ANALISANDO MICRODADOS DO ENADE: UMA PROPOSTA DE TRABALHO DE CIÊNCIA DE DADOS

Universidade Federal Fluminense

Instituto de Computação

Introdução a Ciência de Dados

Professor: José Viterbo

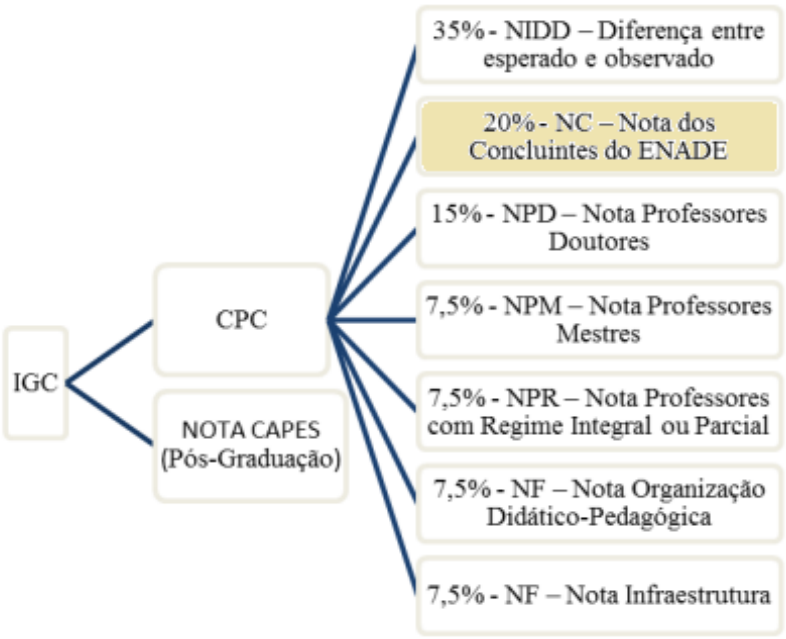
Autor: Jorge Felipe Campos Chagas

1. INTRODUÇÃO

A partir da década de 90 houve uma grande expansão no sistema de ensino superior do país, um fator que contribuiu para esse fenômeno foi a LDBEN(Lei de Diretrizes e Base Educacionais na Educação Nacional), pois apontou melhoras tanto para o ensino presencial como para o ensino à distância. Mais toda essa expansão precisa de um controle, desde então o governo vem propondo ferramentas de avaliação para o ensino superior afim de prover melhorias e garantir um padrão de qualidade dos cursos.

Nesse sentido, foi criado o SINAES(Sistema Nacional de Avaliação do Ensino Superior) cujo objetivo e oferecer uma avaliação sistêmica e periódica. Para que o SINAES funcione foram propostas várias ferramentas de avaliação dado a complexidade do ensino superior. Uma dessas ferramentas é o ENADE(Exame Nacional do Desempenho dos Estudantes), uma avaliação em larga escala que busca avaliar o desempenho dos discentes em questões específicas do curso e de conhecimento geral, aspectos socioeconômicos e informações institucionais.

Mas só o ENADE é suficiente para julgar o nível de qualidade de um curso e/ou instituição de ensino superior? Não, abaixo mostro um diagrama estrutural do sistema de avaliação do ensino superior mencionado por Miranda 2011.



Fonte: Miranda 2011

No esquema acima temos o IGC: Índice Geral de Avaliação dos Cursos, o CPC conceito preliminar dos Cursos, Nota Capes que é a nota dada aos cursos de pós-graduação etc. Neste trabalho nosso foco será o conceito ENADE, ou seja, a nota puramente obtida no exame.

E qual o objetivo desse trabalho? Olhar o conceito ENADE obtido pela Universidade Federal Fluminense em 2011 e através dos microdados explorados do INEP sobre o ENADE, para isso, a princípio, observaremos os dados pelo gênero, tempo de prova, nota total, nota específica e quanto tempo o discente levou para sair do ensino médio e ingressar no curso utilizando conceitos de estatística básica e programação.

2. METODOLOGIA

Consultar o arquivo(.csv) com os microdados do ENADE 2011 e realizar o tratamento dos dados através da linguagem python com o auxílio das bibliotecas numpy, panda, matplotlib e scipy as que ajudaram na construção dos dados visuais e o algoritmo do trabalho. Em seguida, fazer a análise dos dados univariável e bivariável com base no que foi proposto no objetivo do parágrafo anterior utilizando conceitos como média, mediana, desvio padrão, quartis, histograma, coeficiente de Pearson, coeficiente de Spearman's e Scatterplots.

3. CÓDIGO e ANÁLISE DE DADOS

Primeira etapa do trabalho consistiu em abrir o arquivo me2011.csv e filtrar as variáveis que iremos analisar. Aqui tive um problema de formatação que foi resolvido ao adicionarmos: "*encoding = "latin1" , sep = ";" , usecols = filtro* "(linha14).O filtro são as colunas que tem as variáveis: 'nu_idade' que é a idade do candidato, 'co_curso' o código do curso, 'tp_sexo' sexo do estudante, 'ano_fim_2g' ano em que o estudante terminou o ensino médio, 'ano_in_gra' ano de ingresso no curso, 'nt_ger' nota geral obtida no exame e 'nt_ce' nota obtida na prova de conhecimentos técnicos.

```
### Criando nosso data FRame ###
filtro = ['nu_idade','co_curso', 'tp_sexo','ano_fim_2g','ano_in_gra','nt_ger','nt_ce']
df = pd.read_csv('C:/Users/jungl/Documents/PythonProjects/me2011.csv',encoding = "latin1" , sep = ";" , usecols = filtro)
```

