Report of the Final project

Group member: Zihao Li, Jake White, Maxwell Gehred

Github Link: https://github.com/jwhitedubs/FinalProject_STAT433

Introduction

Happiness is one of the few emotions which is valued almost universally. No matter where you were raised, it's likely that living a happy life is important to you. Indeed, most actions taken by both states and individuals are ultimately in pursuit of happiness. For example, one might get a degree to make more money in order to improve their long-term wellbeing. Or a state might implement a welfare program, with the ultimate goal of making their citizens happier.

Considering how much work is done to increase well-being, it's surprising how little attention is paid to stats which measure happiness. If many of a state's policies work to make their citizens happier, it would make sense to measure their wellbeing to test if their interventions are actually effective. Economists pay close attention to GDP figures to assess the state of the economy, why not pay similar attention to wellbeing? After all, few of us are interested in economic growth for its own sake. If we care about a state's happiness we should measure it directly, instead of relying on proxies like GDP. This way, we can better identify the factors which affect a state's wellbeing. That is the goal of this project.

Thesis

Happiness is widely valued and sought after. If we care about increasing wellbeing, it's important to understand the factors which affect it. This study will investigate five in particular: population density, pollution, birth rate, income inequality, and working hours. This will be done using data collected from "Our World in Data". Our group is interested in regressing the average values of all 5 factors respectively onto the mean value of the life ladder score. In doing so, we hope to analyze the correlation (if any) between each factor and the life ladder. This could provide better insight into what affects happiness and what efforts should be taken to improve world happiness. - edited by Jake

Data Cleaning

Happiness Metric

Helliwell, John F., et al. "World Happiness Report 2019." *The World Happiness Report*, 20 Mar. 2019, https://worldhappiness.report/ed/2019/?utm_source=link_wwwv9&utm_campaign=item_319340&utm_medium=copy.

The World Happiness Report measures subjective well-being. The "Life Ladder" score represents a national average answer to the following question: "Please imagine a ladder, with

steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?". The answers are self reported via an at-home survey.
Only year, country, and life ladder scores were used in our analysis. Data cleaning consisted simply of filtering other variables out.

Suicide Rates

"Gho | by Category | Suicide Rate Estimates, Crude - Estimates by Country." *World Health Organization*, World Health Organization, https://apps.who.int/gho/data/node.main.MHSUICIDE?lang=en.

According to this report, suicide rates are derived using self reported data from WHO member states and modeling from the Global Burden of Disease study from the Instititue of Health Metrics and Evaluation. Due to the use of standardized categories and definitions WHO estimates may differ from the official estimates of member states. These estimates represent the best evidence available, and may not be endorsed by member states. The report states their method for deriving estimates involves the "redistribution of deaths of unknown sex/age and deaths assigned to ill-defined codes, interpolation/extrapolation of number of deaths for missing years, scaling of total deaths by age and sex to WHO all-cause envelopes for 2000–2019, and use of population estimates from the UN Population Division"

Data was cleaned by pivoting the data such that country, year, and sex are individual columens.

Air Pollution

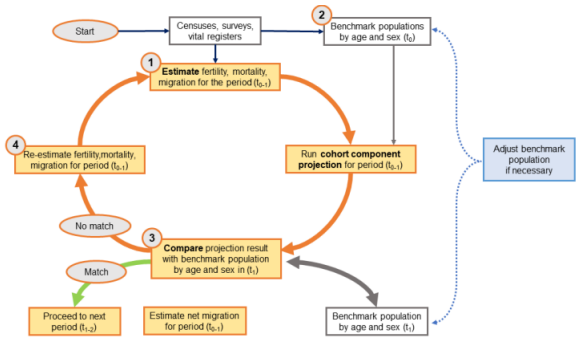
"GBD Results." *Institute for Health Metrics and Evaluation*, https://vizhub.healthdata.org/gbd-results/.

Data is gathered from the IHME Global Burden of Disease survey. The study considers both anthropogenic and natural air pollution and its effect on death rates. It includes deaths caused by both particulate matter pollution and ozone pollution. The data is represented as a percentage of deaths caused by particulate matter pollution and ozone pollution, so, critically, this data will be affected by the way people die, besides by air pollution.

Fertility Rate

"World Population Prospects - Population Division." *United Nations*, United Nations, https://population.un.org/wpp/Download/Standard/MostUsed/.

Fertility rate estimates the average number of children per woman by year. It is estimated using census data from each country as a starting point. Census data is then analyzed for it's completeness and edited according to a protocol laid out in detail here (Pg 3). The outline for the process is diagrammed below.



Data cleaning consisted of selecting for fertility rate, and renaming the associated variables.

Income Inequality

Data Catalog, https://datacatalog.worldbank.org/search/dataset/0041738.

