Pontifícia Universidade Católica de Minas Gerais
Pós Graduação em Ciência dos Dados e Big Data
Campus Praça da Liberdade

Análise Descritiva de Dados
Qualidade e Custo de Vida nas Cidades

Disciplina: Estatística Geral - Teoria e Aplicações

Professor: Wagner Rogério

Grupo II
Amanda M. P. Amorim
Carlos R. Cardoso
Frederico Augusto C.V. da Costa
Ricardo Lempke
Sérgio R. I. Yoshioka
Thaís Ferreira Araújo

1. Introdução

Este documento apresenta inicialmente uma análise descritiva dos dados relacionados à qualidade e custo de vida em algumas das principais cidades e países ao redor do mundo. Os dados utilizados foram extraídos do dataset Movehub City Ranking[1] publicado na plataforma online de datascience Kaggle. Esse dataset se baseia no ranking de cidades do MoveHub[2], plataforma online que permite a consulta das melhores cidades para viver a partir de critérios de pesquisa submetidos pelos seus usuários.

Posteriormente é realizada uma exploração dos dados a fim de identificar um problema e propor a solução por meio de métodos estatísticos.

Ao final do trabalho é feita uma análise das variáveis, executados testes de correlação e clusterização, com o objetivo de propor um modelo para solução do problema.

2. Descrição do Dataset

2.1. Dicionário de Dados

O Dataset Movehub City Ranking é composto por três tabelas, armazenadas em arquivo texto no formato CSV:

Tabela: movehubqualityoflife.csv

Descrição: Registra para cada cidade parâmetros relacionados à percepção de qualidade de vida por parte de seus habitantes.

Parâmetro	Descrição
City	nome da cidade
Movehub Rating	nota geral da cidade a partir da combinação de todos os demais parâmetros
Purchase Power	comparação entre o custo de vida médio e o salário médio local

[1] disponível em: https://www.kaggle.com/blitzr/movehub-city-rankings/home
[2] disponível em:https://www.movehub.com/city-rankings/

Health Care	traduz a percepção dos habitantes sobre a qualidade e facilidade de acesso aos serviços de saúde
Pollution	traduz a percepção dos habitantes sobre a poluição, incluindo ar, água e poluição sonora
Quality of Life	traduz a qualidade de vida da cidade por meio de um balanceamento dos demais parâmetros
Crime Rating	traduz a percepção dos habitantes sobre a segurança

Amostra:

	City	Movehub Rating	Purchase Power	Health Care	Pollution	Quality of Life	Crime Rating
2	Aachen	81.64	60.55	73.25	11.69	90.52	15.34
3	Aberdeen	81.89	49.7	82.86	34.31	76.77	24.22
4	Abu Dhabi	86.4	68.03	48.02	53.42	80.8	10.86
5	Addis Ababa	59.88	6.38	63.89	85.59	28.41	26.04
6	Adelaide	87.29	72.03	56.25	12.01	91.54	41.32

	Movehub Rating	Purchase Power	Health Care	Pollution	Quality of Life	Crime Rating
count	216.000000	216.000000	216.000000	216.000000	216.000000	216.000000
mean	79.676713	46.477176	66.442824	45.240370	59.994537	41.338611
std	6.501011	20.614519	14.416412	25.369741	22.019376	16.416409
min	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
max	100.000000	91.850000	95.960000	92.420000	97.910000	85.700000

Tabela: movehubcostofliving.csv

Descrição: Registra para cada cidade valores cotados em Libras Esterlinas (GBP) para diferentes parâmetros.

Parâmetro	Descrição
City	nome da cidade
Capuccino	preço médio do capuccino

Cinema	preço médio do cinema
Wine	preço médio do vinho
Gasoline	preço médio da gasolina
Avg Rent	preço médio do aluguel residencial
Avg Disposable Income	renda média disponível após dedução de taxas e impostos

Amostra:

	City	Cappuccino	Cinema	Wine	Gasoline	Avg Rent	Avg Disposable Income
2	Aachen	2.05	6.88	4.26	1.33	767.23	1619.72
3	Aberdeen	1.99	6.98	5.98	1.37	1195.74	1743.78
4	Abu Dhabi	2.67	6.23	13.73	0.3	1779.93	2135.92
5	Addis Ababa	0.46	2.29	4.18	0.72	653.77	124.22
6	Adelaide	2.49	11.42	10.08	0.95	1382.26	2911.69

	Cappuccino	Cinema	Wine	Gasoline	Avg Rent	Avg Disposable Income
count	216.000000	216.000000	216.000000	216.000000	216.000000	216.000000
mean	1.981481	6.775602	7.079722	1.001898	1092.979213	1413.530463
std	0.737131	5.632751	3.325691	0.351713	664.778486	912.013027
min	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
max	4.480000	79.490000	26.150000	1.690000	5052.310000	4266.110000

Tabela: cities.csv

Descrição: Registra as cidades e seus respectivos países.

