PUC Minas
DIRETORIA DE
EDUCAÇÃO CONTINUADA

Pós Graduação *Lato Sensu*

Curso: Ciências de Dados e Big Data
Disciplina: Estatística Geral

Alunos

- Alessandro De Almeida Sícoli - 96191
- Matheus D'Almeida Vieira - 96463
- Pedro Araújo Júnior -96452
- Paulo Rogério Vieira Matos - 96287
- Raquel Cardoso Lemos - 873568
- Rúbia Martins - 97393

Tema
Move Hub City Ranking

Origem dos dados

MoveHub

Empresa inserida no setor logístico, cujo negócio são mudanças imobiliárias para qualquer lugar do mundo.

NUMBEO

Bases de dados sobre o custo de vida em todo o mundo, utilizando para isso inputs de usuários da página e consulta bianual de preços de mercado.

Origem de dados

CENTRAL INTELLIGENCE AGENCY

O **World Factbook** da **CIA** fornece informações sobre a história, pessoas, governo, economia, geografia, comunicações, transporte, questões militares e transnacionais para 267 entidades mundiais.

World Health Organization

WHO - World Health Organization ou OMS - Organização Mundial de Saúde: página oficial do braço sobre saúde pública da ONU - Organização das Nações Unidas.

As bases de dados

- **Cities.csv**: Obtida da Wikipédia, este arquivo contém uma lista de países e respectivas cidades. Uma linha completa deste conjunto de dados representa uma cidade e um país.

	City	Country
0	Oakland	United States
1	Oakville	Canada
2	Oaxaca de Juárez	Mexico
3	Oberhausen	Germany
4	Obihiro	Japan
5	Obninsk	Russia
6	Oceanside	United States
7	Odawara	Japan
8	Odense	Denmark
9	Odessa	Ukraine

Haviam 3543 linhas e 2 colunas que, após tratamento, restaram 3.517 linhas distintas sem contar cabeçalhos

As bases de dados

- **Cost Of Living.csv**: Este conjunto de dados diz respeito à índices relacionados ao custo de vida das cidades que compõem o MoveHub.
- Contém 216 linhas e 7 colunas

	City	Cappuccino	Cinema	Wine	Gasoline	Avg Rent	Avg Disposable Income
0	Lausanne	3.15	12.59	8.40	1.32	1714.00	4266.11
1	Zurich	3.28	12.59	8.40	1.31	2378.61	4197.55
2	Geneva	2.80	12.94	10.49	1.28	2607.95	3917.72
3	Basel	3.50	11.89	7.35	1.25	1649.29	3847.76
4	Perth	2.87	11.43	10.08	0.97	2083.14	3358.55

As bases de dados

- **City:** Nome de todas as cidades em um determinado país;
- **Cappucino:** Métrica/Indicador do Índice Cappucino;
- **Cinema**: Preço médio do cinema na cidade, em Libras;
- **Wine:** Preço da garrafa de vinho, em Libras;
- **Gasoline:** Preço da gasolina, em Libra;
- **Avg Rent:** Preço médio do aluguel, em Libras;
- **Avg Disposable Income:** Renda média disponível da população, em Libras.

As bases de dados

Análise descritiva da base Cost Of Living

	Cappuccino	Cinema	Wine	Gasoline	Avg Rent	Avg Disposable Income
Contagem	216	216	216	216	216	216
Média	1.981481	6.775602	7.079722	1.001898	1092.979213	1413.530463
Desvio Padrão	0.737131	5.632751	3.325691	0.351713	664.778486	912.013027
Mínimo	0.460000	1.810000	2.130000	0.070000	120.680000	120.680000
25%	1.320000	4.397500	4.260000	0.735000	609.015000	549.860000
50%	2.085000	6.540000	6.540000	0.950000	980.650000	1535.415000
75%	2.490000	7.850000	8.472500	1.320000	1388.095000	2053.812500
Máximo	4.480000	79.490000	26.150000	1.690000	5052.310000	4266.110000

As bases de dados

- **Quality of Life.csv**: Este conjunto de dados diz respeito a índices relacionados a qualidade de vida das cidades que compõem o MoveHub.
- Contém 216 linhas e 7 colunas

	City	Movehub Rating	Purchase Power	Health Care	Pollution	Quality of Life	Crime Rating
0	Caracas	65.18	11.25	44.44	83.45	8.61	85.70
1	Johannesburg	84.08	53.99	59.98	47.39	51.26	83.93
2	Fortaleza	80.17	52.28	45.46	66.32	36.68	78.65
3	Saint Louis	85.25	80.40	77.29	31.33	87.51	78.13
4	Mexico City	75.07	24.28	61.76	18.95	27.91	77.86

As bases de dados

- **City:** Nome de todas as cidades em um determinado país;
- **Movehub Rating:** Uma combinação de todas as pontuações para uma classificação geral de uma cidade ou país.
- **Purchase Power:** Compara o custo médio de vida com o salário médio local.
- **Health Care:** Compilado de como os cidadãos se sentem sobre o seu acesso aos cuidados de saúde e sua qualidade.

As bases de dados

- **Pollution:** Uma pontuação de quão poluída as pessoas encontram uma cidade, incluindo poluição do ar, da água e ruído.
- **Quality of Life:** Um equilíbrio de cuidados de saúde, poluição, poder de compra, taxa de criminalidade são utilizados para se calcular um ranking global de qualidade de vida.
- **Crime Rating:** Sensação de segurança. Quanto menor a pontuação, mais seguras as pessoas se sentem nesta cidade.

As bases de dados

Análise descritiva da base Quality of Life

	Movehub Rating	Purchase Power	Health Care	Pollution	Quality of Life	Crime Rating
Contagem	216.000000	216.000000	216.000000	216.000000	216.000000	216.000000
Média	79.676713	46.477176	66.442824	45.240370	59.994537	41.338611
Desvio Padrão	6.501011	20.614519	14.416412	25.369741	22.019376	16.416409
Mínimo	59.880000	6.380000	20.830000	0.000000	5.290000	9.110000
25%	75.070000	28.815000	59.420000	24.410000	42.752500	29.375000
50%	81.060000	49.220000	67.685000	37.210000	65.150000	41.140000
75%	84.020000	61.607500	77.207500	67.675000	78.617500	51.327500
Máximo	100.000000	91.850000	95.960000	92.420000	97.910000	85.700000

PROBLEMA

- Verificar se realmente os dados que compõem o "Cost Of Life" são formadores do cálculo do campo "Purchase Power" que consta na base "Quality of life".

SUGESTÃO DA SOLUÇÃO

- Realizar uma regressão linear múltipla indexada pela cidade. Neste caso, nossa variável dependente seria a “Quality of Life" e as demais colunas as variáveis independentes.

Testando a solução

- Juntamos as bases de dados das guias “Cost_Life” e “Society”, indexadas por cidade e, utilizando o módulo pandas e statsmodels, utilizamos as variáveis de “Cost_Life” como nossas variáveis dependentes.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

df = pd.read_excel('Dados Moviehub.xlsx')

independentes = df
independentes.drop(df.columns[[7,8,9,10,11,12]], axis=1, inplace=True)

dependente = pd.DataFrame(df.PurchasePower)

X = independentes[['Cappuccino','Cinema', 'Wine', 'Gasoline', 'AvgRent', \
                   'AvgDisposableIncome']]

y = dependente['PurchasePower']

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

model.summary()
```

5. VALIDAÇÃO DA SOLUÇÃO

- Tratamento dos dados em plataforma tabular (excel)
- Saneamento da base de dados
- Utilização do coeficiente de Pearson
- Visualização em gráfico de dispersão