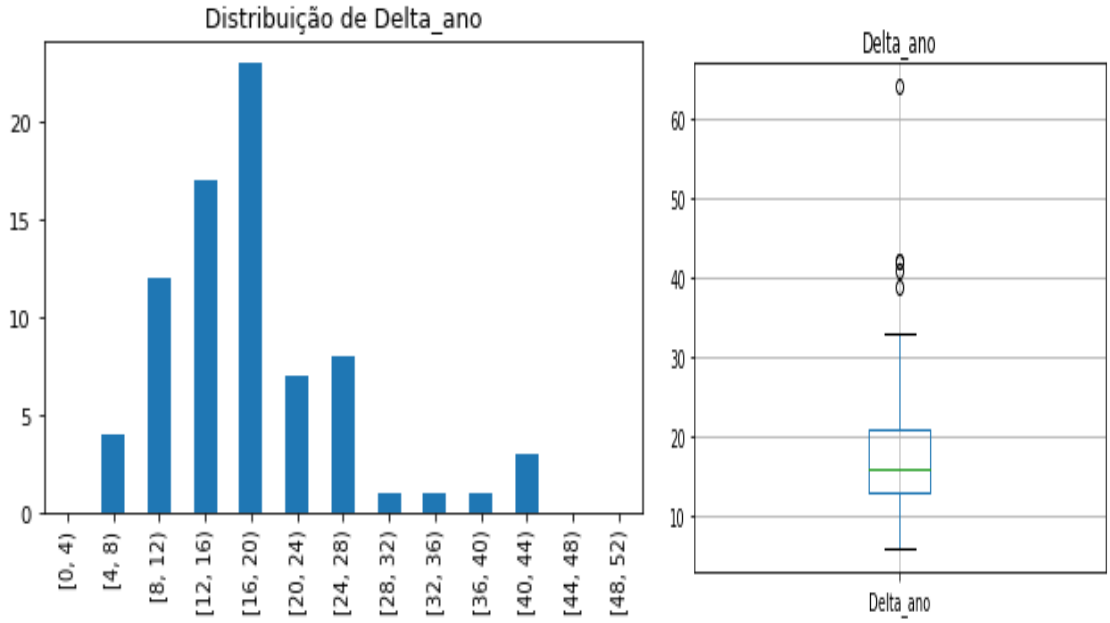
Então fiz a filtragem por curso, removi a “sujeira” dos dados e reindexei meu dataframe;

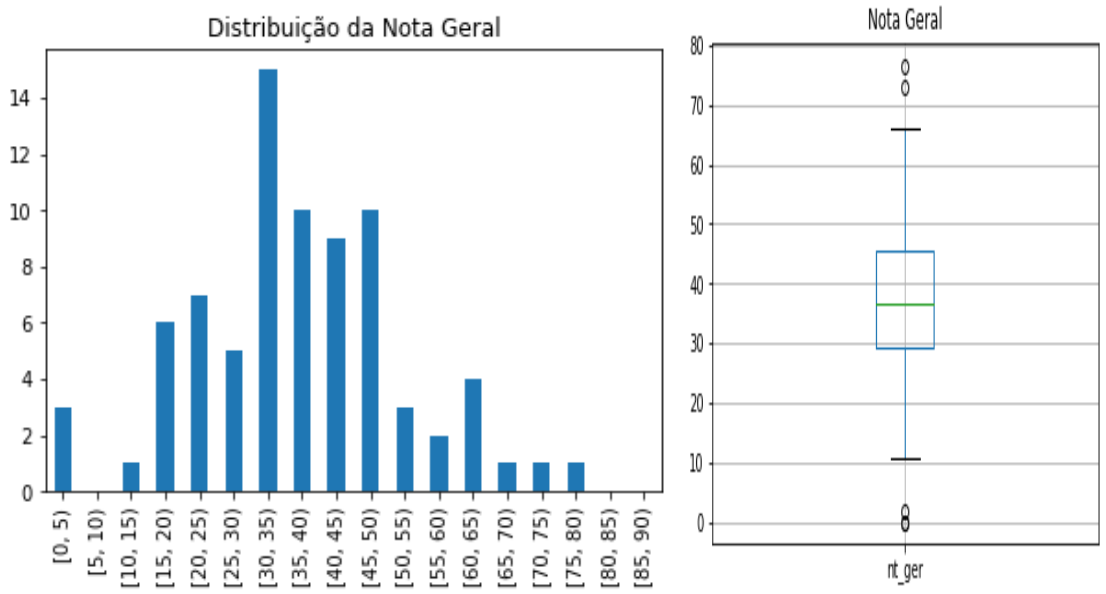
```
### Filtrando df pelo código do curso de Ciência da Computação-UFF ###
filtro_curso_ICUFF = 12710
df = df[(df.co_curso == filtro_curso_ICUFF)]
df = df.dropna()
```

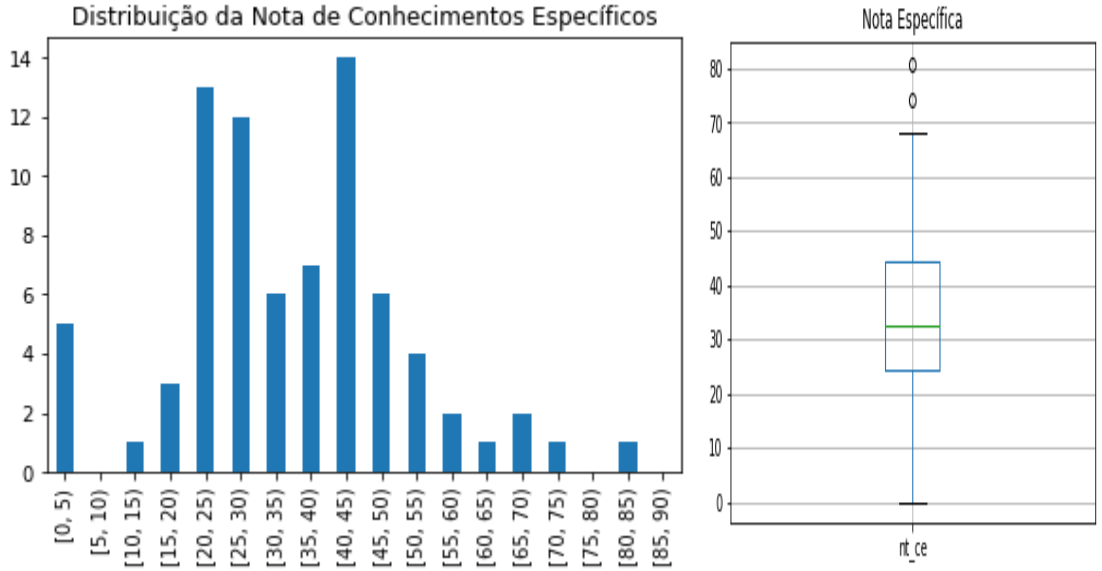
e criamos uma variável chamada ‘Delta_ano’. A finalidade dessa variável e avaliar a sequência de aprendizado do estudante. Para isso adotamos os seguintes critérios: quanto tempo ele demorou para sair do ensino médio e ingressar na universidade, há quanto tempo ele terminou o ensino médio e quanto tempo ele levou para chegar ao final do curso de ciência da computação. Acredito que o menor valor dessa variável deve ser atribuído a estudantes que tiveram facilidade de aprendizado tanto no ensino médio como na graduação, logo há uma expectativa de que ele vá melhor no exame do enade.

```
df['Delta_ano'] =  df['ano_in_gra'] - df['ano_fim_2g']  + df['nu_idade'] + 2010 - df['ano_fim_2g'] + 2011 - df['ano_in_gra'] – 22
```