Income Inequality is measured by country in terms of a Ginis coefficient. The Ginis coefficient is a measure of the variation between two variables, in this case, income. A higher Ginis coefficient indicates greater income inequality. These coefficients are generated from income data taken directly from representative samples of household surveys. The world bank data catalog contains standardized Ginis coefficients made up of data from nine separate studies, which generate their own Ginis coefficients. The standardization process is described in depth here.

Cleaning the dataset was relatively straightforward. We first filtered for countries whose income inequality was tracked via Ginis coefficient. Ginis coefficient are typically displayed as decimals, so we divided our values by 100. Lastly, we selected for just the country, year, and mutated Ginis coefficient to remove irrelevant variables.

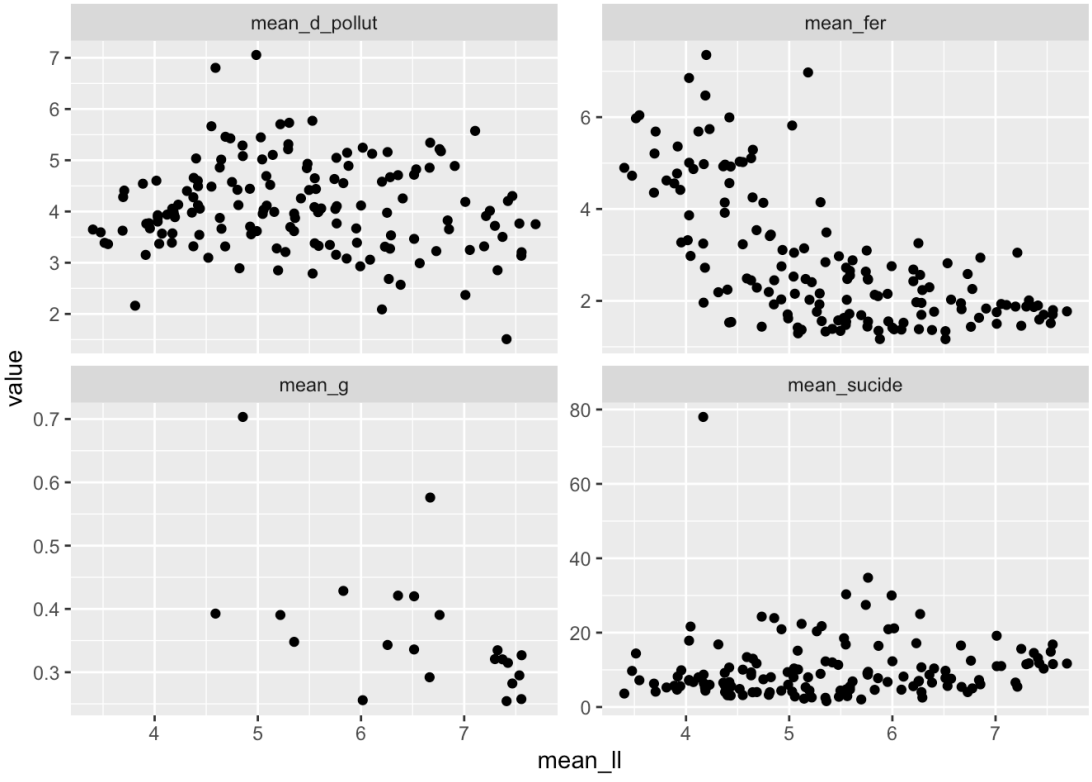
Method

In order to assess which factors impact subjective wellbeing, we searched for correlations between our factors and life ladder scores. To do this we used linear regression to test for a statistically significant relationship between a given factor and our life ladder scores, since linear regression is the most obvious approach to decide whether two factors have a certain correlation. We chose a linear model as default, because there's little reason to believe that wellbeing would have any other relationship with any of our factors. For example, it's implausible that wellbeing increases exponentially as pollution declines. We decided that linear relations, in each case, are far more likely.

If GDP per capita were highly correlated with our factors, and was significantly correlated with subjective well being then our analysis would mistake our factors for driving subjective well being, when GDP per capita was the real driver. To address this concern, we controlled for GDP per capita by splitting the countries with above and below average GDP per capita, and repeating the method described above.

Result

We first found that the mean of life ladder has a negative correlation with the mean of fertility rate and the mean of income inequality over countries.



