Data Analysis on World Happiness Report 2023

Author: Praveen Manimaran

Class: PSTAT 100

In [32]:
```
import numpy as np
import pandas as pd
import altair as alt
from sklearn.decomposition import PCA
import statsmodels.api as sm
import matplotlib.pyplot as plt
# disable row limit for plotting
alt.data_transformers.disable_max_rows()
# uncomment to ensure graphics display with pdf export
alt.renderers.enable('mimetype')
```

Out[32]:
```
RendererRegistry.enable('mimetype')
```

Data Description:

The dataset chosen for this analysis is the World Happiness Report 2023, which provides insights into the various factors that affect subjective well-being across different countries such as life ladder scores, log GDP per capita, social support, healthy life expectancy at birth, freedom to make life choices, generosity, perceptions of corruption, positive affect, and negative affect. The dataset is from the Gallup World Poll (GWP) which which collects information on various well-being indicators from annually conducted surveys in each country. The GWP collects responses from 1000 individuals per year in each country, to construct population-representative national averages. The dataset uses the main indicator as "Life Ladder" which is individual's self-evaluation on their current overall life satisfaction. We can make inferences about this dataset to find out which key variables impact individual's self-evaluation on their life satisfaction.

Name	Variable description	Type
Country name	Name of Country	Categorical
Year	Year that survey took place	Categorical
Life Ladder	Weighted averages for individuals response to evaluate their current life as a whole using the image of a ladder, with the best possible life for them as a 10 and worst possible as a 0	Numeric
Log GDP Per Capita	The natural log of GDP per capita; GDP per capita is in terms of Purchasing Power Parity (PPP)	Numeric
Social Support	National average of the binary responses (0=no, 1=yes) 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?”	Numeric
Healthy Life Expectancy at Birth	Healthy life expectancy at birth based on data from the World Health Organization (WHO)	Numeric
Freedom to Make Life Choices	National average of binary responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”	Numeric
Generosity	Residual of regressing the national average of GWP responses to the donation question “Have you donated money to a charity in the past month?” on log GDP per capita.	Numeric
Perceptions of Corruption	The average of binary answers to two GWP questions: “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?”	Numeric
Positive Affect	The average of previous-day affect measures for laughter, enjoyment, and interest.	Numeric
Negative Affect	The average of previous-day affect measures for worry, sadness, and anger.	Numeric

In [17]:
```
#Reading in data file
whr = pd.read_csv('data/whr-2023.csv')
whr.head(10) #Look at first 10 lines of code
```

Out[17]:

	Country name	year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Positive affect	Negative affect
0	Afghanistan	2008	3.724	7.350	0.451	50.500	0.718	0.168	0.882	0.414	0.258
1	Afghanistan	2009	4.402	7.509	0.552	50.800	0.679	0.191	0.850	0.481	0.237
2	Afghanistan	2010	4.758	7.614	0.539	51.100	0.600	0.121	0.707	0.517	0.275
3	Afghanistan	2011	3.832	7.581	0.521	51.400	0.496	0.164	0.731	0.480	0.267
4	Afghanistan	2012	3.783	7.661	0.521	51.700	0.531	0.238	0.776	0.614	0.268
5	Afghanistan	2013	3.572	7.680	0.484	52.000	0.578	0.063	0.823	0.547	0.273
6	Afghanistan	2014	3.131	7.671	0.526	52.300	0.509	0.106	0.871	0.492	0.375
7	Afghanistan	2015	3.983	7.654	0.529	52.600	0.389	0.082	0.881	0.491	0.339
8	Afghanistan	2016	4.220	7.650	0.559	52.925	0.523	0.044	0.793	0.501	0.348
9	Afghanistan	2017	2.662	7.648	0.491	53.250	0.427	-0.119	0.954	0.435	0.371

In [82]:
```
#Select Countries Only in North America
na_data = whr[whr['Country name'].isin(['Canada', 'United States', 'Mexico', 'Guatemala', 'Haiti', 'Cuba', 'Dominic
                                        'Honduras', 'Nicaragua', 'El Salvador', 'Costa Rica', 'Panama', '
                                         'Trinidad and Tobago', 'Belize'])]
```

Question of Interest:

Motivation:

The World Happiness Report Dataset tries to understand the potential factors that could be attributed to people's life ladder scores. They use various survey questions regarding social support, healthy life expectancy at birth, freedom to make life choices, generosity, perceptions of corruption, positive affect, and negative affect to see if there is any relationship between life ladder and any of these factors. I think that this is essentially the main goal of the dataset, so it is clear that this is the direction I should follow. However, I thought that the dataset could be more meaningful to me and those living in the United States/its neighboring countries to understand why these factors could impact individuals living close to us rather than include everyone across the globe who may or may not share the same living situations as the majority of those in North America.