Parâmetro	Descrição
Cidade	nome da cidade
País	nome do país

Amostra:

	City	Country
2	A Coruña	Spain
3	Aachen	Germany
4	Aalborg	Denmark
5	Aarhus	Denmark
6	Aba	Nigeria

2.2. Classificação das variáveis

Conforme definição da estatística descritiva, foram classificados os dados disponíveis no dataset Movehub City Ranking:

População: Cidades e seus respectivos países

Qualitativas

nominal: *nenhuma*

ordinal: *nenhuma*

Quantitativas

contínua: Movehub Rating, Purchase Power, Health Care, Pollution, Quality of Life, Crime Rating, Capuccino, Cinema, Wine, Gasoline, Avg Rent, Avg Disposable Income
discreta: *nenhuma*

3. Problemas

Com base na descrição e exploração dos dados (disponível no Anexo I) foram definidos os problemas:

3.1. Problema Principal

Como estimar o ranking de uma cidade que não foi estimada pelo MoveHub por meio da coleta de seus dados?

3.2. Problema Secundário

Como saber se os dados processados pelo MoveHub e seus índices são confiáveis?

4. Proposta de Solução (Sugestão)

A seguir a proposta de solução dos problemas anteriormente expostos:

4.1. Informações de custo de vida

4.1.1. Validação das informações do poder de compra

Validar se a métrica de Purchase Power é fortemente correlacionada a Renda Média Disponível.

4.1.2. Geração da informação de poder de compra

1) Selecionar dentre as informações de:
    - Preço do vinho, capuccino, cinema, gasolina e aluguel
    - Renda média disponível

Combinar as variáveis mais relevantes para definir o custo de vida da cidade. Como na base não há a informação do custo de vida, será utilizado o Poder de Compra.

4.2. Informações de qualidade de vida

4.2.1. Validação das informações do MoveHub Rating

Validar se a métrica do MoveHub Rating é fortemente correlacionada ao Health Care e Purchase Power; e fortemente negativamente correlacionada ao crime e poluição.

4.2.2. Validação das informações do Quality of Life

Validar se a métrica do Quality of Life é fortemente correlacionada ao Health Care e Purchase Power; e fortemente negativamente correlacionada ao crime e poluição.

4.2.3. Selecionar o que melhor representa a melhor cidade para se viver

Dentre as colunas Quality of Life e MoveHub Rating, selecionar a que melhor representa a qualidade de vida.

4.3. Geração da informação da qualidade de vida com os dados primários

Combinar as variáveis mais relevantes para definir a qualidade de vida de uma cidade.

5. Validação estatística

Nesta seção, será tratada a proposta de solução de forma a validar estatisticamente com os dados disponíveis.

5.1. Informações de poder de compra

A respeito das informações do poder de compra, foi avaliada a informação da renda disponível, e posteriormente avaliado o modelo de inferência do poder de compra por meio das informações primárias (preços de produtos, renda disponível, etc). Não é possível verificar de forma direta o custo de vida, porque esse dado não está disponível na base. Por isso, na geração do modelo para poder de compra, foram utilizados os custos do vinho, cappucino, cinema, aluguel e gasolina.

5.1.1. Validação das informações do poder de compra

```
> cor.test(dados$Purchase.Power, dados$Avg.Disposable.Income)

        Pearson's product-moment correlation

data:  dados$Purchase.Power and dados$Avg.Disposable.Income
t = 22.389, df = 214, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7921823 0.8730709
sample estimates:
      cor
0.8371464
```

Há forte correlação (0.83) do poder de compra com a renda disponível. Além de haver significância do mesmo, visto que o p-valor (2,2 x 10^-16) é muito menor do que o valor de referência (0,05).

Assim, pode-se evidenciar que o poder de compra pode ter sido gerado com a informação da renda disponível e para tanto a informação parece ser confiável.