In the scatter plots shown above, we set the X-axis to be the mean value of each factor (pollution, fertility rate, income inequality, and suicide rate) and Y-axis to be the mean value of the life ladder score. Each point represents a country and the number of points in the scatter plot do not exceed the number of countries in the data frame. As shown above, there's a clear negative correlation between the mean value of life ladder and mean value of fertility rate & income inequality. To support our claim of negative correlations within the fertility rate and income inequality scatterplots, we fit linear models to each of the factors and the summary is shown below.

```
summary(life_lm_f)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_fer, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65416 -0.57622 -0.07141  0.52326  1.90647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.76124    0.14578   46.38   <2e-16 ***
## mean_fer    -0.47760    0.04555  -10.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8285 on 152 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.4197, Adjusted R-squared:  0.4159
## F-statistic: 109.9 on 1 and 152 DF,  p-value: < 2.2e-16
```

```
summary(life_lm_p)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_d_pollut, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04966 -0.89178 -0.01804  0.78735  2.25236
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.8669     0.4238  13.845   <2e-16 ***
## mean_d_pollut  -0.1139     0.1014  -1.123    0.263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.087 on 151 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.00828,    Adjusted R-squared:  0.001713
## F-statistic: 1.261 on 1 and 151 DF,  p-value: 0.2633
```

```
summary(life_lm_g)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_g, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8407 -0.3730  0.2637  0.5733  1.1270
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3305     0.6198  13.440 1.79e-11 ***
## mean_g       -4.8385     1.6390  -2.952  0.00788 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7936 on 20 degrees of freedom
##   (143 observations deleted due to missingness)
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.2687
## F-statistic: 8.715 on 1 and 20 DF,  p-value: 0.007879
```

```
summary(life_lm_s)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_sucide, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92829 -0.89644 -0.08992  0.86056  2.28240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.294438   0.144404  36.664   <2e-16 ***
## mean_sucide 0.009856   0.010773   0.915    0.362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.101 on 138 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.006028,   Adjusted R-squared:  -0.001175
## F-statistic: 0.8369 on 1 and 138 DF,  p-value: 0.3619
```

Note that “s, g, f, p” represents suicide rate, income inequality, fertility rate and pollution respectively. The variable lm_life_x represents the linear model between the mean of the life ladder score and the mean of the factors, where x is in {s, g, f, p}. As we can see from the summary of the linear model, the p values of lm_life_f and lm_life_g are approaching 0, which reflects that the negative correlations between the mean of life ladder and mean of income inequality and fertility rate are statistically significant respectively. On the other hand, the p values of lm_life_p and lm_life_s are greater than 0.05, meaning that the correlation shown in these linear models are not statistically significant. We notice that there is an outlier in the scatter plot of the mean of suicide rate. Consequently, we tried to find out which country it is that has a significant higher mean suicide rate than other countries.

#outlier of susicide rate

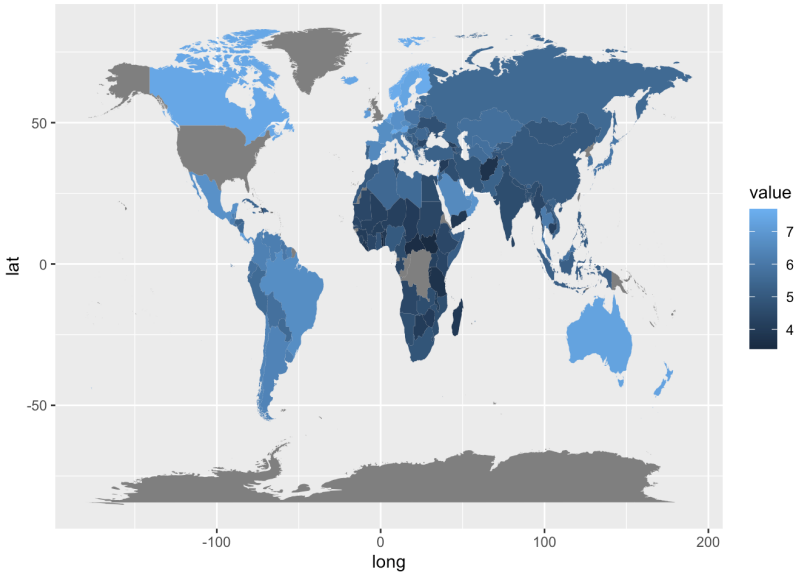
```
df1 %>%
  arrange(desc(mean_sucide))
```

