Crescimento melhor do que as suas complementares.

Conjunto mais Votado - Por categoria e geral

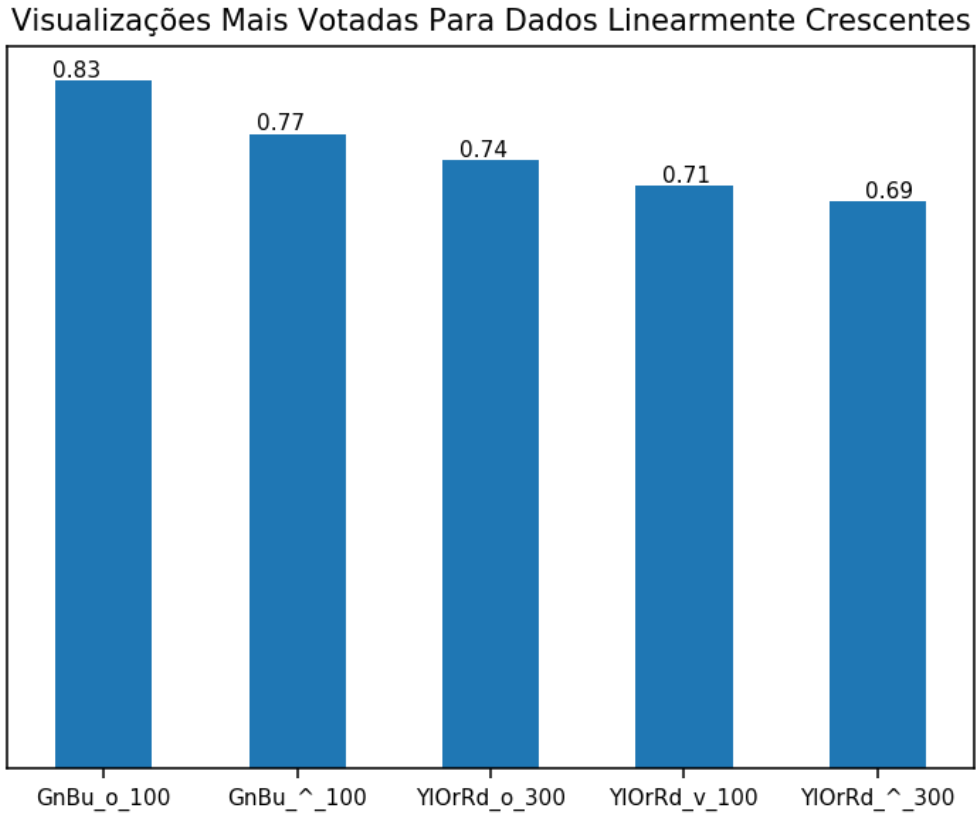

Figura 5 - As cinco visualizações com maiores pontuações para dados linearmente crescentes.

Para o conjunto de dados linearmente crescente, a visualização mais votada foi a com mapa de cores variando do verde para o azul, círculo e tamanho 100.

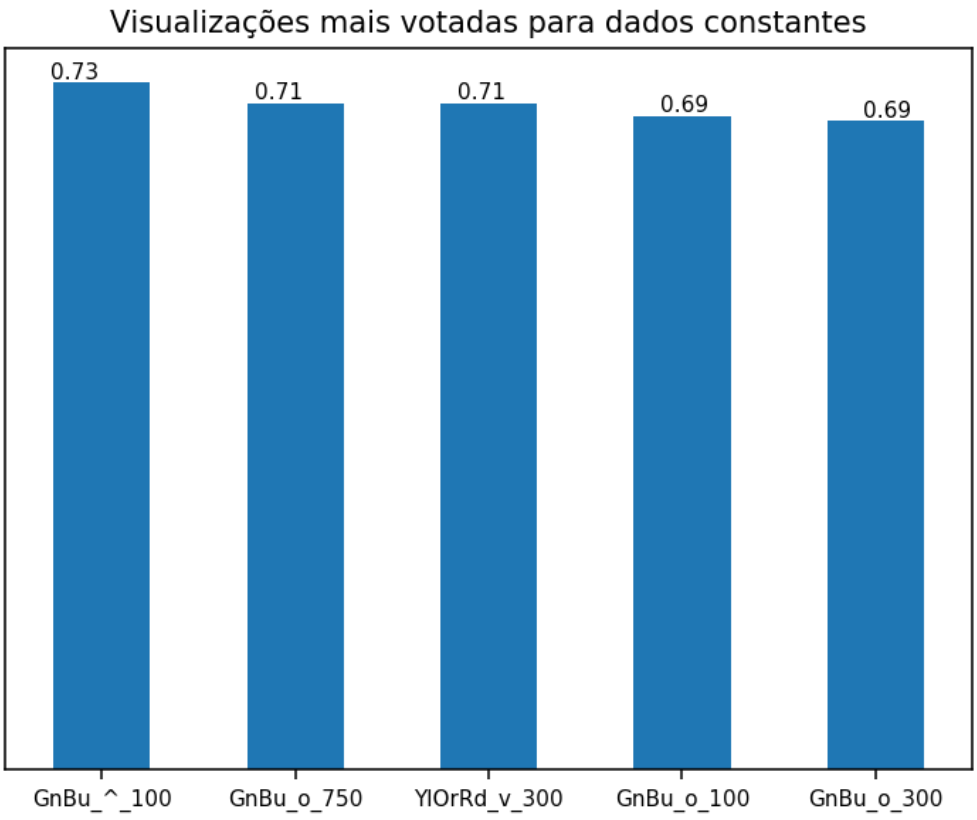

Figura 6 - As cinco visualizações com maiores pontuações para dados constantes.

