I EXPOTECH

Ênfase em Análise de Dados em Big Data

Análise de Dados

Mortalidade por Tabagismo

Maria Eduarda de Souza Cabral e Pedro Rigo de Oliveira

I Exposição de Projetos de Tecnologia da Informação do Campus UNESA - Niterói

Introdução

Este trabalho de extensão aborda a análise de dados de Regressão Simples e uma análise descritiva sobre a correlação entre o consumo de tabaco e a mortalidade em países específicos ao longo de um período. O estudo inclui uma representação gráfica dos resultados. A primeira etapa compreendeu uma análise de regressão linear e uma análise descritiva para o período de 1990 a 2012.

Regressão Linear e Big Data

Big Data envolve o processamento de dados complexos para descobrir insights valiosos, enquanto a Regressão Linear é um modelo matemático que descreve relações entre variáveis. Utilizando a regressão linear simples, o projeto aplicou uma equação de linha reta ($Y = \beta_0 + \beta_1 X + \varepsilon$) para previsões futuras. Na análise de Big Data, a Regressão Linear é crucial para compreender relações entre variáveis e tomar decisões estratégicas fundamentadas.

```
x = df[['CIGARROS POR DIA']]
y = df['TAXA DE MORTE']
x = x.apply(lambda x: x * 365)
model = LinearRegression()
model.fit(x, y)
coeficienteangular = model.coef[0]
intercepto = model.intercept_
previsoes = model.predict(x)
plt.scatter (x, y, color='blue')
plt.plot (x, previsoes,color='red')
plt.xlabel ('Número médio de cigarros fumados por pessoa por ano')
plt.ylabel ('Taxa de Mortalidade (por 100.000 habitantes)')
plt.title('Regressão Linear')
plt.gca() . invert_xaxis()
plt.gca() . invert_yaxis()
plt.show()
```

Metodologia

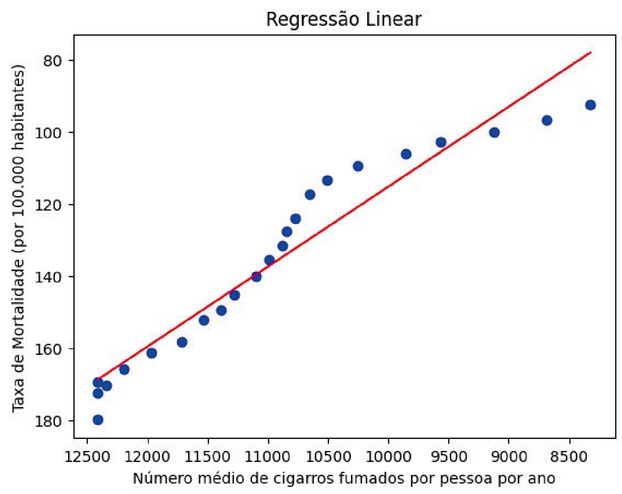
Python é uma linguagem de programação versátil conhecida por sua legibilidade e sintaxe simples, amplamente utilizada em ciência de dados devido à variedade de bibliotecas especializadas disponíveis.
NumPy é uma biblioteca essencial para computação científica, suportando arrays multidimensionais e matrizes, juntamente com funções matemáticas de alto nível.
Pandas, uma biblioteca de código aberto, oferece estruturas de dados fáceis de usar e ferramentas para análise de conjuntos de dados tabulares, facilitando a manipulação e análise de dados complexos.
Scikit-learn, também de código aberto, é uma biblioteca de aprendizado de máquina que oferece uma variedade de algoritmos supervisionados e não supervisionados.
Matplotlib, por sua vez, é uma biblioteca de visualização de dados usada para criar gráficos de alta qualidade em 2D.

Resultados da Regressão Linear com Python

O primeiro passo da análise envolveu a importação de dados e a limpeza utilizando ferramentas da biblioteca pandas da linguagem Python. Nessa etapa, duas tabelas principais foram criadas: tb_taxamorte e tbcigarroporpessoa. Ambas passaram pelo processo de normalização e tiveram suas colunas separadas por ponto e vírgula (;).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression tb_taxamorte =
pd.read_csv('https://raw.githubusercontent.com/MariaESCabral/NF_Topicos_Big_Data_Python/main/death-rate-smoking%20(1).csv', sep=';')
tb_cigarroporpessoa =
pd.read_csv('https://raw.githubusercontent.com/MariaESCabral/NF_Topicos_Big_Data_Python/main/consumption-per-smoker-per-day-bounds-brasil.csv', sep=';')
tb_taxamorte = tb_taxamorte.iloc[:, 2:]
tb_cigarroporpessoa = tb_cigarroporpessoa.iloc[:, 2:] tb_taxamorte.columns =
['ANO', 'TAXA DE MORTE'] tb_cigarroporpessoa.columns = ['ANO', 'CIGARROS POR DIA']
df = tb_cigarroporpessoa.merge(tb_taxamorte, on='ANO', how='inner')
df = df.drop(columns=['ANO']) df['CIGARROS POR DIA'] = df['CIGARROS POR DIA'].str.replace(',', '.') .astype(float)
df['TAXA DE MORTE'] = df['TAXA DE MORTE'].str.replace(',', '.') .astype(float)
```

Após a importação e normalização das tabelas, colunas irrelevantes foram removidas e as restantes foram renomeadas para melhor compreensão. As tabelas foram unidas usando o método 'inner', descartando anos com valores NaN. A coluna 'anos' foi então removida, pois se tornou desnecessária. Em seguida, as vírgulas foram removidas e os valores foram convertidos para o formato float.
Posteriormente, a variável dependente (y) foi estabelecida como a taxa de mortalidade por doenças cardiovasculares, enquanto as variáveis independentes (x) foram definidas como o consumo per capita de tabaco. É destacada a influência significativa de x em y, ressaltando a importância desses parâmetros na análise de regressão linear simples. Os valores de x foram multiplicados por 365 para obter uma média anual, e um modelo de regressão linear foi criado.
Durante 1990 a 2012, uma análise das variáveis x e y revelou uma clara ligação entre o aumento do número de fumantes e um consequente aumento no risco de mortalidade relacionado ao tabagismo. Isso destaca a relação direta entre o aumento de fumantes e os impactos negativos do tabaco. Um gráfico ilustrativo desse processo pode ser encontrado abaixo:



Análise descritiva

Para finalizar o processo fizemos a análise descritiva dos resultados obtidos.
Média de cigarros por dia: 29.92173904347826
Média da taxa de mortalidade: 135.6416072173913
Moda de cigarros por dia: 34.0
Moda da taxa de mortalidade: 92.3854
Mediana de cigarros por dia: 30.1
Mediana da taxa de mortalidade: 135.54193
Desvio padrão de cigarros por dia: 3.274412815659105
Desvio padrão da taxa de mortalidade: 27.503813717945338

Referências

Our World in Data: https://ourworldindata.org/smoking
Dados Estatísticos IBGE: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.htm
Exemplos de Gráficos de Dispersão para regressão linear simples e múltipla: https://edisciplinas.usp.br/pluginfile.php/4466762/mod_resource/content/1/AULA%2024%2025%20%20CORRELA%C3%87%C3%83O%20E%20REGRESS%C3%83O.pdf
Programa Nacional de Controle do Tabagismo: https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/programa-nacional-decontrole-do-tabagismo-e-outros-fatores-de-risco-de-cancer.pdf