Na segunda etapa coletei as primeiras amostras para análise e ver como os dados estão distribuídos. Primeiro esbocei o histograma das seguintes variáveis: ‘Delta_ano’ , ‘nt_ger’ e ‘nt_ce’. Abaixo mostro o histograma e o boxplot respectivamente dessas variáveis e o trecho de código utilizado.





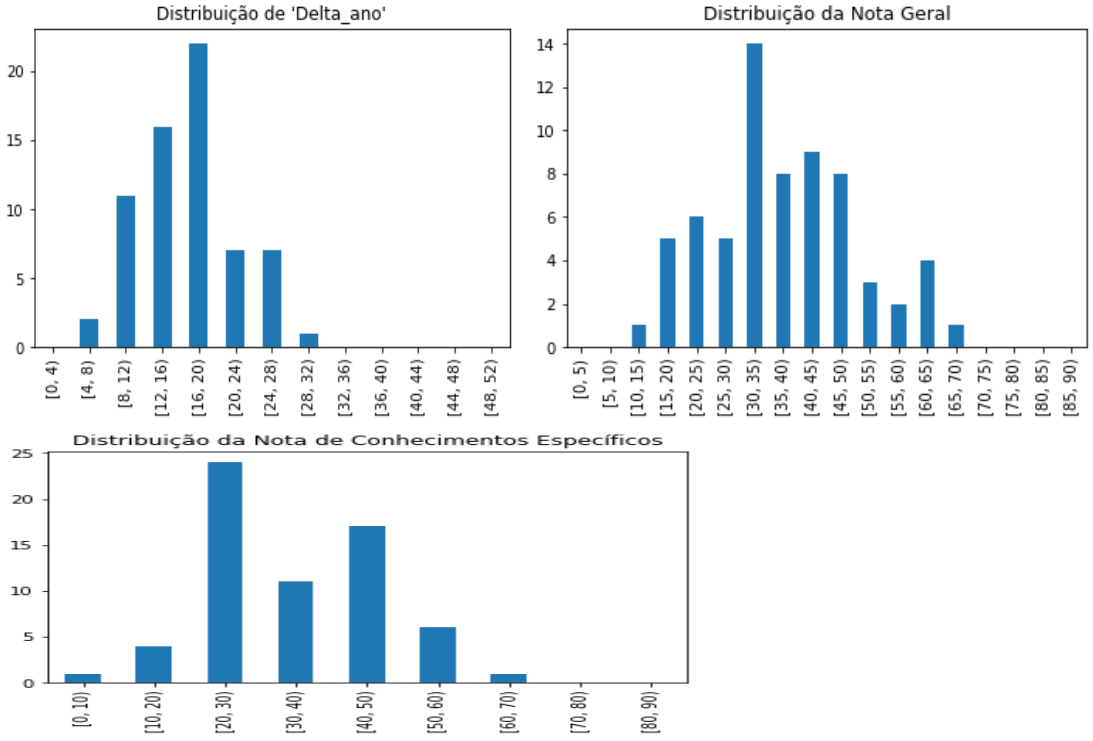


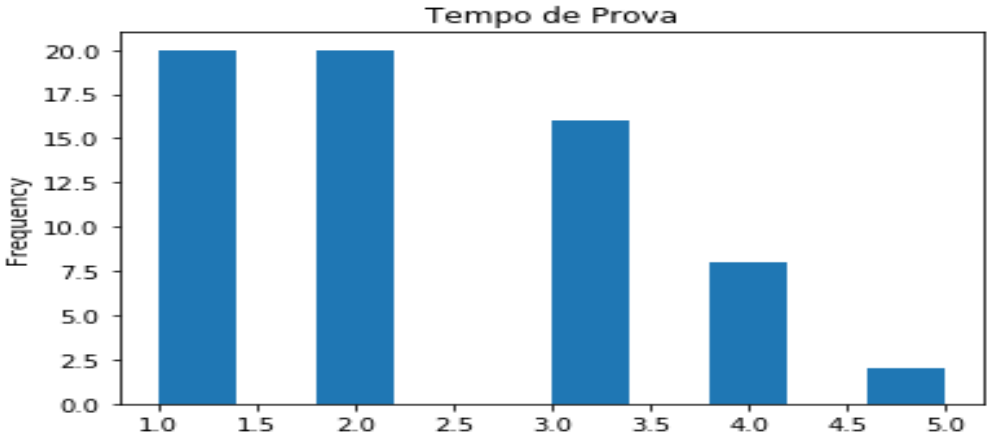

```
from scipy import stats
step = 4
bin_range = np.arange(0, 50+step, step)
out , bins = pd.cut(df['Delta_ano'],bins=bin_range ,include_lowest=True, right=False, retbins=True)
out.value_counts(sort=False).plot.bar(title = "Distribuição de 'Delta_ano'")
### HISTOGRAMA DAS NOTAS ####
# Nota Geral
step = 5
bin_range = np.arange(0, 90+step, step)
out , bins = pd.cut(df['nt_ger'],bins=bin_range ,include_lowest=True, right=False, retbins=True)
out.value_counts(sort=False).plot.bar(title = "Distribuição da Nota Geral")
```

