Countries GDP per capita and their Happiness Index Rating

Sabawoon Popal

MTH 225

Professor Nara Yoon

December 15, 2021

Countries GDP per capita and their Happiness Index Rating

Introduction

My proposed research question for the final project is: Are people happier in countries that have a higher GDP per capita? My parameter of interest is the correlation of a countries GDP per capita, and the happiness score it received in the given dataset. I chose this topic because in general, it's possible for someone to not be financially wealthy and still be happy. However, is the same true for a country's wealth, and the happiness of its citizens? While a number of 3rd world countries are unstable, are there a good number that indeed have citizens living content?

Data Collection

Existing data was collected from Kaggle.com, posted by Sustainable Development Solutions Network, and the key variables were calculated by using data derived from the Gallup World Poll. The following dataset provided in the research consisted of one table, consisting of multiple quantitative categories converted to ratings such as happiness score, and the ratings of all countries GDP per capita, social support, life expectancy, freedom, generosity, and the people's perception of corruption in their respective countries. For this research topic, the cases we have at hand are the countries, where the variables we will be focusing on are happiness score (Score), and GDP per capita for each country. These are the variables we need to analyze whether or not there is an association with GDP per capita and a countries happiness score. rho (the other categories are not significant to this research topic). The reason why our parameter type is rho is because both Score and GDP per capita are quantitative variables. The scores for each country were calculated based on a question that was given for respondents to answer, where a score of 10 meant having a perfect life, and 0 being the worst life anyone can live through. The values of GDP per capita, and the scores were available for 156 different countries. The sample is the collection of countries reported in this dataset (156), represented as sample_size (Appendix 2, Value Storing, 4). Because this is a sample size, and not a population parameter, this number isn't how many countries there are in the world. That being said, the population on the other hand is all of the countries in the world.

Exploratory Data Analysis

It is important to understand range of values we have for each of the quantitative categories because when performing a confidence interval or hypothesis test. One of the things that is key to understanding the data values that this dataset is providing us is by having a summary of each of the variables, meaning the variables minimum, 25th percentile, both the mean and median, 75th percentile, and maximum. Below are the two summary figures (Appendix 2, Exploratory Data Analysis, 3) for GDP per capita and Happiness Score, respectively:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.6028  0.9600  0.9051  1.2325  1.6840 
```

(Top: Summary for GDP per capita)

(Bottom: Summary for Happiness Score)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.853   4.545   5.380   5.407   6.184   7.769 
```

Based on the summaries above, we find that the range for the GDP per capita variable is simply the GDP per capita maximum because the minimum value recorded for a country in the GDP per capita category is 0. As for the range of Happiness Score, 7.769 – 2.853 = 4.916. This entails that the happiest recorded country in the report has a happiness score 4.916 higher than that of the least happiest recorded country. One of the most important statistics is the correlation in this report between the GDP per capita and the happiness score. The correlation will help us determine the confidence level for breaking the association and determining the hypothesis test results later on in this report. A scatterplot is a great visual (Appendix 2, Exploratory Data Analysis, for dealing with possible correlations between two variables. Below is the scatterplot of all 156 countries data included in the report:

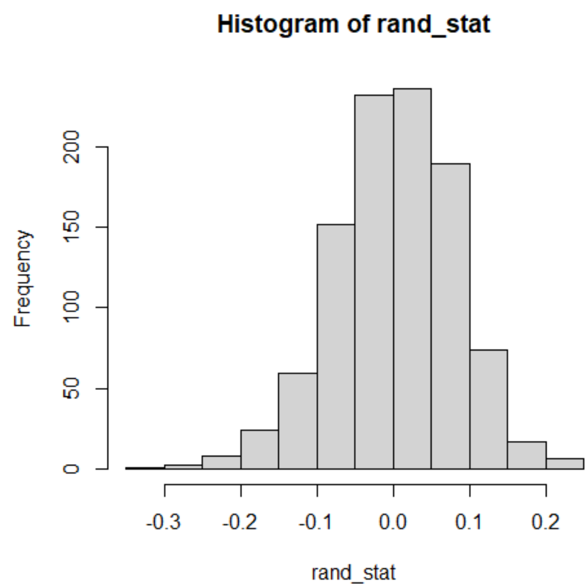
Just by looking at this scatterplot, we can quickly make an inference that this correlation between GDP per capita and the Happiness Score for countries is a moderately positive correlation. To find the exact correlation of the dataset, I assigned the GDP per capita category values to be identifier (x), and making Happiness Scores to be identifier (y). In R, to find the correlation of two quantitative variables, you must use the cor() method with parameters x and y. After doing so, we are returned the value of 0.7938829. I saved this value as (correlation) in R (Appendix 2, Value Storing, 3). This shows that there is a relatively strong association between a countries GDP per capita and the happiness score reported for those countries. However, while there appears to be an association, it is very crucial to remember that association is not causation. Just because the increase of one variable associates with the increase off the other does NOT mean that the increasing of value of one variable causes the increase of another. Getting the correlation for this dataset is very significant, because we can use this as our statistic for determining the confidence interval and hypothesis test in the next sections. Visuals such as histograms, and boxplots would not be supported with a two quantitative variable case study just like this one.

