Trabalho 1 - SI803 Visualização da Informação

Brenda Cristina de Souza Silva 194836
Mariana Albino Q. A. Queiroz 202919

INTRODUÇÃO

Para este projeto, foi proposto que fosse feita a análise visual de dois *datasets*, em 3 tipos diferentes de técnicas para visualizar dados multivariados, sendo as técnicas: Multidimensional Scaling (MDS), Scatterplot Matrices e Barras justapostas.

Como os dois datasets analisados se tratam de diferentes assuntos, dividiremos o trabalho em duas partes, sendo cada parte destinada a um dataset diferente.

REFERENCIAL TEÓRICO DAS TÉCNICAS UTILIZADAS

Multidimensional Scaling (MDS): é uma técnica que encontra uma projeção de pontos de baixa dimensão (no nosso caso, bidimensional), onde ela tenta ajustar as distâncias entre os pontos da melhor forma possível. O ajuste perfeito é normalmente impossível de obter, pois os dados são de alta dimensão ou as distâncias não são euclidianas.

Scatterplot Matrices: Uma matriz de gráfico de dispersão é uma coleção de gráficos de dispersão organizados em uma grade (ou matriz). Cada gráfico de dispersão mostra a relação entre um par de variáveis.

Barras justapostas (Table Lens): O Table Lens é um método comum de visualização de informações para explorar dinamicamente grandes quantidades de dados tabulares. Emprestando a partir do modelo de planilha, a Table Lens exibe valores de dados regionais em colunas e linhas sem barras de rolagem e sem ocultar quaisquer dados e preenchendo as células com barras pequenas horizontais em escala e coloridas. Cada linha na Table Lens representa uma única região e as colunas representam um indicador específico.

PARTE 1

O primeiro dataset utilizado para visualização chama-se Graduate Admissions, do arquivo “Admission_Predict_Ver1.1.csv”. Este dataset se trata de parâmetros de candidatos para ingressar em programas de mestrado em

universidades dos Estados Unidos, visto de uma perspectiva indiana. Estes parâmetros são:

1. Pontos no GRE, sendo este uma prova padronizada e utilizadas por várias universidades dos EUA para avaliação dos ingressantes, o que poderia ser comparado ao nosso vestibular.
2. Pontos no TOEFL, prova aplicada para testar a proficiência em inglês do ingressante.
3. University Rating, sendo de 0 a 5, trata-se de uma pontuação para avaliar a qualidade de uma universidade.
4. Força da declaração de propósito (SOP) e carta de recomendação (LOR), avaliadas de 0 a 5.
5. GPA, também avaliado de 0 a 5, trata-se da média da pontuação escolar do ingressante.
6. Experiência em pesquisa do ingressante (Research), valores 0 ou 1, 0 para negativo e 1 para positivo.
7. Chance de ser aceito no programa, varia de 0 a 1.

RESULTADOS

Barras justapostas (Table Lens): pudemos ver que de 100 candidatos, boa parte dos advindos de universidades de maior pontuação têm experiência em pesquisa, boas notas no TOEFL e maior chance de ser admitido. A barras estão em ordem decrescente por avaliação e por tom da cor roxa para notas do TOEFL, sendo da mais escura as mais altas e as mais claras, as notas mais baixas.

