189 Final Project

> Daniel Mathew / Fangke Jiang / Aadhya Naveen / Eric Gu / Aman Kar / Yongxin Li

Examining the happiness index would be a good way to quantitatively measure people’s well-being around the world.

Many studies have shown that happiness is an important indicator for assessing overall life satisfaction, and we found that the topic would be interesting to research. We can identify which elements—such as GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption—most strongly influence happiness levels.

This helps in understanding how different aspects of society and the economy interact to affect well-being. This type of analysis contributes to the academic and scientific understanding of happiness, providing empirical evidence to support theories in psychology, economics, and social sciences. It fosters interdisciplinary collaboration in researching and addressing one of humanity's most fundamental pursuits: the quest for happiness.

> Reason choose World Happiness Report(WHR):

The WHR delves into the multifaceted concept of happiness on a global scale. The report goes beyond merely measuring subjective well-being to explore the underlying factors and conditions contributing to people’s happiness and life satisfaction.

The report covers:

Subjective Well-being: this accesses people’s self-reported happiness levels and life satisfaction by considering various factors such as health, social support, income, freedom, trust in government, and generosity.

Country Rankings: by analysis of happiness levels across nations, the ranking offers insight into which countries are effectively fostering the conditions for happiness among their citizens.

Determinants of Happiness: including economic factors like GDP per capita, social factors like social support networks and community engagement, health indicators like life expectancy and access to healthcare, also psychological factors like personal freedoms and perceptions of corruption.

Policy Implications: the report prioritizes policies and initiatives for policymakers to improve citizens' well-being and quality of life. It also encourages governments to go beyond traditional economic metrics and consider holistic measures of progress and prosperity.

Global Trends and Patterns: it also highlights global trends and patterns in happiness levels through longitudinal analysis and comparison. It reveals the impact of social changes, cultural differences, and economic and social policies on people's well-being.

Sustainability and Development: by emphasizing the importance of sustainable development and social well-being, the report recognized that economic growth alone is not enough to promote happiness; and advocated for policies that prioritize social inclusion, environmental sustainability, and equitable distribution of resources.

In conclusion, the WHR serves as a comprehensive resource for understanding the complex interplay of factors that contribute to happiness and well-being on a global scale, offering valuable insights for our research.

EDA

> We will be focusing on the Excel sheets of year 2011 - 2022 that contain the countries and scores such as life expectancy, GDP, freedom, etc. These scores are calculated on a specific index curated for the happiness reports.

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy.stats as stats
```

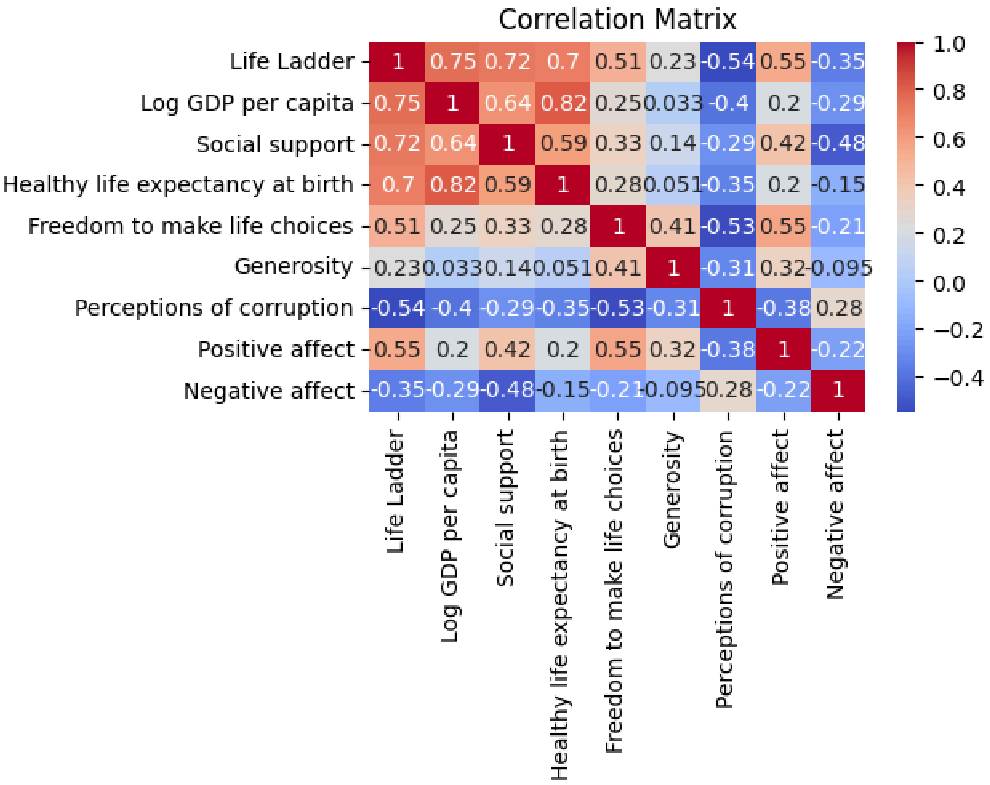
In [6]:
```
data = pd.read_csv('happy_data.csv')
filtered_data = data[data['year'].between(2011, 2022)]
# Group by 'Country name' and filter out groups that don't have all years from 2011 to 2022
continuous_countries = filtered_data.groupby('Country name').filter(lambda x: len(x['year'].unique()) == (2022-2
continuous_country=continuous_countries.groupby(['Country name'])
```

> We're going to focus on the analysis of mostly the top 5 and bottom 5 countries with features form the dataset: 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.

In [7]:
```
numerical_columns = ['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']
```

In [8]:
```
# Compute the correlation matrix for numerical variables
correlation_matrix = continuous_countries[numerical_columns].corr()

# Use seaborn to visualize the correlation matrix
plt.figure(figsize=(5, 3))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



The correlation coefficient values range from -1 to 1, indicated by the color scale on the right. A coefficient close to 1 (red shades) means there is a strong positive correlation, which means that as one variable increases, the other variable tends to also increase. A coefficient close to -1 (blue shades) means there is a strong negative correlation, which means that as one variable increases, the other variable tends to decrease. A coefficient around 0 (white or light shades) means there is no linear correlation between the variables.

