CENTRO UNIVERSITÁRIO INSTITUTO DE EDUCAÇÃO SUPERIOR DE BRASÍLIA

Bacharelado em
Ciência de Dados e Inteligência Artificial

Tassio Lucian de Jesus Sales

1922120021

Análise de Cluster do ENEM 2019

Brasília
2021

Lista de ilustrações

Figura 1 – Frequência de inscritos por ano da base de dados completa. 7
Figura 2 – Caracterítica da base de dados filtrada para o ENEM 2019. 8
Figura 3 – Variáveis e tipos de dados da base filtrada ENEM 2019. 8
Figura 4 – Estatísticas de Resuma da idade dos candidatos do ENEM 2019. 9
Figura 5 – Distribuição da idade dos candidatos do ENEM 2019. 9
Figura 6 – Gráfico de barras da frequência dos candidatos do ENEM 2019 por Sexo. 10
Figura 7 – Gráfico de Donut do percentual da frequência dos candidatos do ENEM 2019 por Sexo. 10
Figura 8 – Frequência dos candidatos do ENEM 2019 por Cor e Raça. 11
Figura 9 – Frequência dos candidatos do ENEM 2019 por Tipo de Escola. 12
Figura 10 – Frequência dos candidatos do ENEM 2019 por Região no Brasil. 13
Figura 11 – Frequência dos candidatos do ENEM 2019 por Unidade da Federação (UF) no Brasil. 14
Figura 12 – Estatística de Resumo das notas dos candidatos do ENEM 2019 no Brasil. 15
Figura 13 – Fluxo de geração de Estatística de Resumo das notas dos candidatos do ENEM 2019 no Brasil por UF. 17
Figura 14 – Fluxo de geração do Cluster pelo SAS Enterprise Miner. 18
Figura 15 – Segment Size pelo método Ward. 19
Figura 16 – Mean Statistics pelo método Ward. 19
Figura 17 – Segment Plot pelo método Ward. 20
Figura 18 – Cluster Distance pelo método Ward. 20
Figura 19 – Input Means pelo método Ward. 21
Figura 20 – Melhores notas por estado - Cluster 1. 21
Figura 21 – Piores notas por estado - Cluster 5. 21
Figura 22 – Clusters do estado de SP. 22
Figura 23 – Clusters do estado de AM. 22
Figura 24 – Segment Size pelo método Centróid. 23
Figura 25 – Mean Statistics pelo método Centróide. 23
Figura 26 – Segment Plot pelo método Centróid. 24
Figura 27 – Cluster Distance pelo método Centróid. 24
Figura 28 – Input Means pelo método Centróid. 25
Figura 29 – Melhores notas por estado - Cluster 6. 25
Figura 30 – Piores notas por estado - Cluster 7. 25
Figura 31 – Clusters do estado de RJ. 26
Figura 32 – Clusters do estado de TO. 26
Figura 33 – Segment Size pelo método Average. 27

Figura 34 – Mean Statistics pelo método Average. 27
Figura 35 – Segment Plot pelo método Average. 28
Figura 36 – Cluster Distance pelo método Average. 28
Figura 37 – Input Means pelo método Average. 29
Figura 38 – Melhores notas por estado - Cluster 2, 4 e 6. 29
Figura 39 – Piores notas por estado - Cluster 9. 29
Figura 40 – Clusters do estado de RS. 30
Figura 41 – Clusters do estado de AP. 30

Lista de tabelas

Tabela 1 – Dicionário da amostra de dados fornecida dos ENEMs 2017-2019 *(INEP)* 36

Sumário

1 INTRODUÇÃO 6

2 DESENVOLVIMENTO 7
2.1 Visão Geral dos dados 7
2.2 Análise Exploratória ENEM 2019 9

3 ANÁLISE DE CLUSTER 17
3.1 Nó de Cluster 18
3.2 Ward 19
3.3 Centróide 23
3.4 Average 27

4 CONCLUSÕES 32

REFERÊNCIAS 33

ANEXOS 35

ANEXO A – DICIONÁRIO DE DADOS 36

ANEXO B – CÓDIGOS SAS BASE 37

1 INTRODUÇÃO

A **Ciência de Dados** é uma área nova de ciência, multidisciplinar, que se baseia na integração das áreas de matemática, estatística e computação. Em termos simples, resume-se na exploração e análise de dados visando à extração de informações e conhecimento a partir dos dados. A Ciência dos Dados em muitos aspectos é uma consequência da necessidade de analisar grandes bancos de dados (Big Data), até pouco tempo conhecidos como Very Large Database (VLDB), que possuem características de grandes volumes de dados, alta velocidade de geração e armazenamento de novos dados, variedade de tipos de dados, alta qualidade nos dados e o processamento de seus dados agregam valor nas organizações. A Ciência de Dados é um campo interdisciplinar que exige a formação de um novo profissional que possua as habilidades e competências no uso da matemática, da estatística e da computação, aplicados na extração de informações e conhecimentos dos dados (HAN; KAMBER; PEI, 2012).

O **Aprendizado de Máquina** tornou-se um dos tópicos mais importantes dentro das organizações de desenvolvimento que buscam maneiras inovadoras de aproveitar ativos de dados para ajudar a empresa a obter um novo nível de compreensão. Com os modelos de aprendizado de máquina apropriados, as organizações têm a capacidade de prever continuamente as mudanças nos negócios para que possam prever o que está por vir. Como os dados são adicionados constantemente, os modelos de aprendizado de máquina garantem que a solução seja atualizada constantemente. O valor é direto: se você usar as fontes de dados mais adequadas e em constante mudança no contexto do aprendizado de máquina, terá a oportunidade de prever o futuro (GRUS, 2016).

O Exame Nacional do Ensino Médio (Enem) avalia o desempenho escolar ao final da educação básica. Realizado anualmente pelo Inep, desde 1998, o Enem colabora para o acesso à educação superior – por meio do Sisu, do Prouni e de convênios com instituições portuguesas – e a programas de financiamento e apoio estudantil, caso do Fies. Os resultados também permitem o desenvolvimento de estudos e indicadores educacionais. Qualquer pessoa pode fazer o Enem, entretanto, participantes “treineiros” podem usar o resultado somente para autoavaliação de conhecimentos (PATRíCIA., 2017).

O objetivo da **Análise de Cluster** é segmentar um conjunto de dados num número de subgrupos homogêneos ou clustering. É formar grupos baseados no princípio de que esses grupos devem ser os mais homogêneos em si e os mais heterogêneos entre si, maximizar a similaridade intra-classe e minimizar similaridade inter-classe. Não existem classes predefinidas para classificar os registros em estudo. Os registros são agrupados em função de suas similaridades básicas. Seleciona-se um conjunto de atributos (variáveis) e em função da similaridade desses atributos são formados os grupos.

2 Desenvolvimento

Para este projeto, foi utilizado a ferramenta **SAS ou Statistical Analysis System (Sistema de Análise Estatística)**, é uma empresa pioneira em Business intelligence e de uma família de softwares gerenciadores de bancos de dados comercializados por ela. Sendo aqui utilizado os produtos **SAS Enterprise Guide** para a etapa de Análise Exploratória e Visualização de Dados e **SAS Enterprise Miner** para a criação dos Clusters.

Para tal, uma amostra de dados dos ENEMs de 2017-2019 foi disponibilizada pelo professor, onde apartir desta, como verifica-se na Figura 1, vemos que nos três anos 8.909.563 candidatos estão considerados na amostra completa. Destaca-se que estão incluídos neste total somente os candidatos que realizaram todas as provas, excluindo os *treineiros* e candidatos que não realizaram pelo menos uma prova.

Outro fato relevante que podemos observar ainda na Figura 1 é o número de candidatos em 2018 foi o maior dentre os três anos, com 3.360.936 candidatos, representando dentre os anos cerca de 37,72% da base completa. Observa-se que de 2017 para 2018 há um aumento de mais de 10% no número de candidatos e uma queda de mais de 3% entre 2018 e 2019.

NU_ANO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2017	2461392	27.63	2461392	27.63
2018	3360936	37.72	5822328	65.35
2019	3087235	34.65	8909563	100.00

Figura 1 – Frequência de inscritos por ano da base de dados completa.

2.1 Visão Geral dos dados

Como vimos no objetivo da Análise de Cluster, necessitou-se da realização de um filtro na base completa, para que se pudesse ter apenas os registros do ENEM do ano de 2019, sendo assim, a partir deste ponto, é demonstrado apenas os dados do ENEm referente ao ano de 2019. A seguir na Figura 2 é demonstrado a característica da amostra de dados filtrada para o ENEM do ano de 2019.

Vemos que a base possui 3.087.235 observações para o ENEM 2019, com 20 variáveis. Por meio desta visualização da característica da base, vemos ainda que o SAS fornece informações como, onde que a base foi verificada, ou seja, por qual **Library** ela foi lida, neste caso pelo *DADOS.ENEM_2019*, qual o tipo de base, no caso *DATA* ou dados, qual a

Data Set Name	DADOS.ENEM_2019	**Observations**	3087235
Member Type	DATA	**Variables**	20
Engine	BASE	**Indexes**	0
Created	21/11/2021 17:12:59	**Observation Length**	200
Last Modified	21/11/2021 17:12:59	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	YES
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Figura 2 – Caracterítica da base de dados filtrada para o ENEM 2019.

Engine que neste mostra que é *BASE*, ou seja, como foi gerada a partir do SAS, ela é uma base de dados do tipo *BASE ou SAS DATASET*, como também outras informações, de data da criação, data da modificação, por meio de qual **Sistema Operacional o Software (SO)** foi executado, neste caso *WINDOWS_64*, o **encoding** neste caso o *wlatin Western (Windows)*, dentre outras mais informações.

Portando, agora com a base de daods filtrado somente para observações que contenham o ENEM do ano de 2019, e por meio do auxílio da ferramenta **SAS Enterprise Guide**, realizou-se uma análise exploratória e demonstrado todas as variáveis que se tem nesta base de dados, essas variáveis com seus atributos, ou seja, na figura 3, tem o nome de cada variável na base de dados, o npumero da variável, o tipo do dado, o formato, a label ou nome lógico e o tamanho.

Name	Variable Number	Type	Format	Label	Length
COR_RACA	10	Character			13
CO_MUNICIPIO_RESIDENCIA	7	Numeric	BEST	CODIGO_MUNICIPIO_DV	8
LATITUDE	20	Numeric	F		8
LONGITUDE	19	Numeric	F		8
NOTA_MEDIA	18	Numeric			8
NO_MUNICIPIO_RESIDENCIA	8	Character	$CHAR	NOME_MUNICIPIO	32
NU_ANO	1	Numeric			8
NU_IDADE	12	Numeric			8
NU_INSCRICAO	2	Numeric			8
NU_NOTA_CH	14	Numeric			8
NU_NOTA_CN	13	Numeric			8
NU_NOTA_LC	15	Numeric			8
NU_NOTA_MT	16	Numeric			8
NU_NOTA_REDACAO	17	Numeric			8
Regiao_Nome	3	Character	$CHAR		12
SG_UF_RESIDENCIA	4	Character	$CHAR	UF	2
Sexo	9	Character			9
Tipo_Escola	11	Character			13
UF_Capital	6	Character	$CHAR		3
UF_Nome	5	Character	$CHAR		19

Figura 3 – Variáveis e tipos de dados da base filtrada ENEM 2019.

2.2 Análise Exploratória ENEM 2019

Analysis Variable : NU_IDADE					
Mean	Minimum	Maximum	N	N Miss	Median
21.8953105	2.0000000	91.0000000	3087235	0	19.0000000

Figura 4 – Estatísticas de Resuma da idade dos candidatos do ENEM 2019.

A variável NU_IDADE de acordo com Figura 4, vemos que está não possui valores missing, observamos ainda que a maior idade segunda essa amostra de dados é de candidato com idade com 91 anos e a menor idade é de um candidato com 2 anos, sendo portanto algum erro na base. A média das idades nessa edição do ENEM de 2019 está entre 21.89 anos, sendo boa parte dos participantes mais jovens, como vemos no histogram da Figura 5, causa geralmente comum por serem candidatos que estão se preparando para concluir o ensino médio, pois é necessário se fazer o Exame Nacional do Ensino Médio. Na variável NU_IDADE percebe-se que que a maior quantidade de inscritos está na faixa etária de 17 a 20 anos, sendo a moda candidatos com 19 anos, ou seja, a maioria dos candidatos possuem idade de 19 anos nessa amostra de dados.