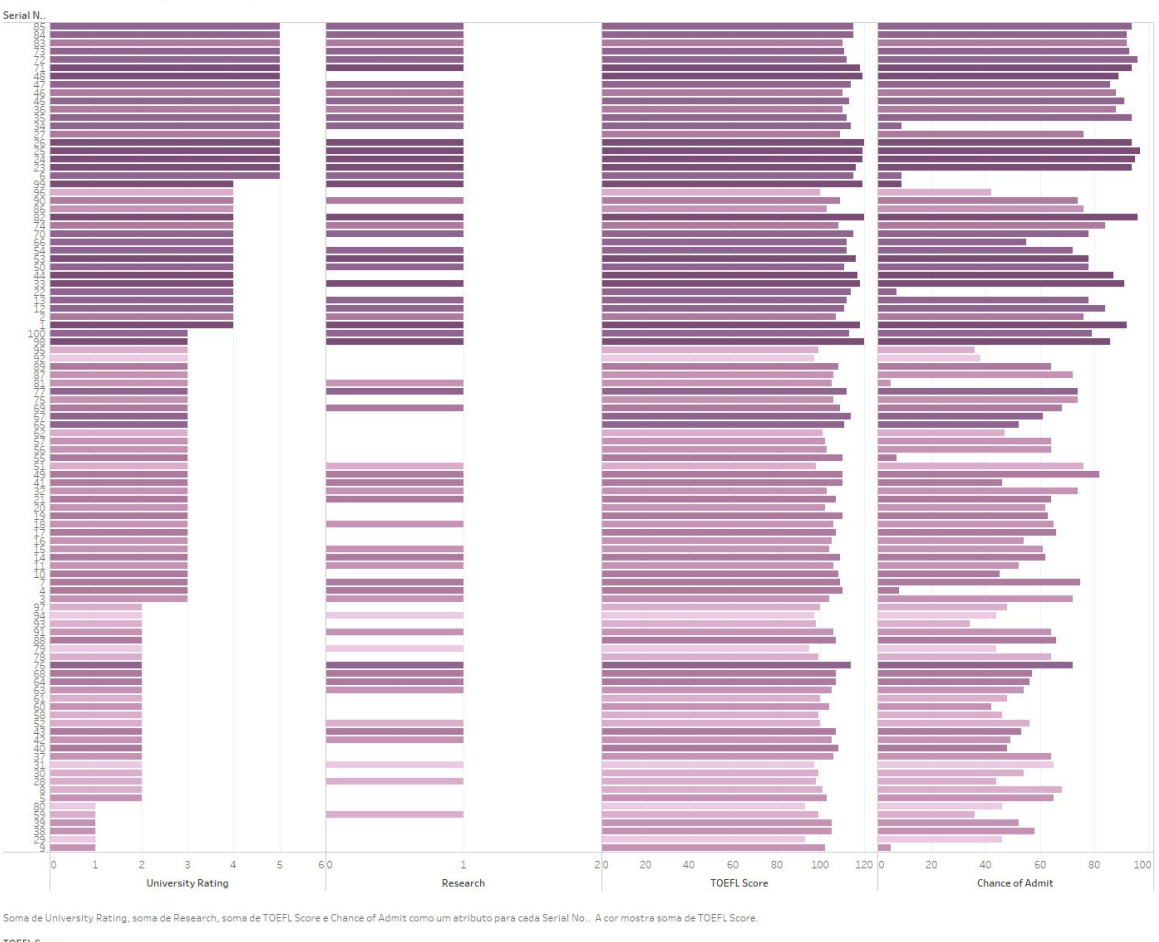

Figura 1 - Barras Justapostas da parte 1. A primeira coluna se refere ao parâmetro University Rating, a segunda a Research, seguidas por pontos no TOEFL e chance de ser admitido.

Multidimensional Scaling (MDS): neste tipo de representação, podemos ver que os candidatos são praticamente divididos em maior e menor chance de aprovação. Na área em amarelo estão os candidatos com maior chance de aprovação, e na azul, com menor chance. Os candidatos também estão separados por formato, sendo um X os que têm experiência com pesquisa, e em O os que não têm. A legenda de cada ponto, em números, é a pontuação de GRE de cada candidato. Podemos observar nesta representação que boa parte dos candidatos com maior pontuação GRE está no grupo amarelo, ou seja, no grupo com maior chance de aprovação. É possível observar também que as maiores notas no GRE estão com formato de X, ou seja, são candidatos com experiência em pesquisa.

