UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO
APRESENTAÇÃO E ANÁLISE ESTATÍSTICA DE DADOS - IMD 0184 -T01
Discentes: João Paulo de Oliveira Câmara Fernandes
Nathália Kênia Cabral Justino
Docente: Ismênia Magalhães

ATIVIDADES DA SEMANA 1 E 2

Tema – Qual o seu tema de estudo?
Avaliação de docentes da UFRN.

Problema – O que você pretende responder com sua análise?
A baixa avaliação de alguns docentes da UFRN.

Justificativa – Por que isso é relevante?
Tomar medidas corretivas e preventivas junto aos docentes com baixo índice de avaliação, visando uma melhoria na educação da instituição.

Objetivo geral e Objetivos Específicos – Que objetivos você pretende atingir?
Objetivo geral - melhorar o ensino e a educação da instituição UFRN. Objetivo específico - selecionar docentes com qualificação abaixo do desejado e fazer uma reciclagem com esse profissional, indicando caminhos para uma melhor abordagem de conteúdo lecionado.

Hipóteses – Que suposições guiam o seu trabalho?
● Os professores não são preparados adequadamente para dar aulas. ● Alguns professores formados em outra época não detêm o conhecimento sobre novas tecnologias educacionais para manter os alunos motivados a aprender. ● Professores desmotivados não conseguem passar para o aluno o conteúdo programado de forma satisfatória.

Banco de dados – Onde você pode encontrar dados a respeito?

Portal Brasileiro de Dados Abertos https://dados.gov.br/dataset/avaliacoes-de-docencia

Banco de dados – Qual base de dados você escolheu para tratar o seu problema?

Base de dados do Portal Brasileiro de Dados Abertos com frequência de atualização semestral, os dados estão acumulados desde o ano de 2013. Estes podem ser obtidos em https://dados.gov.br/dataset, filtrando pela organização “Universidade Federal do Rio Grande do Norte". Optamos por esta base pois ela possui dicionário de dados que mostra a descrição e obrigatoriedade de todos os campos; e informações adicionais sobre o ano, a turma, o período e a quantidade de discentes que essa turma tinha, o que facilita a identificação de variáveis que possam colaborar numa sugestão de solução para o problema.

Variável resposta – Qual variável você acredita que pode te ajudar a responder suas perguntas?

Atuação Profissional - Variável do tipo quantitativa. A variável explícita da nota que os discentes atribuíram ao professor naquele semestre naquela disciplina. É um indicador que o professor pode melhorar sua didática.

Resumo da sua variável resposta – Como é que se comporta a sua variável?

Menor valor – 0.5 pontos; Maior valor – 9.99 pontos. A média foi de 9.14. Desvio padrão de 0.75 pontos.

Traduzindo – O que tudo isso quis dizer?

Apesar de a média ser alta, existem notas bem próximas a zero, o que significa dizer que existem professores que não estão se esforçando o mínimo para dar aulas e colaborar com o aprendizado do discente. Algo precisa ser feito para mitigar isso.

Variáveis Explicativas – Que outras variáveis você pode dispor para te ajudar na análise?

- Quantidade de discentes pode interferir no resultado do professor
- Postura profissional, quanto maior, provavelmente maior será sua atuação profissional
- Autoavaliação do aluno influencia, de forma que a baixa dedicação do aluno pode resultar em uma reprovação, gerando uma avaliação negativa ao professor

ATIVIDADE DA SEMANA 2

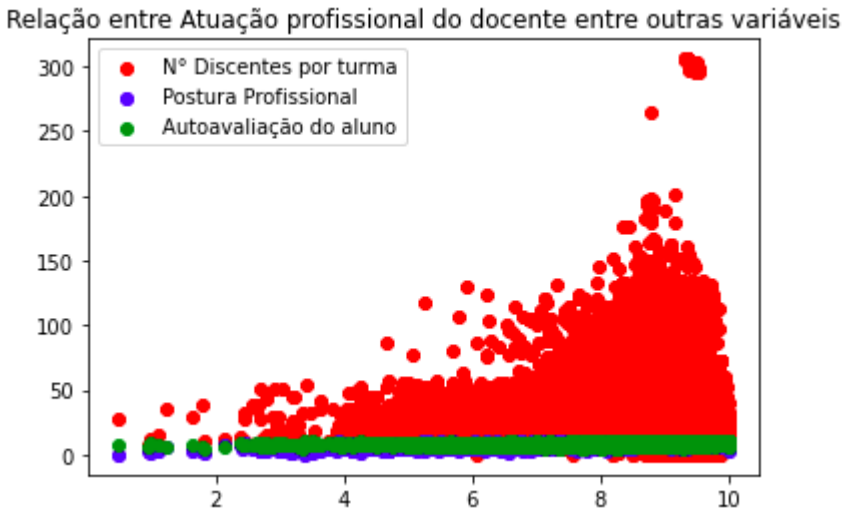
Escolhemos um conjunto de 3 variáveis para estudar e avaliar se está ou não relacionada com o estudo que vem sendo feito, que no caso, é a avaliação da atuação profissional do docente. Todas as variáveis são do tipo quantitativa, são elas: quantidade de discentes na turma, postura profissional do docente e autoavaliação do aluno na disciplina..

```
In [23]: dados1.describe()
```