Life Ladder and Log GDP per capita have a correlation of 0.75, which is strong and positive, indicating that higher GDP per capita is associated with higher scores on the Life Ladder (often a measure of well-being or happiness).

Life Ladder and Perceptions of corruption have a correlation of -0.54, which is a moderate negative correlation, suggesting that higher perceptions of corruption are associated with lower Life Ladder scores.

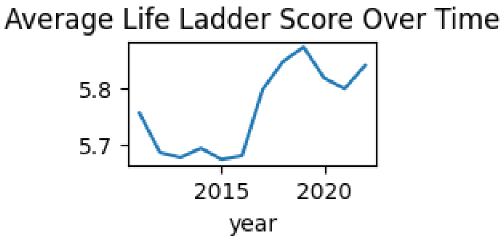
Generosity and Positive affect have a correlation of 0.32, a weak positive correlation, indicating that there is a slight tendency for higher generosity to be associated with higher positive affect, but the relationship is not very strong.

Life Ladder

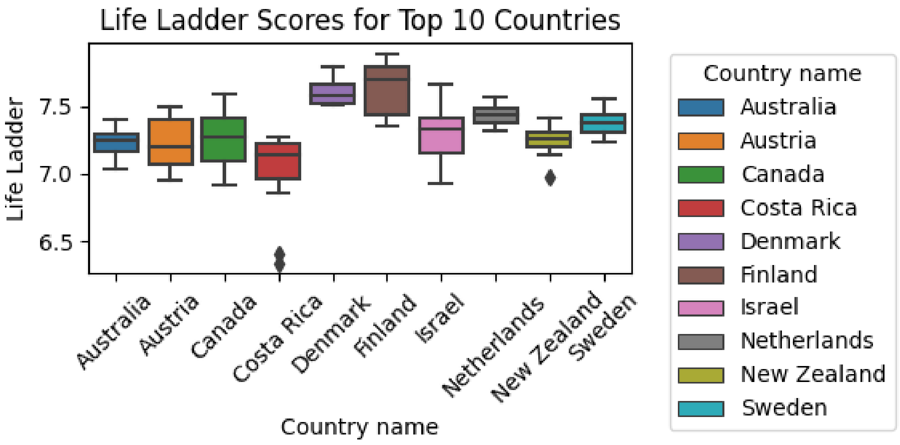
The Life Ladder essentially asks respondents to rate their current life satisfaction on a scale, typically ranging from 0 to 10, with 0 representing the worst possible life and 10 representing the best possible life.

This measure provides a quantitative way to understand and compare levels of happiness across different populations, regions, or time periods.

In [9]:
```
plt.figure(figsize=(2, 1))
continuous_countries.groupby('year')['Life Ladder'].mean().plot()
plt.title('Average Life Ladder Score Over Time')
plt.show()
```



In [24]:
```
average_scores = continuous_countries.groupby('Country name')['Life Ladder'].mean().sort_values(ascending=False)
top_10_countries = average_scores.head(10).index.tolist()
top_10_data = continuous_countries[continuous_countries['Country name'].isin(top_10_countries)]
plt.figure(figsize=(6, 3))
sns.boxplot(x='Country name', y='Life Ladder', data=top_10_data, hue='Country name', dodge=False)
plt.title('Life Ladder Scores for Top 10 Countries')
plt.xticks(rotation=45)
plt.legend(title='Country name', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



The median Life Ladder score varies between the countries. Some countries, like Denmark, have a wide interquartile range, which suggests more variation in the Life Ladder scores within the country. There are outliers for some countries (like Israel and Finland), indicating that there are scores that are significantly different from the rest. The central tendency (median) and spread can be quickly compared across the countries. For example, Sweden seems to have one of the highest median Life Ladder scores, whereas Costa Rica has one of the lowest.

In [11]:

```
top_5_countries = continuous_countries.groupby('Country name')['Life Ladder'].mean().nlargest(5).index
top_5_countries_data = continuous_countries[continuous_countries['Country name'].isin(top_5_countries)]
```

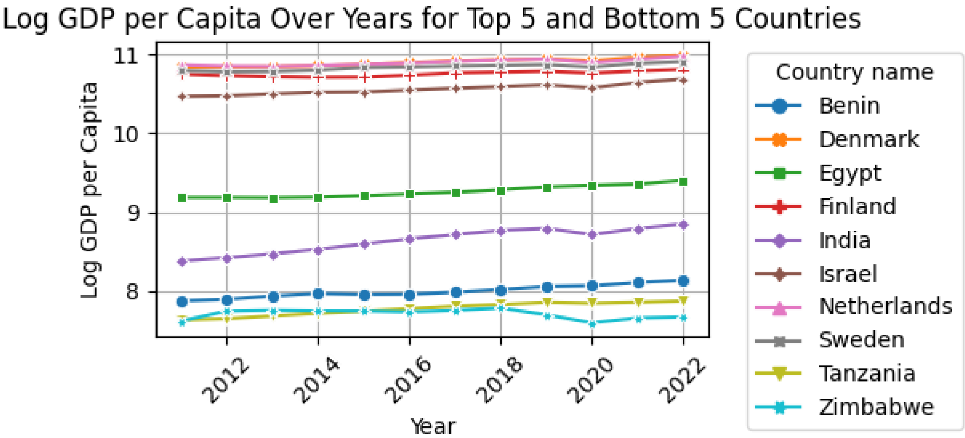
Log GDP per capita

The log GDP per capita allows us to analyze economic growth rate in a standardized manner(in economic). By calculatign the difference in logarithms between two time periods, we can approximate the percentage change in CDP per capita, providing insights into changes in the standard of living over time.

We can see the growth rate against time by log GDP.

In [12]:

```
country_scores = continuous_countries.groupby('Country name')['Life Ladder'].mean()
top_5_countries = country_scores.nlargest(5).index
bottom_5_countries = country_scores.nsmallest(5).index
top_bottom_countries_data = continuous_countries[
    continuous_countries['Country name'].isin(top_5_countries) |
    continuous_countries['Country name'].isin(bottom_5_countries)
]
# Plot the trend of 'Log GDP per capita' over the years for these countries
plt.figure(figsize=(6, 3))
sns.lineplot(x='year', y='Log GDP per capita', hue='Country name', style='Country name', markers=True, dashes=Fa
)
plt.title('Log GDP per Capita Over Years for Top 5 and Bottom 5 Countries')
plt.xlabel('Year')
plt.ylabel('Log GDP per Capita')
plt.legend(title='Country name', bbox_to_anchor=(1.05, 1), loc=2)
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



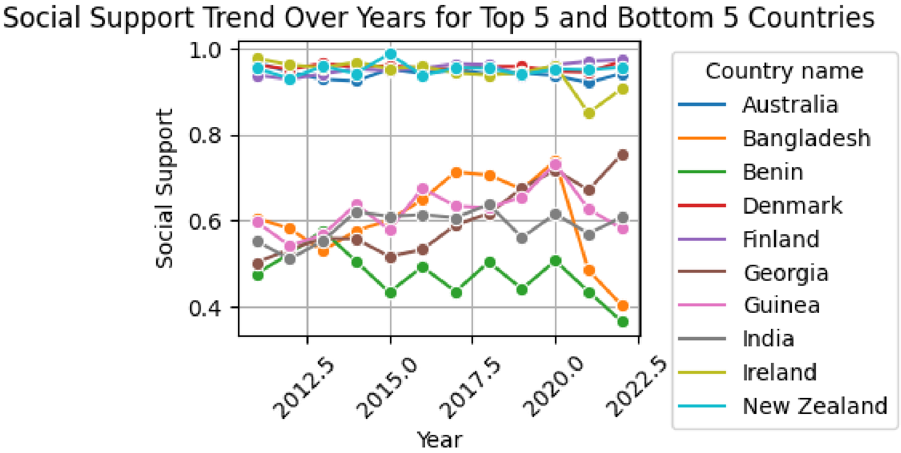
Social Support

The social support is national average of the binary responses, “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?” It is the backbone of human connections, like having a safty net made of friends, family, and community - people who care about you.

It's the comfort you feel when you share your joys and sruggles with them; the helping hand offered when you're facing challenges, whether it's someone listening to you vent, giving ou advice, or lending a hand with pratical tasks.

Socail support is knowing you're not alone and tha tthere are people who have you back, cherring you on through life's ups and downs. it's aout building connections, feeling understood, and knowng you can rely on others when you need them.

```
In [13]: country_social_support = continuous_countries.groupby('Country name')['Social support'].mean()
top_5_countries = country_social_support.nlargest(5).index
bottom_5_countries = country_social_support.nsmallest(5).index
selected_countries = top_5_countries.union(bottom_5_countries)
selected_countries_data = continuous_countries[continuous_countries['Country name'].isin(selected_countries)]
plt.figure(figsize=(5, 3))
sns.lineplot(x='year', y='Social support', hue='Country name', data=selected_countries_data, marker='o'
)
plt.title('Social Support Trend Over Years for Top 5 and Bottom 5 Countries')
plt.xlabel('Year')
plt.ylabel('Social Support')
plt.legend(title='Country name', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



Healthy life expectancy at birth

Healthy life expectancy at birth assesses the average number of years a person is expected to live in good health, without significant illness or disability, from birth onwards.

This metric provides insight into the overall well-being and quality of life within a population, considering both lifespan and healthspan. It helps policymakers allocate resources to promote health and prevent disease effectively.

Freedom to make life choices

The freedom to make life choices is the liberty individuals have to determine their own courses of action without interference from external sources.

This encompasses decisions about education, career, relationships, lifestyle, and more. It is considered fundamental for personal autonomy and happiness, and societies that uphold this freedom tend to be more equitable and prosperous.

Generosity

Generosity is the act of giving to others without expecting anything in return. It involves sharing resources, time, or support for the benefit of others and fosters empathy and compassion within communities.

Perceptions of corruption

Perceptions of corruption refer to how people view the level of dishonesty or misuse of power within institutions or societies. These perceptions affect trust in leadership and institutions and can influence economic development. They are often measured through surveys or indices and require efforts to promote transparency and accountability to address effectively.

Positive affect

Positive affect refers to experiencing emotions like joy, happiness, and contentment, contributing to overall well-being. It's linked to better health, lower stress, and improved cognitive function.

Negative affect

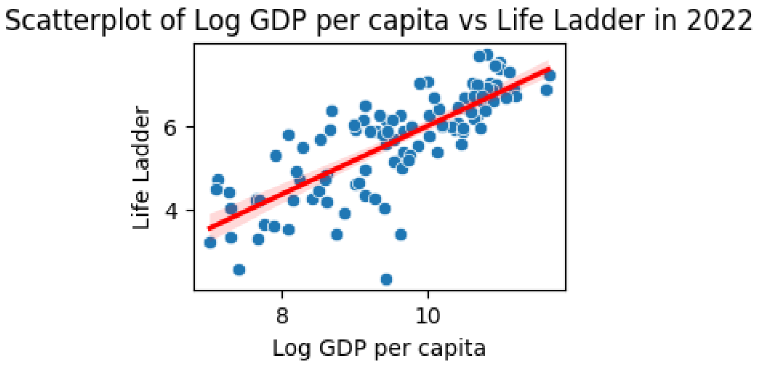
Negative affect refers to experiencing unpleasant emotions like sadness, anger, and anxiety, which can detract from well-being. It's linked to poorer health, increased risk of mental health disorders, and lower quality of life.

In [16]:

```
data2022 = data[data['year'] == 2022]
plt.figure(figsize=(3, 2))
sns.scatterplot(data=data2022, x='Log GDP per capita', y='Life Ladder')
sns.regplot(data=data2022, x='Log GDP per capita', y='Life Ladder', scatter=False, color='red')
plt.title('Scatterplot of Log GDP per capita vs Life Ladder in 2022')
```

Out[16]:

```
Text(0.5, 1.0, 'Scatterplot of Log GDP per capita vs Life Ladder in 2022')
```



GDP per capita (logged) tends to be higher in countries with a higher reported life ladder.

In [17]:

```
correlation_coef = data2022['Log GDP per capita'].corr(data2022['Life Ladder'])
print("Correlation coefficient between 'Log GDP per capita' and 'Life Ladder':", correlation_coef)
```

```
Correlation coefficient between 'Log GDP per capita' and 'Life Ladder': 0.7954744182227044
```

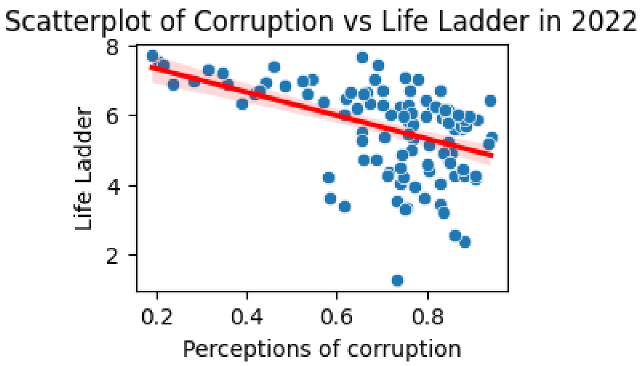
There seems to be a positive correlation between GDP and Happiness with a slope of 0.8.

In [20]:

```
plt.figure(figsize=(3, 2))
sns.scatterplot(data=data2022, x='Perceptions of corruption', y='Life Ladder')
sns.regplot(data=data2022, x='Perceptions of corruption', y='Life Ladder', scatter=False, color='red')
plt.title('Scatterplot of Corruption vs Life Ladder in 2022')
```

Out[20]:

```
Text(0.5, 1.0, 'Scatterplot of Corruption vs Life Ladder in 2022')
```



In [21]:

```
correlation_coef = data2022['Perceptions of corruption'].corr(data2022['Life Ladder'])
print("Correlation coefficient between 'Life Ladder' and 'Perceptions of Corruption':", correlation_coef)
```

```
Correlation coefficient between 'Life Ladder' and 'Perceptions of Corruption': -0.4646060025509597
```

In [18]:

```
data2022 = data[data['year'] == 2022]
```

There seems to be a negative correlation between Happiness and corruption with a slope of -0.5. (not as strong as relationship with GDP)

Analyses / Hypotheses Testing

This following section of the report covers the various analyses and tests used to create and diagnose different models. We first start with a baseline model using every feature in the dataset to predict our dependent variable `Life Ladder`. After that, we regress `Life Ladder` on each variable in its own model to analyze its significance and effects on the happiness score. Finally, we utilize more advanced statistical techniques to improve the model and its predictive capabilities.

Model Building

In [26]:
```
data = pd.read_csv('happy_data.csv')
```

In [27]:
```
filtered_data = data[(data['year'] >= 2011) & (data['year'] <= 2022)].groupby('Country name').filter(lambda x: l
```

In [28]:
```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean')
temp_data = filtered_data[['year', 'Life Ladder', 'Log GDP per capita', 'Social support', 'Healthy life expectan

imputed_data = pd.DataFrame(imputer.fit_transform(temp_data), columns=temp_data.columns)
```

In [29]:
```
imputed_data['year'] = imputed_data.year.apply(int)

test_df = imputed_data[imputed_data.year == 2022]
train_df = imputed_data[imputed_data.year != 2022]
```

In [30]:
```
country_names = filtered_data['Country name'].reset_index().drop('index', axis=1)
train_df['Country name'] = country_names
```