5.1.2. Geração da informação de poder de compra

Criando o modelo de regressão linear para predizer o custo de vida com base nas informações disponíveis, foi verificado que a informação do preço do capuccino e cinema não é interessante para o modelo:

```
Call:
lm(formula = dados$Purchase.Power ~ dados$Cappuccino + dados$Cinema +
    dados$Wine + dados$Gasoline + dados$Avg.Rent + dados$Avg.Disposable.Income,
    data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-34.805  -5.718   0.851   5.723  33.430

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 40.824796   3.043394  13.414  < 2e-16 ***
dados$Cappuccino            -1.941242   1.403575  -1.383  0.16812
dados$Cinema                 0.038540   0.126357   0.305  0.76066
dados$Wine                  -0.867456   0.263539  -3.292  0.00117 **
dados$Gasoline             -11.718988   2.141301  -5.473 1.26e-07 ***
dados$Avg.Rent              -0.008418   0.001399  -6.017 7.87e-09 ***
dados$Avg.Disposable.Income  0.025696   0.001168  21.997  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.398 on 209 degrees of freedom
Multiple R-squared:  0.798,     Adjusted R-squared:  0.7922
F-statistic: 137.6 on 6 and 209 DF,  p-value: < 2.2e-16
```

Esta escolha de variáveis para o modelo é validada pelo stepwise:

```
> sel.varPP <- stepAIC(modPP, direction = "both")
Start:  AIC=974.78
dados$Purchase.Power ~ dados$Cappuccino + dados$Cinema + dados$Wine +
    dados$Gasoline + dados$Avg.Rent + dados$Avg.Disposable.Income

                              Df Sum of Sq   RSS     AIC
- dados$Cinema                 1         8 18467  972.87
- dados$Cappuccino             1       169 18628  974.74
<none>                                     18459  974.78
- dados$Wine                   1       957 19416  983.69
- dados$Gasoline               1      2645 21104 1001.70
- dados$Avg.Rent               1      3197 21656 1007.28
- dados$Avg.Disposable.Income  1     42737 61196 1231.66

Step:  AIC=972.87
dados$Purchase.Power ~ dados$Cappuccino + dados$Wine + dados$Gasoline +
    dados$Avg.Rent + dados$Avg.Disposable.Income

                              Df Sum of Sq   RSS     AIC
- dados$Cappuccino             1       168 18635  972.83
<none>                                     18467  972.87
+ dados$Cinema                 1         8 18459  974.78
- dados$Wine                   1       961 19429  981.83
- dados$Gasoline               1      2637 21104  999.70
- dados$Avg.Rent               1      3240 21707 1005.79
- dados$Avg.Disposable.Income  1     45430 63898 1238.99
```

```
Step:  AIC=972.83
dados$Purchase.Power ~ dados$Wine + dados$Gasoline + dados$Avg.Rent +
    dados$Avg.Disposable.Income

                              Df Sum of Sq   RSS     AIC
<none>                                     18635  972.83
+ dados$Cappuccino             1       168 18467  972.87
+ dados$Cinema                 1         7 18628  974.74
- dados$Wine                   1      1511 20146  987.66
- dados$Gasoline               1      3063 21698 1003.69
- dados$Avg.Rent               1      3519 22154 1008.19
- dados$Avg.Disposable.Income  1     58519 77154 1277.71
>
```

Que gerou o modelo, também retirando o Capuccino e o Cinema:

dados$Purchase.Power ~ dados$Wine + dados$Gasoline + dados$Avg.Rent +
dados$Avg.Disposable.Income

5.2. Informações de qualidade de vida

Nesta seção serão avaliados as métricas de MoveHub Rating e Qualidade de vida e definido qual delas representa melhor a escolha da melhor cidade para se viver, com os dados disponíveis.

5.2.1. Validação das informações do MoveHub Rating

Validar se a métrica do MoveHub Rating é fortemente correlacionada ao Health Care e Purchase Power; e fortemente negativamente correlacionada ao crime e poluição.

```
> cor.test(dados$Movehub.Rating, dados$Health.Care)

        Pearson's product-moment correlation

data:  dados$Movehub.Rating and dados$Health.Care
t = 5.7678, df = 214, p-value = 2.787e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2453157 0.4769367
sample estimates:
      cor
0.3667969
```

```
> cor.test(dados$Movehub.Rating, dados$Purchase.Power)

        Pearson's product-moment correlation

data:  dados$Movehub.Rating and dados$Purchase.Power
t = 21.736, df = 214, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7828167 0.8670803
sample estimates:
      cor
0.8296145
```

```
> cor.test(dados$Movehub.Rating, dados$Crime.Rating)

        Pearson's product-moment correlation

data:  dados$Movehub.Rating and dados$Crime.Rating
t = -2.757, df = 214, p-value = 0.006337
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.31100903 -0.05302329
sample estimates:
       cor
-0.1852053
```

```
> cor.test(dados$Movehub.Rating, dados$Pollution)

        Pearson's product-moment correlation

data:  dados$Movehub.Rating and dados$Pollution
t = -4.1299, df = 214, p-value = 5.203e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3910081 -0.1434051
sample estimates:
       cor
-0.2716968
```

O poder de compra foi fortemente correlacionado com o MoveHub Rating (0,82), as demais tiveram correlações fracas, contudo todas tiveram o nível de significância, visto que o p-valor foi menor que o valor de referência de 0,05.