Final Question: What factors would have the greatest effect on individuals' current life evaluation in North America?

Satisfactory Answer:

A satisfactory answer could be that the ability to have Freedom to make life choices and the countries' log GDP have a strong positive linear relationship with Life Ladder and are strongly correlated.

Exploratory Data Analysis

Compare Average Life Ladder By Countries

In [83]:
```
# compute average life ladder
avg_life_ladder_by_country = na_data[['Country name', 'Life Ladder']].groupby('Country name').mean()

# compute sample sizes
country_n = na_data['Country name'].value_counts().rename('n')

# concatenate
tbl_1 = pd.concat([avg_life_ladder_by_country, country_n], axis = 1)

# print
tbl_1
```

Out[83]:

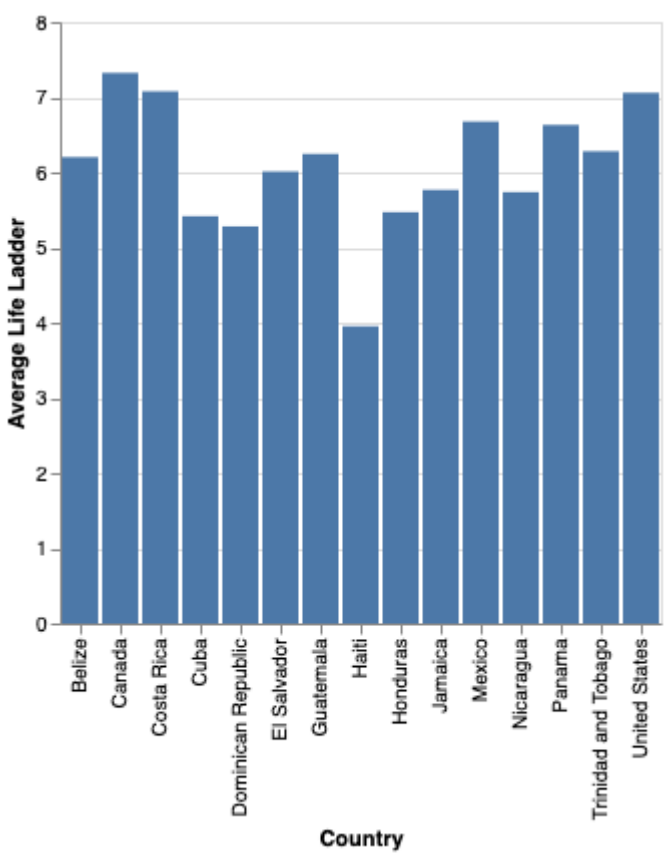
Country name	Life Ladder	n
Belize	6.203500	2
Canada	7.323647	17
Costa Rica	7.078471	17
Cuba	5.418000	1
Dominican Republic	5.281824	17
El Salvador	6.011882	17
Guatemala	6.246333	15
Haiti	3.954182	11
Honduras	5.471375	16
Jamaica	5.767778	9
Mexico	6.676353	17
Nicaragua	5.738118	17
Panama	6.631500	16
Trinidad and Tobago	6.281400	5
United States	7.059118	17

We notice that some countries don't have a lot of observations such as Cuba, Belize, and Trinidad and Tobago. This could definitely have some impact on our analysis.

