PSTAT 100 Final Project: What affects happiness around the world?

By Ethan Choi and Alex Zhao

Data Description

The data we chose to work with was the World Happiness Report 2023. This report contains the happiness level (also called life ladder score) of 165 countries around the world from years between 2005 and 2022, as well as a variety of socioeconomic factors for each country and year. [INSERT HOW DATA WAS SAMPLED HERE]. Each of the socioeconomic variables are described in the table below:

Variable name	Description
`Life Ladder`	National average of answers to question: "How would you rate your current life on a scale from 0-10?" With 0 being the worst and 10 being the best.
`Log GDP per capita`	Logarithm of GDP per capita in terms of Purchasing Power Parity adjusted to constant 2017 international dollars.
`Social Support`	National average of answers to 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?" With 0=No and 1=Yes.
`Healthy life expectancy at birth`	Healthy life expectancy based on time series analysis of World Health Organization data (in years).
`Freedom to make life choices`	National average of answers to question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life??" With 0=Dissatisfied and 1=Satisfied.
`Generosity`	Residual of regressing the national average of answers to the question "Have you donated money to a charity in the past month?" on log GDP per capita.
`Perceptions of corruption`	National average of answers to the two questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" With 0=No and 1=Yes.
`Positive affect`	National average of previous-day affect measures for laughter, enjoyment, and interest. Questions are asked in the form of "Did you experience the following feelings during a lot of the day yesterday?"
`Negative affect`	National average of previous-day affect measures for worry, sadness, and anger. Questions are asked in the form of "Did you experience the following feelings during a lot of the day yesterday?"

Question of Interest

Overall, our question of interest is "Has happiness around the world improved over time and which variables affect happiness the most?". This question interested us because we are hopeful that peoples' happiness has improved over time and wanted to provide a definitive answer. We also believe it is important to identify variables that can explain trends in different countries' happiness levels. In that way, we can discover why certain countries may be happier than others.

In order to tackle this problem we first derived some overall trends in the data such as, has there been an overall improvement in happiness around the world, which country has seen the greatest improvement in happiness, and which country has been the happiest over time. We then attempted to explain these trends using principal component analysis and multiple linear regression.

A satisfactory answer to our question would be showing and increase or decrease in happiness over time as well as a list of variables that explain the most variation (derived from principal component analysis), and [INSERT POSSIBLE RESULTS OF REGRESSION HERE]

Data Analysis

Exploratory data analysis

In [1]:
```
import numpy as np
import pandas as pd
import altair as alt
from scipy import linalg
from statsmodels.multivariate.pca import PCA
# 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]:
```
# import tidy world happiness data
happiness = pd.read_csv('data/whr-2023.csv')
happiness.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 corruptic
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	0.70
3	Afghanistan	2011	3.832	7.581	0.521	51.4	0.496	0.164	0.73
4	Afghanistan	2012	3.783	7.661	0.521	51.7	0.531	0.238	0.77

In [3]:
```
# how many observations are in the data set
happiness.shape
```

Out[3]:
```
(2199, 11)
```

In [4]:
```
# how many countries are in the data set
happiness['Country name'].nunique()
```

Out[4]:
```
165
```

In [5]:
```
# which years do the observations come from
np.sort(happiness['year'].unique())
```

Out[5]:
```
array([2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015,
       2016, 2017, 2018, 2019, 2020, 2021, 2022])
```

In [6]:
```
# how often are variable values missing
happiness.isna().mean()
```

Out[6]:
```
Country name                        0.000000
year                                0.000000
Life Ladder                         0.000000
Log GDP per capita                  0.009095
Social support                      0.005912
Healthy life expectancy at birth    0.024557
Freedom to make life choices        0.015007
Generosity                          0.033197
Perceptions of corruption           0.052751
Positive affect                     0.010914
Negative affect                     0.007276
dtype: float64
```

Our dataset is comprised of 2199 observations, all of which were collected between 2005-2022, and 11 variables. It is important to note that this data was collected from 165 different countries, but not all 165 of those countries participated in the report each year from 2005-2022. For example, we do not have an observation from Afghanistan in 2005. In terms of missing variable values though, this data set is constructed quite well as none of the variables are missing over 6% of the time.

In [7]:
```
# summary statistics
happiness_summary = happiness.drop(columns = ['Country name', 'year']).aggre

# print the dataframe
happiness_summary
```

Out[7]:

	mean	std
Life Ladder	5.479227	1.125527
Log GDP per capita	9.389760	1.153402
Social support	0.810681	0.120953
Healthy life expectancy at birth	63.294582	6.901104
Freedom to make life choices	0.747847	0.140137
Generosity	0.000091	0.161079
Perceptions of corruption	0.745208	0.185835
Positive affect	0.652148	0.105913
Negative affect	0.271493	0.086872

The above summary statistics provide us with a good baseline for comparing countries' happiness levels and the factors that affect those levels. For example, if a country has a life ladder score greater than 5.48 we know that that country is happier than the average country.