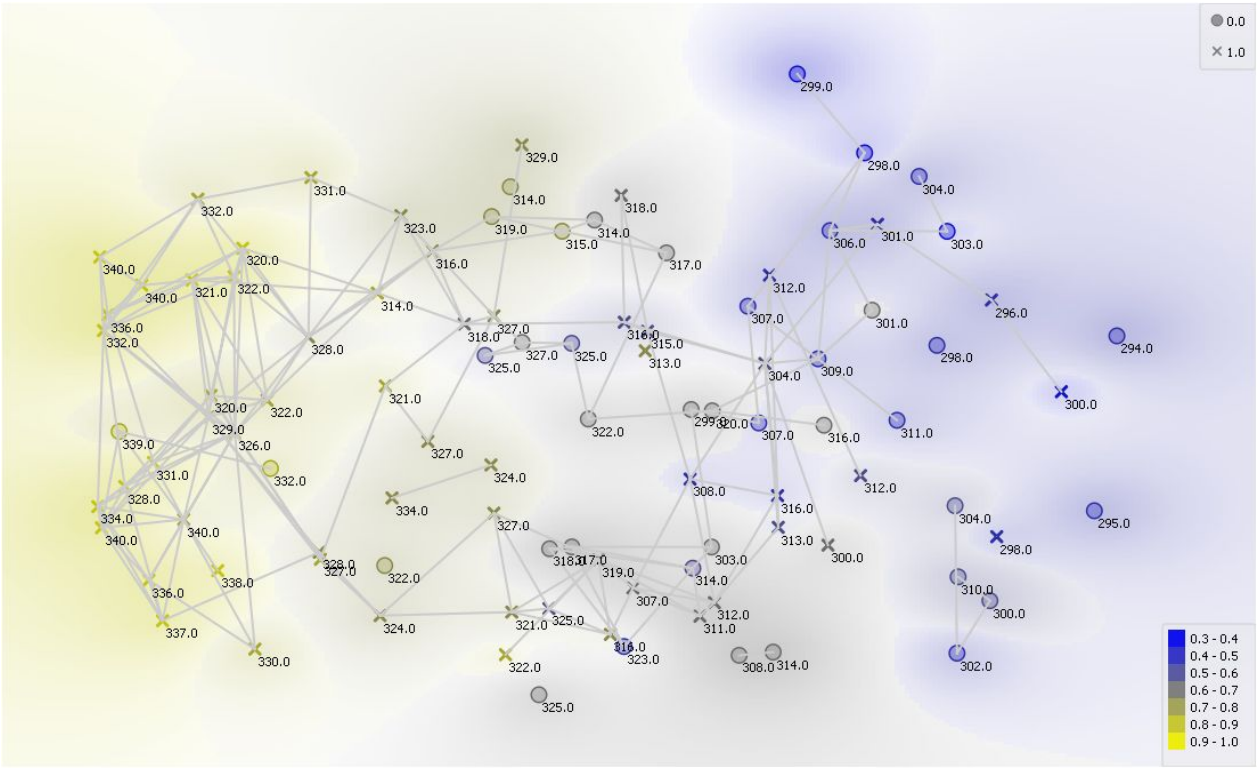

Figura 2 - Multidimensional Scaling (MDS) da parte 1.

Scatterplot Matrices: É possível notar que há uma relação entre o parâmetro TOEFL (Test of English as a Foreign Language) e Chance of Admit (Chance de Admissão) através do gráfico, visto que os valores formam uma aglomeração. Já entre LOR (Letter of Recommendation) e Chance of Admit os valores se encontram dispersos entre si, indicando que a pontuação do candidato no TOEFL tem maior influência na chance de admissão.