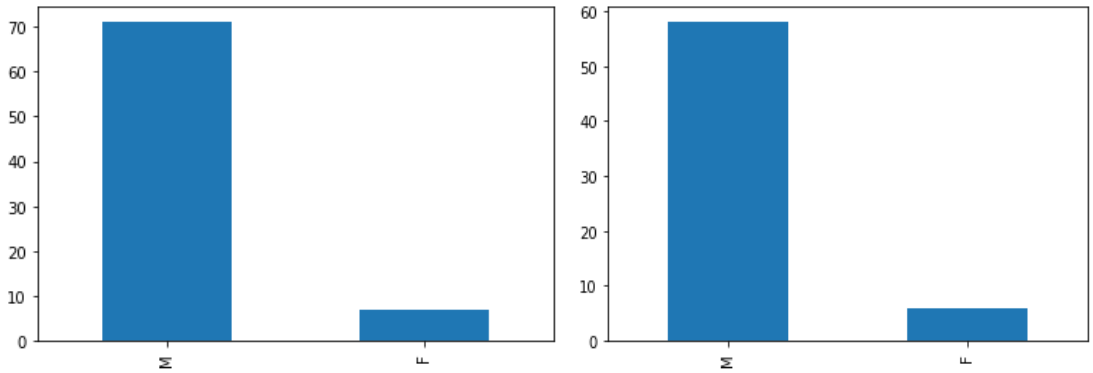
```
# Nota de Conhecimentos Específicos
step = 5
bin_range = np.arange(0, 90+step, step)
out , bins = pd.cut(df['nt_ce'],bins=bin_range ,include_lowest=True, right=False, retbins=True)
out.value_counts(sort=False).plot.bar(title = "Distribuição da Nota de Conhecimentos Específicos")
df.boxplot(column='Delta_ano').set_title("Delta_ano")
df.boxplot(column='nt_ce').set_title("Nota Específica")
df.boxplot(column='nt_ger').set_title("Nota Geral")
```

Então de posse da primeira amostragem de dados foi necessário fazer um novo tratamento de dados e passei o df atual para um df_tratado. O número de estudantes foi reduzido de 78 para 66. Abaixo segue o código e o resultado após o tratamento.

```
df_tratado = df[((df.nt_ger - df.nt_ger.mean()) / df.nt_ger.std()).abs() < 2.2]
print('##############################')
df_tratado = df_tratado[((df_tratado.nt_ce - df_tratado.nt_ce.mean()) / df_tratado.nt_ce.std()).abs() < 3]
print('##############################')
df_tratado = df_tratado[((df_tratado.Delta_ano - df_tratado.Delta_ano.mean()) / df_tratado.Delta_ano.std()).abs() < 1.5]
print('##############################')
df_tratado = df_tratado[((df_tratado.nu_idade - df_tratado.nu_idade.mean()) / df_tratado.nu_idade.std()).abs() < 2.8]
df_tratado.index = range(len(df_tratado))a = df_tratado['tp_sexo'].value_counts()
```





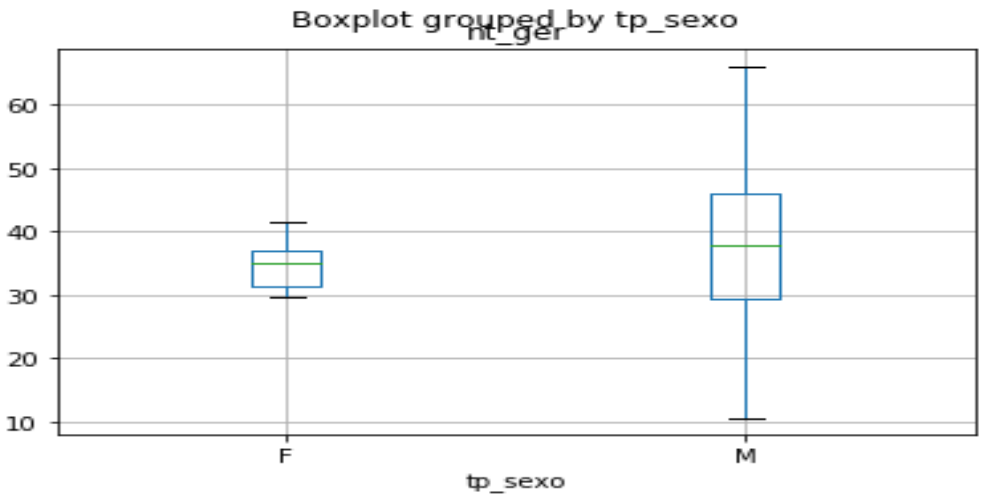


Estudantes por Sexo Antes e Depois

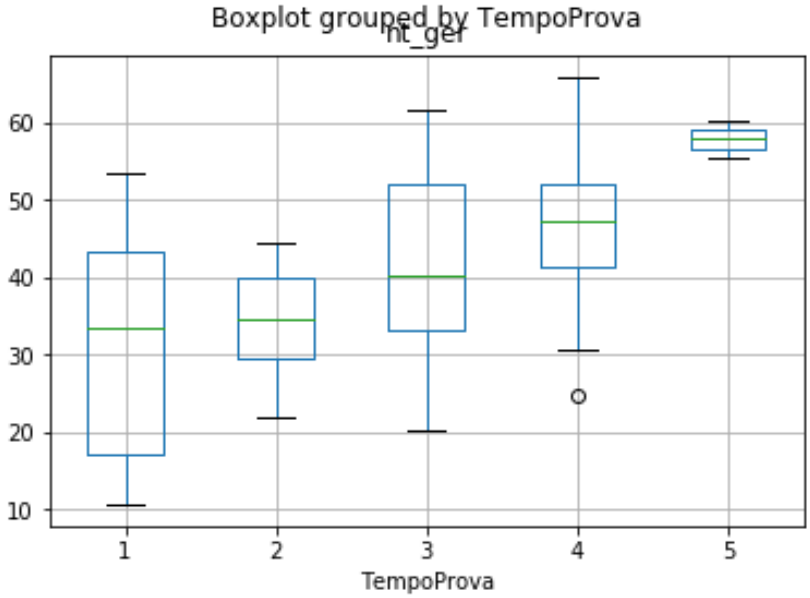
Antes: Masculino(M) = 71 , Feminino (7) = 7 => F/M ~ 9,86%

