Poli 312 Final Project: GDP and Life Satisfaction in Different Regions of the World

Aviel Fradkine 260859909

2022-12-08

Data origins and preparation

The code I used to generate this dataset is submitted in the file "data_preparation.Rmd"

For this project, I obtained my data from the World Happiness Report for 2022 (henceforth, WHR) (click here for the report page), which draws from subjective well-being data from the Gallup World Poll surveys of cantril life ladder data, and from the World Bank Development Indicators for GDP data (see p. 21 in the 2022 WHR). I used the data they publish that is used to generate the results found in "table 2.1" in the WHR (p. 20). From this data, I selected country, year, life_ladder score, and log gdp. I also made two modifications to the data-set. First, I converted from log gdp per capita to gdp per capita and added it as a column to the data-set. Second, I coded each country with its region corresponding to the World Development Indicators classification.

The survey data used in the World Happiness Report is not complete for each country for each year, since the GWP survey are conducted in waves and don't include each country for each and every wave. Furthermore, around 1% of observations in the data provided by the WHR are missing gdp data due to absence of such data for those countries in those years, but this is not a concern to us because data availability does not seem to be associated with systematic differences in a relationships between GDP and happiness scores.

Additionally, for the purposes of this project we will be treating each country-year survey observation as an independent observation, even though there is likely serial correlation between countries from year to year. Given the large number of countries we draw from, this decision is intended to preserve more information about the relationship between life ladder scores and log gdp for each cluster of observations, where a cluster is a unique country in the dataset (of which we have 166).

Finally, I must acknowledge my gratitude to the My World in Data project, whose article on happiness and life satisfaction (click here for link) inspired this report and first tipped me off about the uniqueness of Latin America & the Caribbean as a region in terms of the uniquely high life satisfaction given its level of GDP.

Questions (and Answers)

Gross Domestic Product (GDP) has been often criticized as an insufficient metric of true human well-being. Higher levels of economic output, it is argued, do not in themselves entail human well-being, an important component of which is overall life satisfaction. In this project, we will attempt to investigate the relationship between GDP (an economic metric) and self-reported levels of life satisfaction to better understand the relationship between the two. Further, we will aim to understand how GDP is associated with self-reported levels of life satisfaction across different regions of the world.

To investigate this relationship, we will draw on data presented in the World Happiness Report for 2022. In this survey, a sample of residents from each country (the sample varies from year to year) are asked to "evaluate their current life as a whole using the mental image of a ladder, with the best possible life for them as a 10 and worst possible as a 0. Each respondent provides a numerical response on this scale, referred to as the Cantril ladder". (WHR, p. 15) The life_ladder score for each observation is a weighted average of answers among the sampled residents of a country over three year. The GDP data is gathered from the World Development Indicators published by the World Bank and the "region" variable encodes each country with its region classification assigned by the World Development Indicators classification.

The dataset we will be using is called "happiness.csv". Here is a table of the variables in the dataset:

Variables	Description
country	Name of country
year	Year
life_ladder	World Happiness Report cantril life satisfaction score
log_gdp	Log GDP per capita
gdp	GDP per capita
region	Region grouping according to World Bank Development Indicator classification

1. Data description:

First, transform 'region' into a factor variable. Plot life_ladder and gdp per capita on one plot with a different color for each region. Do the same but this time use log_gdp instead of gdp. Finally, for ease of interpretation, obtain the average life ladder score and the log_gdp for each country across all the observations for each country, and then plot the average life ladder and gdp scores on a plot with a different colour for each region. Why would one use log_gdp instead of gdp in this analysis? Based on the plots, does there seem to exist a relationship between income and self-reported happiness, and is it different for different regions?

Answer:

First, since GDP has an exponential pattern of growth, it often makes sense to model differences in GDP on a log scale, which normalizes the data into a linear trend over time.