Graph of Average Life Ladder by Countries

In [84]:
```
tbl_mod = tbl_1.reset_index().rename(columns={'index': 'Country name'})

fig_1 = alt.Chart(tbl_mod).mark_bar().encode(
    x=alt.X('Country name', title='Country'),
    y=alt.Y('Life Ladder', title='Average Life Ladder'),
)

fig_1
```

Out[84]:

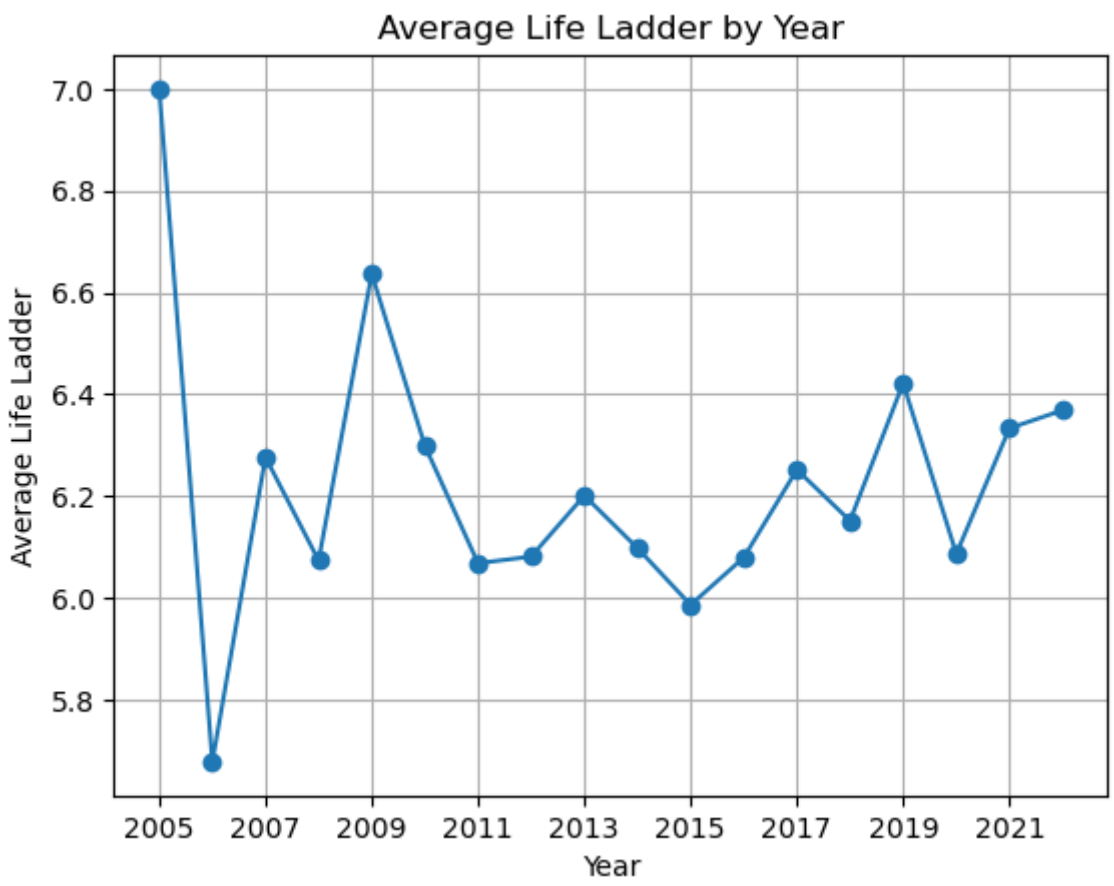


We notice that Haiti is much lower and has a much larger difference in average life ladder compared to other countries.

Plotting Average Life Ladder By Year

In [86]:
```
# Group the data by year and calculate the mean life ladder
avg_life_ladder_by_year = na_data.groupby('year')['Life Ladder'].mean()

# Plot the average life ladder by year
plt.plot(avg_life_ladder_by_year.index, avg_life_ladder_by_year.values, marker='o')
plt.xlabel('Year')
plt.ylabel('Average Life Ladder')
plt.title('Average Life Ladder by Year')
plt.grid(True)

# Set the x-axis tick locations and labels
```

```
years = range(2005, 2023, 2)
plt.xticks(years)

plt.show()
```

