Documentação Etapa 2 CCF 425

Nome: Murillo Santhiago Souza Jacob
Matrícula: 4243

Etapa 1 - Trabalhando Em Cima das Colunas do Dataframe

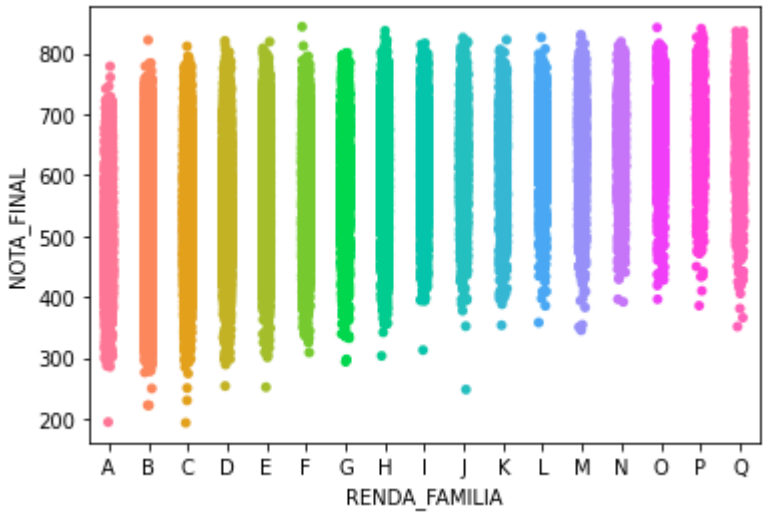
Os microdados do enem, vieram com uma série de necessidade de tratamento, tanto de nome quanto de informação. Vieram com uma série de valores NaN, sem contar que a quantidade de linhas é extremamente grande, a ponto de travar o kernel python que o notebook utiliza. Visto isso, limitei inicialmente a 500 mil linhas, valor que pode ser alterado no futuro, e dropei os valores NaN, e momentaneamente, mantive valores que não foram informados inicialmente.

Além disso, diversos campos vieram com valores nada descritivos no nome das colunas, como Q001 - Q025, fora os valores que substituem valores categóricos por valores numéricos, o que pode confundir e muito na hora de construir o modelo e realizar a análise exploratória, dito isso, também alteramos esses valores para os respectivos dados categóricos.

Outro problema é nas questões 1-25, que foram renomeadas, porém, os valores também estão abreviados, e muitas perguntas do tipo quantidade de eletrodomésticos por exemplo, podem ser convertidas para numéricos, porém, nessa primeira análise exploratória, manterei dessa forma para facilitar a análise. Fora isso, também foi construída a coluna nota final, com base na média das outras notas

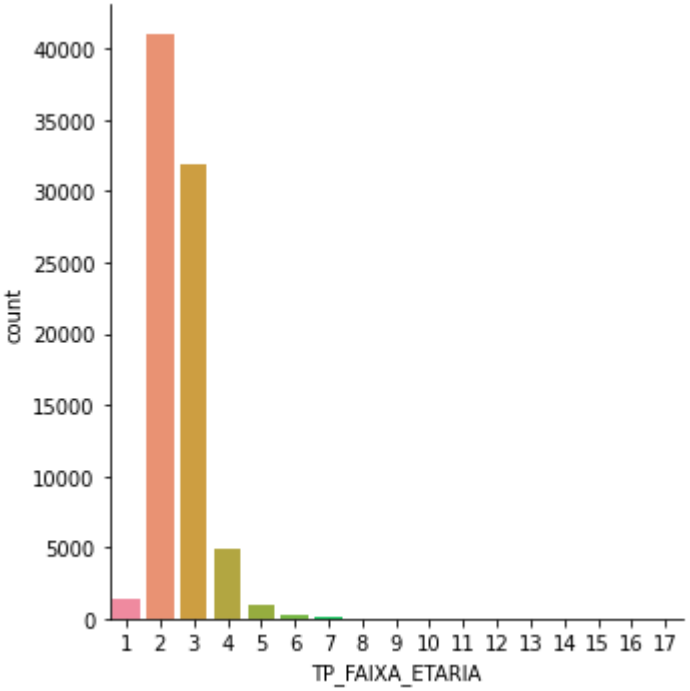
Etapa 2 - Análise Exploratória

Começamos inicialmente com a análise exploratória da renda x nota, pois é algo que pode afetar diretamente a nota final, e tivemos o seguinte resultado:



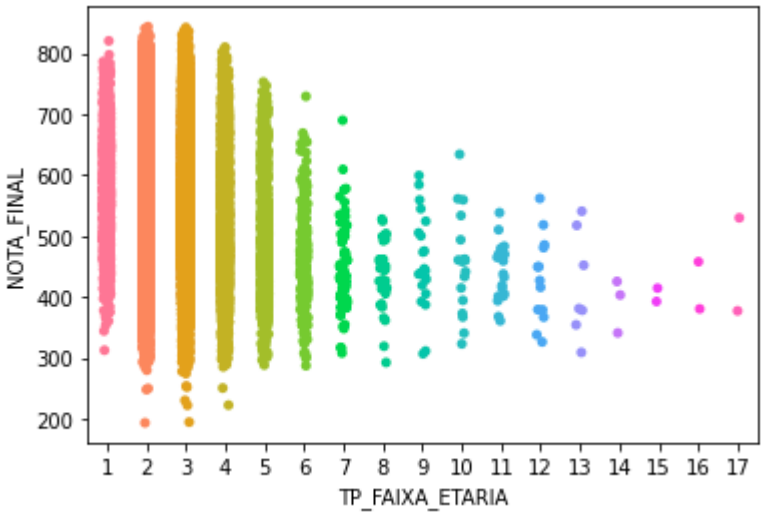
Podemos observar que quanto mais a renda aumenta, maior a concentração de notas na parte de cima da tabela, existe uma correlação clara entre essas duas variáveis.

Já na questão da faixa etária, tirando o gráfico de count dos valores, vemos uma predominância enorme entre o público jovem, como pode ser observado aqui:



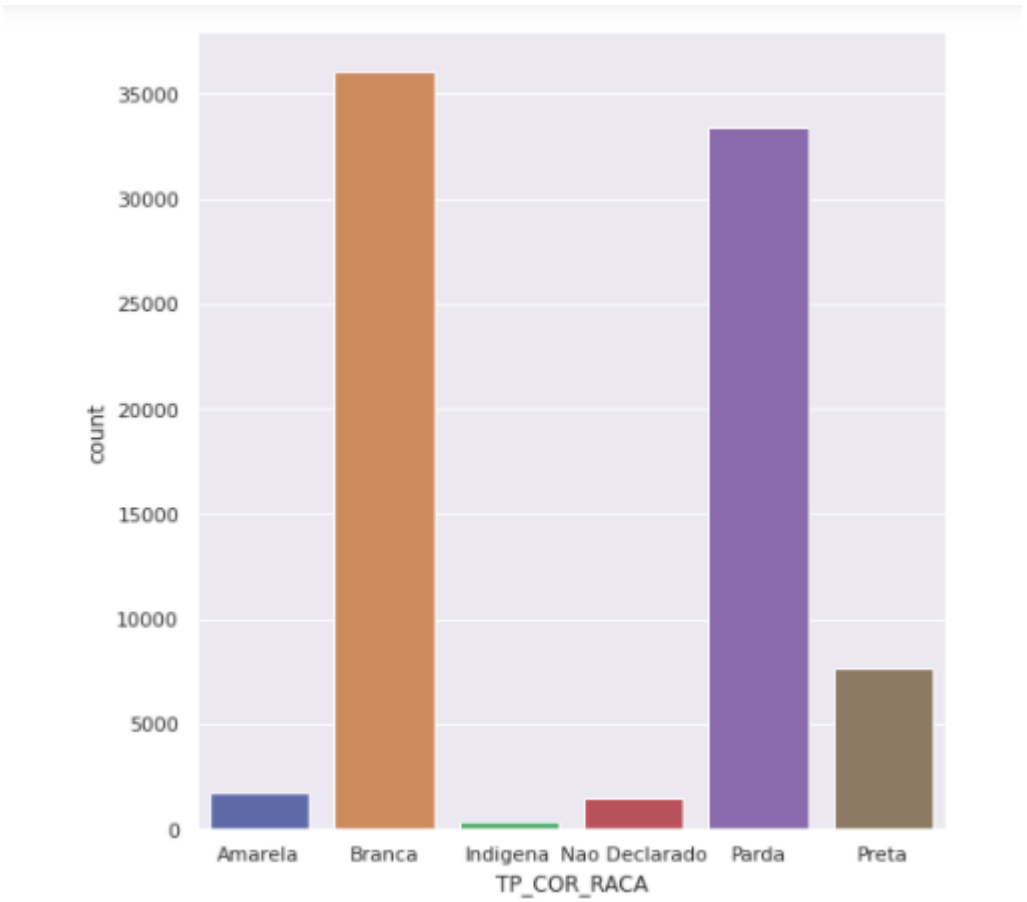
Como vemos, temos uma predominância muito grande entre os indivíduos de 17 e 18 anos, e uma taxa bem menor dos indivíduos de faixa etária mais avançada.