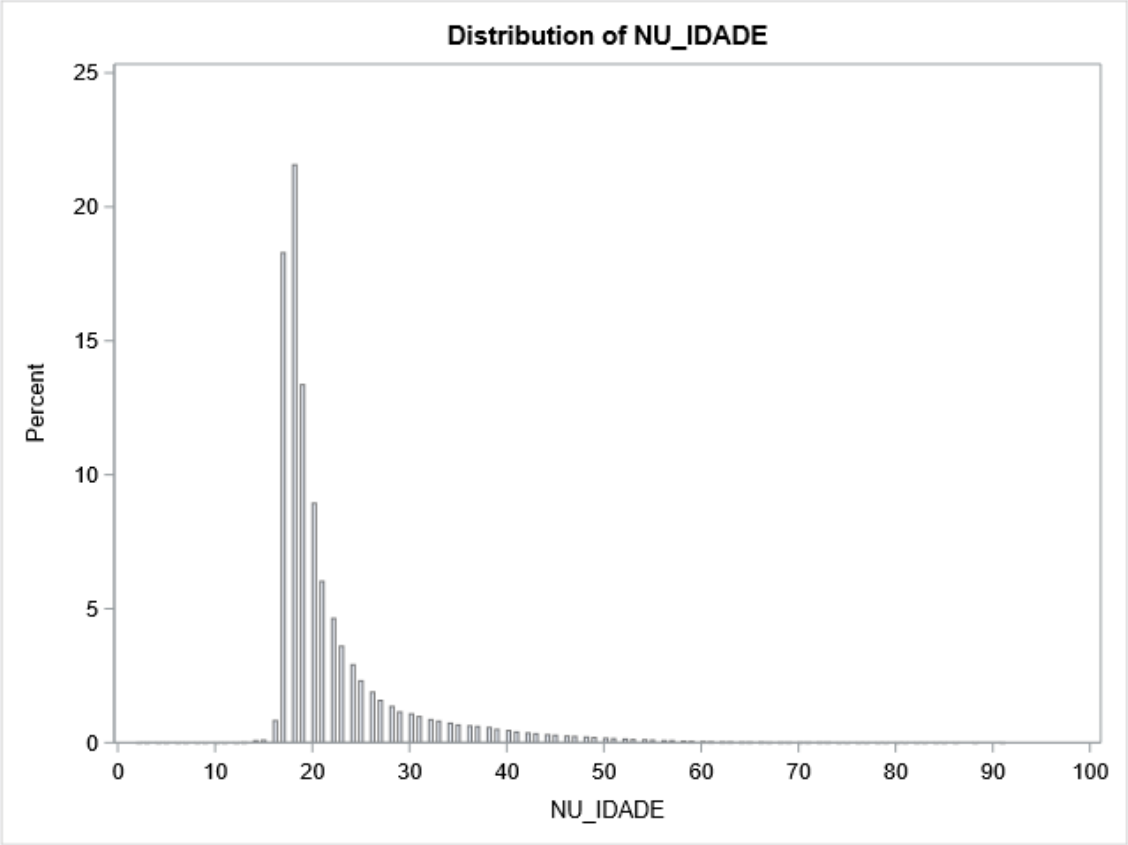

Figura 5 – Distribuição da idade dos candidatos do ENEM 2019.

Para a variável Sexo, podemos perceber pela Figura 6 e Figura 7 que a maioria dos candidatos do ENEM 2019 são do Sexo Feminino com 1.823.376 candidatas representando cerca de 59,06% da base de dados respectivamente, e do sexo Masculino com 1.263.859 candidatos representando cerca de 40,94% da base de dados respectivamente nas figuras apresentadas a seguir.

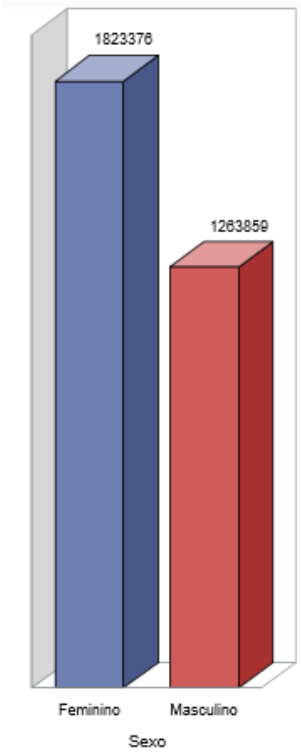

Figura 6 – Gráfico de barras da frequência dos candidatos do ENEM 2019 por Sexo.

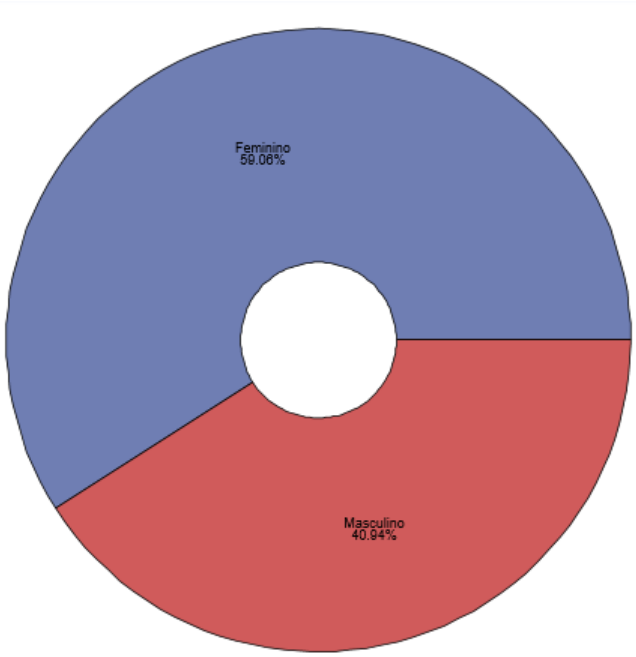

Figura 7 – Gráfico de Donut do percentual da frequência dos candidatos do ENEM 2019 por Sexo.

Vemos na variável COR_RACA os candidatos auto declarados de cor e raça Parda sendo maioria nessa edição com 1.438.579 participantes, candidatos autodeclarados de cor e raça Branca vem em seguida entre os participantes com 1.100.997 participantes. Observamos ainda que candidatos autodeclarados de cor e raça Indígena ainda são menoria no ENEM, com 17.929 participantes dos candidatos dessa amostra de dados, como demonstrado na Figura 8 a seguir.

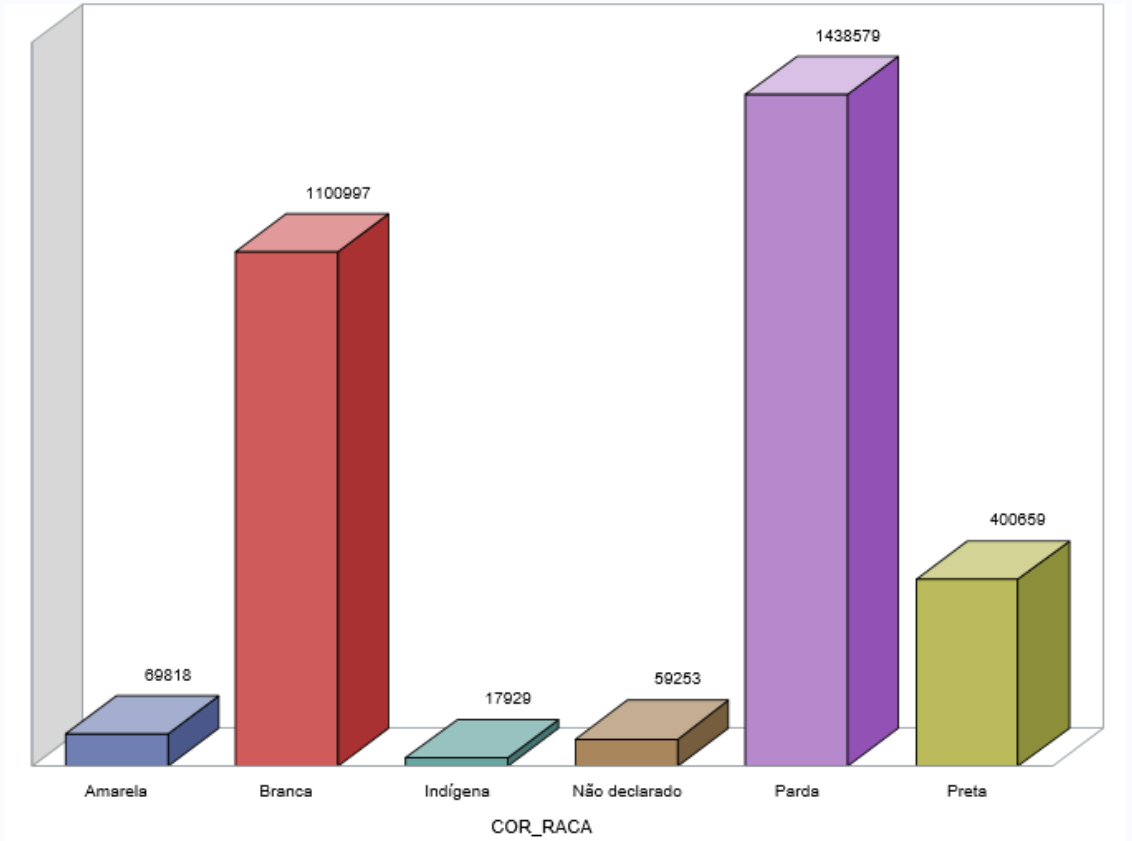

Figura 8 – Frequência dos candidatos do ENEM 2019 por Cor e Raça.

Com a Figura 9 onde mostra **a variável Tipo_Escola**, vemos que a maioria dos candidatos não reponderam qual tipo de escola pertecem/pertenceram, sendo 1.911.499 candidatos desta amostra de dados. Ainda percebe-se que cerca de 969.625 candidatos são de escola pública, seguido de 206.111 candidatos de escola privada.

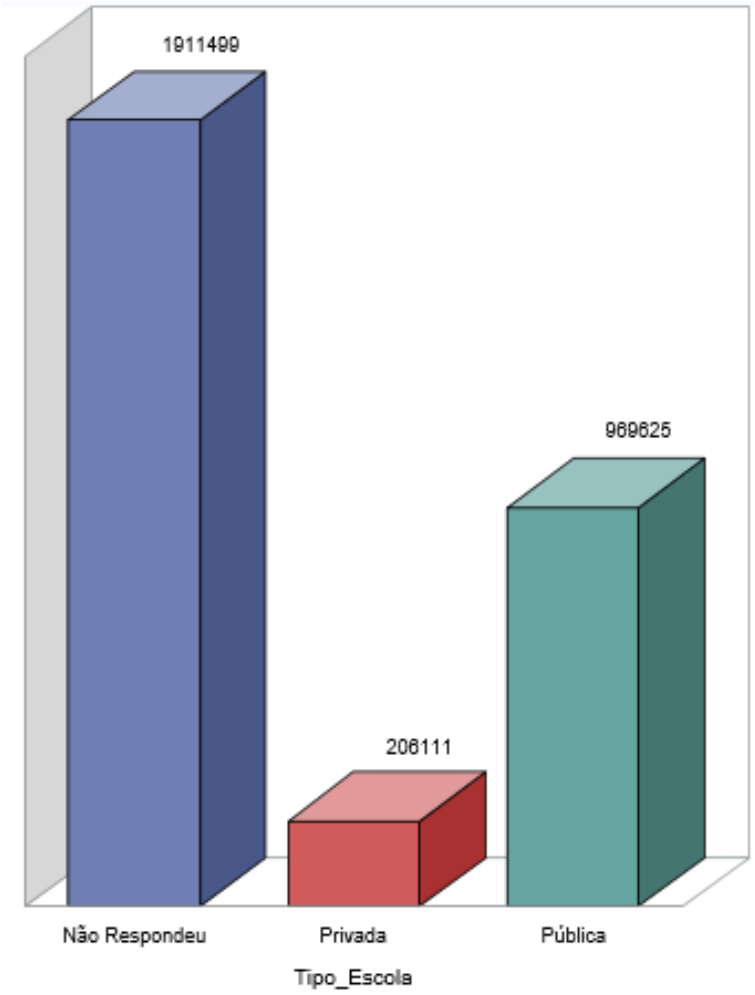

Figura 9 – Frequência dos candidatos do ENEM 2019 por Tipo de Escola.

Para a **variável Região_Nome**, vemos na Figura 9 que a região Sudeste possui a maior quantidade de inscritos no ENEM com 1.085.058 candidatos seguido de candidatos da região Nordeste com 1.077.635 candidatos, vemos ainda que a região Centro-Oeste possui o menor quantitativo de candidatos neste edição com 242.268 candidatos.

