Nome : Gabriel Augusto Nascimento da Silva Costa
Tratamento e Análise de Estatísticas Descritivas – Microdados – ENEM2019
Repositório GITHUB : Tratamento-de-dados---ENEM---2019

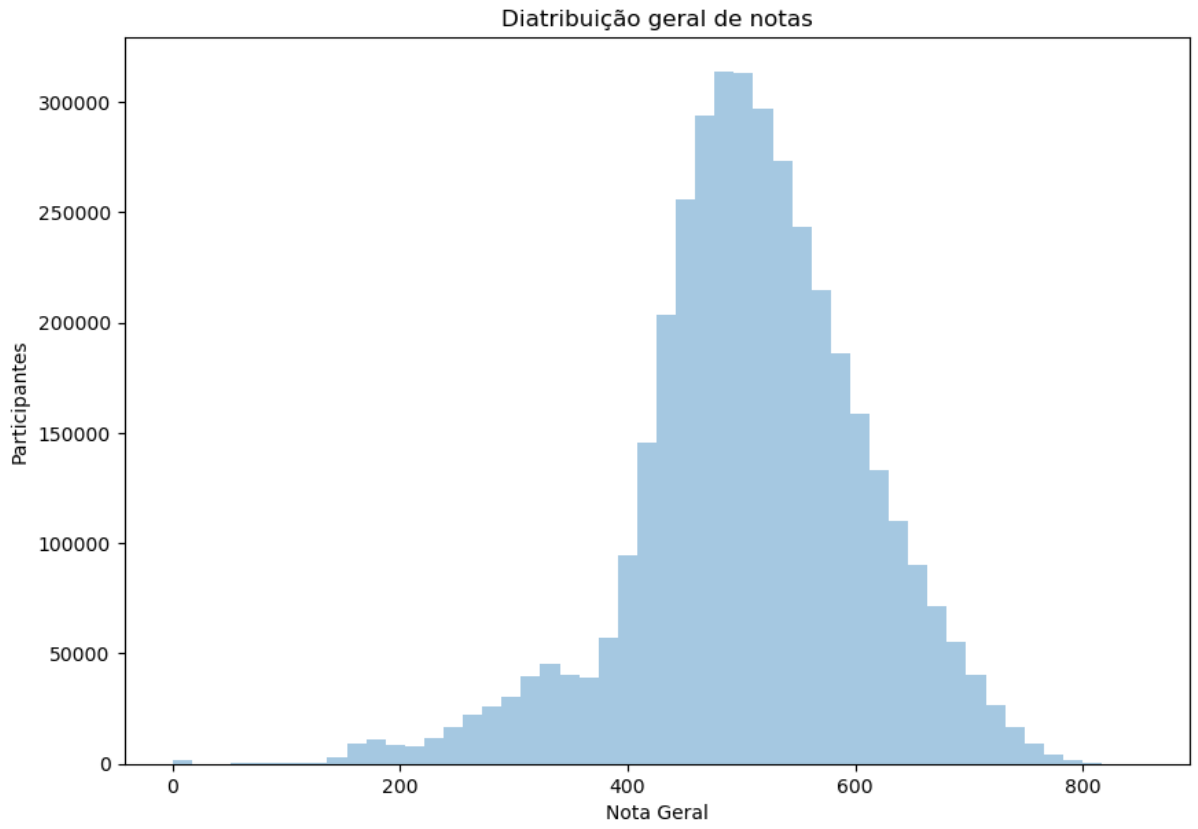
Código preliminar para carregamento e tratamento dos dados : Load_Microdados.py
Plotagem e análises: Análise_Estatística.py , Análise_gabarito.py

obs: Foi-se necessário o particionamento dos dados e o uso de plataformas de compilação remota (Google Collab) em algumas análises, devido a magnitude do tamanho do Dataset e as limitações do hardware disponível.

Começando a partir de algumas inferências mais diretas e fundamentais, seguem algumas distribuições básicas do dataset.

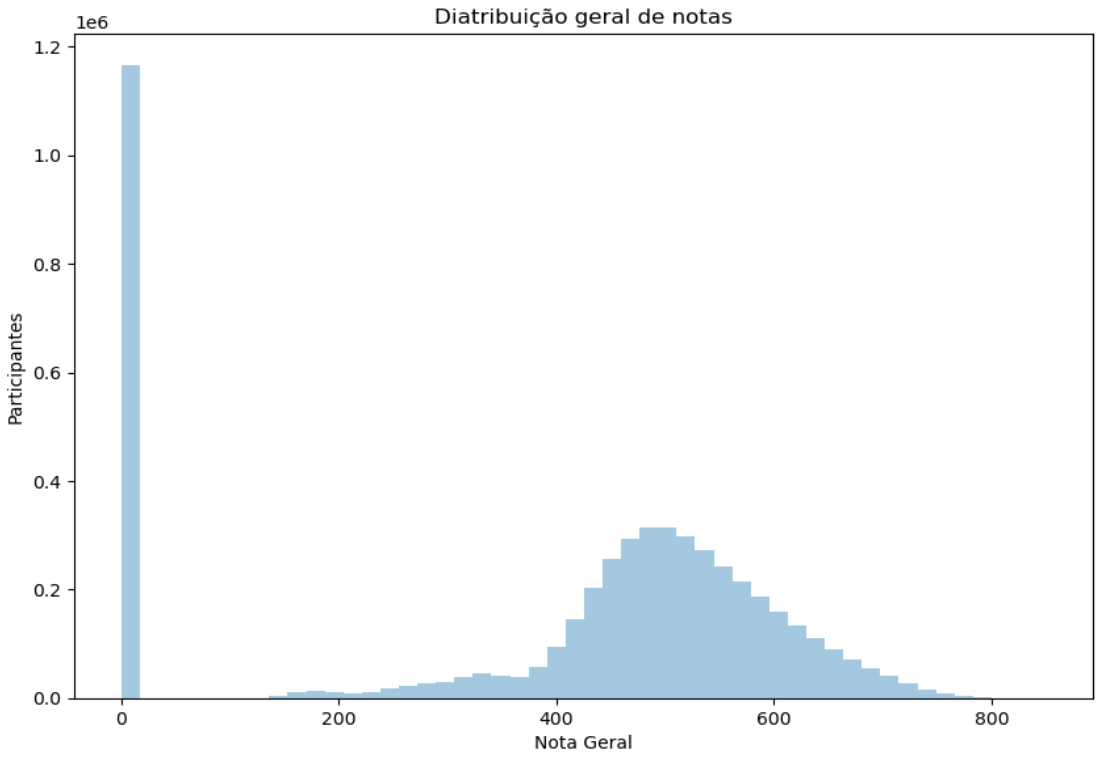
Distribuição da nota geral:

Para esta distribuição foram considerados apenas os participantes que tiveram presença integral durante toda a aplicação da prova, ou seja, os valores faltantes (missing values) devidos ao não comparecimento foram desconsiderados.



Histograma de variável contínua 'nota_geral' com presença integral
Média -> 508.72
Mediana -> 508.88

Em contraproposta, se considerarmos o não comparecimento como 0, temos uma distribuição bastante diferente:



Histograma de variável contínua 'nota_geral' considerando as faltas como 0
Média -> 392.01
Mediana -> 477.12

Baseando-se na última distribuição, é possível inferir que há um grande número de não comparecimentos nas provas, como demonstra a tabela de ocorrências a seguir, em que cada fileira representa uma combinação comum entre diversos candidatos (Ex: Primeira fileira = Presente em todas as provas; Última fileira = Eliminado em todas as provas).

Ciências Naturais	Ciências Humanas	Matemática	Linguagens e Códigos	Contagem	Porcentagem
Presente	Presente	Presente	Presente	3702008	72,65577683
Faltou	Faltou	Faltou	Faltou	1160010	22,76640885
Faltou	Presente	Faltou	Presente	219245	4,302912309
Presente	Faltou	Presente	Faltou	8027	0,157538266
Faltou	Eliminado	Faltou	Eliminado	3670	0,072027586
Eliminado	Presente	Eliminado	Presente	1892	0,037132478

Presente	Eliminado	Presente	Eliminado	398	0,007811166
Eliminado	Faltou	Eliminado	Faltou	16	0,000314017
Eliminado	Eliminado	Eliminado	Eliminado	4	7,85042E-05

É interessante notar que o número de faltas totais (não comparecer em nenhuma prova) representa quase 23% das inscrições, já casos em que há eliminação representam pouco mais de 0.1%. Por esse motivo, a maioria das análises posteriores consideram apenas os candidatos que fizeram a prova.

Comparação de nota por matéria:

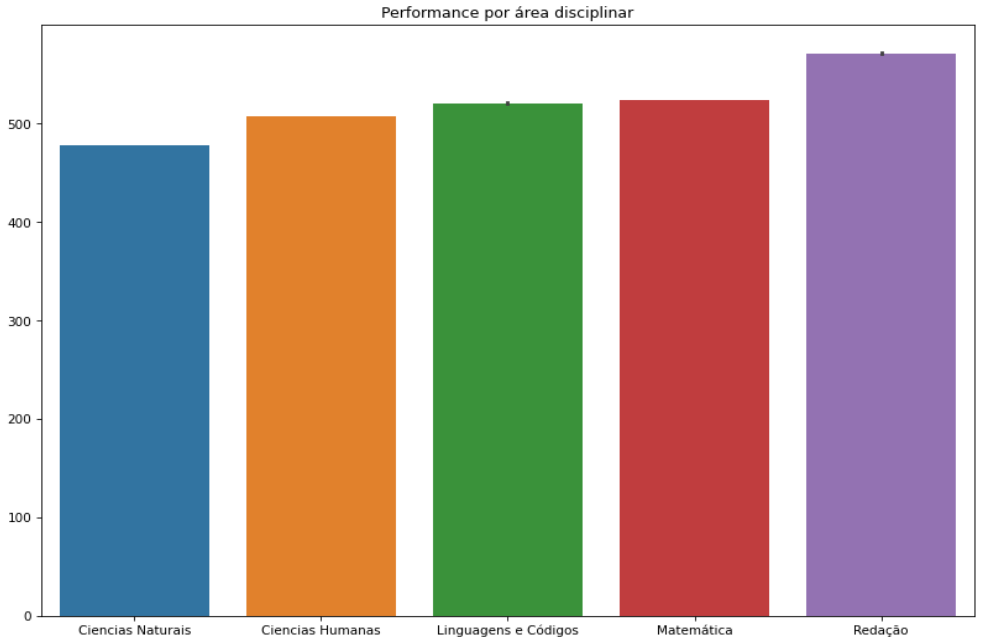
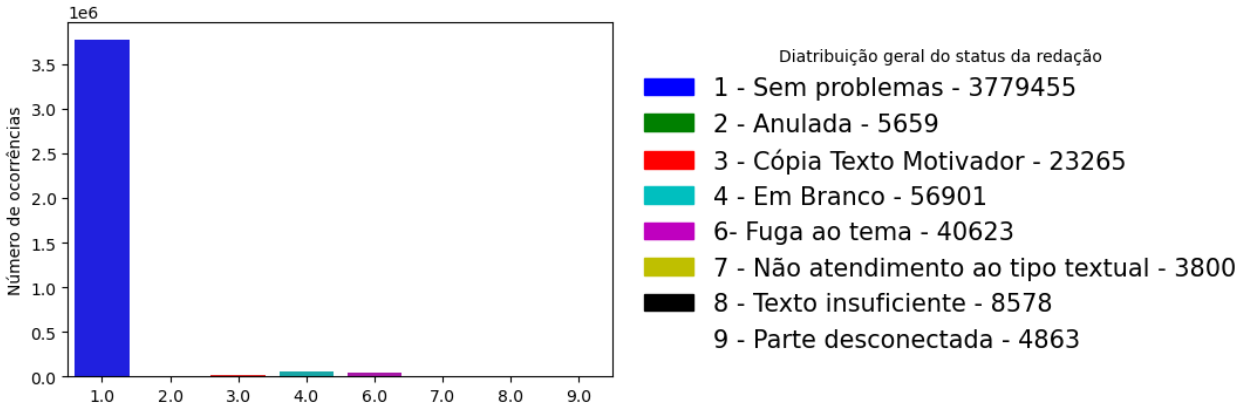