Has happiness around the world improved over time?

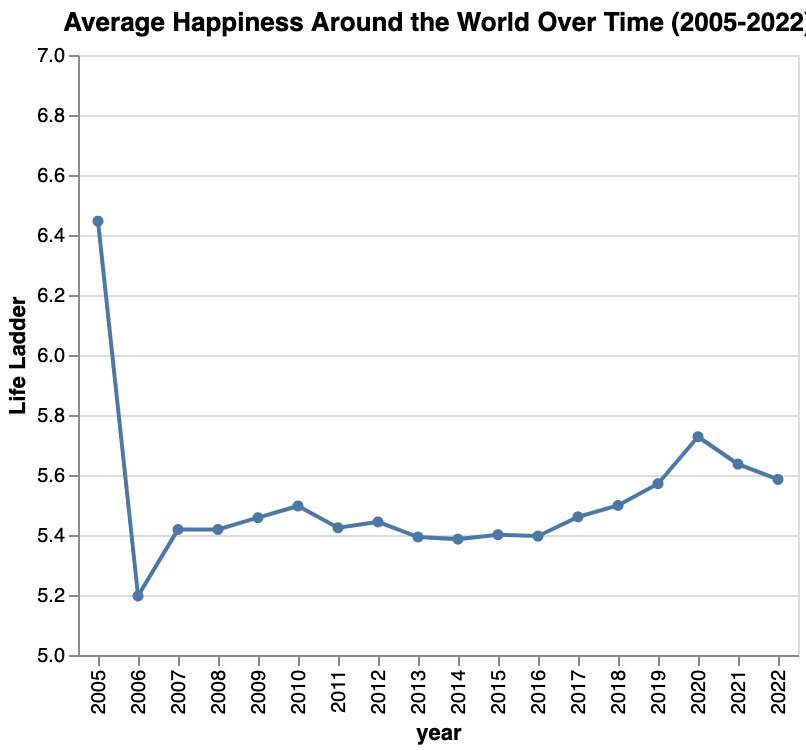
In [8]:
```
# Average life ladder score across all countries each year from 2005-2022
happiness_byyear = happiness.iloc[:,[1,2]].groupby('year').mean().reset_inde
happiness_byyear
```

Out[8]:

	year	Life Ladder
0	2005	6.446259
1	2006	5.196899
2	2007	5.418275
3	2008	5.418509
4	2009	5.457667
5	2010	5.496806
6	2011	5.424082
7	2012	5.443617
8	2013	5.393294
9	2014	5.386264
10	2015	5.400944
11	2016	5.396447
12	2017	5.460408
13	2018	5.498674
14	2019	5.570965
15	2020	5.727517
16	2021	5.636246
17	2022	5.585140

In [9]:
```
# Plot of above dataframe
alt.Chart(happiness_byyear).mark_line(point=True).encode(
    x = 'year:N',
    y = alt.Y('Life Ladder', scale=alt.Scale(domain=[5,7]))).properties(
    title = 'Average Happiness Around the World Over Time (2005-2022)')
```

Out[9]:



The ladder score in 2005 (6.45) appears to be an outlier, therefore we checked the sample size of countries used to obtain this score.

In [10]:
```
# Number of countries surveryed in 2005
len(happiness[happiness['year']==2005])
```

Out[10]: 27

Only 27 countries out of the 165 total countries that appeared in this report had happiness data from 2005. This led us to look at the number of countries surveyed each year.

In [11]:
```
# Checking amount of countries surveyed each year
results = []

for year in range(2005, 2023):
    result_tuple = (year, len(happiness[happiness['year'] == year]))
    results.append(result_tuple)

df = pd.DataFrame(results, columns=['Year', 'Number of Countries Surveyed']

# Display the resulting DataFrame.
df['Proportion of Total Countries'] = df['Number of Countries Surveyed']/165
df
```

Out[11]:

	Year	Number of Countries Surveyed	Proportion of Total Countries
0	2005	27	0.163636
1	2006	89	0.539394
2	2007	102	0.618182
3	2008	110	0.666667
4	2009	114	0.690909
5	2010	124	0.751515
6	2011	146	0.884848
7	2012	141	0.854545
8	2013	136	0.824242
9	2014	144	0.872727
10	2015	142	0.860606
11	2016	141	0.854545
12	2017	147	0.890909
13	2018	141	0.854545
14	2019	143	0.866667
15	2020	116	0.703030
16	2021	122	0.739394
17	2022	114	0.690909

