Projeto Exemplar de Prof. Jim

James Hunter, Ph.D.

5/30/2020

Este projeto é um pequeno exemplo de um projeto de aula, seja projeto individual ou projeto em grupo. Vai ter todos os elementos de um projeto para a matéria. Só vai faltar a profundidade de explicação em texto que estou esperando dos alunos. Pode copiar a estrutura do arquivo e dos blocos e até do código se for útil para seu projeto.

Este projeto de exemplo vai usar os dados do pacote `gapminder`. O propósito do estudo é ver se a expectativa de vida fica maior em países mais ricos que em países mais pobres. Ou seja, se expectativa de vida (`lifeExp`) cresce com PIB per capita (`gdpPercap`). Eu vou focar em dois anos, 1972 e 1992, e comparar os dados e os resultados da regressão linear para esses anos para ver se tiver diferenças interessantes.

VSS: Quem quer usar Gapminder, utilize o site: https://www.gapminder.org/data/ para fazer o download de variáveis que lhe interessem. O pacote tem muito poucas variáveis.

Carregar Pacotes

```
library(tidyverse)
library(gapminder)
library(summarytools)
library(caret)
library(ggsci)
```

Esses são os pacotes de R necessário para analisar os dados de Gapminder. Vocês podem descrever porque estão usando os pacotes que usam.

Carregar Dados

Limitar os dados para anos 1972 e 1992

```
gm <- gapminder %>%
  filter(year %in% c(1972, 1992))
```

Pode explicar aqui de onde vieram os dados e o que eles significam. Quais são as unidades para cada variável?

Limpeza dos Dados

`year` é realmente um fator, não está usado numericamente.

```
# informação básica

glimpse(gm)
```

```
## Rows: 284
## Columns: 6
```

```
## $ country   <fct> Afghanistan, Afghanistan, Albania, Albania, Algeria, Alge...
## $ continent <fct> Asia, Asia, Europe, Europe, Africa, Africa, Africa, Afric...
## $ year      <int> 1972, 1992, 1972, 1992, 1972, 1992, 1972, 1992, 1972, 199...
## $ lifeExp   <dbl> 36.088, 41.674, 67.690, 71.581, 54.518, 67.744, 37.928, 4...
## $ pop       <int> 13079460, 16317921, 2263554, 3326498, 14760787, 26298373,...
## $ gdpPercap <dbl> 739.9811, 649.3414, 3313.4222, 2497.4379, 4182.6638, 5023...
```

```
gm %>%
  select(-1) %>%
  dfSummary(graph.col = FALSE, na.col = FALSE)
```

```
## Data Frame Summary
## gm
## Dimensions: 284 x 5
## Duplicates: 0
##
## ----------------------------------------------------------------------------------------------
## No   Variable     Stats / Values                          Freqs (% of Valid)    Valid
## ---- ------------ --------------------------------------- --------------------- --------
## 1    continent    1. Africa                               104 (36.6%)           284
##      [factor]     2. Americas                              50 (17.6%)           (100%)
##                   3. Asia                                  66 (23.2%)
##                   4. Europe                                60 (21.1%)
##                   5. Oceania                                4 ( 1.4%)
##
## 2    year         Min  : 1972                             1972 : 142 (50.0%)    284
##      [integer]    Mean : 1982                             1992 : 142 (50.0%)    (100%)
##                   Max  : 1992
##
## 3    lifeExp      Mean (sd) : 60.9 (11.7)                 283 distinct values   284
##      [numeric]    min < med < max:                                              (100%)
##                   23.6 < 62.8 < 79.4
##                   IQR (CV) : 19.9 (0.2)
##
## 4    pop          Mean (sd) : 30590448.7 (108016544.7)    284 distinct values   284
##      [integer]    min < med < max:                                              (100%)
##                   76595 < 7379128 < 1164970000
##                   IQR (CV) : 17271138 (3.5)
##
## 5    gdpPercap    Mean (sd) : 7464.3 (9862.1)             284 distinct values   284
##      [numeric]    min < med < max:                                              (100%)
##                   347 < 3770.4 < 109347.9
##                   IQR (CV) : 8367.7 (1.3)
## ----------------------------------------------------------------------------------------------
```

```
# Limpar nomes de variáveis - life_exp

gm <- janitor::clean_names(gm)

# Mude year to factor
gm <- gm %>%
  mutate(year = factor(year))
```