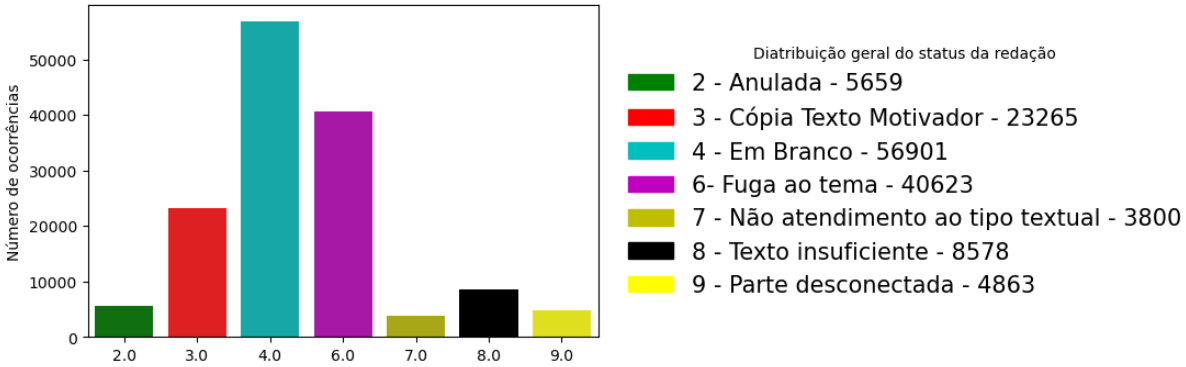
Gráfico em barras comparando a tendência central da nota nas diferentes disciplinas

Apesar de a relação não necessariamente nos permitir realizar alguma inferência mais generalista, podemos nos informar sobre a dificuldade mais expressiva pertinente a aplicação da prova de 2019, especificamente.

Distribuição dos status da redação:



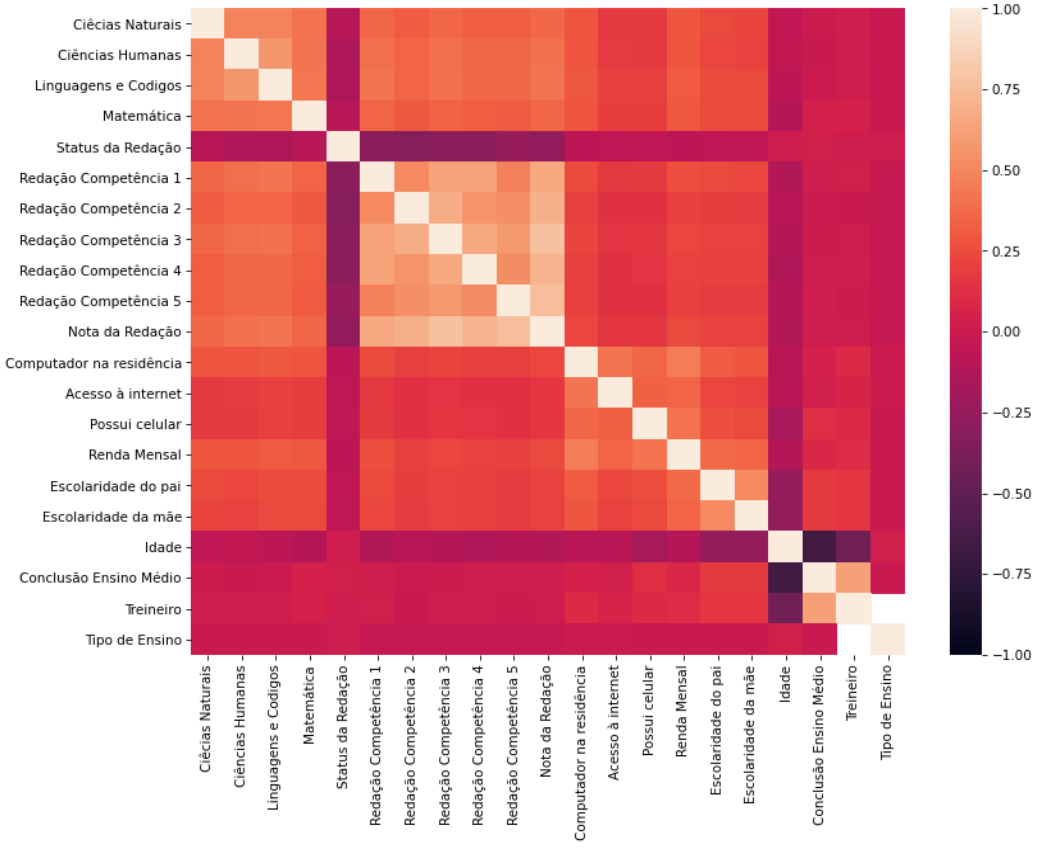
Como é possível observar, ocorrências anormais durante a redação representam um pequena minoria de todos os casos. Para mais fácil visualização segue abaixo uma comparação somente dos status anormais.