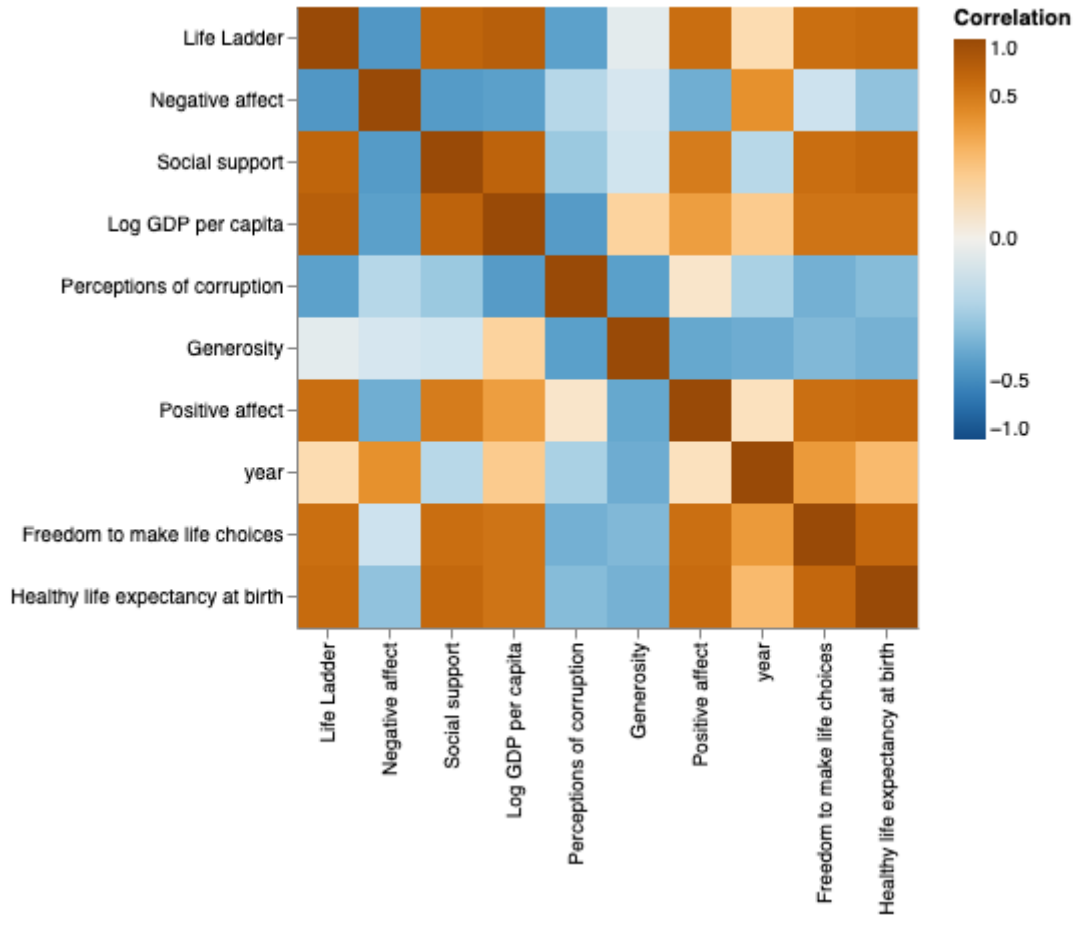


We can see that most of the average life ladders by year appear to be 6 and 6.7 with few outliers in 2005 and 2006.

Correlation Matrix of Various Factors on Life Ladder

In [87]:
```
# Select only numeric columns for correlation calculation
numeric_columns = na_data.select_dtypes(include=[np.number])

# Calculate correlation matrix
corr_mx = numeric_columns.corr()

# Melt corr_mx
corr_mx_long = corr_mx.reset_index().rename(columns={'index': 'row'}).melt(
    id_vars='row',
    var_name='col',
    value_name='Correlation')

# Construct heatmap
heatmap = alt.Chart(corr_mx_long).mark_rect().encode(
    x=alt.X('col', title='', sort=alt.EncodingSortField(field='Correlation', order='ascending')),
    y=alt.Y('row', title='', sort=alt.EncodingSortField(field='Correlation', order='ascending')),
    color=alt.Color('Correlation',
                    scale=alt.Scale(scheme='blueorange', domain=(-1, 1), type='sqrt'),
                    legend=alt.Legend(tickCount=5))
).properties(
    width=310,
    height=310
)

# Display the heatmap
heatmap
```

Out[87]:



From the correlation matrix/heatmap, we notice that life ladder has a strong positive correlation with positive affect and slight positive correlation with freedom to make life choices, social support, and healthy life expectancy at birth. There also appears to be a strong negative correlation between life ladder and negative affect & perception of corruption. These results seem to make sense intuitively.