Análise Exploratório dos Dados

```
descr(gm, stats = c("mean", "sd", "min", "med", "max", "IQR", "CV"))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## Non-numerical variable(s) ignored: country, continent, year
```

```
## Descriptive Statistics
## gm
## N: 284
##
##                gdp_percap   life_exp            pop
## ------------- ------------ ---------- ---------------
##          Mean      7464.35      60.90     30590448.65
##       Std.Dev      9862.06      11.75    108016544.67
##           Min       347.00      23.60        76595.00
##        Median      3770.42      62.84      7379128.00
##           Max    109347.87      79.36   1164970000.00
##           IQR      8367.74      19.93     17271138.00
##            CV         1.32       0.19            3.53
```

```
# agrupado por ano
gm %>%
  group_by(year) %>%
  descr(stats = c("mean", "sd", "min", "med", "max", "IQR", "CV"))
```

```
## Non-numerical variable(s) ignored: country, continent
```

```
## Descriptive Statistics
## gm
## Group: year = 1972
## N: 142
##
##                gdp_percap   life_exp            pop
## ------------- ------------ ---------- --------------
##          Mean      6770.08      57.65    25189979.99
##       Std.Dev     10614.38      11.38    88646816.96
##           Min       357.00      35.40       76595.00
##        Median      3339.13      56.53     5877996.50
##           Max    109347.87      74.72   862030000.00
##           IQR      8251.65      20.75    12328008.00
##            CV         1.57       0.20           3.52
##
```

```
## Group: year = 1992
## N: 142
##
##                gdp_percap   life_exp             pop
## ------------- ------------ ---------- ---------------
##          Mean      8158.61      64.16     35990917.32
##       Std.Dev      9031.85      11.23    124502588.83
##           Min       347.00      23.60       125911.00
##        Median      4386.09      67.70      8688686.50
##           Max     34932.92      79.36   1164970000.00
##           IQR      9413.69      16.46     19099390.50
##            CV         1.11       0.17            3.46
```

```r
# média e sd por país

gm %>%
  group_by(country) %>%
  summarise(mean_pop = mean(pop),
            sd_pop = sd(pop),
            mean_gdp = mean(gdp_percap),
            sd_gdp = sd(gdp_percap),
            mean_life = mean(life_exp),
            sd_life = sd(life_exp)) %>%
  ungroup() %>%
  knitr::kable()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

country	mean_pop	sd_pop	mean_gdp	sd_gdp	mean_life	sd_life
Afghanistan	14698690.5	2289937.73	694.6613	64.091954	38.88100	3.9498985
Albania	2795026.0	751614.91	2905.4300	576.988023	69.63550	2.7513525
Algeria	20529580.0	8158305.30	4602.9402	594.360642	61.13100	9.3521943
Angola	7315423.0	2008982.29	4050.5668	2012.031560	39.28750	1.9226233
Argentina	29369373.0	6490637.80	9375.7286	95.190585	69.46650	3.3962339
Australia	15329488.5	3044078.43	20106.6982	4692.457721	74.74500	3.9810112
Austria	7729585.0	262172.57	21851.8221	7340.046338	73.33500	3.8254477
Bahrain	380145.5	211206.43	18652.1188	542.294884	67.95050	6.5768002
Bangladesh	92231937.0	30366901.54	734.0219	146.778777	50.63500	7.6127116
Belgium	9877361.0	237956.99	21123.8571	6295.673699	73.95000	3.5496760
Benin	3871539.0	1569963.73	1138.5023	74.536693	50.46650	4.8825723
Bolivia	5729661.5	1645846.89	2971.0155	13.174563	53.33550	9.3642151
Bosnia and Herzegovina	4037506.5	309014.86	2703.4756	221.598996	69.81400	3.3432009
Botswana	980982.5	511424.17	5108.8614	4023.791514	59.38450	4.7524647
Brazil	128408016.0	38986980.09	5967.9972	1389.161868	63.28050	5.3407775
Bulgaria	8617353.0	58199.13	6450.0589	208.505255	71.04500	0.2050610
Burkina Faso	7156094.5	2435570.62	893.2444	54.459099	46.92550	4.7156951
Burundi	4669609.5	1611675.25	547.8997	118.511361	44.39650	0.4801255
Cambodia	8800350.0	1908826.27	551.9636	184.327994	48.06000	10.9502556
Cameroon	9744099.5	3851004.65	1738.6549	77.086483	50.68150	5.1371308
Canada	25404001.0	4411640.62	22656.7276	5213.012798	75.41500	3.5850314
Central African Republic	2596192.0	946012.71	908.9594	227.764574	46.42650	4.1995072
Chad	5164242.5	1789226.94	1081.0841	32.554975	48.64650	4.3522422
Chile	11645259.0	2726228.98	6545.0752	1486.410245	68.78350	7.5554360
China	1013500000.0	214210928.29	1166.3421	692.175561	65.90444	3.9393767
