# Course Project PSTAT 100: Brian Levin, Daniel Costello, Anabelle Gutman

# Course project guidelines

Your assignment for the course project is to formulate and answer a question of your choosing based on one of the following datasets:

1. ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present
2. World Happiness Report 2023: indices related to happiness and wellbeing by country 2008-present
3. Any dataset from the class assignments or mini projects

A good question is one that you want to answer. It should be a question with contextual meaning, not a purely technical matter. It should be clear enough to answer, but not so specific or narrow that your analysis is a single line of code. It should require you to do some nontrivial exploratory analysis, descriptive analysis, and possibly some statistical modeling. You aren't required to use any specific methods, but it should take a bit of work to answer the question. There may be multiple answers or approaches to contrast based on different ways of interpreting the question or different ways of analyzing the data. If your question is answerable in under 15 minutes, or your answer only takes a few sentences to explain, the question probably isn't nuanced enough.

## Deliverable

Prepare and submit a jupyter notebook that summarizes your work. Your notebook should contain the following sections/contents:

- **Data description**: write up a short summary of the dataset you chose to work with following the conventions introduced in previous assignments. Cover the sampling if applicable and data semantics, but focus on providing high-level context and not technical details; don't report preprocessing steps or describe tabular layouts, etc.
- **Question of interest**: motivate and formulate your question; explain what a satisfactory answer might look like.
- **Data analysis**: provide a walkthrough with commentary of the steps you took to investigate and answer the question. This section can and should include code cells and text cells, but you should try to focus on presenting the analysis clearly by organizing cells according to the high-level steps in your analysis so that it is easy to skim. For example, if you fit a regression model, include formulating the explanatory

variable matrix and response, fitting the model, extracting coefficients, and perhaps even visualization all in one cell; don't separate these into 5-6 substeps.

- **Summary of findings**: answer your question by interpreting the results of your analysis, referring back as appropriate. This can be a short paragraph or a bulleted list.

## Evaluation

Your work will be evaluated on the following criteria:

1. Thoughtfulness: does your question reflect some thoughtful consideration of the dataset and its nuances, or is it more superficial?
2. Thoroughness: is your analysis an end-to-end exploration, or are there a lot of loose ends or unexplained choices?
3. Mistakes or oversights: is your work free from obvious errors or omissions, or are there mistakes and things you've overlooked?
4. Clarity of write-up: is your report well-organized with commented codes and clear writing, or does it require substantial effort to follow?

In [1]:
```
import numpy as np
import pandas as pd
import altair as alt
import statsmodels.api as sm
# disable row limit for plotting
alt.data_transformers.disable_max_rows()
# uncomment to ensure graphics display with pdf export
# alt.renderers.enable('mimetype')
```

Out[1]:
```
DataTransformerRegistry.enable('default')
```

In [2]:
```
wh = pd.read_csv('data/whr-2023.csv')
continents = pd.read_csv('continents2.csv')
wh2 = wh.dropna()
wh.head()
```

Out[2]:

| | Country name | year | Life Ladder | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity | Perceptior corruptio |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2008 | 3.724 | 7.350 | 0.451 | 50.5 | 0.718 | 0.168 | 0.88 |
| **1** | Afghanistan | 2009 | 4.402 | 7.509 | 0.552 | 50.8 | 0.679 | 0.191 | 0.85 |
| **2** | Afghanistan | 2010 | 4.758 | 7.614 | 0.539 | 51.1 | 0.600 | 0.121 | 0.7( |
| **3** | Afghanistan | 2011 | 3.832 | 7.581 | 0.521 | 51.4 | 0.496 | 0.164 | 0.7: |
| **4** | Afghanistan | 2012 | 3.783 | 7.661 | 0.521 | 51.7 | 0.531 | 0.238 | 0.77 |

# Data Description

In this project, we explore the World Happiness Report datasets from 2005-2022. These reports and the data therein are a publication of the Sustainable Development Solutions Network using data collected via the Gallup World Poll data. The data were collected in support of the World Happiness Report's goal to meet rising worldwide demand to bring more attention to happiness and well-being as a criteria for government policy.

In all, the data comprise 2199 total measurements of countries and years across 165 unqiue countries. The key identifiers in this dataset are country and year and the dataset is made up of measurements of several variables in each country for each year from 2005-2022. The data tracks measurements related to physical health, mental health, generosity, economic productivity, governmental corruption, and freedom. In our analysis, we will focus primarily on the effects of each variable on log GDP per capita, a proxy for economic productivity, and life ladder, a proxy for happiness. Each variable in the dataset is defined below by a brief description or a measurement of national average responses to Gallup World Poll questions:

| Variable Name | Question of Interest / Variable Description | Type |
|---|---|---|
| Country name | Name of country | str |
| year | Year of measurement | int |
| Life Ladder | National average response to the Gallup World Poll (GWP) 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?” | float |
| Log GDP per capita | Log Gross Domestic Product per capita of country of interest in a given year at constant 2017 international dollar prices (Note: 2022 values are projections based on growth rates) | float |
| Social support | National average of the binary responses to the GWP question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?” | float |
| Healthy life expectancy at birth | World Health Organization life expectancies (Note: Only 2010, 2015, and 2019 are available in the WHO data. Interpolation and extrapolation are used to match the range of this dataset) | float |
| Freedom to make life choices | National average of responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” | float |
| Generosity | Residual of regressing national average of response to the GWP question “Have you donated money to a charity in the past month?” on GDP per capita | float |
| Perceptions of corruption | National average of the survey responses to two questions in the GWP: “Is corruption widespread throughoutthe government or not” and “Is corruption widespread within businesses or not?” The overall perception is | float |

| Variable Name | Question of Interest / Variable Description | Type |
|---|---|---|
| | the average of the two binary responses. The corruption perception at the national level is the average response of the overall perception at the individual level. | |
| Positive affect | Positive affect is defined as the average of three positive affect measures in GWP: laugh, enjoyment and doing interesting things in the Gallup World Poll. The average is of 3 binary responses to the following questions: “Did you smile or laugh a lot yesterday?”, and “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?”, “Did you learn or do something interesting yesterday?” | float |
| Negative affect | Average of three negative affect measures in GWP: worry, sadness and anger. The average is of 3 binary responses to the following questions: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?”, “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?”, and “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?” | float |

# Question of Interest

In this analysis, we sought to explore and better understand how the relationships between each of the variables listed above differ between GDP and life ladder (a proxy for happiness).

GDP and happiness are positively correlated. Despite this correlation, GDP and happiness have opposite relationships with many social welfare indicators. Moreover, Africa has the largest proportional disparity in happiness and log GDP per capita. How does the relationship between social welfare and GDP differ from that of social welfare and happiness? How these relationships differ between Africa and the rest of the world?