Data Pre-Processing

Filling in Missing Values Through Imputation

In [25]:
```
#Dropping Country Name and Year
reg_na_data = na_data.copy().drop(columns = ['Country name','year'])

#Filling in Missing Values Through Imputation
reg_na_data['Log GDP per capita']=reg_na_data['Log GDP per capita'].fillna(reg_na_data['Log GDP per capita'].mean()
reg_na_data['Social support']=reg_na_data['Social support'].fillna(reg_na_data['Social support'].mean())
reg_na_data['Healthy life expectancy at birth']=reg_na_data['Healthy life expectancy at birth'].fillna(reg_na_data[
reg_na_data['Freedom to make life choices']=reg_na_data['Freedom to make life choices'].fillna(reg_na_data['Freedom
reg_na_data['Generosity']=reg_na_data['Generosity'].fillna(reg_na_data['Generosity'].mean())
reg_na_data['Perceptions of corruption']=reg_na_data['Perceptions of corruption'].fillna(reg_na_data['Perceptions o
reg_na_data['Positive affect']=reg_na_data['Positive affect'].fillna(reg_na_data['Positive affect'].mean())
reg_na_data['Negative affect']=reg_na_data['Negative affect'].fillna(reg_na_data['Negative affect'].mean())
reg_na_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 194 entries, 179 to 2087
Data columns (total 9 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Life Ladder                       194 non-null    float64
 1   Log GDP per capita                194 non-null    float64
 2   Social support                    194 non-null    float64
 3   Healthy life expectancy at birth  194 non-null    float64
 4   Freedom to make life choices      194 non-null    float64
 5   Generosity                        194 non-null    float64
 6   Perceptions of corruption         194 non-null    float64
 7   Positive affect                   194 non-null    float64
 8   Negative affect                   194 non-null    float64
dtypes: float64(9)
memory usage: 15.2 KB
```

Multiple Linear Regression

In [89]:
```
#Obtain the explanatory variable matrix and response vector needed to fit the linear model.
x = sm.tools.add_constant(reg_na_data.copy().drop(columns = ['Life Ladder']))
y = np.asarray(reg_na_data['Life Ladder'])

# Fitting Regression Model
# fit model
mlr = sm.OLS(endog = y, exog = x)
rslt = mlr.fit()

# retreive estimates and std errors
coef_tbl = pd.DataFrame({
    'estimate': rslt.params.values,
```

```
    'standard error': np.sqrt(rslt.cov_params().values.diagonal())
    },
    index = x.columns.values
)

coef_tbl.loc['error variance', 'estimate'] = rslt.scale

# display table
coef_tbl
```

Out[89]:

	estimate	standard error
const	-1.694083	0.913522
Log GDP per capita	0.460391	0.065345
Social support	0.764093	0.718055
Healthy life expectancy at birth	0.003873	0.007425
Freedom to make life choices	0.341581	0.442800
Generosity	0.212073	0.313845
Perceptions of corruption	-1.270808	0.346318
Positive affect	5.249615	0.779981
Negative affect	-2.815876	0.927177
error variance	0.226675	NaN