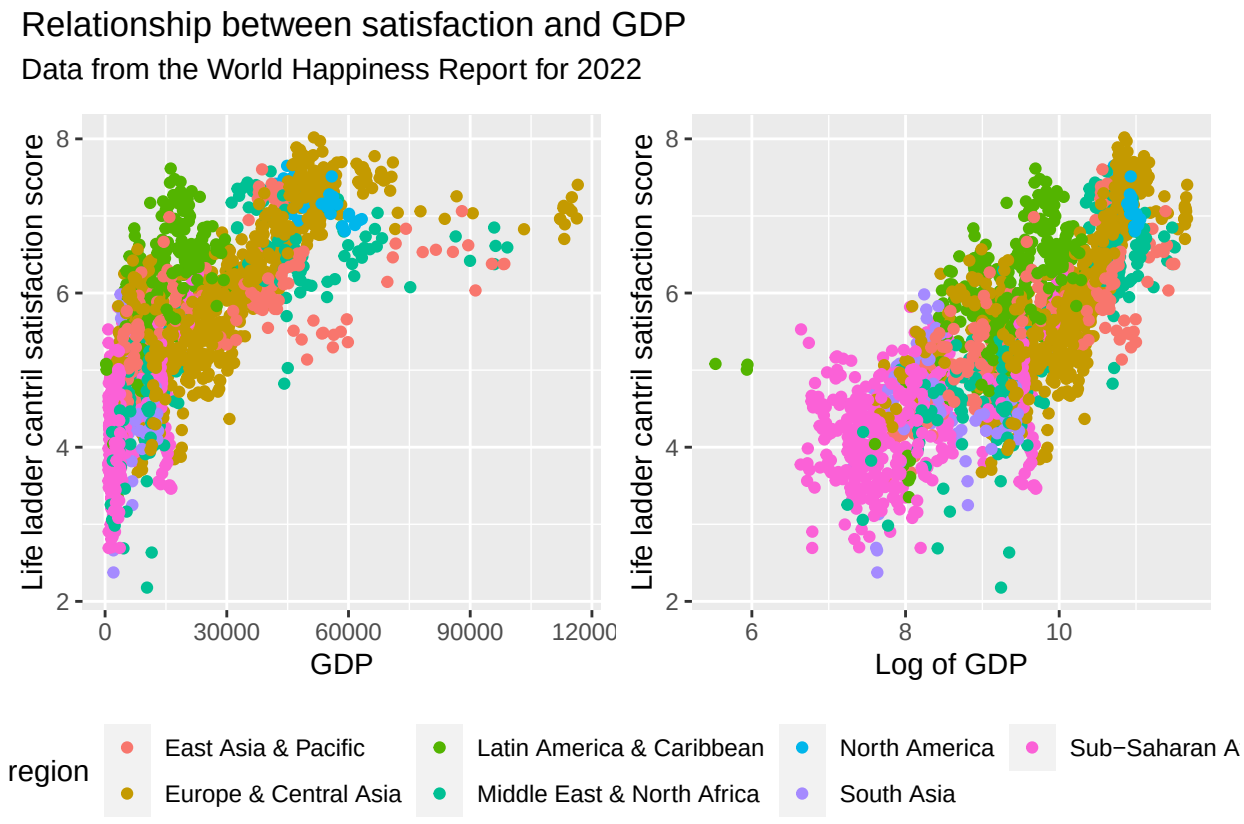
Second, we can see a positive relationship between life ladder scores and GDP; countries with higher GDP scores also have higher average life ladder satisfaction scores. While it is hard to make out any pattern in the data with so many observations, in the plot of average GDP and average life_ladder score by country, we can observe various "clusters" for each region. For example, Latin American & Carribean countries have higher levels of reported life satisfaction at each income level than do other countries at the same income level, Sub-Saharan African countries are clustered in the bottom left of the gdp and life_ladder score plot (with notable spread in life_ladder scores between the countries with the lowest GDP), and North America only includes two countries in the region, which we should keep in mind when we later try to interpret any modeling results.

```
# convert the region variable into a factor variable
h$region <- as.factor(h$region)

# plot the life_ladder scores against gdp
plot <- ggplot(h, aes(x = gdp,
                      y = life_ladder, color = region)) +
        geom_point() +
        labs(x = "GDP",
             y = "Life ladder cantril satisfaction score")

# plot the life_ladder scores against log gdp
log_plot <- ggplot(h, aes(x = log_gdp,
                      y = life_ladder, colour = region)) +
        geom_point() +
        labs(x = "Log of GDP",
             y = "Life ladder cantril satisfaction score")
```