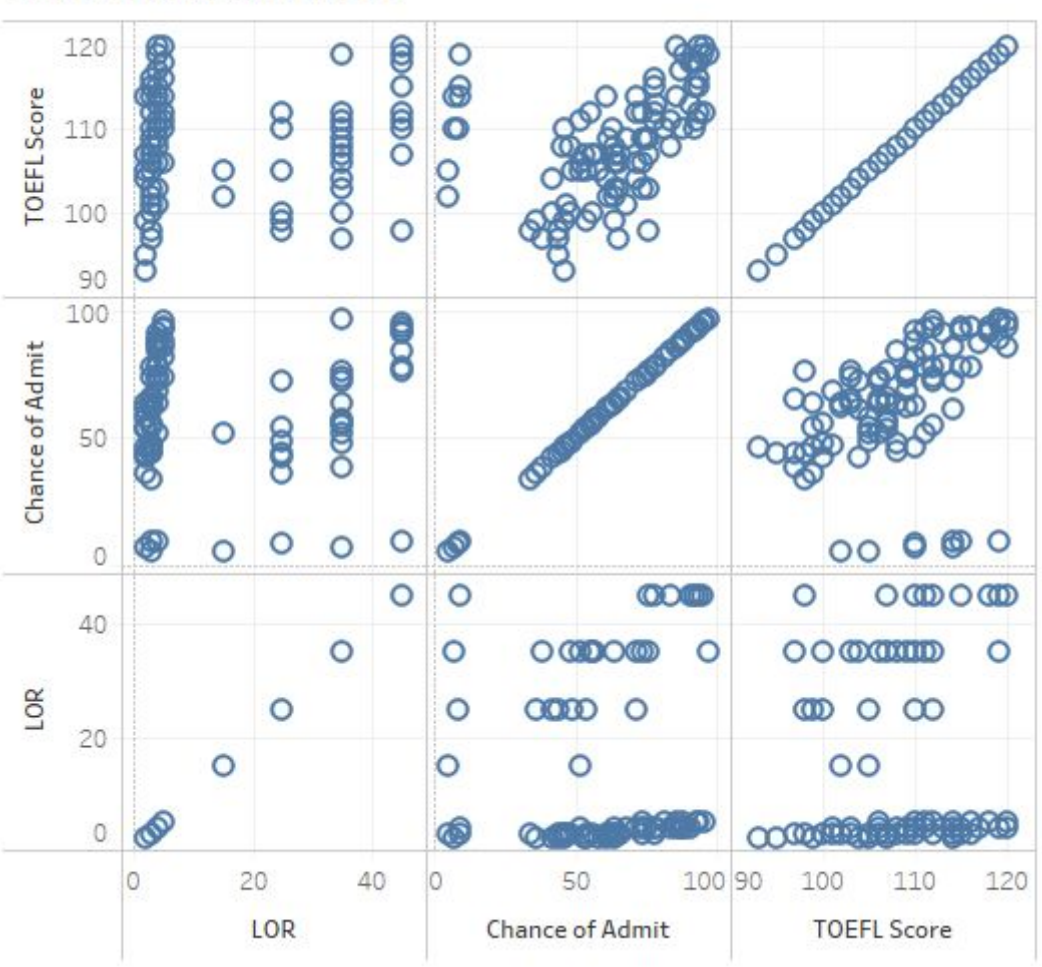

Figura 3 - Scatterplot Matrices da parte 1.

PARTE 2

O segundo dataset a ser trabalho se chama U.S. Education Datasets: Unification Project, do arquivo “states_all.csv”. Este dataset se trata das receitas e despesas federais, estaduais e locais das escolas dos EUA, por estado. Os dados deste dataset são de 1992 a 2017, e abrangem outros tópicos, como notas obtidas nas provas NAEP (provas nacionais para avaliação da educação em redação e matemática).

Para realização da análise visual desses dados, os seguintes parâmetros foram utilizados:

1. TOTAL_EXPENDITURE: despesa total da escola.
2. GRADES_4_G: total de alunos da quarta série.
3. GRADES_8_G: total de alunos da oitava série.
4. TOTAL_REVENUE: gastos totais da escola.

RESULTADOS

Barras justapostas (Table Lens): é possível observar uma forte correlação entre os gastos totais das escolas e a permanência de alunos na escola, demonstrando a importância de investimentos na área da educação. À proporção que os gastos aumentam, o número de alunos nas escolas se mantém ou amplifica. As colunas dos seguintes gráficos representam respectivamente Gastos totais do estado, quantidade de alunos da quarta série e quantidade de alunos da oitava série