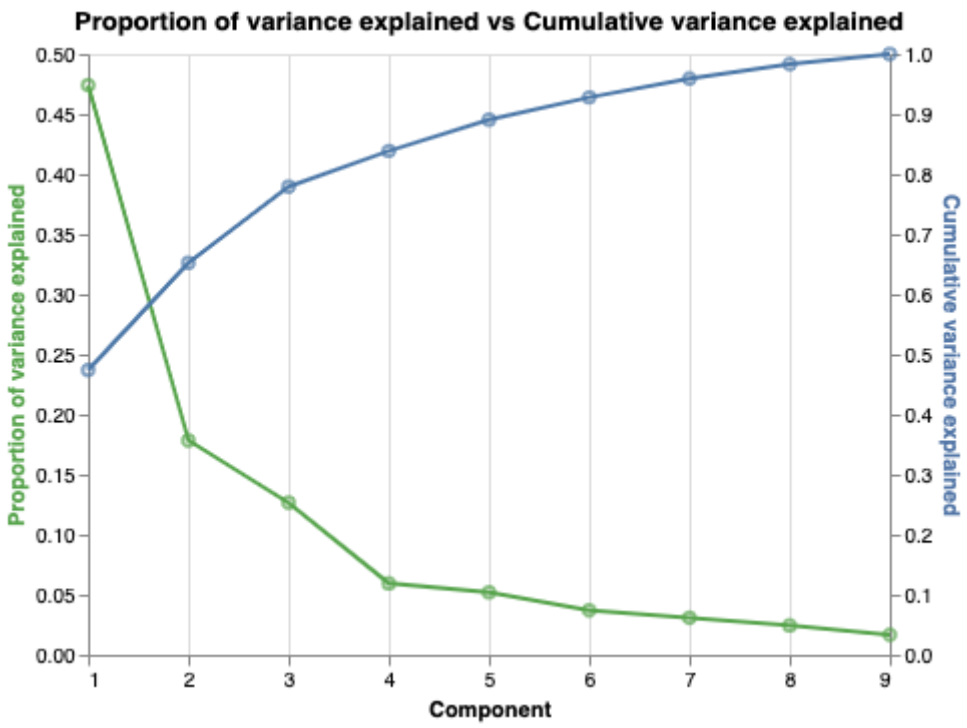
We can see that higher levels of social support, freedom to make life choices, generosity, log GDP per Capita, and Positive affect has a positive relationship with Life Ladder. Higher levels of positive affect are associated with a significantly higher Life Ladder compared to other factors/variables. We can also see that Perceptions of corruption and negative affect have a strong negative relationship with Life Ladder.

Principle Component Analysis

In [90]:

```
## (i) center and scale data

na_data_raw = reg_na_data

# center and scale the relative abundances
na_data2 = (na_data_raw - na_data_raw.mean())/na_data_raw.std()

## (ii) compute pcs
pca = PCA(n_components = na_data2.shape[1])
pca.fit(na_data2)

# Variance ratios
pca_var_explained = pd.DataFrame({'Proportion of variance explained': pca.explained_variance_ratio_})
pca_var_explained['Component'] = np.arange(1, pca_var_explained.shape[0] + 1)
pca_var_explained['Cumulative variance explained'] = pca_var_explained['Proportion of variance explained'].cumsum()

# Plotting the results
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'))
).properties(title='Proportion of variance explained')

cum_var_base = base.encode(
    y=alt.Y('Cumulative variance explained', axis=alt.Axis(titleColor='#5276A7'))
).properties(title='Cumulative variance explained')

prop_var = prop_var_base.mark_line(stroke='#57A44C') + prop_var_base.mark_point(color='#57A44C')
cum_var = cum_var_base.mark_line() + cum_var_base.mark_point()

Fig6 = alt.layer(prop_var, cum_var).resolve_scale(y='independent').properties(
    title='Proportion of variance explained vs Cumulative variance explained'
)

Fig6
```

Out[90]:



This plot indicates that the first 3 principal components explain more than 70% of variation and covariation. Components after the 3rd principal component do not contribute much to the total variance explained. So, the principal components selected will be PC1, PC2, and PC3.