# Exploratory Analysis

## Missingness

In [3]:
```
#missing values
pd.DataFrame(wh.isna().mean(axis = 0)).reset_index().rename(columns = {'inde
                                                                       0:'Pro
```

Out[3]:

| | Variable Name | Proportion of Observations Missing |
|---|---|---|
| **0** | Country name | 0.000000 |
| **1** | year | 0.000000 |
| **2** | Life Ladder | 0.000000 |
| **3** | Log GDP per capita | 0.009095 |
| **4** | Social support | 0.005912 |
| **5** | Healthy life expectancy at birth | 0.024557 |
| **6** | Freedom to make life choices | 0.015007 |
| **7** | Generosity | 0.033197 |
| **8** | Perceptions of corruption | 0.052751 |
| **9** | Positive affect | 0.010914 |
| **10** | Negative affect | 0.007276 |

The table above shows the proportion of missing values for each column in the dataset. No column has more than 5.3% of values missing, as perceptions of corruption has the highest proportion of missing values at 5.28% and generosity second with 3.32% of values missing. Our analysis will focus on the effects of each variable on life ladder and log GDP per capita, both of which have less than 1% of values missing.

In [4]:
```
pd.DataFrame(wh.set_index(
    'Country name').isna().sum(
    axis = 1).groupby(
    'Country name').sum().sort_values(
    ascending = False).head(20)).reset_index().rename(
    columns = {0:'Total Missing Values'})
```

Out[4]:

| | Country name | Total Missing Values |
|---|---|---|
| **0** | China | 22 |
| **1** | Jordan | 21 |
| **2** | United Arab Emirates | 20 |
| **3** | Kosovo | 19 |
| **4** | Taiwan Province of China | 18 |
| **5** | Saudi Arabia | 17 |
| **6** | Kuwait | 15 |
| **7** | Bahrain | 13 |
| **8** | State of Palestine | 13 |
| **9** | Qatar | 13 |
| **10** | Vietnam | 12 |
| **11** | Hong Kong S.A.R. of China | 12 |
| **12** | Turkmenistan | 12 |
| **13** | Somaliland region | 12 |
| **14** | Egypt | 9 |
| **15** | Algeria | 9 |
| **16** | South Sudan | 8 |
| **17** | Malta | 7 |
| **18** | Venezuela | 7 |
| **19** | Cambodia | 4 |

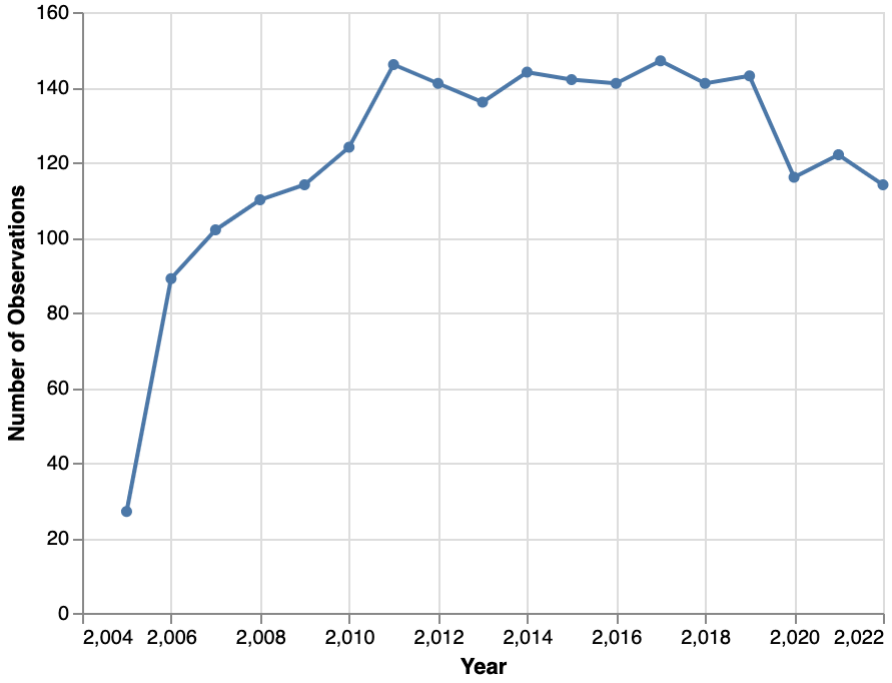
The table above shows the top 20 countries with the highest totals of missing values in the dataset. Perhaps most notably, nearly all countries on the list are located in Africa and Asia, which could have a negative impact on our certainty regarding analysis by country later on.

## Sample Size

In [5]:

```
nByYear = pd.DataFrame(wh.year.value_counts()).reset_index()
alt.Chart(nByYear).mark_line(point = True).encode(
    x = alt.X('year', title = 'Year'),
    y = alt.Y('count', title = 'Number of Observations')
)
```

Out[5]:




The graph above shows the number of observations in the dataset each year, where one observation corresponds to one set of measurements of each variable in an individual country in a given year. While the number of observations is largely constant from 2007 onwards, with over 100 observations each year, there were 89 observations in 2006 and just 27 observations in 2005. Again, this implies higher levels of uncertainty for any measurements or analysis in these years, as there are fewer observations upon which our analysis can be based, increasing the risk of introducing bias into our measurements.

## Correlation

In [6]:

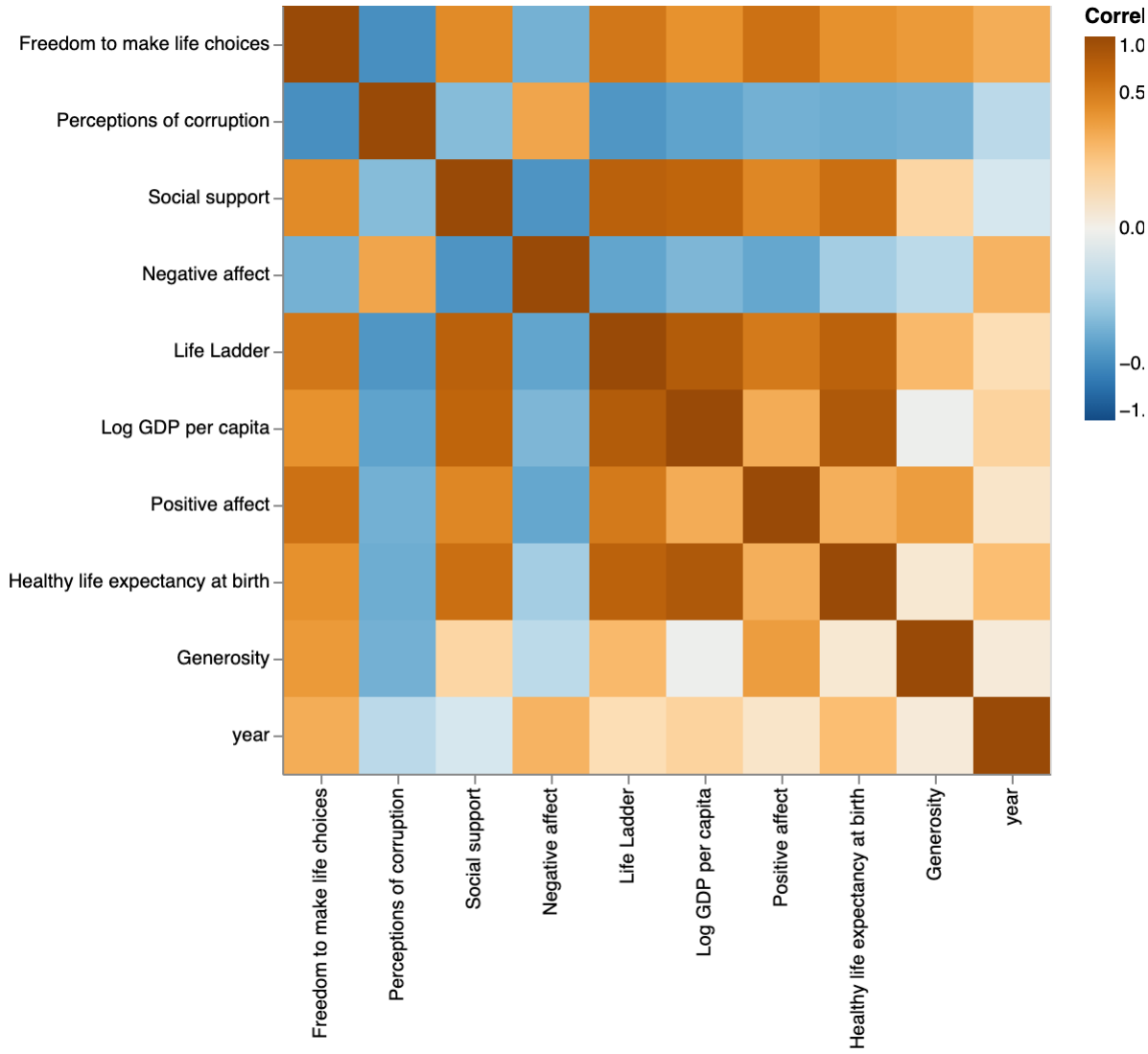
```
#print correlation matrix
corr_mx = wh.drop(columns = ['Country name']).corr()
corr_mx
```

Out[6]:

| | year | Life Ladder | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity |
|---|---|---|---|---|---|---|---|
| **year** | 1.000000 | 0.045947 | 0.077767 | -0.029741 | 0.163500 | 0.234135 | 0.005641 |
| **Life Ladder** | 0.045947 | 1.000000 | 0.784868 | 0.721662 | 0.713499 | 0.534493 | 0.181630 |
| **Log GDP per capita** | 0.077767 | 0.784868 | 1.000000 | 0.683590 | 0.818126 | 0.367525 | -0.000854 |
| **Social support** | -0.029741 | 0.721662 | 0.683590 | 1.000000 | 0.597659 | 0.409326 | 0.068572 |
| **Healthy life expectancy at birth** | 0.163500 | 0.713499 | 0.818126 | 0.597659 | 1.000000 | 0.373465 | 0.010775 |
| **Freedom to make life choices** | 0.234135 | 0.534493 | 0.367525 | 0.409326 | 0.373465 | 1.000000 | 0.325030 |
| **Generosity** | 0.005641 | 0.181630 | -0.000854 | 0.068572 | 0.010775 | 0.325030 | 1.000000 |
| **Perceptions of corruption** | -0.081394 | -0.431500 | -0.352847 | -0.222551 | -0.299016 | -0.476517 | -0.279435 |
| **Positive affect** | 0.019226 | 0.518169 | 0.237933 | 0.431038 | 0.223048 | 0.578680 | 0.307097 |
| **Negative affect** | 0.205329 | -0.339969 | -0.247541 | -0.441837 | -0.140700 | -0.275438 | -0.080801 |

In [7]:
```
#print correlation heat map
# melt to long form
corr_mx_long = corr_mx.reset_index().rename(
    columns = {'index': 'row'}
).melt(
    id_vars = 'row',
    var_name = 'col',
    value_name = 'Correlation'
)

# visualize
alt.Chart(corr_mx_long).mark_rect().encode(
    x = alt.X('col', title = '', sort = {'field': 'Correlation', 'order': 'a
    y = alt.Y('row', title = '', sort = {'field': 'Correlation', 'order': 'a
    color = alt.Color('Correlation',
                      scale = alt.Scale(scheme = 'blueorange',
                                        domain = (-1, 1),
                                        type = 'sqrt'),
                      legend = alt.Legend(tickCount = 5))
).properties(width = 400, height = 400)
```

Out[7]:




The above data frame and heatmap show show the correlation matrix numerically and graphically, respectively.
From these, we see that among all variables (omitting log GDP per capita), social support and healthy life expectancy at birth have high positive correlation with Life Ladder, while perceptions of corruption and negative effect have medium negative correaltion with Life Ladder.
While perceptions of corruption and negative effect also have negative correlation with log GDP per capita, the correlation coefficients between both and log GDP per capita are smaller than that of Life Ladder. Again, social support and healthy life expectancy at birth both have high positive correlation with log GDP per capita, though unlike Life Ladder, healthy life expectancy at birth has a higher correlation coefficient than social support with log GDP per capita.

## Central Tendencies

In [8]:
```
means_by_country = wh.groupby("Country name")[["Log GDP per capita", "Life L
print("Countries with 5 highest Log GDP per capita")
```

```
print(means_by_country.sort_values(by = 'Log GDP per capita',
                                   ascending = False).head(5))
print("")
print("Countries with 5 lowest Log GDP per capita")
print(means_by_country.sort_values(by = 'Log GDP per capita',
                                   ascending = True).head(5))
print("")
print("")
print("Countries with 5 highest average Life Ladder values (happiness)")
print(means_by_country.sort_values(by = 'Life Ladder',
                                   ascending = False).head(5))
print("")
print("Countries with 5 lowest average Life Ladder values (happiness)")
print(means_by_country.sort_values(by = 'Life Ladder',
                                   ascending = True).head(5))
```