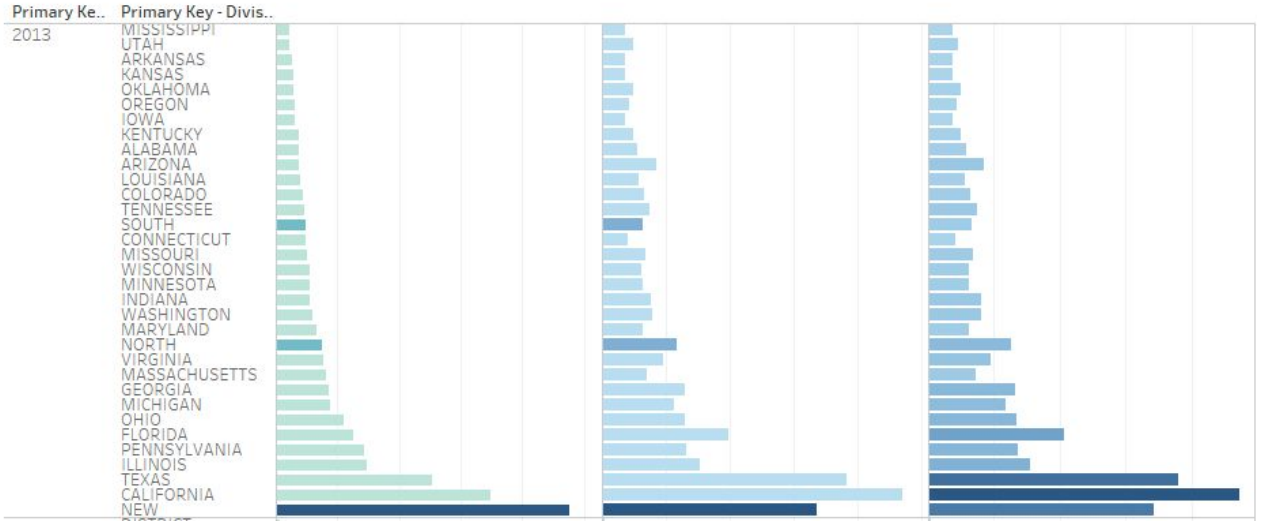

Figura 4 - Barras justapostas da parte 2. Ano 2013.

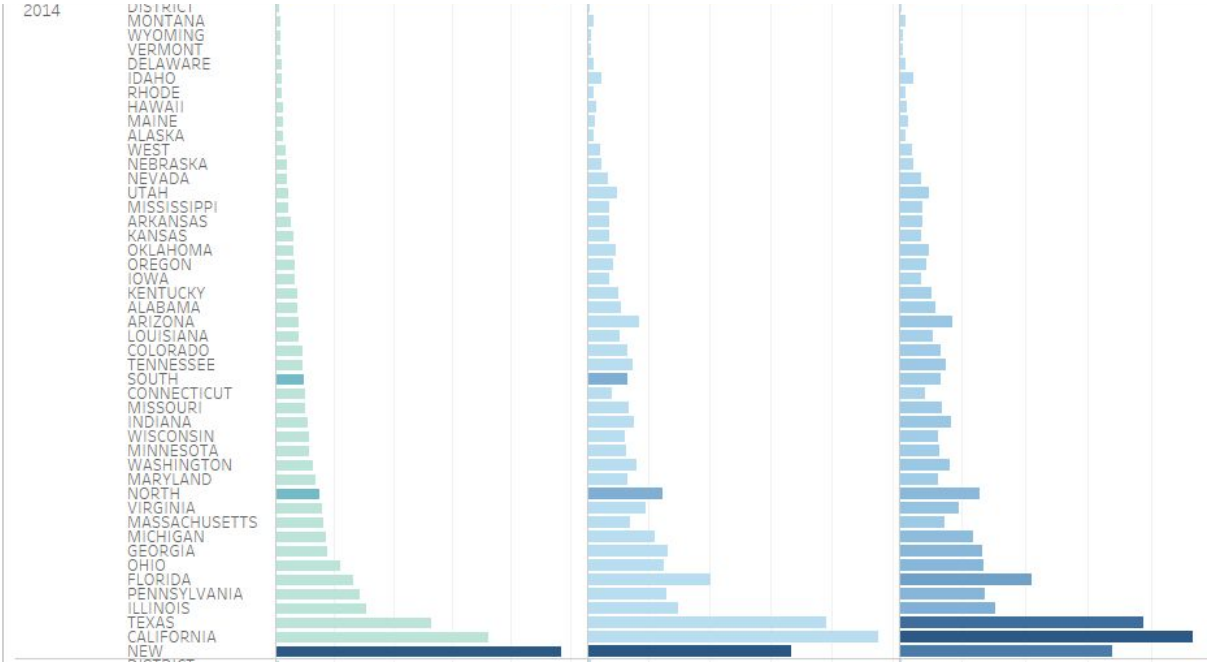

Figura 5 - Barras justapostas da parte 2. Ano 2014.