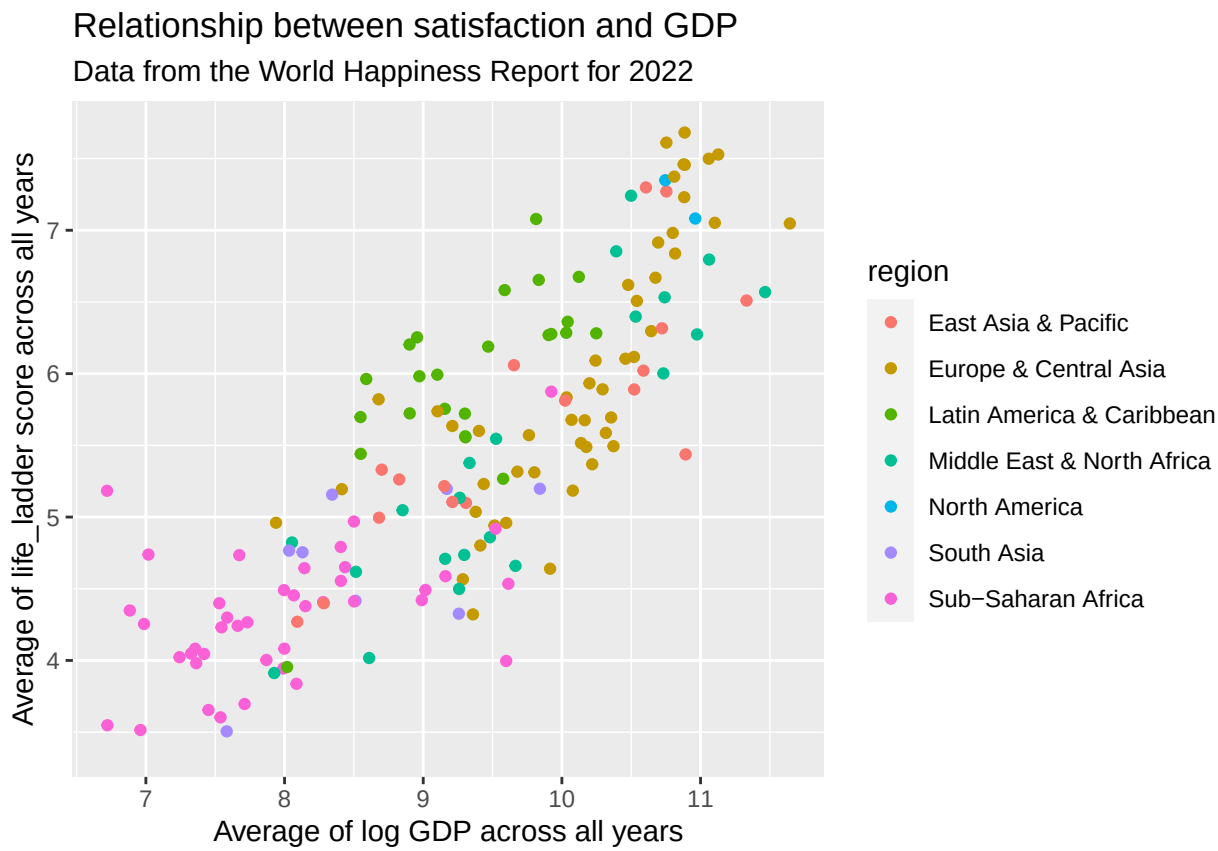
```
# create averages for each by group
means <- h %>%
        group_by(country) %>%
        summarize(avg_hap = mean(life_ladder),
                  avg_gdp = mean(log_gdp, na.rm = T)) # THINK ABOUT WHAT THIS NA RM DOES

# code in region based on country
means <- means %>%
        mutate(region = countrycode(sourcevar = means$country,
                                    origin = 'country.name',
                                    destination = 'region'))
# plot
avg_plot <- ggplot(means, aes(x = avg_gdp,
                      y = avg_hap, color = region)) +
```

```
        geom_point() +
        labs(title = "Relationship between satisfaction and GDP",
             subtitle = "Data from the World Happiness Report for 2022",
             x = "Average of log GDP across all years",
             y = "Average of life_ladder score across all years")
```

```
plot + log_plot + plot_annotation(
        title = "Relationship between satisfaction and GDP",
        subtitle = "Data from the World Happiness Report for 2022") +
        plot_layout(guides = "collect") & theme(legend.position = "bottom")
```




```
avg_plot
```


Relationship between satisfaction and GDP
Data from the World Happiness Report for 2022
Average of life_ladder score across all years
7
6
5
4
7
8
9
10
11
Average of log GDP across all years
region
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub−Saharan Africa

2. Statistical Inference:

Split the data into two groups using the median log_gdp in the dataset: low GDP (below median log gdp) and high gdp (log gdp above median). Run a 95% significance level t-test to check whether self-reported happiness is higher among the high gdp countries than among the low gdp countries. Does the t-test confirm your findings from the previous part on the relationship between log GDP and life ladder scores across countries?