5.2.2. Validação das informações do Quality of Life

Validar se a métrica do Quality of Life é fortemente correlacionada ao Health Care e Purchase Power, e fortemente negativamente correlacionada ao crime e poluição.

```
> cor.test(dados$Quality.of.Life, dados$Health.Care)

        Pearson's product-moment correlation

data:  dados$Quality.of.Life and dados$Health.Care
t = 8.1711, df = 214, p-value = 2.649e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3788180 0.5831799
sample estimates:
      cor
0.4876507
```

```
> cor.test(dados$Quality.of.Life, dados$Purchase.Power)

        Pearson's product-moment correlation

data:  dados$Quality.of.Life and dados$Purchase.Power
t = 23.111, df = 214, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8019164 0.8792715
sample estimates:
      cor
0.8449566
```

```
> cor.test(dados$Quality.of.Life, dados$Crime.Rating)

        Pearson's product-moment correlation

data:  dados$Quality.of.Life and dados$Crime.Rating
t = -6.9092, df = 214, p-value = 5.495e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5303232 -0.3113193
sample estimates:
       cor
-0.4270639
```

```
> cor.test(dados$Quality.of.Life, dados$Pollution)

        Pearson's product-moment correlation

data:  dados$Quality.of.Life and dados$Pollution
t = -5.2005, df = 214, p-value = 4.637e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4484055 -0.2109006
sample estimates:
       cor
-0.3349631
```

O poder de compra foi fortemente correlacionado com a qualidade de vida (0,84), os demais dados tiveram correlações fracas, entretanto, mais fortes que o MoveHub Rating. Contudo todas tiveram o nível de significância, visto que o p-valor foi menor que o valor de referência de 0,05.

5.2.3. Selecionar o que melhor representa a melhor cidade para se viver

Conforme os dados anteriormente analisados, a teoria é que o MoveHub Rating deve utilizar mais informações do que as disponíveis na base de dados.

Desta forma, para buscar maiores evidências, foram criados os modelos do MoveHub Rating e da qualidade de vida, que segue:

```
Call:
lm(formula = dados$Quality.of.Life ~ dados$Purchase.Power + dados$Health.Care +
    dados$Pollution + dados$Crime.Rating, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-22.6595  -6.4556  -0.1708   6.2853  25.6953

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         22.86481    4.07004   5.618 6.06e-08 ***
dados$Purchase.Power  0.76865    0.03247  23.676  < 2e-16 ***
dados$Health.Care     0.27848    0.04651   5.987 9.07e-09 ***
dados$Pollution      -0.09587    0.02538  -3.777 0.000207 ***
dados$Crime.Rating   -0.30869    0.03955  -7.806 2.72e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.104 on 211 degrees of freedom
Multiple R-squared:  0.8322,    Adjusted R-squared:  0.829
F-statistic: 261.7 on 4 and 211 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = dados$Movehub.Rating ~ dados$Purchase.Power + dados$Health.Care +
    dados$Pollution + dados$Crime.Rating, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7229  -1.8294  -0.4123   1.1654  18.8300

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         66.2725301  1.5892152  41.701   <2e-16 ***
dados$Purchase.Power  0.2452972  0.0126766  19.350   <2e-16 ***
dados$Health.Care     0.0457144  0.0181612   2.517   0.0126 *
dados$Pollution      -0.0225498  0.0099117  -2.275   0.0239 *
dados$Crime.Rating   -0.0003331  0.0154416  -0.022   0.9828
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.555 on 211 degrees of freedom
Multiple R-squared:  0.7065,    Adjusted R-squared:  0.701
F-statistic:   127 on 4 and 211 DF,  p-value: < 2.2e-16
```

Assim, nota-se que o modelo para estimar a qualidade de vida foi mais rico do que o que infere o MoveHub Rating. É importante ressaltar que a métrica do MoveHub Rating não é pior, mas sim, menos adequada para os dados que temos disponível.

5.3. Geração da informação da qualidade de vida com os dados primários

Além de definir a melhor cidade para se viver por meio do poder de compra, que é uma informação secundária (gerada a partir de outras informações primárias, como no modelo apresentado em 5.1.2), foi avaliado a criação de um modelo somente com as variáveis primárias.