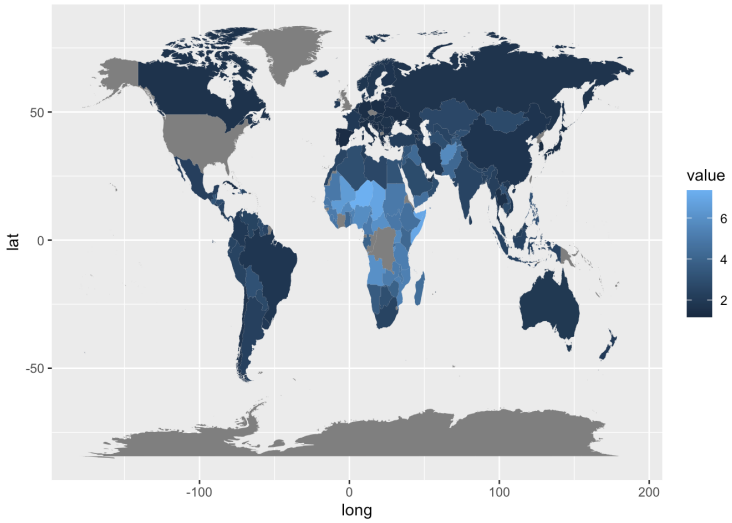
```
## # A tibble: 165 × 6
##    country      mean_ll mean_fer mean_sucide mean_d_pollut  mean_g
##    <chr>          <dbl>    <dbl>       <dbl>          <dbl>   <dbl>
##  1 Lesotho         4.17     3.25          78           3.39 NaN
##  2 Lithuania       5.76     1.55          34.8         3.77 NaN
##  3 Belarus         5.55     1.55          30.3         4.65 NaN
##  4 Guyana          5.99     2.76          30           2.93 NaN
##  5 Kazakhstan      5.74     2.64          27.5         4.63 NaN
##  6 Suriname        6.27     2.57          25           2.68 NaN
##  7 Ukraine         4.73     1.44          24.3         5.43 NaN
##  8 South Africa    4.85     2.45          23.9         5.08   0.703
##  9 Hungary         5.12     1.37          22.4         4.52 NaN
## 10 Latvia          5.32     1.56          21.8         3.70 NaN
## # … with 155 more rows
```

```
#The country is Lesotho
```

As the table shown above, Lesotho has significant higher mean suicide rate than other countries. We think that this could be due to the population differences between Lesotho and the other countries, being that Lesotho is a rather small country.

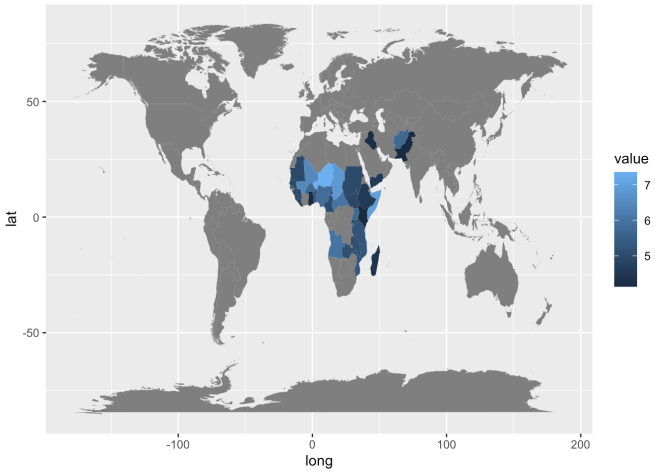
To illustrate the finding that we mentioned above further, we plot the geographical distribution of mean life ladder (top) and mean fertility rate (bottom).





These two geographic plots help show that countries that have a higher life ladder rate tend to have a lower fertility rate, which is represented by the opposite shade of color of the same country in these two plots.

We also notice that countries with a higher fertility rate tend to have a lower GDP per capita. The following plot shows the geographical distribution of the countries that have fertility rates >= 4.