Answer: While there are a number of ways to do split the dataset, I chose to make a binary dummy variable with a value of 1 if a country has log_gdp above the median log_gdp across all observations, and 0 if it has log_gdp below this value. I then ran a Welch Two Sample t-test for a difference in means between the two groups defined by the binary dummy variable.

Running the t-test at the 95% significance level, we see that there is a statistically significant difference in average life_ladder life satisfaction score between the two groups. This supports our findings in the previous part, that countries with higher GDP tend to have higher cantril life satisfaction scores.

```
median_gdp <- mean(h$log_gdp, na.rm = T)
h <- h %>%
        mutate(gdp_group = ifelse(log_gdp < median_gdp, 0, 1))

t.test(h$life_ladder[h$gdp_group == 1],
       h$life_ladder[h$gdp_group == 0],
       data = h)
```

```
    Welch Two Sample t-test

data:  h$life_ladder[h$gdp_group == 1] and h$life_ladder[h$gdp_group == 0]
t = 37.576, df = 2059.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.343005 1.490908
sample estimates:
mean of x mean of y
 6.140790  4.723834
```

3. Data Modeling:

Run a regression of life_ladder on log_gdp including dummy variables for each region and set 'Latin America & Carribean' as your baseline. Then, run a regression in which you interact region and log_gdp, again setting 'Latin America & Caribbean' as your baseline. How do the two regressions differ and what is the proper interpretation of the coefficients and intercepts for each one? Which model do you prefer? Based on these regression results, what can you say about the relationship between life_ladder score and log_gdp for the 'Latin America & Caribbean' countries compared to countries in other regions?

Answer: First, I should note that I couldn't get rmarkdown to output a pdf when using stargazer, so I apologize for using the default summary output in my solution, instead of using the stargazer output.

The regressions provide two different models of the relationship between life_ladder and log_gdp in each country. In the first regression, the one without interaction terms, we assume a constant slope for the relationship between log gdp and life_ladder score across all regions but allow the intercepts for each group to vary. In the second regression, we allow both the slope and the intercept to vary, allowing for the possibility that in addition to different intercepts, countries in different regions also exhibit a different relationship between log_gdp and life_ladder scores. Note that we have set the Latin America and Caribbean region as our baseline for these two regressions. Thus, a proper interpretation of the constant term in both regressions is that it corresponds to the predicted life_ladder cantril score for a country in the Latin America and Caribbean region with 0 log_gdp.

For the first regression, interpretation is rather simple. For all countries, a unit increase in log_gdp is associated with a 0.724 increase in life_ladder cantril score. Furthermore, for countries outside of the Latin America & Carribean region, the coefficient on the "regionRegion Name" term (so, for example "regionEurope & Central Africa") in the output corresponds to the difference in intercept for the regression line for that region and for the Latin America & Caribbean region. We thus see, that with the exception of the North America region (which we saw only has two countries, and whose GDP values are very high, so it might not be so significant), Latin America & Caribbean countries have the highest "baseline" level of life_ladder scores. This indicates that for each income level, Latin America & Carribean countries should, on average, have the highest corresponding level of life_ladder cantril satisfaction scores.

For the second regression, interpretation is somewhat more complicated. For countries in every region other than Latin America and Caribbean, the slope of the regression line is given by the sum of the log_gdp coefficent and the coefficient on log_gdp:regionRegionName. So, for example, the slope for the relationship between life_ladder and log_gdp for a country in South Asia is 0.68369 - 0.51933 = 0.16436 while for a country in Middle East & North Africa it is 0.6839 + 0.24150 = 0.9254. (This corresponds to a 0.1643 and a 0.9245 increase in life_ladder score for a unit increase in log_gdp). For a country in Latin America and Caribbean, the slope is 0.6839 (since it is the baseline factor). Meanwhile, the regionRegionName coefficient corresponds to the predicted intercept for the relationship between life_ladder score and GDP for each region other than Latin America & Caribbean. We can see that allowing the slope to vary between region groups also means that the intercept of the line of best fit must vary widely between groups. The most extreme example of this is the intercept on the North America region, which, due to the negative slope between Canada (avg_hap = 7.35, avg_gdp = 10.7) and the United States (avg_hap = 10.7, avg_gdp = 11), has to be extremely high to fit the data with a linear line, and thus we end up with a very high intercept (outside the upper bound of 10 on the cantril ladder, even).