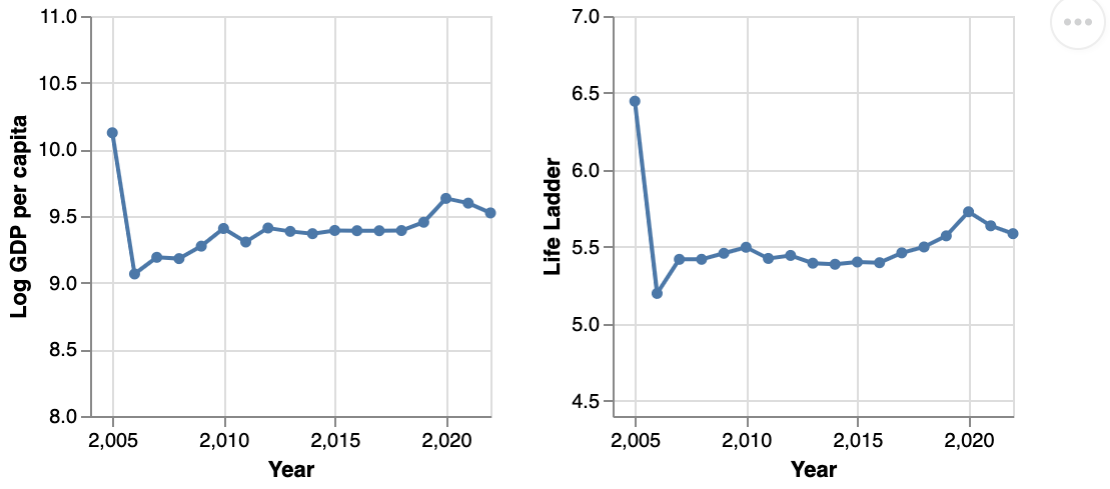
```
Countries with 5 highest Log GDP per capita
              Log GDP per capita  Life Ladder
Country name
Luxembourg              11.643250     7.062250
Qatar                   11.551800     6.569200
Singapore               11.346929     6.510143
Ireland                 11.139500     7.040375
Switzerland             11.134583     7.474583

Countries with 5 lowest Log GDP per capita
                         Log GDP per capita  Life Ladder
Country name
Burundi                            6.682200     3.548200
Congo (Kinshasa)                   6.870667     4.221556
Central African Republic           6.894800     3.515000
Somalia                            6.916000     5.183333
Niger                              6.985000     4.270667


Countries with 5 highest average Life Ladder values (happiness)
              Log GDP per capita  Life Ladder
Country name
Denmark                 10.890588     7.673529
Finland                 10.758267     7.619067
Norway                  11.063583     7.481750
Switzerland             11.134583     7.474583
Iceland                 10.882200     7.458600

Countries with 5 lowest average Life Ladder values (happiness)
                         Log GDP per capita  Life Ladder
Country name
Afghanistan                        7.585615     3.346643
South Sudan                             NaN     3.402000
Central African Republic           6.894800     3.515000
Burundi                            6.682200     3.548200
Rwanda                             7.427667     3.654417
```

The above tables show each country ranked by highest and lowest mean log GDP per capita and life ladder over the full domain of the dataset (all observations (years) for each country). Among the countries with the highest average log GDP per capita and life

ladder values, only Switzerland remains in the top 5 on each list. Meanwhile, among the countries with the lowest average log GDP per capita and life ladder values, both the Central African Republic and Burundi rank among the top 5. Additionally, we see the lists for highest average log GDP per capita and life ladder dominated by European countries, while the lists for the lowest average log GDP per capita and life ladder are comprised nearly entirely of African countries.

In [9]:
```
#means by year graph
means_by_year = wh.drop(columns = ['Country name']).groupby("year").mean()

plotMeansByYear1 = alt.Chart(means_by_year.reset_index()).mark_line(point =
    x = alt.X('year',
              title = 'Year'),
    y = alt.Y('Log GDP per capita', scale = alt.Scale(domain = [8, 11]))
).properties(
    width = 200,
    height = 200
)

plotMeansByYear2 = alt.Chart(means_by_year.reset_index()).mark_line(point =
    x = alt.X('year',
              title = 'Year'),
    y = alt.Y('Life Ladder', scale = alt.Scale(domain = [4.5, 7]))
).properties(
    width = 200,
    height = 200
)

plotMeansByYear1 | plotMeansByYear2
```

Out[9]:

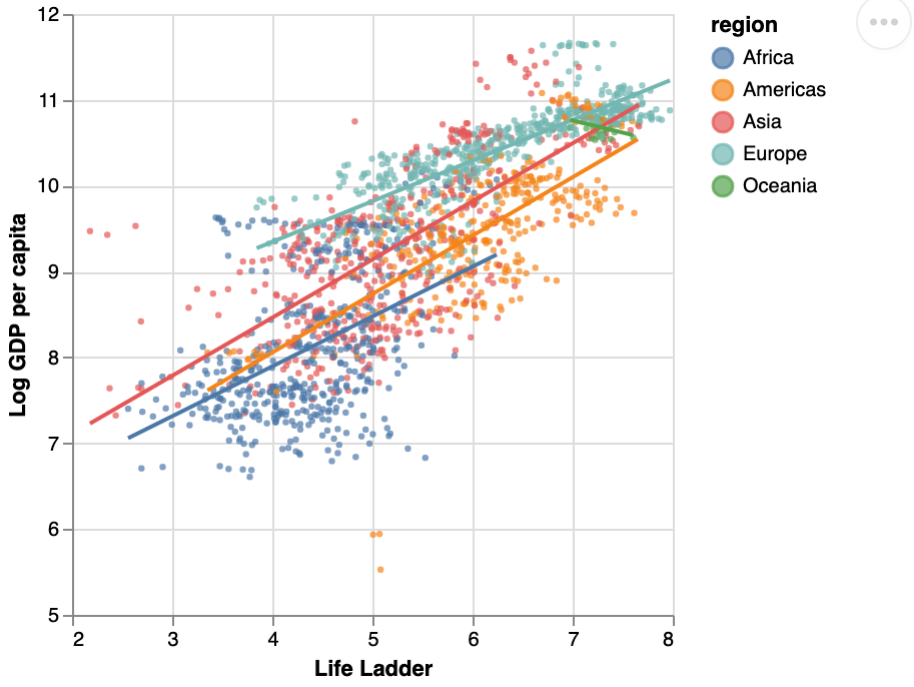



The above plots show the progression of average log GDP per capita and life ladder among all countries over time. We see a highly similar distribution in the two graphs, with a high in 2005 and largely constant values otherwise, save for a small increase in 2020. It should be noted, however, that the sample size in the dataset for 2005 is incredibly low (as detailed earlier) and, as a result, the high measurements in 2005 shown in the plots above may be skewed by bias introduced by this low sample size.