Segundo o que o gráfico nos mostra, após os casos em que a redação é deixada em branco, os erros mais comumente cometidos pelos candidatos são, respectivamente, fuga ao tema e cópia do texto motivador. Lembrando que em ambos os gráficos, foram considerados apenas os candidatos que compareceram.

Estudo da correlação entre as variáveis:

O objetivo principal dessa análise é tentar descobrir relações, mesmo que sutis, entre a performance do candidato e diferentes variáveis fornecidas pelo dataset.
Foram utilizados 3 métodos de correlação estatística para a construção de um heatmap (mapa de calor), na tentativa de localizar correlações latentes entre os parâmetros.
Os métodos aferidos foram a correlação de Pearson, Kendall Tau e Spearman. Após conferência na literatura a respeito dos equacionamentos, constatou-se que o método da correlação Tau de Kendall é o mais adequado para o estudo de associações entre variáveis contínuas e ordinais.



As variáveis analisadas foram selecionadas como aquelas as quais foram julgadas como mais pertinentes em relação a sua influência na performance do candidato.

Kendall's tau-B values:

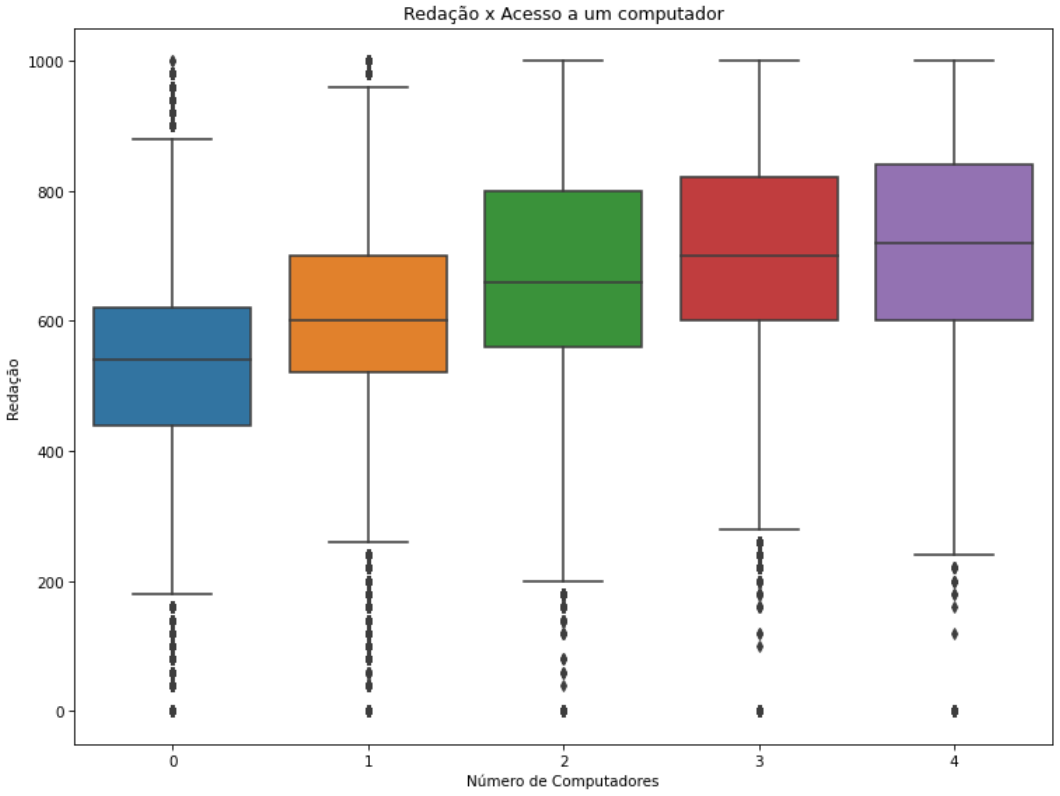
- Less than + or - 0.10: very weak
- + or -0.10 to 0.19: weak
- + or - 0.20 to 0.29: moderate
- + or - 0.30 or above: strong

A interpretação do gráfico se dá pela tabela de valores Tau. Ou seja, quanto mais o valor da correlação estiver próximo aos extremos (1 para correlação positiva, e -1 para correlação negativa) mais forte é a associação entre as variáveis. Tendo em mente a tabela ao lado, podemos nos guiar pelo dataset mais facilmente a procura de relações expressivas.