E, sobre a correlação entre a nota e idade, temos o seguinte gráfico:

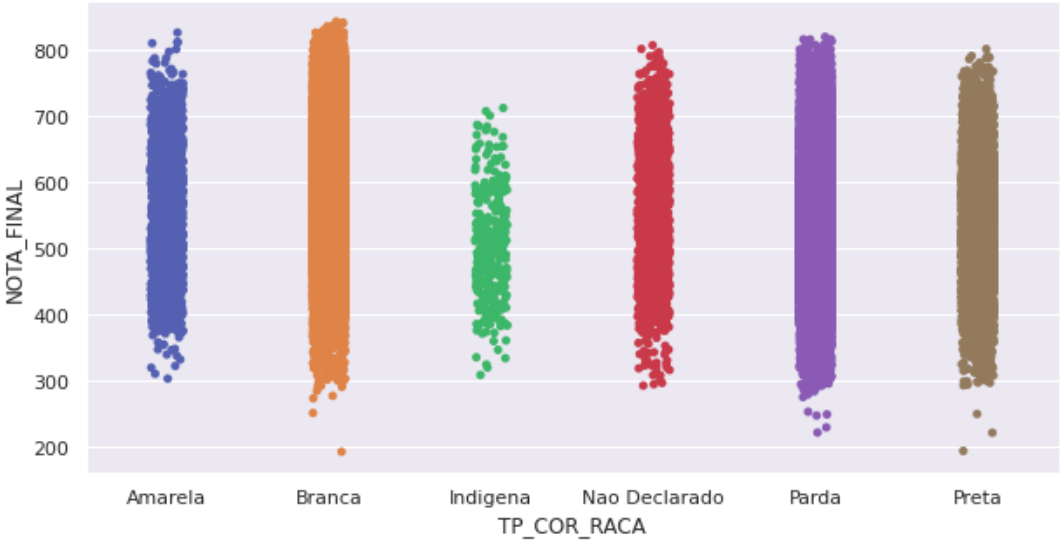


Como podemos ver, as notas de pessoas mais velhas tendem a ser mais baixas, porém, mais esporádicas, visto que essas pessoas ocupam uma quantidade menor na taxa de matriculados. Agora, será que essas notas são mais baixas por uma dificuldade devido a idade, ou devido a quantidade menor de pessoas?

Já analisando a questão das raças, temos a seguinte diistribuição:

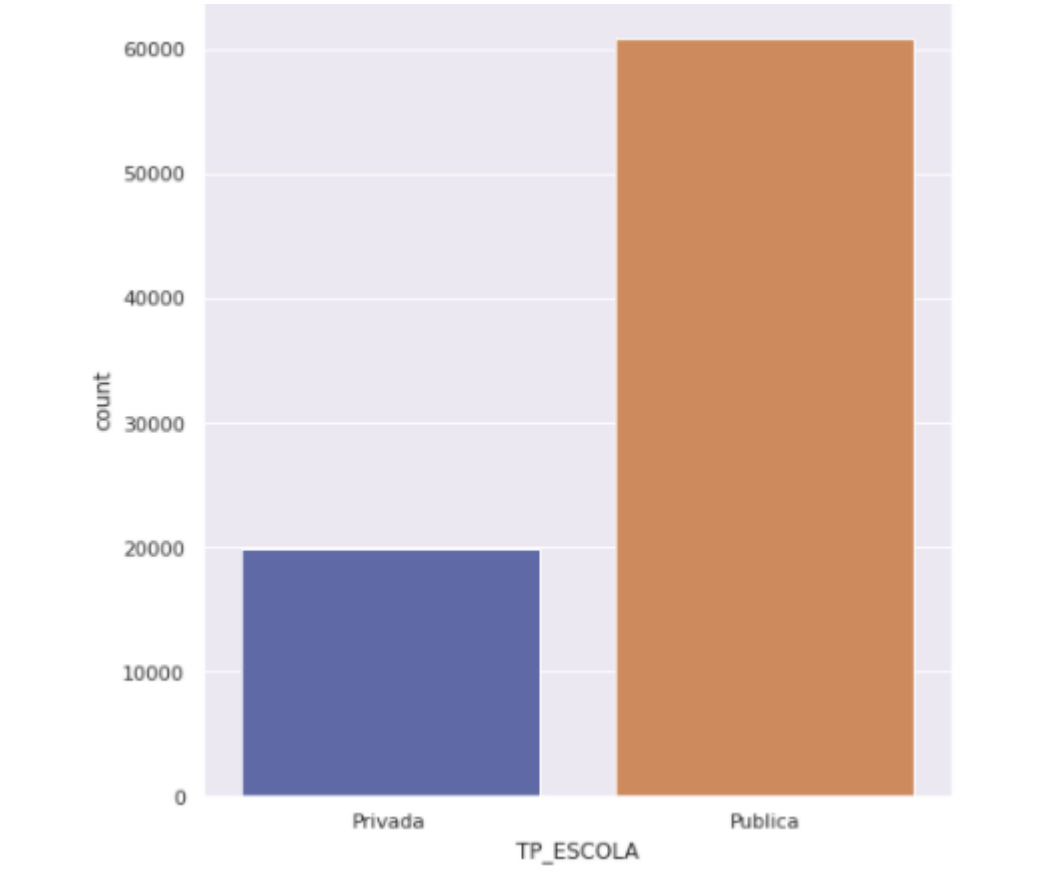


Como podemos ver, temos uma concentração muito maior de brancos e pardos do que o restante, vamos ver o impacto disso nas notas:

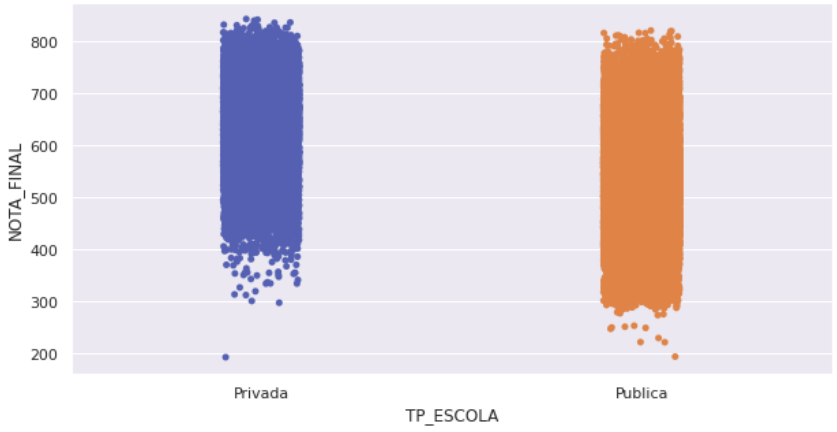


Aqui, vemos que os indígenas ficam um pouco abaixo da maioria, e pessoas brancas tem o maior range. Será que isso pode ser por oportunidades a mais do que os demais? Algo que é muito palpável, visto que as cotas têm um papel importante para corrigir essa diferença, e é aplicada no enem.