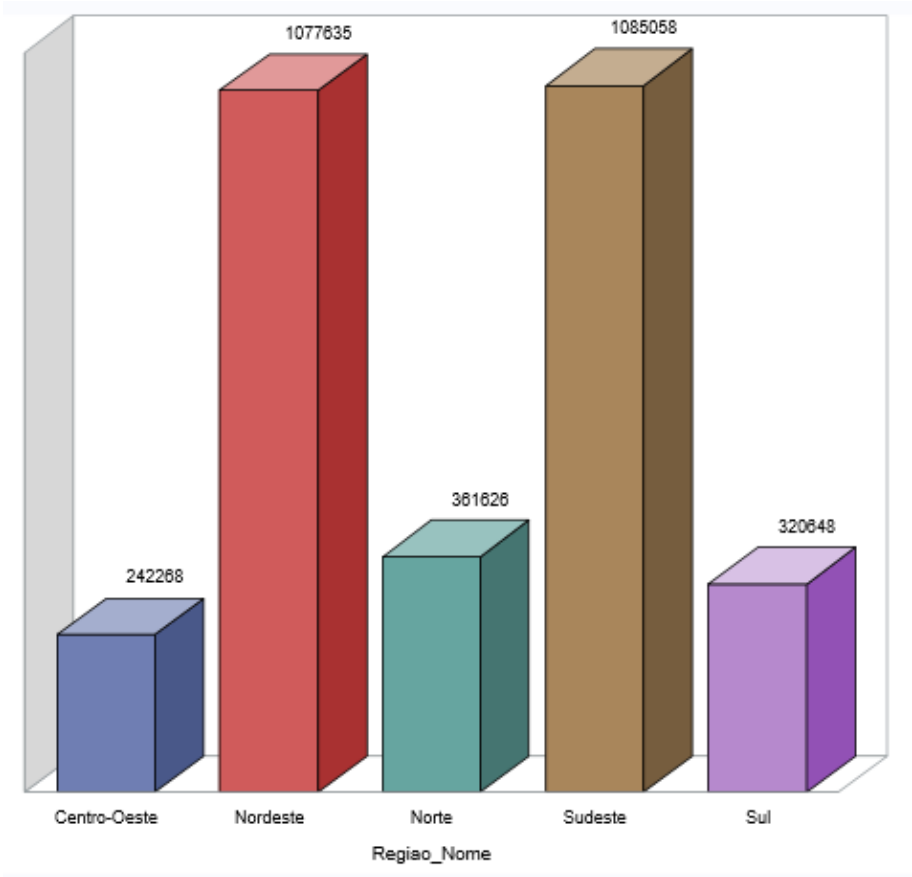

Figura 10 – Frequência dos candidatos do ENEM 2019 por Região no Brasil.

Já na Figura 11 apresentasse a **variável SG_UF_RESIDENCIA** e igualmente informado na base de dados a **variável UF_Nome** o estado de SP é o que apresenta a maior quantidade de inscritos no ENEM 2019, com mais de 490 mil candidatos, seguido do estado de MG com mais de 325 mil candidatos e ainda o estado de BA com mais de 242 mil candidatos, representando estes 3 estados como os maiores em quantidade de candidatos inscritos nesta edição do ENEM, como também estados onde suas capitais são as mais populosas do país, segundo pesquisas e projeções do IBGE. Podemos ver ainda que os estado do AC, AP, RR e TO são os menores em quantidade de candidatos participantes do ENEM 2019 por estados.

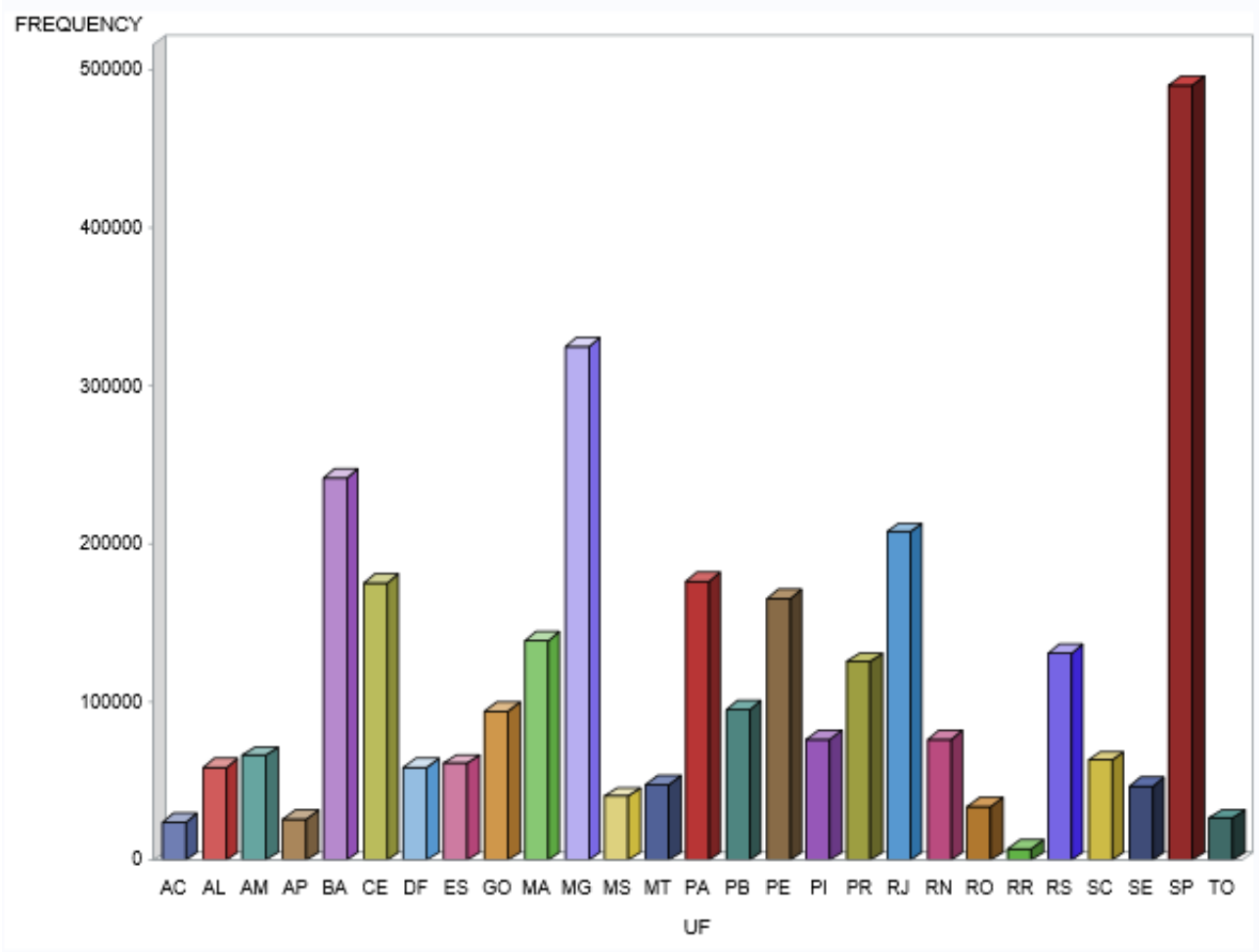

Figura 11 – Frequência dos candidatos do ENEM 2019 por Unidade da Federação (UF) no Brasil.

A partir da tabela da Figura 12, verifica-se as estatísticas básicas de resumo das notas do ENEM 2019 segundo informações contidas na amostra de dados fornecida. Percebe-se que todas as notas não possuem candidatos que zeraram suas notas, ou seja, por ser uma amostra, provavelmente não pegou a parcela de candidatos que obtiveram 0 (zero) ou não realizaram a prova, que por consequência iriam zerar algum nota nesta edição do ENEM.

A nota mais baixa de **Ciências Naturais** no ENEM 2019 foi de 327,90 pontos, a maior nota foi de 860,90 pontos, a média foi de 478,54 pontos. Já a mediana foi de 470,70 pontos em Ciências Naturais. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 470,70 pontos e 50% das notas foram maiores que 470,70 pontos, vemos ainda a moda com 410,70 pontos, e ainda verifica-se que 90% da base de dados, os candidatos obtiveram notas menores ou iguais a 583 pontos, sendo os outros 10% candidatos que obtiveram notas superiores a 583 pontos.

Variable	Mean	Minimum	Maximum	Mode	N	N Miss	Lower Quartile	Median	Upper Quartile	90th Pctl
NU_NOTA_CN	478.5455301	327.9000000	860.9000000	410.7000000	3087235	0	418.7000000	470.7000000	532.9000000	583.0000000
NU_NOTA_CH	511.7636389	315.9000000	835.1000000	541.8000000	3087235	0	453.1000000	514.7000000	569.4000000	613.7000000
NU_NOTA_LC	523.8292513	322.0000000	801.7000000	532.9000000	3087235	0	487.2000000	528.3000000	566.9000000	598.4000000
NU_NOTA_MT	522.8187069	359.0000000	985.5000000	446.4000000	3087235	0	434.7000000	500.0000000	596.8000000	680.1000000
NU_NOTA_REDACAO	596.2201582	120.0000000	1000.00	600.0000000	3087235	0	500.0000000	580.0000000	680.0000000	820.0000000
NOTA_MEDIA	526.6354571	307.2800000	850.8200000	488.2000000	3087235	0	467.6200000	516.1000000	576.9400000	640.0000000

Figura 12 – Estatística de Resumo das notas dos candidatos do ENEM 2019 no Brasil.

A nota mais baixa de **Ciências Humanas** no ENEM 2019 foi de 315,90 pontos, a maior nota foi de 835,10 pontos, a média foi de 511,76 pontos. Já a mediana foi de 514,70 pontos em Ciências Humanas. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 514,70 pontos e 50% das notas foram maiores que 514,70 pontos, vemos ainda a moda com 541,80 pontos, e ainda pode-se perceber que 75% da base de dados ou também conhecido como 3º Quartil, os candidatos obtiveram notas menores ou iguais a 569,40 pontos, sendo os outros 25% candidatos que obtiveram notas superiores a 569,40 pontos.

A nota mais baixa de **Linguages e Códigos** no ENEM 2019 foi de 322 pontos, a maior nota foi de 801,70 pontos, a média foi de 523,82 pontos. Já a mediana foi de 528,30 pontos em Linguages e Códigos. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 528,30 pontos e 50% das notas foram maiores que 528,30 pontos, vemos ainda a moda com 532,90 pontos, e ainda pode-se perceber que 75% da base de dados ou também conhecido como 3º Quartil, os candidatos obtiveram notas menores ou iguais a 566,90 pontos, sendo os outros 25% candidatos que obtiveram notas superiores a 566,90 pontos.

A nota mais baixa de **Matemática** no ENEM 2019 foi de 359 pontos, a maior nota foi de 985,50 pontos, a média foi de 522,81 pontos. Já a mediana foi de 500 pontos em Linguages e Códigos. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 500 pontos e 50% das notas foram maiores que 500 pontos, vemos ainda a moda com 446,40 pontos, e ainda pode-se perceber que 25% da base de dados ou também conhecido como 1º Quartil, os candidatos obtiveram notas menores ou iguais a 434,70 pontos, sendo os outros 75% candidatos que obtiveram notas superiores a 434,70 pontos.

A nota mais baixa de **Redação** no ENEM 2019 foi de 120 pontos, a maior nota foi de 1.000 pontos, a média foi de 596,22 pontos. Já a mediana foi de 580 pontos em Redação. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 580 pontos e 50% das notas foram maiores que 580 pontos, vemos ainda a moda com 600 pontos, e ainda pode-se perceber que 75% da base de dados ou também conhecido como

3º Quartil, os candidatos obtiveram notas menores ou iguais a 680 pontos, sendo os outros 25% candidatos que obtiveram notas superiores a 680 pontos.

A **Nota Média** mais baixa no ENEM 2019 foi de 307,28 pontos, a maior nota foi de 850,82 pontos, a média foi de 525,63 pontos. Já a mediana da Nota Média foi de 515,10 pontos. Por ser a mediana reflete a nota central de todas as notas ou distribuição das notas, ou seja, é a nota onde 50% das notas obtidas no ENEM foram menores que 515,10 pontos e 50% das notas foram maiores que 515,10 pontos, vemos ainda a moda com 488,20 pontos, e ainda pode-se perceber que 75% da base de dados ou também conhecido como 3º Quartil, os candidatos obtiveram notas menores ou iguais a 576,94 pontos, sendo os outros 25% candidatos que obtiveram notas superiores a 576,94 pontos.

3 Análise de Cluster

Com base nas definições sobre Análise de Cluster, precisamos classificar os 5.570 municípios do Brasil em no máximo 10 clusters. Para tal devemos utilizar um conjunto de variáveis do ENEM 2019. Utilizando as 5 notas dos candidatos do ENEM 2019 identifique e calcule um conjunto de estatísticas descritivas e as utilize como variáveis para determinação dos Cluster. A partir deste ponto, é definido algumas estatísticas de resumo das notas do ENEM 2019 por município como demonstrado na Figura 13, como parte da preparação dos dados para utilização do SAS Miner como variáveis input de definição do cluster por município.