Colombia	28372805.5	8244745.57	4354.6543	1541.484705	65.02200	4.8069119

country	mean_pop	sd_pop	mean_gdp	sd_gdp	mean_life	sd_life
Comoros	352228.0	144534.04	1592.2425	488.377656	53.44150	6.3604255
Congo, Dem. Rep.	32339906.0	13197776.13	681.3076	316.201810	45.76850	0.3118341
Congo, Rep.	1874765.5	755624.91	3614.6961	567.868155	55.67000	1.0790449
Costa Rica	2504006.0	946405.86	5639.2816	736.995745	71.78100	5.5606877
Cote d'Ivoire	9422146.0	4738251.83	2013.1375	516.277979	50.92250	1.5860405
Croatia	4359661.5	190001.71	8805.9425	506.497231	71.06850	2.0626305
Cuba	9777304.0	1337783.80	5449.1446	203.221575	72.56850	2.6099311
Czech Republic	10088930.0	320704.04	13702.7374	840.444224	71.34500	1.4919953
Denmark	5081494.5	127135.68	22636.4735	5331.961764	74.40000	1.3152186
Djibouti	281502.0	145174.68	3035.6843	931.299342	47.98500	5.1180389
Dominican Republic	6011255.0	1894941.52	2617.0444	604.109406	64.04400	6.2409245
Ecuador	8523522.5	3146443.45	6192.3487	1288.849106	64.20450	7.6487741
Egypt	47104807.5	17391136.43	2909.3817	1252.107245	57.40550	8.8649977
El Salvador	4532776.0	1049166.86	4482.2389	53.750233	62.50250	6.0747544
Equatorial Guinea	332720.5	77947.92	902.2336	325.016525	44.03050	4.9702536
Eritrea	2964313.5	995785.25	548.5914	48.461070	47.06650	4.1358676
Ethiopia	41429465.5	15074234.59	493.7987	102.453040	45.80300	3.2357206
Finland	4840348.0	283819.93	17503.0204	4446.491858	73.28500	3.4153258
France	54553089.5	3989623.03	20405.4939	6078.717295	74.92000	3.5921024
Gabon	761858.0	316615.55	12462.0530	1499.214239	55.02800	8.9632856
Gambia	771242.5	359410.36	710.8556	63.966593	45.47600	10.1370828
Germany	79657426.0	1329838.75	22260.7417	6002.716369	73.53500	3.5850314
Ghana	12816429.0	4896444.34	1051.6419	179.013666	53.68800	5.3923963
Greece	9607028.5	1015971.73	15133.1630	3405.897736	74.68500	3.3163308
Guatemala	6818265.0	2359875.54	4235.4296	288.529668	58.55550	6.8129738
Guinea	5400980.5	2248024.69	768.0073	37.251946	43.70900	6.8829774
Guinea-Bissau	838149.5	300928.38	782.8822	52.810070	39.87600	4.7941840
Haiti	5512491.5	1151439.25	1555.3832	140.111391	51.56550	4.9829815
Honduras	4021246.5	1493551.65	2805.7685	390.218474	60.14150	8.8494414
Hong Kong, China	4972698.0	1211978.19	16536.7656	11626.019791	74.80050	3.9605051
Hungary	10371387.5	32107.60	10352.1423	259.488701	69.46500	0.4171930
Iceland	234143.5	35169.37	20471.2278	6608.852184	76.61500	3.0476302
India	719500000.0	215667568.26	944.2197	311.391641	55.43700	6.7684261
Indonesia	153049000.0	44925322.24	1747.1244	899.463154	55.94200	9.5303852
Iran	45505986.5	21060449.28	8424.7359	1681.616895	60.48800	7.4302781
Iraq	13961705.5	5515715.03	6660.8391	4122.713191	58.20550	1.7755451
Ireland	3291080.5	377143.18	13544.7942	5676.683400	73.37350	2.9606561
Israel	4016221.5	1301541.05	15419.2274	3722.627508	74.28000	3.7476659
Italy	55603205.5	1750289.39	17141.4593	6890.310869	74.81500	3.7123106
Jamaica	2188117.0	269409.10	7419.4065	20.481778	70.38300	1.9558574
Japan	115758771.0	12120514.51	20801.8407	8517.885184	76.39000	4.2002143
Jordan	2740480.0	1593718.28	2771.2250	933.902328	62.27150	8.1225356
Kenya	18532662.0	9175243.64	1282.1408	84.542926	56.42200	4.0488934
Korea, Dem. Rep.	17746308.0	4193237.96	3713.8425	17.283107	66.98050	4.2391052
Korea, Rep.	38655225.0	7283518.04	7567.5777	6415.864132	67.42800	6.8108525
Kuwait	1130014.5	407407.35	72140.3933	52619.313935	71.45100	5.2877445
Lebanon	2950006.0	381820.69	7188.5956	421.136880	67.35650	2.7372103
Lesotho	1459987.0	485369.41	737.0339	340.050960	54.72600	7.0130851
Liberia	1697801.0	304300.57	719.8142	117.650218	41.70800	1.2812775
Libya	3274189.0	1541934.02	15325.8179	8040.764854	60.76400	11.3009806
Madagascar	9646412.5	3626018.83	1394.6196	500.551551	48.53250	5.2064272
Malawi	7372623.0	3735823.32	573.9110	15.147611	45.59300	5.4121953

country	mean_pop	sd_pop	mean_gdp	sd_gdp	mean_life	sd_life
Malaysia	14880482.0	4863508.73	5063.5038	3131.647256	66.85150	5.4327014
Mali	7122186.5	1830032.65	660.1916	111.472201	44.18250	5.9474751
Mauritania	1726125.5	556266.06	1474.1108	159.439849	53.38500	6.9975287
Mauritius	973768.0	173147.82	4316.8690	2462.690064	66.34450	4.8090332
Mexico	72047662.0	22717032.88	8140.8955	1883.009523	66.90800	6.4304291
Mongolia	1816651.0	701663.47	1603.5720	257.146481	57.51250	5.3153217
Montenegro	574649.5	66427.73	7390.8765	548.060774	73.03550	3.3934054