Já na questão de escola pública/privada, temos o seguinte:



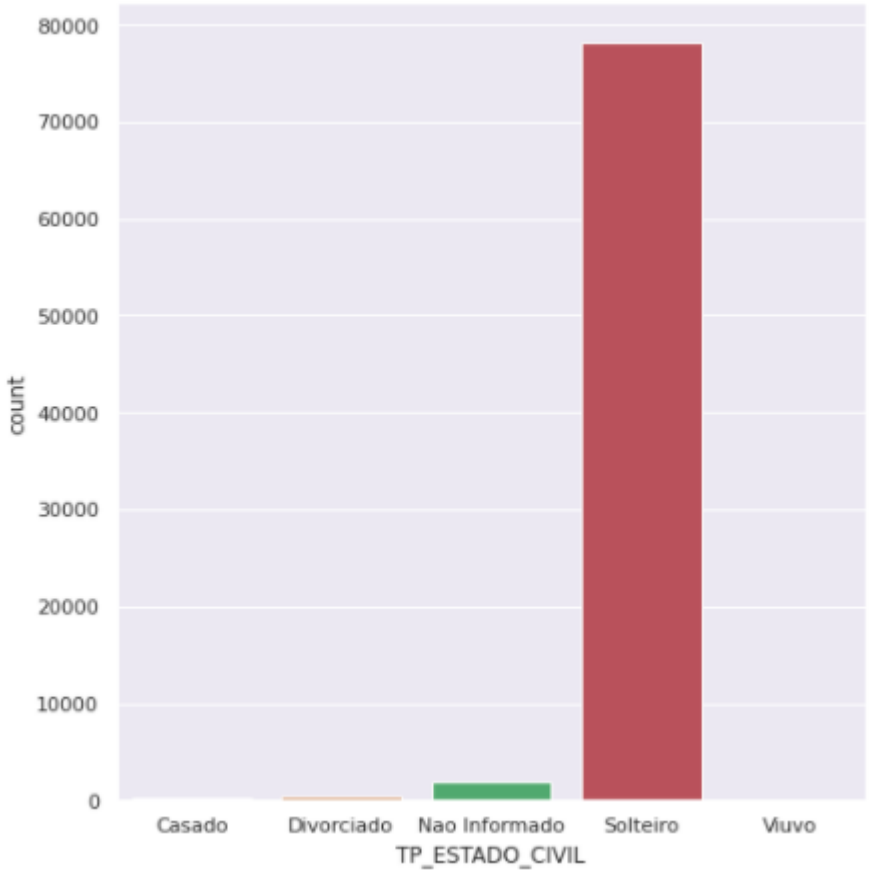
Vemos que temos muito mais alunos de escola pública do que privada, cerca de aproximadamente, o triplo. E visto isso, vamos analisar o impacto disso nas notas.



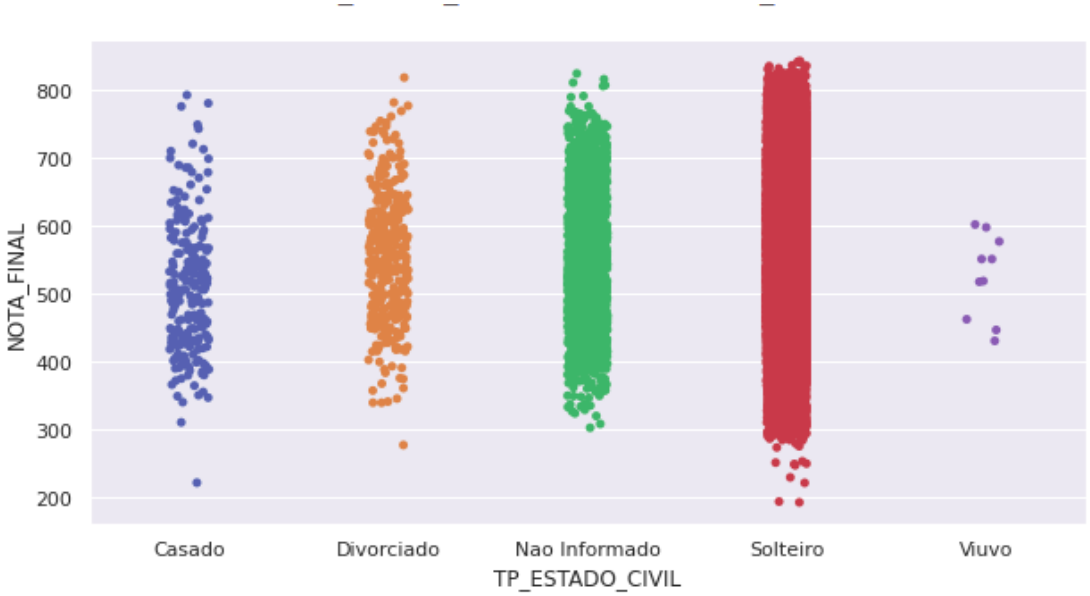
Aqui, podemos ver que a nota de pessoas da escola privada acaba tendo uma média maior do que as notas da escola pública, o que pode ser justificado devido a diferença de nível de ensino durante o ensino médio dos dois tipos de instituições. Sabemos que no Brasil, o ensino básico carece de recursos e de infraestrutura. Esses valores também podem ser esperados, visto que temos uma cota para instituições públicas.