2005 is the only year in which less than 50% of the 165 total countries in the report, were surveyed. Since only about 16% of these countries were surveyed in 2005, we decided to remove observations from 2005 when deciding whether or not happiness improved around the world over time. This is because we thought that data from only 27 countries was not sufficient enough to represent the entire world. So, we replotted the happiness data grouped by year, removing observations from 2005, and got the following results:

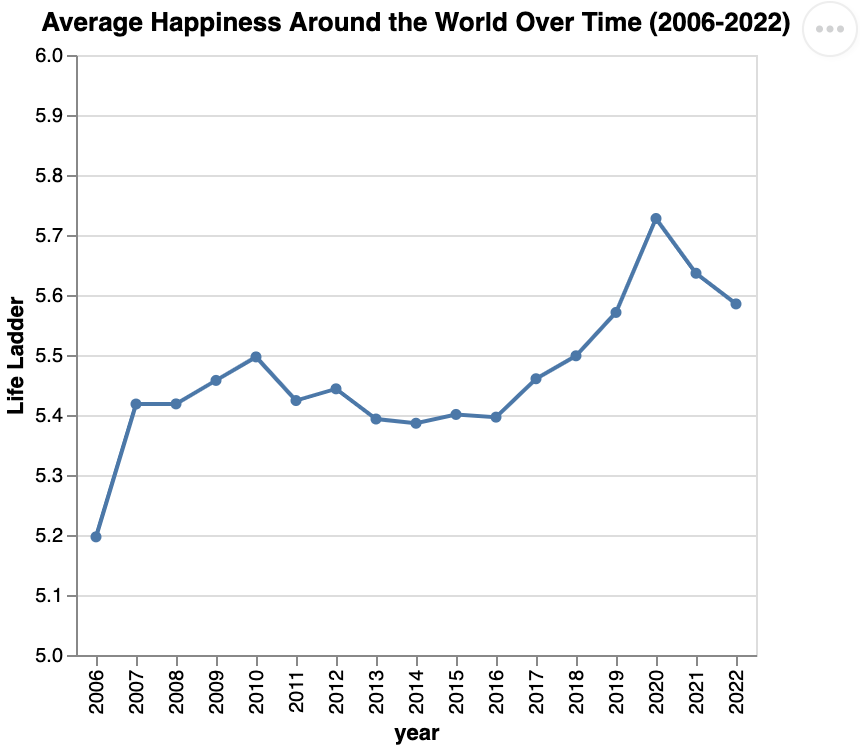
In [12]:
```
# Average life ladder score across all countries each year form 2006-2022
happiness_byyear_no2005 = happiness_byyear[happiness_byyear['year']!=2005]
happiness_byyear_no2005
```

Out[12]:

	year	Life Ladder
1	2006	5.196899
2	2007	5.418275
3	2008	5.418509
4	2009	5.457667
5	2010	5.496806
6	2011	5.424082
7	2012	5.443617
8	2013	5.393294
9	2014	5.386264
10	2015	5.400944
11	2016	5.396447
12	2017	5.460408
13	2018	5.498674
14	2019	5.570965
15	2020	5.727517
16	2021	5.636246
17	2022	5.585140

In [13]:
```
alt.Chart(happiness_byyear_no2005).mark_line(point=True).encode(
    x = 'year:N',
    y = alt.Y('Life Ladder', scale=alt.Scale(domain=[5,6]))).properties(
    title = 'Average Happiness Around the World Over Time (2006-2022)')
```

Out[13]:



From this graph, we can see that since 2006, there has been a slight increase in the overall happiness in the world with the life ladder score increasing from approximately 5.2 in 2006 to approximately 5.59 in 2022.

Which country's happiness has improved the most over time?

In [14]:
```
afghanistan = happiness[happiness['Country name'] == 'Afghanistan']
afghanistan.loc[len(afghanistan)-1,'Life Ladder']-afghanistan.loc[0,'Life La
```

Out[14]:
```
-2.4430000000000005
```

In [15]:
```
countries = happiness['Country name'].unique()
results = []

for country in countries:
    country_data = happiness[happiness['Country name'] == country].reset_inc
    difference = country_data.loc[len(country_data) - 1, 'Life Ladder'] - co
    result_tuple = (country, difference)
    results.append(result_tuple)

df = pd.DataFrame(results, columns=['Country', 'Increase in Happiness'])

df[df['Increase in Happiness']==df['Increase in Happiness'].max()]
```