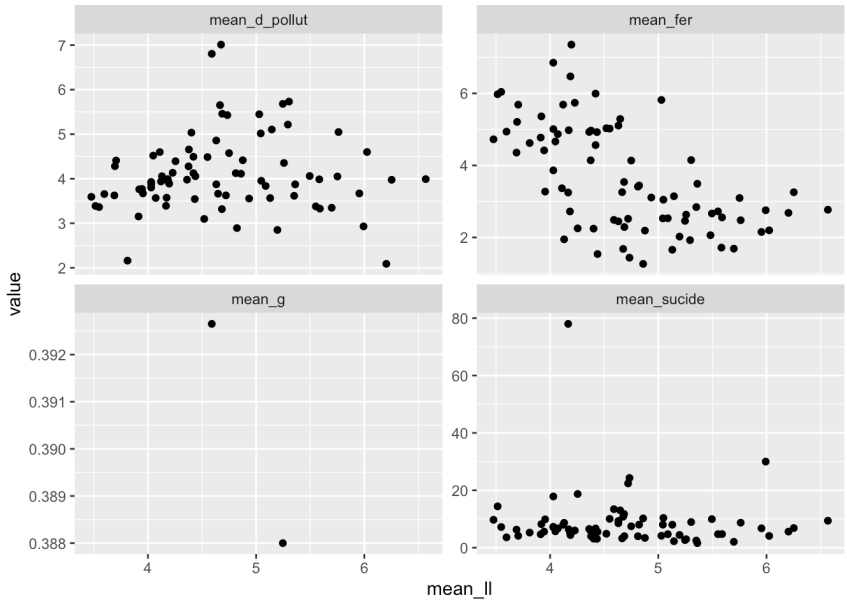


This geographic plot shows that countries that have fertility rates >= 4 are mostly in Africa, which means that most of these countries have lower GDP per capita. As a result, we want to control the GDP per capita of countries and see if the correlations that we mentioned above still exist.

```
max_gdp = max(df$log_gdp, na.rm = T)
min_gdp = min(df$log_gdp, na.rm = T)
mean_gdp = mean(df$log_gdp, na.rm = T)

df_lower_gdp = df %>%
  filter(log_gdp >= min_gdp & log_gdp <= mean_gdp) %>%
  group_by(country) %>%
  summarise(mean_ll = mean(ll),
            mean_fer = mean(ferate, na.rm = T),
            mean_sucide = mean(rate, na.rm = T),
            mean_d_pollut = mean(death, na.rm = T),
            mean_g = mean(g_coef, na.rm = T))
```

Within the code above, we divide the countries by the mean GDP per capita over all of the countries. In other words, countries that have GDP per capita greater than the minimum GDP per capita among all countries and less than the mean GDP per capita over all of the countries are categorized as countries with lower GDP per capita. Similarly, countries that have GDP per capita smaller than the maximum GDP per capita among all countries and greater than the mean GDP per capita over all of the countries are categorized as countries with higher GDP per capita. In hopes of performing this "split", we are able to control the GDP per capita of countries and see if the correlations that we've found still exist in the countries with higher GDP per capita and with lower GDP per capita respectively.



The above scatter plot shows the mean of life ladder and the mean of each factor (pollution, fertility rate, income inequality, and suicide rate) in countries with lower GDP per capita. As we could see from the plot, the negative correlation between mean life ladder and mean fertility rate still exist in countries with lower GDP per capita. However, since there are only two points presented in the plot of mean income inequality, we could ignore the correlation between the mean life ladder and mean income inequality in countries with lower GDP per capita, since there are not enough data points to construct a truly significant linear model. To support our claim, we fit linear models to each of the factors and the summary is shown below.

```
summary(life_lm_fl)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_fer, data = df_lower_gdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0640 -0.4665  0.0256  0.3296  1.6105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.75495    0.17489  32.906  < 2e-16 ***
## mean_fer    -0.28809    0.04441  -6.487 7.31e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5823 on 78 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.3505, Adjusted R-squared:  0.3421
## F-statistic: 42.09 on 1 and 78 DF,  p-value: 7.311e-09
```

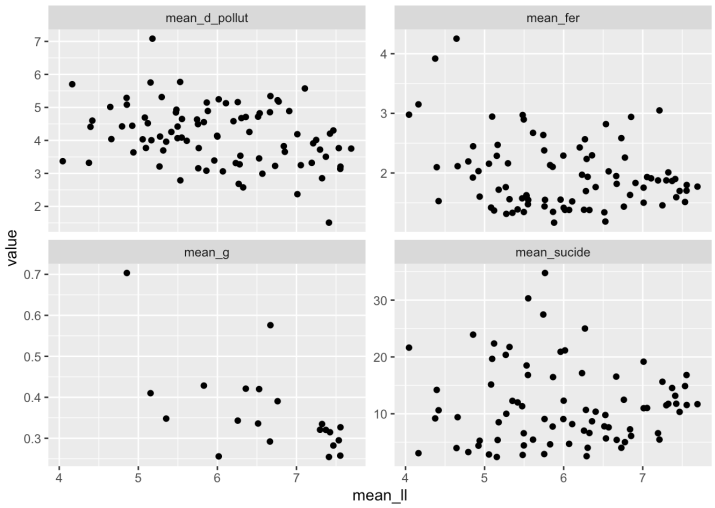
```
summary(life_lm_pl)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_d_pollut, data = df_lower_gdp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19175 -0.53167 -0.09655  0.47265  1.88084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.50507    0.39686  11.352   <2e-16 ***
## mean_d_pollut  0.04554    0.09464   0.481    0.632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7206 on 77 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.002998,   Adjusted R-squared:  -0.00995
## F-statistic: 0.2316 on 1 and 77 DF,  p-value: 0.6317
```

```
summary(life_lm_sl)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_sucide, data = df_lower_gdp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19260 -0.51418 -0.04474  0.43691  1.89651
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.719341   0.112131  42.088   <2e-16 ***
## mean_sucide -0.005121   0.008699  -0.589    0.558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7144 on 71 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.004857,   Adjusted R-squared:  -0.00916
## F-statistic: 0.3465 on 1 and 71 DF,  p-value: 0.558
```

Note that “s, g, f, p” represents suicide rate, income inequality, fertility rate and pollution respectively, and lm_life_xl represents the linear model between the mean of life ladder and the mean of the factors in countries with lower GDP per capita, where x is in {s, g, f, p}. As we could see from the summary of the linear model, the p value of lm_life_fl is approaching 0, which reflects that the negative correlations between the mean of life ladder and mean of fertility rate of countries with lower GDP per capita are statistically significant. Inversely, we notice that the p values of lm_life_pl and lm_life_sl are greater than 0.05, meaning that the correlation shown in these linear models are not statistically significant.