Out[23]:

	qtd_discentes	postura_profissional_media	atuacao_profissional_media	autoavaliacao_aluno_media
count	118158.000000	118158.000000	118158.000000	102783.000000
mean	22.174495	9.452506	9.146739	9.083089
std	17.003041	0.682489	0.754393	0.588967
min	0.000000	0.340000	0.500000	4.340000
25%	9.000000	9.290000	8.830000	8.710000
50%	18.000000	9.670000	9.330000	9.120000
75%	31.000000	9.880000	9.670000	9.540000
max	306.000000	10.000000	10.000000	10.000000

Abaixo o comportamento gráfico das 3 variáveis:



A seguir, apresentamos o resultado da correlação r:

```
In [30]: r1= dados.atuacao_profissional_media.corr(dados.qtd_discentes)

         r2= dados.atuacao_profissional_media.corr(dados.postura_profissional_media)

         r3= dados.atuacao_profissional_media.corr(dados.autoavaliacao_aluno_media)
```

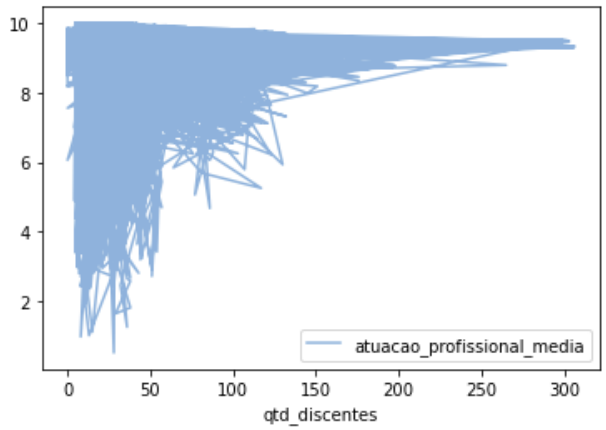
```
In [31]: print(r1)
         print(r2)
         print(r3)
```

```
-0.19827627762727265
0.687586435561898
0.594033681499128
```

A quantidade de alunos não têm uma relação forte com a avaliação atribuída ao docente, mas podemos concluir que temos uma pequena relação de quanto mais alunos na turma, menor é avaliação da atuação profissional do docente. Supõe-se que isso acontece pela maior dificuldade de gerir uma turma com mais alunos. Segue o gráfico que mostra essa suposição.

```
dados1.plot(x= 'qtd_discentes', y='atuacao_profissional_media', alpha=0.5)
```

```
<AxesSubplot:xlabel='qtd_discentes'>
```



A postura profissional do docente apresentada durante as aulas pode ser considerada uma boa relação com a atuação profissional do mesmo.

A autoavaliação do aluno na disciplina chega a ser uma boa relação com a avaliação da atuação do docente. Quanto maior a dedicação na disciplina, maior o resultado, e menores são as queixas com o docente. Ao mesmo tempo, também podemos concluir que quanto pior for a atuação profissional do docente durante as aulas, menor vai ser o interesse pela disciplina por parte do aluno, resultando em uma autoavaliação com notas mais baixas.

Também foi gerada a matriz de correlação de Pearson para algumas colunas da tabela para uma análise mais completa.

```
dados1.corr(method='pearson')
```

	qtd_discentes	postura_profissional_media	atuacao_profissional_media	autoavaliacao_aluno_media
qtd_discentes	1.000000	-0.132268	-0.198276	-0.359126
postura_profissional_media	-0.132268	1.000000	0.687586	0.489315
atuacao_profissional_media	-0.198276	0.687586	1.000000	0.594034
autoavaliacao_aluno_media	-0.359126	0.489315	0.594034	1.000000