Morocco	21229454.5	6461237.00	2439.1211	719.730247	59.12750	8.8607551
Mozambique	11485163.5	2369610.28	567.9073	222.046364	42.30600	2.7973144
Myanmar	34506464.0	8541954.57	352.0000	7.071068	56.19500	4.4194174
Namibia	1188017.5	517935.21	3775.3095	41.335377	57.93300	5.7501923
Nepal	16369401.0	5595771.54	786.2642	157.651034	49.84900	8.3127473
Netherlands	14252059.0	1304166.53	22792.8476	5654.170030	75.58500	2.5950819
New Zealand	3183387.0	359616.12	17204.6811	1638.569818	74.11000	3.1395541
Nicaragua	3100423.5	1297562.86	3429.3725	1780.807093	60.49700	7.5603857
Niger	6726540.0	2356472.95	767.6960	263.769576	43.96850	4.8401459
Nigeria	73552164.5	28018511.53	1659.1185	55.536606	45.14650	3.2887536
Norway	4109680.5	249858.30	26465.3583	10607.029970	75.83000	2.1071782
Oman	1372129.0	768029.69	14617.3727	5655.912638	61.67000	13.4732126
Pakistan	94695462.5	35877949.66	1510.8842	651.875012	56.38350	6.2996143
Panama	2050690.5	614202.14	5991.4964	887.060781	69.33900	4.4165890
Paraguay	3549024.5	1322177.25	3359.8745	1183.041335	67.02000	1.7041273
Peru	18192574.5	5993259.59	5192.1041	1054.611834	60.95300	7.7852457
Philippines	54017953.5	18622099.02	2134.3490	205.025574	62.26150	5.9347472
Poland	35705121.0	3769693.73	7872.6941	189.239980	70.92000	0.0989949
Portugal	9449065.0	676863.82	12614.7570	5080.575808	72.06000	3.9597980
Puerto Rico	3216154.0	521875.92	11882.3144	3902.200852	73.03550	1.2381440
Reunion	541912.0	113531.65	5574.4572	745.005767	68.94450	6.6050844
Romania	21729837.5	1509233.86	7304.9122	999.145063	69.28500	0.1060660
Rwanda	5641162.0	2332096.15	663.8246	103.582609	34.09950	14.8499495
Sao Tome and Principe	101253.0	34871.68	1480.8815	73.685788	59.61100	4.4279027
Saudi Arabia	11709306.5	7405600.74	24839.5232	2.962155	61.32700	10.5231631
Senegal	6448308.0	2629888.51	1482.8057	162.502109	52.00550	8.7546891
Serbia	9069842.5	1069929.63	9923.5679	846.406288	70.17950	2.0923290
Sierra Leone	3569948.5	977130.35	1211.2280	201.570323	36.86650	2.0739442
Singapore	2694132.5	766125.45	16683.8237	11435.426323	72.65450	4.4314382
Slovak Republic	4948160.5	501660.44	9586.3177	124.238593	70.86500	0.7283200
Slovenia	1846860.0	215455.44	13299.1015	1294.875582	71.73000	2.7011479
Somalia	4969980.0	1597805.35	1090.7682	231.659375	40.31550	0.9298454
South Africa	31949984.5	11333754.27	7495.5159	382.469376	57.79200	5.7926188
Spain	37031299.5	3561185.62	14620.9079	5631.619878	75.31500	3.1890516
Sri Lanka	15301896.5	3231709.21	1683.5674	664.923401	67.71050	3.7738289
Sudan	21412303.5	9638267.77	1575.9249	118.409084	49.31950	5.9913158
Swaziland	721224.5	340994.47	3458.9295	133.067438	54.01300	6.3088067
Sweden	8420580.0	421841.52	20856.0207	4276.576290	76.44000	2.4324473
Switzerland	6698423.5	420054.66	29533.3217	3306.726356	75.90500	3.0052038
Syria	9960117.0	4608844.22	2955.9829	543.849794	63.27250	8.4520474
Taiwan	17956478.5	3861424.57	9639.0909	7886.456685	71.82500	3.4436100
Tanzania	20656033.0	8413778.74	870.8338	63.853584	49.03000	1.9940411
Thailand	47971624.0	12297253.02	3070.6277	2186.754314	63.85150	4.8740870
Togo	2901952.0	1195860.40	1341.9795	435.126137	53.91000	5.8704005
Trinidad and Tobago	1079434.0	147410.55	6995.2712	531.347975	67.88100	2.8015571

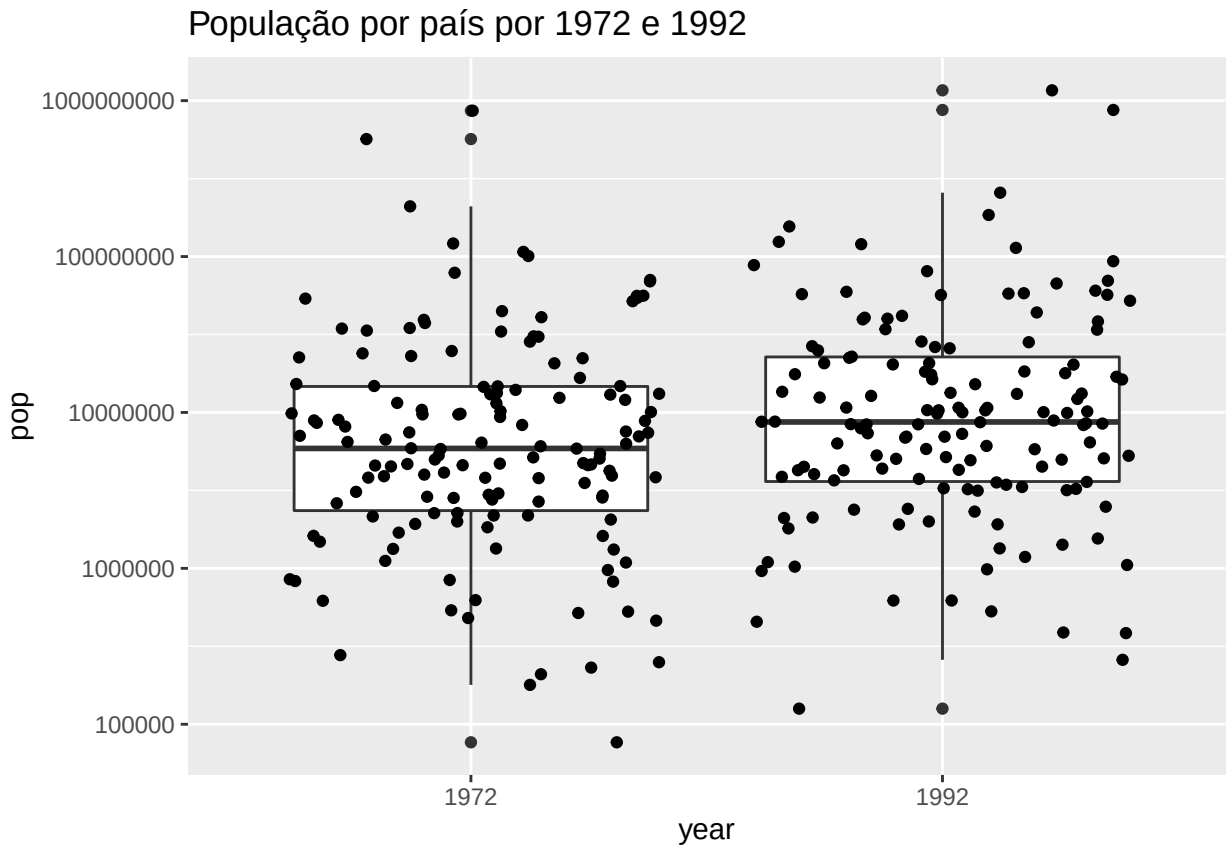
country	mean_pop	sd_pop	mean_gdp	sd_gdp	mean_life	sd_life
Tunisia	6913292.0	2276579.78	3543.0031	1116.828612	62.80150	10.1816305
Turkey	47836048.5	14627345.93	4564.5223	1575.187758	61.57550	6.4636631
Uganda	14221237.5	5700627.69	797.4533	216.774241	49.92050	1.5492710
United Kingdom	56972674.5	1263846.60	19300.1045	4815.380301	74.21500	3.1183409
United States	233395094.5	33232738.15	26904.9841	7211.001628	73.71500	3.3587572
Uruguay	2989394.0	226087.49	6920.2068	1720.812147	70.71250	2.8842886
Venezuela	15890606.0	6187123.52	10619.5930	161.691739	68.43100	3.8452467
Vietnam	57297871.0	17879699.84	844.2624	204.722619	58.95800	12.3093148
West Bank and Gaza	1597175.5	717859.75	4575.5320	2039.469537	63.12500	9.3239100
Yemen, Rep.	10387536.0	4215008.37	1572.2719	434.481509	47.72350	11.1376389
Zambia	6443830.0	2739802.60	1492.1914	397.827914	48.10350	2.8333769
Zimbabwe	8282737.5	3424663.10	746.3915	74.911875	58.00600	3.3531004

O que quer dizer todos esses dados nestes resumos?

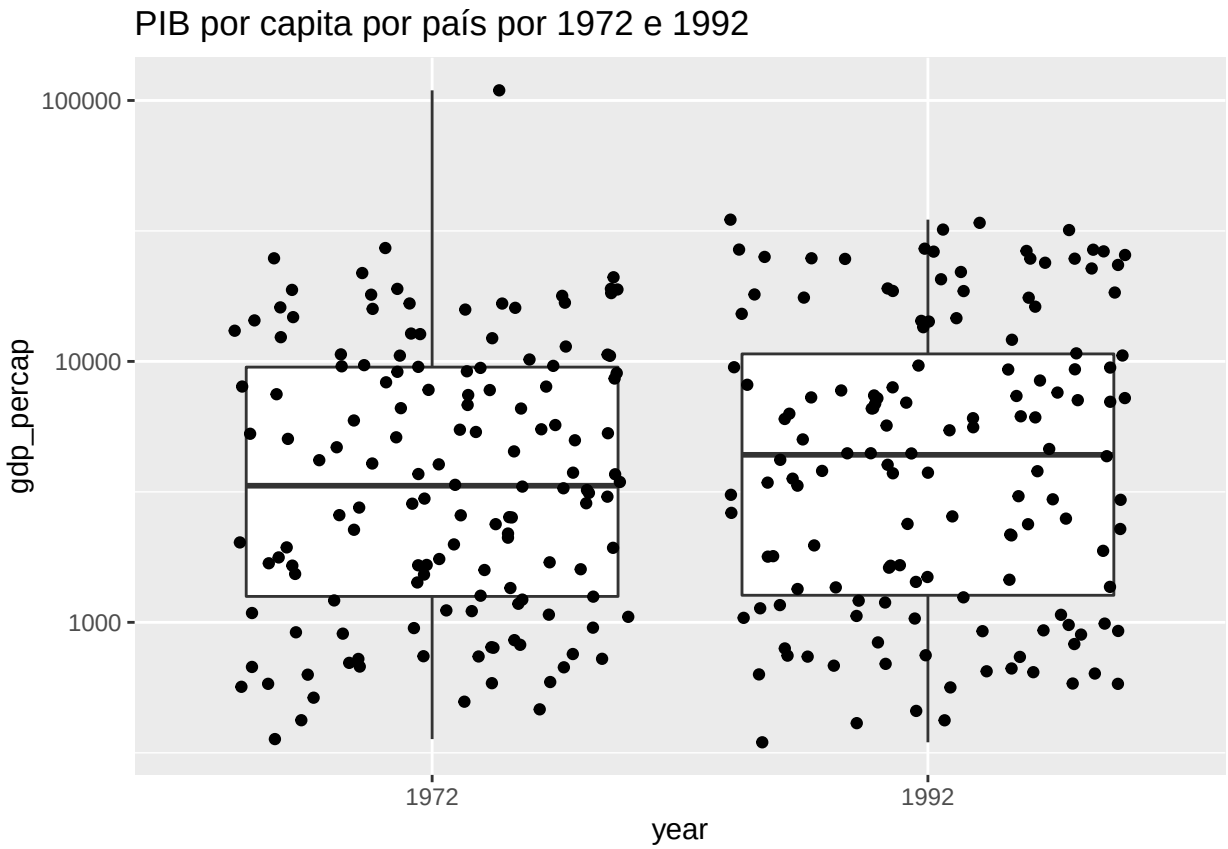
Visualizações

Boxplots dos variáveis `pop`, `life_exp` e `gdp_percap`