Similarly, the above scatter plot shows the mean of life ladder and mean of each factor (pollution, fertility rate, income inequality, and suicide rate) in countries with higher GDP per capita. As we could see from the plot, not only the negative correlation between mean life ladder and mean fertility rate still exist in countries with higher GDP per capita, but there are also negative correlations between the mean of life ladder and the mean of income inequality & pollution respectively in these countries. However, since there are a few points presented in the plot of mean income inequality we could not fully conclude that such a correlation is well supported by the data, since the correlation might change as we fill in the missing value to the data frame. This stands even if our previous linear model would show a negative correlation for it. To support our claim, we fit linear models to each of the factors and the summary is shown below.

```
summary(life_lm_fh)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_fer, data = df_higher_gdp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77186 -0.70903 -0.09216  0.75120  1.64926
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8216     0.3330  20.488   <2e-16 ***
## mean_fer     -0.4130     0.1624  -2.543   0.0128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.892 on 85 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.07072,    Adjusted R-squared:  0.05979
## F-statistic: 6.469 on 1 and 85 DF,  p-value: 0.01279
```

```
summary(life_lm_gh)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_g, data = df_higher_gdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3821 -0.3326  0.2266  0.5074  1.0083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3705     0.5363  15.609 2.73e-12 ***
## mean_g       -4.7018     1.4169  -3.318  0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6869 on 19 degrees of freedom
##   (70 observations deleted due to missingness)
## Multiple R-squared:  0.3669, Adjusted R-squared:  0.3336
## F-statistic: 11.01 on 1 and 19 DF,  p-value: 0.003613
```

```
summary(life_lm_ph)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_d_pollut, data = df_higher_gdp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20927 -0.64036 -0.02235  0.70715  1.55667
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.2947     0.4454   16.38   <2e-16 ***
## mean_d_pollut  -0.3090     0.1047   -2.95   0.0041 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8813 on 85 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.09288,    Adjusted R-squared:  0.08221
## F-statistic: 8.703 on 1 and 85 DF,  p-value: 0.004103
```

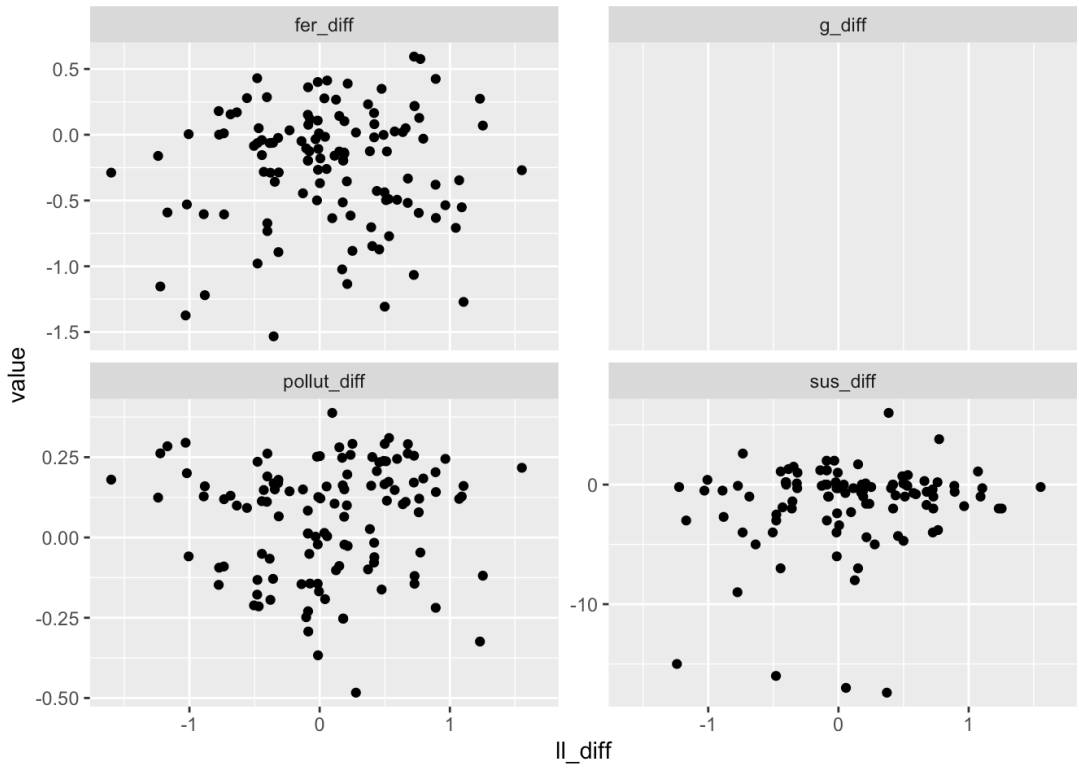
```
summary(life_lm_sh)
```

```
##
## Call:
## lm(formula = mean_ll ~ mean_sucide, data = df_higher_gdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9392 -0.6808 -0.0179  0.7052  1.6787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.048646   0.202112  29.927   <2e-16 ***
## mean_sucide -0.003017   0.015128  -0.199    0.842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9378 on 78 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.0005095,  Adjusted R-squared:  -0.0123
## F-statistic: 0.03976 on 1 and 78 DF,  p-value: 0.8425
```