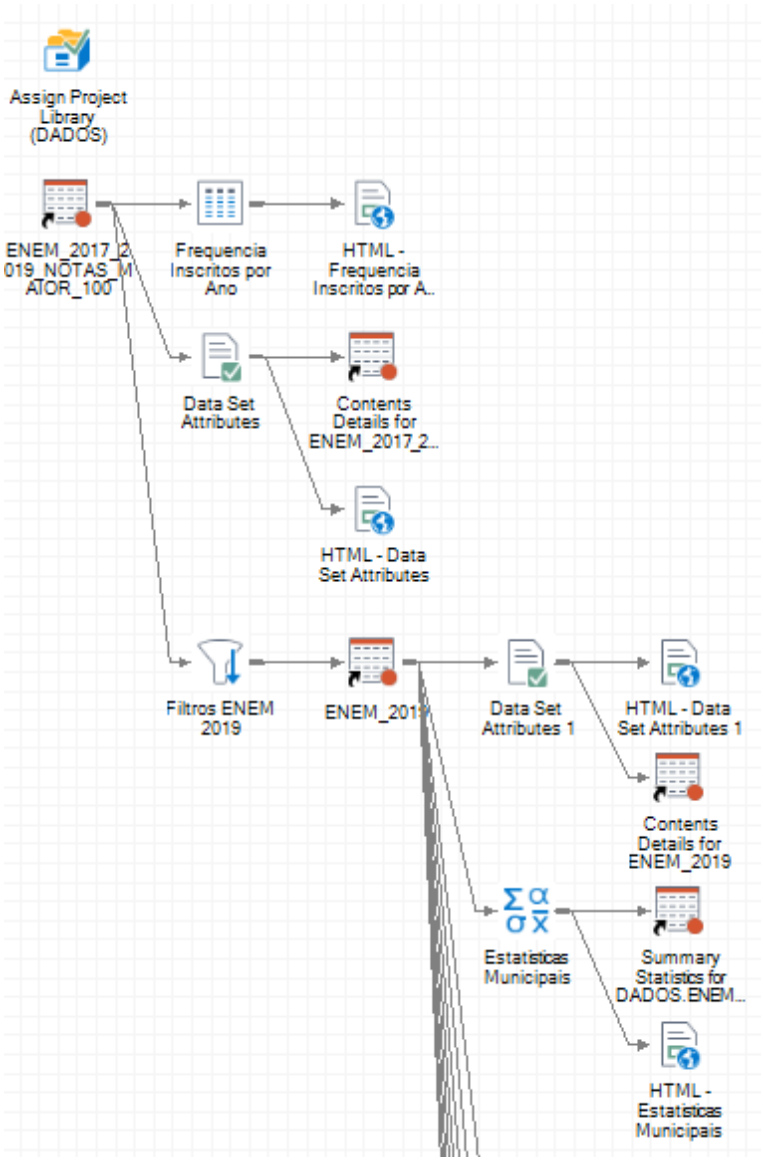

Figura 13 – Fluxo de geração de Estatística de Resumo das notas dos candidatos do ENEM 2019 no Brasil por UF.

3.1 Nó de Cluster

Na ferramenta do SAS Enterprise Miner, dispõe de nós/blocos de tarefas, em que por meio deste, realizam determinadas tarefas conforme a necessidade do projeto. A seguir na Figura 14, mostra o fluxo resumido das tarefas realizadas no Miner.

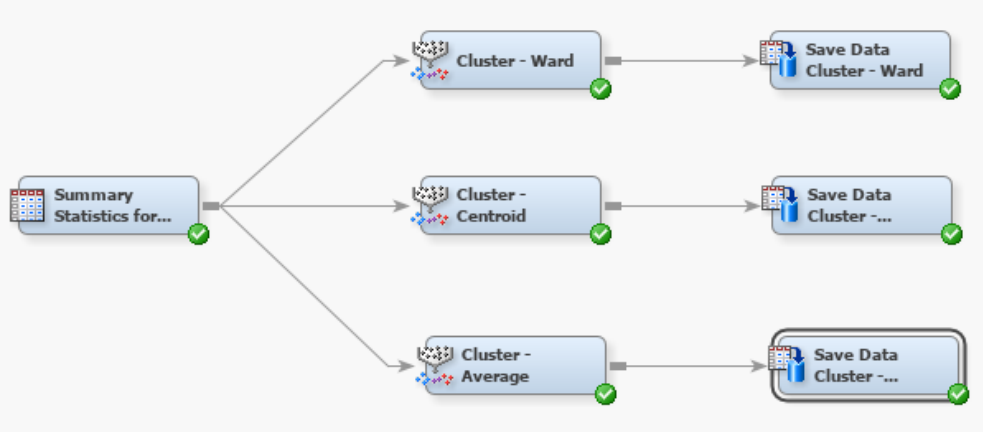

Figura 14 – Fluxo de geração do Cluster pelo SAS Enterprise Miner.

O **nó Cluster** é utilizado para executar clustering de observação, que pode ser usado para segmentar bancos de dados. Clusteringcoloca objetos em grupos ou clusters sugeridos pelos dados. Os objetos em cada cluster tendem a ser semelhantes auns aos outros em certo sentido, e os objetos em grupos diferentes tendem a ser diferentes. Se clusters óbvios ouagrupamentos podem ser desenvolvidos antes da análise, então a análise de agrupamento pode ser realizada simplesmenteclassificar os dados.

Os métodos de agrupamento do nó Cluster realizam análise de agrupamento disjunta com base na análise da distância euclidiana calculadas a partir de uma ou mais variáveis quantitativas e sementes que são geradas e atualizadas pelo algoritmo. Você pode especificar o critério de agrupamento que é usado para medir a distância entre os dados observações e sementes.

O **Método de Clustering** se selecionado o automático como sua propriedade Método de Especificação, Método de Clustering especifica como o SAS Enterprise Miner calcula as distâncias de cluster, sendo possível 3 métodos, estes serão descritos e apresentados resultados da análise de cluster por cada um destes métodos. O resultado fornecido pelo nó de cluster, pode ser observado de várias formas, por gráficos, tabelas, visuais em que permitem ter a verdadeira noção do cluster formado e seu componente, sendo assim, a seguir é demonstrado os métodos, como os resultados em vários através de gráficos e ainda um breve introdução do que é o gráfico.

Lembrando, que para cada uma das implementações de métodos de clustering, definiu-se as variáveis NU_NOTA_MT_Mean, NU_NOTA_CN_Mean e NU_NOTA_REDACAO_Mean para realizar a análise dos clusters gerados por cada método a partir destas variáveis.

3.2 Ward

O método Ward é o método padrão do Miner, sendo a distância entre dois clusters, é a ANOVA soma dos quadrados entre os dois clusters a soma de todas as variáveis. Em cada geração, a soma dos quadrados dentro do cluster é minimizado em todas as partições obtidas pela fusão de dois clusters de uma geração anterior.

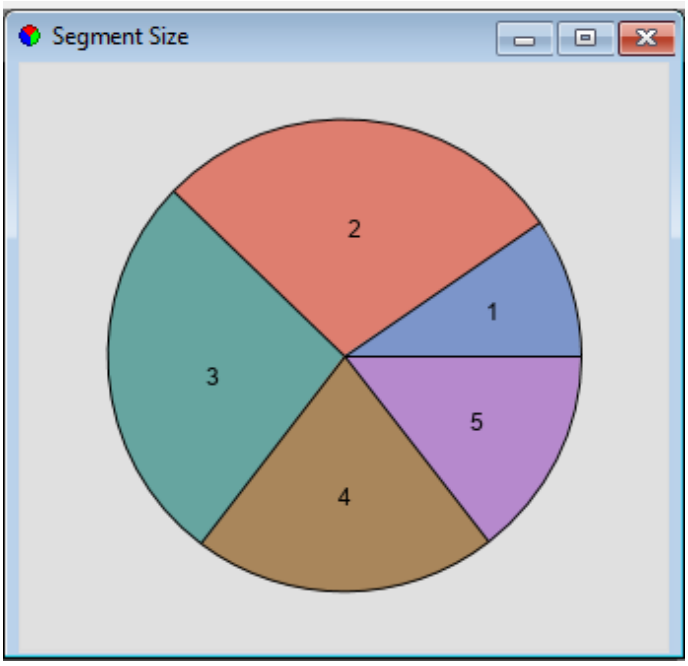

Figura 15 – Segment Size pelo método Ward.

Segment Id	Frequency of Cluster ▼	NU_NOTA_MT_Mean	NU_NOTA_CN_Mean	NU_NOTA_REDACAO_Mean	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster
2	1577	511.3899	467.4669	576.3661	0.466142	2.361727	3
3	1511	484.6785	448.9306	560.3739	0.418392	2.377151	2
4	1156	536.8438	485.0934	602.2024	0.45557	3.785787	2
5	806	465.5456	435.2347	518.1064	0.475271	3.283443	3
1	520	566.1578	503.5491	637.7033	0.53029	4.818568	4

Figura 16 – Mean Statistics pelo método Ward.

Por meio da Figura 15, verifica-se que o método de Ward gerou 5 clusters, onde que já identificamos por este, que o cluster identificado pelo segmento 2, 3 e 4 são os maiores para este método. Dentre estes três segmentos, é possível perceber cuidadosamente o cluster de segmento 2 como o maior dentre eles, como pode-se verificar na Figura 16 o cluster de segmento 2 com o maior número de frequência de dados sendo 1.577 valores.

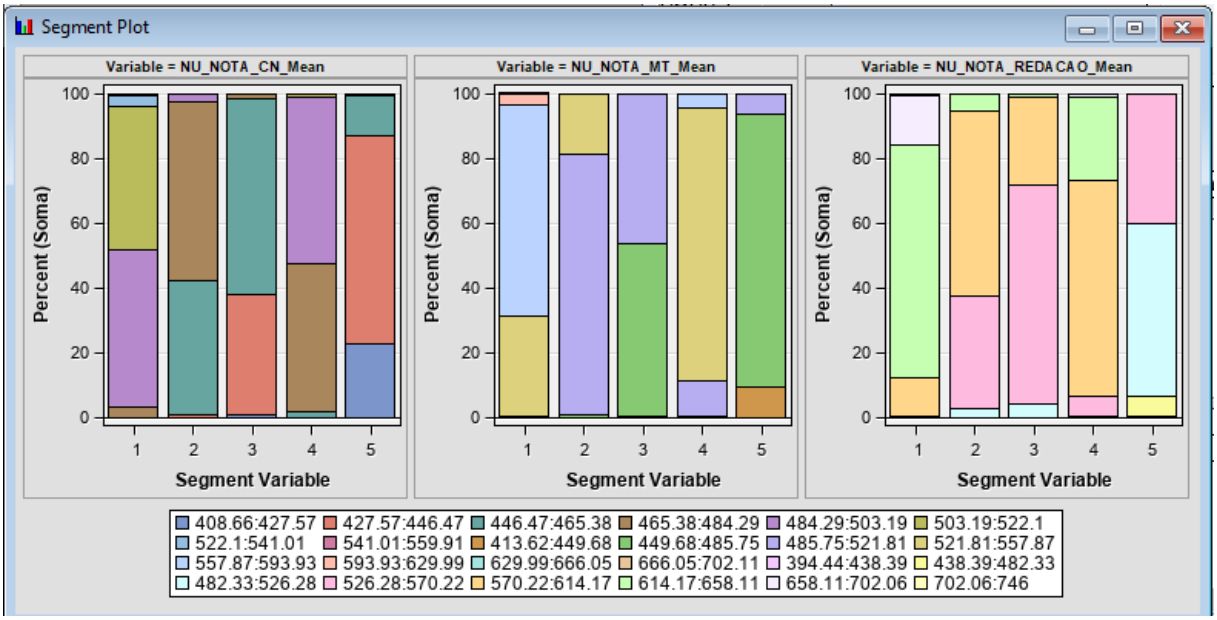

Figura 17 – Segment Plot pelo método Ward.

Já com o gráfico demonstrado na Figura 17, vemos que para a nota de Ciência da Natureza, o cluster de segmento 2, possui 55.16% das notas entre (465.35;484.29), e 41.66% das notas entre (446.47;465.38). Já para a nota de Matemática, o cluster 2, possui 80.21% de notas entre (485.75;521.81) e 18.70% de notas entre (521.81;557.87) pontos. Vemos ainda que para a nota de Redação, o cluster de segmento 2 possui 57.13% das notas entre (570.22;614.17) e 35.06% de notas entre (526.28;570.22).