Depois:Masculino(M) =60, Feminino(F) = 6 => F/M ~ 10%

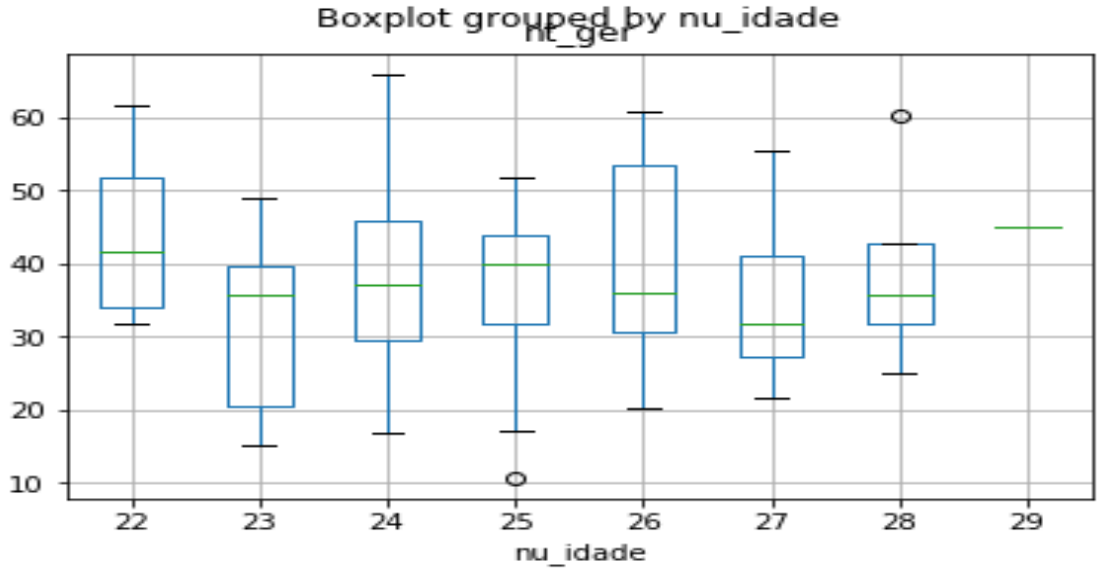
Adicionamos a coluna tempo de prova e após vimos que a proporção por genêro permaneceu inalterada com uma pequena diferença de 0,14%. Já nos novos gráficos mostrados pode se perceber a remoção dos valores extremos.