Até o momento, tudo que foi especulado, está se mostrando nos dados, portanto, vamos continuar a análise.

Agora, analisaremos o estado civil e o impacto disso nas notas, pra começar com a distribuição dos mesmos:



Vemos que a grande maioria é solteira, seguido de pessoas que não informaram, e um pequeno índice de divorciados. Vemos também que casados e viúvos são mínimos. Já com relação a distribuição das notas:

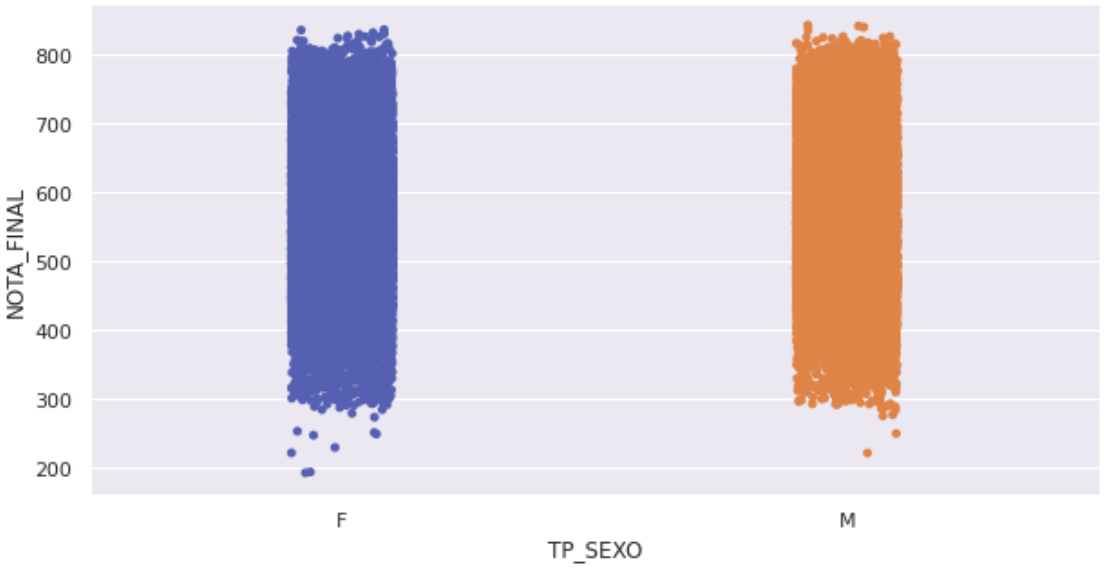


Aqui, vemos os viúvos bem abaixo da média, sendo que o seu maior valor de nota é de cerca de 600. Os solteiros com as maiores, mas também com as menores notas, o que é justificado por serem a maioria, e os casados e divorciados se mantendo entre 400 - 800.