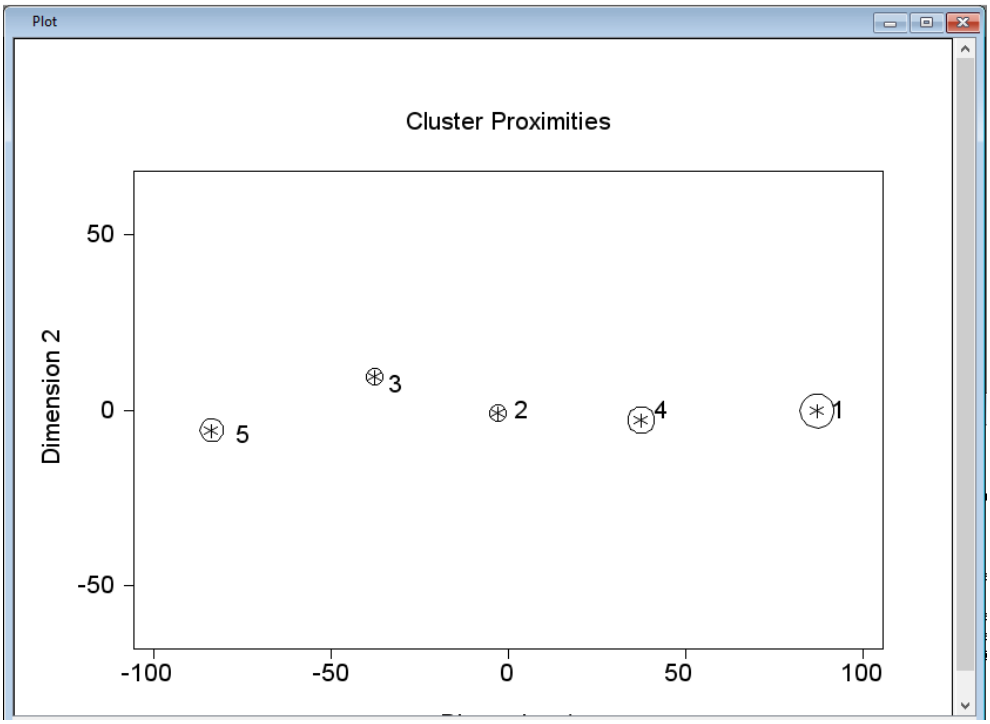

Figura 18 – Cluster Distance pelo método Ward.

Pelo gráfico apresentado na Figura 18, percebe-se que os cluster de segmento 1 e 5, são os mais afastados dos demais, mas também é possível identificar a semelhança de informação com o gráfico da Figura 15 que os maiores cluster 2, 3 e 4 também possuem uma proximidade entre si.

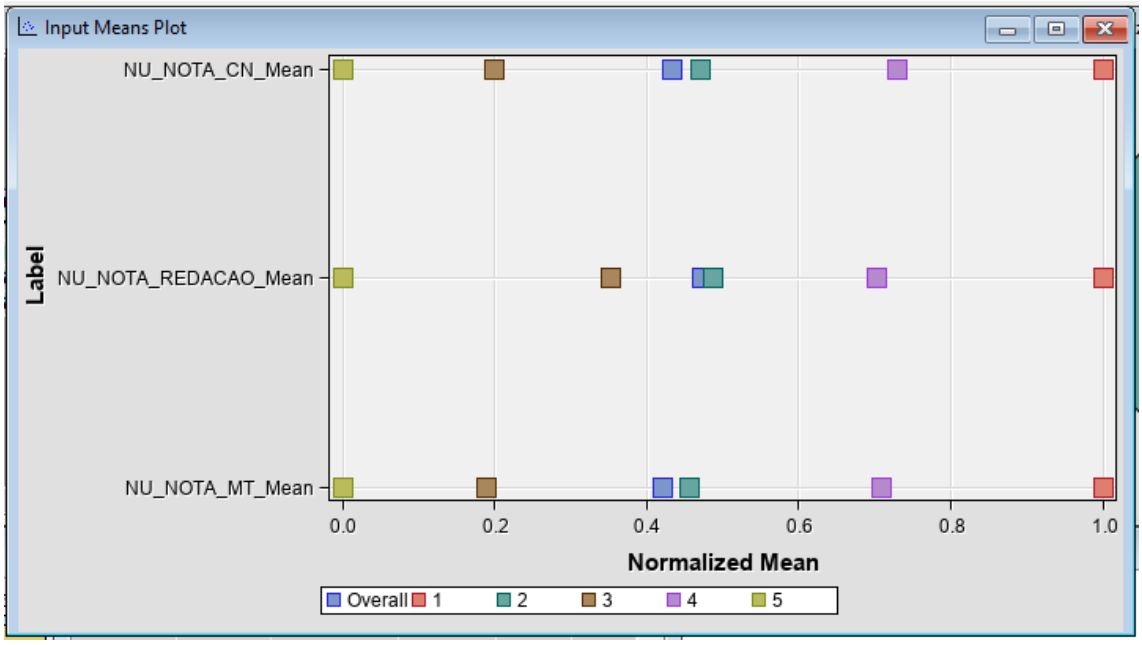

Figura 19 – Input Means pelo método Ward.

Por fim no gráfico demonstrado na figura 19, verifica-se que os clusters 1 e 5 confirmar o disposto na Figura 18, onde estes são os mais distantes entre os demais, pois sere, os clusters de valores da extremidade, ou seja, o cluster 5 possui as menores notas entre Matemática, Ciências da Natureza e Redação, já o cluster 1 possui as maiores notas de Matemática, Ciências da Natureza e Redação.

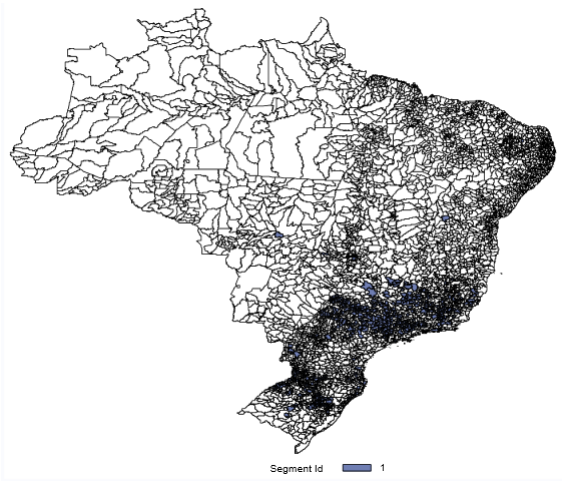

Figura 20 – Melhores notas por estado - Cluster 1.

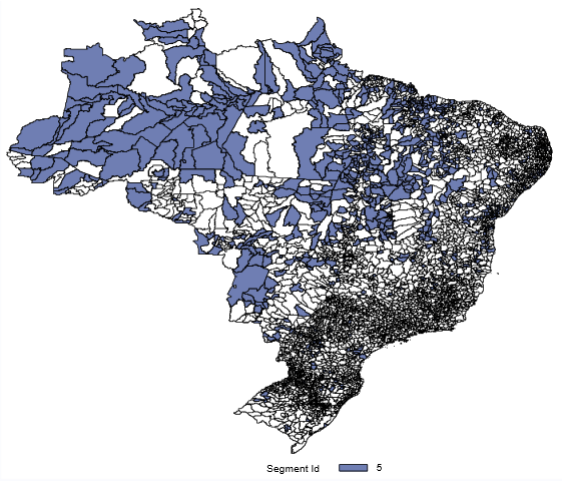

Figura 21 – Piores notas por estado - Cluster 5.

Analisando a Figura 20, vemos que as Regiões Sudeste e Sul do Brasil obtiveram as melhores notas registradas nesta edição do ENEM 2019, como vemos ainda, o mapa do Brasil está filtrado pelo Cluster 1, que referência a informação passada na Figura 19, onde este Cluster são das melhores notas obtidas em Matemática, Ciências da Natureza e Redação.

Já na Figura 22, onde mostrado o estado de São Paula (SP) e seus Clusters distribuídos nos município do estado. Percebe-se que o estado de São Paulo, faz parte da Região Sudeste, conforme o ocorrido no mapa do Brasil da Figura 20, onde demonstra as melhores notas obtidas em Matemática, Ciências da Natureza e Redação no estado, mas verificamos ainda que no estado de São Paulo, as melhores notas foram obtidas entre o Centro e Noroeste do estado, mostrando que nem sempre as melhores notas serem somente na capital do estado.

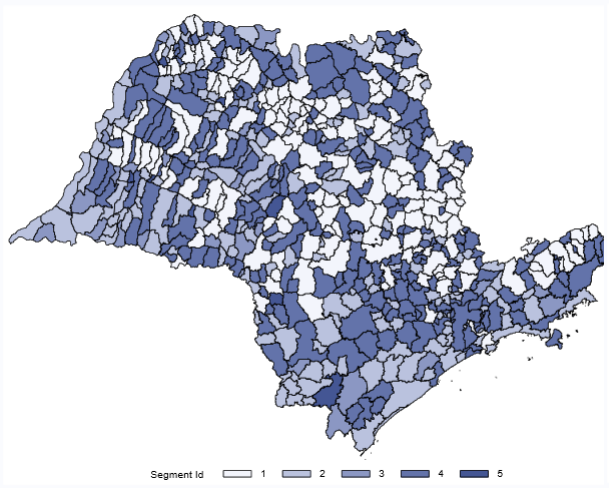

Figura 22 – Clusters do estado de SP.

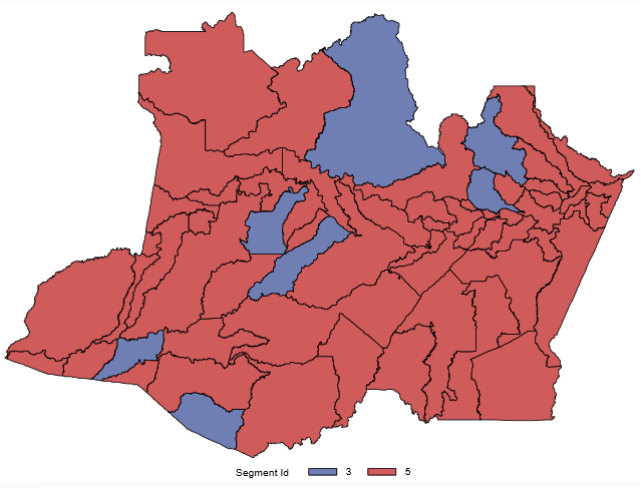

Figura 23 – Clusters do estado de AM.

Analisando a Figura 21, percebe-se que este se refere ao Cluster 5, onde demonstrado na Figura 19 este Cluster é composto pelas piores notas obtidas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que as piores notas foram obtidas na Região Norte, por ser a região d e menor concentração da populacional, e falta de recursos em cidades dos estados desta região, tem grande impacto nas notas obtidas por estado do Brasil, como pode-se perceber.

Ainda, podemos analisar a Figura 23, onde mostra o estado do Amazonas (AM), este é um dos estados que compõe a Região Norte do país, que por sua vez faz parte dos estados que obtiveram as piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que no estado do Amazonas, possui apenas dois Clusters, o Cluster 3 e o Cluster 5, onde estes são como demonstrado na Figura 19, são compostos pelas piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação.

3.3 Centróide

Pelo método do Centróide, a distância entre dois clusters é definida como a distância euclidiana (quadrada) entre seus centróides ou meios.

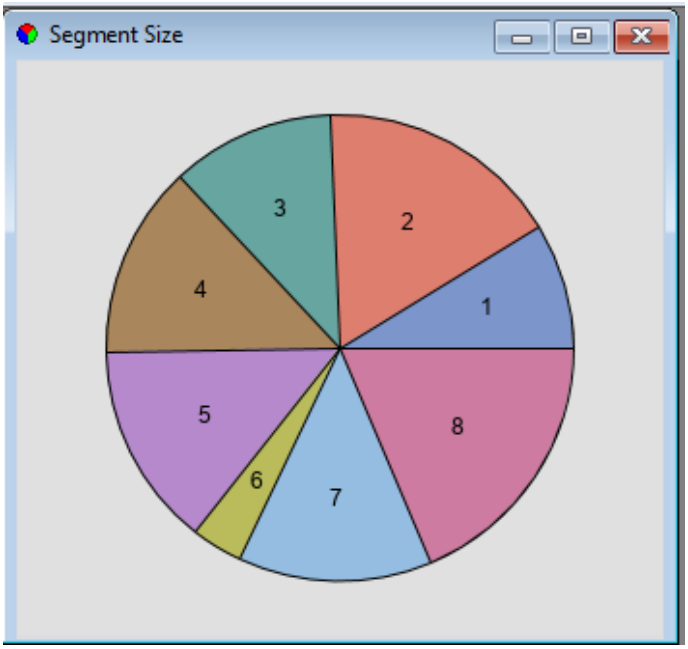

Figura 24 – Segment Size pelo método Centróid.