Assim, gerando o modelo com as variáveis primárias disponíveis na base, tivemos:

```
Call:
lm(formula = dados$Quality.of.Life ~ dados$Cappuccino + dados$Cinema +
    dados$Wine + dados$Gasoline + dados$Avg.Rent + dados$Avg.Disposable.Income +
    dados$Health.Care + dados$Pollution + dados$Crime.Rating,
    data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-35.512  -6.766   0.769   6.941  25.429

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 57.063942   5.642852  10.113  < 2e-16 ***
dados$Cappuccino            -0.413533   1.465753  -0.282 0.778127
dados$Cinema                -0.016948   0.133244  -0.127 0.898911
dados$Wine                  -1.131106   0.276430  -4.092 6.14e-05 ***
dados$Gasoline              -8.706797   2.411928  -3.610 0.000385 ***
dados$Avg.Rent              -0.009500   0.001474  -6.447 7.97e-10 ***
dados$Avg.Disposable.Income  0.021900   0.001261  17.372  < 2e-16 ***
dados$Health.Care            0.277950   0.051586   5.388 1.94e-07 ***
dados$Pollution             -0.111191   0.029385  -3.784 0.000202 ***
dados$Crime.Rating          -0.324496   0.044164  -7.348 4.64e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.783 on 206 degrees of freedom
Multiple R-squared:  0.8109,    Adjusted R-squared:  0.8026
F-statistic: 98.14 on 9 and 206 DF,  p-value: < 2.2e-16
```

Que como na geração do modelo de poder de compra, descartou as variáveis do preço do Capuccino e do Cinema.

Analogamente, foi avaliado o stepwise, para confirmar esta decisão:

```
> sel.varQ2 <- stepAIC(modQ2, direction = "both")
Start:  AIC=994.99
dados$Quality.of.Life ~ dados$Cappuccino + dados$Cinema + dados$Wine +
    dados$Gasoline + dados$Avg.Rent + dados$Avg.Disposable.Income +
    dados$Health.Care + dados$Pollution + dados$Crime.Rating

                               Df Sum of Sq   RSS     AIC
- dados$Cinema                  1       1.5 19716  993.01
- dados$Cappuccino              1       7.6 19723  993.08
<none>                                     19715  994.99
- dados$Gasoline                1    1247.1 20962 1006.24
- dados$Pollution               1    1370.3 21085 1007.51
- dados$Wine                    1    1602.4 21317 1009.87
- dados$Health.Care             1    2778.4 22493 1021.47
- dados$Avg.Rent                1    3977.5 23692 1032.69
- dados$Crime.Rating            1    5166.7 24882 1043.27
- dados$Avg.Disposable.Income   1   28882.5 48597 1187.87

Step:  AIC=993.01
dados$Quality.of.Life ~ dados$Cappuccino + dados$Wine + dados$Gasoline +
    dados$Avg.Rent + dados$Avg.Disposable.Income + dados$Health.Care +
    dados$Pollution + dados$Crime.Rating

                               Df Sum of Sq   RSS     AIC
- dados$Cappuccino              1       7.6 19724  991.09
<none>                                     19716  993.01
+ dados$Cinema                  1       1.5 19715  994.99
- dados$Gasoline                1    1248.4 20965 1004.27
- dados$Pollution               1    1388.9 21105 1005.71
- dados$Wine                    1    1685.3 21402 1008.72
- dados$Health.Care             1    2781.9 22498 1019.52
- dados$Avg.Rent                1    3997.5 23714 1030.88
- dados$Crime.Rating            1    5167.0 24884 1041.28
- dados$Avg.Disposable.Income   1   30241.9 49958 1191.83

Step:  AIC=991.09
dados$Quality.of.Life ~ dados$Wine + dados$Gasoline + dados$Avg.Rent +
    dados$Avg.Disposable.Income + dados$Health.Care + dados$Pollution +
    dados$Crime.Rating

                               Df Sum of Sq   RSS     AIC
<none>                                     19724  991.09
+ dados$Cappuccino              1         8 19716  993.01
+ dados$Cinema                  1         2 19723  993.08
- dados$Gasoline                1      1326 21050 1003.14
- dados$Pollution               1      1381 21105 1003.71
- dados$Wine                    1      2053 21777 1010.48
- dados$Health.Care             1      2787 22511 1017.64
- dados$Avg.Rent                1      4110 23834 1029.98
- dados$Crime.Rating            1      5164 24888 1039.32
- dados$Avg.Disposable.Income   1     39650 59374 1227.13
```

Que também sugeriu remover do modelo o preço do Capuccino e do Cinema.
Gerando o modelo sem estas variáveis, tem-se:

```
Call:
lm(formula = dados$Quality.of.Life ~ dados$Wine + dados$Gasoline +
    dados$Avg.Rent + dados$Avg.Disposable.Income + dados$Health.Care +
    dados$Pollution + dados$Crime.Rating)

Residuals:
    Min      1Q  Median      3Q     Max
-35.440  -6.960   0.752   7.045  25.533

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 56.746642   5.520952  10.278  < 2e-16 ***
dados$Wine                  -1.166331   0.250694  -4.652 5.83e-06 ***
dados$Gasoline              -8.832136   2.362151  -3.739 0.000239 ***
dados$Avg.Rent              -0.009535   0.001448  -6.584 3.67e-10 ***
dados$Avg.Disposable.Income  0.021691   0.001061  20.448  < 2e-16 ***
dados$Health.Care            0.278300   0.051334   5.421 1.63e-07 ***
dados$Pollution             -0.109967   0.028813  -3.817 0.000178 ***
dados$Crime.Rating          -0.324171   0.043930  -7.379 3.75e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.738 on 208 degrees of freedom
Multiple R-squared:  0.8108,    Adjusted R-squared:  0.8044
F-statistic: 127.3 on 7 and 208 DF,  p-value: < 2.2e-16
```