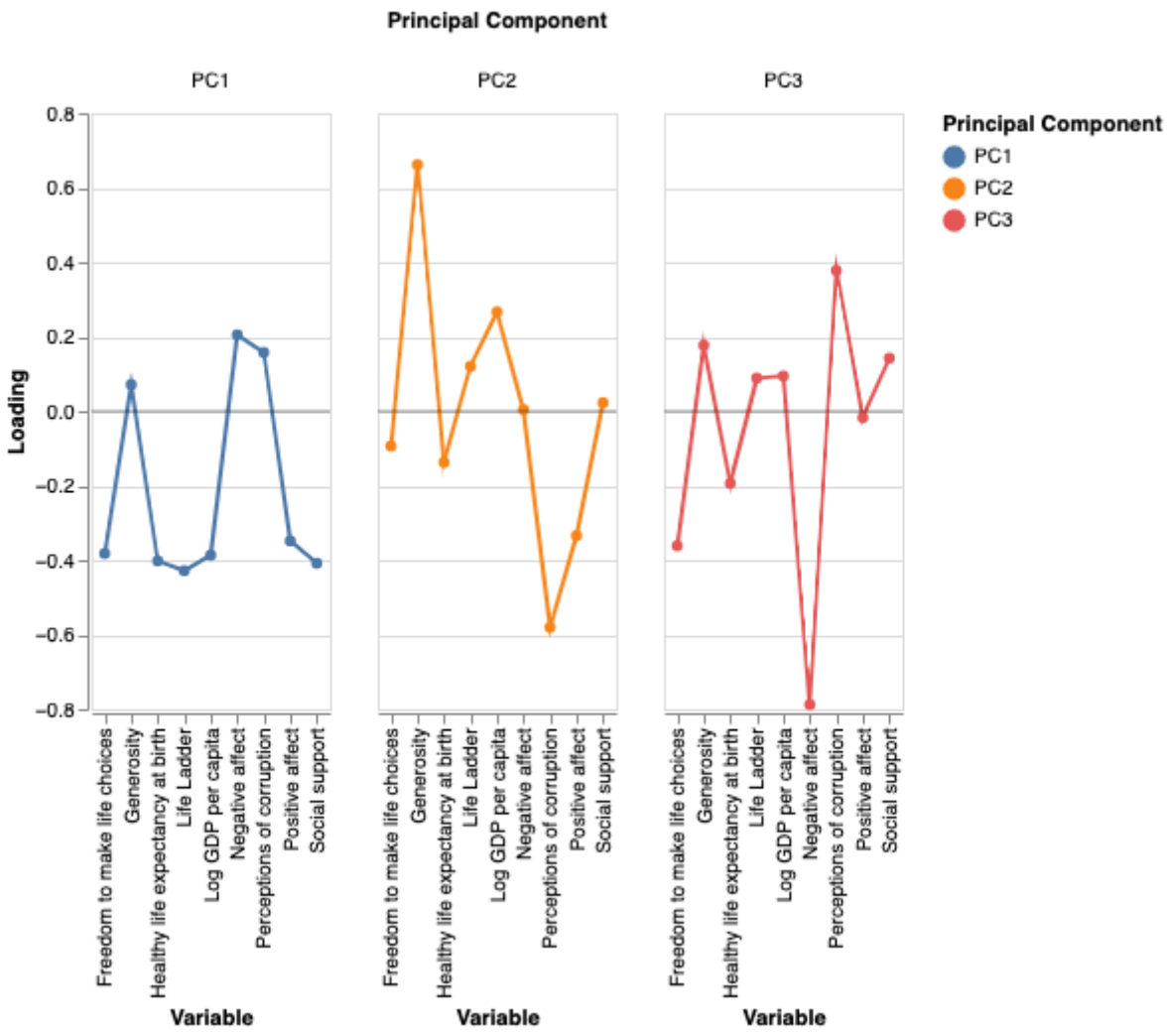
In [59]:
```
loading_df = pd.DataFrame(pca.components_).transpose().rename(
    columns = {0: 'PC1', 1: 'PC2', 2: 'PC3'}).loc[:, [
    'PC1', 'PC2', 'PC3']]

loading_df['Variable'] = na_data2.columns.values

loading_df
```

Out[59]:

	PC1	PC2	PC3	Variable
0	-0.428159	0.120665	0.088972	Life Ladder
1	-0.386413	0.266506	0.094447	Log GDP per capita
2	-0.408143	0.022784	0.142445	Social support
3	-0.401862	-0.137297	-0.193598	Healthy life expectancy at birth
4	-0.382030	-0.093481	-0.361343	Freedom to make life choices
5	0.071611	0.661870	0.177188	Generosity
6	0.157811	-0.580173	0.377870	Perceptions of corruption
7	-0.348280	-0.334089	-0.017228	Positive affect
8	0.204968	0.004205	-0.787621	Negative affect

In [80]:
```
# melt from wide to long
loading_plot_df = loading_df.melt(
    id_vars = 'Variable',
    var_name = 'Principal Component',
    value_name = 'Loading'
)

# add a column of zeros to encode for x = 0 line to plot
loading_plot_df['zero'] = np.repeat(0, len(loading_plot_df))

# create base layer
base = alt.Chart(loading_plot_df)


# create line at zero
rule = base.mark_rule().encode(y = alt.Y('zero', title = 'Loading'), size = alt.value(0.05))


# create lines + points for loadings
loadings = base.mark_line(point = True).encode(
    y = alt.Y('Loading', title = ''),
    x = 'Variable',
    color = 'Principal Component'
)

# layer
loading_plot = (loadings + rule).properties(width = 120)

# show
loading_plot.facet(column = 'Principal Component')
```