```
box_pop <- ggplot(gm, mapping = aes(x = year, y = pop)) +
  geom_boxplot() +
  geom_jitter() +
  scale_y_log10() +
  ggtitle("População por país por 1972 e 1992")

box_gdp <- ggplot(gm, mapping = aes(x = year, y = gdp_percap)) +
  geom_boxplot() +
  geom_jitter() +
  scale_y_log10() +
  ggtitle("PIB por capita por país por 1972 e 1992")

box_life <- ggplot(gm, mapping = aes(x = year, y = life_exp)) +
  geom_boxplot() +
  geom_jitter() +
  scale_y_log10() +
  ggtitle("Expectativa de vida por país por 1972 e 1992")

box_pop
```


População por país por 1972 e 1992
1000000000
100000000
10000000
1000000
100000
pop
1972
1992
year

```
box_gdp
```


PIB por capita por país por 1972 e 1992
100000
10000
1000
gdp_percap
1972
1992
year

```
box_life
```

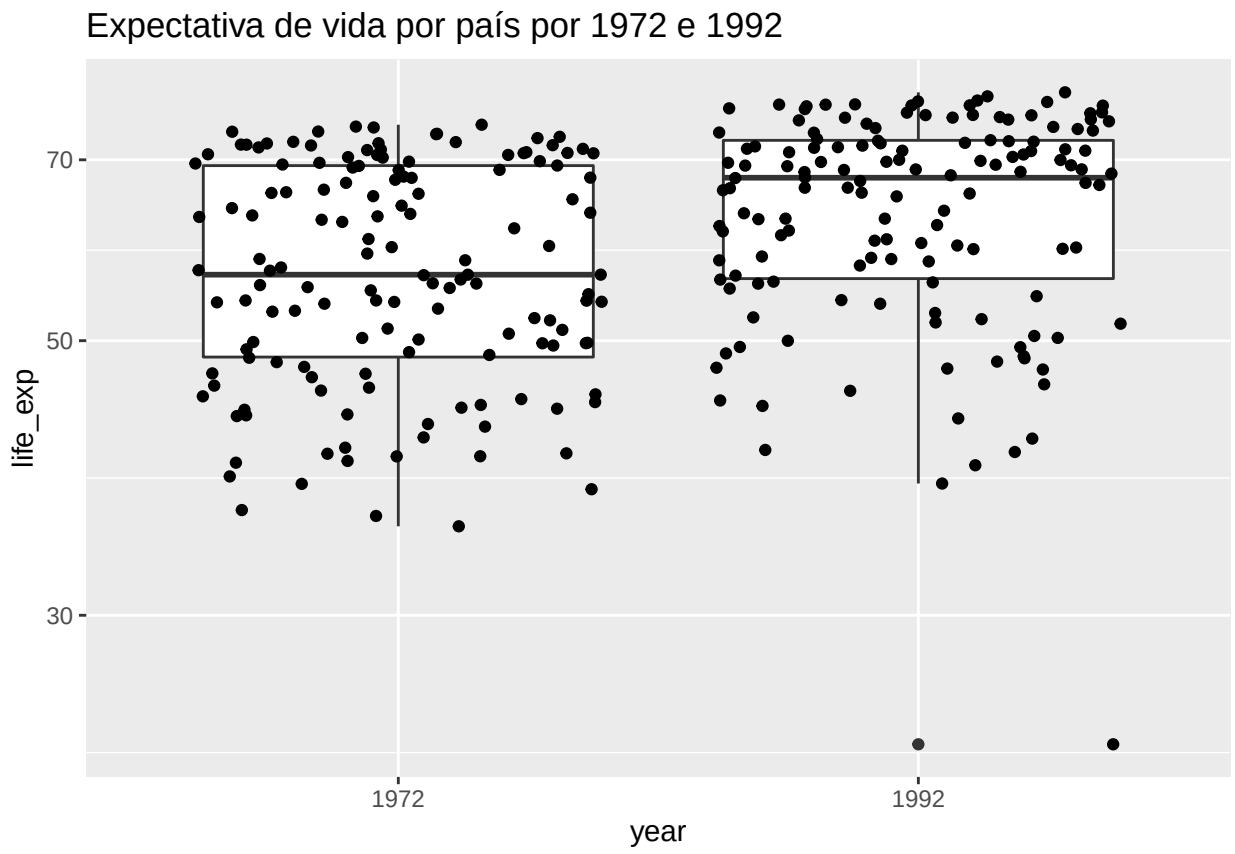
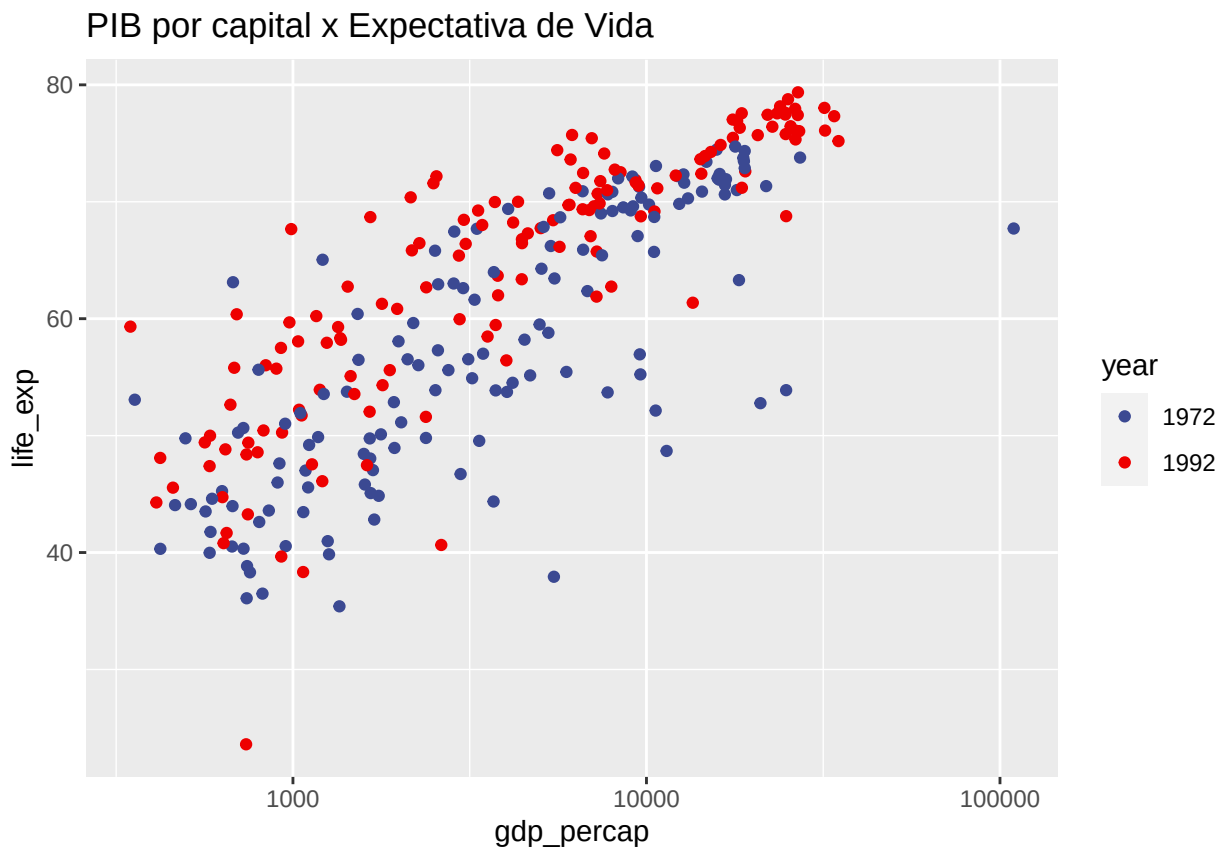


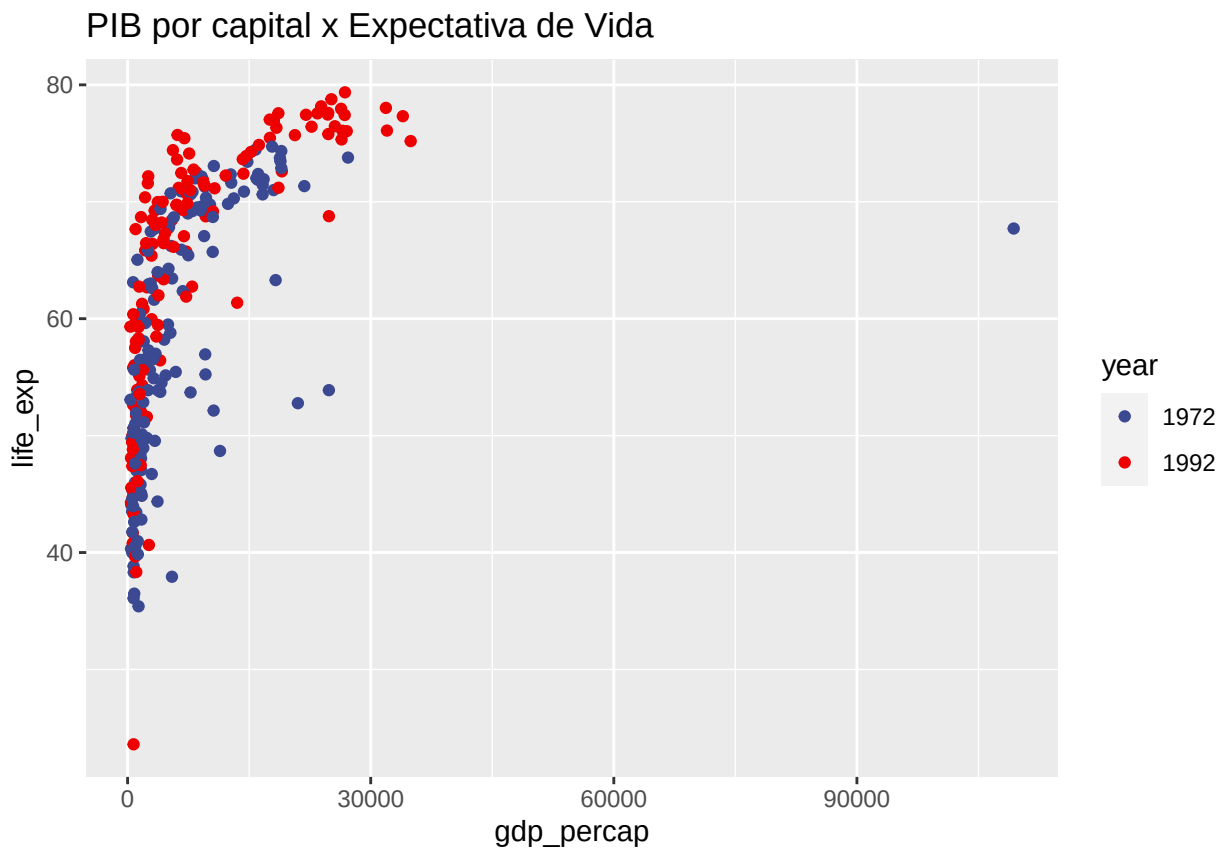
Gráfico de `life_exp` x `gdp_percap`