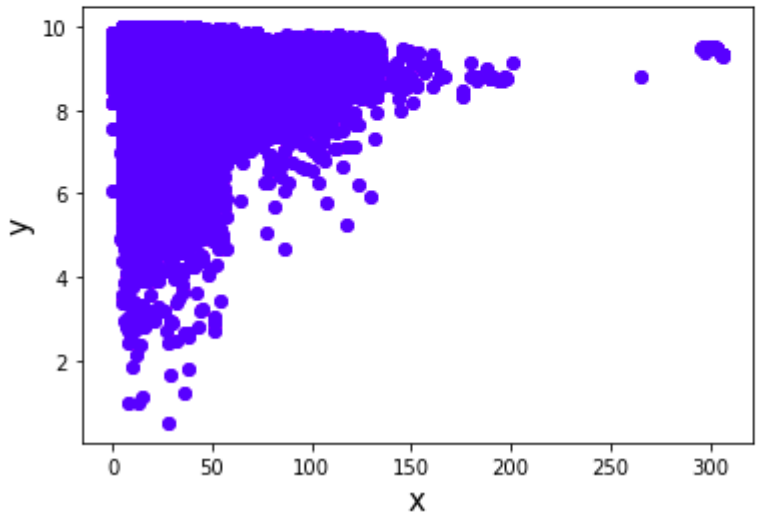
REFERÊNCIA

Pandas - Data Correlations. Disponível em <https://www.w3schools.com/python/pandas/pandas_correlations.asp>. Último acesso em 14/01/2022.

Atividade 3

Vamos fazer o estudo das relações entre a variável “atuação profissional” do docente em relação à “quantidade de alunos por turma” que esse docente dá aula. Definimos como o eixo Y sendo a média da atuação profissional do docente enquanto nosso eixo X será a quantidade de alunos por turma.

```
5]: # define os dados
    x = np.array(dados1["qtd_discentes"])
    y = np.array(dados1["atuacao_profissional_media"])
    plt.plot(x, y, 'bo')
    plt.xlabel("x", fontsize = 15)
    plt.ylabel("y", fontsize = 15)
    plt.show(True)
```



Vamos aplicar o modelo de regressão linear simples para fazer o estudo dessas variáveis. Podemos supor um modelo de regressão linear simples, como:

$$y_i \approx \beta_0 + \beta_1 x_i$$

Como há outros fatores, além de xi que afetam os valores de yi, podemos escrever:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

onde ϵi é uma variável aleatória que indica o erro na aproximação. O objetivo dos métodos de regressão é encontrar o melhor valor de β0 e β1 que minimizem o erro no ajuste. Ou seja, queremos encontrar a linha no plano x−y que melhor se ajusta aos dados observados. Estimando os coeficiente através do método dos momentos ou dos mínimos quadrados, obtemos:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - x)^2} = \frac{Sxy}{Sxx} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Vamos implementar uma função para realizar a estimação.

```
:
from statistics import variance
import math

def estimate_coef(x, y):
    # número de observações/pontos
    n = np.size(x)

    # médias de x e y
    m_x, m_y = np.mean(x), np.mean(y)

    # calculating cross-deviation and deviation about x
    SS_xy = np.sum(y*x) - n*m_y*m_x
    SS_xx = np.sum(x*x) - n*m_x*m_x

    # calcula os coeficientes de regressão
    b_1 = SS_xy / SS_xx
    b_0 = m_y - b_1*m_x

    return(b_0, b_1)

# função para mostrar os dados e o ajuste linear
def plot_regression_line(x, y, b):
    # mostra os dados
    plt.scatter(x, y, color = "b", marker = "o", s = 50)

    # prediz os valores
    y_pred = b[0] + b[1]*x

    # mostra a reta de regressão
    plt.plot(x, y_pred, color = "r")

    plt.xlabel('x', fontsize = 15)
    plt.ylabel('y', fontsize = 15)
    plt.show(True)
```

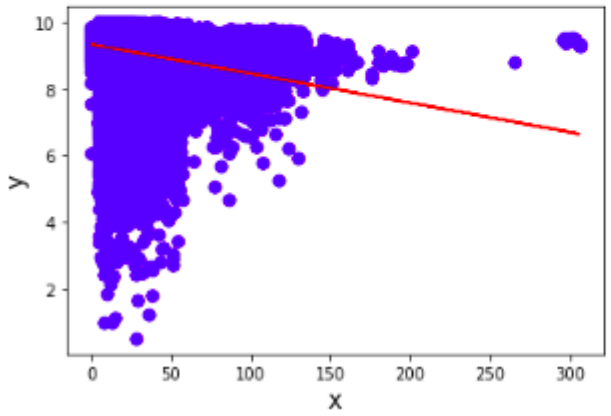
Assim, aplicando ao conjunto de dados:

```
import numpy as np

# estima os coeficientes
b = estimate_coef(x, y)
print("Estimated coefficients:\nb_0 = {}  \nb_1 = {}".format(b[0], b[1]))

# mostra o ajuste linear
plot_regression_line(x, y, b)
```

```
Estimated coefficients:
b_0 = 9.341810887561747
b_1 = -0.008797142717979265
```



Esse foi o modelo de regressão linear que mais se ajustou em nossos pontos, resultando na reta:

$$Y = 9.34 - 0.0088X$$

Já para quantificar a acurácia do modelo, usamos o erro padrão residual (residual standard error):