Considerando os dados constantes, a visualização com maior pontuação foi com mapa de core do verde para o azul, com triângulo para cima e marcador do tamanho 100.

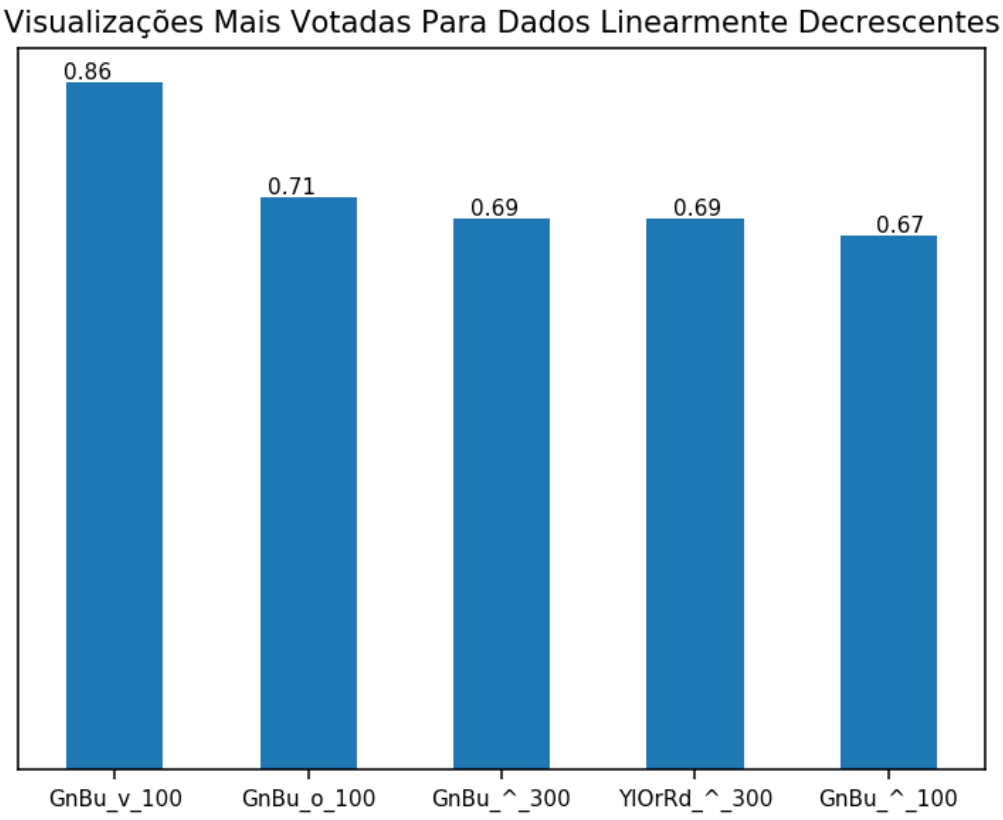

Figura 7 - As cinco visualizações com maiores pontuações para dados linearmente decrescentes.

Para dados linearmente decrescentes a visualização com maior pontuação, independente do parâmetro analisado foi a com mapa de cores do verde para o azul, com triângulo para baixo e marcador do menor tamanho.

Conclusões

Pelo nosso estudo, os resultados estão de acordo com os guias de boas práticas definidas por Jock D. Mackinlay em "Automating the Design of Graphical Presentations of Relational Information". Os parâmetros votados em geral seguiram as normas observadas por Mckinlay e este fato pode indicar ou que a população pesquisada estava acostumada com este tipo de visualização ou que o guia de boas práticas para visualização de gráficos é realmente efetiva.

Ademais, o uso de escala de cor para visualização em formato eletrônico foi muito bem avaliada, especialmente em comparação a escalas de cinza e este recurso deveria ser melhor explorado.

Esperava-se que existisse uma certa dessemelhança entre as visualizações mais votadas em conjunto com os parâmetros mais votados dado dois fatores:

- Estrutural: Não foi analisado as opções mostradas mas sim qual parâmetro está sendo variado. Ou seja, o conjunto pode ser votado mas apenas um dos parâmetros mostrados será avaliado e, assim, não valorizará os outros parâmetros da visualização.
- Perceptivo: a percepção do conjunto pode ser diferente da percepção individual. O conjunto pode ser mais votado em detrimento de um dos seus parâmetros.

Lições Aprendidas

- Em projetos de pesquisa como esse, o aumento no número de parâmetros observados pode acarretar em um crescimento descontrolado do número de amostras necessárias para que se tenha significância estatística dos resultados. Além disso, com o crescimento do número de parâmetros a análise se torna excessivamente complexa. Ou seja, o processo de escolha dos parâmetros estudados é muito importante para o sucesso do experimento;
- Pesquisas envolvendo participação pública exigem grande planejamento em relação ao formato empregado em sua condução. Se as entrevistas forem feitas pessoalmente, é necessário o planejamento de local e horário para tal operação. Por outro lado, se forem virtuais é necessário que um servidor seja mantido sem interrupções e aceitando conexões simultâneas para que os usuários acessem sem limitações;

Trabalhos Futuros

Este é um vasto campo de pesquisa e há muito a ser explorado, valendo ressaltar::

- Pesquisa com maior número de parâmetros
- Análise de escolhas para cada par apresentado;
- Pesquisa com maior número de participantes;
- Análise do tempo gasto por gráfico e adicionar na ponderação;
- Análise de respostas levando em conta gênero e formação profissional;
- Adicionar questionamentos relacionados à compreensão do gráfico dado uma característica dada, como: “o desmatamento está aumentando, qual visualização deixa isso mais claro?”.