## Life Ladder vs Log GDP By Region

In [10]:
```
# continents analysis
continents = continents.rename(columns = {'name' : 'Country name'})

merged_data = pd.merge(wh2, continents, on=['Country name'])

columns_to_keep = ['Country name', 'year', 'Life Ladder', 'Log GDP per capit
                   'Healthy life expectancy at birth', 'Freedom to make life
                   'Perceptions of corruption', 'Positive affect', 'Negative

clean_data = merged_data[columns_to_keep]

# Log GDP per capita and Life Ladder scatter plot
scatter_gdp_ll = alt.Chart(clean_data).mark_circle(size = 10).encode(
    alt.X('Life Ladder', scale=alt.Scale(domain=[2,8])),
    alt.Y('Log GDP per capita', scale=alt.Scale(domain=[5,12])),
    color = 'region',
)
final_plot = scatter_gdp_ll + scatter_gdp_ll.transform_regression(
    'Life Ladder','Log GDP per capita', groupby=['region']).mark_line()
final_plot.properties(
    width = 300,
    height = 300
)
```

Out[10]:




The above plot shows the relationship between life ladder and log GDP per capita. In nearly every region (except Oceania), there is a decidedly positive relationship between the two variables.

Most importantly for our analysis, however, is that Africa tends to have lower values of

both life ladder and log GDP per capita than any other region. Furthermore, Africa tends to be more volatile in its measurements of both variables.

# Regression Analysis

# Happiness and GDP Relationship

In [11]:
```
x = wh2[['Log GDP per capita','Social support','Healthy life expectancy at b
                                        'Generosity','Perceptions of corrupti
x = sm.add_constant(x)
y = wh2['Life Ladder']

mlr = sm.OLS(endog = y, exog = x)
rslt = mlr.fit()

# retreive estimates and std errors
life_ladder_coef_tbl = pd.DataFrame({
    'estimate': rslt.params.values,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
    },
    index = x.columns
)
life_ladder_coef_tbl.loc['error variance', 'estimate'] = rslt.scale

# Defining Variables
x2 = wh2[['Life Ladder','Social support','Healthy life expectancy at birth',
                                        'Generosity','Perceptions of corrupti
x2 = sm.add_constant(x2)
y2 = wh2['Log GDP per capita']

mlr2 = sm.OLS(endog = y2, exog = x2)
rslt2 = mlr2.fit()

# retreive estimates and std errors
gdp_coef_tbl = pd.DataFrame({
    'estimate': rslt2.params.values,
    'standard error': np.sqrt(rslt2.cov_params().values.diagonal())
    },
    index = x2.columns
)
gdp_coef_tbl.loc['error variance', 'estimate'] = rslt2.scale

# display
life_ladder_gdp_coef_tbl = life_ladder_coef_tbl.merge(gdp_coef_tbl, left_inc
    columns = {'estimate_x':'Life Ladder Coef Est','standard error_x':'Stanc
               'standard error_y':'Standard Error'})
life_ladder_gdp_coef_tbl
```

Out[11]:

| | Life Ladder Coef Est | Standard Error | GDP Coef Est | Standard Error |
|---|---|---|---|---|
| **const** | -2.736199 | 0.175728 | 2.472879 | 0.170338 |
| **Social support** | 1.819961 | 0.161093 | 1.934710 | 0.153896 |
| **Healthy life expectancy at birth** | 0.027565 | 0.003180 | 0.075526 | 0.002608 |
| **Freedom to make life choices** | 0.378575 | 0.120621 | -0.204951 | 0.116273 |
| **Generosity** | 0.350698 | 0.083120 | -0.499648 | 0.079550 |
| **Perceptions of corruption** | -0.697080 | 0.080655 | -0.408430 | 0.078544 |
| **Positive affect** | 2.313041 | 0.150995 | -1.445686 | 0.150274 |
| **Negative affect** | -0.014655 | 0.168789 | -0.250143 | 0.162326 |
| **error variance** | 0.287786 | NaN | 0.266493 | NaN |

Consider the above regressions of Life Ladder and GDP against the above variables. Notably, the coefficients of Freedom to make life choices, Generosity, Positive affect have opposite sign, implying that these variables are correlated with more happiness and less GDP. This peculiarity motivated further analysis, specifically with regard to Africa.

Generosity and Positive affect have similar meanings and similar relationships with happiness and GDP. This similar behavior and interpretation motivated grouping variables together to create Mental Health, Physical Health, and Governmental Support indexes. These indexes facilitate visualization and interpretation of the above 7 variables.

In [12]:
```
merged_data = pd.merge(wh2, continents, on=['Country name'])

wh_summary = merged_data.drop(columns = ['alpha-2', 'alpha-3', 'country-code
                                         'sub-region', 'intermediate-region'
                                         'sub-region-code','intermediate-reg


Africa = wh_summary[wh_summary['region'] == 'Africa']
Not_Africa = wh_summary[wh_summary['region'] != 'Africa']

# Happiness Regression - Africa first

x4 = Africa[['Log GDP per capita','Social support','Healthy life expectancy
                                       'Generosity','Perceptions of corrupti
x4 = sm.add_constant(x4)
y4 = Africa['Life Ladder']

mlr = sm.OLS(endog = y4, exog = x4)
rslt = mlr.fit()

# retreive estimates and std errors
life_ladder_coef_africa = pd.DataFrame({
    'estimate': rslt.params.values,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
```

```
    },
    index = x4.columns
)
life_ladder_coef_africa.loc['error variance', 'estimate'] = rslt.scale

x5 = Not_Africa[['Log GDP per capita','Social support','Healthy life expecta
                                        'Generosity','Perceptions of corrupti
x5 = sm.add_constant(x5)
y5 = Not_Africa['Life Ladder']

mlr = sm.OLS(endog = y5, exog = x5)
rslt = mlr.fit()

# retreive estimates and std errors
life_ladder_coef_not_africa = pd.DataFrame({
    'estimate': rslt.params.values,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
    },
    index = x5.columns
)
life_ladder_coef_not_africa.loc['error variance', 'estimate'] = rslt.scale

# display
happiness_coef_table_africa_first = life_ladder_coef_africa.merge(
    life_ladder_coef_not_africa, left_index = True, right_index = True).drop
    columns = ['standard error_x','standard error_y']).rename(columns = {'es
                                                        'estimate_y':'

# GDP Regression - Africa first

x6 = Africa[['Life Ladder','Social support','Healthy life expectancy at birt
                                        'Generosity','Perceptions of corrupti
x6 = sm.add_constant(x6)
y6 = Africa['Log GDP per capita']

mlr = sm.OLS(endog = y6, exog = x6)
rslt = mlr.fit()

# retreive estimates and std errors
gdp_coef_africa = pd.DataFrame({
    'estimate': rslt.params.values,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
    },
    index = x6.columns
)
gdp_coef_africa.loc['error variance', 'estimate'] = rslt.scale

x7 = Not_Africa[['Life Ladder','Social support','Healthy life expectancy at
                                        'Generosity','Perceptions of corrupti
x7 = sm.add_constant(x7)
y7 = Not_Africa['Log GDP per capita']

mlr = sm.OLS(endog = y7, exog = x7)
rslt = mlr.fit()

# retreive estimates and std errors
```

```
gdp_coef_not_africa = pd.DataFrame({
    'estimate': rslt.params.values,
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
    },
    index = x7.columns
)
gdp_coef_africa.loc['error variance', 'estimate'] = rslt.scale

# display
gdp_coef_table_africa_first = gdp_coef_africa.merge(
    gdp_coef_not_africa, left_index = True, right_index = True).drop(
    columns = ['standard error_x','standard error_y']).rename(columns = {'es
                                                                'estimate_y':'
gdp_coef_table_africa_first

happiness_coef_table_africa_first.merge(gdp_coef_table_africa_first, left_ir
```