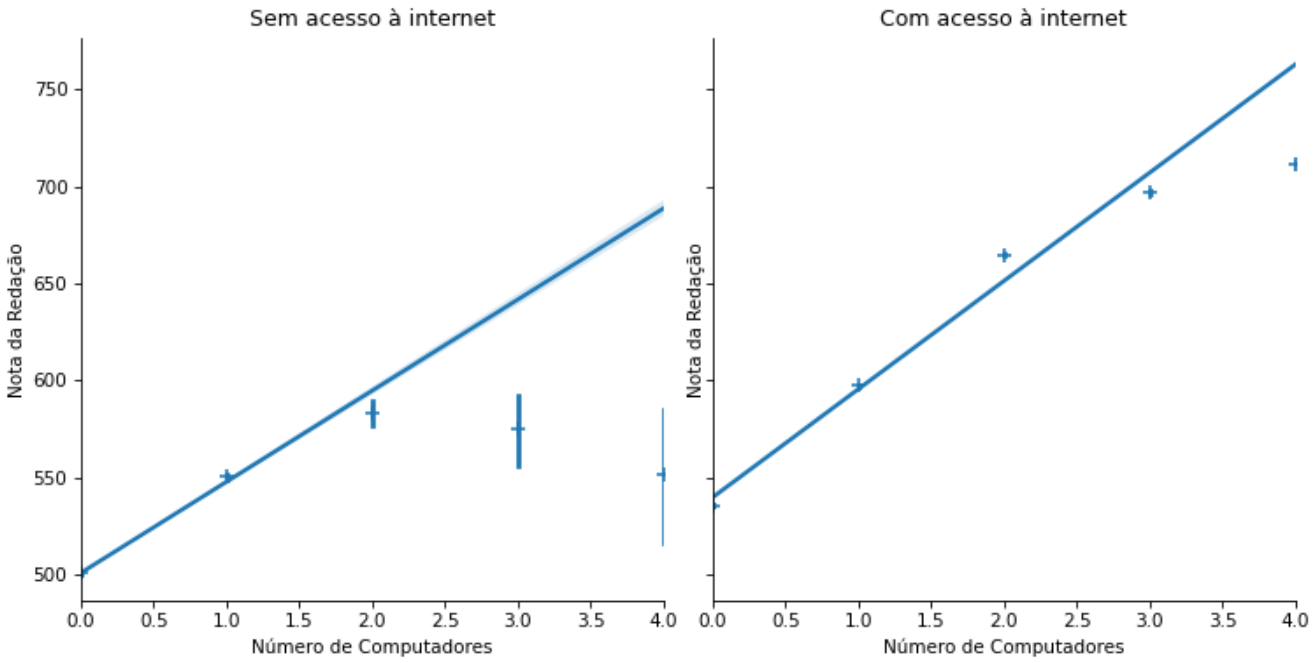
Inferências a partir da performance e suas correlações a partir do mapa de calor:

Acesso à informação e redação:

Se manter informado das notícias e do panorama de acontecimentos do Brasil e do mundo é essencial para um bom embasamento em seus argumentos na redação, uma vez que é muito comum o seu tema envolver conhecimentos de mundo ou de questões às quais requerem um certo nível de acesso à informação quanto aos fatos, ocorrências e eventos da atualidade.
É possível enxergar essa relação a partir da correlação entre as variáveis as quais acusam esse acesso (número de computadores na residência, celular, internet...) e a nota na redação.



Boxplot Computadores na Residência x Nota da Redação -> Kendall Tau = 0.229654

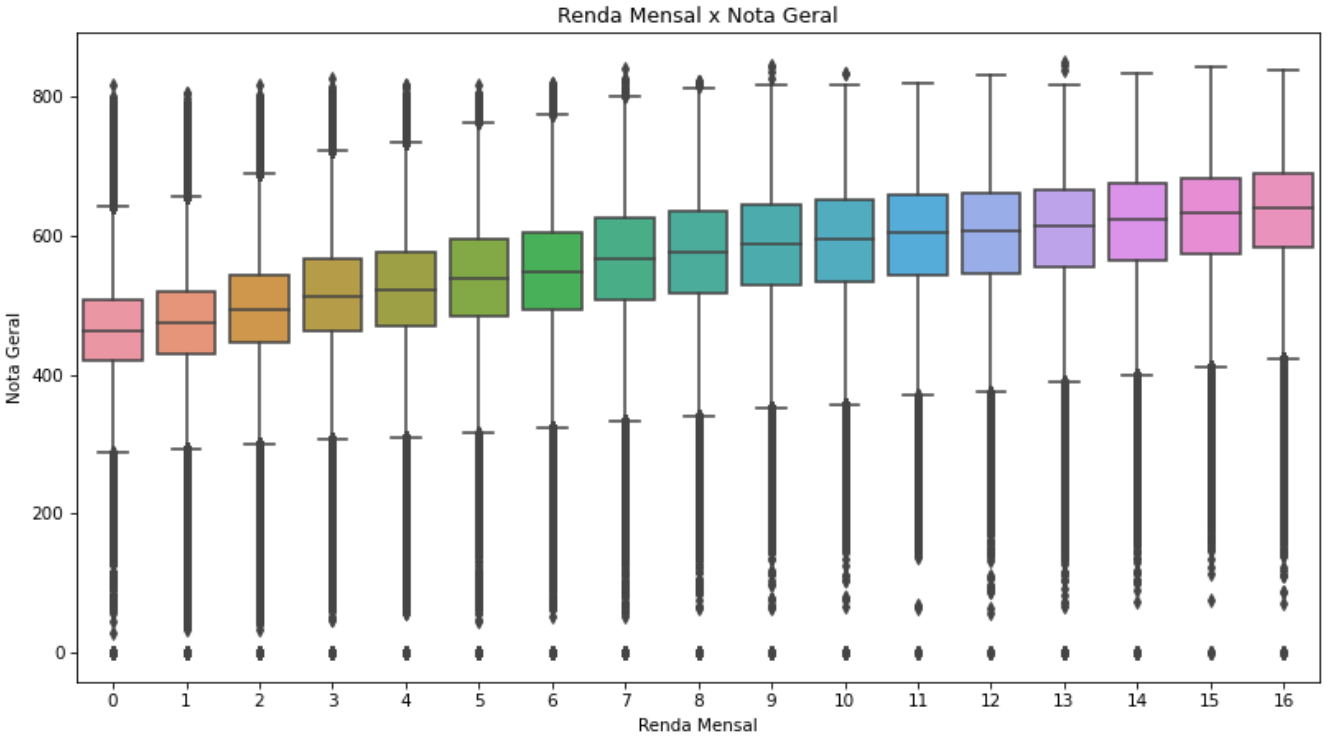


Regressão Linear -> Computadores na Residência x Nota da Redação - >com e sem acesso à internet

A relação positivamente correlacionada se mantém em ambos os casos de acesso a internet, porém observa-se que os candidatos com acesso à mesma têm a tendência central da nota mais elevada.