Segment Id	Frequency of Cluster	NU_NOTA_MT_Mean	NU_NOTA_CN_Mean	NU_NOTA_REDACAO_Mean	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster
8	1041	479.4982	445.2602	565.0157	0.344341	2.179038	4
2	938	506.5469	462.9827	597.2364	0.386949	2.30906	3
5	792	537.3939	485.6505	599.3974	0.36524	2.367637	1
4	740	498.922	459.4605	543.1651	0.374046	2.098645	8
7	738	464.5447	434.3319	517.1346	0.453647	3.260508	8
3	632	523.3257	477.1705	564.2286	0.365826	3.431136	5
1	485	552.6703	494.9385	630.4993	0.425593	2.31802	5
6	204	579.8842	511.5626	646.6457	0.516139	4.224391	1

Figura 25 – Mean Statistics pelo método Centróide.

Por meio da Figura 24, verifica-se que o método de Centroid gerou 8 clusters, onde que já identificamos por este, que o cluster identificado pelo segmento 8, 2 e 5 são os maiores para este método. Dentre estes três segmentos, é possível perceber cuidadosamente o cluster de segmento 8 como o maior dentre eles, como pode-se verificar na Figura 25 o cluster de segmento 8 com o maior número de frequência de dados sendo 1.041 valores.

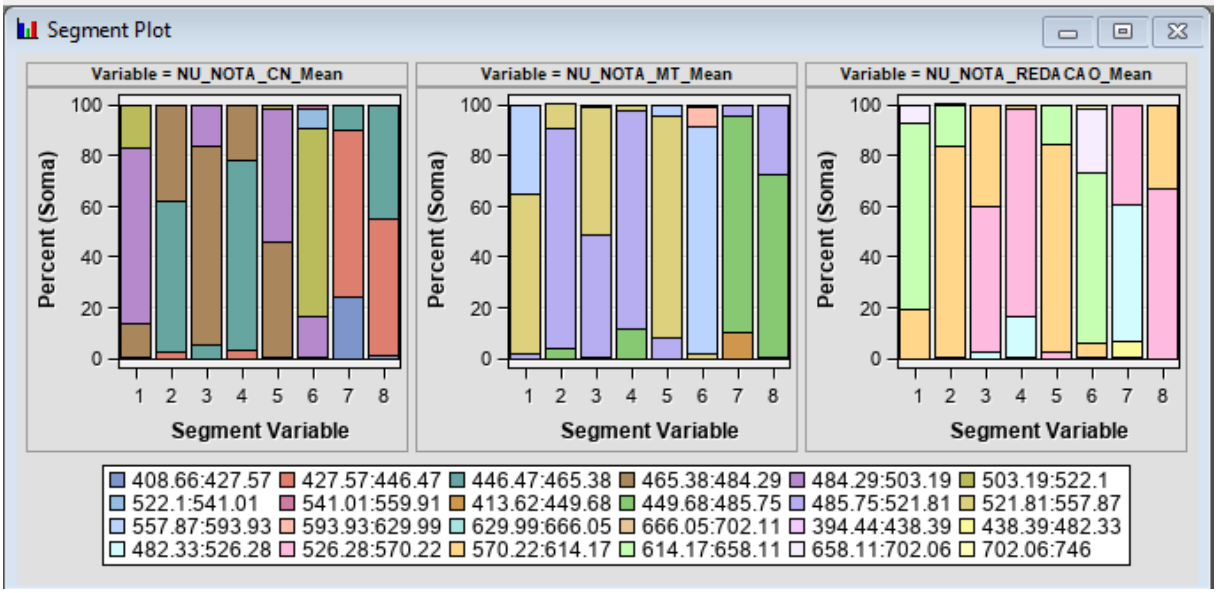

Figura 26 – Segment Plot pelo método Centróid.

Já com o gráfico demonstrado na Figura 26, vemos que para a nota de Ciência da Natureza, o cluster de segmento 8, possui 53.50% das notas entre (427.57;446.47), e 44.95% das notas entre (446.47;465.38). Já para a nota de Matemática, o cluster 8, possui 72.14% de notas entre (449.68;485.75) e 27.56% de notas entre (485.75;521.81) pontos. Vemos ainda que para a nota de Redação, o cluster de segmento 8 possui 66.47% das notas entre (526.28;570.22) e 32.94% de notas entre (570.22;614.17).

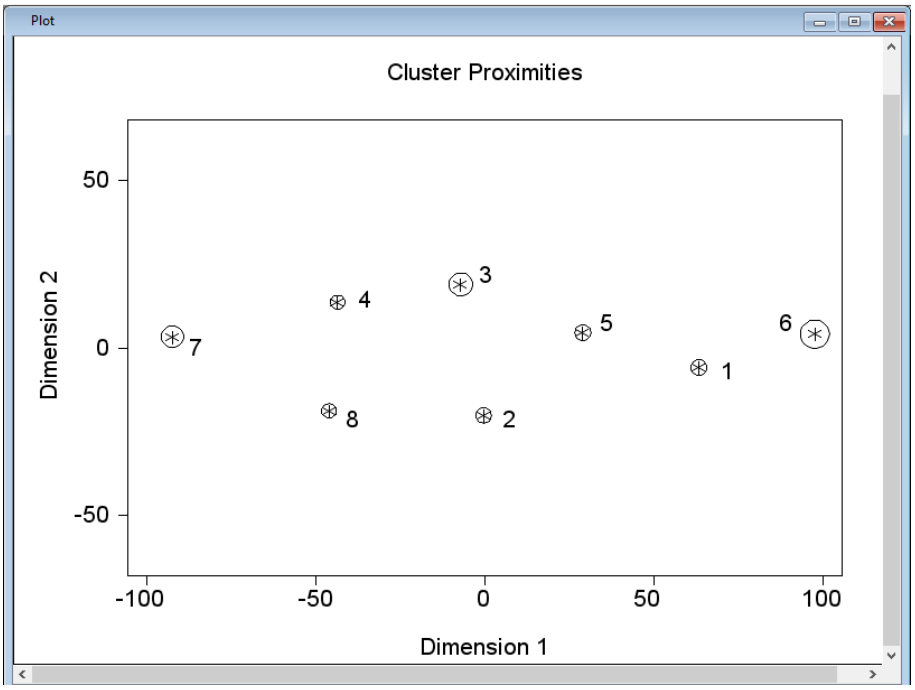

Figura 27 – Cluster Distance pelo método Centróid.

Pelo gráfico apresentado na Figura 27, percebe-se que os cluster de segmento 6 e 7, são os mais afastados dos demais, mas também é possível identificar a semelhança de informação com o gráfico da Figura 24 que os maiores cluster 8, 2 e 5 também possuem uma proximidade entre si.

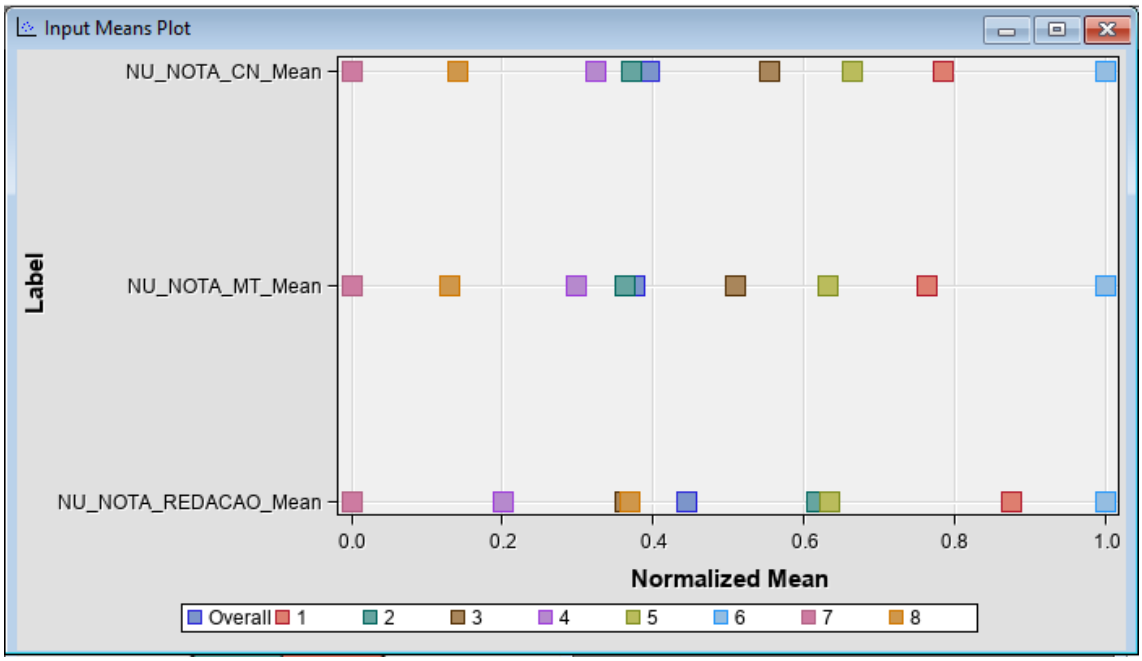

Figura 28 – Input Means pelo método Centróid.

Por fim no gráfico demonstrado na figura 28, verifica-se que os clusters 6 e 7 confirmar o disposto na Figura 27, onde estes são os mais distantes entre os demais, pois sere, os clusters de valores da extremidade, ou seja, o cluster 7 possui as menores notas entre Matemática, Ciências da Natureza e Redação, já o cluster 6 possui as maiores notas de Matemática, Ciências da Natureza e Redação.

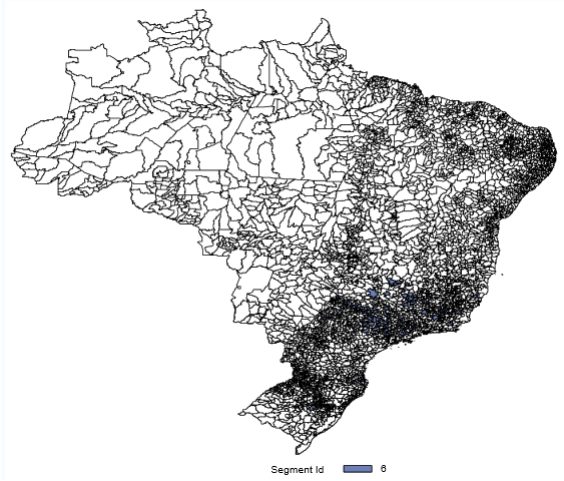

Figura 29 – Melhores notas por estado - Cluster 6.

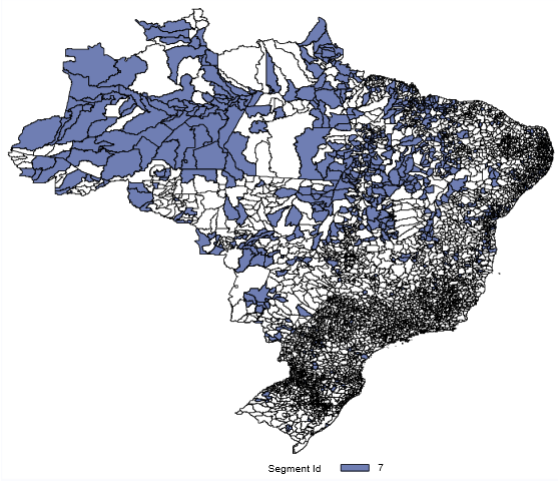

Figura 30 – Piores notas por estado - Cluster 7.

Analisando a Figura 29, vemos que as Regiões Sudeste e Sul do Brasil obtiveram as melhores notas registradas nesta edição do ENEM 2019, como vemos ainda, o mapa do Brasil está filtrado pelo Cluster 6, que referência a informação passada na Figura 28, onde este Cluster 6 corresponde as melhores notas obtidas em Matemática, Ciências da Natureza e Redação.

Já na Figura 31, onde é demonstrado o estado do Rio de Janeiro (RJ) e seus Clusters distribuídos nos município do estado. Percebe-se que o estado do Rio de Janeiro, faz parte da Região Sudeste, conforme o ocorrido no mapa do Brasil da Figura 29, onde demonstra as melhores notas obtidas em Matemática, Ciências da Natureza e Redação no estado, mas verificamos ainda que no estado do Rio de Janeiro, as melhores notas, foram obtidas entre as Regiões Norte e Região dos Lagos do estado, mostrando que nem sempre as melhores notas seram somente na Região Metropolitana do estado.