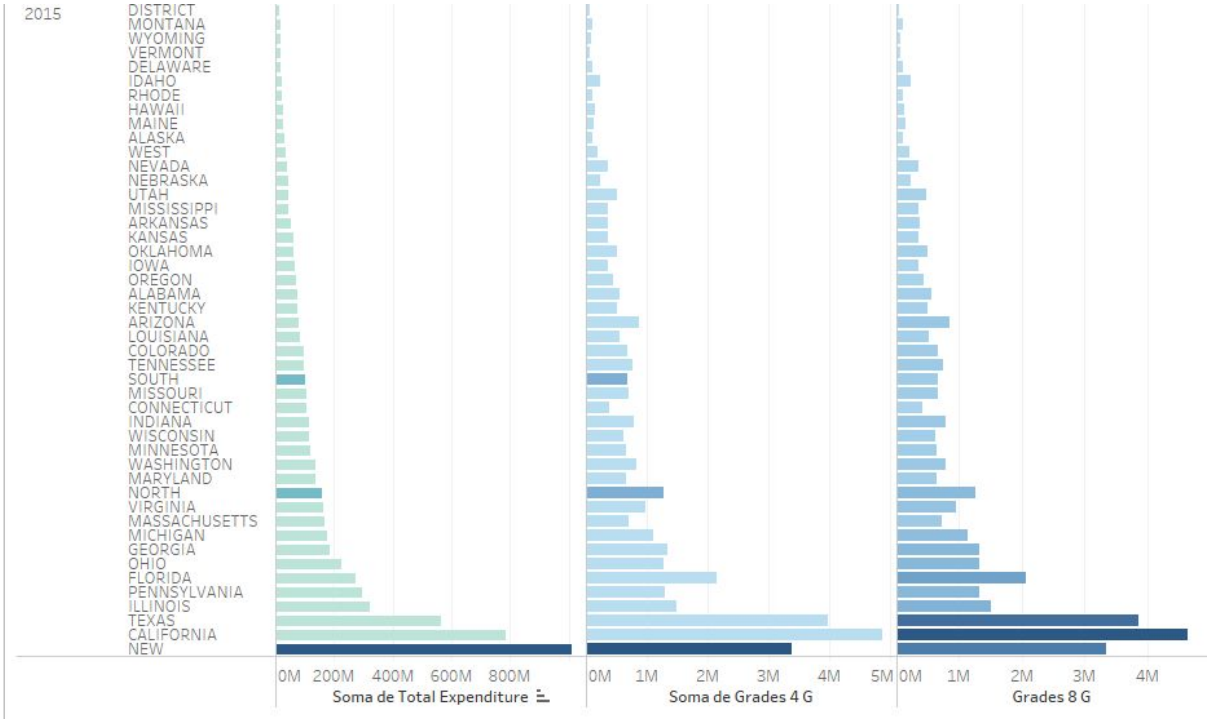

Figura 5 - Barras justapostas da parte 2. Ano 2015.

Multidimensional Scaling (MDS): Para esta representação, usamos cores para dividir os estados com maior receita dos com menor receita, sendo amarelo para as maiores e azul para menores, e usamos do tamanho dos pontos para referenciar as notas dos estudantes do 4 grau em redação, sendo os maiores pontos com as maiores notas, e os menores pontos o contrário. Podemos ver que, não necessariamente as escolas com maior receita tem as maiores médias em notas de redação.