Renda Mensal e Nota Geral:

Partindo do mesmo princípio da correlação anterior, a renda mensal se mostrou fortemente correlacionada (Tau entre valores de 0.30 ou maior) com a performance em **todas** as

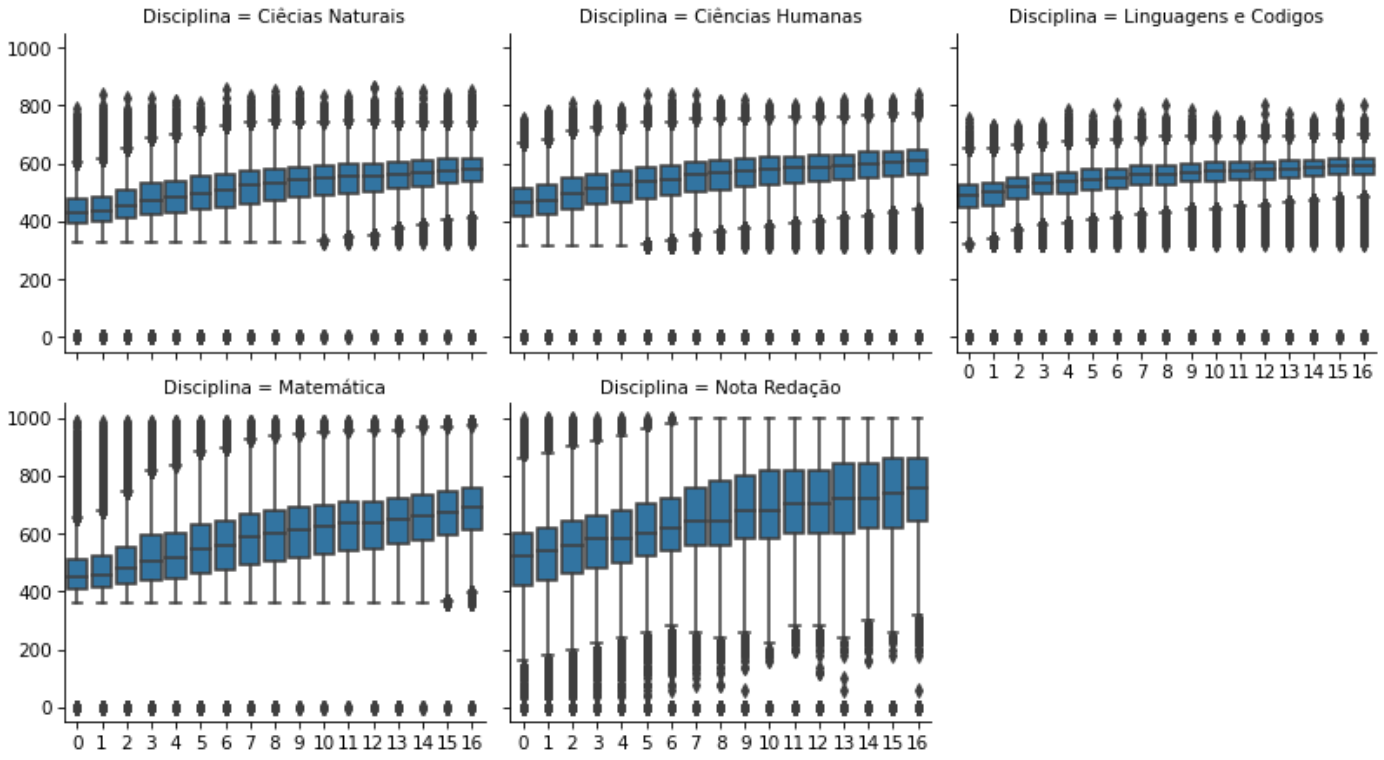


disciplinas, uma vez que parâmetros como, tipo de ensino e acesso à informação são intrinsecamente dependentes da renda mensal familiar.

Boxplot Renda Mensal x Nota Geral -> Kendall Tau = 0.312418

Abaixo segue a relação entre a nota de cada disciplina e a renda mensal da família dos candidatos. Segundo o dicionário fornecido a renda mensal está distribuída ordinal e categoricamente em 16 níveis.

Nenhuma renda.
Até R$ 998,00.
De R$ 998,01 até R$ 1.497,00.
De R$ 1.497,01 até R$ 1.996,00.
De R$ 1.996,01 até R$ 2.495,00.
De R$ 2.495,01 até R$ 2.994,00.
De R$ 2.994,01 até R$ 3.992,00.
De R$ 3.992,01 até R$ 4.990,00.
De R$ 4.990,01 até R$ 5.988,00.
De R$ 5.988,01 até R$ 6.986,00.
De R$ 6.986,01 até R$ 7.984,00.
De R$ 7.984,01 até R$ 8.982,00.
De R$ 8.982,01 até R$ 9.980,00.
De R$ 9.980,01 até R$ 11.976,00.
De R$ 11.976,01 até R$ 14.970,00.
De R$ 14.970,01 até R$ 19.960,00.
Mais de R$ 19.960,00.