```
/var/folders/4s/my3r02s97f370pt47nlylk_h0000gn/T/ipykernel_15442/2325985746.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  train_df['Country name'] = country_names
```

The above data cleaning and wrangling was done to ensure our data has no missing values and there are no breaks across years. We filtered the countries so that the only countries in our train and test datasets were countries that had data for every year between 2011 and 2022. If a country had at least one missing value in that year range, we did not include that so we can regress on the data. We imputed the missing values in X, our features, with the means of the columns the missing value is in.

Baseline

$$y = \text{Life Ladder}$$
$$X = \text{All Features}$$

In [31]:
```
X = train_df.drop(['year', 'Life Ladder', 'Country name'], axis=1) # ADDED drop year
y = train_df['Life Ladder']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
```

The summary for the above model provides statistics for our baseline model using all the available features. The R^2 value is 0.799, a decent value showing that almost 80% of the variance in our data is represented by the model. However, at $\alpha = 0.05$ significance level, `Negative affect` and `Generosity` are not statistically significant since their p-values are greater than 0.05. There is also strong evidence for multicollinearity in our model from the summary. Although this model might be a good fit of our dataset, there is evidence that it won't predict well on unseen data.

In [32]:
```
def rmse(predictions, targets):
    # Ensure both inputs are numpy arrays
    predictions = np.array(predictions)
    targets = np.array(targets)
    # Calculate squared differences
    differences = predictions - targets
    squared_differences = differences ** 2
    # Calculate mean squared error
    mean_squared_error = np.mean(squared_differences)
    # Calculate RMSE
    rmse_value = np.sqrt(mean_squared_error)
    return rmse_value
```

In [33]:

```
# Training Accuracy (RMSE)
rmse(model.predict(X), y) # can replace this with
```

Out[33]:

```
0.4769128351152439
```

In [34]:

```
test_X = test_df.drop(['year', 'Life Ladder'], axis=1)
test_X = sm.add_constant(test_X)
test_y = test_df['Life Ladder']
```

In [35]:

```
# Test Accuracy (RMSE)
rmse(model.predict(test_X), test_y)
```

Out[35]:

```
0.5138524192133357
```

The **RMSE** (root mean squared error) values measuring the accuracy of the model are fairly promising. A test accuracy of 0.514 means that the model, on average, predicts happiness scores that are 0.514 units away from the actual values.

In [36]:

```
def standardize(x):
    return (x - x.mean()) / x.std()
```

In [37]:

```
def diagnostics(model):
    shapiro_p_value = stats.shapiro(model.resid).pvalue
    print(f"Shapiro-Wilk test p-value: {shapiro_p_value}")
    het_p_value = het_breuschpagan(model.resid, model.model.exog)[3]
    print(f"Breush-Pagan test p-value: {het_p_value}")
    fig, ax = plt.subplots(1, 1, figsize=(5,3))
    sm.qqplot(model.resid, line='r', ax=ax)
    ax.set_title('QQ-plot')
    fig, ax = plt.subplots(1, 1, figsize=(5,3))
    ax.scatter(range(len(model.resid)), standardize(model.resid))
    ax.set_title('Scatterplot of Standardized residuals')
    fig, ax = plt.subplots(1,1,figsize=(5,3))
    ax.set_title('Residuals vs Fitted Values')
    ax.scatter(model.fittedvalues, standardize(model.resid))
    plt.show()
```

The model we have fitted as failed all of the diagnostics tests. While the plots themselves don't seem to be that glaring, running the respective tests like the Shapiro-Wilk test and the Breusch-Pagan test result in very small values. This means that there is statistically significant evidence for the hypotheses that the model's residuals are not normal and the variances are heteroscedastic.

Fitting Linear Regression Models For Each Independent Variable

Here, we fit models for each independent variable to classify the variables that are the most helpful to predict `Life Ladder`. We found that all of the variables are statistically significant, meaning there is evidence that they have an effect on `Life Ladder`. But it was apparent that `Log GDP per capita` was the strongest variable in predictive the dependent variable. The model $y = \beta_0 + \beta_{\text{Log GDP per capita}}$ resulted in an $\mathrm{R}^2 = 0.569$, the highest value of all the individual models. `Social support` came second with an $\mathrm{R}^2 = 0.514$. With this analysis, we understood that all of the features were valid on their own, but throwing them into a model together introduced other sources of error.

In [38]:

```
ftrs = ['Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choice
models = {}
for ftr in ftrs:
    X_sub = train_df[ftr]
    X_sub = sm.add_constant(X_sub)
    mod = sm.OLS(y, X_sub).fit()
    models[ftr] = mod
```

Trying to model `Life Ladder` with each feature on its own results in summaries such as above. All of the p-values were significant with an $\alpha = 0.05$, meaning that there is statistically significant evidence that each of the variables has a meaningful effect on the response variable.

In [39]:

```
for ftr in models:
    print(f'Shapiro-Wilk test p-value for feature {ftr}: {stats.shapiro(models[ftr].resid).pvalue}')
```

```
Shapiro-Wilk test p-value for feature Log GDP per capita: 1.8087470380123705e-05
Shapiro-Wilk test p-value for feature Social support: 3.613380927802723e-09
Shapiro-Wilk test p-value for feature Healthy life expectancy at birth: 1.6359323895542843e-09
Shapiro-Wilk test p-value for feature Freedom to make life choices: 1.5925035812616728e-12
Shapiro-Wilk test p-value for feature Positive affect: 1.4850648922398735e-12
Shapiro-Wilk test p-value for feature Negative affect: 4.5301588869062215e-11
Shapiro-Wilk test p-value for feature Generosity: 2.36359248145801e-11
Shapiro-Wilk test p-value for feature Perceptions of corruption: 8.664470271086636e-15
```

In [40]:

```
def rmse(model):
    return np.sqrt(np.mean(model.resid**2))
```

In [41]:

```
for ftr in models:
    print(f"Feature: {ftr} RMSE: {rmse(models[ftr])}")
```