Assim, o modelo final proposto para inferir o índice de melhor cidade para se viver com os dados primários é:

Y = 56,75 -1,17 * Vinho - 8,83 * Gasolina - 0,01 * Aluguel + 0,02 * Renda + 0,28 * Saúde - 0,11 * Poluição - 0,32 * Crime

5.4. Considerações adicionais

Toda a avaliação foi levando em considerações os dados disponíveis, porém caso seja possível evoluir a análise para melhorar o modelo proposto, a acurácia provavelmente seria melhor.

Desta forma, caso se insira a informações de IDH dos países, pode-se verificar que a base de dados se concentrou basicamente em países de alto índice de IDH, conforme gráfico abaixo (Gráfico 1):

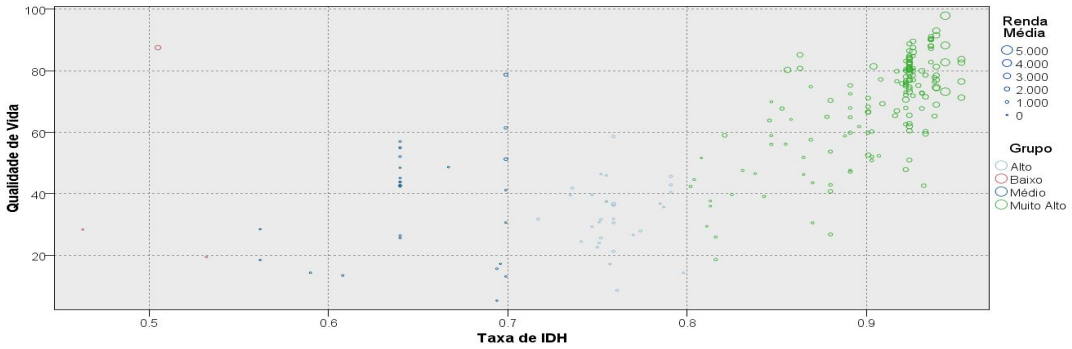

Gráfico 1: Relação entre IDH, Qualidade de Vida e Renda Média
Fonte: Autor

Verifica-se no Gráfico 1 que existe uma clusterização dos países e, principalmente, um comportamento não linear da qualidade de vida em função do IDH, sendo que países com alto IDH e rendas maiores, também, geralmente, representam resultados de qualidade de vida maiores.
Esta informação pode ser constatada variável a variável, conforme tabela abaixo (Tabela 1). Nesta tabela, verifica-se que itens relacionados a consumo e renda, apresentam valores mais altos, bem como itens sociais, apresentam resultados mais baixos.

Tabela 1: Estatística descritiva das variáveis clusterisadas
Fonte: Autor

Item avaliado	IDH Clusterizado	Valor médio	Variância	Número de Dados	Relação ao cluster Alto e Muito Alto
Valor do Cappuccino	Alto e Muito Alto	2,119	0,460	186,000	100%
	Médio	1,069	0,118	27,000	50%
	Baixo	1,680	1,225	3,000	79%
Valor da entradade de Cinema	Alto e Muito Alto	7,383	33,993	186,000	100%
	Médio	2,771	0,420	27,000	38%
	Baixo	5,123	6,021	3,000	69%
Valor do Vinho	Alto e Muito Alto	7,281	11,472	186,000	100%
	Médio	5,793	7,465	27,000	80%
	Baixo	6,190	3,459	3,000	85%
Valor da Gasolina	Alto e Muito Alto	1,038	0,130	186,000	100%
	Médio	0,799	0,024	27,000	77%
	Baixo	0,577	0,026	3,000	56%
Custo médio do Aluguel	Alto e Muito Alto	1.190,478	432.150,430	186,000	100%
	Médio	438,040	57.770,126	27,000	37%
	Baixo	942,517	73.817,916	3,000	79%
Renda Média	Alto e Muito Alto	1.572,979	754.144,510	186,000	100%
	Médio	375,153	82.558,745	27,000	24%
	Baixo	873,123	1.224.865,488	3,000	56%
Indice de mobilidade urbana	Alto e Muito Alto	80,525	37,305	186,000	100%
	Médio	74,623	34,103	27,000	93%
	Baixo	72,533	160,912	3,000	90%
Poder de Compra	Alto e Muito Alto	49,144	394,450	186,000	100%
	Médio	29,504	212,964	27,000	60%
	Baixo	33,897	1.640,110	3,000	69%
Qualidade da saúde	Alto e Muito Alto	67,547	190,222	186,000	100%
	Médio	58,694	288,526	27,000	87%
	Baixo	67,740	69,258	3,000	100%
Indice de Poluição	Alto e Muito Alto	42,314	617,440	186,000	100%
	Médio	65,109	374,386	27,000	154%
	Baixo	47,850	1.073,753	3,000	113%
Qualidade de Vida	Alto e Muito Alto	63,467	414,545	186,000	100%
	Médio	37,727	327,962	27,000	59%
	Baixo	45,133	1.366,773	3,000	71%
Taxa de criminalidade	Alto e Muito Alto	40,062	241,457	186,000	100%
	Médio	48,632	368,481	27,000	121%
	Baixo	54,870	701,611	3,000	137%