```
ggplot(gm, mapping = aes(x = gdp_percap, y = life_exp, colour = year)) +
  geom_point() +
  ggtitle("PIB por capital x Expectativa de Vida") +
  scale_x_log10() +
  scale_color_aaas()
```



O que diz este gráfico? As cores vêm de do pacote `ggsci` e uma paleta que uso frequentemente (`scale_color_aaas()`).

Vocês repararam que a escala de eixo x é logarítmico. O que está significa? Eu fiz isso para espalhar mais os pontos. A escala original de `gdp_percap` fez os pontos ficar só no lado esquerdo do gráfico e sugeriu que esta variável não é muito linear. Mais, tem um ponto no extremo (Kuwait de 1972). Este ponto está distorcendo a distribuição de `gdp_percap`. Para este análise, vou tirar Kuwait de análise para refletir melhor a distribuição geral.



Limpeza Final - Tirar Kuwait da Análise

Vou mostrar os cálculos mas vocês precisam mostrar aqui o porque disso. Porque vocês fizeram ajustes finais.

```
gm_mod <- gm %>%
  filter(country != "Kuwait")

descr(gm_mod, stats = c("mean", "sd", "min", "med", "max", "IQR", "CV"))
```

```
## Non-numerical variable(s) ignored: country, continent, year
```

```
## Descriptive Statistics
## gm_mod
## N: 282
##
##                gdp_percap   life_exp            pop
## ------------- ------------ ---------- --------------
##          Mean      7005.65      60.83     30799387.90
##       Std.Dev      7623.37      11.75    108371561.79
##           Min       347.00      23.60        76595.00
##        Median      3745.86      62.74      7428840.50
##           Max     33965.66      79.36   1164970000.00
##           IQR      8315.90      19.95     17201812.25
##            CV         1.09       0.19            3.52
```

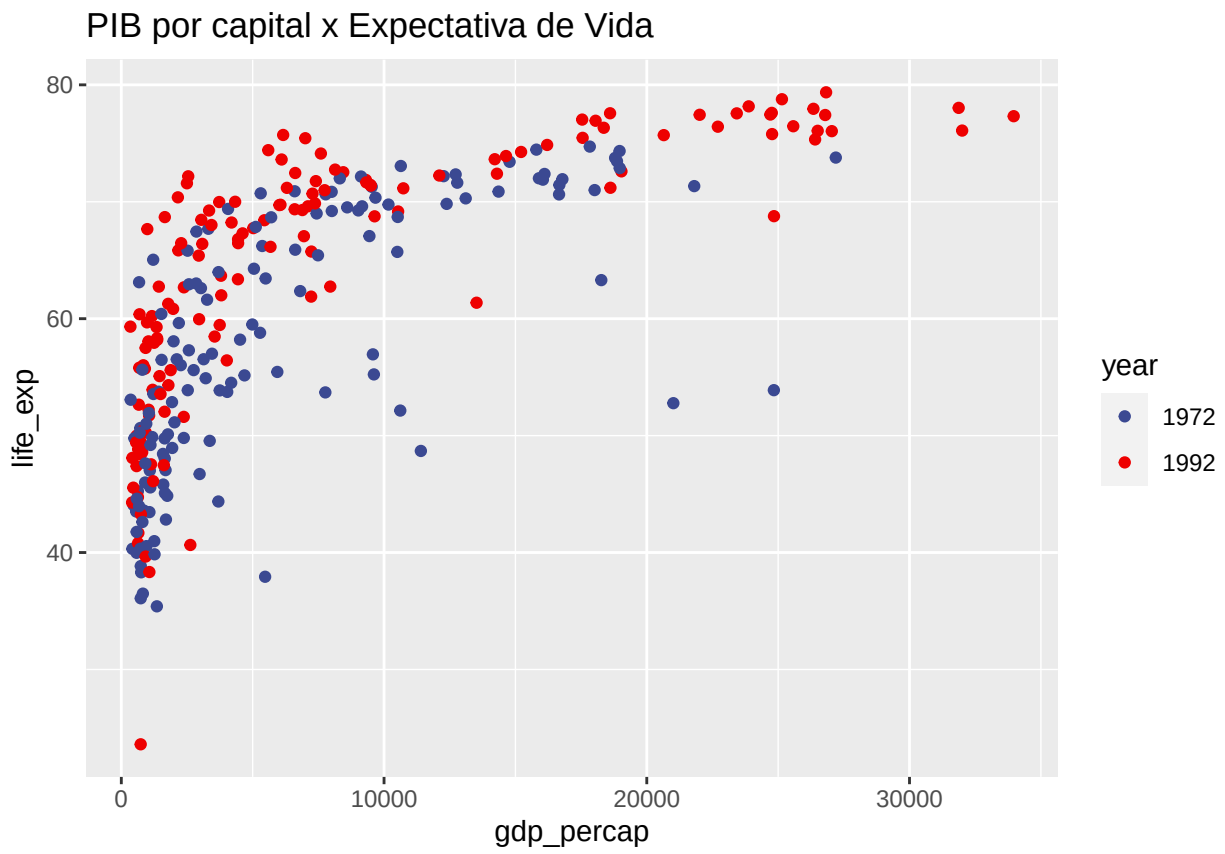
```
# agrupado por ano
gm_mod %>%
```

```
  group_by(year) %>%
  descr(stats = c("mean", "sd", "min", "med", "max", "IQR", "CV"))
```

```
## Non-numerical variable(s) ignored: country, continent
```

```
## Descriptive Statistics
## gm_mod
## Group: year = 1972
## N: 141
##
##                 gdp_percap   life_exp            pop
## ------------- ------------ ---------- --------------
##          Mean      6042.58      57.58    25362661.16
##       Std.Dev      6146.40      11.39    88938878.23
##           Min       357.00      35.40       76595.00
##        Median      3313.42      56.53     5894858.00
##           Max     27195.11      74.72   862030000.00
##           IQR      8188.46      20.82    12092489.00
##            CV         1.02       0.20           3.51
##
## Group: year = 1992
## N: 141
##
##                 gdp_percap   life_exp             pop
## ------------- ------------ ---------- ---------------
##          Mean      7968.72      64.08     36236114.65
##       Std.Dev      8774.97      11.23    124912037.19
##           Min       347.00      23.60       125911.00
##        Median      4332.72      67.66      8718867.00
##           Max     33965.66      79.36   1164970000.00
##           IQR      9288.72      16.51     19128587.00
##            CV         1.10       0.18            3.45
```

```
ggplot(gm_mod, mapping = aes(x = gdp_percap, y = life_exp, colour = year)) +
  geom_point() +
  ggtitle("PIB por capital x Expectativa de Vida") +
  scale_color_aaas()
```



Análise de `life_exp` x `gdp_percap`

O seguinte vai ser uma regressão linear simples entre as duas variáveis de interesse. Deve ser seguido por uma análise do que ela significa. Vou continuar com a transformação logarítmica por causa da tendência quadrática clara na variável `gdp_percap`.

```
fit_basic <- lm(life_exp ~ gdp_percap, data = gm_mod)
summary(fit_basic)
```

```
##
## Call:
## lm(formula = life_exp ~ gdp_percap, data = gm_mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.5306  -5.6795   0.7769   6.3258  16.1143
##
## Coefficients:
##               Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 53.34182178  0.68658281   77.69   <2e-16 ***
## gdp_percap   0.00106874  0.00006638   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.483 on 280 degrees of freedom
## Multiple R-squared:  0.4807, Adjusted R-squared:  0.4789
```