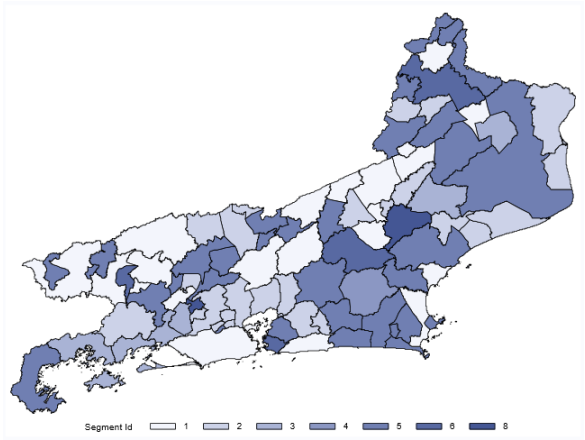

Figura 31 – Clusters do estado de RJ.

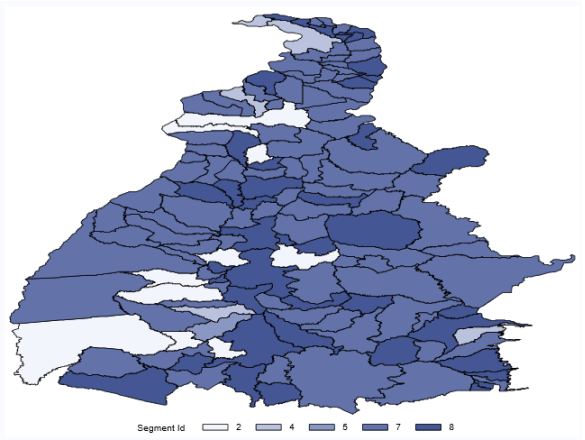

Figura 32 – Clusters do estado de TO.

Analisando a Figura 30, percebe-se que este se refere ao Cluster 7, onde demonstrado na Figura 28 este Cluster é composto pelas piores notas obtidas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que as piores notas foram obtidas na Região Norte do país, por ser a região de menor concentração da populacional, e falta de recursos em cidades dos estados desta região, tem grande impacto nas notas obtidas por estado do Brasil, como pode-se perceber.

Ainda, podemos analisar a Figura 32, onde mostra o estado do Tocantins (TO), este é um dos estados que compõe a Região Norte do país, que por sua vez faz parte dos estados que obtiveram as piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que no estado do Tocantins, possui Cluster 7 bem concentrado pelas Regiões Norte, Noroeste, Centro-Oeste, Central e Jalapão e Sudeste do estado, onde estes são como demonstrado na Figura 32, são compostos pelas piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação.

3.4 Average

Já no método de Average, a distância entre dois clusters é a distância média entre pares de observações, um em cada cluster.

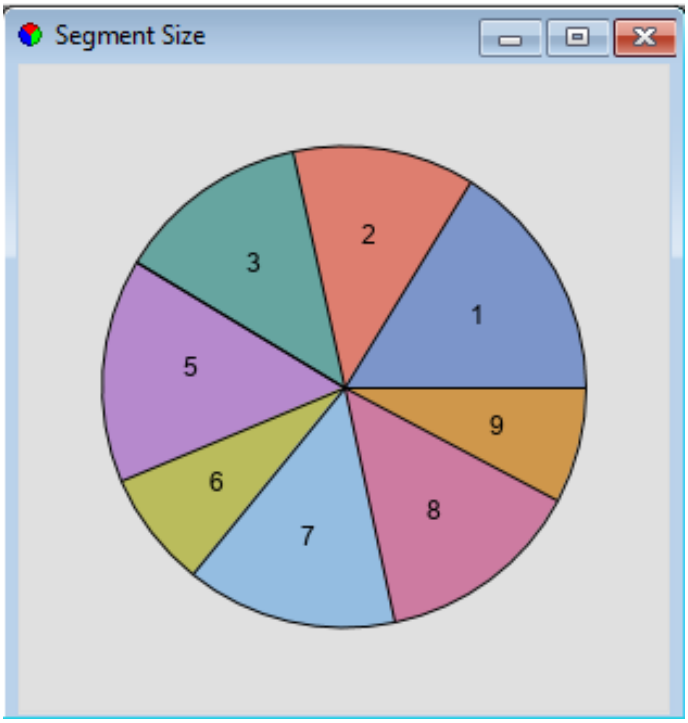

Figura 33 – Segment Size pelo método Average.

Segment Id	Frequency of Cluster	NU_NOTA_MT_Mean	NU_NOTA_CN_Mean	NU_NOTA_REDACAO_Mean	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster
1	894	508.0246	464.2139	595.8253	0.363764	1.897955	5
5	839	484.3794	448.3938	573.5128	0.331777	2.139154	7
7	796	475.1257	442.8622	539.3884	0.340383	1.600019	5
8	768	532.1656	483.0431	578.129	0.375746	3.456125	2
3	719	506.6283	464.7492	545.9832	0.382468	2.016748	5
2	687	540.258	486.9546	618.9693	0.401842	2.274474	8
9	439	460.4584	430.619	508.0315	0.452191	2.981815	7
6	427	569.7805	505.7497	637.6669	0.498858	3.205352	2
4	1	702.1125	547.8125	695	.	0	6

Figura 34 – Mean Statistics pelo método Average.

Por meio da Figura 33, verifica-se que o método de Centroid gerou 9 clusters, onde que já identificamos por este, que o cluster identificado pelo segmento 1, 5 e 7 são os maiores para este método. Dentre estes três segmentos, é possível perceber cuidadosamente o cluster de segmento 1 como o maior dentre eles, como pode-se verificar na Figura 34 o cluster de segmento 1 com o maior número de frequência de dados sendo 894 valores.

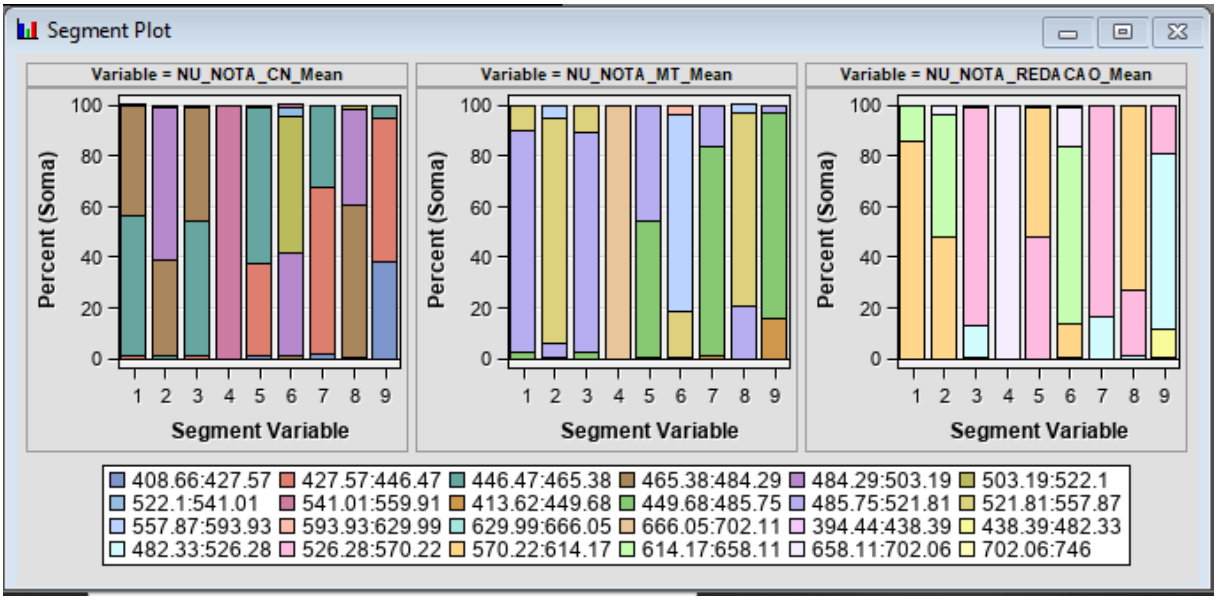

Figura 35 – Segment Plot pelo método Average.

Já com o gráfico demonstrado na Figura 35, vemos que para a nota de Ciência da Natureza, o cluster de segmento 1, possui 55.36% das notas entre (446.47;465.38), e 43.62% das notas entre (465.38;484.29). Já para a nota de Matemática, o cluster 1, possui 88.03% de notas entre (485.75;521.81) e 9.84% de notas entre (521.81;557.87) pontos. Vemos ainda que para a nota de Redação, o cluster de segmento 1 possui 85 34% das notas entre (570.22;614.17) e 14.65% de notas entre (614.17;658.11).

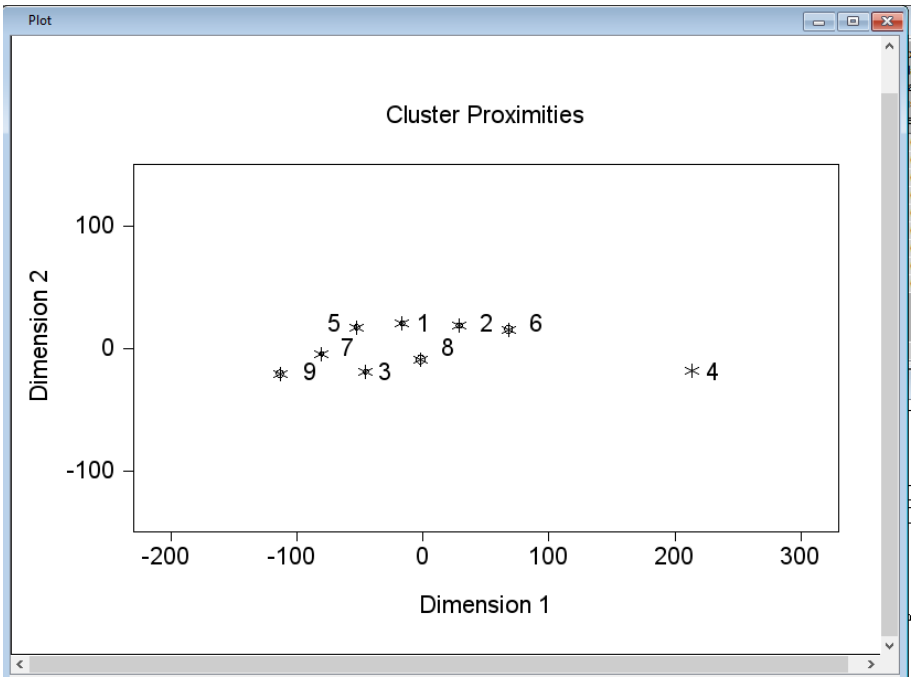

Figura 36 – Cluster Distance pelo método Average.

Pelo gráfico apresentado na Figura 36, percebe-se que o cluster de segmento 4, é o mais afastado dentre os demais, mas também é possível identificar a semelhança de informação com o gráfico da Figura 33 e 34 que o cluster 4 é o menor dentro todos, possuindo apenas 1 dados. Mas podemos ainda verificar que os clusters 5, 7, 9 e 3 como também os clusters 1, 8, 2 e 6 possuem uma proximidade entre si.

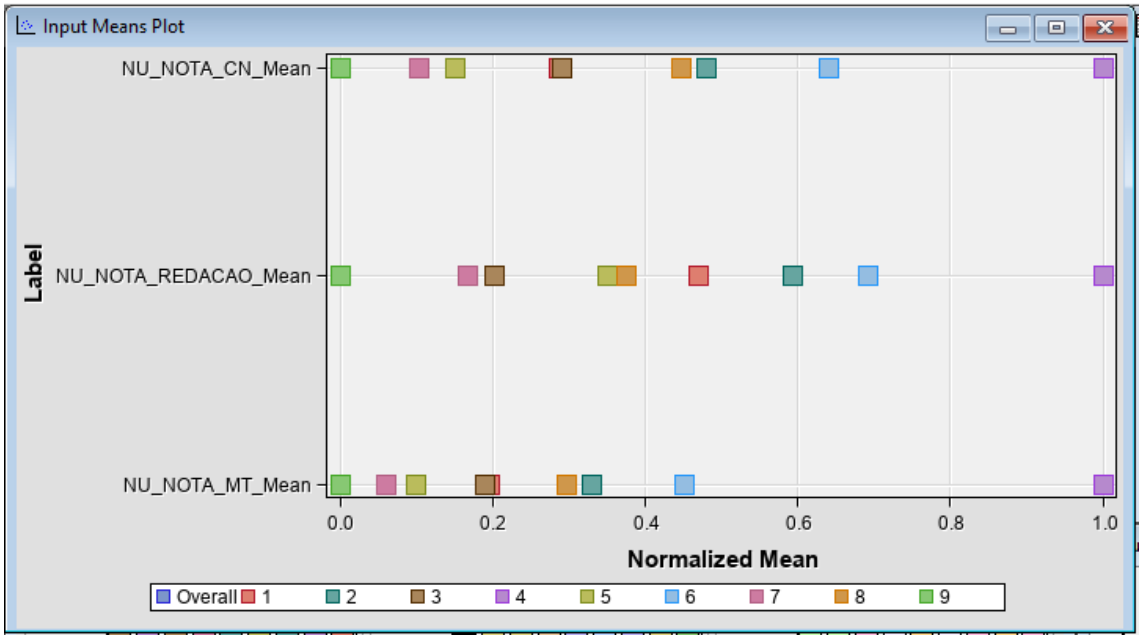

Figura 37 – Input Means pelo método Average.