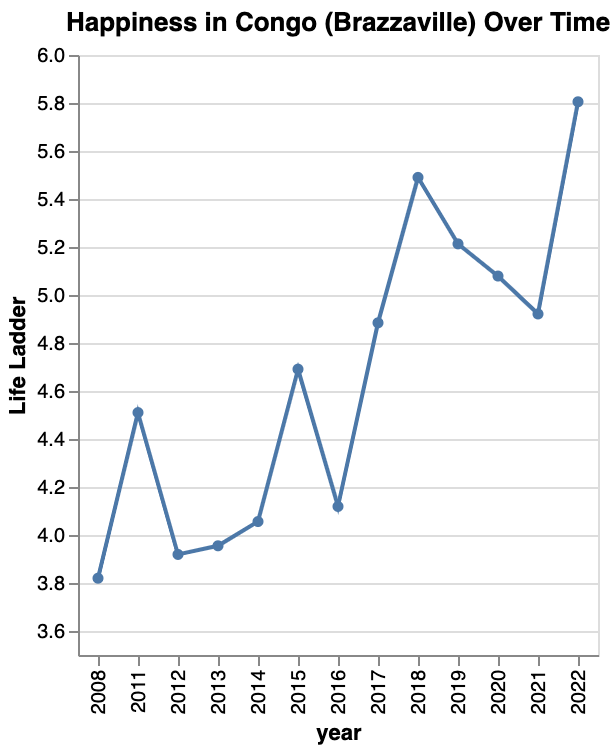
Out[15]:

	Country	Increase in Happiness
32	Congo (Brazzaville)	1.985

In [16]:
```
alt.Chart(happiness[happiness['Country name'] == 'Congo (Brazzaville)']).mar
    x = 'year:N',
```

```
    y = alt.Y('Life Ladder', scale=alt.Scale(domain=[3.5,6]))).properties(
    title = 'Happiness in Congo (Brazzaville) Over Time')
```

Out[16]:



In order to quantify the "greatest improvement" in happiness over time, we took the most oldest life ladder score from each country and subtracted it from that country's most recent life ladder score. Thus, we found that Congo (Brazzaville), also known as the Repubilic of Congo (not to be confused with the Democratic Republic of Congo), has seen the greatest improvement in happiness with their Life Ladder score increasing by nearly 2 points between 2008 and 2022.

Which country has consistently been the happiest?

In [17]:
```
# Which country has consistently been the happiest
happiness_bycountry = happiness.iloc[:,[0,2]].groupby('Country name').mean()
happiness_bycountry
happiness_bycountry[happiness_bycountry['Life Ladder']==happiness_bycountry[
```

Out[17]:

	Country name	Life Ladder
39	Denmark	7.673529

In [18]:
```
happiness[happiness['Country name']=='Denmark']
```

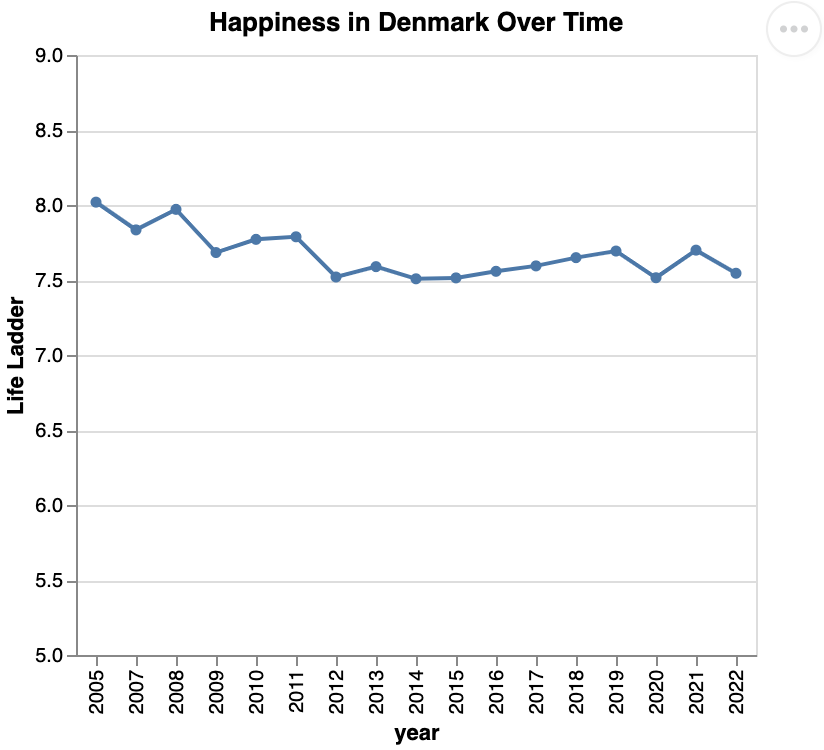
Out[18]:

	Country name	year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptio corruptio
505	Denmark	2005	8.019	10.849	0.972	68.300	0.971	NaN	0.23
506	Denmark	2007	7.834	10.889	0.954	68.740	0.932	0.236	0.20
507	Denmark	2008	7.971	10.878	0.954	68.960	0.970	0.268	0.24
508	Denmark	2009	7.683	10.822	0.939	69.180	0.949	0.259	0.20
509	Denmark	2010	7.771	10.836	0.975	69.400	0.944	0.238	0.17
510	Denmark	2011	7.788	10.845	0.962	69.620	0.935	0.293	0.22
511	Denmark	2012	7.520	10.844	0.951	69.840	0.933	0.135	0.18
512	Denmark	2013	7.589	10.849	0.965	70.060	0.920	0.211	0.17
513	Denmark	2014	7.508	10.860	0.956	70.280	0.942	0.114	0.23
514	Denmark	2015	7.514	10.876	0.960	70.500	0.941	0.218	0.19
515	Denmark	2016	7.558	10.900	0.954	70.625	0.948	0.134	0.21
516	Denmark	2017	7.594	10.922	0.952	70.750	0.955	0.151	0.18
517	Denmark	2018	7.649	10.936	0.958	70.875	0.935	0.013	0.15
518	Denmark	2019	7.693	10.948	0.958	71.000	0.963	0.016	0.17
519	Denmark	2020	7.515	10.924	0.947	71.125	0.938	0.047	0.21
520	Denmark	2021	7.699	10.968	0.945	71.250	0.933	0.131	0.17
521	Denmark	2022	7.545	10.994	0.970	71.375	0.930	0.224	0.20

In [19]:

```
alt.Chart(happiness[happiness['Country name'] == 'Denmark']).mark_line(point
    x = 'year:N',
    y = alt.Y('Life Ladder', scale=alt.Scale(domain=[5,9]))).properties(
    title = 'Happiness in Denmark Over Time')
```

Out[19]:



In order to establish whcih country has *consistently* been the happiest, we simply looked at the average life ladder score of each country between 2005 and 2022. We found that Denmark has consistently been the most happy country between 2005 and 2022 with an average life ladder score of about 7.67 during this time period.

Now that we have established that: there has been a slight increase in the overall hapiness in the world, Congo (Brazzaville) has seen the largest increase in happiness, and Denmark has consistenly been the most happy country, we wanted to figure out *why* these trends occured. In order to do so we performed principal component analysis and multiple linear regressions.

Principal Component Analysis

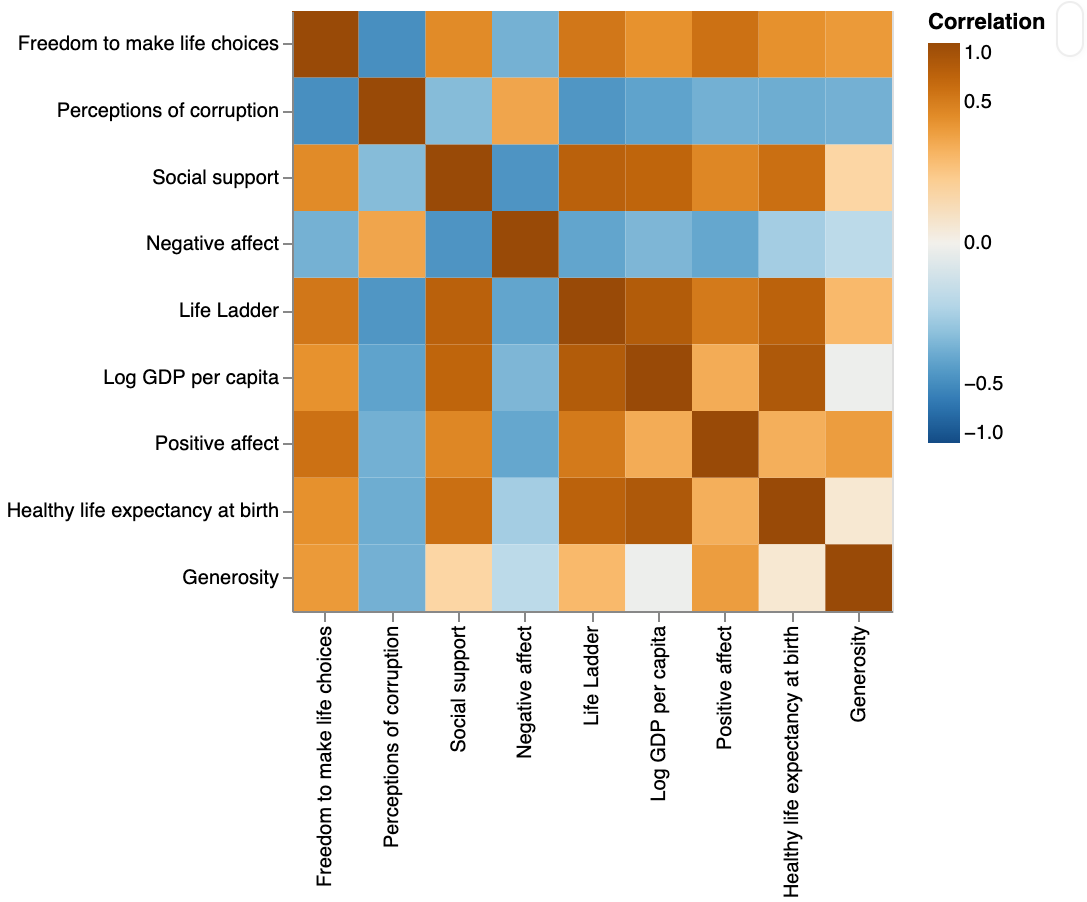
In [20]:
```
# Creating correlation matrix
x_mx = happiness.drop(columns = ['Country name', 'year'])
corr_mx = x_mx.corr()
```