```
## F-statistic: 259.2 on 1 and 280 DF,  p-value: < 2.2e-16
```

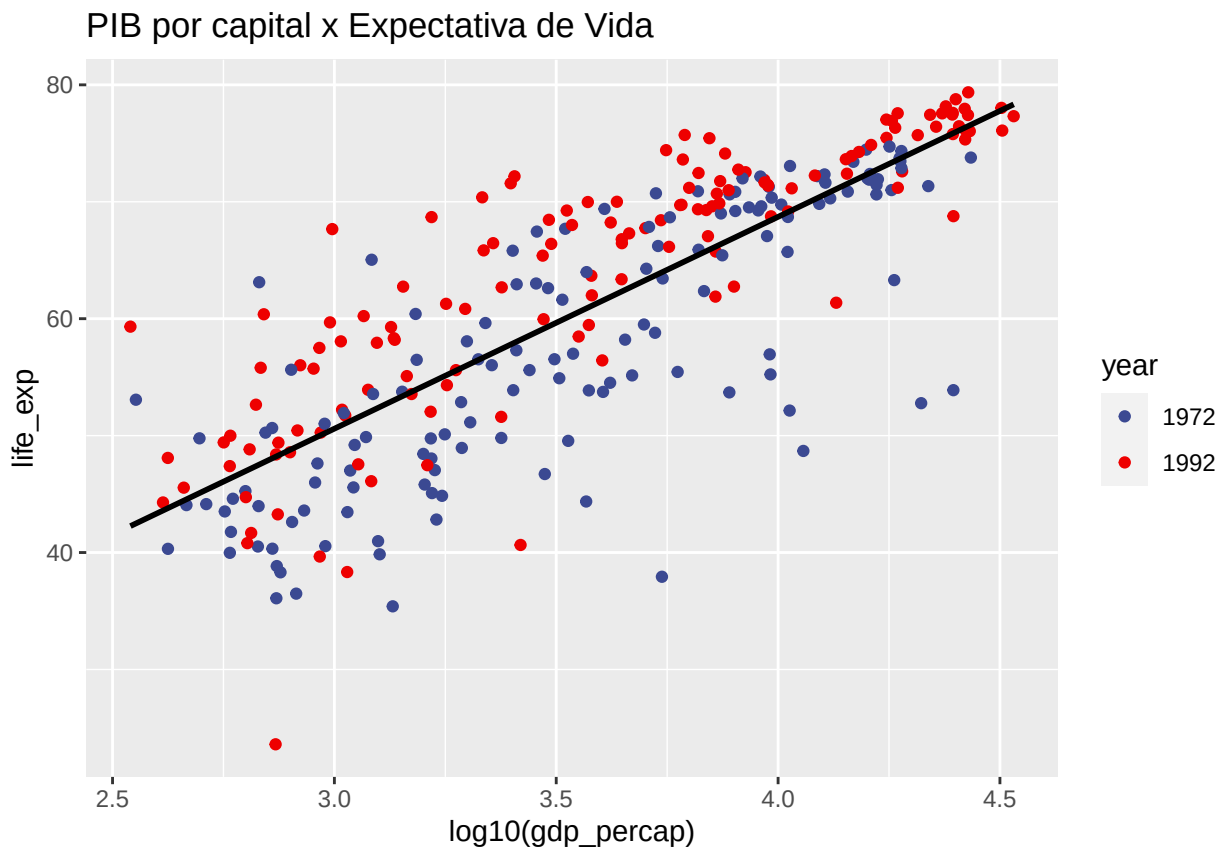
```
fit_log <- lm(life_exp ~ log10(gdp_percap), data = gm_mod)
summary(fit_log)
```

```
##
## Call:
## lm(formula = life_exp ~ log10(gdp_percap), data = gm_mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0407  -3.4566   0.9535   4.0868  17.1562
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.7636     2.8132  -1.338    0.182
## log10(gdp_percap)  18.1187     0.7807  23.209   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.884 on 280 degrees of freedom
## Multiple R-squared:  0.658,  Adjusted R-squared:  0.6567
## F-statistic: 538.6 on 1 and 280 DF,  p-value: < 2.2e-16
```

Gráfico do Modelo

```
ggplot(gm_mod, mapping = aes(x = log10(gdp_percap), y = life_exp, colour = year)) +
  geom_point() +
  ggtitle("PIB por capital x Expectativa de Vida") +
  scale_color_aaas() +
  stat_smooth(method = "lm", color = "black", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Reparem que o R^2 aumenta bastante quando a transformação está aplicada. Só vou testar o modelo `fit_log` porque este é o modelo que acho melhor. Também, precisa anotar o que é o resultado do teste-F do modelo e a teste da inclinação da linha (se a inclinação não é igual a 0)

Validação do Modelo

Aqui deve aplicar pelo menos duas técnicas (gráficos) de validação: a plotagem dos resíduos contra os valores previstos para expectativa de vida pelo modelo

Gráfico dos Resíduos

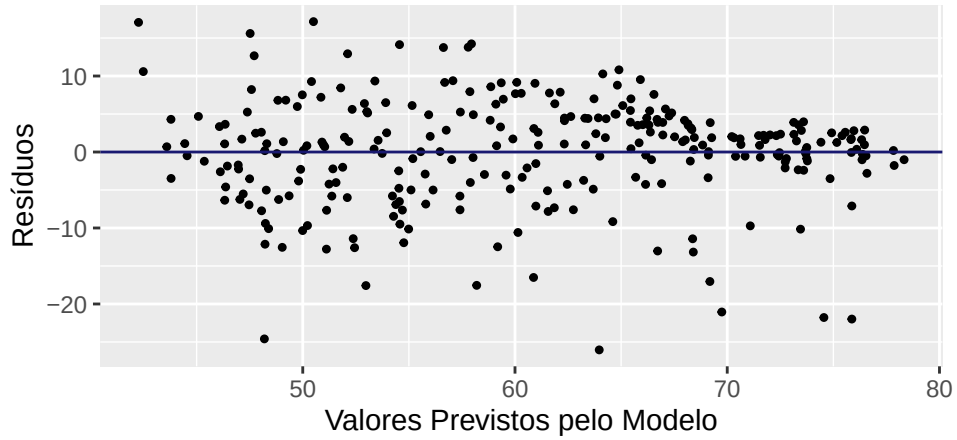
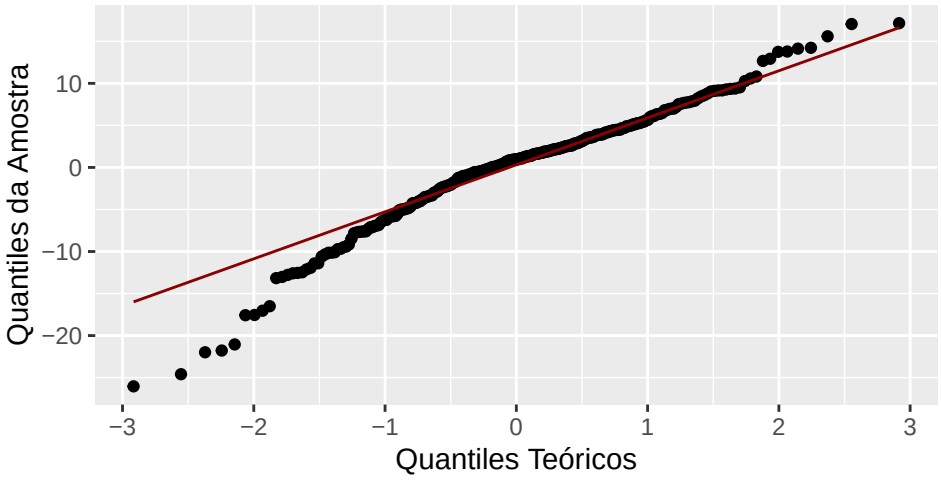


Gráfico Q-Q

```
grqq <- ggplot(data = fit_log, aes(sample = .resid))
grqq <- grqq + stat_qq()
grqq <- grqq + stat_qq_line(color = "darkred")
grqq <- grqq + labs(x = "Quantiles Teóricos",
                    y = "Quantiles da Amostra")
grqq
```



Aqui, os valores dos resíduos divergem da linha que indica normalidade dos resíduos, especialmente no lado esquerda. Este indica *skewness*, ou seja mais valores neste lado da distribuição. O resultado disso é que a curva dos resíduos não é muito normal nas extremidades. O que está poder querer dizer? É para você decider e relatar se esse acontece com sua análise no seu relatório.