Note that “s, g, f, p” represents suicide rate, income inequality, fertility rate and pollution respectively, and lm_life_xh represents the linear model between the mean of life ladder and the mean of the factors in countries with higher GDP per capita, where x is in {s, g, f, p}. As we could see from the summary of the linear model, the p values of lm_life_fh, lm_life_gh , and lm_life_ph approach to 0, which reflects that the negative correlations between the mean of life ladder and mean of income inequality & fertility rate & pollution in countries with higher GDP per capita are statistically significant respectively. Similar to the previous models, we notice that the p values of lm_life_sh are greater than 0.05, meaning that the correlation shown in these linear models are not statistically significant.

However, only analyzing the correlation between the mean of life ladder and the mean of each factor respectively could only represent the overall trend between countries globally. As a result, we want to also include time as a measurement and analyze the correlation between the differences of life ladder and the differences of each factor over a certain period of time within countries respectively, and see if the negative correlations that we’ve found in terms of the means still exist. Unfortunately, we could not find a mutual span of years among the countries in the data frame. For example, for some countries, the latest year on record is 2009, and for some countries, the earliest year on record is 2011. As a result, we decided to use the earliest year on record for each country as the starting year of the time span, and add 10 years to the starting year as the ending year of the time span. For example, if the earliest year on record of a country is 2006, then the time span of the country should be 2006-2016. We are able to control the period

of time we are analyzing for each country. Thus, all the following analyses are based on a time span of 10 years for each country.



The above scatter plot shows the differences of life ladder and differences of each factor (pollution, fertility rate, income inequality, and suicide rate) of countries in 10-year period. As we could see from the plot, we could not see a clear pattern in the plot of fertility differences (top left), of pollution differences (bottom left), and of suicide rate differences (bottom right). Moreover, there is no point showing up on the plot of income inequality differences, since no countries have a gini index over 10 years on record in the data frame. As a result, we fit linear models to each factor and the summary of these models are presented below.

```
summary(life_lm_time_f)
```

```
##
## Call:
## lm(formula = ll_diff ~ fer_diff, data = df_diff)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67331 -0.42522  0.00606  0.43613  1.47852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09691    0.06259   1.548    0.124
## fer_diff     0.08468    0.12407   0.683    0.496
##
## Residual standard error: 0.6009 on 117 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.003965,   Adjusted R-squared:  -0.004548
## F-statistic: 0.4658 on 1 and 117 DF,  p-value: 0.4963
```