Out[12]:

| | African Happiness Coef Est | Not African Happiness Coef Est | African GDP Coef Est | Not African GDP Coef Est |
|---|---|---|---|---|
| **const** | -1.114083 | -1.719584 | -0.095927 | 3.153304 |
| **Social support** | 1.011854 | 2.166394 | 1.737504 | 1.801010 |
| **Healthy life expectancy at birth** | 0.017067 | 0.014618 | 0.071029 | 0.078234 |
| **Freedom to make life choices** | 0.536276 | 0.117508 | 0.970453 | -0.735895 |
| **Generosity** | 0.323473 | 0.122309 | -2.618468 | -0.207237 |
| **Perceptions of corruption** | 0.380820 | -0.920271 | 1.416082 | -0.729552 |
| **Positive affect** | 1.399803 | 2.717731 | 0.468547 | -1.444068 |
| **Negative affect** | 0.675630 | -0.570658 | -0.011065 | -0.205156 |

Consider the above coefficients of four regressions of happiness and GDP within Africa and excluding Africa. Most notably, the aforementioned peculiarity was reversed - Freedom to make life choices, Generosity, and Perceptions of corruption no longer have opposite relationships with happiness and GDP. It is quite surprising that these mental health and governmental support indicators are positively correlated with GDP in Africa since they are negatively correlated in the rest of the world.

The above regressions show that African countries' production disproportionately benefit from mental health and governmental support improvements.

# Indexed Variable Analysis

In this analysis, we combined all variables (aside from identifying variables (country) and response variables (log GDP per capita and life ladder)) into three index variables

covering aspects daily life. Our motivation for this methodology is explained in the regression analysis section prior. In short, it faciliated visualization and improved interpretive value for the variables in the dataset. The variables created are defined below:

- Mental Health: Comprised of the average of social support, generosity, and net positive affect, defined as the difference between positive and negative affect. A higher mental health index value would indicate more supportive, generous, and generally mentally positive citizens within a country.
- Physical Health: Comprised solely of a rescaled healthy life expectancy at birth value. We use life expectancy as a proxy for overall physical health and rescale the variable such that it exists in $[0, 1]$ to match the other index variables. A higher physical health index value is directly proportional to higher life expectancy within a country.
- Governmental Support: Comprised of the average of freedom to make life choices and 1 minus perceptions of corruption. We subtract perceptions of corruption from 1 to invert the original quantity, which tracks whether or not respondees believe there to be governmental and economic corruption. The new quantity thus tracks whether or not respondees DON'T believe there to be such corruption. Therefore, a higher governmental support index value would indicate more confidence in a country's government and higher levels of perceived support from the government to the populous.

In [13]:
```
#create index variables for broad analysis
wh_indexed = clean_data.copy()
wh_indexed

#create mental health index, with social support, generosity, and net positi
wh_indexed["Mental Health Index"] = (wh_indexed["Social support"] +
                                     wh_indexed["Generosity"] +
                                     (wh_indexed["Positive affect"] - wh_ind

#rescale healthy life expentancy to a 0 to 1 scale to match other variables
wh_indexed["Physical Health Index"] = (wh_indexed["Healthy life expectancy a
                                       np.min(wh_indexed['Healthy life expec
    np.max(wh_indexed['Healthy life expectancy at birth']) -
    np.min(wh_indexed['Healthy life expectancy at birth']))

#perceptions of corruption and freedom to make life choices can be used for
wh_indexed['Governmental Support Index'] = ((1 - wh_indexed["Perceptions of
                                             wh_indexed['Freedom to make life

#filter only for response variables, identification variables, and newly cre
wh_indexed = wh_indexed.filter(['Country name', 'year', 'region', 'Life Ladd
                                'Mental Health Index', 'Physical Health Inde
wh_indexed.head()
```

Out[13]:

| | Country name | year | region | Life Ladder | Log GDP per capita | Mental Health Index | Physical Health Index | Governmental Support Index |
|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2008 | Asia | 3.724 | 7.350 | 0.258333 | 0.646152 | 0.4180 |
| **1** | Afghanistan | 2009 | Asia | 4.402 | 7.509 | 0.329000 | 0.650579 | 0.4145 |
| **2** | Afghanistan | 2010 | Asia | 4.758 | 7.614 | 0.300667 | 0.655007 | 0.4465 |
| **3** | Afghanistan | 2011 | Asia | 3.832 | 7.581 | 0.299333 | 0.659435 | 0.3825 |
| **4** | Afghanistan | 2012 | Asia | 3.783 | 7.661 | 0.368333 | 0.663862 | 0.3775 |

In [14]:

```
x_index = sm.tools.add_constant(wh_indexed[['year', 'Mental Health Index', '
                                            'Governmental Support Index']])
x_index
gdp = wh_indexed['Log GDP per capita']
ladder = wh_indexed['Life Ladder']

lr_gdp = sm.OLS(endog = gdp, exog = x_index)
rslt_gdp = lr_gdp.fit()
lr_ladder = sm.OLS(endog = ladder, exog = x_index)
rslt_ladder = lr_ladder.fit()

coef_tbl = pd.DataFrame({
    'estimate, log GDP response':rslt_gdp.params.values,
    'estimate, life ladder response':rslt_ladder.params.values,
    'standard error, log GDP response':np.sqrt(rslt_gdp.cov_params().values.
    'standard error, life ladder response':np.sqrt(rslt_ladder.cov_params().
    index = x_index.columns,
)
coef_tbl["contains zero (95%), log GDP response"] = ((coef_tbl['estimate, lo
                                                      2*coef_tbl['standard e
                                                     (coef_tbl['estimate, lo
                                                      2*coef_tbl['standard e

coef_tbl["contains zero (95%), life ladder response"] = ((coef_tbl['estimate
                                                          2*coef_tbl['standard e
                                                         (coef_tbl['estimate, li
                                                          2*coef_tbl['standard e


coef_tbl
```