O comportamento observado era esperado quando relacionado ao comportamento econômico real, visto que países mais desenvolvidos e com melhor IDH, geralmente se concentram em regiões com maior qualidade de vida e custos de vida maiores, conforme podemos observar na imagem abaixo.

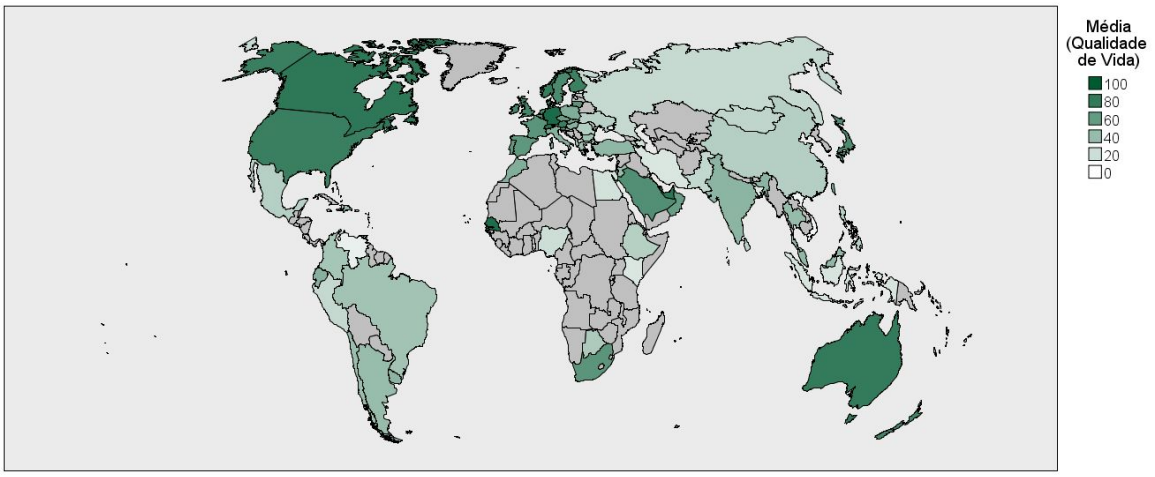

Figura 1: Países do mundo que se concentram as cidades analisadas por qualidade de vida média
Fonte: Autor

Os países geralmente analisados se concentram no hemisfério norte, o qual também detém a maioria dos países desenvolvidos do mundo e, consequentemente, países com maior qualidade de vida. A priori, estes países com câmbio forte e custo de vida geralmente mais elevado, explicar-se-á a relação da qualidade de vida com o poder econômico e renda média, não sendo itens específicos de consumo os mais relevantes para determinar uma equação de qualidade de vida.
Desta forma, se utilizássemos apenas a parte do gráfico que representa a maioria dos dados (~80% - Grupo de alto e muito alto IDH), teríamos um resultado melhor da regressão.