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

```
#funcao que calcula o RSE
def RSE(x,y,b):
    n = len(y)
    RSE = 0
    for i in range(0,n):
        y_pred = b[0]+ x[i]*b[1] # valor predito
        RSE = RSE + (y[i]-y_pred)**2
    RSE = math.sqrt(RSE/(n-2))
    return RSE
print('RSE:', RSE(x,y,b))
```

```
RSE: 0.7394182707176781
```

Outra medida importante é o coeficiente R^2, que mede a proporção da variabilidade em Y que pode ser explicada a partir de X.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad 0 \leq R^2 \leq 1$$

```
def R2(x,y,b):
    n = len(y)
    c1 = 0
    c2 = 0
    ym = np.mean(y)
    for i in range(0,n):
        y_pred = b[0]+ x[i]*b[1] # valor predito
        c1 = c1 + (y[i]-y_pred)**2
        c2 = c2 + (y[i]-ym)**2
    R2 = 1 - c1/c2
    return R2

print('R2:', R2(x,y,b))
```

```
R2: 0.03931348226985931
```

Quanto mais próximo de um, melhor é o ajuste da regressão linear. Como o resultado deu mais próximo a zero, podemos concluir que o ajuste da regressão linear não se adequou tão bem aos nossos dados.

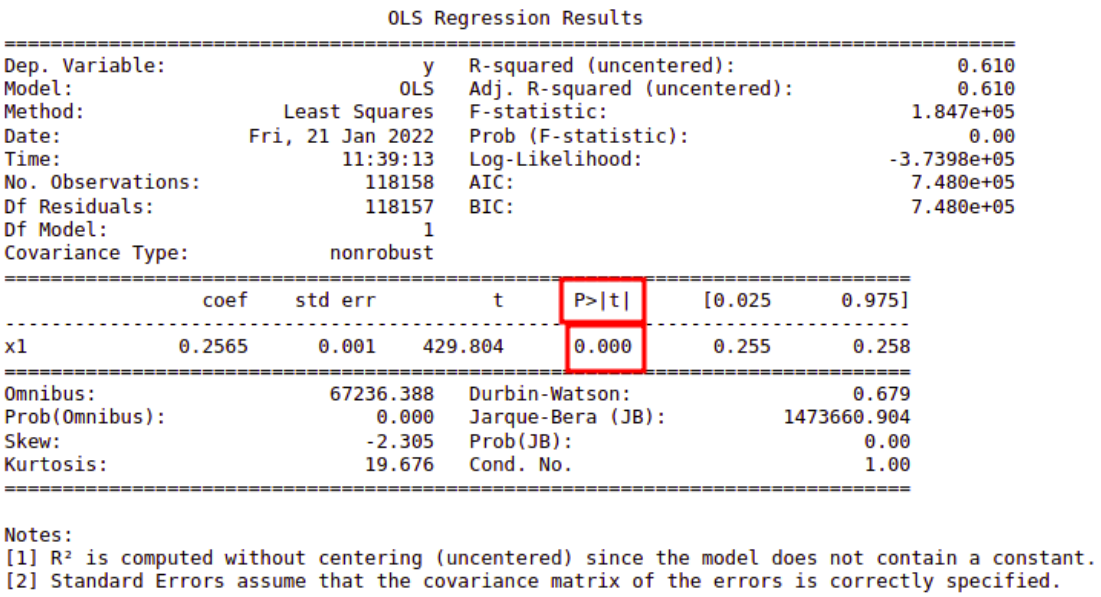
Testando as Hipóteses

- H0: Não há relação entre X e Y. H0: b1 = 0
- Ha: Há alguma relação entre X e Y. H0: b1 != 0

Para fazer o teste de hipóteses usamos a biblioteca “statsmodels” do python para imprimir o sumário dessa análise.

```
]: import statsmodels.api as sm
   est = sm.OLS(y, x)
   est2 = est.fit()
   print(est2.summary())
```

Nosso sumário:



OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.610
Model:	OLS	Adj. R-squared (uncentered):	0.610
Method:	Least Squares	F-statistic:	1.847e+05
Date:	Fri, 21 Jan 2022	Prob (F-statistic):	0.00
Time:	11:39:13	Log-Likelihood:	-3.7398e+05
No. Observations:	118158	AIC:	7.480e+05
Df Residuals:	118157	BIC:	7.480e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.2565	0.001	429.804	0.000	0.255	0.258

Omnibus:	67236.388	Durbin-Watson:	0.679
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1473660.904
Skew:	-2.305	Prob(JB):	0.00
Kurtosis:	19.676	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Como P > |t| deu praticamente zero, a gente rejeita a hipótese nula, ou seja, b1 > 0. Sendo assim, podemos concluir que há sim alguma relação entre X e Y, que no nosso caso, foram a quantidade de alunos por turma e a média da atuação profissional do docente naquela turma.