At face value, using the regression with interaction terms would seem to be more informative because it allows the relationship between log_gdp and life_ladder score to vary across regions. But, we can see that the results we obtain are not so useful for explaining the relationship between log_gdp and life_ladder score, because log_gdp data is only realistically observed in the world between 5 and 12, and the OLS process, in creating a linear line of best fit, provides intercept coefficients and according slope coefficients which, due to the limitations on our log_gdp range and the resulting lines of best fit, do not reflect accurately the actual marginal effect of an increase in log_gdp on happiness at each gdp level. Thus, I prefer the model with no interaction terms.

Finally, we can conclude that, with the exception of the weird (n country = 2) region of North America,

Latin America & Caribbean countries seem to have the highest baseline level of life_ladder scores. That is, at every income level (were it right to extrapolate outside the range of log_gdp actually observed in Latin American & Caribbean countries), Latin America & Caribbean countries would have the highest level of life_ladder scores.

```
# set latin america and the caribbean as our baselines
h$region <- relevel(h$region, ref = 'Latin America & Caribbean')
reg.region <- lm(life_ladder ~ log_gdp + region, h)
reg.region.interact <- lm(life_ladder ~ log_gdp*region, h)

summary(reg.region)
```

```
Call:
lm(formula = life_ladder ~ log_gdp + region, data = h)

Residuals:
     Min       1Q   Median       3Q      Max
-2.91044 -0.42199  0.00433  0.44790  2.19816

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -0.76883    0.17342  -4.433 9.77e-06 ***
log_gdp                          0.72434    0.01814  39.937  < 2e-16 ***
regionEast Asia & Pacific       -0.58826    0.05588 -10.527  < 2e-16 ***
regionEurope & Central Asia     -0.57687    0.04615 -12.499  < 2e-16 ***
regionMiddle East & North Africa -0.83705    0.05601 -14.945  < 2e-16 ***
regionNorth America              0.12247    0.12294   0.996    0.319
regionSouth Asia                -0.72863    0.07803  -9.338  < 2e-16 ***
regionSub-Saharan Africa        -0.71062    0.05366 -13.243  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.6454 on 2054 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.666, Adjusted R-squared:  0.6649
F-statistic: 585.1 on 7 and 2054 DF,  p-value: < 2.2e-16
```

```
summary(reg.region.interact)
```

```
Call:
lm(formula = life_ladder ~ log_gdp * region, data = h)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8209 -0.3980 -0.0094  0.4237  1.8536

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -0.38895    0.49110  -0.792 0.428448
log_gdp                                0.68369    0.05242  13.043  < 2e-16
regionEast Asia & Pacific             -0.83374    0.62795  -1.328 0.184417
regionEurope & Central Asia           -3.30921    0.58131  -5.693 1.43e-08
regionMiddle East & North Africa      -3.16291    0.64031  -4.940 8.46e-07
regionNorth America                   22.65765   10.03211   2.259 0.024019
regionSouth Asia                       3.58194    1.07207   3.341 0.000849
regionSub-Saharan Africa               2.29072    0.57893   3.957 7.86e-05
log_gdp:regionEast Asia & Pacific      0.02682    0.06597   0.407 0.684337
log_gdp:regionEurope & Central Asia    0.27343    0.06074   4.501 7.13e-06
log_gdp:regionMiddle East & North Africa  0.24150    0.06730   3.588 0.000341
log_gdp:regionNorth America           -2.07059    0.92462  -2.239 0.025237
log_gdp:regionSouth Asia              -0.51933    0.12504  -4.153 3.41e-05
log_gdp:regionSub-Saharan Africa      -0.38188    0.06483  -5.891 4.48e-09

(Intercept)
log_gdp                               ***
regionEast Asia & Pacific
regionEurope & Central Asia           ***
regionMiddle East & North Africa      ***
regionNorth America                   *
regionSouth Asia                      ***
regionSub-Saharan Africa              ***
log_gdp:regionEast Asia & Pacific
log_gdp:regionEurope & Central Asia   ***
log_gdp:regionMiddle East & North Africa ***
log_gdp:regionNorth America           *
log_gdp:regionSouth Asia              ***
log_gdp:regionSub-Saharan Africa      ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6124 on 2048 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.7001,    Adjusted R-squared:  0.6982
F-statistic: 367.8 on 13 and 2048 DF,  p-value: < 2.2e-16
```