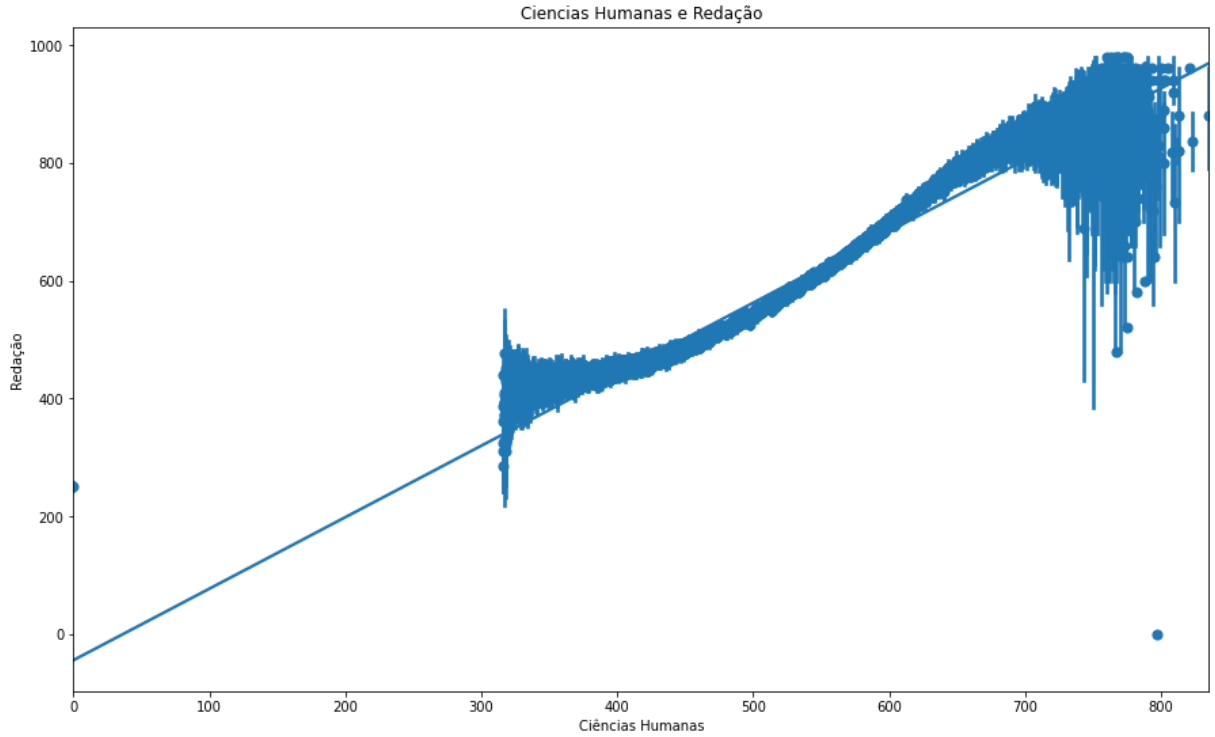


Boxplot Renda Mensal x Nota de todas as disciplinas -> Kendall Tau (respectivamente) = [0.293 0.291 0.309 0.303 0.238]

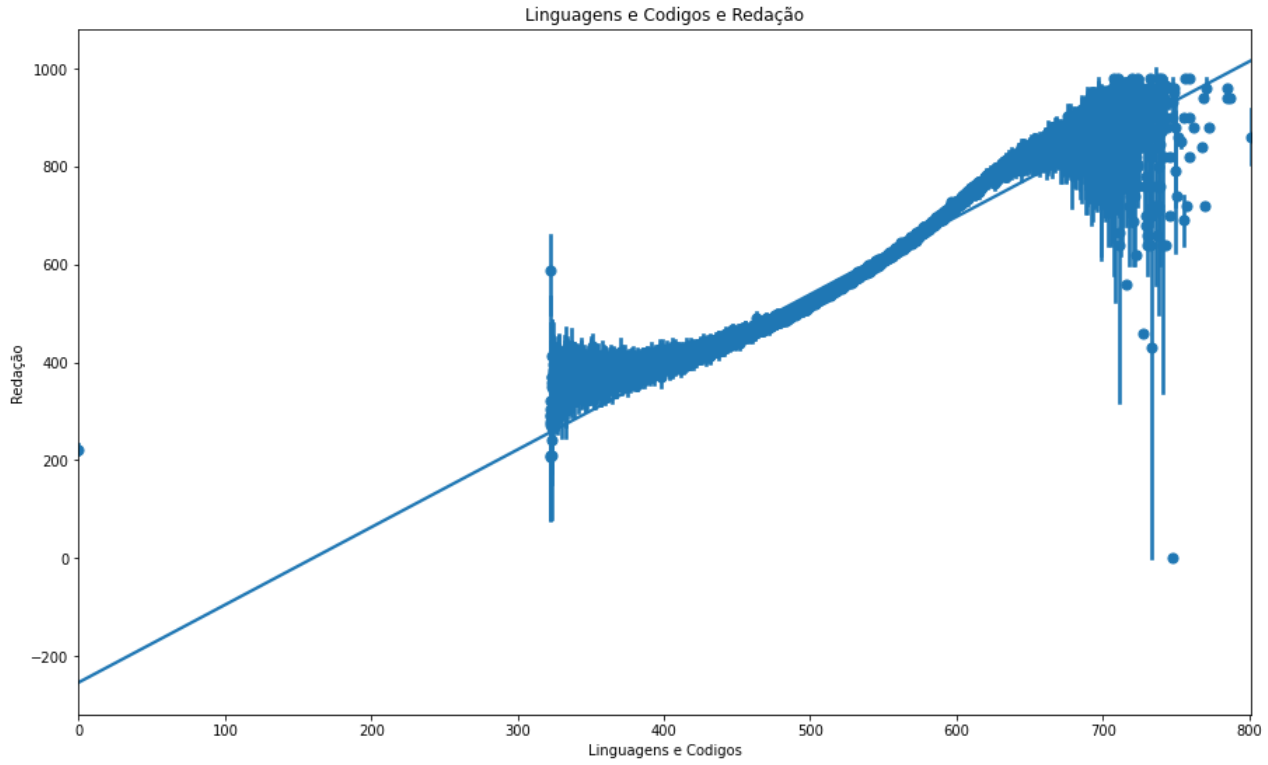
Através dos gráficos e dos valores das correlações podemos verificar que matemática é a disciplina mais influenciada pela renda mensal.