In [21]:
```
corr_mx.loc[:, 'Life Ladder'].sort_values()
```

Out[21]:
```
Perceptions of corruption          -0.431500
Negative affect                    -0.339969
Generosity                          0.181630
Positive affect                     0.518169
Freedom to make life choices        0.534493
Healthy life expectancy at birth    0.713499
Social support                      0.721662
Log GDP per capita                  0.784868
Life Ladder                         1.000000
Name: Life Ladder, dtype: float64
```

In [57]:
```
# Melting corr_mx
corr_mx_long = corr_mx.reset_index().rename(
    columns = {'index': 'row'}
).melt(
    id_vars = 'row',
    var_name = 'col',
    value_name = 'Correlation'
)

# Constructing heatmap
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', # diverging g
                                        domain = (-1, 1), # ensure white = 0
                                        type = 'sqrt'), # adjust gradient sc
                      legend = alt.Legend(tickCount = 5)) # add ticks to colo
).properties(width = 300, height = 300)
```

Out[57]:



From this heat map we can see that the variables Log GDP per Capita, Social support, and Healthy life expectancy at birth are all strongly *positively* correlated with countries' life ladder score. Likewise, we can see that Perceptions of corruption and Negative affect are strongly *negatively* correlated with a country's life ladder score.

In [23]:
```
# Checking how many rows have missing values in our data set
missing_values = x_mx.isnull()
missing_values.any(axis=1).sum()/len(x_mx)
```

Out[23]:
```
0.10959527057753524
```

In order to do principal component analysis, we cannot have any missing values in our dataset. However, as seen above, approximately 11% of our observations have one or more missing values. In order to deal with this missingness we decided to delete the rows containing missing values. We did not feel mean imputation was appropriate because variable values can differ so much between countries and/or regions. Thus, we did not think that substituting an overall mean of each variable to fill missing values was the correct course of action.

In [30]:
```
# Removing rows with missing values
x_mx_nona = x_mx.dropna()
```

In [31]:
```
pca = PCA(data = x_mx_nona, standardize = True)
```

In [33]:
```
# Computing variance ratios
var_ratios = pca.eigenvals/pca.eigenvals.sum()
var_ratios
```

Out[33]:
```
0    0.468191
1    0.168663
2    0.104558
3    0.083123
4    0.069340
5    0.040740
6    0.030082
7    0.020415
8    0.014889
Name: eigenvals, dtype: float64
```

In [58]:
```
# Creating dataframe with proportion of variance explained by each comonent
pca_var_explained = pd.DataFrame({
    'Component': np.arange(1, 10),
    'Proportion of variance explained': var_ratios})

pca_var_explained['Cumulative variance explained'] = var_ratios.cumsum()

pca_var_explained
```