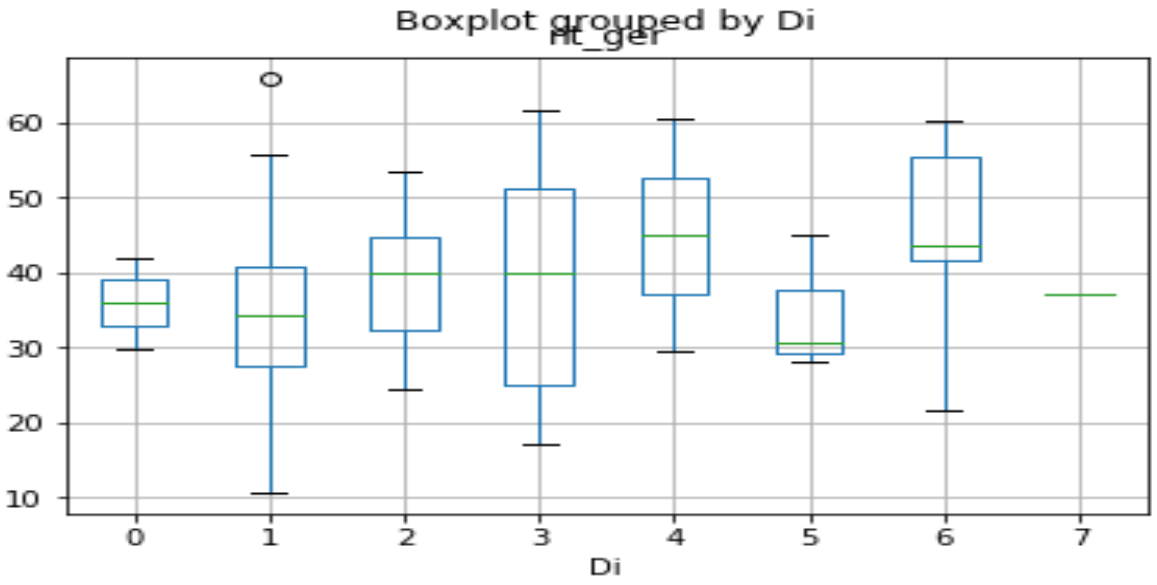


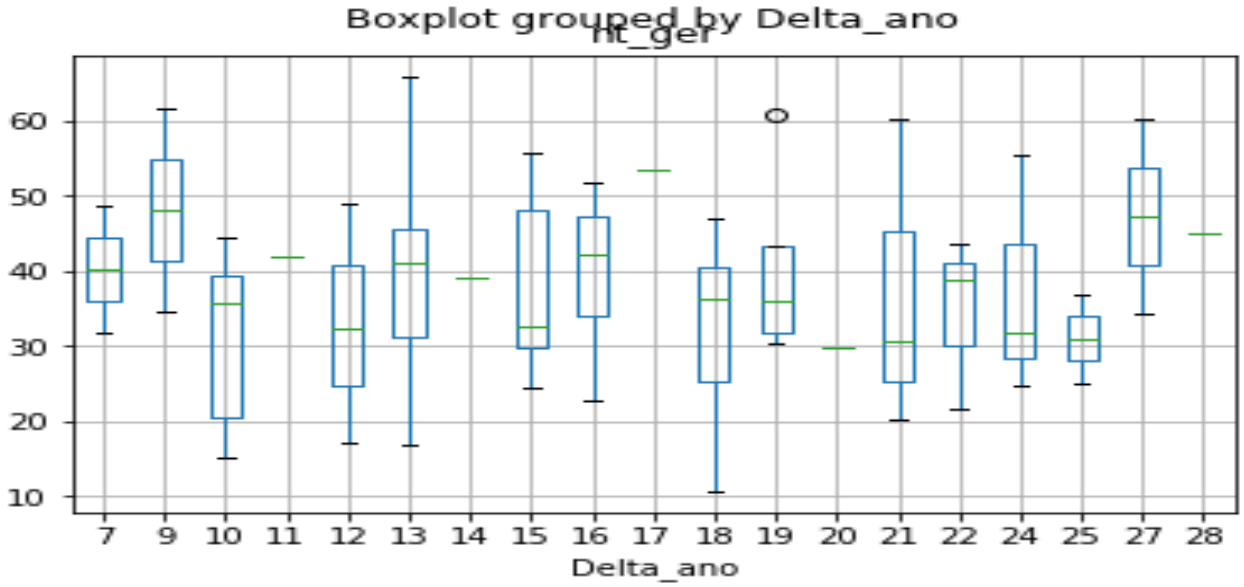
Não teve uma diferença de desempenho grande entre o gênero levando em conta que os homens formam ~90% da amostra.

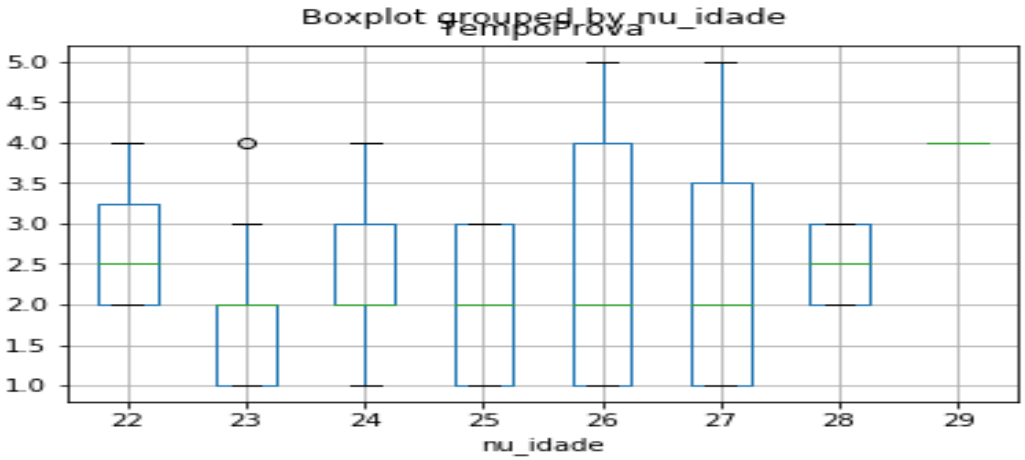


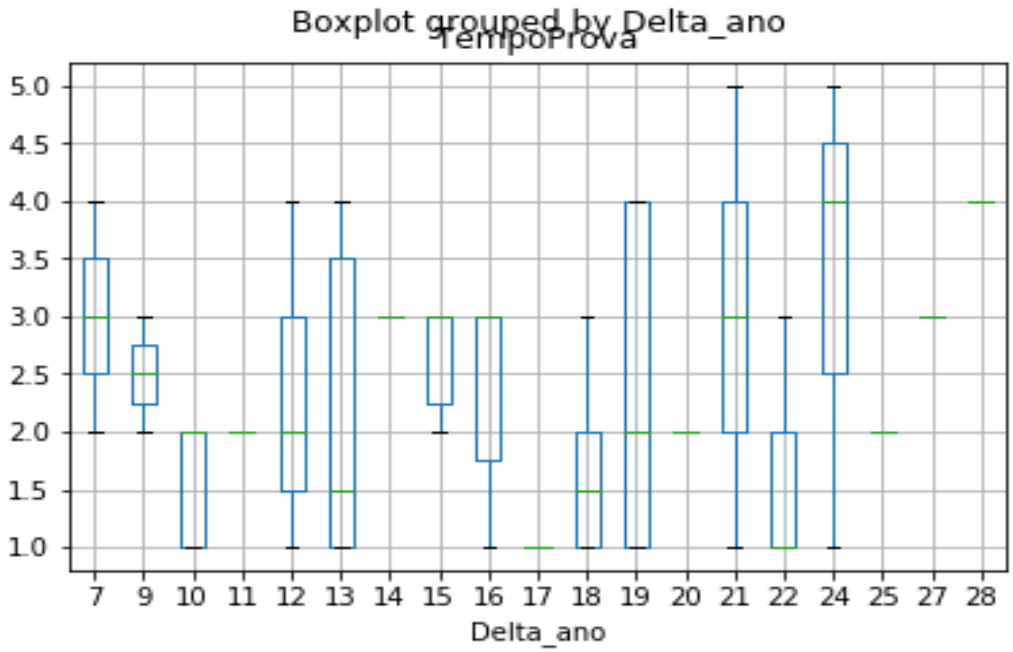
No eixo tempo de prova temos 1 para quem acabou a prova em até 1h, 2 para que realizou entre 1h e 2h, 3 para quem realizou a prova entre 2h e 3h e 4 para quem realizou a prova entre 3h e 4h e 5 para quem realizou a prova por 4 horas e não conseguiu terminar todas as questões. Se observarmos a mediana e os valores extremos o resultado foi próximo do esperado, pois com maior o tempo de prova melhor foi a nota. Tivemos uma “anomalia” que foi no tempo 4 com 1 outlier.