Ciências Humanas, Linguagens e Códigos x Redação:

Devido a natureza dessas disciplinas e dos temas de redação, é de se esperar que haja correlação entre essas variáveis, podemos observar que tanto Ciências Humanas quanto Linguagens e Códigos exercem grande influência sobre a performance dos candidatos na redação, ou seja, estatisticamente, candidatos que dominam essas disciplinas tendem a ir melhor em competências como domínio da língua portuguesa e aplicação de conceitos históricos em suas argumentações.



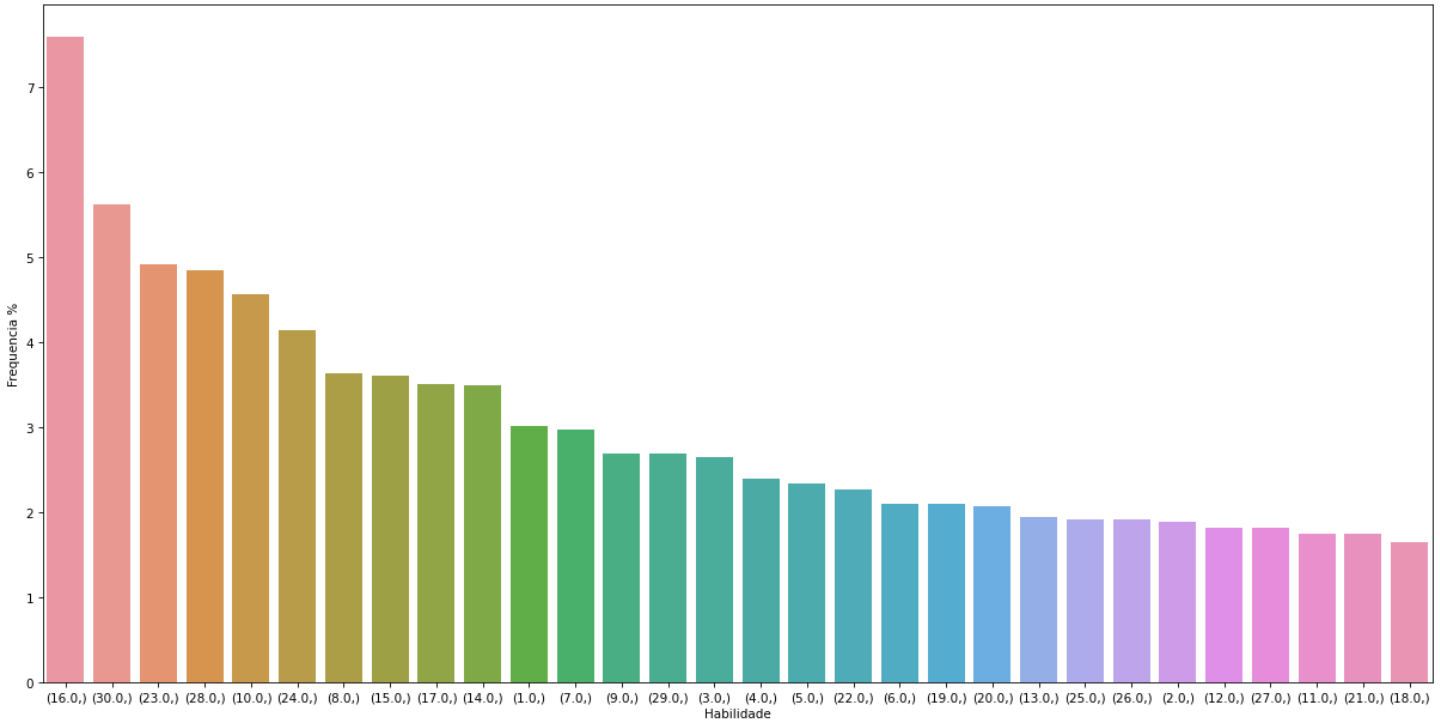
Plots de regressão entre as notas de Ciências Humanas e a nota da redação, Kendall Tau = [0.402]



Plots de regressão entre as notas de Linguagens e Códigos e a nota da redação, Kendall Tau = [0.415]

Análise do gabarito, competências e habilidades:

Seguindo a mesma linha de raciocínio da análise acima, é interessante portanto, tentar descobrir quais as habilidades dessas disciplinas (Ciências Humanas e Linguagens e Códigos) as quais os candidatos mais tem dificuldade. Segundo a matriz de referência do ENEM, disponibilizado pelo INEP, cada questão de diferentes disciplinas possuem uma série de competências e habilidades distintas, os quais devem ser dominados pelo candidato para que este consiga acertá-las.
Segue abaixo esta análise referente a disciplina de Linguagens e Códigos:



Distribuição da freqûencia de erro referente a diferentes habilidades

Para desenvolver a distribuição acima, foi necessário iterar ao longo de todas as questões, cartelas, gabaritos e códigos de todos os candidatos, a fim de coletarmos as habilidades das questões cuja os candidatos não tiveram sucesso.
Como podemos observar as habilidades de número 16 e 30 foram as mais erradas pelos candidatos, isso pode significar um peso maior inerente a ela mesma, ou algo relacionado ao modelo da prova de 2019 (Questões mais difíceis correspondendo a essas habilidades nesse ano específico)

H16 - Relacionar informações sobre concepções artísticas e procedimentos de construção do texto literário.

H30 - Relacionar as tecnologias de comunicação e informação ao desenvolvimento das sociedades e ao conhecimento que elas produzem.