```
Feature: Log GDP per capita RMSE: 0.6982085251972117
Feature: Social support RMSE: 0.7411591313742049
Feature: Healthy life expectancy at birth RMSE: 0.7569027250829592
Feature: Freedom to make life choices RMSE: 0.911758768844117
Feature: Positive affect RMSE: 0.8900205409440642
Feature: Negative affect RMSE: 1.0015828208293682
Feature: Generosity RMSE: 1.0327974129991921
Feature: Perceptions of corruption RMSE: 0.8975260093130294
```

Looking at the QQ-plots and scatterplots of the residuals, we see that the models by themselves do not fit the assumptions of linear regression. The residuals do not follow a normal distribution and the variances of the residuals are not random. The RMSEs are also higher than for the baseline model, which makes sense since there are less explanatory variables that can be fitted.

Improving Baseline Model

Our main improvements on the baseline will be to narrow down the most important features to be able to predict on unseen data more effectively. We will be testing subsets of the variables to best predict `Life Ladder`.

In [42]:

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

exog = model.model.exog
names = model.params.index
for i in range(1, exog.shape[1]):
    print(f'VIF: {names[i]}: {variance_inflation_factor(exog, i): .3f}')
```

```
VIF: Log GDP per capita:  3.510
VIF: Social support:  2.462
VIF: Healthy life expectancy at birth:  3.195
VIF: Freedom to make life choices:  1.845
VIF: Positive affect:  1.666
VIF: Negative affect:  1.420
VIF: Generosity:  1.256
VIF: Perceptions of corruption:  1.662
```

We first check for multicollinearity to check for the variables that may be highly correlated with each other. Using the Variance Inflation Factor, we found that none of the variables have a VIF greater than 5, which is the threshold for variables that are highly correlated with each other. However, looking at the correlation matrix of these variables, there is high correlation between factors like `Log GDP per capita` and `Healthy life expectancy at birth`.

Using Forward Stepwise Selection for Feature Selection

Now we use forward stepwise selection to grab a subset of the variables to put in the final model based on the AIC factor. We use this factor because it takes into account the number of parameters used in the model, similar to the adjusted-R^2 value.

In [43]:
```
criterion = lambda X: sm.OLS(y, sm.add_constant(X)).fit().aic
```

In [44]:
```
X = X.drop('const', axis=1)
```

In [45]:
```
def add(df, selected_columns, columns, criterion):
    best_criterion = np.inf
    best_column = None
    unselected = list(columns - selected_columns)
    for column in unselected:
        new_columns = list(selected_columns.union({column}))
        current_criterion = criterion(df[new_columns])
        if current_criterion < best_criterion:
            best_criterion = current_criterion
            best_column = column
    return selected_columns.union({best_column}), best_criterion
```

In [46]:
```
def forward(df, criterion):
    selected_columns = set()
    columns = set(X.columns)
    best_criterion = np.inf
    while len(selected_columns) < len(columns):
        potential_columns, current_criterion = add(df, selected_columns, columns, criterion)
        if current_criterion > best_criterion:
            break
        else:
            selected_columns = potential_columns
            best_criterion = current_criterion
    return selected_columns
```

In [47]:
```
selected_columns = forward(X, criterion)
```

In [48]:
```
X_subset = X[selected_columns]
X_subset = sm.add_constant(X_subset)
model_subset = sm.OLS(y, X_subset).fit()
```

```
/var/folders/4s/my3r02s97f370pt47nlylk_h0000gn/T/ipykernel_15442/3816458979.py:1: FutureWarning: Passing a set
as an indexer is deprecated and will raise in a future version. Use a list instead.
  X_subset = X[selected_columns]
```

In [49]:
```
rmse(model_subset)
```

Out[49]:
```
0.4769894324224717
```

The model created using forward stepwise selection only used the columns `Log GDP per capita`, `Social support`, `Healthy life expectancy at birth`, `Positive affect`, `Generosity`, `Freedom to make life choices`, and `Perceptions of corruption`. `Generosity` is still not statistically significant, and there is evidence of multicollinearity.

Removing Generosity from Model Due to P-Value > 0.05

In [50]:
```
selected_columns.remove('Generosity')
```

In [51]:
```
X_no_gen = X[selected_columns]
X_no_gen = sm.add_constant(X_no_gen)
model_no_gen = sm.OLS(y, X_no_gen).fit()
```

```
/var/folders/4s/my3r02s97f370pt47nlylk_h0000gn/T/ipykernel_15442/3221184279.py:1: FutureWarning: Passing a set
as an indexer is deprecated and will raise in a future version. Use a list instead.
  X_no_gen = X[selected_columns]
```

In [52]:
```
rmse(model_no_gen)
```

Out[52]:
```
0.47783536011275707
```

Once we remove `Generosity`, R^2 decreases very slightly, by 0.001 and the RMSE also increases very slightly. There is still evidence of multicollinearity, but now our model contains only statistically significant coefficients.

Remove Healthy Life Expectancy due to high correlation with Log GDP per capita

In [53]:

```
selected_columns.remove('Healthy life expectancy at birth')
```

In [54]:

```
X_less = X[selected_columns]
X_less = sm.add_constant(X_less)
model_less = sm.OLS(y, X_less).fit()
```

```
/var/folders/4s/my3r02s97f370pt47nlylk_h0000gn/T/ipykernel_15442/2164697216.py:1: FutureWarning: Passing a set
as an indexer is deprecated and will raise in a future version. Use a list instead.
  X_less = X[selected_columns]
```

In [55]:

```
rmse(model_less)
```

Out[55]:

```
0.48755101947783624
```

Finally, after removing `Healthy life expectancy at birth`, R^2 again decreases slightly and RMSE increases slightly. However, we get rid of the multicollinearity warning because we remove a factor that had a high correlation with `Log GDP per capita`. Running diagnostics on these models, such as the QQ-plots and scatterplots of the residuals which you can see in our source code, show how these models still don't meet the assumptions needed for regression. Although the new models we fitted don't necessarily meet these assumptions due to the restrictions around the independence and normality of residuals, we get a model that is statistically significant. With real world data, it is hard to force the data to meet the assumptions necessary for statistical analysis and we usually have to work with the data we have.