Boxplot grouped by Di
nt_ger
60
50
40
30
20
10
0
1
2
3
4
5
6
7
Di


Boxplot grouped by Delta_ano
nt_ger
60
50
40
30
20
10
7
9
10
11
12
13
14
15
16
17
18
19
20
21
22
24
25
27
28
Delta_ano


Boxplot grouped by nu_idade
TempoProva
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
22
23
24
25
26
27
28
29
nu_idade



Acima temos 5 gráficos em sequência , 3 primeiros mostrando o desempenho por idade, por Delta_Ano e por Di(Tempo que levou pra sair do Ensino Médio e Ingressar na Universidade). Aqui era esperado que quanto o menor o valor da varável no eixo x (x vs nt_ger) melhor deveria ser a nota. Olhando para os coeficientes de Pearson e Spearman vi que a relação entre essas variáveis é fraca ou muito o que explica a dispersão da amostra.

stats.spearmanr(df_tratado['nt_ger'],df_tratado['Delta_ano'])

SpearmanrResult(correlation = -0.0043168329134158296, pvalue=0.97255803088483805)

stats.spearmanr(df_tratado['nt_ger'],df_tratado['nu_idade'])

SpearmanrResult(correlation = 0.025407161142960533, pvalue=0.83952771528419867)

stats.spearmanr(df_tratado['nt_ger'],di['Di'])

SpearmanrResult(correlation = 0.17715657384937544, pvalue=0.15473287882146752)

stats.pearsonr(df_tratado['nt_ger'],di['Di'])

(0.18049535063621658, 0.14698173667925332)

stats.pearsonr(df_tratado['nt_ger'],df_tratado['Delta_ano'])

(0.037609653393199487, 0.76432173013034366)

stats.pearsonr(df_tratado['nt_ger'],df_tratado['nu_idade'])

(0.048575887917606697, 0.69851852959003513)

Entretanto, quando fazemos um contraste do desempenho com o tempo de duração da prova já temos valores mais satisfatórios;

stats.spearmanr(df_tratado['nt_ger'],df_tratado['TempoProva'])

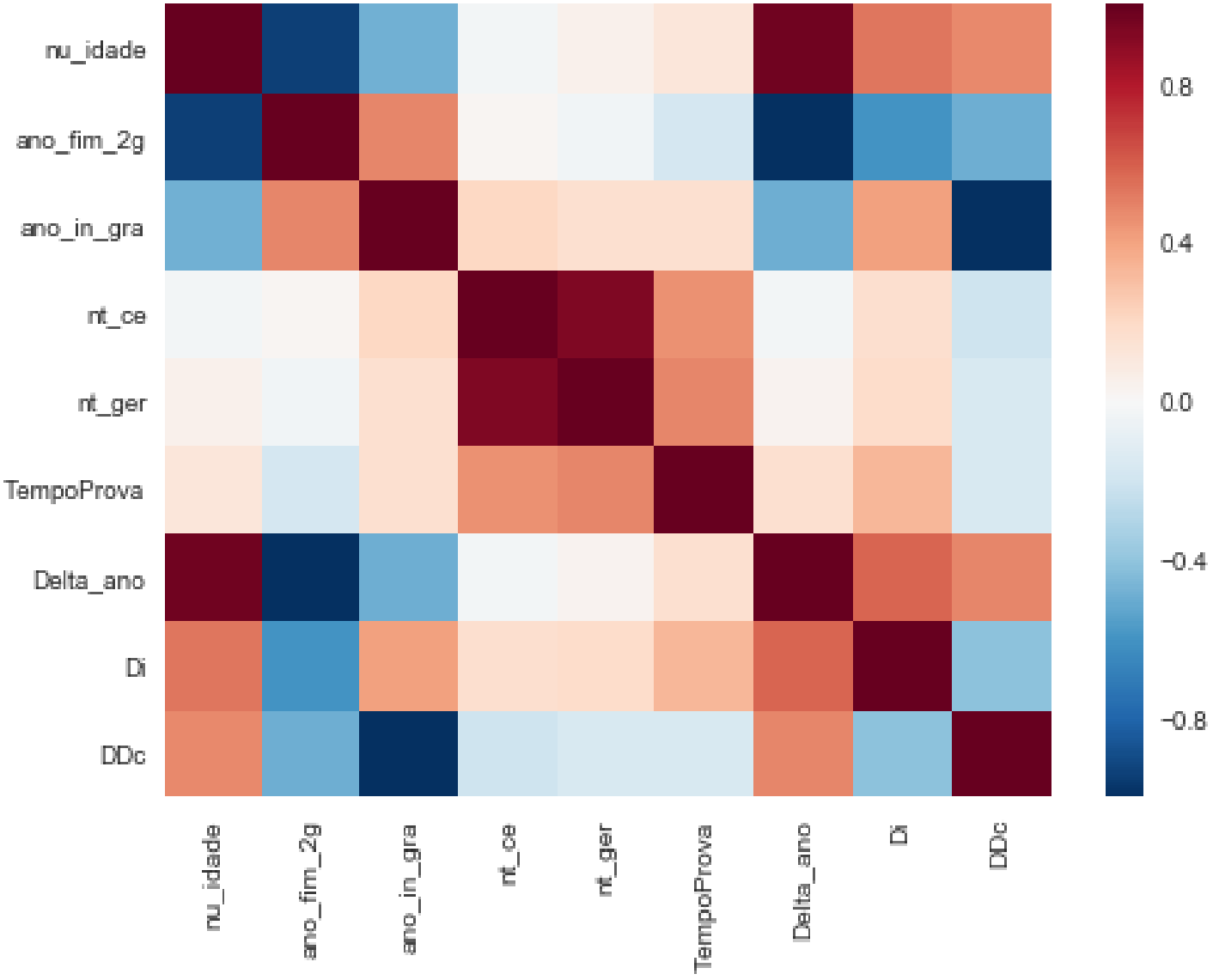
SpearmanrResult(correlation = 0.4162417057755286, pvalue=0.00050870215094906172)

stats.pearsonr(df_tratado['nt_ger'],df_tratado['TempoProva'])

Out[553]: (0.43443833756345335, 0.00026767261960761039)

pois o estudante que ficou com o 5 e porque teve dificuldade de fazer a prova inteira o que justifica um desempenho menor de quem fez em 4, já o 1 poderia apresentar um pequeno aumento em relação ao 2, pois esse período está compreendido entre 1 e 2 horas e 2 entre 2 e 3 horas, sendo 2 horas um tempo razoável de prova.