Confidence Interval

Obtaining the confidence interval when the statistic is a correlation coefficient is very different compared to finding confidence intervals for means, and proportions as examples. The first step in finding the confidence interval with a correlation coefficient using a confidence level of 95% is by conducting the fisher's transformation (Zr), defined as Zr = ln((1 + r)/(1 – r))/2. To imitate this function in R, "ln" can be represented as "log", since the log function in R refers to natural log (Appendix 2, Confidence Interval, 1). With the value we obtained for Zr (1.081847), we are now able to find the lower, and upper limits of the exponent we need to plug in which would then help us find the confidence interval. The upper and lower limits share similar functions, where the lower limit value is defined as LL = Zr – (1.96/sqrt(n-3)), and the upper limit value being UL = Zr + (1.96/sqrt(n-3)). Remember that we now have Zr stored as an actual value (we found this through the fishers transformation). Knowing that n represents the sample size, and our sample size is sample_size, we can plug this into n. LL = Zr – (1.96/sqrt(sample_size-3)), and UL = Zr + (1.96/sqrt(sample_size-3)) (Appendix 2, Confidence Interval, . For LL, we get 0.92339, and for UL, we get 1.240303. We can use both of these values to find the confidence interval at 95% confidence level. The confidence level can be defined as [((e^2LL)-1)/(e^2LL+1) , ((e^2UL)-1)/((e^2UL)+1)]. In R, we can store euler's number, e, as notated in the confidence interval, as the value of eulers number (approx. 2.71828) (Appendix 2, Value Storing, 5). After plugging UL and LL with their respective values in that confidence interval (Appendix 2, Confidence Interval, 4-5), we get [0.727497, 0.845542]. This is the confidence interval for the correlation between GDP per capita and happiness score for each country. In deeper context, we are 95% confident that the true population correlation of a countries GDP per capita and its happiness score is between 0.727497 and 0.845542. To ensure that we got the right results when we conducted this confidence interval, we can use the correlation test provided in R (Appendix 2, Confidence Interval, 5). After applying the correlation test, we get [0.7274985, 0.8455412]. The values start to differ at the millionths place of both the lower and upper end of the confidence interval given by R, otherwise these values are nearly identical, and this holds no significance in the way we interpret our results.

Hypothesis Testing

To begin the hypothesis testing, we must first understand that because we are working with correlation, we will focus our hypothesis testing on "breaking the association". That being said, our null hypothesis (H_0) is that there is no positive correlation between a countries GDP per capita and the happiness score it received. On the other hand, our alternative hypothesis (H_a) states that there is a positive correlation between a countries GDP per capita and the happiness score it received. Our null and alternative hypothesis respectively can be written as H_0 $\rho = 0$, and $H_a : \rho > 0$. Our sample data is the correlation of countries GDP per capita and happiness scores, which we obtained from previous section. Before we can conclude whether the null or alternative hypothesis is true, we must create a randomization distribution where we can use to see where our correlation from the sample would land around compared to the distribution. I chose to create 1000 sampling distributions, as defined by rand_n (Appendix 2, Hypothesis Test, 1). Using the code to generate the randomization distribution (Appendix 2, Hypothesis Test, 2), we can display the data generated by a histogram (Appendix 2, Hypothesis Test, 3), as shown below:



Now that we have our histogram of randomization distribution, we can use this to figure out whether or not we should reject the null hypothesis. Finding the p-value (Appendix 2, Hypothesis Test, 4) to be 0 and significance level of 0.05, this gives us solid evidence in favor of the alternative hypothesis. Therefore, we can reject the null hypothesis, and accept the alternative hypothesis, that a countries GDP per capita positively correlates to their happiness score. In context, there is sufficient evidence that shows that the correlation between a countries GDP per capita and its happiness score is positive.