Interpretation

Based on the resulted RMSE(Root Mean Square Error), the RMSE of the final model decreased to 0.4875 from the RMSE of the baseline model 0.514. The decreasing RMSE illustrates that the final model acheives better performance than the baseline model in predicting the model accuracy.

The final model R squared value is around 0.79, showing that 79% of the Life Ladder variance can be explained by the features 'Log GDP per capita', 'Social support', 'Freedom to make life choices', 'Positive affect', and 'Perceptions of corruption'. With the RMSE 0.485, the final model relatively well predicting the 2022 data. Since there are limited features in the current dataset, as more features included in the original dataset, the prediction can be better.

In the baseline model, the Life Ladder value is predicted based on all features including 'Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices', 'Positive affect', 'Negative affect', 'Generosity', and 'Perceptions of corruption'. In the final model, the features are chosen based on whether the features are statistically significant evidence to the response variable Life Ladder. While performing the OLS regression on each of the features, all of their p-values are less than 0.05, showing that all features are statistically significant to the response variable. It is interesting to see that when fitting all features on the OLS regression, the p-value of 'Negative affect' changes to 0.59, and the p-value of 'Generosity' changes to 0.072, which both are greater than 0.05. The 'Negative affect' and 'Generosity' features therefore are not statistically significant to the response features and removed from the final model. The reason explaining this interesting consequence might due to the low VIF and high RMSE compared with other features. 'Negative affect' has VIF 1.420 and 'Generosity' has VIF 1.256. Both are significantly lower than the others. Having low VIF illustrates that features aren't significant multicollinearity and are relatively independent from each other. These values are suppose to add unique information to the model. Additionally, both features have RMSE value greater than 1 whereas others are less than 1. The difference can be explained as 'Negative affect' and 'Generosity' themselves are poor at explaining the variance of the response variable. The combination of the low VIF and high RMSE can lead to the high p-value. The OLS regression might not be ideal for 'Negative affect' and 'Generosity' features. Perhaps non-linear or high order model can better fit the two features.

After removing both 'Negative affect' and 'Generosity' in the final model, the condition number remains significantly high, which indicates the instability of the regression coefficients with respect to the change in data. Strong multicollinearity can result in the issue. While performing the correlation between the remaining features, the features 'Healthy life expectancy at birth' and 'Log GDP per capita' have strong correlation around 0.81. The high correlation value can lead to large variation in coefficient estimates with small data changes. Information provided by both features can be redundant in model prediction. Thus, to make the condition number lower and produce stable coefficients, the feature 'Healthy life expectancy at birth' is removed as well.

Conclusion and Future work

Based on the rigorous explanations of the introduction, EDA, two OLS regression model, and interpretation of the model results, this paper provides an efficient OLS regression model predicting the 2022 Life Ladder of each country and successfully determining the most important features for world's happiness. The R squared of the final model is around 0.79 with a low RMSE value 0.4875. By comparing the R squared value between each feature and the life ladder result, we find out that Log GDP per capita is most important feature affecting life ladder/ world's happiness. Based on the efficient final model, we are able to accurately predict the world's happiness for each country, not only just the year 2022 but also 2023 and forward. We wish that by using the predicted result, government and citizens can realize how much of their happiness will be in the future. While predicting that they might experience a low happiness, government right now can modify policies like increasing Log GDP per capita to improve their own happiness. We hope that this machine learning model can help every country's citizens to improve their happiness and live on an enjoyable life experience.

Moving on, we plan on improving the model to reach lower RMSE value to better predict world's happiness. Since we use mean imputation method to replace missing values, in future years, we hope to incorporate more complete datasets. In addition, as other features including educational resources and job oppurtunities can play an important role towards world's happiness, we would also like to add more features and figure out their relationship with the world's happiness to improve on the model. We can also utilize non-linear and high-dimensional models in predicting the life ladder. Ideally with all the improvements, we can build an ideal model to improve world's happiness and benefit everyone!

Other factors affecting happiness (external resources)：

> **Gallup** conducts surveys worldwide to measure global happiness and well-being. Their reports often include an analysis of various factors affecting happiness

https://www.gallup.com/home.aspx

"World Unhappier, More Stressed Out Than Ever" by Ray, unveils a sobering reality of heightened negative experiences globally, particularly during the second year of the pandemic. Stress, worry, sadness, and physical pain surged to **new highs**, contributing to an overall decrease in positive emotions.

Afghanistan emerged as a poignant example of this trend, experiencing unprecedented levels of negative experiences and record lows in positive daily emotions. These findings underscore a concerning long-term trajectory, **indicating a decade-long rise in negative emotions worldwide**. While the pandemic exacerbated these trends, deeper societal issues are at play.

The report emphasizes the critical need for policymakers to address the **root causes** of this decline in happiness and prioritize mental health and well-being in their agendas to steer toward a more positive future.