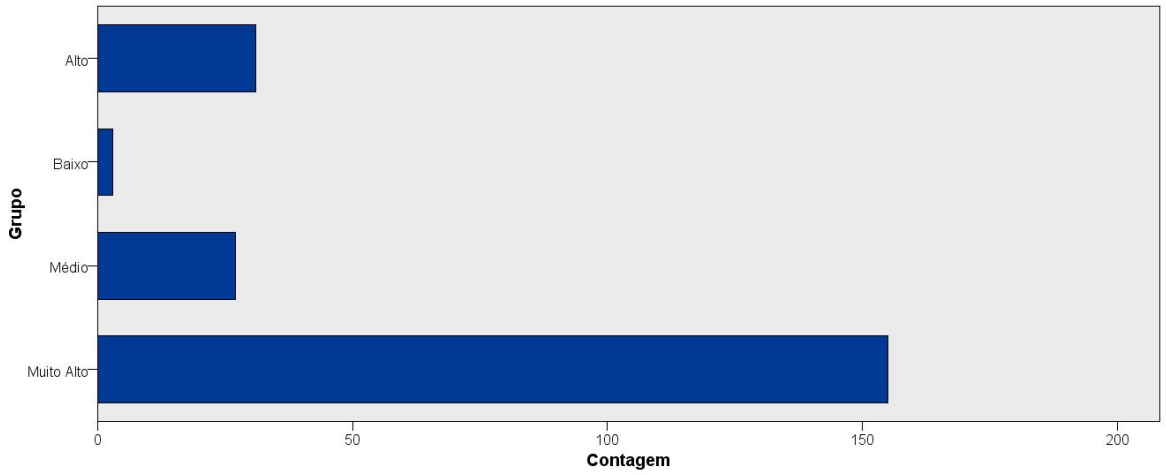

Gráfico 2: Volume de dados por clusters criados
Fonte: Autor

Neste cenário, verificando as correlações das variáveis (Tabela 2 - coluna 2), a significância (Tabela 2 - coluna 3) e a matriz de covariância, podemos determinar variáveis que melhor se adequaria para a regressão (Gráfico 3), conforme cenários abaixo.

Tabela 2: Correlações de Pearson das variáveis com a qualidade de vida
Fonte: Autor

Correlações de Pearson

Valor do Cappuccino	0.465	Forte
Valor da entradade de Cinema	0.228	Forte
Valor do Vinho	0.054	Fraco
Valor da Gasolina	0.084	Fraco
Custo médio do Aluguel	0.275	Forte
Renda Média	0.774	Forte
Indice de mobilidade urbana	0.701	Forte
Poder de Compra	0.826	Forte
Qualidade da saúde	0.448	Forte
Indice de Poluição	-0.307	Forte
Taxa de criminalidade	-0.468	Forte
Taxa de IDH	0.833	Forte

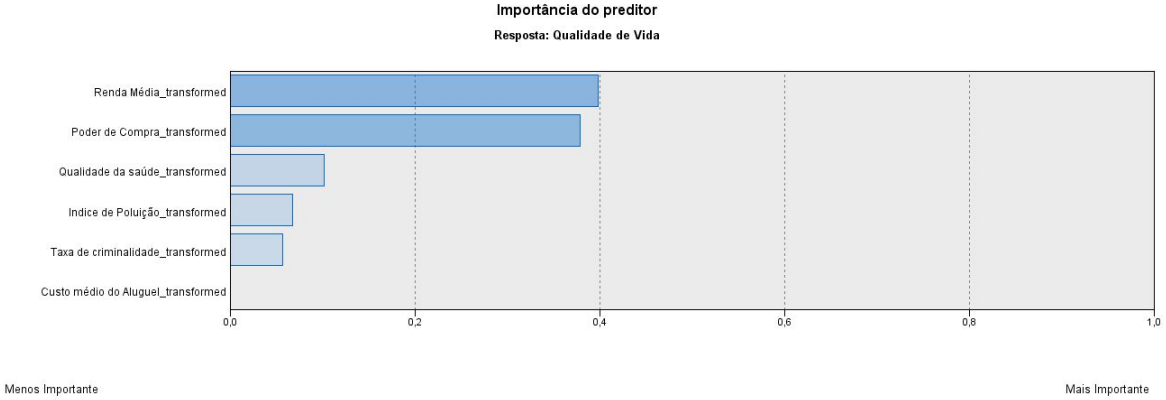

Gráfico 3: Variáveis mais importantes para determinação do modelo em ordem crescente
Fonte: Autor

Neste cenário, normalizando as variáveis através do método do máximo valor, o resultado da equação ficaria assim:

Qualidade de vida = Custo médio do Aluguel * -3,846 + Renda Média * 9,075 + Poder de Compra * 8,865 + Qualidade da saúde * 2,472 + Indice de Poluição * -1,939 + Taxa de criminalidade * -4,79 + 63,63

É importante salientar que outras considerações podem ser feitas para melhor o modelo, contudo, isto carece de mais análise e avaliações. Este cenário foi escrito basicamente para exemplificar que existem situações que podem melhor o modelo e avaliações, não sendo um cenário final para o modelo proposto.

6. Anexos: Análise utilizando Python (Jupyter Notebook)

Foi criado um Jupyter Notebook com uma outra visão da análise, na qual uma exploração da base foi desenvolvida e um modelo muito similar ao anterior (Item 5.4) criado. O mesmo segue anexo para melhor visualização.
Analise_Dados_Movehub_City.ipypnb