4. Prediction:

Calculate the predicted cantril life_ladder score for a representative country in each region for log_gdp equal to 5.5, 6.5, 7.5, 8.5, 9.5, 10.5, and 11.5 using the regression with no interaction terms from the previous part. Put the results in a table. Then do the same with the regression with interaction terms. How are we to interpret this table? What can you say about the relationship between life_ladder scores for each region? Are there limitations on our predictions? What additional limitations are there on using the regression with interaction terms for these predictions?

Answer: Generating new data and then applying it to the predict function, we get two tables, one for each regression. The results in the table should be interpreted as the predicted average life_ladder score for a country given its log_gdp level depending on the region to which it belongs. At each level of log_gdp, the table provides us a way of comparing how the predicted life_ladder score would vary depending on the region to which a hypothetical country with that log_gdp level belongs. Based on the predicted values in the table that uses the coefficients from the regression with no interaction terms, we can see that Latin America and & Caribbean countries generally have the highest level of predicted life_ladder life satisfaction cantril scores at each level of log_gdp.

A general caution we should have is that we are here trying to extrapolate outside the existing data for many regions (e.g. countries in the Subsaharan Africa region which only have lower levels of log_gdp in our data and North America which has only high GDP countries). The predictions are most accurate when they cover levels of log_gdp which are actually observed in the dataset for countries within that dataset, and we should be cautious about assigning any significance to predicted happiness scores for regions which have few or no countries with the specified log_gdp level. For example, the predicted levels of happiness for Sub-Sahara Africa region given higher income levels are predictions which extrapolate outside the range of our data for typically low log_gdp countries in that region.

Looking at the model with interaction terms, we can note that given the aforementioned (see previous question) variation in the way the lines are fitted to the existing data under OLS and the odd coefficients that result, we obtain predicted life_ladder cantril scores that are even more questionable and harder to interpet. For example, given the very high level of GDP for the two countries in the dataset for the North America region (both Canada and America have some of the highest log gdp among all the countries in the dataset), the interaction model assigns a very large intercept to that line and then decreases it with a significant slope. Using that slope coefficient to predict cantril scores at low levels of log_gdp provides highly unrealistic predicted cantril scores. Thus, the interaction model is a poor choice for generating predictions about predicted cantril score at each level of log_gdp. The second table of predictions is less preferable than the first, and if we seek to make predictions (especially predictions outside the range of data we observe initially) we should use the coefficients from the model with no interaction terms.

```
# vector of regions
regions <- unique(h$region)

#vector of desired values
log_gdp <- seq(from = 5.5, to = 11.5, by = 1)

# create region log gdp pairs for prediction
newdata <- data.frame(region = 0, log_gdp = 0)
for (region in regions){
        for(gdp_val in log_gdp){
                new_row <- data.frame(region = region, log_gdp = gdp_val)
                print(new_row)
                newdata <- rbind(newdata, data.frame(new_row))
        }
}
# drop first row
newdata <- newdata[2:nrow(newdata), ]
```

```
# re-index rownames from 1 to length of nrow after having dropped row
rownames(newdata) <- 1:nrow(newdata)
```

```
#with no interaction
pred_hap <- round(predict(reg.region, newdata = newdata), 3)
pred_hap_region <- cbind(newdata, pred_hap)
by_country <- pivot_wider(pred_hap_region, names_from = log_gdp,
                          values_from = c(pred_hap))
# with interaction
pred_hap_int <- round(predict(reg.region.interact, newdata = newdata), 3)
pred_hap_region_int <- cbind(newdata, pred_hap_int)
by_country_int <- pivot_wider(pred_hap_region_int, names_from = log_gdp,
                          values_from = c(pred_hap_int))
```

```
kable(by_country,
      caption = "Predicted life ladder scores at log GDP levels by region, no interactions model")
```