> The Organisation for Economic Co-operation and Development (OECD) provides data on well-being indicators for its member countries, including factors such as income, health, and work-life balance. https://www.oecdbetterlifeindex.org/#/21111111111

OECD took a closer look at the specific factors that affect happiness：

Housing: Adequate housing is fundamental for meeting basic needs and fostering a sense of security and belonging. Beyond providing shelter, it should offer privacy, comfort, and space for family life. Affordability is a key concern.

Income: While money isn't everything, it significantly impacts living standards and well-being. Higher income levels can improve access to education, healthcare, and housing, contributing to overall quality of life.

Jobs: Employment not only provides economic security but also fosters social connections, self-esteem, and skill development. Societies with high employment rates tend to be more stable and healthier.

Community: Social connections play a crucial role in well-being. Spending time with friends and engaging in meaningful relationships can boost positive emotions and reduce negative ones.

Education: Education equips individuals with the skills and knowledge needed to thrive in society. It correlates with better health, civic engagement, and overall happiness.

Environment: A clean and green environment promotes mental well-being and physical health. Access to natural spaces is essential for quality of life and economic productivity. Preserving the environment is vital for future generations.

Civic Engagement: Trust in government and transparency in decision-making are essential for social cohesion. Openness and accountability contribute to better governance and public services.

Health: Good health is paramount for a fulfilling life, enabling access to education, work, and social relationships. It brings numerous benefits, including increased productivity, reduced healthcare costs, and longer life expectancy.

Life Satisfaction: Subjective well-being measures offer insights into personal fulfillment and social conditions. Surveys help gauge life satisfaction and happiness, complementing objective data.

Safety: Personal security is vital for well-being, impacting physical and mental health. Crime can lead to trauma and anxiety, highlighting the importance of safety measures and crime prevention.

Work-Life Balance: Balancing work and personal life is essential for overall well-being, particularly for families. Supportive work policies can help individuals strike a healthier balance between their professional and personal lives.

As Aristotle believed, happiness requires a certain amount of moral virtue and sufficient external material goods. He particularly emphasized the importance of character, defining happiness as "the activity of the soul according to virtue." People’s happiness often depends on the spirit of society: whether it is trustworthy; how generous they are to others; whether the system guarantees individual freedom; and material conditions, such as income and health status.

“To have a society with high average life satisfaction, we need a society with high average eudaimonia.” (Helliwell, Layard, & Sachs, 2023) Not only do individuals and families play a key role in fostering happiness, but institutions of all kinds also play a vital role. The government is the institution primarily responsible for improving the well-being of its citizens as a whole, improving the quality of life of its people through policies, the provision of infrastructure, and the protection of human rights; In addition, it maintains the physical and mental health of the people through healthcare services and mental health support, improves the well-being of students by providing quality educational resources and fostering student interest, and promotes social well-being by providing good working conditions and social activities. Not only that, environmental protection agencies should also strive to protect the natural environment and resources to create good living conditions. Together, these institutions form the pillars of social well-being, working in tandem with each other to achieve the goals of the happiness revolution.

As Helliwell, Layard, & Sachs mentioned at the end, "... the Sustainable Development Goals for 2030 and beyond should put much greater operational and ethical emphasis on well-being. "As The Times progress, people's happiness also increases. But it cannot be denied that there are still some neglected aspects of society that can be done better to improve people's happiness, such as optimizing educational resources and providing more jobs.

Refereneces:

Arias, Elizabeth, Betzaida Tejada-Vera, Kenneth D. Kochanek, and Farida B. Ahmad. "Provisional Life Expectancy Estimates for 2021." August 2022, *Vital Statistics Rapid Release*.

CMHA BC and Anxiety Canada, "Social Support." *Here to Help*, n.d.

Diener, Ed, and Shigehiro Oishi. "The Pursuit of Happiness: Evolutionary Insights and Cultural Perspectives." *Psychological Bulletin*, vol. 126, no. 2, 2000, pp. 267–302.

Gallup. "Understanding How Gallup Uses the Cantril Scale." *Gallup News*, Gallup.

Gallup. Ray, Julie. “World Unhappier, More Stressed out than Ever.” *Gallup.Com*, Gallup.

Growthecon. "Preliminaries: Lines." *Growthecon Study Guide*.

Helliwell, J. F., Huang, H., Norton, M., Goff, L., & Wang, S. (2023). World Happiness, Trust and Social Connections in Times of Crisis. In *World Happiness Report 2023* (11th ed., Chapter 2). Sustainable Development Solutions Network.

Helliwell, J. F., Layard, R., & Sachs, J. D. (2023). The Happiness Agenda: The Next 10 Years. In *World Happiness Report 2023* (11th ed., Chapter 1). Sustainable Development Solutions Network.

Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (Eds.). (2023). *World Happiness Report 2023* (11th ed.). Sustainable Development Solutions Network.

Plys, Ekaterina, and Olivier Desrichard. "Associations Between Positive and Negative Affect and the Way People Perceive Their Health Goals." *Frontiers in Psychology*, Mar 3, 2020.

Ray, Julie. “World Unhappier, More Stressed out than Ever.” *Gallup.Com*, Gallup.

Taking Charge of Your Wellbeing. "Social Support." *University of Minnesota Center for Spirituality & Healing*.

"Corruption Perceptions Index." *Wikipedia*, Wikimedia Foundation, 10 Feb. 2022.

"Freedom of Choice." *Wikipedia*, Wikimedia Foundation, 10 Jan. 2024.

World Bank. "Statistical Capacity Indicators: GDP Per Capita." *World Bank Databank Glossary*, World Bank.

Veenhoven, Ruut. "Happiness and Life Satisfaction as Indicators of Quality of Life: A Review of Current Research." *World Database of Happiness*.