Conclusion

Our hypothesis test helps answer our main question: Are people happier in countries that have a higher GDP per capita? Considering we obtained an extremely small p-value of 0, this only tells us that the countries with a higher GDP per capita have happier people living in them. We also conducted a 95% confidence interval that tells us that the population correlation of a countries GDP per capita and the happiness of its people is between 0.727497 and 0.845542.

Countries GDP per capita and their Happiness Index Rating

Appendix 1

Each countries GDP per capita, and their happiness score as a data frame (Ranked by Score variable. Continued on the next page.)

	Happiness_Rating_by_Country..GDP.per.capita.	Happiness_Rating_by_Country.Score
1	1.340	7.769
2	1.383	7.600
3	1.488	7.554
4	1.380	7.494
5	1.396	7.488
6	1.452	7.480
7	1.387	7.343
8	1.303	7.307
9	1.365	7.278
10	1.376	7.246
11	1.372	7.228
12	1.034	7.167
13	1.276	7.139
14	1.609	7.090
15	1.333	7.054
16	1.499	7.021
17	1.373	6.985
18	1.356	6.923
19	1.433	6.892
20	1.269	6.852
21	1.503	6.825
22	1.300	6.726
23	1.070	6.595
24	1.324	6.592
25	1.368	6.446
26	1.159	6.444
27	0.800	6.436
28	1.403	6.375
29	1.684	6.374
30	1.286	6.354
31	1.149	6.321
32	1.004	6.300
33	1.124	6.293
34	1.572	6.262
35	0.794	6.253
36	1.294	6.223
37	1.362	6.199
38	1.246	6.198
39	1.231	6.192
40	1.206	6.182
41	0.745	6.174
42	1.238	6.149
43	0.985	6.125
44	1.258	6.118
45	0.694	6.105
46	0.882	6.100
47	1.092	6.086
48	1.162	6.070
49	1.263	6.046
50	0.912	6.028
51	1.500	6.021
52	1.050	6.008
53	1.187	5.940
54	1.301	5.895
55	1.237	5.893
56	0.831	5.890
57	1.120	5.888
58	1.327	5.886
59	0.642	5.860

60	1.173	5.809
61	0.776	5.779
62	1.201	5.758
63	0.855	5.743
64	1.263	5.718
65	0.960	5.697
66	1.221	5.693
67	0.677	5.653
68	1.183	5.648
69	0.807	5.631
70	1.004	5.603
71	0.685	5.529
72	1.044	5.525
73	1.051	5.523
74	0.493	5.467
75	1.155	5.432
76	1.438	5.430
77	1.015	5.425
78	0.945	5.386
79	1.183	5.373
80	1.221	5.339
81	1.067	5.323
82	1.181	5.287
83	0.948	5.285
84	0.983	5.274
85	0.696	5.265
86	0.551	5.261
87	1.052	5.247
88	1.002	5.211
89	0.801	5.208
90	1.043	5.208
91	0.987	5.197
92	0.931	5.192
93	1.029	5.191
94	0.741	5.175
95	0.813	5.082
96	0.549	5.044
97	1.092	5.011
98	0.611	4.996
99	0.569	4.944
100	0.446	4.913
101	0.837	4.906
102	0.393	4.883
103	0.673	4.812
104	1.057	4.799
105	0.764	4.796
106	0.960	4.722
107	0.947	4.719
108	0.960	4.707
109	0.574	4.700
110	0.657	4.696
111	0.450	4.681
112	0.000	4.668
113	0.879	4.639
114	0.138	4.628
115	0.331	4.587
116	0.850	4.559
117	1.100	4.548
118	0.380	4.534
119	0.886	4.519