Partindo para análise entre gêneros, temos a seguinte distribuição:

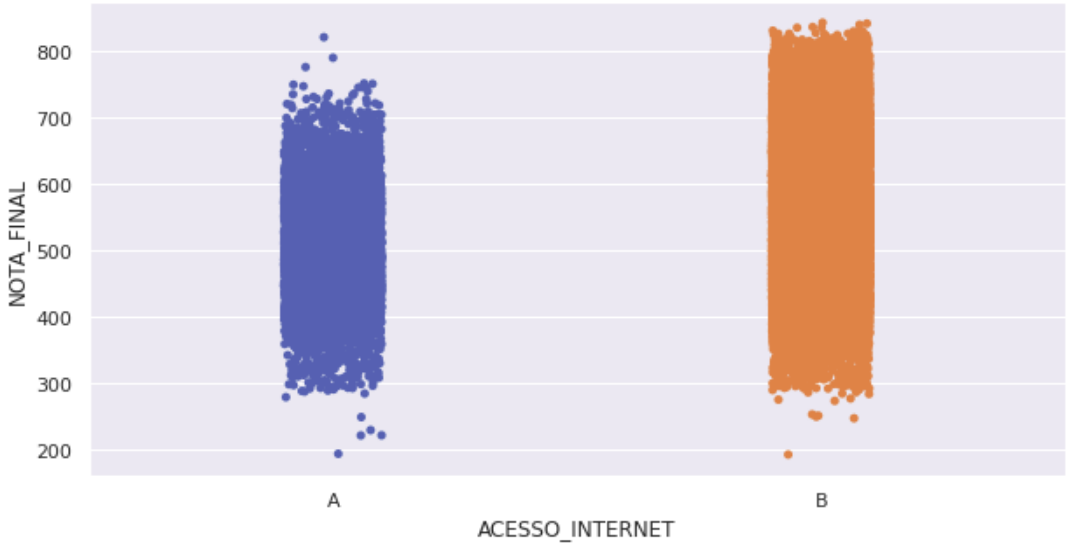
Vemos uma quantidade significativamente maior de mulheres do que de homens.

E na distribuição entre notas, temos:



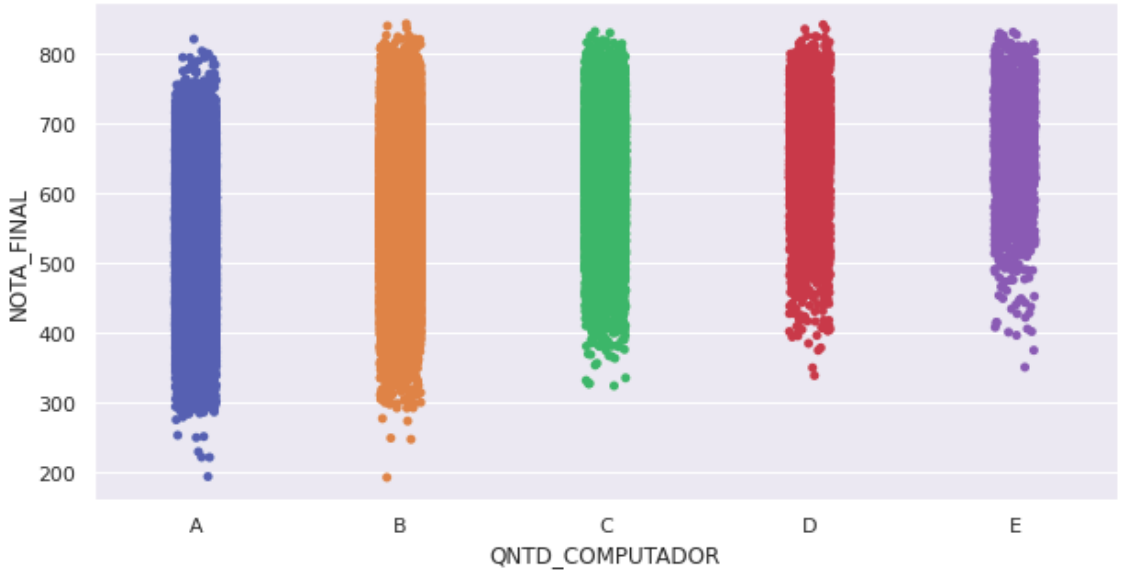
Aqui, vemos uma distribuição bem semelhante entre os gêneros, com ambos ocupando praticamente a mesma faixa. Como dropamos os valores NaN, as pessoas que faltaram foram descartadas, logo descartamos essa abordagem.