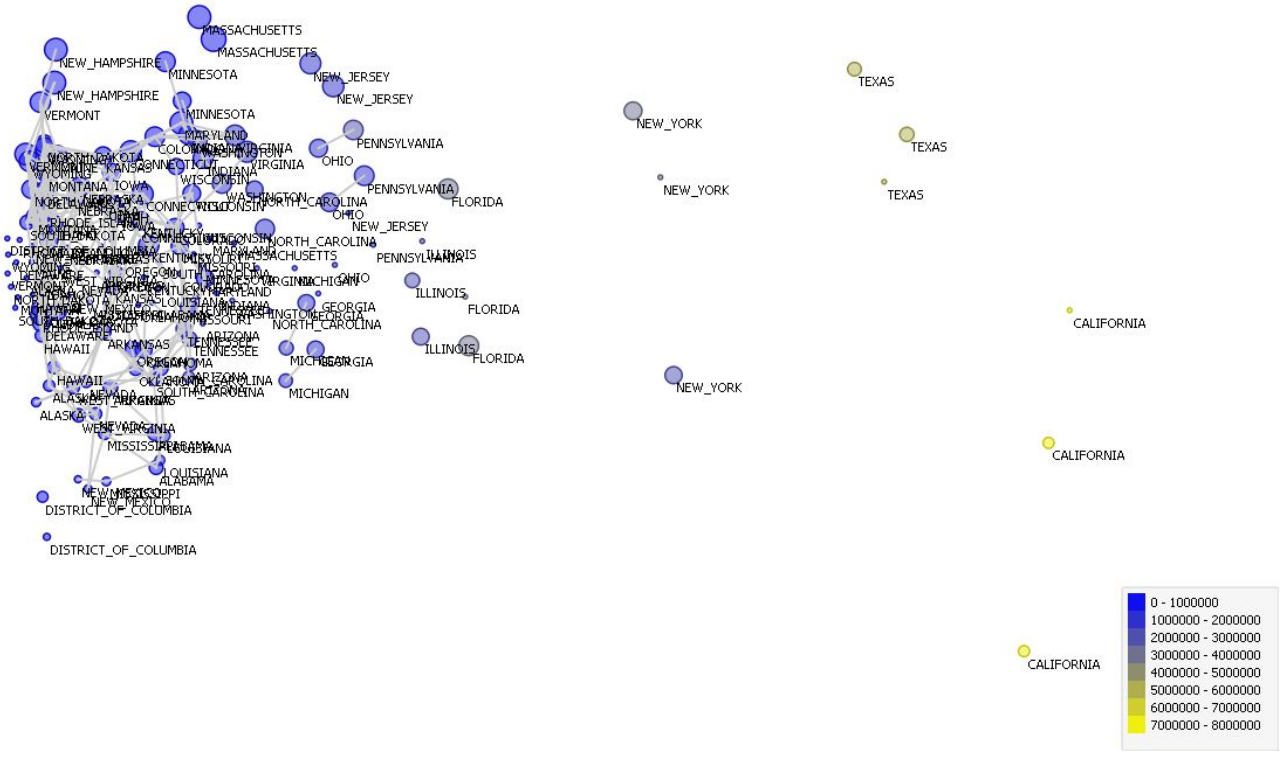

Figura 6 - Multidimensional Scaling (MDS) da parte 2.

Scatterplot Matrices: É plausível constatar a relação entre o Total Expenditure (Custo total) e Gades 4 G(quantidade de alunos na quarta série) na figura 1, visto que quanto menor os gastos menor o número de alunos. Também é notável a relação quase ideal entre a receita da escola (Total Revenue) e os gastos, demonstrando que quase o mesmo montante recebido é gasto.