Table 2: Predicted life ladder scores at log GDP levels by region, no interactions model

region	5.5	6.5	7.5	8.5	9.5	10.5	11.5
South Asia	2.486	3.211	3.935	4.659	5.384	6.108	6.832
Europe & Central Asia	2.638	3.363	4.087	4.811	5.536	6.260	6.984
Middle East & North Africa	2.378	3.102	3.827	4.551	5.275	6.000	6.724
Sub-Saharan Africa	2.504	3.229	3.953	4.677	5.402	6.126	6.850
Latin America & Caribbean	3.215	3.939	4.664	5.388	6.112	6.837	7.561
East Asia & Pacific	2.627	3.351	4.075	4.800	5.524	6.248	6.973
North America	3.338	4.062	4.786	5.511	6.235	6.959	7.684

```
kable(by_country_int,
      caption = "Predicted life ladder scores at log GDP levels by region, interactions model")
```

Table 3: Predicted life ladder scores at log GDP levels by region, interactions model

region	5.5	6.5	7.5	8.5	9.5	10.5	11.5
South Asia	4.097	4.261	4.426	4.590	4.754	4.919	5.083
Europe & Central Asia	1.566	2.523	3.480	4.437	5.395	6.352	7.309
Middle East & North Africa	1.537	2.462	3.387	4.312	5.237	6.163	7.088
Sub-Saharan Africa	3.562	3.864	4.165	4.467	4.769	5.071	5.373
Latin America & Caribbean	3.371	4.055	4.739	5.422	6.106	6.790	7.473
East Asia & Pacific	2.685	3.396	4.106	4.817	5.527	6.238	6.948
North America	14.641	13.254	11.867	10.480	9.093	7.706	6.319

5. Limitations:

Given the results we have observed, can we provide a causal interpretation to the findings we have about the relationship between log_gdp and life_ladder cantril life satisfaction scores? Further, what should we make of the differences we observe by region in relationship between log_gdp and life_ladder scores?

Answer: No. The data are observed through surveys and not in a randomized control setting. We cannot randomly assign different levels of log_gdp to different countries and see what their happiness scores would be. All that we have is observational data on the levels of GDP and self reported life satisfaction among different country. Furthermore, we do not know what determines life_ladder scores. GDP might be correlated with other variables which actually cause life satisfaction that is reflected in the cantril scores, and thus we cannot assume that differences in GDP are what determines differences in life satisfaction. (This would be an instance of ommitted variable bias). Indeed, the research we have done which shows differences in "baseline" level of happiness across different regions suggests that there might be other important factors (which might vary in different regions even among countries with the same level of gdp) that affect life satisfaction. This creates further reservations about any interpretation we might try to provide as to the effect of GDP on life ladder cantril scores. Indeed, even our conclusion about the different "baseline" levels of life_ladder scores between countries in different regions should not be taken as a causal relationship where belonging to a certain region is directly affects happiness (e.g. it seems more plausible to assume that there are other unobserved factors at play here which vary by region).

Thus, while our research does support the conclusion that there exists a statistically significant association between GDP and self-reported measures of well-being (at least cantril life ladder scores), we should be careful not to cast this as a causal conclusion as to the drivers of self reported well-being, which our analysis does not address.

Appendix

I did not ask for this in the questions, but here is a scatterplot with average cantril scores and log gdp values by country from part 1, but with lines of best fit overlaid. (Note that these lines reflect the model with interaction terms, since the slopes are allowed to vary by region.) This is a good way to see that Latin America & Caribbean region has the highest "baseline" level of life satisfaction for each level of log_gdp, as demonstrated by the height of the line of best fit for that region compared to the others. We can also see that the slope for the relationship between life_ladder scores and log_gdp varies somewhat for different regions, e.g. between Sub-Saharan Africa and East Asia and Pacific, but overall there seems to be a consistent upwards relationship between life ladder and log_gdp for countries in all regions with mostly similar slopes. Finally, we can see that for some regions, the relationship is better approximated with a non linear line of best fit (e.g. Europe & Central Asia) and also we can see that for North America there is insufficient data to draw any meaningful conclusion on the relationship between log_gdp and life_ladder cantril scores.

```
## plot with linear lines
avg_plot_lin <- ggplot(means, aes(x = avg_gdp,
                    y = avg_hap, color = region)) +
        geom_smooth(method = "lm", fill = NA) +
        geom_point() +
        labs(title = "Relationship between life satisfaction and GDP",
             subtitle = "Data from the World Happiness Report for 2022",
             x = "Average of log(GDP) across all years",
             y = "Average of life_ladder cantril score across all years")
avg_plot_lin
```