Também podemos ver a discrepância nas notas de pessoas que possuem e não possuem acesso a internet abaixo:



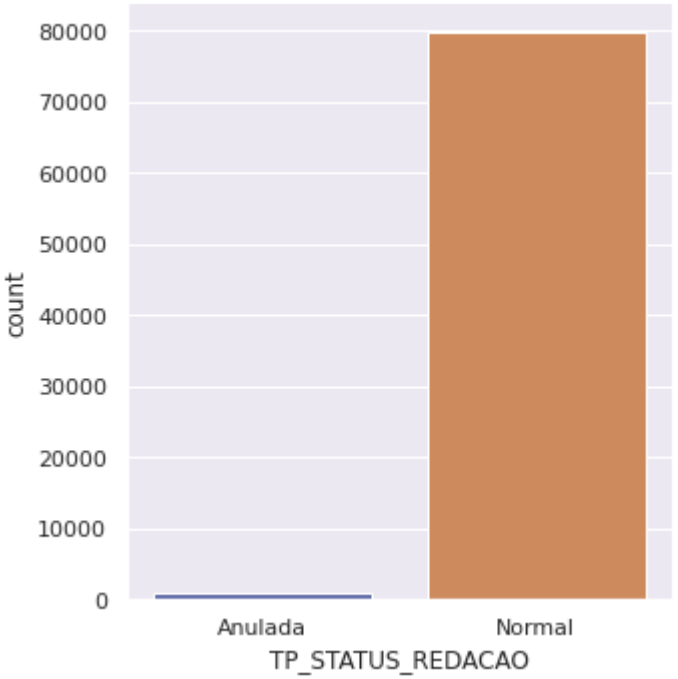
A discrepância é relativamente alta, e é bem evidente que ela existia, o acesso a internet acaba sendo um grande canal de estudos, e sem o mesmo, acaba sendo difícil se preparar para a prova.

Aqui, vemos a relação entre a quantidade de computadores e a nota

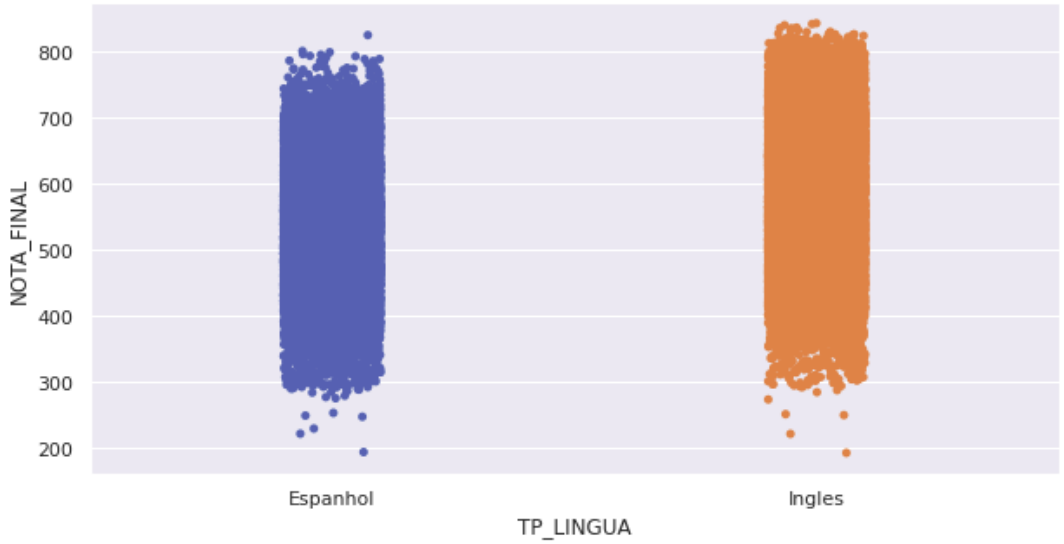


Vemos uma relevância significativa nesse quesito, em que quanto mais computadores, maior a tendência de agrupamento no topo, mostrando que esse dado é significativo e importante.

Já a relação entre redações anuladas e relações não anuladas, temos o seguinte:



Já na questão de escolha de lingua estrangeira, achamos o seguinte:



Vemos que as pessoas que escolheram inglês tiraram notas um pouco mais elevadas do que as que escolheram espanhol.