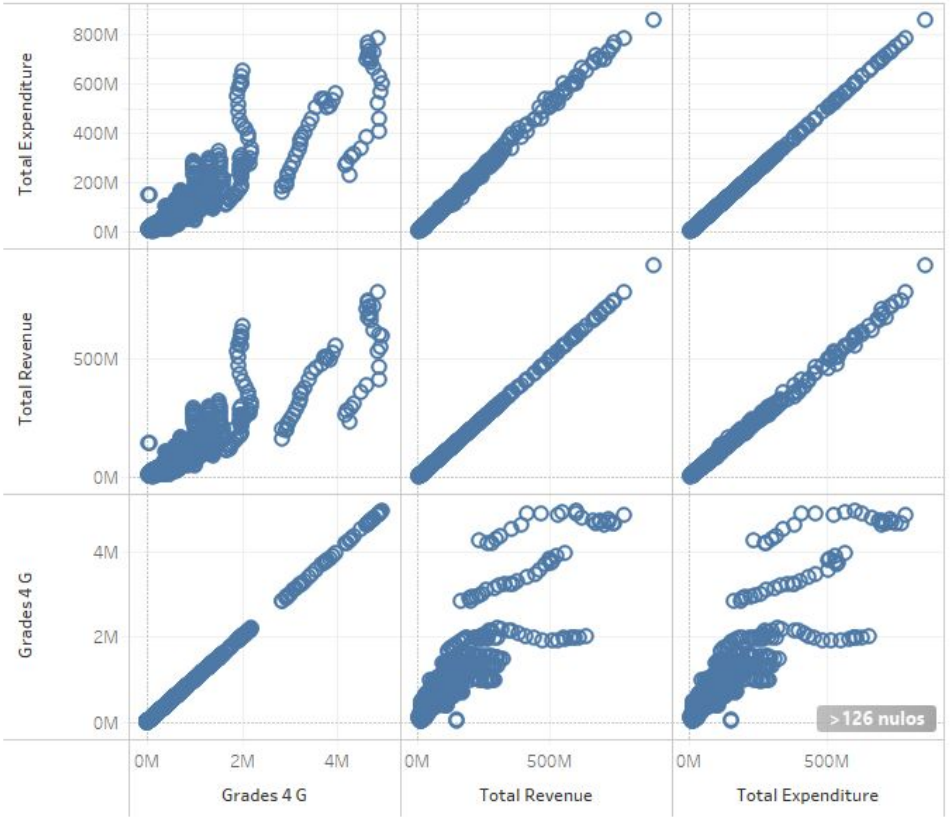

Figura 7 - Scatterplot da parte 2 com todos anos.

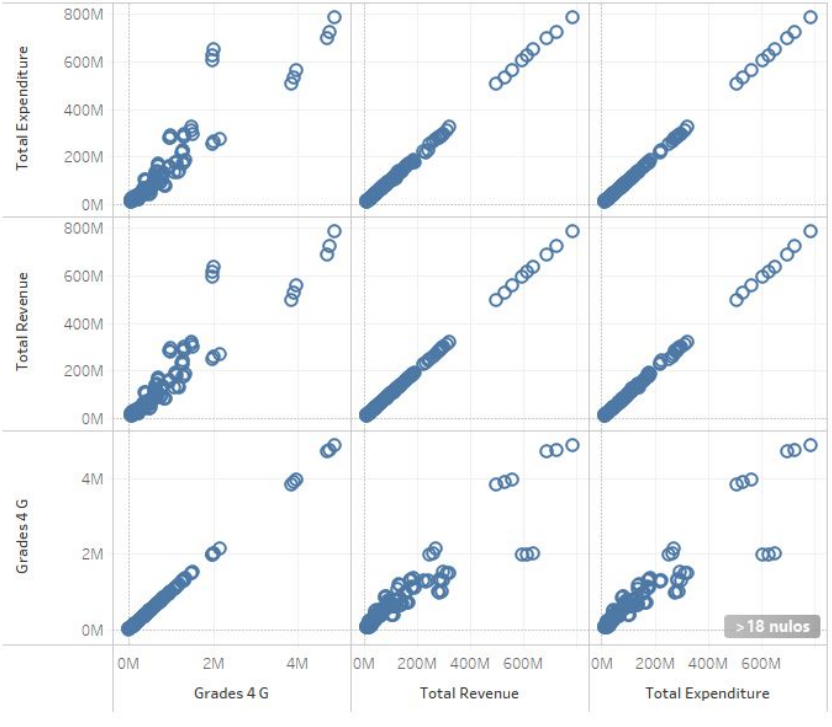

Figura 8 - Scatterplot da parte 2,entre os anos de 2013 a 2015.

Com o objetivo de averiguar os outliers nas intersecções dos parâmetros,Total Revenue com Grades 4 G e Total Expenditure com Grades 4 G, geramos outro scatterplot, como mostra a figura 8, no qual identificamos que as separações são relacionadas com a quantidade de habitantes de cada estado, e como esperado, estados com maior população, possuem maior alunos matriculados, tal ligação foi encontrada a partir de pesquisas externas aos datasets.

CONSIDERAÇÕES FINAIS

Durante a apresentação recebemos a orientação de que, com o propósito de gerar um MDS (Multidimensional Scaling) mais equitativo, normalizar todos os valores dos parâmetros.

REFERÊNCIAS

Multidimensional scaling (MDS). Orange Visual Programming 3 documentation. Disponível em: <https://docs.biolab.si//3/visual-programming/widgets/unsupervised/mds.html> Acessado em: 10 de maio de 2019.

Scatterplot Matrix. JMP Statistical Discovery from SAS. Disponível em: <https://www.jmp.com/support/help/14-2/scatterplot-matrix.shtml> Acessado em: 10 de maio de 2019.

Table Lens. Ncomva User guide. Disponível em: <http://www.ncomva.se/guide/?chapter=Visualizations&section=Table%20Lens#_General> Acessado em: 10 de maio de 2019.

Graduate Admission data. Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

U.S Education Datasets: Unification Project. U.S. Census Bureau and the National Center for Education Statistics (NCES).