Out[80]:



For PC1, Generosity, Negative affect, and Perceptions of corruption are up-weighted because they have positive loadings. However, Generosity has a loading quite lower than the other two and closer to zero, so it may not up-weight PC1 very much. We cannot conclude that it does not up-weight PC1 because it is not extremely close to zero. Since all other variables have negative weights, they are down-weighted. Their weights are also quite close to one another. These observations imply that higher value of Negative affect, Perceptions of corruption, and Generosity correspond with lower values of Freedom to make life choices, Healthy life expectancy at birth, Life Ladder, Log GDP per capita, Positive affect, and Social support. This makes sense as those two are opposite ends of negative and positive life experiences (except for Generosity). It is quite surprising, though, that Generosity does not have a negative weight.

For PC2, Generosity, Life Ladder, and Log GDP per capita are up-weighted due to their positive loadings. Generosity appears to have a much higher positive loading than the other variables. Negative affect and Social support have loadings extremely close to zero, so it seems reasonable to claim that they do not up-weight or down-weight PC2. All other variables have negative loadings, so they are down-weighted. However, Perceptions of corruption appears to have a much higher negative loading than the other variables. This means that higher values of Generosity, Life Ladder, and Log GDP per capita correspond with lower values of Freedom to make life choices, Healthy life expectancy at birth, Perceptions of corruption, and Positive affect. This principal component appears to be measuring one's economic status in relation to the other variables. It makes sense that those with greater wealth would have higher Log GDP per capita. They would also have a greater ability to be generous (explaining the high positive loading) and their perception of their Life Ladder would be higher. Also, the negative loadings make sense because those who are not as well off would not have as much freedom to make life choices and they also tend to have lower life expectancy. Their perceptions of corruption would also be greater because they do not tend to be the ones in power (which would explain why this had such a high negative loading). Also, it is quite surprising that poorer individuals appear to have a higher Positive affect. Perhaps this is due to not having to worry about maintaining economic status.

For PC3, Generosity, Life Ladder, Log GDP per capita, Perceptions of Corruption, and Social support have positive weights so they are up-weighted. Perceptions of corruption appears to have a much higher positive loading than the other variables. Positive affect is has a loading extremely close to zero so it seems reasonable to claim that it does not up-weight or down-weight PC3. All other variables have negative loadings, so they are down-weighted. Negative affect appears to have a significantly higher negative loading than the other variables. This means that higher values of Generosity, Life Ladder, Log GDP per capita, Perceptions of corruption, and Social support correspond with lower values of Freedom to make life choices, Healthy life expectancy, and Negative affect. This principal component is quite close to PC2; however it is different in that now Social support is taken into account. It appears that those who are well off with greater Social support also seem to have greater Perceptions of corruption (explaining the high positive loading) than those who are poorer. The big difference this makes from PC2 is that this greater presence of Social support and Perceptions of corruption causes those who are in better economic positions to have much lower Negative affect than those who have lower economic status. This explains the loading for Negative affect being vastly greater than all other loadings in the negative direction. Perhaps this principal component suggests the great effect Social support and Perceptions of corruption can have on whether an individual experiences more negative emotions. This also is apparent in the fact that Positive affect was greater for those who are poorer in PC2.

Summary of findings

Life Ladder (life satisfaction) tends to be positively associated with higher levels of Log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, and Positive affect. Conversely, Negative affect, Perceptions of corruption, and Generosity are negatively associated with life satisfaction. Through Principal Component Analysis, I was able to also find out that economic factors, social support, and perceptions of corruption played crucial roles in determining individuals' life satisfaction. For

example, those who are more wealthy are more likely to be generous and have higher Log GDP per capita making their Life Ladder much higher.

In conclusion, higher economic status, stronger social support, higher levels of positive affect, and lower perceived corruption appear to have a positive impact on a person's self-evaluation of their overall well-being in North America.