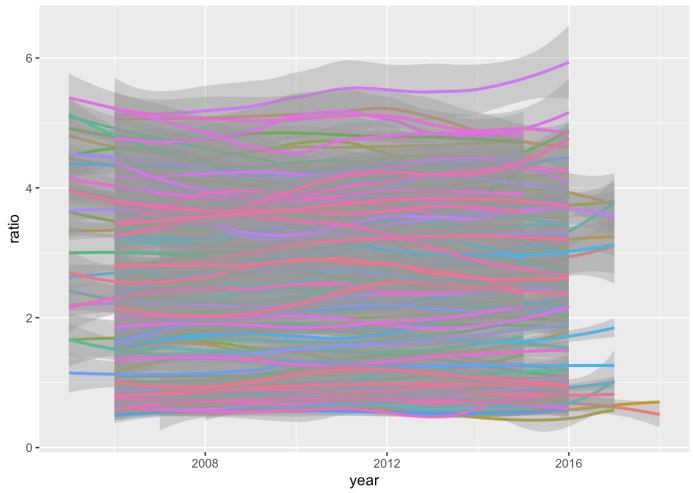
```
summary(life_lm_time_p)
```

```
##
## Call:
## lm(formula = ll_diff ~ pollut_diff, data = df_diff)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68228 -0.42185 -0.00265  0.41492  1.46718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06220    0.05793   1.074    0.285
## pollut_diff  0.10688    0.31261   0.342    0.733
##
## Residual standard error: 0.5976 on 116 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.001007,   Adjusted R-squared:  -0.007605
## F-statistic: 0.1169 on 1 and 116 DF,  p-value: 0.7331
```

```
summary(life_lm_time_s)
```

```
##
## Call:
## lm(formula = ll_diff ~ sus_diff, data = df_diff)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34414 -0.43444  0.01025  0.38917  1.43202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12551    0.05994   2.094   0.0387 *
## sus_diff     0.02473    0.01488   1.661   0.0996 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5674 on 105 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.02561,    Adjusted R-squared:  0.01633
## F-statistic:  2.76 on 1 and 105 DF,  p-value: 0.09964
```

Note that "s, g, f, p" represents suicide rate, income inequality, fertility rate and pollution respectively, and life_lm_time_x represents the linear model between the differences of life ladder and the differences of the factors of countries in 10-year period, where x is in {s, g, f, p}. As we could see from the summary of the linear model, the p values of life_lm_time_s, life_lm_time_f, life_lm_time_p are all greater than 0.05, showing that we failed to conclude that there's any statistically significant correlation between the differences of life ladder and the differences of these factors of countries in 10-year period.



To illustrate my point further, the graph above shows the ratio of life ladder / fertility rate over 10-year periods for each country. In other words, each color of the line in the graph represents a country and how the ratio changes in these 10 years. As a result, if there is a significant negative correlation between the differences of life ladder and the differences of fertility rate of countries in a 10-year period, we are expecting some drastic changes in these lines over time. In other words, these lines should either approach to 0 or increase to some large values, since the denominator and numerator should change towards different directions. However, as we could see from the graph, most of the lines only have some smooth fluctuations over time, which indicates that such negative correlation does not exist.

Conclusion

In this project we found that in general, the mean life ladder has a negative correlation between the mean of fertility rate and income inequality across countries respectively. Moreover, the negative correlation between the mean of the life ladder score and the mean of fertility rate still exist across countries with lower GDP per capita, and the negative correlation between the mean of life ladder and the mean of fertility rate and income inequality and pollution are significant across countries with higher GDP per capita. However, we fail to find any significant negative correlation between the differences of life ladder and the differences of pollution, income inequality, and suicide rate within countries in a 10-year period respectively. Our explanation to such a phenomenon is that the mean value of each factor (fertility rate, pollution, income inequality, suicide rate) could be an indicator of the overall development to the country. As a result, the negative correlations between the mean values of life ladder and each factor could reflect that people's happiness is correlated with the overall development of their countries positively. In other words, people tend to be happier in highly-developed countries than in lower-developing countries. Moreover, by controlling the GDP, we also find that to some extent, more factors will impact people's opinions about their own happiness as their countries have more GDP per capita. On the other hand, the differences of each factor in a 10-year period fail to reflect the change of the overall development of countries in these 10 years, since there are many other factors that could reflect the development of a county, such as GDP per capita, life expectancy, social support and so on. As a result, the change of one factor does not necessarily indicate the change of the overall development of the country. For example, if the life expectancy of a country increases, and every other factor decreases over 10 years, the overall development of the country will decrease over 10 years. For these reasons, we could not conclude if there is any significant correlation between the change of life ladder and the change of each factor over time. Consequently, based on the evidence that we found, we could conclude that people's happiness depends on the overall development of their own countries, and people's opinion about their own happiness could be more and more demanding as their countries continue to have an increase in GDP per capita.

In terms of the limitations of our project, the first one is that the correlation between the mean of the life ladder score and the mean of income inequality may not be reliable. Since there are only a few countries that have a gini index recorded in the data frame when we tried to join the table, even though the linear model suggests that there's a negative correlation between the two, we could not fully conclude that such a correlation is well supported by the data, since the correlation might change as we fill in the missing value to the data frame. Our second limitation of the project is that we cannot conclude the correlation of diff between life ladder and each factor in a specific period of time (ex. 2008-2018), since we don't have the data that share any mutual years among these countries.

We want to further investigate the correlation between the differences of life ladder and GDP per capita in a 10-year period within countries. If we could find a positive correlation between the differences of life ladder and GDP per capita in a 10-year period within countries, this finding could further support our thesis statement.

Link to Github: https://github.com/Ky1ey1/final_projectt/tree/main