**Estatística Com Python**

**Parte I – Frequências e Medidas:**

Todos os arquivos serão baixados, mas deixo o link do drive de qualquer forma: <https://drive.google.com/drive/folders/1MoAf3dLinpNBXVGIIyetaY5qfSwZ31Gt>.

1. **Aula 1 – Qual Seu Tipo de Dado:**
   1. Dependendo do tipo de dados análises diferentes serão feitas. Importante saber quais são os tipos.
      1. Quantitativos: Expressa quantidade de dados. Contagem, mensurações. Ex.:
         1. Idade;
         2. Renda;
         3. Altura.
      2. Qualitativos: Expressa qualidade dos dados. Ex.:
         1. UF;
         2. Sexo;
         3. Cor;
         4. Anos de Estudo.
      3. Podemos dividir os qualitativos em ordinais e nominais, e os quantitativos em discretos e contínuos.
   2. O que aprendemos:
      1. A identificar a versão de uma biblioteca do Python;
      2. A ler um dataset no formato CSV e criar um DataFrame pandas, com o conteúdo deste dataset;
      3. A identificar e classificar as variáveis de um dataset como quantitativas ou qualitativas;
      4. Que as variáveis qualitativas se dividem em ordinais e nominais;
      5. Que as variáveis quantitativas se dividem em discretas e contínuas.
2. **Aula 2 – Distribuição de Frequência:**
   1. S.value\_counts(): conta a quantidade de vezes que uma variável aparece na series. Podemos passar o parâmetro normalize = True, onde ele devolve a porcentagem de cada variável, onde ele soma elas e divide cada um pela soma:

dados['Sexo'].value\_counts(normalize=True) \* 100



* + 1. Podemos usar a porcentagem e a frequência para criar um df de distribuição e frequência qualitativas:

dist\_freq\_qualitativas = pd.DataFrame({'Frequência': frequencia, 'Porcentagem (%)': percentual})



* + 1. Df.rename\_axis(‘nome\_index’, axis=’columns’/’rows’, inplace=True): Renomeia a coluna do index:

dist\_freq\_qualitativas.rename(index= {0: 'Masculino', 1: 'Feminino'}, inplace=True)

dist\_freq\_qualitativas.rename\_axis('Sexo', axis='columns', inplace=True)

dist\_freq\_qualitativas



* 1. Pd.crosstab(variável\_linha, variável\_coluna): Cria um df onde faz o cruzamento dos dados da linha e da coluna passadas. No caso utilizamos cor e sexo, então ele irá mostrar na tabela quantas pessoas do sexo masculino tiveram de cada cor, bem como os femininos:

sexo = {0: 'Masculino',

        1: 'Feminino'}

cor = {0: 'Indígena',

        2: 'Branca',

        4: 'Preta',

        6: 'Amarela',

        8: 'Parda',

        9: 'Sem declaração'}

frequencia = pd.crosstab(dados.Sexo, dados.Cor)

frequencia.rename(index=sexo, columns=cor, inplace=True)

frequencia



* + 1. Podemos passar o mesmo parâmetro (normalize=True) que para value\_counts() e já receber o valor das porcentagens:

frequencia = pd.crosstab(dados.Sexo, dados.Cor, normalize=True) \* 100

frequencia



* + 1. Podemos ainda fazer algo bem mais legal. Podemos passar o valor de outra variável do nosso dataset para ser analisado nesses cruzamentos, ou seja, se escolhermos média da renda, por exemplo, ao invés de mostrar quantos homens e mulheres existem de cada cor, ele passa a exibir, em média, quanto os homens e mulheres ganham de renda. Para isso utilizaremos os parâmetros aggfunc=análise\_estat’ e values = dados[‘index’]:

frequencia = pd.crosstab(dados.Sexo, dados.Cor, aggfunc = 'mean', values = dados.Renda)

frequencia



* 1. Pd.cut(variável, bins=[’distribuição’], labels=labels, include\_lowest=True/False): criamos uma lista com o intervalo colocado nos bins a partir da variável escolhida fazendo relação com as labels que determinamos, como no caso da renda, onde temos a classificação de A a E da renda populacional dos nossos dados analisados. Utilizamos essa função para mostrar quais pessoas estão em qual classificação a partir da renda:

classes = [0, 1576, 3152, 7880, 15760, 200000]

labels = ['E', 'D', 'C', 'B', 'A']

pd.cut(dados.Renda, bins = classes, labels = labels, include\_lowest=True)