119	0.886	4.519
120	0.308	4.516
121	0.512	4.509
122	0.570	4.490
123	0.204	4.466
124	0.921	4.461
125	0.562	4.456
126	1.043	4.437
127	0.094	4.418
128	0.385	4.390
129	0.268	4.374
130	0.949	4.366
131	0.710	4.360
132	0.350	4.350
133	0.820	4.332
134	0.336	4.286
135	0.811	4.212
136	0.332	4.189
137	0.913	4.166
138	0.578	4.107
139	0.275	4.085
140	0.755	4.015
141	0.073	3.975
142	0.274	3.973
143	0.274	3.933
144	0.489	3.802
145	0.046	3.775
146	0.366	3.663
147	0.323	3.597
148	1.041	3.488
149	0.619	3.462
150	0.191	3.410
151	0.287	3.380
152	0.359	3.334
153	0.476	3.231
154	0.350	3.203
155	0.026	3.083
156	0.306	2.853

Appendix 2

Importing Dataset Downloaded in Excel:

1. **library(readxl)** // Allows R to read excel files.
2. **Happiness_Rating_by_Country <- read_excel("Happiness Rating by Country.xlsx")**

   // Initializes the dataset name in R as Happiness_Rating_by_Country
3. **View(Happiness_Rating_by_Country)** // Attempts to import

Value Storing

1. **y = Happiness_Rating_by_Country$Score** // Happiness score of every country recorded.
   a. **156**
2. **x = Happiness_Rating_by_Country$'GDP per capita'** // GDP per capita of every country recorded.
   a. **156**
3. **correlation = cor(x, y)** // The correlation of Score and GDP per capita (included in analysis of Exploratory Data Analysis)
   a. **.79388**
4. **sample_size <-length(x) OR USE length(y)** // Amount of cases for each variable. (Test both code out to ensure both lengths are the same.
5. **e <- exp(1)** // Eulers number needed for confidence interval.
   a. **2.71828182845905**

Countries GDP per capita and their Happiness Index Rating

Exploratory Data Analysis

1. **lin.reg <- lm(y ~ x, data = Happiness_Rating_by_Country)** // Depending on GDP per capita, shows the linear regression numbers for Score and GDP per capita.
2. **lin.reg$coefficient** // Shows only the intercept and linear rate.
3. **summary(x) , summary(y)** // Shows the summary of both Score and GDP per capita.
4. **plot(x, y, main = "Scatterplot of Happiness Score to GDP Per Capita", xlab = "GDP/Capita", ylab = "Score")** // Shows a scatterplot of all cases (x, y) = (GDP/Capita, Score)

Confidence Interval

1. **Zr <- log((1+correlation)/(1-correlation))/2** // Fisher transformation for confidence interval.
   a. **1.08184656**
2. **LL <- Zr – (1.96/sqrt(sample_size-3)), UL <- Zr + (1.96/sqrt(sample_size-3))** // Lower and upperbounds of the confidence interval for correlation coefficient.
3. **(e^(2*LL)-1)/(e^(2*LL)+1)** // Lower end of our confidence interval
   a. **0.7274972**
4. **(e^(2*UL)-1)/(e^(2*UL)+1)** // Upper end of our confidence interval
   a. **0.845542**

Countries GDP per capita and their Happiness Index Rating

Hypothesis Test

1. **rand_n <- 1000** // Variable for 1000 randomization distributions.
2. **for(ii in 1:rand_n){**

   Score_shuffle <- sample(y) // Generates 1000 random numbers from the sample provided.

   new_stat <- cor(x, Score_shuffle) // Generates a new correlation given the shuffling of scores.

   rand_stat <- append(rand_stat, new_stat)} // Assigns randomization statistic.
3. **hist(rand_stat)** // Histogram of the randomization distribution.
4. **sum(rand_stat > correlation) / rand_n** // P-value for right tail

Appendix A Code

1. **dataframe = data.frame(x, y)** // Shows all the cases and their data for GDP per capita and happiness score

<u>Appendix 3</u>

Derived Data:

[1] ***https://www.kaggle.com/unsdsn/world-happiness?select=2019.csv***