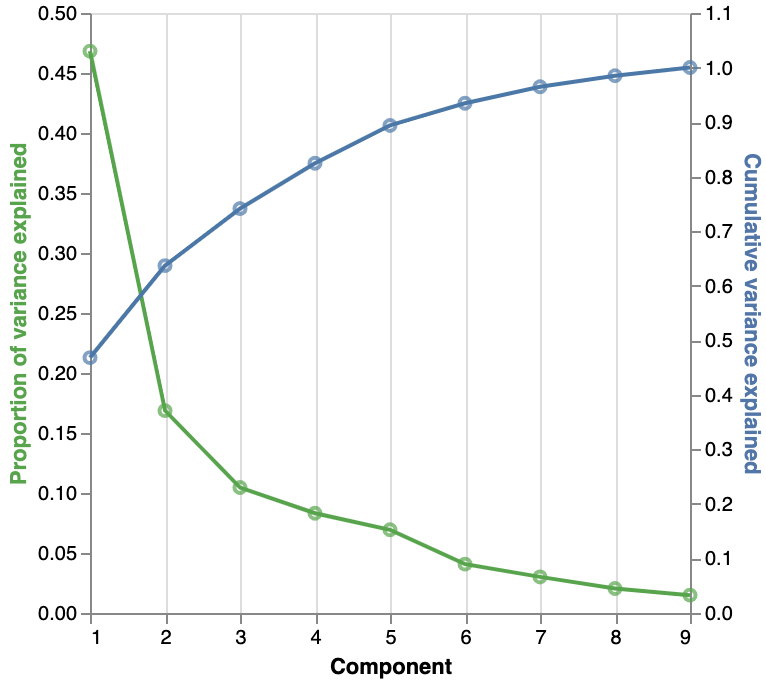
Out[58]:

	Component	Proportion of variance explained	Cumulative variance explained
0	1	0.468191	0.468191
1	2	0.168663	0.636853
2	3	0.104558	0.741411
3	4	0.083123	0.824534
4	5	0.069340	0.893874
5	6	0.040740	0.934614
6	7	0.030082	0.964695
7	8	0.020415	0.985111
8	9	0.014889	1.000000

In [40]:
```
# Plotting variance explained by each component and cumulitave variance
base = alt.Chart(pca_var_explained).encode(
    x = 'Component')

prop_var_base = base.encode(
    y = alt.Y('Proportion of variance explained',
              axis = alt.Axis(titleColor = '#57A44C'))
)

cum_var_base = base.encode(
    y = alt.Y('Cumulative variance explained', axis = alt.Axis(titleColor =
)

prop_var = prop_var_base.mark_line(stroke = '#57A44C') + prop_var_base.mark_
```

```
cum_var = cum_var_base.mark_line() + cum_var_base.mark_point()

var_explained_plot = alt.layer(prop_var, cum_var).resolve_scale(y = 'indeper

# display
var_explained_plot
```

Out[40]:



From this graph, we decided to use 4 principal componenets as over 80% of total variance can be explained by these 4 componenets.

In [56]:
```
# Subsetting only the 4 component loadings that we chose
loading_df = pca.loadings.iloc[:, 0:4]
loading_df = loading_df.rename(columns = dict(zip(loading_df.columns, ['PC'
loading_df
```

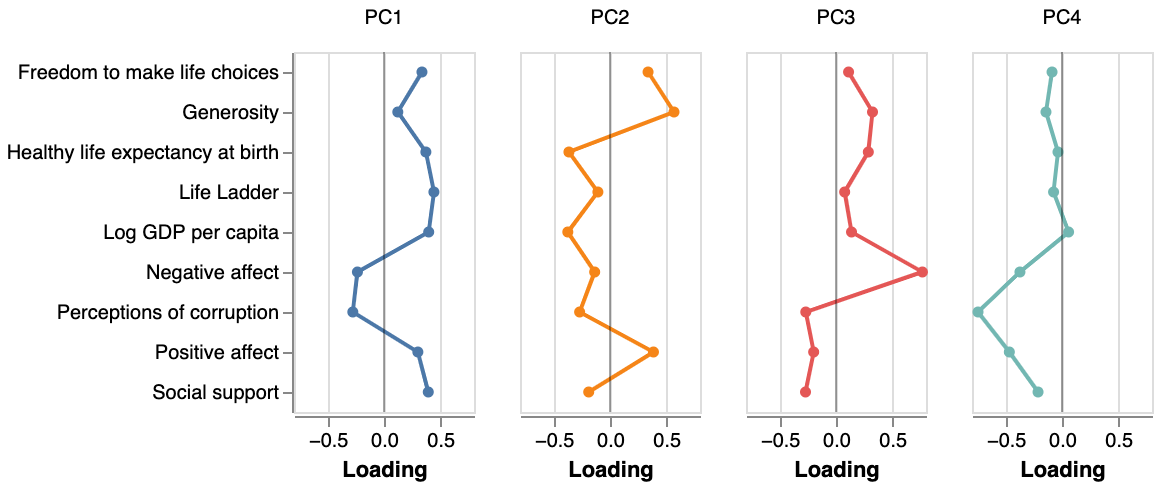
Out[56]:

	PC1	PC2	PC3	PC4
Life Ladder	0.443544	-0.107710	0.076934	-0.074852
Log GDP per capita	0.397397	-0.375113	0.138024	0.058534
Social support	0.393455	-0.189181	-0.271101	-0.212935
Healthy life expectancy at birth	0.372133	-0.365084	0.287211	-0.035678
Freedom to make life choices	0.337298	0.338298	0.111823	-0.088597
Generosity	0.122678	0.568857	0.324524	-0.142566
Perceptions of corruption	-0.277184	-0.269913	-0.268638	-0.746495
Positive affect	0.300972	0.386196	-0.198559	-0.468106
Negative affect	-0.236262	-0.136264	0.767875	-0.373903

In [62]:
```
# Plotting each principal component's loadings
loading_plot_df = loading_df.reset_index().melt(
```

```
    id_vars = 'index',
    var_name = 'Principal Component',
    value_name = 'Loading'
).rename(columns = {'index': 'Variable'})

loading_plot_df['zero'] = np.repeat(0, len(loading_plot_df))

base = alt.Chart(loading_plot_df)

loadings = base.mark_line(point = True).encode(
    y = alt.Y('Variable', title = ''),
    x = 'Loading',
    color = 'Principal Component'
)

rule = base.mark_rule().encode(x = alt.X('zero', title = 'Loading'), size =

loading_plot = (loadings + rule).properties(width = 90)

loading_plot.facet(column = alt.Column('Principal Component', title = ''))
```

Out[62]:



Within a principal component, the variables with the largest loadings/weights (in terms of absolute value) are the ones that are most influential to that component. And since these componenets together capture over 80% of the total variation in the data, the heaviest weighted variables in each componenet are the ones that drive variation in our data.

We interpreted the first principal component (PC1) as a representation of the difference between well-being and danger. This is because, in PC1, the variables that are most influential are: Life ladder(positive), Log GDP per capita (positive), Social support (positive), Perceptions of corruption (negative), and Negative affect (negative). This means that a country with a larger value of PC1 would have a higher than average life ladder score, log GDP per capita, and social support while also having a lower than average perception of corruption and negative affect score. Likewise, a country with a smaller value of PC1 would have a lower than average life ladder score, log GDP per capita, and social support while also having a higher than average perception of corruption and negative affect score.

Next, we interpreted the second principal component (PC2) as a reprentation of the difference between people's willingness to help others and a country's overall wealth. This is because, in PC2, the variables that are most influential are: Generosity (positive), Positive affect (positive), Log GDP per capita (negative), and Healthy life expectancy at birth (negative). This means that a country with a larger value of PC2 would have a higher than average Generosity score and Positive affect score while also having a lower than average Log GDP per capita and Healthy life expectancy at birth. Likewise, a country with a smaller value of PC2 would have a lower than average Generosity score and Positive affect score while also having a higher than average Log GDP per capita and Healthy life expectancy at birth.

Then, we interpreted the third principal component (PC3) as a representaion of the average between people's generosity and negative feelings. This is because, in PC3, the variables that are most influential are Negative affect (positive) and Generosity (positive). This means that a country with a larger value of PC3 would have a higher than average Negative affect score and Generosity score. Likewise, a country with a smaller value of PC3 would have a lower than average Negative affect score and Generosity score.

Finally, we interpreted the fourth principal component (PC4) as a representation of the average between people's perceptions of corruption, positive feelings, and negative feelings. This is because, in PC4, the variables that are most influential are Perceptions of corruption (negative), Positive affect (negative), and Negative affect (negative). This means that a country with a larger value of PC4 would have a higher than average Perception of corruption, Positive affect score, and Negative affect score. Likewise, a country with a smaller value of PC4 would have a lower than average Perception of corruption, Positive affect score, and Negative affect score.

So, from this principal componenet analysis we have determined that the variables that drive the most variation in our data are Life Ladder, Log GDP per capita, Social support, Generosity, Negative affect, and Perceptions of corruption.

Summary of findings

Throughout this project we have found that there has been a slight increase in overall happiness in the world. [INSERT BREAKDOWN BY COUNTRY/REGION HERE IF YOU WANT]. We also found that the variables that drove the variation in our data were: Life Ladder, Log GDP per capita, Social support, Generosity, Negative affect, and Perceptions of corruption. [INSERT ANY OTHER FINDINGS FROM REGRESSION HERE]