* + 1. Porém Podemos notar que ele classifica indivíduo por indivíduo. Para saber a frequência podemos colocar o cut dentro do pd.value\_counts(), sabendo quantas pessoas estão em cada classificação:

frequencia = pd.value\_counts(pd.cut(dados.Renda,

                                    bins = classes,

                                    labels = labels,

                                    include\_lowest=True))

frequencia



* + 1. Para saber o percentual, basta colocar o normalize=True e multiplicar por 100, como estávamos fazendo com os outros:

percentual = pd.value\_counts(pd.cut(dados.Renda,

                                    bins = classes,

                                    labels = labels,

                                    include\_lowest=True), normalize=True) \* 100

percentual



* + 1. Criamos a tabela de frequência e porcentagem com pd.DataFrame():

dist\_freq\_quantitativas\_personalizadas = pd.DataFrame({'Frequência': frequencia,

                                                       'Porcentagem (%)': percentual})

dist\_freq\_quantitativas\_personalizadas



* + 1. E alteramos a ordem do index para crescente, ou seja, de A para baixo com o sorted\_index(ascending=True):

dist\_freq\_quantitativas\_personalizadas.sort\_index(ascending=False)



* 1. Biblioteca numpy possuí diversas fórmulas matemáticas que podemos usar.
     1. Com isso podemos fazer vários cálculos como log10 dentre vários outros, permitindo com que façamos cálculos como tamanho de classes:



* + - 1. N = valor de linhas do nosso df. Basta atribuir o df.shape[0] em n.
    1. Tendo a fórmula nós podemos executar a conta:

n = dados.shape[0]

k = 1 + (10/3) \* np.log10(n)

k = int(k.round(0))

* + - 1. Arredondamos porque não tem como fazer 17.26.... classes, apenas números inteiros.
    1. Com isso podemos fazer um cut passando k como as bins para ter essa classe de amplitude fixa. De resto, a criação da frequência é a mesma:

frequencia = pd.value\_counts(pd.cut(x = dados.Renda, bins = k, include\_lowest = True), sort = False)

frequencia



* + 1. O mesmo vale para o percentual:

percentual = pd.value\_counts(pd.cut(x = dados.Renda, bins = k, include\_lowest = True), sort = False, normalize=True) \* 100

percentual



dist\_freq\_quantitativas\_amplitude\_fixa = pd.DataFrame({'Frequência': frequencia, 'Porcentagem (%)': percentual})

dist\_freq\_quantitativas\_amplitude\_fixa



* 1. Sns.distplot(dados, kde=bool): cria um histograma com a variável do df passada.
     1. O kde é a exibição de densidade do gráfico. Quando true mostra a linha traçada em cima das colunas no fundo, quando false, só as colunas.
     2. Podemos atribuir esse gráfico a uma variável (ax, por exemplo) para ficar com fácil acesso.
     3. A partir disso podemos configurar o tamanho de exibição do gráfico com ax.set\_size\_inches(x, y).
     4. O Título com ax.set\_titles(‘titulo’, fontsize=n).
     5. E as labels com ax.set\_xlabels(‘label’, fontsize=n) e ax.set\_ylabels(‘label’, fontsize=n).
     6. Exemplo:

ax = sns.displot(dados.Altura, kde=True)

ax.figure.set\_size\_inches(12,6)

ax.set\_titles('Distribuição de Frequências - Altura - KDE', fontsize = 18)

ax.set\_xlabels("Metros", fontsize = 14)

ax



Kde = false:



* + 1. Podemos obter um resultado parecido com o pandas usando a função dados.Altura.hist.(bins=n\_barras, figsize=(x,y)):

dados.Altura.hist(bins = 50, figsize=(12,6))



* + 1. Com pandas ainda podemos fazer um df[‘variavel’].plot.bar(width=n, color=’color’, alpha=n). Alpah é a transparência da cor das barras:

dist\_freq\_quantitativas\_personalizadas.Frequência.plot.bar(width = 1, color = 'blue', alpha = 0.2)