Out[14]:

| | estimate, log GDP response | estimate, life ladder response | standard error, log GDP response | standard error, life ladder response | contains zero (95%), log GDP response | contains zero (95%), life ladder response |
|---|---|---|---|---|---|---|
| **const** | 33.510038 | 26.702022 | 6.635247 | 6.603223 | False | False |
| **year** | -0.015916 | -0.014124 | 0.003304 | 0.003288 | False | False |
| **Mental Health Index** | 1.054347 | 3.629738 | 0.170521 | 0.169698 | False | False |
| **Physical Health Index** | 8.658070 | 6.093910 | 0.156727 | 0.155971 | False | False |
| **Governmental Support Index** | 0.519750 | 1.374129 | 0.132318 | 0.131679 | False | False |

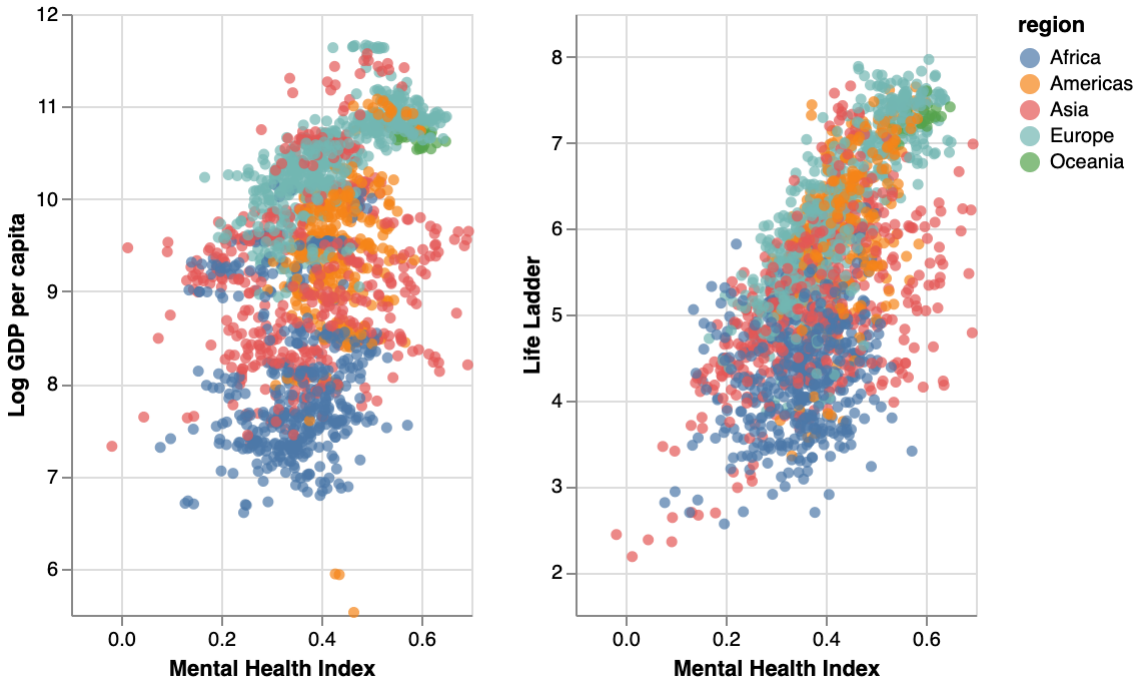
The above table shows the results of two regression models: One with a log GDP response and the other with a life ladder response. Both have the same set of predictors: mental health index, physical health index, governmental support index, and year (to control for temporal effects).
The regression results showed significance at the 95% level for all coefficients. However, there was little difference in coefficients between each model in year. This lack of difference in temporal effects was intuitive, as early we showed that log GDP per capita and life ladder vary highly similarly with time.
The models also showed positive effects of each index on both log GDP per capita and life ladder. Again, this is an intuitive result. We expected to see that improvements in physical health, mental health, and governmental support would, on average, yield improvements in economic productivity and overall happiness.
A shortcoming in our analysis is the lack of comparability between coefficients due to differences in units in log GDP per capita and life ladder. The sign of the coefficients, however, is comparable across models and confirmed our beliefs.
Within each model, however, each index is on the same scale, we are able to draw more apt comparisons:

- Life Ladder Response Findings
  - The mental health index coefficient is far higher than that of governmental support, implying that improvements in mental health tend to yield larger improvements in happiness (life ladder) than improvements in governmental support.
- Log GDP Per Capita Findings
  - A similar trend to the life ladder response model is found with log GDP per capita. However, the physical health index coefficient is far higher relative to all other coefficients than in the life ladder response model. From this, we inferred that improvements in life expectancy yield large improvements in log GDP per capita, relative to those improvements brought about by governmental support or mental health.

## Visualizations

In [15]:
```
mental_gdp = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Mental Health Index', scale = alt.Scale(zero = False)),
    y = alt.X('Log GDP per capita', scale = alt.Scale(domain = [5.5,12])),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
mental_gdp

mental_ll = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Mental Health Index', scale = alt.Scale(zero = False)),
    y = alt.Y('Life Ladder', scale = alt.Scale(domain = [1.5,8.5])),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
mental_ll

mental_gdp | mental_ll
```

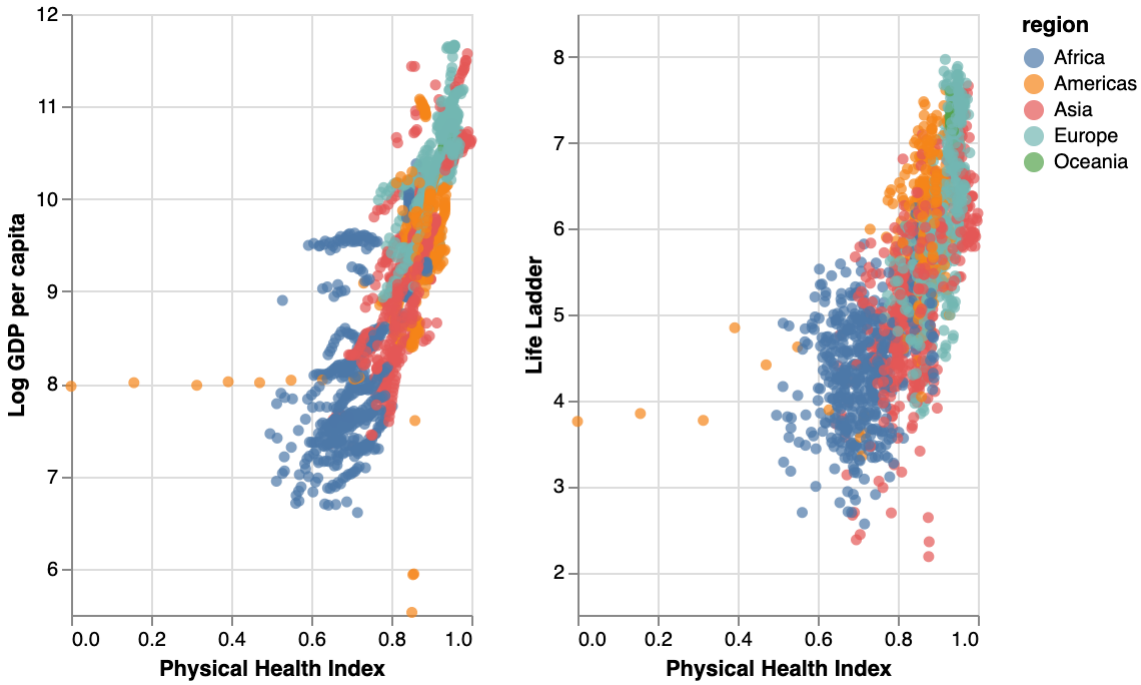
Out[15]:




When comparing how the mental health index impacts log GDP per capita and Life Ladder you can recognize interesting trends. As Life Ladder increases there tends to be a positive increase in mental health score as well. Although, when looking at how log GDP per capita impacts mental health it's a lot more scattered, with inconsistent trends in most continents, including Africa. This shows that as log GDP per capita increases, there is little change in the citizens mental health.

In [16]:
```
physical_gdp = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Physical Health Index', scale = alt.Scale(zero = False)),
    y = alt.Y('Log GDP per capita', scale = alt.Scale(domain = [5.5,12])),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
physical_gdp

physical_ll = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Physical Health Index', scale = alt.Scale(zero = False)),
    y = alt.Y('Life Ladder', scale = alt.Scale(domain = [1.5,8.5])),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
physical_ll

physical_gdp | physical_ll
```

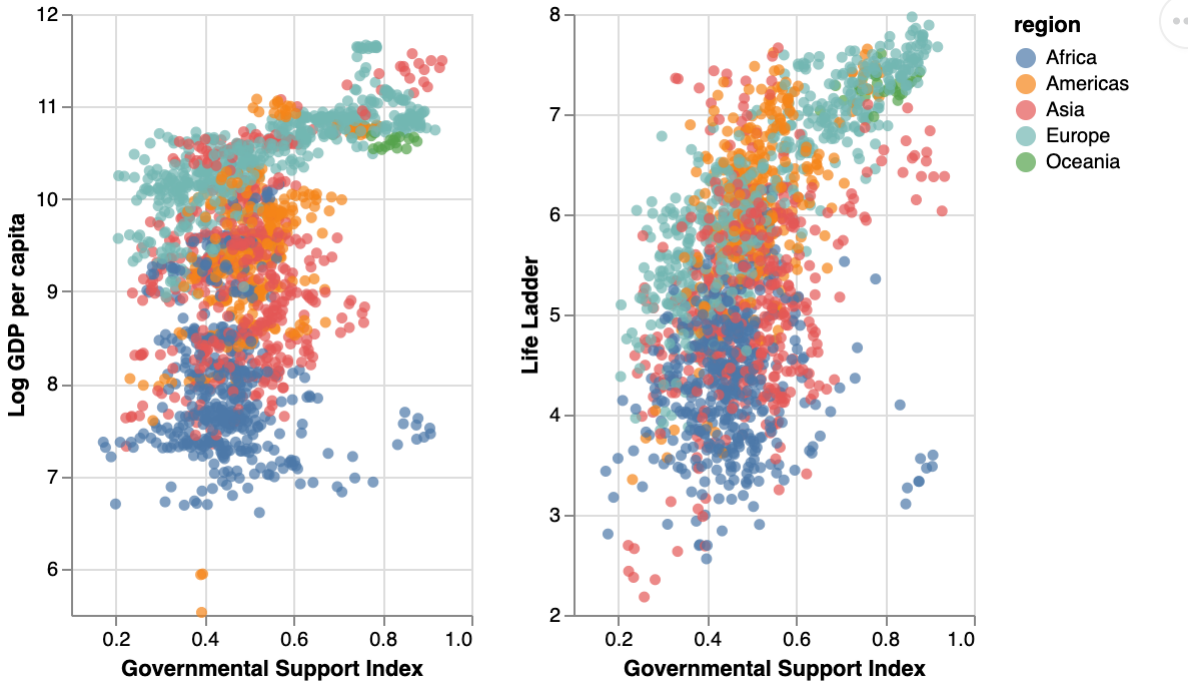
Out[16]:




When looking at how how the physical health index impacts log GDP per capita and Life Ladder the trends appear to be more straightforward amongst continents, with Africa being the biggest exception. As log GDP per capita increases there is a pretty positive increase in physical health universally. Although, when looking at how Life Ladder impacts physical health there's still a positive slope, but the data is more scattered, with the most inconsistent trend in Africa. This is a very interesting result. I initially expected physical health and Life Ladder to be more strictly positively correlated, but it's eye-opening to have my initial assumption refuted.

In [17]:
```
governmental_gdp = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Governmental Support Index', scale = alt.Scale(zero = False)),
    y = alt.Y('Log GDP per capita', scale = alt.Scale(domain = [5.5,12])),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
governmental_gdp

governmental_ll = alt.Chart(wh_indexed).mark_circle(size = 30).encode(
    x = alt.X('Governmental Support Index', scale = alt.Scale(zero = False)),
    y = alt.Y('Life Ladder', scale = alt.Scale(zero = False)),
    color = 'region'
).properties(
    width = 200,
    height = 300
)
governmental_ll

governmental_gdp | governmental_ll
```

Out[17]:




When comparing how the Governmental Support Index impacts log GDP per capita and Life Ladder you can see more interesting trends, especially within Africa. As Life Ladder increases there tends to be an overall positive increase in mental health score amongst all continents, with Africa having the smallest slope. Although, when looking at how log GDP per capita impacts Governmental Support Index it's very scattered and uncorrelated for the most part, with inconsistent trends in almost all continents. The most inconsistent data though is within Africa because you can see that as the log GDP per capita increases, the Governmental Support Index actually decreases contrary to the circumstances for every other continent.

# Summary

In all, our analysis demonstrated log GDP per capita is greatly impacted by the physical health of citizens and is positively impacted by both mental health and governmental support. Simiarly, life ladder, a proxy for happiness, is positively impacted by the same three factors, though seemingly more heavily impacted by mental health relative to log GDP per capita.
Furthermore, exploratory analysis showed that African countries typically maintained lower levels of log GDP per capita and happiness (life ladder). The regression models developed previously show that African countries' production disproportionately benefits from mental health and governmental support improvements.