Por fim plotei a matriz de correlação para dar uma visão geral da análise e verificar se o único fator “relevante” proposto no nosso objetivo de análise foi realmente o resultado encontrado até o momento.



Da matriz observa-se que:

nu_idade só tem relação forte com Delta_ano, pois e uma variável de composição da mesma, e também possui relação com o intervalo de tempo entre o Ensino Médio e a UFF (Di) e a duração no curso de graduação (DDc) e influência quase nula na nossa variável de desempenho nt_ger.

Ano_fim_2g e ano_in_gra são variáveis complementares na análise e serviram para compor outras variáveis temporais.(Di, DDc, Delta_ano)Entretanto, Ano_fim_2g tem uma fraca relação com Di.

Delta_ano: continua expressando sua fraca relação com a nota geral.

Di apresenta alguma relação com o tempo de prova, pode ser pelo fato de ter passado por um exame similar como o vestibular a menos tempo.

Nt_ce acabou como um dado optativo, caso Delta_ano desse o retorno que eu esperava eu iria olhar para a diferença de desempenho geral e específico em relação ao Delta_ano.

4. CONSIDERAÇÕES FINAIS

O tratamento de dados melhorou nossa análise? Sim , olhando os dados obtidos antes do tratamento e após o tratamento temos que os valores ficaram coesos, e os respectivos desvios diminuíram, tive uma redução de 12 elementos.

	nu_idade	ano_fim_2g	ano_in_gra	nt_ce	nt_ger	TempoProva
count	66.000000	66.000000	66.000000	66.000000	66.000000	66.000000
mean	24.848485	2003.757576	2006.000000	35.010606	37.381818	2.272727
std	1.684704	1.745858	1.539231	13.603813	12.617764	1.116939
min	22.000000	2000.000000	2003.000000	4.100000	10.600000	1.000000
25%	24.000000	2003.000000	2005.000000	24.300000	29.975000	1.000000
50%	25.000000	2004.000000	2006.000000	34.450000	37.050000	2.000000
75%	26.000000	2005.000000	2007.000000	43.025000	45.000000	3.000000
max	29.000000	2007.000000	2009.000000	67.900000	66.000000	5.000000

	Delta_ano
count	66.000000
mean	16.333333
std	5.106055
min	7.000000
25%	13.000000
50%	16.000000
75%	19.000000
max	28.000000

<< Tratado

	nu_idade	ano_fim_2g	ano_in_gra	nt_ce	nt_ger
count	78.000000	78.000000	78.000000	78.000000	78.000000
mean	25.525641	2003.153846	2006.000000	34.433333	36.666667
std	3.206019	3.112926	1.690309	16.450577	15.273572
min	21.000000	1988.000000	2002.000000	0.000000	0.000000
25%	24.000000	2002.000000	2005.000000	24.300000	29.200000
50%	25.000000	2004.000000	2006.000000	32.400000	36.500000
75%	26.000000	2005.000000	2007.000000	44.175000	45.600000
max	41.000000	2007.000000	2011.000000	80.800000	76.600000

	TempoProva	Delta_ano
count	78.000000	78.000000
mean	2.307692	18.217949
std	1.143107	9.378959
min	1.000000	6.000000
25%	1.000000	13.000000
50%	2.000000	16.000000
75%	3.000000	21.000000
max	5.000000	64.000000

<< Primário

De acordo com a análise feita o único fator relevante que foi analisado que tem uma relação com o desempenho do estudante foi o tempo de prova, os outros dados propostos no nosso objetivo não foram relevantes para analisar o desempenho do estudante. Por quê eu tinha uma expectativa com Delta_ano?

A ideia do Delta_ano seguiu a linha de raciocínio, se o estudante foi bom aluno no ensino médio e conseguiu obter aprovação direta do EM para o Ensino superior e na sequência chega ao fim do curso no período correto, criou-se então a expectativa de que o estudante nesse perfil teria maior probabilidade de ir bem na prova do ENADE. Mas não foi possível chegar a essa conclusão apenas uma pequena parte da amostra seguiu esse padrão, logo não é relevante.

Quais outros fatores poderiam ser analisados? Condições socioeconômicas? Desempenho do aluno na graduação? Desempenho do aluno no Ensino Médio?

5. REFERÊNCIAS

BRASIL, Lei nº 10.861, de 14 de abril de 2004. Institui o sistema nacional de avaliação da educação superior – SINAES e dá outras Providências. Brasília,14abril.2004.Disponívelem:<http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm>

Femi Anthony – Mastering Pandas 2015

https://pandas.pydata.org/

http://dados.gov.br/dataset/microdados-do-exame-nacional-de-desempenho-de-estudantes-enade

https://stackoverflow.com/questions/

MIRANDA, G. J. Relações entre as qualificações do professor e o desempenho discente nos cursos de graduação em Contabilidade no Brasil. 2011. Tese de Doutorado. Tese de doutorado em Ciências Contábeis, Programa de Pós-Graduação em Ciências Contábeis, Departamento de Contabilidade e Atuária, FEA/USP, São Paulo, SP, Brasil.

Roxy Peck, “Statistics: Learning from Data” 2014