Por fim no gráfico demonstrado na figura 37, verifica-se que o cluster 4 confirma o disposto na Figura 36, onde este é o mais distantes entre os demais, por ser, o cluster de valor da extremidade, ou seja, o cluster 4 possui as maiores notas de Matemática, Ciências da Natureza e Redação, já o cluster 9 as menores notas entre Matemática, Ciências da Natureza e Redação.

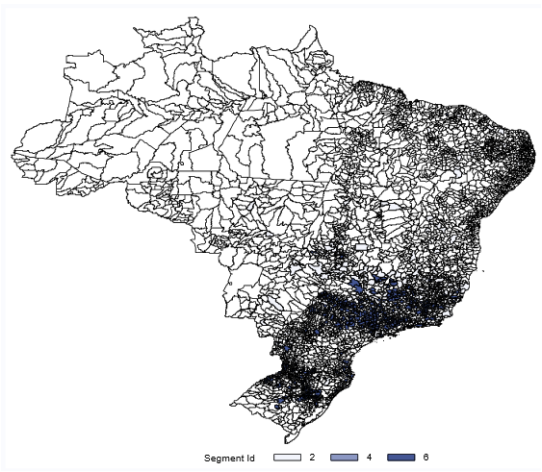

Figura 38 – Melhores notas por estado - Cluster 2, 4 e 6.

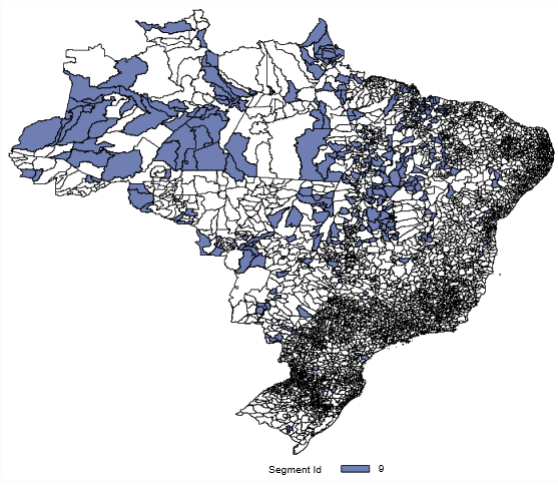

Figura 39 – Piores notas por estado - Cluster 9.

Analisando a Figura 38, vemos que as Regiões Sudeste e Sul do Brasil obtiveram as melhores notas registradas nesta edição do ENEM 2019, como vemos ainda, o mapa do Brasil está filtrado pelos Clusters 2, 4 e 6, que referência a informação passada na Figura 37, onde estes Clusters são os que correspondem as melhores notas obtidas em Matemática, Ciências da Natureza e Redação.

Já na Figura 40, onde é demonstrado o estado do Rio Grande do Sul (RS) e seus Clusters distribuídos nos município do estado. Percebe-se que o estado do Rio Grande do Sul, faz parte da Região Sul do país, conforme o ocorrido no mapa do Brasil da Figura 38, onde demonstra as melhores notas obtidas em Matemática, Ciências da Natureza e Redação no estado, mas verificamos ainda que no estado do Rio Grande do Sul, as melhores notas, foram obtidas entre as Regiões Noroeste, Sudeste, Sudoeste e Região Metropolitana do estado. Uma curiosidade que se mostra na Figura 40, mas pelas melhores notas compostas pelos Clusters 2, 4 e 6 estarem bem esparsas, ficou escondido num olhar rápido, onde o Cluster 4 possui apenas um Município do estado do Rio Grande do Sul com nota superior aos demais Clusters, que por sua vez, corresponde ao Cluster da maior nota obtida como mostrado na figura 37, este Município é o Westfália, encontrado por uma análise mais aprofundada pela ferramenta Guide.

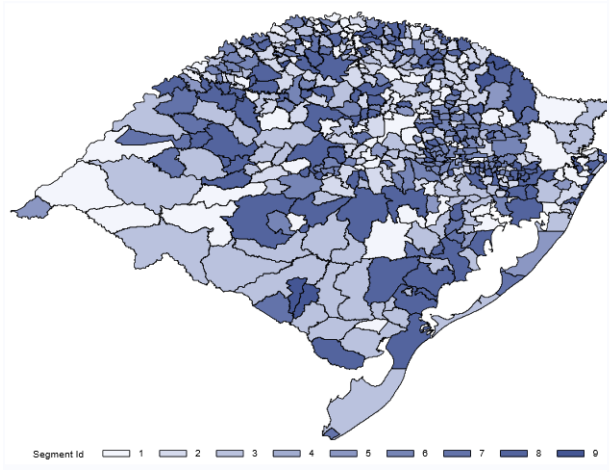

Figura 40 – Clusters do estado de RS.

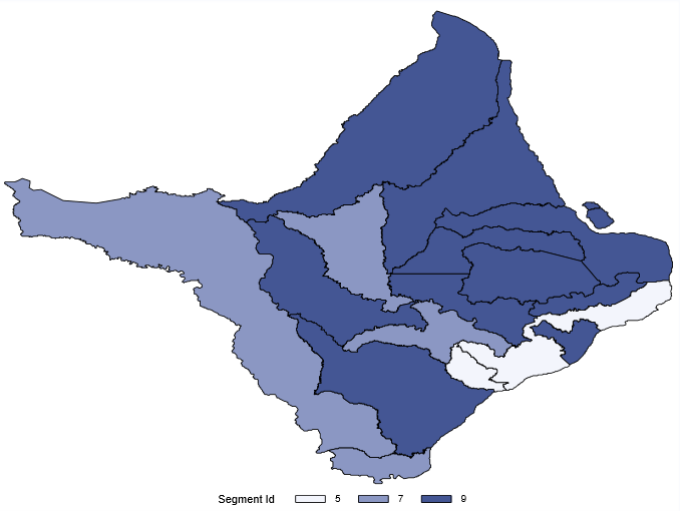

Figura 41 – Clusters do estado de AP.

Analisando a Figura 39, percebe-se que este se refere ao Cluster 9, onde demonstrado na Figura 37 este Cluster é composto pelas piores notas obtidas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que as piores notas foram obtidas na Região Norte do país, por ser a região de menor concentração da populacional, e falta de recursos em cidades dos estados desta região, tem grande impacto nas notas obtidas por estado do Brasil, como pode-se perceber.

Ainda, podemos analisar a Figura 41, onde mostra o estado do Amapá (AP), este é um dos estados que compõe a Região Norte do país, que por sua vez faz parte dos estados

que obtiveram as piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação. Vemos que no estado do Amapá, possui apenas três Clusters, o Cluster 5, 7 e 9 bem concentrado pelas Regiões Norte e Sul do estado, onde estes são como demonstrado na Figura 37, os compostos pelas piores notas no ENEM 2019 em Matemática, Ciências da Natureza e Redação.

4 Conclusões

Incrível perceber a disparidade nas notas por estados no Brasil, onde todos os dias vê-se notícias sobre desigualdades econômicas, raciais, e também regionais, onde nem sempre melhores notas são concentradas exclusivamente em Regiões Metropolitanas dos estados, mas percebeu-se que há talentos e incentivos, que com dedicação e oportunidade de estudo, obtiveram excelentes resultados em suas notas em Municípios fora da Região Metropolitana ou Capital, ou seja, Município de Regiões Norte ou Interior dos estados.

Com este projeto percebi a importância da Análise de Clusters, onde por meio deste recurso, grandes descobertas são possíveis, dentre elas, as que se referem a características populacionais. Para tal, também é muito importante definir uma ferramenta em que auxilie nessa tarefa, sempre observando o que se têm disponível com o que pode ser adquirido, como também a regra de negócio que deseja analisas/observar para a devida correção necessária, assim possibilitando análises e tomada de decisões mais assertivas baseadas em dados.

Assim, é possível estabelecer uma diretriz para a melhoria na organização e distribuição de recursos entre os estados brasileiro, como também por municípios, aumentando a proporção de população com acesso a educação de uma melhor qualidade e recursos sociais, fora o benefício em que se pode diminuir a taxa de desigualdade social e educacional dos brasileiros, com políticas públicas que colaborem para a redução do analfabetismo no país.

Referências

GRUS, J. ***Data Science do Zero. Primeiras Regras com o Python***. 1. ed. Rio de Janeiro: Alta Books, 2016. Citado na página 6.

HAN, J.; KAMBER, M.; PEI, J. ***Data Mining and Machine Learning***. 3. ed. Waltham, MA, USA: Morgan Kaufmann, 2012. Citado na página 6.

PATRíCIA., F. L. P. e B. ***Manual de Análise de Dados – Estatística e Modelagem Multivariada com Excel, SPSS e Stata.*** 1. ed. Rio de Janeiro: GEN LTC, 2017. Citado na página 6.

Anexos

ANEXO A – Dicionário de Dados

A seguir estão descritos todos os bancos de dados, suas tabelas e variáveis.

NOME DAS VARIÁVEIS	DESCRIÇÃO	TAMANHO	TIPO
COR_RACA	Não declarado Branca Preta Parda Amarela Indígena	13	Caractere
NO_MUNICIPIO_RESIDENCIA	Nome do município de residência	150	Caractere
Região_Nome	Centro-Oeste Nordeste Norte Sudeste Sul	12	Caractere
Sexo	Feminino Masculino	9	Caractere
Tipo_Escola	Não Respondeu Pública Privada Exterior	13	Caractere
UF_Capital	Sim Não	3	Caractere
UF_Nome	Nome da Unidade da Federação - UF	19	Caractere
SG_UF_RESIDENCIA	Sigla da Unidade da Federação - UF	2	Caractere
CO_MUNICIPIO_RESIDENCIA	Código do município	8	Numérico
LATITUDE	Latitude do município	8	Numérico
LONGITUDE	Longitude do município	8	Numérico
NU_ANO	2017 2018 2019	8	Numérico
NU_IDADE	Idade do candidato em anos	8	Numérico
NU_INSCRICAO	Numero de inscrição do candidato	8	Numérico
NU_NOTA_CH	Nota de Ciências Humanas	8	Numérico
NU_NOTA_CN	Nota de Ciências da Natureza	8	Numérico
NU_NOTA_LC	Nota de Linguagens e Códigos	8	Numérico
NU_NOTA_MT	Nota de Metemática	8	Numérico
NU_NOTA_REDACAO	Nota de Redação	8	Numérico
NOTA_MEDIA	Nota Média	8	Numérico

[1] Extraído dos dados dos ENEMS 2017-2019 *INEP*.

Tabela 1 – Dicionário da amostra de dados fornecida dos ENEMs 2017-2019 *(INEP)*

ANEXO B – Códigos SAS Base

```
PROC MAPIMPORT OUT=MapaBrasil
     DATAFILE='C:\Users\Usuario\Documents\IESB\5_periodo\Mineracao_de
Run;

Data MapaBrasil;
     Set MapaBrasil;
     Rename CD_Mun = CO_MUNICIPIO_DV;
Run;

/* ----------------------------------------
   Sele   o da Unidade Geogr fica
   ---------------------------------------- */
Data Cluster_Brasil;
     SET DADOS.CLUSTER_AVERAGE_TRAIN;

     Length CO_MUNICIPIO_DV $7.;
     CO_MUNICIPIO_DV = put(CO_MUNICIPIO_RESIDENCIA, z7.);

    IF UF_Nome = 'Rio Grande do Sul';
    *IF _SEGMENT_ in (4, 6, 2);
Run;


proc gmap data = Cluster_Brasil  map=MapaBrasil
     STRETCH UNIFORM;  /* ALL - Imprime sempre o Mapa todo */
     id CO_MUNICIPIO_DV;
     choro _SEGMENT_ No_Municipio_Residencia;
     Title1 '*** Mapas Tem ticos do Brasil, Estados e Municipios ***
     Footnote1 '*** Gera   o SAS ***';
Run;
```